

Contents

<i>Preface</i>	<i>page</i> xxii
P.1 Emphasis on Foundations	xxii
P.2 Glimpse of History	xxiv
P.3 Organization of the Text	xxvi
P.4 How to Use the Text	xxix
P.5 Simulation Datasets	xxxii
P.6 Acknowledgments	xxxv
<i>Notation</i>	xl
VOLUME I FOUNDATIONS	1
1 Matrix Theory	3
1.1 Symmetric Matrices	3
1.2 Positive-Definite Matrices	7
1.3 Range Spaces and Nullspaces	9
1.4 Schur Complements	13
1.5 Cholesky Factorization	16
1.6 QR Decomposition	19
1.7 Singular Value Decomposition	21
1.8 Square-Root Matrices	24
1.9 Kronecker Products	26
1.10 Vector and Matrix Norms	31
1.11 Perturbation Bounds on Eigenvalues	39
1.12 Stochastic Matrices	40
1.13 Complex-Valued Matrices	41
1.14 Commentaries and Discussion	43
<i>Problems</i>	49
1.A Proof of Spectral Theorem	52
1.B Constructive Proof of SVD	54
<i>References</i>	55
2 Vector Differentiation	61
2.1 Gradient Vectors	61
2.2 Hessian Matrices	64

2.3	Matrix Differentiation	65
2.4	Commentaries and Discussion	67
	<i>Problems</i>	67
	<i>References</i>	69
3	Random Variables	70
3.1	Probability Density Functions	70
3.2	Mean and Variance	73
3.3	Dependent Random Variables	79
3.4	Random Vectors	95
3.5	Properties of Covariance Matrices	98
3.6	Illustrative Applications	99
3.7	Complex-Valued Variables	108
3.8	Commentaries and Discussion	111
	<i>Problems</i>	114
3.A	Convergence of Random Variables	121
3.B	Concentration Inequalities	124
	<i>References</i>	130
4	Gaussian Distribution	134
4.1	Scalar Gaussian Variables	134
4.2	Vector Gaussian Variables	136
4.3	Useful Gaussian Manipulations	140
4.4	Jointly Distributed Gaussian Variables	146
4.5	Gaussian Processes	152
4.6	Circular Gaussian Distribution	157
4.7	Commentaries and Discussion	159
	<i>Problems</i>	162
	<i>References</i>	167
5	Exponential Distributions	169
5.1	Definition	169
5.2	Special Cases	171
5.3	Useful Properties	180
5.4	Conjugate Priors	185
5.5	Commentaries and Discussion	189
	<i>Problems</i>	190
5.A	Derivation of Properties	194
	<i>References</i>	196
6	Entropy and Divergence	198
6.1	Information and Entropy	198
6.2	Kullback–Leibler Divergence	205
6.3	Maximum Entropy Distribution	211

6.4	Moment Matching	212
6.5	Fisher Information Matrix	215
6.6	Natural Gradients	218
6.7	Evidence Lower Bound	229
6.8	Commentaries and Discussion	233
	<i>Problems</i>	236
	<i>References</i>	239
7	Random Processes	241
7.1	Stationary Processes	241
7.2	Power Spectral Density	246
7.3	Spectral Factorization	253
7.4	Commentaries and Discussion	256
	<i>Problems</i>	258
	<i>References</i>	260
8	Convex Functions	262
8.1	Convex Sets	262
8.2	Convexity	264
8.3	Strict Convexity	266
8.4	Strong Convexity	267
8.5	Hessian Matrix Conditions	269
8.6	Subgradient Vectors	273
8.7	Jensen Inequality	280
8.8	Conjugate Functions	282
8.9	Bregman Divergence	286
8.10	Commentaries and Discussion	291
	<i>Problems</i>	294
	<i>References</i>	300
9	Convex Optimization	303
9.1	Convex Optimization Problems	303
9.2	Equality Constraints	311
9.3	Motivating the KKT Conditions	313
9.4	Projection onto Convex Sets	316
9.5	Commentaries and Discussion	323
	<i>Problems</i>	324
	<i>References</i>	329
10	Lipschitz Conditions	331
10.1	Mean-Value Theorem	331
10.2	δ -Smooth Functions	333
10.3	Commentaries and Discussion	338
	<i>Problems</i>	339

