

Inference and Learning from Data

Volume I

This extraordinary three-volume work, written in an engaging and rigorous style by a world authority in the field, provides an accessible, comprehensive introduction to the full spectrum of mathematical and statistical techniques underpinning contemporary methods in data-driven learning and inference.

This first volume, *Foundations*, introduces core topics in inference and learning, such as matrix theory, linear algebra, random variables, convex optimization, stochastic optimization, and decentralized methods, and prepares students for studying their practical application in later volumes.

A consistent structure and pedagogy is employed throughout this volume to reinforce student understanding, with over 600 end-of-chapter problems (including solutions for instructors), 180 solved examples, 100 figures, datasets, and downloadable Matlab code. Supported by sister volumes *Inference* and *Learning*, and unique in its scale and depth, this textbook sequence is ideal for early-career researchers and graduate students across many courses in signal processing, machine learning, statistical analysis, data science, and inference.

Ali H. Sayed is Professor and Dean of Engineering at École Polytechnique Fédérale de Lausanne (EPFL), Switzerland. He has also served as Distinguished Professor and Chairman of Electrical Engineering at the University of California, Los Angeles (UCLA), USA, and as President of the IEEE Signal Processing Society. He is a member of the US National Academy of Engineering (NAE) and The World Academy of Sciences (TWAS), and a recipient of several awards, including the 2022 IEEE Fourier Award and the 2020 IEEE Norbert Wiener Society Award. He is a Fellow of the IEEE, EURASIP, and AAAS.

Inference and Learning from Data

Volume I: Foundations

ALI H. SAYED

École Polytechnique Fédérale de Lausanne
University of California at Los Angeles



Shaftesbury Road, Cambridge CB2 8EA, United Kingdom
One Liberty Plaza, 20th Floor, New York, NY 10006, USA
477 Williamstown Road, Port Melbourne, VIC 3207, Australia
314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India
103 Penang Road, #05–06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of Cambridge University Press & Assessment, a department of the University of Cambridge.

We share the University's mission to contribute to society through the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/highereducation/isbn/9781009218122
DOI: 10.1017/9781009218146

© Ali H. Sayed 2023

This publication is in copyright. Subject to statutory exception and to the provisions of relevant collective licensing agreements, no reproduction of any part may take place without the written permission of Cambridge University Press & Assessment.

First published 2023

Printed in the United Kingdom by Bell and Bain Ltd

A catalogue record for this publication is available from the British Library.

ISBN - 3 Volume Set 978-1-009-21810-8 Hardback
ISBN - Volume I 978-1-009-21812-2 Hardback
ISBN - Volume II 978-1-009-21826-9 Hardback
ISBN - Volume III 978-1-009-21828-3 Hardback

Additional resources for this publication at www.cambridge.org/sayed-vol1

Cambridge University Press & Assessment has no responsibility for the persistence or accuracy of URLs for external or third-party internet websites referred to in this publication and does not guarantee that any content on such websites is, or will remain, accurate or appropriate.

In loving memory of my parents

Contents

VOLUME I FOUNDATIONS

	<i>Preface</i>	<i>page</i> xxvii
	P.1 Emphasis on Foundations	xxvii
	P.2 Glimpse of History	xxix
	P.3 Organization of the Text	xxxix
	P.4 How to Use the Text	xxxiv
	P.5 Simulation Datasets	xxxvii
	P.6 Acknowledgments	xl
	<i>Notation</i>	xl
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
	<i>Problems</i>	47
	1.A Proof of Spectral Theorem	50
	1.B Constructive Proof of SVD	52
	<i>References</i>	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
	<i>Problems</i>	65
	<i>References</i>	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
	<i>Problems</i>	112
3.A	Convergence of Random Variables	119
3.B	Concentration Inequalities	122
	<i>References</i>	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
	<i>Problems</i>	160
	<i>References</i>	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
	<i>Problems</i>	189
5.A	Derivation of Properties	192
	<i>References</i>	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
	<i>Problems</i>	234
	<i>References</i>	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
	<i>Problems</i>	257
	<i>References</i>	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
	<i>Problems</i>	293
	<i>References</i>	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
	<i>Problems</i>	323
	<i>References</i>	328
10	Lipschitz Conditions	330
	10.1 Mean-Value Theorem	330
	10.2 δ -Smooth Functions	332
	10.3 Commentaries and Discussion	337
	<i>Problems</i>	338
	<i>References</i>	340

