<table>
<thead>
<tr>
<th>Chapter</th>
<th>Section</th>
<th>Title</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>3.1</td>
<td>Probability Density Functions</td>
<td>70</td>
</tr>
<tr>
<td></td>
<td>3.2</td>
<td>Mean and Variance</td>
<td>70</td>
</tr>
<tr>
<td></td>
<td>3.3</td>
<td>Dependent Random Variables</td>
<td>73</td>
</tr>
<tr>
<td></td>
<td>3.4</td>
<td>Random Vectors</td>
<td>79</td>
</tr>
<tr>
<td></td>
<td>3.5</td>
<td>Properties of Covariance Matrices</td>
<td>98</td>
</tr>
<tr>
<td></td>
<td>3.6</td>
<td>Illustrative Applications</td>
<td>99</td>
</tr>
<tr>
<td></td>
<td>3.7</td>
<td>Complex-Valued Variables</td>
<td>108</td>
</tr>
<tr>
<td></td>
<td>3.8</td>
<td>Commentaries and Discussion</td>
<td>111</td>
</tr>
<tr>
<td></td>
<td>3.A</td>
<td>Convergence of Random Variables</td>
<td>114</td>
</tr>
<tr>
<td></td>
<td>3.B</td>
<td>Concentration Inequalities</td>
<td>124</td>
</tr>
<tr>
<td>4</td>
<td>4.1</td>
<td>Scalar Gaussian Variables</td>
<td>134</td>
</tr>
<tr>
<td></td>
<td>4.2</td>
<td>Vector Gaussian Variables</td>
<td>134</td>
</tr>
<tr>
<td></td>
<td>4.3</td>
<td>Useful Gaussian Manipulations</td>
<td>136</td>
</tr>
<tr>
<td></td>
<td>4.4</td>
<td>Jointly Distributed Gaussian Variables</td>
<td>140</td>
</tr>
<tr>
<td></td>
<td>4.5</td>
<td>Gaussian Processes</td>
<td>146</td>
</tr>
<tr>
<td></td>
<td>4.6</td>
<td>Circular Gaussian Distribution</td>
<td>152</td>
</tr>
<tr>
<td></td>
<td>4.7</td>
<td>Commentaries and Discussion</td>
<td>157</td>
</tr>
<tr>
<td></td>
<td>Problems</td>
<td></td>
<td>159</td>
</tr>
<tr>
<td></td>
<td>References</td>
<td></td>
<td>162</td>
</tr>
<tr>
<td>5</td>
<td>5.1</td>
<td>Definition</td>
<td>169</td>
</tr>
<tr>
<td></td>
<td>5.2</td>
<td>Special Cases</td>
<td>169</td>
</tr>
<tr>
<td></td>
<td>5.3</td>
<td>Useful Properties</td>
<td>171</td>
</tr>
<tr>
<td></td>
<td>5.4</td>
<td>Conjugate Priors</td>
<td>180</td>
</tr>
<tr>
<td></td>
<td>5.5</td>
<td>Commentaries and Discussion</td>
<td>185</td>
</tr>
<tr>
<td></td>
<td>Problems</td>
<td></td>
<td>189</td>
</tr>
<tr>
<td></td>
<td>References</td>
<td></td>
<td>190</td>
</tr>
<tr>
<td>6</td>
<td>6.1</td>
<td>Information and Entropy</td>
<td>194</td>
</tr>
<tr>
<td></td>
<td>6.2</td>
<td>Kullback–Leibler Divergence</td>
<td>198</td>
</tr>
<tr>
<td></td>
<td>6.3</td>
<td>Maximum Entropy Distribution</td>
<td>205</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>211</td>
</tr>
</tbody>
</table>
## References

341

### 11 Proximal Operator

11.1 Definition and Properties 342  
11.2 Proximal Point Algorithm 348  
11.3 Proximal Gradient Algorithm 350  
11.4 Convergence Results 355  
11.5 Douglas–Rachford Algorithm 357  
11.6 Commentaries and Discussion 359  

*Problems*

11.A Convergence under Convexity 367  
11.B Convergence under Strong Convexity 370  

*References* 373

### 12 Gradient-Descent Method

12.1 Empirical and Stochastic Risks 376  
12.2 Conditions on Risk Function 380  
12.3 Constant Step Sizes 382  
12.4 Iteration-Dependent Step Sizes 393  
12.5 Coordinate-Descent Method 403  
12.6 Alternating Projection Algorithm 414  
12.7 Commentaries and Discussion 419  

