Deep Learning in Science
Theory, Algorithms, and Applications

Pierre Baldi

© Draft date October 5, 2020

Contents

Contents i

Preface ix

1 Introduction 3
- 1.1 Carbon-Based and Silicon-Based Computing . . . 4
- 1.2 Early Beginnings Until the Late 1940s . . . 8
  - 1.2.1 Linear Regression . . . 8
  - 1.2.2 Neuroscience Origins . . . 9
  - 1.2.3 The Deep Learning Problem . . . 12
  - 1.2.4 Hebbian Learning . . . 13
- 1.3 From 1950 to 1980 . . . 14
  - 1.3.1 Shallow Non-Linear Learning . . . 14
  - 1.3.2 First Forays into Deep Architectures and their Challenges . . . 15
- 1.4 From 1980 to Today . . . 17
- 1.5 Roadmap . . . 19
- 1.6 Exercises . . . 19

2 Basic Concepts 23
- 2.1 Synapses . . . 24
- 2.2 Units or Neurons . . . 25
- 2.3 Activations . . . 25
- 2.4 Transfer Functions . . . 26
  - 2.4.1 Identity and Linear Transfer Functions . . . 26
  - 2.4.2 Polynomial Transfer Functions . . . 27
  - 2.4.3 Max (Pool) Transfer Functions . . . 27
  - 2.4.4 Threshold Transfer Functions . . . 27
  - 2.4.5 Rectified and Piece-Wise Linear Transfer Functions . . . 29
  - 2.4.6 Sigmoidal Transfer Functions . . . 30
  - 2.4.7 Softmax Transfer Functions . . . 30
- 2.5 Discrete versus Continuous Time . . . 31
- 2.6 Networks and Architectures . . . 32

2.6.1 The Proper Definition of Deep . . . 33
2.6.2 Feedforward Architectures . . . 33
2.6.3 Recurrent Architectures . . . 34
2.6.4 Layered Architectures . . . 34
2.6.5 Weight Sharing, Convolutional, and Siamese Architectures . . . 35
2.7 Functional and Cardinal Capacity of Architectures . . . 36
2.8 The Bayesian Statistical Framework . . . 38
2.8.1 Variational Approaches . . . 41
2.9 Information Theory . . . 42
2.10 Data and Learning Settings . . . 45
2.10.1 Supervised Learning . . . 45
2.10.2 Unsupervised Learning . . . 46
2.10.3 Semi-supervised Learning . . . 46
2.10.4 Self-supervised Learning . . . 46
2.10.5 Transfer Learning . . . 46
2.10.6 On-line versus Batch Learning . . . 47
2.10.7 Reinforcement Learning . . . 47
2.11 Learning Rules . . . 47
2.12 Computational Complexity Theory . . . 48
2.13 Exercises . . . 50

3 Shallow Networks and Shallow Learning 53
3.1 Supervised Shallow Networks and their Design . . . 54
3.1.1 Regression . . . 55
3.1.2 Classification . . . 56
3.1.3 k-Classification . . . 56
3.1.4 Prior Distributions and Regularization . . . 57
3.1.5 Probabilistic Neural Networks . . . 59
3.1.6 Independence of Units During Learning in Shallow Networks . 60
3.2 Capacity of Shallow Networks . . . 60
3.2.1 Functional Capacity . . . 60
3.2.2 The Capacity of Linear Threshold Gates . . . 61
3.2.3 The Capacity of Polynomial Threshold Gates . . . 64
3.2.4 The Capacity of Other Units . . . 66
3.3 Shallow Learning . . . 68
3.3.1 Gradient Descent . . . 68
3.3.2 The Linear Case . . . 69
3.3.3 The Logistic Case . . . 70
3.3.4 The Perceptron (Linear Threshold Function) Case . . . 70
3.3.5 Data Normalization, Weight Initializations, Learning Rates, and Noise . . . 72

3.4 Extensions of Shallow Learning . . . 73
3.4.1 Top Layer of Deep Architectures . . . 73
3.4.2 Extreme Machines . . . 74
3.4.3 Support Vector Machines . . . 75
3.5 Exercises . . . 75