x	Contents	
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
	<i>Problems</i>	362
	11.A Convergence under Convexity	366
	11.B Convergence under Strong Convexity	369
	<i>References</i>	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
	<i>Problems</i>	425
	12.A Zeroth-Order Optimization	433
	<i>References</i>	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
	<i>Problems</i>	466
	<i>References</i>	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
	<i>Problems</i>	498
	14.A Deterministic Inequality Recursion	501
	<i>References</i>	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
	<i>Problems</i>	541
	<i>References</i>	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
	<i>Problems</i>	586
	16.A Switching Expectation and Differentiation	590
	<i>References</i>	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
	<i>Problems</i>	630
	17.A Regret Analysis for ADAM	632
	<i>References</i>	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
	<i>Problems</i>	675
	18.A Averaging over Mini-Batches	677
	18.B Auxiliary Variance Result	679
	<i>References</i>	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
	<i>Problems</i>	716
	19.A Stochastic Inequality Recursion	720
	19.B Proof of Theorem 19.5	722
	<i>References</i>	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
	<i>Problems</i>	753
	<i>References</i>	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
	<i>Problems</i>	775
	<i>References</i>	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
	<i>Problems</i>	808

	22.A Proof of Theorem 22.2	810
	22.B Proof of Theorem 22.3	813
	<i>References</i>	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
	<i>Problems</i>	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
	<i>References</i>	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
	<i>Problems</i>	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
	<i>References</i>	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
	<i>Problems</i>	943
	25.A Proof of Lemma 25.1	947
	25.B Proof of Property (25.71)	949
	25.C Convergence of Primal Algorithms	949
	<i>References</i>	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
	<i>Problems</i>	998
	26.A Convergence of Primal–Dual Algorithms	1000
	<i>References</i>	1006
	<i>Author Index</i>	1009
	<i>Subject Index</i>	1033
	VOLUME II INFERENCE	
	<i>Preface</i>	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
	<i>Notation</i>	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
	<i>Problems</i>	1085
	27.A Circular Gaussian Distribution	1088
	<i>References</i>	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
	<i>Problems</i>	1116
	<i>References</i>	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
	<i>Problems</i>	1145
29.A	Consistency of Normal Equations	1151
	<i>References</i>	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
	<i>Problems</i>	1204
	<i>References</i>	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
	<i>Problems</i>	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
	<i>References</i>	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
	<i>Problems</i>	1310
	32.A Exponential Mixture Models	1312
	<i>References</i>	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
	<i>Problems</i>	1348
	<i>References</i>	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
	<i>Problems</i>	1378
	<i>References</i>	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
	<i>Problems</i>	1401
	<i>References</i>	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

	<i>Problems</i>	1467
	<i>References</i>	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
	<i>Problems</i>	1515
	<i>References</i>	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
	<i>Problems</i>	1557
	<i>References</i>	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
	<i>Problems</i>	1605
	<i>References</i>	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
	<i>Problems</i>	1638
	<i>References</i>	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
	<i>Problems</i>	1679
	<i>References</i>	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
	<i>Problems</i>	1735
	<i>References</i>	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
	<i>Problems</i>	1796
43.A	Proof of the Hammersley–Clifford Theorem	1799
43.B	Equivalence of Markovian Properties	1803
	<i>References</i>	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
	<i>Problems</i>	1850
	<i>References</i>	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
	<i>Problems</i>	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
		<i>References</i>	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
		<i>Problems</i>	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
		<i>References</i>	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
		<i>Problems</i>	1999
	47.A	Convergence of SARSA(0) Algorithm	2001
	47.B	Convergence of Q-Learning Algorithm	2003
		<i>References</i>	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
	<i>Problems</i>	2043
	<i>References</i>	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
	<i>Problems</i>	2109
49.A	Proof of Policy Gradient Theorem	2113
49.B	Proof of Consistency Theorem	2117
	<i>References</i>	2118
	<i>Author Index</i>	2121
	<i>Subject Index</i>	2145
	VOLUME III LEARNING	
	<i>Preface</i>	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
	<i>Notation</i>	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
	<i>Problems</i>	2202
50.A	Minimum-Norm Solution	2210
50.B	Equivalence in Linear Estimation	2211