*Problems* 426  
12.A Zeroth-Order Optimization 434  

*References* 437

### 13 Conjugate Gradient Method

13.1 Linear Systems of Equations 442  
13.2 Nonlinear Optimization 455  
13.3 Convergence Analysis 460  
13.4 Commentaries and Discussion 466  

*Problems* 467  

*References* 470

### 14 Subgradient Method

14.1 Subgradient Algorithm 472  
14.2 Conditions on Risk Function 477  
14.3 Convergence Behavior 480  
14.4 Pocket Variable 484  
14.5 Exponential Smoothing 487  
14.6 Iteration-Dependent Step-Sizes 490  
14.7 Coordinate-Descent Algorithms 494  
14.8 Commentaries and Discussion 497  

*Problems* 499
Contents

14.A Deterministic Inequality Recursion 502

References 506

15 Proximal and Mirror-Descent Methods 508
15.1 Proximal Gradient Method 508
15.2 Projection Gradient Method 516
15.3 Mirror-Descent Method 520
15.4 Comparison of Convergence Rates 538
15.5 Commentaries and Discussion 540

Problems 542

References 545

16 Stochastic Optimization 548
16.1 Stochastic Gradient Algorithm 549
16.2 Stochastic Subgradient Algorithm 566
16.3 Stochastic Proximal Gradient Algorithm 570
16.4 Gradient Noise 576
16.5 Regret Analysis 577
16.6 Commentaries and Discussion 583

Problems 587
16.A Switching Expectation and Differentiation 591

References 596

17 Adaptive Gradient Methods 600
17.1 Motivation 600
17.2 AdaGrad Algorithm 604
17.3 RMSprop Algorithm 609
17.4 ADAM Algorithm 611
17.5 Momentum Acceleration Methods 615
17.6 Federated Learning 620
17.7 Commentaries and Discussion 627

Problems 631
17.A Regret Analysis for ADAM 633

References 641

18 Gradient Noise 643
18.1 Motivation 643
18.2 Smooth Risk Functions 646
18.3 Gradient Noise for Smooth Risks 649
18.4 Nonsmooth Risk Functions 661
18.5 Gradient Noise for Nonsmooth Risks 666
18.6 Commentaries and Discussion 674

Problems 676
18.A Averaging over Mini-Batches 678
18.B Auxiliary Variance Result 680

References 682

19 Convergence Analysis I: Stochastic Gradient Algorithms 684
19.1 Problem Setting 684
19.2 Convergence under Uniform Sampling 687
19.3 Convergence of Mini-Batch Implementation 692
19.4 Convergence under Vanishing Step Sizes 693
19.5 Convergence under Random Reshuffling 699
19.6 Convergence under Importance Sampling 702
19.7 Convergence of Stochastic Conjugate Gradient 708
19.8 Commentaries and Discussion 713

Problems 717

19.A Stochastic Inequality Recursion 721
19.B Proof of Theorem 19.5 723

References 728

20 Convergence Analysis II: Stochastic Subgradient Algorithms 731
20.1 Problem Setting 731
20.2 Convergence under Uniform Sampling 736
20.3 Convergence with Pocket Variables 739
20.4 Convergence with Exponential Smoothing 741
20.5 Convergence of Mini-Batch Implementation 746
20.6 Convergence under Vanishing Step Sizes 748
20.7 Commentaries and Discussion 751

Problems 754

References 755

21 Convergence Analysis III: Stochastic Proximal Algorithms 757
21.1 Problem Setting 757
21.2 Convergence under Uniform Sampling 762
21.3 Convergence of Mini-Batch Implementation 766
21.4 Convergence under Vanishing Step Sizes 767
21.5 Stochastic Projection Gradient 770
21.6 Mirror Descent Algorithm 772
21.7 Commentaries and Discussion 776

Problems 777

References 778

22 Variance-Reduced Methods I: Uniform Sampling 780
22.1 Problem Setting 780
22.2 Naïve Stochastic Gradient Algorithm 783
22.3 Stochastic Average-Gradient Algorithm (SAGA) 786
22.4 Stochastic Variance-Reduced Gradient Algorithm (SVRG) 794
## Contents

25.B Proof of Property (25.71) 950  
25.C Convergence of Primal Algorithms 950  
  References 966  

