

Contents

Preface	xiii
I Preliminaries	1
1 Mathematical Preliminaries	3
1.1 Linear Algebra and Differentiable Calculus	3
1.1.1 Minimization of Quadratic Forms	3
1.1.2 Inverting a 2×2 Matrix	4
1.1.3 Inverting Matrices Defined by Blocks, Matrix Inversion Lemma . .	4
1.1.4 Eigenvalue and Singular Value Decomposition	6
1.1.5 Differential Calculus	7
1.2 Concentration Inequalities	7
1.2.1 Hoeffding's Inequality	10
1.2.2 McDiarmid's Inequality	13
1.2.3 Bernstein's Inequality (♦)	14
1.2.4 Expectation of the Maximum	16
1.2.5 Estimation of Expectations through Quadrature (♦♦)	18
1.2.6 Concentration Inequalities for Random Matrices (♦♦)	19
2 Introduction to Supervised Learning	21
2.1 From Training Data to Predictions	22
2.2 Decision Theory	25
2.2.1 Supervised Learning Problems and Loss Functions	25
2.2.2 Risks	27
2.2.3 Bayes Risk and Bayes Predictor	28
2.3 Learning from Data	30
2.3.1 Local Averaging	31
2.3.2 Empirical Risk Minimization	32
2.4 Statistical Learning Theory	36

2.4.1	Measures of Performance	36
2.4.2	Notions of Consistency over Classes of Problems	36
2.5	“No Free Lunch” Theorems (♦)	38
2.6	Quest for Adaptivity	39
2.7	Beyond Supervised Learning	40
2.8	Summary–Book Outline	41
3	Linear Least-Squares Regression	45
3.1	Introduction	45
3.2	Least-Squares Framework	46
3.3	Ordinary Least-Squares Estimator	47
3.3.1	Closed-Form Solution	47
3.3.2	Geometric Interpretation	48
3.3.3	Numerical Resolution	49
3.4	Statistical Analysis of Ordinary Least-Squares	49
3.5	Fixed Design Setting	50
3.5.1	Statistical Properties of the OLS Estimator	52
3.5.2	Experiments	54
3.6	Ridge Least-Squares Regression	56
3.7	Lower Bound (♦)	60
3.8	Random Design Analysis	63
3.8.1	Gaussian Designs	64
3.8.2	General Designs (♦♦)	65
3.9	Principal Component Analysis (♦)	66
3.10	Conclusion	68
II	Generalization Bounds for Learning Algorithms	69
4	Empirical Risk Minimization	71
4.1	Convexification of the Risk	72
4.1.1	Convex Surrogates	73
4.1.2	Geometric Interpretation of the Support Vector Machine (♦)	74
4.1.3	Conditional Φ -risk and Classification Calibration (♦)	76
4.1.4	Relation between Risk and Φ -risk (♦♦)	79
4.2	Risk Minimization Decomposition	84
4.3	Approximation Error	84
4.4	Estimation Error	85
4.4.1	Application of McDiarmid’s Inequality	86
4.4.2	Easy Case I: Quadratic Functions	87
4.4.3	Easy Case II: Finite Number of Models	88
4.4.4	Beyond Finitely Many Models through Covering Numbers (♦)	89
4.5	Rademacher Complexity	91
4.5.1	Symmetrization	92

4.5.2	Lipschitz-Continuous Losses	94
4.5.3	Ball-Constrained Linear Predictions	96
4.5.4	Putting Things Together (Linear Predictions)	97
4.5.5	From Constrained to Regularized Estimation (♦)	98
4.5.6	Extensions and Improvements	102
4.6	Model Selection (♦)	103
4.6.1	Structural Risk Minimization (♦)	104
4.6.2	Selection Based on Validation Set (♦)	104
4.7	Relation with Asymptotic Statistics (♦)	105
4.8	Summary	107
5	Optimization for Machine Learning	109
5.1	Optimization in Machine Learning	109
5.2	Gradient Descent	111
5.2.1	Simplest Analysis: Ordinary Least-Squares	112
5.2.2	Convex Functions and Their Properties	116
5.2.3	Analysis of Gradient Descent for Strongly Convex and Smooth Functions	119
5.2.4	Analysis of Gradient Descent for Convex and Smooth Functions (♦)	124
5.2.5	Beyond Gradient Descent (♦)	126
5.2.6	Nonconvex Objective Functions (♦)	129
5.3	Gradient Methods on Nonsmooth Problems	130
5.4	Stochastic Gradient Descent	134
5.4.1	Strongly Convex Problems (♦)	139
5.4.2	Adaptive Methods (♦)	141
5.4.3	Bias-Variance Trade-offs for Least-Squares (♦)	143
5.4.4	Variance Reduction (♦)	146
5.5	Conclusion	151
6	Local Averaging Methods	155
6.1	Introduction	155
6.2	Local Averaging Methods	157
6.2.1	Linear Estimators	157
6.2.2	Partition Estimators	158
6.2.3	Nearest-Neighbors	160
6.2.4	Nadaraya-Watson Estimator (aka Kernel Regression) (♦)	162
6.3	Generic Simplest Consistency Analysis	163
6.3.1	Fixed Partition	165
6.3.2	k -nearest Neighbor	168
6.3.3	Kernel Regression (Nadaraya-Watson) (♦)	170
6.4	Universal Consistency (♦)	174
6.5	Adaptivity (♦♦)	177
6.6	Conclusion	178

