Contents

VOLUME I FOUNDATIONS

Preface *page* xxvii
P.1 Emphasis on Foundations xxvii
P.2 Glimpse of History xxix
P.3 Organization of the Text xxxi
P.4 How to Use the Text xxxiv
P.5 Simulation Datasets xxxvii
P.6 Acknowledgments xl
Notation xlv

1 **Matrix Theory** 1
1.1 Symmetric Matrices 1
1.2 Positive-Definite Matrices 5
1.3 Range Spaces and Nullspaces 7
1.4 Schur Complements 11
1.5 Cholesky Factorization 14
1.6 QR Decomposition 18
1.7 Singular Value Decomposition 20
1.8 Square-Root Matrices 22
1.9 Kronecker Products 24
1.10 Vector and Matrix Norms 30
1.11 Perturbation Bounds on Eigenvalues 37
1.12 Stochastic Matrices 38
1.13 Complex-Valued Matrices 39
1.14 Commentaries and Discussion 41
Problems 47
1.A Proof of Spectral Theorem 50
1.B Constructive Proof of SVD 52
References 53

2 **Vector Differentiation** 59
2.1 Gradient Vectors 59
2.2 Hessian Matrices 62

	2.3	Matrix Differentiation	63
	2.4	Commentaries and Discussion	65
		Problems	65
		References	67
3	**Random Variables**		68
	3.1	Probability Density Functions	68
	3.2	Mean and Variance	71
	3.3	Dependent Random Variables	77
	3.4	Random Vectors	93
	3.5	Properties of Covariance Matrices	96
	3.6	Illustrative Applications	97
	3.7	Complex-Valued Variables	106
	3.8	Commentaries and Discussion	109
		Problems	112
	3.A	Convergence of Random Variables	119
	3.B	Concentration Inequalities	122
		References	128
4	**Gaussian Distribution**		132
	4.1	Scalar Gaussian Variables	132
	4.2	Vector Gaussian Variables	134
	4.3	Useful Gaussian Manipulations	138
	4.4	Jointly Distributed Gaussian Variables	144
	4.5	Gaussian Processes	150
	4.6	Circular Gaussian Distribution	155
	4.7	Commentaries and Discussion	157
		Problems	160
		References	165
5	**Exponential Distributions**		167
	5.1	Definition	167
	5.2	Special Cases	169
	5.3	Useful Properties	178
	5.4	Conjugate Priors	183
	5.5	Commentaries and Discussion	187
		Problems	189
	5.A	Derivation of Properties	192
		References	195
6	**Entropy and Divergence**		196
	6.1	Information and Entropy	196
	6.2	Kullback–Leibler Divergence	204
	6.3	Maximum Entropy Distribution	209

	6.4	Moment Matching	211
	6.5	Fisher Information Matrix	213
	6.6	Natural Gradients	217
	6.7	Evidence Lower Bound	227
	6.8	Commentaries and Discussion	231
		Problems	234
		References	237
7	**Random Processes**		240
	7.1	Stationary Processes	240
	7.2	Power Spectral Density	245
	7.3	Spectral Factorization	252
	7.4	Commentaries and Discussion	255
		Problems	257
		References	259
8	**Convex Functions**		261
	8.1	Convex Sets	261
	8.2	Convexity	263
	8.3	Strict Convexity	265
	8.4	Strong Convexity	266
	8.5	Hessian Matrix Conditions	268
	8.6	Subgradient Vectors	272
	8.7	Jensen Inequality	279
	8.8	Conjugate Functions	281
	8.9	Bregman Divergence	285
	8.10	Commentaries and Discussion	290
		Problems	293
		References	299
9	**Convex Optimization**		302
	9.1	Convex Optimization Problems	302
	9.2	Equality Constraints	310
	9.3	Motivating the KKT Conditions	312
	9.4	Projection onto Convex Sets	315
	9.5	Commentaries and Discussion	322
		Problems	323
		References	328
10	**Lipschitz Conditions**		330
	10.1	Mean-Value Theorem	330
	10.2	δ-Smooth Functions	332
	10.3	Commentaries and Discussion	337
		Problems	338
		References	340