**26 Decentralized Optimization II: Primal-Dual Methods** 970  
26.1 Motivation 970  
26.2 EXTRA Algorithm 971  
26.3 EXACT Diffusion Algorithm 973  
26.4 Distributed Inexact Gradient Algorithm 976  
26.5 Augmented Decentralized Gradient Method 979  
26.6 ATC Tracking Method 980  
26.7 Unified Decentralized Algorithm 984  
26.8 Convergence Performance 986  
26.9 Dual Method 988  
26.10 Decentralized Nonconvex Optimization 991  
26.11 Commentaries and Discussion 996  
  Problems 999  
  References 1007  

**VOLUME II INFEERENCE** 1011  

27 Mean-Square-Error Inference 1013  
27.1 Inference without Observations 1014  
27.2 Inference with Observations 1017  
27.3 Gaussian Random Variables 1026  
27.4 Bias–Variance Relation 1031  
27.5 Commentaries and Discussion 1041  
  Problems 1045  
27.A Circular Gaussian Distribution 1048  
  References 1049  

28 Bayesian Inference 1052  
28.1 Bayesian Formulation 1052  
28.2 Maximum A-Posteriori Inference 1054  
28.3 Bayes Classifier 1057  
28.4 Logistic Regression Inference 1066  
28.5 Discriminative and Generative Models 1070  
28.6 Commentaries and Discussion 1073  
  Problems 1076  
  References 1079  

29 Linear Regression 1081  
29.1 Regression Model 1081
29.2 Centering and Augmentation 1088
29.3 Vector Estimation 1091
29.4 Linear Models 1094
29.5 Data Fusion 1096
29.6 Minimum-Variance Unbiased Estimation 1099
29.7 Commentaries and Discussion 1103
   Problems 1105
29.A Consistency of Normal Equations 1111
   References 1112

30 Kalman Filter 1114
30.1 Uncorrelated Observations 1114
30.2 Innovations Process 1117
30.3 State-Space Model 1119
30.4 Measurement- and Time-Update Forms 1131
30.5 Steady-State Filter 1137
30.6 Smoothing Filters 1141
30.7 Ensemble Kalman Filter 1145
30.8 Nonlinear Filtering 1152
30.9 Commentaries and Discussion 1162
   Problems 1165
   References 1169

31 Maximum Likelihood 1172
31.1 Problem Formulation 1172
31.2 Gaussian Distribution 1175
31.3 Multinomial Distribution 1184
31.4 Exponential Family of Distributions 1187
31.5 Cramer–Rao Lower Bound 1190
31.6 Model Selection 1198
31.7 Commentaries and Discussion 1211
   Problems 1220
31.A Derivation of the Cramer–Rao Bound 1225
31.B Derivation of the AIC Formulation 1227
31.C Derivation of the BIC Formulation 1231
   References 1233

32 Expectation Maximization 1236
32.1 Motivation 1236
32.2 Derivation of the EM Algorithm 1242
32.3 Gaussian Mixture Models 1247
32.4 Bernoulli Mixture Models 1262
32.5 Commentaries and Discussion 1268
   Problems 1270
Contents

37.4 Estimating Model Parameters 1460
37.5 Commentaries and Discussion 1474

Problems 1475
References 1475

38 Hidden Markov Models 1477
38.1 Gaussian Mixture Models 1477
38.2 Markov Chains 1482
38.3 Forward–Backward Recursions 1498
38.4 Validation and Prediction Tasks 1507
38.5 Commentaries and Discussion 1511

Problems 1517
References 1520

39 Decoding Hidden Markov Models 1523
39.1 Decoding States 1523
39.2 Decoding Transition Probabilities 1525
39.3 Normalization and Scaling 1529
39.4 Viterbi Algorithm 1534
39.5 EM Algorithm for Dependent Observations 1546
39.6 Commentaries and Discussion 1564

Problems 1565
References 1567

40 Independent Component Analysis 1569
40.1 Problem Formulation 1570
40.2 Maximum-Likelihood Formulation 1577
40.3 Mutual Information Formulation 1582
40.4 Maximum Kurtosis Formulation 1587
40.5 Projection Pursuit 1594
40.6 Commentaries and Discussion 1597

Problems 1598
References 1600

41 Bayesian Networks 1603
41.1 Curse of Dimensionality 1604
41.2 Probabilistic Graphical Models 1607
41.3 Active and Blocked Pathways 1621
41.4 Conditional Independence Relations 1630
41.5 Commentaries and Discussion 1637

