

Contents

1	Introduction	1
1.1	The framework	1
1.2	The possibilities and challenges	2
1.3	About the book	3
1.3.1	Organization of the book	3
1.4	Some examples	4
1.4.1	Prediction and biomarker discovery in genomics	5
2	Lasso for linear models	7
2.1	Organization of the chapter	7
2.2	Introduction and preliminaries	8
2.2.1	The Lasso estimator	9
2.3	Orthonormal design	10
2.4	Prediction	11
2.4.1	Practical aspects about the Lasso for prediction	12
2.4.2	Some results from asymptotic theory	13
2.5	Variable screening and $\ \hat{\beta} - \beta^0\ _q$ -norms	14
2.5.1	Tuning parameter selection for variable screening	17
2.5.2	Motif regression for DNA binding sites	18
2.6	Variable selection	19
2.6.1	Neighborhood stability and irrepresentable condition	22
2.7	Key properties and corresponding assumptions: a summary	23
2.8	The adaptive Lasso: a two-stage procedure	25
2.8.1	An illustration: simulated data and motif regression	25
2.8.2	Orthonormal design	27
2.8.3	The adaptive Lasso: variable selection under weak conditions	28
2.8.4	Computation	29
2.8.5	Multi-step adaptive Lasso	30
2.8.6	Non-convex penalty functions	32
2.9	Thresholding the Lasso	33
2.10	The relaxed Lasso	34

2.11	Degrees of freedom of the Lasso	34
2.12	Path-following algorithms	36
2.12.1	Coordinatewise optimization and shooting algorithms	38
2.13	Elastic net: an extension	41
	Problems	42
3	Generalized linear models and the Lasso	45
3.1	Organization of the chapter	45
3.2	Introduction and preliminaries	45
3.2.1	The Lasso estimator: penalizing the negative log-likelihood	46
3.3	Important examples of generalized linear models	47
3.3.1	Binary response variable and logistic regression	47
3.3.2	Poisson regression	49
3.3.3	Multi-category response variable and multinomial distribution	50
	Problems	53
4	The group Lasso	55
4.1	Organization of the chapter	55
4.2	Introduction and preliminaries	56
4.2.1	The group Lasso penalty	56
4.3	Factor variables as covariates	58
4.3.1	Prediction of splice sites in DNA sequences	59
4.4	Properties of the group Lasso for generalized linear models	61
4.5	The generalized group Lasso penalty	64
4.5.1	Groupwise prediction penalty and parametrization invariance	65
4.6	The adaptive group Lasso	66
4.7	Algorithms for the group Lasso	67
4.7.1	Block coordinate descent	68
4.7.2	Block coordinate gradient descent	72
	Problems	75
5	Additive models and many smooth univariate functions	77
5.1	Organization of the chapter	77
5.2	Introduction and preliminaries	78
5.2.1	Penalized maximum likelihood for additive models	78
5.3	The sparsity-smoothness penalty	79
5.3.1	Orthogonal basis and diagonal smoothing matrices	80
5.3.2	Natural cubic splines and Sobolev spaces	81
5.3.3	Computation	82
5.4	A sparsity-smoothness penalty of group Lasso type	85
5.4.1	Computational algorithm	86
5.4.2	Alternative approaches	88
5.5	Numerical examples	89
5.5.1	Simulated example	89