4 Two Layer Networks and Universal Approximation 83
4.1 Functional Capacity . . . 84
4.1.1 The Linear Model . . . 84
4.1.2 The Unrestricted Boolean Model . . . 85
4.2 Universal Approximation Properties . . . 86
4.2.1 Boolean Setting and Threshold Gates . . . 86
4.2.2 The Classification Setting . . . 86
4.2.3 The Regression Setting . . . 86
4.3 The Capacity of $A(n, m, 1)$ Architectures . . . 89
4.4 Exercises . . . 92

5 Autoencoders 95
5.1 A General Autoencoder Framework . . . 97
5.2 General Autoencoder Properties . . . 99
5.3 Linear Autoencoders . . . 101
5.3.1 Useful Reminders . . . 102
5.3.2 Group Invariances . . . 104
5.3.3 Fixed-Layer and Convexity Results . . . 104
5.3.4 Critical Points and the Landscape of $\mathcal{E}$ . . . 107
5.3.5 Learning Algorithms . . . 112
5.3.6 Generalization Properties . . . 113
5.4 Non-Linear Autoencoders: Unrestricted Boolean Case . . . 113
5.4.1 Analysis of the Unrestricted Boolean Autoencoder . . . 113
5.4.2 Boolean Autoencoder Problem Complexity . . . 117
5.5 Other Autoencoders and Autoencoder Properties . . . 119
5.5.1 Threshold Gate, Sigmoidal, and Mixed Autoencoders . . . 119
5.5.2 Inverting, Redressing, and De-Noising Autoencoders . . . 121
5.5.3 Probabilistic Autoencoders and Variational Autoencoders . . . 122
5.5.4 Expansive Autoencoders . . . 124
5.5.5 Composition of Autoencoders . . . 125
5.6 Exercises . . . 126

6 Deep Networks and Backpropagation 131
6.1 Why Deep? . . . 132
6.2 Functional Capacity: Deep Linear Case . . . 134

6.3 Functional Capacity: Deep Unrestricted Boolean Case . . . 136
6.4 Cardinal Capacity: Deep Feedforward Architectures . . . 136
6.5 Other Notions of Capacity . . . 138
6.6 Learning by Backpropagation . . . 140
6.6.1 Backpropagation . . . 140
6.6.2 Deep Targets . . . 142
6.6.3 Backpropagation through Probabilistic Layers . . . 143
6.6.4 Backpropagation in Linear Networks . . . 143
6.7 The Optimality of Backpropagation . . . 144
6.7.1 Stochastic Gradient Descent . . . 144
6.7.2 Using Backpropagation to Compute Gradients . . . 145
6.8 Architecture Design . . . 147
6.8.1 Design . . . 147
6.8.2 Invariances . . . 148
6.8.3 Ensembles . . . 149
6.9 Practical Training Issues . . . 150
6.10 The Bias-Variance Decomposition . . . 155
6.11 Dropout . . . 156
6.11.1 Ensemble Averaging: Linear Networks . . . 157
6.11.2 Ensemble Averaging: Non-linear Networks . . . 158
6.11.3 Adaptive Regularization: Linear Unit . . . 160
6.11.4 Adaptive Regularization: Linear Networks . . . 162
6.12 Model Compression and Dark Knowledge . . . 163
6.13 Multiplicative Interactions: Gating and Attention . . . 164
6.14 Unsupervised Learning and Generative Models . . . 167
6.14.1 Generative Adversarial Networks (GANs) . . . 168
6.14.2 Autoregressive Generative Models (ARMs) . . . 168
6.14.3 Flow Models . . . 169
6.15 Exercises . . . 172

7 The Local Learning Principle 179
7.1 Virtualization and Learning in the Machine . . . 180
7.2 The Neuronal View . . . 181
7.3 The Synaptic View: the Local Learning Principle . . . 182
7.4 Stratification of Learning Rules . . . 184
7.5 Deep Local Learning and its Fundamental Limitations . . . 186
7.6 Local Deep Learning: the Deep Learning Channel . . . 189
7.7 Local Deep Learning and Deep Targets Equivalence . . . 192
7.8 Exercises . . . 194

