

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 17th July, 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 [243] into this book. We also thank Lukas Gonon, Timo Kröger, Siyu Liang, and Joshua Lee Padgett 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 17th July, 2025

Arnulf Jentzen
Benno Kuckuck
Philippe von Wurstemberger

Contents

Preface	3
Introduction	17
I Artificial neural networks (ANNs)	21
1 Basics on ANNs	23
1.1 Fully-connected feedforward ANNs (vectorized description)	23
1.1.1 Affine functions	24
1.1.2 Vectorized description of fully-connected feedforward ANNs	25
1.1.3 Weight and bias parameters of fully-connected feedforward ANNs	27
1.2 Activation functions	28
1.2.1 Multi-dimensional versions	29
1.2.2 Single hidden layer fully-connected feedforward ANNs	30
1.2.3 Rectified linear unit (ReLU) activation	31
1.2.4 Clipping activation	36
1.2.5 Softplus activation	37
1.2.6 Gaussian error linear unit (GELU) activation	40
1.2.7 Standard logistic activation	41
1.2.8 Swish and sigmoid linear unit (SiLU) activation	44
1.2.9 Hyperbolic tangent activation	46
1.2.10 Softsign activation	47
1.2.11 Leaky rectified linear unit (leaky ReLU) activation	48
1.2.12 Exponential linear unit (ELU) activation	50
1.2.13 Rectified power unit (RePU) activation	52
1.2.14 Sine activation	53
1.2.15 Heaviside activation	54
1.2.16 Softmax activation	55
1.3 Fully-connected feedforward ANNs (structured description)	56
1.3.1 Structured description of fully-connected feedforward ANNs	56
1.3.2 Realizations of fully-connected feedforward ANNs	57

1.3.3	On the connection to the vectorized description	61
1.4	Convolutional ANNs (CNNs)	64
1.4.1	Discrete convolutions	65
1.4.2	Structured description of feedforward CNNs	65
1.4.3	Realizations of feedforward CNNs	65
1.5	Residual ANNs (ResNets)	71
1.5.1	Structured description of fully-connected ResNets	72
1.5.2	Realizations of fully-connected ResNets	72
1.6	Recurrent ANNs (RNNs)	75
1.6.1	Description of RNNs	76
1.6.2	Vectorized description of simple fully-connected RNNs	77
1.6.3	Long short-term memory (LSTM) RNNs	78
1.7	Further types of ANNs	79
1.7.1	ANNS with encoder-decoder architectures: autoencoders	79
1.7.2	Transformers and the attention mechanism	79
1.7.3	Graph neural networks (GNNs)	81
1.7.4	Neural operators	81
2	ANN calculus	83
2.1	Compositions of ANNs	83
2.1.1	Compositions of ANNs	83
2.1.2	Elementary properties of compositions of ANNs	84
2.1.3	Associativity of compositions of ANNs	86
2.1.4	Powers of ANNs	90
2.2	Parallelizations of ANNs	90
2.2.1	Parallelizations of ANNs with the same length	90
2.2.2	ANN representations for the identities	95
2.2.3	Extensions of ANNs	97
2.2.4	Parallelizations of ANNs with different lengths	100
2.3	Scalar multiplications of ANNs	102
2.3.1	Affine transformations as ANNs	102
2.3.2	Scalar multiplications of ANNs	104
2.4	Sums of ANNs with the same length	105
2.4.1	Sums of vectors as ANNs	105
2.4.2	Concatenation of vectors as ANNs	107
2.4.3	Sums of ANNs	109
II	Approximation	113
3	One-dimensional ANN approximation results	115

3.1	Linear interpolation of one-dimensional functions	115
3.1.1	On the modulus of continuity	115
3.1.2	Linear interpolation of one-dimensional functions	117
3.2	Linear interpolation with ANNs	121
3.2.1	Activation functions as ANNs	121
3.2.2	Representations for ReLU ANNs with one hidden neuron	123
3.2.3	ReLU ANN representations for linear interpolations	123
3.3	ANN approximations results for one-dimensional functions	126
3.3.1	Constructive ANN approximation results	126
3.3.2	Convergence rates for the approximation error	130
4	Multi-dimensional ANN approximation results	135
4.1	Approximations through supremal convolutions	135
4.2	ANN representations	138
4.2.1	ANN representations for the 1-norm	138
4.2.2	RePU ANN representations for the identity	140
4.2.3	ANN representations for maxima	143
4.2.4	ANN representations for maximum convolutions	148
4.3	ANN approximations results for multi-dimensional functions	152
4.3.1	Constructive ANN approximation results	152
4.3.2	Covering number estimates	152
4.3.3	Convergence rates for the approximation error	155
4.4	Refined ANN approximations results for multi-dimensional functions	164
4.4.1	Rectified clipped ANNs	164
4.4.2	Embedding ANNs in larger architectures	165
4.4.3	Approximation through ANNs with variable architectures	172
4.4.4	Refined convergence rates for the approximation error	175
III	Optimization	181
5	Optimization through gradient flow (GF) trajectories	183
5.1	Introductory comments for the training of ANNs	183
5.2	Basics for GFs	185
5.2.1	GF ordinary differential equations (ODEs)	185
5.2.2	Direction of negative gradients	186
5.3	Regularity properties for ANNs	192
5.3.1	On the differentiability of compositions of parametric functions	192
5.3.2	On the differentiability of realizations of ANNs	193
5.4	Loss functions	195
5.4.1	Absolute error loss	195

5.4.2	Mean squared error loss	196
5.4.3	Huber error loss	198
5.4.4	Cross-entropy loss	201
5.4.5	Kullback–Leibler divergence loss	205
5.5	GF optimization in the training of ANNs	209
5.6	Critical points in optimization problems	210
5.6.1	Local and global minimizers	210
5.6.2	Local and global maximizers	211
5.6.3	Critical points	211
5.7	Conditions on objective functions in optimization problems	213
5.7.1	Convexity	214
5.7.2	Strict convexity	216
5.7.3	Monotonicity	217
5.7.4	Subgradients	220
5.7.5	Strong convexity	220
5.7.6	Coercivity	226
5.8	Lyapunov-type functions for GFs	229
5.8.1	Gronwall differential inequalities	229
5.8.2	Lyapunov-type functions for ODEs	230
5.8.3	On Lyapunov-type functions and coercivity-type conditions	231
5.8.4	On a linear growth condition	233
5.9	Optimization through flows of ODEs	234
5.9.1	Approximation of local minimum points through GFs	234
5.9.2	Existence and uniqueness of solutions of ODEs	236
5.9.3	Approximation of local minimum points through GFs revisited	239
5.9.4	Approximation error with respect to the objective function	240
6	Deterministic gradient descent (GD) optimization methods	241
6.1	GD optimization	241
6.1.1	GD optimization in the training of ANNs	242
6.1.2	Euler discretizations for GF ODEs	243
6.1.3	Lyapunov-type stability for GD optimization	245
6.1.4	Error analysis for GD optimization	249
6.2	Explicit midpoint optimization	270
6.2.1	Explicit midpoint discretizations for GF ODEs	271
6.3	Momentum optimization	274
6.3.1	Alternative definitions	275
6.3.2	Relationships between different definitions	278
6.3.3	Representations for momentum optimization	286
6.3.4	Bias-adjusted momentum optimization	289
6.3.5	Error analysis for momentum optimization	291

6.3.6	Numerical comparisons for GD and momentum optimization	307
6.4	Nesterov accelerated momentum optimization	312
6.4.1	Alternative definitions	312
6.4.2	Relationships between different definitions	315
6.4.3	Bias-adjusted Nesterov accelerated momentum optimization	323
6.4.4	Shifted representations	324
6.4.5	Simplified Nesterov accelerated momentum optimization	333
6.5	Adaptive gradient (Adagrad) optimization	334
6.6	Root mean square propagation (RMSprop) optimization	336
6.6.1	Representations of the mean square terms in RMSprop	337
6.6.2	Bias-adjusted RMSprop optimization	338
6.7	Adadelta optimization	341
6.8	Adaptive moment estimation (Adam) optimization	342
6.8.1	Adamax optimization	343
6.9	Nesterov accelerated adaptive moment estimation (Nadam) optimization .	344
6.9.1	Simplified Nadam optimization	345
6.9.2	Nadamax optimization	347
6.10	Adam with decoupled weight decay (AdamW) optimization	348
6.10.1	Adam with L^2 -regularization optimization	349
6.11	Shampoo optimization	350
6.12	Momentum orthogonalized by Newton-Schulz (Muon) optimization . .	352
6.13	AMSGrad optimization	355
6.14	Compact summary of deterministic GD optimization methods	356
7	Stochastic gradient descent (SGD) optimization methods	361
7.1	Introductory comments for the training of ANNs with SGD	361
7.2	SGD optimization	363
7.2.1	SGD optimization in the training of ANNs	364
7.2.2	Non-convergence of SGD for not appropriately decaying learning rates	374
7.2.3	Convergence rates for SGD for quadratic objective functions	385
7.2.4	Convergence rates for SGD for coercive objective functions	388
7.2.5	Measurability of SGD processes	389
7.3	Explicit midpoint optimization	390
7.4	Momentum optimization	393
7.4.1	Alternative definitions	396
7.4.2	Bias-adjusted momentum optimization	399
7.5	Nesterov accelerated momentum optimization	401
7.5.1	Alternative definitions	403
7.5.2	Bias-adjusted Nesterov accelerated momentum optimization	406
7.5.3	Shifted representations	408
7.5.4	Simplified Nesterov accelerated momentum optimization	413

7.6	Adagrad optimization	414
7.7	RMSprop optimization	417
7.7.1	Bias-adjusted RMSprop optimization	419
7.8	Adadelta optimization	421
7.9	Adam optimization	423
7.9.1	Adamax optimization	436
7.10	Nadam optimization	438
7.10.1	Simplified Nadam optimization	440
7.10.2	Nadamax optimization	441
7.11	AdamW optimization	442
7.11.1	Adam with L^2 -regularization optimization	444
7.12	Shampoo optimization	445
7.13	Muon optimization	446
7.14	AMSGrad optimization	449
7.15	Compact summary of SGD optimization methods	450
8	Backpropagation	457
8.1	Backpropagation for parametric functions	457
8.2	Backpropagation for ANNs	462
9	Kurdyka–Łojasiewicz (KL) inequalities	469
9.1	Standard KL functions	469
9.2	Convergence analysis using standard KL functions (regular regime)	470
9.3	Standard KL inequalities for monomials	473
9.4	Standard KL inequalities around non-critical points	474
9.5	Standard KL inequalities with increased exponents	475
9.6	Standard KL inequalities for coercive-type functions	476
9.7	Standard KL inequalities for one-dimensional polynomials	478
9.8	Power series and analytic functions	482
9.9	Standard KL inequalities for one-dimensional analytic functions	485
9.10	Standard KL inequalities for analytic functions	491
9.11	Counterexamples	491
9.12	Convergence analysis for solutions of GF ODEs	494
9.12.1	Abstract local convergence results for GF processes	494
9.12.2	Abstract global convergence results for GF processes	500
9.13	Convergence analysis for GD processes	504
9.13.1	One-step descent property for GD processes	505
9.13.2	Abstract local convergence results for GD processes	506
9.14	On the analyticity of realization functions of ANNs	512
9.15	Standard KL inequalities for empirical risks in the training of ANNs with analytic activation functions	515

9.16	Generalized KL-inequalities	518
9.16.1	Fréchet subgradients and limiting Fréchet subgradients	518
9.16.2	Non-smooth slope	524
9.16.3	Generalized KL functions	524
9.17	Non-convergence for stochastic gradient descent	525
10	ANNs with batch normalization	527
10.1	Batch normalization (BN)	527
10.2	Structured description of ANNs with BN (training)	530
10.3	Realizations of fully-connected feedforward ANNs with BN (training) . . .	531
10.4	Structured description of ANNs with BN (inference)	532
10.5	Realizations of ANNs with BN (inference)	532
10.6	On the connection between BN for training and BN for inference	533
11	Optimization through random initializations	535
11.1	Analysis of the optimization error	535
11.1.1	The complementary distribution function formula	535
11.1.2	Estimates for the optimization error involving complementary distribution functions	536
11.2	Strong convergences rates for the optimization error	537
11.2.1	Properties of the gamma and the beta function	537
11.2.2	Product measurability of continuous random fields	542
11.2.3	Strong convergences rates for the optimization error	545
11.3	Strong convergences rates for the optimization error involving ANNs . . .	548
11.3.1	Local Lipschitz continuity estimates for the parametrization functions of ANNs	548
11.3.2	Strong convergences rates for the optimization error involving ANNs	556
IV	Generalization	559
12	Probabilistic generalization error estimates	561
12.1	Concentration inequalities for random variables	561
12.1.1	Markov's inequality	561
12.1.2	A first concentration inequality	562
12.1.3	Moment-generating functions	564
12.1.4	Chernoff bounds	565
12.1.5	Hoeffding's inequality	567
12.1.6	A strengthened Hoeffding's inequality	573
12.2	Covering number estimates	574
12.2.1	Entropy quantities	574

12.2.2	Inequalities for packing entropy quantities in metric spaces	575
12.2.3	Inequalities for covering entropy quantities in metric spaces	577
12.2.4	Inequalities for entropy quantities in finite-dimensional vector spaces	580
12.3	Empirical risk minimization	587
12.3.1	Concentration inequalities for random fields	587
12.3.2	Uniform estimates for the statistical learning error	592
13	Strong generalization error estimates	597
13.1	Monte Carlo estimates	597
13.2	Uniform strong error estimates for random fields	600
13.3	Strong convergence rates for the generalisation error	605
V	Composed error analysis	613
14	Overall error decomposition	615
14.1	Bias-variance decomposition	615
14.1.1	Risk minimization for measurable functions	616
14.2	Overall error decomposition	618
15	Composed error estimates	621
15.1	Full strong error analysis for the training of ANNs	621
15.2	Full strong error analysis with optimization via SGD with random initializations	630
VI	Deep learning for partial differential equations (PDEs)	635
16	Physics-informed neural networks (PINNs)	637
16.1	Reformulation of PDE problems as stochastic optimization problems . . .	638
16.2	Derivation of PINNs and deep Galerkin methods (DGMs)	639
16.3	Implementation of PINNs	641
16.4	Implementation of DGMs	644
17	Deep Kolmogorov methods (DKMs)	649
17.1	Stochastic optimization problems for expectations of random variables .	649
17.2	Stochastic optimization problems for expectations of random fields . . .	650
17.3	Feynman–Kac formulas	652
17.3.1	Feynman–Kac formulas providing existence of solutions	652
17.3.2	Feynman–Kac formulas providing uniqueness of solutions	658
17.4	Reformulation of PDE problems as stochastic optimization problems . . .	663
17.5	Derivation of DKMs	666
17.6	Implementation of DKMs	668

Contents

18 Further deep learning methods for PDEs	671
18.1 Deep learning methods based on strong formulations of PDEs	671
18.2 Deep learning methods based on weak formulations of PDEs	672
18.3 Deep learning methods based on stochastic representations of PDEs	673
18.4 Error analyses for deep learning methods for PDEs	675
Index of abbreviations	677
List of figures	679
List of source codes	681
List of definitions	683
List of exercises	689
Bibliography	693

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

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

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

In the context of a learning problem of the type (1) the objective then is to approximately compute the output $\mathcal{E}(x_{M+1})$ of the $(M+1)$ -th input data x_{M+1} without using explicit knowledge of the function $\mathcal{E}: \mathbb{R}^d \rightarrow \mathbb{R}$ but instead by using the knowledge of the M input-output data pairs

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

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

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

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

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

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

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

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

We think of the set

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

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

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

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

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

or one could think of (7) as the parametrization associated to trigonometric polynomials. However, in the context of *deep supervised learning* one neither chooses (7) as parametrization of polynomials nor as parametrization of trigonometric polynomials, but instead one chooses (7) as a parametrization associated to *deep ANNs*. In Chapter 1 in Part I we present different types of such deep ANN parametrization functions in all mathematical details.

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

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

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

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

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

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

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 $\vartheta \in \mathbb{R}^d$ of (10) in the sense that $\mathcal{L}(\vartheta) \approx \inf_{\theta \in \mathbb{R}^d} \mathcal{L}(\theta)$, which is, however, typically not a minimizer of (10) in the sense that $\mathcal{L}(\vartheta) = \inf_{\theta \in \mathbb{R}^d} \mathcal{L}(\theta)$ (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).

Contents

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, 40, 42, 62, 65, 100, 171, 189, 199, 388, 394, 410, 452] and the references therein. The specific presentation of Sections 1.1 and 1.3 is based on [19, 20, 25, 166, 187].

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 1.1 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).

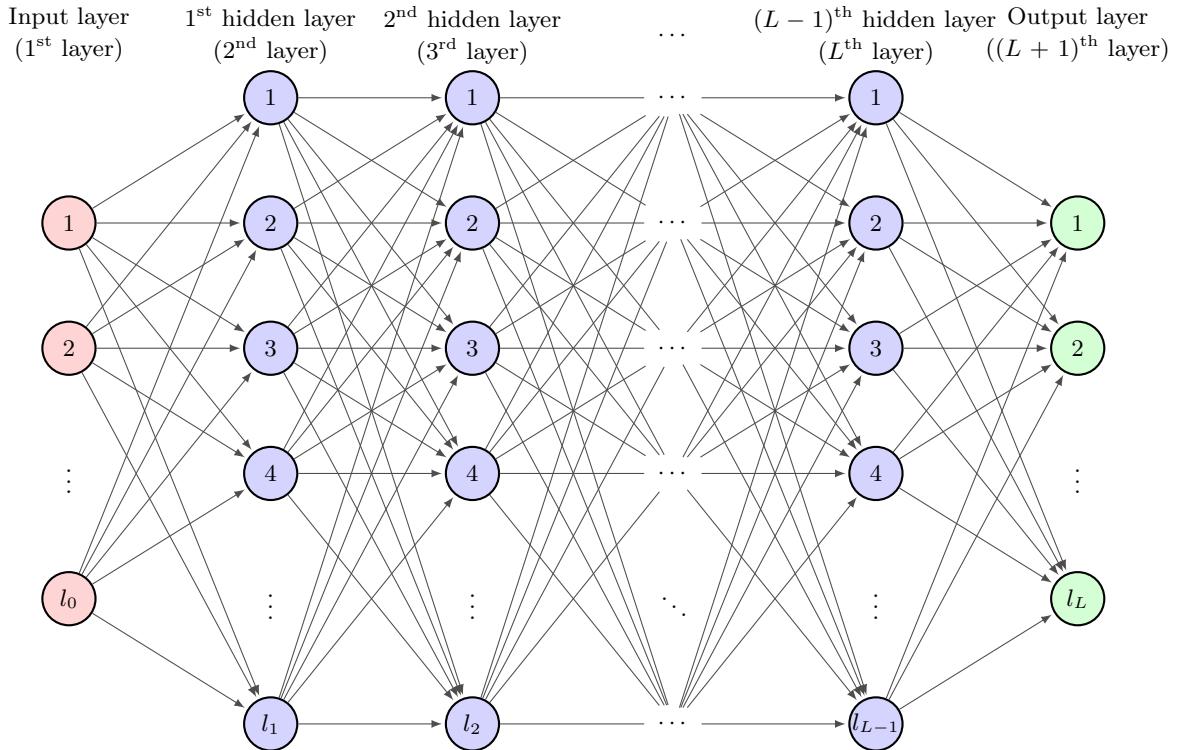


Figure 1.1: Graphical illustration of a fully-connected feedforward **ANN** consisting of $L \in \mathbb{N}$ affine transformations (i.e., consisting of $L + 1$ layers: one input layer, $L - 1$ hidden layers, and one output layer) with $l_0 \in \mathbb{N}$ neurons on the input layer (i.e., with l_0 -dimensional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

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

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

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

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

(cf. Definition 1.1.1).

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

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

The proof for Example 1.1.2 is thus complete. \square

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

1.1.2 Vectorized description of fully-connected feedforward ANNs

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

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

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

$\mathcal{N}_{\Psi_1, \Psi_2, \dots, \Psi_L}^{\theta, l_0} : \mathbb{R}^{l_0} \rightarrow \mathbb{R}^{l_L}$ the function given by

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

and we call $\mathcal{N}_{\Psi_1, \Psi_2, \dots, \Psi_L}^{\theta, l_0}$ the realization function of the fully-connected feedforward ANN associated to θ with $L + 1$ layers with dimensions (l_0, l_1, \dots, l_L) and activation functions $(\Psi_1, \Psi_2, \dots, \Psi_L)$ (we call $\mathcal{N}_{\Psi_1, \Psi_2, \dots, \Psi_L}^{\theta, l_0}$ the realization of the fully-connected feedforward ANN associated to θ with $L + 1$ layers with dimensions (l_0, l_1, \dots, l_L) and activations $(\Psi_1, \Psi_2, \dots, \Psi_L)$) (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 $\Psi : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ satisfy for all $x = (x_1, x_2) \in \mathbb{R}^2$ that

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

Then

$$(\mathcal{N}_{\Psi, \text{id}_{\mathbb{R}}}^{\theta, 1})(2) = 12 \quad (1.7)$$

(cf. Definition 1.1.3).

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

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

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

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

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

Prove or disprove the following statement: It holds that

$$(\mathcal{N}_{\Psi, \text{id}_{\mathbb{R}}}^{\theta, 1})(-1) = -1 \quad (1.10)$$

(cf. Definition 1.1.3).

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

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

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

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

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

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

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

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

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

1.1.3 Weight and bias parameters of fully-connected feedforward ANNs

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

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

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

$$W_k = \begin{pmatrix} \theta_{v_{k-1}+1} & \theta_{v_{k-1}+2} & \dots & \theta_{v_{k-1}+l_{k-1}} \\ \theta_{v_{k-1}+l_{k-1}+1} & \theta_{v_{k-1}+l_{k-1}+2} & \dots & \theta_{v_{k-1}+2l_{k-1}} \\ \theta_{v_{k-1}+2l_{k-1}+1} & \theta_{v_{k-1}+2l_{k-1}+2} & \dots & \theta_{v_{k-1}+3l_{k-1}} \\ \vdots & \vdots & \vdots & \vdots \\ \theta_{v_{k-1}+(l_{k-1}-1)l_{k-1}+1} & \theta_{v_{k-1}+(l_{k-1}-1)l_{k-1}+2} & \dots & \theta_{v_{k-1}+l_k l_{k-1}} \end{pmatrix} \quad (1.14)$$

$$\text{and } b_k = \underbrace{(\theta_{v_{k-1}+l_k l_{k-1}+1}, \theta_{v_{k-1}+l_k l_{k-1}+2}, \dots, \theta_{v_{k-1}+l_k l_{k-1}+l_k})}_{\text{bias parameters}}, \quad (1.15)$$

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

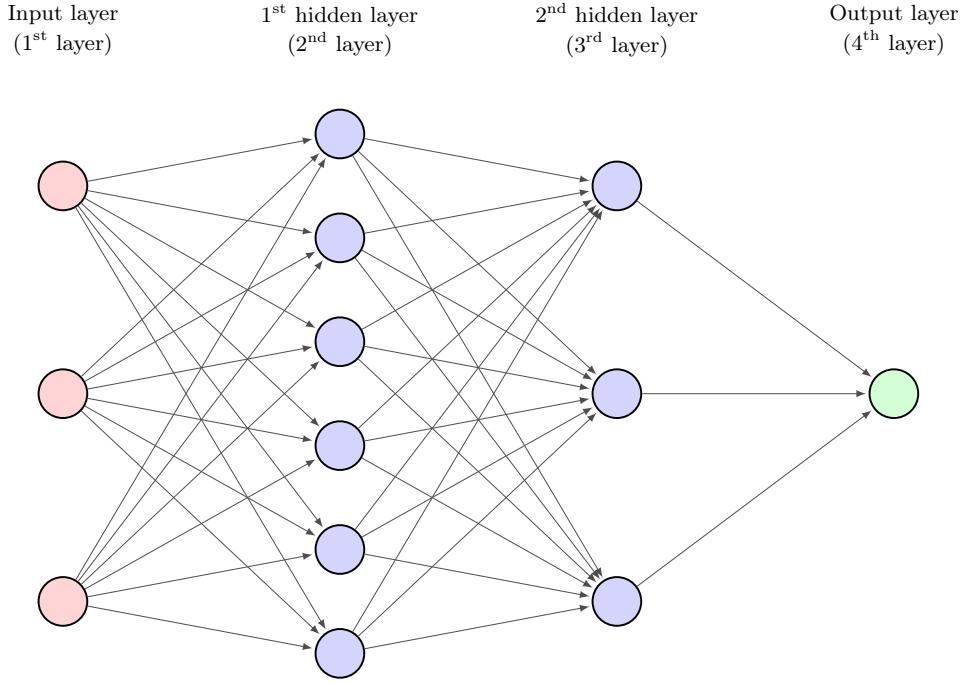


Figure 1.2: Graphical illustration of an ANN. The ANN has 2 hidden layers and length $L = 3$ with 3 neurons in the input layer (corresponding to $l_0 = 3$), 6 neurons in the first hidden layer (corresponding to $l_1 = 6$), 3 neurons in the second hidden layer (corresponding to $l_2 = 3$), and one neuron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

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

and

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

(cf. Definitions 1.1.1 and 1.1.3).

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 $T \in \mathbb{N}$, $d_1, d_2, \dots, d_T \in \mathbb{N}$ and let $\psi: \mathbb{R} \rightarrow \mathbb{R}$ be a function. Then we denote by*

$$\mathfrak{M}_{\psi, d_1, d_2, \dots, d_T}: \mathbb{R}^{d_1 \times d_2 \times \dots \times d_T} \rightarrow \mathbb{R}^{d_1 \times d_2 \times \dots \times d_T} \quad (1.17)$$

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

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

and we call $\mathfrak{M}_{\psi, d_1, d_2, \dots, d_T}$ the $d_1 \times d_2 \times \dots \times d_T$ -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*

$$A = ((1 \ -1), (-2 \ 2), (3 \ -3)) \quad (1.19)$$

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

$$\mathfrak{M}_{\psi, 3, 1, 2}(A) = ((1 \ 1), (4 \ 4), (9 \ 9)) \quad (1.20)$$

Proof for Example 1.2.2. Note that (1.18) establishes (1.20). The proof for Example [1.2.2](#) is thus complete. \square

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

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

and let $\psi: \mathbb{R} \rightarrow \mathbb{R}$ 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 $\theta = (\theta_1, \theta_2, \dots, \theta_{14}) \in \mathbb{R}^{14}$ satisfy

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

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

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

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

1.2.2 Single hidden layer fully-connected feedforward ANNs

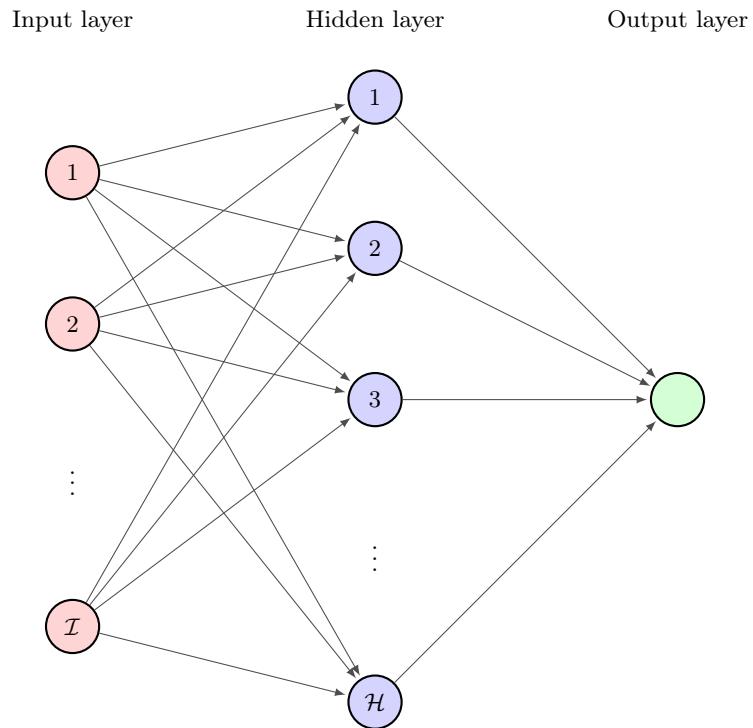


Figure 1.3: Graphical illustration of a fully-connected feedforward ANN consisting of two affine transformations (i.e., consisting of 3 layers: one input layer, one hidden layer, and one output layer) with $\mathcal{I} \in \mathbb{N}$ neurons on the input layer (i.e., with \mathcal{I} -dimensional input 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 $\mathcal{I}, \mathcal{H} \in \mathbb{N}$, $\theta = (\theta_1, \dots, \theta_{\mathcal{H}\mathcal{I}+2\mathcal{H}+1}) \in \mathbb{R}^{\mathcal{H}\mathcal{I}+2\mathcal{H}+1}$, $x = (x_1, \dots, x_{\mathcal{I}}) \in \mathbb{R}^{\mathcal{I}}$ and let $\psi: \mathbb{R} \rightarrow \mathbb{R}$ be a function. Then*

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

(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

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

The proof of Lemma 1.2.3 is thus complete. \square

1.2.3 Rectified linear unit (ReLU) activation

In this subsection we formulate the **ReLU** function which is one of the most frequently used activation functions in deep learning applications (cf., for example, LeCun et al. [277]).

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

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

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

```

1 import matplotlib.pyplot as plt
2
3 def setup_axis(xlim, ylim):
4     _, ax = plt.subplots()
5
6     ax.set_aspect("equal")
7     ax.set_xlim(xlim)
8     ax.set_ylim(ylim)
9     ax.spines["left"].set_position("zero")
10    ax.spines["bottom"].set_position("zero")

```

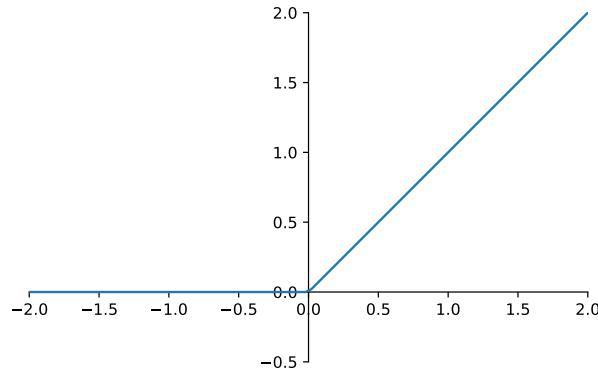


Figure 1.4 ([plots/relu.pdf](#)): A plot of the ReLU activation function

```

11     ax.spines["right"].set_color("none")
12     ax.spines["top"].set_color("none")
13     for s in ax.spines.values():
14         s.set_zorder(0)
15
16     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

```

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((-2,2), (-.5,2))
7
8 x = np.linspace(-2, 2, 100)
9
10 ax.plot(x, tf.keras.activations.relu(x))
11
12 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 $\mathfrak{R}_d: \mathbb{R}^d \rightarrow \mathbb{R}^d$ the function given by*

$$\mathfrak{R}_d = \mathfrak{M}_{\mathfrak{r},d} \quad (1.27)$$

1.2. Activation functions

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

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

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

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

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

The proof of Lemma 1.2.6 is thus complete. \square

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

$$(\mathcal{N}_{\mathfrak{R}_{l_1}, \mathfrak{R}_{l_2}, \dots, \mathfrak{R}_{l_H}, \text{id}_{\mathbb{R}}}^{\theta, 1})(x) = x \quad (1.30)$$

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

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

(cf. Definitions 1.1.3 and 1.2.5).

Proof of Lemma 1.2.7. Note that (1.1), (1.5), and (1.26) imply that for all $x \in \mathbb{R}$ it holds that

$$\begin{aligned} (\mathcal{N}_{\mathfrak{R}_2, \text{id}_{\mathbb{R}}}^{\theta, 1})(x) &= \max\{x + 0, 0\} - \max\{-x + 0, 0\} + 0 \\ &= \max\{x, 0\} - \max\{-x, 0\} = x \end{aligned} \quad (1.32)$$

(cf. Definitions 1.1.3 and 1.2.5). The proof of Lemma 1.2.7 is thus complete. \square

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

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

(cf. Definitions 1.1.3 and 1.2.5).

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

$$(\mathcal{N}_{\mathfrak{R}_{l_1}, \mathfrak{R}_{l_2}, \dots, \mathfrak{R}_{l_H}, \text{id}_{\mathbb{R}}}^{\theta, 1})(x) = e^x \quad (1.34)$$

(cf. Definitions 1.1.3 and 1.2.5).

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

$$(\mathcal{N}_{\mathfrak{R}_{l_1}, \mathfrak{R}_{l_2}, \dots, \mathfrak{R}_{l_H}, \text{id}_{\mathbb{R}}}^{\theta, 2})(x, y) = \max\{x, y\} \quad (1.35)$$

(cf. Definitions 1.1.3 and 1.2.5).

Exercise 1.2.7. 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

$$(\mathcal{N}_{\mathfrak{R}_{l_1}, \mathfrak{R}_{l_2}, \text{id}_{\mathbb{R}}}^{\theta, 1})(x) = x \quad (1.36)$$

(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.

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*

$$(\mathcal{N}_{\mathfrak{R}_2, \mathfrak{R}_2, \text{id}_{\mathbb{R}}}^{\theta, 1})(x) = x \quad (1.37)$$

(cf. Definitions 1.1.3 and 1.2.5).

Proof of Lemma 1.2.8. Observe that (1.1), (1.5), and (1.26) prove that for all $x \in \mathbb{R}$ it holds that

$$\begin{aligned} (\mathcal{N}_{\mathfrak{R}_2, \mathfrak{R}_2, \text{id}_{\mathbb{R}}}^{\theta, 1})(x) &= (1 \ -1) \mathfrak{R}_2 \left(\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \mathfrak{R}_2 \left(\begin{pmatrix} 1 \\ -1 \end{pmatrix} x + \begin{pmatrix} 0 \\ 0 \end{pmatrix} \right) + \begin{pmatrix} 0 \\ 0 \end{pmatrix} \right) + 0 \\ &= (1 \ -1) \mathfrak{R}_2 \left(\begin{pmatrix} \max\{x, 0\} - \max\{-x, 0\} \\ -\max\{x, 0\} + \max\{-x, 0\} \end{pmatrix} \right) \\ &= (1 \ -1) \mathfrak{R}_2 \begin{pmatrix} x \\ x \end{pmatrix} \\ &= \max\{x, 0\} - \max\{-x, 0\} = x. \end{aligned} \quad (1.38)$$

(cf. Definitions 1.1.3 and 1.2.5). The proof of Lemma 1.2.8 is thus complete. \square

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

$$(\mathcal{N}_{\mathfrak{R}_{l_1}, \mathfrak{R}_{l_2}, \dots, \mathfrak{R}_{l_H}, \text{id}_{\mathbb{R}}}^{\theta, 3})(x, y, z) = \max\{x, y, z\} \quad (1.39)$$

(cf. Definitions 1.1.3 and 1.2.5).

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

$$(\mathcal{N}_{\mathfrak{R}_{l_1}, \mathfrak{R}_{l_2}, \dots, \mathfrak{R}_{l_H}, \text{id}_{\mathbb{R}}}^{\theta, k})(x_1, x_2, \dots, x_k) = \max\{x_1, x_2, \dots, x_k\} \quad (1.40)$$

(cf. Definitions 1.1.3 and 1.2.5).

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

$$(\mathcal{N}_{\mathfrak{R}_{l_1}, \mathfrak{R}_{l_2}, \dots, \mathfrak{R}_{l_H}, \text{id}_{\mathbb{R}}}^{\theta, 1})(x) = \max\{x, \frac{x}{2}\} \quad (1.41)$$

(cf. Definitions 1.1.3 and 1.2.5).

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

$$(\mathcal{N}_{\mathfrak{R}_l, \text{id}_{\mathbb{R}}}^{\theta, 1})(x) = \begin{cases} 1 & : x \leq 2 \\ x - 1 & : 2 < x \leq 3 \\ 5 - x & : 3 < x \leq 4 \\ 1 & : x > 4 \end{cases} \quad (1.42)$$

(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

$$(\mathcal{N}_{\mathfrak{R}_l, \text{id}_{\mathbb{R}}}^{\theta, 1})(x) = \begin{cases} -2 & : x \leq 1 \\ 2x - 4 & : 1 < x \leq 3 \\ 2 & : x > 3 \end{cases} \quad (1.43)$$

(cf. Definitions 1.1.3 and 1.2.5).

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

$$(\mathcal{N}_{\mathfrak{R}_{l_1}, \mathfrak{R}_{l_2}, \dots, \mathfrak{R}_{l_H}, \text{id}_{\mathbb{R}}}^{\theta, 1})(x) = \begin{cases} 0 & : x \leq 1 \\ x - 1 & : 1 \leq x \leq 2 \\ 1 & : x \geq 2 \end{cases} \quad (1.44)$$

(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

$$(\mathcal{N}_{\mathfrak{R}_l, \text{id}_{\mathbb{R}}}^{\theta, 1})(x) = x^2 \quad (1.45)$$

(cf. Definitions 1.1.3 and 1.2.5).

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

$$\sup_{x \in [-3, -2]} |(\mathcal{N}_{\mathfrak{R}_{l_1}, \mathfrak{R}_{l_2}, \dots, \mathfrak{R}_{l_H}, \text{id}_{\mathbb{R}}}^{\theta, 1})(x) - (x + 2)^2| \leq \frac{1}{4} \quad (1.46)$$

(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 $\mathbf{c}_{u,v}: \mathbb{R} \rightarrow \mathbb{R}$ the function which satisfies for all $x \in \mathbb{R}$ that*

$$\mathbf{c}_{u,v}(x) = \max\{u, \min\{x, v\}\}. \quad (1.47)$$

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

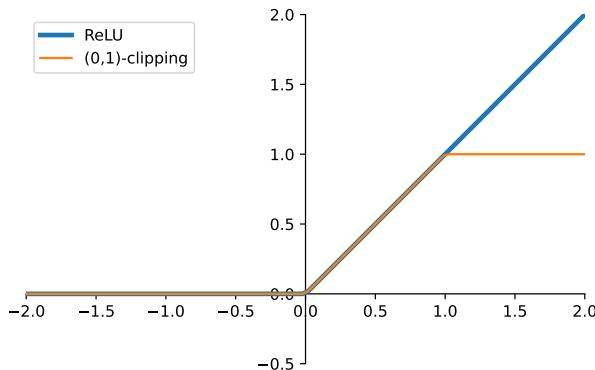


Figure 1.5 ([plots/clipping.pdf](#)): A plot of the $(0, 1)$ -clipping activation function and the ReLU activation function

```

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((-2,2), (-.5,2))
7
8 x = np.linspace(-2, 2, 100)
9
10 ax.plot(x, tf.keras.activations.relu(x), linewidth=3, label='ReLU')
11 ax.plot(x, tf.keras.activations.relu(x, max_value=1),
12         label='(0,1)-clipping')
13 ax.legend()
14
15 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 $\mathfrak{C}_{u,v,d}: \mathbb{R}^d \rightarrow \mathbb{R}^d$ the function given by*

$$\mathfrak{C}_{u,v,d} = \mathfrak{M}_{\mathfrak{c}_{u,v,d}} \quad (1.48)$$

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 $a: \mathbb{R} \rightarrow \mathbb{R}$ is the function from \mathbb{R} to \mathbb{R} which satisfies for all $x \in \mathbb{R}$ that*

$$a(x) = \ln(1 + \exp(x)). \quad (1.49)$$

```

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((-4,4), (-.5,4))
7
8 x = np.linspace(-4, 4, 100)
9
10 ax.plot(x, tf.keras.activations.relu(x), label='ReLU')
11 ax.plot(x, tf.keras.activations.softplus(x), label='softplus')
12 ax.legend()
13
14 plt.savefig("../plots/softplus.pdf", bbox_inches='tight')

```

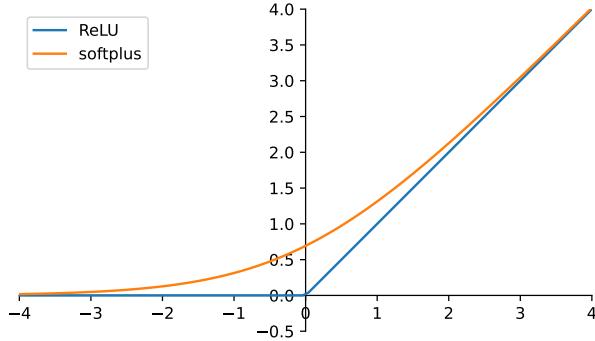


Figure 1.6 ([plots/softplus.pdf](#)): A plot of the softplus activation function and the **ReLU** activation function

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 $\lim_{x \rightarrow -\infty} a(x) = 0$,*
 - (iii) *it holds that $\lim_{x \rightarrow \infty} a(x) = \infty$, and*
 - (iv) *it holds that $a(0) = \ln(2)$*
- (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

$$\begin{aligned} x &= \ln(\exp(x)) \leq \ln(1 + \exp(x)) = \ln(\exp(0) + \exp(x)) \\ &\leq \ln(\exp(x) + \exp(x)) = \ln(2 \exp(x)) \leq \ln(\exp(1) \exp(x)) \\ &= \ln(\exp(x + 1)) = x + 1. \end{aligned} \tag{1.50}$$

The proof of Lemma 1.2.12 is thus complete. □

Note that Lemma 1.2.12 ensures that $\mathfrak{s}(0) = \ln(2) = 0.693\dots$ (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 $A: \mathbb{R}^d \rightarrow \mathbb{R}^d$ be a function. Then A is the d -dimensional softplus activation function if and only if it holds for all $x = (x_1, \dots, x_d) \in \mathbb{R}^d$ that*

$$A(x) = (\ln(1 + \exp(x_1)), \ln(1 + \exp(x_2)), \dots, \ln(1 + \exp(x_d))) \quad (1.51)$$

(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 (1.49) establish that for all $x = (x_1, \dots, x_d) \in \mathbb{R}^d$ it holds that

$$\mathfrak{M}_{a,d}(x) = (\ln(1 + \exp(x_1)), \ln(1 + \exp(x_2)), \dots, \ln(1 + \exp(x_d))) \quad (1.52)$$

(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 (1.51). The proof of Lemma 1.2.14 is thus complete. \square

Exercise 1.2.16. 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 $\mathfrak{d}, H \in \mathbb{N}$, $l_1, l_2, \dots, l_H \in \mathbb{N}$, $\theta \in \mathbb{R}^\mathfrak{d}$ with $\mathfrak{d} \geq 2l_1 + [\sum_{k=2}^H l_k(l_{k-1} + 1)] + l_H + 1$ such that for all $x \in \mathbb{R}$ it holds that

$$(\mathcal{N}_{A_{l_1}, A_{l_2}, \dots, A_{l_H}, \text{id}_{\mathbb{R}}}^{\theta, 1})(x) = x \quad (1.53)$$

(cf. Definition 1.1.3).

The statement of the next lemma, Lemma 1.2.15, provides a partial answer to Exercise 1.2.16 (cf. also Lemma 2.2.8 below).

Lemma 1.2.15 (Real identity for the softplus activation function). *Let a be the softplus activation function and let $\theta = (1, -1, 0, 0, 1, -1, 0) \in \mathbb{R}^7$ (cf. Definition 1.2.11). Then it holds for all $x \in \mathbb{R}$ that*

$$(\mathcal{N}_{\mathfrak{M}_{a,2}, \text{id}_{\mathbb{R}}}^{\theta, 1})(x) = x \quad (1.54)$$

(cf. Definitions 1.1.3 and 1.2.1).

Proof of Lemma 1.2.15. Observe that (1.1), (1.5), and (1.49) demonstrate that here for all $x \in \mathbb{R}$ it holds that

$$\begin{aligned}
 (\mathcal{N}_{\mathfrak{M}_{a,2}, \text{id}_{\mathbb{R}}}^{\theta,1})(x) &= \ln(1 + \exp(x + 0)) - \ln(1 + \exp(-x + 0)) + 0 \\
 &= \ln(1 + \exp(x)) - \ln(1 + \exp(-x)) \\
 &= \ln\left(\frac{1+\exp(x)}{1+\exp(-x)}\right) \\
 &= \ln\left(\frac{\exp(x)(1+\exp(-x))}{1+\exp(-x)}\right) \\
 &= \ln(\exp(x)) = x
 \end{aligned} \tag{1.55}$$

(cf. Definitions 1.1.3 and 1.2.1). The proof of Lemma 1.2.15 is thus complete. \square

1.2.6 Gaussian error linear unit (GELU) activation

Another popular activation function is the **GELU** activation function first introduced in Hendrycks & Gimpel [203]. This activation function is the subject of the next definition.

Definition 1.2.16 (**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 $a: \mathbb{R} \rightarrow \mathbb{R}$ is the function from \mathbb{R} to \mathbb{R} which satisfies for all $x \in \mathbb{R}$ that*

$$a(x) = \frac{x}{\sqrt{2\pi}} \left[\int_{-\infty}^x \exp(-z^2/2) dz \right]. \tag{1.56}$$

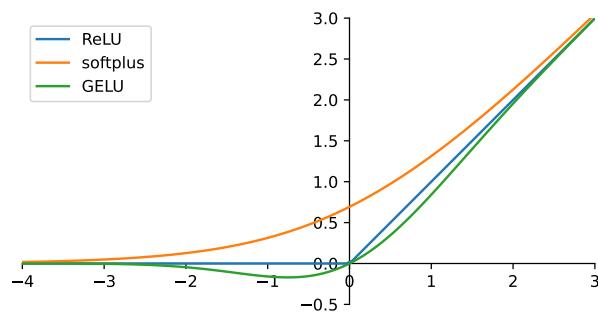


Figure 1.7 ([plots/gelu.pdf](#)): A plot of the **GELU** activation function, the **ReLU** activation function, and the **softplus** activation function

```

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((-4,3), (-.5,3))
7
8 x = np.linspace(-4, 3, 100)
9
10 ax.plot(x, tf.keras.activations.relu(x), label='ReLU')
11 ax.plot(x, tf.keras.activations.softplus(x), label='softplus')
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 1.2.17. Let $x \in \mathbb{R}$ and let a be the **GELU** activation function (cf. Definition 1.2.16). Then the following two statements are equivalent:

- (i) It holds that $a(x) > 0$.
- (ii) It holds that $\tau(x) > 0$ (cf. Definition 1.2.4).

Proof of Lemma 1.2.17. Note that (1.26) and (1.56) show that ((i) \leftrightarrow (ii)). The proof of Lemma 1.2.17 is thus complete. \square

Definition 1.2.18 (Multi-dimensional **GELU** activation functions). Let $d \in \mathbb{N}$ and let a be the **GELU** activation function (cf. Definition 1.2.16). 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 1.2.19 (Standard logistic activation function). We say that a is the standard logistic activation function if and only if it holds that $a: \mathbb{R} \rightarrow \mathbb{R}$ is the function from \mathbb{R} to \mathbb{R} which satisfies for all $x \in \mathbb{R}$ that

$$a(x) = \frac{1}{1 + \exp(-x)} = \frac{\exp(x)}{\exp(x) + 1}. \quad (1.57)$$

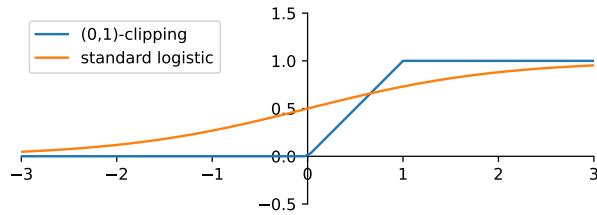


Figure 1.8 ([plots/logistic.pdf](#)): A plot of the standard logistic activation function and the $(0, 1)$ -clipping activation function

```

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((-3,3), (-.5,1.5))
7
8 x = np.linspace(-3, 3, 100)
9
10 ax.plot(x, tf.keras.activations.relu(x, max_value=1),
11           label='(0,1)-clipping')
12 ax.plot(x, tf.keras.activations.sigmoid(x),
13           label='standard logistic')
14 ax.legend()
15
16 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 1.2.20 (Multi-dimensional standard logistic activation functions). Let $d \in \mathbb{N}$ and let a be the standard logistic activation function (cf. Definition 1.2.19). 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 1.2.21 (Logistic ODE). Let a be the standard logistic activation function (cf. Definition 1.2.19). Then

- (i) it holds that $a: \mathbb{R} \rightarrow \mathbb{R}$ is infinitely often differentiable and

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

$$a(0) = 1/2, \quad a'(x) = a(x)(1 - a(x)) = a(x) - [a(x)]^2, \quad \text{and} \quad (1.58)$$

$$a''(x) = a(x)(1 - a(x))(1 - 2a(x)) = 2[a(x)]^3 - 3[a(x)]^2 + a(x). \quad (1.59)$$

Proof of Proposition 1.2.21. Note that (1.57) implies item (i). Next note that (1.57) ensures that for all $x \in \mathbb{R}$ it holds that

$$\begin{aligned} a'(x) &= \frac{\exp(-x)}{(1 + \exp(-x))^2} = a(x) \left(\frac{\exp(-x)}{1 + \exp(-x)} \right) \\ &= a(x) \left(\frac{1 + \exp(-x) - 1}{1 + \exp(-x)} \right) = a(x) \left(1 - \frac{1}{1 + \exp(-x)} \right) \\ &= a(x)(1 - a(x)). \end{aligned} \quad (1.60)$$

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

$$\begin{aligned} a''(x) &= [a(x)(1 - a(x))]' = a'(x)(1 - a(x)) + a(x)(1 - a(x))' \\ &= a'(x)(1 - a(x)) - a(x)a'(x) = a'(x)(1 - 2a(x)) \\ &= a(x)(1 - a(x))(1 - 2a(x)) \\ &= (a(x) - [a(x)]^2)(1 - 2a(x)) = a(x) - [a(x)]^2 - 2[a(x)]^2 + 2[a(x)]^3 \\ &= 2[a(x)]^3 - 3[a(x)]^2 + a(x). \end{aligned} \quad (1.61)$$

This establishes item (ii). The proof of Proposition 1.2.21 is thus complete. \square

1.2.7.2 Integral of the standard logistic activation function

Lemma 1.2.22 (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 1.2.19). Then it holds for all $x \in \mathbb{R}$ that*

$$\int_{-\infty}^x \mathfrak{l}(y) dy = \int_{-\infty}^x \left(\frac{1}{1 + e^{-y}} \right) dy = \ln(1 + \exp(x)) = \mathfrak{s}(x). \quad (1.62)$$

Proof of Lemma 1.2.22. Observe that (1.49) implies that for all $x \in \mathbb{R}$ it holds that

$$\mathfrak{s}'(x) = \left[\frac{1}{1 + \exp(x)} \right] \exp(x) = \mathfrak{l}(x). \quad (1.63)$$

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

$$\int_w^x \underbrace{\mathfrak{l}(y)}_{\geq 0} dy = \mathfrak{s}(x) - \mathfrak{s}(w). \quad (1.64)$$

Combining this with the fact that $\lim_{w \rightarrow -\infty} \mathfrak{s}(w) = 0$ establishes (1.62). The proof of Lemma 1.2.22 is thus complete. \square

1.2.8 Swish and sigmoid linear unit (**SiLU**) activation

In this section we introduce the swish activation function and the *sigmoid linear unit* (**SiLU**) activation function which is a special case of the swish activation function.

Definition 1.2.23 (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 $a: \mathbb{R} \rightarrow \mathbb{R}$ is the function from \mathbb{R} to \mathbb{R} which satisfies for all $x \in \mathbb{R}$ that*

$$a(x) = \frac{x}{1 + \exp(-\beta x)}. \quad (1.65)$$

Definition 1.2.24 (**SiLU** activation functions). *We say that a is the **SiLU** activation function if and only if it holds that $a: \mathbb{R} \rightarrow \mathbb{R}$ is the function from \mathbb{R} to \mathbb{R} which satisfies for all $x \in \mathbb{R}$ that*

$$a(x) = \frac{x}{1 + \exp(-x)}. \quad (1.66)$$

Lemma 1.2.25 (Relation between the **SiLU** activation function and the swish activation function). *Let \mathfrak{S} be the **SiLU** activation function and let \mathfrak{s} be the swish activation function with parameter 1 (cf. Definitions 1.2.23 and 1.2.24). Then $\mathfrak{S} = \mathfrak{s}$.*

Proof of Lemma 1.2.25. Observe that (1.66) and (1.65) establish that $\mathfrak{S} = \mathfrak{s}$. The proof of Lemma 1.2.25 is thus complete. The proof of Lemma 1.2.25 is thus complete. \square

```

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((-4,3), (-.5,3))
7
8 x = np.linspace(-4, 3, 100)
9
10 ax.plot(x, tf.keras.activations.relu(x), label='ReLU')
11 ax.plot(x, tf.keras.activations.gelu(x), label='GELU')
12 ax.plot(x, tf.keras.activations.swish(x), label='swish')
13 ax.legend()
14

```

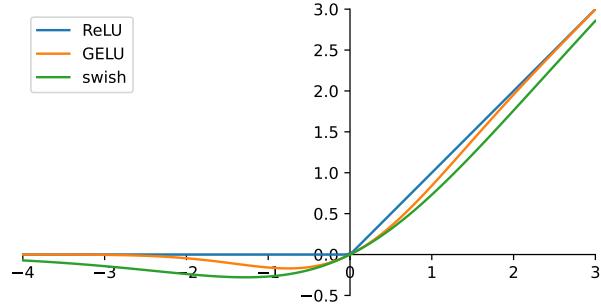


Figure 1.9 ([plots/swish.pdf](#)): A plot of the swish activation function with parameter 1, the GELU activation function, and the ReLU activation function

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 1.2.26 (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 1.2.19 and 1.2.23). Then it holds for all $x \in \mathbb{R}$ that*

$$\mathfrak{s}(x) = x\mathfrak{l}(\beta x). \quad (1.67)$$

Proof of Lemma 1.2.26. Observe that (1.65) and (1.57) establish (1.67). The proof of Lemma 1.2.26 is thus complete. \square

Definition 1.2.27 (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 1.2.23). 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).*

Definition 1.2.28 (Multi-dimensional SiLU activation functions). *Let $d \in \mathbb{N}$ and let a be the SiLU activation function (cf. Definition 1.2.24). Then we say that A is the d -dimensional SiLU activation function activation function if and only if $A = \mathfrak{M}_{a,d}$ (cf. Definition 1.2.1).*

1.2.9 Hyperbolic tangent activation

Definition 1.2.29 (Hyperbolic tangent activation function). We denote by $\tanh: \mathbb{R} \rightarrow \mathbb{R}$ the function which satisfies for all $x \in \mathbb{R}$ that

$$\tanh(x) = \frac{\exp(x) - \exp(-x)}{\exp(x) + \exp(-x)} \quad (1.68)$$

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

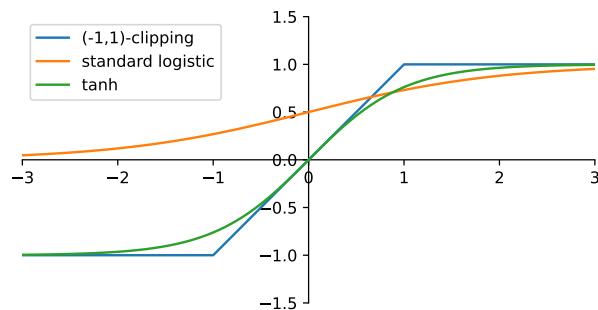


Figure 1.10 ([plots/tanh.pdf](#)): A plot of the hyperbolic tangent, the $(-1, 1)$ -clipping activation function, and the standard logistic activation function

```

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((-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,
11           label='(-1,1)-clipping')
12 ax.plot(x, tf.keras.activations.sigmoid(x),
13           label='standard logistic')
14 ax.plot(x, tf.keras.activations.tanh(x), label='tanh')
15 ax.legend()
16
17 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 1.2.30 (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 1.2.29).

Lemma 1.2.31. Let a be the standard logistic activation function (cf. Definition 1.2.19). Then it holds for all $x \in \mathbb{R}$ that

$$\tanh(x) = 2a(2x) - 1 \quad (1.69)$$

(cf. Definitions 1.2.19 and 1.2.29).

Proof of Lemma 1.2.31. Observe that (1.57) and (1.68) ensure that for all $x \in \mathbb{R}$ it holds that

$$\begin{aligned} 2a(2x) - 1 &= 2\left(\frac{\exp(2x)}{\exp(2x) + 1}\right) - 1 = \frac{2\exp(2x) - (\exp(2x) + 1)}{\exp(2x) + 1} \\ &= \frac{\exp(2x) - 1}{\exp(2x) + 1} = \frac{\exp(x)(\exp(x) - \exp(-x))}{\exp(x)(\exp(x) + \exp(-x))} \\ &= \frac{\exp(x) - \exp(-x)}{\exp(x) + \exp(-x)} = \tanh(x). \end{aligned} \quad (1.70)$$

The proof of Lemma 1.2.31 is thus complete. \square

Exercise 1.2.17. Let a be the standard logistic activation function (cf. Definition 1.2.19). Prove or disprove the following statement: There exists $L \in \{2, 3, \dots\}$, $\mathfrak{d}, l_1, l_2, \dots, l_{L-1} \in \mathbb{N}$, $\theta \in \mathbb{R}^{\mathfrak{d}}$ with $\mathfrak{d} \geq 2l_1 + [\sum_{k=2}^{L-1} l_k(l_{k-1} + 1)] + (l_{L-1} + 1)$ such that for all $x \in \mathbb{R}$ it holds that

$$(\mathcal{N}_{\mathfrak{M}_{a,l_1}, \mathfrak{M}_{a,l_2}, \dots, \mathfrak{M}_{a,l_{L-1}}, \text{id}_{\mathbb{R}}}^{\theta, 1})(x) = \tanh(x) \quad (1.71)$$

(cf. Definitions 1.1.3, 1.2.1, and 1.2.29).

1.2.10 Softsign activation

Definition 1.2.32 (Softsign activation function). We say that a is the softsign activation function if and only if it holds that $a: \mathbb{R} \rightarrow \mathbb{R}$ is the function from \mathbb{R} to \mathbb{R} which satisfies for all $x \in \mathbb{R}$ that

$$a(x) = \frac{x}{|x| + 1}. \quad (1.72)$$

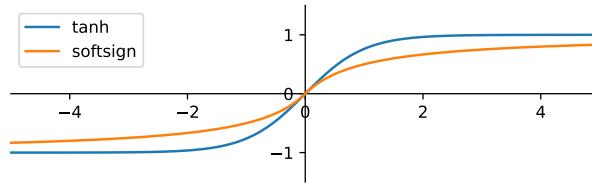


Figure 1.11 ([plots/softsign.pdf](#)): A plot of the softsign activation function and the hyperbolic tangent

```

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((-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 1.2.33 (Multi-dimensional softsign activation functions). *Let $d \in \mathbb{N}$ and let a be the softsign activation function (cf. Definition 1.2.32). 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 1.2.34 (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 $a: \mathbb{R} \rightarrow \mathbb{R}$ is the function from \mathbb{R} to \mathbb{R} which satisfies for all $x \in \mathbb{R}$ that*

$$a(x) = \begin{cases} x & : x > 0 \\ \gamma x & : x \leq 0. \end{cases} \quad (1.73)$$

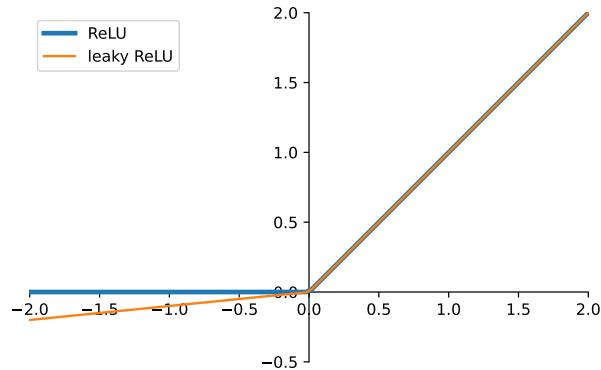


Figure 1.12 ([plots/leaky_relu.pdf](#)): A plot of the leaky ReLU activation function with leak factor $1/10$ and the ReLU activation function

```

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((-2,2), (-.5,2))
7
8 x = np.linspace(-2, 2, 100)
9
10 ax.plot(x, tf.keras.activations.relu(x), linewidth=3, label='ReLU')
11 ax.plot(x, tf.keras.activations.relu(x, alpha=0.1),
12         label='leaky ReLU')
13 ax.legend()
14
15 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 1.2.35. Let $\gamma \in [0, 1]$ and let $a: \mathbb{R} \rightarrow \mathbb{R}$ be a function. Then a is the leaky ReLU activation function with leak factor γ if and only if it holds for all $x \in \mathbb{R}$ that

$$a(x) = \max\{x, \gamma x\} \quad (1.74)$$

(cf. Definition 1.2.34).

Proof of Lemma 1.2.35. Note that the fact that $\gamma \leq 1$ and (1.73) imply (1.74). The proof of Lemma 1.2.35 is thus complete. \square

Lemma 1.2.36. 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, 1.2.16, 1.2.19, 1.2.23, 1.2.32, and 1.2.34). Then

- (i) it holds for all $f \in \{\mathfrak{r}, \mathfrak{c}_{u,v}, \tanh, a_1, a_2, \dots, a_5\}$ that $\limsup_{x \rightarrow -\infty} |f'(x)| = 0$ and
 - (ii) it holds that $\lim_{x \rightarrow -\infty} l'(x) = \gamma$
- (cf. Definitions 1.2.4, 1.2.9, and 1.2.29).

Proof of Lemma 1.2.36. Note that (1.26), (1.47), (1.49), (1.56), (1.57), (1.65), (1.68), and (1.72) prove item (i). Observe that (1.73) establishes item (ii). The proof of Lemma 1.2.36 is thus complete. \square

Definition 1.2.37 (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 1.2.34). 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. [86]. This activation function is the subject of the next notion.

Definition 1.2.38 (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 $a: \mathbb{R} \rightarrow \mathbb{R}$ is the function from \mathbb{R} to \mathbb{R} which satisfies for all $x \in \mathbb{R}$ that

$$a(x) = \begin{cases} x & : x > 0 \\ \gamma(1 - \exp(x)) & : x \leq 0. \end{cases} \quad (1.75)$$

```

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, 2), (-1, 2))
7

```

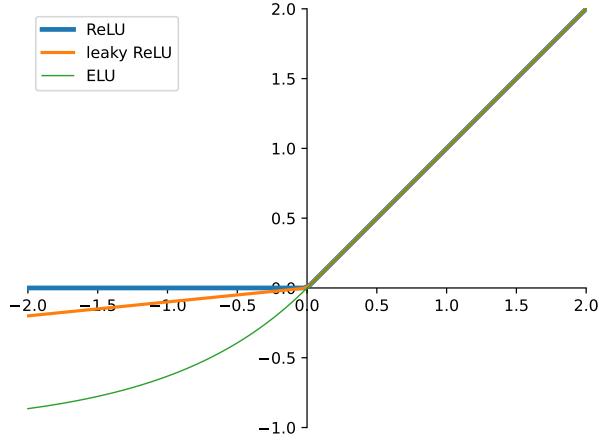


Figure 1.13 ([plots/elu.pdf](#)): A plot of the **ELU** activation function with asymptotic -1 , the leaky **ReLU** activation function with leak factor $1/10$, and the **ReLU** activation function

```

8 x = np.linspace(-2, 2, 100)
9
10 ax.plot(x, tf.keras.activations.relu(x), linewidth=3, label='ReLU')
11 ax.plot(x, tf.keras.activations.relu(x, alpha=0.1), linewidth=2,
12         label='leaky ReLU')
13 ax.plot(x, tf.keras.activations.elu(x), linewidth=0.9, label='ELU')
14 ax.legend()
15 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 1.2.39. Let $\gamma \in (-\infty, 0]$ and let a be the **ELU** activation function with asymptotic γ (cf. Definition 1.2.38). Then

$$\limsup_{x \rightarrow -\infty} a(x) = \liminf_{x \rightarrow -\infty} a(x) = \gamma. \quad (1.76)$$

Proof of Lemma 1.2.39. Observe that (1.75) shows (1.76). The proof of Lemma 1.2.39 is thus complete. \square

Definition 1.2.40 (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 1.2.38). 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 1.2.41 (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 $a: \mathbb{R} \rightarrow \mathbb{R}$ is the function from \mathbb{R} to \mathbb{R} which satisfies for all $x \in \mathbb{R}$ that

$$a(x) = (\max\{x, 0\})^p. \quad (1.77)$$

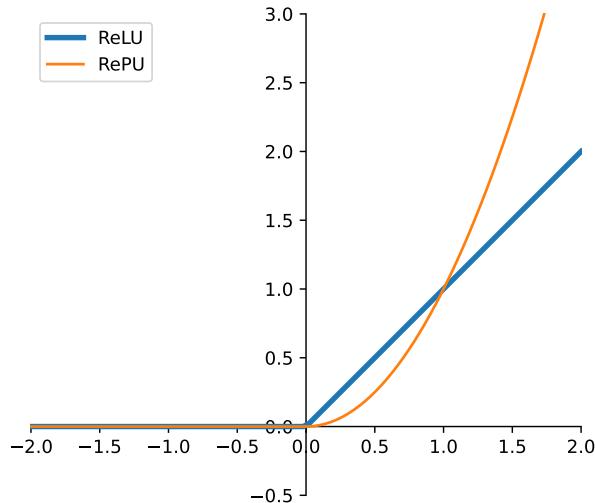


Figure 1.14 ([plots/repu.pdf](#)): A plot of the RePU activation function with power 2 and the ReLU activation function

```

1 import numpy as np
2 import tensorflow as tf
3 import matplotlib.pyplot as plt
4 import plot_util
5
6 ax = plot_util.setup_axis((-2,2), (-.5,3))
7 ax.set_ylim(-.5, 3)
8
9 x = np.linspace(-2, 2, 100)
10
11 ax.plot(x, tf.keras.activations.relu(x), linewidth=3, label='ReLU')

```

```

12 ax.plot(x, tf.keras.activations.relu(x)**2, label='RePU')
13 ax.legend()
14
15 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 1.2.42 (Multi-dimensional RePU activation functions). *Let $d, p \in \mathbb{N}$ and let a be the RePU activation function with power p (cf. Definition 1.2.41). 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. [401]. This is formulated in the next notion.

Definition 1.2.43 (Sine activation function). *We say that a is the sine activation function if and only if it holds that $a: \mathbb{R} \rightarrow \mathbb{R}$ is the function from \mathbb{R} to \mathbb{R} which satisfies for all $x \in \mathbb{R}$ that*

$$a(x) = \sin(x). \quad (1.78)$$

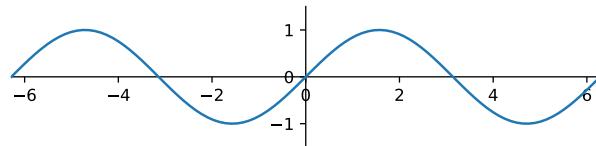


Figure 1.15 ([plots/sine.pdf](#)): A plot of the sine activation function

```

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 1.2.44 (Multi-dimensional sine activation functions). *Let $d \in \mathbb{N}$ and let a be the sine activation function (cf. Definition 1.2.43). 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 1.2.45 (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 $a: \mathbb{R} \rightarrow \mathbb{R}$ is the function from \mathbb{R} to \mathbb{R} which satisfies for all $x \in \mathbb{R}$ that*

$$a(x) = \mathbb{1}_{[0,\infty)}(x) = \begin{cases} 1 & : x \geq 0 \\ 0 & : x < 0. \end{cases} \quad (1.79)$$

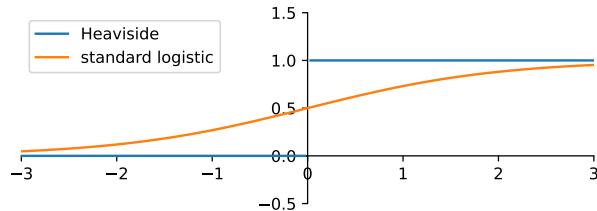


Figure 1.16 ([plots/heaviside.pdf](#)): A plot of the Heaviside activation function and the standard logistic activation function

```

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((-3,3), (-.5,1.5))
7
8 x = np.linspace(-3, 3, 100)
9
10 ax.plot(x[0:50], [0]*50, 'C0')
11 ax.plot(x[50:100], [1]*50, 'C0', label='Heaviside')

```

```

12 ax.plot(x, tf.keras.activations.sigmoid(x), 'C1',
13         label='standard logistic')
14 ax.legend()
15
16 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 1.2.46 (Multi-dimensional Heaviside activation functions). *Let $d \in \mathbb{N}$ and let a be the Heaviside activation function (cf. Definition 1.2.45). 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 1.2.47 (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 $A: \mathbb{R}^d \rightarrow \mathbb{R}^d$ is the function from \mathbb{R}^d to \mathbb{R}^d which satisfies for all $x = (x_1, \dots, x_d) \in \mathbb{R}^d$ that*

$$A(x) = \left(\frac{\exp(x_1)}{\left(\sum_{i=1}^d \exp(x_i)\right)}, \frac{\exp(x_2)}{\left(\sum_{i=1}^d \exp(x_i)\right)}, \dots, \frac{\exp(x_d)}{\left(\sum_{i=1}^d \exp(x_i)\right)} \right). \quad (1.80)$$

Lemma 1.2.48. *Let $d \in \mathbb{N}$ and let $A = (A_1, \dots, A_d)$ be the d -dimensional softmax activation function (cf. Definition 1.2.47). Then*

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

$$\sum_{k=1}^d A_k(x) = 1. \quad (1.81)$$

tum

(cf. Definition 1.2.47).

Proof of Lemma 1.2.48. Observe that (1.80) demonstrates that for all $x = (x_1, \dots, x_d) \in \mathbb{R}^d$ it holds that

$$\sum_{k=1}^d A_k(x) = \sum_{k=1}^d \frac{\exp(x_k)}{\left(\sum_{i=1}^d \exp(x_i)\right)} = \frac{\sum_{k=1}^d \exp(x_k)}{\sum_{i=1}^d \exp(x_i)} = 1. \quad (1.82)$$

The proof of Lemma 1.2.48 is thus complete. \square

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*

$$\mathbf{N} = \bigcup_{L \in \mathbb{N}} \bigcup_{l_0, l_1, \dots, l_L \in \mathbb{N}} \left(\bigtimes_{k=1}^L (\mathbb{R}^{l_k \times l_{k-1}} \times \mathbb{R}^{l_k}) \right), \quad (1.83)$$

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

$$\mathcal{P}(\Phi) = \sum_{k=1}^L l_k(l_{k-1} + 1), \quad \mathcal{L}(\Phi) = L, \quad \mathcal{I}(\Phi) = l_0, \quad \mathcal{O}(\Phi) = l_L, \quad \text{and} \quad \mathcal{H}(\Phi) = L - 1, \quad (1.84)$$

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

$$\mathbb{D}_n(\Phi) = \begin{cases} l_n & : n \leq L \\ 0 & : n > L, \end{cases} \quad (1.85)$$

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

$$\mathcal{D}(\Phi) = (\mathbb{D}_0(\Phi), \mathbb{D}_1(\Phi), \dots, \mathbb{D}_{\mathcal{L}(\Phi)}(\Phi)), \quad (1.86)$$

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

$$\mathcal{W}_{n,\Phi} = W_n \quad \text{and} \quad \mathcal{B}_{n,\Phi} = B_n. \quad (1.87)$$

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

$$\Phi \in \mathbf{N} \quad (1.88)$$

(cf. Definition 1.3.1).

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

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

(ii) *it holds that*

$$\mathcal{I}(\Phi) = \mathbb{D}_0(\Phi) \quad \text{and} \quad \mathcal{O}(\Phi) = \mathbb{D}_{\mathcal{L}(\Phi)}(\Phi), \quad (1.89)$$

and

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

$$\mathcal{W}_{n,\Phi} \in \mathbb{R}^{\mathbb{D}_n(\Phi) \times \mathbb{D}_{n-1}(\Phi)} \quad \text{and} \quad \mathcal{B}_{n,\Phi} \in \mathbb{R}^{\mathbb{D}_n(\Phi)}. \quad (1.90)$$

Proof of Lemma 1.3.3. Note that the assumption that

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

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

$$\Phi \in \left(\bigtimes_{k=1}^L (\mathbb{R}^{l_k \times l_{k-1}} \times \mathbb{R}^{l_k}) \right). \quad (1.91)$$

Observe that (1.91), (1.84), and (1.85) imply that

$$\mathcal{L}(\Phi) = L, \quad \mathcal{I}(\Phi) = l_0 = \mathbb{D}_0(\Phi), \quad \text{and} \quad \mathcal{O}(\Phi) = l_L = \mathbb{D}_L(\Phi). \quad (1.92)$$

This shows that

$$\mathcal{D}(\Phi) = (l_0, l_1, \dots, l_L) \in \mathbb{N}^{L+1} = \mathbb{N}^{\mathcal{L}(\Phi)+1}. \quad (1.93)$$

Next note that (1.91), (1.85), and (1.87) ensure that for all $n \in \{1, 2, \dots, \mathcal{L}(\Phi)\}$ it holds that

$$\mathcal{W}_{n,\Phi} \in \mathbb{R}^{l_n \times l_{n-1}} = \mathbb{R}^{\mathbb{D}_n(\Phi) \times \mathbb{D}_{n-1}(\Phi)} \quad \text{and} \quad \mathcal{B}_{n,\Phi} \in \mathbb{R}^{l_n} = \mathbb{R}^{\mathbb{D}_n(\Phi)}. \quad (1.94)$$

The proof of Lemma 1.3.3 is thus complete. \square

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 $a: \mathbb{R} \rightarrow \mathbb{R}$ be a function (cf. Definition 1.3.1). Then we denote by

$$\mathcal{R}_a^{\mathbf{N}}(\Phi): \mathbb{R}^{\mathcal{I}(\Phi)} \rightarrow \mathbb{R}^{\mathcal{O}(\Phi)} \quad (1.95)$$

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

$$\forall k \in \{1, 2, \dots, \mathcal{L}(\Phi)\}: x_k = \mathfrak{M}_{a \mathbb{1}_{(0, \mathcal{L}(\Phi))}(k) + \text{id}_{\mathbb{R}}} \mathbb{1}_{\{\mathcal{L}(\Phi)\}}(k), \mathbb{D}_k(\Phi)} (\mathcal{W}_{k, \Phi} x_{k-1} + \mathcal{B}_{k, \Phi}) \quad (1.96)$$

that

$$(\mathcal{R}_a^{\mathbf{N}}(\Phi))(x_0) = x_{\mathcal{L}(\Phi)} \quad (1.97)$$

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 $L \in \mathbb{N}$, $l_0, l_1, \dots, l_L \in \mathbb{N}$, let $a: \mathbb{R} \rightarrow \mathbb{R}$ be a function, and consider the function

$$\mathcal{f}: \left(\bigtimes_{k=1}^L (\mathbb{R}^{l_k \times l_{k-1}} \times \mathbb{R}^{l_k}) \right) \times \mathbb{R}^{l_0} \rightarrow \mathbb{R}^{l_L} \quad (1.98)$$

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

$$\mathcal{f}(\Phi, x) = \mathcal{R}_a^{\mathbf{N}}(\Phi)(x) \quad (1.99)$$

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

Exercise 1.3.1. Let

$$\Phi = ((W_1, B_1), (W_2, B_2), (W_3, B_3)) \in (\mathbb{R}^{2 \times 1} \times \mathbb{R}^2) \times (\mathbb{R}^{3 \times 2} \times \mathbb{R}^3) \times (\mathbb{R}^{1 \times 3} \times \mathbb{R}^1) \quad (1.100)$$

satisfy

$$W_1 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \quad B_1 = \begin{pmatrix} 3 \\ 4 \end{pmatrix}, \quad W_2 = \begin{pmatrix} -1 & 2 \\ 3 & -4 \\ -5 & 6 \end{pmatrix}, \quad B_2 = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \quad (1.101)$$

$$W_3 = \begin{pmatrix} -1 & 1 & -1 \end{pmatrix}, \quad \text{and} \quad B_3 = (-4). \quad (1.102)$$

Prove or disprove the following statement: It holds that

$$(\mathcal{R}_{\mathbf{r}}^{\mathbf{N}}(\Phi))(-1) = 0 \quad (1.103)$$

(cf. Definitions 1.2.4 and 1.3.4).

Exercise 1.3.2. Let a be the standard logistic activation function (cf. Definition 1.2.19). Prove or disprove the following statement: There exists $\Phi \in \mathbf{N}$ such that

$$\mathcal{R}_{\tanh}^{\mathbf{N}}(\Phi) = a \quad (1.104)$$

(cf. Definitions 1.2.29, 1.3.1, and 1.3.4).

```

1 import torch
2 import torch.nn as nn
3 import torch.nn.functional as F
4
5
6 # To define a neural network, we define a class that inherits from
7 # torch.nn.Module
8 class FullyConnectedANN(nn.Module):
9     def __init__(self):
10         super().__init__()
11         # In the constructor, we define the weights and biases.
12         # Wrapping the tensors in torch.nn.Parameter objects tells
13         # PyTorch that these are parameters that should be
14         # optimized during training.
15         self.W1 = nn.Parameter(
16             torch.Tensor([[1, 0], [0, -1], [-2, 2]]))
17         self.B1 = nn.Parameter(torch.Tensor([0, 2, -1]))
18         self.W2 = nn.Parameter(torch.Tensor([[1, -2, 3]]))
19         self.B2 = nn.Parameter(torch.Tensor([1]))
20
21
22     # The realization function of the network
23     def forward(self, x0):
24         x1 = F.relu(self.W1 @ x0 + self.B1)
25         x2 = self.W2 @ x1 + self.B2
26         return x2
27
28
29 model = FullyConnectedANN()
30
31 x0 = torch.Tensor([1, 2])
32 # Print the output of the realization function for input x0
33 print(model.forward(x0))
34

```

```

35 # As a consequence of inheriting from torch.nn.Module we can just
36 # "call" the model itself (which will call the forward method
37 # implicitly)
38 print(model(x0))
39
40 # Wrapping a tensor in a Parameter object and assigning it to an
41 # instance variable of the Module makes PyTorch register it as a
42 # parameter. We can access all parameters via the parameters
43 # method.
44 for p in model.parameters():
45     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 $\left(\left(\begin{pmatrix} 1 & 0 \\ 0 & -1 \\ -2 & 2 \end{pmatrix}, \begin{pmatrix} 0 \\ 2 \\ -1 \end{pmatrix}\right), ((1 \rightarrow 3), (1))\right) \in (\mathbb{R}^{3 \times 2} \times \mathbb{R}^3) \times (\mathbb{R}^{1 \times 3} \times \mathbb{R}^1) \subseteq \mathbf{N}$ using the ReLU activation function after the hidden layer.

```

1 import torch
2 import torch.nn as nn
3
4
5 class FullyConnectedANN(nn.Module):
6     def __init__(self):
7         super().__init__()
8         # Define the layers of the network in terms of Modules.
9         # nn.Linear(3, 20) represents an affine function defined
10        # by a 20x3 weight matrix and a 20-dimensional bias vector.
11        self.affine1 = nn.Linear(3, 20)
12        # The torch.nn.ReLU class simply wraps the
13        # torch.nn.functional.relu function as a Module.
14        self.activation1 = nn.ReLU()
15        self.affine2 = nn.Linear(20, 30)
16        self.activation2 = nn.ReLU()
17        self.affine3 = nn.Linear(30, 1)
18
19    def forward(self, x0):
20        x1 = self.activation1(self.affine1(x0))
21        x2 = self.activation2(self.affine2(x1))
22        x3 = self.affine3(x2)
23        return x3
24
25
26 model = FullyConnectedANN()
27
28 x0 = torch.Tensor([1, 2, 3])
29 print(model(x0))
30
31 # Assigning a Module to an instance variable of a Module registers

```

1.3. Fully-connected feedforward ANNs (structured description)

```
32 # all of the former's parameters as parameters of the latter
33 for p in model.parameters():
34     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.

```
1 import torch
2 import torch.nn as nn
3
4 # A Module whose forward method is simply a composition of Modules
5 # can be represented using the torch.nn.Sequential class
6 model = nn.Sequential(
7     nn.Linear(3, 20),
8     nn.ReLU(),
9     nn.Linear(20, 30),
10    nn.ReLU(),
11    nn.Linear(30, 1),
12)
13
14 # Prints a summary of the model architecture
15 print(model)
16
17 x0 = torch.Tensor([1, 2, 3])
18 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 $\mathcal{T}: \mathbf{N} \rightarrow (\bigcup_{d \in \mathbb{N}} \mathbb{R}^d)$ the function which satisfies for all $\Phi \in \mathbf{N}$, $k \in \{1, 2, \dots, \mathcal{L}(\Phi)\}$, $d \in \mathbb{N}$, $\theta = (\theta_1, \dots, \theta_d) \in \mathbb{R}^d$ with*

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

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

$$\begin{pmatrix} \theta_{(\sum_{i=1}^{k-1} l_i(l_{i-1}+1)) + 1} & \theta_{(\sum_{i=1}^{k-1} l_i(l_{i-1}+1)) + 2} & \cdots & \theta_{(\sum_{i=1}^{k-1} l_i(l_{i-1}+1)) + l_{k-1}} \\ \theta_{(\sum_{i=1}^{k-1} l_i(l_{i-1}+1)) + l_{k-1} + 1} & \theta_{(\sum_{i=1}^{k-1} l_i(l_{i-1}+1)) + l_{k-1} + 2} & \cdots & \theta_{(\sum_{i=1}^{k-1} l_i(l_{i-1}+1)) + 2l_{k-1}} \\ \theta_{(\sum_{i=1}^{k-1} l_i(l_{i-1}+1)) + 2l_{k-1} + 1} & \theta_{(\sum_{i=1}^{k-1} l_i(l_{i-1}+1)) + 2l_{k-1} + 2} & \cdots & \theta_{(\sum_{i=1}^{k-1} l_i(l_{i-1}+1)) + 3l_{k-1}} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{(\sum_{i=1}^{k-1} l_i(l_{i-1}+1)) + (l_{k-1})l_{k-1} + 1} & \theta_{(\sum_{i=1}^{k-1} l_i(l_{i-1}+1)) + (l_{k-1})l_{k-1} + 2} & \cdots & \theta_{(\sum_{i=1}^{k-1} l_i(l_{i-1}+1)) + l_k l_{k-1}} \end{pmatrix} \quad (1.105)$$

(cf. Definition 1.3.1).

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

$$\Phi = \left(\left(\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}, \begin{pmatrix} 10 \\ 11 \\ 12 \end{pmatrix} \right), \left(\begin{pmatrix} 13 & 14 & 15 \\ 16 & 17 & 18 \end{pmatrix}, \begin{pmatrix} 19 \\ 20 \end{pmatrix} \right) \right). \quad (1.106)$$

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

Proof for Example 1.3.7. Observe that (1.105) establishes (1.106). The proof for Example 1.3.7 is thus complete. \square

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

$$\begin{aligned} & \mathcal{T}((W, B)) \\ &= (W_{1,1}, W_{1,2}, \dots, W_{1,b}, W_{2,1}, W_{2,2}, \dots, W_{2,b}, \dots, W_{a,1}, W_{a,2}, \dots, W_{a,b}, B_1, B_2, \dots, B_a) \end{aligned} \quad (1.107)$$

(cf. Definition 1.3.6).

Proof of Lemma 1.3.8. Observe that (1.105) establishes (1.107). The proof of Lemma 1.3.8 is thus complete. \square

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

$$\begin{aligned} & \mathcal{T}\left(\left((W_1, B_1), (W_2, B_2), \dots, (W_L, B_L)\right)\right) \\ &= \left(W_{1,1,1}, W_{1,1,2}, \dots, W_{1,1,l_0}, \dots, W_{1,l_1,1}, W_{1,l_1,2}, \dots, W_{1,l_1,l_0}, B_{1,1}, B_{1,2}, \dots, B_{1,l_1}, \right. \\ & \quad W_{2,1,1}, W_{2,1,2}, \dots, W_{2,1,l_1}, \dots, W_{2,l_2,1}, W_{2,l_2,2}, \dots, W_{2,l_2,l_1}, B_{2,1}, B_{2,2}, \dots, B_{2,l_2}, \\ & \quad \dots, \\ & \quad \left.W_{L,1,1}, W_{L,1,2}, \dots, W_{L,1,l_{L-1}}, \dots, W_{L,l_L,1}, W_{L,l_L,2}, \dots, W_{L,l_L,l_{L-1}}, B_{L,1}, B_{L,2}, \dots, B_{L,l_L}\right) \end{aligned} \tag{1.108}$$

(cf. Definition 1.3.6).

Proof of Lemma 1.3.9. Note that (1.105) implies (1.108). The proof of Lemma 1.3.9 is thus complete. \square

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 $a: \mathbb{R} \rightarrow \mathbb{R}$ be a function and let $\Phi \in \mathbf{N}$. (cf. Definition 1.3.1).

Then

$$\mathcal{R}_a^{\mathbf{N}}(\Phi) = \begin{cases} \mathcal{N}_{\text{id}_{\mathbb{R}^{\mathcal{O}(\Phi)}}}^{\mathcal{T}(\Phi), \mathcal{I}(\Phi)} & : \mathcal{H}(\Phi) = 0 \\ \mathcal{N}_{\mathfrak{M}_{a, \mathbb{D}_1(\Phi)}, \mathfrak{M}_{a, \mathbb{D}_2(\Phi)}, \dots, \mathfrak{M}_{a, \mathbb{D}_{\mathcal{H}(\Phi)}(\Phi)}, \text{id}_{\mathbb{R}^{\mathcal{O}(\Phi)}}}^{\mathcal{T}(\Phi), \mathcal{I}(\Phi)} & : \mathcal{H}(\Phi) > 0 \end{cases} \tag{1.109}$$

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

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

$$\mathcal{L}(\Phi) = L \quad \text{and} \quad \mathcal{D}(\Phi) = (l_0, l_1, \dots, l_L). \tag{1.110}$$

Note that (1.105) shows that for all $k \in \{1, 2, \dots, L\}$, $x \in \mathbb{R}^{l_{k-1}}$ it holds that

$$\mathcal{W}_{k,\Phi}x + \mathcal{B}_{k,\Phi} = (\mathcal{A}_{l_k, l_{k-1}}^{\mathcal{T}(\Phi), \sum_{i=1}^{k-1} l_i(l_{i-1}+1)})(x) \tag{1.111}$$

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

$$x_{L-1} = \begin{cases} x_0 & : L = 1 \\ (\mathfrak{M}_{a,l_{L-1}} \circ \mathcal{A}_{l_{L-1},l_{L-2}}^{\mathcal{T}(\Phi), \sum_{i=1}^{L-2} l_i(l_{i-1}+1)} \\ \quad \circ \mathfrak{M}_{a,l_{L-2}} \circ \mathcal{A}_{l_{L-2},l_{L-3}}^{\mathcal{T}(\Phi), \sum_{i=1}^{L-3} l_i(l_{i-1}+1)} \circ \dots \circ \mathfrak{M}_{a,l_1} \circ \mathcal{A}_{l_1,l_0}^{\mathcal{T}(\Phi),0})(x_0) & : L > 1 \end{cases} \quad (1.112)$$

(cf. Definition 1.2.1). This, (1.111), (1.5), and (1.97) prove that for all $x_0 \in \mathbb{R}^{l_0}$, $x_1 \in \mathbb{R}^{l_1}, \dots, x_L \in \mathbb{R}^{l_L}$ with $\forall k \in \{1, 2, \dots, L\}: x_k = \mathfrak{M}_{a\mathbb{1}_{(0,L)}(k)+\text{id}_{\mathbb{R}}\mathbb{1}_{\{L\}}(k),l_k}(\mathcal{W}_{k,\Phi}x_{k-1} + \mathcal{B}_{k,\Phi})$ it holds that

$$\begin{aligned} (\mathcal{R}_a^N(\Phi))(x_0) &= x_L = \mathcal{W}_{L,\Phi}x_{L-1} + \mathcal{B}_{L,\Phi} = (\mathcal{A}_{l_L,l_{L-1}}^{\mathcal{T}(\Phi), \sum_{i=1}^{L-1} l_i(l_{i-1}+1)})(x_{L-1}) \\ &= \begin{cases} (\mathcal{N}_{\text{id}_{\mathbb{R}^{l_L}}}^{\mathcal{T}(\Phi),l_0})(x_0) & : L = 1 \\ (\mathcal{N}_{\mathfrak{M}_{a,l_1},\mathfrak{M}_{a,l_2},\dots,\mathfrak{M}_{a,l_{L-1}},\text{id}_{\mathbb{R}^{l_L}}}}^{\mathcal{T}(\Phi),l_0})(x_0) & : L > 1 \end{cases} \quad (1.113) \end{aligned}$$

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

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], [62, Chapter 16], [65, Section 4.2], [171, Chapter 9], and [284] for other introductions on CNNs.

CNNs were introduced in LeCun et al. [276] for *computer vision* (CV) applications. The first successful modern CNN architecture is widely considered to be the *AlexNet* architecture proposed in Krizhevsky et al. [271]. A few other very successful early CNN architectures for CV include [159, 200, 217, 299, 310, 392, 399, 411]. 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, 150, 259, 451, 455, 458] for applications of CNNs to *natural language processing* (NLP), we refer, for instance, to [1, 61, 81, 380, 417]

for applications of CNNs to audio processing, and we refer, for example, to [48, 111, 250, 369, 429, 461] for applications of CNNs to time series analysis. Finally, for approximation results for feedforward CNNs we refer, for instance, to Petersen & Voigtlander [354] and the references therein.

1.4.1 Discrete convolutions

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

$$\mathfrak{d}_t = a_t - w_t + 1. \quad (1.114)$$

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

$$(A * W)_{i_1, i_2, \dots, i_T} = \sum_{r_1=1}^{w_1} \sum_{r_2=1}^{w_2} \dots \sum_{r_T=1}^{w_T} A_{i_1-1+r_1, i_2-1+r_2, \dots, i_T-1+r_T} W_{r_1, r_2, \dots, r_T}. \quad (1.115)$$

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

$$\begin{aligned} \mathbf{C} = & \bigcup_{T, L \in \mathbb{N}} \bigcup_{l_0, l_1, \dots, l_L \in \mathbb{N}} \bigcup_{(c_{k,t})_{(k,t) \in \{1, 2, \dots, L\} \times \{1, 2, \dots, T\}} \subseteq \mathbb{N}} \left(\bigtimes_{k=1}^L \left((\mathbb{R}^{c_{k,1} \times c_{k,2} \times \dots \times c_{k,T}})^{l_k \times l_{k-1}} \times \mathbb{R}^{l_k} \right) \right). \end{aligned} \quad (1.116)$$

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

$$\Phi \in \mathbf{C} \quad (1.117)$$

(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 $\mathbf{I}^{d_1, d_2, \dots, d_T} = (\mathbf{I}_{i_1, i_2, \dots, i_T}^{d_1, d_2, \dots, d_T})_{(i_1, i_2, \dots, i_T) \in (\times_{t=1}^T \{1, 2, \dots, d_t\})} \in \mathbb{R}^{d_1 \times d_2 \times \dots \times d_T}$ 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

$$\mathbf{I}_{i_1, i_2, \dots, i_T}^{d_1, d_2, \dots, d_T} = 1. \quad (1.118)$$

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 $a: \mathbb{R} \rightarrow \mathbb{R}$ be a function. Then we denote by

$$\mathcal{R}_a^{\mathbf{C}}(\Phi): \left(\bigcup_{\substack{d_1, d_2, \dots, d_T \in \mathbb{N} \\ \forall t \in \{1, 2, \dots, T\}: d_t - \sum_{k=1}^L (c_{k,t} - 1) \geq 1}} (\mathbb{R}^{d_1 \times d_2 \times \dots \times d_T})^{l_0} \right) \rightarrow \left(\bigcup_{d_1, d_2, \dots, d_T \in \mathbb{N}} (\mathbb{R}^{d_1 \times d_2 \times \dots \times d_T})^{l_L} \right) \quad (1.119)$$

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

$$\forall k \in \{1, 2, \dots, L\}, t \in \{1, 2, \dots, T\}: \mathfrak{d}_{k,t} = \mathfrak{d}_{k-1,t} - c_{k,t} + 1 \quad (1.120)$$

and

$$\begin{aligned} \forall k \in \{1, 2, \dots, L\}, n \in \{1, 2, \dots, l_k\}: \\ x_{k,n} = \mathfrak{M}_{a \mathbb{1}_{(0,L)}(k) + \text{id}_{\mathbb{R}} \mathbb{1}_{\{L\}}(k), \mathfrak{d}_{k,1}, \mathfrak{d}_{k,2}, \dots, \mathfrak{d}_{k,T}}(B_{k,n} \mathbf{I}^{\mathfrak{d}_{k,1}, \mathfrak{d}_{k,2}, \dots, \mathfrak{d}_{k,T}} + \sum_{m=1}^{l_{k-1}} x_{k-1,m} * W_{k,n,m}) \end{aligned} \quad (1.121)$$

that

$$(\mathcal{R}_a^{\mathbf{C}}(\Phi))(x_0) = x_L \quad (1.122)$$

and we call $\mathcal{R}_a^{\mathbf{C}}(\Phi)$ the realization function of the feedforward CNN Φ with activation function a (we call $\mathcal{R}_a^{\mathbf{C}}(\Phi)$ 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).

```

1 import torch
2 import torch.nn as nn
3
4

```

```

5  class ConvolutionalANN(nn.Module):
6      def __init__(self):
7          super().__init__()
8          # The convolutional layer defined here takes any tensor of
9          # shape (1, n, m) [a single input] or (N, 1, n, m) [a batch
10         # of N inputs] where N, n, m are natural numbers satisfying
11         # n >= 3 and m >= 3.
12         self.conv1 = nn.Conv2d(
13             in_channels=1, out_channels=5, kernel_size=(3, 3)
14         )
15         self.activation1 = nn.ReLU()
16         self.conv2 = nn.Conv2d(
17             in_channels=5, out_channels=5, kernel_size=(5, 3)
18         )
19
20     def forward(self, x0):
21         x1 = self.activation1(self.conv1(x0))
22         print(x1.shape)
23         x2 = self.conv2(x1)
24         print(x2.shape)
25         return x2
26
27
28 model = ConvolutionalANN()
29 x0 = torch.rand(1, 20, 20)
30 # This will print the shapes of the outputs of the two layers of
31 # the model, in this case:
32 # torch.Size([5, 18, 18])
33 # torch.Size([5, 14, 16])
34 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 $\Phi \in (\bigtimes_{k=1}^L ((\mathbb{R}^{c_{k,1} \times c_{k,2} \times \dots \times c_{k,T}})^{l_k \times l_{k-1}} \times \mathbb{R}^{l_k})) = ((\mathbb{R}^{3 \times 3})^{5 \times 1} \times \mathbb{R}^5) \times ((\mathbb{R}^{3 \times 5})^{5 \times 5} \times \mathbb{R}^5)$. 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

$$\Phi \in \left(\bigtimes_{k=1}^L ((\mathbb{R}^{c_{k,1} \times c_{k,2} \times \dots \times c_{k,T}})^{l_k \times l_{k-1}} \times \mathbb{R}^{l_k}) \right) = ((\mathbb{R}^{2 \times 2})^{2 \times 1} \times \mathbb{R}^2) \times ((\mathbb{R}^{1 \times 1})^{1 \times 2} \times \mathbb{R}^1) \quad (1.123)$$

satisfy

$$\Phi = \left(\left(\begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \end{pmatrix} \right), ((((-2) \quad (2)), (3))) \right). \quad (1.124)$$

Then

$$(\mathcal{R}_r^C(\Phi)) \left(\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix} \right) = \begin{pmatrix} 11 & 15 \\ 23 & 27 \end{pmatrix} \quad (1.125)$$

(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

$$x_0 = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}, \quad x_{1,1} = \mathfrak{M}_{r,2,2} \left(\mathbf{I}^{2,2} + x_0 * \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \right), \quad (1.126)$$

$$x_{1,2} = \mathfrak{M}_{r,2,2} \left((-1)\mathbf{I}^{2,2} + x_0 * \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right), \quad (1.127)$$

$$\text{and} \quad x_2 = \mathfrak{M}_{id_{\mathbb{R}},2,2} (3\mathbf{I}^{2,2} + x_{1,1} * (-2) + x_{1,2} * (2)). \quad (1.128)$$

Note that (1.122), (1.124), (1.126), (1.127), and (1.128) imply that

$$(\mathcal{R}_r^C(\Phi)) \left(\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix} \right) = (\mathcal{R}_r^C(\Phi))(x_0) = x_2. \quad (1.129)$$

Next observe that (1.126) ensures that

$$\begin{aligned} x_{1,1} &= \mathfrak{M}_{r,2 \times 2} \left(\mathbf{I}^{2,2} + x_0 * \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \right) = \mathfrak{M}_{r,2 \times 2} \left(\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \right) \\ &= \mathfrak{M}_{r,2 \times 2} \left(\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \right) = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}. \end{aligned} \quad (1.130)$$

Furthermore, note that (1.127) establishes that

$$\begin{aligned} x_{1,2} &= \mathfrak{M}_{\mathfrak{r},2 \times 2} \left((-1)\mathbf{I}^{2,2} + x_0 * \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right) = \mathfrak{M}_{\mathfrak{r},2 \times 2} \left(\begin{pmatrix} -1 & -1 \\ -1 & -1 \end{pmatrix} + \begin{pmatrix} 6 & 8 \\ 12 & 14 \end{pmatrix} \right) \\ &= \mathfrak{M}_{\mathfrak{r},2 \times 2} \left(\begin{pmatrix} 5 & 7 \\ 11 & 13 \end{pmatrix} \right) = \begin{pmatrix} 5 & 7 \\ 11 & 13 \end{pmatrix}. \end{aligned} \quad (1.131)$$

Moreover, observe that this, (1.130), and (1.128) demonstrate that

$$\begin{aligned} x_2 &= \mathfrak{M}_{\text{id}_{\mathbb{R}},2 \times 2} \left(3\mathbf{I}^{2,2} + x_{1,1} * (-2) + x_{1,2} * (2) \right) \\ &= \mathfrak{M}_{\text{id}_{\mathbb{R}},2 \times 2} \left(3\mathbf{I}^{2,2} + \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} * (-2) + \begin{pmatrix} 5 & 7 \\ 11 & 13 \end{pmatrix} * (2) \right) \\ &= \mathfrak{M}_{\text{id}_{\mathbb{R}},2 \times 2} \left(\begin{pmatrix} 3 & 3 \\ 3 & 3 \end{pmatrix} + \begin{pmatrix} -2 & -2 \\ -2 & -2 \end{pmatrix} + \begin{pmatrix} 10 & 14 \\ 22 & 26 \end{pmatrix} \right) \\ &= \mathfrak{M}_{\text{id}_{\mathbb{R}},2 \times 2} \left(\begin{pmatrix} 11 & 15 \\ 23 & 27 \end{pmatrix} \right) = \begin{pmatrix} 11 & 15 \\ 23 & 27 \end{pmatrix}. \end{aligned} \quad (1.132)$$

This and (1.129) establish (1.125). The proof for Example 1.4.6 is thus complete. \square

```

1 import torch
2 import torch.nn as nn
3
4
5 model = nn.Sequential(
6     nn.Conv2d(in_channels=1, out_channels=2, kernel_size=(2, 2)),
7     nn.ReLU(),
8     nn.Conv2d(in_channels=2, out_channels=1, kernel_size=(1, 1)),
9 )
10
11 with torch.no_grad():
12     model[0].weight.set_(
13         torch.Tensor([[[[0, 0], [0, 0]], [[1, 0], [0, 1]]]])
14     )
15     model[0].bias.set_(torch.Tensor([1, -1]))
16     model[2].weight.set_(torch.Tensor([[[-2]], [[2]]]))
17     model[2].bias.set_(torch.Tensor([3]))
18
19 x0 = torch.Tensor([[1, 2, 3], [4, 5, 6], [7, 8, 9]])
20 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 (1.124)) in PYTORCH and verifying (1.125).

Exercise 1.4.1. Let

$$\Phi = (((W_{1,n,m})_{(n,m) \in \{1,2,3\} \times \{1\}}, (B_{1,n})_{n \in \{1,2,3\}}), ((W_{2,n,m})_{(n,m) \in \{1\} \times \{1,2,3\}}, (B_{2,n})_{n \in \{1\}})) \in ((\mathbb{R}^2)^{3 \times 1} \times \mathbb{R}^3) \times ((\mathbb{R}^3)^{1 \times 3} \times \mathbb{R}^1) \quad (1.133)$$

satisfy

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

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

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

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

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

Exercise 1.4.2. Let

$$\Phi = (((W_{1,n,m})_{(n,m) \in \{1,2,3\} \times \{1\}}, (B_{1,n})_{n \in \{1,2,3\}}), ((W_{2,n,m})_{(n,m) \in \{1\} \times \{1,2,3\}}, (B_{2,n})_{n \in \{1\}})) \in ((\mathbb{R}^3)^{3 \times 1} \times \mathbb{R}^3) \times ((\mathbb{R}^2)^{1 \times 3} \times \mathbb{R}^1) \quad (1.137)$$

satisfy

$$W_{1,1,1} = (1, 1, 1), \quad W_{1,2,1} = (2, -2, -2), \quad (1.138)$$

$$W_{1,3,1} = (-3, -3, 3), \quad (B_{1,n})_{n \in \{1,2,3\}} = (3, -2, -1), \quad (1.139)$$

$$W_{2,1,1} = (2, -1), \quad W_{2,1,2} = (-1, 2), \quad W_{2,1,3} = (-1, 0), \quad \text{and} \quad B_{2,1} = -2 \quad (1.140)$$

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

$$(\mathcal{R}_{\mathbf{r}}^{\mathbf{C}}(\Phi))(v) \quad (1.141)$$

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 $\mathbb{R}^{\mathcal{I}(\Phi)} \subseteq \text{Domain}(\mathcal{R}_a^{\mathbf{C}}(\Psi))$ and

$$(\mathcal{R}_a^{\mathbf{C}}(\Psi))(x) = (\mathcal{R}_a^{\mathbf{N}}(\Phi))(x) \quad (1.142)$$

(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 $\langle \cdot, \cdot \rangle : [\bigcup_{d \in \mathbb{N}} (\mathbb{R}^d \times \mathbb{R}^d)] \rightarrow \mathbb{R}$ the function which satisfies for all $d \in \mathbb{N}$, $x = (x_1, \dots, x_d), y = (y_1, \dots, y_d) \in \mathbb{R}^d$ that

$$\langle x, y \rangle = \sum_{i=1}^d x_i y_i. \quad (1.143)$$

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

$$(\mathcal{R}_a^C(\Phi))(x) = \left(\left(\langle \mathbf{e}_k^{(\mathcal{O}(\Phi))}, (\mathcal{R}_a^N(\Phi))((x_{i,j})_{i \in \{1, 2, \dots, \mathcal{I}(\Phi)\}}) \rangle \right)_{j \in \{1, 2, \dots, D\}} \right)_{k \in \{1, 2, \dots, \mathcal{O}(\Phi)\}} \quad (1.144)$$

(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. [200] as an attempt to improve the performance of deep **ANNs** which typically are much harder to train than shallow **ANNs** (cf., for example, [30, 160, 348]). The **ResNets** in He et al. [200] 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. [405, 406] (cf. also [278, 312, 366, 411, 419]). In addition, we refer to [201, 217, 425, 438, 448] for a few successful **ANN** architectures building on the **ResNets** in He et al. [200].

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

$$\mathbf{R} = \bigcup_{L \in \mathbb{N}} \bigcup_{l_0, l_1, \dots, l_L \in \mathbb{N}} \bigcup_{S \subseteq \{(r, k) \in (\mathbb{N}_0)^2 : r < k \leq L\}} \left(\left(\bigtimes_{k=1}^L (\mathbb{R}^{l_k \times l_{k-1}} \times \mathbb{R}^{l_k}) \right) \times \left(\bigtimes_{(r, k) \in S} \mathbb{R}^{l_k \times l_r} \right) \right). \quad (1.145)$$

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

$$\Phi \in \mathbf{R} \quad (1.146)$$

(cf. Definition 1.5.1).

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

$$\#(\bigtimes_{(r, k) \in S} \mathbb{R}^{l_k \times l_r}) = \begin{cases} 1 & : S = \emptyset \\ \infty & : S \neq \emptyset. \end{cases} \quad (1.147)$$

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

$$\#(\bigtimes_{(r, k) \in S} \mathbb{R}^{l_k \times l_r}) = \#\{f \in F(S, \bigcup_{(r, k) \in S} \mathbb{R}^{l_k \times l_r}) : (\forall (r, k) \in S : f(r, k) \in \mathbb{R}^{l_k \times l_r})\}. \quad (1.148)$$

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

$$\#(\bigtimes_{(r, k) \in \emptyset} \mathbb{R}^{l_k \times l_r}) = \#(F(\emptyset, \emptyset)) = 1. \quad (1.149)$$

Next note that (1.148) establishes that for all $(R, K) \in S$ it holds that

$$\#(\bigtimes_{(r, k) \in S} \mathbb{R}^{l_k \times l_r}) \geq \#(F(\{(R, K)\}, \mathbb{R}^{l_K \times l_R})) = \infty. \quad (1.150)$$

Combining this and (1.149) establishes (1.147). The proof of Lemma 1.5.3 is thus complete. \square

1.5.2 Realizations of fully-connected ResNets

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

$$\mathcal{R}_a^{\mathbf{R}}(\Phi) : \mathbb{R}^{l_0} \rightarrow \mathbb{R}^{l_L} \quad (1.151)$$

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

$\forall k \in \{1, 2, \dots, L\}$:

$$x_k = \mathfrak{M}_{a \mathbb{1}_{(0,L)}(k) + \text{id}_{\mathbb{R}} \mathbb{1}_{\{L\}}(k), l_k}(W_k x_{k-1} + B_k + \sum_{r \in \mathbb{N}_0, (r,k) \in S} V_{r,k} x_r) \quad (1.152)$$

that

$$(\mathcal{R}_a^{\mathbf{R}}(\Phi))(x_0) = x_L \quad (1.153)$$

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 $I_d \in \mathbb{R}^{d \times d}$ the identity matrix in $\mathbb{R}^{d \times d}$.

```

1 import torch
2 import torch.nn as nn
3
4 class ResidualANN(nn.Module):
5     def __init__(self):
6         super().__init__()
7         self.affine1 = nn.Linear(3, 10)
8         self.activation1 = nn.ReLU()
9         self.affine2 = nn.Linear(10, 20)
10        self.activation2 = nn.ReLU()
11        self.affine3 = nn.Linear(20, 10)
12        self.activation3 = nn.ReLU()
13        self.affine4 = nn.Linear(10, 1)
14
15    def forward(self, x0):
16        x1 = self.activation1(self.affine1(x0))
17        x2 = self.activation2(self.affine2(x1))
18        x3 = self.activation3(x1 + self.affine3(x2))
19        x4 = self.affine4(x3)
20        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 $l_0 = 3$, $l_1 = 10$, $l_2 = 20$, $l_3 = 10$, $l_4 = 1$, $\Phi = ((W_1, B_1), (W_2, B_2), (W_3, B_3), (W_4, B_4)) \in (\bigtimes_{k=1}^4 (\mathbb{R}^{l_k \times l_{k-1}} \times \mathbb{R}^{l_k}))$, $S = \{(1, 3)\}$, $V = (V_{r,k})_{(r,k) \in S} \in \bigtimes_{(r,k) \in S} \mathbb{R}^{l_k \times l_r}$, and $V_{1,3} = I_{10}$ (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

$$\Phi = ((W_1, B_1), (W_2, B_2), (W_3, B_3), (W_4, B_4)) \in (\bigtimes_{k=1}^4 (\mathbb{R}^{l_k \times l_{k-1}} \times \mathbb{R}^{l_k})) \quad (1.154)$$

satisfy

$$W_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad B_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad W_2 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \quad B_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad (1.155)$$

$$W_3 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad B_3 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad W_4 = \begin{pmatrix} 2 & 2 \end{pmatrix}, \quad \text{and} \quad B_4 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad (1.156)$$

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

$$V_{0,4} = \begin{pmatrix} -1 \end{pmatrix}. \quad (1.157)$$

Then

$$(\mathcal{R}_{\mathfrak{r}}^{\mathbf{R}}(\Phi, V))(5) = 28 \quad (1.158)$$

(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

$$x_k = \mathfrak{M}_{\mathfrak{r}1_{(0,4)}(k) + \text{id}_{\mathbb{R}1_{\{4\}}(k), l_k}}(W_k x_{k-1} + B_k + \sum_{r \in \mathbb{N}_0, (r,k) \in S} V_{r,k} x_r). \quad (1.159)$$

Observe that (1.159) shows that

$$(\mathcal{R}_{\mathfrak{r}}^{\mathbf{R}}(\Phi, V))(5) = x_4. \quad (1.160)$$

Next note that (1.159) ensures that

$$x_1 = \mathfrak{M}_{\mathfrak{r},1}(W_1 x_0 + B_1) = \mathfrak{M}_{\mathfrak{r},1}(5), \quad (1.161)$$

$$x_2 = \mathfrak{M}_{\mathfrak{r},2}(W_2 x_1 + B_2) = \mathfrak{M}_{\mathfrak{r},1}\left(\begin{pmatrix} 1 \\ 2 \end{pmatrix}(5) + \begin{pmatrix} 0 \\ 1 \end{pmatrix}\right) = \mathfrak{M}_{\mathfrak{r},1}\left(\begin{pmatrix} 5 \\ 11 \end{pmatrix}\right) = \begin{pmatrix} 5 \\ 11 \end{pmatrix}, \quad (1.162)$$

$$x_3 = \mathfrak{M}_{\mathfrak{r},2}(W_3 x_2 + B_3) = \mathfrak{M}_{\mathfrak{r},1}\left(\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 5 \\ 11 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \end{pmatrix}\right) = \mathfrak{M}_{\mathfrak{r},1}\left(\begin{pmatrix} 5 \\ 11 \end{pmatrix}\right) = \begin{pmatrix} 5 \\ 11 \end{pmatrix}, \quad (1.163)$$

$$\begin{aligned} \text{and } x_4 &= \mathfrak{M}_{\mathfrak{r},1}(W_4 x_3 + B_4 + V_{0,4} x_0) \\ &= \mathfrak{M}_{\mathfrak{r},1}\left((2 \ 2)\begin{pmatrix} 5 \\ 11 \end{pmatrix} + (1) + (-1)(5)\right) = \mathfrak{M}_{\mathfrak{r},1}(28) = 28. \end{aligned} \quad (1.164)$$

This and (1.160) establish (1.158). The proof for Example 1.5.6 is thus complete. \square

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

$$\Phi = ((W_1, B_1), (W_2, B_2), (W_3, B_3)) \in (\bigtimes_{k=1}^3 (\mathbb{R}^{l_k \times l_{k-1}} \times \mathbb{R}^{l_k})) \quad (1.165)$$

satisfy

$$W_1 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \quad B_1 = \begin{pmatrix} 3 \\ 4 \end{pmatrix}, \quad W_2 = \begin{pmatrix} -1 & 2 \\ 3 & -4 \\ -5 & 6 \end{pmatrix}, \quad B_2 = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \quad (1.166)$$

$$W_3 = \begin{pmatrix} -1 & 1 & -1 \end{pmatrix}, \quad \text{and} \quad B_3 = \begin{pmatrix} -4 \end{pmatrix}, \quad (1.167)$$

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

$$V_{0,3} = \begin{pmatrix} 1 \end{pmatrix} \quad \text{and} \quad V_{1,3} = \begin{pmatrix} 3 & -2 \end{pmatrix}. \quad (1.168)$$

Prove or disprove the following statement: It holds that

$$(\mathcal{R}_{\mathfrak{r}}^{\mathbf{R}}(\Phi, V))(-1) = 0 \quad (1.169)$$

(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], [62, Chapter 17] [65, Chapter 5], and [171, 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, 79, 80, 409] and the references therein), language generation (cf., for instance, [53, 176, 252, 361] and the references therein), and speech recognition (cf., for example, [6, 84, 177, 179, 381] and the references therein), time series prediction analysis including stock market prediction (cf., for instance, [136, 139, 393, 397] and the references therein) and weather prediction (cf., for example, [373, 396, 428] and the references therein) and video analysis (cf., for instance, [114, 249, 327, 422] and the references therein).

1.6.1 Description of RNNs

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

$$\mathfrak{R}_{f,T,\mathbb{I}}(x_1, x_2, \dots, x_T) = (y_1, y_2, \dots, y_T) \quad (1.170)$$

and we call $\mathfrak{R}_{f,T,\mathbb{I}}$ 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 $\mathfrak{N} = (\mathfrak{N}_\theta)_{\theta \in \mathbb{R}^\mathfrak{d}}: \mathbb{R}^\mathfrak{d} \times X \times I \rightarrow Y \times I$ 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*

$$R = \mathfrak{R}_{\mathfrak{N}_\theta, T, \mathbb{I}} \quad (1.171)$$

(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 $\mathfrak{x}, \mathfrak{y}, \mathfrak{i} \in \mathbb{N}$, $\theta \in \mathbb{R}^{(\mathfrak{x}+\mathfrak{i}+1)\mathfrak{i}+(\mathfrak{i}+1)\mathfrak{y}}$ and let $\Psi_1: \mathbb{R}^{\mathfrak{i}} \rightarrow \mathbb{R}^{\mathfrak{i}}$ and $\Psi_2: \mathbb{R}^{\mathfrak{y}} \rightarrow \mathbb{R}^{\mathfrak{y}}$ 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 $r: \mathbb{R}^{\mathfrak{x}} \times \mathbb{R}^{\mathfrak{i}} \rightarrow \mathbb{R}^{\mathfrak{y}} \times \mathbb{R}^{\mathfrak{i}}$ is the function from $\mathbb{R}^{\mathfrak{x}} \times \mathbb{R}^{\mathfrak{i}}$ to $\mathbb{R}^{\mathfrak{y}} \times \mathbb{R}^{\mathfrak{i}}$ which satisfies for all $x \in \mathbb{R}^{\mathfrak{x}}$, $i \in \mathbb{R}^{\mathfrak{i}}$ that

$$r(x, i) = \left((\mathcal{N}_{\Psi_1, \Psi_2}^{\theta, \mathfrak{x}+\mathfrak{i}})(x, i), (\mathcal{N}_{\Psi_1}^{\theta, \mathfrak{x}+\mathfrak{i}})(x, i) \right) \quad (1.172)$$

(cf. Definition 1.1.3).

Definition 1.6.4 (Vectorized description of simple fully-connected RNNs). Let $\mathfrak{x}, \mathfrak{y}, \mathfrak{i}, T \in \mathbb{N}$, $\theta \in \mathbb{R}^{(\mathfrak{x}+\mathfrak{i}+1)\mathfrak{i}+(\mathfrak{i}+1)\mathfrak{y}}$, $\mathbb{I} \in \mathbb{R}^{\mathfrak{i}}$ and let $\Psi_1: \mathbb{R}^{\mathfrak{i}} \rightarrow \mathbb{R}^{\mathfrak{i}}$ and $\Psi_2: \mathbb{R}^{\mathfrak{y}} \rightarrow \mathbb{R}^{\mathfrak{y}}$ 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 $r: \mathbb{R}^{\mathfrak{x}} \times \mathbb{R}^{\mathfrak{i}} \rightarrow \mathbb{R}^{\mathfrak{y}} \times \mathbb{R}^{\mathfrak{i}}$ 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

$$R = \mathfrak{R}_{r, T, \mathbb{I}} \quad (1.173)$$

(cf. Definitions 1.6.1 and 1.6.3).

Lemma 1.6.5. Let $\mathfrak{x}, \mathfrak{y}, \mathfrak{i}, \mathfrak{d}, T \in \mathbb{N}$, $\theta \in \mathbb{R}^{\mathfrak{d}}$, $\mathbb{I} \in \mathbb{R}^{\mathfrak{i}}$ satisfy $\mathfrak{d} = (\mathfrak{x} + \mathfrak{i} + 1)\mathfrak{i} + (\mathfrak{i} + 1)\mathfrak{y}$, let $\Psi_1: \mathbb{R}^{\mathfrak{i}} \rightarrow \mathbb{R}^{\mathfrak{i}}$ and $\Psi_2: \mathbb{R}^{\mathfrak{y}} \rightarrow \mathbb{R}^{\mathfrak{y}}$ be functions, and let $\mathfrak{N} = (\mathfrak{N}_{\vartheta})_{\vartheta \in \mathbb{R}^{\mathfrak{d}}} : \mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\mathfrak{x}} \times \mathbb{R}^{\mathfrak{i}} \rightarrow \mathbb{R}^{\mathfrak{y}} \times \mathbb{R}^{\mathfrak{i}}$ satisfy for all $\vartheta \in \mathbb{R}^{\mathfrak{d}}$ that \mathfrak{N}_{ϑ} 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*

$$R = \mathfrak{R}_{\mathfrak{N}_\theta, T, \mathbb{I}} \quad (1.174)$$

(cf. Definition 1.6.1).

Proof of Lemma 1.6.5. Observe that (1.171), (1.173), and (1.174) prove that ((i) \leftrightarrow (ii) \leftrightarrow (iii)). The proof of Lemma 1.6.5 is thus complete. \square

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 $\mathfrak{M}_{t,2}$ and $\text{id}_{\mathbb{R}}$, 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 $R_{T,\alpha}(1, 1, \dots, 1)$ 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 [211]. Loosely speaking, LSTM nodes were invented to attempt to 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, 348]). 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 [211] and, for instance, to the excellent explanations in [139, 176, 339]. For a few selected references on LSTM networks in the literature we refer, for example, to [11, 80, 139, 154, 155, 176, 178–181, 307, 350, 381, 388, 409, 446] 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 [51, 208, 210, 267, 377] for foundational references introducing and refining the idea of autoencoders,
- we refer, for example, to [423, 424, 437] for so-called *denoising autoencoders* which add random perturbation to the input data in the training of autoencoders,
- we refer, for instance, to [53, 113, 260] for so-called *variational autoencoders* which use techniques from bayesian statistics in the training of autoencoders,
- we refer, for example, [313, 370] for autoencoders involving convolutions, and
- we refer, for instance, [124, 311] for *adversarial autoencoders* which combine the principles of autoencoders with the paradigm of generative adversarial networks (see Goodfellow et al. [172]).

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 *Transformer ANNs* by the authors of the seminal paper Vaswani et al. [418] called "Attention is all you need". Roughly speaking, Transformer 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. Transformer ANNs are the basis for many recently very successful *large language models* (LLMs), such as, *generative pre-trained transformers* (GPTs) in [56, 340, 362, 363] which are the models behind the famous *ChatGPT* application, *Bidirectional Encoder Representations from Transformers* (BERT) models in Devlin et al. [110], and many others (cf., for example, [94, 281, 364, 439, 443] 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, [116, 254, 295, 425]), protein structure prediction (cf., for example, [246]), multimodal learning (cf., for instance, [301]), and long sequence time-series forecasting (cf., for example, [462]). Moreover, we refer, for instance, to [84, 307], [164, Chapter 17], and [171, 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 [432] 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 [383, 436, 460, 463] for overview articles on [GNNs](#),
- we refer, for example, to [173, 387] for foundational articles for [GNNs](#),
- we refer, for instance, to [420, 447] for applications of attention mechanisms (cf. Section 1.7.2 above) to [GNNs](#),
- we refer, for example, to [57, 98, 433, 445] for [GNNs](#) involving convolutions on graphs, and
- we refer, for instance, to [16, 158, 382, 389, 435] 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. [287] (cf. also Kovachki et al. [266]). 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. [286, 288] which are based on integral transformations with parametric integration kernels. We refer, for example, to [55, 265, 285, 431] 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. [302]. 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 [121, 174, 263, 275, 293, 316, 355, 413, 427, 434, 453] for extensions and theoretical results on [deepONets](#). For a comparison between [deepONets](#) and [FNOs](#) we refer, for example, to Lu et al. [303].

A further natural approach is to employ [CNNs](#) (see Section 1.4) to develop neural operator architectures. We refer, for instance, to [193, 202, 258, 371, 464] for such [CNN](#)-based neural operators. Finally, we refer, for example, to [69, 97, 101, 141, 142, 240, 289, 294, 320, 365, 390, 440] 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), parallelizations (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 [ANNs](#). The 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 [134, 166, 187, 188, 192, 241, 341, 349, 353] and the references therein. The specific presentation of this chapter is based on Grohs et al. [187, 188].

2.1 Compositions of ANNs

2.1.1 Compositions of ANNs

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

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

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

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

(cf. Definition 1.3.1).

2.1.2 Elementary properties of compositions of ANNs

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

(i) *it holds that*

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

(ii) *it holds that*

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

(iii) *it holds that*

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

(iv) *it holds that*

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

and

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

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

(cf. Definitions 1.3.4 and 2.1.1).

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

$$X_a = \left\{ x = (x_0, x_1, \dots, x_L) \in \mathbb{R}^{\mathbb{D}_0(\Phi \bullet \Psi)} \times \mathbb{R}^{\mathbb{D}_1(\Phi \bullet \Psi)} \times \dots \times \mathbb{R}^{\mathbb{D}_L(\Phi \bullet \Psi)} : \right. \\ \left. (\forall k \in \{1, 2, \dots, L\} : x_k = \mathfrak{M}_{a \mathbb{1}_{(0,L)}(k) + \text{id}_{\mathbb{R}} \mathbb{1}_{\{L\}}(k), \mathbb{D}_k(\Phi \bullet \Psi)}(\mathcal{W}_{k, \Phi \bullet \Psi} x_{k-1} + \mathcal{B}_{k, \Phi \bullet \Psi})) \right\}. \quad (2.8)$$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

2.1.3 Associativity of compositions of 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

$$(\Phi_1 \bullet \Phi_2) \bullet \Phi_3 = \Phi_1 \bullet (\Phi_2 \bullet \Phi_3) \quad (2.21)$$

(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

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

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

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

2.1. Compositions of ANNs

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

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

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

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

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

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

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

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

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

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

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

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

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

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

This and (2.23) imply that

$$(\Phi_1 \bullet \Phi_2) \bullet \Phi_3 = \Phi_1 \bullet (\Phi_2 \bullet \Phi_3). \quad (2.31)$$

The proof of Lemma 2.1.3 is thus complete. \square

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*

$$(\Phi_1 \bullet \Phi_2) \bullet \Phi_3 = \Phi_1 \bullet (\Phi_2 \bullet \Phi_3) \quad (2.32)$$

(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

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

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

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

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

$$\begin{aligned} &(\mathcal{W}_{k-\mathcal{L}(\Phi_3)+1,\Phi_1 \bullet \Phi_2}, \mathcal{B}_{k-\mathcal{L}(\Phi_3)+1,\Phi_1 \bullet \Phi_2}) \\ &= \begin{cases} (\mathcal{W}_{k-\mathcal{L}(\Phi_3)+1,\Phi_2}, \mathcal{B}_{k-\mathcal{L}(\Phi_3)+1,\Phi_2}) & : k - \mathcal{L}(\Phi_3) + 1 < \mathcal{L}(\Phi_2) \\ (\mathcal{W}_{1,\Phi_1} \mathcal{W}_{\mathcal{L}(\Phi_2),\Phi_2}, \mathcal{W}_{1,\Phi_1} \mathcal{B}_{\mathcal{L}(\Phi_2),\Phi_2} + \mathcal{B}_{1,\Phi_1}) & : k - \mathcal{L}(\Phi_3) + 1 = \mathcal{L}(\Phi_2) \\ (\mathcal{W}_{k-\mathcal{L}(\Phi_3)+1-\mathcal{L}(\Phi_2)+1,\Phi_1}, \mathcal{B}_{k-\mathcal{L}(\Phi_3)+1-\mathcal{L}(\Phi_2)+1,\Phi_1}) & : k - \mathcal{L}(\Phi_3) + 1 > \mathcal{L}(\Phi_2) \end{cases} \quad (2.35) \\ &88 \quad = \begin{cases} (\mathcal{W}_{k-\mathcal{L}(\Phi_3)+1,\Phi_2}, \mathcal{B}_{k-\mathcal{L}(\Phi_3)+1,\Phi_2}) & : k < \mathcal{L}(\Phi_2) + \mathcal{L}(\Phi_3) - 1 \\ (\mathcal{W}_{1,\Phi_1} \mathcal{W}_{\mathcal{L}(\Phi_2),\Phi_2}, \mathcal{W}_{1,\Phi_1} \mathcal{B}_{\mathcal{L}(\Phi_2),\Phi_2} + \mathcal{B}_{1,\Phi_1}) & : k = \mathcal{L}(\Phi_2) + \mathcal{L}(\Phi_3) - 1 \\ (\mathcal{W}_{k-\mathcal{L}(\Phi_3)-\mathcal{L}(\Phi_2)+2,\Phi_1}, \mathcal{B}_{k-\mathcal{L}(\Phi_3)-\mathcal{L}(\Phi_2)+2,\Phi_1}) & : k > \mathcal{L}(\Phi_2) + \mathcal{L}(\Phi_3) - 1. \end{cases} \end{aligned}$$

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

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

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 $k \in \{1, 2, \dots, \mathcal{L}(\Phi_1 \bullet (\Phi_2 \bullet \Phi_3))\}$ 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_1} \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}, \mathcal{W}_{1,\Phi_1} \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_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_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} \tag{2.37}
 \end{aligned}$$

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

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

$$(\Phi_1 \bullet \Phi_2) \bullet \Phi_3 = \Phi_1 \bullet (\Phi_2 \bullet \Phi_3) \quad (2.40)$$

(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. \square

2.1.4 Powers of ANNs

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

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

(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

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

(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. \square

2.2 Parallelizations of ANNs

2.2.1 Parallelizations of ANNs with the same length

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

$$\mathbf{P}_n: \{\Phi = (\Phi_1, \dots, \Phi_n) \in \mathbf{N}^n: \mathcal{L}(\Phi_1) = \mathcal{L}(\Phi_2) = \dots = \mathcal{L}(\Phi_n)\} \rightarrow \mathbf{N} \quad (2.43)$$

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

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

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

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

(2.44)

(cf. Definition 1.3.1).

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

(i) *it holds that*

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

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

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

and

(iii) *it holds that*

$$\mathcal{D}(\mathbf{P}_n(\Phi)) = \mathcal{D}(\Phi_1) + \mathcal{D}(\Phi_2) + \dots + \mathcal{D}(\Phi_n)$$
(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 $k \in \{1, 2, \dots, L\}$ it holds that

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

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

(i) *it holds that*

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

and

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

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

(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 $j \in \{1, 2, \dots, n\}$ let

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

and let

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

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

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

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

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

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

(x^1, x^2, \dots, x^n) it holds that

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

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

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

Induction, and (1.97) hence demonstrate that for all $k \in \{1, 2, \dots, L\}$, $\mathfrak{x} = (\mathfrak{x}_0, \mathfrak{x}_1, \dots, \mathfrak{x}_L) \in \mathfrak{X}$, $x^1 = (x_0^1, x_1^1, \dots, x_L^1) \in X^1$, $x^2 = (x_0^2, x_1^2, \dots, x_L^2) \in X^2$, \dots , $x^n = (x_0^n, x_1^n, \dots, x_L^n) \in X^n$ with $\mathfrak{x}_0 = (x_0^1, x_0^2, \dots, x_0^n)$ it holds that

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

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

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

$$\mathcal{P}(\mathbf{P}_n(\Phi_1, \Phi_2, \dots, \Phi_n)) \leq \frac{1}{2} [\sum_{j=1}^n \mathcal{P}(\Phi_j)]^2 \quad (2.58)$$

(cf. Definition 2.2.1).

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

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

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

The proof of Proposition 2.2.4 is thus complete. \square

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

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

(cf. Definition 2.2.1).

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

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

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

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

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

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

Hence, we obtain that

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

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

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

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

$$2(ab + 1) = ab + 1 + (a - 1)(b - 1) + a + b \geq ab + a + b + 1 = (a + 1)(b + 1) \quad (2.66)$$

show that

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

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

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

$$\mathcal{P}(\mathbf{P}_n(\Phi_1, \Phi_2, \dots, \Phi_n)) \leq n \left[\sum_{i=1}^n \mathcal{P}(\Phi_i) \right]. \quad (2.68)$$

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

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

2.2.2 ANN representations for the identities

In this section, we show how multi-dimensional identities can be represented as realizations of ANNs with ReLU and softplus activation functions. In Section 4.2.2 we will also show such a representation for the RePU activation function. For this further ANN calculus tools will be needed which are introduced later in this chapter.

Definition 2.2.6 (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

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

and

$$\mathfrak{I}_d = \mathbf{P}_d(\mathfrak{I}_1, \mathfrak{I}_1, \dots, \mathfrak{I}_1) \quad (2.71)$$

(cf. Definitions 1.3.1 and 2.2.1).

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

(i) it holds that

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

and

(ii) it holds that

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

(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

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

This, (2.71), and Lemma 2.2.2 prove that

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

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

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

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

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

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

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

$$\mathcal{R}_a^{\mathbf{N}}(\mathfrak{I}_d) = \text{id}_{\mathbb{R}^d} \quad (2.78)$$

(cf. Definitions 1.3.4 and 2.2.6).

Proof of Lemma 2.2.8. Note that (1.49) and (2.70) ensure that for all $x \in \mathbb{R}$ it holds that

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

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

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

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

2.2.3 Extensions of ANNs

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

$$\mathcal{E}_{L,\mathbb{I}}: \{\Phi \in \mathbf{N}: (\mathcal{L}(\Phi) \leq L \text{ and } \mathcal{O}(\Phi) = \mathcal{I}(\mathbb{I}))\} \rightarrow \mathbf{N} \quad (2.81)$$

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

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

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

Lemma 2.2.10 (Length of extensions of 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

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

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

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

(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

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

(cf. Definition 2.1.6). Combining this with (1.84) and Lemma 1.3.3 implies that

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

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

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

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

$$\Psi^{\bullet 0} = (I_d, 0) \in \mathbb{R}^{d \times d} \times \mathbb{R}^d \quad (2.88)$$

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

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

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

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

(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

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

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

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

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

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

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

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

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*

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

(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

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

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 $\forall x \in \mathbb{R}^d: (\mathcal{R}_a^{\mathbf{N}}(\mathbb{I}^{\bullet 0}))(x) = x$. 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 $\forall x \in \mathbb{R}^d: (\mathcal{R}_a^{\mathbf{N}}(\mathbb{I}^{\bullet n}))(x) = x$ it holds that

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

and

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

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

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

and

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

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

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

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

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

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

(cf. Definition 2.2.9).

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

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

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

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

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

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

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

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

2.2.4 Parallelizations of ANNs with different lengths

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

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

(cf. Definition 1.3.1). Then we denote by

$$P_{n,\Psi}: \{\Phi = (\Phi_1, \dots, \Phi_n) \in \mathbf{N}^n : (\forall j \in \{1, 2, \dots, n\}: \mathcal{O}(\Phi_j) = \mathcal{I}(\Psi_j))\} \rightarrow \mathbf{N} \quad (2.108)$$

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

$$P_{n,\Psi}(\Phi) = P_n(\mathcal{E}_{\max_{k \in \{1, 2, \dots, n\}} \mathcal{L}(\Phi_k), \Psi_1}(\Phi_1), \dots, \mathcal{E}_{\max_{k \in \{1, 2, \dots, n\}} \mathcal{L}(\Phi_k), \Psi_n}(\Phi_n)) \quad (2.109)$$

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

Lemma 2.2.14 (Realizations for parallelizations of ANNs with different length). Let $a \in C(\mathbb{R}, \mathbb{R})$, $n \in \mathbb{N}$, $\mathbb{I} = (\mathbb{I}_1, \dots, \mathbb{I}_n)$, $\Phi = (\Phi_1, \dots, \Phi_n) \in \mathbf{N}^n$ satisfy for all $j \in \{1, 2, \dots, n\}$, $x \in \mathbb{R}^{\mathcal{O}(\Phi_j)}$ that $\mathcal{H}(\mathbb{I}_j) = 1$, $\mathcal{I}(\mathbb{I}_j) = \mathcal{O}(\mathbb{I}_j) = \mathcal{O}(\Phi_j)$, and $(\mathcal{R}_a^{\mathbf{N}}(\mathbb{I}_j))(x) = x$ (cf. Definitions 1.3.1 and 1.3.4). Then

(i) it holds that

$$\mathcal{R}_a^{\mathbf{N}}(P_{n,\mathbb{I}}(\Phi)) \in C(\mathbb{R}^{[\sum_{j=1}^n \mathcal{I}(\Phi_j)]}, \mathbb{R}^{[\sum_{j=1}^n \mathcal{O}(\Phi_j)]}) \quad (2.110)$$

and

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

$$\begin{aligned} & (\mathcal{R}_a^{\mathbf{N}}(P_{n,\mathbb{I}}(\Phi)))(x_1, x_2, \dots, x_n) \\ &= ((\mathcal{R}_a^{\mathbf{N}}(\Phi_1))(x_1), (\mathcal{R}_a^{\mathbf{N}}(\Phi_2))(x_2), \dots, (\mathcal{R}_a^{\mathbf{N}}(\Phi_n))(x_n)) \in \mathbb{R}^{[\sum_{j=1}^n \mathcal{O}(\Phi_j)]} \end{aligned} \quad (2.111)$$

(cf. Definition 2.2.13).

Proof of Lemma 2.2.14. Throughout this proof, let $L \in \mathbb{N}$ satisfy $L = \max_{j \in \{1, 2, \dots, n\}} \mathcal{L}(\Phi_j)$. Note that item (ii) in Lemma 2.2.10, the assumption that for all $j \in \{1, 2, \dots, n\}$ 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 $j \in \{1, 2, \dots, n\}$ it holds that $\mathcal{L}(\mathcal{E}_{L,\mathbb{I}_j}(\Phi_j)) = L$ and $\mathcal{R}_a^{\mathbf{N}}(\mathcal{E}_{L,\mathbb{I}_j}(\Phi_j)) \in C(\mathbb{R}^{\mathcal{I}(\Phi_j)}, \mathbb{R}^{\mathcal{O}(\Phi_j)})$ and
- (II) that for all $j \in \{1, 2, \dots, n\}$, $x \in \mathbb{R}^{\mathcal{I}(\Phi_j)}$ it holds that

$$(\mathcal{R}_a^{\mathbf{N}}(\mathcal{E}_{L,\mathbb{I}_j}(\Phi_j)))(x) = (\mathcal{R}_a^{\mathbf{N}}(\Phi_j))(x) \quad (2.112)$$

(cf. Definition 2.2.9). Items (i) and (ii) in Proposition 2.2.3 therefore imply

(A) that

$$\mathcal{R}_a^N(\mathbf{P}_n(\mathcal{E}_{L,\mathbb{I}_1}(\Phi_1), \mathcal{E}_{L,\mathbb{I}_2}(\Phi_2), \dots, \mathcal{E}_{L,\mathbb{I}_n}(\Phi_n))) \in C(\mathbb{R}^{[\sum_{j=1}^n \mathcal{I}(\Phi_j)]}, \mathbb{R}^{[\sum_{j=1}^n \mathcal{O}(\Phi_j)]}) \quad (2.113)$$

and

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

$$\begin{aligned} & (\mathcal{R}_a^N(\mathbf{P}_n(\mathcal{E}_{L,\mathbb{I}_1}(\Phi_1), \mathcal{E}_{L,\mathbb{I}_2}(\Phi_2), \dots, \mathcal{E}_{L,\mathbb{I}_n}(\Phi_n))))(x_1, x_2, \dots, x_n) \\ &= \left((\mathcal{R}_a^N(\mathcal{E}_{L,\mathbb{I}_1}(\Phi_1)))(x_1), (\mathcal{R}_a^N(\mathcal{E}_{L,\mathbb{I}_2}(\Phi_2)))(x_2), \dots, (\mathcal{R}_a^N(\mathcal{E}_{L,\mathbb{I}_n}(\Phi_n)))(x_n) \right) \quad (2.114) \\ &= \left((\mathcal{R}_a^N(\Phi_1))(x_1), (\mathcal{R}_a^N(\Phi_2))(x_2), \dots, (\mathcal{R}_a^N(\Phi_n))(x_n) \right) \end{aligned}$$

(cf. Definition 2.2.1). Combining this with (2.109) and the fact that $L = \max_{j \in \{1, 2, \dots, n\}} \mathcal{L}(\Phi_j)$ ensures

(C) that

$$\mathcal{R}_a^N(\mathbf{P}_{n,\mathbb{I}}(\Phi)) \in C(\mathbb{R}^{[\sum_{j=1}^n \mathcal{I}(\Phi_j)]}, \mathbb{R}^{[\sum_{j=1}^n \mathcal{O}(\Phi_j)]}) \quad (2.115)$$

and

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

$$\begin{aligned} & (\mathcal{R}_a^N(\mathbf{P}_{n,\mathbb{I}}(\Phi)))(x_1, x_2, \dots, x_n) \\ &= (\mathcal{R}_a^N(\mathbf{P}_n(\mathcal{E}_{L,\mathbb{I}_1}(\Phi_1), \mathcal{E}_{L,\mathbb{I}_2}(\Phi_2), \dots, \mathcal{E}_{L,\mathbb{I}_n}(\Phi_n))))(x_1, x_2, \dots, x_n) \quad (2.116) \\ &= \left((\mathcal{R}_a^N(\Phi_1))(x_1), (\mathcal{R}_a^N(\Phi_2))(x_2), \dots, (\mathcal{R}_a^N(\Phi_n))(x_n) \right). \end{aligned}$$

This establishes items items (i) and (ii). The proof of Lemma 2.2.14 is thus complete. \square

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

$$F_d(x) = (\max\{|x_1|\}, \max\{|x_1|, |x_2|\}, \dots, \max\{|x_1|, |x_2|, \dots, |x_d|\}). \quad (2.117)$$

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

$$\mathcal{R}_r^N(\Phi) = F_d \quad (2.118)$$

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

2.3 Scalar multiplications of ANNs

2.3.1 Affine transformations as ANNs

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

$$\mathbf{A}_{W,B} \in (\mathbb{R}^{m \times n} \times \mathbb{R}^m) \subseteq \mathbf{N} \quad (2.119)$$

the fully-connected feedforward [ANN](#) given by

$$\mathbf{A}_{W,B} = (W, B) \quad (2.120)$$

(cf. Definitions [1.3.1](#) and [1.3.2](#)).

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

- (i) it holds that $\mathcal{D}(\mathbf{A}_{W,B}) = (n, m) \in \mathbb{N}^2$,
- (ii) it holds for all $a \in C(\mathbb{R}, \mathbb{R})$ that $\mathcal{R}_a^{\mathbf{N}}(\mathbf{A}_{W,B}) \in C(\mathbb{R}^n, \mathbb{R}^m)$, and
- (iii) it holds for all $a \in C(\mathbb{R}, \mathbb{R})$, $x \in \mathbb{R}^n$ that

$$(\mathcal{R}_a^{\mathbf{N}}(\mathbf{A}_{W,B}))(x) = Wx + B \quad (2.121)$$

(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

$$\mathcal{D}(\mathbf{A}_{W,B}) = (n, m) \in \mathbb{N}^2. \quad (2.122)$$

This establishes item (i). Furthermore, observe that the fact that

$$\mathbf{A}_{W,B} = (W, B) \in (\mathbb{R}^{m \times n} \times \mathbb{R}^m) \quad (2.123)$$

and [\(1.97\)](#) imply that for all $a \in C(\mathbb{R}, \mathbb{R})$, $x \in \mathbb{R}^n$ it holds that $\mathcal{R}_a^{\mathbf{N}}(\mathbf{A}_{W,B}) \in C(\mathbb{R}^n, \mathbb{R}^m)$ and

$$(\mathcal{R}_a^{\mathbf{N}}(\mathbf{A}_{W,B}))(x) = Wx + B. \quad (2.124)$$

This proves items (ii) and (iii). The proof of Lemma [2.3.2](#) is thus complete. \square

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

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

$$\mathcal{D}(\mathbf{A}_{W,B} \bullet \Phi) = (\mathbb{D}_0(\Phi), \mathbb{D}_1(\Phi), \dots, \mathbb{D}_{\mathcal{H}(\Phi)}(\Phi), m), \quad (2.125)$$

(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 $\mathcal{R}_a^{\mathbf{N}}(\mathbf{A}_{W,B} \bullet \Phi) \in C(\mathbb{R}^{\mathcal{I}(\Phi)}, \mathbb{R}^m)$,

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

$$(\mathcal{R}_a^{\mathbf{N}}(\mathbf{A}_{W,B} \bullet \Phi))(x) = W((\mathcal{R}_a^{\mathbf{N}}(\Phi))(x)) + B, \quad (2.126)$$

(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

$$\mathcal{D}(\Phi \bullet \mathbf{A}_{W,B}) = (n, \mathbb{D}_1(\Phi), \mathbb{D}_2(\Phi), \dots, \mathbb{D}_{\mathcal{L}(\Phi)}(\Phi)), \quad (2.127)$$

(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 $\mathcal{R}_a^{\mathbf{N}}(\Phi \bullet \mathbf{A}_{W,B}) \in C(\mathbb{R}^n, \mathbb{R}^{\mathcal{O}(\Phi)})$, and

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

$$(\mathcal{R}_a^{\mathbf{N}}(\Phi \bullet \mathbf{A}_{W,B}))(x) = (\mathcal{R}_a^{\mathbf{N}}(\Phi))(Wx + B) \quad (2.128)$$

(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 $\mathcal{R}_a^{\mathbf{N}}(\mathbf{A}_{W,B}) \in C(\mathbb{R}^n, \mathbb{R}^m)$ and

$$(\mathcal{R}_a^{\mathbf{N}}(\mathbf{A}_{W,B}))(x) = Wx + B \quad (2.129)$$

(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. \square

2.3.2 Scalar multiplications of ANNs

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

$$\lambda \circledast \Phi = \mathbf{A}_{\lambda I_{\mathcal{O}(\Phi)}, 0} \bullet \Phi \quad (2.130)$$

(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 $\mathcal{R}_a^{\mathbf{N}}(\lambda \circledast \Phi) \in C(\mathbb{R}^{\mathcal{I}(\Phi)}, \mathbb{R}^{\mathcal{O}(\Phi)})$, and

2.4. Sums of ANNs with the same length

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

$$\mathcal{R}_a^N(\lambda \circledast \Phi) = \lambda [\mathcal{R}_a^N(\Phi)] \quad (2.131)$$

(cf. Definitions 1.3.4 and 2.3.4).

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

$$L = \mathcal{L}(\Phi) \quad \text{and} \quad (l_0, l_1, \dots, l_L) = \mathcal{D}(\Phi). \quad (2.132)$$

Observe that item (i) in Lemma 2.3.2 shows that

$$\mathcal{D}(\mathbf{A}_{\lambda I_{\mathcal{O}(\Phi)}, 0}) = (\mathcal{O}(\Phi), \mathcal{O}(\Phi)) \quad (2.133)$$

(cf. Definitions 1.5.5 and 2.3.1). Combining this and item (i) in Lemma 2.3.3 ensures that

$$\mathcal{D}(\lambda \circledast \Phi) = \mathcal{D}(\mathbf{A}_{\lambda I_{\mathcal{O}(\Phi)}, 0} \bullet \Phi) = (l_0, l_1, \dots, l_{L-1}, \mathcal{O}(\Phi)) = \mathcal{D}(\Phi) \quad (2.134)$$

(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 $\mathcal{R}_a^N(\lambda \circledast \Phi) \in C(\mathbb{R}^{\mathcal{I}(\Phi)}, \mathbb{R}^{\mathcal{O}(\Phi)})$ and

$$\begin{aligned} (\mathcal{R}_a^N(\lambda \circledast \Phi))(x) &= (\mathcal{R}_a^N(\mathbf{A}_{\lambda I_{\mathcal{O}(\Phi)}, 0} \bullet \Phi))(x) \\ &= \lambda I_{\mathcal{O}(\Phi)}((\mathcal{R}_a^N(\Phi))(x)) \\ &= \lambda ((\mathcal{R}_a^N(\Phi))(x)) \end{aligned} \quad (2.135)$$

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

2.4 Sums of ANNs with the same length

2.4.1 Sums of vectors as ANNs

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

$$\mathbb{S}_{m,n} \in (\mathbb{R}^{m \times (mn)} \times \mathbb{R}^m) \subseteq \mathbf{N} \quad (2.136)$$

the ANN given by

$$\mathbb{S}_{m,n} = \mathbf{A}_{(I_m \ I_m \ \dots \ I_m), 0} \quad (2.137)$$

(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 $\mathcal{R}_a^{\mathbf{N}}(\mathbb{S}_{m,n}) \in C(\mathbb{R}^{mn}, \mathbb{R}^m)$, and
- (iii) it holds for all $a \in C(\mathbb{R}, \mathbb{R})$, $x_1, x_2, \dots, x_n \in \mathbb{R}^m$ that

$$(\mathcal{R}_a^{\mathbf{N}}(\mathbb{S}_{m,n}))(x_1, x_2, \dots, x_n) = \sum_{k=1}^n x_k \quad (2.138)$$

(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

$$\mathcal{D}(\mathbb{S}_{m,n}) = (mn, m) \in \mathbb{N}^2 \quad (2.139)$$

(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 $a \in C(\mathbb{R}, \mathbb{R})$, $x_1, x_2, \dots, x_n \in \mathbb{R}^m$ it holds that $\mathcal{R}_a^{\mathbf{N}}(\mathbb{S}_{m,n}) \in C(\mathbb{R}^{mn}, \mathbb{R}^m)$ and

$$\begin{aligned} (\mathcal{R}_a^{\mathbf{N}}(\mathbb{S}_{m,n}))(x_1, x_2, \dots, x_n) &= (\mathcal{R}_a^{\mathbf{N}}(\mathbf{A}_{(I_m \ I_m \ \dots \ I_m), 0}))(x_1, x_2, \dots, x_n) \\ &= (I_m \ I_m \ \dots \ I_m)(x_1, x_2, \dots, x_n) = \sum_{k=1}^n x_k \end{aligned} \quad (2.140)$$

(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. \square

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 $\mathcal{R}_a^{\mathbf{N}}(\mathbb{S}_{m,n} \bullet \Phi) \in C(\mathbb{R}^{\mathcal{I}(\Phi)}, \mathbb{R}^m)$ and
- (ii) it holds for all $x \in \mathbb{R}^{\mathcal{I}(\Phi)}$, $y_1, y_2, \dots, y_n \in \mathbb{R}^m$ with $(\mathcal{R}_a^{\mathbf{N}}(\Phi))(x) = (y_1, y_2, \dots, y_n)$ that

$$(\mathcal{R}_a^{\mathbf{N}}(\mathbb{S}_{m,n} \bullet \Phi))(x) = \sum_{k=1}^n y_k \quad (2.141)$$

(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 $x_1, x_2, \dots, x_n \in \mathbb{R}^m$ it holds that $\mathcal{R}_a^{\mathbf{N}}(\mathbb{S}_{m,n}) \in C(\mathbb{R}^{mn}, \mathbb{R}^m)$ and

$$(\mathcal{R}_a^{\mathbf{N}}(\mathbb{S}_{m,n}))(x_1, x_2, \dots, x_n) = \sum_{k=1}^n x_k \quad (2.142)$$

(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. \square

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 $x_1, x_2, \dots, x_n \in \mathbb{R}^{\mathcal{I}(\Phi)}$ that*

$$(\mathcal{R}_a^{\mathbf{N}}(\Phi \bullet \mathbb{S}_{\mathcal{I}(\Phi), n}))(x_1, x_2, \dots, x_n) = (\mathcal{R}_a^{\mathbf{N}}(\Phi))\left(\sum_{k=1}^n x_k\right) \quad (2.143)$$

(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 $m \in \mathbb{N}$, $x_1, x_2, \dots, x_n \in \mathbb{R}^m$ it holds that $\mathcal{R}_a^{\mathbf{N}}(\mathbb{S}_{m,n}) \in C(\mathbb{R}^{mn}, \mathbb{R}^m)$ and

$$(\mathcal{R}_a^{\mathbf{N}}(\mathbb{S}_{m,n}))(x_1, x_2, \dots, x_n) = \sum_{k=1}^n x_k \quad (2.144)$$

(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. \square

2.4.2 Concatenation of vectors as ANNs

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

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

$$\mathbb{T}_{m,n} \in (\mathbb{R}^{(mn) \times m} \times \mathbb{R}^{mn}) \subseteq \mathbf{N} \quad (2.145)$$

the fully-connected feedforward ANN given by

$$\mathbb{T}_{m,n} = \mathbf{A}_{(I_m \ I_m \ \dots \ I_m)^*, 0} \quad (2.146)$$

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

$$(\mathcal{R}_a^{\mathbf{N}}(\mathbb{T}_{m,n}))(x) = (x, x, \dots, x) \quad (2.147)$$

(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

$$\mathcal{D}(\mathbb{T}_{m,n}) = (m, mn) \in \mathbb{N}^2 \quad (2.148)$$

(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 $\mathcal{R}_a^{\mathbf{N}}(\mathbb{T}_{m,n}) \in C(\mathbb{R}^m, \mathbb{R}^{mn})$ and

$$\begin{aligned} (\mathcal{R}_a^{\mathbf{N}}(\mathbb{T}_{m,n}))(x) &= (\mathcal{R}_a^{\mathbf{N}}(\mathbf{A}_{(I_m \ I_m \ \dots \ I_m)^*, 0}))(x) \\ &= (I_m \ I_m \ \dots \ I_m)^* x = (x, x, \dots, x) \end{aligned} \quad (2.149)$$

(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. \square

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 $\mathcal{R}_a^{\mathbf{N}}(\mathbb{T}_{\mathcal{O}(\Phi), n} \bullet \Phi) \in C(\mathbb{R}^{\mathcal{I}(\Phi)}, \mathbb{R}^{n\mathcal{O}(\Phi)})$ and
- (ii) it holds for all $x \in \mathbb{R}^{\mathcal{I}(\Phi)}$ that

$$(\mathcal{R}_a^{\mathbf{N}}(\mathbb{T}_{\mathcal{O}(\Phi), n} \bullet \Phi))(x) = ((\mathcal{R}_a^{\mathbf{N}}(\Phi))(x), (\mathcal{R}_a^{\mathbf{N}}(\Phi))(x), \dots, (\mathcal{R}_a^{\mathbf{N}}(\Phi))(x)) \quad (2.150)$$

(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 $\mathcal{R}_a^{\mathbf{N}}(\mathbb{T}_{m,n}) \in C(\mathbb{R}^m, \mathbb{R}^{mn})$ and

$$(\mathcal{R}_a^{\mathbf{N}}(\mathbb{T}_{m,n}))(x) = (x, x, \dots, x) \quad (2.151)$$

(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. \square

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

$$(\mathcal{R}_a^{\mathbf{N}}(\Phi \bullet \mathbb{T}_{m,n}))(x) = (\mathcal{R}_a^{\mathbf{N}}(\Phi))(x, x, \dots, x) \quad (2.152)$$

(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

$$(\mathcal{R}_a^{\mathbf{N}}(\mathbb{T}_{m,n}))(x) = (x, x, \dots, x) \quad (2.153)$$

(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. \square

2.4.3 Sums of ANNs

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

$$\mathcal{L}(\Phi_k) = \mathcal{L}(\Phi_m), \quad \mathcal{I}(\Phi_k) = \mathcal{I}(\Phi_m), \quad \text{and} \quad \mathcal{O}(\Phi_k) = \mathcal{O}(\Phi_m) \quad (2.154)$$

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

$$\bigoplus_{k=m}^n \Phi_k = (\mathbb{S}_{\mathcal{O}(\Phi_m), n-m+1} \bullet [\mathbf{P}_{n-m+1}(\Phi_m, \Phi_{m+1}, \dots, \Phi_n)] \bullet \mathbb{T}_{\mathcal{I}(\Phi_m), n-m+1}) \in \mathbf{N} \quad (2.155)$$

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

$$\mathcal{L}(\Phi_k) = \mathcal{L}(\Phi_m), \quad \mathcal{I}(\Phi_k) = \mathcal{I}(\Phi_m), \quad \text{and} \quad \mathcal{O}(\Phi_k) = \mathcal{O}(\Phi_m) \quad (2.156)$$

(cf. Definition 1.3.1). Then

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

(ii) it holds that

$$\mathcal{D}\left(\bigoplus_{k=m}^n \Phi_k\right) = \left(\mathcal{I}(\Phi_m), \sum_{k=m}^n \mathbb{D}_1(\Phi_k), \sum_{k=m}^n \mathbb{D}_2(\Phi_k), \dots, \sum_{k=m}^n \mathbb{D}_{\mathcal{L}(\Phi_m)-1}(\Phi_k), \mathcal{O}(\Phi_m)\right), \quad (2.157)$$

and

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

$$\mathcal{R}_a^N\left(\bigoplus_{k=m}^n \Phi_k\right) = \sum_{k=m}^n (\mathcal{R}_a^N(\Phi_k)) \quad (2.158)$$

(cf. Definitions 1.3.4 and 2.4.10).

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

$$\begin{aligned} & \mathcal{D}(\mathbf{P}_{n-m+1}(\Phi_m, \Phi_{m+1}, \dots, \Phi_n)) \\ &= \left(\sum_{k=m}^n \mathbb{D}_0(\Phi_k), \sum_{k=m}^n \mathbb{D}_1(\Phi_k), \dots, \sum_{k=m}^n \mathbb{D}_{\mathcal{L}(\Phi_m)-1}(\Phi_k), \sum_{k=m}^n \mathbb{D}_{\mathcal{L}(\Phi_m)}(\Phi_k) \right) \\ &= \left((n-m+1)\mathcal{I}(\Phi_m), \sum_{k=m}^n \mathbb{D}_1(\Phi_k), \sum_{k=m}^n \mathbb{D}_2(\Phi_k), \dots, \sum_{k=m}^n \mathbb{D}_{\mathcal{L}(\Phi_m)-1}(\Phi_k), \right. \\ &\quad \left. (n-m+1)\mathcal{O}(\Phi_m) \right) \end{aligned} \quad (2.159)$$

(cf. Definition 2.2.1). Furthermore, note that item (i) in Lemma 2.4.2 shows that

$$\mathcal{D}(\mathbb{S}_{\mathcal{O}(\Phi_m), n-m+1}) = ((n-m+1)\mathcal{O}(\Phi_m), \mathcal{O}(\Phi_m)) \quad (2.160)$$

(cf. Definition 2.4.1). This, (2.159), and item (i) in Proposition 2.1.2 ensure that

$$\begin{aligned} & \mathcal{D}(\mathbb{S}_{\mathcal{O}(\Phi_m), n-m+1} \bullet [\mathbf{P}_{n-m+1}(\Phi_m, \Phi_{m+1}, \dots, \Phi_n)]) \\ &= \left((n-m+1)\mathcal{I}(\Phi_m), \sum_{k=m}^n \mathbb{D}_1(\Phi_k), \sum_{k=m}^n \mathbb{D}_2(\Phi_k), \dots, \sum_{k=m}^n \mathbb{D}_{\mathcal{L}(\Phi_m)-1}(\Phi_k), \mathcal{O}(\Phi_m) \right). \end{aligned} \quad (2.161)$$

Moreover, observe that item (i) in Lemma 2.4.7 proves that

$$\mathcal{D}(\mathbb{T}_{\mathcal{I}(\Phi_m), n-m+1}) = (\mathcal{I}(\Phi_m), (n-m+1)\mathcal{I}(\Phi_m)) \quad (2.162)$$

(cf. Definitions 2.1.1 and 2.4.6). Combining this, (2.161), and item (i) in Proposition 2.1.2 implies that

$$\begin{aligned} & \mathcal{D}\left(\bigoplus_{k=m}^n \Phi_k\right) \\ &= \mathcal{D}(\mathbb{S}_{\mathcal{O}(\Phi_m), (n-m+1)} \bullet [\mathbf{P}_{n-m+1}(\Phi_m, \Phi_{m+1}, \dots, \Phi_n)] \bullet \mathbb{T}_{\mathcal{I}(\Phi_m), (n-m+1)}) \\ &= \left(\mathcal{I}(\Phi_m), \sum_{k=m}^n \mathbb{D}_1(\Phi_k), \sum_{k=m}^n \mathbb{D}_2(\Phi_k), \dots, \sum_{k=m}^n \mathbb{D}_{\mathcal{L}(\Phi_m)-1}(\Phi_k), \mathcal{O}(\Phi_m) \right) \end{aligned} \quad (2.163)$$

(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

$$\mathcal{R}_a^N([\mathbf{P}_{n-m+1}(\Phi_m, \Phi_{m+1}, \dots, \Phi_n)] \bullet \mathbb{T}_{\mathcal{I}(\Phi_m), n-m+1}) \in C(\mathbb{R}^{\mathcal{I}(\Phi_m)}, \mathbb{R}^{(n-m+1)\mathcal{O}(\Phi_m)}) \quad (2.164)$$

and

$$\begin{aligned} & (\mathcal{R}_a^N([\mathbf{P}_{n-m+1}(\Phi_m, \Phi_{m+1}, \dots, \Phi_n)] \bullet \mathbb{T}_{\mathcal{I}(\Phi_m), n-m+1}))(x) \\ &= (\mathcal{R}_a^N(\mathbf{P}_{n-m+1}(\Phi_m, \Phi_{m+1}, \dots, \Phi_n)))(x, x, \dots, x) \end{aligned} \quad (2.165)$$

(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

$$\begin{aligned} & (\mathcal{R}_a^N([\mathbf{P}_{n-m+1}(\Phi_m, \Phi_{m+1}, \dots, \Phi_n)] \bullet \mathbb{T}_{\mathcal{I}(\Phi_m), n-m+1}))(x) \\ &= ((\mathcal{R}_a^N(\Phi_m))(x), (\mathcal{R}_a^N(\Phi_{m+1}))(x), \dots, (\mathcal{R}_a^N(\Phi_n))(x)) \in \mathbb{R}^{(n-m+1)\mathcal{O}(\Phi_m)}. \end{aligned} \quad (2.166)$$

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 $\mathcal{R}_a^N(\bigoplus_{k=m}^n \Phi_k) \in C(\mathbb{R}^{\mathcal{I}(\Phi_m)}, \mathbb{R}^{\mathcal{O}(\Phi_m)})$ and

$$\begin{aligned} & \left(\mathcal{R}_a^N \left(\bigoplus_{k=m}^n \Phi_k \right) \right) (x) \\ &= (\mathcal{R}_a^N(\mathbb{S}_{\mathcal{O}(\Phi_m), n-m+1} \bullet [\mathbf{P}_{n-m+1}(\Phi_m, \Phi_{m+1}, \dots, \Phi_n)] \bullet \mathbb{T}_{\mathcal{I}(\Phi_m), n-m+1}))(x) \\ &= \sum_{k=m}^n (\mathcal{R}_a^N(\Phi_k))(x). \end{aligned} \quad (2.167)$$

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

Part II

Approximation

Chapter 3

One-dimensional ANN approximation results

In learning problems **ANNs** are heavily used with the aim to approximate certain target functions. In this chapter we review basic **ReLU ANN** approximation results for a class of one-dimensional target functions (see Section 3.3). **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, 43, 92, 214, 215] for early universal **ANN** approximation results, cf., for example, [28, 45, 182, 353, 395, 444] 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, [134, 186, 273, 391] 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. [220, 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 $f: A \rightarrow \mathbb{R}$ be a function. Then we denote by $w_f: [0, \infty] \rightarrow [0, \infty]$ the function which satisfies for all*

$h \in [0, \infty]$ that

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

and we call w_f the modulus of continuity of f .

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

- (i) *it holds that w_f is non-decreasing,*
 - (ii) *it holds that f is uniformly continuous if and only if $\lim_{h \searrow 0} w_f(h) = 0$,*
 - (iii) *it holds that f is globally bounded if and only if $w_f(\infty) < \infty$, and*
 - (iv) *it holds for all $x, y \in A$ that $|f(x) - f(y)| \leq w_f(|x - y|)$*
- (cf. Definition 3.1.1).*

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

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

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

(cf. Definition 3.1.1).

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

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

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

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

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

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

The proof of Lemma 3.1.3 is thus complete. \square

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

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

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

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

(cf. Definition 3.1.1).

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

$$\begin{aligned} w_f(h) &= \sup\left(\left[\cup_{x,y \in A, |x-y| \leq h} \{|f(x) - f(y)|\}\right] \cup \{0\}\right) \\ &\leq \sup\left(\left[\cup_{x,y \in A, |x-y| \leq h} \{L|x - y|\}\right] \cup \{0\}\right) \\ &\leq \sup(\{Lh, 0\}) = Lh \end{aligned} \quad (3.8)$$

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

3.1.2 Linear interpolation of one-dimensional functions

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

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

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

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

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

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

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

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

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

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

and

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

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

(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 $k \in \{1, 2, \dots, K\}$, $x \in [\mathfrak{x}_{k-1}, \mathfrak{x}_k]$ it holds that

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

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

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

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

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

and

(ii) it holds that

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

(cf. Definitions 3.1.1 and 3.1.5).

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

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

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

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

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

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

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

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

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

$$|\mathbf{l}(x) - \mathbf{l}(y)| \leq L|x - y|. \quad (3.22)$$

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

$$|\mathbf{l}(x) - \mathbf{l}(y)| = 0 \leq L|x - y|, \quad (3.23)$$

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

$$|\mathbf{l}(x) - \mathbf{l}(y)| = 0 \leq L|x - y|, \quad (3.24)$$

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

$$|\mathbf{l}(x) - \mathbf{l}(y)| \leq L|x - y|. \quad (3.25)$$

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

$$\begin{aligned} |\mathfrak{l}(x) - f(x)| &= \left| \left(\frac{\mathfrak{x}_k - x}{\mathfrak{x}_k - \mathfrak{x}_{k-1}} \right) f(\mathfrak{x}_{k-1}) + \left(\frac{x - \mathfrak{x}_{k-1}}{\mathfrak{x}_k - \mathfrak{x}_{k-1}} \right) f(\mathfrak{x}_k) - f(x) \right| \\ &= \left| \left(\frac{\mathfrak{x}_k - x}{\mathfrak{x}_k - \mathfrak{x}_{k-1}} \right) (f(\mathfrak{x}_{k-1}) - f(x)) + \left(\frac{x - \mathfrak{x}_{k-1}}{\mathfrak{x}_k - \mathfrak{x}_{k-1}} \right) (f(\mathfrak{x}_k) - f(x)) \right| \quad (3.26) \\ &\leq \left(\frac{\mathfrak{x}_k - x}{\mathfrak{x}_k - \mathfrak{x}_{k-1}} \right) |f(\mathfrak{x}_{k-1}) - f(x)| + \left(\frac{x - \mathfrak{x}_{k-1}}{\mathfrak{x}_k - \mathfrak{x}_{k-1}} \right) |f(\mathfrak{x}_k) - f(x)|. \end{aligned}$$

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

$$\begin{aligned} |\mathfrak{l}(x) - f(x)| &\leq w_f(|\mathfrak{x}_k - \mathfrak{x}_{k-1}|) \left(\frac{\mathfrak{x}_k - x}{\mathfrak{x}_k - \mathfrak{x}_{k-1}} + \frac{x - \mathfrak{x}_{k-1}}{\mathfrak{x}_k - \mathfrak{x}_{k-1}} \right) \\ &= w_f(|\mathfrak{x}_k - \mathfrak{x}_{k-1}|) \leq w_f(\max_{j \in \{1, 2, \dots, K\}} |\mathfrak{x}_j - \mathfrak{x}_{j-1}|). \end{aligned} \quad (3.27)$$

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

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

$$|f(x) - f(y)| \leq L|x - y|. \quad (3.28)$$

Then

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

$$|(\mathcal{L}_{\mathfrak{x}_0, \mathfrak{x}_1, \dots, \mathfrak{x}_K}^{f(\mathfrak{x}_0), f(\mathfrak{x}_1), \dots, f(\mathfrak{x}_K)})(x) - (\mathcal{L}_{\mathfrak{x}_0, \mathfrak{x}_1, \dots, \mathfrak{x}_K}^{f(\mathfrak{x}_0), f(\mathfrak{x}_1), \dots, f(\mathfrak{x}_K)})(y)| \leq L|x - y| \quad (3.29)$$

and

(ii) it holds that

$$\sup_{x \in [\mathfrak{x}_0, \mathfrak{x}_K]} |(\mathcal{L}_{\mathfrak{x}_0, \mathfrak{x}_1, \dots, \mathfrak{x}_K}^{f(\mathfrak{x}_0), f(\mathfrak{x}_1), \dots, f(\mathfrak{x}_K)})(x) - f(x)| \leq L \left(\max_{k \in \{1, 2, \dots, K\}} |\mathfrak{x}_k - \mathfrak{x}_{k-1}| \right) \quad (3.30)$$

(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

$$0 \leq \frac{|f(\mathfrak{x}_K) - f(\mathfrak{x}_0)|}{(\mathfrak{x}_K - \mathfrak{x}_0)} \leq \frac{L|\mathfrak{x}_K - \mathfrak{x}_0|}{(\mathfrak{x}_K - \mathfrak{x}_0)} = L. \quad (3.31)$$

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

$$\begin{aligned} & |(\mathcal{L}_{\mathfrak{x}_0, \mathfrak{x}_1, \dots, \mathfrak{x}_K}^{f(\mathfrak{x}_0), f(\mathfrak{x}_1), \dots, f(\mathfrak{x}_K)})(x) - (\mathcal{L}_{\mathfrak{x}_0, \mathfrak{x}_1, \dots, \mathfrak{x}_K}^{f(\mathfrak{x}_0), f(\mathfrak{x}_1), \dots, f(\mathfrak{x}_K)})(y)| \\ & \leq \left(\max_{k \in \{1, 2, \dots, K\}} \left(\frac{L|\mathfrak{x}_k - \mathfrak{x}_{k-1}|}{|\mathfrak{x}_k - \mathfrak{x}_{k-1}|} \right) \right) |x - y| = L|x - y|. \end{aligned} \quad (3.32)$$

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

$$\begin{aligned} \sup_{x \in [\mathfrak{x}_0, \mathfrak{x}_K]} |(\mathcal{L}_{\mathfrak{x}_0, \mathfrak{x}_1, \dots, \mathfrak{x}_K}^{f(\mathfrak{x}_0), f(\mathfrak{x}_1), \dots, f(\mathfrak{x}_K)})(x) - f(x)| & \leq w_f \left(\max_{k \in \{1, 2, \dots, K\}} |\mathfrak{x}_k - \mathfrak{x}_{k-1}| \right) \\ & \leq L \left(\max_{k \in \{1, 2, \dots, K\}} |\mathfrak{x}_k - \mathfrak{x}_{k-1}| \right). \end{aligned} \quad (3.33)$$

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

3.2 Linear interpolation with ANNs

3.2.1 Activation functions as ANNs

Definition 3.2.1 (Activation functions as ANNs). *Let $n \in \mathbb{N}$. Then we denote by*

$$\mathbf{i}_n \in ((\mathbb{R}^{n \times n} \times \mathbb{R}^n) \times (\mathbb{R}^{n \times n} \times \mathbb{R}^n)) \subseteq \mathbf{N} \quad (3.34)$$

the fully-connected feedforward ANN given by

$$\mathbf{i}_n = ((\mathbf{I}_n, 0), (\mathbf{I}_n, 0)) \quad (3.35)$$

(cf. Definitions 1.3.1 and 1.5.5).

Lemma 3.2.2 (Realization functions of activation ANNs). *Let $n \in \mathbb{N}$. Then*

- (i) *it holds that $\mathcal{D}(\mathbf{i}_n) = (n, n, n) \in \mathbb{N}^3$ and*
- (ii) *it holds for all $a \in C(\mathbb{R}, \mathbb{R})$ that*

$$\mathcal{R}_a^{\mathbf{N}}(\mathbf{i}_n) = \mathfrak{M}_{a,n} \quad (3.36)$$

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

$$\mathcal{D}(\mathbf{i}_n) = (n, n, n) \in \mathbb{N}^3 \quad (3.37)$$

(cf. Definitions 1.3.1 and 3.2.1). This establishes item (i). Note that (1.97) and the fact that

$$\mathbf{i}_n = ((\mathbf{I}_n, 0), (\mathbf{I}_n, 0)) \in ((\mathbb{R}^{n \times n} \times \mathbb{R}^n) \times (\mathbb{R}^{n \times n} \times \mathbb{R}^n)) \quad (3.38)$$

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

$$(\mathcal{R}_a^{\mathbf{N}}(\mathbf{i}_n))(x) = \mathbf{I}_n(\mathfrak{M}_{a,n}(\mathbf{I}_n x + 0)) + 0 = \mathfrak{M}_{a,n}(x). \quad (3.39)$$

This proves item (ii). The proof of Lemma 3.2.2 is thus complete. \square

Lemma 3.2.3 (Compositions of activation ANNs with general ANNs). *Let $\Phi \in \mathbf{N}$ (cf. Definition 1.3.1). Then*

(i) *it holds that*

$$\begin{aligned} & \mathcal{D}(\mathbf{i}_{\mathcal{O}(\Phi)} \bullet \Phi) \\ &= (\mathbb{D}_0(\Phi), \mathbb{D}_1(\Phi), \mathbb{D}_2(\Phi), \dots, \mathbb{D}_{\mathcal{L}(\Phi)-1}(\Phi), \mathbb{D}_{\mathcal{L}(\Phi)}(\Phi), \mathbb{D}_{\mathcal{L}(\Phi)}(\Phi)) \in \mathbb{N}^{\mathcal{L}(\Phi)+2}, \end{aligned} \quad (3.40)$$

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

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

(iv) *it holds that*

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

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

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

(cf. Definitions 1.2.1, 1.3.4, 2.1.1, and 3.2.1).

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

$$\mathcal{R}_a^{\mathbf{N}}(\mathbf{i}_n) = \mathfrak{M}_{a,n} \quad (3.42)$$

(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. \square

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

$$\mathbf{H} = h \circledast (\mathbf{i}_1 \bullet \mathbf{A}_{\alpha, \beta}) \quad (3.43)$$

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

(cf. Definitions 1.2.4 and 1.3.4).

Proof of Lemma 3.2.4. Note that Lemma 2.3.2 implies that

$$\mathbf{A}_{\alpha, \beta} = (\alpha, \beta), \quad \mathcal{D}(\mathbf{A}_{\alpha, \beta}) = (1, 1) \in \mathbb{N}^2, \quad \mathcal{R}_{\mathbf{r}}^{\mathbf{N}}(\mathbf{A}_{\alpha, \beta}) \in C(\mathbb{R}, \mathbb{R}), \quad (3.44)$$

and $\forall x \in \mathbb{R}: (\mathcal{R}_{\mathbf{r}}^{\mathbf{N}}(\mathbf{A}_{\alpha, \beta}))(x) = \alpha x + \beta$ (cf. Definitions 1.2.4 and 1.3.4). Proposition 2.1.2, Lemma 3.2.2, Lemma 3.2.3, (1.26), (1.97), and (2.2) therefore demonstrate that

$$\mathbf{i}_1 \bullet \mathbf{A}_{\alpha, \beta} = ((\alpha, \beta), (1, 0)), \quad \mathcal{D}(\mathbf{i}_1 \bullet \mathbf{A}_{\alpha, \beta}) = (1, 1, 1) \in \mathbb{N}^3, \quad \mathcal{R}_{\mathbf{r}}^{\mathbf{N}}(\mathbf{i}_1 \bullet \mathbf{A}_{\alpha, \beta}) \in C(\mathbb{R}, \mathbb{R}), \quad (3.45)$$

$$\text{and } \forall x \in \mathbb{R}: (\mathcal{R}_{\mathbf{r}}^{\mathbf{N}}(\mathbf{i}_1 \bullet \mathbf{A}_{\alpha, \beta}))(x) = \mathbf{r}(\mathcal{R}_{\mathbf{r}}^{\mathbf{N}}(\mathbf{A}_{\alpha, \beta})(x)) = \max\{\alpha x + \beta, 0\}. \quad (3.46)$$

This, Lemma 2.3.5, and (2.130) show that

$$\mathbf{H} = h \circledast (\mathbf{i}_1 \bullet \mathbf{A}_{\alpha, \beta}) = ((\alpha, \beta), (h, 0)), \quad \mathcal{D}(\mathbf{H}) = (1, 1, 1), \quad \mathcal{R}_{\mathbf{r}}^{\mathbf{N}}(\mathbf{H}) \in C(\mathbb{R}, \mathbb{R}), \quad (3.47)$$

$$\text{and } (\mathcal{R}_{\mathbf{r}}^{\mathbf{N}}(\mathbf{H}))(x) = h((\mathcal{R}_{\mathbf{r}}^{\mathbf{N}}(\mathbf{i}_1 \bullet \mathbf{A}_{\alpha, \beta}))(x)) = h \max\{\alpha x + \beta, 0\}. \quad (3.48)$$

This proves items (i), (ii), (iii), and (iv). The proof of Lemma 3.2.4 is thus complete. \square

3.2.3 ReLU ANN representations for linear interpolations

Proposition 3.2.5 (ReLU ANN representations for linear interpolations). Let $K \in \mathbb{N}$, $f_0, f_1, \dots, f_K, \mathfrak{x}_0, \mathfrak{x}_1, \dots, \mathfrak{x}_K \in \mathbb{R}$ satisfy $\mathfrak{x}_0 < \mathfrak{x}_1 < \dots < \mathfrak{x}_K$ and let $\mathbf{F} \in \mathbf{N}$ satisfy

$$\mathbf{F} = \mathbf{A}_{1, f_0} \bullet \left(\bigoplus_{k=0}^K \left(\left(\frac{(f_{\min\{k+1, K\}} - f_k)}{(\mathfrak{x}_{\min\{k+1, K\}} - \mathfrak{x}_{\min\{k, K-1\}})} - \frac{(f_k - f_{\max\{k-1, 0\}})}{(\mathfrak{x}_{\max\{k, 1\}} - \mathfrak{x}_{\max\{k-1, 0\}})} \right) \circledast (\mathbf{i}_1 \bullet \mathbf{A}_{1, -\mathfrak{x}_k}) \right) \right) \quad (3.49)$$

(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 $\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{F}) = \mathcal{L}_{\mathfrak{x}_0, \mathfrak{x}_1, \dots, \mathfrak{x}_K}^{f_0, f_1, \dots, f_K}$, and
- (iii) it holds that $\mathcal{P}(\mathbf{F}) = 3K + 4$
 (cf. Definitions 1.2.4, 1.3.4, and 3.1.5).

Proof of Proposition 3.2.5. Throughout this proof, let $c_0, c_1, \dots, c_K \in \mathbb{R}$ satisfy for all $k \in \{0, 1, \dots, K\}$ that

$$c_k = \frac{(f_{\min\{k+1, K\}} - f_k)}{(\mathfrak{x}_{\min\{k+1, K\}} - \mathfrak{x}_{\min\{k, K-1\}})} - \frac{(f_k - f_{\max\{k-1, 0\}})}{(\mathfrak{x}_{\max\{k, 1\}} - \mathfrak{x}_{\max\{k-1, 0\}})} \quad (3.50)$$

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

$$\Phi_k = c_k \circledast (\mathfrak{i}_1 \bullet \mathbf{A}_{1, -\mathfrak{x}_k}). \quad (3.51)$$

Observe that Lemma 3.2.4 ensures that for all $k \in \{0, 1, \dots, K\}$ it holds that

$$\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\Phi_k) \in C(\mathbb{R}, \mathbb{R}), \quad \mathcal{D}(\Phi_k) = (1, 1, 1) \in \mathbb{N}^3, \quad (3.52)$$

$$\text{and } \forall x \in \mathbb{R}: (\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\Phi_k))(x) = c_k \max\{x - \mathfrak{x}_k, 0\} \quad (3.53)$$

(cf. Definitions 1.2.4 and 1.3.4). This, Lemma 2.3.3, Lemma 2.4.11, and (3.49) establish that

$$\mathcal{D}(\mathbf{F}) = (1, K+1, 1) \in \mathbb{N}^3 \quad \text{and} \quad \mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{F}) \in C(\mathbb{R}, \mathbb{R}). \quad (3.54)$$

This proves item (i). Note that item (i) and (1.84) imply that

$$\mathcal{P}(\mathbf{F}) = 2(K+1) + (K+2) = 3K+4. \quad (3.55)$$

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

$$(\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{F}))(x) = f_0 + \sum_{k=0}^K (\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\Phi_k))(x) = f_0 + \sum_{k=0}^K c_k \max\{x - \mathfrak{x}_k, 0\}. \quad (3.56)$$

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

$$(\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{F}))(x) = f_0 + 0 = f_0. \quad (3.57)$$

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

$$\sum_{n=0}^{k-1} c_n = \frac{f_k - f_{k-1}}{\mathfrak{x}_k - \mathfrak{x}_{k-1}}. \quad (3.58)$$

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

$$\sum_{n=0}^0 c_n = c_0 = \frac{f_1 - f_0}{\mathfrak{x}_1 - \mathfrak{x}_0}. \quad (3.59)$$

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

$$\sum_{n=0}^{k-1} c_n = c_{k-1} + \sum_{n=0}^{k-2} c_n = \frac{f_k - f_{k-1}}{\mathfrak{x}_k - \mathfrak{x}_{k-1}} - \frac{f_{k-1} - f_{k-2}}{\mathfrak{x}_{k-1} - \mathfrak{x}_{k-2}} + \frac{f_{k-1} - f_{k-2}}{\mathfrak{x}_{k-1} - \mathfrak{x}_{k-2}} = \frac{f_k - f_{k-1}}{\mathfrak{x}_k - \mathfrak{x}_{k-1}}. \quad (3.60)$$

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

$$\begin{aligned} (\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{F}))(x) - (\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{F}))(\mathfrak{x}_{k-1}) &= \sum_{n=0}^K c_n (\max\{x - \mathfrak{x}_n, 0\} - \max\{\mathfrak{x}_{k-1} - \mathfrak{x}_n, 0\}) \\ &= \sum_{n=0}^{k-1} c_n [(x - \mathfrak{x}_n) - (\mathfrak{x}_{k-1} - \mathfrak{x}_n)] = \sum_{n=0}^{k-1} c_n (x - \mathfrak{x}_{k-1}) \quad (3.61) \\ &= \left(\frac{f_k - f_{k-1}}{\mathfrak{x}_k - \mathfrak{x}_{k-1}} \right) (x - \mathfrak{x}_{k-1}). \end{aligned}$$

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

$$(\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{F}))(x) = f_{k-1} + \left(\frac{f_k - f_{k-1}}{\mathfrak{x}_k - \mathfrak{x}_{k-1}} \right) (x - \mathfrak{x}_{k-1}). \quad (3.62)$$

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

$$(\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{F}))(x) = (\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{F}))(\mathfrak{x}_0) + (\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{F}))(x) - (\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{F}))(\mathfrak{x}_0) = f_0 + \left(\frac{f_1 - f_0}{\mathfrak{x}_1 - \mathfrak{x}_0} \right) (x - \mathfrak{x}_0). \quad (3.63)$$

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 $\forall y \in [\mathfrak{x}_{k-2}, \mathfrak{x}_{k-1}]$: $(\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{F}))(y) = f_{k-2} + \left(\frac{f_{k-1} - f_{k-2}}{\mathfrak{x}_{k-1} - \mathfrak{x}_{k-2}} \right) (y - \mathfrak{x}_{k-2})$ it holds that

$$\begin{aligned} (\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{F}))(x) &= (\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{F}))(\mathfrak{x}_{k-1}) + (\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{F}))(x) - (\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{F}))(\mathfrak{x}_{k-1}) \\ &= f_{k-2} + \left(\frac{f_{k-1} - f_{k-2}}{\mathfrak{x}_{k-1} - \mathfrak{x}_{k-2}} \right) (\mathfrak{x}_{k-1} - \mathfrak{x}_{k-2}) + \left(\frac{f_k - f_{k-1}}{\mathfrak{x}_k - \mathfrak{x}_{k-1}} \right) (x - \mathfrak{x}_{k-1}) \quad (3.64) \\ &= f_{k-1} + \left(\frac{f_k - f_{k-1}}{\mathfrak{x}_k - \mathfrak{x}_{k-1}} \right) (x - \mathfrak{x}_{k-1}). \end{aligned}$$

Induction thus implies (3.62). Moreover, note that (3.50) and (3.58) demonstrate that

$$\sum_{n=0}^K c_n = c_K + \sum_{n=0}^{K-1} c_n = -\frac{f_K - f_{K-1}}{\mathfrak{x}_K - \mathfrak{x}_{K-1}} + \frac{f_K - f_{K-1}}{\mathfrak{x}_K - \mathfrak{x}_{K-1}} = 0. \quad (3.65)$$

The fact that for all $k \in \{0, 1, \dots, K\}$ 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

$$\begin{aligned} (\mathcal{R}_{\mathfrak{r}}^N(\mathbf{F}))(x) - (\mathcal{R}_{\mathfrak{r}}^N(\mathbf{F}))(\mathfrak{x}_K) &= \left[\sum_{n=0}^K c_n (\max\{x - \mathfrak{x}_n, 0\} - \max\{\mathfrak{x}_K - \mathfrak{x}_n, 0\}) \right] \\ &= \sum_{n=0}^K c_n [(x - \mathfrak{x}_n) - (\mathfrak{x}_K - \mathfrak{x}_n)] = \sum_{n=0}^K c_n (x - \mathfrak{x}_K) = 0. \end{aligned} \quad (3.66)$$

This and (3.62) ensure that for all $x \in [\mathfrak{x}_K, \infty)$ it holds that

$$(\mathcal{R}_{\mathfrak{r}}^N(\mathbf{F}))(x) = (\mathcal{R}_{\mathfrak{r}}^N(\mathbf{F}))(\mathfrak{x}_K) = f_{K-1} + \left(\frac{f_K - f_{K-1}}{\mathfrak{x}_K - \mathfrak{x}_{K-1}} \right) (\mathfrak{x}_K - \mathfrak{x}_{K-1}) = f_K. \quad (3.67)$$

Combining this, (3.57), (3.62), and (3.11) establishes item (ii). The proof of Proposition 3.2.5 is thus complete. \square

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

$$\sup_{x \in [-2\pi, 2\pi]} |\cos(x) - (\mathcal{R}_{\mathfrak{r}}^N(\Phi))(x)| \leq \frac{1}{2} \quad (3.68)$$

(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 $\forall x, y, u, v \in \mathbb{R}: (\mathcal{R}_{\mathfrak{r}}^N(\Phi))(x, y, u, v) = \max\{x, y, u, v\}$ (cf. Definitions 1.2.4, 1.3.1, and 1.3.4).

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

3.3 ANN approximations results for one-dimensional functions

3.3.1 Constructive ANN approximation results

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

$$|f(x) - f(y)| \leq L|x - y|, \quad (3.69)$$

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

$$\mathbf{F} = \mathbf{A}_{1,f(\mathfrak{x}_0)} \bullet \left(\bigoplus_{k=0}^K \left(\left(\frac{K(f(\mathfrak{x}_{\min\{k+1,K\}}) - 2f(\mathfrak{x}_k) + f(\mathfrak{x}_{\max\{k-1,0\}}))}{(b-a)} \right) \otimes (\mathbf{i}_1 \bullet \mathbf{A}_{1,-\mathfrak{x}_k}) \right) \right) \quad (3.70)$$

(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 $\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{F}) = \mathcal{L}_{\mathfrak{x}_0, \mathfrak{x}_1, \dots, \mathfrak{x}_K}^{f(\mathfrak{x}_0), f(\mathfrak{x}_1), \dots, f(\mathfrak{x}_K)}$,
- (iii) it holds for all $x, y \in \mathbb{R}$ that $|\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{F})(x) - \mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{F})(y)| \leq L|x - y|$,
- (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 $k \in \{0, 1, \dots, K\}$ it holds that

$$\mathfrak{x}_{\min\{k+1,K\}} - \mathfrak{x}_{\min\{k,K-1\}} = \mathfrak{x}_{\max\{k,1\}} - \mathfrak{x}_{\max\{k-1,0\}} = (b-a)K^{-1} \quad (3.71)$$

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

$$\begin{aligned} & \frac{(f(\mathfrak{x}_{\min\{k+1,K\}}) - f(\mathfrak{x}_k))}{(\mathfrak{x}_{\min\{k+1,K\}} - \mathfrak{x}_{\min\{k,K-1\}})} - \frac{(f(\mathfrak{x}_k) - f(\mathfrak{x}_{\max\{k-1,0\}}))}{(\mathfrak{x}_{\max\{k,1\}} - \mathfrak{x}_{\max\{k-1,0\}})} \\ &= \frac{K(f(\mathfrak{x}_{\min\{k+1,K\}}) - 2f(\mathfrak{x}_k) + f(\mathfrak{x}_{\max\{k-1,0\}}))}{(b-a)}. \end{aligned} \quad (3.72)$$

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

$$|f(x) - f(y)| \leq L|x - y| \quad (3.73)$$

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

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

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

$$\mathfrak{x}_k - \mathfrak{x}_{k-1} = \frac{(b-a)}{K} \quad (3.75)$$

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

$$|(\mathcal{R}_{\mathbf{r}}^{\mathbf{N}}(\mathbf{F}))(x) - f(x)| \leq L \left(\max_{k \in \{1, 2, \dots, K\}} |\mathfrak{x}_k - \mathfrak{x}_{k-1}| \right) = \frac{L(b-a)}{K}. \quad (3.76)$$

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

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

$$|f(x) - f(y)| \leq L|x - y|, \quad (3.77)$$

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

$$\mathbf{F} = \mathbf{A}_{1,f(\xi)} \bullet (0 \circledast (\mathbf{i}_1 \bullet \mathbf{A}_{1,-\xi})) \quad (3.78)$$

(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 $\mathcal{R}_{\mathbf{r}}^{\mathbf{N}}(\mathbf{F}) \in C(\mathbb{R}, \mathbb{R})$,
- (iii) it holds for all $x \in \mathbb{R}$ that $(\mathcal{R}_{\mathbf{r}}^{\mathbf{N}}(\mathbf{F}))(x) = f(\xi)$,
- (iv) it holds that $\sup_{x \in [a,b]} |(\mathcal{R}_{\mathbf{r}}^{\mathbf{N}}(\mathbf{F}))(x) - f(x)| \leq L \max\{\xi - a, b - \xi\}$, and
- (v) it holds that $\mathcal{P}(\mathbf{F}) = 4$

(cf. Definitions 1.2.4 and 1.3.4).

Proof of Lemma 3.3.2. Note that items (i) and (ii) in Lemma 2.3.3, and items (ii) and (iii) in Lemma 3.2.4 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

$$\begin{aligned} (\mathcal{R}_{\mathbf{r}}^{\mathbf{N}}(\mathbf{F}))(x) &= (\mathcal{R}_{\mathbf{r}}^{\mathbf{N}}(0 \circledast (\mathbf{i}_1 \bullet \mathbf{A}_{1,-\xi}))) (x) + f(\xi) \\ &= 0((\mathcal{R}_{\mathbf{r}}^{\mathbf{N}}(\mathbf{i}_1 \bullet \mathbf{A}_{1,-\xi}))(x)) + f(\xi) = f(\xi) \end{aligned} \quad (3.79)$$

(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

$$|f(x) - f(y)| \leq L|x - y| \quad (3.80)$$

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

$$|(\mathcal{R}_{\mathbf{r}}^{\mathbf{N}}(\mathbf{F}))(x) - f(x)| = |f(\xi) - f(x)| \leq L|x - \xi| \leq L \max\{\xi - a, b - \xi\}. \quad (3.81)$$

This proves item (iv). Observe that (1.84) and item (i) ensure that

$$\mathcal{P}(\mathbf{F}) = 1(1+1) + 1(1+1) = 4. \quad (3.82)$$

This establishes item (v). The proof of Lemma 3.3.2 is thus complete. \square

Corollary 3.3.3 (Explicit ANN approximations with prescribed error tolerances). *Let $\varepsilon \in (0, \infty)$, $L, a \in \mathbb{R}$, $b \in (a, \infty)$, $K \in \mathbb{N}_0 \cap [\frac{L(b-a)}{\varepsilon}, \frac{L(b-a)}{\varepsilon} + 1)$, $\mathbf{r}_0, \mathbf{r}_1, \dots, \mathbf{r}_K \in \mathbb{R}$ satisfy for all $k \in \{0, 1, \dots, K\}$ that $\mathbf{r}_k = a + \frac{k(b-a)}{\max\{K, 1\}}$, let $f: [a, b] \rightarrow \mathbb{R}$ satisfy for all $x, y \in [a, b]$ that*

$$|f(x) - f(y)| \leq L|x - y|, \quad (3.83)$$

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

$$\mathbf{F} = \mathbf{A}_{1,f(\mathbf{r}_0)} \bullet \left(\bigoplus_{k=0}^K \left(\left(\frac{K(f(\mathbf{r}_{\min\{k+1, K\}}) - 2f(\mathbf{r}_k) + f(\mathbf{r}_{\max\{k-1, 0\}}))}{(b-a)} \right) \otimes (\mathbf{i}_1 \bullet \mathbf{A}_{1,-\mathbf{r}_k}) \right) \right) \quad (3.84)$$

(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 $\mathcal{R}_{\mathbf{r}}^{\mathbf{N}}(\mathbf{F}) \in C(\mathbb{R}, \mathbb{R})$,
 - (iii) it holds for all $x, y \in \mathbb{R}$ that $|(\mathcal{R}_{\mathbf{r}}^{\mathbf{N}}(\mathbf{F}))(x) - (\mathcal{R}_{\mathbf{r}}^{\mathbf{N}}(\mathbf{F}))(y)| \leq L|x - y|$,
 - (iv) it holds that $\sup_{x \in [a, b]} |(\mathcal{R}_{\mathbf{r}}^{\mathbf{N}}(\mathbf{F}))(x) - f(x)| \leq \frac{L(b-a)}{\max\{K, 1\}} \leq \varepsilon$, and
 - (v) it holds that $\mathcal{P}(\mathbf{F}) = 3K + 4 \leq 3L(b-a)\varepsilon^{-1} + 7$
- (cf. Definitions 1.2.4, 1.3.1, and 1.3.4).

Proof of Corollary 3.3.3. Note that the assumption that $K \in \mathbb{N}_0 \cap [\frac{L(b-a)}{\varepsilon}, \frac{L(b-a)}{\varepsilon} + 1)$ implies that

$$\frac{L(b-a)}{\max\{K, 1\}} \leq \varepsilon. \quad (3.85)$$

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

$$K \leq 1 + \frac{L(b-a)}{\varepsilon}, \quad (3.86)$$

demonstrate that

$$\mathcal{P}(\mathbf{F}) = 3K + 4 \leq \frac{3L(b-a)}{\varepsilon} + 7. \quad (3.87)$$

This establishes item (v). The proof of Corollary 3.3.3 is thus complete. \square

3.3.2 Convergence rates for the approximation error

Definition 3.3.4 (Quasi vector norms). Let $d \in \mathbb{N}$, $p \in (0, \infty]$, $\theta = (\theta_1, \dots, \theta_d) \in \mathbb{R}^d$. Then we denote by $\|\theta\|_p \in \mathbb{R}$ the real number given by

$$\|\theta\|_p = \begin{cases} \left[\sum_{i=1}^d |\theta_i|^p \right]^{1/p} & : p \in (0, \infty) \\ \max_{i \in \{1, 2, \dots, d\}} |\theta_i| & : p = \infty. \end{cases} \quad (3.88)$$

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 $f: [a, b] \rightarrow \mathbb{R}$ satisfy for all $x, y \in [a, b]$ that

$$|f(x) - f(y)| \leq L|x - y|. \quad (3.89)$$

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 $K \in \mathbb{N}_0 \cap [\frac{L(b-a)}{\varepsilon}, \frac{L(b-a)}{\varepsilon} + 1)$, $\mathfrak{x}_0, \mathfrak{x}_1, \dots, \mathfrak{x}_K \in [a, b]$, $c_0, c_1, \dots, c_K \in \mathbb{R}$ satisfy for all $k \in \{0, 1, \dots, K\}$ that

$$\mathfrak{x}_k = a + \frac{k(b-a)}{\max\{K, 1\}} \quad \text{and} \quad c_k = \frac{K(f(\mathfrak{x}_{\min\{k+1, K\}}) - 2f(\mathfrak{x}_k) + f(\mathfrak{x}_{\max\{k-1, 0\}}))}{(b-a)}, \quad (3.90)$$

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

$$\mathbf{F} = \mathbf{A}_{1,f(\mathfrak{x}_0)} \bullet \left(\bigoplus_{k=0}^K (c_k \circledast (\mathbf{i}_1 \bullet \mathbf{A}_{1,-\mathfrak{x}_k})) \right) \quad (3.91)$$

(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 $\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{F}) \in C(\mathbb{R}, \mathbb{R})$,
- (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 $\sup_{x \in [a,b]} |(\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{F}))(x) - f(x)| \leq \varepsilon$, 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

$$K \leq 1 + \frac{L(b-a)}{\varepsilon} \quad (3.92)$$

prove items (ii) and (iii). Note that item (iii) and items (I) and (V) ensure that

$$\mathcal{P}(\mathbf{F}) = 3K + 4 = 3(K+1) + 1 = 3(\mathbb{D}_1(\mathbf{F})) + 1 \leq \frac{3L(b-a)}{\varepsilon} + 7. \quad (3.93)$$

This establishes item (vi). Observe that Lemma 3.2.4 implies that for all $k \in \{0, 1, \dots, K\}$ it holds that

$$c_k \circledast (\mathbf{i}_1 \bullet \mathbf{A}_{1,-\mathfrak{x}_k}) = ((1, -\mathfrak{x}_k), (c_k, 0)). \quad (3.94)$$

Combining this with (2.155), (2.146), (2.137), and (2.2) demonstrates that

$$\begin{aligned} \mathbf{F} &= \left(\left(\begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}, \begin{pmatrix} -\mathfrak{x}_0 \\ -\mathfrak{x}_1 \\ \vdots \\ -\mathfrak{x}_K \end{pmatrix} \right), ((c_0 \ c_1 \ \cdots \ c_K), f(\mathfrak{x}_0)) \right) \\ &\in (\mathbb{R}^{(K+1) \times 1} \times \mathbb{R}^{K+1}) \times (\mathbb{R}^{1 \times (K+1)} \times \mathbb{R}). \end{aligned} \quad (3.95)$$

Lemma 1.3.9 therefore shows that

$$\|\mathcal{T}(\mathbf{F})\|_{\infty} = \max\{|\mathfrak{x}_0|, |\mathfrak{x}_1|, \dots, |\mathfrak{x}_K|, |c_0|, |c_1|, \dots, |c_K|, |f(\mathfrak{x}_0)|, 1\} \quad (3.96)$$

(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

$$|f(x) - f(y)| \leq L|x - y|, \quad (3.97)$$

and the fact that for all $k \in \mathbb{N} \cap (0, K + 1)$ it holds that

$$\mathfrak{x}_k - \mathfrak{x}_{k-1} = \frac{(b - a)}{\max\{K, 1\}} \quad (3.98)$$

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

$$\begin{aligned} |c_k| &\leq \frac{K(|f(\mathfrak{x}_{\min\{k+1, K\}}) - f(\mathfrak{x}_k)| + |f(\mathfrak{x}_{\max\{k-1, 0\}}) - f(\mathfrak{x}_k)|)}{(b - a)} \\ &\leq \frac{KL(|\mathfrak{x}_{\min\{k+1, K\}} - \mathfrak{x}_k| + |\mathfrak{x}_{\max\{k-1, 0\}} - \mathfrak{x}_k|)}{(b - a)} \\ &\leq \frac{2KL(b - a)[\max\{K, 1\}]^{-1}}{(b - a)} \leq 2L. \end{aligned} \quad (3.99)$$

This and (3.96) establish item (vii). The proof of Corollary 3.3.5 is thus complete. \square

Corollary 3.3.6 (Implicit one-dimensional ANN approximations with prescribed error tolerances and asymptotic parameter bounds). *Let $L, a \in \mathbb{R}$, $b \in [a, \infty)$ and let $f: [a, b] \rightarrow \mathbb{R}$ satisfy for all $x, y \in [a, b]$ that*

$$|f(x) - f(y)| \leq L|x - y|. \quad (3.100)$$

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

$$\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{F}) \in C(\mathbb{R}, \mathbb{R}), \quad \sup_{x \in [a, b]} |(\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{F}))(x) - f(x)| \leq \varepsilon, \quad \mathcal{H}(\mathbf{F}) = 1, \quad (3.101)$$

$$\|\mathcal{T}(\mathbf{F})\|_{\infty} \leq \max\{1, |a|, |b|, 2L, |f(a)|\}, \quad \text{and} \quad \mathcal{P}(\mathbf{F}) \leq \mathfrak{C}\varepsilon^{-1} \quad (3.102)$$

(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

$$\mathfrak{C} = 3L(b - a) + 7. \quad (3.103)$$

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

$$3L(b - a)\varepsilon^{-1} + 7 \leq 3L(b - a)\varepsilon^{-1} + 7\varepsilon^{-1} = \mathfrak{C}\varepsilon^{-1}. \quad (3.104)$$

This and Corollary 3.3.5 demonstrate that for all $\varepsilon \in (0, 1]$ there exists $\mathbf{F} \in \mathbf{N}$ such that

$$\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{F}) \in C(\mathbb{R}, \mathbb{R}), \quad \sup_{x \in [a, b]} |(\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{F}))(x) - f(x)| \leq \varepsilon, \quad \mathcal{H}(\mathbf{F}) = 1, \quad (3.105)$$

$$\|\mathcal{T}(\mathbf{F})\|_{\infty} \leq \max\{1, |a|, |b|, 2L, |f(a)|\}, \quad \text{and} \quad \mathcal{P}(\mathbf{F}) \leq 3L(b - a)\varepsilon^{-1} + 7 \leq \mathfrak{C}\varepsilon^{-1} \quad (3.106)$$

(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. \square

Corollary 3.3.7 (Implicit one-dimensional ANN approximations with prescribed error tolerances and asymptotic parameter bounds). *Let $L, a \in \mathbb{R}$, $b \in [a, \infty)$ and let $f: [a, b] \rightarrow \mathbb{R}$ satisfy for all $x, y \in [a, b]$ that*

$$|f(x) - f(y)| \leq L|x - y|. \quad (3.107)$$

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

$$\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{F}) \in C(\mathbb{R}, \mathbb{R}), \quad \sup_{x \in [a, b]} |(\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{F}))(x) - f(x)| \leq \varepsilon, \quad \text{and} \quad \mathcal{P}(\mathbf{F}) \leq \mathfrak{C}\varepsilon^{-1} \quad (3.108)$$

(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. \square

Exercise 3.3.1. Let $f: [-2, 3] \rightarrow \mathbb{R}$ satisfy for all $x \in [-2, 3]$ that

$$f(x) = x^2 + 2 \sin(x). \quad (3.109)$$

Prove or disprove the following statement: There exist $c \in \mathbb{R}$ and $\mathbf{F} = (\mathbf{F}_\varepsilon)_{\varepsilon \in (0, 1]}: (0, 1] \rightarrow \mathbf{N}$ such that for all $\varepsilon \in (0, 1]$ it holds that

$$\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{F}_\varepsilon) \in C(\mathbb{R}, \mathbb{R}), \quad \sup_{x \in [-2, 3]} |(\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{F}_\varepsilon))(x) - f(x)| \leq \varepsilon, \quad \text{and} \quad \mathcal{P}(\mathbf{F}_\varepsilon) \leq c\varepsilon^{-1} \quad (3.110)$$

(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

$$\sup_{x \in [0, 10]} |\sqrt{x} - (\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\Phi))(x)| \leq \frac{1}{4} \quad (3.111)$$

(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], [238, Sections 3 and 4.2], and [243, 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 $\delta: E \times E \rightarrow [0, \infty)$ is a function from $E \times E$ to $[0, \infty)$ which satisfies that*

(i) *it holds that*

$$\{(x, y) \in E^2: \delta(x, y) = 0\} = \bigcup_{x \in E} \{(x, x)\} \quad (4.1)$$

(positive definiteness),

(ii) *it holds for all $x, y \in E$ that*

$$\delta(x, y) = \delta(y, x) \quad (4.2)$$

(symmetry), and

(iii) *it holds for all $x, y, z \in E$ that*

$$\delta(x, z) \leq \delta(x, y) + \delta(y, z) \quad (4.3)$$

(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*

$$\mathcal{E} = (E, \delta) \quad (4.4)$$

(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 $f: D \rightarrow \mathbb{R}$ satisfy for all $x \in D, y \in \mathcal{M}$ that $|f(x) - f(y)| \leq L\delta(x, y)$, and let $F: E \rightarrow \mathbb{R} \cup \{\infty\}$ satisfy for all $x \in E$ that*

$$F(x) = \sup_{y \in \mathcal{M}} [f(y) - L\delta(x, y)] \quad (4.5)$$

(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

$$|F(x) - f(x)| \leq 2L \left[\inf_{y \in \mathcal{M}} \delta(x, y) \right]. \quad (4.6)$$

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

$$f(y) + L\delta(x, y) \geq f(x) \geq f(y) - L\delta(x, y). \quad (4.7)$$

Hence, we obtain that for all $x \in D$ it holds that

$$f(x) \geq \sup_{y \in \mathcal{M}} [f(y) - L\delta(x, y)] = F(x). \quad (4.8)$$

This establishes item (ii). Moreover, note that (4.5) implies that for all $x \in \mathcal{M}$ it holds that

$$F(x) \geq f(x) - L\delta(x, x) = f(x). \quad (4.9)$$

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

$$f(y) - L\delta(x, y) \leq f(z) + L\delta(y, z) - L\delta(x, y) \leq f(z) + L\delta(x, z). \quad (4.10)$$

Hence, we obtain that for all $x \in E$, $z \in \mathcal{M}$ it holds that

$$F(x) = \sup_{y \in \mathcal{M}} [f(y) - L\delta(x, y)] \leq f(z) + L\delta(x, z) < \infty. \quad (4.11)$$

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

$$\begin{aligned} F(x) - F(y) &= \left[\sup_{v \in \mathcal{M}} (f(v) - L\delta(x, v)) \right] - \left[\sup_{w \in \mathcal{M}} (f(w) - L\delta(y, w)) \right] \\ &= \sup_{v \in \mathcal{M}} \left[f(v) - L\delta(x, v) - \sup_{w \in \mathcal{M}} (f(w) - L\delta(y, w)) \right] \\ &\leq \sup_{v \in \mathcal{M}} [f(v) - L\delta(x, v) - (f(v) - L\delta(y, v))] \\ &= \sup_{v \in \mathcal{M}} (L\delta(y, v) - L\delta(x, v)) \\ &\leq \sup_{v \in \mathcal{M}} (L\delta(y, x) + L\delta(x, v) - L\delta(x, v)) = L\delta(x, y). \end{aligned} \quad (4.12)$$

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 $\forall x \in D, y \in \mathcal{M}: |f(x) - f(y)| \leq L\delta(x, y)$ ensure that for all $x \in D$ it holds that

$$\begin{aligned} |F(x) - f(x)| &= \inf_{y \in \mathcal{M}} |F(x) - F(y) + f(y) - f(x)| \\ &\leq \inf_{y \in \mathcal{M}} (|F(x) - F(y)| + |f(y) - f(x)|) \\ &\leq \inf_{y \in \mathcal{M}} (2L\delta(x, y)) = 2L \left[\inf_{y \in \mathcal{M}} \delta(x, y) \right]. \end{aligned} \quad (4.13)$$

This establishes item (v). The proof of Proposition 4.1.3 is thus complete. \square

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 $f: E \rightarrow \mathbb{R}$ satisfy for all $x \in E$, $y \in \mathcal{M}$ that $|f(x) - f(y)| \leq L\delta(x, y)$, and let $F: E \rightarrow \mathbb{R} \cup \{\infty\}$ satisfy for all $x \in E$ that*

$$F(x) = \sup_{y \in \mathcal{M}} [f(y) - L\delta(x, y)] \quad (4.14)$$

. 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

$$|F(x) - f(x)| \leq 2L \left[\inf_{y \in \mathcal{M}} \delta(x, y) \right]. \quad (4.15)$$

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. \square

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

$$\sup_{x, y \in [0, 2\pi]} |\sin(x) \sin(y) - (\mathcal{R}_r^{\mathbf{N}}(\Phi))(x, y)| \leq \frac{1}{5}. \quad (4.16)$$

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

$$\mathbb{L}_1 = \left(\left(\begin{pmatrix} 1 \\ -1 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \end{pmatrix} \right), ((1 \ 1), (0)) \right) \in (\mathbb{R}^{2 \times 1} \times \mathbb{R}^2) \times (\mathbb{R}^{1 \times 2} \times \mathbb{R}^1) \quad (4.17)$$

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 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 $\mathcal{R}_r^{\mathbf{N}}(\mathbb{L}_d) \in C(\mathbb{R}^d, \mathbb{R})$, and

(iii) it holds for all $x \in \mathbb{R}^d$ that $(\mathcal{R}_{\tau}^N(\mathbb{L}_d))(x) = \|x\|_1$
 (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

$$\mathcal{D}(\mathbf{P}_d(\mathbb{L}_1, \mathbb{L}_1, \dots, \mathbb{L}_1)) = (d, 2d, d) \quad (4.18)$$

(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

$$\mathcal{D}(\mathbb{L}_d) = \mathcal{D}(\mathbb{S}_{1,d} \bullet \mathbf{P}_d(\mathbb{L}_1, \mathbb{L}_1, \dots, \mathbb{L}_1)) = (d, 2d, 1) \quad (4.19)$$

(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

$$(\mathcal{R}_{\tau}^N(\mathbb{L}_1))(x) = \tau(x) + \tau(-x) = \max\{x, 0\} + \max\{-x, 0\} = |x| = \|x\|_1 \quad (4.20)$$

(cf. Definitions 1.2.4, 1.3.4, and 3.3.4). Combining this and Proposition 2.2.3 shows that for all $x = (x_1, \dots, x_d) \in \mathbb{R}^d$ it holds that

$$(\mathcal{R}_{\tau}^N(\mathbf{P}_d(\mathbb{L}_1, \mathbb{L}_1, \dots, \mathbb{L}_1)))(x) = (|x_1|, |x_2|, \dots, |x_d|). \quad (4.21)$$

This and Lemma 2.4.2 demonstrate that for all $x = (x_1, \dots, x_d) \in \mathbb{R}^d$ it holds that

$$\begin{aligned} (\mathcal{R}_{\tau}^N(\mathbb{L}_d))(x) &= (\mathcal{R}_{\tau}^N(\mathbb{S}_{1,d} \bullet \mathbf{P}_d(\mathbb{L}_1, \mathbb{L}_1, \dots, \mathbb{L}_1)))(x) \\ &= (\mathcal{R}_{\tau}^N(\mathbb{S}_{1,d}))(|x_1|, |x_2|, \dots, |x_d|) = \sum_{k=1}^d |x_k| = \|x\|_1. \end{aligned} \quad (4.22)$$

This establishes items (ii) and (iii). The proof of Proposition 4.2.2 is thus complete. \square

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 $\mathcal{W}_{2,\mathbb{L}_d} = (1 \ 1 \ \cdots \ 1) \in \mathbb{R}^{1 \times (2d)}$
 (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 $\mathbb{L}_d = \mathbb{S}_{1,d} \bullet \mathbf{P}_d(\mathbb{L}_1, \mathbb{L}_1, \dots, \mathbb{L}_1)$ 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

$$\mathcal{W}_{1,\mathbb{L}_1} = \begin{pmatrix} 1 \\ -1 \end{pmatrix} \quad \text{and} \quad \mathcal{W}_{1,\mathbb{L}_d} = \begin{pmatrix} \mathcal{W}_{1,\mathbb{L}_1} & 0 & \cdots & 0 \\ 0 & \mathcal{W}_{1,\mathbb{L}_1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathcal{W}_{1,\mathbb{L}_1} \end{pmatrix} \in \mathbb{R}^{(2d) \times d} \quad (4.23)$$

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 $\mathbb{L}_d = \mathbb{S}_{1,d} \bullet \mathbf{P}_d(\mathbb{L}_1, \mathbb{L}_1, \dots, \mathbb{L}_1)$ show that

$$\begin{aligned} \mathcal{W}_{2,\mathbb{L}_d} &= \mathcal{W}_{1,\mathbb{S}_{1,d}} \mathcal{W}_{2,\mathbf{P}_d(\mathbb{L}_1, \mathbb{L}_1, \dots, \mathbb{L}_1)} \\ &= \underbrace{\begin{pmatrix} 1 & 1 & \cdots & 1 \end{pmatrix}}_{\in \mathbb{R}^{1 \times d}} \underbrace{\begin{pmatrix} \mathcal{W}_{2,\mathbb{L}_1} & 0 & \cdots & 0 \\ 0 & \mathcal{W}_{2,\mathbb{L}_1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathcal{W}_{2,\mathbb{L}_1} \end{pmatrix}}_{\in \mathbb{R}^{d \times (2d)}} \\ &= (1 \ 1 \ \cdots \ 1) \in \mathbb{R}^{1 \times (2d)}. \end{aligned} \quad (4.24)$$

This establishes item (v). The proof of Lemma 4.2.3 is thus complete. \square

4.2.2 RePU ANN representations for the identity

In Section 2.2.2 we observed that ReLU and softplus ANNs can express the real identity function (cf. Lemmas 2.2.7 and 2.2.8). In this subsection we show that the 1-norm ANN in Definition 4.2.1 can also be used to explicitly specify RePU ANN representations for the real identity function. We note that the remaining parts of this chapter are concerned with ReLU ANNs and do not use the material presented in this subsection.

Lemma 4.2.4. *Let $p \in \mathbb{N} \setminus \{1\}$, $b_1, b_2, \dots, b_p \in \mathbb{R}$ satisfy $b_1 < b_2 < \dots < b_p$. Then*

(i) *there exist unique $c_0, c_1, \dots, c_p \in \mathbb{R}$ which satisfy for all $k \in \{0, 1, \dots, p\}$ that*

$$\mathbb{1}_{\{p\}}(k) c_0 + \sum_{i=1}^p c_i (b_i)^k = \mathbb{1}_{\{p-1\}}(k) p^{-1} \quad (4.25)$$

and

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

$$c_0 + \sum_{i=1}^p c_i(x + b_i)^p = x. \quad (4.26)$$

Proof of Lemma 4.2.4. Throughout this proof let $\mathbf{B} = (\mathbf{B}_{i,j})_{i,j \in \{1,2,\dots,p+1\}} \in \mathbb{R}^{(p+1) \times (p+1)}$ satisfy for all $i, j \in \{1, 2, \dots, p\}$ that $\mathbf{B}_{1,i+1} = 1$, $\mathbf{B}_{i,1} = 0$, $\mathbf{B}_{p+1,1} = 1$, and $\mathbf{B}_{i+1,j+1} = (b_j)^i$ and let $\mathbf{D} = (\mathbf{D}_1, \mathbf{D}_2, \dots, \mathbf{D}_{p+1}) \in \mathbb{R}^{p+1}$ satisfy for all $k \in \{1, 2, \dots, p+1\}$ that $\mathbf{D}_k = \mathbb{1}_{\{p\}}(k) p^{-1}$. Note that, for instance, Horn and Johnson [212, Eq. (0.9.11.2)] and the assumption that $b_1 < b_2 < \dots < b_p$ show that

$$\det(\mathbf{B}) = (-1)^{p+1} \det((\mathbf{B}_{i,j+1})_{i,j \in \{1,2,\dots,p\}}) = (-1)^p \left[\prod_{\substack{i,j \in \{1,2,\dots,p\} \\ i < j}} (b_j - b_i) \right] \neq 0. \quad (4.27)$$

This establishes that there exists a unique $\mathbf{C} = (c_0, c_1, \dots, c_p) \in \mathbb{R}^{p+1}$ such that $\mathbf{BC} = \mathbf{D}$. This proves item (i). Furthermore, observe that item (i) and the binomial theorem ensure that for all $x \in \mathbb{R}$ it holds that

$$\begin{aligned} c_0 + \sum_{i=1}^p c_i(x + b_i)^p &= c_0 + \sum_{i=1}^p c_i \left[\sum_{j=0}^p \binom{p}{j} x^{p-j} (b_i)^j \right] \\ &= c_0 + \sum_{j=0}^p \binom{p}{j} \left[\sum_{i=1}^p c_i (b_i)^j \right] x^{p-j} \\ &= c_0 + \sum_{j=0}^p \binom{p}{j} [\mathbb{1}_{\{p-1\}}(j) p^{-1} - \mathbb{1}_{\{p\}}(j) c_0] x^{p-j} \\ &= c_0 + \binom{p}{p-1} p^{-1} x^{p-(p-1)} - \binom{p}{p} c_0 = x. \end{aligned} \quad (4.28)$$

This establishes item (ii). The proof of Lemma 4.2.4 is thus complete. \square

Lemma 4.2.5 (Shallow RePU ANN representation for the one-dimensional identity function). Let $p \in \mathbb{N} \setminus \{1\}$, $b_1, b_2, \dots, b_p, c_0, c_1, \dots, c_p \in \mathbb{R}$ satisfy for all $k \in \{0, 1, \dots, p\}$ that $b_1 < b_2 < \dots < b_p$ and $\mathbb{1}_{\{p\}}(k) c_0 + \sum_{i=1}^p c_i (b_i)^k = \mathbb{1}_{\{p-1\}}(k) p^{-1}$, let a be the RePU activation function with power p , let $I \in \mathbf{N}$ satisfy

$$I = \begin{cases} \mathfrak{I}_1 & : p \in \{1, 3, 5, \dots\} \\ \mathbb{L}_1 & : p \in \{2, 4, 6, \dots\}, \end{cases} \quad (4.29)$$

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

$$\Psi = \mathbf{A}_{1,c_0} \bullet \left(\bigoplus_{i=1}^p \left(c_i \circledast (I \bullet \mathbf{A}_{1,b_i}) \right) \right) \quad (4.30)$$

(cf. Lemma 4.2.4 and Definitions 1.2.41, 1.3.1, 2.1.1, 2.2.6, 2.3.1, 2.3.4, 2.4.10, and 4.2.1). Then

- (i) it holds for all $x \in \mathbb{R}$ that $(\mathcal{R}_a^N(I))(x) = x^p$,
 - (ii) it holds that $\mathcal{D}(\Psi) = (1, 2p, 1) \in \mathbb{N}^3$,
 - (iii) it holds that $\mathcal{R}_a^N(\Psi) \in C(\mathbb{R}, \mathbb{R})$, and
 - (iv) it holds for all $x \in \mathbb{R}$ that $(\mathcal{R}_a^N(\Psi))(x) = x$
- (cf. Definition 1.3.4).

Proof of Lemma 4.2.5. First, note that the fact that

$$I = \left(\left(\begin{pmatrix} 1 \\ -1 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \end{pmatrix} \right), \left((1 \quad (-1)^p), 0 \right) \right) \in ((\mathbb{R}^{2 \times 1} \times \mathbb{R}^2) \times (\mathbb{R}^{1 \times 2} \times \mathbb{R}^1)) \quad (4.31)$$

implies that for all $x \in \mathbb{R}$ it holds that

$$\begin{aligned} (\mathcal{R}_a^N(I))(x) &= a(x) + (-1)^p a(-x) = (\max\{x, 0\})^p + (-1)^p (\max\{-x, 0\})^p \\ &= (\max\{x, 0\})^p + (\min\{x, 0\})^p = x^p \end{aligned} \quad (4.32)$$

(cf. Definition 1.3.4). Furthermore, observe that item (iv) in Lemma 2.3.3, item (i) in Lemma 2.3.5, the fact that $\mathcal{D}(I) = (1, 2, 1)$, and the fact that for all $i \in \{1, 2, \dots, p\}$ it holds that $\mathcal{D}(\mathbf{A}_{1,b_i}) = (1, 1)$ demonstrate that for all $i \in \{1, 2, \dots, p\}$ it holds that

$$\mathcal{D}(c_i \circledast (I \bullet \mathbf{A}_{1,b_i})) = \mathcal{D}(I \bullet \mathbf{A}_{1,b_i}) = (1, 2, 1). \quad (4.33)$$

Combining this with item (ii) in Lemma 2.4.11 proves that

$$\mathcal{D}\left(\bigoplus_{i=1}^p (c_i \circledast (I \bullet \mathbf{A}_{1,b_i}))\right) = (1, 2p, 1). \quad (4.34)$$

Item (i) in Lemma 2.3.3 hence shows that

$$\mathcal{D}(\Psi) = \mathcal{D}\left(\mathbf{A}_{1,c_0} \bullet \left(\bigoplus_{i=1}^p (c_i \circledast (I \bullet \mathbf{A}_{1,b_i}))\right)\right) = (1, 2p, 1). \quad (4.35)$$

This establishes item (ii). Moreover, note that item (v) in Proposition 2.1.2, item (iii) in Lemma 2.3.3, item (iii) in Lemma 2.4.11, item (iii) in Lemma 2.3.5, and item (iii) in Lemma 2.3.2 establish that for all $x \in \mathbb{R}$ it holds that

$$\begin{aligned} (\mathcal{R}_a^N(\Psi))(x) &= c_0 + \left(\mathcal{R}_a^N\left(\bigoplus_{i=1}^p (c_i \circledast (I \bullet \mathbf{A}_{1,b_i}))\right)\right)(x) \\ &= c_0 + \sum_{i=1}^p c_i (\mathcal{R}_a^N(I))(\mathcal{R}_a^N(\mathbf{A}_{1,b_i})(x)) = c_0 + \sum_{i=1}^p c_i (x + b_i)^p. \end{aligned} \quad (4.36)$$

This, the hypothesis that $b_1 < b_2 < \dots < b_p$, the hypothesis that for all $k \in \{0, 1, \dots, p\}$ it holds that $\mathbb{1}_{\{p\}}(k) c_0 + \sum_{i=1}^p c_i (b_i)^k = \mathbb{1}_{\{p-1\}}(k) p^{-1}$, and Lemma 4.2.4 ensure that for all $x \in \mathbb{R}$ it holds that

$$(\mathcal{R}_a^N(\Psi))(x) = x. \quad (4.37)$$

This establishes items (iii) and (iv). The proof of Lemma 4.2.5 is thus complete. \square

4.2.3 ANN representations for maxima

Lemma 4.2.6 (Unique existence of maxima ANNs). *There exist unique $(\phi_d)_{d \in \mathbb{N}} \subseteq N$ 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*

$$\phi_2 = \left(\left(\begin{pmatrix} 1 & -1 \\ 0 & 1 \\ 0 & -1 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \right), ((1 \ 1 \ -1), (0)) \right) \in (\mathbb{R}^{3 \times 2} \times \mathbb{R}^3) \times (\mathbb{R}^{1 \times 3} \times \mathbb{R}^1), \quad (4.38)$$

(v) *it holds for all $d \in \{2, 3, 4, \dots\}$ that $\phi_{2d} = \phi_d \bullet (\mathbf{P}_d(\phi_2, \phi_2, \dots, \phi_2))$, and*

(vi) *it holds for all $d \in \{2, 3, 4, \dots\}$ that $\phi_{2d-1} = \phi_d \bullet (\mathbf{P}_d(\phi_2, \phi_2, \dots, \phi_2, \mathfrak{J}_1))$*

(cf. Definitions 1.3.1, 2.1.1, 2.2.1, 2.2.6, and 2.3.1).

Proof of Lemma 4.2.6. Throughout this proof, let $\psi \in N$ satisfy

$$\psi = \left(\left(\begin{pmatrix} 1 & -1 \\ 0 & 1 \\ 0 & -1 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \right), ((1 \ 1 \ -1), (0)) \right) \in (\mathbb{R}^{3 \times 2} \times \mathbb{R}^3) \times (\mathbb{R}^{1 \times 3} \times \mathbb{R}^1) \quad (4.39)$$

(cf. Definition 1.3.1). Observe that (4.39) and Lemma 2.2.7 imply that

$$\mathcal{I}(\psi) = 2, \quad \mathcal{O}(\psi) = \mathcal{I}(\mathfrak{J}_1) = \mathcal{O}(\mathfrak{J}_1) = 1, \quad \text{and} \quad \mathcal{L}(\psi) = \mathcal{L}(\mathfrak{J}_1) = 2. \quad (4.40)$$

Lemma 2.2.2 and Lemma 2.2.7 therefore demonstrate that for all $d \in \mathbb{N} \cap (1, \infty)$ it holds that

$$\mathcal{I}(\mathbf{P}_d(\psi, \psi, \dots, \psi)) = 2d, \quad \mathcal{O}(\mathbf{P}_d(\psi, \psi, \dots, \psi)) = d, \quad (4.41)$$

$$\mathcal{I}(\mathbf{P}_d(\psi, \psi, \dots, \psi, \mathfrak{J}_1)) = 2d - 1, \quad \text{and} \quad \mathcal{O}(\mathbf{P}_d(\psi, \psi, \dots, \psi, \mathfrak{J}_1)) = d \quad (4.42)$$

(cf. Definitions 2.2.1 and 2.2.6). Combining (4.40), Proposition 2.1.2, and induction hence proves 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

$$\phi_d = \begin{cases} \mathbf{A}_{1,0} & : d = 1 \\ \psi & : d = 2 \\ \phi_{d/2} \bullet (\mathbf{P}_{d/2}(\psi, \psi, \dots, \psi)) & : d \in \{4, 6, 8, \dots\} \\ \phi_{(d+1)/2} \bullet (\mathbf{P}_{(d+1)/2}(\psi, \psi, \dots, \psi, \mathfrak{J}_1)) & : d \in \{3, 5, 7, \dots\}. \end{cases} \quad (4.43)$$

The proof of Lemma 4.2.6 is thus complete. \square

Definition 4.2.7 (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*

$$\mathbb{M}_2 = \left(\left(\begin{pmatrix} 1 & -1 \\ 0 & 1 \\ 0 & -1 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \right), ((1 \ 1 \ -1), (0)) \right) \in (\mathbb{R}^{3 \times 2} \times \mathbb{R}^3) \times (\mathbb{R}^{1 \times 3} \times \mathbb{R}^1), \quad (4.44)$$

- (v) *it holds for all $d \in \{2, 3, 4, \dots\}$ that $\mathbb{M}_{2d} = \mathbb{M}_d \bullet (\mathbf{P}_d(\mathbb{M}_2, \mathbb{M}_2, \dots, \mathbb{M}_2))$, and*
- (vi) *it holds for all $d \in \{2, 3, 4, \dots\}$ that $\mathbb{M}_{2d-1} = \mathbb{M}_d \bullet (\mathbf{P}_d(\mathbb{M}_2, \mathbb{M}_2, \dots, \mathbb{M}_2, \mathfrak{J}_1))$*

(cf. Definitions 1.3.1, 2.1.1, 2.2.1, 2.2.6, and 2.3.1 and Lemma 4.2.6).

Definition 4.2.8 (Floor and ceiling of real numbers). *We denote by $\lceil \cdot \rceil: \mathbb{R} \rightarrow \mathbb{Z}$ and $\lfloor \cdot \rfloor: \mathbb{R} \rightarrow \mathbb{Z}$ the functions which satisfy for all $x \in \mathbb{R}$ that*

$$\lceil x \rceil = \min(\mathbb{Z} \cap [x, \infty)) \quad \text{and} \quad \lfloor x \rfloor = \max(\mathbb{Z} \cap (-\infty, x]). \quad (4.45)$$

Exercise 4.2.1. Prove or disprove the following statement: For all $n \in \{3, 5, 7, \dots\}$ it holds that $\lceil \log_2(n+1) \rceil = \lceil \log_2(n) \rceil$.

Proposition 4.2.9 (Properties of 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 $\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbb{M}_d) \in C(\mathbb{R}^d, \mathbb{R})$, and*
- (iv) *it holds for all $x = (x_1, \dots, x_d) \in \mathbb{R}^d$ that $(\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbb{M}_d))(x) = \max\{x_1, x_2, \dots, x_d\}$ (cf. Definitions 1.2.4, 1.3.1, 1.3.4, 4.2.7, and 4.2.8).*

Proof of Proposition 4.2.9. Throughout this proof, assume without loss of generality that $d > 1$. Note that (4.44) ensures that

$$\mathcal{H}(\mathbb{M}_2) = 1 \quad (4.46)$$

(cf. Definitions 1.3.1 and 4.2.7). This and (2.44) demonstrate that for all $\mathfrak{d} \in \{2, 3, 4, \dots\}$ it holds that

$$\mathcal{H}(\mathbf{P}_{\mathfrak{d}}(\mathbb{M}_2, \mathbb{M}_2, \dots, \mathbb{M}_2)) = \mathcal{H}(\mathbf{P}_{\mathfrak{d}}(\mathbb{M}_2, \mathbb{M}_2, \dots, \mathbb{M}_2, \mathfrak{I}_1)) = \mathcal{H}(\mathbb{M}_2) = 1 \quad (4.47)$$

(cf. Definitions 2.2.1 and 2.2.6). Combining this with Proposition 2.1.2 establishes that for all $\mathfrak{d} \in \{3, 4, 5, \dots\}$ it holds that

$$\mathcal{H}(\mathbb{M}_{\mathfrak{d}}) = \mathcal{H}(\mathbb{M}_{\lceil \mathfrak{d}/2 \rceil}) + 1 \quad (4.48)$$

(cf. Definition 4.2.8). This assures that for all $\mathfrak{d} \in \{4, 6, 8, \dots\}$ with $\mathcal{H}(\mathbb{M}_{\mathfrak{d}/2}) = \lceil \log_2(\mathfrak{d}/2) \rceil$ it holds that

$$\begin{aligned} \mathcal{H}(\mathbb{M}_{\mathfrak{d}}) &= \mathcal{H}(\mathbb{M}_{\lceil \mathfrak{d}/2 \rceil}) + 1 = \mathcal{H}(\mathbb{M}_{\mathfrak{d}/2}) + 1 \\ &= \lceil \log_2(\mathfrak{d}/2) \rceil + 1 = \lceil \log_2(\mathfrak{d}) - 1 \rceil + 1 = \lceil \log_2(\mathfrak{d}) \rceil. \end{aligned} \quad (4.49)$$

Furthermore, note that (4.48) and the fact that for all $\mathfrak{d} \in \{3, 5, 7, \dots\}$ it holds that $\lceil \log_2(\mathfrak{d}+1) \rceil = \lceil \log_2(\mathfrak{d}) \rceil$ show that for all $\mathfrak{d} \in \{3, 5, 7, \dots\}$ with $\mathcal{H}(\mathbb{M}_{\lceil \mathfrak{d}/2 \rceil}) = \lceil \log_2(\lceil \mathfrak{d}/2 \rceil) \rceil$ it holds that

$$\begin{aligned} \mathcal{H}(\mathbb{M}_{\mathfrak{d}}) &= \mathcal{H}(\mathbb{M}_{\lceil \mathfrak{d}/2 \rceil}) + 1 = \lceil \log_2(\lceil \mathfrak{d}/2 \rceil) \rceil + 1 = \lceil \log_2((\mathfrak{d}+1)/2) \rceil + 1 \\ &= \lceil \log_2(\mathfrak{d}+1) - 1 \rceil + 1 = \lceil \log_2(\mathfrak{d}+1) \rceil = \lceil \log_2(\mathfrak{d}) \rceil. \end{aligned} \quad (4.50)$$

Combining this and (4.49) demonstrates that for all $\mathfrak{d} \in \{3, 4, 5, \dots\}$ with $\forall k \in \{2, 3, \dots, \mathfrak{d}-1\}: \mathcal{H}(\mathbb{M}_k) = \lceil \log_2(k) \rceil$ it holds that

$$\mathcal{H}(\mathbb{M}_{\mathfrak{d}}) = \lceil \log_2(\mathfrak{d}) \rceil. \quad (4.51)$$

The fact that $\mathcal{H}(\mathbb{M}_2) = 1$ and induction hence establish item (i). Observe that the fact that $\mathcal{D}(\mathbb{M}_2) = (2, 3, 1)$ assure that for all $i \in \mathbb{N}$ it holds that

$$\mathbb{D}_i(\mathbb{M}_2) \leq 3 = 3 \lceil \frac{2}{2^i} \rceil. \quad (4.52)$$

Moreover, note that Proposition 2.1.2 and Lemma 2.2.2 imply that for all $\mathfrak{d} \in \{2, 3, 4, \dots\}$, $i \in \mathbb{N}$ it holds that

$$\mathbb{D}_i(\mathbb{M}_{2\mathfrak{d}}) = \mathbb{D}_i(\mathbb{M}_{\mathfrak{d}} \bullet (\mathbf{P}_{\mathfrak{d}}(\mathbb{M}_2, \mathbb{M}_2, \dots, \mathbb{M}_2))) = \begin{cases} 3\mathfrak{d} & : i = 1 \\ \mathbb{D}_{i-1}(\mathbb{M}_{\mathfrak{d}}) & : i \geq 2 \end{cases} \quad (4.53)$$

and

$$\mathbb{D}_i(\mathbb{M}_{2\mathfrak{d}-1}) = \mathbb{D}_i(\mathbb{M}_{\mathfrak{d}} \bullet (\mathbf{P}_{\mathfrak{d}}(\mathbb{M}_2, \mathbb{M}_2, \dots, \mathbb{M}_2, \mathfrak{I}_1))) = \begin{cases} 3\mathfrak{d} - 1 & : i = 1 \\ \mathbb{D}_{i-1}(\mathbb{M}_{\mathfrak{d}}) & : i \geq 2. \end{cases} \quad (4.54)$$

This and (4.51) assure that for all $\mathfrak{d} \in \{2, 4, 6, \dots\}$ it holds that

$$\mathbb{D}_1(\mathbb{M}_{\mathfrak{d}}) = 3(\frac{\mathfrak{d}}{2}) = 3 \lceil \frac{\mathfrak{d}}{2} \rceil. \quad (4.55)$$

In addition, observe that (4.54) establishes that for all $\mathfrak{d} \in \{3, 5, 7, \dots\}$ it holds that

$$\mathbb{D}_1(\mathbb{M}_{\mathfrak{d}}) = 3 \lceil \frac{\mathfrak{d}}{2} \rceil - 1 \leq 3 \lceil \frac{\mathfrak{d}}{2} \rceil. \quad (4.56)$$

This and (4.55) show that for all $\mathfrak{d} \in \{2, 3, 4, \dots\}$ it holds that

$$\mathbb{D}_1(\mathbb{M}_{\mathfrak{d}}) \leq 3 \lceil \frac{\mathfrak{d}}{2} \rceil. \quad (4.57)$$

Next note that (4.53) ensures that for all $\mathfrak{d} \in \{4, 6, 8, \dots\}$, $i \in \{2, 3, 4, \dots\}$ 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

$$\mathbb{D}_i(\mathbb{M}_{\mathfrak{d}}) = \mathbb{D}_{i-1}(\mathbb{M}_{\mathfrak{d}/2}) \leq 3 \lceil (\mathfrak{d}/2) \frac{1}{2^{i-1}} \rceil = 3 \lceil \frac{\mathfrak{d}}{2^i} \rceil. \quad (4.58)$$

Furthermore, observe that (4.54) and the fact that for all $\mathfrak{d} \in \{3, 5, 7, \dots\}$, $i \in \mathbb{N}$ it holds that $\lceil \frac{\mathfrak{d}+1}{2^i} \rceil = \lceil \frac{\mathfrak{d}}{2^i} \rceil$ show that for all $\mathfrak{d} \in \{3, 5, 7, \dots\}$, $i \in \{2, 3, 4, \dots\}$ 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

$$\mathbb{D}_i(\mathbb{M}_{\mathfrak{d}}) = \mathbb{D}_{i-1}(\mathbb{M}_{\lceil \mathfrak{d}/2 \rceil}) \leq 3 \lceil \lceil \mathfrak{d}/2 \rceil \frac{1}{2^{i-1}} \rceil = 3 \lceil \frac{\mathfrak{d}+1}{2^i} \rceil = 3 \lceil \frac{\mathfrak{d}}{2^i} \rceil. \quad (4.59)$$

This, (4.57), and (4.58) establish that for all $\mathfrak{d} \in \{3, 4, 5, \dots\}$, $i \in \mathbb{N}$ with $\forall k \in \{2, 3, \dots, \mathfrak{d}-1\}$, $j \in \mathbb{N}$: $\mathbb{D}_j(\mathbb{M}_k) \leq 3 \lceil \frac{k}{2^j} \rceil$ it holds that

$$\mathbb{D}_i(\mathbb{M}_{\mathfrak{d}}) \leq 3 \lceil \frac{\mathfrak{d}}{2^i} \rceil. \quad (4.60)$$

Combining this and (4.52) with induction establishes item (ii). Note that (4.44) ensures that for all $x = (x_1, x_2) \in \mathbb{R}^2$ it holds that

$$\begin{aligned} (\mathcal{R}_r^N(\mathbb{M}_2))(x) &= \max\{x_1 - x_2, 0\} + \max\{x_2, 0\} - \max\{-x_2, 0\} \\ &= \max\{x_1 - x_2, 0\} + x_2 = \max\{x_1, x_2\} \end{aligned} \quad (4.61)$$

(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 $\mathfrak{d} \in \{2, 3, 4, \dots\}$, $x = (x_1, x_2, \dots, x_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$ it holds that

$$\mathcal{R}_r^N(\mathbb{M}_{\mathfrak{d}}) \in C(\mathbb{R}^{\mathfrak{d}}, \mathbb{R}) \quad \text{and} \quad (\mathcal{R}_r^N(\mathbb{M}_{\mathfrak{d}}))(x) = \max\{x_1, x_2, \dots, x_{\mathfrak{d}}\}. \quad (4.62)$$

This establishes items (iii) and (iv). The proof of Proposition 4.2.9 is thus complete. \square

Lemma 4.2.10. *Let $d \in \mathbb{N}$, $i \in \{1, 2, \dots, \mathcal{L}(\mathbb{M}_d)\}$ (cf. Definitions 1.3.1 and 4.2.7). 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 4.2.10. Throughout this proof, assume without loss of generality that $d > 2$ (cf. items (iii) and (iv) in Definition 4.2.7) 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

$$A_1 = \begin{pmatrix} 1 & -1 \\ 0 & 1 \\ 0 & -1 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 1 & 1 & -1 \end{pmatrix}, \quad C_1 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \quad \text{and} \quad C_2 = \begin{pmatrix} 1 & -1 \end{pmatrix}. \quad (4.63)$$

Note that items (iv), (v), and (vi) in Definition 4.2.7 assure that for all $\mathfrak{d} \in \{2, 3, 4, \dots\}$ it holds that

$$\mathcal{W}_{1,\mathbb{M}_{2\mathfrak{d}-1}} = \underbrace{\begin{pmatrix} A_1 & 0 & \cdots & 0 & 0 \\ 0 & A_1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & A_1 & 0 \\ 0 & 0 & \cdots & 0 & C_1 \end{pmatrix}}_{\in \mathbb{R}^{(3\mathfrak{d}-1) \times (2\mathfrak{d}-1)}}, \quad \mathcal{W}_{1,\mathbb{M}_{2\mathfrak{d}}} = \underbrace{\begin{pmatrix} A_1 & 0 & \cdots & 0 \\ 0 & A_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_1 \end{pmatrix}}_{\in \mathbb{R}^{(3\mathfrak{d}) \times (2\mathfrak{d})}}, \quad (4.64)$$

$$\mathcal{B}_{1,\mathbb{M}_{2\mathfrak{d}-1}} = 0 \in \mathbb{R}^{3\mathfrak{d}-1}, \quad \text{and} \quad \mathcal{B}_{1,\mathbb{M}_{2\mathfrak{d}}} = 0 \in \mathbb{R}^{3\mathfrak{d}}.$$

This and (4.63) proves item (iii). Furthermore, note that (4.64) and item (iv) in Definition 4.2.7 imply that for all $\mathfrak{d} \in \{2, 3, 4, \dots\}$ it holds that $\mathcal{B}_{1,\mathbb{M}_\mathfrak{d}} = 0$. Items (iv), (v), and (vi) in Definition 4.2.7 hence ensure that for all $\mathfrak{d} \in \{2, 3, 4, \dots\}$ it holds that

$$\mathcal{W}_{2,\mathbb{M}_{2\mathfrak{d}-1}} = \underbrace{\mathcal{W}_{1,\mathbb{M}_\mathfrak{d}} \begin{pmatrix} A_2 & 0 & \cdots & 0 & 0 \\ 0 & A_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & A_2 & 0 \\ 0 & 0 & \cdots & 0 & C_2 \end{pmatrix}}_{\in \mathbb{R}^{\mathfrak{d} \times (3\mathfrak{d}-1)}}, \quad \mathcal{W}_{2,\mathbb{M}_{2\mathfrak{d}}} = \underbrace{\mathcal{W}_{1,\mathbb{M}_\mathfrak{d}} \begin{pmatrix} A_2 & 0 & \cdots & 0 & 0 \\ 0 & A_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & A_2 & 0 \end{pmatrix}}_{\in \mathbb{R}^{\mathfrak{d} \times (3\mathfrak{d})}},$$

$$\mathcal{B}_{2,\mathbb{M}_{2\mathfrak{d}-1}} = \mathcal{B}_{1,\mathbb{M}_\mathfrak{d}} = 0, \quad \text{and} \quad \mathcal{B}_{2,\mathbb{M}_{2\mathfrak{d}}} = \mathcal{B}_{1,\mathbb{M}_\mathfrak{d}} = 0. \quad (4.65)$$

Combining this and item (iv) in Definition 4.2.7 shows that for all $\mathfrak{d} \in \{2, 3, 4, \dots\}$ it holds that $\mathcal{B}_{2,\mathbb{M}_\mathfrak{d}} = 0$. Moreover, note that (2.2) demonstrates that for all $\mathfrak{d} \in \{2, 3, 4, \dots, \mathfrak{d}\}$, $i \in \{3, 4, \dots, \mathcal{L}(\mathbb{M}_\mathfrak{d}) + 1\}$ it holds that

$$\mathcal{W}_{i,\mathbb{M}_{2\mathfrak{d}-1}} = \mathcal{W}_{i,\mathbb{M}_{2\mathfrak{d}}} = \mathcal{W}_{i-1,\mathbb{M}_\mathfrak{d}} \quad \text{and} \quad \mathcal{B}_{i,\mathbb{M}_{2\mathfrak{d}-1}} = \mathcal{B}_{i,\mathbb{M}_{2\mathfrak{d}}} = \mathcal{B}_{i-1,\mathbb{M}_\mathfrak{d}}. \quad (4.66)$$

This, (4.63), (4.64), (4.65), the fact that for all $\mathfrak{d} \in \{2, 3, 4, \dots\}$ it holds that $\mathcal{B}_{2,\mathbb{M}_\mathfrak{d}} = 0$, and induction establish items (i) and (ii). The proof of Lemma 4.2.10 is thus complete. \square

4.2.4 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

$$\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbb{M}_d \bullet \mathbf{P}_d(\mathbb{L}_1, \dots, \mathbb{L}_1))(x) = \|x\|_\infty \quad (4.67)$$

(cf. Definitions 1.2.4, 1.3.4, 2.1.1, 2.2.1, 3.3.4, 4.2.1, and 4.2.7).

Lemma 4.2.11. Let $d, K \in \mathbb{N}$, $L \in [0, \infty)$, $\mathfrak{x}_1, \mathfrak{x}_2, \dots, \mathfrak{x}_K \in \mathbb{R}^d$, $\mathfrak{y} = (\mathfrak{y}_1, \dots, \mathfrak{y}_K) \in \mathbb{R}^K$, $\Phi \in \mathbf{N}$ satisfy

$$\Phi = \mathbb{M}_K \bullet \mathbf{A}_{-L\mathbb{I}_K, \mathfrak{y}} \bullet \mathbf{P}_K(\mathbb{L}_d \bullet \mathbf{A}_{\mathbb{I}_d, -\mathfrak{x}_1}, \mathbb{L}_d \bullet \mathbf{A}_{\mathbb{I}_d, -\mathfrak{x}_2}, \dots, \mathbb{L}_d \bullet \mathbf{A}_{\mathbb{I}_d, -\mathfrak{x}_K}) \bullet \mathbb{T}_{d,K} \quad (4.68)$$

(cf. Definitions 1.3.1, 1.5.5, 2.1.1, 2.2.1, 2.3.1, 2.4.6, 4.2.1, and 4.2.7). 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 $i \in \{2, 3, 4, \dots\}$ that $\mathbb{D}_i(\Phi) \leq 3 \lceil \frac{K}{2^{i-1}} \rceil$,
- (vi) it holds that $\|\mathcal{T}(\Phi)\|_\infty \leq \max\{1, L, \max_{k \in \{1, 2, \dots, K\}} \|\mathfrak{x}_k\|_\infty, 2\|\mathfrak{y}\|_\infty\}$, and
- (vii) it holds for all $x \in \mathbb{R}^d$ that $(\mathcal{R}_{\mathfrak{r}}^N(\Phi))(x) = \max_{k \in \{1, 2, \dots, K\}} (\mathfrak{y}_k - L\|x - \mathfrak{x}_k\|_1)$
(cf. Definitions 1.2.4, 1.3.4, 1.3.6, 3.3.4, and 4.2.8).

Proof of Lemma 4.2.11. Throughout this proof, let $\Psi_k \in \mathbf{N}$, $k \in \{1, 2, \dots, K\}$, satisfy for all $k \in \{1, 2, \dots, K\}$ that $\Psi_k = \mathbb{L}_d \bullet \mathbf{A}_{I_d, -\mathfrak{x}_k}$, let $\Xi \in \mathbf{N}$ satisfy

$$\Xi = \mathbf{A}_{-L I_K, \mathfrak{y}} \bullet \mathbf{P}_K(\Psi_1, \Psi_2, \dots, \Psi_K) \bullet \mathbb{T}_{d, K}, \quad (4.69)$$

and let $\|\cdot\|: \bigcup_{m, n \in \mathbb{N}} \mathbb{R}^{m \times n} \rightarrow [0, \infty)$ satisfy for all $m, n \in \mathbb{N}$, $M = (M_{i,j})_{i \in \{1, \dots, m\}, j \in \{1, \dots, n\}} \in \mathbb{R}^{m \times n}$ that $\|M\| = \max_{i \in \{1, \dots, m\}, j \in \{1, \dots, n\}} |M_{i,j}|$. Observe that (4.68) 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

$$\mathcal{H}(\Xi) = \mathcal{H}(\mathbf{A}_{-L I_K, \mathfrak{y}}) + \mathcal{H}(\mathbf{P}_K(\Psi_1, \Psi_2, \dots, \Psi_K)) + \mathcal{H}(\mathbb{T}_{d, K}) = \mathcal{H}(\Psi_1) = \mathcal{H}(\mathbb{L}_d) = 1. \quad (4.70)$$

Proposition 2.1.2 and Proposition 4.2.9 hence ensure that

$$\mathcal{H}(\Phi) = \mathcal{H}(\mathbb{M}_K \bullet \Xi) = \mathcal{H}(\mathbb{M}_K) + \mathcal{H}(\Xi) = \lceil \log_2(K) \rceil + 1 \quad (4.71)$$

(cf. Definition 4.2.8). This establishes item (iii). Next observe that the fact that $\mathcal{H}(\Xi) = 1$, Proposition 2.1.2, and Proposition 4.2.9 assure that for all $i \in \{2, 3, 4, \dots\}$ it holds that

$$\mathbb{D}_i(\Phi) = \mathbb{D}_{i-1}(\mathbb{M}_K) \leq 3 \lceil \frac{K}{2^{i-1}} \rceil. \quad (4.72)$$

This proves item (v). Furthermore, note that Proposition 2.1.2, Proposition 2.2.4, and Proposition 4.2.2 assure that

$$\mathbb{D}_1(\Phi) = \mathbb{D}_1(\Xi) = \mathbb{D}_1(\mathbf{P}_K(\Psi_1, \Psi_2, \dots, \Psi_K)) = \sum_{i=1}^K \mathbb{D}_1(\Psi_i) = \sum_{i=1}^K \mathbb{D}_1(\mathbb{L}_d) = 2dK. \quad (4.73)$$

This establishes item (iv). Moreover, observe that (2.2) and Lemma 4.2.10 imply that

$$\Phi = \left((\mathcal{W}_{1, \Xi}, \mathcal{B}_{1, \Xi}), (\mathcal{W}_{1, \mathbb{M}_K} \mathcal{W}_{2, \Xi}, \mathcal{W}_{1, \mathbb{M}_K} \mathcal{B}_{2, \Xi}), \right. \\ \left. (\mathcal{W}_{2, \mathbb{M}_K}, 0), \dots, (\mathcal{W}_{\mathcal{L}(\mathbb{M}_K), \mathbb{M}_K}, 0) \right). \quad (4.74)$$

Next note that the fact that for all $k \in \{1, 2, \dots, K\}$ it holds that $\mathcal{W}_{1,\Psi_k} = \mathcal{W}_{1,\mathbf{A}_{I_d,-\mathfrak{x}_k}} \mathcal{W}_{1,\mathbb{L}_d} = \mathcal{W}_{1,\mathbb{L}_d}$ assures that

$$\begin{aligned}\mathcal{W}_{1,\Xi} &= \mathcal{W}_{1,\mathbf{P}_K(\Psi_1, \Psi_2, \dots, \Psi_K)} \mathcal{W}_{1,\mathbb{T}_{d,K}} = \begin{pmatrix} \mathcal{W}_{1,\Psi_1} & 0 & \cdots & 0 \\ 0 & \mathcal{W}_{1,\Psi_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathcal{W}_{1,\Psi_K} \end{pmatrix} \begin{pmatrix} \mathbf{I}_d \\ \mathbf{I}_d \\ \vdots \\ \mathbf{I}_d \end{pmatrix} \quad (4.75) \\ &= \begin{pmatrix} \mathcal{W}_{1,\Psi_1} \\ \mathcal{W}_{1,\Psi_2} \\ \vdots \\ \mathcal{W}_{1,\Psi_K} \end{pmatrix} = \begin{pmatrix} \mathcal{W}_{1,\mathbb{L}_d} \\ \mathcal{W}_{1,\mathbb{L}_d} \\ \vdots \\ \mathcal{W}_{1,\mathbb{L}_d} \end{pmatrix}.\end{aligned}$$

Lemma 4.2.3 hence demonstrates that $\|\mathcal{W}_{1,\Xi}\| = 1$. In addition, note that (2.2) implies that

$$\mathcal{B}_{1,\Xi} = \mathcal{W}_{1,\mathbf{P}_K(\Psi_1, \Psi_2, \dots, \Psi_K)} \mathcal{B}_{1,\mathbb{T}_{d,K}} + \mathcal{B}_{1,\mathbf{P}_K(\Psi_1, \Psi_2, \dots, \Psi_K)} = \mathcal{B}_{1,\mathbf{P}_K(\Psi_1, \Psi_2, \dots, \Psi_K)} = \begin{pmatrix} \mathcal{B}_{1,\Psi_1} \\ \mathcal{B}_{1,\Psi_2} \\ \vdots \\ \mathcal{B}_{1,\Psi_K} \end{pmatrix}. \quad (4.76)$$

Furthermore, observe that Lemma 4.2.3 implies that for all $k \in \{1, 2, \dots, K\}$ it holds that

$$\mathcal{B}_{1,\Psi_k} = \mathcal{W}_{1,\mathbb{L}_d} \mathcal{B}_{1,\mathbf{A}_{I_d,-\mathfrak{x}_k}} + \mathcal{B}_{1,\mathbb{L}_d} = -\mathcal{W}_{1,\mathbb{L}_d} \mathfrak{x}_k. \quad (4.77)$$

This, (4.76), and Lemma 4.2.3 show that

$$\|\mathcal{B}_{1,\Xi}\|_\infty = \max_{k \in \{1, 2, \dots, K\}} \|\mathcal{B}_{1,\Psi_k}\|_\infty = \max_{k \in \{1, 2, \dots, K\}} \|\mathcal{W}_{1,\mathbb{L}_d} \mathfrak{x}_k\|_\infty = \max_{k \in \{1, 2, \dots, K\}} \|\mathfrak{x}_k\|_\infty \quad (4.78)$$

(cf. Definition 3.3.4). Combining this, (4.74), Lemma 4.2.10, and the fact that $\|\mathcal{W}_{1,\Xi}\| = 1$ shows that

$$\begin{aligned}\|\mathcal{T}(\Phi)\|_\infty &= \max\{\|\mathcal{W}_{1,\Xi}\|, \|\mathcal{B}_{1,\Xi}\|_\infty, \|\mathcal{W}_{1,\mathbb{M}_K} \mathcal{W}_{2,\Xi}\|, \|\mathcal{W}_{1,\mathbb{M}_K} \mathcal{B}_{2,\Xi}\|_\infty, 1\} \\ &= \max\{1, \max_{k \in \{1, 2, \dots, K\}} \|\mathfrak{x}_k\|_\infty, \|\mathcal{W}_{1,\mathbb{M}_K} \mathcal{W}_{2,\Xi}\|, \|\mathcal{W}_{1,\mathbb{M}_K} \mathcal{B}_{2,\Xi}\|_\infty\}\end{aligned} \quad (4.79)$$

(cf. Definition 1.3.6). Next note that Lemma 4.2.3 ensures that for all $k \in \{1, 2, \dots, K\}$ it holds that $\mathcal{B}_{2,\Psi_k} = \mathcal{B}_{2,\mathbb{L}_d} = 0$. Hence, we obtain that $\mathcal{B}_{2,\mathbf{P}_K(\Psi_1, \Psi_2, \dots, \Psi_K)} = 0$. This implies that

$$\mathcal{B}_{2,\Xi} = \mathcal{W}_{1,\mathbf{A}_{-L\mathbb{I}_K,\mathfrak{y}}} \mathcal{B}_{2,\mathbf{P}_K(\Psi_1, \Psi_2, \dots, \Psi_K)} + \mathcal{B}_{1,\mathbf{A}_{-L\mathbb{I}_K,\mathfrak{y}}} = \mathcal{B}_{1,\mathbf{A}_{-L\mathbb{I}_K,\mathfrak{y}}} = \mathfrak{y}. \quad (4.80)$$

In addition, observe that the fact that for all $k \in \{1, 2, \dots, K\}$ it holds that $\mathcal{W}_{2,\Psi_k} = \mathcal{W}_{2,\mathbb{L}_d}$ assures that

$$\begin{aligned} \mathcal{W}_{2,\Xi} &= \mathcal{W}_{1,\mathbf{A}_{-L\mathbb{I}_K,\mathfrak{y}}} \mathcal{W}_{2,\mathbf{P}_K(\Psi_1, \Psi_2, \dots, \Psi_K)} = -L\mathcal{W}_{2,\mathbf{P}_K(\Psi_1, \Psi_2, \dots, \Psi_K)} \\ &= -L \begin{pmatrix} \mathcal{W}_{2,\Psi_1} & 0 & \cdots & 0 \\ 0 & \mathcal{W}_{2,\Psi_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathcal{W}_{2,\Psi_K} \end{pmatrix} = \begin{pmatrix} -L\mathcal{W}_{2,\mathbb{L}_d} & 0 & \cdots & 0 \\ 0 & -L\mathcal{W}_{2,\mathbb{L}_d} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -L\mathcal{W}_{2,\mathbb{L}_d} \end{pmatrix}. \end{aligned} \quad (4.81)$$

Item (v) in Lemma 4.2.3 and Lemma 4.2.10 hence imply that

$$\|\mathcal{W}_{1,\mathbb{M}_K} \mathcal{W}_{2,\Xi}\| = L \|\mathcal{W}_{1,\mathbb{M}_K}\| \leq L. \quad (4.82)$$

Moreover, observe that (4.80) and Lemma 4.2.10 show that

$$\|\mathcal{W}_{1,\mathbb{M}_K} \mathcal{B}_{2,\Xi}\|_\infty \leq 2 \|\mathcal{B}_{2,\Xi}\|_\infty = 2 \|\mathfrak{y}\|_\infty. \quad (4.83)$$

Combining this with (4.79) and (4.82) establishes item (vi). Next observe that Proposition 4.2.2 and Lemma 2.3.3 show that for all $x \in \mathbb{R}^d$, $k \in \{1, 2, \dots, K\}$ it holds that

$$(\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\Psi_k))(x) = (\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbb{L}_d) \circ \mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{A}_{\mathbb{I}_d, -\mathfrak{x}_k}))(x) = \|x - \mathfrak{x}_k\|_1. \quad (4.84)$$

This, Proposition 2.2.3, and Proposition 2.1.2 imply that for all $x \in \mathbb{R}^d$ it holds that

$$(\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{P}_K(\Psi_1, \Psi_2, \dots, \Psi_K) \bullet \mathbb{T}_{d,K}))(x) = (\|x - \mathfrak{x}_1\|_1, \|x - \mathfrak{x}_2\|_1, \dots, \|x - \mathfrak{x}_K\|_1). \quad (4.85)$$

(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

$$\begin{aligned} (\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\Xi))(x) &= (\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{A}_{-L\mathbb{I}_K,\mathfrak{y}}) \circ \mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbf{P}_K(\Psi_1, \Psi_2, \dots, \Psi_K) \bullet \mathbb{T}_{d,K}))(x) \\ &= (\mathfrak{y}_1 - L\|x - \mathfrak{x}_1\|_1, \mathfrak{y}_2 - L\|x - \mathfrak{x}_2\|_1, \dots, \mathfrak{y}_K - L\|x - \mathfrak{x}_K\|_1). \end{aligned} \quad (4.86)$$

Proposition 2.1.2 and Proposition 4.2.9 hence demonstrate that for all $x \in \mathbb{R}^d$ it holds that

$$\begin{aligned} (\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\Phi))(x) &= (\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbb{M}_K) \circ \mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\Xi))(x) \\ &= (\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathbb{M}_K))(\mathfrak{y}_1 - L\|x - \mathfrak{x}_1\|_1, \mathfrak{y}_2 - L\|x - \mathfrak{x}_2\|_1, \dots, \mathfrak{y}_K - L\|x - \mathfrak{x}_K\|_1) \\ &= \max_{k \in \{1, 2, \dots, K\}} (\mathfrak{y}_k - L\|x - \mathfrak{x}_k\|_1). \end{aligned} \quad (4.87)$$

This establishes item (vii). The proof of Lemma 4.2.11 is thus complete. \square

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 $\mathfrak{x}_1, \mathfrak{x}_2, \dots, \mathfrak{x}_K \in E$, let $f: E \rightarrow \mathbb{R}$ 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 $\mathfrak{y} = (f(\mathfrak{x}_1), f(\mathfrak{x}_2), \dots, f(\mathfrak{x}_K))$ and

$$\Phi = \mathbb{M}_K \bullet \mathbf{A}_{-L\mathbf{I}_K, \mathfrak{y}} \bullet \mathbf{P}_K(\mathbb{L}_d \bullet \mathbf{A}_{\mathbf{I}_d, -\mathfrak{x}_1}, \mathbb{L}_d \bullet \mathbf{A}_{\mathbf{I}_d, -\mathfrak{x}_2}, \dots, \mathbb{L}_d \bullet \mathbf{A}_{\mathbf{I}_d, -\mathfrak{x}_K}) \bullet \mathbb{T}_{d,K} \quad (4.88)$$

(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 4.2.7). Then

$$\sup_{x \in E} |(\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}})(\Phi))(x) - f(x)| \leq 2L[\sup_{x \in E} (\min_{k \in \{1, 2, \dots, K\}} \|x - \mathfrak{x}_k\|_1)] \quad (4.89)$$

(cf. Definitions 1.2.4 and 1.3.4).

Proof of Proposition 4.3.1. Throughout this proof, let $F: \mathbb{R}^d \rightarrow \mathbb{R}$ satisfy for all $x \in \mathbb{R}^d$ that

$$F(x) = \max_{k \in \{1, 2, \dots, K\}} (f(\mathfrak{x}_k) - L\|x - \mathfrak{x}_k\|_1). \quad (4.90)$$

Observe that Corollary 4.1.4, (4.90), and the assumption that for all $x, y \in E$ it holds that $|f(x) - f(y)| \leq L\|x - y\|_1$ establish that

$$\sup_{x \in E} |F(x) - f(x)| \leq 2L[\sup_{x \in E} (\min_{k \in \{1, 2, \dots, K\}} \|x - \mathfrak{x}_k\|_1)]. \quad (4.91)$$

Moreover, note that Lemma 4.2.11 ensures that for all $x \in E$ it holds that $F(x) = (\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}})(\Phi))(x)$. Combining this and (4.91) establishes (4.89). The proof of Proposition 4.3.1 is thus complete. \square

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

$$\sup_{v=(x,y) \in [0,2]^2} |x^2 + y^2 - 2x - 2y + 2 - (\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}})(\Phi))(v)| \leq \frac{3}{8}. \quad (4.92)$$

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

$$\mathcal{C}^{(E,\delta),r} = \min \left(\left\{ n \in \mathbb{N}_0 : \left[\exists A \subseteq E : \left(\begin{array}{l} (|A| \leq n) \wedge (\forall x \in E : \\ \exists a \in A : \delta(a, x) \leq r) \end{array} \right) \right] \right\} \cup \{\infty\} \right) \quad (4.93)$$

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 (X, d) be a metric space, let $n \in \mathbb{N}$, $r \in [0, \infty]$, assume $X \neq \emptyset$, and let $A \subseteq X$ satisfy $|A| \leq n$ and $\forall x \in X : \exists a \in A : d(a, x) \leq r$. Then there exist $x_1, x_2, \dots, x_n \in X$ such that

$$X \subseteq \left[\bigcup_{i=1}^n \{v \in X : d(x_i, v) \leq r\} \right]. \quad (4.94)$$

Proof of Lemma 4.3.3. Note that the assumption that $X \neq \emptyset$ and the assumption that $|A| \leq n$ imply that there exist $x_1, x_2, \dots, x_n \in X$ which satisfy $A \subseteq \{x_1, x_2, \dots, x_n\}$. This and the assumption that $\forall x \in X : \exists a \in A : d(a, x) \leq r$ ensure that

$$X \subseteq \left[\bigcup_{a \in A} \{v \in X : d(a, v) \leq r\} \right] \subseteq \left[\bigcup_{i=1}^n \{v \in X : d(x_i, v) \leq r\} \right]. \quad (4.95)$$

The proof of Lemma 4.3.3 is thus complete. \square

Lemma 4.3.4. Let (X, d) be a metric space, let $n \in \mathbb{N}$, $r \in [0, \infty]$, $x_1, x_2, \dots, x_n \in X$ satisfy $X \subseteq \left[\bigcup_{i=1}^n \{v \in X : d(x_i, v) \leq r\} \right]$. Then there exists $A \subseteq X$ such that $|A| \leq n$ and

$$\forall x \in X : \exists a \in A : d(a, x) \leq r. \quad (4.96)$$

Proof of Lemma 4.3.4. Throughout this proof, let $A = \{x_1, x_2, \dots, x_n\}$. Note that the assumption that $X \subseteq \left[\bigcup_{i=1}^n \{v \in X : d(x_i, v) \leq r\} \right]$ implies that for all $v \in X$ there exists $i \in \{1, 2, \dots, n\}$ such that $d(x_i, v) \leq r$. Hence, we obtain that

$$\forall x \in X : \exists a \in A : d(a, x) \leq r. \quad (4.97)$$

The proof of Lemma 4.3.4 is thus complete. \square

Lemma 4.3.5. Let (X, d) be a metric space, let $n \in \mathbb{N}$, $r \in [0, \infty]$, and assume $X \neq \emptyset$. Then the following two statements are equivalent:

- (i) There exists $A \subseteq X$ such that $|A| \leq n$ and $\forall x \in X : \exists a \in A : d(a, x) \leq r$.
- (ii) There exist $x_1, x_2, \dots, x_n \in X$ such that $X \subseteq [\bigcup_{i=1}^n \{v \in X : d(x_i, v) \leq r\}]$.

Proof of Lemma 4.3.5. Note that Lemma 4.3.3 and Lemma 4.3.4 prove that ((i) \leftrightarrow (ii)). The proof of Lemma 4.3.5 is thus complete. \square

Lemma 4.3.6. Let (E, δ) be a metric space and let $r \in [0, \infty]$. Then

$$\mathcal{C}^{(E, \delta), r} = \begin{cases} 0 & : X = \emptyset \\ \inf \left(\left\{ n \in \mathbb{N} : \left(\exists x_1, x_2, \dots, x_n \in E : \right. \right. \right. \\ \quad \left. \left. \left. E \subseteq \left[\bigcup_{m=1}^n \{v \in E : d(x_m, v) \leq r\} \right] \right) \right\} \cup \{\infty\} \right) & : X \neq \emptyset \end{cases} \quad (4.98)$$

(cf. Definition 4.3.2).

Proof of Lemma 4.3.6. Throughout this proof, assume without loss of generality that $E \neq \emptyset$. Observe that Lemma 4.3.5 establishes (4.98). The proof of Lemma 4.3.6 is thus complete. \square

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 4.3.7. 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 $\delta_p : ([a, b]^d) \times ([a, b]^d) \rightarrow [0, \infty)$ 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

$$\mathcal{C}^{([a, b]^d, \delta_p), r} \leq \left(\left\lceil \frac{d^{1/p}(b-a)}{2r} \right\rceil \right)^d \leq \begin{cases} 1 & : r \geq d(b-a)/2 \\ \left(\frac{d(b-a)}{r} \right)^d & : r < d(b-a)/2. \end{cases} \quad (4.99)$$

(cf. Definitions 4.2.8 and 4.3.2).

Proof of Lemma 4.3.7. Throughout this proof, let $(\mathfrak{N}_p)_{p \in [1, \infty)} \subseteq \mathbb{N}$ satisfy for all $p \in [1, \infty)$ that

$$\mathfrak{N}_p = \left\lceil \frac{d^{1/p}(b-a)}{2r} \right\rceil, \quad (4.100)$$

for every $N \in \mathbb{N}$, $i \in \{1, 2, \dots, N\}$ let $g_{N,i} \in [a, b]$ be given by

$$g_{N,i} = a + \frac{(i-1/2)(b-a)}{N} \quad (4.101)$$

and for every $p \in [1, \infty)$ let $A_p \subseteq [a, b]^d$ be given by

$$A_p = \{g_{\mathfrak{N}_p,1}, g_{\mathfrak{N}_p,2}, \dots, g_{\mathfrak{N}_p,\mathfrak{N}_p}\}^d \quad (4.102)$$

(cf. Definition 4.2.8). 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

$$|x - g_{N,i}| = a + \frac{(i-1/2)(b-a)}{N} - x \leq a + \frac{(i-1/2)(b-a)}{N} - \left(a + \frac{(i-1)(b-a)}{N}\right) = \frac{b-a}{2N}. \quad (4.103)$$

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

$$|x - g_{N,i}| = x - \left(a + \frac{(i-1/2)(b-a)}{N}\right) \leq a + \frac{i(b-a)}{N} - \left(a + \frac{(i-1/2)(b-a)}{N}\right) = \frac{b-a}{2N}. \quad (4.104)$$

Combining this with (4.103) 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

$$|x - y| \leq \frac{b-a}{2N}. \quad (4.105)$$

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

$$\delta_p(x, y) = \|x - y\|_p = \left(\sum_{i=1}^d |x_i - y_i|^p \right)^{1/p} \leq \left(\sum_{i=1}^d \frac{(b-a)^p}{(2\mathfrak{N}_p)^p} \right)^{1/p} = \frac{d^{1/p}(b-a)}{2\mathfrak{N}_p} \leq \frac{d^{1/p}(b-a)2r}{2d^{1/p}(b-a)} = r. \quad (4.106)$$

Combining this with (4.93), (4.102), (4.100), and the fact that $\forall x \in [0, \infty) : \lceil x \rceil \leq \mathbb{1}_{(0,1]}(x) + 2x\mathbb{1}_{(1,\infty)}(x) = \mathbb{1}_{(0,r]}(rx) + 2x\mathbb{1}_{(r,\infty)}(rx)$ yields that for all $p \in [1, \infty)$ it holds that

$$\begin{aligned} \mathcal{C}^{([a,b]^d, \delta_p), r} &\leq |A_p| = (\mathfrak{N}_p)^d = \left(\left\lceil \frac{d^{1/p}(b-a)}{2r} \right\rceil \right)^d \leq \left(\left\lceil \frac{d(b-a)}{2r} \right\rceil \right)^d \\ &\leq \left(\mathbb{1}_{(0,r]} \left(\frac{d(b-a)}{2} \right) + \frac{2d(b-a)}{2r} \mathbb{1}_{(r,\infty)} \left(\frac{d(b-a)}{2} \right) \right)^d \\ &= \mathbb{1}_{(0,r]} \left(\frac{d(b-a)}{2} \right) + \left(\frac{d(b-a)}{r} \right)^d \mathbb{1}_{(r,\infty)} \left(\frac{d(b-a)}{2} \right) \end{aligned} \quad (4.107)$$

(cf. Definition 4.3.2). The proof of Lemma 4.3.7 is thus complete. \square

4.3.3 Convergence rates for the approximation error

Lemma 4.3.8. Let $d \in \mathbb{N}$, $L, a \in \mathbb{R}$, $b \in (a, \infty)$, let $f: [a, b]^d \rightarrow \mathbb{R}$ satisfy for all $x, y \in [a, b]^d$ that $|f(x) - f(y)| \leq L\|x - y\|_1$, and let $\mathbf{F} = \mathbf{A}_{0, f((a+b)/2, (a+b)/2, \dots, (a+b)/2)} \in \mathbb{R}^{1 \times d} \times \mathbb{R}^1$ (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 $\sup_{x \in [a, b]^d} |(\mathcal{R}_r^N(\mathbf{F}))(x) - f(x)| \leq \frac{dL(b-a)}{2}$
- (cf. Definitions 1.2.4, 1.3.1, 1.3.4, and 1.3.6).

Proof of Lemma 4.3.8. 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

$$(\mathcal{R}_r^N(\mathbf{F}))(x) = f((a+b)/2, (a+b)/2, \dots, (a+b)/2). \quad (4.108)$$

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 $x = (x_1, \dots, x_d) \in [a, b]^d$ it holds that

$$\begin{aligned} |(\mathcal{R}_r^N(\mathbf{F}))(x) - f(x)| &= |f((a+b)/2, (a+b)/2, \dots, (a+b)/2) - f(x)| \\ &\leq L\|((a+b)/2, (a+b)/2, \dots, (a+b)/2) - x\|_1 \\ &= L \sum_{i=1}^d |(a+b)/2 - x_i| \leq \sum_{i=1}^d \frac{L(b-a)}{2} = \frac{dL(b-a)}{2}. \end{aligned} \quad (4.109)$$

This and the fact that $\|\mathcal{T}(\mathbf{F})\|_\infty = |f((a+b)/2, (a+b)/2, \dots, (a+b)/2)| \leq \sup_{x \in [a, b]^d} |f(x)|$ complete the proof of Lemma 4.3.8. \square

Proposition 4.3.9. Let $d \in \mathbb{N}$, $L, a \in \mathbb{R}$, $b \in (a, \infty)$, $r \in (0, d/4)$, let $f: [a, b]^d \rightarrow \mathbb{R}$ and $\delta: [a, b]^d \times [a, b]^d \rightarrow \mathbb{R}$ 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 $K \in \mathbb{N}$, $\mathfrak{x}_1, \mathfrak{x}_2, \dots, \mathfrak{x}_K \in [a, b]^d$, $\mathfrak{y} \in \mathbb{R}^K$, $\mathbf{F} \in \mathbf{N}$ satisfy $K = \mathcal{C}([a, b]^d, \delta), (b-a)r$, $\sup_{x \in [a, b]^d} [\min_{k \in \{1, 2, \dots, K\}} \delta(x, \mathfrak{x}_k)] \leq (b-a)r$, $\mathfrak{y} = (f(\mathfrak{x}_1), f(\mathfrak{x}_2), \dots, f(\mathfrak{x}_K))$, and

$$\mathbf{F} = \mathbb{M}_K \bullet \mathbf{A}_{-L\mathbb{I}_K, \mathfrak{y}} \bullet \mathbf{P}_K (\mathbb{L}_d \bullet \mathbf{A}_{\mathbb{I}_d, -\mathfrak{x}_1}, \mathbb{L}_d \bullet \mathbf{A}_{\mathbb{I}_d, -\mathfrak{x}_2}, \dots, \mathbb{L}_d \bullet \mathbf{A}_{\mathbb{I}_d, -\mathfrak{x}_K}) \bullet \mathbb{T}_{d, K} \quad (4.110)$$

(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, 4.2.7, 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 $\mathcal{H}(\mathbf{F}) \leq \lceil d \log_2\left(\frac{3d}{4r}\right) \rceil + 1$,
 - (iv) it holds that $\mathbb{D}_1(\mathbf{F}) \leq 2d\left(\frac{3d}{4r}\right)^d$,
 - (v) it holds for all $i \in \{2, 3, 4, \dots\}$ that $\mathbb{D}_i(\mathbf{F}) \leq 3\left(\frac{3d}{4r}\right)^d \frac{1}{2^{i-1}}$,
 - (vi) it holds that $\mathcal{P}(\mathbf{F}) \leq 35\left(\frac{3d}{4r}\right)^{2d} d^2$,
 - (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 $\sup_{x \in [a,b]^d} |(\mathcal{R}_{\mathfrak{r}}^N(\mathbf{F}))(x) - f(x)| \leq 2L(b-a)r$
- (cf. Definitions 1.2.4, 1.3.4, 1.3.6, and 4.2.8).