5.5.2	Motif regression	90
5.6	Prediction and variable selection	91
5.7	Generalized additive models	92
5.8	Linear model with varying coefficients	93
5.8.1	Properties for prediction	95
5.8.2	Multivariate linear model	95
5.9	Multitask learning	95
	Problems	97
6	Theory for the Lasso	99
6.1	Organization of this chapter	99
6.2	Least squares and the Lasso	101
6.2.1	Introduction	101
6.2.2	The result assuming the truth is linear	102
6.2.3	Linear approximation of the truth	108
6.2.4	A further refinement: handling smallish coefficients	112
6.3	The setup for general convex loss	114
6.4	The margin condition	119
6.5	Generalized linear model without penalty	122
6.6	Consistency of the Lasso for general loss	126
6.7	An oracle inequality	128
6.8	The ℓ_q -error for $1 \leq q \leq 2$	135
6.8.1	Application to least squares assuming the truth is linear	136
6.8.2	Application to general loss and a sparse approximation of the truth	137
6.9	The weighted Lasso	139
6.10	The adaptively weighted Lasso	141
6.11	Concave penalties	144
6.11.1	Sparsity oracle inequalities for least squares with ℓ_r -penalty	146
6.11.2	Proofs for this section (Section 6.11)	147
6.12	Compatibility and (random) matrices	150
6.13	On the compatibility condition	156
6.13.1	Direct bounds for the compatibility constant	158
6.13.2	Bounds using $\ \beta_S\ _1^2 \leq s\ \beta_S\ _2^2$	161
6.13.3	Sets \mathcal{N} containing S	167
6.13.4	Restricted isometry	169
6.13.5	Sparse eigenvalues	170
6.13.6	Further coherence notions	172
6.13.7	An overview of the various eigenvalue flavored constants	174
	Problems	178
7	Variable selection with the Lasso	183
7.1	Introduction	183
7.2	Some results from literature	184
7.3	Organization of this chapter	185

7.4	The beta-min condition	187
7.5	The irrerepresentable condition in the noiseless case	189
7.5.1	Definition of the irrerepresentable condition	190
7.5.2	The KKT conditions	190
7.5.3	Necessity and sufficiency for variable selection	191
7.5.4	The irrerepresentable condition implies the compatibility condition	195
7.5.5	The irrerepresentable condition and restricted regression	197
7.5.6	Selecting a superset of the true active set	199
7.5.7	The weighted irrerepresentable condition	200
7.5.8	The weighted irrerepresentable condition and restricted regression	201
7.5.9	The weighted Lasso with “ideal” weights	203
7.6	Definition of the adaptive and thresholded Lasso	204
7.6.1	Definition of adaptive Lasso	204
7.6.2	Definition of the thresholded Lasso	205
7.6.3	Order symbols	206
7.7	A recollection of the results obtained in Chapter 6	206
7.8	The adaptive Lasso and thresholding: invoking sparse eigenvalues	210
7.8.1	The conditions on the tuning parameters	210
7.8.2	The results	211
7.8.3	Comparison with the Lasso	213
7.8.4	Comparison between adaptive and thresholded Lasso	214
7.8.5	Bounds for the number of false negatives	215
7.8.6	Imposing beta-min conditions	216
7.9	The adaptive Lasso without invoking sparse eigenvalues	218
7.9.1	The condition on the tuning parameter	219
7.9.2	The results	219
7.10	Some concluding remarks	221
7.11	Technical complements for the noiseless case without sparse eigenvalues	222
7.11.1	Prediction error for the noiseless (weighted) Lasso	222
7.11.2	The number of false positives of the noiseless (weighted) Lasso	224
7.11.3	Thresholding the noiseless initial estimator	225
7.11.4	The noiseless adaptive Lasso	227
7.12	Technical complements for the noisy case without sparse eigenvalues	232
7.13	Selection with concave penalties	237
	Problems	241
8	Theory for ℓ_1/ℓ_2-penalty procedures	249
8.1	Introduction	249
8.2	Organization and notation of this chapter	250
8.3	Regression with group structure	252
8.3.1	The loss function and penalty	253

