

Lecture notes on sparsity

Sara van de Geer

February 2016

These notes contain (parts of) six chapters of “Estimation and Testing under Sparsity” (Springer, to appear).

Contents

1	The Lasso	5
1.1	The linear model with $p < n$	5
1.2	The linear model with $p \geq n$	6
1.3	Notation	7
1.4	The Lasso, KKT and two point inequality	7
1.5	Dual norm and decomposability	9
1.6	Compatibility	9
1.7	A sharp oracle inequality	10
1.8	Including a bound for the ℓ_1 -error and allowing many small values.	11
1.9	The ℓ_1 -restricted oracle	15
1.10	Weak sparsity	16
1.11	Complements	17
1.11.1	An alternative bound for the ℓ_1 -error	17
1.11.2	When there are coefficients left unpenalized	17
1.11.3	A direct proof of Theorem 1.7.1.	18
2	The square-root Lasso	21
2.1	Introduction	21
2.2	KKT and two point inequality for the square-root Lasso	22
2.3	A proposition assuming no overfitting	22
2.4	Showing the square-root Lasso does not overfit	23
2.5	A sharp oracle inequality for the square-root Lasso	25
2.6	A bound for the mean ℓ_1 -error	26
2.7	Comparison with scaled Lasso	27
2.8	The multivariate square-root Lasso	29
3	Structured sparsity	31
3.1	The Ω -structured sparsity estimator	31
3.2	Dual norms and KKT-conditions for structured sparsity	32
3.3	Two point inequality	33
3.4	Weak decomposability and Ω -triangle property	33
3.5	Ω -compatibility	35
3.6	A sharp oracle inequality with structured sparsity	36
3.7	Norms stronger than ℓ_1	37
3.8	Structured sparsity and square-root loss	38
3.8.1	Assuming there is no overfitting	38

3.8.2	Showing there is no overfitting	39
3.8.3	A sharp oracle inequality	39
3.9	Norms generated from cones	40
3.10	Complements	44
3.10.1	The case where some coefficients are not penalized	44
3.10.2	The sorted ℓ_1 -norm	44
3.10.3	A direct proof of Theorem 3.6.1	45
4	Empirical process theory for dual norms	47
4.1	Introduction	47
4.2	The dual norm of ℓ_1 and the scaled version	47
4.3	Dual norms generated from cones	49
4.4	A generalized Bernstein inequality	50
4.5	Bounds for weighted sums of squared Gaussians	51
4.6	The special case of χ^2 -random variables	52
4.7	The wedge dual norm	53
5	General loss with norm-penalty	55
5.1	Introduction	55
5.2	Two point inequality, convex conjugate and two point margin	56
5.3	Triangle property and effective sparsity	58
5.4	Two versions of weak decomposability	60
5.5	A sharp oracle inequality	61
5.6	Localizing (or a non-sharp oracle inequality)	63
6	Some worked-out examples	67
6.1	The Lasso and square-root Lasso completed	67
6.2	Least squares loss with Ω -structured sparsity completed	68
6.3	Logistic regression	71
6.3.1	Logistic regression with fixed, bounded design	72
6.4	Trace regression with nuclear norm penalization	72
6.4.1	Some useful matrix inequalities	73
6.4.2	Dual norm of the nuclear norm and its triangle property	74
6.4.3	An oracle result for trace regression with least squares loss	76
6.4.4	Robust matrix completion	76
6.5	Sparse principal components	78
6.5.1	Two-point margin and two point inequality for sparse PCA	79
6.5.2	Effective sparsity and dual-norm inequality for sparse PCA	81
6.5.3	A sharp oracle inequality for sparse PCA	81

Chapter 1

The Lasso

1.1 The linear model with $p < n$

Let X be an $n \times p$ input matrix and $Y \in \mathbb{R}^n$ be an n -vector of responses. The linear model is

$$Y = X\beta^0 + \epsilon,$$

where $\beta^0 \in \mathbb{R}^p$ is an unknown vector of coefficients and $\epsilon \in \mathbb{R}^n$ is a mean-zero noise vector. This is a standard model in regression and $X\beta^0$ is often called the regression of Y on X . The least squares method, usually credited to Gauss, is to estimate the unknown β^0 by minimizing the Euclidean distance between Y and the space spanned by the columns in X :

$$\hat{\beta}_{\text{LS}} := \arg \min_{\beta \in \mathbb{R}^p} \|Y - X\beta\|_2^2.$$

The least squares estimator $\hat{\beta}_{\text{LS}}$ is thus obtained by taking the coefficients of the projection of Y on the column space of X . If X has full rank p we can write it as

$$\hat{\beta}_{\text{LS}} = (X^T X)^{-1} X^T Y.$$

The estimated regression is then the projection vector

$$X\hat{\beta}_{\text{LS}} = X(X^T X)^{-1} X^T Y.$$

If the entries $\epsilon_1, \dots, \epsilon_n$ of the noise vector ϵ are uncorrelated and have common variance σ_0^2 one may verify that

$$\mathbf{E}\|X(\hat{\beta}_{\text{LS}} - \beta^0)\|_2^2 = \sigma_0^2 p.$$

We refer to the normalized quantity $\|X(\hat{\beta}_{\text{LS}} - \beta^0)\|_2^2/n$ as the *prediction error*: if we use $X\hat{\beta}_{\text{LS}}$ as prediction of a new (unobserved) response vector Y_{new} when the input is X , then on average the squared error made is

$$\mathbf{E}\|Y_{\text{new}} - (X\hat{\beta}_{\text{LS}})\|_2^2/n = \mathbf{E}\|X(\hat{\beta}_{\text{LS}} - \beta^0)\|_2^2/n + \sigma_0^2.$$

The first term in the above right-hand side is due to the estimation of β^0 whereas the second term σ_0^2 is due to the noise in the new observation. We neglect the unavoidable second term in our terminology. The mean prediction error is then

$$\mathbf{E}\|X(\hat{\beta}_{\text{LS}} - \beta^0)\|_2^2/n = \sigma_0^2 \times \frac{p}{n} = \sigma_0^2 \times \frac{\text{number of parameters}}{\text{number of observations}}.$$

In this monograph we are mainly concerned with models where $p > n$ or even $p \gg n$. Clearly, the just described least squares method then breaks down. This chapter studies the so-called Lasso estimator $\hat{\beta}$ when possibly $p > n$. Aim is to show that

$$\|X(\hat{\beta} - \beta^0)\|_2^2/n = \mathcal{O}_{\mathbf{P}}\left(\frac{s_0 \log p}{n}\right) \quad (1.1)$$

where s_0 is the number of non-zero coefficients of β^0 (or the number of in absolute value “large enough” coefficients of β^0). The *active set* $S_0 := \{j : \beta_j^0 \neq 0\}$ is however not assumed to be known, nor its size $s_0 = |S_0|$.

1.2 The linear model with $p \geq n$

Let $Y \in \mathbb{R}^n$ be an n -vector of real-valued observations and let X be a given $n \times p$ design matrix. We concentrate from now on mainly on the high-dimensional situation, which is the situation $p \geq n$ or even $p \gg n$.

Write the expectation of the response Y as

$$f^0 := \mathbf{E}Y.$$

The matrix X is fixed in this chapter, i.e., we consider the case of fixed design. The entries of the vector f^0 are thus the (conditional) expectation of Y given X . Let $\epsilon := Y - f^0$ be the noise term.

The linear model is

$$f^0 = X\beta^0$$

where β^0 is an unknown vector of coefficients. Thus this model assumes there is a solution β^0 of the equation $f^0 = X\beta^0$. In the high-dimensional situation with $\text{rank}(X) = n$ this is always the case: the linear model is never misspecified. When there are several solutions we may take for instance a sparsest solution, that is, a solution with the smallest number of non-zero coefficients. Alternatively one may prefer a basis pursuit solution (Chen et al. [1998])

$$\beta^0 := \arg \min \left\{ \|\beta\|_1 : X\beta = f^0 \right\}$$

where $\|\beta\|_1 := \sum_{j=1}^p |\beta_j|$ denotes the ℓ_1 -norm of the vector β . We do not express in our notation that basis pursuit may not generate a unique solution¹.