Proof of Proposition 4.3.9. 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 (4.110), Lemma 4.2.11, 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 $i \in \{2, 3, 4, \dots\}$ that $\mathbb{D}_i(\mathbf{F}) \leq 3\left(\frac{K}{2^{i-1}}\right)$,
- (VI) it holds that $\|\mathcal{T}(\mathbf{F})\|_\infty \leq \max\{1, L, \max_{k \in \{1, 2, \dots, K\}} \|\mathfrak{x}_k\|_\infty, 2[\max_{k \in \{1, 2, \dots, K\}} |f(\mathfrak{x}_k)|]\}$, and
- (VII) it holds that $\sup_{x \in [a,b]^d} |(\mathcal{R}_{\mathfrak{r}}^N(\mathbf{F}))(x) - f(x)| \leq 2L\left[\sup_{x \in [a,b]^d} (\min_{k \in \{1, 2, \dots, K\}} \delta(x, \mathfrak{x}_k))\right]$

(cf. Definitions 1.2.4, 1.3.4, 1.3.6, and 4.2.8). Note that items (I) and (II) establish items (i) and (ii). Next observe that Lemma 4.3.7 and the fact that $\frac{d}{2r} \geq 2$ imply that

$$K = \mathcal{C}^{([a,b]^d, \delta), (b-a)r} \leq \left(\left\lceil \frac{d(b-a)}{2(b-a)r} \right\rceil \right)^d = \left(\left\lceil \frac{d}{2r} \right\rceil \right)^d \leq \left(\frac{3}{2} \left(\frac{d}{2r} \right) \right)^d = \left(\frac{3d}{4r} \right)^d. \quad (4.111)$$

Combining this with item (III) assures that

$$\mathcal{H}(\mathbf{F}) = \lceil \log_2(K) \rceil + 1 \leq \left\lceil \log_2 \left(\left(\frac{3d}{4r} \right)^d \right) \right\rceil + 1 = \lceil d \log_2 \left(\frac{3d}{4r} \right) \rceil + 1. \quad (4.112)$$

This establishes item (iii). Moreover, note that (4.111) and item (IV) imply that

$$\mathbb{D}_1(\mathbf{F}) = 2dK \leq 2d \left(\frac{3d}{4r} \right)^d. \quad (4.113)$$

This establishes item (iv). In addition, observe that item (V) and (4.111) 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

$$\frac{K}{2^{i-1}} \geq \frac{K}{2^{\mathcal{H}(\mathbf{F})-1}} = \frac{K}{2^{\lceil \log_2(K) \rceil}} \geq \frac{K}{2^{\log_2(K)+1}} = \frac{K}{2K} = \frac{1}{2}. \quad (4.114)$$

Item (V) and (4.111) hence show that for all $i \in \mathbb{N} \cap (1, \mathcal{H}(\mathbf{F}))$ it holds that

$$\mathbb{D}_i(\mathbf{F}) \leq 3 \lceil \frac{K}{2^{i-1}} \rceil \leq \frac{3K}{2^{i-2}} \leq \left(\frac{3d}{4r} \right)^d \frac{3}{2^{i-2}}. \quad (4.115)$$

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) prove that

$$\begin{aligned} \|\mathcal{T}(\mathbf{F})\|_\infty &\leq \max\{1, L, \max_{k \in \{1, 2, \dots, K\}} \|\mathfrak{x}_k\|_\infty, 2[\max_{k \in \{1, 2, \dots, K\}} |f(\mathfrak{x}_k)|]\} \\ &\leq \max\{1, L, |a|, |b|, 2[\sup_{x \in [a, b]^d} |f(x)|]\}. \end{aligned} \quad (4.116)$$

This establishes item (vii). Moreover, observe that the assumption that

$$\sup_{x \in [a, b]^d} [\min_{k \in \{1, 2, \dots, K\}} \delta(x, \mathfrak{x}_k)] \leq (b - a)r \quad (4.117)$$

and item (VII) demonstrate that

$$\sup_{x \in [a, b]^d} |(\mathcal{R}_{\mathfrak{r}}^N(\mathbf{F}))(x) - f(x)| \leq 2L [\sup_{x \in [a, b]^d} (\min_{k \in \{1, 2, \dots, K\}} \delta(x, \mathfrak{x}_k))] \leq 2L(b - a)r. \quad (4.118)$$

This establishes item (viii). It thus remains to prove item (vi). For this note that items (I) and (II), (4.113), and (4.115) show that

$$\begin{aligned} \mathcal{P}(\mathbf{F}) &= \sum_{i=1}^{\mathcal{L}(\mathbf{F})} \mathbb{D}_i(\mathbf{F})(\mathbb{D}_{i-1}(\mathbf{F}) + 1) \\ &\leq 2d \left(\frac{3d}{4r} \right)^d (d+1) + \left(\frac{3d}{4r} \right)^d 3 \left(2d \left(\frac{3d}{4r} \right)^d + 1 \right) \\ &\quad + \left[\sum_{i=3}^{\mathcal{L}(\mathbf{F})-1} \left(\frac{3d}{4r} \right)^d \frac{3}{2^{i-2}} \left(\left(\frac{3d}{4r} \right)^d \frac{3}{2^{i-3}} + 1 \right) \right] + \left(\frac{3d}{4r} \right)^d \frac{3}{2^{\mathcal{L}(\mathbf{F})-3}} + 1. \end{aligned} \quad (4.119)$$

Next note that the fact that $\frac{3d}{4r} \geq 3$ ensures that

$$\begin{aligned} & 2d\left(\frac{3d}{4r}\right)^d(d+1) + \left(\frac{3d}{4r}\right)^d 3\left(2d\left(\frac{3d}{4r}\right)^d + 1\right) + \left(\frac{3d}{4r}\right)^d \frac{3}{2^{\mathcal{L}(\mathbf{F})-3}} + 1 \\ & \leq \left(\frac{3d}{4r}\right)^{2d} (2d(d+1) + 3(2d+1) + \frac{3}{2^{1-3}} + 1) \\ & \leq \left(\frac{3d}{4r}\right)^{2d} d^2 (4 + 9 + 12 + 1) = 26\left(\frac{3d}{4r}\right)^{2d} d^2. \end{aligned} \quad (4.120)$$

Moreover, observe that the fact that $\frac{3d}{4r} \geq 3$ implies that

$$\begin{aligned} \sum_{i=3}^{\mathcal{L}(\mathbf{F})-1} \left(\frac{3d}{4r}\right)^d \frac{3}{2^{i-2}} \left(\left(\frac{3d}{4r}\right)^d \frac{3}{2^{i-3}} + 1\right) & \leq \left(\frac{3d}{4r}\right)^{2d} \sum_{i=3}^{\mathcal{L}(\mathbf{F})-1} \frac{3}{2^{i-2}} \left(\frac{3}{2^{i-3}} + 1\right) \\ & = \left(\frac{3d}{4r}\right)^{2d} \sum_{i=3}^{\mathcal{L}(\mathbf{F})-1} \left[\frac{9}{2^{2i-5}} + \frac{3}{2^{i-2}}\right] \\ & = \left(\frac{3d}{4r}\right)^{2d} \sum_{i=0}^{\mathcal{L}(\mathbf{F})-4} \left[\frac{9}{2}(4^{-i}) + \frac{3}{2}(2^{-i})\right] \\ & \leq \left(\frac{3d}{4r}\right)^{2d} \left(\frac{9}{2}\left(\frac{1}{1-4^{-1}}\right) + \frac{3}{2}\left(\frac{1}{1-2^{-1}}\right)\right) = 9\left(\frac{3d}{4r}\right)^{2d}. \end{aligned} \quad (4.121)$$

Combining this, (4.119), and (4.120) demonstrates that

$$\mathcal{P}(\mathbf{F}) \leq 26\left(\frac{3d}{4r}\right)^{2d} d^2 + 9\left(\frac{3d}{4r}\right)^{2d} \leq 35\left(\frac{3d}{4r}\right)^{2d} d^2. \quad (4.122)$$

This establishes item (vi). The proof of Proposition 4.3.9 is thus complete. \square

Proposition 4.3.10. Let $d \in \mathbb{N}$, $L, a \in \mathbb{R}$, $b \in (a, \infty)$, $r \in (0, \infty)$ and let $f: [a, b]^d \rightarrow \mathbb{R}$ 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 $\mathcal{H}(\mathbf{F}) \leq (\lceil d \log_2(\frac{3d}{4r}) \rceil + 1) \mathbb{1}_{(0, d/4)}(r)$,
- (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 $i \in \{2, 3, 4, \dots\}$ that $\mathbb{D}_i(\mathbf{F}) \leq 3\lceil\left(\frac{3d}{4r}\right)^d \frac{1}{2^{i-1}}\rceil$,
- (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 $\sup_{x \in [a,b]^d} |(\mathcal{R}_{\mathbf{r}}^{\mathbf{N}}(\mathbf{F}))(x) - f(x)| \leq 2L(b-a)r$
 (cf. Definitions 1.2.4, 1.3.1, 1.3.4, 1.3.6, and 4.2.8).

Proof of Proposition 4.3.10. Throughout this proof, assume without loss of generality that $r < d/4$ (cf. Lemma 4.3.8), let $\delta: [a,b]^d \times [a,b]^d \rightarrow \mathbb{R}$ satisfy for all $x, y \in [a,b]^d$ that

$$\delta(x, y) = \|x - y\|_1, \quad (4.123)$$

and let $K \in \mathbb{N} \cup \{\infty\}$ satisfy

$$K = \mathcal{C}^{([a,b]^d, \delta), (b-a)r}. \quad (4.124)$$

Note that Lemma 4.3.7 assures that $K < \infty$. This and (4.93) ensure that there exist $\mathfrak{x}_1, \mathfrak{x}_2, \dots, \mathfrak{x}_K \in [a,b]^d$ such that

$$\sup_{x \in [a,b]^d} [\min_{k \in \{1, 2, \dots, K\}} \delta(x, \mathfrak{x}_k)] \leq (b-a)r. \quad (4.125)$$

Combining this with Proposition 4.3.9 establishes items (i), (ii), (iii), (iv), (v), (vi), (vii), and (viii). The proof of Proposition 4.3.10 is thus complete. \square

Proposition 4.3.11 (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 $f: [a, b]^d \rightarrow \mathbb{R}$ satisfy for all $x, y \in [a, b]^d$ that*

$$|f(x) - f(y)| \leq L\|x - y\|_1 \quad (4.126)$$

(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 $\mathcal{H}(\mathbf{F}) \leq d \left(\log_2 \left(\max \left\{ \frac{3dL(b-a)}{2}, 1 \right\} \right) + \log_2(\varepsilon^{-1}) \right) + 2$,
- (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 $i \in \{2, 3, 4, \dots\}$ that $\mathbb{D}_i(\mathbf{F}) \leq \varepsilon^{-d} 3 \left(\frac{(3dL(b-a))^d}{2^i} + 1 \right)$,
- (vi) it holds that $\mathcal{P}(\mathbf{F}) \leq \varepsilon^{-2d} 9 (3d \max\{L(b-a), 1\})^{2d} d^2$,
- (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 $\sup_{x \in [a,b]^d} |(\mathcal{R}_{\mathbf{r}}^{\mathbf{N}}(\mathbf{F}))(x) - f(x)| \leq \varepsilon$
 (cf. Definitions 1.2.4, 1.3.1, 1.3.4, and 1.3.6).

Proof of Proposition 4.3.11. Throughout this proof, assume without loss of generality that

$$L(b-a) \neq 0. \quad (4.127)$$

Observe that (4.127) 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

$$|f(x) - f(y)| \leq L\|x - y\|_1, \quad (4.128)$$

ensures that $L > 0$. Proposition 4.3.10 therefore 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 $\mathcal{H}(\mathbf{F}) \leq (\lceil d \log_2(\frac{3dL(b-a)}{2\varepsilon}) \rceil + 1) \mathbb{1}_{(0,d/4)}(\frac{\varepsilon}{2L(b-a)})$,
- (IV) it holds that $\mathbb{D}_1(\mathbf{F}) \leq 2d(\frac{3dL(b-a)}{2\varepsilon})^d \mathbb{1}_{(0,d/4)}(\frac{\varepsilon}{2L(b-a)}) + \mathbb{1}_{[d/4,\infty)}(\frac{\varepsilon}{2L(b-a)})$,
- (V) it holds for all $i \in \{2, 3, 4, \dots\}$ that $\mathbb{D}_i(\mathbf{F}) \leq 3 \lceil (\frac{3dL(b-a)}{2\varepsilon})^d \frac{1}{2^{i-1}} \rceil$,
- (VI) it holds that $\mathcal{P}(\mathbf{F}) \leq 35(\frac{3dL(b-a)}{2\varepsilon})^{2d} d^2 \mathbb{1}_{(0,d/4)}(\frac{\varepsilon}{2L(b-a)}) + (d+1) \mathbb{1}_{[d/4,\infty)}(\frac{\varepsilon}{2L(b-a)})$,
- (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 $\sup_{x \in [a,b]^d} |(\mathcal{R}_{\mathbf{r}}^{\mathbf{N}}(\mathbf{F}))(x) - f(x)| \leq \varepsilon$

(cf. Definitions 1.2.4, 1.3.1, 1.3.4, 1.3.6, and 4.2.8). Note that item (III) assures that

$$\begin{aligned} \mathcal{H}(\mathbf{F}) &\leq (d(\log_2(\frac{3dL(b-a)}{2}) + \log_2(\varepsilon^{-1})) + 2) \mathbb{1}_{(0,d/4)}(\frac{\varepsilon}{2L(b-a)}) \\ &\leq d(\max\{\log_2(\frac{3dL(b-a)}{2}), 0\} + \log_2(\varepsilon^{-1})) + 2. \end{aligned} \quad (4.129)$$

Furthermore, observe that item (IV) ensures that

$$\begin{aligned} \mathbb{D}_1(\mathbf{F}) &\leq d(\frac{3d \max\{L(b-a), 1\}}{\varepsilon})^d \mathbb{1}_{(0,d/4)}(\frac{\varepsilon}{2L(b-a)}) + \mathbb{1}_{[d/4,\infty)}(\frac{\varepsilon}{2L(b-a)}) \\ &\leq \varepsilon^{-d} d(3d \max\{L(b-a), 1\})^d. \end{aligned} \quad (4.130)$$

Moreover, note that item (V) establishes that for all $i \in \{2, 3, 4, \dots\}$ it holds that

$$\mathbb{D}_i(\mathbf{F}) \leq 3 \left(\left(\frac{3dL(b-a)}{2\varepsilon} \right)^d \frac{1}{2^{i-1}} + 1 \right) \leq \varepsilon^{-d} 3 \left(\frac{(3dL(b-a))^d}{2^i} + 1 \right). \quad (4.131)$$

In addition, observe that item (VI) ensures that

$$\begin{aligned}\mathcal{P}(\mathbf{F}) &\leq 9\left(\frac{3d \max\{L(b-a), 1\}}{\varepsilon}\right)^{2d} d^2 \mathbb{1}_{(0, d/4)}\left(\frac{\varepsilon}{2L(b-a)}\right) + (d+1)\mathbb{1}_{[d/4, \infty)}\left(\frac{\varepsilon}{2L(b-a)}\right) \\ &\leq \varepsilon^{-2d} 9(3d \max\{L(b-a), 1\})^{2d} d^2.\end{aligned}\quad (4.132)$$

Combining this, (4.129), (4.130), and (4.131) with items (I), (II), (VII), and (VIII) establishes items (i), (ii), (iii), (iv), (v), (vi), (vii), and (viii). The proof of Proposition 4.3.11 is thus complete. \square

Corollary 4.3.12 (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 $f: [a, b]^d \rightarrow \mathbb{R}$ satisfy for all $x, y \in [a, b]^d$ that*

$$|f(x) - f(y)| \leq L\|x - y\|_1 \quad (4.133)$$

(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

$$\mathcal{H}(\mathbf{F}) \leq \mathfrak{C}(\log_2(\varepsilon^{-1}) + 1), \quad \|\mathcal{T}(\mathbf{F})\|_\infty \leq \max\{1, L, |a|, |b|, 2[\sup_{x \in [a, b]^d} |f(x)|]\}, \quad (4.134)$$

$$\mathcal{R}_r^N(\mathbf{F}) \in C(\mathbb{R}^d, \mathbb{R}), \quad \sup_{x \in [a, b]^d} |(\mathcal{R}_r^N(\mathbf{F}))(x) - f(x)| \leq \varepsilon, \quad \text{and} \quad \mathcal{P}(\mathbf{F}) \leq \mathfrak{C}\varepsilon^{-2d} \quad (4.135)$$

(cf. Definitions 1.2.4, 1.3.1, 1.3.4, and 1.3.6).

Proof of Corollary 4.3.12. Throughout this proof, let $\mathfrak{C} \in \mathbb{R}$ satisfy

$$\mathfrak{C} = 9(3d \max\{L(b-a), 1\})^{2d} d^2. \quad (4.136)$$

Note that items (i), (ii), (iii), (vi), (vii), and (viii) in Proposition 4.3.11 and the fact that for all $\varepsilon \in (0, 1]$ it holds that

$$\begin{aligned}d\left(\log_2\left(\max\left\{\frac{3dL(b-a)}{2}, 1\right\}\right) + \log_2(\varepsilon^{-1})\right) + 2 &\leq d\left(\max\left\{\frac{3dL(b-a)}{2}, 1\right\} + \log_2(\varepsilon^{-1})\right) + 2 \\ &\leq d \max\{3dL(b-a), 1\} + 2 + d \log_2(\varepsilon^{-1}) \\ &\leq \mathfrak{C}(\log_2(\varepsilon^{-1}) + 1)\end{aligned}\quad (4.137)$$

imply that for every $\varepsilon \in (0, 1]$ there exists $\mathbf{F} \in \mathbf{N}$ such that

$$\mathcal{H}(\mathbf{F}) \leq \mathfrak{C}(\log_2(\varepsilon^{-1}) + 1), \quad \|\mathcal{T}(\mathbf{F})\|_\infty \leq \max\{1, L, |a|, |b|, 2[\sup_{x \in [a, b]^d} |f(x)|]\}, \quad (4.138)$$

$$\mathcal{R}_r^N(\mathbf{F}) \in C(\mathbb{R}^d, \mathbb{R}), \quad \sup_{x \in [a, b]^d} |(\mathcal{R}_r^N(\mathbf{F}))(x) - f(x)| \leq \varepsilon, \quad \text{and} \quad \mathcal{P}(\mathbf{F}) \leq \mathfrak{C}\varepsilon^{-2d} \quad (4.139)$$

(cf. Definitions 1.2.4, 1.3.1, 1.3.4, and 1.3.6). The proof of Corollary 4.3.12 is thus complete. \square

Lemma 4.3.13 (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*

$$\|x\|_p \geq \|x\|_q \quad (4.140)$$

(cf. Definition 3.3.4).

Proof of Lemma 4.3.13. Throughout this proof, assume without loss of generality that $q < \infty$, let $e_1, e_2, \dots, e_d \in \mathbb{R}^d$ satisfy $e_1 = (1, 0, \dots, 0)$, $e_2 = (0, 1, 0, \dots, 0)$, \dots , $e_d = (0, \dots, 0, 1)$, let $r \in \mathbb{R}$ satisfy

$$r = p^{-1}q, \quad (4.141)$$

and let $x = (x_1, \dots, x_d)$, $y = (y_1, \dots, y_d) \in \mathbb{R}^d$ satisfy for all $i \in \{1, 2, \dots, d\}$ that

$$y_i = |x_i|^p. \quad (4.142)$$

Observe that (4.142), the fact that

$$y = \sum_{i=1}^d y_i e_i, \quad (4.143)$$

and the fact that for all $v, w \in \mathbb{R}^d$ it holds that

$$\|v + w\|_r \leq \|v\|_r + \|w\|_r \quad (4.144)$$

(cf. Definition 3.3.4) ensures that

$$\begin{aligned} \|x\|_q &= \left[\sum_{i=1}^d |x_i|^q \right]^{1/q} = \left[\sum_{i=1}^d |x_i|^{pr} \right]^{1/q} = \left[\sum_{i=1}^d |y_i|^r \right]^{1/q} = \left[\sum_{i=1}^d |y_i|^r \right]^{1/(pr)} = \|y\|_r^{1/p} \\ &= \left\| \sum_{i=1}^d y_i e_i \right\|_r^{1/p} \leq \left[\sum_{i=1}^d \|y_i e_i\|_r \right]^{1/p} = \left[\sum_{i=1}^d |y_i| \|e_i\|_r \right]^{1/p} = \left[\sum_{i=1}^d |y_i| \right]^{1/p} \\ &= \|y\|_1^{1/p} = \|x\|_p. \end{aligned} \quad (4.145)$$

This establishes (4.140). The proof of Lemma 4.3.13 is thus complete. □

Corollary 4.3.14 (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 $f: [a, b]^d \rightarrow \mathbb{R}$ satisfy for all $x, y \in [a, b]^d$ that*

$$|f(x) - f(y)| \leq L\|x - y\|_1 \quad (4.146)$$

(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

$$\mathcal{R}_{\tau}^{\mathbf{N}}(\mathbf{F}) \in C(\mathbb{R}^d, \mathbb{R}), \quad \sup_{x \in [a, b]^d} |(\mathcal{R}_{\tau}^{\mathbf{N}}(\mathbf{F}))(x) - f(x)| \leq \varepsilon, \quad \text{and} \quad \mathcal{P}(\mathbf{F}) \leq \mathfrak{C}\varepsilon^{-2d} \quad (4.147)$$

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

Proof of Corollary 4.3.14. Note that Corollary 4.3.12 establishes (4.147). The proof of Corollary 4.3.14 is thus complete. \square

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 example, Corollaries 4.3.12 and 4.3.14 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 $L, \mathfrak{d} \in \mathbb{N}$, $u \in [-\infty, \infty)$, $v \in (u, \infty]$, $\mathbf{l} = (l_0, l_1, \dots, l_L) \in \mathbb{N}^{L+1}$, $\theta \in \mathbb{R}^{\mathfrak{d}}$ satisfy

$$\mathfrak{d} \geq \sum_{k=1}^L l_k(l_{k-1} + 1). \quad (4.148)$$

4.4. Refined ANN approximations results for multi-dimensional functions

Then we denote by $\mathcal{N}_{u,v}^{\theta,1}: \mathbb{R}^{l_0} \rightarrow \mathbb{R}^{l_L}$ the function which satisfies for all $x \in \mathbb{R}^{l_0}$ that

$$\mathcal{N}_{u,v}^{\theta,1}(x) = \begin{cases} (\mathcal{N}_{\mathfrak{C}_{u,v,l_L}}^{\theta,l_0})(x) & : L = 1 \\ (\mathcal{N}_{\mathfrak{R}_{l_1}, \mathfrak{R}_{l_2}, \dots, \mathfrak{R}_{l_{L-1}}, \mathfrak{C}_{u,v,l_L}}^{\theta,l_0})(x) & : L > 1 \end{cases} \quad (4.149)$$

(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

$$(\mathcal{N}_{-\infty, \infty}^{\mathcal{T}(\Phi), \mathcal{D}(\Phi)})(x) = (\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\Phi))(x) \quad (4.150)$$

(cf. Definitions 1.2.4, 1.3.4, 1.3.6, and 4.4.1).

Proof of Lemma 4.4.2. Observe that Proposition 1.3.10, (4.149), (1.27), and the fact that for all $d \in \mathbb{N}$ it holds that $\mathfrak{C}_{-\infty, \infty, d} = \text{id}_{\mathbb{R}^d}$ imply (4.150) (cf. Definition 1.2.10). The proof of Lemma 4.4.2 is thus complete. \square

4.4.2 Embedding ANNs in larger architectures

Lemma 4.4.3. Let $a \in C(\mathbb{R}, \mathbb{R})$, $L \in \mathbb{N}$, $l_0, l_1, \dots, l_L, \mathfrak{l}_1, \dots, \mathfrak{l}_L \in \mathbb{N}$ satisfy for all $k \in \{1, 2, \dots, L\}$ that $\mathfrak{l}_0 = l_0$, $\mathfrak{l}_L = l_L$, and $\mathfrak{l}_k \geq l_k$, for every $k \in \{1, 2, \dots, L\}$ let $W_k = (W_{k,i,j})_{(i,j) \in \{1, 2, \dots, l_k\} \times \{1, 2, \dots, l_{k-1}\}} \in \mathbb{R}^{l_k \times l_{k-1}}$, $\mathcal{W}_k = (\mathcal{W}_{k,i,j})_{(i,j) \in \{1, 2, \dots, \mathfrak{l}_k\} \times \{1, 2, \dots, \mathfrak{l}_{k-1}\}} \in \mathbb{R}^{\mathfrak{l}_k \times \mathfrak{l}_{k-1}}$, $B_k = (B_{k,i})_{i \in \{1, 2, \dots, l_k\}} \in \mathbb{R}^{l_k}$, $\mathcal{B}_k = (\mathcal{B}_{k,i})_{i \in \{1, 2, \dots, \mathfrak{l}_k\}} \in \mathbb{R}^{\mathfrak{l}_k}$, assume for all $k \in \{1, 2, \dots, L\}$, $i \in \{1, 2, \dots, l_k\}$, $j \in \mathbb{N} \cap (0, l_{k-1}]$ that

$$\mathcal{W}_{k,i,j} = W_{k,i,j} \quad \text{and} \quad \mathcal{B}_{k,i} = B_{k,i}, \quad (4.151)$$

and assume for all $k \in \{1, 2, \dots, L\}$, $i \in \{1, 2, \dots, l_k\}$, $j \in \mathbb{N} \cap (l_{k-1}, \mathfrak{l}_{k-1} + 1)$ that $\mathcal{W}_{k,i,j} = 0$. Then

$$\mathcal{R}_a^{\mathbf{N}}(((W_1, B_1), (W_2, B_2), \dots, (W_L, B_L))) = \mathcal{R}_a^{\mathbf{N}}(((\mathcal{W}_1, \mathcal{B}_1), (\mathcal{W}_2, \mathcal{B}_2), \dots, (\mathcal{W}_L, \mathcal{B}_L))) \quad (4.152)$$

(cf. Definition 1.3.4).

Proof of Lemma 4.4.3. Throughout this proof, let $\pi_k: \mathbb{R}^{\mathfrak{l}_k} \rightarrow \mathbb{R}^{l_k}$, $k \in \{0, 1, \dots, L\}$, satisfy for all $k \in \{0, 1, \dots, L\}$, $x = (x_1, x_2, \dots, x_{\mathfrak{l}_k})$ that

$$\pi_k(x) = (x_1, x_2, \dots, x_{l_k}). \quad (4.153)$$

Note that the assumption that $\mathfrak{l}_0 = l_0$ and $\mathfrak{l}_L = l_L$ proves that

$$\mathcal{R}_a^{\mathbf{N}}(((W_1, B_1), (W_2, B_2), \dots, (W_L, B_L))) \in C(\mathbb{R}^{\mathfrak{l}_0}, \mathbb{R}^{\mathfrak{l}_L}) \quad (4.154)$$

(cf. Definition 1.3.4). Furthermore, observe that the assumption that for all $k \in \{1, 2, \dots, l\}$, $i \in \{1, 2, \dots, l_k\}$, $j \in \mathbb{N} \cap (l_{k-1}, l_{k-1} + 1)$ it holds that $\mathcal{W}_{k,i,j} = 0$ shows that for all $k \in \{1, 2, \dots, L\}$, $x = (x_1, \dots, x_{l_{k-1}}) \in \mathbb{R}^{l_{k-1}}$ it holds that

$$\begin{aligned} & \pi_k(\mathcal{W}_k x + \mathcal{B}_k) \\ &= \left(\left[\sum_{i=1}^{l_{k-1}} \mathcal{W}_{k,1,i} x_i \right] + \mathcal{B}_{k,1}, \left[\sum_{i=1}^{l_{k-1}} \mathcal{W}_{k,2,i} x_i \right] + \mathcal{B}_{k,2}, \dots, \left[\sum_{i=1}^{l_{k-1}} \mathcal{W}_{k,l_k,i} x_i \right] + \mathcal{B}_{k,l_k} \right) \quad (4.155) \\ &= \left(\left[\sum_{i=1}^{l_{k-1}} \mathcal{W}_{k,1,i} x_i \right] + \mathcal{B}_{k,1}, \left[\sum_{i=1}^{l_{k-1}} \mathcal{W}_{k,2,i} x_i \right] + \mathcal{B}_{k,2}, \dots, \left[\sum_{i=1}^{l_{k-1}} \mathcal{W}_{k,l_k,i} x_i \right] + \mathcal{B}_{k,l_k} \right). \end{aligned}$$

Combining this with the assumption that for all $k \in \{1, 2, \dots, L\}$, $i \in \{1, 2, \dots, l_k\}$, $j \in \mathbb{N} \cap (0, l_{k-1}]$ it holds that $\mathcal{W}_{k,i,j} = W_{k,i,j}$ and $\mathcal{B}_{k,i} = B_{k,i}$ demonstrates that for all $k \in \{1, 2, \dots, L\}$, $x = (x_1, \dots, x_{l_{k-1}}) \in \mathbb{R}^{l_{k-1}}$ it holds that

$$\begin{aligned} & \pi_k(\mathcal{W}_k x + \mathcal{B}_k) \\ &= \left(\left[\sum_{i=1}^{l_{k-1}} W_{k,1,i} x_i \right] + B_{k,1}, \left[\sum_{i=1}^{l_{k-1}} W_{k,2,i} x_i \right] + B_{k,2}, \dots, \left[\sum_{i=1}^{l_{k-1}} W_{k,l_k,i} x_i \right] + B_{k,l_k} \right) \quad (4.156) \\ &= W_k \pi_{k-1}(x) + B_k. \end{aligned}$$

Hence, we obtain that for all $x_0 \in \mathbb{R}^{l_0}$, $x_1 \in \mathbb{R}^{l_1}, \dots, x_{L-1} \in \mathbb{R}^{l_{L-1}}$, $k \in \mathbb{N} \cap (0, L)$ with $\forall m \in \mathbb{N} \cap (0, L) : x_m = \mathfrak{M}_{a,l_m}(\mathcal{W}_m x_{m-1} + \mathcal{B}_m)$ it holds that

$$\pi_k(x_k) = \mathfrak{M}_{a,l_k}(\pi_k(\mathcal{W}_k x_{k-1} + \mathcal{B}_k)) = \mathfrak{M}_{a,l_k}(W_k \pi_{k-1}(x_{k-1}) + B_k) \quad (4.157)$$

(cf. Definition 1.2.1). Induction, the assumption that $l_0 = l_0$ and $l_L = l_L$, and (4.156) therefore ensure that for all $x_0 \in \mathbb{R}^{l_0}$, $x_1 \in \mathbb{R}^{l_1}, \dots, x_{L-1} \in \mathbb{R}^{l_{L-1}}$ with $\forall k \in \mathbb{N} \cap (0, L) : x_k = \mathfrak{M}_{a,l_k}(\mathcal{W}_k x_{k-1} + \mathcal{B}_k)$ it holds that

$$\begin{aligned} & (\mathcal{R}_a^{\mathbf{N}}(((W_1, B_1), (W_2, B_2), \dots, (W_L, B_L))))(x_0) \\ &= (\mathcal{R}_a^{\mathbf{N}}(((W_1, B_1), (W_2, B_2), \dots, (W_L, B_L))))(\pi_0(x_0)) \\ &= W_L \pi_{L-1}(x_{L-1}) + B_L \quad (4.158) \\ &= \pi_L(\mathcal{W}_L x_{L-1} + \mathcal{B}_L) = \mathcal{W}_L x_{L-1} + \mathcal{B}_L \\ &= (\mathcal{R}_a^{\mathbf{N}}(((\mathcal{W}_1, \mathcal{B}_1), (\mathcal{W}_2, \mathcal{B}_2), \dots, (\mathcal{W}_L, \mathcal{B}_L))))(x_0). \end{aligned}$$

The proof of Lemma 4.4.3 is thus complete. \square

Lemma 4.4.4. Let $a \in C(\mathbb{R}, \mathbb{R})$, $L \in \mathbb{N}$, $l_0, l_1, \dots, l_L, \mathfrak{l}_0, \mathfrak{l}_1, \dots, \mathfrak{l}_L \in \mathbb{N}$ satisfy for all $k \in \{1, 2, \dots, L\}$ that

$$\mathfrak{l}_0 = l_0, \quad \mathfrak{l}_L = l_L, \quad \text{and} \quad \mathfrak{l}_k \geq l_k \quad (4.159)$$

and let $\Phi \in \mathbf{N}$ satisfy $\mathcal{D}(\Phi) = (l_0, l_1, \dots, l_L)$ (cf. Definition 1.3.1). Then there exists $\Psi \in \mathbf{N}$ such that

$$\mathcal{D}(\Psi) = (\mathfrak{l}_0, \mathfrak{l}_1, \dots, \mathfrak{l}_L), \quad \|\mathcal{T}(\Psi)\|_\infty = \|\mathcal{T}(\Phi)\|_\infty, \quad \text{and} \quad \mathcal{R}_a^{\mathbf{N}}(\Psi) = \mathcal{R}_a^{\mathbf{N}}(\Phi) \quad (4.160)$$

(cf. Definitions 1.3.4, 1.3.6, and 3.3.4).

Proof of Lemma 4.4.4. Throughout this proof, let $B_k = (B_{k,i})_{i \in \{1, 2, \dots, l_k\}} \in \mathbb{R}^{l_k}$, $k \in \{1, 2, \dots, L\}$, and $W_k = (W_{k,i,j})_{(i,j) \in \{1, 2, \dots, l_k\} \times \{1, 2, \dots, l_{k-1}\}} \in \mathbb{R}^{l_k \times l_{k-1}}$, $k \in \{1, 2, \dots, L\}$, satisfy

$$\Phi = ((W_1, B_1), (W_2, B_2), \dots, (W_L, B_L)) \quad (4.161)$$

and let $\mathfrak{W}_k = (\mathfrak{W}_{k,i,j})_{(i,j) \in \{1, 2, \dots, \mathfrak{l}_k\} \times \{1, 2, \dots, \mathfrak{l}_{k-1}\}} \in \mathbb{R}^{\mathfrak{l}_k \times \mathfrak{l}_{k-1}}$, $k \in \{1, 2, \dots, L\}$, and $\mathfrak{B}_k = (\mathfrak{B}_{k,i})_{i \in \{1, 2, \dots, \mathfrak{l}_k\}} \in \mathbb{R}^{\mathfrak{l}_k}$, $k \in \{1, 2, \dots, L\}$, satisfy for all $k \in \{1, 2, \dots, L\}$, $i \in \{1, 2, \dots, \mathfrak{l}_k\}$, $j \in \{1, 2, \dots, \mathfrak{l}_{k-1}\}$ that

$$\mathfrak{W}_{k,i,j} = \begin{cases} W_{k,i,j} & : (i \leq l_k) \wedge (j \leq l_{k-1}) \\ 0 & : (i > l_k) \vee (j > l_{k-1}) \end{cases} \quad \text{and} \quad \mathfrak{B}_{k,i} = \begin{cases} B_{k,i} & : i \leq l_k \\ 0 & : i > l_k. \end{cases} \quad (4.162)$$

Note that (1.83) establishes that $((\mathfrak{W}_1, \mathfrak{B}_1), (\mathfrak{W}_2, \mathfrak{B}_2), \dots, (\mathfrak{W}_L, \mathfrak{B}_L)) \in (\bigtimes_{i=1}^L (\mathbb{R}^{\mathfrak{l}_i \times \mathfrak{l}_{i-1}} \times \mathbb{R}^{\mathfrak{l}_i})) \subseteq \mathbf{N}$ and

$$\mathcal{D}(((\mathfrak{W}_1, \mathfrak{B}_1), (\mathfrak{W}_2, \mathfrak{B}_2), \dots, (\mathfrak{W}_L, \mathfrak{B}_L))) = (\mathfrak{l}_0, \mathfrak{l}_1, \dots, \mathfrak{l}_L). \quad (4.163)$$

Furthermore, observe that Lemma 1.3.9 and (4.162) imply that

$$\|\mathcal{T}(((\mathfrak{W}_1, \mathfrak{B}_1), (\mathfrak{W}_2, \mathfrak{B}_2), \dots, (\mathfrak{W}_L, \mathfrak{B}_L)))\|_\infty = \|\mathcal{T}(\Phi)\|_\infty \quad (4.164)$$

(cf. Definitions 1.3.6 and 3.3.4). Moreover, note that Lemma 4.4.3 proves that

$$\begin{aligned} \mathcal{R}_a^{\mathbf{N}}(\Phi) &= \mathcal{R}_a^{\mathbf{N}}(((W_1, B_1), (W_2, B_2), \dots, (W_L, B_L))) \\ &= \mathcal{R}_a^{\mathbf{N}}(((\mathfrak{W}_1, \mathfrak{B}_1), (\mathfrak{W}_2, \mathfrak{B}_2), \dots, (\mathfrak{W}_L, \mathfrak{B}_L))) \end{aligned} \quad (4.165)$$

(cf. Definition 1.3.4). The proof of Lemma 4.4.4 is thus complete. \square

Lemma 4.4.5. Let $L, \mathfrak{L} \in \mathbb{N}$, $l_0, l_1, \dots, l_L, \mathfrak{l}_0, \mathfrak{l}_1, \dots, \mathfrak{l}_{\mathfrak{L}} \in \mathbb{N}$, $\Phi_1 = ((W_1, B_1), (W_2, B_2), \dots, (W_L, B_L)) \in (\bigtimes_{k=1}^L (\mathbb{R}^{l_k \times l_{k-1}} \times \mathbb{R}^{l_k}))$, $\Phi_2 = ((\mathfrak{W}_1, \mathfrak{B}_1), (\mathfrak{W}_2, \mathfrak{B}_2), \dots, (\mathfrak{W}_{\mathfrak{L}}, \mathfrak{B}_{\mathfrak{L}})) \in (\bigtimes_{k=1}^{\mathfrak{L}} (\mathbb{R}^{\mathfrak{l}_k \times \mathfrak{l}_{k-1}} \times \mathbb{R}^{\mathfrak{l}_k}))$. Then

$$\|\mathcal{T}(\Phi_1 \bullet \Phi_2)\|_{\infty} \leq \max\{\|\mathcal{T}(\Phi_1)\|_{\infty}, \|\mathcal{T}(\Phi_2)\|_{\infty}, \|\mathcal{T}((W_1 \mathfrak{W}_{\mathfrak{L}}, W_1 \mathfrak{B}_{\mathfrak{L}} + B_1))\|_{\infty}\} \quad (4.166)$$

(cf. Definitions 1.3.6, 2.1.1, and 3.3.4).

Proof of Lemma 4.4.5. Observe that (2.2) and Lemma 1.3.9 establish (4.166). The proof of Lemma 4.4.5 is thus complete. \square

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

$$\|\mathcal{T}(\mathcal{E}_{L, \mathfrak{J}_d}(\Phi))\|_{\infty} \leq \max\{1, \|\mathcal{T}(\Phi)\|_{\infty}\} \quad (4.167)$$

(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 $l_0, l_1, \dots, l_{L-\mathcal{L}(\Phi)+1} \in \mathbb{N}$ satisfy

$$(l_0, l_1, \dots, l_{L-\mathcal{L}(\Phi)+1}) = (d, 2d, 2d, \dots, 2d, d). \quad (4.168)$$

Note that Lemma 2.2.7 shows that $\mathcal{D}(\mathfrak{J}_d) = (d, 2d, d) \in \mathbb{N}^3$ (cf. Definition 2.2.6). Item (i) in Lemma 2.2.10 hence demonstrates that

$$\begin{aligned} \mathcal{L}((\mathfrak{J}_d)^{\bullet(L-\mathcal{L}(\Phi))}) &= L - \mathcal{L}(\Phi) + 1 \\ \text{and} \quad \mathcal{D}((\mathfrak{J}_d)^{\bullet(L-\mathcal{L}(\Phi))}) &= (l_0, l_1, \dots, l_{L-\mathcal{L}(\Phi)+1}) \in \mathbb{N}^{L-\mathcal{L}(\Phi)+2} \end{aligned} \quad (4.169)$$

(cf. Definition 2.1.1). This ensures that there exist $W_k \in \mathbb{R}^{l_k \times l_{k-1}}$, $k \in \{1, 2, \dots, L-\mathcal{L}(\Phi)+1\}$, and $B_k \in \mathbb{R}^{l_k}$, $k \in \{1, 2, \dots, L-\mathcal{L}(\Phi)+1\}$, which satisfy

$$(\mathfrak{J}_d)^{\bullet(L-\mathcal{L}(\Phi))} = ((W_1, B_1), (W_2, B_2), \dots, (W_{L-\mathcal{L}(\Phi)+1}, B_{L-\mathcal{L}(\Phi)+1})). \quad (4.170)$$

Furthermore, observe that (2.44), (2.70), (2.71), (2.2), and (2.41) imply that

$$W_1 = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ -1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & -1 \end{pmatrix} \in \mathbb{R}^{(2d) \times d} \quad (4.171)$$

$$\text{and } W_{L-\mathcal{L}(\Phi)+1} = \begin{pmatrix} 1 & -1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & -1 \end{pmatrix} \in \mathbb{R}^{d \times (2d)}.$$

Moreover, note that (2.44), (2.70), (2.71), (2.2), and (2.41) prove that for all $k \in \mathbb{N} \cap (1, L - \mathcal{L}(\Phi) + 1)$ it holds that

$$W_k = \underbrace{\begin{pmatrix} 1 & 0 & \cdots & 0 \\ -1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & -1 \end{pmatrix}}_{\in \mathbb{R}^{(2d) \times d}} \underbrace{\begin{pmatrix} 1 & -1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & -1 \end{pmatrix}}_{\in \mathbb{R}^{d \times (2d)}} \quad (4.172)$$

$$= \begin{pmatrix} 1 & -1 & 0 & 0 & \cdots & 0 & 0 \\ -1 & 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & -1 & \cdots & 0 & 0 \\ 0 & 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & -1 \\ 0 & 0 & 0 & 0 & \cdots & -1 & 1 \end{pmatrix} \in \mathbb{R}^{(2d) \times (2d)}.$$

In addition, observe 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

$$B_k = 0 \in \mathbb{R}^{2d} \quad \text{and} \quad B_{L-\mathcal{L}(\Phi)+1} = 0 \in \mathbb{R}^d. \quad (4.173)$$

Combining this, (4.171), and (4.172) shows that

$$\|\mathcal{T}((\mathfrak{I}_d)^{\bullet(L-\mathcal{L}(\Phi))})\|_\infty = 1 \quad (4.174)$$

(cf. Definitions 1.3.6 and 3.3.4). Next note that (4.171) demonstrates that for all $k \in \mathbb{N}$, $\mathfrak{W} = (w_{i,j})_{(i,j) \in \{1,2,\dots,d\} \times \{1,2,\dots,k\}} \in \mathbb{R}^{d \times k}$ it holds that

$$W_1 \mathfrak{W} = \begin{pmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,k} \\ -w_{1,1} & -w_{1,2} & \cdots & -w_{1,k} \\ w_{2,1} & w_{2,2} & \cdots & w_{2,k} \\ -w_{2,1} & -w_{2,2} & \cdots & -w_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ w_{d,1} & w_{d,2} & \cdots & w_{d,k} \\ -w_{d,1} & -w_{d,2} & \cdots & -w_{d,k} \end{pmatrix} \in \mathbb{R}^{(2d) \times k}. \quad (4.175)$$

Furthermore, observe that (4.171) and (4.173) ensure that for all $\mathfrak{B} = (b_1, b_2, \dots, b_d) \in \mathbb{R}^d$ it holds that

$$W_1 \mathfrak{B} + B_1 = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ -1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & -1 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_d \end{pmatrix} = \begin{pmatrix} b_1 \\ -b_1 \\ b_2 \\ -b_2 \\ \vdots \\ b_d \\ -b_d \end{pmatrix} \in \mathbb{R}^{2d}. \quad (4.176)$$

Combining this with (4.175) implies that for all $k \in \mathbb{N}$, $\mathfrak{W} \in \mathbb{R}^{d \times k}$, $\mathfrak{B} \in \mathbb{R}^d$ it holds that

$$\|\mathcal{T}((W_1 \mathfrak{W}, W_1 \mathfrak{B} + B_1))\|_\infty = \|\mathcal{T}((\mathfrak{W}, \mathfrak{B}))\|_\infty. \quad (4.177)$$

This, Lemma 4.4.5, and (4.174) prove that

$$\begin{aligned} \|\mathcal{T}(\mathcal{E}_{L, \mathfrak{I}_d}(\Phi))\|_\infty &= \|\mathcal{T}((\mathfrak{I}_d)^{\bullet(L-\mathcal{L}(\Phi))} \bullet \Phi)\|_\infty \\ &\leq \max\{\|\mathcal{T}((\mathfrak{I}_d)^{\bullet(L-\mathcal{L}(\Phi))})\|_\infty, \|\mathcal{T}(\Phi)\|_\infty\} = \max\{1, \|\mathcal{T}(\Phi)\|_\infty\} \end{aligned} \quad (4.178)$$

(cf. Definition 2.2.9). The proof of Lemma 4.4.6 is thus complete. \square

Lemma 4.4.7. Let $L, \mathfrak{L} \in \mathbb{N}$, $l_0, l_1, \dots, l_L, \mathfrak{l}_0, \mathfrak{l}_1, \dots, \mathfrak{l}_{\mathfrak{L}} \in \mathbb{N}$ satisfy

$$\mathfrak{L} \geq L, \quad \mathfrak{l}_0 = l_0, \quad \text{and} \quad \mathfrak{l}_{\mathfrak{L}} = l_L, \quad (4.179)$$

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 $\mathcal{D}(\Phi) = (l_0, l_1, \dots, l_L)$ (cf. Definition 1.3.1). Then there exists $\Psi \in \mathbf{N}$ such that

$$\mathcal{D}(\Psi) = (\mathfrak{l}_0, \mathfrak{l}_1, \dots, \mathfrak{l}_{\mathfrak{L}}), \quad \|\mathcal{T}(\Psi)\|_\infty \leq \max\{1, \|\mathcal{T}(\Phi)\|_\infty\}, \quad \text{and} \quad \mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\Psi) = \mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\Phi) \quad (4.180)$$

(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 $\Xi = \mathcal{E}_{\mathfrak{L}, \mathfrak{J}_{l_L}}(\Phi)$ (cf. Definitions 2.2.6 and 2.2.9). Note that item (i) in Lemma 2.2.7 establishes that $\mathcal{D}(\mathfrak{J}_{l_L}) = (l_L, 2l_L, l_L) \in \mathbb{N}^3$. Combining this with Lemma 2.2.12 shows that $\mathcal{D}(\Xi) \in \mathbb{N}^{\mathfrak{L}+1}$ and

$$\mathcal{D}(\Xi) = \begin{cases} (l_0, l_1, \dots, l_L) & : \mathfrak{L} = L \\ (l_0, l_1, \dots, l_{L-1}, 2l_L, 2l_L, \dots, 2l_L, l_L) & : \mathfrak{L} > L. \end{cases} \quad (4.181)$$

Furthermore, observe that Lemma 4.4.6 (applied with $d \curvearrowright l_L$, $L \curvearrowright \mathfrak{L}$, $\Phi \curvearrowright \Phi$ in the notation of Lemma 4.4.6) demonstrates that

$$\|\mathcal{T}(\Xi)\|_\infty \leq \max\{1, \|\mathcal{T}(\Phi)\|_\infty\} \quad (4.182)$$

(cf. Definitions 1.3.6 and 3.3.4). Moreover, note that item (ii) in Lemma 2.2.7 ensures that for all $x \in \mathbb{R}^{l_L}$ it holds that

$$(\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\mathfrak{J}_{l_L}))(x) = x \quad (4.183)$$

(cf. Definitions 1.2.4 and 1.3.4). This and item (ii) in Lemma 2.2.11 prove that

$$\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\Xi) = \mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\Phi). \quad (4.184)$$

In addition, observe that (4.181), 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 $a \curvearrowright \mathfrak{r}$, $L \curvearrowright \mathfrak{L}$, $(l_0, l_1, \dots, l_L) \curvearrowright \mathcal{D}(\Xi)$, $(\mathfrak{l}_0, \mathfrak{l}_1, \dots, \mathfrak{l}_{\mathfrak{L}}) \curvearrowright (l_0, l_1, \dots, l_L)$, $\Phi \curvearrowright \Xi$ in the notation of Lemma 4.4.4) imply that there exists $\Psi \in \mathbf{N}$ such that

$$\mathcal{D}(\Psi) = (\mathfrak{l}_0, \mathfrak{l}_1, \dots, \mathfrak{l}_{\mathfrak{L}}), \quad \|\mathcal{T}(\Psi)\|_\infty = \|\mathcal{T}(\Xi)\|_\infty, \quad \text{and} \quad \mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\Psi) = \mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\Xi). \quad (4.185)$$

Combining this with (4.182) and (4.184) proves (4.180). The proof of Lemma 4.4.7 is thus complete. \square

Lemma 4.4.8. Let $u \in [-\infty, \infty)$, $v \in (u, \infty]$, $L, \mathfrak{L}, d, \mathfrak{d} \in \mathbb{N}$, $\theta \in \mathbb{R}^d$, $l_0, l_1, \dots, l_L, \mathfrak{l}_0, \mathfrak{l}_1, \dots, \mathfrak{l}_{\mathfrak{L}} \in \mathbb{N}$ satisfy that

$$d \geq \sum_{i=1}^L l_i(l_{i-1} + 1), \quad \mathfrak{d} \geq \sum_{i=1}^{\mathfrak{L}} \mathfrak{l}_i(\mathfrak{l}_{i-1} + 1), \quad \mathfrak{L} \geq L, \quad \mathfrak{l}_0 = l_0, \quad \text{and} \quad \mathfrak{l}_{\mathfrak{L}} = l_L, \quad (4.186)$$

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

$$\|\vartheta\|_\infty \leq \max\{1, \|\theta\|_\infty\} \quad \text{and} \quad \mathcal{N}_{u,v}^{\vartheta, (\mathfrak{l}_0, \mathfrak{l}_1, \dots, \mathfrak{l}_{\mathfrak{L}})} = \mathcal{N}_{u,v}^{\theta, (l_0, l_1, \dots, l_L)} \quad (4.187)$$

(cf. Definitions 3.3.4 and 4.4.1).

Proof of Lemma 4.4.8. Throughout this proof, let $\eta_1, \eta_2, \dots, \eta_d \in \mathbb{R}$ satisfy

$$\theta = (\eta_1, \eta_2, \dots, \eta_d) \quad (4.188)$$

and let $\Phi \in (\bigtimes_{i=1}^L \mathbb{R}^{l_i \times l_{i-1}} \times \mathbb{R}^{l_i})$ satisfy

$$\mathcal{T}(\Phi) = (\eta_1, \eta_2, \dots, \eta_{\mathcal{P}(\Phi)}) \quad (4.189)$$

(cf. Definitions 1.3.1 and 1.3.6). Note that Lemma 4.4.7 establishes that there exists $\Psi \in \mathbf{N}$ which satisfies

$$\mathcal{D}(\Psi) = (\mathfrak{l}_0, \mathfrak{l}_1, \dots, \mathfrak{l}_{\mathfrak{L}}), \quad \|\mathcal{T}(\Psi)\|_\infty \leq \max\{1, \|\mathcal{T}(\Phi)\|_\infty\}, \quad \text{and} \quad \mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\Psi) = \mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\Phi) \quad (4.190)$$

(cf. Definitions 1.2.4, 1.3.4, and 3.3.4). Next let $\vartheta = (\vartheta_1, \dots, \vartheta_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$ satisfy

$$(\vartheta_1, \vartheta_2, \dots, \vartheta_{\mathcal{P}(\Psi)}) = \mathcal{T}(\Psi) \quad \text{and} \quad \forall i \in \mathbb{N} \cap (\mathcal{P}(\Psi), \mathfrak{d} + 1): \vartheta_i = 0. \quad (4.191)$$

Observe that (4.188), (4.189), (4.190), and (4.191) show that

$$\|\vartheta\|_\infty = \|\mathcal{T}(\Psi)\|_\infty \leq \max\{1, \|\mathcal{T}(\Phi)\|_\infty\} \leq \max\{1, \|\theta\|_\infty\}. \quad (4.192)$$

Furthermore, note that Lemma 4.4.2 and (4.189) demonstrate that for all $x \in \mathbb{R}^{l_0}$ it holds that

$$(\mathcal{N}_{\infty, \infty}^{\theta, (l_0, l_1, \dots, l_L)})(x) = (\mathcal{N}_{\infty, \infty}^{\mathcal{T}(\Phi), \mathcal{D}(\Phi)})(x) = (\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\Phi))(x) \quad (4.193)$$

(cf. Definition 4.4.1). Moreover, observe that Lemma 4.4.2, (4.190), and (4.191) ensure that for all $x \in \mathbb{R}^{l_0}$ it holds that

$$(\mathcal{N}_{\infty, \infty}^{\vartheta, (l_0, l_1, \dots, l_{\mathfrak{L}})})(x) = (\mathcal{N}_{\infty, \infty}^{\mathcal{T}(\Psi), \mathcal{D}(\Psi)})(x) = (\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\Psi))(x). \quad (4.194)$$

Combining this and (4.193) with (4.190) and the assumption that $\mathfrak{l}_0 = l_0$ and $\mathfrak{l}_{\mathfrak{L}} = l_L$ implies that

$$\mathcal{N}_{\infty, \infty}^{\theta, (l_0, l_1, \dots, l_L)} = \mathcal{N}_{\infty, \infty}^{\vartheta, (l_0, l_1, \dots, l_{\mathfrak{L}})}. \quad (4.195)$$

Therefore, we obtain that

$$\mathcal{N}_{u,v}^{\theta, (l_0, l_1, \dots, l_L)} = \mathfrak{C}_{u,v, l_L} \circ \mathcal{N}_{\infty, \infty}^{\theta, (l_0, l_1, \dots, l_L)} = \mathfrak{C}_{u,v, l_{\mathfrak{L}}} \circ \mathcal{N}_{\infty, \infty}^{\vartheta, (l_0, l_1, \dots, l_{\mathfrak{L}})} = \mathcal{N}_{u,v}^{\vartheta, (l_0, l_1, \dots, l_{\mathfrak{L}})} \quad (4.196)$$

(cf. Definition 1.2.10). This and (4.192) prove (4.187). The proof of Lemma 4.4.8 is thus complete. \square

4.4.3 Approximation through ANNs with variable architectures

Corollary 4.4.9. Let $d, K, \mathbf{d}, \mathbf{L} \in \mathbb{N}$, $\mathbf{l} = (l_0, l_1, \dots, l_{\mathbf{L}}) \in \mathbb{N}^{\mathbf{L}+1}$, $L \in [0, \infty)$ satisfy that

$$\mathbf{L} \geq \lceil \log_2(K) \rceil + 2, \quad l_0 = d, \quad l_{\mathbf{L}} = 1, \quad l_1 \geq 2dK, \quad \text{and} \quad \mathbf{d} \geq \sum_{i=1}^{\mathbf{L}} l_i(l_{i-1} + 1), \quad (4.197)$$

assume for all $i \in \mathbb{N} \cap (1, \mathbf{L})$ that $l_i \geq 3 \lceil \frac{K}{2^{i-1}} \rceil$, let $E \subseteq \mathbb{R}^d$ be a set, let $\mathfrak{x}_1, \mathfrak{x}_2, \dots, \mathfrak{x}_K \in E$, and let $f: E \rightarrow \mathbb{R}$ satisfy for all $x, y \in E$ that $|f(x) - f(y)| \leq L \|x - y\|_1$ (cf. Definitions 3.3.4 and 4.2.8). Then there exists $\theta \in \mathbb{R}^{\mathbf{d}}$ such that

$$\|\theta\|_\infty \leq \max\{1, L, \max_{k \in \{1, 2, \dots, K\}} \|\mathfrak{x}_k\|_\infty, 2 \max_{k \in \{1, 2, \dots, K\}} |f(\mathfrak{x}_k)|\} \quad (4.198)$$

and

$$\sup_{x \in E} |f(x) - \mathcal{N}_{-\infty, \infty}^{\theta, \mathbf{l}}(x)| \leq 2L \left[\sup_{x \in E} (\inf_{k \in \{1, 2, \dots, K\}} \|x - \mathfrak{x}_k\|_1) \right] \quad (4.199)$$

(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 $\mathfrak{y} = (f(\mathfrak{x}_1), f(\mathfrak{x}_2), \dots, f(\mathfrak{x}_K))$ and

$$\Phi = \mathbb{M}_K \bullet \mathbf{A}_{-L \mathbf{I}_K, \mathfrak{y}} \bullet \mathbf{P}_K (\mathbb{L}_d \bullet \mathbf{A}_{\mathbf{I}_d, -\mathfrak{x}_1}, \mathbb{L}_d \bullet \mathbf{A}_{\mathbf{I}_d, -\mathfrak{x}_2}, \dots, \mathbb{L}_d \bullet \mathbf{A}_{\mathbf{I}_d, -\mathfrak{x}_K}) \bullet \mathbb{T}_{d, K} \quad (4.200)$$

(cf. Definitions 1.3.1, 1.5.5, 2.1.1, 2.2.1, 2.3.1, 2.4.6, 4.2.1, and 4.2.7). Note that Lemma 4.2.11 and Proposition 4.3.1 establish 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 $i \in \{2, 3, \dots, \mathcal{L}(\Phi) - 1\}$ that $\mathbb{D}_i(\Phi) \leq 3 \lceil \frac{K}{2^{i-1}} \rceil$,
- (VI) it holds that $\|\mathcal{T}(\Phi)\|_\infty \leq \max\{1, L, \max_{k \in \{1, 2, \dots, K\}} \|\mathfrak{x}_k\|_\infty, 2 \max_{k \in \{1, 2, \dots, K\}} |f(\mathfrak{x}_k)|\}$, and
- (VII) it holds that $\sup_{x \in E} |f(x) - (\mathcal{R}_{\mathfrak{x}}^{\mathbf{N}}(\Phi))(x)| \leq 2L \left[\sup_{x \in E} (\inf_{k \in \{1, 2, \dots, K\}} \|x - \mathfrak{x}_k\|_1) \right]$

(cf. Definitions 1.2.4, 1.3.4, and 1.3.6). Furthermore, observe that the fact that $\mathbf{L} \geq \lceil \log_2(K) \rceil + 2 = \mathcal{L}(\Phi)$, the fact that $l_0 = d = \mathbb{D}_0(\Phi)$, the fact that $l_1 \geq 2dK = \mathbb{D}_1(\Phi)$, the fact that for all $i \in \{1, 2, \dots, \mathcal{L}(\Phi) - 1\} \setminus \{1\}$ it holds that $l_i \geq 3 \lceil \frac{K}{2^{i-1}} \rceil \geq \mathbb{D}_i(\Phi)$, the fact that for all $i \in \mathbb{N} \cap (\mathcal{L}(\Phi) - 1, \mathbf{L})$ it holds that $l_i \geq 3 \lceil \frac{K}{2^{i-1}} \rceil \geq 2 = 2\mathbb{D}_{\mathcal{L}(\Phi)}(\Phi)$, the fact that $l_{\mathbf{L}} = 1 = \mathbb{D}_{\mathcal{L}(\Phi)}(\Phi)$, and Lemma 4.4.8 show that there exists $\theta \in \mathbb{R}^{\mathbf{d}}$ which satisfies that

$$\|\theta\|_\infty \leq \max\{1, \|\mathcal{T}(\Phi)\|_\infty\} \quad \text{and} \quad \mathcal{N}_{-\infty, \infty}^{\theta, (l_0, l_1, \dots, l_{\mathbf{L}})} = \mathcal{N}_{-\infty, \infty}^{\mathcal{T}(\Phi), \mathcal{D}(\Phi)}. \quad (4.201)$$

This and item (VI) demonstrate that

$$\|\theta\|_\infty \leq \max\{1, L, \max_{k \in \{1, 2, \dots, K\}} \|\mathfrak{x}_k\|_\infty, 2 \max_{k \in \{1, 2, \dots, K\}} |f(\mathfrak{x}_k)|\}. \quad (4.202)$$

Moreover, note that (4.201), Lemma 4.4.2, and item (VII) ensure that

$$\begin{aligned} \sup_{x \in E} |f(x) - \mathcal{N}_{-\infty, \infty}^{\theta, (\mathbf{l}_0, \mathbf{l}_1, \dots, \mathbf{l}_L)}(x)| &= \sup_{x \in E} |f(x) - \mathcal{N}_{-\infty, \infty}^{\mathcal{T}(\Phi), \mathcal{D}(\Phi)}(x)| \\ &= \sup_{x \in E} |f(x) - (\mathcal{R}_{\mathbf{r}}^{\mathbf{N}}(\Phi))(x)| \\ &\leq 2L [\sup_{x \in E} (\inf_{k \in \{1, 2, \dots, K\}} \|x - \mathfrak{x}_k\|_1)] \end{aligned} \quad (4.203)$$

(cf. Definition 4.4.1). The proof of Corollary 4.4.9 is thus complete. \square

Corollary 4.4.10. Let $d, K, \mathbf{d}, \mathbf{L} \in \mathbb{N}$, $\mathbf{l} = (\mathbf{l}_0, \mathbf{l}_1, \dots, \mathbf{l}_L) \in \mathbb{N}^{\mathbf{L}+1}$, $L \in [0, \infty)$, $u \in [-\infty, \infty)$, $v \in (u, \infty]$ satisfy that

$$\mathbf{L} \geq \lceil \log_2 K \rceil + 2, \quad \mathbf{l}_0 = d, \quad \mathbf{l}_L = 1, \quad \mathbf{l}_1 \geq 2dK, \quad \text{and} \quad \mathbf{d} \geq \sum_{i=1}^{\mathbf{L}} \mathbf{l}_i (\mathbf{l}_{i-1} + 1), \quad (4.204)$$

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 $\mathfrak{x}_1, \mathfrak{x}_2, \dots, \mathfrak{x}_K \in E$, and let $f: E \rightarrow ([u, v] \cap \mathbb{R})$ satisfy for all $x, y \in E$ that $|f(x) - f(y)| \leq L \|x - y\|_1$ (cf. Definitions 3.3.4 and 4.2.8). Then there exists $\theta \in \mathbb{R}^{\mathbf{d}}$ such that

$$\|\theta\|_\infty \leq \max\{1, L, \max_{k \in \{1, 2, \dots, K\}} \|\mathfrak{x}_k\|_\infty, 2 \max_{k \in \{1, 2, \dots, K\}} |f(\mathfrak{x}_k)|\} \quad (4.205)$$

and

$$\sup_{x \in E} |f(x) - \mathcal{N}_{u, v}^{\theta, \mathbf{l}}(x)| \leq 2L [\sup_{x \in E} (\inf_{k \in \{1, 2, \dots, K\}} \|x - \mathfrak{x}_k\|_1)]. \quad (4.206)$$

(cf. Definition 4.4.1).

Proof of Corollary 4.4.10. Observe that Corollary 4.4.9 implies that there exists $\theta \in \mathbb{R}^{\mathbf{d}}$ such that

$$\|\theta\|_\infty \leq \max\{1, L, \max_{k \in \{1, 2, \dots, K\}} \|\mathfrak{x}_k\|_\infty, 2 \max_{k \in \{1, 2, \dots, K\}} |f(\mathfrak{x}_k)|\} \quad (4.207)$$

and

$$\sup_{x \in E} |f(x) - \mathcal{N}_{-\infty, \infty}^{\theta, \mathbf{l}}(x)| \leq 2L [\sup_{x \in E} (\inf_{k \in \{1, 2, \dots, K\}} \|x - \mathfrak{x}_k\|_1)]. \quad (4.208)$$

Furthermore, note that the assumption that $f(E) \subseteq [u, v]$ proves that for all $x \in E$ it holds that

$$f(x) = \mathfrak{c}_{u, v}(f(x)) \quad (4.209)$$

(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 (4.208) hence establish that

$$\begin{aligned} \sup_{x \in E} |f(x) - \mathcal{N}_{u, v}^{\theta, \mathbf{l}}(x)| &= \sup_{x \in E} |\mathfrak{c}_{u, v}(f(x)) - \mathfrak{c}_{u, v}(\mathcal{N}_{-\infty, \infty}^{\theta, \mathbf{l}}(x))| \\ &\leq \sup_{x \in E} |f(x) - \mathcal{N}_{-\infty, \infty}^{\theta, \mathbf{l}}(x)| \leq 2L [\sup_{x \in E} (\inf_{k \in \{1, 2, \dots, K\}} \|x - \mathfrak{x}_k\|_1)]. \end{aligned} \quad (4.210)$$

The proof of Corollary 4.4.10 is thus complete. \square

4.4.4 Refined convergence rates for the approximation error

Lemma 4.4.11. Let $d, \mathbf{d}, \mathbf{L} \in \mathbb{N}$, $L, a \in \mathbb{R}$, $b \in (a, \infty)$, $u \in [-\infty, \infty)$, $v \in (u, \infty]$, $\mathbf{l} = (\mathbf{l}_0, \mathbf{l}_1, \dots, \mathbf{l}_{\mathbf{L}}) \in \mathbb{N}^{\mathbf{L}+1}$, 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 $f: [a, b]^d \rightarrow ([u, v] \cap \mathbb{R})$ 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

$$\sup_{x \in [a, b]^d} |\mathcal{M}_{u,v}^{\vartheta, \mathbf{l}}(x) - f(x)| \leq \frac{dL(b-a)}{2} \quad (4.211)$$

(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 $\mathbf{m} = (\mathbf{m}_1, \dots, \mathbf{m}_d) \in [a, b]^d$ satisfy for all $i \in \{1, 2, \dots, d\}$ that

$$\mathbf{m}_i = \frac{a+b}{2}, \quad (4.212)$$

and let $\vartheta = (\vartheta_1, \dots, \vartheta_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$ satisfy for all $i \in \{1, 2, \dots, \mathfrak{d}\} \setminus \{\mathfrak{d}\}$ that $\vartheta_i = 0$ and $\vartheta_{\mathfrak{d}} = f(\mathbf{m})$. Observe that the assumption that $\mathbf{l}_{\mathbf{L}} = 1$ and the fact that $\forall i \in \{1, 2, \dots, \mathfrak{d}-1\}: \vartheta_i = 0$ show that for all $x = (x_1, \dots, x_{\mathbf{l}_{\mathbf{L}-1}}) \in \mathbb{R}^{\mathbf{l}_{\mathbf{L}-1}}$ it holds that

$$\begin{aligned} \mathcal{A}_{1, \mathbf{l}_{\mathbf{L}-1}}^{\vartheta, \sum_{i=1}^{\mathbf{L}-1} \mathbf{l}_i(\mathbf{l}_{i-1} + 1)}(x) &= \left[\sum_{i=1}^{\mathbf{L}-1} \vartheta_{[\sum_{i=1}^{\mathbf{L}-1} \mathbf{l}_i(\mathbf{l}_{i-1} + 1)]+i} x_i \right] + \vartheta_{[\sum_{i=1}^{\mathbf{L}-1} \mathbf{l}_i(\mathbf{l}_{i-1} + 1)]+\mathbf{l}_{\mathbf{L}-1}+1} \\ &= \left[\sum_{i=1}^{\mathbf{L}-1} \vartheta_{[\sum_{i=1}^{\mathbf{L}} \mathbf{l}_i(\mathbf{l}_{i-1} + 1)]-(\mathbf{l}_{\mathbf{L}-1}-i+1)} x_i \right] + \vartheta_{\sum_{i=1}^{\mathbf{L}} \mathbf{l}_i(\mathbf{l}_{i-1} + 1)} \\ &= \left[\sum_{i=1}^{\mathbf{L}-1} \vartheta_{\mathfrak{d}-(\mathbf{l}_{\mathbf{L}-1}-i+1)} x_i \right] + \vartheta_{\mathfrak{d}} = \vartheta_{\mathfrak{d}} = f(\mathbf{m}) \end{aligned} \quad (4.213)$$

(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

$$\begin{aligned} (\mathfrak{C}_{u,v, \mathbf{l}_{\mathbf{L}}} \circ \mathcal{A}_{\mathbf{l}_{\mathbf{L}}, \mathbf{l}_{\mathbf{L}-1}}^{\vartheta, \sum_{i=1}^{\mathbf{L}-1} \mathbf{l}_i(\mathbf{l}_{i-1} + 1)})(x) &= (\mathfrak{C}_{u,v, 1} \circ \mathcal{A}_{1, \mathbf{l}_{\mathbf{L}-1}}^{\vartheta, \sum_{i=1}^{\mathbf{L}-1} \mathbf{l}_i(\mathbf{l}_{i-1} + 1)})(x) \\ &= \mathfrak{c}_{u,v}(f(\mathbf{m})) = \max\{u, \min\{f(\mathbf{m}), v\}\} \\ &= \max\{u, f(\mathbf{m})\} = f(\mathbf{m}) \end{aligned} \quad (4.214)$$

(cf. Definitions 1.2.9 and 1.2.10). This ensures for all $x \in \mathbb{R}^d$ that

$$\mathcal{M}_{u,v}^{\vartheta, \mathbf{l}}(x) = f(\mathbf{m}). \quad (4.215)$$

Furthermore, note that (4.212) implies that for all $x \in [a, \mathbf{m}_1]$, $\mathfrak{x} \in [\mathbf{m}_1, b]$ it holds that

$$\begin{aligned} |\mathbf{m}_1 - x| &= \mathbf{m}_1 - x = (a+b)/2 - x \leq (a+b)/2 - a = (b-a)/2 \\ \text{and } |\mathbf{m}_1 - \mathfrak{x}| &= \mathfrak{x} - \mathbf{m}_1 = \mathfrak{x} - (a+b)/2 \leq b - (a+b)/2 = (b-a)/2. \end{aligned} \quad (4.216)$$

The assumption that $\forall x, y \in [a, b]^d: |f(x) - f(y)| \leq L\|x - y\|_1$ and (4.215) therefore prove that for all $x = (x_1, \dots, x_d) \in [a, b]^d$ it holds that

$$\begin{aligned} |\mathcal{N}_{u,v}^{\vartheta,\mathbf{l}}(x) - f(x)| &= |f(\mathbf{m}) - f(x)| \leq L\|\mathbf{m} - x\|_1 = L \sum_{i=1}^d |\mathbf{m}_i - x_i| \\ &= L \sum_{i=1}^d |\mathbf{m}_i - x_i| \leq \sum_{i=1}^d \frac{L(b-a)}{2} = \frac{dL(b-a)}{2}. \end{aligned} \quad (4.217)$$

This and the fact that $\|\vartheta\|_\infty = \max_{i \in \{1, 2, \dots, \mathbf{d}\}} |\vartheta_i| = |f(\mathbf{m})| \leq \sup_{x \in [a, b]^d} |f(x)|$ establish (4.211). The proof of Lemma 4.4.11 is thus complete. \square

Proposition 4.4.12. Let $d, \mathbf{d}, \mathbf{L} \in \mathbb{N}$, $A \in (0, \infty)$, $L, a \in \mathbb{R}$, $b \in (a, \infty)$, $u \in [-\infty, \infty)$, $v \in (u, \infty]$, $\mathbf{l} = (\mathbf{l}_0, \mathbf{l}_1, \dots, \mathbf{l}_{\mathbf{L}}) \in \mathbb{N}^{\mathbf{L}+1}$, assume

$$\mathbf{L} \geq 1 + (\lceil \log_2(A/(2d)) \rceil + 1) \mathbb{1}_{(6^d, \infty)}(A), \quad \mathbf{l}_0 = d, \quad \mathbf{l}_1 \geq A \mathbb{1}_{(6^d, \infty)}(A), \quad \mathbf{l}_{\mathbf{L}} = 1, \quad (4.218)$$

and $\mathbf{d} \geq \sum_{i=1}^{\mathbf{L}} \mathbf{l}_i (\mathbf{l}_{i-1} + 1)$, assume for all $i \in \{1, 2, \dots, \mathbf{L}\} \setminus \{1, \mathbf{L}\}$ that

$$\mathbf{l}_i \geq 3^{\lceil A/(2^i d) \rceil} \mathbb{1}_{(6^d, \infty)}(A), \quad (4.219)$$

and let $f: [a, b]^d \rightarrow ([u, v] \cap \mathbb{R})$ satisfy for all $x, y \in [a, b]^d$ that

$$|f(x) - f(y)| \leq L\|x - y\|_1 \quad (4.220)$$

(cf. Definitions 3.3.4 and 4.2.8). 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

$$\sup_{x \in [a, b]^d} |\mathcal{N}_{u,v}^{\vartheta,\mathbf{l}}(x) - f(x)| \leq \frac{3dL(b-a)}{A^{1/d}} \quad (4.221)$$

(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 $\mathfrak{Z} = \lfloor (\frac{A}{2d})^{1/d} \rfloor \in \mathbb{Z}$. Observe that the fact that for all $k \in \mathbb{N}$ it holds that $2k \leq 2(2^{k-1}) = 2^k$ shows that $3^d = 6^d/2^d \leq A/(2d)$. Hence, we obtain that

$$2 \leq \frac{2}{3} \left(\frac{A}{2d} \right)^{1/d} \leq \left(\frac{A}{2d} \right)^{1/d} - 1 < \mathfrak{Z}. \quad (4.222)$$

In the next step let $r = \frac{d(b-a)}{2\mathfrak{Z}} \in (0, \infty)$, let $\delta: [a, b]^d \times [a, b]^d \rightarrow \mathbb{R}$ 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). Note that (4.222) and Lemma 4.3.7 demonstrate that

$$K = \max\{2, \mathcal{C}^{([a, b]^d, \delta), r}\} \leq \max\left\{2, \left(\lceil \frac{d(b-a)}{2r} \rceil \right)^d\right\} = \max\{2, (\lceil \mathfrak{Z} \rceil)^d\} = \mathfrak{Z}^d < \infty. \quad (4.223)$$

This ensures that

$$4 \leq 2dK \leq 2d\mathfrak{Z}^d \leq \frac{2dA}{2d} = A. \quad (4.224)$$

Combining this and the fact that $\mathbf{L} \geq 1 + (\lceil \log_2(A/(2d)) \rceil + 1)\mathbb{1}_{(6^d, \infty)}(A) = \lceil \log_2(A/(2d)) \rceil + 2$ therefore implies that $\lceil \log_2(K) \rceil \leq \lceil \log_2(A/(2d)) \rceil \leq \mathbf{L} - 2$. This, (4.224), the assumption that $\mathbf{l}_1 \geq A\mathbb{1}_{(6^d, \infty)}(A) = A$, and the assumption that $\forall i \in \{2, 3, \dots, \mathbf{L} - 1\}: \mathbf{l}_i \geq 3\lceil A/(2^{i-1}d) \rceil \mathbb{1}_{(6^d, \infty)}(A) = 3\lceil A/(2^{i-1}d) \rceil$ prove that for all $i \in \{2, 3, \dots, \mathbf{L} - 1\}$ it holds that

$$\mathbf{L} \geq \lceil \log_2(K) \rceil + 2, \quad \mathbf{l}_1 \geq A \geq 2dK, \quad \text{and} \quad \mathbf{l}_i \geq 3\lceil \frac{A}{2^{i-1}d} \rceil \geq 3\lceil \frac{K}{2^{i-1}} \rceil. \quad (4.225)$$

Let $\mathfrak{x}_1, \mathfrak{x}_2, \dots, \mathfrak{x}_K \in [a, b]^d$ satisfy

$$\sup_{x \in [a, b]^d} [\inf_{k \in \{1, 2, \dots, K\}} \delta(x, \mathfrak{x}_k)] \leq r. \quad (4.226)$$

Observe that (4.225), 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 $\forall x, y \in [a, b]^d: |f(x) - f(y)| \leq L\|x - y\|_1$, and Corollary 4.4.10 establish that there exists $\vartheta \in \mathbb{R}^{\mathbf{d}}$ such that

$$\|\vartheta\|_\infty \leq \max\{1, L, \max_{k \in \{1, 2, \dots, K\}} \|\mathfrak{x}_k\|_\infty, 2 \max_{k \in \{1, 2, \dots, K\}} |f(\mathfrak{x}_k)|\} \quad (4.227)$$

and

$$\begin{aligned} \sup_{x \in [a, b]^d} |\mathcal{N}_{u, v}^{\vartheta, 1}(x) - f(x)| &\leq 2L[\sup_{x \in [a, b]^d} (\inf_{k \in \{1, 2, \dots, K\}} \|x - \mathfrak{x}_k\|_1)] \\ &= 2L[\sup_{x \in [a, b]^d} (\inf_{k \in \{1, 2, \dots, K\}} \delta(x, \mathfrak{x}_k))]. \end{aligned} \quad (4.228)$$

Note that (4.227) shows that

$$\|\vartheta\|_\infty \leq \max\{1, L, |a|, |b|, 2 \sup_{x \in [a, b]^d} |f(x)|\}. \quad (4.229)$$

Furthermore, observe that (4.228), (4.222), (4.226), and the fact that for all $k \in \mathbb{N}$ it holds that $2k \leq 2(2^{k-1}) = 2^k$ demonstrate that

$$\begin{aligned} \sup_{x \in [a, b]^d} |\mathcal{N}_{u, v}^{\vartheta, 1}(x) - f(x)| &\leq 2L[\sup_{x \in [a, b]^d} (\inf_{k \in \{1, 2, \dots, K\}} \delta(x, \mathfrak{x}_k))] \\ &\leq 2Lr = \frac{dL(b-a)}{3} \leq \frac{dL(b-a)}{\frac{2}{3} \left(\frac{A}{2d}\right)^{1/d}} = \frac{(2d)^{1/d} 3dL(b-a)}{2A^{1/d}} \leq \frac{3dL(b-a)}{A^{1/d}}. \end{aligned} \quad (4.230)$$

Combining this with (4.229) ensures (4.221). The proof of Proposition 4.4.12 is thus complete. \square

Corollary 4.4.13. Let $d \in \mathbb{N}$, $a \in \mathbb{R}$, $b \in (a, \infty)$, $L \in (0, \infty)$ and let $f: [a, b]^d \rightarrow \mathbb{R}$ satisfy for all $x, y \in [a, b]^d$ that

$$|f(x) - f(y)| \leq L\|x - y\|_1 \quad (4.231)$$

(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

$$\mathcal{H}(\mathbf{F}) \leq \max\{0, d(\log_2(\varepsilon^{-1}) + \log_2(d) + \log_2(3L(b-a)) + 1)\}, \quad (4.232)$$

$$\|\mathcal{T}(\mathbf{F})\|_\infty \leq \max\{1, L, |a|, |b|, 2[\sup_{x \in [a,b]^d} |f(x)|]\}, \quad \mathcal{R}_{\mathbf{r}}^{\mathbf{N}}(\mathbf{F}) \in C(\mathbb{R}^d, \mathbb{R}), \quad (4.233)$$

$$\sup_{x \in [a,b]^d} |(\mathcal{R}_{\mathbf{r}}^{\mathbf{N}}(\mathbf{F}))(x) - f(x)| \leq \varepsilon, \quad \text{and} \quad \mathcal{P}(\mathbf{F}) \leq \mathfrak{C}\varepsilon^{-2d} \quad (4.234)$$

(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

$$\mathfrak{C} = \frac{9}{8}(3dL(b-a))^{2d} + (d+22)(3dL(b-a))^d + d + 11, \quad (4.235)$$

for every $\varepsilon \in (0, 1]$ let $A_\varepsilon \in (0, \infty)$, $\mathbf{L}_\varepsilon \in \mathbb{N}$, $\mathbf{l}^{(\varepsilon)} = (\mathbf{l}_0^{(\varepsilon)}, \mathbf{l}_1^{(\varepsilon)}, \dots, \mathbf{l}_{\mathbf{L}_\varepsilon}^{(\varepsilon)}) \in \mathbb{N}^{\mathbf{L}_\varepsilon+1}$ satisfy

$$A_\varepsilon = \left(\frac{3dL(b-a)}{\varepsilon} \right)^d, \quad \mathbf{L}_\varepsilon = 1 + (\lceil \log_2(\frac{A_\varepsilon}{2d}) \rceil + 1) \mathbb{1}_{(6^d, \infty)}(A_\varepsilon), \quad (4.236)$$

$$\mathbf{l}_0^{(\varepsilon)} = d, \quad \mathbf{l}_1^{(\varepsilon)} = \lfloor A_\varepsilon \rfloor \mathbb{1}_{(6^d, \infty)}(A_\varepsilon) + 1, \quad \text{and} \quad \mathbf{l}_{\mathbf{L}_\varepsilon}^{(\varepsilon)} = 1, \quad (4.237)$$

and assume for all $\varepsilon \in (0, 1]$, $i \in \{2, 3, \dots, \mathbf{L}_\varepsilon - 1\}$ that

$$\mathbf{l}_i^{(\varepsilon)} = 3 \lceil \frac{A_\varepsilon}{2^i d} \rceil \mathbb{1}_{(6^d, \infty)}(A_\varepsilon) \quad (4.238)$$

(cf. Definition 4.2.8). Observe that the fact that for all $\varepsilon \in (0, 1]$ it holds that $\mathbf{L}_\varepsilon \geq 1 + (\lceil \log_2(\frac{A_\varepsilon}{2d}) \rceil + 1) \mathbb{1}_{(6^d, \infty)}(A_\varepsilon)$, the fact that for all $\varepsilon \in (0, 1]$ it holds that $\mathbf{l}_0^{(\varepsilon)} = d$, the fact that for all $\varepsilon \in (0, 1]$ it holds that $\mathbf{l}_1^{(\varepsilon)} \geq A_\varepsilon \mathbb{1}_{(6^d, \infty)}(A_\varepsilon)$, the fact that for all $\varepsilon \in (0, 1]$ it holds that $\mathbf{l}_{\mathbf{L}_\varepsilon}^{(\varepsilon)} = 1$, the fact that for all $\varepsilon \in (0, 1]$, $i \in \{2, 3, \dots, \mathbf{L}_\varepsilon - 1\}$ it holds that $\mathbf{l}_i^{(\varepsilon)} \geq 3 \lceil \frac{A_\varepsilon}{2^i d} \rceil \mathbb{1}_{(6^d, \infty)}(A_\varepsilon)$, Proposition 4.4.12, and Lemma 4.4.2 imply that for all $\varepsilon \in (0, 1]$ there exists $\mathbf{F}_\varepsilon \in (\bigtimes_{i=1}^{\mathbf{L}_\varepsilon} (\mathbb{R}^{\mathbf{l}_i^{(\varepsilon)} \times \mathbf{l}_{i-1}^{(\varepsilon)}} \times \mathbb{R}^{\mathbf{l}_i^{(\varepsilon)}})) \subseteq \mathbf{N}$ which satisfies $\|\mathcal{T}(\mathbf{F}_\varepsilon)\|_\infty \leq \max\{1, L, |a|, |b|, 2[\sup_{x \in [a,b]^d} |f(x)|]\}$ and

$$\sup_{x \in [a,b]^d} |(\mathcal{R}_{\mathbf{r}}^{\mathbf{N}}(\mathbf{F}_\varepsilon))(x) - f(x)| \leq \frac{3dL(b-a)}{(A_\varepsilon)^{1/d}} = \varepsilon. \quad (4.239)$$

(cf. Definitions 1.2.4, 1.3.1, 1.3.4, and 1.3.6). Furthermore, note that the fact that $d \geq 1$ proves that for all $\varepsilon \in (0, 1]$ it holds that

$$\begin{aligned} \mathcal{H}(\mathbf{F}_\varepsilon) &= \mathbf{L}_\varepsilon - 1 = (\lceil \log_2(\frac{A_\varepsilon}{2d}) \rceil + 1) \mathbb{1}_{(6^d, \infty)}(A_\varepsilon) \\ &= \lceil \log_2(\frac{A_\varepsilon}{d}) \rceil \mathbb{1}_{(6^d, \infty)}(A_\varepsilon) \leq \max\{0, \log_2(A_\varepsilon) + 1\}. \end{aligned} \quad (4.240)$$

Combining this and the fact that for all $\varepsilon \in (0, 1]$ it holds that

$$\log_2(A_\varepsilon) = d \log_2\left(\frac{3dL(b-a)}{\varepsilon}\right) = d(\log_2(\varepsilon^{-1}) + \log_2(d) + \log_2(3L(b-a))) \quad (4.241)$$

establishes that for all $\varepsilon \in (0, 1]$ it holds that

$$\mathcal{H}(\mathbf{F}_\varepsilon) \leq \max\{0, d(\log_2(\varepsilon^{-1}) + \log_2(d) + \log_2(3L(b-a))) + 1\}. \quad (4.242)$$

Moreover, observe that (4.237) and (4.238) show that for all $\varepsilon \in (0, 1]$ it holds that

$$\begin{aligned} \mathcal{P}(\mathbf{F}_\varepsilon) &= \sum_{i=1}^{\mathbf{L}_\varepsilon} l_i^{(\varepsilon)}(l_{i-1}^{(\varepsilon)} + 1) \\ &\leq (\lfloor A_\varepsilon \rfloor + 1)(d + 1) + 3 \lceil \frac{A_\varepsilon}{4d} \rceil (\lfloor A_\varepsilon \rfloor + 2) \\ &\quad + \max\{\lfloor A_\varepsilon \rfloor + 1, 3 \lceil \frac{A_\varepsilon}{2^{\mathbf{L}_\varepsilon-1}d} \rceil\} + 1 + \sum_{i=3}^{\mathbf{L}_\varepsilon-1} 3 \lceil \frac{A_\varepsilon}{2^i d} \rceil (3 \lceil \frac{A_\varepsilon}{2^{i-1}d} \rceil + 1) \\ &\leq (A_\varepsilon + 1)(d + 1) + 3\left(\frac{A_\varepsilon}{4} + 1\right)(A_\varepsilon + 2) + 3A_\varepsilon + 4 + \sum_{i=3}^{\mathbf{L}_\varepsilon-1} 3\left(\frac{A_\varepsilon}{2^i} + 1\right)\left(\frac{3A_\varepsilon}{2^{i-1}} + 4\right). \end{aligned} \quad (4.243)$$

In addition, note that the fact that $\forall x \in (0, \infty): \log_2(x) = \log_2(x/2) + 1 \leq x/2 + 1$ demonstrates that for all $\varepsilon \in (0, 1]$ it holds that

$$\mathbf{L}_\varepsilon \leq 2 + \log_2\left(\frac{A_\varepsilon}{d}\right) \leq 3 + \frac{A_\varepsilon}{2d} \leq 3 + \frac{A_\varepsilon}{2}. \quad (4.244)$$

This ensures that for all $\varepsilon \in (0, 1]$ it holds that

$$\begin{aligned} &\sum_{i=3}^{\mathbf{L}_\varepsilon-1} 3\left(\frac{A_\varepsilon}{2^i} + 1\right)\left(\frac{3A_\varepsilon}{2^{i-1}} + 4\right) \\ &\leq 9(A_\varepsilon)^2 \left[\sum_{i=3}^{\mathbf{L}_\varepsilon-1} 2^{1-2i} \right] + 12A_\varepsilon \left[\sum_{i=3}^{\mathbf{L}_\varepsilon-1} 2^{-i} \right] + 9A_\varepsilon \left[\sum_{i=3}^{\mathbf{L}_\varepsilon-1} 2^{1-i} \right] + 12(\mathbf{L}_\varepsilon - 3) \\ &\leq \frac{9(A_\varepsilon)^2}{8} \left[\sum_{i=1}^{\infty} 4^{-i} \right] + 3A_\varepsilon \left[\sum_{i=1}^{\infty} 2^{-i} \right] + \frac{9A_\varepsilon}{2} \left[\sum_{i=1}^{\infty} 2^{-i} \right] + 6A_\varepsilon \\ &= \frac{3}{8}(A_\varepsilon)^2 + 3A_\varepsilon + \frac{9}{2}A_\varepsilon + 6A_\varepsilon = \frac{3}{8}(A_\varepsilon)^2 + \frac{27}{2}A_\varepsilon. \end{aligned} \quad (4.245)$$

This and (4.243) imply that for all $\varepsilon \in (0, 1]$ it holds that

$$\begin{aligned} \mathcal{P}(\mathbf{F}_\varepsilon) &\leq \left(\frac{3}{4} + \frac{3}{8}\right)(A_\varepsilon)^2 + (d + 1 + \frac{9}{2} + 3 + \frac{27}{2})A_\varepsilon + d + 1 + 6 + 4 \\ &= \frac{9}{8}(A_\varepsilon)^2 + (d + 22)A_\varepsilon + d + 11. \end{aligned} \quad (4.246)$$

Combining this, (4.235), and (4.236) proves that

$$\begin{aligned}\mathcal{P}(\mathbf{F}_\varepsilon) &\leq \frac{9}{8} (3dL(b-a))^{2d} \varepsilon^{-2d} + (d+22) (3dL(b-a))^d \varepsilon^{-d} + d + 11 \\ &\leq \left[\frac{9}{8} (3dL(b-a))^{2d} + (d+22) (3dL(b-a))^d + d + 11 \right] \varepsilon^{-2d} = \mathfrak{C} \varepsilon^{-2d}. \end{aligned}\quad (4.247)$$

Combining this with (4.239) and (4.242) establishes (4.232), (4.233), and (4.234). The proof of Corollary 4.4.13 is thus complete. \square

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 (4.234) 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 (4.234).

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 instance, to [14, 15, 28, 73, 127, 167] and the references therein for such high-dimensional ANN approximation results in the context of general classes of target functions and we refer, for example, to [3, 29, 35, 129, 134, 168–170, 184, 186, 216, 220, 241, 273, 374] 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 instance, [32, 70, 71, 213, 269, 384] 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, 46, 132, 227, 236, 237, 272] 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 $d, M \in \mathbb{N}$, $\mathcal{E} \in C(\mathbb{R}^d, \mathbb{R})$, $x_1, x_2, \dots, x_{M+1} \in \mathbb{R}^d$, $y_1, y_2, \dots, y_M \in \mathbb{R}$ satisfy for all $m \in \{1, 2, \dots, M\}$ that

$$y_m = \mathcal{E}(x_m) \tag{5.1}$$

and let $\mathcal{L}: C(\mathbb{R}^d, \mathbb{R}) \rightarrow [0, \infty)$ satisfy for all $\phi \in C(\mathbb{R}^d, \mathbb{R})$ that

$$\mathcal{L}(\phi) = \frac{1}{M} \left[\sum_{m=1}^M |\phi(x_m) - y_m|^2 \right]. \quad (5.2)$$

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 $\mathcal{E}: \mathbb{R}^d \rightarrow \mathbb{R}$ as an unknown function which relates input and output data through (5.1), we think of $x_1, x_2, \dots, x_{M+1} \in \mathbb{R}^d$ as the available known input data, we think of $y_1, y_2, \dots, y_M \in \mathbb{R}$ as the available known output data, and we have that the function $\mathcal{L}: C(\mathbb{R}^d, \mathbb{R}) \rightarrow [0, \infty)$ 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

$$\mathcal{L}(\mathcal{E}) = 0 \quad (5.3)$$

and we are trying to approximate the function \mathcal{E} by computing an approximate minimizer of the function $\mathcal{L}: C(\mathbb{R}^d, \mathbb{R}) \rightarrow [0, \infty)$. In order to make this optimization problem amenable to numerical computations, we consider a spatially discretized version of the optimization problem associated to (5.2) by employing parametrizations of [ANNs](#) (cf. (7) in the [introduction](#)).

More formally, let $a: \mathbb{R} \rightarrow \mathbb{R}$ be differentiable, let $h \in \mathbb{N}$, $l_1, l_2, \dots, l_h, \mathfrak{d} \in \mathbb{N}$ satisfy $\mathfrak{d} = l_1(d+1) + [\sum_{k=2}^h l_k(l_{k-1}+1)] + l_h + 1$, and consider the parametrization function

$$\mathbb{R}^{\mathfrak{d}} \ni \theta \mapsto \mathcal{N}_{\mathfrak{M}_{a,l_1}, \mathfrak{M}_{a,l_2}, \dots, \mathfrak{M}_{a,l_h}, \text{id}_{\mathbb{R}}}^{\theta, d} \in C(\mathbb{R}^d, \mathbb{R}) \quad (5.4)$$

(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 $i \in \{1, 2, \dots, h\}$ 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

$$\mathbb{R}^d \ni x \mapsto \mathcal{N}_{\mathfrak{M}_{a,l_1}, \mathfrak{M}_{a,l_2}, \dots, \mathfrak{M}_{a,l_h}, \text{id}_{\mathbb{R}}}^{\theta, d} \in \mathbb{R} \quad (5.5)$$

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 $a: \mathbb{R} \rightarrow \mathbb{R}$ is applied. We restrict ourselves in this section to a differentiable activation function as this differentiability property allows us to consider gradients (cf. (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 $\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \rightarrow [0, \infty)$ which satisfies for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that

$$\mathcal{L}(\theta) = \frac{1}{M} \left[\sum_{m=1}^M |(\mathcal{N}_{\mathfrak{M}_{a,l_1}, \mathfrak{M}_{a,l_2}, \dots, \mathfrak{M}_{a,l_h}, \text{id}_{\mathbb{R}}}^{\theta, d})(x_m) - y_m|^2 \right] \quad (5.6)$$

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 $\xi \in \mathbb{R}^d$, let $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, and let $\theta = (\theta_n)_{n \in \mathbb{N}_0} : \mathbb{N}_0 \rightarrow \mathbb{R}^d$ satisfy for all $n \in \mathbb{N}$ that

$$\theta_0 = \xi \quad \text{and} \quad \theta_n = \theta_{n-1} - \gamma_n(\nabla \mathcal{L})(\theta_{n-1}). \quad (5.7)$$

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 $\Theta = (\Theta_t)_{t \in [0, \infty)} : [0, \infty) \rightarrow \mathbb{R}^d$ be a continuously differentiable function which satisfies for all $t \in [0, \infty)$ that

$$\Theta_0 = \xi \quad \text{and} \quad \dot{\Theta}_t = \frac{\partial}{\partial t} \Theta_t = -(\nabla \mathcal{L})(\Theta_t). \quad (5.8)$$

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 $\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$ be a function, and let $\mathcal{G}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a $\mathcal{B}(\mathbb{R}^{\mathfrak{d}})/\mathcal{B}(\mathbb{R}^{\mathfrak{d}})$ -measurable function which satisfies for all $U \in \{V \subseteq \mathbb{R}^{\mathfrak{d}}: V \text{ is open}\}$, $\theta \in U$ with $\mathcal{L}|_U \in C^1(U, \mathbb{R}^{\mathfrak{d}})$ that

$$\mathcal{G}(\theta) = (\nabla \mathcal{L})(\theta). \quad (5.9)$$

Then we say that Θ is a *GF* trajectory for the objective function \mathcal{L} with generalized gradient \mathcal{G} and initial value ξ (we say that Θ is a *GF* trajectory for the objective function \mathcal{L} with initial value ξ , we say that Θ is a solution of the *GF ODE* for the objective function \mathcal{L} with generalized gradient \mathcal{G} and initial value ξ , we say that Θ is a solution of the *GF ODE* for the objective function \mathcal{L} with initial value ξ) if and only if it holds that $\Theta: [0, \infty) \rightarrow \mathbb{R}^{\mathfrak{d}}$ is a continuous function from $[0, \infty)$ to $\mathbb{R}^{\mathfrak{d}}$ which satisfies for all $t \in [0, \infty)$ that $\int_0^t \|\mathcal{G}(\Theta_s)\|_2 ds < \infty$ and

$$\Theta_t = \xi - \int_0^t \mathcal{G}(\Theta_s) ds \quad (5.10)$$

(cf. Definition 3.3.4).

5.2.2 Direction of negative gradients

Lemma 5.2.2. Let $\mathfrak{d} \in \mathbb{N}$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$, $\theta \in \mathbb{R}^{\mathfrak{d}}$, $r \in (0, \infty)$ and let $\mathcal{G}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $v \in \mathbb{R}^{\mathfrak{d}}$ that

$$\mathcal{G}(v) = \lim_{h \rightarrow 0} \left(\frac{\mathcal{L}(\theta + hv) - \mathcal{L}(\theta)}{h} \right) = [\mathcal{L}'(\theta)](v). \quad (5.11)$$

Then

(i) it holds that

$$\sup_{v \in \{w \in \mathbb{R}^{\mathfrak{d}}: \|w\|_2=r\}} \mathcal{G}(v) = r \|(\nabla \mathcal{L})(\theta)\|_2 = \begin{cases} 0 & : (\nabla \mathcal{L})(\theta) = 0 \\ \mathcal{G}\left(\frac{r(\nabla \mathcal{L})(\theta)}{\|(\nabla \mathcal{L})(\theta)\|_2}\right) & : (\nabla \mathcal{L})(\theta) \neq 0 \end{cases} \quad (5.12)$$

and

(ii) it holds that

$$\inf_{v \in \{\mathbf{w} \in \mathbb{R}^{\mathfrak{d}} : \|v\|_2 = r\}} \mathcal{G}(v) = -r \|\nabla \mathcal{L}(\theta)\|_2 = \begin{cases} 0 & : (\nabla \mathcal{L})(\theta) = 0 \\ \mathcal{G}\left(\frac{-r(\nabla \mathcal{L})(\theta)}{\|\nabla \mathcal{L}(\theta)\|_2}\right) & : (\nabla \mathcal{L})(\theta) \neq 0 \end{cases} \quad (5.13)$$

(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

$$\mathcal{G}(v) = \langle (\nabla \mathcal{L})(\theta), v \rangle \quad (5.14)$$

(cf. Definition 1.4.7). The Cauchy–Schwarz inequality hence ensures that for all $v \in \mathbb{R}^{\mathfrak{d}}$ with $\|v\|_2 = r$ it holds that

$$\begin{aligned} -r \|\nabla \mathcal{L}(\theta)\|_2 &= -\|\nabla \mathcal{L}(\theta)\|_2 \|v\|_2 \leq -\langle -(\nabla \mathcal{L})(\theta), v \rangle \\ &= \mathcal{G}(v) \leq \|\nabla \mathcal{L}(\theta)\|_2 \|v\|_2 = r \|\nabla \mathcal{L}(\theta)\|_2 \end{aligned} \quad (5.15)$$

(cf. Definition 3.3.4). Furthermore, observe that (5.14) shows that for all $c \in \mathbb{R}$ it holds that

$$\mathcal{G}(c(\nabla \mathcal{L})(\theta)) = \langle (\nabla \mathcal{L})(\theta), c(\nabla \mathcal{L})(\theta) \rangle = c \|\nabla \mathcal{L}(\theta)\|_2^2. \quad (5.16)$$

Combining this and (5.15) proves item (i) and item (ii). The proof of Lemma 5.2.2 is thus complete. \square

Lemma 5.2.3. Let $\mathfrak{d} \in \mathbb{N}$, $\Theta \in C([0, \infty), \mathbb{R}^{\mathfrak{d}})$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ and assume for all $t \in [0, \infty)$ that $\Theta_t = \Theta_0 - \int_0^t (\nabla \mathcal{L})(\Theta_s) ds$. 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 $\dot{\Theta}_t = -(\nabla \mathcal{L})(\Theta_t)$, and
- (iii) it holds for all $t \in [0, \infty)$ that

$$\mathcal{L}(\Theta_t) = \mathcal{L}(\Theta_0) - \int_0^t \|\nabla \mathcal{L}(\Theta_s)\|_2^2 ds \quad (5.17)$$

(cf. Definition 3.3.4).

Proof of Lemma 5.2.3. Note 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

$$\mathcal{L}(\Theta_t) = \mathcal{L}(\Theta_0) + \int_0^t \langle (\nabla \mathcal{L})(\Theta_s), \dot{\Theta}_s \rangle ds = \mathcal{L}(\Theta_0) - \int_0^t \|\nabla \mathcal{L}(\Theta_s)\|_2^2 ds \quad (5.18)$$

(cf. Definitions 1.4.7 and 3.3.4). This establishes item (iii). The proof of Lemma 5.2.3 is thus complete. \square

Corollary 5.2.4 (Illustration for the negative GF). *Let $\mathfrak{d} \in \mathbb{N}$, $\Theta \in C([0, \infty), \mathbb{R}^{\mathfrak{d}})$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ and assume for all $t \in [0, \infty)$ that $\Theta(t) = \Theta(0) - \int_0^t (\nabla \mathcal{L})(\Theta(s)) ds$. 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*

$$(\mathcal{L} \circ \Theta)'(t) = -\|(\nabla \mathcal{L})(\Theta(t))\|_2^2, \quad (5.19)$$

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*

$$(\mathcal{L} \circ \Theta)'(\tau) \leq (\mathcal{L} \circ \Xi)'(\tau) \quad (5.20)$$

(cf. Definition 3.3.4).

Proof of Corollary 5.2.4. Observe that Lemma 5.2.3 and the fundamental theorem of calculus imply items (i) and (ii). Note 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

$$\begin{aligned} (\mathcal{L} \circ \Xi)'(t) &= [\mathcal{L}'(\Xi(t))](\Xi'(t)) \\ &\geq \inf_{v \in \{w \in \mathbb{R}^{\mathfrak{d}} : \|w\|_2 = \|\Xi'(t)\|_2\}} [\mathcal{L}'(\Xi(t))](v) \\ &= -\|\Xi'(t)\|_2 \|(\nabla \mathcal{L})(\Xi(t))\|_2 \end{aligned} \quad (5.21)$$

(cf. Definition 3.3.4). Lemma 5.2.3 therefore 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

$$\begin{aligned} (\mathcal{L} \circ \Xi)'(\tau) &\geq -\|\Xi'(\tau)\|_2 \|(\nabla \mathcal{L})(\Xi(\tau))\|_2 \geq -\|\Theta'(\tau)\|_2 \|(\nabla \mathcal{L})(\Theta(\tau))\|_2 \\ &= -\|(\nabla \mathcal{L})(\Theta(\tau))\|_2^2 = (\mathcal{L} \circ \Theta)'(\tau). \end{aligned} \quad (5.22)$$

This establishes item (iii). The proof of Corollary 5.2.4 is thus complete. \square

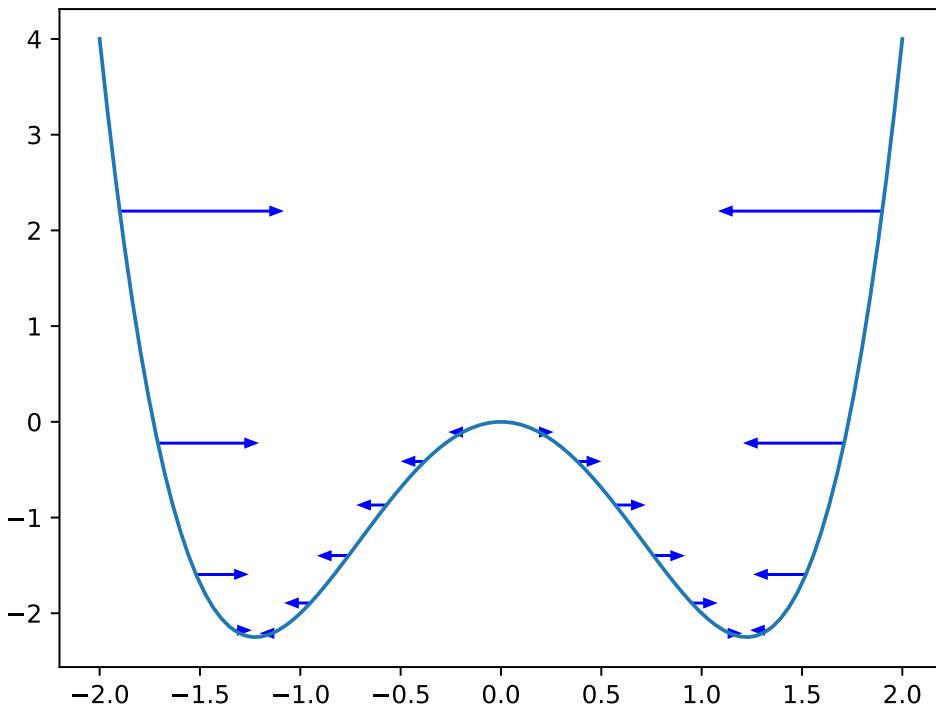


Figure 5.1 ([plots/gradient_plot1.pdf](#)): Illustration of negative gradients in a one-dimensional example. The plot shows the graph of the function $[-2, 2] \ni x \mapsto x^4 - 3x^2 \in \mathbb{R}$ with the value of the negative gradient, scaled by $\frac{1}{20}$, indicated by horizontal arrows at several points. The PYTHON code used to produce this plot is given in Source code 5.1.

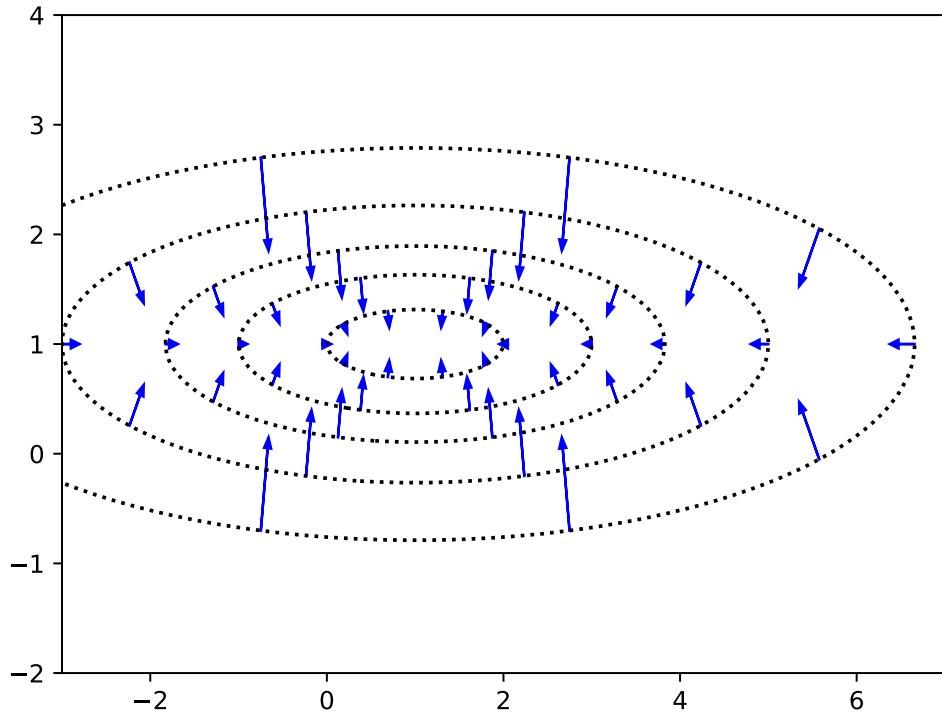


Figure 5.2 ([plots/gradient_plot2.pdf](#)): Illustration of negative gradients in a two-dimensional example. The plot shows contour lines of the function $\mathbb{R}^2 \ni (x, y) \mapsto \frac{1}{2}|x - 1|^2 + 5|y - 1|^2 \in \mathbb{R}$ with arrows indicating the direction and magnitude, scaled by $\frac{1}{20}$, 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.

```

1 import numpy as np
2 import matplotlib.pyplot as plt
3
4 def f(x):
5     return x**4 - 3 * x**2
6
7 def nabla_f(x):
8     return 4 * x**3 - 6 * x
9
10 plt.figure()
11 # Plot graph of f
12 x = np.linspace(-2, 2, 100)
13

```

```

14 plt.plot(x,f(x))
15
16 # Plot arrows
17 for x in np.linspace(-1.9,1.9,21):
18     d = nabla_f(x)
19     plt.arrow(x, f(x), -.05 * d, 0,
20               length_includes_head=True, head_width=0.08,
21               head_length=0.05, color='b')
22
23 plt.savefig("../plots/gradient_plot1.pdf")

```

Source code 5.1 ([code/gradient_plot1.py](#)): PYTHON code used to create Figure 5.1

```

1 import numpy as np
2 import matplotlib.pyplot as plt
3
4 K = [1., 10.]
5 vartheta = np.array([1., 1.])
6
7 def f(x, y):
8     result = K[0] / 2. * np.abs(x - vartheta[0])**2 \
9     + K[1] / 2. * np.abs(y - vartheta[1])**2
10    return result
11
12 def nabla_f(x):
13     return K * (x - vartheta)
14
15 plt.figure()
16
17 # Plot contour lines of f
18 x = np.linspace(-3., 7., 100)
19 y = np.linspace(-2., 4., 100)
20 X, Y = np.meshgrid(x, y)
21 Z = f(X, Y)
22 cp = plt.contour(X, Y, Z, colors="black",
23                   levels = [0.5,2,4,8,16],
24                   linestyles=":")
25
26 # Plot arrows along contour lines
27 for l in [0.5,2,4,8,16]:
28     for d in np.linspace(0, 2.*np.pi, 10, endpoint=False):
29         x = np.cos(d) / ((K[0] / (2*l))**.5) + vartheta[0]
30         y = np.sin(d) / ((K[1] / (2*l))**.5) + vartheta[1]
31         grad = nabla_f(np.array([x,y]))
32         plt.arrow(x, y, -.05 * grad[0], -.05 * grad[1],
33                   length_includes_head=True, head_width=.08,
34                   head_length=.1, color='b')
35
36 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 $A_1: \mathbb{R}^{l_1} \rightarrow \mathbb{R}^{l_1} \times \mathbb{R}^{l_2}$ and $A_2: \mathbb{R}^{l_2} \rightarrow \mathbb{R}^{l_1} \times \mathbb{R}^{l_2}$ 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 $B_k: \mathbb{R}^{l_1} \times \mathbb{R}^{l_2} \rightarrow \mathbb{R}^{l_k}$ 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 $F_k: \mathbb{R}^{\mathfrak{d}_k} \rightarrow \mathbb{R}^{l_k}$ be differentiable, and let $f: \mathbb{R}^{\mathfrak{d}_1} \times \mathbb{R}^{\mathfrak{d}_2} \rightarrow \mathbb{R}^{l_1} \times \mathbb{R}^{l_2}$ satisfy for all $x_1 \in \mathbb{R}^{\mathfrak{d}_1}$, $x_2 \in \mathbb{R}^{\mathfrak{d}_2}$ that

$$f(x_1, x_2) = (F_1(x_1), F_2(x_2)). \quad (5.23)$$

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. Observe 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

$$\begin{aligned} (A_1 \circ F_1 \circ B_1 + A_2 \circ F_2 \circ B_2)(x_1, x_2) &= (A_1 \circ F_1)(x_1) + (A_2 \circ F_2)(x_2) \\ &= (F_1(x_1), 0) + (0, F_2(x_2)) \\ &= (F_1(x_1), F_2(x_2)). \end{aligned} \quad (5.24)$$

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. \square

Lemma 5.3.2. Let $\mathfrak{d}_1, \mathfrak{d}_2, l_0, l_1, l_2 \in \mathbb{N}$, let $A: \mathbb{R}^{\mathfrak{d}_1} \times \mathbb{R}^{\mathfrak{d}_2} \times \mathbb{R}^{l_0} \rightarrow \mathbb{R}^{\mathfrak{d}_2} \times \mathbb{R}^{\mathfrak{d}_1+l_0}$ and $B: \mathbb{R}^{\mathfrak{d}_2} \times \mathbb{R}^{\mathfrak{d}_1+l_0} \rightarrow \mathbb{R}^{\mathfrak{d}_2} \times \mathbb{R}^{l_1}$ 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

$$A(\theta_1, \theta_2, x) = (\theta_2, (\theta_1, x)) \quad \text{and} \quad B(\theta_2, (\theta_1, x)) = (\theta_2, F_1(\theta_1, x)), \quad (5.25)$$

for every $k \in \{1, 2\}$ let $F_k: \mathbb{R}^{\mathfrak{d}_k} \times \mathbb{R}^{l_{k-1}} \rightarrow \mathbb{R}^{l_k}$ be differentiable, and let $f: \mathbb{R}^{\mathfrak{d}_1} \times \mathbb{R}^{\mathfrak{d}_2} \times \mathbb{R}^{l_0} \rightarrow \mathbb{R}^{l_2}$ 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

$$f(\theta_1, \theta_2, x) = (F_2(\theta_2, \cdot) \circ F_1(\theta_1, \cdot))(x). \quad (5.26)$$

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. Note 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

$$f(\theta_1, \theta_2, x) = F_2(\theta_2, F_1(\theta_1, x)) = F_2(B(\theta_2, (\theta_1, x))) = F_2(B(A(\theta_1, \theta_2, x))). \quad (5.27)$$

Observe 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. \square

5.3.2 On the differentiability of realizations of ANNs

Lemma 5.3.3 (Differentiability of realization functions of ANNs). *Let $L \in \mathbb{N}$, $l_0, l_1, \dots, l_L \in \mathbb{N}$, for every $k \in \{1, 2, \dots, L\}$ let $\mathfrak{d}_k = l_k(l_{k-1} + 1)$, for every $k \in \{1, 2, \dots, L\}$ let $\Psi_k: \mathbb{R}^{l_k} \rightarrow \mathbb{R}^{l_k}$ be differentiable, and for every $k \in \{1, 2, \dots, L\}$ let $F_k: \mathbb{R}^{\mathfrak{d}_k} \times \mathbb{R}^{l_{k-1}} \rightarrow \mathbb{R}^{l_k}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}_k}$, $x \in \mathbb{R}^{l_{k-1}}$ that*

$$F_k(\theta, x) = \Psi_k(\mathcal{A}_{l_k, l_{k-1}}^{\theta, 0}(x)) \quad (5.28)$$

(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

$$(\mathcal{N}_{\Psi_1, \Psi_2, \dots, \Psi_L}^{(\theta_1, \theta_2, \dots, \theta_L), l_0})(x) = (F_L(\theta_L, \cdot) \circ F_{L-1}(\theta_{L-1}, \cdot) \circ \dots \circ F_1(\theta_1, \cdot))(x) \quad (5.29)$$

and

- (ii) it holds that

$$\mathbb{R}^{\mathfrak{d}_1 + \mathfrak{d}_2 + \dots + \mathfrak{d}_L} \times \mathbb{R}^{l_0} \ni (\theta, x) \mapsto (\mathcal{N}_{\Psi_1, \Psi_2, \dots, \Psi_L}^{\theta, l_0})(x) \in \mathbb{R}^{l_L} \quad (5.30)$$

is differentiable

(cf. Definition 1.1.3).

Proof of Lemma 5.3.3. Note that (1.1) shows that for all $\theta_1 \in \mathbb{R}^{\mathfrak{d}_1}$, $\theta_2 \in \mathbb{R}^{\mathfrak{d}_2}$, ..., $\theta_L \in \mathbb{R}^{\mathfrak{d}_L}$,

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

$$\mathcal{A}_{l_k, l_{k-1}}^{(\theta_1, \theta_2, \dots, \theta_L), \sum_{j=1}^{k-1} \mathfrak{d}_j} = \mathcal{A}_{l_k, l_{k-1}}^{\theta_k, 0}. \quad (5.31)$$

Hence, we obtain that for all $\theta_1 \in \mathbb{R}^{\mathfrak{d}_1}, \theta_2 \in \mathbb{R}^{\mathfrak{d}_2}, \dots, \theta_L \in \mathbb{R}^{\mathfrak{d}_L}, k \in \{1, 2, \dots, L\}$ it holds that

$$F_k(\theta_k, x) = (\Psi_k \circ \mathcal{A}_{l_k, l_{k-1}}^{(\theta_1, \theta_2, \dots, \theta_L), \sum_{j=1}^{k-1} \mathfrak{d}_j})(x). \quad (5.32)$$

Combining this with (1.5) establishes item (i). Observe that the assumption that for all $k \in \{1, 2, \dots, L\}$ 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 $\mathbb{R}^{m(n+1)} \times \mathbb{R}^n \ni (\theta, x) \mapsto \mathcal{A}_{m,n}^{\theta, 0}(x) \in \mathbb{R}^m$ is differentiable, and the chain rule ensure that for all $k \in \{1, 2, \dots, L\}$ it holds that F_k is differentiable. Lemma 5.3.2 and induction hence prove that

$$\begin{aligned} \mathbb{R}^{\mathfrak{d}_1} \times \mathbb{R}^{\mathfrak{d}_2} \times \dots \times \mathbb{R}^{\mathfrak{d}_L} \times \mathbb{R}^{l_0} &\ni (\theta_1, \theta_2, \dots, \theta_L, x) \\ &\mapsto (F_L(\theta_L, \cdot) \circ F_{L-1}(\theta_{L-1}, \cdot) \circ \dots \circ F_1(\theta_1, \cdot))(x) \in \mathbb{R}^{l_L} \end{aligned} \quad (5.33)$$

is differentiable. This and item (i) prove item (ii). The proof of Lemma 5.3.3 is thus complete. \square

Lemma 5.3.4 (Differentiability of the empirical risk function). *Let $L, \mathfrak{d} \in \mathbb{N} \setminus \{1\}$, $M, l_0, l_1, \dots, l_L \in \mathbb{N}$, $x_1, x_2, \dots, x_M \in \mathbb{R}^{l_0}$, $y_1, y_2, \dots, y_M \in \mathbb{R}^{l_L}$ satisfy $\mathfrak{d} = \sum_{k=1}^L l_k(l_{k-1} + 1)$, for every $k \in \{1, 2, \dots, L\}$ let $\Psi_k: \mathbb{R}^{l_k} \rightarrow \mathbb{R}^{l_k}$ be differentiable, and let $\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that*

$$\mathcal{L}(\theta) = \frac{1}{M} \left[\sum_{m=1}^M \mathbf{L}((\mathcal{N}_{\Psi_1, \Psi_2, \dots, \Psi_L}^{\theta, l_0})(x_m), y_m) \right] \quad (5.34)$$

(cf. Definition 1.1.3). Then \mathcal{L} is differentiable.

Proof of Lemma 5.3.4. Note that Lemma 5.3.3 and Lemma 5.3.1 (applied with $\mathfrak{d}_1 \curvearrowright \mathfrak{d} + l_0$, $\mathfrak{d}_2 \curvearrowright l_L$, $l_1 \curvearrowright l_L$, $l_2 \curvearrowright l_L$, $F_1 \curvearrowright (\mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{l_0} \ni (\theta, x) \mapsto (\mathcal{N}_{\Psi_1, \Psi_2, \dots, \Psi_L}^{\theta, l_0})(x) \in \mathbb{R}^{l_L})$, $F_2 \curvearrowright \text{id}_{\mathbb{R}^{l_L}}$ in the notation of Lemma 5.3.1) demonstrate that

$$\mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{l_0} \times \mathbb{R}^{l_L} \ni (\theta, x, y) \mapsto ((\mathcal{N}_{\Psi_1, \Psi_2, \dots, \Psi_L}^{\theta, l_0})(x), y) \in \mathbb{R}^{l_L} \times \mathbb{R}^{l_L} \quad (5.35)$$

is differentiable. The assumption that \mathbf{L} is differentiable and the chain rule therefore ensure that for all $x \in \mathbb{R}^{l_0}$, $y \in \mathbb{R}^{l_L}$ it holds that

$$\mathbb{R}^{\mathfrak{d}} \ni \theta \mapsto \mathbf{L}((\mathcal{N}_{\Psi_1, \Psi_2, \dots, \Psi_L}^{\theta, l_0})(x_m), y_m) \in \mathbb{R} \quad (5.36)$$

is differentiable. This implies that \mathcal{L} is differentiable. The proof of Lemma 5.3.4 is thus complete. \square

Lemma 5.3.5. Let $a: \mathbb{R} \rightarrow \mathbb{R}$ 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. Observe that the assumption that a is differentiable, Lemma 5.3.1, and induction prove 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. \square

Corollary 5.3.6. Let $L, \mathfrak{d} \in \mathbb{N} \setminus \{1\}$, $M, l_0, l_1, \dots, l_L \in \mathbb{N}$, $x_1, x_2, \dots, x_M \in \mathbb{R}^{l_0}$, $y_1, y_2, \dots, y_M \in \mathbb{R}^{l_L}$ satisfy $\mathfrak{d} = \sum_{k=1}^L l_k(l_{k-1} + 1)$, let $a: \mathbb{R} \rightarrow \mathbb{R}$ and $\mathbf{L}: \mathbb{R}^{l_L} \times \mathbb{R}^{l_L} \rightarrow \mathbb{R}$ be differentiable, and let $\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that

$$\mathcal{L}(\theta) = \frac{1}{M} \left[\sum_{m=1}^M \mathbf{L}\left((\mathcal{N}_{\mathfrak{M}_{a,l_1}, \mathfrak{M}_{a,l_2}, \dots, \mathfrak{M}_{a,l_{L-1}}, \text{id}_{\mathbb{R}^{l_L}}}^{\theta, l_0})(x_m), y_m\right) \right] \quad (5.37)$$

(cf. Definitions 1.1.3 and 1.2.1). Then \mathcal{L} is differentiable.

Proof of Corollary 5.3.6. Note that Lemma 5.3.5, and Lemma 5.3.4 establish that \mathcal{L} is differentiable. The proof of Corollary 5.3.6 is thus complete. \square

Corollary 5.3.7. Let $L, \mathfrak{d} \in \mathbb{N} \setminus \{1\}$, $M, l_0, l_1, \dots, l_L \in \mathbb{N}$, $x_1, x_2, \dots, x_M \in \mathbb{R}^{l_0}$, $y_1, y_2, \dots, y_M \in (0, \infty)^{l_L}$ satisfy $\mathfrak{d} = \sum_{k=1}^L l_k(l_{k-1} + 1)$, let A be the l_L -dimensional softmax activation function, let $a: \mathbb{R} \rightarrow \mathbb{R}$ and $\mathbf{L}: (0, \infty)^{l_L} \times (0, \infty)^{l_L} \rightarrow \mathbb{R}$ be differentiable, and let $\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that

$$\mathcal{L}(\theta) = \frac{1}{M} \left[\sum_{m=1}^M \mathbf{L}\left((\mathcal{N}_{\mathfrak{M}_{a,l_1}, \mathfrak{M}_{a,l_2}, \dots, \mathfrak{M}_{a,l_{L-1}}, A}^{\theta, l_0})(x_m), y_m\right) \right] \quad (5.38)$$

(cf. Definitions 1.1.3, 1.2.1, and 1.2.47 and Lemma 1.2.48). Then \mathcal{L} is differentiable.

Proof of Corollary 5.3.7. Observe that Lemma 5.3.5, the fact that A is differentiable, and Lemma 5.3.4 show that \mathcal{L} is differentiable. The proof of Corollary 5.3.7 is thus complete. \square

5.4 Loss functions

5.4.1 Absolute error loss

Definition 5.4.1. Let $d \in \mathbb{N}$ and let $\|\cdot\|: \mathbb{R}^d \rightarrow [0, \infty)$ 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 $\mathbf{L}: \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ is the function from

$\mathbb{R}^d \times \mathbb{R}^d$ to \mathbb{R} which satisfies for all $x, y \in \mathbb{R}^d$ that

$$\mathbf{L}(x, y) = \|x - y\|. \quad (5.39)$$

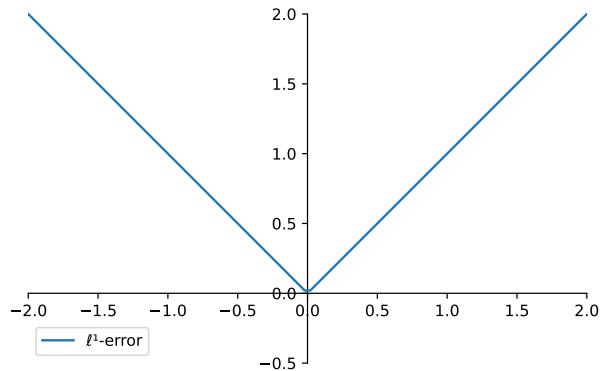


Figure 5.3 ([plots/l1loss.pdf](#)): A plot of the function $\mathbb{R} \ni x \mapsto \mathbf{L}(x, 0) \in [0, \infty)$ where \mathbf{L} is the ℓ^1 -error loss function based on $\mathbb{R} \ni x \mapsto |x| \in [0, \infty)$ (cf. Definition 5.4.1).

```

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((-2,2), (-.5,2))
7
8 x = np.linspace(-2, 2, 100)
9
10 mae_loss = tf.keras.losses.MeanAbsoluteError(
11     reduction=tf.keras.losses.Reduction.NONE)
12 zero = tf.zeros([100,1])
13
14 ax.plot(x, mae_loss(x.reshape([100,1]),zero),
15          label=' $\ell^1$ -error')
16 ax.legend()
17
18 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 $\|\cdot\|: \mathbb{R}^d \rightarrow [0, \infty)$ 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 $\mathbf{L}: \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ is the function from $\mathbb{R}^d \times \mathbb{R}^d$ to \mathbb{R} which satisfies for all $x, y \in \mathbb{R}^d$ that

$$\mathbf{L}(x, y) = \|x - y\|^2. \quad (5.40)$$

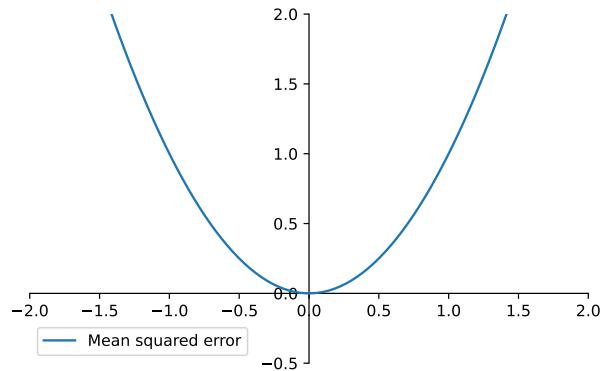


Figure 5.4 ([plots/mseloss.pdf](#)): A plot of the function $\mathbb{R} \ni x \mapsto \mathbf{L}(x, 0) \in [0, \infty)$ where \mathbf{L} is the mean squared error loss function based on $\mathbb{R} \ni x \mapsto |x| \in [0, \infty)$ (cf. Definition 5.4.2).

```

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((-2,2), (-.5,2))
7
8 x = np.linspace(-2, 2, 100)
9
10 mse_loss = tf.keras.losses.MeanSquaredError(
11     reduction=tf.keras.losses.Reduction.NONE)
12 zero = tf.zeros([100,1])
13
14 ax.plot(x, mse_loss(x.reshape([100,1]), zero),
15          label='Mean squared error')
16 ax.legend()
17
18 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

$$\mathbf{L}(u, v) = \mathbf{L}(x, y) + \mathbf{L}'(x, y)(u - x, v - y) + \frac{1}{2}\mathbf{L}^{(2)}(x, y)((u - x, v - y), (u - x, v - y)). \quad (5.41)$$

Proof of Lemma 5.4.3. Note that (5.40) implies that for all $x = (x_1, \dots, x_d), y = (y_1, \dots, y_d) \in \mathbb{R}^d$ it holds that

$$\mathbf{L}(x, y) = \|x - y\|_2^2 = \langle x - y, x - y \rangle = \sum_{i=1}^d (x_i - y_i)^2. \quad (5.42)$$

Hence, 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

$$(\nabla \mathbf{L})(x, y) = (2(x - y), -2(x - y)) \in \mathbb{R}^{2d}. \quad (5.43)$$

This implies that for all $x, y, h, k \in \mathbb{R}^d$ it holds that

$$\mathbf{L}'(x, y)(h, k) = \langle 2(x - y), h \rangle + \langle -2(x - y), k \rangle = 2\langle x - y, h - k \rangle. \quad (5.44)$$

Furthermore, observe 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

$$(\text{Hess}_{(x,y)} \mathbf{L}) = \begin{pmatrix} 2\mathbf{I}_d & -2\mathbf{I}_d \\ -2\mathbf{I}_d & 2\mathbf{I}_d \end{pmatrix}. \quad (5.45)$$

Therefore, we obtain that for all $x, y, h, k \in \mathbb{R}^d$ it holds that

$$\mathbf{L}^{(2)}(x, y)((h, k), (h, k)) = 2\langle h, h \rangle - 2\langle h, k \rangle - 2\langle k, h \rangle + 2\langle k, k \rangle = 2\|h - k\|_2^2. \quad (5.46)$$

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

$$\begin{aligned} & \mathbf{L}(x, y) + \mathbf{L}'(x, y)(h, k) + \frac{1}{2}\mathbf{L}^{(2)}(x, y)((h, k), (h, k)) \\ &= \|x - y\|_2^2 + 2\langle x - y, h - k \rangle + \|h - k\|_2^2 \\ &= \|x - y + (h - k)\|_2^2 \\ &= \mathbf{L}(x + h, y + k). \end{aligned} \quad (5.47)$$

This implies items (i) and (ii). The proof of Lemma 5.4.3 is thus complete. \square

5.4.3 Huber error loss

Definition 5.4.4. Let $d \in \mathbb{N}$, $\delta \in [0, \infty)$ and let $\|\cdot\|: \mathbb{R}^d \rightarrow [0, \infty)$ 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 $\mathbf{L}: \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ is the function from $\mathbb{R}^d \times \mathbb{R}^d$ to \mathbb{R} which satisfies for all $x, y \in \mathbb{R}^d$ that

$$\mathbf{L}(x, y) = \begin{cases} \frac{1}{2}\|x - y\|^2 & : \|x - y\| \leq \delta \\ \delta(\|x - y\| - \frac{\delta}{2}) & : \|x - y\| > \delta. \end{cases} \quad (5.48)$$

Lemma 5.4.5. Let $\delta \in [0, \infty)$ and let $\mathbf{H}: \mathbb{R} \rightarrow [0, \infty)$ satisfy for all $z \in \mathbb{R}$ that

$$\mathbf{H}(z) = \begin{cases} \frac{1}{2}z^2 & : z \leq \delta \\ \delta(z - \frac{\delta}{2}) & : z > \delta. \end{cases} \quad (5.49)$$

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

$$f(z) = \frac{1}{2}z^2 \quad \text{and} \quad g(z) = \delta(z - \frac{\delta}{2}). \quad (5.50)$$

Note that (5.50) demonstrates that

$$g(\delta) = \delta(\delta - \frac{\delta}{2}) = \frac{1}{2}\delta^2 = f(\delta). \quad (5.51)$$

Combining this with the fact that for all $z \in \mathbb{R}$ it holds that

$$\mathbf{H}(z) = \begin{cases} f(z) & : z \leq \delta \\ g(z) & : z > \delta \end{cases} \quad (5.52)$$

proves that \mathbf{H} is continuous. The proof of Lemma 5.4.5 is thus complete. \square

Corollary 5.4.6. Let $d \in \mathbb{N}$, $\delta \in [0, \infty)$, let $\|\cdot\|: \mathbb{R}^d \rightarrow [0, \infty)$ 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 $\mathbf{H}: \mathbb{R} \rightarrow [0, \infty)$ satisfy for all $z \in \mathbb{R}$ that

$$\mathbf{H}(z) = \begin{cases} \frac{1}{2}z^2 & : z \leq \delta \\ \delta(z - \frac{\delta}{2}) & : z > \delta. \end{cases} \quad (5.53)$$

Observe that (5.48) ensures that for all $x, y \in \mathbb{R}^d$ it holds that

$$\mathbf{L}(x, y) = \mathbf{H}(\|x - y\|). \quad (5.54)$$

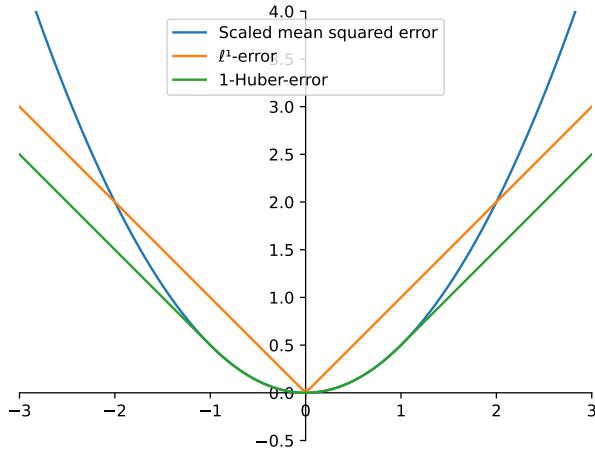


Figure 5.5 ([plots/huberloss.pdf](#)): A plot of the functions $\mathbb{R} \ni x \mapsto \mathbf{L}_i(x, 0) \in [0, \infty)$, $i \in \{1, 2, 3\}$, where \mathbf{L}_0 is the mean squared error loss function based on $\mathbb{R} \ni x \mapsto |x| \in [0, \infty)$, where $\mathbf{L}_1: \mathbb{R}^d \times \mathbb{R}^d \rightarrow [0, \infty)$ 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, note that Lemma 5.4.5 implies 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) establishes that \mathbf{L} is continuous. The proof of Corollary 5.4.6 is thus complete. \square

```

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), (-.5,4))
7
8 x = np.linspace(-3, 3, 100)
9
10 mse_loss = tf.keras.losses.MeanSquaredError(
11     reduction=tf.keras.losses.Reduction.NONE)
12 mae_loss = tf.keras.losses.MeanAbsoluteError(
13     reduction=tf.keras.losses.Reduction.NONE)
14 huber_loss = tf.keras.losses.Huber(
15     reduction=tf.keras.losses.Reduction.NONE)
16
17 zero = tf.zeros([100,1])
18
19 ax.plot(x, mse_loss(x.reshape([100,1]),zero)/2.,
20         label='Scaled mean squared error')
21 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')

```

```

22     label=' $\ell^1$ -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 $\mathbf{L}: [0, \infty)^d \times [0, \infty)^d \rightarrow (-\infty, \infty]$ 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

$$\mathbf{L}(x, y) = - \sum_{i=1}^d \lim_{\mathbf{x} \searrow x_i} [\ln(\mathbf{x}) y_i]. \quad (5.55)$$

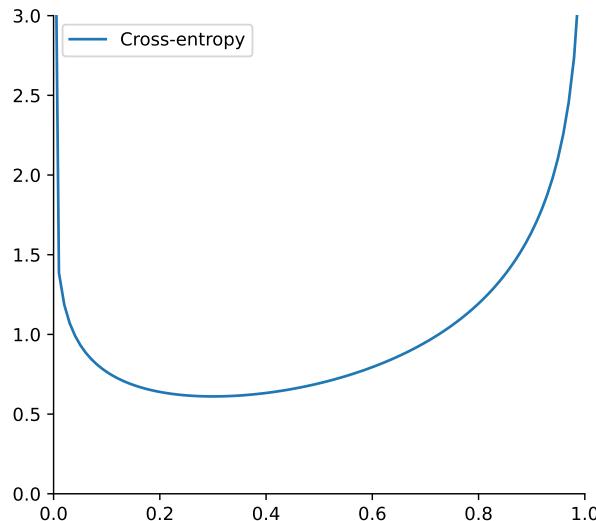


Figure 5.6 ([plots/crossentropyloss.pdf](#)): A plot of the function $(0, 1) \ni x \mapsto \mathbf{L}\left((x, 1-x), \left(\frac{3}{10}, \frac{7}{10}\right)\right) \in \mathbb{R}$ where \mathbf{L} is the 2-dimensional cross-entropy loss function (cf. Definition 5.4.7).

```

1 import numpy as np
2 import tensorflow as tf

```

```

3 import matplotlib.pyplot as plt
4 import plot_util
5
6 ax = plot_util.setup_axis((0,1), (0,3))
7
8 ax.set_aspect(.3)
9
10 x = np.linspace(0, 1, 100)
11
12 cce_loss = tf.keras.losses.CategoricalCrossentropy(
13     reduction=tf.keras.losses.Reduction.NONE)
14 y = tf.constant([[0.3, 0.7]] * 100, shape=(100, 2))
15
16 X = tf.stack([x, 1-x], axis=1)
17
18 ax.plot(x, cce_loss(y, X), label='Cross-entropy')
19 ax.legend()
20
21 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 $x = (x_1, \dots, x_d)$, $y = (y_1, \dots, y_d) \in [0, \infty)^d$ that

$$(\mathbf{L}(x, y) = \infty) \leftrightarrow (\exists i \in \{1, 2, \dots, d\}: [(x_i = 0) \wedge (y_i \neq 0)]), \quad (5.56)$$

(ii) it holds for all $x = (x_1, \dots, x_d)$, $y = (y_1, \dots, y_d) \in [0, \infty)^d$ with $\forall i \in \{1, 2, \dots, d\}: [(x_i \neq 0) \vee (y_i = 0)]$ that

$$\mathbf{L}(x, y) = - \sum_{\substack{i \in \{1, 2, \dots, d\}, \\ y_i \neq 0}} \ln(x_i) y_i \in \mathbb{R}, \quad (5.57)$$

and

(iii) it holds for all $x = (x_1, \dots, x_d) \in (0, \infty)^d$, $y = (y_1, \dots, y_d) \in [0, \infty)^d$ that

$$\mathbf{L}(x, y) = - \sum_{i=1}^d \ln(x_i) y_i \in \mathbb{R}. \quad (5.58)$$

Proof of Lemma 5.4.8. Observe that (5.55) and the fact that for all $a, b \in [0, \infty)$ it holds that

$$\lim_{\mathbf{a} \searrow a} [\ln(\mathbf{a})b] = \begin{cases} 0 & : b = 0 \\ \ln(a)b & : (a \neq 0) \wedge (b \neq 0) \\ -\infty & : (a = 0) \wedge (b \neq 0) \end{cases} \quad (5.59)$$

prove items (i), (ii), and (iii). The proof of Lemma 5.4.8 is thus complete. \square

Lemma 5.4.9. Let $d \in \mathbb{N}$, let \mathbf{L} be the d -dimensional cross-entropy loss function, let $x = (x_1, \dots, x_d)$, $y = (y_1, \dots, y_d) \in [0, \infty)^d$ satisfy $\sum_{i=1}^d x_i = \sum_{i=1}^d y_i$ and $x \neq y$, and let $f: [0, 1] \rightarrow (-\infty, \infty]$ satisfy for all $h \in [0, 1]$ that

$$f(h) = \mathbf{L}(x + h(y - x), y) \quad (5.60)$$

(cf. Definition 5.4.7). Then f is strictly decreasing.

Proof of Lemma 5.4.9. Throughout this proof, let $g: [0, 1] \rightarrow (-\infty, \infty]$ satisfy for all $h \in [0, 1)$ that

$$g(h) = f(1 - h) \quad (5.61)$$

and let $J = \{i \in \{1, 2, \dots, d\}: y_i \neq 0\}$. Note that (5.60) shows that for all $h \in [0, 1)$ it holds that

$$g(h) = \mathbf{L}(x + (1 - h)(y - x), y) = \mathbf{L}(y + h(x - y), y). \quad (5.62)$$

Furthermore, observe that the fact that for all $i \in J$ it holds that $x_i \in [0, \infty)$ and $y_i \in (0, \infty)$ demonstrates that for all $i \in J$, $h \in [0, 1)$ it holds that

$$y_i + h(x_i - y_i) = (1 - h)y_i + hx_i \geq (1 - h)y_i > 0. \quad (5.63)$$

This, (5.62), and item (ii) in Lemma 5.4.8 ensure that for all $h \in [0, 1)$ it holds that

$$g(h) = - \sum_{i \in J} \ln(y_i + h(x_i - y_i))y_i \in \mathbb{R}. \quad (5.64)$$

The chain rule hence implies 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

$$g'(h) = - \sum_{i \in J} \frac{y_i(x_i - y_i)}{y_i + h(x_i - y_i)}. \quad (5.65)$$

This and the chain rule establish that for all $h \in [0, 1)$ it holds that

$$g''(h) = \sum_{i \in J} \frac{y_i(x_i - y_i)^2}{(y_i + h(x_i - y_i))^2}. \quad (5.66)$$

Moreover, note that the fact that for all $z = (z_1, \dots, z_d) \in [0, \infty)^d$ with $\sum_{i=1}^d z_i = \sum_{i=1}^d y_i$ and $\forall i \in J: z_i = y_i$ it holds that

$$\begin{aligned} \sum_{i \in \{1, 2, \dots, d\} \setminus J} z_i &= \left[\sum_{i \in \{1, 2, \dots, d\}} z_i \right] - \left[\sum_{i \in J} z_i \right] \\ &= \left[\sum_{i \in \{1, 2, \dots, d\}} y_i \right] - \left[\sum_{i \in J} z_i \right] \\ &= \sum_{i \in J} (y_i - z_i) = 0 \end{aligned} \tag{5.67}$$

proves that for all $z = (z_1, \dots, z_d) \in [0, \infty)^d$ with $\sum_{i=1}^d z_i = \sum_{i=1}^d y_i$ and $\forall i \in J: z_i = y_i$ it holds that $z = y$. The assumption that $\sum_{i=1}^d x_i = \sum_{i=1}^d y_i$ and $x \neq y$ therefore ensures 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 hence demonstrates that for all $h \in (0, 1)$ it holds that

$$g'(h) = g'(0) + \int_0^h g''(\xi) d\xi > g'(0). \tag{5.69}$$

In addition, observe that (5.65) and the assumption that $\sum_{i=1}^d x_i = \sum_{i=1}^d y_i$ imply that

$$\begin{aligned} g'(0) &= - \sum_{i \in J} \frac{y_i(x_i - y_i)}{y_i} = \sum_{i \in J} (y_i - x_i) = \left[\sum_{i \in J} y_i \right] - \left[\sum_{i \in J} x_i \right] \\ &= \left[\sum_{i \in \{1, 2, \dots, d\}} y_i \right] - \left[\sum_{i \in J} x_i \right] = \left[\sum_{i \in \{1, 2, \dots, d\}} x_i \right] - \left[\sum_{i \in J} x_i \right] = \left[\sum_{i \in \{1, 2, \dots, d\} \setminus J} x_i \right] \geq 0. \end{aligned} \tag{5.70}$$

Combining this and (5.69) establishes that for all $h \in (0, 1)$ it holds that

$$g'(h) > 0. \tag{5.71}$$

Therefore, we obtain that g is strictly increasing. This and (5.61) prove that $f|_{(0,1]}$ is strictly decreasing. Next note that (5.61) and (5.64) ensure 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

$$f(0) = \infty. \tag{5.73}$$

Observe 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

$$f(0) < \infty. \quad (5.74)$$

Note that (5.74) and items (i) and (ii) in Lemma 5.4.8 demonstrate that

$$0 \notin \cup_{i \in J} \{x_i\} \quad \text{and} \quad f(0) = - \sum_{i \in J} \ln(x_i + 0(y_i - x_i))y_i \in \mathbb{R}. \quad (5.75)$$

Combining this with (5.72) implies that $f([0, 1]) \subseteq \mathbb{R}$ and

$$([0, 1] \ni h \mapsto f(h) \in \mathbb{R}) \in C([0, 1], \mathbb{R}). \quad (5.76)$$

This and the fact that $f|_{(0,1]}$ is strictly decreasing establish 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. \square

Corollary 5.4.10. Let $d \in \mathbb{N}$, let $A = \{x = (x_1, \dots, x_d) \in [0, 1]^d : \sum_{i=1}^d x_i = 1\}$, 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

$$\{x \in A : \mathbf{L}(x, y) = \inf_{z \in A} \mathbf{L}(z, y)\} = \{y\} \quad (5.77)$$

and

(ii) it holds that

$$\inf_{z \in A} \mathbf{L}(z, y) = \mathbf{L}(y, y) = - \sum_{\substack{i \in \{1, 2, \dots, d\}, \\ y_i \neq 0}} \ln(y_i)y_i. \quad (5.78)$$

Proof of Corollary 5.4.10. Observe that Lemma 5.4.9 shows that for all $x \in A \setminus \{y\}$ it holds that

$$\mathbf{L}(x, y) = \mathbf{L}(x + 0(y - x), y) > \mathbf{L}(x + 1(y - x), y) = \mathbf{L}(y, y). \quad (5.79)$$

This and item (ii) in Lemma 5.4.8 prove items (i) and (ii). The proof of Corollary 5.4.10 is thus complete. \square

5.4.5 Kullback–Leibler divergence loss

Lemma 5.4.11. Let $z \in (0, \infty)$. Then

(i) it holds that

$$\liminf_{x \searrow 0} |\ln(x)x| = 0 \quad (5.80)$$

and

(ii) it holds for all $y \in [0, \infty)$ that

$$\liminf_{\mathbf{y} \searrow \mathbf{y}} [\ln(\frac{z}{y})\mathbf{y}] = \limsup_{\mathbf{y} \searrow \mathbf{y}} [\ln(\frac{z}{y})\mathbf{y}] = \begin{cases} 0 & : y = 0 \\ \ln(\frac{z}{y})y & : y > 0 \end{cases} \quad (5.81)$$

Proof of Lemma 5.4.11. Throughout this proof, let $f: (0, \infty) \rightarrow \mathbb{R}$ and $g: (0, \infty) \rightarrow \mathbb{R}$ satisfy for all $x \in (0, \infty)$ that

$$f(x) = \ln(x^{-1}) \quad \text{and} \quad g(x) = x. \quad (5.82)$$

Note that the chain rule ensures that for all $x \in (0, \infty)$ it holds that f is differentiable and

$$f'(x) = -x^{-2}(x^{-1})^{-1} = -x^{-1}. \quad (5.83)$$

Combining this, the fact that $\lim_{x \rightarrow \infty} |f(x)| = \infty = \lim_{x \rightarrow \infty} |g(x)|$, 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 $\lim_{x \rightarrow \infty} \frac{-x^{-1}}{1} = 0$ with l'Hôpital's rule shows that

$$\liminf_{x \rightarrow \infty} \frac{f(x)}{g(x)} = 0 = \limsup_{x \rightarrow \infty} \frac{f(x)}{g(x)}. \quad (5.84)$$

This demonstrates that

$$\liminf_{x \searrow 0} \frac{f(x^{-1})}{g(x^{-1})} = 0 = \limsup_{x \searrow 0} \frac{f(x^{-1})}{g(x^{-1})}. \quad (5.85)$$

The fact that for all $x \in (0, \infty)$ it holds that $\frac{f(x^{-1})}{g(x^{-1})} = \ln(x)x$ hence establishes item (i). Observe that item (i) and the fact that for all $x \in (0, \infty)$ it holds that $\ln(\frac{z}{x})x = \ln(z)x - \ln(x)x$ prove item (ii). The proof of Lemma 5.4.11 is thus complete. \square

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 $\mathbf{L}: [0, \infty)^d \times [0, \infty)^d \rightarrow (-\infty, \infty]$ 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

$$\mathbf{L}(x, y) = - \sum_{i=1}^d \lim_{\mathbf{x} \searrow x_i} \lim_{\mathbf{y} \searrow y_i} [\ln(\frac{\mathbf{x}}{\mathbf{y}})\mathbf{y}] \quad (5.86)$$

(cf. Lemma 5.4.11).

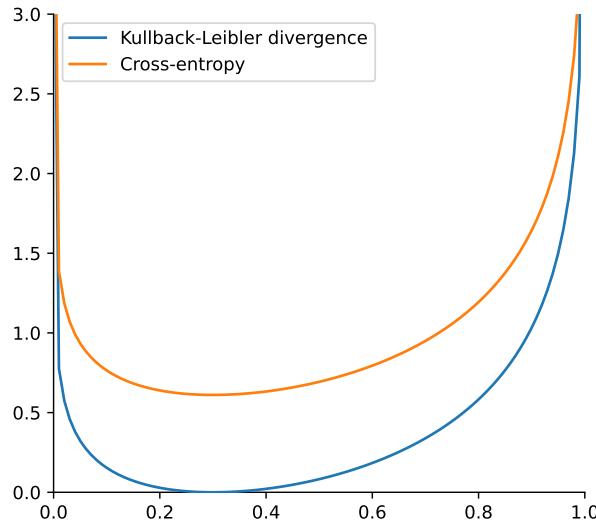


Figure 5.7 ([plots/kldloss.pdf](#)): A plot of the functions $(0, 1) \ni x \mapsto \mathbf{L}_i((x, 1 - x), (\frac{3}{10}, \frac{7}{10})) \in \mathbb{R}$, $i \in \{1, 2\}$, where \mathbf{L}_1 is the 2-dimensional Kullback–Leibler divergence loss function and where \mathbf{L}_2 is the 2-dimensional cross-entropy loss function (cf. Definitions 5.4.7 and 5.4.12).

```

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((0,1), (0,3))
7
8 ax.set_aspect(.3)
9
10 x = np.linspace(0, 1, 100)
11
12 kld_loss = tf.keras.losses.KLDivergence(
13     reduction=tf.keras.losses.Reduction.NONE)
14 cce_loss = tf.keras.losses.CategoricalCrossentropy(
15     reduction=tf.keras.losses.Reduction.NONE)
16 y = tf.constant([[0.3, 0.7]] * 100, shape=(100, 2))
17
18 X = tf.stack([x, 1-x], axis=1)
19
20 ax.plot(x, kld_loss(y, X), label='Kullback-Leibler divergence')
21 ax.plot(x, cce_loss(y, X), label='Cross-entropy')
22 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}_{CE} be the d -dimensional cross-entropy loss function, and let \mathbf{L}_{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

$$\mathbf{L}_{\text{CE}}(x, y) = \mathbf{L}_{\text{KLD}}(x, y) + \mathbf{L}_{\text{CE}}(y, y). \quad (5.87)$$

Proof of Lemma 5.4.13. Note that Lemma 5.4.11 implies that for all $a, b \in [0, \infty)$ it holds that

$$\begin{aligned}
 \lim_{\mathbf{a} \searrow a} \lim_{\mathbf{b} \searrow b} [\ln(\frac{\mathbf{a}}{\mathbf{b}})\mathbf{b}] &= \lim_{\mathbf{a} \searrow a} \lim_{\mathbf{b} \searrow b} [\ln(\mathbf{a})\mathbf{b} - \ln(\mathbf{b})\mathbf{b}] \\
 &= \lim_{\mathbf{a} \searrow a} \left[\ln(\mathbf{a})b - \lim_{\mathbf{b} \searrow b} [\ln(\mathbf{b})\mathbf{b}] \right] \\
 &= \left(\lim_{\mathbf{a} \searrow a} [\ln(\mathbf{a})b] \right) - \left(\lim_{\mathbf{b} \searrow b} [\ln(\mathbf{b})\mathbf{b}] \right).
 \end{aligned} \quad (5.88)$$

This and (5.86) ensure that for all $x = (x_1, \dots, x_d), y = (y_1, \dots, y_d) \in [0, \infty)^d$ it holds that

$$\begin{aligned}
 \mathbf{L}_{\text{KLD}}(x, y) &= - \sum_{i=1}^d \lim_{\mathbf{x} \searrow x_i} \lim_{\mathbf{y} \searrow y_i} [\ln(\frac{\mathbf{x}}{\mathbf{y}})\mathbf{y}] \\
 &= - \left(\sum_{i=1}^d \lim_{\mathbf{x} \searrow x_i} [\ln(\mathbf{x})y_i] \right) + \left(\sum_{i=1}^d \lim_{\mathbf{y} \searrow y_i} [\ln(\mathbf{y})\mathbf{y}] \right).
 \end{aligned} \quad (5.89)$$

Furthermore, observe that Lemma 5.4.11 shows that for all $b \in [0, \infty)$ it holds that

$$\lim_{\mathbf{b} \searrow b} [\ln(\mathbf{b})\mathbf{b}] = \begin{cases} 0 & : b = 0 \\ \ln(b)b & : b > 0 \end{cases} = \lim_{\mathbf{b} \searrow b} [\ln(\mathbf{b})b]. \quad (5.90)$$

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

$$\mathbf{L}_{\text{KLD}}(x, y) = - \left(\sum_{i=1}^d \lim_{\mathbf{x} \searrow x_i} [\ln(\mathbf{x})y_i] \right) + \left(\sum_{i=1}^d \lim_{\mathbf{y} \searrow y_i} [\ln(\mathbf{y})\mathbf{y}] \right) = \mathbf{L}_{\text{CE}}(x, y) - \mathbf{L}_{\text{CE}}(y, y). \quad (5.91)$$

Therefore, we obtain (5.87). The proof of Lemma 5.4.13 is thus complete. \square

Lemma 5.4.14. Let $d \in \mathbb{N}$, let \mathbf{L} be the d -dimensional Kullback–Leibler divergence loss function, let $x = (x_1, \dots, x_d)$, $y = (y_1, \dots, y_d) \in [0, \infty)^d$ satisfy $\sum_{i=1}^d x_i = \sum_{i=1}^d y_i$ and $x \neq y$, and let $f: [0, 1] \rightarrow (-\infty, \infty]$ satisfy for all $h \in [0, 1]$ that

$$f(h) = \mathbf{L}(x + h(y - x), y) \quad (5.92)$$

(cf. Definition 5.4.12). Then f is strictly decreasing.

Proof of Lemma 5.4.14. Note that Lemma 5.4.9 and Lemma 5.4.13 establish that f is strictly decreasing. The proof of Lemma 5.4.14 is thus complete. \square

Corollary 5.4.15. Let $d \in \mathbb{N}$, let $A = \{x = (x_1, \dots, x_d) \in [0, 1]^d : \sum_{i=1}^d x_i = 1\}$, 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

$$\{x \in A : \mathbf{L}(x, y) = \inf_{z \in A} \mathbf{L}(z, y)\} = \{y\} \quad (5.93)$$

and

(ii) it holds that $\inf_{z \in A} \mathbf{L}(z, y) = \mathbf{L}(y, y) = 0$.

Proof of Corollary 5.4.15. Observe that Lemma 5.4.13 and Lemma 5.4.13 prove items (i) and (ii). The proof of Corollary 5.4.15 is thus complete. \square

5.5 GF optimization in the training of ANNs

Example 5.5.1. Let $d, L, \mathfrak{d} \in \mathbb{N}$, $l_1, l_2, \dots, l_L \in \mathbb{N}$ satisfy

$$\mathfrak{d} = l_1(d+1) + \left[\sum_{k=2}^L l_k(l_{k-1}+1) \right], \quad (5.94)$$

let $a: \mathbb{R} \rightarrow \mathbb{R}$ be continuously differentiable, let $M \in \mathbb{N}$, $x_1, x_2, \dots, x_M \in \mathbb{R}^d$, $y_1, y_2, \dots, y_M \in \mathbb{R}^{l_L}$, let $\mathbf{L}: \mathbb{R}^{l_L} \times \mathbb{R}^{l_L} \rightarrow \mathbb{R}$ be the mean squared error loss function based on $\mathbb{R}^d \ni x \mapsto \|x\|_2 \in [0, \infty)$, let $\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \rightarrow [0, \infty)$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that

$$\mathcal{L}(\theta) = \frac{1}{M} \left[\sum_{m=1}^M \mathbf{L}\left((\mathcal{N}_{\mathfrak{M}_{a,l_1}, \mathfrak{M}_{a,l_2}, \dots, \mathfrak{M}_{a,l_{L-1}}, \text{id}_{\mathbb{R}^{l_L}}}^{\theta, d})(x_m), y_m \right) \right], \quad (5.95)$$

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

$$\Theta_t = \xi - \int_0^t (\nabla \mathcal{L})(\Theta_s) \, ds \quad (5.96)$$

(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 \mathcal{L} with initial value ξ (cf. Definition 5.2.1).

Proof for Example 5.5.1. Note that (5.9), (5.10), and (5.96) demonstrate that Θ is a GF trajectory for the objective function \mathcal{L} with initial value ξ (cf. Definition 5.2.1). The proof for Example 5.5.1 is thus complete. \square

Example 5.5.2. Let $d, L, \mathfrak{d} \in \mathbb{N}$, $l_1, l_2, \dots, l_L \in \mathbb{N}$ satisfy

$$\mathfrak{d} = l_1(d+1) + \left[\sum_{k=2}^L l_k(l_{k-1} + 1) \right], \quad (5.97)$$

let $a: \mathbb{R} \rightarrow \mathbb{R}$ be continuously differentiable, let $A: \mathbb{R}^{l_L} \rightarrow \mathbb{R}^{l_L}$ be the l_L -dimensional softmax activation function, let $M \in \mathbb{N}$, $x_1, x_2, \dots, x_M \in \mathbb{R}^d$, $y_1, y_2, \dots, y_M \in [0, \infty)^{l_L}$, 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 $\mathcal{L}_i: \mathbb{R}^{\mathfrak{d}} \rightarrow [0, \infty)$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that

$$\mathcal{L}_i(\theta) = \frac{1}{M} \left[\sum_{m=1}^M \mathbf{L}_i((\mathcal{N}_{\mathfrak{M}_{a, l_1}, \mathfrak{M}_{a, l_2}, \dots, \mathfrak{M}_{a, l_{L-1}}, A}^{\theta, d})(x_m), y_m) \right], \quad (5.98)$$

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

$$\Theta_t^i = \xi - \int_0^t (\nabla \mathcal{L}_i)(\Theta_s^i) \, ds \quad (5.99)$$

(cf. Definitions 1.1.3, 1.2.1, 1.2.47, 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 \mathcal{L}_j with initial value ξ (cf. Definition 5.2.1).

Proof for Example 5.5.2. Observe that Lemma 5.4.13 implies that for all $x, y \in (0, \infty)^{l_L}$ it holds that

$$(\nabla_x \mathbf{L}_1)(x, y) = (\nabla_x \mathbf{L}_2)(x, y). \quad (5.100)$$

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

$$(\nabla \mathcal{L}_1)(x) = (\nabla \mathcal{L}_2)(x). \quad (5.101)$$

This, (5.9), (5.10), and (5.99) ensure that for all $i \in \{1, 2\}$ it holds that Θ^i is a GF trajectory for the objective function \mathcal{L}_j with initial value ξ (cf. Definition 5.2.1). The proof for Example 5.5.2 is thus complete. \square

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 $\mathcal{L}: O \rightarrow \mathbb{R}$ be a function. Then we say that ϑ is a local minimum point of \mathcal{L} (we say that ϑ is a local minimizer of \mathcal{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

$$\mathcal{L}(\vartheta) \leq \mathcal{L}(\theta) \quad (5.102)$$

(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 $\mathcal{L}: O \rightarrow \mathbb{R}$ be a function. Then we say that ϑ is a global minimum point of \mathcal{L} (we say that ϑ is a global minimizer of \mathcal{L}) if and only if it holds for all $\theta \in O$ that

$$\mathcal{L}(\vartheta) \leq \mathcal{L}(\theta). \quad (5.103)$$

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 $\mathcal{L}: O \rightarrow \mathbb{R}$ be a function. Then we say that ϑ is a local maximum point of \mathcal{L} (we say that ϑ is a local maximizer of \mathcal{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

$$\mathcal{L}(\vartheta) \geq \mathcal{L}(\theta) \quad (5.104)$$

(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 $\mathcal{L}: O \rightarrow \mathbb{R}$ be a function. Then we say that ϑ is a global maximum point of \mathcal{L} (we say that ϑ is a global maximizer of \mathcal{L}) if and only if it holds for all $\theta \in O$ that

$$\mathcal{L}(\vartheta) \geq \mathcal{L}(\theta). \quad (5.105)$$

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 $\mathcal{L}: O \rightarrow \mathbb{R}$ be differentiable at ϑ . Then we say that ϑ is a critical point of \mathcal{L} if and only if it holds that

$$(\nabla \mathcal{L})(\vartheta) = 0. \quad (5.106)$$

Lemma 5.6.6. Let $\mathfrak{d} \in \mathbb{N}$, let $O \subseteq \mathbb{R}^{\mathfrak{d}}$ be open, let $\vartheta \in O$, let $\mathcal{L}: O \rightarrow \mathbb{R}$ be a function, assume that \mathcal{L} is differentiable at ϑ , and assume that $(\nabla \mathcal{L})(\vartheta) \neq 0$. Then there exists $\theta \in O$ such that $\mathcal{L}(\theta) < \mathcal{L}(\vartheta)$.

Proof of Lemma 5.6.6. Throughout this proof, let $v \in \mathbb{R}^{\mathfrak{d}} \setminus \{0\}$ satisfy $v = -(\nabla \mathcal{L})(\vartheta)$, let $\delta \in (0, \infty)$ satisfy for all $t \in (-\delta, \delta)$ that

$$\vartheta + tv = \vartheta - t(\nabla \mathcal{L})(\vartheta) \in O, \quad (5.107)$$

and let $L: (-\delta, \delta) \rightarrow \mathbb{R}$ satisfy for all $t \in (-\delta, \delta)$ that

$$L(t) = \mathcal{L}(\vartheta + tv). \quad (5.108)$$

Note that for all $t \in (0, \delta)$ it holds that

$$\begin{aligned} \left| \left[\frac{L(t) - L(0)}{t} \right] + \|v\|_2^2 \right| &= \left| \left[\frac{\mathcal{L}(\vartheta + tv) - \mathcal{L}(\vartheta)}{t} \right] + \|(\nabla \mathcal{L})(\vartheta)\|_2^2 \right| \\ &= \left| \left[\frac{\mathcal{L}(\vartheta + tv) - \mathcal{L}(\vartheta)}{t} \right] + \langle (\nabla \mathcal{L})(\vartheta), (\nabla \mathcal{L})(\vartheta) \rangle \right| \\ &= \left| \left[\frac{\mathcal{L}(\vartheta + tv) - \mathcal{L}(\vartheta)}{t} \right] - \langle (\nabla \mathcal{L})(\vartheta), v \rangle \right|. \end{aligned} \quad (5.109)$$

Therefore, we obtain that for all $t \in (0, \delta)$ it holds that

$$\begin{aligned} \left| \left[\frac{L(t) - L(0)}{t} \right] + \|v\|_2^2 \right| &= \left| \left[\frac{\mathcal{L}(\vartheta + tv) - \mathcal{L}(\vartheta)}{t} \right] - \mathcal{L}'(\vartheta)v \right| \\ &= \left| \frac{\mathcal{L}(\vartheta + tv) - \mathcal{L}(\vartheta) - \mathcal{L}'(\vartheta)tv}{t} \right| = \frac{|\mathcal{L}(\vartheta + tv) - \mathcal{L}(\vartheta) - \mathcal{L}'(\vartheta)tv|}{t}. \end{aligned} \quad (5.110)$$

The assumption that \mathcal{L} is differentiable at ϑ hence demonstrates that

$$\limsup_{t \searrow 0} \left| \left[\frac{L(t) - L(0)}{t} \right] + \|v\|_2^2 \right| = 0. \quad (5.111)$$

The fact that $\|v\|_2^2 > 0$ therefore demonstrates that there exists $t \in (0, \delta)$ which satisfies

$$\left| \left[\frac{L(t) - L(0)}{t} \right] + \|v\|_2^2 \right| < \frac{\|v\|_2^2}{2}. \quad (5.112)$$

Note that the triangle inequality, the fact that $\|v\|_2^2 > 0$, and (5.112) prove that

$$\begin{aligned} \frac{L(t) - L(0)}{t} &= \left[\frac{L(t) - L(0)}{t} + \|v\|_2^2 \right] - \|v\|_2^2 \leq \left| \left[\frac{L(t) - L(0)}{t} \right] + \|v\|_2^2 \right| - \|v\|_2^2 \\ &< \frac{\|v\|_2^2}{2} - \|v\|_2^2 = -\frac{\|v\|_2^2}{2} < 0. \end{aligned} \quad (5.113)$$

This ensures that

$$\mathcal{L}(\vartheta + tv) = L(t) < L(0) = \mathcal{L}(\vartheta). \quad (5.114)$$

The proof of Lemma 5.6.6 is thus complete. \square

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 $\mathcal{L}: O \rightarrow \mathbb{R}$ be a function, assume that \mathcal{L} is differentiable at ϑ , and assume*

$$\mathcal{L}(\vartheta) = \inf_{\theta \in O} \mathcal{L}(\theta). \quad (5.115)$$

Then $(\nabla \mathcal{L})(\vartheta) = 0$.

Proof of Lemma 5.6.7. We prove Lemma 5.6.7 by contradiction. We thus assume that $(\nabla \mathcal{L})(\vartheta) \neq 0$. Lemma 5.6.6 then implies that there exists $\theta \in O$ such that $\mathcal{L}(\theta) < \mathcal{L}(\vartheta)$. Combining this with (5.115) shows that

$$\mathcal{L}(\theta) < \mathcal{L}(\vartheta) = \inf_{w \in O} \mathcal{L}(w) \leq \mathcal{L}(\theta). \quad (5.116)$$

The proof of Lemma 5.6.7 is thus complete. \square

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 $\mathcal{L}: O \rightarrow \mathbb{R}$ be differentiable at ϑ , and assume that ϑ is a local minimum point of \mathcal{L} . Then ϑ is a critical point of \mathcal{L} (cf. Definition 5.6.5).*

Proof of Corollary 5.6.8. Observe that Lemma 5.6.7 shows that $(\nabla \mathcal{L})(\vartheta) = 0$. The proof of Corollary 5.6.8 is thus complete. \square

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 example, to [149].

5.7.1 Convexity

Definition 5.7.1 (Convex functions). Let $\mathfrak{d} \in \mathbb{N}$ and let $\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$ be a function. Then we say that \mathcal{L} is a convex function (we say that \mathcal{L} is convex) if and only if it holds for all $v, w \in \mathbb{R}^{\mathfrak{d}}$, $t \in (0, 1)$ that

$$\mathcal{L}(tv + (1 - t)w) \leq t\mathcal{L}(v) + (1 - t)\mathcal{L}(w). \quad (5.117)$$

Lemma 5.7.2 (Equivalence for convex functions). Let $\mathfrak{d} \in \mathbb{N}$ and let $\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$ be a function. Then the following three statements are equivalent:

(i) It holds that \mathcal{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

$$\mathcal{L}(\theta + tv) \leq \mathcal{L}(\theta) + t(\mathcal{L}(\theta + v) - \mathcal{L}(\theta)). \quad (5.118)$$

(iii) It holds for all $\theta, v \in \mathbb{R}^{\mathfrak{d}}$, $t \in (0, 1)$ that

$$t(\mathcal{L}(\theta + v) - \mathcal{L}(\theta + tv)) - (1 - t)(\mathcal{L}(\theta + tv) - \mathcal{L}(\theta)) \geq 0. \quad (5.119)$$

Proof of Lemma 5.7.2. Note that (5.117) establishes that ((i) \leftrightarrow (ii)) and ((i) \leftrightarrow (iii)). The proof of Lemma 5.7.2 is thus complete. \square

Lemma 5.7.3 (Equivalence for differentiable convex functions). Let $\mathfrak{d} \in \mathbb{N}$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$. Then the following three statements are equivalent:

(i) It holds that \mathcal{L} is convex (cf. Definition 5.7.1).

(ii) It holds for all $v, w \in \mathbb{R}^{\mathfrak{d}}$ that

$$\mathcal{L}(v) \geq \mathcal{L}(w) + \langle (\nabla \mathcal{L})(w), v - w \rangle \quad (5.120)$$

(cf. Definition 1.4.7).

(iii) It holds for all $v, w \in \mathbb{R}^{\mathfrak{d}}$ that

$$\langle (\nabla \mathcal{L})(v) - (\nabla \mathcal{L})(w), v - w \rangle \geq 0 \quad (5.121)$$

(cf. Definition 1.4.7).

Proof of Lemma 5.7.3. We first prove that ((i) \rightarrow (ii)). For this assume that \mathcal{L} is convex (cf. Definition 5.7.1). Observe that the assumption that \mathcal{L} is convex and item (ii) in Lemma 5.7.2 demonstrate that for all $v, w \in \mathbb{R}^d$, $t \in (0, 1)$ it holds that

$$\mathcal{L}(w + t(v - w)) \leq \mathcal{L}(w) + t(\mathcal{L}(v) - \mathcal{L}(w)). \quad (5.122)$$

Hence, we obtain that for all $v, w \in \mathbb{R}^d$, $t \in (0, 1)$ it holds that

$$\mathcal{L}(v) \geq \mathcal{L}(w) + \frac{\mathcal{L}(w + t(v - w)) - \mathcal{L}(w)}{t}. \quad (5.123)$$

Combining this and the assumption that \mathcal{L} is differentiable establishes that for all $v, w \in \mathbb{R}^d$ it holds that

$$\mathcal{L}(v) \geq \mathcal{L}(w) + \limsup_{t \rightarrow 0} \frac{\mathcal{L}(w + t(v - w)) - \mathcal{L}(w)}{t} = \mathcal{L}(w) + \langle (\nabla \mathcal{L})(w), v - w \rangle \quad (5.124)$$

(cf. Definition 1.4.7). This proves that ((i) \rightarrow (ii)).

In the next step we prove that ((ii) \rightarrow (iii)). For this assume that for all $v, w \in \mathbb{R}^d$ it holds that

$$\mathcal{L}(v) \geq \mathcal{L}(w) + \langle (\nabla \mathcal{L})(w), v - w \rangle. \quad (5.125)$$

Note that (5.125) proves that for all $v, w \in \mathbb{R}^d$ it holds that

$$\begin{aligned} \mathcal{L}(v) + \mathcal{L}(w) &\geq \mathcal{L}(w) + \langle (\nabla \mathcal{L})(w), v - w \rangle + \mathcal{L}(v) + \langle (\nabla \mathcal{L})(v), w - v \rangle \\ &= \mathcal{L}(v) + \mathcal{L}(w) - \langle (\nabla \mathcal{L})(v) - (\nabla \mathcal{L})(w), v - w \rangle \end{aligned} \quad (5.126)$$

This implies that for all $v, w \in \mathbb{R}^d$ it holds that

$$\langle (\nabla \mathcal{L})(v) - (\nabla \mathcal{L})(w), v - w \rangle \geq 0. \quad (5.127)$$

This proves that ((ii) \rightarrow (iii)).

In the next step we prove that ((iii) \rightarrow (i)). For this assume that for all $v, w \in \mathbb{R}^d$ it holds that

$$\langle (\nabla \mathcal{L})(v) - (\nabla \mathcal{L})(w), v - w \rangle \geq 0. \quad (5.128)$$

Observe that (5.128) ensures that for all $\theta, v \in \mathbb{R}^d$, $\alpha, \beta \in \mathbb{R}$ with $\alpha > \beta$ it holds that

$$\begin{aligned} &\langle (\nabla \mathcal{L})(\theta + \alpha v) - (\nabla \mathcal{L})(\theta + \beta v), v \rangle \\ &= (\alpha - \beta)^{-1} \langle (\nabla \mathcal{L})(\theta + \alpha v) - (\nabla \mathcal{L})(\theta + \beta v), (\alpha - \beta)v \rangle \geq 0. \end{aligned} \quad (5.129)$$

Combining this and the fundamental theorem of calculus shows that for all $\theta, v \in \mathbb{R}^{\vartheta}$, $t \in (0, 1)$ it holds that

$$\begin{aligned}
 & t(\mathcal{L}(\theta + v) - \mathcal{L}(\theta + tv)) - (1-t)(\mathcal{L}(\theta + tv) - \mathcal{L}(\theta)) \\
 &= t \left(\int_t^1 \langle (\nabla \mathcal{L})(\theta + sv), v \rangle ds \right) - (1-t) \left(\int_0^t \langle (\nabla \mathcal{L})(\theta + sv), v \rangle ds \right) \\
 &= t(1-t) \left(\int_0^1 \langle (\nabla \mathcal{L})(\theta + (t+s(1-t))v), v \rangle ds \right) \\
 &\quad - (1-t)t \left(\int_0^1 \langle (\nabla \mathcal{L})(\theta + stv), v \rangle ds \right) \\
 &= t(1-t) \left(\int_0^1 \langle (\nabla \mathcal{L})(\theta + (t+s(1-t))v) - (\nabla \mathcal{L})(\theta + stv), v \rangle ds \right) \\
 &\geq 0.
 \end{aligned} \tag{5.130}$$

This and item (iii) in Lemma 5.7.2 demonstrate that \mathcal{L} is convex. This proves that ((iii) \rightarrow (i)). The proof of Lemma 5.7.3 is thus complete. \square

5.7.2 Strict convexity

Definition 5.7.4 (Strictly convex functions). Let $\vartheta \in \mathbb{N}$ and let $\mathcal{L}: \mathbb{R}^{\vartheta} \rightarrow \mathbb{R}$ be a function. Then we say that \mathcal{L} is a strictly convex function (we say that \mathcal{L} is strictly convex) if and only if it holds for all $v, w \in \mathbb{R}^{\vartheta}$, $t \in (0, 1)$ with $v \neq w$ that

$$\mathcal{L}(tv + (1-t)w) < t\mathcal{L}(v) + (1-t)\mathcal{L}(w). \tag{5.131}$$

Lemma 5.7.5 (Strictly convex functions are convex). Let $\vartheta \in \mathbb{N}$, $\vartheta \in \mathbb{R}^{\vartheta}$ and let $\mathcal{L}: \mathbb{R}^{\vartheta} \rightarrow \mathbb{R}$ be strictly convex (cf. Definition 5.7.4). Then it holds that \mathcal{L} is convex (cf. Definition 5.7.1).

Proof of Lemma 5.7.5. Note that (5.117) and (5.131) establish that \mathcal{L} is convex (cf. Definition 5.7.1). The proof of Lemma 5.7.5 is thus complete. \square

Lemma 5.7.6 (Global minima of strictly convex functions are unique). Let $\vartheta \in \mathbb{N}$, $\vartheta \in \mathbb{R}^{\vartheta}$, let $\mathcal{L}: \mathbb{R}^{\vartheta} \rightarrow \mathbb{R}$ be strictly convex, and assume that ϑ is a global minimum point of \mathcal{L} (cf. Definitions 5.6.2 and 5.7.4). Then it holds for all global minimum points

$v \in \mathbb{R}^{\mathfrak{d}}$ of \mathcal{L} that

$$v = \vartheta. \quad (5.132)$$

Proof of Lemma 5.7.6. Observe that (5.131), the assumption that ϑ is a global minimum point of \mathcal{L} , and the assumption that v is a global minimum point of \mathcal{L} prove that for all $v \in \mathbb{R}^{\mathfrak{d}} \setminus \{\vartheta\}$ with $\mathcal{L}(v) = \inf_{\theta \in \mathbb{R}^{\mathfrak{d}}} \mathcal{L}(\theta)$ it holds that

$$\mathcal{L}(\vartheta) \leq \mathcal{L}\left(\frac{v+\vartheta}{2}\right) < \frac{1}{2}[\mathcal{L}(v) + \mathcal{L}(\vartheta)] = \mathcal{L}(v) \quad (5.133)$$

Hence, we obtain that

$$\left\{ v \in \mathbb{R}^{\mathfrak{d}} \setminus \{\vartheta\} : \mathcal{L}(v) = \inf_{\theta \in \mathbb{R}^{\mathfrak{d}}} \mathcal{L}(\theta) \right\} = \{\vartheta\}. \quad (5.134)$$

The proof of Lemma 5.7.6 is thus complete. \square

Example 5.7.7. Let $f: \mathbb{R} \rightarrow \mathbb{R}$ and $g: \mathbb{R} \rightarrow \mathbb{R}$ satisfy for all $x \in \mathbb{R}$ that

$$f(x) = \exp(x) \quad \text{and} \quad g(x) = 0. \quad (5.135)$$

Then

- (i) it holds that f is strictly convex,
 - (ii) it holds that g is convex, and
 - (iii) it holds that g is not strictly convex
- (cf. Definitions 5.7.1 and 5.7.4).

5.7.3 Monotonicity

Definition 5.7.8 (Monotonically increasing functions). Let $\mathfrak{d} \in \mathbb{N}$ and let $\mathcal{G}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a function. Then we say that \mathcal{G} is a monotonically increasing function (we say that \mathcal{G} is monotonically increasing) if and only if it holds for all $v, w \in \mathbb{R}^{\mathfrak{d}}$ that

$$\langle \mathcal{G}(v) - \mathcal{G}(w), v - w \rangle \geq 0 \quad (5.136)$$

(cf. Definition 1.4.7).

Definition 5.7.9 (Monotonically decreasing functions). Let $\mathfrak{d} \in \mathbb{N}$ and let $\mathcal{G}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a function. Then we say that \mathcal{G} is a monotonically decreasing function (we say that \mathcal{G} is monotonically decreasing) if and only if it holds for all $v, w \in \mathbb{R}^{\mathfrak{d}}$ that

$$\langle \mathcal{G}(v) - \mathcal{G}(w), v - w \rangle \leq 0 \quad (5.137)$$

(cf. Definition 1.4.7).

Lemma 5.7.10 (Equivalence for monotonically increasing and decreasing functions). Let $\mathfrak{d} \in \mathbb{N}$ and let $\mathcal{G}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a function. Then the following two statements are equivalent:

- (i) It holds that \mathcal{G} is monotonically increasing (cf. Definition 5.7.8).
- (ii) It holds that $-\mathcal{G}$ is monotonically decreasing (cf. Definition 5.7.9).

Proof of Lemma 5.7.10. Note that (5.136) and (5.137) imply that ((i) \leftrightarrow (ii)). The proof of Lemma 5.7.10 is thus complete. \square

Lemma 5.7.11 (Convexity and monotonicity). Let $\mathfrak{d} \in \mathbb{N}$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$. Then the following three statements are equivalent:

- (i) It holds that \mathcal{L} is convex (cf. Definition 5.7.1).
- (ii) It holds that $\nabla \mathcal{L}$ is monotonically increasing (cf. Definition 5.7.8).
- (iii) It holds that $-(\nabla \mathcal{L})$ is monotonically decreasing (cf. Definition 5.7.9).

Proof of Lemma 5.7.11. Observe that Lemma 5.7.3 and Lemma 5.7.10 establish that ((i) \leftrightarrow (ii)) and that ((i) \leftrightarrow (iii)). The proof of Lemma 5.7.11 is thus complete. \square

Definition 5.7.12 (Generalized monotonically increasing functions). Let $\mathfrak{d} \in \mathbb{N}$, $c \in \mathbb{R}$ and let $\mathcal{G}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a function. Then we say that \mathcal{G} is a c -generalized monotonically increasing function (we say that \mathcal{G} is c -generalized monotonically increasing) if and only if it holds for all $v, w \in \mathbb{R}^{\mathfrak{d}}$ that

$$\langle \mathcal{G}(v) - \mathcal{G}(w), v - w \rangle \geq c \|v - w\|_2^2 \quad (5.138)$$

(cf. Definitions 1.4.7 and 3.3.4).

Definition 5.7.13 (Generalized monotonically decreasing functions). *Let $\mathfrak{d} \in \mathbb{N}$, $c \in \mathbb{R}$ and let $\mathcal{G}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a function. Then we say that \mathcal{G} is a c -generalized monotonically decreasing function (we say that \mathcal{G} is c -generalized monotonically decreasing) if and only if it holds for all $v, w \in \mathbb{R}^{\mathfrak{d}}$ that*

$$\langle \mathcal{G}(v) - \mathcal{G}(w), v - w \rangle \leq -c\|v - w\|_2^2. \quad (5.139)$$

(cf. Definitions 1.4.7 and 3.3.4).

Lemma 5.7.14 (Equivalence for monotonically increasing and decreasing functions). *Let $\mathfrak{d} \in \mathbb{N}$, $c \in \mathbb{R}$ and let $\mathcal{G}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a function. Then the following two statements are equivalent:*

- (i) *It holds that \mathcal{G} is c -generalized monotonically increasing (cf. Definition 5.7.12).*
- (ii) *It holds that $-\mathcal{G}$ is c -generalized monotonically decreasing (cf. Definition 5.7.13).*

Proof of Lemma 5.7.14. Note that (5.138) and (5.139) ensure that (i) \leftrightarrow (ii). The proof of Lemma 5.7.14 is thus complete. \square

Lemma 5.7.15 (Equivalence for differentiable monotonically increasing functions). *Let $\mathfrak{d} \in \mathbb{N}$, $c \in \mathbb{R}$ and let $\mathcal{G} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R}^{\mathfrak{d}})$. Then the following two statements are equivalent:*

- (i) *It holds that \mathcal{G} is c -generalized monotonically increasing (cf. Definition 5.7.12).*
- (ii) *It holds for all $v, w \in \mathbb{R}^{\mathfrak{d}}$ that*

$$\langle \mathcal{G}'(v)w, w \rangle \geq c\|w\|_2^2 \quad (5.140)$$

(cf. Definitions 1.4.7 and 3.3.4).

Proof of Lemma 5.7.15. We first prove that (i) \rightarrow (ii). For this assume that \mathcal{G} is c -generalized monotonically increasing (cf. Definition 5.7.12). Observe that (5.138) and the fact that \mathcal{G} is differentiable show that for all $v, w \in \mathbb{R}^{\mathfrak{d}}$ it holds that

$$\begin{aligned} \langle \mathcal{G}'(v)w, w \rangle &= \limsup_{t \rightarrow 0} \left(\left\langle \frac{\mathcal{G}(v + tw) - \mathcal{G}(v)}{t}, w \right\rangle \right) \\ &= \limsup_{t \rightarrow 0} \left(\frac{1}{t^2} \langle \mathcal{G}(v + tw) - \mathcal{G}(v), tw \rangle \right) \\ &\geq \limsup_{t \rightarrow 0} \left(\frac{c}{t^2} \|tw\|_2^2 \right) = c\|w\|_2^2 \end{aligned} \quad (5.141)$$

(cf. Definitions 1.4.7 and 3.3.4). This proves that ((i) \rightarrow (ii)).

In the next step, we prove that ((ii) \rightarrow (i)). For this assume that for all $v, w \in \mathbb{R}^{\mathfrak{d}}$ it holds that

$$\langle \mathcal{G}'(v)w, w \rangle \geq c\|w\|_2^2. \quad (5.142)$$

Note that (5.142) and the fundamental theorem of calculus demonstrate that for all $v, w \in \mathbb{R}^{\mathfrak{d}}$ it holds that

$$\begin{aligned} \langle \mathcal{G}(v) - \mathcal{G}(w), v - w \rangle &= \left\langle \int_0^1 \mathcal{G}'(w + t(v - w))(v - w) dt, v - w \right\rangle \\ &= \int_0^1 \langle \mathcal{G}'(w + t(v - w))(v - w), v - w \rangle dt \\ &\geq \int_0^1 c\|v - w\|_2^2 dt = c\|v - w\|_2^2. \end{aligned} \quad (5.143)$$

This proves that ((ii) \rightarrow (i)). The proof of Lemma 5.7.15 is thus complete. \square

5.7.4 Subgradients

Definition 5.7.16 (Subgradients). Let $\mathfrak{d} \in \mathbb{N}$, $g, \theta \in \mathbb{R}^{\mathfrak{d}}$ and let $\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$ be a function. Then we say that g is a subgradient of \mathcal{L} at θ if and only if it holds for all $v \in \mathbb{R}^{\mathfrak{d}}$ that

$$\mathcal{L}(v) \geq \mathcal{L}(\theta) + \langle g, v - \theta \rangle \quad (5.144)$$

(cf. Definition 1.4.7).

Lemma 5.7.17 (Convexity and subgradients). Let $\mathfrak{d} \in \mathbb{N}$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$. Then the following two statements are equivalent:

- (i) It holds that \mathcal{L} is convex (cf. Definition 5.7.1).
- (ii) It holds for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that $(\nabla \mathcal{L})(\theta)$ is a subgradient of \mathcal{L} at θ (cf. Definition 5.7.16).

Proof of Lemma 5.7.17. Observe that Lemma 5.7.3 proves that ((i) \leftrightarrow (ii)). The proof of Lemma 5.7.17 is thus complete. \square

5.7.5 Strong convexity

Definition 5.7.18 (Generalized convex functions). Let $\mathfrak{d} \in \mathbb{N}$, $c \in \mathbb{R}$ and let $\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$ be a function. Then we say that \mathcal{L} is a c -generalized convex function (we say that \mathcal{L} is c -generalized convex) if and only if it holds that

$$\mathbb{R}^{\mathfrak{d}} \ni \theta \mapsto \mathcal{L}(\theta) - \frac{c}{2} \|\theta\|_2^2 \in \mathbb{R} \quad (5.145)$$

is convex (cf. Definitions 3.3.4 and 5.7.1).

Definition 5.7.19 (Strongly convex functions). Let $\mathfrak{d} \in \mathbb{N}$ and let $\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$ be a function. Then we say that \mathcal{L} is a strongly convex function (we say that \mathcal{L} is strongly convex) if and only if there exists $c \in (0, \infty)$ such that \mathcal{L} is c -generalized convex (cf. Definitions 5.7.18 and 5.7.19).

Lemma 5.7.20 (Equivalence for generalized convex functions). Let $\mathfrak{d} \in \mathbb{N}$, $c \in \mathbb{R}$ and let $\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$ be a function. Then the following four statements are equivalent:

(i) It holds that \mathcal{L} is c -generalized convex (cf. Definition 5.7.18).

(ii) It holds for all $v, w \in \mathbb{R}^{\mathfrak{d}}$, $t \in (0, 1)$ that

$$\mathcal{L}(tv + (1-t)w) \leq t\mathcal{L}(v) + (1-t)\mathcal{L}(w) - \frac{c}{2}[t(1-t)\|v-w\|_2^2] \quad (5.146)$$

(cf. Definition 3.3.4).

(iii) It holds for all $\theta, v \in \mathbb{R}^{\mathfrak{d}}$, $t \in (0, 1)$ that

$$\mathcal{L}(\theta + tv) \leq \mathcal{L}(\theta) + t(\mathcal{L}(\theta + v) - \mathcal{L}(\theta)) - \frac{c}{2}[t(1-t)\|v\|_2^2] \quad (5.147)$$

(cf. Definition 3.3.4).

(iv) It holds for all $\theta, v \in \mathbb{R}^{\mathfrak{d}}$, $t \in (0, 1)$ that

$$t(\mathcal{L}(\theta+v) - \mathcal{L}(\theta+tv)) - (1-t)(\mathcal{L}(\theta+tv) - \mathcal{L}(\theta)) \geq \frac{c}{2}[t(1-t)\|v-w\|_2^2] \quad (5.148)$$

(cf. Definition 3.3.4).

Proof of Lemma 5.7.20. Note that (5.117) and (5.145) imply that \mathcal{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

$$\mathcal{L}(tv + (1-t)w) - \frac{c}{2}\|tv + (1-t)w\|_2^2 \leq t(\mathcal{L}(v) - \frac{c}{2}\|v\|_2^2) + (1-t)(\mathcal{L}(w) - \frac{c}{2}\|w\|_2^2) \quad (5.149)$$

(cf. Definitions 3.3.4 and 5.7.18). This establishes that \mathcal{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

$$\begin{aligned}\mathcal{L}(tv + (1-t)w) &\leq t\mathcal{L}(v) + (1-t)\mathcal{L}(w) \\ &\quad - \frac{c}{2}(t\|v\|_2^2 + (1-t)\|w\|_2^2 - \|tv + (1-t)w\|_2^2).\end{aligned}\tag{5.150}$$

Moreover, observe that the fact that for all $t \in (0, 1)$ it holds that

$$(1-t) - (1-t)^2 = 1 - t - t^2 + 2t - 1 = t(1-t)\tag{5.151}$$

ensures that for all $v, w \in \mathbb{R}^{\mathfrak{d}}$, $t \in (0, 1)$ it holds that

$$\begin{aligned}&t\|v\|_2^2 + (1-t)\|w\|_2^2 - \|tv + (1-t)w\|_2^2 \\ &= t\|v\|_2^2 + (1-t)\|w\|_2^2 - (t^2\|v\|_2^2 + (1-t)^2\|w\|_2^2 + 2t(1-t)\langle v, w \rangle) \\ &= (t - t^2)\|v\|_2^2 + (1 - t - (1-t)^2)\|w\|_2^2 - 2t(1-t)\langle v, w \rangle \\ &= t(1-t)(\|v\|_2^2 + \|w\|_2^2 - 2\langle v, w \rangle) \\ &= t(1-t)\|v - w\|_2^2.\end{aligned}\tag{5.152}$$

(cf. Definition 1.4.7). Combining this and (5.150) shows that \mathcal{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

$$\mathcal{L}(tv + (1-t)w) \leq t\mathcal{L}(v) + (1-t)\mathcal{L}(w) - \frac{c}{2}[t(1-t)\|v - w\|_2^2].\tag{5.153}$$

Hence, we obtain that ((i) \leftrightarrow (ii)). Furthermore, note that (5.146) proves that ((ii) \leftrightarrow (iii)) and that ((iii) \leftrightarrow (iv)). The proof of Lemma 5.7.20 is thus complete. \square

Lemma 5.7.21 (Strongly convex functions are strictly convex). *Let $\mathfrak{d} \in \mathbb{N}$ and let $\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$ be strongly convex (cf. Definition 5.7.19). Then it holds that \mathcal{L} is strictly convex (cf. Definition 5.7.4).*

Proof of Lemma 5.7.21. Observe that (5.131) and item (ii) in Lemma 5.7.20 demonstrate that \mathcal{L} is strictly convex (cf. Definition 5.7.4). The proof of Lemma 5.7.21 is thus complete. \square

Corollary 5.7.22 (Strongly convex functions have a unique minimum point). *Let $\mathfrak{d} \in \mathbb{N}$ and let $\mathcal{L} \in C(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ be strongly convex (cf. Definition 5.7.19). Then there exists a unique $\vartheta \in \mathbb{R}^{\mathfrak{d}}$ such that ϑ is a global minimum point of \mathcal{L} (cf. Definition 5.6.2).*

Proof of Corollary 5.7.22. Throughout this proof, for every $r \in (0, \infty)$ let

$$B_r = \{v \in \mathbb{R}^d : \|v\|_2 \leq r\}, \quad (5.154)$$

let

$$S = \{v \in \mathbb{R}^d : \|v\|_2 = 1\} \quad (5.155)$$

let $c \in (0, \infty)$ satisfy that \mathcal{L} is c -generalized convex, and let $C \in \mathbb{R}$ satisfy

$$C = \left[\inf_{\theta \in S} \mathcal{L}(\theta) \right] - \mathcal{L}(0) - \frac{c}{2} \quad (5.156)$$

(cf. Definitions 3.3.4 and 5.7.18). Note that item (iii) in Lemma 5.7.20 and the fact that \mathcal{L} is c -generalized convex imply that for all $v \in \mathbb{R}^d \setminus B_1$ it holds that

$$\mathcal{L}\left(0 + \frac{1}{\|v\|_2} v\right) \leq \mathcal{L}(0) + \frac{1}{\|v\|_2} (\mathcal{L}(v) - \mathcal{L}(0)) - \frac{c}{2} \left[\frac{1}{\|v\|_2} \left(1 - \frac{1}{\|v\|_2}\right) \|v\|_2^2 \right] \quad (5.157)$$

This establishes that for all $v \in \mathbb{R}^d \setminus B_1$ it holds that

$$\begin{aligned} \mathcal{L}(v) &\geq \mathcal{L}(0) + \|v\|_2 \left(\mathcal{L}\left(\frac{1}{\|v\|_2} v\right) - \mathcal{L}(0) \right) + \frac{c}{2} \left(1 - \frac{1}{\|v\|_2}\right) \|v\|_2^2 \\ &\geq \mathcal{L}(0) + \left(\left[\inf_{\theta \in S} \mathcal{L}(\theta) \right] - \mathcal{L}(0) - \frac{c}{2} \right) \|v\|_2 + \frac{c}{2} \|v\|_2^2 \\ &= \mathcal{L}(0) + C \|v\|_2 + \frac{c}{2} \|v\|_2^2. \end{aligned} \quad (5.158)$$

Hence, we obtain that

$$\begin{aligned} \limsup_{r \rightarrow \infty} \left(\inf_{v \in \mathbb{R}^d \setminus B_r} \mathcal{L}(v) \right) &\geq \limsup_{r \rightarrow \infty} \left(\inf_{v \in \mathbb{R}^d \setminus B_r} \mathcal{L}(0) + C \|v\|_2 + \frac{c}{2} \|v\|_2^2 \right) \\ &= \limsup_{r \rightarrow \infty} \left(\inf_{s \in (r, \infty)} \mathcal{L}(0) + Cs + \frac{cs^2}{2} \right) \\ &= \infty. \end{aligned} \quad (5.159)$$

This ensures that there exists $r \in (0, \infty)$ which satisfies that

$$\inf_{v \in \mathbb{R}^d} \mathcal{L}(v) = \inf_{v \in B_r} \mathcal{L}(v). \quad (5.160)$$

Combining the fact that B_r is compact and the assumption that \mathcal{L} is continuous therefore shows that there exists $\vartheta \in B_r$ which satisfies that

$$\mathcal{L}(\vartheta) = \inf_{v \in B_r} \mathcal{L}(v) = \inf_{v \in \mathbb{R}^d} \mathcal{L}(v). \quad (5.161)$$

This, Lemma 5.7.6, and Lemma 5.7.21 prove that ϑ is the unique global minimum point of \mathcal{L} (cf. Definition 5.6.2). The proof of Corollary 5.7.22 is thus complete. \square

Proposition 5.7.23 (Equivalence for differentiable generalized-convex functions). *Let $\mathfrak{d} \in \mathbb{N}$, $c \in \mathbb{R}$, $\mathcal{L} \in C^1(\mathbb{R}^\mathfrak{d}, \mathbb{R})$. Then the following six statements are equivalent:*

(i) *It holds that \mathcal{L} is c -generalized-convex (cf. Definition 5.7.18).*

(ii) *It holds for all $v, w \in \mathbb{R}^\mathfrak{d}$ that*

$$\mathcal{L}(v) \geq \mathcal{L}(w) + \langle (\nabla \mathcal{L})(w), v - w \rangle + \frac{c}{2} \|v - w\|_2^2 \quad (5.162)$$

(cf. Definitions 1.4.7 and 3.3.4).

(iii) *It holds for all $v, w \in \mathbb{R}^\mathfrak{d}$ that*

$$\langle (\nabla \mathcal{L})(v) - (\nabla \mathcal{L})(w), v - w \rangle \geq c \|v - w\|_2^2 \quad (5.163)$$

(cf. Definitions 1.4.7 and 3.3.4).

(iv) *It holds for all $\theta \in \mathbb{R}^\mathfrak{d}$ that $(\nabla \mathcal{L})(\theta) - c\theta$ is a subgradient of $\mathbb{R}^\mathfrak{d} \ni v \mapsto \mathcal{L}(v) - \frac{c}{2} \|v\|_2^2 \in \mathbb{R}$ at θ (cf. Definitions 3.3.4 and 5.7.16).*

(v) *It holds that $\nabla \mathcal{L}$ is c -monotonically increasing (cf. Definition 5.7.12).*

(vi) *It holds that $-\nabla \mathcal{L}$ is c -monotonically decreasing (cf. Definition 5.7.13).*

Proof of Proposition 5.7.23. We first prove that ((i) \rightarrow (ii)). For this assume that \mathcal{L} is c -generalized convex. Observe that the assumption that \mathcal{L} is c -generalized convex and Lemma 5.7.20 demonstrate that for all $v, w \in \mathbb{R}^\mathfrak{d}$, $t \in (0, 1)$ it holds that

$$\mathcal{L}(w + t(v - w)) \leq \mathcal{L}(w) + t(\mathcal{L}(v) - \mathcal{L}(w)) - \frac{c}{2}[t(1-t)\|w - v\|_2^2]. \quad (5.164)$$

(cf. Definitions 3.3.4 and 5.7.18). Hence, we obtain that for all $v, w \in \mathbb{R}^\mathfrak{d}$, $t \in (0, 1)$ it holds that

$$\mathcal{L}(v) \geq \mathcal{L}(w) + \frac{\mathcal{L}(w + t(v - w)) - \mathcal{L}(w)}{t} + \frac{c}{2}[(1-t)\|v - w\|_2^2] \quad (5.165)$$

Combining this and the assumption that \mathcal{L} is differentiable implies that for all $v, w \in \mathbb{R}^\mathfrak{d}$ it holds that

$$\begin{aligned} \mathcal{L}(v) &\geq \mathcal{L}(w) + \limsup_{t \rightarrow 0} \left(\frac{\mathcal{L}(w + t(v - w)) - \mathcal{L}(w)}{t} + \frac{c}{2}[(1-t)\|v - w\|_2^2] \right) \\ &= \mathcal{L}(w) + \langle (\nabla \mathcal{L})(w), v - w \rangle + \frac{c}{2} \|v - w\|_2^2. \end{aligned} \quad (5.166)$$

(cf. Definition 1.4.7). This proves that ((i) \rightarrow (ii)).

In the next step we prove that ((ii) \rightarrow (iii)). For this assume that for all $v, w \in \mathbb{R}^\mathfrak{d}$ it holds that

$$\mathcal{L}(v) \geq \mathcal{L}(w) + \langle (\nabla \mathcal{L})(w), v - w \rangle + \frac{c}{2} \|v - w\|_2^2. \quad (5.167)$$

Note that (5.167) establishes that for all $v, w \in \mathbb{R}^d$ it holds that

$$\begin{aligned}\mathcal{L}(v) + \mathcal{L}(w) &\geq \mathcal{L}(w) + \langle (\nabla \mathcal{L})(w), v - w \rangle + \frac{c}{2} \|v - w\|_2^2 \\ &\quad + \mathcal{L}(v) + \langle (\nabla \mathcal{L})(v), w - v \rangle + \frac{c}{2} \|w - v\|_2^2 \\ &= \mathcal{L}(v) + \mathcal{L}(w) - \langle (\nabla \mathcal{L})(v) - (\nabla \mathcal{L})(w), v - w \rangle + c \|w - v\|_2^2.\end{aligned}\tag{5.168}$$

This ensures that for all $v, w \in \mathbb{R}^d$ it holds that

$$\langle (\nabla \mathcal{L})(v) - (\nabla \mathcal{L})(w), v - w \rangle \geq c \|v - w\|_2^2.\tag{5.169}$$

This proves that ((ii) \rightarrow (iii)).

In the next step we prove that ((iii) \rightarrow (i)). For this assume that for all $v, w \in \mathbb{R}^d$ it holds that

$$\langle (\nabla \mathcal{L})(v) - (\nabla \mathcal{L})(w), v - w \rangle \geq c \|v - w\|_2^2.\tag{5.170}$$

Observe that (5.170) shows that for all $\theta, v \in \mathbb{R}^d$, $\alpha, \beta \in \mathbb{R}$ with $\alpha > \beta$ it holds that

$$\begin{aligned}&\langle (\nabla \mathcal{L})(\theta + \alpha v) - (\nabla \mathcal{L})(\theta + \beta v), v \rangle \\ &= (\alpha - \beta)^{-1} \langle (\nabla \mathcal{L})(\theta + \alpha v) - (\nabla \mathcal{L})(\theta + \beta v), (\alpha - \beta)v \rangle \\ &\geq (\alpha - \beta)^{-1} c \|(\alpha - \beta)v\|_2^2 = (\alpha - \beta)c \|v\|_2^2.\end{aligned}\tag{5.171}$$

Combining this and the fundamental theorem of calculus proves that for all $\theta, v \in \mathbb{R}^d$, $t \in (0, 1)$ it holds that

$$\begin{aligned}&t(\mathcal{L}(\theta + v) - \mathcal{L}(\theta + tv)) - (1-t)(\mathcal{L}(\theta + tv) - \mathcal{L}(\theta)) \\ &= t \left(\int_t^1 \langle (\nabla \mathcal{L})(\theta + sv), v \rangle ds \right) - (1-t) \left(\int_0^t \langle (\nabla \mathcal{L})(\theta + sv), v \rangle ds \right) \\ &= t(1-t) \left(\int_0^1 \langle (\nabla \mathcal{L})(\theta + (t+s(1-t))v), v \rangle ds \right) \\ &\quad - (1-t)t \left(\int_0^1 \langle (\nabla \mathcal{L})(\theta + stv), v \rangle ds \right) \\ &= t(1-t) \left(\int_0^1 \langle (\nabla \mathcal{L})(\theta + (t+s(1-t))v) - (\nabla \mathcal{L})(\theta + stv), v \rangle ds \right) \\ &\geq t(1-t) \left(\int_0^1 (t+s-2st)c \|v\|_2^2 ds \right) \\ &= t(1-t)(t + \frac{1}{2} - t)c \|v\|_2^2 = \frac{c}{2}[t(1-t)\|v\|_2^2]\end{aligned}\tag{5.172}$$

This and Lemma 5.7.20 demonstrate that \mathcal{L} is c -generalized convex. This proves that ((iii) \rightarrow (i)).

Note that Lemma 5.7.17 implies that ((i) \leftrightarrow (iv)). In the next step we observe that (5.138) establishes that ((iii) \leftrightarrow (v)). Next, note that Lemma 5.7.14 ensures that ((v) \leftrightarrow (vi)). The proof of Proposition 5.7.23 is thus complete. \square

Proposition 5.7.24 (Equivalence for two times differentiable generalized-convex functions). *Let $\mathfrak{d} \in \mathbb{N}$, $c \in \mathbb{R}$, $\mathcal{L} \in \mathbb{C}^2(\mathbb{R}^\mathfrak{d}, \mathbb{R})$. Then the following two statements are equivalent:*

- (i) *It holds that \mathcal{L} is c -generalized-convex (cf. Definition 5.7.18).*
- (ii) *It holds for all $v, w \in \mathbb{R}^\mathfrak{d}$ that*

$$\langle (\text{Hess } \mathcal{L})(v)w, w \rangle \geq c\|w\|_2^2 \quad (5.173)$$

(cf. Definitions 1.4.7 and 3.3.4).

Proof of Proposition 5.7.24. Observe that Lemma 5.7.15 and Proposition 5.7.23 and the fact that $\nabla \mathcal{L}$ is continuously differentiable prove that ((i) \leftrightarrow (ii)). The proof of Proposition 5.7.24 is thus complete. \square

5.7.6 Coercivity

Definition 5.7.25 (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 $\mathcal{L}: O \rightarrow \mathbb{R}$ be a function. Then we say that \mathcal{L} satisfies a coercivity-type condition with coercivity constant c at ϑ if and only if*

- (i) *it holds that \mathcal{L} is differentiable and*
- (ii) *it holds for all $\theta \in O$ that*

$$\langle \theta - \vartheta, (\nabla \mathcal{L})(\theta) \rangle \geq c\|\theta - \vartheta\|_2^2 \quad (5.174)$$

(cf. Definitions 1.4.7 and 3.3.4).

Definition 5.7.26 (Coercive-type functions). *Let $\mathfrak{d} \in \mathbb{N}$ and let $\mathcal{L}: \mathbb{R}^\mathfrak{d} \rightarrow \mathbb{R}$ be a function. Then we say that \mathcal{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 \mathcal{L} satisfies a coercivity-type condition at ϑ with coercivity constant c (cf. Definition 5.7.25).*

Corollary 5.7.27 (Strongly convex functions are coercive). *Let $\mathfrak{d} \in \mathbb{N}$, $c \in (0, \infty)$ and let $\mathcal{L} \in \mathbb{C}^1(\mathbb{R}^\mathfrak{d}, \mathbb{R})$ be c -generalized convex (cf. Definition 5.7.18). Then*

- (i) *there exists a unique $\vartheta \in \mathbb{R}^\mathfrak{d}$ such that ϑ is a global minimum point of \mathcal{L} and*

(ii) it holds that \mathcal{L} satisfies a coercivity-type condition at ϑ with coercivity constant c (cf. Definitions 5.6.2 and 5.7.25).

Proof of Corollary 5.7.27. Note that Corollary 5.7.22 and the assumption that \mathcal{L} is c -generalized convex and continuous show that there exists a unique $\vartheta \in \mathbb{R}^{\mathfrak{d}}$ which satisfies that ϑ is a global minimum point of \mathcal{L} (cf. Definition 5.6.2). This establishes item (i). Next we combine item (iii) in Proposition 5.7.23 and the assumption that \mathcal{L} is c -generalized convex to obtain that for all $v, w \in \mathbb{R}^{\mathfrak{d}}$ it holds that

$$\langle (\nabla \mathcal{L})(v) - (\nabla \mathcal{L})(w), v - w \rangle \geq c \|v - w\|_2^2 \quad (5.175)$$

(cf. Definitions 1.4.7 and 3.3.4). In addition, observe that Corollary 5.6.8 demonstrates that

$$(\nabla \mathcal{L})(\vartheta) = 0. \quad (5.176)$$

Combining this and (5.175) implies that it holds for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that

$$\langle \theta - \vartheta, (\nabla \mathcal{L})(\theta) \rangle = \langle \theta - \vartheta, (\nabla \mathcal{L})(\theta) - (\nabla \mathcal{L})(\vartheta) \rangle \geq c \|\theta - \vartheta\|_2^2. \quad (5.177)$$

This and (5.174) establish that \mathcal{L} satisfies a coercivity-type condition at ϑ with coercivity constant c (cf. Definition 5.7.25). The proof of Corollary 5.7.27 is thus complete. \square

Corollary 5.7.28. Let $\mathfrak{d} \in \mathbb{N}$, $\vartheta \in \mathbb{R}^{\mathfrak{d}}$ and let $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ be strongly convex (cf. Definition 5.7.19). Then it holds that \mathcal{L} is a coercive-type function (cf. Definition 5.7.26).

Proof of Corollary 5.7.28. Note that Corollary 5.7.27 ensures that \mathcal{L} is a coercive-type function (cf. Definition 5.7.26). The proof of Corollary 5.7.28 is thus complete. \square

Lemma 5.7.29 (A sufficient condition for a local minimum point). Let $\mathfrak{d} \in \mathbb{N}$, $c \in (0, \infty)$, $r \in (0, \infty]$, $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, $\mathbb{B} = \{w \in \mathbb{R}^{\mathfrak{d}} : \|w - \vartheta\|_2 \leq r\}$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ satisfy for all $\theta \in \mathbb{B}$ that

$$\langle \theta - \vartheta, (\nabla \mathcal{L})(\theta) \rangle \geq c \|\theta - \vartheta\|_2^2 \quad (5.178)$$

(cf. Definitions 1.4.7 and 3.3.4). Then

- (i) it holds for all $\theta \in \mathbb{B}$ that $\mathcal{L}(\theta) - \mathcal{L}(\vartheta) \geq \frac{c}{2} \|\theta - \vartheta\|_2^2$,
- (ii) it holds that $\{\theta \in \mathbb{B} : \mathcal{L}(\theta) = \inf_{w \in \mathbb{B}} \mathcal{L}(w)\} = \{\vartheta\}$, and
- (iii) it holds that $(\nabla \mathcal{L})(\vartheta) = 0$.

Proof of Lemma 5.7.29. Throughout this proof, let B be the set given by

$$B = \{w \in \mathbb{R}^{\mathfrak{d}} : \|w - \vartheta\|_2 < r\}. \quad (5.179)$$

Note that (5.178) implies that for all $v \in \mathbb{R}^{\mathfrak{d}}$ with $\|v\|_2 \leq r$ it holds that

$$\langle (\nabla \mathcal{L})(\vartheta + v), v \rangle \geq c\|v\|_2^2. \quad (5.180)$$

The fundamental theorem of calculus hence demonstrates that for all $\theta \in \mathbb{B}$ it holds that

$$\begin{aligned} \mathcal{L}(\theta) - \mathcal{L}(\vartheta) &= [\mathcal{L}(\vartheta + t(\theta - \vartheta))]_{t=0}^{t=1} \\ &= \int_0^1 \mathcal{L}'(\vartheta + t(\theta - \vartheta))(\theta - \vartheta) dt \\ &= \int_0^1 \langle (\nabla \mathcal{L})(\vartheta + t(\theta - \vartheta)), t(\theta - \vartheta) \rangle \frac{1}{t} dt \\ &\geq \int_0^1 c\|t(\theta - \vartheta)\|_2^2 \frac{1}{t} dt = c\|\theta - \vartheta\|_2^2 \left[\int_0^1 t dt \right] = \frac{c}{2}\|\theta - \vartheta\|_2^2. \end{aligned} \quad (5.181)$$

This proves item (i). Next observe that (5.181) ensures that for all $\theta \in \mathbb{B} \setminus \{\vartheta\}$ it holds that

$$\mathcal{L}(\theta) \geq \mathcal{L}(\vartheta) + \frac{c}{2}\|\theta - \vartheta\|_2^2 > \mathcal{L}(\vartheta). \quad (5.182)$$

Hence, we obtain for all $\theta \in \mathbb{B} \setminus \{\vartheta\}$ that

$$\inf_{w \in \mathbb{B}} \mathcal{L}(w) = \mathcal{L}(\vartheta) < \mathcal{L}(\theta). \quad (5.183)$$

This establishes item (ii). It thus remains thus remains to prove item (iii). For this observe that item (ii) ensures that

$$\{\theta \in B : \mathcal{L}(\theta) = \inf_{w \in B} \mathcal{L}(w)\} = \{\vartheta\}. \quad (5.184)$$

Combining this, the fact that B is open, and Lemma 5.6.7 (applied with $\mathfrak{d} \curvearrowright \mathfrak{d}$, $O \curvearrowright B$, $\vartheta \curvearrowright \vartheta$, $\mathcal{L} \curvearrowright \mathcal{L}|_B$ in the notation of Lemma 5.6.7) assures that $(\nabla \mathcal{L})(\vartheta) = 0$. This establishes item (iii). The proof of Lemma 5.7.29 is thus complete. \square

Example 5.7.30. Let $\mathfrak{d} \in \mathbb{N}$, $\xi \in \mathbb{R}^{\mathfrak{d}}$, $\vartheta = (\vartheta_1, \dots, \vartheta_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$, $\kappa, \lambda_1, \lambda_2, \dots, \lambda_{\mathfrak{d}} \in (0, \infty)$ satisfy $\kappa = \min\{\lambda_1, \lambda_2, \dots, \lambda_{\mathfrak{d}}\}$ and let $\mathcal{L} : \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$ satisfy for all $\theta = (\theta_1, \dots, \theta_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$ that

$$\mathcal{L}(\theta) = \frac{1}{2} \left[\sum_{i=1}^{\mathfrak{d}} \lambda_i |\theta_i - \vartheta_i|^2 \right]. \quad (5.185)$$

Then

- (i) it holds that \mathcal{L} is κ -generalized convex,
- (ii) it holds that \mathcal{L} is strongly convex,
- (iii) it holds that \mathcal{L} satisfies a coercivity-type condition at ϑ with coercivity constant κ , and
- (iv) it holds that \mathcal{L} is a coercive-type function

(cf. Definitions 5.7.18, 5.7.19, 5.7.25, and 5.7.26).

Proof for Example 5.7.30. Observe that (5.185) proves that for all $\theta = (\theta_1, \dots, \theta_d) \in \mathbb{R}^d$ it holds that

$$(\nabla \mathcal{L})(\theta) = (\lambda_1(\theta_1 - \vartheta_1), \dots, \lambda_d(\theta_d - \vartheta_d)). \quad (5.186)$$

Hence, we obtain that for all $v = (v_1, \dots, v_d), w = (w_1, \dots, w_d) \in \mathbb{R}^d$ it holds that

$$\begin{aligned} \langle (\nabla \mathcal{L})(v) - (\nabla \mathcal{L})(w), v - w \rangle &= \sum_{i=1}^d \lambda_i(v_i - w_i)(v_i - w_i) \\ &\geq \kappa \sum_{i=1}^d (v_i - w_i)^2 = \kappa \|v - w\|_2^2 \end{aligned} \quad (5.187)$$

(cf. Definitions 1.4.7 and 3.3.4). Combining this and Lemma 5.7.20 shows that \mathcal{L} is κ -generalized convex (cf. Definition 5.7.18). This establishes item (i). Note that item (i) and the fact that $(\nabla \mathcal{L})(\vartheta) = 0$ establish items (ii), (iii), and (iv). The proof for Example 5.7.30 is thus complete. \square

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 instance, Henry [204, 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*

$$\epsilon'(t) \leq \alpha \epsilon(t) + \beta(t). \quad (5.188)$$

Then it holds for all $t \in [0, T]$ that

$$\epsilon(t) \leq e^{\alpha t} \epsilon(0) + \int_0^t e^{\alpha(t-s)} \beta(s) ds. \quad (5.189)$$

Proof of Lemma 5.8.1. Throughout this proof, let $v: [0, T] \rightarrow \mathbb{R}$ satisfy for all $t \in [0, T]$ that

$$v(t) = e^{\alpha t} \left[\int_0^t e^{-\alpha s} \beta(s) ds \right] \quad (5.190)$$

and let $u: [0, T] \rightarrow \mathbb{R}$ satisfy for all $t \in [0, T]$ that

$$u(t) = [\epsilon(t) - v(t)]e^{-\alpha t}. \quad (5.191)$$

Observe that the product rule and the fundamental theorem of calculus demonstrate that for all $t \in [0, T]$ it holds that $v \in C^1([0, T], \mathbb{R})$ and

$$v'(t) = \alpha e^{\alpha t} \left[\int_0^t e^{-\alpha s} \beta(s) ds \right] + e^{\alpha t} [e^{-\alpha t} \beta(t)] = \alpha v(t) + \beta(t). \quad (5.192)$$

The assumption that $\epsilon \in C^1([0, T], \mathbb{R})$ and the product rule hence ensure that for all $t \in [0, T]$ it holds that $u \in C^1([0, T], \mathbb{R})$ and

$$\begin{aligned} u'(t) &= [\epsilon'(t) - v'(t)]e^{-\alpha t} - [\epsilon(t) - v(t)]\alpha e^{-\alpha t} \\ &= [\epsilon'(t) - v'(t) - \alpha \epsilon(t) + \alpha v(t)]e^{-\alpha t} \\ &= [\epsilon'(t) - \alpha v(t) - \beta(t) - \alpha \epsilon(t) + \alpha v(t)]e^{-\alpha t} \\ &= [\epsilon'(t) - \beta(t) - \alpha \epsilon(t)]e^{-\alpha t}. \end{aligned} \quad (5.193)$$

Combining this with the assumption that for all $t \in [0, T]$ it holds that $\epsilon'(t) \leq \alpha \epsilon(t) + \beta(t)$ proves that for all $t \in [0, T]$ it holds that

$$u'(t) \leq [\alpha \epsilon(t) + \beta(t) - \beta(t) - \alpha \epsilon(t)]e^{-\alpha t} = 0. \quad (5.194)$$

This and the fundamental theorem of calculus imply that for all $t \in [0, T]$ it holds that

$$u(t) = u(0) + \int_0^t u'(s) ds \leq u(0) + \int_0^t 0 ds = u(0) = \epsilon(0). \quad (5.195)$$

Combining this, (5.190), and (5.191) shows that for all $t \in [0, T]$ it holds that

$$\epsilon(t) = e^{\alpha t} u(t) + v(t) \leq e^{\alpha t} \epsilon(0) + v(t) = e^{\alpha t} \epsilon(0) + \int_0^t e^{\alpha(t-s)} \beta(s) ds. \quad (5.196)$$

The proof of Lemma 5.8.1 is thus complete. \square

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 $\beta \in C(O, \mathbb{R})$, $\mathcal{G} \in C(O, \mathbb{R}^{\mathfrak{d}})$, $V \in C^1(O, \mathbb{R})$ satisfy for all $\theta \in O$ that*

$$V'(\theta)\mathcal{G}(\theta) = \langle (\nabla V)(\theta), \mathcal{G}(\theta) \rangle \leq \alpha V(\theta) + \beta(\theta), \quad (5.197)$$

and let $\Theta \in C([0, T], O)$ satisfy for all $t \in [0, T]$ that $\Theta_t = \Theta_0 + \int_0^t \mathcal{G}(\Theta_s) ds$ (cf. Definition 1.4.7). Then it holds for all $t \in [0, T]$ that

$$V(\Theta_t) \leq e^{\alpha t} V(\Theta_0) + \int_0^t e^{\alpha(t-s)} \beta(\Theta_s) ds. \quad (5.198)$$

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

$$\epsilon(t) = V(\Theta_t) \quad \text{and} \quad b(t) = \beta(\Theta_t). \quad (5.199)$$

Note that (5.197), (5.199), the fundamental theorem of calculus, and the chain rule ensure that for all $t \in [0, T]$ it holds that

$$\epsilon'(t) = \frac{d}{dt}(V(\Theta_t)) = V'(\Theta_t)\dot{\Theta}_t = V'(\Theta_t)\mathcal{G}(\Theta_t) \leq \alpha V(\Theta_t) + \beta(\Theta_t) = \alpha\epsilon(t) + b(t). \quad (5.200)$$

Lemma 5.8.1 and (5.199) therefore demonstrate that for all $t \in [0, T]$ it holds that

$$V(\Theta_t) = \epsilon(t) \leq e^{\alpha t}\epsilon(0) + \int_0^t e^{\alpha(t-s)}b(s) ds = e^{\alpha t}V(\Theta_0) + \int_0^t e^{\alpha(t-s)}\beta(\Theta_s) ds. \quad (5.201)$$

The proof of Proposition 5.8.2 is thus complete. \square

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 $\mathcal{G} \in C(O, \mathbb{R}^{\mathfrak{d}})$, $V \in C^1(O, \mathbb{R})$ satisfy for all $\theta \in O$ that*

$$V'(\theta)\mathcal{G}(\theta) = \langle (\nabla V)(\theta), \mathcal{G}(\theta) \rangle \leq \alpha V(\theta), \quad (5.202)$$

and let $\Theta \in C([0, T], O)$ satisfy for all $t \in [0, T]$ that $\Theta_t = \Theta_0 + \int_0^t \mathcal{G}(\Theta_s) ds$ (cf. Definition 1.4.7). Then it holds for all $t \in [0, T]$ that

$$V(\Theta_t) \leq e^{\alpha t}V(\Theta_0). \quad (5.203)$$

Proof of Corollary 5.8.3. Observe that Proposition 5.8.2 and (5.202) imply (5.203). The proof of Corollary 5.8.3 is thus complete. \square

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 $V: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that*

$$V(\theta) = \|\theta - \vartheta\|_2^2 \quad (5.204)$$

(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

$$(\nabla V)(\theta) = 2(\theta - \vartheta). \quad (5.205)$$

Proof of Lemma 5.8.4. Throughout this proof, let $\vartheta_1, \dots, \vartheta_{\mathfrak{d}} \in \mathbb{R}$ satisfy $\vartheta = (\vartheta_1, \dots, \vartheta_{\mathfrak{d}})$. Note that the fact that for all $\theta = (\theta_1, \dots, \theta_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$ it holds that

$$V(\theta) = \sum_{i=1}^{\mathfrak{d}} (\theta_i - \vartheta_i)^2 \quad (5.206)$$

implies that for all $\theta = (\theta_1, \dots, \theta_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$ it holds that $V \in C^\infty(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ and

$$(\nabla V)(\theta) = \begin{pmatrix} \left(\frac{\partial V}{\partial \theta_1}\right)(\theta) \\ \vdots \\ \left(\frac{\partial V}{\partial \theta_{\mathfrak{d}}}\right)(\theta) \end{pmatrix} = \begin{pmatrix} 2(\theta_1 - \vartheta_1) \\ \vdots \\ 2(\theta_{\mathfrak{d}} - \vartheta_{\mathfrak{d}}) \end{pmatrix} = 2(\theta - \vartheta). \quad (5.207)$$

The proof of Lemma 5.8.4 is thus complete. \square

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 5.7.25.

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 $\mathcal{L} \in C^1(O, \mathbb{R})$ satisfy for all $\theta \in O$ that*

$$\langle \theta - \vartheta, (\nabla \mathcal{L})(\theta) \rangle \geq c \|\theta - \vartheta\|_2^2, \quad (5.208)$$

and let $\Theta \in C([0, T], O)$ satisfy for all $t \in [0, T]$ that $\Theta_t = \Theta_0 - \int_0^t (\nabla \mathcal{L})(\Theta_s) ds$ (cf. Definitions 1.4.7 and 3.3.4). Then it holds for all $t \in [0, T]$ that

$$\|\Theta_t - \vartheta\|_2 \leq e^{-ct} \|\Theta_0 - \vartheta\|_2. \quad (5.209)$$

Proof of Corollary 5.8.5. Throughout this proof, let $\mathcal{G}: O \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $\theta \in O$ that

$$\mathcal{G}(\theta) = -(\nabla \mathcal{L})(\theta) \quad (5.210)$$

and let $V: O \rightarrow \mathbb{R}$ satisfy for all $\theta \in O$ that

$$V(\theta) = \|\theta - \vartheta\|_2^2. \quad (5.211)$$

Observe that Lemma 5.8.4 and (5.208) ensure that for all $\theta \in O$ it holds that $V \in C^1(O, \mathbb{R})$ and

$$\begin{aligned} V'(\theta)\mathcal{G}(\theta) &= \langle (\nabla V)(\theta), \mathcal{G}(\theta) \rangle = \langle 2(\theta - \vartheta), \mathcal{G}(\theta) \rangle \\ &= -2\langle \theta - \vartheta, (\nabla \mathcal{L})(\theta) \rangle \leq -2c\|\theta - \vartheta\|_2^2 = -2cV(\theta). \end{aligned} \quad (5.212)$$

Corollary 5.8.3 hence proves that for all $t \in [0, T]$ it holds that

$$\|\Theta_t - \vartheta\|_2^2 = V(\Theta_t) \leq e^{-2ct} V(\Theta_0) = e^{-2ct} \|\Theta_0 - \vartheta\|_2^2. \quad (5.213)$$

The proof of Corollary 5.8.5 is thus complete. \square

5.8.4 On a linear growth condition

Lemma 5.8.6 (On a linear growth condition). *Let $\mathfrak{d} \in \mathbb{N}$, $L \in \mathbb{R}$, $r \in (0, \infty]$, $\vartheta \in \mathbb{R}^\mathfrak{d}$, $\mathbb{B} = \{w \in \mathbb{R}^\mathfrak{d} : \|w - \vartheta\|_2 \leq r\}$, $\mathcal{L} \in C^1(\mathbb{R}^\mathfrak{d}, \mathbb{R})$ satisfy for all $\theta \in \mathbb{B}$ that*

$$\|(\nabla \mathcal{L})(\theta)\|_2 \leq L\|\theta - \vartheta\|_2 \quad (5.214)$$

(cf. Definition 3.3.4). Then it holds for all $\theta \in \mathbb{B}$ that

$$\mathcal{L}(\theta) - \mathcal{L}(\vartheta) \leq \frac{L}{2}\|\theta - \vartheta\|_2^2. \quad (5.215)$$

Proof of Lemma 5.8.6. Observe that (5.214), the Cauchy-Schwarz inequality, and the fundamental theorem of calculus ensure that for all $\theta \in \mathbb{B}$ it holds that

$$\begin{aligned} \mathcal{L}(\theta) - \mathcal{L}(\vartheta) &= [\mathcal{L}(\vartheta + t(\theta - \vartheta))]_{t=0}^{t=1} \\ &= \int_0^1 \mathcal{L}'(\vartheta + t(\theta - \vartheta))(\theta - \vartheta) dt \\ &= \int_0^1 \langle (\nabla \mathcal{L})(\vartheta + t(\theta - \vartheta)), \theta - \vartheta \rangle dt \\ &\leq \int_0^1 \|(\nabla \mathcal{L})(\vartheta + t(\theta - \vartheta))\|_2 \|\theta - \vartheta\|_2 dt \\ &\leq \int_0^1 L\|\vartheta + t(\theta - \vartheta) - \vartheta\|_2 \|\theta - \vartheta\|_2 dt \\ &= L\|\theta - \vartheta\|_2^2 \left[\int_0^1 t dt \right] = \frac{L}{2}\|\theta - \vartheta\|_2^2 \end{aligned} \quad (5.216)$$

(cf. Definition 1.4.7). The proof of Lemma 5.8.6 is thus complete. \square

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 $\mathfrak{d} \in \mathbb{N}$, $c, T \in (0, \infty)$, $r \in (0, \infty]$, $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, $\mathbb{B} = \{w \in \mathbb{R}^{\mathfrak{d}} : \|w - \vartheta\|_2 \leq r\}$, $\xi \in \mathbb{B}$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ satisfy for all $\theta \in \mathbb{B}$ that*

$$\langle \theta - \vartheta, (\nabla \mathcal{L})(\theta) \rangle \geq c \|\theta - \vartheta\|_2^2, \quad (5.217)$$

and let $\Theta \in C([0, T], \mathbb{R}^{\mathfrak{d}})$ satisfy for all $t \in [0, T]$ that $\Theta_t = \xi - \int_0^t (\nabla \mathcal{L})(\Theta_s) ds$ (cf. Definitions 1.4.7 and 3.3.4). Then

- (i) *it holds that $\{\theta \in \mathbb{B} : \mathcal{L}(\theta) = \inf_{w \in \mathbb{B}} \mathcal{L}(w)\} = \{\vartheta\}$,*
- (ii) *it holds for all $t \in [0, T]$ that $\|\Theta_t - \vartheta\|_2 \leq e^{-ct} \|\xi - \vartheta\|_2$, and*
- (iii) *it holds for all $t \in [0, T]$ that*

$$0 \leq \frac{c}{2} \|\Theta_t - \vartheta\|_2^2 \leq \mathcal{L}(\Theta_t) - \mathcal{L}(\vartheta). \quad (5.218)$$

Proof of Proposition 5.9.1. Throughout this proof, let $V: \mathbb{R}^{\mathfrak{d}} \rightarrow [0, \infty)$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that $V(\theta) = \|\theta - \vartheta\|_2^2$, let $\epsilon: [0, T] \rightarrow [0, \infty)$ satisfy for all $t \in [0, T]$ that $\epsilon(t) = \|\Theta_t - \vartheta\|_2^2 = V(\Theta_t)$, and let $\tau \in [0, T]$ be the real number given by

$$\tau = \inf(\{t \in [0, T] : \Theta_t \notin \mathbb{B}\} \cup \{T\}) = \inf(\{t \in [0, T] : \epsilon(t) > r^2\} \cup \{T\}). \quad (5.219)$$

Note that (5.217) and item (ii) in Lemma 5.7.29 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

$$(\nabla V)(\theta) = 2(\theta - \vartheta). \quad (5.220)$$

Moreover, observe that the fundamental theorem of calculus (see, for example, Coleman [88, Theorem 3.9]) and the fact that $\mathbb{R}^{\mathfrak{d}} \ni v \mapsto (\nabla \mathcal{L})(v) \in \mathbb{R}^{\mathfrak{d}}$ and $\Theta: [0, T] \rightarrow \mathbb{R}^{\mathfrak{d}}$ are continuous functions ensure that for all $t \in [0, T]$ it holds that $\Theta \in C^1([0, T], \mathbb{R}^{\mathfrak{d}})$ and

$$\frac{d}{dt}(\Theta_t) = -(\nabla \mathcal{L})(\Theta_t). \quad (5.221)$$

Combining (5.217) and (5.220) hence demonstrates that for all $t \in [0, \tau]$ it holds that $\epsilon \in C^1([0, T], [0, \infty))$ and

$$\begin{aligned} \epsilon'(t) &= \frac{d}{dt}(V(\Theta_t)) = V'(\Theta_t)(\frac{d}{dt}(\Theta_t)) \\ &= \langle (\nabla V)(\Theta_t), \frac{d}{dt}(\Theta_t) \rangle \\ &= \langle 2(\Theta_t - \vartheta), -(\nabla \mathcal{L})(\Theta_t) \rangle \\ &= -2 \langle (\Theta_t - \vartheta), (\nabla \mathcal{L})(\Theta_t) \rangle \\ &\leq -2c \|\Theta_t - \vartheta\|_2^2 = -2c\epsilon(t). \end{aligned} \quad (5.222)$$

The Gronwall inequality, for instance, in Lemma 5.8.1 therefore implies that for all $t \in [0, \tau]$ it holds that

$$\epsilon(t) \leq \epsilon(0)e^{-2ct}. \quad (5.223)$$

Hence, we obtain for all $t \in [0, \tau]$ that

$$\|\Theta_t - \vartheta\|_2 = \sqrt{\epsilon(t)} \leq \sqrt{\epsilon(0)}e^{-ct} = \|\Theta_0 - \vartheta\|_2 e^{-ct} = \|\xi - \vartheta\|_2 e^{-ct}. \quad (5.224)$$

In the next step we prove that

$$\tau > 0. \quad (5.225)$$

In our proof of (5.225) we distinguish between the case $\varepsilon(0) = 0$ and the case $\varepsilon(0) > 0$. We first prove (5.225) in the case

$$\varepsilon(0) = 0. \quad (5.226)$$

Note that (5.226), the assumption that $r \in (0, \infty]$, and the fact that $\epsilon: [0, T] \rightarrow [0, \infty)$ is a continuous function show that

$$\tau = \inf(\{t \in [0, T]: \epsilon(t) > r^2\} \cup \{T\}) > 0. \quad (5.227)$$

This establishes (5.225) in the case $\varepsilon(0) = 0$. In the next step we prove (5.225) in the case

$$\varepsilon(0) > 0. \quad (5.228)$$

Observe that (5.222) and the assumption that $c \in (0, \infty)$ assure that for all $t \in [0, \tau]$ with $\epsilon(t) > 0$ it holds that

$$\epsilon'(t) \leq -2c\epsilon(t) < 0. \quad (5.229)$$

Combining this with (5.228) shows that

$$\epsilon'(0) < 0. \quad (5.230)$$

The fact that $\epsilon': [0, T] \rightarrow [0, \infty)$ is a continuous function and the assumption that $T \in (0, \infty)$ therefore demonstrate that

$$\inf(\{t \in [0, T]: \epsilon'(t) > 0\} \cup \{T\}) > 0. \quad (5.231)$$

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 $s < \inf(\{t \in [0, T]: \epsilon'(t) > 0\} \cup \{T\})$ it holds that

$$\epsilon(s) = \epsilon(0) + \int_0^s \epsilon'(u) du \leq \epsilon(0) = \|\xi - \vartheta\|_2^2 \leq r^2. \quad (5.232)$$

Combining this with (5.231) proves that

$$\tau = \inf(\{s \in [0, T]: \epsilon(s) > r^2\} \cup \{T\}) > 0. \quad (5.233)$$

This establishes (5.225) in the case $\varepsilon(0) > 0$. Note that (5.224), (5.225), and the assumption that $c \in (0, \infty)$ demonstrate that

$$\|\Theta_\tau - \vartheta\|_2 \leq \|\xi - \vartheta\|_2 e^{-c\tau} < r. \quad (5.234)$$

The fact that $\epsilon: [0, T] \rightarrow [0, \infty)$ is a continuous function, (5.219), and (5.225) hence assure that $\tau = T$. Combining this with (5.224) proves that for all $t \in [0, T]$ it holds that

$$\|\Theta_t - \vartheta\|_2 \leq \|\xi - \vartheta\|_2 e^{-ct}. \quad (5.235)$$

This establishes item (ii). It thus remains to prove item (iii). For this observe that (5.217) and item (i) in Lemma 5.7.29 demonstrate that for all $\theta \in \mathbb{B}$ it holds that

$$0 \leq \frac{c}{2} \|\theta - \vartheta\|_2^2 \leq \mathcal{L}(\theta) - \mathcal{L}(\vartheta). \quad (5.236)$$

Combining this and item (ii) implies that for all $t \in [0, T]$ it holds that

$$0 \leq \frac{c}{2} \|\Theta_t - \vartheta\|_2^2 \leq \mathcal{L}(\Theta_t) - \mathcal{L}(\vartheta) \quad (5.237)$$

This establishes item (iii). The proof of Proposition 5.9.1 is thus complete. \square

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 $\|\cdot\|: \mathbb{R}^\mathfrak{d} \rightarrow [0, \infty)$ be a norm, and let $\mathcal{G}: \mathbb{R}^\mathfrak{d} \rightarrow \mathbb{R}^\mathfrak{d}$ be locally Lipschitz continuous. Then there exist a unique real number $\tau \in (0, T]$ and a unique continuous function $\Theta: [0, \tau) \rightarrow \mathbb{R}^\mathfrak{d}$ such that for all $t \in [0, \tau)$ it holds that*

$$\liminf_{s \nearrow \tau} [\|\Theta_s\| + \frac{1}{(T-s)}] = \infty \quad \text{and} \quad \Theta_t = \xi + \int_0^t \mathcal{G}(\Theta_s) ds. \quad (5.238)$$

Proof of Lemma 5.9.2. Note that, for example, Teschl [415, Theorem 2.2 and Corollary 2.16] implies (5.238) (cf., for instance, [5, Theorem 7.6] and [233, Theorem 1.1]). The proof of Lemma 5.9.2 is thus complete. \square

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 $\|\cdot\|: \mathbb{R}^\mathfrak{d} \rightarrow [0, \infty)$ be a norm, and let $\mathcal{G}: \mathbb{R}^\mathfrak{d} \rightarrow \mathbb{R}^\mathfrak{d}$ be locally Lipschitz continuous. Then there exist a unique extended real number $\tau \in (0, \infty]$ and a unique continuous function $\Theta: [0, \tau) \rightarrow \mathbb{R}^\mathfrak{d}$ such that for all $t \in [0, \tau)$ it holds that*

$$\liminf_{s \nearrow \tau} [\|\Theta_s\| + s] = \infty \quad \text{and} \quad \Theta_t = \xi + \int_0^t \mathcal{G}(\Theta_s) ds. \quad (5.239)$$

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 $\Theta^{(n)}: [0, \tau_n) \rightarrow \mathbb{R}^d$, $n \in \mathbb{N}$, such that for all $n \in \mathbb{N}$, $t \in [0, \tau_n)$ it holds that

$$\liminf_{s \nearrow \tau_n} \left[\|\Theta_s^{(n)}\| + \frac{1}{(n-s)} \right] = \infty \quad \text{and} \quad \Theta_t^{(n)} = \xi + \int_0^t \mathcal{G}(\Theta_s^{(n)}) \, ds. \quad (5.240)$$

This shows that for all $n \in \mathbb{N}$, $t \in [0, \min\{\tau_{n+1}, n\})$ it holds that

$$\liminf_{s \nearrow \tau_{n+1}} \left[\|\Theta_s^{(n+1)}\| + \frac{1}{(n+1-s)} \right] = \infty \quad \text{and} \quad \Theta_t^{(n+1)} = \xi + \int_0^t \mathcal{G}(\Theta_s^{(n+1)}) \, ds. \quad (5.241)$$

Hence, we obtain that for all $n \in \mathbb{N}$, $t \in [0, \min\{\tau_{n+1}, n\})$ it holds that

$$\liminf_{s \nearrow \min\{\tau_{n+1}, n\}} \left[\|\Theta_s^{(n+1)}\| + \frac{1}{(n-s)} \right] = \infty \quad (5.242)$$

$$\text{and} \quad \Theta_t^{(n+1)} = \xi + \int_0^t \mathcal{G}(\Theta_s^{(n+1)}) \, ds. \quad (5.243)$$

Combining this with (5.240) demonstrates that for all $n \in \mathbb{N}$ it holds that

$$\tau_n = \min\{\tau_{n+1}, n\} \quad \text{and} \quad \Theta^{(n)} = \Theta^{(n+1)}|_{[0, \min\{\tau_{n+1}, n\}]}. \quad (5.244)$$

Therefore, we obtain that for all $n \in \mathbb{N}$ it holds that

$$\tau_n \leq \tau_{n+1} \quad \text{and} \quad \Theta^{(n)} = \Theta^{(n+1)}|_{[0, \tau_n]}. \quad (5.245)$$

Next let $\mathbf{t} \in (0, \infty]$ be the extended real number given by

$$\mathbf{t} = \lim_{n \rightarrow \infty} \tau_n \quad (5.246)$$

and let $\Theta: [0, \mathbf{t}) \rightarrow \mathbb{R}^d$ satisfy for all $n \in \mathbb{N}$, $t \in [0, \tau_n)$ that

$$\Theta_t = \Theta_t^{(n)}. \quad (5.247)$$

Observe that for all $t \in [0, \mathbf{t})$ there exists $n \in \mathbb{N}$ such that $t \in [0, \tau_n)$. This, (5.240), and (5.245) assure that for all $t \in [0, \mathbf{t})$ it holds that $\Theta \in C([0, \mathbf{t}), \mathbb{R}^d)$ and

$$\Theta_t = \xi + \int_0^t \mathcal{G}(\Theta_s) \, ds. \quad (5.248)$$

In addition, note that (5.244) ensures that for all $n \in \mathbb{N}$, $k \in \mathbb{N} \cap [n, \infty)$ it holds that

$$\min\{\tau_{k+1}, n\} = \min\{\tau_{k+1}, k, n\} = \min\{\min\{\tau_{k+1}, k\}, n\} = \min\{\tau_k, n\}. \quad (5.249)$$

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

$$\min\{\tau_k, n\} = \min\{\tau_{k-1}, n\} = \dots = \min\{\tau_{n+1}, n\} = \min\{\tau_n, n\} = \tau_n. \quad (5.250)$$

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

$$\min\{\mathbf{t}, n\} = \min\left\{\lim_{k \rightarrow \infty} \tau_k, n\right\} = \lim_{k \rightarrow \infty} (\min\{\tau_k, n\}) = \lim_{k \rightarrow \infty} \tau_n = \tau_n. \quad (5.251)$$

Therefore, we obtain that for all $n \in \mathbb{N}$ with $\mathbf{t} < n$ it holds that

$$\tau_n = \min\{\mathbf{t}, n\} = \mathbf{t}. \quad (5.252)$$

This, (5.240), and (5.247) demonstrate that for all $n \in \mathbb{N}$ with $\mathbf{t} < n$ it holds that

$$\begin{aligned} \liminf_{s \nearrow \mathbf{t}} \|\Theta_s\| &= \liminf_{s \nearrow \tau_n} \|\Theta_s\| = \liminf_{s \nearrow \tau_n} \|\Theta_s^{(n)}\| \\ &= -\frac{1}{(n-\mathbf{t})} + \liminf_{s \nearrow \tau_n} \left[\|\Theta_s^{(n)}\| + \frac{1}{(n-\mathbf{t})} \right] \\ &= -\frac{1}{(n-\mathbf{t})} + \liminf_{s \nearrow \tau_n} \left[\|\Theta_s^{(n)}\| + \frac{1}{(n-s)} \right] = \infty. \end{aligned} \quad (5.253)$$

Therefore, we obtain that

$$\liminf_{s \nearrow \mathbf{t}} [\|\Theta_s\| + s] = \infty. \quad (5.254)$$

Next note that for all $\hat{\mathbf{t}} \in (0, \infty]$, $\hat{\Theta} \in C([0, \hat{\mathbf{t}}], \mathbb{R}^d)$, $n \in \mathbb{N}$, $t \in [0, \min\{\hat{\mathbf{t}}, n\}]$ with $\liminf_{s \nearrow \hat{\mathbf{t}}} [\|\hat{\Theta}_s\| + s] = \infty$ and $\forall s \in [0, \hat{\mathbf{t}}]: \hat{\Theta}_s = \xi + \int_0^s \mathcal{G}(\hat{\Theta}_u) du$ it holds that

$$\liminf_{s \nearrow \min\{\hat{\mathbf{t}}, n\}} \left[\|\hat{\Theta}_s\| + \frac{1}{(n-s)} \right] = \infty \quad \text{and} \quad \hat{\Theta}_t = \xi + \int_0^t \mathcal{G}(\hat{\Theta}_s) ds. \quad (5.255)$$

This and (5.240) prove that for all $\hat{\mathbf{t}} \in (0, \infty]$, $\hat{\Theta} \in C([0, \hat{\mathbf{t}}], \mathbb{R}^d)$, $n \in \mathbb{N}$ with $\liminf_{t \nearrow \hat{\mathbf{t}}} [\|\hat{\Theta}_t\| + t] = \infty$ and $\forall t \in [0, \hat{\mathbf{t}}]: \hat{\Theta}_t = \xi + \int_0^t \mathcal{G}(\hat{\Theta}_s) ds$ it holds that

$$\min\{\hat{\mathbf{t}}, n\} = \tau_n \quad \text{and} \quad \hat{\Theta}|_{[0, \tau_n]} = \Theta^{(n)}. \quad (5.256)$$

Combining (5.248) and (5.254) hence assures that for all $\hat{\mathbf{t}} \in (0, \infty]$, $\hat{\Theta} \in C([0, \hat{\mathbf{t}}], \mathbb{R}^d)$, $n \in \mathbb{N}$ with $\liminf_{t \nearrow \hat{\mathbf{t}}} [\|\hat{\Theta}_t\| + t] = \infty$ and $\forall t \in [0, \hat{\mathbf{t}}]: \hat{\Theta}_t = \xi + \int_0^t \mathcal{G}(\hat{\Theta}_s) ds$ it holds that

$$\min\{\hat{\mathbf{t}}, n\} = \tau_n = \min\{\mathbf{t}, n\} \quad \text{and} \quad \hat{\Theta}|_{[0, \tau_n]} = \Theta^{(n)} = \Theta|_{[0, \tau_n]}. \quad (5.257)$$

This and (5.246) show that for all $\hat{\mathbf{t}} \in (0, \infty]$, $\hat{\Theta} \in C([0, \hat{\mathbf{t}}], \mathbb{R}^d)$ with $\liminf_{t \nearrow \hat{\mathbf{t}}} [\|\hat{\Theta}_t\| + t] = \infty$ and $\forall t \in [0, \hat{\mathbf{t}}]: \hat{\Theta}_t = \xi + \int_0^t \mathcal{G}(\hat{\Theta}_s) ds$ it holds that

$$\hat{\mathbf{t}} = \mathbf{t} \quad \text{and} \quad \hat{\Theta} = \Theta. \quad (5.258)$$

Combining this, (5.248), and (5.254) completes the proof of Lemma 5.9.3. \square

5.9.3 Approximation of local minimum points through GFs revisited

Theorem 5.9.4 (Approximation of local minimum points through GFs revisited). Let $\mathfrak{d} \in \mathbb{N}$, $c \in (0, \infty)$, $r \in (0, \infty]$, $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, $\mathbb{B} = \{w \in \mathbb{R}^{\mathfrak{d}} : \|w - \vartheta\|_2 \leq r\}$, $\xi \in \mathbb{B}$, $\mathcal{L} \in C^2(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ satisfy for all $\theta \in \mathbb{B}$ that

$$\langle \theta - \vartheta, (\nabla \mathcal{L})(\theta) \rangle \geq c \|\theta - \vartheta\|_2^2 \quad (5.259)$$

(cf. Definitions 1.4.7 and 3.3.4). Then

- (i) there exists a unique continuous function $\Theta: [0, \infty) \rightarrow \mathbb{R}^{\mathfrak{d}}$ such that for all $t \in [0, \infty)$ it holds that

$$\Theta_t = \xi - \int_0^t (\nabla \mathcal{L})(\Theta_s) ds, \quad (5.260)$$

(ii) it holds that $\{\theta \in \mathbb{B} : \mathcal{L}(\theta) = \inf_{w \in \mathbb{B}} \mathcal{L}(w)\} = \{\vartheta\}$,

(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

$$0 \leq \frac{c}{2} \|\Theta_t - \vartheta\|_2^2 \leq \mathcal{L}(\Theta_t) - \mathcal{L}(\vartheta). \quad (5.261)$$

Proof of Theorem 5.9.4. First, observe that the assumption that $\mathcal{L} \in C^2(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ ensures that

$$\mathbb{R}^{\mathfrak{d}} \ni \theta \mapsto -(\nabla \mathcal{L})(\theta) \in \mathbb{R}^{\mathfrak{d}} \quad (5.262)$$

is continuously differentiable. The fundamental theorem of calculus hence implies that

$$\mathbb{R}^{\mathfrak{d}} \ni \theta \mapsto -(\nabla \mathcal{L})(\theta) \in \mathbb{R}^{\mathfrak{d}} \quad (5.263)$$

is locally Lipschitz continuous. Combining this with Lemma 5.9.3 (applied with $\mathcal{G} \curvearrowleft (\mathbb{R}^{\mathfrak{d}} \ni \theta \mapsto -(\nabla \mathcal{L})(\theta) \in \mathbb{R}^{\mathfrak{d}})$ 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 $\Theta: [0, \tau) \rightarrow \mathbb{R}^{\mathfrak{d}}$ such that for all $t \in [0, \tau)$ it holds that

$$\liminf_{s \nearrow \tau} [\|\Theta_s\|_2 + s] = \infty \quad \text{and} \quad \Theta_t = \xi - \int_0^t (\nabla \mathcal{L})(\Theta_s) ds. \quad (5.264)$$

Next observe that Proposition 5.9.1 proves that for all $t \in [0, \tau)$ it holds that

$$\|\Theta_t - \vartheta\|_2 \leq e^{-ct} \|\xi - \vartheta\|_2. \quad (5.265)$$

This implies that

$$\begin{aligned} \liminf_{s \nearrow \tau} \|\Theta_s\|_2 &\leq \left[\liminf_{s \nearrow \tau} \|\Theta_s - \vartheta\|_2 \right] + \|\vartheta\|_2 \\ &\leq \left[\liminf_{s \nearrow \tau} e^{-cs} \|\xi - \vartheta\|_2 \right] + \|\vartheta\|_2 \leq \|\xi - \vartheta\|_2 + \|\vartheta\|_2 < \infty. \end{aligned} \quad (5.266)$$

This and (5.264) demonstrate that

$$\tau = \infty. \quad (5.267)$$

This and (5.264) 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. \square

5.9.4 Approximation error with respect to the objective function

Corollary 5.9.5 (Approximation error with respect to the objective function). *Let $\mathfrak{d} \in \mathbb{N}$, $c, L \in (0, \infty)$, $r \in (0, \infty]$, $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, $\mathbb{B} = \{w \in \mathbb{R}^{\mathfrak{d}} : \|w - \vartheta\|_2 \leq r\}$, $\xi \in \mathbb{B}$, $\mathcal{L} \in C^2(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ satisfy for all $\theta \in \mathbb{B}$ that*

$$\langle \theta - \vartheta, (\nabla \mathcal{L})(\theta) \rangle \geq c \|\theta - \vartheta\|_2^2 \quad \text{and} \quad \|(\nabla \mathcal{L})(\theta)\|_2 \leq L \|\theta - \vartheta\|_2 \quad (5.268)$$

(cf. Definitions 1.4.7 and 3.3.4). Then

- (i) there exists a unique continuous function $\Theta: [0, \infty) \rightarrow \mathbb{R}^{\mathfrak{d}}$ such that for all $t \in [0, \infty)$ it holds that

$$\Theta_t = \xi - \int_0^t (\nabla \mathcal{L})(\Theta_s) ds, \quad (5.269)$$

- (ii) it holds that $\{\theta \in \mathbb{B} : \mathcal{L}(\theta) = \inf_{w \in \mathbb{B}} \mathcal{L}(w)\} = \{\vartheta\}$,

- (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

$$0 \leq \frac{c}{2} \|\Theta_t - \vartheta\|_2^2 \leq \mathcal{L}(\Theta_t) - \mathcal{L}(\vartheta) \leq \frac{L}{2} \|\Theta_t - \vartheta\|_2^2 \leq \frac{L}{2} e^{-2ct} \|\xi - \vartheta\|_2^2. \quad (5.270)$$

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. \square

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 example, the books [9, Chapter 5], [54, Chapter 9], [59, Chapter 3], [171, Sections 4.3 and 5.9 and Chapter 8], [322], and [394, Chapter 14] and the references therein and, for instance, the survey articles [33, 50, 128, 375, 407] 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, [322, Section 1.2.3], [54, Section 9.3], and [59, 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 $\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$ be differentiable, let $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$, and let $\Theta: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a function. Then we say*

that Θ is the **GD** process for the objective function \mathcal{L} with learning rates $(\gamma_n)_{n \in \mathbb{N}}$ and initial value ξ if and only if it holds for all $n \in \mathbb{N}$ that

$$\Theta_0 = \xi \quad \text{and} \quad \Theta_n = \Theta_{n-1} - \gamma_n(\nabla \mathcal{L})(\Theta_{n-1}). \quad (6.1)$$

Algorithm 6.1.2: GD optimization method

Input: $\mathfrak{d}, N \in \mathbb{N}$, $\mathcal{L} \in C^1(\mathbb{R}^\mathfrak{d}, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $\xi \in \mathbb{R}^\mathfrak{d}$