8.3.2	The empirical process	254
8.3.3	The group Lasso compatibility condition	255
8.3.4	A group Lasso sparsity oracle inequality	256
8.3.5	Extensions	258
8.4	High-dimensional additive model	258
8.4.1	The loss function and penalty	258
8.4.2	The empirical process	260
8.4.3	The smoothed Lasso compatibility condition	264
8.4.4	A smoothed group Lasso sparsity oracle inequality	265
8.4.5	On the choice of the penalty	270
8.5	Linear model with time-varying coefficients	275
8.5.1	The loss function and penalty	275
8.5.2	The empirical process	277
8.5.3	The compatibility condition for the time-varying coefficients model	278
8.5.4	A sparsity oracle inequality for the time-varying coefficients model	279
8.6	Multivariate linear model and multitask learning	281
8.6.1	The loss function and penalty	281
8.6.2	The empirical process	282
8.6.3	The multitask compatibility condition	283
8.6.4	A multitask sparsity oracle inequality	284
8.7	The approximation condition for the smoothed group Lasso	286
8.7.1	Sobolev smoothness	286
8.7.2	Diagonalized smoothness	287
	Problems	288
9	Non-convex loss functions and ℓ_1-regularization	293
9.1	Organization of the chapter	293
9.2	Finite mixture of regressions model	294
9.2.1	Finite mixture of Gaussian regressions model	294
9.2.2	ℓ_1 -penalized maximum likelihood estimator	295
9.2.3	Properties of the ℓ_1 -penalized maximum likelihood estimator	299
9.2.4	Selection of the tuning parameters	300
9.2.5	Adaptive ℓ_1 -penalization	301
9.2.6	Riboflavin production with bacillus subtilis	301
9.2.7	Simulated example	303
9.2.8	Numerical optimization	304
9.2.9	GEM algorithm for optimization	304
9.2.10	Proof of Proposition 9.2	308
9.3	Linear mixed effects models	310
9.3.1	The model and ℓ_1 -penalized estimation	311
9.3.2	The Lasso in linear mixed effects models	312
9.3.3	Estimation of the random effects coefficients	312
9.3.4	Selection of the regularization parameter	313

9.3.5	Properties of the Lasso in linear mixed effects models	313
9.3.6	Adaptive ℓ_1 -penalized maximum likelihood estimator	314
9.3.7	Computational algorithm	314
9.3.8	Numerical results	317
9.4	Theory for ℓ_1 -penalization with non-convex negative log-likelihood	320
9.4.1	The setting and notation	320
9.4.2	Oracle inequality for the Lasso for non-convex loss functions	323
9.4.3	Theory for finite mixture of regressions models	326
9.4.4	Theory for linear mixed effects models	329
9.5	Proofs for Section 9.4	332
9.5.1	Proof of Lemma 9.1	332
9.5.2	Proof of Lemma 9.2	333
9.5.3	Proof of Theorem 9.1	335
9.5.4	Proof of Lemma 9.3	337
	Problems	337
10	Stable solutions	339
10.1	Organization of the chapter	339
10.2	Introduction, stability and subsampling	340
10.2.1	Stability paths for linear models	341
10.3	Stability selection	346
10.3.1	Choice of regularization and error control	346
10.4	Numerical results	351
10.5	Extensions	352
10.5.1	Randomized Lasso	352
10.6	Improvements from a theoretical perspective	354
10.7	Proofs	355
10.7.1	Sample splitting	355
10.7.2	Proof of Theorem 10.1	356
	Problems	358
11	P-values for linear models and beyond	359
11.1	Organization of the chapter	359
11.2	Introduction, sample splitting and high-dimensional variable selection	360
11.3	Multi sample splitting and familywise error control	363
11.3.1	Aggregation over multiple p-values	364
11.3.2	Control of familywise error	365
11.4	Multi sample splitting and false discovery rate	367
11.4.1	Control of false discovery rate	368
11.5	Numerical results	369
11.5.1	Simulations and familywise error control	369
11.5.2	Familywise error control for motif regression in computational biology	372
11.5.3	Simulations and false discovery rate control	372