¹A suitable notation that expresses the non-uniqueness is $\beta^0 \in \arg \min\{\|\beta\|_1 : X\beta = f^0\}$. In our analysis, non-uniqueness is not a major concern.

Aim is to construct an estimator $\hat{\beta}$ of β^0 . When $p > n$ the least squares estimator $\hat{\beta}_{\text{LS}}$ will not work: it will just reproduce the data by returning the estimator $X\hat{\beta}_{\text{LS}} = Y$. This is called an instance of *overfitting*. Least squares loss with an ℓ_1 -regularization penalty can overcome the overfitting problem. This method is called the Lasso. The Lasso estimator $\hat{\beta}$ is presented in more detail in (1.3) in Section 1.4.

1.3 Notation

For a vector $v \in \mathbb{R}^n$ we use the notation $\|v\|_n^2 := v^T v/n = \|v\|_2^2/n$, where $\|\cdot\|_2$ is the ℓ_2 -norm. Write the (normalized) Gram matrix as $\hat{\Sigma} := X^T X/n$. Thus $\|X\beta\|_n^2 = \beta^T \hat{\Sigma} \beta$, $\beta \in \mathbb{R}^p$.

For a vector $\beta \in \mathbb{R}^p$ we denote its ℓ_1 -norm by $\|\beta\|_1 := \sum_{j=1}^p |\beta_j|$. Its ℓ_∞ -norm is denoted by $\|\beta\|_\infty := \max_{1 \leq j \leq p} |\beta_j|$,

Let $S \subset \{1, \dots, p\}$ be an index set. The vector $\beta_S \in \mathbb{R}^p$ with the set S as subscript is defined as

$$\beta_{j,S} := \beta_j \mathbf{1}\{j \in S\}, \quad j = 1, \dots, p. \quad (1.2)$$

Thus β_S is a p -vector with entries equal to zero at the indexes $j \notin S$. We will sometimes identify β_S with the vector $\{\beta_j\}_{j \in S} \in \mathbb{R}^{|S|}$. The vector β_{-S} has all entries inside the set S set to zero, i.e. $\beta_{-S} = \beta_{S^c}$ where $S^c = \{j \in \{1, \dots, p\} : j \notin S\}$ is the complement of the set S . The notation (1.2) allows us to write $\beta = \beta_S + \beta_{-S}$.

The *active set* S_β of a vector $\beta \in \mathbb{R}^p$ is $S_\beta := \{j : \beta_j \neq 0\}$. For a solution β^0 of $X\beta^0 = f^0$, we denote its active set by $S_0 := S_{\beta^0}$ and the cardinality of this active set by $s_0 := |S_0|$.

The j -th column of X is denoted by X_j , $j = 1, \dots, p$ (and if there is little risk of confusion we also write X_i as the i -th row of the matrix X , $i = 1, \dots, n$). For a set $S \subset \{1, \dots, p\}$ the matrix with only columns in the set S is denoted by $X_S := \{X_j\}_{j \in S}$. To fix the ordering of the columns here, we put them in increasing in j ordering. The ‘‘complement’’ matrix of X_S is denoted by $X_{-S} := \{X_j\}_{j \notin S}$. Moreover, for $j \in \{1, \dots, p\}$, we let $X_{-j} := \{X_k\}_{k \neq j}$.

1.4 The Lasso, KKT and two point inequality

The Lasso estimator (Tibshirani [1996]) $\hat{\beta}$ is a solution of the minimization problem

$$\hat{\beta} := \arg \min_{\beta \in \mathbb{R}^p} \left\{ \|Y - X\beta\|_n^2 + 2\lambda \|\beta\|_1 \right\}. \quad (1.3)$$

This estimator is the starting point from which we study more general norm-penalized estimators. The Lasso itself will be the object of study in the rest

of this chapter and in other chapters as well. Although “Lasso” refers to a method rather than an estimator, we refer to $\hat{\beta}$ as “the Lasso”. It is generally not uniquely defined but we do not express this in our notation. This is justified in the sense that the theoretical results which we will present will hold for any solution of minimization problem (1.3). The parameter $\lambda \geq 0$ is a given tuning parameter: large values will lead to a sparser solution $\hat{\beta}$, that is, a solution with more entries set to zero. In an asymptotic sense, λ will be “small”, it will generally be of order $\sqrt{\log p/n}$.

This Lasso $\hat{\beta}$ satisfies the *Karush-Kuhn-Tucker conditions* or *KKT-conditions* which say that

$$X^T(Y - X\hat{\beta})/n = \lambda\hat{z} \quad (1.4)$$

where \hat{z} is a p -dimensional vector with $\|\hat{z}\|_\infty \leq 1$ and with $\hat{z}_j = \text{sign}(\hat{\beta}_j)$ if $\hat{\beta}_j \neq 0$. The latter can also be written as

$$\hat{z}^T \hat{\beta} = \|\hat{\beta}\|_1.$$

The KKT-conditions follow from sub-differential calculus which defines the sub-differential of the absolute value function $x \mapsto |x|$ as

$$\partial|x| = \{\text{sign}(x)\}\{x \neq 0\} + [-1, 1]\{x = 0\}.$$

Thus, $\hat{z} \in \partial\|\hat{\beta}\|_1$.

The KKT-conditions may be interpreted as the Lasso version of the *normal equations* which are true for the least squares estimator. The KKT-conditions will play an important role. They imply the *almost orthogonality* of X on the one hand and the residuals $Y - X\hat{\beta}$ on the other, in the sense that

$$\|X^T(Y - X\hat{\beta})\|_\infty/n \leq \lambda.$$

Recall that λ will (generally) be “small”. Furthermore, the KKT-conditions are equivalent to: for any $\beta \in \mathbb{R}^p$

$$(\beta - \hat{\beta})^T X^T(Y - X\hat{\beta})/n \leq \lambda\|\beta\|_1 - \lambda\|\hat{\beta}\|_1.$$

We will often refer to this inequality as the *two point inequality*. As we will see in the proofs this is useful in conjunction with the *two point margin*: for any β and β'

$$2(\beta' - \beta)^T \hat{\Sigma}(\beta' - \beta^0) = \|X(\beta' - \beta^0)\|_n^2 - \|X(\beta - \beta^0)\|_n^2 + \|X(\beta' - \beta)\|_n^2.$$

Thus the two point inequality can be written in the alternative form as

$$\|Y - X\hat{\beta}\|_n^2 - \|Y - X\beta\|_n^2 + \|X(\hat{\beta} - \beta)\|_n^2 \leq 2\lambda\|\beta\|_1 - 2\lambda\|\hat{\beta}\|_1, \quad \forall \beta.$$

The two point inequality was proved more generally by [Güler [1991], Lemma 2.2] and further extended by [Chen and Teboulle [1993], Lemma 3.2], see also Lemma 3.3.1 in Section 3.3 or more generally Lemma 5.2.1 in Section 5.2.

Another important inequality will be the *convex conjugate* inequality: for any $a, b \in \mathbb{R}$

$$2ab \leq a^2 + b^2.$$

As a further look-ahead: in the case of loss functions other than least squares, we will be facing convex functions that are not necessarily quadratic and then the convex conjugate inequality is a consequence of Definition 5.2.2 in Section 5.2.

1.5 Dual norm and decomposability

As we will see, we will need a bound for the random quantity $\epsilon^T X(\hat{\beta} - \beta^0)/n$ in terms of $\|\hat{\beta} - \beta^0\|_1$, or modifications thereof. Here one may apply the dual norm inequality. The dual norm of $\|\cdot\|_1$ is the ℓ_∞ -norm $\|\cdot\|_\infty$. The *dual norm inequality* says that for any two vectors w and β

$$|w^T \beta| \leq \|w\|_\infty \|\beta\|_1.$$

Another important ingredient of the arguments to come is the *decomposability* of the ℓ_1 -norm:

$$\|\beta'\|_1 = \|\beta'_S\|_1 + \|\beta'_{-S}\|_1 \quad \forall \beta'.$$

The decomposability implies what we call the *triangle property*:

$$\|\beta\|_1 - \|\beta'\|_1 \leq \|\beta_S - \beta'_S\|_1 + \|\beta_{-S}\|_1 - \|\beta'_{-S}\|_1,$$

where β and β' are any two vectors and $S \subset \{1, \dots, p\}$ is any index set. The importance of triangle property is was highlighted in van de Geer [2001] in the context of adaptive estimation. It has been invoked at first to derive non-sharp oracle inequalities (see Bühlmann and van de Geer [2011] and its references).