Problems 1639
References 1640

42 Inference over Graphs 1642
42.1 Probabilistic Inference 1642
42.2 Inference by Enumeration 1645
42.3 Inference by Variable Elimination 1651
42.4 Chow–Liu Algorithm 1658
42.5 Graphical LASSO 1665
42.6 Learning Graph Parameters 1671
42.7 Commentaries and Discussion 1693

Problems 1694
References 1697

43 Undirected Graphs 1699
43.1 Cliques and Potentials 1699
43.2 Representation Theorem 1711
43.3 Factor Graphs 1715
43.4 Message-Passing Algorithms 1720
43.5 Commentaries and Discussion 1752

Problems 1755
43.A Proof of Hammersley–Clifford Theorem 1758
43.B Equivalence of Markovian Properties 1762

References 1763

44 Markov Decision Processes 1766
44.1 MDP Model 1766
44.2 Discounted Rewards 1780
44.3 Policy Evaluation 1784
44.4 Linear Function Approximation 1799
44.5 Commentaries and Discussion 1807

Problems 1809
References 1810

45 Value and Policy Iterations 1812
45.1 Value Iteration 1812
45.2 Policy Iteration 1825
45.3 Partially Observable MDP 1838
45.4 Commentaries and Discussion 1852

Problems 1859
45.A Optimal Policy and State-Action Values 1862
45.B Convergence of Value Iteration 1864
45.C Proof of $\epsilon$–Optimality 1865
45.D Convergence of Policy Iteration 1866
45.E Piecewise Linear Property 1868
45.F Bellman Principle of Optimality 1869

References 1873
## Temporal Difference Learning

46.1 Model-Based Learning 1877
46.2 Monte–Carlo Policy Evaluation 1879
46.3 TD(0) Algorithm 1887
46.4 Look-Ahead TD Algorithm 1895
46.5 TD(λ) Algorithm 1899
46.6 True Online TD(λ) Algorithm 1908
46.7 Off-Policy Learning 1911
46.8 Commentaries and Discussion 1916

### Problems
46.A Useful Convergence Result 1918
46.B Convergence of TD(0) Algorithm 1919
46.C Convergence of TD(λ) Algorithm 1922
46.D Equivalence of Offline Implementations 1926

### References
1928

## Q−Learning

47.1 SARSA(0) Algorithm 1930
47.2 Look-Ahead SARSA Algorithm 1934
47.3 SARSA(λ) Algorithm 1935
47.4 Off-Policy Learning 1938
47.5 Optimal Policy Extraction 1939
47.6 Q−Learning Algorithm 1941
47.7 Exploration versus Exploitation 1944
47.8 Q−Learning with Replay Buffer 1952
47.9 Double Q−Learning 1953
47.10 Commentaries and Discussion 1955

### Problems
47.A Convergence of SARSA(0) Algorithm 1960
47.B Convergence of Q−Learning Algorithm 1962

### References
1964

## Value Function Approximation

48.1 Stochastic Gradient TD-Learning 1967
48.2 Least-Squares TD-Learning 1977
48.3 Projected Bellman Learning 1978
48.4 SARSA Methods 1985
48.5 Deep Q−Learning 1991
48.6 Commentaries and Discussion 2000

### Problems
2002

### References
2004

## Policy Gradient Methods

49.1 Policy Model 2006
49.2 Finite–Difference Method 2007
49.3 Score Function 2009
49.4 Objective Functions 2011
49.5 Policy Gradient Theorem 2016
49.6 Actor–Critic Algorithms 2018
49.7 Natural Gradient Policy 2030
49.8 Trust Region Policy Optimization 2033
49.9 Deep Reinforcement Learning 2052
49.10 Soft Learning 2057
49.11 Commentaries and Discussion 2065

Problems 2067

49.A Proof of Policy Gradient Theorem 2071
49.B Proof of Consistency Theorem 2075

References 2077

VOLUME III LEARNING 2081

50 Least-Squares Problems 2083
50.1 Motivation 2083
50.2 Normal Equations 2088
50.3 Recursive Least-Squares 2105
50.4 Implicit Bias 2113
50.5 Commentaries and Discussion 2115

Problems 2120

50.A Minimum-Norm Solution 2128
50.B Equivalence in Linear Estimation 2129
50.C Extended Least-Squares 2130

References 2135

51 Regularization 2138
51.1 Three Challenges 2139
51.2 $\ell_2$–Regularization 2142
51.3 $\ell_1$–Regularization 2147
51.4 Soft Thresholding 2151
51.5 Commentaries and Discussion 2159

