Deep Learning

Ian Goodfellow
Yoshua Bengio
Aaron Courville
Contents

Website viii
Acknowledgments ix
Notation xii

1 Introduction 1
  1.1 Who Should Read This Book? 8
  1.2 Historical Trends in Deep Learning 12

I Applied Math and Machine Learning Basics 27

2 Linear Algebra 29
  2.1 Scalars, Vectors, Matrices and Tensors 29
  2.2 Multiplying Matrices and Vectors 32
  2.3 Identity and Inverse Matrices 34
  2.4 Linear Dependence and Span 35
  2.5 Norms 37
  2.6 Special Kinds of Matrices and Vectors 38
  2.7 Eigendecomposition 40
  2.8 Singular Value Decomposition 42
  2.9 The Moore-Penrose Pseudoinverse 43
  2.10 The Trace Operator 44
  2.11 The Determinant 45
  2.12 Example: Principal Components Analysis 45

3 Probability and Information Theory 51
  3.1 Why Probability? 52
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.2 Random Variables</td>
<td>54</td>
</tr>
<tr>
<td>3.3 Probability Distributions</td>
<td>54</td>
</tr>
<tr>
<td>3.4 Marginal Probability</td>
<td>56</td>
</tr>
<tr>
<td>3.5 Conditional Probability</td>
<td>57</td>
</tr>
<tr>
<td>3.6 The Chain Rule of Conditional Probabilities</td>
<td>57</td>
</tr>
<tr>
<td>3.7 Independence and Conditional Independence</td>
<td>58</td>
</tr>
<tr>
<td>3.8 Expectation, Variance and Covariance</td>
<td>58</td>
</tr>
<tr>
<td>3.9 Common Probability Distributions</td>
<td>60</td>
</tr>
<tr>
<td>3.10 Useful Properties of Common Functions</td>
<td>65</td>
</tr>
<tr>
<td>3.11 Bayes’ Rule</td>
<td>68</td>
</tr>
<tr>
<td>3.12 Technical Details of Continuous Variables</td>
<td>69</td>
</tr>
<tr>
<td>3.13 Information Theory</td>
<td>71</td>
</tr>
<tr>
<td>3.14 Structured Probabilistic Models</td>
<td>73</td>
</tr>
<tr>
<td>4 Numerical Computation</td>
<td>78</td>
</tr>
<tr>
<td>4.1 Overflow and Underflow</td>
<td>78</td>
</tr>
<tr>
<td>4.2 Poor Conditioning</td>
<td>80</td>
</tr>
<tr>
<td>4.3 Gradient-Based Optimization</td>
<td>80</td>
</tr>
<tr>
<td>4.4 Constrained Optimization</td>
<td>91</td>
</tr>
<tr>
<td>4.5 Example: Linear Least Squares</td>
<td>94</td>
</tr>
<tr>
<td>5 Machine Learning Basics</td>
<td>96</td>
</tr>
<tr>
<td>5.1 Learning Algorithms</td>
<td>97</td>
</tr>
<tr>
<td>5.2 Capacity, Overfitting and Underfitting</td>
<td>108</td>
</tr>
<tr>
<td>5.3 Hyperparameters and Validation Sets</td>
<td>118</td>
</tr>
<tr>
<td>5.4 Estimators, Bias and Variance</td>
<td>120</td>
</tr>
<tr>
<td>5.5 Maximum Likelihood Estimation</td>
<td>129</td>
</tr>
<tr>
<td>5.6 Bayesian Statistics</td>
<td>133</td>
</tr>
<tr>
<td>5.7 Supervised Learning Algorithms</td>
<td>137</td>
</tr>
<tr>
<td>5.8 Unsupervised Learning Algorithms</td>
<td>142</td>
</tr>
<tr>
<td>5.9 Stochastic Gradient Descent</td>
<td>149</td>
</tr>
<tr>
