
Contents

<i>List of illustrations</i>	xv
1 Introduction	1
1.1 Classical versus high-dimensional theory	1
1.2 What can go wrong in high dimensions?	2
1.2.1 Linear discriminant analysis	2
1.2.2 Covariance estimation	5
1.2.3 Nonparametric regression	7
1.3 What can help us in high dimensions?	9
1.3.1 Sparsity in vectors	10
1.3.2 Structure in covariance matrices	11
1.3.3 Structured forms of regression	12
1.4 What is the non-asymptotic viewpoint?	14
1.5 Overview of the book	15
1.5.1 Chapter structure and synopses	15
1.5.2 Recommended background	17
1.5.3 Teaching possibilities and a flow diagram	17
1.6 Bibliographic details and background	19
2 Basic tail and concentration bounds	21
2.1 Classical bounds	21
2.1.1 From Markov to Chernoff	21
2.1.2 Sub-Gaussian variables and Hoeffding bounds	22
2.1.3 Sub-exponential variables and Bernstein bounds	25
2.1.4 Some one-sided results	31
2.2 Martingale-based methods	32
2.2.1 Background	33
2.2.2 Concentration bounds for martingale difference sequences	35
2.3 Lipschitz functions of Gaussian variables	40
2.4 Appendix A: Equivalent versions of sub-Gaussian variables	45
2.5 Appendix B: Equivalent versions of sub-exponential variables	48
2.6 Bibliographic details and background	49
2.7 Exercises	50
3 Concentration of measure	58
3.1 Concentration by entropic techniques	58
3.1.1 Entropy and its properties	58
3.1.2 Herbst argument and its extensions	60

x	<i>Contents</i>	
	3.1.3 Separately convex functions and the entropic method	62
	3.1.4 Tensorization and separately convex functions	64
3.2	A geometric perspective on concentration	67
	3.2.1 Concentration functions	67
	3.2.2 Connection to Lipschitz functions	70
	3.2.3 From geometry to concentration	72
3.3	Wasserstein distances and information inequalities	76
	3.3.1 Wasserstein distances	76
	3.3.2 Transportation cost and concentration inequalities	78
	3.3.3 Tensorization for transportation cost	80
	3.3.4 Transportation cost inequalities for Markov chains	82
	3.3.5 Asymmetric coupling cost	84
3.4	Tail bounds for empirical processes	87
	3.4.1 A functional Hoeffding inequality	87
	3.4.2 A functional Bernstein inequality	89
3.5	Bibliographic details and background	91
3.6	Exercises	92
4	Uniform laws of large numbers	98
4.1	Motivation	98
	4.1.1 Uniform convergence of cumulative distribution functions	98
	4.1.2 Uniform laws for more general function classes	101
4.2	A uniform law via Rademacher complexity	104
	4.2.1 Necessary conditions with Rademacher complexity	107
4.3	Upper bounds on the Rademacher complexity	109
	4.3.1 Classes with polynomial discrimination	109
	4.3.2 Vapnik–Chervonenkis dimension	111
	4.3.3 Controlling the VC dimension	115
4.4	Bibliographic details and background	117
4.5	Exercises	117
5	Metric entropy and its uses	121
5.1	Covering and packing	121
5.2	Gaussian and Rademacher complexity	132
5.3	Metric entropy and sub-Gaussian processes	134
	5.3.1 Upper bound by one-step discretization	135
	5.3.2 Some examples of discretization bounds	137
	5.3.3 Chaining and Dudley’s entropy integral	139
5.4	Some Gaussian comparison inequalities	143
	5.4.1 A general comparison result	143
	5.4.2 Slepian and Sudakov–Fernique inequalities	145
	5.4.3 Gaussian contraction inequality	146
5.5	Sudakov’s lower bound	148
5.6	Chaining and Orlicz processes	150
5.7	Bibliographic details and background	153
5.8	Exercises	154
6	Random matrices and covariance estimation	159
6.1	Some preliminaries	159

