

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 and Span	33
2.5 Norms	35
2.6 Special Kinds of Matrices and Vectors	36
2.7 Eigendecomposition	38
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	67
3.13	Technical Details of Continuous Variables	67
3.14	Structured Probabilistic Models	69
3.15	Example: Naive Bayes	70
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	124
5.7	Bayesian Statistics	127
5.8	Supervised Learning Algorithms	134
5.9	Unsupervised Learning Algorithms	139
5.10	Weakly Supervised Learning	142
5.11	Building a Machine Learning Algorithm	143
5.12	The Curse of Dimensionality and Statistical Limitations of Local Generalization	145
II	Deep Networks: Modern Practices	156
6	Feedforward Deep Networks	158
6.1	MLPs from the 1980's	159
6.2	Estimating Conditional Statistics	163
6.3	Parametrizing a Learned Predictor	163
6.4	Flow Graphs and Back-Propagation	175

6.5	Back-propagation through Random Operations and Graphical Models	188
6.6	Universal Approximation Properties and Depth	192
6.7	Feature / Representation Learning	195
6.8	Piecewise Linear Hidden Units	197
6.9	Historical Notes	199
7	Regularization of Deep or Distributed Models	201
7.1	Regularization from a Bayesian Perspective	203
7.2	Classical Regularization: Parameter Norm Penalty	204
7.3	Classical Regularization as Constrained Optimization	212
7.4	Regularization and Under-Constrained Problems	213
7.5	Dataset Augmentation	214
7.6	Classical Regularization as Noise Robustness	216
7.7	Early Stopping as a Form of Regularization	220
7.8	Parameter Tying and Parameter Sharing	227
7.9	Sparse Representations	228
7.10	Bagging and Other Ensemble Methods	230
7.11	Dropout	232
7.12	Multi-Task Learning	235
7.13	Adversarial Training	236
8	Optimization for Training Deep Models	240
8.1	Optimization for Model Training	241
8.2	Challenges in Neural Network Optimization	246
8.3	Optimization Algorithms I: Basic Algorithms	259
8.4	Optimization Algorithms II: Adaptive Learning Rates	265
8.5	Optimization Algorithms III: Approximate Second-Order Methods	270
8.6	Optimization Algorithms IV: Natural Gradient Methods	280
8.7	Optimization Strategies and Meta-Algorithms	282
9	Convolutional Networks	296
9.1	The Convolution Operation	297
9.2	Motivation	300
9.3	Pooling	306
9.4	Convolution and Pooling as an Infinitely Strong Prior	309
9.5	Variants of the Basic Convolution Function	310
9.6	Structured Outputs	316
9.7	Data Types	317
9.8	Efficient Convolution Algorithms	319

9.9	Random or Unsupervised Features	320
9.10	The Neuroscientific Basis for Convolutional Networks	321
9.11	Convolutional Networks and the History of Deep Learning	327
10	Sequence Modeling: Recurrent and Recursive Nets	330
10.1	Unfolding Flow Graphs and Sharing Parameters	331
10.2	Recurrent Neural Networks	333
10.3	Bidirectional RNNs	348
10.4	Encoder-Decoder Sequence-to-Sequence Architectures	348
10.5	Deep Recurrent Networks	350
10.6	Recursive Neural Networks	352
10.7	The Challenge of Long-Term Dependencies	353
11	Practical methodology	371
11.1	Default Baseline Models	373
11.2	Selecting Hyperparameters	374
11.3	Debugging Strategies	383
12	Applications	388
12.1	Large Scale Deep Learning	388
12.2	Computer Vision	396
12.3	Speech Recognition	401
12.4	Natural Language Processing and Neural Language Models . . .	405
12.5	Structured Outputs	421
12.6	Other Applications	423
III	Deep Learning Research	432
13	Structured Probabilistic Models for Deep Learning	434
13.1	The Challenge of Unstructured Modeling	435
13.2	Using Graphs to Describe Model Structure	439
13.3	Advantages of Structured Modeling	453
13.4	Learning about Dependencies	454
13.5	Inference and Approximate Inference over Latent Variables . . .	456
13.6	The Deep Learning Approach to Structured Probabilistic Models	457
14	Monte Carlo Methods	462
14.1	Markov Chain Monte Carlo Methods	462
14.2	The Difficulty of Mixing between Well-Separated Modes	464

15	Linear Factor Models and Auto-Encoders	466
15.1	Regularized Auto-Encoders	467
15.2	Denoising Auto-encoders	470
15.3	Representational Power, Layer Size and Depth	472
15.4	Reconstruction Distribution	473
15.5	Linear Factor Models	474
15.6	Probabilistic PCA and Factor Analysis	475
15.7	Reconstruction Error as Log-Likelihood	479
15.8	Sparse Representations	480
15.9	Denoising Auto-Encoders	485
15.10	Contractive Auto-Encoders	490
16	Representation Learning	493
16.1	Greedy Layerwise Unsupervised Pre-Training	494
16.2	Transfer Learning and Domain Adaptation	501
16.3	Semi-Supervised Learning	508
16.4	Semi-Supervised Learning and Disentangling Underlying Causal Factors	509
16.5	Assumption of Underlying Factors and Distributed Representation	511
16.6	Exponential Gain in Representational Efficiency from Distributed Representations	515
16.7	Exponential Gain in Representational Efficiency from Depth . . .	517
16.8	Priors regarding the Underlying Factors	520
17	The Manifold Perspective on Representation Learning	523
17.1	Manifold Interpretation of PCA and Linear Auto-Encoders . . .	531
17.2	Manifold Interpretation of Sparse Coding	534
17.3	The Entropy Bias from Maximum Likelihood	534
17.4	Manifold Learning via Regularized Auto-Encoders	535
17.5	Tangent Distance, Tangent-Prop, and Manifold Tangent Classifier	536
18	Confronting the Partition Function	540
18.1	The Log-Likelihood Gradient of Energy-Based Models	541
18.2	Stochastic Maximum Likelihood and Contrastive Divergence . . .	543
18.3	Pseudolikelihood	550
18.4	Score Matching and Ratio Matching	552
18.5	Denoising Score Matching	554
18.6	Noise-Contrastive Estimation	554
18.7	Estimating the Partition Function	556

19	Approximate inference	564
19.1	Inference as Optimization	566
19.2	Expectation Maximization	567
19.3	MAP Inference: Sparse Coding as a Probabilistic Model	568
19.4	Sequence Modeling with Graphical Models	569
19.5	Combining Neural Networks and Search	579
19.6	Variational Inference and Learning	584
19.7	Stochastic Inference	588
19.8	Learned Approximate Inference	588
20	Deep Generative Models	590
20.1	Boltzmann Machines	590
20.2	Restricted Boltzmann Machines	593
20.3	Training Restricted Boltzmann Machines	596
20.4	Deep Belief Networks	600
20.5	Deep Boltzmann Machines	603
20.6	Boltzmann Machines for Real-Valued Data	614
20.7	Convolutional Boltzmann Machines	617
20.8	Other Boltzmann Machines	618
20.9	Directed Generative Nets	618
20.10	Auto-Regressive Networks	621
20.11	A Generative View of Autoencoders	626
20.12	Generative Stochastic Networks	632
20.13	Methodological Notes	634
	Bibliography	638
	Index	686