50.C	Extended Least-Squares	2212
	<i>References</i>	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
	<i>Problems</i>	2245
51.A	Constrained Formulations for Regularization	2250
51.B	Expression for LASSO Solution	2253
	<i>References</i>	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
	<i>Problems</i>	2282
52.A	Performance of the NN Classifier	2284
	<i>References</i>	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
	<i>Problems</i>	2310
	<i>References</i>	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
	<i>Problems</i>	2337
	<i>References</i>	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
	<i>Problems</i>	2354
	<i>References</i>	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
	<i>Problems</i>	2379
	<i>References</i>	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
	<i>Problems</i>	2414
57.A	Maximum Likelihood Solution	2417
57.B	Alternative Optimization Problem	2421
	<i>References</i>	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
	<i>Problems</i>	2446
58.A	Orthogonal Matching Pursuit	2448
	<i>References</i>	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
	<i>Problems</i>	2488

	59.A Generalized Linear Models	2492
	<i>References</i>	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
	<i>Problems</i>	2517
	60.A Counting Theorem	2520
	60.B Boolean Functions	2526
	<i>References</i>	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
	<i>Problems</i>	2553
	<i>References</i>	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
	<i>Problems</i>	2581
	<i>References</i>	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
	<i>Problems</i>	2640
	<i>References</i>	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
	<i>Problems</i>	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
	<i>References</i>	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
	<i>Problems</i>	2781
	65.A Derivation of Batch Normalization Algorithm	2787
	<i>References</i>	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
	<i>Problems</i>	2834
	<i>References</i>	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
	<i>Problems</i>	2887
67.A	Derivation of Training Algorithm	2888
	<i>References</i>	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
	<i>Problems</i>	2963
	<i>References</i>	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
	<i>Problems</i>	3037
	<i>References</i>	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
	<i>Problems</i>	3061
	<i>References</i>	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
	<i>Problems</i>	3095
	<i>References</i>	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
	<i>Problems</i>	3136
72.A	Matching Networks	3138
72.B	Prototypical Networks	3144
	<i>References</i>	3146
	<i>Author Index</i>	3149
	<i>Subject Index</i>	3173

Preface

Learning directly from data is critical to a host of disciplines in engineering and the physical, social, and life sciences. Modern society is literally driven by an interconnected web of data exchanges at rates unseen before, and it relies heavily on decisions inferred from patterns in data. There is nothing fundamentally wrong with this approach, except that the inference and learning methodologies need to be anchored on solid foundations, be fair and reliable in their conclusions, and be robust to unwarranted imperfections and malicious interference.

P.1 EMPHASIS ON FOUNDATIONS

Given the explosive interest in data-driven learning methods, it is not uncommon to encounter claims of superior designs in the literature that are substantiated mainly by sporadic simulations and the potential for “life-changing” applications rather than by an approach that is founded on the well-tested scientific principle to inquiry. For this reason, one of the main objectives of this text is to highlight, in a unified and formal manner, the firm mathematical and statistical pillars that underlie many popular data-driven learning and inference methods. This is a nontrivial task given the wide scope of techniques that exist, and which have often been motivated independently of each other. It is nevertheless important for practitioners and researchers alike to remain cognizant of the common foundational threads that run across these methods. It is also imperative that progress in the domain remains grounded on firm theory. As the aphorism often attributed to Lewin (1945) states, “*there is nothing more practical than a good theory.*” According to Bedeian (2016), this saying has an even older history.