Contents

xi

	6.1.1 Notation and basic facts	159
	6.1.2 Set-up of covariance estimation	160
6.2	Wishart matrices and their behavior	161
6.3	Covariance matrices from sub-Gaussian ensembles	165
6.4	Bounds for general matrices	168
	6.4.1 Background on matrix analysis	168
	6.4.2 Tail conditions for matrices	169
	6.4.3 Matrix Chernoff approach and independent decompositions	172
	6.4.4 Upper tail bounds for random matrices	174
	6.4.5 Consequences for covariance matrices	179
6.5	Bounds for structured covariance matrices	180
	6.5.1 Unknown sparsity and thresholding	180
	6.5.2 Approximate sparsity	183
6.6	Appendix: Proof of Theorem 6.1	185
6.7	Bibliographic details and background	188
6.8	Exercises	189
7	Sparse linear models in high dimensions	194
7.1	Problem formulation and applications	194
	7.1.1 Different sparsity models	194
	7.1.2 Applications of sparse linear models	196
7.2	Recovery in the noiseless setting	199
	7.2.1 ℓ_1 -based relaxation	200
	7.2.2 Exact recovery and restricted nullspace	200
	7.2.3 Sufficient conditions for restricted nullspace	202
7.3	Estimation in noisy settings	206
	7.3.1 Restricted eigenvalue condition	207
	7.3.2 Bounds on ℓ_2 -error for hard sparse models	209
	7.3.3 Restricted nullspace and eigenvalues for random designs	213
7.4	Bounds on prediction error	216
7.5	Variable or subset selection	218
	7.5.1 Variable selection consistency for the Lasso	219
	7.5.2 Proof of Theorem 7.21	222
7.6	Appendix: Proof of Theorem 7.16	224
7.7	Bibliographic details and background	227
7.8	Exercises	229
8	Principal component analysis in high dimensions	236
8.1	Principal components and dimension reduction	236
	8.1.1 Interpretations and uses of PCA	237
	8.1.2 Perturbations of eigenvalues and eigenspaces	241
8.2	Bounds for generic eigenvectors	242
	8.2.1 A general deterministic result	242
	8.2.2 Consequences for a spiked ensemble	245
8.3	Sparse principal component analysis	248
	8.3.1 A general deterministic result	249
	8.3.2 Consequences for the spiked model with sparsity	252
8.4	Bibliographic details and background	255
8.5	Exercises	256

9	Decomposability and restricted strong convexity	259
9.1	A general regularized M -estimator	259
9.2	Decomposable regularizers and their utility	269
	9.2.1 Definition and some examples	269
	9.2.2 A key consequence of decomposability	272
9.3	Restricted curvature conditions	276
	9.3.1 Restricted strong convexity	277
9.4	Some general theorems	279
	9.4.1 Guarantees under restricted strong convexity	280
	9.4.2 Bounds under Φ^* -curvature	284
9.5	Bounds for sparse vector regression	286
	9.5.1 Generalized linear models with sparsity	286
	9.5.2 Bounds under restricted strong convexity	287
	9.5.3 Bounds under ℓ_∞ -curvature conditions	288
9.6	Bounds for group-structured sparsity	290
9.7	Bounds for overlapping decomposition-based norms	293
9.8	Techniques for proving restricted strong convexity	297
	9.8.1 Lipschitz cost functions and Rademacher complexity	298
	9.8.2 A one-sided bound via truncation	302
9.9	Appendix: Star-shaped property	306
9.10	Bibliographic details and background	306
9.11	Exercises	307
10	Matrix estimation with rank constraints	312
10.1	Matrix regression and applications	312
10.2	Analysis of nuclear norm regularization	317
	10.2.1 Decomposability and subspaces	317
	10.2.2 Restricted strong convexity and error bounds	319
	10.2.3 Bounds under operator norm curvature	320
10.3	Matrix compressed sensing	321
10.4	Bounds for phase retrieval	326
10.5	Multivariate regression with low-rank constraints	329
10.6	Matrix completion	330
10.7	Additive matrix decompositions	337
10.8	Bibliographic details and background	341
10.9	Exercises	343
11	Graphical models for high-dimensional data	347
11.1	Some basics	347
	11.1.1 Factorization	347
	11.1.2 Conditional independence	350
	11.1.3 Hammersley–Clifford equivalence	351
	11.1.4 Estimation of graphical models	352
11.2	Estimation of Gaussian graphical models	352
	11.2.1 Graphical Lasso: ℓ_1 -regularized maximum likelihood	353
	11.2.2 Neighborhood-based methods	359
11.3	Graphical models in exponential form	365
	11.3.1 A general form of neighborhood regression	366
	11.3.2 Graph selection for Ising models	367