Output: N -th step of the **GD** process for the objective function \mathcal{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  $n = 1, \dots, N$  do
3:    $\Theta \leftarrow \Theta - \gamma_n(\nabla \mathcal{L})(\Theta)$ 
4: return  $\Theta$ 

```

Exercise 6.1.1. Let $\xi = (\xi_1, \xi_2, \xi_3) \in \mathbb{R}^3$ satisfy $\xi = (1, 2, 3)$, let $\mathcal{L}: \mathbb{R}^3 \rightarrow \mathbb{R}$ satisfy for all $\theta = (\theta_1, \theta_2, \theta_3) \in \mathbb{R}^3$ that

$$\mathcal{L}(\theta) = 2(\theta_1)^2 + (\theta_2 + 1)^2 + (\theta_3 - 1)^2, \quad (6.2)$$

and let Θ be the **GD** process for the objective function \mathcal{L} with learning rates $\mathbb{N} \ni n \mapsto \frac{1}{2^n}$, and initial value ξ (cf. Definition 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 $\mathcal{L}: \mathbb{R}^3 \rightarrow \mathbb{R}$ satisfy for all $\theta = (\theta_1, \theta_2, \theta_3) \in \mathbb{R}^3$ that

$$\mathcal{L}(\theta) = (\theta_1)^2 + (\theta_2 - 1)^2 + 2(\theta_3 + 1)^2,$$

and let Θ be the **GD** process for the objective function \mathcal{L} with learning rates $\mathbb{N} \ni n \mapsto \frac{1}{3} \in [0, \infty)$ 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 $d, h, \mathfrak{d} \in \mathbb{N}$, $l_1, l_2, \dots, l_h \in \mathbb{N}$ satisfy $\mathfrak{d} = l_1(d+1) + [\sum_{k=2}^h l_k(l_{k-1}+1)] + l_h + 1$, let $a: \mathbb{R} \rightarrow \mathbb{R}$ be differentiable, let $M \in \mathbb{N}$, $x_1, x_2, \dots, x_M \in \mathbb{R}^d$, $y_1, y_2, \dots, y_M \in \mathbb{R}$,

let $\mathcal{L}: \mathbb{R}^d \rightarrow \mathbb{R}$ satisfy for all $\theta \in \mathbb{R}^d$ that

$$\mathcal{L}(\theta) = \frac{1}{M} \left[\sum_{m=1}^M \left| (\mathcal{N}_{\mathfrak{M}_{a,l_1}, \mathfrak{M}_{a,l_2}, \dots, \mathfrak{M}_{a,l_h}, \text{id}_{\mathbb{R}}})^{\theta,d}(x_m) - y_m \right|^2 \right], \quad (6.3)$$

let $\xi \in \mathbb{R}^d$, let $(\gamma_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, and let $\Theta: \mathbb{N}_0 \rightarrow \mathbb{R}^d$ satisfy for all $n \in \mathbb{N}$ that

$$\Theta_0 = \xi \quad \text{and} \quad \Theta_n = \Theta_{n-1} - \gamma_n (\nabla \mathcal{L})(\Theta_{n-1}) \quad (6.4)$$

(cf. Definitions 1.1.3 and 1.2.1 and Corollary 5.3.6). Then Θ is the **GD** process for the objective function \mathcal{L} with learning rates $(\gamma_n)_{n \in \mathbb{N}}$ and initial value ξ .

Proof for Example 6.1.3. Observe that (6.1) and (6.4) demonstrate that Θ is the **GD** process for the objective function \mathcal{L} with learning rates $(\gamma_n)_{n \in \mathbb{N}}$ and initial value ξ . The proof for Example 6.1.3 is thus complete. \square

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

$$f(b) = \left[\sum_{n=0}^{N-1} \frac{f^{(n)}(a)(b-a)^n}{n!} \right] + \int_0^1 \frac{f^{(N)}(a+r(b-a))(b-a)^N(1-r)^{N-1}}{(N-1)!} dr. \quad (6.5)$$

Proof of Theorem 6.1.4. Note that the fundamental theorem of calculus assures that for all $g \in C^1([0, 1], \mathbb{R})$ it holds that

$$g(1) = g(0) + \int_0^1 g'(r) dr = g(0) + \int_0^1 \frac{g'(r)(1-r)^0}{0!} dr. \quad (6.6)$$

Furthermore, observe that integration by parts ensures that for all $n \in \mathbb{N}$, $g \in C^{n+1}([0, 1], \mathbb{R})$ it holds that

$$\begin{aligned} \int_0^1 \frac{g^{(n)}(r)(1-r)^{n-1}}{(n-1)!} dr &= - \left[\frac{g^{(n)}(r)(1-r)^n}{n!} \right]_{r=0}^{r=1} + \int_0^1 \frac{g^{(n+1)}(r)(1-r)^n}{n!} dr \\ &= \frac{g^{(n)}(0)}{n!} + \int_0^1 \frac{g^{(n+1)}(r)(1-r)^n}{n!} dr. \end{aligned} \quad (6.7)$$

Combining this with (6.6) and induction shows that for all $g \in C^N([0, 1], \mathbb{R})$ it holds that

$$g(1) = \left[\sum_{n=0}^{N-1} \frac{g^{(n)}(0)}{n!} \right] + \int_0^1 \frac{g^{(N)}(r)(1-r)^{N-1}}{(N-1)!} dr. \quad (6.8)$$

This establishes (6.5). The proof of Theorem 6.1.4 is thus complete. \square

Lemma 6.1.5 (Local error of the Euler method). *Let $\mathfrak{d} \in \mathbb{N}$, $T, \gamma, c \in [0, \infty)$, $\mathcal{G} \in C^1(\mathbb{R}^\mathfrak{d}, \mathbb{R}^\mathfrak{d})$, $\Theta \in C([0, \infty), \mathbb{R}^\mathfrak{d})$, $\theta \in \mathbb{R}^\mathfrak{d}$ satisfy for all $x, y \in \mathbb{R}^\mathfrak{d}$, $t \in [0, \infty)$ that*

$$\Theta_t = \Theta_0 + \int_0^t \mathcal{G}(\Theta_s) ds, \quad \theta = \Theta_T + \gamma \mathcal{G}(\Theta_T), \quad (6.9)$$

$$\|\mathcal{G}(x)\|_2 \leq c, \quad \text{and} \quad \|\mathcal{G}'(x)y\|_2 \leq c\|y\|_2 \quad (6.10)$$

(cf. Definition 3.3.4). Then

$$\|\Theta_{T+\gamma} - \theta\|_2 \leq c^2\gamma^2. \quad (6.11)$$

Proof of Lemma 6.1.5. Note that the fundamental theorem of calculus, the hypothesis that $\mathcal{G} \in C^1(\mathbb{R}^\mathfrak{d}, \mathbb{R}^\mathfrak{d})$, and (6.9) show that for all $t \in (0, \infty)$ it holds that $\Theta \in C^1([0, \infty), \mathbb{R}^\mathfrak{d})$ and

$$\dot{\Theta}_t = \mathcal{G}(\Theta_t). \quad (6.12)$$

Combining this with the hypothesis that $\mathcal{G} \in C^1(\mathbb{R}^\mathfrak{d}, \mathbb{R}^\mathfrak{d})$ and the chain rule ensures that for all $t \in (0, \infty)$ it holds that $\Theta \in C^2([0, \infty), \mathbb{R}^\mathfrak{d})$ and

$$\ddot{\Theta}_t = \mathcal{G}'(\Theta_t)\dot{\Theta}_t = \mathcal{G}'(\Theta_t)\mathcal{G}(\Theta_t). \quad (6.13)$$

Theorem 6.1.4 and (6.12) hence imply that

$$\begin{aligned} \Theta_{T+\gamma} &= \Theta_T + \gamma \dot{\Theta}_T + \int_0^1 (1-r)\gamma^2 \ddot{\Theta}_{T+r\gamma} dr \\ &= \Theta_T + \gamma \mathcal{G}(\Theta_T) + \gamma^2 \int_0^1 (1-r)\mathcal{G}'(\Theta_{T+r\gamma})\mathcal{G}(\Theta_{T+r\gamma}) dr. \end{aligned} \quad (6.14)$$

This and (6.9) demonstrate that

$$\begin{aligned} &\|\Theta_{T+\gamma} - \theta\|_2 \\ &= \left\| \Theta_T + \gamma \mathcal{G}(\Theta_T) + \gamma^2 \int_0^1 (1-r)\mathcal{G}'(\Theta_{T+r\gamma})\mathcal{G}(\Theta_{T+r\gamma}) dr - (\Theta_T + \gamma \mathcal{G}(\Theta_T)) \right\|_2 \\ &\leq \gamma^2 \int_0^1 (1-r)\|\mathcal{G}'(\Theta_{T+r\gamma})\mathcal{G}(\Theta_{T+r\gamma})\|_2 dr \\ &\leq c^2\gamma^2 \int_0^1 r dr = \frac{c^2\gamma^2}{2} \leq c^2\gamma^2. \end{aligned} \quad (6.15)$$

The proof of Lemma 6.1.5 is thus complete. \square

Corollary 6.1.6 (Local error of the Euler method for GF ODEs). *Let $\mathfrak{d} \in \mathbb{N}$, $T, \gamma, c \in [0, \infty)$, $\mathcal{L} \in C^2(\mathbb{R}^\mathfrak{d}, \mathbb{R})$, $\Theta \in C([0, \infty), \mathbb{R}^\mathfrak{d})$, $\theta \in \mathbb{R}^\mathfrak{d}$ satisfy for all $x, y \in \mathbb{R}^\mathfrak{d}$, $t \in [0, \infty)$ that*

$$\Theta_t = \Theta_0 - \int_0^t (\nabla \mathcal{L})(\Theta_s) ds, \quad \theta = \Theta_T - \gamma(\nabla \mathcal{L})(\Theta_T), \quad (6.16)$$

$$\|(\nabla \mathcal{L})(x)\|_2 \leq c, \quad \text{and} \quad \|(\text{Hess } \mathcal{L})(x)y\|_2 \leq c\|y\|_2 \quad (6.17)$$

(cf. Definition 3.3.4). Then

$$\|\Theta_{T+\gamma} - \theta\|_2 \leq c^2 \gamma^2. \quad (6.18)$$

Proof of Corollary 6.1.6. Throughout this proof, let $\mathcal{G}: \mathbb{R}^\mathfrak{d} \rightarrow \mathbb{R}^\mathfrak{d}$ satisfy for all $\theta \in \mathbb{R}^\mathfrak{d}$ that

$$\mathcal{G}(\theta) = -(\nabla \mathcal{L})(\theta). \quad (6.19)$$

Note that the fact that for all $t \in [0, \infty)$ it holds that $\Theta_t = \Theta_0 + \int_0^t \mathcal{G}(\Theta_s) ds$, the fact that $\theta = \Theta_T + \gamma \mathcal{G}(\Theta_T)$, the fact that for all $x \in \mathbb{R}^\mathfrak{d}$ it holds that $\|\mathcal{G}(x)\|_2 \leq c$, the fact that for all $x, y \in \mathbb{R}^\mathfrak{d}$ it holds that $\|\mathcal{G}'(x)y\|_2 \leq c\|y\|_2$, and Lemma 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. \square

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 $V: \mathbb{R}^\mathfrak{d} \rightarrow \mathbb{R}$, $\Phi: \mathbb{R}^\mathfrak{d} \times [0, \infty) \rightarrow \mathbb{R}^\mathfrak{d}$, and $\varepsilon: [0, c] \rightarrow [0, \infty)$ satisfy for all $\theta \in \mathbb{R}^\mathfrak{d}$, $t \in [0, c]$ that*

$$V(\Phi(\theta, t)) \leq \varepsilon(t)V(\theta), \quad (6.20)$$

and let $\Theta: \mathbb{N}_0 \rightarrow \mathbb{R}^\mathfrak{d}$ satisfy for all $n \in \mathbb{N}$ that

$$\Theta_0 = \xi \quad \text{and} \quad \Theta_n = \Phi(\Theta_{n-1}, \gamma_n). \quad (6.21)$$

Then it holds for all $n \in \mathbb{N}_0$ that

$$V(\Theta_n) \leq \left[\prod_{k=1}^n \varepsilon(\gamma_k) \right] V(\xi). \quad (6.22)$$

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

$$\begin{aligned} V(\Theta_{n+1}) &= V(\Phi(\Theta_n, \gamma_{n+1})) \leq \varepsilon(\gamma_{n+1})V(\Theta_n) \\ &\leq \varepsilon(\gamma_{n+1}) \left(\left[\prod_{k=1}^n \varepsilon(\gamma_k) \right] V(\xi) \right) = \left[\prod_{k=1}^{n+1} \varepsilon(\gamma_k) \right] V(\xi). \end{aligned} \quad (6.23)$$

Induction thus establishes (6.22). The proof of Proposition 6.1.7 is thus complete. \square

Corollary 6.1.8 (On quadratic Lyapunov-type functions for the GD optimization method). *Let $\mathfrak{d} \in \mathbb{N}$, $\vartheta, \xi \in \mathbb{R}^\mathfrak{d}$, $c \in (0, \infty)$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, c]$, $\mathcal{L} \in C^1(\mathbb{R}^\mathfrak{d}, \mathbb{R})$, let $\|\cdot\|: \mathbb{R}^\mathfrak{d} \rightarrow [0, \infty)$ be a norm, let $\varepsilon: [0, c] \rightarrow [0, \infty)$ satisfy for all $\theta \in \mathbb{R}^\mathfrak{d}$, $t \in [0, c]$ that*

$$\|\theta - t(\nabla \mathcal{L})(\theta) - \vartheta\|^2 \leq \varepsilon(t)\|\theta - \vartheta\|^2, \quad (6.24)$$

and let $\Theta: \mathbb{N}_0 \rightarrow \mathbb{R}^\mathfrak{d}$ satisfy for all $n \in \mathbb{N}$ that

$$\Theta_0 = \xi \quad \text{and} \quad \Theta_n = \Theta_{n-1} - \gamma_n(\nabla \mathcal{L})(\Theta_{n-1}). \quad (6.25)$$

Then it holds for all $n \in \mathbb{N}_0$ that

$$\|\Theta_n - \vartheta\|^2 \leq \left[\prod_{k=1}^n [\varepsilon(\gamma_k)]^{1/2} \right] \|\xi - \vartheta\|^2. \quad (6.26)$$

Proof of Corollary 6.1.8. Throughout this proof, let $V: \mathbb{R}^\mathfrak{d} \rightarrow \mathbb{R}$ and $\Phi: \mathbb{R}^\mathfrak{d} \times [0, \infty) \rightarrow \mathbb{R}^\mathfrak{d}$ satisfy for all $\theta \in \mathbb{R}^\mathfrak{d}$, $t \in [0, \infty)$ that

$$V(\theta) = \|\theta - \vartheta\|^2 \quad \text{and} \quad \Phi(\theta, t) = \theta - t(\nabla \mathcal{L})(\theta). \quad (6.27)$$

Observe that Proposition 6.1.7 (applied with $V \curvearrowright V$, $\Phi \curvearrowright \Phi$ in the notation of Proposition 6.1.7) and (6.27) imply that for all $n \in \mathbb{N}_0$ it holds that

$$\|\Theta_n - \vartheta\|^2 = V(\Theta_n) \leq \left[\prod_{k=1}^n \varepsilon(\gamma_k) \right] V(\xi) = \left[\prod_{k=1}^n \varepsilon(\gamma_k) \right] \|\xi - \vartheta\|^2. \quad (6.28)$$

This establishes (6.26). The proof of Corollary 6.1.8 is thus complete. \square

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 (5.268) in Corollary 5.9.5 above. Results similar to Lemma 6.1.9 can, for example, be found in [109, Remark 2.1] and [232, 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 $\langle\cdot, \cdot\rangle: \mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$ be a scalar product, let $\|\cdot\|: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$ satisfy for all $v \in \mathbb{R}^{\mathfrak{d}}$ that $\|v\| = \sqrt{\langle v, v \rangle}$, and let $c, L \in (0, \infty)$, $r \in (0, \infty]$, $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, $\mathbb{B} = \{w \in \mathbb{R}^{\mathfrak{d}}: \|w - \vartheta\| \leq r\}$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ satisfy for all $\theta \in \mathbb{B}$ that*

$$\langle\langle\theta - \vartheta, (\nabla \mathcal{L})(\theta)\rangle\rangle \geq c\|\theta - \vartheta\|^2 \quad \text{and} \quad \|(\nabla \mathcal{L})(\theta)\| \leq L\|\theta - \vartheta\|. \quad (6.29)$$

Then

- (i) it holds that $c \leq L$,
- (ii) it holds for all $\theta \in \mathbb{B}$, $\gamma \in [0, \infty)$ that

$$\|\theta - \gamma(\nabla \mathcal{L})(\theta) - \vartheta\|^2 \leq (1 - 2\gamma c + \gamma^2 L^2)\|\theta - \vartheta\|^2, \quad (6.30)$$

(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

$$\|\theta - \gamma(\nabla \mathcal{L})(\theta) - \vartheta\|^2 \leq (1 - c\gamma)\|\theta - \vartheta\|^2. \quad (6.31)$$

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

$$\begin{aligned} 0 \leq \|\theta - \gamma(\nabla \mathcal{L})(\theta) - \vartheta\|^2 &= \|(\theta - \vartheta) - \gamma(\nabla \mathcal{L})(\theta)\|^2 \\ &= \|\theta - \vartheta\|^2 - 2\gamma \langle\langle\theta - \vartheta, (\nabla \mathcal{L})(\theta)\rangle\rangle + \gamma^2 \|(\nabla \mathcal{L})(\theta)\|^2 \\ &\leq \|\theta - \vartheta\|^2 - 2\gamma c\|\theta - \vartheta\|^2 + \gamma^2 L^2\|\theta - \vartheta\|^2 \\ &= (1 - 2\gamma c + \gamma^2 L^2)\|\theta - \vartheta\|^2. \end{aligned} \quad (6.32)$$

This establishes item (ii). Moreover, note that the fact that $\mathbb{B} \setminus \{\vartheta\} \neq \emptyset$ and (6.32) assure that for all $\gamma \in [0, \infty)$ it holds that

$$1 - 2\gamma c + \gamma^2 L^2 \geq 0. \quad (6.33)$$

Hence, we obtain that

$$\begin{aligned} 1 - \frac{c^2}{L^2} &= 1 - \frac{2c^2}{L^2} + \frac{c^2}{L^2} = 1 - 2\left[\frac{c}{L^2}\right]c + \left[\frac{c^2}{L^4}\right]L^2 \\ &= 1 - 2\left[\frac{c}{L^2}\right]c + \left[\frac{c}{L^2}\right]^2L^2 \geq 0. \end{aligned} \quad (6.34)$$

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

$$0 \leq 1 - 2\gamma c + \gamma^2 L^2 = 1 - \underbrace{\gamma}_{>0} \underbrace{(2c - \gamma L^2)}_{>0} < 1. \quad (6.35)$$

This proves item (iii). In addition, note that for all $\gamma \in [0, \frac{c}{L^2}]$ it holds that

$$1 - 2\gamma c + \gamma^2 L^2 \leq 1 - 2\gamma c + \gamma \left[\frac{c}{L^2}\right] L^2 = 1 - c\gamma. \quad (6.36)$$

Combining this with (6.32) establishes item (iv). The proof of Lemma 6.1.9 is thus complete. \square

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 $\mathcal{G}: \mathbb{R}^\mathfrak{d} \rightarrow \mathbb{R}^\mathfrak{d}$ such that $\|\theta - \vartheta\|_2 \leq r$, $\forall \xi \in \{w \in \mathbb{R}^\mathfrak{d}: \|w - \vartheta\|_2 \leq r\}: \|\xi - \gamma \mathcal{G}(\xi) - \vartheta\|_2 \leq \varepsilon \|\xi - \vartheta\|_2$, and

$$\langle \theta - \vartheta, \mathcal{G}(\theta) \rangle < \min\left\{\frac{1-\varepsilon^2}{2\gamma}, \frac{\gamma}{2}\right\} \max\{\|\theta - \vartheta\|_2^2, \|\mathcal{G}(\theta)\|_2^2\}. \quad (6.37)$$

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 $\mathcal{G}: \mathbb{R}^\mathfrak{d} \rightarrow \mathbb{R}^\mathfrak{d}$ which satisfies $\forall \theta \in \{w \in \mathbb{R}^\mathfrak{d}: \|w - \vartheta\|_2 \leq r\}: \langle \theta - \vartheta, \mathcal{G}(\theta) \rangle \geq \frac{1}{2} \max\{\|\theta - \vartheta\|_2^2, \|\mathcal{G}(\theta)\|_2^2\}$ it holds that

$$\forall \theta \in \{w \in \mathbb{R}^\mathfrak{d}: \|w - \vartheta\|_2 \leq r\}: (\langle \theta - \vartheta, \mathcal{G}(\theta) \rangle \geq \frac{1}{2} \|\theta - \vartheta\|_2^2 \wedge \|\mathcal{G}(\theta)\|_2 \leq 2\|\theta - \vartheta\|_2). \quad (6.38)$$

Exercise 6.1.5. Prove or disprove the following statement: For all $\mathfrak{d} \in \mathbb{N}$, $c \in (0, \infty)$, $r \in (0, \infty]$, $\vartheta, v \in \mathbb{R}^\mathfrak{d}$, $\mathcal{L} \in C^1(\mathbb{R}^\mathfrak{d}, \mathbb{R})$, $s, t \in [0, 1]$ such that $\|v\|_2 \leq r$, $s \leq t$, and $\forall \theta \in \{w \in \mathbb{R}^\mathfrak{d}: \|w - \vartheta\|_2 \leq r\}: \langle \theta - \vartheta, (\nabla \mathcal{L})(\theta) \rangle \geq c\|\theta - \vartheta\|_2^2$ it holds that

$$\mathcal{L}(\vartheta + tv) - \mathcal{L}(\vartheta + sv) \geq \frac{c}{2}(t^2 - s^2)\|v\|_2^2. \quad (6.39)$$

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 $\mathcal{L} \in C^1(\mathbb{R}^\mathfrak{d}, \mathbb{R})$ 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 $\mathcal{L}(\vartheta + tv) - \mathcal{L}(\vartheta + sv) \geq c(t^2 - s^2)\|v\|_2^2$ it holds that

$$\forall \theta \in \{w \in \mathbb{R}^\mathfrak{d}: \|w - \vartheta\|_2 \leq r\}: \langle \theta - \vartheta, (\nabla \mathcal{L})(\theta) \rangle \geq 2c\|\theta - \vartheta\|_2^2. \quad (6.40)$$

Exercise 6.1.7. Let $\mathfrak{d} \in \mathbb{N}$ and for every $v \in \mathbb{R}^\mathfrak{d}$, $R \in [0, \infty]$ let $\mathbb{B}_R(v) = \{w \in \mathbb{R}^\mathfrak{d}: \|w - v\|_2 \leq R\}$. Prove or disprove the following statement: For all $r \in (0, \infty]$, $\vartheta \in \mathbb{R}^\mathfrak{d}$, $\mathcal{L} \in C^1(\mathbb{R}^\mathfrak{d}, \mathbb{R})$ 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

$$\langle \theta - \vartheta, (\nabla \mathcal{L})(\theta) \rangle \geq c \|\theta - \vartheta\|_2^2. \quad (6.41)$$

(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

$$\mathcal{L}(\vartheta + t(v - \vartheta)) - \mathcal{L}(\vartheta + s(v - \vartheta)) \geq c(t^2 - s^2) \|v - \vartheta\|_2^2. \quad (6.42)$$

Exercise 6.1.8. Let $\mathfrak{d} \in \mathbb{N}$ and for every $v \in \mathbb{R}^\mathfrak{d}$, $R \in [0, \infty]$ let $\mathbb{B}_R(v) = \{w \in \mathbb{R}^\mathfrak{d} : \|v - w\|_2 \leq R\}$. Prove or disprove the following statement: For all $r \in (0, \infty]$, $\vartheta \in \mathbb{R}^\mathfrak{d}$, $\mathcal{L} \in C^1(\mathbb{R}^\mathfrak{d}, \mathbb{R})$ 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

$$\langle \theta - \vartheta, (\nabla \mathcal{L})(\theta) \rangle \geq c \|\theta - \vartheta\|_2^2 \quad \text{and} \quad \|(\nabla \mathcal{L})(\theta)\|_2 \leq L \|\theta - \vartheta\|_2. \quad (6.43)$$

(ii) There exist $\gamma \in (0, \infty)$, $\varepsilon \in (0, 1)$ such that for all $\theta \in \mathbb{B}_r(\vartheta)$ it holds that

$$\|\theta - \gamma(\nabla \mathcal{L})(\theta) - \vartheta\|_2 \leq \varepsilon \|\theta - \vartheta\|_2. \quad (6.44)$$

(iii) There exists $c \in (0, \infty)$ such that for all $\theta \in \mathbb{B}_r(\vartheta)$ it holds that

$$\langle \theta - \vartheta, (\nabla \mathcal{L})(\theta) \rangle \geq c \max\{\|\theta - \vartheta\|_2^2, \|(\nabla \mathcal{L})(\theta)\|_2^2\}. \quad (6.45)$$

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 $\mathfrak{d} \in \mathbb{N}$, $c, L \in (0, \infty)$, $r \in (0, \infty]$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \frac{2c}{L^2}]$, $\vartheta \in \mathbb{R}^\mathfrak{d}$, $\mathbb{B} = \{w \in \mathbb{R}^\mathfrak{d} : \|w - \vartheta\|_2 \leq r\}$, $\xi \in \mathbb{B}$, $\mathcal{L} \in C^1(\mathbb{R}^\mathfrak{d}, \mathbb{R})$ satisfy for all $\theta \in \mathbb{B}$ that*

$$\langle \theta - \vartheta, (\nabla \mathcal{L})(\theta) \rangle \geq c \|\theta - \vartheta\|_2^2 \quad \text{and} \quad \|(\nabla \mathcal{L})(\theta)\|_2 \leq L \|\theta - \vartheta\|_2, \quad (6.46)$$

and let $\Theta: \mathbb{N}_0 \rightarrow \mathbb{R}^\mathfrak{d}$ satisfy for all $n \in \mathbb{N}$ that

$$\Theta_0 = \xi \quad \text{and} \quad \Theta_n = \Theta_{n-1} - \gamma_n(\nabla \mathcal{L})(\Theta_{n-1}) \quad (6.47)$$

(cf. Definitions 1.4.7 and 3.3.4). Then

- (i) it holds that $\{\theta \in \mathbb{B} : \mathcal{L}(\theta) = \inf_{w \in \mathbb{B}} \mathcal{L}(w)\} = \{\vartheta\}$,
- (ii) it holds for all $n \in \mathbb{N}$ that $0 \leq 1 - 2c\gamma_n + (\gamma_n)^2 L^2 \leq 1$,
- (iii) it holds for all $n \in \mathbb{N}$ that $\|\Theta_n - \vartheta\|_2 \leq (1 - 2c\gamma_n + (\gamma_n)^2 L^2)^{1/2} \|\Theta_{n-1} - \vartheta\|_2 \leq r$,
- (iv) it holds for all $n \in \mathbb{N}_0$ that

$$\|\Theta_n - \vartheta\|_2 \leq \left[\prod_{k=1}^n (1 - 2c\gamma_k + (\gamma_k)^2 L^2)^{1/2} \right] \|\xi - \vartheta\|_2, \quad (6.48)$$

and

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

$$0 \leq \mathcal{L}(\Theta_n) - \mathcal{L}(\vartheta) \leq \frac{L}{2} \|\Theta_n - \vartheta\|_2^2 \leq \frac{L}{2} \left[\prod_{k=1}^n (1 - 2c\gamma_k + (\gamma_k)^2 L^2) \right] \|\xi - \vartheta\|_2^2. \quad (6.49)$$

Proof of Proposition 6.1.10. First, note that (6.46) and item (ii) in Lemma 5.7.29 prove item (i). Moreover, observe 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

$$1 - 2c \left[\frac{2c}{L^2} \right] + \left[\frac{2c}{L^2} \right]^2 L^2 = 1 - \frac{4c^2}{L^2} + \left[\frac{4c^2}{L^4} \right] L^2 = 1 - \frac{4c^2}{L^2} + \frac{4c^2}{L^2} = 1 \quad (6.50)$$

and establish item (ii). Next we claim that for all $n \in \mathbb{N}$ it holds that

$$\|\Theta_n - \vartheta\|_2 \leq (1 - 2c\gamma_n + (\gamma_n)^2 L^2)^{1/2} \|\Theta_{n-1} - \vartheta\|_2 \leq r. \quad (6.51)$$

We now prove (6.51) by induction on $n \in \mathbb{N}$. For the base case $n = 1$ note that (6.47), the assumption that $\Theta_0 = \xi \in \mathbb{B}$, item (ii) in Lemma 6.1.9, and item (ii) ensure that

$$\begin{aligned} \|\Theta_1 - \vartheta\|_2^2 &= \|\Theta_0 - \gamma_1(\nabla \mathcal{L})(\Theta_0) - \vartheta\|_2^2 \\ &\leq (1 - 2c\gamma_1 + (\gamma_1)^2 L^2) \|\Theta_0 - \vartheta\|_2^2 \\ &\leq \|\Theta_0 - \vartheta\|_2^2 \leq r^2. \end{aligned} \quad (6.52)$$

This establishes (6.51) in the base case $n = 1$. For the induction step observe 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

$$\begin{aligned} \|\Theta_{n+1} - \vartheta\|_2^2 &= \|\Theta_n - \gamma_{n+1}(\nabla \mathcal{L})(\Theta_n) - \vartheta\|_2^2 \\ &\leq \underbrace{(1 - 2c\gamma_{n+1} + (\gamma_{n+1})^2 L^2)}_{\in [0,1]} \|\Theta_n - \vartheta\|_2^2 \\ &\leq \|\Theta_n - \vartheta\|_2^2 \leq r^2. \end{aligned} \quad (6.53)$$

This demonstrates that for all $n \in \mathbb{N}$ with $\|\Theta_n - \vartheta\|_2 \leq r$ it holds that

$$\|\Theta_{n+1} - \vartheta\|_2 \leq (1 - 2c\gamma_{n+1} + (\gamma_{n+1})^2 L^2)^{1/2} \|\Theta_n - \vartheta\|_2 \leq r. \quad (6.54)$$

Induction thus proves (6.51). Next note that (6.51) establishes item (iii). Moreover, observe that induction, item (ii), and item (iii) prove item (iv). Furthermore, note that item (iii) and the fact that $\Theta_0 = \xi \in \mathbb{B}$ ensure that for all $n \in \mathbb{N}_0$ it holds that $\Theta_n \in \mathbb{B}$. Combining this, (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. \square

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 $\mathfrak{d} \in \mathbb{N}$, $c, L \in (0, \infty)$, $r \in (0, \infty]$, $\gamma \in (0, \frac{2c}{L^2})$, $\vartheta \in \mathbb{R}^\mathfrak{d}$, $\mathbb{B} = \{w \in \mathbb{R}^\mathfrak{d}: \|w - \vartheta\|_2 \leq r\}$, $\xi \in \mathbb{B}$, $\mathcal{L} \in C^1(\mathbb{R}^\mathfrak{d}, \mathbb{R})$ satisfy for all $\theta \in \mathbb{B}$ that*

$$\langle \theta - \vartheta, (\nabla \mathcal{L})(\theta) \rangle \geq c \|\theta - \vartheta\|_2^2 \quad \text{and} \quad \|(\nabla \mathcal{L})(\theta)\|_2 \leq L \|\theta - \vartheta\|_2, \quad (6.55)$$

and let $\Theta: \mathbb{N}_0 \rightarrow \mathbb{R}^\mathfrak{d}$ satisfy for all $n \in \mathbb{N}$ that

$$\Theta_0 = \xi \quad \text{and} \quad \Theta_n = \Theta_{n-1} - \gamma (\nabla \mathcal{L})(\Theta_{n-1}) \quad (6.56)$$

(cf. Definitions 1.4.7 and 3.3.4). Then

- (i) it holds that $\{\theta \in \mathbb{B}: \mathcal{L}(\theta) = \inf_{w \in \mathbb{B}} \mathcal{L}(w)\} = \{\vartheta\}$,
- (ii) it holds that $0 \leq 1 - 2c\gamma + \gamma^2 L^2 < 1$,
- (iii) it holds for all $n \in \mathbb{N}_0$ that

$$\|\Theta_n - \vartheta\|_2 \leq [1 - 2c\gamma + \gamma^2 L^2]^{n/2} \|\xi - \vartheta\|_2, \quad (6.57)$$

and

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

$$0 \leq \mathcal{L}(\Theta_n) - \mathcal{L}(\vartheta) \leq \frac{L}{2} \|\Theta_n - \vartheta\|_2^2 \leq \frac{L}{2} [1 - 2c\gamma + \gamma^2 L^2]^n \|\xi - \vartheta\|_2^2. \quad (6.58)$$

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. \square

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 $\mathfrak{d} \in \mathbb{N}$, $\alpha \in (0, \infty)$, $\gamma \in \mathbb{R}$, $\vartheta \in \mathbb{R}^\mathfrak{d}$, $\xi \in \mathbb{R}^\mathfrak{d} \setminus \{\vartheta\}$, let $\mathcal{L}: \mathbb{R}^\mathfrak{d} \rightarrow \mathbb{R}$ satisfy for all $\theta \in \mathbb{R}^\mathfrak{d}$ that*

$$\mathcal{L}(\theta) = \frac{\alpha}{2} \|\theta - \vartheta\|_2^2, \quad (6.59)$$

and let $\Theta: \mathbb{N}_0 \rightarrow \mathbb{R}^\mathfrak{d}$ satisfy for all $n \in \mathbb{N}$ that

$$\Theta_0 = \xi \quad \text{and} \quad \Theta_n = \Theta_{n-1} - \gamma(\nabla \mathcal{L})(\Theta_{n-1}) \quad (6.60)$$

(cf. Definition 3.3.4). Then

- (i) it holds for all $\theta \in \mathbb{R}^\mathfrak{d}$ that $\langle \theta - \vartheta, (\nabla \mathcal{L})(\theta) \rangle = \alpha \|\theta - \vartheta\|_2^2$,
- (ii) it holds for all $\theta \in \mathbb{R}^\mathfrak{d}$ that $\|(\nabla \mathcal{L})(\theta)\|_2 = \alpha \|\theta - \vartheta\|_2$,
- (iii) it holds for all $n \in \mathbb{N}_0$ that $\|\Theta_n - \vartheta\|_2 = |1 - \gamma \alpha|^n \|\xi - \vartheta\|_2$, and
- (iv) it holds that

$$\liminf_{n \rightarrow \infty} \|\Theta_n - \vartheta\|_2 = \limsup_{n \rightarrow \infty} \|\Theta_n - \vartheta\|_2 = \begin{cases} 0 & : \gamma \in (0, 2/\alpha) \\ \|\xi - \vartheta\|_2 & : \gamma \in \{0, 2/\alpha\} \\ \infty & : \gamma \in \mathbb{R} \setminus [0, 2/\alpha] \end{cases} \quad (6.61)$$

(cf. Definition 1.4.7).

Proof of Theorem 6.1.12. First of all, observe that Lemma 5.8.4 ensures that for all $\theta \in \mathbb{R}^\mathfrak{d}$ it holds that $\mathcal{L} \in C^\infty(\mathbb{R}^\mathfrak{d}, \mathbb{R})$ and

$$(\nabla \mathcal{L})(\theta) = \frac{\alpha}{2}(2(\theta - \vartheta)) = \alpha(\theta - \vartheta). \quad (6.62)$$

This proves item (ii). Moreover, observe that (6.62) assures that for all $\theta \in \mathbb{R}^\mathfrak{d}$ it holds that

$$\langle \theta - \vartheta, (\nabla \mathcal{L})(\theta) \rangle = \langle \theta - \vartheta, \alpha(\theta - \vartheta) \rangle = \alpha \|\theta - \vartheta\|_2^2 \quad (6.63)$$

(cf. Definition 1.4.7). This establishes item (i). Note that (6.60) and (6.62) demonstrate that for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned} \Theta_n - \vartheta &= \Theta_{n-1} - \gamma(\nabla \mathcal{L})(\Theta_{n-1}) - \vartheta \\ &= \Theta_{n-1} - \gamma \alpha(\Theta_{n-1} - \vartheta) - \vartheta \\ &= (1 - \gamma \alpha)(\Theta_{n-1} - \vartheta). \end{aligned} \quad (6.64)$$

The assumption that $\Theta_0 = \xi$ and induction hence prove that for all $n \in \mathbb{N}_0$ it holds that

$$\Theta_n - \vartheta = (1 - \gamma\alpha)^n(\Theta_0 - \vartheta) = (1 - \gamma\alpha)^n(\xi - \vartheta). \quad (6.65)$$

Therefore, we obtain for all $n \in \mathbb{N}_0$ that

$$\|\Theta_n - \vartheta\|_2 = |1 - \gamma\alpha|^n \|\xi - \vartheta\|_2. \quad (6.66)$$

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 $t \in \mathbb{R} \setminus [0, 2/\alpha]$ it holds that $|1 - t\alpha| \in (1, \infty)$, and the fact that $\|\xi - \vartheta\|_2 > 0$ establishes item (iv). The proof of Theorem 6.1.12 is thus complete. \square

Exercise 6.1.9. Let $\mathcal{L}: \mathbb{R} \rightarrow \mathbb{R}$ satisfy for all $\theta \in \mathbb{R}$ that

$$\mathcal{L}(\theta) = 2\theta^2 \quad (6.67)$$

and let $\Theta: \mathbb{N}_0 \rightarrow \mathbb{R}$ satisfy for all $n \in \mathbb{N}$ that $\Theta_0 = 1$ and

$$\Theta_n = \Theta_{n-1} - n^{-2}(\nabla \mathcal{L})(\Theta_{n-1}). \quad (6.68)$$

Prove or disprove the following statement: It holds that

$$\limsup_{n \rightarrow \infty} |\Theta_n| = 0. \quad (6.69)$$

Exercise 6.1.10. Let $\mathcal{L}: \mathbb{R} \rightarrow \mathbb{R}$ satisfy for all $\theta \in \mathbb{R}$ that

$$\mathcal{L}(\theta) = 4\theta^2 \quad (6.70)$$

and for every $r \in (1, \infty)$ let $\Theta^{(r)}: \mathbb{N}_0 \rightarrow \mathbb{R}$ satisfy for all $n \in \mathbb{N}$ that $\Theta_0^{(r)} = 1$ and

$$\Theta_n^{(r)} = \Theta_{n-1}^{(r)} - n^{-r}(\nabla \mathcal{L})(\Theta_{n-1}^{(r)}). \quad (6.71)$$

Prove or disprove the following statement: It holds for all $r \in (1, \infty)$ that

$$\liminf_{n \rightarrow \infty} |\Theta_n^{(r)}| > 0. \quad (6.72)$$

Exercise 6.1.11. Let $\mathcal{L}: \mathbb{R} \rightarrow \mathbb{R}$ satisfy for all $\theta \in \mathbb{R}$ that

$$\mathcal{L}(\theta) = 5\theta^2 \quad (6.73)$$

and for every $r \in (1, \infty)$ let $\Theta^{(r)} = (\Theta_n^{(r)})_{n \in \mathbb{N}_0}: \mathbb{N}_0 \rightarrow \mathbb{R}$ satisfy for all $n \in \mathbb{N}$ that $\Theta_0^{(r)} = 1$ and

$$\Theta_n^{(r)} = \Theta_{n-1}^{(r)} - n^{-r}(\nabla \mathcal{L})(\Theta_{n-1}^{(r)}). \quad (6.74)$$

Prove or disprove the following statement: It holds for all $r \in (1, \infty)$ that

$$\liminf_{n \rightarrow \infty} |\Theta_n^{(r)}| > 0. \quad (6.75)$$

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 $\mathfrak{d} \in \mathbb{N}$, $\mathcal{L} \in C^1(\mathbb{R}^\mathfrak{d}, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq \mathbb{R}$, $c, L \in (0, \infty)$, $\xi, \vartheta \in \mathbb{R}^\mathfrak{d}$ satisfy for all $\theta \in \mathbb{R}^\mathfrak{d}$ that*

$$\langle \theta - \vartheta, (\nabla \mathcal{L})(\theta) \rangle \geq c \|\theta - \vartheta\|_2^2, \quad \|(\nabla \mathcal{L})(\theta)\|_2 \leq L \|\theta - \vartheta\|_2, \quad (6.76)$$

$$\text{and } 0 < \liminf_{n \rightarrow \infty} \gamma_n \leq \limsup_{n \rightarrow \infty} \gamma_n < \frac{2c}{L^2}, \quad (6.77)$$

and let $\Theta: \mathbb{N}_0 \rightarrow \mathbb{R}^\mathfrak{d}$ satisfy for all $n \in \mathbb{N}$ that

$$\Theta_0 = \xi \quad \text{and} \quad \Theta_n = \Theta_{n-1} - \gamma_n (\nabla \mathcal{L})(\Theta_{n-1}) \quad (6.78)$$

(cf. Definitions 1.4.7 and 3.3.4). Then

- (i) it holds that $\{\theta \in \mathbb{R}^\mathfrak{d}: \mathcal{L}(\theta) = \inf_{w \in \mathbb{R}^\mathfrak{d}} \mathcal{L}(w)\} = \{\vartheta\}$,
- (ii) there exist $\epsilon \in (0, 1)$, $C \in \mathbb{R}$ such that for all $n \in \mathbb{N}_0$ it holds that

$$\|\Theta_n - \vartheta\|_2 \leq \epsilon^n C, \quad (6.79)$$

and

- (iii) there exist $\epsilon \in (0, 1)$, $C \in \mathbb{R}$ such that for all $n \in \mathbb{N}_0$ it holds that

$$0 \leq \mathcal{L}(\Theta_n) - \mathcal{L}(\vartheta) \leq \epsilon^n C. \quad (6.80)$$

Proof of Corollary 6.1.13. Throughout this proof, let $\alpha, \beta \in \mathbb{R}$ satisfy

$$0 < \alpha < \liminf_{n \rightarrow \infty} \gamma_n \leq \limsup_{n \rightarrow \infty} \gamma_n < \beta < \frac{2c}{L^2} \quad (6.81)$$

(cf. (6.77)), let $m \in \mathbb{N}$ satisfy for all $n \in \mathbb{N}$ that $\gamma_{m+n} \in [\alpha, \beta]$, and let $h: \mathbb{R} \rightarrow \mathbb{R}$ satisfy for all $t \in \mathbb{R}$ that

$$h(t) = 1 - 2ct + t^2 L^2. \quad (6.82)$$

Observe that (6.76) and item (ii) in Lemma 5.7.29 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

$$h'(t) \leq -2c + 2\left[\frac{c}{L^2}\right]L^2 = 0. \quad (6.83)$$

6.1. GD optimization

The fundamental theorem of calculus hence assures that for all $t \in [\alpha, \beta] \cap (-\infty, \frac{c}{L^2}]$ it holds that

$$h(t) = h(\alpha) + \int_{\alpha}^t h'(s) ds \leq h(\alpha) + \int_{\alpha}^t 0 ds = h(\alpha) \leq \max\{h(\alpha), h(\beta)\}. \quad (6.84)$$

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

$$h'(t) \geq h'(\frac{c}{L^2}) = -2c + 2[\frac{c}{L^2}]L^2 = 0. \quad (6.85)$$

The fundamental theorem of calculus hence ensures that for all $t \in [\alpha, \beta] \cap [\frac{c}{L^2}, \infty)$ it holds that

$$\max\{h(\alpha), h(\beta)\} \geq h(\beta) = h(t) + \int_t^{\beta} h'(s) ds \geq h(t) + \int_t^{\beta} 0 ds = h(t). \quad (6.86)$$

Combining this and (6.84) establishes that for all $t \in [\alpha, \beta]$ it holds that

$$h(t) \leq \max\{h(\alpha), h(\beta)\}. \quad (6.87)$$

Moreover, observe that the fact that $\alpha, \beta \in (0, \frac{2c}{L^2})$ and item (iii) in Lemma 6.1.9 ensure that

$$\{h(\alpha), h(\beta)\} \subseteq [0, 1]. \quad (6.88)$$

Hence, we obtain that

$$\max\{h(\alpha), h(\beta)\} \in [0, 1]. \quad (6.89)$$

This implies that there exists $\varepsilon \in \mathbb{R}$ such that

$$0 \leq \max\{h(\alpha), h(\beta)\} < \varepsilon < 1. \quad (6.90)$$

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 $\mathfrak{d} \curvearrowright \mathfrak{d}$, $c \curvearrowright c$, $L \curvearrowright L$, $r \curvearrowright \infty$, $(\gamma_n)_{n \in \mathbb{N}} \curvearrowright (\gamma_{m+n})_{n \in \mathbb{N}}$, $\vartheta \curvearrowright \vartheta$, $\xi \curvearrowright \Theta_m$, $\mathcal{L} \curvearrowright \mathcal{L}$ in the notation of Proposition 6.1.10), (6.76), (6.78), and (6.87) demonstrate that for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned} \|\Theta_{m+n} - \vartheta\|_2 &\leq \left[\prod_{k=1}^n (1 - 2c\gamma_{m+k} + (\gamma_{m+k})^2 L^2)^{1/2} \right] \|\Theta_m - \vartheta\|_2 \\ &= \left[\prod_{k=1}^n (h(\gamma_{m+k}))^{1/2} \right] \|\Theta_m - \vartheta\|_2 \\ &\leq (\max\{h(\alpha), h(\beta)\})^{n/2} \|\Theta_m - \vartheta\|_2 \\ &\leq \varepsilon^{n/2} \|\Theta_m - \vartheta\|_2. \end{aligned} \quad (6.91)$$

This shows that for all $n \in \mathbb{N}$ with $n > m$ it holds that

$$\|\Theta_n - \vartheta\|_2 \leq \varepsilon^{(n-m)/2} \|\Theta_m - \vartheta\|_2. \quad (6.92)$$

The fact that for all $n \in \mathbb{N}_0$ with $n \leq m$ it holds that

$$\|\Theta_n - \vartheta\|_2 = \left[\frac{\|\Theta_n - \vartheta\|_2}{\varepsilon^{n/2}} \right] \varepsilon^{n/2} \leq \left[\max \left\{ \frac{\|\Theta_k - \vartheta\|_2}{\varepsilon^{k/2}} : k \in \{0, 1, \dots, m\} \right\} \right] \varepsilon^{n/2} \quad (6.93)$$

hence assures that for all $n \in \mathbb{N}_0$ it holds that

$$\begin{aligned} \|\Theta_n - \vartheta\|_2 &\leq \max \left\{ \left[\max \left\{ \frac{\|\Theta_k - \vartheta\|_2}{\varepsilon^{k/2}} : k \in \{0, 1, \dots, m\} \right\} \right] \varepsilon^{n/2}, \varepsilon^{(n-m)/2} \|\Theta_m - \vartheta\|_2 \right\} \\ &= (\varepsilon^{1/2})^n \left[\max \left\{ \max \left\{ \frac{\|\Theta_k - \vartheta\|_2}{\varepsilon^{k/2}} : k \in \{0, 1, \dots, m\} \right\}, \varepsilon^{-m/2} \|\Theta_m - \vartheta\|_2 \right\} \right] \\ &= (\varepsilon^{1/2})^n \left[\max \left\{ \frac{\|\Theta_k - \vartheta\|_2}{\varepsilon^{k/2}} : k \in \{0, 1, \dots, m\} \right\} \right]. \end{aligned} \quad (6.94)$$

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

$$0 \leq \mathcal{L}(\Theta_n) - \mathcal{L}(\vartheta) \leq \frac{L}{2} \|\Theta_n - \vartheta\|_2^2 \leq \frac{\varepsilon^n L}{2} \left[\max \left\{ \frac{\|\Theta_k - \vartheta\|_2^2}{\varepsilon^k} : k \in \{0, 1, \dots, m\} \right\} \right]. \quad (6.95)$$

This establishes item (iii). The proof of Corollary 6.1.13 is thus complete. \square

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 $\mathfrak{d} \in \mathbb{N}$, $c, L \in (0, \infty)$, $r \in (0, \infty]$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \frac{c}{L^2}]$, $\vartheta \in \mathbb{R}^\mathfrak{d}$, $\mathbb{B} = \{w \in \mathbb{R}^\mathfrak{d} : \|w - \vartheta\|_2 \leq r\}$, $\xi \in \mathbb{B}$, $\mathcal{L} \in C^1(\mathbb{R}^\mathfrak{d}, \mathbb{R})$ satisfy for all $\theta \in \mathbb{B}$ that*

$$\langle \theta - \vartheta, (\nabla \mathcal{L})(\theta) \rangle \geq c \|\theta - \vartheta\|_2^2 \quad \text{and} \quad \|(\nabla \mathcal{L})(\theta)\|_2 \leq L \|\theta - \vartheta\|_2, \quad (6.96)$$

and let $\Theta: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $n \in \mathbb{N}$ that

$$\Theta_0 = \xi \quad \text{and} \quad \Theta_n = \Theta_{n-1} - \gamma_n (\nabla \mathcal{L})(\Theta_{n-1}) \quad (6.97)$$

(cf. Definitions 1.4.7 and 3.3.4). Then

- (i) it holds that $\{\theta \in \mathbb{B}: \mathcal{L}(\theta) = \inf_{w \in \mathbb{B}} \mathcal{L}(w)\} = \{\vartheta\}$,
- (ii) it holds for all $n \in \mathbb{N}$ that $0 \leq 1 - c\gamma_n \leq 1$,
- (iii) it holds for all $n \in \mathbb{N}_0$ that

$$\|\Theta_n - \vartheta\|_2 \leq \left[\prod_{k=1}^n (1 - c\gamma_k)^{1/2} \right] \|\xi - \vartheta\|_2 \leq \exp\left(-\frac{c}{2} [\sum_{k=1}^n \gamma_k]\right) \|\xi - \vartheta\|_2, \quad (6.98)$$

and

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

$$0 \leq \mathcal{L}(\Theta_n) - \mathcal{L}(\vartheta) \leq \frac{L}{2} \left[\prod_{k=1}^n (1 - c\gamma_k) \right] \|\xi - \vartheta\|_2^2 \leq \frac{L}{2} \exp\left(-c [\sum_{k=1}^n \gamma_k]\right) \|\xi - \vartheta\|_2^2. \quad (6.99)$$

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

$$0 \leq 1 - 2c\gamma_n + (\gamma_n)^2 L^2 \leq 1 - 2c\gamma_n + \gamma_n \left[\frac{c}{L^2} \right] L^2 = 1 - 2c\gamma_n + \gamma_n c = 1 - c\gamma_n \leq 1. \quad (6.100)$$

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. \square

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 $\mathfrak{d} \in \mathbb{N}$, $c, L \in (0, \infty)$, $r \in (0, \infty]$, $\gamma \in (0, \frac{c}{L^2}]$, $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, $\mathbb{B} = \{w \in \mathbb{R}^{\mathfrak{d}}: \|w - \vartheta\|_2 \leq r\}$, $\xi \in \mathbb{B}$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ satisfy for all $\theta \in \mathbb{B}$ that*

$$\langle \theta - \vartheta, (\nabla \mathcal{L})(\theta) \rangle \geq c \|\theta - \vartheta\|_2^2 \quad \text{and} \quad \|(\nabla \mathcal{L})(\theta)\|_2 \leq L \|\theta - \vartheta\|_2, \quad (6.101)$$

and let $\Theta: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $n \in \mathbb{N}$ that

$$\Theta_0 = \xi \quad \text{and} \quad \Theta_n = \Theta_{n-1} - \gamma (\nabla \mathcal{L})(\Theta_{n-1}) \quad (6.102)$$

(cf. Definitions 1.4.7 and 3.3.4). Then

- (i) it holds that $\{\theta \in \mathbb{B}: \mathcal{L}(\theta) = \inf_{w \in \mathbb{B}} \mathcal{L}(w)\} = \{\vartheta\}$,
- (ii) it holds that $0 \leq 1 - c\gamma < 1$,
- (iii) it holds for all $n \in \mathbb{N}_0$ that $\|\Theta_n - \vartheta\|_2 \leq (1 - c\gamma)^{n/2} \|\xi - \vartheta\|_2$, and
- (iv) it holds for all $n \in \mathbb{N}_0$ that $0 \leq \mathcal{L}(\Theta_n) - \mathcal{L}(\vartheta) \leq \frac{L}{2} (1 - c\gamma)^n \|\xi - \vartheta\|_2^2$.

Proof of Corollary 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. \square

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 instance, 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*

$$\mathcal{S} = \{\lambda \in \mathbb{C}: (\exists v \in \mathbb{C}^{\mathfrak{d}} \setminus \{0\}): Av = \lambda v\}. \quad (6.103)$$

Then

- (i) it holds that $\mathcal{S} = \{\lambda \in \mathbb{R}: (\exists v \in \mathbb{R}^{\mathfrak{d}} \setminus \{0\}): Av = \lambda v\} \subseteq \mathbb{R}$,

- (ii) it holds that

$$\sup_{v \in \mathbb{R}^{\mathfrak{d}} \setminus \{0\}} \left[\frac{\|Av\|_2}{\|v\|_2} \right] = \max_{\lambda \in \mathcal{S}} |\lambda|, \quad (6.104)$$

and

- (iii) it holds for all $v \in \mathbb{R}^{\mathfrak{d}}$ that

$$\min(\mathcal{S}) \|v\|_2^2 \leq \langle v, Av \rangle \leq \max(\mathcal{S}) \|v\|_2^2 \quad (6.105)$$

6.1. GD optimization

(cf. Definitions 1.4.7 and 3.3.4).

Proof of Lemma 6.1.16. Throughout this proof, let $e_1, e_2, \dots, e_{\mathfrak{d}} \in \mathbb{R}^{\mathfrak{d}}$ be the vectors given by

$$e_1 = (1, 0, \dots, 0), \quad e_2 = (0, 1, 0, \dots, 0), \quad \dots, \quad e_{\mathfrak{d}} = (0, \dots, 0, 1). \quad (6.106)$$

Observe that the spectral theorem for symmetric matrices (see, for example, Petersen [351, Theorem 4.3.4]) proves that there exist $(\mathfrak{d} \times \mathfrak{d})$ -matrices $\Lambda = (\Lambda_{i,j})_{(i,j) \in \{1,2,\dots,\mathfrak{d}\}^2}$, $O = (O_{i,j})_{(i,j) \in \{1,2,\dots,\mathfrak{d}\}^2} \in \mathbb{R}^{\mathfrak{d} \times \mathfrak{d}}$ such that $\mathcal{S} = \{\Lambda_{1,1}, \Lambda_{2,2}, \dots, \Lambda_{\mathfrak{d},\mathfrak{d}}\}$, $O^*O = OO^* = I_{\mathfrak{d}}$, $A = O\Lambda O^*$, and

$$\Lambda = \begin{pmatrix} \Lambda_{1,1} & & 0 \\ & \ddots & \\ 0 & & \Lambda_{\mathfrak{d},\mathfrak{d}} \end{pmatrix} \in \mathbb{R}^{\mathfrak{d} \times \mathfrak{d}} \quad (6.107)$$

(cf. Definition 1.5.5). Hence, we obtain that $\mathcal{S} \subseteq \mathbb{R}$. Next note that the assumption that $\mathcal{S} = \{\lambda \in \mathbb{C}: (\exists v \in \mathbb{C}^{\mathfrak{d}} \setminus \{0\}): Av = \lambda v\}$ ensures that for every $\lambda \in \mathcal{S}$ there exists $v \in \mathbb{C}^{\mathfrak{d}} \setminus \{0\}$ such that

$$A\Re(v) + iA\Im(v) = Av = \lambda v = \lambda\Re(v) + i\lambda\Im(v). \quad (6.108)$$

The fact that $\mathcal{S} \subseteq \mathbb{R}$ therefore demonstrates that for every $\lambda \in \mathcal{S}$ there exists $v \in \mathbb{R}^{\mathfrak{d}} \setminus \{0\}$ such that $Av = \lambda v$. This and the fact that $\mathcal{S} \subseteq \mathbb{R}$ ensure that $\mathcal{S} \subseteq \{\lambda \in \mathbb{R}: (\exists v \in \mathbb{R}^{\mathfrak{d}} \setminus \{0\}): Av = \lambda v\}$. Combining this and the fact that $\{\lambda \in \mathbb{R}: (\exists v \in \mathbb{R}^{\mathfrak{d}} \setminus \{0\}): Av = \lambda v\} \subseteq \mathcal{S}$ proves item (i). Furthermore, note that (6.107) assures that for all $v = (v_1, v_2, \dots, v_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$ it holds that

$$\begin{aligned} \|\Lambda v\|_2 &= \left[\sum_{i=1}^{\mathfrak{d}} |\Lambda_{i,i} v_i|^2 \right]^{1/2} \leq \left[\sum_{i=1}^{\mathfrak{d}} \max\{|\Lambda_{1,1}|^2, \dots, |\Lambda_{\mathfrak{d},\mathfrak{d}}|^2\} |v_i|^2 \right]^{1/2} \\ &= \left[\max\{|\Lambda_{1,1}|, \dots, |\Lambda_{\mathfrak{d},\mathfrak{d}}|\}^2 \|v\|_2^2 \right]^{1/2} \\ &= \max\{|\Lambda_{1,1}|, \dots, |\Lambda_{\mathfrak{d},\mathfrak{d}}|\} \|v\|_2 \\ &= (\max_{\lambda \in \mathcal{S}} |\lambda|) \|v\|_2 \end{aligned} \quad (6.109)$$

(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

$$\begin{aligned} \|Av\|_2 &= \|O\Lambda O^*v\|_2 = \|\Lambda O^*v\|_2 \\ &\leq (\max_{\lambda \in \mathcal{S}} |\lambda|) \|O^*v\|_2 \\ &= (\max_{\lambda \in \mathcal{S}} |\lambda|) \|v\|_2. \end{aligned} \quad (6.110)$$

This implies that

$$\sup_{v \in \mathbb{R}^{\mathfrak{d}} \setminus \{0\}} \left[\frac{\|Av\|_2}{\|v\|_2} \right] \leq \sup_{v \in \mathbb{R}^{\mathfrak{d}} \setminus \{0\}} \left[\frac{(\max_{\lambda \in \mathcal{S}} |\lambda|) \|v\|_2}{\|v\|_2} \right] = \max_{\lambda \in \mathcal{S}} |\lambda|. \quad (6.111)$$

In addition, note that the fact that $\mathcal{S} = \{\Lambda_{1,1}, \Lambda_{2,2}, \dots, \Lambda_{\mathfrak{d},\mathfrak{d}}\}$ ensures that there exists $j \in \{1, 2, \dots, \mathfrak{d}\}$ such that

$$|\Lambda_{j,j}| = \max_{\lambda \in \mathcal{S}} |\lambda|. \quad (6.112)$$

Next observe that the fact that $A = O\Lambda O^*$, the fact that O is an orthogonal matrix, and (6.112) imply that

$$\begin{aligned} \sup_{v \in \mathbb{R}^{\mathfrak{d}} \setminus \{0\}} \left[\frac{\|Av\|_2}{\|v\|_2} \right] &\geq \frac{\|AOe_j\|_2}{\|Oe_j\|_2} = \|O\Lambda O^* Oe_j\|_2 = \|O\Lambda e_j\|_2 \\ &= \|\Lambda e_j\|_2 = \|\Lambda_{j,j} e_j\|_2 = |\Lambda_{j,j}| = \max_{\lambda \in \mathcal{S}} |\lambda|. \end{aligned} \quad (6.113)$$

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 $v = (v_1, v_2, \dots, v_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$ it holds that

$$\begin{aligned} \langle v, \Lambda v \rangle &= \sum_{i=1}^{\mathfrak{d}} \Lambda_{i,i} |v_i|^2 \leq \sum_{i=1}^{\mathfrak{d}} \max\{\Lambda_{1,1}, \dots, \Lambda_{\mathfrak{d},\mathfrak{d}}\} |v_i|^2 \\ &= \max\{\Lambda_{1,1}, \dots, \Lambda_{\mathfrak{d},\mathfrak{d}}\} \|v\|_2^2 = \max(\mathcal{S}) \|v\|_2^2 \end{aligned} \quad (6.114)$$

(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

$$\begin{aligned} \langle v, Av \rangle &= \langle v, O\Lambda O^* v \rangle = \langle O^* v, \Lambda O^* v \rangle \\ &\leq \max(\mathcal{S}) \|O^* v\|_2^2 = \max(\mathcal{S}) \|v\|_2^2. \end{aligned} \quad (6.115)$$

Moreover, observe that (6.107) implies that for all $v = (v_1, v_2, \dots, v_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$ it holds that

$$\begin{aligned} \langle v, \Lambda v \rangle &= \sum_{i=1}^{\mathfrak{d}} \Lambda_{i,i} |v_i|^2 \geq \sum_{i=1}^{\mathfrak{d}} \min\{\Lambda_{1,1}, \dots, \Lambda_{\mathfrak{d},\mathfrak{d}}\} |v_i|^2 \\ &= \min\{\Lambda_{1,1}, \dots, \Lambda_{\mathfrak{d},\mathfrak{d}}\} \|v\|_2^2 = \min(\mathcal{S}) \|v\|_2^2. \end{aligned} \quad (6.116)$$

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

$$\begin{aligned} \langle v, Av \rangle &= \langle v, O\Lambda O^* v \rangle = \langle O^* v, \Lambda O^* v \rangle \\ &\geq \min(\mathcal{S}) \|O^* v\|_2^2 = \min(\mathcal{S}) \|v\|_2^2. \end{aligned} \quad (6.117)$$

Combining this with (6.115) establishes item (iii). The proof of Lemma 6.1.16 is thus complete. \square

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 $\|\cdot\|: \mathbb{R}^{\mathfrak{d} \times \mathfrak{d}} \rightarrow [0, \infty)$ 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 $\lambda, \alpha \in (0, \infty)$, $\beta \in [\alpha, \infty)$, $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, $\mathcal{L} \in C^2(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ satisfy for all $v, w \in \mathbb{R}^{\mathfrak{d}}$ that*

$$(\nabla \mathcal{L})(\vartheta) = 0, \quad \|(\text{Hess } \mathcal{L})(v) - (\text{Hess } \mathcal{L})(w)\| \leq \lambda \|v - w\|_2, \quad (6.118)$$

$$\text{and } \{\mu \in \mathbb{R}: (\exists u \in \mathbb{R}^{\mathfrak{d}} \setminus \{0\}: [(\text{Hess } \mathcal{L})(\vartheta)]u = \mu u)\} \subseteq [\alpha, \beta] \quad (6.119)$$

(cf. Definition 3.3.4). Then it holds for all $\theta \in \{w \in \mathbb{R}^{\mathfrak{d}}: \|w - \vartheta\|_2 \leq \frac{\alpha}{\lambda}\}$ that

$$\langle \theta - \vartheta, (\nabla \mathcal{L})(\theta) \rangle \geq \frac{\alpha}{2} \|\theta - \vartheta\|_2^2 \quad \text{and} \quad \|(\nabla \mathcal{L})(\theta)\|_2 \leq \frac{3\beta}{2} \|\theta - \vartheta\|_2 \quad (6.120)$$

(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

$$\mathbb{B} = \left\{ w \in \mathbb{R}^{\mathfrak{d}}: \|w - \vartheta\|_2 \leq \frac{\alpha}{\lambda} \right\} \quad (6.121)$$

and let $\mathcal{S} \subseteq \mathbb{C}$ be the set given by

$$\mathcal{S} = \{\mu \in \mathbb{C}: (\exists u \in \mathbb{C}^{\mathfrak{d}} \setminus \{0\}: [(\text{Hess } \mathcal{L})(\vartheta)]u = \mu u)\}. \quad (6.122)$$

Note that the fact that $(\text{Hess } \mathcal{L})(\vartheta) \in \mathbb{R}^{\mathfrak{d} \times \mathfrak{d}}$ is a symmetric matrix, item (i) in Lemma 6.1.16, and (6.119) imply that

$$\mathcal{S} = \{\mu \in \mathbb{R}: (\exists u \in \mathbb{R}^{\mathfrak{d}} \setminus \{0\}: [(\text{Hess } \mathcal{L})(\vartheta)]u = \mu u)\} \subseteq [\alpha, \beta]. \quad (6.123)$$

Next observe that the assumption that $(\nabla \mathcal{L})(\vartheta) = 0$ and the fundamental theorem of

calculus ensure that for all $\theta, w \in \mathbb{R}^{\mathfrak{d}}$ it holds that

$$\begin{aligned}
 \langle w, (\nabla \mathcal{L})(\theta) \rangle &= \langle w, (\nabla \mathcal{L})(\theta) - (\nabla \mathcal{L})(\vartheta) \rangle \\
 &= \left\langle w, [(\nabla \mathcal{L})(\vartheta + t(\theta - \vartheta))]_{t=0}^{t=1} \right\rangle \\
 &= \left\langle w, \int_0^1 [(\text{Hess } \mathcal{L})(\vartheta + t(\theta - \vartheta))](\theta - \vartheta) dt \right\rangle \\
 &= \int_0^1 \langle w, [(\text{Hess } \mathcal{L})(\vartheta + t(\theta - \vartheta))](\theta - \vartheta) \rangle dt \\
 &= \langle w, [(\text{Hess } \mathcal{L})(\vartheta)](\theta - \vartheta) \rangle \\
 &\quad + \int_0^1 \langle w, [(\text{Hess } \mathcal{L})(\vartheta + t(\theta - \vartheta)) - (\text{Hess } \mathcal{L})(\vartheta)](\theta - \vartheta) \rangle dt
 \end{aligned} \tag{6.124}$$

(cf. Definition 1.4.7). The fact that $(\text{Hess } \mathcal{L})(\vartheta) \in \mathbb{R}^{\mathfrak{d} \times \mathfrak{d}}$ 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

$$\begin{aligned}
 &\langle \theta - \vartheta, (\nabla \mathcal{L})(\theta) \rangle \\
 &\geq \langle \theta - \vartheta, [(\text{Hess } \mathcal{L})(\vartheta)](\theta - \vartheta) \rangle \\
 &\quad - \left| \int_0^1 \langle \theta - \vartheta, [(\text{Hess } \mathcal{L})(\vartheta + t(\theta - \vartheta)) - (\text{Hess } \mathcal{L})(\vartheta)](\theta - \vartheta) \rangle dt \right| \\
 &\geq \min(\mathcal{S}) \|\theta - \vartheta\|_2^2 \\
 &\quad - \int_0^1 \|\theta - \vartheta\|_2 \|[(\text{Hess } \mathcal{L})(\vartheta + t(\theta - \vartheta)) - (\text{Hess } \mathcal{L})(\vartheta)](\theta - \vartheta)\|_2 dt.
 \end{aligned} \tag{6.125}$$

Combining this with (6.123) and (6.118) shows that for all $\theta \in \mathbb{B}$ it holds that

$$\begin{aligned}
 &\langle \theta - \vartheta, (\nabla \mathcal{L})(\theta) \rangle \\
 &\geq \alpha \|\theta - \vartheta\|_2^2 \\
 &\quad - \int_0^1 \|\theta - \vartheta\|_2 \|[(\text{Hess } \mathcal{L})(\vartheta + t(\theta - \vartheta)) - (\text{Hess } \mathcal{L})(\vartheta)](\theta - \vartheta)\|_2 dt \\
 &\geq \alpha \|\theta - \vartheta\|_2^2 - \left[\int_0^1 \lambda \|\vartheta + t(\theta - \vartheta) - \vartheta\|_2 dt \right] \|\theta - \vartheta\|_2^2 \\
 &= \left(\alpha - \left[\int_0^1 t dt \right] \lambda \|\theta - \vartheta\|_2 \right) \|\theta - \vartheta\|_2^2 = \left(\alpha - \frac{\lambda}{2} \|\theta - \vartheta\|_2 \right) \|\theta - \vartheta\|_2^2 \\
 &\geq \left(\alpha - \frac{\lambda \alpha}{2\lambda} \right) \|\theta - \vartheta\|_2^2 = \frac{\alpha}{2} \|\theta - \vartheta\|_2^2.
 \end{aligned} \tag{6.126}$$

Moreover, observe that (6.118), (6.123), (6.124), the fact that $(\text{Hess } \mathcal{L})(\vartheta) \in \mathbb{R}^{\mathfrak{d} \times \mathfrak{d}}$ 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

$$\begin{aligned}
 & \langle w, (\nabla \mathcal{L})(\theta) \rangle \\
 & \leq |\langle w, [(\text{Hess } \mathcal{L})(\vartheta)](\theta - \vartheta) \rangle| \\
 & \quad + \left| \int_0^1 \langle w, [(\text{Hess } \mathcal{L})(\vartheta + t(\theta - \vartheta)) - (\text{Hess } \mathcal{L})(\vartheta)](\theta - \vartheta) \rangle dt \right| \\
 & \leq \|w\|_2 \|[(\text{Hess } \mathcal{L})(\vartheta)](\theta - \vartheta)\|_2 \\
 & \quad + \int_0^1 \|w\|_2 \|[(\text{Hess } \mathcal{L})(\vartheta + t(\theta - \vartheta)) - (\text{Hess } \mathcal{L})(\vartheta)](\theta - \vartheta)\|_2 dt \\
 & \leq \left[\sup_{v \in \mathbb{R}^{\mathfrak{d}} \setminus \{0\}} \frac{\|[(\text{Hess } \mathcal{L})(\vartheta)]v\|_2}{\|v\|_2} \right] \|\theta - \vartheta\|_2 \\
 & \quad + \int_0^1 \|(\text{Hess } \mathcal{L})(\vartheta + t(\theta - \vartheta)) - (\text{Hess } \mathcal{L})(\vartheta)\| \|\theta - \vartheta\|_2 dt \\
 & \leq \max(\mathcal{S}) \|\theta - \vartheta\|_2 + \left[\int_0^1 \lambda \|\vartheta + t(\theta - \vartheta) - \vartheta\|_2 dt \right] \|\theta - \vartheta\|_2 \\
 & \leq \left(\beta + \lambda \left[\int_0^1 t dt \right] \|\theta - \vartheta\|_2 \right) \|\theta - \vartheta\|_2 = \left(\beta + \frac{\lambda}{2} \|\theta - \vartheta\|_2 \right) \|\theta - \vartheta\|_2 \\
 & \leq \left(\beta + \frac{\lambda \alpha}{2\lambda} \right) \|\theta - \vartheta\|_2 = \left[\frac{2\beta + \alpha}{2} \right] \|\theta - \vartheta\|_2 \leq \frac{3\beta}{2} \|\theta - \vartheta\|_2.
 \end{aligned} \tag{6.127}$$

Therefore, we obtain for all $\theta \in \mathbb{B}$ that

$$\|(\nabla \mathcal{L})(\theta)\|_2 = \sup_{w \in \mathbb{R}^{\mathfrak{d}}, \|w\|_2=1} [\langle w, (\nabla \mathcal{L})(\theta) \rangle] \leq \frac{3\beta}{2} \|\theta - \vartheta\|_2. \tag{6.128}$$

Combining this and (6.126) establishes (6.120). The proof of Proposition 6.1.17 is thus complete. \square

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 instance, be found in Nesterov [322, 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 $\|\cdot\|: \mathbb{R}^{\mathfrak{d} \times \mathfrak{d}} \rightarrow \mathbb{R}$ 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 $\lambda, \alpha \in (0, \infty)$, $\beta \in [\alpha, \infty)$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \frac{4\alpha}{9\beta^2}]$, $\vartheta, \xi \in \mathbb{R}^{\mathfrak{d}}$, $\mathcal{L} \in C^2(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ satisfy for all $v, w \in \mathbb{R}^{\mathfrak{d}}$ that*

$$(\nabla \mathcal{L})(\vartheta) = 0, \quad \|(\text{Hess } \mathcal{L})(v) - (\text{Hess } \mathcal{L})(w)\| \leq \lambda \|v - w\|_2, \tag{6.129}$$

$$\{\mu \in \mathbb{R}: (\exists u \in \mathbb{R}^{\mathfrak{d}} \setminus \{0\}: [(\text{Hess } \mathcal{L})(\vartheta)]u = \mu u)\} \subseteq [\alpha, \beta], \tag{6.130}$$

and $\|\xi - \vartheta\|_2 \leq \frac{\alpha}{\lambda}$, and let $\Theta: \mathbb{N}_0 \rightarrow \mathbb{R}^d$ satisfy for all $n \in \mathbb{N}$ that

$$\Theta_0 = \xi \quad \text{and} \quad \Theta_n = \Theta_{n-1} - \gamma_n(\nabla \mathcal{L})(\Theta_{n-1}) \quad (6.131)$$

(cf. Definition 3.3.4). Then

- (i) it holds that $\{\theta \in \mathbb{B}: \mathcal{L}(\theta) = \inf_{w \in \mathbb{B}} \mathcal{L}(w)\} = \{\vartheta\}$,
- (ii) it holds for all $k \in \mathbb{N}$ 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

$$\|\Theta_n - \vartheta\|_2 \leq \left[\prod_{k=1}^n \left[1 - \alpha \gamma_k + \frac{9\beta^2(\gamma_k)^2}{4} \right]^{1/2} \right] \|\xi - \vartheta\|_2, \quad (6.132)$$

and

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

$$0 \leq \mathcal{L}(\Theta_n) - \mathcal{L}(\vartheta) \leq \frac{3\beta}{4} \left[\prod_{k=1}^n \left[1 - \alpha \gamma_k + \frac{9\beta^2(\gamma_k)^2}{4} \right] \right] \|\xi - \vartheta\|_2^2. \quad (6.133)$$

Proof of Corollary 6.1.18. Note that (6.129), (6.130), and Proposition 6.1.17 prove that for all $\theta \in \{w \in \mathbb{R}^d: \|w - \vartheta\|_2 \leq \frac{\alpha}{\lambda}\}$ it holds that

$$\langle \theta - \vartheta, (\nabla \mathcal{L})(\theta) \rangle \geq \frac{\alpha}{2} \|\theta - \vartheta\|_2^2 \quad \text{and} \quad \|(\nabla \mathcal{L})(\theta)\|_2 \leq \frac{3\beta}{2} \|\theta - \vartheta\|_2 \quad (6.134)$$

(cf. Definition 1.4.7). Combining this, the assumption that

$$\|\xi - \vartheta\|_2 \leq \frac{\alpha}{\lambda}, \quad (6.135)$$

(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. \square

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 $\mathcal{L}: \mathbb{R}^d \rightarrow \mathbb{R}$ 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}^d and that the rank of the Hessian of \mathcal{L} is on this set of local minimum points locally (at least) $d - k$ (cf. Fehrman et al. [138] 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 [235]).

6.1.4.6 Equivalent conditions on the objective function

Lemma 6.1.20. *Let $d \in \mathbb{N}$, let $\langle\langle \cdot, \cdot \rangle\rangle: \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ be a scalar product, let $\|\cdot\|: \mathbb{R}^d \rightarrow \mathbb{R}$ satisfy for all $v \in \mathbb{R}^d$ that $\|v\| = \sqrt{\langle\langle v, v \rangle\rangle}$, let $\gamma \in (0, \infty)$, $\varepsilon \in (0, 1)$, $r \in (0, \infty]$, $\vartheta \in \mathbb{R}^d$, $\mathbb{B} = \{w \in \mathbb{R}^d : \|w - \vartheta\| \leq r\}$, and let $\mathcal{G}: \mathbb{R}^d \rightarrow \mathbb{R}^d$ satisfy for all $\theta \in \mathbb{B}$ that*

$$\|\theta - \gamma \mathcal{G}(\theta) - \vartheta\| \leq \varepsilon \|\theta - \vartheta\|. \quad (6.136)$$

Then it holds for all $\theta \in \mathbb{B}$ that

$$\begin{aligned} \langle\langle \theta - \vartheta, \mathcal{G}(\theta) \rangle\rangle &\geq \max \left\{ \left[\frac{1-\varepsilon^2}{2\gamma} \right] \|\theta - \vartheta\|^2, \frac{\gamma}{2} \|\mathcal{G}(\theta)\|^2 \right\} \\ &\geq \min \left\{ \frac{1-\varepsilon^2}{2\gamma}, \frac{\gamma}{2} \right\} \max \left\{ \|\theta - \vartheta\|^2, \|\mathcal{G}(\theta)\|^2 \right\}. \end{aligned} \quad (6.137)$$

Proof of Lemma 6.1.20. First, note that (6.136) ensures that for all $\theta \in \mathbb{B}$ it holds that

$$\begin{aligned} \varepsilon^2 \|\theta - \vartheta\|^2 &\geq \|\theta - \gamma \mathcal{G}(\theta) - \vartheta\|^2 = \|\theta - \vartheta - \gamma \mathcal{G}(\theta)\|^2 \\ &= \|\theta - \vartheta\|^2 - 2\gamma \langle\langle \theta - \vartheta, \mathcal{G}(\theta) \rangle\rangle + \gamma^2 \|\mathcal{G}(\theta)\|^2. \end{aligned} \quad (6.138)$$

Hence, we obtain for all $\theta \in \mathbb{B}$ that

$$\begin{aligned} 2\gamma \langle\langle \theta - \vartheta, \mathcal{G}(\theta) \rangle\rangle &\geq (1 - \varepsilon^2) \|\theta - \vartheta\|^2 + \gamma^2 \|\mathcal{G}(\theta)\|^2 \\ &\geq \max \left\{ (1 - \varepsilon^2) \|\theta - \vartheta\|^2, \gamma^2 \|\mathcal{G}(\theta)\|^2 \right\} \geq 0. \end{aligned} \quad (6.139)$$

This demonstrates that for all $\theta \in \mathbb{B}$ it holds that

$$\begin{aligned} \langle\langle \theta - \vartheta, \mathcal{G}(\theta) \rangle\rangle &\geq \frac{1}{2\gamma} \max \left\{ (1 - \varepsilon^2) \|\theta - \vartheta\|^2, \gamma^2 \|\mathcal{G}(\theta)\|^2 \right\} \\ &= \max \left\{ \left[\frac{1-\varepsilon^2}{2\gamma} \right] \|\theta - \vartheta\|^2, \frac{\gamma}{2} \|\mathcal{G}(\theta)\|^2 \right\} \\ &\geq \min \left\{ \frac{1-\varepsilon^2}{2\gamma}, \frac{\gamma}{2} \right\} \max \left\{ \|\theta - \vartheta\|^2, \|\mathcal{G}(\theta)\|^2 \right\}. \end{aligned} \quad (6.140)$$

The proof of Lemma 6.1.20 is thus complete. \square

Lemma 6.1.21. *Let $d \in \mathbb{N}$, let $\langle\langle \cdot, \cdot \rangle\rangle: \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ be a scalar product, let $\|\cdot\|: \mathbb{R}^d \rightarrow \mathbb{R}$ satisfy for all $v \in \mathbb{R}^d$ that $\|v\| = \sqrt{\langle\langle v, v \rangle\rangle}$, let $c \in (0, \infty)$, $r \in (0, \infty]$, $\vartheta \in \mathbb{R}^d$,*

$\mathbb{B} = \{w \in \mathbb{R}^d : \|w - \vartheta\| \leq r\}$, and let $\mathcal{G} : \mathbb{R}^d \rightarrow \mathbb{R}^d$ satisfy for all $\theta \in \mathbb{B}$ that

$$\langle\langle \theta - \vartheta, \mathcal{G}(\theta) \rangle\rangle \geq c \max\{\|\theta - \vartheta\|^2, \|\mathcal{G}(\theta)\|^2\}. \quad (6.141)$$

Then it holds for all $\theta \in \mathbb{B}$ that

$$\langle\langle \theta - \vartheta, \mathcal{G}(\theta) \rangle\rangle \geq c \|\theta - \vartheta\|^2 \quad \text{and} \quad \|\mathcal{G}(\theta)\| \leq \frac{1}{c} \|\theta - \vartheta\|. \quad (6.142)$$

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

$$\|\mathcal{G}(\theta)\|^2 \leq \max\{\|\theta - \vartheta\|^2, \|\mathcal{G}(\theta)\|^2\} \leq \frac{1}{c} \langle\langle \theta - \vartheta, \mathcal{G}(\theta) \rangle\rangle \leq \frac{1}{c} \|\theta - \vartheta\| \|\mathcal{G}(\theta)\|. \quad (6.143)$$

Therefore, we obtain for all $\theta \in \mathbb{B}$ that

$$\|\mathcal{G}(\theta)\| \leq \frac{1}{c} \|\theta - \vartheta\|. \quad (6.144)$$

Combining this with (6.141) completes the proof of Lemma 6.1.21. \square

Lemma 6.1.22. Let $d \in \mathbb{N}$, $c \in (0, \infty)$, $r \in (0, \infty]$, $\vartheta \in \mathbb{R}^d$, $\mathbb{B} = \{w \in \mathbb{R}^d : \|w - \vartheta\|_2 \leq r\}$, $\mathcal{L} \in C^1(\mathbb{R}^d, \mathbb{R})$ satisfy for all $\theta \in \mathbb{B}$ that

$$\langle \theta - \vartheta, (\nabla \mathcal{L})(\theta) \rangle \geq c \|\theta - \vartheta\|_2^2. \quad (6.145)$$

Then it holds for all $v \in \mathbb{R}^d$, $s, t \in [0, 1]$ with $\|v\|_2 \leq r$ and $s \leq t$ that

$$\mathcal{L}(\vartheta + tv) - \mathcal{L}(\vartheta + sv) \geq \frac{c}{2}(t^2 - s^2) \|v\|_2^2. \quad (6.146)$$

Proof of Lemma 6.1.22. First of all, observe that (6.145) implies that for all $v \in \mathbb{R}^d$ with $\|v\|_2 \leq r$ it holds that

$$\langle (\nabla \mathcal{L})(\vartheta + v), v \rangle \geq c \|v\|_2^2. \quad (6.147)$$

The fundamental theorem of calculus hence ensures that for all $v \in \mathbb{R}^d$, $s, t \in [0, 1]$ with $\|v\|_2 \leq r$ and $s \leq t$ it holds that

$$\begin{aligned} \mathcal{L}(\vartheta + tv) - \mathcal{L}(\vartheta + sv) &= [\mathcal{L}(\vartheta + hv)]_{h=s}^{h=t} \\ &= \int_s^t \mathcal{L}'(\vartheta + hv)v dh \\ &= \int_s^t \frac{1}{h} \langle (\nabla \mathcal{L})(\vartheta + hv), hv \rangle dh \\ &\geq \int_s^t \frac{c}{h} \|hv\|_2^2 dh \\ &= c \left[\int_s^t h dh \right] \|v\|_2^2 = \frac{c}{2}(t^2 - s^2) \|v\|_2^2. \end{aligned} \quad (6.148)$$

The proof of Lemma 6.1.22 is thus complete. \square

Lemma 6.1.23. Let $\mathfrak{d} \in \mathbb{N}$, $c \in (0, \infty)$, $r \in (0, \infty]$, $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, $\mathbb{B} = \{w \in \mathbb{R}^{\mathfrak{d}} : \|w - \vartheta\|_2 \leq r\}$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ satisfy for all $v \in \mathbb{R}^{\mathfrak{d}}$, $s, t \in [0, 1]$ with $\|v\|_2 \leq r$ and $s \leq t$ that

$$\mathcal{L}(v + tv) - \mathcal{L}(v + sv) \geq c(t^2 - s^2)\|v\|_2^2 \quad (6.149)$$

(cf. Definition 3.3.4). Then it holds for all $\theta \in \mathbb{B}$ that

$$\langle \theta - \vartheta, (\nabla \mathcal{L})(\theta) \rangle \geq 2c\|\theta - \vartheta\|_2^2 \quad (6.150)$$

(cf. Definition 1.4.7).

Proof of Lemma 6.1.23. Observe that (6.149) ensures that for all $s \in (0, r] \cap \mathbb{R}$, $\theta \in \mathbb{R}^{\mathfrak{d}} \setminus \{\vartheta\}$ with $\|\theta - \vartheta\|_2 < s$ it holds that

$$\begin{aligned} \langle \theta - \vartheta, (\nabla \mathcal{L})(\theta) \rangle &= \mathcal{L}'(\theta)(\theta - \vartheta) = \lim_{h \searrow 0} \left(\frac{1}{h} [\mathcal{L}(\theta + h(\theta - \vartheta)) - \mathcal{L}(\theta)] \right) \\ &= \lim_{h \searrow 0} \left(\frac{1}{h} \left[\mathcal{L} \left(\vartheta + \frac{(1+h)\|\theta-\vartheta\|_2}{s} \left(\frac{s}{\|\theta-\vartheta\|_2} (\theta - \vartheta) \right) \right) \right. \right. \\ &\quad \left. \left. - \mathcal{L} \left(\vartheta + \frac{\|\theta-\vartheta\|_2}{s} \left(\frac{s}{\|\theta-\vartheta\|_2} (\theta - \vartheta) \right) \right) \right] \right) \\ &\geq \limsup_{h \searrow 0} \left(\frac{c}{h} \left(\left[\frac{(1+h)\|\theta-\vartheta\|_2}{s} \right]^2 - \left[\frac{\|\theta-\vartheta\|_2}{s} \right]^2 \right) \left\| \frac{s}{\|\theta-\vartheta\|_2} (\theta - \vartheta) \right\|_2^2 \right) \\ &= c \left[\limsup_{h \searrow 0} \left(\frac{(1+h)^2 - 1}{h} \right) \right] \left[\frac{\|\theta-\vartheta\|_2}{s} \right]^2 \left\| \frac{s}{\|\theta-\vartheta\|_2} (\theta - \vartheta) \right\|_2^2 \\ &= c \left[\limsup_{h \searrow 0} \left(\frac{2h+h^2}{h} \right) \right] \|\theta - \vartheta\|_2^2 \\ &= c \left[\limsup_{h \searrow 0} (2 + h) \right] \|\theta - \vartheta\|_2^2 = 2c\|\theta - \vartheta\|_2^2 \end{aligned} \quad (6.151)$$

(cf. Definition 1.4.7). Hence, we obtain that for all $\theta \in \mathbb{R}^{\mathfrak{d}} \setminus \{\vartheta\}$ with $\|\theta - \vartheta\|_2 < r$ it holds that

$$\langle \theta - \vartheta, (\nabla \mathcal{L})(\theta) \rangle \geq 2c\|\theta - \vartheta\|_2^2. \quad (6.152)$$

Combining this with the fact that the function

$$\mathbb{R}^{\mathfrak{d}} \ni v \mapsto (\nabla \mathcal{L})(v) \in \mathbb{R}^{\mathfrak{d}} \quad (6.153)$$

is continuous establishes (6.150). The proof of Lemma 6.1.23 is thus complete. \square

Lemma 6.1.24. Let $\mathfrak{d} \in \mathbb{N}$, $L \in (0, \infty)$, $r \in (0, \infty]$, $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, $\mathbb{B} = \{w \in \mathbb{R}^{\mathfrak{d}} : \|w - \vartheta\|_2 \leq r\}$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ satisfy for all $\theta \in \mathbb{B}$ that

$$\|(\nabla \mathcal{L})(\theta)\|_2 \leq L\|\theta - \vartheta\|_2 \quad (6.154)$$

(cf. Definition 3.3.4). Then it holds for all $v, w \in \mathbb{B}$ that

$$|\mathcal{L}(v) - \mathcal{L}(w)| \leq L \max\{\|v - \vartheta\|_2, \|w - \vartheta\|_2\} \|v - w\|_2. \quad (6.155)$$

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

$$\begin{aligned} |\mathcal{L}(v) - \mathcal{L}(w)| &= \left| [\mathcal{L}(w + h(v - w))]_{h=0}^{h=1} \right| \\ &= \left| \int_0^1 \mathcal{L}'(w + h(v - w))(v - w) dh \right| \\ &= \left| \int_0^1 \langle (\nabla \mathcal{L})(w + h(v - w)), v - w \rangle dh \right| \\ &\leq \int_0^1 \|(\nabla \mathcal{L})(hv + (1-h)w)\|_2 \|v - w\|_2 dh \\ &\leq \int_0^1 L\|hv + (1-h)w - \vartheta\|_2 \|v - w\|_2 dh \\ &\leq \int_0^1 L(h\|v - \vartheta\|_2 + (1-h)\|w - \vartheta\|_2) \|v - w\|_2 dh \\ &= L\|v - w\|_2 \left[\int_0^1 (h\|v - \vartheta\|_2 + h\|w - \vartheta\|_2) dh \right] \\ &= L(\|v - \vartheta\|_2 + \|w - \vartheta\|_2) \|v - w\|_2 \left[\int_0^1 h dh \right] \\ &\leq L \max\{\|v - \vartheta\|_2, \|w - \vartheta\|_2\} \|v - w\|_2 \end{aligned} \quad (6.156)$$

(cf. Definition 1.4.7). The proof of Lemma 6.1.24 is thus complete. \square

Lemma 6.1.25. Let $\mathfrak{d} \in \mathbb{N}$, $L \in (0, \infty)$, $r \in (0, \infty]$, $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, $\mathbb{B} = \{w \in \mathbb{R}^{\mathfrak{d}} : \|w - \vartheta\|_2 \leq r\}$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ satisfy for all $v, w \in \mathbb{B}$ that

$$|\mathcal{L}(v) - \mathcal{L}(w)| \leq L \max\{\|v - \vartheta\|_2, \|w - \vartheta\|_2\} \|v - w\|_2 \quad (6.157)$$

6.1. GD optimization

(cf. Definition 3.3.4). Then it holds for all $\theta \in \mathbb{B}$ that

$$\|(\nabla \mathcal{L})(\theta)\|_2 \leq L\|\theta - \vartheta\|_2. \quad (6.158)$$

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

$$\begin{aligned} \|(\nabla \mathcal{L})(\theta)\|_2 &= \sup_{w \in \mathbb{R}^{\mathfrak{d}}, \|w\|_2=1} [\mathcal{L}'(\theta)(w)] \\ &= \sup_{w \in \mathbb{R}^{\mathfrak{d}}, \|w\|_2=1} \left[\lim_{h \searrow 0} [\frac{1}{h} (\mathcal{L}(\theta + hw) - \mathcal{L}(\theta))] \right] \\ &\leq \sup_{w \in \mathbb{R}^{\mathfrak{d}}, \|w\|_2=1} \left[\liminf_{h \searrow 0} \left[\frac{L}{h} \max\{\|\theta + hw - \vartheta\|_2, \|\theta - \vartheta\|_2\} \|\theta + hw - \theta\|_2 \right] \right] \\ &= \sup_{w \in \mathbb{R}^{\mathfrak{d}}, \|w\|_2=1} \left[\liminf_{h \searrow 0} \left[L \max\{\|\theta + hw - \vartheta\|_2, \|\theta - \vartheta\|_2\} \frac{1}{h} \|hw\|_2 \right] \right] \\ &= \sup_{w \in \mathbb{R}^{\mathfrak{d}}, \|w\|_2=1} \left[\liminf_{h \searrow 0} \left[L \max\{\|\theta + hw - \vartheta\|_2, \|\theta - \vartheta\|_2\} \right] \right] \\ &= \sup_{w \in \mathbb{R}^{\mathfrak{d}}, \|w\|_2=1} [L\|\theta - \vartheta\|_2] = L\|\theta - \vartheta\|_2. \end{aligned} \quad (6.159)$$

The fact that the function $\mathbb{R}^{\mathfrak{d}} \ni v \mapsto (\nabla \mathcal{L})(v) \in \mathbb{R}^{\mathfrak{d}}$ is continuous therefore establishes (6.158). The proof of Lemma 6.1.25 is thus complete. \square

Corollary 6.1.26. Let $\mathfrak{d} \in \mathbb{N}$, $r \in (0, \infty]$, $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, $\mathbb{B} = \{w \in \mathbb{R}^{\mathfrak{d}} : \|w - \vartheta\|_2 \leq r\}$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ (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

$$\langle \theta - \vartheta, (\nabla \mathcal{L})(\theta) \rangle \geq c\|\theta - \vartheta\|_2^2 \quad \text{and} \quad \|(\nabla \mathcal{L})(\theta)\|_2 \leq L\|\theta - \vartheta\|_2. \quad (6.160)$$

(ii) There exist $\gamma \in (0, \infty)$, $\varepsilon \in (0, 1)$ such that for all $\theta \in \mathbb{B}$ it holds that

$$\|\theta - \gamma(\nabla \mathcal{L})(\theta) - \vartheta\|_2 \leq \varepsilon\|\theta - \vartheta\|_2. \quad (6.161)$$

(iii) There exists $c \in (0, \infty)$ such that for all $\theta \in \mathbb{B}$ it holds that

$$\langle \theta - \vartheta, (\nabla \mathcal{L})(\theta) \rangle \geq c \max\{\|\theta - \vartheta\|_2^2, \|(\nabla \mathcal{L})(\theta)\|_2^2\}. \quad (6.162)$$

(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

$$\mathcal{L}(\vartheta + t(v - \vartheta)) - \mathcal{L}(\vartheta + s(v - \vartheta)) \geq c(t^2 - s^2)\|v - \vartheta\|_2^2 \quad (6.163)$$

$$\text{and} \quad |\mathcal{L}(v) - \mathcal{L}(w)| \leq L \max\{\|v - \vartheta\|_2, \|w - \vartheta\|_2\} \|v - w\|_2 \quad (6.164)$$

(cf. Definition 1.4.7).

Proof of Corollary 6.1.26. Observe that items (ii) and (iii) in Lemma 6.1.9 prove that ((i) \rightarrow (ii)). Note that Lemma 6.1.20 demonstrates that ((ii) \rightarrow (iii)). Observe that Lemma 6.1.21 establishes that ((iii) \rightarrow (i)). Note that Lemma 6.1.22 and Lemma 6.1.24 show that ((i) \rightarrow (iv)). Observe that Lemma 6.1.23 and Lemma 6.1.25 establish that ((iv) \rightarrow (i)). The proof of Corollary 6.1.26 is thus complete. \square

6.2 Explicit midpoint 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 example, Zhang et al. [454]). 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 $\mathcal{L}: \mathbb{R}^\mathfrak{d} \rightarrow \mathbb{R}$ be differentiable, let $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $\xi \in \mathbb{R}^\mathfrak{d}$, and let $\Theta: \mathbb{N}_0 \rightarrow \mathbb{R}^\mathfrak{d}$ be a function. Then we say that Θ is the explicit midpoint **GD** process for the objective function \mathcal{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

$$\Theta_0 = \xi \quad \text{and} \quad \Theta_n = \Theta_{n-1} - \gamma_n (\nabla \mathcal{L})(\Theta_{n-1}) - \frac{\gamma_n}{2} (\nabla \mathcal{L})(\Theta_{n-1}). \quad (6.165)$$

Algorithm 6.2.2: Explicit midpoint **GD** optimization method

Input: $\mathfrak{d}, N \in \mathbb{N}$, $\mathcal{L} \in C^1(\mathbb{R}^\mathfrak{d}, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $\xi \in \mathbb{R}^\mathfrak{d}$

Output: N -th step of the explicit midpoint **GD** process for the objective function \mathcal{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  $n = 1, \dots, N$  do
3:    $\Theta \leftarrow \Theta - \gamma_n (\nabla \mathcal{L})(\Theta) - \frac{\gamma_n}{2} (\nabla \mathcal{L})(\Theta)$ 
4: return  $\Theta$ 

```

6.2.1 Explicit midpoint discretizations for GF ODEs

Lemma 6.2.3 (Local error of the explicit midpoint method). *Let $\mathfrak{d} \in \mathbb{N}$, $T, \gamma, c \in [0, \infty)$, $\mathcal{G} \in C^2(\mathbb{R}^\mathfrak{d}, \mathbb{R}^\mathfrak{d})$, $\Theta \in C([0, \infty), \mathbb{R}^\mathfrak{d})$, $\theta \in \mathbb{R}^\mathfrak{d}$ satisfy for all $x, y, z \in \mathbb{R}^\mathfrak{d}$, $t \in [0, \infty)$ that*

$$\Theta_t = \Theta_0 + \int_0^t \mathcal{G}(\Theta_s) \, ds, \quad \theta = \Theta_T + \gamma \mathcal{G}(\Theta_T + \frac{\gamma}{2} \mathcal{G}(\Theta_T)), \quad (6.166)$$

$$\|\mathcal{G}(x)\|_2 \leq c, \quad \|\mathcal{G}'(x)y\|_2 \leq c\|y\|_2, \quad \text{and} \quad \|\mathcal{G}''(x)(y, z)\|_2 \leq c\|y\|_2\|z\|_2 \quad (6.167)$$

(cf. Definition 3.3.4). Then

$$\|\Theta_{T+\gamma} - \theta\|_2 \leq c^3 \gamma^3. \quad (6.168)$$

Proof of Lemma 6.2.3. Note that the fundamental theorem of calculus, the assumption that $\mathcal{G} \in C^2(\mathbb{R}^\mathfrak{d}, \mathbb{R}^\mathfrak{d})$, and (6.166) establish that for all $t \in [0, \infty)$ it holds that $\Theta \in C^1([0, \infty), \mathbb{R}^\mathfrak{d})$ and

$$\dot{\Theta}_t = \mathcal{G}(\Theta_t). \quad (6.169)$$

Combining this with the assumption that $\mathcal{G} \in C^2(\mathbb{R}^\mathfrak{d}, \mathbb{R}^\mathfrak{d})$ 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

$$\ddot{\Theta}_t = \mathcal{G}'(\Theta_t) \dot{\Theta}_t = \mathcal{G}'(\Theta_t) \mathcal{G}(\Theta_t). \quad (6.170)$$

Theorem 6.1.4 and (6.169) therefore ensure that

$$\begin{aligned} \Theta_{T+\frac{\gamma}{2}} &= \Theta_T + \left[\frac{\gamma}{2} \right] \dot{\Theta}_T + \int_0^1 (1-r) \left[\frac{\gamma}{2} \right]^2 \ddot{\Theta}_{T+r\gamma/2} \, dr \\ &= \Theta_T + \left[\frac{\gamma}{2} \right] \mathcal{G}(\Theta_T) + \frac{\gamma^2}{4} \int_0^1 (1-r) \mathcal{G}'(\Theta_{T+r\gamma/2}) \mathcal{G}(\Theta_{T+r\gamma/2}) \, dr. \end{aligned} \quad (6.171)$$

Hence, we obtain that

$$\Theta_{T+\frac{\gamma}{2}} - \Theta_T - \left[\frac{\gamma}{2} \right] \mathcal{G}(\Theta_T) = \frac{\gamma^2}{4} \int_0^1 (1-r) \mathcal{G}'(\Theta_{T+r\gamma/2}) \mathcal{G}(\Theta_{T+r\gamma/2}) \, dr. \quad (6.172)$$

Combining this, the fact that for all $x, y \in \mathbb{R}^\mathfrak{d}$ it holds that $\|\mathcal{G}(x) - \mathcal{G}(y)\|_2 \leq c\|x - y\|_2$, and (6.167) ensures that

$$\begin{aligned} \|\mathcal{G}(\Theta_{T+\frac{\gamma}{2}}) - \mathcal{G}(\Theta_T + \frac{\gamma}{2} \mathcal{G}(\Theta_T))\|_2 &\leq c \|\Theta_{T+\frac{\gamma}{2}} - \Theta_T - \frac{\gamma}{2} \mathcal{G}(\Theta_T)\|_2 \\ &\leq \frac{c\gamma^2}{4} \int_0^1 (1-r) \|\mathcal{G}'(\Theta_{T+r\gamma/2}) \mathcal{G}(\Theta_{T+r\gamma/2})\|_2 \, dr \\ &\leq \frac{c^3 \gamma^2}{4} \int_0^1 r \, dr = \frac{c^3 \gamma^2}{8}. \end{aligned} \quad (6.173)$$

Furthermore, note that (6.169), (6.170), the hypothesis that $\mathcal{G} \in C^2(\mathbb{R}^d, \mathbb{R}^d)$, the product rule, and the chain rule show that for all $t \in [0, \infty)$ it holds that $\Theta \in C^3([0, \infty), \mathbb{R}^d)$ and

$$\begin{aligned}\ddot{\Theta}_t &= \mathcal{G}''(\Theta_t)(\dot{\Theta}_t, \mathcal{G}(\Theta_t)) + \mathcal{G}'(\Theta_t)\mathcal{G}'(\Theta_t)\dot{\Theta}_t \\ &= \mathcal{G}''(\Theta_t)(\mathcal{G}(\Theta_t), \mathcal{G}(\Theta_t)) + \mathcal{G}'(\Theta_t)\mathcal{G}'(\Theta_t)\mathcal{G}(\Theta_t).\end{aligned}\quad (6.174)$$

Theorem 6.1.4, (6.169), and (6.170) therefore imply that for all $s, t \in [0, \infty)$ it holds that

$$\begin{aligned}\Theta_s &= \Theta_t + (s - t)\dot{\Theta}_t + \left[\frac{(s - t)^2}{2} \right] \ddot{\Theta}_t + \int_0^1 \left[\frac{(1 - r)^2(s - t)^3}{2} \right] \ddot{\Theta}_{t+r(s-t)} dr \\ &= \Theta_t + (s - t)\mathcal{G}(\Theta_t) + \left[\frac{(s - t)^2}{2} \right] \mathcal{G}'(\Theta_t)\mathcal{G}(\Theta_t) \\ &\quad + \frac{(s - t)^3}{2} \int_0^1 (1 - r)^2 (\mathcal{G}''(\Theta_{t+r(s-t)})(\mathcal{G}(\Theta_{t+r(s-t)}), \mathcal{G}(\Theta_{t+r(s-t)})) \\ &\quad + \mathcal{G}'(\Theta_{t+r(s-t)})\mathcal{G}'(\Theta_{t+r(s-t)})\mathcal{G}(\Theta_{t+r(s-t)})) dr.\end{aligned}\quad (6.175)$$

This establishes that

$$\begin{aligned}\Theta_{T+\gamma} - \Theta_T &= \Theta_{T+\frac{\gamma}{2}} + \left[\frac{\gamma}{2} \right] \mathcal{G}(\Theta_{T+\frac{\gamma}{2}}) + \left[\frac{\gamma^2}{8} \right] \mathcal{G}'(\Theta_{T+\frac{\gamma}{2}})\mathcal{G}(\Theta_{T+\frac{\gamma}{2}}) \\ &\quad + \frac{\gamma^3}{16} \int_0^1 (1 - r)^2 (\mathcal{G}''(\Theta_{T+(1+r)\gamma/2})(\mathcal{G}(\Theta_{T+(1+r)\gamma/2}), \mathcal{G}(\Theta_{T+(1+r)\gamma/2})) \\ &\quad + \mathcal{G}'(\Theta_{T+(1+r)\gamma/2})\mathcal{G}'(\Theta_{T+(1+r)\gamma/2})\mathcal{G}(\Theta_{T+(1+r)\gamma/2})) dr \\ &\quad - \left[\Theta_{T+\frac{\gamma}{2}} - \left[\frac{\gamma}{2} \right] \mathcal{G}(\Theta_{T+\frac{\gamma}{2}}) + \left[\frac{\gamma^2}{8} \right] \mathcal{G}'(\Theta_{T+\frac{\gamma}{2}})\mathcal{G}(\Theta_{T+\frac{\gamma}{2}}) \right. \\ &\quad \left. - \frac{\gamma^3}{16} \int_0^1 (1 - r)^2 (\mathcal{G}''(\Theta_{T+(1-r)\gamma/2})(\mathcal{G}(\Theta_{T+(1-r)\gamma/2}), \mathcal{G}(\Theta_{T+(1-r)\gamma/2})) \right. \\ &\quad \left. + \mathcal{G}'(\Theta_{T+(1-r)\gamma/2})\mathcal{G}'(\Theta_{T+(1-r)\gamma/2})\mathcal{G}(\Theta_{T+(1-r)\gamma/2})) dr \right] \\ &= \gamma \mathcal{G}(\Theta_{T+\frac{\gamma}{2}}) + \frac{\gamma^3}{16} \int_0^1 (1 - r)^2 (\mathcal{G}''(\Theta_{T+(1+r)\gamma/2})(\mathcal{G}(\Theta_{T+(1+r)\gamma/2}), \mathcal{G}(\Theta_{T+(1+r)\gamma/2})) \\ &\quad + \mathcal{G}'(\Theta_{T+(1+r)\gamma/2})\mathcal{G}'(\Theta_{T+(1+r)\gamma/2})\mathcal{G}(\Theta_{T+(1+r)\gamma/2}) \\ &\quad + \mathcal{G}''(\Theta_{T+(1-r)\gamma/2})(\mathcal{G}(\Theta_{T+(1-r)\gamma/2}), \mathcal{G}(\Theta_{T+(1-r)\gamma/2})) \\ &\quad + \mathcal{G}'(\Theta_{T+(1-r)\gamma/2})\mathcal{G}'(\Theta_{T+(1-r)\gamma/2})\mathcal{G}(\Theta_{T+(1-r)\gamma/2})) dr.\end{aligned}\quad (6.176)$$

This, (6.167), and (6.173) show that

$$\begin{aligned}
 \|\Theta_{T+\gamma} - \theta\|_2 &= \left\| \Theta_{T+\gamma} - \Theta_T - \gamma \mathcal{G}(\Theta_T + \frac{\gamma}{2} \mathcal{G}(\Theta_T)) \right\|_2 \\
 &\leq \left\| \Theta_{T+\gamma} - [\Theta_T + \gamma \mathcal{G}(\Theta_{T+\frac{\gamma}{2}})] \right\|_2 + \gamma \left\| \gamma \mathcal{G}(\Theta_{T+\frac{\gamma}{2}}) - \mathcal{G}(\Theta_T + \frac{\gamma}{2} \mathcal{G}(\Theta_T)) \right\|_2 \\
 &\leq \gamma \left\| \mathcal{G}(\Theta_{T+\frac{\gamma}{2}}) - \mathcal{G}(\Theta_T + \frac{\gamma}{2} \mathcal{G}(\Theta_T)) \right\|_2 \\
 &\quad + \frac{\gamma^3}{16} \int_0^1 (1-r)^2 \left(\left\| \mathcal{G}''(\Theta_{T+(1+r)\gamma/2})(\mathcal{G}(\Theta_{T+(1+r)\gamma/2}), \mathcal{G}(\Theta_{T+(1+r)\gamma/2})) \right\|_2 \right. \\
 &\quad + \left\| \mathcal{G}'(\Theta_{T+(1+r)\gamma/2}) \mathcal{G}'(\Theta_{T+(1+r)\gamma/2}) \mathcal{G}(\Theta_{T+(1+r)\gamma/2}) \right\|_2 \\
 &\quad + \left\| \mathcal{G}''(\Theta_{T+(1-r)\gamma/2})(\mathcal{G}(\Theta_{T+(1-r)\gamma/2}), \mathcal{G}(\Theta_{T+(1-r)\gamma/2})) \right\|_2 \\
 &\quad \left. + \left\| \mathcal{G}'(\Theta_{T+(1-r)\gamma/2}) \mathcal{G}'(\Theta_{T+(1-r)\gamma/2}) \mathcal{G}(\Theta_{T+(1-r)\gamma/2}) \right\|_2 \right) dr \\
 &\leq \frac{c^3 \gamma^3}{8} + \frac{c^3 \gamma^3}{4} \int_0^1 r^2 dr = \frac{5c^3 \gamma^3}{24} \leq c^3 \gamma^3.
 \end{aligned} \tag{6.177}$$

The proof of Lemma 6.2.3 is thus complete. \square

Corollary 6.2.4 (Local error of the explicit midpoint method for GF ODEs). *Let $\mathfrak{d} \in \mathbb{N}$, $T, \gamma, c \in [0, \infty)$, $\mathcal{L} \in C^3(\mathbb{R}^\mathfrak{d}, \mathbb{R})$, $\Theta \in C([0, \infty), \mathbb{R}^\mathfrak{d})$, $\theta \in \mathbb{R}^\mathfrak{d}$ satisfy for all $x, y, z \in \mathbb{R}^\mathfrak{d}$, $t \in [0, \infty)$ that*

$$\Theta_t = \Theta_0 - \int_0^t (\nabla \mathcal{L})(\Theta_s) ds, \quad \theta = \Theta_T - \gamma (\nabla \mathcal{L})(\Theta_T - \frac{\gamma}{2} (\nabla \mathcal{L})(\Theta_T)), \tag{6.178}$$

$$\|(\nabla \mathcal{L})(x)\|_2 \leq c, \quad \|(\text{Hess } \mathcal{L})(x)y\|_2 \leq c\|y\|_2, \quad \text{and} \quad \|(\nabla \mathcal{L})''(x)(y, z)\|_2 \leq c\|y\|_2\|z\|_2 \tag{6.179}$$

(cf. Definition 3.3.4). Then

$$\|\Theta_{T+\gamma} - \theta\|_2 \leq c^3 \gamma^3. \tag{6.180}$$

Proof of Corollary 6.2.4. Throughout this proof, let $\mathcal{G}: \mathbb{R}^\mathfrak{d} \rightarrow \mathbb{R}^\mathfrak{d}$ satisfy for all $\theta \in \mathbb{R}^\mathfrak{d}$ that

$$\mathcal{G}(\theta) = -(\nabla \mathcal{L})(\theta). \tag{6.181}$$

Observe that the fact that for all $t \in [0, \infty)$ it holds that

$$\Theta_t = \Theta_0 + \int_0^t \mathcal{G}(\Theta_s) ds, \tag{6.182}$$

the fact that

$$\theta = \Theta_T + \gamma \mathcal{G}(\Theta_T + \frac{\gamma}{2} \mathcal{G}(\Theta_T)), \tag{6.183}$$

the fact that for all $x \in \mathbb{R}^d$ it holds that $\|\mathcal{G}(x)\|_2 \leq c$, the fact that for all $x, y \in \mathbb{R}^d$ it holds that $\|\mathcal{G}'(x)y\|_2 \leq c\|y\|_2$, the fact that for all $x, y, z \in \mathbb{R}^d$ it holds that

$$\|\mathcal{G}''(x)(y, z)\|_2 \leq c\|y\|_2\|z\|_2, \quad (6.184)$$

and Lemma 6.2.3 demonstrate that

$$\|\Theta_{T+\gamma} - \theta\|_2 \leq c^3\gamma^3. \quad (6.185)$$

The proof of Corollary 6.2.4 is thus complete. \square

6.3 Momentum optimization

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 [375] and Sections 6.4, 6.5, 6.6, 6.7, 6.8, 6.9, 6.10, and 6.13 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 [357]. 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 [261, 326] and (7) in [117]. 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 $d \in \mathbb{N}$, let $\mathcal{L}: \mathbb{R}^d \rightarrow \mathbb{R}$ be differentiable, let $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $\xi \in \mathbb{R}^d$, and let $\Theta: \mathbb{N}_0 \rightarrow \mathbb{R}^d$ be a function. Then we say that Θ is the momentum GD process for the objective function \mathcal{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 (1st version) for the objective function \mathcal{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 $\mathbf{m}: \mathbb{N}_0 \rightarrow \mathbb{R}^d$ such that for all $n \in \mathbb{N}$ it holds that*

$$\Theta_0 = \xi, \quad \mathbf{m}_0 = 0, \quad (6.186)$$

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + (1 - \alpha_n)(\nabla \mathcal{L})(\Theta_{n-1}), \quad (6.187)$$

$$\text{and} \quad \Theta_n = \Theta_{n-1} - \gamma_n \mathbf{m}_n. \quad (6.188)$$

Algorithm 6.3.2: Momentum GD optimization method

Input: $\mathfrak{d}, N \in \mathbb{N}$, $\mathcal{L} \in C^1(\mathbb{R}^\mathfrak{d}, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $\xi \in \mathbb{R}^\mathfrak{d}$

Output: N -th step of the momentum **GD** process for the objective function \mathcal{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  $n = 1, \dots, N$  do
3:    $\mathbf{m} \leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n)(\nabla \mathcal{L})(\Theta)$ 
4:    $\Theta \leftarrow \Theta - \gamma_n \mathbf{m}$ 
5: return  $\Theta$ 

```

Exercise 6.3.1. Let $\mathcal{L}: \mathbb{R} \rightarrow \mathbb{R}$ satisfy for all $\theta \in \mathbb{R}$ that $\mathcal{L}(\theta) = 2\theta^2$ and let Θ be the momentum **GD** process for the objective function \mathcal{L} 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 $\mathcal{L}: \mathbb{R}^2 \rightarrow \mathbb{R}$ satisfy for all $\theta = (\theta_1, \theta_2) \in \mathbb{R}^2$ that

$$\mathcal{L}(\theta) = (\theta_1 - 3)^2 + \frac{1}{2}(\theta_2 - 2)^2 + \theta_1 + \theta_2,$$

and let Θ be the momentum **GD** process for the objective function \mathcal{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

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 differences between the methods lie in two aspects:

- Whether the momentum terms are accumulated over the gradients of the objective function or over the increments of the optimization process and
- whether the momentum terms are given as weighted averages or as general linear combinations.

The method in Definition 6.3.3 below can, for instance, be found in [118, Algorithm 2]. The method in Definition 6.3.5 below can, for example, be found in (9) in [357], (2) in [360], and (4) in [375]. Some relationships between these definitions are discussed in Section 6.3.2 below.

Definition 6.3.3 (Momentum GD optimization method (2nd version)). *Let $\mathfrak{d} \in \mathbb{N}$, let $\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$ 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 $\Theta: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a function. Then we say that Θ is the momentum GD process (2nd version) for the objective function \mathcal{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 $\mathbf{m}: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ such that for all $n \in \mathbb{N}$ it holds that*

$$\Theta_0 = \xi, \quad \mathbf{m}_0 = 0, \quad (6.189)$$

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + (\nabla \mathcal{L})(\Theta_{n-1}), \quad (6.190)$$

$$\text{and} \quad \Theta_n = \Theta_{n-1} - \gamma_n \mathbf{m}_n. \quad (6.191)$$

Algorithm 6.3.4: Momentum GD optimization method (2nd version)

Input: $\mathfrak{d}, N \in \mathbb{N}$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$
Output: N -th step of the momentum GD process (2nd version) for the objective function \mathcal{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  $n = 1, \dots, N$  do
3:    $\mathbf{m} \leftarrow \alpha_n \mathbf{m} + (\nabla \mathcal{L})(\Theta)$ 
4:    $\Theta \leftarrow \Theta - \gamma_n \mathbf{m}$ 
5: return  $\Theta$ 

```

Definition 6.3.5 (Momentum GD optimization method (3rd version)). *Let $\mathfrak{d} \in \mathbb{N}$, let $\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$ 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 $\Theta: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a function. Then we say that Θ is the momentum GD process (3rd version) for the objective function \mathcal{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 $\mathbf{m}: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ such that for all $n \in \mathbb{N}$ it holds that*

$$\Theta_0 = \xi, \quad \mathbf{m}_0 = 0, \quad (6.192)$$

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + (1 - \alpha_n) \gamma_n (\nabla \mathcal{L})(\Theta_{n-1}), \quad (6.193)$$

$$\text{and} \quad \Theta_n = \Theta_{n-1} - \mathbf{m}_n. \quad (6.194)$$

Algorithm 6.3.6: Momentum GD optimization method (3rd version)

Input: $\mathfrak{d}, N \in \mathbb{N}$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $\xi \in \mathbb{R}^{\mathfrak{d}}$

Output: N -th step of the momentum GD process (3rd version) for the objective function \mathcal{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  $n = 1, \dots, N$  do
3:    $\mathbf{m} \leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n) \gamma_n (\nabla \mathcal{L})(\Theta)$ 
4:    $\Theta \leftarrow \Theta - \mathbf{m}$ 
5: return  $\Theta$ 

```

Definition 6.3.7 (Momentum GD optimization method (4th version)). Let $\mathfrak{d} \in \mathbb{N}$, let $\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$ 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 $\Theta: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a function. Then we say that Θ is the momentum GD process (4th version) for the objective function \mathcal{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 $\mathbf{m}: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ such that for all $n \in \mathbb{N}$ it holds that

$$\Theta_0 = \xi, \quad \mathbf{m}_0 = 0, \quad (6.195)$$

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + \gamma_n (\nabla \mathcal{L})(\Theta_{n-1}), \quad (6.196)$$

$$\text{and} \quad \Theta_n = \Theta_{n-1} - \mathbf{m}_n. \quad (6.197)$$

Algorithm 6.3.8: Momentum GD optimization method (4th version)

Input: $\mathfrak{d}, N \in \mathbb{N}$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$

Output: N -th step of the momentum GD process (4th version) for the objective function \mathcal{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  $n = 1, \dots, N$  do
3:    $\mathbf{m} \leftarrow \alpha_n \mathbf{m} + \gamma_n (\nabla \mathcal{L})(\Theta)$ 
4:    $\Theta \leftarrow \Theta - \mathbf{m}$ 
5: return  $\Theta$ 

```

6.3.2 Relationships between different definitions

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 $\mathfrak{d} \in \mathbb{N}$, $(\mathbf{a}_n^{(1)})_{n \in \mathbb{N}} \subseteq (0, \infty)$, $(\mathbf{a}_n^{(2)})_{n \in \mathbb{N}} \subseteq (0, \infty)$, $(\mathbf{b}_n^{(1)})_{n \in \mathbb{N}} \subseteq (0, \infty)$, $(\mathbf{b}_n^{(2)})_{n \in \mathbb{N}} \subseteq (0, \infty)$, $(\mathbf{c}_n^{(1)})_{n \in \mathbb{N}} \subseteq (0, \infty)$, $(\mathbf{c}_n^{(2)})_{n \in \mathbb{N}} \subseteq (0, \infty)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ satisfy for all $n \in \mathbb{N}$ that*

$$\mathbf{b}_n^{(1)} \mathbf{c}_n^{(1)} = \mathbf{b}_n^{(2)} \mathbf{c}_n^{(2)} \quad \text{and} \quad \frac{\mathbf{a}_{n+1}^{(1)} \mathbf{b}_n^{(1)}}{\mathbf{b}_{n+1}^{(1)}} = \frac{\mathbf{a}_{n+1}^{(2)} \mathbf{b}_n^{(2)}}{\mathbf{b}_{n+1}^{(2)}}, \quad (6.198)$$

and for every $i \in \{1, 2\}$ let $\Theta^{(i)}: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ and $\mathbf{m}^{(i)}: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $n \in \mathbb{N}$ that

$$\Theta_0^{(i)} = \xi, \quad \mathbf{m}_0^{(i)} = 0, \quad (6.199)$$

$$\mathbf{m}_n^{(i)} = \mathbf{a}_n^{(i)} \mathbf{m}_{n-1}^{(i)} + \mathbf{b}_n^{(i)} (\nabla \mathcal{L})(\Theta_{n-1}^{(i)}), \quad (6.200)$$

$$\text{and} \quad \Theta_n^{(i)} = \Theta_{n-1}^{(i)} - \mathbf{c}_n^{(i)} \mathbf{m}_n^{(i)}. \quad (6.201)$$

Then

$$\Theta^{(1)} = \Theta^{(2)}. \quad (6.202)$$

Proof of Proposition 6.3.9. Throughout this proof, let $\mathcal{G}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that

$$\mathcal{G}(\theta) = (\nabla \mathcal{L})(\theta). \quad (6.203)$$

Note that the fact that for all $n \in \mathbb{N}$ it holds that

$$\mathbf{c}_{n+1}^{(1)} = \frac{\mathbf{c}_{n+1}^{(2)} \mathbf{b}_{n+1}^{(2)}}{\mathbf{b}_{n+1}^{(1)}}, \quad \frac{\mathbf{c}_n^{(2)}}{\mathbf{c}_n^{(1)}} = \frac{\mathbf{b}_{n+1}^{(1)}}{\mathbf{b}_{n+1}^{(2)}}, \quad \text{and} \quad \frac{\mathbf{b}_{n+1}^{(2)} \mathbf{a}_{n+1}^{(1)} \mathbf{b}_n^{(1)}}{\mathbf{b}_{n+1}^{(1)} \mathbf{b}_n^{(2)}} = \mathbf{a}_{n+1}^{(2)} \quad (6.204)$$

proves that for all $n \in \mathbb{N}$ it holds that

$$\frac{\mathbf{c}_{n+1}^{(1)} \mathbf{a}_{n+1}^{(1)} \mathbf{c}_n^{(2)}}{\mathbf{c}_n^{(1)}} = \frac{\mathbf{c}_{n+1}^{(2)} \mathbf{b}_{n+1}^{(2)} \mathbf{a}_n^{(1)} \mathbf{b}_n^{(1)}}{\mathbf{b}_{n+1}^{(1)} \mathbf{b}_n^{(2)}} = \mathbf{c}_{n+1}^{(2)} \mathbf{a}_{n+1}^{(2)}. \quad (6.205)$$

Furthermore, observe that (6.199) implies that

$$\mathbf{m}_0^{(1)} = 0 = \mathbf{m}_0^{(2)} \quad \text{and} \quad \Theta_0^{(1)} = \xi = \Theta_0^{(2)}. \quad (6.206)$$

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

$$\mathbf{c}_n^{(1)} \mathbf{m}_n^{(1)} = \mathbf{c}_n^{(2)} \mathbf{m}_n^{(2)} \quad \text{and} \quad \Theta_n^{(1)} = \Theta_n^{(2)}. \quad (6.207)$$

We now prove (6.207) by induction on $n \in \mathbb{N}$. For the base case $n = 1$ note that (6.198), (6.199), and (6.206) ensure that

$$\begin{aligned} \mathbf{c}_1^{(1)} \mathbf{m}_1^{(1)} &= \mathbf{c}_1^{(1)} (\mathbf{a}_1^{(1)} \mathbf{m}_0^{(1)} + \mathbf{b}_1^{(1)} \mathcal{G}(\Theta_0^{(1)})) \\ &= \mathbf{c}_1^{(1)} \mathbf{b}_1^{(1)} \mathcal{G}(\Theta_0^{(1)}) \\ &= \mathbf{c}_1^{(2)} \mathbf{b}_1^{(2)} \mathcal{G}(\Theta_0^{(2)}) \\ &= \mathbf{c}_1^{(2)} (\mathbf{a}_1^{(2)} \mathbf{m}_0^{(2)} + \mathbf{b}_1^{(2)} \mathcal{G}(\Theta_0^{(2)})) \\ &= \mathbf{c}_1^{(2)} \mathbf{m}_1^{(2)}. \end{aligned} \quad (6.208)$$

This, (6.201), and (6.206) establishes

$$\Theta_1^{(1)} = \Theta_0^{(1)} - \mathbf{c}_1^{(1)} \mathbf{m}_1^{(1)} = \Theta_0^{(2)} - \mathbf{c}_1^{(2)} \mathbf{m}_1^{(2)} = \Theta_1^{(2)}. \quad (6.209)$$

Combining this and (6.208) establishes (6.207) in the base case $n = 1$. For the induction step $\mathbb{N} \ni n \rightarrow n+1 \in \{2, 3, \dots\}$ let $n \in \mathbb{N}$ and assume that

$$\mathbf{c}_n^{(1)} \mathbf{m}_n^{(1)} = \mathbf{c}_n^{(2)} \mathbf{m}_n^{(2)} \quad \text{and} \quad \Theta_n^{(1)} = \Theta_n^{(2)}. \quad (6.210)$$

Observe that (6.198), (6.200), (6.205), and (6.210) show that

$$\begin{aligned} \mathbf{c}_{n+1}^{(1)} \mathbf{m}_{n+1}^{(1)} &= \mathbf{c}_{n+1}^{(1)} (\mathbf{a}_{n+1}^{(1)} \mathbf{m}_n^{(1)} + \mathbf{b}_{n+1}^{(1)} \mathcal{G}(\Theta_n^{(1)})) \\ &= \frac{\mathbf{c}_{n+1}^{(1)} \mathbf{a}_{n+1}^{(1)} \mathbf{c}_n^{(2)}}{\mathbf{c}_n^{(1)}} \mathbf{m}_n^{(2)} + \mathbf{c}_{n+1}^{(1)} \mathbf{b}_{n+1}^{(1)} \mathcal{G}(\Theta_n^{(2)}) \\ &= \mathbf{c}_{n+1}^{(2)} \mathbf{a}_{n+1}^{(2)} \mathbf{m}_n^{(2)} + \mathbf{c}_{n+1}^{(2)} \mathbf{b}_{n+1}^{(2)} \mathcal{G}(\Theta_n^{(2)}) \\ &= \mathbf{c}_{n+1}^{(2)} (\mathbf{a}_{n+1}^{(2)} \mathbf{m}_n^{(2)} + \mathbf{b}_{n+1}^{(2)} \mathcal{G}(\Theta_n^{(2)})) \\ &= \mathbf{c}_{n+1}^{(2)} \mathbf{m}_{n+1}^{(2)}. \end{aligned} \quad (6.211)$$

This, (6.201), and (6.210) demonstrate that

$$\Theta_{n+1}^{(1)} = \Theta_n^{(1)} - \mathbf{c}_{n+1}^{(1)} \mathbf{m}_{n+1}^{(1)} = \Theta_n^{(2)} - \mathbf{c}_{n+1}^{(2)} \mathbf{m}_{n+1}^{(2)} = \Theta_{n+1}^{(2)}. \quad (6.212)$$

Induction thus proves (6.207). Combining (6.206) and (6.207) establishes (6.202). The proof of Proposition 6.3.9 is thus complete. \square

Corollary 6.3.10 (Comparison of the 1st and 2nd version of the momentum GD optimization method). *Let $\mathfrak{d} \in \mathbb{N}$, $(\gamma_n^{(1)})_{n \in \mathbb{N}} \subseteq (0, \infty)$, $(\gamma_n^{(2)})_{n \in \mathbb{N}} \subseteq (0, \infty)$, $(\alpha_n^{(1)})_{n \in \mathbb{N}} \subseteq (0, 1)$, $(\alpha_n^{(2)})_{n \in \mathbb{N}} \subseteq (0, \infty)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ satisfy for all $n \in \mathbb{N}$ that*

$$\gamma_n^{(1)}(1 - \alpha_n^{(1)}) = \gamma_n^{(2)} \quad \text{and} \quad \frac{\alpha_{n+1}^{(1)}(1 - \alpha_n^{(1)})}{1 - \alpha_{n+1}^{(1)}} = \alpha_{n+1}^{(2)}, \quad (6.213)$$

for every $i \in \{1, 2\}$ let $\Theta^{(i)}: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ be the momentum GD process (i^{th} version) for the objective function \mathcal{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

$$\Theta^{(1)} = \Theta^{(2)}. \quad (6.214)$$

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

$$\mathfrak{a}_n^{(1)} = \alpha_n^{(1)}, \quad \mathfrak{b}_n^{(1)} = 1 - \alpha_n^{(1)}, \quad \mathfrak{c}_n^{(1)} = \gamma_n^{(1)}, \quad (6.215)$$

$$\mathfrak{a}_n^{(2)} = \alpha_n^{(2)}, \quad \mathfrak{b}_n^{(2)} = 1, \quad \text{and} \quad \mathfrak{c}_n^{(2)} = \gamma_n^{(2)}. \quad (6.216)$$

Note 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

$$\Theta_0^{(i)} = \xi, \quad \mathbf{m}_0^{(i)} = 0, \quad (6.217)$$

$$\mathbf{m}_n^{(i)} = \mathfrak{a}_n^{(i)} \mathbf{m}_{n-1}^{(i)} + \mathfrak{b}_n^{(i)} (\nabla \mathcal{L})(\Theta_{n-1}^{(i)}), \quad (6.218)$$

$$\text{and} \quad \Theta_n^{(i)} = \Theta_{n-1}^{(i)} - \mathfrak{c}_n^{(i)} \mathbf{m}_n^{(i)}. \quad (6.219)$$

Furthermore, observe that (6.213), (6.215), and (6.216) implies that for all $n \in \mathbb{N}$ it holds that

$$\mathfrak{b}_n^{(1)} \mathfrak{c}_n^{(1)} = (1 - \alpha_n^{(1)}) \gamma_n^{(1)} = \gamma_n^{(2)} = \mathfrak{b}_n^{(2)} \mathfrak{c}_n^{(2)}. \quad (6.220)$$

Moreover, note that (6.213), (6.215), and (6.216) ensures that for all $n \in \mathbb{N}$ it holds that

$$\frac{\mathfrak{a}_{n+1}^{(1)} \mathfrak{b}_n^{(1)}}{\mathfrak{b}_{n+1}^{(1)}} = \frac{\alpha_{n+1}^{(1)}(1 - \alpha_n^{(1)})}{1 - \alpha_{n+1}^{(1)}} = \alpha_{n+1}^{(2)} = \frac{\mathfrak{a}_{n+1}^{(2)} \mathfrak{b}_n^{(2)}}{\mathfrak{b}_{n+1}^{(2)}}. \quad (6.221)$$

Combining this, (6.217), (6.218), (6.219), and (6.220) with Proposition 6.3.9 establishes (6.214). The proof of Corollary 6.3.10 is thus complete. \square

Lemma 6.3.11 (Comparison of the 1st and 3rd version of the momentum GD optimization method). Let $\mathfrak{d} \in \mathbb{N}$, $(\gamma_n^{(1)})_{n \in \mathbb{N}} \subseteq (0, \infty)$, $(\gamma_n^{(3)})_{n \in \mathbb{N}} \subseteq (0, \infty)$, $(\alpha_n^{(1)})_{n \in \mathbb{N}} \subseteq (0, 1)$, $(\alpha_n^{(3)})_{n \in \mathbb{N}} \subseteq (0, 1)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ satisfy for all $n \in \mathbb{N}$ that

$$\gamma_n^{(1)}(1 - \alpha_n^{(1)}) = \gamma_n^{(3)}(1 - \alpha_n^{(3)}) \quad \text{and} \quad \frac{\gamma_{n+1}^{(1)} \alpha_{n+1}^{(1)}}{\gamma_n^{(1)}} = \alpha_{n+1}^{(3)}, \quad (6.222)$$

for every $i \in \{1, 3\}$ let $\Theta^{(i)} : \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ be the momentum GD process (i^{th} version) for the objective function \mathcal{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

$$\Theta^{(1)} = \Theta^{(3)}. \quad (6.223)$$

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

$$\mathfrak{a}_n^{(1)} = \alpha_n^{(1)}, \quad \mathfrak{b}_n^{(1)} = 1 - \alpha_n^{(1)}, \quad \mathfrak{c}_n^{(1)} = \gamma_n^{(1)}, \quad (6.224)$$

$$\mathfrak{a}_n^{(3)} = \alpha_n^{(3)}, \quad \mathfrak{b}_n^{(3)} = (1 - \alpha_n^{(3)})\gamma_n^{(3)}, \quad \text{and} \quad \mathfrak{c}_n^{(3)} = 1. \quad (6.225)$$

Observe that (6.186), (6.187), (6.188), (6.192), (6.193), and (6.194) show that for all $i \in \{1, 3\}$, $n \in \mathbb{N}$ it holds that

$$\Theta_0^{(i)} = \xi, \quad \mathbf{m}_0^{(i)} = 0, \quad (6.226)$$

$$\mathbf{m}_n^{(i)} = \mathfrak{a}_n^{(i)} \mathbf{m}_{n-1}^{(i)} + \mathfrak{b}_n^{(i)} (\nabla \mathcal{L})(\Theta_{n-1}^{(i)}), \quad (6.227)$$

$$\text{and} \quad \Theta_n^{(i)} = \Theta_{n-1}^{(i)} - \mathfrak{c}_n^{(i)} \mathbf{m}_n^{(i)}. \quad (6.228)$$

Furthermore, note that (6.222), (6.224), and (6.225) demonstrates that for all $n \in \mathbb{N}$ it holds that

$$\mathfrak{b}_n^{(1)} \mathfrak{c}_n^{(1)} = (1 - \alpha_n^{(1)})\gamma_n^{(1)} = (1 - \alpha_n^{(3)})\gamma_n^{(3)} = \mathfrak{b}_n^{(3)} \mathfrak{c}_n^{(3)}. \quad (6.229)$$

Moreover, observe that (6.222), (6.224), and (6.225) proves that for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned} \frac{\mathfrak{a}_{n+1}^{(1)} \mathfrak{b}_n^{(1)}}{\mathfrak{b}_{n+1}^{(1)}} &= \frac{\alpha_{n+1}^{(1)}(1 - \alpha_n^{(1)})}{1 - \alpha_{n+1}^{(1)}} = \frac{\alpha_{n+1}^{(1)} \gamma_n^{(3)}(1 - \alpha_n^{(3)}) \gamma_{n+1}^{(1)}}{\gamma_n^{(1)} \gamma_{n+1}^{(3)}(1 - \alpha_{n+1}^{(3)})} \\ &= \frac{\alpha_{n+1}^{(3)} \gamma_n^{(3)}(1 - \alpha_n^{(3)})}{\gamma_{n+1}^{(3)}(1 - \alpha_{n+1}^{(3)})} = \frac{\mathfrak{a}_{n+1}^{(3)} \mathfrak{b}_n^{(3)}}{\mathfrak{b}_{n+1}^{(3)}}. \end{aligned} \quad (6.230)$$

Combining this, (6.222), (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. \square

Lemma 6.3.12 (Comparison of the 1st and 4th version of the momentum GD optimization method). Let $\mathfrak{d} \in \mathbb{N}$, $(\gamma_n^{(1)})_{n \in \mathbb{N}} \subseteq (0, \infty)$, $(\gamma_n^{(4)})_{n \in \mathbb{N}} \subseteq (0, \infty)$, $(\alpha_n^{(1)})_{n \in \mathbb{N}} \subseteq (0, 1)$, $(\alpha_n^{(4)})_{n \in \mathbb{N}} \subseteq (0, \infty)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ satisfy for all $n \in \mathbb{N}$ that

$$\gamma_n^{(1)}(1 - \alpha_n^{(1)}) = \gamma_n^{(4)} \quad \text{and} \quad \frac{\gamma_{n+1}^{(1)} \alpha_{n+1}^{(1)}}{\gamma_n^{(1)}} = \alpha_{n+1}^{(4)}, \quad (6.231)$$

for every $i \in \{1, 4\}$ let $\Theta^{(i)}: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ be the momentum GD process (i^{th} version) for the objective function \mathcal{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

$$\Theta^{(1)} = \Theta^{(4)}. \quad (6.232)$$

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

$$\mathfrak{a}_n^{(1)} = \alpha_n^{(1)}, \quad \mathfrak{b}_n^{(1)} = 1 - \alpha_n^{(1)}, \quad \mathfrak{c}_n^{(1)} = \gamma_n^{(1)}, \quad (6.233)$$

$$\mathfrak{a}_n^{(4)} = \alpha_n^{(4)}, \quad \mathfrak{b}_n^{(4)} = \gamma_n^{(4)}, \quad \text{and} \quad \mathfrak{c}_n^{(4)} = 1. \quad (6.234)$$

Note 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

$$\Theta_0^{(i)} = \xi, \quad \mathbf{m}_0^{(i)} = 0, \quad (6.235)$$

$$\mathbf{m}_n^{(i)} = \mathfrak{a}_n^{(i)} \mathbf{m}_{n-1}^{(i)} + \mathfrak{b}_n^{(i)} (\nabla \mathcal{L})(\Theta_{n-1}^{(i)}), \quad (6.236)$$

$$\text{and} \quad \Theta_n^{(i)} = \Theta_{n-1}^{(i)} - \mathfrak{c}_n^{(i)} \mathbf{m}_n^{(i)}. \quad (6.237)$$

Furthermore, observe that (6.231), (6.233), and (6.234) establishes that for all $n \in \mathbb{N}$ it holds that

$$\mathfrak{b}_n^{(1)} \mathfrak{c}_n^{(1)} = (1 - \alpha_n^{(1)}) \gamma_n^{(1)} = \gamma_n^{(4)} = \mathfrak{b}_n^{(4)} \mathfrak{c}_n^{(4)}. \quad (6.238)$$

Moreover, note that (6.231), (6.233), and (6.234) shows that for all $n \in \mathbb{N}$ it holds that

$$\frac{\mathfrak{a}_{n+1}^{(1)} \mathfrak{b}_n^{(1)}}{\mathfrak{b}_{n+1}^{(1)}} = \frac{\alpha_{n+1}^{(1)} (1 - \alpha_n^{(1)})}{1 - \alpha_{n+1}^{(1)}} = \frac{\alpha_{n+1}^{(1)} \gamma_n^{(4)} \gamma_{n+1}^{(1)}}{\gamma_n^{(1)} \gamma_{n+1}^{(4)}} = \frac{\alpha_{n+1}^{(4)} \gamma_n^{(4)}}{\gamma_{n+1}^{(4)}} = \frac{\mathfrak{a}_{n+1}^{(4)} \mathfrak{b}_n^{(4)}}{\mathfrak{b}_{n+1}^{(4)}}. \quad (6.239)$$

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. \square

Corollary 6.3.13 (Comparison of the 2nd and 3rd version of the momentum SGD optimization method). Let $\mathfrak{d} \in \mathbb{N}$, $(\gamma_n^{(2)})_{n \in \mathbb{N}} \subseteq (0, \infty)$, $(\gamma_n^{(3)})_{n \in \mathbb{N}} \subseteq (0, \infty)$, $(\alpha_n^{(2)})_{n \in \mathbb{N}} \subseteq (0, \infty)$, $(\alpha_n^{(3)})_{n \in \mathbb{N}} \subseteq (0, 1)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ satisfy for all $n \in \mathbb{N}$ that

$$\gamma_n^{(2)} = \gamma_n^{(3)}(1 - \alpha_n^{(3)}) \quad \text{and} \quad \frac{\gamma_{n+1}^{(2)} \alpha_{n+1}^{(2)}}{\gamma_n^{(2)}} = \alpha_{n+1}^{(3)}, \quad (6.240)$$

for every $i \in \{2, 3\}$ let $\Theta^{(i)} : \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ be the momentum GD process (i^{th} version) for the objective function \mathcal{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

$$\Theta^{(2)} = \Theta^{(3)}. \quad (6.241)$$

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

$$\mathfrak{a}_n^{(2)} = \alpha_n^{(2)}, \quad \mathfrak{b}_n^{(2)} = 1, \quad \mathfrak{c}_n^{(2)} = \gamma_n^{(2)}, \quad (6.242)$$

$$\mathfrak{a}_n^{(3)} = \alpha_n^{(3)}, \quad \mathfrak{b}_n^{(3)} = (1 - \alpha_n^{(3)})\gamma_n^{(3)}, \quad \text{and} \quad \mathfrak{c}_n^{(3)} = 1. \quad (6.243)$$

Observe 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

$$\Theta_0^{(i)} = \xi, \quad \mathbf{m}_0^{(i)} = 0, \quad (6.244)$$

$$\mathbf{m}_n^{(i)} = \mathfrak{a}_n^{(i)} \mathbf{m}_{n-1}^{(i)} + \mathfrak{b}_n^{(i)} (\nabla \mathcal{L})(\Theta_{n-1}^{(i)}), \quad (6.245)$$

$$\text{and} \quad \Theta_n^{(i)} = \Theta_{n-1}^{(i)} - \mathfrak{c}_n^{(i)} \mathbf{m}_n^{(i)}. \quad (6.246)$$

Furthermore, note that (6.240), (6.242), and (6.243) implies that for all $n \in \mathbb{N}$ it holds that

$$\mathfrak{b}_n^{(2)} \mathfrak{c}_n^{(2)} = \gamma_n^{(2)} = \gamma_n^{(3)}(1 - \alpha_n^{(3)}) = \mathfrak{b}_n^{(3)} \mathfrak{c}_n^{(3)}. \quad (6.247)$$

Moreover, observe that (6.240), (6.242), and (6.243) ensures that for all $n \in \mathbb{N}$ it holds that

$$\frac{\mathfrak{a}_{n+1}^{(2)} \mathfrak{b}_n^{(2)}}{\mathfrak{b}_{n+1}^{(2)}} = \alpha_{n+1}^{(2)} = \frac{\alpha_{n+1}^{(3)} \gamma_n^{(3)}(1 - \alpha_n^{(3)})}{\gamma_{n+1}^{(3)}(1 - \alpha_{n+1}^{(3)})} = \frac{\mathfrak{a}_{n+1}^{(3)} \mathfrak{b}_n^{(3)}}{\mathfrak{b}_{n+1}^{(3)}}. \quad (6.248)$$

Combining this, (6.244), (6.245), (6.246), and (6.247) with Proposition 6.3.9 establishes (6.241). The proof of Corollary 6.3.13 is thus complete. \square

Lemma 6.3.14 (Comparison of the 2nd and 4th version of the momentum GD optimization method). Let $\mathfrak{d} \in \mathbb{N}$, $(\gamma_n^{(2)})_{n \in \mathbb{N}} \subseteq (0, \infty)$, $(\gamma_n^{(4)})_{n \in \mathbb{N}} \subseteq (0, \infty)$, $(\alpha_n^{(2)})_{n \in \mathbb{N}} \subseteq (0, 1)$, $(\alpha_n^{(4)})_{n \in \mathbb{N}} \subseteq (0, 1)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ satisfy for all $n \in \mathbb{N}$ that

$$\gamma_n^{(2)} = \gamma_n^{(4)} \quad \text{and} \quad \frac{\gamma_{n+1}^{(2)} \alpha_{n+1}^{(2)}}{\gamma_n^{(2)}} = \alpha_{n+1}^{(4)}, \quad (6.249)$$

for every $i \in \{2, 4\}$ let $\Theta^{(i)}: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ be the momentum GD process (i^{th} version) for the objective function \mathcal{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

$$\Theta^{(2)} = \Theta^{(4)}. \quad (6.250)$$

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

$$\mathfrak{a}_n^{(2)} = \alpha_n^{(2)}, \quad \mathfrak{b}_n^{(2)} = 1, \quad \mathfrak{c}_n^{(2)} = \gamma_n^{(2)}, \quad (6.251)$$

$$\mathfrak{a}_n^{(4)} = \alpha_n^{(4)}, \quad \mathfrak{b}_n^{(4)} = \gamma_n^{(4)}, \quad \text{and} \quad \mathfrak{c}_n^{(4)} = 1. \quad (6.252)$$

Note that (6.189), (6.190), (6.191), (6.195), (6.196), and (6.197) show that for all $i \in \{2, 4\}$, $n \in \mathbb{N}$ it holds that

$$\Theta_0^{(i)} = \xi, \quad \mathbf{m}_0^{(i)} = 0, \quad (6.253)$$

$$\mathbf{m}_n^{(i)} = \mathfrak{a}_n^{(i)} \mathbf{m}_{n-1}^{(i)} + \mathfrak{b}_n^{(i)} (\nabla \mathcal{L})(\Theta_{n-1}^{(i)}), \quad (6.254)$$

$$\text{and} \quad \Theta_n^{(i)} = \Theta_{n-1}^{(i)} - \mathfrak{c}_n^{(i)} \mathbf{m}_n^{(i)}. \quad (6.255)$$

Furthermore, observe that (6.249), (6.251), and (6.252) demonstrates that for all $n \in \mathbb{N}$ it holds that

$$\mathfrak{b}_n^{(2)} \mathfrak{c}_n^{(2)} = \gamma_n^{(2)} = \gamma_n^{(4)} = \mathfrak{b}_n^{(4)} \mathfrak{c}_n^{(4)}. \quad (6.256)$$

Moreover, note that (6.249), (6.251), and (6.252) proves that for all $n \in \mathbb{N}$ it holds that

$$\frac{\mathfrak{a}_{n+1}^{(2)} \mathfrak{b}_n^{(2)}}{\mathfrak{b}_{n+1}^{(2)}} = \alpha_{n+1}^{(2)} = \frac{\alpha_{n+1}^{(4)} \gamma_n^{(4)}}{\gamma_{n+1}^{(4)}} = \frac{\mathfrak{a}_{n+1}^{(4)} \mathfrak{b}_n^{(4)}}{\mathfrak{b}_{n+1}^{(4)}}. \quad (6.257)$$

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. \square

Corollary 6.3.15 (Comparison of the 3rd and 4th version of the momentum **GD** optimization method). Let $\mathfrak{d} \in \mathbb{N}$, $(\gamma_n^{(3)})_{n \in \mathbb{N}} \subseteq (0, \infty)$, $(\gamma_n^{(4)})_{n \in \mathbb{N}} \subseteq (0, \infty)$, $(\alpha_n^{(3)})_{n \in \mathbb{N}} \subseteq (0, 1)$, $(\alpha_n^{(4)})_{n \in \mathbb{N}} \subseteq (0, \infty)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ satisfy for all $n \in \mathbb{N}$ that

$$\gamma_n^{(3)}(1 - \alpha_n^{(3)}) = \gamma_n^{(4)} \quad \text{and} \quad \alpha_{n+1}^{(3)} = \alpha_{n+1}^{(4)}, \quad (6.258)$$

for every $i \in \{3, 4\}$ let $\Theta^{(i)}: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ be the momentum **GD** process (i^{th} version) for the objective function \mathcal{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

$$\Theta^{(3)} = \Theta^{(4)}. \quad (6.259)$$

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

$$\mathfrak{a}_n^{(3)} = \alpha_n^{(3)}, \quad \mathfrak{b}_n^{(3)} = (1 - \alpha_n^{(3)})\gamma_n^{(3)}, \quad \mathfrak{c}_n^{(3)} = 1 \quad (6.260)$$

$$\mathfrak{a}_n^{(4)} = \alpha_n^{(4)}, \quad \mathfrak{b}_n^{(4)} = \gamma_n^{(4)}, \quad \text{and} \quad \mathfrak{c}_n^{(4)} = 1, \quad (6.261)$$

Observe 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

$$\Theta_0^{(i)} = \xi, \quad \mathbf{m}_0^{(i)} = 0, \quad (6.262)$$

$$\mathbf{m}_n^{(i)} = \mathfrak{a}_n^{(i)}\mathbf{m}_{n-1}^{(i)} + \mathfrak{b}_n^{(i)}(\nabla \mathcal{L})(\Theta_{n-1}^{(i)}), \quad (6.263)$$

$$\text{and} \quad \Theta_n^{(i)} = \Theta_{n-1}^{(i)} - \mathfrak{c}_n^{(i)}\mathbf{m}_n^{(i)}. \quad (6.264)$$

Furthermore, note that (6.258), (6.260), and (6.261) establishes that for all $n \in \mathbb{N}$ it holds that

$$\mathfrak{b}_n^{(3)}\mathfrak{c}_n^{(3)} = \gamma_n^{(3)}(1 - \alpha_n^{(3)}) = \gamma_n^{(4)} = \mathfrak{b}_n^{(4)}\mathfrak{c}_n^{(4)}. \quad (6.265)$$

Moreover, observe that (6.258), (6.260), and (6.261) shows that for all $n \in \mathbb{N}$ it holds that

$$\frac{\mathfrak{a}_{n+1}^{(3)}\mathfrak{b}_n^{(3)}}{\mathfrak{b}_{n+1}^{(3)}} = \frac{\alpha_{n+1}^{(3)}(1 - \alpha_n^{(3)})\gamma_n^{(3)}}{(1 - \alpha_{n+1}^{(3)})\gamma_{n+1}^{(3)}} = \frac{\alpha_{n+1}^{(4)}\gamma_n^{(4)}}{\gamma_{n+1}^{(4)}} = \frac{\mathfrak{a}_{n+1}^{(4)}\mathfrak{b}_n^{(4)}}{\mathfrak{b}_{n+1}^{(4)}}. \quad (6.266)$$

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. \square

6.3.3 Representations for momentum optimization

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*

$$m_n = \alpha_n m_{n-1} + 1 - \alpha_n. \quad (6.267)$$

Then it holds for all $n \in \mathbb{N}_0$ that

$$m_n = 1 - \prod_{k=1}^n \alpha_k. \quad (6.268)$$

Proof of Lemma 6.3.16. We prove (6.268) by induction on $n \in \mathbb{N}_0$. For the base case $n = 0$ note that the assumption that $m_0 = 0$ proves that

$$m_0 = 0 = 1 - \prod_{k=1}^0 \alpha_k. \quad (6.269)$$

This establishes (6.268) in the base case $n = 0$. For the induction step note that (6.267) establishes that for all $n \in \mathbb{N}_0$ with $m_n = 1 - \prod_{k=1}^n \alpha_k$ it holds that

$$\begin{aligned} m_{n+1} &= \alpha_{n+1} m_n + 1 - \alpha_{n+1} = \alpha_{n+1} \left[1 - \prod_{k=1}^n \alpha_k \right] + 1 - \alpha_{n+1} \\ &= \alpha_{n+1} - \prod_{k=1}^{n+1} \alpha_k + 1 - \alpha_{n+1} = 1 - \prod_{k=1}^{n+1} \alpha_k. \end{aligned} \quad (6.270)$$

Induction hence establishes (6.268). The proof of Lemma 6.3.16 is thus complete. \square

Lemma 6.3.17 (An explicit representation of momentum terms). *Let $\mathfrak{d} \in \mathbb{N}$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq \mathbb{R}$, $(a_{n,k})_{(n,k) \in (\mathbb{N}_0)^2} \subseteq \mathbb{R}$, $(\mathcal{G}_n)_{n \in \mathbb{N}_0} \subseteq \mathbb{R}^\mathfrak{d}$, $(\mathbf{m}_n)_{n \in \mathbb{N}_0} \subseteq \mathbb{R}^\mathfrak{d}$ satisfy for all $n \in \mathbb{N}$, $k \in$*

$\{0, 1, \dots, n-1\}$ that

$$\mathbf{m}_0 = 0, \quad \mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + (1 - \alpha_n) \mathcal{G}_{n-1}, \quad \text{and} \quad a_{n,k} = (1 - \alpha_{k+1}) \left[\prod_{l=k+2}^n \alpha_l \right] \quad (6.271)$$

Then

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

$$\mathbf{m}_n = \sum_{k=0}^{n-1} a_{n,k} \mathcal{G}_k \quad (6.272)$$

and

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

$$\sum_{k=0}^{n-1} a_{n,k} = 1 - \prod_{k=1}^n \alpha_k. \quad (6.273)$$

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

$$m_n = \sum_{k=0}^{n-1} a_{n,k}. \quad (6.274)$$

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

$$\mathbf{m}_0 = 0 = \sum_{k=0}^{-1} a_{0,k} \mathcal{G}_k. \quad (6.275)$$

This establishes item (i) in the base case $n = 0$. For the induction step observe that (6.271)

shows that for all $n \in \mathbb{N}_0$ with $\mathbf{m}_n = \sum_{k=0}^{n-1} a_{n,k} \mathcal{G}_k$ it holds that

$$\begin{aligned}
 \mathbf{m}_{n+1} &= \alpha_{n+1} \mathbf{m}_n + (1 - \alpha_{n+1}) \mathcal{G}_n \\
 &= \left[\sum_{k=0}^{n-1} \alpha_{n+1} a_{n,k} \mathcal{G}_k \right] + (1 - \alpha_{n+1}) \mathcal{G}_n \\
 &= \left[\sum_{k=0}^{n-1} \alpha_{n+1} (1 - \alpha_{k+1}) \left[\prod_{l=k+2}^n \alpha_l \right] \mathcal{G}_k \right] + (1 - \alpha_{n+1}) \mathcal{G}_n \\
 &= \left[\sum_{k=0}^{n-1} (1 - \alpha_{k+1}) \left[\prod_{l=k+2}^{n+1} \alpha_l \right] \mathcal{G}_k \right] + (1 - \alpha_{n+1}) \mathcal{G}_n \\
 &= \sum_{k=0}^n (1 - \alpha_{k+1}) \left[\prod_{l=k+2}^{n+1} \alpha_l \right] \mathcal{G}_k = \sum_{k=0}^n a_{n+1,k} \mathcal{G}_k.
 \end{aligned} \tag{6.276}$$

Induction thus proves item (i). Furthermore, note 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. \square

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 $n \in \mathbb{N}$, $k \in \{0, 1, \dots, n-1\}$ that*

$$a_{n,k} = (1 - \alpha_{k+1}) \left[\prod_{l=k+2}^n \alpha_l \right], \tag{6.279}$$

*let $\mathcal{L} \in C^1(\mathbb{R}^\mathfrak{d}, \mathbb{R})$, and let Θ be the momentum **GD** process for the objective function \mathcal{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 $n \in \mathbb{N}$, $k \in \{0, 1, \dots, n-1\}$ that $0 \leq a_{n,k} \leq 1$,

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

$$\sum_{k=0}^{n-1} a_{n,k} = 1 - \prod_{k=1}^n \alpha_k, \quad (6.280)$$

and

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

$$\Theta_n = \Theta_{n-1} - \gamma_n \left[\sum_{k=0}^{n-1} a_{n,k} (\nabla \mathcal{L})(\Theta_k) \right]. \quad (6.281)$$

Proof of Corollary 6.3.18. Throughout this proof, let $\mathbf{m}: \mathbb{N}_0 \rightarrow \mathbb{R}^\mathfrak{d}$ satisfy for all $n \in \mathbb{N}$ that

$$\mathbf{m}_0 = 0 \quad \text{and} \quad \mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + (1 - \alpha_n) (\nabla \mathcal{L})(\Theta_{n-1}). \quad (6.282)$$

Observe that (6.279) establishes item (i). Note that (6.279), (6.282), and Lemma 6.3.17 establish that for all $n \in \mathbb{N}_0$ it holds that

$$\mathbf{m}_n = \sum_{k=0}^{n-1} a_{n,k} (\nabla \mathcal{L})(\Theta_k) \quad \text{and} \quad \sum_{k=0}^{n-1} a_{n,k} = 1 - \prod_{k=1}^n \alpha_k. \quad (6.283)$$

This proves item (ii). Observe that (6.186), (6.187), (6.188), (6.282), and (6.283) ensure that for all $n \in \mathbb{N}$ it holds that

$$\Theta_n = \Theta_{n-1} - \gamma_n \mathbf{m}_n = \Theta_{n-1} - \gamma_n \left[\sum_{k=0}^{n-1} a_{n,k} (\nabla \mathcal{L})(\Theta_k) \right]. \quad (6.284)$$

This establishes item (iii). The proof of Corollary 6.3.18 is thus complete. \square

6.3.4 Bias-adjusted momentum optimization

Definition 6.3.19 (Bias-adjusted momentum GD optimization method). Let $\mathfrak{d} \in \mathbb{N}$, let $\mathcal{L}: \mathbb{R}^\mathfrak{d} \rightarrow \mathbb{R}$ 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 $\Theta: \mathbb{N}_0 \rightarrow \mathbb{R}^\mathfrak{d}$ be a function. Then we say that Θ is the bias-adjusted momentum GD process for the objective function \mathcal{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 $\mathbf{m}: \mathbb{N}_0 \rightarrow \mathbb{R}^\mathfrak{d}$ such that for all $n \in \mathbb{N}$ it holds that

$$\Theta_0 = \xi, \quad \mathbf{m}_0 = 0, \quad (6.285)$$

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + (1 - \alpha_n) (\nabla \mathcal{L})(\Theta_{n-1}), \quad (6.286)$$

$$\text{and} \quad \Theta_n = \Theta_{n-1} - \frac{\gamma_n \mathbf{m}_n}{1 - \prod_{l=1}^n \alpha_l}. \quad (6.287)$$

Algorithm 6.3.20: Bias-adjusted momentum GD optimization method

Input: $\mathfrak{d}, N \in \mathbb{N}$, $\mathcal{L} \in C^1(\mathbb{R}^\mathfrak{d}, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, 1)$, $\xi \in \mathbb{R}^\mathfrak{d}$
Output: N -th step of the bias-adjusted momentum GD process for the objective function \mathcal{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)

```

1: Initialization:  $\Theta \leftarrow \xi; \mathbf{m} \leftarrow 0 \in \mathbb{R}^\mathfrak{d}$ 
2: for  $n = 1, \dots, N$  do
3:    $\mathbf{m} \leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n)(\nabla \mathcal{L})(\Theta)$ 
4:    $\Theta \leftarrow \Theta - \frac{\gamma_n \mathbf{m}}{1 - \prod_{l=1}^n \alpha_l}$ 
5: return  $\Theta$ 
```

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 $n \in \mathbb{N}$, $k \in \{0, 1, \dots, n-1\}$ that

$$a_{n,k} = \frac{(1 - \alpha_{k+1}) [\prod_{l=k+2}^n \alpha_l]}{1 - \prod_{l=1}^n \alpha_l}, \quad (6.288)$$

let $\mathcal{L} \in C^1(\mathbb{R}^\mathfrak{d}, \mathbb{R})$, and let Θ be the bias-adjusted momentum GD process for the objective function \mathcal{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 $n \in \mathbb{N}$, $k \in \{0, 1, \dots, n-1\}$ that $0 \leq a_{n,k} \leq 1$,

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

$$\sum_{k=0}^{n-1} a_{n,k} = 1, \quad (6.289)$$

and

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

$$\Theta_n = \Theta_{n-1} - \gamma_n \left[\sum_{k=0}^{n-1} a_{n,k} (\nabla \mathcal{L})(\Theta_k) \right]. \quad (6.290)$$

Proof of Corollary 6.3.21. Throughout this proof, let $\mathbf{m}: \mathbb{N}_0 \rightarrow \mathbb{R}^d$ satisfy for all $n \in \mathbb{N}$ that

$$\mathbf{m}_0 = 0 \quad \text{and} \quad \mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + (1 - \alpha_n) (\nabla \mathcal{L})(\Theta_{n-1}) \quad (6.291)$$

and let $(b_{n,k})_{(n,k) \in (\mathbb{N}_0)^2} \subseteq \mathbb{R}$ satisfy for all $n \in \mathbb{N}$, $k \in \{0, 1, \dots, n-1\}$ that

$$b_{n,k} = (1 - \alpha_{k+1}) \left[\prod_{l=k+2}^n \alpha_l \right]. \quad (6.292)$$

Note that (6.288) implies item (i). Observe that (6.288), (6.291), (6.292), and Lemma 6.3.17 show that for all $n \in \mathbb{N}$ it holds that

$$\mathbf{m}_n = \sum_{k=0}^{n-1} b_{n,k} (\nabla \mathcal{L})(\Theta_k) \quad \text{and} \quad \sum_{k=0}^{n-1} a_{n,k} = \frac{\sum_{k=0}^{n-1} b_{n,k}}{1 - \prod_{k=1}^n \alpha_k} = \frac{1 - \prod_{k=1}^n \alpha_k}{1 - \prod_{k=1}^n \alpha_k} = 1. \quad (6.293)$$

This proves item (ii). Note 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

$$\begin{aligned} \Theta_n &= \Theta_{n-1} - \frac{\gamma_n \mathbf{m}_n}{1 - \prod_{l=1}^n \alpha_l} = \Theta_{n-1} - \gamma_n \left[\sum_{k=0}^{n-1} \left[\frac{b_{n,k}}{1 - \prod_{l=1}^n \alpha_l} \right] (\nabla \mathcal{L})(\Theta_k) \right] \\ &= \Theta_{n-1} - \gamma_n \left[\sum_{k=0}^{n-1} a_{n,k} (\nabla \mathcal{L})(\Theta_k) \right]. \end{aligned} \quad (6.294)$$

This establishes item (iii). The proof of Corollary 6.3.21 is thus complete. \square

6.3.5 Error analysis for momentum optimization

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 speaking, that the momentum **GD** optimization method outperforms 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 instance, be found in [50, Section 7.1] and [357].

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 $\mathfrak{d} \in \mathbb{N}$, $\xi \in \mathbb{R}^{\mathfrak{d}}$, $\vartheta = (\vartheta_1, \dots, \vartheta_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$, $\kappa, \mathcal{K}, \lambda_1, \lambda_2, \dots, \lambda_{\mathfrak{d}} \in (0, \infty)$ satisfy $\kappa = \min\{\lambda_1, \lambda_2, \dots, \lambda_{\mathfrak{d}}\}$ and $\mathcal{K} = \max\{\lambda_1, \lambda_2, \dots, \lambda_{\mathfrak{d}}\}$, let $\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$ satisfy for all $\theta = (\theta_1, \dots, \theta_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$ that*

$$\mathcal{L}(\theta) = \frac{1}{2} \left[\sum_{i=1}^{\mathfrak{d}} \lambda_i |\theta_i - \vartheta_i|^2 \right], \quad (6.295)$$

and let $\Theta: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $n \in \mathbb{N}$ that

$$\Theta_0 = \xi \quad \text{and} \quad \Theta_n = \Theta_{n-1} - \frac{2}{\mathcal{K}+\kappa} (\nabla \mathcal{L})(\Theta_{n-1}). \quad (6.296)$$

Then it holds for all $n \in \mathbb{N}_0$ that

$$\|\Theta_n - \vartheta\|_2 \leq \left[\frac{\mathcal{K}-\kappa}{\mathcal{K}+\kappa} \right]^n \|\xi - \vartheta\|_2 \quad (6.297)$$

(cf. Definition 3.3.4).

Proof of Lemma 6.3.22. Throughout this proof, let $\Theta^{(i)}: \mathbb{N}_0 \rightarrow \mathbb{R}$, $i \in \{1, 2, \dots, \mathfrak{d}\}$, satisfy for all $n \in \mathbb{N}_0$ that $\Theta_n = (\Theta_n^{(1)}, \Theta_n^{(2)}, \dots, \Theta_n^{(\mathfrak{d})})$. Note that (6.295) implies that for all $\theta = (\theta_1, \dots, \theta_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$, $i \in \{1, 2, \dots, \mathfrak{d}\}$ it holds that

$$\left(\frac{\partial \mathcal{L}}{\partial \theta_i} \right)(\theta) = \lambda_i (\theta_i - \vartheta_i). \quad (6.298)$$

Combining this and (6.296) ensures that for all $n \in \mathbb{N}$, $i \in \{1, 2, \dots, \mathfrak{d}\}$ it holds that

$$\begin{aligned} \Theta_n^{(i)} - \vartheta_i &= \Theta_{n-1}^{(i)} - \frac{2}{\mathcal{K}+\kappa} \left(\frac{\partial \mathcal{L}}{\partial \theta_i} \right)(\Theta_{n-1}) - \vartheta_i \\ &= \Theta_{n-1}^{(i)} - \vartheta_i - \frac{2}{\mathcal{K}+\kappa} [\lambda_i (\Theta_{n-1}^{(i)} - \vartheta_i)] \\ &= \left(1 - \frac{2\lambda_i}{\mathcal{K}+\kappa} \right) (\Theta_{n-1}^{(i)} - \vartheta_i). \end{aligned} \quad (6.299)$$

Hence, we obtain that for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned}
 \|\Theta_n - \vartheta\|_2^2 &= \sum_{i=1}^{\mathfrak{d}} |\Theta_n^{(i)} - \vartheta_i|^2 \\
 &= \sum_{i=1}^{\mathfrak{d}} \left[\left| 1 - \frac{2\lambda_i}{\mathcal{K}+\kappa} \right|^2 |\Theta_{n-1}^{(i)} - \vartheta_i|^2 \right] \\
 &\leq \left[\max \left\{ \left| 1 - \frac{2\lambda_1}{\mathcal{K}+\kappa} \right|^2, \dots, \left| 1 - \frac{2\lambda_{\mathfrak{d}}}{\mathcal{K}+\kappa} \right|^2 \right\} \right] \left[\sum_{i=1}^{\mathfrak{d}} |\Theta_{n-1}^{(i)} - \vartheta_i|^2 \right] \\
 &= \left[\max \left\{ \left| 1 - \frac{2\lambda_1}{\mathcal{K}+\kappa} \right|, \dots, \left| 1 - \frac{2\lambda_{\mathfrak{d}}}{\mathcal{K}+\kappa} \right| \right\} \right]^2 \|\Theta_{n-1} - \vartheta\|_2^2
 \end{aligned} \tag{6.300}$$

(cf. Definition 3.3.4). Moreover, note that the fact that for all $i \in \{1, 2, \dots, \mathfrak{d}\}$ it holds that $\lambda_i \geq \kappa$ implies that for all $i \in \{1, 2, \dots, \mathfrak{d}\}$ it holds that

$$1 - \frac{2\lambda_i}{\mathcal{K}+\kappa} \leq 1 - \frac{2\kappa}{\mathcal{K}+\kappa} = \frac{\mathcal{K}+\kappa-2\kappa}{\mathcal{K}+\kappa} = \frac{\mathcal{K}-\kappa}{\mathcal{K}+\kappa} \geq 0. \tag{6.301}$$

In addition, observe that the fact that for all $i \in \{1, 2, \dots, \mathfrak{d}\}$ it holds that $\lambda_i \leq \mathcal{K}$ implies that for all $i \in \{1, 2, \dots, \mathfrak{d}\}$ it holds that

$$1 - \frac{2\lambda_i}{\mathcal{K}+\kappa} \geq 1 - \frac{2\mathcal{K}}{\mathcal{K}+\kappa} = \frac{\mathcal{K}+\kappa-2\mathcal{K}}{\mathcal{K}+\kappa} = -\left[\frac{\mathcal{K}-\kappa}{\mathcal{K}+\kappa} \right] \leq 0. \tag{6.302}$$

This and (6.301) ensure that for all $i \in \{1, 2, \dots, \mathfrak{d}\}$ it holds that

$$\left| 1 - \frac{2\lambda_i}{\mathcal{K}+\kappa} \right| \leq \frac{\mathcal{K}-\kappa}{\mathcal{K}+\kappa}. \tag{6.303}$$

Combining this with (6.300) demonstrates that for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned}
 \|\Theta_n - \vartheta\|_2 &\leq \left[\max \left\{ \left| 1 - \frac{2\lambda_1}{\mathcal{K}+\kappa} \right|, \dots, \left| 1 - \frac{2\lambda_{\mathfrak{d}}}{\mathcal{K}+\kappa} \right| \right\} \right] \|\Theta_{n-1} - \vartheta\|_2 \\
 &\leq \left[\frac{\mathcal{K}-\kappa}{\mathcal{K}+\kappa} \right] \|\Theta_{n-1} - \vartheta\|_2.
 \end{aligned} \tag{6.304}$$

Induction therefore establishes that for all $n \in \mathbb{N}_0$ it holds that

$$\|\Theta_n - \vartheta\|_2 \leq \left[\frac{\mathcal{K}-\kappa}{\mathcal{K}+\kappa} \right]^n \|\Theta_0 - \vartheta\|_2 = \left[\frac{\mathcal{K}-\kappa}{\mathcal{K}+\kappa} \right]^n \|\xi - \vartheta\|_2. \tag{6.305}$$

The proof of Lemma 6.3.22 is thus complete. \square

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 different choice of the learning rate.

Lemma 6.3.23 (Lower bound for the convergence rate of **GD** for quadratic objective functions). Let $\mathfrak{d} \in \mathbb{N}$, $\xi = (\xi_1, \dots, \xi_{\mathfrak{d}})$, $\vartheta = (\vartheta_1, \dots, \vartheta_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$, $\gamma, \kappa, \mathcal{K}, \lambda_1, \lambda_2, \dots, \lambda_{\mathfrak{d}} \in (0, \infty)$ satisfy $\kappa = \min\{\lambda_1, \lambda_2, \dots, \lambda_{\mathfrak{d}}\}$ and $\mathcal{K} = \max\{\lambda_1, \lambda_2, \dots, \lambda_{\mathfrak{d}}\}$, let $\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$ satisfy for all $\theta = (\theta_1, \theta_2, \dots, \theta_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$ that

$$\mathcal{L}(\theta) = \frac{1}{2} \left[\sum_{i=1}^{\mathfrak{d}} \lambda_i |\theta_i - \vartheta_i|^2 \right], \quad (6.306)$$

and let $\Theta: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $n \in \mathbb{N}$ that

$$\Theta_0 = \xi \quad \text{and} \quad \Theta_n = \Theta_{n-1} - \gamma(\nabla \mathcal{L})(\Theta_{n-1}). \quad (6.307)$$

Then it holds for all $n \in \mathbb{N}_0$ that

$$\begin{aligned} \|\Theta_n - \vartheta\|_2 &\geq [\max\{\gamma\mathcal{K} - 1, 1 - \gamma\kappa\}]^n [\min\{|\xi_1 - \vartheta_1|, \dots, |\xi_{\mathfrak{d}} - \vartheta_{\mathfrak{d}}|\}] \\ &\geq \left[\frac{\mathcal{K} - \kappa}{\mathcal{K} + \kappa}\right]^n [\min\{|\xi_1 - \vartheta_1|, \dots, |\xi_{\mathfrak{d}} - \vartheta_{\mathfrak{d}}|\}] \end{aligned} \quad (6.308)$$

(cf. Definition 3.3.4).

Proof of Lemma 6.3.23. Throughout this proof, let $\Theta^{(i)}: \mathbb{N}_0 \rightarrow \mathbb{R}$, $i \in \{1, 2, \dots, \mathfrak{d}\}$, satisfy for all $n \in \mathbb{N}_0$ that $\Theta_n = (\Theta_n^{(1)}, \dots, \Theta_n^{(\mathfrak{d})})$ and let $\iota, \mathcal{I} \in \{1, 2, \dots, \mathfrak{d}\}$ satisfy $\lambda_{\iota} = \kappa$ and $\lambda_{\mathcal{I}} = \mathcal{K}$. Observe that (6.306) implies that for all $\theta = (\theta_1, \dots, \theta_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$, $i \in \{1, 2, \dots, \mathfrak{d}\}$ it holds that

$$\left(\frac{\partial \mathcal{L}}{\partial \theta_i}\right)(\theta) = \lambda_i(\theta_i - \vartheta_i). \quad (6.309)$$

Combining this with (6.307) implies that for all $n \in \mathbb{N}$, $i \in \{1, 2, \dots, \mathfrak{d}\}$ it holds that

$$\begin{aligned} \Theta_n^{(i)} - \vartheta_i &= \Theta_{n-1}^{(i)} - \gamma \left(\frac{\partial \mathcal{L}}{\partial \theta_i}\right)(\Theta_{n-1}) - \vartheta_i \\ &= \Theta_{n-1}^{(i)} - \vartheta_i - \gamma \lambda_i (\Theta_{n-1}^{(i)} - \vartheta_i) \\ &= (1 - \gamma \lambda_i) (\Theta_{n-1}^{(i)} - \vartheta_i). \end{aligned} \quad (6.310)$$

Induction and (6.307) hence prove that for all $n \in \mathbb{N}_0$, $i \in \{1, 2, \dots, \mathfrak{d}\}$ it holds that

$$\Theta_n^{(i)} - \vartheta_i = (1 - \gamma \lambda_i)^n (\Theta_0^{(i)} - \vartheta_i) = (1 - \gamma \lambda_i)^n (\xi_i - \vartheta_i). \quad (6.311)$$

This shows that for all $n \in \mathbb{N}_0$ it holds that

$$\begin{aligned}
 \|\Theta_n - \vartheta\|_2^2 &= \sum_{i=1}^{\mathfrak{d}} |\Theta_n^{(i)} - \vartheta_i|^2 = \sum_{i=1}^{\mathfrak{d}} \left[|1 - \gamma \lambda_i|^{2n} |\xi_i - \vartheta_i|^2 \right] \\
 &\geq [\min\{|\xi_1 - \vartheta_1|^2, \dots, |\xi_{\mathfrak{d}} - \vartheta_{\mathfrak{d}}|^2\}] \left[\sum_{i=1}^{\mathfrak{d}} |1 - \gamma \lambda_i|^{2n} \right] \\
 &\geq [\min\{|\xi_1 - \vartheta_1|^2, \dots, |\xi_{\mathfrak{d}} - \vartheta_{\mathfrak{d}}|^2\}] [\max\{|1 - \gamma \lambda_1|^{2n}, \dots, |1 - \gamma \lambda_{\mathfrak{d}}|^{2n}\}] \\
 &= [\min\{|\xi_1 - \vartheta_1|, \dots, |\xi_{\mathfrak{d}} - \vartheta_{\mathfrak{d}}|\}]^2 [\max\{|1 - \gamma \lambda_1|, \dots, |1 - \gamma \lambda_{\mathfrak{d}}|\}]^{2n}
 \end{aligned} \tag{6.312}$$

(cf. Definition 3.3.4). Furthermore, note that

$$\begin{aligned}
 \max\{|1 - \gamma \lambda_1|, \dots, |1 - \gamma \lambda_{\mathfrak{d}}|\} &\geq \max\{|1 - \gamma \lambda_I|, |1 - \gamma \lambda_{\ell}|\} \\
 &= \max\{|1 - \gamma \mathcal{K}|, |1 - \gamma \kappa|\} = \max\{1 - \gamma \mathcal{K}, \gamma \mathcal{K} - 1, 1 - \gamma \kappa, \gamma \kappa - 1\} \\
 &= \max\{\gamma \mathcal{K} - 1, 1 - \gamma \kappa\}.
 \end{aligned} \tag{6.313}$$

In addition, observe that for all $\alpha \in (-\infty, \frac{2}{\mathcal{K}+\kappa}]$ it holds that

$$\max\{\alpha \mathcal{K} - 1, 1 - \alpha \kappa\} \geq 1 - \alpha \kappa \geq 1 - \left[\frac{2}{\mathcal{K}+\kappa}\right] \kappa = \frac{\mathcal{K}+\kappa-2\kappa}{\mathcal{K}+\kappa} = \frac{\mathcal{K}-\kappa}{\mathcal{K}+\kappa}. \tag{6.314}$$

Moreover, note that for all $\alpha \in [\frac{2}{\mathcal{K}+\kappa}, \infty)$ it holds that

$$\max\{\alpha \mathcal{K} - 1, 1 - \alpha \kappa\} \geq \alpha \mathcal{K} - 1 \geq \left[\frac{2}{\mathcal{K}+\kappa}\right] \mathcal{K} - 1 = \frac{2\mathcal{K}-(\mathcal{K}+\kappa)}{\mathcal{K}+\kappa} = \frac{\mathcal{K}-\kappa}{\mathcal{K}+\kappa}. \tag{6.315}$$

Combining this, (6.313), and (6.314) proves that

$$\max\{|1 - \gamma \lambda_1|, \dots, |1 - \gamma \lambda_{\mathfrak{d}}|\} \geq \max\{\gamma \mathcal{K} - 1, 1 - \gamma \kappa\} \geq \frac{\mathcal{K}-\kappa}{\mathcal{K}+\kappa} \geq 0. \tag{6.316}$$

This and (6.312) demonstrate that for all $n \in \mathbb{N}_0$ it holds that

$$\begin{aligned}
 \|\Theta_n - \vartheta\|_2 &\geq [\max\{|1 - \gamma \lambda_1|, \dots, |1 - \gamma \lambda_{\mathfrak{d}}|\}]^n [\min\{|\xi_1 - \vartheta_1|, \dots, |\xi_{\mathfrak{d}} - \vartheta_{\mathfrak{d}}|\}] \\
 &\geq [\max\{\gamma \mathcal{K} - 1, 1 - \gamma \kappa\}]^n [\min\{|\xi_1 - \vartheta_1|, \dots, |\xi_{\mathfrak{d}} - \vartheta_{\mathfrak{d}}|\}] \\
 &\geq \left[\frac{\mathcal{K}-\kappa}{\mathcal{K}+\kappa}\right]^n [\min\{|\xi_1 - \vartheta_1|, \dots, |\xi_{\mathfrak{d}} - \vartheta_{\mathfrak{d}}|\}].
 \end{aligned} \tag{6.317}$$

The proof of Lemma 6.3.23 is thus complete. \square

6.3.5.2 Error analysis for momentum GD optimization 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 example, be found in Silvester [398, Theorem 3].

Lemma 6.3.24 (A special case of Gelfand's spectral radius formula for real matrices). Let $\mathfrak{d} \in \mathbb{N}$, $A \in \mathbb{R}^{\mathfrak{d} \times \mathfrak{d}}$, $\mathcal{S} = \{\lambda \in \mathbb{C}: (\exists v \in \mathbb{C}^{\mathfrak{d}} \setminus \{0\}: Av = \lambda v)\}$ and let $\|\cdot\|: \mathbb{R}^{\mathfrak{d}} \rightarrow [0, \infty)$ be a norm. Then

$$\liminf_{n \rightarrow \infty} \left(\left[\sup_{v \in \mathbb{R}^{\mathfrak{d}} \setminus \{0\}} \frac{\|A^n v\|}{\|v\|} \right]^{1/n} \right) = \limsup_{n \rightarrow \infty} \left(\left[\sup_{v \in \mathbb{R}^{\mathfrak{d}} \setminus \{0\}} \frac{\|A^n v\|}{\|v\|} \right]^{1/n} \right) = \max_{\lambda \in \mathcal{S} \cup \{0\}} |\lambda|. \quad (6.318)$$

Proof of Lemma 6.3.24. Note that, for instance, Einsiedler & Ward [133, Theorem 11.6] establishes (6.318) (cf., for example, Tropp [416]). The proof of Lemma 6.3.24 is thus complete. \square

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

$$\det \underbrace{\begin{pmatrix} A & B \\ C & D \end{pmatrix}}_{\in \mathbb{R}^{(2\mathfrak{d}) \times (2\mathfrak{d})}} = \det(AD - BC) \quad (6.319)$$

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

$$\mathcal{D}_x = D - x \mathbf{I}_{\mathfrak{d}} \quad (6.320)$$

(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

$$\det \begin{pmatrix} X & Y \\ 0 & Z \end{pmatrix} = \det(X) \det(Z) = \det \begin{pmatrix} X & 0 \\ Y & Z \end{pmatrix} \quad (6.321)$$

(cf., for instance, Petersen [351, Proposition 5.5.3 and Proposition 5.5.4]) imply that for all $x \in \mathbb{C}$ it holds that

$$\begin{aligned} \det \left(\begin{pmatrix} A & B \\ C & \mathcal{D}_x \end{pmatrix} \begin{pmatrix} \mathcal{D}_x & 0 \\ -C & \mathbf{I}_{\mathfrak{d}} \end{pmatrix} \right) &= \det \begin{pmatrix} (A\mathcal{D}_x - BC) & B \\ (C\mathcal{D}_x - \mathcal{D}_x C) & \mathcal{D}_x \end{pmatrix} \\ &= \det \begin{pmatrix} (A\mathcal{D}_x - BC) & B \\ 0 & \mathcal{D}_x \end{pmatrix} \\ &= \det(A\mathcal{D}_x - BC) \det(\mathcal{D}_x). \end{aligned} \quad (6.322)$$

Moreover, note that (6.321) and the multiplicative property of the determinant (see, for

example, Petersen [351, (1) in Proposition 5.5.2]) imply that for all $x \in \mathbb{C}$ it holds that

$$\begin{aligned} \det\left(\begin{pmatrix} A & B \\ C & \mathcal{D}_x \end{pmatrix} \begin{pmatrix} \mathcal{D}_x & 0 \\ -C & I_{\mathfrak{d}} \end{pmatrix}\right) &= \det\left(\begin{pmatrix} A & B \\ C & \mathcal{D}_x \end{pmatrix}\right) \det\left(\begin{pmatrix} \mathcal{D}_x & 0 \\ -C & I_{\mathfrak{d}} \end{pmatrix}\right) \\ &= \det\left(\begin{pmatrix} A & B \\ C & \mathcal{D}_x \end{pmatrix}\right) \det(\mathcal{D}_x) \det(I_{\mathfrak{d}}) \\ &= \det\left(\begin{pmatrix} A & B \\ C & \mathcal{D}_x \end{pmatrix}\right) \det(\mathcal{D}_x). \end{aligned} \quad (6.323)$$

Combining this and (6.322) demonstrates that for all $x \in \mathbb{C}$ it holds that

$$\det\left(\begin{pmatrix} A & B \\ C & \mathcal{D}_x \end{pmatrix}\right) \det(\mathcal{D}_x) = \det(A\mathcal{D}_x - BC) \det(\mathcal{D}_x). \quad (6.324)$$

Hence, we obtain for all $x \in \mathbb{C}$ that

$$\left(\det\left(\begin{pmatrix} A & B \\ C & \mathcal{D}_x \end{pmatrix}\right) - \det(A\mathcal{D}_x - BC) \right) \det(\mathcal{D}_x) = 0. \quad (6.325)$$

This implies that for all $x \in \mathbb{C}$ with $\det(\mathcal{D}_x) \neq 0$ it holds that

$$\det\left(\begin{pmatrix} A & B \\ C & \mathcal{D}_x \end{pmatrix}\right) - \det(A\mathcal{D}_x - BC) = 0. \quad (6.326)$$

Moreover, note that the fact that $\mathbb{C} \ni x \mapsto \det(D - x I_{\mathfrak{d}}) \in \mathbb{C}$ is a polynomial function of degree \mathfrak{d} ensures that $\{x \in \mathbb{C}: \det(\mathcal{D}_x) = 0\} = \{x \in \mathbb{C}: \det(D - x I_{\mathfrak{d}}) = 0\}$ is a finite set. Combining this and (6.326) with the fact that the function

$$\mathbb{C} \ni x \mapsto \det\left(\begin{pmatrix} A & B \\ C & \mathcal{D}_x \end{pmatrix}\right) - \det(A\mathcal{D}_x - BC) \in \mathbb{C} \quad (6.327)$$

is continuous shows that for all $x \in \mathbb{C}$ it holds that

$$\det\left(\begin{pmatrix} A & B \\ C & \mathcal{D}_x \end{pmatrix}\right) - \det(A\mathcal{D}_x - BC) = 0. \quad (6.328)$$

Hence, we obtain for all $x \in \mathbb{C}$ that

$$\det\left(\begin{pmatrix} A & B \\ C & \mathcal{D}_x \end{pmatrix}\right) = \det(A\mathcal{D}_x - BC). \quad (6.329)$$

This establishes that

$$\det\left(\begin{pmatrix} A & B \\ C & D \end{pmatrix}\right) = \det\left(\begin{pmatrix} A & B \\ C & \mathcal{D}_0 \end{pmatrix}\right) = \det(A\mathcal{D}_0 - BC) = \det(A\mathcal{D}_0 - BC). \quad (6.330)$$

The proof of Lemma 6.3.25 is thus completed. \square

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 $\mathfrak{d} \in \mathbb{N}$, $\xi \in \mathbb{R}^{\mathfrak{d}}$, $\vartheta = (\vartheta_1, \dots, \vartheta_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$, $\kappa, \mathcal{K}, \lambda_1, \lambda_2, \dots, \lambda_{\mathfrak{d}} \in (0, \infty)$ satisfy $\kappa = \min\{\lambda_1, \lambda_2, \dots, \lambda_{\mathfrak{d}}\}$ and $\mathcal{K} = \max\{\lambda_1, \lambda_2, \dots, \lambda_{\mathfrak{d}}\}$, let $\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$ satisfy for all $\theta = (\theta_1, \dots, \theta_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$ that*

$$\mathcal{L}(\theta) = \frac{1}{2} \left[\sum_{i=1}^{\mathfrak{d}} \lambda_i |\theta_i - \vartheta_i|^2 \right], \quad (6.331)$$

and let $\Theta: \mathbb{N}_0 \cup \{-1\} \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $n \in \mathbb{N}$ that $\Theta_{-1} = \Theta_0 = \xi$ and

$$\Theta_n = \Theta_{n-1} - \frac{4}{(\sqrt{\mathcal{K}} + \sqrt{\kappa})^2} (\nabla \mathcal{L})(\Theta_{n-1}) + \left[\frac{\sqrt{\mathcal{K}} - \sqrt{\kappa}}{\sqrt{\mathcal{K}} + \sqrt{\kappa}} \right]^2 (\Theta_{n-1} - \Theta_{n-2}). \quad (6.332)$$

Then

- (i) it holds that $\Theta|_{\mathbb{N}_0}: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ is the momentum **GD** process for the objective function \mathcal{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

$$\|\Theta_n - \vartheta\|_2 \leq \mathfrak{c} \left[\frac{\sqrt{\mathcal{K}} - \sqrt{\kappa}}{\sqrt{\mathcal{K}} + \sqrt{\kappa}} + \varepsilon \right]^n \quad (6.333)$$

(cf. Definitions 3.3.4 and 6.3.1).

Proof of Proposition 6.3.26. Throughout this proof, let $\varepsilon \in (0, \infty)$, let $\|\cdot\|: \mathbb{R}^{(2\mathfrak{d}) \times (2\mathfrak{d})} \rightarrow [0, \infty)$ satisfy for all $B \in \mathbb{R}^{(2\mathfrak{d}) \times (2\mathfrak{d})}$ that

$$\|B\| = \sup_{v \in \mathbb{R}^{2\mathfrak{d}} \setminus \{0\}} \left[\frac{\|Bv\|_2}{\|v\|_2} \right], \quad (6.334)$$

let $\Theta^{(i)}: \mathbb{N}_0 \rightarrow \mathbb{R}$, $i \in \{1, 2, \dots, \mathfrak{d}\}$, satisfy for all $n \in \mathbb{N}_0$ that $\Theta_n = (\Theta_n^{(1)}, \dots, \Theta_n^{(\mathfrak{d})})$, let $\mathbf{m}: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $n \in \mathbb{N}_0$ that

$$\mathbf{m}_n = -\sqrt{\mathcal{K}\kappa}(\Theta_n - \Theta_{n-1}), \quad (6.335)$$

let $\varrho \in (0, \infty)$, $\alpha \in [0, 1)$ be given by

$$\varrho = \frac{4}{(\sqrt{\mathcal{K}} + \sqrt{\kappa})^2} \quad \text{and} \quad \alpha = \left[\frac{\sqrt{\mathcal{K}} - \sqrt{\kappa}}{\sqrt{\mathcal{K}} + \sqrt{\kappa}} \right]^2, \quad (6.336)$$

let $M \in \mathbb{R}^{\mathfrak{d} \times \mathfrak{d}}$ be the diagonal $(\mathfrak{d} \times \mathfrak{d})$ -matrix given by

$$M = \begin{pmatrix} (1 - \varrho \lambda_1 + \alpha) & & 0 \\ & \ddots & \\ 0 & & (1 - \varrho \lambda_{\mathfrak{d}} + \alpha) \end{pmatrix}, \quad (6.337)$$

let $A \in \mathbb{R}^{2\mathfrak{d} \times 2\mathfrak{d}}$ be the $((2\mathfrak{d}) \times (2\mathfrak{d}))$ -matrix given by

$$A = \begin{pmatrix} M & (-\alpha I_{\mathfrak{d}}) \\ I_{\mathfrak{d}} & 0 \end{pmatrix}, \quad (6.338)$$

and let $\mathcal{S} \subseteq \mathbb{C}$ be the set given by

$$\mathcal{S} = \{\mu \in \mathbb{C}: (\exists v \in \mathbb{C}^{2\mathfrak{d}} \setminus \{0\}: Av = \mu v)\} = \{\mu \in \mathbb{C}: \det(A - \mu I_{2\mathfrak{d}}) = 0\} \quad (6.339)$$

(cf. Definition 1.5.5). Observe that (6.332), (6.335), and the fact that

$$\begin{aligned} \frac{(\sqrt{\mathcal{K}} + \sqrt{\kappa})^2 - (\sqrt{\mathcal{K}} - \sqrt{\kappa})^2}{4} &= \frac{1}{4} \left[(\sqrt{\mathcal{K}} + \sqrt{\kappa} + \sqrt{\mathcal{K}} - \sqrt{\kappa})(\sqrt{\mathcal{K}} + \sqrt{\kappa} - [\sqrt{\mathcal{K}} - \sqrt{\kappa}]) \right] \\ &= \frac{1}{4} \left[(2\sqrt{\mathcal{K}})(2\sqrt{\kappa}) \right] = \sqrt{\mathcal{K}\kappa} \end{aligned} \quad (6.340)$$

assure that for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned} \mathbf{m}_n &= -\sqrt{\mathcal{K}\kappa}(\Theta_n - \Theta_{n-1}) \\ &= -\sqrt{\mathcal{K}\kappa} \left(\Theta_{n-1} - \left[\frac{4}{(\sqrt{\mathcal{K}} + \sqrt{\kappa})^2} \right] (\nabla \mathcal{L})(\Theta_{n-1}) + \left[\frac{\sqrt{\mathcal{K}} - \sqrt{\kappa}}{\sqrt{\mathcal{K}} + \sqrt{\kappa}} \right]^2 (\Theta_{n-1} - \Theta_{n-2}) - \Theta_{n-1} \right) \\ &= \sqrt{\mathcal{K}\kappa} \left(\left[\frac{4}{(\sqrt{\mathcal{K}} + \sqrt{\kappa})^2} \right] (\nabla \mathcal{L})(\Theta_{n-1}) - \left[\frac{\sqrt{\mathcal{K}} - \sqrt{\kappa}}{\sqrt{\mathcal{K}} + \sqrt{\kappa}} \right]^2 (\Theta_{n-1} - \Theta_{n-2}) \right) \\ &= \frac{(\sqrt{\mathcal{K}} + \sqrt{\kappa})^2 - (\sqrt{\mathcal{K}} - \sqrt{\kappa})^2}{4} \left[\frac{4}{(\sqrt{\mathcal{K}} + \sqrt{\kappa})^2} \right] (\nabla \mathcal{L})(\Theta_{n-1}) \\ &\quad - \sqrt{\mathcal{K}\kappa} \left[\frac{\sqrt{\mathcal{K}} - \sqrt{\kappa}}{\sqrt{\mathcal{K}} + \sqrt{\kappa}} \right]^2 (\Theta_{n-1} - \Theta_{n-2}) \\ &= \left[1 - \frac{(\sqrt{\mathcal{K}} - \sqrt{\kappa})^2}{(\sqrt{\mathcal{K}} + \sqrt{\kappa})^2} \right] (\nabla \mathcal{L})(\Theta_{n-1}) + \left[\frac{\sqrt{\mathcal{K}} - \sqrt{\kappa}}{\sqrt{\mathcal{K}} + \sqrt{\kappa}} \right]^2 \left[-\sqrt{\mathcal{K}\kappa}(\Theta_{n-1} - \Theta_{n-2}) \right] \\ &= \left[1 - \left[\frac{\sqrt{\mathcal{K}} - \sqrt{\kappa}}{\sqrt{\mathcal{K}} + \sqrt{\kappa}} \right]^2 \right] (\nabla \mathcal{L})(\Theta_{n-1}) + \left[\frac{\sqrt{\mathcal{K}} - \sqrt{\kappa}}{\sqrt{\mathcal{K}} + \sqrt{\kappa}} \right]^2 \mathbf{m}_{n-1}. \end{aligned} \quad (6.341)$$

Moreover, note that (6.335) implies that for all $n \in \mathbb{N}_0$ it holds that

$$\begin{aligned} \Theta_n &= \Theta_{n-1} + (\Theta_n - \Theta_{n-1}) \\ &= \Theta_{n-1} - \frac{1}{\sqrt{\mathcal{K}\kappa}} \left(\left[-\sqrt{\mathcal{K}\kappa} \right] (\Theta_n - \Theta_{n-1}) \right) = \Theta_{n-1} - \frac{1}{\sqrt{\mathcal{K}\kappa}} \mathbf{m}_n. \end{aligned} \quad (6.342)$$

In addition, observe that the assumption that $\Theta_{-1} = \Theta_0 = \xi$ and (6.335) ensure that

$$\mathbf{m}_0 = -\sqrt{\mathcal{K}\kappa} (\Theta_0 - \Theta_{-1}) = 0. \quad (6.343)$$

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 $\theta = (\theta_1, \dots, \theta_d) \in \mathbb{R}^d$, $i \in \{1, 2, \dots, d\}$ it holds that

$$\left(\frac{\partial \mathcal{L}}{\partial \theta_i}\right)(\theta) = \lambda_i(\theta_i - \vartheta_i). \quad (6.344)$$

This, (6.332), and (6.336) imply that for all $n \in \mathbb{N}$, $i \in \{1, 2, \dots, d\}$ it holds that

$$\begin{aligned} \Theta_n^{(i)} - \vartheta_i &= \Theta_{n-1}^{(i)} - \varrho \left(\frac{\partial \mathcal{L}}{\partial \theta_i} \right) (\Theta_{n-1}) + \alpha (\Theta_{n-1}^{(i)} - \Theta_{n-2}^{(i)}) - \vartheta_i \\ &= (\Theta_{n-1}^{(i)} - \vartheta_i) - \varrho \lambda_i (\Theta_{n-1}^{(i)} - \vartheta_i) + \alpha ((\Theta_{n-1}^{(i)} - \vartheta_i) - (\Theta_{n-2}^{(i)} - \vartheta_i)) \\ &= (1 - \varrho \lambda_i + \alpha) (\Theta_{n-1}^{(i)} - \vartheta_i) - \alpha (\Theta_{n-2}^{(i)} - \vartheta_i). \end{aligned} \quad (6.345)$$

Combining this with (6.337) demonstrates that for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned} \mathbb{R}^d \ni (\Theta_n - \vartheta) &= M(\Theta_{n-1} - \vartheta) - \alpha(\Theta_{n-2} - \vartheta) \\ &= \underbrace{(M - (-\alpha I_d))}_{\in \mathbb{R}^{d \times 2d}} \underbrace{\begin{pmatrix} \Theta_{n-1} - \vartheta \\ \Theta_{n-2} - \vartheta \end{pmatrix}}_{\in \mathbb{R}^{2d}}. \end{aligned} \quad (6.346)$$

This and (6.338) assure that for all $n \in \mathbb{N}$ it holds that

$$\mathbb{R}^{2d} \ni \begin{pmatrix} \Theta_n - \vartheta \\ \Theta_{n-1} - \vartheta \end{pmatrix} = \begin{pmatrix} M & (-\alpha I_d) \\ I_d & 0 \end{pmatrix} \begin{pmatrix} \Theta_{n-1} - \vartheta \\ \Theta_{n-2} - \vartheta \end{pmatrix} = A \begin{pmatrix} \Theta_{n-1} - \vartheta \\ \Theta_{n-2} - \vartheta \end{pmatrix}. \quad (6.347)$$

Induction hence proves that for all $n \in \mathbb{N}_0$ it holds that

$$\mathbb{R}^{2d} \ni \begin{pmatrix} \Theta_n - \vartheta \\ \Theta_{n-1} - \vartheta \end{pmatrix} = A^n \begin{pmatrix} \Theta_0 - \vartheta \\ \Theta_{-1} - \vartheta \end{pmatrix} = A^n \begin{pmatrix} \xi - \vartheta \\ \xi - \vartheta \end{pmatrix}. \quad (6.348)$$

This implies that for all $n \in \mathbb{N}_0$ it holds that

$$\begin{aligned} \|\Theta_n - \vartheta\|_2 &\leq \sqrt{\|\Theta_n - \vartheta\|_2^2 + \|\Theta_{n-1} - \vartheta\|_2^2} \\ &= \left\| \begin{pmatrix} \Theta_n - \vartheta \\ \Theta_{n-1} - \vartheta \end{pmatrix} \right\|_2 \\ &= \left\| A^n \begin{pmatrix} \xi - \vartheta \\ \xi - \vartheta \end{pmatrix} \right\|_2 \\ &\leq \|A^n\| \left\| \begin{pmatrix} \xi - \vartheta \\ \xi - \vartheta \end{pmatrix} \right\|_2 \\ &= \|A^n\| \sqrt{\|\xi - \vartheta\|_2^2 + \|\xi - \vartheta\|_2^2} \\ &= \|A^n\| \sqrt{2} \|\xi - \vartheta\|_2. \end{aligned} \quad (6.349)$$

Next note that (6.339) and Lemma 6.3.24 demonstrate that

$$\limsup_{n \rightarrow \infty} (\|A^n\|)^{1/n} = \liminf_{n \rightarrow \infty} (\|A^n\|)^{1/n} = \max_{\mu \in \mathcal{S} \cup \{0\}} |\mu|. \quad (6.350)$$

This implies that there exists $m \in \mathbb{N}$ which satisfies for all $n \in \mathbb{N} \cap [m, \infty)$ that

$$(\|A^n\|)^{1/n} \leq \varepsilon + \max_{\mu \in \mathcal{S} \cup \{0\}} |\mu|. \quad (6.351)$$

Note that (6.351) implies that for all $n \in \mathbb{N} \cap [m, \infty)$ it holds that

$$\|A^n\| \leq \left[\varepsilon + \max_{\mu \in \mathcal{S} \cup \{0\}} |\mu| \right]^n. \quad (6.352)$$

Furthermore, note that for all $n \in \mathbb{N} \cap [0, m)$ it holds that

$$\begin{aligned} \|A^n\| &= \left[\varepsilon + \max_{\mu \in \mathcal{S} \cup \{0\}} |\mu| \right]^n \left[\frac{\|A^n\|}{(\varepsilon + \max_{\mu \in \mathcal{S} \cup \{0\}} |\mu|)^n} \right] \\ &\leq \left[\varepsilon + \max_{\mu \in \mathcal{S} \cup \{0\}} |\mu| \right]^n \left[\max \left(\left\{ \frac{\|A^k\|}{(\varepsilon + \max_{\mu \in \mathcal{S} \cup \{0\}} |\mu|)^k} : k \in \mathbb{N}_0 \cap [0, m) \right\} \cup \{1\} \right) \right]. \end{aligned} \quad (6.353)$$

Combining this and (6.352) proves that for all $n \in \mathbb{N}_0$ it holds that

$$\|A^n\| \leq \left[\varepsilon + \max_{\mu \in \mathcal{S} \cup \{0\}} |\mu| \right]^n \left[\max \left(\left\{ \frac{\|A^k\|}{(\varepsilon + \max_{\mu \in \mathcal{S} \cup \{0\}} |\mu|)^k} : k \in \mathbb{N}_0 \cap [0, m) \right\} \cup \{1\} \right) \right]. \quad (6.354)$$

Next observe that Lemma 6.3.25, (6.338), and the fact that for all $\mu \in \mathbb{C}$ it holds that $I_{\mathfrak{d}}(-\mu I_{\mathfrak{d}}) = -\mu I_{\mathfrak{d}} = (-\mu I_{\mathfrak{d}})I_{\mathfrak{d}}$ ensure that for all $\mu \in \mathbb{C}$ it holds that

$$\begin{aligned} \det(A - \mu I_{2\mathfrak{d}}) &= \det \begin{pmatrix} (M - \mu I_{\mathfrak{d}}) & (-\alpha I_{\mathfrak{d}}) \\ I_{\mathfrak{d}} & -\mu I_{\mathfrak{d}} \end{pmatrix} \\ &= \det((M - \mu I_{\mathfrak{d}})(-\mu I_{\mathfrak{d}}) - (-\alpha I_{\mathfrak{d}})I_{\mathfrak{d}}) \\ &= \det((M - \mu I_{\mathfrak{d}})(-\mu I_{\mathfrak{d}}) + \alpha I_{\mathfrak{d}}). \end{aligned} \quad (6.355)$$

This and (6.337) demonstrate that for all $\mu \in \mathbb{C}$ it holds that

$$\begin{aligned} \det(A - \mu I_{2\mathfrak{d}}) &= \det \begin{pmatrix} ((1 - \varrho\lambda_1 + \alpha - \mu)(-\mu) + \alpha) & 0 \\ 0 & ((1 - \varrho\lambda_{\mathfrak{d}} + \alpha - \mu)(-\mu) + \alpha) \end{pmatrix} \\ &= \prod_{i=1}^{\mathfrak{d}} ((1 - \varrho\lambda_i + \alpha - \mu)(-\mu) + \alpha) \\ &= \prod_{i=1}^{\mathfrak{d}} (\mu^2 - (1 - \varrho\lambda_i + \alpha)\mu + \alpha). \end{aligned} \quad (6.356)$$

Moreover, note that for all $\mu \in \mathbb{C}$, $i \in \{1, 2, \dots, \mathfrak{d}\}$ it holds that

$$\begin{aligned} \mu^2 - (1 - \varrho\lambda_i + \alpha)\mu + \alpha &= \mu^2 - 2\mu \left[\frac{(1-\varrho\lambda_i+\alpha)}{2} \right] + \left[\frac{(1-\varrho\lambda_i+\alpha)}{2} \right]^2 + \alpha - \left[\frac{(1-\varrho\lambda_i+\alpha)}{2} \right]^2 \\ &= \left[\mu - \frac{(1-\varrho\lambda_i+\alpha)}{2} \right]^2 + \alpha - \frac{1}{4}[1 - \varrho\lambda_i + \alpha]^2 \\ &= \left[\mu - \frac{(1-\varrho\lambda_i+\alpha)}{2} \right]^2 - \frac{1}{4} \left[[1 - \varrho\lambda_i + \alpha]^2 - 4\alpha \right]. \end{aligned} \quad (6.357)$$

Hence, we obtain that for all $i \in \{1, 2, \dots, \mathfrak{d}\}$ it holds that

$$\begin{aligned} &\{\mu \in \mathbb{C}: \mu^2 - (1 - \varrho\lambda_i + \alpha)\mu + \alpha = 0\} \\ &= \left\{ \mu \in \mathbb{C}: \left[\mu - \frac{(1-\varrho\lambda_i+\alpha)}{2} \right]^2 = \frac{1}{4} \left[[1 - \varrho\lambda_i + \alpha]^2 - 4\alpha \right] \right\} \\ &= \left\{ \frac{(1-\varrho\lambda_i+\alpha)+\sqrt{[1-\varrho\lambda_i+\alpha]^2-4\alpha}}{2}, \frac{(1-\varrho\lambda_i+\alpha)-\sqrt{[1-\varrho\lambda_i+\alpha]^2-4\alpha}}{2}, \right\} \\ &= \bigcup_{s \in \{-1, 1\}} \left\{ \frac{1}{2} \left[1 - \varrho\lambda_i + \alpha + s\sqrt{(1 - \varrho\lambda_i + \alpha)^2 - 4\alpha} \right] \right\}. \end{aligned} \quad (6.358)$$

Combining this, (6.339), and (6.356) demonstrates that

$$\begin{aligned} \mathcal{S} &= \{\mu \in \mathbb{C}: \det(A - \mu I_{2\mathfrak{d}}) = 0\} \\ &= \left\{ \mu \in \mathbb{C}: \left[\prod_{i=1}^{\mathfrak{d}} (\mu^2 - (1 - \varrho\lambda_i + \alpha)\mu + \alpha) = 0 \right] \right\} \\ &= \bigcup_{i=1}^{\mathfrak{d}} \{\mu \in \mathbb{C}: \mu^2 - (1 - \varrho\lambda_i + \alpha)\mu + \alpha = 0\} \\ &= \bigcup_{i=1}^{\mathfrak{d}} \bigcup_{s \in \{-1, 1\}} \left\{ \frac{1}{2} \left[1 - \varrho\lambda_i + \alpha + s\sqrt{(1 - \varrho\lambda_i + \alpha)^2 - 4\alpha} \right] \right\}. \end{aligned} \quad (6.359)$$

Moreover, observe that the fact that for all $i \in \{1, 2, \dots, \mathfrak{d}\}$ it holds that $\lambda_i \geq \kappa$ and (6.336) ensure that for all $i \in \{1, 2, \dots, \mathfrak{d}\}$ it holds that

$$\begin{aligned} 1 - \varrho\lambda_i + \alpha &\leq 1 - \varrho\kappa + \alpha = 1 - \left[\frac{4}{(\sqrt{\kappa} + \sqrt{\kappa})^2} \right] \kappa + \frac{(\sqrt{\kappa} - \sqrt{\kappa})^2}{(\sqrt{\kappa} + \sqrt{\kappa})^2} \\ &= \frac{(\sqrt{\kappa} + \sqrt{\kappa})^2 - 4\kappa + (\sqrt{\kappa} - \sqrt{\kappa})^2}{(\sqrt{\kappa} + \sqrt{\kappa})^2} = \frac{\kappa + 2\sqrt{\kappa}\sqrt{\kappa} + \kappa - 4\kappa + \kappa - 2\sqrt{\kappa}\sqrt{\kappa} + \kappa}{(\sqrt{\kappa} + \sqrt{\kappa})^2} \\ &= \frac{2\kappa - 2\kappa}{(\sqrt{\kappa} + \sqrt{\kappa})^2} = \frac{2(\sqrt{\kappa} - \sqrt{\kappa})(\sqrt{\kappa} + \sqrt{\kappa})}{(\sqrt{\kappa} + \sqrt{\kappa})^2} = 2 \left[\frac{\sqrt{\kappa} - \sqrt{\kappa}}{\sqrt{\kappa} + \sqrt{\kappa}} \right] \geq 0. \end{aligned} \quad (6.360)$$

In addition, note that the fact that for all $i \in \{1, 2, \dots, \mathfrak{d}\}$ it holds that $\lambda_i \leq \mathcal{K}$ and (6.336)

6.3. Momentum optimization

assure that for all $i \in \{1, 2, \dots, d\}$ it holds that

$$\begin{aligned}
1 - \varrho \lambda_i + \alpha &\geq 1 - \varrho \mathcal{K} + \alpha = 1 - \left[\frac{4}{(\sqrt{\mathcal{K}} + \sqrt{\kappa})^2} \right] \mathcal{K} + \frac{(\sqrt{\mathcal{K}} - \sqrt{\kappa})^2}{(\sqrt{\mathcal{K}} + \sqrt{\kappa})^2} \\
&= \frac{(\sqrt{\mathcal{K}} + \sqrt{\kappa})^2 - 4\mathcal{K} + (\sqrt{\mathcal{K}} - \sqrt{\kappa})^2}{(\sqrt{\mathcal{K}} + \sqrt{\kappa})^2} = \frac{\mathcal{K} + 2\sqrt{\mathcal{K}}\sqrt{\kappa} - 4\mathcal{K} + \mathcal{K} - 2\sqrt{\mathcal{K}}\sqrt{\kappa} + \kappa}{(\sqrt{\mathcal{K}} + \sqrt{\kappa})^2} \\
&= \frac{-2\mathcal{K} + 2\kappa}{(\sqrt{\mathcal{K}} + \sqrt{\kappa})^2} = -2 \left[\frac{\mathcal{K} - \kappa}{(\sqrt{\mathcal{K}} + \sqrt{\kappa})^2} \right] = -2 \left[\frac{(\sqrt{\mathcal{K}} - \sqrt{\kappa})(\sqrt{\mathcal{K}} + \sqrt{\kappa})}{(\sqrt{\mathcal{K}} + \sqrt{\kappa})^2} \right] \\
&= -2 \left[\frac{\sqrt{\mathcal{K}} - \sqrt{\kappa}}{\sqrt{\mathcal{K}} + \sqrt{\kappa}} \right] \leq 0.
\end{aligned} \tag{6.361}$$

Combining this, (6.360), and (6.336) implies that for all $i \in \{1, 2, \dots, d\}$ it holds that

$$(1 - \varrho \lambda_i + \alpha)^2 \leq \left[2 \left(\frac{\sqrt{\mathcal{K}} - \sqrt{\kappa}}{\sqrt{\mathcal{K}} + \sqrt{\kappa}} \right) \right]^2 = 4 \left[\frac{\sqrt{\mathcal{K}} - \sqrt{\kappa}}{\sqrt{\mathcal{K}} + \sqrt{\kappa}} \right]^2 = 4\alpha. \tag{6.362}$$

This and (6.359) demonstrate that

$$\begin{aligned}
\max_{\mu \in \mathcal{S} \cup \{0\}} |\mu| &= \max_{\mu \in \mathcal{S}} |\mu| \\
&= \max_{i \in \{1, 2, \dots, d\}} \max_{s \in \{-1, 1\}} \left| \frac{1}{2} \left[1 - \varrho \lambda_i + \alpha + s \sqrt{(1 - \varrho \lambda_i + \alpha)^2 - 4\alpha} \right] \right| \\
&= \frac{1}{2} \left[\max_{i \in \{1, 2, \dots, d\}} \max_{s \in \{-1, 1\}} \left| \left[1 - \varrho \lambda_i + \alpha + s \sqrt{(-1)(4\alpha - [1 - \varrho \lambda_i + \alpha]^2)} \right] \right| \right] \\
&= \frac{1}{2} \left[\max_{i \in \{1, 2, \dots, d\}} \max_{s \in \{-1, 1\}} \left| \left[1 - \varrho \lambda_i + \alpha + s \sqrt{4\alpha - (1 - \varrho \lambda_i + \alpha)^2} \right] \right|^2 \right]^{1/2}.
\end{aligned} \tag{6.363}$$

Combining this with (6.362) proves that

$$\begin{aligned}
\max_{\mu \in \mathcal{S} \cup \{0\}} |\mu| &= \frac{1}{2} \left[\max_{i \in \{1, 2, \dots, d\}} \max_{s \in \{-1, 1\}} \left(|1 - \varrho \lambda_i + \alpha|^2 + |s \sqrt{4\alpha - (1 - \varrho \lambda_i + \alpha)^2}|^2 \right) \right]^{1/2} \\
&= \frac{1}{2} \left[\max_{i \in \{1, 2, \dots, d\}} \max_{s \in \{-1, 1\}} ((1 - \varrho \lambda_i + \alpha)^2 + 4\alpha - (1 - \varrho \lambda_i + \alpha)^2) \right]^{1/2} \\
&= \frac{1}{2} [4\alpha]^{1/2} = \sqrt{\alpha}.
\end{aligned} \tag{6.364}$$

Combining (6.349) and (6.354) hence ensures that for all $n \in \mathbb{N}_0$ it holds that

$$\begin{aligned}
\|\Theta_n - \vartheta\|_2 &\leq \sqrt{2} \|\xi - \vartheta\|_2 \|A^n\| \\
&\leq \sqrt{2} \|\xi - \vartheta\|_2 \left[\varepsilon + \max_{\mu \in \mathcal{S} \cup \{0\}} |\mu| \right]^n \\
&\quad \cdot \left[\max \left(\left\{ \frac{\|A^k\|}{(\varepsilon + \max_{\mu \in \mathcal{S} \cup \{0\}} |\mu|)^k} \in \mathbb{R} : k \in \mathbb{N}_0 \cap [0, m) \right\} \cup \{1\} \right) \right] \\
&= \sqrt{2} \|\xi - \vartheta\|_2 [\varepsilon + \alpha^{1/2}]^n \left[\max \left(\left\{ \frac{\|A^k\|}{(\varepsilon + \alpha^{1/2})^k} \in \mathbb{R} : k \in \mathbb{N}_0 \cap [0, m) \right\} \cup \{1\} \right) \right]^{303} \\
&= \sqrt{2} \|\xi - \vartheta\|_2 \left[\varepsilon + \frac{\sqrt{\mathcal{K}} - \sqrt{\kappa}}{\sqrt{\mathcal{K}} + \sqrt{\kappa}} \right]^n \left[\max \left(\left\{ \frac{\|A^k\|}{(\varepsilon + \alpha^{1/2})^k} \in \mathbb{R} : k \in \mathbb{N}_0 \cap [0, m) \right\} \cup \{1\} \right) \right].
\end{aligned} \tag{6.365}$$

This establishes item (ii). The proof of Proposition 6.3.26 it thus completed. \square

6.3.5.3 Comparison of the convergence speeds of GD and momentum optimization

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*

$$\frac{\sqrt{\mathcal{K}} - \sqrt{\kappa}}{\sqrt{\mathcal{K}} + \sqrt{\kappa}} < \frac{\mathcal{K} - \kappa}{\mathcal{K} + \kappa}. \quad (6.366)$$

Proof of Lemma 6.3.27. Note that the fact that $\mathcal{K} - \kappa > 0 < 2\sqrt{\mathcal{K}}\sqrt{\kappa}$ ensures that

$$\frac{\sqrt{\mathcal{K}} - \sqrt{\kappa}}{\sqrt{\mathcal{K}} + \sqrt{\kappa}} = \frac{(\sqrt{\mathcal{K}} - \sqrt{\kappa})(\sqrt{\mathcal{K}} + \sqrt{\kappa})}{(\sqrt{\mathcal{K}} + \sqrt{\kappa})^2} = \frac{\mathcal{K} - \kappa}{\mathcal{K} + 2\sqrt{\mathcal{K}}\sqrt{\kappa} + \kappa} < \frac{\mathcal{K} - \kappa}{\mathcal{K} + \kappa}. \quad (6.367)$$

The proof of Lemma 6.3.27 it thus completed. \square

Corollary 6.3.28 (Convergence speed comparisons between the GD optimization method and the momentum GD optimization method). *Let $\mathfrak{d} \in \mathbb{N}$, $\kappa, \mathcal{K}, \lambda_1, \lambda_2, \dots, \lambda_{\mathfrak{d}} \in (0, \infty)$, $\xi = (\xi_1, \dots, \xi_{\mathfrak{d}})$, $\vartheta = (\vartheta_1, \dots, \vartheta_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$ satisfy $\kappa = \min\{\lambda_1, \lambda_2, \dots, \lambda_{\mathfrak{d}}\} < \max\{\lambda_1, \lambda_2, \dots, \lambda_{\mathfrak{d}}\} = \mathcal{K}$, let $\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$ satisfy for all $\theta = (\theta_1, \dots, \theta_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$ that*

$$\mathcal{L}(\theta) = \frac{1}{2} \left[\sum_{i=1}^{\mathfrak{d}} \lambda_i |\theta_i - \vartheta_i|^2 \right], \quad (6.368)$$

for every $\gamma \in (0, \infty)$ let $\Theta^{\gamma}: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $n \in \mathbb{N}$ that

$$\Theta_0^{\gamma} = \xi \quad \text{and} \quad \Theta_n^{\gamma} = \Theta_{n-1}^{\gamma} - \gamma(\nabla \mathcal{L})(\Theta_{n-1}^{\gamma}), \quad (6.369)$$

and let $\mathcal{M}: \mathbb{N}_0 \cup \{-1\} \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $n \in \mathbb{N}$ that $\mathcal{M}_{-1} = \mathcal{M}_0 = \xi$ and

$$\mathcal{M}_n = \mathcal{M}_{n-1} - \frac{4}{(\sqrt{\mathcal{K}} + \sqrt{\kappa})^2} (\nabla \mathcal{L})(\mathcal{M}_{n-1}) + \left[\frac{\sqrt{\mathcal{K}} - \sqrt{\kappa}}{\sqrt{\mathcal{K}} + \sqrt{\kappa}} \right]^2 (\mathcal{M}_{n-1} - \mathcal{M}_{n-2}). \quad (6.370)$$

Then

(i) there exist $\gamma, \mathbf{c} \in (0, \infty)$ such that for all $n \in \mathbb{N}_0$ it holds that

$$\|\Theta_n^\gamma - \vartheta\|_2 \leq \mathbf{c} \left[\frac{\kappa - \kappa}{\kappa + \kappa} \right]^n, \quad (6.371)$$

(ii) it holds for all $\gamma \in (0, \infty), n \in \mathbb{N}_0$ that

$$\|\Theta_n^\gamma - \vartheta\|_2 \geq [\min\{|\xi_1 - \vartheta_1|, \dots, |\xi_d - \vartheta_d|\}] \left[\frac{\kappa - \kappa}{\kappa + \kappa} \right]^n, \quad (6.372)$$

(iii) for every $\varepsilon \in (0, \infty)$ there exists $\mathbf{c} \in (0, \infty)$ such that for all $n \in \mathbb{N}_0$ it holds that

$$\|\mathcal{M}_n - \vartheta\|_2 \leq \mathbf{c} \left[\frac{\sqrt{\kappa} - \sqrt{\kappa}}{\sqrt{\kappa} + \sqrt{\kappa}} + \varepsilon \right]^n, \quad (6.373)$$

and

(iv) it holds that $\frac{\sqrt{\kappa} - \sqrt{\kappa}}{\sqrt{\kappa} + \sqrt{\kappa}} < \frac{\kappa - \kappa}{\kappa + \kappa}$

(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. \square

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 $\lambda_1, \lambda_2, \dots, \lambda_d \in (0, \infty)$ in the objective function in (6.368) satisfy the assumption that

$$\min\{\lambda_1, \dots, \lambda_d\} < \max\{\lambda_1, \dots, \lambda_d\}. \quad (6.374)$$

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 $\min\{\lambda_1, \dots, \lambda_d\} = \max\{\lambda_1, \dots, \lambda_d\}$.

Lemma 6.3.29 (Concurrence of the **GD** optimization method and the momentum **GD** optimization method). *Let $d \in \mathbb{N}$, $\xi, \vartheta \in \mathbb{R}^d$, $\alpha \in (0, \infty)$, let $\mathcal{L}: \mathbb{R}^d \rightarrow \mathbb{R}$ satisfy for all $\theta \in \mathbb{R}^d$ that*

$$\mathcal{L}(\theta) = \frac{\alpha}{2} \|\theta - \vartheta\|_2^2, \quad (6.375)$$

let $\Theta: \mathbb{N}_0 \rightarrow \mathbb{R}^d$ satisfy for all $n \in \mathbb{N}$ that

$$\Theta_0 = \xi \quad \text{and} \quad \Theta_n = \Theta_{n-1} - \frac{2}{(\alpha+\alpha)} (\nabla \mathcal{L})(\Theta_{n-1}), \quad (6.376)$$

and let $\mathcal{M}: \mathbb{N}_0 \cup \{-1\} \rightarrow \mathbb{R}^d$ satisfy for all $n \in \mathbb{N}$ that $\mathcal{M}_{-1} = \mathcal{M}_0 = \xi$ and

$$\mathcal{M}_n = \mathcal{M}_{n-1} - \frac{4}{(\sqrt{\alpha}+\sqrt{\alpha})^2} (\nabla \mathcal{L})(\mathcal{M}_{n-1}) + \left[\frac{\sqrt{\alpha}-\sqrt{\alpha}}{\sqrt{\alpha}+\sqrt{\alpha}} \right]^2 (\mathcal{M}_{n-1} - \mathcal{M}_{n-2}) \quad (6.377)$$

(cf. Definition 3.3.4). Then

(i) it holds that $\mathcal{M}|_{\mathbb{N}_0}: \mathbb{N}_0 \rightarrow \mathbb{R}^d$ is the momentum GD process for the objective function \mathcal{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

$$\mathcal{M}_n = \mathcal{M}_{n-1} - \frac{4}{(2\sqrt{\alpha})^2} (\nabla \mathcal{L})(\mathcal{M}_{n-1}) = \mathcal{M}_{n-1} - \frac{1}{\alpha} (\nabla \mathcal{L})(\mathcal{M}_{n-1}). \quad (6.378)$$

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

$$\Theta_n = \Theta_{n-1} - \frac{1}{\alpha} (\nabla \mathcal{L})(\Theta_{n-1}). \quad (6.379)$$

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 $\theta \in \mathbb{R}^d$ it holds that

$$(\nabla \mathcal{L})(\theta) = \frac{\alpha}{2} (2(\theta - \vartheta)) = \alpha(\theta - \vartheta). \quad (6.380)$$

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

$$\Theta_n = \vartheta. \quad (6.381)$$

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

$$\Theta_1 = \Theta_0 - \frac{1}{\alpha} (\nabla \mathcal{L})(\Theta_0) = \xi - \frac{1}{\alpha} (\alpha(\xi - \vartheta)) = \xi - (\xi - \vartheta) = \vartheta. \quad (6.382)$$

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

$$\Theta_{n+1} = \Theta_n - \frac{1}{\alpha} (\nabla \mathcal{L})(\Theta_n) = \vartheta - \frac{1}{\alpha} (\alpha(\vartheta - \vartheta)) = \vartheta. \quad (6.383)$$

Induction thus proves (6.381). Combining (6.381) and item (ii) establishes item (iii). The proof of Lemma 6.3.29 is thus complete. \square

6.3.6 Numerical comparisons for GD and momentum optimization

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

$$\vartheta = \begin{pmatrix} \vartheta_1 \\ \vartheta_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad \text{and} \quad \xi = \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} = \begin{pmatrix} 5 \\ 3 \end{pmatrix}, \quad (6.384)$$

let $\mathcal{L}: \mathbb{R}^2 \rightarrow \mathbb{R}$ satisfy for all $\theta = (\theta_1, \theta_2) \in \mathbb{R}^2$ that

$$\mathcal{L}(\theta) = \left(\frac{\kappa}{2}\right)|\theta_1 - \vartheta_1|^2 + \left(\frac{\mathcal{K}}{2}\right)|\theta_2 - \vartheta_2|^2, \quad (6.385)$$

let $\Theta: \mathbb{N}_0 \rightarrow \mathbb{R}^d$ satisfy for all $n \in \mathbb{N}$ that $\Theta_0 = \xi$ and

$$\begin{aligned} \Theta_n &= \Theta_{n-1} - \frac{2}{\kappa+\mathcal{K}}(\nabla \mathcal{L})(\Theta_{n-1}) = \Theta_{n-1} - \frac{2}{11}(\nabla \mathcal{L})(\Theta_{n-1}) \\ &= \Theta_{n-1} - 0.1\overline{8}(\nabla \mathcal{L})(\Theta_{n-1}) \approx \Theta_{n-1} - 0.18(\nabla \mathcal{L})(\Theta_{n-1}), \end{aligned} \quad (6.386)$$

and let $\mathcal{M}: \mathbb{N}_0 \rightarrow \mathbb{R}^d$ and $\mathbf{m}: \mathbb{N}_0 \rightarrow \mathbb{R}^d$ 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

$$\begin{aligned} \mathbf{m}_n &= 0.5 \mathbf{m}_{n-1} + (1 - 0.5)(\nabla \mathcal{L})(\mathcal{M}_{n-1}) \\ &= 0.5 (\mathbf{m}_{n-1} + (\nabla \mathcal{L})(\mathcal{M}_{n-1})). \end{aligned} \quad (6.387)$$

Then

(i) it holds for all $\theta = (\theta_1, \theta_2) \in \mathbb{R}^2$ that

$$(\nabla \mathcal{L})(\theta) = \begin{pmatrix} \kappa(\theta_1 - \vartheta_1) \\ \mathcal{K}(\theta_2 - \vartheta_2) \end{pmatrix} = \begin{pmatrix} \theta_1 - 1 \\ 10(\theta_2 - 1) \end{pmatrix}, \quad (6.388)$$

(ii) it holds that

$$\Theta_0 = \begin{pmatrix} 5 \\ 3 \end{pmatrix}, \quad (6.389)$$

$$\begin{aligned} \Theta_1 &= \Theta_0 - \frac{2}{11}(\nabla \mathcal{L})(\Theta_0) \approx \Theta_0 - 0.18(\nabla \mathcal{L})(\Theta_0) \\ &= \begin{pmatrix} 5 \\ 3 \end{pmatrix} - 0.18 \begin{pmatrix} 5 - 1 \\ 10(3 - 1) \end{pmatrix} = \begin{pmatrix} 5 - 0.18 \cdot 4 \\ 3 - 0.18 \cdot 10 \cdot 2 \end{pmatrix} \\ &= \begin{pmatrix} 5 - 0.72 \\ 3 - 3.6 \end{pmatrix} = \begin{pmatrix} 4.28 \\ -0.6 \end{pmatrix}, \end{aligned} \quad (6.390)$$

$$\begin{aligned}
 \Theta_2 &\approx \Theta_1 - 0.18(\nabla \mathcal{L})(\Theta_1) = \begin{pmatrix} 4.28 \\ -0.6 \end{pmatrix} - 0.18 \begin{pmatrix} 4.28 - 1 \\ 10(-0.6 - 1) \end{pmatrix} \\
 &= \begin{pmatrix} 4.28 - 0.18 \cdot 3.28 \\ -0.6 - 0.18 \cdot 10 \cdot (-1.6) \end{pmatrix} = \begin{pmatrix} 4.10 - 0.18 \cdot 2 - 0.18 \cdot 0.28 \\ -0.6 + 1.8 \cdot 1.6 \end{pmatrix} \\
 &= \begin{pmatrix} 4.10 - 0.36 - 2 \cdot 9 \cdot 4 \cdot 7 \cdot 10^{-4} \\ -0.6 + 1.6 \cdot 1.6 + 0.2 \cdot 1.6 \end{pmatrix} = \begin{pmatrix} 3.74 - 9 \cdot 56 \cdot 10^{-4} \\ -0.6 + 2.56 + 0.32 \end{pmatrix} \\
 &= \begin{pmatrix} 3.74 - 504 \cdot 10^{-4} \\ 2.88 - 0.6 \end{pmatrix} = \begin{pmatrix} 3.6896 \\ 2.28 \end{pmatrix} \approx \begin{pmatrix} 3.69 \\ 2.28 \end{pmatrix},
 \end{aligned} \tag{6.391}$$

$$\begin{aligned}
 \Theta_3 &\approx \Theta_2 - 0.18(\nabla \mathcal{L})(\Theta_2) \approx \begin{pmatrix} 3.69 \\ 2.28 \end{pmatrix} - 0.18 \begin{pmatrix} 3.69 - 1 \\ 10(2.28 - 1) \end{pmatrix} \\
 &= \begin{pmatrix} 3.69 - 0.18 \cdot 2.69 \\ 2.28 - 0.18 \cdot 10 \cdot 1.28 \end{pmatrix} = \begin{pmatrix} 3.69 - 0.2 \cdot 2.69 + 0.02 \cdot 2.69 \\ 2.28 - 1.8 \cdot 1.28 \end{pmatrix} \\
 &= \begin{pmatrix} 3.69 - 0.538 + 0.0538 \\ 2.28 - 1.28 - 0.8 \cdot 1.28 \end{pmatrix} = \begin{pmatrix} 3.7438 - 0.538 \\ 1 - 1.28 + 0.2 \cdot 1.28 \end{pmatrix} \\
 &= \begin{pmatrix} 3.2058 \\ 0.256 - 0.280 \end{pmatrix} = \begin{pmatrix} 3.2058 \\ -0.024 \end{pmatrix} \approx \begin{pmatrix} 3.21 \\ -0.02 \end{pmatrix},
 \end{aligned} \tag{6.392}$$

⋮

and

(iii) it holds that

$$\mathcal{M}_0 = \begin{pmatrix} 5 \\ 3 \end{pmatrix}, \tag{6.393}$$

$$\begin{aligned}
 \mathbf{m}_1 &= 0.5 (\mathbf{m}_0 + (\nabla \mathcal{L})(\mathcal{M}_0)) = 0.5 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 5 - 1 \\ 10(3 - 1) \end{pmatrix} \right) \\
 &= \begin{pmatrix} 0.5(0 + 4) \\ 0.5(0 + 10 \cdot 2) \end{pmatrix} = \begin{pmatrix} 2 \\ 10 \end{pmatrix},
 \end{aligned} \tag{6.394}$$

$$\mathcal{M}_1 = \mathcal{M}_0 - 0.3 \mathbf{m}_1 = \begin{pmatrix} 5 \\ 3 \end{pmatrix} - 0.3 \begin{pmatrix} 2 \\ 10 \end{pmatrix} = \begin{pmatrix} 4.4 \\ 0 \end{pmatrix}, \tag{6.395}$$

$$\begin{aligned}\mathbf{m}_2 &= 0.5 (\mathbf{m}_1 + (\nabla \mathcal{L})(\mathcal{M}_1)) = 0.5 \left(\begin{pmatrix} 2 \\ 10 \end{pmatrix} + \begin{pmatrix} 4.4 - 1 \\ 10(0 - 1) \end{pmatrix} \right) \\ &= \begin{pmatrix} 0.5(2 + 3.4) \\ 0.5(10 - 10) \end{pmatrix} = \begin{pmatrix} 2.7 \\ 0 \end{pmatrix},\end{aligned}\tag{6.396}$$

$$\mathcal{M}_2 = \mathcal{M}_1 - 0.3 \mathbf{m}_2 = \begin{pmatrix} 4.4 \\ 0 \end{pmatrix} - 0.3 \begin{pmatrix} 2.7 \\ 0 \end{pmatrix} = \begin{pmatrix} 4.4 - 0.81 \\ 0 \end{pmatrix} = \begin{pmatrix} 3.59 \\ 0 \end{pmatrix},\tag{6.397}$$

$$\begin{aligned}\mathbf{m}_3 &= 0.5 (\mathbf{m}_2 + (\nabla \mathcal{L})(\mathcal{M}_2)) = 0.5 \left(\begin{pmatrix} 2.7 \\ 0 \end{pmatrix} + \begin{pmatrix} 3.59 - 1 \\ 10(0 - 1) \end{pmatrix} \right) \\ &= \begin{pmatrix} 0.5(2.7 + 2.59) \\ 0.5(0 - 10) \end{pmatrix} = \begin{pmatrix} 0.5 \cdot 5.29 \\ 0.5(-10) \end{pmatrix} \\ &= \begin{pmatrix} 2.5 + 0.145 \\ -5 \end{pmatrix} = \begin{pmatrix} 2.645 \\ -5 \end{pmatrix} \approx \begin{pmatrix} 2.65 \\ -5 \end{pmatrix},\end{aligned}\tag{6.398}$$

$$\begin{aligned}\mathcal{M}_3 &= \mathcal{M}_2 - 0.3 \mathbf{m}_3 \approx \begin{pmatrix} 3.59 \\ 0 \end{pmatrix} - 0.3 \begin{pmatrix} 2.65 \\ -5 \end{pmatrix} \\ &= \begin{pmatrix} 3.59 - 0.795 \\ 1.5 \end{pmatrix} = \begin{pmatrix} 3 - 0.205 \\ 1.5 \end{pmatrix} = \begin{pmatrix} 2.795 \\ 1.5 \end{pmatrix} \approx \begin{pmatrix} 2.8 \\ 1.5 \end{pmatrix},\end{aligned}\tag{6.399}$$

⋮

```

1 # Example for GD and momentum GD
2
3 import numpy as np
4 import matplotlib.pyplot as plt
5
6 # Number of steps for the schemes
7 N = 8
8
9 # Problem setting
10 d = 2
11 K = [1., 10.]
12
13 vartheta = np.array([1., 1.])
14 xi = np.array([5., 3.])
15

```

```

16 def f(x, y):
17     result = K[0] / 2. * np.abs(x - vartheta[0])**2 \
18     + K[1] / 2. * np.abs(y - vartheta[1])**2
19     return result
20
21 def nabla_f(x):
22     return K * (x - vartheta)
23
24 # Coefficients for GD
25 gamma_GD = 2 / (K[0] + K[1])
26
27 # Coefficients for momentum
28 gamma_momentum = 0.3
29 alpha = 0.5
30
31 # Placeholder for processes
32 Theta = np.zeros((N+1, d))
33 M = np.zeros((N+1, d))
34 m = np.zeros((N+1, d))
35
36 Theta[0] = xi
37 M[0] = xi
38
39 # Perform gradient descent
40 for i in range(N):
41     Theta[i+1] = Theta[i] - gamma_GD * nabla_f(Theta[i])
42
43 # Perform momentum GD
44 for i in range(N):
45     m[i+1] = alpha * m[i] + (1 - alpha) * nabla_f(M[i])
46     M[i+1] = M[i] - gamma_momentum * m[i+1]
47
48
49 ### Plot ###
50 plt.figure()
51
52 # Plot the gradient descent process
53 plt.plot(Theta[:, 0], Theta[:, 1],
54           label = "GD", color = "c",
55           linestyle = "--", marker = "*")
56
57 # Plot the momentum gradient descent process
58 plt.plot(M[:, 0], M[:, 1],
59           label = "Momentum", color = "orange", marker = "*")
60
61 # Target value
62 plt.scatter(vartheta[0], vartheta[1],
63             label = "vartheta", color = "red", marker = "x")
64

```

```

65 # Plot contour lines of f
66 x = np.linspace(-3., 7., 100)
67 y = np.linspace(-2., 4., 100)
68 X, Y = np.meshgrid(x, y)
69 Z = f(X, Y)
70 cp = plt.contour(X, Y, Z, colors="black",
71                   levels = [0.5,2,4,8,16],
72                   linestyles=":")
73
74 plt.legend()
75 plt.savefig("../plots/GD_momentum_plots.pdf")

```

Source code 6.1 ([code/example_GD_momentum_plots.py](#)): PYTHON code for Figure 6.1

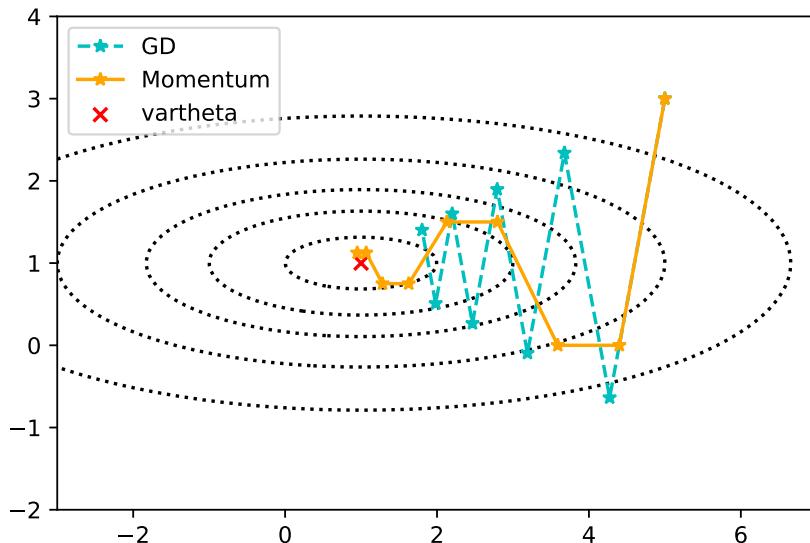


Figure 6.1 ([plots/GD_momentum_plots.pdf](#)): Result of a call of PYTHON code 6.1

Exercise 6.3.3. Let $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, 1]$ satisfy for all $n \in \mathbb{N}$ that $\gamma_n = \frac{1}{n}$ and $\alpha_n = \frac{1}{2}$, let $\mathcal{L}: \mathbb{R} \rightarrow \mathbb{R}$ satisfy for all $\theta \in \mathbb{R}$ that $\mathcal{L}(\theta) = \theta^2$, and let Θ be the momentum GD process for the objective function \mathcal{L} with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, and initial value 1 (cf. Definition 6.3.1). Specify Θ_1 , Θ_2 , Θ_3 , and Θ_4 explicitly and prove that your results are correct!

6.4 Nesterov accelerated momentum optimization

In this section we review the Nesterov accelerated GD optimization method, which was first introduced in Nesterov [321] (cf., for instance, Sutskever et al. [408]). The Nesterov accelerated GD optimization method can be viewed as building on the momentum GD optimization method (see Definition 6.3.1) by attempting to provide some kind of foresight to the scheme. A similar perspective is to see the Nesterov accelerated GD optimization method as a combination of the momentum GD optimization method (see Definition 6.3.1) and the explicit midpoint GD optimization method (see Definition 6.2.1).

Definition 6.4.1 (Nesterov accelerated GD optimization method). *Let $\mathfrak{d} \in \mathbb{N}$, let $\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$ 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 $\Theta: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a function. Then we say that Θ is the Nesterov accelerated GD process for the objective function \mathcal{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 Nesterov accelerated GD process (1st version) for the objective function \mathcal{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 $\mathbf{m}: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ such that for all $n \in \mathbb{N}$ it holds that*

$$\Theta_0 = \xi, \quad \mathbf{m}_0 = 0, \quad (6.400)$$

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + (1 - \alpha_n)(\nabla \mathcal{L})(\Theta_{n-1} - \gamma_n \alpha_n \mathbf{m}_{n-1}), \quad (6.401)$$

$$\text{and} \quad \Theta_n = \Theta_{n-1} - \gamma_n \mathbf{m}_n. \quad (6.402)$$

Algorithm 6.4.2: Nesterov accelerated GD optimization method

Input: $\mathfrak{d}, N \in \mathbb{N}$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $\xi \in \mathbb{R}^{\mathfrak{d}}$

Output: N -th step of the Nesterov accelerated GD process for the objective function \mathcal{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 6.4.1)

- 1: **Initialization:** $\Theta \leftarrow \xi; \mathbf{m} \leftarrow 0 \in \mathbb{R}^{\mathfrak{d}}$
- 2: **for** $n = 1, \dots, N$ **do**
- 3: $\mathbf{m} \leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n)(\nabla \mathcal{L})(\Theta - \gamma_n \alpha_n \mathbf{m})$
- 4: $\Theta \leftarrow \Theta - \gamma_n \mathbf{m}$
- 5: **return** Θ

6.4.1 Alternative definitions

In analogy to the alternative definitions of the momentum GD optimization method in Section 6.3.1 above, we now provide alternative definitions of the Nesterov accelerated

GD optimization method. Relationships between the considered versions of the Nesterov accelerated **GD** optimization method are discussed in Section 6.4.2 below.

Definition 6.4.3 (Nesterov accelerated **GD** optimization method (2nd version)). *Let $\mathfrak{d} \in \mathbb{N}$, let $\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$ 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 $\Theta: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a function. Then we say that Θ is the Nesterov accelerated **GD** process (2nd version) for the objective function \mathcal{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 $\mathbf{m}: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ such that for all $n \in \mathbb{N}$ it holds that*

$$\Theta_0 = \xi, \quad \mathbf{m}_0 = 0, \quad (6.403)$$

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + (\nabla \mathcal{L})(\Theta_{n-1} - \gamma_n \alpha_n \mathbf{m}_{n-1}), \quad (6.404)$$

$$\text{and} \quad \Theta_n = \Theta_{n-1} - \gamma_n \mathbf{m}_n. \quad (6.405)$$

Algorithm 6.4.4: Nesterov accelerated **GD optimization method (2nd version)**

Input: $\mathfrak{d}, N \in \mathbb{N}$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$

Output: N -th step of the Nesterov accelerated **GD** process (2nd version) for the objective function \mathcal{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 6.4.3)

- 1: **Initialization:** $\Theta \leftarrow \xi; \mathbf{m} \leftarrow 0 \in \mathbb{R}^{\mathfrak{d}}$
- 2: **for** $n = 1, \dots, N$ **do**
- 3: $\mathbf{m} \leftarrow \alpha_n \mathbf{m} + (\nabla \mathcal{L})(\Theta - \gamma_n \alpha_n \mathbf{m})$
- 4: $\Theta \leftarrow \Theta - \gamma_n \mathbf{m}$
- 5: **return** Θ

Definition 6.4.5 (Nesterov accelerated **GD** optimization method (3rd version)). *Let $\mathfrak{d} \in \mathbb{N}$, let $\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$ 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 $\Theta: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a function. Then we say that Θ is the Nesterov accelerated **GD** process (3rd version) for the objective function \mathcal{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 $\mathbf{m}: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ such that for all $n \in \mathbb{N}$ it holds that*

$$\Theta_0 = \xi, \quad \mathbf{m}_0 = 0, \quad (6.406)$$

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + (1 - \alpha_n) \gamma_n (\nabla \mathcal{L})(\Theta_{n-1} - \alpha_n \mathbf{m}_{n-1}), \quad (6.407)$$

$$and \quad \Theta_n = \Theta_{n-1} - \mathbf{m}_n. \quad (6.408)$$

Algorithm 6.4.6: Nesterov accelerated GD optimization method (3rd version)

Input: $\mathfrak{d}, N \in \mathbb{N}$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $\xi \in \mathbb{R}^{\mathfrak{d}}$
Output: N -th step of the Nesterov accelerated GD process (3rd version) for the objective function \mathcal{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 6.4.5)

```

1: Initialization:  $\Theta \leftarrow \xi; \mathbf{m} \leftarrow 0 \in \mathbb{R}^{\mathfrak{d}}$ 
2: for  $n = 1, \dots, N$  do
3:    $\mathbf{m} \leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n) \gamma_n (\nabla \mathcal{L})(\Theta - \alpha_n \mathbf{m})$ 
4:    $\Theta \leftarrow \Theta - \mathbf{m}$ 
5: return  $\Theta$ 
```

Definition 6.4.7 (Nesterov accelerated GD optimization method (4th version)). Let $\mathfrak{d} \in \mathbb{N}$, let $\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$ 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 $\Theta: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a function. Then we say that Θ is the Nesterov accelerated GD process (4th version) for the objective function \mathcal{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 $\mathbf{m}: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ such that for all $n \in \mathbb{N}$ it holds that

$$\Theta_0 = \xi, \quad \mathbf{m}_0 = 0, \quad (6.409)$$

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + \gamma_n (\nabla \mathcal{L})(\Theta_{n-1} - \alpha_n \mathbf{m}_{n-1}), \quad (6.410)$$

$$and \quad \Theta_n = \Theta_{n-1} - \mathbf{m}_n. \quad (6.411)$$

Algorithm 6.4.8: Nesterov accelerated GD optimization method (4th version)

Input: $\mathfrak{d}, N \in \mathbb{N}$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$
Output: N -th step of the Nesterov accelerated GD process (4th version) for the objective function \mathcal{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 6.4.7)

```

1: Initialization:  $\Theta \leftarrow \xi; \mathbf{m} \leftarrow 0 \in \mathbb{R}^{\mathfrak{d}}$ 
2: for  $n = 1, \dots, N$  do
3:    $\mathbf{m} \leftarrow \alpha_n \mathbf{m} + \gamma_n (\nabla \mathcal{L})(\Theta - \alpha_n \mathbf{m})$ 
4:    $\Theta \leftarrow \Theta - \mathbf{m}$ 
```

5: **return** Θ

6.4.2 Relationships between different definitions

In this section we discuss relationships between the different versions of the Nesterov accelerated GD optimization method introduced in Definitions 6.4.1, 6.4.3, 6.4.5, and 6.4.7 above.

Proposition 6.4.9 (Comparison of general Nesterov-type GD optimization methods). *Let $\mathfrak{d} \in \mathbb{N}$, $(\mathfrak{a}_n^{(1)})_{n \in \mathbb{N}} \subseteq (0, \infty)$, $(\mathfrak{a}_n^{(2)})_{n \in \mathbb{N}} \subseteq (0, \infty)$, $(\mathfrak{b}_n^{(1)})_{n \in \mathbb{N}} \subseteq (0, \infty)$, $(\mathfrak{b}_n^{(2)})_{n \in \mathbb{N}} \subseteq (0, \infty)$, $(\mathfrak{c}_n^{(1)})_{n \in \mathbb{N}} \subseteq (0, \infty)$, $(\mathfrak{c}_n^{(2)})_{n \in \mathbb{N}} \subseteq (0, \infty)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ satisfy for all $n \in \mathbb{N}$ that*

$$\mathfrak{b}_n^{(1)} \mathfrak{c}_n^{(1)} = \mathfrak{b}_n^{(2)} \mathfrak{c}_n^{(2)} \quad \text{and} \quad \frac{\mathfrak{a}_{n+1}^{(1)} \mathfrak{b}_n^{(1)}}{\mathfrak{b}_{n+1}^{(1)}} = \frac{\mathfrak{a}_{n+1}^{(2)} \mathfrak{b}_n^{(2)}}{\mathfrak{b}_{n+1}^{(2)}}, \quad (6.412)$$

and for every $i \in \{1, 2\}$ let $\Theta^{(i)}: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ and $\mathbf{m}^{(i)}: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $n \in \mathbb{N}$ that

$$\Theta_0^{(i)} = \xi, \quad \mathbf{m}_0^{(i)} = 0, \quad (6.413)$$

$$\mathbf{m}_n^{(i)} = \mathfrak{a}_n^{(i)} \mathbf{m}_{n-1}^{(i)} + \mathfrak{b}_n^{(i)} (\nabla \mathcal{L})(\Theta_{n-1}^{(i)} - \mathfrak{c}_n^{(i)} \mathfrak{a}_n^{(i)} \mathbf{m}_{n-1}^{(i)}), \quad (6.414)$$

$$\text{and} \quad \Theta_n^{(i)} = \Theta_{n-1}^{(i)} - \mathfrak{c}_n^{(i)} \mathbf{m}_n^{(i)}. \quad (6.415)$$

Then

$$\Theta^{(1)} = \Theta^{(2)}. \quad (6.416)$$

Proof of Proposition 6.4.9. Throughout this proof, let $\mathcal{G}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that

$$\mathcal{G}(\theta) = (\nabla \mathcal{L})(\theta). \quad (6.417)$$

Observe that the fact that for all $n \in \mathbb{N}$ it holds that

$$\mathfrak{c}_{n+1}^{(1)} = \frac{\mathfrak{c}_{n+1}^{(2)} \mathfrak{b}_{n+1}^{(2)}}{\mathfrak{b}_{n+1}^{(1)}}, \quad \frac{\mathfrak{c}_n^{(2)}}{\mathfrak{c}_n^{(1)}} = \frac{\mathfrak{b}_{n+1}^{(1)}}{\mathfrak{b}_{n+1}^{(2)}}, \quad \text{and} \quad \frac{\mathfrak{b}_{n+1}^{(2)} \mathfrak{a}_{n+1}^{(1)} \mathfrak{b}_n^{(1)}}{\mathfrak{b}_{n+1}^{(1)} \mathfrak{b}_n^{(2)}} = \mathfrak{a}_{n+1}^{(2)} \quad (6.418)$$

proves that for all $n \in \mathbb{N}$ it holds that

$$\frac{\mathfrak{c}_{n+1}^{(1)} \mathfrak{a}_{n+1}^{(1)} \mathfrak{c}_n^{(2)}}{\mathfrak{c}_n^{(1)}} = \frac{\mathfrak{c}_{n+1}^{(2)} \mathfrak{b}_{n+1}^{(2)} \mathfrak{a}_n^{(1)} \mathfrak{b}_n^{(1)}}{\mathfrak{b}_{n+1}^{(1)} \mathfrak{b}_n^{(2)}} = \mathfrak{c}_{n+1}^{(2)} \mathfrak{a}_{n+1}^{(2)}. \quad (6.419)$$

Furthermore, note that (6.413) establishes that

$$\mathbf{m}_0^{(1)} = 0 = \mathbf{m}_0^{(2)} \quad \text{and} \quad \Theta_0^{(1)} = \xi = \Theta_0^{(2)}. \quad (6.420)$$

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

$$\mathfrak{c}_n^{(1)} \mathbf{m}_n^{(1)} = \mathfrak{c}_n^{(2)} \mathbf{m}_n^{(2)} \quad \text{and} \quad \Theta_n^{(1)} = \Theta_n^{(2)}. \quad (6.421)$$

We now prove (6.421) by induction on $n \in \mathbb{N}$. For the base case $n = 1$ observe that (6.412), (6.413), and (6.420) ensure that

$$\begin{aligned} \mathfrak{c}_1^{(1)} \mathbf{m}_1^{(1)} &= \mathfrak{c}_1^{(1)} (\mathfrak{a}_1^{(1)} \mathbf{m}_0^{(1)} + \mathfrak{b}_1^{(1)} \mathcal{G}(\Theta_0^{(1)} - \mathfrak{c}_1^{(1)} \mathfrak{a}_1^{(1)} \mathbf{m}_0^{(1)})) \\ &= \mathfrak{c}_1^{(1)} \mathfrak{b}_1^{(1)} \mathcal{G}(\Theta_0^{(1)}) \\ &= \mathfrak{c}_1^{(2)} \mathfrak{b}_1^{(2)} \mathcal{G}(\Theta_0^{(2)}) \\ &= \mathfrak{c}_1^{(2)} (\mathfrak{a}_1^{(2)} \mathbf{m}_0^{(2)} + \mathfrak{b}_1^{(2)} \mathcal{G}(\Theta_0^{(2)} - \mathfrak{c}_1^{(2)} \mathfrak{a}_1^{(2)} \mathbf{m}_0^{(2)})) \\ &= \mathfrak{c}_1^{(2)} \mathbf{m}_1^{(2)}. \end{aligned} \quad (6.422)$$

This, (6.415), and (6.420) imply that

$$\Theta_1^{(1)} = \Theta_0^{(1)} - \mathfrak{c}_1^{(1)} \mathbf{m}_1^{(1)} = \Theta_0^{(2)} - \mathfrak{c}_1^{(2)} \mathbf{m}_1^{(2)} = \Theta_1^{(2)}. \quad (6.423)$$

Combining this and (6.422) establishes (6.421) in the base case $n = 1$. For the induction step $\mathbb{N} \ni n \rightarrow n+1 \in \{2, 3, \dots\}$ let $n \in \mathbb{N}$ and assume that

$$\mathfrak{c}_n^{(1)} \mathbf{m}_n^{(1)} = \mathfrak{c}_n^{(2)} \mathbf{m}_n^{(2)} \quad \text{and} \quad \Theta_n^{(1)} = \Theta_n^{(2)}. \quad (6.424)$$

Note that (6.412), (6.414), (6.419), and (6.424) show that

$$\begin{aligned} \mathfrak{c}_{n+1}^{(1)} \mathbf{m}_{n+1}^{(1)} &= \mathfrak{c}_{n+1}^{(1)} (\mathfrak{a}_{n+1}^{(1)} \mathbf{m}_n^{(1)} + \mathfrak{b}_{n+1}^{(1)} \mathcal{G}(\Theta_n^{(1)} - \mathfrak{c}_{n+1}^{(1)} \mathfrak{a}_{n+1}^{(1)} \mathbf{m}_n^{(1)})) \\ &= \frac{\mathfrak{c}_{n+1}^{(1)} \mathfrak{a}_{n+1}^{(1)} \mathfrak{c}_n^{(2)}}{\mathfrak{c}_n^{(1)}} \mathbf{m}_n^{(2)} + \mathfrak{c}_{n+1}^{(1)} \mathfrak{b}_{n+1}^{(1)} \mathcal{G}\left(\Theta_n^{(2)} - \frac{\mathfrak{c}_{n+1}^{(1)} \mathfrak{a}_{n+1}^{(1)} \mathfrak{c}_n^{(2)}}{\mathfrak{c}_n^{(1)}} \mathbf{m}_n^{(2)}\right) \\ &= \mathfrak{c}_{n+1}^{(2)} \mathfrak{a}_{n+1}^{(2)} \mathbf{m}_n^{(2)} + \mathfrak{c}_{n+1}^{(2)} \mathfrak{b}_{n+1}^{(2)} \mathcal{G}(\Theta_n^{(2)} - \mathfrak{c}_{n+1}^{(2)} \mathfrak{a}_{n+1}^{(2)} \mathbf{m}_n^{(2)}) \\ &= \mathfrak{c}_{n+1}^{(2)} (\mathfrak{a}_{n+1}^{(2)} \mathbf{m}_n^{(2)} + \mathfrak{b}_{n+1}^{(2)} \mathcal{G}(\Theta_n^{(2)} - \mathfrak{c}_{n+1}^{(2)} \mathfrak{a}_{n+1}^{(2)} \mathbf{m}_n^{(2)})) \\ &= \mathfrak{c}_{n+1}^{(2)} \mathbf{m}_{n+1}^{(2)}. \end{aligned} \quad (6.425)$$

This, (6.415), and (6.424) demonstrate that

$$\Theta_{n+1}^{(1)} = \Theta_n^{(1)} - \mathfrak{c}_{n+1}^{(1)} \mathbf{m}_{n+1}^{(1)} = \Theta_n^{(2)} - \mathfrak{c}_{n+1}^{(2)} \mathbf{m}_{n+1}^{(2)} = \Theta_{n+1}^{(2)}. \quad (6.426)$$

Induction thus proves (6.421). Combining (6.420) and (6.421) establishes (6.416). The proof of Proposition 6.4.9 is thus complete. \square

Corollary 6.4.10 (Comparison of the 1st and 2nd version of the momentum **GD** optimization method). *Let $\mathfrak{d} \in \mathbb{N}$, $(\gamma_n^{(1)})_{n \in \mathbb{N}} \subseteq (0, \infty)$, $(\gamma_n^{(2)})_{n \in \mathbb{N}} \subseteq (0, \infty)$, $(\alpha_n^{(1)})_{n \in \mathbb{N}} \subseteq (0, 1)$, $(\alpha_n^{(2)})_{n \in \mathbb{N}} \subseteq (0, \infty)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ satisfy for all $n \in \mathbb{N}$ that*

$$\gamma_n^{(1)}(1 - \alpha_n^{(1)}) = \gamma_n^{(2)} \quad \text{and} \quad \frac{\alpha_{n+1}^{(1)}(1 - \alpha_n^{(1)})}{1 - \alpha_{n+1}^{(1)}} = \alpha_{n+1}^{(2)}, \quad (6.427)$$

for every $i \in \{1, 2\}$ let $\Theta^{(i)}: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ be the Nesterov accelerated **GD** process (i^{th} version) for the objective function \mathcal{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 6.4.1 and 6.4.3). Then

$$\Theta^{(1)} = \Theta^{(2)}. \quad (6.428)$$

Proof of Corollary 6.4.10. 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

$$\mathfrak{a}_n^{(1)} = \alpha_n^{(1)}, \quad \mathfrak{b}_n^{(1)} = 1 - \alpha_n^{(1)}, \quad \mathfrak{c}_n^{(1)} = \gamma_n^{(1)}, \quad (6.429)$$

$$\mathfrak{a}_n^{(2)} = \alpha_n^{(2)}, \quad \mathfrak{b}_n^{(2)} = 1, \quad \text{and} \quad \mathfrak{c}_n^{(2)} = \gamma_n^{(2)}. \quad (6.430)$$

Observe that (6.400), (6.401), (6.402), (6.403), (6.404), and (6.405) prove that for all $i \in \{1, 2\}$, $n \in \mathbb{N}$ it holds that

$$\Theta_0^{(i)} = \xi, \quad \mathbf{m}_0^{(i)} = 0, \quad (6.431)$$

$$\mathbf{m}_n^{(i)} = \mathfrak{a}_n^{(i)} \mathbf{m}_{n-1}^{(i)} + \mathfrak{b}_n^{(i)} (\nabla \mathcal{L})(\Theta_{n-1}^{(i)} - \mathfrak{c}_n^{(i)} \mathfrak{a}_n^{(i)} \mathbf{m}_{n-1}^{(i)}), \quad (6.432)$$

$$\text{and} \quad \Theta_n^{(i)} = \Theta_{n-1}^{(i)} - \mathfrak{c}_n^{(i)} \mathbf{m}_n^{(i)}. \quad (6.433)$$

Furthermore, note that (6.427), (6.429), and (6.430) establishes that for all $n \in \mathbb{N}$ it holds that

$$\mathfrak{b}_n^{(1)} \mathfrak{c}_n^{(1)} = (1 - \alpha_n^{(1)}) \gamma_n^{(1)} = \gamma_n^{(2)} = \mathfrak{b}_n^{(2)} \mathfrak{c}_n^{(2)}. \quad (6.434)$$

Moreover, observe that (6.427), (6.429), and (6.430) ensures that for all $n \in \mathbb{N}$ it holds that

$$\frac{\mathfrak{a}_{n+1}^{(1)} \mathfrak{b}_n^{(1)}}{\mathfrak{b}_{n+1}^{(1)}} = \frac{\alpha_{n+1}^{(1)}(1 - \alpha_n^{(1)})}{1 - \alpha_{n+1}^{(1)}} = \alpha_{n+1}^{(2)} = \frac{\mathfrak{a}_{n+1}^{(2)} \mathfrak{b}_n^{(2)}}{\mathfrak{b}_{n+1}^{(2)}}. \quad (6.435)$$

Combining this, (6.431), (6.432), (6.433), and (6.434) with Proposition 6.4.9 implies (6.428). The proof of Corollary 6.4.10 is thus complete. \square

Lemma 6.4.11 (Comparison of the 1st and 3rd version of the momentum GD optimization method). Let $\mathfrak{d} \in \mathbb{N}$, $(\gamma_n^{(1)})_{n \in \mathbb{N}} \subseteq (0, \infty)$, $(\gamma_n^{(3)})_{n \in \mathbb{N}} \subseteq (0, \infty)$, $(\alpha_n^{(1)})_{n \in \mathbb{N}} \subseteq (0, 1)$, $(\alpha_n^{(3)})_{n \in \mathbb{N}} \subseteq (0, 1)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ satisfy for all $n \in \mathbb{N}$ that

$$\gamma_n^{(1)}(1 - \alpha_n^{(1)}) = \gamma_n^{(3)}(1 - \alpha_n^{(3)}) \quad \text{and} \quad \frac{\gamma_{n+1}^{(1)} \alpha_{n+1}^{(1)}}{\gamma_n^{(1)}} = \alpha_{n+1}^{(3)}, \quad (6.436)$$

for every $i \in \{1, 3\}$ let $\Theta^{(i)}: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ be the Nesterov accelerated GD process (i^{th} version) for the objective function \mathcal{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 6.4.1 and 6.4.7). Then

$$\Theta^{(1)} = \Theta^{(3)}. \quad (6.437)$$

Proof of Lemma 6.4.11. 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

$$\mathfrak{a}_n^{(1)} = \alpha_n^{(1)}, \quad \mathfrak{b}_n^{(1)} = 1 - \alpha_n^{(1)}, \quad \mathfrak{c}_n^{(1)} = \gamma_n^{(1)}, \quad (6.438)$$

$$\mathfrak{a}_n^{(3)} = \alpha_n^{(3)}, \quad \mathfrak{b}_n^{(3)} = (1 - \alpha_n^{(3)})\gamma_n^{(3)}, \quad \text{and} \quad \mathfrak{c}_n^{(3)} = 1. \quad (6.439)$$

Note that (6.400), (6.401), (6.402), (6.406), (6.407), and (6.408) show that for all $i \in \{1, 3\}$, $n \in \mathbb{N}$ it holds that

$$\Theta_0^{(i)} = \xi, \quad \mathbf{m}_0^{(i)} = 0, \quad (6.440)$$

$$\mathbf{m}_n^{(i)} = \mathfrak{a}_n^{(i)} \mathbf{m}_{n-1}^{(i)} + \mathfrak{b}_n^{(i)} (\nabla \mathcal{L})(\Theta_{n-1}^{(i)} - \mathfrak{c}_n^{(i)} \mathfrak{a}_n^{(i)} \mathbf{m}_{n-1}^{(i)}), \quad (6.441)$$

$$\text{and} \quad \Theta_n^{(i)} = \Theta_{n-1}^{(i)} - \mathfrak{c}_n^{(i)} \mathbf{m}_n^{(i)}. \quad (6.442)$$

Furthermore, observe that (6.436), (6.438), and (6.439) demonstrates that for all $n \in \mathbb{N}$ it holds that

$$\mathfrak{b}_n^{(1)} \mathfrak{c}_n^{(1)} = (1 - \alpha_n^{(1)})\gamma_n^{(1)} = (1 - \alpha_n^{(3)})\gamma_n^{(3)} = \mathfrak{b}_n^{(3)} \mathfrak{c}_n^{(3)}. \quad (6.443)$$

Moreover, note that (6.436), (6.438), and (6.439) proves that for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned} \frac{\mathfrak{a}_{n+1}^{(1)} \mathfrak{b}_n^{(1)}}{\mathfrak{b}_{n+1}^{(1)}} &= \frac{\alpha_{n+1}^{(1)}(1 - \alpha_n^{(1)})}{1 - \alpha_{n+1}^{(1)}} = \frac{\alpha_{n+1}^{(1)} \gamma_n^{(3)}(1 - \alpha_n^{(3)}) \gamma_{n+1}^{(1)}}{\gamma_n^{(1)} \gamma_{n+1}^{(3)}(1 - \alpha_{n+1}^{(3)})} \\ &= \frac{\alpha_{n+1}^{(3)} \gamma_n^{(3)}(1 - \alpha_n^{(3)})}{\gamma_{n+1}^{(3)}(1 - \alpha_{n+1}^{(3)})} = \frac{\mathfrak{a}_{n+1}^{(3)} \mathfrak{b}_n^{(3)}}{\mathfrak{b}_{n+1}^{(3)}}. \end{aligned} \quad (6.444)$$

Combining this, (6.440), (6.441), (6.442), and (6.443) with Proposition 6.4.9 establishes (6.437). The proof of Lemma 6.4.11 is thus complete. \square

Lemma 6.4.12 (Comparison of the 1st and 4th version of the momentum **GD** optimization method). Let $\mathfrak{d} \in \mathbb{N}$, $(\gamma_n^{(1)})_{n \in \mathbb{N}} \subseteq (0, \infty)$, $(\gamma_n^{(4)})_{n \in \mathbb{N}} \subseteq (0, \infty)$, $(\alpha_n^{(1)})_{n \in \mathbb{N}} \subseteq (0, 1)$, $(\alpha_n^{(4)})_{n \in \mathbb{N}} \subseteq (0, \infty)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ satisfy for all $n \in \mathbb{N}$ that

$$\gamma_n^{(1)}(1 - \alpha_n^{(1)}) = \gamma_n^{(4)} \quad \text{and} \quad \frac{\gamma_{n+1}^{(1)} \alpha_{n+1}^{(1)}}{\gamma_n^{(1)}} = \alpha_{n+1}^{(4)}, \quad (6.445)$$

for every $i \in \{1, 4\}$ let $\Theta^{(i)}: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ be the Nesterov accelerated **GD** process (i^{th} version) for the objective function \mathcal{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 6.4.1 and 6.4.5). Then

$$\Theta^{(1)} = \Theta^{(4)}. \quad (6.446)$$

Proof of Lemma 6.4.12. 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

$$\mathfrak{a}_n^{(1)} = \alpha_n^{(1)}, \quad \mathfrak{b}_n^{(1)} = 1 - \alpha_n^{(1)}, \quad \mathfrak{c}_n^{(1)} = \gamma_n^{(1)}, \quad (6.447)$$

$$\mathfrak{a}_n^{(4)} = \alpha_n^{(4)}, \quad \mathfrak{b}_n^{(4)} = \gamma_n^{(4)}, \quad \text{and} \quad \mathfrak{c}_n^{(4)} = 1. \quad (6.448)$$

Observe that (6.400), (6.401), (6.402), (6.409), (6.410), and (6.411) ensure that for all $i \in \{1, 4\}$, $n \in \mathbb{N}$ it holds that

$$\Theta_0^{(i)} = \xi, \quad \mathbf{m}_0^{(i)} = 0, \quad (6.449)$$

$$\mathbf{m}_n^{(i)} = \mathfrak{a}_n^{(i)} \mathbf{m}_{n-1}^{(i)} + \mathfrak{b}_n^{(i)} (\nabla \mathcal{L})(\Theta_{n-1}^{(i)} - \mathfrak{c}_n^{(i)} \mathfrak{a}_n^{(i)} \mathbf{m}_{n-1}^{(i)}), \quad (6.450)$$

$$\text{and} \quad \Theta_n^{(i)} = \Theta_{n-1}^{(i)} - \mathfrak{c}_n^{(i)} \mathbf{m}_n^{(i)}. \quad (6.451)$$

Furthermore, note that (6.445), (6.447), and (6.448) implies that for all $n \in \mathbb{N}$ it holds that

$$\mathfrak{b}_n^{(1)} \mathfrak{c}_n^{(1)} = (1 - \alpha_n^{(1)}) \gamma_n^{(1)} = \gamma_n^{(4)} = \mathfrak{b}_n^{(4)} \mathfrak{c}_n^{(4)}. \quad (6.452)$$

Moreover, observe that (6.445), (6.447), and (6.448) shows that for all $n \in \mathbb{N}$ it holds that

$$\frac{\mathfrak{a}_{n+1}^{(1)} \mathfrak{b}_n^{(1)}}{\mathfrak{b}_{n+1}^{(1)}} = \frac{\alpha_{n+1}^{(1)} (1 - \alpha_n^{(1)})}{1 - \alpha_{n+1}^{(1)}} = \frac{\alpha_{n+1}^{(1)} \gamma_n^{(4)} \gamma_{n+1}^{(1)}}{\gamma_n^{(1)} \gamma_{n+1}^{(4)}} = \frac{\alpha_{n+1}^{(4)} \gamma_n^{(4)}}{\gamma_{n+1}^{(4)}} = \frac{\mathfrak{a}_{n+1}^{(4)} \mathfrak{b}_n^{(4)}}{\mathfrak{b}_{n+1}^{(4)}}. \quad (6.453)$$

Combining this, (6.449), (6.450), (6.451), and (6.452) with Proposition 6.4.9 demonstrates (6.446). The proof of Lemma 6.4.12 is thus complete. \square

Corollary 6.4.13 (Comparison of the 2nd and 3rd version of the momentum SGD optimization method). *Let $\mathfrak{d} \in \mathbb{N}$, $(\gamma_n^{(2)})_{n \in \mathbb{N}} \subseteq (0, \infty)$, $(\gamma_n^{(3)})_{n \in \mathbb{N}} \subseteq (0, \infty)$, $(\alpha_n^{(2)})_{n \in \mathbb{N}} \subseteq (0, \infty)$, $(\alpha_n^{(3)})_{n \in \mathbb{N}} \subseteq (0, 1)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ satisfy for all $n \in \mathbb{N}$ that*

$$\gamma_n^{(2)} = \gamma_n^{(3)}(1 - \alpha_n^{(3)}) \quad \text{and} \quad \frac{\gamma_{n+1}^{(2)} \alpha_{n+1}^{(2)}}{\gamma_n^{(2)}} = \alpha_{n+1}^{(3)}, \quad (6.454)$$

for every $i \in \{2, 3\}$ let $\Theta^{(i)}: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ be the Nesterov accelerated GD process (i^{th} version) for the objective function \mathcal{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 6.4.3 and 6.4.7). Then

$$\Theta^{(2)} = \Theta^{(3)}. \quad (6.455)$$

Proof of Corollary 6.4.13. 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

$$\mathfrak{a}_n^{(2)} = \alpha_n^{(2)}, \quad \mathfrak{b}_n^{(2)} = 1, \quad \mathfrak{c}_n^{(2)} = \gamma_n^{(2)}, \quad (6.456)$$

$$\mathfrak{a}_n^{(3)} = \alpha_n^{(3)}, \quad \mathfrak{b}_n^{(3)} = (1 - \alpha_n^{(3)})\gamma_n^{(3)}, \quad \text{and} \quad \mathfrak{c}_n^{(3)} = 1. \quad (6.457)$$

Note that (6.403), (6.404), (6.405), (6.406), (6.407), and (6.408) prove that for all $i \in \{2, 3\}$, $n \in \mathbb{N}$ it holds that

$$\Theta_0^{(i)} = \xi, \quad \mathbf{m}_0^{(i)} = 0, \quad (6.458)$$

$$\mathbf{m}_n^{(i)} = \mathfrak{a}_n^{(i)} \mathbf{m}_{n-1}^{(i)} + \mathfrak{b}_n^{(i)} (\nabla \mathcal{L})(\Theta_{n-1}^{(i)} - \mathfrak{c}_n^{(i)} \mathfrak{a}_n^{(i)} \mathbf{m}_{n-1}^{(i)}), \quad (6.459)$$

$$\text{and} \quad \Theta_n^{(i)} = \Theta_{n-1}^{(i)} - \mathfrak{c}_n^{(i)} \mathbf{m}_n^{(i)}. \quad (6.460)$$

Furthermore, observe that (6.454), (6.456), and (6.457) establishes that for all $n \in \mathbb{N}$ it holds that

$$\mathfrak{b}_n^{(2)} \mathfrak{c}_n^{(2)} = \gamma_n^{(2)} = \gamma_n^{(3)}(1 - \alpha_n^{(3)}) = \mathfrak{b}_n^{(3)} \mathfrak{c}_n^{(3)}. \quad (6.461)$$

Moreover, note that (6.454), (6.456), and (6.457) ensures that for all $n \in \mathbb{N}$ it holds that

$$\frac{\mathfrak{a}_{n+1}^{(2)} \mathfrak{b}_n^{(2)}}{\mathfrak{b}_{n+1}^{(2)}} = \alpha_{n+1}^{(2)} = \frac{\alpha_{n+1}^{(3)} \gamma_n^{(3)} (1 - \alpha_n^{(3)})}{\gamma_{n+1}^{(3)} (1 - \alpha_{n+1}^{(3)})} = \frac{\mathfrak{a}_{n+1}^{(3)} \mathfrak{b}_n^{(3)}}{\mathfrak{b}_{n+1}^{(3)}}. \quad (6.462)$$

Combining this, (6.458), (6.459), (6.460), and (6.461) with Proposition 6.4.9 implies (6.455). The proof of Corollary 6.4.13 is thus complete. \square

Lemma 6.4.14 (Comparison of the 2nd and 4th version of the momentum **GD** optimization method). Let $\mathfrak{d} \in \mathbb{N}$, $(\gamma_n^{(2)})_{n \in \mathbb{N}} \subseteq (0, \infty)$, $(\gamma_n^{(4)})_{n \in \mathbb{N}} \subseteq (0, \infty)$, $(\alpha_n^{(2)})_{n \in \mathbb{N}} \subseteq (0, 1)$, $(\alpha_n^{(4)})_{n \in \mathbb{N}} \subseteq (0, 1)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ satisfy for all $n \in \mathbb{N}$ that

$$\gamma_n^{(2)} = \gamma_n^{(4)} \quad \text{and} \quad \frac{\gamma_{n+1}^{(2)} \alpha_{n+1}^{(2)}}{\gamma_n^{(2)}} = \alpha_{n+1}^{(4)}, \quad (6.463)$$

for every $i \in \{2, 4\}$ let $\Theta^{(i)}: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ be the Nesterov accelerated **GD** process (i^{th} version) for the objective function \mathcal{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 6.4.3 and 6.4.5). Then

$$\Theta^{(2)} = \Theta^{(4)}. \quad (6.464)$$

Proof of Lemma 6.4.14. 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

$$\mathfrak{a}_n^{(2)} = \alpha_n^{(2)}, \quad \mathfrak{b}_n^{(2)} = 1, \quad \mathfrak{c}_n^{(2)} = \gamma_n^{(2)}, \quad (6.465)$$

$$\mathfrak{a}_n^{(4)} = \alpha_n^{(4)}, \quad \mathfrak{b}_n^{(4)} = \gamma_n^{(4)}, \quad \text{and} \quad \mathfrak{c}_n^{(4)} = 1. \quad (6.466)$$

Observe that (6.403), (6.404), (6.405), (6.409), (6.410), and (6.411) show that for all $i \in \{2, 4\}$, $n \in \mathbb{N}$ it holds that

$$\Theta_0^{(i)} = \xi, \quad \mathbf{m}_0^{(i)} = 0, \quad (6.467)$$

$$\mathbf{m}_n^{(i)} = \mathfrak{a}_n^{(i)} \mathbf{m}_{n-1}^{(i)} + \mathfrak{b}_n^{(i)} (\nabla \mathcal{L})(\Theta_{n-1}^{(i)} - \mathfrak{c}_n^{(i)} \mathfrak{a}_n^{(i)} \mathbf{m}_{n-1}^{(i)}), \quad (6.468)$$

$$\text{and} \quad \Theta_n^{(i)} = \Theta_{n-1}^{(i)} - \mathfrak{c}_n^{(i)} \mathbf{m}_n^{(i)}. \quad (6.469)$$

Furthermore, note that (6.463), (6.465), and (6.466) demonstrates that for all $n \in \mathbb{N}$ it holds that

$$\mathfrak{b}_n^{(2)} \mathfrak{c}_n^{(2)} = \gamma_n^{(2)} = \gamma_n^{(4)} = \mathfrak{b}_n^{(4)} \mathfrak{c}_n^{(4)}. \quad (6.470)$$

Moreover, observe that (6.463), (6.465), and (6.466) proves that for all $n \in \mathbb{N}$ it holds that

$$\frac{\mathfrak{a}_{n+1}^{(2)} \mathfrak{b}_n^{(2)}}{\mathfrak{b}_{n+1}^{(2)}} = \alpha_{n+1}^{(2)} = \frac{\alpha_{n+1}^{(4)} \gamma_n^{(4)}}{\gamma_{n+1}^{(4)}} = \frac{\mathfrak{a}_{n+1}^{(4)} \mathfrak{b}_n^{(4)}}{\mathfrak{b}_{n+1}^{(4)}}. \quad (6.471)$$

Combining this, (6.467), (6.468), (6.469), and (6.470) with Proposition 6.4.9 establishes (6.464). The proof of Lemma 6.4.14 is thus complete. \square

Corollary 6.4.15 (Comparison of the 3rd and 4th version of the momentum GD optimization method). *Let $\mathfrak{d} \in \mathbb{N}$, $(\gamma_n^{(3)})_{n \in \mathbb{N}} \subseteq (0, \infty)$, $(\gamma_n^{(4)})_{n \in \mathbb{N}} \subseteq (0, \infty)$, $(\alpha_n^{(3)})_{n \in \mathbb{N}} \subseteq (0, 1)$, $(\alpha_n^{(4)})_{n \in \mathbb{N}} \subseteq (0, \infty)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ satisfy for all $n \in \mathbb{N}$ that*

$$\gamma_n^{(3)}(1 - \alpha_n^{(3)}) = \gamma_n^{(4)} \quad \text{and} \quad \alpha_{n+1}^{(3)} = \alpha_{n+1}^{(4)}, \quad (6.472)$$

for every $i \in \{3, 4\}$ let $\Theta^{(i)}: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ be the Nesterov accelerated GD process (i^{th} version) for the objective function \mathcal{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 6.4.5 and 6.4.7). Then

$$\Theta^{(3)} = \Theta^{(4)}. \quad (6.473)$$

Proof of Corollary 6.4.15. 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

$$\mathfrak{a}_n^{(3)} = \alpha_n^{(3)}, \quad \mathfrak{b}_n^{(3)} = (1 - \alpha_n^{(3)})\gamma_n^{(3)}, \quad \mathfrak{c}_n^{(3)} = 1 \quad (6.474)$$

$$\mathfrak{a}_n^{(4)} = \alpha_n^{(4)}, \quad \mathfrak{b}_n^{(4)} = \gamma_n^{(4)}, \quad \text{and} \quad \mathfrak{c}_n^{(4)} = 1, \quad (6.475)$$

Note that (6.406), (6.407), (6.408), (6.409), (6.410), and (6.411) ensure that for all $i \in \{3, 4\}$, $n \in \mathbb{N}$ it holds that

$$\Theta_0^{(i)} = \xi, \quad \mathbf{m}_0^{(i)} = 0, \quad (6.476)$$

$$\mathbf{m}_n^{(i)} = \mathfrak{a}_n^{(i)}\mathbf{m}_{n-1}^{(i)} + \mathfrak{b}_n^{(i)}(\nabla \mathcal{L})(\Theta_{n-1}^{(i)} - \mathfrak{c}_n^{(i)}\mathfrak{a}_n^{(i)}\mathbf{m}_{n-1}^{(i)}), \quad (6.477)$$

$$\text{and} \quad \Theta_n^{(i)} = \Theta_{n-1}^{(i)} - \mathfrak{c}_n^{(i)}\mathbf{m}_n^{(i)}. \quad (6.478)$$

Furthermore, observe that (6.472), (6.474), and (6.475) implies that for all $n \in \mathbb{N}$ it holds that

$$\mathfrak{b}_n^{(3)}\mathfrak{c}_n^{(3)} = \gamma_n^{(3)}(1 - \alpha_n^{(3)}) = \gamma_n^{(4)} = \mathfrak{b}_n^{(4)}\mathfrak{c}_n^{(4)}. \quad (6.479)$$

Moreover, note that (6.472), (6.474), and (6.475) shows that for all $n \in \mathbb{N}$ it holds that

$$\frac{\mathfrak{a}_{n+1}^{(3)}\mathfrak{b}_n^{(3)}}{\mathfrak{b}_{n+1}^{(3)}} = \frac{\alpha_{n+1}^{(3)}(1 - \alpha_n^{(3)})\gamma_n^{(3)}}{(1 - \alpha_{n+1}^{(3)})\gamma_{n+1}^{(3)}} = \frac{\alpha_{n+1}^{(4)}\gamma_n^{(4)}}{\gamma_{n+1}^{(4)}} = \frac{\mathfrak{a}_{n+1}^{(4)}\mathfrak{b}_n^{(4)}}{\mathfrak{b}_{n+1}^{(4)}}. \quad (6.480)$$

Combining this, (6.476), (6.477), (6.478), and (6.479) with Proposition 6.4.9 demonstrates (6.473). The proof of Corollary 6.4.15 is thus complete. \square

6.4.3 Bias-adjusted Nesterov accelerated momentum optimization

In this section, we introduce a bias-adjusted version of the Nesterov accelerated GD optimization method. Roughly speaking, the bias-adjusted Nesterov accelerated GD optimization method is obtained by adding the same kind of foresight to the bias-adjusted momentum GD optimization method in Definition 6.3.19 as the foresight that is added to the momentum GD optimization method in Definition 6.3.1 to obtain the Nesterov accelerated GD optimization method in Definition 6.4.1.

Definition 6.4.16 (Bias-adjusted Nesterov accelerated GD optimization method). *Let $\mathfrak{d} \in \mathbb{N}$, let $\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$ 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 $\Theta: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a function. Then we say that Θ is the bias-adjusted Nesterov accelerated GD process for the objective function \mathcal{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 $\mathbf{m}: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ such that for all $n \in \mathbb{N}$ it holds that*

$$\Theta_0 = \xi, \quad \mathbf{m}_0 = 0, \quad (6.481)$$

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + (1 - \alpha_n)(\nabla \mathcal{L})\left(\Theta_{n-1} - \frac{\gamma_n \alpha_n \mathbf{m}_{n-1}}{1 - \prod_{l=1}^n \alpha_l}\right), \quad (6.482)$$

$$\text{and} \quad \Theta_n = \Theta_{n-1} - \frac{\gamma_n \mathbf{m}_n}{1 - \prod_{l=1}^n \alpha_l}. \quad (6.483)$$

Algorithm 6.4.17: Bias-adjusted Nesterov accelerated GD optimization method

Input: $\mathfrak{d}, N \in \mathbb{N}$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, 1)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$

Output: N -th step of the bias-adjusted Nesterov accelerated GD process for the objective function \mathcal{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 6.4.16)

```

1: Initialization:  $\Theta \leftarrow \xi; \mathbf{m} \leftarrow 0 \in \mathbb{R}^{\mathfrak{d}}$ 
2: for  $n = 1, \dots, N$  do
3:    $\mathbf{m} \leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n)(\nabla \mathcal{L})(\Theta - \frac{\gamma_n \alpha_n \mathbf{m}}{1 - \prod_{l=1}^n \alpha_l})$ 
4:    $\Theta \leftarrow \Theta - \frac{\gamma_n \mathbf{m}}{1 - \prod_{l=1}^n \alpha_l}$ 
5: return  $\Theta$ 

```

6.4.4 Shifted representations

In this section, we introduce shifted representations of all the Nesterov accelerated GD optimization methods introduced in this section (cf. Definitions 6.4.1, 6.4.3, 6.4.5, 6.4.7, and 6.4.16). Roughly speaking, the shifted representations are obtained by taking the point at which the gradient of the objective function is evaluated as the current state of the optimization process.

6.4.4.1 Shifted representation for the first version of Nesterov accelerated momentum optimization

Lemma 6.4.18 (Shifting the Nesterov accelerated GD process). *Let $\mathfrak{d} \in \mathbb{N}$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $\xi \in \mathbb{R}^{\mathfrak{d}}$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$, let $\Theta: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ be the Nesterov accelerated GD process for the objective function \mathcal{L} with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, and initial value ξ , let $\mathbf{m}: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $n \in \mathbb{N}$ that $\mathbf{m}_0 = 0$ and*

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + (1 - \alpha_n)(\nabla \mathcal{L})(\Theta_{n-1} - \gamma_n \alpha_n \mathbf{m}_{n-1}), \quad (6.484)$$

and let $\Psi: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $n \in \mathbb{N}_0$ that

$$\Psi_n = \Theta_n - \gamma_{n+1} \alpha_{n+1} \mathbf{m}_n \quad (6.485)$$

(cf. Definition 6.4.1). Then it holds for all $n \in \mathbb{N}$ that

$$\Psi_0 = \xi, \quad \mathbf{m}_0 = 0, \quad (6.486)$$

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + (1 - \alpha_n)(\nabla \mathcal{L})(\Psi_{n-1}), \quad (6.487)$$

$$\text{and} \quad \Psi_n = \Psi_{n-1} - \gamma_{n+1} \alpha_{n+1} \mathbf{m}_n - \gamma_n (1 - \alpha_n)(\nabla \mathcal{L})(\Psi_{n-1}). \quad (6.488)$$

Proof of Lemma 6.4.18. Observe that (6.400), (6.401), (6.402), and (6.484) prove that for all $n \in \mathbb{N}$ it holds that

$$\Theta_n = \Theta_{n-1} - \gamma_n \mathbf{m}_n. \quad (6.489)$$

Furthermore, note that (6.484), the fact that $\Theta_0 = \xi$, and the assumption that $\mathbf{m}_0 = 0$ establish that

$$\Psi_0 = \Theta_0 - \gamma_1 \alpha_1 \mathbf{m}_0 = \xi. \quad (6.490)$$

Moreover, observe that (6.484) and (6.485) ensure that for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned} \mathbf{m}_n &= \alpha_n \mathbf{m}_{n-1} + (1 - \alpha_n)(\nabla \mathcal{L})(\Theta_{n-1} - \gamma_n \alpha_n \mathbf{m}_{n-1}) \\ &= \alpha_n \mathbf{m}_{n-1} + (1 - \alpha_n)(\nabla \mathcal{L})(\Psi_{n-1}). \end{aligned} \quad (6.491)$$

This, (6.485), and (6.489) imply that for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned}\Psi_n &= \Theta_n - \gamma_{n+1}\alpha_{n+1}\mathbf{m}_n \\ &= \Theta_{n-1} - \gamma_n\mathbf{m}_n - \gamma_{n+1}\alpha_{n+1}\mathbf{m}_n \\ &= \Theta_{n-1} - \gamma_n(\alpha_n\mathbf{m}_{n-1} + (1 - \alpha_n)(\nabla \mathcal{L})(\Psi_{n-1})) - \gamma_{n+1}\alpha_{n+1}\mathbf{m}_n \\ &= \Psi_{n-1} - \gamma_n(1 - \alpha_n)(\nabla \mathcal{L})(\Psi_{n-1}) - \gamma_{n+1}\alpha_{n+1}\mathbf{m}_n.\end{aligned}\tag{6.492}$$

Combining this, (6.490), and (6.491) establishes (6.486), (6.487), and (6.488). The proof of Lemma 6.4.18 is thus complete. \square

Definition 6.4.19 (Shifted Nesterov accelerated **GD** optimization method). *Let $\mathfrak{d} \in \mathbb{N}$, let $\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$ 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 $\Theta: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a function. Then we say that Θ is the shifted Nesterov accelerated **GD** process for the objective function \mathcal{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 $\mathbf{m}: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ such that for all $n \in \mathbb{N}$ it holds that*

$$\Theta_0 = \xi, \quad \mathbf{m}_0 = 0, \tag{6.493}$$

$$\mathbf{m}_n = \alpha_n\mathbf{m}_{n-1} + (1 - \alpha_n)(\nabla \mathcal{L})(\Theta_{n-1}), \quad \text{and} \tag{6.494}$$

$$\Theta_n = \Theta_{n-1} - \gamma_{n+1}\alpha_{n+1}\mathbf{m}_n - \gamma_n(1 - \alpha_n)(\nabla \mathcal{L})(\Theta_{n-1}). \tag{6.495}$$

Algorithm 6.4.20: Shifted Nesterov accelerated **GD** optimization method

Input: $\mathfrak{d}, N \in \mathbb{N}$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $\xi \in \mathbb{R}^{\mathfrak{d}}$

Output: N -th step of the shifted Nesterov accelerated **GD** process for the objective function \mathcal{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 6.4.19)

- 1: **Initialization:** $\Theta \leftarrow \xi$; $\mathbf{m} \leftarrow 0 \in \mathbb{R}^{\mathfrak{d}}$
- 2: **for** $n = 1, \dots, N$ **do**
- 3: $\mathbf{m} \leftarrow \alpha_n\mathbf{m} + (1 - \alpha_n)(\nabla \mathcal{L})(\Theta)$
- 4: $\Theta \leftarrow \Theta - \gamma_{n+1}\alpha_{n+1}\mathbf{m} - \gamma_n(1 - \alpha_n)(\nabla \mathcal{L})(\Theta)$
- 5: **return** Θ

6.4.4.2 Shifted representation for the second version of Nesterov accelerated momentum optimization

Lemma 6.4.21 (Shifting the Nesterov accelerated **GD** process (2nd version)). *Let $\mathfrak{d} \in \mathbb{N}$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$, let $\Theta: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ be the*

Nesterov accelerated **GD** process (2^{nd} version) for the objective function \mathcal{L} with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, and initial value ξ , let $\mathbf{m}: \mathbb{N}_0 \rightarrow \mathbb{R}^d$ satisfy for all $n \in \mathbb{N}$ that $\mathbf{m}_0 = 0$ and

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + (\nabla \mathcal{L})(\Theta_{n-1} - \gamma_n \alpha_n \mathbf{m}_{n-1}), \quad (6.496)$$

and let $\Psi: \mathbb{N}_0 \rightarrow \mathbb{R}^d$ satisfy for all $n \in \mathbb{N}_0$ that

$$\Psi_n = \Theta_n - \gamma_{n+1} \alpha_{n+1} \mathbf{m}_n \quad (6.497)$$

(cf. Definition 6.4.3). Then it holds for all $n \in \mathbb{N}$ that

$$\Psi_0 = \xi, \quad \mathbf{m}_0 = 0, \quad (6.498)$$

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + (\nabla \mathcal{L})(\Psi_{n-1}), \quad (6.499)$$

$$\text{and } \Psi_n = \Psi_{n-1} - \gamma_{n+1} \alpha_{n+1} \mathbf{m}_n - \gamma_n (\nabla \mathcal{L})(\Psi_{n-1}). \quad (6.500)$$

Proof of Lemma 6.4.21. Note that (6.403), (6.404), (6.405), and (6.496) show that for all $n \in \mathbb{N}$ it holds that

$$\Theta_n = \Theta_{n-1} - \gamma_n \mathbf{m}_n. \quad (6.501)$$

Furthermore, observe that (6.496), the fact that $\Theta_0 = \xi$, and the assumption that $\mathbf{m}_0 = 0$ demonstrate that

$$\Psi_0 = \Theta_0 - \gamma_1 \alpha_1 \mathbf{m}_0 = \xi. \quad (6.502)$$

Moreover, note that (6.496) and (6.497) prove that for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned} \mathbf{m}_n &= \alpha_n \mathbf{m}_{n-1} + (\nabla \mathcal{L})(\Theta_{n-1} - \gamma_n \alpha_n \mathbf{m}_{n-1}) \\ &= \alpha_n \mathbf{m}_{n-1} + (\nabla \mathcal{L})(\Psi_{n-1}). \end{aligned} \quad (6.503)$$

In addition, observe that this, (6.496), (6.497), and (6.501) establish that for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned} \Psi_n &= \Theta_n - \gamma_{n+1} \alpha_{n+1} \mathbf{m}_n \\ &= \Theta_{n-1} - \gamma_n \mathbf{m}_n - \gamma_{n+1} \alpha_{n+1} \mathbf{m}_n \\ &= \Theta_{n-1} - \gamma_n (\alpha_n \mathbf{m}_{n-1} + (\nabla \mathcal{L})(\Psi_{n-1})) - \gamma_{n+1} \alpha_{n+1} \mathbf{m}_n \\ &= \Psi_{n-1} - \gamma_n (\nabla \mathcal{L})(\Psi_{n-1}) - \gamma_{n+1} \alpha_{n+1} \mathbf{m}_n. \end{aligned} \quad (6.504)$$

Combining this, (6.502), and (6.503) establishes (6.498), (6.499), and (6.500). The proof of Lemma 6.4.21 is thus complete. \square

Definition 6.4.22 (Shifted Nesterov accelerated **GD** optimization method (2nd version)). Let $\mathfrak{d} \in \mathbb{N}$, let $\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$ 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 $\Theta: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a function. Then we say that Θ is the shifted Nesterov accelerated **GD** process (2nd version) for the objective function \mathcal{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 $\mathbf{m}: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ such that for all $n \in \mathbb{N}$ it holds that

$$\Theta_0 = \xi, \quad \mathbf{m}_0 = 0, \quad (6.505)$$

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + (\nabla \mathcal{L})(\Theta_{n-1}), \quad \text{and} \quad (6.506)$$

$$\Theta_n = \Theta_{n-1} - \gamma_{n+1} \alpha_{n+1} \mathbf{m}_n - \gamma_n (\nabla \mathcal{L})(\Theta_{n-1}). \quad (6.507)$$

Algorithm 6.4.23: Shifted Nesterov accelerated **GD optimization method (2nd version)**

Input: $\mathfrak{d}, N \in \mathbb{N}$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$

Output: N -th step of the shifted Nesterov accelerated **GD** process (2nd version) for the objective function \mathcal{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 6.4.22)

```

1: Initialization:  $\Theta \leftarrow \xi; \mathbf{m} \leftarrow 0 \in \mathbb{R}^{\mathfrak{d}}$ 
2: for  $n = 1, \dots, N$  do
3:    $\mathbf{m} \leftarrow \alpha_n \mathbf{m} + (\nabla \mathcal{L})(\Theta)$ 
4:    $\Theta \leftarrow \Theta - \gamma_{n+1} \alpha_{n+1} \mathbf{m} - \gamma_n (\nabla \mathcal{L})(\Theta)$ 
5: return  $\Theta$ 

```

6.4.4.3 Shifted representation for the third version of Nesterov accelerated momentum optimization

Lemma 6.4.24 (Shifting the Nesterov accelerated **GD** process (3rd version)). Let $\mathfrak{d} \in \mathbb{N}$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$, let $\Theta: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ be the Nesterov accelerated **GD** process (3rd version) for the objective function \mathcal{L} with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, and initial value ξ , let $\mathbf{m}: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $n \in \mathbb{N}$ that $\mathbf{m}_0 = 0$ and

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + (1 - \alpha_n) \gamma_n (\nabla \mathcal{L})(\Theta_{n-1} - \alpha_n \mathbf{m}_{n-1}), \quad (6.508)$$

and let $\Psi: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $n \in \mathbb{N}_0$ that

$$\Psi_n = \Theta_n - \alpha_{n+1}\mathbf{m}_n \quad (6.509)$$

(cf. Definition 6.4.5). Then it holds for all $n \in \mathbb{N}$ that

$$\Psi_0 = \xi, \quad \mathbf{m}_0 = 0, \quad (6.510)$$

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + (1 - \alpha_n) \gamma_n (\nabla \mathcal{L})(\Psi_{n-1}), \quad (6.511)$$

$$\text{and } \Psi_n = \Psi_{n-1} - \alpha_{n+1}\mathbf{m}_n - (1 - \alpha_n) \gamma_n (\nabla \mathcal{L})(\Psi_{n-1}). \quad (6.512)$$

Proof of Lemma 6.4.24. Note that (6.406), (6.407), (6.408), and (6.508) ensure that for all $n \in \mathbb{N}$ it holds that

$$\Theta_n = \Theta_{n-1} - \mathbf{m}_n. \quad (6.513)$$

Furthermore, observe that (6.508), the fact that $\Theta_0 = \xi$, and the assumption that $\mathbf{m}_0 = 0$ imply that

$$\Psi_0 = \Theta_0 - \alpha_1 \mathbf{m}_0 = \xi. \quad (6.514)$$

Moreover, note that (6.508) and (6.509) show that for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned} \mathbf{m}_n &= \alpha_n \mathbf{m}_{n-1} + (1 - \alpha_n) \gamma_n (\nabla \mathcal{L})(\Theta_{n-1} - \alpha_n \mathbf{m}_{n-1}) \\ &= \alpha_n \mathbf{m}_{n-1} + (1 - \alpha_n) \gamma_n (\nabla \mathcal{L})(\Psi_{n-1}). \end{aligned} \quad (6.515)$$

In addition, observe that this, (6.508), (6.509), and (6.513) demonstrate that for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned} \Psi_n &= \Theta_n - \alpha_{n+1}\mathbf{m}_n \\ &= \Theta_{n-1} - \mathbf{m}_n - \alpha_{n+1}\mathbf{m}_n \\ &= \Theta_{n-1} - (\alpha_n \mathbf{m}_{n-1} + (1 - \alpha_n) \gamma_n (\nabla \mathcal{L})(\Psi_{n-1})) - \alpha_{n+1}\mathbf{m}_n \\ &= \Psi_{n-1} - (1 - \alpha_n) \gamma_n (\nabla \mathcal{L})(\Psi_{n-1}) - \alpha_{n+1}\mathbf{m}_n. \end{aligned} \quad (6.516)$$

Combining this, (6.514), and (6.515) establishes (6.510), (6.511), and (6.512). The proof of Lemma 6.4.24 is thus complete. \square

Definition 6.4.25 (Shifted Nesterov accelerated GD optimization method (3rd version)). Let $\mathfrak{d} \in \mathbb{N}$, let $\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$ 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 $\Theta: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a function. Then we say that Θ is the shifted Nesterov accelerated GD process (3rd version) for the objective function \mathcal{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

$\mathbf{m}: \mathbb{N}_0 \rightarrow \mathbb{R}^d$ such that for all $n \in \mathbb{N}$ it holds that

$$\Theta_0 = \xi, \quad \mathbf{m}_0 = 0, \quad (6.517)$$

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + (1 - \alpha_n) \gamma_n (\nabla \mathcal{L})(\Theta_{n-1}), \quad (6.518)$$

$$\text{and} \quad \Theta_n = \Theta_{n-1} - \alpha_{n+1} \mathbf{m}_n - (1 - \alpha_n) \gamma_n (\nabla \mathcal{L})(\Theta_{n-1}). \quad (6.519)$$

Algorithm 6.4.26: Shifted Nesterov accelerated GD optimization method (3rd version)

Input: $d, N \in \mathbb{N}$, $\mathcal{L} \in C^1(\mathbb{R}^d, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $\xi \in \mathbb{R}^d$

Output: N -th step of the shifted Nesterov accelerated GD process (3rd version) for the objective function \mathcal{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 6.4.25)

- 1: **Initialization:** $\Theta \leftarrow \xi; \mathbf{m} \leftarrow 0 \in \mathbb{R}^d$
- 2: **for** $n = 1, \dots, N$ **do**
- 3: $\mathbf{m} \leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n) \gamma_n (\nabla \mathcal{L})(\Theta)$
- 4: $\Theta \leftarrow \Theta - \alpha_{n+1} \mathbf{m} - (1 - \alpha_n) \gamma_n (\nabla \mathcal{L})(\Theta)$
- 5: **return** Θ

6.4.4.4 Shifted representation for the fourth version of Nesterov accelerated GD optimization

Lemma 6.4.27 (Shifting the Nesterov accelerated GD process (4th version)). *Let $d \in \mathbb{N}$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $\xi \in \mathbb{R}^d$, $\mathcal{L} \in C^1(\mathbb{R}^d, \mathbb{R})$, let $\Theta: \mathbb{N}_0 \rightarrow \mathbb{R}^d$ be the Nesterov accelerated GD process (4th version) for the objective function \mathcal{L} with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, and initial value ξ , let $\mathbf{m}: \mathbb{N}_0 \rightarrow \mathbb{R}^d$ satisfy for all $n \in \mathbb{N}$ that $\mathbf{m}_0 = 0$ and*

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + \gamma_n (\nabla \mathcal{L})(\Theta_{n-1} - \alpha_n \mathbf{m}_{n-1}), \quad (6.520)$$

and let $\Psi: \mathbb{N}_0 \rightarrow \mathbb{R}^d$ satisfy for all $n \in \mathbb{N}_0$ that

$$\Psi_n = \Theta_n - \alpha_{n+1} \mathbf{m}_n \quad (6.521)$$

(cf. Definition 6.4.7). Then it holds for all $n \in \mathbb{N}$ that

$$\Psi_0 = \xi, \quad \mathbf{m}_0 = 0, \quad (6.522)$$

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + \gamma_n (\nabla \mathcal{L})(\Psi_{n-1}), \quad (6.523)$$

$$\text{and} \quad \Psi_n = \Psi_{n-1} - \alpha_{n+1} \mathbf{m}_n - \gamma_n (\nabla \mathcal{L})(\Psi_{n-1}). \quad (6.524)$$

Proof of Lemma 6.4.27. Note that (6.409), (6.410), (6.411), and (6.520) prove that for all $n \in \mathbb{N}$ it holds that

$$\Theta_n = \Theta_{n-1} - \mathbf{m}_n. \quad (6.525)$$

Furthermore, observe that (6.520), the fact that $\Theta_0 = \xi$, and the assumption that $\mathbf{m}_0 = 0$ establish that

$$\Psi_0 = \Theta_0 - \alpha_1 \mathbf{m}_0 = \xi. \quad (6.526)$$

Moreover, note that (6.520) and (6.521) ensure that for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned} \mathbf{m}_n &= \alpha_n \mathbf{m}_{n-1} + \gamma_n (\nabla \mathcal{L})(\Theta_{n-1} - \alpha_n \mathbf{m}_{n-1}) \\ &= \alpha_n \mathbf{m}_{n-1} + \gamma_n (\nabla \mathcal{L})(\Psi_{n-1}). \end{aligned} \quad (6.527)$$

In addition, observe that this, (6.520), (6.521), and (6.525) imply that for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned} \Psi_n &= \Theta_n - \alpha_{n+1} \mathbf{m}_n \\ &= \Theta_{n-1} - \mathbf{m}_n - \alpha_{n+1} \mathbf{m}_n \\ &= \Theta_{n-1} - (\alpha_n \mathbf{m}_{n-1} + \gamma_n (\nabla \mathcal{L})(\Psi_{n-1})) - \alpha_{n+1} \mathbf{m}_n \\ &= \Psi_{n-1} - \gamma_n (\nabla \mathcal{L})(\Psi_{n-1}) - \alpha_{n+1} \mathbf{m}_n. \end{aligned} \quad (6.528)$$

Combining this, (6.526), and (6.527) establishes (6.522), (6.523), and (6.524). The proof of Lemma 6.4.27 is thus complete. \square

Definition 6.4.28 (Shifted Nesterov accelerated GD optimization method (4th version)). Let $\mathfrak{d} \in \mathbb{N}$, let $\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$ 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 $\Theta: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a function. Then we say that Θ is the shifted Nesterov accelerated GD process (3rd version) for the objective function \mathcal{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 $\mathbf{m}: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ such that for all $n \in \mathbb{N}$ it holds that

$$\Theta_0 = \xi, \quad \mathbf{m}_0 = 0, \quad (6.529)$$

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + \gamma_n (\nabla \mathcal{L})(\Theta_{n-1}), \quad (6.530)$$

$$\text{and} \quad \Theta_n = \Theta_{n-1} - \alpha_{n+1} \mathbf{m}_n - \gamma_n (\nabla \mathcal{L})(\Theta_{n-1}). \quad (6.531)$$

Algorithm 6.4.29: Shifted Nesterov accelerated GD optimization method (4th version)

Input: $\mathfrak{d}, N \in \mathbb{N}$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$

Output: N -th step of the shifted Nesterov accelerated GD process (4th version) for the objective function \mathcal{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 6.4.28)

```

1: Initialization:  $\Theta \leftarrow \xi; \mathbf{m} \leftarrow 0 \in \mathbb{R}^{\mathfrak{d}}$ 
2: for  $n = 1, \dots, N$  do
3:    $\mathbf{m} \leftarrow \alpha_n \mathbf{m} + \gamma_n (\nabla \mathcal{L})(\Theta)$ 
4:    $\Theta \leftarrow \Theta - \alpha_{n+1} \mathbf{m} - \gamma_n (\nabla \mathcal{L})(\Theta)$ 
5: return  $\Theta$ 

```

6.4.4.5 Shifted representation for the bias-adjusted Nesterov accelerated GD optimization

Lemma 6.4.30 (Shifting the bias-adjusted Nesterov accelerated GD process). *Let $\mathfrak{d} \in \mathbb{N}$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $\xi \in \mathbb{R}^{\mathfrak{d}}$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$, let $\Theta: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ be the bias-adjusted Nesterov accelerated GD process for the objective function \mathcal{L} with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, and initial value ξ , let $\mathbf{m}: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $n \in \mathbb{N}$ that $\mathbf{m}_0 = 0$ and*

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + (1 - \alpha_n) (\nabla \mathcal{L}) \left(\Theta_{n-1} - \frac{\gamma_n \alpha_n \mathbf{m}_{n-1}}{1 - \prod_{l=1}^n \alpha_l} \right), \quad (6.532)$$

and let $\Psi: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $n \in \mathbb{N}_0$ that

$$\Psi_n = \Theta_n - \frac{\gamma_{n+1} \alpha_{n+1} \mathbf{m}_n}{1 - \prod_{l=1}^{n+1} \alpha_l} \quad (6.533)$$

(cf. Definition 6.4.16). Then it holds for all $n \in \mathbb{N}$ that

$$\Psi_0 = \xi, \quad \mathbf{m}_0 = 0, \quad (6.534)$$

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + (1 - \alpha_n) (\nabla \mathcal{L})(\Psi_{n-1}), \quad \text{and} \quad (6.535)$$

$$\Psi_n = \Psi_{n-1} - \frac{\gamma_{n+1} \alpha_{n+1} \mathbf{m}_n}{1 - \prod_{l=1}^{n+1} \alpha_l} - \frac{\gamma_n (1 - \alpha_n) (\nabla \mathcal{L})(\Psi_{n-1})}{1 - \prod_{l=1}^n \alpha_l}. \quad (6.536)$$

Proof of Lemma 6.4.30. Note that (6.481), (6.482), (6.483), and (6.532) show that for all $n \in \mathbb{N}$ it holds that

$$\Theta_n = \Theta_{n-1} - \frac{\gamma_n \mathbf{m}_n}{1 - \prod_{l=1}^n \alpha_l}. \quad (6.537)$$

Furthermore, observe that (6.532), the fact that $\Theta_0 = \xi$, and the assumption that $\mathbf{m}_0 = 0$ demonstrate that

$$\Psi_0 = \Theta_0 - \alpha_1 \mathbf{m}_0 = \xi. \quad (6.538)$$

Moreover, note that (6.532) and (6.533) prove that for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned} \mathbf{m}_n &= \alpha_n \mathbf{m}_{n-1} + (1 - \alpha_n)(\nabla \mathcal{L})(\Theta_{n-1} - \frac{\gamma_n \alpha_n \mathbf{m}_{n-1}}{1 - \prod_{l=1}^n \alpha_l}) \\ &= \alpha_n \mathbf{m}_{n-1} + (1 - \alpha_n)(\nabla \mathcal{L})(\Psi_{n-1}). \end{aligned} \quad (6.539)$$

In addition, observe that this, (6.532), (6.533), and (6.537) establish that for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned} \Psi_n &= \Theta_n - \frac{\gamma_{n+1} \alpha_{n+1} \mathbf{m}_n}{1 - \prod_{l=1}^{n+1} \alpha_l} \\ &= \Theta_{n-1} - \frac{\gamma_n \mathbf{m}_n}{1 - \prod_{l=1}^n \alpha_l} - \frac{\gamma_{n+1} \alpha_{n+1} \mathbf{m}_n}{1 - \prod_{l=1}^{n+1} \alpha_l} \\ &= \Theta_{n-1} - \frac{\gamma_n (\alpha_n \mathbf{m}_{n-1} + (1 - \alpha_n)(\nabla \mathcal{L})(\Psi_{n-1}))}{1 - \prod_{l=1}^n \alpha_l} - \frac{\gamma_{n+1} \alpha_{n+1} \mathbf{m}_n}{1 - \prod_{l=1}^{n+1} \alpha_l} \\ &= \Psi_{n-1} - \frac{\gamma_n (1 - \alpha_n)(\nabla \mathcal{L})(\Psi_{n-1})}{1 - \prod_{l=1}^n \alpha_l} - \frac{\gamma_{n+1} \alpha_{n+1} \mathbf{m}_n}{1 - \prod_{l=1}^{n+1} \alpha_l}. \end{aligned} \quad (6.540)$$

Combining this, (6.538), and (6.539) establishes (6.534), (6.535), and (6.536). The proof of Lemma 6.4.30 is thus complete. \square

Definition 6.4.31 (Shifted bias-adjusted Nesterov accelerated **GD** optimization method). Let $\mathfrak{d} \in \mathbb{N}$, let $\mathcal{L}: \mathbb{R}^\mathfrak{d} \rightarrow \mathbb{R}$ 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 $\Theta: \mathbb{N}_0 \rightarrow \mathbb{R}^\mathfrak{d}$ be a function. Then we say that Θ is the shifted bias-adjusted Nesterov accelerated **GD** process for the objective function \mathcal{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 $\mathbf{m}: \mathbb{N}_0 \rightarrow \mathbb{R}^\mathfrak{d}$ such that for all $n \in \mathbb{N}$ it holds that

$$\Theta_0 = \xi, \quad \mathbf{m}_0 = 0, \quad (6.541)$$

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + (1 - \alpha_n)(\nabla \mathcal{L})(\Theta_{n-1}), \quad \text{and} \quad (6.542)$$

$$\Theta_n = \Theta_{n-1} - \frac{\gamma_n(1-\alpha_n)(\nabla \mathcal{L})(\Theta_{n-1})}{1 - \prod_{l=1}^n \alpha_l} - \frac{\gamma_{n+1}\alpha_{n+1}\mathbf{m}_n}{1 - \prod_{l=1}^{n+1} \alpha_l}. \quad (6.543)$$

Algorithm 6.4.32: Shifted bias-adjusted Nesterov accelerated **GD** optimization method

Input: $\mathfrak{d}, N \in \mathbb{N}$, $\mathcal{L} \in C^1(\mathbb{R}^\mathfrak{d}, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $\xi \in \mathbb{R}^\mathfrak{d}$

Output: N -th step of the Nesterov accelerated **GD** process for the objective function \mathcal{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 6.4.31)

```

1: Initialization:  $\Theta \leftarrow \xi; \mathbf{m} \leftarrow 0 \in \mathbb{R}^\mathfrak{d}$ 
2: for  $n = 1, \dots, N$  do
3:    $\mathbf{m} \leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n)(\nabla \mathcal{L})(\Theta)$ 
4:    $\Theta \leftarrow \Theta - \frac{\gamma_n(1-\alpha_n)(\nabla \mathcal{L})(\Theta)}{1 - \prod_{l=1}^n \alpha_l} - \frac{\gamma_{n+1}\alpha_{n+1}\mathbf{m}_n}{1 - \prod_{l=1}^{n+1} \alpha_l}$ 
5: return  $\Theta$ 

```

6.4.5 Simplified Nesterov accelerated momentum optimization

For reasons of algorithmic simplicity, in several deep learning libraries including PYTORCH (see [359] and cf., for example, [31, Section 3.5]) optimization with Nesterov momentum is not implemented such that it precisely corresponds to any of the definitions presented above. Rather, an alternative definition for Nesterov accelerated **GD** optimization is used, which we present in Definition 6.4.33 below. Roughly speaking, the simplified version of Nesterov accelerated **GD** optimization in Definition 6.4.33 employed by PYTORCH is obtained by reducing some indices in the update rule of the shifted Nesterov accelerated **GD** optimization method (2nd version) in Definition 6.4.22 so that for each update step only one learning rate and one momentum decay factor are used.

