Contents

1 Deep Learning: A (Currently) Black-Box Model 1
1.1 Definitions and Terminology 1
1.2 Advances in Deep Learning 7
1.3 Applications of Deep Learning 8
1.4 The Status Quo of Deep Learning Theory 10
1.5 Notations 12

Part I Background

2 Statistical Learning Theory 17
2.1 Glivenko-Cantelli Theorem and Concentration Inequalities 17
2.1.1 Glivenko-Cantelli Theorem 18
2.1.2 Concentration Inequality 24
2.2 Probably Approximately Correct (PAC) Learning 26
2.2.1 The PAC Learning Model 26
2.2.2 Generalities 28
2.2.3 Insights from Glivenko-Cantelli Theorem 32
2.3 PAC-Bayesian Learning 33
References 38

3 Hypothesis Complexity 39
3.1 Worst-Case Bounds Based on the Rademacher Complexity 39
3.2 Worst-Case Bounds Based on the Vapnik-Chervonenkis (VC) Dimension 43
References 46

4 Algorithmic Stability 47
4.1 Definition of Algorithmic Stability 47
4.2 Algorithmic Stability and Generalization Error Bounds 48
4.3 Uniform Stability of Regularized Learning 56
References 58

Part II Deep Learning Theory

5 Capacity and Complexity 61
5.1 Worst-Case Bounds Based on the VC Dimension 61
5.2 Rademacher Complexity and Covering Number 62
5.3 Generalization Bound of ResNet 66
5.3.1 Stem-Vine Framework 68
5.3.2 Generalization Bound 70
5.3.3 Proofs of Section 5.3 81
5.4 Vacuous Generalization Guarantee in Deep Learning 90
References 90

6 Stochastic Gradient Descent as an Implicit Regularization 95
6.1 Stochastic Gradient Methods (SGMs) 95
6.2 Generalization Bounds on Convex Loss Surfaces 96
6.3 Generalization Bounds on Nonconvex Loss Surfaces 97
6.4 Generalization Bounds Relying on Data-Dependent Priors 102
6.5 The Role of Learning Rate and Batch Size in Shaping Generalizability 104
6.5.1 Theoretical Evidence 105
6.5.2 Empirical Evidence 114
6.6 Interplay of Optimization and Bayesian Inference 117
References 119

7 The Geometry of the Loss Surfaces 123
7.1 Linear Networks Have No Spurious Local Minima 123
7.2 Nonlinear Activations Bring Infinite Spurious Local Minima 124
7.2.1 Neural Networks Have Infinite Spurious Local Minima 126
7.2.2 A Big Picture of the Loss Surface 130
7.2.3 Proofs of Sect. 7.2 135
7.3 Proofs of Theorems 7.7, 7.8, Corollaries 7.1, and 7.2 167
7.3.1 Squared Loss 167
7.3.2 Smooth and Multilinear Partition 168
7.3.3 Every Local Minimum Is Globally Minimal Within a Cell 168
7.3.4 Equivalence Classes of Local Minimum Valleys in Cells 172
7.4 Geometric Structure of the Loss Surface 174
References 176

8 Linear Partition in the Input Space 179
8.1 Preliminaries 179
8.2 Neural Networks Act as Hash Encoders 180
8.3 Factors that Influence the Encoding Properties 182

8.3.1 Relationship Between Model Size and Encoding Properties 182
8.3.2 Relationship Between Training Time and Encoding Properties 185
8.3.3 Relationship Between Sample Size and Encoding Properties 186
8.3.4 Layerwise Ablation Study 187
8.3.5 Impacts of Regularization, Random Data, and Random Labels 189
8.3.6 Activation Hash Phase Chart 191
8.4 Additional Experimental Implementation Details 192
8.5 Additional Experimental Results 194
References 195

9 Reflecting on the Role of Overparameterization: Is it Solely Harmful? 197
9.1 Double Descent and Benign Overfitting 197
9.2 Neural Tangent Kernels (NTKs) 200
9.3 Loss Surface and Optimization 202
9.4 Generalization and Learnability 203
References 204

10 Theoretical Foundations for Specific Architectures 207
10.1 Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) 207
10.2 Equivariant Neural Networks 208
10.2.1 Group CNNs 209
10.2.2 Steerable Neural Networks 210
10.2.3 Nonlinearities in Equivariant Networks 213
10.2.4 Generalization of Equivariant Networks 214
10.2.5 Generalization Bounds of Equivariant Networks 214
10.2.6 Approximation of Equivariant Networks 216
References 217

Part III Deep Learning Theory form the Trust-worthy Facet

11 Privacy Preservation 223
11.1 Differential Privacy 223
11.2 The Interplay of Generalizability and Privacy-Preserving Ability 225
11.2.1 Preliminaries 227
11.2.2 Generalization Bounds for Iterative Differentially Private Algorithms 229
11.2.3 Applications 254
References 260

12 Algorithmic Fairness 263
12.1 Definitions of Fairness 263
12.2 Fairness-Aware Algorithms 265
12.2.1 Preprocessing Methods 265
12.2.2 In-processing Methods 265
12.2.3 Postprocessing Methods 267
References 267

13 Adversarial Robustness 269
13.1 Adversarial Attacks and Defences 269
13.2 Interplay Between Adversarial Robustness, Privacy Preservation, and Generalizability 271
13.2.1 Measurement of Robustness 273
13.2.2 Privacy–Robustness Trade-Off 279
13.2.3 Generalization–Robustness Trade-Off 284
References 290