Contents

xiii

11.4	Graphs with corrupted or hidden variables	368
11.4.1	Gaussian graph estimation with corrupted data	368
11.4.2	Gaussian graph selection with hidden variables	373
11.5	Bibliographic details and background	376
11.6	Exercises	378
12	Reproducing kernel Hilbert spaces	383
12.1	Basics of Hilbert spaces	383
12.2	Reproducing kernel Hilbert spaces	385
12.2.1	Positive semidefinite kernel functions	386
12.2.2	Feature maps in $\ell^2(\mathbb{N})$	387
12.2.3	Constructing an RKHS from a kernel	388
12.2.4	A more abstract viewpoint and further examples	390
12.3	Mercer's theorem and its consequences	394
12.4	Operations on reproducing kernel Hilbert spaces	400
12.4.1	Sums of reproducing kernels	400
12.4.2	Tensor products	403
12.5	Interpolation and fitting	405
12.5.1	Function interpolation	405
12.5.2	Fitting via kernel ridge regression	407
12.6	Distances between probability measures	409
12.7	Bibliographic details and background	411
12.8	Exercises	412
13	Nonparametric least squares	416
13.1	Problem set-up	416
13.1.1	Different measures of quality	416
13.1.2	Estimation via constrained least squares	417
13.1.3	Some examples	418
13.2	Bounding the prediction error	420
13.2.1	Bounds via metric entropy	425
13.2.2	Bounds for high-dimensional parametric problems	427
13.2.3	Bounds for nonparametric problems	429
13.2.4	Proof of Theorem 13.5	430
13.3	Oracle inequalities	432
13.3.1	Some examples of oracle inequalities	434
13.3.2	Proof of Theorem 13.13	437
13.4	Regularized estimators	439
13.4.1	Oracle inequalities for regularized estimators	439
13.4.2	Consequences for kernel ridge regression	439
13.4.3	Proof of Corollary 13.18	443
13.4.4	Proof of Theorem 13.17	444
13.5	Bibliographic details and background	448
13.6	Exercises	449
14	Localization and uniform laws	453
14.1	Population and empirical L^2 -norms	453
14.1.1	A uniform law with localization	454
14.1.2	Specialization to kernel classes	458

xiv	<i>Contents</i>	
	14.1.3 Proof of Theorem 14.1	460
14.2	A one-sided uniform law	462
	14.2.1 Consequences for nonparametric least squares	466
	14.2.2 Proof of Theorem 14.12	468
14.3	A uniform law for Lipschitz cost functions	469
	14.3.1 General prediction problems	469
	14.3.2 Uniform law for Lipschitz cost functions	472
14.4	Some consequences for nonparametric density estimation	475
	14.4.1 Density estimation via the nonparametric maximum likelihood estimate	475
	14.4.2 Density estimation via projections	477
14.5	Appendix: Population and empirical Rademacher complexities	480
14.6	Bibliographic details and background	481
14.7	Exercises	482
15	Minimax lower bounds	485
15.1	Basic framework	485
	15.1.1 Minimax risks	486
	15.1.2 From estimation to testing	487
	15.1.3 Some divergence measures	489
15.2	Binary testing and Le Cam's method	491
	15.2.1 Bayes error and total variation distance	491
	15.2.2 Le Cam's convex hull method	497
15.3	Fano's method	500
	15.3.1 Kullback–Leibler divergence and mutual information	501
	15.3.2 Fano lower bound on minimax risk	501
	15.3.3 Bounds based on local packings	503
	15.3.4 Local packings with Gaussian entropy bounds	506
	15.3.5 Yang–Barron version of Fano's method	512
15.4	Appendix: Basic background in information theory	515
15.5	Bibliographic details and background	518
15.6	Exercises	519
	<i>References</i>	524
	<i>Subject index</i>	540
	<i>Author index</i>	548