Problems 2162

51.A Constrained Formulations for Regularization 2167
51.B Expression for LASSO Solution 2170

References 2174

52 Nearest-Neighbor Rule 2176
52.1 Bayes Classifier 2178
52.2 $k$–NN Classifier 2181
52.3 Performance Guarantee 2184
52.4 k–Means Algorithm 2186
52.5 Commentaries and Discussion 2195
  Problems 2198
52.A Performance of the NN Classifier 2200
  References 2203

53 Self-Organizing Maps 2206
53.1 Grid Arrangements 2206
53.2 Training Algorithm 2209
53.3 Visualization 2218
53.4 Commentaries and Discussion 2225
  Problems 2226
  References 2227

54 Decision Trees 2229
54.1 Trees and Attributes 2229
54.2 Selecting Attributes 2233
54.3 Constructing a Tree 2243
54.4 Commentaries and Discussion 2251
  Problems 2253
  References 2254

55 Naïve Bayes Classifier 2257
55.1 Independence Condition 2257
55.2 Modeling the Conditional Distribution 2259
55.3 Estimating the Priors 2260
55.4 Gaussian Naïve Classifier 2267
55.5 Commentaries and Discussion 2268
  Problems 2270
  References 2272

56 Linear Discriminant Analysis 2273
56.1 Discriminant Functions 2273
56.2 Linear Discriminant Algorithm 2276
56.3 Minimum Distance Classifier 2278
56.4 Fisher Discriminant Analysis 2281
56.5 Commentaries and Discussion 2294
  Problems 2295
  References 2297

57 Principal Component Analysis 2299
57.1 Data Preprocessing 2299
57.2 Dimensionality Reduction 2301
57.3 Subspace Interpretations 2312
Contents

57.4 Sparse PCA 2315
57.5 Probabilistic PCA 2320
57.6 Commentaries and Discussion 2327
Problems 2330
57.A Maximum-Likelihood Solution 2333
57.B Alternative Optimization Problem 2337
References 2338

58 Dictionary Learning 2340
58.1 Learning under Regularization 2341
58.2 Learning under Constraints 2346
58.3 K-SVD Approach 2348
58.4 Nonnegative Matrix Factorization 2351
58.5 Commentaries and Discussion 2359
Problems 2362
58.A Orthogonal Matching Pursuit 2364
References 2369

59 Logistic Regression 2372
59.1 Logistic Model 2372
59.2 Logistic Empirical Risk 2374
59.3 Multiclass Classification 2379
59.4 Active Learning 2386
59.5 Domain Adaptation 2391
59.6 Commentaries and Discussion 2399
Problems 2403
59.A Generalized Linear Models 2407
References 2411

60 Perceptron 2414
60.1 Linear Separability 2414
60.2 Perceptron Empirical Risk 2416
60.3 Termination in Finite Steps 2422
60.4 Pocket Perceptron 2424
60.5 Commentaries and Discussion 2428
Problems 2432
60.A Counting Theorem 2435
60.B Boolean Functions 2441
References 2443

