Contents

Preface ix

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 . . . 9
1.2.2 McDiarmid's inequality . . . 12
1.2.3 Bernstein's inequality (♦) . . . 14
1.2.4 Expectation of the maximum . . . 16
1.2.5 Estimation of expectations through quadrature (♦) . . . 17
1.2.6 Concentration inequalities for matrices (♦♦) . . . 18

2 Introduction to supervised learning 21

2.1 From training data to predictions . . . 22
2.2 Decision theory . . . 25
2.2.1 Loss functions . . . 25
2.2.2 Risks . . . 26
2.2.3 Bayes risk and Bayes predictor . . . 27
2.3 Learning from data . . . 30
2.3.1 Local averaging . . . 30
2.3.2 Empirical risk minimization . . . 31
2.4 Statistical learning theory . . . 33
2.4.1 Measures of performance . . . 35
2.4.2 Notions of consistency over classes of problems . . . 35
2.5 No free lunch theorems (♦) . . . 36

2.6 Quest for adaptivity 38
2.7 Beyond supervised learning 39
2.8 Summary - book outline 40

3 Linear least-squares regression **43**
3.1 Introduction 43
3.2 Least-squares framework 44
3.3 Ordinary least-squares (OLS) estimator 45
3.3.1 Closed-form solution 45
3.3.2 Geometric interpretation 46
3.3.3 Numerical resolution 47
3.4 Statistical analysis of OLS 47
3.5 Fixed design setting 48
3.5.1 Statistical properties of the OLS estimator 50
3.5.2 Experiments 52
3.6 Ridge least-squares regression 53
3.7 Lower-bound (♦) 57
3.8 Random design analysis 59
3.8.1 Gaussian designs 61
3.8.2 General designs (♦♦) 61
3.9 Principal component analysis (♦) 63
3.10 Conclusion 65

II Generalization bounds for learning algorithms **67**

4 Empirical risk minimization **69**
4.1 Convexification of the risk 70
4.1.1 Convex surrogates 71
4.1.2 Geometric interpretation of the support vector machine (♦) 72
4.1.3 Conditional Φ-risk and classification calibration (♦) 74
4.1.4 Relationship between risk and Φ-risk (♦♦) 76
4.2 Risk minimization decomposition 80
4.3 Approximation error 80
4.4 Estimation error 81
4.4.1 Application of McDiarmid's inequality 82
4.4.2 Easy case I: quadratic functions 83
4.4.3 Easy case II: Finite number of models 84
4.4.4 Beyond finitely many models through covering numbers (♦) 85
4.5 Rademacher complexity 86
4.5.1 Symmetrization 87
4.5.2 Lipschitz-continuous losses 89
4.5.3 Ball-constrained linear predictions 91
4.5.4 Putting things together (linear predictions) 92
4.5.5 From constrained to regularized estimation (♦) 93

4.5.6 Extensions and improvements . . . 96
4.6 Model selection (♦) . . . 97
4.6.1 Structural risk minimization . . . 98
4.6.2 Selection based on validation set . . . 99
4.7 Relationship with asymptotic statistics (♦) . . . 99
4.8 Summary . . . 101

5 Optimization for machine learning 103
5.1 Optimization in machine learning . . . 103
5.2 Gradient descent . . . 105
5.2.1 Simplest analysis: ordinary least-squares . . . 106
5.2.2 Convex functions and their properties . . . 110
5.2.3 Analysis of GD for strongly convex and smooth functions . . . 112
5.2.4 Analysis of GD for convex and smooth functions (♦) . . . 117
5.2.5 Beyond gradient descent (♦) . . . 120
5.2.6 Non-convex objective functions (♦) . . . 122
5.3 Gradient methods on non-smooth problems . . . 123
5.4 Convergence rate of stochastic gradient descent (SGD) . . . 127
5.4.1 Strongly convex problems (♦) . . . 132
5.4.2 Adaptive methods (♦) . . . 134
5.4.3 Bias-variance trade-offs for least-squares (♦) . . . 135
5.4.4 Variance reduction (♦) . . . 138
5.5 Conclusion . . . 143

6 Local averaging methods 145
6.1 Introduction . . . 145
6.2 Local averaging methods . . . 147
6.2.1 Linear estimators . . . 147
6.2.2 Partition estimators . . . 148
6.2.3 Nearest-neighbors . . . 150
6.2.4 Nadaraya-Watson estimator a.k.a. kernel regression (♦) . . . 151
6.3 Generic “simplest” consistency analysis . . . 153
6.3.1 Fixed partition . . . 155
6.3.2 k-nearest neighbor . . . 158
6.3.3 Kernel regression (Nadaraya-Watson) (♦) . . . 160
6.4 Universal consistency (♦) . . . 163
6.5 Adaptivity (♦♦) . . . 166
6.6 Conclusion . . . 167

