

## Inference and Learning from Data

### Volume II

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 second volume, *Inference*, builds on the foundational topics established in Volume I to introduce students to techniques for inferring unknown variables and quantities, including Bayesian inference, Markov chain Monte Carlo methods, maximum likelihood, variational inference, hidden Markov models, Bayesian networks, and reinforcement learning.

A consistent structure and pedagogy is employed throughout this volume to reinforce student understanding, with over 350 end-of-chapter problems (including solutions for instructors), 180 solved examples, almost 200 figures, datasets, and downloadable Matlab code. Supported by sister volumes *Foundations* 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 II: Inference

ALI H. SAYED

*École Polytechnique Fédérale de Lausanne*

*University of California at Los Angeles*



CAMBRIDGE  
UNIVERSITY PRESS

Cambridge University Press & Assessment  
978-1-009-21826-9 — Inference and Learning from Data  
Volume 2  
Ali H. Sayed  
Frontmatter  
[More Information](#)



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](http://www.cambridge.org)  
Information on this title: [www.cambridge.org/highereducation/isbn/9781009218269](http://www.cambridge.org/highereducation/isbn/9781009218269)  
DOI: 10.1017/9781009218245

© 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-vol2](http://www.cambridge.org/sayed-vol2).

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>	xlv
<b>1</b>	<b>Matrix Theory</b>	<b>1</b>
	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
<b>2</b>	<b>Vector Differentiation</b>	<b>59</b>
	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
<b>3</b>	<b>Random Variables</b>	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
<b>4</b>	<b>Gaussian Distribution</b>	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
<b>5</b>	<b>Exponential Distributions</b>	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
<b>6</b>	<b>Entropy and Divergence</b>	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
<b>7</b>	<b>Random Processes</b>	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
<b>8</b>	<b>Convex Functions</b>	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
<b>9</b>	<b>Convex Optimization</b>	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
<b>10</b>	<b>Lipschitz Conditions</b>	330
10.1	Mean-Value Theorem	330
10.2	$\delta$ -Smooth Functions	332
10.3	Commentaries and Discussion	337
	<i>Problems</i>	338
	<i>References</i>	340

<b>11</b>	<b>Proximal Operator</b>	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
<b>12</b>	<b>Gradient-Descent Method</b>	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
<b>13</b>	<b>Conjugate Gradient Method</b>	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
<b>14</b>	<b>Subgradient Method</b>	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



<b>15</b>	<b>Proximal and Mirror-Descent Methods</b>	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
<b>16</b>	<b>Stochastic Optimization</b>	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
<b>17</b>	<b>Adaptive Gradient Methods</b>	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
<b>18</b>	<b>Gradient Noise</b>	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

<b>19</b>	<b>Convergence Analysis I: Stochastic Gradient Algorithms</b>	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
<b>20</b>	<b>Convergence Analysis II: Stochastic Subgradient Algorithms</b>	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
<b>21</b>	<b>Convergence Analysis III: Stochastic Proximal Algorithms</b>	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
<b>22</b>	<b>Variance-Reduced Methods I: Uniform Sampling</b>	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
<b>23</b>	<b>Variance-Reduced Methods II: Random Reshuffling</b>	<b>816</b>
	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
<b>24</b>	<b>Nonconvex Optimization</b>	<b>852</b>
	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
<b>25</b>	<b>Decentralized Optimization I: Primal Methods</b>	<b>902</b>
	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

<b>26</b>	<b>Decentralized Optimization II: Primal–Dual Methods</b>	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
	<b>VOLUME II INFERENCE</b>	
	<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
<b>27</b>	<b>Mean-Square-Error Inference</b>	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
<b>28</b>	<b>Bayesian Inference</b>	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
<b>29</b>	<b>Linear Regression</b>	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
<b>30</b>	<b>Kalman Filter</b>	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
<b>31</b>	<b>Maximum Likelihood</b>	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

<b>32</b>	<b>Expectation Maximization</b>	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
<b>33</b>	<b>Predictive Modeling</b>	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
<b>34</b>	<b>Expectation Propagation</b>	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
<b>35</b>	<b>Particle Filters</b>	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
<b>36</b>	<b>Variational Inference</b>	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
<b>37</b>	<b>Latent Dirichlet Allocation</b>	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
<b>38</b>	<b>Hidden Markov Models</b>	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
<b>39</b>	<b>Decoding Hidden Markov Models</b>	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
<b>40</b>	<b>Independent Component Analysis</b>	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
<b>41</b>	<b>Bayesian Networks</b>	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
<b>42</b>	<b>Inference over Graphs</b>	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
<b>43</b>	<b>Undirected Graphs</b>	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
<b>44</b>	<b>Markov Decision Processes</b>	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
<b>45</b>	<b>Value and Policy Iterations</b>	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 $\epsilon$ -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
<b>46</b>	<b>Temporal Difference Learning</b>	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( $\lambda$ ) Algorithm	1940
46.6	True Online TD( $\lambda$ ) 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( $\lambda$ ) Algorithm	1963
46.D	Equivalence of Offline Implementations	1967
	<i>References</i>	1969
<b>47</b>	<b>Q-Learning</b>	1971
47.1	SARSA(0) Algorithm	1971
47.2	Look-Ahead SARSA Algorithm	1975
47.3	SARSA( $\lambda$ ) 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
<b>48</b>	<b>Value Function Approximation</b>	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
<b>49</b>	<b>Policy Gradient Methods</b>	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
	<b>VOLUME III LEARNING</b>	
	<i>Preface</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>	xlv
<b>50</b>	<b>Least-Squares Problems</b>	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
<b>51</b>	<b>Regularization</b>		2221
	51.1	Three Challenges	2222
	51.2	$\ell_2$ -Regularization	2225
	51.3	$\ell_1$ -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
<b>52</b>	<b>Nearest-Neighbor Rule</b>		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
<b>53</b>	<b>Self-Organizing Maps</b>		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
<b>54</b>	<b>Decision Trees</b>		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
<b>55</b>	<b>Naïve Bayes Classifier</b>		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
<b>56</b>	<b>Linear Discriminant Analysis</b>	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
<b>57</b>	<b>Principal Component Analysis</b>	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
<b>58</b>	<b>Dictionary Learning</b>	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
<b>59</b>	<b>Logistic Regression</b>	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
<b>60</b>		<b>Perceptron</b>	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
<b>61</b>		<b>Support Vector Machines</b>	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
<b>62</b>		<b>Bagging and Boosting</b>	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
<b>63</b>		<b>Kernel Methods</b>	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

<b>64</b>	<b>Generalization Theory</b>	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
<b>65</b>	<b>Feedforward Neural Networks</b>	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
<b>66</b>	<b>Deep Belief Networks</b>	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
<b>67</b>	<b>Convolutional Networks</b>	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
<b>68</b>	<b>Generative Networks</b>	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
<b>69</b>	<b>Recurrent Networks</b>	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
<b>70</b>	<b>Explainable Learning</b>	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
<b>71</b>	<b>Adversarial Attacks</b>	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
<b>72</b>	<b>Meta Learning</b>	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

**L**earning 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