Rigorous data analysis, and conclusions derived from experimentation and theory, have been driving science since time immemorial. As reported by Heath (1912), the Greek scientist Archimedes of Syracuse devised the now famous Archimedes’ Principle about the volume displaced by an immersed object from observing how the level of water in a tub rose when he sat in it. In the account by Hall (1970), Gauss’ formulation of the least-squares problem was driven by his desire to predict the future location of the planetoid Ceres from observations of its location over 41 prior days. There are numerous similar examples by notable scientists where experimentation led to hypotheses and from there

to substantiated theories and well-founded design methodologies. Science is also full of progress in the reverse direction, where theories have been developed first to be validated only decades later through experimentation and data analysis. Einstein (1916) postulated the existence of gravitational waves over 100 years ago. It took until 2016 to detect them! Regardless of which direction one follows, experimentation to theory or the reverse, the match between solid theory and rigorous data analysis has enabled science and humanity to march confidently toward the immense progress that permeates our modern world today.

For similar reasons, data-driven learning and inference should be developed with strong theoretical guarantees. Otherwise, the confidence in their reliability can be shaken if there is over-reliance on “proof by simulation or experience.” Whenever possible, we explain the underlying models and statistical theories for a large number of methods covered in this text. A good grasp of these theories will enable practitioners and researchers to devise variations with greater mastery. We weave through the foundations in a coherent and cohesive manner, and show how the various methods blend together techniques that may appear decoupled but are actually facets of the same common methodology. In this process, we discover that a good number of techniques are well-grounded and meet proven performance guarantees, while other methods are driven by ingenious insights but lack solid justifications and cannot be guaranteed to be “fail-proof.”

Researchers on learning and inference methods are of course aware of the limitations of some of their approaches, so much so that we encounter today many studies, for example, on the topic of “explainable machine learning.” The objective here is to understand why learning algorithms produce certain recommendations. While this is an important area of inquiry, it nevertheless highlights one interesting shift in paradigm. In the past, the emphasis would have been on designing inference methods that respond to the input data in certain desirable and controllable ways. Today, in many instances, the emphasis is to stick to the available algorithms (often, out of convenience) and try to understand or explain why they are responding in certain ways to the input!

Writing this text has been a rewarding journey that took me from the early days of statistical mathematical theory to the modern state of affairs in learning theory. One can only stand in awe at the wondrous ideas that have been introduced by notable researchers along this trajectory. At the same time, one observes with some concern an emerging trend in recent years where solid foundations receive less attention in lieu of “speed publishing” and over-reliance on “illustration by simulation.” This is of course not the norm and most researchers in the field stay honest to the scientific approach to inquiry and design. After concluding this comprehensive text, I stand humbled at the realization of “*how little we know!*” There are countless questions that remain open, and even for many of the questions that have been answered, their answers rely on assumptions or (over)simplifications. It is understandable that the complexity of the problems we face today has increased manifold, and ingenious approximations become necessary to enable tractable solutions.

P.2 GLIMPSE OF HISTORY

Reading through the text, the alert reader will quickly realize that the core foundations of modern-day machine learning, data analytics, and inference methods date back for at least two centuries, with contributions arising from a range of fields including mathematics, statistics, optimization theory, information theory, signal processing, communications, control, and computer science. For the benefit of the reader, I reproduce here with permission from IEEE some historical remarks from the editorial I published in Sayed (2018). I explained there that these disciplines have generated a string of “big ideas” that are driving today multi-faceted efforts in the age of “big data” and machine learning. Generations of students in the statistical sciences and engineering have been trained in the art of modeling, problem solving, and optimization. Their algorithms power everything from cell phones, to spacecraft, robotic explorers, imaging devices, automated systems, computing machines, and also recommender systems. These students mastered the foundations of their fields and have been well prepared to contribute to the explosive growth of data analysis and machine learning solutions.

As the list below shows, many well-known engineering and statistical methods have actually been motivated by data-driven inquiries, even from times remote. The list is a tour of some older historical contributions, which is of course biased by my personal preferences and is not intended to be exhaustive. It is only meant to illustrate how concepts from statistics and the information sciences have always been at the center of promoting big ideas for data and machine learning. Readers will encounter these concepts in various chapters in the text. Readers will also encounter additional historical accounts in the concluding remarks of each chapter, and in particular comments on newer contributions and contributors.