<td>5.10 Building a Machine Learning Algorithm</td>
<td>151</td>
</tr>
<tr>
<td>5.11 Challenges Motivating Deep Learning</td>
<td>152</td>
</tr>
<tr>
<td>II Deep Networks: Modern Practices</td>
<td>162</td>
</tr>
<tr>
<td>6 Deep Feedforward Networks</td>
<td>164</td>
</tr>
<tr>
<td>6.1 Example: Learning XOR</td>
<td>167</td>
</tr>
<tr>
<td>6.2 Gradient-Based Learning</td>
<td>172</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
</tr>
<tr>
<td>---------</td>
<td>-----------------------------------------------------------------------</td>
</tr>
<tr>
<td>6.3</td>
<td>Hidden Units</td>
</tr>
<tr>
<td>6.4</td>
<td>Architecture Design</td>
</tr>
<tr>
<td>6.5</td>
<td>Back-Propagation and Other Differentiation Algorithms</td>
</tr>
<tr>
<td>6.6</td>
<td>Historical Notes</td>
</tr>
<tr>
<td>7</td>
<td>Regularization for Deep Learning</td>
</tr>
<tr>
<td>7.1</td>
<td>Parameter Norm Penalties</td>
</tr>
<tr>
<td>7.2</td>
<td>Norm Penalties as Constrained Optimization</td>
</tr>
<tr>
<td>7.3</td>
<td>Regularization and Under-Constrained Problems</td>
</tr>
<tr>
<td>7.4</td>
<td>Dataset Augmentation</td>
</tr>
<tr>
<td>7.5</td>
<td>Noise Robustness</td>
</tr>
<tr>
<td>7.6</td>
<td>Semi-Supervised Learning</td>
</tr>
<tr>
<td>7.7</td>
<td>Multitask Learning</td>
</tr>
<tr>
<td>7.8</td>
<td>Early Stopping</td>
</tr>
<tr>
<td>7.9</td>
<td>Parameter Tying and Parameter Sharing</td>
</tr>
<tr>
<td>7.10</td>
<td>Sparse Representations</td>
</tr>
<tr>
<td>7.11</td>
<td>Bagging and Other Ensemble Methods</td>
</tr>
<tr>
<td>7.12</td>
<td>Dropout</td>
</tr>
<tr>
<td>7.13</td>
<td>Adversarial Training</td>
</tr>
<tr>
<td>7.14</td>
<td>Tangent Distance, Tangent Prop and Manifold</td>
</tr>
<tr>
<td>8</td>
<td>Optimization for Training Deep Models</td>
</tr>
<tr>
<td>8.1</td>
<td>How Learning Differs from Pure Optimization</td>
</tr>
<tr>
<td>8.2</td>
<td>Challenges in Neural Network Optimization</td>
</tr>
<tr>
<td>8.3</td>
<td>Basic Algorithms</td>
</tr>
<tr>
<td>8.4</td>
<td>Parameter Initialization Strategies</td>
</tr>
<tr>
<td>8.5</td>
<td>Algorithms with Adaptive Learning Rates</td>
</tr>
<tr>
<td>8.6</td>
<td>Approximate Second-Order Methods</td>
</tr>
<tr>
<td>8.7</td>
<td>Optimization Strategies and Meta-Algorithms</td>
</tr>
<tr>
<td>9</td>
<td>Convolution Networks</td>
</tr>
<tr>
<td>9.1</td>
<td>The Convolution Operation</td>
</tr>
<tr>
<td>9.2</td>
<td>Motivation</td>
</tr>
<tr>
<td>9.3</td>
<td>Pooling</td>
</tr>
<tr>
<td>9.4</td>
<td>Convolution and Pooling as an Infinitely Strong Prior</td>
</tr>
<tr>
<td>9.5</td>
<td>Variants of the Basic Convolution Function</td>
</tr>
<tr>
<td>9.6</td>
<td>Structured Outputs</td>
</tr>
<tr>
<td>9.7</td>
<td>Data Types</td>
</tr>
</tbody>
</table>
CONTENTS