11	**Proximal Operator**		341
	11.1	Definition and Properties	341
	11.2	Proximal Point Algorithm	347
	11.3	Proximal Gradient Algorithm	349
	11.4	Convergence Results	354
	11.5	Douglas–Rachford Algorithm	356
	11.6	Commentaries and Discussion	358
		Problems	362
	11.A	Convergence under Convexity	366
	11.B	Convergence under Strong Convexity	369
		References	372
12	**Gradient-Descent Method**		375
	12.1	Empirical and Stochastic Risks	375
	12.2	Conditions on Risk Function	379
	12.3	Constant Step Sizes	381
	12.4	Iteration-Dependent Step-Sizes	392
	12.5	Coordinate-Descent Method	402
	12.6	Alternating Projection Algorithm	413
	12.7	Commentaries and Discussion	418
		Problems	425
	12.A	Zeroth-Order Optimization	433
		References	436
13	**Conjugate Gradient Method**		441
	13.1	Linear Systems of Equations	441
	13.2	Nonlinear Optimization	454
	13.3	Convergence Analysis	459
	13.4	Commentaries and Discussion	465
		Problems	466
		References	469
14	**Subgradient Method**		471
	14.1	Subgradient Algorithm	471
	14.2	Conditions on Risk Function	475
	14.3	Convergence Behavior	479
	14.4	Pocket Variable	483
	14.5	Exponential Smoothing	486
	14.6	Iteration-Dependent Step Sizes	489
	14.7	Coordinate-Descent Algorithms	493
	14.8	Commentaries and Discussion	496
		Problems	498
	14.A	Deterministic Inequality Recursion	501
		References	505

15	**Proximal and Mirror-Descent Methods**	507
	15.1 Proximal Gradient Method	507
	15.2 Projection Gradient Method	515
	15.3 Mirror-Descent Method	519
	15.4 Comparison of Convergence Rates	537
	15.5 Commentaries and Discussion	539
	Problems	541
	References	544
16	**Stochastic Optimization**	547
	16.1 Stochastic Gradient Algorithm	548
	16.2 Stochastic Subgradient Algorithm	565
	16.3 Stochastic Proximal Gradient Algorithm	569
	16.4 Gradient Noise	574
	16.5 Regret Analysis	576
	16.6 Commentaries and Discussion	582
	Problems	586
	16.A Switching Expectation and Differentiation	590
	References	595
17	**Adaptive Gradient Methods**	599
	17.1 Motivation	599
	17.2 AdaGrad Algorithm	603
	17.3 RMSprop Algorithm	608
	17.4 ADAM Algorithm	610
	17.5 Momentum Acceleration Methods	614
	17.6 Federated Learning	619
	17.7 Commentaries and Discussion	626
	Problems	630
	17.A Regret Analysis for ADAM	632
	References	640
18	**Gradient Noise**	642
	18.1 Motivation	642
	18.2 Smooth Risk Functions	645
	18.3 Gradient Noise for Smooth Risks	648
	18.4 Nonsmooth Risk Functions	660
	18.5 Gradient Noise for Nonsmooth Risks	665
	18.6 Commentaries and Discussion	673
	Problems	675
	18.A Averaging over Mini-Batches	677
	18.B Auxiliary Variance Result	679
	References	681

19	**Convergence Analysis I: Stochastic Gradient Algorithms**	683
	19.1 Problem Setting	683
	19.2 Convergence under Uniform Sampling	686
	19.3 Convergence of Mini-Batch Implementation	691
	19.4 Convergence under Vanishing Step Sizes	692
	19.5 Convergence under Random Reshuffling	698
	19.6 Convergence under Importance Sampling	701
	19.7 Convergence of Stochastic Conjugate Gradient	707
	19.8 Commentaries and Discussion	712
	Problems	716
	19.A Stochastic Inequality Recursion	720
	19.B Proof of Theorem 19.5	722
	References	727
20	**Convergence Analysis II: Stochastic Subgradient Algorithms**	730
	20.1 Problem Setting	730
	20.2 Convergence under Uniform Sampling	735
	20.3 Convergence with Pocket Variables	738
	20.4 Convergence with Exponential Smoothing	740
	20.5 Convergence of Mini-Batch Implementation	745
	20.6 Convergence under Vanishing Step Sizes	747
	20.7 Commentaries and Discussion	750
	Problems	753
	References	754
21	**Convergence Analysis III: Stochastic Proximal Algorithms**	756
	21.1 Problem Setting	756
	21.2 Convergence under Uniform Sampling	761
	21.3 Convergence of Mini-Batch Implementation	765
	21.4 Convergence under Vanishing Step Sizes	766
	21.5 Stochastic Projection Gradient	769
	21.6 Mirror-Descent Algorithm	771
	21.7 Commentaries and Discussion	774
	Problems	775
	References	776
22	**Variance-Reduced Methods I: Uniform Sampling**	779
	22.1 Problem Setting	779
	22.2 Naïve Stochastic Gradient Algorithm	782
	22.3 Stochastic Average-Gradient Algorithm (SAGA)	785
	22.4 Stochastic Variance-Reduced Gradient Algorithm (SVRG)	793
	22.5 Nonsmooth Risk Functions	799
	22.6 Commentaries and Discussion	806
	Problems	808

