Contents

Acknowledgments **vii**

Notation **ix**

1 Introduction **1**
1.1 Who Should Read This Book? 8
1.2 Historical Trends in Deep Learning 11

I Applied Math and Machine Learning Basics **26**

2 Linear Algebra **28**
2.1 Scalars, Vectors, Matrices and Tensors 28
2.2 Multiplying Matrices and Vectors 30
2.3 Identity and Inverse Matrices 32
2.4 Linear Dependence, Span and Rank 33
2.5 Norms 35
2.6 Special Kinds of Matrices and Vectors 36
2.7 Eigendecomposition 37
2.8 Singular Value Decomposition 40
2.9 The Moore-Penrose Pseudoinverse 41
2.10 The Trace Operator 42
2.11 Determinant 43
2.12 Example: Principal Components Analysis 43

3 Probability and Information Theory **48**
3.1 Why Probability? 48
3.2 Random Variables 51
3.3 Probability Distributions 51
3.4 Marginal Probability 53
3.5 Conditional Probability 53

3.6 The Chain Rule of Conditional Probabilities 54
3.7 Independence and Conditional Independence 54
3.8 Expectation, Variance and Covariance 55
3.9 Information Theory 56
3.10 Common Probability Distributions 59
3.11 Useful Properties of Common Functions 65
3.12 Bayes' Rule 66
3.13 Technical Details of Continuous Variables 67
3.14 Structured Probabilistic Models 68
3.15 Example: Naive Bayes 71

4 Numerical Computation 77
4.1 Overflow and Underflow 77
4.2 Poor Conditioning 78
4.3 Gradient-Based Optimization 79
4.4 Constrained Optimization 88
4.5 Example: Linear Least Squares 90

5 Machine Learning Basics 92
5.1 Learning Algorithms 92
5.2 Example: Linear Regression 100
5.3 Generalization, Capacity, Overfitting and Underfitting 103
5.4 Hyperparameters and Validation Sets 113
5.5 Estimators, Bias and Variance 115
5.6 Maximum Likelihood Estimation 123
5.7 Bayesian Statistics 126
5.8 Supervised Learning Algorithms 133
5.9 Unsupervised Learning Algorithms 138
5.10 Weakly Supervised Learning 141
5.11 Building a Machine Learning Algorithm 142
5.12 The Curse of Dimensionality and Statistical Limitations of Local Generalization 143

II Modern Practical Deep Networks 155

6 Feedforward Deep Networks 157
6.1 Vanilla MLPs 158
6.2 Estimating Conditional Statistics 162
6.3 Parametrizing a Learned Predictor 162
6.4 Flow Graphs and Back-Propagation 174

6.5 Universal Approximation Properties and Depth 188
6.6 Feature / Representation Learning 191
6.7 Piecewise Linear Hidden Units 192
6.8 Historical Notes 194

7 Regularization of Deep or Distributed Models 196
7.1 Regularization from a Bayesian Perspective 198
7.2 Classical Regularization: Parameter Norm Penalty 199
7.3 Classical Regularization as Constrained Optimization 207
7.4 Regularization and Under-Constrained Problems 208
7.5 Dataset Augmentation 210
7.6 Classical Regularization as Noise Robustness 211
7.7 Early Stopping as a Form of Regularization 217
7.8 Parameter Tying and Parameter Sharing 223
7.9 Sparse Representations 224
7.10 Bagging and Other Ensemble Methods 226
7.11 Dropout 227
7.12 Multi-Task Learning 232
7.13 Adversarial Training 234

8 Optimization for Training Deep Models 236
8.1 Optimization for Model Training 236
8.2 Challenges in Optimization 241
8.3 Optimization Algorithms I: Basic Algorithms 250
8.4 Optimization Algorithms II: Adaptive Learning Rates 256
8.5 Optimization Algorithms III: Approximate Second-Order Methods261
8.6 Optimization Algorithms IV: Natural Gradient Methods 262
8.7 Optimization Strategies and Meta-Algorithms 262
8.8 Hints, Global Optimization and Curriculum Learning 270

9 Convolutional Networks 274
9.1 The Convolution Operation 275
9.2 Motivation 278
9.3 Pooling 282
9.4 Convolution and Pooling as an Infinitely Strong Prior 287
9.5 Variants of the Basic Convolution Function 288
9.6 Structured Outputs 295
9.7 Convolutional Modules 295
9.8 Data Types 295
9.9 Efficient Convolution Algorithms 297
9.10 Random or Unsupervised Features 298

9.11 The Neuroscientific Basis for Convolutional Networks 299
9.12 Convolutional Networks and the History of Deep Learning 305