Let me start with Gauss himself, who in 1795 at the young age of 18, was fitting lines and hyperplanes to astronomical data and invented the least-squares criterion for regression analysis – see the collection of his works in Gauss (1903). He even devised the recursive least-squares solution to address what was a “big” data problem for him at the time: He had to avoid tedious repeated calculations by hand as more observational data became available. What a wonderful big idea for a data-driven problem! Of course, Gauss had many other big ideas.

de Moivre (1730), Laplace (1812), and Lyapunov (1901) worked on the central limit theorem. The theorem deals with the limiting distribution of averages of “large” amounts of data. The result is also related to the law of “large” numbers, which even has the qualification “large” in its name. Again, big ideas motivated by “large” data problems.

Bayes (ca mid-1750s) and Laplace (1774) appear to have independently discovered the Bayes rule, which updates probabilities conditioned on observations – see the article by Bayes and Price (1763). The rule forms the backbone of much of statistical signal analysis, Bayes classifiers, Naïve classifiers, and Bayesian networks. Again, a big idea for data-driven inference.

Fourier (1822), whose tools are at the core of disciplines in the information sciences, developed the phenomenal Fourier representation for signals. It is meant to transform data from one domain to another to facilitate the extraction and visualization of information. A big transformative idea for data.

Forward to modern times. The fast Fourier transform (FFT) is another example of an algorithm driven by challenges posed by data size. Its modern version is due to Cooley and Tukey (1965). Their algorithm revolutionized the field of discrete-time signal processing, and FFT processors have become common components in many modern electronic devices. Even Gauss had a role to play here, having proposed an early version of the algorithm some 160 years before, again motivated by a data-driven problem while trying to fit astronomical data onto trigonometric polynomials. A big idea for a data-driven problem.

Closer to the core of statistical mathematical theory, both Kolmogorov (1939) and Wiener (1949) laid out the foundations of modern statistical signal analysis and optimal prediction methods. Their theories taught us how to extract information optimally from data, leading to further refinements by Wiener's student Levinson (1947) and more dramatically by Kalman (1960). The innovations approach by Kailath (1968) exploited to great effect the concept of orthogonalization of the data and recursive constructions. The Kalman filter is applied across many domains today, including in financial analysis from market data. Kalman's work was an outgrowth of the model-based approach to system theory advanced by Zadeh (1954). The concept of a recursive solution from streaming data was a novelty in Kalman's filter; the same concept is commonplace today in most online learning techniques. Again, big ideas for recursive inference from data.

Cauchy (1847) early on, and Robbins and Monro (1951) a century later, developed the powerful gradient-descent method for root finding, which is also recursive in nature. Their techniques have grown to motivate huge advances in stochastic approximation theory. Notable contributions that followed include the work by Rosenblatt (1957) on the perceptron algorithm for single-layer networks, and the impactful delta rule by Widrow and Hoff (1960), widely known as the LMS algorithm in the signal processing literature. Subsequent work on multilayer neural networks grew out of the desire to increase the approximation power of single-layer networks, culminating with the backpropagation method of Werbos (1974). Many of these techniques form the backbone of modern learning algorithms. Again, big ideas for recursive online learning.

Shannon (1948a, b) contributed fundamental insights to data representation, sampling, coding, and communications. His concepts of entropy and information measure helped quantify the amount of uncertainty in data and are used, among other areas, in the design of decision trees for classification purposes and in driving learning algorithms for neural networks. Nyquist (1928) contributed to the understanding of data representations as well. Big ideas for data sampling and data manipulation.

Bellman (1957a, b), a towering system-theorist, introduced dynamic programming and the notion of the curse of dimensionality, both of which are core

underpinnings of many results in learning theory, reinforcement learning, and the theory of Markov decision processes. Viterbi's algorithm (1967) is one notable example of a dynamic programming solution, which has revolutionized communications and has also found applications in hidden Markov models widely used in speech recognition nowadays. Big ideas for conquering complex data problems by dividing them into simpler problems.

Kernel methods, building on foundational results by Mercer (1909) and Aron-szajn (1950), have found widespread applications in learning theory since the mid-1960s with the introduction of the kernel perceptron algorithm. They have also been widely used in estimation theory by Parzen (1962), Kailath (1971), and others. Again, a big idea for learning from data.