7	Kernel Methods	179
7.1	Introduction	180
7.2	Representer Theorem	181
7.3	Kernels	183
7.3.1	Linear and Polynomial Kernels	186
7.3.2	Translation-Invariant Kernels on $[0, 1]$	187
7.3.3	Translation-Invariant Kernels on $\mathbb{R}^d$	191
7.3.4	Beyond Vectorial Input Spaces (♦)	194
7.4	Algorithms	196
7.4.1	Representer Theorem	196
7.4.2	Column Sampling	197
7.4.3	Random Features	198
7.4.4	Dual Algorithms (♦)	199
7.4.5	Stochastic Gradient Descent (♦)	200
7.4.6	Kernelization of Linear Algorithms	201
7.5	Generalization Guarantees–Lipschitz-continuous Losses	202
7.5.1	Risk Decomposition	203
7.5.2	Approximation Error for Translation-Invariant Kernels on $\mathbb{R}^d$	205
7.6	Theoretical Analysis of Ridge Regression (♦)	208
7.6.1	Kernel Ridge Regression as a Linear Estimator	208
7.6.2	Bias and Variance Decomposition (♦)	209
7.6.3	Relating Empirical and Population Covariance Operators	212
7.6.4	Analysis for Well-Specified Problems (♦)	214
7.6.5	Analysis beyond Well-Specified Problems (♦)	216
7.6.6	Balancing Bias and Variance (♦)	217
7.7	Experiments	218
7.8	Conclusion	220
8	Sparse Methods	221
8.1	Introduction	221
8.1.1	Dedicated Proof Technique for Constrained Least-Squares	223
8.1.2	Probabilistic and Combinatorial Lemmas	224
8.2	Variable Selection by the ℓ_0 -penalty	226
8.2.1	Assuming That k Is Known	226
8.2.2	Sparsity-Adaptive Estimation (Unknown k) (♦)	228
8.3	Variable Selection by ℓ_1 -regularization	231
8.3.1	Intuition and Algorithms	231
8.3.2	Slow Rates–Random Design	234
8.3.3	Slow Rates–Fixed Design (Square Loss)	236
8.3.4	Fast Rates–Fixed Design (♦)	238
8.3.5	Zoo of Conditions (♦♦)	239
8.3.6	Fast Rates–Random Design (♦)	241
8.4	Experiments	243
8.5	Extensions	243

8.6	Conclusion	245
9	Neural Networks	247
9.1	Introduction	247
9.2	Single Hidden-Layer Neural Network	249
9.2.1	Optimization	251
9.2.2	Rectified Linear Units and Homogeneity	253
9.2.3	Estimation Error	253
9.3	Approximation Properties	256
9.3.1	Universal Approximation Property in One Dimension	256
9.3.2	Infinitely Many Neurons and the Variation Norm	257
9.3.3	Variation Norm in One Dimension	260
9.3.4	Variation Norm in an Arbitrary Dimension	263
9.3.5	Precise Approximation Properties	265
9.3.6	From the Variation Norm to a Finite Number of Neurons (♦)	266
9.4	Generalization Performance for Neural Networks	269
9.5	Relationship with Kernel Methods (♦)	271
9.5.1	From a Banach Space $\mathcal{F}_1$ to a Hilbert Space $\mathcal{F}_2$ (♦)	271
9.5.2	Kernel Function (♦♦)	273
9.5.3	Upper Bound on RKHS Norm (♦♦)	275
9.6	Experiments	277
9.7	Extensions	278
9.8	Conclusion	279

III Special Topics 281

10	Ensemble Learning	283
10.1	Averaging/Bagging	284
10.1.1	Independent Datasets	284
10.1.2	Bagging	286
10.2	Random Projections and Averaging	288
10.2.1	Gaussian Sketching	290
10.2.2	Random Projections	292
10.3	Boosting	298
10.3.1	Problem Setup	298
10.3.2	Incremental Learning	301
10.3.3	Matching Pursuit	302
10.3.4	Adaboost	303
10.3.5	Greedy Algorithm Based on Gradient Boosting	304
10.3.6	Convergence of Expected Risk	308
10.3.7	Experiments	310
10.4	Conclusion	311