10 Sequence Modeling: Recurrent and Recursive Nets 308
10.1 Unfolding Flow Graphs and Sharing Parameters 309
10.2 Recurrent Neural Networks 311
10.3 Bidirectional RNNs 322
10.4 Encoder-Decoder Sequence-to-Sequence Architectures 323
10.5 Deep Recurrent Networks 325
10.6 Recursive Neural Networks 326
10.7 Auto-Regressive Networks 328
10.8 Facing the Challenge of Long-Term Dependencies 334
10.9 Handling Temporal Dependencies with n-grams, HMMs, CRFs and Other Graphical Models 347
10.10 Combining Neural Networks and Search 358

11 Practical methodology 364
11.1 Basic Machine Learning Methodology 364
11.2 Selecting Hyperparameters 365
11.3 Debugging Strategies 373

12 Applications 376
12.1 Large Scale Deep Learning 376
12.2 Computer Vision 384
12.3 Speech Recognition 391
12.4 Natural Language Processing and Neural Language Models . . . 393
12.5 Structured Outputs 408
12.6 Other Applications 409

III Deep Learning Research 410

13 Structured Probabilistic Models for Deep Learning 412
13.1 The Challenge of Unstructured Modeling 413
13.2 Using Graphs to Describe Model Structure 417
13.3 Advantages of Structured Modeling 431
13.4 Learning About Dependencies 432
13.5 Inference and Approximate Inference Over Latent Variables . . . 434
13.6 The Deep Learning Approach to Structured Probabilistic Models 435

14 Monte Carlo Methods 440
14.1 Markov Chain Monte Carlo Methods 440

14.2 The Difficulty of Mixing Between Well-Separated Modes 442

15 Linear Factor Models and Auto-Encoders 444

15.1 Regularized Auto-Encoders 445
15.2 Denoising Auto-encoders 448
15.3 Representational Power, Layer Size and Depth 450
15.4 Reconstruction Distribution 451
15.5 Linear Factor Models 452
15.6 Probabilistic PCA and Factor Analysis 453
15.7 Reconstruction Error as Log-Likelihood 457
15.8 Sparse Representations 458
15.9 Denoising Auto-Encoders 463
15.10 Contractive Auto-Encoders 468

16 Representation Learning 471

16.1 Greedy Layerwise Unsupervised Pre-Training 472
16.2 Transfer Learning and Domain Adaptation 479
16.3 Semi-Supervised Learning 486
16.4 Semi-Supervised Learning and Disentangling Underlying Causal Factors 487
16.5 Assumption of Underlying Factors and Distributed Representation 489
16.6 Exponential Gain in Representational Efficiency from Distributed Representations 493
16.7 Exponential Gain in Representational Efficiency from Depth . . . 495
16.8 Priors Regarding The Underlying Factors 497

17 The Manifold Perspective on Representation Learning 501

17.1 Manifold Interpretation of PCA and Linear Auto-Encoders . . . 509
17.2 Manifold Interpretation of Sparse Coding 512
17.3 The Entropy Bias from Maximum Likelihood 512
17.4 Manifold Learning via Regularized Auto-Encoders 513
17.5 Tangent Distance, Tangent-Prop, and Manifold Tangent Classifier 514

18 Confronting the Partition Function 518

18.1 The Log-Likelihood Gradient of Energy-Based Models 519
18.2 Stochastic Maximum Likelihood and Contrastive Divergence . . . 521
18.3 Pseudolikelihood 528
18.4 Score Matching and Ratio Matching 530
18.5 Denoising Score Matching 532
18.6 Noise-Contrastive Estimation 532
18.7 Estimating the Partition Function 534

19 Approximate inference 542
19.1 Inference as Optimization 544
19.2 Expectation Maximization 545
19.3 MAP Inference: Sparse Coding as a Probabilistic Model 546
19.4 Variational Inference and Learning 547
19.5 Stochastic Inference 551
19.6 Learned Approximate Inference 551

20 Deep Generative Models 553
20.1 Boltzmann Machines 553
20.2 Restricted Boltzmann Machines 556
20.3 Training Restricted Boltzmann Machines 559
20.4 Deep Belief Networks 563
20.5 Deep Boltzmann Machines 566
20.6 Boltzmann Machines for Real-Valued Data 577
20.7 Convolutional Boltzmann Machines 580
20.8 Other Boltzmann Machines 581
20.9 Directed Generative Nets 581
20.10 A Generative View of Autoencoders 583
20.11 Generative Stochastic Networks 589
20.12 Methodological Notes 591

Bibliography 595

Index 637