1.6 Compatibility

We will need a notion of *compatibility* between the ℓ_1 -norm and the Euclidean norm $\|\cdot\|_n$. This allows us to identify β^0 to a certain extent.

Definition 1.6.1 (*van de Geer [2007], Bühlmann and van de Geer [2011]*) For a constant $L > 0$ and an index set S , the compatibility constant is

$$\hat{\phi}^2(L, S) := \min \left\{ |S| \|X\beta_S - X\beta_{-S}\|_n^2 : \|\beta_S\|_1 = 1, \|\beta_{-S}\|_1 \leq L \right\}.$$

We call L the stretching factor: generally $L \geq 1$.

Example 1.6.1 Let $S = \{j\}$ be the j -th variable for some $j \in \{1, \dots, p\}$. Then

$$\hat{\phi}^2(L, \{j\}) = \min \left\{ \|X_j - X_{-j}\gamma_j\|_n^2 : \gamma_j \in \mathbb{R}^{p-1}, \|\gamma_j\|_1 \leq L \right\}.$$

Note that the unrestricted minimum $\min\{\|X_j - X_{-j}\gamma_j\|_n : \gamma_j \in \mathbb{R}^{p-1}\}$ is the length of the anti-projection of the first variable X_j on the space spanned by the remaining variables X_{-j} . In the high-dimensional situation this unrestricted minimum will generally be zero. The ℓ_1 -restriction $\|\gamma_j\|_1 \leq L$ potentially takes care that the ℓ_1 -restricted minimum $\hat{\phi}(L, \{j\})$ is strictly positive. The ℓ_1 -restricted minimization is the dual formulation for the Lasso which we consider in the next section.

The compatibility constant $\hat{\phi}^2(L, S)$ measures the distance between the signed convex hull of the variables in X_S and linear combinations of variables in X_{-S} satisfying an ℓ_1 -restriction (that is, the latter are restricted to lie within the signed convex hull of $L \times X_{-S}$). Loosely speaking one may think of this as an ℓ_1 -variant of “(1– canonical correlation)”.

For general S one always has $\hat{\phi}^2(L, \{j\}) \geq \hat{\phi}^2(L, S)/|S|$ for all $j \in S$. The more general case $\underline{S} \subset S$ is treated in the next lemma. It says that the larger the set S the larger the *effective sparsity*² $|S|/\hat{\phi}^2(L, S)$.

Lemma 1.6.1 *For all L and $\underline{S} \subset S$ it holds that*

$$|\underline{S}|/\hat{\phi}^2(L, \underline{S}) \leq |S|/\hat{\phi}^2(L, S).$$

Proof of Lemma 1.6.1. Let

$$\|Xb\|_n^2 := \min\left\{\|X\beta\|_n^2 : \|\beta_{\underline{S}}\|_1 = 1, \|\beta_{-\underline{S}}\|_1 \leq L\right\} = \frac{\hat{\phi}^2(L, \underline{S})}{|\underline{S}|}.$$

Then $\|b_S\|_1 \geq \|b_{\underline{S}}\|_1 = 1$ and $\|b_{-S}\|_1 \leq \|b_{-\underline{S}}\|_1 \leq L$. Thus, writing $c = b/\|b_S\|_1$, we have $\|c_S\|_1 = 1$ and $\|c_{-S}\|_1 = \|b_{-S}\|_1/\|b_S\|_1 \leq \|b_{-S}\|_1 \leq L$. Therefore

$$\begin{aligned} \|Xb\|_n^2 &= \|b_S\|_1^2 \|Xc\|_n^2 \\ &\geq \|b_S\|_1^2 \min\left\{\|X\beta\|_n^2 : \|\beta_S\|_1 = 1, \|\beta_{-S}\|_1 \leq L\right\} \\ &= \|b_S\|_1^2 \hat{\phi}^2(L, S)/|S| \geq \hat{\phi}^2(L, S)/|S|. \end{aligned}$$

□

1.7 A sharp oracle inequality

Let us summarize what are the main ingredients of the proof of Theorems 1.7.1 and 1.8.1 below:

- the two point margin
- two point inequality

²or non-sparsity actually

- the dual norm inequality
- the triangle property, or decomposability
- the convex conjugate inequality
- compatibility

Finally, to control the ℓ_∞ -norm of the random vector $X^T \epsilon$ occurring below in Theorem 1.7.1 (and onwards) we will use

- empirical process theory,

see Lemma 4.2.1 for the case of Gaussian errors ϵ . See also Corollary 6.1.1 for a complete picture in the Gaussian case.

The paper Koltchinskii et al. [2011] (see also Koltchinskii [2011]) nicely combines ingredients such as the above to arrive at general sharp oracle inequalities for nuclear-norm penalized estimators for example. Theorem 1.7.1 below is a special case of their results. The sharpness refers to the constant 1 in front of $\|X(\beta - \beta^0)\|_n^2$ in the right-hand side of the result of the theorem.

Theorem 1.7.1 (*Koltchinskii et al. [2011]*) *Let λ_ϵ satisfy*

$$\lambda_\epsilon \geq \|X^T \epsilon\|_\infty / n.$$

Define for $\lambda > \lambda_\epsilon$

$$\underline{\lambda} := \lambda - \lambda_\epsilon, \quad \bar{\lambda} := \lambda + \lambda_\epsilon$$

and

$$L := \bar{\lambda} / \underline{\lambda}.$$

Then

$$\|X(\hat{\beta} - \beta^0)\|_n^2 \leq \min_S \left\{ \min_{\beta \in \mathbb{R}^p, S_\beta = S} \|X(\beta - \beta^0)\|_n^2 + \bar{\lambda}^2 |S| / \hat{\phi}^2(L, S) \right\}.$$

Theorem 1.7.1 follows from Theorem 1.8.1 below by taking there $\delta = 0$. It also follows the general case given in Theorem 5.5.1. However, a reader preferring to first consult a direct derivation before looking at generalizations may consider the the proof given in Subsection 1.11.3. We call the set of β 's over which we minimize, as in Theorem 1.7.1 “candidate oracles”. The minimizer is then called the “oracle”. Note that the stretching factor L is indeed larger than one and depends on the tuning parameter and the noise level λ_ϵ . If there is no noise, $L = 1$ (as then $\lambda_\epsilon = 0$). (However, with noise, it is not always a must to take $L > 1$.)

1.8 Including a bound for the ℓ_1 -error and allowing many small values.

We will now show that if one increases the stretching factor L in the compatibility constant one can establish a bound for the ℓ_1 -estimation error. We

moreover will no longer insist that for candidate oracles β it holds that $S = S_\beta$ as is done in Theorem 1.7.1, that is, we allow β to be non-sparse but then its small coefficients should have small ℓ_1 -norm. The result is a special case of the results for general loss and penalty given in Theorem 5.5.1.

Theorem 1.8.1 *Let λ_ϵ satisfy*

$$\lambda_\epsilon \geq \|X^T \epsilon\|_\infty / n.$$

Let $0 \leq \delta < 1$ be arbitrary and define for $\lambda > \lambda_\epsilon$

$$\underline{\lambda} := \lambda - \lambda_\epsilon, \quad \bar{\lambda} := \lambda + \lambda_\epsilon + \delta \underline{\lambda}$$

and

$$L := \frac{\bar{\lambda}}{(1 - \delta)\underline{\lambda}}.$$

Then for all $\beta \in \mathbb{R}^p$ and all sets S

$$2\delta \underline{\lambda} \|\hat{\beta} - \beta\|_1 + \|X(\hat{\beta} - \beta^0)\|_n^2 \leq \|X(\beta - \beta^0)\|_n^2 + \frac{\bar{\lambda}^2 |S|}{\hat{\phi}^2(L, S)} + 4\lambda \|\beta_{-S}\|_1. \quad (1.5)$$

The proof of this result invokes the ingredients we have outlined in the previous sections:

- the two point margin,
- two point inequality,
- the dual norm inequality,
- the triangle property,
- the convex conjugate inequality
- compatibility.