7 Kernel methods 169
7.1 Introduction . . . 170
7.2 Representer theorem . . . 170
7.3 Kernels . . . 173
7.3.1 Linear and polynomial kernels . . . 175
7.3.2 Translation-invariant kernels on $[0, 1]$ . . . 177

7.3.3 Translation-invariant kernels on $\mathbb{R}^d$ . . . 179
7.3.4 Beyond vectorial input spaces (♦) . . . 183
7.4 Algorithms . . . 184
7.4.1 Representer theorem . . . 184
7.4.2 Column sampling . . . 185
7.4.3 Random features . . . 186
7.4.4 Dual algorithms (♦) . . . 187
7.4.5 Stochastic gradient descent (♦) . . . 188
7.4.6 “Kernelization” of linear algorithms . . . 188
7.5 Generalization guarantees - Lipschitz-continuous losses . . . 189
7.5.1 Risk decomposition . . . 190
7.5.2 Approximation error for translation-invariant kernels on $\mathbb{R}^d$ . . . 192
7.6 Theoretical analysis of ridge regression (♦) . . . 195
7.6.1 Kernel ridge regression as a “linear” estimator . . . 195
7.6.2 Bias and variance decomposition (♦) . . . 196
7.6.3 Relating empirical and population covariance operators . . . 198
7.6.4 Analysis for well-specified problems (♦) . . . 200
7.6.5 Analysis beyond well-specified problems (♦) . . . 202
7.6.6 Balancing bias and variance (♦) . . . 202
7.7 Experiments . . . 204
7.8 Conclusion . . . 204

8 Sparse methods **207**
8.1 Introduction . . . 207
8.1.1 Dedicated proof technique for constrained least-squares . . . 209
8.1.2 Probabilistic and combinatorial lemmas . . . 210
8.2 Variable selection by the ℓ_0-penalty . . . 212
8.2.1 Assuming k is known . . . 212
8.2.2 Estimating k (♦) . . . 214
8.3 Variable selection by ℓ_1-regularization . . . 216
8.3.1 Intuition and algorithms . . . 217
8.3.2 Slow rates - random design . . . 221
8.3.3 Slow rates - fixed design (square loss) . . . 221
8.3.4 Fast rates (♦) . . . 224
8.3.5 Zoo of conditions (♦♦) . . . 225
8.3.6 Random design (♦) . . . 227
8.4 Experiments . . . 228
8.5 Extensions . . . 229
8.6 Conclusion . . . 230

9 Neural networks **233**
9.1 Introduction . . . 233
9.2 Single hidden layer neural network . . . 235
9.2.1 Optimization . . . 236
9.2.2 Rectified linear units and homogeneity . . . 237

9.2.3 Estimation error . . . 239
9.3 Approximation properties . . . 241
9.3.1 Universal approximation property in one dimension . . . 242
9.3.2 Infinitely many neurons and variation norm . . . 243
9.3.3 Variation norm in one dimension . . . 244
9.3.4 Variation norm in arbitrary dimension . . . 248
9.3.5 Precise approximation properties . . . 249
9.3.6 From the variation norm to a finite number of neurons (♦) . . . 250
9.4 Generalization performance for neural networks . . . 253
9.5 Relationship with kernel methods (♦) . . . 255
9.5.1 From a Banach space $\mathcal{F}_1$ to a Hilbert space $\mathcal{F}_2$ (♦) . . . 255
9.5.2 Kernel function (♦♦) . . . 257
9.5.3 Upper-bound on RKHS norm (♦♦) . . . 258
9.6 Experiments . . . 259
9.7 Extensions . . . 260
9.8 Conclusion . . . 261

III Special topics 263

10 Ensemble learning 265
10.1 Averaging / bagging . . . 266
10.1.1 Independent datasets . . . 266
10.1.2 Bagging . . . 268
10.2 Random projections and averaging . . . 270
10.2.1 Gaussian sketching . . . 271
10.2.2 Random projections . . . 273
10.3 Boosting . . . 278
10.3.1 Problem set-up . . . 278
10.3.2 Incremental learning . . . 281
10.3.3 Matching pursuit . . . 282
10.3.4 Adaboost . . . 283
10.3.5 Greedy algorithm based on gradient boosting . . . 284
10.3.6 Convergence of expected risk . . . 287
10.3.7 Experiments . . . 290
10.4 Conclusion . . . 290