Definition 6.4.33 (Simplified Nesterov accelerated **GD** optimization method). *Let $\mathfrak{d} \in \mathbb{N}$, let $\mathcal{L}: \mathbb{R}^\mathfrak{d} \rightarrow \mathbb{R}$ 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 $\Theta: \mathbb{N}_0 \rightarrow \mathbb{R}^\mathfrak{d}$ be a function. Then we say that Θ is the simplified Nesterov accelerated **GD** process for the objective function \mathcal{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 $\mathbf{m}: \mathbb{N}_0 \rightarrow \mathbb{R}^\mathfrak{d}$ such that for all $n \in \mathbb{N}$ it holds that*

$$\Theta_0 = \xi, \quad \mathbf{m}_0 = 0, \quad (6.544)$$

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + (\nabla \mathcal{L})(\Theta_{n-1}), \quad \text{and} \quad (6.545)$$

$$\Theta_n = \Theta_{n-1} - \gamma_n \alpha_n \mathbf{m}_n - \gamma_n (\nabla \mathcal{L})(\Theta_{n-1}). \quad (6.546)$$

Algorithm 6.4.34: Simplified Nesterov accelerated GD optimization method

Input: $\mathfrak{d}, N \in \mathbb{N}$, $\mathcal{L} \in C^1(\mathbb{R}^\mathfrak{d}, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $\xi \in \mathbb{R}^\mathfrak{d}$
Output: N -th step of the simplified Nesterov accelerated GD process for the objective function \mathcal{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 6.4.33)

```

1: Initialization:  $\Theta \leftarrow \xi$ ;  $\mathbf{m} \leftarrow 0 \in \mathbb{R}^\mathfrak{d}$ 
2: for  $n = 1, \dots, N$  do
3:    $\mathbf{m} \leftarrow \alpha_n \mathbf{m} + (\nabla \mathcal{L})(\Theta)$ 
4:    $\Theta \leftarrow \Theta - \gamma_n \alpha_n \mathbf{m} - \gamma_n (\nabla \mathcal{L})(\Theta)$ 
5: return  $\Theta$ 
```

6.5 Adaptive gradient (Adagrad) optimization

In this section we review the *adaptive gradient* (Adagrad) GD optimization method. Roughly speaking, the idea of the Adagrad GD optimization method is to modify the plain-vanilla GD optimization method by adapting the learning rates separately for every component of the optimization process. Adagrad was first presented in Duchi et al. [123] in the context of stochastic optimization. For pedagogical purposes we present in this section a deterministic version of Adagrad optimization and we refer to Section 7.6 below for the original stochastic version of Adagrad optimization.

Definition 6.5.1 (Adagrad GD optimization method). *Let $\mathfrak{d} \in \mathbb{N}$, let $\mathcal{L}: \mathbb{R}^\mathfrak{d} \rightarrow \mathbb{R}$ be differentiable, let $\mathcal{G} = (\mathcal{G}_1, \dots, \mathcal{G}_\mathfrak{d}): \mathbb{R}^\mathfrak{d} \rightarrow \mathbb{R}^\mathfrak{d}$ satisfy for all $\theta \in \mathbb{R}^\mathfrak{d}$ that*

$$\mathcal{G}(\theta) = (\nabla \mathcal{L})(\theta), \quad (6.547)$$

let $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $\varepsilon \in (0, \infty)$, $\xi \in \mathbb{R}^\mathfrak{d}$, and let $\Theta = (\Theta^{(1)}, \dots, \Theta^{(\mathfrak{d})}): \mathbb{N}_0 \rightarrow \mathbb{R}^\mathfrak{d}$ be a function. Then we say that Θ is the Adagrad GD process for the objective function \mathcal{L} with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, regularizing factor ε , and initial value ξ if and only if it

holds for all $n \in \mathbb{N}$, $i \in \{1, 2, \dots, \mathfrak{d}\}$ that

$$\Theta_0 = \xi \quad \text{and} \quad \Theta_n^{(i)} = \Theta_{n-1}^{(i)} - \gamma_n \left[\varepsilon + \left[\sum_{k=0}^{n-1} |\mathcal{G}_i(\Theta_k)|^2 \right]^{1/2} \right]^{-1} \mathcal{G}_i(\Theta_{n-1}). \quad (6.548)$$

Definition 6.5.2 (Componentwise operations). For every $d \in \mathbb{N}$, $c \in \mathbb{R}$, $v = (v_1, \dots, v_d) \in \mathbb{R}^d$ we denote by $c + v$ and $v + c$ the vectors which satisfy

$$c + v = v + c = (c + v_1, c + v_2, \dots, c + v_d), \quad (6.549)$$

for every $d \in \mathbb{N}$, $p \in \mathbb{R}$, $v = (v_1, \dots, v_d) \in [0, \infty)^d$ with $(p > 0) \vee (v \in (0, \infty)^d)$ we denote by v^p the vector which satisfies

$$v^p = ((v_1)^p, (v_2)^p, \dots, (v_d)^p), \quad (6.550)$$

for every $d \in \mathbb{N}$, $v = (v_1, \dots, v_d) \in \mathbb{R}^d$ we denote by $|v|$ the vector which satisfies

$$|v| = (|v_1|, |v_2|, \dots, |v_d|), \quad (6.551)$$

for every $d \in \mathbb{N}$, $v = (v_1, \dots, v_d)$, $w = (w_1, \dots, w_d) \in \mathbb{R}^d$ we denote by $vw \in \mathbb{R}^d$ the vector which satisfies

$$vw = (v_1 w_1, v_2 w_2, \dots, v_d w_d), \quad (6.552)$$

for every $d \in \mathbb{N}$, $v = (v_1, \dots, v_d) \in \mathbb{R}^d$, $w = (w_1, \dots, w_d) \in (\mathbb{R} \setminus \{0\})^d$ we denote by $\frac{v}{w} \in \mathbb{R}^d$ the vector which satisfies

$$\frac{v}{w} = \left(\frac{v_1}{w_1}, \frac{v_2}{w_2}, \dots, \frac{v_d}{w_d} \right), \quad (6.553)$$

and for every $d \in \mathbb{N}$, $v = (v_1, \dots, v_d)$, $w = (w_1, \dots, w_d) \in \mathbb{R}^d$ we denote by $\max\{v, w\} \in \mathbb{R}^d$ the vector which satisfies

$$\max\{v, w\} = (\max\{v_1, w_1\}, \max\{v_2, w_2\}, \dots, \max\{v_d, w_d\}). \quad (6.554)$$

Algorithm 6.5.3: Adagrad GD optimization method

Input: $\mathfrak{d}, N \in \mathbb{N}$, $\mathcal{L} \in C^1(\mathbb{R}^\mathfrak{d}, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $\varepsilon \in (0, \infty)$, $\xi \in \mathbb{R}^\mathfrak{d}$

Output: N -th step of the Adagrad GD process for the objective function \mathcal{L} with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, regularizing factor ε , and initial value ξ (cf. Definition 6.5.1)

```

1: Initialization:  $\Theta \leftarrow \xi; \mathbb{M} \leftarrow 0 \in \mathbb{R}^{\mathfrak{d}}$ 
2: for  $n = 1, \dots, N$  do # (cf. Definition 6.5.2)
3:    $\mathbb{M} \leftarrow \mathbb{M} + [(\nabla \mathcal{L})(\Theta)]^2$ 
4:    $\Theta \leftarrow \Theta - \gamma_n [\varepsilon + \mathbb{M}^{1/2}]^{-1} (\nabla \mathcal{L})(\Theta)$ 
5: return  $\Theta$ 

```

6.6 Root mean square propagation (RMSprop) optimization

In this section we review the *root mean square propagation* (RMSprop) GD optimization method. Roughly speaking, the RMSprop GD optimization method is a modification of the Adagrad GD optimization method where the sum over the squares of previous partial derivatives of the objective function (cf. (6.548) in Definition 6.5.1) is replaced by an exponentially decaying average over the squares of previous partial derivatives of the objective function (cf. (6.556) and (6.557) in Definition 6.6.1). RMSprop optimization was introduced by Geoffrey Hinton in his coursera class on *Neural Networks for Machine Learning* (see Hinton et al. [209]) in the context of stochastic optimization. As in the case of Adagrad optimization, we present for pedagogical purposes first a deterministic version of RMSprop optimization in this section and we refer to Section 7.7 below for the original stochastic version of RMSprop optimization.

Definition 6.6.1 (RMSprop GD optimization method). *Let $\mathfrak{d} \in \mathbb{N}$, let $\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$ be differentiable, let $\mathcal{G} = (\mathcal{G}_1, \dots, \mathcal{G}_{\mathfrak{d}}): \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that*

$$\mathcal{G}(\theta) = (\nabla \mathcal{L})(\theta), \quad (6.555)$$

let $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\beta_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $\varepsilon \in (0, \infty)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$, and let $\Theta = (\Theta^{(1)}, \dots, \Theta^{(\mathfrak{d})}): \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a function. Then we say that Θ is the RMSprop GD process for the objective function \mathcal{L} with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, second moment decay factors $(\beta_n)_{n \in \mathbb{N}}$, regularizing factor ε , and initial value ξ if and only if there exists $\mathbb{M} = (\mathbb{M}^{(1)}, \dots, \mathbb{M}^{(\mathfrak{d})}): \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ such that for all $n \in \mathbb{N}$, $i \in \{1, 2, \dots, \mathfrak{d}\}$ it holds that

$$\Theta_0 = \xi, \quad \mathbb{M}_0 = 0, \quad \mathbb{M}_n^{(i)} = \beta_n \mathbb{M}_{n-1}^{(i)} + (1 - \beta_n) |\mathcal{G}_i(\Theta_{n-1})|^2, \quad (6.556)$$

$$\text{and} \quad \Theta_n^{(i)} = \Theta_{n-1}^{(i)} - \gamma_n [\varepsilon + \mathbb{M}_n^{(i)}]^{1/2} \mathcal{G}_i(\Theta_{n-1}). \quad (6.557)$$

Algorithm 6.6.2: RMSprop GD optimization method

Input: $\mathfrak{d}, N \in \mathbb{N}$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\beta_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $\varepsilon \in (0, \infty)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$

Output: N -th step of the RMSprop GD process for the objective function \mathcal{L} with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, second moment decay factors $(\beta_n)_{n \in \mathbb{N}}$, regularizing factor ε , and initial value ξ (cf. Definition 6.6.1)

```

1: Initialization:  $\Theta \leftarrow \xi; \mathbb{M} \leftarrow 0 \in \mathbb{R}^{\mathfrak{d}}$ 
2: for  $n = 1, \dots, N$  do # (cf. Definition 6.5.2)
3:    $\mathbb{M} \leftarrow \beta_n \mathbb{M} + (1 - \beta_n)[(\nabla \mathcal{L})(\Theta)]^2$ 
4:    $\Theta \leftarrow \Theta - \gamma_n [\varepsilon + \mathbb{M}^{1/2}]^{-1} (\nabla \mathcal{L})(\Theta)$ 
5: return  $\Theta$ 

```

Remark 6.6.3. In Hinton et al. [209] it is proposed to choose $0.9 = \beta_1 = \beta_2 = \dots$ as default values for the second moment decay factors $(\beta_n)_{n \in \mathbb{N}} \subseteq [0, 1]$ in Definition 7.7.1.

Moreover, we note that in Hinton et al. [209] the regularizing factor ε is omitted in the definition of RMSprop. We chose the include the regularizing factor ε in the definition of RMSprop in Definition 6.6.1 as it is usually implemented in machine learning libraries (cf., e.g., [358]).

6.6.1 Representations of the mean square terms in RMSprop

Lemma 6.6.4 (On a representation of the second order terms in RMSprop). *Let $\mathfrak{d} \in \mathbb{N}$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\beta_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $(b_{n,k})_{(n,k) \in (\mathbb{N}_0)^2} \subseteq \mathbb{R}$, $\varepsilon \in (0, \infty)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$ satisfy for all $n \in \mathbb{N}$, $k \in \{0, 1, \dots, n-1\}$ that*

$$b_{n,k} = (1 - \beta_{k+1}) \left[\prod_{l=k+2}^n \beta_l \right], \quad (6.558)$$

let $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$, $\mathcal{G} = (\mathcal{G}_1, \dots, \mathcal{G}_{\mathfrak{d}}) \in C(\mathbb{R}^{\mathfrak{d}}, \mathbb{R}^{\mathfrak{d}})$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that

$$\mathcal{G}(\theta) = (\nabla \mathcal{L})(\theta), \quad (6.559)$$

and let $\Theta = (\Theta^{(1)}, \dots, \Theta^{(\mathfrak{d})}) : \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ be the RMSprop GD process for the objective function \mathcal{L} with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, second moment decay factors $(\beta_n)_{n \in \mathbb{N}}$, regularizing factor ε , and initial value ξ (cf. Definition 6.6.1). Then

(i) it holds for all $n \in \mathbb{N}$, $k \in \{0, 1, \dots, n-1\}$ that $0 \leq b_{n,k} \leq 1$,

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

$$\sum_{k=0}^{n-1} b_{n,k} = 1 - \prod_{k=1}^n \beta_k, \quad (6.560)$$

and

(iii) it holds for all $n \in \mathbb{N}$, $i \in \{1, 2, \dots, \mathfrak{d}\}$ that

$$\Theta_n^{(i)} = \Theta_{n-1}^{(i)} - \gamma_n \left[\varepsilon + \left[\sum_{k=0}^{n-1} b_{n,k} |\mathcal{G}_i(\Theta_k)|^2 \right]^{1/2} \right]^{-1} \mathcal{G}_i(\Theta_{n-1}). \quad (6.561)$$

Proof of Lemma 6.6.4. Throughout this proof, let $\mathbb{M} = (\mathbb{M}^{(1)}, \dots, \mathbb{M}^{(\mathfrak{d})}) : \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $n \in \mathbb{N}$, $i \in \{1, 2, \dots, \mathfrak{d}\}$ that $\mathbb{M}_0^{(i)} = 0$ and

$$\mathbb{M}_n^{(i)} = \beta_n \mathbb{M}_{n-1}^{(i)} + (1 - \beta_n) |\mathcal{G}_i(\Theta_{n-1})|^2. \quad (6.562)$$

Note that (6.558) implies item (i). Furthermore, observe that (6.558), (6.562), and Lemma 6.3.17 show that for all $n \in \mathbb{N}$, $i \in \{1, 2, \dots, \mathfrak{d}\}$ it holds that

$$\mathbb{M}_n^{(i)} = \sum_{k=0}^{n-1} b_{n,k} |\mathcal{G}_i(\Theta_k)|^2 \quad \text{and} \quad \sum_{k=0}^{n-1} b_{n,k} = 1 - \prod_{k=1}^n \beta_k. \quad (6.563)$$

This proves item (ii). Moreover, note that (6.556), (6.557), (6.562), and (6.563) ensure that for all $n \in \mathbb{N}$, $i \in \{1, 2, \dots, \mathfrak{d}\}$ it holds that

$$\begin{aligned} \Theta_n^{(i)} &= \Theta_{n-1}^{(i)} - \gamma_n \left[\varepsilon + [\mathbb{M}_n^{(i)}]^{1/2} \right]^{-1} \mathcal{G}_i(\Theta_{n-1}) \\ &= \Theta_{n-1}^{(i)} - \gamma_n \left[\varepsilon + \left[\sum_{k=0}^{n-1} b_{n,k} |\mathcal{G}_i(\Theta_k)|^2 \right]^{1/2} \right]^{-1} \mathcal{G}_i(\Theta_{n-1}). \end{aligned} \quad (6.564)$$

This establishes item (iii). The proof of Lemma 6.6.4 is thus complete. \square

6.6.2 Bias-adjusted RMSprop optimization

Definition 6.6.5 (Bias-adjusted RMSprop GD optimization method). Let $\mathfrak{d} \in \mathbb{N}$, let $\mathcal{L} : \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$ be differentiable, let $\mathcal{G} = (\mathcal{G}_1, \dots, \mathcal{G}_{\mathfrak{d}}) : \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that

$$\mathcal{G}(\theta) = (\nabla \mathcal{L})(\theta), \quad (6.565)$$

let $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\beta_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $\varepsilon \in (0, \infty)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$, and let $\Theta = (\Theta^{(1)}, \dots,$

$\Theta^{(\mathfrak{d})}: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a function. Then we say that Θ is the bias-adjusted RMSprop GD process for the objective function \mathcal{L} with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, second moment decay factors $(\beta_n)_{n \in \mathbb{N}}$, regularizing factor ε , and initial value ξ if and only if there exists $\mathbb{M} = (\mathbb{M}^{(1)}, \dots, \mathbb{M}^{(\mathfrak{d})}): \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ such that for all $n \in \mathbb{N}$, $i \in \{1, 2, \dots, \mathfrak{d}\}$ it holds that

$$\Theta_0 = \xi, \quad \mathbb{M}_0 = 0, \quad \mathbb{M}_n^{(i)} = \beta_n \mathbb{M}_{n-1}^{(i)} + (1 - \beta_n) |\mathcal{G}_i(\Theta_{n-1})|^2, \quad (6.566)$$

$$\text{and} \quad \Theta_n^{(i)} = \Theta_{n-1}^{(i)} - \gamma_n \left[\varepsilon + \left[\frac{\mathbb{M}_n^{(i)}}{1 - \prod_{k=1}^n \beta_k} \right]^{1/2} \right]^{-1} \mathcal{G}_i(\Theta_{n-1}). \quad (6.567)$$

Algorithm 6.6.6: Bias-adjusted RMSprop GD optimization method

Input: $\mathfrak{d}, N \in \mathbb{N}$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\beta_n)_{n \in \mathbb{N}} \subseteq [0, 1)$, $\varepsilon \in (0, \infty)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$

Output: N -th step of the bias-adjusted RMSprop GD process for the objective function \mathcal{L} with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, second moment decay factors $(\beta_n)_{n \in \mathbb{N}}$, regularizing factor ε , and initial value ξ (cf. Definition 6.6.5)

```

1: Initialization:  $\Theta \leftarrow \xi; \mathbb{M} \leftarrow 0 \in \mathbb{R}^{\mathfrak{d}}$ 
2: for  $n = 1, \dots, N$  do # (cf. Definition 6.5.2)
3:    $\mathbb{M} \leftarrow \beta_n \mathbb{M} + (1 - \beta_n) [(\nabla \mathcal{L})(\Theta)]^2$ 
4:    $\Theta \leftarrow \Theta - \gamma_n \left[ \varepsilon + \left[ \frac{\mathbb{M}}{1 - \prod_{k=1}^n \beta_k} \right]^{1/2} \right]^{-1} (\nabla \mathcal{L})(\Theta)$ 
5: return  $\Theta$ 

```

Lemma 6.6.7 (On a representation of the second order terms in bias-adjusted RMSprop). Let $\mathfrak{d} \in \mathbb{N}$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\beta_n)_{n \in \mathbb{N}} \subseteq [0, 1)$, $(b_{n,k})_{(n,k) \in (\mathbb{N}_0)^2} \subseteq \mathbb{R}$, $\varepsilon \in (0, \infty)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$ satisfy for all $n \in \mathbb{N}$, $k \in \{0, 1, \dots, n-1\}$ that

$$b_{n,k} = \frac{(1 - \beta_{k+1}) [\prod_{l=k+2}^n \beta_l]}{1 - \prod_{l=1}^n \beta_l}, \quad (6.568)$$

let $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$, $\mathcal{G} = (\mathcal{G}_1, \dots, \mathcal{G}_{\mathfrak{d}}) \in C(\mathbb{R}^{\mathfrak{d}}, \mathbb{R}^{\mathfrak{d}})$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that

$$\mathcal{G}(\theta) = (\nabla \mathcal{L})(\theta), \quad (6.569)$$

and let $\Theta = (\Theta^{(1)}, \dots, \Theta^{(\mathfrak{d})}): \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ be the bias-adjusted RMSprop GD process for the objective function \mathcal{L} with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, second moment decay factors $(\beta_n)_{n \in \mathbb{N}}$, regularizing factor ε , and initial value ξ (cf. Definition 6.6.5). Then

(i) it holds for all $n \in \mathbb{N}$, $k \in \{0, 1, \dots, n-1\}$ that $0 \leq b_{n,k} \leq 1$,

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

$$\sum_{k=0}^{n-1} b_{n,k} = 1, \quad (6.570)$$

and

(iii) it holds for all $n \in \mathbb{N}$, $i \in \{1, 2, \dots, \mathfrak{d}\}$ that

$$\Theta_n^{(i)} = \Theta_{n-1}^{(i)} - \gamma_n \left[\varepsilon + \left[\sum_{k=0}^{n-1} b_{n,k} |\mathcal{G}_i(\Theta_k)|^2 \right]^{1/2} \right]^{-1} \mathcal{G}_i(\Theta_{n-1}). \quad (6.571)$$

Proof of Lemma 6.6.7. Throughout this proof, let $\mathbb{M} = (\mathbb{M}^{(1)}, \dots, \mathbb{M}^{(\mathfrak{d})}) : \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $n \in \mathbb{N}$, $i \in \{1, 2, \dots, \mathfrak{d}\}$ that $\mathbb{M}_0^{(i)} = 0$ and

$$\mathbb{M}_n^{(i)} = \beta_n \mathbb{M}_{n-1}^{(i)} + (1 - \beta_n) |\mathcal{G}_i(\Theta_{n-1})|^2 \quad (6.572)$$

and let $(B_{n,k})_{(n,k) \in (\mathbb{N}_0)^2} \subseteq \mathbb{R}$ satisfy for all $n \in \mathbb{N}$, $k \in \{0, 1, \dots, n-1\}$ that

$$B_{n,k} = (1 - \beta_{k+1}) \left[\prod_{l=k+2}^n \beta_l \right]. \quad (6.573)$$

Observe that (6.568) proves item (i). Note that (6.568), (6.572), (6.573), and Lemma 6.3.17 establish that for all $n \in \mathbb{N}$, $i \in \{1, 2, \dots, \mathfrak{d}\}$ it holds that

$$\mathbb{M}_n^{(i)} = \sum_{k=0}^{n-1} B_{n,k} |\mathcal{G}_i(\Theta_k)|^2 \quad \text{and} \quad \sum_{k=0}^{n-1} b_{n,k} = \frac{\sum_{k=0}^{n-1} B_{n,k}}{1 - \prod_{l=1}^n \beta_l} = \frac{1 - \prod_{k=1}^n \beta_k}{1 - \prod_{k=1}^n \beta_k} = 1. \quad (6.574)$$

This proves item (ii). Observe that (6.566), (6.567), (6.572), and (6.574) demonstrate that for all $n \in \mathbb{N}$, $i \in \{1, 2, \dots, \mathfrak{d}\}$ it holds that

$$\begin{aligned} \Theta_n^{(i)} &= \Theta_{n-1}^{(i)} - \gamma_n \left[\varepsilon + \left[\frac{\mathbb{M}_n^{(i)}}{1 - \prod_{k=1}^n \beta_k} \right]^{1/2} \right]^{-1} \mathcal{G}_i(\Theta_{n-1}) \\ &= \Theta_{n-1}^{(i)} - \gamma_n \left[\varepsilon + \left[\frac{\sum_{k=0}^{n-1} B_{n,k} |\mathcal{G}_i(\Theta_k)|^2}{1 - \prod_{k=1}^n \beta_k} \right]^{1/2} \right]^{-1} \mathcal{G}_i(\Theta_{n-1}) \\ &= \Theta_{n-1}^{(i)} - \gamma_n \left[\varepsilon + \left[\sum_{k=0}^{n-1} \left(\frac{(1 - \beta_{k+1}) [\prod_{l=k+2}^n \beta_l]}{1 - \prod_{l=1}^n \beta_l} \right) |\mathcal{G}_i(\Theta_k)|^2 \right]^{1/2} \right]^{-1} \mathcal{G}_i(\Theta_{n-1}) \\ &= \Theta_{n-1}^{(i)} - \gamma_n \left[\varepsilon + \left[\sum_{k=0}^{n-1} b_{n,k} |\mathcal{G}_i(\Theta_k)|^2 \right]^{1/2} \right]^{-1} \mathcal{G}_i(\Theta_{n-1}). \end{aligned} \quad (6.575)$$

This establishes item (iii). The proof of Lemma 6.6.7 is thus complete. \square

6.7 Adadelta optimization

The Adadelta **GD** optimization method reviewed in this section is an extension of the RMSprop **GD** optimization method. Like the RMSprop **GD** optimization method, the Adadelta **GD** optimization method adapts the learning rates for every component of the optimization process separately. To do this, the Adadelta **GD** optimization method uses two exponentially decaying averages: one over the squares of the past partial derivatives of the objective function as does the RMSprop **GD** optimization method (cf. (6.578) below) and another one over the squares of the past increments (cf. (6.580) below). As in the case of Adagrad and RMSprop optimization, Adadelta optimization was introduced in a stochastic setting (see Zeiler [450]), but for pedagogical purposes we present in this section a deterministic version of Adadelta optimization. We refer to Section 7.8 below for the original stochastic version of Adadelta optimization.

Definition 6.7.1 (Adadelta **GD** optimization method). *Let $\mathfrak{d} \in \mathbb{N}$, let $\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$ be differentiable, let $\mathcal{G} = (\mathcal{G}_1, \dots, \mathcal{G}_{\mathfrak{d}}): \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that*

$$\mathcal{G}(\theta) = (\nabla \mathcal{L})(\theta), \quad (6.576)$$

*let $(\beta_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $(\delta_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $\varepsilon \in (0, \infty)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$, and let $\Theta = (\Theta^{(1)}, \dots, \Theta^{(\mathfrak{d})}): \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a function. Then we say that Θ is the Adadelta **GD** process for the objective function \mathcal{L} with second moment decay factors $(\beta_n)_{n \in \mathbb{N}}$, delta decay factors $(\delta_n)_{n \in \mathbb{N}}$, regularizing factor ε , and initial value ξ if and only if there exist $\mathbb{M} = (\mathbb{M}^{(1)}, \dots, \mathbb{M}^{(\mathfrak{d})}): \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ and $\Delta = (\Delta^{(1)}, \dots, \Delta^{(\mathfrak{d})}): \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ such that for all $n \in \mathbb{N}$, $i \in \{1, 2, \dots, \mathfrak{d}\}$ it holds that*

$$\Theta_0 = \xi, \quad \mathbb{M}_0 = 0, \quad \Delta_0 = 0, \quad (6.577)$$

$$\mathbb{M}_n^{(i)} = \beta_n \mathbb{M}_{n-1}^{(i)} + (1 - \beta_n) |\mathcal{G}_i(\Theta_{n-1})|^2, \quad (6.578)$$

$$\Theta_n^{(i)} = \Theta_{n-1}^{(i)} - \left[\frac{\varepsilon + \Delta_{n-1}^{(i)}}{\varepsilon + \mathbb{M}_n^{(i)}} \right]^{1/2} \mathcal{G}_i(\Theta_{n-1}), \quad (6.579)$$

$$\text{and} \quad \Delta_n^{(i)} = \delta_n \Delta_{n-1}^{(i)} + (1 - \delta_n) |\Theta_n^{(i)} - \Theta_{n-1}^{(i)}|^2. \quad (6.580)$$

Algorithm 6.7.2: Adadelta **GD optimization method**

Input: $\mathfrak{d}, N \in \mathbb{N}$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$, $(\beta_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $(\delta_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $\varepsilon \in (0, \infty)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$

Output: N -th step of the Adadelta **GD** process for the objective function \mathcal{L} with second moment decay factors $(\beta_n)_{n \in \mathbb{N}}$, delta decay factors $(\delta_n)_{n \in \mathbb{N}}$, regularizing factor ε , and initial value ξ (cf. Definition 6.7.1)

```

1: Initialization:  $\Theta \leftarrow \xi; \Xi \leftarrow \xi; \mathbb{M} \leftarrow 0 \in \mathbb{R}^{\mathfrak{d}}; \Delta \leftarrow 0 \in \mathbb{R}^{\mathfrak{d}}$ 
2: for  $n = 1, \dots, N$  do # (cf. Definition 6.5.2)
3:    $\mathbb{M} \leftarrow \beta_n \mathbb{M} + (1 - \beta_n)[(\nabla \mathcal{L})(\Theta)]^2$ 
4:    $\Theta \leftarrow \Theta - [\frac{\varepsilon + \Delta}{\varepsilon + \mathbb{M}}]^{1/2} (\nabla \mathcal{L})(\Theta)$ 
5:    $\Delta \leftarrow \delta_n \Delta + (1 - \delta_n)[\Theta - \Xi]^2$ 
6:    $\Xi \leftarrow \Theta$ 
7: return  $\Theta$ 

```

6.8 Adaptive moment estimation (Adam) optimization

In this section we introduce the *adaptive moment estimation* (**Adam**) **GD** optimization method (see Kingma & Ba [261]). Roughly speaking, the **Adam** **GD** optimization method can be viewed as a combination of the bias-adjusted momentum **GD** optimization method (see Section 6.3.4) and the bias-adjusted **RMSprop** **GD** optimization method (see Section 6.6.2). As in the case of previously considered optimization methods, **Adam** optimization was introduced in a stochastic setting in Kingma & Ba [261], but for pedagogical purposes we present in this section a deterministic version of **Adam** optimization. We refer to Section 7.9 below for the original stochastic version of **Adam** optimization.

Definition 6.8.1 (**Adam** **GD** optimization method). *Let $\mathfrak{d} \in \mathbb{N}$, let $\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$ be differentiable, let $\mathcal{G} = (\mathcal{G}_1, \dots, \mathcal{G}_{\mathfrak{d}}): \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that*

$$\mathcal{G}(\theta) = (\nabla \mathcal{L})(\theta), \quad (6.581)$$

*let $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $(\beta_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $\varepsilon \in (0, \infty)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$, and let $\Theta = (\Theta^{(1)}, \dots, \Theta^{(\mathfrak{d})}): \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a function. Then we say that Θ is the **Adam** **GD** process for the objective function \mathcal{L} with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, second moment decay factors $(\beta_n)_{n \in \mathbb{N}}$, regularizing factor ε , and initial value ξ if and only if there exist $\mathbf{m} = (\mathbf{m}^{(1)}, \dots, \mathbf{m}^{(\mathfrak{d})}): \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ and $\mathbb{M} = (\mathbb{M}^{(1)}, \dots, \mathbb{M}^{(\mathfrak{d})}): \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ such that for all $n \in \mathbb{N}$, $i \in \{1, 2, \dots, \mathfrak{d}\}$ it holds that*

$$\Theta_0 = \xi, \quad \mathbf{m}_0 = 0, \quad \mathbb{M}_0 = 0, \quad (6.582)$$

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + (1 - \alpha_n) \mathcal{G}(\Theta_{n-1}), \quad (6.583)$$

$$\mathbb{M}_n^{(i)} = \beta_n \mathbb{M}_{n-1}^{(i)} + (1 - \beta_n) |\mathcal{G}_i(\Theta_{n-1})|^2, \quad (6.584)$$

$$\text{and} \quad \Theta_n^{(i)} = \Theta_{n-1}^{(i)} - \gamma_n \left[\varepsilon + \left[\frac{\mathbb{M}_n^{(i)}}{(1 - \prod_{l=1}^n \beta_l)} \right]^{1/2} \right]^{-1} \left[\frac{\mathbf{m}_n^{(i)}}{(1 - \prod_{l=1}^n \alpha_l)} \right]. \quad (6.585)$$

Algorithm 6.8.2: Adam GD optimization method

Input: $\mathfrak{d}, N \in \mathbb{N}$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $(\beta_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $\varepsilon \in (0, \infty)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$

Output: N -th step of the Adam GD process for the objective function \mathcal{L} with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, second moment decay factors $(\beta_n)_{n \in \mathbb{N}}$, regularizing factor ε , and initial value ξ (cf. Definition 6.8.1)

```

1: Initialization:  $\Theta \leftarrow \xi; \mathbf{m} \leftarrow 0 \in \mathbb{R}^{\mathfrak{d}}; \mathbb{M} \leftarrow 0 \in \mathbb{R}^{\mathfrak{d}}$ 
2: for  $n = 1, \dots, N$  do # (cf. Definition 6.5.2)
3:    $\mathbf{m} \leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n) (\nabla \mathcal{L})(\Theta)$ 
4:    $\mathbb{M} \leftarrow \beta_n \mathbb{M} + (1 - \beta_n) [(\nabla \mathcal{L})(\Theta)]^2$ 
5:    $\Theta \leftarrow \Theta - \left[ \varepsilon + \left[ \frac{\mathbb{M}}{1 - \prod_{k=1}^n \beta_k} \right]^{1/2} \right]^{-1} \left[ \frac{\gamma_n \mathbf{m}}{1 - \prod_{k=1}^n \alpha_k} \right]$ 
6: return  $\Theta$ 

```

6.8.1 Adamax optimization

In this section we consider the deterministic Adamax GD optimization method which was introduced together with the Adam GD optimization method in Kingma & Ba [261]. We refer to Section 7.9.1 below for the original stochastic version of Adamax optimization.

Definition 6.8.3 (Adamax GD optimization method). *Let $\mathfrak{d} \in \mathbb{N}$, let $\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$ be differentiable, let $\mathcal{G} = (\mathcal{G}_1, \dots, \mathcal{G}_{\mathfrak{d}}): \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that*

$$\mathcal{G}(\theta) = (\nabla \mathcal{L})(\theta), \quad (6.586)$$

let $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $(\beta_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $\varepsilon \in (0, \infty)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$, and let $\Theta = (\Theta^{(1)}, \dots, \Theta^{(\mathfrak{d})}): \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a function. Then we say that Θ is the Adamax GD process for the objective function \mathcal{L} with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, second moment decay factors $(\beta_n)_{n \in \mathbb{N}}$, regularizing factor ε , and initial value ξ if and only if there exist $\mathbf{m} = (\mathbf{m}^{(1)}, \dots, \mathbf{m}^{(\mathfrak{d})}): \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ and $\mathbb{M} = (\mathbb{M}^{(1)}, \dots, \mathbb{M}^{(\mathfrak{d})}): \mathbb{N}_0 \rightarrow$

\mathbb{R}^d such that for all $n \in \mathbb{N}$, $i \in \{1, 2, \dots, d\}$ it holds that

$$\Theta_0 = \xi, \quad \mathbf{m}_0 = 0, \quad \mathbb{M}_0 = 0, \quad (6.587)$$

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + (1 - \alpha_n) \mathcal{G}(\Theta_{n-1}), \quad (6.588)$$

$$\mathbb{M}_n^{(i)} = \max\{\beta_n \mathbb{M}_{n-1}^{(i)}, |\mathcal{G}_i(\Theta_{n-1})|\}, \quad (6.589)$$

$$\text{and} \quad \Theta_n^{(i)} = \Theta_{n-1}^{(i)} - \gamma_n \left[\varepsilon + \mathbb{M}_n^{(i)} \right]^{-1} \left[\frac{\mathbf{m}_n^{(i)}}{(1 - \prod_{l=1}^n \alpha_l)} \right]. \quad (6.590)$$

Algorithm 6.8.4: Adamax GD optimization method

Input: $d, N \in \mathbb{N}$, $\mathcal{L} \in C^1(\mathbb{R}^d, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $(\beta_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $\varepsilon \in (0, \infty)$, $\xi \in \mathbb{R}^d$

Output: N -th step of the Adamax GD process for the objective function \mathcal{L} with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, second moment decay factors $(\beta_n)_{n \in \mathbb{N}}$, regularizing factor ε , and initial value ξ (cf. Definition 6.8.3)

```

1: Initialization:  $\Theta \leftarrow \xi; \mathbf{m} \leftarrow 0 \in \mathbb{R}^d; \mathbb{M} \leftarrow 0 \in \mathbb{R}^d$ 
2: for  $n = 1, \dots, N$  do # (cf. Definition 6.5.2)
3:    $\mathbf{m} \leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n) (\nabla \mathcal{L})(\Theta)$ 
4:    $\mathbb{M} \leftarrow \max\{\beta_n \mathbb{M}, |(\nabla \mathcal{L})(\Theta)|\}$ 
5:    $\Theta \leftarrow \Theta - \gamma_n [\varepsilon + \mathbb{M}]^{-1} \left[ \frac{\mathbf{m}}{1 - \prod_{k=1}^n \alpha_k} \right]$ 
6: return  $\Theta$ 

```

6.9 Nesterov accelerated adaptive moment estimation (Nadam) optimization

In this section we review the *Nesterov-accelerated adaptive moment estimation* (Nadam) GD optimization method (cf. Dozat [117, 118]). Roughly speaking, the Nadam GD optimization method can be viewed as a combination of the shifted bias-adjusted Nesterov GD optimization method in Definition 6.4.31 and the bias-adjusted RMSprop GD optimization method in Definition 6.6.5. Alternatively, it can be seen as adding Nesterov acceleration to the Adam GD optimization method in Definition 6.8.1. As in the case of previously considered optimization methods, the Nadam GD optimization method was introduced in a stochastic setting in Dozat [117, 118], but for pedagogical purposes we present in this section a deterministic version of Nadam optimization. We refer to Section 7.10 below for the original stochastic version of Nadam optimization.

Definition 6.9.1 (Nadam GD optimization method). Let $\mathfrak{d} \in \mathbb{N}$, let $\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$ be differentiable, let $\mathcal{G} = (\mathcal{G}_1, \dots, \mathcal{G}_{\mathfrak{d}}): \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that

$$\mathcal{G}(\theta) = (\nabla \mathcal{L})(\theta), \quad (6.591)$$

let $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, 1)$, $(\beta_n)_{n \in \mathbb{N}} \subseteq [0, 1)$, $\varepsilon \in (0, \infty)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$, and let $\Theta = (\Theta^{(1)}, \dots, \Theta^{(\mathfrak{d})}): \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a function. Then we say that Θ is the Nadam GD process for the objective function \mathcal{L} with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, second moment decay factors $(\beta_n)_{n \in \mathbb{N}}$, regularizing factor ε , and initial value ξ if and only if there exist $\mathbf{m} = (\mathbf{m}^{(1)}, \dots, \mathbf{m}^{(\mathfrak{d})}): \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ and $\mathbb{M} = (\mathbb{M}^{(1)}, \dots, \mathbb{M}^{(\mathfrak{d})}): \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ such that for all $n \in \mathbb{N}$, $i \in \{1, 2, \dots, \mathfrak{d}\}$ it holds that

$$\Theta_0 = \xi, \quad \mathbf{m}_0 = 0, \quad \mathbb{M}_0 = 0, \quad (6.592)$$

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + (1 - \alpha_n) \mathcal{G}(\Theta_{n-1}), \quad (6.593)$$

$$\mathbb{M}_n^{(i)} = \beta_n \mathbb{M}_{n-1}^{(i)} + (1 - \beta_n) |\mathcal{G}_i(\Theta_{n-1})|^2, \quad \text{and} \quad (6.594)$$

$$\Theta_n^{(i)} = \Theta_{n-1}^{(i)} - \left[\varepsilon + \left[\frac{\mathbb{M}_n^{(i)}}{1 - \prod_{l=1}^n \beta_l} \right]^{1/2} \right]^{-1} \left[\left[\frac{\gamma_n(1-\alpha_n)}{1 - \prod_{l=1}^n \alpha_l} \right] \mathcal{G}_i(\Theta_{n-1}) + \left[\frac{\gamma_{n+1}\alpha_{n+1}}{1 - \prod_{l=1}^{n+1} \alpha_l} \right] \mathbf{m}_n^{(i)} \right]. \quad (6.595)$$

Algorithm 6.9.2: Nadam GD optimization method

Input: $\mathfrak{d}, N \in \mathbb{N}$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, 1)$, $(\beta_n)_{n \in \mathbb{N}} \subseteq [0, 1)$, $\varepsilon \in (0, \infty)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$

Output: N -th step of the Nadam GD process for the objective function \mathcal{L} with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, second moment decay factors $(\beta_n)_{n \in \mathbb{N}}$, regularizing factor ε , and initial value ξ (cf. Definition 6.9.1)

```

1: Initialization:  $\Theta \leftarrow \xi; \mathbf{m} \leftarrow 0 \in \mathbb{R}^{\mathfrak{d}}; \mathbb{M} \leftarrow 0 \in \mathbb{R}^{\mathfrak{d}}$ 
2: for  $n = 1, \dots, N$  do # (cf. Definition 6.5.2)
3:    $\mathbf{m} \leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n) (\nabla \mathcal{L})(\Theta)$ 
4:    $\mathbb{M} \leftarrow \beta_n \mathbb{M} + (1 - \beta_n) [(\nabla \mathcal{L})(\Theta)]^2$ 
5:    $\Theta \leftarrow \Theta - \left[ \varepsilon + \left[ \frac{\mathbb{M}}{1 - \prod_{k=1}^n \beta_k} \right]^{1/2} \right]^{-1} \left[ \left[ \frac{\gamma_n(1-\alpha_n)}{1 - \prod_{k=1}^n \alpha_k} \right] (\nabla \mathcal{L})(\Theta) + \left[ \frac{\gamma_{n+1}\alpha_{n+1}}{1 - \prod_{k=1}^{n+1} \alpha_k} \right] \mathbf{m} \right]$ 
6: return  $\Theta$ 

```

6.9.1 Simplified Nadam optimization

For reasons of algorithmic simplicity, in Dozat [117, 118] a simplified version of the Nadam GD optimization method has been proposed. Roughly speaking, the simplified version

presented in Definition 6.9.3 below is obtained by reducing the index of the second learning rate in the update step of the Nadam GD optimization method in Definition 6.9.1 above to ensure that for each update step only one learning rate is used.

Definition 6.9.3 (Simplified Nadam GD optimization method). *Let $\mathfrak{d} \in \mathbb{N}$, let $\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$ be differentiable, let $\mathcal{G} = (\mathcal{G}_1, \dots, \mathcal{G}_{\mathfrak{d}}): \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that*

$$\mathcal{G}(\theta) = (\nabla \mathcal{L})(\theta), \quad (6.596)$$

let $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, 1)$, $(\beta_n)_{n \in \mathbb{N}} \subseteq [0, 1)$, $\varepsilon \in (0, \infty)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$, and let $\Theta = (\Theta^{(1)}, \dots, \Theta^{(\mathfrak{d})}): \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a function. Then we say that Θ is the simplified Nadam GD process for the objective function \mathcal{L} with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, second moment decay factors $(\beta_n)_{n \in \mathbb{N}}$, regularizing factor ε , and initial value ξ if and only if there exist $\mathbf{m} = (\mathbf{m}^{(1)}, \dots, \mathbf{m}^{(\mathfrak{d})}): \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ and $\mathbb{M} = (\mathbb{M}^{(1)}, \dots, \mathbb{M}^{(\mathfrak{d})}): \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ such that for all $n \in \mathbb{N}$, $i \in \{1, 2, \dots, \mathfrak{d}\}$ it holds that

$$\Theta_0 = \xi, \quad \mathbf{m}_0 = 0, \quad \mathbb{M}_0 = 0, \quad (6.597)$$

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + (1 - \alpha_n) \mathcal{G}(\Theta_{n-1}), \quad (6.598)$$

$$\mathbb{M}_n^{(i)} = \beta_n \mathbb{M}_{n-1}^{(i)} + (1 - \beta_n) |\mathcal{G}_i(\Theta_{n-1})|^2, \quad \text{and} \quad (6.599)$$

$$\Theta_n^{(i)} = \Theta_{n-1}^{(i)} - \gamma_n \left[\varepsilon + \left[\frac{\mathbb{M}_n^{(i)}}{(1 - \prod_{l=1}^n \beta_l)} \right]^{1/2} \right]^{-1} \left[\left[\frac{1 - \alpha_n}{1 - \prod_{l=1}^n \alpha_l} \right] \mathcal{G}_i(\Theta_{n-1}) + \left[\frac{\alpha_{n+1}}{1 - \prod_{l=1}^{n+1} \alpha_l} \right] \mathbf{m}_n^{(i)} \right]. \quad (6.600)$$

Algorithm 6.9.4: Simplified Nadam GD optimization method

Input: $\mathfrak{d}, N \in \mathbb{N}$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, 1)$, $(\beta_n)_{n \in \mathbb{N}} \subseteq [0, 1)$, $\varepsilon \in (0, \infty)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$

Output: N -th step of the simplified Simplified Nadam GD process for the objective function \mathcal{L} with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, second moment decay factors $(\beta_n)_{n \in \mathbb{N}}$, regularizing factor ε , and initial value ξ (cf. Definition 6.9.3)

```

1: Initialization:  $\Theta \leftarrow \xi; \mathbf{m} \leftarrow 0 \in \mathbb{R}^{\mathfrak{d}}; \mathbb{M} \leftarrow 0 \in \mathbb{R}^{\mathfrak{d}}$ 
2: for  $n = 1, \dots, N$  do # (cf. Definition 6.5.2)
3:    $\mathbf{m} \leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n) (\nabla \mathcal{L})(\Theta)$ 
4:    $\mathbb{M} \leftarrow \beta_n \mathbb{M} + (1 - \beta_n) [(\nabla \mathcal{L})(\Theta)]^2$ 
5:    $\Theta \leftarrow \Theta - \gamma_n \left[ \varepsilon + \left[ \frac{\mathbb{M}}{1 - \prod_{k=1}^n \beta_k} \right]^{1/2} \right]^{-1} \left[ \left[ \frac{1 - \alpha_n}{1 - \prod_{k=1}^n \alpha_k} \right] (\nabla \mathcal{L})(\Theta) + \left[ \frac{\alpha_{n+1}}{1 - \prod_{k=1}^{n+1} \alpha_k} \right] \mathbf{m} \right]$ 
6: return  $\Theta$ 

```

6.9.2 Nadamax optimization

In this section we consider the deterministic Nadamax **GD** optimization method which was introduced together with the **Nadam GD** optimization method in Dozat [117, 118]. We refer to Section 7.10.2 below for the original stochastic version of Nadamax optimization.

Definition 6.9.5 (Nadamax **GD** optimization method). *Let $\mathfrak{d} \in \mathbb{N}$, let $\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$ be differentiable, let $\mathcal{G} = (\mathcal{G}_1, \dots, \mathcal{G}_{\mathfrak{d}}): \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that*

$$\mathcal{G}(\theta) = (\nabla \mathcal{L})(\theta), \quad (6.601)$$

*let $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $(\beta_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $\varepsilon \in (0, \infty)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$, and let $\Theta = (\Theta^{(1)}, \dots, \Theta^{(\mathfrak{d})}): \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a function. Then we say that Θ is the Nadamax **GD** process for the objective function \mathcal{L} with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, second moment decay factors $(\beta_n)_{n \in \mathbb{N}}$, regularizing factor ε , and initial value ξ if and only if there exist $\mathbf{m}: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ and $\mathbb{M}: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ such that for all $n \in \mathbb{N}$, $i \in \{1, 2, \dots, \mathfrak{d}\}$ it holds that*

$$\Theta_0 = \xi, \quad \mathbf{m}_0 = 0, \quad \mathbb{M}_0 = 0, \quad (6.602)$$

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + (1 - \alpha_n) \mathcal{G}(\Theta_{n-1}), \quad (6.603)$$

$$\mathbb{M}_n^{(i)} = \max\{\beta_n \mathbb{M}_{n-1}^{(i)}, |\mathcal{G}_i(\Theta_{n-1})|\}, \quad \text{and} \quad (6.604)$$

$$\Theta_n^{(i)} = \Theta_{n-1}^{(i)} - [\varepsilon + \mathbb{M}_n^{(i)}]^{-1} \left[\left[\frac{\gamma_n(1-\alpha_n)}{1 - \prod_{l=1}^n \alpha_l} \right] \mathcal{G}_i(\Theta_{n-1}) + \left[\frac{\gamma_{n+1}\alpha_{n+1}}{1 - \prod_{l=1}^{n+1} \alpha_l} \right] \mathbf{m}_n^{(i)} \right]. \quad (6.605)$$

Algorithm 6.9.6: Nadamax **GD** optimization method

Input: $\mathfrak{d}, N \in \mathbb{N}$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $(\beta_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $\varepsilon \in (0, \infty)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$

Output: N -th step of the Nadamax **GD** process for the objective function \mathcal{L} with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, second moment decay factors $(\beta_n)_{n \in \mathbb{N}}$, regularizing factor ε , and initial value ξ (cf. Definition 6.9.5)

- 1: **Initialization:** $\Theta \leftarrow \xi; \mathbf{m} \leftarrow 0 \in \mathbb{R}^{\mathfrak{d}}; \mathbb{M} \leftarrow 0 \in \mathbb{R}^{\mathfrak{d}}$
- 2: **for** $n = 1, \dots, N$ **do** *# (cf. Definition 6.5.2)*
- 3: $\mathbf{m} \leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n) (\nabla \mathcal{L})(\Theta)$
- 4: $\mathbb{M} \leftarrow \max\{\beta_n \mathbb{M}, |(\nabla \mathcal{L})(\Theta)|\}$
- 5: $\Theta \leftarrow \Theta - [\varepsilon + \mathbb{M}]^{-1} \left[\left[\frac{\gamma_n(1-\alpha_n)}{1 - \prod_{l=1}^n \alpha_l} \right] (\nabla \mathcal{L})(\Theta) + \left[\frac{\gamma_{n+1}\alpha_{n+1}}{1 - \prod_{l=1}^{n+1} \alpha_l} \right] \mathbf{m} \right]$
- 6: **return** Θ

6.10 Adam with decoupled weight decay (AdamW) optimization

In this section we introduce the *Adam with decoupled weight decay* (AdamW) GD optimization method (see Loshchilov & Hutter [300]). Roughly speaking, the AdamW GD optimization method can be viewed as modification of the Adam GD optimization method in Section 6.8 with a weight decay term added to the update step. As in the case of previously considered optimization methods, the AdamW GD optimization method was introduced in a stochastic setting in Loshchilov & Hutter [300], but for pedagogical purposes we present in this section a deterministic version of AdamW optimization. We refer to Section 7.11 below for the original stochastic version of AdamW optimization.

Definition 6.10.1 (AdamW GD optimization method). *Let $\mathfrak{d} \in \mathbb{N}$, let $\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$ be differentiable, let $\mathcal{G} = (\mathcal{G}_1, \dots, \mathcal{G}_{\mathfrak{d}}): \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that*

$$\mathcal{G}(\theta) = (\nabla \mathcal{L})(\theta), \quad (6.606)$$

let $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, 1)$, $(\beta_n)_{n \in \mathbb{N}} \subseteq [0, 1)$, $\lambda \in \mathbb{R}$, $\varepsilon \in (0, \infty)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$, and let $\Theta = (\Theta^{(1)}, \dots, \Theta^{(\mathfrak{d})}): \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a function. Then we say that Θ is the AdamW GD process for the objective function \mathcal{L} with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, second moment decay factors $(\beta_n)_{n \in \mathbb{N}}$, weight decay factor λ , regularizing factor ε , and initial value ξ if and only if there exist $\mathbf{m} = (\mathbf{m}^{(1)}, \dots, \mathbf{m}^{(\mathfrak{d})}): \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ and $\mathbb{M} = (\mathbb{M}^{(1)}, \dots, \mathbb{M}^{(\mathfrak{d})}): \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ such that for all $n \in \mathbb{N}$, $i \in \{1, 2, \dots, \mathfrak{d}\}$ it holds that

$$\Theta_0 = \xi, \quad \mathbf{m}_0 = 0, \quad \mathbb{M}_0 = 0, \quad (6.607)$$

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + (1 - \alpha_n) \mathcal{G}(\Theta_{n-1}), \quad (6.608)$$

$$\mathbb{M}_n^{(i)} = \beta_n \mathbb{M}_{n-1}^{(i)} + (1 - \beta_n) |\mathcal{G}_i(\Theta_{n-1})|^2, \quad \text{and} \quad (6.609)$$

$$\Theta_n^{(i)} = \Theta_{n-1}^{(i)} - \gamma_n \left(\left[\varepsilon + \left[\frac{\mathbb{M}_n^{(i)}}{(1 - \prod_{l=1}^n \beta_l)} \right]^{1/2} \right]^{-1} \left[\frac{\mathbf{m}_n^{(i)}}{(1 - \prod_{l=1}^n \alpha_l)} \right] + \lambda \Theta_{n-1}^{(i)} \right). \quad (6.610)$$

Algorithm 6.10.2: AdamW GD optimization method

Input: $\mathfrak{d}, N \in \mathbb{N}$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, 1)$, $(\beta_n)_{n \in \mathbb{N}} \subseteq [0, 1)$, $\lambda \in \mathbb{R}$, $\varepsilon \in (0, \infty)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$, $\xi \in \mathbb{R}^{\mathfrak{d}}$

Output: N -th step of the AdamW GD process for the objective function \mathcal{L} with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, second moment decay factors $(\beta_n)_{n \in \mathbb{N}}$, weight decay factor λ , regularizing factor ε , and initial value ξ (cf. Definition 6.10.1)

```

1: Initialization:  $\Theta \leftarrow \xi; \mathbf{m} \leftarrow 0 \in \mathbb{R}^{\mathfrak{d}}; \mathbb{M} \leftarrow 0 \in \mathbb{R}^{\mathfrak{d}}$ 
2: for  $n = 1, \dots, N$  do # (cf. Definition 6.5.2)
3:    $\mathbf{m} \leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n)(\nabla \mathcal{L})(\Theta)$ 
4:    $\mathbb{M} \leftarrow \beta_n \mathbb{M} + (1 - \beta_n)[(\nabla \mathcal{L})(\Theta)]^2$ 
5:    $\Theta \leftarrow \Theta - \gamma_n \left( \left[ \varepsilon + \left[ \frac{\mathbb{M}}{1 - \prod_{k=1}^n \beta_k} \right]^{1/2} \right]^{-1} \left[ \frac{\mathbf{m}}{1 - \prod_{k=1}^n \alpha_k} \right] + \lambda \Theta \right)$ 
6: return  $\Theta$ 

```

6.10.1 Adam with L^2 -regularization optimization

As an alternative way to regularize the parameters in the [Adam GD](#) optimization method, in Loshchilov & Hutter [300] it was also suggested to add a L^2 -regularization term to the objective function. This results in the [Adam GD](#) optimization method with L^2 -regularization in [Definition 6.10.3](#) below. We refer to [Section 7.11.1](#) below for the original stochastic version of [Adam GD](#) optimization with L^2 -regularization.

Definition 6.10.3 ([Adam GD](#) optimization method with L^2 -regularization). *Let $\mathfrak{d} \in \mathbb{N}$, let $\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$ be differentiable, let $\mathcal{G} = (\mathcal{G}_1, \dots, \mathcal{G}_{\mathfrak{d}}): \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that*

$$\mathcal{G}(\theta) = (\nabla \mathcal{L})(\theta), \quad (6.611)$$

let $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $(\beta_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $\lambda \in \mathbb{R}$, $\varepsilon \in (0, \infty)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$, and let $\Theta = (\Theta^{(1)}, \dots, \Theta^{(\mathfrak{d})}): \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a function. Then we say that Θ is the [Adam GD](#) process for the objective function \mathcal{L} with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, second moment decay factors $(\beta_n)_{n \in \mathbb{N}}$, L^2 -regularization factor λ , regularizing factor ε , and initial value ξ if and only if there exist $\mathbf{m} = (\mathbf{m}^{(1)}, \dots, \mathbf{m}^{(\mathfrak{d})}): \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ and $\mathbb{M} = (\mathbb{M}^{(1)}, \dots, \mathbb{M}^{(\mathfrak{d})}): \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ such that for all $n \in \mathbb{N}$, $i \in \{1, 2, \dots, \mathfrak{d}\}$ it holds that

$$\Theta_0 = \xi, \quad \mathbf{m}_0 = 0, \quad \mathbb{M}_0 = 0, \quad (6.612)$$

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + (1 - \alpha_n)(\mathcal{G}(\Theta_{n-1}) + \lambda \Theta_{n-1}), \quad (6.613)$$

$$\mathbb{M}_n^{(i)} = \beta_n \mathbb{M}_{n-1}^{(i)} + (1 - \beta_n) |\mathcal{G}_i(\Theta_{n-1}) + \lambda \Theta_{n-1}|^2, \quad (6.614)$$

$$\text{and } \Theta_n^{(i)} = \Theta_{n-1}^{(i)} - \gamma_n \left[\varepsilon + \left[\frac{\mathbb{M}_n^{(i)}}{(1 - \prod_{l=1}^n \beta_l)} \right]^{1/2} \right]^{-1} \left[\frac{\mathbf{m}_n^{(i)}}{(1 - \prod_{l=1}^n \alpha_l)} \right]. \quad (6.615)$$

Algorithm 6.10.4: Adam GD optimization method with L^2 -regularization

Input: $\mathfrak{d}, N \in \mathbb{N}$, $\mathcal{L} \in C^1(\mathbb{R}^\mathfrak{d}, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $(\beta_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $\lambda \in \mathbb{R}$, $\varepsilon \in (0, \infty)$, $\xi \in \mathbb{R}^\mathfrak{d}$

Output: N -th step of the Adam GD process for the objective function \mathcal{L} with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, second moment decay factors $(\beta_n)_{n \in \mathbb{N}}$, L^2 -regularization factor λ , regularizing factor ε , and initial value ξ (cf. Definition 6.10.3)

```

1: Initialization:  $\Theta \leftarrow \xi; \mathbf{m} \leftarrow 0 \in \mathbb{R}^\mathfrak{d}; \mathbb{M} \leftarrow 0 \in \mathbb{R}^\mathfrak{d}$ 
2: for  $n = 1, \dots, N$  do # (cf. Definition 6.5.2)
3:    $\mathbf{m} \leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n)((\nabla \mathcal{L})(\Theta) + \lambda \Theta)$ 
4:    $\mathbb{M} \leftarrow \beta_n \mathbb{M} + (1 - \beta_n)[(\nabla \mathcal{L})(\Theta) + \lambda \Theta]^2$ 
5:    $\Theta \leftarrow \Theta - \left[ \varepsilon + \left[ \frac{\mathbb{M}}{1 - \prod_{k=1}^n \beta_k} \right]^{1/2} \right]^{-1} \left[ \frac{\gamma_n \mathbf{m}}{1 - \prod_{k=1}^n \alpha_k} \right]$ 
6: return  $\Theta$ 

```

6.11 Shampoo optimization

In this section we introduce the Shampoo GD optimization method (cf. Gupta et al. [194]). The Shampoo GD optimization method was introduced in Gupta et al. [194] for loss functions which are defined on spaces of multidimensional matrices (sometimes called tensors). However, for simplicity we present in this section the Shampoo GD optimization method only in the case of loss functions which are defined on spaces of (two-dimensional) matrices (cf. Definition 6.11.4 below). Roughly speaking, the Shampoo GD optimization method aims to apply suitable preconditioners to transform the gradient of the loss function before doing a step in the direction of the negative gradient. The use of preconditioners motivated the authors in Gupta et al. [194] to name their method *Shampoo*. We refer to Section 7.12 below for the stochastic version of Shampoo optimization.

To ensure that the preconditioners used in the definition of the Shampoo GD optimization method in Definition 6.11.4 are well-defined, we first show in Lemma 6.11.2 and Corollary 6.11.3 below that sums of symmetric positive definite matrices remain symmetric positive definite and thus their roots are well-defined.

Definition 6.11.1 (Symmetric positive definite matrix). *Let $d \in \mathbb{N}$, $A \in \mathbb{R}^{d \times d}$. Then we say that A is a symmetric positive definite matrix if and only if it holds for all*

$v \in \mathbb{R}^d \setminus \{0\}$ that

$$A^* = A \quad \text{and} \quad \langle Av, v \rangle > 0 \quad (6.616)$$

(cf. Definition 1.4.7).

Lemma 6.11.2. Let $n, m \in \mathbb{N}$, let $A \in \mathbb{R}^{n \times n}$ be a symmetric positive definite matrix, and let $B \in \mathbb{R}^{n \times m}$ (cf. Definition 6.11.1). Then $A + B^*B$ is a symmetric positive definite matrix.

Proof of Lemma 6.11.2. Note that the assumption that A is a symmetric positive definite matrix implies that for all $v \in \mathbb{R}^n \setminus \{0\}$ it holds that

$$\langle (A + B^*B)v, v \rangle = \langle Av, v \rangle + \langle B^*Bv, v \rangle = \langle Av, v \rangle + \langle Bv, Bv \rangle = \langle Av, v \rangle + \|Bv\|_2^2 > 0 \quad (6.617)$$

(cf. Definitions 1.4.7 and 3.3.4). Moreover, observe that the fact that A is a symmetric positive definite matrix shows that

$$(A + B^*B)^* = A^* + (B^*B)^* = A + B^*B. \quad (6.618)$$

This and (6.617) ensure that $A + B^*B$ is a symmetric positive definite matrix. The proof of Lemma 6.11.2 is thus complete. \square

Corollary 6.11.3. Let $n, m \in \mathbb{N}$, $\varepsilon \in (0, \infty)$, let $(B_k)_{k \in \mathbb{N}_0} \subseteq \mathbb{R}^{n \times m}$, and let $(A_k)_{k \in \mathbb{N}_0} \subseteq \mathbb{R}^{n \times n}$ satisfy for all $k \in \mathbb{N}$ that $A_0 = \varepsilon I_n$ and $A_k = A_{k-1} + (B_{k-1})^*B_{k-1}$. Then it holds for all $k \in \mathbb{N}_0$ that A_k is a symmetric positive definite matrix (cf. Definitions 1.5.5 and 6.11.1).

Proof of Corollary 6.11.3. Note that induction and Lemma 6.11.2 prove that for all $k \in \mathbb{N}_0$ it holds that A_k is a symmetric positive definite matrix (cf. Definition 6.11.1). The proof of Corollary 6.11.3 is thus complete. \square

Definition 6.11.4 (Shampoo GD optimization method). Let $\mathfrak{d}_1, \mathfrak{d}_2 \in \mathbb{N}$, let $\mathcal{L}: \mathbb{R}^{\mathfrak{d}_1 \times \mathfrak{d}_2} \rightarrow \mathbb{R}$ be differentiable, let $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $\varepsilon \in (0, \infty)$, $\xi \in \mathbb{R}^{\mathfrak{d}_1 \times \mathfrak{d}_2}$, and let $\Theta: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}_1 \times \mathfrak{d}_2}$ be a function. Then we say that Θ is the Shampoo GD process for the objective function \mathcal{L} with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, regularizing factor ε , and initial value ξ if and only if there exist $\mathbf{L}: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}_1 \times \mathfrak{d}_1}$ and $\mathbf{R}: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}_2 \times \mathfrak{d}_2}$ such that for all $n \in \mathbb{N}$ it holds that

$$\Theta_0 = \xi, \quad \mathbf{L}_0 = \varepsilon I_{\mathfrak{d}_1}, \quad \mathbf{R}_0 = \varepsilon I_{\mathfrak{d}_2}, \quad (6.619)$$

$$\mathbf{L}_n = \mathbf{L}_{n-1} + (\nabla \mathcal{L})(\Theta_{n-1})((\nabla \mathcal{L})(\Theta_{n-1}))^*, \quad (6.620)$$

$$\mathbf{R}_n = \mathbf{R}_{n-1} + ((\nabla \mathcal{L})(\Theta_{n-1}))^*(\nabla \mathcal{L})(\Theta_{n-1}), \quad (6.621)$$

and $\Theta_n = \Theta_{n-1} - \gamma_n(\mathbf{L}_n)^{-1/4}(\nabla \mathcal{L})(\Theta_{n-1})(\mathbf{R}_n)^{-1/4}$ (6.622)
 (cf. Definition 1.5.5 and Corollary 6.11.3).

Algorithm 6.11.5: Shampoo GD optimization method

Input: $\mathfrak{d}_1, \mathfrak{d}_2, N \in \mathbb{N}, \mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}_1} \times \mathbb{R}^{\mathfrak{d}_2}, \mathbb{R}), (\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty), \varepsilon \in (0, \infty), \xi \in \mathbb{R}^{\mathfrak{d}_1 \times \mathfrak{d}_2}$
Output: N -th step of the Shampoo GD process for the objective function \mathcal{L} with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, regularizing factor ε , and initial value ξ (cf. Definition 6.11.4)

```

1: Initialization:  $\Theta \leftarrow \xi; \mathbf{L} \leftarrow \varepsilon I_{\mathfrak{d}_1}; \mathbf{R} \leftarrow \varepsilon I_{\mathfrak{d}_2}$ 
2: for  $n = 1, \dots, N$  do
3:    $\mathbf{L} \leftarrow \mathbf{L} + (\nabla \mathcal{L})(\Theta)((\nabla \mathcal{L})(\Theta))^*$ 
4:    $\mathbf{R} \leftarrow \mathbf{R} + ((\nabla \mathcal{L})(\Theta))^*(\nabla \mathcal{L})(\Theta)$ 
5:    $\Theta \leftarrow \Theta - \gamma_n(\mathbf{L})^{-1/4}(\nabla \mathcal{L})(\Theta)(\mathbf{R})^{-1/4}$ 
6: return  $\Theta$ 
```

6.12 Momentum orthogonalized by Newton-Schulz (Muon) optimization

In this section we review the *momentum orthogonalized by Newton-Schulz (Muon)* GD optimization method (cf. Jordan et al. [245]). The Muon GD optimization method in Jordan et al. [245] (cf. Definition 6.12.5 below) employs the Newton-Schulz method (cf. Definition 6.12.4 below) to approximately orthogonalize momentum terms. We note that the version of the momentum GD optimization method considered in Jordan et al. [245] (and Definition 6.12.5 below respectively) corresponds to the 2nd version of the momentum GD optimization method introduced in this book (cf. Definition 6.3.3 above). For work related to the Muon GD optimization method we refer, for instance, to [37, 282, 292]. We refer to Section 7.13 below for the stochastic version of Muon optimization.

To better understand the aim of the Muon GD optimization method we first define an idealized version of the Muon GD optimization method in Definition 6.12.2 which employs an exact orthogonalization of momentum terms instead of an approximation by the Newton-Schulz method.

Definition 6.12.1 (Hilbert-Schmidt norm). *Let $n, m \in \mathbb{N}$, $A = (A_{i,j})_{(i,j) \in \{1, \dots, n\} \times \{1, \dots, m\}} \in \mathbb{R}^{n \times m}$. Then we denote by $\|A\|_{HS} \in \mathbb{R}$ the real number given by*

$$\|A\|_{HS} = \left[\sum_{i=1}^n \sum_{j=1}^m |A_{i,j}|^2 \right]^{1/2}. \quad (6.623)$$

Definition 6.12.2 (Muon GD optimization method). Let $\mathfrak{d}_1, \mathfrak{d}_2 \in \mathbb{N}$, let $\mathcal{L}: \mathbb{R}^{\mathfrak{d}_1 \times \mathfrak{d}_2} \rightarrow \mathbb{R}$ be differentiable, let $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $\xi \in \mathbb{R}^{\mathfrak{d}_1 \times \mathfrak{d}_2}$, $\mathcal{O} = \{O \in \mathbb{R}^{\mathfrak{d}_1 \times \mathfrak{d}_2} : ((OO^* = I_{\mathfrak{d}_1}) \vee (O^*O = I_{\mathfrak{d}_2}))\}$, let $\Pi: \mathbb{R}^{\mathfrak{d}_1 \times \mathfrak{d}_2} \rightarrow \mathcal{O}$ satisfy for all $A \in \mathbb{R}^{\mathfrak{d}_1 \times \mathfrak{d}_2}$ that

$$\|\Pi(A) - A\|_{HS} = \inf_{O \in \mathcal{O}} \|O - A\|_{HS}, \quad (6.624)$$

and let $\Theta: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}_1 \times \mathfrak{d}_2}$ be a function (cf. Definitions 1.5.5 and 6.12.1). Then we say that Θ is the Muon GD process for the objective function \mathcal{L} with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, orthogonal projection Π , and initial value ξ if and only if there exists $\mathbf{m}: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}_1 \times \mathfrak{d}_2}$ such that for all $n \in \mathbb{N}$ it holds that

$$\Theta_0 = \xi, \quad \mathbf{m}_0 = 0, \quad (6.625)$$

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + (\nabla \mathcal{L})(\Theta_{n-1}), \quad (6.626)$$

$$\text{and} \quad \Theta_n = \Theta_{n-1} - \gamma_n \Pi(\mathbf{m}_n). \quad (6.627)$$

Algorithm 6.12.3: Idealized Muon GD optimization method

Input: $\mathfrak{d}_1, \mathfrak{d}_2, N \in \mathbb{N}$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}_1} \times \mathbb{R}^{\mathfrak{d}_2}, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $\xi \in \mathbb{R}^{\mathfrak{d}_1 \times \mathfrak{d}_2}$, $\Pi: \mathbb{R}^{\mathfrak{d}_1 \times \mathfrak{d}_2} \rightarrow \{O \in \mathbb{R}^{\mathfrak{d}_1 \times \mathfrak{d}_2} : ((OO^* = I_{\mathfrak{d}_1}) \vee (O^*O = I_{\mathfrak{d}_2}))\}$ with $\forall A \in \mathbb{R}^{\mathfrak{d}_1 \times \mathfrak{d}_2} : \|\Pi(A) - A\|_{HS} = \inf_{O \in \{O \in \mathbb{R}^{\mathfrak{d}_1 \times \mathfrak{d}_2} : ((OO^* = I_{\mathfrak{d}_1}) \vee (O^*O = I_{\mathfrak{d}_2}))\}} \|O - A\|_{HS}$

Output: N -th step of the idealized Muon GD process for the objective function \mathcal{L} with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, orthogonal projection Π , and initial value ξ (cf. Definition 6.12.2)

- 1: **Initialization:** $\Theta \leftarrow \xi; \mathbf{m} \leftarrow 0 \in \mathbb{R}^{\mathfrak{d}}$
- 2: **for** $n = 1, \dots, N$ **do**
- 3: $\mathbf{m} \leftarrow \alpha_n \mathbf{m} + (\nabla \mathcal{L})(\Theta)$
- 4: $\Theta \leftarrow \Theta - \gamma_n \Pi(\mathbf{m})$
- 5: **return** Θ

Definition 6.12.4 (Newton-Schulz method). Let $a, b, c \in \mathbb{R}$, $\varepsilon \in (0, \infty)$, $n, m \in \mathbb{N}$, $A \in \mathbb{R}^{n \times m}$. Then we denote by $(\text{NS}_{a,b,c,\varepsilon}(A, k))_{k \in \mathbb{N}_0} \subseteq \mathbb{R}^{n \times m}$ the matrices which satisfy

for all $k \in \mathbb{N}$ that $\text{NS}_{a,b,c,\varepsilon}(A, 0) = (\|A\|_{HS} + \varepsilon)^{-1}A$ and

$$\begin{aligned} \text{NS}_{a,b,c,\varepsilon}(A, k) &= a \text{NS}_{a,b,c,\varepsilon}(A, k-1) \\ &\quad + b \text{NS}_{a,b,c,\varepsilon}(A, k-1)(\text{NS}_{a,b,c,\varepsilon}(A, k-1))^* \text{NS}_{a,b,c,\varepsilon}(A, k-1) \\ &\quad + c \text{NS}_{a,b,c,\varepsilon}(A, k-1)(\text{NS}_{a,b,c,\varepsilon}(A, k-1))^* \\ &\quad \text{NS}_{a,b,c,\varepsilon}(A, k-1)(\text{NS}_{a,b,c,\varepsilon}(A, k-1))^* \text{NS}_{a,b,c,\varepsilon}(A, k-1) \end{aligned} \quad (6.628)$$

and for every $k \in \mathbb{N}$ we call $\text{NS}_{a,b,c,\varepsilon}(A, k)$ the k -th iteration of the Newton-Schulz method applied to A with polynomial coefficients a, b, c and regularization parameter ε (cf. Definition 6.12.1).

Definition 6.12.5 (**Muon GD** optimization method). Let $\mathfrak{d}_1, \mathfrak{d}_2, K \in \mathbb{N}$, let $\mathcal{L}: \mathbb{R}^{\mathfrak{d}_1 \times \mathfrak{d}_2} \rightarrow \mathbb{R}$ be differentiable, let $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $a, b, c \in \mathbb{R}$, $\varepsilon \in (0, \infty)$, $\xi \in \mathbb{R}^{\mathfrak{d}_1 \times \mathfrak{d}_2}$, and let $\Theta: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}_1 \times \mathfrak{d}_2}$ be a function. Then we say that Θ is the **Muon GD** process for the objective function \mathcal{L} with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, Newton-Schulz method with polynomial coefficients a, b, c , regularization parameter ε , and N iterations, and initial value ξ if and only if there exists $\mathbf{m}: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}_1 \times \mathfrak{d}_2}$ such that for all $n \in \mathbb{N}$ it holds that

$$\Theta_0 = \xi, \quad \mathbf{m}_0 = 0, \quad (6.629)$$

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + (\nabla \mathcal{L})(\Theta_{n-1}), \quad (6.630)$$

$$\text{and} \quad \Theta_n = \Theta_{n-1} - \gamma_n \text{NS}_{a,b,c,\varepsilon}(\mathbf{m}_n, K) \quad (6.631)$$

(cf. Definition 6.12.4).

Algorithm 6.12.6: **Muon GD** optimization method

Input: $\mathfrak{d}_1, \mathfrak{d}_2, N, K \in \mathbb{N}$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}_1} \times \mathbb{R}^{\mathfrak{d}_2}, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $a, b, c \in \mathbb{R}$, $\varepsilon \in (0, \infty)$, $\xi \in \mathbb{R}^{\mathfrak{d}_1 \times \mathfrak{d}_2}$

Output: N -th step of the **Muon GD** process for the objective function \mathcal{L} with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, Newton-Schulz method with polynomial coefficients a, b, c , regularization parameter ε , and K iterations, and initial value ξ (cf. Definition 6.12.5)

- 1: **Initialization:** $\Theta \leftarrow \xi$; $\mathbf{m} \leftarrow 0 \in \mathbb{R}^{\mathfrak{d}}$
- 2: **for** $n = 1, \dots, N$ **do**
- 3: $\mathbf{m} \leftarrow \alpha_n \mathbf{m} + (\nabla \mathcal{L})(\Theta)$
- 4: $\Theta \leftarrow \Theta - \gamma_n \text{NS}_{a,b,c,\varepsilon}(\mathbf{m}, K)$

5: **return** Θ

Remark 6.12.7. In Jordan et al. [245] it is proposed to choose $a = 3.4445$, $b = -4.7750$, $c = 2.0315$, $\varepsilon = 10^{-7}$, and $N = 5$ as values for the polynomial coefficients a, b, c , regularization parameter ε , and number of iterations N for the Newton-Schulz method in Definition 6.12.5.

Remark 6.12.8. In Definition 6.12.5 we have defined the Muon GD optimization method for loss functions that take matrices as inputs. For loss functions that are defined on more complicated domains (such as, for example, ANNs) the Muon GD optimization method can be applied independently on components of the domain which are spaces of matrices and another GD-type optimization method can be applied to the remaining components. For example, in Jordan et al. [245] it is suggested to apply the Muon GD optimization method to all ANN parameter matrices except for the matrices in the input and output layers and to train all remaining parameters with the AdamW GD optimization method.

6.13 AMSGrad optimization

In this section we consider the AMSGrad GD optimization method (see Reddi et al. [372]). As in the case of previously considered optimization methods, the AMSGrad GD optimization method was introduced in a stochastic setting in Reddi et al. [372], but for pedagogical purposes we present in this section a deterministic version of AMSGrad optimization. We refer to Section 7.14 below for the original stochastic version of AMSGrad optimization.

Definition 6.13.1 (AMSgrad GD optimization method). *Let $\mathfrak{d} \in \mathbb{N}$, let $\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$ be differentiable, let $\mathcal{G} = (\mathcal{G}_1, \dots, \mathcal{G}_{\mathfrak{d}}): \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that*

$$\mathcal{G}(\theta) = (\nabla \mathcal{L})(\theta), \quad (6.632)$$

let $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $(\beta_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $\varepsilon \in (0, \infty)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$, and let $\Theta = (\Theta^{(1)}, \dots, \Theta^{(\mathfrak{d})}): \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a function. Then we say that Θ is the AMSgrad GD process for the objective function \mathcal{L} with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, second moment decay factors $(\beta_n)_{n \in \mathbb{N}}$, regularizing factor ε , and initial value ξ if and only if there exist $\mathbf{m} = (\mathbf{m}^{(1)}, \dots, \mathbf{m}^{(\mathfrak{d})}): \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$, $\mathbb{M} = (\mathbb{M}^{(1)}, \dots, \mathbb{M}^{(\mathfrak{d})}): \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$, and $\mathfrak{M} = (\mathfrak{M}^{(1)}, \dots, \mathfrak{M}^{(\mathfrak{d})}): \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ such that for all $n \in \mathbb{N}$, $i \in \{1, 2, \dots, \mathfrak{d}\}$ it holds that

$$\Theta_0 = \xi, \quad \mathbf{m}_0 = 0, \quad \mathbb{M}_0 = 0, \quad \mathfrak{M}_0 = 0, \quad (6.633)$$

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + (1 - \alpha_n) \mathcal{G}(\Theta_{n-1}), \quad (6.634)$$

$$\mathbb{M}_n^{(i)} = \beta_n \mathbb{M}_{n-1}^{(i)} + (1 - \beta_n) |\mathcal{G}_i(\Theta_{n-1})|^2, \quad (6.635)$$

$$\mathfrak{M}_n^{(i)} = \max\{\mathfrak{M}_{n-1}^{(i)}, \mathbb{M}_n^{(i)}\}, \quad \text{and} \quad (6.636)$$

$$\Theta_n^{(i)} = \Theta_{n-1}^{(i)} - \frac{\gamma_n \mathbf{m}_n^{(i)}}{(\mathfrak{M}_n^{(i)})^{1/2} + \varepsilon}. \quad (6.637)$$

Algorithm 6.13.2: AMSgrad GD optimization method

Input: $\mathfrak{d}, N \in \mathbb{N}$, $\mathcal{L} \in C^1(\mathbb{R}^\mathfrak{d}, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $(\beta_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $\varepsilon \in (0, \infty)$, $\xi \in \mathbb{R}^\mathfrak{d}$

Output: N -th step of the AMSgrad GD process for the objective function \mathcal{L} with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, second moment decay factors $(\beta_n)_{n \in \mathbb{N}}$, regularizing factor ε , and initial value ξ (cf. Definition 6.13.1)

```

1: Initialization:  $\Theta \leftarrow \xi$ ;  $\mathbf{m} \leftarrow 0 \in \mathbb{R}^\mathfrak{d}$ ;  $\mathbb{M} \leftarrow 0 \in \mathbb{R}^\mathfrak{d}$ ;  $\mathfrak{M} \leftarrow 0 \in \mathbb{R}^\mathfrak{d}$ 
2: for  $n = 1, \dots, N$  do # (cf. Definition 6.5.2)
3:    $\mathbf{m} \leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n)(\nabla \mathcal{L})(\Theta)$ 
4:    $\mathbb{M} \leftarrow \beta_n \mathbb{M} + (1 - \beta_n)[(\nabla \mathcal{L})(\Theta)]^2$ 
5:    $\mathfrak{M} \leftarrow \max\{\mathfrak{M}, \mathbb{M}\}$ 
6:    $\Theta \leftarrow \Theta - \gamma_n [\varepsilon + \mathfrak{M}^{1/2}]^{-1} \mathbf{m}$ 
7: return  $\Theta$ 

```

6.14 Compact summary of deterministic GD optimization methods

In this section we provide an overview over all deterministic GD-type optimization methods considered in Chapter 6. Roughly speaking, in this summary we provide for each considered method the iteration step of the respective pseudo-code. The formulas in this summary make use of the componentwise operations in Definition 6.5.2.

GD optimization method (Definition 6.1.1)

$$\Theta \leftarrow \Theta - \gamma_n (\nabla \mathcal{L})(\Theta)$$

Explicit midpoint GD optimization method (Definition 6.2.1)

$$\Theta \leftarrow \Theta - \gamma_n (\nabla \mathcal{L}) \left(\Theta - \frac{\gamma_n}{2} (\nabla \mathcal{L})(\Theta) \right)$$

Momentum GD optimization method (Definition 6.3.1)

$$\begin{aligned} \mathbf{m} &\leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n)(\nabla \mathcal{L})(\Theta) \\ \Theta &\leftarrow \Theta - \gamma_n \mathbf{m} \end{aligned}$$

Momentum GD optimization method (2nd version) (Definition 6.3.3)

$$\begin{aligned}\mathbf{m} &\leftarrow \alpha_n \mathbf{m} + (\nabla \mathcal{L})(\Theta) \\ \Theta &\leftarrow \Theta - \gamma_n \mathbf{m}\end{aligned}$$

Momentum GD optimization method (3rd version) (Definition 6.3.5)

$$\begin{aligned}\mathbf{m} &\leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n) \gamma_n (\nabla \mathcal{L})(\Theta) \\ \Theta &\leftarrow \Theta - \mathbf{m}\end{aligned}$$

Momentum GD optimization method (4th version) (Definition 6.3.7)

$$\begin{aligned}\mathbf{m} &\leftarrow \alpha_n \mathbf{m} + \gamma_n (\nabla \mathcal{L})(\Theta) \\ \Theta &\leftarrow \Theta - \mathbf{m}\end{aligned}$$

Bias-adjusted momentum GD optimization method (Definition 6.3.19)

$$\begin{aligned}\mathbf{m} &\leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n) (\nabla \mathcal{L})(\Theta) \\ \Theta &\leftarrow \Theta - \frac{\gamma_n \mathbf{m}}{1 - \prod_{l=1}^n \alpha_l}\end{aligned}$$

Nesterov accelerated GD optimization method (Definition 6.4.1)

$$\begin{aligned}\mathbf{m} &\leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n) (\nabla \mathcal{L})(\Theta - \gamma_n \alpha_n \mathbf{m}) \\ \Theta &\leftarrow \Theta - \gamma_n \mathbf{m}\end{aligned}$$

Nesterov accelerated GD optimization method (2nd version) (Definition 6.4.3)

$$\begin{aligned}\mathbf{m} &\leftarrow \alpha_n \mathbf{m} + (\nabla \mathcal{L})(\Theta - \gamma_n \alpha_n \mathbf{m}) \\ \Theta &\leftarrow \Theta - \gamma_n \mathbf{m}\end{aligned}$$

Nesterov accelerated GD optimization method (3rd version) (Definition 6.4.5)

$$\begin{aligned}\mathbf{m} &\leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n) \gamma_n (\nabla \mathcal{L})(\Theta - \alpha_n \mathbf{m}) \\ \Theta &\leftarrow \Theta - \mathbf{m}\end{aligned}$$

Nesterov accelerated GD optimization method (4th version) (Definition 6.4.7)

$$\begin{aligned}\mathbf{m} &\leftarrow \alpha_n \mathbf{m} + \gamma_n (\nabla \mathcal{L})(\Theta - \alpha_n \mathbf{m}) \\ \Theta &\leftarrow \Theta - \mathbf{m}\end{aligned}$$

Bias-adjusted Nesterov accelerated GD optimization method (Definition 6.4.16)

$$\begin{aligned}\mathbf{m} &\leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n) (\nabla \mathcal{L})(\Theta - \frac{\gamma_n \alpha_n \mathbf{m}}{1 - \prod_{l=1}^n \alpha_l}) \\ \Theta &\leftarrow \Theta - \frac{\gamma_n \mathbf{m}}{1 - \prod_{l=1}^n \alpha_l}\end{aligned}$$

Shifted Nesterov accelerated GD optimization method (Definition 6.4.19)

$$\begin{aligned}\mathbf{m} &\leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n)(\nabla \mathcal{L})(\Theta) \\ \Theta &\leftarrow \Theta - \gamma_{n+1} \alpha_{n+1} \mathbf{m} - \gamma_n (1 - \alpha_n)(\nabla \mathcal{L})(\Theta)\end{aligned}$$

Shifted Nesterov accelerated GD optimization method (2nd version) (Definition 6.4.22)

$$\begin{aligned}\mathbf{m} &\leftarrow \alpha_n \mathbf{m} + (\nabla \mathcal{L})(\Theta) \\ \Theta &\leftarrow \Theta - \gamma_{n+1} \alpha_{n+1} \mathbf{m} - \gamma_n (\nabla \mathcal{L})(\Theta)\end{aligned}$$

Shifted Nesterov accelerated GD optimization method (3rd version) (Definition 6.4.25)

$$\begin{aligned}\mathbf{m} &\leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n) \gamma_n (\nabla \mathcal{L})(\Theta) \\ \Theta &\leftarrow \Theta - \alpha_{n+1} \mathbf{m} - (1 - \alpha_n) \gamma_n (\nabla \mathcal{L})(\Theta)\end{aligned}$$

Shifted Nesterov accelerated GD optimization method (4th version) (Definition 6.4.28)

$$\begin{aligned}\mathbf{m} &\leftarrow \alpha_n \mathbf{m} + \gamma_n (\nabla \mathcal{L})(\Theta) \\ \Theta &\leftarrow \Theta - \alpha_{n+1} \mathbf{m} - \gamma_n (\nabla \mathcal{L})(\Theta)\end{aligned}$$

Shifted bias-adjusted Nesterov accelerated GD optimization method (Definition 6.4.31)

$$\begin{aligned}\mathbf{m} &\leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n)(\nabla \mathcal{L})(\Theta) \\ \Theta &\leftarrow \Theta - \frac{\gamma_n(1 - \alpha_n)(\nabla \mathcal{L})(\Theta)}{1 - \prod_{l=1}^n \alpha_l} - \frac{\gamma_{n+1} \alpha_{n+1} \mathbf{m}}{1 - \prod_{l=1}^{n+1} \alpha_l}\end{aligned}$$

Simplified Nesterov accelerated GD optimization method (Definition 6.4.33)

$$\begin{aligned}\mathbf{m} &\leftarrow \alpha_n \mathbf{m} + (\nabla \mathcal{L})(\Theta) \\ \Theta &\leftarrow \Theta - \gamma_n \alpha_n \mathbf{m} - \gamma_n (\nabla \mathcal{L})(\Theta)\end{aligned}$$

Adagrad GD optimization method (Definition 6.5.1)

$$\begin{aligned}\mathbb{M} &\leftarrow \mathbb{M} + [(\nabla \mathcal{L})(\Theta)]^2 \\ \Theta &\leftarrow \Theta - \gamma_n [\varepsilon + \mathbb{M}^{1/2}]^{-1} (\nabla \mathcal{L})(\Theta)\end{aligned}$$

RMSprop GD optimization method (Definition 6.6.1)

$$\begin{aligned}\mathbb{M} &\leftarrow \beta_n \mathbb{M} + (1 - \beta_n)[(\nabla \mathcal{L})(\Theta)]^2 \\ \Theta &\leftarrow \Theta - \gamma_n [\varepsilon + \mathbb{M}^{1/2}]^{-1} (\nabla \mathcal{L})(\Theta)\end{aligned}$$

Bias-adjusted RMSprop GD optimization method (Definition 6.6.5)

$$\begin{aligned}\mathbb{M} &\leftarrow \beta_n \mathbb{M} + (1 - \beta_n)[(\nabla \mathcal{L})(\Theta)]^2 \\ \Theta &\leftarrow \Theta - \gamma_n \left[\varepsilon + \left[\frac{\mathbb{M}}{1 - \prod_{k=1}^n \beta_k} \right]^{1/2} \right]^{-1} (\nabla \mathcal{L})(\Theta)\end{aligned}$$

Adadelta GD optimization method (Definition 6.7.1)

$$\begin{aligned}\mathbb{M} &\leftarrow \beta_n \mathbb{M} + (1 - \beta_n)[(\nabla \mathcal{L})(\Theta)]^2 \\ \Theta &\leftarrow \Theta - \left[\frac{\varepsilon + \Delta}{\varepsilon + \mathbb{M}} \right]^{1/2} (\nabla \mathcal{L})(\Theta) \\ \Delta &\leftarrow \delta_n \Delta + (1 - \delta_n)[\Theta - \Xi]^2 \\ \Xi &\leftarrow \Theta\end{aligned}$$

Adam GD optimization method (Definition 6.8.1)

$$\begin{aligned}\mathbf{m} &\leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n)(\nabla \mathcal{L})(\Theta) \\ \mathbb{M} &\leftarrow \beta_n \mathbb{M} + (1 - \beta_n)[(\nabla \mathcal{L})(\Theta)]^2 \\ \Theta &\leftarrow \Theta - \left[\varepsilon + \left[\frac{\mathbb{M}}{1 - \prod_{k=1}^n \beta_k} \right]^{1/2} \right]^{-1} \left[\frac{\gamma_n \mathbf{m}}{1 - \prod_{k=1}^n \alpha_k} \right]\end{aligned}$$

Adamax GD optimization method (Definition 6.8.3)

$$\begin{aligned}\mathbf{m} &\leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n)(\nabla \mathcal{L})(\Theta) \\ \mathbb{M} &\leftarrow \max\{\beta_n \mathbb{M}, |(\nabla \mathcal{L})(\Theta)|\} \\ \Theta &\leftarrow \Theta - \gamma_n [\varepsilon + \mathbb{M}]^{-1} \left[\frac{\mathbf{m}}{1 - \prod_{k=1}^n \alpha_k} \right]\end{aligned}$$

Nadam GD optimization method (Definition 6.9.1)

$$\begin{aligned}\mathbf{m} &\leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n)(\nabla \mathcal{L})(\Theta) \\ \mathbb{M} &\leftarrow \beta_n \mathbb{M} + (1 - \beta_n)[(\nabla \mathcal{L})(\Theta)]^2 \\ \Theta &\leftarrow \Theta - \left[\varepsilon + \left[\frac{\mathbb{M}}{1 - \prod_{k=1}^n \beta_k} \right]^{1/2} \right]^{-1} \left[\left[\frac{\gamma_n(1 - \alpha_n)}{1 - \prod_{k=1}^n \alpha_k} \right] (\nabla \mathcal{L})(\Theta) + \left[\frac{\gamma_{n+1} \alpha_{n+1}}{1 - \prod_{k=1}^{n+1} \alpha_k} \right] \mathbf{m} \right]\end{aligned}$$

Simplified Nadam GD optimization method (Definition 6.9.3)

$$\begin{aligned}\mathbf{m} &\leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n)(\nabla \mathcal{L})(\Theta) \\ \mathbb{M} &\leftarrow \beta_n \mathbb{M} + (1 - \beta_n)[(\nabla \mathcal{L})(\Theta)]^2 \\ \Theta &\leftarrow \Theta - \gamma_n \left[\varepsilon + \left[\frac{\mathbb{M}}{1 - \prod_{k=1}^n \beta_k} \right]^{1/2} \right]^{-1} \left[\left[\frac{1 - \alpha_n}{1 - \prod_{k=1}^n \alpha_k} \right] (\nabla \mathcal{L})(\Theta) + \left[\frac{\alpha_{n+1}}{1 - \prod_{k=1}^{n+1} \alpha_k} \right] \mathbf{m} \right]\end{aligned}$$

kil

Nadamax GD optimization method (Definition 6.9.5)

$$\begin{aligned}\mathbf{m} &\leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n)(\nabla \mathcal{L})(\Theta) \\ \mathbb{M} &\leftarrow \max\{\beta_n \mathbb{M}, |(\nabla \mathcal{L})(\Theta)|\} \\ \Theta &\leftarrow \Theta - [\varepsilon + \mathbb{M}]^{-1} \left[\left[\frac{\gamma_n(1-\alpha_n)}{1-\prod_{l=1}^n \alpha_l} \right] (\nabla \mathcal{L})(\Theta) + \left[\frac{\gamma_{n+1}\alpha_{n+1}}{1-\prod_{l=1}^{n+1} \alpha_l} \right] \mathbf{m} \right]\end{aligned}$$

Adam GD optimization method with L^2 -regularization (Definition 6.10.3)

$$\begin{aligned}\mathbf{m} &\leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n)((\nabla \mathcal{L})(\Theta) + \lambda \Theta) \\ \mathbb{M} &\leftarrow \beta_n \mathbb{M} + (1 - \beta_n)[(\nabla \mathcal{L})(\Theta) + \lambda \Theta]^2 \\ \Theta &\leftarrow \Theta - \left[\varepsilon + \left[\frac{\mathbb{M}}{1-\prod_{k=1}^n \beta_k} \right]^{1/2} \right]^{-1} \left[\frac{\gamma_n \mathbf{m}}{1-\prod_{k=1}^n \alpha_k} \right]\end{aligned}$$

AdamW GD optimization method (Definition 6.10.1)

$$\begin{aligned}\mathbf{m} &\leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n)(\nabla \mathcal{L})(\Theta) \\ \mathbb{M} &\leftarrow \beta_n \mathbb{M} + (1 - \beta_n)[(\nabla \mathcal{L})(\Theta)]^2 \\ \Theta &\leftarrow \Theta - \gamma_n \left(\left[\varepsilon + \left[\frac{\mathbb{M}}{1-\prod_{k=1}^n \beta_k} \right]^{1/2} \right]^{-1} \left[\frac{\mathbf{m}}{1-\prod_{k=1}^n \alpha_k} \right] + \lambda \Theta \right)\end{aligned}$$

Shampoo GD optimization method (Definition 6.11.4)

$$\begin{aligned}\mathbf{L} &\leftarrow \mathbf{L} + (\nabla \mathcal{L})(\Theta)((\nabla \mathcal{L})(\Theta))^* \\ \mathbf{R} &\leftarrow \mathbf{R} + ((\nabla \mathcal{L})(\Theta))^*(\nabla \mathcal{L})(\Theta) \\ \Theta &\leftarrow \Theta - \gamma_n (\mathbf{L})^{-1/4} (\nabla \mathcal{L})(\Theta) (\mathbf{R})^{-1/4}\end{aligned}$$

Idealized Muon GD optimization method (Definition 6.12.2)

$$\begin{aligned}\mathbf{m} &\leftarrow \alpha_n \mathbf{m} + (\nabla \mathcal{L})(\Theta) \\ \Theta &\leftarrow \Theta - \gamma_n \Pi(\mathbf{m})\end{aligned}$$

Muon GD optimization method (Definition 6.12.5)

$$\begin{aligned}\mathbf{m} &\leftarrow \alpha_n \mathbf{m} + (\nabla \mathcal{L})(\Theta) \\ \Theta &\leftarrow \Theta - \gamma_n \text{NS}_{a,b,c,\varepsilon}(\mathbf{m}, K)\end{aligned}$$

AMSGrad GD optimization method (Definition 6.13.1)

$$\begin{aligned}\mathbf{m} &\leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n)(\nabla \mathcal{L})(\Theta) \\ \mathbb{M} &\leftarrow \beta_n \mathbb{M} + (1 - \beta_n)[(\nabla \mathcal{L})(\Theta)]^2 \\ \mathfrak{M} &\leftarrow \max\{\mathfrak{M}, \mathbb{M}\} \\ \Theta &\leftarrow \Theta - \gamma_n [\varepsilon + \mathfrak{M}^{1/2}]^{-1} \mathbf{m}\end{aligned}$$

Chapter 7

Stochastic gradient descent (SGD) optimization methods

This chapter reviews and studies SGD-type optimization methods such as the classical plain-vanilla SGD optimization method (see Section 7.2) as well as more sophisticated SGD-type optimization methods including SGD-type optimization methods with momenta (cf. Sections 7.4, 7.5, and 7.9 below) and SGD-type optimization methods with adaptive modifications of the learning rates (cf. Sections 7.6, 7.7, 7.8, and 7.9 below).

For a brief list of resources in the scientific literature providing reviews on gradient based optimization methods we refer to the beginning of Chapter 6.

7.1 Introductory comments for the training of ANNs with SGD

In Chapter 6 we have introduced and studied deterministic GD-type optimization methods. In deep learning algorithms usually not deterministic GD-type optimization methods but stochastic variants of GD-type optimization methods are employed. Such SGD-type optimization methods can be viewed as suitable Monte Carlo approximations of deterministic GD-type methods and in this section we now roughly sketch some of the main ideas of such SGD-type optimization methods. To do this, we now briefly recall the deep supervised learning framework developed in the [introduction](#) and Section 5.1 above.

Specifically, let $d, M \in \mathbb{N}$, $\mathcal{E} \in C(\mathbb{R}^d, \mathbb{R})$, $x_1, x_2, \dots, x_{M+1} \in \mathbb{R}^d$, $y_1, y_2, \dots, y_M \in \mathbb{R}$ satisfy for all $m \in \{1, 2, \dots, M\}$ that

$$y_m = \mathcal{E}(x_m). \quad (7.1)$$

As in the [introduction](#) and in Section 5.1 we think of $M \in \mathbb{N}$ as the number of available known input-output data pairs, we think of $d \in \mathbb{N}$ as the dimension of the input data, we

think of $\mathcal{E}: \mathbb{R}^d \rightarrow \mathbb{R}$ as an unknown function which we want to approximate, we think of $x_1, x_2, \dots, x_{M+1} \in \mathbb{R}^d$ as the available known input data, we think of $y_1, y_2, \dots, y_M \in \mathbb{R}$ as the available known output data, and we are trying to use the available known input-output data pairs to approximate the unknown function \mathcal{E} by means of ANNs.

Specifically, let $a: \mathbb{R} \rightarrow \mathbb{R}$ be differentiable, let $h \in \mathbb{N}$, $l_1, l_2, \dots, l_h, \mathfrak{d} \in \mathbb{N}$ satisfy $\mathfrak{d} = l_1(d+1) + [\sum_{k=2}^h l_k(l_{k-1}+1)] + l_h + 1$, and let $\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \rightarrow [0, \infty)$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that

$$\mathcal{L}(\theta) = \frac{1}{M} \left[\sum_{m=1}^M |(\mathcal{N}_{\mathfrak{M}_{a,l_1}, \mathfrak{M}_{a,l_2}, \dots, \mathfrak{M}_{a,l_h}, \text{id}_{\mathbb{R}}}^{\vartheta, d})(x_m) - y_m|^2 \right] \quad (7.2)$$

(cf. Definitions 1.1.3 and 1.2.1). Note that h is the number of hidden layers of the ANNs in (7.2), note for every $i \in \{1, 2, \dots, h\}$ that $l_i \in \mathbb{N}$ is the number of neurons in the i -th hidden layer of the ANNs in (7.2), and note that \mathfrak{d} is the number of real parameters used to describe the ANNs in (7.2). We recall that we are trying to approximate the function \mathcal{E} by, first, computing an approximate minimizer $\vartheta \in \mathbb{R}^{\mathfrak{d}}$ of the function $\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \rightarrow [0, \infty)$ and, thereafter, employing the realization

$$\mathbb{R}^d \ni x \mapsto \mathcal{N}_{\mathfrak{M}_{a,l_1}, \mathfrak{M}_{a,l_2}, \dots, \mathfrak{M}_{a,l_h}, \text{id}_{\mathbb{R}}}^{\vartheta, d} \in \mathbb{R} \quad (7.3)$$

of the ANN associated to the approximate minimizer $\vartheta \in \mathbb{R}^{\mathfrak{d}}$ as an approximation of \mathcal{E} .

Deep learning algorithms typically solve optimization problems of the type (7.2) by means of gradient based optimization methods, which aim to minimize the considered objective function by performing successive steps based on the direction of the negative gradient of the objective function. We recall that one of the simplest gradient based optimization method is the plain-vanilla GD optimization method which performs successive steps in the direction of the negative gradient. In the context of the optimization problem in (7.2) this GD optimization method reads as follows. Let $\xi \in \mathbb{R}^{\mathfrak{d}}$, let $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, and let $\theta = (\theta_n)_{n \in \mathbb{N}_0}: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $n \in \mathbb{N}$ that

$$\theta_0 = \xi \quad \text{and} \quad \theta_n = \theta_{n-1} - \gamma_n (\nabla \mathcal{L})(\theta_{n-1}). \quad (7.4)$$

Note that the process $(\theta_n)_{n \in \mathbb{N}_0}$ is the GD process for the objective function \mathcal{L} with learning rates $(\gamma_n)_{n \in \mathbb{N}}$ and initial value ξ (cf. Definition 6.1.1). Moreover, observe that the assumption that a is differentiable ensures that \mathcal{L} in (7.4) is also differentiable (see Section 5.3.2 above for details).

In typical practical deep learning applications the number M of available known input-output data pairs is very large, say, for instance, $M \geq 10^6$. As a consequence it is typically computationally prohibitively expensive to determine the exact gradient of the objective function to perform steps of deterministic GD-type optimization methods. As a remedy for this, deep learning algorithms usually employ stochastic variants of GD-type optimization methods, where in each step of the optimization method the precise gradient of the objective function is replaced by a Monte Carlo approximation of the gradient of the objective function.

We now sketch this approach for the **GD** optimization method in (7.4) resulting in the popular **SGD** optimization method applied to (7.2).

Specifically, let $J \in \mathbb{N}$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, for every $n, j \in \mathbb{N}$, let $\mathbf{m}_{n,j} : \Omega \rightarrow \mathbb{R}$ be a $\{1, 2, \dots, M\}$ -uniformly distributed random variable, let $\ell : \mathbb{R}^d \times \mathbb{R} \rightarrow \mathbb{R}$ satisfy for all $\theta \in \mathbb{R}^d$, $m \in \{1, 2, \dots, M\}$ that

$$\ell(\theta, m) = |(\mathcal{N}_{\mathbf{m}_{a,l_1}, \mathbf{m}_{a,l_2}, \dots, \mathbf{m}_{a,l_h}, \text{id}_{\mathbb{R}}})^{\theta, d}(x_m) - y_m|^2, \quad (7.5)$$

and let $\Theta = (\Theta_n)_{n \in \mathbb{N}_0} : \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^d$ satisfy for all $n \in \mathbb{N}$ that

$$\Theta_0 = \xi \quad \text{and} \quad \Theta_n = \Theta_{n-1} - \gamma_n \left[\frac{1}{J} \sum_{j=1}^J (\nabla_{\theta} \ell)(\Theta_{n-1}, \mathbf{m}_{n,j}) \right]. \quad (7.6)$$

The stochastic process $(\Theta_n)_{n \in \mathbb{N}_0}$ is an **SGD** process for the minimization problem associated to (7.2) with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, constant number of Monte Carlo samples (batch sizes) J , initial value ξ , and data $(\mathbf{m}_{n,j})_{(n,j) \in \mathbb{N}^2}$ (see Definition 7.2.1 below for the precise definition). Note that in (7.6) in each step $n \in \mathbb{N}$ we only employ a Monte Carlo approximation

$$\frac{1}{J} \sum_{j=1}^J (\nabla_{\theta} \ell)(\Theta_{n-1}, \mathbf{m}_{n,j}) \approx \frac{1}{M} \sum_{m=1}^M (\nabla_{\theta} \ell)(\Theta_{n-1}, m) = (\nabla \mathcal{L})(\Theta_{n-1}) \quad (7.7)$$

of the exact gradient of the objective function. Nonetheless, in deep learning applications the **SGD** optimization method (or other **SGD**-type optimization methods) typically result in good approximate minimizers of the objective function. Note that employing approximate gradients in the **SGD** optimization method in (7.6) means that performing any step of the **SGD** process involves the computation of a sum with only J summands, while employing the exact gradient in the **GD** optimization method in (7.4) means that performing any step of the process involves the computation of a sum with M summands. In deep learning applications when M is very large (for example, $M \geq 10^6$) and J is chosen to be reasonably small (for instance, $J = 128$), this means that performing steps of the **SGD** process is much more computationally affordable than performing steps of the **GD** process. Combining this with the fact that **SGD**-type optimization methods do in the training of **ANNs** often find good approximate minimizers (cf., for example, Remark 9.15.5 and [106, 412]) is the key reason making the **SGD** optimization method and other **SGD**-type optimization methods the optimization methods chosen in almost all deep learning applications. It is the topic of this chapter to introduce and study **SGD**-type optimization methods such as the plain-vanilla **SGD** optimization method in (7.6) above.

7.2 SGD optimization

In the next notion we present the promised stochastic version of the plain-vanilla **GD** optimization method from Section 6.1, that is, in the next notion we present the plain-

vanilla SGD optimization method.

Definition 7.2.1 (SGD optimization method). Let $\mathfrak{d} \in \mathbb{N}$, let (S, \mathcal{S}) be a measurable space, let $\ell = (\ell(\theta, x))_{(\theta, x) \in \mathbb{R}^{\mathfrak{d}} \times S}: \mathbb{R}^{\mathfrak{d}} \times S \rightarrow \mathbb{R}$ be measurable, assume for all $x \in S$ that $\ell(\cdot, x)$ is differentiable, let $g: \mathbb{R}^{\mathfrak{d}} \times S \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}, x \in S$ that

$$g(\theta, x) = (\nabla_{\theta}\ell)(\theta, x), \quad (7.8)$$

let $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\xi: \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a random variable, for every $n, j \in \mathbb{N}$ let $X_{n,j}: \Omega \rightarrow S$ be a random variable, and let $\Theta: \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a function. Then we say that Θ is the SGD process for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ if and only if it holds for all $n \in \mathbb{N}$ that

$$\Theta_0 = \xi \quad \text{and} \quad \Theta_n = \Theta_{n-1} - \gamma_n \left[\frac{1}{J_n} \sum_{j=1}^{J_n} g(\Theta_{n-1}, X_{n,j}) \right]. \quad (7.9)$$

Algorithm 7.2.2: SGD optimization method

Input: $\mathfrak{d}, \mathbf{d}, N \in \mathbb{N}$, $\ell = (\ell(\theta, x))_{(\theta, x) \in \mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\mathbf{d}}} \in C^1(\mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\mathbf{d}}, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, probability space $(\Omega, \mathcal{F}, \mathbb{P})$, random variable $\xi: \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$, random variables $X_{n,j}: \Omega \rightarrow \mathbb{R}^{\mathbf{d}}$ for $n, j \in \mathbb{N}$

Output: N -th step of the SGD process for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ (cf. Definition 7.2.1)

```

1: Initialization:  $\Theta \leftarrow \xi$ 
2: for  $n = 1, \dots, N$  do
3:    $g \leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_{\theta}\ell)(\Theta, X_{n,j})$ 
4:    $\Theta \leftarrow \Theta - \gamma_n g$ 
5: return  $\Theta$ 

```

7.2.1 SGD optimization in the training of ANNs

In the next example we apply the SGD optimization method in the context of the training of 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). Note that this is a very similar framework as the one developed in Section 7.1.

Example 7.2.3. Let $d, h, \mathfrak{d} \in \mathbb{N}$, $l_1, l_2, \dots, l_h \in \mathbb{N}$ satisfy $\mathfrak{d} = l_1(d+1) + [\sum_{k=2}^h l_k(l_{k-1}+1)] + l_h + 1$, let $a: \mathbb{R} \rightarrow \mathbb{R}$ be differentiable, let $M \in \mathbb{N}$, $x_1, x_2, \dots, x_M \in \mathbb{R}^d$, $y_1, y_2, \dots, y_M \in \mathbb{R}$,

let $\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \rightarrow [0, \infty)$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that

$$\mathcal{L}(\theta) = \frac{1}{M} \left[\sum_{m=1}^M \left| (\mathcal{N}_{\mathfrak{M}_{a,l_1}, \mathfrak{M}_{a,l_2}, \dots, \mathfrak{M}_{a,l_h}, \text{id}_{\mathbb{R}}}^{\theta,d})(x_m) - y_m \right|^2 \right], \quad (7.10)$$

let $\ell: \mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{d+1} \rightarrow \mathbb{R}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}, x \in \mathbb{R}^d, y \in \mathbb{R}$ that

$$\ell(\theta, (x, y)) = \left| (\mathcal{N}_{\mathfrak{M}_{a,l_1}, \mathfrak{M}_{a,l_2}, \dots, \mathfrak{M}_{a,l_h}, \text{id}_{\mathbb{R}}}^{\theta,d})(x) - y \right|^2, \quad (7.11)$$

let $\xi \in \mathbb{R}^{\mathfrak{d}}$, let $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, let $\vartheta: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $n \in \mathbb{N}$ that

$$\vartheta_0 = \xi \quad \text{and} \quad \vartheta_n = \vartheta_{n-1} - \gamma_n (\nabla \mathcal{L})(\vartheta_{n-1}), \quad (7.12)$$

let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, for every $n, j \in \mathbb{N}$, let $\mathbf{m}_{n,j}: \Omega \rightarrow \mathbb{R}$ be a $\{1, 2, \dots, M\}$ -uniformly distributed random variable, for every $n, j \in \mathbb{N}$ let $X_{n,j}: \Omega \rightarrow \mathbb{R}^d$ and $Y_{n,j}: \Omega \rightarrow \mathbb{R}$ satisfy

$$X_{n,j} = x_{\mathbf{m}_{n,j}} \quad \text{and} \quad Y_{n,j} = y_{\mathbf{m}_{n,j}}, \quad (7.13)$$

and let $\Theta: \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $n \in \mathbb{N}$ that

$$\Theta_0 = \xi \quad \text{and} \quad \Theta_n = \Theta_{n-1} - \gamma_n \left[\frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_{\theta} \ell)(\Theta_{n-1}, (X_{n,j}, Y_{n,j})) \right] \quad (7.14)$$

(cf. Definitions 1.1.3 and 1.2.1 and Corollary 5.3.6). Then

- (i) it holds that ϑ is the **GD** process for the objective function \mathcal{L} with learning rates $(\gamma_n)_{n \in \mathbb{N}}$ and initial value ξ ,
- (ii) it holds that Θ is the **SGD** process for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, batch sizes $(J_n)_{n \in \mathbb{N}}$, initial value ξ , and data $((X_{n,j}, Y_{n,j}))_{(n,j) \in \mathbb{N}^2}$, and
- (iii) it holds for all $n \in \mathbb{N}, \theta \in \mathbb{R}^{\mathfrak{d}}$ that

$$\mathbb{E} \left[\theta - \gamma_n \left[\frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_{\theta} \ell)(\theta, (X_{n,j}, Y_{n,j})) \right] \right] = \theta - \gamma_n (\nabla \mathcal{L})(\theta) \quad (7.15)$$

(cf. Definitions 6.1.1 and 7.2.1).

Proof for Example 7.2.3. Observe that (7.12) establishes item (i). Note that (7.14) proves item (ii). Observe that (7.10), (7.11), (7.13), and the assumption that for all $n, j \in \mathbb{N}$ it

holds that $\mathbf{m}_{n,j}$ is uniformly distributed on $\{1, 2, \dots, M\}$ demonstrate that for all $\theta \in \mathbb{R}^{\mathfrak{d}}$, $n, j \in \mathbb{N}$ it holds that

$$\begin{aligned}
 \mathbb{E}[(\nabla_{\theta}\ell)(\theta, (X_{n,j}, Y_{n,j}))] &= \mathbb{E}[(\nabla_{\theta}\ell)(\theta, (\mathbf{x}_{\mathbf{m}_{n,j}}, \mathbf{y}_{\mathbf{m}_{n,j}}))] \\
 &= \mathbb{E}\left[\sum_{m=1}^M(\nabla_{\theta}\ell)(\theta, (\mathbf{x}_m, \mathbf{y}_m))\mathbb{1}_{\{\mathbf{m}_{n,j}=m\}}\right] \\
 &= \sum_{m=1}^M(\nabla_{\theta}\ell)(\theta, (\mathbf{x}_m, \mathbf{y}_m))\mathbb{E}\left[\mathbb{1}_{\{\mathbf{m}_{n,j}=m\}}\right] \\
 &= \sum_{m=1}^M(\nabla_{\theta}\ell)(\theta, (\mathbf{x}_m, \mathbf{y}_m))\mathbb{P}(\{\mathbf{m}_{n,j}=m\}) \\
 &= \frac{1}{M}\left[\sum_{m=1}^M(\nabla_{\theta}\ell)(\theta, (\mathbf{x}_m, \mathbf{y}_m))\right] \\
 &= \frac{1}{M}\left[\sum_{m=1}^M\left|(\nabla_{\theta}\mathcal{N}_{\mathfrak{M}_{a,l_1}, \mathfrak{M}_{a,l_2}, \dots, \mathfrak{M}_{a,l_h}, \text{id}_{\mathbb{R}}}^{\theta, d})(\mathbf{x}_m) - \mathbf{y}_m\right|^2\right] \\
 &= (\nabla\mathcal{L})(\theta).
 \end{aligned} \tag{7.16}$$

Hence, we obtain for all $n \in \mathbb{N}$, $\theta \in \mathbb{R}^{\mathfrak{d}}$ that

$$\begin{aligned}
 &\mathbb{E}\left[\theta - \gamma_n\left[\frac{1}{J_n}\sum_{j=1}^{J_n}(\nabla_{\theta}\ell)(\theta, (X_{n,j}, Y_{n,j}))\right]\right] \\
 &= \theta - \gamma_n\left[\frac{1}{J_n}\sum_{j=1}^{J_n}\mathbb{E}[(\nabla_{\theta}\ell)(\theta, (X_{n,j}, Y_{n,j}))]\right] \\
 &= \theta - \gamma_n\left[\frac{1}{J_n}\sum_{j=1}^{J_n}(\nabla\mathcal{L})(\theta)\right] \\
 &= \theta - \gamma_n(\nabla\mathcal{L})(\theta).
 \end{aligned} \tag{7.17}$$

The proof for Example 7.2.3 is thus complete. \square

Example 7.2.4. Let $d, h, \mathfrak{d} \in \mathbb{N}$, $l_1, l_2, \dots, l_h \in \mathbb{N}$ satisfy $\mathfrak{d} = l_1(d+1) + [\sum_{k=2}^h l_k(l_{k-1}+1)] + l_h + 1$, let $a: \mathbb{R} \rightarrow \mathbb{R}$ be differentiable, let $M \in \mathbb{N}$, $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_M \in \mathbb{R}^d$, $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_M \in \mathbb{R}$, let $\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \rightarrow [0, \infty)$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that

$$\mathcal{L}(\theta) = \frac{1}{M}\left[\sum_{m=1}^M\left|(\mathcal{N}_{\mathfrak{M}_{a,l_1}, \mathfrak{M}_{a,l_2}, \dots, \mathfrak{M}_{a,l_h}, \text{id}_{\mathbb{R}}}^{\theta, d})(\mathbf{x}_m) - \mathbf{y}_m\right|^2\right], \tag{7.18}$$

let $S = \{1, 2, \dots, M\}$, let $\ell: \mathbb{R}^d \times S \rightarrow \mathbb{R}$ satisfy for all $\theta \in \mathbb{R}^d$, $m \in S$ that

$$\ell(\theta, m) = |(\mathcal{N}_{\mathfrak{M}_{a,l_1}, \mathfrak{M}_{a,l_2}, \dots, \mathfrak{M}_{a,l_h}, \text{id}_{\mathbb{R}}})^{\theta,d}(x_m) - y_m|^2, \quad (7.19)$$

let $\xi \in \mathbb{R}^d$, let $(\gamma_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, let $\vartheta: \mathbb{N}_0 \rightarrow \mathbb{R}^d$ satisfy for all $n \in \mathbb{N}$ that

$$\vartheta_0 = \xi \quad \text{and} \quad \vartheta_n = \vartheta_{n-1} - \gamma_n (\nabla \mathcal{L})(\vartheta_{n-1}), \quad (7.20)$$

let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, for every $n, j \in \mathbb{N}$ let $\mathbf{m}_{n,j}: \Omega \rightarrow S$ be a uniformly distributed random variable, and let $\Theta: \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^d$ satisfy for all $n \in \mathbb{N}$ that

$$\Theta_0 = \xi \quad \text{and} \quad \Theta_n = \Theta_{n-1} - \gamma_n \left[\frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_\theta \ell)(\Theta_{n-1}, \mathbf{m}_{n,j}) \right] \quad (7.21)$$

(cf. Definitions 1.1.3 and 1.2.1 and Corollary 5.3.6). Then

- (i) it holds that ϑ is the **GD** process for the objective function \mathcal{L} with learning rates $(\gamma_n)_{n \in \mathbb{N}}$ and initial value ξ ,
- (ii) it holds that Θ is the **SGD** process for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, batch sizes $(J_n)_{n \in \mathbb{N}}$, initial value ξ , and data $(\mathbf{m}_{n,j})_{(n,j) \in \mathbb{N}^2}$, and
- (iii) it holds for all $n \in \mathbb{N}$, $\theta \in \mathbb{R}^d$ that

$$\mathbb{E} \left[\theta - \gamma_n \left[\frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_\theta \ell)(\theta, \mathbf{m}_{n,j}) \right] \right] = \theta - \gamma_n (\nabla \mathcal{L})(\theta) \quad (7.22)$$

(cf. Definitions 6.1.1 and 7.2.1).

Proof for Example 7.2.4. Note that (7.20) implies item (i). Observe that (7.21) proves item (ii). Note that (7.19), (7.18), and the assumption that for all $n, j \in \mathbb{N}$ it holds that $\mathbf{m}_{n,j}$ is uniformly distributed on $\{1, 2, \dots, M\}$ show that for all $\theta \in \mathbb{R}^d$, $n \in \mathbb{N}$ it holds that

$$\begin{aligned} \mathbb{E}[(\nabla_\theta \ell)(\theta, (x_{\mathbf{m}_{n,j}}, y_{\mathbf{m}_{n,j}}))] &= \frac{1}{M} \left[\sum_{m=1}^M (\nabla_\theta \ell)(\theta, (x_m, y_m)) \right] \\ &= \frac{1}{M} \left[\sum_{m=1}^M |(\mathcal{N}_{\mathfrak{M}_{a,l_1}, \mathfrak{M}_{a,l_2}, \dots, \mathfrak{M}_{a,l_h}, \text{id}_{\mathbb{R}}})^{\theta,d}(x_m) - y_m|^2 \right] = (\nabla \mathcal{L})(\theta). \end{aligned} \quad (7.23)$$

Therefore, we obtain for all $n \in \mathbb{N}$, $\theta \in \mathbb{R}^d$ that

$$\begin{aligned}\mathbb{E}\left[\theta - \gamma_n \left[\frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_\theta \ell)(\theta, \mathbf{m}_{n,j}) \right]\right] &= \theta - \gamma_n \left[\frac{1}{J_n} \sum_{j=1}^{J_n} \mathbb{E}[(\nabla_\theta \ell)(\theta, \mathbf{m}_{n,j})] \right] \\ &= \theta - \gamma_n \left[\frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla \mathcal{L})(\theta) \right] \\ &= \theta - \gamma_n (\nabla \mathcal{L})(\theta).\end{aligned}\tag{7.24}$$

The proof for Example 7.2.4 is thus complete. \square

Source codes 7.1 and 7.2 give two concrete implementations in PYTORCH of the framework described in Example 7.2.4 with different data and network architectures. The plots generated by these codes can be found in Figures 7.1 and 7.2, respectively. They show the approximations of the respective target functions by the realization functions of the ANNs at various points during the training.

```

1 import torch
2 import torch.nn as nn
3 import numpy as np
4 import matplotlib.pyplot as plt
5
6 M = 10000 # number of training samples
7
8 # We fix a random seed. This is not necessary for training a
9 # neural network, but we use it here to ensure that the same
10 # plot is created on every run.
11 torch.manual_seed(0)
12
13 # Here, we define the training set.
14 # Create a tensor of shape (M, 1) with entries sampled from a
15 # uniform distribution on [-2 * pi, 2 * pi]
16 X = (torch.rand((M, 1)) - 0.5) * 4 * np.pi
17 # We use the sine as the target function, so this defines the
18 # desired outputs.
19 Y = torch.sin(X)
20
21 J = 32 # the batch size
22 N = 100000 # the number of SGD iterations
23
24 loss = nn.MSELoss() # the mean squared error loss function
25 gamma = 0.003 # the learning rate
26
27 # Define a network with a single hidden layer of 200 neurons and
28 # tanh activation function
29 net = nn.Sequential(

```

```

30     nn.Linear(1, 200), nn.Tanh(), nn.Linear(200, 1)
31 )
32
33 # Set up a 3x3 grid of plots
34 fig, axs = plt.subplots(
35     3,
36     3,
37     figsize=(12, 8),
38     sharex="col",
39     sharey="row",
40 )
41
42 # Plot the target function
43 x = torch.linspace(-2 * np.pi, 2 * np.pi, 1000).reshape((1000, 1))
44 y = torch.sin(x)
45 for ax in axs.flatten():
46     ax.plot(x, y, label="Target")
47     ax.set_xlim([-2 * np.pi, 2 * np.pi])
48     ax.set_ylim([-1.1, 1.1])
49
50 plot_after = [1, 30, 100, 300, 1000, 3000, 10000, 30000, 100000]
51
52 # The training loop
53 for n in range(N):
54     # Choose J samples randomly from the training set
55     indices = torch.randint(0, M, (J,))
56     X_batch = X[indices]
57     Y_batch = Y[indices]
58
59     net.zero_grad() # Zero out the gradients
60
61     loss_val = loss(net(X_batch), Y_batch) # Compute the loss
62     loss_val.backward() # Compute the gradients
63
64     # Update the parameters
65     with torch.no_grad():
66         for p in net.parameters():
67             # Subtract the scaled gradient in-place
68             p.sub_(gamma * p.grad)
69
70     if n + 1 in plot_after:
71         # Plot the realization function of the ANN
72         i = plot_after.index(n + 1)
73         ax = axs[i // 3][i % 3]
74         ax.set_title(f"Batch {n+1}")
75
76         with torch.no_grad():
77             ax.plot(x, net(x), label="ANN realization")
78

```

```

79 | axs[0][0].legend(loc="upper right")
80 |
81 | plt.tight_layout()
82 | plt.savefig("../plots/sgd.pdf", bbox_inches="tight")

```

Source code 7.1 ([code/optimization_methods/sgd.py](#)): PYTHON code implementing the SGD optimization method in the training of an ANN as described in Example 7.2.4 in PYTORCH. In this code a fully-connected feedforward ANN with a single hidden layer with 200 neurons using the hyperbolic tangent activation function is trained so that the realization function approximates the target function $\sin: \mathbb{R} \rightarrow \mathbb{R}$. Example 7.2.4 is implemented with $d = 1$, $h = 1$, $\mathfrak{d} = 301$, $l_1 = 200$, $a = \tanh$, $M = 10000$, $x_1, x_2, \dots, x_M \in \mathbb{R}$, $y_i = \sin(x_i)$ for all $i \in \{1, 2, \dots, M\}$, $\gamma_n = 0.003$ for all $n \in \mathbb{N}$, and $J_n = 32$ for all $n \in \mathbb{N}$ in the notation of Example 7.2.4. The plot generated by this code is shown in Figure 7.1.

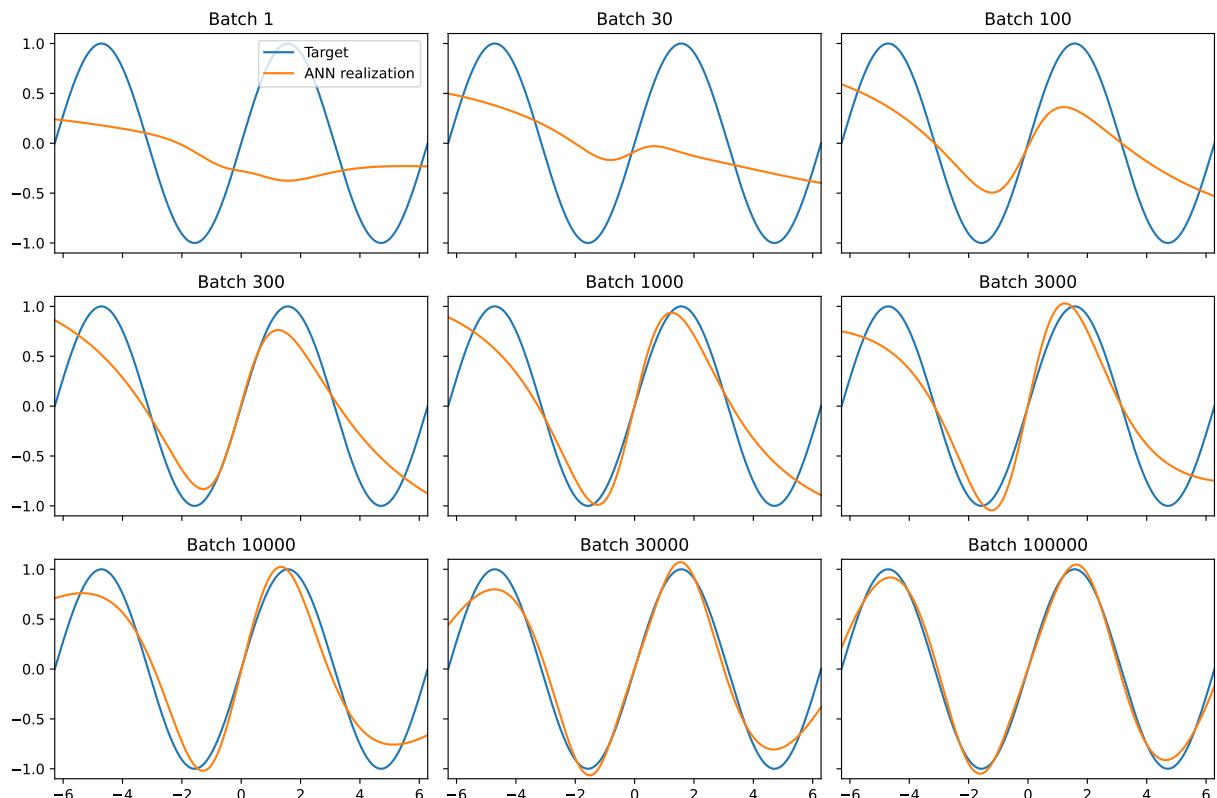


Figure 7.1 ([plots/sgd.pdf](#)): A plot showing the realization function of an ANN at several points during training with the SGD optimization method. This plot is generated by the code in Source code 7.1.

```

1 import torch
2 import torch.nn as nn
3 import numpy as np
4 import matplotlib.pyplot as plt
5
6 def plot_heatmap(ax, g):
7     x = np.linspace(-2 * np.pi, 2 * np.pi, 100)
8     y = np.linspace(-2 * np.pi, 2 * np.pi, 100)
9     x, y = np.meshgrid(x, y)
10
11    # flatten the grid to [num_points, 2] and convert to tensor
12    grid = np.vstack([x.flatten(), y.flatten()]).T
13    grid_torch = torch.from_numpy(grid).float()
14
15    # pass the grid through the network
16    z = g(grid_torch)
17
18    # reshape the predictions back to a 2D grid
19    Z = z.numpy().reshape(x.shape)
20
21    # plot the heatmap
22    ax.imshow(Z, origin='lower', extent=(-2 * np.pi, 2 * np.pi,
23                                         -2 * np.pi, 2 * np.pi))
24
25 M = 10000
26
27 def f(x):
28     return torch.sin(x).prod(dim=1, keepdim=True)
29
30 torch.manual_seed(0)
31 X = torch.rand((M, 2)) * 4 * np.pi - 2 * np.pi
32 Y = f(X)
33
34 J = 32
35
36 N = 100000
37
38 loss = nn.MSELoss()
39 gamma = 0.05
40
41 fig, axs = plt.subplots(
42     3, 3, figsize=(12, 12), sharex="col", sharey="row",
43 )
44
45 net = nn.Sequential(
46     nn.Linear(2, 50),
47     nn.Softplus(),
48     nn.Linear(50, 50),

```

```

49     nn.Softplus(),
50     nn.Linear(50, 1)
51 )
52
53 plot_after = [0, 100, 300, 1000, 3000, 10000, 30000, 100000]
54
55 for n in range(N + 1):
56     indices = torch.randint(0, M, (J,))
57
58     x = X[indices]
59     y = Y[indices]
60
61     net.zero_grad()
62
63     loss_val = loss(net(x), y)
64     loss_val.backward()
65
66     with torch.no_grad():
67         for p in net.parameters():
68             p.sub_(gamma * p.grad)
69
70     if n in plot_after:
71         i = plot_after.index(n)
72
73         with torch.no_grad():
74             plot_heatmap(axs[i // 3][i % 3], net)
75             axs[i // 3][i % 3].set_title(f"Batch {n}")
76
77     with torch.no_grad():
78         plot_heatmap(axs[2][2], f)
79         axs[2][2].set_title("Target")
80
81 plt.tight_layout()
82 plt.savefig("../plots/sgd2.pdf", bbox_inches="tight")

```

Source code 7.2 ([code/optimization_methods/sgd2.py](#)): PYTHON code implementing the SGD optimization method in the training of an ANN as described in Example 7.2.4 in PYTORCH. In this code a fully-connected feedforward ANN with two hidden layers with 50 neurons each using the softplus activation function is trained so that the realization function approximates the target function $f: \mathbb{R}^2 \rightarrow \mathbb{R}$ which satisfies for all $x, y \in \mathbb{R}$ that $f(x, y) = \sin(x) \sin(y)$. Example 7.2.4 is implemented with $d = 1$, $h = 2$, $\mathfrak{d} = 2701$, $l_1 = l_2 = 50$, a being the softplus activation function, $M = 10000$, $x_1, x_2, \dots, x_M \in \mathbb{R}^2$, $y_i = f(x_i)$ for all $i \in \{1, 2, \dots, M\}$, $\gamma_n = 0.003$ for all $n \in \mathbb{N}$, and $J_n = 32$ for all $n \in \mathbb{N}$ in the notation of Example 7.2.4. The plot generated by this code is shown in Figure 7.2.

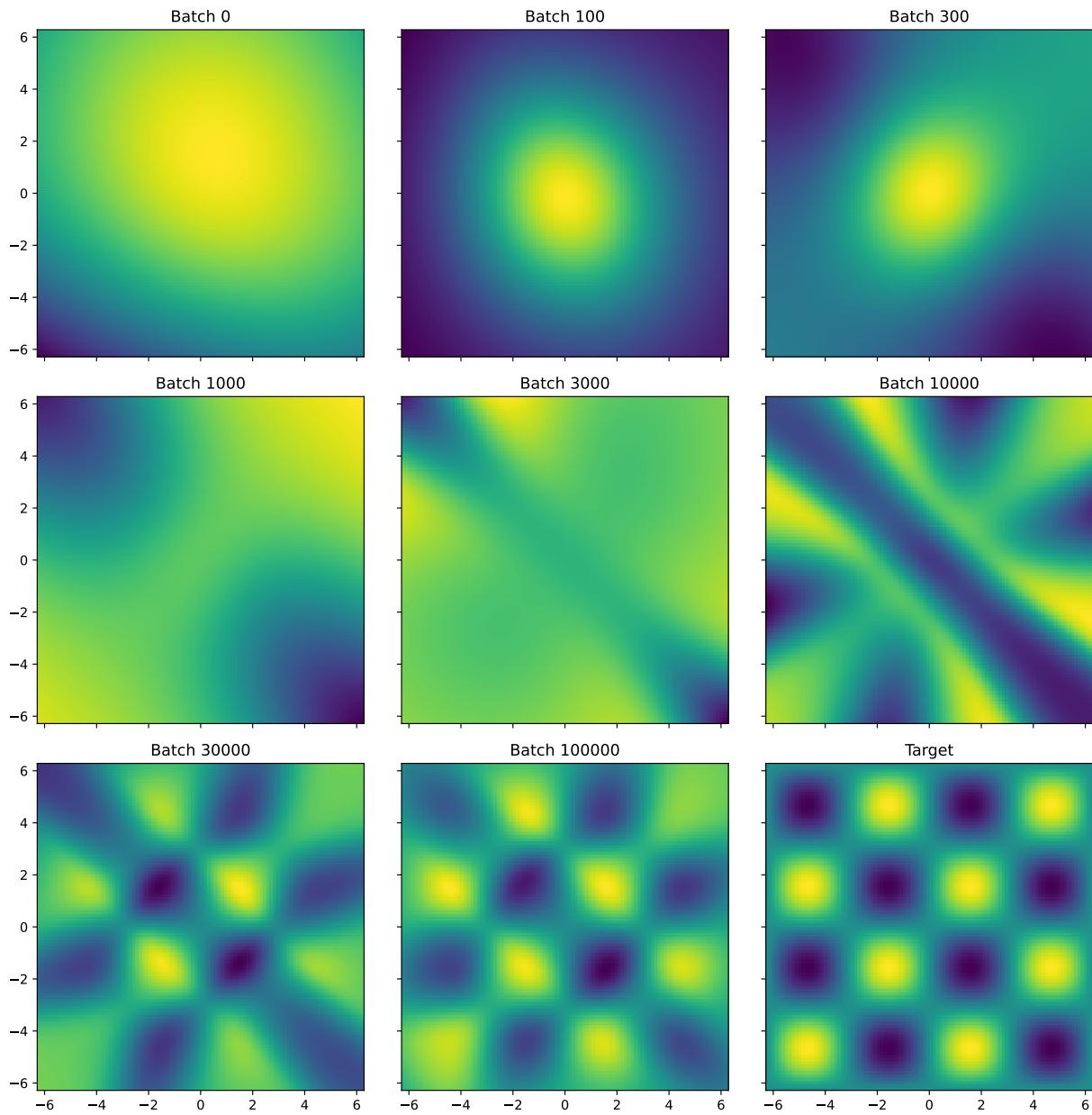


Figure 7.2 ([plots/sgd2.pdf](#)): A plot showing the realization function of an ANN at several points during training with the SGD optimization method. This plot is generated by the code in Source code 7.2.

7.2.2 Non-convergence of SGD for not appropriately decaying learning rates

In this section we present two results that, roughly speaking, motivate that the sequence of learning rates of the SGD optimization method should be chosen such that they converge to zero (see Corollary 7.2.12 below) but not too fast (see Lemma 7.2.15 below).

7.2.2.1 Bias-variance decomposition of the mean square error

Lemma 7.2.5 (Bias-variance decomposition of the mean square error). *Let $d \in \mathbb{N}$, $\vartheta \in \mathbb{R}^d$, let $\langle\langle \cdot, \cdot \rangle\rangle: \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ be a scalar product, let $\|\cdot\|: \mathbb{R}^d \rightarrow [0, \infty)$ satisfy for all $v \in \mathbb{R}^d$ that*

$$\|v\| = \sqrt{\langle\langle v, v \rangle\rangle}, \quad (7.25)$$

let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $Z: \Omega \rightarrow \mathbb{R}^d$ be a random variable with $\mathbb{E}[\|Z\|] < \infty$. Then

$$\mathbb{E}[\|Z - \vartheta\|^2] = \mathbb{E}[\|Z - \mathbb{E}[Z]\|^2] + \|\mathbb{E}[Z] - \vartheta\|^2. \quad (7.26)$$

Proof of Lemma 7.2.5. Observe that the assumption that $\mathbb{E}[\|Z\|] < \infty$ and the Cauchy-Schwarz inequality ensure that

$$\begin{aligned} \mathbb{E}[|\langle\langle Z - \mathbb{E}[Z], \mathbb{E}[Z] - \vartheta \rangle\rangle|] &\leq \mathbb{E}[\|Z - \mathbb{E}[Z]\| \|\mathbb{E}[Z] - \vartheta\|] \\ &\leq (\mathbb{E}[\|Z\|] + \|\mathbb{E}[Z]\|) \|\mathbb{E}[Z] - \vartheta\| < \infty. \end{aligned} \quad (7.27)$$

The linearity of the expectation hence proves that

$$\begin{aligned} \mathbb{E}[\|Z - \vartheta\|^2] &= \mathbb{E}[\|(Z - \mathbb{E}[Z]) + (\mathbb{E}[Z] - \vartheta)\|^2] \\ &= \mathbb{E}[\|Z - \mathbb{E}[Z]\|^2 + 2\langle\langle Z - \mathbb{E}[Z], \mathbb{E}[Z] - \vartheta \rangle\rangle + \|\mathbb{E}[Z] - \vartheta\|^2] \\ &= \mathbb{E}[\|Z - \mathbb{E}[Z]\|^2] + 2\langle\langle \mathbb{E}[Z] - \mathbb{E}[Z], \mathbb{E}[Z] - \vartheta \rangle\rangle + \|\mathbb{E}[Z] - \vartheta\|^2 \\ &= \mathbb{E}[\|Z - \mathbb{E}[Z]\|^2] + \|\mathbb{E}[Z] - \vartheta\|^2. \end{aligned} \quad (7.28)$$

The proof of Lemma 7.2.5 is thus complete. \square

7.2.2.2 Non-convergence of SGD for constant learning rates

In this section we present Lemma 7.2.11, Corollary 7.2.12, and Lemma 7.2.13. Our proof of Lemma 7.2.11 employs the auxiliary results in Lemmas 7.2.6, 7.2.7, 7.2.8, 7.2.9, and 7.2.10 below. Lemma 7.2.6 recalls an elementary and well known property for the expectation of the product of independent random variables (see, for instance, Klenke [262, Theorem 5.4]). In the elementary Lemma 7.2.10 we prove under suitable hypotheses the measurability of certain derivatives of a function. A result similar to Lemma 7.2.10 can, for example, be found in Jentzen et al. [281, Lemma 4.4].

Lemma 7.2.6. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X, Y: \Omega \rightarrow \mathbb{R}$ be independent random variables with $\mathbb{E}[|X| + |Y|] < \infty$. Then

- (i) it holds that $\mathbb{E}[|XY|] = \mathbb{E}[|X|]\mathbb{E}[|Y|] < \infty$ and
- (ii) it holds that $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.

Proof of Lemma 7.2.6. Note that the fact that $(X, Y)(\mathbb{P}) = (X(\mathbb{P})) \otimes (Y(\mathbb{P}))$, the integral transformation theorem, Fubini's theorem, and the assumption that $\mathbb{E}[|X| + |Y|] < \infty$ establish that

$$\begin{aligned}
 \mathbb{E}[|XY|] &= \int_{\Omega} |X(\omega)Y(\omega)| \mathbb{P}(d\omega) \\
 &= \int_{\mathbb{R} \times \mathbb{R}} |xy| ((X, Y)(\mathbb{P}))(dx, dy) \\
 &= \int_{\mathbb{R}} \left[\int_{\mathbb{R}} |xy| (X(\mathbb{P}))(dx) \right] (Y(\mathbb{P}))(dy) \\
 &= \int_{\mathbb{R}} |y| \left[\int_{\mathbb{R}} |x| (X(\mathbb{P}))(dx) \right] (Y(\mathbb{P}))(dy) \\
 &= \left[\int_{\mathbb{R}} |x| (X(\mathbb{P}))(dx) \right] \left[\int_{\mathbb{R}} |y| (Y(\mathbb{P}))(dy) \right] \\
 &= \mathbb{E}[|X|]\mathbb{E}[|Y|] < \infty.
 \end{aligned} \tag{7.29}$$

This proves item (i). Observe that item (i), the fact that $(X, Y)(\mathbb{P}) = (X(\mathbb{P})) \otimes (Y(\mathbb{P}))$, the integral transformation theorem, and Fubini's theorem demonstrate that

$$\begin{aligned}
 \mathbb{E}[XY] &= \int_{\Omega} X(\omega)Y(\omega) \mathbb{P}(d\omega) \\
 &= \int_{\mathbb{R} \times \mathbb{R}} xy ((X, Y)(\mathbb{P}))(dx, dy) \\
 &= \int_{\mathbb{R}} \left[\int_{\mathbb{R}} xy (X(\mathbb{P}))(dx) \right] (Y(\mathbb{P}))(dy) \\
 &= \int_{\mathbb{R}} y \left[\int_{\mathbb{R}} x (X(\mathbb{P}))(dx) \right] (Y(\mathbb{P}))(dy) \\
 &= \left[\int_{\mathbb{R}} x (X(\mathbb{P}))(dx) \right] \left[\int_{\mathbb{R}} y (Y(\mathbb{P}))(dy) \right] \\
 &= \mathbb{E}[X]\mathbb{E}[Y].
 \end{aligned} \tag{7.30}$$

This establishes item (ii). The proof of Lemma 7.2.6 is thus complete. \square

Lemma 7.2.7. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $d \in \mathbb{N}$, let $\langle\langle \cdot, \cdot \rangle\rangle: \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ be a scalar product, let $\|\cdot\|: \mathbb{R}^d \rightarrow [0, \infty)$ satisfy for all $v \in \mathbb{R}^d$ that

$$\|v\| = \sqrt{\langle\langle v, v \rangle\rangle}, \quad (7.31)$$

let $X: \Omega \rightarrow \mathbb{R}^d$ be a random variable, assume $\mathbb{E}[\|X\|^2] < \infty$, let $e_1, e_2, \dots, e_d \in \mathbb{R}^d$ satisfy for all $i, j \in \{1, 2, \dots, d\}$ that $\langle\langle e_i, e_j \rangle\rangle = \mathbb{1}_{\{i\}}(j)$, and for every random variable $Y: \Omega \rightarrow \mathbb{R}^d$ with $\mathbb{E}[\|Y\|^2] < \infty$ let $\text{Cov}(Y) \in \mathbb{R}^{d \times d}$ satisfy

$$\text{Cov}(Y) = (\mathbb{E}[\langle\langle e_i, Y - \mathbb{E}[Y] \rangle\rangle \langle\langle e_j, Y - \mathbb{E}[Y] \rangle\rangle])_{(i,j) \in \{1, 2, \dots, d\}^2}. \quad (7.32)$$

Then

$$\text{Trace}(\text{Cov}(X)) = \mathbb{E}[\|X - \mathbb{E}[X]\|^2]. \quad (7.33)$$

Proof of Lemma 7.2.7. First, note that the fact that $\forall i, j \in \{1, 2, \dots, d\}: \langle\langle e_i, e_j \rangle\rangle = \mathbb{1}_{\{i\}}(j)$ implies that for all $v \in \mathbb{R}^d$ it holds that $\sum_{i=1}^d \langle\langle e_i, v \rangle\rangle e_i = v$. Combining this with the fact that $\forall i, j \in \{1, 2, \dots, d\}: \langle\langle e_i, e_j \rangle\rangle = \mathbb{1}_{\{i\}}(j)$ shows that

$$\begin{aligned} \text{Trace}(\text{Cov}(X)) &= \sum_{i=1}^d \mathbb{E}[\langle\langle e_i, X - \mathbb{E}[X] \rangle\rangle \langle\langle e_i, X - \mathbb{E}[X] \rangle\rangle] \\ &= \sum_{i=1}^d \sum_{j=1}^d \mathbb{E}[\langle\langle e_i, X - \mathbb{E}[X] \rangle\rangle \langle\langle e_j, X - \mathbb{E}[X] \rangle\rangle \langle\langle e_i, e_j \rangle\rangle] \\ &= \mathbb{E}[\langle\langle \sum_{i=1}^d \langle\langle e_i, X - \mathbb{E}[X] \rangle\rangle e_i, \sum_{j=1}^d \langle\langle e_j, X - \mathbb{E}[X] \rangle\rangle e_j \rangle\rangle] \\ &= \mathbb{E}[\langle\langle X - \mathbb{E}[X], X - \mathbb{E}[X] \rangle\rangle] = \mathbb{E}[\|X - \mathbb{E}[X]\|^2]. \end{aligned} \quad (7.34)$$

The proof of Lemma 7.2.7 is thus complete. \square

Lemma 7.2.8. Let $d, n \in \mathbb{N}$, let $\langle\langle \cdot, \cdot \rangle\rangle: \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ be a scalar product, let $\|\cdot\|: \mathbb{R}^d \rightarrow [0, \infty)$ satisfy for all $v \in \mathbb{R}^d$ that

$$\|v\| = \sqrt{\langle\langle v, v \rangle\rangle}, \quad (7.35)$$

let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $X_k: \Omega \rightarrow \mathbb{R}^d$, $k \in \{1, 2, \dots, n\}$, be independent random variables, and assume $\sum_{k=1}^n \mathbb{E}[\|X_k\|] < \infty$. Then

$$\mathbb{E}\left[\left\|\sum_{k=1}^n (X_k - \mathbb{E}[X_k])\right\|^2\right] = \sum_{k=1}^n \mathbb{E}[\|X_k - \mathbb{E}[X_k]\|^2]. \quad (7.36)$$

Proof of Lemma 7.2.8. First, observe that Lemma 7.2.6 and the assumption that $\mathbb{E}[\|X_1\| + \|X_2\| + \dots + \|X_n\|] < \infty$ ensure that for all $k_1, k_2 \in \{1, 2, \dots, n\}$ with $k_1 \neq k_2$ it holds that

$$\mathbb{E}[|\langle X_{k_1} - \mathbb{E}[X_{k_1}], X_{k_2} - \mathbb{E}[X_{k_2}] \rangle|] \leq \mathbb{E}[\|X_{k_1} - \mathbb{E}[X_{k_1}]\| \|X_{k_2} - \mathbb{E}[X_{k_2}]\|] < \infty \quad (7.37)$$

and

$$\begin{aligned} & \mathbb{E}[\langle X_{k_1} - \mathbb{E}[X_{k_1}], X_{k_2} - \mathbb{E}[X_{k_2}] \rangle] \\ &= \langle \mathbb{E}[X_{k_1} - \mathbb{E}[X_{k_1}]], \mathbb{E}[X_{k_2} - \mathbb{E}[X_{k_2}]] \rangle \\ &= \langle \mathbb{E}[X_{k_1}] - \mathbb{E}[X_{k_1}], \mathbb{E}[X_{k_2}] - \mathbb{E}[X_{k_2}] \rangle = 0. \end{aligned} \quad (7.38)$$

Therefore, we obtain that

$$\begin{aligned} & \mathbb{E}\left[\left\|\sum_{k=1}^n (X_k - \mathbb{E}[X_k])\right\|^2\right] \\ &= \mathbb{E}\left[\langle\langle \sum_{k_1=1}^n (X_{k_1} - \mathbb{E}[X_{k_1}]), \sum_{k_2=1}^n (X_{k_2} - \mathbb{E}[X_{k_2}]) \rangle\rangle\right] \\ &= \mathbb{E}\left[\sum_{k_1, k_2=1}^n \langle\langle X_{k_1} - \mathbb{E}[X_{k_1}], X_{k_2} - \mathbb{E}[X_{k_2}] \rangle\rangle\right] \\ &= \mathbb{E}\left[\left(\sum_{k=1}^n \|X_k - \mathbb{E}[X_k]\|^2\right) + \left(\sum_{\substack{k_1, k_2 \in \{1, 2, \dots, n\}, \\ k_1 \neq k_2}} \langle\langle X_{k_1} - \mathbb{E}[X_{k_1}], X_{k_2} - \mathbb{E}[X_{k_2}] \rangle\rangle\right)\right] \\ &= \left(\sum_{k=1}^n \mathbb{E}[\|X_k - \mathbb{E}[X_k]\|^2]\right) + \left(\sum_{\substack{k_1, k_2 \in \{1, 2, \dots, n\}, \\ k_1 \neq k_2}} \mathbb{E}[\langle\langle X_{k_1} - \mathbb{E}[X_{k_1}], X_{k_2} - \mathbb{E}[X_{k_2}] \rangle\rangle]\right) \\ &= \sum_{k=1}^n \mathbb{E}[\|X_k - \mathbb{E}[X_k]\|^2]. \end{aligned} \quad (7.39)$$

The proof of Lemma 7.2.8 is thus complete. \square

Lemma 7.2.9 (Factorization lemma for independent random variables). *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $(\mathbb{X}, \mathcal{X})$ and $(\mathbb{Y}, \mathcal{Y})$ be measurable spaces, let $X: \Omega \rightarrow \mathbb{X}$ and $Y: \Omega \rightarrow \mathbb{Y}$ be independent random variables, let $\Phi: \mathbb{X} \times \mathbb{Y} \rightarrow [0, \infty]$ be $(\mathcal{X} \otimes \mathcal{Y})/\mathcal{B}([0, \infty])$ -measurable, and let $\phi: \mathbb{Y} \rightarrow [0, \infty]$ satisfy for all $y \in \mathbb{Y}$ that*

$$\phi(y) = \mathbb{E}[\Phi(X, y)]. \quad (7.40)$$

Then

- (i) it holds that the function ϕ is $\mathcal{Y}/\mathcal{B}([0, \infty])$ -measurable and

(ii) it holds that

$$\mathbb{E}[\Phi(X, Y)] = \mathbb{E}[\phi(Y)]. \quad (7.41)$$

Proof of Lemma 7.2.9. First, note that Fubini's theorem (cf., for instance, Klenke [262], (14.6) in Theorem 14.16]), the assumption that the function $X: \Omega \rightarrow \mathbb{X}$ is \mathcal{F}/\mathcal{X} -measurable, and the assumption that the function $\Phi: \mathbb{X} \times \mathbb{Y} \rightarrow [0, \infty]$ is $(\mathcal{X} \otimes \mathcal{Y})/\mathcal{B}([0, \infty])$ -measurable prove that the function

$$\mathbb{Y} \ni y \mapsto \phi(y) = \mathbb{E}[\Phi(X, y)] = \int_{\Omega} \Phi(X(\omega), y) \mathbb{P}(d\omega) \in [0, \infty] \quad (7.42)$$

is $\mathcal{Y}/\mathcal{B}([0, \infty])$ -measurable. This establishes item (i). Observe that the integral transformation theorem, the fact that $(X, Y)(\mathbb{P}) = (X(\mathbb{P})) \otimes (Y(\mathbb{P}))$, and Fubini's theorem demonstrate that

$$\begin{aligned} \mathbb{E}[\Phi(X, Y)] &= \int_{\Omega} \Phi(X(\omega), Y(\omega)) \mathbb{P}(d\omega) \\ &= \int_{\mathbb{X} \times \mathbb{Y}} \Phi(x, y) ((X, Y)(\mathbb{P}))(dx, dy) \\ &= \int_{\mathbb{Y}} \left[\int_{\mathbb{X}} \Phi(x, y) (X(\mathbb{P}))(dx) \right] (Y(\mathbb{P}))(dy) \\ &= \int_{\mathbb{Y}} \mathbb{E}[\Phi(X, y)] (Y(\mathbb{P}))(dy) \\ &= \int_{\mathbb{Y}} \phi(y) (Y(\mathbb{P}))(dy) = \mathbb{E}[\phi(Y)]. \end{aligned} \quad (7.43)$$

This proves item (ii). The proof of Lemma 7.2.9 is thus complete. \square

Lemma 7.2.10. Let $\mathfrak{d} \in \mathbb{N}$, let (S, \mathcal{S}) be a measurable space, let $\ell = (\ell(\theta, x))_{(\theta, x) \in \mathbb{R}^{\mathfrak{d}} \times S}$: $\mathbb{R}^{\mathfrak{d}} \times S \rightarrow \mathbb{R}$ be $(\mathcal{B}(\mathbb{R}^{\mathfrak{d}}) \otimes \mathcal{S})/\mathcal{B}(\mathbb{R})$ -measurable, and assume for every $x \in S$ that the function $\mathbb{R}^{\mathfrak{d}} \ni \theta \mapsto \ell(\theta, x) \in \mathbb{R}$ is differentiable. Then the function

$$\mathbb{R}^{\mathfrak{d}} \times S \ni (\theta, x) \mapsto (\nabla_{\theta} \ell)(\theta, x) \in \mathbb{R}^{\mathfrak{d}} \quad (7.44)$$

is $(\mathcal{B}(\mathbb{R}^{\mathfrak{d}}) \otimes \mathcal{S})/\mathcal{B}(\mathbb{R}^{\mathfrak{d}})$ -measurable.

Proof of Lemma 7.2.10. Throughout this proof, let $\varrho = (\varrho_1, \dots, \varrho_{\mathfrak{d}}): \mathbb{R}^{\mathfrak{d}} \times S \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$, $x \in S$ that

$$\varrho(\theta, x) = (\nabla_{\theta} \ell)(\theta, x). \quad (7.45)$$

The assumption that the function $\ell: \mathbb{R}^{\mathfrak{d}} \times S \rightarrow \mathbb{R}$ is $(\mathcal{B}(\mathbb{R}^{\mathfrak{d}}) \otimes \mathcal{S})/\mathcal{B}(\mathbb{R})$ -measurable implies that for all $i \in \{1, 2, \dots, \mathfrak{d}\}$, $h \in \mathbb{R} \setminus \{0\}$ it holds that the function

$$\mathbb{R}^{\mathfrak{d}} \times S \ni (\theta, x) = ((\theta_1, \dots, \theta_{\mathfrak{d}}), x) \mapsto \left(\frac{\ell((\theta_1, \dots, \theta_{i-1}, \theta_i + h, \theta_{i+1}, \dots, \theta_{\mathfrak{d}}), x) - \ell(\theta, x)}{h} \right) \in \mathbb{R} \quad (7.46)$$

is $(\mathcal{B}(\mathbb{R}^{\mathfrak{d}}) \otimes \mathcal{S})/\mathcal{B}(\mathbb{R})$ -measurable. The fact that for all $i \in \{1, 2, \dots, \mathfrak{d}\}$, $\theta = (\theta_1, \dots, \theta_{\mathfrak{d}}) \in \mathbb{R}^{\mathfrak{d}}$, $x \in S$ it holds that

$$\varrho_i(\theta, x) = \lim_{n \rightarrow \infty} \left(\frac{\ell((\theta_1, \dots, \theta_{i-1}, \theta_i + 2^{-n}, \theta_{i+1}, \dots, \theta_{\mathfrak{d}}), x) - \ell(\theta, x)}{2^{-n}} \right) \quad (7.47)$$

hence shows that for all $i \in \{1, 2, \dots, \mathfrak{d}\}$ it holds that the function $\varrho_i: \mathbb{R}^{\mathfrak{d}} \times S \rightarrow \mathbb{R}$ is $(\mathcal{B}(\mathbb{R}^{\mathfrak{d}}) \otimes \mathcal{S})/\mathcal{B}(\mathbb{R})$ -measurable. This ensures that ϱ is $(\mathcal{B}(\mathbb{R}^{\mathfrak{d}}) \otimes \mathcal{S})/\mathcal{B}(\mathbb{R}^{\mathfrak{d}})$ -measurable. The proof of Lemma 7.2.10 is thus complete. \square

Lemma 7.2.11. *Let $\mathfrak{d} \in \mathbb{N}$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq (0, \infty)$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, let $\langle\!\langle \cdot, \cdot \rangle\!\rangle: \mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$ be a scalar product, let $\|\cdot\|: \mathbb{R}^{\mathfrak{d}} \rightarrow [0, \infty)$ satisfy for all $v \in \mathbb{R}^{\mathfrak{d}}$ that*

$$\|\cdot\| = \sqrt{\langle\!\langle v, v \rangle\!\rangle}, \quad (7.48)$$

let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\xi: \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a random variable, let (S, \mathcal{S}) be a measurable space, let $X_{n,j}: \Omega \rightarrow S$, $j \in \{1, 2, \dots, J_n\}$, $n \in \mathbb{N}$, be i.i.d. random variables, assume that ξ and $(X_{n,j})_{(n,j) \in \{(k,l) \in \mathbb{N}^2 : l \leq J_k\}}$ are independent, let $\ell = (\ell(\theta, x))_{(\theta,x) \in \mathbb{R}^{\mathfrak{d}} \times S}: \mathbb{R}^{\mathfrak{d}} \times S \rightarrow \mathbb{R}$ be $(\mathcal{B}(\mathbb{R}^{\mathfrak{d}}) \otimes \mathcal{S})/\mathcal{B}(\mathbb{R})$ -measurable, assume for all $x \in S$ that $(\mathbb{R}^{\mathfrak{d}} \ni \theta \mapsto \ell(\theta, x) \in \mathbb{R}) \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$, assume for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that $\mathbb{E}[\|(\nabla_{\theta}\ell)(\theta, X_{1,1})\|] < \infty$ (cf. Lemma 7.2.10), let $\mathcal{V}: \mathbb{R}^{\mathfrak{d}} \rightarrow [0, \infty]$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that

$$\mathcal{V}(\theta) = \mathbb{E}[\|(\nabla_{\theta}\ell)(\theta, X_{1,1}) - \mathbb{E}[(\nabla_{\theta}\ell)(\theta, X_{1,1})]\|^2], \quad (7.49)$$

and let $\Theta: \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ be the stochastic process which satisfies for all $n \in \mathbb{N}$ that

$$\Theta_0 = \xi \quad \text{and} \quad \Theta_n = \Theta_{n-1} - \gamma_n \left[\frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_{\theta}\ell)(\Theta_{n-1}, X_{n,j}) \right]. \quad (7.50)$$

Then it holds for all $n \in \mathbb{N}$, $\vartheta \in \mathbb{R}^{\mathfrak{d}}$ that

$$(\mathbb{E}[\|\Theta_n - \vartheta\|^2])^{1/2} \geq \left[\frac{\gamma_n}{(J_n)^{1/2}} \right] (\mathbb{E}[\mathcal{V}(\Theta_{n-1})])^{1/2}. \quad (7.51)$$

Proof of Lemma 7.2.11. Throughout this proof, for every $n \in \mathbb{N}$ let $\phi_n: \mathbb{R}^{\mathfrak{d}} \rightarrow [0, \infty]$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that

$$\phi_n(\theta) = \mathbb{E} \left[\left\| \theta - \frac{\gamma_n}{J_n} \left[\sum_{j=1}^{J_n} (\nabla_{\theta}\ell)(\theta, X_{n,j}) \right] - \vartheta \right\|^2 \right]. \quad (7.52)$$

Note that Lemma 7.2.5 establishes that for all $\vartheta \in \mathbb{R}^{\mathfrak{d}}$ and all random variables $Z: \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ with $\mathbb{E}[\|Z\|] < \infty$ it holds that

$$\mathbb{E}[\|Z - \vartheta\|^2] = \mathbb{E}[\|Z - \mathbb{E}[Z]\|^2] + \|\mathbb{E}[Z] - \vartheta\|^2 \geq \mathbb{E}[\|Z - \mathbb{E}[Z]\|^2]. \quad (7.53)$$

Therefore, we obtain for all $n \in \mathbb{N}$, $\theta \in \mathbb{R}^d$ that

$$\begin{aligned}\phi_n(\theta) &= \mathbb{E} \left[\left\| \frac{\gamma_n}{J_n} \left[\sum_{j=1}^{J_n} (\nabla_\theta \ell)(\theta, X_{n,j}) \right] - (\theta - \vartheta) \right\|^2 \right] \\ &\geq \mathbb{E} \left[\left\| \frac{\gamma_n}{J_n} \left[\sum_{j=1}^{J_n} (\nabla_\theta \ell)(\theta, X_{n,j}) \right] - \mathbb{E} \left[\frac{\gamma_n}{J_n} \left[\sum_{j=1}^{J_n} (\nabla_\theta \ell)(\theta, X_{n,j}) \right] \right] \right\|^2 \right] \\ &= \frac{(\gamma_n)^2}{(J_n)^2} \mathbb{E} \left[\left\| \sum_{j=1}^{J_n} ((\nabla_\theta \ell)(\theta, X_{n,j}) - \mathbb{E}[(\nabla_\theta \ell)(\theta, X_{n,j})]) \right\|^2 \right].\end{aligned}\quad (7.54)$$

Lemma 7.2.8, the fact that $X_{n,j}: \Omega \rightarrow S$, $j \in \{1, 2, \dots, J_n\}$, $n \in \mathbb{N}$, are i.i.d. random variables, and the fact that for all $n \in \mathbb{N}$, $j \in \{1, 2, \dots, J_n\}$, $\theta \in \mathbb{R}^d$ it holds that

$$\mathbb{E}[\|(\nabla_\theta \ell)(\theta, X_{n,j})\|] = \mathbb{E}[\|(\nabla_\theta \ell)(\theta, X_{1,1})\|] < \infty \quad (7.55)$$

hence demonstrate that for all $n \in \mathbb{N}$, $\theta \in \mathbb{R}^d$ it holds that

$$\begin{aligned}\phi_n(\theta) &\geq \frac{(\gamma_n)^2}{(J_n)^2} \left[\sum_{j=1}^{J_n} \mathbb{E} \left[\|(\nabla_\theta \ell)(\theta, X_{n,j}) - \mathbb{E}[(\nabla_\theta \ell)(\theta, X_{n,j})]\|^2 \right] \right] \\ &= \frac{(\gamma_n)^2}{(J_n)^2} \left[\sum_{j=1}^{J_n} \mathbb{E} \left[\|(\nabla_\theta \ell)(\theta, X_{1,1}) - \mathbb{E}[(\nabla_\theta \ell)(\theta, X_{1,1})]\|^2 \right] \right] \\ &= \frac{(\gamma_n)^2}{(J_n)^2} \left[\sum_{j=1}^{J_n} \mathcal{V}(\theta) \right] = \frac{(\gamma_n)^2}{(J_n)^2} [J_n \mathcal{V}(\theta)] = \left(\frac{(\gamma_n)^2}{J_n} \right) \mathcal{V}(\theta).\end{aligned}\quad (7.56)$$

Furthermore, observe that (7.50), (7.52), the fact that for all $n \in \mathbb{N}$ it holds that Θ_{n-1} and $(X_{n,j})_{j \in \{1, 2, \dots, J_n\}}$ are independent random variables, and Lemma 7.2.9 prove that for all $n \in \mathbb{N}$, $\vartheta \in \mathbb{R}^d$ it holds that

$$\begin{aligned}\mathbb{E}[\|\Theta_n - \vartheta\|^2] &= \mathbb{E} \left[\left\| \Theta_{n-1} - \frac{\gamma_n}{J_n} \left[\sum_{j=1}^{J_n} (\nabla_\theta \ell)(\Theta_{n-1}, X_{n,j}) \right] - \vartheta \right\|^2 \right] \\ &= \mathbb{E}[\phi_n(\Theta_{n-1})].\end{aligned}\quad (7.57)$$

Combining this with (7.56) implies that for all $n \in \mathbb{N}$, $\vartheta \in \mathbb{R}^d$ it holds that

$$\mathbb{E}[\|\Theta_n - \vartheta\|^2] \geq \mathbb{E} \left[\left(\frac{(\gamma_n)^2}{J_n} \right) \mathcal{V}(\Theta_{n-1}) \right] = \left(\frac{(\gamma_n)^2}{J_n} \right) \mathbb{E}[\mathcal{V}(\Theta_{n-1})]. \quad (7.58)$$

This establishes (7.51). The proof of Lemma 7.2.11 is thus complete. \square

Corollary 7.2.12. Let $\mathfrak{d} \in \mathbb{N}$, $\varepsilon \in (0, \infty)$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq (0, \infty)$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, let $\langle \cdot, \cdot \rangle: \mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$ be a scalar product, let $\|\cdot\|: \mathbb{R}^{\mathfrak{d}} \rightarrow [0, \infty)$ satisfy for all $v \in \mathbb{R}^{\mathfrak{d}}$ that

$$\|v\| = \sqrt{\langle v, v \rangle}, \quad (7.59)$$

let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\xi: \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a random variable, let (S, \mathcal{S}) be a measurable space, let $X_{n,j}: \Omega \rightarrow S$, $j \in \{1, 2, \dots, J_n\}$, $n \in \mathbb{N}$, be i.i.d. random variables, assume that ξ and $(X_{n,j})_{(n,j) \in \{(k,l) \in \mathbb{N}^2 : l \leq J_k\}}$ are independent, let $\ell = (\ell(\theta, x))_{(\theta,x) \in \mathbb{R}^{\mathfrak{d}} \times S}: \mathbb{R}^{\mathfrak{d}} \times S \rightarrow \mathbb{R}$ be $(\mathcal{B}(\mathbb{R}^{\mathfrak{d}}) \otimes \mathcal{S})/\mathcal{B}(\mathbb{R})$ -measurable, assume for all $x \in S$ that $(\mathbb{R}^{\mathfrak{d}} \ni \theta \mapsto \ell(\theta, x) \in \mathbb{R}) \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$, assume for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that $\mathbb{E}[\|(\nabla_{\theta}\ell)(\theta, X_{1,1})\|] < \infty$ (cf. Lemma 7.2.10) and

$$(\mathbb{E}[\|(\nabla_{\theta}\ell)(\theta, X_{1,1}) - \mathbb{E}[(\nabla_{\theta}\ell)(\theta, X_{1,1})]\|^2])^{1/2} \geq \varepsilon, \quad (7.60)$$

and let $\Theta: \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ be the stochastic process which satisfies for all $n \in \mathbb{N}$ that

$$\Theta_0 = \xi \quad \text{and} \quad \Theta_n = \Theta_{n-1} - \gamma_n \left[\frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_{\theta}\ell)(\Theta_{n-1}, X_{n,j}) \right]. \quad (7.61)$$

Then

(i) it holds for all $n \in \mathbb{N}$, $\vartheta \in \mathbb{R}^{\mathfrak{d}}$ that

$$(\mathbb{E}[\|\Theta_n - \vartheta\|^2])^{1/2} \geq \varepsilon \left(\frac{\gamma_n}{(J_n)^{1/2}} \right) \quad (7.62)$$

and

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

$$\liminf_{n \rightarrow \infty} (\mathbb{E}[\|\Theta_n - \vartheta\|^2])^{1/2} \geq \varepsilon \left(\liminf_{n \rightarrow \infty} \left[\frac{\gamma_n}{(J_n)^{1/2}} \right] \right). \quad (7.63)$$

Proof of Corollary 7.2.12. Throughout this proof, let $\mathcal{V}: \mathbb{R}^{\mathfrak{d}} \rightarrow [0, \infty]$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that

$$\mathcal{V}(\theta) = \mathbb{E}[\|(\nabla_{\theta}\ell)(\theta, X_{1,1}) - \mathbb{E}[(\nabla_{\theta}\ell)(\theta, X_{1,1})]\|^2]. \quad (7.64)$$

Note that (7.60) shows that for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ it holds that

$$\mathcal{V}(\theta) \geq \varepsilon^2. \quad (7.65)$$

Lemma 7.2.11 therefore ensures that for all $n \in \mathbb{N}$, $\vartheta \in \mathbb{R}^{\mathfrak{d}}$ it holds that

$$(\mathbb{E}[\|\Theta_n - \vartheta\|^2])^{1/2} \geq \frac{\gamma_n}{(J_n)^{1/2}} (\mathbb{E}[\mathcal{V}(\Theta_{n-1})])^{1/2} \geq \left[\frac{\gamma_n}{(J_n)^{1/2}} \right] (\varepsilon^2)^{1/2} = \frac{\gamma_n \varepsilon}{(J_n)^{1/2}}. \quad (7.66)$$

This demonstrates item (i). Observe that item (i) implies item (ii). The proof of Corollary 7.2.12 is thus complete. \square

Lemma 7.2.13 (Lower bound for the SGD optimization method). *Let $\mathfrak{d} \in \mathbb{N}$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq (0, \infty)$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\xi: \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a random variable, let $X_{n,j}: \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$, $j \in \{1, 2, \dots, J_n\}$, $n \in \mathbb{N}$, be i.i.d. random variables with $\mathbb{E}[\|X_{1,1}\|_2] < \infty$, assume that ξ and $(X_{n,j})_{(n,j) \in \{(k,l) \in \mathbb{N}^2 : l \leq J_k\}}$ are independent, let $\ell = (\ell(\theta, x))_{(\theta,x) \in \mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\mathfrak{d}}} : \mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$ satisfy for all $\theta, x \in \mathbb{R}^{\mathfrak{d}}$ that*

$$\ell(\theta, x) = \frac{1}{2}\|\theta - x\|_2^2, \quad (7.67)$$

and let $\Theta: \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ be the stochastic process which satisfies for all $n \in \mathbb{N}$ that

$$\Theta_0 = \xi \quad \text{and} \quad \Theta_n = \Theta_{n-1} - \gamma_n \left[\frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_{\theta} \ell)(\Theta_{n-1}, X_{n,j}) \right]. \quad (7.68)$$

Then

(i) it holds for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that

$$\mathbb{E}[(\nabla_{\theta} \ell)(\theta, X_{1,1})] < \infty, \quad (7.69)$$

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

$$\mathbb{E}\left[\|(\nabla_{\theta} \ell)(\theta, X_{1,1}) - \mathbb{E}[(\nabla_{\theta} \ell)(\theta, X_{1,1})]\|_2^2\right] = \mathbb{E}[\|X_{1,1} - \mathbb{E}[X_{1,1}]\|_2^2], \quad (7.70)$$

and

(iii) it holds for all $n \in \mathbb{N}$, $\vartheta \in \mathbb{R}^{\mathfrak{d}}$ that

$$(\mathbb{E}[\|\Theta_n - \vartheta\|_2^2])^{1/2} \geq (\mathbb{E}[\|X_{1,1} - \mathbb{E}[X_{1,1}]\|_2^2])^{1/2} \left[\frac{\gamma_n}{(J_n)^{1/2}} \right]. \quad (7.71)$$

Proof of Lemma 7.2.13. First, note that (7.67) and Lemma 5.8.4 prove that for all $\theta, x \in \mathbb{R}^{\mathfrak{d}}$ it holds that

$$(\nabla_{\theta} \ell)(\theta, x) = \frac{1}{2}(2(\theta - x)) = \theta - x. \quad (7.72)$$

The assumption that $\mathbb{E}[\|X_{1,1}\|_2] < \infty$ hence implies that for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ it holds that

$$\mathbb{E}[(\nabla_{\theta} \ell)(\theta, X_{1,1})] = \mathbb{E}[\|\theta - X_{1,1}\|_2] \leq \|\theta\|_2 + \mathbb{E}[\|X_{1,1}\|_2] < \infty. \quad (7.73)$$

This establishes item (i). Furthermore, observe that (7.72) and item (i) show that for all $\theta \in \mathbb{R}^{\vartheta}$ it holds that

$$\begin{aligned} & \mathbb{E}[\|(\nabla_{\theta}\ell)(\theta, X_{1,1}) - \mathbb{E}[(\nabla_{\theta}\ell)(\theta, X_{1,1})]\|_2^2] \\ &= \mathbb{E}[\|(\theta - X_{1,1}) - \mathbb{E}[\theta - X_{1,1}]\|_2^2] = \mathbb{E}[\|X_{1,1} - \mathbb{E}[X_{1,1}]\|_2^2]. \end{aligned} \quad (7.74)$$

This proves item (ii). Note that item (i) in Corollary 7.2.12 and items (i) and (ii) establish item (iii). The proof of Lemma 7.2.13 is thus complete. \square

7.2.2.3 Non-convergence of GD for summable learning rates

In the next auxiliary result, Lemma 7.2.14 below, we recall a well known lower bound for the natural logarithm.

Lemma 7.2.14 (A lower bound for the natural logarithm). *It holds for all $x \in (0, \infty)$ that*

$$\ln(x) \geq \frac{(x-1)}{x}. \quad (7.75)$$

Proof of Lemma 7.2.14. First, observe that the fundamental theorem of calculus ensures that for all $x \in [1, \infty)$ it holds that

$$\ln(x) = \ln(x) - \ln(1) = \int_1^x \frac{1}{t} dt \geq \int_1^x \frac{1}{x} dt = \frac{(x-1)}{x}. \quad (7.76)$$

Furthermore, note that the fundamental theorem of calculus demonstrates that for all $x \in (0, 1]$ it holds that

$$\begin{aligned} \ln(x) &= \ln(x) - \ln(1) = -(\ln(1) - \ln(x)) = -\left[\int_x^1 \frac{1}{t} dt\right] \\ &= \int_x^1 \left(-\frac{1}{t}\right) dt \geq \int_x^1 \left(-\frac{1}{x}\right) dt = (1-x)\left(-\frac{1}{x}\right) = \frac{(x-1)}{x}. \end{aligned} \quad (7.77)$$

This and (7.76) prove (7.75). The proof of Lemma 7.2.14 is thus complete. \square

Lemma 7.2.15 (**GD** fails to converge for a summable sequence of learning rates). *Let $\vartheta \in \mathbb{N}$, $\vartheta \in \mathbb{R}^{\vartheta}$, $\xi \in \mathbb{R}^{\vartheta} \setminus \{\vartheta\}$, $\alpha \in (0, \infty)$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty) \setminus \{1/\alpha\}$ satisfy $\sum_{n=1}^{\infty} \gamma_n < \infty$, let $\mathcal{L}: \mathbb{R}^{\vartheta} \rightarrow \mathbb{R}$ satisfy for all $\theta \in \mathbb{R}^{\vartheta}$ that*

$$\mathcal{L}(\theta) = \frac{\alpha}{2} \|\theta - \vartheta\|_2^2, \quad (7.78)$$

and let $\Theta: \mathbb{N}_0 \rightarrow \mathbb{R}^{\vartheta}$ satisfy for all $n \in \mathbb{N}$ that $\Theta_0 = \xi$ and

$$\Theta_n = \Theta_{n-1} - \gamma_n (\nabla \mathcal{L})(\Theta_{n-1}). \quad (7.79)$$

Then

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

$$\Theta_n - \vartheta = \left[\prod_{k=1}^n (1 - \gamma_k \alpha) \right] (\xi - \vartheta), \quad (7.80)$$

(ii) it holds that

$$\liminf_{n \rightarrow \infty} \left[\prod_{k=1}^n |1 - \gamma_k \alpha| \right] > 0, \quad (7.81)$$

and

(iii) it holds that

$$\liminf_{n \rightarrow \infty} \|\Theta_n - \vartheta\|_2 > 0. \quad (7.82)$$

Proof of Lemma 7.2.15. Throughout this proof, let $m \in \mathbb{N}$ satisfy for all $k \in \mathbb{N} \cap [m, \infty)$ that $\gamma_k < 1/(2\alpha)$. Observe that Lemma 5.8.4 implies that for all $\theta \in \mathbb{R}^\vartheta$ it holds that

$$(\nabla \mathcal{L})(\theta) = \frac{\alpha}{2}(2(\theta - \vartheta)) = \alpha(\theta - \vartheta). \quad (7.83)$$

Therefore, we obtain for all $n \in \mathbb{N}$ that

$$\begin{aligned} \Theta_n - \vartheta &= \Theta_{n-1} - \gamma_n (\nabla \mathcal{L})(\Theta_{n-1}) - \vartheta \\ &= \Theta_{n-1} - \gamma_n \alpha (\Theta_{n-1} - \vartheta) - \vartheta \\ &= (1 - \gamma_n \alpha) (\Theta_{n-1} - \vartheta). \end{aligned} \quad (7.84)$$

Induction hence shows that for all $n \in \mathbb{N}$ it holds that

$$\Theta_n - \vartheta = \left[\prod_{k=1}^n (1 - \gamma_k \alpha) \right] (\Theta_0 - \vartheta), \quad (7.85)$$

This and the assumption that $\Theta_0 = \xi$ establish item (i). Note that the fact that for all $k \in \mathbb{N}$ it holds that $\gamma_k \alpha \neq 1$ ensures that

$$\prod_{k=1}^{m-1} |1 - \gamma_k \alpha| > 0. \quad (7.86)$$

Moreover, note that the fact that for all $k \in \mathbb{N} \cap [m, \infty)$ it holds that $\gamma_k \alpha \in [0, 1/2)$ assures that for all $k \in \mathbb{N} \cap [m, \infty)$ it holds that

$$(1 - \gamma_k \alpha) \in (1/2, 1]. \quad (7.87)$$

This, Lemma 7.2.14, and the assumption that $\sum_{n=1}^{\infty} \gamma_n < \infty$ demonstrate that for all $n \in \mathbb{N} \cap [m, \infty)$ it holds that

$$\begin{aligned} \ln \left(\prod_{k=m}^n |1 - \gamma_k \alpha| \right) &= \sum_{k=m}^n \ln(1 - \gamma_k \alpha) \\ &\geq \sum_{k=m}^n \frac{(1 - \gamma_k \alpha) - 1}{(1 - \gamma_k \alpha)} = \sum_{k=m}^n \left[-\frac{\gamma_k \alpha}{(1 - \gamma_k \alpha)} \right] \\ &\geq \sum_{k=m}^n \left[-\frac{\gamma_k \alpha}{(\frac{1}{2})} \right] = -2\alpha \left[\sum_{k=m}^n \gamma_k \right] \geq -2\alpha \left[\sum_{k=1}^{\infty} \gamma_k \right] > -\infty. \end{aligned} \quad (7.88)$$

Combining this with (7.86) proves that for all $n \in \mathbb{N} \cap [m, \infty)$ it holds that

$$\begin{aligned} \prod_{k=1}^n |1 - \gamma_k \alpha| &= \left[\prod_{k=1}^{m-1} |1 - \gamma_k \alpha| \right] \exp \left(\ln \left(\prod_{k=m}^n |1 - \gamma_k \alpha| \right) \right) \\ &\geq \left[\prod_{k=1}^{m-1} |1 - \gamma_k \alpha| \right] \exp \left(-2\alpha \left[\sum_{k=1}^{\infty} \gamma_k \right] \right) > 0. \end{aligned} \quad (7.89)$$

Therefore, we obtain that

$$\liminf_{n \rightarrow \infty} \left[\prod_{k=1}^n |1 - \gamma_k \alpha| \right] \geq \left[\prod_{k=1}^{m-1} |1 - \gamma_k \alpha| \right] \exp \left(-2\alpha \left[\sum_{k=1}^{\infty} \gamma_k \right] \right) > 0. \quad (7.90)$$

This establishes item (ii). Observe that items (i) and (ii) and the assumption that $\xi \neq \vartheta$ imply that

$$\begin{aligned} \liminf_{n \rightarrow \infty} \|\Theta_n - \vartheta\|_2 &= \liminf_{n \rightarrow \infty} \left\| \left[\prod_{k=1}^n (1 - \gamma_k \alpha) \right] (\xi - \vartheta) \right\|_2 \\ &= \liminf_{n \rightarrow \infty} \left(\left\| \prod_{k=1}^n (1 - \gamma_k \alpha) \right\| \|\xi - \vartheta\|_2 \right) \\ &= \|\xi - \vartheta\|_2 \left(\liminf_{n \rightarrow \infty} \left[\prod_{k=1}^n |1 - \gamma_k \alpha| \right] \right) > 0. \end{aligned} \quad (7.91)$$

This proves item (iii). The proof of Lemma 7.2.15 is thus complete. \square

7.2.3 Convergence rates for SGD for quadratic objective functions

Example 7.2.16 below, in particular, provides an error analysis for the SGD optimization method in the case of one specific stochastic optimization problem (see (7.92) below). More

general error analyses for the SGD optimization method can, for example, be found in [232, 242] and the references therein (cf. Section 7.2.3 below).

Example 7.2.16 (Example of an SGD process). *Let $\mathfrak{d} \in \mathbb{N}$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $X_n: \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$, $n \in \mathbb{N}$, be i.i.d. random variables with $\mathbb{E}[\|X_1\|_2^2] < \infty$, let $\ell = (\ell(\theta, x))_{(\theta, x) \in \mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\mathfrak{d}}} : \mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$ and $\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$ satisfy for all $\theta, x \in \mathbb{R}^{\mathfrak{d}}$ that*

$$\ell(\theta, x) = \frac{1}{2}\|\theta - x\|_2^2 \quad \text{and} \quad \mathcal{L}(\theta) = \mathbb{E}[\ell(\theta, X_1)], \quad (7.92)$$

and let $\Theta: \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ be the stochastic process which satisfies for all $n \in \mathbb{N}$ that $\Theta_0 = 0$ and

$$\Theta_n = \Theta_{n-1} - \frac{1}{n}(\nabla_{\theta}\ell)(\Theta_{n-1}, X_n) \quad (7.93)$$

(cf. Definition 3.3.4). Then

(i) it holds that $\{\theta \in \mathbb{R}^{\mathfrak{d}}: \mathcal{L}(\theta) = \inf_{w \in \mathbb{R}^{\mathfrak{d}}} \mathcal{L}(w)\} = \{\mathbb{E}[X_1]\}$,

(ii) it holds for all $n \in \mathbb{N}$ that $\Theta_n = \frac{1}{n}(X_1 + X_2 + \dots + X_n)$,

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

$$(\mathbb{E}[\|\Theta_n - \mathbb{E}[X_1]\|_2^2])^{1/2} = (\mathbb{E}[\|X_1 - \mathbb{E}[X_1]\|_2^2])^{1/2} n^{-1/2}, \quad (7.94)$$

and

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

$$\mathbb{E}[\mathcal{L}(\Theta_n)] - \mathcal{L}(\mathbb{E}[X_1]) = \frac{1}{2}\mathbb{E}[\|X_1 - \mathbb{E}[X_1]\|_2^2]n^{-1}. \quad (7.95)$$

Proof for Example 7.2.16. Note that the assumption that $\mathbb{E}[\|X_1\|_2^2] < \infty$ and Lemma 7.2.5 show that for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ it holds that

$$\begin{aligned} \mathcal{L}(\theta) &= \mathbb{E}[\ell(\theta, X_1)] = \frac{1}{2}\mathbb{E}[\|X_1 - \theta\|_2^2] \\ &= \frac{1}{2}(\mathbb{E}[\|X_1 - \mathbb{E}[X_1]\|_2^2] + \|\theta - \mathbb{E}[X_1]\|_2^2). \end{aligned} \quad (7.96)$$

This establishes item (i). Observe that Lemma 5.8.4 ensures that for all $\theta, x \in \mathbb{R}^{\mathfrak{d}}$ it holds that

$$(\nabla_{\theta}\ell)(\theta, x) = \frac{1}{2}(2(\theta - x)) = \theta - x. \quad (7.97)$$

This and (7.93) assure that for all $n \in \mathbb{N}$ it holds that

$$\Theta_n = \Theta_{n-1} - \frac{1}{n}(\Theta_{n-1} - X_n) = (1 - \frac{1}{n})\Theta_{n-1} + \frac{1}{n}X_n = \frac{(n-1)}{n}\Theta_{n-1} + \frac{1}{n}X_n. \quad (7.98)$$

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

$$\Theta_n = \frac{1}{n}(X_1 + X_2 + \dots + X_n). \quad (7.99)$$

We now prove (7.99) by induction on $n \in \mathbb{N}$. For the base case $n = 1$ note that (7.98) implies that

$$\Theta_1 = \left(\begin{smallmatrix} 0 \\ 1 \end{smallmatrix}\right) \Theta_0 + X_1 = \left(\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right) (X_1). \quad (7.100)$$

This establishes (7.99) in the base case $n = 1$. For the induction step note that (7.98) demonstrates that for all $n \in \{2, 3, 4, \dots\}$ with $\Theta_{n-1} = \frac{1}{(n-1)}(X_1 + X_2 + \dots + X_{n-1})$ it holds that

$$\begin{aligned} \Theta_n &= \frac{(n-1)}{n} \Theta_{n-1} + \frac{1}{n} X_n = \left[\frac{(n-1)}{n}\right] \left[\frac{1}{(n-1)}\right] (X_1 + X_2 + \dots + X_{n-1}) + \frac{1}{n} X_n \\ &= \frac{1}{n} (X_1 + X_2 + \dots + X_{n-1}) + \frac{1}{n} X_n = \frac{1}{n} (X_1 + X_2 + \dots + X_n). \end{aligned} \quad (7.101)$$

Induction hence implies (7.99). Furthermore, note that (7.99) proves item (ii). Observe that Lemma 7.2.8, item (ii), and the fact that $(X_n)_{n \in \mathbb{N}}$ are i.i.d. random variables with $\mathbb{E}[\|X_1\|_2] < \infty$ show that for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned} \mathbb{E}[\|\Theta_n - \mathbb{E}[X_1]\|_2^2] &= \mathbb{E}[\|\frac{1}{n}(X_1 + X_2 + \dots + X_n) - \mathbb{E}[X_1]\|_2^2] \\ &= \mathbb{E}\left[\left\|\frac{1}{n}\left[\sum_{k=1}^n (X_k - \mathbb{E}[X_1])\right]\right\|_2^2\right] \\ &= \frac{1}{n^2} \left(\mathbb{E}\left[\left\|\sum_{k=1}^n (X_k - \mathbb{E}[X_k])\right\|_2^2\right] \right) \\ &= \frac{1}{n^2} \left[\sum_{k=1}^n \mathbb{E}[\|X_k - \mathbb{E}[X_k]\|_2^2] \right] \\ &= \frac{1}{n^2} \left[n \mathbb{E}[\|X_1 - \mathbb{E}[X_1]\|_2^2] \right] \\ &= \frac{\mathbb{E}[\|X_1 - \mathbb{E}[X_1]\|_2^2]}{n}. \end{aligned} \quad (7.102)$$

This establishes item (iii). It thus remains to prove item (iv). For this note that (7.96) and (7.102) ensure that for all $n \in \mathbb{N}$ it holds that

$$\begin{aligned} \mathbb{E}[\mathcal{L}(\Theta_n)] - \mathcal{L}(\mathbb{E}[X_1]) &= \mathbb{E}\left[\frac{1}{2}(\mathbb{E}[\|\mathbb{E}[X_1] - X_1\|_2^2] + \|\Theta_n - \mathbb{E}[X_1]\|_2^2)\right] \\ &\quad - \frac{1}{2}(\mathbb{E}[\|\mathbb{E}[X_1] - X_1\|_2^2] + \|\mathbb{E}[X_1] - \mathbb{E}[X_1]\|_2^2) \\ &= \frac{1}{2} \mathbb{E}[\|\Theta_n - \mathbb{E}[X_1]\|_2^2] \\ &= \frac{1}{2} \mathbb{E}[\|X_1 - \mathbb{E}[X_1]\|_2^2] n^{-1}. \end{aligned} \quad (7.103)$$

This proves item (iv). The proof for Example 7.2.16 is thus complete. □

The next result, Theorem 7.2.17 below, specifies strong and weak convergence rates for the SGD optimization method in dependence on the asymptotic behavior of the sequence of learning rates. The statement and the proof of Theorem 7.2.17 can be found in Jentzen et al. [242, Theorem 1.1].

Theorem 7.2.17 (Convergence rates in dependence of learning rates). *Let $\mathfrak{d} \in \mathbb{N}$, $\alpha, \gamma, \nu \in (0, \infty)$, $\xi \in \mathbb{R}^{\mathfrak{d}}$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $X_n: \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$, $n \in \mathbb{N}$, be i.i.d. random variables with $\mathbb{E}[\|X_1\|_2^2] < \infty$ and $\mathbb{P}(X_1 = \mathbb{E}[X_1]) < 1$, let $(r_{\varepsilon,i})_{(\varepsilon,i) \in (0,\infty) \times \{0,1\}} \subseteq \mathbb{R}$ satisfy for all $\varepsilon \in (0, \infty)$, $i \in \{0, 1\}$ that*

$$r_{\varepsilon,i} = \begin{cases} \nu/2 & : \nu < 1 \\ \min\{1/2, \gamma\alpha + (-1)^i\varepsilon\} & : \nu = 1 \\ 0 & : \nu > 1, \end{cases} \quad (7.104)$$

let $\ell = (\ell(\theta, x))_{(\theta,x) \in \mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\mathfrak{d}}} : \mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$ and $\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$ be the functions which satisfy for all $\theta, x \in \mathbb{R}^{\mathfrak{d}}$ that

$$\ell(\theta, x) = \frac{\alpha}{2}\|\theta - x\|_2^2 \quad \text{and} \quad \mathcal{L}(\theta) = \mathbb{E}[\ell(\theta, X_1)], \quad (7.105)$$

and let $\Theta: \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ be the stochastic process which satisfies for all $n \in \mathbb{N}$ that

$$\Theta_0 = \xi \quad \text{and} \quad \Theta_n = \Theta_{n-1} - \frac{\gamma}{n^{\nu}}(\nabla_{\theta}\ell)(\Theta_{n-1}, X_n). \quad (7.106)$$

Then

- (i) there exists a unique $\vartheta \in \mathbb{R}^{\mathfrak{d}}$ which satisfies that $\{\theta \in \mathbb{R}^{\mathfrak{d}} : \mathcal{L}(\theta) = \inf_{w \in \mathbb{R}^{\mathfrak{d}}} \mathcal{L}(w)\} = \{\vartheta\}$,
- (ii) for every $\varepsilon \in (0, \infty)$ there exist $c_0, c_1 \in (0, \infty)$ such that for all $n \in \mathbb{N}$ it holds that

$$c_0 n^{-r_{\varepsilon,0}} \leq (\mathbb{E}[\|\Theta_n - \vartheta\|_2^2])^{1/2} \leq c_1 n^{-r_{\varepsilon,1}}, \quad (7.107)$$

and

- (iii) for every $\varepsilon \in (0, \infty)$ there exist $c_0, c_1 \in (0, \infty)$ such that for all $n \in \mathbb{N}$ it holds that

$$c_0 n^{-2r_{\varepsilon,0}} \leq \mathbb{E}[\mathcal{L}(\Theta_n)] - \mathcal{L}(\vartheta) \leq c_1 n^{-2r_{\varepsilon,1}}. \quad (7.108)$$

Proof of Theorem 7.2.17. Note that Jentzen et al. [242, Theorem 1.1] establishes items (i), (ii), and (iii). The proof of Theorem 7.2.17 is thus complete. \square

7.2.4 Convergence rates for SGD for coercive objective functions

The statement and the proof of the next result, Theorem 7.2.18 below, can be found in Jentzen et al. [232, Theorem 1.1].

Theorem 7.2.18. Let $\mathfrak{d} \in \mathbb{N}$, $p, \alpha, \kappa, c \in (0, \infty)$, $\nu \in (0, 1)$, $q = \min(\{2, 4, 6, \dots\} \cap [p, \infty))$, $\xi, \vartheta \in \mathbb{R}^{\mathfrak{d}}$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let (S, \mathcal{S}) be a measurable space, let $X_n: \Omega \rightarrow S$, $n \in \mathbb{N}$, be i.i.d. random variables, let $\ell = (\ell(\theta, x))_{\theta \in \mathbb{R}^{\mathfrak{d}}, x \in S}: \mathbb{R}^{\mathfrak{d}} \times S \rightarrow \mathbb{R}$ be $(\mathcal{B}(\mathbb{R}^{\mathfrak{d}}) \otimes \mathcal{S})/\mathcal{B}(\mathbb{R})$ -measurable, assume for all $x \in S$ that $(\mathbb{R}^{\mathfrak{d}} \ni \theta \mapsto \ell(\theta, x) \in \mathbb{R}) \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$, assume for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that

$$\mathbb{E}[|\ell(\theta, X_1)| + \|(\nabla_{\theta}\ell)(\theta, X_1)\|_2] < \infty, \quad (7.109)$$

$$\langle \theta - \vartheta, \mathbb{E}[(\nabla_{\theta}\ell)(\theta, X_1)] \rangle \geq c \max\{\|\theta - \vartheta\|_2^2, \|\mathbb{E}[(\nabla_{\theta}\ell)(\theta, X_1)]\|_2^2\}, \quad (7.110)$$

$$\text{and } \mathbb{E}[\|(\nabla_{\theta}\ell)(\theta, X_1) - \mathbb{E}[(\nabla_{\theta}\ell)(\theta, X_1)]\|_2^q] \leq \kappa(1 + \|\theta\|_2^q), \quad (7.111)$$

let $\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that $\mathcal{L}(\theta) = \mathbb{E}[\ell(\theta, X_1)]$, and let $\Theta: \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ be the stochastic process which satisfies for all $n \in \mathbb{N}$ that

$$\Theta_0 = \xi \quad \text{and} \quad \Theta_n = \Theta_{n-1} - \frac{\alpha}{n^{\nu}} (\nabla_{\theta}\ell)(\Theta_{n-1}, X_n) \quad (7.112)$$

(cf. Definitions 1.4.7 and 3.3.4). Then

(i) it holds that $\{\theta \in \mathbb{R}^{\mathfrak{d}}: \mathcal{L}(\theta) = \inf_{w \in \mathbb{R}^{\mathfrak{d}}} \mathcal{L}(w)\} = \{\vartheta\}$ and

(ii) there exists $c \in \mathbb{R}$ such that for all $n \in \mathbb{N}$ it holds that

$$(\mathbb{E}[\|\Theta_n - \vartheta\|_2^p])^{1/p} \leq cn^{-\nu/2}. \quad (7.113)$$

Proof of Theorem 7.2.18. Observe that Jentzen et al. [232, Theorem 1.1] proves items (i) and (ii). The proof of Theorem 7.2.18 is thus complete. \square

7.2.5 Measurability of SGD processes

Lemma 7.2.19. Let $\mathfrak{d}, \mathbf{d} \in \mathbb{N}$, let $\ell = (\ell(\theta, x))_{(\theta, x) \in \mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\mathbf{d}}}: \mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\mathbf{d}} \rightarrow \mathbb{R}$ be differentiable, and let $\mathcal{g} = (\mathcal{g}_1, \dots, \mathcal{g}_{\mathfrak{d}}): \mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\mathbf{d}} \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$, $x \in \mathbb{R}^{\mathbf{d}}$ that

$$\mathcal{g}(\theta, x) = (\nabla_{\theta}\ell)(\theta, x). \quad (7.114)$$

Then \mathcal{g} is measurable.

Proof of Lemma 7.2.19. Throughout this proof, let $e_1, e_2, \dots, e_{\mathfrak{d}} \in \mathbb{R}^{\mathfrak{d}}$ satisfy

$$e_1 = (1, 0, \dots, 0), \quad e_2 = (0, 1, \dots, 0), \quad \dots, \quad e_{\mathfrak{d}} = (0, 0, \dots, 1) \quad (7.115)$$

and for every $i \in \{1, 2, \dots, \mathfrak{d}\}$, $n \in \mathbb{N}$ let $g_{i,n}: \mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\mathbf{d}} \rightarrow \mathbb{R}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$, $x \in \mathbb{R}^{\mathbf{d}}$ that

$$g_{i,n}(\theta, x) = n[\ell(\theta + \frac{1}{n}e_i, x) - \ell(\theta, x)]. \quad (7.116)$$

Note that the fact that ℓ is measurable demonstrates that for all $i \in \{1, 2, \dots, \mathfrak{d}\}$, $n \in \mathbb{N}$ it holds that $g_{i,n}$ is measurable. Furthermore, observe that the fact that ℓ is differentiable implies that for all $\theta \in \mathbb{R}^{\mathfrak{d}}$, $x \in \mathbb{R}^{\mathbf{d}}$, $i \in \{1, 2, \dots, \mathfrak{d}\}$, it holds that

$$g_i(\theta, x) = \limsup_{n \rightarrow \infty} g_{i,n}(\theta, x). \quad (7.117)$$

Combining this with the fact that for all $i \in \{1, 2, \dots, \mathfrak{d}\}$, $n \in \mathbb{N}$ it holds that $g_{i,n}$ is measurable and, for instance, [262, Theorem 1.92] shows that for all $i \in \{1, 2, \dots, \mathfrak{d}\}$ it holds that g_i is measurable. This and, for example, [262, Theorem 1.90] ensure that \mathcal{g} is measurable. The proof of Lemma 7.2.19 is thus complete. \square

Corollary 7.2.20. *Let $\mathfrak{d}, \mathbf{d} \in \mathbb{N}$, let $\ell: \mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\mathbf{d}} \rightarrow \mathbb{R}$ be differentiable, let $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, let $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\xi: \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a random variable, for every $n, j \in \mathbb{N}$ let $X_{n,j}: \Omega \rightarrow \mathbb{R}^{\mathbf{d}}$ be a random variable, and let $\Theta: \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ be the SGD process for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ (cf. Definition 7.2.1). Then Θ is a stochastic process.*

Proof of Corollary 7.2.20. Note that (7.9), Lemma 7.2.10, the fact that ξ is a random variable, and induction establish that for all $n \in \mathbb{N}_0$ it holds that Θ_n is a random variable. The proof of Corollary 7.2.20 is thus complete. \square

7.3 Explicit midpoint optimization

In this section we introduce the stochastic version of the explicit midpoint GD optimization method from Section 6.2.

Definition 7.3.1 (Explicit midpoint SGD optimization method). *Let $\mathfrak{d} \in \mathbb{N}$, let (S, \mathcal{S}) be a measurable space, let $\ell = (\ell(\theta, x))_{(\theta,x) \in \mathbb{R}^{\mathfrak{d}} \times S}: \mathbb{R}^{\mathfrak{d}} \times S \rightarrow \mathbb{R}$ be measurable, assume for all $x \in S$ that $\ell(\cdot, x)$ is differentiable, let $\mathcal{g}: \mathbb{R}^{\mathfrak{d}} \times S \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$, $x \in S$ that*

$$\mathcal{g}(\theta, x) = (\nabla_{\theta} \ell)(\theta, x), \quad (7.118)$$

let $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\xi: \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a random variable, for every $n, j \in \mathbb{N}$ let $X_{n,j}: \Omega \rightarrow S$ be a random variable, and let $\Theta: \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a function. Then we say that Θ is the explicit midpoint SGD process for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$,

and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ if and only if it holds for all $n \in \mathbb{N}$ that

$$\Theta_0 = \xi \quad \text{and} \quad \Theta_n = \Theta_{n-1} - \gamma_n \left[\frac{1}{J_n} \sum_{j=1}^{J_n} \mathcal{G} \left(\Theta_{n-1} - \frac{\gamma_n}{2} \left[\frac{1}{J_n} \sum_{j=1}^{J_n} \mathcal{G}(\Theta_{n-1}, X_{n,j}) \right], X_{n,j} \right) \right]. \quad (7.119)$$

Algorithm 7.3.2: SGD optimization method

Input: $\mathbf{d}, \mathbf{d}, N \in \mathbb{N}$, $\ell = (\ell(\theta, x))_{(\theta,x) \in \mathbb{R}^d \times \mathbb{R}^d} \in C^1(\mathbb{R}^d \times \mathbb{R}^d, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, probability space $(\Omega, \mathcal{F}, \mathbb{P})$, random variable $\xi: \Omega \rightarrow \mathbb{R}^d$, random variables $X_{n,j}: \Omega \rightarrow \mathbb{R}^d$ for $n, j \in \mathbb{N}$

Output: N -th step of the SGD process for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ (cf. Definition 7.2.1)

```

1: Initialization:  $\Theta \leftarrow \xi$ 
2: for  $n = 1, \dots, N$  do
3:    $g \leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_\theta \ell)(\Theta, X_{n,j})$ 
4:    $\Theta \leftarrow \Theta - \gamma_n g$ 
5: return  $\Theta$ 
```

An implementation of the explicit midpoint SGD optimization method in PYTORCH is given in Source code 7.3.

```

1 import torch
2 import torch.nn as nn
3 import numpy as np
4
5 net = nn.Sequential(
6     nn.Linear(1, 200), nn.ReLU(), nn.Linear(200, 1)
7 )
8
9 M = 1000
10
11 X = torch.rand((M, 1)) * 4 * np.pi - 2 * np.pi
12 Y = torch.sin(X)
13
14 J = 64
15
16 N = 150000
17
18 loss = nn.MSELoss()
19 lr = 0.003
20
21 for n in range(N):
```

```

22     indices = torch.randint(0, M, (J,))
23
24     x = X[indices]
25     y = Y[indices]
26
27     net.zero_grad()
28
29     # Remember the original parameters
30     params = [p.clone().detach() for p in net.parameters()]
31     # Compute the loss
32     loss_val = loss(net(x), y)
33     # Compute the gradients with respect to the parameters
34     loss_val.backward()
35
36     with torch.no_grad():
37         # Make a half-step in the direction of the negative
38         # gradient
39         for p in net.parameters():
40             if p.grad is not None:
41                 p.sub_(0.5 * lr * p.grad)
42
43     net.zero_grad()
44     # Compute the loss and the gradients at the midpoint
45     loss_val = loss(net(x), y)
46     loss_val.backward()
47
48     with torch.no_grad():
49         # Subtract the scaled gradient at the midpoint from the
50         # original parameters
51         for param, midpoint_param in zip(
52             params, net.parameters()
53         ):
54             param.sub_(lr * midpoint_param.grad)
55
56         # Copy the new parameters into the model
57         for param, p in zip(params, net.parameters()):
58             p.copy_(param)
59
60     if n % 1000 == 0:
61         with torch.no_grad():
62             x = torch.rand((1000, 1)) * 4 * np.pi - 2 * np.pi
63             y = torch.sin(x)
64             loss_val = loss(net(x), y)
65             print(f"Iteration: {n+1}, Loss: {loss_val}")

```

Source code 7.3 ([code/optimization_methods/midpoint_sgd.py](#)): PYTHON code implementing the explicit midpoint SGD optimization method in PyTorch

7.4 Momentum optimization

In this section we introduce the stochastic version of the momentum **GD** optimization method from Section 6.3 (cf. Polyak [357] and, for instance, [117, 261]).

Definition 7.4.1 (Momentum **SGD** optimization method). Let $\mathfrak{d} \in \mathbb{N}$, let (S, \mathcal{S}) be a measurable space, let $\ell = (\ell(\theta, x))_{(\theta, x) \in \mathbb{R}^{\mathfrak{d}} \times S} : \mathbb{R}^{\mathfrak{d}} \times S \rightarrow \mathbb{R}$ be measurable, assume for all $x \in S$ that $\ell(\cdot, x)$ is differentiable, let $\varphi : \mathbb{R}^{\mathfrak{d}} \times S \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}, x \in S$ that

$$\varphi(\theta, x) = (\nabla_{\theta}\ell)(\theta, x), \quad (7.120)$$

let $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\xi : \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a random variable, for every $n, j \in \mathbb{N}$ let $X_{n,j} : \Omega \rightarrow S$ be a random variable, and let $\Theta : \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a function. Then we say that Θ is the momentum **SGD** process for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ (we say that Θ is the momentum **SGD** process (1st version) for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$) if and only if there exists $\mathbf{m} : \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ such that for all $n \in \mathbb{N}$ it holds that

$$\Theta_0 = \xi, \quad \mathbf{m}_0 = 0, \quad (7.121)$$

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + (1 - \alpha_n) \left[\frac{1}{J_n} \sum_{j=1}^{J_n} \varphi(\Theta_{n-1}, X_{n,j}) \right], \quad (7.122)$$

$$\text{and} \quad \Theta_n = \Theta_{n-1} - \gamma_n \mathbf{m}_n. \quad (7.123)$$

Algorithm 7.4.2: Momentum **SGD** optimization method

Input: $\mathfrak{d}, \mathbf{d}, N \in \mathbb{N}$, $\ell = (\ell(\theta, x))_{(\theta, x) \in \mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\mathfrak{d}}} \in C^1(\mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\mathfrak{d}}, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, probability space $(\Omega, \mathcal{F}, \mathbb{P})$, random variable $\xi : \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$, random variables $X_{n,j} : \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ for $n, j \in \mathbb{N}$

Output: N -th step of the momentum **SGD** process for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ (cf. Definition 7.4.1)

- 1: **Initialization:** $\Theta \leftarrow \xi; \mathbf{m} \leftarrow 0 \in \mathbb{R}^{\mathfrak{d}}$
- 2: **for** $n = 1, \dots, N$ **do**
- 3: $g \leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_{\theta}\ell)(\Theta, X_{n,j})$
- 4: $\mathbf{m} \leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n) g$
- 5: $\Theta \leftarrow \Theta - \gamma_n \mathbf{m}$

6: **return** Θ

An implementation in PYTORCH of the momentum SGD optimization method as described in Definition 7.4.1 above is given in Source code 7.4. This code produces a plot which illustrates how different choices of the momentum decay rate and of the learning rate influence the progression of the loss during the training of a simple ANN with a single hidden layer, learning an approximation of the sine function. We note that while Source code 7.4 serves to illustrate a concrete implementation of the momentum SGD optimization method, for applications it is generally much preferable to use PYTORCH's built-in implementation of the momentum SGD optimization method in the `torch.optim.SGD` optimizer, rather than implementing it from scratch.

```

1 import torch
2 import torch.nn as nn
3 import numpy as np
4 import matplotlib.pyplot as plt
5
6 M = 10000
7
8 torch.manual_seed(0)
9 X = torch.rand((M, 1)) * 4 * np.pi - 2 * np.pi
10 Y = torch.sin(X)
11
12 J = 64
13
14 N = 100000
15
16 loss = nn.MSELoss()
17 lr = 0.01
18 alpha = 0.999
19
20 fig, axs = plt.subplots(1, 4, figsize=(12, 3), sharey='row')
21
22 net = nn.Sequential(
23     nn.Linear(1, 200), nn.ReLU(), nn.Linear(200, 1)
24 )
25
26 for i, alpha in enumerate([0, 0.9, 0.99, 0.999]):
27     print(f"alpha = {alpha}")
28
29     for lr in [0.1, 0.03, 0.01, 0.003]:
30         torch.manual_seed(0)
31         net.apply(
32             lambda m: m.reset_parameters()
33             if isinstance(m, nn.Linear)
34             else None
35         )

```

7.4. Momentum optimization

```
36     momentum = [
37         p.clone().detach().zero_() for p in net.parameters()
38     ]
39
40     losses = []
41     print(f"lr = {lr}")
42
43     for n in range(N):
44         indices = torch.randint(0, M, (J,))
45
46         x = X[indices]
47         y = Y[indices]
48
49         net.zero_grad()
50
51         loss_val = loss(net(x), y)
52         loss_val.backward()
53
54         with torch.no_grad():
55             for m, p in zip(momentum, net.parameters()):
56                 m.mul_(alpha)
57                 m.add_((1 - alpha) * p.grad)
58                 p.sub_(lr * m)
59
60         if n % 100 == 0:
61             with torch.no_grad():
62                 x = (torch.rand((1000, 1)) - 0.5) * 4 * np.pi
63                 y = torch.sin(x)
64                 loss_val = loss(net(x), y)
65                 losses.append(loss_val.item())
66
67         axs[i].plot(losses, label=f"\gamma = {lr}")
68
69         axs[i].set_yscale("log")
70         axs[i].set_ylim([1e-6, 1])
71         axs[i].set_title(f"\alpha = {alpha}")
72
73     axs[0].legend()
74
75     plt.tight_layout()
76     plt.savefig("../plots/sgd_momentum.pdf", bbox_inches='tight')
```

Source code 7.4 ([code/optimization_methods/momentum_sgd.py](#)): PYTHON code implementing the SGD optimization method with classical momentum in PYTORCH

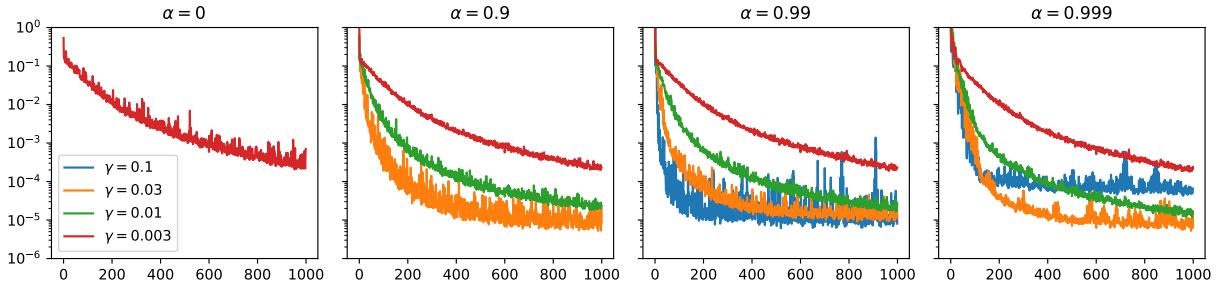


Figure 7.3 ([plots/sgd_momentum.pdf](#)): A plot showing the influence of the momentum decay rate and learning rate on the loss during the training of an ANN using the SGD optimization method with classical momentum

7.4.1 Alternative definitions

In this section we introduce the stochastic versions of the alternative momentum GD optimization methods from Section 6.3.1.

Definition 7.4.3 (Momentum SGD optimization method (2nd version)). Let $\mathfrak{d} \in \mathbb{N}$, let (S, \mathcal{S}) be a measurable space, let $\ell = (\ell(\theta, x))_{(\theta, x) \in \mathbb{R}^\mathfrak{d} \times S}: \mathbb{R}^\mathfrak{d} \times S \rightarrow \mathbb{R}$ be measurable, assume for all $x \in S$ that $\ell(\cdot, x)$ is differentiable, let $\varphi: \mathbb{R}^\mathfrak{d} \times S \rightarrow \mathbb{R}^\mathfrak{d}$ satisfy for all $\theta \in \mathbb{R}^\mathfrak{d}, x \in S$ that

$$\varphi(\theta, x) = (\nabla_\theta \ell)(\theta, x), \quad (7.124)$$

let $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\xi: \Omega \rightarrow \mathbb{R}^\mathfrak{d}$ be a random variable, for every $n, j \in \mathbb{N}$ let $X_{n,j}: \Omega \rightarrow S$ be a random variable, and let $\Theta: \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^\mathfrak{d}$ be a function. Then we say that Θ is the momentum SGD process (2nd version) for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ if and only if there exists $\mathbf{m}: \mathbb{N}_0 \rightarrow \mathbb{R}^\mathfrak{d}$ such that for all $n \in \mathbb{N}$ it holds that it holds for all $n \in \mathbb{N}$ that

$$\Theta_0 = \xi, \quad \mathbf{m}_0 = 0, \quad (7.125)$$

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + \left[\frac{1}{J_n} \sum_{j=1}^{J_n} \varphi(\Theta_{n-1}, X_{n,j}) \right], \quad (7.126)$$

$$\text{and} \quad \Theta_n = \Theta_{n-1} - \gamma_n \mathbf{m}_n. \quad (7.127)$$

Algorithm 7.4.4: Momentum SGD optimization method (2nd version)

Input: $\mathfrak{d}, \mathbf{d}, N \in \mathbb{N}$, $\ell = (\ell(\theta, x))_{(\theta, x) \in \mathbb{R}^\mathfrak{d} \times \mathbb{R}^\mathfrak{d}} \in C^1(\mathbb{R}^\mathfrak{d} \times \mathbb{R}^\mathfrak{d}, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, probability space $(\Omega, \mathcal{F}, \mathbb{P})$, random variable $\xi: \Omega \rightarrow \mathbb{R}^\mathfrak{d}$,

random variables $X_{n,j}: \Omega \rightarrow \mathbb{R}^d$ for $n, j \in \mathbb{N}$

Output: N -th step of the momentum SGD process (2nd version) for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ (cf. Definition 7.4.3)

```

1: Initialization:  $\Theta \leftarrow \xi; \mathbf{m} \leftarrow 0 \in \mathbb{R}^d$ 
2: for  $n = 1, \dots, N$  do
3:    $g \leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_\theta \ell)(\Theta, X_{n,j})$ 
4:    $\mathbf{m} \leftarrow \alpha_n \mathbf{m} + g$ 
5:    $\Theta \leftarrow \Theta - \gamma_n \mathbf{m}$ 
6: return  $\Theta$ 

```

Definition 7.4.5 (Momentum SGD optimization method (3rd version)). Let $d \in \mathbb{N}$, let (S, \mathcal{S}) be a measurable space, let $\ell = (\ell(\theta, x))_{(\theta,x) \in \mathbb{R}^d \times S}: \mathbb{R}^d \times S \rightarrow \mathbb{R}$ be measurable, assume for all $x \in S$ that $\ell(\cdot, x)$ is differentiable, let $\varphi: \mathbb{R}^d \times S \rightarrow \mathbb{R}^d$ satisfy for all $\theta \in \mathbb{R}^d$, $x \in S$ that

$$\varphi(\theta, x) = (\nabla_\theta \ell)(\theta, x), \quad (7.128)$$

let $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\xi: \Omega \rightarrow \mathbb{R}^d$ be a random variable, for every $n, j \in \mathbb{N}$ let $X_{n,j}: \Omega \rightarrow S$ be a random variable, and let $\Theta: \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^d$ be a function. Then we say that Θ is the momentum SGD process (3rd version) for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ if and only if there exists $\mathbf{m}: \mathbb{N}_0 \rightarrow \mathbb{R}^d$ such that for all $n \in \mathbb{N}$ it holds that it holds for all $n \in \mathbb{N}$ that

$$\Theta_0 = \xi, \quad \mathbf{m}_0 = 0, \quad (7.129)$$

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + (1 - \alpha_n) \gamma_n \left[\frac{1}{J_n} \sum_{j=1}^{J_n} \varphi(\Theta_{n-1}, X_{n,j}) \right], \quad (7.130)$$

$$\text{and} \quad \Theta_n = \Theta_{n-1} - \mathbf{m}_n. \quad (7.131)$$

Algorithm 7.4.6: Momentum SGD optimization method (3rd version)

Input: $d, d, N \in \mathbb{N}$, $\ell = (\ell(\theta, x))_{(\theta,x) \in \mathbb{R}^d \times \mathbb{R}^d} \in C^1(\mathbb{R}^d \times \mathbb{R}^d, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, probability space $(\Omega, \mathcal{F}, \mathbb{P})$, random variable $\xi: \Omega \rightarrow \mathbb{R}^d$, random variables $X_{n,j}: \Omega \rightarrow \mathbb{R}^d$ for $n, j \in \mathbb{N}$

Output: N -th step of the momentum SGD process (3rd version) for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, initial value ξ , batch

sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ (cf. Definition 7.4.5)

```

1: Initialization:  $\Theta \leftarrow \xi; \mathbf{m} \leftarrow 0 \in \mathbb{R}^{\mathfrak{d}}$ 
2: for  $n = 1, \dots, N$  do
3:    $g \leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_{\theta} \ell)(\Theta, X_{n,j})$ 
4:    $\mathbf{m} \leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n) \gamma_n g$ 
5:    $\Theta \leftarrow \Theta - \mathbf{m}$ 
6: return  $\Theta$ 
```

Definition 7.4.7 (Momentum SGD optimization method (4th version)). Let $\mathfrak{d} \in \mathbb{N}$, let (S, \mathcal{S}) be a measurable space, let $\ell = (\ell(\theta, x))_{(\theta,x) \in \mathbb{R}^{\mathfrak{d}} \times S}: \mathbb{R}^{\mathfrak{d}} \times S \rightarrow \mathbb{R}$ be measurable, assume for all $x \in S$ that $\ell(\cdot, x)$ is differentiable, let $\mathcal{G}: \mathbb{R}^{\mathfrak{d}} \times S \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}, x \in S$ that

$$\mathcal{G}(\theta, x) = (\nabla_{\theta} \ell)(\theta, x), \quad (7.132)$$

let $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\xi: \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a random variable, for every $n, j \in \mathbb{N}$ let $X_{n,j}: \Omega \rightarrow S$ be a random variable, and let $\Theta: \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a function. Then we say that Θ is the momentum SGD process (4th version) for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ if and only if there exists $\mathbf{m}: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ such that for all $n \in \mathbb{N}$ it holds that it holds for all $n \in \mathbb{N}$ that

$$\Theta_0 = \xi, \quad \mathbf{m}_0 = 0, \quad (7.133)$$

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + \gamma_n \left[\frac{1}{J_n} \sum_{j=1}^{J_n} \mathcal{G}(\Theta_{n-1}, X_{n,j}) \right], \quad (7.134)$$

$$\text{and} \quad \Theta_n = \Theta_{n-1} - \mathbf{m}_n. \quad (7.135)$$

Algorithm 7.4.8: Momentum SGD optimization method (4th version)

Input: $\mathfrak{d}, \mathbf{d}, N \in \mathbb{N}$, $\ell = (\ell(\theta, x))_{(\theta,x) \in \mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\mathbf{d}}} \in C^1(\mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\mathbf{d}}, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, probability space $(\Omega, \mathcal{F}, \mathbb{P})$, random variable $\xi: \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$, random variables $X_{n,j}: \Omega \rightarrow \mathbb{R}^{\mathbf{d}}$ for $n, j \in \mathbb{N}$

Output: N -th step of the momentum SGD process (4th version) for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ (cf. Definition 7.4.7)

```

1: Initialization:  $\Theta \leftarrow \xi; \mathbf{m} \leftarrow 0 \in \mathbb{R}^{\mathfrak{d}}$ 
```

```

2: for  $n = 1, \dots, N$  do
3:    $g \leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_\theta \ell)(\Theta, X_{n,j})$ 
4:    $\mathbf{m} \leftarrow \alpha_n \mathbf{m} + \gamma_n g$ 
5:    $\Theta \leftarrow \Theta - \mathbf{m}$ 
6: return  $\Theta$ 

```

7.4.2 Bias-adjusted momentum optimization

In this section we introduce the stochastic version of the bias-adjusted momentum **GD** optimization method from Section 6.3.4.

Definition 7.4.9 (Bias-adjusted momentum **SGD** optimization method). Let $\mathfrak{d} \in \mathbb{N}$, let (S, \mathcal{S}) be a measurable space, let $\ell = (\ell(\theta, x))_{(\theta, x) \in \mathbb{R}^\mathfrak{d} \times S} : \mathbb{R}^\mathfrak{d} \times S \rightarrow \mathbb{R}$ be measurable, assume for all $x \in S$ that $\ell(\cdot, x)$ is differentiable, let $\varphi : \mathbb{R}^\mathfrak{d} \times S \rightarrow \mathbb{R}^\mathfrak{d}$ satisfy for all $\theta \in \mathbb{R}^\mathfrak{d}$, $x \in S$ that

$$\varphi(\theta, x) = (\nabla_\theta \ell)(\theta, x), \quad (7.136)$$

let $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, 1)$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\xi : \Omega \rightarrow \mathbb{R}^\mathfrak{d}$ be a random variable, for every $n, j \in \mathbb{N}$ let $X_{n,j} : \Omega \rightarrow S$ be a random variable, and let $\Theta : \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^\mathfrak{d}$ be a function. Then we say that Θ is the bias-adjusted momentum **SGD** process for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ if and only if there exists $\mathbf{m} : \mathbb{N}_0 \rightarrow \mathbb{R}^\mathfrak{d}$ such that for all $n \in \mathbb{N}$ it holds that

$$\Theta_0 = \xi, \quad \mathbf{m}_0 = 0, \quad (7.137)$$

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + (1 - \alpha_n) \left[\frac{1}{J_n} \sum_{j=1}^{J_n} \varphi(\Theta_{n-1}, X_{n,j}) \right], \quad (7.138)$$

$$\text{and} \quad \Theta_n = \Theta_{n-1} - \frac{\gamma_n \mathbf{m}_n}{1 - \prod_{l=1}^n \alpha_l}. \quad (7.139)$$

Algorithm 7.4.10: Bias-adjusted momentum **SGD** optimization method

Input: $\mathfrak{d}, \mathbf{d}, N \in \mathbb{N}$, $\ell = (\ell(\theta, x))_{(\theta, x) \in \mathbb{R}^\mathfrak{d} \times \mathbb{R}^\mathfrak{d}} \in C^1(\mathbb{R}^\mathfrak{d} \times \mathbb{R}^\mathfrak{d}, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, 1)$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, probability space $(\Omega, \mathcal{F}, \mathbb{P})$, random variable $\xi : \Omega \rightarrow \mathbb{R}^\mathfrak{d}$, random variables $X_{n,j} : \Omega \rightarrow \mathbb{R}^\mathfrak{d}$ for $n, j \in \mathbb{N}$

Output: N -th step of the bias-adjusted momentum **SGD** process for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ (cf. Definition 7.4.9)

```

1: Initialization:  $\Theta \leftarrow \xi$ ;  $\mathbf{m} \leftarrow 0 \in \mathbb{R}^d$ 
2: for  $n = 1, \dots, N$  do
3:    $g \leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_{\theta} \ell)(\Theta, X_{n,j})$ 
4:    $\mathbf{m} \leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n)g$ 
5:    $\Theta \leftarrow \Theta - \frac{\gamma_n \mathbf{m}}{1 - \prod_{l=1}^n \alpha_l}$ 
6: return  $\Theta$ 

```

An implementation of the bias-adjusted momentum SGD optimization method in PYTORCH is given in Source code 7.5.

```

1 import torch
2 import torch.nn as nn
3 import numpy as np
4
5 net = nn.Sequential(
6     nn.Linear(1, 200), nn.ReLU(), nn.Linear(200, 1)
7 )
8
9 M = 1000
10
11 X = torch.rand((M, 1)) * 4 * np.pi - 2 * np.pi
12 Y = torch.sin(X)
13
14 J = 64
15
16 N = 150000
17
18 loss = nn.MSELoss()
19 lr = 0.01
20 alpha = 0.99
21 adj = 1
22
23 momentum = [p.clone().detach().zero_() for p in net.parameters()]
24
25 for n in range(N):
26     indices = torch.randint(0, M, (J,))
27
28     x = X[indices]
29     y = Y[indices]
30
31     net.zero_grad()
32
33     loss_val = loss(net(x), y)
34     loss_val.backward()
35

```

```

36     adj *= alpha
37
38     with torch.no_grad():
39         for m, p in zip(momentum, net.parameters()):
40             m.mul_(alpha)
41             m.add_((1-alpha) * p.grad)
42             p.sub_(lr * m / (1 - adj))
43
44     if n % 1000 == 0:
45         with torch.no_grad():
46             x = torch.rand((1000, 1)) * 4 * np.pi - 2 * np.pi
47             y = torch.sin(x)
48             loss_val = loss(net(x), y)
49             print(f"Iteration: {n+1}, Loss: {loss_val}")

```

Source code 7.5 ([code/optimization_methods/momentum_sgd_bias_adj.py](#)):
 PYTHON code implementing the bias-adjusted momentum SGD optimization method
 in PYTORCH

7.5 Nesterov accelerated momentum optimization

In this section we introduce the stochastic version of the Nesterov accelerated GD optmization method from Section 6.4 (cf. [321, 408]).

Definition 7.5.1 (Nesterov accelerated SGD optimization method). *Let $\mathfrak{d} \in \mathbb{N}$, let (S, \mathcal{S}) be a measurable space, let $\ell = (\ell(\theta, x))_{(\theta, x) \in \mathbb{R}^{\mathfrak{d}} \times S}: \mathbb{R}^{\mathfrak{d}} \times S \rightarrow \mathbb{R}$ be measurable, assume for all $x \in S$ that $\ell(\cdot, x)$ is differentiable, let $\varphi: \mathbb{R}^{\mathfrak{d}} \times S \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}, x \in S$ that*

$$\varphi(\theta, x) = (\nabla_{\theta} \ell)(\theta, x), \quad (7.140)$$

let $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\xi: \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a random variable, for every $n, j \in \mathbb{N}$ let $X_{n,j}: \Omega \rightarrow S$ be a random variable, and let $\Theta: \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a function. Then we say that Θ is the Nesterov accelerated SGD process for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ (we say that Θ is the Nesterov accelerated SGD process (1st version) for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$) if and only if there exists $\mathbf{m}: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ such that for all $n \in \mathbb{N}$ it holds that

$$\Theta_0 = \xi, \quad \mathbf{m}_0 = 0, \quad (7.141)$$

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + (1 - \alpha_n) \left[\frac{1}{J_n} \sum_{j=1}^{J_n} \mathcal{g}(\Theta_{n-1} - \gamma_n \alpha_n \mathbf{m}_{n-1}, X_{n,j}) \right], \quad (7.142)$$

$$and \quad \Theta_n = \Theta_{n-1} - \gamma_n \mathbf{m}_n. \quad (7.143)$$

Algorithm 7.5.2: Nesterov accelerated SGD optimization method

Input: $\mathbf{d}, \mathbf{d}, N \in \mathbb{N}$, $\ell = (\ell(\theta, x))_{(\theta, x) \in \mathbb{R}^{\mathbf{d}} \times \mathbb{R}^{\mathbf{d}}} \in C^1(\mathbb{R}^{\mathbf{d}} \times \mathbb{R}^{\mathbf{d}}, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, probability space $(\Omega, \mathcal{F}, \mathbb{P})$, random variable $\xi: \Omega \rightarrow \mathbb{R}^{\mathbf{d}}$, random variables $X_{n,j}: \Omega \rightarrow \mathbb{R}^{\mathbf{d}}$ for $n, j \in \mathbb{N}$

Output: N -th step of the Nesterov accelerated SGD process for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ (cf. Definition 7.5.1)

```

1: Initialization:  $\Theta \leftarrow \xi$ ;  $\mathbf{m} \leftarrow 0 \in \mathbb{R}^{\mathbf{d}}$ 
2: for  $n = 1, \dots, N$  do
3:    $g \leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_{\theta} \ell)(\Theta - \gamma_n \alpha_n \mathbf{m}, X_{n,j})$ 
4:    $\mathbf{m} \leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n) g$ 
5:    $\Theta \leftarrow \Theta - \gamma_n \mathbf{m}$ 
6: return  $\Theta$ 
```

An implementation of the Nesterov accelerated SGD optimization method in PYTORCH is given in Source code 7.6.

```

1 import torch
2 import torch.nn as nn
3 import numpy as np
4
5 net = nn.Sequential(
6     nn.Linear(1, 200), nn.ReLU(), nn.Linear(200, 1)
7 )
8
9 M = 1000
10
11 X = torch.rand((M, 1)) * 4 * np.pi - 2 * np.pi
12 Y = torch.sin(X)
13
14 J = 64
15
16 N = 150000
17
18 loss = nn.MSELoss()
19 lr = 0.003
20 alpha = 0.999
```

```

21 m = [p.clone().detach().zero_() for p in net.parameters()]
22
23 for n in range(N):
24     indices = torch.randint(0, M, (J,))
25
26     x = X[indices]
27     y = Y[indices]
28
29     net.zero_grad()
30
31     # Remember the original parameters
32     params = [p.clone().detach() for p in net.parameters()]
33
34     for p, m_p in zip(params, m):
35         p.sub_(lr * alpha * m_p)
36
37     # Compute the loss
38     loss_val = loss(net(x), y)
39     # Compute the gradients with respect to the parameters
40     loss_val.backward()
41
42     with torch.no_grad():
43         for p, m_p, q in zip(net.parameters(), m, params):
44             m_p.mul_(alpha)
45             m_p.add_((1 - alpha) * p.grad)
46             q.sub_(lr * m_p)
47             p.copy_(q)
48
49     if n % 1000 == 0:
50         with torch.no_grad():
51             x = torch.rand((1000, 1)) * 4 * np.pi - 2 * np.pi
52             y = torch.sin(x)
53             loss_val = loss(net(x), y)
54             print(f"Iteration: {n+1}, Loss: {loss_val}")
55

```

Source code 7.6 ([code/optimization_methods/nesterov_sgd.py](#)): PYTHON code implementing the Nesterov accelerated SGD optimization method in PYTORCH

7.5.1 Alternative definitions

In this section we introduce the stochastic versions of the alternative Nesterov accelerated GD optimization methods from Section 6.4.1.

Definition 7.5.3 (Nesterov accelerated SGD optimization method (2nd version)). *Let $\mathfrak{d} \in \mathbb{N}$, let (S, \mathcal{S}) be a measurable space, let $\ell = (\ell(\theta, x))_{(\theta, x) \in \mathbb{R}^{\mathfrak{d}} \times S}: \mathbb{R}^{\mathfrak{d}} \times S \rightarrow \mathbb{R}$ be*

measurable, assume for all $x \in S$ that $\ell(\cdot, x)$ is differentiable, let $\varphi: \mathbb{R}^d \times S \rightarrow \mathbb{R}^d$ satisfy for all $\theta \in \mathbb{R}^d$, $x \in S$ that

$$\varphi(\theta, x) = (\nabla_\theta \ell)(\theta, x), \quad (7.144)$$

let $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, \infty]$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\xi: \Omega \rightarrow \mathbb{R}^d$ be a random variable, for every $n, j \in \mathbb{N}$ let $X_{n,j}: \Omega \rightarrow S$ be a random variable, and let $\Theta: \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^d$ be a function. Then we say that Θ is the Nesterov accelerated SGD process (2nd version) for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ if and only if there exists $\mathbf{m}: \mathbb{N}_0 \rightarrow \mathbb{R}^d$ such that for all $n \in \mathbb{N}$ it holds that

$$\Theta_0 = \xi, \quad \mathbf{m}_0 = 0, \quad (7.145)$$

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + \left[\frac{1}{J_n} \sum_{j=1}^{J_n} \varphi(\Theta_{n-1} - \gamma_n \alpha_n \mathbf{m}_{n-1}, X_{n,j}) \right], \quad (7.146)$$

$$\text{and} \quad \Theta_n = \Theta_{n-1} - \gamma_n \mathbf{m}_n. \quad (7.147)$$

Algorithm 7.5.4: Nesterov accelerated SGD optimization method (2nd version)

Input: $d, d, N \in \mathbb{N}$, $\ell = (\ell(\theta, x))_{(\theta,x) \in \mathbb{R}^d \times \mathbb{R}^d} \in C^1(\mathbb{R}^d \times \mathbb{R}^d, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, probability space $(\Omega, \mathcal{F}, \mathbb{P})$, random variable $\xi: \Omega \rightarrow \mathbb{R}^d$, random variables $X_{n,j}: \Omega \rightarrow \mathbb{R}^d$ for $n, j \in \mathbb{N}$

Output: N -th step of the Nesterov accelerated SGD process (2nd version) for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ (cf. Definition 7.5.3)

```

1: Initialization:  $\Theta \leftarrow \xi; \mathbf{m} \leftarrow 0 \in \mathbb{R}^d$ 
2: for  $n = 1, \dots, N$  do
3:    $g \leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_\theta \ell)(\Theta - \gamma_n \alpha_n \mathbf{m}, X_{n,j})$ 
4:    $\mathbf{m} \leftarrow \alpha_n \mathbf{m} + g$ 
5:    $\Theta \leftarrow \Theta - \gamma_n \mathbf{m}$ 
6: return  $\Theta$ 

```

Definition 7.5.5 (Nesterov accelerated SGD optimization method (3rd version)). Let $d \in \mathbb{N}$, let (S, \mathcal{S}) be a measurable space, let $\ell = (\ell(\theta, x))_{(\theta,x) \in \mathbb{R}^d \times S}: \mathbb{R}^d \times S \rightarrow \mathbb{R}$ be measurable, assume for all $x \in S$ that $\ell(\cdot, x)$ is differentiable, let $\varphi: \mathbb{R}^d \times S \rightarrow \mathbb{R}^d$

satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$, $x \in S$ that

$$\mathcal{g}(\theta, x) = (\nabla_{\theta}\ell)(\theta, x), \quad (7.148)$$

let $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\xi: \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a random variable, for every $n, j \in \mathbb{N}$ let $X_{n,j}: \Omega \rightarrow S$ be a random variable, and let $\Theta: \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a function. Then we say that Θ is the Nesterov accelerated SGD process (3rd version) for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ if and only if there exists $\mathbf{m}: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ such that for all $n \in \mathbb{N}$ it holds that

$$\Theta_0 = \xi, \quad \mathbf{m}_0 = 0, \quad (7.149)$$

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + (1 - \alpha_n) \gamma_n \left[\frac{1}{J_n} \sum_{j=1}^{J_n} \mathcal{g}(\Theta_{n-1} - \alpha_n \mathbf{m}_{n-1}, X_{n,j}) \right], \quad (7.150)$$

$$\text{and} \quad \Theta_n = \Theta_{n-1} - \mathbf{m}_n. \quad (7.151)$$

Algorithm 7.5.6: Nesterov accelerated SGD optimization method (3rd version)

Input: $\mathfrak{d}, \mathbf{d}, N \in \mathbb{N}$, $\ell = (\ell(\theta, x))_{(\theta,x) \in \mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\mathbf{d}}} \in C^1(\mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\mathbf{d}}, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, probability space $(\Omega, \mathcal{F}, \mathbb{P})$, random variable $\xi: \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$, random variables $X_{n,j}: \Omega \rightarrow \mathbb{R}^{\mathbf{d}}$ for $n, j \in \mathbb{N}$

Output: N -th step of the Nesterov accelerated SGD process (3rd version) for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ (cf. Definition 7.5.5)

```

1: Initialization:  $\Theta \leftarrow \xi; \mathbf{m} \leftarrow 0 \in \mathbb{R}^{\mathfrak{d}}$ 
2: for  $n = 1, \dots, N$  do
3:    $g \leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_{\theta}\ell)(\Theta - \alpha_n \mathbf{m}, X_{n,j})$ 
4:    $\mathbf{m} \leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n) \gamma_n g$ 
5:    $\Theta \leftarrow \Theta - \mathbf{m}$ 
6: return  $\Theta$ 

```

Definition 7.5.7 (Nesterov accelerated SGD optimization method (4th version)). Let $\mathfrak{d} \in \mathbb{N}$, let (S, \mathcal{S}) be a measurable space, let $\ell = (\ell(\theta, x))_{(\theta,x) \in \mathbb{R}^{\mathfrak{d}} \times S}: \mathbb{R}^{\mathfrak{d}} \times S \rightarrow \mathbb{R}$ be measurable, assume for all $x \in S$ that $\ell(\cdot, x)$ is differentiable, let $\mathcal{g}: \mathbb{R}^{\mathfrak{d}} \times S \rightarrow \mathbb{R}^{\mathfrak{d}}$

satisfy for all $\theta \in \mathbb{R}^d$, $x \in S$ that

$$g(\theta, x) = (\nabla_\theta \ell)(\theta, x), \quad (7.152)$$

let $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, \infty]$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\xi: \Omega \rightarrow \mathbb{R}^d$ be a random variable, for every $n, j \in \mathbb{N}$ let $X_{n,j}: \Omega \rightarrow S$ be a random variable, and let $\Theta: \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^d$ be a function. Then we say that Θ is the Nesterov accelerated SGD process (4th version) for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ if and only if there exists $\mathbf{m}: \mathbb{N}_0 \rightarrow \mathbb{R}^d$ such that for all $n \in \mathbb{N}$ it holds that

$$\Theta_0 = \xi, \quad \mathbf{m}_0 = 0, \quad (7.153)$$

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + \gamma_n \left[\frac{1}{J_n} \sum_{j=1}^{J_n} g(\Theta_{n-1} - \alpha_n \mathbf{m}_{n-1}, X_{n,j}) \right], \quad (7.154)$$

$$\text{and} \quad \Theta_n = \Theta_{n-1} - \mathbf{m}_n. \quad (7.155)$$

Algorithm 7.5.8: Nesterov accelerated SGD optimization method (4th version)

Input: $d, d, N \in \mathbb{N}$, $\ell = (\ell(\theta, x))_{(\theta,x) \in \mathbb{R}^d \times \mathbb{R}^d} \in C^1(\mathbb{R}^d \times \mathbb{R}^d, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, probability space $(\Omega, \mathcal{F}, \mathbb{P})$, random variable $\xi: \Omega \rightarrow \mathbb{R}^d$, random variables $X_{n,j}: \Omega \rightarrow \mathbb{R}^d$ for $n, j \in \mathbb{N}$

Output: N -th step of the Nesterov accelerated SGD process (4th version) for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ (cf. Definition 7.5.7)

```

1: Initialization:  $\Theta \leftarrow \xi$ ;  $\mathbf{m} \leftarrow 0 \in \mathbb{R}^d$ 
2: for  $n = 1, \dots, N$  do
3:    $g \leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_\theta \ell)(\Theta - \alpha_n \mathbf{m}, X_{n,j})$ 
4:    $\mathbf{m} \leftarrow \alpha_n \mathbf{m} + \gamma_n g$ 
5:    $\Theta \leftarrow \Theta - \mathbf{m}$ 
6: return  $\Theta$ 

```

7.5.2 Bias-adjusted Nesterov accelerated momentum optimization

In this section we introduce the stochastic version of the bias-adjusted Nesterov accelerated GD optimization method from Section 6.4.3.

Definition 7.5.9 (Bias-adjusted Nesterov accelerated SGD optimization method). Let $\mathfrak{d} \in \mathbb{N}$, let (S, \mathcal{S}) be a measurable space, let $\ell = (\ell(\theta, x))_{(\theta, x) \in \mathbb{R}^{\mathfrak{d}} \times S}: \mathbb{R}^{\mathfrak{d}} \times S \rightarrow \mathbb{R}$ be measurable, assume for all $x \in S$ that $\ell(\cdot, x)$ is differentiable, let $\varphi: \mathbb{R}^{\mathfrak{d}} \times S \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}, x \in S$ that

$$\varphi(\theta, x) = (\nabla_{\theta}\ell)(\theta, x), \quad (7.156)$$

let $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, 1)$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\xi: \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a random variable, for every $n, j \in \mathbb{N}$ let $X_{n,j}: \Omega \rightarrow S$ be a random variable, and let $\Theta: \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a function. Then we say that Θ is the bias-adjusted Nesterov accelerated SGD process for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ if and only if there exists $\mathbf{m}: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ such that for all $n \in \mathbb{N}$ it holds that

$$\Theta_0 = \xi, \quad \mathbf{m}_0 = 0, \quad (7.157)$$

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + (1 - \alpha_n) \left[\frac{1}{J_n} \sum_{j=1}^{J_n} \varphi \left(\Theta_{n-1} - \frac{\gamma_n \alpha_n \mathbf{m}_{n-1}}{1 - \prod_{l=1}^n \alpha_l}, X_{n,j} \right), \right] \quad (7.158)$$

$$\text{and} \quad \Theta_n = \Theta_{n-1} - \frac{\gamma_n \mathbf{m}_n}{1 - \prod_{l=1}^n \alpha_l}. \quad (7.159)$$

Algorithm 7.5.10: Bias-adjusted Nesterov accelerated SGD optimization method

Input: $\mathfrak{d}, \mathbf{d}, N \in \mathbb{N}$, $\ell = (\ell(\theta, x))_{(\theta, x) \in \mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\mathfrak{d}}} \in C^1(\mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\mathfrak{d}}, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, 1)$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, probability space $(\Omega, \mathcal{F}, \mathbb{P})$, random variable $\xi: \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$, random variables $X_{n,j}: \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ for $n, j \in \mathbb{N}$

Output: N -th step of the bias-adjusted Nesterov accelerated SGD process for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ (cf. Definition 7.5.9)

- 1: **Initialization:** $\Theta \leftarrow \xi; \mathbf{m} \leftarrow 0 \in \mathbb{R}^{\mathfrak{d}}$
- 2: **for** $n = 1, \dots, N$ **do**
- 3: $g \leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_{\theta}\ell)(\Theta - \frac{\gamma_n \alpha_n \mathbf{m}}{1 - \prod_{l=1}^n \alpha_l}, X_{n,j})$
- 4: $\mathbf{m} \leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n)g$
- 5: $\Theta \leftarrow \Theta - \frac{\gamma_n \mathbf{m}}{1 - \prod_{l=1}^n \alpha_l}$
- 6: **return** Θ

7.5.3 Shifted representations

In this section we introduce the stochastic versions of the shifted representations of the Nesterov accelerated GD optimization methods from Section 6.4.4.

7.5.3.1 Shifted representation for the first version of Nesterov accelerated momentum optimization

Definition 7.5.11 (Shifted Nesterov accelerated SGD optimization method). Let $\mathfrak{d} \in \mathbb{N}$, let (S, \mathcal{S}) be a measurable space, let $\ell = (\ell(\theta, x))_{(\theta, x) \in \mathbb{R}^{\mathfrak{d}} \times S}: \mathbb{R}^{\mathfrak{d}} \times S \rightarrow \mathbb{R}$ be measurable, assume for all $x \in S$ that $\ell(\cdot, x)$ is differentiable, let $\varrho: \mathbb{R}^{\mathfrak{d}} \times S \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}, x \in S$ that

$$\varrho(\theta, x) = (\nabla_{\theta}\ell)(\theta, x), \quad (7.160)$$

let $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\xi: \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a random variable, for every $n, j \in \mathbb{N}$ let $X_{n,j}: \Omega \rightarrow S$ be a random variable, and let $\Theta: \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a function. Then we say that Θ is the shifted Nesterov accelerated SGD process for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ if and only if there exists $\mathbf{m}: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ such that for all $n \in \mathbb{N}$ it holds that

$$\Theta_0 = \xi, \quad \mathbf{m}_0 = 0, \quad (7.161)$$

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + (1 - \alpha_n) \left[\frac{1}{J_n} \sum_{j=1}^{J_n} \varrho(\Theta_{n-1}, X_{n,j}) \right] \quad \text{and} \quad (7.162)$$

$$\Theta_n = \Theta_{n-1} - \gamma_{n+1} \alpha_{n+1} \mathbf{m}_n - \gamma_n (1 - \alpha_n) \left[\frac{1}{J_n} \sum_{j=1}^{J_n} \varrho(\Theta_{n-1}, X_{n,j}) \right]. \quad (7.163)$$

Algorithm 7.5.12: Shifted Nesterov accelerated SGD optimization method

Input: $\mathfrak{d}, \mathbf{d}, N \in \mathbb{N}$, $\ell = (\ell(\theta, x))_{(\theta, x) \in \mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\mathfrak{d}}} \in C^1(\mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\mathfrak{d}}, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, probability space $(\Omega, \mathcal{F}, \mathbb{P})$, random variable $\xi: \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$, random variables $X_{n,j}: \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ for $n, j \in \mathbb{N}$

Output: N -th step of the shifted Nesterov accelerated SGD process for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ (cf. Definition 7.5.11)

- 1: **Initialization:** $\Theta \leftarrow \xi; \mathbf{m} \leftarrow 0 \in \mathbb{R}^{\mathfrak{d}}$
- 2: **for** $n = 1, \dots, N$ **do**

```

3:    $g \leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_\theta \ell)(\Theta, X_{n,j})$ 
4:    $\mathbf{m} \leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n) g$ 
5:    $\Theta \leftarrow \Theta - \gamma_{n+1} \alpha_{n+1} \mathbf{m} - \gamma_n (1 - \alpha_n) g$ 
6: return  $\Theta$ 

```

7.5.3.2 Shifted representation for the second version of Nesterov accelerated momentum optimization

Definition 7.5.13 (Shifted Nesterov accelerated SGD optimization method (2nd version)). Let $\mathfrak{d} \in \mathbb{N}$, let (S, \mathcal{S}) be a measurable space, let $\ell = (\ell(\theta, x))_{(\theta, x) \in \mathbb{R}^\mathfrak{d} \times S}: \mathbb{R}^\mathfrak{d} \times S \rightarrow \mathbb{R}$ be measurable, assume for all $x \in S$ that $\ell(\cdot, x)$ is differentiable, let $\varphi: \mathbb{R}^\mathfrak{d} \times S \rightarrow \mathbb{R}^\mathfrak{d}$ satisfy for all $\theta \in \mathbb{R}^\mathfrak{d}$, $x \in S$ that

$$\varphi(\theta, x) = (\nabla_\theta \ell)(\theta, x), \quad (7.164)$$

let $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\xi: \Omega \rightarrow \mathbb{R}^\mathfrak{d}$ be a random variable, for every $n, j \in \mathbb{N}$ let $X_{n,j}: \Omega \rightarrow S$ be a random variable, and let $\Theta: \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^\mathfrak{d}$ be a function. Then we say that Θ is the shifted Nesterov accelerated SGD process (2nd version) for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ if and only if there exists $\mathbf{m}: \mathbb{N}_0 \rightarrow \mathbb{R}^\mathfrak{d}$ such that for all $n \in \mathbb{N}$ it holds that

$$\Theta_0 = \xi, \quad \mathbf{m}_0 = 0, \quad (7.165)$$

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + \frac{1}{J_n} \sum_{j=1}^{J_n} \varphi(\Theta_{n-1}, X_{n,j}), \quad \text{and} \quad (7.166)$$

$$\Theta_n = \Theta_{n-1} - \gamma_{n+1} \alpha_{n+1} \mathbf{m}_n - \gamma_n \left[\frac{1}{J_n} \sum_{j=1}^{J_n} \varphi(\Theta_{n-1}, X_{n,j}) \right]. \quad (7.167)$$

Algorithm 7.5.14: Shifted Nesterov accelerated SGD optimization method (2nd version)

Input: $\mathfrak{d}, \mathbf{d}, N \in \mathbb{N}$, $\ell = (\ell(\theta, x))_{(\theta, x) \in \mathbb{R}^\mathfrak{d} \times \mathbb{R}^\mathbf{d}} \in C^1(\mathbb{R}^\mathfrak{d} \times \mathbb{R}^\mathbf{d}, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, probability space $(\Omega, \mathcal{F}, \mathbb{P})$, random variable $\xi: \Omega \rightarrow \mathbb{R}^\mathfrak{d}$, random variables $X_{n,j}: \Omega \rightarrow \mathbb{R}^\mathbf{d}$ for $n, j \in \mathbb{N}$

Output: N -th step of the shifted Nesterov accelerated SGD process (2nd version) for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, initial

value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ (cf. Definition 7.5.13)

```

1: Initialization:  $\Theta \leftarrow \xi; \mathbf{m} \leftarrow 0 \in \mathbb{R}^{\mathfrak{d}}$ 
2: for  $n = 1, \dots, N$  do
3:    $g \leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_{\theta} \ell)(\Theta, X_{n,j})$ 
4:    $\mathbf{m} \leftarrow \alpha_n \mathbf{m} + g$ 
5:    $\Theta \leftarrow \Theta - \gamma_{n+1} \alpha_{n+1} \mathbf{m} - \gamma_n g$ 
6: return  $\Theta$ 
```

7.5.3.3 Shifted representation for the third version of Nesterov accelerated momentum optimization

Definition 7.5.15 (Shifted Nesterov accelerated SGD optimization method (3rd version)). Let $\mathfrak{d} \in \mathbb{N}$, let (S, \mathcal{S}) be a measurable space, let $\ell = (\ell(\theta, x))_{(\theta, x) \in \mathbb{R}^{\mathfrak{d}} \times S}: \mathbb{R}^{\mathfrak{d}} \times S \rightarrow \mathbb{R}$ be measurable, assume for all $x \in S$ that $\ell(\cdot, x)$ is differentiable, let $\mathcal{G}: \mathbb{R}^{\mathfrak{d}} \times S \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}, x \in S$ that

$$\mathcal{G}(\theta, x) = (\nabla_{\theta} \ell)(\theta, x), \quad (7.168)$$

let $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\xi: \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a random variable, for every $n, j \in \mathbb{N}$ let $X_{n,j}: \Omega \rightarrow S$ be a random variable, and let $\Theta: \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a function. Then we say that Θ is the shifted Nesterov accelerated SGD process (3rd version) for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ if and only if there exists $\mathbf{m}: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ such that for all $n \in \mathbb{N}$ it holds that

$$\Theta_0 = \xi, \quad \mathbf{m}_0 = 0, \quad (7.169)$$

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + (1 - \alpha_n) \gamma_n \left[\frac{1}{J_n} \sum_{j=1}^{J_n} \mathcal{G}(\Theta_{n-1}, X_{n,j}) \right], \quad (7.170)$$

$$\text{and} \quad \Theta_n = \Theta_{n-1} - \alpha_{n+1} \mathbf{m}_n - (1 - \alpha_n) \gamma_n \left[\frac{1}{J_n} \sum_{j=1}^{J_n} \mathcal{G}(\Theta_{n-1}, X_{n,j}) \right]. \quad (7.171)$$

Algorithm 7.5.16: Shifted Nesterov accelerated SGD optimization method (3rd version)

Input: $\mathfrak{d}, \mathbf{d}, N \in \mathbb{N}$, $\ell = (\ell(\theta, x))_{(\theta, x) \in \mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\mathbf{d}}} \in C^1(\mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\mathbf{d}}, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, probability space $(\Omega, \mathcal{F}, \mathbb{P})$, random variable $\xi: \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$, random variables $X_{n,j}: \Omega \rightarrow \mathbb{R}^{\mathbf{d}}$ for $n, j \in \mathbb{N}$

Output: N -th step of the shifted Nesterov accelerated SGD process (3rd version) for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ (cf. Definition 7.5.15)

```

1: Initialization:  $\Theta \leftarrow \xi; \mathbf{m} \leftarrow 0 \in \mathbb{R}^{\mathfrak{d}}$ 
2: for  $n = 1, \dots, N$  do
3:    $g \leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_{\theta} \ell)(\Theta, X_{n,j})$ 
4:    $\mathbf{m} \leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n) \gamma_n g$ 
5:    $\Theta \leftarrow \Theta - \alpha_{n+1} \mathbf{m} - (1 - \alpha_n) \gamma_n g$ 
6: return  $\Theta$ 

```

7.5.3.4 Shifted representation for the fourth version of Nesterov accelerated momentum optimization

Definition 7.5.17 (Shifted Nesterov accelerated SGD optimization method (4th version)). Let $\mathfrak{d} \in \mathbb{N}$, let (S, \mathcal{S}) be a measurable space, let $\ell = (\ell(\theta, x))_{(\theta, x) \in \mathbb{R}^{\mathfrak{d}} \times S}: \mathbb{R}^{\mathfrak{d}} \times S \rightarrow \mathbb{R}$ be measurable, assume for all $x \in S$ that $\ell(\cdot, x)$ is differentiable, let $\varrho: \mathbb{R}^{\mathfrak{d}} \times S \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$, $x \in S$ that

$$\varrho(\theta, x) = (\nabla_{\theta} \ell)(\theta, x), \quad (7.172)$$

let $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\xi: \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a random variable, for every $n, j \in \mathbb{N}$ let $X_{n,j}: \Omega \rightarrow S$ be a random variable, and let $\Theta: \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a function. Then we say that Θ is the shifted Nesterov accelerated SGD process (4th version) for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ if and only if there exists $\mathbf{m}: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ such that for all $n \in \mathbb{N}$ it holds that

$$\Theta_0 = \xi, \quad \mathbf{m}_0 = 0, \quad (7.173)$$

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + \gamma_n \left[\frac{1}{J_n} \sum_{j=1}^{J_n} \varrho(\Theta_{n-1}, X_{n,j}) \right], \quad (7.174)$$

$$\text{and} \quad \Theta_n = \Theta_{n-1} - \alpha_{n+1} \mathbf{m}_n - \gamma_n \left[\frac{1}{J_n} \sum_{j=1}^{J_n} \mathcal{g}(\Theta_{n-1}, X_{n,j}) \right]. \quad (7.175)$$

Algorithm 7.5.18: Shifted Nesterov accelerated SGD optimization method (4th version)

Input: $\mathfrak{d}, \mathbf{d}, N \in \mathbb{N}$, $\ell = (\ell(\theta, x))_{(\theta, x) \in \mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\mathbf{d}}} \in C^1(\mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\mathbf{d}}, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, probability space $(\Omega, \mathcal{F}, \mathbb{P})$, random variable $\xi: \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$, random variables $X_{n,j}: \Omega \rightarrow \mathbb{R}^{\mathbf{d}}$ for $n, j \in \mathbb{N}$

Output: N -th step of the shifted Nesterov accelerated SGD process (4th version) for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ (cf. Definition 7.5.17)

```

1: Initialization:  $\Theta \leftarrow \xi; \mathbf{m} \leftarrow 0 \in \mathbb{R}^{\mathfrak{d}}$ 
2: for  $n = 1, \dots, N$  do
3:    $g \leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_{\theta} \ell)(\Theta, X_{n,j})$ 
4:    $\mathbf{m} \leftarrow \alpha_n \mathbf{m} + \gamma_n g$ 
5:    $\Theta \leftarrow \Theta - \alpha_{n+1} \mathbf{m} - \gamma_n g$ 
6: return  $\Theta$ 

```

7.5.3.5 Shifted representation for the bias-adjusted Nesterov accelerated momentum optimization

Definition 7.5.19 (Shifted bias-adjusted Nesterov accelerated SGD optimization method). Let $\mathfrak{d} \in \mathbb{N}$, let (S, \mathcal{S}) be a measurable space, let $\ell = (\ell(\theta, x))_{(\theta, x) \in \mathbb{R}^{\mathfrak{d}} \times S}: \mathbb{R}^{\mathfrak{d}} \times S \rightarrow \mathbb{R}$ be measurable, assume for all $x \in S$ that $\ell(\cdot, x)$ is differentiable, let $\mathcal{g}: \mathbb{R}^{\mathfrak{d}} \times S \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$, $x \in S$ that

$$\mathcal{g}(\theta, x) = (\nabla_{\theta} \ell)(\theta, x), \quad (7.176)$$

let $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, 1)$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\xi: \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a random variable, for every $n, j \in \mathbb{N}$ let $X_{n,j}: \Omega \rightarrow S$ be a random variable, and let $\Theta: \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a function. Then we say that Θ is the shifted bias-adjusted Nesterov accelerated SGD process for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ if and only if there exists $\mathbf{m}: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ such that for all $n \in \mathbb{N}$ it holds that

$$\Theta_0 = \xi, \quad \mathbf{m}_0 = 0, \quad (7.177)$$

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + (1 - \alpha_n) \left[\frac{1}{J_n} \sum_{j=1}^{J_n} \varphi(\Theta_{n-1}, X_{n,j}) \right], \quad \text{and} \quad (7.178)$$

$$\Theta_n = \Theta_{n-1} - \frac{\gamma_n(1 - \alpha_n)}{1 - \prod_{l=1}^n \alpha_l} \left[\frac{1}{J_n} \sum_{j=1}^{J_n} \varphi(\Theta_{n-1}, X_{n,j}) \right] - \frac{\gamma_{n+1}\alpha_{n+1}\mathbf{m}_n}{1 - \prod_{l=1}^{n+1} \alpha_l}. \quad (7.179)$$

Algorithm 7.5.20: Shifted bias-adjusted Nesterov accelerated SGD optimization method

Input: $\mathfrak{d}, \mathbf{d}, N \in \mathbb{N}$, $\ell = (\ell(\theta, x))_{(\theta, x) \in \mathbb{R}^\mathfrak{d} \times \mathbb{R}^\mathbf{d}} \in C^1(\mathbb{R}^\mathfrak{d} \times \mathbb{R}^\mathbf{d}, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, probability space $(\Omega, \mathcal{F}, \mathbb{P})$, random variable $\xi: \Omega \rightarrow \mathbb{R}^\mathfrak{d}$, random variables $X_{n,j}: \Omega \rightarrow \mathbb{R}^\mathbf{d}$ for $n, j \in \mathbb{N}$

Output: N -th step of the Nesterov accelerated SGD process for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ (cf. Definition 7.5.19)

```

1: Initialization:  $\Theta \leftarrow \xi; \mathbf{m} \leftarrow 0 \in \mathbb{R}^\mathfrak{d}$ 
2: for  $n = 1, \dots, N$  do
3:    $g \leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_\theta \ell)(\Theta, X_{n,j})$ 
4:    $\mathbf{m} \leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n)g$ 
5:    $\Theta \leftarrow \Theta - \frac{\gamma_n(1 - \alpha_n)g}{1 - \prod_{l=1}^n \alpha_l} - \frac{\gamma_{n+1}\alpha_{n+1}\mathbf{m}}{1 - \prod_{l=1}^{n+1} \alpha_l}$ 
6: return  $\Theta$ 

```

7.5.4 Simplified Nesterov accelerated momentum optimization

In this section we introduce the stochastic version of the simplified Nesterov accelerated GD optimization method from Section 6.4.5.

Definition 7.5.21 (Simplified Nesterov accelerated SGD optimization method). Let $\mathfrak{d} \in \mathbb{N}$, let (S, \mathcal{S}) be a measurable space, let $\ell = (\ell(\theta, x))_{(\theta, x) \in \mathbb{R}^\mathfrak{d} \times S}: \mathbb{R}^\mathfrak{d} \times S \rightarrow \mathbb{R}$ be measurable, assume for all $x \in S$ that $\ell(\cdot, x)$ is differentiable, let $\varphi: \mathbb{R}^\mathfrak{d} \times S \rightarrow \mathbb{R}^\mathfrak{d}$ satisfy for all $\theta \in \mathbb{R}^\mathfrak{d}$, $x \in S$ that

$$\varphi(\theta, x) = (\nabla_\theta \ell)(\theta, x), \quad (7.180)$$

let $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\xi: \Omega \rightarrow \mathbb{R}^\mathfrak{d}$ be a random variable, for every $n, j \in \mathbb{N}$ let $X_{n,j}: \Omega \rightarrow S$ be a random variable, and let $\Theta: \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^\mathfrak{d}$ be a function. Then we say that Θ is the

simplified Nesterov accelerated SGD process for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ if and only if there exists $\mathbf{m}: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ such that for all $n \in \mathbb{N}$ it holds that

$$\Theta_0 = \xi, \quad \mathbf{m}_0 = 0, \quad (7.181)$$

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + \frac{1}{J_n} \sum_{j=1}^{J_n} \mathcal{g}(\Theta_{n-1}, X_{n,j}), \quad \text{and} \quad (7.182)$$

$$\Theta_n = \Theta_{n-1} - \gamma_n \alpha_n \mathbf{m}_n - \gamma_n \left[\frac{1}{J_n} \sum_{j=1}^{J_n} \mathcal{g}(\Theta_{n-1}, X_{n,j}) \right]. \quad (7.183)$$

Algorithm 7.5.22: Simplified Nesterov accelerated SGD optimization method

Input: $\mathfrak{d}, \mathbf{d}, N \in \mathbb{N}$, $\ell = (\ell(\theta, x))_{(\theta,x) \in \mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\mathbf{d}}} \in C^1(\mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\mathbf{d}}, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, probability space $(\Omega, \mathcal{F}, \mathbb{P})$, random variable $\xi: \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$, random variables $X_{n,j}: \Omega \rightarrow \mathbb{R}^{\mathbf{d}}$ for $n, j \in \mathbb{N}$

Output: N -th step of the simplified Nesterov accelerated SGD process for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ (cf. Definition 7.5.21)

```

1: Initialization:  $\Theta \leftarrow \xi; \mathbf{m} \leftarrow 0 \in \mathbb{R}^{\mathfrak{d}}$ 
2: for  $n = 1, \dots, N$  do
3:    $g \leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_{\theta} \ell)(\Theta, X_{n,j})$ 
4:    $\mathbf{m} \leftarrow \alpha_n \mathbf{m} + g$ 
5:    $\Theta \leftarrow \Theta - \gamma_n \alpha_n \mathbf{m} - \gamma_n g$ 
6: return  $\Theta$ 

```

The simplified Nesterov accelerated SGD optimization method as described in Definition 7.5.21 is implemented in PYTORCH in the form of the `torch.optim.SGD` optimizer with the `nesterov=True` option.

