

Contents

Preface	3
Introduction	15
I Artificial neural networks (ANNs)	19
1 Basics on ANNs	21
1.1 Fully-connected feedforward ANNs (vectorized description)	21
1.1.1 Affine functions	23
1.1.2 Vectorized description of fully-connected feedforward ANNs	23
1.1.3 Weight and bias parameters of fully-connected feedforward ANNs	25
1.2 Activation functions	26
1.2.1 Multidimensional versions	27
1.2.2 Single hidden layer fully-connected feedforward ANNs	28
1.2.3 Rectified linear unit (ReLU) activation	29
1.2.4 Clipping activation	34
1.2.5 Softplus activation	35
1.2.6 Gaussian error linear unit (GELU) activation	37
1.2.7 Standard logistic activation	38
1.2.8 Swish activation	40
1.2.9 Hyperbolic tangent activation	42
1.2.10 Softsign activation	43
1.2.11 Leaky rectified linear unit (leaky ReLU) activation	44
1.2.12 Exponential linear unit (ELU) activation	46
1.2.13 Rectified power unit (RePU) activation	47
1.2.14 Sine activation	49
1.2.15 Heaviside activation	49
1.2.16 Softmax activation	51
1.3 Fully-connected feedforward ANNs (structured description)	51
1.3.1 Structured description of fully-connected feedforward ANNs	52
1.3.2 Realizations of fully-connected feedforward ANNs	53

1.3.3	On the connection to the vectorized description	57
1.4	Convolutional ANNs (CNNs)	59
1.4.1	Discrete convolutions	60
1.4.2	Structured description of feedforward CNNs	60
1.4.3	Realizations of feedforward CNNs	60
1.5	Residual ANNs (ResNets)	66
1.5.1	Structured description of fully-connected ResNets	66
1.5.2	Realizations of fully-connected ResNets	67
1.6	Recurrent ANNs (RNNs)	70
1.6.1	Description of RNNs	70
1.6.2	Vectorized description of simple fully-connected RNNs	71
1.6.3	Long short-term memory (LSTM) RNNs	72
1.7	Further types of ANNs	72
1.7.1	ANNs with encoder-decoder architectures: autoencoders	73
1.7.2	Transformers and the attention mechanism	73
1.7.3	Graph neural networks (GNNs)	74
1.7.4	Neural operators	75
2	ANN calculus	77
2.1	Compositions of fully-connected feedforward ANNs	77
2.1.1	Compositions of fully-connected feedforward ANNs	77
2.1.2	Elementary properties of compositions of fully-connected feedforward ANNs	78
2.1.3	Associativity of compositions of fully-connected feedforward ANNs	80
2.1.4	Powers of fully-connected feedforward ANNs	84
2.2	Parallelizations of fully-connected feedforward ANNs	84
2.2.1	Parallelizations of fully-connected feedforward ANNs with the same length	84
2.2.2	Representations of the identities with ReLU activation functions	89
2.2.3	Extensions of fully-connected feedforward ANNs	90
2.2.4	Parallelizations of fully-connected feedforward ANNs with different lengths	94
2.3	Scalar multiplications of fully-connected feedforward ANNs	96
2.3.1	Affine transformations as fully-connected feedforward ANNs	96
2.3.2	Scalar multiplications of fully-connected feedforward ANNs	97
2.4	Sums of fully-connected feedforward ANNs with the same length	98
2.4.1	Sums of vectors as fully-connected feedforward ANNs	98
2.4.2	Concatenation of vectors as fully-connected feedforward ANNs	100
2.4.3	Sums of fully-connected feedforward ANNs	102

II	Approximation	105
3	One-dimensional ANN approximation results	107
3.1	Linear interpolation of one-dimensional functions	107
3.1.1	On the modulus of continuity	107
3.1.2	Linear interpolation of one-dimensional functions	109
3.2	Linear interpolation with fully-connected feedforward ANNs	113
3.2.1	Activation functions as fully-connected feedforward ANNs	113
3.2.2	Representations for ReLU ANNs with one hidden neuron	114
3.2.3	ReLU ANN representations for linear interpolations	115
3.3	ANN approximations results for one-dimensional functions	118
3.3.1	Constructive ANN approximation results	118
3.3.2	Convergence rates for the approximation error	122
4	Multi-dimensional ANN approximation results	127
4.1	Approximations through supremal convolutions	127
4.2	ANN representations	130
4.2.1	ANN representations for the 1-norm	130
4.2.2	ANN representations for maxima	132
4.2.3	ANN representations for maximum convolutions	137
4.3	ANN approximations results for multi-dimensional functions	141
4.3.1	Constructive ANN approximation results	141
4.3.2	Covering number estimates	141
4.3.3	Convergence rates for the approximation error	143
4.4	Refined ANN approximations results for multi-dimensional functions	152
4.4.1	Rectified clipped ANNs	152
4.4.2	Embedding ANNs in larger architectures	153
4.4.3	Approximation through ANNs with variable architectures	160
4.4.4	Refined convergence rates for the approximation error	162
III	Optimization	169
5	Optimization through gradient flow (GF) trajectories	171
5.1	Introductory comments for the training of ANNs	171
5.2	Basics for GFs	173
5.2.1	GF ordinary differential equations (ODEs)	173
5.2.2	Direction of negative gradients	174
5.3	Regularity properties for ANNs	180
5.3.1	On the differentiability of compositions of parametric functions	180
5.3.2	On the differentiability of realizations of ANNs	181