22.A Proof of Theorem 22.2 810
22.B Proof of Theorem 22.3 813
References 815

23 Variance-Reduced Methods II: Random Reshuffling 816
23.1 Amortized Variance-Reduced Gradient Algorithm (AVRG) 816
23.2 Evolution of Memory Variables 818
23.3 Convergence of SAGA 822
23.4 Convergence of AVRG 827
23.5 Convergence of SVRG 830
23.6 Nonsmooth Risk Functions 831
23.7 Commentaries and Discussion 832
Problems 833
23.A Proof of Lemma 23.3 834
23.B Proof of Lemma 23.4 838
23.C Proof of Theorem 23.1 842
23.D Proof of Lemma 23.5 845
23.E Proof of Theorem 23.2 849
References 851

24 Nonconvex Optimization 852
24.1 First- and Second-Order Stationarity 852
24.2 Stochastic Gradient Optimization 860
24.3 Convergence Behavior 865
24.4 Commentaries and Discussion 872
Problems 874
24.A Descent in the Large Gradient Regime 876
24.B Introducing a Short-Term Model 877
24.C Descent Away from Strict Saddle Points 888
24.D Second-Order Convergence Guarantee 897
References 900

25 Decentralized Optimization I: Primal Methods 902
25.1 Graph Topology 903
25.2 Weight Matrices 909
25.3 Aggregate and Local Risks 913
25.4 Incremental, Consensus, and Diffusion 918
25.5 Formal Derivation as Primal Methods 935
25.6 Commentaries and Discussion 940
Problems 943
25.A Proof of Lemma 25.1 947
25.B Proof of Property (25.71) 949
25.C Convergence of Primal Algorithms 949
References 965

26 Decentralized Optimization II: Primal–Dual Methods 969
26.1 Motivation 969
26.2 EXTRA Algorithm 970
26.3 EXACT Diffusion Algorithm 972
26.4 Distributed Inexact Gradient Algorithm 975
26.5 Augmented Decentralized Gradient Method 978
26.6 ATC Tracking Method 979
26.7 Unified Decentralized Algorithm 983
26.8 Convergence Performance 985
26.9 Dual Method 987
26.10 Decentralized Nonconvex Optimization 990
26.11 Commentaries and Discussion 995
Problems 998
26.A Convergence of Primal–Dual Algorithms 1000
References 1006

Author Index 1009
Subject Index 1033

VOLUME II INFERENCE

Preface xxvii
P.1 Emphasis on Foundations xxvii
P.2 Glimpse of History xxix
P.3 Organization of the Text xxxi
P.4 How to Use the Text xxxiv
P.5 Simulation Datasets xxxvii
P.6 Acknowledgments xl
Notation xlv

27 Mean-Square-Error Inference 1053
27.1 Inference without Observations 1054
27.2 Inference with Observations 1057
27.3 Gaussian Random Variables 1066
27.4 Bias–Variance Relation 1072
27.5 Commentaries and Discussion 1082
Problems 1085
27.A Circular Gaussian Distribution 1088
References 1090

28 Bayesian Inference 1092
28.1 Bayesian Formulation 1092
28.2 Maximum A-Posteriori Inference 1094
28.3 Bayes Classifier 1097
28.4 Logistic Regression Inference 1106