7.6 Adagrad optimization

In this section we introduce the stochastic version of the Adagrad GD optimization method from Section 6.5 (cf. Duchi et al. [123]).

Definition 7.6.1 (Adagrad SGD optimization method). Let $\mathfrak{d} \in \mathbb{N}$, let (S, \mathcal{S}) be a

measurable space, let $\ell = (\ell(\theta, x))_{(\theta, x) \in \mathbb{R}^d \times S} : \mathbb{R}^d \times S \rightarrow \mathbb{R}$ be measurable, assume for all $x \in S$ that $\ell(\cdot, x)$ is differentiable, let $\mathcal{G} : \mathbb{R}^d \times S \rightarrow \mathbb{R}^d$ satisfy for all $\theta \in \mathbb{R}^d$, $x \in S$ that

$$\mathcal{G}(\theta, x) = (\nabla_\theta \ell)(\theta, x), \quad (7.184)$$

let $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $\varepsilon \in (0, \infty)$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\xi : \Omega \rightarrow \mathbb{R}^d$ be a random variable, for every $n, j \in \mathbb{N}$ let $X_{n,j} : \Omega \rightarrow S$ be a random variable, and let $\Theta : \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^d$ be a function. Then we say that Θ is the **Adagrad SGD** process for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, regularizing factor ε , initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ if and only if it holds for all $n \in \mathbb{N}$, $i \in \{1, 2, \dots, d\}$ that

$$\Theta_n^{(i)} = \Theta_{n-1}^{(i)} - \gamma_n \left[\varepsilon + \left[\sum_{k=1}^n \left(\frac{1}{J_k} \sum_{j=1}^{J_k} \mathcal{G}_i(\Theta_{k-1}, X_{k,j}) \right)^2 \right]^{1/2} \right]^{-1} \left[\frac{1}{J_n} \sum_{j=1}^{J_n} \mathcal{G}_i(\Theta_{n-1}, X_{n,j}) \right]. \quad (7.185)$$

Algorithm 7.6.2: Adagrad SGD optimization method

Input: $d, d, N \in \mathbb{N}$, $\ell = (\ell(\theta, x))_{(\theta, x) \in \mathbb{R}^d \times \mathbb{R}^d} \in C^1(\mathbb{R}^d \times \mathbb{R}^d, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $\varepsilon \in (0, \infty)$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, probability space $(\Omega, \mathcal{F}, \mathbb{P})$, random variable $\xi : \Omega \rightarrow \mathbb{R}^d$, random variables $X_{n,j} : \Omega \rightarrow \mathbb{R}^d$ for $n, j \in \mathbb{N}$

Output: N -th step of the **Adagrad SGD** process for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, regularizing factor ε , initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ (cf. Definition 7.6.1)

```

1: Initialization:  $\Theta \leftarrow \xi; M \leftarrow 0 \in \mathbb{R}^d$ 
2: for  $n = 1, \dots, N$  do  $\#$  (cf. Definition 6.5.2)
3:    $g \leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_\theta \ell)(\Theta, X_{n,j})$ 
4:    $M \leftarrow M + g^2$ 
5:    $\Theta \leftarrow \Theta - \gamma_n [\varepsilon + M^{1/2}]^{-1} g$ 
6: return  $\Theta$ 

```

An implementation in PYTORCH of the **Adagrad SGD** optimization method as described in Definition 7.6.1 above is given in Source code 7.7. The **Adagrad SGD** optimization method as described in Definition 7.6.1 above is also available in PYTORCH in the form of the built-in `torch.optim.Adagrad` optimizer (which, for applications, is generally much preferable to implementing it from scratch).

```

1 import torch
2 import torch.nn as nn
3 import numpy as np

```

```

4
5     net = nn.Sequential(
6         nn.Linear(1, 200), nn.ReLU(), nn.Linear(200, 1)
7     )
8
9     M = 1000
10
11    X = torch.rand((M, 1)) * 4 * np.pi - 2 * np.pi
12    Y = torch.sin(X)
13
14    J = 64
15
16    N = 150000
17
18    loss = nn.MSELoss()
19    lr = 0.02
20    eps = 1e-10
21
22    sum_sq_grad = [p.clone().detach().fill_(eps) for p in net.
23                    parameters()]
24
25    for n in range(N):
26        indices = torch.randint(0, M, (J,))
27
28        x = X[indices]
29        y = Y[indices]
30
31        net.zero_grad()
32
33        loss_val = loss(net(x), y)
34        loss_val.backward()
35
36        with torch.no_grad():
37            for a, p in zip(sum_sq_grad, net.parameters()):
38                a.add_(p.grad * p.grad)
39                p.sub_(lr * a.rsqrt() * p.grad)
40
41        if n % 1000 == 0:
42            with torch.no_grad():
43                x = torch.rand((1000, 1)) * 4 * np.pi - 2 * np.pi
44                y = torch.sin(x)
45                loss_val = loss(net(x), y)
46                print(f"Iteration: {n+1}, Loss: {loss_val}")

```

Source code 7.7 ([code/optimization_methods/adagrad.py](#)): PYTHON code implementing the Adagrad SGD optimization method in PYTORCH

7.7 RMSprop optimization

In this section we introduce the stochastic version of the RMSprop GD optimization method from Section 6.6 (cf. Hinton et al. [209]).

Definition 7.7.1 (RMSprop SGD optimization method). Let $\mathfrak{d} \in \mathbb{N}$, let (S, \mathcal{S}) be a measurable space, let $\ell = (\ell(\theta, x))_{(\theta, x) \in \mathbb{R}^{\mathfrak{d}} \times S} : \mathbb{R}^{\mathfrak{d}} \times S \rightarrow \mathbb{R}$ be measurable, assume for all $x \in S$ that $\ell(\cdot, x)$ is differentiable, let $\varphi = (\varphi_1, \dots, \varphi_{\mathfrak{d}}) : \mathbb{R}^{\mathfrak{d}} \times S \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}, x \in S$ that

$$\varphi(\theta, x) = (\nabla_{\theta} \ell)(\theta, x), \quad (7.186)$$

let $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\beta_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $\varepsilon \in (0, \infty)$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\xi : \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a random variable, for every $n, j \in \mathbb{N}$ let $X_{n,j} : \Omega \rightarrow S$ be a random variable, and let $\Theta = (\Theta^{(1)}, \dots, \Theta^{(\mathfrak{d})}) : \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a function. Then we say that Θ is the RMSprop SGD process for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, second moment decay factors $(\beta_n)_{n \in \mathbb{N}}$, regularizing factor ε , initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ if and only if there exist $\mathbb{M} = (\mathbb{M}^{(1)}, \dots, \mathbb{M}^{(\mathfrak{d})}) : \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ such that for all $n \in \mathbb{N}, i \in \{1, 2, \dots, \mathfrak{d}\}$ it holds that

$$\Theta_0 = \xi, \quad \mathbb{M}_0 = 0, \quad (7.187)$$

$$\mathbb{M}_n^{(i)} = \beta_n \mathbb{M}_{n-1}^{(i)} + (1 - \beta_n) \left[\frac{1}{J_n} \sum_{j=1}^{J_n} \varphi_i(\Theta_{n-1}, X_{n,j}) \right]^2, \quad (7.188)$$

$$\text{and} \quad \Theta_n^{(i)} = \Theta_{n-1}^{(i)} - \gamma_n \left[\varepsilon + [\mathbb{M}_n^{(i)}]^{1/2} \right]^{-1} \left[\frac{1}{J_n} \sum_{j=1}^{J_n} \varphi_i(\Theta_{n-1}, X_{n,j}) \right]. \quad (7.189)$$

Algorithm 7.7.2: RMSprop SGD optimization method

Input: $\mathfrak{d}, \mathbf{d}, N \in \mathbb{N}$, $\ell = (\ell(\theta, x))_{(\theta, x) \in \mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\mathbf{d}}} \in C^1(\mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\mathbf{d}}, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\beta_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $\varepsilon \in (0, \infty)$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, probability space $(\Omega, \mathcal{F}, \mathbb{P})$, random variable $\xi : \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$, random variables $X_{n,j} : \Omega \rightarrow \mathbb{R}^{\mathbf{d}}$ for $n, j \in \mathbb{N}$

Output: N -th step of the RMSprop SGD process for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, second moment decay factors $(\beta_n)_{n \in \mathbb{N}}$, regularizing factor ε , initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ (cf. Definition 7.7.1)

- 1: **Initialization:** $\Theta \leftarrow \xi; \mathbb{M} \leftarrow 0 \in \mathbb{R}^{\mathfrak{d}}$
- 2: **for** $n = 1, \dots, N$ **do** $\#$ (cf. Definition 6.5.2)
- 3: $g \leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_{\theta} \ell)(\Theta, X_{n,j})$
- 4: $\mathbb{M} \leftarrow \beta_n \mathbb{M} + (1 - \beta_n) g^2$

```

5:      $\Theta \leftarrow \Theta - \gamma_n [\varepsilon + \mathbb{M}^{1/2}]^{-1} g$ 
6: return  $\Theta$ 
```

An implementation in PYTORCH of the RMSprop SGD optimization method as described in Definition 7.7.1 above is given in Source code 7.8. The RMSprop SGD optimization method as described in Definition 7.7.1 above is also available in PYTORCH in the form of the built-in `torch.optim.RMSprop` optimizer (which, for applications, is generally much preferable to implementing it from scratch).

```

1 import torch
2 import torch.nn as nn
3 import numpy as np
4
5 net = nn.Sequential(
6     nn.Linear(1, 200), nn.ReLU(), nn.Linear(200, 1)
7 )
8
9 M = 1000
10
11 X = torch.rand((M, 1)) * 4 * np.pi - 2 * np.pi
12 Y = torch.sin(X)
13
14 J = 64
15
16 N = 150000
17
18 loss = nn.MSELoss()
19 lr = 0.001
20 beta = 0.9
21 eps = 1e-10
22
23 moments = [p.clone().detach().zero_() for p in net.parameters()]
24
25 for n in range(N):
26     indices = torch.randint(0, M, (J,))
27
28     x = X[indices]
29     y = Y[indices]
30
31     net.zero_grad()
32
33     loss_val = loss(net(x), y)
34     loss_val.backward()
35
36     with torch.no_grad():
37         for m, p in zip(moments, net.parameters()):
38             m.mul_(beta)
```

```

39     m.add_((1 - beta) * p.grad * p.grad)
40     p.sub_(lr * (eps + m).rsqrt() * p.grad)
41
42     if n % 1000 == 0:
43         with torch.no_grad():
44             x = torch.rand((1000, 1)) * 4 * np.pi - 2 * np.pi
45             y = torch.sin(x)
46             loss_val = loss(net(x), y)
47             print(f"Iteration: {n+1}, Loss: {loss_val}")

```

Source code 7.8 ([code/optimization_methods/rmsprop.py](#)): PYTHON code implementing the RMSprop SGD optimization method in PYTORCH

7.7.1 Bias-adjusted RMSprop optimization

Definition 7.7.3 (Bias-adjusted RMSprop SGD optimization method). Let $\mathfrak{d} \in \mathbb{N}$, let (S, \mathcal{S}) be a measurable space, let $\ell = (\ell(\theta, x))_{(\theta, x) \in \mathbb{R}^{\mathfrak{d}} \times S}: \mathbb{R}^{\mathfrak{d}} \times S \rightarrow \mathbb{R}$ be measurable, assume for all $x \in S$ that $\ell(\cdot, x)$ is differentiable, let $\mathcal{g} = (\mathcal{g}_1, \dots, \mathcal{g}_{\mathfrak{d}}): \mathbb{R}^{\mathfrak{d}} \times S \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}, x \in S$ that

$$\mathcal{g}(\theta, x) = (\nabla_{\theta} \ell)(\theta, x), \quad (7.190)$$

let $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\beta_n)_{n \in \mathbb{N}} \subseteq [0, 1)$, $\varepsilon \in (0, \infty)$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\xi: \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a random variable, for every $n, j \in \mathbb{N}$ let $X_{n,j}: \Omega \rightarrow S$ be a random variable, and let $\Theta = (\Theta^{(1)}, \dots, \Theta^{(\mathfrak{d})}): \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a function. Then we say that Θ is the bias-adjusted RMSprop SGD process for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, second moment decay factors $(\beta_n)_{n \in \mathbb{N}}$, regularizing factor ε , initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ if and only if there exists $\mathbb{M} = (\mathbb{M}^{(1)}, \dots, \mathbb{M}^{(\mathfrak{d})}): \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ such that for all $n \in \mathbb{N}, i \in \{1, 2, \dots, \mathfrak{d}\}$ it holds that

$$\Theta_0 = \xi, \quad \mathbb{M}_0 = 0, \quad (7.191)$$

$$\mathbb{M}_n^{(i)} = \beta_n \mathbb{M}_{n-1}^{(i)} + (1 - \beta_n) \left[\frac{1}{J_n} \sum_{j=1}^{J_n} \mathcal{g}_i(\Theta_{n-1}, X_{n,j}) \right]^2, \quad \text{and} \quad (7.192)$$

$$\Theta_n^{(i)} = \Theta_{n-1}^{(i)} - \gamma_n \left[\varepsilon + \left[\frac{\mathbb{M}_n^{(i)}}{(1 - \prod_{l=1}^n \beta_l)} \right]^{1/2} \right]^{-1} \left[\frac{1}{J_n} \sum_{j=1}^{J_n} \mathcal{g}_i(\Theta_{n-1}, X_{n,j}) \right]. \quad (7.193)$$

An implementation in PYTORCH of the bias-adjusted RMSprop SGD optimization method as described in Definition 7.7.3 above is given in Source code 7.9.

```

1 import torch

```

```

2 import torch.nn as nn
3 import numpy as np
4
5 net = nn.Sequential(
6     nn.Linear(1, 200), nn.ReLU(), nn.Linear(200, 1)
7 )
8
9 M = 1000
10
11 X = torch.rand((M, 1)) * 4 * np.pi - 2 * np.pi
12 Y = torch.sin(X)
13
14 J = 64
15
16 N = 150000
17
18 loss = nn.MSELoss()
19 lr = 0.001
20 beta = 0.9
21 eps = 1e-10
22 adj = 1
23
24 moments = [p.clone().detach().zero_() for p in net.parameters()]
25
26 for n in range(N):
27     indices = torch.randint(0, M, (J,))
28
29     x = X[indices]
30     y = Y[indices]
31
32     net.zero_grad()
33
34     loss_val = loss(net(x), y)
35     loss_val.backward()
36
37     with torch.no_grad():
38         adj *= beta
39         for m, p in zip(moments, net.parameters()):
40             m.mul_(beta)
41             m.add_((1 - beta) * p.grad * p.grad)
42             p.sub_(lr * (eps + (m / (1 - adj)).sqrt()).reciprocal()
43             * p.grad)
44
45     if n % 1000 == 0:
46         with torch.no_grad():
47             x = torch.rand((1000, 1)) * 4 * np.pi - 2 * np.pi
48             y = torch.sin(x)
49             loss_val = loss(net(x), y)
50             print(f"Iteration: {n+1}, Loss: {loss_val}")

```

Source code 7.9 ([code/optimization_methods/rmsprop_bias_adj.py](#)): PYTHON code implementing the bias-adjusted RMSprop SGD optimization method in PYTORCH

7.8 Adadelta optimization

In this section we introduce the stochastic version of the Adadelta GD optimization method from Section 6.7 (cf. Zeiler [450]).

Definition 7.8.1 (Adadelta SGD optimization method). Let $\mathfrak{d} \in \mathbb{N}$, let (S, \mathcal{S}) be a measurable space, let $\ell = (\ell(\theta, x))_{(\theta, x) \in \mathbb{R}^{\mathfrak{d}} \times S}: \mathbb{R}^{\mathfrak{d}} \times S \rightarrow \mathbb{R}$ be measurable, assume for all $x \in S$ that $\ell(\cdot, x)$ is differentiable, let $\varphi = (\varphi_1, \dots, \varphi_{\mathfrak{d}}): \mathbb{R}^{\mathfrak{d}} \times S \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}, x \in S$ that

$$\varphi(\theta, x) = (\nabla_{\theta} \ell)(\theta, x), \quad (7.194)$$

let $(\beta_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $(\delta_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $\varepsilon \in (0, \infty)$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\xi: \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a random variable, for every $n, j \in \mathbb{N}$ let $X_{n,j}: \Omega \rightarrow S$ be a random variable, and let $\Theta = (\Theta^{(1)}, \dots, \Theta^{(\mathfrak{d})}): \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a function. Then we say that Θ is the Adadelta SGD process for the loss function ℓ with second moment decay factors $(\beta_n)_{n \in \mathbb{N}}$, delta decay factors $(\delta_n)_{n \in \mathbb{N}}$, regularizing factor ε , initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ if and only if there exist $\mathbb{M} = (\mathbb{M}^{(1)}, \dots, \mathbb{M}^{(\mathfrak{d})}): \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ and $\Delta = (\Delta^{(1)}, \dots, \Delta^{(\mathfrak{d})}): \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ such that for all $n \in \mathbb{N}, i \in \{1, 2, \dots, \mathfrak{d}\}$ it holds that

$$\Theta_0 = \xi, \quad \mathbb{M}_0 = 0, \quad \Delta_0 = 0, \quad (7.195)$$

$$\mathbb{M}_n^{(i)} = \beta_n \mathbb{M}_{n-1}^{(i)} + (1 - \beta_n) \left[\frac{1}{J_n} \sum_{j=1}^{J_n} \varphi_i(\Theta_{n-1}, X_{n,j}) \right]^2, \quad (7.196)$$

$$\Theta_n^{(i)} = \Theta_{n-1}^{(i)} - \left(\frac{\varepsilon + \Delta_{n-1}^{(i)}}{\varepsilon + \mathbb{M}_n^{(i)}} \right)^{1/2} \left[\frac{1}{J_n} \sum_{j=1}^{J_n} \varphi_i(\Theta_{n-1}, X_{n,j}) \right], \quad (7.197)$$

$$\text{and} \quad \Delta_n^{(i)} = \delta_n \Delta_{n-1}^{(i)} + (1 - \delta_n) |\Theta_n^{(i)} - \Theta_{n-1}^{(i)}|^2. \quad (7.198)$$

Algorithm 7.8.2: Adadelta SGD optimization method

Input: $\mathfrak{d}, \mathbf{d}, N \in \mathbb{N}$, $\ell = (\ell(\theta, x))_{(\theta, x) \in \mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\mathbf{d}}} \in C^1(\mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\mathbf{d}}, \mathbb{R})$, $(\beta_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $(\delta_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $\varepsilon \in (0, \infty)$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, probability space $(\Omega, \mathcal{F}, \mathbb{P})$, random variable

$\xi: \Omega \rightarrow \mathbb{R}^d$, random variables $X_{n,j}: \Omega \rightarrow \mathbb{R}^d$ for $n, j \in \mathbb{N}$

Output: N -th step of the Adadelta [SGD](#) process for the loss function ℓ with second moment decay factors $(\beta_n)_{n \in \mathbb{N}}$, delta decay factors $(\delta_n)_{n \in \mathbb{N}}$, regularizing factor ε , initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ (cf. Definition 7.8.1)

```

1: Initialization:  $\Theta \leftarrow \xi; \Xi \leftarrow \xi; \mathbb{M} \leftarrow 0 \in \mathbb{R}^d; \Delta \leftarrow 0 \in \mathbb{R}^d$ 
2: for  $n = 1, \dots, N$  do # (cf. Definition 6.5.2)
3:    $g \leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_\theta \ell)(\Theta, X_{n,j})$ 
4:    $\mathbb{M} \leftarrow \beta_n \mathbb{M} + (1 - \beta_n) g^2$ 
5:    $\Theta \leftarrow \Theta - [\frac{\varepsilon + \Delta}{\varepsilon + \mathbb{M}}]^{1/2} g$ 
6:    $\Delta \leftarrow \delta_n \Delta + (1 - \delta_n)[\Theta - \Xi]^2$ 
7:    $\Xi \leftarrow \Theta$ 
8: return  $\Theta$ 

```

An implementation in PYTORCH of the Adadelta [SGD](#) optimization method as described in Definition 7.8.1 above is given in Source code 7.10. The Adadelta [SGD](#) optimization method as described in Definition 7.8.1 above is also available in PYTORCH in the form of the built-in `torch.optim.Adadelta` optimizer (which, for applications, is generally much preferable to implementing it from scratch).

```

1 import torch
2 import torch.nn as nn
3 import numpy as np
4
5 net = nn.Sequential(
6     nn.Linear(1, 200), nn.ReLU(), nn.Linear(200, 1)
7 )
8
9 M = 1000
10
11 X = torch.rand((M, 1)) * 4 * np.pi - 2 * np.pi
12 Y = torch.sin(X)
13
14 J = 64
15
16 N = 150000
17
18 loss = nn.MSELoss()
19 beta = 0.9
20 delta = 0.9
21 eps = 1e-10
22
23 moments = [p.clone().detach().zero_() for p in net.parameters()]
24 Delta = [p.clone().detach().zero_() for p in net.parameters()]

```

```

25
26     for n in range(N):
27         indices = torch.randint(0, M, (J,))
28
29         x = X[indices]
30         y = Y[indices]
31
32         net.zero_grad()
33
34         loss_val = loss(net(x), y)
35         loss_val.backward()
36
37         with torch.no_grad():
38             for m, D, p in zip(moments, Delta, net.parameters()):
39                 m.mul_(beta)
40                 m.add_((1 - beta) * p.grad * p.grad)
41                 inc = ((eps + D) / (eps + m)).sqrt() * p.grad
42                 p.sub_(inc)
43                 D.mul_(delta)
44                 D.add_((1 - delta) * inc * inc)
45
46         if n % 1000 == 0:
47             with torch.no_grad():
48                 x = torch.rand((1000, 1)) * 4 * np.pi - 2 * np.pi
49                 y = torch.sin(x)
50                 loss_val = loss(net(x), y)
51                 print(f"Iteration: {n+1}, Loss: {loss_val}")

```

Source code 7.10 ([code/optimization_methods/adadelta.py](#)): PYTHON code implementing the Adadelta SGD optimization method in PYTORCH

7.9 Adam optimization

In this section we introduce the stochastic version of the [Adam GD](#) optimization method from Section 6.8 (cf. Kingma & Ba [261]).

Definition 7.9.1 ([Adam SGD](#) optimization method). Let $\mathfrak{d} \in \mathbb{N}$, let (S, \mathcal{S}) be a measurable space, let $\ell = (\ell(\theta, x))_{(\theta, x) \in \mathbb{R}^{\mathfrak{d}} \times S}: \mathbb{R}^{\mathfrak{d}} \times S \rightarrow \mathbb{R}$ be measurable, assume for all $x \in S$ that $\ell(\cdot, x)$ is differentiable, let $\varrho = (\varrho_1, \dots, \varrho_{\mathfrak{d}}): \mathbb{R}^{\mathfrak{d}} \times S \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}, x \in S$ that

$$\varrho(\theta, x) = (\nabla_{\theta} \ell)(\theta, x), \quad (7.199)$$

let $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $(\beta_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $\varepsilon \in (0, \infty)$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\xi: \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a random variable, for every $n, j \in \mathbb{N}$

let $X_{n,j}: \Omega \rightarrow S$ be a random variable, and let $\Theta = (\Theta^{(1)}, \dots, \Theta^{(\mathfrak{d})}): \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a function. Then we say that Θ is the **Adam SGD** process for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, second moment decay factors $(\beta_n)_{n \in \mathbb{N}}$, regularizing factor ε , initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ if and only if there exist $\mathbf{m} = (\mathbf{m}^{(1)}, \dots, \mathbf{m}^{(\mathfrak{d})}): \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ and $\mathbb{M} = (\mathbb{M}^{(1)}, \dots, \mathbb{M}^{(\mathfrak{d})}): \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ such that for all $n \in \mathbb{N}$, $i \in \{1, 2, \dots, \mathfrak{d}\}$ it holds that

$$\Theta_0 = \xi, \quad \mathbf{m}_0 = 0, \quad \mathbb{M}_0 = 0, \quad (7.200)$$

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + (1 - \alpha_n) \left[\frac{1}{J_n} \sum_{j=1}^{J_n} \mathcal{g}(\Theta_{n-1}, X_{n,j}) \right], \quad (7.201)$$

$$\mathbb{M}_n^{(i)} = \beta_n \mathbb{M}_{n-1}^{(i)} + (1 - \beta_n) \left[\frac{1}{J_n} \sum_{j=1}^{J_n} \mathcal{g}_i(\Theta_{n-1}, X_{n,j}) \right]^2, \quad (7.202)$$

$$\text{and} \quad \Theta_n^{(i)} = \Theta_{n-1}^{(i)} - \gamma_n \left[\varepsilon + \left[\frac{\mathbb{M}_n^{(i)}}{(1 - \prod_{l=1}^n \beta_l)} \right]^{1/2} \right]^{-1} \left[\frac{\mathbf{m}_n^{(i)}}{(1 - \prod_{l=1}^n \alpha_l)} \right]. \quad (7.203)$$

Algorithm 7.9.2: Adam SGD optimization method

Input: $\mathfrak{d}, \mathbf{d}, N \in \mathbb{N}$, $\ell = (\ell(\theta, x))_{(\theta,x) \in \mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\mathfrak{d}}} \in C^1(\mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\mathfrak{d}}, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $(\beta_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $\varepsilon \in (0, \infty)$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, probability space $(\Omega, \mathcal{F}, \mathbb{P})$, random variable $\xi: \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$, random variables $X_{n,j}: \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ for $n, j \in \mathbb{N}$

Output: N -th step of the Adam SGD process for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, second moment decay factors $(\beta_n)_{n \in \mathbb{N}}$, regularizing factor ε , initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ (cf. Definition 7.9.1)

```

1: Initialization:  $\Theta \leftarrow \xi; \mathbf{m} \leftarrow 0 \in \mathbb{R}^{\mathfrak{d}}; \mathbb{M} \leftarrow 0 \in \mathbb{R}^{\mathfrak{d}}$ 
2: for  $n = 1, \dots, N$  do # (cf. Definition 6.5.2)
3:    $g \leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_{\theta} \ell)(\Theta, X_{n,j})$ 
4:    $\mathbf{m} \leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n) g$ 
5:    $\mathbb{M} \leftarrow \beta_n \mathbb{M} + (1 - \beta_n) g^2$ 
6:    $\Theta \leftarrow \Theta - \left[ \varepsilon + \left[ \frac{\mathbb{M}}{1 - \prod_{k=1}^n \beta_k} \right]^{1/2} \right]^{-1} \left[ \frac{\gamma_n \mathbf{m}}{1 - \prod_{k=1}^n \alpha_k} \right]$ 
7: return  $\Theta$ 

```

Remark 7.9.3. In Kingma & Ba [261] it is proposed to choose

$$0.001 = \gamma_1 = \gamma_2 = \dots, \quad 0.9 = \alpha_1 = \alpha_2 = \dots, \quad 0.999 = \beta_1 = \beta_2 = \dots, \quad (7.204)$$

7.9. Adam optimization

and $10^{-8} = \varepsilon$ as default values for $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $(\beta_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $\varepsilon \in (0, \infty)$ in Definition 7.9.1.

An implementation in PYTORCH of the Adam SGD optimization method as described in Definition 7.9.1 above is given in Source code 7.11. The Adam SGD optimization method as described in Definition 7.9.1 above is also available in PYTORCH in the form of the built-in `torch.optim.Adam` optimizer (which, for applications, is generally much preferable to implementing it from scratch).

```
1 import torch
2 import torch.nn as nn
3 import numpy as np
4
5 net = nn.Sequential(
6     nn.Linear(1, 200), nn.ReLU(), nn.Linear(200, 1)
7 )
8
9 M = 1000
10
11 X = torch.rand((M, 1)) * 4 * np.pi - 2 * np.pi
12 Y = torch.sin(X)
13
14 J = 64
15
16 N = 150000
17
18 loss = nn.MSELoss()
19 lr = 0.0001
20 alpha = 0.9
21 beta = 0.999
22 eps = 1e-8
23 adj = 1.
24 adj2 = 1.
25
26 m = [p.clone().detach().zero_() for p in net.parameters()]
27 MM = [p.clone().detach().zero_() for p in net.parameters()]
28
29 for n in range(N):
30     indices = torch.randint(0, M, (J,))
31
32     x = X[indices]
33     y = Y[indices]
34
35     net.zero_grad()
36
37     loss_val = loss(net(x), y)
38     loss_val.backward()
39
40     with torch.no_grad():
```

```

41         adj *= alpha
42         adj2 *= beta
43         for m_p, M_p, p in zip(m, MM, net.parameters()):
44             m_p.mul_(alpha)
45             m_p.add_((1 - alpha) * p.grad)
46             M_p.mul_(beta)
47             M_p.add_((1 - beta) * p.grad * p.grad)
48             p.sub_(lr * m_p / ((1 - adj) * (eps + (M_p / (1 - adj2)
49             ).sqrt())))
50
51     if n % 1000 == 0:
52         with torch.no_grad():
53             x = torch.rand((1000, 1)) * 4 * np.pi - 2 * np.pi
54             y = torch.sin(x)
55             loss_val = loss(net(x), y)
56             print(f"Iteration: {n+1}, Loss: {loss_val}")

```

Source code 7.11 ([code/optimization_methods/adam.py](#)): PYTHON code implementing the Adam SGD optimization method in PYTORCH

Whereas Source code 7.11 and the other source codes presented in this chapter so far served mostly to elucidate the definitions of the various optimization methods introduced in this chapter by giving example implementations, in Source code 7.12 we demonstrate how an actual machine learning problem might be solved using the built-in functionality of PYTORCH. This code trains a neural network with 3 convolutional layers and 2 fully connected layers (with each hidden layer followed by a ReLU activation function) on the MNIST dataset (introduced in Bottou et al. [49]), which consists of 28×28 pixel grayscale images of handwritten digits from 0 to 9 and the corresponding labels and is one of the most commonly used benchmarks for training machine learning systems in the literature. Source code 7.12 uses the cross-entropy loss function and the Adam SGD optimization method and outputs a graph showing the progression of the average loss on the training set and on a test set that is not used for training as well as the accuracy of the model's predictions over the course of the training, see Figure 7.4.

```

1 import torch
2 import torchvision.datasets as datasets
3 import torchvision.transforms as transforms
4 import torch.nn as nn
5 import torch.utils.data as data
6 import torch.optim as optim
7 import matplotlib.pyplot as plt
8 from matplotlib.ticker import ScalarFormatter, NullFormatter
9
10 # We use the GPU if available. Otherwise, we use the CPU.
11 device = torch.device(
12     "cuda" if torch.cuda.is_available() else "cpu"
13 )

```

7.9. Adam optimization

```
14 # We fix a random seed. This is not necessary for training a
15 # neural network, but we use it here to ensure that the same
16 # plot is created on every run.
17 torch.manual_seed(0)
18
19
20 # The torch.utils.data.Dataset class is an abstraction for a
21 # collection of instances that has a length and can be indexed
22 # (usually by integers).
23 # The torchvision.datasets module contains functions for loading
24 # popular machine learning datasets, possibly downloading and
25 # transforming the data.
26
27 # Here we load the MNIST dataset, containing 28x28 grayscale images
28 # of handwritten digits with corresponding labels in
29 # {0, 1, ..., 9}.
30
31 # First load the training portion of the data set, downloading it
32 # from an online source to the local folder ./data (if it is not
33 # yet there) and transforming the data to PyTorch Tensors.
34 mnist_train = datasets.MNIST(
35     "./data",
36     train=True,
37     transform=transforms.ToTensor(),
38     download=True,
39 )
40 # Next load the test portion
41 mnist_test = datasets.MNIST(
42     "./data",
43     train=False,
44     transform=transforms.ToTensor(),
45     download=True,
46 )
47
48 # The data.utils.DataLoader class allows iterating datasets for
49 # training and validation. It supports, e.g., batching and
50 # shuffling of datasets.
51
52 # Construct a DataLoader that when iterating returns minibatches
53 # of 64 instances drawn from a random permutation of the training
54 # dataset
55 train_loader = data.DataLoader(
56     mnist_train, batch_size=64, shuffle=True
57 )
58 # The loader for the test dataset does not need shuffling
59 test_loader = data.DataLoader(
60     mnist_test, batch_size=64, shuffle=False
61 )
62
```

```

63 # Define a neural network with 3 convolutional layers, each
64 # followed by a ReLU activation and then two affine layers,
65 # the first followed by a ReLU activation
66 net = nn.Sequential( # input shape (N, 1, 28, 28)
67     nn.Conv2d(1, 5, 5), # (N, 5, 24, 24)
68     nn.ReLU(),
69     nn.Conv2d(5, 5, 5), # (N, 5, 20, 20)
70     nn.ReLU(),
71     nn.Conv2d(5, 3, 5), # (N, 3, 16, 16)
72     nn.ReLU(),
73     nn.Flatten(), # (N, 3 * 16 * 16) = (N, 768)
74     nn.Linear(768, 128), # (N, 128)
75     nn.ReLU(),
76     nn.Linear(128, 10), # output shape (N, 10)
77 ).to(device)
78
79 # Define the loss function. For every natural number d, for
80 # e_1, e_2, ..., e_d the standard basis vectors in R^d, for L the
81 # d-dimensional cross-entropy loss function, and for A the
82 # d-dimensional softmax activation function, the function loss_fn
83 # defined here satisfies for all x in R^d and all natural numbers
84 # i in [0,d) that
85 # loss_fn(x, i) = L(A(x), e_i).
86 # The function loss_fn also accepts batches of inputs, in which
87 # case it will return the mean of the corresponding outputs.
88 loss_fn = nn.CrossEntropyLoss()
89
90 # Define the optimizer. We use the Adam SGD optimization method.
91 optimizer = optim.Adam(net.parameters(), lr=1e-3)
92
93 # This function computes the average loss of the model over the
94 # entire test set and the accuracy of the model's predictions.
95 def compute_test_loss_and_accuracy():
96     total_test_loss = 0.0
97     correct_count = 0
98     with torch.no_grad():
99         # On each iteration the test_loader will yield a
100        # minibatch of images with corresponding labels
101        for images, labels in test_loader:
102            # Move the data to the device
103            images = images.to(device)
104            labels = labels.to(device)
105            # Compute the output of the neural network on the
106            # current minibatch
107            output = net(images)
108            # Compute the mean of the cross-entropy losses
109            loss = loss_fn(output, labels)
110            # For the cumulative total_test_loss, we multiply loss
111            # with the batch size (usually 64, as specified above,

```

7.9. Adam optimization

```
112     # but might be less for the final batch).
113     total_test_loss += loss.item() * images.size(0)
114     # For each input, the predicted label is the index of
115     # the maximal component in the output vector.
116     pred_labels = torch.max(output, dim=1).indices
117     # pred_labels == labels compares the two vectors
118     # componentwise and returns a vector of booleans.
119     # Summing over this vector counts the number of True
120     # entries.
121     correct_count += torch.sum(
122         pred_labels == labels
123     ).item()
124     avg_test_loss = total_test_loss / len(mnist_test)
125     accuracy = correct_count / len(mnist_test)
126     return (avg_test_loss, accuracy)
127
128
129 # Initialize a list that holds the computed loss on every
130 # batch during training
131 train_losses = []
132
133 # Every 10 batches, we will compute the loss on the entire test
134 # set as well as the accuracy of the model's predictions on the
135 # entire test set. We do this for the purpose of illustrating in
136 # the produced plot the generalization capability of the ANN.
137 # Computing these losses and accuracies so frequently with such a
138 # relatively large set of datapoints (compared to the training
139 # set) is extremely computationally expensive, however (most of
140 # the training runtime will be spent computing these values) and
141 # so is not advisable during normal neural network training.
142 # Usually, the test set is only used at the very end to judge the
143 # performance of the final trained network. Often, a third set of
144 # datapoints, called the validation set (not used to train the
145 # network directly nor to evaluate it at the end) is used to
146 # judge overfitting or to tune hyperparameters.
147 test_interval = 10
148 test_losses = []
149 accuracies = []
150
151 # We run the training for 5 epochs, i.e., 5 full iterations
152 # through the training set.
153 i = 0
154 for e in range(5):
155     for images, labels in train_loader:
156         # Move the data to the device
157         images = images.to(device)
158         labels = labels.to(device)
159
160         # Zero out the gradients
```

```

161     optimizer.zero_grad()
162     # Compute the output of the neural network on the current
163     # minibatch
164     output = net(images)
165     # Compute the cross entropy loss
166     loss = loss_fn(output, labels)
167     # Compute the gradients
168     loss.backward()
169     # Update the parameters of the neural network
170     optimizer.step()

171
172     # Append the current loss to the list of training losses.
173     # Note that tracking the training loss comes at
174     # essentially no computational cost (since we have to
175     # compute these values anyway) and so is typically done
176     # during neural network training to gauge the training
177     # progress.
178     train_losses.append(loss.item())

179
180     if (i + 1) % test_interval == 0:
181         # Compute the average loss on the test set and the
182         # accuracy of the model and add the values to the
183         # corresponding list
184         test_loss, accuracy = compute_test_loss_and_accuracy()
185         test_losses.append(test_loss)
186         accuracies.append(accuracy)

187
188         i += 1

189
190 fig, ax1 = plt.subplots(figsize=(12, 8))
191 # We plot the training losses, test losses, and accuracies in the
192 # same plot, but using two different y-axes
193 ax2 = ax1.twinx()

194
195 # Use a logarithmic scale for the losses
196 ax1.set_yscale("log")
197 # Use a logit scale for the accuracies
198 ax2.set_yscale("logit")
199 ax2.set_ylim((0.3, 0.99))
200 N = len(test_losses) * test_interval
201 ax2.set_xlim((0, N))
202 # Plot the training losses
203 (training_loss_line,) = ax1.plot(
204     train_losses,
205     label="Training loss (left axis)",
206 )
207 # Plot test losses
208 (test_loss_line,) = ax1.plot(
209     range(0, N, test_interval),

```

```

210     test_losses,
211     label="Test loss (left axis)",
212 )
213 # Plot the accuracies
214 (accuracies_line,) = ax2.plot(
215     range(0, N, test_interval),
216     accuracies,
217     label="Accuracy (right axis)",
218     color="red",
219 )
220 ax2.yaxis.set_major_formatter(ScalarFormatter())
221 ax2.yaxis.set_minor_formatter(NullFormatter())
222
223 # Put all the labels in a common legend
224 lines = [training_loss_line, test_loss_line, accuracies_line]
225 labels = [l.get_label() for l in lines]
226 ax2.legend(lines, labels)
227
228 plt.tight_layout()
229 plt.savefig("../plots/mnist.pdf", bbox_inches="tight")

```

Source code 7.12 ([code/mnist.py](#)): PYTHON code training an ANN on the MNIST dataset in PYTORCH. This code produces a plot showing the progression of the average loss on the test set and the accuracy of the model's predictions, see Figure 7.4.

Source code 7.13 compares the performance of several of the optimization methods introduced in this chapter, namely the plain vanilla SGD optimization method introduced in Definition 7.2.1, the momentum SGD optimization method introduced in Definition 7.4.1, the simplified Nesterov accelerated SGD optimization method introduced in Definition 7.5.21, the Adagrad SGD optimization method introduced in Definition 7.6.1, the RMSprop SGD optimization method introduced in Definition 7.7.1, the Adadelta SGD optimization method introduced in Definition 7.8.1, and the Adam SGD optimization method introduced in Definition 7.9.1, during training of an ANN on the MNIST dataset. The code produces two plots showing the progression of the training loss as well as the accuracy of the model's predictions on the test set, see Figure 7.5. Note that this compares the performance of the optimization methods only on one particular problem and without any efforts towards choosing good hyperparameters for the considered optimization methods. Thus, the results are not necessarily representative of the performance of these optimization methods in general.

```

1 import torch
2 import torchvision.datasets as datasets
3 import torchvision.transforms as transforms
4 import torch.nn as nn
5 import torch.utils.data as data
6 import torch.optim as optim
7 import matplotlib.pyplot as plt

```

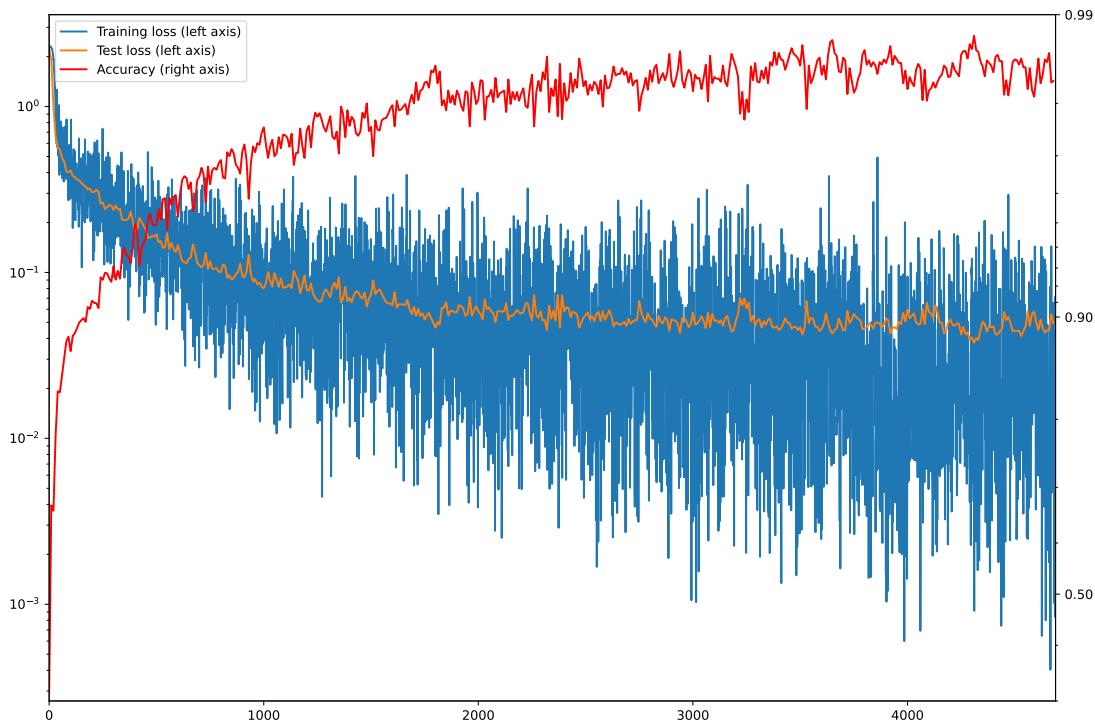


Figure 7.4 ([plots/mnist.pdf](#)): The plot produced by Source code 7.12, showing the average loss over each minibatch used during training (training loss) as well as the average loss over the test set and the accuracy of the model's predictions over the course of the training.

```

8  from matplotlib.ticker import ScalarFormatter, NullFormatter
9  import copy
10
11 # Set device as GPU if available or CPU otherwise
12 device = torch.device(
13     "cuda" if torch.cuda.is_available() else "cpu"
14 )
15
16 # Fix a random seed
17 torch.manual_seed(0)
18
19 # Load the MNIST training and test datasets
20 mnist_train = datasets.MNIST(
21     "./data",
22     train=True,
23     transform=transforms.ToTensor(),
24     download=True,
25 )
26 mnist_test = datasets.MNIST(

```

```

27     "./data",
28     train=False,
29     transform=transforms.ToTensor(),
30     download=True,
31 )
32 train_loader = data.DataLoader(
33     mnist_train, batch_size=64, shuffle=True
34 )
35 test_loader = data.DataLoader(
36     mnist_test, batch_size=64, shuffle=False
37 )
38
39 # Define a neural network
40 net = nn.Sequential( # input shape (N, 1, 28, 28)
41     nn.Conv2d(1, 5, 5), # (N, 5, 24, 24)
42     nn.ReLU(),
43     nn.Conv2d(5, 5, 3), # (N, 5, 22, 22)
44     nn.ReLU(),
45     nn.Conv2d(5, 3, 3), # (N, 3, 20, 20)
46     nn.ReLU(),
47     nn.Flatten(), # (N, 3 * 16 * 16) = (N, 1200)
48     nn.Linear(1200, 128), # (N, 128)
49     nn.ReLU(),
50     nn.Linear(128, 10), # output shape (N, 10)
51 ).to(device)
52
53 # Save the initial state of the neural network
54 initial_state = copy.deepcopy(net.state_dict())
55
56 # Define the loss function
57 loss_fn = nn.CrossEntropyLoss()
58
59 # Define the optimizers that we want to compare. Each entry in the
60 # list is a tuple of a label (for the plot) and an optimizer
61 optimizers = [
62     # For SGD we use a learning rate of 0.001
63     (
64         "SGD",
65         optim.SGD(net.parameters(), lr=1e-3),
66     ),
67     (
68         "SGD with momentum",
69         optim.SGD(net.parameters(), lr=1e-3, momentum=0.9),
70     ),
71     (
72         "Nesterov SGD",
73         optim.SGD(
74             net.parameters(), lr=1e-3, momentum=0.9, nesterov=True
75         ),

```

```

76     ),
77     # For the adaptive optimization methods we use the default
78     # hyperparameters
79     (
80         "RMSprop",
81         optim.RMSprop(net.parameters())),
82     ),
83     (
84         "Adagrad",
85         optim.Adagrad(net.parameters())),
86     ),
87     (
88         "Adadelta",
89         optim.Adadelta(net.parameters())),
90     ),
91     (
92         "Adam",
93         optim.Adam(net.parameters())),
94     ),
95 ]
96
97 def compute_test_loss_and_accuracy():
98     total_test_loss = 0.0
99     correct_count = 0
100    with torch.no_grad():
101        for images, labels in test_loader:
102            images = images.to(device)
103            labels = labels.to(device)
104
105            output = net(images)
106            loss = loss_fn(output, labels)
107
108            total_test_loss += loss.item() * images.size(0)
109            pred_labels = torch.max(output, dim=1).indices
110            correct_count += torch.sum(
111                pred_labels == labels
112            ).item()
113
114    avg_test_loss = total_test_loss / len(mnist_test)
115    accuracy = correct_count / len(mnist_test)
116
117    return (avg_test_loss, accuracy)
118
119
120 loss_plots = []
121 accuracy_plots = []
122
123 test_interval = 100
124

```

```

125 | for _, optimizer in optimizers:
126 |     train_losses = []
127 |     accuracies = []
128 |     print(optimizer)
129 |
130 |     with torch.no_grad():
131 |         net.load_state_dict(initial_state)
132 |
133 |     i = 0
134 |     for e in range(5):
135 |         print(f"Epoch {e+1}")
136 |         for images, labels in train_loader:
137 |             images = images.to(device)
138 |             labels = labels.to(device)
139 |
140 |             optimizer.zero_grad()
141 |             output = net(images)
142 |             loss = loss_fn(output, labels)
143 |             loss.backward()
144 |             optimizer.step()
145 |
146 |             train_losses.append(loss.item())
147 |
148 |             if (i + 1) % test_interval == 0:
149 |                 (
150 |                     test_loss,
151 |                     accuracy,
152 |                 ) = compute_test_loss_and_accuracy()
153 |                 print(accuracy)
154 |                 accuracies.append(accuracy)
155 |
156 |             i += 1
157 |
158 |     loss_plots.append(train_losses)
159 |     accuracy_plots.append(accuracies)
160 |
161 | WINDOW = 200
162 |
163 | _, (ax1, ax2) = plt.subplots(2, 1, figsize=(10, 12))
164 | ax1.set_yscale("log")
165 | ax2.set_yscale("logit")
166 | ax2.yaxis.set_major_formatter(ScalarFormatter())
167 | ax2.yaxis.set_minor_formatter(NullFormatter())
168 | for (label, _), train_losses, accuracies in zip(
169 |     optimizers, loss_plots, accuracy_plots
170 | ):
171 |     ax1.plot(
172 |         [
173 |             sum(train_losses[max(0, i-WINDOW) : i]) / min(i, WINDOW)

```

```

174         for i in range(1, len(train_losses))
175     ],
176     label=label,
177 )
178 ax2.plot(
179     range(0, len(accuracies) * test_interval, test_interval),
180     accuracies,
181     label=label,
182 )
183
184 ax1.legend()
185
186 plt.tight_layout()
187 plt.savefig("../plots/mnist_optim.pdf", bbox_inches="tight")

```

Source code 7.13 ([code/mnist_optim.py](#)): PYTHON code comparing the performance of several optimization methods during training of an ANN on the MNIST dataset. See Figure 7.5 for the plots produced by this code.

7.9.1 Adamax optimization

In this section we introduce the stochastic version of the Adamax GD optimization method from Section 6.8.1 (cf. Kingma & Ba [261]).

Definition 7.9.4 (Adamax SGD optimization method). *Let $\mathfrak{d} \in \mathbb{N}$, let (S, \mathcal{S}) be a measurable space, let $\ell = (\ell(\theta, x))_{(\theta, x) \in \mathbb{R}^{\mathfrak{d}} \times S}: \mathbb{R}^{\mathfrak{d}} \times S \rightarrow \mathbb{R}$ be measurable, assume for all $x \in S$ that $\ell(\cdot, x)$ is differentiable, let $\mathbf{g} = (g_1, \dots, g_{\mathfrak{d}}): \mathbb{R}^{\mathfrak{d}} \times S \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}, x \in S$ that*

$$\mathbf{g}(\theta, x) = (\nabla_{\theta} \ell)(\theta, x), \quad (7.205)$$

let $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, 1)$, $(\beta_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $\varepsilon \in (0, \infty)$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\xi: \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a random variable, for every $n, j \in \mathbb{N}$ let $X_{n,j}: \Omega \rightarrow S$ be a random variable, and let $\Theta = (\Theta^{(1)}, \dots, \Theta^{(\mathfrak{d})}): \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a function. Then we say that Θ is the Adamax SGD process for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, second moment decay factors $(\beta_n)_{n \in \mathbb{N}}$, regularizing factor ε , initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ if and only if there exist $\mathbf{m} = (\mathbf{m}^{(1)}, \dots, \mathbf{m}^{(\mathfrak{d})}): \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ and $\mathbb{M} = (\mathbb{M}^{(1)}, \dots, \mathbb{M}^{(\mathfrak{d})}): \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ such that for all $n \in \mathbb{N}, i \in \{1, 2, \dots, \mathfrak{d}\}$ it holds that

$$\Theta_0 = \xi, \quad \mathbf{m}_0 = 0, \quad \mathbb{M}_0 = 0, \quad (7.206)$$

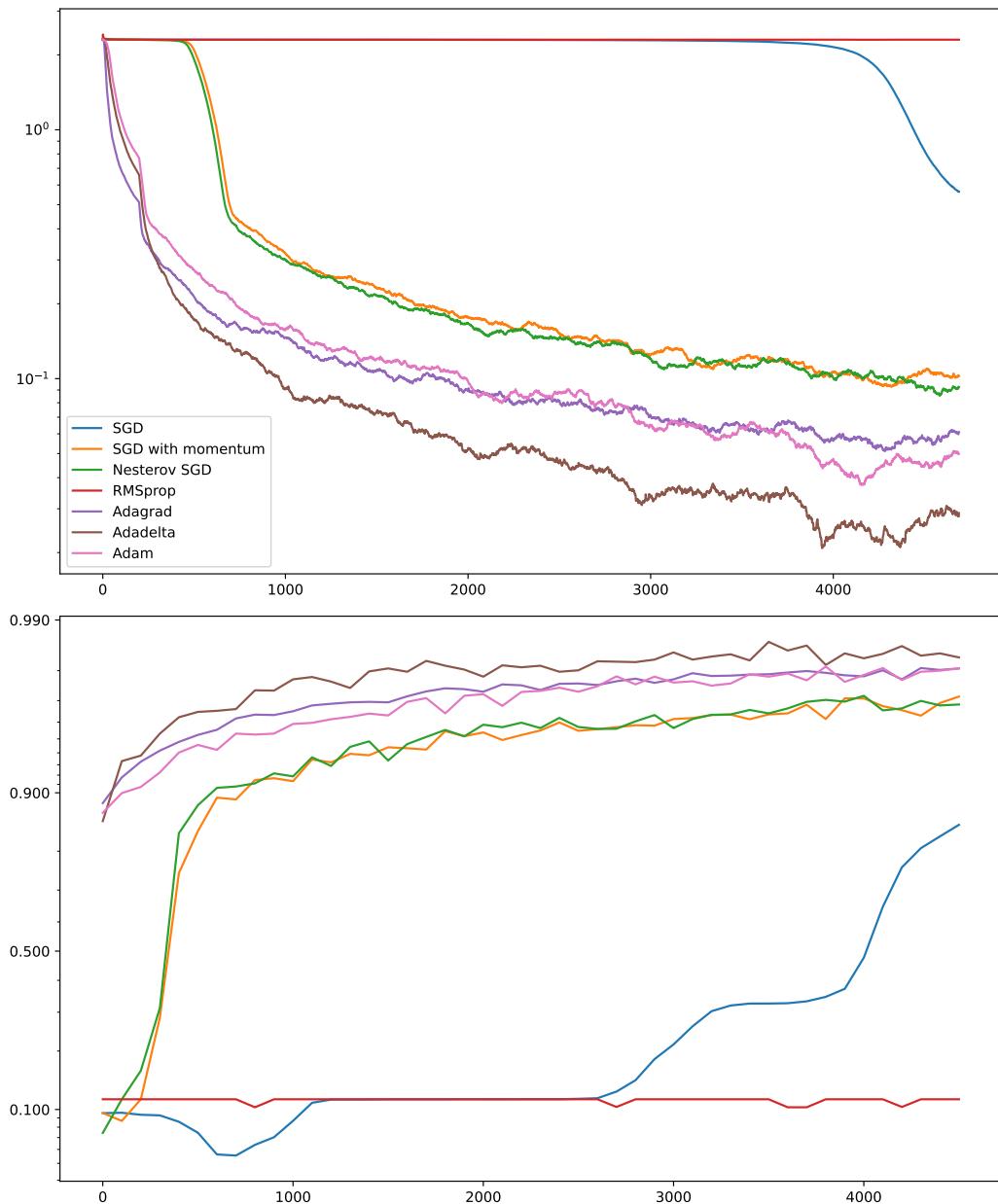


Figure 7.5 ([plots/mnist_optim.pdf](#)): The plots produced by Source code 7.13. The upper plot shows the progression of the training loss during the training of the ANNs. More precisely, each line shows a moving average of the training loss over 200 minibatches during the training of an ANN with the corresponding optimization method. The lower plot shows the accuracy of the ANN’s predictions on the test set over the course of the training with each optimization method.

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + (1 - \alpha_n) \left[\frac{1}{J_n} \sum_{j=1}^{J_n} \mathcal{g}(\Theta_{n-1}, X_{n,j}) \right], \quad (7.207)$$

$$\mathbb{M}_n^{(i)} = \max \left\{ \beta_n \mathbb{M}_{n-1}^{(i)}, \left| \frac{1}{J_n} \sum_{j=1}^{J_n} \mathcal{g}_i(\Theta_{n-1}, X_{n,j}) \right| \right\}, \quad (7.208)$$

$$\text{and} \quad \Theta_n^{(i)} = \Theta_{n-1}^{(i)} - \gamma_n \left[\varepsilon + \mathbb{M}_n^{(i)} \right]^{-1} \left[\frac{\mathbf{m}_n^{(i)}}{(1 - \prod_{l=1}^n \alpha_l)} \right]. \quad (7.209)$$

Algorithm 7.9.5: Adamax SGD optimization method

Input: $\mathfrak{d}, \mathbf{d}, N \in \mathbb{N}$, $\ell = (\ell(\theta, x))_{(\theta,x) \in \mathbb{R}^\mathfrak{d} \times \mathbb{R}^\mathbf{d}} \in C^1(\mathbb{R}^\mathfrak{d} \times \mathbb{R}^\mathbf{d}, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $(\beta_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $\varepsilon \in (0, \infty)$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, probability space $(\Omega, \mathcal{F}, \mathbb{P})$, random variable $\xi: \Omega \rightarrow \mathbb{R}^\mathfrak{d}$, random variables $X_{n,j}: \Omega \rightarrow \mathbb{R}^\mathbf{d}$ for $n, j \in \mathbb{N}$

Output: N -th step of the Adamax SGD process for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, second moment decay factors $(\beta_n)_{n \in \mathbb{N}}$, regularizing factor ε , initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ (cf. Definition 7.9.4)

```

1: Initialization:  $\Theta \leftarrow \xi; \mathbf{m} \leftarrow 0 \in \mathbb{R}^\mathfrak{d}; \mathbb{M} \leftarrow 0 \in \mathbb{R}^\mathfrak{d}$ 
2: for  $n = 1, \dots, N$  do # (cf. Definition 6.5.2)
3:    $g \leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_\theta \ell)(\Theta, X_{n,j})$ 
4:    $\mathbf{m} \leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n) g$ 
5:    $\mathbb{M} \leftarrow \max \{ \beta_n \mathbb{M}, |g| \}$ 
6:    $\Theta \leftarrow \Theta - \gamma_n [\varepsilon + \mathbb{M}]^{-1} \left[ \frac{\mathbf{m}}{1 - \prod_{k=1}^n \alpha_k} \right]$ 
7: return  $\Theta$ 

```

7.10 Nadam optimization

In this section we introduce the stochastic version of the Nadam GD optimization method from Section 6.9 (cf. Dozat [117, 118]).

Definition 7.10.1 (Nadam SGD optimization method). Let $\mathfrak{d} \in \mathbb{N}$, let (S, \mathcal{S}) be a measurable space, let $\ell = (\ell(\theta, x))_{(\theta,x) \in \mathbb{R}^\mathfrak{d} \times S}: \mathbb{R}^\mathfrak{d} \times S \rightarrow \mathbb{R}$ be measurable, assume for all $x \in S$ that $\ell(\cdot, x)$ is differentiable, let $\mathcal{g} = (\mathcal{g}_1, \dots, \mathcal{g}_\mathfrak{d}): \mathbb{R}^\mathfrak{d} \times S \rightarrow \mathbb{R}^\mathfrak{d}$ satisfy for all $\theta \in \mathbb{R}^\mathfrak{d}$, $x \in S$ that

$$\mathcal{g}(\theta, x) = (\nabla_\theta \ell)(\theta, x), \quad (7.210)$$

let $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, 1)$, $(\beta_n)_{n \in \mathbb{N}} \subseteq [0, 1)$, $\varepsilon \in (0, \infty)$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\xi: \Omega \rightarrow \mathbb{R}^d$ be a random variable, for every $n, j \in \mathbb{N}$ let $X_{n,j}: \Omega \rightarrow S$ be a random variable, and let $\Theta = (\Theta^{(1)}, \dots, \Theta^{(d)}): \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^d$ be a function. Then we say that Θ is the **Nadam SGD** process for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, second moment decay factors $(\beta_n)_{n \in \mathbb{N}}$, regularizing factor ε , initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ if and only if there exist $\mathbf{m} = (\mathbf{m}^{(1)}, \dots, \mathbf{m}^{(d)}): \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^d$ and $\mathbb{M} = (\mathbb{M}^{(1)}, \dots, \mathbb{M}^{(d)}): \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^d$ such that for all $n \in \mathbb{N}$, $i \in \{1, 2, \dots, d\}$ it holds that

$$\Theta_0 = \xi, \quad \mathbf{m}_0 = 0, \quad \mathbb{M}_0 = 0, \quad (7.211)$$

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + (1 - \alpha_n) \left[\frac{1}{J_n} \sum_{j=1}^{J_n} \mathcal{L}(\Theta_{n-1}, X_{n,j}) \right], \quad (7.212)$$

$$\mathbb{M}_n^{(i)} = \beta_n \mathbb{M}_{n-1}^{(i)} + (1 - \beta_n) \left[\frac{1}{J_n} \sum_{j=1}^{J_n} \mathcal{L}_i(\Theta_{n-1}, X_{n,j}) \right]^2, \quad \text{and} \quad (7.213)$$

$$\begin{aligned} \Theta_n^{(i)} = \Theta_{n-1}^{(i)} & - \left[\varepsilon + \left[\frac{\mathbb{M}_n^{(i)}}{(1 - \prod_{l=1}^n \beta_l)} \right]^{1/2} \right]^{-1} \left[\left[\frac{\gamma_n (1 - \alpha_n)}{1 - \prod_{l=1}^n \alpha_l} \right] \left[\frac{1}{J_n} \sum_{j=1}^{J_n} \mathcal{L}_i(\Theta_{n-1}, X_{n,j}) \right] \right. \\ & \left. + \left[\frac{\gamma_{n+1} \alpha_{n+1}}{1 - \prod_{l=1}^{n+1} \alpha_l} \right] \mathbf{m}_n^{(i)} \right]. \end{aligned} \quad (7.214)$$

Algorithm 7.10.2: Nadam SGD optimization method

Input: $d, d, N \in \mathbb{N}$, $\ell = (\ell(\theta, x))_{(\theta,x) \in \mathbb{R}^d \times \mathbb{R}^d} \in C^1(\mathbb{R}^d \times \mathbb{R}^d, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, 1)$, $(\beta_n)_{n \in \mathbb{N}} \subseteq [0, 1)$, $\varepsilon \in (0, \infty)$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, probability space $(\Omega, \mathcal{F}, \mathbb{P})$, random variable $\xi: \Omega \rightarrow \mathbb{R}^d$, random variables $X_{n,j}: \Omega \rightarrow \mathbb{R}^d$ for $n, j \in \mathbb{N}$

Output: N -th step of the **Nadam SGD** process for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, second moment decay factors $(\beta_n)_{n \in \mathbb{N}}$, regularizing factor ε , initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ (cf. Definition 7.10.1)

- 1: **Initialization:** $\Theta \leftarrow \xi$; $\mathbf{m} \leftarrow 0 \in \mathbb{R}^d$; $\mathbb{M} \leftarrow 0 \in \mathbb{R}^d$
- 2: **for** $n = 1, \dots, N$ **do** $\#$ (cf. Definition 6.5.2)
- 3: $g \leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_\theta \ell)(\Theta, X_{n,j})$
- 4: $\mathbf{m} \leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n) g$
- 5: $\mathbb{M} \leftarrow \beta_n \mathbb{M} + (1 - \beta_n) g^2$

```

6:    $\Theta \leftarrow \Theta - \left[ \varepsilon + \left[ \frac{\mathbb{M}}{1 - \prod_{k=1}^n \beta_k} \right]^{1/2} \right]^{-1} \left[ \left[ \frac{\gamma_n(1-\alpha_n)}{1 - \prod_{k=1}^n \alpha_k} \right] g + \left[ \frac{\gamma_{n+1}\alpha_{n+1}}{1 - \prod_{k=1}^{n+1} \alpha_k} \right] \mathbf{m} \right]$ 
7: return  $\Theta$ 

```

7.10.1 Simplified Nadam optimization

In this section we introduce the stochastic version of the simplified Nadam GD optimization method from Section 6.9.1.

Definition 7.10.3 (Simplified Nadam SGD optimization method). *Let $\mathfrak{d} \in \mathbb{N}$, let (S, \mathcal{S}) be a measurable space, let $\ell = (\ell(\theta, x))_{(\theta, x) \in \mathbb{R}^\mathfrak{d} \times S}: \mathbb{R}^\mathfrak{d} \times S \rightarrow \mathbb{R}$ be measurable, assume for all $x \in S$ that $\ell(\cdot, x)$ is differentiable, let $\mathcal{G} = (\mathcal{G}_1, \dots, \mathcal{G}_{\mathfrak{d}}): \mathbb{R}^\mathfrak{d} \times S \rightarrow \mathbb{R}^\mathfrak{d}$ satisfy for all $\theta \in \mathbb{R}^\mathfrak{d}$, $x \in S$ that*

$$\mathcal{G}(\theta, x) = (\nabla_\theta \ell)(\theta, x), \quad (7.215)$$

let $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, 1)$, $(\beta_n)_{n \in \mathbb{N}} \subseteq [0, 1)$, $\varepsilon \in (0, \infty)$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\xi: \Omega \rightarrow \mathbb{R}^\mathfrak{d}$ be a random variable, for every $n, j \in \mathbb{N}$ let $X_{n,j}: \Omega \rightarrow S$ be a random variable, and let $\Theta = (\Theta^{(1)}, \dots, \Theta^{(\mathfrak{d})}): \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^\mathfrak{d}$ be a function. Then we say that Θ is the simplified Simplified Nadam SGD process for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, second moment decay factors $(\beta_n)_{n \in \mathbb{N}}$, regularizing factor ε , initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ if and only if there exist $\mathbf{m} = (\mathbf{m}^{(1)}, \dots, \mathbf{m}^{(\mathfrak{d})}): \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^\mathfrak{d}$ and $\mathbb{M} = (\mathbb{M}^{(1)}, \dots, \mathbb{M}^{(\mathfrak{d})}): \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^\mathfrak{d}$ such that for all $n \in \mathbb{N}$, $i \in \{1, 2, \dots, \mathfrak{d}\}$ it holds that

$$\Theta_0 = \xi, \quad \mathbf{m}_0 = 0, \quad \mathbb{M}_0 = 0, \quad (7.216)$$

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + (1 - \alpha_n) \left[\frac{1}{J_n} \sum_{j=1}^{J_n} \mathcal{G}(\Theta_{n-1}, X_{n,j}) \right], \quad (7.217)$$

$$\mathbb{M}_n^{(i)} = \beta_n \mathbb{M}_{n-1}^{(i)} + (1 - \beta_n) \left[\frac{1}{J_n} \sum_{j=1}^{J_n} \mathcal{G}_i(\Theta_{n-1}, X_{n,j}) \right]^2, \quad \text{and} \quad (7.218)$$

$$\begin{aligned} \Theta_n^{(i)} &= \Theta_{n-1}^{(i)} - \gamma_n \left[\varepsilon + \left[\frac{\mathbb{M}_n^{(i)}}{(1 - \prod_{l=1}^n \beta_l)} \right]^{1/2} \right]^{-1} \left[\left[\frac{1 - \alpha_n}{1 - \prod_{l=1}^n \alpha_l} \right] \left[\frac{1}{J_n} \sum_{j=1}^{J_n} \mathcal{G}_i(\Theta_{n-1}, X_{n,j}) \right] \right. \\ &\quad \left. + \left[\frac{\alpha_{n+1}}{1 - \prod_{l=1}^{n+1} \alpha_l} \right] \mathbf{m}_n^{(i)} \right]. \end{aligned} \quad (7.219)$$

Algorithm 7.10.4: Simplified Nadam SGD optimization method

Input: $\mathfrak{d}, \mathbf{d}, N \in \mathbb{N}$, $\ell = (\ell(\theta, x))_{(\theta, x) \in \mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\mathbf{d}}} \in C^1(\mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\mathbf{d}}, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $(\beta_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $\varepsilon \in (0, \infty)$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, probability space $(\Omega, \mathcal{F}, \mathbb{P})$, random variable $\xi: \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$, random variables $X_{n,j}: \Omega \rightarrow \mathbb{R}^{\mathbf{d}}$ for $n, j \in \mathbb{N}$

Output: N -th step of the simplified Nadam SGD process for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, second moment decay factors $(\beta_n)_{n \in \mathbb{N}}$, regularizing factor ε , initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ (cf. Definition 7.10.3)

```

1: Initialization:  $\Theta \leftarrow \xi; \mathbf{m} \leftarrow 0 \in \mathbb{R}^{\mathfrak{d}}; \mathbb{M} \leftarrow 0 \in \mathbb{R}^{\mathfrak{d}}$ 
2: for  $n = 1, \dots, N$  do  $\#$  (cf. Definition 6.5.2)
3:    $g \leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_{\theta} \ell)(\Theta, X_{n,j})$ 
4:    $\mathbf{m} \leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n) g$ 
5:    $\mathbb{M} \leftarrow \beta_n \mathbb{M} + (1 - \beta_n) g^2$ 
6:    $\Theta \leftarrow \Theta - \gamma_n \left[ \varepsilon + \left[ \frac{\mathbb{M}}{1 - \prod_{k=1}^n \beta_k} \right]^{1/2} \right]^{-1} \left[ \left[ \frac{1 - \alpha_n}{1 - \prod_{k=1}^n \alpha_k} \right] g + \left[ \frac{\alpha_{n+1}}{1 - \prod_{k=1}^{n+1} \alpha_k} \right] \mathbf{m} \right]$ 
7: return  $\Theta$ 

```

7.10.2 Nadamax optimization

In this section we introduce the stochastic version of the Nadamax GD optimization method from Section 6.8.1 (cf. Dozat [117, 118]).

Definition 7.10.5 (Nadamax SGD optimization method). Let $\mathfrak{d} \in \mathbb{N}$, let (S, \mathcal{S}) be a measurable space, let $\ell = (\ell(\theta, x))_{(\theta, x) \in \mathbb{R}^{\mathfrak{d}} \times S}: \mathbb{R}^{\mathfrak{d}} \times S \rightarrow \mathbb{R}$ be measurable, assume for all $x \in S$ that $\ell(\cdot, x)$ is differentiable, let $\mathcal{G} = (\mathcal{G}_1, \dots, \mathcal{G}_{\mathfrak{d}}): \mathbb{R}^{\mathfrak{d}} \times S \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$, $x \in S$ that

$$\mathcal{G}(\theta, x) = (\nabla_{\theta} \ell)(\theta, x), \quad (7.220)$$

let $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $(\beta_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $\varepsilon \in (0, \infty)$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\xi: \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a random variable, for every $n, j \in \mathbb{N}$ let $X_{n,j}: \Omega \rightarrow S$ be a random variable, and let $\Theta = (\Theta^{(1)}, \dots, \Theta^{(\mathfrak{d})}): \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a function. Then we say that Θ is the Nadamax SGD process for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, second moment decay factors $(\beta_n)_{n \in \mathbb{N}}$, regularizing factor ε , initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ if and only if there exist $\mathbf{m} = (\mathbf{m}^{(1)}, \dots, \mathbf{m}^{(\mathfrak{d})}): \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ and $\mathbb{M} =$

$(\mathbb{M}^{(1)}, \dots, \mathbb{M}^{(\mathfrak{d})}): \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ such that for all $n \in \mathbb{N}$, $i \in \{1, 2, \dots, \mathfrak{d}\}$ it holds that

$$\Theta_0 = \xi, \quad \mathbf{m}_0 = 0, \quad \mathbb{M}_0 = 0, \quad (7.221)$$

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + (1 - \alpha_n) \left[\frac{1}{J_n} \sum_{j=1}^{J_n} \mathcal{g}_i(\Theta_{n-1}, X_{n,j}) \right], \quad (7.222)$$

$$\mathbb{M}_n^{(i)} = \max \left\{ \beta_n \mathbb{M}_{n-1}^{(i)}, \left| \frac{1}{J_n} \sum_{j=1}^{J_n} \mathcal{g}_i(\Theta_{n-1}, X_{n,j}) \right| \right\}, \quad \text{and} \quad (7.223)$$

$$\Theta_n^{(i)} = \Theta_{n-1}^{(i)} - [\varepsilon + \mathbb{M}_n^{(i)}]^{-1} \left[\left[\frac{\gamma_n(1-\alpha_n)}{1-\prod_{l=1}^n \alpha_l} \right] \left[\frac{1}{J_n} \sum_{j=1}^{J_n} \mathcal{g}_i(\Theta_{n-1}, X_{n,j}) \right] + \left[\frac{\gamma_{n+1}\alpha_{n+1}}{1-\prod_{l=1}^{n+1} \alpha_l} \right] \mathbf{m}_n^{(i)} \right]. \quad (7.224)$$

Algorithm 7.10.6: Nadamax SGD optimization method

Input: $\mathfrak{d}, \mathbf{d}, N \in \mathbb{N}$, $\ell = (\ell(\theta, x))_{(\theta,x) \in \mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\mathbf{d}}} \in C^1(\mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\mathbf{d}}, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $(\beta_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $\varepsilon \in (0, \infty)$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, probability space $(\Omega, \mathcal{F}, \mathbb{P})$, random variable $\xi: \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$, random variables $X_{n,j}: \Omega \rightarrow \mathbb{R}^{\mathbf{d}}$ for $n, j \in \mathbb{N}$

Output: N -th step of the Nadamax SGD process for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, second moment decay factors $(\beta_n)_{n \in \mathbb{N}}$, regularizing factor ε , initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ (cf. Definition 7.10.5)

```

1: Initialization:  $\Theta \leftarrow \xi; \mathbf{m} \leftarrow 0 \in \mathbb{R}^{\mathfrak{d}}; \mathbb{M} \leftarrow 0 \in \mathbb{R}^{\mathfrak{d}}$ 
2: for  $n = 1, \dots, N$  do # (cf. Definition 6.5.2)
3:    $g \leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_{\theta} \ell)(\Theta, X_{n,j})$ 
4:    $\mathbf{m} \leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n) g$ 
5:    $\mathbb{M} \leftarrow \max \{ \beta_n \mathbb{M}, |g| \}$ 
6:    $\Theta \leftarrow \Theta - [\varepsilon + \mathbb{M}]^{-1} \left[ \left[ \frac{\gamma_n(1-\alpha_n)}{1-\prod_{l=1}^n \alpha_l} \right] g + \left[ \frac{\gamma_{n+1}\alpha_{n+1}}{1-\prod_{l=1}^{n+1} \alpha_l} \right] \mathbf{m} \right]$ 
7: return  $\Theta$ 

```

7.11 AdamW optimization

In this section we introduce the stochastic version of the AdamW GD optimization method from Section 6.10 (cf. Loshchilov & Hutter [300]).

Definition 7.11.1 (**AdamW SGD** optimization method). Let $\mathfrak{d} \in \mathbb{N}$, let (S, \mathcal{S}) be a measurable space, let $\ell = (\ell(\theta, x))_{(\theta, x) \in \mathbb{R}^{\mathfrak{d}} \times S}: \mathbb{R}^{\mathfrak{d}} \times S \rightarrow \mathbb{R}$ be measurable, assume for all $x \in S$ that $\ell(\cdot, x)$ is differentiable, let $\varphi = (\varphi_1, \dots, \varphi_{\mathfrak{d}}): \mathbb{R}^{\mathfrak{d}} \times S \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}, x \in S$ that

$$\varphi(\theta, x) = (\nabla_{\theta} \ell)(\theta, x), \quad (7.225)$$

let $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, 1)$, $(\beta_n)_{n \in \mathbb{N}} \subseteq [0, 1)$, $\lambda \in \mathbb{R}$, $\varepsilon \in (0, \infty)$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\xi: \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a random variable, for every $n, j \in \mathbb{N}$ let $X_{n,j}: \Omega \rightarrow S$ be a random variable, and let $\Theta = (\Theta^{(1)}, \dots, \Theta^{(\mathfrak{d})}): \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a function. Then we say that Θ is the **AdamW SGD** process for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, second moment decay factors $(\beta_n)_{n \in \mathbb{N}}$, weight decay factor λ , regularizing factor ε , initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ if and only if there exist $\mathbf{m} = (\mathbf{m}^{(1)}, \dots, \mathbf{m}^{(\mathfrak{d})}): \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ and $\mathbb{M} = (\mathbb{M}^{(1)}, \dots, \mathbb{M}^{(\mathfrak{d})}): \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ such that for all $n \in \mathbb{N}$, $i \in \{1, 2, \dots, \mathfrak{d}\}$ it holds that

$$\Theta_0 = \xi, \quad \mathbf{m}_0 = 0, \quad \mathbb{M}_0 = 0, \quad (7.226)$$

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + (1 - \alpha_n) \left[\frac{1}{J_n} \sum_{j=1}^{J_n} \varphi_i(\Theta_{n-1}, X_{n,j}) \right], \quad (7.227)$$

$$\mathbb{M}_n^{(i)} = \beta_n \mathbb{M}_{n-1}^{(i)} + (1 - \beta_n) \left[\frac{1}{J_n} \sum_{j=1}^{J_n} \varphi_i(\Theta_{n-1}, X_{n,j}) \right]^2, \quad \text{and} \quad (7.228)$$

$$\Theta_n^{(i)} = \Theta_{n-1}^{(i)} - \gamma_n \left(\left[\varepsilon + \left[\frac{\mathbb{M}_n^{(i)}}{(1 - \prod_{l=1}^n \beta_l)} \right]^{1/2} \right]^{-1} \left[\frac{\mathbf{m}_n^{(i)}}{(1 - \prod_{l=1}^n \alpha_l)} \right] + \lambda \Theta_{n-1}^{(i)} \right). \quad (7.229)$$

Algorithm 7.11.2: AdamW SGD optimization method

Input: $\mathfrak{d}, \mathbf{d}, N \in \mathbb{N}$, $\ell = (\ell(\theta, x))_{(\theta, x) \in \mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\mathbf{d}}} \in C^1(\mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\mathbf{d}}, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, 1)$, $(\beta_n)_{n \in \mathbb{N}} \subseteq [0, 1)$, $\lambda \in \mathbb{R}$, $\varepsilon \in (0, \infty)$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, probability space $(\Omega, \mathcal{F}, \mathbb{P})$, random variable $\xi: \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$, random variables $X_{n,j}: \Omega \rightarrow \mathbb{R}^{\mathbf{d}}$ for $n, j \in \mathbb{N}$

Output: N -th step of the **AdamW SGD** process for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, second moment decay factors $(\beta_n)_{n \in \mathbb{N}}$, weight decay factor λ , regularizing factor ε , initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ (cf. Definition 7.11.1)

```

1: Initialization:  $\Theta \leftarrow \xi; \mathbf{m} \leftarrow 0 \in \mathbb{R}^{\mathfrak{d}}; \mathbb{M} \leftarrow 0 \in \mathbb{R}^{\mathfrak{d}}$ 
2: for  $n = 1, \dots, N$  do  $\#$  (cf. Definition 6.5.2)
3:    $g \leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_{\theta} \ell)(\Theta, X_{n,j})$ 

```

```

4:    $\mathbf{m} \leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n) g$ 
5:    $\mathbb{M} \leftarrow \beta_n \mathbb{M} + (1 - \beta_n) g^2$ 
6:    $\Theta \leftarrow \Theta - \gamma_n \left( \left[ \varepsilon + \left[ \frac{\mathbb{M}}{1 - \prod_{k=1}^n \beta_k} \right]^{1/2} \right]^{-1} \left[ \frac{\mathbf{m}}{1 - \prod_{k=1}^n \alpha_k} \right] + \lambda \Theta \right)$ 
7: return  $\Theta$ 

```

7.11.1 Adam with L^2 -regularization optimization

In this section we introduce the stochastic version of the [Adam GD](#) optimization method with L^2 -regularization from Section [6.10.1](#) (cf. Loshchilov & Hutter [300]).

Definition 7.11.3 ([Adam SGD](#) optimization method with L^2 -regularization). Let $\mathfrak{d} \in \mathbb{N}$, let (S, \mathcal{S}) be a measurable space, let $\ell = (\ell(\theta, x))_{(\theta, x) \in \mathbb{R}^\mathfrak{d} \times S}: \mathbb{R}^\mathfrak{d} \times S \rightarrow \mathbb{R}$ be measurable, assume for all $x \in S$ that $\ell(\cdot, x)$ is differentiable, let $\varphi = (\varphi_1, \dots, \varphi_\mathfrak{d}): \mathbb{R}^\mathfrak{d} \times S \rightarrow \mathbb{R}^\mathfrak{d}$ satisfy for all $\theta \in \mathbb{R}^\mathfrak{d}$, $x \in S$ that

$$\varphi(\theta, x) = (\nabla_\theta \ell)(\theta, x), \quad (7.230)$$

let $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, 1)$, $(\beta_n)_{n \in \mathbb{N}} \subseteq [0, 1)$, $\lambda \in \mathbb{R}$, $\varepsilon \in (0, \infty)$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\xi: \Omega \rightarrow \mathbb{R}^\mathfrak{d}$ be a random variable, for every $n, j \in \mathbb{N}$ let $X_{n,j}: \Omega \rightarrow S$ be a random variable, and let $\Theta = (\Theta^{(1)}, \dots, \Theta^{(\mathfrak{d})}): \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^\mathfrak{d}$ be a function. Then we say that Θ is the [Adam SGD](#) process for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, second moment decay factors $(\beta_n)_{n \in \mathbb{N}}$, L^2 -regularization factor λ , regularizing factor ε , initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ if and only if there exist $\mathbf{m} = (\mathbf{m}^{(1)}, \dots, \mathbf{m}^{(\mathfrak{d})}): \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^\mathfrak{d}$ and $\mathbb{M} = (\mathbb{M}^{(1)}, \dots, \mathbb{M}^{(\mathfrak{d})}): \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^\mathfrak{d}$ such that for all $n \in \mathbb{N}$, $i \in \{1, 2, \dots, \mathfrak{d}\}$ it holds that

$$\Theta_0 = \xi, \quad \mathbf{m}_0 = 0, \quad \mathbb{M}_0 = 0, \quad (7.231)$$

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + (1 - \alpha_n) \left[\lambda \Theta_{n-1} + \frac{1}{J_n} \sum_{j=1}^{J_n} \varphi(\Theta_{n-1}, X_{n,j}) \right], \quad (7.232)$$

$$\mathbb{M}_n^{(i)} = \beta_n \mathbb{M}_{n-1}^{(i)} + (1 - \beta_n) \left[\lambda \Theta_{n-1}^{(i)} + \frac{1}{J_n} \sum_{j=1}^{J_n} \varphi_i(\Theta_{n-1}, X_{n,j}) \right]^2, \quad (7.233)$$

$$\Theta_n^{(i)} = \Theta_{n-1}^{(i)} - \gamma_n \left[\varepsilon + \left[\frac{\mathbb{M}_n^{(i)}}{(1 - \prod_{l=1}^n \beta_l)} \right]^{1/2} \right]^{-1} \left[\frac{\mathbf{m}_n^{(i)}}{(1 - \prod_{l=1}^n \alpha_l)} \right]. \quad (7.234)$$

Algorithm 7.11.4: Adam SGD optimization method with L^2 -regularization

Input: $\mathbf{d}, \mathbf{d}, N \in \mathbb{N}$, $\ell = (\ell(\theta, x))_{(\theta, x) \in \mathbb{R}^{\mathbf{d}} \times \mathbb{R}^{\mathbf{d}}} \in C^1(\mathbb{R}^{\mathbf{d}} \times \mathbb{R}^{\mathbf{d}}, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $(\beta_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $\lambda \in \mathbb{R}$, $\varepsilon \in (0, \infty)$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, probability space $(\Omega, \mathcal{F}, \mathbb{P})$, random variable $\xi: \Omega \rightarrow \mathbb{R}^{\mathbf{d}}$, random variables $X_{n,j}: \Omega \rightarrow \mathbb{R}^{\mathbf{d}}$ for $n, j \in \mathbb{N}$

Output: N -th step of the Adam SGD process for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, second moment decay factors $(\beta_n)_{n \in \mathbb{N}}$, L^2 -regularization factor λ , regularizing factor ε , initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ (cf. Definition 7.11.3)

```

1: Initialization:  $\Theta \leftarrow \xi; \mathbf{m} \leftarrow 0 \in \mathbb{R}^{\mathbf{d}}; \mathbb{M} \leftarrow 0 \in \mathbb{R}^{\mathbf{d}}$ 
2: for  $n = 1, \dots, N$  do  $\#$  (cf. Definition 6.5.2)
3:    $g \leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_{\theta} \ell)(\Theta, X_{n,j})$ 
4:    $\mathbf{m} \leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n)(g + \lambda \Theta)$ 
5:    $\mathbb{M} \leftarrow \beta_n \mathbb{M} + (1 - \beta_n)[g + \lambda \Theta]^2$ 
6:    $\Theta \leftarrow \Theta - \left[ \varepsilon + \left[ \frac{\mathbb{M}}{1 - \prod_{k=1}^n \beta_k} \right]^{1/2} \right]^{-1} \left[ \frac{\gamma_n \mathbf{m}}{1 - \prod_{k=1}^n \alpha_k} \right]$ 
7: return  $\Theta$ 

```

Remark 7.11.5. In the PYTORCH implementation of the Adam SGD optimization method with L^2 -regularization in Definition 7.11.3 the default value for $\lambda \in \mathbb{R}$ is $\lambda = 0.01$.

7.12 Shampoo optimization

In this section we introduce the stochastic version of the Shampoo GD optimization method from Section 6.11 (cf. Gupta et al. [194]).

Definition 7.12.1 (Shampoo SGD optimization method). Let $\mathbf{d}_1, \mathbf{d}_2 \in \mathbb{N}$, let (S, \mathcal{S}) be a measurable space, let $\ell = (\ell(\theta, x))_{(\theta, x) \in \mathbb{R}^{\mathbf{d}_1 \times \mathbf{d}_2} \times S} : \mathbb{R}^{\mathbf{d}_1 \times \mathbf{d}_2} \times S \rightarrow \mathbb{R}$ be measurable, assume for all $x \in S$ that $\ell(\cdot, x)$ is differentiable, let $\mathcal{G} : \mathbb{R}^{\mathbf{d}_1 \times \mathbf{d}_2} \times S \rightarrow \mathbb{R}^{\mathbf{d}_1 \times \mathbf{d}_2}$ satisfy for all $\theta \in \mathbb{R}^{\mathbf{d}_1 \times \mathbf{d}_2}$, $x \in S$ that

$$\mathcal{G}(\theta, x) = (\nabla_{\theta} \ell)(\theta, x), \quad (7.235)$$

let $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $\varepsilon \in (0, \infty)$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\xi : \Omega \rightarrow \mathbb{R}^{\mathbf{d}_1 \times \mathbf{d}_2}$ be a random variable, for every $n, j \in \mathbb{N}$ let $X_{n,j} : \Omega \rightarrow S$ be a random variable, and let $\Theta : \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^{\mathbf{d}_1 \times \mathbf{d}_2}$ be a function. Then we say that Θ is the Shampoo SGD process for the objective function \mathcal{L} with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, regularizing factor ε , initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ if and only if there exist

$\mathbf{L}: \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^{\mathfrak{d}_1 \times \mathfrak{d}_1}$ and $\mathbf{R}: \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^{\mathfrak{d}_2 \times \mathfrak{d}_2}$ such that for all $n \in \mathbb{N}$ it holds that

$$\Theta_0 = \xi, \quad \mathbf{L}_0 = \varepsilon \mathbf{I}_{\mathfrak{d}_1}, \quad \mathbf{R}_0 = \varepsilon \mathbf{I}_{\mathfrak{d}_2}, \quad (7.236)$$

$$\mathbf{L}_n = \mathbf{L}_{n-1} + \left[\frac{1}{J_n} \sum_{j=1}^{J_n} \mathcal{g}(\Theta_{n-1}, X_{n,j}) \right] \left[\frac{1}{J_n} \sum_{j=1}^{J_n} \mathcal{g}(\Theta_{n-1}, X_{n,j}) \right]^*, \quad (7.237)$$

$$\mathbf{R}_n = \mathbf{R}_{n-1} + \left[\frac{1}{J_n} \sum_{j=1}^{J_n} \mathcal{g}(\Theta_{n-1}, X_{n,j}) \right]^* \left[\frac{1}{J_n} \sum_{j=1}^{J_n} \mathcal{g}(\Theta_{n-1}, X_{n,j}) \right], \quad (7.238)$$

$$\text{and} \quad \Theta_n = \Theta_{n-1} - \gamma_n (\mathbf{L}_n)^{-1/4} \left[\frac{1}{J_n} \sum_{j=1}^{J_n} \mathcal{g}(\Theta_{n-1}, X_{n,j}) \right] (\mathbf{R}_n)^{-1/4} \quad (7.239)$$

(cf. Definition 1.5.5 and Corollary 6.11.3).

Algorithm 7.12.2: Shampoo SGD optimization method

Input: $\mathfrak{d}_1, \mathfrak{d}_2, \mathbf{d}, N \in \mathbb{N}$, $\ell = (\ell(\theta, x))_{(\theta, x) \in \mathbb{R}^{\mathfrak{d}_1 \times \mathfrak{d}_2} \times \mathbb{R}^{\mathbf{d}}} \in C^1(\mathbb{R}^{\mathfrak{d}_1 \times \mathfrak{d}_2} \times \mathbb{R}^{\mathbf{d}}, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $\varepsilon \in (0, \infty)$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, probability space $(\Omega, \mathcal{F}, \mathbb{P})$, random variable $\xi: \Omega \rightarrow \mathbb{R}^{\mathfrak{d}_1 \times \mathfrak{d}_2}$, random variables $X_{n,j}: \Omega \rightarrow \mathbb{R}^{\mathbf{d}}$ for $n, j \in \mathbb{N}$

Output: N -th step of the Shampoo SGD process for the objective function \mathcal{L} with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, regularizing factor ε , initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ (cf. Definition 7.12.1)

```

1: Initialization:  $\Theta \leftarrow \xi; \mathbf{L} \leftarrow \varepsilon \mathbf{I}_{\mathfrak{d}_1}; \mathbf{R} \leftarrow \varepsilon \mathbf{I}_{\mathfrak{d}_2}$ 
2: for  $n = 1, \dots, N$  do
3:    $g \leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_\theta \ell)(\Theta, X_{n,j})$ 
4:    $\mathbf{L} \leftarrow \mathbf{L} + gg^*$ 
5:    $\mathbf{R} \leftarrow \mathbf{R} + g^*g$ 
6:    $\Theta \leftarrow \Theta - \gamma_n (\mathbf{L})^{-1/4} g (\mathbf{R})^{-1/4}$ 
7: return  $\Theta$ 

```

7.13 Muon optimization

In this section we introduce the stochastic version of the Muon GD optimization method from Section 6.12 (cf. Jordan et al. [245]).

Definition 7.13.1 (Idealized Muon SGD optimization method). Let $\mathfrak{d}_1, \mathfrak{d}_2, K \in \mathbb{N}$,

let (S, \mathcal{S}) be a measurable space, let $\ell = (\ell(\theta, x))_{(\theta, x) \in \mathbb{R}^{\mathfrak{d}_1 \times \mathfrak{d}_2} \times S} : \mathbb{R}^{\mathfrak{d}_1 \times \mathfrak{d}_2} \times S \rightarrow \mathbb{R}$ be measurable, assume for all $x \in S$ that $\ell(\cdot, x)$ is differentiable, let $\varphi : \mathbb{R}^{\mathfrak{d}_1 \times \mathfrak{d}_2} \times S \rightarrow \mathbb{R}^{\mathfrak{d}_1 \times \mathfrak{d}_2}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}_1 \times \mathfrak{d}_2}$, $x \in S$ that

$$\varphi(\theta, x) = (\nabla_\theta \ell)(\theta, x), \quad (7.240)$$

let $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, $\mathcal{O} = \{O \in \mathbb{R}^{\mathfrak{d}_1 \times \mathfrak{d}_2} : ((OO^* = I_{\mathfrak{d}_1}) \vee (O^*O = I_{\mathfrak{d}_2}))\}$, let $\Pi : \mathbb{R}^{\mathfrak{d}_1 \times \mathfrak{d}_2} \rightarrow \mathcal{O}$ satisfy for all $A \in \mathbb{R}^{\mathfrak{d}_1 \times \mathfrak{d}_2}$ that

$$\|\Pi(A) - A\|_{HS} = \inf_{O \in \mathcal{O}} \|O - A\|_{HS}, \quad (7.241)$$

let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\xi : \Omega \rightarrow \mathbb{R}^{\mathfrak{d}_1 \times \mathfrak{d}_2}$ be a random variable, for every $n, j \in \mathbb{N}$ let $X_{n,j} : \Omega \rightarrow S$ be a random variable, and let $\Theta : \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^{\mathfrak{d}_1 \times \mathfrak{d}_2}$ be a function (cf. Definitions 1.5.5 and 6.12.1). Then we say that Θ is the idealized Muon SGD process for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, orthogonal projection Π , initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ if and only if there exists $\mathbf{m} : \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ such that for all $n \in \mathbb{N}$ it holds that it holds for all $n \in \mathbb{N}$ that

$$\Theta_0 = \xi, \quad \mathbf{m}_0 = 0, \quad (7.242)$$

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + \left[\frac{1}{J_n} \sum_{j=1}^{J_n} \varphi(\Theta_{n-1}, X_{n,j}) \right], \quad (7.243)$$

$$\text{and} \quad \Theta_n = \Theta_{n-1} - \gamma_n \Pi(\mathbf{m}_n). \quad (7.244)$$

Algorithm 7.13.2: Idealized Muon SGD optimization method

Input: $\mathfrak{d}_1, \mathfrak{d}_2, \mathbf{d}, N \in \mathbb{N}$, $\ell = (\ell(\theta, x))_{(\theta, x) \in \mathbb{R}^{\mathfrak{d}_1 \times \mathfrak{d}_2} \times \mathbb{R}^{\mathbf{d}}} \in C^1(\mathbb{R}^{\mathfrak{d}_1 \times \mathfrak{d}_2} \times \mathbb{R}^{\mathbf{d}}, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, probability space $(\Omega, \mathcal{F}, \mathbb{P})$, random variable $\xi : \Omega \rightarrow \mathbb{R}^{\mathfrak{d}_1 \times \mathfrak{d}_2}$, random variables $X_{n,j} : \Omega \rightarrow \mathbb{R}^{\mathbf{d}}$ for $n, j \in \mathbb{N}$, $\Pi : \mathbb{R}^{\mathfrak{d}_1 \times \mathfrak{d}_2} \rightarrow \{O \in \mathbb{R}^{\mathfrak{d}_1 \times \mathfrak{d}_2} : ((OO^* = I_{\mathfrak{d}_1}) \vee (O^*O = I_{\mathfrak{d}_2}))\}$ with $\forall A \in \mathbb{R}^{\mathfrak{d}_1 \times \mathfrak{d}_2} : \|\Pi(A) - A\|_{HS} = \inf_{O \in \{O \in \mathbb{R}^{\mathfrak{d}_1 \times \mathfrak{d}_2} : ((OO^* = I_{\mathfrak{d}_1}) \vee (O^*O = I_{\mathfrak{d}_2}))\}} \|\mathcal{O} - A\|_{HS}$

Output: N -th step of the idealized Muon SGD process for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, orthogonal projection Π , initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ (cf. Definitions 1.5.5, 6.12.1, and 7.13.1)

- 1: **Initialization:** $\Theta \leftarrow \xi; \mathbf{m} \leftarrow 0 \in \mathbb{R}^{\mathfrak{d}}$
- 2: **for** $n = 1, \dots, N$ **do**
- 3: $g \leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_\theta \ell)(\Theta, X_{n,j})$

```

4:    $\mathbf{m} \leftarrow \alpha_n \mathbf{m} + g$ 
5:    $\Theta \leftarrow \Theta - \gamma_n \Pi(\mathbf{m})$ 
6: return  $\Theta$ 

```

Definition 7.13.3 (**Muon SGD** optimization method). Let $\mathfrak{d}_1, \mathfrak{d}_2, K \in \mathbb{N}$, let (S, \mathcal{S}) be a measurable space, let $\ell = (\ell(\theta, x))_{(\theta, x) \in \mathbb{R}^{\mathfrak{d}_1 \times \mathfrak{d}_2} \times S}: \mathbb{R}^{\mathfrak{d}_1 \times \mathfrak{d}_2} \times S \rightarrow \mathbb{R}$ be measurable, assume for all $x \in S$ that $\ell(\cdot, x)$ is differentiable, let $\mathcal{G}: \mathbb{R}^{\mathfrak{d}_1 \times \mathfrak{d}_2} \times S \rightarrow \mathbb{R}^{\mathfrak{d}_1 \times \mathfrak{d}_2}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}_1 \times \mathfrak{d}_2}$, $x \in S$ that

$$\mathcal{G}(\theta, x) = (\nabla_\theta \ell)(\theta, x), \quad (7.245)$$

let $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $a, b, c \in \mathbb{R}$, $\varepsilon \in (0, \infty)$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\xi: \Omega \rightarrow \mathbb{R}^{\mathfrak{d}_1 \times \mathfrak{d}_2}$ be a random variable, for every $n, j \in \mathbb{N}$ let $X_{n,j}: \Omega \rightarrow S$ be a random variable, and let $\Theta: \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^{\mathfrak{d}_1 \times \mathfrak{d}_2}$ be a function. Then we say that Θ is the **Muon SGD** process for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, Newton-Schulz method with polynomial coefficients a, b, c , regularization parameter ε , and K iterations, initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ if and only if there exists $\mathbf{m}: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ such that for all $n \in \mathbb{N}$ it holds that it holds for all $n \in \mathbb{N}$ that

$$\Theta_0 = \xi, \quad \mathbf{m}_0 = 0, \quad (7.246)$$

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + \left[\frac{1}{J_n} \sum_{j=1}^{J_n} \mathcal{G}(\Theta_{n-1}, X_{n,j}) \right], \quad (7.247)$$

$$\text{and} \quad \Theta_n = \Theta_{n-1} - \gamma_n \text{NS}_{a,b,c,\varepsilon}(\mathbf{m}_n, K). \quad (7.248)$$

Algorithm 7.13.4: **Muon SGD** optimization method

Input: $\mathfrak{d}_1, \mathfrak{d}_2, \mathbf{d}, N, K \in \mathbb{N}$, $\ell = (\ell(\theta, x))_{(\theta, x) \in \mathbb{R}^{\mathfrak{d}_1 \times \mathfrak{d}_2} \times \mathbb{R}^{\mathfrak{d}}}: \mathbb{R}^{\mathfrak{d}_1 \times \mathfrak{d}_2} \times \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $a, b, c \in \mathbb{R}$, $\varepsilon \in (0, \infty)$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, probability space $(\Omega, \mathcal{F}, \mathbb{P})$, random variable $\xi: \Omega \rightarrow \mathbb{R}^{\mathfrak{d}_1 \times \mathfrak{d}_2}$, random variables $X_{n,j}: \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ for $n, j \in \mathbb{N}$

Output: N -th step of the **Muon SGD** process for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, Newton-Schulz method with polynomial coefficients a, b, c , regularization parameter ε , and K iterations, initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ (cf. Definition 7.13.3)

```

1: Initialization:  $\Theta \leftarrow \xi; \mathbf{m} \leftarrow 0 \in \mathbb{R}^{\mathfrak{d}}$ 
2: for  $n = 1, \dots, N$  do
3:    $g \leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_\theta \ell)(\Theta, X_{n,j})$ 

```

```

4:    $\mathbf{m} \leftarrow \alpha_n \mathbf{m} + g$ 
5:    $\Theta \leftarrow \Theta - \gamma_n \text{NS}_{a,b,c,\varepsilon}(\mathbf{m}, K)$ 
6: return  $\Theta$ 

```

7.14 AMSGrad optimization

In this section we introduce the stochastic version of the AMSGrad [GD](#) optimization method from Section 6.13 (cf. Reddi et al. [372]).

Definition 7.14.1 (AMSGrad [SGD](#) optimization method). *Let $\mathfrak{d} \in \mathbb{N}$, let (S, \mathcal{S}) be a measurable space, let $\ell = (\ell(\theta, x))_{(\theta, x) \in \mathbb{R}^{\mathfrak{d}} \times S}: \mathbb{R}^{\mathfrak{d}} \times S \rightarrow \mathbb{R}$ be measurable, assume for all $x \in S$ that $\ell(\cdot, x)$ is differentiable, let $\varphi = (\varphi_1, \dots, \varphi_{\mathfrak{d}}): \mathbb{R}^{\mathfrak{d}} \times S \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}, x \in S$ that*

$$\varphi(\theta, x) = (\nabla_{\theta} \ell)(\theta, x), \quad (7.249)$$

let $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $(\beta_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $\varepsilon \in (0, \infty)$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\xi: \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a random variable, for every $n, j \in \mathbb{N}$ let $X_{n,j}: \Omega \rightarrow S$ be a random variable, and let $\Theta = (\Theta^{(1)}, \dots, \Theta^{(\mathfrak{d})}): \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ be a function. Then we say that Θ is the AMSGrad [SGD](#) process for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, second moment decay factors $(\beta_n)_{n \in \mathbb{N}}$, regularizing factor ε , initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ if and only if there exist $\mathbf{m} = (\mathbf{m}^{(1)}, \dots, \mathbf{m}^{(\mathfrak{d})}): \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ and $\mathbb{M} = (\mathbb{M}^{(1)}, \dots, \mathbb{M}^{(\mathfrak{d})}): \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ and $\mathfrak{M} = (\mathfrak{M}^{(1)}, \dots, \mathfrak{M}^{(\mathfrak{d})}): \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ such that for all $n \in \mathbb{N}, i \in \{1, 2, \dots, \mathfrak{d}\}$ it holds that

$$\Theta_0 = \xi, \quad \mathbf{m}_0 = 0, \quad \mathbb{M}_0 = 0, \quad \mathfrak{M}_0 = 0, \quad (7.250)$$

$$\mathbf{m}_n = \alpha_n \mathbf{m}_{n-1} + (1 - \alpha_n) \left[\frac{1}{J_n} \sum_{j=1}^{J_n} \varphi(\Theta_{n-1}, X_{n,j}) \right], \quad (7.251)$$

$$\mathbb{M}_n^{(i)} = \beta_n \mathbb{M}_{n-1}^{(i)} + (1 - \beta_n) \left[\frac{1}{J_n} \sum_{j=1}^{J_n} \varphi_i(\Theta_{n-1}, X_{n,j}) \right]^2, \quad (7.252)$$

$$\mathfrak{M}_n^{(i)} = \max\{\mathfrak{M}_{n-1}^{(i)}, \mathbb{M}_n^{(i)}\}, \quad \text{and} \quad (7.253)$$

$$\Theta_n^{(i)} = \Theta_{n-1}^{(i)} - \gamma_n [\varepsilon + \mathfrak{M}_n^{1/2}]^{-1} \mathbf{m}_n^{(i)}. \quad (7.254)$$

Algorithm 7.14.2: AMSgrad SGD optimization method

Input: $\mathbf{d}, \mathbf{d}, N \in \mathbb{N}$, $\ell = (\ell(\theta, x))_{(\theta, x) \in \mathbb{R}^{\mathbf{d}} \times \mathbb{R}^{\mathbf{d}}} \in C^1(\mathbb{R}^{\mathbf{d}} \times \mathbb{R}^{\mathbf{d}}, \mathbb{R})$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, $(\alpha_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $(\beta_n)_{n \in \mathbb{N}} \subseteq [0, 1]$, $\varepsilon \in (0, \infty)$, $(J_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, probability space $(\Omega, \mathcal{F}, \mathbb{P})$, random variable $\xi: \Omega \rightarrow \mathbb{R}^{\mathbf{d}}$, random variables $X_{n,j}: \Omega \rightarrow \mathbb{R}^{\mathbf{d}}$ for $n, j \in \mathbb{N}$

Output: N -th step of the AMSgrad SGD process for the loss function ℓ with learning rates $(\gamma_n)_{n \in \mathbb{N}}$, momentum decay factors $(\alpha_n)_{n \in \mathbb{N}}$, second moment decay factors $(\beta_n)_{n \in \mathbb{N}}$, regularizing factor ε , initial value ξ , batch sizes $(J_n)_{n \in \mathbb{N}}$, and data $(X_{n,j})_{(n,j) \in \mathbb{N}^2}$ (cf. Definition 7.14.1)

```

1: Initialization:  $\Theta \leftarrow \xi; \mathbf{m} \leftarrow 0 \in \mathbb{R}^{\mathbf{d}}; \mathbb{M} \leftarrow 0 \in \mathbb{R}^{\mathbf{d}}; \mathfrak{M} \leftarrow 0 \in \mathbb{R}^{\mathbf{d}}$ 
2: for  $n = 1, \dots, N$  do # (cf. Definition 6.5.2)
3:    $g \leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_{\theta} \ell)(\Theta, X_{n,j})$ 
4:    $\mathbf{m} \leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n) g$ 
5:    $\mathbb{M} \leftarrow \beta_n \mathbb{M} + (1 - \beta_n) g^2$ 
6:    $\mathfrak{M} \leftarrow \max\{\mathfrak{M}, \mathbb{M}\}$ 
7:    $\Theta \leftarrow \Theta - \gamma_n [\varepsilon + \mathfrak{M}^{1/2}]^{-1} \mathbf{m}$ 
8: return  $\Theta$ 

```

Remark 7.14.3 (Analysis of accelerated SGD-type optimization methods). In the literature there are numerous research articles which study the accelerated SGD-type optimization methods reviewed in this chapter. In particular, we refer, for example, to [156, 291, 297, 360, 408] and the references therein for articles on SGD-type optimization methods with momentum and we refer, for instance, to [99, 163, 308, 372, 459] and the references therein for articles on adaptive SGD-type optimization methods.

7.15 Compact summary of SGD optimization methods

In this section we provide an overview over all SGD-type optimization methods considered in this chapter. Roughly speaking, in this summary we provide for each considered method only the iteration step of the respective pseudo-code. The formulas in this summary make use of the componentwise operations in Definition 6.5.2.

SGD optimization method (Definition 7.2.1)

$$\begin{aligned} g &\leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_{\theta} \ell)(\Theta, X_{n,j}) \\ \Theta &\leftarrow \Theta - \gamma_n g \end{aligned}$$

Explicit midpoint SGD optimization method (Definition 7.3.1)

$$\begin{aligned} g_1 &\leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_{\theta} \ell)(\Theta_{n-1}, X_{n,j}) \\ g_2 &\leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_{\theta} \ell)(\Theta_{n-1} - \frac{\gamma_n}{2} g_1, X_{n,j}) \\ \Theta &\leftarrow \Theta - \gamma_n g_2 \end{aligned}$$

Momentum SGD optimization method (Definition 7.4.1)

$$\begin{aligned} g &\leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_{\theta} \ell)(\Theta, X_{n,j}) \\ \mathbf{m} &\leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n) g \\ \Theta &\leftarrow \Theta - \gamma_n \mathbf{m} \end{aligned}$$

Momentum SGD optimization method (2nd version) (Definition 7.4.3)

$$\begin{aligned} g &\leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_{\theta} \ell)(\Theta, X_{n,j}) \\ \mathbf{m} &\leftarrow \alpha_n \mathbf{m} + g \\ \Theta &\leftarrow \Theta - \gamma_n \mathbf{m} \end{aligned}$$

Momentum SGD optimization method (3rd version) (Definition 7.4.5)

$$\begin{aligned} g &\leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_{\theta} \ell)(\Theta, X_{n,j}) \\ \mathbf{m} &\leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n) \gamma_n g \\ \Theta &\leftarrow \Theta - \mathbf{m} \end{aligned}$$

Momentum SGD optimization method (4th version) (Definition 7.4.7)

$$\begin{aligned} g &\leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_{\theta} \ell)(\Theta, X_{n,j}) \\ \mathbf{m} &\leftarrow \alpha_n \mathbf{m} + \gamma_n g \\ \Theta &\leftarrow \Theta - \mathbf{m} \end{aligned}$$

Bias-adjusted momentum SGD optimization method (Definition 7.4.9)

$$\begin{aligned} g &\leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_{\theta} \ell)(\Theta, X_{n,j}) \\ \mathbf{m} &\leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n) g \\ \Theta &\leftarrow \Theta - \frac{\gamma_n \mathbf{m}}{1 - \prod_{l=1}^n \alpha_l} \end{aligned}$$

Nesterov accelerated SGD optimization method (Definition 7.5.1)

$$\begin{aligned} g &\leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_{\theta} \ell)(\Theta - \gamma_n \alpha_n \mathbf{m}, X_{n,j}) \\ \mathbf{m} &\leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n) g \\ \Theta &\leftarrow \Theta - \gamma_n \mathbf{m} \end{aligned}$$

Nesterov accelerated SGD optimization method (2nd version) (Definition 7.5.3)

$$\begin{aligned} g &\leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_{\theta} \ell)(\Theta - \gamma_n \alpha_n \mathbf{m}, X_{n,j}) \\ \mathbf{m} &\leftarrow \alpha_n \mathbf{m} + g \\ \Theta &\leftarrow \Theta - \gamma_n \mathbf{m} \end{aligned}$$

Nesterov accelerated SGD optimization method (3rd version) (Definition 7.5.5)

$$\begin{aligned} g &\leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_{\theta} \ell)(\Theta - \alpha_n \mathbf{m}, X_{n,j}) \\ \mathbf{m} &\leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n) \gamma_n g \\ \Theta &\leftarrow \Theta - \mathbf{m} \end{aligned}$$

Nesterov accelerated SGD optimization method (4th version) (Definition 7.5.7)

$$\begin{aligned} g &\leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_{\theta} \ell)(\Theta - \alpha_n \mathbf{m}, X_{n,j}) \\ \mathbf{m} &\leftarrow \alpha_n \mathbf{m} + \gamma_n g \\ \Theta &\leftarrow \Theta - \mathbf{m} \end{aligned}$$

Shifted Nesterov accelerated SGD optimization method (Definition 7.5.11)

$$\begin{aligned} g &\leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_{\theta} \ell)(\Theta, X_{n,j}) \\ \mathbf{m} &\leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n) g \\ \Theta &\leftarrow \Theta - \gamma_{n+1} \alpha_{n+1} \mathbf{m} - \gamma_n (1 - \alpha_n) g \end{aligned}$$

Shifted Nesterov accelerated SGD optimization method (2nd version) (Definition 7.5.13)

$$\begin{aligned} g &\leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_{\theta} \ell)(\Theta, X_{n,j}) \\ \mathbf{m} &\leftarrow \alpha_n \mathbf{m} + g \\ \Theta &\leftarrow \Theta - \gamma_{n+1} \alpha_{n+1} \mathbf{m} - \gamma_n g \end{aligned}$$

Shifted Nesterov accelerated SGD optimization method (3rd version) (Definition 7.5.15)

$$\begin{aligned} g &\leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_{\theta} \ell)(\Theta, X_{n,j}) \\ \mathbf{m} &\leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n) \gamma_n g \\ \Theta &\leftarrow \Theta - \alpha_{n+1} \mathbf{m} - (1 - \alpha_n) \gamma_n g \end{aligned}$$

Shifted Nesterov accelerated SGD optimization method (4th version) (Definition 7.5.17)

$$\begin{aligned} g &\leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_\theta \ell)(\Theta, X_{n,j}) \\ \mathbf{m} &\leftarrow \alpha_n \mathbf{m} + \gamma_n g \\ \Theta &\leftarrow \Theta - \alpha_{n+1} \mathbf{m} - \gamma_n g \end{aligned}$$

Bias-adjusted Nesterov accelerated SGD optimization method (Definition 7.5.9)

$$\begin{aligned} g &\leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_\theta \ell)(\Theta - \frac{\gamma_n \alpha_n \mathbf{m}}{1 - \prod_{l=1}^n \alpha_l}, X_{n,j}) \\ \mathbf{m} &\leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n)g \\ \Theta &\leftarrow \Theta - \frac{\gamma_n \mathbf{m}}{1 - \prod_{l=1}^n \alpha_l} \end{aligned}$$

Shifted bias-adjusted Nesterov accelerated SGD optimization method (Definition 7.5.19)

$$\begin{aligned} g &\leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_\theta \ell)(\Theta, X_{n,j}) \\ \mathbf{m} &\leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n)g \\ \Theta &\leftarrow \Theta - \frac{\gamma_n (1 - \alpha_n) g}{1 - \prod_{l=1}^n \alpha_l} - \frac{\gamma_{n+1} \alpha_{n+1} \mathbf{m}}{1 - \prod_{l=1}^{n+1} \alpha_l} \end{aligned}$$

Simplified Nesterov accelerated SGD optimization method (Definition 7.5.21)

$$\begin{aligned} g &\leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_\theta \ell)(\Theta, X_{n,j}) \\ \mathbf{m} &\leftarrow \alpha_n \mathbf{m} + g \\ \Theta &\leftarrow \Theta - \gamma_n \alpha_n \mathbf{m} - \gamma_n g \end{aligned}$$

Adagrad SGD optimization method (Definition 7.6.1)

$$\begin{aligned} g &\leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_\theta \ell)(\Theta, X_{n,j}) \\ \mathbb{M} &\leftarrow \mathbb{M} + g^2 \\ \Theta &\leftarrow \Theta - \gamma_n [\varepsilon + \mathbb{M}^{1/2}]^{-1} g \end{aligned}$$

RMSprop SGD optimization method (Definition 7.7.1)

$$\begin{aligned} g &\leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_\theta \ell)(\Theta, X_{n,j}) \\ \mathbb{M} &\leftarrow \beta_n \mathbb{M} + (1 - \beta_n) g^2 \\ \Theta &\leftarrow \Theta - \gamma_n [\varepsilon + \mathbb{M}^{1/2}]^{-1} g \end{aligned}$$

Bias-adjusted RMSprop SGD optimization method (Definition 7.7.3)

$$\begin{aligned} g &\leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_\theta \ell)(\Theta, X_{n,j}) \\ \mathbb{M} &\leftarrow \beta_n \mathbb{M} + (1 - \beta_n) g^2 \\ \Theta &\leftarrow \Theta - \gamma_n \left[\varepsilon + \left[\frac{\mathbb{M}}{1 - \prod_{k=1}^n \beta_k} \right]^{1/2} \right]^{-1} g \end{aligned}$$

Adadelta SGD optimization method (Definition 7.8.1)

$$\begin{aligned} g &\leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_\theta \ell)(\Theta, X_{n,j}) \\ \mathbb{M} &\leftarrow \beta_n \mathbb{M} + (1 - \beta_n) g^2 \\ \Theta &\leftarrow \Theta - \left[\frac{\varepsilon + \Delta}{\varepsilon + \mathbb{M}} \right]^{1/2} g \\ \Delta &\leftarrow \delta_n \Delta + (1 - \delta_n) [\Theta - \Xi]^2 \\ \Xi &\leftarrow \Theta \end{aligned}$$

Adam SGD optimization method (Definition 7.9.1)

$$\begin{aligned} g &\leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_\theta \ell)(\Theta, X_{n,j}) \\ \mathbf{m} &\leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n) g \\ \mathbb{M} &\leftarrow \beta_n \mathbb{M} + (1 - \beta_n) g^2 \\ \Theta &\leftarrow \Theta - \left[\varepsilon + \left[\frac{\mathbb{M}}{1 - \prod_{k=1}^n \beta_k} \right]^{1/2} \right]^{-1} \left[\frac{\gamma_n \mathbf{m}}{1 - \prod_{k=1}^n \alpha_k} \right] \end{aligned}$$

Adamax SGD optimization method (Definition 7.9.4)

$$\begin{aligned} g &\leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_\theta \ell)(\Theta, X_{n,j}) \\ \mathbf{m} &\leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n) g \\ \mathbb{M} &\leftarrow \max\{\beta_n \mathbb{M}, |g|\} \\ \Theta &\leftarrow \Theta - \gamma_n [\varepsilon + \mathbb{M}]^{-1} \left[\frac{\mathbf{m}}{1 - \prod_{k=1}^n \alpha_k} \right] \end{aligned}$$

Nadam SGD optimization method (Definition 7.10.1)

$$\begin{aligned} g &\leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_\theta \ell)(\Theta, X_{n,j}) \\ \mathbf{m} &\leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n) g \\ \mathbb{M} &\leftarrow \beta_n \mathbb{M} + (1 - \beta_n) g^2 \\ \Theta &\leftarrow \Theta - \left[\varepsilon + \left[\frac{\mathbb{M}}{1 - \prod_{k=1}^n \beta_k} \right]^{1/2} \right]^{-1} \left[\left[\frac{\gamma_n(1 - \alpha_n)}{1 - \prod_{k=1}^n \alpha_k} \right] g + \left[\frac{\gamma_{n+1} \alpha_{n+1}}{1 - \prod_{k=1}^{n+1} \alpha_k} \right] \mathbf{m} \right] \end{aligned}$$

Simplified Nadam SGD optimization method (Definition 7.10.3)

$$\begin{aligned} g &\leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_\theta \ell)(\Theta, X_{n,j}) \\ \mathbf{m} &\leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n)g \\ \mathbb{M} &\leftarrow \beta_n \mathbb{M} + (1 - \beta_n)g^2 \\ \Theta &\leftarrow \Theta - \gamma_n \left[\varepsilon + \left[\frac{\mathbb{M}}{1 - \prod_{k=1}^n \beta_k} \right]^{1/2} \right]^{-1} \left[\left[\frac{1 - \alpha_n}{1 - \prod_{k=1}^n \alpha_k} \right] g + \left[\frac{\alpha_{n+1}}{1 - \prod_{k=1}^{n+1} \alpha_k} \right] \mathbf{m} \right] \end{aligned}$$

Nadamax SGD optimization method (Definition 7.10.5)

$$\begin{aligned} g &\leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_\theta \ell)(\Theta, X_{n,j}) \\ \mathbf{m} &\leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n)g \\ \mathbb{M} &\leftarrow \max\{\beta_n \mathbb{M}, |g|\} \\ \Theta &\leftarrow \Theta - [\varepsilon + \mathbb{M}]^{-1} \left[\left[\frac{\gamma_n(1 - \alpha_n)}{1 - \prod_{l=1}^n \alpha_l} \right] g + \left[\frac{\gamma_{n+1}\alpha_{n+1}}{1 - \prod_{l=1}^{n+1} \alpha_l} \right] \mathbf{m} \right] \end{aligned}$$

Adam SGD optimization method with L^2 -regularization (Definition 7.11.3)

$$\begin{aligned} g &\leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_\theta \ell)(\Theta, X_{n,j}) \\ \mathbf{m} &\leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n)(g + \lambda \Theta) \\ \mathbb{M} &\leftarrow \beta_n \mathbb{M} + (1 - \beta_n)[g + \lambda \Theta]^2 \\ \Theta &\leftarrow \Theta - \left[\varepsilon + \left[\frac{\mathbb{M}}{1 - \prod_{k=1}^n \beta_k} \right]^{1/2} \right]^{-1} \left[\frac{\gamma_n \mathbf{m}}{1 - \prod_{k=1}^n \alpha_k} \right] \end{aligned}$$

Shampoo SGD optimization method (Definition 7.12.1)

$$\begin{aligned} g &\leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_\theta \ell)(\Theta, X_{n,j}) \\ \mathbf{L} &\leftarrow \mathbf{L} + gg^* \\ \mathbf{R} &\leftarrow \mathbf{R} + g^*g \\ \Theta &\leftarrow \Theta - \gamma_n (\mathbf{L})^{-1/4} g (\mathbf{R})^{-1/4} \end{aligned}$$

Idealized Muon SGD optimization method (Definition 7.13.1)

$$\begin{aligned} g &\leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_\theta \ell)(\Theta, X_{n,j}) \\ \mathbf{m} &\leftarrow \alpha_n \mathbf{m} + g \\ \Theta &\leftarrow \Theta - \gamma_n \Pi(\mathbf{m}) \end{aligned}$$

Muon SGD optimization method (Definition 7.13.3)

$$\begin{aligned} g &\leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_\theta \ell)(\Theta, X_{n,j}) \\ \mathbf{m} &\leftarrow \alpha_n \mathbf{m} + g \\ \Theta &\leftarrow \Theta - \gamma_n \text{NS}_{a,b,c,\varepsilon}(\mathbf{m}, K) \end{aligned}$$

AdamW SGD optimization method (Definition 7.11.1)

$$\begin{aligned} g &\leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_\theta \ell)(\Theta, X_{n,j}) \\ \mathbf{m} &\leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n) g \\ \mathbb{M} &\leftarrow \beta_n \mathbb{M} + (1 - \beta_n) g^2 \\ \Theta &\leftarrow \Theta - \gamma_n \left(\left[\varepsilon + \left[\frac{\mathbb{M}}{1 - \prod_{k=1}^n \beta_k} \right]^{1/2} \right]^{-1} \left[\frac{\mathbf{m}}{1 - \prod_{k=1}^n \alpha_k} \right] + \lambda \Theta \right) \end{aligned}$$

AMSgrad SGD optimization method (Definition 7.14.1)

$$\begin{aligned} g &\leftarrow \frac{1}{J_n} \sum_{j=1}^{J_n} (\nabla_\theta \ell)(\Theta, X_{n,j}) \\ \mathbf{m} &\leftarrow \alpha_n \mathbf{m} + (1 - \alpha_n) g \\ \mathbb{M} &\leftarrow \beta_n \mathbb{M} + (1 - \beta_n) g^2 \\ \mathfrak{M} &\leftarrow \max\{\mathfrak{M}, \mathbb{M}\} \\ \Theta &\leftarrow \Theta - \gamma_n [\varepsilon + \mathfrak{M}^{1/2}]^{-1} \mathbf{m} \end{aligned}$$

Chapter 8

Backpropagation

In Chapters 6 and 7 we reviewed common deterministic and stochastic **GD**-type optimization methods used for the training of **ANNs**. The specific implementation of such methods requires efficient explicit computations of gradients. The most popular and somehow most natural method to explicitly compute such gradients in the case of the training of **ANNs** is the *backpropagation* method. In this chapter we derive and present this method in detail.

Further material on the backpropagation method can, for example, be found in the books and overview articles [183], [4, Section 11.7], [62, Section 6.2.3], [65, Section 3.2.3], [100, Section 5.6], and [394, Section 20.6].

8.1 Backpropagation for parametric functions

Proposition 8.1.1 (Backpropagation for parametric functions). *Let $L \in \mathbb{N}$, l_0, l_1, \dots, l_L , $\mathfrak{d}_1, \mathfrak{d}_2, \dots, \mathfrak{d}_L \in \mathbb{N}$, for every $k \in \{1, 2, \dots, L\}$ let $F_k = (F_k(\theta_k, x_{k-1}))_{(\theta_k, x_{k-1}) \in \mathbb{R}^{\mathfrak{d}_k} \times \mathbb{R}^{l_{k-1}}} : \mathbb{R}^{\mathfrak{d}_k} \times \mathbb{R}^{l_{k-1}} \rightarrow \mathbb{R}^{l_k}$ be differentiable, for every $k \in \{1, 2, \dots, L\}$ let $f_k = (f_k(\theta_k, \theta_{k+1}, \dots, \theta_L, x_{k-1}))_{(\theta_k, \theta_{k+1}, \dots, \theta_L, x_{k-1}) \in \mathbb{R}^{\mathfrak{d}_k} \times \mathbb{R}^{\mathfrak{d}_{k+1}} \times \dots \times \mathbb{R}^{\mathfrak{d}_L} \times \mathbb{R}^{l_{k-1}}} : \mathbb{R}^{\mathfrak{d}_k} \times \mathbb{R}^{\mathfrak{d}_{k+1}} \times \dots \times \mathbb{R}^{\mathfrak{d}_L} \times \mathbb{R}^{l_{k-1}} \rightarrow \mathbb{R}^{l_L}$ satisfy for all $\theta = (\theta_k, \theta_{k+1}, \dots, \theta_L) \in \mathbb{R}^{\mathfrak{d}_k} \times \mathbb{R}^{\mathfrak{d}_{k+1}} \times \dots \times \mathbb{R}^{\mathfrak{d}_L}$, $x_{k-1} \in \mathbb{R}^{l_{k-1}}$ that*

$$f_k(\theta, x_{k-1}) = (F_L(\theta_L, \cdot) \circ F_{L-1}(\theta_{L-1}, \cdot) \circ \dots \circ F_k(\theta_k, \cdot))(x_{k-1}), \quad (8.1)$$

let $\vartheta = (\vartheta_1, \vartheta_2, \dots, \vartheta_L) \in \mathbb{R}^{\mathfrak{d}_1} \times \mathbb{R}^{\mathfrak{d}_2} \times \dots \times \mathbb{R}^{\mathfrak{d}_L}$, $\mathfrak{x}_0 \in \mathbb{R}^{l_0}$, $\mathfrak{x}_1 \in \mathbb{R}^{l_1}, \dots, \mathfrak{x}_L \in \mathbb{R}^{l_L}$ satisfy for all $k \in \{1, 2, \dots, L\}$ that

$$\mathfrak{x}_k = F_k(\vartheta_k, \mathfrak{x}_{k-1}), \quad (8.2)$$

and let $D_k \in \mathbb{R}^{l_L \times l_{k-1}}$, $k \in \{1, 2, \dots, L+1\}$, satisfy for all $k \in \{1, 2, \dots, L\}$ that

$D_{L+1} = \mathbf{I}_{l_L}$ and

$$D_k = D_{k+1} \left[\left(\frac{\partial F_k}{\partial x_{k-1}} \right) (\vartheta_k, \mathbf{x}_{k-1}) \right] \quad (8.3)$$

(cf. Definition 1.5.5). Then

(i) it holds for all $k \in \{1, 2, \dots, L\}$ that $f_k: \mathbb{R}^{\mathfrak{d}_k} \times \mathbb{R}^{\mathfrak{d}_{k+1}} \times \dots \times \mathbb{R}^{\mathfrak{d}_L} \times \mathbb{R}^{l_{k-1}} \rightarrow \mathbb{R}^{l_L}$ is differentiable,

(ii) it holds for all $k \in \{1, 2, \dots, L\}$ that

$$D_k = \left(\frac{\partial f_k}{\partial x_{k-1}} \right) ((\vartheta_k, \vartheta_{k+1}, \dots, \vartheta_L), \mathbf{x}_{k-1}), \quad (8.4)$$

and

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

$$\left(\frac{\partial f_1}{\partial \theta_k} \right) (\vartheta, \mathbf{x}_0) = D_{k+1} \left[\left(\frac{\partial F_k}{\partial \theta_k} \right) (\vartheta_k, \mathbf{x}_{k-1}) \right]. \quad (8.5)$$

Proof of Proposition 8.1.1. Note that (8.1), the fact that for all $k \in \mathbb{N} \cap (0, L)$, $(\theta_k, \theta_{k+1}, \dots, \theta_L) \in \mathbb{R}^{\mathfrak{d}_k} \times \mathbb{R}^{\mathfrak{d}_{k+1}} \times \dots \times \mathbb{R}^{\mathfrak{d}_L}$, $x_{k-1} \in \mathbb{R}^{l_{k-1}}$ it holds that

$$f_k((\theta_k, \theta_{k+1}, \dots, \theta_L), x_{k-1}) = (f_{k+1}((\theta_{k+1}, \theta_{k+2}, \dots, \theta_L), \cdot) \circ F_k(\theta_k, \cdot))(x_{k-1}), \quad (8.6)$$

the assumption that for all $k \in \{1, 2, \dots, L\}$ it holds that $F_k: \mathbb{R}^{\mathfrak{d}_k} \times \mathbb{R}^{l_{k-1}} \rightarrow \mathbb{R}^{l_k}$ is differentiable, Lemma 5.3.2, and induction imply that for all $k \in \{1, 2, \dots, L\}$ it holds that

$$f_k: \mathbb{R}^{\mathfrak{d}_k} \times \mathbb{R}^{\mathfrak{d}_{k+1}} \times \dots \times \mathbb{R}^{\mathfrak{d}_L} \times \mathbb{R}^{l_{k-1}} \rightarrow \mathbb{R}^{l_L} \quad (8.7)$$

is differentiable. This proves item (i). Next we prove (8.4) by induction on $k \in \{L, L-1, \dots, 1\}$. Note that (8.3), the assumption that $D_{L+1} = \mathbf{I}_{l_L}$, and the fact that $f_L = F_L$ show that

$$D_L = D_{L+1} \left[\left(\frac{\partial F_L}{\partial x_{L-1}} \right) (\vartheta_L, \mathbf{x}_{L-1}) \right] = \left(\frac{\partial f_L}{\partial x_{L-1}} \right) (\vartheta_L, \mathbf{x}_{L-1}). \quad (8.8)$$

This establishes (8.4) in the base case $k = L$. For the induction step note that (8.3), the chain rule, and the fact that for all $k \in \mathbb{N} \cap (0, L)$, $x_{k-1} \in \mathbb{R}^{l_{k-1}}$ it holds that

$$f_k((\vartheta_k, \vartheta_{k+1}, \dots, \vartheta_L), x_{k-1}) = f_{k+1}((\vartheta_{k+1}, \vartheta_{k+2}, \dots, \vartheta_L), F_k(\vartheta_k, x_{k-1})) \quad (8.9)$$

prove that for all $k \in \mathbb{N} \cap (0, L)$ with $D_{k+1} = \left(\frac{\partial f_{k+1}}{\partial x_k} \right) ((\vartheta_{k+1}, \vartheta_{k+2}, \dots, \vartheta_L), \mathfrak{x}_k)$ it holds that

$$\begin{aligned}
 & \left(\frac{\partial f_k}{\partial x_{k-1}} \right) ((\vartheta_k, \vartheta_{k+1}, \dots, \vartheta_L), \mathfrak{x}_{k-1}) \\
 &= (\mathbb{R}^{l_{k-1}} \ni x_{k-1} \mapsto f_k((\vartheta_k, \vartheta_{k+1}, \dots, \vartheta_L), x_{k-1}) \in \mathbb{R}^{l_L})'(\mathfrak{x}_{k-1}) \\
 &= (\mathbb{R}^{l_{k-1}} \ni x_{k-1} \mapsto f_{k+1}((\vartheta_{k+1}, \vartheta_{k+2}, \dots, \vartheta_L), F_k(\vartheta_k, x_{k-1})) \in \mathbb{R}^{l_L})'(\mathfrak{x}_{k-1}) \\
 &= \left[(\mathbb{R}^{l_{k-1}} \ni x_k \mapsto f_{k+1}((\vartheta_{k+1}, \vartheta_{k+2}, \dots, \vartheta_L), x_k)) \in \mathbb{R}^{l_L} \right]'(F_k(\vartheta_k, \mathfrak{x}_{k-1})) \quad (8.10) \\
 &\quad \left[(\mathbb{R}^{l_{k-1}} \ni x_{k-1} \mapsto F_k(\vartheta_k, x_{k-1})) \in \mathbb{R}^{l_k} \right]'(\mathfrak{x}_{k-1}) \\
 &= \left[\left(\frac{\partial f_{k+1}}{\partial x_k} \right) ((\vartheta_{k+1}, \vartheta_{k+2}, \dots, \vartheta_L), \mathfrak{x}_k) \right] \left[\left(\frac{\partial F_k}{\partial x_{k-1}} \right) (\vartheta_k, \mathfrak{x}_{k-1}) \right] \\
 &= D_{k+1} \left[\left(\frac{\partial F_k}{\partial x_{k-1}} \right) (\vartheta_k, \mathfrak{x}_{k-1}) \right] = D_k.
 \end{aligned}$$

Induction thus proves (8.4). This establishes item (ii). Moreover, observe that (8.1) and (8.2) establish that for all $k \in \mathbb{N} \cap (0, L)$, $\theta_k \in \mathbb{R}^{l_k}$ it holds that

$$\begin{aligned}
 & f_1((\vartheta_1, \dots, \vartheta_{k-1}, \theta_k, \vartheta_{k+1}, \dots, \vartheta_L), \mathfrak{x}_0) \\
 &= (F_L(\vartheta_L, \cdot) \circ \dots \circ F_{k+1}(\vartheta_{k+1}, \cdot) \circ F_k(\theta_k, \cdot) \circ F_{k-1}(\vartheta_{k-1}, \cdot) \circ \dots \circ F_1(\vartheta_1, \cdot))(\mathfrak{x}_0) \quad (8.11) \\
 &= (f_{k+1}((\vartheta_{k+1}, \vartheta_{k+2}, \dots, \vartheta_L), F_k(\theta_k, \cdot)))((F_{k-1}(\vartheta_{k-1}, \cdot) \circ \dots \circ F_1(\vartheta_1, \cdot))(\mathfrak{x}_0)) \\
 &= f_{k+1}((\vartheta_{k+1}, \vartheta_{k+2}, \dots, \vartheta_L), F_k(\theta_k, \mathfrak{x}_{k-1})).
 \end{aligned}$$

Combining this with the chain rule, (8.2), and (8.4) demonstrates that for all $k \in \mathbb{N} \cap (0, L)$ it holds that

$$\begin{aligned}
 \left(\frac{\partial f_1}{\partial \theta_k} \right) (\vartheta, \mathfrak{x}_0) &= (\mathbb{R}^{n_k} \ni \theta_k \mapsto f_{k+1}((\vartheta_{k+1}, \vartheta_{k+2}, \dots, \vartheta_L), F_k(\theta_k, \mathfrak{x}_{k-1})) \in \mathbb{R}^{l_L})'(\vartheta_k) \\
 &= \left[(\mathbb{R}^{l_k} \ni x_k \mapsto f_{k+1}((\vartheta_{k+1}, \vartheta_{k+2}, \dots, \vartheta_L), x_k)) \in \mathbb{R}^{l_L} \right]'(F_k(\vartheta_k, \mathfrak{x}_{k-1})) \\
 &\quad \left[(\mathbb{R}^{n_k} \ni \theta_k \mapsto F_k(\theta_k, \mathfrak{x}_{k-1})) \in \mathbb{R}^{l_k} \right]'(\vartheta_k) \quad (8.12) \\
 &= \left[\left(\frac{\partial f_{k+1}}{\partial x_k} \right) ((\vartheta_{k+1}, \vartheta_{k+2}, \dots, \vartheta_L), \mathfrak{x}_k) \right] \left[\left(\frac{\partial F_k}{\partial \theta_k} \right) (\vartheta_k, \mathfrak{x}_{k-1}) \right] \\
 &= D_{k+1} \left[\left(\frac{\partial F_k}{\partial \theta_k} \right) (\vartheta_k, \mathfrak{x}_{k-1}) \right]. \quad 459
 \end{aligned}$$

Furthermore, observe that (8.1) and the fact that $D_{L+1} = \mathbf{I}_{l_L}$ ensure that

$$\begin{aligned} \left(\frac{\partial f_1}{\partial \theta_L} \right)(\vartheta, \mathfrak{x}_0) &= (\mathbb{R}^{n_L} \ni \theta_L \mapsto F_L(\theta_L, \mathfrak{x}_{L-1})) \in \mathbb{R}^{l_L}'(\vartheta_L) \\ &= \left[\left(\frac{\partial F_L}{\partial \theta_L} \right)(\vartheta_L, \mathfrak{x}_{L-1}) \right] \\ &= D_{L+1} \left[\left(\frac{\partial F_L}{\partial \theta_L} \right)(\vartheta_L, \mathfrak{x}_{L-1}) \right]. \end{aligned} \quad (8.13)$$

Combining this and (8.12) establishes item (iii). The proof of Proposition 8.1.1 is thus complete. \square

Corollary 8.1.2 (Backpropagation for parametric functions with loss). *Let $L \in \mathbb{N}$, $l_0, l_1, \dots, l_L, \mathfrak{d}_1, \mathfrak{d}_2, \dots, \mathfrak{d}_L \in \mathbb{N}$, $\vartheta = (\vartheta_1, \vartheta_2, \dots, \vartheta_L) \in \mathbb{R}^{\mathfrak{d}_1} \times \mathbb{R}^{\mathfrak{d}_2} \times \dots \times \mathbb{R}^{\mathfrak{d}_L}$, $\mathfrak{x}_0 \in \mathbb{R}^{l_0}$, $\mathfrak{x}_1 \in \mathbb{R}^{l_1}, \dots, \mathfrak{x}_L \in \mathbb{R}^{l_L}$, $\mathfrak{y} \in \mathbb{R}^{l_L}$, let $\mathfrak{C} = (\mathfrak{C}(x, y))_{(x, y) \in \mathbb{R}^{l_L} \times \mathbb{R}^{l_L}} : \mathbb{R}^{l_L} \times \mathbb{R}^{l_L} \rightarrow \mathbb{R}$ be differentiable, for every $k \in \{1, 2, \dots, L\}$ let $F_k = (F_k(\theta_k, x_{k-1}))_{(\theta_k, x_{k-1}) \in \mathbb{R}^{\mathfrak{d}_k} \times \mathbb{R}^{l_{k-1}}} : \mathbb{R}^{\mathfrak{d}_k} \times \mathbb{R}^{l_{k-1}} \rightarrow \mathbb{R}^{l_k}$ be differentiable, let $\mathcal{L} = (\mathcal{L}(\theta_1, \theta_2, \dots, \theta_L))_{(\theta_1, \theta_2, \dots, \theta_L) \in \mathbb{R}^{\mathfrak{d}_1} \times \mathbb{R}^{\mathfrak{d}_2} \times \dots \times \mathbb{R}^{\mathfrak{d}_L}} : \mathbb{R}^{\mathfrak{d}_1} \times \mathbb{R}^{\mathfrak{d}_2} \times \dots \times \mathbb{R}^{\mathfrak{d}_L} \rightarrow \mathbb{R}$ satisfy for all $\theta = (\theta_1, \theta_2, \dots, \theta_L) \in \mathbb{R}^{\mathfrak{d}_1} \times \mathbb{R}^{\mathfrak{d}_2} \times \dots \times \mathbb{R}^{\mathfrak{d}_L}$ that*

$$\mathcal{L}(\theta) = (\mathfrak{C}(\cdot, \mathfrak{y}) \circ F_L(\theta_L, \cdot) \circ F_{L-1}(\theta_{L-1}, \cdot) \circ \dots \circ F_1(\theta_1, \cdot))(\mathfrak{x}_0), \quad (8.14)$$

assume for all $k \in \{1, 2, \dots, L\}$ that

$$\mathfrak{x}_k = F_k(\vartheta_k, \mathfrak{x}_{k-1}), \quad (8.15)$$

and let $D_k \in \mathbb{R}^{l_{k-1}}$, $k \in \{1, 2, \dots, L+1\}$, satisfy for all $k \in \{1, 2, \dots, L\}$ that

$$D_{L+1} = (\nabla_x \mathfrak{C})(\mathfrak{x}_L, \mathfrak{y}) \quad \text{and} \quad D_k = \left[\left(\frac{\partial F_k}{\partial x_{k-1}} \right)(\vartheta_k, \mathfrak{x}_{k-1}) \right]^* D_{k+1}. \quad (8.16)$$

Then

- (i) it holds that $\mathcal{L} : \mathbb{R}^{\mathfrak{d}_1} \times \mathbb{R}^{\mathfrak{d}_2} \times \dots \times \mathbb{R}^{\mathfrak{d}_L} \rightarrow \mathbb{R}$ is differentiable and
- (ii) it holds for all $k \in \{1, 2, \dots, L\}$ that

$$(\nabla_{\theta_k} \mathcal{L})(\vartheta) = \left[\left(\frac{\partial F_k}{\partial \theta_k} \right)(\vartheta_k, \mathfrak{x}_{k-1}) \right]^* D_{k+1}. \quad (8.17)$$

Proof of Corollary 8.1.2. Throughout this proof, let $\mathbf{D}_k \in \mathbb{R}^{l_L \times l_{k-1}}$, $k \in \{1, 2, \dots, L+1\}$,

satisfy for all $k \in \{1, 2, \dots, L\}$ that $\mathbf{D}_{L+1} = \mathbf{I}_{l_L}$ and

$$\mathbf{D}_k = \mathbf{D}_{k+1} \left[\left(\frac{\partial F_k}{\partial x_{k-1}} \right) (\vartheta_k, \mathfrak{x}_{k-1}) \right] \quad (8.18)$$

and let $f = (f(\theta_1, \theta_2, \dots, \theta_L))_{(\theta_1, \theta_2, \dots, \theta_L) \in \mathbb{R}^{\mathfrak{d}_1} \times \mathbb{R}^{\mathfrak{d}_2} \times \dots \times \mathbb{R}^{\mathfrak{d}_L}} : \mathbb{R}^{\mathfrak{d}_1} \times \mathbb{R}^{\mathfrak{d}_2} \times \dots \times \mathbb{R}^{\mathfrak{d}_L} \rightarrow \mathbb{R}^{l_L}$ satisfy for all $\theta = (\theta_1, \theta_2, \dots, \theta_L) \in \mathbb{R}^{\mathfrak{d}_1} \times \mathbb{R}^{\mathfrak{d}_2} \times \dots \times \mathbb{R}^{\mathfrak{d}_L}$ that

$$f(\theta) = (F_L(\theta_L, \cdot) \circ F_{L-1}(\theta_{L-1}, \cdot) \circ \dots \circ F_1(\theta_1, \cdot))(\mathfrak{x}_0) \quad (8.19)$$

(cf. Definition 1.5.5). Note that item (i) in Proposition 8.1.1 ensures that $f : \mathbb{R}^{\mathfrak{d}_1} \times \mathbb{R}^{\mathfrak{d}_2} \times \dots \times \mathbb{R}^{\mathfrak{d}_L} \rightarrow \mathbb{R}^{l_L}$ is differentiable. This, the assumption that $\mathfrak{C} : \mathbb{R}^{l_L} \times \mathbb{R}^{l_L} \rightarrow \mathbb{R}$ is differentiable, and the fact that $\mathcal{L} = \mathfrak{C}(\cdot, \mathfrak{y}) \circ f$ show that $\mathcal{L} : \mathbb{R}^{\mathfrak{d}_1} \times \mathbb{R}^{\mathfrak{d}_2} \times \dots \times \mathbb{R}^{\mathfrak{d}_L} \rightarrow \mathbb{R}$ is differentiable. This establishes item (i). Next we claim that for all $k \in \{1, 2, \dots, L+1\}$ it holds that

$$[D_k]^* = \left[\left(\frac{\partial \mathfrak{C}}{\partial x} \right) (\mathfrak{x}_L, \mathfrak{y}) \right] \mathbf{D}_k. \quad (8.20)$$

We now prove (8.20) by induction on $k \in \{L+1, L, \dots, 1\}$. For the base case $k = L+1$ note that (8.16) and (8.18) establish that

$$\begin{aligned} [D_{L+1}]^* &= [(\nabla_x \mathfrak{C})(\mathfrak{x}_L, \mathfrak{y})]^* = \left(\frac{\partial \mathfrak{C}}{\partial x} \right) (\mathfrak{x}_L, \mathfrak{y}) \\ &= \left[\left(\frac{\partial \mathfrak{C}}{\partial x} \right) (\mathfrak{x}_L, \mathfrak{y}) \right] \mathbf{I}_{l_L} = \left[\left(\frac{\partial \mathfrak{C}}{\partial x} \right) (\mathfrak{x}_L, \mathfrak{y}) \right] \mathbf{D}_{L+1}. \end{aligned} \quad (8.21)$$

This establishes (8.20) in the base case $k = L+1$. For the induction step observe (8.16) and (8.18) demonstrate that for all $k \in \{L, L-1, \dots, 1\}$ with $[D_{k+1}]^* = \left[\left(\frac{\partial \mathfrak{C}}{\partial x} \right) (\mathfrak{x}_L, \mathfrak{y}) \right] \mathbf{D}_{k+1}$ it holds that

$$\begin{aligned} [D_k]^* &= [D_{k+1}]^* \left[\left(\frac{\partial F_k}{\partial x_{k-1}} \right) (\vartheta_k, \mathfrak{x}_{k-1}) \right] \\ &= \left[\left(\frac{\partial \mathfrak{C}}{\partial x} \right) (\mathfrak{x}_L, \mathfrak{y}) \right] \mathbf{D}_{k+1} \left[\left(\frac{\partial F_k}{\partial x_{k-1}} \right) (\vartheta_k, \mathfrak{x}_{k-1}) \right] = \left[\left(\frac{\partial \mathfrak{C}}{\partial x} \right) (\mathfrak{x}_L, \mathfrak{y}) \right] \mathbf{D}_k. \end{aligned} \quad (8.22)$$

Induction thus establishes (8.20). Furthermore, note that item (iii) in Proposition 8.1.1 shows that for all $k \in \{1, 2, \dots, L\}$ it holds that

$$\left(\frac{\partial f}{\partial \theta_k} \right) (\vartheta) = \mathbf{D}_{k+1} \left[\left(\frac{\partial F_k}{\partial \theta_k} \right) (\vartheta_k, \mathfrak{x}_{k-1}) \right]. \quad (8.23)$$

Combining this with chain rule, the fact that $\mathcal{L} = \mathfrak{C}(\cdot, \mathfrak{y}) \circ f$, and (8.20) ensures that for all $k \in \{1, 2, \dots, L\}$ it holds that

$$\begin{aligned}\left(\frac{\partial \mathcal{L}}{\partial \theta_k}\right)(\vartheta) &= \left[\left(\frac{\partial \mathfrak{C}}{\partial x}\right)(f(\vartheta), \mathfrak{y})\right] \left[\left(\frac{\partial f}{\partial \theta_k}\right)(\vartheta)\right] \\ &= \left[\left(\frac{\partial \mathfrak{C}}{\partial x}\right)(\mathfrak{x}_L, \mathfrak{y})\right] \mathbf{D}_{k+1} \left[\left(\frac{\partial F_k}{\partial \theta_k}\right)(\vartheta_k, \mathfrak{x}_{k-1})\right] \\ &= [D_{k+1}]^* \left[\left(\frac{\partial F_k}{\partial \theta_k}\right)(\vartheta_k, \mathfrak{x}_{k-1})\right].\end{aligned}\quad (8.24)$$

Hence, we obtain that for all $k \in \{1, 2, \dots, L\}$ it holds that

$$(\nabla_{\theta_k} \mathcal{L})(\vartheta) = \left[\left(\frac{\partial \mathcal{L}}{\partial \theta_k}\right)(\vartheta)\right]^* = \left[\left(\frac{\partial F_k}{\partial \theta_k}\right)(\vartheta_k, \mathfrak{x}_{k-1})\right]^* D_{k+1}. \quad (8.25)$$

This establishes item (ii). The proof of Corollary 8.1.2 is thus complete. \square

8.2 Backpropagation for ANNs

Definition 8.2.1 (Diagonal matrices). We denote by $\text{diag}: (\bigcup_{d \in \mathbb{N}} \mathbb{R}^d) \rightarrow (\bigcup_{d \in \mathbb{N}} \mathbb{R}^{d \times d})$ the function which satisfies for all $d \in \mathbb{N}$, $x = (x_1, \dots, x_d) \in \mathbb{R}^d$ that

$$\text{diag}(x) = \begin{pmatrix} x_1 & 0 & \cdots & 0 \\ 0 & x_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & x_d \end{pmatrix} \in \mathbb{R}^{d \times d}. \quad (8.26)$$

Corollary 8.2.2 (Backpropagation for ANNs). Let $L \in \mathbb{N}$, $l_0, l_1, \dots, l_L \in \mathbb{N}$, $\Phi = ((\mathbf{W}_1, \mathbf{B}_1), \dots, (\mathbf{W}_L, \mathbf{B}_L)) \in \bigtimes_{k=1}^L (\mathbb{R}^{l_k \times l_{k-1}} \times \mathbb{R}^{l_k})$, let $\mathfrak{C} = (\mathfrak{C}(x, y))_{(x, y) \in \mathbb{R}^{l_L} \times \mathbb{R}^{l_L}}: \mathbb{R}^{l_L} \times \mathbb{R}^{l_L} \rightarrow \mathbb{R}$ and $a: \mathbb{R} \rightarrow \mathbb{R}$ be differentiable, let $\mathfrak{x}_0 \in \mathbb{R}^{l_0}$, $\mathfrak{x}_1 \in \mathbb{R}^{l_1}, \dots, \mathfrak{x}_L \in \mathbb{R}^{l_L}$, $\mathfrak{y} \in \mathbb{R}^{l_L}$ satisfy for all $k \in \{1, 2, \dots, L\}$ that

$$\mathfrak{x}_k = \mathfrak{M}_{a \mathbb{1}_{[0, L]}(k) + \text{id}_{\mathbb{R}} \mathbb{1}_{\{L\}}(k), l_k}(\mathbf{W}_k \mathfrak{x}_{k-1} + \mathbf{B}_k), \quad (8.27)$$

let $\mathcal{L} = (\mathcal{L}((W_1, B_1), \dots, (W_L, B_L)))_{((W_1, B_1), \dots, (W_L, B_L)) \in \bigtimes_{k=1}^L (\mathbb{R}^{l_k \times l_{k-1}} \times \mathbb{R}^{l_k})} : \bigtimes_{k=1}^L (\mathbb{R}^{l_k \times l_{k-1}} \times \mathbb{R}^{l_k}) \rightarrow \mathbb{R}$ satisfy for all $\Psi \in \bigtimes_{k=1}^L (\mathbb{R}^{l_k \times l_{k-1}} \times \mathbb{R}^{l_k})$ that

$$\mathcal{L}(\Psi) = \mathfrak{C}((\mathcal{R}_a^N(\Psi))(\mathfrak{x}_0), \mathfrak{y}), \quad (8.28)$$

and let $D_k \in \mathbb{R}^{l_{k-1}}$, $k \in \{1, 2, \dots, L+1\}$, satisfy for all $k \in \{1, 2, \dots, L-1\}$ that

$$D_{L+1} = (\nabla_x \mathfrak{C})(\mathfrak{x}_L, \mathfrak{y}), \quad D_L = [\mathbf{W}_L]^* D_{L+1}, \quad \text{and} \quad (8.29)$$

$$D_k = [\mathbf{W}_k]^* [\text{diag}(\mathfrak{M}_{a', l_k}(\mathbf{W}_k \mathfrak{x}_{k-1} + \mathbf{B}_k))] D_{k+1} \quad (8.30)$$

(cf. Definitions 1.2.1, 1.3.4, and 8.2.1). Then

(i) it holds that $\mathcal{L}: \bigtimes_{k=1}^L (\mathbb{R}^{l_k \times l_{k-1}} \times \mathbb{R}^{l_k}) \rightarrow \mathbb{R}$ is differentiable,

(ii) it holds that $(\nabla_{B_L} \mathcal{L})(\Phi) = D_{L+1}$,

(iii) it holds for all $k \in \{1, 2, \dots, L-1\}$ that

$$(\nabla_{B_k} \mathcal{L})(\Phi) = [\text{diag}(\mathfrak{M}_{a', l_k}(\mathbf{W}_k \mathfrak{x}_{k-1} + \mathbf{B}_k))] D_{k+1}, \quad (8.31)$$

(iv) it holds that $(\nabla_{W_L} \mathcal{L})(\Phi) = D_{L+1}[\mathfrak{x}_{L-1}]^*$, and

(v) it holds for all $k \in \{1, 2, \dots, L-1\}$ that

$$(\nabla_{W_k} \mathcal{L})(\Phi) = [\text{diag}(\mathfrak{M}_{a', l_k}(\mathbf{W}_k \mathfrak{x}_{k-1} + \mathbf{B}_k))] D_{k+1}[\mathfrak{x}_{k-1}]^*. \quad (8.32)$$

Proof of Corollary 8.2.2. Throughout this proof, for every $k \in \{1, 2, \dots, L\}$ let

$$\begin{aligned} F_k &= (F_k^{(m)})_{m \in \{1, 2, \dots, l_k\}} \\ &= (F_k(((W_{k,i,j})_{(i,j) \in \{1, 2, \dots, l_k\} \times \{1, 2, \dots, l_{k-1}\}}, B_k), \\ &\quad x_{k-1}))_{((W_{k,i,j})_{(i,j) \in \{1, 2, \dots, l_k\} \times \{1, 2, \dots, l_{k-1}\}}, B_k), x_{k-1}) \in (\mathbb{R}^{l_k \times l_{k-1}} \times \mathbb{R}^{l_{k-1}}) \times \mathbb{R}^{l_{k-1}}} \\ &: (\mathbb{R}^{l_k \times l_{k-1}} \times \mathbb{R}^{l_{k-1}}) \times \mathbb{R}^{l_{k-1}} \rightarrow \mathbb{R}^{l_k} \end{aligned} \quad (8.33)$$

satisfy for all $(W_k, B_k) \in \mathbb{R}^{l_k \times l_{k-1}} \times \mathbb{R}^{l_{k-1}}$, $x_{k-1} \in \mathbb{R}^{l_{k-1}}$ that

$$F_k((W_k, B_k), x_{k-1}) = \mathfrak{M}_{a \mathbb{1}_{[0,L]}(k) + \text{id}_{\mathbb{R}} \mathbb{1}_{\{L\}}(k), l_k}(W_k x_{k-1} + B_k) \quad (8.34)$$

and for every $d \in \mathbb{N}$ let $\mathbf{e}_1^{(d)}, \mathbf{e}_2^{(d)}, \dots, \mathbf{e}_d^{(d)} \in \mathbb{R}^d$ satisfy $\mathbf{e}_1^{(d)} = (1, 0, \dots, 0)$, $\mathbf{e}_2^{(d)} = (0, 1, 0, \dots, 0)$, \dots , $\mathbf{e}_d^{(d)} = (0, \dots, 0, 1)$. Observe that the assumption that a is differentiable and (8.27) imply that $\mathcal{L}: \bigtimes_{k=1}^L (\mathbb{R}^{l_k \times l_{k-1}} \times \mathbb{R}^{l_k}) \rightarrow \mathbb{R}$ is differentiable. This establishes item (i). Next note that (1.97), (8.28), and (8.34) ensure that for all $\Psi = ((W_1, B_1), \dots, (W_L, B_L)) \in \bigtimes_{k=1}^L (\mathbb{R}^{l_k \times l_{k-1}} \times \mathbb{R}^{l_k})$ it holds that

$$\mathcal{L}(\Psi) = (\mathfrak{C}(\cdot, \mathfrak{y}) \circ F_L((W_L, B_L), \cdot) \circ F_{L-1}((W_{L-1}, B_{L-1}), \cdot) \circ \dots \circ F_1((W_1, B_1), \cdot))(\mathfrak{x}_0). \quad (8.35)$$

Moreover, observe that (8.27) and (8.34) imply that for all $k \in \{1, 2, \dots, L\}$ it holds that

$$\mathfrak{x}_k = F_k((\mathbf{W}_k, \mathbf{B}_k), \mathfrak{x}_{k-1}). \quad (8.36)$$

In addition, observe that (8.34) establishes that

$$\left(\frac{\partial F_L}{\partial x_{L-1}} \right) ((\mathbf{W}_L, \mathbf{B}_L), \mathfrak{x}_{L-1}) = \mathbf{W}_L. \quad (8.37)$$

Moreover, note that (8.34) proves that for all $k \in \{1, 2, \dots, L-1\}$ it holds that

$$\left(\frac{\partial F_k}{\partial x_{k-1}} \right) ((\mathbf{W}_k, \mathbf{B}_k), \mathfrak{x}_{k-1}) = [\text{diag}(\mathfrak{M}_{a',l_k}(\mathbf{W}_k \mathfrak{x}_{k-1} + \mathbf{B}_k))] \mathbf{W}_k. \quad (8.38)$$

Combining this and (8.37) with (8.29) and (8.30) demonstrates that for all $k \in \{1, 2, \dots, L\}$ it holds that

$$D_{L+1} = (\nabla_x \mathfrak{C})(\mathfrak{x}_L, \mathfrak{y}) \quad \text{and} \quad D_k = \left[\left(\frac{\partial F_k}{\partial x_{k-1}} \right) (\vartheta_k, \mathfrak{x}_{k-1}) \right]^* D_{k+1}. \quad (8.39)$$

Next note that this, (8.35), (8.36), and Corollary 8.1.2 establish that for all $k \in \{1, 2, \dots, L\}$ it holds that

$$(\nabla_{B_k} \mathcal{L})(\Phi) = \left[\left(\frac{\partial F_k}{\partial B_k} \right) ((\mathbf{W}_k, \mathbf{B}_k), \mathfrak{x}_{k-1}) \right]^* D_{k+1} \quad \text{and} \quad (8.40)$$

$$(\nabla_{W_k} \mathcal{L})(\Phi) = \left[\left(\frac{\partial F_k}{\partial W_k} \right) ((\mathbf{W}_k, \mathbf{B}_k), \mathfrak{x}_{k-1}) \right]^* D_{k+1}. \quad (8.41)$$

Moreover, observe that (8.34) implies that

$$\left(\frac{\partial F_L}{\partial B_L} \right) ((\mathbf{W}_L, \mathbf{B}_L), \mathfrak{x}_{L-1}) = \mathbf{I}_{l_L} \quad (8.42)$$

(cf. Definition 1.5.5). Combining this with (8.40) demonstrates that

$$(\nabla_{B_L} \mathcal{L})(\Phi) = [\mathbf{I}_{l_L}]^* D_{L+1} = D_{L+1}. \quad (8.43)$$

This establishes item (ii). Furthermore, note that (8.34) shows that for all $k \in \{1, 2, \dots, L-1\}$ it holds that

$$\left(\frac{\partial F_k}{\partial B_k} \right) ((\mathbf{W}_k, \mathbf{B}_k), \mathfrak{x}_{k-1}) = \text{diag}(\mathfrak{M}_{a',l_k}(\mathbf{W}_k \mathfrak{x}_{k-1} + \mathbf{B}_k)). \quad (8.44)$$

Combining this with (8.40) implies that for all $k \in \{1, 2, \dots, L-1\}$ it holds that

$$\begin{aligned} (\nabla_{B_k} \mathcal{L})(\Phi) &= [\text{diag}(\mathfrak{M}_{a',l_k}(\mathbf{W}_k \mathfrak{x}_{k-1} + \mathbf{B}_k))]^* D_{k+1} \\ &= [\text{diag}(\mathfrak{M}_{a',l_k}(\mathbf{W}_k \mathfrak{x}_{k-1} + \mathbf{B}_k))] D_{k+1}. \end{aligned} \quad (8.45)$$

This establishes item (iii). In addition, observe that (8.34) ensures that for all $m, i \in \{1, 2, \dots, l_L\}$, $j \in \{1, 2, \dots, l_{L-1}\}$ it holds that

$$\left(\frac{\partial F_L^{(m)}}{\partial W_{L,i,j}} \right) ((\mathbf{W}_L, \mathbf{B}_L), \mathfrak{x}_{L-1}) = \mathbb{1}_{\{m\}}(i) \langle \mathfrak{x}_{L-1}, \mathbf{e}_j^{(l_{L-1})} \rangle \quad (8.46)$$

(cf. Definition 1.4.7). Combining this with (8.41) demonstrates that

$$\begin{aligned} &(\nabla_{W_L} \mathcal{L})(\Phi) \\ &= \left(\sum_{m=1}^{l_L} \left[\left(\frac{\partial F_L^{(m)}}{\partial W_{L,i,j}} \right) ((\mathbf{W}_L, \mathbf{B}_L), \mathfrak{x}_{L-1}) \right] \langle D_{L+1}, \mathbf{e}_m^{(l_L)} \rangle \right)_{(i,j) \in \{1, 2, \dots, l_L\} \times \{1, 2, \dots, l_{L-1}\}} \\ &= \left(\sum_{m=1}^{l_L} \mathbb{1}_{\{m\}}(i) \langle \mathbf{e}_j^{(l_{L-1})}, \mathfrak{x}_{L-1} \rangle \langle \mathbf{e}_m^{(l_L)}, D_{L+1} \rangle \right)_{(i,j) \in \{1, 2, \dots, l_L\} \times \{1, 2, \dots, l_{L-1}\}} \\ &= \left(\langle \mathbf{e}_j^{(l_{L-1})}, \mathfrak{x}_{L-1} \rangle \langle \mathbf{e}_i^{(l_L)}, D_{L+1} \rangle \right)_{(i,j) \in \{1, 2, \dots, l_L\} \times \{1, 2, \dots, l_{L-1}\}} \\ &= D_{L+1}[\mathfrak{x}_{L-1}]^*. \end{aligned} \quad (8.47)$$

This establishes item (iv). Moreover, note that (8.34) proves that for all $k \in \{1, 2, \dots, L-1\}$, $m, i \in \{1, 2, \dots, l_k\}$, $j \in \{1, 2, \dots, l_{k-1}\}$ it holds that

$$\left(\frac{\partial F_k^{(m)}}{\partial W_{k,i,j}} \right) ((\mathbf{W}_k, \mathbf{B}_k), \mathfrak{x}_{k-1}) = \mathbb{1}_{\{m\}}(i) a'(\langle \mathbf{e}_i^{(l_k)}, \mathbf{W}_k \mathfrak{x}_{k-1} + \mathbf{B}_k \rangle) \langle \mathbf{e}_j^{(l_{k-1})}, \mathfrak{x}_{k-1} \rangle. \quad (8.48)$$

Combining this with (8.41) demonstrates that for all $k \in \{1, 2, \dots, L-1\}$ it holds that

$$\begin{aligned} &(\nabla_{W_k} \mathcal{L})(\Phi) \\ &= \left(\sum_{m=1}^{l_k} \left[\left(\frac{\partial F_k^{(m)}}{\partial W_{k,i,j}} \right) ((\mathbf{W}_k, \mathbf{B}_k), \mathfrak{x}_{k-1}) \right] \langle \mathbf{e}_m^{(l_k)}, D_{k+1} \rangle \right)_{(i,j) \in \{1, 2, \dots, l_k\} \times \{1, 2, \dots, l_{k-1}\}} \\ &= \left(\sum_{m=1}^{l_k} \mathbb{1}_{\{m\}}(i) a'(\langle \mathbf{e}_i^{(l_k)}, \mathbf{W}_k \mathfrak{x}_{k-1} + \mathbf{B}_k \rangle) \langle \mathbf{e}_j^{(l_{k-1})}, \mathfrak{x}_{k-1} \rangle \langle \mathbf{e}_m^{(l_k)}, D_{k+1} \rangle \right)_{(i,j) \in \{1, 2, \dots, l_k\} \times \{1, 2, \dots, l_{k-1}\}} \\ &= \left(a'(\langle \mathbf{e}_i^{(l_k)}, \mathbf{W}_k \mathfrak{x}_{k-1} + \mathbf{B}_k \rangle) \langle \mathbf{e}_j^{(l_{k-1})}, \mathfrak{x}_{k-1} \rangle \langle \mathbf{e}_i^{(l_k)}, D_{k+1} \rangle \right)_{(i,j) \in \{1, 2, \dots, l_k\} \times \{1, 2, \dots, l_{k-1}\}} \\ &= [\text{diag}(\mathfrak{M}_{a',l_k}(\mathbf{W}_k \mathfrak{x}_{k-1} + \mathbf{B}_k))] D_{k+1}[\mathfrak{x}_{k-1}]^*. \end{aligned} \quad (8.49)$$

This establishes item (v). The proof of Corollary 8.2.2 is thus complete. \square

Corollary 8.2.3 (Backpropagation for ANNs with minibatches). *Let $L, M \in \mathbb{N}$, $l_0, l_1, \dots, l_L \in \mathbb{N}$, $\Phi = ((\mathbf{W}_1, \mathbf{B}_1), \dots, (\mathbf{W}_L, \mathbf{B}_L)) \in \bigtimes_{k=1}^L (\mathbb{R}^{l_k \times l_{k-1}} \times \mathbb{R}^{l_k})$, let $a: \mathbb{R} \rightarrow \mathbb{R}$ and $\mathfrak{C} = (\mathfrak{C}(x, y))_{(x, y) \in \mathbb{R}^{l_L} \times \mathbb{R}^{l_L}}: \mathbb{R}^{l_L} \times \mathbb{R}^{l_L} \rightarrow \mathbb{R}$ be differentiable, for every $m \in \{1, 2, \dots, M\}$ let $\mathfrak{x}_0^{(m)} \in \mathbb{R}^{l_0}$, $\mathfrak{x}_1^{(m)} \in \mathbb{R}^{l_1}$, \dots , $\mathfrak{x}_L^{(m)} \in \mathbb{R}^{l_L}$, $\mathfrak{y}^{(m)} \in \mathbb{R}^{l_L}$ satisfy for all $k \in \{1, 2, \dots, L\}$ that*

$$\mathfrak{x}_k^{(m)} = \mathfrak{M}_{a \mathbb{1}_{[0, L)}(k) + \text{id}_{\mathbb{R}} \mathbb{1}_{\{L\}}(k), l_k}(\mathbf{W}_k \mathfrak{x}_{k-1}^{(m)} + \mathbf{B}_k), \quad (8.50)$$

let $\mathcal{L} = (\mathcal{L}((W_1, B_1), \dots, (W_L, B_L)))_{((W_1, B_1), \dots, (W_L, B_L)) \in \bigtimes_{k=1}^L (\mathbb{R}^{l_k \times l_{k-1}} \times \mathbb{R}^{l_k})}: \bigtimes_{k=1}^L (\mathbb{R}^{l_k \times l_{k-1}} \times \mathbb{R}^{l_k}) \rightarrow \mathbb{R}$ satisfy for all $\Psi \in \bigtimes_{k=1}^L (\mathbb{R}^{l_k \times l_{k-1}} \times \mathbb{R}^{l_k})$ that

$$\mathcal{L}(\Psi) = \frac{1}{M} \left[\sum_{m=1}^M \mathfrak{C}((\mathcal{R}_a^{\mathbf{N}}(\Psi))(\mathfrak{x}_0^{(m)}), \mathfrak{y}^{(m)}) \right], \quad (8.51)$$

and for every $m \in \{1, 2, \dots, M\}$ let $D_k^{(m)} \in \mathbb{R}^{l_{k-1}}$, $k \in \{1, 2, \dots, L+1\}$, satisfy for all $k \in \{1, 2, \dots, L-1\}$ that

$$D_{L+1}^{(m)} = (\nabla_x \mathfrak{C})(\mathfrak{x}_L^{(m)}, \mathfrak{y}^{(m)}), \quad D_L^{(m)} = [\mathbf{W}_L]^* D_{L+1}^{(m)}, \quad \text{and} \quad (8.52)$$

$$D_k^{(m)} = [\mathbf{W}_k]^* [\text{diag}(\mathfrak{M}_{a', l_k}(\mathbf{W}_k \mathfrak{x}_{k-1}^{(m)} + \mathbf{B}_k))] D_{k+1}^{(m)} \quad (8.53)$$

(cf. Definitions 1.2.1, 1.3.4, and 8.2.1). Then

(i) it holds that $\mathcal{L}: \bigtimes_{k=1}^L (\mathbb{R}^{l_k \times l_{k-1}} \times \mathbb{R}^{l_k}) \rightarrow \mathbb{R}$ is differentiable,

(ii) it holds that $(\nabla_{B_L} \mathcal{L})(\Phi) = \frac{1}{M} [\sum_{m=1}^M D_{L+1}^{(m)}]$,

(iii) it holds for all $k \in \{1, 2, \dots, L-1\}$ that

$$(\nabla_{B_k} \mathcal{L})(\Phi) = \frac{1}{M} \left[\sum_{m=1}^M [\text{diag}(\mathfrak{M}_{a', l_k}(\mathbf{W}_k \mathfrak{x}_{k-1}^{(m)} + \mathbf{B}_k))] D_{k+1}^{(m)} \right], \quad (8.54)$$

(iv) it holds that $(\nabla_{W_L} \mathcal{L})(\Phi) = \frac{1}{M} [\sum_{m=1}^M D_{L+1}^{(m)} [\mathfrak{x}_{L-1}^{(m)}]^]$, and*

(v) it holds for all $k \in \{1, 2, \dots, L-1\}$ that

$$(\nabla_{W_k} \mathcal{L})(\Phi) = \frac{1}{M} \left[\sum_{m=1}^M [\text{diag}(\mathfrak{M}_{a', l_k}(\mathbf{W}_k \mathfrak{x}_{k-1}^{(m)} + \mathbf{B}_k))] D_{k+1}^{(m)} [\mathfrak{x}_{k-1}^{(m)}]^* \right]. \quad (8.55)$$

Proof of Corollary 8.2.3. Throughout this proof, let $\mathbf{L}^{(m)}: \bigtimes_{k=1}^L (\mathbb{R}^{l_k \times l_{k-1}} \times \mathbb{R}^{l_k}) \rightarrow \mathbb{R}$,

$m \in \{1, 2, \dots, M\}$, satisfy for all $m \in \{1, 2, \dots, M\}$, $\Psi \in \bigtimes_{k=1}^L (\mathbb{R}^{l_k \times l_{k-1}} \times \mathbb{R}^{l_k})$ that

$$\mathbf{L}^{(m)}(\Psi) = \mathfrak{C}((\mathcal{R}_a^{\mathbf{N}}(\Psi))(\mathfrak{x}_0^{(m)}), \mathfrak{y}^{(m)}). \quad (8.56)$$

Note that (8.56) and (8.51) ensure that for all $\Psi \in \bigtimes_{k=1}^L (\mathbb{R}^{l_k \times l_{k-1}} \times \mathbb{R}^{l_k})$ it holds that

$$\mathcal{L}(\Psi) = \frac{1}{M} \left[\sum_{m=1}^M \mathbf{L}^{(m)}(\Psi) \right]. \quad (8.57)$$

Corollary 8.2.2 hence establishes items (i), (ii), (iii), (iv), and (v). The proof of Corollary 8.2.3 is thus complete. \square

Corollary 8.2.4 (Backpropagation for ANNs with quadratic loss and minibatches). *Let $L, M \in \mathbb{N}$, $l_0, l_1, \dots, l_L \in \mathbb{N}$, $\Phi = ((\mathbf{W}_1, \mathbf{B}_1), \dots, (\mathbf{W}_L, \mathbf{B}_L)) \in \bigtimes_{k=1}^L (\mathbb{R}^{l_k \times l_{k-1}} \times \mathbb{R}^{l_k})$, let $a: \mathbb{R} \rightarrow \mathbb{R}$ be differentiable, for every $m \in \{1, 2, \dots, M\}$ let $\mathfrak{x}_0^{(m)} \in \mathbb{R}^{l_0}$, $\mathfrak{x}_1^{(m)} \in \mathbb{R}^{l_1}, \dots, \mathfrak{x}_L^{(m)} \in \mathbb{R}^{l_L}$, $\mathfrak{y}^{(m)} \in \mathbb{R}^{l_L}$ satisfy for all $k \in \{1, 2, \dots, L\}$ that*

$$\mathfrak{x}_k^{(m)} = \mathfrak{M}_{a \mathbb{1}_{[0, L]}(k) + \text{id}_{\mathbb{R}} \mathbb{1}_{\{L\}}(k), l_k}(\mathbf{W}_k \mathfrak{x}_{k-1}^{(m)} + \mathbf{B}_k), \quad (8.58)$$

let $\mathcal{L} = (\mathcal{L}((W_1, B_1), \dots, (W_L, B_L)))_{((W_1, B_1), \dots, (W_L, B_L)) \in \bigtimes_{k=1}^L (\mathbb{R}^{l_k \times l_{k-1}} \times \mathbb{R}^{l_k})}: \bigtimes_{k=1}^L (\mathbb{R}^{l_k \times l_{k-1}} \times \mathbb{R}^{l_k}) \rightarrow \mathbb{R}$ satisfy for all $\Psi \in (\bigtimes_{k=1}^L (\mathbb{R}^{l_k \times l_{k-1}} \times \mathbb{R}^{l_k}))$ that

$$\mathcal{L}(\Psi) = \frac{1}{M} \left[\sum_{m=1}^M \|(\mathcal{R}_a^{\mathbf{N}}(\Psi))(\mathfrak{x}_0^{(m)}) - \mathfrak{y}^{(m)}\|_2^2 \right], \quad (8.59)$$

and for every $m \in \{1, 2, \dots, M\}$ let $D_k^{(m)} \in \mathbb{R}^{l_{k-1}}$, $k \in \{1, 2, \dots, L+1\}$, satisfy for all $k \in \{1, 2, \dots, L-1\}$ that

$$D_{L+1}^{(m)} = 2(\mathfrak{x}_L^{(m)} - \mathfrak{y}^{(m)}), \quad D_L^{(m)} = [\mathbf{W}_L]^* D_{L+1}^{(m)}, \quad \text{and} \quad (8.60)$$

$$D_k^{(m)} = [\mathbf{W}_k]^* [\text{diag}(\mathfrak{M}_{a', l_k}(\mathbf{W}_k \mathfrak{x}_{k-1}^{(m)} + \mathbf{B}_k))] D_{k+1}^{(m)} \quad (8.61)$$

(cf. Definitions 1.2.1, 1.3.4, 3.3.4, and 8.2.1). Then

(i) it holds that $\mathcal{L}: \bigtimes_{k=1}^L (\mathbb{R}^{l_k \times l_{k-1}} \times \mathbb{R}^{l_k}) \rightarrow \mathbb{R}$ is differentiable,

(ii) it holds that $(\nabla_{B_L} \mathcal{L})(\Phi) = \frac{1}{M} [\sum_{m=1}^M D_{L+1}^{(m)}]$,

(iii) it holds for all $k \in \{1, 2, \dots, L-1\}$ that

$$(\nabla_{B_k} \mathcal{L})(\Phi) = \frac{1}{M} \left[\sum_{m=1}^M [\text{diag}(\mathfrak{M}_{a', l_k}(\mathbf{W}_k \mathfrak{x}_{k-1}^{(m)} + \mathbf{B}_k))] D_{k+1}^{(m)} \right], \quad (8.62)$$

(iv) it holds that $(\nabla_{W_L} \mathcal{L})(\Phi) = \frac{1}{M} [\sum_{m=1}^M D_{L+1}^{(m)} [\mathfrak{x}_{L-1}^{(m)}]^*]$, and

(v) it holds for all $k \in \{1, 2, \dots, L-1\}$ that

$$(\nabla_{W_k} \mathcal{L})(\Phi) = \frac{1}{M} \left[\sum_{m=1}^M [\text{diag}(\mathfrak{M}_{a',l_k}(\mathbf{W}_k \mathfrak{x}_{k-1}^{(m)} + \mathbf{B}_k))] D_{k+1}^{(m)} [\mathfrak{x}_{k-1}^{(m)}]^* \right]. \quad (8.63)$$

Proof of Corollary 8.2.4. Throughout this proof, let $\mathfrak{C} = (\mathfrak{C}(x, y))_{(x,y) \in \mathbb{R}^{l_L} \times \mathbb{R}^{l_L}} : \mathbb{R}^{l_L} \times \mathbb{R}^{l_L} \rightarrow \mathbb{R}$ satisfy for all $x, y \in \mathbb{R}^{l_L}$ that

$$\mathfrak{C}(x, y) = \|x - y\|_2^2, \quad (8.64)$$

Observe that (8.64) establishes that for all $m \in \{1, 2, \dots, M\}$ it holds that

$$(\nabla_x \mathfrak{C})(\mathfrak{x}_L^{(m)}, \mathfrak{y}^{(m)}) = 2(\mathfrak{x}_L^{(m)} - \mathfrak{y}^{(m)}) = D_{L+1}^{(m)}. \quad (8.65)$$

Combining this, (8.58), (8.59), (8.60), and (8.61) with Corollary 8.2.3 establishes items (i), (ii), (iii), (iv), and (v). The proof of Corollary 8.2.4 is thus complete. \square

Chapter 9

Kurdyka–Łojasiewicz (KL) inequalities

In Chapter 5 (GF trajectories), Chapter 6 (deterministic GD-type processes), and Chapter 7 (SGD-type processes) we reviewed and studied gradient based processes for the approximate solution of certain optimization problems. In particular, we sketched the approach of general Lyapunov-type functions as well as the special case where the Lyapunov-type function is the squared standard norm around a minimizer resulting in the coercivity-type conditions used in several convergence results in Chapters 5, 6, and 7. However, the coercivity-type conditions in Chapters 5, 6, and 7 are usually too restrictive to cover the situation of the training of ANNs (cf., for instance, item (ii) in Lemma 5.7.29, [235], item (vi) in Corollary 29], and [224, Corollary 2.19]).

In this chapter we introduce another general class of Lyapunov-type functions which does indeed cover the mathematical analysis of many of the ANN training situations. Specifically, in this chapter we study Lyapunov-type functions that are given by suitable fractional powers of differences of the risk function (cf., for example (9.7) in the proof of Proposition 9.2.1 below). In that case the resulting Lyapunov-type conditions (cf., for instance, (9.1), (9.3), and (9.10) below) are referred to as KL inequalities in the literature.

Further investigations related to KL inequalities in the scientific literature can, for example, be found in [39, 46, 87, 106, 132, 236]. The specific presentation of this chapter is in parts closely based on [234, Sections 3, 7, and 8].

9.1 Standard KL functions

Definition 9.1.1 (Standard KL inequalities). *Let $\mathfrak{d} \in \mathbb{N}$, $c \in \mathbb{R}$, $\alpha \in (0, \infty)$, let $U \subseteq \mathbb{R}^{\mathfrak{d}}$ be open, let $\vartheta \in U$, and let $\mathcal{L}: U \rightarrow \mathbb{R}$ be a function. Then we say that \mathcal{L} satisfies the standard KL inequality at ϑ with exponent α and constant c (we say that \mathcal{L} satisfies the standard KL inequality at ϑ) if and only if*

(i) it holds that \mathcal{L} is differentiable and

(ii) it holds for all $\theta \in U$ that

$$|\mathcal{L}(\vartheta) - \mathcal{L}(\theta)|^\alpha \leq c \|(\nabla \mathcal{L})(\theta)\|_2 \quad (9.1)$$

(cf. Definition 3.3.4).

Definition 9.1.2 (Standard KL functions). Let $\mathfrak{d} \in \mathbb{N}$ and let $\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$ be a function. Then we say that \mathcal{L} is a standard KL function if and only if for all $\vartheta \in \mathbb{R}^{\mathfrak{d}}$ there exist $\varepsilon, c \in (0, \infty)$, $\alpha \in (0, 1)$ such that $\mathcal{L}|_{\{\theta \in \mathbb{R}^{\mathfrak{d}}: \|\theta - \vartheta\|_2 < \varepsilon\}}$ satisfies the standard KL inequality at ϑ with exponent α and constant c (cf. Definitions 3.3.4 and 9.1.1).

9.2 Convergence analysis using standard KL functions (regular regime)

Proposition 9.2.1. Let $\mathfrak{d} \in \mathbb{N}$, $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, $\mathfrak{c}, \mathfrak{C}, \varepsilon \in (0, \infty)$, $\alpha \in (0, 1)$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$, let $O \subseteq \mathbb{R}^{\mathfrak{d}}$ satisfy

$$O = \{\theta \in \mathbb{R}^{\mathfrak{d}}: \|\theta - \vartheta\|_2 < \varepsilon\} \setminus \{\vartheta\} \quad \text{and} \quad \mathfrak{c} = \mathfrak{C}^2 [\sup_{\theta \in O} |\mathcal{L}(\theta) - \mathcal{L}(\vartheta)|]^{2-2\alpha}, \quad (9.2)$$

assume for all $\theta \in O$ that $\mathcal{L}(\theta) > \mathcal{L}(\vartheta)$ and

$$|\mathcal{L}(\theta) - \mathcal{L}(\vartheta)|^\alpha \leq \mathfrak{C} \|(\nabla \mathcal{L})(\theta)\|_2, \quad (9.3)$$

and let $\Theta \in C([0, \infty), O)$ satisfy for all $t \in [0, \infty)$ that

$$\Theta_t = \Theta_0 - \int_0^t (\nabla \mathcal{L})(\Theta_s) \, ds \quad (9.4)$$

(cf. Definition 3.3.4). Then there exists $\psi \in \mathbb{R}^{\mathfrak{d}}$ such that

(i) it holds that $\mathcal{L}(\psi) = \mathcal{L}(\vartheta)$,

(ii) it holds for all $t \in [0, \infty)$ that

$$0 \leq \mathcal{L}(\Theta_t) - \mathcal{L}(\psi) \leq [(\mathcal{L}(\Theta_0) - \mathcal{L}(\psi))^{-1} + \mathfrak{c}^{-1}t]^{-1}, \quad (9.5)$$

and

(iii) it holds for all $t \in [0, \infty)$ that

$$\begin{aligned} \|\Theta_t - \psi\|_2 &\leq \int_t^\infty \|(\nabla \mathcal{L})(\Theta_s)\|_2 ds \\ &\leq \mathfrak{C}(1-\alpha)^{-1} [\mathcal{L}(\Theta_t) - \mathcal{L}(\psi)]^{1-\alpha} \\ &\leq \mathfrak{C}(1-\alpha)^{-1} [(\mathcal{L}(\Theta_0) - \mathcal{L}(\psi))^{-1} + \mathfrak{c}^{-1}t]^{\alpha-1}. \end{aligned} \quad (9.6)$$

Proof of Proposition 9.2.1. Throughout this proof, let $V: O \rightarrow \mathbb{R}$ and $U: O \rightarrow \mathbb{R}$ satisfy for all $\theta \in O$ that

$$V(\theta) = -|\mathcal{L}(\theta) - \mathcal{L}(\vartheta)|^{-1} \quad \text{and} \quad U(\theta) = |\mathcal{L}(\theta) - \mathcal{L}(\vartheta)|^{1-\alpha}. \quad (9.7)$$

Observe that the assumption that for all $\theta \in O$ it holds that $|\mathcal{L}(\theta) - \mathcal{L}(\vartheta)|^\alpha \leq \mathfrak{C}\|(\nabla \mathcal{L})(\theta)\|_2$ demonstrates that for all $\theta \in O$ it holds that

$$\|(\nabla \mathcal{L})(\theta)\|_2^2 \geq \mathfrak{C}^{-2}|\mathcal{L}(\theta) - \mathcal{L}(\vartheta)|^{2\alpha}. \quad (9.8)$$

Furthermore, note that (9.7) ensures that for all $\theta \in O$ it holds that $V \in C^1(O, \mathbb{R})$ and

$$(\nabla V)(\theta) = |\mathcal{L}(\theta) - \mathcal{L}(\vartheta)|^{-2}(\nabla \mathcal{L})(\theta). \quad (9.9)$$

Combining this with (9.8) implies that for all $\theta \in O$ it holds that

$$\begin{aligned} \langle (\nabla V)(\theta), -(\nabla \mathcal{L})(\theta) \rangle &= -|\mathcal{L}(\theta) - \mathcal{L}(\vartheta)|^{-2}\|(\nabla \mathcal{L})(\theta)\|_2^2 \\ &\leq -\mathfrak{C}^{-2}|\mathcal{L}(\theta) - \mathcal{L}(\vartheta)|^{2\alpha-2} \leq -\mathfrak{c}^{-1}. \end{aligned} \quad (9.10)$$

The assumption that for all $t \in [0, \infty)$ it holds that $\Theta_t \in O$, the assumption that for all $t \in [0, \infty)$ it holds that $\Theta_t = \Theta_0 - \int_0^t (\nabla \mathcal{L})(\Theta_s) ds$, and Proposition 5.8.2 therefore establish that for all $t \in [0, \infty)$ it holds that

$$\begin{aligned} -|\mathcal{L}(\Theta_t) - \mathcal{L}(\vartheta)|^{-1} &= V(\Theta_t) \leq V(\Theta_0) + \int_0^t -\mathfrak{c}^{-1} ds = V(\Theta_0) - \mathfrak{c}^{-1}t \\ &= -|\mathcal{L}(\Theta_0) - \mathcal{L}(\vartheta)|^{-1} - \mathfrak{c}^{-1}t. \end{aligned} \quad (9.11)$$

Hence, we obtain for all $t \in [0, \infty)$ that

$$0 < \mathcal{L}(\Theta_t) - \mathcal{L}(\vartheta) \leq [|\mathcal{L}(\Theta_0) - \mathcal{L}(\vartheta)|^{-1} + \mathfrak{c}^{-1}t]^{-1}. \quad (9.12)$$

Moreover, observe that (9.7) ensures that for all $\theta \in O$ it holds that $U \in C^1(O, \mathbb{R})$ and

$$(\nabla U)(\theta) = (1-\alpha)|\mathcal{L}(\theta) - \mathcal{L}(\vartheta)|^{-\alpha}(\nabla \mathcal{L})(\theta). \quad (9.13)$$

The assumption that for all $\theta \in O$ it holds that $|\mathcal{L}(\theta) - \mathcal{L}(\vartheta)|^\alpha \leq \mathfrak{C}\|(\nabla\mathcal{L})(\theta)\|_2$ therefore shows that for all $\theta \in O$ it holds that

$$\begin{aligned} \langle (\nabla U)(\theta), -(\nabla\mathcal{L})(\theta) \rangle &= -(1-\alpha)|\mathcal{L}(\theta) - \mathcal{L}(\vartheta)|^{-\alpha}\|(\nabla\mathcal{L})(\theta)\|_2^2 \\ &\leq -\mathfrak{C}^{-1}(1-\alpha)\|(\nabla\mathcal{L})(\theta)\|_2. \end{aligned} \quad (9.14)$$

Combining this, the assumption that for all $t \in [0, \infty)$ it holds that $\Theta_t \in O$, the fact that for all $s, t \in [0, \infty)$ it holds that

$$\Theta_{s+t} = \Theta_s - \int_0^t (\nabla\mathcal{L})(\Theta_{s+u}) du, \quad (9.15)$$

and Proposition 5.8.2 (applied for every $s \in [0, \infty)$, $t \in (s, \infty)$ with $\mathfrak{d} \curvearrowright \mathfrak{d}$, $T \curvearrowright t-s$, $\alpha \curvearrowright 0$, $O \curvearrowright O$, $\beta \curvearrowright (O \ni \theta \mapsto -\mathfrak{C}^{-1}(1-\alpha)\|(\nabla\mathcal{L})(\theta)\|_2 \in \mathbb{R})$, $\mathcal{G} \curvearrowright (\nabla\mathcal{L})|_O$, $\Theta \curvearrowright ([0, t-s] \ni u \mapsto \Theta_{s+u} \in O)$ in the notation of Proposition 5.8.2) ensures that for all $s, t \in [0, \infty)$ with $s < t$ it holds that

$$\begin{aligned} 0 &< |\mathcal{L}(\Theta_t) - \mathcal{L}(\vartheta)|^{1-\alpha} = U(\Theta_t) \\ &\leq U(\Theta_s) + \int_s^t -\mathfrak{C}^{-1}(1-\alpha)\|(\nabla\mathcal{L})(\Theta_u)\|_2 du \\ &= |\mathcal{L}(\Theta_s) - \mathcal{L}(\vartheta)|^{1-\alpha} - \mathfrak{C}^{-1}(1-\alpha) \left[\int_s^t \|(\nabla\mathcal{L})(\Theta_u)\|_2 du \right]. \end{aligned} \quad (9.16)$$

This proves that for all $s, t \in [0, \infty)$ with $s < t$ it holds that

$$\int_s^t \|(\nabla\mathcal{L})(\Theta_u)\|_2 du \leq \mathfrak{C}(1-\alpha)^{-1}|\mathcal{L}(\Theta_s) - \mathcal{L}(\vartheta)|^{1-\alpha}. \quad (9.17)$$

Hence, we obtain that

$$\int_0^\infty \|(\nabla\mathcal{L})(\Theta_s)\|_2 ds \leq \mathfrak{C}(1-\alpha)^{-1}|\mathcal{L}(\Theta_0) - \mathcal{L}(\vartheta)|^{1-\alpha} < \infty \quad (9.18)$$

This demonstrates that

$$\limsup_{r \rightarrow \infty} \int_r^\infty \|(\nabla\mathcal{L})(\Theta_s)\|_2 ds = 0. \quad (9.19)$$

In addition, note that the fundamental theorem of calculus and the assumption that for all $t \in [0, \infty)$ it holds that $\Theta_t = \Theta_0 - \int_0^t (\nabla\mathcal{L})(\Theta_s) ds$ establish that for all $r, s, t \in [0, \infty)$ with $r \leq s \leq t$ it holds that

$$\|\Theta_t - \Theta_s\|_2 = \left\| \int_s^t (\nabla\mathcal{L})(\Theta_u) du \right\|_2 \leq \int_s^t \|(\nabla\mathcal{L})(\Theta_u)\|_2 du \leq \int_r^\infty \|(\nabla\mathcal{L})(\Theta_u)\|_2 du. \quad (9.20)$$

This and (9.19) demonstrate that there exists $\psi \in \mathbb{R}^{\mathfrak{d}}$ which satisfies

$$\limsup_{t \rightarrow \infty} \|\Theta_t - \psi\|_2 = 0. \quad (9.21)$$

Combining this and the assumption that \mathcal{L} is continuous with (9.12) demonstrates that

$$\mathcal{L}(\psi) = \mathcal{L}\left(\lim_{t \rightarrow \infty} \Theta_t\right) = \lim_{t \rightarrow \infty} \mathcal{L}(\Theta_t) = \mathcal{L}(\vartheta). \quad (9.22)$$

Next observe that (9.21), (9.17), and (9.20) imply that for all $t \in [0, \infty)$ it holds that

$$\begin{aligned} \|\Theta_t - \psi\|_2 &= \left\| \Theta_t - \left[\lim_{s \rightarrow \infty} \Theta_s \right] \right\|_2 \\ &= \lim_{s \rightarrow \infty} \|\Theta_t - \Theta_s\|_2 \\ &\leq \int_t^\infty \|(\nabla \mathcal{L})(\Theta_u)\|_2 du \\ &\leq \mathfrak{C}(1-\alpha)^{-1} |\mathcal{L}(\Theta_t) - \mathcal{L}(\vartheta)|^{1-\alpha}. \end{aligned} \quad (9.23)$$

Combining this with (9.12) and (9.22) establishes items (i), (ii), and (iii). The proof of Proposition 9.2.1 is thus complete. \square

9.3 Standard KL inequalities for monomials

Lemma 9.3.1 (Standard KL inequalities for monomials). *Let $\mathfrak{d} \in \mathbb{N}$, $p \in (1, \infty)$, $\varepsilon, c, \alpha \in (0, \infty)$ satisfy $c \geq p^{-1} \varepsilon^{p(\alpha-1)+1}$ and $\alpha \geq 1 - \frac{1}{p}$ and let $\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that*

$$\mathcal{L}(\theta) = \|\theta\|_2^p. \quad (9.24)$$

Then

- (i) *it holds that $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ and*
- (ii) *it holds for all $\theta \in \{v \in \mathbb{R}^{\mathfrak{d}} : \|v\|_2 \leq \varepsilon\}$ that*

$$|\mathcal{L}(0) - \mathcal{L}(\theta)|^\alpha \leq c \|(\nabla \mathcal{L})(\theta)\|_2. \quad (9.25)$$

Proof of Lemma 9.3.1. First, note that the fact that for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ it holds that

$$\mathcal{L}(\theta) = (\|\theta\|_2^2)^{p/2} \quad (9.26)$$

shows that for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ it holds that $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ and

$$\|(\nabla \mathcal{L})(\theta)\|_2 = p \|\theta\|_2^{p-1}. \quad (9.27)$$

Furthermore, observe that the assumption that $\alpha \geq 1 - \frac{1}{p}$ ensures that $p(\alpha - 1) + 1 \geq 0$. The assumption that $c \geq p^{-1}\varepsilon^{p(\alpha-1)+1}$ therefore ensures that for all $\theta \in \{v \in \mathbb{R}^{\mathfrak{d}} : \|v\|_2 \leq \varepsilon\}$ it holds that

$$\|\theta\|_2^{p\alpha} \|\theta\|_2^{-(p-1)} = \|\theta\|_2^{p(\alpha-1)+1} \leq \varepsilon^{p(\alpha-1)+1} \leq cp. \quad (9.28)$$

Combining (9.27) and (9.28) proves that for all $\theta \in \{v \in \mathbb{R}^{\mathfrak{d}} : \|v\|_2 \leq \varepsilon\}$ it holds that

$$|\mathcal{L}(0) - \mathcal{L}(\theta)|^\alpha = \|\theta\|_2^{p\alpha} \leq cp\|\theta\|_2^{p-1} = c\|(\nabla \mathcal{L})(\theta)\|_2. \quad (9.29)$$

This completes the proof of Lemma 9.3.1. \square

9.4 Standard KL inequalities around non-critical points

Lemma 9.4.1 (Standard KL inequality around non-critical points). *Let $\mathfrak{d} \in \mathbb{N}$, let $U \subseteq \mathbb{R}^{\mathfrak{d}}$ be open, and let $\mathcal{L} \in C^1(U, \mathbb{R})$, $\vartheta \in U$, $c \in [0, \infty)$, $\alpha \in (0, \infty)$ satisfy for all $\theta \in U$ that*

$$\max\{|\mathcal{L}(\vartheta) - \mathcal{L}(\theta)|^\alpha, c\|(\nabla \mathcal{L})(\vartheta) - (\nabla \mathcal{L})(\theta)\|_2\} \leq \frac{c\|(\nabla \mathcal{L})(\vartheta)\|_2}{2} \quad (9.30)$$

(cf. Definition 3.3.4). Then it holds for all $\theta \in U$ that

$$|\mathcal{L}(\vartheta) - \mathcal{L}(\theta)|^\alpha \leq c\|(\nabla \mathcal{L})(\theta)\|_2. \quad (9.31)$$

Proof of Lemma 9.4.1. Note that (9.30) and the triangle inequality ensure that for all $\theta \in U$ it holds that

$$\begin{aligned} & c\|(\nabla \mathcal{L})(\vartheta)\|_2 \\ &= c\|(\nabla \mathcal{L})(\theta) + [(\nabla \mathcal{L})(\vartheta) - (\nabla \mathcal{L})(\theta)]\|_2 \\ &\leq c\|(\nabla \mathcal{L})(\theta)\|_2 + c\|(\nabla \mathcal{L})(\vartheta) - (\nabla \mathcal{L})(\theta)\|_2 \leq c\|(\nabla \mathcal{L})(\theta)\|_2 + \frac{c\|(\nabla \mathcal{L})(\vartheta)\|_2}{2}. \end{aligned} \quad (9.32)$$

Hence, we obtain for all $\theta \in U$ that

$$\frac{c\|(\nabla \mathcal{L})(\vartheta)\|_2}{2} \leq c\|(\nabla \mathcal{L})(\theta)\|_2. \quad (9.33)$$

Combining this with (9.30) establishes that for all $\theta \in U$ it holds that

$$|\mathcal{L}(\vartheta) - \mathcal{L}(\theta)|^\alpha \leq \frac{c\|(\nabla \mathcal{L})(\vartheta)\|_2}{2} \leq c\|(\nabla \mathcal{L})(\theta)\|_2. \quad (9.34)$$

The proof of Lemma 9.4.1 is thus complete. \square

Corollary 9.4.2 (Standard KL inequality around non-critical points). *Let $\mathfrak{d} \in \mathbb{N}$, let $U \subseteq \mathbb{R}^{\mathfrak{d}}$ be open, let $\mathcal{L} \in C^1(U, \mathbb{R})$, $\vartheta \in U$, $c, \alpha \in (0, \infty)$ satisfy $(\nabla \mathcal{L})(\vartheta) \neq 0$. Then there exists $\varepsilon \in (0, 1)$ such that for all $\theta \in \{v \in U : \|v - \vartheta\|_2 < \varepsilon\}$ it holds that*

$$|\mathcal{L}(\vartheta) - \mathcal{L}(\theta)|^{\alpha} \leq c \|(\nabla \mathcal{L})(\theta)\|_2 \quad (9.35)$$

(cf. Definition 3.3.4).

Proof of Corollary 9.4.2. Observe that the assumption that $\mathcal{L} \in C^1(U, \mathbb{R})$ ensures that

$$\limsup_{\varepsilon \searrow 0} (\sup_{\theta \in \{v \in U : \|v - \vartheta\|_2 < \varepsilon\}} \|(\nabla \mathcal{L})(\vartheta) - (\nabla \mathcal{L})(\theta)\|_2) = 0 \quad (9.36)$$

(cf. Definition 3.3.4). Combining this and the fact that $\alpha > 0$ with the fact that \mathcal{L} is continuous demonstrates that

$$\limsup_{\varepsilon \searrow 0} (\sup_{\theta \in \{v \in U : \|v - \vartheta\|_2 < \varepsilon\}} \max\{|\mathcal{L}(\vartheta) - \mathcal{L}(\theta)|^{\alpha}, c \|(\nabla \mathcal{L})(\vartheta) - (\nabla \mathcal{L})(\theta)\|_2\}) = 0. \quad (9.37)$$

The fact that $c > 0$ and the fact that $\|(\nabla \mathcal{L})(\vartheta)\|_2 > 0$ therefore demonstrate that there exists $\varepsilon \in (0, 1)$ which satisfies

$$\sup_{\theta \in \{v \in U : \|v - \vartheta\|_2 < \varepsilon\}} \max\{|\mathcal{L}(\vartheta) - \mathcal{L}(\theta)|^{\alpha}, c \|(\nabla \mathcal{L})(\vartheta) - (\nabla \mathcal{L})(\theta)\|_2\} < \frac{c \|(\nabla \mathcal{L})(\vartheta)\|_2}{2}. \quad (9.38)$$

Note that (9.38) ensures that for all $\theta \in \{v \in U : \|v - \vartheta\|_2 < \varepsilon\}$ it holds that

$$\max\{|\mathcal{L}(\vartheta) - \mathcal{L}(\theta)|^{\alpha}, c \|(\nabla \mathcal{L})(\vartheta) - (\nabla \mathcal{L})(\theta)\|_2\} \leq \frac{c \|(\nabla \mathcal{L})(\vartheta)\|_2}{2}. \quad (9.39)$$

This and Lemma 9.4.1 establish (9.35). The proof of Corollary 9.4.2 is thus complete. \square

9.5 Standard KL inequalities with increased exponents

Lemma 9.5.1 (Standard KL inequalities with increased exponents). *Let $\mathfrak{d} \in \mathbb{N}$, let $U \subseteq \mathbb{R}^{\mathfrak{d}}$ be a set, let $\vartheta \in U$, $\mathfrak{c}, \alpha \in (0, \infty)$, let $\mathcal{L}: U \rightarrow \mathbb{R}$ and $\mathcal{G}: U \rightarrow \mathbb{R}$ satisfy for all $\theta \in U$ that*

$$|\mathcal{L}(\vartheta) - \mathcal{L}(\theta)|^{\alpha} \leq \mathfrak{c} |\mathcal{G}(\theta)|, \quad (9.40)$$

and let $\beta \in (\alpha, \infty)$, $\mathfrak{C} \in \mathbb{R}$ satisfy $\mathfrak{C} = \mathfrak{c} (\sup_{\theta \in U} |\mathcal{L}(\vartheta) - \mathcal{L}(\theta)|^{\beta-\alpha})$. Then it holds for all $\theta \in U$ that

$$|\mathcal{L}(\vartheta) - \mathcal{L}(\theta)|^{\beta} \leq \mathfrak{C} |\mathcal{G}(\theta)|. \quad (9.41)$$

Proof of Lemma 9.5.1. Observe that (9.40) implies that for all $\theta \in U$ it holds that

$$\begin{aligned} |\mathcal{L}(\vartheta) - \mathcal{L}(\theta)|^{\beta} &= |\mathcal{L}(\vartheta) - \mathcal{L}(\theta)|^{\alpha} |\mathcal{L}(\vartheta) - \mathcal{L}(\theta)|^{\beta-\alpha} \leq (\mathfrak{c} |\mathcal{G}(\theta)|) |\mathcal{L}(\vartheta) - \mathcal{L}(\theta)|^{\beta-\alpha} \\ &= (\mathfrak{c} |\mathcal{L}(\vartheta) - \mathcal{L}(\theta)|^{\beta-\alpha}) |\mathcal{G}(\theta)| \leq \mathfrak{C} |\mathcal{G}(\theta)|. \end{aligned} \quad (9.42)$$

This establishes (9.41). The proof of Lemma 9.5.1 is thus complete. \square

Corollary 9.5.2 (Standard KL inequalities with increased exponents). *Let $\mathfrak{d} \in \mathbb{N}$, $\mathcal{L} \in C^1(\mathbb{R}^\mathfrak{d}, \mathbb{R})$, $\vartheta \in \mathbb{R}^\mathfrak{d}$, $\varepsilon, \mathfrak{c}, \alpha \in (0, \infty)$, $\beta \in [\alpha, \infty)$ satisfy for all $\theta \in \{v \in \mathbb{R}^\mathfrak{d} : \|v - \vartheta\|_2 < \varepsilon\}$ that*

$$|\mathcal{L}(\vartheta) - \mathcal{L}(\theta)|^\alpha \leq \mathfrak{c} \|(\nabla \mathcal{L})(\theta)\|_2 \quad (9.43)$$

(cf. Definition 3.3.4). Then there exists $\mathfrak{C} \in (0, \infty)$ such that for all $\theta \in \{v \in \mathbb{R}^\mathfrak{d} : \|v - \vartheta\|_2 < \varepsilon\}$ it holds that

$$|\mathcal{L}(\vartheta) - \mathcal{L}(\theta)|^\beta \leq \mathfrak{C} \|(\nabla \mathcal{L})(\theta)\|_2. \quad (9.44)$$

Proof of Corollary 9.5.2. Note that Lemma 9.5.1 establishes (9.44). The proof of Corollary 9.5.2 is thus complete. \square

9.6 Standard KL inequalities for coercive-type functions

Lemma 9.6.1 (On a growth bound on the gradient). *Let $\mathfrak{d} \in \mathbb{N}$, $\vartheta \in \mathbb{R}^\mathfrak{d}$, $L, \rho \in (0, \infty)$, $r \in (0, \infty]$, $\mathbb{B} = \{v \in \mathbb{R}^\mathfrak{d} : \|v - \vartheta\|_2 < r\}$, $\mathcal{L} \in C^1(\mathbb{R}^\mathfrak{d}, \mathbb{R})$ satisfy for all $\theta \in \mathbb{B}$ that*

$$\|(\nabla \mathcal{L})(\theta)\|_2 \leq L \|\theta - \vartheta\|_2^\rho \quad (9.45)$$

(cf. Definition 3.3.4). Then it holds for all $\theta \in \mathbb{B}$ that

$$|\mathcal{L}(\theta) - \mathcal{L}(\vartheta)| \leq L \|\theta - \vartheta\|_2^{\rho+1}. \quad (9.46)$$

Proof of Lemma 9.6.1. Observe that (9.45), the fundamental theorem of calculus, and the Cauchy-Schwarz inequality assure that for all $\theta \in \mathbb{B}$ it holds that

$$\begin{aligned} |\mathcal{L}(\theta) - \mathcal{L}(\vartheta)| &= \left| \left[\mathcal{L}(\vartheta + h(\theta - \vartheta)) \right]_{h=0}^{h=1} \right| \\ &= \left| \int_0^1 \mathcal{L}'(\vartheta + h(\theta - \vartheta))(\theta - \vartheta) dh \right| \\ &= \left| \int_0^1 \langle (\nabla \mathcal{L})(\vartheta + h(\theta - \vartheta)), \theta - \vartheta \rangle dh \right| \\ &\leq \int_0^1 \|(\nabla \mathcal{L})(\vartheta + h(\theta - \vartheta))\|_2 \|\theta - \vartheta\|_2 dh \\ &\leq \int_0^1 L \|\vartheta + h(\theta - \vartheta) - \vartheta\|_2^\rho \|\theta - \vartheta\|_2 dh \\ &= L \|\theta - \vartheta\|_2^{\rho+1} \int_0^1 h^\rho dh = \frac{L \|\theta - \vartheta\|_2^{\rho+1}}{\rho + 1} \leq L \|\theta - \vartheta\|_2^{\rho+1}. \end{aligned} \quad (9.47)$$

The proof of Lemma 9.6.1 is thus complete. \square

Lemma 9.6.2 (Explicit KL inequalities for coercive-type functions). *Let $\mathfrak{d} \in \mathbb{N}$, $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, $c, \rho \in (0, \infty)$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that*

$$\langle \theta - \vartheta, (\nabla \mathcal{L})(\theta) \rangle \geq c \|\theta - \vartheta\|_2^2 \quad (9.48)$$

and assume that $\nabla \mathcal{L}$ is locally ρ -Hölder continuous (cf. Definitions 1.4.7 and 3.3.4). Then

(i) *for every $\theta \in \mathbb{R}^{\mathfrak{d}} \setminus \{\vartheta\}$, $\alpha, \mathfrak{C} \in (0, \infty)$ there exists $\varepsilon \in (0, 1)$ such that for all $w \in \{v \in \mathbb{R}^{\mathfrak{d}} : \|v - \theta\|_2 < \varepsilon\}$ it holds that*

$$|\mathcal{L}(\theta) - \mathcal{L}(w)|^{\alpha} \leq \mathfrak{C} \|(\nabla \mathcal{L})(w)\|_2 \quad (9.49)$$

and

(ii) *for every $\theta \in \mathbb{R}^{\mathfrak{d}}$, $\alpha \in [1/(1 + \rho), \infty)$ there exist $\mathfrak{C} \in (0, \infty)$, $\varepsilon \in (0, 1)$ such that for all $w \in \{v \in \mathbb{R}^{\mathfrak{d}} : \|v - \theta\|_2 < \varepsilon\}$ it holds that*

$$|\mathcal{L}(\theta) - \mathcal{L}(w)|^{\alpha} \leq \mathfrak{C} \|(\nabla \mathcal{L})(w)\|_2 \quad (9.50)$$

Proof of Lemma 9.6.2. Throughout this proof, let $L \in (0, \infty)$, $\epsilon \in (0, 1)$ satisfy for all $\theta, w \in \{v \in \mathbb{R}^{\mathfrak{d}} : \|v - \vartheta\|_2 < \epsilon\}$ that

$$\|(\nabla \mathcal{L})(\theta) - (\nabla \mathcal{L})(w)\|_2 \leq L \|\theta - w\|_2^{\rho}. \quad (9.51)$$

Observe that the Cauchy-Schwarz inequality and (9.48) show that for all $\theta \in \mathbb{R}^{\mathfrak{d}} \setminus \{\vartheta\}$ it holds that

$$\|\theta - \vartheta\|_2 \|(\nabla \mathcal{L})(\theta)\|_2 \geq \langle \theta - \vartheta, (\nabla \mathcal{L})(\theta) \rangle \geq c \|\theta - \vartheta\|_2^2. \quad (9.52)$$

This ensures that for all $\theta \in \mathbb{R}^{\mathfrak{d}} \setminus \{\vartheta\}$ it holds that

$$\|(\nabla \mathcal{L})(\theta)\|_2 \geq c \|\theta - \vartheta\|_2 > 0. \quad (9.53)$$

Hence, we obtain that for all $\theta \in \mathbb{R}^{\mathfrak{d}} \setminus \{\vartheta\}$ it holds that $(\nabla \mathcal{L})(\theta) \neq 0$. Corollary 9.4.2 therefore proves that for all $\theta \in \mathbb{R}^{\mathfrak{d}} \setminus \{\vartheta\}$, $\alpha, \mathfrak{C} \in (0, \infty)$ there exists $\varepsilon \in (0, 1)$ such that for all $\theta \in \{v \in \mathbb{R}^{\mathfrak{d}} : \|v - \theta\|_2 < \varepsilon\}$ it holds that

$$|\mathcal{L}(\theta) - \mathcal{L}(\theta)|^{\alpha} \leq \mathfrak{C} \|(\nabla \mathcal{L})(\theta)\|_2. \quad (9.54)$$

This establishes item (i). Furthermore, note that (9.48) and item (iii) in Lemma 5.7.29 establish that $(\nabla \mathcal{L})(\vartheta) = 0$. This and (9.51) demonstrate that for all $\theta \in \{v \in \mathbb{R}^{\mathfrak{d}} : \|v - \vartheta\|_2 < \epsilon\}$ it holds that

$$\|(\nabla \mathcal{L})(\theta)\|_2 = \|(\nabla \mathcal{L})(\theta) - (\nabla \mathcal{L})(\vartheta)\|_2 \leq L \|\theta - \vartheta\|_2^{\rho}. \quad (9.55)$$

Combining this with (9.53) and Lemma 9.6.1 implies that for all $\theta \in \{v \in \mathbb{R}^{\mathfrak{d}} : \|v - \vartheta\|_2 < \epsilon\}$ it holds that

$$|\mathcal{L}(\theta) - \mathcal{L}(\vartheta)| \leq L\|\theta - \vartheta\|_2^{\rho+1} \leq L\left(\frac{1}{c}\|(\nabla \mathcal{L})(\theta)\|_2\right)^{\rho+1}. \quad (9.56)$$

Hence, we obtain that for all $\theta \in \{v \in \mathbb{R}^{\mathfrak{d}} : \|v - \vartheta\|_2 < \epsilon\}$ it holds that

$$|\mathcal{L}(\theta) - \mathcal{L}(\vartheta)|^{1/(1+\rho)} \leq L^{1/(1+\rho)}c\|(\nabla \mathcal{L})(\theta)\|_2. \quad (9.57)$$

This and Corollary 9.5.2 show that for all $\alpha \in [1/(1+\rho), \infty)$ there exists $\mathfrak{C} \in (0, \infty)$ such that for all $\theta \in \{v \in \mathbb{R}^{\mathfrak{d}} : \|v - \vartheta\|_2 < \epsilon\}$ it holds that

$$|\mathcal{L}(\theta) - \mathcal{L}(\vartheta)|^{\alpha} \leq \mathfrak{C}\|(\nabla \mathcal{L})(\theta)\|_2. \quad (9.58)$$

This and (9.54) establish item (ii). The proof of Lemma 9.6.2 is thus complete. \square

Proposition 9.6.3 (Coercive-type functions are standard KL functions). *Let $\mathfrak{d} \in \mathbb{N}$ and let $\mathcal{L} \in C^2(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ be a coercive-type function (cf. Definition 5.7.26). Then \mathcal{L} is a standard KL function (cf. Definition 9.1.2).*

Proof of Proposition 9.6.3. Observe that the fact that $\nabla \mathcal{L}$ is continuously differentiable ensures that $\nabla \mathcal{L}$ is locally 1-Hölder continuous. Combining this, the assumption that \mathcal{L} is a coercive-type function, and item (ii) in Lemma 9.6.2 proves that for every $\theta \in \mathbb{R}^{\mathfrak{d}}$, $\alpha \in [1/2, \infty)$ there exist $c \in (0, \infty)$, $\varepsilon \in (0, 1)$ such that for all $w \in \{v \in \mathbb{R}^{\mathfrak{d}} : \|v - \theta\|_2 < \varepsilon\}$ it holds that

$$|\mathcal{L}(\theta) - \mathcal{L}(w)|^{\alpha} \leq c\|(\nabla \mathcal{L})(w)\|_2. \quad (9.59)$$

This and (9.1) establish that \mathcal{L} is a standard KL function. The proof of Proposition 9.6.3 is thus complete. \square

9.7 Standard KL inequalities for one-dimensional polynomials

Definition 9.7.1 (Polynomial). *Let $d, \delta \in \mathbb{N}$ and let $p: \mathbb{R}^d \rightarrow \mathbb{R}^{\delta}$ be a function. Then we say that p is a polynomial if and only if there exist $N \in \mathbb{N}$ and $c = (c_{\alpha})_{\alpha \in \{0, 1, \dots, N\}^d}: \{0, 1, \dots, N\}^d \rightarrow \mathbb{R}^{\delta}$ such that for all $\theta = (\theta_1, \dots, \theta_d) \in \mathbb{R}^d$ it holds that*

$$p(\theta) = \sum_{\alpha=(\alpha_1, \dots, \alpha_d) \in \{0, 1, \dots, N\}^d} c_{\alpha}(\theta_1)^{\alpha_1}(\theta_2)^{\alpha_2} \cdots (\theta_d)^{\alpha_d}. \quad (9.60)$$

Corollary 9.7.2 (Reparametrization). *Let $\vartheta \in \mathbb{R}$, $N \in \mathbb{N}$, $p \in C^\infty(\mathbb{R}, \mathbb{R})$ satisfy for all $\theta \in \mathbb{R}$ that $p^{(N+1)}(\theta) = 0$ and let $\beta_0, \beta_1, \dots, \beta_N \in \mathbb{R}$ satisfy for all $n \in \{0, 1, \dots, N\}$ that $\beta_n = \frac{p^{(n)}(\vartheta)}{n!}$. Then it holds for all $\theta \in \mathbb{R}$ that*

$$p(\theta) = \sum_{n=0}^N \beta_n (\theta - \vartheta)^n. \quad (9.61)$$

Proof of Corollary 9.7.2. Note that Theorem 6.1.4 establishes (9.61). The proof of Corollary 9.7.2 is thus complete. \square

Corollary 9.7.3 (Equivalent conditions for one-dimensional polynomials). *Let $p: \mathbb{R} \rightarrow \mathbb{R}$ be a function. Then the following four statements are equivalent.*

- (i) *It holds that p is a polynomial (cf. Definition 9.7.1).*
- (ii) *There exists $N \in \mathbb{N}$ such that for all $\theta \in \mathbb{R}$ it holds that $p \in C^N(\mathbb{R}, \mathbb{R})$ and $p^{(N)}(\theta) = 0$.*
- (iii) *For every $\vartheta \in \mathbb{R}$ there exist $N \in \mathbb{N}$, $\beta_0, \beta_1, \dots, \beta_N \in \mathbb{R}$ such that for all $\theta \in \mathbb{R}$ it holds that*

$$p(\theta) = \sum_{n=0}^N \beta_n (\theta - \vartheta)^n. \quad (9.62)$$

- (iv) *There exists $N \in \mathbb{N}$, $\beta_0, \beta_1, \dots, \beta_N \in \mathbb{R}$ such that for all $\theta \in \mathbb{R}$ it holds that*

$$p(\theta) = \sum_{n=0}^N \beta_n \theta^n. \quad (9.63)$$

Proof of Corollary 9.7.3. Observe that (9.60) and (9.63) establish that ((i) \leftrightarrow (iv)). Note that the fact that for all $N \in \mathbb{N}$, $g \in C^N(\mathbb{R}, \mathbb{R})$ with $\forall \theta \in \mathbb{R}: g^{(N)}(\theta) = 0$ it holds that $g \in C^\infty(\mathbb{R}, \mathbb{R})$ and Corollary 9.7.2 establish that ((ii) \rightarrow (iii)). Observe that (9.62) and (9.63) establish that ((iii) \rightarrow (iv)). Note that (9.63) establishes that ((iv) \rightarrow (ii)). The proof of Corollary 9.7.3 is thus complete. \square

Corollary 9.7.4 (Quantitative standard KL inequalities for non-constant one-dimensional polynomials). *Let $\vartheta \in \mathbb{R}$, $N \in \mathbb{N}$, $p \in C^\infty(\mathbb{R}, \mathbb{R})$ satisfy for all $\theta \in \mathbb{R}$ that $p^{(N+1)}(\theta) = 0$, let $\beta_0, \beta_1, \dots, \beta_N \in \mathbb{R}$ satisfy for all $n \in \{0, 1, \dots, N\}$ that $\beta_n = \frac{p^{(n)}(\vartheta)}{n!}$, and let $m \in \{1, 2, \dots, N\}$, $\alpha \in [0, 1]$, $c, \varepsilon \in \mathbb{R}$ satisfy*

$$\beta_m \neq 0 = \sum_{n=1}^{m-1} |\beta_n|, \quad \alpha \geq 1 - m^{-1}, \quad c = 2 \left[\sum_{n=1}^N \frac{|\beta_n|^\alpha}{|\beta_m m|} \right], \quad (9.64)$$

and $\varepsilon = \frac{1}{2}[\sum_{n=1}^N \frac{|\beta_n n|}{|\beta_m m|}]^{-1}$. Then it holds for all $\theta \in [\vartheta - \varepsilon, \vartheta + \varepsilon]$ that

$$|p(\theta) - p(\vartheta)|^\alpha \leq c|p'(\theta)|. \quad (9.65)$$

Proof of Corollary 9.7.4. Observe that Corollary 9.7.2 ensures that for all $\theta \in \mathbb{R}$ it holds that

$$p(\theta) - p(\vartheta) = \sum_{n=1}^N \beta_n (\theta - \vartheta)^n. \quad (9.66)$$

Therefore, we obtain for all $\theta \in \mathbb{R}$ that

$$p'(\theta) = \sum_{n=1}^N \beta_n n (\theta - \vartheta)^{n-1} \quad (9.67)$$

Hence, we obtain for all $\theta \in \mathbb{R}$ that

$$p(\theta) - p(\vartheta) = \sum_{n=m}^N \beta_n (\theta - \vartheta)^n \quad \text{and} \quad p'(\theta) = \sum_{n=m}^N \beta_n n (\theta - \vartheta)^{n-1}. \quad (9.68)$$

Therefore, we obtain for all $\theta \in \mathbb{R}$ that

$$|p(\theta) - p(\vartheta)|^\alpha \leq \sum_{n=m}^N (|\beta_n|^\alpha |\theta - \vartheta|^{n\alpha}). \quad (9.69)$$

The fact that for all $n \in \{m, m+1, \dots, N\}$, $\theta \in \mathbb{R}$ with $|\theta - \vartheta| \leq 1$ it holds that $|\theta - \vartheta|^{n\alpha} \leq |\theta - \vartheta|^{n(1-m^{-1})} \leq |\theta - \vartheta|^{m(1-m^{-1})} = |\theta - \vartheta|^{m-1}$ hence demonstrates that for all $\theta \in \mathbb{R}$ with $|\theta - \vartheta| \leq 1$ it holds that

$$\begin{aligned} |p(\theta) - p(\vartheta)|^\alpha &\leq \sum_{n=m}^N |\beta_n|^\alpha |\theta - \vartheta|^{n\alpha} \leq \sum_{n=m}^N |\beta_n|^\alpha |\theta - \vartheta|^{m-1} \\ &= |\theta - \vartheta|^{m-1} \left[\sum_{n=m}^N |\beta_n|^\alpha \right] = |\theta - \vartheta|^{m-1} \left[\sum_{n=1}^N |\beta_n|^\alpha \right]. \end{aligned} \quad (9.70)$$

Therefore, we obtain for all $\theta \in \mathbb{R}$ with $|\theta - \vartheta| \leq 1$ that

$$|p(\theta) - p(\vartheta)|^\alpha \leq |\theta - \vartheta|^{m-1} \left[\sum_{n=1}^N |\beta_n|^\alpha \right] = \frac{c}{2} |\theta - \vartheta|^{m-1} |\beta_m m|. \quad (9.71)$$

Furthermore, note that (9.68) ensures that for all $\theta \in \mathbb{R}$ with $|\theta - \vartheta| \leq 1$ it holds that

$$\begin{aligned} |p'(\theta)| &= \left| \sum_{n=m}^N \beta_n n (\theta - \vartheta)^{n-1} \right| \geq |\beta_m m| |\theta - \vartheta|^{m-1} - \left| \sum_{n=m+1}^N \beta_n n (\theta - \vartheta)^{n-1} \right| \\ &\geq |\theta - \vartheta|^{m-1} |\beta_m m| - \left[\sum_{n=m+1}^N |\theta - \vartheta|^{n-1} |\beta_n n| \right] \\ &\geq |\theta - \vartheta|^{m-1} |\beta_m m| - \left[\sum_{n=m+1}^N |\theta - \vartheta|^m |\beta_n n| \right] \\ &= |\theta - \vartheta|^{m-1} |\beta_m m| - |\theta - \vartheta|^m \left[\sum_{n=m+1}^N |\beta_n n| \right]. \end{aligned} \quad (9.72)$$

Hence, we obtain for all $\theta \in \mathbb{R}$ with $|\theta - \vartheta| \leq \frac{1}{2} \left[\sum_{n=m}^N \frac{|\beta_n n|}{|\beta_m m|} \right]^{-1}$ that

$$\begin{aligned} |p'(\theta)| &\geq |\theta - \vartheta|^{m-1} \left(|\beta_m m| - |\theta - \vartheta| \left[\sum_{n=m+1}^N |\beta_n n| \right] \right) \\ &\geq |\theta - \vartheta|^{m-1} \left(|\beta_m m| - \frac{|\beta_m m|}{2} \left(|\theta - \vartheta| \left[\sum_{n=m}^N \frac{2|\beta_n n|}{|\beta_m m|} \right] \right) \right) \\ &\geq |\theta - \vartheta|^{m-1} \left(|\beta_m m| - \frac{|\beta_m m|}{2} \right) = \frac{1}{2} |\theta - \vartheta|^{m-1} |\beta_m m|. \end{aligned} \quad (9.73)$$

Combining this with (9.71) implies that for all $\theta \in \mathbb{R}$ with $|\theta - \vartheta| \leq \frac{1}{2} \left[\sum_{n=m}^N \frac{|\beta_n n|}{|\beta_m m|} \right]^{-1}$ it holds that

$$|p(\theta) - p(\vartheta)|^\alpha \leq \frac{c}{2} |\theta - \vartheta|^{m-1} |\beta_m m| \leq c |p'(\theta)|. \quad (9.74)$$

This establishes (9.65). The proof of Corollary 9.7.4 is thus complete. \square

Corollary 9.7.5 (Quantitative standard KL inequalities for general one-dimensional polynomials). *Let $\vartheta \in \mathbb{R}$, $N \in \mathbb{N}$, $p \in C^\infty(\mathbb{R}, \mathbb{R})$ satisfy for all $\theta \in \mathbb{R}$ that $p^{(N+1)}(\theta) = 0$, let $\beta_0, \beta_1, \dots, \beta_N \in \mathbb{R}$ satisfy for all $n \in \{0, 1, \dots, N\}$ that $\beta_n = \frac{p^{(n)}(\vartheta)}{n!}$, let $\rho \in \mathbb{R}$ satisfy $\rho = \mathbb{1}_{\{0\}}(\sum_{n=1}^N |\beta_n|) + \min(((\bigcup_{n=1}^N \{|\beta_n n|\}) \setminus \{0\}) \cup \{\sum_{n=1}^N |\beta_n n|\})$, and let $\alpha \in (0, 1]$, $c, \varepsilon \in [0, \infty)$ satisfy*

$$\alpha \geq 1 - N^{-1}, \quad c \geq 2\rho^{-1} [\sum_{n=1}^N |\beta_n|^\alpha], \quad \text{and} \quad \varepsilon \leq \rho [\mathbb{1}_{\{0\}}(\sum_{n=1}^N |\beta_n|) + 2(\sum_{n=1}^N |\beta_n n|)]^{-1}. \quad (9.75)$$

Then it holds for all $\theta \in [\vartheta - \varepsilon, \vartheta + \varepsilon]$ that

$$|p(\theta) - p(\vartheta)|^\alpha \leq c |p'(\theta)|. \quad (9.76)$$

Proof of Corollary 9.7.5. Throughout this proof, assume without loss of generality that

$$\sup_{\theta \in \mathbb{R}} |p(\theta) - p(\vartheta)| > 0. \quad (9.77)$$

Observe that Corollary 9.7.2 and (9.77) ensure that $\sum_{n=1}^N |\beta_n| > 0$. Therefore, we obtain that there exists $m \in \{1, 2, \dots, N\}$ which satisfies

$$|\beta_m| > 0 = \sum_{n=1}^{m-1} |\beta_n|. \quad (9.78)$$

Note that (9.78), the fact that $\alpha \geq 1 - N^{-1}$, and Corollary 9.7.4 show that for all $\theta \in \mathbb{R}$ with $|\theta - \vartheta| \leq \frac{1}{2} [\sum_{n=1}^N \frac{|\beta_n n|}{|\beta_m m|}]^{-1}$ it holds that

$$|p(\theta) - p(\vartheta)|^\alpha \leq \left[\sum_{n=1}^N \frac{2|\beta_n|^\alpha}{|\beta_m m|} \right] |p'(\theta)| \leq \left[\frac{2}{\rho} \left[\sum_{n=1}^N |\beta_n|^\alpha \right] \right] |p'(\theta)| \leq c |p'(\theta)|. \quad (9.79)$$

This establishes (9.76). The proof of Corollary 9.7.5 is thus complete. \square

Corollary 9.7.6. *Let $\vartheta \in \mathbb{R}$, $N \in \mathbb{N}$, $\alpha \in [1 - \frac{1}{N}, \infty) \cap (0, \infty)$, $p \in C^\infty(\mathbb{R}, \mathbb{R})$ satisfy for all $\theta \in \mathbb{R}$ that $p^{(N+1)}(\theta) = 0$. Then there exist $\varepsilon, c \in (0, \infty)$ such that for all $\theta \in [\vartheta - \varepsilon, \vartheta + \varepsilon]$ it holds that*

$$|p(\theta) - p(\vartheta)|^\alpha \leq c |p'(\theta)|. \quad (9.80)$$

Proof of Corollary 9.7.6. Observe that Corollary 9.5.2 and Corollary 9.7.5 ensure (9.80). The proof of Corollary 9.7.6 is thus complete. \square

Corollary 9.7.7 (Qualitative standard **KL** inequalities for general one-dimensional polynomials). *Let $\vartheta \in \mathbb{R}$, $N \in \mathbb{N}$, $p \in C^\infty(\mathbb{R}, \mathbb{R})$ satisfy for all $\theta \in \mathbb{R}$ that $p^{(N)}(\theta) = 0$. Then there exist $\varepsilon, c \in (0, \infty)$, $\alpha \in (0, 1)$ such that for all $\theta \in [\vartheta - \varepsilon, \vartheta + \varepsilon]$ it holds that*

$$|p(\theta) - p(\vartheta)|^\alpha \leq c|p'(\theta)|. \quad (9.81)$$

Proof of Corollary 9.7.7. Note that Corollary 9.7.5 proves (9.81). The proof of Corollary 9.7.7 is thus complete. \square

Corollary 9.7.8. *Let $\mathcal{L}: \mathbb{R} \rightarrow \mathbb{R}$ be a polynomial. Then \mathcal{L} is a standard **KL** function (cf. Definition 9.1.2).*

Proof of Corollary 9.7.8. Observe that (9.1) and Corollary 9.7.7 establish that \mathcal{L} is a standard **KL** function (cf. Definition 9.1.2). The proof of Corollary 9.7.8 is thus complete. \square

9.8 Power series and analytic functions

Definition 9.8.1 (Analytic functions). *Let $m, n \in \mathbb{N}$, let $U \subseteq \mathbb{R}^m$ be open, and let $f: U \rightarrow \mathbb{R}^n$ be a function. Then we say that f is analytic if and only if*

- (i) *it holds that $f \in C^\infty(U, \mathbb{R}^n)$ and*
- (ii) *for all $\vartheta \in U$ there exists $\varepsilon \in (0, \infty)$ such that for all $\theta \in \{v \in U: \|\vartheta - v\|_2 < \varepsilon\}$ it holds that*

$$\limsup_{K \rightarrow \infty} \left\| f(\theta) - \sum_{k=0}^K \frac{1}{k!} f^{(k)}(\vartheta)(\theta - \vartheta, \theta - \vartheta, \dots, \theta - \vartheta) \right\|_2 = 0 \quad (9.82)$$

(cf. Definition 3.3.4).

Lemma 9.8.2 (Higher derivatives of multidimensional functions). *Let $k, n, m \in \mathbb{N}$, let $U \subseteq \mathbb{R}^m$ be open, and let $f \in C^k(U, \mathbb{R}^n)$, $\theta = (\theta_1, \dots, \theta_m) \in U$, $v^1 = (v_1^1, \dots, v_m^1)$,*

$v^2 = (v_1^2, \dots, v_m^2), \dots, v^k = (v_1^k, \dots, v_m^k) \in \mathbb{R}^m$. Then

$$f^{(k)}(\theta)(v^1, v^2, \dots, v^k) = \sum_{i_1, i_2, \dots, i_k \in \{1, \dots, m\}} v_{i_1}^1 v_{i_2}^2 \cdots v_{i_k}^k \left(\frac{\partial^k f}{\partial \theta_{i_1} \partial \theta_{i_2} \cdots \partial \theta_{i_k}} \right)(\theta). \quad (9.83)$$

Proof of Lemma 9.8.2. Throughout this proof, let $e_1, e_2, \dots, e_m \in \mathbb{R}^m$ satisfy

$$e_1 = (1, 0, \dots, 0), \quad e_2 = (0, 1, 0, \dots, 0), \quad \dots, \quad e_m = (0, \dots, 0, 1). \quad (9.84)$$

Note that the fact that $f^{(k)}(\theta)$ is k -linear and the fact that for all $i_1, i_2, \dots, i_k \in \{1, \dots, m\}$ it holds that $f^{(k)}(\theta)(e_{i_1}, e_{i_2}, \dots, e_{i_k}) = \left(\frac{\partial^k f}{\partial \theta_{i_1} \partial \theta_{i_2} \cdots \partial \theta_{i_k}} \right)(\theta)$ establish

$$\begin{aligned} f^{(k)}(\theta)(v^1, v^2, \dots, v^k) &= f^{(k)}(\theta) \left(\sum_{i_1=1}^m v_{i_1}^1 e_{i_1}, \sum_{i_2=1}^m v_{i_2}^2 e_{i_2}, \dots, \sum_{i_k=1}^m v_{i_k}^k e_{i_k} \right) \\ &= \sum_{i_1, i_2, \dots, i_k \in \{1, \dots, m\}} f^{(k)}(\theta)(v_{i_1}^1 e_{i_1}, v_{i_2}^2 e_{i_2}, \dots, v_{i_k}^k e_{i_k}) \\ &= \sum_{i_1, i_2, \dots, i_k \in \{1, \dots, m\}} v_{i_1}^1 v_{i_2}^2 \cdots v_{i_k}^k f^{(k)}(\theta)(e_{i_1}, e_{i_2}, \dots, e_{i_k}) \\ &= \sum_{i_1, i_2, \dots, i_k \in \{1, \dots, m\}} v_{i_1}^1 v_{i_2}^2 \cdots v_{i_k}^k \left(\frac{\partial^k f}{\partial \theta_{i_1} \partial \theta_{i_2} \cdots \partial \theta_{i_k}} \right)(\theta). \end{aligned} \quad (9.85)$$

The proof of Lemma 9.8.2 is thus complete. \square

Lemma 9.8.3 (One-dimensional analytic functions). *Let $U \subseteq \mathbb{R}$ be open, let $f: U \rightarrow \mathbb{R}$ be a function. Then the following two statements are equivalent:*

- (i) *It holds that f is analytic (cf. Definition 9.8.1).*
- (ii) *For every $\vartheta \in U$ there exists $\varepsilon \in (0, \infty)$ such that for all $\theta \in U \cap (\vartheta - \varepsilon, \vartheta + \varepsilon)$ it holds that $f \in C^\infty(U, \mathbb{R})$ and*

$$\limsup_{K \rightarrow \infty} \left| f(\theta) - \sum_{k=0}^K \frac{f^{(k)}(\vartheta)(\theta - \vartheta)^k}{k!} \right| = 0. \quad (9.86)$$

Proof of Lemma 9.8.3. Observe that (9.82) and Lemma 9.8.2 establish that ((i) \leftrightarrow (ii)). The proof of Lemma 9.8.3 is thus complete. \square

Proposition 9.8.4 (Power series). *Let $m, n \in \mathbb{N}$, $\varepsilon \in (0, \infty)$, let $U \subseteq \mathbb{R}^m$ satisfy $U = \{w \in \mathbb{R}^m : \|w\|_2 \leq \varepsilon\}$, for every $k \in \mathbb{N}$ let $A_k: (\mathbb{R}^m)^k \rightarrow \mathbb{R}^n$ be k -linear and*

symmetric, and let $f: U \rightarrow \mathbb{R}^n$ satisfy for all $\theta \in U$ that

$$\limsup_{K \rightarrow \infty} \left\| f(\theta) - f(0) - \sum_{k=1}^K A_k(\theta, \theta, \dots, \theta) \right\|_2 = 0 \quad (9.87)$$

(cf. Definition 3.3.4). Then

(i) it holds for all $\theta \in \{w \in U : \|w\|_2 < \varepsilon\}$ that $\sum_{k=1}^{\infty} \|A_k(\theta, \theta, \dots, \theta)\|_2 < \infty$ and

$$f(\theta) = f(0) + \sum_{k=1}^{\infty} A_k(\theta, \theta, \dots, \theta), \quad (9.88)$$

(ii) it holds that $f|_{\{w \in U : \|w\|_2 < \varepsilon\}}$ is infinitely often differentiable,

(iii) it holds for all $\theta \in \{w \in U : \|w\|_2 < \varepsilon\}$, $l \in \mathbb{N}$, $v_1, v_2, \dots, v_l \in \mathbb{R}^m$ that

$$\sum_{k=l}^{\infty} \left(\left[\frac{k!}{(k-l)!} \right] \|A_k(v_1, v_2, \dots, v_l, \theta, \theta, \dots, \theta)\|_2 \right) < \infty \quad (9.89)$$

and

$$f^{(l)}(\theta)(v_1, \dots, v_l) = \sum_{k=l}^{\infty} \left(\left[\frac{k!}{(k-l)!} \right] A_k(v_1, v_2, \dots, v_l, \theta, \theta, \dots, \theta) \right), \quad (9.90)$$

and

(iv) it holds for all $k \in \mathbb{N}$ that $f^{(k)}(0) = k! A_k$.

Proof of Proposition 9.8.4. Throughout this proof, for every $K \in \mathbb{N}_0$ let $F_K: \mathbb{R}^m \rightarrow \mathbb{R}^n$ satisfy for all $\theta \in \mathbb{R}^m$ that

$$F_K(\theta) = f(0) + \sum_{k=1}^K A_k(\theta, \theta, \dots, \theta). \quad (9.91)$$

Note that (9.87) and (9.91) ensure that for all $\theta \in U$ it holds that

$$\limsup_{K \rightarrow \infty} \|f(\theta) - F_K(\theta)\|_2 = 0. \quad (9.92)$$

Hence, we obtain for all $\theta \in U$ that

$$\limsup_{K \rightarrow \infty} \|F_{K+1}(\theta) - F_K(\theta)\|_2 = 0. \quad (9.93)$$

This demonstrates for all $\theta \in U$ that

$$\sup_{k \in \mathbb{N}} \|A_k(\theta, \theta, \dots, \theta)\|_2 = \sup_{K \in \mathbb{N}_0} \|F_{K+1}(\theta) - F_K(\theta)\|_2 < \infty. \quad (9.94)$$

Therefore, we obtain for all $\theta \in \{w \in U : \|w\|_2 < \varepsilon\} \setminus \{0\}$ that

$$\begin{aligned} \sum_{k=1}^{\infty} \|A_k(\theta, \theta, \dots, \theta)\|_2 &= \sum_{k=1}^{\infty} \left(\left[\frac{\|\theta\|_2}{\varepsilon} \right]^k \|A_k\left(\frac{\varepsilon\theta}{\|\theta\|_2}, \frac{\varepsilon\theta}{\|\theta\|_2}, \dots, \frac{\varepsilon\theta}{\|\theta\|_2}\right)\|_2 \right) \\ &\leq \left[\sum_{k=1}^{\infty} \left[\frac{\|\theta\|_2}{\varepsilon} \right]^k \right] \left[\sup_{k \in \mathbb{N}} \|A_k\left(\frac{\varepsilon\theta}{\|\theta\|_2}, \frac{\varepsilon\theta}{\|\theta\|_2}, \dots, \frac{\varepsilon\theta}{\|\theta\|_2}\right)\|_2 \right] < \infty. \end{aligned} \quad (9.95)$$

This implies that for all $\theta \in \{w \in U : \|w\|_2 < \varepsilon\}$ it holds that

$$\sum_{k=1}^{\infty} \|A_k(\theta, \theta, \dots, \theta)\|_2 < \infty. \quad (9.96)$$

Combining this with (9.87) establishes item (i). Observe that, for instance, Krantz & Parks [268, Proposition 2.2.3] shows items (ii) and (iii). Note that (9.90) implies item (iv). The proof of Proposition 9.8.4 is thus complete. \square

Proposition 9.8.5 (Characterization for analytic functions). *Let $m, n \in \mathbb{N}$, let $U \subseteq \mathbb{R}^m$ be open, and let $f \in C^\infty(U, \mathbb{R}^n)$. Then the following three statements are equivalent:*

(i) *It holds that f is analytic (cf. Definition 9.8.1).*

(ii) *It holds for all $\vartheta \in U$ that there exists $\varepsilon \in (0, \infty)$ such that for all $\theta \in \{w \in U : \|\vartheta - w\|_2 < \varepsilon\}$ it holds that $\sum_{k=0}^{\infty} \frac{1}{k!} \|f^{(k)}(\vartheta)(\theta - \vartheta, \theta - \vartheta, \dots, \theta - \vartheta)\|_2 < \infty$ and*

$$f(\theta) = \sum_{k=0}^{\infty} \frac{1}{k!} f^{(k)}(\vartheta)(\theta - \vartheta, \theta - \vartheta, \dots, \theta - \vartheta). \quad (9.97)$$

(iii) *It holds for all compact $\mathfrak{C} \subseteq U$ that there exists $c \in \mathbb{R}$ such that for all $\theta \in \mathfrak{C}$, $k \in \mathbb{N}$, $v \in \mathbb{R}^m$ it holds that*

$$\|f^{(k)}(\theta)(v, v, \dots, v)\|_2 \leq k! c^k \|v\|_2^k. \quad (9.98)$$

Proof of Proposition 9.8.5. The equivalence is a direct consequence from Proposition 9.8.4. The proof of Proposition 9.8.5 is thus complete. \square

9.9 Standard KL inequalities for one-dimensional analytic functions

In Section 9.7 above we have seen that one-dimensional polynomials are standard KL functions (see Corollary 9.7.8). In this section we verify that one-dimensional analytic functions are also standard KL functions (see Corollary 9.9.6 below). The main arguments for this statement are presented in the proof of Lemma 9.9.2 and are inspired by [135].

Lemma 9.9.1. Let $\varepsilon \in (0, \infty)$, let $(a_k)_{k \in \mathbb{N}} \subseteq \mathbb{R}$, and let $f: [-\varepsilon, \varepsilon] \rightarrow \mathbb{R}$ satisfy for all $\theta \in [-\varepsilon, \varepsilon]$ that

$$\limsup_{K \rightarrow \infty} \left| f(\theta) - f(0) - \sum_{k=1}^K a_k \theta^k \right| = 0. \quad (9.99)$$

Then

(i) it holds for all $\theta \in (-\varepsilon, \varepsilon)$ that $\sum_{k=1}^{\infty} |a_k| |\theta|^k < \infty$ and

$$f(\theta) = f(0) + \sum_{k=1}^{\infty} a_k \theta^k, \quad (9.100)$$

(ii) it holds that $f|_{(-\varepsilon, \varepsilon)}$ is infinitely often differentiable,

(iii) it holds for all $\theta \in (-\varepsilon, \varepsilon)$, $l \in \mathbb{N}$ that $\sum_{k=l}^{\infty} \left[\frac{k!}{(k-l)!} \right] |a_k| |\theta|^{k-l} < \infty$ and

$$f^{(l)}(\theta) = \sum_{k=l}^{\infty} \left[\frac{k!}{(k-l)!} \right] a_k \theta^{k-l}, \quad (9.101)$$

and

(iv) it holds for all $k \in \mathbb{N}$ that $f^{(k)}(0) = k! a_k$.

Proof of Lemma 9.9.1. Observe that Proposition 9.8.4 (applied with $m \curvearrowleft 1$, $n \curvearrowleft 1$, $\varepsilon \curvearrowleft \varepsilon$, $U \curvearrowleft [-\varepsilon, \varepsilon]$, $(A_k)_{k \in \mathbb{N}} \curvearrowleft ((\mathbb{R}^k \ni (v_1, v_2, \dots, v_k) \mapsto a_k v_1 v_2 \cdots v_k \in \mathbb{R}))_{k \in \mathbb{N}}$, $f \curvearrowleft f$ in the notation of Proposition 9.8.4) establishes items (i), (ii), (iii), and (iv). The proof of Lemma 9.9.1 is thus complete. \square

Lemma 9.9.2. Let $\varepsilon, \delta \in (0, 1)$, $N \in \mathbb{N} \setminus \{1\}$, let $(a_k)_{k \in \mathbb{N}} \subseteq \mathbb{R}$ satisfy $N = \min(\{k \in \mathbb{N}: a_k \neq 0\} \cup \{\infty\})$, let $\mathcal{L}: [-\varepsilon, \varepsilon] \rightarrow \mathbb{R}$ satisfy for all $\theta \in [-\varepsilon, \varepsilon]$ that

$$\limsup_{K \rightarrow \infty} \left| \mathcal{L}(\theta) - \mathcal{L}(0) - \left[\sum_{k=1}^K a_k \theta^k \right] \right| = 0, \quad (9.102)$$

and let $M \in \mathbb{N} \cap (N, \infty)$ satisfy for all $k \in \mathbb{N} \cap [M, \infty)$ that $k|a_k| \leq (2\varepsilon^{-1})^k$ and

$$\delta = \min \left\{ \frac{\varepsilon}{4}, |a_N| \left[(2\varepsilon^{-1})^{N+1} + (\max_{k \in \{1, 2, \dots, M\}} |2ka_k|) \right]^{-1} \right\}. \quad (9.103)$$

Then it holds for all $\theta \in (-\delta, \delta)$ that

$$|\mathcal{L}(\theta) - \mathcal{L}(0)|^{\frac{N-1}{N}} \leq 2|a_N|^{-\frac{1}{N}} |\mathcal{L}'(\theta)|. \quad (9.104)$$

Proof of Lemma 9.9.2. Note that the fact that for all $k \in \mathbb{N} \cap [M, \infty)$ it holds that $|a_k| \leq k|a_k| \leq (2\varepsilon^{-1})^k$ ensures that for all $\theta \in \mathbb{R}$ it holds that

$$\begin{aligned}
 & \sum_{k=N+1}^{\infty} |a_k| |\theta|^k \\
 &= |\theta|^{N+1} \left[\sum_{k=N+1}^{\infty} |a_k| |\theta|^{k-N-1} \right] \\
 &= |\theta|^{N+1} \left[\left(\sum_{k=N+1}^M |a_k| |\theta|^{k-N-1} \right) + \left(\sum_{k=M+1}^{\infty} |a_k| |\theta|^{k-N-1} \right) \right] \\
 &\leq |\theta|^{N+1} \left[\left(\max_{k \in \{1, 2, \dots, M\}} |a_k| \right) \left(\sum_{k=N+1}^M |\theta|^{k-N-1} \right) + \left(\sum_{k=M+1}^{\infty} (2\varepsilon^{-1})^k |\theta|^{k-N-1} \right) \right] \\
 &= |\theta|^{N+1} \left[\left(\max_{k \in \{1, 2, \dots, M\}} |a_k| \right) \left(\sum_{k=0}^{M-N-1} |\theta|^k \right) + (2\varepsilon^{-1})^{N+1} \left(\sum_{k=M-N}^{\infty} (2\varepsilon^{-1}|\theta|)^k \right) \right] \\
 &\leq |\theta|^{N+1} \left[\left(\max_{k \in \{1, 2, \dots, M\}} |a_k| \right) \left(\sum_{k=0}^{\infty} |\theta|^k \right) + (2\varepsilon^{-1})^{N+1} \left(\sum_{k=M-N}^{\infty} (2\varepsilon^{-1}|\theta|)^k \right) \right]. \tag{9.105}
 \end{aligned}$$

Hence, we obtain for all $\theta \in (-\frac{\varepsilon}{4}, \frac{\varepsilon}{4})$ that

$$\begin{aligned}
 \sum_{k=N+1}^{\infty} |a_k| |\theta|^k &\leq |\theta|^{N+1} \left[\left(\max_{k \in \{1, 2, \dots, M\}} |a_k| \right) \left(\sum_{k=0}^{\infty} \left| \frac{1}{4} \right|^k \right) + (2\varepsilon^{-1})^{N+1} \left(\sum_{k=1}^{\infty} \left| \frac{1}{2} \right|^k \right) \right] \\
 &\leq |\theta|^{N+1} \left[2 \left(\max_{k \in \{1, 2, \dots, M\}} |a_k| \right) + (2\varepsilon^{-1})^{N+1} \right]. \tag{9.106}
 \end{aligned}$$

This and (9.103) ensure for all $\theta \in (-\delta, \delta)$ that

$$\sum_{k=N+1}^{\infty} |a_k| |\theta|^k \leq |a_N| |\theta|^N. \tag{9.107}$$

Combining this and (9.102) proves for all $\theta \in (-\delta, \delta)$ that

$$|\mathcal{L}(\theta) - \mathcal{L}(0)| = \left| \sum_{k=N}^{\infty} a_k \theta^k \right| \leq |a_N| |\theta|^N + \left[\sum_{k=N+1}^{\infty} |a_k| |\theta|^k \right] \leq 2|a_N| |\theta|^N. \tag{9.108}$$

Next observe that the assumption that for all $k \in \mathbb{N} \cap [M, \infty)$ it holds that $k|a_k| \leq (2\varepsilon^{-1})^k$

establishes that for all $\theta \in \mathbb{R}$ it holds that

$$\begin{aligned}
 & \sum_{k=N+1}^{\infty} k|a_k||\theta|^{k-1} \\
 &= |\theta|^N \left[\left(\sum_{k=N+1}^{\infty} k|a_k||\theta|^{k-N-1} \right) \right] \\
 &\leq |\theta|^N \left[\left(\sum_{k=N+1}^M k|a_k||\theta|^{k-N-1} \right) + \left(\sum_{k=M+1}^{\infty} (2\varepsilon^{-1})^k |\theta|^{k-N-1} \right) \right] \\
 &\leq |\theta|^N \left[\left(\max_{k \in \{1, 2, \dots, M\}} k|a_k| \right) \left(\sum_{k=N+1}^M |\theta|^{k-N-1} \right) + \left(\sum_{k=M+1}^{\infty} (2\varepsilon^{-1})^k |\theta|^{k-N-1} \right) \right] \\
 &= |\theta|^N \left[\left(\max_{k \in \{1, 2, \dots, M\}} k|a_k| \right) \left(\sum_{k=0}^{M-N-1} |\theta|^k \right) + \left(\sum_{k=M-N}^{\infty} (2\varepsilon^{-1})^{k+N+1} |\theta|^k \right) \right] \\
 &\leq |\theta|^N \left[\left(\max_{k \in \{1, 2, \dots, M\}} k|a_k| \right) \left(\sum_{k=0}^{\infty} |\theta|^k \right) + (2\varepsilon^{-1})^{N+1} \left(\sum_{k=M-N}^{\infty} (2\varepsilon^{-1}|\theta|)^k \right) \right]. \tag{9.109}
 \end{aligned}$$

Therefore, we obtain for all $\theta \in (-\frac{\varepsilon}{4}, \frac{\varepsilon}{4})$ that

$$\begin{aligned}
 & \sum_{k=N+1}^{\infty} k|a_k||\theta|^{k-1} \\
 &\leq |\theta|^N \left[\left(\max_{k \in \{1, 2, \dots, M\}} k|a_k| \right) \left(\sum_{k=0}^{\infty} \left| \frac{1}{4} \right|^k \right) + (2\varepsilon^{-1})^{N+1} \left(\sum_{k=1}^{\infty} \left| \frac{1}{2} \right|^k \right) \right] \\
 &\leq |\theta|^N \left[2 \left(\max_{k \in \{1, 2, \dots, M\}} k|a_k| \right) + (2\varepsilon^{-1})^{N+1} \right]. \tag{9.110}
 \end{aligned}$$

Combining this and (9.103) demonstrates for all $\theta \in (-\delta, \delta)$ that

$$\sum_{k=N+1}^{\infty} k|a_k||\theta|^{k-1} \leq |a_N||\theta|^{N-1}. \tag{9.111}$$

Hence, we obtain for all $K \in \mathbb{N} \cap [N, \infty)$, $\theta \in (-\delta, \delta)$ that

$$\left| \sum_{k=1}^K k a_k \theta^{k-1} \right| = \left| \sum_{k=N}^K k a_k \theta^{k-1} \right| \geq N |a_N| |\theta|^{N-1} - \sum_{k=N+1}^{\infty} k |a_k| |\theta|^{k-1} \geq (N-1) |a_N| |\theta|^{N-1}. \tag{9.112}$$

Lemma 9.9.1 and (9.102) therefore imply that for all $\theta \in (-\delta, \delta)$ it holds that $\sum_{k=1}^{\infty} k |a_k \theta^{k-1}| < \infty$ and

$$|\mathcal{L}'(\theta)| = \left| \sum_{k=1}^{\infty} k a_k \theta^{k-1} \right| \geq (N-1) |a_N| |\theta|^{N-1}. \tag{9.113}$$

Combining this with (9.108) shows that for all $\theta \in (-\delta, \delta)$ it holds that

$$\begin{aligned} |\mathcal{L}(\theta) - \mathcal{L}(0)|^{\frac{N-1}{N}} &\leq [2|a_N||\theta|^N]^{\frac{N-1}{N}} \\ &\leq |2a_N|^{\frac{N-1}{N}}|\theta|^{N-1} \\ &\leq |2a_N|^{\frac{N-1}{N}}(N-1)^{-1}|a_N|^{-1}|\mathcal{L}'(\theta)| \\ &\leq 2|a_N|^{-\frac{1}{N}}(N-1)^{-1}|\mathcal{L}'(\theta)| \\ &\leq 2|a_N|^{-\frac{1}{N}}|\mathcal{L}'(\theta)|. \end{aligned} \quad (9.114)$$

The proof of Lemma 9.9.2 is thus complete. \square

Corollary 9.9.3. Let $\varepsilon \in (0, \infty)$, let $(a_k)_{k \in \mathbb{N}} \subseteq \mathbb{R}$, and let $\mathcal{L}: [-\varepsilon, \varepsilon] \rightarrow \mathbb{R}$ satisfy for all $\theta \in [-\varepsilon, \varepsilon]$ that

$$\limsup_{K \rightarrow \infty} \left| \mathcal{L}(\theta) - \mathcal{L}(0) - \sum_{k=1}^K a_k \theta^k \right| = 0. \quad (9.115)$$

Then there exist $\delta \in (0, \varepsilon)$, $c \in (0, \infty)$, $\alpha \in (0, 1)$ such that for all $\theta \in (-\delta, \delta)$ it holds that

$$|\mathcal{L}(\theta) - \mathcal{L}(0)|^\alpha \leq c |\mathcal{L}'(\theta)|. \quad (9.116)$$

Proof of Corollary 9.9.3. Throughout this proof, assume without loss of generality that $\varepsilon < 1$, let $N \in \mathbb{N} \cup \{\infty\}$ satisfy $N = \min(\{k \in \mathbb{N}: a_k \neq 0\} \cup \{\infty\})$, and assume without loss of generality that $1 < N < \infty$ (cf. items (iii) and (iv) in Lemma 9.9.1 and Corollary 9.4.2). Note that item (iii) in Lemma 9.9.1 ensures that for all $\theta \in (-\varepsilon, \varepsilon)$ it holds that

$$\sum_{k=1}^{\infty} k|a_k||\theta|^{k-1} < \infty. \quad (9.117)$$

Hence, we obtain that

$$\sum_{k=1}^{\infty} k|a_k| \left| \frac{\varepsilon}{2} \right|^k < \infty. \quad (9.118)$$

Therefore, we obtain that

$$\limsup_{k \rightarrow \infty} \left(k|a_k| \left| \frac{\varepsilon}{2} \right|^k \right) = 0. \quad (9.119)$$

This ensures that there exists $M \in \mathbb{N} \cap (N, \infty)$ which satisfies for all $k \in \mathbb{N} \cap [M, \infty)$ that

$$k|a_k| \leq (2\varepsilon^{-1})^k. \quad (9.120)$$

Lemma 9.9.2 hence shows that for all $\theta \in \{w \in \mathbb{R}: |w| < \min\{\frac{\varepsilon}{4}, |a_N|[(\max_{k \in \{1, 2, \dots, M\}} |2ka_k|) + (2\varepsilon^{-1})^{N+1}]^{-1}\}\}$ it holds that

$$|\mathcal{L}(\theta) - \mathcal{L}(0)|^{\frac{N-1}{N}} \leq 2|a_N|^{-\frac{1}{N}}|\mathcal{L}'(\theta)|. \quad (9.121)$$

The proof of Corollary 9.9.3 is thus complete. \square

Corollary 9.9.4. Let $\varepsilon \in (0, \infty)$, $\vartheta \in \mathbb{R}$, let $(a_k)_{k \in \mathbb{N}} \subseteq \mathbb{R}$, and let $\mathcal{L}: [\vartheta - \varepsilon, \vartheta + \varepsilon] \rightarrow \mathbb{R}$ satisfy for all $\theta \in [\vartheta - \varepsilon, \vartheta + \varepsilon]$ that

$$\limsup_{K \rightarrow \infty} \left| \mathcal{L}(\theta) - \mathcal{L}(\vartheta) - \sum_{k=1}^K a_k (\theta - \vartheta)^k \right| = 0. \quad (9.122)$$

Then there exist $\delta \in (0, \varepsilon)$, $c \in (0, \infty)$, $\alpha \in (0, 1)$ such that for all $\theta \in (\vartheta - \delta, \vartheta + \delta)$ it holds that

$$|\mathcal{L}(\theta) - \mathcal{L}(\vartheta)|^\alpha \leq c |\mathcal{L}'(\theta)|. \quad (9.123)$$

Proof of Corollary 9.9.4. Throughout this proof, let $\mathcal{M}: [-\varepsilon, \varepsilon] \rightarrow \mathbb{R}$ satisfy for all $\theta \in [-\varepsilon, \varepsilon]$ that $\mathcal{M}(\theta) = \mathcal{L}(\theta + \vartheta)$. Observe that (9.122) and the fact that for all $\theta \in [-\varepsilon, \varepsilon]$ it holds that $\theta + \vartheta \in [\vartheta - \varepsilon, \vartheta + \varepsilon]$ prove that for all $\theta \in [-\varepsilon, \varepsilon]$ it holds that

$$\begin{aligned} & \limsup_{K \rightarrow \infty} \left| \mathcal{M}(\theta) - \mathcal{M}(0) - \sum_{k=1}^K a_k \theta^k \right| \\ &= \limsup_{K \rightarrow \infty} \left| \mathcal{L}(\theta + \vartheta) - \mathcal{L}(\vartheta) - \sum_{k=1}^K a_k ((\theta + \vartheta) - \vartheta)^k \right| = 0. \end{aligned} \quad (9.124)$$

Corollary 9.9.3 therefore establishes that there exist $\delta \in (0, \varepsilon)$, $c \in (0, \infty)$, $\alpha \in (0, 1)$ which satisfy for all $\theta \in (-\delta, \delta)$ that

$$|\mathcal{M}(\theta) - \mathcal{M}(0)|^\alpha \leq c |\mathcal{M}'(\theta)|. \quad (9.125)$$

Hence, we obtain for all $\theta \in (-\delta, \delta)$ that

$$|\mathcal{L}(\theta + \vartheta) - \mathcal{L}(\vartheta)|^\alpha \leq c |\mathcal{L}'(\theta + \vartheta)|. \quad (9.126)$$

This implies (9.123). The proof of Corollary 9.9.4 is thus complete. \square

Corollary 9.9.5. Let $U \subseteq \mathbb{R}$ be open, let $\mathcal{L}: U \rightarrow \mathbb{R}$ be analytic, and let $\vartheta \in U$ (cf. Definition 9.8.1). Then there exist $\varepsilon, c \in (0, \infty)$, $\alpha \in (0, 1)$ such that for all $\theta \in (\vartheta - \varepsilon, \vartheta + \varepsilon)$ it holds that

$$|\mathcal{L}(\theta) - \mathcal{L}(\vartheta)|^\alpha \leq c |(\nabla \mathcal{L})(\theta)|. \quad (9.127)$$

Proof of Corollary 9.9.5. Note that Lemma 9.8.3 and Corollary 9.9.4 establish (9.127). The proof of Corollary 9.9.5 is thus complete. \square

Corollary 9.9.6. Let $\mathcal{L}: \mathbb{R} \rightarrow \mathbb{R}$ be analytic (cf. Definition 9.8.1). Then \mathcal{L} is a standard KL function (cf. Definition 9.1.2).

Proof of Corollary 9.9.6. Observe that (9.1) and Corollary 9.9.5 establish that \mathcal{L} is a standard KL function (cf. Definition 9.1.2). The proof of Corollary 9.9.6 is thus complete. \square

9.10 Standard KL inequalities for analytic functions

Theorem 9.10.1 (Standard KL inequalities for analytic functions). Let $\mathfrak{d} \in \mathbb{N}$, let $U \subseteq \mathbb{R}^{\mathfrak{d}}$ be open, let $\mathcal{L}: U \rightarrow \mathbb{R}$ be analytic, and let $\vartheta \in U$ (cf. Definition 9.8.1). Then there exist $\varepsilon, c \in (0, \infty)$, $\alpha \in (0, 1)$ such that for all $\theta \in \{v \in U : \|\vartheta - v\|_2 < \varepsilon\}$ it holds that

$$|\mathcal{L}(\vartheta) - \mathcal{L}(\theta)|^\alpha \leq c \|(\nabla \mathcal{L})(\theta)\|_2 \quad (9.128)$$

(cf. Definition 3.3.4).

Proof of Theorem 9.10.1. Note that Łojasiewicz [298, Proposition 1] demonstrates (9.128) (cf., for example, also Bierstone & Milman [39, Proposition 6.8]). The proof of Theorem 9.10.1 is thus complete. \square

Corollary 9.10.2. Let $\mathfrak{d} \in \mathbb{N}$ and let $\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$ be analytic (cf. Definition 9.8.1). Then \mathcal{L} is a standard KL function (cf. Definition 9.1.2).

Proof of Corollary 9.10.2. Observe that (9.1) and Theorem 9.10.1 establish that \mathcal{L} is a standard KL function (cf. Definition 9.1.2). The proof of Corollary 9.10.2 is thus complete. \square

9.11 Counterexamples

Example 9.11.1 (Example of a smooth function that is not a standard KL function). Let $\mathcal{L}: \mathbb{R} \rightarrow \mathbb{R}$ satisfy for all $\theta \in \mathbb{R}$ that

$$\mathcal{L}(\theta) = \begin{cases} \exp(-\theta^{-1}) & : \theta > 0 \\ 0 & : \theta \leq 0. \end{cases} \quad (9.129)$$

Then

- (i) it holds that $\mathcal{L} \in C^\infty(\mathbb{R}, \mathbb{R})$,
- (ii) it holds for all $\theta \in (0, \infty)$ that $\mathcal{L}'(\theta) = \theta^{-2} \exp(-\theta^{-1})$,

(iii) it holds for all $\alpha \in (0, 1)$, $\varepsilon \in (0, \infty)$ that

$$\sup_{\theta \in (0, \varepsilon)} \left(\frac{|\mathcal{L}(\theta) - \mathcal{L}(0)|^\alpha}{|\mathcal{L}'(\theta)|} \right) = \infty, \quad (9.130)$$

and

(iv) it holds that \mathcal{L} is not a standard **KL** function

(cf. Definition 9.1.2).

Proof for Example 9.11.1. Throughout this proof, let

$$P = \{f \in C((0, \infty), \mathbb{R}) : f \text{ is a polynomial}\} \quad (9.131)$$

and for every $f \in C((0, \infty), \mathbb{R})$ let $G_f : (0, \infty) \rightarrow \mathbb{R}$ satisfy for all $\theta \in (0, \infty)$ that

$$G_f(\theta) = f(\theta^{-1}) \exp(-\theta^{-1}). \quad (9.132)$$

Note that the chain rule and the product rule ensure that for all $f \in C^1((0, \infty), \mathbb{R})$, $\theta \in (0, \infty)$ it holds that $G_f \in C^1((0, \infty), \mathbb{R})$ and

$$\begin{aligned} (G_f)'(\theta) &= -f'(\theta^{-1})\theta^{-2} \exp(-\theta^{-1}) + f(\theta^{-1})\theta^{-2} \exp(-\theta^{-1}) \\ &= (f(\theta^{-1}) - f'(\theta^{-1}))\theta^{-2} \exp(-\theta^{-1}). \end{aligned} \quad (9.133)$$

Therefore, we obtain for all $p \in P$ that there exists $q \in P$ such that

$$(G_p)' = G_q. \quad (9.134)$$

Combining this and (9.133) with induction implies that for all $p \in P$, $n \in \mathbb{N}$ it holds that

$$G_p \in C^\infty((0, \infty), \mathbb{R}) \quad \text{and} \quad (\exists q \in P : (G_p)^{(n)} = G_q). \quad (9.135)$$

This and the fact that for all $p \in P$ it holds that $\lim_{\theta \searrow 0} G_p(\theta) = 0$ show that for all $p \in P$, $n \in \mathbb{N}$ it holds that

$$\lim_{\theta \searrow 0} (G_p)^{(n)}(\theta) = 0. \quad (9.136)$$

The fact that $\mathcal{L}|_{(0, \infty)} = G_{(0, \infty) \ni \theta \mapsto 1 \in \mathbb{R}}$ and (9.133) hence establish items (i) and (ii). Observe that (9.129) and the fact that for all $\theta \in (0, \infty)$ it holds that

$$\exp(\theta) = \sum_{k=0}^{\infty} \frac{\theta^k}{k!} \geq \frac{\theta^3}{3!} = \frac{\theta^3}{6} \quad (9.137)$$

ensure that for all $\alpha \in (0, 1)$, $\varepsilon \in (0, \infty)$, $\theta \in (0, \varepsilon)$ it holds that

$$\begin{aligned} \frac{|\mathcal{L}(\theta) - \mathcal{L}(0)|^\alpha}{|\mathcal{L}'(\theta)|} &= \frac{|\mathcal{L}(\theta)|^\alpha}{|\mathcal{L}'(\theta)|} = \frac{\theta^2 |\mathcal{L}(\theta)|^\alpha}{\mathcal{L}(\theta)} = \theta^2 |\mathcal{L}(\theta)|^{\alpha-1} \\ &= \theta^2 \exp\left(\frac{(1-\alpha)}{\theta}\right) \geq \frac{\theta^2(1-\alpha)^3}{6\theta^3} = \frac{(1-\alpha)^3}{6\theta}. \end{aligned} \quad (9.138)$$

Therefore, we obtain for all $\alpha \in (0, 1)$, $\varepsilon \in (0, \infty)$ that

$$\sup_{\theta \in (0, \varepsilon)} \left(\frac{|\mathcal{L}(\theta) - \mathcal{L}(0)|^\alpha}{|\mathcal{L}'(\theta)|} \right) \geq \sup_{\theta \in (0, \varepsilon)} \left(\frac{(1-\alpha)^3}{6\theta} \right) = \infty. \quad (9.139)$$

This establishes items (iii) and (iv). The proof for Example 9.11.1 is thus complete. \square

Example 9.11.2 (Example of a differentiable function that fails to satisfy the standard KL inequality). Let $\mathcal{L}: \mathbb{R} \rightarrow \mathbb{R}$ satisfy for all $\theta \in \mathbb{R}$ that

$$\mathcal{L}(\theta) = \int_0^{\max\{\theta, 0\}} x |\sin(x^{-1})| dx. \quad (9.140)$$

Then

(i) it holds that $\mathcal{L} \in C^1(\mathbb{R}, \mathbb{R})$,

(ii) it holds for all $c \in \mathbb{R}$, $\alpha, \varepsilon \in (0, \infty)$ that there exist $\theta \in (0, \varepsilon)$ such that

$$|\mathcal{L}(\theta) - \mathcal{L}(0)|^\alpha > c |\mathcal{L}'(\theta)|, \quad (9.141)$$

and

(iii) it holds for all $c \in \mathbb{R}$, $\alpha, \varepsilon \in (0, \infty)$ that we do not have that \mathcal{L} satisfies the standard KL inequality at 0 on $[0, \varepsilon]$ with exponent α and constant c

(cf. Definition 9.1.1).

Proof for Example 9.11.2. Throughout this proof, let $\mathcal{G}: \mathbb{R} \rightarrow \mathbb{R}$ satisfy for all $\theta \in \mathbb{R}$ that

$$\mathcal{G}(\theta) = \begin{cases} \theta |\sin(\theta^{-1})| & : \theta > 0 \\ 0 & : \theta \leq 0. \end{cases} \quad (9.142)$$

Note that (9.142) ensures that for all $k \in \mathbb{N}$ it holds that

$$\mathcal{G}((k\pi)^{-1}) = (k\pi)^{-1} |\sin(k\pi)| = 0. \quad (9.143)$$

Furthermore, observe that (9.142) proves for all $\theta \in (0, \infty)$ that

$$|\mathcal{G}(\theta) - \mathcal{G}(0)| = |\theta \sin(\theta^{-1})| \leq |\theta|. \quad (9.144)$$

Hence, we obtain that \mathcal{G} is continuous. This, (9.140), and the fundamental theorem of calculus ensure that \mathcal{L} is continuously differentiable with

$$\mathcal{L}' = \mathcal{G}. \quad (9.145)$$

Combining this with (9.143) demonstrates that for all $c \in \mathbb{R}$, $\alpha \in (0, \infty)$, $k \in \mathbb{N}$ it holds that

$$|\mathcal{L}((k\pi)^{-1}) - \mathcal{L}(0)|^\alpha = [\mathcal{L}((k\pi)^{-1})]^\alpha > 0 = c|\mathcal{G}((k\pi)^{-1})| = c|\mathcal{L}'((k\pi)^{-1})|. \quad (9.146)$$

The proof for Example 9.11.2 is thus complete. \square

Exercise 9.11.1. Let $\mathcal{L}: \mathbb{R} \rightarrow \mathbb{R}$ satisfy for all $\theta \in \mathbb{R}$ that

$$\mathcal{L}(\theta) = \begin{cases} \exp(-\theta^{-2}) & : \theta > 0 \\ 0 & : \theta \leq 0. \end{cases}$$

Prove or disprove the following statement: It holds that \mathcal{L} is a standard **KL** function.

9.12 Convergence analysis for solutions of GF ODEs

In this section we employ standard **KL** inequalities to establish convergence of solutions of **GF ODEs** to critical points (see Section 5.6.3). The specific presentation of this section is closely based on [234, Section 7].

9.12.1 Abstract local convergence results for GF processes

Lemma 9.12.1. *Let $\mathfrak{d} \in \mathbb{N}$, $\Theta \in C([0, \infty), \mathbb{R}^{\mathfrak{d}})$, $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$, let $\mathcal{G}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that $\mathcal{G}(\theta) = (\nabla \mathcal{L})(\theta)$, and assume for all $t \in [0, \infty)$ that $\Theta_t = \Theta_0 - \int_0^t \mathcal{G}(\Theta_s) ds$. Then it holds for all $t \in [0, \infty)$ that*

$$\mathcal{L}(\Theta_t) = \mathcal{L}(\Theta_0) - \int_0^t \|\mathcal{G}(\Theta_s)\|_2^2 ds \quad (9.147)$$

(cf. Definition 3.3.4).