Pearson and Fisher launched the modern field of mathematical statistical signal analysis with the introduction of methods such as principal component analysis (PCA) by Pearson (1901) and maximum likelihood and linear discriminant analysis by Fisher (1912, 1922, 1925). These methods are at the core of statistical signal processing. Pearson (1894, 1896) also had one of the earliest studies of fitting a mixture of Gaussian models to biological data. Mixture models have now become an important tool in modern learning algorithms. Big ideas for data-driven inference.

Markov (1913) introduced the formalism of Markov chains, which is widely used today as a powerful modeling tool in a variety of fields including word and speech recognition, handwriting recognition, natural language processing, spam filtering, gene analysis, and web search. Markov chains are also used in Google's PageRank algorithm. Markov's motivation was to study letter patterns in texts. He laboriously went through the first 20,000 letters of a classical Russian novel and counted pairs of vowels, consonants, vowels followed by a consonant, and consonants followed by a vowel. A "big" data problem for his time. Great ideas (and great patience) for data-driven inquiries.

And the list goes on, with many modern-day and ongoing contributions by statisticians, engineers, and computer scientists to network science, distributed processing, compressed sensing, randomized algorithms, optimization, multi-agent systems, intelligent systems, computational imaging, speech processing, forensics, computer visions, privacy and security, and so forth. We provide additional historical accounts about these contributions and contributors at the end of the chapters.

P.3 ORGANIZATION OF THE TEXT

The text is organized into three volumes, with a sizable number of problems and solved examples. The table of contents provides details on what is covered in each volume. Here we provide a condensed summary listing the three main themes:

1. (**Volume I: Foundations**). The first volume covers the *foundations* needed for a solid grasp of inference and learning methods. Many important topics are covered in this part, in a manner that prepares readers for the study of inference and learning methods in the second and third volumes. Topics include: matrix theory, linear algebra, random variables, Gaussian and exponential distributions, entropy and divergence, Lipschitz conditions, convexity, convex optimization, proximal operators, gradient-descent, mirror-descent, conjugate-gradient, subgradient methods, stochastic optimization, adaptive gradient methods, variance-reduced methods, distributed optimization, and nonconvex optimization. Interestingly enough, the following concepts occur time and again in all three volumes and the reader is well-advised to develop familiarity with them: convexity, sample mean and law of large numbers, Gaussianity, Bayes rule, entropy, Kullback–Leibler divergence, gradient-descent, least squares, regularization, and maximum-likelihood. The last three concepts are discussed in the initial chapters of the second volume.
2. (**Volume II: Inference**). The second volume covers inference methods. By “inference” we mean techniques that infer some unknown variable or quantity from observations. The difference we make between “inference” and “learning” in our treatment is that inference methods will target situations where some prior information is known about the underlying signal models or signal distributions (such as their joint probability density functions or generative models). The performance by many of these inference methods will be the ultimate goal that learning algorithms, studied in the third volume, will attempt to emulate. Topics covered here include: mean-square-error inference, Bayesian inference, maximum-likelihood estimation, expectation maximization, expectation propagation, Kalman filters, particle filters, posterior modeling and prediction, Markov chain Monte Carlo methods, sampling methods, variational inference, latent Dirichlet allocation, hidden Markov models, independent component analysis, Bayesian networks, inference over directed and undirected graphs, Markov decision processes, dynamic programming, and reinforcement learning.
3. (**Volume III: Learning**). The third volume covers learning methods. Here, again, we are interested in inferring some unknown variable or quantity from observations. The difference, however, is that the inference will now be solely data-driven, i.e., based on available data and not on any assumed knowledge about signal distributions or models. The designer is only given a collection of observations that arise from the underlying (unknown) distribution. New phenomena arise related to generalization power, overfitting, and underfitting depending on how representative the data is and how complex or simple the approximate models are. The target is to use the data to learn about the quantity of interest (its value or evolution). Topics covered here include: least-squares methods, regularization, nearest-neighbor rule, self-organizing maps, decision trees, naïve Bayes classifier, linear discrimi-