28.5 Discriminative and Generative Models 1110
28.6 Commentaries and Discussion 1113
Problems 1116
References 1119

29 Linear Regression 1121
29.1 Regression Model 1121
29.2 Centering and Augmentation 1128
29.3 Vector Estimation 1131
29.4 Linear Models 1134
29.5 Data Fusion 1136
29.6 Minimum-Variance Unbiased Estimation 1139
29.7 Commentaries and Discussion 1143
Problems 1145
29.A Consistency of Normal Equations 1151
References 1153

30 Kalman Filter 1154
30.1 Uncorrelated Observations 1154
30.2 Innovations Process 1157
30.3 State-Space Model 1159
30.4 Measurement- and Time-Update Forms 1171
30.5 Steady-State Filter 1177
30.6 Smoothing Filters 1181
30.7 Ensemble Kalman Filter 1185
30.8 Nonlinear Filtering 1191
30.9 Commentaries and Discussion 1201
Problems 1204
References 1208

31 Maximum Likelihood 1211
31.1 Problem Formulation 1211
31.2 Gaussian Distribution 1214
31.3 Multinomial Distribution 1223
31.4 Exponential Family of Distributions 1226
31.5 Cramer–Rao Lower Bound 1229
31.6 Model Selection 1237
31.7 Commentaries and Discussion 1251
Problems 1259
31.A Derivation of the Cramer–Rao Bound 1265
31.B Derivation of the AIC Formulation 1266
31.C Derivation of the BIC Formulation 1271
References 1273

32 Expectation Maximization 1276
32.1 Motivation 1276
32.2 Derivation of the EM Algorithm 1282
32.3 Gaussian Mixture Models 1287
32.4 Bernoulli Mixture Models 1302
32.5 Commentaries and Discussion 1308
Problems 1310
32.A Exponential Mixture Models 1312
References 1316

33 Predictive Modeling 1319
33.1 Posterior Distributions 1320
33.2 Laplace Method 1328
33.3 Markov Chain Monte Carlo Method 1333
33.4 Commentaries and Discussion 1346
Problems 1348
References 1349

34 Expectation Propagation 1352
34.1 Factored Representation 1352
34.2 Gaussian Sites 1357
34.3 Exponential Sites 1371
34.4 Assumed Density Filtering 1375
34.5 Commentaries and Discussion 1378
Problems 1378
References 1379

35 Particle Filters 1380
35.1 Data Model 1380
35.2 Importance Sampling 1385
35.3 Particle Filter Implementations 1393
35.4 Commentaries and Discussion 1400
Problems 1401
References 1403

36 Variational Inference 1405
36.1 Evaluating Evidences 1405
36.2 Evaluating Posterior Distributions 1411
36.3 Mean-Field Approximation 1413
36.4 Exponential Conjugate Models 1440
36.5 Maximizing the ELBO 1454
36.6 Stochastic Gradient Solution 1458
36.7 Black Box Inference 1461
36.8 Commentaries and Discussion 1467

Problems 1467
References 1470

37 Latent Dirichlet Allocation 1472
37.1 Generative Model 1473
37.2 Coordinate-Ascent Solution 1482
37.3 Maximizing the ELBO 1493
37.4 Estimating Model Parameters 1500
37.5 Commentaries and Discussion 1514
Problems 1515
References 1515

38 Hidden Markov Models 1517
38.1 Gaussian Mixture Models 1517
38.2 Markov Chains 1522
38.3 Forward–Backward Recursions 1538
38.4 Validation and Prediction Tasks 1547
38.5 Commentaries and Discussion 1551
Problems 1557
References 1560

39 Decoding Hidden Markov Models 1563
39.1 Decoding States 1563
39.2 Decoding Transition Probabilities 1565
39.3 Normalization and Scaling 1569
39.4 Viterbi Algorithm 1574
39.5 EM Algorithm for Dependent Observations 1586
39.6 Commentaries and Discussion 1604
Problems 1605
References 1607

40 Independent Component Analysis 1609
40.1 Problem Formulation 1610
40.2 Maximum-Likelihood Formulation 1617
40.3 Mutual Information Formulation 1622
40.4 Maximum Kurtosis Formulation 1627
40.5 Projection Pursuit 1634
40.6 Commentaries and Discussion 1637
Problems 1638
References 1640

41 Bayesian Networks 1643
41.1 Curse of Dimensionality 1644
41.2 Probabilistic Graphical Models 1647

41.3 Active and Blocked Pathways 1661
41.4 Conditional Independence Relations 1670
41.5 Commentaries and Discussion 1677
Problems 1679
References 1680