Proof of Lemma 9.12.1. Note that Lemma 5.2.3 implies (9.147). This completes the proof of Lemma 9.12.1. \square

Proposition 9.12.2. *Let $\mathfrak{d} \in \mathbb{N}$, $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, $\mathfrak{c} \in \mathbb{R}$, $\mathfrak{C}, \varepsilon \in (0, \infty)$, $\alpha \in (0, 1)$, $\Theta \in C([0, \infty), \mathbb{R}^{\mathfrak{d}})$, $\mathcal{L} \in C(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$, let $\mathcal{G}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}^{\mathfrak{d}}$ be $\mathcal{B}(\mathbb{R}^{\mathfrak{d}})/\mathcal{B}(\mathbb{R}^{\mathfrak{d}})$ -measurable, assume for*

all $t \in [0, \infty)$ that

$$\mathcal{L}(\Theta_t) = \mathcal{L}(\Theta_0) - \int_0^t \|\mathcal{G}(\Theta_s)\|_2^2 ds \quad \text{and} \quad \Theta_t = \Theta_0 - \int_0^t \mathcal{G}(\Theta_s) ds, \quad (9.148)$$

and assume for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ with $\|\theta - \vartheta\|_2 < \varepsilon$ that

$$|\mathcal{L}(\theta) - \mathcal{L}(\vartheta)|^\alpha \leq \mathfrak{C} \|\mathcal{G}(\theta)\|_2, \quad \mathfrak{c} = |\mathcal{L}(\Theta_0) - \mathcal{L}(\vartheta)|, \quad \mathfrak{C}(1 - \alpha)^{-1} \mathfrak{c}^{1-\alpha} + \|\Theta_0 - \vartheta\|_2 < \varepsilon, \quad (9.149)$$

and $\inf_{t \in \{s \in [0, \infty) : \forall r \in [0, s] : \|\Theta_r - \vartheta\|_2 < \varepsilon\}} \mathcal{L}(\Theta_t) \geq \mathcal{L}(\vartheta)$ (cf. Definition 3.3.4). Then there exists $\psi \in \mathbb{R}^{\mathfrak{d}}$ such that

(i) it holds that $\mathcal{L}(\psi) = \mathcal{L}(\vartheta)$,

(ii) it holds for all $t \in [0, \infty)$ that $\|\Theta_t - \vartheta\|_2 < \varepsilon$,

(iii) it holds for all $t \in [0, \infty)$ that $0 \leq \mathcal{L}(\Theta_t) - \mathcal{L}(\psi) \leq \mathfrak{C}^2 \mathfrak{c}^2 (\mathbb{1}_{\{0\}}(\mathfrak{c}) + \mathfrak{C}^2 \mathfrak{c} + \mathfrak{c}^{2\alpha} t)^{-1}$, and

(iv) it holds for all $t \in [0, \infty)$ that

$$\begin{aligned} \|\Theta_t - \psi\|_2 &\leq \int_t^\infty \|\mathcal{G}(\Theta_s)\|_2 ds \leq \mathfrak{C}(1 - \alpha)^{-1} [\mathcal{L}(\Theta_t) - \mathcal{L}(\psi)]^{1-\alpha} \\ &\leq \mathfrak{C}^{3-2\alpha} \mathfrak{c}^{2-2\alpha} (1 - \alpha)^{-1} (\mathbb{1}_{\{0\}}(\mathfrak{c}) + \mathfrak{C}^2 \mathfrak{c} + \mathfrak{c}^{2\alpha} t)^{\alpha-1}. \end{aligned} \quad (9.150)$$

Proof of Proposition 9.12.2. Throughout this proof, let $\mathbf{L} : [0, \infty) \rightarrow \mathbb{R}$ satisfy for all $t \in [0, \infty)$ that

$$\mathbf{L}(t) = \mathcal{L}(\Theta_t) - \mathcal{L}(\vartheta), \quad (9.151)$$

let $\mathbb{B} \subseteq \mathbb{R}^{\mathfrak{d}}$ satisfy

$$\mathbb{B} = \{\theta \in \mathbb{R}^{\mathfrak{d}} : \|\theta - \vartheta\|_2 < \varepsilon\}, \quad (9.152)$$

let $T \in [0, \infty]$ satisfy

$$T = \inf(\{t \in [0, \infty) : \Theta_t \notin \mathbb{B}\} \cup \{\infty\}), \quad (9.153)$$

let $\tau \in [0, T]$ satisfy

$$\tau = \inf(\{t \in [0, T) : \mathbf{L}(t) = 0\} \cup \{T\}), \quad (9.154)$$

let $\mathcal{G} = (\mathcal{G}_t)_{t \in [0, \infty)} : [0, \infty) \rightarrow [0, \infty]$ satisfy for all $t \in [0, \infty)$ that $\mathcal{G}_t = \int_t^\infty \|\mathcal{G}(\Theta_s)\|_2 ds$, and let $\mathfrak{D} \in \mathbb{R}$ satisfy $\mathfrak{D} = \mathfrak{C}^2 \mathfrak{c}^{(2-2\alpha)}$. In the first step of our proof of items (i), (ii), (iii), and (iv) we show that for all $t \in [0, \infty)$ it holds that

$$\Theta_t \in \mathbb{B}. \quad (9.155)$$

For this we observe that (9.149), the triangle inequality, and the assumption that for all $t \in [0, \infty)$ it holds that $\Theta_t = \Theta_0 - \int_0^t \mathcal{G}(\Theta_s) ds$ establish that for all $t \in [0, \infty)$ it holds that

$$\begin{aligned} \|\Theta_t - \vartheta\|_2 &\leq \|\Theta_t - \Theta_0\|_2 + \|\Theta_0 - \vartheta\|_2 \leq \left\| \int_0^t \mathcal{G}(\Theta_s) ds \right\|_2 + \|\Theta_0 - \vartheta\|_2 \\ &\leq \int_0^t \|\mathcal{G}(\Theta_s)\|_2 ds + \|\Theta_0 - \vartheta\|_2 < \int_0^t \|\mathcal{G}(\Theta_s)\|_2 ds - \mathfrak{C}(1-\alpha)^{-1} |\mathcal{L}(\Theta_0) - \mathcal{L}(\vartheta)|^{1-\alpha} + \varepsilon. \end{aligned} \quad (9.156)$$

To establish (9.155), it is thus sufficient to prove that $\int_0^T \|\mathcal{G}(\Theta_s)\|_2 ds \leq \mathfrak{C}(1-\alpha)^{-1} |\mathcal{L}(\Theta_0) - \mathcal{L}(\vartheta)|^{1-\alpha}$. We will accomplish this by employing an appropriate differential inequality for a fractional power of the function \mathbf{L} in (9.151) (see (9.161) below for details). For this we need several technical preparations. More formally, note that (9.151) and the assumption that for all $t \in [0, \infty)$ it holds that

$$\mathcal{L}(\Theta_t) = \mathcal{L}(\Theta_0) - \int_0^t \|\mathcal{G}(\Theta_s)\|_2^2 ds \quad (9.157)$$

demonstrate that for almost all $t \in [0, \infty)$ it holds that \mathbf{L} is differentiable at t and satisfies

$$\mathbf{L}'(t) = \frac{\partial}{\partial t} (\mathcal{L}(\Theta_t)) = -\|\mathcal{G}(\Theta_t)\|_2^2. \quad (9.158)$$

Furthermore, observe that the assumption that $\inf_{t \in \{s \in [0, \infty) : \forall r \in [0, s] : \|\Theta_r - \vartheta\|_2 < \varepsilon\}} \mathcal{L}(\Theta_t) \geq \mathcal{L}(\vartheta)$ implies that for all $t \in [0, T)$ it holds that

$$\mathbf{L}(t) \geq 0. \quad (9.159)$$

Combining this with (9.149), (9.151), and (9.154) shows that for all $t \in [0, \tau)$ it holds that

$$0 < [\mathbf{L}(t)]^\alpha = |\mathcal{L}(\Theta_t) - \mathcal{L}(\vartheta)|^\alpha \leq \mathfrak{C} \|\mathcal{G}(\Theta_t)\|_2. \quad (9.160)$$

The chain rule and (9.158) therefore ensure that for almost all $t \in [0, \tau)$ it holds that

$$\begin{aligned} \frac{\partial}{\partial t} ([\mathbf{L}(t)]^{1-\alpha}) &= (1-\alpha)[\mathbf{L}(t)]^{-\alpha}(-\|\mathcal{G}(\Theta_t)\|_2^2) \\ &\leq -(1-\alpha)\mathfrak{C}^{-1}\|\mathcal{G}(\Theta_t)\|_2^{-1}\|\mathcal{G}(\Theta_t)\|_2^2 = -\mathfrak{C}^{-1}(1-\alpha)\|\mathcal{G}(\Theta_t)\|_2. \end{aligned} \quad (9.161)$$

Moreover, note that (9.157) proves that $[0, \infty) \ni t \mapsto \mathbf{L}(t) \in \mathbb{R}$ is absolutely continuous. This and the fact that for all $r \in (0, \infty)$ it holds that $[r, \infty) \ni y \mapsto y^{1-\alpha} \in \mathbb{R}$ is Lipschitz continuous establish that for all $t \in [0, \tau)$ it holds that $[0, t] \ni s \mapsto [\mathbf{L}(s)]^{1-\alpha} \in \mathbb{R}$ is absolutely continuous. Combining this with (9.161) demonstrates that for all $s, t \in [0, \tau)$ with $s \leq t$ it holds that

$$\int_s^t \|\mathcal{G}(\Theta_u)\|_2 \, du \leq -\mathfrak{C}(1-\alpha)^{-1}([\mathbf{L}(t)]^{1-\alpha} - [\mathbf{L}(s)]^{1-\alpha}) \leq \mathfrak{C}(1-\alpha)^{-1}[\mathbf{L}(s)]^{1-\alpha}. \quad (9.162)$$

In the next step we observe that (9.157) implies that $[0, \infty) \ni t \mapsto \mathcal{L}(\Theta_t) \in \mathbb{R}$ is non-increasing. This and (9.151) show that \mathbf{L} is non-increasing. Combining (9.154) and (9.159) hence ensures that for all $t \in [\tau, T]$ it holds that $\mathbf{L}(t) = 0$. Therefore, we obtain that for all $t \in (\tau, T)$ it holds that

$$\mathbf{L}'(t) = 0. \quad (9.163)$$

This and (9.158) prove that for almost all $t \in (\tau, T)$ it holds that

$$\mathcal{G}(\Theta_t) = 0. \quad (9.164)$$

Combining this with (9.162) establishes that for all $s, t \in [0, T)$ with $s \leq t$ it holds that

$$\int_s^t \|\mathcal{G}(\Theta_u)\|_2 du \leq \mathfrak{C}(1-\alpha)^{-1} [\mathbf{L}(s)]^{1-\alpha}. \quad (9.165)$$

Hence, we obtain that for all $t \in [0, T)$ it holds that

$$\int_0^t \|\mathcal{G}(\Theta_u)\|_2 du \leq \mathfrak{C}(1-\alpha)^{-1} [\mathbf{L}(0)]^{1-\alpha}. \quad (9.166)$$

In addition, note that (9.149) demonstrates that $\Theta_0 \in \mathbb{B}$. Combining this with (9.153) implies that $T > 0$. This, (9.166), and (9.149) show that

$$\int_0^T \|\mathcal{G}(\Theta_u)\|_2 du \leq \mathfrak{C}(1-\alpha)^{-1} [\mathbf{L}(0)]^{1-\alpha} < \varepsilon < \infty. \quad (9.167)$$

Combining (9.153) and (9.156) hence ensures that

$$T = \infty. \quad (9.168)$$

This proves (9.155). In the next step of our proof of items (i), (ii), (iii), and (iv) we verify that $\Theta_t \in \mathbb{R}^{\mathfrak{d}}$, $t \in [0, \infty)$, is convergent (see (9.170) below). For this observe that the assumption that for all $t \in [0, \infty)$ it holds that $\Theta_t = \Theta_0 - \int_0^t \mathcal{G}(\Theta_s) ds$ establishes that for all $r, s, t \in [0, \infty)$ with $r \leq s \leq t$ it holds that

$$\|\Theta_t - \Theta_s\|_2 = \left\| \int_s^t \mathcal{G}(\Theta_u) du \right\|_2 \leq \int_s^t \|\mathcal{G}(\Theta_u)\|_2 du \leq \int_r^\infty \|\mathcal{G}(\Theta_u)\|_2 du = \varrho_r. \quad (9.169)$$

Next note that (9.167) and (9.168) demonstrate that $\infty > \varrho_0 \geq \limsup_{r \rightarrow \infty} \varrho_r = 0$. Combining this with (9.169) implies that there exist $\psi \in \mathbb{R}^{\mathfrak{d}}$ which satisfies

$$\limsup_{t \rightarrow \infty} \|\Theta_t - \psi\|_2 = 0. \quad (9.170)$$

In the next step of our proof of items (i), (ii), (iii), and (iv) we show that $\mathcal{L}(\Theta_t)$, $t \in [0, \infty)$, converges to $\mathcal{L}(\psi)$ with convergence order 1. We accomplish this by bringing a suitable

differential inequality for the reciprocal of the function \mathbf{L} in (9.151) into play (see (9.173) below for details). More specifically, observe that (9.158), (9.168), (9.153), and (9.149) show that for almost all $t \in [0, \infty)$ it holds that

$$\mathbf{L}'(t) = -\|\mathcal{G}(\Theta_t)\|_2^2 \leq -\mathfrak{C}^{-2}[\mathbf{L}(t)]^{2\alpha}. \quad (9.171)$$

Therefore, we obtain that \mathbf{L} is non-increasing. This ensures that for all $t \in [0, \infty)$ it holds that $\mathbf{L}(t) \leq \mathbf{L}(0)$. This and the fact that for all $t \in [0, \tau)$ it holds that $\mathbf{L}(t) > 0$ prove that for almost all $t \in [0, \tau)$ it holds that

$$\mathbf{L}'(t) \leq -\mathfrak{C}^{-2}[\mathbf{L}(t)]^{(2\alpha-2)}[\mathbf{L}(t)]^2 \leq -\mathfrak{C}^{-2}[\mathbf{L}(0)]^{(2\alpha-2)}[\mathbf{L}(t)]^2 = -\mathfrak{D}^{-1}[\mathbf{L}(t)]^2. \quad (9.172)$$

Hence, we obtain that for almost all $t \in [0, \tau)$ it holds that

$$\frac{\mathfrak{d}}{dt} \left(\frac{\mathfrak{D}}{\mathbf{L}(t)} \right) = - \left(\frac{\mathfrak{D} \mathbf{L}'(t)}{[\mathbf{L}(t)]^2} \right) \geq 1. \quad (9.173)$$

Furthermore, note that the fact that for all $t \in [0, \tau)$ it holds that $[0, t] \ni s \mapsto \mathbf{L}(s) \in (0, \infty)$ is absolutely continuous establishes that for all $t \in [0, \tau)$ it holds that $[0, t] \ni s \mapsto \mathfrak{D}[\mathbf{L}(s)]^{-1} \in (0, \infty)$ is absolutely continuous. This and (9.173) demonstrate that for all $t \in [0, \tau)$ it holds that

$$\frac{\mathfrak{D}}{\mathbf{L}(t)} - \frac{\mathfrak{D}}{\mathbf{L}(0)} \geq t. \quad (9.174)$$

Therefore, we obtain that for all $t \in [0, \tau)$ it holds that

$$\frac{\mathfrak{D}}{\mathbf{L}(t)} \geq \frac{\mathfrak{D}}{\mathbf{L}(0)} + t. \quad (9.175)$$

Hence, we obtain that for all $t \in [0, \tau)$ it holds that

$$\mathfrak{D} \left(\frac{\mathfrak{D}}{\mathbf{L}(0)} + t \right)^{-1} \geq \mathbf{L}(t). \quad (9.176)$$

This implies that for all $t \in [0, \tau)$ it holds that

$$\mathbf{L}(t) \leq \mathfrak{D} (\mathfrak{D}[\mathbf{L}(0)]^{-1} + t)^{-1} = \mathfrak{C}^2 \mathfrak{c}^{2-2\alpha} (\mathfrak{C}^2 \mathfrak{c}^{1-2\alpha} + t)^{-1} = \mathfrak{C}^2 \mathfrak{c}^2 (\mathfrak{C}^2 \mathfrak{c} + \mathfrak{c}^{2\alpha} t)^{-1}. \quad (9.177)$$

The fact that for all $t \in [\tau, \infty)$ it holds that $\mathbf{L}(t) = 0$ and (9.154) therefore show that for all $t \in [0, \infty)$ it holds that

$$0 \leq \mathbf{L}(t) \leq \mathfrak{C}^2 \mathfrak{c}^2 (\mathbb{1}_{\{0\}}(\mathfrak{c}) + \mathfrak{C}^2 \mathfrak{c} + \mathfrak{c}^{2\alpha} t)^{-1}. \quad (9.178)$$

Moreover, observe that (9.170) and the assumption that $\mathcal{L} \in C(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ ensure that $\limsup_{t \rightarrow \infty} |\mathcal{L}(\Theta_t) - \mathcal{L}(\psi)| = 0$. Combining this with (9.178) proves that $\mathcal{L}(\psi) = \mathcal{L}(\vartheta)$. This and (9.178) establish that for all $t \in [0, \infty)$ it holds that

$$0 \leq \mathcal{L}(\Theta_t) - \mathcal{L}(\psi) \leq \mathfrak{C}^2 \mathfrak{c}^2 (\mathbb{1}_{\{0\}}(\mathfrak{c}) + \mathfrak{C}^2 \mathfrak{c} + \mathfrak{c}^{2\alpha} t)^{-1}. \quad (9.179)$$

In the final step of our proof of items (i), (ii), (iii), and (iv) we establish convergence rates for the real numbers $\|\Theta_t - \psi\|_2$, $t \in [0, \infty)$. Note that (9.170), (9.169), and (9.165) demonstrate that for all $t \in [0, \infty)$ it holds that

$$\|\Theta_t - \psi\|_2 = \|\Theta_t - [\lim_{s \rightarrow \infty} \Theta_s]\|_2 = \lim_{s \rightarrow \infty} \|\Theta_t - \Theta_s\|_2 \leq g_t \leq \mathfrak{C}(1-\alpha)^{-1}[\mathbf{L}(t)]^{1-\alpha}. \quad (9.180)$$

This and (9.179) imply that for all $t \in [0, \infty)$ it holds that

$$\begin{aligned} \|\Theta_t - \psi\|_2 &\leq g_t \leq \mathfrak{C}(1-\alpha)^{-1}[\mathcal{L}(\Theta_t) - \mathcal{L}(\psi)]^{1-\alpha} \\ &\leq \mathfrak{C}(1-\alpha)^{-1}[\mathfrak{C}^2 \mathfrak{c}^2 (\mathbb{1}_{\{0\}}(\mathfrak{c}) + \mathfrak{C}^2 \mathfrak{c} + \mathfrak{c}^{2\alpha} t)^{-1}]^{1-\alpha} \\ &= \mathfrak{C}^{3-2\alpha} \mathfrak{c}^{2-2\alpha} (1-\alpha)^{-1} (\mathbb{1}_{\{0\}}(\mathfrak{c}) + \mathfrak{C}^2 \mathfrak{c} + \mathfrak{c}^{2\alpha} t)^{\alpha-1}. \end{aligned} \quad (9.181)$$

Combining this with (9.155) and (9.179) proves items (i), (ii), (iii), and (iv). The proof of Proposition 9.12.2 is thus complete. \square

Corollary 9.12.3. *Let $\mathfrak{d} \in \mathbb{N}$, $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, $\mathfrak{c} \in [0, 1]$, $\mathfrak{C}, \varepsilon \in (0, \infty)$, $\alpha \in (0, 1)$, $\Theta \in C([0, \infty), \mathbb{R}^{\mathfrak{d}})$, $\mathcal{L} \in C(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$, let $\mathcal{G}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}^{\mathfrak{d}}$ be $\mathcal{B}(\mathbb{R}^{\mathfrak{d}})/\mathcal{B}(\mathbb{R}^{\mathfrak{d}})$ -measurable, assume for all $t \in [0, \infty)$ that*

$$\mathcal{L}(\Theta_t) = \mathcal{L}(\Theta_0) - \int_0^t \|\mathcal{G}(\Theta_s)\|_2^2 ds \quad \text{and} \quad \Theta_t = \Theta_0 - \int_0^t \mathcal{G}(\Theta_s) ds, \quad (9.182)$$

and assume for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ with $\|\theta - \vartheta\|_2 < \varepsilon$ that

$$|\mathcal{L}(\theta) - \mathcal{L}(\vartheta)|^\alpha \leq \mathfrak{C}\|\mathcal{G}(\theta)\|_2, \quad \mathfrak{c} = |\mathcal{L}(\Theta_0) - \mathcal{L}(\vartheta)|, \quad \mathfrak{C}(1-\alpha)^{-1} \mathfrak{c}^{1-\alpha} + \|\Theta_0 - \vartheta\|_2 < \varepsilon, \quad (9.183)$$

and $\inf_{t \in \{s \in [0, \infty) : \forall r \in [0, s] : \|\Theta_r - \vartheta\|_2 < \varepsilon\}} \mathcal{L}(\Theta_t) \geq \mathcal{L}(\vartheta)$ (cf. Definition 3.3.4). Then there exists $\psi \in \mathbb{R}^{\mathfrak{d}}$ such that for all $t \in [0, \infty)$ it holds that $\mathcal{L}(\psi) = \mathcal{L}(\vartheta)$, $\|\Theta_t - \vartheta\|_2 < \varepsilon$, $0 \leq \mathcal{L}(\Theta_t) - \mathcal{L}(\psi) \leq (1 + \mathfrak{C}^{-2}t)^{-1}$, and

$$\|\Theta_t - \psi\|_2 \leq \int_t^\infty \|\mathcal{G}(\Theta_s)\|_2 ds \leq \mathfrak{C}(1-\alpha)^{-1} (1 + \mathfrak{C}^{-2}t)^{\alpha-1}. \quad (9.184)$$

Proof of Corollary 9.12.3. Observe that Proposition 9.12.2 shows that there exists $\psi \in \mathbb{R}^{\mathfrak{d}}$ which satisfies that

- (i) it holds that $\mathcal{L}(\psi) = \mathcal{L}(\vartheta)$,
- (ii) it holds for all $t \in [0, \infty)$ that $\|\Theta_t - \vartheta\|_2 < \varepsilon$,
- (iii) it holds for all $t \in [0, \infty)$ that $0 \leq \mathcal{L}(\Theta_t) - \mathcal{L}(\psi) \leq \mathfrak{C}^2 \mathfrak{c}^2 (\mathbb{1}_{\{0\}}(\mathfrak{c}) + \mathfrak{C}^2 \mathfrak{c} + \mathfrak{c}^{2\alpha} t)^{-1}$, and

(iv) it holds for all $t \in [0, \infty)$ that

$$\begin{aligned}\|\Theta_t - \psi\|_2 &\leq \int_t^\infty \|\mathcal{G}(\Theta_s)\|_2 ds \leq \mathfrak{C}(1-\alpha)^{-1}[\mathcal{L}(\Theta_t) - \mathcal{L}(\psi)]^{1-\alpha} \\ &\leq \mathfrak{C}^{3-2\alpha} \mathfrak{c}^{2-2\alpha} (1-\alpha)^{-1} (\mathbb{1}_{\{0\}}(\mathfrak{c}) + \mathfrak{C}^2 \mathfrak{c} + \mathfrak{c}^{2\alpha} t)^{\alpha-1}.\end{aligned}\quad (9.185)$$

Note that item (iii) and the assumption that $\mathfrak{c} \leq 1$ ensure that for all $t \in [0, \infty)$ it holds that

$$0 \leq \mathcal{L}(\Theta_t) - \mathcal{L}(\psi) \leq \mathfrak{c}^2 (\mathfrak{C}^{-2} \mathbb{1}_{\{0\}}(\mathfrak{c}) + \mathfrak{c} + \mathfrak{C}^{-2} \mathfrak{c}^{2\alpha} t)^{-1} \leq (1 + \mathfrak{C}^{-2} t)^{-1}. \quad (9.186)$$

This and item (iv) establish that for all $t \in [0, \infty)$ it holds that

$$\begin{aligned}\|\Theta_t - \psi\|_2 &\leq \int_t^\infty \|\mathcal{G}(\Theta_s)\|_2 ds \leq \mathfrak{C}(1-\alpha)^{-1}[\mathcal{L}(\Theta_t) - \mathcal{L}(\psi)]^{1-\alpha} \\ &\leq \mathfrak{C}(1-\alpha)^{-1} (1 + \mathfrak{C}^{-2} t)^{\alpha-1}.\end{aligned}\quad (9.187)$$

Combining this with item (i), item (ii), and (9.186) proves (9.184). The proof of Corollary 9.12.3 is thus complete. \square

9.12.2 Abstract global convergence results for GF processes

Proposition 9.12.4. *Let $\mathfrak{d} \in \mathbb{N}$, $\Theta \in C([0, \infty), \mathbb{R}^{\mathfrak{d}})$, let $\mathcal{L} \in C(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ be a standard KL function, let $\mathcal{G}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}^{\mathfrak{d}}$ be $\mathcal{B}(\mathbb{R}^{\mathfrak{d}})/\mathcal{B}(\mathbb{R}^{\mathfrak{d}})$ -measurable, assume for all $t \in [0, \infty)$ that*

$$\mathcal{L}(\Theta_t) = \mathcal{L}(\Theta_0) - \int_0^t \|\mathcal{G}(\Theta_s)\|_2^2 ds \quad \text{and} \quad \Theta_t = \Theta_0 - \int_0^t \mathcal{G}(\Theta_s) ds, \quad (9.188)$$

and assume $\liminf_{t \rightarrow \infty} \|\Theta_t\|_2 < \infty$ (cf. Definitions 3.3.4 and 9.1.2). Then there exist $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, $\mathfrak{C}, \tau, \beta \in (0, \infty)$ such that for all $t \in [\tau, \infty)$ it holds that

$$\|\Theta_t - \vartheta\|_2 \leq (1 + \mathfrak{C}(t - \tau))^{-\beta} \quad \text{and} \quad 0 \leq \mathcal{L}(\Theta_t) - \mathcal{L}(\vartheta) \leq (1 + \mathfrak{C}(t - \tau))^{-1}. \quad (9.189)$$

Proof of Proposition 9.12.4. Observe that (9.188) demonstrates that $[0, \infty) \ni t \mapsto \mathcal{L}(\Theta_t) \in \mathbb{R}$ is non-increasing. Hence, we obtain that there exists $\mathbf{m} \in [-\infty, \infty)$ which satisfies

$$\mathbf{m} = \limsup_{t \rightarrow \infty} \mathcal{L}(\Theta_t) = \liminf_{t \rightarrow \infty} \mathcal{L}(\Theta_t) = \inf_{t \in [0, \infty)} \mathcal{L}(\Theta_t). \quad (9.190)$$

Furthermore, note that the assumption that $\liminf_{t \rightarrow \infty} \|\Theta_t\|_2 < \infty$ implies that there exist $\vartheta \in \mathbb{R}^{\mathfrak{d}}$ and $\delta = (\delta_n)_{n \in \mathbb{N}}: \mathbb{N} \rightarrow [0, \infty)$ which satisfy

$$\liminf_{n \rightarrow \infty} \delta_n = \infty \quad \text{and} \quad \limsup_{n \rightarrow \infty} \|\Theta_{\delta_n} - \vartheta\|_2 = 0. \quad (9.191)$$

Observe that (9.190), (9.191), and the fact that \mathcal{L} is continuous show that

$$\mathcal{L}(\vartheta) = \mathbf{m} \in \mathbb{R} \quad \text{and} \quad \forall t \in [0, \infty) : \mathcal{L}(\Theta_t) \geq \mathcal{L}(\vartheta). \quad (9.192)$$

Next let $\varepsilon, \mathfrak{C} \in (0, \infty)$, $\alpha \in (0, 1)$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ with $\|\theta - \vartheta\|_2 < \varepsilon$ that

$$|\mathcal{L}(\theta) - \mathcal{L}(\vartheta)|^\alpha \leq \mathfrak{C} \|\mathcal{G}(\theta)\|_2. \quad (9.193)$$

Note that (9.191) and the fact that \mathcal{L} is continuous demonstrate that there exist $n \in \mathbb{N}$, $\mathfrak{c} \in [0, 1]$ which satisfy

$$\mathfrak{c} = |\mathcal{L}(\Theta_{\delta_n}) - \mathcal{L}(\vartheta)| \quad \text{and} \quad \mathfrak{C}(1 - \alpha)^{-1} \mathfrak{c}^{1-\alpha} + \|\Theta_{\delta_n} - \vartheta\|_2 < \varepsilon. \quad (9.194)$$

Next let $\Phi : [0, \infty) \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $t \in [0, \infty)$ that

$$\Phi_t = \Theta_{\delta_n+t}. \quad (9.195)$$

Observe that (9.188), (9.192), and (9.195) ensure that for all $t \in [0, \infty)$ it holds that

$$\mathcal{L}(\Phi_t) = \mathcal{L}(\Phi_0) - \int_0^t \|\mathcal{G}(\Phi_s)\|_2^2 ds, \quad \Phi_t = \Phi_0 - \int_0^t \mathcal{G}(\Phi_s) ds, \quad \text{and} \quad \mathcal{L}(\Phi_t) \geq \mathcal{L}(\vartheta). \quad (9.196)$$

Combining this with (9.193), (9.194), (9.195), and Corollary 9.12.3 (applied with $\Theta \curvearrowright \Phi$ in the notation of Corollary 9.12.3) establishes that there exists $\psi \in \mathbb{R}^{\mathfrak{d}}$ which satisfies for all $t \in [0, \infty)$ that

$$0 \leq \mathcal{L}(\Phi_t) - \mathcal{L}(\psi) \leq (1 + \mathfrak{C}^{-2}t)^{-1}, \quad \|\Phi_t - \psi\|_2 \leq \mathfrak{C}(1 - \alpha)^{-1}(1 + \mathfrak{C}^{-2}t)^{\alpha-1}, \quad (9.197)$$

and $\mathcal{L}(\psi) = \mathcal{L}(\vartheta)$. Note that (9.195) and (9.197) establish for all $t \in [0, \infty)$ that $0 \leq \mathcal{L}(\Theta_{\delta_n+t}) - \mathcal{L}(\psi) \leq (1 + \mathfrak{C}^{-2}t)^{-1}$ and $\|\Theta_{\delta_n+t} - \psi\|_2 \leq \mathfrak{C}(1 - \alpha)^{-1}(1 + \mathfrak{C}^{-2}t)^{\alpha-1}$. Therefore, we obtain for all $\tau \in [\delta_n, \infty)$, $t \in [\tau, \infty)$ that

$$\begin{aligned} 0 \leq \mathcal{L}(\Theta_t) - \mathcal{L}(\psi) &\leq (1 + \mathfrak{C}^{-2}(t - \delta_n))^{-1} = (1 + \mathfrak{C}^{-2}(t - \tau) + \mathfrak{C}^{-2}(\tau - \delta_n))^{-1} \\ &\leq (1 + \mathfrak{C}^{-2}(t - \tau))^{-1} \end{aligned} \quad (9.198)$$

and

$$\begin{aligned} \|\Theta_t - \psi\|_2 &\leq \mathfrak{C}(1 - \alpha)^{-1}(1 + \mathfrak{C}^{-2}(t - \delta_n))^{\alpha-1} \\ &= \left[[\mathfrak{C}(1 - \alpha)^{-1}]^{\frac{1}{\alpha-1}} (1 + \mathfrak{C}^{-2}(t - \delta_n)) \right]^{\alpha-1} \\ &= \left[[\mathfrak{C}(1 - \alpha)^{-1}]^{\frac{1}{\alpha-1}} [1 + \mathfrak{C}^{-2}(\tau - \delta_n)] + \left[[\mathfrak{C}(1 - \alpha)^{-1}]^{\frac{1}{1-\alpha}} \mathfrak{C}^2 \right]^{-1} (t - \tau) \right]^{\alpha-1}. \end{aligned} \quad (9.199)$$

Next let $\mathcal{C}, \tau \in (0, \infty)$ satisfy

$$\mathcal{C} = \max\{\mathfrak{C}^2, [\mathfrak{C}(1 - \alpha)^{-1}]^{\frac{1}{1-\alpha}} \mathfrak{C}^2\} \quad \text{and} \quad \tau = \delta_n + \mathfrak{C}^2 [\mathfrak{C}(1 - \alpha)^{-1}]^{\frac{1}{1-\alpha}}. \quad (9.200)$$

Observe that (9.198), (9.199), and (9.200) demonstrate for all $t \in [\tau, \infty)$ that

$$0 \leq \mathcal{L}(\Theta_t) - \mathcal{L}(\psi) \leq (1 + \mathfrak{C}^{-2}(t - \tau))^{-1} \leq (1 + \mathcal{C}^{-1}(t - \tau))^{-1} \quad (9.201)$$

and

$$\begin{aligned} \|\Theta_t - \psi\|_2 &\leq \left[[\mathfrak{C}(1 - \alpha)^{-1}]^{\frac{1}{\alpha-1}} [1 + \mathfrak{C}^{-2}(\tau - \delta_n)] + \mathcal{C}^{-1}(t - \tau) \right]^{\alpha-1} \\ &= \left[[\mathfrak{C}(1 - \alpha)^{-1}]^{\frac{1}{\alpha-1}} [1 + [\mathfrak{C}(1 - \alpha)^{-1}]^{\frac{1}{1-\alpha}}] + \mathcal{C}^{-1}(t - \tau) \right]^{\alpha-1} \\ &\leq [1 + \mathcal{C}^{-1}(t - \tau)]^{\alpha-1}. \end{aligned} \quad (9.202)$$

The proof of Proposition 9.12.4 is thus complete. \square

Corollary 9.12.5. Let $\mathfrak{d} \in \mathbb{N}$, $\Theta \in C([0, \infty), \mathbb{R}^\mathfrak{d})$, let $\mathcal{L} \in C(\mathbb{R}^\mathfrak{d}, \mathbb{R})$ be a standard **KL** function, let $\mathcal{G}: \mathbb{R}^\mathfrak{d} \rightarrow \mathbb{R}^\mathfrak{d}$ be $\mathcal{B}(\mathbb{R}^\mathfrak{d})/\mathcal{B}(\mathbb{R}^\mathfrak{d})$ -measurable, assume for all $t \in [0, \infty)$ that

$$\mathcal{L}(\Theta_t) = \mathcal{L}(\Theta_0) - \int_0^t \|\mathcal{G}(\Theta_s)\|_2^2 ds \quad \text{and} \quad \Theta_t = \Theta_0 - \int_0^t \mathcal{G}(\Theta_s) ds, \quad (9.203)$$

and assume $\liminf_{t \rightarrow \infty} \|\Theta_t\|_2 < \infty$ (cf. Definitions 3.3.4 and 9.1.2). Then there exist $\vartheta \in \mathbb{R}^\mathfrak{d}$, $\mathcal{C}, \beta \in (0, \infty)$ which satisfy for all $t \in [0, \infty)$ that

$$\|\Theta_t - \vartheta\|_2 \leq \mathcal{C}(1 + t)^{-\beta} \quad \text{and} \quad 0 \leq \mathcal{L}(\Theta_t) - \mathcal{L}(\vartheta) \leq \mathcal{C}(1 + t)^{-1}. \quad (9.204)$$

Proof of Corollary 9.12.5. Note that Proposition 9.12.4 demonstrates that there exist $\vartheta \in \mathbb{R}^\mathfrak{d}$, $\mathfrak{C}, \tau, \beta \in (0, \infty)$ which satisfy for all $t \in [\tau, \infty)$ that

$$\|\Theta_t - \vartheta\|_2 \leq (1 + \mathfrak{C}(t - \tau))^{-\beta} \quad \text{and} \quad 0 \leq \mathcal{L}(\Theta_t) - \mathcal{L}(\vartheta) \leq (1 + \mathfrak{C}(t - \tau))^{-1}. \quad (9.205)$$

In the following let $\mathcal{C} \in (0, \infty)$ satisfy

$$\mathcal{C} = \max \left\{ 1 + \tau, (1 + \tau)^\beta, \mathfrak{C}^{-1}, \mathfrak{C}^{-\beta}, (1 + \tau)^\beta (\sup_{s \in [0, \tau]} \|\Theta_s - \vartheta\|_2), (1 + \tau)(\mathcal{L}(\Theta_0) - \mathcal{L}(\vartheta)) \right\}. \quad (9.206)$$

Observe that (9.205), (9.206), and the fact that $[0, \infty) \ni t \mapsto \mathcal{L}(\Theta_t) \in \mathbb{R}$ is non-increasing prove for all $t \in [0, \tau]$ that

$$\|\Theta_t - \vartheta\|_2 \leq \sup_{s \in [0, \tau]} \|\Theta_s - \vartheta\|_2 \leq \mathcal{C}(1 + \tau)^{-\beta} \leq \mathcal{C}(1 + t)^{-\beta} \quad (9.207)$$

and

$$0 \leq \mathcal{L}(\Theta_t) - \mathcal{L}(\vartheta) \leq \mathcal{L}(\Theta_0) - \mathcal{L}(\vartheta) \leq \mathcal{C}(1 + \tau)^{-1} \leq \mathcal{C}(1 + t)^{-1}. \quad (9.208)$$

Furthermore, note that (9.205) and (9.206) demonstrate for all $t \in [\tau, \infty)$ that

$$\begin{aligned} \|\Theta_t - \vartheta\|_2 &\leq (1 + \mathcal{C}(t - \tau))^{-\beta} = \mathcal{C}(\mathcal{C}^{1/\beta} + \mathcal{C}^{1/\beta}\mathcal{C}(t - \tau))^{-\beta} \\ &\leq \mathcal{C}(\mathcal{C}^{1/\beta} + t - \tau)^{-\beta} \leq \mathcal{C}(1 + t)^{-\beta}. \end{aligned} \quad (9.209)$$

Moreover, observe that (9.205) and (9.206) demonstrate for all $t \in [\tau, \infty)$ that

$$0 \leq \mathcal{L}(\Theta_t) - \mathcal{L}(\vartheta) \leq \mathcal{C}(\mathcal{C} + \mathcal{C}\mathcal{C}(t - \tau))^{-1} \leq \mathcal{C}(\mathcal{C} - \tau + t)^{-1} \leq \mathcal{C}(1 + t)^{-1}. \quad (9.210)$$

The proof of Corollary 9.12.5 is thus complete. \square

Corollary 9.12.6. *Let $\mathfrak{d} \in \mathbb{N}$, $\Theta \in C([0, \infty), \mathbb{R}^{\mathfrak{d}})$, let $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ be a standard KL function, assume for all $t \in [0, \infty)$ that*

$$\Theta_t = \Theta_0 - \int_0^t (\nabla \mathcal{L})(\Theta_s) ds, \quad (9.211)$$

and assume $\liminf_{t \rightarrow \infty} \|\Theta_t\|_2 < \infty$ (cf. Definitions 3.3.4 and 9.1.2). Then there exist $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, $\mathcal{C}, \beta \in (0, \infty)$ which satisfy for all $t \in [0, \infty)$ that

$$\|\Theta_t - \vartheta\|_2 \leq \mathcal{C}(1 + t)^{-\beta}, \quad 0 \leq \mathcal{L}(\Theta_t) - \mathcal{L}(\vartheta) \leq \mathcal{C}(1 + t)^{-1}, \quad \text{and} \quad (\nabla \mathcal{L})(\vartheta) = 0. \quad (9.212)$$

Proof of Corollary 9.12.6. Note that Lemma 9.12.1 implies that for all $t \in [0, \infty)$ it holds that

$$\mathcal{L}(\Theta_t) = \mathcal{L}(\Theta_0) - \int_0^t \|(\nabla \mathcal{L})(\Theta_s)\|_2^2 ds. \quad (9.213)$$

Corollary 9.12.5 hence establishes that there exist $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, $\mathcal{C}, \beta \in (0, \infty)$ which satisfy for all $t \in [0, \infty)$ that

$$\|\Theta_t - \vartheta\|_2 \leq \mathcal{C}(1 + t)^{-\beta} \quad \text{and} \quad 0 \leq \mathcal{L}(\Theta_t) - \mathcal{L}(\vartheta) \leq \mathcal{C}(1 + t)^{-1}. \quad (9.214)$$

This shows that

$$\limsup_{t \rightarrow \infty} \|\Theta_t - \vartheta\|_2 = 0. \quad (9.215)$$

Combining this with the assumption that $\mathcal{L} \in C^1(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ ensures that

$$\limsup_{t \rightarrow \infty} \|(\nabla \mathcal{L})(\Theta_t) - (\nabla \mathcal{L})(\vartheta)\|_2 = 0. \quad (9.216)$$

Therefore, we obtain that

$$\limsup_{t \rightarrow \infty} | \|(\nabla \mathcal{L})(\Theta_t)\|_2 - \|(\nabla \mathcal{L})(\vartheta)\|_2 | = 0. \quad (9.217)$$

Furthermore, observe that (9.213) and (9.214) ensure that

$$\int_0^\infty \|(\nabla \mathcal{L})(\Theta_s)\|_2^2 ds < \infty. \quad (9.218)$$

This and (9.217) demonstrate that

$$(\nabla \mathcal{L})(\vartheta) = 0. \quad (9.219)$$

Combining this with (9.214) establishes (9.212). The proof of Corollary 9.12.6 is thus complete. \square

Corollary 9.12.7. Let $\mathfrak{d} \in \mathbb{N}$, $\Theta \in C([0, \infty), \mathbb{R}^\mathfrak{d})$, let $\mathcal{L}: \mathbb{R}^\mathfrak{d} \rightarrow \mathbb{R}$ be analytic, assume for all $t \in [0, \infty)$ that

$$\Theta_t = \Theta_0 - \int_0^t (\nabla \mathcal{L})(\Theta_s) ds, \quad (9.220)$$

and assume $\liminf_{t \rightarrow \infty} \|\Theta_t\|_2 < \infty$ (cf. Definitions 3.3.4 and 9.8.1). Then there exist $\vartheta \in \mathbb{R}^\mathfrak{d}$, $\mathcal{C}, \beta \in (0, \infty)$ which satisfy for all $t \in [0, \infty)$ that

$$\|\Theta_t - \vartheta\|_2 \leq \mathcal{C}(1+t)^{-\beta}, \quad 0 \leq \mathcal{L}(\Theta_t) - \mathcal{L}(\vartheta) \leq \mathcal{C}(1+t)^{-1}, \quad \text{and} \quad (\nabla \mathcal{L})(\vartheta) = 0. \quad (9.221)$$

Proof of Corollary 9.12.7. Note that Theorem 9.10.1 establishes that for all $\vartheta \in \mathbb{R}^\mathfrak{d}$ there exist $\varepsilon, \mathfrak{C} \in (0, \infty)$, $\alpha \in (0, 1)$ such that for all $\theta \in \mathbb{R}^\mathfrak{d}$ with $\|\theta - \vartheta\|_2 < \varepsilon$ it holds that

$$|\mathcal{L}(\theta) - \mathcal{L}(\vartheta)|^\alpha \leq \mathfrak{C} \|(\nabla \mathcal{L})(\theta)\|_2. \quad (9.222)$$

Corollary 9.12.6 hence establishes (9.221). The proof of Corollary 9.12.7 is thus complete. \square

Exercise 9.12.1. Prove or disprove the following statement: For all $\mathfrak{d} \in \mathbb{N}$, $L \in (0, \infty)$, $\gamma \in [0, L^{-1}]$, all open and convex sets $U \subseteq \mathbb{R}^\mathfrak{d}$, and all $\mathcal{L} \in C^1(U, \mathbb{R})$, $\theta \in U$ with $\theta - \gamma(\nabla \mathcal{L})(\theta) \in U$ and $\forall v, w \in U: \|(\nabla \mathcal{L})(v) - (\nabla \mathcal{L})(w)\|_2 \leq L\|v - w\|_2$ it holds that

$$\mathcal{L}(\theta - \gamma(\nabla \mathcal{L})(\theta)) \leq \mathcal{L}(\theta) - \frac{\gamma}{2} \|(\nabla \mathcal{L})(\theta)\|_2^2 \quad (9.223)$$

(cf. Definition 3.3.4).

9.13 Convergence analysis for GD processes

In this section we employ standard **KL** inequalities to establish convergence of the **GD** method to critical points (see Section 5.6.3). The specific presentation of this section is closely based on [234, Section 8].

9.13.1 One-step descent property for GD processes

Lemma 9.13.1. Let $\mathfrak{d} \in \mathbb{N}$, $L \in \mathbb{R}$, let $U \subseteq \mathbb{R}^{\mathfrak{d}}$ be open and convex, let $\mathcal{L} \in C^1(U, \mathbb{R})$, and assume for all $v, w \in U$ that

$$\|(\nabla \mathcal{L})(v) - (\nabla \mathcal{L})(w)\|_2 \leq L\|v - w\|_2 \quad (9.224)$$

(cf. Definition 3.3.4). Then it holds for all $v, w \in U$ that

$$\mathcal{L}(w) \leq \mathcal{L}(v) + \langle (\nabla \mathcal{L})(v), w - v \rangle + \frac{L}{2}\|v - w\|_2^2 \quad (9.225)$$

(cf. Definition 1.4.7).

Proof of Lemma 9.13.1. Observe that the fundamental theorem of calculus, the Cauchy-Schwarz inequality, and (9.224) prove that for all $v, w \in U$ we have that

$$\begin{aligned} & \mathcal{L}(w) - \mathcal{L}(v) \\ &= [\mathcal{L}(v + r(w - v))]_{r=0}^{r=1} = \int_0^1 \langle (\nabla \mathcal{L})(v + r(w - v)), w - v \rangle dr \\ &= \langle (\nabla \mathcal{L})(v), w - v \rangle + \int_0^1 \langle (\nabla \mathcal{L})(v + r(w - v)) - (\nabla \mathcal{L})(v), w - v \rangle dr \\ &\leq \langle (\nabla \mathcal{L})(v), w - v \rangle + \int_0^1 |\langle (\nabla \mathcal{L})(v + r(w - v)) - (\nabla \mathcal{L})(v), w - v \rangle| dr \quad (9.226) \\ &\leq \langle (\nabla \mathcal{L})(v), w - v \rangle + \left[\int_0^1 \|(\nabla \mathcal{L})(v + r(w - v)) - (\nabla \mathcal{L})(v)\|_2 dr \right] \|w - v\|_2 \\ &\leq \langle (\nabla \mathcal{L})(v), w - v \rangle + L\|w - v\|_2 \left[\int_0^1 \|r(w - v)\|_2 dr \right] \\ &= \langle (\nabla \mathcal{L})(v), w - v \rangle + \frac{L}{2}\|v - w\|_2^2 \end{aligned}$$

(cf. Definition 1.4.7). The proof of Lemma 9.13.1 is thus complete. \square

Corollary 9.13.2. Let $\mathfrak{d} \in \mathbb{N}$, $L, \gamma \in \mathbb{R}$, let $U \subseteq \mathbb{R}^{\mathfrak{d}}$ be open and convex, let $\mathcal{L} \in C^1(U, \mathbb{R})$, and assume for all $v, w \in U$ that

$$\|(\nabla \mathcal{L})(v) - (\nabla \mathcal{L})(w)\|_2 \leq L\|v - w\|_2 \quad (9.227)$$

(cf. Definition 3.3.4). Then it holds for all $\theta \in U$ with $\theta - \gamma(\nabla \mathcal{L})(\theta) \in U$ that

$$\mathcal{L}(\theta - \gamma(\nabla \mathcal{L})(\theta)) \leq \mathcal{L}(\theta) + \gamma\left(\frac{L\gamma}{2} - 1\right)\|(\nabla \mathcal{L})(\theta)\|_2^2. \quad (9.228)$$

Proof of Corollary 9.13.2. Observe that Lemma 9.13.1 ensures that for all $\theta \in U$ with $\theta - \gamma(\nabla \mathcal{L})(\theta) \in U$ it holds that

$$\begin{aligned}\mathcal{L}(\theta - \gamma(\nabla \mathcal{L})(\theta)) &\leq \mathcal{L}(\theta) + \langle (\nabla \mathcal{L})(\theta), -\gamma(\nabla \mathcal{L})(\theta) \rangle + \frac{L}{2} \|\gamma(\nabla \mathcal{L})(\theta)\|_2^2 \\ &= \mathcal{L}(\theta) - \gamma \|(\nabla \mathcal{L})(\theta)\|_2^2 + \frac{L\gamma^2}{2} \|(\nabla \mathcal{L})(\theta)\|_2^2.\end{aligned}\quad (9.229)$$

This establishes (9.228). The proof of Corollary 9.13.2 is thus complete. \square

Corollary 9.13.3. Let $\mathfrak{d} \in \mathbb{N}$, $L \in (0, \infty)$, $\gamma \in [0, L^{-1}]$, let $U \subseteq \mathbb{R}^\mathfrak{d}$ be open and convex, let $\mathcal{L} \in C^1(U, \mathbb{R})$, and assume for all $v, w \in U$ that

$$\|(\nabla \mathcal{L})(v) - (\nabla \mathcal{L})(w)\|_2 \leq L\|v - w\|_2 \quad (9.230)$$

(cf. Definition 3.3.4). Then it holds for all $\theta \in U$ with $\theta - \gamma(\nabla \mathcal{L})(\theta) \in U$ that

$$\mathcal{L}(\theta - \gamma(\nabla \mathcal{L})(\theta)) \leq \mathcal{L}(\theta) - \frac{\gamma}{2} \|(\nabla \mathcal{L})(\theta)\|_2^2 \leq \mathcal{L}(\theta). \quad (9.231)$$

Proof of Corollary 9.13.3. Note that Corollary 9.13.2, the fact that $\gamma \geq 0$, and the fact that $\frac{L\gamma}{2} - 1 \leq -\frac{1}{2}$ establish (9.231). The proof of Corollary 9.13.3 is thus complete. \square

Exercise 9.13.1. Let $(\gamma_n)_{n \in \mathbb{N}} \subseteq (0, \infty)$ satisfy for all $n \in \mathbb{N}$ that $\gamma_n = \frac{1}{n+1}$ and let $\mathcal{L}: \mathbb{R} \rightarrow \mathbb{R}$ satisfy for all $\theta \in \mathbb{R}$ that

$$\mathcal{L}(\theta) = 2\theta + \sin(\theta). \quad (9.232)$$

Prove or disprove the following statement: For every $\Theta = (\Theta_k)_{k \in \mathbb{N}_0}: \mathbb{N}_0 \rightarrow \mathbb{R}$ with $\forall k \in \mathbb{N}: \Theta_k = \Theta_{k-1} - \gamma_k(\nabla \mathcal{L})(\Theta_{k-1})$ and every $n \in \mathbb{N}$ it holds that

$$\mathcal{L}(\Theta_n) \leq \mathcal{L}(\Theta_{n-1}) - \frac{1}{n+1} \left(1 - \frac{3}{2(n+1)}\right) |2 + \cos(\Theta_{n-1})|^2. \quad (9.233)$$

Exercise 9.13.2. Let $\mathcal{L}: \mathbb{R} \rightarrow \mathbb{R}$ satisfy for all $\theta \in \mathbb{R}$ that

$$\mathcal{L}(\theta) = 4\theta + 3\sin(\theta). \quad (9.234)$$

Prove or disprove the following statement: For every $\Theta = (\Theta_n)_{n \in \mathbb{N}_0}: \mathbb{N}_0 \rightarrow \mathbb{R}$ with $\forall n \in \mathbb{N}: \Theta_n = \Theta_{n-1} - \frac{1}{n+1}(\nabla \mathcal{L})(\Theta_{n-1})$ and every $k \in \mathbb{N}$ it holds that

$$\mathcal{L}(\Theta_k) < \mathcal{L}(\Theta_{k-1}). \quad (9.235)$$

9.13.2 Abstract local convergence results for GD processes

Proposition 9.13.4. Let $\mathfrak{d} \in \mathbb{N}$, $\mathfrak{c} \in \mathbb{R}$, $\varepsilon, L, \mathfrak{C} \in (0, \infty)$, $\alpha \in (0, 1)$, $\gamma \in (0, L^{-1}]$, $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, let $\mathbb{B} \subseteq \mathbb{R}^{\mathfrak{d}}$ satisfy $\mathbb{B} = \{\theta \in \mathbb{R}^{\mathfrak{d}} : \|\theta - \vartheta\|_2 < \varepsilon\}$, let $\mathcal{L} \in C(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ satisfy $\mathcal{L}|_{\mathbb{B}} \in C^1(\mathbb{B}, \mathbb{R})$, let $\mathcal{G}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $\theta \in \mathbb{B}$ that $\mathcal{G}(\theta) = (\nabla \mathcal{L})(\theta)$, assume $\mathcal{G}(\vartheta) = 0$, assume for all $v, w \in \mathbb{B}$ that

$$\|\mathcal{G}(v) - \mathcal{G}(w)\|_2 \leq L\|v - w\|_2, \quad (9.236)$$

let $\Theta: \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $n \in \mathbb{N}_0$ that $\Theta_{n+1} = \Theta_n - \gamma \mathcal{G}(\Theta_n)$, and assume for all $\theta \in \mathbb{B}$ that

$$|\mathcal{L}(\theta) - \mathcal{L}(\vartheta)|^{\alpha} \leq \mathfrak{C}\|\mathcal{G}(\theta)\|_2, \quad \mathfrak{c} = |\mathcal{L}(\Theta_0) - \mathcal{L}(\vartheta)|, \quad 2\mathfrak{C}(1-\alpha)^{-1}\mathfrak{c}^{1-\alpha} + \|\Theta_0 - \vartheta\|_2 < \frac{\varepsilon}{\gamma L + 1}, \quad (9.237)$$

and $\inf_{n \in \{m \in \mathbb{N}_0 : \forall k \in \mathbb{N}_0 \cap [0, m] : \Theta_k \in \mathbb{B}\}} \mathcal{L}(\Theta_n) \geq \mathcal{L}(\vartheta)$ (cf. Definition 3.3.4). Then there exists $\psi \in \mathcal{L}^{-1}(\{\mathcal{L}(\vartheta)\}) \cap \mathcal{G}^{-1}(\{0\}) \cap \mathbb{B}$ such that

- (i) it holds for all $n \in \mathbb{N}_0$ that $\Theta_n \in \mathbb{B}$,
- (ii) it holds for all $n \in \mathbb{N}_0$ that $0 \leq \mathcal{L}(\Theta_n) - \mathcal{L}(\psi) \leq 2\mathfrak{C}^2\mathfrak{c}^2(\mathbb{1}_{\{0\}}(\mathfrak{c}) + \mathfrak{c}^{2\alpha}n\gamma + 2\mathfrak{C}^2\mathfrak{c})^{-1}$, and
- (iii) it holds for all $n \in \mathbb{N}_0$ that

$$\begin{aligned} \|\Theta_n - \psi\|_2 &\leq \sum_{k=n}^{\infty} \|\Theta_{k+1} - \Theta_k\|_2 \leq 2\mathfrak{C}(1-\alpha)^{-1}|\mathcal{L}(\Theta_n) - \mathcal{L}(\psi)|^{1-\alpha} \\ &\leq 2^{2-\alpha}\mathfrak{C}^{3-2\alpha}\mathfrak{c}^{2-2\alpha}(1-\alpha)^{-1}(\mathbb{1}_{\{0\}}(\mathfrak{c}) + \mathfrak{c}^{2\alpha}n\gamma + 2\mathfrak{C}^2\mathfrak{c})^{\alpha-1}. \end{aligned} \quad (9.238)$$

Proof of Proposition 9.13.4. Throughout this proof, let $T \in \mathbb{N}_0 \cup \{\infty\}$ satisfy

$$T = \inf(\{n \in \mathbb{N}_0 : \Theta_n \notin \mathbb{B}\} \cup \{\infty\}), \quad (9.239)$$

let $\mathbb{L}: \mathbb{N}_0 \rightarrow \mathbb{R}$ satisfy for all $n \in \mathbb{N}_0$ that $\mathbb{L}(n) = \mathcal{L}(\Theta_n) - \mathcal{L}(\vartheta)$, and let $\tau \in \mathbb{N}_0 \cup \{\infty\}$ satisfy

$$\tau = \inf(\{n \in \mathbb{N}_0 \cap [0, T) : \mathbb{L}(n) = 0\} \cup \{T\}). \quad (9.240)$$

Observe that the assumption that $\mathcal{G}(\vartheta) = 0$ demonstrates for all $\theta \in \mathbb{B}$ that

$$\gamma\|\mathcal{G}(\theta)\|_2 = \gamma\|\mathcal{G}(\theta) - \mathcal{G}(\vartheta)\|_2 \leq \gamma L\|\theta - \vartheta\|_2. \quad (9.241)$$

This, the fact that $\|\Theta_0 - \vartheta\|_2 < \varepsilon$, and the fact that

$$\|\Theta_1 - \vartheta\|_2 \leq \|\Theta_1 - \Theta_0\|_2 + \|\Theta_0 - \vartheta\|_2 = \gamma\|\mathcal{G}(\Theta_0)\|_2 + \|\Theta_0 - \vartheta\|_2 \leq (\gamma L + 1)\|\Theta_0 - \vartheta\|_2 < \varepsilon \quad (9.242)$$

ensure that $T \geq 2$. Next note that the assumption that

$$\inf_{n \in \{m \in \mathbb{N}_0 : \forall k \in \mathbb{N}_0 \cap [0, m] : \Theta_k \in \mathbb{B}\}} \mathcal{L}(\Theta_n) \geq \mathcal{L}(\vartheta) \quad (9.243)$$

implies for all $n \in \mathbb{N}_0 \cap [0, T)$ that

$$\mathbb{L}(n) \geq 0. \quad (9.244)$$

Furthermore, observe that the fact that $\mathbb{B} \subseteq \mathbb{R}^{\mathfrak{d}}$ is open and convex, Corollary 9.13.3, and (9.237) demonstrate for all $n \in \mathbb{N}_0 \cap [0, T - 1)$ that

$$\begin{aligned} \mathbb{L}(n+1) - \mathbb{L}(n) &= \mathcal{L}(\Theta_{n+1}) - \mathcal{L}(\Theta_n) \leq -\frac{\gamma}{2} \|\mathcal{G}(\Theta_n)\|_2^2 = -\frac{1}{2} \|\mathcal{G}(\Theta_n)\|_2 \|\gamma \mathcal{G}(\Theta_n)\|_2 \\ &= -\frac{1}{2} \|\mathcal{G}(\Theta_n)\|_2 \|\Theta_{n+1} - \Theta_n\|_2 \leq -(2\mathfrak{C})^{-1} |\mathcal{L}(\Theta_n) - \mathcal{L}(\vartheta)|^\alpha \|\Theta_{n+1} - \Theta_n\|_2 \\ &= -(2\mathfrak{C})^{-1} [\mathbb{L}(n)]^\alpha \|\Theta_{n+1} - \Theta_n\|_2 \leq 0. \end{aligned} \quad (9.245)$$

Therefore, we obtain that

$$\mathbb{N}_0 \cap [0, T) \ni n \mapsto \mathbb{L}(n) \in [0, \infty) \quad (9.246)$$

is non-increasing. Combining this with (9.240) shows for all $n \in \mathbb{N}_0 \cap [\tau, T)$ that

$$\mathbb{L}(n) = 0. \quad (9.247)$$

This and (9.245) demonstrate for all $n \in \mathbb{N}_0 \cap [\tau, T - 1)$ that

$$0 = \mathbb{L}(n+1) - \mathbb{L}(n) \leq -\frac{\gamma}{2} \|\mathcal{G}(\Theta_n)\|_2^2 \leq 0. \quad (9.248)$$

The fact that $\gamma > 0$ hence ensures for all $n \in \mathbb{N}_0 \cap [\tau, T - 1)$ that $\mathcal{G}(\Theta_n) = 0$. Therefore, we obtain for all $n \in \mathbb{N}_0 \cap [\tau, T)$ that

$$\Theta_n = \Theta_\tau. \quad (9.249)$$

Moreover, note that (9.240) and (9.245) ensure for all $n \in \mathbb{N}_0 \cap [0, \tau) \cap [0, T - 1)$ that

$$\begin{aligned} \|\Theta_{n+1} - \Theta_n\|_2 &\leq \frac{2\mathfrak{C}(\mathbb{L}(n) - \mathbb{L}(n+1))}{[\mathbb{L}(n)]^\alpha} = 2\mathfrak{C} \int_{\mathbb{L}(n+1)}^{\mathbb{L}(n)} [\mathbb{L}(n)]^{-\alpha} du \\ &\leq 2\mathfrak{C} \int_{\mathbb{L}(n+1)}^{\mathbb{L}(n)} u^{-\alpha} du = \frac{2\mathfrak{C}([\mathbb{L}(n)]^{1-\alpha} - [\mathbb{L}(n+1)]^{1-\alpha})}{1-\alpha}. \end{aligned} \quad (9.250)$$

This and (9.249) establish for all $n \in \mathbb{N}_0 \cap [0, T - 1)$ that

$$\|\Theta_{n+1} - \Theta_n\|_2 \leq \frac{2\mathfrak{C}([\mathbb{L}(n)]^{1-\alpha} - [\mathbb{L}(n+1)]^{1-\alpha})}{1-\alpha}. \quad (9.251)$$

Combining this with the triangle inequality proves for all $m, n \in \mathbb{N}_0 \cap [0, T)$ with $m \leq n$ that

$$\begin{aligned} \|\Theta_n - \Theta_m\|_2 &\leq \sum_{k=m}^{n-1} \|\Theta_{k+1} - \Theta_k\|_2 \leq \frac{2\mathfrak{C}}{1-\alpha} \left[\sum_{k=m}^{n-1} ([\mathbb{L}(k)]^{1-\alpha} - [\mathbb{L}(k+1)]^{1-\alpha}) \right] \\ &= \frac{2\mathfrak{C}([\mathbb{L}(m)]^{1-\alpha} - [\mathbb{L}(n)]^{1-\alpha})}{1-\alpha} \leq \frac{2\mathfrak{C}[\mathbb{L}(m)]^{1-\alpha}}{1-\alpha}. \end{aligned} \quad (9.252)$$

This and (9.237) demonstrate for all $n \in \mathbb{N}_0 \cap [0, T)$ that

$$\|\Theta_n - \Theta_0\|_2 \leq \frac{2\mathfrak{C}[\mathbb{L}(0)]^{1-\alpha}}{1-\alpha} = \frac{2\mathfrak{C}|\mathcal{L}(\Theta_0) - \mathcal{L}(\vartheta)|^{1-\alpha}}{1-\alpha} = 2\mathfrak{C}(1-\alpha)^{-1}\mathfrak{c}^{1-\alpha}. \quad (9.253)$$

Combining this with (9.241), (9.237), and the triangle inequality demonstrates for all $n \in \mathbb{N}_0 \cap [0, T)$ that

$$\begin{aligned} \|\Theta_{n+1} - \vartheta\|_2 &\leq \|\Theta_{n+1} - \Theta_n\|_2 + \|\Theta_n - \vartheta\|_2 = \gamma\|\mathcal{G}(\Theta_n)\|_2 + \|\Theta_n - \vartheta\|_2 \\ &\leq (\gamma L + 1)\|\Theta_n - \vartheta\|_2 \leq (\gamma L + 1)(\|\Theta_n - \Theta_0\|_2 + \|\Theta_0 - \vartheta\|_2) \\ &\leq (\gamma L + 1)(2\mathfrak{C}(1-\alpha)^{-1}\mathfrak{c}^{1-\alpha} + \|\Theta_0 - \vartheta\|_2) < \varepsilon. \end{aligned} \quad (9.254)$$

Hence, we obtain that

$$T = \infty. \quad (9.255)$$

Combining this with (9.237), (9.252), and (9.246) implies that

$$\sum_{k=0}^{\infty} \|\Theta_{k+1} - \Theta_k\|_2 = \lim_{n \rightarrow \infty} \left[\sum_{k=0}^n \|\Theta_{k+1} - \Theta_k\|_2 \right] \leq \frac{2\mathfrak{C}[\mathbb{L}(0)]^{1-\alpha}}{1-\alpha} = \frac{2\mathfrak{C}\mathfrak{c}^{1-\alpha}}{1-\alpha} < \varepsilon < \infty. \quad (9.256)$$

Therefore, we obtain that there exists $\psi \in \mathbb{R}^{\mathfrak{d}}$ which satisfies

$$\limsup_{n \rightarrow \infty} \|\Theta_n - \psi\|_2 = 0. \quad (9.257)$$

Observe that (9.254), (9.255), and (9.257) show that

$$\|\psi - \vartheta\|_2 \leq (\gamma L + 1)(2\mathfrak{C}(1-\alpha)^{-1}\mathfrak{c}^{1-\alpha} + \|\Theta_0 - \vartheta\|_2) < \varepsilon. \quad (9.258)$$

Hence, we obtain that

$$\psi \in \mathbb{B}. \quad (9.259)$$

Next note that (9.245), (9.237), and the fact that for all $n \in \mathbb{N}_0$ it holds that $\mathbb{L}(n) \leq \mathbb{L}(0) = \mathfrak{c}$ ensure that for all $n \in \mathbb{N}_0 \cap [0, \tau)$ we have that

$$-\mathbb{L}(n) \leq \mathbb{L}(n+1) - \mathbb{L}(n) \leq -\frac{\gamma}{2}\|\mathcal{G}(\Theta_n)\|_2^2 \leq -\frac{\gamma}{2\mathfrak{C}^2}[\mathbb{L}(n)]^{2\alpha} \leq -\frac{\gamma}{2\mathfrak{C}^2\mathfrak{c}^{2-2\alpha}}[\mathbb{L}(n)]^2. \quad (9.260)$$

This ensures for all $n \in \mathbb{N}_0 \cap [0, \tau)$ that

$$0 < \mathbb{L}(n) \leq \frac{2\mathfrak{C}^2\mathfrak{c}^{2-2\alpha}}{\gamma}. \quad (9.261)$$

Combining this and (9.260) demonstrates for all $n \in \mathbb{N}_0 \cap [0, \tau - 1)$ that

$$\begin{aligned} \frac{1}{\mathbb{L}(n)} - \frac{1}{\mathbb{L}(n+1)} &\leq \frac{1}{\mathbb{L}(n)} - \frac{1}{\mathbb{L}(n)(1 - \frac{\gamma}{2\mathfrak{C}^2\mathfrak{c}^{2-2\alpha}}\mathbb{L}(n))} = \frac{(1 - \frac{\gamma}{2\mathfrak{C}^2\mathfrak{c}^{2-2\alpha}}\mathbb{L}(n)) - 1}{\mathbb{L}(n)(1 - \frac{\gamma}{2\mathfrak{C}^2\mathfrak{c}^{2-2\alpha}}\mathbb{L}(n))} \\ &= \frac{-\frac{\gamma}{2\mathfrak{C}^2\mathfrak{c}^{2-2\alpha}}}{(1 - \frac{\gamma}{2\mathfrak{C}^2\mathfrak{c}^{2-2\alpha}}\mathbb{L}(n))} = -\frac{1}{(\frac{2\mathfrak{C}^2\mathfrak{c}^{2-2\alpha}}{\gamma} - \mathbb{L}(n))} < -\frac{\gamma}{2\mathfrak{C}^2\mathfrak{c}^{2-2\alpha}}. \end{aligned} \quad (9.262)$$

Therefore, we get for all $n \in \mathbb{N}_0 \cap [0, \tau)$ that

$$\frac{1}{\mathbb{L}(n)} = \frac{1}{\mathbb{L}(0)} + \sum_{k=0}^{n-1} \left[\frac{1}{\mathbb{L}(k+1)} - \frac{1}{\mathbb{L}(k)} \right] > \frac{1}{\mathbb{L}(0)} + \frac{n\gamma}{2\mathfrak{C}^2\mathfrak{c}^{2-2\alpha}} = \frac{1}{\mathfrak{c}} + \frac{n\gamma}{2\mathfrak{C}^2\mathfrak{c}^{2-2\alpha}}. \quad (9.263)$$

Therefore, we obtain for all $n \in \mathbb{N}_0 \cap [0, \tau)$ that $\mathbb{L}(n) < \frac{2\mathfrak{C}^2\mathfrak{c}^{2-2\alpha}}{n\gamma + 2\mathfrak{C}^2\mathfrak{c}^{1-2\alpha}}$. Combining this with the fact that for all $n \in \mathbb{N}_0 \cap [\tau, \infty)$ it holds that $\mathbb{L}(n) = 0$ establishes that for all $n \in \mathbb{N}_0$ we have that

$$\mathbb{L}(n) \leq \frac{2\mathfrak{C}^2\mathfrak{c}^2}{\mathbb{1}_{\{0\}}(\mathfrak{c}) + \mathfrak{c}^{2\alpha}n\gamma + 2\mathfrak{C}^2\mathfrak{c}}. \quad (9.264)$$

This, (9.257), and the assumption that \mathcal{L} is continuous prove that

$$\mathcal{L}(\psi) = \lim_{n \rightarrow \infty} \mathcal{L}(\Theta_n) = \mathcal{L}(\vartheta). \quad (9.265)$$

Combining this with (9.264) demonstrates for all $n \in \mathbb{N}_0$ that

$$0 \leq \mathcal{L}(\Theta_n) - \mathcal{L}(\psi) \leq \frac{2\mathfrak{C}^2\mathfrak{c}^2}{\mathbb{1}_{\{0\}}(\mathfrak{c}) + \mathfrak{c}^{2\alpha}n\gamma + 2\mathfrak{C}^2\mathfrak{c}}. \quad (9.266)$$

Furthermore, observe that the fact that $\mathbb{B} \ni \theta \mapsto \mathcal{G}(\theta) \in \mathbb{R}^{\mathfrak{d}}$ is continuous, the fact that $\psi \in \mathbb{B}$, and (9.257) imply that

$$\mathcal{G}(\psi) = \lim_{n \rightarrow \infty} \mathcal{G}(\Theta_n) = \lim_{n \rightarrow \infty} (\gamma^{-1}(\Theta_n - \Theta_{n+1})) = 0. \quad (9.267)$$

Next note that (9.264) and (9.252) ensure for all $n \in \mathbb{N}_0$ that

$$\begin{aligned} \|\Theta_n - \psi\|_2 &= \lim_{m \rightarrow \infty} \|\Theta_n - \Theta_m\|_2 \leq \sum_{k=n}^{\infty} \|\Theta_{k+1} - \Theta_k\|_2 \leq \frac{2\mathfrak{C}[\mathbb{L}(n)]^{1-\alpha}}{1-\alpha} \\ &\leq \frac{2^{2-\alpha}\mathfrak{C}^{3-2\alpha}\mathfrak{c}^{2-2\alpha}}{(1-\alpha)(\mathbb{1}_{\{0\}}(\mathfrak{c}) + \mathfrak{c}^{2\alpha}n\gamma + 2\mathfrak{C}^2\mathfrak{c})^{1-\alpha}}. \end{aligned} \quad (9.268)$$

Combining this with (9.265), (9.255), (9.267), and (9.266) establishes items (i), (ii), and (iii). The proof of Proposition 9.13.4 is thus complete. \square

Corollary 9.13.5. Let $\mathfrak{d} \in \mathbb{N}$, $\mathfrak{c} \in [0, 1]$, $\varepsilon, L, \mathfrak{C} \in (0, \infty)$, $\alpha \in (0, 1)$, $\gamma \in (0, L^{-1}]$, $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, let $\mathbb{B} \subseteq \mathbb{R}^{\mathfrak{d}}$ satisfy $\mathbb{B} = \{\theta \in \mathbb{R}^{\mathfrak{d}} : \|\theta - \vartheta\|_2 < \varepsilon\}$, let $\mathcal{L} \in C(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$ satisfy $\mathcal{L}|_{\mathbb{B}} \in C^1(\mathbb{B}, \mathbb{R})$, let $\mathcal{G} : \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $\theta \in \mathbb{B}$ that $\mathcal{G}(\theta) = (\nabla \mathcal{L})(\theta)$, assume for all $v, w \in \mathbb{B}$ that

$$\|\mathcal{G}(v) - \mathcal{G}(w)\|_2 \leq L\|v - w\|_2, \quad (9.269)$$

let $\Theta = (\Theta_n)_{n \in \mathbb{N}_0} : \mathbb{N}_0 \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $n \in \mathbb{N}_0$ that

$$\Theta_{n+1} = \Theta_n - \gamma \mathcal{G}(\Theta_n), \quad (9.270)$$

and assume for all $\theta \in \mathbb{B}$ that

$$|\mathcal{L}(\theta) - \mathcal{L}(\vartheta)|^\alpha \leq \mathfrak{C} \|\mathcal{G}(\theta)\|_2, \quad \mathfrak{c} = |\mathcal{L}(\Theta_0) - \mathcal{L}(\vartheta)|, \quad 2\mathfrak{C}(1-\alpha)^{-1}\mathfrak{c}^{1-\alpha} + \|\Theta_0 - \vartheta\|_2 < \frac{\varepsilon}{\gamma L+1}, \quad (9.271)$$

and $\mathcal{L}(\theta) \geq \mathcal{L}(\vartheta)$. Then there exists $\psi \in \mathcal{L}^{-1}(\{\mathcal{L}(\vartheta)\}) \cap \mathcal{G}^{-1}(\{0\})$ such that for all $n \in \mathbb{N}_0$ it holds that $\Theta_n \in \mathbb{B}$, $0 \leq \mathcal{L}(\Theta_n) - \mathcal{L}(\psi) \leq 2(2 + \mathfrak{C}^{-2}\gamma n)^{-1}$, and

$$\|\Theta_n - \psi\|_2 \leq \sum_{k=n}^{\infty} \|\Theta_{k+1} - \Theta_k\|_2 \leq 2^{2-\alpha}\mathfrak{C}(1-\alpha)^{-1}(2 + \mathfrak{C}^{-2}\gamma n)^{\alpha-1}. \quad (9.272)$$

Proof of Corollary 9.13.5. Observe that the fact that $\mathcal{L}(\vartheta) = \inf_{\theta \in \mathbb{B}} \mathcal{L}(\theta)$ ensures that $\mathcal{G}(\vartheta) = (\nabla \mathcal{L})(\vartheta) = 0$ and $\inf_{n \in \{m \in \mathbb{N}_0 : \forall k \in \mathbb{N}_0 \cap [0, m] : \Theta_k \in \mathbb{B}\}} \mathcal{L}(\Theta_n) \geq \mathcal{L}(\vartheta)$. Combining this with Proposition 9.13.4 shows that there exists $\psi \in \mathcal{L}^{-1}(\{\mathcal{L}(\vartheta)\}) \cap \mathcal{G}^{-1}(\{0\})$ such that

- (I) it holds for all $n \in \mathbb{N}_0$ that $\Theta_n \in \mathbb{B}$,
- (II) it holds for all $n \in \mathbb{N}_0$ that $0 \leq \mathcal{L}(\Theta_n) - \mathcal{L}(\psi) \leq \frac{2\mathfrak{C}^2\mathfrak{c}^2}{\mathbb{1}_{\{0\}}(\mathfrak{c}) + \mathfrak{c}^{2\alpha}n\gamma + 2\mathfrak{C}^2\mathfrak{c}}$, and
- (III) it holds for all $n \in \mathbb{N}_0$ that

$$\begin{aligned} \|\Theta_n - \psi\|_2 &\leq \sum_{k=n}^{\infty} \|\Theta_{k+1} - \Theta_k\|_2 \leq \frac{2\mathfrak{C}|\mathcal{L}(\Theta_n) - \mathcal{L}(\psi)|^{1-\alpha}}{1-\alpha} \\ &\leq \frac{2^{2-\alpha}\mathfrak{C}^{3-2\alpha}\mathfrak{c}^{2-2\alpha}}{(1-\alpha)(\mathbb{1}_{\{0\}}(\mathfrak{c}) + \mathfrak{c}^{2\alpha}n\gamma + 2\mathfrak{C}^2\mathfrak{c})^{1-\alpha}}. \end{aligned} \quad (9.273)$$

Note that item (II) and the assumption that $\mathfrak{c} \leq 1$ ensure for all $n \in \mathbb{N}_0$ that

$$0 \leq \mathcal{L}(\Theta_n) - \mathcal{L}(\psi) \leq 2\mathfrak{c}^2(\mathfrak{C}^{-2}\mathbb{1}_{\{0\}}(\mathfrak{c}) + \mathfrak{C}^{-2}\mathfrak{c}^{2\alpha}n\gamma + 2\mathfrak{c})^{-1} \leq 2(2 + \mathfrak{C}^{-2}\gamma n)^{-1}. \quad (9.274)$$

This and item (III) demonstrate for all $n \in \mathbb{N}_0$ that

$$\|\Theta_n - \psi\|_2 \leq \sum_{k=n}^{\infty} \|\Theta_{k+1} - \Theta_k\|_2 \leq \frac{2\mathfrak{C}|\mathcal{L}(\Theta_n) - \mathcal{L}(\psi)|^{1-\alpha}}{1-\alpha} \leq \left[\frac{2^{2-\alpha}\mathfrak{C}}{1-\alpha} \right] (2 + \mathfrak{C}^{-2}\gamma n)^{\alpha-1}. \quad (9.275)$$

The proof of Corollary 9.13.5 is thus complete. \square

Exercise 9.13.3. Let $\mathcal{L} \in C^1(\mathbb{R}, \mathbb{R})$ satisfy for all $\theta \in \mathbb{R}$ that

$$\mathcal{L}(\theta) = \theta^4 + \int_0^1 (\sin(x) - \theta x)^2 dx. \quad (9.276)$$

Prove or disprove the following statement: For every continuous $\Theta = (\Theta_t)_{t \in [0, \infty)} : [0, \infty) \rightarrow \mathbb{R}$ with $\sup_{t \in [0, \infty)} |\Theta_t| < \infty$ and $\forall t \in [0, \infty) : \Theta_t = \Theta_0 - \int_0^t (\nabla \mathcal{L})(\Theta_s) ds$ there exists $\vartheta \in \mathbb{R}$ such that

$$\limsup_{t \rightarrow \infty} |\Theta_t - \vartheta| = 0. \quad (9.277)$$

Exercise 9.13.4. Let $\mathcal{L} \in C^\infty(\mathbb{R}, \mathbb{R})$ satisfy for all $\theta \in \mathbb{R}$ that

$$\mathcal{L}(\theta) = \int_0^1 (\sin(x) - \theta x + \theta^2)^2 dx. \quad (9.278)$$

Prove or disprove the following statement: For every $\Theta \in C([0, \infty), \mathbb{R})$ with $\sup_{t \in [0, \infty)} |\Theta_t| < \infty$ and $\forall t \in [0, \infty) : \Theta_t = \Theta_0 - \int_0^t (\nabla \mathcal{L})(\Theta_s) ds$ there exists $\vartheta \in \mathbb{R}$, $\mathcal{C}, \beta \in (0, \infty)$ such that for all $t \in [0, \infty)$ it holds that

$$|\Theta_t - \vartheta| = \mathcal{C}(1+t)^{-\beta}. \quad (9.279)$$

9.14 On the analyticity of realization functions of ANNs

Proposition 9.14.1 (Compositions of analytic functions). *Let $l, m, n \in \mathbb{N}$, let $U \subseteq \mathbb{R}^l$ and $V \subseteq \mathbb{R}^m$ be open, let $f : U \rightarrow \mathbb{R}^m$ and $g : V \rightarrow \mathbb{R}^n$ be analytic, and assume $f(U) \subseteq V$ (cf. Definition 9.8.1). Then*

$$U \ni u \mapsto g(f(u)) \in \mathbb{R}^n \quad (9.280)$$

is analytic.

Proof of Proposition 9.14.1. Observe that Faà di Bruno's formula (cf., for instance, Fraenkel [140]) establishes that $f \circ g$ is analytic (cf. also, for example, Krantz & Parks [268, Proposition 2.8]). The proof of Proposition 9.14.1 is thus complete. \square

Lemma 9.14.2. *Let $\mathfrak{d}_1, \mathfrak{d}_2, l_1, l_2 \in \mathbb{N}$, for every $k \in \{1, 2\}$ let $F_k : \mathbb{R}^{\mathfrak{d}_k} \rightarrow \mathbb{R}^{l_k}$ be analytic, and let $f : \mathbb{R}^{\mathfrak{d}_1} \times \mathbb{R}^{\mathfrak{d}_2} \rightarrow \mathbb{R}^{l_1} \times \mathbb{R}^{l_2}$ satisfy for all $x_1 \in \mathbb{R}^{\mathfrak{d}_1}, x_2 \in \mathbb{R}^{\mathfrak{d}_2}$ that*

$$f(x_1, x_2) = (F_1(x_1), F_2(x_2)) \quad (9.281)$$

(cf. Definition 9.8.1). Then f is analytic.

Proof of Lemma 9.14.2. Throughout this proof, let $A_1 : \mathbb{R}^{l_1} \rightarrow \mathbb{R}^{l_1} \times \mathbb{R}^{l_2}$ and $A_2 : \mathbb{R}^{l_2} \rightarrow \mathbb{R}^{l_1} \times \mathbb{R}^{l_2}$ satisfy for all $x_1 \in \mathbb{R}^{l_1}, x_2 \in \mathbb{R}^{l_2}$ that

$$A_1(x_1) = (x_1, 0) \quad \text{and} \quad A_2(x_2) = (0, x_2) \quad (9.282)$$

and for every $k \in \{1, 2\}$ let $B_k: \mathbb{R}^{l_1} \times \mathbb{R}^{l_2} \rightarrow \mathbb{R}^{l_k}$ satisfy for all $x_1 \in \mathbb{R}^{l_1}, x_2 \in \mathbb{R}^{l_2}$ that

$$B_k(x_1, x_2) = x_k. \quad (9.283)$$

Note that item (i) in Lemma 5.3.1 establishes that

$$f = A_1 \circ F_1 \circ B_1 + A_2 \circ F_2 \circ B_2. \quad (9.284)$$

This, the fact that A_1, A_2, F_1, F_2, B_1 , and B_2 are analytic, and Proposition 9.14.1 establishes that f is differentiable. The proof of Lemma 9.14.2 is thus complete. \square

Lemma 9.14.3. Let $\mathfrak{d}_1, \mathfrak{d}_2, l_0, l_1, l_2 \in \mathbb{N}$, for every $k \in \{1, 2\}$ let $F_k: \mathbb{R}^{\mathfrak{d}_k} \times \mathbb{R}^{l_{k-1}} \rightarrow \mathbb{R}^{l_k}$ be analytic, and let $f: \mathbb{R}^{\mathfrak{d}_1} \times \mathbb{R}^{\mathfrak{d}_2} \times \mathbb{R}^{l_0} \rightarrow \mathbb{R}^{l_2}$ 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

$$f(\theta_1, \theta_2, x) = (F_2(\theta_2, \cdot) \circ F_1(\theta_1, \cdot))(x) \quad (9.285)$$

(cf. Definition 9.8.1). Then f is analytic.

Proof of Lemma 9.14.3. Throughout this proof, let $A: \mathbb{R}^{\mathfrak{d}_1} \times \mathbb{R}^{\mathfrak{d}_2} \times \mathbb{R}^{l_0} \rightarrow \mathbb{R}^{\mathfrak{d}_2} \times \mathbb{R}^{\mathfrak{d}_1+l_0}$ and $B: \mathbb{R}^{\mathfrak{d}_2} \times \mathbb{R}^{\mathfrak{d}_1+l_0} \rightarrow \mathbb{R}^{\mathfrak{d}_2} \times \mathbb{R}^{l_1}$ 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

$$A(\theta_1, \theta_2, x) = (\theta_2, (\theta_1, x)) \quad \text{and} \quad B(\theta_2, (\theta_1, x)) = (\theta_2, F_1(\theta_1, x)), \quad (9.286)$$

Observe that item (i) in Lemma 5.3.2 proves that

$$f = F_2 \circ B \circ A. \quad (9.287)$$

Furthermore, note that Lemma 9.14.2 (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 9.14.2) demonstrates that B is analytic. Combining this, the fact that A is analytic, the fact that F_2 is analytic, and (9.287) with Proposition 9.14.1 demonstrates that f is analytic. The proof of Lemma 9.14.3 is thus complete. \square

Corollary 9.14.4 (Analyticity of realization functions of ANNs). Let $L \in \mathbb{N}, l_0, l_1, \dots, l_L \in \mathbb{N}$ and for every $k \in \{1, 2, \dots, L\}$ let $\Psi_k: \mathbb{R}^{l_k} \rightarrow \mathbb{R}^{l_k}$ be analytic (cf. Definition 9.8.1). Then

$$\mathbb{R}^{\sum_{k=1}^L l_k(l_{k-1}+1)} \times \mathbb{R}^{l_0} \ni (\theta, x) \mapsto (\mathcal{N}_{\Psi_1, \Psi_2, \dots, \Psi_L}^{\theta, l_0})(x) \in \mathbb{R}^{l_L} \quad (9.288)$$

is analytic (cf. Definition 1.1.3).

Proof of Corollary 9.14.4. Throughout this proof, for every $k \in \{1, 2, \dots, L\}$ let $\mathfrak{d}_k = l_k(l_{k-1} + 1)$ and for every $k \in \{1, 2, \dots, L\}$ let $F_k: \mathbb{R}^{\mathfrak{d}_k} \times \mathbb{R}^{l_{k-1}} \rightarrow \mathbb{R}^{l_k}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}_k}$, $x \in \mathbb{R}^{l_{k-1}}$ that

$$F_k(\theta, x) = \Psi_k(\mathcal{A}_{l_k, l_{k-1}}^{\theta, 0}(x)) \quad (9.289)$$

(cf. Definition 1.1.1). Observe that item (i) in Lemma 5.3.3 implies that for all $\theta_1 \in \mathbb{R}^{\mathfrak{d}_1}$, $\theta_2 \in \mathbb{R}^{\mathfrak{d}_2}, \dots, \theta_L \in \mathbb{R}^{\mathfrak{d}_L}$, $x \in \mathbb{R}^{l_0}$ it holds that

$$(\mathcal{N}_{\Psi_1, \Psi_2, \dots, \Psi_L}^{(\theta_1, \theta_2, \dots, \theta_L), l_0})(x) = (F_L(\theta_L, \cdot) \circ F_{L-1}(\theta_{L-1}, \cdot) \circ \dots \circ F_1(\theta_1, \cdot))(x) \quad (9.290)$$

(cf. Definition 1.1.3). Note that the assumption that for all $k \in \{1, 2, \dots, L\}$ it holds that Ψ_k is analytic, the fact that for all $m, n \in \mathbb{N}$, $\theta \in \mathbb{R}^{m(n+1)}$ it holds that $\mathbb{R}^{m(n+1)} \times \mathbb{R}^n \ni (\theta, x) \mapsto \mathcal{A}_{m,n}^{\theta, 0}(x) \in \mathbb{R}^m$ is analytic, and Proposition 9.14.1 ensure that for all $k \in \{1, 2, \dots, L\}$ it holds that F_k is analytic. Lemma 5.3.2 and induction hence show that

$$\begin{aligned} \mathbb{R}^{\mathfrak{d}_1} \times \mathbb{R}^{\mathfrak{d}_2} \times \dots \times \mathbb{R}^{\mathfrak{d}_L} \times \mathbb{R}^{l_0} &\ni (\theta_1, \theta_2, \dots, \theta_L, x) \\ &\mapsto (F_L(\theta_L, \cdot) \circ F_{L-1}(\theta_{L-1}, \cdot) \circ \dots \circ F_1(\theta_1, \cdot))(x) \in \mathbb{R}^{l_L} \end{aligned} \quad (9.291)$$

is analytic. This and (9.290) ensure that

$$\mathbb{R}^{\sum_{k=1}^L l_k(l_{k-1}+1)} \times \mathbb{R}^{l_0} \ni (\theta, x) \mapsto \mathcal{N}_{\Psi_1, \Psi_2, \dots, \Psi_L}^{\theta, l_0}(x) \in \mathbb{R}^{l_L} \quad (9.292)$$

is analytic. The proof of Corollary 9.14.4 is thus complete. \square

Corollary 9.14.5 (Analyticity of the empirical risk function). *Let $L, \mathfrak{d} \in \mathbb{N} \setminus \{1\}$, $M, l_0, l_1, \dots, l_L \in \mathbb{N}$, $x_1, x_2, \dots, x_M \in \mathbb{R}^{l_0}$, $y_1, y_2, \dots, y_M \in \mathbb{R}^{l_L}$ satisfy $\mathfrak{d} = \sum_{k=1}^L l_k(l_{k-1} + 1)$, let $a: \mathbb{R} \rightarrow \mathbb{R}$ and $\mathbf{L}: \mathbb{R}^{l_L} \times \mathbb{R}^{l_L} \rightarrow \mathbb{R}$ be analytic, let $\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that*

$$\mathcal{L}(\theta) = \frac{1}{M} \left[\sum_{m=1}^M \mathbf{L}((\mathcal{N}_{\mathfrak{M}_{a, l_1}, \mathfrak{M}_{a, l_2}, \dots, \mathfrak{M}_{a, l_{L-1}}, \text{id}_{\mathbb{R}^{l_L}}}^{\theta, l_0})(x_m), y_m) \right] \quad (9.293)$$

(cf. Definitions 1.1.3, 1.2.1, and 9.8.1). Then \mathcal{L} is analytic.

Proof of Corollary 9.14.5. Observe that the assumption that a is analytic, Lemma 9.14.2, and induction establish that for all $m \in \mathbb{N}$ it holds that $\mathfrak{M}_{a,m}$ is analytic. This, Corollary 9.14.4 and Lemma 9.14.2 (applied with $\mathfrak{d}_1 \curvearrowleft \mathfrak{d} + l_0$, $\mathfrak{d}_2 \curvearrowleft l_L$, $l_1 \curvearrowleft l_L$, $l_2 \curvearrowleft l_L$, $F_1 \curvearrowleft (\mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{l_0} \ni (\theta, x) \mapsto (\mathcal{N}_{\mathfrak{M}_{a, l_1}, \mathfrak{M}_{a, l_2}, \dots, \mathfrak{M}_{a, l_{L-1}}, \text{id}_{\mathbb{R}^{l_L}}}^{\theta, l_0})(x) \in \mathbb{R}^{l_L})$, $F_2 \curvearrowleft \text{id}_{\mathbb{R}^{l_L}}$ in the notation of Lemma 9.14.2) ensure that

$$\mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{l_0} \times \mathbb{R}^{l_L} \ni (\theta, x, y) \mapsto ((\mathcal{N}_{\mathfrak{M}_{a, l_1}, \mathfrak{M}_{a, l_2}, \dots, \mathfrak{M}_{a, l_{L-1}}, \text{id}_{\mathbb{R}^{l_L}}}^{\theta, l_0})(x), y) \in \mathbb{R}^{l_L} \times \mathbb{R}^{l_L} \quad (9.294)$$

is analytic. The assumption that \mathbf{L} is differentiable and the chain rule hence establish that for all $x \in \mathbb{R}^{l_0}$, $y \in \mathbb{R}^{l_L}$ it holds that

$$\mathbb{R}^{\mathfrak{d}} \ni \theta \mapsto \mathbf{L}((\mathcal{N}_{\mathfrak{M}_{a,l_1}, \mathfrak{M}_{a,l_2}, \dots, \mathfrak{M}_{a,l_{L-1}}, \text{id}_{\mathbb{R}^{l_L}}}^{\theta, l_0})(x_m), y_m) \in \mathbb{R} \quad (9.295)$$

is analytic. This proves (9.293). The proof of Corollary 9.14.5 is thus complete. \square

9.15 Standard KL inequalities for empirical risks in the training of ANNs with analytic activation functions

Theorem 9.15.1 (Empirical risk minimization for ANNs with analytic activation functions). *Let $L, \mathfrak{d} \in \mathbb{N} \setminus \{1\}$, $M, l_0, l_1, \dots, l_L \in \mathbb{N}$, $x_1, x_2, \dots, x_M \in \mathbb{R}^{l_0}$, $y_1, y_2, \dots, y_M \in \mathbb{R}^{l_L}$ satisfy $\mathfrak{d} = \sum_{k=1}^L l_k(l_{k-1} + 1)$, let $a: \mathbb{R} \rightarrow \mathbb{R}$ and $\mathbf{L}: \mathbb{R}^{l_L} \times \mathbb{R}^{l_L} \rightarrow \mathbb{R}$ be analytic, let $\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that*

$$\mathcal{L}(\theta) = \frac{1}{M} \left[\sum_{m=1}^M \mathbf{L}(\mathcal{N}_{\mathfrak{M}_{a,l_1}, \mathfrak{M}_{a,l_2}, \dots, \mathfrak{M}_{a,l_{L-1}}, \text{id}_{\mathbb{R}^{l_L}}}^{\theta, l_0}(x_m), y_m) \right], \quad (9.296)$$

and let $\Theta \in C([0, \infty), \mathbb{R}^{\mathfrak{d}})$ satisfy

$$\liminf_{t \rightarrow \infty} \|\Theta_t\|_2 < \infty \quad \text{and} \quad \forall t \in [0, \infty): \Theta_t = \Theta_0 - \int_0^t (\nabla \mathcal{L})(\Theta_s) \, ds \quad (9.297)$$

(cf. Definitions 1.1.3, 1.2.1, 3.3.4, and 9.8.1). Then there exist $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, $c, \beta \in (0, \infty)$ such that for all $t \in (0, \infty)$ it holds that

$$\|\Theta_t - \vartheta\|_2 \leq ct^{-\beta}, \quad 0 \leq \mathcal{L}(\Theta_t) - \mathcal{L}(\vartheta) \leq ct^{-1}, \quad \text{and} \quad (\nabla \mathcal{L})(\vartheta) = 0. \quad (9.298)$$

Proof of Theorem 9.15.1. Note that Corollary 9.14.5 demonstrates that \mathcal{L} is analytic. Combining this with Corollary 9.12.7 establishes (9.298). The proof of Theorem 9.15.1 is thus complete. \square

Lemma 9.15.2. *Let $a: \mathbb{R} \rightarrow \mathbb{R}$ be the softplus activation function (cf. Definition 1.2.11). Then a is analytic (cf. Definition 9.8.1).*

Proof of Lemma 9.15.2. Throughout this proof, let $f: \mathbb{R} \rightarrow (0, \infty)$ satisfy for all $x \in \mathbb{R}$ that $f(x) = 1 + \exp(x)$. Observe that the fact that $\mathbb{R} \ni x \mapsto \exp(x) \in \mathbb{R}$ is analytic implies that f is analytic (cf. Definition 9.8.1). Combining this and the fact that $(0, \infty) \ni x \mapsto \ln(x) \in \mathbb{R}$ is analytic with Proposition 9.14.1 and (1.49) demonstrates that a is analytic. The proof of Lemma 9.15.2 is thus complete. \square

Lemma 9.15.3. 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 \mathbf{L} is analytic (cf. Definition 9.8.1).

Proof of Lemma 9.15.3. Note that Lemma 5.4.3 shows that \mathbf{L} is analytic (cf. Definition 9.8.1). The proof of Lemma 9.15.3 is thus complete. \square

Corollary 9.15.4 (Empirical risk minimization for ANNs with softplus activation). Let $L, \mathfrak{d} \in \mathbb{N} \setminus \{1\}$, $M, l_0, l_1, \dots, l_L \in \mathbb{N}$, $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_M \in \mathbb{R}^{l_0}$, $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_M \in \mathbb{R}^{l_L}$ satisfy $\mathfrak{d} = \sum_{k=1}^L l_k(l_{k-1} + 1)$, let a be the softplus activation function, let $\mathcal{L}: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that

$$\mathcal{L}(\theta) = \frac{1}{M} \left[\sum_{m=1}^M \left\| \mathbf{y}_m - \mathcal{N}_{\mathfrak{M}_{a,l_1}, \mathfrak{M}_{a,l_2}, \dots, \mathfrak{M}_{a,l_{L-1}}, \text{id}_{\mathbb{R}^{l_L}}}^{\theta, l_0}(\mathbf{x}_m) \right\|_2^2 \right], \quad (9.299)$$

and let $\Theta \in C([0, \infty), \mathbb{R}^{\mathfrak{d}})$ satisfy

$$\liminf_{t \rightarrow \infty} \|\Theta_t\|_2 < \infty \quad \text{and} \quad \forall t \in [0, \infty): \Theta_t = \Theta_0 - \int_0^t (\nabla \mathcal{L})(\Theta_s) \, ds \quad (9.300)$$

(cf. Definitions 1.1.3, 1.2.1, 1.2.11, and 3.3.4). Then there exist $\vartheta \in \mathbb{R}^{\mathfrak{d}}$, $c, \beta \in (0, \infty)$ such that for all $t \in (0, \infty)$ it holds that

$$\|\Theta_t - \vartheta\|_2 \leq ct^{-\beta}, \quad 0 \leq \mathcal{L}(\Theta_t) - \mathcal{L}(\vartheta) \leq ct^{-1} \quad \text{and} \quad (\nabla \mathcal{L})(\vartheta) = 0. \quad (9.301)$$

Proof of Corollary 9.15.4. Observe that Lemma 9.15.2, Lemma 9.15.3, and Theorem 9.15.1 ensure (9.301). The proof of Corollary 9.15.4 is thus complete. \square

Remark 9.15.5 (Convergence to a good suboptimal critical point whose risk value is close to the optimal risk value). Corollary 9.15.4 establishes convergence of a non-divergent GF trajectory in the training of fully-connected feedforward ANNs to a critical point $\vartheta \in \mathbb{R}^{\mathfrak{d}}$ of the objective function. In several scenarios in the training of ANNs such limiting critical points seem to be with high probability not global minimum points but suboptimal critical points at which the value of the objective function is, however, not far away from the minimal value of the objective function (cf. Ibragimov et al. [227] and also [151, 430]). In view of this, there has been an increased interest in landscape analyses associated to the objective function to gather more information on critical points of the objective function (cf., for instance, [12, 75, 82, 83, 95, 119, 147, 226, 227, 253, 332, 378, 379, 386, 402–404, 421, 456, 457] and the references therein).

In general in most cases it remains an open problem to rigorously prove that the value of the objective function at the limiting critical point is indeed with high probability close

to the minimal/infimal value¹ of the objective function and thereby establishing a full convergence analysis. However, in the so-called overparametrized regime where there are much more ANN parameters than input-output training data pairs, several convergence analyses for the training of ANNs have been achieved (cf., for instance, [77, 78, 120, 229] and the references therein).

Remark 9.15.6 (Almost surely excluding strict saddle points). We also note that in several situations it has been shown that the limiting critical point of the considered GF trajectory with random initialization or of the considered GD process with random initialization is almost surely not a saddle points but a local minimizers; cf., for example, [74, 279, 280, 342, 343].

Remark 9.15.7 (A priori bounds and existence of minimizers). Under the assumption that the considered GF trajectory is non-divergent in the sense that

$$\liminf_{t \rightarrow \infty} \|\Theta_t\|_2 < \infty \quad (9.302)$$

(see (9.300) above) we have that Corollary 9.15.4 establishes convergence of a GF trajectory in the training of fully-connected feedforward ANNs to a critical point $\vartheta \in \mathbb{R}^d$ of the objective function (see (9.301) above). Such kind of non-divergence and slightly stronger boundedness assumptions, respectively, are very common hypotheses in convergence results for gradient based optimization methods in the training of ANNs (cf., for instance, [2, 8, 46, 106, 107, 132, 236, 412], Section 9.12.2, and Theorem 9.15.1 in the context of the KL approach and [96, 107, 237, 315] in the context of other approaches).

In most scenarios in the training of ANNs it remains an open problem to prove or disprove such non-divergence and boundedness assumptions. In Gallon et al. [148] the condition in (9.302) has been disproved and divergence of GF trajectories in the training of shallow fully-connected feedforward ANNs has been established for specific target functions; see also Petersen et al. [352].

The question of non-divergence of gradient based optimization methods seems to be closely related to the question whether there exist minimizers in the optimization landscape of the objective function. We refer to [105, 108, 236, 247] for results proving the existence of minimizers in optimization landscapes for the training of ANNs and we refer to [148, 352] for results disproving the existence of minimizers in optimization landscapes for the training of ANNs. We also refer to, for example, [131, 227] for strongly simplified ANN training scenarios where non-divergence and boundedness conditions of the form (9.302) have been established.

¹It is of interest to note that it seems to strongly depend on the activation function, the architecture of the ANN, and the underlying probability distribution of the data of the considered learning problem whether the infimal value of the objective function is also a minimal value of the objective function or whether there exists no minimal value of the objective function (cf., for example, [105, 148] and Remark 9.15.7 below).

9.16 Generalized KL-inequalities

In this section we present and study suitable generalized gradients (Fréchet subgradients and limiting Fréchet subgradients) and we briefly present generalized KL inequalities that are based on such generalized gradients. The specific presentation of this section is based on [234, Section 3.8].

9.16.1 Fréchet subgradients and limiting Fréchet subgradients

Definition 9.16.1 (Fréchet subgradients and limiting Fréchet subgradients). *Let $\mathfrak{d} \in \mathbb{N}$, $\mathcal{L} \in C(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$, $\theta \in \mathbb{R}^{\mathfrak{d}}$. Then we denote by $(\mathcal{DL})(\theta) \subseteq \mathbb{R}^{\mathfrak{d}}$ the set given by*

$$(\mathcal{DL})(\theta) = \left\{ v \in \mathbb{R}^{\mathfrak{d}} : \left[\liminf_{\mathbb{R}^{\mathfrak{d}} \setminus \{0\} \ni h \rightarrow 0} \left(\frac{\mathcal{L}(\theta + h) - \mathcal{L}(\theta) - \langle v, h \rangle}{\|h\|_2} \right) \geq 0 \right] \right\}, \quad (9.303)$$

we call $(\mathcal{DL})(\theta)$ the set of Fréchet subgradients of f at θ , we denote by $(\mathcal{DL})(\theta) \subseteq \mathbb{R}^{\mathfrak{d}}$ the set given by

$$(\mathcal{DL})(\theta) = \bigcap_{\varepsilon \in (0, \infty)} \overline{\left[\bigcup_{v \in \{z \in \mathbb{R}^{\mathfrak{d}} : \|\theta - z\|_2 < \varepsilon\}} (\mathcal{DL})(v) \right]}, \quad (9.304)$$

and we call $(\mathcal{DL})(\theta)$ the set of limiting Fréchet subgradients of f at θ (cf. Definitions 1.4.7 and 3.3.4).

Lemma 9.16.2 (Convex differentials). *Let $\mathfrak{d} \in \mathbb{N}$, $\mathcal{L} \in C(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$, $\theta, a \in \mathbb{R}^{\mathfrak{d}}$, $b \in \mathbb{R}$, $\varepsilon \in (0, \infty)$ and let $A: \mathbb{R}^{\mathfrak{d}} \rightarrow \mathbb{R}$ satisfy for all $v \in \{w \in \mathbb{R}^{\mathfrak{d}} : \|w - \theta\|_2 < \varepsilon\}$ that*

$$A(v) = \langle a, v \rangle + b \leq \mathcal{L}(v) \quad \text{and} \quad A(\theta) = \mathcal{L}(\theta) \quad (9.305)$$

(cf. Definitions 1.4.7 and 3.3.4). Then

- (i) *it holds for all $v \in \{w \in \mathbb{R}^{\mathfrak{d}} : \|w - \theta\|_2 < \varepsilon\}$ that $A(v) = \langle a, v - \theta \rangle + \mathcal{L}(\theta)$ and*
- (ii) *it holds that $a \in (\mathcal{DL})(\theta)$*

(cf. Definition 9.16.1).

Proof of Lemma 9.16.2. Note that (9.305) establishes for all $v \in \{w \in \mathbb{R}^{\mathfrak{d}} : \|w - \theta\|_2 < \varepsilon\}$ that

$$\begin{aligned} A(v) &= [A(v) - A(\theta)] + A(\theta) = [(\langle a, v \rangle + b) - (\langle a, \theta \rangle + b)] + A(\theta) \\ &= \langle a, v - \theta \rangle + A(\theta) = \langle a, v - \theta \rangle + \mathcal{L}(\theta). \end{aligned} \quad (9.306)$$

This establishes item (i). Observe that (9.305) and item (i) ensure for all $h \in \{w \in \mathbb{R}^d : 0 < \|w\|_2 < \varepsilon\}$ that

$$\frac{\mathcal{L}(\theta + h) - \mathcal{L}(\theta) - \langle a, h \rangle}{\|h\|_2} = \frac{\mathcal{L}(\theta + h) - A(\theta + h)}{\|h\|_2} \geq 0. \quad (9.307)$$

This and (9.303) establish item (ii). The proof of Lemma 9.16.2 is thus complete. \square

Lemma 9.16.3 (Properties of Fréchet subgradients). *Let $d \in \mathbb{N}$, $\mathcal{L} \in C(\mathbb{R}^d, \mathbb{R})$. Then*

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

$$(\mathcal{D}\mathcal{L})(\theta) = \left\{ v \in \mathbb{R}^d : [\exists z = (z_1, z_2) : \mathbb{N} \rightarrow \mathbb{R}^d \times \mathbb{R}^d : ([\forall k \in \mathbb{N} : z_2(k) \in (\mathcal{D}\mathcal{L})(z_1(k))] \wedge [\limsup_{k \rightarrow \infty} (\|z_1(k) - \theta\|_2 + \|z_2(k) - v\|_2) = 0])]\right\}, \quad (9.308)$$

(ii) *it holds for all $\theta \in \mathbb{R}^d$ that $(\mathcal{D}\mathcal{L})(\theta) \subseteq (\mathcal{D}\mathcal{L})(\theta)$,*

(iii) *it holds for all $\theta \in \{v \in \mathbb{R}^d : \mathcal{L} \text{ is differentiable at } v\}$ that $(\mathcal{D}\mathcal{L})(\theta) = \{(\nabla \mathcal{L})(\theta)\}$,*

(iv) *it holds for all $\theta \in \bigcup_{U \subseteq \mathbb{R}^d, U \text{ is open}, \mathcal{L}|_U \in C^1(U, \mathbb{R})} U$ that $(\mathcal{D}\mathcal{L})(\theta) = \{(\nabla \mathcal{L})(\theta)\}$, and*

(v) *it holds for all $\theta \in \mathbb{R}^d$ that $(\mathcal{D}\mathcal{L})(\theta)$ is closed.*

(cf. Definitions 3.3.4 and 9.16.1).

Proof of Lemma 9.16.3. Throughout this proof, for every $\theta, v \in \mathbb{R}^d$ let $Z^{\theta, v} = (Z_1^{\theta, v}, Z_2^{\theta, v}) : \mathbb{N} \rightarrow \mathbb{R}^d \times \mathbb{R}^d$ satisfy for all $k \in \mathbb{N}$ that

$$Z_1^{\theta, v}(k) = \theta \quad \text{and} \quad Z_2^{\theta, v}(k) = v. \quad (9.309)$$

Note that (9.304) proves that for all $\theta \in \mathbb{R}^d$, $v \in (\mathcal{D}\mathcal{L})(\theta)$, $\varepsilon \in (0, \infty)$ it holds that

$$v \in \overline{[\bigcup_{w \in \{u \in \mathbb{R}^d : \|\theta - u\|_2 < \varepsilon\}} (\mathcal{D}\mathcal{L})(w)]}. \quad (9.310)$$

This implies that for all $\theta \in \mathbb{R}^d$, $v \in (\mathcal{D}\mathcal{L})(\theta)$ and all $\varepsilon, \delta \in (0, \infty)$ there exists $V \in (\bigcup_{w \in \{u \in \mathbb{R}^d : \|\theta - u\|_2 < \varepsilon\}} (\mathcal{D}\mathcal{L})(w))$ such that

$$\|v - V\|_2 < \delta. \quad (9.311)$$

Therefore, we obtain that for all $\theta \in \mathbb{R}^d$, $v \in (\mathcal{D}\mathcal{L})(\theta)$, $\varepsilon, \delta \in (0, \infty)$ there exist $w \in \{u \in \mathbb{R}^d : \|\theta - u\|_2 < \varepsilon\}$, $V \in (\mathcal{D}\mathcal{L})(w)$ such that $\|v - V\|_2 < \delta$. This demonstrates that for all $\theta \in \mathbb{R}^d$, $v \in (\mathcal{D}\mathcal{L})(\theta)$, $\varepsilon, \delta \in (0, \infty)$ there exist $\Theta \in \mathbb{R}^d$, $V \in (\mathcal{D}\mathcal{L})(\Theta)$ such that

$$\|\theta - \Theta\|_2 < \varepsilon \quad \text{and} \quad \|v - V\|_2 < \delta. \quad (9.312)$$

Hence, we obtain that for all $\theta \in \mathbb{R}^{\mathfrak{d}}$, $v \in (\mathcal{DL})(\theta)$, $k \in \mathbb{N}$ there exist $z_1, z_2 \in \mathbb{R}^{\mathfrak{d}}$ such that

$$z_2 \in (\mathcal{DL})(z_1) \quad \text{and} \quad \|z_1 - \theta\|_2 + \|z_2 - v\|_2 < \frac{1}{k}. \quad (9.313)$$

Furthermore, observe that for all $\theta, v \in \mathbb{R}^{\mathfrak{d}}$, $\varepsilon \in (0, \infty)$ and all $z = (z_1, z_2) : \mathbb{N} \rightarrow \mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\mathfrak{d}}$ with $\limsup_{k \rightarrow \infty} (\|z_1(k) - \theta\|_2 + \|z_2(k) - v\|_2) = 0$ and $\forall k \in \mathbb{N} : z_2(k) \in (\mathcal{DL})(z_1(k))$ there exist $\Theta, V \in \mathbb{R}^{\mathfrak{d}}$ such that

$$V \in (\mathcal{DL})(\Theta) \quad \text{and} \quad \|\Theta - \theta\|_2 + \|V - v\|_2 < \varepsilon. \quad (9.314)$$

Therefore, we obtain that for all $\theta, v \in \mathbb{R}^{\mathfrak{d}}$, $\varepsilon, \delta \in (0, \infty)$ and all $z = (z_1, z_2) : \mathbb{N} \rightarrow \mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\mathfrak{d}}$ with $\limsup_{k \rightarrow \infty} (\|z_1(k) - \theta\|_2 + \|z_2(k) - v\|_2) = 0$ and $\forall k \in \mathbb{N} : z_2(k) \in (\mathcal{DL})(z_1(k))$ there exist $\Theta, V \in \mathbb{R}^{\mathfrak{d}}$ such that

$$V \in (\mathcal{DL})(\Theta), \quad \|\theta - \Theta\|_2 < \varepsilon, \quad \text{and} \quad \|v - V\|_2 < \delta. \quad (9.315)$$

This shows that for all $\theta, v \in \mathbb{R}^{\mathfrak{d}}$, $\varepsilon, \delta \in (0, \infty)$ and all $z = (z_1, z_2) : \mathbb{N} \rightarrow \mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\mathfrak{d}}$ with $\limsup_{k \rightarrow \infty} (\|z_1(k) - \theta\|_2 + \|z_2(k) - v\|_2) = 0$ and $\forall k \in \mathbb{N} : z_2(k) \in (\mathcal{DL})(z_1(k))$ there exist $w \in \{u \in \mathbb{R}^{\mathfrak{d}} : \|\theta - u\|_2 < \varepsilon\}$, $V \in (\mathcal{DL})(w)$ such that $\|v - V\|_2 < \delta$. Hence, we obtain that for all $\theta, v \in \mathbb{R}^{\mathfrak{d}}$, $\varepsilon, \delta \in (0, \infty)$ and all $z = (z_1, z_2) : \mathbb{N} \rightarrow \mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\mathfrak{d}}$ with $\limsup_{k \rightarrow \infty} (\|z_1(k) - \theta\|_2 + \|z_2(k) - v\|_2) = 0$ and $\forall k \in \mathbb{N} : z_2(k) \in (\mathcal{DL})(z_1(k))$ there exists $V \in [\bigcup_{w \in \{u \in \mathbb{R}^{\mathfrak{d}} : \|\theta - u\|_2 < \varepsilon\}} (\mathcal{DL})(w)]$ such that

$$\|v - V\|_2 < \delta. \quad (9.316)$$

This ensures that for all $\theta, v \in \mathbb{R}^{\mathfrak{d}}$, $\varepsilon \in (0, \infty)$ and all $z = (z_1, z_2) : \mathbb{N} \rightarrow \mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\mathfrak{d}}$ with $\limsup_{k \rightarrow \infty} (\|z_1(k) - \theta\|_2 + \|z_2(k) - v\|_2) = 0$ and $\forall k \in \mathbb{N} : z_2(k) \in (\mathcal{DL})(z_1(k))$ it holds that

$$v \in \overline{\left[\bigcup_{w \in \{u \in \mathbb{R}^{\mathfrak{d}} : \|\theta - u\|_2 < \varepsilon\}} (\mathcal{DL})(w) \right]}. \quad (9.317)$$

This and (9.304) establish that for all $\theta, v \in \mathbb{R}^{\mathfrak{d}}$ and all $z = (z_1, z_2) : \mathbb{N} \rightarrow \mathbb{R}^{\mathfrak{d}} \times \mathbb{R}^{\mathfrak{d}}$ with $\limsup_{k \rightarrow \infty} (\|z_1(k) - \theta\|_2 + \|z_2(k) - v\|_2) = 0$ and $\forall k \in \mathbb{N} : z_2(k) \in (\mathcal{DL})(z_1(k))$ it holds that

$$v \in (\mathcal{DL})(\theta). \quad (9.318)$$

Combining this with (9.313) proves item (i). Note that (9.309) implies that for all $\theta \in \mathbb{R}^{\mathfrak{d}}$, $v \in (\mathcal{DL})(\theta)$ it holds that

$$\left[\forall k \in \mathbb{N} : \left(Z_2^{\theta, v}(k) \in (\mathcal{DL})(Z_1^{\theta, v}(k)) \right) \right] \wedge \left[\limsup_{k \rightarrow \infty} (\|Z_1^{\theta, v}(k) - \theta\|_2 + \|Z_2^{\theta, v}(k) - v\|_2) = 0 \right] \quad (9.319)$$

(cf. Definitions 3.3.4 and 9.16.1). Combining this with item (i) establishes item (ii). Observe that the fact that for all $a \in \mathbb{R}$ it holds that $-a \leq |a|$ demonstrates that for all

$\theta \in \{v \in \mathbb{R}^d : \mathcal{L} \text{ is differentiable at } v\}$ it holds that

$$\begin{aligned} \liminf_{\mathbb{R}^d \setminus \{0\} \ni h \rightarrow 0} \left(\frac{\mathcal{L}(\theta+h) - \mathcal{L}(\theta) - \langle (\nabla \mathcal{L})(\theta), h \rangle}{\|h\|_2} \right) &\geq - \left| \liminf_{\mathbb{R}^d \setminus \{0\} \ni h \rightarrow 0} \left(\frac{\mathcal{L}(\theta+h) - \mathcal{L}(\theta) - \langle (\nabla \mathcal{L})(\theta), h \rangle}{\|h\|_2} \right) \right| \\ &\geq - \left[\limsup_{\mathbb{R}^d \setminus \{0\} \ni h \rightarrow 0} \left(\frac{|\mathcal{L}(\theta+h) - \mathcal{L}(\theta) - \langle (\nabla \mathcal{L})(\theta), h \rangle|}{\|h\|_2} \right) \right] = 0 \end{aligned} \quad (9.320)$$

(cf. Definition 1.4.7). This demonstrates that for all $\theta \in \{v \in \mathbb{R}^d : \mathcal{L} \text{ is differentiable at } v\}$ it holds that

$$(\nabla \mathcal{L})(\theta) \in (\mathcal{D}\mathcal{L})(\theta). \quad (9.321)$$

Moreover, note that for all $w \in \mathbb{R}^d \setminus \{0\}$ it holds that

$$\begin{aligned} \liminf_{\mathbb{R}^d \setminus \{0\} \ni h \rightarrow 0} \left(\frac{\langle w, h \rangle}{\|h\|_2} \right) &= \sup_{\varepsilon \in (0, \infty)} \inf_{h \in \{u \in \mathbb{R}^d : \|u\|_2 \leq \varepsilon\}} \left(\frac{\langle w, h \rangle}{\|h\|_2} \right) \\ &\leq \sup_{\varepsilon \in (0, \infty)} \left(\frac{\langle w, -\varepsilon \|w\|_2^{-1} w \rangle}{\|-\varepsilon \|w\|_2^{-1} w\|_2} \right) = \sup_{\varepsilon \in (0, \infty)} (\langle w, -\|w\|_2^{-1} w \rangle) = -\|w\|_2 < 0. \end{aligned} \quad (9.322)$$

Therefore, we obtain for all $\theta \in \{v \in \mathbb{R}^d : \mathcal{L} \text{ is differentiable at } v\}$, $u \in (\mathcal{D}\mathcal{L})(\theta)$ that

$$\begin{aligned} 0 &\leq \liminf_{\mathbb{R}^d \setminus \{0\} \ni h \rightarrow 0} \left(\frac{\mathcal{L}(\theta+h) - \mathcal{L}(\theta) - \langle u, h \rangle}{\|h\|_2} \right) \\ &= \liminf_{\mathbb{R}^d \setminus \{0\} \ni h \rightarrow 0} \left(\frac{\mathcal{L}(\theta+h) - \mathcal{L}(\theta) - \langle (\nabla \mathcal{L})(\theta), h \rangle - \langle u - (\nabla \mathcal{L})(\theta), h \rangle}{\|h\|_2} \right) \\ &\leq \liminf_{\mathbb{R}^d \setminus \{0\} \ni h \rightarrow 0} \left(\frac{|\mathcal{L}(\theta+h) - \mathcal{L}(\theta) - \langle (\nabla \mathcal{L})(\theta), h \rangle| + \langle (\nabla \mathcal{L})(\theta) - u, h \rangle}{\|h\|_2} \right) \\ &\leq \left[\liminf_{\mathbb{R}^d \setminus \{0\} \ni h \rightarrow 0} \left(\frac{\langle (\nabla \mathcal{L})(\theta) - u, h \rangle}{\|h\|_2} \right) \right] + \left[\limsup_{\mathbb{R}^d \setminus \{0\} \ni h \rightarrow 0} \left(\frac{|\mathcal{L}(\theta+h) - \mathcal{L}(\theta) - \langle (\nabla \mathcal{L})(\theta), h \rangle|}{\|h\|_2} \right) \right] \\ &= \liminf_{\mathbb{R}^d \setminus \{0\} \ni h \rightarrow 0} \left(\frac{\langle (\nabla \mathcal{L})(\theta) - u, h \rangle}{\|h\|_2} \right) \leq -\|(\nabla \mathcal{L})(\theta) - u\|_2. \end{aligned} \quad (9.323)$$

Combining this with (9.321) proves item (iii). Observe that items (ii) and (iii) show that for all open $U \subseteq \mathbb{R}^n$ and all $\theta \in U$ with $\mathcal{L}|_U \in C^1(U, \mathbb{R})$ it holds that

$$\{(\nabla \mathcal{L})(\theta)\} = (\mathcal{D}\mathcal{L})(\theta) \subseteq (\mathcal{D}\mathcal{L})(\theta). \quad (9.324)$$

In addition, note that for all open $U \subseteq \mathbb{R}^d$, all $\theta \in U$, $v \in \mathbb{R}^d$ and all $z = (z_1, z_2) : \mathbb{N} \rightarrow \mathbb{R}^d \times \mathbb{R}^d$ with $\limsup_{k \rightarrow \infty} (\|z_1(k) - \theta\|_2 + \|z_2(k) - v\|_2) = 0$ and $\forall k \in \mathbb{N} : z_2(k) \in (\mathcal{D}\mathcal{L})(z_1(k))$ there exists $K \in \mathbb{N}$ such that for all $k \in \mathbb{N} \cap [K, \infty)$ it holds that

$$z_1(k) \in U. \quad (9.325)$$

Combining this with item (iii) ensures that for all open $U \subseteq \mathbb{R}^d$, all $\theta \in U$, $v \in \mathbb{R}^d$ and all $z = (z_1, z_2) : \mathbb{N} \rightarrow \mathbb{R}^d \times \mathbb{R}^d$ with $\mathcal{L}|_U \in C^1(U, \mathbb{R})$, $\limsup_{k \rightarrow \infty} (\|z_1(k) - \theta\|_2 + \|z_2(k) - v\|_2) = 0$

and $\forall k \in \mathbb{N}: z_2(k) \in (\mathcal{DL})(z_1(k))$ there exists $K \in \mathbb{N}$ such that $\forall k \in \mathbb{N} \cap [K, \infty): z_1(k) \in U$ and

$$\begin{aligned} & \limsup_{\mathbb{N} \cap [K, \infty) \ni k \rightarrow \infty} (\|z_1(k) - \theta\|_2 + \|(\nabla \mathcal{L})(z_1(k)) - v\|_2) \\ &= \limsup_{k \rightarrow \infty} (\|z_1(k) - \theta\|_2 + \|z_2(k) - v\|_2) = 0. \end{aligned} \quad (9.326)$$

This and item (i) imply that for all open $U \subseteq \mathbb{R}^d$ and all $\theta \in U$, $v \in (\mathcal{DL})(\theta)$ with $\mathcal{L}|_U \in C^1(U, \mathbb{R})$ it holds that

$$v = (\nabla \mathcal{L})(\theta). \quad (9.327)$$

Combining this with (9.324) establishes item (iv). Observe that (9.304) demonstrates that for all $\theta \in \mathbb{R}^d$ it holds that

$$\mathbb{R}^d \setminus ((\mathcal{DL})(\theta)) = \bigcup_{\varepsilon \in (0, \infty)} \left(\overline{\mathbb{R}^d \setminus \left(\bigcup_{v \in \{z \in \mathbb{R}^d : \|\theta - z\|_2 < \varepsilon\}} (\mathcal{DL})(v) \right)} \right) \quad (9.328)$$

Hence, we obtain for all $\theta \in \mathbb{R}^d$ that $\mathbb{R}^d \setminus ((\mathcal{DL})(\theta))$ is open. This proves item (v). The proof of Lemma 9.16.3 is thus complete. \square

Lemma 9.16.4 (Fréchet subgradients for maxima). *Let $c \in \mathbb{R}$ and let $\mathcal{L}: \mathbb{R} \rightarrow \mathbb{R}$ satisfy for all $\theta \in \mathbb{R}$ that $\mathcal{L}(\theta) = \max\{\theta, c\}$. Then*

- (i) *it holds for all $\theta \in (-\infty, c)$ that $(\mathcal{DL})(\theta) = \{0\}$,*
- (ii) *it holds for all $\theta \in (c, \infty)$ that $(\mathcal{DL})(\theta) = \{1\}$, and*
- (iii) *it holds that $(\mathcal{DL})(c) = [0, 1]$*
(cf. Definition 9.16.1).

Proof of Lemma 9.16.4. Note that item (iii) in Lemma 9.16.3 establishes items (i) and (ii). Observe that Lemma 9.16.2 establishes

$$[0, 1] \subseteq (\mathcal{DL})(c). \quad (9.329)$$

Furthermore, note that the assumption that for all $\theta \in \mathbb{R}$ it holds that $\mathcal{L}(\theta) = \max\{\theta, c\}$ shows that for all $a \in (1, \infty)$, $h \in (0, \infty)$ it holds that

$$\frac{\mathcal{L}(c+h) - \mathcal{L}(c) - ah}{|h|} = \frac{(c+h) - c - ah}{h} = 1 - a < 0. \quad (9.330)$$

Moreover, observe that the assumption that for all $\theta \in \mathbb{R}$ it holds that $\mathcal{L}(\theta) = \max\{\theta, c\}$ ensures that for all $a, h \in (-\infty, 0)$, it holds that

$$\frac{\mathcal{L}(c+h) - \mathcal{L}(c) - ah}{|h|} = \frac{c - c - ah}{-h} = a < 0. \quad (9.331)$$

Combining this with (9.330) demonstrates that

$$(\mathcal{DL})(c) \subseteq [0, 1]. \quad (9.332)$$

This and (9.329) establish item (iii). The proof of Lemma 9.16.4 is thus complete. \square

Lemma 9.16.5 (Limits of limiting Fréchet subgradients). *Let $\mathfrak{d} \in \mathbb{N}$, $\mathcal{L} \in C(\mathbb{R}^\mathfrak{d}, \mathbb{R})$, let $(\theta_k)_{k \in \mathbb{N}_0} \subseteq \mathbb{R}^\mathfrak{d}$ and $(v_k)_{k \in \mathbb{N}_0} \subseteq \mathbb{R}^\mathfrak{d}$ satisfy*

$$\limsup_{k \rightarrow \infty} (\|\theta_k - \theta_0\|_2 + \|v_k - v_0\|_2) = 0, \quad (9.333)$$

and assume for all $k \in \mathbb{N}$ that $v_k \in (\mathcal{DL})(\theta_k)$ (cf. Definitions 3.3.4 and 9.16.1). Then $v_0 \in (\mathcal{DL})(\theta_0)$.

Proof of Lemma 9.16.5. Note that item (i) in Lemma 9.16.3 and the fact that for all $k \in \mathbb{N}$ it holds that $v_k \in (\mathcal{DL})(\theta_k)$ imply that for every $k \in \mathbb{N}$ there exists $z^{(k)} = (z_1^{(k)}, z_2^{(k)}) : \mathbb{N} \rightarrow \mathbb{R}^\mathfrak{d} \times \mathbb{R}^\mathfrak{d}$ which satisfies for all $n \in \mathbb{N}$ that

$$z_2^{(k)}(n) \in (\mathcal{DL})(z_1^{(k)}(n)) \text{ and } \limsup_{w \rightarrow \infty} (\|z_1^{(k)}(w) - \theta_k\|_2 + \|z_2^{(k)}(w) - v_k\|_2) = 0. \quad (9.334)$$

Observe that (9.334) demonstrates that there exists $w = (w_k)_{k \in \mathbb{N}} : \mathbb{N} \rightarrow \mathbb{N}$ which satisfies for all $k \in \mathbb{N}$ that

$$\|z_1^{(k)}(w_k) - \theta_k\|_2 + \|z_2^{(k)}(w_k) - v_k\|_2 \leq 2^{-k}. \quad (9.335)$$

Next let $Z = (Z_1, Z_2) : \mathbb{N} \rightarrow \mathbb{R}^\mathfrak{d} \times \mathbb{R}^\mathfrak{d}$ satisfy for all $j \in \{1, 2\}$, $k \in \mathbb{N}$ that

$$Z_j(k) = z_j^{(k)}(w_k). \quad (9.336)$$

Note that (9.334), (9.335), (9.336), and the assumption that $\limsup_{k \rightarrow \infty} (\|\theta_k - \theta_0\|_2 + \|v_k - v_0\|_2) = 0$ prove that

$$\begin{aligned} & \limsup_{k \rightarrow \infty} (\|Z_1(k) - \theta_0\|_2 + \|Z_2(k) - v_0\|_2) \\ & \leq [\limsup_{k \rightarrow \infty} (\|Z_1(k) - \theta_k\|_2 + \|Z_2(k) - v_k\|_2)] \\ & \quad + [\limsup_{k \rightarrow \infty} (\|\theta_k - \theta_0\|_2 + \|v_k - v_0\|_2)] \\ & = \limsup_{k \rightarrow \infty} (\|Z_1(k) - \theta_k\|_2 + \|Z_2(k) - v_k\|_2) \\ & = \limsup_{k \rightarrow \infty} (\|z_1^{(k)}(w_k) - \theta_k\|_2 + \|z_2^{(k)}(w_k) - v_k\|_2) \\ & \leq \limsup_{k \rightarrow \infty} (2^{-k}) = 0. \end{aligned} \quad (9.337)$$

Furthermore, observe that (9.334) and (9.336) establish that for all $k \in \mathbb{N}$ it holds that $Z_2(k) \in (\mathcal{DL})(Z_1(k))$. Combining this and (9.337) with item (i) in Lemma 9.16.3 proves that $v_0 \in (\mathcal{DL})(\theta_0)$. The proof of Lemma 9.16.5 is thus complete. \square

Exercise 9.16.1. Prove or disprove the following statement: It holds for all $\mathfrak{d} \in \mathbb{N}$, $\mathcal{L} \in C^1(\mathbb{R}^\mathfrak{d}, \mathbb{R})$, $\theta \in \mathbb{R}^\mathfrak{d}$ that $(\mathcal{D}\mathcal{L})(\theta) = (\mathcal{D}\mathcal{L})(\theta)$ (cf. Definition 9.16.1).

Exercise 9.16.2. Prove or disprove the following statement: There exists $\mathfrak{d} \in \mathbb{N}$ such that for all $\mathcal{L} \in C(\mathbb{R}^\mathfrak{d}, \mathbb{R})$, $\theta \in \mathbb{R}^\mathfrak{d}$ it holds that $(\mathcal{D}\mathcal{L})(\theta) \subseteq (\mathcal{D}\mathcal{L})(\theta)$ (cf. Definition 9.16.1).

Exercise 9.16.3. Prove or disprove the following statement: It holds for all $\mathfrak{d} \in \mathbb{N}$, $\mathcal{L} \in C(\mathbb{R}^\mathfrak{d}, \mathbb{R})$, $\theta \in \mathbb{R}^\mathfrak{d}$ that $(\mathcal{D}\mathcal{L})(\theta)$ is convex (cf. Definition 9.16.1).

Exercise 9.16.4. Prove or disprove the following statement: It holds for all $\mathfrak{d} \in \mathbb{N}$, $\mathcal{L} \in C(\mathbb{R}^\mathfrak{d}, \mathbb{R})$, $\theta \in \mathbb{R}^n$ that $(\mathcal{D}\mathcal{L})(\theta)$ is convex (cf. Definition 9.16.1).

Exercise 9.16.5. For every $\alpha \in (0, \infty)$, $s \in \{-1, 1\}$ let $\mathcal{L}_{\alpha,s}: \mathbb{R} \rightarrow \mathbb{R}$ satisfy for all $\theta \in \mathbb{R}$ that

$$\mathcal{L}_{\alpha,s}(\theta) = \begin{cases} \theta & : \theta > 0 \\ s|\theta|^\alpha & : \theta \leq 0. \end{cases} \quad (9.338)$$

For every $\alpha \in (0, \infty)$, $s \in \{-1, 1\}$, $\theta \in \mathbb{R}$ specify $(\mathcal{D}\mathcal{L}_{\alpha,s})(\theta)$ and $(\mathcal{D}\mathcal{L}_{\alpha,s})(\theta)$ explicitly and prove that your results are correct (cf. Definition 9.16.1)!

9.16.2 Non-smooth slope

Definition 9.16.6 (Non-smooth slope). Let $\mathfrak{d} \in \mathbb{N}$, $\mathcal{L} \in C(\mathbb{R}^\mathfrak{d}, \mathbb{R})$. Then we denote by $\mathbb{S}_f: \mathbb{R}^\mathfrak{d} \rightarrow [0, \infty]$ the function which satisfies for all $\theta \in \mathbb{R}^\mathfrak{d}$ that

$$\mathbb{S}_{\mathcal{L}}(\theta) = \inf(\{r \in \mathbb{R}: (\exists v \in (\mathcal{D}\mathcal{L})(\theta): r = \|v\|_2)\} \cup \{\infty\}) \quad (9.339)$$

and we call \mathbb{S}_f the non-smooth slope of f (cf. Definitions 3.3.4 and 9.16.1).

9.16.3 Generalized KL functions

Definition 9.16.7 (Generalized KL inequalities). Let $\mathfrak{d} \in \mathbb{N}$, $c \in \mathbb{R}$, $\alpha \in (0, \infty)$, $\mathcal{L} \in C(\mathbb{R}^\mathfrak{d}, \mathbb{R})$, let $U \subseteq \mathbb{R}^\mathfrak{d}$ be a set, and let $\vartheta \in U$. Then we say that \mathcal{L} satisfies the generalized KL inequality at ϑ on U with exponent α and constant c (we say that \mathcal{L} satisfies the generalized KL inequality at ϑ) if and only if for all $\theta \in U$ it holds that

$$|\mathcal{L}(\vartheta) - \mathcal{L}(\theta)|^\alpha \leq c |\mathbb{S}_{\mathcal{L}}(\theta)| \quad (9.340)$$

(cf. Definition 9.16.6).

Definition 9.16.8 (Generalized **KL** functions). *Let $\mathfrak{d} \in \mathbb{N}$, $\mathcal{L} \in C(\mathbb{R}^{\mathfrak{d}}, \mathbb{R})$. Then we say that \mathcal{L} is a generalized **KL** function if and only if for all $\vartheta \in \mathbb{R}^{\mathfrak{d}}$ there exist $\varepsilon, c \in (0, \infty)$, $\alpha \in (0, 1)$ such that for all $\theta \in \{v \in \mathbb{R}^{\mathfrak{d}} : \|v - \vartheta\|_2 < \varepsilon\}$ it holds that*

$$|\mathcal{L}(\vartheta) - \mathcal{L}(\theta)|^{\alpha} \leq c |\mathbb{S}_{\mathcal{L}}(\theta)| \quad (9.341)$$

(cf. Definitions 3.3.4 and 9.16.6).

Remark 9.16.9 (Examples and convergence results for generalized **KL** functions). In Theorem 9.10.1 and Corollary 9.14.5 above we have seen that in the case of an analytic activation function we have that the associated empirical risk function is also analytic and therefore a standard **KL** function. In deep learning algorithms often deep **ANNs** with non-analytic activation functions such as the **ReLU** activation (cf. Section 1.2.3) and the leaky **ReLU** activation (cf. Section 1.2.11) are used. In the case of such non-differentiable activation functions, the associated risk function is typically not a standard **KL** function. However, under suitable assumptions on the target function and the underlying probability measure of the input data of the considered learning problem, using Bolte et al. [46, Theorem 3.1] one can verify in the case of such non-differentiable activation functions that the risk function is a generalized **KL** function in the sense of Definition 9.16.8 above; cf., for instance, [132, 236]. Similar as for standard **KL** functions (cf., for example, Dereich & Kassing [106] and Sections 9.12 and 9.13) one can then also develop a convergence theory for gradient based optimization methods for generalized **KL** function (cf., for instance, Bolte et al. [46, Section 4] and Corollary 9.12.5).

Remark 9.16.10 (Further convergence analyses). We refer, for example, to [2, 7, 8, 46, 106, 412] and the references therein for convergence analyses under **KL**-type conditions for gradient based optimization methods in the literature. Beyond the **KL** approach reviewed in this chapter there are also several other approaches in the literature with which one can conclude convergence of gradient based optimization methods to suitable generalized critical points; cf., for instance, [47, 67, 96] and the references therein.

9.17 Non-convergence for stochastic gradient descent

In Sections 9.12 and 9.13 above we present convergence results for gradient based optimization procedures to critical points. In general one cannot expect that these limiting critical points are global minimizers. This is the subject of the next statement, Theorem 9.17.1 below. Theorem 9.17.1 shows in particular that the probability to converge to a global minimizer converges to zero as the number of parameters of the **ANN** architecture converges to infinity.

Theorem 9.17.1. Let $d \in \mathbb{N}$, $a \in \mathbb{R}$, $b \in (a, \infty)$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, for every $m, n \in \mathbb{N}_0$ let $X_n^m: \Omega \rightarrow [a, b]^d$ and $Y_n^m: \Omega \rightarrow \mathbb{R}$ be random variables, assume for all $i \in \mathbb{N}$, $j \in \mathbb{N} \setminus \{i\}$ that $\mathbb{P}(X_0^i = X_0^j) = 0$, let $f: \mathbb{N} \rightarrow \mathbb{N}$ be a function, for every $k \in \mathbb{N}_0$ let $\mathfrak{d}_k, L_k \in \mathbb{N}$, $\ell_k = (\ell_k^0, \ell_k^1, \dots, \ell_k^{L_k}) \in \mathbb{N}^{L_k+1}$ satisfy $\ell_k^0 = d$, $\ell_k^{L_k} = 1$, $\mathfrak{d}_k = \sum_{i=1}^L \ell_k^i (\ell_k^{i-1} + 1)$, and $\max\{\ell_k^1, \ell_k^2, \dots, \ell_k^{L_k}\} \leq f(\ell_k^1)$, let $\mathbb{A}_r: \mathbb{R} \rightarrow \mathbb{R}$, $r \in \mathbb{N}_0$, satisfy for all $x \in \mathbb{R}$ that there exists $m \in \mathbb{N}$ such that $(\cup_{r=m}^{\infty} \{\mathbb{A}_r\}) \subseteq C^1(\mathbb{R}, \mathbb{R})$ and

$$\sum_{r=m}^{\infty} (|\mathbb{A}_0(x) - \mathbb{A}_r(x)| + |\mathbb{1}_{(0,\infty)}(x) - (\mathbb{A}_r)'(x)| + |\max\{0, x\} - \mathbb{A}_0(x)|) = 0, \quad (9.342)$$

for every $r, k \in \mathbb{N}_0$, $v \in \{1, 2, \dots, L_k\}$ let $\Psi_v^{r,k}: \mathbb{R}^{\ell_k^v} \rightarrow \mathbb{R}^{\ell_k^v}$ satisfy

$$\Psi_v^{r,k} = \begin{cases} \mathfrak{M}_{\mathbb{A}_r, \ell_k^v} & : v < L_k \\ \text{id}_{\mathbb{R}^{\ell_k^v}} & : v = L_k, \end{cases} \quad (9.343)$$

for every $k, n \in \mathbb{N}_0$ let $M_n^k \in \mathbb{N}$, $\gamma_n^k \in \mathbb{R}$, for every $r, k, n \in \mathbb{N}_0$ let $\mathcal{L}_n^{r,k}: \mathbb{R}^{\mathfrak{d}_k} \times \Omega \rightarrow \mathbb{R}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}_k}$ that

$$\mathcal{L}_n^{r,k}(\theta) = \frac{1}{M_n^k} \left[\sum_{m=1}^{M_n^k} |(\mathcal{N}_{\Psi_1^{r,k}, \Psi_2^{r,k}, \dots, \Psi_{L_k}^{r,k}}^{\theta, d})(X_n^m) - Y_n^m|^2 \right], \quad (9.344)$$

for every $k, n \in \mathbb{N}_0$ let $\mathfrak{G}_n^k: \mathbb{R}^{\mathfrak{d}_k} \times \Omega \rightarrow \mathbb{R}^{\mathfrak{d}_k}$ satisfy for all $\omega \in \Omega$, $\theta \in \{v \in \mathbb{R}^{\mathfrak{d}_k}: (\nabla_v \mathcal{L}_n^{r,k}(v, \omega))_{r \in \mathbb{N}}$ is convergent} that

$$\mathfrak{G}_n^k(\theta, \omega) = \lim_{r \rightarrow \infty} [\nabla_{\theta} \mathcal{L}_n^{r,k}(\theta, \omega)] \quad (9.345)$$

and let $\Theta_n^k: \Omega \rightarrow \mathbb{R}^{\mathfrak{d}_k}$ be a random variable, assume for all $k, n \in \mathbb{N}$ that

$$\Theta_n^k = \Theta_{n-1}^k - \gamma_n^k \mathfrak{G}_n^k(\Theta_{n-1}^k), \quad (9.346)$$

assume $\liminf_{k \rightarrow \infty} \mathfrak{d}_k = \infty$ and $\liminf_{k \rightarrow \infty} \mathbb{P}(\inf_{\theta \in \mathbb{R}^{\mathfrak{d}_k}} \mathcal{L}_0^{0,k}(\theta) > 0) = 1$, for every $k \in \mathbb{N}$ let $c_k \in (0, \infty)$, and assume for all $k \in \mathbb{N}$ that $c_k \Theta_0^k$ is standard normal (cf. Definitions 1.1.3 and 1.2.1). Then

$$\liminf_{k \rightarrow \infty} \mathbb{P} \left(\inf_{n \in \mathbb{N}_0} \mathcal{L}_0^{0,k}(\Theta_n^k) > \inf_{\theta \in \mathbb{R}^{\mathfrak{d}_k}} \mathcal{L}_0^{0,k}(\theta) \right) = 1. \quad (9.347)$$

Theorem 9.17.1 is, in a slightly modified form, proved as [198, Theorem 1.1] (cf. also [102, Theorem 1.1]). We also refer, for example, to [72, 103, 104, 148, 190, 239, 305, 372] for further lower bounds and non-convergence results for SGD optimization methods.

Chapter 10

ANNs with batch normalization

In data-driven learning problems popular methods that aim to accelerate ANN training procedures are BN methods. In this chapter we rigorously review such methods in detail. In the literature BN methods have first been introduced in Ioffe & Szegedi [228].

Further investigation on BN techniques and applications of such methods can, for instance, be found in [4, Section 12.3.3], [137, Section 6.2.3], [171, Section 8.7.1], and [41, 385].

10.1 Batch normalization (BN)

Definition 10.1.1 (Batch). *Let $d, M \in \mathbb{N}$. Then we say that x is a batch of d -dimensional data points of size M (we say that x is a batch of M d -dimensional data points, we say that x is a batch) if and only if it holds that $x \in (\mathbb{R}^d)^M$.*

Definition 10.1.2 (Batch mean). *Let $d, M \in \mathbb{N}$, $x = (x^{(m)})_{m \in \{1, 2, \dots, M\}} \in (\mathbb{R}^d)^M$. Then we denote by $\text{Batchmean}(x) = (\text{Batchmean}_1(x), \dots, \text{Batchmean}_d(x)) \in \mathbb{R}^d$ the vector given by*

$$\text{Batchmean}(x) = \frac{1}{M} \left[\sum_{m=1}^M x^{(m)} \right] \quad (10.1)$$

and we call $\text{Batchmean}(x)$ the batch mean of the batch x .

Definition 10.1.3 (Batch variance). *Let $d, M \in \mathbb{N}$, $x = ((x_i^{(m)})_{i \in \{1, 2, \dots, d\}})_{m \in \{1, 2, \dots, M\}} \in$*

$(\mathbb{R}^d)^M$. Then we denote by

$$\text{Batchvar}(x) = (\text{Batchvar}_1(x), \dots, \text{Batchvar}_d(x)) \in \mathbb{R}^d \quad (10.2)$$

the vector which satisfies for all $i \in \{1, 2, \dots, d\}$ that

$$\text{Batchvar}_i(x) = \frac{1}{M} \left[\sum_{m=1}^M (x_i^{(m)} - \text{Batchmean}_i(x))^2 \right] \quad (10.3)$$

and we call $\text{Batchvar}(x)$ the batch variance of the batch x (cf. Definition 10.1.2).

Lemma 10.1.4. Let $d, M \in \mathbb{N}$, $x = (x^{(m)})_{m \in \{1, 2, \dots, M\}} = ((x_i^{(m)})_{i \in \{1, 2, \dots, d\}})_{m \in \{1, 2, \dots, M\}} \in (\mathbb{R}^d)^M$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $U: \Omega \rightarrow \{1, 2, \dots, M\}$ be a $\{1, 2, \dots, M\}$ -uniformly distributed random variable. Then

- (i) it holds that $\text{Batchmean}(x) = \mathbb{E}[x^{(U)}]$ and
- (ii) it holds for all $i \in \{1, 2, \dots, d\}$ that $\text{Batchvar}_i(x) = \text{Var}(x_i^{(U)})$.

Proof of Lemma 10.1.4. Note that (10.1) proves item (i). Furthermore, note that item (i) and (10.3) show item (ii). The proof of Lemma 10.1.4 is thus complete. \square

Definition 10.1.5 (**BN** operations for given batch mean and batch variance). Let $d \in \mathbb{N}$, $\varepsilon \in (0, \infty)$, $\beta = (\beta_1, \dots, \beta_d)$, $\gamma = (\gamma_1, \dots, \gamma_d)$, $\mu = (\mu_1, \dots, \mu_d) \in \mathbb{R}^d$, $V = (V_1, \dots, V_d) \in [0, \infty)^d$. Then we denote by

$$\text{batchnorm}_{\beta, \gamma, \mu, V, \varepsilon}: \mathbb{R}^d \rightarrow \mathbb{R}^d \quad (10.4)$$

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

$$\text{batchnorm}_{\beta, \gamma, \mu, V, \varepsilon}(x) = \left(\gamma_i \left[\frac{x_i - \mu_i}{\sqrt{V_i + \varepsilon}} \right] + \beta_i \right)_{i \in \{1, 2, \dots, d\}} \quad (10.5)$$

and we call $\text{batchnorm}_{\beta, \gamma, \mu, V, \varepsilon}$ the **BN** operation with mean parameter β , standard deviation parameter γ , and regularization parameter ε given the batch mean μ and batch variance V .

Definition 10.1.6 (Batch normalization). Let $d \in \mathbb{N}$, $\varepsilon \in (0, \infty)$, $\beta, \gamma \in \mathbb{R}^d$. Then we denote by

$$\text{Batchnorm}_{\beta, \gamma, \varepsilon}: (\bigcup_{M \in \mathbb{N}} (\mathbb{R}^d)^M) \rightarrow (\bigcup_{M \in \mathbb{N}} (\mathbb{R}^d)^M) \quad (10.6)$$

10.1. Batch normalization (BN)

the function which satisfies for all $M \in \mathbb{N}$, $x = (x^{(m)})_{m \in \{1, 2, \dots, M\}} \in (\mathbb{R}^d)^M$ that

$$\text{Batchnorm}_{\beta, \gamma, \varepsilon}(x) = (\text{batchnorm}_{\beta, \gamma, \text{Batchmean}(x), \text{Batchvar}(x), \varepsilon}(x^{(m)}))_{m \in \{1, 2, \dots, M\}} \in (\mathbb{R}^d)^M \quad (10.7)$$

and we call $\text{Batchnorm}_{\beta, \gamma, \varepsilon}$ the **BN** with mean parameter β , standard deviation parameter γ , and regularization parameter ε (cf. Definitions 10.1.2, 10.1.3, and 10.1.5).

Lemma 10.1.7. Let $d, M \in \mathbb{N}$, $\beta = (\beta_1, \dots, \beta_d)$, $\gamma = (\gamma_1, \dots, \gamma_d) \in \mathbb{R}^d$. Then

(i) it holds for all $\varepsilon \in (0, \infty)$, $x = ((x_i^{(m)})_{i \in \{1, 2, \dots, d\}})_{m \in \{1, 2, \dots, M\}} \in (\mathbb{R}^d)^M$ that

$$\text{Batchnorm}_{\beta, \gamma, \varepsilon}(x) = \left(\left(\gamma_i \left[\frac{x_i^{(m)} - \text{Batchmean}_i(x)}{\sqrt{\text{Batchvar}_i(x) + \varepsilon}} \right] + \beta_i \right)_{i \in \{1, 2, \dots, d\}} \right)_{m \in \{1, 2, \dots, M\}}, \quad (10.8)$$

(ii) it holds for all $\varepsilon \in (0, \infty)$, $x \in (\mathbb{R}^d)^M$ that

$$\text{Batchmean}(\text{Batchnorm}_{\beta, \gamma, \varepsilon}(x)) = \beta, \quad (10.9)$$

and

(iii) it holds for all $x = ((x_i^{(m)})_{i \in \{1, 2, \dots, d\}})_{m \in \{1, 2, \dots, M\}} \in (\mathbb{R}^d)^M$, $i \in \{1, 2, \dots, d\}$ with $\#(\bigcup_{m=1}^M \{x_i^{(m)}\}) > 1$ that

$$\limsup_{\varepsilon \searrow 0} |\text{Batchvar}_i(\text{Batchnorm}_{\beta, \gamma, \varepsilon}(x)) - (\gamma_i)^2| = 0 \quad (10.10)$$

(cf. Definitions 10.1.2, 10.1.3, and 10.1.6).

Proof of Lemma 10.1.7. Note that (10.1), (10.3), (10.5), and (10.7) imply item (i). In addition, note that item (i) ensures that for all $\varepsilon \in (0, \infty)$, $x = ((x_i^{(m)})_{i \in \{1, 2, \dots, d\}})_{m \in \{1, 2, \dots, M\}} \in (\mathbb{R}^d)^M$, $i \in \{1, 2, \dots, d\}$ it holds that

$$\begin{aligned} \text{Batchmean}_i(\text{Batchnorm}_{\beta, \gamma, \varepsilon}(x)) &= \frac{1}{M} \sum_{m=1}^M \left(\gamma_i \left[\frac{x_i^{(m)} - \text{Batchmean}_i(x)}{\sqrt{\text{Batchvar}_i(x) + \varepsilon}} \right] + \beta_i \right) \\ &= \gamma_i \left[\frac{\frac{1}{M} \left(\sum_{m=1}^M x_i^{(m)} \right) - \text{Batchmean}_i(x)}{\sqrt{\text{Batchvar}_i(x) + \varepsilon}} \right] + \beta_i \\ &= \gamma_i \left[\frac{\text{Batchmean}_i(x) - \text{Batchmean}_i(x)}{\sqrt{\text{Batchvar}_i(x) + \varepsilon}} \right] + \beta_i = \beta_i \end{aligned} \quad (10.11)$$

(cf. Definitions 10.1.2, 10.1.3, and 10.1.6). This implies item (ii). Furthermore, observe that (10.11) and item (i) establish that for all $\varepsilon \in (0, \infty)$, $x = ((x_i^{(m)})_{i \in \{1, 2, \dots, d\}})_{m \in \{1, 2, \dots, M\}} \in (\mathbb{R}^d)^M$, $i \in \{1, 2, \dots, d\}$ it holds that

$$\begin{aligned} & \text{Batchvar}_i(\text{Batchnorm}_{\beta, \gamma, \varepsilon}(x)) \\ &= \frac{1}{M} \sum_{m=1}^M \left[\gamma_i \left[\frac{x_i^{(m)} - \text{Batchmean}_i(x)}{\sqrt{\text{Batchvar}_i(x) + \varepsilon}} \right] + \beta_i - \text{Batchmean}_i(\text{Batchnorm}_{\beta, \gamma, \varepsilon}(x)) \right]^2 \\ &= \frac{1}{M} \sum_{m=1}^M (\gamma_i)^2 \left[\frac{x_i^{(m)} - \text{Batchmean}_i(x)}{\sqrt{\text{Batchvar}_i(x) + \varepsilon}} \right]^2 \\ &= (\gamma_i)^2 \left[\frac{\frac{1}{M} \sum_{m=1}^M (x_i^{(m)} - \text{Batchmean}_i(x))^2}{\text{Batchvar}_i(x) + \varepsilon} \right] = (\gamma_i)^2 \left[\frac{\text{Batchvar}_i(x)}{\text{Batchvar}_i(x) + \varepsilon} \right]. \end{aligned} \quad (10.12)$$

Combining this with the fact that for all $x = ((x_i^{(m)})_{i \in \{1, 2, \dots, d\}})_{m \in \{1, 2, \dots, M\}} \in (\mathbb{R}^d)^M$, $i \in \{1, 2, \dots, d\}$ with $\#(\bigcup_{m=1}^M \{x_i^{(m)}\}) > 1$ it holds that

$$\text{Batchvar}_i(x) > 0 \quad (10.13)$$

implies item (iii). The proof of Lemma 10.1.7 is thus complete. \square

10.2 Structured description of ANNs with BN for training

Definition 10.2.1 (Structured description of fully-connected feedforward ANNs with BN). We denote by \mathbf{B} the set given by

$$\mathbf{B} = \bigcup_{L \in \mathbb{N}} \bigcup_{l_0, l_1, \dots, l_L \in \mathbb{N}} \bigcup_{N \subseteq \{0, 1, \dots, L\}} \left(\left(\bigtimes_{k=1}^L (\mathbb{R}^{l_k \times l_{k-1}} \times \mathbb{R}^{l_k}) \right) \times \left(\bigtimes_{k \in N} (\mathbb{R}^{l_k})^2 \right) \right). \quad (10.14)$$

Definition 10.2.2 (Fully-connected feedforward ANNs with BN). We say that Φ is a fully-connected feedforward ANN with BN (we say that Φ is an ANN with BN) if and only if it holds that

$$\Phi \in \mathbf{B} \quad (10.15)$$

(cf. Definition 10.2.1).

10.3 Realizations of fully-connected feedforward ANNs with BN for training

In the next definition we apply the multi-dimensional version of Definition 1.2.1 with batches as input. For this we implicitly identify batches with matrices. This identification is exemplified in the following exercise.

Exercise 10.3.1. Let $l_0 = 2$, $l_1 = 3$, $M = 4$, $W \in \mathbb{R}^{l_1 \times l_0}$, $B \in \mathbb{R}^{l_1}$, $y \in (\mathbb{R}^{l_0})^M$, $x \in (\mathbb{R}^{l_1})^M$ satisfy

$$W = \begin{pmatrix} 3 & -1 \\ -1 & 3 \\ 3 & -1 \end{pmatrix}, \quad B = \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}, \quad y = \left(\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ -2 \end{pmatrix}, \begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix} \right), \quad (10.16)$$

and $x = \mathfrak{M}_{r,l_1,M}(Wy + (B, B, B, B))$ (cf. Definitions 1.2.1 and 1.2.4). Prove the following statement: It holds that

$$x = \left(\begin{pmatrix} 0 \\ 2 \\ 0 \end{pmatrix}, \begin{pmatrix} 4 \\ 0 \\ 4 \end{pmatrix}, \begin{pmatrix} 9 \\ 0 \\ 9 \end{pmatrix}, \begin{pmatrix} 0 \\ 3 \\ 0 \end{pmatrix} \right). \quad (10.17)$$

Definition 10.3.1 (Realizations associated to fully-connected feedforward ANNs with BN). Let $\varepsilon \in (0, \infty)$, $a \in C(\mathbb{R}, \mathbb{R})$. Then we denote by

$$\mathcal{R}_{a,\varepsilon}^{\mathbf{B}} : \mathbf{B} \rightarrow \left(\bigcup_{k,l \in \mathbb{N}} C(\bigcup_{M \in \mathbb{N}} (\mathbb{R}^k)^M, \bigcup_{M \in \mathbb{N}} (\mathbb{R}^l)^M) \right) \quad (10.18)$$

the function which satisfies for all $L, M \in \mathbb{N}$, $l_0, l_1, \dots, l_L \in \mathbb{N}$, $N \subseteq \{0, 1, \dots, L\}$, $\Phi = (((W_k, B_k))_{k \in \{1, 2, \dots, L\}}, ((\beta_k, \gamma_k))_{k \in N}) \in (\bigtimes_{k=1}^L (\mathbb{R}^{l_k \times l_{k-1}} \times \mathbb{R}^{l_k})) \times (\bigtimes_{k \in N} (\mathbb{R}^{l_k})^2)$, $x_0, y_0 \in (\mathbb{R}^{l_0})^M$, $x_1, y_1 \in (\mathbb{R}^{l_1})^M$, \dots , $x_L, y_L \in (\mathbb{R}^{l_L})^M$ with

$$\forall k \in \{0, 1, \dots, L\} : \quad y_k = \begin{cases} \text{Batchnorm}_{\beta_k, \gamma_k, \varepsilon}(x_k) & : k \in N \\ x_k & : k \notin N \end{cases} \quad \text{and} \quad (10.19)$$

$$\forall k \in \{1, 2, \dots, L\} : \quad x_k = \mathfrak{M}_{a \mathbb{1}_{(0,L)}(k) + \text{id}_{\mathbb{R}} \mathbb{1}_{\{L\}}(k), l_k, M}(W_k y_{k-1} + (B_k, B_k, \dots, B_k)) \quad (10.20)$$

that

$$\mathcal{R}_{a,\varepsilon}^{\mathbf{B}}(\Phi) \in C(\bigcup_{B \in \mathbb{N}} (\mathbb{R}^{l_0})^B, \bigcup_{B \in \mathbb{N}} (\mathbb{R}^{l_L})^B) \quad \text{and} \quad (\mathcal{R}_{a,\varepsilon}^{\mathbf{B}}(\Phi))(x_0) = y_L \in (\mathbb{R}^{l_L})^M \quad (10.21)$$

and for every $\Phi \in \mathbf{B}$ we call $\mathcal{R}_{a,\varepsilon}^{\mathbf{B}}(\Phi)$ the realization function of the fully-connected feedforward ANN with BN Φ with activation function a and BN regularization parameter

ε (we call $\mathcal{R}_{a,\varepsilon}^{\mathbf{B}}(\Phi)$ the realization of the fully-connected feedforward ANN with BN Φ with activation a and BN regularization parameter ε) (cf. Definitions 1.2.1, 10.1.6, and 10.2.1).

10.4 Structured description of ANNs with BN for inference

Definition 10.4.1 (Structured description of fully-connected feedforward ANNs with BN for given batch means and batch variances). We denote by \mathbf{b} the set given by

$$\mathbf{b} = \bigcup_{L \in \mathbb{N}} \bigcup_{l_0, l_1, \dots, l_L \in \mathbb{N}} \bigcup_{N \subseteq \{0, 1, \dots, L\}} \left((\bigtimes_{k=1}^L (\mathbb{R}^{l_k \times l_{k-1}} \times \mathbb{R}^{l_k})) \times (\bigtimes_{k \in N} ((\mathbb{R}^{l_k})^3 \times [0, \infty)^{l_k})) \right). \quad (10.22)$$

Definition 10.4.2 (Fully-connected feedforward ANNs with BN for given batch means and batch variances). We say that Φ is a fully-connected feedforward ANN with BN for given batch means and batch variances (we say that Φ is an ANN with BN for given batch means and batch variances) if and only if it holds that

$$\Phi \in \mathbf{b} \quad (10.23)$$

(cf. Definition 10.4.1).

10.5 Realizations of ANNs with BN for inference

Definition 10.5.1 (Realizations associated to fully-connected feedforward ANNs with BN for given batch means and batch variances). Let $\varepsilon \in (0, \infty)$, $a \in C(\mathbb{R}, \mathbb{R})$. Then we denote by

$$\mathcal{R}_{a,\varepsilon}^{\mathbf{b}}: \mathbf{b} \rightarrow (\bigcup_{k,l \in \mathbb{N}} C(\mathbb{R}^k, \mathbb{R}^l)) \quad (10.24)$$

the function which satisfies for all $L \in \mathbb{N}$, $l_0, l_1, \dots, l_L \in \mathbb{N}$, $N \subseteq \{0, 1, \dots, L\}$, $\Phi = (((W_k, B_k))_{k \in \{1, 2, \dots, L\}}, ((\beta_k, \gamma_k, \mu_k, V_k))_{k \in N}) \in (\bigtimes_{k=1}^L (\mathbb{R}^{l_k \times l_{k-1}} \times \mathbb{R}^{l_k})) \times (\bigtimes_{k \in N} ((\mathbb{R}^{l_k})^3 \times [0, \infty)^{l_k}))$, $x_0, y_0 \in \mathbb{R}^{l_0}$, $x_1, y_1 \in \mathbb{R}^{l_1}$, ..., $x_L, y_L \in \mathbb{R}^{l_L}$ with

$$\forall k \in \{0, 1, \dots, L\}: \quad y_k = \begin{cases} \text{batchnorm}_{\beta_k, \gamma_k, \mu_k, V_k, \varepsilon}(x_k) & : k \in N \\ x_k & : k \notin N \end{cases} \quad \text{and} \quad (10.25)$$

$$\forall k \in \{1, 2, \dots, L\}: \quad x_k = \mathfrak{M}_{a\mathbb{1}_{(0,L)}(k) + \text{id}_{\mathbb{R}}\mathbb{1}_{\{L\}}(k), l_k}(W_k y_{k-1} + B_k) \quad (10.26)$$

that

$$\mathcal{R}_{a,\varepsilon}^{\mathbf{b}}(\Phi) \in C(\mathbb{R}^{l_0}, \mathbb{R}^{l_L}) \quad \text{and} \quad (\mathcal{R}_{a,\varepsilon}^{\mathbf{b}}(\Phi))(x_0) = y_L \quad (10.27)$$

and for every $\Phi \in \mathbf{b}$ we call $\mathcal{R}_{a,\varepsilon}^{\mathbf{b}}(\Phi)$ the realization function of the fully-connected feedforward ANN with BN for given batch means and batch variances Φ with activation function a and BN regularization parameter ε (cf. Definitions 10.1.5 and 10.4.1).

10.6 On the connection between BN for training and BN for inference

Definition 10.6.1 (Fully-connected feed-forward ANNs with BN for given batch means and batch variances associated to fully-connected feedforward ANNs with BN and given input batches). Let $\varepsilon \in (0, \infty)$, $a \in C(\mathbb{R}, \mathbb{R})$, $L, M \in \mathbb{N}$, $l_0, l_1, \dots, l_L \in \mathbb{N}$, $N \subseteq \{0, 1, \dots, L\}$, $\Phi = (((W_k, B_k))_{k \in \{1, 2, \dots, L\}}, ((\beta_k, \gamma_k))_{k \in N}) \in (\bigtimes_{k=1}^L (\mathbb{R}^{l_k \times l_{k-1}} \times \mathbb{R}^{l_k})) \times (\bigtimes_{k \in N} (\mathbb{R}^{l_k})^2)$, $\mathfrak{x} \in (\mathbb{R}^{l_0})^M$. Then we say that Ψ is the fully-connected feedforward ANNs with BN for given batch means and batch variances associated to $(\Phi, \mathfrak{x}, a, \varepsilon)$ if and only if there exists $x_0, y_0 \in (\mathbb{R}^{l_0})^M$, $x_1, y_1 \in (\mathbb{R}^{l_1})^M, \dots, x_L, y_L \in (\mathbb{R}^{l_L})^M$ such that

- (i) it holds that $x_0 = \mathfrak{x}$,
- (ii) it holds for all $k \in \{0, 1, \dots, L\}$ that

$$y_k = \begin{cases} \text{Batchnorm}_{\beta_k, \gamma_k, \varepsilon}(x_k) & : k \in N \\ x_k & : k \notin N, \end{cases} \quad (10.28)$$

- (iii) it holds for all $k \in \{1, 2, \dots, L\}$ that

$$x_k = \mathfrak{M}_{a\mathbb{1}_{(0,L)}(k) + \text{id}_{\mathbb{R}}\mathbb{1}_{\{L\}}(k), l_k, M}(W_k y_{k-1} + (B_k, B_k, \dots, B_k)), \quad (10.29)$$

and

- (iv) it holds that

$$\begin{aligned} \Psi &= (((W_k, B_k))_{k \in \{1, 2, \dots, L\}}, ((\beta_k, \gamma_k, \text{Batchmean}(x_k), \text{Batchvar}(x_k)))_{k \in N}) \\ &\in (\bigtimes_{k=1}^L (\mathbb{R}^{l_k \times l_{k-1}} \times \mathbb{R}^{l_k})) \times (\bigtimes_{k \in N} (\mathbb{R}^{l_k})^4) \end{aligned} \quad (10.30)$$

(cf. Definitions 1.2.1, 10.1.2, 10.1.3, and 10.1.6).

Lemma 10.6.2. Let $\varepsilon \in (0, \infty)$, $a \in C(\mathbb{R}, \mathbb{R})$, $L, M \in \mathbb{N}$, $l_0, l_1, \dots, l_L \in \mathbb{N}$, $N \subseteq \{0, 1, \dots, L\}$, $\Phi = (((W_k, B_k))_{k \in \{1, 2, \dots, L\}}, ((\beta_k, \gamma_k))_{k \in N}) \in (\bigtimes_{k=1}^L (\mathbb{R}^{l_k \times l_{k-1}} \times \mathbb{R}^{l_k})) \times (\bigtimes_{k \in N} (\mathbb{R}^{l_k})^2)$, $x = (x^{(m)})_{m \in \{1, 2, \dots, M\}} \in (\mathbb{R}^{l_0})^M$ and let Ψ be the fully-connected feedforward ANN with BN for given batch means and batch variances associated to $(\Phi, x, a, \varepsilon)$ (cf. Definition 10.6.1). Then

$$(\mathcal{R}_{a, \varepsilon}^{\mathbf{B}}(\Phi))(x) = ((\mathcal{R}_{a, \varepsilon}^{\mathbf{b}}(\Psi))(x^{(m)}))_{m \in \{1, 2, \dots, M\}} \quad (10.31)$$

(cf. Definitions 10.3.1 and 10.5.1).

Proof of Lemma 10.6.2. Observe that (10.19), (10.20), (10.21), (10.25), (10.26), (10.27), (10.28), (10.29), and (10.30) establish (10.31). The proof of Lemma 10.6.2 is thus complete. \square

Exercise 10.6.1. Let $l_0 = 2$, $l_1 = 3$, $l_2 = 1$, $N = \{0, 1\}$, $\gamma_0 = (2, 2)$, $\beta_0 = (0, 0)$, $\gamma_1 = (1, 1, 1)$, $\beta_1 = (0, 1, 0)$, $x = ((0, 1), (1, 0), (-2, 2), (2, -2))$, $\Phi \in \mathbf{B}$ satisfy

$$\begin{aligned} \Phi &= \left(\left(\left(\begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{pmatrix}, \begin{pmatrix} -1 \\ -1 \end{pmatrix} \right), \left(\begin{pmatrix} -1 & 1 & -1 \\ 1 & -1 & 1 \end{pmatrix}, (-2) \right) \right), ((\gamma_k, \beta_k))_{k \in N} \right) \\ &\in (\bigtimes_{k=1}^2 (\mathbb{R}^{l_k \times l_{k-1}} \times \mathbb{R}^{l_k})) \times (\bigtimes_{k \in N} (\mathbb{R}^{l_k})^2) \end{aligned} \quad (10.32)$$

and let $\Psi \in \mathbf{b}$ be the fully-connected feedforward ANNs with BN for given batch means and batch variances associated to $(\Phi, x, \mathfrak{r}, 0.01)$. Compute $(\mathcal{R}_{\mathfrak{r}, \frac{1}{100}}^{\mathbf{B}}(\Phi))(x)$ and $(\mathcal{R}_{\mathfrak{r}, \frac{1}{100}}^{\mathbf{b}}(\Psi))(-1, 1)$ explicitly and prove that your results are correct (cf. Definitions 1.2.4, 10.2.1, 10.3.1, 10.4.1, 10.5.1, and 10.6.1)!

Chapter 11

Optimization through random initializations

In addition to minimizing an objective function through iterative steps of an SGD-type optimization method, another approach to minimize an objective function is to sample different random initializations, to iteratively calculate SGD optimization processes starting at these random initializations, and, thereafter, to pick a SGD trajectory with the smallest final evaluation of the objective function. The approach to consider different random initializations is reviewed and analyzed within this chapter in detail. The specific presentation of this chapter is strongly based on Jentzen & Welti [243, Section 5].

11.1 Analysis of the optimization error

11.1.1 The complementary distribution function formula

Lemma 11.1.1 (Complementary distribution function formula). *Let $\mu: \mathcal{B}([0, \infty)) \rightarrow [0, \infty]$ be a sigma-finite measure. Then*

$$\int_0^\infty x \mu(dx) = \int_0^\infty \mu([x, \infty)) dx = \int_0^\infty \mu((x, \infty)) dx. \quad (11.1)$$

Proof of Lemma 11.1.1. First, note that

$$\begin{aligned} \int_0^\infty x \mu(dx) &= \int_0^\infty \left[\int_0^x dy \right] \mu(dx) = \int_0^\infty \left[\int_0^\infty \mathbb{1}_{(-\infty, x]}(y) dy \right] \mu(dx) \\ &= \int_0^\infty \int_0^\infty \mathbb{1}_{[y, \infty)}(x) dy \mu(dx). \end{aligned} \quad (11.2)$$

Furthermore, observe that the fact that $[0, \infty)^2 \ni (x, y) \mapsto \mathbb{1}_{[y, \infty)}(x) \in \mathbb{R}$ is $(\mathcal{B}([0, \infty)) \otimes \mathcal{B}([0, \infty))) / \mathcal{B}(\mathbb{R})$ -measurable, the assumption that μ is a sigma-finite measure, and Fubini's

theorem demonstrate that

$$\int_0^\infty \int_0^\infty \mathbb{1}_{[y,\infty)}(x) dy \mu(dx) = \int_0^\infty \int_0^\infty \mathbb{1}_{[y,\infty)}(x) \mu(dx) dy = \int_0^\infty \mu([y, \infty)) dy. \quad (11.3)$$

Combining this with (11.2) proves that for all $\varepsilon \in (0, \infty)$ it holds that

$$\begin{aligned} \int_0^\infty x \mu(dx) &= \int_0^\infty \mu([y, \infty)) dy \geq \int_0^\infty \mu((y, \infty)) dy \\ &\geq \int_0^\infty \mu([y + \varepsilon, \infty)) dy = \int_\varepsilon^\infty \mu([y, \infty)) dy. \end{aligned} \quad (11.4)$$

Beppo Levi's monotone convergence theorem therefore shows that

$$\begin{aligned} \int_0^\infty x \mu(dx) &= \int_0^\infty \mu([y, \infty)) dy \geq \int_0^\infty \mu((y, \infty)) dy \\ &\geq \sup_{\varepsilon \in (0, \infty)} \left[\int_\varepsilon^\infty \mu([y, \infty)) dy \right] \\ &= \sup_{\varepsilon \in (0, \infty)} \left[\int_0^\infty \mu([y, \infty)) \mathbb{1}_{(\varepsilon, \infty)}(y) dy \right] = \int_0^\infty \mu([y, \infty)) dy. \end{aligned} \quad (11.5)$$

The proof of Lemma 11.1.1 is thus complete. \square

11.1.2 Estimates for the optimization error involving complementary distribution functions

Lemma 11.1.2. Let (E, δ) be a metric space, let $x \in E$, $K \in \mathbb{N}$, $p, L \in (0, \infty)$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\mathcal{R}: E \times \Omega \rightarrow \mathbb{R}$ be $(\mathcal{B}(E) \otimes \mathcal{F})/\mathcal{B}(\mathbb{R})$ -measurable, assume for all $y \in E$, $\omega \in \Omega$ that $|\mathcal{R}(x, \omega) - \mathcal{R}(y, \omega)| \leq L\delta(x, y)$, and let $X_k: \Omega \rightarrow E$, $k \in \{1, 2, \dots, K\}$, be i.i.d. random variables. Then

$$\mathbb{E}[\min_{k \in \{1, 2, \dots, K\}} |\mathcal{R}(X_k) - \mathcal{R}(x)|^p] \leq L^p \int_0^\infty [\mathbb{P}(\delta(X_1, x) > \varepsilon^{1/p})]^K d\varepsilon. \quad (11.6)$$

Proof of Lemma 11.1.2. Throughout this proof, let $Y: \Omega \rightarrow [0, \infty)$ satisfy for all $\omega \in \Omega$ that $Y(\omega) = \min_{k \in \{1, 2, \dots, K\}} [\delta(X_k(\omega), x)]^p$. Note that the fact that Y is a random variable, the assumption that $\forall y \in E, \omega \in \Omega: |\mathcal{R}(x, \omega) - \mathcal{R}(y, \omega)| \leq L\delta(x, y)$, and Lemma 11.1.1 imply that

$$\begin{aligned} \mathbb{E}[\min_{k \in \{1, 2, \dots, K\}} |\mathcal{R}(X_k) - \mathcal{R}(x)|^p] &\leq L^p \mathbb{E}[\min_{k \in \{1, 2, \dots, K\}} [\delta(X_k, x)]^p] \\ &= L^p \mathbb{E}[Y] = L^p \int_0^\infty y \mathbb{P}_Y(dy) = L^p \int_0^\infty \mathbb{P}_Y((\varepsilon, \infty)) d\varepsilon \\ &= L^p \int_0^\infty \mathbb{P}(Y > \varepsilon) d\varepsilon = L^p \int_0^\infty \mathbb{P}(\min_{k \in \{1, 2, \dots, K\}} [\delta(X_k, x)]^p > \varepsilon) d\varepsilon. \end{aligned} \quad (11.7)$$

Furthermore, observe that the assumption that X_k , $k \in \{1, 2, \dots, K\}$, are i.i.d. random variables ensures that for all $\varepsilon \in (0, \infty)$ it holds that

$$\begin{aligned} \mathbb{P}\left(\min_{k \in \{1, 2, \dots, K\}} [\delta(X_k, x)]^p > \varepsilon\right) &= \mathbb{P}\left(\forall k \in \{1, 2, \dots, K\}: [\delta(X_k, x)]^p > \varepsilon\right) \\ &= \prod_{k=1}^K \mathbb{P}([\delta(X_k, x)]^p > \varepsilon) = [\mathbb{P}([\delta(X_1, x)]^p > \varepsilon)]^K = [\mathbb{P}(\delta(X_1, x) > \varepsilon^{1/p})]^K. \end{aligned} \quad (11.8)$$

Combining this with (11.7) proves (11.6). The proof of Lemma 11.1.2 is thus complete. \square

11.2 Strong convergences rates for the optimization error

11.2.1 Properties of the gamma and the beta function

Lemma 11.2.1. *Let $\Gamma: (0, \infty) \rightarrow (0, \infty)$ and $\mathbb{B}: (0, \infty)^2 \rightarrow (0, \infty)$ satisfy for all $x, y \in (0, \infty)$ that $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$ and $\mathbb{B}(x, y) = \int_0^1 t^{x-1} (1-t)^{y-1} dt$. Then*

- (i) *it holds for all $x \in (0, \infty)$ that $\Gamma(x+1) = x \Gamma(x)$,*
- (ii) *it holds that $\Gamma(1) = \Gamma(2) = 1$, and*
- (iii) *it holds for all $x, y \in (0, \infty)$ that $\mathbb{B}(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$.*

Proof of Lemma 11.2.1. Throughout this proof, let $x, y \in (0, \infty)$, let $\Phi: (0, \infty) \times (0, 1) \rightarrow (0, \infty)^2$ satisfy for all $u \in (0, \infty)$, $v \in (0, 1)$ that

$$\Phi(u, v) = (u(1-v), uv), \quad (11.9)$$

and let $f: (0, \infty)^2 \rightarrow (0, \infty)$ satisfy for all $s, t \in (0, \infty)$ that

$$f(s, t) = s^{(x-1)} t^{(y-1)} e^{-(s+t)}. \quad (11.10)$$

Note that the integration by parts formula establishes that for all $x \in (0, \infty)$ it holds that

$$\begin{aligned} \Gamma(x+1) &= \int_0^\infty t^{((x+1)-1)} e^{-t} dt = - \int_0^\infty t^x [-e^{-t}] dt \\ &= - \left([t^x e^{-t}]_{t=0}^{t=\infty} - x \int_0^\infty t^{(x-1)} e^{-t} dt \right) = x \int_0^\infty t^{(x-1)} e^{-t} dt = x \cdot \Gamma(x). \end{aligned} \quad (11.11)$$

This proves item (i). Furthermore, observe that

$$\Gamma(1) = \int_0^\infty t^0 e^{-t} dt = [-e^{-t}]_{t=0}^{t=\infty} = 1. \quad (11.12)$$

This and item (i) establish item (ii). Moreover, note that the integral transformation theorem with the diffeomorphism $(1, \infty) \ni t \mapsto \frac{1}{t} \in (0, 1)$ demonstrates that

$$\begin{aligned} B(x, y) &= \int_0^1 t^{(x-1)} (1-t)^{(y-1)} dt = \int_1^\infty \left[\frac{1}{t}\right]^{(x-1)} \left[1 - \frac{1}{t}\right]^{(y-1)} \frac{1}{t^2} dt \\ &= \int_1^\infty t^{(-x-1)} \left[\frac{t-1}{t}\right]^{(y-1)} dt = \int_1^\infty t^{(-x-y)} (t-1)^{(y-1)} dt \\ &= \int_0^\infty (t+1)^{(-x-y)} t^{(y-1)} dt = \int_0^\infty \frac{t^{(y-1)}}{(t+1)^{(x+y)}} dt. \end{aligned} \quad (11.13)$$

In addition, observe that the fact that for all $(u, v) \in (0, \infty) \times (0, 1)$ it holds that

$$\Phi'(u, v) = \begin{bmatrix} 1-v & -u \\ v & u \end{bmatrix} \quad (11.14)$$

shows that for all $(u, v) \in (0, \infty) \times (0, 1)$ it holds that

$$\det(\Phi'(u, v)) = (1-v)u - v(-u) = u - vu + vu = u \in (0, \infty). \quad (11.15)$$

This, the fact that

$$\begin{aligned} \Gamma(x) \cdot \Gamma(y) &= \left[\int_0^\infty t^{(x-1)} e^{-t} dt \right] \left[\int_0^\infty t^{(y-1)} e^{-t} dt \right] \\ &= \left[\int_0^\infty s^{(x-1)} e^{-s} ds \right] \left[\int_0^\infty t^{(y-1)} e^{-t} dt \right] \\ &= \int_0^\infty \int_0^\infty s^{(x-1)} t^{(y-1)} e^{-(s+t)} dt ds \\ &= \int_{(0,\infty)^2} f(s, t) d(s, t), \end{aligned} \quad (11.16)$$

and the integral transformation theorem imply that

$$\begin{aligned} \Gamma(x) \cdot \Gamma(y) &= \int_{(0,\infty) \times (0,1)} f(\Phi(u, v)) |\det(\Phi'(u, v))| d(u, v) \\ &= \int_0^\infty \int_0^1 (u(1-v))^{(x-1)} (uv)^{(y-1)} e^{-(u(1-v)+uv)} u dv du \\ &= \int_0^\infty \int_0^1 u^{(x+y-1)} e^{-u} v^{(y-1)} (1-v)^{(x-1)} dv du \\ &= \left[\int_0^\infty u^{(x+y-1)} e^{-u} du \right] \left[\int_0^1 v^{(y-1)} (1-v)^{(x-1)} dv \right] \\ &= \Gamma(x+y) B(y, x). \end{aligned} \quad (11.17)$$

This establishes item (iii). The proof of Lemma 11.2.1 is thus complete. \square

Lemma 11.2.2. *It holds for all $\alpha, x \in [0, 1]$ that $(1 - x)^\alpha \leq 1 - \alpha x$.*

Proof of Lemma 11.2.2. Note that the fact that for all $y \in [0, \infty)$ it holds that $[0, \infty) \ni z \mapsto y^z \in [0, \infty)$ is convex ensures that for all $\alpha, x \in [0, 1]$ it holds that

$$\begin{aligned} (1 - x)^\alpha &\leq \alpha(1 - x)^1 + (1 - \alpha)(1 - x)^0 \\ &= \alpha - \alpha x + 1 - \alpha = 1 - \alpha x. \end{aligned} \quad (11.18)$$

The proof of Lemma 11.2.2 is thus complete. \square

Proposition 11.2.3. *Let $\Gamma: (0, \infty) \rightarrow (0, \infty)$ and $\lfloor \cdot \rfloor: (0, \infty) \rightarrow \mathbb{N}_0$ satisfy for all $x \in (0, \infty)$ that $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$ and $\lfloor x \rfloor = \max([0, x] \cap \mathbb{N}_0)$. Then*

- (i) *it holds that $\Gamma: (0, \infty) \rightarrow (0, \infty)$ is convex,*
- (ii) *it holds for all $x \in (0, \infty)$ that $\Gamma(x+1) = x\Gamma(x) \leq x^{\lfloor x \rfloor} \leq \max\{1, x^x\}$,*
- (iii) *it holds for all $x \in (0, \infty)$, $\alpha \in [0, 1]$ that*

$$(\max\{x + \alpha - 1, 0\})^\alpha \leq \frac{x}{(x + \alpha)^{1-\alpha}} \leq \frac{\Gamma(x + \alpha)}{\Gamma(x)} \leq x^\alpha, \quad (11.19)$$

and

- (iv) *it holds for all $x \in (0, \infty)$, $\alpha \in [0, \infty)$ that*

$$(\max\{x + \min\{\alpha - 1, 0\}, 0\})^\alpha \leq \frac{\Gamma(x + \alpha)}{\Gamma(x)} \leq (x + \max\{\alpha - 1, 0\})^\alpha. \quad (11.20)$$

Proof of Proposition 11.2.3. Throughout this proof, let $\lfloor \cdot \rfloor: [0, \infty) \rightarrow \mathbb{N}_0$ satisfy for all $x \in [0, \infty)$ that $\lfloor x \rfloor = \max([0, x] \cap \mathbb{N}_0)$. Observe that the fact that for all $t \in (0, \infty)$ it holds that $\mathbb{R} \ni x \mapsto t^x \in (0, \infty)$ is convex proves that for all $x, y \in (0, \infty)$, $\alpha \in [0, 1]$ it holds that

$$\begin{aligned} \Gamma(\alpha x + (1 - \alpha)y) &= \int_0^\infty t^{\alpha x + (1 - \alpha)y - 1} e^{-t} dt = \int_0^\infty t^{\alpha x + (1 - \alpha)y} t^{-1} e^{-t} dt \\ &\leq \int_0^\infty (\alpha t^x + (1 - \alpha)t^y) t^{-1} e^{-t} dt \\ &= \alpha \int_0^\infty t^{x-1} e^{-t} dt + (1 - \alpha) \int_0^\infty t^{y-1} e^{-t} dt \\ &= \alpha \Gamma(x) + (1 - \alpha) \Gamma(y). \end{aligned} \quad (11.21)$$

This establishes item (i). Furthermore, note that item (ii) in Lemma 11.2.1 and item (i) demonstrate that for all $\alpha \in [0, 1]$ it holds that

$$\Gamma(\alpha + 1) = \Gamma(\alpha \cdot 2 + (1 - \alpha) \cdot 1) \leq \alpha \Gamma(2) + (1 - \alpha) \Gamma(1) = \alpha + (1 - \alpha) = 1. \quad (11.22)$$

This shows for all $x \in (0, 1]$ that

$$\Gamma(x+1) \leq 1 = x^{\lfloor x \rfloor} = \max\{1, x^x\}. \quad (11.23)$$

Induction, item (i) in Lemma 11.2.1, and the fact that $\forall x \in (0, \infty): x - \lfloor x \rfloor \in (0, 1]$ hence imply that for all $x \in [1, \infty)$ it holds that

$$\Gamma(x+1) = \left[\prod_{i=1}^{\lfloor x \rfloor} (x-i+1) \right] \Gamma(x - \lfloor x \rfloor + 1) \leq x^{\lfloor x \rfloor} \Gamma(x - \lfloor x \rfloor + 1) \leq x^{\lfloor x \rfloor} \leq x^x = \max\{1, x^x\}. \quad (11.24)$$

Combining this and (11.23) with item (i) in Lemma 11.2.1 proves item (ii). Moreover, observe that Hölder's inequality and item (i) in Lemma 11.2.1 ensure that for all $x \in (0, \infty)$, $\alpha \in [0, 1]$ it holds that

$$\begin{aligned} \Gamma(x+\alpha) &= \int_0^\infty t^{x+\alpha-1} e^{-t} dt = \int_0^\infty t^{\alpha x} e^{-\alpha t} t^{(1-\alpha)x-(1-\alpha)} e^{-(1-\alpha)t} dt \\ &= \int_0^\infty [t^x e^{-t}]^\alpha [t^{x-1} e^{-t}]^{1-\alpha} dt \\ &\leq \left(\int_0^\infty t^x e^{-t} dt \right)^\alpha \left(\int_0^\infty t^{x-1} e^{-t} dt \right)^{1-\alpha} \\ &= [\Gamma(x+1)]^\alpha [\Gamma(x)]^{1-\alpha} = x^\alpha [\Gamma(x)]^\alpha [\Gamma(x)]^{1-\alpha} \\ &= x^\alpha \Gamma(x). \end{aligned} \quad (11.25)$$

This and item (i) in Lemma 11.2.1 establish that for all $x \in (0, \infty)$, $\alpha \in [0, 1]$ it holds that

$$x \Gamma(x) = \Gamma(x+1) = \Gamma(x+\alpha+(1-\alpha)) \leq (x+\alpha)^{1-\alpha} \Gamma(x+\alpha). \quad (11.26)$$

Combining (11.25) and (11.26) demonstrates that for all $x \in (0, \infty)$, $\alpha \in [0, 1]$ it holds that

$$\frac{x}{(x+\alpha)^{1-\alpha}} \leq \frac{\Gamma(x+\alpha)}{\Gamma(x)} \leq x^\alpha. \quad (11.27)$$

In addition, note that item (i) in Lemma 11.2.1 and (11.27) show that for all $x \in (0, \infty)$, $\alpha \in [0, 1]$ it holds that

$$\frac{\Gamma(x+\alpha)}{\Gamma(x+1)} = \frac{\Gamma(x+\alpha)}{x \Gamma(x)} \leq x^{\alpha-1}. \quad (11.28)$$

This implies for all $\alpha \in [0, 1]$, $x \in (\alpha, \infty)$ that

$$\frac{\Gamma(x)}{\Gamma(x+(1-\alpha))} = \frac{\Gamma((x-\alpha)+\alpha)}{\Gamma((x-\alpha)+1)} \leq (x-\alpha)^{\alpha-1} = \frac{1}{(x-\alpha)^{1-\alpha}}. \quad (11.29)$$

This, in turn, proves for all $\alpha \in [0, 1]$, $x \in (1-\alpha, \infty)$ that

$$(x+\alpha-1)^\alpha = (x-(1-\alpha))^\alpha \leq \frac{\Gamma(x+\alpha)}{\Gamma(x)}. \quad (11.30)$$

Next observe that Lemma 11.2.2 ensures that for all $x \in (0, \infty)$, $\alpha \in [0, 1]$ it holds that

$$\begin{aligned}
 (\max\{x + \alpha - 1, 0\})^\alpha &= (x + \alpha)^\alpha \left(\frac{\max\{x + \alpha - 1, 0\}}{x + \alpha} \right)^\alpha \\
 &= (x + \alpha)^\alpha \left(\max \left\{ 1 - \frac{1}{x + \alpha}, 0 \right\} \right)^\alpha \\
 &\leq (x + \alpha)^\alpha \left(1 - \frac{\alpha}{x + \alpha} \right) = (x + \alpha)^\alpha \left(\frac{x}{x + \alpha} \right) \\
 &= \frac{x}{(x + \alpha)^{1-\alpha}}.
 \end{aligned} \tag{11.31}$$

This and (11.27) establish item (iii). Furthermore, note that induction, item (i) in Lemma 11.2.1, the fact that $\forall \alpha \in [0, \infty) : \alpha - \lfloor \alpha \rfloor \in [0, 1)$, and item (iii) demonstrate that for all $x \in (0, \infty)$, $\alpha \in [0, \infty)$ it holds that

$$\begin{aligned}
 \frac{\Gamma(x + \alpha)}{\Gamma(x)} &= \left[\prod_{i=1}^{\lfloor \alpha \rfloor} (x + \alpha - i) \right] \frac{\Gamma(x + \alpha - \lfloor \alpha \rfloor)}{\Gamma(x)} \leq \left[\prod_{i=1}^{\lfloor \alpha \rfloor} (x + \alpha - i) \right] x^{\alpha - \lfloor \alpha \rfloor} \\
 &\leq (x + \alpha - 1)^{\lfloor \alpha \rfloor} x^{\alpha - \lfloor \alpha \rfloor} \\
 &\leq (x + \max\{\alpha - 1, 0\})^{\lfloor \alpha \rfloor} (x + \max\{\alpha - 1, 0\})^{\alpha - \lfloor \alpha \rfloor} \\
 &= (x + \max\{\alpha - 1, 0\})^\alpha.
 \end{aligned} \tag{11.32}$$

Moreover, observe that the fact that $\forall \alpha \in [0, \infty) : \alpha - \lfloor \alpha \rfloor \in [0, 1)$, item (iii), induction, and item (i) in Lemma 11.2.1 show that for all $x \in (0, \infty)$, $\alpha \in [0, \infty)$ it holds that

$$\begin{aligned}
 \frac{\Gamma(x + \alpha)}{\Gamma(x)} &= \frac{\Gamma(x + \lfloor \alpha \rfloor + \alpha - \lfloor \alpha \rfloor)}{\Gamma(x)} \\
 &\geq (\max\{x + \lfloor \alpha \rfloor + \alpha - \lfloor \alpha \rfloor - 1, 0\})^{\alpha - \lfloor \alpha \rfloor} \left[\frac{\Gamma(x + \lfloor \alpha \rfloor)}{\Gamma(x)} \right] \\
 &= (\max\{x + \alpha - 1, 0\})^{\alpha - \lfloor \alpha \rfloor} \left[\prod_{i=1}^{\lfloor \alpha \rfloor} (x + \lfloor \alpha \rfloor - i) \right] \frac{\Gamma(x)}{\Gamma(x)} \\
 &\geq (\max\{x + \alpha - 1, 0\})^{\alpha - \lfloor \alpha \rfloor} x^{\lfloor \alpha \rfloor} \\
 &= (\max\{x + \alpha - 1, 0\})^{\alpha - \lfloor \alpha \rfloor} (\max\{x, 0\})^{\lfloor \alpha \rfloor} \\
 &\geq (\max\{x + \min\{\alpha - 1, 0\}, 0\})^{\alpha - \lfloor \alpha \rfloor} (\max\{x + \min\{\alpha - 1, 0\}, 0\})^{\lfloor \alpha \rfloor} \\
 &= (\max\{x + \min\{\alpha - 1, 0\}, 0\})^\alpha.
 \end{aligned} \tag{11.33}$$

Combining this with (11.32) proves item (iv). The proof of Proposition 11.2.3 is thus complete. \square

Corollary 11.2.4. Let $\mathbb{B}: (0, \infty)^2 \rightarrow (0, \infty)$ satisfy for all $x, y \in (0, \infty)$ that $\mathbb{B}(x, y) = \int_0^1 t^{x-1} (1-t)^{y-1} dt$ and let $\Gamma: (0, \infty) \rightarrow (0, \infty)$ satisfy for all $x \in (0, \infty)$ that $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$. Then it holds for all $x, y \in (0, \infty)$ with $x + y > 1$ that

$$\frac{\Gamma(x)}{(y + \max\{x - 1, 0\})^x} \leq \mathbb{B}(x, y) \leq \frac{\Gamma(x)}{(y + \min\{x - 1, 0\})^x} \leq \frac{\max\{1, x^x\}}{x(y + \min\{x - 1, 0\})^x}. \quad (11.34)$$

Proof of Corollary 11.2.4. Note that item (iii) in Lemma 11.2.1 implies that for all $x, y \in (0, \infty)$ it holds that

$$\mathbb{B}(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(y+x)}. \quad (11.35)$$

Furthermore, observe that the fact that for all $x, y \in (0, \infty)$ with $x + y > 1$ it holds that $y + \min\{x - 1, 0\} > 0$ and item (iv) in Proposition 11.2.3 ensure that for all $x, y \in (0, \infty)$ with $x + y > 1$ it holds that

$$0 < (y + \min\{x - 1, 0\})^x \leq \frac{\Gamma(y+x)}{\Gamma(y)} \leq (y + \max\{x - 1, 0\})^x. \quad (11.36)$$

Combining this with (11.35) and item (ii) in Proposition 11.2.3 establishes that for all $x, y \in (0, \infty)$ with $x + y > 1$ it holds that

$$\frac{\Gamma(x)}{(y + \max\{x - 1, 0\})^x} \leq \mathbb{B}(x, y) \leq \frac{\Gamma(x)}{(y + \min\{x - 1, 0\})^x} \leq \frac{\max\{1, x^x\}}{x(y + \min\{x - 1, 0\})^x}. \quad (11.37)$$

The proof of Corollary 11.2.4 is thus complete. \square

11.2.2 Product measurability of continuous random fields

Lemma 11.2.5 (Projections in metric spaces). Let (E, d) be a metric space, let $n \in \mathbb{N}$, $e_1, e_2, \dots, e_n \in E$, and let $P: E \rightarrow E$ satisfy for all $x \in E$ that

$$P(x) = e_{\min\{k \in \{1, 2, \dots, n\} : d(x, e_k) = \min\{d(x, e_1), d(x, e_2), \dots, d(x, e_n)\}\}}. \quad (11.38)$$

Then

(i) it holds for all $x \in E$ that

$$d(x, P(x)) = \min_{k \in \{1, 2, \dots, n\}} d(x, e_k) \quad (11.39)$$

and

(ii) it holds for all $A \subseteq E$ that $P^{-1}(A) \in \mathcal{B}(E)$.

Proof of Lemma 11.2.5. Throughout this proof, let $D = (D_1, \dots, D_n): E \rightarrow \mathbb{R}^n$ satisfy for all $x \in E$ that

$$D(x) = (D_1(x), D_2(x), \dots, D_n(x)) = (d(x, e_1), d(x, e_2), \dots, d(x, e_n)). \quad (11.40)$$

Note that (11.38) demonstrates that for all $x \in E$ it holds that

$$\begin{aligned} d(x, P(x)) &= d(x, e_{\min\{k \in \{1, 2, \dots, n\}: d(x, e_k) = \min\{d(x, e_1), d(x, e_2), \dots, d(x, e_n)\}\}}) \\ &= \min_{k \in \{1, 2, \dots, n\}} d(x, e_k). \end{aligned} \quad (11.41)$$

This proves item (i). It thus remains to prove item (ii). For this observe that the fact that $d: E \times E \rightarrow [0, \infty)$ is continuous shows that $D: E \rightarrow \mathbb{R}^n$ is continuous. Therefore, we obtain that $D: E \rightarrow \mathbb{R}^n$ is $\mathcal{B}(E)/\mathcal{B}(\mathbb{R}^n)$ -measurable. Furthermore, note that item (i) implies that for all $k \in \{1, 2, \dots, n\}$, $x \in P^{-1}(\{e_k\})$ it holds that

$$d(x, e_k) = d(x, P(x)) = \min_{l \in \{1, 2, \dots, n\}} d(x, e_l). \quad (11.42)$$

Hence, we obtain that for all $k \in \{1, 2, \dots, n\}$, $x \in P^{-1}(\{e_k\})$ it holds that

$$k \geq \min\{l \in \{1, 2, \dots, n\}: d(x, e_l) = \min\{d(x, e_1), d(x, e_2), \dots, d(x, e_n)\}\}. \quad (11.43)$$

Moreover, observe that (11.38) ensures that for all $k \in \{1, 2, \dots, n\}$, $x \in P^{-1}(\{e_k\})$ it holds that

$$\begin{aligned} \min\left\{l \in \{1, 2, \dots, n\}: d(x, e_l) = \min_{u \in \{1, 2, \dots, n\}} d(x, e_u)\right\} \\ \in \{l \in \{1, 2, \dots, n\}: e_l = e_k\} \subseteq \{k, k+1, \dots, n\}. \end{aligned} \quad (11.44)$$

Therefore, we obtain that for all $k \in \{1, 2, \dots, n\}$, $x \in P^{-1}(\{e_k\})$ with $e_k \notin (\bigcup_{l \in \mathbb{N} \cap [0, k)} \{e_l\})$ it holds that

$$\min\left\{l \in \{1, 2, \dots, n\}: d(x, e_l) = \min_{u \in \{1, 2, \dots, n\}} d(x, e_u)\right\} \geq k. \quad (11.45)$$

Combining this with (11.43) establishes that for all $k \in \{1, 2, \dots, n\}$, $x \in P^{-1}(\{e_k\})$ with $e_k \notin (\bigcup_{l \in \mathbb{N} \cap [0, k)} \{e_l\})$ it holds that

$$\min\left\{l \in \{1, 2, \dots, n\}: d(x, e_l) = \min_{u \in \{1, 2, \dots, n\}} d(x, e_u)\right\} = k. \quad (11.46)$$

Hence, we obtain that for all $k \in \{1, 2, \dots, n\}$ with $e_k \notin (\bigcup_{l \in \mathbb{N} \cap [0, k)} \{e_l\})$ it holds that

$$P^{-1}(\{e_k\}) \subseteq \left\{ x \in E : \min \left\{ l \in \{1, 2, \dots, n\} : d(x, e_l) = \min_{u \in \{1, 2, \dots, n\}} d(x, e_u) \right\} = k \right\}. \quad (11.47)$$

This and (11.38) demonstrate that for all $k \in \{1, 2, \dots, n\}$ with $e_k \notin (\bigcup_{l \in \mathbb{N} \cap [0, k)} \{e_l\})$ it holds that

$$P^{-1}(\{e_k\}) = \left\{ x \in E : \min \left\{ l \in \{1, 2, \dots, n\} : d(x, e_l) = \min_{u \in \{1, 2, \dots, n\}} d(x, e_u) \right\} = k \right\}. \quad (11.48)$$

Combining (11.40) with the fact that $D: E \rightarrow \mathbb{R}^n$ is $\mathcal{B}(E)/\mathcal{B}(\mathbb{R}^n)$ -measurable therefore proves that for all $k \in \{1, 2, \dots, n\}$ with $e_k \notin (\bigcup_{l \in \mathbb{N} \cap [0, k)} \{e_l\})$ it holds that

$$\begin{aligned} P^{-1}(\{e_k\}) &= \left\{ x \in E : \min \left\{ l \in \{1, 2, \dots, n\} : d(x, e_l) = \min_{u \in \{1, 2, \dots, n\}} d(x, e_u) \right\} = k \right\} \\ &= \left\{ x \in E : \min \left\{ l \in \{1, 2, \dots, n\} : D_l(x) = \min_{u \in \{1, 2, \dots, n\}} D_u(x) \right\} = k \right\} \\ &= \left\{ x \in E : \left(\begin{array}{l} \forall l \in \mathbb{N} \cap [0, k) : D_k(x) < D_l(x) \text{ and} \\ \forall l \in \{1, 2, \dots, n\} : D_k(x) \leq D_l(x) \end{array} \right) \right\} \\ &= \left[\bigcap_{l=1}^{k-1} \underbrace{\{x \in E : D_k(x) < D_l(x)\}}_{\in \mathcal{B}(E)} \right] \cap \left[\bigcap_{l=1}^n \underbrace{\{x \in E : D_k(x) \leq D_l(x)\}}_{\in \mathcal{B}(E)} \right] \in \mathcal{B}(E). \end{aligned} \quad (11.49)$$

Hence, we obtain that for all $f \in \{e_1, e_2, \dots, e_n\}$ it holds that

$$P^{-1}(\{f\}) \in \mathcal{B}(E). \quad (11.50)$$

Therefore, we obtain that for all $A \subseteq E$ it holds that

$$P^{-1}(A) = P^{-1}(A \cap \{e_1, e_2, \dots, e_n\}) = \bigcup_{f \in A \cap \{e_1, e_2, \dots, e_n\}} \underbrace{P^{-1}(\{f\})}_{\in \mathcal{B}(E)} \in \mathcal{B}(E). \quad (11.51)$$

This establishes item (ii). The proof of Lemma 11.2.5 is thus complete. \square

Lemma 11.2.6. Let (E, d) be a separable metric space, let (\mathcal{E}, δ) be a metric space, let (Ω, \mathbb{F}) be a measurable space, let $X: E \times \Omega \rightarrow \mathcal{E}$, assume for all $e \in E$ that $\Omega \ni \omega \mapsto X(e, \omega) \in \mathcal{E}$ is $\mathbb{F}/\mathcal{B}(\mathcal{E})$ -measurable, and assume for all $\omega \in \Omega$ that $E \ni e \mapsto X(e, \omega) \in \mathcal{E}$ is continuous. Then $X: E \times \Omega \rightarrow \mathcal{E}$ is $(\mathcal{B}(E) \otimes \mathbb{F})/\mathcal{B}(\mathcal{E})$ -measurable.

Proof of Lemma 11.2.6. Throughout this proof, let $e = (e_m)_{m \in \mathbb{N}}: \mathbb{N} \rightarrow E$ satisfy

$$\overline{\{e_m: m \in \mathbb{N}\}} = E, \quad (11.52)$$

let $P_n: E \rightarrow E$, $n \in \mathbb{N}$, satisfy for all $n \in \mathbb{N}$, $x \in E$ that

$$P_n(x) = e_{\min\{k \in \{1, 2, \dots, n\}: d(x, e_k) = \min\{d(x, e_1), d(x, e_2), \dots, d(x, e_n)\}\}}, \quad (11.53)$$

and let $\mathcal{X}_n: E \times \Omega \rightarrow \mathcal{E}$, $n \in \mathbb{N}$, satisfy for all $n \in \mathbb{N}$, $x \in E$, $\omega \in \Omega$ that

$$\mathcal{X}_n(x, \omega) = X(P_n(x), \omega). \quad (11.54)$$

Note that (11.54) shows that for all $n \in \mathbb{N}$, $B \in \mathcal{B}(\mathcal{E})$ it holds that

$$\begin{aligned} (\mathcal{X}_n)^{-1}(B) &= \{(x, \omega) \in E \times \Omega: \mathcal{X}_n(x, \omega) \in B\} \\ &= \bigcup_{y \in \text{Im}(P_n)} \left([(\mathcal{X}_n)^{-1}(B)] \cap [(P_n)^{-1}(\{y\}) \times \Omega] \right) \\ &= \bigcup_{y \in \text{Im}(P_n)} \left\{ (x, \omega) \in E \times \Omega: [\mathcal{X}_n(x, \omega) \in B \text{ and } x \in (P_n)^{-1}(\{y\})] \right\} \\ &= \bigcup_{y \in \text{Im}(P_n)} \left\{ (x, \omega) \in E \times \Omega: [X(P_n(x), \omega) \in B \text{ and } x \in (P_n)^{-1}(\{y\})] \right\}. \end{aligned} \quad (11.55)$$

Item (ii) in Lemma 11.2.5 hence implies that for all $n \in \mathbb{N}$, $B \in \mathcal{B}(\mathcal{E})$ it holds that

$$\begin{aligned} (\mathcal{X}_n)^{-1}(B) &= \bigcup_{y \in \text{Im}(P_n)} \left\{ (x, \omega) \in E \times \Omega: [X(y, \omega) \in B \text{ and } x \in (P_n)^{-1}(\{y\})] \right\} \\ &= \bigcup_{y \in \text{Im}(P_n)} \left(\{(x, \omega) \in E \times \Omega: X(y, \omega) \in B\} \cap [(P_n)^{-1}(\{y\}) \times \Omega] \right) \\ &= \bigcup_{y \in \text{Im}(P_n)} \left(\underbrace{[E \times ((X(y, \cdot))^{-1}(B))]}_{\in (\mathcal{B}(E) \otimes \mathbb{F})} \cap \underbrace{[(P_n)^{-1}(\{y\}) \times \Omega]}_{\in (\mathcal{B}(E) \otimes \mathbb{F})} \right) \in (\mathcal{B}(E) \otimes \mathbb{F}). \end{aligned} \quad (11.56)$$

This ensures that for all $n \in \mathbb{N}$ it holds that \mathcal{X}_n is $(\mathcal{B}(E) \otimes \mathbb{F})/\mathcal{B}(\mathcal{E})$ -measurable. Furthermore, observe that item (i) in Lemma 11.2.5 and the assumption that for all $\omega \in \Omega$ it holds that $E \ni x \mapsto X(x, \omega) \in \mathcal{E}$ is continuous demonstrate that for all $x \in E$, $\omega \in \Omega$ it holds that

$$\lim_{n \rightarrow \infty} \mathcal{X}_n(x, \omega) = \lim_{n \rightarrow \infty} X(P_n(x), \omega) = X(x, \omega). \quad (11.57)$$

Combining this with the fact that for all $n \in \mathbb{N}$ it holds that $X_n: E \times \Omega \rightarrow \mathcal{E}$ is $(\mathcal{B}(E) \otimes \mathbb{F})/\mathcal{B}(\mathcal{E})$ -measurable proves that $X: E \times \Omega \rightarrow \mathcal{E}$ is $(\mathcal{B}(E) \otimes \mathbb{F})/\mathcal{B}(\mathcal{E})$ -measurable. The proof of Lemma 11.2.6 is thus complete. \square

11.2.3 Strong convergences rates for the optimization error

Proposition 11.2.7. Let $\mathbf{d}, K \in \mathbb{N}$, $L, \alpha \in \mathbb{R}$, $\beta \in (\alpha, \infty)$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\mathcal{R}: [\alpha, \beta]^{\mathbf{d}} \times \Omega \rightarrow \mathbb{R}$ be a random field, assume for all $\theta, \vartheta \in [\alpha, \beta]^{\mathbf{d}}$, $\omega \in \Omega$ that $|\mathcal{R}(\theta, \omega) - \mathcal{R}(\vartheta, \omega)| \leq L\|\theta - \vartheta\|_{\infty}$, let $\Theta_k: \Omega \rightarrow [\alpha, \beta]^{\mathbf{d}}$, $k \in \{1, 2, \dots, K\}$, be i.i.d. random variables, and assume that Θ_1 is continuously uniformly distributed on $[\alpha, \beta]^{\mathbf{d}}$ (cf. Definition 3.3.4). Then

- (i) it holds that \mathcal{R} is $(\mathcal{B}([\alpha, \beta]^{\mathbf{d}}) \otimes \mathcal{F})/\mathcal{B}(\mathbb{R})$ -measurable and
- (ii) it holds for all $\theta \in [\alpha, \beta]^{\mathbf{d}}$, $p \in (0, \infty)$ that

$$\begin{aligned} (\mathbb{E}[\min_{k \in \{1, 2, \dots, K\}} |\mathcal{R}(\Theta_k) - \mathcal{R}(\theta)|^p])^{1/p} &\leq \frac{L(\beta - \alpha) \max\{1, (p/\mathbf{d})^{1/\mathbf{d}}\}}{K^{1/\mathbf{d}}} \\ &\leq \frac{L(\beta - \alpha) \max\{1, p\}}{K^{1/\mathbf{d}}}. \end{aligned} \quad (11.58)$$

Proof of Proposition 11.2.7. Throughout this proof, assume without loss of generality that $L > 0$, let $\delta: ([\alpha, \beta]^{\mathbf{d}}) \times ([\alpha, \beta]^{\mathbf{d}}) \rightarrow [0, \infty)$ satisfy for all $\theta, \vartheta \in [\alpha, \beta]^{\mathbf{d}}$ that

$$\delta(\theta, \vartheta) = \|\theta - \vartheta\|_{\infty}, \quad (11.59)$$

let $\mathbb{B}: (0, \infty)^2 \rightarrow (0, \infty)$ satisfy for all $x, y \in (0, \infty)$ that

$$\mathbb{B}(x, y) = \int_0^1 t^{x-1} (1-t)^{y-1} dt, \quad (11.60)$$

and let $\Theta_{1,1}, \Theta_{1,2}, \dots, \Theta_{1,\mathbf{d}}: \Omega \rightarrow [\alpha, \beta]$ satisfy $\Theta_1 = (\Theta_{1,1}, \Theta_{1,2}, \dots, \Theta_{1,\mathbf{d}})$. First, note that the assumption that for all $\theta, \vartheta \in [\alpha, \beta]^{\mathbf{d}}$, $\omega \in \Omega$ it holds that

$$|\mathcal{R}(\theta, \omega) - \mathcal{R}(\vartheta, \omega)| \leq L\|\theta - \vartheta\|_{\infty} \quad (11.61)$$

establishes that for all $\omega \in \Omega$ it holds that $[\alpha, \beta]^{\mathbf{d}} \ni \theta \mapsto \mathcal{R}(\theta, \omega) \in \mathbb{R}$ is continuous. Combining this with the fact that $([\alpha, \beta]^{\mathbf{d}}, \delta)$ is a separable metric space, the fact that for all $\theta \in [\alpha, \beta]^{\mathbf{d}}$ it holds that $\Omega \ni \omega \mapsto \mathcal{R}(\theta, \omega) \in \mathbb{R}$ is $\mathcal{F}/\mathcal{B}(\mathbb{R})$ -measurable, and Lemma 11.2.6 proves item (i). Observe that the fact that for all $\theta \in [\alpha, \beta]$, $\varepsilon \in [0, \infty)$ it holds that

$$\begin{aligned} \min\{\theta + \varepsilon, \beta\} - \max\{\theta - \varepsilon, \alpha\} &= \min\{\theta + \varepsilon, \beta\} + \min\{\varepsilon - \theta, -\alpha\} \\ &= \min\{\theta + \varepsilon + \min\{\varepsilon - \theta, -\alpha\}, \beta + \min\{\varepsilon - \theta, -\alpha\}\} \\ &= \min\{\min\{2\varepsilon, \theta - \alpha + \varepsilon\}, \min\{\beta - \theta + \varepsilon, \beta - \alpha\}\} \\ &\geq \min\{\min\{2\varepsilon, \alpha - \alpha + \varepsilon\}, \min\{\beta - \beta + \varepsilon, \beta - \alpha\}\} \\ &= \min\{2\varepsilon, \varepsilon, \varepsilon, \beta - \alpha\} = \min\{\varepsilon, \beta - \alpha\} \end{aligned} \quad (11.62)$$

and the assumption that Θ_1 is continuously uniformly distributed on $[\alpha, \beta]^{\mathbf{d}}$ show that for all $\theta = (\theta_1, \theta_2, \dots, \theta_{\mathbf{d}}) \in [\alpha, \beta]^{\mathbf{d}}, \varepsilon \in [0, \infty)$ it holds that

$$\begin{aligned}
 \mathbb{P}(\|\Theta_1 - \theta\|_{\infty} \leq \varepsilon) &= \mathbb{P}\left(\max_{i \in \{1, 2, \dots, \mathbf{d}\}} |\Theta_{1,i} - \theta_i| \leq \varepsilon\right) \\
 &= \mathbb{P}\left(\forall i \in \{1, 2, \dots, \mathbf{d}\}: -\varepsilon \leq \Theta_{1,i} - \theta_i \leq \varepsilon\right) \\
 &= \mathbb{P}\left(\forall i \in \{1, 2, \dots, \mathbf{d}\}: \theta_i - \varepsilon \leq \Theta_{1,i} \leq \theta_i + \varepsilon\right) \\
 &= \mathbb{P}\left(\forall i \in \{1, 2, \dots, \mathbf{d}\}: \max\{\theta_i - \varepsilon, \alpha\} \leq \Theta_{1,i} \leq \min\{\theta_i + \varepsilon, \beta\}\right) \\
 &= \mathbb{P}\left(\Theta_1 \in \left[\bigtimes_{i=1}^{\mathbf{d}} [\max\{\theta_i - \varepsilon, \alpha\}, \min\{\theta_i + \varepsilon, \beta\}]\right]\right) \\
 &= \frac{1}{(\beta - \alpha)^{\mathbf{d}}} \prod_{i=1}^{\mathbf{d}} (\min\{\theta_i + \varepsilon, \beta\} - \max\{\theta_i - \varepsilon, \alpha\}) \\
 &\geq \frac{1}{(\beta - \alpha)^{\mathbf{d}}} [\min\{\varepsilon, \beta - \alpha\}]^{\mathbf{d}} = \min\left\{1, \frac{\varepsilon^{\mathbf{d}}}{(\beta - \alpha)^{\mathbf{d}}}\right\}.
 \end{aligned} \tag{11.63}$$

Therefore, we obtain for all $\theta \in [\alpha, \beta]^{\mathbf{d}}, p \in (0, \infty), \varepsilon \in [0, \infty)$ that

$$\begin{aligned}
 \mathbb{P}(\|\Theta_1 - \theta\|_{\infty} > \varepsilon^{1/p}) &= 1 - \mathbb{P}(\|\Theta_1 - \theta\|_{\infty} \leq \varepsilon^{1/p}) \\
 &\leq 1 - \min\left\{1, \frac{\varepsilon^{d/p}}{(\beta - \alpha)^d}\right\} = \max\left\{0, 1 - \frac{\varepsilon^{d/p}}{(\beta - \alpha)^d}\right\}.
 \end{aligned} \tag{11.64}$$

This, item (i), the assumption that for all $\theta, \vartheta \in [\alpha, \beta]^{\mathbf{d}}, \omega \in \Omega$ it holds that

$$|\mathcal{R}(\theta, \omega) - \mathcal{R}(\vartheta, \omega)| \leq L\|\theta - \vartheta\|_{\infty}, \tag{11.65}$$

the assumption that $\Theta_k, k \in \{1, 2, \dots, K\}$, are i.i.d. random variables, and Lemma 11.1.2 (applied with $(E, \delta) \curvearrowright ([\alpha, \beta]^{\mathbf{d}}, \delta)$, $(X_k)_{k \in \{1, 2, \dots, K\}} \curvearrowright (\Theta_k)_{k \in \{1, 2, \dots, K\}}$ in the notation of Lemma 11.1.2) imply that for all $\theta \in [\alpha, \beta]^{\mathbf{d}}, p \in (0, \infty)$ it holds that

$$\begin{aligned}
 \mathbb{E}[\min_{k \in \{1, 2, \dots, K\}} |\mathcal{R}(\Theta_k) - \mathcal{R}(\theta)|^p] &\leq L^p \int_0^\infty [\mathbb{P}(\|\Theta_1 - \theta\|_{\infty} > \varepsilon^{1/p})]^K d\varepsilon \\
 &\leq L^p \int_0^\infty \left[\max\left\{0, 1 - \frac{\varepsilon^{d/p}}{(\beta - \alpha)^d}\right\}\right]^K d\varepsilon = L^p \int_0^{(\beta - \alpha)^p} \left(1 - \frac{\varepsilon^{d/p}}{(\beta - \alpha)^d}\right)^K d\varepsilon \\
 &= \frac{p}{d} L^p (\beta - \alpha)^p \int_0^1 t^{p/d-1} (1-t)^K dt = \frac{p}{d} L^p (\beta - \alpha)^p \int_0^1 t^{p/d-1} (1-t)^{K+1-1} dt \\
 &= \frac{p}{d} L^p (\beta - \alpha)^p \mathbb{B}(p/d, K+1).
 \end{aligned} \tag{11.66}$$

Corollary 11.2.4 (applied with $x \curvearrowright p/d, y \curvearrowright K+1$ for $p \in (0, \infty)$ in the notation of (11.34) in Corollary 11.2.4) hence ensures that for all $\theta \in [\alpha, \beta]^{\mathbf{d}}, p \in (0, \infty)$ it holds that

$$\begin{aligned}
 \mathbb{E}[\min_{k \in \{1, 2, \dots, K\}} |\mathcal{R}(\Theta_k) - \mathcal{R}(\theta)|^p] &\leq \frac{\frac{p}{d} L^p (\beta - \alpha)^p \max\{1, (p/d)^{p/d}\}}{\frac{p}{d} (K+1 + \min\{p/d - 1, 0\})^{p/d}} \\
 &\leq \frac{L^p (\beta - \alpha)^p \max\{1, (p/d)^{p/d}\}}{K^{p/d}}.
 \end{aligned} \tag{11.67}$$

This demonstrates for all $\theta \in [\alpha, \beta]^d$, $p \in (0, \infty)$ that

$$\begin{aligned} (\mathbb{E}[\min_{k \in \{1, 2, \dots, K\}} |\mathcal{R}(\Theta_k) - \mathcal{R}(\theta)|^p])^{1/p} &\leq \frac{L(\beta - \alpha) \max\{1, (p/d)^{1/d}\}}{K^{1/d}} \\ &\leq \frac{L(\beta - \alpha) \max\{1, p\}}{K^{1/d}}. \end{aligned} \quad (11.68)$$

This establishes item (ii). The proof of Proposition 11.2.7 is thus complete. \square

11.3 Strong convergences rates for the optimization error involving ANNs

11.3.1 Local Lipschitz continuity estimates for the parametrization functions of ANNs

Lemma 11.3.1. *Let $a, x, y \in \mathbb{R}$. Then*

$$|\max\{x, a\} - \max\{y, a\}| \leq \max\{x, y\} - \min\{x, y\} = |x - y|. \quad (11.69)$$

Proof of Lemma 11.3.1. Note that the fact that

$$\begin{aligned} |\max\{x, a\} - \max\{y, a\}| &= |\max\{\max\{x, y\}, a\} - \max\{\min\{x, y\}, a\}| \\ &= \max\{\max\{x, y\}, a\} - \max\{\min\{x, y\}, a\} \\ &= \max\{\max\{x, y\} - \max\{\min\{x, y\}, a\}, a - \max\{\min\{x, y\}, a\}\} \\ &\leq \max\{\max\{x, y\} - \max\{\min\{x, y\}, a\}, a - a\} \\ &= \max\{\max\{x, y\} - \max\{\min\{x, y\}, a\}, 0\} \leq \max\{\max\{x, y\} - \min\{x, y\}, 0\} \\ &= \max\{x, y\} - \min\{x, y\} = |\max\{x, y\} - \min\{x, y\}| = |x - y|. \end{aligned} \quad (11.70)$$

proves (11.69). The proof of Lemma 11.3.1 is thus complete. \square

Corollary 11.3.2. *Let $a, x, y \in \mathbb{R}$. Then*

$$|\min\{x, a\} - \min\{y, a\}| \leq \max\{x, y\} - \min\{x, y\} = |x - y|. \quad (11.71)$$

Proof of Corollary 11.3.2. Observe that Lemma 11.3.1 shows that

$$\begin{aligned} |\min\{x, a\} - \min\{y, a\}| &= |-(\min\{x, a\} - \min\{y, a\})| \\ &= |\max\{-x, -a\} - \max\{-y, -a\}| \\ &\leq |(-x) - (-y)| = |x - y|. \end{aligned} \quad (11.72)$$

The proof of Corollary 11.3.2 is thus complete. \square

Lemma 11.3.3. Let $d \in \mathbb{N}$. Then it holds for all $x, y \in \mathbb{R}^d$ that

$$\|\mathfrak{R}_d(x) - \mathfrak{R}_d(y)\|_\infty \leq \|x - y\|_\infty \quad (11.73)$$

(cf. Definitions 1.2.5 and 3.3.4).

Proof of Lemma 11.3.3. Observe that Lemma 11.3.1 demonstrates (11.73). The proof of Lemma 11.3.3 is thus complete. \square

Lemma 11.3.4. Let $d \in \mathbb{N}$, $u \in [-\infty, \infty)$, $v \in (u, \infty]$. Then it holds for all $x, y \in \mathbb{R}^d$ that

$$\|\mathfrak{C}_{u,v,d}(x) - \mathfrak{C}_{u,v,d}(y)\|_\infty \leq \|x - y\|_\infty \quad (11.74)$$

(cf. Definitions 1.2.10 and 3.3.4).

Proof of Lemma 11.3.4. Note that Lemma 11.3.1, Corollary 11.3.2, and the fact that for all $x \in \mathbb{R}$ it holds that $\max\{-\infty, x\} = x = \min\{x, \infty\}$ imply that for all $x, y \in \mathbb{R}$ it holds that

$$\begin{aligned} |\mathfrak{c}_{u,v}(x) - \mathfrak{c}_{u,v}(y)| &= |\max\{u, \min\{x, v\}\} - \max\{u, \min\{y, v\}\}| \\ &\leq |\min\{x, v\} - \min\{y, v\}| \leq |x - y| \end{aligned} \quad (11.75)$$

(cf. Definition 1.2.9). Therefore, we obtain that for all $x = (x_1, x_2, \dots, x_d), y = (y_1, y_2, \dots, y_d) \in \mathbb{R}^d$ it holds that

$$\begin{aligned} \|\mathfrak{C}_{u,v,d}(x) - \mathfrak{C}_{u,v,d}(y)\|_\infty &= \max_{i \in \{1, 2, \dots, d\}} |\mathfrak{c}_{u,v}(x_i) - \mathfrak{c}_{u,v}(y_i)| \\ &\leq \max_{i \in \{1, 2, \dots, d\}} |x_i - y_i| = \|x - y\|_\infty \end{aligned} \quad (11.76)$$

(cf. Definitions 1.2.10 and 3.3.4). The proof of Lemma 11.3.4 is thus complete. \square

Lemma 11.3.5 (Row sum norm, operator norm induced by the maximum norm). Let $a, b \in \mathbb{N}$, $M = (M_{i,j})_{(i,j) \in \{1, 2, \dots, a\} \times \{1, 2, \dots, b\}} \in \mathbb{R}^{a \times b}$. Then

$$\sup_{v \in \mathbb{R}^b \setminus \{0\}} \left[\frac{\|Mv\|_\infty}{\|v\|_\infty} \right] = \max_{i \in \{1, 2, \dots, a\}} \left[\sum_{j=1}^b |M_{i,j}| \right] \leq b \left[\max_{i \in \{1, 2, \dots, a\}} \max_{j \in \{1, 2, \dots, b\}} |M_{i,j}| \right] \quad (11.77)$$

(cf. Definition 3.3.4).

Proof of Lemma 11.3.5. Observe that

$$\begin{aligned}
 \sup_{v \in \mathbb{R}^b} \left[\frac{\|Mv\|_\infty}{\|v\|_\infty} \right] &= \sup_{v \in \mathbb{R}^b, \|v\|_\infty \leq 1} \|Mv\|_\infty \\
 &= \sup_{v=(v_1, v_2, \dots, v_b) \in [-1, 1]^b} \|Mv\|_\infty \\
 &= \sup_{v=(v_1, v_2, \dots, v_b) \in [-1, 1]^b} \left(\max_{i \in \{1, 2, \dots, a\}} \left| \sum_{j=1}^b M_{i,j} v_j \right| \right) \\
 &= \max_{i \in \{1, 2, \dots, a\}} \left(\sup_{v=(v_1, v_2, \dots, v_b) \in [-1, 1]^b} \left| \sum_{j=1}^b M_{i,j} v_j \right| \right) \\
 &= \max_{i \in \{1, 2, \dots, a\}} \left(\sum_{j=1}^b |M_{i,j}| \right)
 \end{aligned} \tag{11.78}$$

(cf. Definition 3.3.4). The proof of Lemma 11.3.5 is thus complete. \square

Theorem 11.3.6. Let $a \in \mathbb{R}$, $b \in [a, \infty)$, $d, L \in \mathbb{N}$, $l = (l_0, l_1, \dots, l_L) \in \mathbb{N}^{L+1}$ satisfy

$$d \geq \sum_{k=1}^L l_k (l_{k-1} + 1). \tag{11.79}$$

Then it holds for all $\theta, \vartheta \in \mathbb{R}^d$ that

$$\begin{aligned}
 &\sup_{x \in [a, b]^{l_0}} \|\mathcal{N}_{\infty, \infty}^{\theta, l}(x) - \mathcal{N}_{\infty, \infty}^{\vartheta, l}(x)\|_\infty \\
 &\leq \max\{1, |a|, |b|\} \|\theta - \vartheta\|_\infty \left[\prod_{m=0}^{L-1} (l_m + 1) \right] \left[\sum_{n=0}^{L-1} (\max\{1, \|\theta\|_\infty^n\} \|\vartheta\|_\infty^{L-1-n}) \right] \\
 &\leq L \max\{1, |a|, |b|\} (\max\{1, \|\theta\|_\infty, \|\vartheta\|_\infty\})^{L-1} \left[\prod_{m=0}^{L-1} (l_m + 1) \right] \|\theta - \vartheta\|_\infty \\
 &\leq L \max\{1, |a|, |b|\} (\|l\|_\infty + 1)^L (\max\{1, \|\theta\|_\infty, \|\vartheta\|_\infty\})^{L-1} \|\theta - \vartheta\|_\infty
 \end{aligned} \tag{11.80}$$

(cf. Definitions 3.3.4 and 4.4.1).

Proof of Theorem 11.3.6. Throughout this proof, let $\theta_j = (\theta_{j,1}, \theta_{j,2}, \dots, \theta_{j,d}) \in \mathbb{R}^d$, $j \in \{1, 2\}$, let $\mathfrak{d} \in \mathbb{N}$ satisfy

$$\mathfrak{d} = \sum_{k=1}^L l_k (l_{k-1} + 1), \tag{11.81}$$

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

$$\mathcal{T}\left(\left((W_{j,1}, B_{j,1}), (W_{j,2}, B_{j,2}), \dots, (W_{j,L}, B_{j,L})\right)\right) = (\theta_{j,1}, \theta_{j,2}, \dots, \theta_{j,0}), \quad (11.82)$$

let $\phi_{j,k} \in \mathbf{N}$, $k \in \{1, 2, \dots, L\}$, $j \in \{1, 2\}$, satisfy for all $j \in \{1, 2\}$, $k \in \{1, 2, \dots, L\}$ that

$$\phi_{j,k} = \left((W_{j,1}, B_{j,1}), (W_{j,2}, B_{j,2}), \dots, (W_{j,k}, B_{j,k})\right) \in \left[\bigtimes_{i=1}^k (\mathbb{R}^{l_i \times l_{i-1}} \times \mathbb{R}^{l_i})\right], \quad (11.83)$$

let $D = [a, b]^{l_0}$, let $\mathfrak{m}_{j,k} \in [0, \infty)$, $j \in \{1, 2\}$, $k \in \{0, 1, \dots, L\}$, satisfy for all $j \in \{1, 2\}$, $k \in \{0, 1, \dots, L\}$ that

$$\mathfrak{m}_{j,k} = \begin{cases} \max\{1, |a|, |b|\} & : k = 0 \\ \max\{1, \sup_{x \in D} \|(\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\phi_{j,k}))(x)\|_{\infty}\} & : k > 0, \end{cases} \quad (11.84)$$

and let $\mathfrak{e}_k \in [0, \infty)$, $k \in \{0, 1, \dots, L\}$, satisfy for all $k \in \{0, 1, \dots, L\}$ that

$$\mathfrak{e}_k = \begin{cases} 0 & : k = 0 \\ \sup_{x \in D} \|(\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\phi_{1,k}))(x) - (\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\phi_{2,k}))(x)\|_{\infty} & : k > 0 \end{cases} \quad (11.85)$$

(cf. Definitions 1.2.4, 1.3.1, 1.3.4, 1.3.6, and 3.3.4). Note that Lemma 11.3.5 ensures that

$$\begin{aligned} \mathfrak{e}_1 &= \sup_{x \in D} \|(\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\phi_{1,1}))(x) - (\mathcal{R}_{\mathfrak{r}}^{\mathbf{N}}(\phi_{2,1}))(x)\|_{\infty} \\ &= \sup_{x \in D} \|(W_{1,1}x + B_{1,1}) - (W_{2,1}x + B_{2,1})\|_{\infty} \\ &\leq \left[\sup_{x \in D} \|(W_{1,1} - W_{2,1})x\|_{\infty} \right] + \|B_{1,1} - B_{2,1}\|_{\infty} \\ &\leq \left[\sup_{v \in \mathbb{R}^{l_0} \setminus \{0\}} \left(\frac{\|(W_{1,1} - W_{2,1})v\|_{\infty}}{\|v\|_{\infty}} \right) \right] \left[\sup_{x \in D} \|x\|_{\infty} \right] + \|B_{1,1} - B_{2,1}\|_{\infty} \\ &\leq l_0 \|\theta_1 - \theta_2\|_{\infty} \max\{|a|, |b|\} + \|B_{1,1} - B_{2,1}\|_{\infty} \\ &\leq l_0 \|\theta_1 - \theta_2\|_{\infty} \max\{|a|, |b|\} + \|\theta_1 - \theta_2\|_{\infty} \\ &= \|\theta_1 - \theta_2\|_{\infty} (l_0 \max\{|a|, |b|\} + 1) \leq \mathfrak{m}_{1,0} \|\theta_1 - \theta_2\|_{\infty} (l_0 + 1). \end{aligned} \quad (11.86)$$

Furthermore, observe that the triangle inequality establishes that for all $k \in \{1, 2, \dots, L\} \cap$

$(1, \infty)$ it holds that

$$\begin{aligned}
 \epsilon_k &= \sup_{x \in D} \|(\mathcal{R}_{\tau}^{\mathbf{N}}(\phi_{1,k}))(x) - (\mathcal{R}_{\tau}^{\mathbf{N}}(\phi_{2,k}))(x)\|_{\infty} \\
 &= \sup_{x \in D} \left\| \left[W_{1,k} \left(\mathfrak{R}_{l_{k-1}}((\mathcal{R}_{\tau}^{\mathbf{N}}(\phi_{1,k-1}))(x)) \right) + B_{1,k} \right] \right. \\
 &\quad \left. - \left[W_{2,k} \left(\mathfrak{R}_{l_{k-1}}((\mathcal{R}_{\tau}^{\mathbf{N}}(\phi_{2,k-1}))(x)) \right) + B_{2,k} \right] \right\|_{\infty} \\
 &\leq \left[\sup_{x \in D} \left\| W_{1,k} \left(\mathfrak{R}_{l_{k-1}}((\mathcal{R}_{\tau}^{\mathbf{N}}(\phi_{1,k-1}))(x)) \right) - W_{2,k} \left(\mathfrak{R}_{l_{k-1}}((\mathcal{R}_{\tau}^{\mathbf{N}}(\phi_{2,k-1}))(x)) \right) \right\|_{\infty} \right. \\
 &\quad \left. + \|\theta_1 - \theta_2\|_{\infty} \right].
 \end{aligned} \tag{11.87}$$

The triangle inequality hence proves that for all $j \in \{1, 2\}$, $k \in \{1, 2, \dots, L\} \cap (1, \infty)$ it holds that

$$\begin{aligned}
 \epsilon_k &\leq \left[\sup_{x \in D} \left\| (W_{1,k} - W_{2,k}) \left(\mathfrak{R}_{l_{k-1}}((\mathcal{R}_{\tau}^{\mathbf{N}}(\phi_{j,k-1}))(x)) \right) \right\|_{\infty} \right] \\
 &\quad + \left[\sup_{x \in D} \left\| W_{3-j,k} \left(\mathfrak{R}_{l_{k-1}}((\mathcal{R}_{\tau}^{\mathbf{N}}(\phi_{1,k-1}))(x)) - \mathfrak{R}_{l_{k-1}}((\mathcal{R}_{\tau}^{\mathbf{N}}(\phi_{2,k-1}))(x)) \right) \right\|_{\infty} \right] \\
 &\quad + \|\theta_1 - \theta_2\|_{\infty} \\
 &\leq \left[\sup_{v \in \mathbb{R}^{l_{k-1}} \setminus \{0\}} \left(\frac{\|(W_{1,k} - W_{2,k})v\|_{\infty}}{\|v\|_{\infty}} \right) \right] \left[\sup_{x \in D} \left\| \mathfrak{R}_{l_{k-1}}((\mathcal{R}_{\tau}^{\mathbf{N}}(\phi_{j,k-1}))(x)) \right\|_{\infty} \right] \\
 &\quad + \left[\sup_{v \in \mathbb{R}^{l_{k-1}} \setminus \{0\}} \left(\frac{\|W_{3-j,k}v\|_{\infty}}{\|v\|_{\infty}} \right) \right] \left[\sup_{x \in D} \left\| \mathfrak{R}_{l_{k-1}}((\mathcal{R}_{\tau}^{\mathbf{N}}(\phi_{1,k-1}))(x)) \right. \right. \\
 &\quad \left. \left. - \mathfrak{R}_{l_{k-1}}((\mathcal{R}_{\tau}^{\mathbf{N}}(\phi_{2,k-1}))(x)) \right\|_{\infty} \right] + \|\theta_1 - \theta_2\|_{\infty}.
 \end{aligned} \tag{11.88}$$

Lemma 11.3.5 and Lemma 11.3.3 therefore show that for all $j \in \{1, 2\}$, $k \in \{1, 2, \dots, L\} \cap (1, \infty)$ it holds that

$$\begin{aligned}
 \epsilon_k &\leq l_{k-1} \|\theta_1 - \theta_2\|_{\infty} \left[\sup_{x \in D} \left\| \mathfrak{R}_{l_{k-1}}((\mathcal{R}_{\tau}^{\mathbf{N}}(\phi_{j,k-1}))(x)) \right\|_{\infty} \right] + \|\theta_1 - \theta_2\|_{\infty} \\
 &\quad + l_{k-1} \|\theta_{3-j}\|_{\infty} \left[\sup_{x \in D} \left\| \mathfrak{R}_{l_{k-1}}((\mathcal{R}_{\tau}^{\mathbf{N}}(\phi_{1,k-1}))(x)) - \mathfrak{R}_{l_{k-1}}((\mathcal{R}_{\tau}^{\mathbf{N}}(\phi_{2,k-1}))(x)) \right\|_{\infty} \right] \\
 &\leq l_{k-1} \|\theta_1 - \theta_2\|_{\infty} \left[\sup_{x \in D} \left\| (\mathcal{R}_{\tau}^{\mathbf{N}}(\phi_{j,k-1}))(x) \right\|_{\infty} \right] + \|\theta_1 - \theta_2\|_{\infty} \\
 &\quad + l_{k-1} \|\theta_{3-j}\|_{\infty} \left[\sup_{x \in D} \left\| (\mathcal{R}_{\tau}^{\mathbf{N}}(\phi_{1,k-1}))(x) - (\mathcal{R}_{\tau}^{\mathbf{N}}(\phi_{2,k-1}))(x) \right\|_{\infty} \right] \\
 &\leq \|\theta_1 - \theta_2\|_{\infty} (l_{k-1} \mathfrak{m}_{j,k-1} + 1) + l_{k-1} \|\theta_{3-j}\|_{\infty} \epsilon_{k-1}.
 \end{aligned} \tag{11.89}$$

Hence, we obtain that for all $j \in \{1, 2\}$, $k \in \{1, 2, \dots, L\} \cap (1, \infty)$ it holds that

$$\epsilon_k \leq m_{j,k-1} \|\theta_1 - \theta_2\|_\infty (l_{k-1} + 1) + l_{k-1} \|\theta_{3-j}\|_\infty \epsilon_{k-1}. \quad (11.90)$$

Combining this with (11.86), the fact that $\epsilon_0 = 0$, and the fact that $m_{1,0} = m_{2,0}$ demonstrates that for all $j \in \{1, 2\}$, $k \in \{1, 2, \dots, L\}$ it holds that

$$\epsilon_k \leq m_{j,k-1} (l_{k-1} + 1) \|\theta_1 - \theta_2\|_\infty + l_{k-1} \|\theta_{3-j}\|_\infty \epsilon_{k-1}. \quad (11.91)$$

This implies that for all $j = (j_n)_{n \in \{0,1,\dots,L\}} : \{0, 1, \dots, L\} \rightarrow \{1, 2\}$ and all $k \in \{1, 2, \dots, L\}$ it holds that

$$\epsilon_k \leq m_{j_{k-1},k-1} (l_{k-1} + 1) \|\theta_1 - \theta_2\|_\infty + l_{k-1} \|\theta_{3-j_{k-1}}\|_\infty \epsilon_{k-1}. \quad (11.92)$$

Therefore, we obtain that for all $j = (j_n)_{n \in \{0,1,\dots,L\}} : \{0, 1, \dots, L\} \rightarrow \{1, 2\}$ and all $k \in \{1, 2, \dots, L\}$ it holds that

$$\begin{aligned} \epsilon_k &\leq \sum_{n=0}^{k-1} \left(\left[\prod_{m=n+1}^{k-1} (l_m \|\theta_{3-j_m}\|_\infty) \right] m_{j_n,n} (l_n + 1) \|\theta_1 - \theta_2\|_\infty \right) \\ &= \|\theta_1 - \theta_2\|_\infty \left[\sum_{n=0}^{k-1} \left(\left[\prod_{m=n+1}^{k-1} (l_m \|\theta_{3-j_m}\|_\infty) \right] m_{j_n,n} (l_n + 1) \right) \right]. \end{aligned} \quad (11.93)$$

Moreover, note that Lemma 11.3.5 ensures that for all $j \in \{1, 2\}$, $k \in \{1, 2, \dots, L\} \cap (1, \infty)$, $x \in D$ it holds that

$$\begin{aligned} &\|(\mathcal{R}_r^N(\phi_{j,k}))(x)\|_\infty \\ &= \left\| W_{j,k} \left(\mathfrak{R}_{l_{k-1}}((\mathcal{R}_r^N(\phi_{j,k-1}))(x)) \right) + B_{j,k} \right\|_\infty \\ &\leq \left[\sup_{v \in \mathbb{R}^{l_{k-1}} \setminus \{0\}} \frac{\|W_{j,k}v\|_\infty}{\|v\|_\infty} \right] \|\mathfrak{R}_{l_{k-1}}((\mathcal{R}_r^N(\phi_{j,k-1}))(x))\|_\infty + \|B_{j,k}\|_\infty \\ &\leq l_{k-1} \|\theta_j\|_\infty \|\mathfrak{R}_{l_{k-1}}((\mathcal{R}_r^N(\phi_{j,k-1}))(x))\|_\infty + \|\theta_j\|_\infty \\ &\leq l_{k-1} \|\theta_j\|_\infty \|(\mathcal{R}_r^N(\phi_{j,k-1}))(x)\|_\infty + \|\theta_j\|_\infty \\ &= (l_{k-1} \|(\mathcal{R}_r^N(\phi_{j,k-1}))(x)\|_\infty + 1) \|\theta_j\|_\infty \\ &\leq (l_{k-1} m_{j,k-1} + 1) \|\theta_j\|_\infty \leq m_{j,k-1} (l_{k-1} + 1) \|\theta_j\|_\infty. \end{aligned} \quad (11.94)$$

Hence, we obtain for all $j \in \{1, 2\}$, $k \in \{1, 2, \dots, L\} \cap (1, \infty)$ that

$$m_{j,k} \leq \max\{1, m_{j,k-1} (l_{k-1} + 1) \|\theta_j\|_\infty\}. \quad (11.95)$$

In addition, observe that Lemma 11.3.5 establishes that for all $j \in \{1, 2\}$, $x \in D$ it holds that

$$\begin{aligned} \|(\mathcal{R}_{\tau}^N(\phi_{j,1}))(x)\|_\infty &= \|W_{j,1}x + B_{j,1}\|_\infty \\ &\leq \left[\sup_{v \in \mathbb{R}^{l_0} \setminus \{0\}} \frac{\|W_{j,1}v\|_\infty}{\|v\|_\infty} \right] \|x\|_\infty + \|B_{j,1}\|_\infty \\ &\leq l_0 \|\theta_j\|_\infty \|x\|_\infty + \|\theta_j\|_\infty \leq l_0 \|\theta_j\|_\infty \max\{|a|, |b|\} + \|\theta_j\|_\infty \\ &= (l_0 \max\{|a|, |b|\} + 1) \|\theta_j\|_\infty \leq \mathfrak{m}_{1,0}(l_0 + 1) \|\theta_j\|_\infty. \end{aligned} \tag{11.96}$$

Therefore, we obtain that for all $j \in \{1, 2\}$ it holds that

$$\mathfrak{m}_{j,1} \leq \max\{1, \mathfrak{m}_{j,0}(l_0 + 1) \|\theta_j\|_\infty\}. \tag{11.97}$$

Combining this with (11.95) proves that for all $j \in \{1, 2\}$, $k \in \{1, 2, \dots, L\}$ it holds that

$$\mathfrak{m}_{j,k} \leq \max\{1, \mathfrak{m}_{j,k-1}(l_{k-1} + 1) \|\theta_j\|_\infty\}. \tag{11.98}$$

Hence, we obtain that for all $j \in \{1, 2\}$, $k \in \{0, 1, \dots, L\}$ it holds that

$$\mathfrak{m}_{j,k} \leq \mathfrak{m}_{j,0} \left[\prod_{n=0}^{k-1} (l_n + 1) \right] [\max\{1, \|\theta_j\|_\infty\}]^k. \tag{11.99}$$

Combining this with (11.93) shows that for all $j = (j_n)_{n \in \{0, 1, \dots, L\}} : \{0, 1, \dots, L\} \rightarrow \{1, 2\}$ and all $k \in \{1, 2, \dots, L\}$ it holds that

$$\begin{aligned} \mathfrak{e}_k &\leq \|\theta_1 - \theta_2\|_\infty \left[\sum_{n=0}^{k-1} \left(\left[\prod_{m=n+1}^{k-1} (l_m \|\theta_{3-j_m}\|_\infty) \right] \right. \right. \\ &\quad \cdot \left. \left. \left(\mathfrak{m}_{j_n,0} \left[\prod_{v=0}^{n-1} (l_v + 1) \right] \max\{1, \|\theta_{j_n}\|_\infty^n\} (l_n + 1) \right) \right) \right] \\ &= \mathfrak{m}_{1,0} \|\theta_1 - \theta_2\|_\infty \left[\sum_{n=0}^{k-1} \left(\left[\prod_{m=n+1}^{k-1} (l_m \|\theta_{3-j_m}\|_\infty) \right] \left(\left[\prod_{v=0}^n (l_v + 1) \right] \max\{1, \|\theta_{j_n}\|_\infty^n\} \right) \right) \right] \\ 554 &\leq \mathfrak{m}_{1,0} \|\theta_1 - \theta_2\|_\infty \left[\sum_{n=0}^{k-1} \left(\left[\prod_{m=n+1}^{k-1} \|\theta_{3-j_m}\|_\infty \right] \left[\prod_{v=0}^{k-1} (l_v + 1) \right] \max\{1, \|\theta_{j_n}\|_\infty^n\} \right) \right] \\ &= \mathfrak{m}_{1,0} \|\theta_1 - \theta_2\|_\infty \left[\prod_{n=0}^{k-1} (l_n + 1) \right] \left[\sum_{n=0}^{k-1} \left(\left[\prod_{m=n+1}^{k-1} \|\theta_{3-j_m}\|_\infty \right] \max\{1, \|\theta_{j_n}\|_\infty^n\} \right) \right]. \end{aligned} \tag{11.100}$$

Therefore, we obtain that for all $j \in \{1, 2\}$, $k \in \{1, 2, \dots, L\}$ it holds that

$$\begin{aligned} \epsilon_k &\leq m_{1,0} \|\theta_1 - \theta_2\|_\infty \left[\prod_{n=0}^{k-1} (l_n + 1) \right] \left[\sum_{n=0}^{k-1} \left(\left[\prod_{m=n+1}^{k-1} \|\theta_{3-j}\|_\infty \right] \max\{1, \|\theta_j\|_\infty^n\} \right) \right] \\ &= m_{1,0} \|\theta_1 - \theta_2\|_\infty \left[\prod_{n=0}^{k-1} (l_n + 1) \right] \left[\sum_{n=0}^{k-1} (\max\{1, \|\theta_j\|_\infty^n\} \|\theta_{3-j}\|_\infty^{k-1-n}) \right] \\ &\leq k m_{1,0} \|\theta_1 - \theta_2\|_\infty (\max\{1, \|\theta_1\|_\infty, \|\theta_2\|_\infty\})^{k-1} \left[\prod_{m=0}^{k-1} (l_m + 1) \right]. \end{aligned} \quad (11.101)$$

The proof of Theorem 11.3.6 is thus complete. \square

Corollary 11.3.7. Let $a \in \mathbb{R}$, $b \in [a, \infty)$, $u \in [-\infty, \infty)$, $v \in (u, \infty]$, $d, L \in \mathbb{N}$, $l = (l_0, l_1, \dots, l_L) \in \mathbb{N}^{L+1}$ satisfy

$$d \geq \sum_{k=1}^L l_k (l_{k-1} + 1). \quad (11.102)$$

Then it holds for all $\theta, \vartheta \in \mathbb{R}^d$ that

$$\begin{aligned} &\sup_{x \in [a, b]^{l_0}} \|\mathcal{N}_{u,v}^{\theta,l}(x) - \mathcal{N}_{u,v}^{\vartheta,l}(x)\|_\infty \\ &\leq L \max\{1, |a|, |b|\} (\|l\|_\infty + 1)^L (\max\{1, \|\theta\|_\infty, \|\vartheta\|_\infty\})^{L-1} \|\theta - \vartheta\|_\infty \end{aligned} \quad (11.103)$$

(cf. Definitions 3.3.4 and 4.4.1).

Proof of Corollary 11.3.7. Note that Lemma 11.3.4 and Theorem 11.3.6 demonstrate that for all $\theta, \vartheta \in \mathbb{R}^d$ it holds that

$$\begin{aligned} &\sup_{x \in [a, b]^{l_0}} \|\mathcal{N}_{u,v}^{\theta,l}(x) - \mathcal{N}_{u,v}^{\vartheta,l}(x)\|_\infty \\ &= \sup_{x \in [a, b]^{l_0}} \|\mathfrak{C}_{u,v,l_L}(\mathcal{N}_{\infty,\infty}^{\theta,l}(x)) - \mathfrak{C}_{u,v,l_L}(\mathcal{N}_{\infty,\infty}^{\vartheta,l}(x))\|_\infty \\ &\leq \sup_{x \in [a, b]^{l_0}} \|\mathcal{N}_{\infty,\infty}^{\theta,l}(x) - \mathcal{N}_{\infty,\infty}^{\vartheta,l}(x)\|_\infty \\ &\leq L \max\{1, |a|, |b|\} (\|l\|_\infty + 1)^L (\max\{1, \|\theta\|_\infty, \|\vartheta\|_\infty\})^{L-1} \|\theta - \vartheta\|_\infty \end{aligned} \quad (11.104)$$

(cf. Definitions 1.2.10, 3.3.4, and 4.4.1). The proof of Corollary 11.3.7 is thus complete. \square

11.3.2 Strong convergences rates for the optimization error involving ANNs

Lemma 11.3.8. Let $d, \mathbf{d}, \mathbf{L}, M \in \mathbb{N}$, $B, b \in [1, \infty)$, $u \in \mathbb{R}$, $v \in (u, \infty)$, $\mathbf{l} = (\mathbf{l}_0, \mathbf{l}_1, \dots, \mathbf{l}_{\mathbf{L}}) \in \mathbb{N}^{\mathbf{L}+1}$, $D \subseteq [-b, b]^d$, assume $\mathbf{l}_0 = d$, $\mathbf{l}_{\mathbf{L}} = 1$, and $\mathbf{d} \geq \sum_{i=1}^{\mathbf{L}} \mathbf{l}_i(\mathbf{l}_{i-1} + 1)$, let Ω be a set, let $X_j: \Omega \rightarrow D$, $j \in \{1, 2, \dots, M\}$, and $Y_j: \Omega \rightarrow [u, v]$, $j \in \{1, 2, \dots, M\}$, be functions, and let $\mathcal{R}: [-B, B]^{\mathbf{d}} \times \Omega \rightarrow [0, \infty)$ satisfy for all $\theta \in [-B, B]^{\mathbf{d}}$, $\omega \in \Omega$ that

$$\mathcal{R}(\theta, \omega) = \frac{1}{M} \left[\sum_{j=1}^M |\mathcal{N}_{u,v}^{\theta,\mathbf{l}}(X_j(\omega)) - Y_j(\omega)|^2 \right] \quad (11.105)$$

(cf. Definition 4.4.1). Then it holds for all $\theta, \vartheta \in [-B, B]^{\mathbf{d}}$, $\omega \in \Omega$ that

$$|\mathcal{R}(\theta, \omega) - \mathcal{R}(\vartheta, \omega)| \leq 2(v - u)b\mathbf{L}(\|\mathbf{l}\|_{\infty} + 1)^{\mathbf{L}}B^{\mathbf{L}-1}\|\theta - \vartheta\|_{\infty} \quad (11.106)$$

(cf. Definition 3.3.4).

Proof of Lemma 11.3.8. Observe that the fact that for all $x_1, x_2, y \in \mathbb{R}$ it holds that $(x_1 - y)^2 - (x_2 - y)^2 = (x_1 - x_2)((x_1 - y) + (x_2 - y))$, the fact that for all $\theta \in \mathbb{R}^{\mathbf{d}}$, $x \in \mathbb{R}^d$ it holds that $\mathcal{N}_{u,v}^{\theta,\mathbf{l}}(x) \in [u, v]$, and the assumption that for all $j \in \{1, 2, \dots, M\}$, $\omega \in \Omega$ it holds that $Y_j(\omega) \in [u, v]$ imply that for all $\theta, \vartheta \in [-B, B]^{\mathbf{d}}$, $\omega \in \Omega$ it holds that

$$\begin{aligned} & |\mathcal{R}(\theta, \omega) - \mathcal{R}(\vartheta, \omega)| \\ &= \frac{1}{M} \left| \left[\sum_{j=1}^M |\mathcal{N}_{u,v}^{\theta,\mathbf{l}}(X_j(\omega)) - Y_j(\omega)|^2 \right] - \left[\sum_{j=1}^M |\mathcal{N}_{u,v}^{\vartheta,\mathbf{l}}(X_j(\omega)) - Y_j(\omega)|^2 \right] \right| \\ &\leq \frac{1}{M} \left[\sum_{j=1}^M |[\mathcal{N}_{u,v}^{\theta,\mathbf{l}}(X_j(\omega)) - Y_j(\omega)]^2 - [\mathcal{N}_{u,v}^{\vartheta,\mathbf{l}}(X_j(\omega)) - Y_j(\omega)]^2| \right] \\ &= \frac{1}{M} \left[\sum_{j=1}^M \left(|\mathcal{N}_{u,v}^{\theta,\mathbf{l}}(X_j(\omega)) - \mathcal{N}_{u,v}^{\vartheta,\mathbf{l}}(X_j(\omega))| \right. \right. \\ &\quad \cdot \left. \left. |[\mathcal{N}_{u,v}^{\theta,\mathbf{l}}(X_j(\omega)) - Y_j(\omega)] + [\mathcal{N}_{u,v}^{\vartheta,\mathbf{l}}(X_j(\omega)) - Y_j(\omega)]| \right) \right] \quad (11.107) \\ &\leq \frac{2}{M} \left[\sum_{j=1}^M \left([\sup_{x \in D} |\mathcal{N}_{u,v}^{\theta,\mathbf{l}}(x) - \mathcal{N}_{u,v}^{\vartheta,\mathbf{l}}(x)|] [\sup_{y_1, y_2 \in [u, v]} |y_1 - y_2|] \right) \right] \\ &= 2(v - u) [\sup_{x \in D} |\mathcal{N}_{u,v}^{\theta,\mathbf{l}}(x) - \mathcal{N}_{u,v}^{\vartheta,\mathbf{l}}(x)|]. \end{aligned}$$

Furthermore, note that the assumption that $D \subseteq [-b, b]^d$, $\mathbf{d} \geq \sum_{i=1}^{\mathbf{L}} \mathbf{l}_i(\mathbf{l}_{i-1} + 1)$, $\mathbf{l}_0 = d$, $\mathbf{l}_{\mathbf{L}} = 1$, $b \geq 1$, and $B \geq 1$ and Corollary 11.3.7 (applied with $a \curvearrowright -b$, $b \curvearrowright b$, $u \curvearrowright u$, $v \curvearrowright v$, $d \curvearrowright \mathbf{d}$, $L \curvearrowright \mathbf{L}$, $l \curvearrowright \mathbf{l}$ in the notation of Corollary 11.3.7) ensure that for all $\theta, \vartheta \in [-B, B]^{\mathbf{d}}$

it holds that

$$\begin{aligned} \sup_{x \in D} |\mathcal{N}_{u,v}^{\theta,1}(x) - \mathcal{N}_{u,v}^{\vartheta,1}(x)| &\leq \sup_{x \in [-b,b]^d} |\mathcal{N}_{u,v}^{\theta,1}(x) - \mathcal{N}_{u,v}^{\vartheta,1}(x)| \\ &\leq \mathbf{L} \max\{1, b\} (\|\mathbf{l}\|_\infty + 1)^{\mathbf{L}} (\max\{1, \|\theta\|_\infty, \|\vartheta\|_\infty\})^{\mathbf{L}-1} \|\theta - \vartheta\|_\infty \\ &\leq b \mathbf{L} (\|\mathbf{l}\|_\infty + 1)^{\mathbf{L}} B^{\mathbf{L}-1} \|\theta - \vartheta\|_\infty \end{aligned} \quad (11.108)$$

(cf. Definition 3.3.4). This and (11.107) establish that for all $\theta, \vartheta \in [-B, B]^d$, $\omega \in \Omega$ it holds that

$$|\mathcal{R}(\theta, \omega) - \mathcal{R}(\vartheta, \omega)| \leq 2(v-u)b \mathbf{L} (\|\mathbf{l}\|_\infty + 1)^{\mathbf{L}} B^{\mathbf{L}-1} \|\theta - \vartheta\|_\infty. \quad (11.109)$$

The proof of Lemma 11.3.8 is thus complete. \square

Corollary 11.3.9. Let $d, \mathbf{d}, \mathfrak{d}, \mathbf{L}, M, K \in \mathbb{N}$, $B, b \in [1, \infty)$, $u \in \mathbb{R}$, $v \in (u, \infty)$, $\mathbf{l} = (\mathbf{l}_0, \mathbf{l}_1, \dots, \mathbf{l}_K) \in \mathbb{N}^{K+1}$, $D \subseteq [-b, b]^d$, assume $\mathbf{l}_0 = d$, $\mathbf{l}_K = 1$, and $\mathbf{d} \geq \mathfrak{d} = \sum_{i=1}^K \mathbf{l}_i (\mathbf{l}_{i-1} + 1)$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\Theta_k: \Omega \rightarrow [-B, B]^d$, $k \in \{1, 2, \dots, K\}$, be i.i.d. random variables, assume that Θ_1 is continuously uniformly distributed on $[-B, B]^d$, let $X_j: \Omega \rightarrow D$, $j \in \{1, 2, \dots, M\}$, and $Y_j: \Omega \rightarrow [u, v]$, $j \in \{1, 2, \dots, M\}$, be random variables, and let $\mathcal{R}: [-B, B]^d \times \Omega \rightarrow [0, \infty)$ satisfy for all $\theta \in [-B, B]^d$, $\omega \in \Omega$ that

$$\mathcal{R}(\theta, \omega) = \frac{1}{M} \left[\sum_{j=1}^M |\mathcal{N}_{u,v}^{\theta,1}(X_j(\omega)) - Y_j(\omega)|^2 \right] \quad (11.110)$$

(cf. Definition 4.4.1). Then

- (i) it holds that \mathcal{R} is a $(\mathcal{B}([-B, B]^d) \otimes \mathcal{F})/\mathcal{B}([0, \infty))$ -measurable function and
- (ii) it holds for all $\theta \in [-B, B]^d$, $p \in (0, \infty)$ that

$$\begin{aligned} &\left(\mathbb{E} [\min_{k \in \{1, 2, \dots, K\}} |\mathcal{R}(\Theta_k) - \mathcal{R}(\theta)|^p] \right)^{1/p} \\ &\leq \frac{4(v-u)b \mathbf{L} (\|\mathbf{l}\|_\infty + 1)^{\mathbf{L}} B^{\mathbf{L}} \sqrt{\max\{1, p/\mathfrak{d}\}}}{K^{1/\mathfrak{d}}} \\ &\leq \frac{4(v-u)b \mathbf{L} (\|\mathbf{l}\|_\infty + 1)^{\mathbf{L}} B^{\mathbf{L}} \max\{1, p\}}{K^{[\mathbf{L}^{-1}(\|\mathbf{l}\|_\infty + 1)^{-2}]}} \end{aligned} \quad (11.111)$$

(cf. Definition 3.3.4).

Proof of Corollary 11.3.9. Throughout this proof, let $L = 2(v-u)b \mathbf{L} (\|\mathbf{l}\|_\infty + 1)^{\mathbf{L}} B^{\mathbf{L}-1}$, let $P: [-B, B]^d \rightarrow [-B, B]^{\mathfrak{d}}$ satisfy for all $\theta = (\theta_1, \theta_2, \dots, \theta_{\mathfrak{d}}) \in [-B, B]^d$ that $P(\theta) = (\theta_1, \theta_2, \dots, \theta_{\mathfrak{d}})$, and let $R: [-B, B]^{\mathfrak{d}} \times \Omega \rightarrow \mathbb{R}$ satisfy for all $\theta \in [-B, B]^{\mathfrak{d}}$, $\omega \in \Omega$ that

$$R(\theta, \omega) = \frac{1}{M} \left[\sum_{j=1}^M |\mathcal{N}_{u,v}^{\theta,1}(X_j(\omega)) - Y_j(\omega)|^2 \right]. \quad (11.112)$$

Observe that the fact that $\forall \theta \in [-B, B]^{\mathbf{d}}: \mathcal{N}_{u,v}^{\theta,1} = \mathcal{N}_{u,v}^{P(\theta),1}$ proves that for all $\theta \in [-B, B]^{\mathbf{d}}$, $\omega \in \Omega$ it holds that

$$\begin{aligned}\mathcal{R}(\theta, \omega) &= \frac{1}{M} \left[\sum_{j=1}^M |\mathcal{N}_{u,v}^{\theta,1}(X_j(\omega)) - Y_j(\omega)|^2 \right] \\ &= \frac{1}{M} \left[\sum_{j=1}^M |\mathcal{N}_{u,v}^{P(\theta),1}(X_j(\omega)) - Y_j(\omega)|^2 \right] = R(P(\theta), \omega).\end{aligned}\tag{11.113}$$

Furthermore, note that Lemma 11.3.8 (applied with $\mathbf{d} \curvearrowright \mathfrak{d}$, $\mathcal{R} \curvearrowright ([-B, B]^{\mathfrak{d}} \times \Omega \ni (\theta, \omega) \mapsto R(\theta, \omega) \in [0, \infty))$ in the notation of Lemma 11.3.8) shows that for all $\theta, \vartheta \in [-B, B]^{\mathfrak{d}}$, $\omega \in \Omega$ it holds that

$$|R(\theta, \omega) - R(\vartheta, \omega)| \leq 2(v - u)b\mathbf{L}(\|\mathbf{l}\|_\infty + 1)^{\mathbf{L}} B^{\mathbf{L}-1} \|\theta - \vartheta\|_\infty = L\|\theta - \vartheta\|_\infty.\tag{11.114}$$

Moreover, observe that the assumption that X_j , $j \in \{1, 2, \dots, M\}$, and Y_j , $j \in \{1, 2, \dots, M\}$, are random variables demonstrates that $R: [-B, B]^{\mathfrak{d}} \times \Omega \rightarrow \mathbb{R}$ is a random field. This, (11.114), the fact that $P \circ \Theta_k: \Omega \rightarrow [-B, B]^{\mathfrak{d}}$, $k \in \{1, 2, \dots, K\}$, are i.i.d. random variables, the fact that $P \circ \Theta_1$ is continuously uniformly distributed on $[-B, B]^{\mathfrak{d}}$, and Proposition 11.2.7 (applied with $\mathbf{d} \curvearrowright \mathfrak{d}$, $\alpha \curvearrowright -B$, $\beta \curvearrowright B$, $\mathcal{R} \curvearrowright R$, $(\Theta_k)_{k \in \{1, 2, \dots, K\}} \curvearrowright (P \circ \Theta_k)_{k \in \{1, 2, \dots, K\}}$ in the notation of Proposition 11.2.7) imply that for all $\theta \in [-B, B]^{\mathbf{d}}$, $p \in (0, \infty)$ it holds that R is $(\mathcal{B}([-B, B]^{\mathfrak{d}}) \otimes \mathcal{F})/\mathcal{B}(\mathbb{R})$ -measurable and

$$\begin{aligned}&(\mathbb{E}[\min_{k \in \{1, 2, \dots, K\}} |R(P(\Theta_k)) - R(P(\theta))|^p])^{1/p} \\ &\leq \frac{L(2B) \max\{1, (p/\mathfrak{d})^{1/\mathfrak{d}}\}}{K^{1/\mathfrak{d}}} = \frac{4(v - u)b\mathbf{L}(\|\mathbf{l}\|_\infty + 1)^{\mathbf{L}} B^{\mathbf{L}} \max\{1, (p/\mathfrak{d})^{1/\mathfrak{d}}\}}{K^{1/\mathfrak{d}}}.\end{aligned}\tag{11.115}$$

The fact that P is $\mathcal{B}([-B, B]^{\mathbf{d}})/\mathcal{B}([-B, B]^{\mathfrak{d}})$ -measurable and (11.113) hence establish item (i). In addition, note that (11.113), (11.115), and the fact that $2 \leq \mathfrak{d} = \sum_{i=1}^{\mathbf{L}} \mathbf{l}_i(\mathbf{l}_{i-1} + 1) \leq \mathbf{L}(\|\mathbf{l}\|_\infty + 1)^2$ ensure that for all $\theta \in [-B, B]^{\mathbf{d}}$, $p \in (0, \infty)$ it holds that

$$\begin{aligned}&(\mathbb{E}[\min_{k \in \{1, 2, \dots, K\}} |\mathcal{R}(\Theta_k) - \mathcal{R}(\theta)|^p])^{1/p} \\ &= (\mathbb{E}[\min_{k \in \{1, 2, \dots, K\}} |R(P(\Theta_k)) - R(P(\theta))|^p])^{1/p} \\ &\leq \frac{4(v - u)b\mathbf{L}(\|\mathbf{l}\|_\infty + 1)^{\mathbf{L}} B^{\mathbf{L}} \sqrt{\max\{1, p/\mathfrak{d}\}}}{K^{1/\mathfrak{d}}} \\ &\leq \frac{4(v - u)b\mathbf{L}(\|\mathbf{l}\|_\infty + 1)^{\mathbf{L}} B^{\mathbf{L}} \max\{1, p\}}{K^{[\mathbf{L}^{-1}(\|\mathbf{l}\|_\infty + 1)^{-2}]}}.\end{aligned}\tag{11.116}$$

This proves item (ii). The proof of Corollary 11.3.9 is thus complete. \square

Part IV

Generalization

Chapter 12

Probabilistic generalization error estimates

In Chapter 15 below we establish a full error analysis for the training of ANNs in the specific situation of GD-type optimization methods with many independent random initializations (see Corollary 15.2.3). For this combined error analysis we do not only employ estimates for the approximation error (see Part II above) and the optimization error (see Part III above) but we also employ suitable generalization error estimates. Such generalization error estimates are the subject of this chapter (cf. Corollary 12.3.10 below) and the next (cf. Corollary 13.3.3 below). While in this chapter, we treat probabilistic generalization error estimates, in Chapter we will present generalization error estimates in the strong L^p -sense.

In the literature, related generalization error estimates can, for example, be found in the survey articles and books [25, 35, 36, 90, 394] and the references therein. The specific material in Section 12.1 is inspired by Duchi [122], the specific material in Section 12.2 is inspired by Cucker & Smale [90, Section 6 in Chapter I] and Carl & Stephani [63, Section 1.1], and the specific presentation of Section 12.3 is strongly based on Beck et al. [25, Section 3.2].

12.1 Concentration inequalities for random variables

12.1.1 Markov's inequality

Lemma 12.1.1 (Markov inequality). *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space, let $X: \Omega \rightarrow [0, \infty)$ be $\mathcal{F}/\mathcal{B}([0, \infty))$ -measurable, and let $\varepsilon \in (0, \infty)$. Then*

$$\mu(X \geq \varepsilon) \leq \frac{\int_{\Omega} X \, d\mu}{\varepsilon}. \quad (12.1)$$

Proof of Lemma 12.1.1. Observe that the fact that $X \geq 0$ proves that

$$\mathbb{1}_{\{X \geq \varepsilon\}} = \frac{\varepsilon \mathbb{1}_{\{X \geq \varepsilon\}}}{\varepsilon} \leq \frac{X \mathbb{1}_{\{X \geq \varepsilon\}}}{\varepsilon} \leq \frac{X}{\varepsilon}. \quad (12.2)$$

Hence, we obtain that

$$\mu(X \geq \varepsilon) = \int_{\Omega} \mathbb{1}_{\{X \geq \varepsilon\}} d\mu \leq \frac{\int_{\Omega} X d\mu}{\varepsilon}. \quad (12.3)$$

The proof of Lemma 12.1.1 is thus complete. \square

12.1.2 A first concentration inequality

12.1.2.1 On the variance of bounded random variables

Lemma 12.1.2. *Let $x \in [0, 1]$, $y \in \mathbb{R}$. Then*

$$(x - y)^2 \leq (1 - x)y^2 + x(1 - y)^2. \quad (12.4)$$

Proof of Lemma 12.1.2. Observe that the assumption that $x \in [0, 1]$ assures that

$$(1 - x)y^2 + x(1 - y)^2 = y^2 - xy^2 + x - 2xy + xy^2 \geq y^2 + x^2 - 2xy = (x - y)^2. \quad (12.5)$$

This establishes (12.4). The proof of Lemma 12.1.2 is thus complete. \square

Lemma 12.1.3. *It holds that $\sup_{p \in \mathbb{R}} p(1 - p) = \frac{1}{4}$.*

Proof of Lemma 12.1.3. Throughout this proof, let $f: \mathbb{R} \rightarrow \mathbb{R}$ satisfy for all $p \in \mathbb{R}$ that $f(p) = p(1 - p)$. Observe that the fact that $\forall p \in \mathbb{R}: f'(p) = 1 - 2p$ implies that $\{p \in \mathbb{R}: f'(p) = 0\} = \{\frac{1}{2}\}$. Combining this with the fact that f is strictly concave implies that

$$\sup_{p \in \mathbb{R}} p(1 - p) = \sup_{p \in \mathbb{R}} f(p) = f(\frac{1}{2}) = \frac{1}{4}. \quad (12.6)$$

The proof of Lemma 12.1.3 is thus complete. \square

Lemma 12.1.4. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X: \Omega \rightarrow [0, 1]$ be a random variable. Then*

$$\text{Var}(X) \leq \frac{1}{4}. \quad (12.7)$$

Proof of Lemma 12.1.4. Observe that Lemma 12.1.2 implies that

$$\begin{aligned}\text{Var}(X) &= \mathbb{E}[(X - \mathbb{E}[X])^2] \leq \mathbb{E}[(1 - X)(\mathbb{E}[X])^2 + X(1 - \mathbb{E}[X])^2] \\ &= (1 - \mathbb{E}[X])(\mathbb{E}[X])^2 + \mathbb{E}[X](1 - \mathbb{E}[X])^2 \\ &= (1 - \mathbb{E}[X])\mathbb{E}[X](\mathbb{E}[X] + (1 - \mathbb{E}[X])) \\ &= (1 - \mathbb{E}[X])\mathbb{E}[X].\end{aligned}\tag{12.8}$$

This and Lemma 12.1.3 demonstrate that $\text{Var}(X) \leq 1/4$. The proof of Lemma 12.1.4 is thus complete. \square

Lemma 12.1.5. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $a \in \mathbb{R}$, $b \in [a, \infty)$, and let $X: \Omega \rightarrow [a, b]$ be a random variable. Then

$$\text{Var}(X) \leq \frac{(b - a)^2}{4}.\tag{12.9}$$

Proof of Lemma 12.1.5. Throughout this proof, assume without loss of generality that $a < b$. Observe that Lemma 12.1.4 implies that

$$\begin{aligned}\text{Var}(X) &= \mathbb{E}[(X - \mathbb{E}[X])^2] = (b - a)^2 \mathbb{E}\left[\left(\frac{X-a-(\mathbb{E}[X]-a)}{b-a}\right)^2\right] \\ &= (b - a)^2 \mathbb{E}\left[\left(\frac{X-a}{b-a} - \mathbb{E}\left[\frac{X-a}{b-a}\right]\right)^2\right] \\ &= (b - a)^2 \text{Var}\left(\frac{X-a}{b-a}\right) \leq (b - a)^2\left(\frac{1}{4}\right) = \frac{(b - a)^2}{4}.\end{aligned}\tag{12.10}$$

The proof of Lemma 12.1.5 is thus complete. \square

12.1.2.2 A concentration inequality

Lemma 12.1.6. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $N \in \mathbb{N}$, $\varepsilon \in (0, \infty)$, $a_1, a_2, \dots, a_N \in \mathbb{R}$, $b_1 \in [a_1, \infty)$, $b_2 \in [a_2, \infty)$, \dots , $b_N \in [a_N, \infty)$, and let $X_n: \Omega \rightarrow [a_n, b_n]$, $n \in \{1, 2, \dots, N\}$, be independent random variables. Then

$$\mathbb{P}\left(\left|\sum_{n=1}^N (X_n - \mathbb{E}[X_n])\right| \geq \varepsilon\right) \leq \frac{\sum_{n=1}^N (b_n - a_n)^2}{4\varepsilon^2}.\tag{12.11}$$

Proof of Lemma 12.1.6. Note that Lemma 12.1.1 assures that

$$\begin{aligned}\mathbb{P}\left(\left|\sum_{n=1}^N (X_n - \mathbb{E}[X_n])\right| \geq \varepsilon\right) &= \mathbb{P}\left(\left|\sum_{n=1}^N (X_n - \mathbb{E}[X_n])\right|^2 \geq \varepsilon^2\right) \\ &\leq \frac{\mathbb{E}\left[\left|\sum_{n=1}^N (X_n - \mathbb{E}[X_n])\right|^2\right]}{\varepsilon^2}.\end{aligned}\tag{12.12}$$

In addition, note that the assumption that $X_n: \Omega \rightarrow [a_n, b_n]$, $n \in \{1, 2, \dots, N\}$, are independent variables and Lemma 12.1.5 demonstrate that

$$\begin{aligned}\mathbb{E}\left[\left|\sum_{n=1}^N (X_n - \mathbb{E}[X_n])\right|^2\right] &= \sum_{n,m=1}^N \mathbb{E}\left[(X_n - \mathbb{E}[X_n])(X_m - \mathbb{E}[X_m])\right] \\ &= \sum_{n=1}^N \mathbb{E}\left[(X_n - \mathbb{E}[X_n])^2\right] \leq \frac{\sum_{n=1}^N (b_n - a_n)^2}{4}.\end{aligned}\tag{12.13}$$

Combining this with (12.12) establishes

$$\mathbb{P}\left(\left|\sum_{n=1}^N (X_n - \mathbb{E}[X_n])\right| \geq \varepsilon\right) \leq \frac{\sum_{n=1}^N (b_n - a_n)^2}{4\varepsilon^2}\tag{12.14}$$

The proof of Lemma 12.1.6 is thus complete. \square

12.1.3 Moment-generating functions

Definition 12.1.7 (Moment generating functions). *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X: \Omega \rightarrow \mathbb{R}$ be a random variable. Then we denote by $\mathbb{M}_{X,\mathbb{P}}: \mathbb{R} \rightarrow [0, \infty]$ (we denote by $\mathbb{M}_X: \mathbb{R} \rightarrow [0, \infty]$) the function which satisfies for all $t \in \mathbb{R}$ that*

$$\mathbb{M}_{X,\mathbb{P}}(t) = \mathbb{E}[e^{tX}]\tag{12.15}$$

and we call $\mathbb{M}_{X,\mathbb{P}}$ the moment-generating function of X with respect to \mathbb{P} (we call \mathbb{M}_X the moment-generating function of X).