11 From online learning to bandits 291
11.1 First-order online convex optimization . . . 292
11.1.1 Convex case . . . 293
11.1.2 Strongly-convex case (♦) . . . 295
11.1.3 Online mirror descent (♦) . . . 295
11.1.4 Lower bounds (♦♦) . . . 297
11.2 Zero-th order convex optimization . . . 299
11.2.1 Smooth stochastic gradient descent . . . 301

11.2.2 Stochastic smoothing (♦) . . . 303
11.2.3 Extensions . . . 307
11.3 Multi-armed bandits . . . 307
11.3.1 Need for an exploration-exploitation trade-off . . . 308
11.3.2 "Explore-then-commit" . . . 309
11.3.3 Optimism in the face of uncertainty (♦) . . . 310
11.3.4 Adversarial bandits (♦) . . . 313
11.4 Conclusion . . . 315

12 Over-parameterized models **317**
12.1 Implicit bias of gradient descent . . . 318
12.1.1 Least-squares . . . 318
12.1.2 Separable classification . . . 320
12.1.3 Beyond convex problems (♦) . . . 325
12.2 Double descent . . . 327
12.2.1 The double descent phenomenon . . . 327
12.2.2 Empirical evidence . . . 328
12.2.3 Linear regression with Gaussian projections (♦) . . . 329
12.3 Global convergence of gradient descent . . . 334
12.3.1 Mean field limits . . . 335
12.3.2 From linear networks to positive definite matrices . . . 339
12.3.3 Global convergence for positive definite matrices . . . 340
12.3.4 Special case of Oja flow . . . 342
12.4 Lazy regime and neural tangent kernels (♦) . . . 343
12.5 Conclusion . . . 345

13 Structured prediction **347**
13.1 Multi-category classification . . . 348
13.1.1 Extension of classical convex surrogates . . . 348
13.1.2 Generalization bound I: stochastic gradient descent . . . 350
13.1.3 Generalization bound II: Rademacher complexities (♦) . . . 352
13.2 General set-up and examples . . . 354
13.2.1 Examples . . . 355
13.2.2 Structure encoding loss functions . . . 357
13.3 Surrogate methods . . . 358
13.3.1 Score functions and decoding step . . . 359
13.3.2 Fisher consistency and calibration functions . . . 359
13.3.3 Main surrogate frameworks . . . 360
13.4 Smooth/quadratic surrogates . . . 360
13.4.1 Quadratic surrogate . . . 360
13.4.2 Theoretical guarantees . . . 361
13.4.3 Linear estimators and decoding steps . . . 362
13.4.4 Smooth surrogates (♦) . . . 362
13.5 Max-margin formulations . . . 364
13.5.1 Structured SVM . . . 365

13.5.2 Max-min formulations (♦♦) . . . 365
13.6 Generalization bounds (♦) . . . 367
13.7 Experiments . . . 368
13.7.1 Robust regression . . . 368
13.7.2 Ranking . . . 369
13.8 Conclusion . . . 372

14 Probabilistic methods 373
14.1 From empirical risks to log-likelihoods . . . 373
14.1.1 Conditional likelihoods . . . 375
14.1.2 Classical priors . . . 375
14.1.3 Sparse priors . . . 376
14.1.4 On the relationship between MAP and MMSE (♦) . . . 377
14.2 Discriminative vs. generative models . . . 380
14.2.1 Linear discriminant analysis and softmax regression . . . 381
14.2.2 Naive Bayes . . . 381
14.2.3 Maximum likelihood estimations . . . 382
14.3 Bayesian inference . . . 383
14.3.1 Computational handling of posterior distributions . . . 384
14.3.2 Model selection through marginal likelihood . . . 385
14.4 PAC-Bayesian analysis . . . 386
14.4.1 Set-up . . . 386
14.4.2 Uniformly bounded loss functions . . . 387
14.5 Conclusion . . . 389

15 Lower bounds on performance 391
15.1 Statistical lower bounds . . . 392
15.1.1 Minimax lower bounds . . . 392
15.1.2 Reduction to a hypothesis test . . . 393
15.1.3 Review of information theory . . . 395
15.1.4 Lower-bound on hypothesis testing based on information theory . 397
15.1.5 Examples . . . 400
15.1.6 Minimax lower bounds through Bayesian analysis . . . 401
15.2 Optimization lower bounds . . . 404
15.2.1 Convex optimization . . . 404
15.2.2 Non-convex optimization (♦) . . . 406
15.3 Lower bounds for stochastic gradient descent (♦) . . . 409
15.4 Conclusion . . . 411