Deep Learning

Ian Goodfellow
Yoshua Bengio
Aaron Courville
# Contents

<table>
<thead>
<tr>
<th>Website</th>
<th>viii</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acknowledgments</td>
<td>ix</td>
</tr>
<tr>
<td>Notation</td>
<td>xiii</td>
</tr>
</tbody>
</table>

## 1 Introduction

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
CONTENTS

6.3 Hidden Units ................................................. 187
6.4 Architecture Design .......................................... 193
6.5 Back-Propagation and Other Differentiation Algorithms ...................................................................... 200
6.6 Historical Notes ................................................. 220

7 Regularization for Deep Learning 224
7.1 Parameter Norm Penalties ...................................... 226
7.2 Norm Penalties as Constrained Optimization .............. 233
7.3 Regularization and Under-Constrained Problems .......... 235
7.4 Dataset Augmentation .......................................... 236
7.5 Noise Robustness .............................................. 238
7.6 Semi-Supervised Learning ..................................... 240
7.7 Multitask Learning ............................................. 241
7.8 Early Stopping .................................................. 241
7.9 Parameter Tying and Parameter Sharing .................... 249
7.10 Sparse Representations ....................................... 251
7.11 Bagging and Other Ensemble Methods .................... 253
7.12 Dropout ...................................................... 255
7.13 Adversarial Training ......................................... 265
7.14 Tangent Distance, Tangent Prop and Manifold Tangent Classifier ........................................ 267

8 Optimization for Training Deep Models 271
8.1 How Learning Differs from Pure Optimization ............ 272
8.2 Challenges in Neural Network Optimization ............... 279
8.3 Basic Algorithms .............................................. 290
8.4 Parameter Initialization Strategies ......................... 296
8.5 Algorithms with Adaptive Learning Rates ................. 302
8.6 Approximate Second-Order Methods ....................... 307
8.7 Optimization Strategies and Meta-Algorithms ............ 313

9 Convolutional Networks 326
9.1 The Convolution Operation ................................... 327
9.2 Motivation .................................................... 329
9.3 Pooling ...................................................... 335
9.4 Convolution and Pooling as an Infinitely Strong Prior .. 339
9.5 Variants of the Basic Convolution Function ............... 342
9.6 Structured Outputs .......................................... 352
9.7 Data Types .................................................. 354
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.8</td>
<td>Efficient Convolution Algorithms</td>
<td>356</td>
</tr>
<tr>
<td>9.9</td>
<td>Random or Unsupervised Features</td>
<td>356</td>
</tr>
<tr>
<td>9.10</td>
<td>The Neuroscientific Basis for Convolutional Networks</td>
<td>358</td>
</tr>
<tr>
<td>9.11</td>
<td>Convolutional Networks and the History of Deep Learning</td>
<td>365</td>
</tr>
<tr>
<td>10</td>
<td>Sequence Modeling: Recurrent and Recursive Nets</td>
<td>367</td>
</tr>
<tr>
<td>10.1</td>
<td>Unfolding Computational Graphs</td>
<td>369</td>
</tr>
<tr>
<td>10.2</td>
<td>Recurrent Neural Networks</td>
<td>372</td>
</tr>
<tr>
<td>10.3</td>
<td>Bidirectional RNNs</td>
<td>388</td>
</tr>
<tr>
<td>10.4</td>
<td>Encoder-Decoder Sequence-to-Sequence Architectures</td>
<td>390</td>
</tr>
<tr>
<td>10.5</td>
<td>Deep Recurrent Networks</td>
<td>392</td>
</tr>
<tr>
<td>10.6</td>
<td>Recursive Neural Networks</td>
<td>394</td>
</tr>
<tr>
<td>10.7</td>
<td>The Challenge of Long-Term Dependencies</td>
<td>396</td>
</tr>
<tr>
<td>10.8</td>
<td>Echo State Networks</td>
<td>399</td>
</tr>
<tr>
<td>10.9</td>
<td>Leaky Units and Other Strategies for Multiple Time Scales</td>
<td>402</td>
</tr>
<tr>
<td>10.10</td>
<td>The Long Short-Term Memory and Other Gated RNNs</td>
<td>404</td>
</tr>
<tr>
<td>10.11</td>
<td>Optimization for Long-Term Dependencies</td>
<td>408</td>
</tr>
<tr>
<td>10.12</td>
<td>Explicit Memory</td>
<td>412</td>
</tr>
<tr>
<td>11</td>
<td>Practical Methodology</td>
<td>416</td>
</tr>
<tr>
<td>11.1</td>
<td>Performance Metrics</td>
<td>417</td>
</tr>
<tr>
<td>11.2</td>
<td>Default Baseline Models</td>
<td>420</td>
</tr>
<tr>
<td>11.3</td>
<td>Determining Whether to Gather More Data</td>
<td>421</td>
</tr>
<tr>
<td>11.4</td>
<td>Selecting Hyperparameters</td>
<td>422</td>
</tr>
<tr>
<td>11.5</td>
<td>Debugging Strategies</td>
<td>431</td>
</tr>
<tr>
<td>11.6</td>
<td>Example: Multi-Digit Number Recognition</td>
<td>435</td>
</tr>
<tr>
<td>12</td>
<td>Applications</td>
<td>438</td>
</tr>
<tr>
<td>12.1</td>
<td>Large-Scale Deep Learning</td>
<td>438</td>
</tr>
<tr>
<td>12.2</td>
<td>Computer Vision</td>
<td>447</td>
</tr>
<tr>
<td>12.3</td>
<td>Speech Recognition</td>
<td>453</td>
</tr>
<tr>
<td>12.4</td>
<td>Natural Language Processing</td>
<td>456</td>
</tr>
<tr>
<td>12.5</td>
<td>Other Applications</td>
<td>473</td>
</tr>
</tbody>
</table>
## III Deep Learning Research

### 13 Linear Factor Models
- 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
- 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
- 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
- 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
- 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


717


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


732


735


738


Kiros, R., Salakhutdinov, R., and Zemel, R. (2014b). Unifying visual-semantic embeddings with multimodal neural language models. *arXiv:1411.2539 [cs.LG]*. 100, 404


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


746


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


754


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


762


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


769


770
BIBLIOGRAPHY


771
BIBLIOGRAPHY


772