12.1.3.1 Moment-generation function for the sum of independent random variables

Lemma 12.1.8. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $t \in \mathbb{R}$, $N \in \mathbb{N}$, and let $X_n: \Omega \rightarrow \mathbb{R}$, $n \in \{1, 2, \dots, N\}$, be independent random variables. Then*

$$\mathbb{M}_{\sum_{n=1}^N X_n}(t) = \prod_{n=1}^N \mathbb{M}_{X_n}(t).\tag{12.16}$$

Proof of Lemma 12.1.8. Observe that Fubini's theorem ensures that for all $t \in \mathbb{R}$ it holds that

$$\mathbb{M}_{\sum_{n=1}^N X_n}(t) = \mathbb{E}\left[e^{t(\sum_{n=1}^N X_n)}\right] = \mathbb{E}\left[\prod_{n=1}^N e^{tX_n}\right] = \prod_{n=1}^N \mathbb{E}[e^{tX_n}] = \prod_{n=1}^N \mathbb{M}_{X_n}(t).\tag{12.17}$$

The proof of Lemma 12.1.8 is thus complete. \square

12.1.4 Chernoff bounds

12.1.4.1 Probability to cross a barrier

Proposition 12.1.9. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $X: \Omega \rightarrow \mathbb{R}$ be a random variable, and let $\varepsilon \in \mathbb{R}$. Then

$$\mathbb{P}(X \geq \varepsilon) \leq \inf_{\lambda \in [0, \infty)} (e^{-\lambda\varepsilon} \mathbb{E}[e^{\lambda X}]) = \inf_{\lambda \in [0, \infty)} (e^{-\lambda\varepsilon} \mathbb{M}_X(\lambda)). \quad (12.18)$$

Proof of Proposition 12.1.9. Note that Lemma 12.1.1 ensures that for all $\lambda \in [0, \infty)$ it holds that

$$\mathbb{P}(X \geq \varepsilon) \leq \mathbb{P}(\lambda X \geq \lambda\varepsilon) = \mathbb{P}(\exp(\lambda X) \geq \exp(\lambda\varepsilon)) \leq \frac{\mathbb{E}[\exp(\lambda X)]}{\exp(\lambda\varepsilon)} = e^{-\lambda\varepsilon} \mathbb{E}[e^{\lambda X}]. \quad (12.19)$$

The proof of Proposition 12.1.9 is thus complete. \square

Corollary 12.1.10. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $X: \Omega \rightarrow \mathbb{R}$ be a random variable, and let $c, \varepsilon \in \mathbb{R}$. Then

$$\mathbb{P}(X \geq c + \varepsilon) \leq \inf_{\lambda \in [0, \infty)} (e^{-\lambda\varepsilon} \mathbb{M}_{X-c}(\lambda)). \quad (12.20)$$

Proof of Corollary 12.1.10. Throughout this proof, let $Y: \Omega \rightarrow \mathbb{R}$ satisfy

$$Y = X - c. \quad (12.21)$$

Observe that Proposition 12.1.9 and (12.21) ensure that

$$\mathbb{P}(X - c \geq \varepsilon) = \mathbb{P}(Y \geq \varepsilon) \leq \inf_{\lambda \in [0, \infty)} (e^{-\lambda\varepsilon} \mathbb{M}_Y(\lambda)) = \inf_{\lambda \in [0, \infty)} (e^{-\lambda\varepsilon} \mathbb{M}_{X-c}(\lambda)). \quad (12.22)$$

The proof of Corollary 12.1.10 is thus complete. \square

Corollary 12.1.11. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $X: \Omega \rightarrow \mathbb{R}$ be a random variable with $\mathbb{E}[|X|] < \infty$, and let $\varepsilon \in \mathbb{R}$. Then

$$\mathbb{P}(X \geq \mathbb{E}[X] + \varepsilon) \leq \inf_{\lambda \in [0, \infty)} (e^{-\lambda\varepsilon} \mathbb{M}_{X-\mathbb{E}[X]}(\lambda)). \quad (12.23)$$

Proof of Corollary 12.1.11. Observe that Corollary 12.1.10 (applied with $c \curvearrowright \mathbb{E}[X]$ in the notation of Corollary 12.1.10) establishes (12.23). The proof of Corollary 12.1.11 is thus complete. \square

12.1.4.2 Probability to fall below a barrier

Corollary 12.1.12. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $X: \Omega \rightarrow \mathbb{R}$ be a random variable, and let $c, \varepsilon \in \mathbb{R}$. Then

$$\mathbb{P}(X \leq c - \varepsilon) \leq \inf_{\lambda \in [0, \infty)} (e^{-\lambda\varepsilon} \mathbb{M}_{c-X}(\lambda)). \quad (12.24)$$

Proof of Corollary 12.1.12. Throughout this proof, let $\mathfrak{c} \in \mathbb{R}$ satisfy $\mathfrak{c} = -c$ and let $\mathfrak{X}: \Omega \rightarrow \mathbb{R}$ satisfy

$$\mathfrak{X} = -X. \quad (12.25)$$

Observe that Corollary 12.1.10 and (12.25) ensure that

$$\begin{aligned} \mathbb{P}(X \leq c - \varepsilon) &= \mathbb{P}(-X \geq -c + \varepsilon) = \mathbb{P}(\mathfrak{X} \geq \mathfrak{c} + \varepsilon) \leq \inf_{\lambda \in [0, \infty)} (e^{-\lambda\varepsilon} \mathbb{M}_{\mathfrak{X}-\mathfrak{c}}(\lambda)) \\ &= \inf_{\lambda \in [0, \infty)} (e^{-\lambda\varepsilon} \mathbb{M}_{c-X}(\lambda)). \end{aligned} \quad (12.26)$$

The proof of Corollary 12.1.12 is thus complete. \square

12.1.4.3 Sums of independent random variables

Corollary 12.1.13. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\varepsilon \in \mathbb{R}$, $N \in \mathbb{N}$, and let $X_n: \Omega \rightarrow \mathbb{R}$, $n \in \{1, 2, \dots, N\}$, be independent random variables with $\sum_{n=1}^N \mathbb{E}[|X_n|] < \infty$. Then

$$\mathbb{P}\left(\left[\sum_{n=1}^N (X_n - \mathbb{E}[X_n])\right] \geq \varepsilon\right) \leq \inf_{\lambda \in [0, \infty)} \left(e^{-\lambda\varepsilon} \left[\prod_{n=1}^N \mathbb{M}_{X_n - \mathbb{E}[X_n]}(\lambda) \right] \right). \quad (12.27)$$

Proof of Corollary 12.1.13. Throughout this proof, let $Y_n: \Omega \rightarrow \mathbb{R}$, $n \in \{1, 2, \dots, N\}$, satisfy for all $n \in \{1, 2, \dots, N\}$ that

$$Y_n = X_n - \mathbb{E}[X_n]. \quad (12.28)$$

Observe that Proposition 12.1.9, Lemma 12.1.8, and (12.28) ensure that

$$\begin{aligned} \mathbb{P}\left(\left[\sum_{n=1}^N (X_n - \mathbb{E}[X_n])\right] \geq \varepsilon\right) &= \mathbb{P}\left(\left[\sum_{n=1}^N Y_n\right] \geq \varepsilon\right) \leq \inf_{\lambda \in [0, \infty)} \left(e^{-\lambda\varepsilon} \mathbb{M}_{\sum_{n=1}^N Y_n}(\lambda) \right) \\ &= \inf_{\lambda \in [0, \infty)} \left(e^{-\lambda\varepsilon} \left[\prod_{n=1}^N \mathbb{M}_{Y_n}(\lambda) \right] \right) = \inf_{\lambda \in [0, \infty)} \left(e^{-\lambda\varepsilon} \left[\prod_{n=1}^N \mathbb{M}_{X_n - \mathbb{E}[X_n]}(\lambda) \right] \right). \end{aligned} \quad (12.29)$$

The proof of Corollary 12.1.13 is thus complete. \square

12.1.5 Hoeffding's inequality

12.1.5.1 On the moment-generating function for bounded random variables

Lemma 12.1.14. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\lambda, a \in \mathbb{R}$, $b \in (a, \infty)$, $p \in [0, 1]$ satisfy $p = \frac{-a}{(b-a)}$, let $X: \Omega \rightarrow [a, b]$ be a random variable with $\mathbb{E}[X] = 0$, and let $\phi: \mathbb{R} \rightarrow \mathbb{R}$ satisfy for all $x \in \mathbb{R}$ that $\phi(x) = \ln(1 - p + pe^x) - px$. Then

$$\mathbb{E}[e^{\lambda X}] \leq e^{\phi(\lambda(b-a))}. \quad (12.30)$$

Proof of Lemma 12.1.14. Observe that for all $x \in \mathbb{R}$ it holds that

$$\begin{aligned} x(b-a) &= bx - ax = [ab - ax] + [bx - ab] = [a(b-x)] + [b(x-a)] \\ &= a(b-x) + b[b-a-b+x] = a(b-x) + b[(b-a)-(b-x)]. \end{aligned} \quad (12.31)$$

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

$$x = a\left(\frac{b-x}{b-a}\right) + b\left[1 - \left(\frac{b-x}{b-a}\right)\right]. \quad (12.32)$$

This implies that for all $x \in \mathbb{R}$ it holds that

$$\lambda x = \left(\frac{b-x}{b-a}\right)\lambda a + \left[1 - \left(\frac{b-x}{b-a}\right)\right]\lambda b. \quad (12.33)$$

The fact that $\mathbb{R} \ni x \mapsto e^x \in \mathbb{R}$ is convex hence demonstrates that for all $x \in [a, b]$ it holds that

$$e^{\lambda x} = \exp\left(\left(\frac{b-x}{b-a}\right)\lambda a + \left[1 - \left(\frac{b-x}{b-a}\right)\right]\lambda b\right) \leq \left(\frac{b-x}{b-a}\right)e^{\lambda a} + \left[1 - \left(\frac{b-x}{b-a}\right)\right]e^{\lambda b}. \quad (12.34)$$

The assumption that $\mathbb{E}[X] = 0$ therefore assures that

$$\mathbb{E}[e^{\lambda X}] \leq \left(\frac{b}{b-a}\right)e^{\lambda a} + \left[1 - \left(\frac{b}{b-a}\right)\right]e^{\lambda b}. \quad (12.35)$$

Combining this with the fact that

$$\begin{aligned} \frac{b}{(b-a)} &= 1 - \left[1 - \left(\frac{b}{(b-a)}\right)\right] \\ &= 1 - \left[\left(\frac{(b-a)}{(b-a)}\right) - \left(\frac{b}{(b-a)}\right)\right] \\ &= 1 - \left[\frac{-a}{(b-a)}\right] = 1 - p \end{aligned} \quad (12.36)$$

demonstrates that

$$\begin{aligned}
 \mathbb{E}[e^{\lambda X}] &\leq \left(\frac{b}{b-a}\right)e^{\lambda a} + \left[1 - \left(\frac{b}{b-a}\right)\right]e^{\lambda b} \\
 &= (1-p)e^{\lambda a} + [1 - (1-p)]e^{\lambda b} \\
 &= (1-p)e^{\lambda a} + p e^{\lambda b} \\
 &= [(1-p) + p e^{\lambda(b-a)}]e^{\lambda a}.
 \end{aligned} \tag{12.37}$$

Moreover, note that the assumption that $p = \frac{-a}{(b-a)}$ shows that $p(b-a) = -a$. Hence, we obtain that $a = -p(b-a)$. This and (12.37) assure that

$$\begin{aligned}
 \mathbb{E}[e^{\lambda X}] &\leq [(1-p) + p e^{\lambda(b-a)}]e^{-p\lambda(b-a)} = \exp(\ln([(1-p) + p e^{\lambda(b-a)}]e^{-p\lambda(b-a)})) \\
 &= \exp(\ln((1-p) + p e^{\lambda(b-a)}) - p\lambda(b-a)) = \exp(\phi(\lambda(b-a))).
 \end{aligned} \tag{12.38}$$

The proof of Lemma 12.1.14 is thus complete. \square

12.1.5.2 Hoeffding's lemma

Lemma 12.1.15. Let $p \in [0, 1]$ and let $\phi: \mathbb{R} \rightarrow \mathbb{R}$ satisfy for all $x \in \mathbb{R}$ that $\phi(x) = \ln(1 - p + pe^x) - px$. Then it holds for all $x \in \mathbb{R}$ that $\phi(x) \leq \frac{x^2}{8}$.

Proof of Lemma 12.1.15. Observe that the fundamental theorem of calculus ensures that for all $x \in \mathbb{R}$ it holds that

$$\begin{aligned}
 \phi(x) &= \phi(0) + \int_0^x \phi'(y) dy \\
 &= \phi(0) + \phi'(0)x + \int_0^x \int_0^y \phi''(z) dz dy \\
 &\leq \phi(0) + \phi'(0)x + \frac{x^2}{2} \left[\sup_{z \in \mathbb{R}} \phi''(z) \right].
 \end{aligned} \tag{12.39}$$

Moreover, note that for all $x \in \mathbb{R}$ it holds that

$$\phi'(x) = \left[\frac{pe^x}{1-p+pe^x} \right] - p \quad \text{and} \quad \phi''(x) = \left[\frac{pe^x}{1-p+pe^x} \right] - \left[\frac{p^2e^{2x}}{(1-p+pe^x)^2} \right]. \tag{12.40}$$

Hence, we obtain that

$$\phi'(0) = \left[\frac{p}{1-p+p} \right] - p = 0. \tag{12.41}$$

In the next step we combine (12.40) and the fact that for all $a \in \mathbb{R}$ it holds that

$$a(1-a) = a - a^2 = -\left[a^2 - 2a[\tfrac{1}{2}] + [\tfrac{1}{2}]^2\right] + [\tfrac{1}{2}]^2 = \tfrac{1}{4} - [a - \tfrac{1}{2}]^2 \leq \tfrac{1}{4} \tag{12.42}$$

to obtain that for all $x \in \mathbb{R}$ it holds that $\phi''(x) \leq \frac{1}{4}$. This, (12.39), and (12.41) ensure that for all $x \in \mathbb{R}$ it holds that

$$\phi(x) \leq \phi(0) + \phi'(0)x + \frac{x^2}{2} \left[\sup_{z \in \mathbb{R}} \phi''(z) \right] = \phi(0) + \frac{x^2}{2} \left[\sup_{z \in \mathbb{R}} \phi''(z) \right] \leq \phi(0) + \frac{x^2}{8} = \frac{x^2}{8}. \quad (12.43)$$

The proof of Lemma 12.1.15 is thus complete. \square

Lemma 12.1.16. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $a \in \mathbb{R}$, $b \in [a, \infty)$, $\lambda \in \mathbb{R}$, and let $X: \Omega \rightarrow [a, b]$ be a random variable with $\mathbb{E}[X] = 0$. Then

$$\mathbb{E}[\exp(\lambda X)] \leq \exp\left(\frac{\lambda^2(b-a)^2}{8}\right). \quad (12.44)$$

Proof of Lemma 12.1.16. Throughout this proof, assume without loss of generality that $a < b$, let $p \in \mathbb{R}$ satisfy $p = \frac{-a}{(b-a)}$, and let $\phi_r: \mathbb{R} \rightarrow \mathbb{R}$, $r \in [0, 1]$, satisfy for all $r \in [0, 1]$, $x \in \mathbb{R}$ that

$$\phi_r(x) = \ln(1 - r + re^x) - rx. \quad (12.45)$$

Observe that the assumption that $\mathbb{E}[X] = 0$ and the fact that $a \leq \mathbb{E}[X] \leq b$ ensures that $a \leq 0 \leq b$. Combining this with the assumption that $a < b$ implies that

$$0 \leq p = \frac{-a}{(b-a)} \leq \frac{(b-a)}{(b-a)} = 1. \quad (12.46)$$

Lemma 12.1.14 and Lemma 12.1.15 hence demonstrate that

$$\mathbb{E}[e^{\lambda X}] \leq e^{\phi_p(\lambda(b-a))} = \exp(\phi_p(\lambda(b-a))) \leq \exp\left(\frac{(\lambda(b-a))^2}{8}\right) = \exp\left(\frac{\lambda^2(b-a)^2}{8}\right). \quad (12.47)$$

The proof of Lemma 12.1.16 is thus complete. \square

12.1.5.3 Probability to cross a barrier

Lemma 12.1.17. Let $\beta \in (0, \infty)$, $\varepsilon \in [0, \infty)$ and let $f: [0, \infty) \rightarrow [0, \infty)$ satisfy for all $\lambda \in [0, \infty)$ that $f(\lambda) = \beta\lambda^2 - \varepsilon\lambda$. Then

$$\inf_{\lambda \in [0, \infty)} f(\lambda) = f\left(\frac{\varepsilon}{2\beta}\right) = -\frac{\varepsilon^2}{4\beta}. \quad (12.48)$$

Proof of Lemma 12.1.17. Observe that for all $\lambda \in \mathbb{R}$ it holds that

$$f'(\lambda) = 2\beta\lambda - \varepsilon. \quad (12.49)$$

Moreover, note that

$$f\left(\frac{\varepsilon}{2\beta}\right) = \beta \left[\frac{\varepsilon}{2\beta}\right]^2 - \varepsilon \left[\frac{\varepsilon}{2\beta}\right] = \frac{\varepsilon^2}{4\beta} - \frac{\varepsilon^2}{2\beta} = -\frac{\varepsilon^2}{4\beta}. \quad (12.50)$$

Combining this and (12.49) establishes (12.48). The proof of Lemma 12.1.17 is thus complete. \square

Corollary 12.1.18. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $N \in \mathbb{N}$, $\varepsilon \in [0, \infty)$, $a_1, a_2, \dots, a_N \in \mathbb{R}$, $b_1 \in [a_1, \infty)$, $b_2 \in [a_2, \infty)$, \dots , $b_N \in [a_N, \infty)$ satisfy $\sum_{n=1}^N (b_n - a_n)^2 \neq 0$, and let $X_n: \Omega \rightarrow [a_n, b_n]$, $n \in \{1, 2, \dots, N\}$, be independent random variables. Then

$$\mathbb{P}\left(\left[\sum_{n=1}^N (X_n - \mathbb{E}[X_n])\right] \geq \varepsilon\right) \leq \exp\left(\frac{-2\varepsilon^2}{\sum_{n=1}^N (b_n - a_n)^2}\right). \quad (12.51)$$

Proof of Corollary 12.1.18. Throughout this proof, let $\beta \in (0, \infty)$ satisfy

$$\beta = \frac{1}{8} \left[\sum_{n=1}^N (b_n - a_n)^2 \right]. \quad (12.52)$$

Observe that Corollary 12.1.13 ensures that

$$\mathbb{P}\left(\left[\sum_{n=1}^N (X_n - \mathbb{E}[X_n])\right] \geq \varepsilon\right) \leq \inf_{\lambda \in [0, \infty)} \left(e^{-\lambda\varepsilon} \left[\prod_{n=1}^N \mathbb{M}_{X_n - \mathbb{E}[X_n]}(\lambda) \right] \right). \quad (12.53)$$

Moreover, note that Lemma 12.1.16 proves that for all $n \in \{1, 2, \dots, N\}$ it holds that

$$\mathbb{M}_{X_n - \mathbb{E}[X_n]}(\lambda) \leq \exp\left(\frac{\lambda^2[(b_n - \mathbb{E}[X_n]) - (a_n - \mathbb{E}[X_n])]^2}{8}\right) = \exp\left(\frac{\lambda^2(b_n - a_n)^2}{8}\right). \quad (12.54)$$

Combining this with (12.53) and Lemma 12.1.17 ensures that

$$\begin{aligned} \mathbb{P}\left(\left[\sum_{n=1}^N (X_n - \mathbb{E}[X_n])\right] \geq \varepsilon\right) &\leq \inf_{\lambda \in [0, \infty)} \left(\exp\left(\left[\sum_{n=1}^N \left(\frac{\lambda^2(b_n - a_n)^2}{8}\right)\right] - \lambda\varepsilon\right) \right) \\ &= \inf_{\lambda \in [0, \infty)} \left[\exp\left(\lambda^2 \left[\frac{\sum_{n=1}^N (b_n - a_n)^2}{8}\right] - \lambda\varepsilon\right) \right] = \exp\left(\inf_{\lambda \in [0, \infty)} [\beta\lambda^2 - \varepsilon\lambda]\right) \\ &= \exp\left(\frac{-\varepsilon^2}{4\beta}\right) = \exp\left(\frac{-2\varepsilon^2}{\sum_{n=1}^N (b_n - a_n)^2}\right). \end{aligned} \quad (12.55)$$

The proof of Corollary 12.1.18 is thus complete. \square

12.1.5.4 Probability to fall below a barrier

Corollary 12.1.19. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $N \in \mathbb{N}$, $\varepsilon \in [0, \infty)$, $a_1, a_2, \dots, a_N \in \mathbb{R}$, $b_1 \in [a_1, \infty)$, $b_2 \in [a_2, \infty)$, \dots , $b_N \in [a_N, \infty)$ satisfy $\sum_{n=1}^N (b_n - a_n)^2 \neq 0$, and let $X_n: \Omega \rightarrow [a_n, b_n]$, $n \in \{1, 2, \dots, N\}$, be independent random variables. Then

$$\mathbb{P}\left(\left[\sum_{n=1}^N (X_n - \mathbb{E}[X_n])\right] \leq -\varepsilon\right) \leq \exp\left(\frac{-2\varepsilon^2}{\sum_{n=1}^N (b_n - a_n)^2}\right). \quad (12.56)$$

Proof of Corollary 12.1.19. Throughout this proof, let $\mathfrak{X}_n: \Omega \rightarrow [-b_n, -a_n]$, $n \in \{1, 2, \dots, N\}$, satisfy for all $n \in \{1, 2, \dots, N\}$ that

$$\mathfrak{X}_n = -X_n. \quad (12.57)$$

Observe that Corollary 12.1.18 and (12.57) ensure that

$$\begin{aligned} & \mathbb{P}\left(\left[\sum_{n=1}^N (X_n - \mathbb{E}[X_n])\right] \leq -\varepsilon\right) \\ &= \mathbb{P}\left(\left[\sum_{n=1}^N (-X_n - \mathbb{E}[-X_n])\right] \geq \varepsilon\right) \\ &= \mathbb{P}\left(\left[\sum_{n=1}^N (\mathfrak{X}_n - \mathbb{E}[\mathfrak{X}_n])\right] \geq \varepsilon\right) \leq \exp\left(\frac{-2\varepsilon^2}{\sum_{n=1}^N (b_n - a_n)^2}\right). \end{aligned} \quad (12.58)$$

The proof of Corollary 12.1.19 is thus complete. \square

12.1.5.5 Hoeffding's inequality

Corollary 12.1.20. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $N \in \mathbb{N}$, $\varepsilon \in [0, \infty)$, $a_1, a_2, \dots, a_N \in \mathbb{R}$, $b_1 \in [a_1, \infty)$, $b_2 \in [a_2, \infty)$, \dots , $b_N \in [a_N, \infty)$ satisfy $\sum_{n=1}^N (b_n - a_n)^2 \neq 0$, and let $X_n: \Omega \rightarrow [a_n, b_n]$, $n \in \{1, 2, \dots, N\}$, be independent random variables. Then

$$\mathbb{P}\left(\left|\sum_{n=1}^N (X_n - \mathbb{E}[X_n])\right| \geq \varepsilon\right) \leq 2 \exp\left(\frac{-2\varepsilon^2}{\sum_{n=1}^N (b_n - a_n)^2}\right). \quad (12.59)$$

Proof of Corollary 12.1.20. Observe that

$$\begin{aligned}
 & \mathbb{P}\left(\left|\sum_{n=1}^N (X_n - \mathbb{E}[X_n])\right| \geq \varepsilon\right) \\
 &= \mathbb{P}\left(\left\{\left[\sum_{n=1}^N (X_n - \mathbb{E}[X_n])\right] \geq \varepsilon\right\} \cup \left\{\left[\sum_{n=1}^N (X_n - \mathbb{E}[X_n])\right] \leq -\varepsilon\right\}\right) \\
 &\leq \mathbb{P}\left(\left[\sum_{n=1}^N (X_n - \mathbb{E}[X_n])\right] \geq \varepsilon\right) + \mathbb{P}\left(\left[\sum_{n=1}^N (X_n - \mathbb{E}[X_n])\right] \leq -\varepsilon\right).
 \end{aligned} \tag{12.60}$$

Combining this with Corollary 12.1.18 and Corollary 12.1.19 establishes (12.59). The proof of Corollary 12.1.20 is thus complete. \square

Corollary 12.1.21. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $N \in \mathbb{N}$, $\varepsilon \in [0, \infty)$, $a_1, a_2, \dots, a_N \in \mathbb{R}$, $b_1 \in [a_1, \infty)$, $b_2 \in [a_2, \infty)$, \dots , $b_N \in [a_N, \infty)$ satisfy $\sum_{n=1}^N (b_n - a_n)^2 \neq 0$, and let $X_n: \Omega \rightarrow [a_n, b_n]$, $n \in \{1, 2, \dots, N\}$, be independent random variables. Then

$$\mathbb{P}\left(\frac{1}{N} \left| \sum_{n=1}^N (X_n - \mathbb{E}[X_n]) \right| \geq \varepsilon\right) \leq 2 \exp\left(\frac{-2\varepsilon^2 N^2}{\sum_{n=1}^N (b_n - a_n)^2}\right). \tag{12.61}$$

Proof of Corollary 12.1.21. Observe that Corollary 12.1.20 ensures that

$$\begin{aligned}
 \mathbb{P}\left(\frac{1}{N} \left| \sum_{n=1}^N (X_n - \mathbb{E}[X_n]) \right| \geq \varepsilon\right) &= \mathbb{P}\left(\left| \sum_{n=1}^N (X_n - \mathbb{E}[X_n]) \right| \geq \varepsilon N\right) \\
 &\leq 2 \exp\left(\frac{-2(\varepsilon N)^2}{\sum_{n=1}^N (b_n - a_n)^2}\right).
 \end{aligned} \tag{12.62}$$

The proof of Corollary 12.1.21 is thus complete. \square

Exercise 12.1.1. Prove or disprove the following statement: For every probability space $(\Omega, \mathcal{F}, \mathbb{P})$, every $N \in \mathbb{N}$, $\varepsilon \in [0, \infty)$, and every random variable $X = (X_1, X_2, \dots, X_N): \Omega \rightarrow [-1, 1]^N$ with $\forall a = (a_1, a_2, \dots, a_N) \in [-1, 1]^N: \mathbb{P}(\bigcap_{i=1}^N \{X_i \leq a_i\}) = \prod_{i=1}^N \frac{a_i+1}{2}$ it holds that

$$\mathbb{P}\left(\frac{1}{N} \left| \sum_{i=1}^N (X_i - \mathbb{E}[X_i]) \right| \geq \varepsilon\right) \leq 2 \exp\left(\frac{-\varepsilon^2 N}{2}\right). \tag{12.63}$$

Exercise 12.1.2. Prove or disprove the following statement: For every probability space $(\Omega, \mathcal{F}, \mathbb{P})$, every $N \in \mathbb{N}$, and every random variable $X = (X_1, X_2, \dots, X_N): \Omega \rightarrow [-1, 1]^N$ with $\forall a = (a_1, a_2, \dots, a_N) \in [-1, 1]^N: \mathbb{P}(\bigcap_{i=1}^N \{X_i \leq a_i\}) = \prod_{i=1}^N \frac{a_i+1}{2}$ it holds that

$$\mathbb{P}\left(\frac{1}{N} \left| \sum_{n=1}^N (X_n - \mathbb{E}[X_n]) \right| \geq \frac{1}{2}\right) \leq 2 \left[\frac{e}{4}\right]^N. \tag{12.64}$$

Exercise 12.1.3. Prove or disprove the following statement: For every probability space $(\Omega, \mathcal{F}, \mathbb{P})$, every $N \in \mathbb{N}$, and every random variable $X = (X_1, X_2, \dots, X_N) : \Omega \rightarrow [-1, 1]^N$ with $\forall a = (a_1, a_2, \dots, a_N) \in [-1, 1]^N : \mathbb{P}(\bigcap_{i=1}^N \{X_i \leq a_i\}) = \prod_{i=1}^N \frac{a_i+1}{2}$ it holds that

$$\mathbb{P}\left(\frac{1}{N} \left| \sum_{n=1}^N (X_n - \mathbb{E}[X_n]) \right| \geq \frac{1}{2}\right) \leq 2 \left[\frac{e - e^{-3}}{4} \right]^N. \quad (12.65)$$

Exercise 12.1.4. Prove or disprove the following statement: For every probability space $(\Omega, \mathcal{F}, \mathbb{P})$, every $N \in \mathbb{N}$, $\varepsilon \in [0, \infty)$, and every standard normal random variable $X = (X_1, X_2, \dots, X_N) : \Omega \rightarrow \mathbb{R}^N$ it holds that

$$\mathbb{P}\left(\frac{1}{N} \left| \sum_{n=1}^N (X_n - \mathbb{E}[X_n]) \right| \geq \varepsilon\right) \leq 2 \exp\left(\frac{-\varepsilon^2 N}{2}\right). \quad (12.66)$$

12.1.6 A strengthened Hoeffding's inequality

Lemma 12.1.22. Let $f, g : (0, \infty) \rightarrow \mathbb{R}$ satisfy for all $x \in (0, \infty)$ that $f(x) = 2 \exp(-2x)$ and $g(x) = \frac{1}{4x}$. Then

(i) it holds that $\lim_{x \rightarrow \infty} \frac{f(x)}{g(x)} = \lim_{x \searrow 0} \frac{f(x)}{g(x)} = 0$ and

(ii) it holds that $g\left(\frac{1}{2}\right) = \frac{1}{2} < \frac{2}{3} < \frac{2}{e} = f\left(\frac{1}{2}\right)$.

Proof of Lemma 12.1.22. Note that the fact that $\lim_{x \rightarrow \infty} \frac{\exp(-x)}{x^{-1}} = \lim_{x \searrow 0} \frac{\exp(-x)}{x^{-1}} = 0$ establishes item (i). Moreover, observe that the fact that $e < 3$ implies item (ii). The proof of Lemma 12.1.22 is thus complete. \square

Corollary 12.1.23. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $N \in \mathbb{N}$, $\varepsilon \in (0, \infty)$, $a_1, a_2, \dots, a_N \in \mathbb{R}$, $b_1 \in [a_1, \infty)$, $b_2 \in [a_2, \infty)$, \dots , $b_N \in [a_N, \infty)$ satisfy $\sum_{n=1}^N (b_n - a_n)^2 \neq 0$, and let $X_n : \Omega \rightarrow [a_n, b_n]$, $n \in \{1, 2, \dots, N\}$, be independent random variables. Then

$$\mathbb{P}\left(\left| \sum_{n=1}^N (X_n - \mathbb{E}[X_n]) \right| \geq \varepsilon\right) \leq \min\left\{1, 2 \exp\left(\frac{-2\varepsilon^2}{\sum_{n=1}^N (b_n - a_n)^2}\right), \frac{\sum_{n=1}^N (b_n - a_n)^2}{4\varepsilon^2}\right\}. \quad (12.67)$$

Proof of Corollary 12.1.23. Observe that Lemma 12.1.6, Corollary 12.1.20, and the fact that for all $B \in \mathcal{F}$ it holds that $\mathbb{P}(B) \leq 1$ establish (12.67). The proof of Corollary 12.1.23 is thus complete. \square

12.2 Covering number estimates

12.2.1 Entropy quantities

12.2.1.1 Covering radii (Outer entropy numbers)

Definition 12.2.1 (Covering radii). Let (X, d) be a metric space and let $n \in \mathbb{N}$. Then we denote by $\mathcal{C}_{(X,d),n} \in [0, \infty]$ (we denote by $\mathcal{C}_{X,n} \in [0, \infty]$) the extended real number given by

$$\mathcal{C}_{(X,d),n} = \inf \left(\left\{ r \in [0, \infty] : (\exists A \subseteq X : [|A| \leq n] \wedge (\forall x \in X : \exists a \in A : d(a, x) \leq r)) \right\} \right) \quad (12.68)$$

and we call $\mathcal{C}_{(X,d),n}$ the n -covering radius of (X, d) (we call $\mathcal{C}_{X,r}$ the n -covering radius of X).

Lemma 12.2.2. Let (X, d) be a metric space and let $n \in \mathbb{N}$. Then

$$\mathcal{C}_{(X,d),n} = \begin{cases} 0 & : X = \emptyset \\ \inf \left(\left\{ r \in [0, \infty) : \left(\exists x_1, x_2, \dots, x_n \in X : X \subseteq \left[\bigcup_{m=1}^n \{v \in X : d(x_m, v) \leq r\} \right] \right) \right\} \cup \{\infty\} \right) & : X \neq \emptyset \end{cases} \quad (12.69)$$

(cf. Definition 12.2.1).

Proof of Lemma 12.2.2. Throughout this proof, assume without loss of generality that $X \neq \emptyset$ and let $a \in X$. Note that the assumption that d is a metric implies that for all $x \in X$ it holds that $d(a, x) \leq \infty$. Combining this with Lemma 4.3.5 proves (12.69). This completes the proof of Lemma 12.2.2. \square

Exercise 12.2.1. Prove or disprove the following statement: For every metric space (X, d) and every $n, m \in \mathbb{N}$ it holds that $\mathcal{C}_{(X,d),n} < \infty$ if and only if $\mathcal{C}_{(X,d),m} < \infty$ (cf. Definition 12.2.1).

Exercise 12.2.2. Prove or disprove the following statement: For every metric space (X, d) and every $n \in \mathbb{N}$ it holds that (X, d) is bounded if and only if $\mathcal{C}_{(X,d),n} < \infty$ (cf. Definition 12.2.1).

Exercise 12.2.3. Prove or disprove the following statement: For every $n \in \mathbb{N}$ and every metric space (X, d) with $X \neq \emptyset$ it holds that

$$\begin{aligned} \mathcal{C}_{(X,d),n} &= \inf_{x_1, x_2, \dots, x_n \in X} \sup_{v \in X} \min_{i \in \{1, 2, \dots, n\}} d(x_i, v) \\ &= \inf_{x_1, x_2, \dots, x_n \in X} \sup_{x_{n+1} \in X} \min_{i \in \{1, 2, \dots, n\}} d(x_i, x_{n+1}) \end{aligned} \quad (12.70)$$

(cf. Definition 12.2.1).

12.2.1.2 Packing radii (Inner entropy numbers)

Definition 12.2.3 (Packing radii). Let (X, d) be a metric space and let $n \in \mathbb{N}$. Then we denote by $\mathcal{P}_{(X,d),n} \in [0, \infty]$ (we denote by $\mathcal{P}_{X,n} \in [0, \infty]$) the extended real number given by

$$\mathcal{P}_{(X,d),n} = \sup(\{r \in [0, \infty) : (\exists x_1, x_2, \dots, x_{n+1} \in X : [\min_{i,j \in \{1,2,\dots,n+1\}, i \neq j} d(x_i, x_j)] > 2r)\} \cup \{0\}) \quad (12.71)$$

and we call $\mathcal{P}_{(X,d),n}$ the n -packing radius of (X, d) (we call $\mathcal{P}^{X,r}$ the n -packing radius of X).

Exercise 12.2.4. Prove or disprove the following statement: For every $n \in \mathbb{N}$ and every metric space (X, d) with $X \neq \emptyset$ it holds that

$$\mathcal{P}_{(X,d),n} = \frac{1}{2} [\sup_{x_1, x_2, \dots, x_{n+1} \in X} \min_{i,j \in \{1,2,\dots,n+1\}, i \neq j} d(x_i, x_j)] \quad (12.72)$$

(cf. Definition 12.2.3).

12.2.1.3 Packing numbers

Definition 12.2.4 (Packing numbers). Let (X, d) be a metric space and let $r \in [0, \infty]$. Then we denote by $\mathcal{P}^{(X,d),r} \in [0, \infty]$ (we denote by $\mathcal{P}^{X,r} \in [0, \infty]$) the extended real number given by

$$\mathcal{P}^{(X,d),r} = \sup(\{n \in \mathbb{N} : (\exists x_1, x_2, \dots, x_{n+1} \in X : [\min_{i,j \in \{1,2,\dots,n+1\}, i \neq j} d(x_i, x_j)] > 2r)\} \cup \{0\}) \quad (12.73)$$

and we call $\mathcal{P}^{(X,d),r}$ the r -packing number of (X, d) (we call $\mathcal{P}^{X,r}$ the r -packing number of X).

12.2.2 Inequalities for packing entropy quantities in metric spaces

12.2.2.1 Lower bounds for packing radii based on lower bounds for packing numbers

Lemma 12.2.5 (Lower bounds for packing radii). Let (X, d) be a metric space and let $n \in \mathbb{N}$, $r \in [0, \infty]$ satisfy $n \leq \mathcal{P}^{(X,d),r}$ (cf. Definition 12.2.4). Then $r \leq \mathcal{P}_{(X,d),n}$ (cf. Definition 12.2.3).

Proof of Lemma 12.2.5. Note that (12.73) ensures that there exist $x_1, x_2, \dots, x_{n+1} \in X$ such that

$$[\min_{i,j \in \{1,2,\dots,n+1\}, i \neq j} d(x_i, x_j)] > 2r. \quad (12.74)$$

This implies that $\mathcal{P}_{(X,d),n} \geq r$ (cf. Definition 12.2.3). The proof of Lemma 12.2.5 is thus complete. \square

12.2.2.2 Upper bounds for packing numbers based on upper bounds for packing radii

Lemma 12.2.6. Let (X, d) be a metric space and let $n \in \mathbb{N}$, $r \in [0, \infty]$ satisfy $\mathcal{P}_{(X,d),n} < r$ (cf. Definition 12.2.3). Then $\mathcal{P}^{(X,d),r} < n$ (cf. Definition 12.2.4).

Proof of Lemma 12.2.6. Observe that Lemma 12.2.5 establishes that $\mathcal{P}^{(X,d),r} < n$ (cf. Definition 12.2.4). The proof of Lemma 12.2.6 is thus complete. \square

12.2.2.3 Upper bounds for packing radii based on upper bounds for covering radii

Lemma 12.2.7. Let (X, d) be a metric space and let $n \in \mathbb{N}$. Then $\mathcal{P}_{(X,d),n} \leq \mathcal{C}_{(X,d),n}$ (cf. Definitions 12.2.1 and 12.2.3).

Proof of Lemma 12.2.7. Throughout this proof, assume without loss of generality that $\mathcal{C}_{(X,d),n} < \infty$ and $\mathcal{P}_{(X,d),n} > 0$, let $r \in [0, \infty)$, $x_1, x_2, \dots, x_n \in X$ satisfy

$$X \subseteq \left[\bigcup_{m=1}^n \{v \in X : d(x_m, v) \leq r\} \right], \quad (12.75)$$

let $\mathbf{r} \in [0, \infty)$, $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{n+1} \in X$ satisfy

$$[\min_{i,j \in \{1,2,\dots,n+1\}, i \neq j} d(\mathbf{x}_i, \mathbf{x}_j)] > 2\mathbf{r}, \quad (12.76)$$

and let $\varphi: X \rightarrow \{1, 2, \dots, n\}$ satisfy for all $v \in X$ that

$$\varphi(v) = \min\{m \in \{1, 2, \dots, n\} : v \in \{w \in X : d(x_m, w) \leq r\}\} \quad (12.77)$$

(cf. Definitions 12.2.1 and 12.2.3 and Lemma 12.2.2). Observe that (12.77) shows that for all $v \in X$ it holds that

$$v \in \{w \in X : d(x_{\varphi(v)}, w) \leq r\}. \quad (12.78)$$

Hence, we obtain that for all $v \in X$ it holds that

$$d(v, x_{\varphi(v)}) \leq r \quad (12.79)$$

Moreover, note that the fact that $\varphi(\mathbf{x}_1), \varphi(\mathbf{x}_2), \dots, \varphi(\mathbf{x}_{n+1}) \in \{1, 2, \dots, n\}$ ensures that there exist $i, j \in \{1, 2, \dots, n+1\}$ which satisfy

$$i \neq j \quad \text{and} \quad \varphi(\mathbf{x}_i) = \varphi(\mathbf{x}_j). \quad (12.80)$$

The triangle inequality, (12.76), and (12.79) hence show that

$$2r < d(\mathbf{x}_i, \mathbf{x}_j) \leq d(\mathbf{x}_i, x_{\varphi(\mathbf{x}_i)}) + d(x_{\varphi(\mathbf{x}_i)}, \mathbf{x}_j) = d(\mathbf{x}_i, x_{\varphi(\mathbf{x}_i)}) + d(\mathbf{x}_j, x_{\varphi(\mathbf{x}_j)}) \leq 2r. \quad (12.81)$$

This implies that $r < r$. The proof of Lemma 12.2.7 is thus complete. \square

12.2.2.4 Upper bounds for packing radii in balls of metric spaces

Lemma 12.2.8. *Let (X, d) be a metric space, let $n \in \mathbb{N}$, $x \in X$, $r \in (0, \infty]$, and let $S = \{v \in X : d(x, v) \leq r\}$. Then $\mathcal{P}_{(S, d|_{S \times S}), n} \leq r$ (cf. Definition 12.2.3).*

Proof of Lemma 12.2.8. Throughout this proof, assume without loss of generality that $\mathcal{P}_{(S, d|_{S \times S}), n} > 0$ (cf. Definition 12.2.3). Observe that for all $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{n+1} \in S$, $i, j \in \{1, 2, \dots, n+1\}$ it holds that

$$d(\mathbf{x}_i, \mathbf{x}_j) \leq d(\mathbf{x}_i, x) + d(x, \mathbf{x}_j) \leq 2r. \quad (12.82)$$

Hence, we obtain that for all $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{n+1} \in S$ it holds that

$$\min_{i,j \in \{1, 2, \dots, n+1\}, i \neq j} d(\mathbf{x}_i, \mathbf{x}_j) \leq 2r. \quad (12.83)$$

Moreover, note that (12.71) ensures that for all $\rho \in [0, \mathcal{P}_{(S, d|_{S \times S}), n}]$ there exist $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{n+1} \in S$ such that

$$\min_{i,j \in \{1, 2, \dots, n+1\}, i \neq j} d(\mathbf{x}_i, \mathbf{x}_j) > 2\rho. \quad (12.84)$$

This and (12.83) demonstrate that for all $\rho \in [0, \mathcal{P}_{(S, d|_{S \times S}), n}]$ it holds that $2\rho < 2r$. The proof of Lemma 12.2.8 is thus complete. \square

12.2.3 Inequalities for covering entropy quantities in metric spaces

12.2.3.1 Upper bounds for covering numbers based on upper bounds for covering radii

Lemma 12.2.9. *Let (X, d) be a metric space and let $r \in [0, \infty]$, $n \in \mathbb{N}$ satisfy $\mathcal{C}_{(X, d), n} < r$ (cf. Definition 12.2.1). Then $\mathcal{C}^{(X, d), r} \leq n$ (cf. Definition 4.3.2).*

Proof of Lemma 12.2.9. Observe that the assumption that $\mathcal{C}_{(X,d),n} < r$ ensures that there exists $A \subseteq X$ such that $|A| \leq n$ and

$$X \subseteq \left[\bigcup_{a \in A} \{v \in X : d(a, v) \leq r\} \right]. \quad (12.85)$$

This establishes that $\mathcal{C}^{(X,d),r} \leq n$ (cf. Definition 4.3.2). The proof of Lemma 12.2.9 is thus complete. \square

Lemma 12.2.10. *Let (X, d) be a compact metric space and let $r \in [0, \infty]$, $n \in \mathbb{N}$, satisfy $\mathcal{C}_{(X,d),n} \leq r$ (cf. Definition 12.2.1). Then $\mathcal{C}^{(X,d),r} \leq n$ (cf. Definition 4.3.2).*

Proof of Lemma 12.2.10. Throughout this proof, assume without loss of generality that $X \neq \emptyset$ and let $x_{k,m} \in X$, $m \in \{1, 2, \dots, n\}$, $k \in \mathbb{N}$, satisfy for all $k \in \mathbb{N}$ that

$$X \subseteq \left[\bigcup_{m=1}^n \{v \in X : d(x_{k,m}, v) \leq r + \frac{1}{k}\} \right] \quad (12.86)$$

(cf. Lemma 4.3.5). Note that the assumption that (X, d) is a compact metric space demonstrates that there exist $\mathfrak{x} = (\mathfrak{x}_m)_{m \in \{1, 2, \dots, n\}} : \{1, 2, \dots, n\} \rightarrow X$ and $k = (k_l)_{l \in \mathbb{N}} : \mathbb{N} \rightarrow \mathbb{N}$ which satisfy that

$$\limsup_{l \rightarrow \infty} \max_{m \in \{1, 2, \dots, n\}} d(\mathfrak{x}_m, x_{k_l, m}) = 0 \quad \text{and} \quad \limsup_{l \rightarrow \infty} k_l = \infty. \quad (12.87)$$

Next observe that the assumption that d is a metric ensures that for all $v \in X$, $m \in \{1, 2, \dots, n\}$, $l \in \mathbb{N}$ it holds that

$$d(v, \mathfrak{x}_m) \leq d(v, x_{k_l, m}) + d(x_{k_l, m}, \mathfrak{x}_m). \quad (12.88)$$

This and (12.86) prove that for all $v \in X$, $l \in \mathbb{N}$ it holds that

$$\begin{aligned} \min_{m \in \{1, 2, \dots, n\}} d(v, \mathfrak{x}_m) &\leq \min_{m \in \{1, 2, \dots, n\}} [d(v, x_{k_l, m}) + d(x_{k_l, m}, \mathfrak{x}_m)] \\ &\leq [\min_{m \in \{1, 2, \dots, n\}} d(v, x_{k_l, m})] + [\max_{m \in \{1, 2, \dots, n\}} d(x_{k_l, m}, \mathfrak{x}_m)] \\ &\leq [r + \frac{1}{k_l}] + [\max_{m \in \{1, 2, \dots, n\}} d(x_{k_l, m}, \mathfrak{x}_m)]. \end{aligned} \quad (12.89)$$

Hence, we obtain for all $v \in X$ that

$$\min_{m \in \{1, 2, \dots, n\}} d(v, \mathfrak{x}_m) \leq \limsup_{l \rightarrow \infty} ([r + \frac{1}{k_l}] + [\max_{m \in \{1, 2, \dots, n\}} d(x_{k_l, m}, \mathfrak{x}_m)]) = r. \quad (12.90)$$

This establishes that $\mathcal{C}^{(X,d),r} \leq n$ (cf. Definition 4.3.2). The proof of Lemma 12.2.10 is thus complete. \square

12.2.3.2 Upper bounds for covering radii based on upper bounds for covering numbers

Lemma 12.2.11. Let (X, d) be a metric space and let $r \in [0, \infty]$, $n \in \mathbb{N}$ satisfy $\mathcal{C}^{(X,d),r} \leq n$ (cf. Definition 4.3.2). Then $\mathcal{C}_{(X,d),n} \leq r$ (cf. Definition 12.2.1).

Proof of Lemma 12.2.11. Observe that the assumption that $\mathcal{C}^{(X,d),r} \leq n$ ensures that there exists $A \subseteq X$ such that $|A| \leq n$ and

$$X \subseteq \left[\bigcup_{a \in A} \{v \in X : d(a, v) \leq r\} \right]. \quad (12.91)$$

This establishes that $\mathcal{C}_{(X,d),n} \leq r$ (cf. Definition 12.2.1). The proof of Lemma 12.2.11 is thus complete. \square

12.2.3.3 Upper bounds for covering radii based on upper bounds for packing radii

Lemma 12.2.12. Let (X, d) be a metric space and let $n \in \mathbb{N}$. Then $\mathcal{C}_{(X,d),n} \leq 2\mathcal{P}_{(X,d),n}$ (cf. Definitions 12.2.1 and 12.2.3).

Proof of Lemma 12.2.12. Throughout this proof, assume w.l.o.g. that $X \neq \emptyset$, assume without loss of generality that $\mathcal{P}_{(X,d),n} < \infty$, let $r \in [0, \infty]$ satisfy $r > \mathcal{P}_{(X,d),n}$, and let $N \in \mathbb{N}_0 \cup \{\infty\}$ satisfy $N = \mathcal{P}^{(X,d),r}$ (cf. Definitions 12.2.3 and 12.2.4). Observe that Lemma 12.2.6 ensures that

$$N = \mathcal{P}^{(X,d),r} < n. \quad (12.92)$$

Moreover, note that the fact that $N = \mathcal{P}^{(X,d),r}$ and (12.73) demonstrate that for all $x_1, x_2, \dots, x_{N+1}, x_{N+2} \in X$ it holds that

$$\min_{i,j \in \{1, 2, \dots, N+2\}, i \neq j} d(x_i, x_j) \leq 2r. \quad (12.93)$$

In addition, observe that the fact that $N = \mathcal{P}^{(X,d),r}$ and (12.73) imply that there exist $x_1, x_2, \dots, x_{N+1} \in X$ which satisfy that

$$\min(\{d(x_i, x_j) : i, j \in \{1, 2, \dots, N+1\}, i \neq j\} \cup \{\infty\}) > 2r. \quad (12.94)$$

Combining this with (12.93) establishes that for all $v \in X$ it holds that

$$\min_{i \in \{1, 2, \dots, N\}} d(x_i, v) \leq 2r. \quad (12.95)$$

Hence, we obtain that for all $w \in X$ it holds that

$$w \in \left[\bigcup_{m=1}^n \{v \in X : d(x_m, v) \leq 2r\} \right]. \quad (12.96)$$

Therefore, we obtain that

$$X \subseteq \left[\bigcup_{m=1}^n \{v \in X : d(x_i, v) \leq 2r\} \right]. \quad (12.97)$$

Combining this and Lemma 12.2.2 shows that $\mathcal{C}_{(X,d),n} \leq 2r$ (cf. Definition 12.2.1). The proof of Lemma 12.2.12 is thus complete. \square

12.2.3.4 Equivalence of covering and packing radii

Corollary 12.2.13. Let (X, d) be a metric space and let $n \in \mathbb{N}$. Then $\mathcal{P}_{(X,d),n} \leq \mathcal{C}_{(X,d),n} \leq 2\mathcal{P}_{(X,d),n}$ (cf. Definitions 12.2.1 and 12.2.3).

Proof of Corollary 12.2.13. Observe that Lemma 12.2.7 and Lemma 12.2.12 establish that $\mathcal{P}_{(X,d),n} \leq \mathcal{C}_{(X,d),n} \leq 2\mathcal{P}_{(X,d),n}$ (cf. Definitions 12.2.1 and 12.2.3). The proof of Corollary 12.2.13 is thus complete. \square

12.2.4 Inequalities for entropy quantities in finite-dimensional vector spaces

12.2.4.1 Measures induced by Lebesgue–Borel measures

Lemma 12.2.14. Let $(V, \|\cdot\|)$ be a normed vector space, let $N \in \mathbb{N}$, let $b_1, b_2, \dots, b_N \in V$ be a Hamel-basis of V , let $\lambda: \mathcal{B}(\mathbb{R}^N) \rightarrow [0, \infty]$ be the Lebesgue–Borel measure on \mathbb{R}^N , let $\Phi: \mathbb{R}^N \rightarrow V$ satisfy for all $r = (r_1, r_2, \dots, r_N) \in \mathbb{R}^N$ that $\Phi(r) = r_1 b_1 + r_2 b_2 + \dots + r_N b_N$, and let $\nu: \mathcal{B}(V) \rightarrow [0, \infty]$ satisfy for all $A \in \mathcal{B}(V)$ that

$$\nu(A) = \lambda(\Phi^{-1}(A)). \quad (12.98)$$

Then

- (i) it holds that Φ is linear,
- (ii) it holds for all $r = (r_1, r_2, \dots, r_N) \in \mathbb{R}^N$ that $\|\Phi(r)\| \leq [\sum_{n=1}^N \|b_n\|^2]^{1/2} [\sum_{n=1}^N |r_n|^2]^{1/2}$,
- (iii) it holds that $\Phi \in C(\mathbb{R}^N, V)$,
- (iv) it holds that Φ is bijective,
- (v) it holds that $(V, \mathcal{B}(V), \nu)$ is a measure space,
- (vi) it holds for all $r \in (0, \infty)$, $v \in V$, $A \in \mathcal{B}(V)$ that $\nu(\{(ra + v) \in V : a \in A\}) = r^N \nu(A)$,

- (vii) it holds for all $r \in (0, \infty)$ that $\nu(\{v \in V: \|v\| \leq r\}) = r^N \nu(\{v \in V: \|v\| \leq 1\})$, and
- (viii) it holds that $\nu(\{v \in V: \|v\| \leq 1\}) > 0$.

Proof of Lemma 12.2.14. Note that for all $r = (r_1, r_2, \dots, r_N)$, $s = (s_1, s_2, \dots, s_N) \in \mathbb{R}^N$, $\rho \in \mathbb{R}$ it holds that

$$\Phi(\rho r + s) = (\rho r_1 + s_1)b_1 + (\rho r_2 + s_2)b_2 + \dots + (\rho r_N + s_N)b_N = \rho\Phi(r) + \Phi(s). \quad (12.99)$$

This establishes item (i). Next observe that Hölder's inequality shows that for all $r = (r_1, r_2, \dots, r_N) \in \mathbb{R}^N$ it holds that

$$\|\Phi(r)\| = \|r_1 b_1 + r_2 b_2 + \dots + r_N b_N\| \leq \sum_{n=1}^N |r_n| \|b_n\| \leq \left[\sum_{n=1}^N \|b_n\|^2 \right]^{1/2} \left[\sum_{n=1}^N |r_n|^2 \right]^{1/2}. \quad (12.100)$$

This establishes item (ii). Moreover, note that item (ii) proves item (iii). Furthermore, observe that the assumption that $b_1, b_2, \dots, b_N \in V$ is a Hamel-basis of V establishes item (iv). Next note that (12.98) and item (iii) prove item (v). In addition, observe that the integral transformation theorem shows that for all $r \in (0, \infty)$, $v \in \mathbb{R}^N$, $A \in \mathcal{B}(\mathbb{R}^N)$ it holds that

$$\begin{aligned} \lambda(\{(ra + v) \in \mathbb{R}^N : a \in A\}) &= \lambda(\{ra \in \mathbb{R}^N : a \in A\}) = \int_{\mathbb{R}^N} \mathbb{1}_{\{ra \in \mathbb{R}^N : a \in A\}}(x) dx \\ &= \int_{\mathbb{R}^N} \mathbb{1}_A\left(\frac{x}{r}\right) dx = r^N \int_{\mathbb{R}^N} \mathbb{1}_A(x) dx = r^N \lambda(A). \end{aligned} \quad (12.101)$$

Combining item (i) and item (iv) hence demonstrates that for all $r \in (0, \infty)$, $v \in V$, $A \in \mathcal{B}(V)$ it holds that

$$\begin{aligned} \nu(\{(ra + v) \in V : a \in A\}) &= \lambda(\Phi^{-1}(\{(ra + v) \in V : a \in A\})) \\ &= \lambda(\{\Phi^{-1}(ra + v) \in \mathbb{R}^N : a \in A\}) \\ &= \lambda(\{\left[r\Phi^{-1}(a) + \Phi^{-1}(v)\right] \in \mathbb{R}^N : a \in A\}) \\ &= \lambda(\{\left[ra + \Phi^{-1}(v)\right] \in \mathbb{R}^N : a \in \Phi^{-1}(A)\}) \\ &= r^N \lambda(\Phi^{-1}(A)) = r^N \nu(A). \end{aligned} \quad (12.102)$$

This establishes item (vi). Hence, we obtain that for all $r \in (0, \infty)$ it holds that

$$\begin{aligned} \nu(\{v \in V : \|v\| \leq r\}) &= \nu(\{rv \in V : \|v\| \leq 1\}) \\ &= r^N \nu(\{v \in V : \|v\| \leq 1\}) \\ &= r^N \nu(X). \end{aligned} \quad (12.103)$$

This establishes item (vii). Furthermore, observe that (12.103) demonstrates that

$$\begin{aligned}\infty = \lambda(\mathbb{R}^N) &= \nu(V) = \limsup_{r \rightarrow \infty} [\nu(\{v \in V : \|v\| \leq r\})] \\ &= \limsup_{r \rightarrow \infty} [r^N \nu(\{v \in V : \|v\| \leq 1\})].\end{aligned}\tag{12.104}$$

Hence, we obtain that $\nu(\{v \in V : \|v\| \leq 1\}) \neq 0$. This establishes item (viii). The proof of Lemma 12.2.14 is thus complete. \square

12.2.4.2 Upper bounds for packing radii

Lemma 12.2.15. Let $(V, \|\cdot\|)$ be a normed vector space, let $X = \{v \in V : \|v\| \leq 1\}$, let $d: X \times X \rightarrow [0, \infty)$ satisfy for all $v, w \in X$ that $d(v, w) = \|v - w\|$, and let $n, N \in \mathbb{N}$ satisfy $N = \dim(V)$. Then

$$\mathcal{P}_{(X,d),n} \leq 2(n+1)^{-1/N}\tag{12.105}$$

(cf. Definition 12.2.3).

Proof of Lemma 12.2.15. Throughout this proof, assume without loss of generality that $\mathcal{P}_{(X,d),n} > 0$, let $\rho \in [0, \mathcal{P}_{(X,d),n}]$, let $\lambda: \mathcal{B}(\mathbb{R}^N) \rightarrow [0, \infty]$ be the Lebesgue-Borel measure on \mathbb{R}^N , let $b_1, b_2, \dots, b_N \in V$ be a Hamel-basis of V , let $\Phi: \mathbb{R}^N \rightarrow V$ satisfy for all $r = (r_1, r_2, \dots, r_N) \in \mathbb{R}^N$ that

$$\Phi(r) = r_1 b_1 + r_2 b_2 + \dots + r_N b_N,\tag{12.106}$$

and let $\nu: \mathcal{B}(V) \rightarrow [0, \infty]$ satisfy for all $A \in \mathcal{B}(V)$ that

$$\nu(A) = \lambda(\Phi^{-1}(A))\tag{12.107}$$

(cf. Definition 12.2.3). Observe that Lemma 12.2.8 ensures that $\rho < \mathcal{P}_{(X,d),n} \leq 1$. Moreover, note that (12.71) shows that there exist $x_1, x_2, \dots, x_{n+1} \in X$ which satisfy

$$\min_{i,j \in \{1, 2, \dots, n+1\}, i \neq j} \|x_i - x_j\| = \min_{i,j \in \{1, 2, \dots, n+1\}, i \neq j} d(x_i, x_j) > 2\rho.\tag{12.108}$$

Observe that (12.108) ensures that for all $i, j \in \{1, 2, \dots, n+1\}$ with $i \neq j$ it holds that

$$\{v \in V : \|x_i - v\| \leq \rho\} \cap \{v \in V : \|x_j - v\| \leq \rho\} = \emptyset.\tag{12.109}$$

Moreover, note that (12.108) and the fact that $\rho < 1$ show that for all $j \in \{1, 2, \dots, n+1\}$, $w \in \{v \in X : d(x_j, v) \leq \rho\}$ it holds that

$$\|w\| \leq \|w - x_j\| + \|x_j\| \leq \rho + 1 \leq 2.\tag{12.110}$$

Therefore, we obtain that for all $j \in \{1, 2, \dots, n+1\}$ it holds that

$$\{v \in V : \|v - x_j\| \leq \rho\} \subseteq \{v \in V : \|v\| \leq 2\}. \quad (12.111)$$

Next observe that Lemma 12.2.14 ensures that $(V, \mathcal{B}(V), \nu)$ is a measure space. Combining this and (12.109) with (12.111) proves that

$$\begin{aligned} \sum_{j=1}^{n+1} \nu(\{v \in V : \|v - x_j\| \leq \rho\}) &= \nu\left(\bigcup_{j=1}^{n+1} \{v \in V : \|v - x_j\| \leq \rho\}\right) \\ &\leq \nu(\{v \in V : \|v\| \leq 2\}). \end{aligned} \quad (12.112)$$

Lemma 12.2.14 hence shows that

$$\begin{aligned} (n+1)\rho^N \nu(X) &= \sum_{j=1}^{n+1} [\rho^N \nu(\{v \in V : \|v\| \leq 1\})] \\ &= \sum_{j=1}^{n+1} \nu(\{v \in V : \|v\| \leq \rho\}) \\ &= \sum_{j=1}^{n+1} \nu(\{v \in V : \|v - x_j\| \leq \rho\}) \leq \nu(\{v \in V : \|v\| \leq 2\}) \\ &= 2^N \nu(\{v \in V : \|v\| \leq 1\}) = 2^N \nu(X). \end{aligned} \quad (12.113)$$

Next observe that Lemma 12.2.14 demonstrates that $\nu(X) > 0$. Combining this with (12.113) assures that $(n+1)\rho^N \leq 2^N$. Therefore, we obtain that $\rho^N \leq (n+1)^{-1}2^N$. Hence, we obtain that $\rho \leq 2(n+1)^{-1/N}$. The proof of Lemma 12.2.15 is thus complete. \square

12.2.4.3 Upper bounds for covering radii

Corollary 12.2.16. *Let $(V, \|\cdot\|)$ be a normed vector space, let $X = \{v \in V : \|v\| \leq 1\}$, let $d: X \times X \rightarrow [0, \infty)$ satisfy for all $v, w \in X$ that $d(v, w) = \|v - w\|$, and let $n, N \in \mathbb{N}$ satisfy $N = \dim(V)$. Then*

$$\mathcal{C}_{(X,d),n} \leq 4(n+1)^{-1/N} \quad (12.114)$$

(cf. Definition 12.2.1).

Proof of Corollary 12.2.16. Observe that Corollary 12.2.13 and Lemma 12.2.15 establish (12.114). The proof of Corollary 12.2.16 is thus complete. \square

12.2.4.4 Lower bounds for covering radii

Lemma 12.2.17. Let $(V, \|\cdot\|)$ be a normed vector space, let $X = \{v \in V : \|v\| \leq 1\}$, let $d: X \times X \rightarrow [0, \infty)$ satisfy for all $v, w \in X$ that $d(v, w) = \|v - w\|$, and let $n, N \in \mathbb{N}$ satisfy $N = \dim(V)$. Then

$$n^{-1/N} \leq \mathcal{C}_{(X,d),n} \quad (12.115)$$

(cf. Definition 12.2.1).

Proof of Lemma 12.2.17. Throughout this proof, assume without loss of generality that $\mathcal{C}_{(X,d),n} < \infty$, let $\rho \in (\mathcal{C}_{(X,d),n}, \infty)$, let $\lambda: \mathcal{B}(\mathbb{R}^N) \rightarrow [0, \infty]$ be the Lebesgue-Borel measure on \mathbb{R}^N , let $b_1, b_2, \dots, b_N \in V$ be a Hamel-basis of V , let $\Phi: \mathbb{R}^N \rightarrow V$ satisfy for all $r = (r_1, r_2, \dots, r_N) \in \mathbb{R}^N$ that

$$\Phi(r) = r_1 b_1 + r_2 b_2 + \dots + r_N b_N, \quad (12.116)$$

and let $\nu: \mathcal{B}(V) \rightarrow [0, \infty]$ satisfy for all $A \in \mathcal{B}(V)$ that

$$\nu(A) = \lambda(\Phi^{-1}(A)) \quad (12.117)$$

(cf. Definition 12.2.1). The fact that $\rho > \mathcal{C}_{(X,d),n}$ demonstrates that there exist $x_1, x_2, \dots, x_n \in X$ which satisfy

$$X \subseteq \left[\bigcup_{m=1}^n \{v \in X : d(x_m, v) \leq \rho\} \right]. \quad (12.118)$$

Lemma 12.2.14 hence shows that

$$\begin{aligned} \nu(X) &\leq \nu\left(\bigcup_{m=1}^n \{v \in X : d(x_m, v) \leq \rho\}\right) \leq \sum_{m=1}^n \nu(\{v \in X : d(x_m, v) \leq \rho\}) \\ &= \sum_{m=1}^n [\rho^N \nu(\{v \in X : d(x_m, v) \leq 1\})] \leq n\rho^N \nu(X). \end{aligned} \quad (12.119)$$

This and Lemma 12.2.14 demonstrate that $1 \leq n\rho^N$. Hence, we obtain that $\rho^N \geq n^{-1}$. This ensures that $\rho \geq n^{-1/N}$. The proof of Lemma 12.2.17 is thus complete. \square

12.2.4.5 Lower and upper bounds for covering radii

Corollary 12.2.18. Let $(V, \|\cdot\|)$ be a normed vector space, let $X = \{v \in V : \|v\| \leq 1\}$, let $d: X \times X \rightarrow [0, \infty)$ satisfy for all $v, w \in X$ that $d(v, w) = \|v - w\|$, and let $n, N \in \mathbb{N}$ satisfy $N = \dim(V)$. Then

$$n^{-1/N} \leq \mathcal{C}_{(X,d),n} \leq 4(n+1)^{-1/N} \quad (12.120)$$

(cf. Definition 12.2.1).

Proof of Corollary 12.2.18. Observe that Corollary 12.2.16 and Lemma 12.2.17 establish (12.120). The proof of Corollary 12.2.18 is thus complete. \square

12.2.4.6 Scaling property for covering radii

Lemma 12.2.19. Let $(V, \|\cdot\|)$ be a normed vector space, let $d: V \times V \rightarrow [0, \infty)$ satisfy for all $v, w \in V$ that $d(v, w) = \|v - w\|$, let $n \in \mathbb{N}$, $r \in (0, \infty)$, and let $X \subseteq V$ and $\mathfrak{X} \subseteq V$ satisfy $\mathfrak{X} = \{rv \in V : v \in X\}$. Then

$$\mathcal{C}_{(\mathfrak{X}, d|_{\mathfrak{X} \times \mathfrak{X}}), n} = r \mathcal{C}_{(X, d|_{X \times X}), n} \quad (12.121)$$

(cf. Definition 12.2.1).

Proof of Lemma 12.2.19. Throughout this proof, let $\Phi: V \rightarrow V$ satisfy for all $v \in V$ that $\Phi(v) = rv$. Observe that Exercise 12.2.3 shows that

$$\begin{aligned} r \mathcal{C}_{(X, d), n} &= r \left[\inf_{x_1, x_2, \dots, x_n \in X} \sup_{v \in X} \min_{i \in \{1, 2, \dots, n\}} d(x_i, v) \right] \\ &= \inf_{x_1, x_2, \dots, x_n \in X} \sup_{v \in X} \min_{i \in \{1, 2, \dots, n\}} \|rx_i - rv\| \\ &= \inf_{x_1, x_2, \dots, x_n \in X} \sup_{v \in X} \min_{i \in \{1, 2, \dots, n\}} \|\Phi(x_i) - \Phi(v)\| \\ &= \inf_{x_1, x_2, \dots, x_n \in X} \sup_{v \in X} \min_{i \in \{1, 2, \dots, n\}} d(\Phi(x_i), \Phi(v)) \\ &= \inf_{x_1, x_2, \dots, x_n \in X} \sup_{v \in \mathfrak{X}} \min_{i \in \{1, 2, \dots, n\}} d(\Phi(x_i), v) \\ &= \inf_{x_1, x_2, \dots, x_n \in \mathfrak{X}} \sup_{v \in \mathfrak{X}} \min_{i \in \{1, 2, \dots, n\}} d(x_i, v) = \mathcal{C}_{(\mathfrak{X}, d|_{\mathfrak{X} \times \mathfrak{X}}), n} \end{aligned} \quad (12.122)$$

(cf. Definition 12.2.1). This establishes (12.121). The proof of Lemma 12.2.19 is thus complete. \square

12.2.4.7 Upper bounds for covering numbers

Proposition 12.2.20. Let $(V, \|\cdot\|)$ be a normed vector space with $\dim(V) < \infty$, let $r, R \in (0, \infty)$, $X = \{v \in V : \|v\| \leq R\}$, and let $d: X \times X \rightarrow [0, \infty)$ satisfy for all $v, w \in X$ that $d(v, w) = \|v - w\|$. Then

$$\mathcal{C}^{(X, d), r} \leq \begin{cases} 1 & : r \geq R \\ \left[\frac{4R}{r} \right]^{\dim(V)} & : r < R \end{cases} \quad (12.123)$$

(cf. Definition 4.3.2).

Proof of Proposition 12.2.20. Throughout this proof, assume without loss of generality that $\dim(V) > 0$, assume without loss of generality that $r < R$, let $N \in \mathbb{N}$ satisfy $N = \dim(V)$,

let $n \in \mathbb{N}$ satisfy

$$n = \left\lceil \left[\frac{4R}{r} \right]^N - 1 \right\rceil, \quad (12.124)$$

let $\mathfrak{X} = \{v \in V : \|v\| \leq 1\}$, and let $\mathfrak{d} : \mathfrak{X} \times \mathfrak{X} \rightarrow [0, \infty)$ satisfy for all $v, w \in \mathfrak{X}$ that

$$\mathfrak{d}(v, w) = \|v - w\| \quad (12.125)$$

(cf. Definition 4.2.8). Observe that Corollary 12.2.16 proves that

$$\mathcal{C}_{(\mathfrak{X}, \mathfrak{d}), n} \leq 4(n+1)^{-1/N} \quad (12.126)$$

(cf. Definition 12.2.1). The fact that

$$n+1 = \left\lceil \left[\frac{4R}{r} \right]^N - 1 \right\rceil + 1 \geq \left\lceil \left[\frac{4R}{r} \right]^N - 1 \right\rceil + 1 = \left[\frac{4R}{r} \right]^N \quad (12.127)$$

therefore ensures that

$$\mathcal{C}_{(\mathfrak{X}, \mathfrak{d}), n} \leq 4(n+1)^{-1/N} \leq 4 \left[\left[\frac{4R}{r} \right]^N \right]^{-1/N} = 4 \left[\frac{4R}{r} \right]^{-1} = \frac{r}{R}. \quad (12.128)$$

This and Lemma 12.2.19 demonstrate that

$$\mathcal{C}_{(X, d), n} = R \mathcal{C}_{(\mathfrak{X}, \mathfrak{d}), n} \leq R \left[\frac{r}{R} \right] = r. \quad (12.129)$$

Lemma 12.2.10 hence ensures that

$$\mathcal{C}_{(X, d), r} \leq n \leq \left[\frac{4R}{r} \right]^N = \left[\frac{4R}{r} \right]^{\dim(V)} \quad (12.130)$$

(cf. Definition 4.3.2). The proof of Proposition 12.2.20 is thus complete. \square

Proposition 12.2.21. Let $d \in \mathbb{N}$, $a \in \mathbb{R}$, $b \in (a, \infty)$, $r \in (0, \infty)$ and let $\delta : ([a, b]^d) \times ([a, b]^d) \rightarrow [0, \infty)$ satisfy for all $x, y \in [a, b]^d$ that $\delta(x, y) = \|x - y\|_\infty$ (cf. Definition 3.3.4). Then

$$\mathcal{C}_{([a, b]^d, \delta), r} \leq \left(\lceil \frac{b-a}{2r} \rceil \right)^d \leq \begin{cases} 1 & : r \geq (b-a)/2 \\ \left(\frac{b-a}{r} \right)^d & : r < (b-a)/2 \end{cases} \quad (12.131)$$

(cf. Definitions 4.2.8 and 4.3.2).

Proof of Proposition 12.2.21. Throughout this proof, let $\mathfrak{N} \subseteq \mathbb{N}$ satisfy

$$\mathfrak{N} = \left\lceil \frac{b-a}{2r} \right\rceil, \quad (12.132)$$

for every $N \in \mathbb{N}$, $i \in \{1, 2, \dots, N\}$ let $g_{N,i} \in [a, b]$ be given by

$$g_{N,i} = a + \frac{(i-1/2)(b-a)}{N} \quad (12.133)$$

and let $A \subseteq [a, b]^d$ be given by

$$A = \{g_{\mathfrak{N},1}, g_{\mathfrak{N},2}, \dots, g_{\mathfrak{N},\mathfrak{N}}\}^d \quad (12.134)$$

(cf. Definition 4.2.8). 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

$$|x - g_{N,i}| = a + \frac{(i-1/2)(b-a)}{N} - x \leq a + \frac{(i-1/2)(b-a)}{N} - \left(a + \frac{(i-1)(b-a)}{N}\right) = \frac{b-a}{2N}. \quad (12.135)$$

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

$$|x - g_{N,i}| = x - \left(a + \frac{(i-1/2)(b-a)}{N}\right) \leq a + \frac{i(b-a)}{N} - \left(a + \frac{(i-1/2)(b-a)}{N}\right) = \frac{b-a}{2N}. \quad (12.136)$$

Combining this with (12.135) 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

$$|x - y| \leq \frac{b-a}{2N}. \quad (12.137)$$

This shows that for every $x = (x_1, x_2, \dots, x_d) \in [a, b]^d$ there exists $y = (y_1, y_2, \dots, y_d) \in A$ such that

$$\delta(x, y) = \|x - y\|_\infty = \max_{i \in \{1, 2, \dots, d\}} |x_i - y_i| \leq \frac{b-a}{2\mathfrak{N}} \leq \frac{(b-a)2r}{2(b-a)} = r. \quad (12.138)$$

Combining this with (4.93), (12.134), (12.132), and the fact that $\forall x \in [0, \infty]: \lceil x \rceil \leq \mathbb{1}_{(0,r]}(rx) + 2x\mathbb{1}_{(r,\infty)}(rx)$ demonstrates that

$$\mathcal{C}^{([a,b]^d, \delta), r} \leq |A| = (\mathfrak{N})^d = \left(\left\lceil \frac{b-a}{2r} \right\rceil\right)^d \leq \mathbb{1}_{(0,r]} \left(\frac{b-a}{2}\right) + \left(\frac{b-a}{r}\right)^d \mathbb{1}_{(r,\infty)} \left(\frac{b-a}{2}\right) \quad (12.139)$$

(cf. Definition 4.3.2). The proof of Proposition 12.2.21 is thus complete. \square

12.3 Empirical risk minimization

12.3.1 Concentration inequalities for random fields

Lemma 12.3.1. Let (E, d) be a separable metric space and let $F \subseteq E$ be a set. Then

$$(F, d|_{F \times F}) \quad (12.140)$$

is a separable metric space.

Proof of Lemma 12.3.1. Throughout this proof, assume without loss of generality that $F \neq \emptyset$, let $e = (e_n)_{n \in \mathbb{N}}: \mathbb{N} \rightarrow E$ be a sequence of elements in E such that $\{e_n \in E: n \in \mathbb{N}\}$ is dense in E , and let $f = (f_n)_{n \in \mathbb{N}}: \mathbb{N} \rightarrow F$ be a sequence of elements in F such that for all $n \in \mathbb{N}$ it holds that

$$d(f_n, e_n) \leq \begin{cases} 0 & : e_n \in F \\ [\inf_{x \in F} d(x, e_n)] + \frac{1}{2^n} & : e_n \notin F. \end{cases} \quad (12.141)$$

Observe that for all $v \in F \setminus \{e_m \in E: m \in \mathbb{N}\}$, $n \in \mathbb{N}$ it holds that

$$\begin{aligned} \inf_{m \in \mathbb{N}} d(v, f_m) &\leq \inf_{m \in \mathbb{N} \cap [n, \infty)} d(v, f_m) \\ &\leq \inf_{m \in \mathbb{N} \cap [n, \infty)} [d(v, e_m) + d(e_m, f_m)] \\ &\leq \inf_{m \in \mathbb{N} \cap [n, \infty)} \left[d(v, e_m) + [\inf_{x \in F} d(x, e_m)] + \frac{1}{2^m} \right] \\ &\leq \inf_{m \in \mathbb{N} \cap [n, \infty)} \left[2d(v, e_m) + \frac{1}{2^m} \right] \\ &\leq 2 \left[\inf_{m \in \mathbb{N} \cap [n, \infty)} d(v, e_m) \right] + \frac{1}{2^n} = \frac{1}{2^n}. \end{aligned} \quad (12.142)$$

Combining this with the fact that for all $v \in F \cap \{e_m \in E: m \in \mathbb{N}\}$ it holds that $\inf_{m \in \mathbb{N}} d(v, f_m) = 0$ ensures that the set $\{f_n \in F: n \in \mathbb{N}\}$ is dense in F . The proof of Lemma 12.3.1 is thus complete. \square

Lemma 12.3.2. Let (E, \mathcal{E}) be a topological space, assume $E \neq \emptyset$, let $\mathbf{E} \subseteq E$ be an at most countable set, assume that \mathbf{E} is dense in E , let (Ω, \mathcal{F}) be a measurable space, for every $x \in E$ let $f_x: \Omega \rightarrow \mathbb{R}$ be $\mathcal{F}/\mathcal{B}(\mathbb{R})$ -measurable, assume for all $\omega \in \Omega$ that $E \ni x \mapsto f_x(\omega) \in \mathbb{R}$ is continuous, and let $F: \Omega \rightarrow \mathbb{R} \cup \{\infty\}$ satisfy for all $\omega \in \Omega$ that

$$F(\omega) = \sup_{x \in E} f_x(\omega). \quad (12.143)$$

Then

- (i) it holds for all $\omega \in \Omega$ that $F(\omega) = \sup_{x \in \mathbf{E}} f_x(\omega)$ and

(ii) it holds that F is $\mathcal{F}/\mathcal{B}(\mathbb{R} \cup \{\infty\})$ -measurable.

Proof of Lemma 12.3.2. Observe that the assumption that \mathbf{E} is dense in E shows that for all $g \in C(E, \mathbb{R})$ it holds that

$$\sup_{x \in E} g(x) = \sup_{x \in \mathbf{E}} g(x). \quad (12.144)$$

This and the assumption that for all $\omega \in \Omega$ it holds that $E \ni x \mapsto f_x(\omega) \in \mathbb{R}$ is continuous demonstrate that for all $\omega \in \Omega$ it holds that

$$F(\omega) = \sup_{x \in E} f_x(\omega) = \sup_{x \in \mathbf{E}} f_x(\omega). \quad (12.145)$$

This establishes item (i). Furthermore, note that item (i) and the assumption that for all $x \in E$ it holds that $f_x: \Omega \rightarrow \mathbb{R}$ is $\mathcal{F}/\mathcal{B}(\mathbb{R})$ -measurable prove item (ii). The proof of Lemma 12.3.2 is thus complete. \square

Lemma 12.3.3. Let (E, δ) be a separable metric space, let $\varepsilon, L \in \mathbb{R}$, $N \in \mathbb{N}$, $z_1, z_2, \dots, z_N \in E$ satisfy $E \subseteq \bigcup_{i=1}^N \{x \in E : 2L\delta(x, z_i) \leq \varepsilon\}$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $Z_x: \Omega \rightarrow \mathbb{R}$, $x \in E$, be random variables which satisfy for all $x, y \in E$ that $|Z_x - Z_y| \leq L\delta(x, y)$. Then

$$\mathbb{P}(\sup_{x \in E} |Z_x| \geq \varepsilon) \leq \sum_{i=1}^N \mathbb{P}(|Z_{z_i}| \geq \frac{\varepsilon}{2}) \quad (12.146)$$

(cf. Lemma 12.3.2).

Proof of Lemma 12.3.3. Throughout this proof, let $B_1, B_2, \dots, B_N \subseteq E$ satisfy for all $i \in \{1, 2, \dots, N\}$ that $B_i = \{x \in E : 2L\delta(x, z_i) \leq \varepsilon\}$. Observe that the triangle inequality and the assumption that for all $x, y \in E$ it holds that $|Z_x - Z_y| \leq L\delta(x, y)$ show that for all $i \in \{1, 2, \dots, N\}$, $x \in B_i$ it holds that

$$|Z_x| = |Z_x - Z_{z_i} + Z_{z_i}| \leq |Z_x - Z_{z_i}| + |Z_{z_i}| \leq L\delta(x, z_i) + |Z_{z_i}| \leq \frac{\varepsilon}{2} + |Z_{z_i}|. \quad (12.147)$$

Combining this with Lemma 12.3.2 and Lemma 12.3.1 proves that for all $i \in \{1, 2, \dots, N\}$ it holds that

$$\mathbb{P}(\sup_{x \in B_i} |Z_x| \geq \varepsilon) \leq \mathbb{P}\left(\frac{\varepsilon}{2} + |Z_{z_i}| \geq \varepsilon\right) = \mathbb{P}(|Z_{z_i}| \geq \frac{\varepsilon}{2}). \quad (12.148)$$

This, Lemma 12.3.2, and Lemma 12.3.1 establish that

$$\begin{aligned} \mathbb{P}(\sup_{x \in E} |Z_x| \geq \varepsilon) &= \mathbb{P}\left(\sup_{x \in (\bigcup_{i=1}^N B_i)} |Z_x| \geq \varepsilon\right) = \mathbb{P}\left(\bigcup_{i=1}^N \{\sup_{x \in B_i} |Z_x| \geq \varepsilon\}\right) \\ &\leq \sum_{i=1}^N \mathbb{P}(\sup_{x \in B_i} |Z_x| \geq \varepsilon) \leq \sum_{i=1}^N \mathbb{P}(|Z_{z_i}| \geq \frac{\varepsilon}{2}). \end{aligned} \quad (12.149)$$

This completes the proof of Lemma 12.3.3. \square

Lemma 12.3.4. Let (E, δ) be a separable metric space, assume $E \neq \emptyset$, let $\varepsilon, L \in (0, \infty)$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $Z_x: \Omega \rightarrow \mathbb{R}$, $x \in E$, be random variables which satisfy for all $x, y \in E$ that $|Z_x - Z_y| \leq L\delta(x, y)$. Then

$$[\mathcal{C}^{(E, \delta), \frac{\varepsilon}{2L}}]^{-1} \mathbb{P}(\sup_{x \in E} |Z_x| \geq \varepsilon) \leq \sup_{x \in E} \mathbb{P}(|Z_x| \geq \frac{\varepsilon}{2}). \quad (12.150)$$

(cf. Definition 4.3.2 and Lemma 12.3.2).

Proof of Lemma 12.3.4. Throughout this proof, let $N \in \mathbb{N} \cup \{\infty\}$ satisfy $N = \mathcal{C}^{(E, \delta), \frac{\varepsilon}{2L}}$, assume without loss of generality that $N < \infty$, and let $z_1, z_2, \dots, z_N \in E$ satisfy $E \subseteq \bigcup_{i=1}^N \{x \in E : \delta(x, z_i) \leq \frac{\varepsilon}{2L}\}$ (cf. Definition 4.3.2). Observe that Lemma 12.3.2 and Lemma 12.3.3 establish that

$$\mathbb{P}(\sup_{x \in E} |Z_x| \geq \varepsilon) \leq \sum_{i=1}^N \mathbb{P}(|Z_{z_i}| \geq \frac{\varepsilon}{2}) \leq N \left[\sup_{x \in E} \mathbb{P}(|Z_x| \geq \frac{\varepsilon}{2}) \right]. \quad (12.151)$$

This completes the proof of Lemma 12.3.4. \square

Lemma 12.3.5. Let (E, δ) be a separable metric space, assume $E \neq \emptyset$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $L \in \mathbb{R}$, for every $x \in E$ let $Z_x: \Omega \rightarrow \mathbb{R}$ be a random variable with $\mathbb{E}[|Z_x|] < \infty$, and assume for all $x, y \in E$ that $|Z_x - Z_y| \leq L\delta(x, y)$. Then

(i) it holds for all $x, y \in E$, $\eta \in \Omega$ that

$$|(Z_x(\eta) - \mathbb{E}[Z_x]) - (Z_y(\eta) - \mathbb{E}[Z_y])| \leq 2L\delta(x, y) \quad (12.152)$$

and

(ii) it holds that $\Omega \ni \eta \mapsto \sup_{x \in E} |Z_x(\eta) - \mathbb{E}[Z_x]| \in [0, \infty]$ is $\mathcal{F}/\mathcal{B}([0, \infty])$ -measurable.

Proof of Lemma 12.3.5. Observe that the assumption that for all $x, y \in E$ it holds that $|Z_x - Z_y| \leq L\delta(x, y)$ implies that for all $x, y \in E$, $\eta \in \Omega$ it holds that

$$\begin{aligned} |(Z_x(\eta) - \mathbb{E}[Z_x]) - (Z_y(\eta) - \mathbb{E}[Z_y])| &= |(Z_x(\eta) - Z_y(\eta)) + (\mathbb{E}[Z_y] - \mathbb{E}[Z_x])| \\ &\leq |Z_x(\eta) - Z_y(\eta)| + |\mathbb{E}[Z_y] - \mathbb{E}[Z_x]| \\ &\leq L\delta(x, y) + |\mathbb{E}[Z_x] - \mathbb{E}[Z_y]| \\ &= L\delta(x, y) + |\mathbb{E}[Z_x - Z_y]| \\ &\leq L\delta(x, y) + \mathbb{E}[|Z_x - Z_y|] \\ &\leq L\delta(x, y) + L\delta(x, y) = 2L\delta(x, y). \end{aligned} \quad (12.153)$$

This ensures item (i). Note that item (i) shows that for all $\eta \in \Omega$ it holds that $E \ni x \mapsto |Z_x(\eta) - \mathbb{E}[Z_x]| \in \mathbb{R}$ is continuous. Combining this and the assumption that E is separable with Lemma 12.3.2 establishes item (ii). The proof of Lemma 12.3.5 is thus complete. \square

Lemma 12.3.6. Let (E, δ) be a separable metric space, assume $E \neq \emptyset$, let $\varepsilon, L \in (0, \infty)$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $Z_x: \Omega \rightarrow \mathbb{R}$, $x \in E$, be random variables which satisfy for all $x, y \in E$ that $\mathbb{E}[|Z_x|] < \infty$ and $|Z_x - Z_y| \leq L\delta(x, y)$. Then

$$[\mathcal{C}^{(E, \delta), \frac{\varepsilon}{4L}}]^{-1} \mathbb{P}(\sup_{x \in E} |Z_x - \mathbb{E}[Z_x]| \geq \varepsilon) \leq \sup_{x \in E} \mathbb{P}(|Z_x - \mathbb{E}[Z_x]| \geq \frac{\varepsilon}{2}). \quad (12.154)$$

(cf. Definition 4.3.2 and Lemma 12.3.5).

Proof of Lemma 12.3.6. Throughout this proof, let $Y_x: \Omega \rightarrow \mathbb{R}$, $x \in E$, satisfy for all $x \in E$, $\eta \in \Omega$ that $Y_x(\eta) = Z_x(\eta) - \mathbb{E}[Z_x]$. Observe that Lemma 12.3.5 ensures that for all $x, y \in E$ it holds that

$$|Y_x - Y_y| \leq 2L\delta(x, y). \quad (12.155)$$

This and Lemma 12.3.4 (applied with $(E, \delta) \curvearrowleft (E, \delta)$, $\varepsilon \curvearrowleft \varepsilon$, $L \curvearrowleft 2L$, $(\Omega, \mathcal{F}, \mathbb{P}) \curvearrowleft (\Omega, \mathcal{F}, \mathbb{P})$, $(Z_x)_{x \in E} \curvearrowleft (Y_x)_{x \in E}$ in the notation of Lemma 12.3.4) establish (12.154). The proof of Lemma 12.3.6 is thus complete. \square

Lemma 12.3.7. Let (E, δ) be a separable metric space, assume $E \neq \emptyset$, let $M \in \mathbb{N}$, $\varepsilon, L, D \in (0, \infty)$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, for every $x \in E$ let $Y_{x,1}, Y_{x,2}, \dots, Y_{x,M}: \Omega \rightarrow [0, D]$ be independent random variables, assume for all $x, y \in E$, $m \in \{1, 2, \dots, M\}$ that $|Y_{x,m} - Y_{y,m}| \leq L\delta(x, y)$, and let $Z_x: \Omega \rightarrow [0, \infty)$, $x \in E$, satisfy for all $x \in E$ that

$$Z_x = \frac{1}{M} \left[\sum_{m=1}^M Y_{x,m} \right]. \quad (12.156)$$

Then

- (i) it holds for all $x \in E$ that $\mathbb{E}[|Z_x|] \leq D < \infty$,
- (ii) it holds that $\Omega \ni \eta \mapsto \sup_{x \in E} |Z_x(\eta) - \mathbb{E}[Z_x]| \in [0, \infty]$ is $\mathcal{F}/\mathcal{B}([0, \infty])$ -measurable, and
- (iii) it holds that

$$\mathbb{P}(\sup_{x \in E} |Z_x - \mathbb{E}[Z_x]| \geq \varepsilon) \leq 2\mathcal{C}^{(E, \delta), \frac{\varepsilon}{4L}} \exp\left(\frac{-\varepsilon^2 M}{2D^2}\right) \quad (12.157)$$

(cf. Definition 4.3.2).

Proof of Lemma 12.3.7. First, observe that the triangle inequality and the assumption that for all $x, y \in E$, $m \in \{1, 2, \dots, M\}$ it holds that $|Y_{x,m} - Y_{y,m}| \leq L\delta(x, y)$ imply that for all

$x, y \in E$ it holds that

$$\begin{aligned} |Z_x - Z_y| &= \left| \frac{1}{M} \left[\sum_{m=1}^M Y_{x,m} \right] - \frac{1}{M} \left[\sum_{m=1}^M Y_{y,m} \right] \right| = \frac{1}{M} \left| \sum_{m=1}^M (Y_{x,m} - Y_{y,m}) \right| \\ &\leq \frac{1}{M} \left[\sum_{m=1}^M |Y_{x,m} - Y_{y,m}| \right] \leq L\delta(x, y). \end{aligned} \quad (12.158)$$

Next note that the assumption that for all $x \in E$, $m \in \{1, 2, \dots, M\}$, $\omega \in \Omega$ it holds that $|Y_{x,m}(\omega)| \in [0, D]$ ensures that for all $x \in E$ it holds that

$$\mathbb{E}[|Z_x|] = \mathbb{E}\left[\frac{1}{M} \left[\sum_{m=1}^M Y_{x,m} \right]\right] = \frac{1}{M} \left[\sum_{m=1}^M \mathbb{E}[Y_{x,m}] \right] \leq D < \infty. \quad (12.159)$$

This proves item (i). Furthermore, note that item (i), (12.158), and Lemma 12.3.5 establish item (ii). Next observe that (12.156) shows that for all $x \in E$ it holds that

$$|Z_x - \mathbb{E}[Z_x]| = \left| \frac{1}{M} \left[\sum_{m=1}^M Y_{x,m} \right] - \mathbb{E}\left[\frac{1}{M} \left[\sum_{m=1}^M Y_{x,m} \right]\right] \right| = \frac{1}{M} \left| \sum_{m=1}^M (Y_{x,m} - \mathbb{E}[Y_{x,m}]) \right|. \quad (12.160)$$

Combining this with Corollary 12.1.21 (applied with $(\Omega, \mathcal{F}, \mathbb{P}) \curvearrowright (\Omega, \mathcal{F}, \mathbb{P})$, $N \curvearrowright M$, $\varepsilon \curvearrowright \frac{\varepsilon}{2}$, $(a_1, a_2, \dots, a_N) \curvearrowright (0, 0, \dots, 0)$, $(b_1, b_2, \dots, b_N) \curvearrowright (D, D, \dots, D)$, $(X_n)_{n \in \{1, 2, \dots, N\}} \curvearrowright (Y_{x,m})_{m \in \{1, 2, \dots, M\}}$ for $x \in E$ in the notation of Corollary 12.1.21) ensures that for all $x \in E$ it holds that

$$\mathbb{P}(|Z_x - \mathbb{E}[Z_x]| \geq \frac{\varepsilon}{2}) \leq 2 \exp\left(\frac{-2\left[\frac{\varepsilon}{2}\right]^2 M^2}{MD^2}\right) = 2 \exp\left(\frac{-\varepsilon^2 M}{2D^2}\right). \quad (12.161)$$

Combining this, (12.158), and (12.159) with Lemma 12.3.6 establishes item (iii). The proof of Lemma 12.3.7 is thus complete. \square

12.3.2 Uniform estimates for the statistical learning error

Lemma 12.3.8. Let (E, δ) be a separable metric space, assume $E \neq \emptyset$, let $M \in \mathbb{N}$, $\varepsilon, L, D \in (0, \infty)$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $X_{x,m}: \Omega \rightarrow \mathbb{R}$, $x \in E$, $m \in \{1, 2, \dots, M\}$, and $Y_m: \Omega \rightarrow \mathbb{R}$, $m \in \{1, 2, \dots, M\}$, be functions, assume for all $x \in E$ that $(X_{x,m}, Y_m)$, $m \in \{1, 2, \dots, M\}$, are i.i.d. random variables, assume for all $x, y \in E$, $m \in \{1, 2, \dots, M\}$ that $|X_{x,m} - X_{y,m}| \leq L\delta(x, y)$ and $|X_{x,m} - Y_m| \leq D$, let $\mathfrak{E}_x: \Omega \rightarrow [0, \infty)$, $x \in E$, satisfy for all $x \in E$ that

$$\mathfrak{E}_x = \frac{1}{M} \left[\sum_{m=1}^M |X_{x,m} - Y_m|^2 \right], \quad (12.162)$$

and let $\mathcal{E}_x \in [0, \infty)$, $x \in E$, satisfy for all $x \in E$ that $\mathcal{E}_x = \mathbb{E}[|X_{x,1} - Y_1|^2]$. Then $\Omega \ni \omega \mapsto \sup_{x \in E} |\mathfrak{E}_x(\omega) - \mathcal{E}_x| \in [0, \infty]$ is $\mathcal{F}/\mathcal{B}([0, \infty])$ -measurable and

$$\mathbb{P}(\sup_{x \in E} |\mathfrak{E}_x - \mathcal{E}_x| \geq \varepsilon) \leq 2\mathcal{C}^{(E, \delta), \frac{\varepsilon}{8LD}} \exp\left(\frac{-\varepsilon^2 M}{2D^4}\right) \quad (12.163)$$

(cf. Definition 4.3.2).

Proof of Lemma 12.3.8. Throughout this proof, let $\mathcal{E}_{x,m}: \Omega \rightarrow [0, D^2]$, $x \in E$, $m \in \{1, 2, \dots, M\}$, satisfy for all $x \in E$, $m \in \{1, 2, \dots, M\}$ that

$$\mathcal{E}_{x,m} = |X_{x,m} - Y_m|^2. \quad (12.164)$$

Observe that the fact that for all $x_1, x_2, y \in \mathbb{R}$ it holds that $(x_1 - y)^2 - (x_2 - y)^2 = (x_1 - x_2)((x_1 - y) + (x_2 - y))$, the assumption that for all $x \in E$, $m \in \{1, 2, \dots, M\}$ it holds that $|X_{x,m} - Y_m| \leq D$, and the assumption that for all $x, y \in E$, $m \in \{1, 2, \dots, M\}$ it holds that $|X_{x,m} - X_{y,m}| \leq L\delta(x, y)$ imply that for all $x, y \in E$, $m \in \{1, 2, \dots, M\}$ it holds that

$$\begin{aligned} |\mathcal{E}_{x,m} - \mathcal{E}_{y,m}| &= |(X_{x,m} - Y_m)^2 - (X_{y,m} - Y_m)^2| \\ &= |X_{x,m} - X_{y,m}| |(X_{x,m} - Y_m) + (X_{y,m} - Y_m)| \\ &\leq |X_{x,m} - X_{y,m}| (|X_{x,m} - Y_m| + |X_{y,m} - Y_m|) \\ &\leq 2D|X_{x,m} - X_{y,m}| \leq 2LD\delta(x, y). \end{aligned} \quad (12.165)$$

In addition, note that (12.162) and the assumption that for all $x \in E$ it holds that $(X_{x,m}, Y_m)$, $m \in \{1, 2, \dots, M\}$, are i.i.d. random variables show that for all $x \in E$ it holds that

$$\mathbb{E}[\mathfrak{E}_x] = \frac{1}{M} \left[\sum_{m=1}^M \mathbb{E}[|X_{x,m} - Y_m|^2] \right] = \frac{1}{M} \left[\sum_{m=1}^M \mathbb{E}[|X_{x,1} - Y_1|^2] \right] = \frac{1}{M} \left[\sum_{m=1}^M \mathcal{E}_x \right] = \mathcal{E}_x. \quad (12.166)$$

Furthermore, observe that the assumption that for all $x \in E$ it holds that $(X_{x,m}, Y_m)$, $m \in \{1, 2, \dots, M\}$, are i.i.d. random variables ensures that for all $x \in E$ it holds that $\mathcal{E}_{x,m}$, $m \in \{1, 2, \dots, M\}$, are i.i.d. random variables. Combining this, (12.165), and (12.166) with Lemma 12.3.7 (applied with $(E, \delta) \curvearrowright (E, \delta)$, $M \curvearrowright M$, $\varepsilon \curvearrowright \varepsilon$, $L \curvearrowright 2LD$, $D \curvearrowright D^2$, $(\Omega, \mathcal{F}, \mathbb{P}) \curvearrowright (\Omega, \mathcal{F}, \mathbb{P})$, $(Y_{x,m})_{x \in E, m \in \{1, 2, \dots, M\}} \curvearrowright (\mathcal{E}_{x,m})_{x \in E, m \in \{1, 2, \dots, M\}}$, $(Z_x)_{x \in E} = (\mathfrak{E}_x)_{x \in E}$ in the notation of Lemma 12.3.7) establishes (12.163). The proof of Lemma 12.3.8 is thus complete. \square

Proposition 12.3.9. Let $d, \mathfrak{d}, M \in \mathbb{N}$, $R, L, \mathcal{R}, \varepsilon \in (0, \infty)$, let $D \subseteq \mathbb{R}^d$ be a compact set, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $X_m: \Omega \rightarrow D$, $m \in \{1, 2, \dots, M\}$, and $Y_m: \Omega \rightarrow \mathbb{R}$, $m \in \{1, 2, \dots, M\}$, be functions, assume that (X_m, Y_m) , $m \in \{1, 2, \dots, M\}$, are i.i.d. random variables, let $H = (H_\theta)_{\theta \in [-R, R]^{\mathfrak{d}}}: [-R, R]^{\mathfrak{d}} \rightarrow C(D, \mathbb{R})$ satisfy for all

$\theta, \vartheta \in [-R, R]^\mathfrak{d}$, $x \in D$ that $|H_\theta(x) - H_\vartheta(x)| \leq L\|\theta - \vartheta\|_\infty$, assume for all $\theta \in [-R, R]^\mathfrak{d}$, $m \in \{1, 2, \dots, M\}$ that $|H_\theta(X_m) - Y_m| \leq \mathcal{R}$ and $\mathbb{E}[|Y_1|^2] < \infty$, let $\mathcal{E}: C(D, \mathbb{R}) \rightarrow [0, \infty)$ satisfy for all $f \in C(D, \mathbb{R})$ that $\mathcal{E}(f) = \mathbb{E}[|f(X_1) - Y_1|^2]$, and let $\mathfrak{E}: [-R, R]^\mathfrak{d} \times \Omega \rightarrow [0, \infty)$ satisfy for all $\theta \in [-R, R]^\mathfrak{d}$, $\omega \in \Omega$ that

$$\mathfrak{E}(\theta, \omega) = \frac{1}{M} \left[\sum_{m=1}^M |H_\theta(X_m(\omega)) - Y_m(\omega)|^2 \right] \quad (12.167)$$

(cf. Definition 3.3.4). Then $\Omega \ni \omega \mapsto \sup_{\theta \in [-R, R]^\mathfrak{d}} |\mathfrak{E}(\theta, \omega) - \mathcal{E}(H_\theta)| \in [0, \infty]$ is $\mathcal{F}/\mathcal{B}([0, \infty])$ -measurable and

$$\mathbb{P}\left(\sup_{\theta \in [-R, R]^\mathfrak{d}} |\mathfrak{E}(\theta) - \mathcal{E}(H_\theta)| \geq \varepsilon\right) \leq 2 \max\left\{1, \left[\frac{16L\mathcal{R}}{\varepsilon}\right]^\mathfrak{d}\right\} \exp\left(\frac{-\varepsilon^2 M}{2\mathcal{R}^4}\right). \quad (12.168)$$

Proof of Proposition 12.3.9. Throughout this proof, let $B \subseteq \mathbb{R}^\mathfrak{d}$ satisfy $B = [-R, R]^\mathfrak{d} = \{\theta \in \mathbb{R}^\mathfrak{d}: \|\theta\|_\infty \leq R\}$ and let $\delta: B \times B \rightarrow [0, \infty)$ satisfy for all $\theta, \vartheta \in B$ that

$$\delta(\theta, \vartheta) = \|\theta - \vartheta\|_\infty. \quad (12.169)$$

Observe that the assumption that (X_m, Y_m) , $m \in \{1, 2, \dots, M\}$, are i.i.d. random variables and the assumption that for all $\theta \in [-R, R]^\mathfrak{d}$ it holds that H_θ is continuous imply that for all $\theta \in B$ it holds that $(H_\theta(X_m), Y_m)$, $m \in \{1, 2, \dots, M\}$, are i.i.d. random variables. Combining this, the assumption that for all $\theta, \vartheta \in B$, $x \in D$ it holds that $|H_\theta(x) - H_\vartheta(x)| \leq L\|\theta - \vartheta\|_\infty$, and the assumption that for all $\theta \in B$, $m \in \{1, 2, \dots, M\}$ it holds that $|H_\theta(X_m) - Y_m| \leq \mathcal{R}$ with Lemma 12.3.8 (applied with $(E, \delta) \curvearrowright (B, \delta)$, $M \curvearrowright M$, $\varepsilon \curvearrowright \varepsilon$, $L \curvearrowright L$, $D \curvearrowright \mathcal{R}$, $(\Omega, \mathcal{F}, \mathbb{P}) \curvearrowright (\Omega, \mathcal{F}, \mathbb{P})$, $(X_{x,m})_{x \in E, m \in \{1, 2, \dots, M\}} \curvearrowright (H_\theta(X_m))_{\theta \in B, m \in \{1, 2, \dots, M\}}$, $(Y_m)_{m \in \{1, 2, \dots, M\}} \curvearrowright (Y_m)_{m \in \{1, 2, \dots, M\}}$, $(\mathfrak{E}_x)_{x \in E} \curvearrowright ((\Omega \ni \omega \mapsto \mathfrak{E}(\theta, \omega) \in [0, \infty]))_{\theta \in B}$, $(\mathcal{E}_x)_{x \in E} \curvearrowright (\mathcal{E}(H_\theta))_{\theta \in B}$ in the notation of Lemma 12.3.8) establishes that $\Omega \ni \omega \mapsto \sup_{\theta \in B} |\mathfrak{E}(\theta, \omega) - \mathcal{E}(H_\theta)| \in [0, \infty]$ is $\mathcal{F}/\mathcal{B}([0, \infty])$ -measurable and

$$\mathbb{P}\left(\sup_{\theta \in B} |\mathfrak{E}(\theta) - \mathcal{E}(H_\theta)| \geq \varepsilon\right) \leq 2\mathcal{C}^{(B, \delta), \frac{\varepsilon}{8L\mathcal{R}}} \exp\left(\frac{-\varepsilon^2 M}{2\mathcal{R}^4}\right) \quad (12.170)$$

(cf. Definition 4.3.2). Moreover, note that Proposition 12.2.21 (applied with $d \curvearrowright \mathfrak{d}$, $a \curvearrowright -R$, $b \curvearrowright R$, $r \curvearrowright \frac{\varepsilon}{8L\mathcal{R}}$, $\delta \curvearrowright \delta$ in the notation of Proposition 12.2.20) demonstrates that

$$\mathcal{C}^{(B, \delta), \frac{\varepsilon}{8L\mathcal{R}}} \leq \max\left\{1, \left(\frac{16L\mathcal{R}}{\varepsilon}\right)^\mathfrak{d}\right\}. \quad (12.171)$$

This and (12.170) prove (12.168). The proof of Proposition 12.3.9 is thus complete. \square

Corollary 12.3.10. Let $\mathfrak{d}, M, L \in \mathbb{N}$, $u \in \mathbb{R}$, $v \in (u, \infty)$, $R \in [1, \infty)$, $\varepsilon, b \in (0, \infty)$, $l = (l_0, l_1, \dots, l_L) \in \mathbb{N}^{L+1}$ satisfy $l_L = 1$ and $\sum_{k=1}^L l_k(l_{k-1} + 1) \leq \mathfrak{d}$, let $D \subseteq [-b, b]^{l_0}$ be a compact set, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $X_m: \Omega \rightarrow D$, $m \in \{1, 2, \dots, M\}$, and $Y_m: \Omega \rightarrow [u, v]$, $m \in \{1, 2, \dots, M\}$, be functions, assume that (X_m, Y_m) , $m \in \{1, 2, \dots, M\}$, are i.i.d. random variables, let $\mathcal{E}: C(D, \mathbb{R}) \rightarrow [0, \infty)$ satisfy for all $f \in C(D, \mathbb{R})$ that $\mathcal{E}(f) = \mathbb{E}[|f(X_1) - Y_1|^2]$, and let $\mathfrak{E}: [-R, R]^{\mathfrak{d}} \times \Omega \rightarrow [0, \infty)$ satisfy for all $\theta \in [-R, R]^{\mathfrak{d}}$, $\omega \in \Omega$ that

$$\mathfrak{E}(\theta, \omega) = \frac{1}{M} \left[\sum_{m=1}^M |\mathcal{N}_{u,v}^{\theta,l}(X_m(\omega)) - Y_m(\omega)|^2 \right] \quad (12.172)$$

(cf. Definition 4.4.1). Then

(i) it holds that $\Omega \ni \omega \mapsto \sup_{\theta \in [-R, R]^{\mathfrak{d}}} |\mathfrak{E}(\theta, \omega) - \mathcal{E}(\mathcal{N}_{u,v}^{\theta,l}|_D)| \in [0, \infty]$ is $\mathcal{F}/\mathcal{B}([0, \infty])$ -measurable and

(ii) it holds that

$$\begin{aligned} & \mathbb{P}\left(\sup_{\theta \in [-R, R]^{\mathfrak{d}}} |\mathfrak{E}(\theta) - \mathcal{E}(\mathcal{N}_{u,v}^{\theta,l}|_D)| \geq \varepsilon\right) \\ & \leq 2 \max\left\{1, \left[\frac{16L \max\{1, b\} (\|l\|_\infty + 1)^L R^L (v - u)}{\varepsilon}\right]^{\mathfrak{d}}\right\} \exp\left(\frac{-\varepsilon^2 M}{2(v - u)^4}\right). \end{aligned} \quad (12.173)$$

Proof of Corollary 12.3.10. Throughout this proof, let $\mathfrak{L} \in (0, \infty)$ satisfy

$$\mathfrak{L} = L \max\{1, b\} (\|l\|_\infty + 1)^L R^{L-1}. \quad (12.174)$$

Observe that Corollary 11.3.7 (applied with $a \curvearrowleft -b$, $b \curvearrowleft b$, $u \curvearrowleft u$, $v \curvearrowleft v$, $d \curvearrowleft \mathfrak{d}$, $L \curvearrowleft L$, $l \curvearrowleft l$ in the notation of Corollary 11.3.7) and the assumption that $D \subseteq [-b, b]^{l_0}$ show that for all $\theta, \vartheta \in [-R, R]^{\mathfrak{d}}$ it holds that

$$\begin{aligned} & \sup_{x \in D} |\mathcal{N}_{u,v}^{\theta,l}(x) - \mathcal{N}_{u,v}^{\vartheta,l}(x)| \\ & \leq \sup_{x \in [-b, b]^{l_0}} |\mathcal{N}_{u,v}^{\theta,l}(x) - \mathcal{N}_{u,v}^{\vartheta,l}(x)| \\ & \leq L \max\{1, b\} (\|l\|_\infty + 1)^L (\max\{1, \|\theta\|_\infty, \|\vartheta\|_\infty\})^{L-1} \|\theta - \vartheta\|_\infty \\ & \leq L \max\{1, b\} (\|l\|_\infty + 1)^L R^{L-1} \|\theta - \vartheta\|_\infty = \mathfrak{L} \|\theta - \vartheta\|_\infty. \end{aligned} \quad (12.175)$$

Furthermore, observe that the fact that for all $\theta \in \mathbb{R}^{\mathfrak{d}}$, $x \in \mathbb{R}^{l_0}$ it holds that $\mathcal{N}_{u,v}^{\theta,l}(x) \in [u, v]$ and the assumption that for all $m \in \{1, 2, \dots, M\}$, $\omega \in \Omega$ it holds that $Y_m(\omega) \in [u, v]$ demonstrate that for all $\theta \in [-R, R]^{\mathfrak{d}}$, $m \in \{1, 2, \dots, M\}$ it holds that

$$|\mathcal{N}_{u,v}^{\theta,l}(X_m) - Y_m| \leq v - u. \quad (12.176)$$

Combining this and (12.175) with Proposition 12.3.9 (applied with $d \curvearrowleft l_0$, $\mathfrak{d} \curvearrowleft \mathfrak{d}$, $M \curvearrowleft M$, $R \curvearrowleft R$, $L \curvearrowleft \mathfrak{L}$, $\mathcal{R} \curvearrowleft v - u$, $\varepsilon \curvearrowleft \varepsilon$, $D \curvearrowleft D$, $(\Omega, \mathcal{F}, \mathbb{P}) \curvearrowleft (\Omega, \mathcal{F}, \mathbb{P})$, $(X_m)_{m \in \{1, 2, \dots, M\}} \curvearrowleft (X_m)_{m \in \{1, 2, \dots, M\}}$, $(Y_m)_{m \in \{1, 2, \dots, M\}} \curvearrowleft ((\Omega \ni \omega \mapsto Y_m(\omega) \in \mathbb{R}))_{m \in \{1, 2, \dots, M\}}$, $H \curvearrowleft ([-R, R]^{\mathfrak{d}} \ni \theta \mapsto \mathcal{N}_{u,v}^{\theta,l}|_D \in C(D, \mathbb{R}))$, $\mathcal{E} \curvearrowleft \mathcal{E}$, $\mathfrak{E} \curvearrowleft \mathfrak{E}$ in the notation of Proposition 12.3.9) establishes that $\Omega \ni \omega \mapsto \sup_{\theta \in [-R, R]^{\mathfrak{d}}} |\mathfrak{E}(\theta, \omega) - \mathcal{E}(\mathcal{N}_{u,v}^{\theta,l}|_D)| \in [0, \infty]$ is $\mathcal{F}/\mathcal{B}([0, \infty])$ -measurable and

$$\mathbb{P}(\sup_{\theta \in [-R, R]^{\mathfrak{d}}} |\mathfrak{E}(\theta) - \mathcal{E}(\mathcal{N}_{u,v}^{\theta,l}|_D)| \geq \varepsilon) \leq 2 \max \left\{ 1, \left[\frac{16\mathfrak{L}R(v-u)}{\varepsilon} \right]^{\mathfrak{d}} \right\} \exp \left(\frac{-\varepsilon^2 M}{2(v-u)^4} \right). \quad (12.177)$$

The proof of Corollary 12.3.10 is thus complete. \square

Chapter 13

Strong generalization error estimates

In Chapter 12 above we reviewed generalization error estimates in the probabilistic sense. Besides such probabilistic generalization error estimates, generalization error estimates in the strong L^p -sense are also considered in the literature and in our overall error analysis in Chapter 15 below we employ such strong generalization error estimates. These estimates are precisely the subject of this chapter (cf. Corollary 13.3.3 below).

We refer to the beginning of Chapter 12 for a short list of references in the literature dealing with similar generalization error estimates. The specific material in this chapter mostly consists of slightly modified extracts from Jentzen & Welti [243, Section 4].

13.1 Monte Carlo estimates

Proposition 13.1.1. *Let $d, M \in \mathbb{N}$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $X_j: \Omega \rightarrow \mathbb{R}^d$, $j \in \{1, 2, \dots, M\}$, be independent random variables, and assume $\max_{j \in \{1, 2, \dots, M\}} \mathbb{E}[\|X_j\|_2] < \infty$ (cf. Definition 3.3.4). Then*

$$\begin{aligned} & \left(\mathbb{E} \left[\left\| \frac{1}{M} \left[\sum_{j=1}^M X_j \right] - \mathbb{E} \left[\frac{1}{M} \left[\sum_{j=1}^M X_j \right] \right] \right\|_2^2 \right] \right)^{1/2} \\ & \leq \frac{1}{\sqrt{M}} \left[\max_{j \in \{1, 2, \dots, M\}} (\mathbb{E}[\|X_j - \mathbb{E}[X_j]\|_2^2])^{1/2} \right]. \end{aligned} \quad (13.1)$$

Proof of Proposition 13.1.1. Observe that the fact that for all $x \in \mathbb{R}^d$ it holds that $\langle x, x \rangle =$

$\|x\|_2^2$ demonstrates that

$$\begin{aligned}
 & \left\| \frac{1}{M} \left[\sum_{j=1}^M X_j \right] - \mathbb{E} \left[\frac{1}{M} \sum_{j=1}^M X_j \right] \right\|_2^2 \\
 &= \frac{1}{M^2} \left\| \left[\sum_{j=1}^M X_j \right] - \mathbb{E} \left[\sum_{j=1}^M X_j \right] \right\|_2^2 \\
 &= \frac{1}{M^2} \left\| \sum_{j=1}^M (X_j - \mathbb{E}[X_j]) \right\|_2^2 \\
 &= \frac{1}{M^2} \left[\sum_{i,j=1}^M \langle X_i - \mathbb{E}[X_i], X_j - \mathbb{E}[X_j] \rangle \right] \\
 &= \frac{1}{M^2} \left[\sum_{j=1}^M \|X_j - \mathbb{E}[X_j]\|_2^2 \right] + \frac{1}{M^2} \left[\sum_{(i,j) \in \{1,2,\dots,M\}^2, i \neq j} \langle X_i - \mathbb{E}[X_i], X_j - \mathbb{E}[X_j] \rangle \right]
 \end{aligned} \tag{13.2}$$

(cf. Definition 1.4.7). This, the fact that for all independent random variables $Y: \Omega \rightarrow \mathbb{R}^d$ and $Z: \Omega \rightarrow \mathbb{R}^d$ with $\mathbb{E}[\|Y\|_2 + \|Z\|_2] < \infty$ it holds that $\mathbb{E}[|\langle Y, Z \rangle|] < \infty$ and $\mathbb{E}[\langle Y, Z \rangle] = \langle \mathbb{E}[Y], \mathbb{E}[Z] \rangle$, and the assumption that $X_j: \Omega \rightarrow \mathbb{R}^d$, $j \in \{1, 2, \dots, M\}$, are independent random variables prove that

$$\begin{aligned}
 & \mathbb{E} \left[\left\| \frac{1}{M} \left[\sum_{j=1}^M X_j \right] - \mathbb{E} \left[\frac{1}{M} \sum_{j=1}^M X_j \right] \right\|_2^2 \right] \\
 &= \frac{1}{M^2} \left[\sum_{j=1}^M \mathbb{E} [\|X_j - \mathbb{E}[X_j]\|_2^2] \right] + \frac{1}{M^2} \left[\sum_{(i,j) \in \{1,2,\dots,M\}^2, i \neq j} \langle \mathbb{E}[X_i - \mathbb{E}[X_i]], \mathbb{E}[X_j - \mathbb{E}[X_j]] \rangle \right] \\
 &= \frac{1}{M^2} \left[\sum_{j=1}^M \mathbb{E} [\|X_j - \mathbb{E}[X_j]\|_2^2] \right] \\
 &\leq \frac{1}{M} \left[\max_{j \in \{1,2,\dots,M\}} \mathbb{E} [\|X_j - \mathbb{E}[X_j]\|_2^2] \right].
 \end{aligned} \tag{13.3}$$

The proof of Proposition 13.1.1 is thus complete. \square

Definition 13.1.2 (Rademacher family). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let J be a set. Then we say that $(r_j)_{j \in J}$ is a \mathbb{P} -Rademacher family if and only if it holds that $r_j: \Omega \rightarrow \{-1, 1\}$, $j \in J$, are independent random variables with

$$\forall j \in J: \mathbb{P}(r_j = 1) = \mathbb{P}(r_j = -1). \tag{13.4}$$

Definition 13.1.3 (*p*-Kahane–Khintchine constant). Let $p \in (0, \infty)$. Then we denote by $\mathfrak{K}_p \in (0, \infty]$ the extended real number given by

$$\mathfrak{K}_p = \sup \left\{ c \in [0, \infty) : \begin{array}{l} \exists \mathbb{R}\text{-Banach space } (E, \|\cdot\|) : \\ \exists \text{ probability space } (\Omega, \mathcal{F}, \mathbb{P}) : \\ \exists \mathbb{P}\text{-Rademacher family } (r_j)_{j \in \mathbb{N}} : \\ \exists k \in \mathbb{N} : \exists x_1, x_2, \dots, x_k \in E \setminus \{0\} : \\ \left(\mathbb{E} \left[\left\| \sum_{j=1}^k r_j x_j \right\|^p \right] \right)^{1/p} = c \left(\mathbb{E} \left[\left\| \sum_{j=1}^k r_j x_j \right\|^2 \right] \right)^{1/2} \end{array} \right\} \quad (13.5)$$

(cf. Definition 13.1.2).

Lemma 13.1.4. It holds for all $p \in [2, \infty)$ that

$$\mathfrak{K}_p \leq \sqrt{p-1} < \infty \quad (13.6)$$

(cf. Definition 13.1.3).

Proof of Lemma 13.1.4. Note that (13.5) and Grohs et al. [186, Corollary 2.5] imply (13.6). The proof of Lemma 13.1.4 is thus complete. \square

Proposition 13.1.5. Let $d, M \in \mathbb{N}$, $p \in [2, \infty)$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $X_j : \Omega \rightarrow \mathbb{R}^d$, $j \in \{1, 2, \dots, M\}$, be independent random variables, and assume

$$\max_{j \in \{1, 2, \dots, M\}} \mathbb{E}[\|X_j\|_2] < \infty \quad (13.7)$$

(cf. Definition 3.3.4). Then

$$\left(\mathbb{E} \left[\left\| \sum_{j=1}^M X_j \right\| - \mathbb{E} \left[\left\| \sum_{j=1}^M X_j \right\|_2 \right] \right]^p \right)^{1/p} \leq 2\mathfrak{K}_p \left[\sum_{j=1}^M (\mathbb{E}[\|X_j - \mathbb{E}[X_j]\|_2^p])^{2/p} \right]^{1/2} \quad (13.8)$$

(cf. Definition 13.1.3 and Lemma 13.1.4).

Proof of Proposition 13.1.5. Observe that (13.5) and Cox et al. [89, Corollary 5.11] ensure (13.6). The proof of Proposition 13.1.5 is thus complete. \square

Corollary 13.1.6. Let $d, M \in \mathbb{N}$, $p \in [2, \infty)$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $X_j : \Omega \rightarrow \mathbb{R}^d$, $j \in \{1, 2, \dots, M\}$, be independent random variables, and assume

$$\max_{j \in \{1, 2, \dots, M\}} \mathbb{E}[\|X_j\|_2] < \infty \quad (13.9)$$

(cf. Definition 3.3.4). Then

$$\left(\mathbb{E} \left[\left\| \frac{1}{M} \left[\sum_{j=1}^M X_j \right] - \mathbb{E} \left[\frac{1}{M} \sum_{j=1}^M X_j \right] \right\|_2^p \right] \right)^{1/p} \leq \frac{2\sqrt{p-1}}{\sqrt{M}} \left[\max_{j \in \{1, 2, \dots, M\}} (\mathbb{E} [\|X_j - \mathbb{E}[X_j]\|_2^p])^{1/p} \right]. \quad (13.10)$$

Proof of Corollary 13.1.6. Note that Proposition 13.1.5 and Lemma 13.1.4 show that

$$\begin{aligned} & \left(\mathbb{E} \left[\left\| \frac{1}{M} \left[\sum_{j=1}^M X_j \right] - \mathbb{E} \left[\frac{1}{M} \sum_{j=1}^M X_j \right] \right\|_2^p \right] \right)^{1/p} \\ &= \frac{1}{M} \left(\mathbb{E} \left[\left\| \left[\sum_{j=1}^M X_j \right] - \mathbb{E} \left[\sum_{j=1}^M X_j \right] \right\|_2^p \right] \right)^{1/p} \\ &\leq \frac{2\mathfrak{K}_p}{M} \left[\sum_{j=1}^M (\mathbb{E} [\|X_j - \mathbb{E}[X_j]\|_2^p])^{2/p} \right]^{1/2} \\ &\leq \frac{2\mathfrak{K}_p}{M} \left[M \left(\max_{j \in \{1, 2, \dots, M\}} (\mathbb{E} [\|X_j - \mathbb{E}[X_j]\|_2^p])^{2/p} \right) \right]^{1/2} \\ &= \frac{2\mathfrak{K}_p}{\sqrt{M}} \left[\max_{j \in \{1, 2, \dots, M\}} (\mathbb{E} [\|X_j - \mathbb{E}[X_j]\|_2^p])^{1/p} \right] \\ &\leq \frac{2\sqrt{p-1}}{\sqrt{M}} \left[\max_{j \in \{1, 2, \dots, M\}} (\mathbb{E} [\|X_j - \mathbb{E}[X_j]\|_2^p])^{1/p} \right] \end{aligned} \quad (13.11)$$

(cf. Definition 13.1.3). The proof of Corollary 13.1.6 is thus complete. \square

13.2 Uniform strong error estimates for random fields

Lemma 13.2.1. Let (E, δ) be a separable metric space, let $N \in \mathbb{N}$, $r_1, r_2, \dots, r_N \in [0, \infty)$, $z_1, z_2, \dots, z_N \in E$ satisfy

$$E \subseteq \bigcup_{n=1}^N \{x \in E : \delta(x, z_n) \leq r_n\}, \quad (13.12)$$

let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, for every $x \in E$ let $Z_x : \Omega \rightarrow \mathbb{R}$ be a random variable, let $L \in [0, \infty)$ satisfy for all $x, y \in E$ that $|Z_x - Z_y| \leq L\delta(x, y)$, and let $p \in [0, \infty)$. Then

$$\mathbb{E} [\sup_{x \in E} |Z_x|^p] \leq \sum_{n=1}^N \mathbb{E} [(Lr_n + |Z_{z_n}|)^p] \quad (13.13)$$

(cf. Lemma 12.3.2).

Proof of Lemma 13.2.1. Throughout this proof, for every $n \in \{1, 2, \dots, N\}$ let

$$B_n = \{x \in E : \delta(x, z_n) \leq r_n\}. \quad (13.14)$$

Observe that (13.12) and (13.14) establish that

$$E \subseteq \bigcup_{n=1}^N B_n \quad \text{and} \quad E \supseteq \bigcup_{n=1}^N B_n. \quad (13.15)$$

Therefore, we obtain that

$$\sup_{x \in E} |Z_x| = \sup_{x \in (\bigcup_{n=1}^N B_n)} |Z_x| = \max_{n \in \{1, 2, \dots, N\}} \sup_{x \in B_n} |Z_x|. \quad (13.16)$$

Hence, we obtain that

$$\begin{aligned} \mathbb{E}[\sup_{x \in E} |Z_x|^p] &= \mathbb{E}[\max_{n \in \{1, 2, \dots, N\}} \sup_{x \in B_n} |Z_x|^p] \\ &\leq \mathbb{E}\left[\sum_{n=1}^N \sup_{x \in B_n} |Z_x|^p\right] = \sum_{n=1}^N \mathbb{E}[\sup_{x \in B_n} |Z_x|^p]. \end{aligned} \quad (13.17)$$

(cf. Lemma 12.3.2). Furthermore, note that the assumption that for all $x, y \in E$ it holds that $|Z_x - Z_y| \leq L\delta(x, y)$ demonstrates that for all $n \in \{1, 2, \dots, N\}$, $x \in B_n$ it holds that

$$|Z_x| = |Z_x - Z_{z_n} + Z_{z_n}| \leq |Z_x - Z_{z_n}| + |Z_{z_n}| \leq L\delta(x, z_n) + |Z_{z_n}| \leq Lr_n + |Z_{z_n}|. \quad (13.18)$$

This and (13.17) prove that

$$\mathbb{E}[\sup_{x \in E} |Z_x|^p] \leq \sum_{n=1}^N \mathbb{E}[(Lr_n + |Z_{z_n}|)^p]. \quad (13.19)$$

The proof of Lemma 13.2.1 is thus complete. \square

Lemma 13.2.2. Let (E, δ) be a non-empty separable metric space, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, for every $x \in E$ let $Z_x : \Omega \rightarrow \mathbb{R}$ be a random variable, let $L \in (0, \infty)$ satisfy for all $x, y \in E$ that $|Z_x - Z_y| \leq L\delta(x, y)$, and let $p, r \in (0, \infty)$. Then

$$\mathbb{E}[\sup_{x \in E} |Z_x|^p] \leq C^{(E, \delta), r} \left[\sup_{x \in E} \mathbb{E}[(Lr + |Z_x|)^p] \right] \quad (13.20)$$

(cf. Definition 4.3.2 and Lemma 12.3.2).

Proof of Lemma 13.2.2. Throughout this proof, assume without loss of generality that $C^{(E, \delta), r} < \infty$, let $N = C^{(E, \delta), r}$, and let $z_1, z_2, \dots, z_N \in E$ satisfy

$$E \subseteq \bigcup_{n=1}^N \{x \in E : \delta(x, z_n) \leq r\} \quad (13.21)$$

(cf. Definition 4.3.2). Observe that Lemma 13.2.1 (applied with $r_1 \curvearrowleft r, r_2 \curvearrowleft r, \dots, r_N \curvearrowleft r$ in the notation of Lemma 13.2.1) implies that

$$\begin{aligned} \mathbb{E}\left[\sup_{x \in E} |Z_x|^p\right] &\leq \sum_{i=1}^N \mathbb{E}\left[(Lr + |Z_{z_i}|)^p\right] \\ &\leq \sum_{i=1}^N \left[\sup_{x \in E} \mathbb{E}\left[(Lr + |Z_x|)^p\right] \right] = N \left[\sup_{x \in E} \mathbb{E}\left[(Lr + |Z_x|)^p\right] \right]. \end{aligned} \quad (13.22)$$

(cf. Lemma 12.3.2). The proof of Lemma 13.2.2 is thus complete. \square

Lemma 13.2.3. Let (E, δ) be a non-empty separable metric space, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, for every $x \in E$ let $Z_x: \Omega \rightarrow \mathbb{R}$ be a random variable with $\mathbb{E}[|Z_x|] < \infty$, let $L \in (0, \infty)$ satisfy for all $x, y \in E$ that $|Z_x - Z_y| \leq L\delta(x, y)$, and let $p \in [1, \infty)$, $r \in (0, \infty)$. Then

$$\left(\mathbb{E}\left[\sup_{x \in E} |Z_x - \mathbb{E}[Z_x]|^p\right]\right)^{1/p} \leq (\mathcal{C}^{(E, \delta), r})^{1/p} \left[2Lr + \sup_{x \in E} (\mathbb{E}[|Z_x - \mathbb{E}[Z_x]|^p])^{1/p}\right] \quad (13.23)$$

(cf. Definition 4.3.2 and Lemma 12.3.5).

Proof of Lemma 13.2.3. Throughout this proof, for every $x \in E$ let $Y_x: \Omega \rightarrow \mathbb{R}$ satisfy for all $\omega \in \Omega$ that

$$Y_x(\omega) = Z_x(\omega) - \mathbb{E}[Z_x]. \quad (13.24)$$

Note that (13.24) and the triangle inequality ensure that for all $x, y \in E$ it holds that

$$\begin{aligned} |Y_x - Y_y| &= |(Z_x - \mathbb{E}[Z_x]) - (Z_y - \mathbb{E}[Z_y])| \\ &= |(Z_x - Z_y) - (\mathbb{E}[Z_x] - \mathbb{E}[Z_y])| \\ &\leq |Z_x - Z_y| + |\mathbb{E}[Z_x] - \mathbb{E}[Z_y]| \\ &\leq L\delta(x, y) + \mathbb{E}[|Z_x - Z_y|] \leq 2L\delta(x, y). \end{aligned} \quad (13.25)$$

Lemma 13.2.2 (applied with $L \curvearrowleft 2L$, $(\Omega, \mathcal{F}, \mathbb{P}) \curvearrowleft (\Omega, \mathcal{F}, \mathbb{P})$, $(Z_x)_{x \in E} \curvearrowleft (Y_x)_{x \in E}$ in the notation of Lemma 13.2.2) therefore shows that

$$\begin{aligned} \left(\mathbb{E}\left[\sup_{x \in E} |Z_x - \mathbb{E}[Z_x]|^p\right]\right)^{1/p} &= \left(\mathbb{E}\left[\sup_{x \in E} |Y_x|^p\right]\right)^{1/p} \\ &\leq (\mathcal{C}^{(E, \delta), r})^{1/p} \left[\sup_{x \in E} (\mathbb{E}[(2Lr + |Y_x|)^p])^{1/p} \right] \\ &\leq (\mathcal{C}^{(E, \delta), r})^{1/p} \left[2Lr + \sup_{x \in E} (\mathbb{E}[|Y_x|^p])^{1/p} \right] \\ &= (\mathcal{C}^{(E, \delta), r})^{1/p} \left[2Lr + \sup_{x \in E} (\mathbb{E}[|Z_x - \mathbb{E}[Z_x]|^p])^{1/p} \right]. \end{aligned} \quad (13.26)$$

The proof of Lemma 13.2.3 is thus complete. \square

Lemma 13.2.4. Let (E, δ) be a non-empty separable metric space, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $M \in \mathbb{N}$, for every $x \in E$ let $Y_{x,m}: \Omega \rightarrow \mathbb{R}$, $m \in \{1, 2, \dots, M\}$, be independent random variables with $\mathbb{E}[|Y_{x,1}| + |Y_{x,2}| + \dots + |Y_{x,M}|] < \infty$, let $L \in (0, \infty)$ satisfy for all $x, y \in E$, $m \in \{1, 2, \dots, M\}$ that

$$|Y_{x,m} - Y_{y,m}| \leq L\delta(x, y), \quad (13.27)$$

and for every $x \in E$ let $Z_x: \Omega \rightarrow \mathbb{R}$ satisfy

$$Z_x = \frac{1}{M} \left[\sum_{m=1}^M Y_{x,m} \right]. \quad (13.28)$$

Then

- (i) it holds for all $x \in E$ that $\mathbb{E}[|Z_x|] < \infty$,
- (ii) it holds that $\Omega \ni \omega \mapsto \sup_{x \in E} |Z_x(\omega) - \mathbb{E}[Z_x]| \in [0, \infty]$ is $\mathcal{F}/\mathcal{B}([0, \infty])$ -measurable, and
- (iii) it holds for all $p \in [2, \infty)$, $r \in (0, \infty)$ that

$$\begin{aligned} & (\mathbb{E}[\sup_{x \in E} |Z_x - \mathbb{E}[Z_x]|^p])^{1/p} \\ & \leq 2(\mathcal{C}^{(E, \delta), r})^{1/p} \left[Lr + \frac{\sqrt{p-1}}{\sqrt{M}} \left(\sup_{x \in E} \max_{m \in \{1, 2, \dots, M\}} (\mathbb{E}[|Y_{x,m} - \mathbb{E}[Y_{x,m}]|^p])^{1/p} \right) \right] \end{aligned} \quad (13.29)$$

(cf. Definition 4.3.2).

Proof of Lemma 13.2.4. Observe that the assumption that for all $x \in E$, $m \in \{1, 2, \dots, M\}$ it holds that $\mathbb{E}[|Y_{x,m}|] < \infty$ establishes that for all $x \in E$ it holds that

$$\mathbb{E}[|Z_x|] = \mathbb{E}\left[\frac{1}{M} \left[\sum_{m=1}^M Y_{x,m} \right]\right] \leq \frac{1}{M} \left[\sum_{m=1}^M \mathbb{E}[|Y_{x,m}|] \right] \leq \max_{m \in \{1, 2, \dots, M\}} \mathbb{E}[|Y_{x,m}|] < \infty. \quad (13.30)$$

This proves item (i). Note that (13.27) demonstrates that for all $x, y \in E$ it holds that

$$|Z_x - Z_y| = \frac{1}{M} \left| \left[\sum_{m=1}^M Y_{x,m} \right] - \left[\sum_{m=1}^M Y_{y,m} \right] \right| \leq \frac{1}{M} \left[\sum_{m=1}^M |Y_{x,m} - Y_{y,m}| \right] \leq L\delta(x, y). \quad (13.31)$$

Item (i) and Lemma 12.3.5 hence establish item (ii). It thus remains to show item (iii). For this observe that item (i), (13.31), and Lemma 13.2.3 imply that for all $p \in [1, \infty)$, $r \in (0, \infty)$ it holds that

$$(\mathbb{E}[\sup_{x \in E} |Z_x - \mathbb{E}[Z_x]|^p])^{1/p} \leq (\mathcal{C}^{(E, \delta), r})^{1/p} \left[2Lr + \sup_{x \in E} (\mathbb{E}[|Z_x - \mathbb{E}[Z_x]|^p])^{1/p} \right] \quad (13.32)$$

(cf. Definition 4.3.2). Furthermore, note that (13.30) and Corollary 13.1.6 (applied with $d \curvearrowleft 1$, $(X_m)_{m \in \{1, 2, \dots, M\}} \curvearrowleft (Y_{x,m})_{m \in \{1, 2, \dots, M\}}$ for $x \in E$ in the notation of Corollary 13.1.6) ensure that for all $x \in E$, $p \in [2, \infty)$, $r \in (0, \infty)$ it holds that

$$\begin{aligned} (\mathbb{E}[|Z_x - \mathbb{E}[Z_x]|^p])^{1/p} &= \left(\mathbb{E}\left[\left| \frac{1}{M} \left[\sum_{m=1}^M Y_{x,m} \right] - \mathbb{E}\left[\frac{1}{M} \sum_{m=1}^M Y_{x,m} \right] \right|^p \right] \right)^{1/p} \\ &\leq \frac{2\sqrt{p-1}}{\sqrt{M}} \left[\max_{m \in \{1, 2, \dots, M\}} (\mathbb{E}[|Y_{x,m} - \mathbb{E}[Y_{x,m}]|^p])^{1/p} \right]. \end{aligned} \quad (13.33)$$

Combining this with (13.32) shows that for all $p \in [2, \infty)$, $r \in (0, \infty)$ it holds that

$$\begin{aligned} &(\mathbb{E}[\sup_{x \in E} |Z_x - \mathbb{E}[Z_x]|^p])^{1/p} \\ &\leq (\mathcal{C}^{(E, \delta), r})^{1/p} \left[2Lr + \frac{2\sqrt{p-1}}{\sqrt{M}} \left(\sup_{x \in E} \max_{m \in \{1, 2, \dots, M\}} (\mathbb{E}[|Y_{x,m} - \mathbb{E}[Y_{x,m}]|^p])^{1/p} \right) \right] \\ &= 2(\mathcal{C}^{(E, \delta), r})^{1/p} \left[Lr + \frac{\sqrt{p-1}}{\sqrt{M}} \left(\sup_{x \in E} \max_{m \in \{1, 2, \dots, M\}} (\mathbb{E}[|Y_{x,m} - \mathbb{E}[Y_{x,m}]|^p])^{1/p} \right) \right]. \end{aligned} \quad (13.34)$$

The proof of Lemma 13.2.4 is thus complete. \square

Corollary 13.2.5. *Let (E, δ) be a non-empty separable metric space, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $M \in \mathbb{N}$, for every $x \in E$ let $Y_{x,m}: \Omega \rightarrow \mathbb{R}$, $m \in \{1, 2, \dots, M\}$, be independent random variables with $\mathbb{E}[|Y_{x,1}| + |Y_{x,2}| + \dots + |Y_{x,M}|] < \infty$, let $L \in (0, \infty)$ satisfy for all $x, y \in E$, $m \in \{1, 2, \dots, M\}$ that $|Y_{x,m} - Y_{y,m}| \leq L\delta(x, y)$, and for every $x \in E$ let $Z_x: \Omega \rightarrow \mathbb{R}$ satisfy*

$$Z_x = \frac{1}{M} \left[\sum_{m=1}^M Y_{x,m} \right]. \quad (13.35)$$

Then

- (i) *it holds for all $x \in E$ that $\mathbb{E}[|Z_x|] < \infty$,*
- (ii) *it holds that $\Omega \ni \omega \mapsto \sup_{x \in E} |Z_x(\omega) - \mathbb{E}[Z_x]| \in [0, \infty]$ is $\mathcal{F}/\mathcal{B}([0, \infty])$ -measurable, and*
- (iii) *it holds for all $p \in [2, \infty)$, $c \in (0, \infty)$ that*

$$\begin{aligned} &(\mathbb{E}[\sup_{x \in E} |Z_x - \mathbb{E}[Z_x]|^p])^{1/p} \\ &\leq \frac{2\sqrt{p-1}}{\sqrt{M}} \left(\mathcal{C}^{(E, \delta), \frac{c\sqrt{p-1}}{L\sqrt{M}}} \right)^{1/p} \left[c + \sup_{x \in E} \max_{m \in \{1, 2, \dots, M\}} (\mathbb{E}[|Y_{x,m} - \mathbb{E}[Y_{x,m}]|^p])^{1/p} \right] \end{aligned} \quad (13.36)$$

(cf. Definition 4.3.2).

Proof of Corollary 13.2.5. Observe that Lemma 13.2.4 proves items (i) and (ii). Note that Lemma 13.2.4 (applied with $r \curvearrowleft c\sqrt{p-1}/(L\sqrt{M})$ for $c \in (0, \infty)$ in the notation of Lemma 13.2.4) demonstrates that for all $p \in [2, \infty)$, $c \in (0, \infty)$ it holds that

$$\begin{aligned} & (\mathbb{E}[\sup_{x \in E} |Z_x - \mathbb{E}[Z_x]|^p])^{1/p} \\ & \leq 2 \left(\mathcal{C}^{(E, \delta), \frac{c\sqrt{p-1}}{L\sqrt{M}}} \right)^{1/p} \left[L \frac{c\sqrt{p-1}}{L\sqrt{M}} \right. \\ & \quad \left. + \frac{\sqrt{p-1}}{\sqrt{M}} \left(\sup_{x \in E} \max_{m \in \{1, 2, \dots, M\}} (\mathbb{E}[|Y_{x,m} - \mathbb{E}[Y_{x,m}]|^p])^{1/p} \right) \right] \\ & = \frac{2\sqrt{p-1}}{\sqrt{M}} \left(\mathcal{C}^{(E, \delta), \frac{c\sqrt{p-1}}{L\sqrt{M}}} \right)^{1/p} \left[c + \sup_{x \in E} \max_{m \in \{1, 2, \dots, M\}} (\mathbb{E}[|Y_{x,m} - \mathbb{E}[Y_{x,m}]|^p])^{1/p} \right] \end{aligned} \tag{13.37}$$

(cf. Definition 4.3.2). This establishes item (iii). The proof of Corollary 13.2.5 is thus complete. \square

13.3 Strong convergence rates for the generalisation error

Lemma 13.3.1. *Let (E, δ) be a separable metric space, assume $E \neq \emptyset$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $M \in \mathbb{N}$, let $X_{x,m}: \Omega \rightarrow \mathbb{R}$, $m \in \{1, 2, \dots, M\}$, $x \in E$, and $Y_m: \Omega \rightarrow \mathbb{R}$, $m \in \{1, 2, \dots, M\}$, be functions, assume for all $x \in E$ that $(X_{x,m}, Y_m)$, $m \in \{1, 2, \dots, M\}$, are i.i.d. random variables, let $L, b \in (0, \infty)$ satisfy for all $x, y \in E$, $m \in \{1, 2, \dots, M\}$ that*

$$|X_{x,m} - Y_m| \leq b \quad \text{and} \quad |X_{x,m} - X_{y,m}| \leq L\delta(x, y), \tag{13.38}$$

and let $\mathbf{R}: E \rightarrow [0, \infty)$ and $\mathcal{R}: E \times \Omega \rightarrow [0, \infty)$ satisfy for all $x \in E$, $\omega \in \Omega$ that

$$\mathbf{R}(x) = \mathbb{E}[|X_{x,1} - Y_1|^2] \quad \text{and} \quad \mathcal{R}(x, \omega) = \frac{1}{M} \left[\sum_{m=1}^M |X_{x,m}(\omega) - Y_m(\omega)|^2 \right]. \tag{13.39}$$

Then

- (i) it holds that $\Omega \ni \omega \mapsto \sup_{x \in E} |\mathcal{R}(x, \omega) - \mathbf{R}(x)| \in [0, \infty]$ is $\mathcal{F}/\mathcal{B}([0, \infty])$ -measurable and
- (ii) it holds for all $p \in [2, \infty)$, $c \in (0, \infty)$ that

$$(\mathbb{E}[\sup_{x \in E} |\mathcal{R}(x) - \mathbf{R}(x)|^p])^{1/p} \leq \left(\mathcal{C}^{(E, \delta), \frac{cb\sqrt{p-1}}{2L\sqrt{M}}} \right)^{1/p} \left[\frac{2(c+1)b^2\sqrt{p-1}}{\sqrt{M}} \right] \tag{13.40}$$

(cf. Definition 4.3.2).

Proof of Lemma 13.3.1. Throughout this proof, for every $x \in E$, $m \in \{1, 2, \dots, M\}$ let $\mathcal{Y}_{x,m}: \Omega \rightarrow \mathbb{R}$ satisfy $\mathcal{Y}_{x,m} = |X_{x,m} - Y_m|^2$. Observe that the assumption that for all $x \in E$ it holds that $(X_{x,m}, Y_m)$, $m \in \{1, 2, \dots, M\}$, are i.i.d. random variables implies that for all $x \in E$ it holds that

$$\mathbb{E}[\mathcal{R}(x)] = \frac{1}{M} \left[\sum_{m=1}^M \mathbb{E}[|X_{x,m} - Y_m|^2] \right] = \frac{M \mathbb{E}[|X_{x,1} - Y_1|^2]}{M} = \mathbf{R}(x). \quad (13.41)$$

Furthermore, note that the assumption that for all $x \in E$, $m \in \{1, 2, \dots, M\}$ it holds that $|X_{x,m} - Y_m| \leq b$ shows that for all $x \in E$, $m \in \{1, 2, \dots, M\}$ it holds that

$$\mathbb{E}[|\mathcal{Y}_{x,m}|] = \mathbb{E}[|X_{x,m} - Y_m|^2] \leq b^2 < \infty, \quad (13.42)$$

$$\mathcal{Y}_{x,m} - \mathbb{E}[\mathcal{Y}_{x,m}] = |X_{x,m} - Y_m|^2 - \mathbb{E}[|X_{x,m} - Y_m|^2] \leq |X_{x,m} - Y_m|^2 \leq b^2, \quad (13.43)$$

and

$$\mathbb{E}[\mathcal{Y}_{x,m}] - \mathcal{Y}_{x,m} = \mathbb{E}[|X_{x,m} - Y_m|^2] - |X_{x,m} - Y_m|^2 \leq \mathbb{E}[|X_{x,m} - Y_m|^2] \leq b^2. \quad (13.44)$$

Observe that (13.42), (13.43), and (13.44) ensure for all $x \in E$, $m \in \{1, 2, \dots, M\}$, $p \in (0, \infty)$ that

$$(\mathbb{E}[|\mathcal{Y}_{x,m} - \mathbb{E}[\mathcal{Y}_{x,m}]|^p])^{1/p} \leq (\mathbb{E}[b^{2p}])^{1/p} = b^2. \quad (13.45)$$

Moreover, note that (13.38) and the fact that for all $x_1, x_2, y \in \mathbb{R}$ it holds that $(x_1 - y)^2 - (x_2 - y)^2 = (x_1 - x_2)((x_1 - y) + (x_2 - y))$ show that for all $x, y \in E$, $m \in \{1, 2, \dots, M\}$ it holds that

$$\begin{aligned} |\mathcal{Y}_{x,m} - \mathcal{Y}_{y,m}| &= |(X_{x,m} - Y_m)^2 - (X_{y,m} - Y_m)^2| \\ &\leq |X_{x,m} - X_{y,m}|(|X_{x,m} - Y_m| + |X_{y,m} - Y_m|) \\ &\leq 2b|X_{x,m} - X_{y,m}| \leq 2bL\delta(x, y). \end{aligned} \quad (13.46)$$

The fact that for all $x \in E$ it holds that $\mathcal{Y}_{x,m}$, $m \in \{1, 2, \dots, M\}$, are independent random variables, (13.42), and Corollary 13.2.5 (applied with $(Y_{x,m})_{x \in E, m \in \{1, 2, \dots, M\}} \curvearrowright (\mathcal{Y}_{x,m})_{x \in E, m \in \{1, 2, \dots, M\}}$, $L \curvearrowright 2bL$, $(Z_x)_{x \in E} \curvearrowright (\Omega \ni \omega \mapsto \mathcal{R}(x, \omega) \in \mathbb{R})_{x \in E}$ in the notation of Corollary 13.2.5) therefore prove that

(I) it holds that $\Omega \ni \omega \mapsto \sup_{x \in E} |\mathcal{R}(x, \omega) - \mathbf{R}(x)| \in [0, \infty]$ is $\mathcal{F}/\mathcal{B}([0, \infty])$ -measurable and

(II) it holds for all $p \in [2, \infty)$, $c \in (0, \infty)$ that

$$\begin{aligned} (\mathbb{E}[\sup_{x \in E} |\mathcal{R}(x) - \mathbb{E}[\mathcal{R}(x)]|^p])^{1/p} &\leq \frac{2\sqrt{p-1}}{\sqrt{M}} \left(\mathcal{C}^{(E, \delta), \frac{cb^2\sqrt{p-1}}{2bL\sqrt{M}}} \right)^{1/p} [cb^2 \\ &\quad + \sup_{x \in E} \max_{m \in \{1, 2, \dots, M\}} (\mathbb{E}[|\mathcal{Y}_{x,m} - \mathbb{E}[\mathcal{Y}_{x,m}]|^p])^{1/p}]. \end{aligned} \quad (13.47)$$

Observe that item (II), (13.41), (13.42), and (13.45) demonstrate that for all $p \in [2, \infty)$, $c \in (0, \infty)$ it holds that

$$\begin{aligned} (\mathbb{E}[\sup_{x \in E} |\mathcal{R}(x) - \mathbf{R}(x)|^p])^{1/p} &\leq \frac{2\sqrt{p-1}}{\sqrt{M}} \left(\mathcal{C}^{(E, \delta), \frac{cb\sqrt{p-1}}{2L\sqrt{M}}} \right)^{1/p} [cb^2 + b^2] \\ &= \left(\mathcal{C}^{(E, \delta), \frac{cb\sqrt{p-1}}{2L\sqrt{M}}} \right)^{1/p} \left[\frac{2(c+1)b^2\sqrt{p-1}}{\sqrt{M}} \right]. \end{aligned} \quad (13.48)$$

This and item (I) establish items (i) and (ii). The proof of Lemma 13.3.1 is thus complete. \square

Proposition 13.3.2. Let $d \in \mathbb{N}$, $D \subseteq \mathbb{R}^d$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $M \in \mathbb{N}$, let $\mathbb{X}_m = (X_m, Y_m): \Omega \rightarrow (D \times \mathbb{R})$, $m \in \{1, 2, \dots, M\}$, be i.i.d. random variables, let $\alpha \in \mathbb{R}$, $\beta \in (\alpha, \infty)$, $\mathbf{d} \in \mathbb{N}$, let $f = (f_\theta)_{\theta \in [\alpha, \beta]^d}: [\alpha, \beta]^d \rightarrow C(D, \mathbb{R})$, let $L, b \in (0, \infty)$ satisfy for all $\theta, \vartheta \in [\alpha, \beta]^d$, $m \in \{1, 2, \dots, M\}$, $x \in D$ that

$$|f_\theta(X_m) - Y_m| \leq b \quad \text{and} \quad |f_\theta(x) - f_\vartheta(x)| \leq L\|\theta - \vartheta\|_\infty, \quad (13.49)$$

and let $\mathbf{R}: [\alpha, \beta]^d \rightarrow [0, \infty)$ and $\mathcal{R}: [\alpha, \beta]^d \times \Omega \rightarrow [0, \infty)$ satisfy for all $\theta \in [\alpha, \beta]^d$, $\omega \in \Omega$ that

$$\mathbf{R}(\theta) = \mathbb{E}[|f_\theta(X_1) - Y_1|^2] \quad \text{and} \quad \mathcal{R}(\theta, \omega) = \frac{1}{M} \left[\sum_{m=1}^M |f_\theta(X_m(\omega)) - Y_m(\omega)|^2 \right] \quad (13.50)$$

(cf. Definition 3.3.4). Then

(i) it holds that $\Omega \ni \omega \mapsto \sup_{\theta \in [\alpha, \beta]^d} |\mathcal{R}(\theta, \omega) - \mathbf{R}(\theta)| \in [0, \infty]$ is $\mathcal{F}/\mathcal{B}([0, \infty])$ -measurable and

(ii) it holds for all $p \in (0, \infty)$ that

$$\begin{aligned} &(\mathbb{E}[\sup_{\theta \in [\alpha, \beta]^d} |\mathcal{R}(\theta) - \mathbf{R}(\theta)|^p])^{1/p} \\ &\leq \inf_{c, \varepsilon \in (0, \infty)} \left[\frac{2(c+1)b^2 \max\{1, [2\sqrt{M}L(\beta-\alpha)(cb)^{-1}]^\varepsilon\} \sqrt{\max\{1, p, d/\varepsilon\}}}{\sqrt{M}} \right] \\ &\leq \inf_{c \in (0, \infty)} \left[\frac{2(c+1)b^2 \sqrt{e \max\{1, p, d \ln(4ML^2(\beta-\alpha)^2(cb)^{-2})\}}}{\sqrt{M}} \right]. \end{aligned} \quad (13.51)$$

Proof of Proposition 13.3.2. Throughout this proof, let $(\kappa_c)_{c \in (0, \infty)} \subseteq (0, \infty)$ satisfy for all $c \in (0, \infty)$ that

$$\kappa_c = \frac{2\sqrt{M}L(\beta-\alpha)}{cb}, \quad (13.52)$$

let $\mathcal{X}_{\theta,m}: \Omega \rightarrow \mathbb{R}$, $m \in \{1, 2, \dots, M\}$, $\theta \in [\alpha, \beta]^d$, satisfy for all $\theta \in [\alpha, \beta]^d$, $m \in \{1, 2, \dots, M\}$ that

$$\mathcal{X}_{\theta,m} = f_\theta(X_m), \quad (13.53)$$

and let $\delta: [\alpha, \beta]^d \times [\alpha, \beta]^d \rightarrow [0, \infty)$ satisfy for all $\theta, \vartheta \in [\alpha, \beta]^d$ that

$$\delta(\theta, \vartheta) = \|\theta - \vartheta\|_\infty. \quad (13.54)$$

First, note that the assumption that for all $\theta \in [\alpha, \beta]^d$, $m \in \{1, 2, \dots, M\}$ it holds that $|f_\theta(X_m) - Y_m| \leq b$ implies for all $\theta \in [\alpha, \beta]^d$, $m \in \{1, 2, \dots, M\}$ that

$$|\mathcal{X}_{\theta,m} - Y_m| = |f_\theta(X_m) - Y_m| \leq b. \quad (13.55)$$

Furthermore, observe that the assumption that for all $\theta, \vartheta \in [\alpha, \beta]^d$, $x \in D$ it holds that $|f_\theta(x) - f_\vartheta(x)| \leq L\|\theta - \vartheta\|_\infty$ ensures for all $\theta, \vartheta \in [\alpha, \beta]^d$, $m \in \{1, 2, \dots, M\}$ that

$$|\mathcal{X}_{\theta,m} - \mathcal{X}_{\vartheta,m}| = |f_\theta(X_m) - f_\vartheta(X_m)| \leq \sup_{x \in D} |f_\theta(x) - f_\vartheta(x)| \leq L\|\theta - \vartheta\|_\infty = L\delta(\theta, \vartheta). \quad (13.56)$$

The fact that for all $\theta \in [\alpha, \beta]^d$ it holds that $(\mathcal{X}_{\theta,m}, Y_m)$, $m \in \{1, 2, \dots, M\}$, are i.i.d. random variables, (13.55), and Lemma 13.3.1 (applied with $p \curvearrowright q$, $C \curvearrowright C$, $(E, \delta) \curvearrowright ([\alpha, \beta]^d, \delta)$, $(X_{x,m})_{x \in E, m \in \{1, 2, \dots, M\}} \curvearrowright (\mathcal{X}_{\theta,m})_{\theta \in [\alpha, \beta]^d, m \in \{1, 2, \dots, M\}}$ for $p \in [2, \infty)$, $C \in (0, \infty)$ in the notation of Lemma 13.3.1) hence ensure that for all $p \in [2, \infty)$, $c \in (0, \infty)$ it holds that $\Omega \ni \omega \mapsto \sup_{\theta \in [\alpha, \beta]^d} |\mathcal{R}(\theta, \omega) - \mathbf{R}(\theta)| \in [0, \infty]$ is $\mathcal{F}/\mathcal{B}([0, \infty])$ -measurable and

$$(\mathbb{E}[\sup_{\theta \in [\alpha, \beta]^d} |\mathcal{R}(\theta) - \mathbf{R}(\theta)|^p])^{1/p} \leq \left(\mathcal{C}^{([\alpha, \beta]^d, \delta), \frac{cb\sqrt{p-1}}{2L\sqrt{M}}} \right)^{1/p} \left[\frac{2(c+1)b^2\sqrt{p-1}}{\sqrt{M}} \right] \quad (13.57)$$

(cf. Definition 4.3.2). This proves item (i). Note that Proposition 12.2.21 (applied with $d \curvearrowright d$, $a \curvearrowright \alpha$, $b \curvearrowright \beta$, $r \curvearrowright r$ for $r \in (0, \infty)$ in the notation of Proposition 12.2.21) shows that for all $r \in (0, \infty)$ it holds that

$$\begin{aligned} \mathcal{C}^{([\alpha, \beta]^d, \delta), r} &\leq \mathbb{1}_{[0, r]} \left(\frac{\beta-\alpha}{2} \right) + \left(\frac{\beta-\alpha}{r} \right)^d \mathbb{1}_{(r, \infty)} \left(\frac{\beta-\alpha}{2} \right) \\ &\leq \max \left\{ 1, \left(\frac{\beta-\alpha}{r} \right)^d \right\} \left(\mathbb{1}_{[0, r]} \left(\frac{\beta-\alpha}{2} \right) + \mathbb{1}_{(r, \infty)} \left(\frac{\beta-\alpha}{2} \right) \right) \\ &= \max \left\{ 1, \left(\frac{\beta-\alpha}{r} \right)^d \right\}. \end{aligned} \quad (13.58)$$

Therefore, we obtain for all $c \in (0, \infty)$, $p \in [2, \infty)$ that

$$\begin{aligned} \left(\mathcal{C}^{([\alpha, \beta]^d, \delta), \frac{cb\sqrt{p-1}}{2L\sqrt{M}}} \right)^{1/p} &\leq \max \left\{ 1, \left(\frac{2(\beta-\alpha)L\sqrt{M}}{cb\sqrt{p-1}} \right)^{\frac{d}{p}} \right\} \\ &\leq \max \left\{ 1, \left(\frac{2(\beta-\alpha)L\sqrt{M}}{cb} \right)^{\frac{d}{p}} \right\} = \max \left\{ 1, (\kappa_c)^{\frac{d}{p}} \right\}. \end{aligned} \quad (13.59)$$

This, (13.57), and Jensen's inequality demonstrate that for all $c, \varepsilon, p \in (0, \infty)$ it holds that

$$\begin{aligned}
 & (\mathbb{E}[\sup_{\theta \in [\alpha, \beta]^d} |\mathcal{R}(\theta) - \mathbf{R}(\theta)|^p])^{1/p} \\
 & \leq (\mathbb{E}[\sup_{\theta \in [\alpha, \beta]^d} |\mathcal{R}(\theta) - \mathbf{R}(\theta)|^{\max\{2, p, d/\varepsilon\}}])^{\frac{1}{\max\{2, p, d/\varepsilon\}}} \\
 & \leq \max\left\{1, (\kappa_c)^{\frac{d}{\max\{2, p, d/\varepsilon\}}}\right\} \frac{2(c+1)b^2 \sqrt{\max\{2, p, d/\varepsilon\} - 1}}{\sqrt{M}} \\
 & = \max\left\{1, (\kappa_c)^{\min\{d/2, d/p, \varepsilon\}}\right\} \frac{2(c+1)b^2 \sqrt{\max\{1, p-1, d/\varepsilon - 1\}}}{\sqrt{M}} \\
 & \leq \frac{2(c+1)b^2 \max\{1, (\kappa_c)^\varepsilon\} \sqrt{\max\{1, p, d/\varepsilon\}}}{\sqrt{M}}.
 \end{aligned} \tag{13.60}$$

Moreover, observe that the fact that for all $a \in (1, \infty)$ it holds that

$$a^{1/(2 \ln(a))} = e^{\ln(a)/(2 \ln(a))} = e^{1/2} = \sqrt{e} \geq 1 \tag{13.61}$$

establishes that for all $c, p \in (0, \infty)$ with $\kappa_c > 1$ it holds that

$$\begin{aligned}
 & \inf_{\varepsilon \in (0, \infty)} \left[\frac{2(c+1)b^2 \max\{1, (\kappa_c)^\varepsilon\} \sqrt{\max\{1, p, d/\varepsilon\}}}{\sqrt{M}} \right] \\
 & \leq \frac{2(c+1)b^2 \max\{1, (\kappa_c)^{1/(2 \ln(\kappa_c))}\} \sqrt{\max\{1, p, 2d \ln(\kappa_c)\}}}{\sqrt{M}} \\
 & = \frac{2(c+1)b^2 \sqrt{e \max\{1, p, d \ln([\kappa_c]^2)\}}}{\sqrt{M}}.
 \end{aligned} \tag{13.62}$$

The fact that for all $c, p \in (0, \infty)$ with $\kappa_c \leq 1$ it holds that

$$\begin{aligned}
 & \inf_{\varepsilon \in (0, \infty)} \left[\frac{2(c+1)b^2 \max\{1, (\kappa_c)^\varepsilon\} \sqrt{\max\{1, p, d/\varepsilon\}}}{\sqrt{M}} \right] \\
 & = \inf_{\varepsilon \in (0, \infty)} \left[\frac{2(c+1)b^2 \sqrt{\max\{1, p, d/\varepsilon\}}}{\sqrt{M}} \right] \leq \frac{2(c+1)b^2 \sqrt{\max\{1, p\}}}{\sqrt{M}} \\
 & \leq \frac{2(c+1)b^2 \sqrt{e \max\{1, p, d \ln([\kappa_c]^2)\}}}{\sqrt{M}}.
 \end{aligned} \tag{13.63}$$

and (13.60) hence imply that for all $p \in (0, \infty)$ it holds that

$$\begin{aligned}
 & (\mathbb{E}[\sup_{\theta \in [\alpha, \beta]^d} |\mathcal{R}(\theta) - \mathbf{R}(\theta)|^p])^{1/p} \\
 & \leq \inf_{c, \varepsilon \in (0, \infty)} \left[\frac{2(c+1)b^2 \max\{1, (\kappa_c)^\varepsilon\} \sqrt{\max\{1, p, d/\varepsilon\}}}{\sqrt{M}} \right] \\
 & = \inf_{c, \varepsilon \in (0, \infty)} \left[\frac{2(c+1)b^2 \max\{1, [2\sqrt{M}L(\beta-\alpha)(cb)^{-1}]^\varepsilon\} \sqrt{\max\{1, p, d/\varepsilon\}}}{\sqrt{M}} \right] \\
 & \leq \inf_{c \in (0, \infty)} \left[\frac{2(c+1)b^2 \sqrt{e \max\{1, p, d \ln([\kappa_c]^2)\}}}{\sqrt{M}} \right] \\
 & = \inf_{c \in (0, \infty)} \left[\frac{2(c+1)b^2 \sqrt{e \max\{1, p, d \ln(4ML^2(\beta-\alpha)^2(cb)^{-2})\}}}{\sqrt{M}} \right].
 \end{aligned} \tag{13.64}$$

This proves item (ii). The proof of Proposition 13.3.2 is thus complete. \square

Corollary 13.3.3. Let $d, M \in \mathbb{N}$, $b \in [1, \infty)$, $u \in \mathbb{R}$, $v \in [u+1, \infty)$, $D \subseteq [-b, b]^d$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\mathbb{X}_m = (X_m, Y_m): \Omega \rightarrow (D \times [u, v])$, $m \in \{1, 2, \dots, M\}$, be i.i.d. random variables, let $B \in [1, \infty)$, $\mathbf{L}, \mathbf{d} \in \mathbb{N}$, $\mathbf{l} = (\mathbf{l}_0, \mathbf{l}_1, \dots, \mathbf{l}_{\mathbf{L}}) \in \mathbb{N}^{\mathbf{L}+1}$ satisfy $\mathbf{l}_0 = d$, $\mathbf{l}_{\mathbf{L}} = 1$, and $\mathbf{d} \geq \sum_{i=1}^{\mathbf{L}} \mathbf{l}_i (\mathbf{l}_{i-1} + 1)$, let $\mathbf{R}: [-B, B]^{\mathbf{d}} \rightarrow [0, \infty)$ and $\mathcal{R}: [-B, B]^{\mathbf{d}} \times \Omega \rightarrow [0, \infty)$ satisfy for all $\theta \in [-B, B]^{\mathbf{d}}$, $\omega \in \Omega$ that

$$\mathbf{R}(\theta) = \mathbb{E}[|\mathcal{N}_{u,v}^{\theta, \mathbf{l}}(X_1) - Y_1|^2] \quad \text{and} \quad \mathcal{R}(\theta, \omega) = \frac{1}{M} \left[\sum_{m=1}^M |\mathcal{N}_{u,v}^{\theta, \mathbf{l}}(X_m(\omega)) - Y_m(\omega)|^2 \right] \tag{13.65}$$

(cf. Definition 4.4.1). Then

(i) it holds that $\Omega \ni \omega \mapsto \sup_{\theta \in [-B, B]^{\mathbf{d}}} |\mathcal{R}(\theta, \omega) - \mathbf{R}(\theta)| \in [0, \infty]$ is $\mathcal{F}/\mathcal{B}([0, \infty])$ -measurable and

(ii) it holds for all $p \in (0, \infty)$ that

$$\begin{aligned}
 & (\mathbb{E}[\sup_{\theta \in [-B, B]^{\mathbf{d}}} |\mathcal{R}(\theta) - \mathbf{R}(\theta)|^p])^{1/p} \\
 & \leq \frac{9(v-u)^2 \mathbf{L}(\|\mathbf{l}\|_\infty + 1) \sqrt{\max\{p, \ln(4(Mb)^{1/\mathbf{L}}(\|\mathbf{l}\|_\infty + 1)B)\}}}{\sqrt{M}} \\
 & \leq \frac{9(v-u)^2 \mathbf{L}(\|\mathbf{l}\|_\infty + 1)^2 \max\{p, \ln(3MBb)\}}{\sqrt{M}}
 \end{aligned} \tag{13.66}$$

(cf. Definition 3.3.4).

Proof of Corollary 13.3.3. Throughout this proof, let $\mathfrak{d} = \sum_{i=1}^L \mathbf{l}_i(\mathbf{l}_{i-1} + 1) \in \mathbb{N}$, let $L = b\mathbf{L}(\|\mathbf{l}\|_\infty + 1)^{\mathbf{L}}B^{\mathbf{L}-1} \in (0, \infty)$, for every $\theta \in [-B, B]^{\mathfrak{d}}$ let $f_\theta: D \rightarrow \mathbb{R}$ satisfy for all $x \in D$ that

$$f_\theta(x) = \mathcal{N}_{u,v}^{\theta,\mathbf{l}}(x), \quad (13.67)$$

let $\mathcal{R}: [-B, B]^{\mathfrak{d}} \rightarrow [0, \infty)$ satisfy for all $\theta \in [-B, B]^{\mathfrak{d}}$ that

$$\mathcal{R}(\theta) = \mathbb{E}[|f_\theta(X_1) - Y_1|^2] = \mathbb{E}[|\mathcal{N}_{u,v}^{\theta,\mathbf{l}}(X_1) - Y_1|^2], \quad (13.68)$$

and let $R: [-B, B]^{\mathfrak{d}} \times \Omega \rightarrow [0, \infty)$ satisfy for all $\theta \in [-B, B]^{\mathfrak{d}}, \omega \in \Omega$ that

$$R(\theta, \omega) = \frac{1}{M} \left[\sum_{m=1}^M |f_\theta(X_m(\omega)) - Y_m(\omega)|^2 \right] = \frac{1}{M} \left[\sum_{m=1}^M |\mathcal{N}_{u,v}^{\theta,\mathbf{l}}(X_m(\omega)) - Y_m(\omega)|^2 \right] \quad (13.69)$$

(cf. Definition 3.3.4). Note that the fact that for all $\theta \in \mathbb{R}^{\mathfrak{d}}, x \in \mathbb{R}^d$ it holds that $\mathcal{N}_{u,v}^{\theta,\mathbf{l}}(x) \in [u, v]$ and the assumption that for all $m \in \{1, 2, \dots, M\}$ it holds that $Y_m(\Omega) \subseteq [u, v]$ ensure for all $\theta \in [-B, B]^{\mathfrak{d}}, m \in \{1, 2, \dots, M\}$ that

$$|f_\theta(X_m) - Y_m| = |\mathcal{N}_{u,v}^{\theta,\mathbf{l}}(X_m) - Y_m| \leq \sup_{y_1, y_2 \in [u, v]} |y_1 - y_2| = v - u. \quad (13.70)$$

Furthermore, observe that the assumption that $D \subseteq [-b, b]^d$, $\mathbf{l}_0 = d$, and $\mathbf{l}_{\mathbf{L}} = 1$, Corollary 11.3.7 (applied with $a \curvearrowright -b$, $b \curvearrowright b$, $u \curvearrowright u$, $v \curvearrowright v$, $d \curvearrowright \mathfrak{d}$, $L \curvearrowright \mathbf{L}$, $l \curvearrowright \mathbf{l}$ in the notation of Corollary 11.3.7), and the assumption that $b \geq 1$ and $B \geq 1$ show that for all $\theta, \vartheta \in [-B, B]^{\mathfrak{d}}, x \in D$ it holds that

$$\begin{aligned} |f_\theta(x) - f_\vartheta(x)| &\leq \sup_{y \in [-b, b]^d} |\mathcal{N}_{u,v}^{\theta,\mathbf{l}}(y) - \mathcal{N}_{u,v}^{\vartheta,\mathbf{l}}(y)| \\ &\leq \mathbf{L} \max\{1, b\} (\|\mathbf{l}\|_\infty + 1)^{\mathbf{L}} (\max\{1, \|\theta\|_\infty, \|\vartheta\|_\infty\})^{\mathbf{L}-1} \|\theta - \vartheta\|_\infty \\ &\leq b\mathbf{L} (\|\mathbf{l}\|_\infty + 1)^{\mathbf{L}} B^{\mathbf{L}-1} \|\theta - \vartheta\|_\infty = L \|\theta - \vartheta\|_\infty. \end{aligned} \quad (13.71)$$

Moreover, note that the fact that $\mathbf{d} \geq \mathfrak{d}$ and the fact that for all $\theta = (\theta_1, \theta_2, \dots, \theta_{\mathbf{d}}) \in \mathbb{R}^{\mathbf{d}}$ it holds that $\mathcal{N}_{u,v}^{\theta,\mathbf{l}} = \mathcal{N}_{u,v}^{(\theta_1, \theta_2, \dots, \theta_{\mathfrak{d}}), \mathbf{l}}$ demonstrates that for all $\omega \in \Omega$ it holds that

$$\sup_{\theta \in [-B, B]^{\mathbf{d}}} |\mathcal{R}(\theta, \omega) - \mathbf{R}(\theta)| = \sup_{\theta \in [-B, B]^{\mathfrak{d}}} |R(\theta, \omega) - \mathcal{R}(\theta)|. \quad (13.72)$$

In addition, observe that (13.70), (13.71), Proposition 13.3.2 (applied with $\alpha \curvearrowright -B$, $\beta \curvearrowright B$, $\mathbf{d} \curvearrowright \mathfrak{d}$, $b \curvearrowright v - u$, $\mathbf{R} \curvearrowright \mathcal{R}$, $\mathcal{R} \curvearrowright R$ in the notation of Proposition 13.3.2), the fact that

$$v - u \geq (u + 1) - u = 1 \quad (13.73)$$

and the fact that

$$\mathfrak{d} \leq \mathbf{L} \|\mathbf{l}\|_\infty (\|\mathbf{l}\|_\infty + 1) \leq \mathbf{L} (\|\mathbf{l}\|_\infty + 1)^2 \quad (13.74)$$

establish that for all $p \in (0, \infty)$ it holds that $\Omega \ni \omega \mapsto \sup_{\theta \in [-B, B]^{\mathfrak{d}}} |R(\theta, \omega) - \mathcal{R}(\theta)| \in [0, \infty]$ is $\mathcal{F}/\mathcal{B}([0, \infty])$ -measurable and

$$\begin{aligned} & (\mathbb{E}[\sup_{\theta \in [-B, B]^{\mathfrak{d}}} |R(\theta) - \mathcal{R}(\theta)|^p])^{1/p} \\ & \leq \inf_{C \in (0, \infty)} \left[\frac{2(C+1)(v-u)^2 \sqrt{e \max\{1, p, \mathfrak{d} \ln(4ML^2(2B)^2(C[v-u])^{-2})\}}}{\sqrt{M}} \right] \\ & \leq \inf_{C \in (0, \infty)} \left[\frac{2(C+1)(v-u)^2 \sqrt{e \max\{1, p, \mathbf{L}(\|\mathbf{l}\|_\infty + 1)^2 \ln(2^4ML^2B^2C^{-2})\}}}{\sqrt{M}} \right]. \end{aligned} \quad (13.75)$$

Combining this with (13.72) proves item (i). Note that (13.72), (13.75), the fact that $2^6\mathbf{L}^2 \leq 2^6 \cdot 2^{2(\mathbf{L}-1)} = 2^{4+2\mathbf{L}} \leq 2^{4\mathbf{L}+2\mathbf{L}} = 2^{6\mathbf{L}}$, the fact that $3 \geq e$, and the assumption that $B \geq 1$, $\mathbf{L} \geq 1$, $M \geq 1$, and $b \geq 1$ imply that for all $p \in (0, \infty)$ it holds that

$$\begin{aligned} & (\mathbb{E}[\sup_{\theta \in [-B, B]^{\mathfrak{d}}} |\mathcal{R}(\theta) - \mathbf{R}(\theta)|^p])^{1/p} = (\mathbb{E}[\sup_{\theta \in [-B, B]^{\mathfrak{d}}} |R(\theta) - \mathcal{R}(\theta)|^p])^{1/p} \\ & \leq \frac{2(1/2+1)(v-u)^2 \sqrt{e \max\{1, p, \mathbf{L}(\|\mathbf{l}\|_\infty + 1)^2 \ln(2^4ML^2B^22^2)\}}}{\sqrt{M}} \\ & = \frac{3(v-u)^2 \sqrt{e \max\{p, \mathbf{L}(\|\mathbf{l}\|_\infty + 1)^2 \ln(2^6Mb^2\mathbf{L}^2(\|\mathbf{l}\|_\infty + 1)^{2\mathbf{L}}B^{2\mathbf{L}})\}}}{\sqrt{M}} \\ & \leq \frac{3(v-u)^2 \sqrt{e \max\{p, 3\mathbf{L}^2(\|\mathbf{l}\|_\infty + 1)^2 \ln([2^{6\mathbf{L}}Mb^2(\|\mathbf{l}\|_\infty + 1)^{2\mathbf{L}}B^{2\mathbf{L}}]^{1/(3\mathbf{L})})\}}}{\sqrt{M}} \\ & \leq \frac{3(v-u)^2 \sqrt{3 \max\{p, 3\mathbf{L}^2(\|\mathbf{l}\|_\infty + 1)^2 \ln(2^2(Mb^2)^{1/(3\mathbf{L})}(\|\mathbf{l}\|_\infty + 1)B)\}}}{\sqrt{M}} \\ & \leq \frac{9(v-u)^2\mathbf{L}(\|\mathbf{l}\|_\infty + 1) \sqrt{\max\{p, \ln(4(Mb)^{1/\mathbf{L}}(\|\mathbf{l}\|_\infty + 1)B)\}}}{\sqrt{M}}. \end{aligned} \quad (13.76)$$

Next observe that the fact that for all $n \in \mathbb{N}$ it holds that $n \leq 2^{n-1}$ and the fact that $\|\mathbf{l}\|_\infty \geq 1$ ensure that

$$4(\|\mathbf{l}\|_\infty + 1) \leq 2^2 \cdot 2^{(\|\mathbf{l}\|_\infty + 1)-1} = 2^3 \cdot 2^{(\|\mathbf{l}\|_\infty + 1)-2} \leq 3^2 \cdot 3^{(\|\mathbf{l}\|_\infty + 1)-2} = 3^{(\|\mathbf{l}\|_\infty + 1)}. \quad (13.77)$$

Therefore, we obtain that for all $p \in (0, \infty)$ it holds that

$$\begin{aligned} & \frac{9(v-u)^2\mathbf{L}(\|\mathbf{l}\|_\infty + 1) \sqrt{\max\{p, \ln(4(Mb)^{1/\mathbf{L}}(\|\mathbf{l}\|_\infty + 1)B)\}}}{\sqrt{M}} \\ & \leq \frac{9(v-u)^2\mathbf{L}(\|\mathbf{l}\|_\infty + 1) \sqrt{\max\{p, (\|\mathbf{l}\|_\infty + 1) \ln([3^{(\|\mathbf{l}\|_\infty + 1)}(Mb)^{1/\mathbf{L}}B]^{1/(\|\mathbf{l}\|_\infty + 1)})\}}}{\sqrt{M}} \\ & \leq \frac{9(v-u)^2\mathbf{L}(\|\mathbf{l}\|_\infty + 1)^2 \max\{p, \ln(3MBb)\}}{\sqrt{M}}. \end{aligned} \quad (13.78)$$

This and (13.76) establish item (ii). The proof of Corollary 13.3.3 is thus complete. \square

Part V

Composed error analysis

Chapter 14

Overall error decomposition

In Chapter 15 below we 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 training of ANNs in the specific situation of GD-type optimization methods with many independent random initializations. For such a combined error analysis we employ a suitable overall error decomposition for supervised learning problems. It is the subject of this chapter to review and derive this overall error decomposition (see Proposition 14.2.1 below).

In the literature such kind of error decompositions can, for instance, be found in [25, 35, 36, 90, 243]. The specific presentation of this chapter is strongly based on [25, Section 4.1] and [243, Section 6.1].

14.1 Bias-variance decomposition

Lemma 14.1.1 (Bias-variance decomposition). *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let (S, \mathcal{S}) be a measurable space, let $X: \Omega \rightarrow S$ and $Y: \Omega \rightarrow \mathbb{R}$ be random variables with $\mathbb{E}[|Y|^2] < \infty$, and let $\mathbf{r}: \mathcal{L}^2(\mathbb{P}_X; \mathbb{R}) \rightarrow [0, \infty)$ satisfy for all $f \in \mathcal{L}^2(\mathbb{P}_X; \mathbb{R})$ that*

$$\mathbf{r}(f) = \mathbb{E}[|f(X) - Y|^2]. \quad (14.1)$$

Then

(i) it holds for all $f \in \mathcal{L}^2(\mathbb{P}_X; \mathbb{R})$ that

$$\mathbf{r}(f) = \mathbb{E}[|f(X) - \mathbb{E}[Y|X]|^2] + \mathbb{E}[|Y - \mathbb{E}[Y|X]|^2], \quad (14.2)$$

(ii) it holds for all $f, g \in \mathcal{L}^2(\mathbb{P}_X; \mathbb{R})$ that

$$\mathbf{r}(f) - \mathbf{r}(g) = \mathbb{E}[|f(X) - \mathbb{E}[Y|X]|^2] - \mathbb{E}[|g(X) - \mathbb{E}[Y|X]|^2], \quad (14.3)$$

and

(iii) it holds for all $f, g \in \mathcal{L}^2(\mathbb{P}_X; \mathbb{R})$ that

$$\mathbb{E}[|f(X) - \mathbb{E}[Y|X]|^2] = \mathbb{E}[|g(X) - \mathbb{E}[Y|X]|^2] + (\mathbf{r}(f) - \mathbf{r}(g)). \quad (14.4)$$

Proof of Lemma 14.1.1. First, note that (14.1) shows that for all $f \in \mathcal{L}^2(\mathbb{P}_X; \mathbb{R})$ it holds that

$$\begin{aligned} \mathbf{r}(f) &= \mathbb{E}[|f(X) - Y|^2] = \mathbb{E}[|(f(X) - \mathbb{E}[Y|X]) + (\mathbb{E}[Y|X] - Y)|^2] \\ &= \mathbb{E}[|f(X) - \mathbb{E}[Y|X]|^2] + 2\mathbb{E}[(f(X) - \mathbb{E}[Y|X])(\mathbb{E}[Y|X] - Y)] \\ &\quad + \mathbb{E}[|\mathbb{E}[Y|X] - Y|^2] \end{aligned} \quad (14.5)$$

Furthermore, observe that the tower rule demonstrates that for all $f \in \mathcal{L}^2(\mathbb{P}_X; \mathbb{R})$ it holds that

$$\begin{aligned} &\mathbb{E}[(f(X) - \mathbb{E}[Y|X])(\mathbb{E}[Y|X] - Y)] \\ &= \mathbb{E}\left[\mathbb{E}[(f(X) - \mathbb{E}[Y|X])(\mathbb{E}[Y|X] - Y)|X]\right] \\ &= \mathbb{E}\left[(f(X) - \mathbb{E}[Y|X])\mathbb{E}[(\mathbb{E}[Y|X] - Y)|X]\right] \\ &= \mathbb{E}[(f(X) - \mathbb{E}[Y|X])(\mathbb{E}[Y|X] - \mathbb{E}[Y|X])] = 0. \end{aligned} \quad (14.6)$$

Combining this with (14.5) proves that for all $f \in \mathcal{L}^2(\mathbb{P}_X; \mathbb{R})$ it holds that

$$\mathbf{r}(f) = \mathbb{E}[|f(X) - \mathbb{E}[Y|X]|^2] + \mathbb{E}[|\mathbb{E}[Y|X] - Y|^2]. \quad (14.7)$$

This implies that for all $f, g \in \mathcal{L}^2(\mathbb{P}_X; \mathbb{R})$ it holds that

$$\mathbf{r}(f) - \mathbf{r}(g) = \mathbb{E}[|f(X) - \mathbb{E}[Y|X]|^2] - \mathbb{E}[|g(X) - \mathbb{E}[Y|X]|^2]. \quad (14.8)$$

Hence, we obtain that for all $f, g \in \mathcal{L}^2(\mathbb{P}_X; \mathbb{R})$ it holds that

$$\mathbb{E}[|f(X) - \mathbb{E}[Y|X]|^2] = \mathbb{E}[|g(X) - \mathbb{E}[Y|X]|^2] + \mathbf{r}(f) - \mathbf{r}(g). \quad (14.9)$$

Combining this with (14.7) and (14.8) establishes items (i), (ii), and (iii). The proof of Lemma 14.1.1 is thus complete. \square

14.1.1 Risk minimization for measurable functions

Proposition 14.1.2. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let (S, \mathcal{S}) be a measurable space, let $X: \Omega \rightarrow S$ and $Y: \Omega \rightarrow \mathbb{R}$ be random variables, assume $\mathbb{E}[|Y|^2] < \infty$, let $\mathcal{E}: \mathcal{L}^2(\mathbb{P}_X; \mathbb{R}) \rightarrow [0, \infty)$ satisfy for all $f \in \mathcal{L}^2(\mathbb{P}_X; \mathbb{R})$ that

$$\mathcal{E}(f) = \mathbb{E}[|f(X) - Y|^2]. \quad (14.10)$$

Then

$$\begin{aligned} & \{f \in \mathcal{L}^2(\mathbb{P}_X; \mathbb{R}): \mathcal{E}(f) = \inf_{g \in \mathcal{L}^2(\mathbb{P}_X; \mathbb{R})} \mathcal{E}(g)\} \\ &= \{f \in \mathcal{L}^2(\mathbb{P}_X; \mathbb{R}): \mathcal{E}(f) = \mathbb{E}[|\mathbb{E}[Y|X] - Y|^2]\} \\ &= \{f \in \mathcal{L}^2(\mathbb{P}_X; \mathbb{R}): f(X) = \mathbb{E}[Y|X] \text{ } \mathbb{P}\text{-a.s.}\}. \end{aligned} \quad (14.11)$$

Proof of Proposition 14.1.2. Note that Lemma 14.1.1 ensures that for all $g \in \mathcal{L}^2(\mathbb{P}_X; \mathbb{R})$ it holds that

$$\mathcal{E}(g) = \mathbb{E}[|g(X) - \mathbb{E}[Y|X]|^2] + \mathbb{E}[|\mathbb{E}[Y|X] - Y|^2]. \quad (14.12)$$

Therefore, we obtain that for all $g \in \mathcal{L}^2(\mathbb{P}_X; \mathbb{R})$ it holds that

$$\mathcal{E}(g) \geq \mathbb{E}[|\mathbb{E}[Y|X] - Y|^2]. \quad (14.13)$$

Furthermore, observe that (14.12) shows that

$$\begin{aligned} & \{f \in \mathcal{L}^2(\mathbb{P}_X; \mathbb{R}): \mathcal{E}(f) = \mathbb{E}[|\mathbb{E}[Y|X] - Y|^2]\} \\ &= \{f \in \mathcal{L}^2(\mathbb{P}_X; \mathbb{R}): \mathbb{E}[|f(X) - \mathbb{E}[Y|X]|^2] = 0\} \\ &= \{f \in \mathcal{L}^2(\mathbb{P}_X; \mathbb{R}): f(X) = \mathbb{E}[Y|X] \text{ } \mathbb{P}\text{-a.s.}\}. \end{aligned} \quad (14.14)$$

Combining this with (14.13) proves (14.11). The proof of Proposition 14.1.2 is thus complete. \square

Corollary 14.1.3. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let (S, \mathcal{S}) be a measurable space, let $X: \Omega \rightarrow S$ be a random variable, let $\mathcal{M} = \{(f: S \rightarrow \mathbb{R}): f \text{ is } \mathcal{S}/\mathcal{B}(\mathbb{R})\text{-measurable}\}$, let $\varphi \in \mathcal{M}$, and let $\mathcal{E}: \mathcal{M} \rightarrow [0, \infty)$ satisfy for all $f \in \mathcal{M}$ that

$$\mathcal{E}(f) = \mathbb{E}[|f(X) - \varphi(X)|^2]. \quad (14.15)$$

Then

$$\begin{aligned} \{f \in \mathcal{M}: \mathcal{E}(f) = \inf_{g \in \mathcal{M}} \mathcal{E}(g)\} &= \{f \in \mathcal{M}: \mathcal{E}(f) = 0\} \\ &= \{f \in \mathcal{M}: \mathbb{P}(f(X) = \varphi(X)) = 1\}. \end{aligned} \quad (14.16)$$

Proof of Corollary 14.1.3. Note that (14.15) demonstrates that $\mathcal{E}(\varphi) = 0$. Hence, we obtain that

$$\inf_{g \in \mathcal{M}} \mathcal{E}(g) = 0. \quad (14.17)$$

Furthermore, observe that

$$\begin{aligned} \{f \in \mathcal{M}: \mathcal{E}(f) = 0\} &= \{f \in \mathcal{M}: \mathbb{E}[|f(X) - \varphi(X)|^2] = 0\} \\ &= \{f \in \mathcal{M}: \mathbb{P}(\{\omega \in \Omega: f(X(\omega)) \neq \varphi(X(\omega))\}) = 0\} \\ &= \{f \in \mathcal{M}: \mathbb{P}(X^{-1}(\{x \in S: f(x) \neq \varphi(x)\})) = 0\} \\ &= \{f \in \mathcal{M}: \mathbb{P}_X(\{x \in S: f(x) \neq \varphi(x)\}) = 0\}. \end{aligned} \quad (14.18)$$

The proof of Corollary 14.1.3 is thus complete. \square

14.2 Overall error decomposition

Proposition 14.2.1. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $M, d \in \mathbb{N}$, $D \subseteq \mathbb{R}^d$, $u \in \mathbb{R}$, $v \in (u, \infty)$, for every $j \in \{1, 2, \dots, M\}$ let $X_j: \Omega \rightarrow D$ and $Y_j: \Omega \rightarrow [u, v]$ be random variables, let $\mathbf{R}: \mathbb{R}^d \rightarrow \mathbb{R}$ satisfy for all $\theta \in \mathbb{R}^d$ that

$$\mathbf{R}(\theta) = \mathbb{E}[|\mathcal{N}_{u,v}^{\theta,1}(X_1) - Y_1|^2], \quad (14.19)$$

let $\mathbf{d}, \mathbf{L} \in \mathbb{N}$, $\mathbf{l} = (l_0, l_1, \dots, l_L) \in \mathbb{N}^{\mathbf{L}+1}$ satisfy

$$l_0 = d, \quad l_L = 1, \quad \text{and} \quad \mathbf{d} \geq \sum_{i=1}^L l_i(l_{i-1} + 1), \quad (14.20)$$

let $\mathcal{R}: \mathbb{R}^d \times \Omega \rightarrow \mathbb{R}$ satisfy for all $\theta \in \mathbb{R}^d$ that

$$\mathcal{R}(\theta) = \frac{1}{M} \left[\sum_{j=1}^M |\mathcal{N}_{u,v}^{\theta,1}(X_j) - Y_j|^2 \right], \quad (14.21)$$

let $\mathcal{E}: D \rightarrow [u, v]$ be $\mathcal{B}(D)/\mathcal{B}([u, v])$ -measurable, assume \mathbb{P} -a.s. that

$$\mathcal{E}(X_1) = \mathbb{E}[Y_1 | X_1], \quad (14.22)$$

let $B \in [0, \infty)$, for every $k, n \in \mathbb{N}_0$ let $\Theta_{k,n}: \Omega \rightarrow \mathbb{R}^d$ be a function, let $K, N \in \mathbb{N}$, $\mathbf{T} \subseteq \{0, 1, \dots, N\}$, let $\mathbf{k}: \Omega \rightarrow (\mathbb{N}_0)^2$ satisfy for all $\omega \in \Omega$ that

$$\mathbf{k}(\omega) \in \{(k, n) \in \{1, 2, \dots, K\} \times \mathbf{T}: \|\Theta_{k,n}(\omega)\|_\infty \leq B\} \quad (14.23)$$

$$\text{and} \quad \mathcal{R}(\Theta_{\mathbf{k}(\omega)}(\omega)) = \min_{(k,n) \in \{1,2,\dots,K\} \times \mathbf{T}, \|\Theta_{k,n}(\omega)\|_\infty \leq B} \mathcal{R}(\Theta_{k,n}(\omega)) \quad (14.24)$$

(cf. Definitions 3.3.4 and 4.4.1). Then it holds for all $\vartheta \in [-B, B]^d$ that

$$\begin{aligned} & \int_D |\mathcal{N}_{u,v}^{\Theta_k,1}(x) - \mathcal{E}(x)|^2 \mathbb{P}_{X_1}(dx) \\ & \leq [\sup_{x \in D} |\mathcal{N}_{u,v}^{\vartheta,1}(x) - \mathcal{E}(x)|^2] + 2[\sup_{\theta \in [-B, B]^d} |\mathcal{R}(\theta) - \mathbf{R}(\theta)|] \\ & \quad + \min_{(k,n) \in \{1, 2, \dots, K\} \times \mathbf{T}, \|\Theta_{k,n}\|_\infty \leq B} [\mathcal{R}(\Theta_{k,n}) - \mathcal{R}(\vartheta)]. \end{aligned} \quad (14.25)$$

Proof of Proposition 14.2.1. Throughout this proof, let $\mathbf{r}: \mathcal{L}^2(\mathbb{P}_{X_1}; \mathbb{R}) \rightarrow [0, \infty)$ satisfy for all $f \in \mathcal{L}^2(\mathbb{P}_{X_1}; \mathbb{R})$ that

$$\mathbf{r}(f) = \mathbb{E}[|f(X_1) - Y_1|^2]. \quad (14.26)$$

Observe that the assumption that for all $\omega \in \Omega$ it holds that $Y_1(\omega) \in [u, v]$ and the fact that for all $\theta \in \mathbb{R}^d$, $x \in \mathbb{R}^d$ it holds that $\mathcal{N}_{u,v}^{\theta,1}(x) \in [u, v]$ imply that for all $\theta \in \mathbb{R}^d$ it holds that $\mathbb{E}[|Y_1|^2] \leq \max\{u^2, v^2\} < \infty$ and

$$\int_D |\mathcal{N}_{u,v}^{\theta,1}(x)|^2 \mathbb{P}_{X_1}(dx) = \mathbb{E}[|\mathcal{N}_{u,v}^{\theta,1}(X_1)|^2] \leq \max\{u^2, v^2\} < \infty. \quad (14.27)$$

Item (iii) in Lemma 14.1.1 (applied with $(\Omega, \mathcal{F}, \mathbb{P}) \curvearrowleft (\Omega, \mathcal{F}, \mathbb{P})$, $(S, \mathcal{S}) \curvearrowleft (D, \mathcal{B}(D))$, $X \curvearrowleft X_1$, $Y \curvearrowleft (\Omega \ni \omega \mapsto Y_1(\omega) \in \mathbb{R})$, $\mathbf{r} \curvearrowleft \mathbf{r}$, $f \curvearrowleft \mathcal{N}_{u,v}^{\theta,1}|_D$, $g \curvearrowleft \mathcal{N}_{u,v}^{\vartheta,1}|_D$ for $\theta, \vartheta \in \mathbb{R}^d$ in the notation of item (iii) in Lemma 14.1.1) therefore establishes that for all $\theta, \vartheta \in \mathbb{R}^d$ it holds that

$$\begin{aligned} & \int_D |\mathcal{N}_{u,v}^{\theta,1}(x) - \mathcal{E}(x)|^2 \mathbb{P}_{X_1}(dx) \\ & = \mathbb{E}[|\mathcal{N}_{u,v}^{\theta,1}(X_1) - \mathcal{E}(X_1)|^2] = \mathbb{E}[|\mathcal{N}_{u,v}^{\theta,1}(X_1) - \mathbb{E}[Y_1|X_1]|^2] \\ & = \mathbb{E}[|\mathcal{N}_{u,v}^{\vartheta,1}(X_1) - \mathbb{E}[Y_1|X_1]|^2] + \mathbf{r}(\mathcal{N}_{u,v}^{\theta,1}|_D) - \mathbf{r}(\mathcal{N}_{u,v}^{\vartheta,1}|_D) \end{aligned} \quad (14.28)$$

Combining this with (14.26) and (14.19) ensures that for all $\theta, \vartheta \in \mathbb{R}^d$ it holds that

$$\begin{aligned} & \int_D |\mathcal{N}_{u,v}^{\theta,1}(x) - \mathcal{E}(x)|^2 \mathbb{P}_{X_1}(dx) \\ & = \mathbb{E}[|\mathcal{N}_{u,v}^{\vartheta,1}(X_1) - \mathcal{E}(X_1)|^2] + \mathbb{E}[|\mathcal{N}_{u,v}^{\theta,1}(X_1) - Y_1|^2] - \mathbb{E}[|\mathcal{N}_{u,v}^{\vartheta,1}(X_1) - Y_1|^2] \\ & = \int_D |\mathcal{N}_{u,v}^{\vartheta,1}(x) - \mathcal{E}(x)|^2 \mathbb{P}_{X_1}(dx) + \mathbf{R}(\theta) - \mathbf{R}(\vartheta). \end{aligned} \quad (14.29)$$

This shows that for all $\theta, \vartheta \in \mathbb{R}^d$ it holds that

$$\begin{aligned}
 & \int_D |\mathcal{N}_{u,v}^{\theta,1}(x) - \mathcal{E}(x)|^2 \mathbb{P}_{X_1}(dx) \\
 &= \int_D |\mathcal{N}_{u,v}^{\vartheta,1}(x) - \mathcal{E}(x)|^2 \mathbb{P}_{X_1}(dx) - [\mathcal{R}(\theta) - \mathbf{R}(\theta)] + \mathcal{R}(\vartheta) - \mathbf{R}(\vartheta) \\
 &\quad + \mathcal{R}(\theta) - \mathcal{R}(\vartheta) \\
 &\leq \int_D |\mathcal{N}_{u,v}^{\vartheta,1}(x) - \mathcal{E}(x)|^2 \mathbb{P}_{X_1}(dx) + 2[\max_{\eta \in \{\theta, \vartheta\}} |\mathcal{R}(\eta) - \mathbf{R}(\eta)|] \\
 &\quad + \mathcal{R}(\theta) - \mathcal{R}(\vartheta).
 \end{aligned} \tag{14.30}$$

Furthermore, note that (14.23) implies that for all $\omega \in \Omega$ it holds that $\Theta_{k(\omega)}(\omega) \in [-B, B]^d$. Combining (14.30) with (14.24) hence proves that for all $\vartheta \in [-B, B]^d$ it holds that

$$\begin{aligned}
 & \int_D |\mathcal{N}_{u,v}^{\Theta_k,1}(x) - \mathcal{E}(x)|^2 \mathbb{P}_{X_1}(dx) \\
 &\leq \int_D |\mathcal{N}_{u,v}^{\vartheta,1}(x) - \mathcal{E}(x)|^2 \mathbb{P}_{X_1}(dx) + 2[\sup_{\theta \in [-B, B]^d} |\mathcal{R}(\theta) - \mathbf{R}(\theta)|] \\
 &\quad + \mathcal{R}(\Theta_k) - \mathcal{R}(\vartheta) \\
 &= \int_D |\mathcal{N}_{u,v}^{\vartheta,1}(x) - \mathcal{E}(x)|^2 \mathbb{P}_{X_1}(dx) + 2[\sup_{\theta \in [-B, B]^d} |\mathcal{R}(\theta) - \mathbf{R}(\theta)|] \\
 &\quad + \min_{(k,n) \in \{1,2,\dots,K\} \times \mathbf{T}, \|\Theta_{k,n}\|_\infty \leq B} [\mathcal{R}(\Theta_{k,n}) - \mathcal{R}(\vartheta)] \\
 &\leq [\sup_{x \in D} |\mathcal{N}_{u,v}^{\vartheta,1}(x) - \mathcal{E}(x)|^2] + 2[\sup_{\theta \in [-B, B]^d} |\mathcal{R}(\theta) - \mathbf{R}(\theta)|] \\
 &\quad + \min_{(k,n) \in \{1,2,\dots,K\} \times \mathbf{T}, \|\Theta_{k,n}\|_\infty \leq B} [\mathcal{R}(\Theta_{k,n}) - \mathcal{R}(\vartheta)].
 \end{aligned} \tag{14.31}$$

The proof of Proposition 14.2.1 is thus complete. \square

Chapter 15

Composed error estimates

In Part II we have established several estimates for the approximation error, in Part III we have established several estimates for the optimization error, and in Part IV we have established several estimates for the generalization error. In this chapter we employ the error decomposition from Chapter 14 as well as parts of Parts II, III, and IV (see Proposition 4.4.12 and Corollaries 11.3.9 and 13.3.3) to establish estimates for the overall error in the training of ANNs in the specific situation of GD-type optimization methods with many independent random initializations.

In the literature such overall error analyses can, for example, be found in [25, 238, 243]. The material in this chapter consist of slightly modified extracts from Jentzen & Welti [243, Sections 6.2 and 6.3].

15.1 Full strong error analysis for the training of ANNs

Lemma 15.1.1. *Let $d, \mathbf{d}, \mathbf{L} \in \mathbb{N}$, $\mathbf{l} = (\mathbf{l}_0, \mathbf{l}_1, \dots, \mathbf{l}_{\mathbf{L}}) \in \mathbb{N}^{\mathbf{L}+1}$, $u \in [-\infty, \infty)$, $v \in (u, \infty]$, let $D \subseteq \mathbb{R}^d$, assume*

$$\mathbf{l}_0 = d, \quad \mathbf{l}_{\mathbf{L}} = 1, \quad \text{and} \quad \mathbf{d} \geq \sum_{i=1}^{\mathbf{L}} \mathbf{l}_i(\mathbf{l}_{i-1} + 1), \quad (15.1)$$

let $\mathcal{E}: D \rightarrow \mathbb{R}$ be $\mathcal{B}(D)/\mathcal{B}(\mathbb{R})$ -measurable, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $X: \Omega \rightarrow D$, $\mathbf{k}: \Omega \rightarrow (\mathbb{N}_0)^2$, and $\Theta_{k,n}: \Omega \rightarrow \mathbb{R}^{\mathbf{d}}$, $k, n \in \mathbb{N}_0$, be random variables. Then

- (i) *it holds that $\mathbb{R}^{\mathbf{d}} \times \mathbb{R}^d \ni (\theta, x) \mapsto \mathcal{N}_{u,v}^{\theta,1}(x) \in \mathbb{R}$ is $(\mathcal{B}(\mathbb{R}^{\mathbf{d}}) \otimes \mathcal{B}(\mathbb{R}^d))/\mathcal{B}(\mathbb{R})$ -measurable,*
- (ii) *it holds for all $\omega \in \Omega$ that $\mathbb{R}^d \ni x \mapsto \mathcal{N}_{u,v}^{\Theta_{\mathbf{k}(\omega)}(\omega),1}(x) \in \mathbb{R}$ is $\mathcal{B}(\mathbb{R}^d)/\mathcal{B}(\mathbb{R})$ -mesaurable, and*

(iii) it holds for all $p \in [0, \infty)$ that

$$\Omega \ni \omega \mapsto \int_D |\mathcal{N}_{u,v}^{\Theta_{\mathbf{k}(\omega)}(\omega), \mathbf{l}}(x) - \mathcal{E}(x)|^p \mathbb{P}_X(dx) \in [0, \infty] \quad (15.2)$$

is $\mathcal{F}/\mathcal{B}([0, \infty])$ -measurable

(cf. Definition 4.4.1).

Proof of Lemma 15.1.1. Throughout this proof let $\Xi: \Omega \rightarrow \mathbb{R}^d$ satisfy for all $\omega \in \Omega$ that

$$\Xi(\omega) = \Theta_{\mathbf{k}(\omega)}(\omega). \quad (15.3)$$

Observe that the assumption that $\Theta_{k,n}: \Omega \rightarrow \mathbb{R}^d$, $k, n \in \mathbb{N}_0$, and $\mathbf{k}: \Omega \rightarrow (\mathbb{N}_0)^2$ are random variables demonstrates that for all $U \in \mathcal{B}(\mathbb{R}^d)$ it holds that

$$\begin{aligned} \Xi^{-1}(U) &= \{\omega \in \Omega: \Xi(\omega) \in U\} = \{\omega \in \Omega: \Theta_{\mathbf{k}(\omega)}(\omega) \in U\} \\ &= \{\omega \in \Omega: [\exists k, n \in \mathbb{N}_0: ([\Theta_{k,n}(\omega) \in U] \wedge [\mathbf{k}(\omega) = (k, n)])]\} \\ &= \bigcup_{k=0}^{\infty} \bigcup_{n=0}^{\infty} (\{\omega \in \Omega: \Theta_{k,n}(\omega) \in U\} \cap \{\omega \in \Omega: \mathbf{k}(\omega) = (k, n)\}) \\ &= \bigcup_{k=0}^{\infty} \bigcup_{n=0}^{\infty} ([\Theta_{k,n}]^{-1}(U) \cap [\mathbf{k}^{-1}(\{(k, n)\})]) \in \mathcal{F}. \end{aligned} \quad (15.4)$$

This establishes that

$$\Omega \ni \omega \mapsto \Theta_{\mathbf{k}(\omega)}(\omega) \in \mathbb{R}^d \quad (15.5)$$

is $\mathcal{F}/\mathcal{B}(\mathbb{R}^d)$ -measurable. Furthermore, note that Corollary 11.3.7 (applied with $a \curvearrowright -\|x\|_\infty$, $b \curvearrowright \|x\|_\infty$, $u \curvearrowright u$, $v \curvearrowright v$, $d \curvearrowright \mathbf{d}$, $L \curvearrowright \mathbf{L}$, $l \curvearrowright \mathbf{l}$ for $x \in \mathbb{R}^d$ in the notation of Corollary 11.3.7) ensures that for all $\theta, \vartheta \in \mathbb{R}^d$, $x \in \mathbb{R}^d$ it holds that

$$\begin{aligned} |\mathcal{N}_{u,v}^{\theta, \mathbf{l}}(x) - \mathcal{N}_{u,v}^{\vartheta, \mathbf{l}}(x)| &\leq \sup_{y \in [-\|x\|_\infty, \|x\|_\infty]^d} |\mathcal{N}_{u,v}^{\theta, \mathbf{l}}(y) - \mathcal{N}_{u,v}^{\vartheta, \mathbf{l}}(y)| \\ &\leq \mathbf{L} \max\{1, \|x\|_\infty\} (\|\mathbf{l}\|_\infty + 1)^{\mathbf{L}} (\max\{1, \|\theta\|_\infty, \|\vartheta\|_\infty\})^{\mathbf{L}-1} \|\theta - \vartheta\|_\infty \end{aligned} \quad (15.6)$$

(cf. Definitions 3.3.4 and 4.4.1). This shows for all $x \in \mathbb{R}^d$ that

$$\mathbb{R}^d \ni \theta \mapsto \mathcal{N}_{u,v}^{\theta, \mathbf{l}}(x) \in \mathbb{R} \quad (15.7)$$

is continuous. Moreover, observe that the fact that for all $\theta \in \mathbb{R}^d$ it holds that $\mathcal{N}_{u,v}^{\theta, \mathbf{l}} \in C(\mathbb{R}^d, \mathbb{R})$ implies that for all $\theta \in \mathbb{R}^d$ it holds that $\mathcal{N}_{u,v}^{\theta, \mathbf{l}}(x)$ is $\mathcal{B}(\mathbb{R}^d)/\mathcal{B}(\mathbb{R})$ -measurable. This, (15.7), the fact that $(\mathbb{R}^d, \|\cdot\|_\infty|_{\mathbb{R}^d})$ is a separable normed \mathbb{R} -vector space, and Lemma 11.2.6 prove item (i). Note that item (i) and (15.5) demonstrate that

$$\Omega \times \mathbb{R}^d \ni (\omega, x) \mapsto \mathcal{N}_{u,v}^{\Theta_{\mathbf{k}(\omega)}(\omega), \mathbf{l}}(x) \in \mathbb{R} \quad (15.8)$$

is $(\mathcal{F} \otimes \mathcal{B}(\mathbb{R}^d)) / \mathcal{B}(\mathbb{R})$ -measurable. This establishes item (ii). Observe that item (ii) and the assumption that $\mathcal{E}: D \rightarrow \mathbb{R}$ is $\mathcal{B}(D) / \mathcal{B}(\mathbb{R})$ -measurable ensure that for all $p \in [0, \infty)$ it holds that

$$\Omega \times D \ni (\omega, x) \mapsto |\mathcal{N}_{u,v}^{\Theta_{\mathbf{k}(\omega)}(\omega), \mathbf{l}}(x) - \mathcal{E}(x)|^p \in [0, \infty) \quad (15.9)$$

is $(\mathcal{F} \otimes \mathcal{B}(D)) / \mathcal{B}([0, \infty))$ -measurable. Tonelli's theorem therefore proves item (iii). The proof of Lemma 15.1.1 is thus complete. \square

Proposition 15.1.2. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $M, d \in \mathbb{N}$, $b \in [1, \infty)$, $D \subseteq [-b, b]^d$, $u \in \mathbb{R}$, $v \in (u, \infty)$, for every $j \in \mathbb{N}$ let $X_j: \Omega \rightarrow D$ and $Y_j: \Omega \rightarrow [u, v]$ be random variables, assume that (X_j, Y_j) , $j \in \{1, 2, \dots, M\}$, are i.i.d., let $\mathbf{d}, \mathbf{L} \in \mathbb{N}$, $\mathbf{l} = (l_0, l_1, \dots, l_{\mathbf{L}}) \in \mathbb{N}^{\mathbf{L}+1}$ satisfy*

$$l_0 = d, \quad l_{\mathbf{L}} = 1, \quad \text{and} \quad \mathbf{d} \geq \sum_{i=1}^{\mathbf{L}} l_i(l_{i-1} + 1), \quad (15.10)$$

let $\mathcal{R}: \mathbb{R}^{\mathbf{d}} \times \Omega \rightarrow [0, \infty)$ satisfy for all $\theta \in \mathbb{R}^{\mathbf{d}}$ that

$$\mathcal{R}(\theta) = \frac{1}{M} \left[\sum_{j=1}^M |\mathcal{N}_{u,v}^{\theta, \mathbf{l}}(X_j) - Y_j|^2 \right], \quad (15.11)$$

let $\mathcal{E}: D \rightarrow [u, v]$ be $\mathcal{B}(D) / \mathcal{B}([u, v])$ -measurable, assume \mathbb{P} -a.s. that

$$\mathcal{E}(X_1) = \mathbb{E}[Y_1 | X_1], \quad (15.12)$$

let $K \in \mathbb{N}$, $c \in [1, \infty)$, $B \in [c, \infty)$, for every $k, n \in \mathbb{N}_0$ let $\Theta_{k,n}: \Omega \rightarrow \mathbb{R}^{\mathbf{d}}$ be random variables, assume $\bigcup_{k=1}^{\infty} \Theta_{k,0}(\Omega) \subseteq [-B, B]^{\mathbf{d}}$, assume that $\Theta_{k,0}$, $k \in \{1, 2, \dots, K\}$, are i.i.d., assume that $\Theta_{1,0}$ is continuously uniformly distributed on $[-c, c]^{\mathbf{d}}$, let $N \in \mathbb{N}$, $\mathbf{T} \subseteq \{0, 1, \dots, N\}$ satisfy $0 \in \mathbf{T}$, let $\mathbf{k}: \Omega \rightarrow (\mathbb{N}_0)^2$ be a random variable, and assume for all $\omega \in \Omega$ that

$$\mathbf{k}(\omega) \in \{(k, n) \in \{1, 2, \dots, K\} \times \mathbf{T}: \|\Theta_{k,n}(\omega)\|_{\infty} \leq B\} \quad (15.13)$$

$$\text{and} \quad \mathcal{R}(\Theta_{\mathbf{k}(\omega)}(\omega)) = \min_{(k,n) \in \{1,2,\dots,K\} \times \mathbf{T}, \|\Theta_{k,n}(\omega)\|_{\infty} \leq B} \mathcal{R}(\Theta_{k,n}(\omega)) \quad (15.14)$$

(cf. Definitions 3.3.4 and 4.4.1). Then it holds for all $p \in (0, \infty)$ that

$$\begin{aligned} & \left(\mathbb{E} \left[\left(\int_D |\mathcal{N}_{u,v}^{\Theta_{\mathbf{k}}, \mathbf{l}}(x) - \mathcal{E}(x)|^2 \mathbb{P}_{X_1}(dx) \right)^p \right] \right)^{1/p} \\ & \leq \left[\inf_{\theta \in [-c, c]^{\mathbf{d}}} \sup_{x \in D} |\mathcal{N}_{u,v}^{\theta, \mathbf{l}}(x) - \mathcal{E}(x)|^2 \right] + \frac{4(v-u)b\mathbf{L}(\|\mathbf{l}\|_{\infty} + 1)^{\mathbf{L}}c^{\mathbf{L}} \max\{1, p\}}{K^{[\mathbf{L}^{-1}(\|\mathbf{l}\|_{\infty} + 1)^{-2}]}} \\ & \quad + \frac{18 \max\{1, (v-u)^2\} \mathbf{L}(\|\mathbf{l}\|_{\infty} + 1)^2 \max\{p, \ln(3MBb)\}}{\sqrt{M}} \end{aligned} \quad (15.15)$$

(cf. Lemma 15.1.1).

Proof of Proposition 15.1.2. Throughout this proof, let $\mathbf{R}: \mathbb{R}^d \rightarrow [0, \infty)$ satisfy for all $\theta \in \mathbb{R}^d$ that

$$\mathbf{R}(\theta) = \mathbb{E}[|\mathcal{N}_{u,v}^{\theta,1}(X_1) - Y_1|^2]. \quad (15.16)$$

Note that Proposition 14.2.1 shows that for all $\vartheta \in [-B, B]^d$ it holds that

$$\begin{aligned} & \int_D |\mathcal{N}_{u,v}^{\Theta_k,1}(x) - \mathcal{E}(x)|^2 \mathbb{P}_{X_1}(dx) \\ & \leq [\sup_{x \in D} |\mathcal{N}_{u,v}^{\vartheta,1}(x) - \mathcal{E}(x)|^2] + 2[\sup_{\theta \in [-B, B]^d} |\mathcal{R}(\theta) - \mathbf{R}(\theta)|] \\ & \quad + \min_{(k,n) \in \{1,2,\dots,K\} \times \mathbf{T}, \|\Theta_{k,n}\|_\infty \leq B} |\mathcal{R}(\Theta_{k,n}) - \mathcal{R}(\vartheta)|. \end{aligned} \quad (15.17)$$

The assumption that $\bigcup_{k=1}^{\infty} \Theta_{k,0}(\Omega) \subseteq [-B, B]^d$ and the assumption that $0 \in \mathbf{T}$ hence imply that

$$\begin{aligned} & \int_D |\mathcal{N}_{u,v}^{\Theta_k,1}(x) - \mathcal{E}(x)|^2 \mathbb{P}_{X_1}(dx) \\ & \leq [\sup_{x \in D} |\mathcal{N}_{u,v}^{\vartheta,1}(x) - \mathcal{E}(x)|^2] + 2[\sup_{\theta \in [-B, B]^d} |\mathcal{R}(\theta) - \mathbf{R}(\theta)|] \\ & \quad + \min_{k \in \{1,2,\dots,K\}, \|\Theta_{k,0}\|_\infty \leq B} |\mathcal{R}(\Theta_{k,0}) - \mathcal{R}(\vartheta)| \\ & = [\sup_{x \in D} |\mathcal{N}_{u,v}^{\vartheta,1}(x) - \mathcal{E}(x)|^2] + 2[\sup_{\theta \in [-B, B]^d} |\mathcal{R}(\theta) - \mathbf{R}(\theta)|] \\ & \quad + \min_{k \in \{1,2,\dots,K\}} |\mathcal{R}(\Theta_{k,0}) - \mathcal{R}(\vartheta)|. \end{aligned} \quad (15.18)$$

Minkowski's inequality therefore demonstrates that for all $p \in [1, \infty)$, $\vartheta \in [-c, c]^d \subseteq [-B, B]^d$ it holds that

$$\begin{aligned} & \left(\mathbb{E} \left[\left(\int_D |\mathcal{N}_{u,v}^{\Theta_k,1}(x) - \mathcal{E}(x)|^2 \mathbb{P}_{X_1}(dx) \right)^p \right] \right)^{1/p} \\ & \leq (\mathbb{E} [\sup_{x \in D} |\mathcal{N}_{u,v}^{\vartheta,1}(x) - \mathcal{E}(x)|^{2p}])^{1/p} + 2(\mathbb{E} [\sup_{\theta \in [-B, B]^d} |\mathcal{R}(\theta) - \mathbf{R}(\theta)|^p])^{1/p} \\ & \quad + (\mathbb{E} [\min_{k \in \{1,2,\dots,K\}} |\mathcal{R}(\Theta_{k,0}) - \mathcal{R}(\vartheta)|^p])^{1/p} \\ & \leq [\sup_{x \in D} |\mathcal{N}_{u,v}^{\vartheta,1}(x) - \mathcal{E}(x)|^2] + 2(\mathbb{E} [\sup_{\theta \in [-B, B]^d} |\mathcal{R}(\theta) - \mathbf{R}(\theta)|^p])^{1/p} \\ & \quad + \sup_{\theta \in [-c, c]^d} (\mathbb{E} [\min_{k \in \{1,2,\dots,K\}} |\mathcal{R}(\Theta_{k,0}) - \mathcal{R}(\theta)|^p])^{1/p} \end{aligned} \quad (15.19)$$

(cf. item (i) in Corollary 13.3.3 and item (i) in Corollary 11.3.9). Furthermore, observe that Corollary 13.3.3 (applied with $v \curvearrowright \max\{u+1, v\}$, $\mathbf{R} \curvearrowright \mathbf{R}|_{[-B, B]^d}$, $\mathcal{R} \curvearrowright \mathcal{R}|_{[-B, B]^d \times \Omega}$ in the notation of Corollary 13.3.3) establishes that for all $p \in (0, \infty)$ it holds that

$$\begin{aligned} & (\mathbb{E} [\sup_{\theta \in [-B, B]^d} |\mathcal{R}(\theta) - \mathbf{R}(\theta)|^p])^{1/p} \\ & \leq \frac{9(\max\{u+1, v\} - u)^2 \mathbf{L}(\|\mathbf{l}\|_\infty + 1)^2 \max\{p, \ln(3MBb)\}}{\sqrt{M}} \\ & = \frac{9 \max\{1, (v-u)^2\} \mathbf{L}(\|\mathbf{l}\|_\infty + 1)^2 \max\{p, \ln(3MBb)\}}{\sqrt{M}}. \end{aligned} \quad (15.20)$$

Moreover, note that Corollary 11.3.9 (applied with $\mathfrak{d} \curvearrowright \sum_{i=1}^{\mathbf{L}} \mathbf{l}_i(\mathbf{l}_{i-1} + 1)$, $B \curvearrowright c$, $(\Theta_k)_{k \in \{1, 2, \dots, K\}} \curvearrowright (\Omega \ni \omega \mapsto \mathbb{1}_{\{\Theta_{k,0} \in [-c, c]^{\mathfrak{d}}\}}(\omega) \Theta_{k,0}(\omega) \in [-c, c]^{\mathfrak{d}})_{k \in \{1, 2, \dots, K\}}$, $\mathcal{R} \curvearrowright \mathcal{R}|_{[-c, c]^{\mathfrak{d}} \times \Omega}$ in the notation of Corollary 11.3.9) ensures that for all $p \in (0, \infty)$ it holds that

$$\begin{aligned} & \sup_{\theta \in [-c, c]^{\mathfrak{d}}} (\mathbb{E} [\min_{k \in \{1, 2, \dots, K\}} |\mathcal{R}(\Theta_{k,0}) - \mathcal{R}(\theta)|^p])^{1/p} \\ &= \sup_{\theta \in [-c, c]^{\mathfrak{d}}} (\mathbb{E} [\min_{k \in \{1, 2, \dots, K\}} |\mathcal{R}(\mathbb{1}_{\{\Theta_{k,0} \in [-c, c]^{\mathfrak{d}}\}} \Theta_{k,0}) - \mathcal{R}(\theta)|^p])^{1/p} \\ &\leq \frac{4(v-u)b\mathbf{L}(\|\mathbf{l}\|_{\infty} + 1)^{\mathbf{L}} c^{\mathbf{L}} \max\{1, p\}}{K^{[\mathbf{L}^{-1}(\|\mathbf{l}\|_{\infty} + 1)^{-2}]}}. \end{aligned} \quad (15.21)$$

Combining this and (15.20) with (15.19) proves that for all $p \in [1, \infty)$ it holds that

$$\begin{aligned} & \left(\mathbb{E} \left[\left(\int_D |\mathcal{N}_{u,v}^{\Theta_{\mathbf{k}}, \mathbf{l}}(x) - \mathcal{E}(x)|^2 \mathbb{P}_{X_1}(dx) \right)^p \right] \right)^{1/p} \\ &\leq \left[\inf_{\theta \in [-c, c]^{\mathfrak{d}}} \sup_{x \in D} |\mathcal{N}_{u,v}^{\theta, \mathbf{l}}(x) - \mathcal{E}(x)|^2 \right] + \frac{4(v-u)b\mathbf{L}(\|\mathbf{l}\|_{\infty} + 1)^{\mathbf{L}} c^{\mathbf{L}} \max\{1, p\}}{K^{[\mathbf{L}^{-1}(\|\mathbf{l}\|_{\infty} + 1)^{-2}]}} \\ &\quad + \frac{18 \max\{1, (v-u)^2\} \mathbf{L}(\|\mathbf{l}\|_{\infty} + 1)^2 \max\{p, \ln(3MBb)\}}{\sqrt{M}}. \end{aligned} \quad (15.22)$$

In addition, observe that that Jensen's inequality shows that for all $p \in (0, \infty)$ it holds that

$$\begin{aligned} & \left(\mathbb{E} \left[\left(\int_D |\mathcal{N}_{u,v}^{\Theta_{\mathbf{k}}, \mathbf{l}}(x) - \mathcal{E}(x)|^2 \mathbb{P}_{X_1}(dx) \right)^p \right] \right)^{1/p} \\ &\leq \left(\mathbb{E} \left[\left(\int_D |\mathcal{N}_{u,v}^{\Theta_{\mathbf{k}}, \mathbf{l}}(x) - \mathcal{E}(x)|^2 \mathbb{P}_{X_1}(dx) \right)^{\max\{1, p\}} \right] \right)^{\frac{1}{\max\{1, p\}}} \end{aligned} \quad (15.23)$$

This, (15.22), and the fact that $\ln(3MBb) \geq 1$ imply that for all $p \in (0, \infty)$ it holds that

$$\begin{aligned} & \left(\mathbb{E} \left[\left(\int_D |\mathcal{N}_{u,v}^{\Theta_{\mathbf{k}}, \mathbf{l}}(x) - \mathcal{E}(x)|^2 \mathbb{P}_{X_1}(dx) \right)^p \right] \right)^{1/p} \\ &\leq \left[\inf_{\theta \in [-c, c]^{\mathfrak{d}}} \sup_{x \in D} |\mathcal{N}_{u,v}^{\theta, \mathbf{l}}(x) - \mathcal{E}(x)|^2 \right] + \frac{4(v-u)b\mathbf{L}(\|\mathbf{l}\|_{\infty} + 1)^{\mathbf{L}} c^{\mathbf{L}} \max\{1, p\}}{K^{[\mathbf{L}^{-1}(\|\mathbf{l}\|_{\infty} + 1)^{-2}]}} \\ &\quad + \frac{18 \max\{1, (v-u)^2\} \mathbf{L}(\|\mathbf{l}\|_{\infty} + 1)^2 \max\{p, \ln(3MBb)\}}{\sqrt{M}}. \end{aligned} \quad (15.24)$$

The proof of Proposition 15.1.2 is thus complete. \square

Lemma 15.1.3. Let $a, x, p \in (0, \infty)$. Then $ax^p \leq \exp\left(\frac{a^{1/p}px}{e}\right)$.

Proof of Lemma 15.1.3. Note that the fact that for all $y \in \mathbb{R}$ it holds that $y + 1 \leq e^y$ demonstrates that

$$ax^p = (a^{1/p}x)^p = \left[e\left(\frac{a^{1/p}x}{e} - 1 + 1\right) \right]^p \leq \left[e \exp\left(\frac{a^{1/p}x}{e} - 1\right) \right]^p = \exp\left(\frac{a^{1/p}px}{e}\right). \quad (15.25)$$

The proof of Lemma 15.1.3 is thus complete. \square

Lemma 15.1.4. Let $M, c \in [1, \infty)$, $B \in [c, \infty)$. Then $\ln(3MBc) \leq \frac{23B}{18} \ln(eM)$.

Proof of Lemma 15.1.4. Observe that Lemma 15.1.3 and the fact that $2\sqrt{3}/e \leq 23/18$ establish that

$$3B^2 \leq \exp\left(\frac{2\sqrt{3}B}{e}\right) \leq \exp\left(\frac{23B}{18}\right). \quad (15.26)$$

The fact that $B \geq c \geq 1$ and $M \geq 1$ hence ensures that

$$\ln(3MBc) \leq \ln(3B^2M) \leq \ln([eM]^{23B/18}) = \frac{23B}{18} \ln(eM). \quad (15.27)$$

The proof of Lemma 15.1.4 is thus complete. \square

Theorem 15.1.5. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $M, d \in \mathbb{N}$, $a, u \in \mathbb{R}$, $b \in (a, \infty)$, $v \in (u, \infty)$, for every $j \in \mathbb{N}$ let $X_j: \Omega \rightarrow [a, b]^d$ and $Y_j: \Omega \rightarrow [u, v]$ be random variables, assume that (X_j, Y_j) , $j \in \{1, 2, \dots, M\}$, are i.i.d., let $A \in (0, \infty)$, $\mathbf{L} \in \mathbb{N}$ satisfy $\mathbf{L} \geq A\mathbb{1}_{(6^d, \infty)}(A)/(2d) + 1$, let $\mathbf{l} = (\mathbf{l}_0, \mathbf{l}_1, \dots, \mathbf{l}_{\mathbf{L}}) \in \mathbb{N}^{\mathbf{L}+1}$ satisfy for all $i \in \{2, 3, 4, \dots\} \cap [0, \mathbf{L}]$ that

$$\mathbf{l}_0 = d, \quad \mathbf{l}_1 \geq A\mathbb{1}_{(6^d, \infty)}(A), \quad \mathbf{l}_i \geq \mathbb{1}_{(6^d, \infty)}(A) \max\{A/d - 2i + 3, 2\}, \quad \text{and} \quad \mathbf{l}_{\mathbf{L}} = 1, \quad (15.28)$$

let $\mathbf{d} \in \mathbb{N}$ satisfy $\mathbf{d} \geq \sum_{i=1}^{\mathbf{L}} \mathbf{l}_i(\mathbf{l}_{i-1} + 1)$, let $\mathcal{R}: \mathbb{R}^{\mathbf{d}} \times \Omega \rightarrow [0, \infty)$ satisfy for all $\theta \in \mathbb{R}^{\mathbf{d}}$ that

$$\mathcal{R}(\theta) = \frac{1}{M} \left[\sum_{j=1}^M |\mathcal{N}_{u,v}^{\theta, \mathbf{l}}(X_j) - Y_j|^2 \right], \quad (15.29)$$

let $\mathcal{E}: [a, b]^d \rightarrow [u, v]$ satisfy \mathbb{P} -a.s. that

$$\mathcal{E}(X_1) = \mathbb{E}[Y_1|X_1], \quad (15.30)$$

let $L \in \mathbb{R}$ satisfy for all $x, y \in [a, b]^d$ that $|\mathcal{E}(x) - \mathcal{E}(y)| \leq L\|x - y\|_1$, let $K \in \mathbb{N}$, $c \in [\max\{1, L, |a|, |b|, 2|u|, 2|v|\}, \infty)$, $B \in [c, \infty)$, for every $k, n \in \mathbb{N}_0$ let $\Theta_{k,n}: \Omega \rightarrow \mathbb{R}^{\mathbf{d}}$ be a random variable, assume $\bigcup_{k=1}^{\infty} \Theta_{k,0}(\Omega) \subseteq [-B, B]^{\mathbf{d}}$, assume that $\Theta_{k,0}$, $k \in \{1, 2, \dots, K\}$, are i.i.d., assume that $\Theta_{1,0}$ is continuously uniformly distributed on $[-c, c]^{\mathbf{d}}$, let $N \in \mathbb{N}$, $\mathbf{T} \subseteq \{0, 1, \dots, N\}$ satisfy $0 \in \mathbf{T}$, let $\mathbf{k}: \Omega \rightarrow (\mathbb{N}_0)^2$ be a random variable, and assume for all $\omega \in \Omega$ that

$$\mathbf{k}(\omega) \in \{(k, n) \in \{1, 2, \dots, K\} \times \mathbf{T}: \|\Theta_{k,n}(\omega)\|_{\infty} \leq B\} \quad (15.31)$$

$$\text{and} \quad \mathcal{R}(\Theta_{\mathbf{k}(\omega)}(\omega)) = \min_{(k,n) \in \{1, 2, \dots, K\} \times \mathbf{T}, \|\Theta_{k,n}(\omega)\|_{\infty} \leq B} \mathcal{R}(\Theta_{k,n}(\omega)) \quad (15.32)$$

(cf. Definitions 3.3.4 and 4.4.1). Then it holds for all $p \in (0, \infty)$ that

$$\begin{aligned} & \left(\mathbb{E} \left[\left(\int_{[a,b]^d} |\mathcal{N}_{u,v}^{\Theta_k, \mathbf{l}}(x) - \mathcal{E}(x)|^2 \mathbb{P}_{X_1}(dx) \right)^p \right] \right)^{1/p} \\ & \leq \frac{36d^2c^4}{A^{2/d}} + \frac{4\mathbf{L}(\|\mathbf{l}\|_\infty + 1)^{\mathbf{L}} c^{\mathbf{L}+2} \max\{1, p\}}{K^{[\mathbf{L}^{-1}(\|\mathbf{l}\|_\infty + 1)^{-2}]}} \\ & \quad + \frac{23B^3\mathbf{L}(\|\mathbf{l}\|_\infty + 1)^2 \max\{p, \ln(eM)\}}{\sqrt{M}} \end{aligned} \quad (15.33)$$

(cf. Lemma 15.1.1).

Proof of Theorem 15.1.5. Note that the assumption that for all $x, y \in [a, b]^d$ it holds that $|\mathcal{E}(x) - \mathcal{E}(y)| \leq L\|x - y\|_1$ proves that $\mathcal{E}: [a, b]^d \rightarrow [u, v]$ is $\mathcal{B}([a, b]^d)/\mathcal{B}([u, v])$ -measurable. Proposition 15.1.2 (applied with $b \curvearrowright \max\{1, |a|, |b|\}$, $D \curvearrowright [a, b]^d$ in the notation of Proposition 15.1.2) therefore shows that for all $p \in (0, \infty)$ it holds that

$$\begin{aligned} & \left(\mathbb{E} \left[\left(\int_{[a,b]^d} |\mathcal{N}_{u,v}^{\Theta_k, \mathbf{l}}(x) - \mathcal{E}(x)|^2 \mathbb{P}_{X_1}(dx) \right)^p \right] \right)^{1/p} \\ & \leq [\inf_{\theta \in [-c, c]^d} \sup_{x \in [a, b]^d} |\mathcal{N}_{u,v}^{\theta, \mathbf{l}}(x) - \mathcal{E}(x)|^2] \\ & \quad + \frac{4(v-u) \max\{1, |a|, |b|\} \mathbf{L}(\|\mathbf{l}\|_\infty + 1)^{\mathbf{L}} c^{\mathbf{L}} \max\{1, p\}}{K^{[\mathbf{L}^{-1}(\|\mathbf{l}\|_\infty + 1)^{-2}]}} \\ & \quad + \frac{18 \max\{1, (v-u)^2\} \mathbf{L}(\|\mathbf{l}\|_\infty + 1)^2 \max\{p, \ln(3MB \max\{1, |a|, |b|\})\}}{\sqrt{M}}. \end{aligned} \quad (15.34)$$

The fact that $\max\{1, |a|, |b|\} \leq c$ hence implies that for all $p \in (0, \infty)$ it holds that

$$\begin{aligned} & \left(\mathbb{E} \left[\left(\int_{[a,b]^d} |\mathcal{N}_{u,v}^{\Theta_k, \mathbf{l}}(x) - \mathcal{E}(x)|^2 \mathbb{P}_{X_1}(dx) \right)^p \right] \right)^{1/p} \\ & \leq [\inf_{\theta \in [-c, c]^d} \sup_{x \in [a, b]^d} |\mathcal{N}_{u,v}^{\theta, \mathbf{l}}(x) - \mathcal{E}(x)|^2] \\ & \quad + \frac{4(v-u) \mathbf{L}(\|\mathbf{l}\|_\infty + 1)^{\mathbf{L}} c^{\mathbf{L}+1} \max\{1, p\}}{K^{[\mathbf{L}^{-1}(\|\mathbf{l}\|_\infty + 1)^{-2}]}} \\ & \quad + \frac{18 \max\{1, (v-u)^2\} \mathbf{L}(\|\mathbf{l}\|_\infty + 1)^2 \max\{p, \ln(3MBc)\}}{\sqrt{M}}. \end{aligned} \quad (15.35)$$

Furthermore, observe that Proposition 4.4.12 (applied with $f \curvearrowright \mathcal{E}$ in the notation of Proposition 4.4.12) demonstrates that there exists $\vartheta \in \mathbb{R}^d$ such that $\|\vartheta\|_\infty \leq \max\{1, L, |a|, |b|, 2[\sup_{x \in [a, b]^d} |\mathcal{E}(x)|]\}$ and

$$\sup_{x \in [a, b]^d} |\mathcal{N}_{u,v}^{\vartheta, \mathbf{l}}(x) - \mathcal{E}(x)| \leq \frac{3dL(b-a)}{A^{1/d}}. \quad (15.36)$$

The fact that for all $x \in [a, b]^d$ it holds that $\mathcal{E}(x) \in [u, v]$ therefore establishes that

$$\|\vartheta\|_\infty \leq \max\{1, L, |a|, |b|, 2|u|, 2|v|\} \leq c. \quad (15.37)$$

This and (15.36) ensure that

$$\begin{aligned} \inf_{\theta \in [-c, c]^d} \sup_{x \in [a, b]^d} |\mathcal{N}_{u,v}^{\theta, \mathbf{l}}(x) - \mathcal{E}(x)|^2 &\leq \sup_{x \in [a, b]^d} |\mathcal{N}_{u,v}^{\vartheta, \mathbf{l}}(x) - \mathcal{E}(x)|^2 \\ &\leq \left[\frac{3dL(b-a)}{A^{1/d}} \right]^2 = \frac{9d^2L^2(b-a)^2}{A^{2/d}}. \end{aligned} \quad (15.38)$$

Combining this with (15.35) proves that for all $p \in (0, \infty)$ it holds that

$$\begin{aligned} &\left(\mathbb{E} \left[\left(\int_{[a,b]^d} |\mathcal{N}_{u,v}^{\Theta_{\mathbf{k}}, \mathbf{l}}(x) - \mathcal{E}(x)|^2 \mathbb{P}_{X_1}(dx) \right)^p \right] \right)^{1/p} \\ &\leq \frac{9d^2L^2(b-a)^2}{A^{2/d}} + \frac{4(v-u)\mathbf{L}(\|\mathbf{l}\|_\infty + 1)^{\mathbf{L}} c^{\mathbf{L}+1} \max\{1, p\}}{K^{[\mathbf{L}^{-1}(\|\mathbf{l}\|_\infty + 1)^{-2}]}} \\ &\quad + \frac{18 \max\{1, (v-u)^2\} \mathbf{L}(\|\mathbf{l}\|_\infty + 1)^2 \max\{p, \ln(3MBc)\}}{\sqrt{M}}. \end{aligned} \quad (15.39)$$

Moreover, note that the fact that $\max\{1, L, |a|, |b|\} \leq c$ and $(b-a)^2 \leq (|a|+|b|)^2 \leq 2(a^2+b^2)$ shows that

$$9L^2(b-a)^2 \leq 18c^2(a^2+b^2) \leq 18c^2(c^2+c^2) = 36c^4. \quad (15.40)$$

In addition, observe that the fact that $B \geq c \geq 1$, the fact that $M \geq 1$, and Lemma 15.1.4 imply that $\ln(3MBc) \leq \frac{23B}{18} \ln(eM)$. This, (15.40), the fact that $(v-u) \leq 2 \max\{|u|, |v|\} = \max\{2|u|, 2|v|\} \leq c \leq B$, and the fact that $B \geq 1$ demonstrate that for all $p \in (0, \infty)$ it holds that

$$\begin{aligned} &\frac{9d^2L^2(b-a)^2}{A^{2/d}} + \frac{4(v-u)\mathbf{L}(\|\mathbf{l}\|_\infty + 1)^{\mathbf{L}} c^{\mathbf{L}+1} \max\{1, p\}}{K^{[\mathbf{L}^{-1}(\|\mathbf{l}\|_\infty + 1)^{-2}]}} \\ &\quad + \frac{18 \max\{1, (v-u)^2\} \mathbf{L}(\|\mathbf{l}\|_\infty + 1)^2 \max\{p, \ln(3MBc)\}}{\sqrt{M}} \\ &\leq \frac{36d^2c^4}{A^{2/d}} + \frac{4\mathbf{L}(\|\mathbf{l}\|_\infty + 1)^{\mathbf{L}} c^{\mathbf{L}+2} \max\{1, p\}}{K^{[\mathbf{L}^{-1}(\|\mathbf{l}\|_\infty + 1)^{-2}]}} \\ &\quad + \frac{23B^3\mathbf{L}(\|\mathbf{l}\|_\infty + 1)^2 \max\{p, \ln(eM)\}}{\sqrt{M}}. \end{aligned} \quad (15.41)$$

Combining this with (15.39) establishes (15.33). The proof of Theorem 15.1.5 is thus complete. \square

Corollary 15.1.6. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $M, d \in \mathbb{N}$, $a, u \in \mathbb{R}$, $b \in (a, \infty)$, $v \in (u, \infty)$, for every $j \in \mathbb{N}$ let $X_j: \Omega \rightarrow [a, b]^d$ and $Y_j: \Omega \rightarrow [u, v]$ be random variables, assume that (X_j, Y_j) , $j \in \{1, 2, \dots, M\}$, are i.i.d., let $\mathbf{d}, \mathbf{L} \in \mathbb{N}$, $\mathbf{l} = (\mathbf{l}_0, \mathbf{l}_1, \dots, \mathbf{l}_{\mathbf{L}}) \in \mathbb{N}^{\mathbf{L}+1}$, assume*

$$\mathbf{l}_0 = d, \quad \mathbf{l}_{\mathbf{L}} = 1, \quad \text{and} \quad \mathbf{d} \geq \sum_{i=1}^{\mathbf{L}} \mathbf{l}_i(\mathbf{l}_{i-1} + 1), \quad (15.42)$$

let $\mathcal{R}: \mathbb{R}^d \times \Omega \rightarrow [0, \infty)$ satisfy for all $\theta \in \mathbb{R}^d$ that

$$\mathcal{R}(\theta) = \frac{1}{M} \left[\sum_{j=1}^M |\mathcal{N}_{u,v}^{\theta,1}(X_j) - Y_j|^2 \right], \quad (15.43)$$

let $\mathcal{E}: [a, b]^d \rightarrow [u, v]$ satisfy \mathbb{P} -a.s. that

$$\mathcal{E}(X_1) = \mathbb{E}[Y_1|X_1], \quad (15.44)$$

let $L \in \mathbb{R}$ satisfy for all $x, y \in [a, b]^d$ that $|\mathcal{E}(x) - \mathcal{E}(y)| \leq L\|x - y\|_1$, let $K \in \mathbb{N}$, $c \in [\max\{1, L, |a|, |b|, 2|u|, 2|v|\}, \infty)$, $B \in [c, \infty)$, for every $k, n \in \mathbb{N}_0$ let $\Theta_{k,n}: \Omega \rightarrow \mathbb{R}^d$ be a random variable, assume $\bigcup_{k=1}^{\infty} \Theta_{k,0}(\Omega) \subseteq [-B, B]^d$, assume that $\Theta_{k,0}$, $k \in \{1, 2, \dots, K\}$, are i.i.d., assume that $\Theta_{1,0}$ is continuously uniformly distributed on $[-c, c]^d$, let $N \in \mathbb{N}$, $\mathbf{T} \subseteq \{0, 1, \dots, N\}$ satisfy $0 \in \mathbf{T}$, let $\mathbf{k}: \Omega \rightarrow (\mathbb{N}_0)^2$ be a random variable, and assume for all $\omega \in \Omega$ that

$$\mathbf{k}(\omega) \in \{(k, n) \in \{1, 2, \dots, K\} \times \mathbf{T}: \|\Theta_{k,n}(\omega)\|_{\infty} \leq B\} \quad (15.45)$$

$$\text{and } \mathcal{R}(\Theta_{\mathbf{k}(\omega)}(\omega)) = \min_{(k,n) \in \{1, 2, \dots, K\} \times \mathbf{T}, \|\Theta_{k,n}(\omega)\|_{\infty} \leq B} \mathcal{R}(\Theta_{k,n}(\omega)) \quad (15.46)$$

(cf. Definitions 3.3.4 and 4.4.1). Then it holds for all $p \in (0, \infty)$ that

$$\begin{aligned} & \left(\mathbb{E} \left[\left(\int_{[a,b]^d} |\mathcal{N}_{u,v}^{\Theta_{\mathbf{k}},1}(x) - \mathcal{E}(x)|^2 \mathbb{P}_{X_1}(dx) \right)^{p/2} \right] \right)^{1/p} \\ & \leq \frac{6dc^2}{[\min(\{\mathbf{L}\} \cup \{\mathbf{l}_i: i \in \mathbb{N} \cap [0, \mathbf{L}]\})]^{1/d}} + \frac{2\mathbf{L}(\|\mathbf{l}\|_{\infty} + 1)^{\mathbf{L}} c^{\mathbf{L}+1} \max\{1, p\}}{K^{[(2\mathbf{L})^{-1}(\|\mathbf{l}\|_{\infty} + 1)^{-2}]}} \\ & \quad + \frac{5B^2\mathbf{L}(\|\mathbf{l}\|_{\infty} + 1) \max\{p, \ln(eM)\}}{M^{1/4}} \end{aligned} \quad (15.47)$$

(cf. Lemma 15.1.1).

Proof of Corollary 15.1.6. Throughout this proof, let

$$A = \min(\{\mathbf{L}\} \cup \{\mathbf{l}_i: i \in \mathbb{N} \cap [0, \mathbf{L}]\}) \in (0, \infty). \quad (15.48)$$

Note that (15.48) ensures that

$$\begin{aligned} \mathbf{L} & \geq A = A - 1 + 1 \geq (A - 1)\mathbb{1}_{[2,\infty)}(A) + 1 \\ & \geq \left(A - \frac{A}{2}\right)\mathbb{1}_{[2,\infty)}(A) + 1 = \frac{A\mathbb{1}_{[2,\infty)}(A)}{2} + 1 \geq \frac{A\mathbb{1}_{(6^d,\infty)}(A)}{2d} + 1. \end{aligned} \quad (15.49)$$

Furthermore, observe that the assumption that $\mathbf{l}_{\mathbf{L}} = 1$ and (15.48) prove that

$$\mathbf{l}_1 = \mathbf{l}_1 \mathbb{1}_{\{1\}}(\mathbf{L}) + \mathbf{l}_1 \mathbb{1}_{[2,\infty)}(\mathbf{L}) \geq \mathbb{1}_{\{1\}}(\mathbf{L}) + A \mathbb{1}_{[2,\infty)}(\mathbf{L}) = A \geq A \mathbb{1}_{(6^d,\infty)}(A). \quad (15.50)$$

Moreover, note that (15.48) shows that for all $i \in \{2, 3, 4, \dots\} \cap [0, \mathbf{L}]$ it holds that

$$\begin{aligned} \mathbf{l}_i &\geq A \geq A \mathbb{1}_{[2,\infty)}(A) \geq \mathbb{1}_{[2,\infty)}(A) \max\{A - 1, 2\} = \mathbb{1}_{[2,\infty)}(A) \max\{A - 4 + 3, 2\} \\ &\geq \mathbb{1}_{[2,\infty)}(A) \max\{A - 2i + 3, 2\} \geq \mathbb{1}_{(6^d,\infty)}(A) \max\{A/d - 2i + 3, 2\}. \end{aligned} \quad (15.51)$$

Combining this, (15.49), and (15.50) with Theorem 15.1.5 (applied with $p \curvearrowleft p/2$ for $p \in (0, \infty)$ in the notation of Theorem 15.1.5) implies that for all $p \in (0, \infty)$ it holds that

$$\begin{aligned} &\left(\mathbb{E} \left[\left(\int_{[a,b]^d} |\mathcal{N}_{u,v}^{\Theta_{\mathbf{k}}, \mathbf{l}}(x) - \mathcal{E}(x)|^2 \mathbb{P}_{X_1}(\mathrm{d}x) \right)^{p/2} \right] \right)^{2/p} \\ &\leq \frac{36d^2c^4}{A^{2/d}} + \frac{4\mathbf{L}(\|\mathbf{l}\|_\infty + 1)^{\mathbf{L}} c^{\mathbf{L}+2} \max\{1, p/2\}}{K^{[\mathbf{L}^{-1}(\|\mathbf{l}\|_\infty + 1)^{-2}]}} \\ &\quad + \frac{23B^3\mathbf{L}(\|\mathbf{l}\|_\infty + 1)^2 \max\{p/2, \ln(eM)\}}{\sqrt{M}}. \end{aligned} \quad (15.52)$$

This, (15.48), and the fact that $\mathbf{L} \geq 1$, $c \geq 1$, $B \geq 1$, and $\ln(eM) \geq 1$ demonstrate that for all $p \in (0, \infty)$ it holds that

$$\begin{aligned} &\left(\mathbb{E} \left[\left(\int_{[a,b]^d} |\mathcal{N}_{u,v}^{\Theta_{\mathbf{k}}, \mathbf{l}}(x) - \mathcal{E}(x)|^2 \mathbb{P}_{X_1}(\mathrm{d}x) \right)^{p/2} \right] \right)^{1/p} \\ &\leq \frac{6dc^2}{[\min(\{\mathbf{L}\} \cup \{\mathbf{l}_i : i \in \mathbb{N} \cap [0, \mathbf{L}]\})]^{1/d}} + \frac{2[\mathbf{L}(\|\mathbf{l}\|_\infty + 1)^{\mathbf{L}} c^{\mathbf{L}+2} \max\{1, p/2\}]^{1/2}}{K^{[(2\mathbf{L})^{-1}(\|\mathbf{l}\|_\infty + 1)^{-2}]}} \\ &\quad + \frac{5B^3[\mathbf{L}(\|\mathbf{l}\|_\infty + 1)^2 \max\{p/2, \ln(eM)\}]^{1/2}}{M^{1/4}} \\ &\leq \frac{6dc^2}{[\min(\{\mathbf{L}\} \cup \{\mathbf{l}_i : i \in \mathbb{N} \cap [0, \mathbf{L}]\})]^{1/d}} + \frac{2\mathbf{L}(\|\mathbf{l}\|_\infty + 1)^{\mathbf{L}} c^{\mathbf{L}+1} \max\{1, p\}}{K^{[(2\mathbf{L})^{-1}(\|\mathbf{l}\|_\infty + 1)^{-2}]}} \\ &\quad + \frac{5B^2\mathbf{L}(\|\mathbf{l}\|_\infty + 1) \max\{p, \ln(eM)\}}{M^{1/4}}. \end{aligned} \quad (15.53)$$

The proof of Corollary 15.1.6 is thus complete. \square

15.2 Full strong error analysis with optimization via SGD with random initializations

Corollary 15.2.1. let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $M, d \in \mathbb{N}$, $a, u \in \mathbb{R}$, $b \in (a, \infty)$, $v \in (u, \infty)$, for every $k, n, j \in \mathbb{N}_0$ let $X_j^{k,n} : \Omega \rightarrow [a, b]^d$ and $Y_j^{k,n} : \Omega \rightarrow [u, v]$ be random variables, assume that $(X_j^{0,0}, Y_j^{0,0})$, $j \in \{1, 2, \dots, M\}$, are i.i.d., let $\mathbf{d}, \mathbf{L} \in \mathbb{N}$, $\mathbf{l} = (\mathbf{l}_0, \mathbf{l}_1, \dots, \mathbf{l}_{\mathbf{L}}) \in \mathbb{N}^{\mathbf{L}+1}$ satisfy

$$\mathbf{l}_0 = d, \quad \mathbf{l}_{\mathbf{L}} = 1, \quad \text{and} \quad \mathbf{d} \geq \sum_{i=1}^{\mathbf{L}} \mathbf{l}_i(\mathbf{l}_{i-1} + 1), \quad (15.54)$$

for every $k, n \in \mathbb{N}_0$, $J \in \mathbb{N}$ let $\mathcal{R}_J^{k,n}: \mathbb{R}^d \times \Omega \rightarrow [0, \infty)$ satisfy for all $\theta \in \mathbb{R}^d$, $\omega \in \Omega$ that

$$\mathcal{R}_J^{k,n}(\theta, \omega) = \frac{1}{J} \left[\sum_{j=1}^J |\mathcal{N}_{u,v}^{\theta, \mathbf{l}}(X_j^{k,n}(\omega)) - Y_j^{k,n}(\omega)|^2 \right], \quad (15.55)$$

let $\mathcal{E}: [a, b]^d \rightarrow [u, v]$ satisfy \mathbb{P} -a.s. that

$$\mathcal{E}(X_1^{0,0}) = \mathbb{E}[Y_1^{0,0}|X_1^{0,0}], \quad (15.56)$$

let $L \in \mathbb{R}$ satisfy for all $x, y \in [a, b]^d$ that $|\mathcal{E}(x) - \mathcal{E}(y)| \leq L\|x - y\|_1$, let $(\mathbf{J}_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, for every $k, n \in \mathbb{N}$ let $\mathcal{G}^{k,n}: \mathbb{R}^d \times \Omega \rightarrow \mathbb{R}^d$ satisfy for all $\omega \in \Omega$, $\theta \in \{\vartheta \in \mathbb{R}^d: (\mathcal{R}_{\mathbf{J}_n}^{k,n}(\cdot, \omega)): \mathbb{R}^d \rightarrow [0, \infty) \text{ is differentiable at } \vartheta\}\}$ that

$$\mathcal{G}^{k,n}(\theta, \omega) = (\nabla_\theta \mathcal{R}_{\mathbf{J}_n}^{k,n})(\theta, \omega), \quad (15.57)$$

let $K \in \mathbb{N}$, $c \in [\max\{1, L, |a|, |b|, 2|u|, 2|v|\}, \infty)$, $B \in [c, \infty)$, for every $k, n \in \mathbb{N}_0$ let $\Theta_{k,n}: \Omega \rightarrow \mathbb{R}^d$ be a random variable, assume $\bigcup_{k=1}^\infty \Theta_{k,0}(\Omega) \subseteq [-B, B]^d$, assume that $\Theta_{k,0}$, $k \in \{1, 2, \dots, K\}$, are i.i.d., assume that $\Theta_{1,0}$ is continuously uniformly distributed on $[-c, c]^d$, let $(\gamma_n)_{n \in \mathbb{N}} \subseteq \mathbb{R}$ satisfy for all $k, n \in \mathbb{N}$ that

$$\Theta_{k,n} = \Theta_{k,n-1} - \gamma_n \mathcal{G}^{k,n}(\Theta_{k,n-1}), \quad (15.58)$$

let $N \in \mathbb{N}$, $\mathbf{T} \subseteq \{0, 1, \dots, N\}$ satisfy $0 \in \mathbf{T}$, let $\mathbf{k}: \Omega \rightarrow (\mathbb{N}_0)^2$ be a random variable, and assume for all $\omega \in \Omega$ that

$$\mathbf{k}(\omega) \in \{(k, n) \in \{1, 2, \dots, K\} \times \mathbf{T}: \|\Theta_{k,n}(\omega)\|_\infty \leq B\} \quad (15.59)$$

$$\text{and} \quad \mathcal{R}(\Theta_{\mathbf{k}(\omega)}(\omega)) = \min_{(k,n) \in \{1, 2, \dots, K\} \times \mathbf{T}, \|\Theta_{k,n}(\omega)\|_\infty \leq B} \mathcal{R}(\Theta_{k,n}(\omega)) \quad (15.60)$$

(cf. Definitions 3.3.4 and 4.4.1). Then it holds for all $p \in (0, \infty)$ that

$$\begin{aligned} & \left(\mathbb{E} \left[\left(\int_{[a,b]^d} |\mathcal{N}_{u,v}^{\Theta_{\mathbf{k}}, \mathbf{l}}(x) - \mathcal{E}(x)|^2 \mathbb{P}_{X_1^{0,0}}(dx) \right)^{p/2} \right] \right)^{1/p} \\ & \leq \frac{6dc^2}{[\min(\{\mathbf{L}\} \cup \{\mathbf{l}_i: i \in \mathbb{N} \cap [0, \mathbf{L}]\})]^{1/d}} + \frac{2\mathbf{L}(\|\mathbf{l}\|_\infty + 1)^{\mathbf{L}} c^{\mathbf{L}+1} \max\{1, p\}}{K^{[(2\mathbf{L})^{-1}(\|\mathbf{l}\|_\infty + 1)^{-2}]}} \\ & \quad + \frac{5B^2 \mathbf{L}(\|\mathbf{l}\|_\infty + 1) \max\{p, \ln(eM)\}}{M^{1/4}} \end{aligned} \quad (15.61)$$

(cf. Lemma 15.1.1).

Proof of Corollary 15.2.1. Note that Corollary 15.1.6 (applied with $(X_j)_{j \in \mathbb{N}} \curvearrowright (X_j^{0,0})_{j \in \mathbb{N}}$, $(Y_j)_{j \in \mathbb{N}} \curvearrowright (Y_j^{0,0})_{j \in \mathbb{N}}$, $\mathcal{R} \curvearrowright \mathcal{R}_M^{0,0}$ in the notation of Corollary 15.1.6) establishes (15.61). The proof of Corollary 15.2.1 is thus complete. \square

Corollary 15.2.2. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $M, d \in \mathbb{N}$, $a, u \in \mathbb{R}$, $b \in (a, \infty)$, $v \in (u, \infty)$, for every $k, n, j \in \mathbb{N}_0$ let $X_j^{k,n}: \Omega \rightarrow [a, b]^d$ and $Y_j^{k,n}: \Omega \rightarrow [u, v]$ be random variables, assume that $(X_j^{0,0}, Y_j^{0,0})$, $j \in \{1, 2, \dots, M\}$, are i.i.d., let $\mathbf{d}, \mathbf{L} \in \mathbb{N}$, $\mathbf{l} = (\mathbf{l}_0, \mathbf{l}_1, \dots, \mathbf{l}_{\mathbf{L}}) \in \mathbb{N}^{\mathbf{L}+1}$ satisfy

$$\mathbf{l}_0 = d, \quad \mathbf{l}_{\mathbf{L}} = 1, \quad \text{and} \quad \mathbf{d} \geq \sum_{i=1}^{\mathbf{L}} \mathbf{l}_i (\mathbf{l}_{i-1} + 1), \quad (15.62)$$

for every $k, n \in \mathbb{N}_0$, $J \in \mathbb{N}$ let $\mathcal{R}_J^{k,n}: \mathbb{R}^{\mathbf{d}} \times \Omega \rightarrow [0, \infty)$ satisfy for all $\theta \in \mathbb{R}^{\mathbf{d}}$ that

$$\mathcal{R}_J^{k,n}(\theta) = \frac{1}{J} \left[\sum_{j=1}^J |\mathcal{N}_{u,v}^{\theta, \mathbf{l}}(X_j^{k,n}) - Y_j^{k,n}|^2 \right], \quad (15.63)$$

let $\mathcal{E}: [a, b]^d \rightarrow [u, v]$ satisfy \mathbb{P} -a.s. that

$$\mathcal{E}(X_1^{0,0}) = \mathbb{E}[Y_1^{0,0}|X_1^{0,0}], \quad (15.64)$$

let $L \in \mathbb{R}$ satisfy for all $x, y \in [a, b]^d$ that $|\mathcal{E}(x) - \mathcal{E}(y)| \leq L \|x - y\|_1$, let $(\mathbf{J}_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, for every $k, n \in \mathbb{N}$ let $\mathcal{G}^{k,n}: \mathbb{R}^{\mathbf{d}} \times \Omega \rightarrow \mathbb{R}^{\mathbf{d}}$ satisfy for all $\omega \in \Omega$, $\theta \in \{\vartheta \in \mathbb{R}^{\mathbf{d}}: (\mathcal{R}_{\mathbf{J}_n}^{k,n}(\cdot, \omega)): \mathbb{R}^{\mathbf{d}} \rightarrow [0, \infty) \text{ is differentiable at } \vartheta\}$ that

$$\mathcal{G}^{k,n}(\theta, \omega) = (\nabla_{\theta} \mathcal{R}_{\mathbf{J}_n}^{k,n})(\theta, \omega), \quad (15.65)$$

let $K \in \mathbb{N}$, $c \in [\max\{1, L, |a|, |b|, 2|u|, 2|v|\}, \infty)$, $B \in [c, \infty)$, for every $k, n \in \mathbb{N}_0$ let $\Theta_{k,n}: \Omega \rightarrow \mathbb{R}^{\mathbf{d}}$ be a random variable, assume $\bigcup_{k=1}^{\infty} \Theta_{k,0}(\Omega) \subseteq [-B, B]^{\mathbf{d}}$, assume that $\Theta_{k,0}$, $k \in \{1, 2, \dots, K\}$, are i.i.d., assume that $\Theta_{1,0}$ is continuously uniformly distributed on $[-c, c]^{\mathbf{d}}$, let $(\gamma_n)_{n \in \mathbb{N}} \subseteq \mathbb{R}$ satisfy for all $k, n \in \mathbb{N}$ that

$$\Theta_{k,n} = \Theta_{k,n-1} - \gamma_n \mathcal{G}^{k,n}(\Theta_{k,n-1}), \quad (15.66)$$

let $N \in \mathbb{N}$, $\mathbf{T} \subseteq \{0, 1, \dots, N\}$ satisfy $0 \in \mathbf{T}$, let $\mathbf{k}: \Omega \rightarrow (\mathbb{N}_0)^2$ be a random variable, and assume for all $\omega \in \Omega$ that

$$\mathbf{k}(\omega) \in \{(k, n) \in \{1, 2, \dots, K\} \times \mathbf{T}: \|\Theta_{k,n}(\omega)\|_{\infty} \leq B\} \quad (15.67)$$

$$\text{and} \quad \mathcal{R}(\Theta_{\mathbf{k}(\omega)}(\omega)) = \min_{(k,n) \in \{1, 2, \dots, K\} \times \mathbf{T}, \|\Theta_{k,n}(\omega)\|_{\infty} \leq B} \mathcal{R}(\Theta_{k,n}(\omega)) \quad (15.68)$$

(cf. Definitions 3.3.4 and 4.4.1). Then

$$\begin{aligned} & \mathbb{E} \left[\int_{[a,b]^d} |\mathcal{N}_{u,v}^{\Theta_k, \mathbf{l}}(x) - \mathcal{E}(x)| \mathbb{P}_{X_1^{0,0}}(dx) \right] \\ & \leq \frac{6dc^2}{[\min\{\mathbf{L}, \mathbf{l}_1, \mathbf{l}_2, \dots, \mathbf{l}_{\mathbf{L}-1}\}]^{1/d}} + \frac{5B^2 \mathbf{L}(\|\mathbf{l}\|_\infty + 1) \ln(eM)}{M^{1/4}} + \frac{2\mathbf{L}(\|\mathbf{l}\|_\infty + 1)^{\mathbf{L}} c^{\mathbf{L}+1}}{K^{[(2\mathbf{L})^{-1}(\|\mathbf{l}\|_\infty + 1)^{-2}]}} \end{aligned} \quad (15.69)$$

(cf. Lemma 15.1.1).

Proof of Corollary 15.2.2. Observe that Jensen's inequality ensures that

$$\mathbb{E} \left[\int_{[a,b]^d} |\mathcal{N}_{u,v}^{\Theta_k, \mathbf{l}}(x) - \mathcal{E}(x)| \mathbb{P}_{X_1^{0,0}}(dx) \right] \leq \mathbb{E} \left[\left(\int_{[a,b]^d} |\mathcal{N}_{u,v}^{\Theta_k, \mathbf{l}}(x) - \mathcal{E}(x)|^2 \mathbb{P}_{X_1^{0,0}}(dx) \right)^{1/2} \right]. \quad (15.70)$$

This and Corollary 15.2.1 (applied with $p \curvearrowright 1$ in the notation of Corollary 15.2.1) prove (15.69). The proof of Corollary 15.2.2 is thus complete. \square

Corollary 15.2.3. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $M, d \in \mathbb{N}$, for every $k, n, j \in \mathbb{N}_0$ let $X_j^{k,n}: \Omega \rightarrow [0, 1]^d$ and $Y_j^{k,n}: \Omega \rightarrow [0, 1]$ be random variables, assume that $(X_j^{0,0}, Y_j^{0,0})$, $j \in \{1, 2, \dots, M\}$, are i.i.d., for every $k, n \in \mathbb{N}_0$, $J \in \mathbb{N}$ let $\mathcal{R}_J^{k,n}: \mathbb{R}^d \times \Omega \rightarrow [0, \infty)$ satisfy for all $\theta \in \mathbb{R}^d$ that

$$\mathcal{R}_J^{k,n}(\theta, \omega) = \frac{1}{J} \left[\sum_{j=1}^J |\mathcal{N}_{0,1}^{\theta, \mathbf{l}}(X_j^{k,n}(\omega)) - Y_j^{k,n}(\omega)|^2 \right], \quad (15.71)$$

let $\mathbf{d}, \mathbf{L} \in \mathbb{N}$, $\mathbf{l} = (\mathbf{l}_0, \mathbf{l}_1, \dots, \mathbf{l}_{\mathbf{L}}) \in \mathbb{N}^{\mathbf{L}+1}$ satisfy

$$\mathbf{l}_0 = d, \quad \mathbf{l}_{\mathbf{L}} = 1, \quad \text{and} \quad \mathbf{d} \geq \sum_{i=1}^{\mathbf{L}} \mathbf{l}_i (\mathbf{l}_{i-1} + 1), \quad (15.72)$$

let $\mathcal{E}: [0, 1]^d \rightarrow [0, 1]$ satisfy \mathbb{P} -a.s. that

$$\mathcal{E}(X_1^{0,0}) = \mathbb{E}[Y_1^{0,0}|X_1^{0,0}], \quad (15.73)$$

let $c \in [2, \infty)$, satisfy for all $x, y \in [0, 1]^d$ that $|\mathcal{E}(x) - \mathcal{E}(y)| \leq c\|x - y\|_1$, let $(\mathbf{J}_n)_{n \in \mathbb{N}} \subseteq \mathbb{N}$, for every $k, n \in \mathbb{N}$ let $\mathcal{G}^{k,n}: \mathbb{R}^d \times \Omega \rightarrow \mathbb{R}^d$ satisfy for all $\omega \in \Omega$, $\theta \in \{\vartheta \in \mathbb{R}^d: (\mathcal{R}_{\mathbf{J}_n}^{k,n}(\cdot, \omega)): \mathbb{R}^d \rightarrow [0, \infty) \text{ is differentiable at } \vartheta\}$ that

$$\mathcal{G}^{k,n}(\theta, \omega) = (\nabla_\theta \mathcal{R}_{\mathbf{J}_n}^{k,n})(\theta, \omega), \quad (15.74)$$

let $K \in \mathbb{N}$, for every $k, n \in \mathbb{N}_0$ let $\Theta_{k,n}: \Omega \rightarrow \mathbb{R}^d$ be a random variable, assume $\bigcup_{k=1}^\infty \Theta_{k,0}(\Omega) \subseteq [-c, c]^d$, assume that $\Theta_{k,0}$, $k \in \{1, 2, \dots, K\}$, are i.i.d., assume that

$\Theta_{1,0}$ is continuously uniformly distributed on $[-c, c]^d$, let $(\gamma_n)_{n \in \mathbb{N}} \subseteq \mathbb{R}$ satisfy for all $k, n \in \mathbb{N}$ that

$$\Theta_{k,n} = \Theta_{k,n-1} - \gamma_n \mathcal{G}^{k,n}(\Theta_{k,n-1}), \quad (15.75)$$

let $N \in \mathbb{N}$, $\mathbf{T} \subseteq \{0, 1, \dots, N\}$ satisfy $0 \in \mathbf{T}$, let $\mathbf{k}: \Omega \rightarrow (\mathbb{N}_0)^2$ be a random variable, and assume for all $\omega \in \Omega$ that

$$\mathbf{k}(\omega) \in \{(k, n) \in \{1, 2, \dots, K\} \times \mathbf{T} : \|\Theta_{k,n}(\omega)\|_\infty \leq B\} \quad (15.76)$$

$$\text{and } \mathcal{R}(\Theta_{\mathbf{k}(\omega)}(\omega)) = \min_{(k,n) \in \{1,2,\dots,K\} \times \mathbf{T}, \|\Theta_{k,n}(\omega)\|_\infty \leq B} \mathcal{R}(\Theta_{k,n}(\omega)) \quad (15.77)$$

(cf. Definitions 3.3.4 and 4.4.1). Then

$$\begin{aligned} & \mathbb{E} \left[\int_{[0,1]^d} |\mathcal{N}_{0,1}^{\Theta_{\mathbf{k}}, \mathbf{l}}(x) - \mathcal{E}(x)| \mathbb{P}_{X_1^{0,0}}(dx) \right] \\ & \leq \frac{6dc^2}{[\min\{\mathbf{L}, \mathbf{l}_1, \mathbf{l}_2, \dots, \mathbf{l}_{\mathbf{L}-1}\}]^{1/d}} + \frac{5c^2 \mathbf{L}(\|\mathbf{l}\|_\infty + 1) \ln(eM)}{M^{1/4}} + \frac{\mathbf{L}(\|\mathbf{l}\|_\infty + 1)^{\mathbf{L}} c^{\mathbf{L}+1}}{K^{[(2\mathbf{L})^{-1}(\|\mathbf{l}\|_\infty + 1)^{-2}]}} \end{aligned} \quad (15.78)$$

(cf. Lemma 15.1.1).

Proof of Corollary 15.2.3. Note that Corollary 15.2.2 (applied with $a \curvearrowright 0$, $u \curvearrowright 0$, $b \curvearrowright 1$, $v \curvearrowright 1$, $L \curvearrowright c$, $c \curvearrowright c$, $B \curvearrowright c$ in the notation of Corollary 15.2.2), the fact that $c \geq 2$ and $M \geq 1$, and Lemma 15.1.4 show (15.78). The proof of Corollary 15.2.3 is thus complete. \square

Part VI

Deep learning for partial differential equations (PDEs)

Chapter 16

Physics-informed neural networks (PINNs)

Deep learning methods have not only become very popular for data-driven learning problems, but are nowadays also heavily used for solving mathematical equations such as ordinary and partial differential equations (cf., for instance, [125, 196, 368, 400]). In particular, we refer to the overview articles [24, 58, 91, 152, 251, 376] and the references therein for numerical simulations and theoretical investigations for deep learning methods for [PDEs](#).

Often deep learning methods for [PDEs](#) are obtained, first, by reformulating the [PDE](#) problem under consideration as an infinite-dimensional stochastic optimization problem, then, by approximating the infinite-dimensional stochastic optimization problem through finite-dimensional stochastic optimization problems involving deep [ANNs](#) as approximations for the [PDE](#) solution and/or its derivatives, and thereafter, by approximately solving the resulting finite-dimensional stochastic optimization problems through [SGD](#)-type optimization methods.

Among the most basic schemes of such deep learning methods for [PDEs](#) are [PINNs](#) and [DGMs](#); see [368, 400]. In this chapter we present in Theorem 16.1.1 in Section 16.1 a reformulation of [PDE](#) problems as stochastic optimization problems, we use the theoretical considerations from Section 16.1 to briefly sketch in Section 16.2 a possible derivation of [PINNs](#) and [DGMs](#), and we present in Sections 16.3 and 16.4 numerical simulations for [PINNs](#) and [DGMs](#). For simplicity and concreteness we restrict ourselves in this chapter to the case of semilinear heat [PDEs](#). The specific presentation of this chapter is based on Beck et al. [24].

16.1 Reformulation of PDE problems as stochastic optimization problems

Both PINNs and DGMs are based on reformulations of the considered PDEs as suitable infinite-dimensional stochastic optimization problems. In Theorem 16.1.1 below we present the theoretical result behind this reformulation in the special case of semilinear heat PDEs.

Theorem 16.1.1. *Let $T \in (0, \infty)$, $d \in \mathbb{N}$, $g \in C^2(\mathbb{R}^d, \mathbb{R})$, $u \in C^{1,2}([0, T] \times \mathbb{R}^d, \mathbb{R})$, $t \in C([0, T], (0, \infty))$, $x \in C(\mathbb{R}^d, (0, \infty))$, assume that g has at most polynomially growing partial derivatives, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\mathcal{T}: \Omega \rightarrow [0, T]$ and $\mathcal{X}: \Omega \rightarrow \mathbb{R}^d$ be independent random variables, assume for all $A \in \mathcal{B}([0, T])$, $B \in \mathcal{B}(\mathbb{R}^d)$ that*

$$\mathbb{P}(\mathcal{T} \in A) = \int_A t(t) dt \quad \text{and} \quad \mathbb{P}(\mathcal{X} \in B) = \int_B x(x) dx, \quad (16.1)$$

let $f: \mathbb{R} \rightarrow \mathbb{R}$ be Lipschitz continuous, and let $\mathfrak{L}: C^{1,2}([0, T] \times \mathbb{R}^d, \mathbb{R}) \rightarrow [0, \infty]$ satisfy for all $v = (v(t, x))_{(t,x) \in [0,T] \times \mathbb{R}^d} \in C^{1,2}([0, T] \times \mathbb{R}^d, \mathbb{R})$ that

$$\mathfrak{L}(v) = \mathbb{E}[|v(0, \mathcal{X}) - g(\mathcal{X})|^2 + |(\frac{\partial v}{\partial t})(\mathcal{T}, \mathcal{X}) - (\Delta_x v)(\mathcal{T}, \mathcal{X}) - f(v(\mathcal{T}, \mathcal{X}))|^2]. \quad (16.2)$$

Then the following two statements are equivalent:

- (i) *It holds that $\mathfrak{L}(u) = \inf_{v \in C^{1,2}([0,T] \times \mathbb{R}^d, \mathbb{R})} \mathfrak{L}(v)$.*
- (ii) *It holds for all $t \in [0, T]$, $x \in \mathbb{R}^d$ that $u(0, x) = g(x)$ and*

$$(\frac{\partial u}{\partial t})(t, x) = (\Delta_x u)(t, x) + f(u(t, x)). \quad (16.3)$$

Proof of Theorem 16.1.1. Observe that (16.2) implies that for all $v \in C^{1,2}([0, T] \times \mathbb{R}^d, \mathbb{R})$ with $\forall x \in \mathbb{R}^d: u(0, x) = g(x)$ and $\forall t \in [0, T], x \in \mathbb{R}^d: (\frac{\partial v}{\partial t})(t, x) = (\Delta_x v)(t, x) + f(v(t, x))$ it holds that

$$\mathfrak{L}(v) = 0. \quad (16.4)$$

This and the fact that for all $v \in C^{1,2}([0, T] \times \mathbb{R}^d, \mathbb{R})$ it holds that $\mathfrak{L}(v) \geq 0$ establish that (ii) \rightarrow (i). Note that the assumption that f is Lipschitz continuous, the assumption that g is twice continuously differentiable, and the assumption that g has at most polynomially growing partial derivatives demonstrate that there exists $v \in C^{1,2}([0, T] \times \mathbb{R}^d, \mathbb{R})$ which satisfies for all $t \in [0, T]$, $x \in \mathbb{R}^d$ that $v(0, x) = g(x)$ and

$$(\frac{\partial v}{\partial t})(t, x) = (\Delta_x v)(t, x) + f(v(t, x)) \quad (16.5)$$

(cf., for example, Beck et al. [23, Corollary 3.4]). This and (16.4) show that

$$\inf_{v \in C^{1,2}([0,T] \times \mathbb{R}^d, \mathbb{R})} \mathfrak{L}(v) = 0. \quad (16.6)$$

Furthermore, observe that (16.2), (16.1), and the assumption that \mathcal{T} and \mathcal{X} are independent ensure that for all $v \in C^{1,2}([0, T] \times \mathbb{R}^d, \mathbb{R})$ it holds that

$$\mathfrak{L}(v) = \int_{[0,T] \times \mathbb{R}^d} \left(|v(0, x) - g(x)|^2 + \left| \left(\frac{\partial v}{\partial t} \right)(t, x) - (\Delta_x v)(t, x) - f(v(t, x)) \right|^2 \right) \mathfrak{t}(t) \mathfrak{x}(x) dt dx. \quad (16.7)$$

The assumption that \mathfrak{t} and \mathfrak{x} are continuous and the fact that for all $t \in [0, T]$, $x \in \mathbb{R}^d$ it holds that $\mathfrak{t}(t) \geq 0$ and $\mathfrak{x}(x) \geq 0$ hence prove that for all $v \in C^{1,2}([0, T] \times \mathbb{R}^d, \mathbb{R})$, $t \in [0, T]$, $x \in \mathbb{R}^d$ with $\mathfrak{L}(v) = 0$ it holds that

$$\left(|v(0, x) - g(x)|^2 + \left| \left(\frac{\partial v}{\partial t} \right)(t, x) - (\Delta_x v)(t, x) - f(v(t, x)) \right|^2 \right) \mathfrak{t}(t) \mathfrak{x}(x) = 0. \quad (16.8)$$

This and the assumption that for all $t \in [0, T]$, $x \in \mathbb{R}^d$ it holds that $\mathfrak{t}(t) > 0$ and $\mathfrak{x}(x) > 0$ show that for all $v \in C^{1,2}([0, T] \times \mathbb{R}^d, \mathbb{R})$, $t \in [0, T]$, $x \in \mathbb{R}^d$ with $\mathfrak{L}(v) = 0$ it holds that

$$|v(0, x) - g(x)|^2 + \left| \left(\frac{\partial v}{\partial t} \right)(t, x) - (\Delta_x v)(t, x) - f(v(t, x)) \right|^2 = 0. \quad (16.9)$$

Combining this with (16.6) establishes that ((i) \rightarrow (ii)). The proof of Theorem 16.1.1 is thus complete. \square

16.2 Derivation of PINNs and deep Galerkin methods (DGMs)

In this section we employ the reformulation of semilinear PDEs as optimization problems from Theorem 16.1.1 to sketch an informal derivation of deep learning schemes to approximate solutions of semilinear heat PDEs. For this let $T \in (0, \infty)$, $d \in \mathbb{N}$, $u \in C^{1,2}([0, T] \times \mathbb{R}^d, \mathbb{R})$, $g \in C^2(\mathbb{R}^d, \mathbb{R})$ satisfy that g has at most polynomially growing partial derivatives, let $f: \mathbb{R} \rightarrow \mathbb{R}$ be Lipschitz continuous, and assume for all $t \in [0, T]$, $x \in \mathbb{R}^d$ that $u(0, x) = g(x)$ and

$$\left(\frac{\partial u}{\partial t} \right)(t, x) = (\Delta_x u)(t, x) + f(u(t, x)). \quad (16.10)$$

In the framework described in the previous sentence, we think of u as the unknown PDE solution. The objective of this derivation is to develop deep learning methods which aim to approximate the unknown function u .

In the first step we employ Theorem 16.1.1 to reformulate the PDE problem associated to (16.10) as an infinite-dimensional stochastic optimization problem over a function space. For this let $\mathfrak{t} \in C([0, T], (0, \infty))$, $\mathfrak{x} \in C(\mathbb{R}^d, (0, \infty))$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\mathcal{T}: \Omega \rightarrow [0, T]$ and $\mathcal{X}: \Omega \rightarrow \mathbb{R}^d$ be independent random variables, assume for all $A \in \mathcal{B}([0, T])$, $B \in \mathcal{B}(\mathbb{R}^d)$ that

$$\mathbb{P}(\mathcal{T} \in A) = \int_A \mathfrak{t}(t) dt \quad \text{and} \quad \mathbb{P}(\mathcal{X} \in B) = \int_B \mathfrak{x}(x) dx, \quad (16.11)$$

and let $\mathfrak{L}: C^{1,2}([0, T] \times \mathbb{R}^d, \mathbb{R}) \rightarrow [0, \infty]$ satisfy for all $v = (v(t, x))_{(t,x) \in [0,T] \times \mathbb{R}^d} \in C^{1,2}([0, T] \times \mathbb{R}^d, \mathbb{R})$ that

$$\mathfrak{L}(v) = \mathbb{E}[|v(0, \mathcal{X}) - g(\mathcal{X})|^2 + |(\frac{\partial v}{\partial t})(\mathcal{T}, \mathcal{X}) - (\Delta_x v)(\mathcal{T}, \mathcal{X}) - f(v(\mathcal{T}, \mathcal{X}))|^2]. \quad (16.12)$$

Observe that Theorem 16.1.1 assures that the unknown function u satisfies

$$\mathfrak{L}(u) = 0 \quad (16.13)$$

and is thus a minimizer of the optimization problem associated to (16.12). Motivated by this, we consider aim to find approximations of u by computing approximate minimizers of the function $\mathfrak{L}: C^{1,2}([0, T] \times \mathbb{R}^d, \mathbb{R}) \rightarrow [0, \infty]$. Due to its infinite-dimensionality this optimization problem is however not yet amenable to numerical computations.