11.6	Consistent variable selection	374
11.6.1	Single sample split method	374
11.6.2	Multi sample split method	377
11.7	Extensions	377
11.7.1	Other models	378
11.7.2	Control of expected false positive selections	378
11.8	Proofs	379
11.8.1	Proof of Proposition 11.1	379
11.8.2	Proof of Theorem 11.1	380
11.8.3	Proof of Theorem 11.2	382
11.8.4	Proof of Proposition 11.2	384
11.8.5	Proof of Lemma 11.3	384
	Problems	386
12	Boosting and greedy algorithms	387
12.1	Organization of the chapter	387
12.2	Introduction and preliminaries	388
12.2.1	Ensemble methods: multiple prediction and aggregation	388
12.2.2	AdaBoost	389
12.3	Gradient boosting: a functional gradient descent algorithm	389
12.3.1	The generic FGD algorithm	390
12.4	Some loss functions and boosting algorithms	392
12.4.1	Regression	392
12.4.2	Binary classification	393
12.4.3	Poisson regression	396
12.4.4	Two important boosting algorithms	396
12.4.5	Other data structures and models	398
12.5	Choosing the base procedure	398
12.5.1	Componentwise linear least squares for generalized linear models	399
12.5.2	Componentwise smoothing spline for additive models	400
12.5.3	Trees	403
12.5.4	The low-variance principle	404
12.5.5	Initialization of boosting	404
12.6	L_2 Boosting	405
12.6.1	Nonparametric curve estimation: some basic insights about boosting	405
12.6.2	L_2 Boosting for high-dimensional linear models	409
12.7	Forward selection and orthogonal matching pursuit	413
12.7.1	Linear models and squared error loss	414
12.8	Proofs	418
12.8.1	Proof of Theorem 12.1	418
12.8.2	Proof of Theorem 12.2	420
12.8.3	Proof of Theorem 12.3	426
	Problems	430

13 Graphical modeling	433
13.1 Organization of the chapter	433
13.2 Preliminaries about graphical models	434
13.3 Undirected graphical models	434
13.3.1 Markov properties for undirected graphs	434
13.4 Gaussian graphical models	435
13.4.1 Penalized estimation for covariance matrix and edge set ...	436
13.4.2 Nodewise regression	440
13.4.3 Covariance estimation based on undirected graph	442
13.5 Ising model for binary random variables	444
13.6 Faithfulness assumption	445
13.6.1 Failure of faithfulness	446
13.6.2 Faithfulness and Gaussian graphical models	448
13.7 The PC-algorithm: an iterative estimation method	449
13.7.1 Population version of the PC-algorithm	449
13.7.2 Sample version for the PC-algorithm	451
13.8 Consistency for high-dimensional data	453
13.8.1 An illustration	455
13.8.2 Theoretical analysis of the PC-algorithm	456
13.9 Back to linear models	462
13.9.1 Partial faithfulness	463
13.9.2 The PC-simple algorithm	465
13.9.3 Numerical results	468
13.9.4 Asymptotic results in high dimensions	471
13.9.5 Correlation screening (sure independence screening)	474
13.9.6 Proofs	475
Problems	480
14 Probability and moment inequalities	481
14.1 Organization of this chapter	481
14.2 Some simple results for a single random variable	482
14.2.1 Sub-exponential random variables	482
14.2.2 Sub-Gaussian random variables	483
14.2.3 Jensen's inequality for partly concave functions	485
14.3 Bernstein's inequality	486
14.4 Hoeffding's inequality	487
14.5 The maximum of p averages	489
14.5.1 Using Bernstein's inequality	489
14.5.2 Using Hoeffding's inequality	491
14.5.3 Having sub-Gaussian random variables	493
14.6 Concentration inequalities	494
14.6.1 Bousquet's inequality	494
14.6.2 Massart's inequality	496
14.6.3 Sub-Gaussian random variables	496
14.7 Symmetrization and contraction	497

14.8 Concentration inequalities for Lipschitz loss functions 500

14.9 Concentration for squared error loss with random design 504

 14.9.1 The inner product of noise and linear functions 505

 14.9.2 Squared linear functions 505

 14.9.3 Squared error loss 508

14.10 Assuming only lower order moments 508

 14.10.1 Nemirovski moment inequality 509

 14.10.2 A uniform inequality for quadratic forms 510

14.11 Using entropy for concentration in the sub-Gaussian case 511

14.12 Some entropy results 516

 14.12.1 Entropy of finite-dimensional spaces and general convex
 hulls 518

 14.12.2 Sets with restrictions on the coefficients 518

 14.12.3 Convex hulls of small sets: entropy with log-term 519

 14.12.4 Convex hulls of small sets: entropy without log-term 520

 14.12.5 Further refinements 523

 14.12.6 An example: functions with $(m - 1)$ -th derivative of
 bounded variation 523

 14.12.7 Proofs for this section (Section 14.12) 525

Problems 535

Author Index 539

Index 543

References 547