	<i>References</i>	341
11	Proximal Operator	342
	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
	<i>Problems</i>	363
	11.A Convergence under Convexity	367
	11.B Convergence under Strong Convexity	370
	<i>References</i>	373
12	Gradient-Descent Method	376
	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
	<i>Problems</i>	426
	12.A Zeroth-Order Optimization	434
	<i>References</i>	437
13	Conjugate Gradient Method	442
	13.1 Linear Systems of Equations	442
	13.2 Nonlinear Optimization	455
	13.3 Convergence Analysis	460
	13.4 Commentaries and Discussion	466
	<i>Problems</i>	467
	<i>References</i>	470
14	Subgradient Method	472
	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
	<i>Problems</i>	499

14.A	Deterministic Inequality Recursion	502
	<i>References</i>	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
	<i>Problems</i>	542
	<i>References</i>	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
	<i>Problems</i>	587
16.A	Switching Expectation and Differentiation	591
	<i>References</i>	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
	<i>Problems</i>	631
17.A	Regret Analysis for ADAM	633
	<i>References</i>	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
	<i>Problems</i>	676
18.A	Averaging over Mini-Batches	678

18.B	Auxiliary Variance Result	680
	<i>References</i>	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
	<i>Problems</i>	717
19.A	Stochastic Inequality Recursion	721
19.B	Proof of Theorem 19.5	723
	<i>References</i>	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
	<i>Problems</i>	754
	<i>References</i>	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
	<i>Problems</i>	777
	<i>References</i>	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

22.5	Nonsmooth Risk Functions	800
22.6	Commentaries and Discussion	807
	<i>Problems</i>	809
22.A	Proof of Theorem 22.2	811
22.B	Proof of Theorem 22.3	814
	<i>References</i>	816
23	Variance-Reduced Methods II: Random Reshuffling	817
23.1	Amortized Variance-Reduced Gradient Algorithm (AVRG)	817
23.2	Evolution of Memory Variables	819
23.3	Convergence of SAGA	823
23.4	Convergence of AVRG	828
23.5	Convergence of SVRG	831
23.6	Nonsmooth Risk Functions	832
23.7	Commentaries and Discussion	833
	<i>Problems</i>	834
23.A	Proof of Lemma 23.3	835
23.B	Proof of Lemma 23.4	839
23.C	Proof of Theorem 23.1	843
23.D	Proof of Lemma 23.5	846
23.E	Proof of Theorem 23.2	850
	<i>References</i>	852
24	Nonconvex Optimization	853
24.1	First- and Second-Order Stationarity	853
24.2	Stochastic Gradient Optimization	861
24.3	Convergence Behavior	866
24.4	Commentaries and Discussion	873
	<i>Problems</i>	875
24.A	Descent in the Large Gradient Regime	877
24.B	Introducing a Short-Term Model	878
24.C	Descent Away from Strict Saddle Points	889
24.D	Second-Order Convergence Guarantee	898
	<i>References</i>	901
25	Decentralized Optimization I: Primal Methods	903
25.1	Graph Topology	904
25.2	Weight Matrices	910
25.3	Aggregate and Local Risks	914
25.4	Incremental, Consensus, and Diffusion	919
25.5	Formal Derivation as Primal Methods	936
25.6	Commentaries and Discussion	941
	<i>Problems</i>	944
25.A	Proof of Lemma 25.1	948

25.B	Proof of Property (25.71)	950
25.C	Convergence of Primal Algorithms	950
	<i>References</i>	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
	<i>Problems</i>	999
26.A	Convergence of Primal-Dual Algorithms	1001
	<i>References</i>	1007
	VOLUME II INFERENCE	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
	<i>Problems</i>	1045
27.A	Circular Gaussian Distribution	1048
	<i>References</i>	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
	<i>Problems</i>	1076
	<i>References</i>	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
	<i>Problems</i>	1105
29.A	Consistency of Normal Equations	1111
	<i>References</i>	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
	<i>Problems</i>	1165
	<i>References</i>	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
	<i>Problems</i>	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
	<i>References</i>	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
	<i>Problems</i>	1270