For this reason, in the second step, we reduce this infinite-dimensional stochastic optimization problem to a finite-dimensional stochastic optimization problem involving **ANNs**. Specifically, let $a: \mathbb{R} \rightarrow \mathbb{R}$ be differentiable, let $h \in \mathbb{N}$, $l_1, l_2, \dots, l_h, \mathfrak{d} \in \mathbb{N}$ satisfy $\mathfrak{d} = l_1(d+2) + [\sum_{k=2}^h l_k(l_{k-1}+1)] + l_h + 1$, and let $\mathscr{L}: \mathbb{R}^{\mathfrak{d}} \rightarrow [0, \infty)$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that

$$\begin{aligned} \mathscr{L}(\theta) &= \mathfrak{L}(\mathcal{N}_{\mathfrak{M}_{a,l_1}, \mathfrak{M}_{a,l_2}, \dots, \mathfrak{M}_{a,l_h}, \text{id}_{\mathbb{R}}}^{\theta, d+1}) \\ &= \mathbb{E}\left[|\mathcal{N}_{\mathfrak{M}_{a,l_1}, \mathfrak{M}_{a,l_2}, \dots, \mathfrak{M}_{a,l_h}, \text{id}_{\mathbb{R}}}^{\theta, d+1}(0, \mathcal{X}) - g(\mathcal{X})|^2 \right. \\ &\quad + \left| \left(\frac{\partial \mathcal{N}_{\mathfrak{M}_{a,l_1}, \mathfrak{M}_{a,l_2}, \dots, \mathfrak{M}_{a,l_h}, \text{id}_{\mathbb{R}}}^{\theta, d+1}}{\partial t} \right)(\mathcal{T}, \mathcal{X}) - (\Delta_x \mathcal{N}_{\mathfrak{M}_{a,l_1}, \mathfrak{M}_{a,l_2}, \dots, \mathfrak{M}_{a,l_h}, \text{id}_{\mathbb{R}}}^{\theta, d+1})(\mathcal{T}, \mathcal{X}) \right. \\ &\quad \left. - f(\mathcal{N}_{\mathfrak{M}_{a,l_1}, \mathfrak{M}_{a,l_2}, \dots, \mathfrak{M}_{a,l_h}, \text{id}_{\mathbb{R}}}^{\theta, d+1}(\mathcal{T}, \mathcal{X})) \right|^2 \right] \end{aligned} \quad (16.14)$$

(cf. Definitions 1.1.3 and 1.2.1). We can now compute an approximate minimizer of the function \mathfrak{L} by computing an approximate minimizer $\vartheta \in \mathbb{R}^{\mathfrak{d}}$ of the function \mathscr{L} and employing the realization $\mathcal{N}_{\mathfrak{M}_{a,l_1}, \mathfrak{M}_{a,l_2}, \dots, \mathfrak{M}_{a,l_h}, \text{id}_{\mathbb{R}}}^{\vartheta, d+1}$ of the **ANN** associated to this approximate minimizer as an approximate minimizer of \mathfrak{L} .

The third and last step of this derivation is to approximately compute such an approximate minimizer of \mathscr{L} by means of **SGD**-type optimization methods. We now sketch this in the case of the plain-vanilla **SGD** optimization method (cf. Definition 7.2.1). Let $\xi \in \mathbb{R}^{\mathfrak{d}}$, $J \in \mathbb{N}$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, for every $n \in \mathbb{N}$, $j \in \{1, 2, \dots, J\}$ let $\mathfrak{T}_{n,j}: \Omega \rightarrow [0, T]$ and $\mathfrak{X}_{n,j}: \Omega \rightarrow \mathbb{R}^d$ be random variables, assume for all $n \in \mathbb{N}$, $j \in \{1, 2, \dots, J\}$, $A \in \mathcal{B}([0, T])$, $B \in \mathcal{B}(\mathbb{R}^d)$ that

$$\mathbb{P}(\mathcal{T} \in A) = \mathbb{P}(\mathfrak{T}_{n,j} \in A) \quad \text{and} \quad \mathbb{P}(\mathcal{X} \in B) = \mathbb{P}(\mathfrak{X}_{n,j} \in B), \quad (16.15)$$

let $\ell : \mathbb{R}^d \times [0, T] \times \mathbb{R}^d \rightarrow \mathbb{R}$ satisfy for all $\theta \in \mathbb{R}^d$, $t \in [0, T]$, $x \in \mathbb{R}^d$ that

$$\begin{aligned}\ell(\theta, t, x) &= |\mathcal{N}_{\mathfrak{M}_{a,l_1}, \mathfrak{M}_{a,l_2}, \dots, \mathfrak{M}_{a,l_h}, \text{id}_{\mathbb{R}}}^{\theta, d+1}(0, x) - g(x)|^2 \\ &\quad + \left| \left(\frac{\partial \mathcal{N}_{\mathfrak{M}_{a,l_1}, \mathfrak{M}_{a,l_2}, \dots, \mathfrak{M}_{a,l_h}, \text{id}_{\mathbb{R}}}^{\theta, d+1}}{\partial t} \right)(t, x) - (\Delta_x \mathcal{N}_{\mathfrak{M}_{a,l_1}, \mathfrak{M}_{a,l_2}, \dots, \mathfrak{M}_{a,l_h}, \text{id}_{\mathbb{R}}}^{\theta, d+1})(t, x) \right|^2 \\ &\quad - f(\mathcal{N}_{\mathfrak{M}_{a,l_1}, \mathfrak{M}_{a,l_2}, \dots, \mathfrak{M}_{a,l_h}, \text{id}_{\mathbb{R}}}^{\theta, d+1}(t, x)) \right|^2,\end{aligned}\quad (16.16)$$

and let $\Theta = (\Theta_n)_{n \in \mathbb{N}_0} : \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^d$ satisfy for all $n \in \mathbb{N}$ that

$$\Theta_0 = \xi \quad \text{and} \quad \Theta_n = \Theta_{n-1} - \gamma_n \left[\frac{1}{J} \sum_{j=1}^J (\nabla_\theta \ell)(\Theta_{n-1}, \mathfrak{T}_{n,j}, \mathfrak{X}_{n,j}) \right]. \quad (16.17)$$

Finally, the idea of **PINNs** and **DGMs** is then to choose for large enough $n \in \mathbb{N}$ the realization $\mathcal{N}_{\mathfrak{M}_{a,l_1}, \mathfrak{M}_{a,l_2}, \dots, \mathfrak{M}_{a,l_h}, \text{id}_{\mathbb{R}}}^{\Theta_n, d+1}$ as an approximation

$$\mathcal{N}_{\mathfrak{M}_{a,l_1}, \mathfrak{M}_{a,l_2}, \dots, \mathfrak{M}_{a,l_h}, \text{id}_{\mathbb{R}}}^{\Theta_n, d+1} \approx u \quad (16.18)$$

of the unknown solution u of the PDE in (16.10).

The ideas and the resulting schemes in the above derivation were first introduced as **PINNs** in Raissi et al. [368] and as **DGMs** in Sirignano & Spiliopoulos [400]. Very roughly speaking, **PINNs** and **DGMs** in their original form differ in the way the joint distribution of the random variables $(\mathfrak{T}_{n,j}, \mathfrak{X}_{n,j})_{(n,j) \in \mathbb{N} \times \{1, 2, \dots, J\}}$ would be chosen. Loosely speaking, in the case of **PINNs** the originally proposed distribution for $(\mathfrak{T}_{n,j}, \mathfrak{X}_{n,j})_{(n,j) \in \mathbb{N} \times \{1, 2, \dots, J\}}$ would be based on drawing a finite number of samples of the random variable $(\mathcal{T}, \mathcal{X})$ and then having the random variable $(\mathfrak{T}_{n,j}, \mathfrak{X}_{n,j})_{(n,j) \in \mathbb{N} \times \{1, 2, \dots, J\}}$ be randomly chosen among those samples. In the case of **DGMs** the original proposition would be to choose $(\mathfrak{T}_{n,j}, \mathfrak{X}_{n,j})_{(n,j) \in \mathbb{N} \times \{1, 2, \dots, J\}}$ independent and identically distributed. Implementations of **PINNs** and **DGMs** that employ more sophisticated optimization methods, such as the **Adam SGD** optimization method, can be found in the next section.

16.3 Implementation of PINNs

In Source code 16.1 below we present a simple implementation of the **PINN** method, as explained in Section 16.2 above, for finding an approximation of a solution $u \in C^{1,2}([0, 3] \times \mathbb{R}^2)$ of the two-dimensional Allen–Cahn-type semilinear heat equation

$$(\frac{\partial u}{\partial t})(t, x) = \frac{1}{200}(\Delta_x u)(t, x) + u(t, x) - [u(t, x)]^3 \quad (16.19)$$

with $u(0, x) = \sin(\|x\|_2^2)$ for $t \in [0, 3]$, $x \in \mathbb{R}^2$. This implementation follows the original proposal in Raissi et al. [368] in that it first chooses 20000 realizations of the random variable

$(\mathcal{T}, \mathcal{X})$, where \mathcal{T} is continuous uniformly distributed on $[0, 3]$ and where \mathcal{X} is normally distributed on \mathbb{R}^2 with mean $0 \in \mathbb{R}^2$ and covariance $4I_2 \in \mathbb{R}^{2 \times 2}$ (cf. Definition 1.5.5). It then trains a fully connected feed-forward ANN with 4 hidden layers (with 50 neurons on each hidden layer) and using the swish activation function with parameter 1 (cf. Section 1.2.8). The training uses batches of size 256 with each batch chosen from the 20000 realizations of the random variable $(\mathcal{T}, \mathcal{X})$ which were picked beforehand. The training is performed using the Adam SGD optimization method (cf. Section 7.9). A plot of the resulting approximation of the solution u after 20000 training steps is shown in Figure 16.1.

```

1 import torch
2 import matplotlib.pyplot as plt
3 from torch.autograd import grad
4 from matplotlib.gridspec import GridSpec
5 from matplotlib.cm import ScalarMappable
6
7
8 dev = torch.device("cuda:0" if torch.cuda.is_available() else
9                  "cpu")
10
11 T = 3.0 # the time horizon
12 M = 20000 # the number of training samples
13
14 torch.manual_seed(0)
15
16 x_data = torch.randn(M, 2).to(dev) * 2
17 t_data = torch.rand(M, 1).to(dev) * T
18
19 # The initial value
20 def phi(x):
21     return x.square().sum(axis=1, keepdims=True).sin()
22
23 # We use a network with 4 hidden layers of 50 neurons each and the
24 # Swish activation function (called SiLU in PyTorch)
25 N = torch.nn.Sequential(
26     torch.nn.Linear(3, 50), torch.nn.SiLU(),
27     torch.nn.Linear(50, 50), torch.nn.SiLU(),
28     torch.nn.Linear(50, 50), torch.nn.SiLU(),
29     torch.nn.Linear(50, 50), torch.nn.SiLU(),
30     torch.nn.Linear(50, 1),
31 ).to(dev)
32
33 optimizer = torch.optim.Adam(N.parameters(), lr=3e-4)
34
35 J = 256 # the batch size
36
37 for i in range(20000):
38     # Choose a random batch of training samples
39     indices = torch.randint(0, M, (J,))

```

```
40     x = x_data[indices, :]
41     t = t_data[indices, :]
42
43     x1, x2 = x[:, 0:1], x[:, 1:2]
44
45     x1.requires_grad_()
46     x2.requires_grad_()
47     t.requires_grad_()
48
49     optimizer.zero_grad()
50
51     # Denoting by u the realization function of the ANN, compute
52     # u(0, x) for each x in the batch
53     u0 = N(torch.hstack((torch.zeros_like(t), x)))
54     # Compute the loss for the initial condition
55     initial_loss = (u0 - phi(x)).square().mean()
56
57     # Compute the partial derivatives using automatic
58     # differentiation
59     u = N(torch.hstack((t, x1, x2)))
60     ones = torch.ones_like(u)
61     u_t = grad(u, t, ones, create_graph=True)[0]
62     u_x1 = grad(u, x1, ones, create_graph=True)[0]
63     u_x2 = grad(u, x2, ones, create_graph=True)[0]
64     ones = torch.ones_like(u_x1)
65     u_x1x1 = grad(u_x1, x1, ones, create_graph=True)[0]
66     u_x2x2 = grad(u_x2, x2, ones, create_graph=True)[0]
67
68     # Compute the loss for the PDE
69     Laplace = u_x1x1 + u_x2x2
70     pde_loss = (u_t - (0.005 * Laplace + u - u**3)).square().mean()
71
72     # Compute the total loss and perform a gradient step
73     loss = initial_loss + pde_loss
74     loss.backward()
75     optimizer.step()
76
77
78     ### Plot the solution at different times
79
80     mesh = 128
81     a, b = -3, 3
82
83     gs = GridSpec(2, 4, width_ratios=[1, 1, 1, 0.05])
84     fig = plt.figure(figsize=(16, 10), dpi=300)
85
86     x, y = torch.meshgrid(
87         torch.linspace(a, b, mesh),
88         torch.linspace(a, b, mesh),
```

```

89     indexing="xy"
90 )
91 x = x.reshape((mesh * mesh, 1)).to(dev)
92 y = y.reshape((mesh * mesh, 1)).to(dev)
93
94 for i in range(6):
95     t = torch.full((mesh * mesh, 1), i * T / 5).to(dev)
96     z = N(torch.cat((t, x, y), 1))
97     z = z.detach().cpu().numpy().reshape((mesh, mesh))
98
99     ax = fig.add_subplot(gs[i // 3, i % 3])
100    ax.set_title(f"t = {i * T / 5}")
101    ax.imshow(
102        z, cmap="viridis", extent=[a, b, a, b], vmin=-1.2, vmax=1.2
103    )
104
105 # Add the colorbar to the figure
106 norm = plt.Normalize(vmin=-1.2, vmax=1.2)
107 sm = ScalarMappable(cmap="viridis", norm=norm)
108 cax = fig.add_subplot(gs[:, 3])
109 fig.colorbar(sm, cax=cax, orientation='vertical')
110
111 fig.savefig("../plots/pinn.pdf", bbox_inches="tight")

```

Source code 16.1 ([code/pinn.py](#)): A simple implementation in PYTORCH of the PINN method, computing an approximation of the function $u \in C^{1,2}([0,3] \times \mathbb{R}^2, \mathbb{R})$ which satisfies for all $t \in [0, 2]$, $x \in \mathbb{R}^2$ that $(\frac{\partial u}{\partial t})(t, x) = \frac{1}{200}(\Delta_x u)(t, x) + u(t, x) - [u(t, x)]^3$ and $u(0, x) = \sin(\|x\|_2^2)$ (cf. Definition 3.3.4). The plot created by this code is shown in Figure 16.1.

16.4 Implementation of DGMs

In Source code 16.2 below we present a simple implementation of the DGM, as explained in Section 16.2 above, for finding an approximation for a solution $u \in C^{1,2}([0,3] \times \mathbb{R}^2)$ of the two-dimensional Allen–Cahn-type semilinear heat equation

$$(\frac{\partial u}{\partial t})(t, x) = \frac{1}{200}(\Delta_x u)(t, x) + u(t, x) - [u(t, x)]^3 \quad (16.20)$$

with $u(0, x) = \sin(x_1) \sin(x_2)$ for $t \in [0, 3]$, $x = (x_1, x_2) \in \mathbb{R}^2$. As originally proposed in Sirignano & Spiliopoulos [400], this implementation chooses for each training step a batch of 256 realizations of the random variable $(\mathcal{T}, \mathcal{X})$, where \mathcal{T} is continuously uniformly distributed on $[0, 3]$ and where \mathcal{X} is normally distributed on \mathbb{R}^2 with mean $0 \in \mathbb{R}^2$ and covariance $4 I_2 \in \mathbb{R}^{2 \times 2}$ (cf. Definition 1.5.5). Like the PINN implementation in Source code 16.1, it trains a fully connected feed-forward ANN with 4 hidden layers (with 50

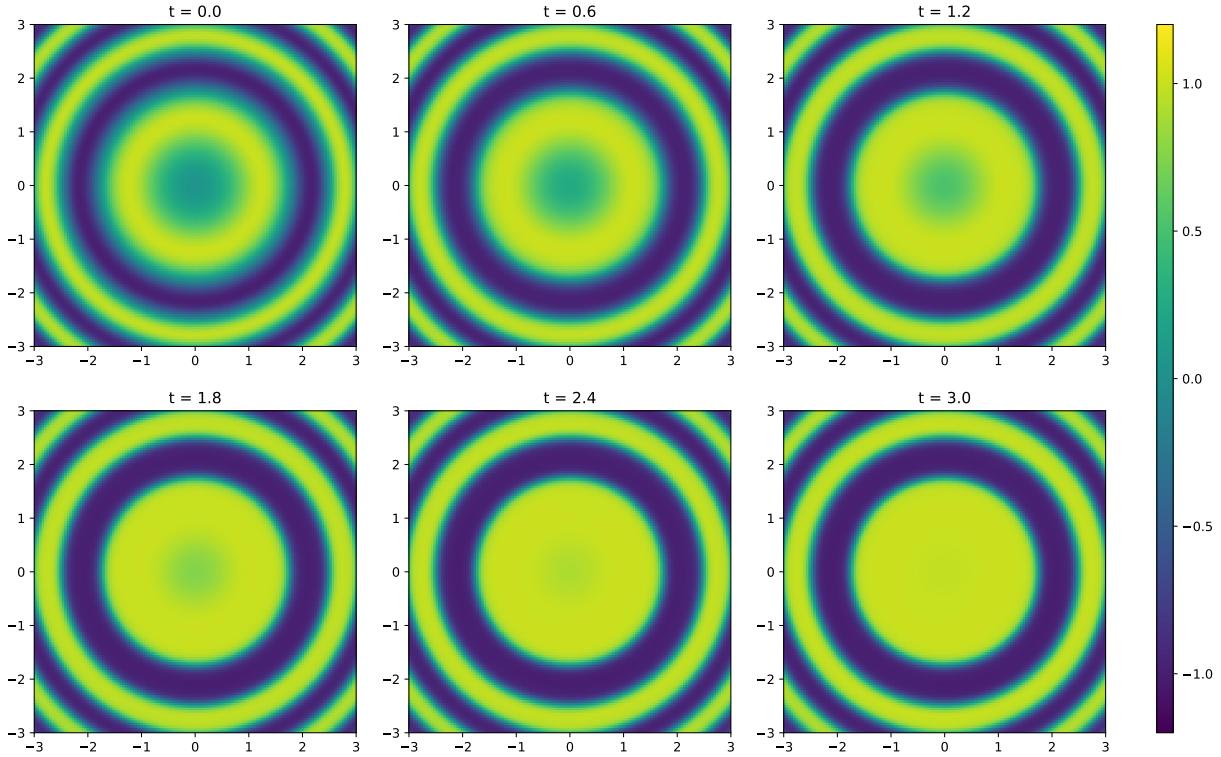


Figure 16.1 ([plots/pinn.pdf](#)): Plots for the functions $[-3, 3]^2 \ni x \mapsto U(t, x) \in \mathbb{R}$, where $t \in \{0, 0.6, 1.2, 1.8, 2.4, 3\}$ and where $U \in C([0, 3] \times \mathbb{R}^2, \mathbb{R})$ is an approximation of the function $u \in C^{1,2}([0, 3] \times \mathbb{R}^2, \mathbb{R})$ which satisfies for all $t \in [0, 3]$, $x \in \mathbb{R}^2$ that $\left(\frac{\partial u}{\partial t}\right)(t, x) = \frac{1}{200}(\Delta u)(t, x) + u(t, x) - [u(t, x)]^3$ and $u(0, x) = \sin(\|x\|_2^2)$ computed by means of the PINN method as implemented in Source code 16.1 (cf. Definition 3.3.4).

neurons on each hidden layer) and using the swish activation function with parameter 1 (cf. Section 1.2.8). The training is performed using the Adam SGD optimization method (cf. Section 7.9). A plot of the resulting approximation of the solution u after 30000 training steps is shown in Figure 16.2.

```

1 import torch
2 import matplotlib.pyplot as plt
3 from torch.autograd import grad
4 from matplotlib.gridspec import GridSpec
5 from matplotlib.cm import ScalarMappable
6
7
8 dev = torch.device("cuda:0" if torch.cuda.is_available() else
9         "cpu")
10
11 T = 3.0 # the time horizon

```

```

12 # The initial value
13 def phi(x):
14     return x.sin().prod(axis=1, keepdims=True)
15
16 torch.manual_seed(0)
17
18 # We use a network with 4 hidden layers of 50 neurons each and the
19 # Swish activation function (called SiLU in PyTorch)
20 N = torch.nn.Sequential(
21     torch.nn.Linear(3, 50), torch.nn.SiLU(),
22     torch.nn.Linear(50, 50), torch.nn.SiLU(),
23     torch.nn.Linear(50, 50), torch.nn.SiLU(),
24     torch.nn.Linear(50, 50), torch.nn.SiLU(),
25     torch.nn.Linear(50, 1),
26 ).to(dev)
27
28
29 optimizer = torch.optim.Adam(N.parameters(), lr=3e-4)
30
31 J = 256 # the batch size
32
33 for i in range(30000):
34     # Choose a random batch of training samples
35     x = torch.randn(J, 2).to(dev) * 2
36     t = torch.rand(J, 1).to(dev) * T
37
38     x1 = x[:, 0:1]
39     x2 = x[:, 1:2]
40
41     x1.requires_grad_()
42     x2.requires_grad_()
43     t.requires_grad_()
44
45     optimizer.zero_grad()
46
47     # Denoting by u the realization function of the ANN, compute
48     # u(0, x) for each x in the batch
49     u0 = N(torch.hstack((torch.zeros_like(t), x)))
50     # Compute the loss for the initial condition
51     initial_loss = (u0 - phi(x)).square().mean()
52
53     # Compute the partial derivatives using automatic
54     # differentiation
55     u = N(torch.hstack((t, x1, x2)))
56     ones = torch.ones_like(u)
57     u_t = grad(u, t, ones, create_graph=True)[0]
58     u_x1 = grad(u, x1, ones, create_graph=True)[0]
59     u_x2 = grad(u, x2, ones, create_graph=True)[0]
60     ones = torch.ones_like(u_x1)

```

```

61     u_x1x1 = grad(u_x1, x1, ones, create_graph=True)[0]
62     u_x2x2 = grad(u_x2, x2, ones, create_graph=True)[0]
63
64     # Compute the loss for the PDE
65     Laplace = u_x1x1 + u_x2x2
66     pde_loss = (u_t - (0.005 * Laplace + u - u**3)).square().mean()
67
68     # Compute the total loss and perform a gradient step
69     loss = initial_loss + pde_loss
70     loss.backward()
71     optimizer.step()
72
73
74     ### Plot the solution at different times
75
76     mesh = 128
77     a, b = -torch.pi, torch.pi
78
79     gs = GridSpec(2, 4, width_ratios=[1, 1, 1, 0.05])
80     fig = plt.figure(figsize=(16, 10), dpi=300)
81
82     x, y = torch.meshgrid(
83         torch.linspace(a, b, mesh),
84         torch.linspace(a, b, mesh),
85         indexing="xy"
86     )
87     x = x.reshape((mesh * mesh, 1)).to(dev)
88     y = y.reshape((mesh * mesh, 1)).to(dev)
89
90     for i in range(6):
91         t = torch.full((mesh * mesh, 1), i * T / 5).to(dev)
92         z = N(torch.cat((t, x, y), 1))
93         z = z.detach().cpu().numpy().reshape((mesh, mesh))
94
95         ax = fig.add_subplot(gs[i // 3, i % 3])
96         ax.set_title(f"t = {i * T / 5}")
97         ax.imshow(
98             z, cmap="viridis", extent=[a, b, a, b], vmin=-1.2, vmax=1.2
99         )
100
101    # Add the colorbar to the figure
102    norm = plt.Normalize(vmin=-1.2, vmax=1.2)
103    sm = ScalarMappable(cmap="viridis", norm=norm)
104    cax = fig.add_subplot(gs[:, 3])
105    fig.colorbar(sm, cax=cax, orientation='vertical')
106
107    fig.savefig("../plots/dgm.pdf", bbox_inches="tight")

```

Source code 16.2 ([code/dgm.py](#)): A simple implementation in PYTORCH of the deep Galerkin method, computing an approximation of the function $u \in C^{1,2}([0, 3] \times \mathbb{R}^2, \mathbb{R})$ which satisfies for all $t \in [0, 3]$, $x = (x_1, x_2) \in \mathbb{R}^2$ that $(\frac{\partial u}{\partial t})(t, x) = \frac{1}{200}(\Delta_x u)(t, x) + u(t, x) - [u(t, x)]^3$ and $u(0, x) = \sin(x_1) \sin(x_2)$. The plot created by this code is shown in Figure 16.2.

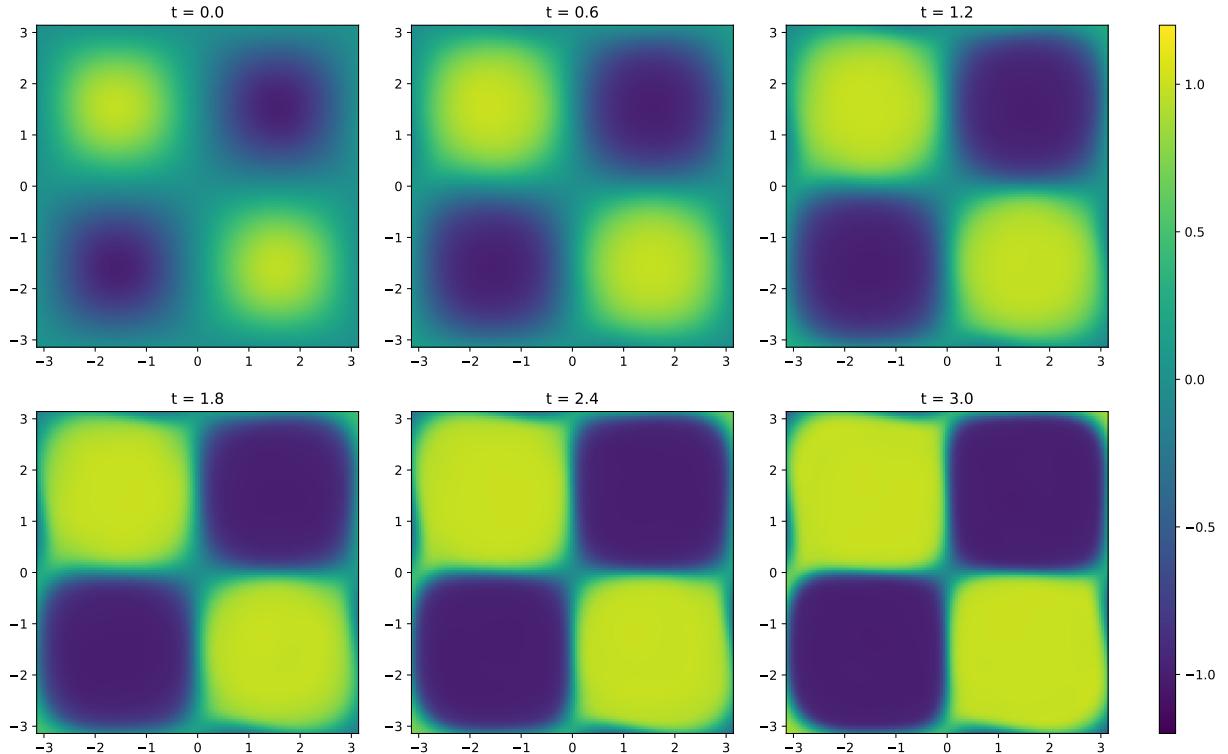


Figure 16.2 ([plots/dgm.pdf](#)): Plots for the functions $[-\pi, \pi]^2 \ni x \mapsto U(t, x) \in \mathbb{R}$, where $t \in \{0, 0.6, 1.2, 1.8, 2.4, 3\}$ and where $U \in C([0, 3] \times \mathbb{R}^2, \mathbb{R})$ is an approximation of the function $u \in C^{1,2}([0, 3] \times \mathbb{R}^2, \mathbb{R})$ which satisfies for all $t \in [0, 3]$, $x = (x_1, x_2) \in \mathbb{R}^2$ that $u(0, x) = \sin(x_1) \sin(x_2)$ and $(\frac{\partial u}{\partial t})(t, x) = \frac{1}{200}(\Delta_x u)(t, x) + u(t, x) - [u(t, x)]^3$ computed by means of Source code 16.2.

Chapter 17

Deep Kolmogorov methods (DKMs)

The PINNs and the DGMs presented in Chapter 16 do, on the one hand, not exploit a lot of structure of the underlying PDE in the process of setting up the associated stochastic optimization problems and have as such the key advantage to be very widely applicable deep learning methods for PDEs. On the other hand, deep learning methods for PDEs that in some way exploit the specific structure of the considered PDE problem often result in more accurate approximations (cf., for instance, Beck et al. [24] and the references therein). In particular, there are several deep learning approximation methods in the literature which exploit in the process of setting up stochastic optimization problems that the PDE itself admits a stochastic representation. In the literature there are a lot of deep learning methods which are based on such stochastic formulations of PDEs and therefore have a strong link to stochastic analysis and formulas of the Feynman–Kac-type (cf., for example, [20, 125, 152, 196, 218, 356] and the references therein).

The schemes in Beck et al. [19], which we refer to as DKMs, belong to the simplest of such deep learning methods for PDEs. In this chapter we present in Sections 17.1, 17.2, 17.3, and 17.4 theoretical considerations leading to a reformulation of heat PDE problems as stochastic optimization problems (see Proposition 17.4.1 below), we use these theoretical considerations to derive DKMs in the specific case of heat equations in Section 17.5, and we present an implementation of DKMs in the case of a simple two-dimensional heat equation in Section 17.6.

Sections 17.1 and 17.2 are slightly modified extracts from Beck et al. [18], Section 17.3 is inspired by Beck et al. [23, Section 2], and Sections 17.4 and 17.5 are inspired by Beck et al. [18].

17.1 Stochastic optimization problems for expectations of random variables

Lemma 17.1.1. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X: \Omega \rightarrow \mathbb{R}$ be a random variable with $\mathbb{E}[|X|^2] < \infty$. Then

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

$$\mathbb{E}[|X - y|^2] = \mathbb{E}[|X - \mathbb{E}[X]|^2] + |\mathbb{E}[X] - y|^2, \quad (17.1)$$

(ii) there exists a unique $z \in \mathbb{R}$ such that

$$\mathbb{E}[|X - z|^2] = \inf_{y \in \mathbb{R}} \mathbb{E}[|X - y|^2], \quad (17.2)$$

and

(iii) it holds that

$$\mathbb{E}[|X - \mathbb{E}[X]|^2] = \inf_{y \in \mathbb{R}} \mathbb{E}[|X - y|^2]. \quad (17.3)$$

Proof of Lemma 17.1.1. Note that Lemma 7.2.5 proves item (i). Observe that item (i) establishes items (ii) and (iii). The proof of Lemma 17.1.1 is thus complete. \square

17.2 Stochastic optimization problems for expectations of random fields

Proposition 17.2.1. Let $d \in \mathbb{N}$, $a \in \mathbb{R}$, $b \in (a, \infty)$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $X = (X_x)_{x \in [a, b]^d}: [a, b]^d \times \Omega \rightarrow \mathbb{R}$ be $(\mathcal{B}([a, b]^d) \otimes \mathcal{F})/\mathcal{B}(\mathbb{R})$ -measurable, assume for every $x \in [a, b]^d$ that $\mathbb{E}[|X_x|^2] < \infty$, and assume that $[a, b]^d \ni x \mapsto \mathbb{E}[X_x] \in \mathbb{R}$ is continuous. Then

(i) there exists a unique $u \in C([a, b]^d, \mathbb{R})$ such that

$$\int_{[a, b]^d} \mathbb{E}[|X_x - u(x)|^2] dx = \inf_{v \in C([a, b]^d, \mathbb{R})} \left(\int_{[a, b]^d} \mathbb{E}[|X_x - v(x)|^2] dx \right) \quad (17.4)$$

and

(ii) it holds for all $x \in [a, b]^d$ that $u(x) = \mathbb{E}[X_x]$.

Proof of Proposition 17.2.1. Note that item (i) in Lemma 17.1.1 and the assumption that for all $x \in [a, b]^d$ it holds that $\mathbb{E}[|X_x|^2] < \infty$ imply that for every function $u: [a, b]^d \rightarrow \mathbb{R}$ and every $x \in [a, b]^d$ it holds that

$$\mathbb{E}[|X_x - u(x)|^2] = \mathbb{E}[|X_x - \mathbb{E}[X_x]|^2] + |\mathbb{E}[X_x] - u(x)|^2. \quad (17.5)$$

Fubini's theorem (see, for instance, Klenke [262, Theorem 14.16]) therefore demonstrates that for all $u \in C([a, b]^d, \mathbb{R})$ it holds that

$$\int_{[a,b]^d} \mathbb{E}[|X_x - u(x)|^2] dx = \int_{[a,b]^d} \mathbb{E}[|X_x - \mathbb{E}[X_x]|^2] dx + \int_{[a,b]^d} |\mathbb{E}[X_x] - u(x)|^2 dx. \quad (17.6)$$

This ensures that

$$\begin{aligned} & \int_{[a,b]^d} \mathbb{E}[|X_x - \mathbb{E}[X_x]|^2] dx \\ & \geq \inf_{v \in C([a,b]^d, \mathbb{R})} \left(\int_{[a,b]^d} \mathbb{E}[|X_x - v(x)|^2] dx \right) \\ & = \inf_{v \in C([a,b]^d, \mathbb{R})} \left(\int_{[a,b]^d} \mathbb{E}[|X_x - \mathbb{E}[X_x]|^2] dx + \int_{[a,b]^d} |\mathbb{E}[X_x] - v(x)|^2 dx \right) \end{aligned} \quad (17.7)$$

The assumption that $[a, b]^d \ni x \mapsto \mathbb{E}[X_x] \in \mathbb{R}$ is continuous hence shows that

$$\begin{aligned} \int_{[a,b]^d} \mathbb{E}[|X_x - \mathbb{E}[X_x]|^2] dx & \geq \inf_{v \in C([a,b]^d, \mathbb{R})} \left(\int_{[a,b]^d} \mathbb{E}[|X_x - \mathbb{E}[X_x]|^2] dx \right) \\ & = \int_{[a,b]^d} \mathbb{E}[|X_x - \mathbb{E}[X_x]|^2] dx. \end{aligned} \quad (17.8)$$

Therefore, we obtain that

$$\int_{[a,b]^d} \mathbb{E}[|X_x - \mathbb{E}[X_x]|^2] dx = \inf_{v \in C([a,b]^d, \mathbb{R})} \left(\int_{[a,b]^d} \mathbb{E}[|X_x - v(x)|^2] dx \right). \quad (17.9)$$

The fact that the function $[a, b]^d \ni x \mapsto \mathbb{E}[X_x] \in \mathbb{R}$ is continuous hence proves that there exists $u \in C([a, b]^d, \mathbb{R})$ such that

$$\int_{[a,b]^d} \mathbb{E}[|X_x - u(x)|^2] dx = \inf_{v \in C([a,b]^d, \mathbb{R})} \left(\int_{[a,b]^d} \mathbb{E}[|X_x - v(x)|^2] dx \right). \quad (17.10)$$

Furthermore, observe that (17.6) and (17.9) establish that for all $u \in C([a, b]^d, \mathbb{R})$ with

$$\int_{[a,b]^d} \mathbb{E}[|X_x - u(x)|^2] dx = \inf_{v \in C([a,b]^d, \mathbb{R})} \left(\int_{[a,b]^d} \mathbb{E}[|X_x - v(x)|^2] dx \right) \quad (17.11)$$

it holds that

$$\begin{aligned} & \int_{[a,b]^d} \mathbb{E}[|X_x - \mathbb{E}[X_x]|^2] dx \\ & = \inf_{v \in C([a,b]^d, \mathbb{R})} \left(\int_{[a,b]^d} \mathbb{E}[|X_x - v(x)|^2] dx \right) = \int_{[a,b]^d} \mathbb{E}[|X_x - u(x)|^2] dx \\ & = \int_{[a,b]^d} \mathbb{E}[|X_x - \mathbb{E}[X_x]|^2] dx + \int_{[a,b]^d} |\mathbb{E}[X_x] - u(x)|^2 dx. \end{aligned} \quad (17.12)$$

Therefore, we obtain that for all $u \in C([a, b]^d, \mathbb{R})$ with

$$\int_{[a,b]^d} \mathbb{E}[|X_x - u(x)|^2] dx = \inf_{v \in C([a,b]^d, \mathbb{R})} \left(\int_{[a,b]^d} \mathbb{E}[|X_x - v(x)|^2] dx \right) \quad (17.13)$$

it holds that

$$\int_{[a,b]^d} |\mathbb{E}[X_x] - u(x)|^2 dx = 0. \quad (17.14)$$

This and the assumption that $[a, b]^d \ni x \mapsto \mathbb{E}[X_x] \in \mathbb{R}$ is continuous imply that for all $y \in [a, b]^d$, $u \in C([a, b]^d, \mathbb{R})$ with

$$\int_{[a,b]^d} \mathbb{E}[|X_x - u(x)|^2] dx = \inf_{v \in C([a,b]^d, \mathbb{R})} \left(\int_{[a,b]^d} \mathbb{E}[|X_x - v(x)|^2] dx \right) \quad (17.15)$$

it holds that $u(y) = \mathbb{E}[X_y]$. Combining this with (17.10) proves items (i) and (ii). The proof of Proposition 17.2.1 is thus complete. \square

17.3 Feynman–Kac formulas

17.3.1 Feynman–Kac formulas providing existence of solutions

Lemma 17.3.1 (A variant of Lebesgue’s theorem on dominated convergence). *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, for every $n \in \mathbb{N}_0$ let $X_n: \Omega \rightarrow \mathbb{R}$ be a random variable, assume for all $\varepsilon \in (0, \infty)$ that*

$$\limsup_{n \rightarrow \infty} \mathbb{P}(|X_n - X_0| > \varepsilon) = 0, \quad (17.16)$$

let $Y: \Omega \rightarrow \mathbb{R}$ be a random variable with $\mathbb{E}[|Y|] < \infty$, and assume for all $n \in \mathbb{N}$ that $\mathbb{P}(|X_n| \leq Y) = 1$. Then

- (i) *it holds that $\limsup_{n \rightarrow \infty} \mathbb{E}[|X_n - X_0|] = 0$,*
- (ii) *it holds that $\mathbb{E}[|X_0|] < \infty$, and*
- (iii) *it holds that $\limsup_{n \rightarrow \infty} |\mathbb{E}[X_n] - \mathbb{E}[X_0]| = 0$.*

Proof of Lemma 17.3.1. Note that, for example, the variant of Lebesgue’s theorem on dominated convergence in Klenke [262, Corollary 6.26] establishes items (i), (ii), and (iii). The proof of Lemma 17.3.1 is thus complete. \square

Proposition 17.3.2. Let $T \in (0, \infty)$, $d, m \in \mathbb{N}$, $B \in \mathbb{R}^{d \times m}$, $\varphi \in C^2(\mathbb{R}^d, \mathbb{R})$ satisfy

$$\sup_{x \in \mathbb{R}^d} \left[\sum_{i,j=1}^d \left(|\varphi(x)| + \left| \left(\frac{\partial}{\partial x_i} \varphi \right)(x) \right| + \left| \left(\frac{\partial^2}{\partial x_i \partial x_j} \varphi \right)(x) \right| \right) \right] < \infty, \quad (17.17)$$

let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $Z: \Omega \rightarrow \mathbb{R}^m$ be a standard normal random variable, and let $u: [0, T] \times \mathbb{R}^d \rightarrow \mathbb{R}$ satisfy for all $t \in [0, T]$, $x \in \mathbb{R}^d$ that

$$u(t, x) = \mathbb{E}[\varphi(x + \sqrt{t}BZ)]. \quad (17.18)$$

Then

- (i) it holds that $u \in C^{1,2}([0, T] \times \mathbb{R}^d, \mathbb{R})$ and
- (ii) it holds for all $t \in [0, T]$, $x \in \mathbb{R}^d$ that

$$\left(\frac{\partial u}{\partial t} \right)(t, x) = \frac{1}{2} \text{Trace}(BB^*(\text{Hess}_x u)(t, x)) \quad (17.19)$$

(cf. Definition 2.4.5).

Proof of Proposition 17.3.2. Throughout this proof, let

$$e_1 = (1, 0, \dots, 0), e_2 = (0, 1, \dots, 0), \dots, e_m = (0, \dots, 0, 1) \in \mathbb{R}^m \quad (17.20)$$

and for every $t \in [0, T]$, $x \in \mathbb{R}^d$ let $\psi_{t,x}: \mathbb{R}^m \rightarrow \mathbb{R}$, satisfy for all $y \in \mathbb{R}^m$ that $\psi_{t,x}(y) = \varphi(x + \sqrt{t}By)$. Note that the assumption that $\varphi \in C^2(\mathbb{R}^d, \mathbb{R})$, the chain rule, Lemma 17.3.1, and (17.17) demonstrate that

- (I) for all $x \in \mathbb{R}^d$ it holds that $(0, T) \ni t \mapsto u(t, x) \in \mathbb{R}$ is differentiable,
- (II) for all $t \in [0, T]$ it holds that $\mathbb{R}^d \ni x \mapsto u(t, x) \in \mathbb{R}$ is twice differentiable,
- (III) for all $t \in (0, T]$, $x \in \mathbb{R}^d$ it holds that

$$\left(\frac{\partial u}{\partial t} \right)(t, x) = \mathbb{E}[\langle (\nabla \varphi)(x + \sqrt{t}BZ), \frac{1}{2\sqrt{t}}BZ \rangle], \quad (17.21)$$

and

- (IV) for all $t \in [0, T]$, $x \in \mathbb{R}^d$ it holds that

$$(\text{Hess}_x u)(t, x) = \mathbb{E}[(\text{Hess } \varphi)(x + \sqrt{t}BZ)] \quad (17.22)$$

(cf. Definition 1.4.7). Note that items (III) and (IV), the assumption that $\varphi \in C^2(\mathbb{R}^d, \mathbb{R})$, the assumption that

$$\sup_{x \in \mathbb{R}^d} \left[\sum_{i,j=1}^d \left(|\varphi(x)| + \left| \left(\frac{\partial}{\partial x_i} \varphi \right)(x) \right| + \left| \left(\frac{\partial^2}{\partial x_i \partial x_j} \varphi \right)(x) \right| \right) \right] < \infty, \quad (17.23)$$

the fact that $\mathbb{E}[\|Z\|_2] < \infty$, and Lemma 17.3.1 ensure that

$$(0, T] \times \mathbb{R}^d \ni (t, x) \mapsto \left(\frac{\partial u}{\partial t}\right)(t, x) \in \mathbb{R} \quad (17.24)$$

and

$$[0, T] \times \mathbb{R}^d \ni (t, x) \mapsto (\text{Hess}_x u)(t, x) \in \mathbb{R}^{d \times d} \quad (17.25)$$

are continuous (cf. Definition 3.3.4). Furthermore, observe that item (IV) and the fact that for all $X \in \mathbb{R}^{m \times d}$, $Y \in \mathbb{R}^{d \times m}$ it holds that $\text{Trace}(XY) = \text{Trace}(YX)$ show that for all $t \in (0, T]$, $x \in \mathbb{R}^d$ it holds that

$$\begin{aligned} \frac{1}{2} \text{Trace}\left(BB^*(\text{Hess}_x u)(t, x)\right) &= \mathbb{E}\left[\frac{1}{2} \text{Trace}\left(BB^*(\text{Hess } \varphi)(x + \sqrt{t}BZ)\right)\right] \\ &= \frac{1}{2} \mathbb{E}\left[\text{Trace}\left(B^*(\text{Hess } \varphi)(x + \sqrt{t}BZ)B\right)\right] = \frac{1}{2} \mathbb{E}\left[\sum_{k=1}^m \langle e_k, B^*(\text{Hess } \varphi)(x + \sqrt{t}BZ)Be_k \rangle\right] \\ &= \frac{1}{2} \mathbb{E}\left[\sum_{k=1}^m \langle Be_k, (\text{Hess } \varphi)(x + \sqrt{t}BZ)Be_k \rangle\right] = \frac{1}{2} \mathbb{E}\left[\sum_{k=1}^m \varphi''(x + \sqrt{t}BZ)(Be_k, Be_k)\right] \\ &= \frac{1}{2t} \mathbb{E}\left[\sum_{k=1}^m (\psi_{t,x})''(Z)(e_k, e_k)\right] = \frac{1}{2t} \mathbb{E}\left[\sum_{k=1}^m \left(\frac{\partial^2}{\partial y_k^2} \psi_{t,x}\right)(Z)\right] = \frac{1}{2t} \mathbb{E}[(\Delta \psi_{t,x})(Z)] \end{aligned} \quad (17.26)$$

(cf. Definition 2.4.5). The assumption that $Z: \Omega \rightarrow \mathbb{R}^m$ is a standard normal random variable and integration by parts hence imply that for all $t \in (0, T]$, $x \in \mathbb{R}^d$ it holds that

$$\begin{aligned} \frac{1}{2} \text{Trace}\left(BB^*(\text{Hess}_x u)(t, x)\right) &= \frac{1}{2t} \int_{\mathbb{R}^m} (\Delta \psi_{t,x})(y) \left[\frac{\exp\left(-\frac{\langle y, y \rangle}{2}\right)}{(2\pi)^{m/2}} \right] dy = \frac{1}{2t} \int_{\mathbb{R}^m} \langle (\nabla \psi_{t,x})(y), y \rangle \left[\frac{\exp\left(-\frac{\langle y, y \rangle}{2}\right)}{(2\pi)^{m/2}} \right] dy \\ &= \frac{1}{2\sqrt{t}} \int_{\mathbb{R}^m} \left\langle B^*(\nabla \varphi)(x + \sqrt{t}By), y \right\rangle \left[\frac{\exp\left(-\frac{\langle y, y \rangle}{2}\right)}{(2\pi)^{m/2}} \right] dy \\ &= \frac{1}{2\sqrt{t}} \mathbb{E}[\langle B^*(\nabla \varphi)(x + \sqrt{t}BZ), Z \rangle] = \mathbb{E}[\langle (\nabla \varphi)(x + \sqrt{t}BZ), \frac{1}{2\sqrt{t}} BZ \rangle]. \end{aligned} \quad (17.27)$$

Item (III) therefore proves that for all $t \in (0, T]$, $x \in \mathbb{R}^d$ it holds that

$$\left(\frac{\partial u}{\partial t}\right)(t, x) = \frac{1}{2} \text{Trace}\left(BB^*(\text{Hess}_x u)(t, x)\right). \quad (17.28)$$

The fundamental theorem of calculus hence establishes that for all $t, s \in (0, T]$, $x \in \mathbb{R}^d$ it holds that

$$u(t, x) - u(s, x) = \int_s^t \left(\frac{\partial u}{\partial t}\right)(r, x) dr = \int_s^t \frac{1}{2} \text{Trace}\left(BB^*(\text{Hess}_x u)(r, x)\right) dr. \quad (17.29)$$

The fact that $[0, T] \times \mathbb{R}^d \ni (t, x) \mapsto (\text{Hess}_x u)(t, x) \in \mathbb{R}^{d \times d}$ is continuous therefore demonstrates for all $t \in (0, T]$, $x \in \mathbb{R}^d$ that

$$\frac{u(t, x) - u(0, x)}{t} = \lim_{s \searrow 0} \left[\frac{u(t, x) - u(s, x)}{t} \right] = \frac{1}{t} \int_0^t \frac{1}{2} \text{Trace}(BB^*(\text{Hess}_x u)(r, x)) dr. \quad (17.30)$$

This and the fact that $[0, T] \times \mathbb{R}^d \ni (t, x) \mapsto (\text{Hess}_x u)(t, x) \in \mathbb{R}^{d \times d}$ is continuous ensure that for all $x \in \mathbb{R}^d$ it holds that

$$\begin{aligned} & \limsup_{t \searrow 0} \left| \frac{u(t, x) - u(0, x)}{t} - \frac{1}{2} \text{Trace}(BB^*(\text{Hess}_x u)(0, x)) \right| \\ & \leq \limsup_{t \searrow 0} \left[\frac{1}{t} \int_0^t \left| \frac{1}{2} \text{Trace}(BB^*(\text{Hess}_x u)(s, x)) - \frac{1}{2} \text{Trace}(BB^*(\text{Hess}_x u)(0, x)) \right| ds \right] \\ & \leq \limsup_{t \searrow 0} \left[\sup_{s \in [0, t]} \left| \frac{1}{2} \text{Trace}(BB^*((\text{Hess}_x u)(s, x) - (\text{Hess}_x u)(0, x))) \right| \right] = 0. \end{aligned} \quad (17.31)$$

Item (I) hence shows that for all $x \in \mathbb{R}^d$ it holds that $[0, T] \ni t \mapsto u(t, x) \in \mathbb{R}$ is differentiable. Combining this with (17.31) and (17.28) ensures that for all $t \in [0, T]$, $x \in \mathbb{R}^d$ it holds that

$$(\frac{\partial u}{\partial t})(t, x) = \frac{1}{2} \text{Trace}(BB^*(\text{Hess}_x u)(t, x)). \quad (17.32)$$

This and the fact that $[0, T] \times \mathbb{R}^d \ni (t, x) \mapsto (\text{Hess}_x u)(t, x) \in \mathbb{R}^{d \times d}$ is continuous prove item (i). Note that (17.32) establishes item (ii). The proof of Proposition 17.3.2 is thus complete. \square

Definition 17.3.3 (Standard Brownian motions). *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Then we say that W is an m -dimensional \mathbb{P} -standard Brownian motion (we say that W is a \mathbb{P} -standard Brownian motion, we say that W is a standard Brownian motion) if and only if there exists $T \in (0, \infty)$ such that*

- (i) *it holds that $m \in \mathbb{N}$,*
- (ii) *it holds that $W: [0, T] \times \Omega \times \mathbb{R}^m$ is a function,*
- (iii) *it holds for all $\omega \in \Omega$ that $[0, T] \ni s \mapsto W_s(\omega) \in \mathbb{R}^m$ is continuous,*
- (iv) *it holds for all $\omega \in \Omega$ that $W_0(\omega) = 0 \in \mathbb{R}^m$,*
- (v) *it holds for all $t_1 \in [0, T]$, $t_2 \in [0, T]$ with $t_1 < t_2$ that $\Omega \ni \omega \mapsto (t_2 - t_1)^{-1/2}(W_{t_2}(\omega) - W_{t_1}(\omega)) \in \mathbb{R}^m$ is a standard normal random variable, and*

- (vi) it holds for all $n \in \{3, 4, 5, \dots\}$, $t_1, t_2, \dots, t_n \in [0, T]$ with $t_1 \leq t_2 \leq \dots \leq t_n$ that $W_{t_2} - W_{t_1}, W_{t_3} - W_{t_2}, \dots, W_{t_n} - W_{t_{n-1}}$ are independent.

```

1 import numpy as np
2 import matplotlib.pyplot as plt
3
4 def generate_brownian_motion(T, N):
5     increments = np.random.randn(N) * np.sqrt(T/N)
6     BM = np.cumsum(increments)
7     BM = np.insert(BM, 0, 0)
8     return BM
9
10 T = 1
11 N = 1000
12 t_values = np.linspace(0, T, N+1)
13
14 fig, axarr = plt.subplots(2, 2)
15
16 for i in range(2):
17     for j in range(2):
18         BM = generate_brownian_motion(T, N)
19         axarr[i, j].plot(t_values, BM)
20
21 plt.tight_layout()
22 plt.savefig('../plots/brownian_motions.pdf')
23 plt.show()
```

Source code 17.1 ([code/brownian_motion.py](#)): PYTHON code producing four trajectories of a one-dimensional standard Brownian motion.

Corollary 17.3.4. Let $T \in (0, \infty)$, $d, m \in \mathbb{N}$, $B \in \mathbb{R}^{d \times m}$, $\varphi \in C^2(\mathbb{R}^d, \mathbb{R})$ satisfy

$$\sup_{x \in \mathbb{R}^d} \left[\sum_{i,j=1}^d \left(|\varphi(x)| + \left| \left(\frac{\partial}{\partial x_i} \varphi \right)(x) \right| + \left| \left(\frac{\partial^2}{\partial x_i \partial x_j} \varphi \right)(x) \right| \right) \right] < \infty, \quad (17.33)$$

let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $W: [0, T] \times \Omega \rightarrow \mathbb{R}^m$ be a standard Brownian motion, and let $u: [0, T] \times \mathbb{R}^d \rightarrow \mathbb{R}$ satisfy for all $t \in [0, T]$, $x \in \mathbb{R}^d$ that

$$u(t, x) = \mathbb{E}[\varphi(x + BW_t)] \quad (17.34)$$

(cf. Definition 17.3.3). Then

- (i) it holds that $u \in C^{1,2}([0, T] \times \mathbb{R}^d, \mathbb{R})$ and

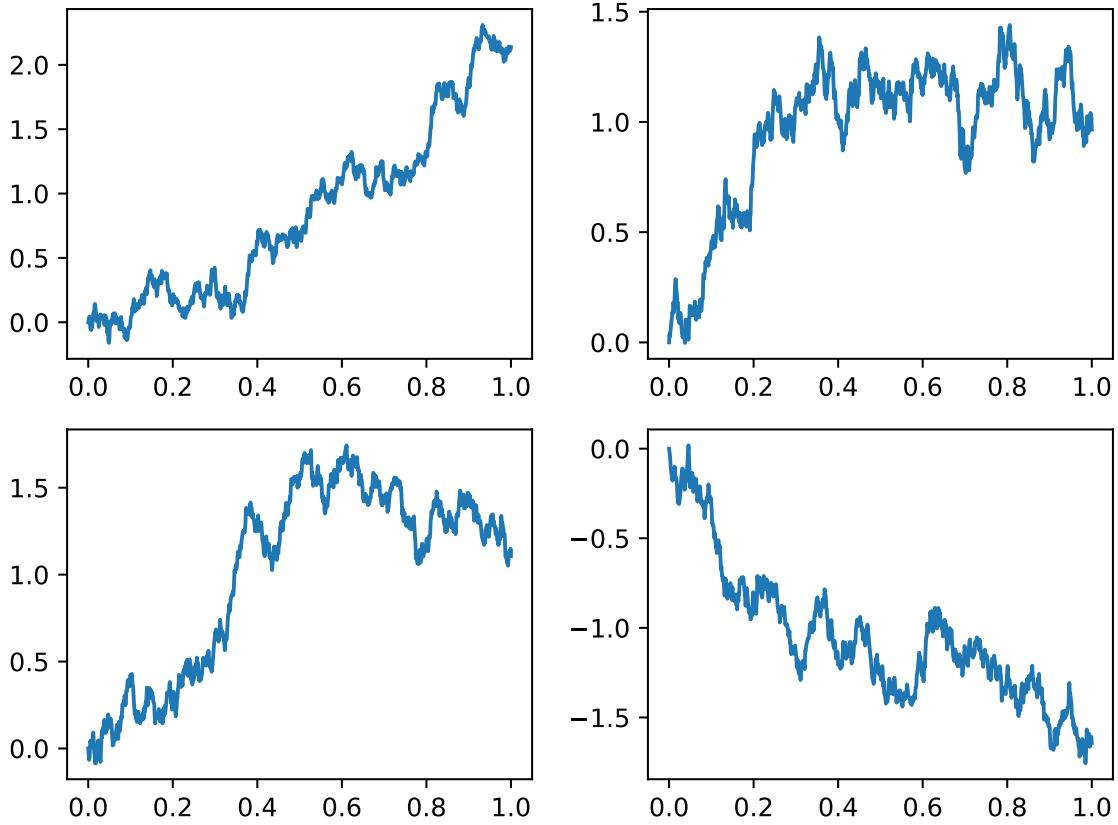


Figure 17.1 ([plots/brownian_motions.pdf](#)): Four trajectories of a one-dimensional standard Brownian motion

(ii) it holds for all $t \in [0, T]$, $x \in \mathbb{R}^d$ that

$$\left(\frac{\partial u}{\partial t}\right)(t, x) = \frac{1}{2} \operatorname{Trace}\left(BB^*(\operatorname{Hess}_x u)(t, x)\right) \quad (17.35)$$

(cf. Definition 2.4.5).

Proof of Corollary 17.3.4. First, observe that the assumption that $W: [0, T] \times \Omega \rightarrow \mathbb{R}^m$ is a standard Brownian motion implies that for all $t \in [0, T]$, $x \in \mathbb{R}^d$ it holds that

$$u(t, x) = \mathbb{E}[\varphi(x + BW_t)] = \mathbb{E}\left[\varphi\left(x + \sqrt{t}B \frac{W_T}{\sqrt{T}}\right)\right]. \quad (17.36)$$

The fact that $\frac{W_T}{\sqrt{T}}: \Omega \rightarrow \mathbb{R}^m$ is a standard normal random variable and Proposition 17.3.2 therefore prove items (i) and (ii). The proof of Corollary 17.3.4 is thus complete. \square

17.3.2 Feynman–Kac formulas providing uniqueness of solutions

Lemma 17.3.5 (A special case of Vitali's convergence theorem). *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $X_n: \Omega \rightarrow \mathbb{R}$, $n \in \mathbb{N}_0$, be random variables with*

$$\mathbb{P}(\limsup_{n \rightarrow \infty} |X_n - X_0| = 0) = 1, \quad (17.37)$$

and let $p \in (1, \infty)$ satisfy $\sup_{n \in \mathbb{N}} \mathbb{E}[|X_n|^p] < \infty$. Then

- (i) *it holds that $\limsup_{n \rightarrow \infty} \mathbb{E}[|X_n - X_0|] = 0$,*
- (ii) *it holds that $\mathbb{E}[|X_0|] < \infty$, and*
- (iii) *it holds that $\limsup_{n \rightarrow \infty} |\mathbb{E}[X_n] - \mathbb{E}[X_0]| = 0$.*

Proof of Lemma 17.3.5. First, note that the assumption that

$$\sup_{n \in \mathbb{N}} \mathbb{E}[|X_n|^p] < \infty \quad (17.38)$$

and, for instance, the consequence of de la Vallée-Poussin's theorem in Klenke [262, Corollary 6.21] demonstrate that $\{X_n: n \in \mathbb{N}\}$ is uniformly integrable. This, (17.37), and Vitali's convergence theorem in, for example, Klenke [262, Theorem 6.25] establish items (i) and (ii). Observe that items (i) and (ii) imply item (iii). The proof of Lemma 17.3.5 is thus complete. \square

Proposition 17.3.6. *Let $d \in \mathbb{N}$, $T, \rho \in (0, \infty)$, $f \in C([0, T] \times \mathbb{R}^d, \mathbb{R})$, let $u \in C^{1,2}([0, T] \times \mathbb{R}^d, \mathbb{R})$ have at most polynomially growing partial derivatives, assume for all $t \in [0, T]$, $x \in \mathbb{R}^d$ that*

$$\left(\frac{\partial u}{\partial t}\right)(t, x) = \rho (\Delta_x u)(t, x) + f(t, x), \quad (17.39)$$

let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $W: [0, T] \times \Omega \rightarrow \mathbb{R}^d$ be a standard Brownian motion (cf. Definition 17.3.3). Then it holds for all $t \in [0, T]$, $x \in \mathbb{R}^d$ that

$$u(t, x) = \mathbb{E} \left[u(0, x + \sqrt{2\rho} W_t) + \int_0^t f(t-s, x + \sqrt{2\rho} W_s) ds \right]. \quad (17.40)$$

Proof of Proposition 17.3.6. Throughout this proof, let $D_1: [0, T] \times \mathbb{R}^d \rightarrow \mathbb{R}$ satisfy for all $t \in [0, T]$, $x \in \mathbb{R}^d$ that

$$D_1(t, x) = \left(\frac{\partial u}{\partial t}\right)(t, x), \quad (17.41)$$

let $D_2 = (D_{2,1}, D_{2,2}, \dots, D_{2,d}) : [0, T] \times \mathbb{R}^d \rightarrow \mathbb{R}^d$ satisfy for all $t \in [0, T]$, $x \in \mathbb{R}^d$ that $D_2(t, x) = (\nabla_x u)(t, x)$, let $H = (H_{i,j})_{i,j \in \{1, 2, \dots, d\}} : [0, T] \times \mathbb{R}^d \rightarrow \mathbb{R}^{d \times d}$ satisfy for all $t \in [0, T]$, $x \in \mathbb{R}^d$ that

$$H(t, x) = (\text{Hess}_x u)(t, x), \quad (17.42)$$

let $\gamma : \mathbb{R}^d \rightarrow \mathbb{R}$ satisfy for all $z \in \mathbb{R}^d$ that

$$\gamma(z) = (2\pi)^{-d/2} \exp\left(-\frac{\|z\|_2^2}{2}\right), \quad (17.43)$$

and let $v_{t,x} : [0, t] \rightarrow \mathbb{R}$, $t \in [0, T]$, $x \in \mathbb{R}^d$, satisfy for all $t \in [0, T]$, $x \in \mathbb{R}^d$, $s \in [0, t]$ that

$$v_{t,x}(s) = \mathbb{E}[u(s, x + \sqrt{2\rho}W_{t-s})] \quad (17.44)$$

(cf. Definition 3.3.4). Note that the assumption that W is a standard Brownian motion ensures that for all $t \in (0, T]$, $s \in [0, t)$ it holds that $(t-s)^{-1/2}W_{t-s} : \Omega \rightarrow \mathbb{R}^d$ is a standard normal random variable. This shows that for all $t \in (0, T]$, $x \in \mathbb{R}^d$, $s \in [0, t)$ it holds that

$$\begin{aligned} v_{t,x}(s) &= \mathbb{E}[u(s, x + \sqrt{2\rho(t-s)}(t-s)^{-1/2}W_{t-s})] \\ &= \int_{\mathbb{R}^d} u(s, x + \sqrt{2\rho(t-s)}z) \gamma(z) dz. \end{aligned} \quad (17.45)$$

The assumption that u has at most polynomially growing partial derivatives, the fact that $(0, \infty) \ni s \mapsto \sqrt{s} \in (0, \infty)$ is differentiable, the chain rule, and Vitali's convergence theorem hence prove that for all $t \in (0, T]$, $x \in \mathbb{R}^d$, $s \in [0, t)$ it holds that $v_{t,x}|_{[0,t)} \in C^1([0, t), \mathbb{R})$ and

$$(v_{t,x})'(s) = \int_{\mathbb{R}^d} \left[D_1(s, x + \sqrt{2\rho(t-s)}z) + \left\langle D_2(s, x + \sqrt{2\rho(t-s)}z), \frac{-\rho z}{\sqrt{2\rho(t-s)}} \right\rangle \right] \gamma(z) dz \quad (17.46)$$

(cf. Definition 1.4.7). Furthermore, observe that the fact that for all $z \in \mathbb{R}^d$ it holds that $(\nabla \gamma)(z) = -\gamma(z)z$ demonstrates that for all $t \in (0, T]$, $x \in \mathbb{R}^d$, $s \in [0, t)$ it holds that

$$\begin{aligned} &\int_{\mathbb{R}^d} \left\langle D_2(s, x + \sqrt{2\rho(t-s)}z), \frac{-\rho z}{\sqrt{2\rho(t-s)}} \right\rangle \gamma(z) dz \\ &= \int_{\mathbb{R}^d} \left\langle D_2(s, x + \sqrt{2\rho(t-s)}z), \frac{\rho(\nabla \gamma)(z)}{\sqrt{2\rho(t-s)}} \right\rangle dz \\ &= \frac{\rho}{\sqrt{2\rho(t-s)}} \sum_{i=1}^d \left[\int_{\mathbb{R}^d} D_{2,i}(s, x + \sqrt{2\rho(t-s)}z) \left(\frac{\partial \gamma}{\partial z_i} \right)(z_1, z_2, \dots, z_d) dz \right]. \end{aligned} \quad (17.47)$$

Moreover, note that integration by parts establishes that for all $t \in (0, T]$, $x \in \mathbb{R}^d$, $s \in [0, t)$,

$i \in \{1, 2, \dots, d\}$, $a \in \mathbb{R}$, $b \in (a, \infty)$ it holds that

$$\begin{aligned} & \int_a^b D_{2,i}(s, x + \sqrt{2\rho(t-s)}(z_1, z_2, \dots, z_d))(\frac{\partial \gamma}{\partial z_i})(z_1, z_2, \dots, z_d) dz_i \\ &= \left[D_{2,i}(s, x + \sqrt{2\rho(t-s)}(z_1, z_2, \dots, z_d))\gamma(z_1, z_2, \dots, z_d) \right]_{z_i=a}^{z_i=b} \\ & \quad - \int_a^b \sqrt{2\rho(t-s)} H_{i,i}(s, x + \sqrt{2\rho(t-s)}(z_1, z_2, \dots, z_d))\gamma(z_1, z_2, \dots, z_d) dz_i. \end{aligned} \quad (17.48)$$

The assumption that u has at most polynomially growing derivatives therefore implies that for all $t \in (0, T]$, $x \in \mathbb{R}^d$, $s \in [0, t)$, $i \in \{1, 2, \dots, d\}$ it holds that

$$\begin{aligned} & \int_{\mathbb{R}} D_{2,i}(s, x + \sqrt{2\rho(t-s)}(z_1, z_2, \dots, z_d))(\frac{\partial \gamma}{\partial z_i})(z_1, z_2, \dots, z_d) dz_i \\ &= -\sqrt{2\rho(t-s)} \int_{\mathbb{R}} H_{i,i}(s, x + \sqrt{2\rho(t-s)}(z_1, z_2, \dots, z_d))\gamma(z_1, z_2, \dots, z_d) dz_i. \end{aligned} \quad (17.49)$$

Combining this with (17.47) and Fubini's theorem ensures that for all $t \in (0, T]$, $x \in \mathbb{R}^d$, $s \in [0, t)$ it holds that

$$\begin{aligned} & \int_{\mathbb{R}^d} \left\langle D_2(s, x + \sqrt{2\rho(t-s)}z), \frac{-\rho z}{\sqrt{2\rho(t-s)}} \right\rangle \gamma(z) dz \\ &= -\rho \sum_{i=1}^d \int_{\mathbb{R}^d} H_{i,i}(s, x + \sqrt{2\rho(t-s)}(z))\gamma(z) dz \\ &= -\int_{\mathbb{R}^d} \rho \text{Trace}(H(s, x + \sqrt{2\rho(t-s)}(z)))\gamma(z) dz. \end{aligned} \quad (17.50)$$

This, (17.46), (17.39), and the fact that for all $t \in (0, T]$, $s \in [0, t)$ it holds that $(t-s)^{-1/2}W_{t-s}: \Omega \rightarrow \mathbb{R}^d$ is a standard normal random variable show that for all $t \in (0, T]$, $x \in \mathbb{R}^d$, $s \in [0, t)$ it holds that

$$\begin{aligned} (v_{t,x})'(s) &= \int_{\mathbb{R}^d} [D_1(s, x + \sqrt{2\rho(t-s)}z) - \rho \text{Trace}(H(s, x + \sqrt{2\rho(t-s)}z))] \gamma(z) dz \\ &= \int_{\mathbb{R}^d} f(s, x + \sqrt{2\rho(t-s)}z) \gamma(z) dz = \mathbb{E}[f(s, x + \sqrt{2\rho}W_{t-s})]. \end{aligned} \quad (17.51)$$

The fact that $W_0 = 0$, the fact that for all $t \in [0, T]$, $x \in \mathbb{R}^d$ it holds that $v_{t,x}: [0, t] \rightarrow \mathbb{R}$ is continuous, and the fundamental theorem of calculus hence prove that for all $t \in [0, T]$, $x \in \mathbb{R}^d$ it holds that

$$\begin{aligned} u(t, x) &= \mathbb{E}[u(t, x + \sqrt{2\rho}W_{t-t})] = v_{t,x}(t) = v_{t,x}(0) + \int_0^t (v_{t,x})'(s) ds \\ &= \mathbb{E}[u(0, x + \sqrt{2\rho}W_t)] + \int_0^t \mathbb{E}[f(s, x + \sqrt{2\rho}W_{t-s})] ds. \end{aligned} \quad (17.52)$$

Fubini's theorem and the fact that u and f are at most polynomially growing therefore establish (17.40). The proof of Proposition 17.3.6 is thus complete. \square

Corollary 17.3.7. Let $d \in \mathbb{N}$, $T, \rho \in (0, \infty)$, $\varrho = \sqrt{2\rho T}$, $a \in \mathbb{R}$, $b \in (a, \infty)$, let $\varphi: \mathbb{R}^d \rightarrow \mathbb{R}$ be a function, let $u \in C^{1,2}([0, T] \times \mathbb{R}^d, \mathbb{R})$ have at most polynomially growing partial derivatives, assume for all $t \in [0, T]$, $x \in \mathbb{R}^d$ that $u(0, x) = \varphi(x)$ and

$$\left(\frac{\partial u}{\partial t}\right)(t, x) = \rho (\Delta_x u)(t, x), \quad (17.53)$$

let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $\mathcal{W}: \Omega \rightarrow \mathbb{R}^d$ be a standard normal random variable. Then

- (i) it holds that $\varphi: \mathbb{R}^d \rightarrow \mathbb{R}$ is twice continuously differentiable with at most polynomially growing partial derivatives and
- (ii) it holds for all $x \in \mathbb{R}^d$ that $u(T, x) = \mathbb{E}[\varphi(\varrho \mathcal{W} + x)]$.

Proof of Corollary 17.3.7. Observe that the assumption that $u \in C^{1,2}([0, T] \times \mathbb{R}^d, \mathbb{R})$ has at most polynomially growing partial derivatives and the fact that for all $x \in \mathbb{R}^d$ it holds that $\varphi(x) = u(0, x)$ imply item (i). Furthermore, note that Proposition 17.3.6 proves item (ii). The proof of Corollary 17.3.7 is thus complete. \square

Definition 17.3.8 (Continuous convolutions). Let $d \in \mathbb{N}$ and let $f: \mathbb{R}^d \rightarrow \mathbb{R}$ and $g: \mathbb{R}^d \rightarrow \mathbb{R}$ be $\mathcal{B}(\mathbb{R}^d)/\mathcal{B}(\mathbb{R})$ -measurable. Then we denote by

$$f \circledast g: \left\{ x \in \mathbb{R}^d : \min \left\{ \int_{\mathbb{R}^d} \max \{0, f(x-y)g(y)\} dy, - \int_{\mathbb{R}^d} \min \{0, f(x-y)g(y)\} dy \right\} < \infty \right\} \rightarrow [-\infty, \infty] \quad (17.54)$$

the function which satisfies for all $x \in \mathbb{R}^d$ with

$$\min \left\{ \int_{\mathbb{R}^d} \max \{0, f(x-y)g(y)\} dy, - \int_{\mathbb{R}^d} \min \{0, f(x-y)g(y)\} dy \right\} < \infty \quad (17.55)$$

that

$$(f \circledast g)(x) = \int_{\mathbb{R}^d} f(x-y)g(y) dy. \quad (17.56)$$

Exercise 17.3.1. Let $d \in \mathbb{N}$, $T \in (0, \infty)$, for every $\sigma \in (0, \infty)$ let $\gamma_\sigma: \mathbb{R}^d \rightarrow \mathbb{R}$ satisfy for all $x \in \mathbb{R}^d$ that

$$\gamma_\sigma(x) = (2\pi\sigma^2)^{-\frac{d}{2}} \exp\left(\frac{-\|x\|_2^2}{2\sigma^2}\right), \quad (17.57)$$

and for every $\rho \in (0, \infty)$, $\varphi \in C^2(\mathbb{R}^d, \mathbb{R})$ with $\sup_{x \in \mathbb{R}^d} [\sum_{i,j=1}^d (|\varphi(x)| + |(\frac{\partial}{\partial x_i} \varphi)(x)| + |(\frac{\partial^2}{\partial x_i \partial x_j} \varphi)(x)|)] < \infty$ let $u_{\rho, \varphi}: [0, T] \times \mathbb{R}^d \rightarrow \mathbb{R}$ satisfy for all $t \in (0, T]$, $x \in \mathbb{R}^d$ that

$$u_{\rho, \varphi}(0, x) = \varphi(x) \quad \text{and} \quad u_{\rho, \varphi}(t, x) = (\varphi \circledast \gamma_{\sqrt{2t\rho}})(x) \quad (17.58)$$

(cf. Definitions 3.3.4 and 17.3.8). Prove or disprove the following statement: For all $\rho \in (0, \infty)$, $\varphi \in C^2(\mathbb{R}^d, \mathbb{R})$ with $\sup_{x \in \mathbb{R}^d} [\sum_{i,j=1}^d (|\varphi(x)| + |(\frac{\partial}{\partial x_i} \varphi)(x)| + |(\frac{\partial^2}{\partial x_i \partial x_j} \varphi)(x)|)] < \infty$ it holds for all $t \in (0, T)$, $x \in \mathbb{R}^d$ that $u_{\rho, \varphi} \in C^{1,2}([0, T] \times \mathbb{R}^d, \mathbb{R})$ and

$$\left(\frac{\partial u_{\rho, \varphi}}{\partial t} \right)(t, x) = \rho (\Delta_x u_{\rho, \varphi})(t, x). \quad (17.59)$$

Exercise 17.3.2. Prove or disprove the following statement: For every $x \in \mathbb{R}$ it holds that

$$e^{-x^2/2} = \frac{1}{\sqrt{2\pi}} \left[\int_{\mathbb{R}} e^{-t^2/2} e^{-ixt} dt \right]. \quad (17.60)$$

Exercise 17.3.3. Let $d \in \mathbb{N}$, $T \in (0, \infty)$, for every $\sigma \in (0, \infty)$ let $\gamma_{\sigma}: \mathbb{R}^d \rightarrow \mathbb{R}$ satisfy for all $x \in \mathbb{R}^d$ that

$$\gamma_{\sigma}(x) = (2\pi\sigma^2)^{-\frac{d}{2}} \exp\left(\frac{-\|x\|_2^2}{2\sigma^2}\right), \quad (17.61)$$

for every $\varphi \in C^2(\mathbb{R}^d, \mathbb{R})$ with $\sup_{x \in \mathbb{R}^d} [\sum_{i,j=1}^d (|\varphi(x)| + |(\frac{\partial}{\partial x_i} \varphi)(x)| + |(\frac{\partial^2}{\partial x_i \partial x_j} \varphi)(x)|)] < \infty$ let $u_{\varphi}: [0, T] \times \mathbb{R}^d \rightarrow \mathbb{R}$ satisfy for all $t \in (0, T]$, $x \in \mathbb{R}^d$ that

$$u_{\varphi}(0, x) = \varphi(x) \quad \text{and} \quad u_{\varphi}(t, x) = (\varphi \circledast \gamma_{\sqrt{2t}})(x), \quad (17.62)$$

and for every $i = (i_1, \dots, i_d) \in \mathbb{N}^d$ let $\psi_i: \mathbb{R}^d \rightarrow \mathbb{R}$ satisfy for all $x = (x_1, \dots, x_d) \in \mathbb{R}^d$ that

$$\psi_i(x) = 2^{\frac{d}{2}} \left[\prod_{k=1}^d \sin(i_k \pi x_k) \right] \quad (17.63)$$

(cf. Definitions 3.3.4 and 17.3.8). Prove or disprove the following statement: For all $i = (i_1, \dots, i_d) \in \mathbb{N}^d$, $t \in [0, T]$, $x \in \mathbb{R}^d$ it holds that

$$u_{\psi_i}(t, x) = \exp(-\pi^2 [\sum_{k=1}^d |i_k|^2] t) \psi_i(x). \quad (17.64)$$

Exercise 17.3.4. Let $d \in \mathbb{N}$, $T \in (0, \infty)$, for every $\sigma \in (0, \infty)$ let $\gamma_{\sigma}: \mathbb{R}^d \rightarrow \mathbb{R}$ satisfy for all $x \in \mathbb{R}^d$ that

$$\gamma_{\sigma}(x) = (2\pi\sigma^2)^{-\frac{d}{2}} \exp\left(\frac{-\|x\|_2^2}{2\sigma^2}\right), \quad (17.65)$$

and for every $i = (i_1, \dots, i_d) \in \mathbb{N}^d$ let $\psi_i: \mathbb{R}^d \rightarrow \mathbb{R}$ satisfy for all $x = (x_1, \dots, x_d) \in \mathbb{R}^d$ that

$$\psi_i(x) = 2^{\frac{d}{2}} \left[\prod_{k=1}^d \sin(i_k \pi x_k) \right] \quad (17.66)$$

(cf. Definition 3.3.4). Prove or disprove the following statement: For every $i = (i_1, \dots, i_d) \in \mathbb{N}^d$, $s \in [0, T]$, $y \in \mathbb{R}^d$ and every function $u \in C^{1,2}([0, T] \times \mathbb{R}^d, \mathbb{R})$ with at most polynomially growing partial derivatives which satisfies for all $t \in (0, T)$, $x \in \mathbb{R}^d$ that $u(0, x) = \psi_i(x)$ and

$$\left(\frac{\partial u}{\partial t}\right)(t, x) = (\Delta_x u)(t, x) \quad (17.67)$$

it holds that

$$u(s, y) = \exp(-\pi^2 [\sum_{k=1}^d |i_k|^2] s) \psi_i(y). \quad (17.68)$$

17.4 Reformulation of PDE problems as stochastic optimization problems

The proof of the next result, Proposition 17.4.1 below, is based on an application of Proposition 17.2.1 and Proposition 17.3.6. A more general result than Proposition 17.4.1 with a detailed proof can, for instance, be found in Beck et al. [18, Proposition 2.7].

Proposition 17.4.1. *Let $d \in \mathbb{N}$, $T, \rho \in (0, \infty)$, $\varrho = \sqrt{2\rho T}$, $a \in \mathbb{R}$, $b \in (a, \infty)$, let $\varphi: \mathbb{R}^d \rightarrow \mathbb{R}$ be a function, let $u \in C^{1,2}([0, T] \times \mathbb{R}^d, \mathbb{R})$ have at most polynomially growing partial derivatives, assume for all $t \in [0, T]$, $x \in \mathbb{R}^d$ that $u(0, x) = \varphi(x)$ and*

$$\left(\frac{\partial u}{\partial t}\right)(t, x) = \rho (\Delta_x u)(t, x), \quad (17.69)$$

let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\mathcal{W}: \Omega \rightarrow \mathbb{R}^d$ be a standard normal random variable, let $\mathcal{X}: \Omega \rightarrow [a, b]^d$ be a continuously uniformly distributed random variable, and assume that \mathcal{W} and \mathcal{X} are independent. Then

(i) *it holds that $\varphi: \mathbb{R}^d \rightarrow \mathbb{R}$ is twice continuously differentiable with at most polynomially growing partial derivatives,*

(ii) *there exists a unique continuous function $U: [a, b]^d \rightarrow \mathbb{R}$ such that*

$$\mathbb{E}[|\varphi(\varrho \mathcal{W} + \mathcal{X}) - U(\mathcal{X})|^2] = \inf_{v \in C([a, b]^d, \mathbb{R})} \mathbb{E}[|\varphi(\varrho \mathcal{W} + \mathcal{X}) - v(\mathcal{X})|^2], \quad (17.70)$$

and

(iii) *it holds for every $x \in [a, b]^d$ that $U(x) = u(T, x)$.*

Proof of Proposition 17.4.1. First, observe that (17.69), the assumption that \mathcal{W} is a standard normal random variable, and Corollary 17.3.7 demonstrate that for all $x \in \mathbb{R}^d$ it holds that $\varphi: \mathbb{R}^d \rightarrow \mathbb{R}$ is twice continuously differentiable with at most polynomially growing partial derivatives and

$$u(T, x) = \mathbb{E}[u(0, \varrho \mathcal{W} + x)] = \mathbb{E}[\varphi(\varrho \mathcal{W} + x)]. \quad (17.71)$$

Furthermore, note that the assumption that \mathcal{W} is a standard normal random variable, the fact that φ is continuous, and the fact that φ has at most polynomially growing partial derivatives and is continuous ensure that

- (I) it holds that $[a, b]^d \times \Omega \ni (x, \omega) \mapsto \varphi(\varrho \mathcal{W}(\omega) + x) \in \mathbb{R}$ is $(\mathcal{B}([a, b]^d) \otimes \mathcal{F})/\mathcal{B}(\mathbb{R})$ -measurable and
- (II) it holds for all $x \in [a, b]^d$ that $\mathbb{E}[|\varphi(\varrho \mathcal{W} + x)|^2] < \infty$.

Proposition 17.2.1 and (17.71) hence ensure that

- (A) there exists a unique continuous function $U: [a, b]^d \rightarrow \mathbb{R}$ which satisfies that

$$\int_{[a, b]^d} \mathbb{E}[|\varphi(\varrho \mathcal{W} + x) - U(x)|^2] dx = \inf_{v \in C([a, b]^d, \mathbb{R})} \left(\int_{[a, b]^d} \mathbb{E}[|\varphi(\varrho \mathcal{W} + x) - v(x)|^2] dx \right) \quad (17.72)$$

and

- (B) it holds for all $x \in [a, b]^d$ that $U(x) = u(T, x)$.

Moreover, observe that the assumption that \mathcal{W} and \mathcal{X} are independent, item (I), and the assumption that \mathcal{X} is continuously uniformly distributed on $[a, b]^d$ show that for all $v \in C([a, b]^d, \mathbb{R})$ it holds that

$$\mathbb{E}[|\varphi(\varrho \mathcal{W} + \mathcal{X}) - v(\mathcal{X})|^2] = \frac{1}{(b-a)^d} \int_{[a, b]^d} \mathbb{E}[|\varphi(\varrho \mathcal{W} + x) - v(x)|^2] dx. \quad (17.73)$$

Combining this with item (A) establishes item (ii). Note that items (A) and (B) and (17.73) imply item (iii). The proof of Proposition 17.4.1 is thus complete. \square

While Proposition 17.4.1 above recasts the solutions of the PDE in (17.69) at a particular point in time as the solutions of a stochastic optimization problem, we can also derive from this a corollary which shows that the solutions of the PDE over an entire timespan are similarly the solutions of a stochastic optimization problem.

Corollary 17.4.2. Let $d \in \mathbb{N}$, $T, \rho \in (0, \infty)$, $\varrho = \sqrt{2\rho}$, $a \in \mathbb{R}$, $b \in (a, \infty)$, let $\varphi: \mathbb{R}^d \rightarrow \mathbb{R}$ be a function, let $u \in C^{1,2}([0, T] \times \mathbb{R}^d, \mathbb{R})$ be a function with at most polynomially growing partial derivatives which satisfies for all $t \in [0, T]$, $x \in \mathbb{R}^d$ that $u(0, x) = \varphi(x)$ and

$$\left(\frac{\partial u}{\partial t} \right)(t, x) = \rho (\Delta_x u)(t, x), \quad (17.74)$$

let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\mathcal{W}: \Omega \rightarrow \mathbb{R}^d$ be a standard normal random variable, let $\tau: \Omega \rightarrow [0, T]$ be a continuously uniformly distributed random variable, let $\mathcal{X}: \Omega \rightarrow [a, b]^d$ be a continuously uniformly distributed random variable, and assume that \mathcal{W} , τ , and \mathcal{X} are independent. Then

(i) there exists a unique $U \in C([0, T] \times [a, b]^d, \mathbb{R})$ which satisfies that

$$\mathbb{E}[|\varphi(\varrho\sqrt{\tau}\mathcal{W} + \mathcal{X}) - U(\tau, \mathcal{X})|^2] = \inf_{v \in C([0, T] \times [a, b]^d, \mathbb{R})} \mathbb{E}[|\varphi(\varrho\sqrt{\tau}\mathcal{W} + \mathcal{X}) - v(\tau, \mathcal{X})|^2] \quad (17.75)$$

and

(ii) it holds for all $t \in [0, T]$, $x \in [a, b]^d$ that $U(t, x) = u(t, x)$.

Proof of Corollary 17.4.2. Throughout this proof, let $F: C([0, T] \times [a, b]^d, \mathbb{R}) \rightarrow [0, \infty]$ satisfy for all $v \in C([0, T] \times [a, b]^d, \mathbb{R})$ that

$$F(v) = \mathbb{E}[|\varphi(\varrho\sqrt{\tau}\mathcal{W} + \mathcal{X}) - v(\tau, \mathcal{X})|^2]. \quad (17.76)$$

Observe that Proposition 17.4.1 proves that for all $v \in C([0, T] \times [a, b]^d, \mathbb{R})$, $s \in [0, T]$ it holds that

$$\mathbb{E}[|\varphi(\varrho\sqrt{s}\mathcal{W} + \mathcal{X}) - v(s, \mathcal{X})|^2] \geq \mathbb{E}[|\varphi(\varrho\sqrt{s}\mathcal{W} + \mathcal{X}) - u(s, \mathcal{X})|^2]. \quad (17.77)$$

Furthermore, note that the assumption that \mathcal{W} , τ , and \mathcal{X} are independent, the assumption that $\tau: \Omega \rightarrow [0, T]$ is continuously uniformly distributed, and Fubini's theorem demonstrate that for all $v \in C([0, T] \times [a, b]^d, \mathbb{R})$ it holds that

$$F(v) = \mathbb{E}[|\varphi(\varrho\sqrt{\tau}\mathcal{W} + \mathcal{X}) - v(\tau, \mathcal{X})|^2] = \int_{[0, T]} \mathbb{E}[|\varphi(\varrho\sqrt{s}\mathcal{W} + \mathcal{X}) - v(s, \mathcal{X})|^2] ds. \quad (17.78)$$

This and (17.77) ensure that for all $v \in C([0, T] \times [a, b]^d, \mathbb{R})$ it holds that

$$F(v) \geq \int_{[0, T]} \mathbb{E}[|\varphi(\varrho\sqrt{s}\mathcal{W} + \mathcal{X}) - u(s, \mathcal{X})|] ds. \quad (17.79)$$

Combining this with (17.78) shows that for all $v \in C([0, T] \times [a, b]^d, \mathbb{R})$ it holds that $F(v) \geq F(u)$. Hence, we obtain that

$$F(u) = \inf_{v \in C([0, T] \times [a, b]^d, \mathbb{R})} F(v). \quad (17.80)$$

This and (17.78) establish that for all $U \in C([0, T] \times [a, b]^d, \mathbb{R})$ with

$$F(U) = \inf_{v \in C([0, T] \times [a, b]^d, \mathbb{R})} F(v) \quad (17.81)$$

it holds that

$$\int_{[0, T]} \mathbb{E}[|\varphi(\varrho\sqrt{s}\mathcal{W} + \mathcal{X}) - U(s, \mathcal{X})|] ds = \int_{[0, T]} \mathbb{E}[|\varphi(\varrho\sqrt{s}\mathcal{W} + \mathcal{X}) - u(s, \mathcal{X})|] ds. \quad (17.82)$$

Combining this with (17.77) implies that for all $U \in C([0, T] \times [a, b]^d, \mathbb{R})$ with $F(U) = \inf_{v \in C([0, T] \times [a, b]^d, \mathbb{R})} F(v)$ there exists $A \subseteq [0, T]$ with $\int_A 1 dx = T$ such that for all $s \in A$ it holds that

$$\mathbb{E}[|\varphi(\varrho\sqrt{s}\mathcal{W} + \mathcal{X}) - U(s, \mathcal{X})|^2] = \mathbb{E}[|\varphi(\varrho\sqrt{s}\mathcal{W} + \mathcal{X}) - u(s, \mathcal{X})|^2]. \quad (17.83)$$

Proposition 17.4.1 therefore demonstrates that for all $U \in C([0, T] \times [a, b]^d, \mathbb{R})$ with $F(U) = \inf_{v \in C([0, T] \times [a, b]^d, \mathbb{R})} F(v)$ there exists $A \subseteq [0, T]$ with $\int_A 1 dx = T$ such that for all $s \in A$ it holds that $U(s) = u(s)$. The fact that $u \in C([0, T] \times [a, b]^d, \mathbb{R})$ hence proves that for all $U \in C([0, T] \times [a, b]^d, \mathbb{R})$ with $F(U) = \inf_{v \in C([0, T] \times [a, b]^d, \mathbb{R})} F(v)$ it holds that $U = u$. Combining this with (17.80) establishes items (i) and (ii). The proof of Corollary 17.4.2 is thus complete. \square

17.5 Derivation of DKMs

In this section we present in the special case of the heat equation a rough derivation of the **DKMs** introduced in Beck et al. [19]. This derivation will proceed along the analogous steps as the derivation of **PINNs** and **DGMs** in Section 16.2. Firstly, we will employ Proposition 17.4.1 to reformulate the **PDE** problem under consideration as an infinite-dimensional stochastic optimization problem, secondly, we will employ **ANNs** to reduce the infinite-dimensional stochastic optimization problem to a finite-dimensional stochastic optimization problem, and thirdly, we will aim to approximately solve this finite-dimensional stochastic optimization problem by means of **SGD**-type optimization methods. We start by introducing the setting of the problem. Let $d \in \mathbb{N}$, $T, \rho \in (0, \infty)$, $a \in \mathbb{R}$, $b \in (a, \infty)$, let $\varphi: \mathbb{R}^d \rightarrow \mathbb{R}$ be a function, let $u \in C^{1,2}([0, T] \times \mathbb{R}^d, \mathbb{R})$ have at most polynomially growing partial derivatives, and assume for all $t \in [0, T]$, $x \in \mathbb{R}^d$ that $u(0, x) = \varphi(x)$ and

$$(\frac{\partial u}{\partial t})(t, x) = \rho (\Delta_x u)(t, x). \quad (17.84)$$

In the framework described in the previous sentence, we think of u as the unknown **PDE** solution. The objective of this derivation is to develop deep learning methods which aim to approximate the unknown **PDE** solution $u(T, \cdot)|_{[a, b]^d}: [a, b]^d \rightarrow \mathbb{R}$ at time T restricted on $[a, b]^d$.

In the first step, we employ Proposition 17.4.1 to recast the unknown target function $u(T, \cdot)|_{[a, b]^d}: [a, b]^d \rightarrow \mathbb{R}$ as the solution of an optimization problem. For this let $\varrho = \sqrt{2\rho T}$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\mathcal{W}: \Omega \rightarrow \mathbb{R}^d$ be a standard normally distributed random variable, let $\mathcal{X}: \Omega \rightarrow [a, b]^d$ be a continuously uniformly distributed random variable, assume that \mathcal{W} and \mathcal{X} are independent, and let $\mathfrak{L}: C([a, b]^d, \mathbb{R}) \rightarrow [0, \infty]$ satisfy for all $v \in C([a, b]^d, \mathbb{R})$ that

$$\mathfrak{L}(v) = \mathbb{E}[|\varphi(\varrho\mathcal{W} + \mathcal{X}) - v(\mathcal{X})|^2]. \quad (17.85)$$

Proposition 17.4.1 then ensures that the unknown target function $u(T, \cdot)|_{[a,b]^d} : [a, b]^d \rightarrow \mathbb{R}$ is the unique global minimizer of the function $\mathfrak{L} : C([a, b]^d, \mathbb{R}) \rightarrow [0, \infty]$. Minimizing \mathfrak{L} is, however, not yet amenable to numerical computations.

In the second step, we therefore reduce this infinite-dimensional stochastic optimization problem to a finite-dimensional stochastic optimization problem involving ANNs. Specifically, let $a : \mathbb{R} \rightarrow \mathbb{R}$ be differentiable, let $h \in \mathbb{N}$, $l_1, l_2, \dots, l_h, \mathfrak{d} \in \mathbb{N}$ satisfy $\mathfrak{d} = l_1(d+1) + [\sum_{k=2}^h l_k(l_{k-1}+1)] + l_h + 1$, and let $\mathcal{L} : \mathbb{R}^{\mathfrak{d}} \rightarrow [0, \infty)$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$ that

$$\begin{aligned}\mathcal{L}(\theta) &= \mathfrak{L}\left(\left(\mathcal{N}_{\mathfrak{M}_{a,l_1}, \mathfrak{M}_{a,l_2}, \dots, \mathfrak{M}_{a,l_h}, \text{id}_{\mathbb{R}}}^{\theta,d}\right)|_{[a,b]^d}\right) \\ &= \mathbb{E}\left[\left|\varphi(\varrho w + \mathcal{X}) - \mathcal{N}_{\mathfrak{M}_{a,l_1}, \mathfrak{M}_{a,l_2}, \dots, \mathfrak{M}_{a,l_h}, \text{id}_{\mathbb{R}}}^{\theta,d}(\mathcal{X})\right|^2\right]\end{aligned}\quad (17.86)$$

(cf. Definitions 1.1.3 and 1.2.1). We can now compute an approximate minimizer of the function \mathfrak{L} by computing an approximate minimizer $\vartheta \in \mathbb{R}^{\mathfrak{d}}$ of the function \mathcal{L} and employing the realization $(\mathcal{N}_{\mathfrak{M}_{a,l_1}, \mathfrak{M}_{a,l_2}, \dots, \mathfrak{M}_{a,l_h}, \text{id}_{\mathbb{R}}}^{\theta,d})|_{[a,b]^d} \in C([a, b]^d, \mathbb{R})$ of the ANN associated to this approximate minimizer restricted on $[a, b]^d$ as an approximate minimizer of \mathfrak{L} .

In the third step, we use SGD-type methods to compute such an approximate minimizer of \mathcal{L} . We now sketch this in the case of the plain-vanilla SGD optimization method (cf. Definition 7.2.1). Let $\xi \in \mathbb{R}^{\mathfrak{d}}$, $J \in \mathbb{N}$, $(\gamma_n)_{n \in \mathbb{N}} \subseteq [0, \infty)$, for every $n \in \mathbb{N}$, $j \in \{1, 2, \dots, J\}$ let $\mathfrak{W}_{n,j} : \Omega \rightarrow \mathbb{R}^d$ be a standard normally distributed random variable and let $\mathfrak{X}_{n,j} : \Omega \rightarrow [a, b]^d$ be a continuously uniformly distributed random variable, let $\ell : \mathbb{R}^{\mathfrak{d}} \times [0, T] \times \mathbb{R}^d \rightarrow \mathbb{R}$ satisfy for all $\theta \in \mathbb{R}^{\mathfrak{d}}$, $w \in \mathbb{R}^d$, $x \in [a, b]^d$ that

$$\ell(\theta, w, x) = \left|\varphi(\varrho w + x) - \mathcal{N}_{\mathfrak{M}_{a,l_1}, \mathfrak{M}_{a,l_2}, \dots, \mathfrak{M}_{a,l_h}, \text{id}_{\mathbb{R}}}^{\theta,d}(x)\right|^2,\quad (17.87)$$

and let $\Theta = (\Theta_n)_{n \in \mathbb{N}_0} : \mathbb{N}_0 \times \Omega \rightarrow \mathbb{R}^{\mathfrak{d}}$ satisfy for all $n \in \mathbb{N}$ that

$$\Theta_0 = \xi \quad \text{and} \quad \Theta_n = \Theta_{n-1} - \gamma_n \left[\frac{1}{J} \sum_{j=1}^J (\nabla_{\theta} \ell)(\Theta_{n-1}, \mathfrak{W}_{n,j}, \mathfrak{X}_{n,j}) \right].\quad (17.88)$$

Finally, the idea of DKMs is to consider for large enough $n \in \mathbb{N}$ the realization function $\mathcal{N}_{\mathfrak{M}_{a,l_1}, \mathfrak{M}_{a,l_2}, \dots, \mathfrak{M}_{a,l_h}, \text{id}_{\mathbb{R}}}^{\Theta_n, d}$ as an approximation

$$(\mathcal{N}_{\mathfrak{M}_{a,l_1}, \mathfrak{M}_{a,l_2}, \dots, \mathfrak{M}_{a,l_h}, \text{id}_{\mathbb{R}}}^{\Theta_n, d})|_{[a,b]^d} \approx u(T, \cdot)|_{[a,b]^d}\quad (17.89)$$

of the unknown solution u of the PDE in (17.84) at time T restricted to $[a, b]^d$.

An implementation in the case of a two-dimensional heat equation of the DKMs derived above that employs the more sophisticated Adam SGD optimization method instead of the SGD optimization method can be found in the next section.

17.6 Implementation of DKMs

In Source code 17.2 below we present a simple implementation of a **DKM**, as explained in Section 17.5 above, for finding an approximation of a solution $u \in C^{1,2}([0, 2] \times \mathbb{R}^2)$ of the two-dimensional heat equation

$$\left(\frac{\partial u}{\partial t}\right)(t, x) = (\Delta_x u)(t, x) \quad (17.90)$$

with $u(0, x) = \cos(x_1) + \cos(x_2)$ for $t \in [0, 2]$, $x = (x_1, x_2) \in \mathbb{R}^2$. This implementation trains a fully connected feed-forward **ANN** with 2 hidden layers (with 50 neurons on each hidden layer) and using the **ReLU** activation function (cf. Section 1.2.3). The training uses batches of size 256 with each batch consisting of 256 randomly chosen realizations of the random variable $(\mathcal{T}, \mathcal{X})$, where \mathcal{T} is continuously uniformly distributed random variable on $[0, 2]$ and where \mathcal{X} is a continuously uniformly distributed random variable on $[-5, 5]^2$. The training is performed using the **Adam SGD** optimization method (cf. Section 7.9). A plot of the resulting approximation of the solution u after 3000 training steps is shown in Figure 16.1.

```

1 import torch
2 import matplotlib.pyplot as plt
3
4 # Use the GPU if available
5 dev = torch.device("cuda" if torch.cuda.is_available() else "cpu")
6
7 # Computes an approximation of E[|phi(sqrt(2*rho*T) W + xi) -
8 # N(xi)|^2] with W a standard normal random variable using the rows
9 # of x as # independent realizations of the random variable xi
10 def loss(N, rho, phi, t, x):
11     W = torch.randn_like(x).to(dev)
12     return (phi(torch.sqrt(2 * rho * t) * W + x) -
13             N(torch.cat((t,x),1))).square().mean()
14
15 d = 2           # the input dimension
16 a, b = -5.0, 5.0 # the domain will be [a,b]^d
17 T = 2.0         # the time horizon
18 rho = 1.0       # the diffusivity
19
20 # Define the initial value
21 def phi(x):
22     return x.cos().sum(axis=1, keepdim=True)
23
24 # Define a neural network with two hidden layers with 50 neurons
25 # each using ReLU activations
26 N = torch.nn.Sequential(
27     torch.nn.Linear(d+1, 50), torch.nn.ReLU(),
28     torch.nn.Linear(50, 50), torch.nn.ReLU(),
29     torch.nn.Linear(50, 1)

```

```

30 | ).to(dev)
31 |
32 # Configure the training parameters and optimization algorithm
33 steps = 3000
34 batch_size = 256
35 optimizer = torch.optim.Adam(N.parameters())
36
37 # Train the network
38 for step in range(steps):
39     # Generate uniformly distributed samples from [a,b]^d
40     x = (torch.rand(batch_size, d) * (b-a) + a).to(dev)
41     t = T * torch.rand(batch_size, 1).to(dev)
42
43     optimizer.zero_grad()
44     # Compute the loss
45     L = loss(N, rho, phi, t, x)
46     # Compute the gradients
47     L.backward()
48     # Apply changes to weights and biases of N
49     optimizer.step()
50
51 # Plot the result at M+1 timesteps
52 M = 5
53 mesh = 128
54
55 def toNumpy(t):
56     return t.detach().cpu().numpy().reshape((mesh,mesh))
57
58 fig, axs = plt.subplots(2,3,subplot_kw=dict(projection='3d'))
59 fig.set_size_inches(16, 10)
60 fig.set_dpi(300)
61
62 for i in range(M+1):
63     x = torch.linspace(a, b, mesh)
64     y = torch.linspace(a, b, mesh)
65     x, y = torch.meshgrid(x, y, indexing='xy')
66     x = x.reshape((mesh*mesh,1)).to(dev)
67     y = y.reshape((mesh*mesh,1)).to(dev)
68     z = N(torch.cat((i*T/M*torch.ones(128*128,1).to(dev), x, y),
69                     1))
70
71     axs[i//3,i%3].set_title(f"t = {i * T / M}")
72     axs[i//3,i%3].set_zlim(-2,2)
73     axs[i//3,i%3].plot_surface(toNumpy(x), toNumpy(y), toNumpy(z),
74                               cmap='viridis')
75
76 fig.savefig(f"../plots/kolmogorov.pdf", bbox_inches='tight')

```

Source code 17.2 ([code/kolmogorov.py](#)): A simple implementation in PYTORCH of the deep Kolmogorov method based on Corollary 17.4.2, computing an approximation of the function $u \in C^{1,2}([0, 2] \times \mathbb{R}^2, \mathbb{R})$ which satisfies for all $t \in [0, 2]$, $x = (x_1, x_2) \in \mathbb{R}^2$ that $(\frac{\partial u}{\partial t})(t, x) = (\Delta_x u)(t, x)$ and $u(0, x) = \cos(x_1) + \cos(x_2)$.

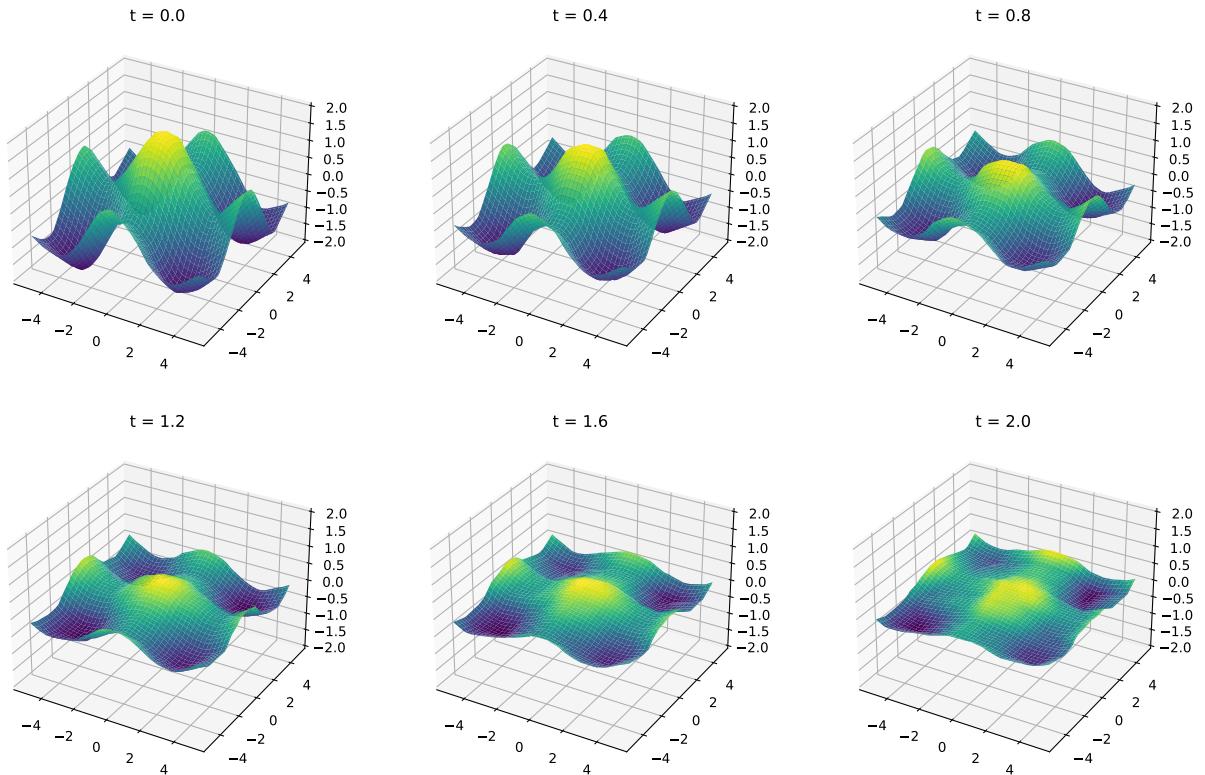


Figure 17.2 ([plots/kolmogorov.pdf](#)): Plots for the functions $[-5, 5]^2 \ni x \mapsto U(t, x) \in \mathbb{R}$, where $t \in \{0, 0.4, 0.8, 1.2, 1.6, 2\}$ and where $U \in C([0, 2] \times \mathbb{R}^2, \mathbb{R})$ is an approximation for the function $u \in C^{1,2}([0, 2] \times \mathbb{R}^2, \mathbb{R})$ satisfies for all $t \in [0, 2]$, $x = (x_1, x_2) \in \mathbb{R}^2$ that $(\frac{\partial u}{\partial t})(t, x) = (\Delta_x u)(t, x)$ and $u(0, x) = \cos(x_1) + \cos(x_2)$ computed by means of Source code 17.2.

Chapter 18

Further deep learning methods for PDEs

Besides PINNs, DGMs, and DKMs reviewed in Chapters 16 and 17 above there are also a large number of other works which propose and study deep learning based approximation methods for various classes of PDEs. In the following we mention a selection of such methods from the literature roughly grouped into three classes. Specifically, we consider deep learning methods for PDEs which employ *strong formulations* of PDEs to set up learning problems in Section 18.1, we consider deep learning methods for PDEs which employ *weak or variational formulations* of PDEs to set up learning problems in Section 18.2, and we consider deep learning methods for PDEs which employ intrinsic *stochastic representations* of PDEs to set up learning problems in Section 18.3. Finally, in Section 18.4 we also point to several theoretical results and error analyses for deep learning methods for PDEs in the literature.

Our selection of references for methods as well as theoretical results is by no means complete. For more complete reviews of the literature on deep learning methods for PDEs and corresponding theoretical results we refer, for example, to the overview articles [24, 58, 91, 126, 152, 251, 376].

18.1 Deep learning methods based on strong formulations of PDEs

There are a number of deep learning based methods for PDEs in the literature that employ residuals of strong formulations of PDEs to set up learning problems (cf., for instance, Theorem 16.1.1 and (16.16) for the residual of the strong formulation in the case of semilinear heat PDEs). Basic methods in this category include the PINNs (see Raissi et al. [368]) and DGMs (see Sirignano & Spiliopoulos [400]) reviewed in Chapter 16 above, the approach proposed in Berg & Nyström [34], the *theory-guided neural networks* (TGNNs) proposed in Wang et al. [426], and the two early methods proposed in [112, 274]. There are also many refinements and adaptions of these basic methods in the literature including

- the *conservative PINNs* (cPINNs) methodology for conservation laws in Jagtap et al. [230] which relies on multiple ANNs representing a PDE solution on respective sub-domains,
- the *extended PINNs* (XPINNs) methodology in Jagtap & Karniadakis [93] which generalizes the domain decomposition idea of Jagtap et al. [230] to other types of PDEs,
- the *Navier-Stokes flow nets* (NSFnets) methodology in Jin et al. [244] which explores the use of PINNs for the incompressible Navier-Stokes PDEs,
- the *Bayesian PINNs* methodology in Yang et al. [442] which combines PINNs with *Bayesian neural networks* (BNNs) from Bayesian learning (cf., for example, [306, 319]),
- the *parareal PINNs* (PPINNs) methodology for time-dependent PDEs with long time horizons in Meng et al. [314] which combines the PINNs methodology with ideas from parareal algorithms (cf., for instance, [44, 309]) in order to split up long-time problems into many independent short-time problems,
- the *SelectNets* methodology in Gu et al. [191] which extends the PINNs methodology by employing a second ANN to adaptively *select* during the training process the points at which the residual of the PDE is considered, and
- the *fractional PINNs* (fPINNs) methodology in Pang et al. [344] which extends the PINNs methodology to PDEs with fractional derivatives such as space-time fractional advection-diffusion equations.

We also refer to the article Lu et al. [304] which introduces an elegant PYTHON library for PINNs called *DeepXDE* and also provides a good introduction to PINNs.

18.2 Deep learning methods based on weak formulations of PDEs

Another group of deep learning methods for PDEs relies on weak or variational formulations of PDEs to set up learning problems. Such methods include

- the *variational PINNs* (VPINNs) methodology in Kharazmi et al. [255, 256] which use the residuals of weak formulations of PDEs for a fixed set of test functions to set up a learning problem,
- the *VarNets* methodology in Khodayi-Mehr & Zavlanos [257] which employs a similar methodology than VPINNs but also consider parametric PDEs,

- the *weak form TGNN* methodology in Xu et al. [441] which further extend the VPINNs methodology by (amongst other adaptions) considering test functions in the weak formulation of PDEs tailored to the considered problem,
- the *deep fourier residual method* in Taylor et al. [414] which is based on minimizing the dual norm of the weak-form residual operator of PDEs by employing Fourier-type representations of this dual norm which can efficiently be approximated using the *discrete sine transform (DST)* and *discrete cosine transform (DCT)*,
- the *weak adversarial networks (WANs)* methodology in Zang et al. [449] (cf. also Bao et al. [13]) which is based on approximating both the solution of the PDE and the test function in the weak formulation of the PDE by ANNs and on using an adversarial approach (cf., for example, Goodfellow et al. [172]) to train both networks to minimize and maximize, respectively, the weak-form residual of the PDE,
- the *Friedrichs learning* methodology in Chen et al. [68] which is similar to the WAN methodology but uses a different minimax formulation for the weak solution related to Friedrichs' theory on symmetric system of PDEs (see Friedrichs [145]),
- the *deep Ritz* method for elliptic PDEs in E & Yu [130] which employs variational minimization problems associated to PDEs to set up a learning problem,
- the *deep Nitsche* method in Liao & Ming [290] which refines the deep Ritz method using Nitsche's method (see Nitsche [333]) to enforce boundary conditions, and
- the *deep domain decomposition method (D3M)* in Li et al. [283] which refines the deep Ritz method using domain decompositions.

We also refer to the *multi-scale deep neural networks (MscaleDNNs)* in Cai et al. [60, 296] for a refined ANN architecture which can be employed in both the strong-form-based PINNs methodology and the variational-form-based deep Ritz methodology.

18.3 Deep learning methods based on stochastic representations of PDEs

A further class of deep learning based methods for PDEs are based on intrinsic links between PDEs and probability theory such as Feynman–Kac-type formulas; cf., for instance, [338, Section 8.2], [248, Section 4.4] for linear Feynman–Kac formulas based on (forward) *stochastic differential equations (SDEs)* and cf., for example, [76, 345–347] for nonlinear Feynman–Kac-type formulas based on *backward stochastic differential equations (BSDEs)*. The DKMs for linear PDEs (see Beck et al. [19]) reviewed in Chapter 17 are one type of such methods based on linear Feynman–Kac formulas. Other methods based on stochastic representations of PDEs include

- the *deep BSDE* methodology in E et al. [125, 196] which suggests to approximate solutions of *semilinear parabolic PDEs* by approximately solving the **BSDE** associated to the considered **PDE** through the nonlinear Feyman-Kac formula (see Pardoux & Peng [345, 346]) using a new deep learning methodology based on
 - reinterpreting the **BSDE** as a stochastic control problem in which the objective is to minimize the distance between the terminal value of the controlled process and the terminal value of the **BSDE**,
 - discretizing the control problem in time, and
 - approximately solving the discrete time control problem by approximating the policy functions at each time steps by means of **ANNs** as proposed in E & Han [195],
- the generalization of the deep BSDE methodology in Han & Long [197] for semilinear and quasilinear parabolic **PDEs** based on *forward backward stochastic differential equations (FBSDEs)*
- the refinements of the deep BSDE methodology in [66, 146, 206, 337, 367] which explore different nontrivial variations and extensions of the original deep BSDE methodology including different **ANN** architectures, initializations, and loss functions,
- the extension of the deep BSDE methodology to fully nonlinear parabolic **PDEs** in Beck et al. [20] which is based on a nonlinear Feyman-Kac formula involving second order **BSDEs** (see Cheridito et al. [76]),
- the *deep backward schemes* for semilinear parabolic **PDEs** in Huré et al. [218] which also rely on **BSDEs** but set up many separate learning problems which are solved inductively backwards in time instead of one single optimization problem,
- the *deep backward schemes* in Pham et al. [356] which extend the methodology in Huré et al. [218] to fully nonlinear parabolic **PDEs**,
- the *deep splitting* method for semilinear parabolic **PDEs** in Beck et al. [17] which iteratively solve for small time increments linear approximations of the semilinear parabolic **PDEs** using **DKMs**,
- the extensions of the deep backwards schemes to *partial integro-differential equations (PIDEs)* in [64, 161],
- the extensions of the deep splitting method to **PIDEs** in [52, 144],
- the methods in Nguwi et al. [328, 329, 331] which are based on representations of **PDE** solutions involving *branching-type processes* (cf., for instance, also [205, 207],

[330] and the references therein for nonlinear Feynman–Kac-type formulas based on such branching-type processes), and

- the methodology for elliptic PDEs in Kremsner et al. [270] which relies on suitable representations of elliptic PDEs involving BSDEs with random terminal times.

18.4 Error analyses for deep learning methods for PDEs

Until today there is not yet any complete error analysis for a GD/SGD based ANN training approximation scheme for PDEs in the literature (cf. also Remark 9.15.5 above). However, there are now several partial error analysis results for deep learning methods for PDEs in the literature (cf., for example, [26, 143, 153, 165, 197, 317, 318] and the references therein).

In particular, there are nowadays a number of results which rigorously establish that ANNs have the fundamental capacity to approximate solutions of certain classes of PDEs without the *curse of dimensionality* (COD) (cf., for instance, [27] and [334, Chapter 1]) in the sense that the number of parameters of the approximating ANN grows at most polynomially in both the reciprocal $1/\varepsilon$ of the prescribed approximation accuracy $\varepsilon \in (0, \infty)$ and the PDE dimension $d \in \mathbb{N}$. We refer, for example, to [10, 35, 38, 134, 168, 169, 184, 186, 188, 216, 241, 273, 374] for such and related ANN approximation results for solutions of linear PDEs and we refer, for instance, to [3, 85, 185, 220, 323] for such and related ANN approximation results for solutions of nonlinear PDEs.

The proofs in the above named ANN approximation results are usually based, first, on considering a suitable algorithm which approximates the considered PDEs without the COD and, thereafter, on constructing ANNs which approximate the considered approximation algorithm. In the context of linear PDEs the employed approximation algorithms are typically standard Monte Carlo methods (cf., for example, [162, 175, 264] and the references therein) and in the context of nonlinear PDEs the employed approximation algorithms are typically nonlinear Monte Carlo methods of the multilevel-Picard-type (cf., for instance, [21, 22, 157, 219, 221–223, 225, 324, 325] and the references therein).

In the literature the above named polynomial growth property in both the reciprocal $1/\varepsilon$ of the prescribed approximation accuracy $\varepsilon \in (0, \infty)$ and the PDE dimension $d \in \mathbb{N}$ is also referred to as *polynomial tractability* (cf., for example, [334, Definition 4.44], [335], and [336]).

Index of abbreviations

ANN (artificial neural network)	3
Adagrad (adaptive gradient)	334
Adam (adaptive moment estimation)	342
AdamW (Adam with decoupled weight decay)	348
BERT (Bidirectional Encoder Representations from Transformers)	80
BN (batch normalization)	4
BNN (Bayesian neural network)	672
BSDE (backward stochastic differential equation)	673
CNN (convolutional ANN)	3
COD (curse of dimensionality)	675
CV (computer vision)	64
D3M (deep domain decomposition method)	673
DCT (discrete cosine transform)	673
DGM (deep Galerkin method)	4
DKM (deep Kolmogorov method)	4
DST (discrete sine transform)	673
ELU (exponential linear unit)	50
FBSDE (forward backward stochastic differential equation)	674
FNO (Fourier neural operator)	81
GD (gradient descent)	3
GELU (Gaussian error linear unit)	23
GF (gradient flow)	3
GNN (graph neural network)	81
GPT (generative pre-trained transformer)	80
KL (Kurdyka–Łojasiewicz)	4
LLM (large language model)	80
LSTM (long short-term memory)	76
MscaleDNN (multi-scale deep neural network)	673
Muon (momentum orthogonalized by Newton-Schulz)	352
NLP (natural language processing)	64
NSFnet (Navier-Stokes flow net)	672

Nadam (Nesterov-accelerated adaptive moment estimation)	344
ODE (ordinary differential equation)	3
PDE (partial differential equation)	4
PIDE (partial integro-differential equation)	674
PINN (physics-informed neural network)	4
PPINN (parareal PINN)	672
RMSprop (root mean square propagation)	336
RNN (recurrent ANN)	3
ReLU (rectified linear unit)	23
RePU (rectified power unit)	52
ResNet (residual ANN)	3
SDE (stochastic differential equation)	673
SGD (stochastic gradient descent)	3
SiLU (sigmoid linear unit)	44
TGNN (theory-guided neural network)	671
VPINN (variational PINN)	672
WAN (weak adversarial network)	673
XPINN (extended PINN)	672
cPINN (conservative PINN)	672
deepONet (deep operator network)	82
fPINN (fractional PINN)	672

List of figures

Figure 1.1	24
Figure 1.2	28
Figure 1.3	30
Figure 1.4: <code>plots/relu.pdf</code>	32
Figure 1.5: <code>plots/clipping.pdf</code>	36
Figure 1.6: <code>plots/softplus.pdf</code>	38
Figure 1.7: <code>plots/gelu.pdf</code>	40
Figure 1.8: <code>plots/logistic.pdf</code>	42
Figure 1.9: <code>plots/swish.pdf</code>	45
Figure 1.10: <code>plots/tanh.pdf</code>	46
Figure 1.11: <code>plots/softsign.pdf</code>	48
Figure 1.12: <code>plots/leaky_relu.pdf</code>	49
Figure 1.13: <code>plots/elu.pdf</code>	51
Figure 1.14: <code>plots/repu.pdf</code>	52
Figure 1.15: <code>plots/sine.pdf</code>	53
Figure 1.16: <code>plots/heaviside.pdf</code>	54
Figure 5.1: <code>plots/gradient_plot1.pdf</code>	189
Figure 5.2: <code>plots/gradient_plot2.pdf</code>	190
Figure 5.3: <code>plots/l1loss.pdf</code>	196
Figure 5.4: <code>plots/mseloss.pdf</code>	197
Figure 5.5: <code>plots/huberloss.pdf</code>	200
Figure 5.6: <code>plots/crossentropyloss.pdf</code>	201
Figure 5.7: <code>plots/kldloss.pdf</code>	207
Figure 6.1: <code>plots/GD_momentum_plots.pdf</code>	311
Figure 7.1: <code>plots/sgd.pdf</code>	370
Figure 7.2: <code>plots/sgd2.pdf</code>	373
Figure 7.3: <code>plots/sgd_momentum.pdf</code>	396
Figure 7.4: <code>plots/mnist.pdf</code>	432
Figure 7.5: <code>plots/mnist_optim.pdf</code>	437
Figure 16.1: <code>plots/pinn.pdf</code>	645
Figure 16.2: <code>plots/dgm.pdf</code>	648

List of figures

Figure 17.1: <code>plots/brownian_motions.pdf</code>	657
Figure 17.2: <code>plots/kolmogorov.pdf</code>	670

List of source codes

Source code 1.1: <code>code/activation_functions/plot_util.py</code>	31
Source code 1.2: <code>code/activation_functions/relu_plot.py</code>	32
Source code 1.3: <code>code/activation_functions/clipping_plot.py</code>	36
Source code 1.4: <code>code/activation_functions/softplus_plot.py</code>	37
Source code 1.5: <code>code/activation_functions/gelu_plot.py</code>	41
Source code 1.6: <code>code/activation_functions/logistic_plot.py</code>	42
Source code 1.7: <code>code/activation_functions/swish_plot.py</code>	44
Source code 1.8: <code>code/activation_functions/tanh_plot.py</code>	46
Source code 1.9: <code>code/activation_functions/softsign_plot.py</code>	48
Source code 1.10: <code>code/activation_functions/leaky_relu_plot.py</code>	49
Source code 1.11: <code>code/activation_functions/elu_plot.py</code>	50
Source code 1.12: <code>code/activation_functions/repu_plot.py</code>	52
Source code 1.13: <code>code/activation_functions/sine_plot.py</code>	53
Source code 1.14: <code>code/activation_functions/heaviside_plot.py</code>	54
Source code 1.15: <code>code/fc-ann-manual.py</code>	59
Source code 1.16: <code>code/fc-ann.py</code>	60
Source code 1.17: <code>code/fc-ann2.py</code>	61
Source code 1.18: <code>code/conv-ann.py</code>	66
Source code 1.19: <code>code/conv-ann-ex.py</code>	69
Source code 1.20: <code>code/res-ann.py</code>	73
Source code 5.1: <code>code/gradient_plot1.py</code>	190
Source code 5.2: <code>code/gradient_plot2.py</code>	191
Source code 5.3: <code>code/loss_functions/l1loss_plot.py</code>	196
Source code 5.4: <code>code/loss_functions/mseloss_plot.py</code>	197
Source code 5.5: <code>code/loss_functions/huberloss_plot.py</code>	200
Source code 5.6: <code>code/loss_functions/crossentropyloss_plot.py</code>	201
Source code 5.7: <code>code/loss_functions/kldloss_plot.py</code>	207
Source code 6.1: <code>code/example_GD_momentum_plots.py</code>	309
Source code 7.1: <code>code/optimization_methods/sgd.py</code>	368
Source code 7.2: <code>code/optimization_methods/sgd2.py</code>	371
Source code 7.3: <code>code/optimization_methods/midpoint_sgd.py</code>	391

Source code 7.4: <code>code/optimization_methods/momentum_sgd.py</code>	394
Source code 7.5: <code>code/optimization_methods/momentum_sgd_bias_adj.py</code>	400
Source code 7.6: <code>code/optimization_methods/nesterov_sgd.py</code>	402
Source code 7.7: <code>code/optimization_methods/adagrad.py</code>	415
Source code 7.8: <code>code/optimization_methods/rmsprop.py</code>	418
Source code 7.9: <code>code/optimization_methods/rmsprop_bias_adj.py</code>	419
Source code 7.10: <code>code/optimization_methods/adadelta.py</code>	422
Source code 7.11: <code>code/optimization_methods/adam.py</code>	425
Source code 7.12: <code>code/mnist.py</code>	426
Source code 7.13: <code>code/mnist_optim.py</code>	431
Source code 16.1: <code>code/pinn.py</code>	642
Source code 16.2: <code>code/dgm.py</code>	645
Source code 17.1: <code>code/brownian_motion.py</code>	656
Source code 17.2: <code>code/kolmogorov.py</code>	668

List of definitions

Chapter 1

Definition 1.1.1: Affine functions	25
Definition 1.1.3: Vectorized description of fully-connected feedforward ANNs	25
Definition 1.2.1: Multi-dimensional versions of one-dimensional functions	29
Definition 1.2.4: ReLU activation function	31
Definition 1.2.5: Multi-dimensional ReLU activation functions	32
Definition 1.2.9: Clipping activation functions	36
Definition 1.2.10: Multi-dimensional clipping activation functions	37
Definition 1.2.11: Softplus activation function.....	37
Definition 1.2.13: Multi-dimensional softplus activation functions	39
Definition 1.2.16: GELU activation function	40
Definition 1.2.18: Multi-dimensional GELU activation functions.....	41
Definition 1.2.19: Standard logistic activation function	41
Definition 1.2.20: Multi-dimensional standard logistic activation functions	42
Definition 1.2.23: Swish activation functions	44
Definition 1.2.24: SiLU activation functions.....	44
Definition 1.2.27: Multi-dimensional swish activation functions.....	45
Definition 1.2.28: Multi-dimensional SiLU activation functions.....	45
Definition 1.2.29: Hyperbolic tangent activation function	46
Definition 1.2.30: Multi-dimensional hyperbolic tangent activation functions	47
Definition 1.2.32: Softsign activation function.....	47
Definition 1.2.33: Multi-dimensional softsign activation functions	48
Definition 1.2.34: Leaky ReLU activation functions	48
Definition 1.2.37: Multi-dimensional leaky ReLU activation functions	50
Definition 1.2.38: ELU activation functions	50
Definition 1.2.40: Multi-dimensional ELU activation functions	51
Definition 1.2.41: RePU activation functions.....	52
Definition 1.2.42: Multi-dimensional RePU activation functions	53
Definition 1.2.43: Sine activation function	53
Definition 1.2.44: Multi-dimensional sine activation functions	54
Definition 1.2.45: Heaviside activation function	54

Definition 1.2.46: Multi-dimensional Heaviside activation functions.....	55
Definition 1.2.47: Softmax activation functions	55
Definition 1.3.1: Structured description of fully-connected feedforward ANNs	56
Definition 1.3.2: Fully-connected feedforward ANNs	57
Definition 1.3.4: Realizations of fully-connected feedforward ANNs	58
Definition 1.3.6: Transformation from the structured to the vectorized description of fully-connected feedforward ANNs	61
Definition 1.4.1: Discrete convolutions	65
Definition 1.4.2: Structured description of feedforward CNNs	65
Definition 1.4.3: Feedforward CNNs	65
Definition 1.4.4: One tensor	66
Definition 1.4.5: Realizations associated to feedforward CNNs	66
Definition 1.4.7: Standard scalar products	71
Definition 1.5.1: Structured description of fully-connected ResNets	72
Definition 1.5.2: Fully-connected ResNets	72
Definition 1.5.4: Realizations associated to fully-connected ResNets	73
Definition 1.5.5: Identity matrices	73
Definition 1.6.1: Function unrolling.....	76
Definition 1.6.2: Description of RNNs	76
Definition 1.6.3: Vectorized description of simple fully-connected RNN nodes.....	77
Definition 1.6.4: Vectorized description of simple fully-connected RNNs	77
Chapter 2	
Definition 2.1.1: Composition of ANNs	83
Definition 2.1.6: Powers of ANNs	90
Definition 2.2.1: Parallelization of ANNs	90
Definition 2.2.6: ReLU identity ANNs	96
Definition 2.2.9: Extensions of ANNs	97
Definition 2.2.13: Parallelization of ANNs with different length	101
Definition 2.3.1: Affine transformation ANNs	103
Definition 2.3.4: Scalar multiplications of ANNs	104
Definition 2.4.1: Sums of vectors as ANNs	105
Definition 2.4.5: Transpose of a matrix	107
Definition 2.4.6: Concatenation of vectors as ANNs	107
Definition 2.4.10: Sums of ANNs with the same length	109
Chapter 3	
Definition 3.1.1: Modulus of continuity	115
Definition 3.1.5: Linear interpolation operator	117
Definition 3.2.1: Activation functions as ANNs	121
Definition 3.3.4: Quasi vector norms.....	130
Chapter 4	

Definition 4.1.1: Metric	135
Definition 4.1.2: Metric space	136
Definition 4.2.1: 1-norm ANN representations	138
Definition 4.2.7: Maxima ANN representations	144
Definition 4.2.8: Floor and ceiling of real numbers	144
Definition 4.3.2: Covering numbers	152
Definition 4.4.1: Rectified clipped ANNs	164

Chapter 5

Definition 5.6.1: Local minimum point	211
Definition 5.6.2: Global minimum point	211
Definition 5.6.3: Local maximum point	211
Definition 5.6.4: Global maximum point	211
Definition 5.6.5: Critical point	212

Chapter 6

Definition 6.1.1: GD optimization method	241
Definition 6.2.1: Explicit midpoint GD optimization method	270
Definition 6.3.1: Momentum GD optimization method	274
Definition 6.3.3: Momentum GD optimization method (2 nd version)	276
Definition 6.3.5: Momentum GD optimization method (3 rd version)	276
Definition 6.3.7: Momentum GD optimization method (4 th version)	277
Definition 6.3.19: Bias-adjusted momentum GD optimization method	289
Definition 6.4.1: Nesterov accelerated GD optimization method	312
Definition 6.4.3: Nesterov accelerated GD optimization method (2 nd version)	313
Definition 6.4.5: Nesterov accelerated GD optimization method (3 rd version)	313
Definition 6.4.7: Nesterov accelerated GD optimization method (4 th version)	314
Definition 6.4.16: Bias-adjusted Nesterov accelerated GD optimization method	323
Definition 6.4.19: Shifted Nesterov accelerated GD optimization method	325
Definition 6.4.22: Shifted Nesterov accelerated GD optimization method (2 nd version)	327
Definition 6.4.25: Shifted Nesterov accelerated GD optimization method (3 rd version)	328
Definition 6.4.28: Shifted Nesterov accelerated GD optimization method (4 th version)	330
Definition 6.4.31: Shifted bias-adjusted Nesterov accelerated GD optimization method	332
Definition 6.4.33: Simplified Nesterov accelerated GD optimization method	333
Definition 6.5.1: Adagrad GD optimization method	334
Definition 6.5.2: Componentwise operations	335
Definition 6.6.1: RMSprop GD optimization method	336
Definition 6.6.5: Bias-adjusted RMSprop GD optimization method	338

Definition 6.7.1: Adadelta GD optimization method	341
Definition 6.8.1: Adam GD optimization method	342
Definition 6.8.3: Adamax GD optimization method	343
Definition 6.9.1: Nadam GD optimization method	345
Definition 6.9.3: Simplified Nadam GD optimization method	346
Definition 6.9.5: Nadamax GD optimization method	347
Definition 6.10.1: AdamW GD optimization method	348
Definition 6.10.3: Adam GD optimization method with L^2 -regularization	349
Definition 6.11.1: Symmetric positive definite matrix	350
Definition 6.11.4: Shampoo GD optimization method	351
Definition 6.12.1: Hilbert-Schmidt norm	352
Definition 6.12.2: Muon GD optimization method	353
Definition 6.12.4: Newton-Schulz method	353
Definition 6.12.5: Muon GD optimization method	354
Definition 6.13.1: AMSgrad GD optimization method	355

Chapter 7

Definition 7.2.1: SGD optimization method	364
Definition 7.3.1: Explicit midpoint SGD optimization method	390
Definition 7.4.1: Momentum SGD optimization method	393
Definition 7.4.3: Momentum SGD optimization method (2 nd version)	396
Definition 7.4.5: Momentum SGD optimization method (3 rd version)	397
Definition 7.4.7: Momentum SGD optimization method (4 th version)	398
Definition 7.4.9: Bias-adjusted momentum SGD optimization method	399
Definition 7.5.1: Nesterov accelerated SGD optimization method	401
Definition 7.5.3: Nesterov accelerated SGD optimization method (2 nd version)	403
Definition 7.5.5: Nesterov accelerated SGD optimization method (3 rd version)	404
Definition 7.5.7: Nesterov accelerated SGD optimization method (4 th version)	405
Definition 7.5.9: Bias-adjusted Nesterov accelerated SGD optimization method	407
Definition 7.5.11: Shifted Nesterov accelerated SGD optimization method	408
Definition 7.5.13: Shifted Nesterov accelerated SGD optimization method (2 nd version)	409
Definition 7.5.15: Shifted Nesterov accelerated SGD optimization method (3 rd version)	410
Definition 7.5.17: Shifted Nesterov accelerated SGD optimization method (4 th version)	411
Definition 7.5.19: Shifted bias-adjusted Nesterov accelerated SGD optimization method	412
Definition 7.5.21: Simplified Nesterov accelerated SGD optimization method	413
Definition 7.6.1: Adagrad SGD optimization method	414
Definition 7.7.1: RMSprop SGD optimization method	417

Definition 7.7.3: Bias-adjusted RMSprop SGD optimization method	419
Definition 7.8.1: Adadelta SGD optimization method	421
Definition 7.9.1: Adam SGD optimization method	423
Definition 7.9.4: Adamax SGD optimization method	436
Definition 7.10.1: Nadam SGD optimization method	438
Definition 7.10.3: Simplified Nadam SGD optimization method	440
Definition 7.10.5: Nadamax SGD optimization method	441
Definition 7.11.1: AdamW SGD optimization method	443
Definition 7.11.3: Adam SGD optimization method with L^2 -regularization	444
Definition 7.12.1: Shampoo SGD optimization method	445
Definition 7.13.1: Idealized Muon SGD optimization method	446
Definition 7.13.3: Muon SGD optimization method	448
Definition 7.14.1: AMSGrad SGD optimization method	449
Chapter 8	
Definition 8.2.1: Diagonal matrices	462
Chapter 9	
Definition 9.1.1: Standard KL inequalities	469
Definition 9.1.2: Standard KL functions	470
Definition 9.7.1: Polynomial	478
Definition 9.8.1: Analytic functions	482
Definition 9.16.1: Fréchet subgradients and limiting Fréchet subgradients	518
Definition 9.16.6: Non-smooth slope	524
Definition 9.16.7: Generalized KL inequalities	524
Definition 9.16.8: Generalized KL functions	525
Chapter 10	
Definition 10.1.1: Batch	527
Definition 10.1.2: Batch mean	527
Definition 10.1.3: Batch variance	527
Definition 10.1.5: BN operations for given batch mean and batch variance	528
Definition 10.1.6: Batch normalization	528
Definition 10.2.1: Structured description of fully-connected feedforward ANNs with BN	530
Definition 10.2.2: Fully-connected feedforward ANNs with BN	530
Definition 10.3.1: Realizations associated to fully-connected feedforward ANNs with BN	531
Definition 10.4.1: Structured description of fully-connected feedforward ANNs with BN for given batch means and batch variances	532
Definition 10.4.2: Fully-connected feedforward ANNs with BN for given batch means and batch variances	532

Definition 10.5.1: Realizations associated to fully-connected feedforward ANNs with BN for given batch means and batch variances	532
Definition 10.6.1: Fully-connected feed-forward ANNs with BN for given batch means and batch variances associated to fully-connected feedforward ANNs with BN and given input batches	533
Chapter 12	
Definition 12.1.7: Moment generating functions.....	564
Definition 12.2.1: Covering radii.....	574
Definition 12.2.3: Packing radii.....	575
Definition 12.2.4: Packing numbers.....	575
Chapter 13	
Definition 13.1.2: Rademacher family.....	598
Definition 13.1.3: p -Kahane–Khintchine constant	599
Chapter 17	
Definition 17.3.3: Standard Brownian motions	655
Definition 17.3.8: Continuous convolutions	661

List of exercises

Exercise 1.1.1	25
Exercise 1.1.2	26
Exercise 1.1.3	27
Exercise 1.1.4	27
Exercise 1.2.1	29
Exercise 1.2.2	30
Exercise 1.2.3 (Real identity)	33
Exercise 1.2.4 (Absolute value)	33
Exercise 1.2.5 (Exponential)	34
Exercise 1.2.6 (Two-dimensional maximum)	34
Exercise 1.2.7 (Real identity with two hidden layers)	34
Exercise 1.2.8 (Three-dimensional maximum)	35
Exercise 1.2.9 (Multi-dimensional maxima)	35
Exercise 1.2.10	35
Exercise 1.2.11 (Hat function)	35
Exercise 1.2.12	35
Exercise 1.2.13	35
Exercise 1.2.14	36
Exercise 1.2.15	36
Exercise 1.2.16 (Real identity)	39
Exercise 1.2.17	47
Exercise 1.3.1	58
Exercise 1.3.2	59
Exercise 1.3.3	63
Exercise 1.3.4	63
Exercise 1.3.5	63
Exercise 1.4.1	70
Exercise 1.4.2	70
Exercise 1.4.3	70
Exercise 1.4.4	71
Exercise 1.5.1	75

Exercise 1.6.1	78
Exercise 2.2.1	95
Exercise 2.2.2	95
Exercise 2.2.3	102
Exercise 3.2.1	126
Exercise 3.2.2	126
Exercise 3.2.3	126
Exercise 3.3.1	133
Exercise 3.3.2	133
Exercise 4.1.1	138
Exercise 4.2.1	144
Exercise 4.2.2	148
Exercise 4.3.1	152
Exercise 4.3.2	154
Exercise 4.3.3	154
Exercise 4.3.4	154
Exercise 4.3.5	154
Exercise 6.1.1	242
Exercise 6.1.2	242
Exercise 6.1.3	248
Exercise 6.1.4	248
Exercise 6.1.5	248
Exercise 6.1.6	248
Exercise 6.1.7	248
Exercise 6.1.8	249
Exercise 6.1.9	253
Exercise 6.1.10	253
Exercise 6.1.11	253
Exercise 6.3.1	275
Exercise 6.3.2	275
Exercise 6.3.3	311
Exercise 9.11.1	494
Exercise 9.12.1	504
Exercise 9.13.1	506
Exercise 9.13.2	506
Exercise 9.13.3	511
Exercise 9.13.4	512
Exercise 9.16.1	524
Exercise 9.16.2	524
Exercise 9.16.3	524

Exercise 9.16.4	524
Exercise 9.16.5	524
Exercise 10.3.1	531
Exercise 10.6.1	534
Exercise 12.1.1	572
Exercise 12.1.2	572
Exercise 12.1.3	573
Exercise 12.1.4	573
Exercise 12.2.1	574
Exercise 12.2.2	574
Exercise 12.2.3	574
Exercise 12.2.4	575
Exercise 17.3.1	661
Exercise 17.3.2	662
Exercise 17.3.3	662
Exercise 17.3.4	662

Bibliography

- [1] ABDEL-HAMID, O., MOHAMED, A., JIANG, H., DENG, L., PENN, G., AND YU, D. Convolutional Neural Networks for Speech Recognition. *IEEE/ACM Trans. Audio, Speech, Language Process.* 22, 10 (2014), pp. 1533–1545. URL: doi.org/10.1109/TASLP.2014.2339736.
- [2] ABSIL, P.-A., MAHONY, R., AND ANDREWS, B. Convergence of the iterates of descent methods for analytic cost functions. *SIAM J. Optim.* 16, 2 (2005), pp. 531–547. URL: doi.org/10.1137/040605266.
- [3] ACKERMANN, J., JENTZEN, A., KRUSE, T., KUCKUCK, B., AND PADGETT, J. L. Deep neural networks with ReLU, leaky ReLU, and softplus activation provably overcome the curse of dimensionality for Kolmogorov partial differential equations with Lipschitz nonlinearities in the L^p -sense. *arXiv:2309.13722* (2023), 52 pp. URL: arxiv.org/abs/2309.13722.
- [4] ALPAYDIN, E. *Introduction to Machine Learning*. 4th ed. MIT Press, Cambridge, Mass., 2020. 712 pp.
- [5] AMANN, H. *Ordinary differential equations*. Walter de Gruyter & Co., Berlin, 1990. xiv+458 pp. URL: doi.org/10.1515/9783110853698.
- [6] AMODEI, D., ANANTHANARAYANAN, S., ANUBHAI, R., BAI, J., BATTENBERG, E., CASE, C., CASPER, J., CATANZARO, B., CHENG, Q., CHEN, G., CHEN, J., CHEN, J., CHEN, Z., CHRZANOWSKI, M., COATES, A., DIAMOS, G., DING, K., DU, N., ELSEN, E., ENGEL, J., FANG, W., FAN, L., FOUGNER, C., GAO, L., GONG, C., HANNUN, A., HAN, T., JOHANNES, L., JIANG, B., JU, C., JUN, B., LEGRESLEY, P., LIN, L., LIU, J., LIU, Y., LI, W., LI, X., MA, D., NARANG, S., NG, A., OZAIR, S., PENG, Y., PRENGER, R., QIAN, S., QUAN, Z., RAIMAN, J., RAO, V., SATHEESH, S., SEETAPUN, D., SENGUPTA, S., SRINET, K., SRIRAM, A., TANG, H., TANG, L., WANG, C., WANG, J., WANG, K., WANG, Y., WANG, Z., WANG, Z., WU, S., WEI, L., XIAO, B., XIE, W., XIE, Y., YOGATAMA, D., YUAN, B., ZHAN, J., AND ZHU, Z. Deep Speech 2 : End-to-End Speech Recognition in English and Mandarin. In *Proceedings of The 33rd International Conference on Machine Learning* (New York, NY, USA, June 20–22, 2016). Ed. by Balcan, M. F.

- and Weinberger, K. Q. Vol. 48. Proceedings of Machine Learning Research. PMLR, 2016, pp. 173–182. URL: proceedings.mlr.press/v48/amodei16.html.
- [7] AN, J. AND LU, J. Convergence of stochastic gradient descent under a local Łojasiewicz condition for deep neural networks. *arXiv:2304.09221* (2023), 14 pp. URL: arxiv.org/abs/2304.09221.
- [8] ATTOUTCH, H. AND BOLTE, J. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Math. Program.* 116, 1–2 (2009), pp. 5–16. URL: doi.org/10.1007/s10107-007-0133-5.
- [9] BACH, F. *Learning Theory from First Principles*. Draft version of April 19, 2023. book draft, to be published by MIT Press. 2023. URL: www.di.ens.fr/%7Efbach/ltpf_book.pdf.
- [10] BAGGENSTOS, J. AND SALIMOVA, D. Approximation properties of residual neural networks for Kolmogorov PDEs. *Discrete Contin. Dyn. Syst. Ser. B* 28, 5 (2023), pp. 3193–3215. URL: doi.org/10.3934/dcdsb.2022210.
- [11] BAHDANAU, D., CHO, K., AND BENGIO, Y. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv:1409.0473* (2014), 15 pp. URL: arxiv.org/abs/1409.0473.
- [12] BALDI, P. AND HORNIK, K. Neural networks and principal component analysis: Learning from examples without local minima. *Neural Networks* 2, 1 (1989), pp. 53–58. URL: [doi.org/10.1016/0893-6080\(89\)90014-2](https://doi.org/10.1016/0893-6080(89)90014-2).
- [13] BAO, G., YE, X., ZANG, Y., AND ZHOU, H. Numerical solution of inverse problems by weak adversarial networks. *Inverse Problems* 36, 11 (2020), Art. No. 115003, 31 pp. URL: doi.org/10.1088/1361-6420/abb447.
- [14] BARRON, A. R. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. Inform. Theory* 39, 3 (1993), pp. 930–945. URL: doi.org/10.1109/18.256500.
- [15] BARRON, A. R. Approximation and estimation bounds for artificial neural networks. *Mach. Learn.* 14, 1 (1994), pp. 115–133. URL: doi.org/10.1007/bf00993164.
- [16] BATTAGLIA, P. W., HAMRICK, J. B., BAPST, V., SANCHEZ-GONZALEZ, A., ZAMBALDI, V., MALINOWSKI, M., TACCHETTI, A., RAPOSO, D., SANTORO, A., FAULKNER, R., GULCEHRE, C., SONG, F., BALLARD, A., GILMER, J., DAHL, G., VASWANI, A., ALLEN, K., NASH, C., LANGSTON, V., DYER, C., HEESS, N., WIERSTRA, D., KOHLI, P., BOTVINICK, M., VINYALS, O., LI, Y., AND PASCANU, R. Relational inductive biases, deep learning, and graph networks. *arXiv:1806.01261* (2018), 40 pp. URL: arxiv.org/abs/1806.01261.

- [17] BECK, C., BECKER, S., CHERIDITO, P., JENTZEN, A., AND NEUFELD, A. Deep splitting method for parabolic PDEs. *SIAM J. Sci. Comput.* 43, 5 (2021), A3135–A3154. URL: doi.org/10.1137/19M1297919.
- [18] BECK, C., BECKER, S., GROHS, P., JAAFARI, N., AND JENTZEN, A. Solving stochastic differential equations and Kolmogorov equations by means of deep learning. *arXiv:1806.00421* (2018), 56 pp. URL: arxiv.org/abs/1806.00421.
- [19] BECK, C., BECKER, S., GROHS, P., JAAFARI, N., AND JENTZEN, A. Solving the Kolmogorov PDE by means of deep learning. *J. Sci. Comput.* 88, 3 (2021), Art. No. 73, 28 pp. URL: doi.org/10.1007/s10915-021-01590-0.
- [20] BECK, C., E, W., AND JENTZEN, A. Machine learning approximation algorithms for high-dimensional fully nonlinear partial differential equations and second-order backward stochastic differential equations. *J. Nonlinear Sci.* 29, 4 (2019), pp. 1563–1619. URL: doi.org/10.1007/s00332-018-9525-3.
- [21] BECK, C., GONON, L., AND JENTZEN, A. Overcoming the curse of dimensionality in the numerical approximation of high-dimensional semilinear elliptic partial differential equations. *arXiv:2003.00596* (2020), 50 pp. URL: arxiv.org/abs/2003.00596.
- [22] BECK, C., HORNUNG, F., HUTZENTHALER, M., JENTZEN, A., AND KRUSE, T. Overcoming the curse of dimensionality in the numerical approximation of Allen-Cahn partial differential equations via truncated full-history recursive multilevel Picard approximations. *J. Numer. Math.* 28, 4 (2020), pp. 197–222. URL: doi.org/10.1515/jnma-2019-0074.
- [23] BECK, C., HUTZENTHALER, M., AND JENTZEN, A. On nonlinear Feynman–Kac formulas for viscosity solutions of semilinear parabolic partial differential equations. *Stoch. Dyn.* 21, 8 (2021), Art. No. 2150048, 68 pp. URL: doi.org/10.1142/S0219493721500489.
- [24] BECK, C., HUTZENTHALER, M., JENTZEN, A., AND KUCKUCK, B. An overview on deep learning-based approximation methods for partial differential equations. *Discrete Contin. Dyn. Syst. Ser. B* 28, 6 (2023), pp. 3697–3746. URL: doi.org/10.3934/dcdsb.2022238.
- [25] BECK, C., JENTZEN, A., AND KUCKUCK, B. Full error analysis for the training of deep neural networks. *Infin. Dimens. Anal. Quantum Probab. Relat. Top.* 25, 2 (2022), Art. No. 2150020, 76 pp. URL: doi.org/10.1142/S021902572150020X.
- [26] BELAK, C., HAGER, O., REIMERS, C., SCHNELL, L., AND WÜRSCHMIDT, M. Convergence Rates for a Deep Learning Algorithm for Semilinear PDEs (2021). Available at SSRN, 42 pp. URL: doi.org/10.2139/ssrn.3981933.
- [27] BELLMAN, R. *Dynamic programming*. Reprint of the 1957 edition. Princeton University Press, Princeton, NJ, 2010, xxx+340 pp. URL: doi.org/10.1515/9781400835386.

- [28] BENEVENTANO, P., CHERIDITO, P., GRAEBER, R., JENTZEN, A., AND KUCKUCK, B. Deep neural network approximation theory for high-dimensional functions. *arXiv:2112.14523* (2021), 82 pp. URL: arxiv.org/abs/2112.14523.
- [29] BENEVENTANO, P., CHERIDITO, P., JENTZEN, A., AND VON WURSTEMBERGER, P. High-dimensional approximation spaces of artificial neural networks and applications to partial differential equations. *arXiv:2012.04326* (2020). URL: arxiv.org/abs/2012.04326.
- [30] BENGIO, Y., SIMARD, P., AND FRASCONI, P. Learning long-term dependencies with gradient descent is difficult. *IEEE Trans. Neural Netw.* 5, 2 (1994), pp. 157–166. URL: doi.org/10.1109/72.279181.
- [31] BENGIO, Y., BOULANGER-LEWANDOWSKI, N., AND PASCANU, R. Advances in optimizing recurrent networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (Vancouver, BC, Canada, May 26–31, 2013). 2013, pp. 8624–8628. URL: doi.org/10.1109/ICASSP.2013.6639349.
- [32] BENTH, F. E., DETERING, N., AND GALIMBERTI, L. Neural networks in Fréchet spaces. *Ann. Math. Artif. Intell.* 91, 1 (2023), pp. 75–103. URL: doi.org/10.1007/s10472-022-09824-z.
- [33] BERCU, B. AND FORT, J.-C. Generic Stochastic Gradient Methods. In *Wiley Encyclopedia of Operations Research and Management Science*. Ed. by Cochran, J. J., Cox Jr., L. A., Keskinocak, P., Kharoufeh, J. P., and Smith, J. C. John Wiley & Sons, Ltd., 2013. URL: doi.org/10.1002/9780470400531.eorms1068.
- [34] BERG, J. AND NYSTRÖM, K. A unified deep artificial neural network approach to partial differential equations in complex geometries. *Neurocomputing* 317 (2018), pp. 28–41. URL: doi.org/10.1016/j.neucom.2018.06.056.
- [35] BERNER, J., GROHS, P., AND JENTZEN, A. Analysis of the Generalization Error: Empirical Risk Minimization over Deep Artificial Neural Networks Overcomes the Curse of Dimensionality in the Numerical Approximation of Black–Scholes Partial Differential Equations. *SIAM J. Math. Data Sci.* 2, 3 (2020), pp. 631–657. URL: doi.org/10.1137/19M125649X.
- [36] BERNER, J., GROHS, P., KUTYNIOK, G., AND PETERSEN, P. The Modern Mathematics of Deep Learning. In *Mathematical Aspects of Deep Learning*. Ed. by Grohs, P. and Kutyniok, G. Cambridge University Press, 2022, pp. 1–111. URL: doi.org/10.1017/9781009025096.002.
- [37] BERNSTEIN, J. AND NEWHOUSE, L. Old Optimizer, New Norm: An Anthology. *arXiv:2409.20325* (2024). URL: doi.org/10.48550/ARXIV.2409.20325.
- [38] BEZNEA, L., CIMPEAN, I., LUPASCU-STAMATE, O., POPESCU, I., AND ZARNESCU, A. From Monte Carlo to neural networks approximations of boundary value problems. *arXiv:2209.01432* (2022), 40 pp. URL: arxiv.org/abs/2209.01432.

- [39] BIERSTONE, E. AND MILMAN, P. D. Semianalytic and subanalytic sets. *Inst. Hautes Études Sci. Publ. Math.* 67 (1988), pp. 5–42. URL: doi.org/10.1007/BF02699126.
- [40] BISHOP, C. M. *Neural networks for pattern recognition*. The Clarendon Press, Oxford University Press, New York, 1995, xviii+482 pp.
- [41] BJORCK, N., GOMES, C. P., SELMAN, B., AND WEINBERGER, K. Q. Understanding Batch Normalization. In *Advances in Neural Information Processing Systems (NeurIPS 2018)*. Ed. by Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. Vol. 31. Curran Associates, Inc., 2018. URL: proceedings.neurips.cc/paper_files/paper/2018/file/36072923bfc3cf47745d704feb489480-Paper.pdf.
- [42] BLUM, A., HOPCROFT, J., AND KANNAN, R. *Foundations of Data Science*. Cambridge University Press, 2020. URL: doi.org/10.1017/9781108755528.
- [43] BLUM, E. K. AND LI, L. K. Approximation theory and feedforward networks. *Neural Networks* 4, 4 (1991), pp. 511–515. URL: [doi.org/10.1016/0893-6080\(91\)90047-9](https://doi.org/10.1016/0893-6080(91)90047-9).
- [44] BLUMERS, A. L., LI, Z., AND KARNIADAKIS, G. E. Supervised parallel-in-time algorithm for long-time Lagrangian simulations of stochastic dynamics: Application to hydrodynamics. *J. Comput. Phys.* 393 (2019), pp. 214–228. URL: doi.org/10.1016/j.jcp.2019.05.016.
- [45] BÖLCSKEI, H., GROHS, P., KUTYNIOK, G., AND PETERSEN, P. Optimal approximation with sparsely connected deep neural networks. *SIAM J. Math. Data Sci.* 1, 1 (2019), pp. 8–45. URL: doi.org/10.1137/18M118709X.
- [46] BOLTE, J., DANIILIDIS, A., AND LEWIS, A. The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM J. Optim.* 17, 4 (2006), pp. 1205–1223. URL: doi.org/10.1137/050644641.
- [47] BOLTE, J. AND PAUWELS, E. Conservative set valued fields, automatic differentiation, stochastic gradient methods and deep learning. *Math. Program.* 188, 1 (2021), pp. 19–51. URL: doi.org/10.1007/s10107-020-01501-5.
- [48] BOROVYKH, A., BOHTE, S., AND OOSTERLEE, C. W. Conditional Time Series Forecasting with Convolutional Neural Networks. *arXiv:1703.04691* (2017), 22 pp. URL: arxiv.org/abs/1703.04691.
- [49] BOTTOU, L., CORTES, C., DENKER, J., DRUCKER, H., GUYON, I., JACKEL, L., LECUN, Y., MULLER, U., SACKINGER, E., SIMARD, P., AND VAPNIK, V. Comparison of classifier methods: a case study in handwritten digit recognition. In *Proceedings of the 12th IAPR International Conference on Pattern Recognition, Vol. 3 - Conference C: Signal Processing (Cat. No.94CH3440-5)* (Jerusalem, Israel, Oct. 9–13, 1994). Vol. 2. 1994, pp. 77–82. URL: doi.org/10.1109/ICPR.1994.576879.

- [50] BOTTOU, L., CURTIS, F. E., AND NOCEDAL, J. Optimization Methods for Large-Scale Machine Learning. *SIAM Rev.* 60, 2 (2018), pp. 223–311. URL: doi.org/10.1137/16M1080173.
- [51] BOURLARD, H. AND KAMP, Y. Auto-association by multilayer perceptrons and singular value decomposition. *Biol. Cybernet.* 59, 4–5 (1988), pp. 291–294. URL: doi.org/10.1007/BF00332918.
- [52] BOUSSANGE, V., BECKER, S., JENTZEN, A., KUCKUCK, B., AND PELLISSIER, L. Deep learning approximations for non-local nonlinear PDEs with Neumann boundary conditions. *arXiv:2205.03672* (2022), 59 pp. URL: arxiv.org/abs/2205.03672.
- [53] BOWMAN, S. R., VILNIS, L., VINYALS, O., DAI, A., JOZEFOWICZ, R., AND BENGIO, S. Generating Sentences from a Continuous Space. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning* (Berlin, Germany, Aug. 7–12, 2016). Ed. by Riezler, S. and Goldberg, Y. Association for Computational Linguistics, 2016, pp. 10–21. URL: doi.org/10.18653/v1/K16-1002.
- [54] BOYD, S. AND VANDENBERGHE, L. *Convex Optimization*. Cambridge University Press, 2004. 727 pp. URL: doi.org/10.1017/CBO9780511804441.
- [55] BRANDSTETTER, J., VAN DEN BERG, R., WELLING, M., AND GUPTA, J. K. Clifford Neural Layers for PDE Modeling. *arXiv:2209.04934* (2022), 58 pp. URL: arxiv.org/abs/2209.04934.
- [56] BROWN, T. B., MANN, B., RYDER, N., SUBBIAH, M., KAPLAN, J., DHARIWAL, P., NEELAKANTAN, A., SHYAM, P., SASTRY, G., ASKELL, A., AGARWAL, S., HERBERT-VOSS, A., KRUEGER, G., HENIGHAN, T., CHILD, R., RAMESH, A., ZIEGLER, D. M., WU, J., WINTER, C., HESSE, C., CHEN, M., SIGLER, E., LITWIN, M., GRAY, S., CHESS, B., CLARK, J., BERNER, C., MCCANDLISH, S., RADFORD, A., SUTSKEVER, I., AND AMODEI, D. Language Models are Few-Shot Learners. *arXiv:2005.14165* (2020), 75 pp. URL: arxiv.org/abs/2005.14165.
- [57] BRUNA, J., ZAREMBA, W., SZLAM, A., AND LECUN, Y. Spectral Networks and Locally Connected Networks on Graphs. *arXiv:1312.6203* (2013), 14 pp. URL: arxiv.org/abs/1312.6203.
- [58] BRUNTON, S. L. AND KUTZ, J. N. Machine Learning for Partial Differential Equations. *arXiv:2303.17078* (2023), 16 pp. URL: arxiv.org/abs/2303.17078.
- [59] BUBECK, S. Convex Optimization: Algorithms and Complexity. *Found. Trends Mach. Learn.* 8, 3–4 (2015), pp. 231–357. URL: doi.org/10.1561/2200000050.
- [60] CAI, W. AND XU, Z.-Q. J. Multi-scale Deep Neural Networks for Solving High Dimensional PDEs. *arXiv:1910.11710* (2019), 14 pp. URL: arxiv.org/abs/1910.11710.

- [61] CAKIR, E., PARASCANDOLO, G., HEITTO LA, T., HUTTUNEN, H., AND VIRTANEN, T. Convolutional Recurrent Neural Networks for Polyphonic Sound Event Detection. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.* 25, 6 (2017), pp. 1291–1303. URL: doi.org/10.1109/TASLP.2017.2690575.
- [62] CALIN, O. *Deep learning architectures—a mathematical approach*. Springer, Cham, 2020, xxx+760 pp. URL: doi.org/10.1007/978-3-030-36721-3.
- [63] CARL, B. AND STEPHANI, I. *Entropy, compactness and the approximation of operators*. Vol. 98. Cambridge University Press, Cambridge, 1990, x+277 pp. URL: doi.org/10.1017/CBO9780511897467.
- [64] CASTRO, J. Deep learning schemes for parabolic nonlocal integro-differential equations. *Partial Differ. Equ. Appl.* 3, 6 (2022), Art. No. 77, 35 pp. URL: doi.org/10.1007/s42985-022-00213-z.
- [65] CATERINI, A. L. AND CHANG, D. E. *Deep neural networks in a mathematical framework*. Springer, Cham, 2018, xiii+84 pp. URL: doi.org/10.1007/978-3-319-75304-1.
- [66] CHAN-WAI-NAM, Q., MIKAEL, J., AND WARIN, X. Machine learning for semi linear PDEs. *J. Sci. Comput.* 79, 3 (2019), pp. 1667–1712. URL: doi.org/10.1007/s10915-019-00908-3.
- [67] CHATTERJEE, S. Convergence of gradient descent for deep neural networks. *arXiv:2203.16462* (2022), 23 pp. URL: arxiv.org/abs/2203.16462.
- [68] CHEN, F., HUANG, J., WANG, C., AND YANG, H. Friedrichs Learning: Weak Solutions of Partial Differential Equations via Deep Learning. *SIAM J. Sci. Comput.* 45, 3 (2023), A1271–A1299. URL: doi.org/10.1137/22M1488405.
- [69] CHEN, K., WANG, C., AND YANG, H. Deep Operator Learning Lessens the Curse of Dimensionality for PDEs. *arXiv:2301.12227* (2023), 21 pp. URL: arxiv.org/abs/2301.12227.
- [70] CHEN, T. AND CHEN, H. Approximations of continuous functionals by neural networks with application to dynamic systems. *IEEE Trans. Neural Netw.* 4, 6 (1993), pp. 910–918. URL: doi.org/10.1109/72.286886.
- [71] CHEN, T. AND CHEN, H. Universal approximation to nonlinear operators by neural networks with arbitrary activation functions and its application to dynamical systems. *IEEE Trans. Neural Netw.* 6, 4 (1995), pp. 911–917. URL: doi.org/10.1109/72.392253.
- [72] CHERIDITO, P., JENTZEN, A., AND ROSSMANEK, F. Non-convergence of stochastic gradient descent in the training of deep neural networks. *J. Complexity* 64 (2021), Paper No. 101540, 10. URL: doi.org/10.1016/j.jco.2020.101540.

- [73] CHERIDITO, P., JENTZEN, A., AND ROSSMANEK, F. Efficient approximation of high-dimensional functions with neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* 33, 7 (2022), pp. 3079–3093. URL: doi.org/10.1109/TNNLS.2021.3049719.
- [74] CHERIDITO, P., JENTZEN, A., AND ROSSMANEK, F. Gradient descent provably escapes saddle points in the training of shallow ReLU networks. *arXiv:2208.02083* (2022), 16 pp. URL: arxiv.org/abs/2208.02083.
- [75] CHERIDITO, P., JENTZEN, A., AND ROSSMANEK, F. Landscape analysis for shallow neural networks: complete classification of critical points for affine target functions. *J. Nonlinear Sci.* 32, 5 (2022), Art. No. 64, 45 pp. URL: doi.org/10.1007/s00332-022-09823-8.
- [76] CHERIDITO, P., SONER, H. M., TOUZI, N., AND VICTOIR, N. Second-order backward stochastic differential equations and fully nonlinear parabolic PDEs. *Comm. Pure Appl. Math.* 60, 7 (2007), pp. 1081–1110. URL: doi.org/10.1002/cpa.20168.
- [77] CHIZAT, L. AND BACH, F. On the Global Convergence of Gradient Descent for Over-parameterized Models using Optimal Transport. In *Advances in Neural Information Processing Systems (NeurIPS 2018)*. Ed. by Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. Vol. 31. Curran Associates, Inc., 2018. URL: proceedings.neurips.cc/paper_files/paper/2018/file/a1afc58c6ca9540d057299ec3016d726-Paper.pdf.
- [78] CHIZAT, L., OYALLON, E., AND BACH, F. On Lazy Training in Differentiable Programming. In *Advances in Neural Information Processing Systems (NeurIPS 2019)*. Ed. by Wallach, H., Larochelle, H., Beygelzimer, A., d’Alché-Buc, F., Fox, E., and Garnett, R. Vol. 32. Curran Associates, Inc., 2019. URL: proceedings.neurips.cc/paper_files/paper/2019/file/ae614c557843b1df326cb29c57225459-Paper.pdf.
- [79] CHO, K., VAN MERRIËNBOER, B., BAHDANAU, D., AND BENGIO, Y. On the Properties of Neural Machine Translation: Encoder–Decoder Approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation* (Doha, Qatar, Oct. 25, 2014). Association for Computational Linguistics, 2014, pp. 103–111. URL: doi.org/10.3115/v1/W14-4012.
- [80] CHO, K., VAN MERRIËNBOER, B., GULCEHRE, C., BAHDANAU, D., BOUGARES, F., SCHWENK, H., AND BENGIO, Y. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. *arXiv:1406.1078* (2014), 15 pp. URL: arxiv.org/abs/1406.1078.
- [81] CHOI, K., FAZEKAS, G., SANDLER, M., AND CHO, K. Convolutional recurrent neural networks for music classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (New Orleans, LA, USA, Mar. 5–9, 2017). 2017, pp. 2392–2396. URL: doi.org/10.1109/ICASSP.2017.7952585.

- [82] CHOROMANSKA, A., HENAFF, M., MATHIEU, M., BEN AROUS, G., AND LECUN, Y. The Loss Surfaces of Multilayer Networks. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics* (San Diego, California, USA, May 9–12, 2015). Ed. by Lebanon, G. and Vishwanathan, S. V. N. Vol. 38. Proceedings of Machine Learning Research. PMLR, 2015, pp. 192–204. URL: proceedings.mlr.press/v38/choromanska15.html.
- [83] CHOROMANSKA, A., LECUN, Y., AND BEN AROUS, G. Open Problem: The landscape of the loss surfaces of multilayer networks. In *Proceedings of The 28th Conference on Learning Theory* (Paris, France, July 3–6, 2015). Ed. by Grünwald, P., Hazan, E., and Kale, S. Vol. 40. Proceedings of Machine Learning Research. PMLR, 2015, pp. 1756–1760. URL: proceedings.mlr.press/v40/Choromanska15.html.
- [84] CHOROWSKI, J. K., BAHDANAU, D., SERDYUK, D., CHO, K., AND BENGIO, Y. Attention-Based Models for Speech Recognition. In *Advances in Neural Information Processing Systems (NeurIPS 2015)*. Ed. by Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R. Vol. 28. Curran Associates, Inc., 2015. URL: proceedings.neurips.cc/paper_files/paper/2015/file/1068c6e4c8051cf4e9ea8072e3189e2-Paper.pdf.
- [85] CIOICA-LICHT, P. A., HUTZENTHALER, M., AND WERNER, P. T. Deep neural networks overcome the curse of dimensionality in the numerical approximation of semilinear partial differential equations. *arXiv:2205.14398* (2022), 34 pp. URL: arxiv.org/abs/2205.14398.
- [86] CLEVERT, D.-A., UNTERTHINER, T., AND HOCHREITER, S. Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). *arXiv:1511.07289* (2015), 14 pp. URL: arxiv.org/abs/1511.07289.
- [87] COLDING, T. H. AND MINICOZZI II, W. P. Łojasiewicz inequalities and applications. In *Surveys in Differential Geometry 2014. Regularity and evolution of nonlinear equations*. Vol. 19. Int. Press, Somerville, MA, 2015, pp. 63–82. URL: doi.org/10.4310/SDG.2014.v19.n1.a3.
- [88] COLEMAN, R. *Calculus on normed vector spaces*. Springer New York, 2012, xi+249 pp. URL: doi.org/10.1007/978-1-4614-3894-6.
- [89] COX, S., HUTZENTHALER, M., JENTZEN, A., VAN NEERVEN, J., AND WELTI, T. Convergence in Hölder norms with applications to Monte Carlo methods in infinite dimensions. *IMA J. Numer. Anal.* 41, 1 (2020), pp. 493–548. URL: doi.org/10.1093/imanum/drz063.
- [90] CUCKER, F. AND SMALE, S. On the mathematical foundations of learning. *Bull. Amer. Math. Soc. (N.S.)* 39, 1 (2002), pp. 1–49. URL: doi.org/10.1090/S0273-0979-01-00923-5.

- [91] CUOMO, S., DI COLA, V. S., GIAMPAOLO, F., ROZZA, G., RAISSI, M., AND PICCIALLI, F. Scientific Machine Learning Through Physics-Informed Neural Networks: Where we are and What's Next. *J. Sci. Comp.* 92, 3 (2022), Art. No. 88, 62 pp. URL: doi.org/10.1007/s10915-022-01939-z.
- [92] CYBENKO, G. Approximation by superpositions of a sigmoidal function. *Math. Control Signals Systems* 2, 4 (1989), pp. 303–314. URL: doi.org/10.1007/BF02551274.
- [93] D. JAGTAP, A. AND EM KARNIADAKIS, G. Extended Physics-Informed Neural Networks (XPINNs): A Generalized Space-Time Domain Decomposition Based Deep Learning Framework for Nonlinear Partial Differential Equations. *Commun. Comput. Phys.* 28, 5 (2020), pp. 2002–2041. URL: doi.org/10.4208/cicp.OA-2020-0164.
- [94] DAI, Z., YANG, Z., YANG, Y., CARBONELL, J., LE, Q., AND SALAKHUTDINOV, R. Transformer-XL: Attentive Language Models beyond a Fixed-Length Context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Florence, Italy, July 28–Aug. 2, 2019). Association for Computational Linguistics, 2019, pp. 2978–2988. URL: doi.org/10.18653/v1/P19-1285.
- [95] DAUPHIN, Y. N., PASCANU, R., GULCEHRE, C., CHO, K., GANGULI, S., AND BENGIO, Y. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Advances in Neural Information Processing Systems*. Ed. by Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K. Vol. 27. Curran Associates, Inc., 2014. URL: proceedings.neurips.cc/paper_files/paper/2014/file/17e23e50bedc63b4095e3d8204ce063b-Paper.pdf.
- [96] DAVIS, D., DRUSVYATSKIY, D., KAKADE, S., AND LEE, J. D. Stochastic subgradient method converges on tame functions. *Found. Comput. Math.* 20, 1 (2020), pp. 119–154. URL: doi.org/10.1007/s10208-018-09409-5.
- [97] DE RYCK, T. AND MISHRA, S. Generic bounds on the approximation error for physics-informed (and) operator learning. *arXiv:2205.11393* (2022), 40 pp. URL: arxiv.org/abs/2205.11393.
- [98] DEFFERRARD, M., BRESSON, X., AND VANDERHEYNST, P. Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering. In *Advances in Neural Information Processing Systems*. Ed. by Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R. Vol. 29. Curran Associates, Inc., 2016. URL: proceedings.neurips.cc/paper_files/paper/2016/file/04df4d434d481c5bb723be1b6df1ee65-Paper.pdf.
- [99] DÉFOSSEZ, A., BOTTOU, L., BACH, F., AND USUNIER, N. A Simple Convergence Proof of Adam and Adagrad. *arXiv:2003.02395* (2020), 30 pp. URL: arxiv.org/abs/2003.02395.

- [100] DEISENROTH, M. P., FAISAL, A. A., AND ONG, C. S. *Mathematics for machine learning*. Cambridge University Press, Cambridge, 2020, xvii+371 pp. URL: doi.org/10.1017/9781108679930.
- [101] DENG, B., SHIN, Y., LU, L., ZHANG, Z., AND KARNIADAKIS, G. E. Approximation rates of DeepONets for learning operators arising from advection–diffusion equations. *Neural Networks* 153 (2022), pp. 411–426. URL: doi.org/10.1016/j.neunet.2022.06.019.
- [102] DEREICH, S., DO, T., JENTZEN, A., AND WEBER, F. Mathematical analysis of the gradients in deep learning. *arXiv:2501.15646* (2025). URL: arxiv.org/abs/2501.15646.
- [103] DEREICH, S., GRAEBER, R., AND JENTZEN, A. Non-convergence of Adam and other adaptive stochastic gradient descent optimization methods for non-vanishing learning rates. *arXiv:2407.08100* (2024). URL: doi.org/10.48550/ARXIV.2407.08100.
- [104] DEREICH, S. AND JENTZEN, A. Convergence rates for the Adam optimizer. *arXiv:2407.21078* (2024). URL: doi.org/10.48550/ARXIV.2407.21078.
- [105] DEREICH, S., JENTZEN, A., AND KASSING, S. On the existence of minimizers in shallow residual ReLU neural network optimization landscapes. *arXiv:2302.14690* (2023), 26 pp. URL: arxiv.org/abs/2302.14690.
- [106] DEREICH, S. AND KASSING, S. Convergence of stochastic gradient descent schemes for Lojasiewicz-landscapes. *arXiv:2102.09385* (2021), 24 pp. URL: arxiv.org/abs/2102.09385.
- [107] DEREICH, S. AND KASSING, S. Cooling down stochastic differential equations: Almost sure convergence. *Stochastic Process. Appl.* 152 (2022), pp. 289–311. URL: doi.org/10.1016/j.spa.2022.06.020.
- [108] DEREICH, S. AND KASSING, S. On the existence of optimal shallow feedforward networks with ReLU activation. *arXiv:2303.03950* (2023), 17 pp. URL: arxiv.org/abs/2303.03950.
- [109] DEREICH, S. AND MÜLLER-GRONBACH, T. General multilevel adaptations for stochastic approximation algorithms. *arXiv:1506.05482* (2017), 33 pages. URL: arxiv.org/abs/1506.05482.
- [110] DEVLIN, J., CHANG, M.-W., LEE, K., AND TOUTANOVA, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (Minneapolis, MN, USA, June 2–7, 2019). Association for Computational Linguistics, 2019, pp. 4171–4186. URL: doi.org/10.18653/v1/N19-1423.

- [111] DING, X., ZHANG, Y., LIU, T., AND DUAN, J. Deep Learning for Event-Driven Stock Prediction. In *Proceedings of the 24th International Conference on Artificial Intelligence* (Buenos Aires, Argentina, July 25–31, 2015). IJCAI’15. AAAI Press, 2015, pp. 2327–2333. URL: www.ijcai.org/Proceedings/15/Papers/329.pdf.
- [112] DISSANAYAKE, M. W. M. G. AND PHAN-THIEN, N. Neural-network-based approximations for solving partial differential equations. *Commun. Numer. Methods Engrg.* 10, 3 (1994), pp. 195–201. URL: doi.org/10.1002/cnm.1640100303.
- [113] DOERSCH, C. Tutorial on Variational Autoencoders. *arXiv:1606.05908* (2016), 23 pp. URL: arxiv.org/abs/1606.05908.
- [114] DONAHUE, J., HENDRICKS, L. A., ROHRBACH, M., VENUGOPALAN, S., GUADARRAMA, S., SAENKO, K., AND DARRELL, T. Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 4 (2017), pp. 677–691. URL: doi.org/10.1109/TPAMI.2016.2599174.
- [115] DOS SANTOS, C. AND GATTI, M. Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers* (Dublin, Ireland, Aug. 23–29, 2014). Dublin City University and Association for Computational Linguistics, 2014, pp. 69–78. URL: aclanthology.org/C14-1008.
- [116] DOSOVITSKIY, A., BEYER, L., KOLESNIKOV, A., WEISSENBORN, D., ZHAI, X., UNTERTHINER, T., DEHGHANI, M., MINDERER, M., HEIGOLD, G., GELLY, S., USZKOREIT, J., AND HOULSBY, N. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv:2010.11929* (2020), 22 pp. URL: arxiv.org/abs/2010.11929.
- [117] DOZAT, T. *Incorporating Nesterov momentum into Adam*. <https://openreview.net/forum?id=OM0jvwB8jIp57ZJjtNEZ>. [Accessed 6-December-2017]. 2016.
- [118] DOZAT, T. *Incorporating Nesterov momentum into Adam*. http://cs229.stanford.edu/proj2015/054_report.pdf. [Accessed 6-December-2017]. 2016.
- [119] DU, S. AND LEE, J. On the Power of Over-parametrization in Neural Networks with Quadratic Activation. In *Proceedings of the 35th International Conference on Machine Learning* (Stockholm, Sweden, July 10–15, 2018). Ed. by Dy, J. and Krause, A. Vol. 80. Proceedings of Machine Learning Research. PMLR, 2018, pp. 1329–1338. URL: proceedings.mlr.press/v80/du18a.html.
- [120] DU, S., LEE, J., LI, H., WANG, L., AND ZHAI, X. Gradient Descent Finds Global Minima of Deep Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning* (Long Beach, CA, USA, June 9–15, 2019). Ed. by Chaudhuri, K. and Salakhutdinov, R. Vol. 97. Proceedings of Machine Learning Research. PMLR, 2019, pp. 1675–1685. URL: proceedings.mlr.press/v97/du19c.html.

- [121] DU, T., HUANG, Z., AND LI, Y. Approximation and Generalization of DeepONets for Learning Operators Arising from a Class of Singularly Perturbed Problems. *arXiv:2306.16833* (2023), 32 pp. URL: arxiv.org/abs/2306.16833.
- [122] DUCHI, J. *Probability Bounds*. https://stanford.edu/~jduchi/projects/probability_bounds.pdf. [Accessed 27-October-2023].
- [123] DUCHI, J., HAZAN, E., AND SINGER, Y. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *J. Mach. Learn. Res.* 12 (2011), pp. 2121–2159. URL: jmlr.org/papers/v12/duchi11a.html.
- [124] DUMOULIN, V., BELGHASI, I., POOLE, B., MASTROPIETRO, O., LAMB, A., ARJOVSKY, M., AND COURVILLE, A. Adversarially Learned Inference. *arXiv:1606.00704* (2016), 18 pp. URL: arxiv.org/abs/1606.00704.
- [125] E, W., HAN, J., AND JENTZEN, A. Deep learning-based numerical methods for high-dimensional parabolic partial differential equations and backward stochastic differential equations. *Commun. Math. Stat.* 5, 4 (2017), pp. 349–380. URL: doi.org/10.1007/s40304-017-0117-6.
- [126] E, W., HAN, J., AND JENTZEN, A. Algorithms for solving high dimensional PDEs: from nonlinear Monte Carlo to machine learning. *Nonlinearity* 35, 1 (2021), p. 278. URL: doi.org/10.1088/1361-6544/ac337f.
- [127] E, W., MA, C., AND WU, L. The Barron space and the flow-induced function spaces for neural network models. *Constr. Approx.* 55, 1 (2022), pp. 369–406. URL: doi.org/10.1007/s00365-021-09549-y.
- [128] E, W., MA, C., WU, L., AND WOJTOWYTSCH, S. Towards a Mathematical Understanding of Neural Network-Based Machine Learning: What We Know and What We Don’t. *CSIAM Trans. Appl. Math.* 1, 4 (2020), pp. 561–615. URL: doi.org/10.4208/csiatam.S0-2020-0002.
- [129] E, W. AND WOJTOWYTSCH, S. Some observations on high-dimensional partial differential equations with Barron data. In *Proceedings of the 2nd Mathematical and Scientific Machine Learning Conference* (Aug. 16–19, 2021). Ed. by Bruna, J., Hesthaven, J., and Zdeborova, L. Vol. 145. Proceedings of Machine Learning Research. PMLR, 2022, pp. 253–269. URL: proceedings.mlr.press/v145/e22a.html.
- [130] E, W. AND YU, B. The deep Ritz method: a deep learning-based numerical algorithm for solving variational problems. *Commun. Math. Stat.* 6, 1 (2018), pp. 1–12. URL: doi.org/10.1007/s40304-018-0127-z.
- [131] EBERLE, S., JENTZEN, A., RIEKERT, A., AND WEISS, G. Normalized gradient flow optimization in the training of ReLU artificial neural networks. *arXiv:2207.06246* (2022), 26 pp. URL: arxiv.org/abs/2207.06246.

- [132] EBERLE, S., JENTZEN, A., RIEKERT, A., AND WEISS, G. S. Existence, uniqueness, and convergence rates for gradient flows in the training of artificial neural networks with ReLU activation. *Electron. Res. Arch.* 31, 5 (2023), pp. 2519–2554. URL: doi.org/10.3934/era.2023128.
- [133] EINSIEDLER, M. AND WARD, T. *Functional analysis, spectral theory, and applications*. Vol. 276. Springer, Cham, 2017, xiv+614 pp. URL: doi.org/10.1007/978-3-319-58540-6.
- [134] ELBRÄCHTER, D., GROHS, P., JENTZEN, A., AND SCHWAB, C. DNN expression rate analysis of high-dimensional PDEs: application to option pricing. *Constr. Approx.* 55, 1 (2022), pp. 3–71. URL: doi.org/10.1007/s00365-021-09541-6.
- [135] *Encyclopedia of Mathematics: Lojasiewicz inequality*. https://encyclopediaofmath.org/wiki/Lojasiewicz_inequality. [Accessed 28-August-2023].
- [136] FABBRI, M. AND MORO, G. Dow Jones Trading with Deep Learning: The Unreasonable Effectiveness of Recurrent Neural Networks. In *Proceedings of the 7th International Conference on Data Science, Technology and Applications* (Porto, Portugal, July 26–28, 2018). Ed. by Bernardino, J. and Quix, C. SciTePress - Science and Technology Publications, 2018. URL: doi.org/10.5220/0006922101420153.
- [137] FAN, J., MA, C., AND ZHONG, Y. A selective overview of deep learning. *Statist. Sci.* 36, 2 (2021), pp. 264–290. URL: doi.org/10.1214/20-sts783.
- [138] FEHRMAN, B., GESS, B., AND JENTZEN, A. Convergence Rates for the Stochastic Gradient Descent Method for Non-Convex Objective Functions. *J. Mach. Learn. Res.* 21, 136 (2020), pp. 1–48. URL: jmlr.org/papers/v21/19-636.html.
- [139] FISCHER, T. AND KRAUSS, C. Deep learning with long short-term memory networks for financial market predictions. *European J. Oper. Res.* 270, 2 (2018), pp. 654–669. URL: doi.org/10.1016/j.ejor.2017.11.054.
- [140] FRAENKEL, L. E. Formulae for high derivatives of composite functions. *Math. Proc. Cambridge Philos. Soc.* 83, 2 (1978), pp. 159–165. URL: doi.org/10.1017/S0305004100054402.
- [141] FRESCA, S., DEDE', L., AND MANZONI, A. A comprehensive deep learning-based approach to reduced order modeling of nonlinear time-dependent parametrized PDEs. *J. Sci. Comput.* 87, 2 (2021), Art. No. 61, 36 pp. URL: doi.org/10.1007/s10915-021-01462-7.
- [142] FRESCA, S. AND MANZONI, A. POD-DL-ROM: enhancing deep learning-based reduced order models for nonlinear parametrized PDEs by proper orthogonal decomposition. *Comput. Methods Appl. Mech. Engrg.* 388 (2022), Art. No. 114181, 27 pp. URL: doi.org/10.1016/j.cma.2021.114181.

- [143] FREY, R. AND KÖCK, V. Convergence Analysis of the Deep Splitting Scheme: the Case of Partial Integro-Differential Equations and the associated FBSDEs with Jumps. *arXiv:2206.01597* (2022), 21 pp. URL: arxiv.org/abs/2206.01597.
- [144] FREY, R. AND KÖCK, V. Deep Neural Network Algorithms for Parabolic PIDEs and Applications in Insurance and Finance. *Computation* 10, 11 (2022). URL: doi.org/10.3390/computation10110201.
- [145] FRIEDRICHHS, K. O. Symmetric positive linear differential equations. *Comm. Pure Appl. Math.* 11 (1958), pp. 333–418. URL: doi.org/10.1002/cpa.3160110306.
- [146] FUJII, M., TAKAHASHI, A., AND TAKAHASHI, M. Asymptotic Expansion as Prior Knowledge in Deep Learning Method for High dimensional BSDEs. *Asia-Pacific Financial Markets* 26, 3 (2019), pp. 391–408. URL: doi.org/10.1007/s10690-019-09271-7.
- [147] FUKUMIZU, K. AND AMARI, S. Local minima and plateaus in hierarchical structures of multilayer perceptrons. *Neural Networks* 13, 3 (2000), pp. 317–327. URL: [doi.org/10.1016/S0893-6080\(00\)00009-5](https://doi.org/10.1016/S0893-6080(00)00009-5).
- [148] GALLON, D., JENTZEN, A., AND LINDNER, F. Blow up phenomena for gradient descent optimization methods in the training of artificial neural networks. *arXiv:2211.15641* (2022), 84 pp. URL: arxiv.org/abs/2211.15641.
- [149] GARRIGOS, G. AND GOWER, R. M. Handbook of Convergence Theorems for (Stochastic) Gradient Methods. *arXiv:2301.11235* (2023). URL: doi.org/10.48550/ARXIV.2301.11235.
- [150] GEHRING, J., AULI, M., GRANGIER, D., YARATS, D., AND DAUPHIN, Y. N. Convolutional Sequence to Sequence Learning. In *Proceedings of the 34th International Conference on Machine Learning* (Sydney, Australia, Aug. 6–11, 2017). Ed. by Precup, D. and Teh, Y. W. Vol. 70. Proceedings of Machine Learning Research. PMLR, 2017, pp. 1243–1252. URL: proceedings.mlr.press/v70/gehring17a.html.
- [151] GENTILE, R. AND WELPER, G. Approximation results for Gradient Descent trained Shallow Neural Networks in 1d. *arXiv:2209.08399* (2022), 49 pp. URL: arxiv.org/abs/2209.08399.
- [152] GERMAIN, M., PHAM, H., AND WARIN, X. Neural networks-based algorithms for stochastic control and PDEs in finance. *arXiv:2101.08068* (2021), 27 pp. URL: arxiv.org/abs/2101.08068.
- [153] GERMAIN, M., PHAM, H., AND WARIN, X. Approximation error analysis of some deep backward schemes for nonlinear PDEs. *SIAM J. Sci. Comput.* 44, 1 (2022), A28–A56. URL: doi.org/10.1137/20M1355355.

- [154] GERS, F. A., SCHMIDHUBER, J., AND CUMMINS, F. Learning to Forget: Continual Prediction with LSTM. *Neural Comput.* 12, 10 (2000), pp. 2451–2471. URL: doi.org/10.1162/089976600300015015.
- [155] GERS, F. A., SCHRAUDOLPH, N. N., AND SCHMIDHUBER, J. Learning precise timing with LSTM recurrent networks. *J. Mach. Learn. Res.* 3, 1 (2003), pp. 115–143. URL: doi.org/10.1162/153244303768966139.
- [156] GESS, B., KASSING, S., AND KONAROVSKYI, V. Stochastic Modified Flows, Mean-Field Limits and Dynamics of Stochastic Gradient Descent. *arXiv:2302.07125* (2023), 24 pp. URL: arxiv.org/abs/2302.07125.
- [157] GILES, M. B., JENTZEN, A., AND WELTI, T. Generalised multilevel Picard approximations. *arXiv:1911.03188* (2019), 61 pp. URL: arxiv.org/abs/1911.03188.
- [158] GILMER, J., SCHOENHOLZ, S. S., RILEY, P. F., VINYALS, O., AND DAHL, G. E. Neural Message Passing for Quantum Chemistry. In *Proceedings of the 34th International Conference on Machine Learning* (Sydney, Australia, Aug. 6–11, 2017). Ed. by Precup, D. and Teh, Y. W. Vol. 70. Proceedings of Machine Learning Research. PMLR, 2017, pp. 1263–1272. URL: proceedings.mlr.press/v70/gilmer17a.html.
- [159] GIRSHICK, R., DONAHUE, J., DARRELL, T., AND MALIK, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition* (Columbus, OH, USA, June 23–28, 2014). CVPR ’14. IEEE Computer Society, 2014, pp. 580–587. URL: doi.org/10.1109/CVPR.2014.81.
- [160] GLOROT, X. AND BENGIO, Y. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* (Chia Laguna Resort, Sardinia, Italy, May 13–15, 2010). Ed. by Teh, Y. W. and Titterington, M. Vol. 9. Proceedings of Machine Learning Research. PMLR, 2010, pp. 249–256. URL: proceedings.mlr.press/v9/glorot10a.html.
- [161] GNOATTO, A., PATACCA, M., AND PICARELLI, A. A deep solver for BSDEs with jumps. *arXiv:2211.04349* (2022), 31 pp. URL: arxiv.org/abs/2211.04349.
- [162] GOBET, E. *Monte-Carlo methods and stochastic processes*. From linear to non-linear. CRC Press, Boca Raton, FL, 2016, xxv+309 pp.
- [163] GODICHON-BAGGIONI, A. AND TARRAGO, P. Non asymptotic analysis of Adaptive stochastic gradient algorithms and applications. *arXiv:2303.01370* (2023), 59 pp. URL: arxiv.org/abs/2303.01370.
- [164] GOLDBERG, Y. *Neural Network Methods for Natural Language Processing*. Springer Cham, 2017, xx+292 pp. URL: doi.org/10.1007/978-3-031-02165-7.

- [165] GONON, L. Random Feature Neural Networks Learn Black-Scholes Type PDEs Without Curse of Dimensionality. *J. Mach. Learn. Res.* 24, 189 (2023), pp. 1–51. URL: jmlr.org/papers/v24/21-0987.html.
- [166] GONON, L., GRAEBER, R., AND JENTZEN, A. The necessity of depth for artificial neural networks to approximate certain classes of smooth and bounded functions without the curse of dimensionality. *arXiv:2301.08284* (2023), 101 pp. URL: arxiv.org/abs/2301.08284.
- [167] GONON, L., GRIGORYEVA, L., AND ORTEGA, J.-P. Approximation bounds for random neural networks and reservoir systems. *Ann. Appl. Probab.* 33, 1 (2023), pp. 28–69. URL: doi.org/10.1214/22-aap1806.
- [168] GONON, L., GROHS, P., JENTZEN, A., KOFLER, D., AND ŠIŠKA, D. Uniform error estimates for artificial neural network approximations for heat equations. *IMA J. Numer. Anal.* 42, 3 (2022), pp. 1991–2054. URL: doi.org/10.1093/imanum/drab027.
- [169] GONON, L. AND SCHWAB, C. Deep ReLU network expression rates for option prices in high-dimensional, exponential Lévy models. *Finance Stoch.* 25, 4 (2021), pp. 615–657. URL: doi.org/10.1007/s00780-021-00462-7.
- [170] GONON, L. AND SCHWAB, C. Deep ReLU neural networks overcome the curse of dimensionality for partial integrodifferential equations. *Anal. Appl. (Singap.)* 21, 1 (2023), pp. 1–47. URL: doi.org/10.1142/S0219530522500129.
- [171] GOODFELLOW, I., BENGIO, Y., AND COURVILLE, A. *Deep learning*. MIT Press, Cambridge, MA, 2016, xxii+775 pp. URL: www.deeplearningbook.org/.
- [172] GOODFELLOW, I. J., POUGET-ABADIE, J., MIRZA, M., XU, B., WARDE-FARLEY, D., OZAIR, S., COURVILLE, A., AND BENGIO, Y. Generative Adversarial Networks. *arXiv:1406.2661* (2014), 9 pp. URL: arxiv.org/abs/1406.2661.
- [173] GORI, M., MONFARDINI, G., AND SCARSELLI, F. A new model for learning in graph domains. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005*. Vol. 2. 2005, 729–734 vol. 2. URL: doi.org/10.1109/IJCNN.2005.1555942.
- [174] GOSWAMI, S., JAGTAP, A. D., BABAE, H., SUSI, B. T., AND KARNIADAKIS, G. E. Learning stiff chemical kinetics using extended deep neural operators. *arXiv:2302.12645* (2023), 21 pp. URL: arxiv.org/abs/2302.12645.
- [175] GRAHAM, C. AND TALAY, D. *Stochastic simulation and Monte Carlo methods*. Vol. 68. Mathematical foundations of stochastic simulation. Springer, Heidelberg, 2013, xvi+260 pp. URL: doi.org/10.1007/978-3-642-39363-1.
- [176] GRAVES, A. Generating Sequences With Recurrent Neural Networks. *arXiv:1308.0850* (2013), 43 pp. URL: arxiv.org/abs/1308.0850.

- [177] GRAVES, A. AND JAITLY, N. Towards End-To-End Speech Recognition with Recurrent Neural Networks. In *Proceedings of the 31st International Conference on Machine Learning* (Beijing, China, June 22–24, 2014). Ed. by Xing, E. P. and Jebara, T. Vol. 32. Proceedings of Machine Learning Research 2. PMLR, 2014, pp. 1764–1772. URL: proceedings.mlr.press/v32/graves14.html.
- [178] GRAVES, A., LIWICKI, M., FERNÁNDEZ, S., BERTOLAMI, R., BUNKE, H., AND SCHMIDHUBER, J. A Novel Connectionist System for Unconstrained Handwriting Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 31, 5 (2009), pp. 855–868. URL: doi.org/10.1109/TPAMI.2008.137.
- [179] GRAVES, A., MOHAMED, A.-R., AND HINTON, G. E. Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (Vancouver, BC, Canada, May 26–31, 2013). 2013, pp. 6645–6649. URL: doi.org/10.1109/ICASSP.2013.6638947.
- [180] GRAVES, A. AND SCHMIDHUBER, J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks* 18, 5 (2005). IJCNN 2005, pp. 602–610. URL: doi.org/10.1016/j.neunet.2005.06.042.
- [181] GREFF, K., SRIVASTAVA, R. K., KOUTNÍK, J., STEUNEBRINK, B. R., AND SCHMIDHUBER, J. LSTM: A Search Space Odyssey. *IEEE Trans. Neural Netw. Learn. Syst.* 28, 10 (2017), pp. 2222–2232. URL: doi.org/10.1109/TNNLS.2016.2582924.
- [182] GRIBONVAL, R., KUTYNIOK, G., NIELSEN, M., AND VOIGTLAENDER, F. Approximation spaces of deep neural networks. *Constr. Approx.* 55, 1 (2022), pp. 259–367. URL: doi.org/10.1007/s00365-021-09543-4.
- [183] GRIEWANK, A. AND WALTHER, A. *Evaluating Derivatives*. 2nd ed. Society for Industrial and Applied Mathematics, 2008. URL: doi.org/10.1137/1.9780898717761.
- [184] GROHS, P. AND HERRMANN, L. Deep neural network approximation for high-dimensional elliptic PDEs with boundary conditions. *IMA J. Numer. Anal.* 42, 3 (May 2021), pp. 2055–2082. URL: doi.org/10.1093/imanum/drab031.
- [185] GROHS, P. AND HERRMANN, L. Deep neural network approximation for high-dimensional parabolic Hamilton-Jacobi-Bellman equations. *arXiv:2103.05744* (2021), 23 pp. URL: arxiv.org/abs/2103.05744.
- [186] GROHS, P., HORNUNG, F., JENTZEN, A., AND VON WURSTEMBERGER, P. A proof that artificial neural networks overcome the curse of dimensionality in the numerical approximation of Black-Scholes partial differential equations. *Mem. Amer. Math. Soc.* 284, 1410 (2023), v+93 pp. URL: doi.org/10.1090/memo/1410.

- [187] GROHS, P., HORNUNG, F., JENTZEN, A., AND ZIMMERMANN, P. Space-time error estimates for deep neural network approximations for differential equations. *Adv. Comput. Math.* 49, 1 (2023), Art. No. 4, 78 pp. URL: doi.org/10.1007/s10444-022-09970-2.
- [188] GROHS, P., JENTZEN, A., AND SALIMOVA, D. Deep neural network approximations for solutions of PDEs based on Monte Carlo algorithms. *Partial Differ. Equ. Appl.* 3, 4 (2022), Art. No. 45, 41 pp. URL: doi.org/10.1007/s42985-021-00100-z.
- [189] Grohs, P. and Kutyniok, G., eds. *Mathematical aspects of deep learning*. Cambridge University Press, Cambridge, 2023, xviii+473 pp. URL: doi.org/10.1016/j.enganabound.2022.10.033.
- [190] GROHS, P. AND VOIGTLAENDER, F. Proof of the Theory-to-Practice Gap in Deep Learning via Sampling Complexity bounds for Neural Network Approximation Spaces. *arXiv:2104.02746* (2021). URL: doi.org/10.48550/ARXIV.2104.02746.
- [191] GU, Y., YANG, H., AND ZHOU, C. SelectNet: Self-paced learning for high-dimensional partial differential equations. *J. Comput. Phys.* 441 (2021), p. 110444. URL: doi.org/10.1016/j.jcp.2021.110444.
- [192] GÜHRING, I., KUTYNIOK, G., AND PETERSEN, P. Error bounds for approximations with deep ReLU neural networks in $W^{s,p}$ norms. *Anal. Appl. (Singap.)* 18, 5 (2020), pp. 803–859. URL: doi.org/10.1142/S0219530519410021.
- [193] GUO, X., LI, W., AND IORIO, F. Convolutional Neural Networks for Steady Flow Approximation. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA, Aug. 13–17, 2016). KDD ’16. New York, NY, USA: Association for Computing Machinery, 2016, pp. 481–490. URL: doi.org/10.1145/2939672.2939738.
- [194] GUPTA, V., KOREN, T., AND SINGER, Y. Shampoo: Preconditioned Stochastic Tensor Optimization. *arXiv:1802.09568* (2018). URL: doi.org/10.48550/ARXIV.1802.09568.
- [195] HAN, J. AND E, W. Deep Learning Approximation for Stochastic Control Problems. *arXiv:1611.07422* (2016), 9 pp. URL: arxiv.org/abs/1611.07422.
- [196] HAN, J., JENTZEN, A., AND E, W. Solving high-dimensional partial differential equations using deep learning. *Proc. Natl. Acad. Sci. USA* 115, 34 (2018), pp. 8505–8510. URL: doi.org/10.1073/pnas.1718942115.
- [197] HAN, J. AND LONG, J. Convergence of the deep BSDE method for coupled FBSDEs. *Probab. Uncertain. Quant. Risk* 5 (2020), Art. No. 5, 33 pp. URL: doi.org/10.1186/s41546-020-00047-w.

- [198] HANNIBAL, S., JENTZEN, A., AND THANG, D. M. Non-convergence to global minimizers in data driven supervised deep learning: Adam and stochastic gradient descent optimization provably fail to converge to global minimizers in the training of deep neural networks with ReLU activation. *arXiv:2410.10533* (2024). URL: doi.org/10.48550/ARXIV.2410.10533.
- [199] HASTIE, T., TIBSHIRANI, R., AND FRIEDMAN, J. *The elements of statistical learning*. 2nd ed. Data mining, inference, and prediction. Springer, New York, 2009, xxii+745 pp. URL: doi.org/10.1007/978-0-387-84858-7.
- [200] HE, K., ZHANG, X., REN, S., AND SUN, J. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Las Vegas, NV, USA, June 27–30, 2016). 2016, pp. 770–778. URL: doi.org/10.1109/CVPR.2016.90.
- [201] HE, K., ZHANG, X., REN, S., AND SUN, J. Identity Mappings in Deep Residual Networks. In *Computer Vision – ECCV 2016, 14th European Conference, Proceedings Part IV* (Amsterdam, The Netherlands, Oct. 11–14, 2016). Ed. by Leibe, B., Matas, J., Sebe, N., and Welling, M. Springer, Cham, 2016, pp. 630–645. URL: doi.org/10.1007/978-3-319-46493-0_38.
- [202] HEISS, C., GÜHRING, I., AND EIGEL, M. Multilevel CNNs for Parametric PDEs. *arXiv:2304.00388* (2023), 42 pp. URL: arxiv.org/abs/2304.00388.
- [203] HENDRYCKS, D. AND GIMPEL, K. Gaussian Error Linear Units (GELUs). *arXiv:1606.08415v4* (2016), 10 pp. URL: arxiv.org/abs/1606.08415.
- [204] HENRY, D. *Geometric theory of semilinear parabolic equations*. Vol. 840. Springer-Verlag, Berlin, 1981, iv+348 pp.
- [205] HENRY-LABORDERE, P. Counterparty Risk Valuation: A Marked Branching Diffusion Approach. *arXiv:1203.2369* (2012), 17 pp. URL: arxiv.org/abs/1203.2369.
- [206] HENRY-LABORDERE, P. Deep Primal-Dual Algorithm for BSDEs: Applications of Machine Learning to CVA and IM (2017). Available at SSRN. URL: doi.org/10.2139/ssrn.3071506.
- [207] HENRY-LABORDÈRE, P. AND TOUZI, N. Branching diffusion representation for nonlinear Cauchy problems and Monte Carlo approximation. *Ann. Appl. Probab.* 31, 5 (2021), pp. 2350–2375. URL: doi.org/10.1214/20-aap1649.
- [208] HINTON, G. E. AND SALAKHUTDINOV, R. R. Reducing the dimensionality of data with neural networks. *Science* 313, 5786 (2006), pp. 504–507. URL: doi.org/10.1126/science.1127647.

- [209] HINTON, G., SRIVASTAVA, N., AND SWERSKY, K. *Lecture 6e: RMSprop: Divide the gradient by a running average of its recent magnitude.* https://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf. [Accessed 01-December-2017].
- [210] HINTON, G. E. AND ZEMEL, R. Autoencoders, Minimum Description Length and Helmholtz Free Energy. In *Advances in Neural Information Processing Systems*. Ed. by Cowan, J., Tesauro, G., and Alspector, J. Vol. 6. Morgan-Kaufmann, 1993. URL: proceedings.neurips.cc/paper_files/paper/1993/file/9e3cf48eccf81a0d57663e129aef3cb-Paper.pdf.
- [211] HOCHREITER, S. AND SCHMIDHUBER, J. Long Short-Term Memory. *Neural Comput.* 9, 8 (1997), pp. 1735–1780. URL: doi.org/10.1162/neco.1997.9.8.1735.
- [212] HORN, R. A. AND JOHNSON, C. R. *Matrix analysis*. Second. Cambridge University Press, Cambridge, 2013, pp. xviii+643.
- [213] HORNIK, K. Some new results on neural network approximation. *Neural Networks* 6, 8 (1993), pp. 1069–1072. URL: [doi.org/10.1016/S0893-6080\(09\)80018-X](https://doi.org/10.1016/S0893-6080(09)80018-X).
- [214] HORNIK, K. Approximation capabilities of multilayer feedforward networks. *Neural Networks* 4, 2 (1991), pp. 251–257. URL: [doi.org/10.1016/0893-6080\(91\)90009-T](https://doi.org/10.1016/0893-6080(91)90009-T).
- [215] HORNIK, K., STINCHCOMBE, M., AND WHITE, H. Multilayer feedforward networks are universal approximators. *Neural Networks* 2, 5 (1989), pp. 359–366. URL: [doi.org/10.1016/0893-6080\(89\)90020-8](https://doi.org/10.1016/0893-6080(89)90020-8).
- [216] HORNUNG, F., JENTZEN, A., AND SALIMOVA, D. Space-time deep neural network approximations for high-dimensional partial differential equations. *arXiv:2006.02199* (2020), 52 pages. URL: arxiv.org/abs/2006.02199.
- [217] HUANG, G., LIU, Z., MAATEN, L. V. D., AND WEINBERGER, K. Q. Densely Connected Convolutional Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Honolulu, HI, USA, July 21–26, 2017). Los Alamitos, CA, USA: IEEE Computer Society, 2017, pp. 2261–2269. URL: doi.org/10.1109/CVPR.2017.243.
- [218] HURÉ, C., PHAM, H., AND WARIN, X. Deep backward schemes for high-dimensional nonlinear PDEs. *Math. Comp.* 89, 324 (2020), pp. 1547–1579. URL: doi.org/10.1090/mcom/3514.
- [219] HUTZENTHALER, M., JENTZEN, A., AND KRUSE, T. Overcoming the curse of dimensionality in the numerical approximation of parabolic partial differential equations with gradient-dependent nonlinearities. *Found. Comput. Math.* 22, 4 (2022), pp. 905–966. URL: doi.org/10.1007/s10208-021-09514-y.

- [220] HUTZENTHALER, M., JENTZEN, A., KRUSE, T., AND NGUYEN, T. A. A proof that rectified deep neural networks overcome the curse of dimensionality in the numerical approximation of semilinear heat equations. *SN Partial Differ. Equ. Appl.* 10, 1 (2020). URL: doi.org/10.1007/s42985-019-0006-9.
- [221] HUTZENTHALER, M., JENTZEN, A., KRUSE, T., AND NGUYEN, T. A. Multilevel Picard approximations for high-dimensional semilinear second-order PDEs with Lipschitz nonlinearities. *arXiv:2009.02484* (2020), 37 pp. URL: arxiv.org/abs/2009.02484.
- [222] HUTZENTHALER, M., JENTZEN, A., KRUSE, T., AND NGUYEN, T. A. Overcoming the curse of dimensionality in the numerical approximation of backward stochastic differential equations. *arXiv:2108.10602* (2021), 34 pp. URL: arxiv.org/abs/2108.10602.
- [223] HUTZENTHALER, M., JENTZEN, A., KRUSE, T., NGUYEN, T. A., AND VON WURSTEMBERGER, P. Overcoming the curse of dimensionality in the numerical approximation of semilinear parabolic partial differential equations. *Proc. A.* 476, 2244 (2020), Art. No. 20190630, 25 pp. URL: doi.org/10.1098/rspa.2019.0630.
- [224] HUTZENTHALER, M., JENTZEN, A., POHL, K., RIEKERT, A., AND SCARPA, L. Convergence proof for stochastic gradient descent in the training of deep neural networks with ReLU activation for constant target functions. *arXiv:2112.07369* (2021), 71 pp. URL: arxiv.org/abs/2112.07369.
- [225] HUTZENTHALER, M., JENTZEN, A., AND VON WURSTEMBERGER, P. Overcoming the curse of dimensionality in the approximative pricing of financial derivatives with default risks. *Electron. J. Probab.* 25 (2020), Art. No. 101, 73 pp. URL: doi.org/10.1214/20-ejp423.
- [226] IBRAGIMOV, S., JENTZEN, A., KRÖGER, T., AND RIEKERT, A. On the existence of infinitely many realization functions of non-global local minima in the training of artificial neural networks with ReLU activation. *arXiv:2202.11481* (2022), 49 pp. URL: arxiv.org/abs/2202.11481.
- [227] IBRAGIMOV, S., JENTZEN, A., AND RIEKERT, A. Convergence to good non-optimal critical points in the training of neural networks: Gradient descent optimization with one random initialization overcomes all bad non-global local minima with high probability. *arXiv:2212.13111* (2022), 98 pp. URL: arxiv.org/abs/2212.13111.
- [228] IOFFE, S. AND SZEGEDY, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proceedings of the 32nd International Conference on Machine Learning – Volume 37* (Lille, France, July 6–11, 2015). Ed. by Bach, F. and Blei, D. ICML’15. JMLR.org, 2015, pp. 448–456.

- [229] JACOT, A., GABRIEL, F., AND HONGLER, C. Neural Tangent Kernel: Convergence and Generalization in Neural Networks. In *Advances in Neural Information Processing Systems*. Ed. by Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. Vol. 31. Curran Associates, Inc., 2018. URL: proceedings.neurips.cc/paper_files/paper/2018/file/5a4be1fa34e62bb8a6ec6b91d2462f5a-Paper.pdf.
- [230] JAGTAP, A. D., KHARAZMI, E., AND KARNIADAKIS, G. E. Conservative physics-informed neural networks on discrete domains for conservation laws: Applications to forward and inverse problems. *Comput. Methods Appl. Mech. Engrg.* 365 (2020), p. 113028. URL: doi.org/10.1016/j.cma.2020.113028.
- [231] JENTZEN, A., KUCKUCK, B., NEUFELD, A., AND VON WURSTEMBERGER, P. Strong error analysis for stochastic gradient descent optimization algorithms. *arXiv:1801.09324* (2018), 75 pages. URL: arxiv.org/abs/1801.09324.
- [232] JENTZEN, A., KUCKUCK, B., NEUFELD, A., AND VON WURSTEMBERGER, P. Strong error analysis for stochastic gradient descent optimization algorithms. *IMA J. Numer. Anal.* 41, 1 (2020), pp. 455–492. URL: doi.org/10.1093/imanum/drz055.
- [233] JENTZEN, A., MAZZONETTO, S., AND SALIMOVA, D. Existence and uniqueness properties for solutions of a class of Banach space valued evolution equations (2018), 28 pp. URL: arxiv.org/abs/1812.06859.
- [234] JENTZEN, A. AND RIEKERT, A. On the existence of global minima and convergence analyses for gradient descent methods in the training of deep neural networks. *arXiv:2112.09684v1* (2021), 93 pp. URL: arxiv.org/abs/2112.09684v1.
- [235] JENTZEN, A. AND RIEKERT, A. A proof of convergence for the gradient descent optimization method with random initializations in the training of neural networks with ReLU activation for piecewise linear target functions. *J. Mach. Learn. Res.* 23, 260 (2022), pp. 1–50. URL: jmlr.org/papers/v23/21-0962.html.
- [236] JENTZEN, A. AND RIEKERT, A. On the Existence of Global Minima and Convergence Analyses for Gradient Descent Methods in the Training of Deep Neural Networks. *J. Mach. Learn.* 1, 2 (2022), pp. 141–246. URL: doi.org/10.4208/jml.220114a.
- [237] JENTZEN, A. AND RIEKERT, A. Convergence analysis for gradient flows in the training of artificial neural networks with ReLU activation. *J. Math. Anal. Appl.* 517, 2 (2023), Art. No. 126601, 43 pp. URL: doi.org/10.1016/j.jmaa.2022.126601.
- [238] JENTZEN, A. AND RIEKERT, A. Strong Overall Error Analysis for the Training of Artificial Neural Networks Via Random Initializations. *Commun. Math. Stat.* (2023). URL: doi.org/10.1007/s40304-022-00292-9.

- [239] JENTZEN, A. AND RIEKERT, A. Non-convergence to global minimizers for Adam and stochastic gradient descent optimization and constructions of local minimizers in the training of artificial neural networks. *arXiv:2402.05155* (2024). To appear in SIAM/ASA J. Uncertain. Quantif. URL: doi.org/10.48550/ARXIV.2402.05155.
- [240] JENTZEN, A., RIEKERT, A., AND VON WURSTEMBERGER, P. Algorithmically Designed Artificial Neural Networks (ADANNs): Higher order deep operator learning for parametric partial differential equations. *arXiv:2302.03286* (2023), 22 pp. URL: arxiv.org/abs/2302.03286.
- [241] JENTZEN, A., SALIMOVA, D., AND WELTI, T. A proof that deep artificial neural networks overcome the curse of dimensionality in the numerical approximation of Kolmogorov partial differential equations with constant diffusion and nonlinear drift coefficients. *Commun. Math. Sci.* 19, 5 (2021), pp. 1167–1205. URL: doi.org/10.4310/CMS.2021.v19.n5.a1.
- [242] JENTZEN, A. AND VON WURSTEMBERGER, P. Lower error bounds for the stochastic gradient descent optimization algorithm: Sharp convergence rates for slowly and fast decaying learning rates. *J. Complexity* 57 (2020), Art. No. 101438. URL: doi.org/10.1016/j.jco.2019.101438.
- [243] JENTZEN, A. AND WELTI, T. Overall error analysis for the training of deep neural networks via stochastic gradient descent with random initialisation. *Appl. Math. Comput.* 455 (2023), Art. No. 127907, 34 pp. URL: doi.org/10.1016/j.amc.2023.127907.
- [244] JIN, X., CAI, S., LI, H., AND KARNIADAKIS, G. E. NSFnets (Navier-Stokes flow nets): Physics-informed neural networks for the incompressible Navier-Stokes equations. *J. Comput. Phys.* 426 (2021), Art. No. 109951. URL: doi.org/10.1016/j.jcp.2020.109951.
- [245] JORDAN, K., JIN, Y., BOZA, V., YOU, J., CESISTA, F., NEWHOUSE, L., AND BERNSTEIN, J. *Muon: An optimizer for hidden layers in neural networks*. 2024. URL: kellerjordan.github.io/posts/muon/.
- [246] JUMPER, J., EVANS, R., PRITZEL, A., GREEN, T., FIGURNOV, M., RON-NEBERGER, O., TUNYASUVUNAKOOL, K., BATES, R., ŽÍDEK, A., POTAPENKO, A., BRIDGLAND, A., MEYER, C., KOHL, S. A. A., BALLARD, A. J., COWIE, A., ROMERA-PAREDES, B., NIKOLOV, S., JAIN, R., ADLER, J., BACK, T., PETERSEN, S., REIMAN, D., CLANCY, E., ZIELINSKI, M., STEINEGGER, M., PACHOLSKA, M., BERGHAMMER, T., BODENSTEIN, S., SILVER, D., VINYALS, O., SENIOR, A. W., KAVUKCUOGLU, K., KOHLI, P., AND HASSABIS, D. Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 7873 (2021), pp. 583–589. URL: doi.org/10.1038/s41586-021-03819-2.

- [247] KAINEN, P. C., KŮRKOVÁ, V., AND VOGT, A. Best approximation by linear combinations of characteristic functions of half-spaces. *J. Approx. Theory* 122, 2 (2003), pp. 151–159. URL: [doi.org/10.1016/S0021-9045\(03\)00072-8](https://doi.org/10.1016/S0021-9045(03)00072-8).
- [248] KARATZAS, I. AND SHREVE, S. E. *Brownian motion and stochastic calculus*. 2nd ed. Vol. 113. Springer-Verlag, New York, 1991, xxiv+470 pp. URL: doi.org/10.1007/978-1-4612-0949-2.
- [249] KAREVAN, Z. AND SUYKENS, J. A. Transductive LSTM for time-series prediction: An application to weather forecasting. *Neural Networks* 125 (2020), pp. 1–9. URL: doi.org/10.1016/j.neunet.2019.12.030.
- [250] KARIM, F., MAJUMDAR, S., DARABI, H., AND CHEN, S. LSTM Fully Convolutional Networks for Time Series Classification. *IEEE Access* 6 (2018), pp. 1662–1669. URL: doi.org/10.1109/ACCESS.2017.2779939.
- [251] KARNIADAKIS, G. E., KEVREKIDIS, I. G., LU, L., PERDIKARIS, P., WANG, S., AND YANG, L. Physics-informed machine learning. *Nat. Rev. Phys.* 3, 6 (2021), pp. 422–440. URL: doi.org/10.1038/s42254-021-00314-5.
- [252] KARPATHY, A., JOHNSON, J., AND FEI-FEI, L. Visualizing and Understanding Recurrent Networks. *arXiv:1506.02078* (2015), 12 pp. URL: arxiv.org/abs/1506.02078.
- [253] KAWAGUCHI, K. Deep Learning without Poor Local Minima. In *Advances in Neural Information Processing Systems*. Ed. by Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R. Vol. 29. Curran Associates, Inc., 2016. URL: proceedings.neurips.cc/paper_files/paper/2016/file/f2fc990265c712c49d51a18a32b39f0c-Paper.pdf.
- [254] KHAN, S., NASEER, M., HAYAT, M., ZAMIR, S. W., KHAN, F. S., AND SHAH, M. Transformers in Vision: A Survey. *ACM Comput. Surv.* 54, 10s (2022), Art. No. 200, 41 pp. URL: doi.org/10.1145/3505244.
- [255] KHARAZMI, E., ZHANG, Z., AND KARNIADAKIS, G. E. Variational Physics-Informed Neural Networks For Solving Partial Differential Equations. *arXiv:1912.00873* (2019), 24 pp. URL: arxiv.org/abs/1912.00873.
- [256] KHARAZMI, E., ZHANG, Z., AND KARNIADAKIS, G. E. M. *hp*-VPINNs: variational physics-informed neural networks with domain decomposition. *Comput. Methods Appl. Mech. Engrg.* 374 (2021), Art. No. 113547, 25 pp. URL: doi.org/10.1016/j.cma.2020.113547.

- [257] KHODAYI-MEHR, R. AND ZAVLANOS, M. VarNet: Variational Neural Networks for the Solution of Partial Differential Equations. In *Proceedings of the 2nd Conference on Learning for Dynamics and Control* (June 10–11, 2020). Ed. by Bayen, A. M., Jadbabaie, A., Pappas, G., Parrilo, P. A., Recht, B., Tomlin, C., and Zeilinger, M. Vol. 120. Proceedings of Machine Learning Research. PMLR, 2020, pp. 298–307. URL: proceedings.mlr.press/v120/khodayi-mehr20a.html.
- [258] KHOO, Y., LU, J., AND YING, L. Solving parametric PDE problems with artificial neural networks. *European J. Appl. Math.* 32, 3 (2021), pp. 421–435. URL: doi.org/10.1017/S0956792520000182.
- [259] KIM, Y. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Doha, Qatar, Oct. 25–29, 2014). Ed. by Moschitti, A., Pang, B., and Daelemans, W. Association for Computational Linguistics, 2014, pp. 1746–1751. URL: doi.org/10.3115/v1/D14-1181.
- [260] KINGMA, D. P. AND WELLING, M. Auto-Encoding Variational Bayes. *arXiv:1312.6114* (2013), 14 pp. URL: arxiv.org/abs/1312.6114.
- [261] KINGMA, D. P. AND BA, J. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980* (2014), 15 pp. URL: arxiv.org/abs/1412.6980.
- [262] KLENKE, A. *Probability Theory*. 2nd ed. Springer-Verlag London Ltd., 2014. xii+638 pp. URL: doi.org/10.1007/978-1-4471-5361-0.
- [263] KONTOLATI, K., GOSWAMI, S., KARNIADAKIS, G. E., AND SHIELDS, M. D. Learning in latent spaces improves the predictive accuracy of deep neural operators. *arXiv:2304.07599* (2023), 22 pp. URL: arxiv.org/abs/2304.07599.
- [264] KORN, R., KORN, E., AND KROISANDT, G. *Monte Carlo methods and models in finance and insurance*. CRC Press, Boca Raton, FL, 2010, xiv+470 pp. URL: doi.org/10.1201/9781420076196.
- [265] KOVACHKI, N., LANTHALER, S., AND MISHRA, S. On universal approximation and error bounds for Fourier neural operators. *J. Mach. Learn. Res.* 22 (2021), Art. No. 290, 76 pp. URL: jmlr.org/papers/v22/21-0806.html.
- [266] KOVACHKI, N., LI, Z., LIU, B., AZIZZADENESHELI, K., BHATTACHARYA, K., STUART, A., AND ANANDKUMAR, A. Neural Operator: Learning Maps Between Function Spaces With Applications to PDEs. *J. Mach. Learn. Res.* 24 (2023), Art. No. 89, 97 pp. URL: jmlr.org/papers/v24/21-1524.html.
- [267] KRAMER, M. A. Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal* 37, 2 (1991), pp. 233–243. URL: doi.org/10.1002/aic.690370209.

- [268] KRANTZ, S. G. AND PARKS, H. R. *A primer of real analytic functions*. 2nd ed. Birkhäuser Boston, Inc., Boston, MA, 2002, xiv+205 pp. URL: doi.org/10.1007/978-0-8176-8134-0.
- [269] KRATSIOS, A. The universal approximation property: characterization, construction, representation, and existence. *Ann. Math. Artif. Intell.* 89, 5–6 (2021), pp. 435–469. URL: doi.org/10.1007/s10472-020-09723-1.
- [270] KREMSNER, S., STEINICKE, A., AND SZÖLGYENYI, M. A Deep Neural Network Algorithm for Semilinear Elliptic PDEs with Applications in Insurance Mathematics. *Risks* 8, 4 (2020), Art. No. 136, 18 pp. URL: doi.org/10.3390/risks8040136.
- [271] KRIZHEVSKY, A., SUTSKEVER, I., AND HINTON, G. E. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*. Ed. by Pereira, F., Burges, C., Bottou, L., and Weinberger, K. Vol. 25. Curran Associates, Inc., 2012. URL: proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.
- [272] KURDYKA, K., MOSTOWSKI, T., AND PARUSIŃSKI, A. Proof of the gradient conjecture of R. Thom. *Ann. of Math.* (2) 152, 3 (2000), pp. 763–792. URL: doi.org/10.2307/2661354.
- [273] KUTYNIOK, G., PETERSEN, P., RASLAN, M., AND SCHNEIDER, R. A theoretical analysis of deep neural networks and parametric PDEs. *Constr. Approx.* 55, 1 (2022), pp. 73–125. URL: doi.org/10.1007/s00365-021-09551-4.
- [274] LAGARIS, I., LIKAS, A., AND FOTIADIS, D. Artificial neural networks for solving ordinary and partial differential equations. *IEEE Trans. Neural Netw.* 9, 5 (1998), pp. 987–1000. URL: doi.org/10.1109/72.712178.
- [275] LANTHALER, S., MOLINARO, R., HADORN, P., AND MISHRA, S. Nonlinear Reconstruction for Operator Learning of PDEs with Discontinuities. *arXiv:2210.01074* (2022), 40 pp. URL: arxiv.org/abs/2210.01074.
- [276] LECUN, Y., BOSER, B., DENKER, J. S., HENDERSON, D., HOWARD, R. E., HUBBARD, W., AND JACKEL, L. D. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Comput.* 1, 4 (1989), pp. 541–551. URL: doi.org/10.1162/neco.1989.1.4.541.
- [277] LECUN, Y., BENGIO, Y., AND HINTON, G. Deep learning. *Nature* 521 (2015), pp. 436–444. URL: doi.org/10.1038/nature14539.
- [278] LEE, C.-Y., XIE, S., GALLAGHER, P., ZHANG, Z., AND TU, Z. Deeply-Supervised Nets. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics* (San Diego, California, USA, May 9–12, 2015). Ed. by Lebanon, G. and Vishwanathan, S. V. N. Vol. 38. Proceedings of Machine Learning Research. PMLR, 2015, pp. 562–570. URL: proceedings.mlr.press/v38/lee15a.html.

- [279] LEE, J. D., PANAGEAS, I., PILIOURAS, G., SIMCHOWITZ, M., JORDAN, M. I., AND RECHT, B. First-order methods almost always avoid strict saddle points. *Math. Program.* 176, 1–2 (2019), pp. 311–337. URL: doi.org/10.1007/s10107-019-01374-3.
- [280] LEE, J. D., SIMCHOWITZ, M., JORDAN, M. I., AND RECHT, B. Gradient Descent Only Converges to Minimizers. In *29th Annual Conference on Learning Theory* (Columbia University, New York, NY, USA, June 23–26, 2016). Ed. by Feldman, V., Rakhlin, A., and Shamir, O. Vol. 49. Proceedings of Machine Learning Research. PMLR, 2016, pp. 1246–1257. URL: proceedings.mlr.press/v49/lee16.html.
- [281] LEWIS, M., LIU, Y., GOYAL, N., GHAZVININEJAD, M., MOHAMED, A., LEVY, O., STOYANOV, V., AND ZETTELEMOYER, L. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *arXiv:1910.13461* (2019). URL: arxiv.org/abs/1910.13461.
- [282] LI, J. AND HONG, M. A Note on the Convergence of Muon and Further. *arXiv:2502.02900* (2025). URL: doi.org/10.48550/ARXIV.2502.02900.
- [283] LI, K., TANG, K., WU, T., AND LIAO, Q. D3M: A Deep Domain Decomposition Method for Partial Differential Equations. *IEEE Access* 8 (2020), pp. 5283–5294. URL: doi.org/10.1109/ACCESS.2019.2957200.
- [284] LI, Z., LIU, F., YANG, W., PENG, S., AND ZHOU, J. A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects. *IEEE Transactions on Neural Networks and Learning Systems* 33, 12 (2022), pp. 6999–7019. URL: doi.org/10.1109/TNNLS.2021.3084827.
- [285] LI, Z., HUANG, D. Z., LIU, B., AND ANANDKUMAR, A. Fourier Neural Operator with Learned Deformations for PDEs on General Geometries. *arXiv:2207.05209* (2022). URL: arxiv.org/abs/2207.05209.
- [286] LI, Z., KOVACHKI, N., AZIZZADENESHELI, K., LIU, B., BHATTACHARYA, K., STUART, A., AND ANANDKUMAR, A. Neural Operator: Graph Kernel Network for Partial Differential Equations. *arXiv:2003.03485* (2020). URL: arxiv.org/abs/2003.03485.
- [287] LI, Z., KOVACHKI, N., AZIZZADENESHELI, K., LIU, B., BHATTACHARYA, K., STUART, A., AND ANANDKUMAR, A. Fourier Neural Operator for Parametric Partial Differential Equations. In *International Conference on Learning Representations*. 2021. URL: openreview.net/forum?id=c8P9NQVtmn0.
- [288] LI, Z., KOVACHKI, N., AZIZZADENESHELI, K., LIU, B., STUART, A., BHATTACHARYA, K., AND ANANDKUMAR, A. Multipole graph neural operator for parametric partial differential equations. *Advances in Neural Information Processing Systems* 33 (2020), pp. 6755–6766.

- [289] LI, Z., ZHENG, H., KOVACHKI, N., JIN, D., CHEN, H., LIU, B., AZIZZADENESHELI, K., AND ANANDKUMAR, A. Physics-Informed Neural Operator for Learning Partial Differential Equations. *arXiv:2111.03794* (2021). URL: arxiv.org/abs/2111.03794.
- [290] LIAO, Y. AND MING, P. Deep Nitsche Method: Deep Ritz Method with Essential Boundary Conditions. *Commun. Comput. Phys.* 29, 5 (2021), pp. 1365–1384. URL: doi.org/10.4208/cicp.OA-2020-0219.
- [291] LIU, C. AND BELKIN, M. Accelerating SGD with momentum for over-parameterized learning. *arXiv:1810.13395* (2018). URL: arxiv.org/abs/1810.13395.
- [292] LIU, J., SU, J., YAO, X., JIANG, Z., LAI, G., DU, Y., QIN, Y., XU, W., LU, E., YAN, J., CHEN, Y., ZHENG, H., LIU, Y., LIU, S., YIN, B., HE, W., ZHU, H., WANG, Y., WANG, J., DONG, M., ZHANG, Z., KANG, Y., ZHANG, H., XU, X., ZHANG, Y., WU, Y., ZHOU, X., AND YANG, Z. Muon is Scalable for LLM Training. *arXiv:2502.16982* (2025). URL: doi.org/10.48550/ARXIV.2502.16982.
- [293] LIU, L. AND CAI, W. DeepPropNet—A Recursive Deep Propagator Neural Network for Learning Evolution PDE Operators. *arXiv:2202.13429* (2022). URL: arxiv.org/abs/2202.13429.
- [294] LIU, Y., KUTZ, J. N., AND BRUNTON, S. L. Hierarchical deep learning of multiscale differential equation time-steppers. *Philos. Trans. Roy. Soc. A* 380, 2229 (2022), Art. No. 20210200, 17 pp. URL: doi.org/10.1098/rsta.2021.0200.
- [295] LIU, Z., LIN, Y., CAO, Y., HU, H., WEI, Y., ZHANG, Z., LIN, S., AND GUO, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (Montreal, QC, Canada, Oct. 10–17, 2021). IEEE Computer Society, 2021, pp. 10012–10022. URL: doi.org/10.1109/ICCV48922.2021.00986.
- [296] LIU, Z., CAI, W., AND XU, Z.-Q. J. Multi-scale deep neural network (MscaleDNN) for solving Poisson-Boltzmann equation in complex domains. *Commun. Comput. Phys.* 28, 5 (2020), pp. 1970–2001.
- [297] LOIZOU, N. AND RICHTÁRIK, P. Momentum and stochastic momentum for stochastic gradient, Newton, proximal point and subspace descent methods. *Comput. Optim. Appl.* 77, 3 (2020), pp. 653–710. URL: doi.org/10.1007/s10589-020-00220-z.
- [298] ŁOJASIEWICZ, S. *Ensembles semi-analytiques*. Unpublished lecture notes. Institut des Hautes Études Scientifiques, 1964. URL: perso.univ-rennes1.fr/michel.coste/Lojasiewicz.pdf.
- [299] LONG, J., SHELHAMER, E., AND DARRELL, T. Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Boston, MA, USA, June 7–12, 2015). IEEE Computer Society, 2015, pp. 3431–3440. URL: doi.org/10.1109/CVPR.2015.7298965.

- [300] LOSHCHILOV, I. AND HUTTER, F. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*. 2019. URL: openreview.net/forum?id=Bkg6RiCqY7.
- [301] LU, J., BATRA, D., PARIKH, D., AND LEE, S. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *Advances in Neural Information Processing Systems*. Ed. by Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. Vol. 32. Curran Associates, Inc., 2019. URL: proceedings.neurips.cc/paper_files/paper/2019/file/c74d97b01eae257e44aa9d5bade97baf-Paper.pdf.
- [302] LU, L., JIN, P., PANG, G., ZHANG, Z., AND KARNIADAKIS, G. E. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nature Machine Intelligence* 3, 3 (2021), pp. 218–229. URL: doi.org/10.1038/s42256-021-00302-5.
- [303] LU, L., MENG, X., CAI, S., MAO, Z., GOSWAMI, S., ZHANG, Z., AND KARNIADAKIS, G. E. A comprehensive and fair comparison of two neural operators (with practical extensions) based on FAIR data. *Comput. Methods Appl. Mech. Engrg.* 393 (2022), Art. No. 114778. URL: doi.org/10.1016/j.cma.2022.114778.
- [304] LU, L., MENG, X., MAO, Z., AND KARNIADAKIS, G. E. DeepXDE: A Deep Learning Library for Solving Differential Equations. *SIAM Rev.* 63, 1 (2021), pp. 208–228. URL: doi.org/10.1137/19M1274067.
- [305] LU, L., YEONJONG, S., YANHUI, S., AND KARNIADAKIS GEORGE, E. Dying ReLU and Initialization: Theory and Numerical Examples. *Commun. Comput. Phys.* 28, 5 (2020), pp. 1671–1706. URL: doi.org/https://doi.org/10.4208/cicp.OA-2020-0165.
- [306] LUO, X. AND KAREEM, A. Bayesian deep learning with hierarchical prior: Predictions from limited and noisy data. *Structural Safety* 84 (2020), p. 101918. URL: doi.org/10.1016/j.strusafe.2019.101918.
- [307] LUONG, M.-T., PHAM, H., AND MANNING, C. D. Effective Approaches to Attention-based Neural Machine Translation. *arXiv:1508.04025* (2015). URL: arxiv.org/abs/1508.04025.
- [308] MA, C., WU, L., AND E, W. A Qualitative Study of the Dynamic Behavior for Adaptive Gradient Algorithms. *arXiv:2009.06125* (2020). URL: arxiv.org/abs/2009.06125.
- [309] MADAY, Y. AND TURINICI, G. A parareal in time procedure for the control of partial differential equations. *C. R. Math. Acad. Sci. Paris* 335, 4 (2002), pp. 387–392. URL: [doi.org/10.1016/S1631-073X\(02\)02467-6](https://doi.org/10.1016/S1631-073X(02)02467-6).

- [310] MAHENDRAN, A. AND VEDALDI, A. Visualizing deep convolutional neural networks using natural pre-images. *Int. J. Comput. Vis.* 120, 3 (2016), pp. 233–255. URL: doi.org/10.1007/s11263-016-0911-8.
- [311] MAKHZANI, A., SHLENS, J., JAITLEY, N., GOODFELLOW, I., AND FREY, B. Adversarial Autoencoders. *arXiv:1511.05644* (2015). URL: arxiv.org/abs/1511.05644.
- [312] MAO, X., SHEN, C., AND YANG, Y.-B. Image Restoration Using Very Deep Convolutional Encoder-Decoder Networks with Symmetric Skip Connections. In *Advances in Neural Information Processing Systems*. Ed. by Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R. Vol. 29. Curran Associates, Inc., 2016. URL: proceedings.neurips.cc/paper_files/paper/2016/file/0ed9422357395a0d4879191c66f4faa2-Paper.pdf.
- [313] MASCI, J., MEIER, U., CIREŞAN, D., AND SCHMIDHUBER, J. Stacked Convolutional Auto-Encoders for Hierarchical Feature Extraction. In *Artificial Neural Networks and Machine Learning – ICANN 2011* (Espoo, Finland, June 14–17, 2011). Ed. by Honkela, T., Duch, W., Girolami, M., and Kaski, S. Springer Berlin Heidelberg, 2011, pp. 52–59.
- [314] MENG, X., LI, Z., ZHANG, D., AND KARNIADAKIS, G. E. PPINN: Parareal physics-informed neural network for time-dependent PDEs. *Comput. Methods Appl. Mech. Engrg.* 370 (2020), p. 113250. URL: doi.org/10.1016/j.cma.2020.113250.
- [315] MERTIKOPOULOS, P., HALLAK, N., KAVIS, A., AND CEVHER, V. On the Almost Sure Convergence of Stochastic Gradient Descent in Non-Convex Problems. In *Advances in Neural Information Processing Systems*. Ed. by Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. Vol. 33. Curran Associates, Inc., 2020, pp. 1117–1128. URL: proceedings.neurips.cc/paper_files/paper/2020/file/0cb5ebb1b34ec343dfe135db691e4a85-Paper.pdf.
- [316] MEURIS, B., QADEER, S., AND STINIS, P. Machine-learning-based spectral methods for partial differential equations. *Scientific Reports* 13, 1 (2023), p. 1739. URL: doi.org/10.1038/s41598-022-26602-3.
- [317] MISHRA, S. AND MOLINARO, R. Estimates on the generalization error of Physics Informed Neural Networks (PINNs) for approximating a class of inverse problems for PDEs. *arXiv:2007.01138* (2020). URL: arxiv.org/abs/2007.01138.
- [318] MISHRA, S. AND MOLINARO, R. Estimates on the generalization error of Physics Informed Neural Networks (PINNs) for approximating PDEs. *arXiv:2006.16144* (2020). URL: arxiv.org/abs/2006.16144.
- [319] NEAL, R. M. *Bayesian Learning for Neural Networks*. Springer New York, 1996. 204 pp. URL: doi.org/10.1007/978-1-4612-0745-0.

- [320] NELSEN, N. H. AND STUART, A. M. The random feature model for input-output maps between Banach spaces. *SIAM J. Sci. Comput.* 43, 5 (2021), A3212–A3243. URL: doi.org/10.1137/20M133957X.
- [321] NESTEROV, Y. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. In *Soviet Mathematics Doklady*. Vol. 27. 1983, pp. 372–376.
- [322] NESTEROV, Y. *Introductory lectures on convex optimization: A basic course*. Vol. 87. Springer, New York, 2013, xviii+236 pp. URL: doi.org/10.1007/978-1-4419-8853-9.
- [323] NEUFELD, A., NGUYEN, T. A., AND WU, S. Deep ReLU neural networks overcome the curse of dimensionality when approximating semilinear partial integro-differential equations. *arXiv:2310.15581* (2023). URL: arxiv.org/abs/2310.15581.
- [324] NEUFELD, A. AND WU, S. Multilevel Picard approximation algorithm for semilinear partial integro-differential equations and its complexity analysis. *arXiv:2205.09639* (2022). URL: arxiv.org/abs/2205.09639.
- [325] NEUFELD, A. AND WU, S. Multilevel Picard algorithm for general semilinear parabolic PDEs with gradient-dependent nonlinearities. *arXiv:2310.12545* (2023). URL: arxiv.org/abs/2310.12545.
- [326] NG, A. *coursera: Improving Deep Neural Networks: Hyperparameter tuning, Regularization and Optimization*. <https://www.coursera.org/learn/deep-neural-network>. [Accessed 6-December-2017].
- [327] NG, J. Y.-H., HAUSKNECHT, M., VIJAYANARASIMHAN, S., VINYALS, O., MONGA, R., AND TODERICI, G. Beyond Short Snippets: Deep Networks for Video Classification. *arXiv:1503.08909* (2015). URL: arxiv.org/abs/1503.08909.
- [328] NGUWI, J. Y., PENENT, G., AND PRIVAULT, N. A deep branching solver for fully nonlinear partial differential equations. *arXiv:2203.03234* (2022). URL: arxiv.org/abs/2203.03234.
- [329] NGUWI, J. Y., PENENT, G., AND PRIVAULT, N. Numerical solution of the incompressible Navier-Stokes equation by a deep branching algorithm. *arXiv:2212.13010* (2022). URL: arxiv.org/abs/2212.13010.
- [330] NGUWI, J. Y., PENENT, G., AND PRIVAULT, N. A fully nonlinear Feynman-Kac formula with derivatives of arbitrary orders. *J. Evol. Equ.* 23, 1 (2023), Art. No. 22, 29 pp. URL: doi.org/10.1007/s00028-023-00873-3.
- [331] NGUWI, J. Y. AND PRIVAULT, N. Numerical solution of the modified and non-Newtonian Burgers equations by stochastic coded trees. *Jpn. J. Ind. Appl. Math.* 40, 3 (2023), pp. 1745–1763. URL: doi.org/10.1007/s13160-023-00611-9.

- [332] NGUYEN, Q. AND HEIN, M. The Loss Surface of Deep and Wide Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning* (Sydney, Australia, Aug. 6–11, 2017). Ed. by Precup, D. and Teh, Y. W. Vol. 70. Proceedings of Machine Learning Research. PMLR, 2017, pp. 2603–2612. URL: proceedings.mlr.press/v70/nguyen17a.html.
- [333] NITSCHE, J. Über ein Variationsprinzip zur Lösung von Dirichlet-Problemen bei Verwendung von Teilräumen, die keinen Randbedingungen unterworfen sind. *Abh. Math. Sem. Univ. Hamburg* 36 (1971), pp. 9–15. URL: doi.org/10.1007/BF02995904.
- [334] NOVAK, E. AND WOŹNIAKOWSKI, H. *Tractability of multivariate problems. Vol. I: Linear information*. Vol. 6. European Mathematical Society (EMS), Zürich, 2008, xii+384 pp. URL: doi.org/10.4171/026.
- [335] NOVAK, E. AND WOŹNIAKOWSKI, H. *Tractability of multivariate problems. Volume II: Standard information for functionals*. Vol. 12. European Mathematical Society (EMS), Zürich, 2010, xviii+657 pp. URL: doi.org/10.4171/084.
- [336] NOVAK, E. AND WOŹNIAKOWSKI, H. *Tractability of multivariate problems. Volume III: Standard information for operators*. Vol. 18. European Mathematical Society (EMS), Zürich, 2012, xviii+586 pp. URL: doi.org/10.4171/116.
- [337] NÜSKEN, N. AND RICHTER, L. Solving high-dimensional Hamilton-Jacobi-Bellman PDEs using neural networks: perspectives from the theory of controlled diffusions and measures on path space. *Partial Differ. Equ. Appl.* 2, 4 (2021), Art. No. 48, 48 pp. URL: doi.org/10.1007/s42985-021-00102-x.
- [338] ØKSENDAL, B. *Stochastic differential equations*. 6th ed. An introduction with applications. Springer-Verlag, Berlin, 2003, xxiv+360 pp. URL: doi.org/10.1007/978-3-642-14394-6.
- [339] OLAH, C. *Understanding LSTM Networks*. <http://colah.github.io/posts/2015-08-Understanding-LSTMs/>. [Accessed 9-October-2023].
- [340] OPENAI. GPT-4 Technical Report. *arXiv:2303.08774* (2023). URL: arxiv.org/abs/2303.08774.
- [341] OPSCHOOR, J. A. A., PETERSEN, P. C., AND SCHWAB, C. Deep ReLU networks and high-order finite element methods. *Anal. Appl. (Singap.)* 18, 5 (2020), pp. 715–770. URL: doi.org/10.1142/S0219530519410136.
- [342] PANAGEAS, I. AND PILIOURAS, G. Gradient Descent Only Converges to Minimizers: Non-Isolated Critical Points and Invariant Regions. *arXiv:1605.00405* (2016). URL: arxiv.org/abs/1605.00405.

- [343] PANAGEAS, I., PILIOURAS, G., AND WANG, X. First-order methods almost always avoid saddle points: The case of vanishing step-sizes. In *Advances in Neural Information Processing Systems*. Ed. by Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. Vol. 32. Curran Associates, Inc., 2019. URL: proceedings.neurips.cc/paper_files/paper/2019/file/3fb04953d95a94367bb133f862402bce-Paper.pdf.
- [344] PANG, G., LU, L., AND KARNIADAKIS, G. E. fPINNs: Fractional Physics-Informed Neural Networks. *SIAM J. Sci. Comput.* 41, 4 (2019), A2603–A2626. URL: doi.org/10.1137/18M1229845.
- [345] PARDOUX, É. AND PENG, S. Backward stochastic differential equations and quasilinear parabolic partial differential equations. In *Stochastic partial differential equations and their applications*. Vol. 176. Lect. Notes Control Inf. Sci. Springer, Berlin, 1992, pp. 200–217. URL: doi.org/10.1007/BFb0007334.
- [346] PARDOUX, É. AND PENG, S. G. Adapted solution of a backward stochastic differential equation. *Systems Control Lett.* 14, 1 (1990), pp. 55–61. URL: [doi.org/10.1016/0167-6911\(90\)90082-6](https://doi.org/10.1016/0167-6911(90)90082-6).
- [347] PARDOUX, E. AND TANG, S. Forward-backward stochastic differential equations and quasilinear parabolic PDEs. *Probab. Theory Related Fields* 114, 2 (1999), pp. 123–150. URL: doi.org/10.1007/s004409970001.
- [348] PASCANU, R., MIKOLOV, T., AND BENGIO, Y. On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on Machine Learning* (Atlanta, GA, USA, June 17–19, 2013). Ed. by Dasgupta, S. and McAllester, D. Vol. 28. Proceedings of Machine Learning Research 3. PMLR, 2013, pp. 1310–1318. URL: proceedings.mlr.press/v28/pascanu13.html.
- [349] PEREKRESTENKO, D., GROHS, P., ELBRÄCHTER, D., AND BÖLCSKEI, H. The universal approximation power of finite-width deep ReLU networks. *arXiv:1806.01528* (2018). URL: arxiv.org/abs/1806.01528.
- [350] PÉREZ-ORTIZ, J. A., GERS, F. A., ECK, D., AND SCHMIDHUBER, J. Kalman filters improve LSTM network performance in problems unsolvable by traditional recurrent nets. *Neural Networks* 16, 2 (2003), pp. 241–250. URL: [doi.org/10.1016/S0893-6080\(02\)00219-8](https://doi.org/10.1016/S0893-6080(02)00219-8).
- [351] PETERSEN, P. *Linear Algebra*. Springer New York, 2012. x+390 pp. URL: doi.org/10.1007/978-1-4614-3612-6.
- [352] PETERSEN, P., RASLAN, M., AND VOIGTLAENDER, F. Topological properties of the set of functions generated by neural networks of fixed size. *Found. Comput. Math.* 21, 2 (2021), pp. 375–444. URL: doi.org/10.1007/s10208-020-09461-0.

- [353] PETERSEN, P. AND VOIGTLAENDER, F. Optimal approximation of piecewise smooth functions using deep ReLU neural networks. *Neural Networks* 108 (2018), pp. 296–330. URL: doi.org/10.1016/j.neunet.2018.08.019.
- [354] PETERSEN, P. AND VOIGTLAENDER, F. Equivalence of approximation by convolutional neural networks and fully-connected networks. *Proc. Amer. Math. Soc.* 148, 4 (2020), pp. 1567–1581. URL: doi.org/10.1090/proc/14789.
- [355] PHAM, H. AND WARIN, X. Mean-field neural networks: learning mappings on Wasserstein space. *arXiv:2210.15179* (2022). URL: arxiv.org/abs/2210.15179.
- [356] PHAM, H., WARIN, X., AND GERMAIN, M. Neural networks-based backward scheme for fully nonlinear PDEs. *Partial Differ. Equ. Appl.* 2, 1 (2021), Art. No. 16, 24 pp. URL: doi.org/10.1007/s42985-020-00062-8.
- [357] POLYAK, B. T. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics* 4, 5 (1964), pp. 1–17.
- [358] Pytorch: RMSprop. <https://pytorch.org/docs/stable/generated/torch.optim.RMSprop.html>. [Accessed 14-April-2025].
- [359] PyTorch: SGD. <https://pytorch.org/docs/stable/generated/torch.optim.SGD.html>. [Accessed 4-September-2023].
- [360] QIAN, N. On the momentum term in gradient descent learning algorithms. *Neural Networks* 12, 1 (1999), pp. 145–151. URL: [doi.org/10.1016/S0893-6080\(98\)00116-6](https://doi.org/10.1016/S0893-6080(98)00116-6).
- [361] RADFORD, A., JOZEFOWICZ, R., AND SUTSKEVER, I. Learning to Generate Reviews and Discovering Sentiment. *arXiv:1704.01444* (2017). URL: arxiv.org/abs/1704.01444.
- [362] RADFORD, A., NARASIMHAN, K., SALIMANS, T., AND SUTSKEVER, I. Improving language understanding by generative pre-training (2018), 12 pp. URL: openai.com/research/language-unsupervised.
- [363] RADFORD, A., WU, J., CHILD, R., LUAN, D., AMODEI, D., AND SUTSKEVER, I. Language Models are Unsupervised Multitask Learners (2019), 24 pp. URL: openai.com/research/better-language-models.
- [364] RAFFEL, C., SHAZER, N., ROBERTS, A., LEE, K., NARANG, S., MATENA, M., ZHOU, Y., LI, W., AND LIU, P. J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* 21, 140 (2020), pp. 1–67. URL: jmlr.org/papers/v21/20-074.html.
- [365] RAFIQ, M., RAFIQ, G., JUNG, H.-Y., AND CHOI, G. S. SSNO: Spatio-Spectral Neural Operator for Functional Space Learning of Partial Differential Equations. *IEEE Access* 10 (2022), pp. 15084–15095. URL: doi.org/10.1109/ACCESS.2022.3148401.

- [366] RAIKO, T., VALPOLA, H., AND LECUN, Y. Deep Learning Made Easier by Linear Transformations in Perceptrons. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics* (La Palma, Canary Islands, Apr. 21–23, 2012). Ed. by Lawrence, N. D. and Girolami, M. Vol. 22. Proceedings of Machine Learning Research. PMLR, 2012, pp. 924–932. URL: proceedings.mlr.press/v22/raiko12.html.
- [367] RAISSI, M. Forward-Backward Stochastic Neural Networks: Deep Learning of High-dimensional Partial Differential Equations. *arXiv:1804.07010* (2018). URL: arxiv.org/abs/1804.07010.
- [368] RAISSI, M., PERDIKARIS, P., AND KARNIADAKIS, G. E. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.* 378 (2019), pp. 686–707. URL: doi.org/10.1016/j.jcp.2018.10.045.
- [369] RAJPURKAR, P., HANNUN, A. Y., HAGHPANAH, M., BOURN, C., AND NG, A. Y. Cardiologist-Level Arrhythmia Detection with Convolutional Neural Networks. *arXiv:1707.01836* (2017). URL: arxiv.org/abs/1707.01836.
- [370] RANZATO, M., HUANG, F. J., BOUREAU, Y.-L., AND LEcUN, Y. Unsupervised Learning of Invariant Feature Hierarchies with Applications to Object Recognition. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*. 2007, pp. 1–8. URL: doi.org/10.1109/CVPR.2007.383157.
- [371] RAONIĆ, B., MOLINARO, R., RYCK, T. D., ROHNER, T., BARTOLUCCI, F., ALAIFARI, R., MISHRA, S., AND DE BÉZENAC, E. Convolutional Neural Operators for robust and accurate learning of PDEs. *arXiv:2302.01178* (2023). URL: arxiv.org/abs/2302.01178.
- [372] REDDI, S. J., KALE, S., AND KUMAR, S. On the Convergence of Adam and Beyond. *arXiv:1904.09237* (2019). URL: arxiv.org/abs/1904.09237.
- [373] REICHSTEIN, M., CAMPS-VALLS, G., STEVENS, B., JUNG, M., DENZLER, J., CARVALHAIS, N., AND PRABHAT. Deep learning and process understanding for data-driven Earth system science. *Nature* 566, 7743 (2019), pp. 195–204. URL: doi.org/10.1038/s41586-019-0912-1.
- [374] REISINGER, C. AND ZHANG, Y. Rectified deep neural networks overcome the curse of dimensionality for nonsmooth value functions in zero-sum games of nonlinear stiff systems. *Anal. Appl. (Singap.)* 18, 6 (2020), pp. 951–999. URL: doi.org/10.1142/S0219530520500116.
- [375] RUDER, S. An overview of gradient descent optimization algorithms. *arXiv:1609.04747* (2016). URL: arxiv.org/abs/1609.04747.
- [376] RUF, J. AND WANG, W. Neural networks for option pricing and hedging: a literature review. *arXiv:1911.05620* (2019). URL: arxiv.org/abs/1911.05620.

- [377] RUMELHART, D. E., HINTON, G. E., AND WILLIAMS, R. J. Learning Internal Representations by Error Propagation. In. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations*. Cambridge, MA, USA: MIT Press, 1986, pp. 318–362.
- [378] SAFRAN, I. AND SHAMIR, O. On the Quality of the Initial Basin in Overspecified Neural Networks. In *Proceedings of The 33rd International Conference on Machine Learning* (New York, NY, USA, June 20–22, 2016). Vol. 48. Proceedings of Machine Learning Research. PMLR, 2016, pp. 774–782. URL: proceedings.mlr.press/v48/safran16.html.
- [379] SAFRAN, I. AND SHAMIR, O. Spurious Local Minima are Common in Two-Layer ReLU Neural Networks. In *Proceedings of the 35th International Conference on Machine Learning* (Stockholm, Sweden, July 10–15, 2018). Vol. 80. Proceedings of Machine Learning Research. ISSN: 2640-3498. PMLR, 2018, pp. 4433–4441. URL: proceedings.mlr.press/v80/safran18a.html.
- [380] SAINATH, T. N., MOHAMED, A., KINGSBURY, B., AND RAMABHADRAN, B. Deep convolutional neural networks for LVCSR. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (Vancouver, BC, Canada, May 26–31, 2013). IEEE Computer Society, 2013, pp. 8614–8618. URL: doi.org/10.1109/ICASSP.2013.6639347.
- [381] SAK, H., SENIOR, A., AND BEAUFAYS, F. Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition. *arXiv:1402.1128* (2014). URL: arxiv.org/abs/1402.1128.
- [382] SANCHEZ-GONZALEZ, A., GODWIN, J., PFAFF, T., YING, R., LESKOVEC, J., AND BATTAGLIA, P. W. Learning to Simulate Complex Physics with Graph Networks. *arXiv:2002.09405* (Feb. 2020). URL: arxiv.org/abs/2002.09405.
- [383] SANCHEZ-LENGELING, B., REIF, E., PEARCE, A., AND WILTSCHKO, A. B. *A Gentle Introduction to Graph Neural Networks*. <https://distill.pub/2021/gnn-intro/>. [Accessed 10-October-2023].
- [384] SANDBERG, I. Approximation theorems for discrete-time systems. *IEEE Trans. Circuits Syst.* 38, 5 (1991), pp. 564–566. URL: doi.org/10.1109/31.76498.
- [385] SANTURKAR, S., TSIPRAS, D., ILYAS, A., AND MADRY, A. How Does Batch Normalization Help Optimization? In *Advances in Neural Information Processing Systems*. Ed. by Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. Vol. 31. Curran Associates, Inc., 2018. URL: proceedings.neurips.cc/paper_files/paper/2018/file/905056c1ac1dad141560467e0a99e1cf-Paper.pdf.

- [386] SARAO MANNELLI, S., VANDEN-EIJNDEN, E., AND ZDEBOROVÁ, L. Optimization and Generalization of Shallow Neural Networks with Quadratic Activation Functions. In *Advances in Neural Information Processing Systems*. Ed. by Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. Vol. 33. Curran Associates, Inc., 2020, pp. 13445–13455. URL: proceedings.neurips.cc/paper_files/paper/2020/file/9b8b50fb590c590ffbf1295ce92258dc-Paper.pdf.
- [387] SCARSELLI, F., GORI, M., TSOI, A. C., HAGENBUCHNER, M., AND MONFARDINI, G. The Graph Neural Network Model. *IEEE Trans. Neural Netw.* 20, 1 (2009), pp. 61–80. URL: doi.org/10.1109/TNN.2008.2005605.
- [388] SCHMIDHUBER, J. Deep learning in neural networks: An overview. *Neural Networks* 61 (2015), pp. 85–117. URL: doi.org/10.1016/j.neunet.2014.09.003.
- [389] SCHÜTT, K. T., SAUCEDA, H. E., KINDERMANS, P.-J., TKATCHENKO, A., AND MÜLLER, K.-R. SchNet – A deep learning architecture for molecules and materials. *The Journal of Chemical Physics* 148, 24 (2018). URL: doi.org/10.1063/1.5019779.
- [390] SCHWAB, C., STEIN, A., AND ZECH, J. Deep Operator Network Approximation Rates for Lipschitz Operators. *arXiv:2307.09835* (2023). URL: arxiv.org/abs/2307.09835.
- [391] SCHWAB, C. AND ZECH, J. Deep learning in high dimension: neural network expression rates for generalized polynomial chaos expansions in UQ. *Anal. Appl. (Singap.)* 17, 1 (2019), pp. 19–55. URL: doi.org/10.1142/S0219530518500203.
- [392] SERMANET, P., EIGEN, D., ZHANG, X., MATHIEU, M., FERGUS, R., AND LECUN, Y. OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks. *arXiv:1312.6229* (2013). URL: arxiv.org/abs/1312.6229.
- [393] SEZER, O. B., GUDELEK, M. U., AND OZBAYOGLU, A. M. Financial time series forecasting with deep learning : A systematic literature review: 2005–2019. *Appl. Soft Comput.* 90 (2020), Art. No. 106181. URL: doi.org/10.1016/j.asoc.2020.106181.
- [394] SHALEV-SHWARTZ, S. AND BEN-DAVID, S. *Understanding Machine Learning. From Theory to Algorithms*. Cambridge University Press, 2014, xvi+397 pp. URL: doi.org/10.1017/CBO9781107298019.
- [395] SHEN, Z., YANG, H., AND ZHANG, S. Deep network approximation characterized by number of neurons. *Commun. Comput. Phys.* 28, 5 (2020), pp. 1768–1811. URL: doi.org/10.4208/cicp.oa-2020-0149.
- [396] SHI, X., CHEN, Z., WANG, H., YEUNG, D.-Y., WONG, W.-K., AND WOO, W.-C. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. In *Advances in Neural Information Processing Systems*. Ed. by Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R. Vol. 28. Curran Associates, Inc., 2015. URL: proceedings.neurips.cc/paper_files/paper/2015/file/07563a3fe3bbe7e3ba84431ad9d055af-Paper.pdf.

- [397] SIAMI-NAMINI, S., TAVAKOLI, N., AND SIAMI NAMIN, A. A Comparison of ARIMA and LSTM in Forecasting Time Series. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)* (Orlando, FL, USA, Dec. 17–20, 2018). IEEE Computer Society, 2018, pp. 1394–1401. URL: doi.org/10.1109/ICMLA.2018.00227.
- [398] SILVESTER, J. R. Determinants of block matrices. *Math. Gaz.* 84, 501 (2000), pp. 460–467. URL: doi.org/10.2307/3620776.
- [399] SIMONYAN, K. AND ZISSERMAN, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556* (2014). URL: arxiv.org/abs/1409.1556.
- [400] SIRIGNANO, J. AND SPILIOPOULOS, K. DGM: A deep learning algorithm for solving partial differential equations. *J. Comput. Phys.* 375 (2018), pp. 1339–1364. URL: doi.org/10.1016/j.jcp.2018.08.029.
- [401] SITZMANN, V., MARTEL, J. N. P., BERGMAN, A. W., LINDELL, D. B., AND WETZSTEIN, G. Implicit Neural Representations with Periodic Activation Functions. *arXiv:2006.09661* (2020). URL: arxiv.org/abs/2006.09661.
- [402] SOLTANOLKOTABI, M., JAVANMARD, A., AND LEE, J. D. Theoretical Insights Into the Optimization Landscape of Over-Parameterized Shallow Neural Networks. *IEEE Trans. Inform. Theory* 65, 2 (2019), pp. 742–769. URL: doi.org/10.1109/TIT.2018.2854560.
- [403] SOUDRY, D. AND CARMON, Y. No bad local minima: Data independent training error guarantees for multilayer neural networks. *arXiv:1605.08361* (2016). URL: arxiv.org/abs/1605.08361.
- [404] SOUDRY, D. AND HOFFER, E. Exponentially vanishing sub-optimal local minima in multilayer neural networks. *arXiv:1702.05777* (2017). URL: arxiv.org/abs/1702.05777.
- [405] SRIVASTAVA, R. K., GREFF, K., AND SCHMIDHUBER, J. Training Very Deep Networks. In *Advances in Neural Information Processing Systems*. Ed. by Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R. Vol. 28. Curran Associates, Inc., 2015. URL: proceedings.neurips.cc/paper_files/paper/2015/file/215a71a12769b056c3c32e7299f1c5ed-Paper.pdf.
- [406] SRIVASTAVA, R. K., GREFF, K., AND SCHMIDHUBER, J. Highway Networks. *arXiv:1505.00387* (2015). URL: arxiv.org/abs/1505.00387.
- [407] SUN, R. Optimization for deep learning: theory and algorithms. *arXiv:1912.08957* (Dec. 2019). URL: arxiv.org/abs/1912.08957.

- [408] SUTSKEVER, I., MARTENS, J., DAHL, G., AND HINTON, G. On the importance of initialization and momentum in deep learning. In *Proceedings of the 30th International Conference on Machine Learning* (Atlanta, GA, USA, June 17–19, 2013). Ed. by Dasgupta, S. and McAllester, D. Vol. 28. Proceedings of Machine Learning Research 3. PMLR, 2013, pp. 1139–1147. URL: proceedings.mlr.press/v28/sutskever13.html.
- [409] SUTSKEVER, I., VINYALS, O., AND LE, Q. V. Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems*. Ed. by Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K. Vol. 27. Curran Associates, Inc., 2014. URL: proceedings.neurips.cc/paper_files/paper/2014/file/a14ac55a4f27472c5d894ec1c3c743d2-Paper.pdf.
- [410] SUTTON, R. S. AND BARTO, A. G. *Reinforcement Learning: An Introduction*. 2nd ed. MIT Press, Cambridge, MA, 2018, xxii+526 pp.
- [411] SZEGEDY, C., LIU, W., JIA, Y., SERMANET, P., REED, S., ANGUELOV, D., ERHAN, D., VANHOUCKE, V., AND RABINOVICH, A. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Boston, MA, USA, June 7–12, 2015). IEEE Computer Society, 2015, pp. 1–9. URL: doi.org/10.1109/CVPR.2015.7298594.
- [412] TADIĆ, V. B. Convergence and convergence rate of stochastic gradient search in the case of multiple and non-isolated extrema. *Stochastic Process. Appl.* 125, 5 (2015), pp. 1715–1755. URL: doi.org/10.1016/j.spa.2014.11.001.
- [413] TAN, L. AND CHEN, L. Enhanced DeepONet for modeling partial differential operators considering multiple input functions. *arXiv:2202.08942* (2022). URL: arxiv.org/abs/2202.08942.
- [414] TAYLOR, J. M., PARDO, D., AND MUGA, I. A deep Fourier residual method for solving PDEs using neural networks. *Comput. Methods Appl. Mech. Engrg.* 405 (2023), Art. No. 115850, 27 pp. URL: doi.org/10.1016/j.cma.2022.115850.
- [415] TESCHL, G. *Ordinary differential equations and dynamical systems*. Vol. 140. American Mathematical Society, Providence, RI, 2012, xii+356 pp. URL: doi.org/10.1090/gsm/140.
- [416] TROPP, J. A. *An Elementary Proof of the Spectral Radius Formula for Matrices*. <http://users.cms.caltech.edu/~jtropp/notes/Tro01-Spectral-Radius.pdf>. [Accessed 16-February-2018]. 2001.
- [417] VAN DEN OORD, A., DIELEMAN, S., AND SCHRAUWEN, B. Deep content-based music recommendation. In *Advances in Neural Information Processing Systems*. Ed. by Burges, C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Vol. 26. Curran Associates, Inc., 2013. URL: proceedings.neurips.cc/paper_files/paper/2013/file/b3ba8f1bee1238a2f37603d90b58898d-Paper.pdf.

- [418] VASWANI, A., SHAZER, N., PARMAR, N., USZKOREIT, J., JONES, L., GOMEZ, A. N., KAISER, Ł., AND POLOSUKHIN, I. Attention is All you Need. In *Advances in Neural Information Processing Systems*. Ed. by Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. Vol. 30. Curran Associates, Inc., 2017. URL: proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fdb053c1c4a845aa-Paper.pdf.
- [419] VATANEN, T., RAIKO, T., VALPOLA, H., AND LECUN, Y. Pushing Stochastic Gradient towards Second-Order Methods – Backpropagation Learning with Transformations in Nonlinearities. In *Neural Information Processing*. Ed. by Lee, M., Hirose, A., Hou, Z.-G., and Kil, R. M. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 442–449.
- [420] VELIČKOVIĆ, P., CUCURULL, G., CASANOVA, A., ROMERO, A., LIÒ, P., AND BENGIO, Y. Graph Attention Networks. *arXiv:1710.10903* (2017). URL: arxiv.org/abs/1710.10903.
- [421] VENTURI, L., BANDEIRA, A. S., AND BRUNA, J. Spurious Valleys in One-hidden-layer Neural Network Optimization Landscapes. *J. Mach. Learn. Res.* 20, 133 (2019), pp. 1–34. URL: jmlr.org/papers/v20/18-674.html.
- [422] VENUGOPALAN, S., ROHRBACH, M., DONAHUE, J., MOONEY, R., DARRELL, T., AND SAENKO, K. Sequence to Sequence – Video to Text. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (Santiago, Chile, Dec. 7–13, 2015). IEEE Computer Society, 2015. URL: doi.org/10.1109/ICCV.2015.515.
- [423] VINCENT, P., LAROCHELLE, H., BENGIO, Y., AND MANZAGOL, P.-A. Extracting and Composing Robust Features with Denoising Autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*. ICML ’08. Helsinki, Finland: Association for Computing Machinery, 2008, pp. 1096–1103. URL: doi.org/10.1145/1390156.1390294.
- [424] VINCENT, P., LAROCHELLE, H., LAJOIE, I., BENGIO, Y., AND MANZAGOL, P.-A. Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. *J. Mach. Learn. Res.* 11, 110 (2010), pp. 3371–3408. URL: jmlr.org/papers/v11/vincent10a.html.
- [425] WANG, F., JIANG, M., QIAN, C., YANG, S., LI, C., ZHANG, H., WANG, X., AND TANG, X. Residual Attention Network for Image Classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Honolulu, HI, USA, July 21–26, 2017). IEEE Computer Society, 2017. URL: doi.org/10.1109/CVPR.2017.683.
- [426] WANG, N., ZHANG, D., CHANG, H., AND LI, H. Deep learning of subsurface flow via theory-guided neural network. *J. Hydrology* 584 (2020), p. 124700. URL: doi.org/10.1016/j.jhydrol.2020.124700.

- [427] WANG, S., WANG, H., AND PERDIKARIS, P. Learning the solution operator of parametric partial differential equations with physics-informed DeepONets. *Science Advances* 7, 40 (2021), eabi8605. URL: doi.org/10.1126/sciadv.abi8605.
- [428] WANG, Y., ZOU, R., LIU, F., ZHANG, L., AND LIU, Q. A review of wind speed and wind power forecasting with deep neural networks. *Appl. Energy* 304 (2021), Art. No. 117766. URL: doi.org/10.1016/j.apenergy.2021.117766.
- [429] WANG, Z., YAN, W., AND OATES, T. Time series classification from scratch with deep neural networks: A strong baseline. In *2017 International Joint Conference on Neural Networks (IJCNN)*. 2017, pp. 1578–1585. URL: doi.org/10.1109/IJCNN.2017.7966039.
- [430] WELPER, G. Approximation Results for Gradient Descent trained Neural Networks. *arXiv:2309.04860* (2023). URL: arxiv.org/abs/2309.04860.
- [431] WEN, G., LI, Z., AZIZZADENESHELI, K., ANANDKUMAR, A., AND BENSON, S. M. U-FNO – An enhanced Fourier neural operator-based deep-learning model for multiphase flow. *arXiv:2109.03697* (2021). URL: arxiv.org/abs/2109.03697.
- [432] WEST, D. *Introduction to Graph Theory*. Prentice Hall, 2001. 588 pp.
- [433] WU, F., SOUZA, A., ZHANG, T., FIFTY, C., YU, T., AND WEINBERGER, K. Simplifying Graph Convolutional Networks. In *Proceedings of the 36th International Conference on Machine Learning* (Long Beach, California, USA, June 9–15, 2019). Ed. by Chaudhuri, K. and Salakhutdinov, R. Vol. 97. Proceedings of Machine Learning Research. PMLR, 2019, pp. 6861–6871. URL: proceedings.mlr.press/v97/wu19e.html.
- [434] WU, K., YAN, X.-B., JIN, S., AND MA, Z. Asymptotic-Preserving Convolutional DeepONets Capture the Diffusive Behavior of the Multiscale Linear Transport Equations. *arXiv:2306.15891* (2023). URL: arxiv.org/abs/2306.15891.
- [435] WU, Z., RAMSUNDAR, B., FEINBERG, E. N., GOMES, J., GENIESSE, C., PAPPU, A. S., LESWING, K., AND PANDE, V. MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.* 9 (2 2018), pp. 513–530. URL: doi.org/10.1039/C7SC02664A.
- [436] WU, Z., PAN, S., CHEN, F., LONG, G., ZHANG, C., AND YU, P. S. A Comprehensive Survey on Graph Neural Networks. *IEEE Trans. Neural Netw. Learn. Syst.* 32, 1 (2021), pp. 4–24. URL: doi.org/10.1109/TNNLS.2020.2978386.
- [437] XIE, J., XU, L., AND CHEN, E. Image Denoising and Inpainting with Deep Neural Networks. In *Advances in Neural Information Processing Systems*. Ed. by Pereira, F., Burges, C., Bottou, L., and Weinberger, K. Vol. 25. Curran Associates, Inc., 2012. URL: proceedings.neurips.cc/paper_files/paper/2012/file/6cdd60ea0045eb7a6ec44c54d29ed402-Paper.pdf.

- [438] XIE, S., GIRSHICK, R., DOLLÁR, P., TU, Z., AND HE, K. Aggregated Residual Transformations for Deep Neural Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Honolulu, HI, USA, July 21–26, 2017). IEEE Computer Society, 2017, pp. 5987–5995. URL: doi.org/10.1109/CVPR.2017.634.
- [439] XIONG, R., YANG, Y., HE, D., ZHENG, K., ZHENG, S., XING, C., ZHANG, H., LAN, Y., WANG, L., AND LIU, T.-Y. On Layer Normalization in the Transformer Architecture. In *Proceedings of the 37th International Conference on Machine Learning* (July 13–18, 2020). ICML’20. JMLR.org, 2020, 975, pp. 10524–10533. URL: proceedings.mlr.press/v119/xiong20b.html.
- [440] XIONG, W., HUANG, X., ZHANG, Z., DENG, R., SUN, P., AND TIAN, Y. Koopman neural operator as a mesh-free solver of non-linear partial differential equations. *arXiv:2301.10022* (2023). URL: arxiv.org/abs/2301.10022.
- [441] XU, R., ZHANG, D., RONG, M., AND WANG, N. Weak form theory-guided neural network (TgNN-wf) for deep learning of subsurface single- and two-phase flow. *J. Comput. Phys.* 436 (2021), Art. No. 110318, 20 pp. URL: doi.org/10.1016/j.jcp.2021.110318.
- [442] YANG, L., MENG, X., AND KARNIADAKIS, G. E. B-PINNs: Bayesian physics-informed neural networks for forward and inverse PDE problems with noisy data. *J. Comput. Phys.* 425 (2021), Art. No. 109913. URL: doi.org/10.1016/j.jcp.2020.109913.
- [443] YANG, Z., DAI, Z., YANG, Y., CARBONELL, J., SALAKHUTDINOV, R., AND LE, Q. V. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *arXiv:1906.08237* (2019). URL: arxiv.org/abs/1906.08237.
- [444] YAROTSKY, D. Error bounds for approximations with deep ReLU networks. *Neural Networks* 94 (2017), pp. 103–114. URL: doi.org/10.1016/j.neunet.2017.07.002.
- [445] YING, R., HE, R., CHEN, K., EKSOMBATCHAI, P., HAMILTON, W. L., AND LESKOVEC, J. Graph Convolutional Neural Networks for Web-Scale Recommender Systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (London, United Kingdom, Aug. 19–23, 2018). KDD ’18. New York, NY, USA: Association for Computing Machinery, 2018, pp. 974–983. URL: doi.org/10.1145/3219819.3219890.
- [446] YU, Y., SI, X., HU, C., AND ZHANG, J. A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures. *Neural Comput.* 31, 7 (July 2019), pp. 1235–1270. URL: doi.org/10.1162/neco_a_01199.

- [447] YUN, S., JEONG, M., KIM, R., KANG, J., AND KIM, H. J. Graph Transformer Networks. In *Advances in Neural Information Processing Systems*. Ed. by Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. Vol. 32. Curran Associates, Inc., 2019. URL: proceedings.neurips.cc/paper_files/paper/2019/file/9d63484abb477c97640154d40595a3bb-Paper.pdf.
- [448] ZAGORUYKO, S. AND KOMODAKIS, N. Wide Residual Networks. *arXiv:1605.07146* (2016). URL: arxiv.org/abs/1605.07146.
- [449] ZANG, Y., BAO, G., YE, X., AND ZHOU, H. Weak adversarial networks for high-dimensional partial differential equations. *J. Comput. Phys.* 411 (2020), pp. 109409, 14. URL: doi.org/10.1016/j.jcp.2020.109409.
- [450] ZEILER, M. D. ADADELTA: An Adaptive Learning Rate Method. *arXiv:1212.5701* (2012). URL: arxiv.org/abs/1212.5701.
- [451] ZENG, D., LIU, K., LAI, S., ZHOU, G., AND ZHAO, J. Relation Classification via Convolutional Deep Neural Network. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin, Ireland: Dublin City University and Association for Computational Linguistics, Aug. 2014, pp. 2335–2344. URL: aclanthology.org/C14-1220.
- [452] ZHANG, A., LIPTON, Z. C., LI, M., AND SMOLA, A. J. *Dive into Deep Learning*. Cambridge University Press, 2023. URL: d2l.ai.
- [453] ZHANG, J., ZHANG, S., SHEN, J., AND LIN, G. Energy-Dissipative Evolutionary Deep Operator Neural Networks. *arXiv:2306.06281* (2023). URL: arxiv.org/abs/2306.06281.
- [454] ZHANG, J., MOKHTARI, A., SRA, S., AND JADBABAIE, A. Direct Runge-Kutta Discretization Achieves Acceleration. *arXiv:1805.00521* (2018). URL: arxiv.org/abs/1805.00521.
- [455] ZHANG, X., ZHAO, J., AND LECUN, Y. Character-level Convolutional Networks for Text Classification. In *Advances in Neural Information Processing Systems*. Ed. by Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R. Vol. 28. Curran Associates, Inc., 2015. URL: proceedings.neurips.cc/paper_files/paper/2015/file/250cf8b51c773f3f8dc8b4be867a9a02-Paper.pdf.
- [456] ZHANG, Y., LI, Y., ZHANG, Z., LUO, T., AND XU, Z.-Q. J. Embedding Principle: a hierarchical structure of loss landscape of deep neural networks. *arXiv:2111.15527* (2021). URL: arxiv.org/abs/2111.15527.
- [457] ZHANG, Y., ZHANG, Z., LUO, T., AND XU, Z.-Q. J. Embedding Principle of Loss Landscape of Deep Neural Networks. *arXiv:2105.14573* (2021). URL: arxiv.org/abs/2105.14573.

- [458] ZHANG, Y. AND WALLACE, B. A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (Taipei, Taiwan, Nov. 27–Dec. 1, 2017). Asian Federation of Natural Language Processing, 2017, pp. 253–263. URL: aclanthology.org/I17-1026.
- [459] ZHANG, Y., CHEN, C., SHI, N., SUN, R., AND LUO, Z.-Q. Adam Can Converge Without Any Modification On Update Rules. *arXiv:2208.09632* (2022). URL: arxiv.org/abs/2208.09632.
- [460] ZHANG, Z., CUI, P., AND ZHU, W. Deep Learning on Graphs: A Survey. *IEEE Trans. Knowledge Data Engrg.* 34, 1 (2022), pp. 249–270. URL: doi.org/10.1109/TKDE.2020.2981333.
- [461] ZHENG, Y., LIU, Q., CHEN, E., GE, Y., AND ZHAO, J. L. Time Series Classification Using Multi-Channels Deep Convolutional Neural Networks. In *Web-Age Information Management*. Ed. by Li, F., Li, G., Hwang, S.-w., Yao, B., and Zhang, Z. Springer, Cham, 2014, pp. 298–310. URL: doi.org/10.1007/978-3-319-08010-9_33.
- [462] ZHOU, H., ZHANG, S., PENG, J., ZHANG, S., LI, J., XIONG, H., AND ZHANG, W. Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 12 (2021), pp. 11106–11115. URL: doi.org/10.1609/aaai.v35i12.17325.
- [463] ZHOU, J., CUI, G., HU, S., ZHANG, Z., YANG, C., LIU, Z., WANG, L., LI, C., AND SUN, M. Graph neural networks: A review of methods and applications. *AI Open* 1 (2020), pp. 57–81. URL: doi.org/10.1016/j.aiopen.2021.01.001.
- [464] ZHU, Y. AND ZABARAS, N. Bayesian deep convolutional encoder-decoder networks for surrogate modeling and uncertainty quantification. *J. Comput. Phys.* 366 (2018), pp. 415–447. URL: doi.org/10.1016/j.jcp.2018.04.018.