Contents

Preface xiii

Notation xvii

Chapter 1. Introduction 1
§1.1. Preliminaries 1
§1.2. Brief Historical Review of Deep Learning 3
§1.3. Overview and Notation 4
§1.4. On the General Task of Machine Learning 6
§1.5. Quick Overview of Supervised Learning 8
§1.6. Bias-Variance Tradeoff and Double Descent 13
§1.7. Some Existing Related Books 17
§1.8. Organization of this Book 17

Part 1. Mathematical Introduction to Deep Learning

Chapter 2. Linear Regression 27
§2.1. Introduction 27
§2.2. Loss Function 29
§2.3. Minimization 30
§2.4. Metric 34
§2.5. Computational Realization 35
§2.6. Brief Concluding Remarks 36
§2.7. Exercises 37

Chapter 3. Logistic Regression 39
§3.1. Introduction 39
§3.2. Formalization of the Problem 41
§3.3. Metric 46
§3.4. Transitions and Scaling 46
§3.5. Normalization 48
§3.6. Perfect Data and Penalization 51
§3.7. Multiclass prediction 53
§3.8. Brief Concluding Remarks 58
§3.9. Exercises 58

Chapter 4. From the Perceptron Model to Kernels to Neural Networks 59
§4.1. Introduction 59
§4.2. Perceptron Model and Stochastic Gradient Descent 60
§4.3. Perceptron Through the Lens of a Kernel 61
§4.4. Linear Regression and Kernels 64
§4.5. From Kernels to Neural Networks 66
§4.6. Brief Concluding Remarks 67

Chapter 5. Feed Forward Neural Networks 69
§5.1. Introduction 69
§5.2. Truth Tables 72
§5.3. Numerical Exploration 84
§5.4. Activation Functions 87
§5.5. Brief Concluding Remarks 90
§5.6. Exercises 91

Chapter 6. Backpropagation 93
§6.1. Introduction 93
§6.2. Introductory Example 94
§6.3. Backpropagation in a More General Case 96
§6.4. Backpropagation for Multilayer Feed Forward Neural Networks 98
§6.5. Backpropagation Applied to a Deep Learning Example 99
§6.6. Vanishing Gradient Problem 102
§6.7. Brief Concluding Remarks 103
§6.8. Exercises 104

Chapter 7. Basics of Stochastic Gradient Descent 105
§7.1. Introduction 105
§7.2. The basic setup 106
§7.3. Stochastic gradient descent algorithm 107
§7.4. Applications to Shallow Neural Networks 114
§7.5. Implementation Examples 120
§7.6. Brief Concluding Remarks 127
§7.7. Exercises 127

Chapter 8. Stochastic Gradient Descent for Multi-layer Networks 129
§8.1. Introduction 129
§8.2. Multi-layer Neural Networks 129
§8.3. Computational Cost 132
§8.4. Vanishing Gradient Problem 133
§8.5. Implementation Example 134
§8.6. Brief Concluding Remarks 135
§8.7. Exercises 136

Chapter 9. Regularization and Dropout 137
§9.1. Introduction 137
§9.2. Regularization by Penalty Terms 137
§9.3. Dropout and its Relation to Regularization 141
§9.4. A Neural Network Example with Dropout Implemented 143
§9.5. Dropout on General Multi-layer Neural Networks 147
§9.6. Brief Concluding Remarks 149
§9.7. Exercises 149

Chapter 10. Batch Normalization 151
§10.1. Introduction 151
§10.2. Batch Normalization Through an Example 152
§10.3. Batch Normalization and Minibatches 157
§10.4. Brief Concluding Remarks 157

Chapter 11. Training, Validation, and Testing 159
§11.1. Introduction 159
§11.2. Polynomials 160
§11.3. Training 160
§11.4. Validation 161

§11.5. Cross-Validation 164
§11.6. Brief Concluding Remarks 167

Chapter 12. Feature Importance 169
§12.1. Introduction 169
§12.2. Feature Permutation 171
§12.3. Shapley Value 173
§12.4. Feature Permutation versus Shapley Value 176
§12.5. Brief Concluding Remarks 177
§12.6. Exercises 177

Chapter 13. Recurrent Neural Networks for Sequential Data 181
§13.1. Introduction 181
§13.2. The Plant-Observer Paradigm 182
§13.3. Jordan Networks 183
§13.4. Elman Networks 184
§13.5. Training and Backpropagation for Recurrent Neural Networks 189
§13.6. Stability 193
§13.7. Advanced Architectures 194
§13.8. Implementation Aspects for Recurrent Neural Networks 198
§13.9. Attention Mechanism and Transformers 202
§13.10. Brief Concluding Remarks 211
§13.11. Exercises 211

Chapter 14. Convolution Neural Networks 213
§14.1. Introduction 213
§14.2. Detection of Known Signal 214
§14.3. Detection of Unknown Signal 220
§14.4. Auxiliary Thoughts 223
§14.5. SGD for Convolution Neural Networks with a Single Channel 226
§14.6. On Convolution Neural Networks with Multiple Channels 228
§14.7. Brief Concluding Remarks 231
§14.8. Exercises 231