Similar ingredients will be used to cook up results with other loss functions and regularization penalties. We remark here that for least squares loss one also may take a different route where the “bias” and “variance” of the Lasso is treated separately.

Proof of Theorem 1.8.1.

- If

$$(\hat{\beta} - \beta)^T \hat{\Sigma}(\hat{\beta} - \beta^0) \leq -\delta \underline{\lambda} \|\hat{\beta} - \beta\|_1 + 2\lambda \|\beta_{-S}\|_1$$

we find from the two point margin

$$\begin{aligned} 2\delta \underline{\lambda} \|\hat{\beta} - \beta\|_1 + \|X(\hat{\beta} - \beta^0)\|_n^2 & \\ &= 2\delta \underline{\lambda} \|\hat{\beta} - \beta\|_1 + \|X(\beta - \beta^0)\|_n^2 - \|X(\beta - \hat{\beta})\|_n^2 + 2(\hat{\beta} - \beta)^T \hat{\Sigma}(\hat{\beta} - \beta^0) \\ &\leq \|X(\beta - \beta^0)\|_n^2 + 4\lambda \|\beta_{-S}\|_1 \end{aligned}$$

and we are done.

- From now on we may therefore assume that

$$(\hat{\beta} - \beta)^T \hat{\Sigma}(\hat{\beta} - \beta^0) \geq -\delta \underline{\lambda} \|\hat{\beta} - \beta\|_1 + 2\lambda \|\beta_{-S}\|_1.$$

By the two point inequality we have

$$(\hat{\beta} - \beta)^T \hat{\Sigma}(\hat{\beta} - \beta^0) \leq (\hat{\beta} - \beta)^T X^T \epsilon / n + \lambda \|\beta\|_1 - \lambda \|\hat{\beta}\|_1.$$

By the dual norm inequality

$$|(\hat{\beta} - \beta)^T X^T \epsilon / n| \leq \lambda_\epsilon \|\hat{\beta} - \beta\|_1.$$

Thus

$$\begin{aligned} (\hat{\beta} - \beta)^T \hat{\Sigma}(\hat{\beta} - \beta^0) &\leq \lambda_\epsilon \|\hat{\beta} - \beta\|_1 + \lambda \|\beta\|_1 - \lambda \|\hat{\beta}\|_1 \\ &\leq \lambda_\epsilon \|\hat{\beta}_S - \beta_S\|_1 + \lambda_\epsilon \|\hat{\beta}_{-S}\|_1 + \lambda_\epsilon \|\beta_{-S}\|_1 + \lambda \|\beta\|_1 - \lambda \|\hat{\beta}\|_1. \end{aligned}$$

By the triangle property and invoking $\underline{\lambda} = \lambda - \lambda_\epsilon$ this implies

$$(\hat{\beta} - \beta)^T \hat{\Sigma}(\hat{\beta} - \beta^0) + \underline{\lambda} \|\hat{\beta}_{-S}\|_1 \leq (\lambda + \lambda_\epsilon) \|\hat{\beta}_S - \beta_S\|_1 + (\lambda + \lambda_\epsilon) \|\beta_{-S}\|_1$$

and so

$$(\hat{\beta} - \beta)^T \hat{\Sigma}(\hat{\beta} - \beta^0) + \underline{\lambda} \|\hat{\beta}_{-S} - \beta_{-S}\|_1 \leq (\lambda + \lambda_\epsilon) \|\hat{\beta}_S - \beta_S\|_1 + 2\underline{\lambda} \|\beta_{-S}\|_1.$$

Hence, invoking $\bar{\lambda} = \lambda + \lambda_\epsilon + \delta \underline{\lambda}$,

$$\begin{aligned} (\hat{\beta} - \beta)^T \hat{\Sigma}(\hat{\beta} - \beta^0) + \underline{\lambda} \|\hat{\beta}_{-S} - \beta_{-S}\|_1 + \delta \underline{\lambda} \|\hat{\beta}_S - \beta_S\|_1 & \quad (1.6) \\ &\leq \bar{\lambda} \|\hat{\beta}_S - \beta_S\|_1 + 2\underline{\lambda} \|\beta_{-S}\|_1. \end{aligned}$$

Since $(\hat{\beta} - \beta)^T \hat{\Sigma}(\hat{\beta} - \beta^0) \geq -\delta \underline{\lambda} \|\hat{\beta} - \beta\|_1 + 2\underline{\lambda} \|\beta_{-S}\|_1$ this gives

$$(1 - \delta) \underline{\lambda} \|\hat{\beta}_{-S} - \beta_{-S}\|_1 \leq \bar{\lambda} \|\hat{\beta}_S - \beta_S\|_1$$

or

$$\|\hat{\beta}_{-S} - \beta_{-S}\|_1 \leq L \|\hat{\beta}_S - \beta_S\|_1.$$

But then by the definition of the compatibility constant

$$\|\hat{\beta}_S - \beta_S\|_1 \leq \sqrt{|S|} \|X(\hat{\beta} - \beta)\|_n / \hat{\phi}(L, S). \quad (1.7)$$

Continue with inequality (1.6) and apply the convex conjugate inequality:

$$\begin{aligned} (\hat{\beta} - \beta)^T \hat{\Sigma}(\hat{\beta} - \beta^0) + \underline{\lambda} \|\hat{\beta}_{-S} - \beta_{-S}\|_1 + \delta \underline{\lambda} \|\hat{\beta}_S - \beta_S\|_1 & \\ &\leq \bar{\lambda} \sqrt{|S|} \|X(\hat{\beta} - \beta)\|_n / \hat{\phi}(L, S) + 2\underline{\lambda} \|\beta_{-S}\|_1 \\ &\leq \frac{1}{2} \frac{\bar{\lambda}^2 |S|}{\hat{\phi}^2(L, S)} + \frac{1}{2} \|X(\hat{\beta} - \beta)\|_n^2 + 2\underline{\lambda} \|\beta_{-S}\|_1. \end{aligned}$$

Invoking the two point margin

$$2(\hat{\beta} - \beta)^T \hat{\Sigma}(\hat{\beta} - \beta^0) = \|X(\hat{\beta} - \beta^0)\|_n^2 - \|X(\beta - \beta^0)\|_n^2 + \|X(\hat{\beta} - \beta)\|_n^2,$$

we obtain

$$\|X(\hat{\beta} - \beta^0)\|_n^2 + 2\underline{\lambda} \|\hat{\beta}_{-S} - \beta_{-S}\|_1 + 2\delta \underline{\lambda} \|\hat{\beta}_S - \beta_S\|_1$$

$$\leq \|X(\beta - \beta^0)\|_n^2 + \bar{\lambda}^2 |S| / \hat{\phi}^2(L, S) + 4\lambda \|\beta_{-S}\|_1.$$

□

What we see from Theorem 1.8.1 is firstly that the tuning parameter λ should be sufficiently large to “overrule” the part due to the noise $\|X^T \epsilon\|_\infty / n$. Since $\|X^T \epsilon\|_\infty / n$ is random, we need to complete the theorem with a bound for this quantity that holds with large probability. See Corollary 6.1.1 in Section 6.1 for this completion for the case of Gaussian errors. One sees there that one may choose $\lambda \asymp \sqrt{\log p/n}$. Secondly, by taking $\beta = \beta^0$ we deduce from the theorem that the prediction error $\|X(\hat{\beta} - \beta^0)\|_n^2$ is bounded by $\bar{\lambda}^2 |S_0| / \hat{\phi}^2(L, S_0)$ where S_0 is the active set of β^0 . In other words, we reached the aim (1.1) of Section 1.1, under the conditions that the part due to the noise behaves like $\sqrt{\log p/n}$ and that the compatibility constant $\hat{\phi}^2(L, S_0)$ stays away from zero.