9.8 Efficient Convolution Algorithms ........................................ 356
9.9 Random or Unsupervised Features ....................................... 356
9.10 The Neuroscientific Basis for Convolutional Networks .......... 358
9.11 Convolutional Networks and the History of Deep Learning .... 365

10 Sequence Modeling: Recurrent and Recursive Nets ............... 367
10.1 Unfolding Computational Graphs ....................................... 369
10.2 Recurrent Neural Networks ............................................. 372
10.3 Bidirectional RNNs ..................................................... 388
10.4 Encoder-Decoder Sequence-to-Sequence Architectures .......... 390
10.5 Deep Recurrent Networks .............................................. 392
10.6 Recursive Neural Networks ............................................ 394
10.7 The Challenge of Long-Term Dependencies ......................... 396
10.8 Echo State Networks ................................................... 399
10.9 Leaky Units and Other Strategies for Multiple Time Scales .... 402
10.10 The Long Short-Term Memory and Other Gated RNNs .......... 404
10.11 Optimization for Long-Term Dependencies ......................... 408
10.12 Explicit Memory ........................................................ 412

11 Practical Methodology ...................................................... 416
11.1 Performance Metrics .................................................... 417
11.2 Default Baseline Models ................................................ 420
11.3 Determining Whether to Gather More Data ......................... 421
11.4 Selecting Hyperparameters ............................................ 422
11.5 Debugging Strategies .................................................. 431
11.6 Example: Multi-Digit Number Recognition ......................... 435

12 Applications ...................................................................... 438
12.1 Large-Scale Deep Learning ............................................ 438
12.2 Computer Vision .......................................................... 447
12.3 Speech Recognition ...................................................... 453
12.4 Natural Language Processing .......................................... 456
12.5 Other Applications ........................................................ 473
III Deep Learning Research 482

13 Linear Factor Models 485
13.1 Probabilistic PCA and Factor Analysis 486
13.2 Independent Component Analysis (ICA) 487
13.3 Slow Feature Analysis 489
13.4 Sparse Coding 492
13.5 Manifold Interpretation of PCA 496

14 Autoencoders 499
14.1 Undercomplete Autoencoders 500
14.2 Regularized Autoencoders 501
14.3 Representational Power, Layer Size and Depth 505
14.4 Stochastic Encoders and Decoders 506
14.5 Denoising Autoencoders 507
14.6 Learning Manifolds with Autoencoders 513
14.7 Contractive Autoencoders 518
14.8 Predictive Sparse Decomposition 521
14.9 Applications of Autoencoders 522

15 Representation Learning 524
15.1 Greedy Layer-Wise Unsupervised Pretraining 526
15.2 Transfer Learning and Domain Adaptation 534
15.3 Semi-Supervised Disentangling of Causal Factors 539
15.4 Distributed Representation 544
15.5 Exponential Gains from Depth 550
15.6 Providing Clues to Discover Underlying Causes 552

16 Structured Probabilistic Models for Deep Learning 555
16.1 The Challenge of Unstructured Modeling 556
16.2 Using Graphs to Describe Model Structure 560
16.3 Sampling from Graphical Models 577
16.4 Advantages of Structured Modeling 579
16.5 Learning about Dependencies 579
16.6 Inference and Approximate Inference 580
16.7 The Deep Learning Approach to Structured Probabilistic Models 581

17 Monte Carlo Methods 587
17.1 Sampling and Monte Carlo Methods 587
## CONTENTS

17.2 Importance Sampling .............................................. 589  
17.3 Markov Chain Monte Carlo Methods .......................... 592  
17.4 Gibbs Sampling .................................................. 596  
17.5 The Challenge of Mixing between Separated Modes .......... 597  