8 The Deep Learning Channel 197

8.1 Random Backpropagation (RBP) and its Variations . . . 199
8.1.1 Variations . . . 199
8.2 Simulations of Random Backpropagation . . . 202
8.3 Understanding Random Backpropagation . . . 204
8.4 Mathematical Analysis of Random Backpropagation . . . 205
8.5 Further Remarks About Learning Channels . . . 211
8.6 Circular Autoencoders . . . 213
8.7 Recirculation: Locality in Both Space and Time . . . 214
8.8 Simulations of Recirculation . . . 217
8.9 Recirculation is Random Backpropagation . . . 218
8.10 Mathematical Analysis of Recirculation . . . 221
8.11 Exercises . . . 225

9 Recurrent Networks 231
9.1 Recurrent Networks . . . 232
9.2 Cardinal Capacity of Recurrent Networks . . . 233
9.3 Symmetric Connections: The Hopfield Model . . . 234
9.4 Symmetric Connections: Boltzmann Machines . . . 237
9.4.1 Restricted Boltzmann Machines (RBMs): . . . 241
9.5 Exercises . . . 242

10 Recursive Networks 247
10.1 Variable-Size Structured Data . . . 248
10.2 Recursive Networks and Design . . . 249
10.2.1 Inner Approaches . . . 251
10.2.2 Outer Approaches . . . 259
10.3 Relationships between Inner and Outer Approaches . . . 261
10.4 Exercises . . . 263

11 Applications in Physics 267
11.1 Deep Learning in the Physical Sciences . . . 268
11.2 Antimatter Physics . . . 273
11.2.1 ASACUSA . . . 274
11.3 High Energy Collider Physics . . . 279
11.3.1 The Large Hadron Collider and the ATLAS Detector . . . 279
11.3.2 Exotic Particle Searches . . . 280
11.3.3 Jet Substructure Classification . . . 281
11.3.4 Decorrelated Tagging with Adversarial Neural Networks . . . 285
11.3.5 Building Sparse Generative Models . . . 290
11.4 Neutrino Physics . . . 291
11.5 Dark Matter Physics . . . 295

11.6 Cosmology and Astrophysics . . . 298
11.7 Climate Physics . . . 303
11.8 Incorporating Physics Knowledge and Constraints . . . 305
11.8.1 Incorporating Physics Constraints . . . 306
11.9 Conclusion: Theoretical Physics . . . 308

12 Applications in Chemistry 311
12.1 Chemical Data and Chemical Space . . . 313
12.2 Prediction of Small Molecule Properties . . . 315
12.2.1 Representations . . . 315
12.2.2 Similarity Measures and Kernels . . . 317
12.2.3 Prediction of 3D Structures . . . 317
12.2.4 Deep Learning: Feedforward Networks . . . 318
12.2.5 Deep Learning: Recursive Inner Approaches . . . 319
12.2.6 Deep Learning: Recursive Outer Approaches . . . 320
12.3 Prediction of Chemical Reactions . . . 320
12.3.1 Representations . . . 320
12.3.2 QM Approaches . . . 322
12.3.3 Rule-Based Approaches . . . 323
12.3.4 Deep Learning Approaches . . . 325

13 Applications in Biology and Medicine 333
13.1 Biomedical Data . . . 334
13.2 Life in a Nutshell . . . 335
13.3 Deep Learning in Proteomics . . . 338
13.3.1 Protein Structures . . . 339
13.3.2 Protein Secondary Structure and Other Structural Features . 342
13.3.3 Protein Contact and Distance Maps . . . 343
13.3.4 Protein Functional Features . . . 346
13.4 Deep Learning in Genomics and Transcriptomics . . . 347
13.5 Deep Learning in Biomedical Imaging . . . 349
13.6 Deep Learning in Health Care . . . 352

14 Conclusion 355
14.1 Explainability and the Black-Box Question . . . 356
14.2 ANNs versus BNNs . . . 359

15 Appendix: Reinforcement Learning 363

16 Appendix: Hints and Remarks for the Exercises 365

Bibliography 371

Index **431**