A third insight from Theorem 1.8.1 is that the Lasso also allows one to bound the estimation error in $\|\cdot\|_1$ -norm, provided that the stretching constant L is taken large enough. This makes sense as a compatibility constant that can stand a larger L tells us that we have good identifiability properties. Here is an example statement for the ℓ_1 -estimation error.

Corollary 1.8.1 *As an example, take $\beta = \beta^0$ and take $S = S_0$ as the active set of β^0 with cardinality $s_0 = |S_0|$. Let us furthermore choose $\lambda = 2\lambda_\epsilon$ and $\delta = 1/5$. The following ℓ_0 -sparsity based bound holds under the conditions of Theorem 1.8.1:*

$$\|\hat{\beta} - \beta^0\|_1 \leq C_0 \frac{\lambda_\epsilon s_0}{\hat{\phi}^2(4, S_0)},$$

where $C_0 = (16/5)^2(5/2)$.

Finally, it is important to note that we do not insist that β^0 is sparse. The result of Theorem 1.8.1 is good if β^0 can be well approximated by a sparse vector β or by a vector β with many smallish coefficients. The smallish coefficients occur in a term proportional to $\|\beta_{-S}\|_1$. By minimizing the bound over all candidate oracles β and all sets S one obtains the following corollary.

Corollary 1.8.2 *Under the conditions of Theorem 1.8.1, and using its notation, we have the following trade-off bound:*

$$\begin{aligned} & 2\delta\lambda \|\hat{\beta} - \beta^0\|_1 + \|X(\hat{\beta} - \beta^0)\|_n^2 \\ & \leq \min_{\beta \in \mathbb{R}^p} \min_{S \subset \{1, \dots, p\}} \left\{ 2\delta\lambda \|\beta - \beta^0\|_1 + \|X(\beta - \beta^0)\|_n^2 + \frac{\bar{\lambda}^2 |S|}{\hat{\phi}^2(L, S)} + 4\lambda \|\beta_{-S}\|_1 \right\}. \end{aligned} \quad (1.8)$$

We will refer to the minimizer (β^*, S_*) in (1.8) as the (or an) *oracle*. Corollary 1.8.2 says that the Lasso mimics the oracle (β^*, S_*) . It trades off approximation error, sparsity and the ℓ_1 -norm $\|\beta_{-S}\|_1$ of smallish coefficients. In general, we will define oracles in a loose sense, not necessarily the overall minimizer over all candidate oracles and furthermore constants in the various appearances may be (somewhat) different.

One can make two types of restrictions on the set of candidate oracles. The first one, considered in the next section (Section 1.9) requires that the pair (β, S) has $S = S_\beta$ so that the term with the smallish coefficients $\|\beta_{-S}\|_1$ vanishes. A second type of restriction is to require $\beta = \beta^0$ but optimize over S , i.e., the consider only candidate oracles (β^0, S) . This is done in Section 1.10.

1.9 The ℓ_1 -restricted oracle

Restricting ourselves to candidate oracles (β, S) with $S = S_\beta$ in Corollary 1.8.2 leads to a trade-off between the ℓ_1 -error $\|\beta - \beta^0\|_1$, the approximation error $\|X(\beta - \beta^0)\|_n^2$ and the sparseness $|S|$ (or rather the *effective sparseness* $|S|/\hat{\phi}^2(L, S)$). To study this let us consider the oracle β^* which trades off approximation error and (effective) sparsity but is meanwhile restricted to have an ℓ_1 -norm at least as large as that of β^0 .

Lemma 1.9.1 *Let for some $\bar{\lambda}$ the vector β^* be defined as*

$$\beta^* := \arg \min \left\{ \|X(\beta - \beta^0)\|_n^2 + \bar{\lambda}^2 |S_\beta| / \hat{\phi}^2(L, S_\beta) : \|\beta\|_1 \geq \|\beta^0\|_1 \right\}.$$

Let $S_ := S_{\beta^*} = \{j : \beta_j^* \neq 0\}$ be the active set of β^* . Then*

$$\bar{\lambda} \|\beta^* - \beta^0\|_1 \leq \|X(\beta^* - \beta^0)\|_n^2 + \frac{\bar{\lambda}^2 |S_*|}{\hat{\phi}^2(1, S_*)}.$$

Proof of Lemma 1.9.1. Since $\|\beta^0\|_1 \leq \|\beta^*\|_1$ we know by the ℓ_1 -triangle property

$$\|\beta_{-S_*}^0\|_1 \leq \|\beta^* - \beta_{S_*}^0\|_1.$$

Hence by the definition of the compatibility constant and by the convex conjugate inequality

$$\bar{\lambda} \|\beta^* - \beta^0\|_1 \leq 2\bar{\lambda} \|\beta^* - \beta_{S_*}^0\|_1 \leq \frac{2\bar{\lambda} \|X(\beta^* - \beta^0)\|_n}{\hat{\phi}(1, S_*)} \leq \|X(\beta^* - \beta^0)\|_n^2 + \frac{\bar{\lambda}^2 |S_*|}{\hat{\phi}^2(1, S_*)}.$$

□

From Lemma 1.9.1 we see that an ℓ_1 -restricted oracle β^* that trades off approximation error and sparseness is also going to be close in ℓ_1 -norm. We have the following corollary for the bound of Theorem 1.8.1.

Corollary 1.9.1 *Let*

$$\lambda_\epsilon \geq \|X^T \epsilon\|_\infty / n.$$

Let $0 \leq \delta < 1$ be arbitrary and define for $\lambda > \lambda_\epsilon$

$$\underline{\lambda} := \lambda - \lambda_\epsilon, \quad \bar{\lambda} := \lambda + \lambda_\epsilon + \delta \underline{\lambda}$$

and

$$L := \frac{\bar{\lambda}}{(1 - \delta)\underline{\lambda}}.$$

Let the vector β^* with active set S_* be defined as in Lemma 1.9.1. We have

$$\lambda \|\hat{\beta} - \beta^0\|_1 \leq \left(\frac{\bar{\lambda} + 2\delta\lambda}{2\delta\bar{\lambda}} \right) \left(\|X(\beta^* - \beta^0)\|_n^2 + \frac{\bar{\lambda}^2 |S_*|}{\hat{\phi}^2(L, S_*)} \right).$$

1.10 Weak sparsity

In the previous section we found a bound for the trade-off in Corollary 1.8.2 by considering the ℓ_1 -restricted oracle. In this section we take an alternative route, where we take in Theorem 1.8.1 candidate oracles (β, S) with the vector β equal to β^0 as in Corollary 1.8.1, but now S not necessarily equal to the active set $S_0 := \{j : \beta_j^0 \neq 0\}$ of β^0 . We define

$$\rho_r^r := \sum_{j=1}^p |\beta_j^0|^r, \quad (1.9)$$

where $0 < r < 1$. The constant $\rho_r > 0$ is assumed to be “not too large”. This is sometimes called *weak sparsity* as opposed to *strong sparsity* which requires “not too many” non-zero coefficients

$$s_0 := \#\{\beta_j^0 \neq 0\}.$$

Observe that this is a limiting case in the sense that

$$\lim_{r \downarrow 0} \rho_r^r = s_0.$$

Lemma 1.10.1 *Suppose β^0 satisfies the weak sparsity condition (1.9) for some $0 < r < 1$ and $\rho_r > 0$. Then for any $\bar{\lambda}$ and λ*

$$\min_S \left\{ \frac{\bar{\lambda}^2 |S|}{\hat{\phi}^2(L, S)} + 4\lambda \|\beta_{-S}^0\|_1 \right\} \leq \frac{5\bar{\lambda}^{2(1-r)} \lambda^r \rho_r^r}{\hat{\phi}^2(L, S_*)},$$

where $S_* := \{j : |\beta_j^0| > \bar{\lambda}^2/\lambda\}$ and assuming $\hat{\phi}(L, S) \leq 1$ for any L and S (to simplify the expressions).