42 Inference over Graphs 1682
42.1 Probabilistic Inference 1682
42.2 Inference by Enumeration 1685
42.3 Inference by Variable Elimination 1691
42.4 Chow–Liu Algorithm 1698
42.5 Graphical LASSO 1705
42.6 Learning Graph Parameters 1711
42.7 Commentaries and Discussion 1733
Problems 1735
References 1737

43 Undirected Graphs 1740
43.1 Cliques and Potentials 1740
43.2 Representation Theorem 1752
43.3 Factor Graphs 1756
43.4 Message-Passing Algorithms 1761
43.5 Commentaries and Discussion 1793
Problems 1796
43.A Proof of the Hammersley–Clifford Theorem 1799
43.B Equivalence of Markovian Properties 1803
References 1804

44 Markov Decision Processes 1807
44.1 MDP Model 1807
44.2 Discounted Rewards 1821
44.3 Policy Evaluation 1825
44.4 Linear Function Approximation 1840
44.5 Commentaries and Discussion 1848
Problems 1850
References 1851

45 Value and Policy Iterations 1853
45.1 Value Iteration 1853
45.2 Policy Iteration 1866
45.3 Partially Observable MDP 1879
45.4 Commentaries and Discussion 1893
Problems 1900
45.A Optimal Policy and State–Action Values 1903

45.B Convergence of Value Iteration 1905
45.C Proof of ϵ-Optimality 1906
45.D Convergence of Policy Iteration 1907
45.E Piecewise Linear Property 1909
45.F Bellman Principle of Optimality 1910
References 1914

46 Temporal Difference Learning 1917
46.1 Model-Based Learning 1918
46.2 Monte Carlo Policy Evaluation 1920
46.3 TD(0) Algorithm 1928
46.4 Look-Ahead TD Algorithm 1936
46.5 TD(λ) Algorithm 1940
46.6 True Online TD(λ) Algorithm 1949
46.7 Off-Policy Learning 1952
46.8 Commentaries and Discussion 1957
Problems 1958
46.A Useful Convergence Result 1959
46.B Convergence of TD(0) Algorithm 1960
46.C Convergence of TD(λ) Algorithm 1963
46.D Equivalence of Offline Implementations 1967
References 1969

47 Q-Learning 1971
47.1 SARSA(0) Algorithm 1971
47.2 Look-Ahead SARSA Algorithm 1975
47.3 SARSA(λ) Algorithm 1977
47.4 Off-Policy Learning 1979
47.5 Optimal Policy Extraction 1980
47.6 Q-Learning Algorithm 1982
47.7 Exploration versus Exploitation 1985
47.8 Q-Learning with Replay Buffer 1993
47.9 Double Q-Learning 1994
47.10 Commentaries and Discussion 1996
Problems 1999
47.A Convergence of SARSA(0) Algorithm 2001
47.B Convergence of Q-Learning Algorithm 2003
References 2005

48 Value Function Approximation 2008
48.1 Stochastic Gradient TD-Learning 2008
48.2 Least-Squares TD-Learning 2018
48.3 Projected Bellman Learning 2019
48.4 SARSA Methods 2026

48.5 Deep Q-Learning 2032
48.6 Commentaries and Discussion 2041
Problems 2043
References 2045

49 Policy Gradient Methods 2047
49.1 Policy Model 2047
49.2 Finite-Difference Method 2048
49.3 Score Function 2050
49.4 Objective Functions 2052
49.5 Policy Gradient Theorem 2057
49.6 Actor–Critic Algorithms 2059
49.7 Natural Gradient Policy 2071
49.8 Trust Region Policy Optimization 2074
49.9 Deep Reinforcement Learning 2093
49.10 Soft Learning 2098
49.11 Commentaries and Discussion 2106
Problems 2109
49.A Proof of Policy Gradient Theorem 2113
49.B Proof of Consistency Theorem 2117
References 2118

Author Index 2121
Subject Index 2145

VOLUME III LEARNING

Preface xxvii
P.1 Emphasis on Foundations xxvii
P.2 Glimpse of History xxix
P.3 Organization of the Text xxxi
P.4 How to Use the Text xxxiv
P.5 Simulation Datasets xxxvii
P.6 Acknowledgments xl
Notation xlv