	32.A Exponential Mixture Models	1272
	<i>References</i>	1275
33	Predictive Modeling	1278
	33.1 Posterior Distributions	1279
	33.2 Laplace Method	1287
	33.3 Markov Chain Monte Carlo Method	1292
	33.4 Commentaries and Discussion	1305
	<i>Problems</i>	1307
	<i>References</i>	1308
34	Expectation Propagation	1311
	34.1 Factored Representation	1311
	34.2 Gaussian Sites	1316
	34.3 Exponential Sites	1330
	34.4 Assumed Density Filtering	1334
	34.5 Commentaries and Discussion	1337
	<i>Problems</i>	1337
	<i>References</i>	1338
35	Particle Filters	1339
	35.1 Data Model	1339
	35.2 Importance Sampling	1344
	35.3 Particle Filter Implementations	1352
	35.4 Commentaries and Discussion	1359
	<i>Problems</i>	1360
	<i>References</i>	1362
36	Variational Inference	1364
	36.1 Evaluating Evidences	1364
	36.2 Evaluating Posterior Distributions	1370
	36.3 Mean-Field Approximation	1372
	36.4 Exponential Conjugate Models	1398
	36.5 Maximizing the ELBO	1414
	36.6 Stochastic Gradient Solution	1418
	36.7 Black Box Inference	1421
	36.8 Commentaries and Discussion	1427
	<i>Problems</i>	1427
	<i>References</i>	1430
37	Latent Dirichlet Allocation	1432
	37.1 Generative Model	1433
	37.2 Coordinate-Ascent Solution	1442
	37.3 Maximizing the ELBO	1453

	37.4 Estimating Model Parameters	1460
	37.5 Commentaries and Discussion	1474
	<i>Problems</i>	1475
	<i>References</i>	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
	<i>Problems</i>	1517
	<i>References</i>	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
	<i>Problems</i>	1565
	<i>References</i>	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
	<i>Problems</i>	1598
	<i>References</i>	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
	<i>Problems</i>	1639
	<i>References</i>	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
	<i>Problems</i>	1694
	<i>References</i>	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
	<i>Problems</i>	1755
43.A	Proof of Hammersley–Clifford Theorem	1758
43.B	Equivalence of Markovian Properties	1762
	<i>References</i>	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
	<i>Problems</i>	1809
	<i>References</i>	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
	<i>Problems</i>	1859
45.A	Optimal Policy and State-Action Values	1862
45.B	Convergence of Value Iteration	1864
45.C	Proof of ϵ -Optimality	1865
45.D	Convergence of Policy Iteration	1866
45.E	Piecewise Linear Property	1868
45.F	Bellman Principle of Optimality	1869
	<i>References</i>	1873

46	Temporal Difference Learning	1876
	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
	<i>Problems</i>	1917
	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
	<i>References</i>	1928
47	<i>Q</i>-Learning	1930
	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 <i>Q</i> -Learning Algorithm	1941
	47.7 Exploration versus Exploitation	1944
	47.8 <i>Q</i> -Learning with Replay Buffer	1952
	47.9 Double <i>Q</i> -Learning	1953
	47.10 Commentaries and Discussion	1955
	<i>Problems</i>	1958
	47.A Convergence of SARSA(0) Algorithm	1960
	47.B Convergence of <i>Q</i> -Learning Algorithm	1962
	<i>References</i>	1964
48	Value Function Approximation	1967
	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 <i>Q</i> -Learning	1991
	48.6 Commentaries and Discussion	2000
	<i>Problems</i>	2002
	<i>References</i>	2004
49	Policy Gradient Methods	2006
	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
	<i>Problems</i>	2067
49.A	Proof of Policy Gradient Theorem	2071
49.B	Proof of Consistency Theorem	2075
	<i>References</i>	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
	<i>Problems</i>	2120
50.A	Minimum-Norm Solution	2128
50.B	Equivalence in Linear Estimation	2129
50.C	Extended Least-Squares	2130
	<i>References</i>	2135
51	Regularization	2138
51.1	Three Challenges	2139
51.2	ℓ_2 -Regularization	2142
51.3	ℓ_1 -Regularization	2147
51.4	Soft Thresholding	2151
51.5	Commentaries and Discussion	2159
	<i>Problems</i>	2162
51.A	Constrained Formulations for Regularization	2167
51.B	Expression for LASSO Solution	2170
	<i>References</i>	2174
52	Nearest-Neighbor Rule	2176
52.1	Bayes Classifier	2178
52.2	k -NN Classifier	2181
52.3	Performance Guarantee	2184

52.4	<i>k</i> -Means Algorithm	2186
52.5	Commentaries and Discussion	2195
	<i>Problems</i>	2198
52.A	Performance of the NN Classifier	2200
	<i>References</i>	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
	<i>Problems</i>	2226
	<i>References</i>	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
	<i>Problems</i>	2253
	<i>References</i>	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
	<i>Problems</i>	2270
	<i>References</i>	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
	<i>Problems</i>	2295
	<i>References</i>	2297
57	Principal Component Analysis	2299
57.1	Data Preprocessing	2299
57.2	Dimensionality Reduction	2301
57.3	Subspace Interpretations	2312