Proof of Lemma 1.10.1. Define $\lambda_* := \bar{\lambda}^2/\lambda$. Then $S_* = \{j : |\beta_j^0| > \lambda_*\}$. We get

$$|S_*| \leq \lambda_*^{-r} \rho_r^r = \bar{\lambda}^{2(1-r)} \lambda^r \rho_r^r.$$

Moreover

$$\|\beta_{-S_*}^0\|_1 \leq \lambda_*^{1-r} \rho_r^r = \bar{\lambda}^{2(1-r)} \lambda^{r-1} \rho_r^r \leq \bar{\lambda}^{2(1-r)} \lambda^{r-1} \rho_r^r / \hat{\phi}^2(L, S_*),$$

since by assumption $\hat{\phi}^2(L, S_*) \leq 1$. □

As a consequence, we obtain bounds for the prediction error and ℓ_1 -error of the Lasso under (weak) sparsity. We only present the bound for the ℓ_1 -error.

We make some arbitrary choices for the constants: we set $\lambda = 2\lambda_\epsilon$ and we choose $\delta = 1/5$.

Corollary 1.10.1 *Assume the ℓ_r -sparsity condition (1.9) for some $0 < r < 1$ and $\rho_r > 0$. Set*

$$S_* := \{j : |\beta_j^0| > 3\lambda_\epsilon\}.$$

Then for $\lambda_\epsilon \geq \|X^T \epsilon\|_\infty/n$ and $\lambda = 2\lambda_\epsilon$, we have the ℓ_r -sparsity based bound

$$\|\hat{\beta} - \beta^0\|_1 \leq C_r \lambda_\epsilon^{1-r} \rho_r^r / \hat{\phi}^2(4, S_*).$$

assuming that $\hat{\phi}(L, S) \leq 1$ for any L and S . The constant $C_r = (16/5)^{2(1-r)}(5^2/2^r)$ depends only on r .

1.11 Complements

1.11.1 An alternative bound for the ℓ_1 -error

Theorem 5.6.1 provides an alternative (and “dirty” in the sense that not much care was paid to optimize the constants) way to prove bounds for the ℓ_1 -error. This route gives a perhaps clearer picture of the relation between the stretching constant L and the parameter δ controlling the ℓ_1 -estimation error.

Corollary 1.11.1 *(Corollary of Theorem 5.6.1.) Let $\hat{\beta}$ be the Lasso*

$$\hat{\beta} := \arg \min_{\beta \in \mathbb{R}^p} \left\{ \|Y - X\beta\|_n^2 + 2\lambda \|\beta\|_1 \right\}.$$

Take $\lambda_\epsilon \geq \|X^T \epsilon\|_\infty/n$ and $\lambda \geq 8\lambda_\epsilon/\delta$. Then for all $\beta \in \mathbb{R}^p$ and sets S

$$\lambda \delta \|\hat{\beta} - \beta\|_1 \leq \frac{2\lambda^2(1+\delta)^2|S|}{\hat{\phi}^2(1/(1-\delta), S)} + 4\|X(\beta - \beta^0)\|_n^2 + 16\lambda \|\beta_{-S}\|_1.$$

1.11.2 When there are coefficients left unpenalized

In most cases one does not penalize the constant term in the regression. More generally, suppose that the set of coefficients that are not penalized have indices $U \subset \{1, \dots, p\}$. The Lasso estimator is then

$$\hat{\beta} := \arg \min_{\beta \in \mathbb{R}^p} \left\{ \|Y - X\beta\|_n^2 + 2\lambda \|\beta_{-U}\|_1 \right\}.$$

The KKT-conditions are now

$$X^T(Y - X\hat{\beta})/n + \lambda \hat{z}_{-U} = 0, \quad \|\hat{z}_{-U}\|_\infty \leq 1, \quad z_{-U}^T \hat{\beta}_{-U} = \|\hat{\beta}_{-U}\|_1.$$

1.11.3 A direct proof of Theorem 1.7.1.

Fix some $\beta \in \mathbb{R}^p$. The derivation of Theorem 1.7.1 is identical to the one of Theorem 1.8.1 except for the fact that we consider the case $\delta = 0$ and $S = S_\beta$. These restrictions lead to a somewhat more transparent argumentation.

- If

$$(\hat{\beta} - \beta)^T \hat{\Sigma}(\hat{\beta} - \beta^0) \leq 0$$

we find from the two point margin

$$\begin{aligned} \|X(\hat{\beta} - \beta^0)\|_n^2 &= \|X(\beta - \beta^0)\|_n^2 - \|X(\beta - \hat{\beta})\|_n^2 + 2(\hat{\beta} - \beta)^T \hat{\Sigma}(\hat{\beta} - \beta^0) \\ &\leq \|X(\beta - \beta^0)\|_n^2. \end{aligned}$$

Hence then we are done.

- Suppose now that

$$(\hat{\beta} - \beta)^T \hat{\Sigma}(\hat{\beta} - \beta^0) \geq 0.$$

By the two point inequality

$$(\beta - \hat{\beta})^T X^T (Y - X\hat{\beta})/n \leq \lambda \|\beta\|_1 - \lambda \|\hat{\beta}\|_1.$$

As $Y = X\beta^0 + \epsilon$

$$(\hat{\beta} - \beta)^T \hat{\Sigma}(\hat{\beta} - \beta^0) + \lambda \|\hat{\beta}\|_1 \leq (\hat{\beta} - \beta)^T X^T \epsilon/n + \lambda \|\beta\|_1.$$

By the dual norm inequality

$$|(\hat{\beta} - \beta)^T X^T \epsilon/n| \leq (\|X^T \epsilon\|_\infty/n) \|\hat{\beta} - \beta\|_1 \leq \lambda_\epsilon \|\hat{\beta} - \beta\|_1.$$

Thus

$$(\hat{\beta} - \beta)^T \hat{\Sigma}(\hat{\beta} - \beta^0) + \lambda \|\hat{\beta}\|_1 \leq \lambda_\epsilon \|\hat{\beta} - \beta\|_1 + \lambda \|\beta\|_1.$$

By the triangle property this implies

$$(\hat{\beta} - \beta)^T \hat{\Sigma}(\hat{\beta} - \beta^0) + (\lambda - \lambda_\epsilon) \|\hat{\beta}_{-S}\|_1 \leq (\lambda + \lambda_\epsilon) \|\hat{\beta}_S - \beta\|_1.$$

or

$$(\hat{\beta} - \beta)^T \hat{\Sigma}(\hat{\beta} - \beta^0) + \underline{\lambda} \|\hat{\beta}_{-S}\|_1 \leq \bar{\lambda} \|\hat{\beta}_S - \beta\|_1. \quad (1.10)$$

Since $(\hat{\beta} - \beta)^T \hat{\Sigma}(\hat{\beta} - \beta^0) \geq 0$ this gives

$$\|\hat{\beta}_{-S}\|_1 \leq (\bar{\lambda}/\underline{\lambda}) \|\hat{\beta}_S - \beta\|_1 = L \|\hat{\beta}_S - \beta\|_1.$$

By the definition of the compatibility constant $\hat{\phi}^2(L, S)$ we then have

$$\|\hat{\beta}_S - \beta\|_1 \leq \sqrt{|S|} \|X(\hat{\beta} - \beta)\|_n / \hat{\phi}(L, S). \quad (1.11)$$

Continue with inequality (1.10) and apply the convex conjugate inequality

$$\begin{aligned} (\hat{\beta} - \beta)^T \hat{\Sigma}(\hat{\beta} - \beta^0) + \underline{\lambda} \|\hat{\beta}_{-S}\|_1 &\leq \bar{\lambda} \sqrt{|S|} \|X(\hat{\beta} - \beta)\|_n / \hat{\phi}(L, S) \\ &\leq \frac{1}{2} \frac{\bar{\lambda}^2 |S|}{\hat{\phi}^2(L, S)} + \frac{1}{2} \|X(\hat{\beta} - \beta)\|_n^2. \end{aligned}$$

Since by the two point margin

$$2(\hat{\beta} - \beta^0)^T \hat{\Sigma}(\hat{\beta} - \beta) = \|X(\hat{\beta} - \beta^0)\|_n^2 - \|X(\beta - \beta^0)\|_n^2 + \|X(\hat{\beta} - \beta)\|_n^2,$$

we obtain

$$\|X(\hat{\beta} - \beta^0)\|_n^2 + 2\lambda\|\hat{\beta}_{-S}\|_1 \leq \|X(\beta - \beta^0)\|_n^2 + \bar{\lambda}^2|S|/\hat{\phi}^2(\mathcal{L}, S).$$

□

Chapter 2

The square-root Lasso

2.1 Introduction

Consider as in the previous chapter the linear model

$$Y = X\beta^0 + \epsilon.$$

In the previous chapter we required that the tuning parameter λ for the Lasso defined in Section 1.4 is chosen at least as large as the *noise level* λ_ϵ where λ_ϵ is a bound for $\|\epsilon^T X\|_\infty/n$. Clearly, if for example the entries in ϵ are i.i.d. with variance σ_0^2 , the choice of λ will depend on the standard deviation σ_0 which will usually be unknown in practice. To avoid this problem, Belloni et al. [2011] introduced (and studied) the square-root Lasso

$$\hat{\beta} := \arg \min_{\beta \in \mathbb{R}^p} \left\{ \|Y - X\beta\|_n + \lambda_0 \|\beta\|_1 \right\}.$$

Again, we do not express in our notation that the estimator is in general not uniquely defined by the above inequality. The results to come hold for any solution.