61 Support Vector Machines 2446
61.1 SVM Empirical Risk 2446
61.2 Convex Quadratic Program 2457
61.3 Cross Validation 2462
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>65.3</td>
<td>Regression and Classification</td>
<td>2642</td>
</tr>
<tr>
<td>65.4</td>
<td>Calculation of Gradient Vectors</td>
<td>2645</td>
</tr>
<tr>
<td>65.5</td>
<td>Backpropagation Algorithm</td>
<td>2653</td>
</tr>
<tr>
<td>65.6</td>
<td>Dropout Strategy</td>
<td>2664</td>
</tr>
<tr>
<td>65.7</td>
<td>Regularized Cross-Entropy Risk</td>
<td>2668</td>
</tr>
<tr>
<td>65.8</td>
<td>Slowdown in Learning</td>
<td>2682</td>
</tr>
<tr>
<td>65.9</td>
<td>Batch Normalization</td>
<td>2683</td>
</tr>
<tr>
<td>65.10</td>
<td>Commentaries and Discussion</td>
<td>2690</td>
</tr>
<tr>
<td>65.A</td>
<td>Derivation of Batch Normalization Algorithm</td>
<td>2701</td>
</tr>
<tr>
<td></td>
<td>Problems</td>
<td>2706</td>
</tr>
<tr>
<td>66</td>
<td>Deep Belief Networks</td>
<td>2711</td>
</tr>
<tr>
<td>66.1</td>
<td>Pre-Training Using Stacked Autoencoders</td>
<td>2711</td>
</tr>
<tr>
<td>66.2</td>
<td>Restricted Boltzmann Machines</td>
<td>2716</td>
</tr>
<tr>
<td>66.3</td>
<td>Contrastive Divergence</td>
<td>2723</td>
</tr>
<tr>
<td>66.4</td>
<td>Pre-Training using Stacked RBMs</td>
<td>2734</td>
</tr>
<tr>
<td>66.5</td>
<td>Deep Generative Model</td>
<td>2737</td>
</tr>
<tr>
<td>66.6</td>
<td>Commentaries and Discussion</td>
<td>2744</td>
</tr>
<tr>
<td></td>
<td>Problems</td>
<td>2748</td>
</tr>
<tr>
<td></td>
<td>References</td>
<td>2750</td>
</tr>
<tr>
<td>67</td>
<td>Convolutional Networks</td>
<td>2752</td>
</tr>
<tr>
<td>67.1</td>
<td>Correlation Layers</td>
<td>2753</td>
</tr>
<tr>
<td>67.2</td>
<td>Pooling</td>
<td>2774</td>
</tr>
<tr>
<td>67.3</td>
<td>Full Network</td>
<td>2783</td>
</tr>
<tr>
<td>67.4</td>
<td>Training Algorithm</td>
<td>2790</td>
</tr>
<tr>
<td>67.5</td>
<td>Commentaries and Discussion</td>
<td>2799</td>
</tr>
<tr>
<td></td>
<td>Problems</td>
<td>2801</td>
</tr>
<tr>
<td>67.A</td>
<td>Derivation of Training Algorithm</td>
<td>2802</td>
</tr>
<tr>
<td></td>
<td>References</td>
<td>2817</td>
</tr>
<tr>
<td>68</td>
<td>Generative Networks</td>
<td>2819</td>
</tr>
<tr>
<td>68.1</td>
<td>Variational Autoencoders</td>
<td>2819</td>
</tr>
<tr>
<td>68.2</td>
<td>Training Variational Autoencoders</td>
<td>2827</td>
</tr>
<tr>
<td>68.3</td>
<td>Conditional Variational Autoencoders</td>
<td>2844</td>
</tr>
<tr>
<td>68.4</td>
<td>Generative Adversarial Networks</td>
<td>2849</td>
</tr>
<tr>
<td>68.5</td>
<td>Training of GANs</td>
<td>2857</td>
</tr>
<tr>
<td>68.6</td>
<td>Conditional GAN</td>
<td>2870</td>
</tr>
<tr>
<td>68.7</td>
<td>Commentaries and Discussion</td>
<td>2874</td>
</tr>
<tr>
<td></td>
<td>Problems</td>
<td>2877</td>
</tr>
<tr>
<td></td>
<td>References</td>
<td>2878</td>
</tr>
<tr>
<td>69</td>
<td>Recurrent Networks</td>
<td>2880</td>
</tr>
</tbody>
</table>
Contents

69.1 Recurrent Neural Networks 2880
69.2 Backpropagation Through Time 2886
69.3 Bidirectional Recurrent Networks 2908
69.4 Vanishing and Exploding Gradients 2915
69.5 Long Short-Term Memory Networks 2917
69.6 Bidirectional LSTMs 2939
69.7 Gated Recurrent Units 2947
69.8 Commentaries and Discussion 2949
  Problems 2950
  References 2953

70 Explainable Learning 2956
70.1 Classifier Model 2956
70.2 Sensitivity Analysis 2960
70.3 Gradient X Input Analysis 2963
70.4 Relevance Analysis 2964
70.5 Commentaries and Discussion 2974
  Problems 2975
  References 2976

71 Adversarial Attacks 2979
71.1 Types of Attacks 2980
71.2 Fast Gradient Sign Method 2984
71.3 Jacobian Saliency Map Approach 2989
71.4 DeepFool Technique 2992
71.5 Black-Box Attacks 3002
71.6 Defense Mechanisms 3005
71.7 Commentaries and Discussion 3007
  Problems 3009
  References 3010

72 Meta Learning 3013
72.1 Network Model 3013
72.2 Siamese Networks 3015
72.3 Relation Networks 3026
72.4 Exploration Models 3032
72.5 Commentaries and Discussion 3051
  Problems 3051
  References 3061

Author Index 3065
Subject Index 3089