Chapter 15. Variational Inference and Generative Models 233
§15.1. Introduction 233
§15.2. Estimating Densities and the Evidence Lower Bound 234
§15.3. Generative Adversarial Networks 239

§15.4. Optimization in GANs 243
§15.5. Wasserstein GANs 246
§15.6. Brief Concluding Remarks 248
§15.7. Exercises 248

Part 2. Advanced Topics and Convergence Results in Deep Learning

Transitioning from Part 1 to Part 2 253
1. Motivating Learning: Part 1. 253
2. Neural Networks and Universal Approximation: Part 1 ⟶ Part 2. 254
3. Training of Neural Networks: Part 1 ⟶ Part 2. 255
4. Optimize Training of Neural Networks: Part 1 ⟶ Part 2. 255
5. Optimization in the Feature Learning Regime: Part 2. 256
6. Selected Topics: Part 1 ⟶ Part 2. 256

Chapter 16. Universal Approximation Theorems 259
§16.1. Introduction 259
§16.2. Basic Universal Approximation Theorems 259
§16.3. Universal Approximation Results Using ReLU Activation Functions 266
§16.4. Brief Concluding Remarks 271
§16.5. Exercises 272

Chapter 17. Convergence Analysis of Gradient Descent 273
§17.1. Introduction 273
§17.2. Convergence Properties under Convexity Assumptions 274
§17.3. Convergence in the Absence of Convexity Assumptions 281
§17.4. Accelerated Gradient Descent Methods 286
§17.5. Brief Concluding Remarks 290
§17.6. Exercises 290

Chapter 18. Convergence Analysis of Stochastic Gradient Descent 293
§18.1. Introduction 293
§18.2. Preliminary calculations 294
§18.3. Convergence Results for SGD 297
§18.4. Comparing SGD with GD 306
§18.5. Variants of Stochastic Gradient Descent 310

§18.6. Brief Concluding Remarks 318
§18.7. Exercises 318

Chapter 19. The Neural Tangent Kernel Regime 321
§19.1. Introduction 321
§19.2. Weight Initialization 322
§19.3. The Linear Asymptotic Regime: Neural Tangent Kernel 326
§19.4. The Linear Asymptotic Regime in the Discrete Time Case 330
§19.5. Preliminary Bounds and Existence of a Limit 335
§19.6. Alternative Representation of the Prelimit Process 345
§19.7. Proof of Main Convergence Results 348
§19.8. Brief Concluding Remarks 351
§19.9. Exercises 351

Chapter 20. Optimization in the Feature Learning Regime: Mean Field Scaling 355
§20.1. Introduction 355
§20.2. Preliminary Thoughts 356
§20.3. Mean Field Limit for Shallow Neural Networks 358
§20.4. Central Limit Theorem Behavior for Shallow Neural Networks 374
§20.5. Deep Neural Networks in Mean Field Scaling 376
§20.6. In Between the Linear and the Nonlinear Regime 381
§20.7. Elements of Generalization Performance 387
§20.8. Brief Concluding Remarks 390
§20.9. Exercises 391

Chapter 21. Reinforcement Learning 393
§21.1. Introduction 393
§21.2. Motivating Reinforcement Learning Through an Example 393
§21.3. Deep Reinforcement Learning 406
§21.4. *Q*-learning 408
§21.5. Convergence Properties of the *Q*-learning Algorithm 411
§21.6. Brief Concluding Remarks 422
§21.7. Exercises 423

Chapter 22. Neural Differential Equations 427
§22.1. Introduction 427

§22.2. Ordinary Differential Equations with Neural Network Dynamics 427
§22.3. Backpropagation Formula from the Euler Discretization 430
§22.4. Training Neural ODEs with Minibatch Datasets 432
§22.5. Neural Stochastic Differential Equations 433
§22.6. Examples in PyTorch 436
§22.7. Brief Concluding Remarks 441

Chapter 23. Distributed Training 443
§23.1. Introduction 443
§23.2. Synchronous Gradient Descent 445
§23.3. Asynchronous Gradient Descent 446
§23.4. Parallel Efficiency 448
§23.5. MPI Communication 449
§23.6. Point-to-point MPI Communication 453
§23.7. Python MPI Communication 453
§23.8. Brief Concluding Remarks 459
§23.9. Exercises 459

Chapter 24. Automatic Differentiation 461
§24.1. Introduction 461
§24.2. Reverse-mode versus Forward-mode Differentiation 462
§24.3. Introduction to PyTorch Automatic Differentiation 464
§24.4. Brief Concluding Remarks 470

Part 3. Appendixes

Appendix A. Background Material in Probability 473
§A.1. Basic Notions in Probability 473
§A.2. Basics on Stochastic Processes 475
§A.3. Notions of Convergence and Tightness 477
§A.4. Convergence in the Skorokhod Space $D_E([0,T])$ 479
§A.5. Some Limiting Results and Concentration Bounds 481
§A.6. Itô Stochastic Integral 483
§A.7. Very Basics of Itô Stochastic Calculus 484

Appendix B. Background Material in Analysis 487
§B.1. Basic Inequalities Used in the Book 487
§B.2. Basic Background in Analysis 488
Bibliography 491
Index 503