50 Least-Squares Problems 2165
50.1 Motivation 2165
50.2 Normal Equations 2170
50.3 Recursive Least-Squares 2187
50.4 Implicit Bias 2195
50.5 Commentaries and Discussion 2197
Problems 2202
50.A Minimum-Norm Solution 2210
50.B Equivalence in Linear Estimation 2211

50.C Extended Least-Squares 2212
References 2217

51 Regularization 2221
51.1 Three Challenges 2222
51.2 ℓ_2-Regularization 2225
51.3 ℓ_1-Regularization 2230
51.4 Soft Thresholding 2234
51.5 Commentaries and Discussion 2242
Problems 2245
51.A Constrained Formulations for Regularization 2250
51.B Expression for LASSO Solution 2253
References 2257

52 Nearest-Neighbor Rule 2260
52.1 Bayes Classifier 2262
52.2 k-NN Classifier 2265
52.3 Performance Guarantee 2268
52.4 k-Means Algorithm 2270
52.5 Commentaries and Discussion 2279
Problems 2282
52.A Performance of the NN Classifier 2284
References 2287

53 Self-Organizing Maps 2290
53.1 Grid Arrangements 2290
53.2 Training Algorithm 2293
53.3 Visualization 2302
53.4 Commentaries and Discussion 2310
Problems 2310
References 2311

54 Decision Trees 2313
54.1 Trees and Attributes 2313
54.2 Selecting Attributes 2317
54.3 Constructing a Tree 2327
54.4 Commentaries and Discussion 2335
Problems 2337
References 2338

55 Naïve Bayes Classifier 2341
55.1 Independence Condition 2341
55.2 Modeling the Conditional Distribution 2343
55.3 Estimating the Priors 2344

55.4 Gaussian Naïve Classifier 2351
55.5 Commentaries and Discussion 2352
Problems 2354
References 2356

56 Linear Discriminant Analysis 2357
56.1 Discriminant Functions 2357
56.2 Linear Discriminant Algorithm 2360
56.3 Minimum Distance Classifier 2362
56.4 Fisher Discriminant Analysis 2365
56.5 Commentaries and Discussion 2378
Problems 2379
References 2381

57 Principal Component Analysis 2383
57.1 Data Preprocessing 2383
57.2 Dimensionality Reduction 2385
57.3 Subspace Interpretations 2396
57.4 Sparse PCA 2399
57.5 Probabilistic PCA 2404
57.6 Commentaries and Discussion 2411
Problems 2414
57.A Maximum Likelihood Solution 2417
57.B Alternative Optimization Problem 2421
References 2422

58 Dictionary Learning 2424
58.1 Learning Under Regularization 2425
58.2 Learning Under Constraints 2430
58.3 K-SVD Approach 2432
58.4 Nonnegative Matrix Factorization 2435
58.5 Commentaries and Discussion 2443
Problems 2446
58.A Orthogonal Matching Pursuit 2448
References 2454

59 Logistic Regression 2457
59.1 Logistic Model 2457
59.2 Logistic Empirical Risk 2459
59.3 Multiclass Classification 2464
59.4 Active Learning 2471
59.5 Domain Adaptation 2476
59.6 Commentaries and Discussion 2484
Problems 2488

59.A Generalized Linear Models 2492
References 2496

60 Perceptron 2499
60.1 Linear Separability 2499
60.2 Perceptron Empirical Risk 2501
60.3 Termination in Finite Steps 2507
60.4 Pocket Perceptron 2509
60.5 Commentaries and Discussion 2513
Problems 2517
60.A Counting Theorem 2520
60.B Boolean Functions 2526
References 2528

61 Support Vector Machines 2530
61.1 SVM Empirical Risk 2530
61.2 Convex Quadratic Program 2541
61.3 Cross Validation 2546
61.4 Commentaries and Discussion 2551
Problems 2553
References 2554

62 Bagging and Boosting 2557
62.1 Bagging Classifiers 2557
62.2 AdaBoost Classifier 2561
62.3 Gradient Boosting 2572
62.4 Commentaries and Discussion 2580
Problems 2581
References 2584