11 From Online Learning to Bandits	313
11.1 First-Order Online Convex Optimization	315
11.1.1 Convex Case	316
11.1.2 Strongly Convex Case (♦)	318
11.1.3 Online Mirror Descent (♦)	319
11.1.4 Lower Bounds (♦♦)	321
11.2 Zeroth-Order Convex Optimization	323
11.2.1 Smooth Stochastic Gradient Descent	325
11.2.2 Stochastic Smoothing (♦)	328
11.2.3 Extensions	331
11.3 Multiarmed Bandits	331
11.3.1 Need for an Exploration-Exploitation Trade-off	333
11.3.2 “Explore-Then-Commit”	333
11.3.3 Optimism in the Face of Uncertainty (♦)	336
11.3.4 Adversarial Bandits (♦)	339
11.4 Conclusion	341
 12 Overparameterized Models	 343
12.1 Implicit Bias of Gradient Descent	344
12.1.1 Least-Squares Regression	344
12.1.2 Separable Classification	346
12.1.3 Beyond Convex Problems (♦)	351
12.1.4 Remarks on Implicit Bias	354
12.2 Double Descent	355
12.2.1 The Double Descent Phenomenon	355
12.2.2 Empirical Evidence	356
12.2.3 Linear Regression with Gaussian Inputs	358
12.2.4 Linear Regression with Gaussian Projections (♦♦)	360
12.3 Global Convergence of Gradient Descent	365
12.3.1 Mean Field Limits	365
12.3.2 From Linear Networks to Positive-Definite Matrices	370
12.3.3 Global Convergence for Positive-Definite Matrices	370
12.3.4 Special Cases	374
12.4 Lazy Regime and Neural Tangent Kernels (♦)	375
12.5 Conclusion	377
 13 Structured Prediction	 379
13.1 Multicategory Classification	380
13.1.1 Extension of Classical Convex Surrogates	380
13.1.2 Generalization Bound I: Stochastic Gradient Descent	383
13.1.3 Generalization Bound II: Rademacher Complexities (♦)	384
13.2 General Setup and Examples	387
13.2.1 Examples	387
13.2.2 Structure Encoding Loss Functions	390

13.3	Surrogate Methods	391
13.3.1	Score Functions and Decoding Step	392
13.3.2	Fisher Consistency and Calibration Functions	392
13.3.3	Main Surrogate Frameworks	393
13.4	Smooth/Quadratic Surrogates	393
13.4.1	Quadratic Surrogate	393
13.4.2	Theoretical Guarantees	394
13.4.3	Linear Estimators and Decoding Steps	395
13.4.4	Smooth Surrogates (♦)	396
13.5	Max-Margin Formulations	398
13.5.1	Structured Support Vector Machines	399
13.5.2	Max-Min Formulations (♦♦)	399
13.6	Generalization Bounds (♦)	402
13.7	Experiments	404
13.7.1	Robust Regression	404
13.7.2	Ranking	404
13.8	Conclusion	407
14	Probabilistic Methods	409
14.1	From Empirical Risks to Log-Likelihoods	409
14.1.1	Conditional Likelihoods	411
14.1.2	Classical Priors	411
14.1.3	Sparse Priors	412
14.1.4	On the Relationship between MAP and MMSE (♦)	413
14.2	Discriminative versus Generative Models	417
14.2.1	Linear Discriminant Analysis and Softmax Regression	417
14.2.2	Naive Bayes	418
14.2.3	Maximum Likelihood Estimations	419
14.3	Bayesian Inference	420
14.3.1	Computational Handling of Posterior Distributions	421
14.3.2	Model Selection through Marginal Likelihood	422
14.4	PAC-Bayesian Analysis	423
14.4.1	Setup	423
14.4.2	Uniformly Bounded Loss Functions	424
14.5	Conclusion	426
15	Lower Bounds	427
15.1	Statistical Lower Bounds	428
15.1.1	Minimax Lower Bounds	428
15.1.2	Reduction to a Hypothesis Test	429
15.1.3	Review of Information Theory	431
15.1.4	Lower Bound on Hypothesis Testing Based on Information Theory	434
15.1.5	Examples	436
15.1.6	Minimax Lower Bounds through Bayesian Analysis	438

15.2 Optimization Lower Bounds	441
15.2.1 Convex Optimization	441
15.2.2 Nonconvex Optimization (♦)	443
15.3 Lower Bounds for Stochastic Gradient Descent (♦)	447
15.4 Conclusion	449
Conclusion	451
References	453
Index	471