18 Confronting the Partition Function ............................. 603  
18.1 The Log-Likelihood Gradient .................................. 604  
18.2 Stochastic Maximum Likelihood and Contrastive Divergence 605  
18.3 Pseudolikelihood .................................................. 613  
18.4 Score Matching and Ratio Matching ........................... 615  
18.5 Denoising Score Matching ....................................... 617  
18.6 Noise-Contrastive Estimation .................................. 618  
18.7 Estimating the Partition Function ............................. 621  

19 Approximate Inference ............................................. 629  
19.1 Inference as Optimization ...................................... 631  
19.2 Expectation Maximization ...................................... 632  
19.3 MAP Inference and Sparse Coding ............................ 633  
19.4 Variational Inference and Learning ......................... 636  
19.5 Learned Approximate Inference ................................ 648  

20 Deep Generative Models ........................................... 651  
20.1 Boltzmann Machines ........................................... 651  
20.2 Restricted Boltzmann Machines ................................. 653  
20.3 Deep Belief Networks .......................................... 657  
20.4 Deep Boltzmann Machines ...................................... 660  
20.5 Boltzmann Machines for Real-Valued Data .................. 673  
20.6 Convolutional Boltzmann Machines ........................... 679  
20.7 Boltzmann Machines for Structured or Sequential Outputs 681  
20.8 Other Boltzmann Machines ..................................... 683  
20.9 Back-Propagation through Random Operations .............. 684  
20.10 Directed Generative Nets ...................................... 688  
20.11 Drawing Samples from Autoencoders ....................... 707  
20.12 Generative Stochastic Networks ............................. 710  
20.13 Other Generation Schemes .................................... 712  
20.14 Evaluating Generative Models ............................... 713  
20.15 Conclusion ..................................................... 716  

Bibliography .......................................................... 717
This book is accompanied by the above website. The website provides a variety of supplementary material, including exercises, lecture slides, corrections of mistakes, and other resources that should be useful to both readers and instructors.
Bibliography


Bengio, Y. and LeCun, Y. (2007). Scaling learning algorithms towards AI. In Large Scale Kernel Machines. 18


Bishop, C. M. (1994). Mixture density networks. 185


Franzius, M., Sprekeler, H., and Wiskott, L. (2007). Slowness and sparseness lead to place, head-direction, and spatial-view cells. 491


735


739


Koren, Y. (2009). The BellKor solution to the Netflix grand prize. 255, 475


748


Lovelace, A. (1842). Notes upon L. F. Menabrea’s “Sketch of the Analytical Engine invented by Charles Babbage”. 1


749


753


754


Olshausen, B. and Field, D. J. (2005). How close are we to understanding V1? *Neural Computation*, 17, 1665–1699. 15


Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann. 52


Distributed Processing*, volume 1, chapter 8, pages 318–362. MIT Press, Cambridge. 19,
23, 221

Distributed Processing: Explorations in the Microstructure of Cognition*. MIT Press,
Cambridge. 16

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy,
Scale Visual Recognition Challenge. 19

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy,

Hall. 84

elements of macaque V1 receptive fields. *Neuron, 46*(6), 945–956. 361

Sainath, T., Mohamed, A., Kingsbury, B., and Ramabhadran, B. (2013). Deep convolu-
tional neural networks for LVCSR. In *ICASSP 2013*. 455

Y. Bengio, D. Schuurmans, C. Williams, J. Lafferty, and A. Culotta, editors, *Advances
in Neural Information Processing Systems 22 (NIPS’09)*. 601

the International Conference on Artificial Intelligence and Statistics*, volume 5, pages
448–455. 22, 23, 527, 660, 663, 668, 669

Approximate Reasoning*. 522

preserving class neighbourhood structure. In *Proceedings of the Eleventh International
Conference on Artificial Intelligence and Statistics (AISTATS’07)*, San Juan, Porto Rico. Omnipress. 525


766


767


Töscher, A., Jahrer, M., and Bell, R. M. (2009). The BigChaos solution to the Netflix grand prize. 475


770


BIBLIOGRAPHY


772