The square-root Lasso can be seen as a method that estimates β^0 and the noise variance σ_0^2 simultaneously. Defining the residuals $\hat{\epsilon} := Y - X\hat{\beta}$ and letting $\hat{\sigma}^2 := \|\hat{\epsilon}\|_n^2$ one clearly has

$$(\hat{\beta}, \hat{\sigma}^2) = \arg \min_{\beta \in \mathbb{R}^p, \sigma^2 > 0} \left\{ \frac{\|Y - X\beta\|_n^2}{\sigma} + \sigma + 2\lambda_0 \|\beta\|_1 \right\} \quad (2.1)$$

(up to uniqueness) provided the minimum is attained at a non-zero value of σ^2 .

We note in passing that the square-root Lasso is *not* a quasi-likelihood estimator as the function $\exp[-z^2/\sigma - \sigma]$, $z \in \mathbb{R}$, is not a density with respect to a dominating measure not depending on $\sigma^2 > 0$. The square-root Lasso is moreover not to be confused with the scaled Lasso. See Section 2.7 for our definition of the latter. The scaled Lasso as we define it there *is* a quasi-likelihood estimator. It is studied in e.g. the paper Sun and Zhang [2010] which comments

on Städler et al. [2010]. In their rejoinder Städler et al. [2010] the name scaled Lasso is used. Some confusion arises as for example Sun and Zhang [2012] call the square-root Lasso the scaled Lasso.

2.2 KKT and two point inequality for the square-root Lasso

When $\hat{\sigma} > 0$ the square-root Lasso $\hat{\beta}$ satisfies the KKT-conditions

$$\frac{X^T(Y - X\hat{\beta})/n}{\|Y - X\hat{\beta}\|_n} = \lambda_0 \hat{z} \quad (2.2)$$

where $\|\hat{z}\|_\infty \leq 1$ and $\hat{z}_j = \text{sign}(\hat{\beta}_j)$ if $\hat{\beta}_j \neq 0$.

These KKT-conditions (2.2) again follow from sub-differential calculus. Indeed, for a fixed $\sigma > 0$ the sub-differential with respect to β of the expression in curly brackets given in (2.1) is equal to

$$-\frac{2X^T(Y - X\beta)/n}{\sigma} + 2\lambda_0 z(\beta)$$

with, for $j = 1, \dots, p$, $z_j(\beta)$ the sub-differential of $\beta_j \mapsto |\beta_j|$. Setting this to zero at $(\hat{\beta}, \hat{\sigma})$ gives the above KKT-conditions (2.2).

2.3 A proposition assuming no overfitting

If $\|\hat{\epsilon}\|_n = 0$ the square-root Lasso returns a degenerate solution which overfits. We assume now that $\|\hat{\epsilon}\|_n > 0$ and show in the next section that this is the case under ℓ_1 -sparsity conditions.

We define

$$\hat{R} := \frac{\|X^T \epsilon\|_\infty}{n \|\epsilon\|_n}.$$

A probability inequality for \hat{R} for the case of normally distributed errors is given in Lemma 4.2.2. See also Corollary 6.1.2 for a complete picture for the Gaussian case.

Proposition 2.3.1 *Suppose $\|\hat{\epsilon}\|_n > 0$. Let $\hat{R} \leq R$ for some constant $R > 0$. Let λ_0 satisfy*

$$\lambda_0 \|\hat{\epsilon}\|_n \geq R \|\epsilon\|_n.$$

Let $0 \leq \delta < 1$ be arbitrary and define

$$\hat{\lambda}_L \|\epsilon\|_n := \lambda_0 \|\hat{\epsilon}\|_n - R \|\epsilon\|_n, \quad \hat{\lambda}_U \|\epsilon\|_n := \lambda_0 \|\hat{\epsilon}\|_n + R \|\epsilon\|_n + \delta \hat{\lambda}_L \|\epsilon\|_n$$

and

$$\hat{L} := \frac{\hat{\lambda}_U}{(1 - \delta)\hat{\lambda}_L}.$$

Then

$$\begin{aligned} & 2\delta\hat{\lambda}_L\|\hat{\beta} - \beta^0\|_1\|\epsilon\|_n + \|X(\hat{\beta} - \beta^0)\|_n^2 \\ \leq & \min_{S \subset \{1, \dots, p\}} \min_{\beta \in \mathbb{R}^p} \left\{ 2\delta\hat{\lambda}_L\|\beta - \beta^0\|_1\|\epsilon\|_n + \|X(\beta - \beta^0)\|_n^2 \right. \\ & \left. + \frac{\hat{\lambda}_U^2\|\epsilon\|_n^2|S|}{\hat{\phi}^2(\hat{L}, S)} + 4\lambda_0\|\hat{\epsilon}\|_n\|\beta_{-S}\|_1 \right\}. \end{aligned}$$

Proof of Proposition 2.3.1. The estimator $\hat{\beta}$ satisfies the KKT-conditions (2.2) which are exactly the KKT-conditions (1.4) but with λ replaced by $\lambda_0\|\hat{\epsilon}\|_n$. This means we can recycle the proof of Theorem 1.8.1. \square

2.4 Showing the square-root Lasso does not overfit

Proposition 2.3.1 is not very useful as such as it assumes $\|\hat{\epsilon}\|_n > 0$ and depends also otherwise on the value of $\|\hat{\epsilon}\|_n$. We therefore provide bounds for this quantity.

Lemma 2.4.1 *Let λ_0 be the tuning parameter used for the square-root Lasso. Suppose that for some $0 < \eta < 1$, some $R > 0$ and some $\underline{\sigma} > 0$, we have*

$$\lambda_0(1 - \eta) \geq R$$

and

$$\lambda_0\|\beta^0\|_1/\underline{\sigma} \leq 2\left(\sqrt{1 + (\eta/2)^2} - 1\right). \quad (2.3)$$

Then on the set where $\hat{R} \leq R$ and $\|\epsilon\|_n \geq \underline{\sigma}$ we have $\left|\|\hat{\epsilon}\|_n/\|\epsilon\|_n - 1\right| \leq \eta$.

The constant $\sqrt{1 + (\eta/2)^2} - 1$ is not essential, one may replace it by a prettier-looking lower bound. Note that it is smaller than $(\eta/2)^2$ but for η small it is approximately equal to $(\eta/2)^2$. In an asymptotic formulation, say with i.i.d. standard normal noise, the conditions of Lemma 2.4.1 are met when $\|\beta^0\|_1 = o(\sqrt{n/\log p})$ and $\lambda_0 \asymp \sqrt{\log p/n}$ is suitably chosen.

The proof of the lemma makes use of the convexity of the least-squares loss function and of the penalty.