63 Kernel Methods 2587
63.1 Motivation 2587
63.2 Nonlinear Mappings 2590
63.3 Polynomial and Gaussian Kernels 2592
63.4 Kernel-Based Perceptron 2595
63.5 Kernel-Based SVM 2604
63.6 Kernel-Based Ridge Regression 2610
63.7 Kernel-Based Learning 2613
63.8 Kernel PCA 2618
63.9 Inference under Gaussian Processes 2623
63.10 Commentaries and Discussion 2634
Problems 2640
References 2646

64	**Generalization Theory**		2650
	64.1	Curse of Dimensionality	2650
	64.2	Empirical Risk Minimization	2654
	64.3	Generalization Ability	2657
	64.4	VC Dimension	2662
	64.5	Bias–Variance Trade-off	2663
	64.6	Surrogate Risk Functions	2667
	64.7	Commentaries and Discussion	2672
		Problems	2679
	64.A	VC Dimension for Linear Classifiers	2686
	64.B	Sauer Lemma	2688
	64.C	Vapnik–Chervonenkis Bound	2694
	64.D	Rademacher Complexity	2701
		References	2711
65	**Feedforward Neural Networks**		2715
	65.1	Activation Functions	2716
	65.2	Feedforward Networks	2721
	65.3	Regression and Classification	2728
	65.4	Calculation of Gradient Vectors	2731
	65.5	Backpropagation Algorithm	2739
	65.6	Dropout Strategy	2750
	65.7	Regularized Cross-Entropy Risk	2754
	65.8	Slowdown in Learning	2768
	65.9	Batch Normalization	2769
	65.10	Commentaries and Discussion	2776
		Problems	2781
	65.A	Derivation of Batch Normalization Algorithm	2787
		References	2792
66	**Deep Belief Networks**		2797
	66.1	Pre-Training Using Stacked Autoencoders	2797
	66.2	Restricted Boltzmann Machines	2802
	66.3	Contrastive Divergence	2809
	66.4	Pre-Training using Stacked RBMs	2820
	66.5	Deep Generative Model	2823
	66.6	Commentaries and Discussion	2830
		Problems	2834
		References	2836
67	**Convolutional Networks**		2838
	67.1	Correlation Layers	2839
	67.2	Pooling	2860
	67.3	Full Network	2869

	67.4	Training Algorithm	2876
	67.5	Commentaries and Discussion	2885
		Problems	2887
	67.A	Derivation of Training Algorithm	2888
		References	2903
68	**Generative Networks**		2905
	68.1	Variational Autoencoders	2905
	68.2	Training Variational Autoencoders	2913
	68.3	Conditional Variational Autoencoders	2930
	68.4	Generative Adversarial Networks	2935
	68.5	Training of GANs	2943
	68.6	Conditional GANs	2956
	68.7	Commentaries and Discussion	2960
		Problems	2963
		References	2964
69	**Recurrent Networks**		2967
	69.1	Recurrent Neural Networks	2967
	69.2	Backpropagation Through Time	2973
	69.3	Bidirectional Recurrent Networks	2995
	69.4	Vanishing and Exploding Gradients	3002
	69.5	Long Short-Term Memory Networks	3004
	69.6	Bidirectional LSTMs	3026
	69.7	Gated Recurrent Units	3034
	69.8	Commentaries and Discussion	3036
		Problems	3037
		References	3040
70	**Explainable Learning**		3042
	70.1	Classifier Model	3042
	70.2	Sensitivity Analysis	3046
	70.3	Gradient X Input Analysis	3049
	70.4	Relevance Analysis	3050
	70.5	Commentaries and Discussion	3060
		Problems	3061
		References	3062
71	**Adversarial Attacks**		3065
	71.1	Types of Attacks	3066
	71.2	Fast Gradient Sign Method	3070
	71.3	Jacobian Saliency Map Approach	3075
	71.4	DeepFool Technique	3078
	71.5	Black-Box Attacks	3088

71.6 Defense Mechanisms 3091
71.7 Commentaries and Discussion 3093
Problems 3095
References 3096

72 Meta Learning 3099
72.1 Network Model 3099
72.2 Siamese Networks 3101
72.3 Relation Networks 3112
72.4 Exploration Models 3118
72.5 Commentaries and Discussion 3136
Problems 3136
72.A Matching Networks 3138
72.B Prototypical Networks 3144
References 3146

Author Index 3149
Subject Index 3173