5.4	Loss functions	183
5.4.1	Absolute error loss	183
5.4.2	Mean squared error loss	184
5.4.3	Huber error loss	186
5.4.4	Cross-entropy loss	188
5.4.5	Kullback–Leibler divergence loss	192
5.5	GF optimization in the training of ANNs	195
5.6	Lyapunov-type functions for GFs	197
5.6.1	Gronwall differential inequalities	197
5.6.2	Lyapunov-type functions for ODEs	198
5.6.3	On Lyapunov-type functions and coercivity-type conditions	199
5.6.4	Sufficient and necessary conditions for local minimum points	200
5.6.5	On a linear growth condition	203
5.7	Optimization through flows of ODEs	203
5.7.1	Approximation of local minimum points through GFs	203
5.7.2	Existence and uniqueness of solutions of ODEs	206
5.7.3	Approximation of local minimum points through GFs revisited	208
5.7.4	Approximation error with respect to the objective function	210
6	Deterministic gradient descent (GD) optimization methods	211
6.1	GD optimization	211
6.1.1	GD optimization in the training of ANNs	212
6.1.2	Euler discretizations for GF ODEs	213
6.1.3	Lyapunov-type stability for GD optimization	215
6.1.4	Error analysis for GD optimization	219
6.2	Explicit midpoint GD optimization	239
6.2.1	Explicit midpoint discretizations for GF ODEs	239
6.3	GD optimization with classical momentum	242
6.3.1	Representations for GD optimization with momentum	244
6.3.2	Bias-adjusted GD optimization with momentum	247
6.3.3	Error analysis for GD optimization with momentum	249
6.3.4	Numerical comparisons for GD optimization with and without momentum	264
6.4	GD optimization with Nesterov momentum	269
6.5	Adagrad GD optimization (Adagrad)	269
6.6	Root mean square propagation GD optimization (RMSprop)	270
6.6.1	Representations of the mean square terms in RMSprop	271
6.6.2	Bias-adjusted root mean square propagation GD optimization	272
6.7	Adadelta GD optimization	274
6.8	Adaptive moment estimation GD optimization (Adam)	275

7	Stochastic gradient descent (SGD) optimization methods	277
7.1	Introductory comments for the training of ANNs with SGD	277
7.2	SGD optimization	279
7.2.1	SGD optimization in the training of ANNs	280
7.2.2	Non-convergence of SGD for not appropriately decaying learning rates	288
7.2.3	Convergence rates for SGD for quadratic objective functions	299
7.2.4	Convergence rates for SGD for coercive objective functions	302
7.3	Explicit midpoint SGD optimization	303
7.4	SGD optimization with classical momentum	305
7.4.1	Bias-adjusted SGD optimization with classical momentum	307
7.5	SGD optimization with Nesterov momentum	310
7.5.1	Simplified SGD optimization with Nesterov momentum	312
7.6	Adagrad SGD optimization (Adagrad)	314
7.7	Root mean square propagation SGD optimization (RMSprop)	316
7.7.1	Bias-adjusted root mean square propagation SGD optimization	318
7.8	Adadelta SGD optimization	320
7.9	Adaptive moment estimation SGD optimization (Adam)	322
8	Backpropagation	337
8.1	Backpropagation for parametric functions	337
8.2	Backpropagation for ANNs	342
9	Kurdyka–Łojasiewicz (KL) inequalities	349
9.1	Standard KL functions	349
9.2	Convergence analysis using standard KL functions (regular regime)	350
9.3	Standard KL inequalities for monomials	353
9.4	Standard KL inequalities around non-critical points	353
9.5	Standard KL inequalities with increased exponents	355
9.6	Standard KL inequalities for one-dimensional polynomials	355
9.7	Power series and analytic functions	358
9.8	Standard KL inequalities for one-dimensional analytic functions	360
9.9	Standard KL inequalities for analytic functions	365
9.10	Counterexamples	365
9.11	Convergence analysis for solutions of GF ODEs	368
9.11.1	Abstract local convergence results for GF processes	368
9.11.2	Abstract global convergence results for GF processes	373
9.12	Convergence analysis for GD processes	378
9.12.1	One-step descent property for GD processes	378
9.12.2	Abstract local convergence results for GD processes	380
9.13	On the analyticity of realization functions of ANNs	385