57.4	Sparse PCA	2315
57.5	Probabilistic PCA	2320
57.6	Commentaries and Discussion	2327
	<i>Problems</i>	2330
57.A	Maximum-Likelihood Solution	2333
57.B	Alternative Optimization Problem	2337
	<i>References</i>	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
	<i>Problems</i>	2362
58.A	Orthogonal Matching Pursuit	2364
	<i>References</i>	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
	<i>Problems</i>	2403
59.A	Generalized Linear Models	2407
	<i>References</i>	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
	<i>Problems</i>	2432
60.A	Counting Theorem	2435
60.B	Boolean Functions	2441
	<i>References</i>	2443
61	Support Vector Machines	2446
61.1	SVM Empirical Risk	2446
61.2	Convex Quadratic Program	2457
61.3	Cross Validation	2462

	61.4	Commentaries and Discussion	2467
		<i>Problems</i>	2469
		<i>References</i>	2470
62		Bagging and Boosting	2473
	62.1	Bagging Classifiers	2473
	62.2	AdaBoost Classifier	2477
	62.3	Gradient Boosting	2488
	62.4	Commentaries and Discussion	2496
		<i>Problems</i>	2497
		<i>References</i>	2500
63		Kernel Methods	2503
	63.1	Motivation	2503
	63.2	Nonlinear Mappings	2506
	63.3	Polynomial and Gaussian Kernels	2508
	63.4	Kernel-Based Perceptron	2511
	63.5	Kernel-Based SVM	2520
	63.6	Kernel-Based Ridge Regression	2526
	63.7	Kernel-Based Learning	2529
	63.8	Kernel PCA	2534
	63.9	Inference under Gaussian Processes	2538
	63.10	Commentaries and Discussion	2549
		<i>Problems</i>	2556
		<i>References</i>	2562
64		Generalization Theory	2565
	64.1	Curse of Dimensionality	2565
	64.2	Empirical Risk Minimization	2569
	64.3	Generalization Ability	2572
	64.4	VC Dimension	2576
	64.5	Bias–Variance Trade-off	2578
	64.6	Surrogate Risk Functions	2582
	64.7	Commentaries and Discussion	2587
		<i>Problems</i>	2594
	64.A	VC Dimension for Linear Classifiers	2601
	64.B	Sauer Lemma	2603
	64.C	Vapnik–Chervonenkis Bound	2609
	64.D	Rademacher Complexity	2615
		<i>References</i>	2625
65		Feedforward Neural Networks	2629
	65.1	Activation Functions	2630
	65.2	Feedforward Networks	2635

65.3	Regression and Classification	2642
65.4	Calculation of Gradient Vectors	2645
65.5	Backpropagation Algorithm	2653
65.6	Dropout Strategy	2664
65.7	Regularized Cross-Entropy Risk	2668
65.8	Slowdown in Learning	2682
65.9	Batch Normalization	2683
65.10	Commentaries and Discussion	2690
	<i>Problems</i>	2695
65.A	Derivation of Batch Normalization Algorithm	2701
	<i>References</i>	2706
66	Deep Belief Networks	2711
66.1	Pre-Training Using Stacked Autoencoders	2711
66.2	Restricted Boltzmann Machines	2716
66.3	Contrastive Divergence	2723
66.4	Pre-Training using Stacked RBMs	2734
66.5	Deep Generative Model	2737
66.6	Commentaries and Discussion	2744
	<i>Problems</i>	2748
	<i>References</i>	2750
67	Convolutional Networks	2752
67.1	Correlation Layers	2753
67.2	Pooling	2774
67.3	Full Network	2783
67.4	Training Algorithm	2790
67.5	Commentaries and Discussion	2799
	<i>Problems</i>	2801
67.A	Derivation of Training Algorithm	2802
	<i>References</i>	2817
68	Generative Networks	2819
68.1	Variational Autoencoders	2819
68.2	Training Variational Autoencoders	2827
68.3	Conditional Variational Autoencoders	2844
68.4	Generative Adversarial Networks	2849
68.5	Training of GANs	2857
68.6	Conditional GAN	2870
68.7	Commentaries and Discussion	2874
	<i>Problems</i>	2877
	<i>References</i>	2878
69	Recurrent Networks	2880

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
	<i>Problems</i>	2950
	<i>References</i>	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
	<i>Problems</i>	2975
	<i>References</i>	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
	<i>Problems</i>	3009
	<i>References</i>	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
	<i>Problems</i>	3051
72.A	Matching Networks	3053
72.B	Prototypical Networks	3059
	<i>References</i>	3061
	<i>Author Index</i>	3065
	<i>Subject Index</i>	3089