9.14	Standard KL inequalities for empirical risks in the training of ANNs with analytic activation functions	388
9.15	Fréchet subdifferentials and limiting Fréchet subdifferentials	390
9.16	Non-smooth slope	396
9.17	Generalized KL functions	396
10	ANNs with batch normalization	399
10.1	Batch normalization (BN)	399
10.2	Structured descr. of fully-connected feedforward ANNs with BN (training)	402
10.3	Realizations of fully-connected feedforward ANNs with BN (training) . . .	402
10.4	Structured descr. of fully-connected feedforward ANNs with BN (inference)	403
10.5	Realizations of fully-connected feedforward ANNs with BN (inference) . .	403
10.6	On the connection between BN for training and BN for inference	404
11	Optimization through random initializations	407
11.1	Analysis of the optimization error	407
11.1.1	The complementary distribution function formula	407
11.1.2	Estimates for the optimization error involving complementary distribution functions	408
11.2	Strong convergences rates for the optimization error	409
11.2.1	Properties of the gamma and the beta function	409
11.2.2	Product measurability of continuous random fields	414
11.2.3	Strong convergences rates for the optimization error	417
11.3	Strong convergences rates for the optimization error involving ANNs . . .	420
11.3.1	Local Lipschitz continuity estimates for the parametrization functions of ANNs	420
11.3.2	Strong convergences rates for the optimization error involving ANNs	427
IV	Generalization	431
12	Probabilistic generalization error estimates	433
12.1	Concentration inequalities for random variables	433
12.1.1	Markov's inequality	433
12.1.2	A first concentration inequality	434
12.1.3	Moment-generating functions	436
12.1.4	Chernoff bounds	436
12.1.5	Hoeffding's inequality	438
12.1.6	A strengthened Hoeffding's inequality	444
12.2	Covering number estimates	445
12.2.1	Entropy quantities	445

12.2.2	Inequalities for packing entropy quantities in metric spaces	448
12.2.3	Inequalities for covering entropy quantities in metric spaces	450
12.2.4	Inequalities for entropy quantities in finite dimensional vector spaces	452
12.3	Empirical risk minimization	459
12.3.1	Concentration inequalities for random fields	459
12.3.2	Uniform estimates for the statistical learning error	464
13	Strong generalization error estimates	469
13.1	Monte Carlo estimates	469
13.2	Uniform strong error estimates for random fields	472
13.3	Strong convergence rates for the generalisation error	476
V	Composed error analysis	485
14	Overall error decomposition	487
14.1	Bias-variance decomposition	487
14.1.1	Risk minimization for measurable functions	488
14.2	Overall error decomposition	490
15	Composed error estimates	493
15.1	Full strong error analysis for the training of ANNs	493
15.2	Full strong error analysis with optimization via SGD with random initializations	502
VI	Deep learning for partial differential equations (PDEs)	507
16	Physics-informed neural networks (PINNs)	509
16.1	Reformulation of PDE problems as stochastic optimization problems	510
16.2	Derivation of PINNs and deep Galerkin methods (DGMs)	511
16.3	Implementation of PINNs	513
16.4	Implementation of DGMs	516
17	Deep Kolmogorov methods (DKMs)	521
17.1	Stochastic optimization problems for expectations of random variables	522
17.2	Stochastic optimization problems for expectations of random fields	522
17.3	Feynman–Kac formulas	524
17.3.1	Feynman–Kac formulas providing existence of solutions	524
17.3.2	Feynman–Kac formulas providing uniqueness of solutions	529
17.4	Reformulation of PDE problems as stochastic optimization problems	534
17.5	Derivation of DKMs	537
17.6	Implementation of DKMs	539

18 Further deep learning methods for PDEs	543
18.1 Deep learning methods based on strong formulations of PDEs	543
18.2 Deep learning methods based on weak formulations of PDEs	544
18.3 Deep learning methods based on stochastic representations of PDEs	545
18.4 Error analyses for deep learning methods for PDEs	547
Index of abbreviations	549
List of figures	551
List of source codes	553
List of definitions	555
Bibliography	559