Proof of Lemma 2.4.1. Suppose $\hat{R} \leq R$ and $\|\epsilon\|_n \geq \underline{\sigma}$. First we note that the inequality (2.3) gives

$$\lambda_0\|\beta^0\|_1/\|\epsilon\|_n \leq 2\left(\sqrt{1 + (\eta/2)^2} - 1\right).$$

For the upper bound for $\|\hat{\epsilon}\|_n$ we use that

$$\|\hat{\epsilon}\|_n + \lambda_0\|\hat{\beta}\|_1 \leq \|\epsilon\|_n + \lambda_0\|\beta^0\|_1$$

by the definition of the estimator. Hence

$$\|\hat{\epsilon}\|_n \leq \|\epsilon\|_n + \lambda_0 \|\beta^0\|_1 \leq \left[1 + 2 \left(\sqrt{1 + (\eta/2)^2} - 1 \right) \right] \|\epsilon\|_n \leq (1 + \eta) \|\epsilon\|_n.$$

For the lower bound for $\|\hat{\epsilon}\|_n$ we use the convexity of both the loss function and the penalty. Define

$$t := \frac{\eta \|\epsilon\|_n}{\eta \|\epsilon\|_n + \|X(\hat{\beta} - \beta^0)\|_n}.$$

Note that $0 < t \leq 1$. Let $\hat{\beta}_t$ be the convex combination $\hat{\beta}_t := t\hat{\beta} + (1-t)\beta^0$. Then

$$\|X(\hat{\beta}_t - \beta^0)\|_n = t \|X(\hat{\beta} - \beta^0)\|_n = \frac{\eta \|\epsilon\|_n \|X(\hat{\beta} - \beta^0)\|_n}{\eta \|\epsilon\|_n + \|X(\hat{\beta} - \beta^0)\|_n} \leq \eta \|\epsilon\|_n.$$

Define $\hat{\epsilon}_t := Y - X\hat{\beta}_t$. Then, by convexity of $\|\cdot\|_n$ and $\|\cdot\|_1$,

$$\begin{aligned} \|\hat{\epsilon}_t\|_n + \lambda_0 \|\hat{\beta}_t\|_1 &\leq t \|\hat{\epsilon}\|_n + t \lambda_0 \|\hat{\beta}\|_1 + (1-t) \|\epsilon\|_n + (1-t) \lambda_0 \|\beta^0\|_1 \\ &\leq \|\epsilon\|_n + \lambda_0 \|\beta^0\|_1 \end{aligned}$$

where in the last step we again used that $\hat{\beta}$ minimizes $\|Y - X\beta\|_n + \lambda_0 \|\beta\|_1$. Taking squares on both sides gives

$$\|\hat{\epsilon}_t\|_n^2 + 2\lambda_0 \|\hat{\beta}_t\|_1 \|\hat{\epsilon}_t\|_n + \lambda_0^2 \|\hat{\beta}_t\|_1^2 \leq \|\epsilon\|_n^2 + 2\lambda_0 \|\beta^0\|_1 \|\epsilon\|_n + \lambda_0^2 \|\beta^0\|_1^2. \quad (2.4)$$

But

$$\begin{aligned} \|\hat{\epsilon}_t\|_n^2 &= \|\epsilon\|_n^2 - 2\epsilon^T X(\hat{\beta}_t - \beta^0)/n + \|X(\hat{\beta}_t - \beta^0)\|_n^2 \\ &\geq \|\epsilon\|_n^2 - 2R \|\hat{\beta}_t - \beta^0\|_1 \|\epsilon\|_n + \|X(\hat{\beta}_t - \beta^0)\|_n^2 \\ &\geq \|\epsilon\|_n^2 - 2R \|\hat{\beta}_t\|_1 \|\epsilon\|_n - 2R \|\beta^0\|_1 \|\epsilon\|_n + \|X(\hat{\beta}_t - \beta^0)\|_n^2. \end{aligned}$$

Moreover, by the triangle inequality

$$\|\hat{\epsilon}_t\|_n \geq \|\epsilon\|_n - \|X(\hat{\beta}_t - \beta^0)\|_n \geq (1 - \eta) \|\epsilon\|_n.$$

Inserting these two inequalities into (2.4) gives

$$\begin{aligned} \|\epsilon\|_n^2 - 2R \|\hat{\beta}_t\|_1 \|\epsilon\|_n &\leq \|\epsilon\|_n^2 - 2R \|\beta^0\|_1 \|\epsilon\|_n \\ &\quad + \|X(\hat{\beta}_t - \beta^0)\|_n^2 + 2\lambda_0 (1 - \eta) \|\hat{\beta}_t\|_1 \|\epsilon\|_n + \lambda_0^2 \|\hat{\beta}_t\|_1^2 \\ &\leq \|\epsilon\|_n^2 + 2\lambda_0 \|\beta^0\|_1 \|\epsilon\|_n + \lambda_0^2 \|\beta^0\|_1^2 \end{aligned}$$

which implies by the assumption $\lambda_0(1 - \eta) \geq R$

$$\begin{aligned} \|X(\hat{\beta}_t - \beta^0)\|_n^2 &\leq 2(\lambda_0 + R) \|\beta^0\|_1 \|\epsilon\|_n + \lambda_0^2 \|\beta^0\|_1^2 \\ &\leq 4\lambda_0 \|\beta^0\|_1 \|\epsilon\|_n + \lambda_0^2 \|\beta^0\|_1^2 \end{aligned}$$

2.5. A SHARP ORACLE INEQUALITY FOR THE SQUARE-ROOT LASSO 25

where in the last inequality we used $R \leq (1 - \eta)\lambda_0 \leq \lambda_0$. But continuing we see that we can write the last expression as

$$4\lambda_0\|\beta^0\|_1\|\epsilon\|_n + \lambda_0^2\|\beta^0\|_1^2 = \left((\lambda_0\|\beta_0\|_1/\|\epsilon_n\|_n + 2)^2 - 4 \right) \|\epsilon\|_n^2.$$

Again invoke the ℓ_1 -sparsity condition

$$\lambda_0\|\beta^0\|_1/\|\epsilon\|_n \leq 2\left(\sqrt{1 + (\eta/2)^2} - 1\right)$$

to get

$$\left((\lambda_0\|\beta_0\|_1/\|\epsilon_n\|_n + 2)^2 - 4 \right) \|\epsilon\|_n^2 \leq \frac{\eta^2}{4} \|\epsilon\|_n^2.$$

We thus established that

$$\|X(\hat{\beta}_t - \beta^0)\|_n \leq \frac{\eta\|\epsilon\|_n}{2}.$$

Rewrite this to

$$\frac{\eta\|\epsilon\|_n\|X(\hat{\beta} - \beta^0)\|_n}{\eta\|\epsilon\|_n + \|X(\hat{\beta} - \beta^0)\|_n} \leq \frac{\eta\|\epsilon\|_n}{2},$$

and rewrite this in turn to

$$\eta\|\epsilon\|_n\|X(\hat{\beta} - \beta^0)\|_n \leq \frac{\eta^2\|\epsilon\|_n^2}{2} + \frac{\eta\|\epsilon\|_n\|X(\hat{\beta} - \beta^0)\|_n}{2}$$

or

$$\|X(\hat{\beta} - \beta^0)\|_n \leq \eta\|\epsilon\|_n.$$

But then, by repeating the argument, also

$$\|\hat{\epsilon}\|_n \geq \|\epsilon\|_n - \|X(\hat{\beta} - \beta^0)\|_n \geq (1 - \eta)\|\epsilon\|_n.$$

□

2.5 A sharp oracle inequality for the square-root Lasso

We combine the results of the two previous sections.

Theorem 2.5.1 *Assume the ℓ_1 -sparsity (2.3) for some $0 < \eta < 1$ and $\underline{\sigma} > 0$, i.e.*

$$\lambda_0\|\beta^0\|_1/\underline{\sigma} \leq 2\left(\sqrt{1 + (\eta/2)^2} - 1\right).$$

Let λ_0 satisfy for some $R > 0$

$$\lambda_0(1 - \eta) > R.$$

Let $0 \leq \delta < 1$ be arbitrary and define

$$\underline{\lambda}_0 := \lambda_0(1 - \eta) - R,$$

$$\bar{\lambda}_0 := \lambda_0(1 + \eta) + R + \delta\lambda_0$$

and

$$L := \frac{\bar{\lambda}_0}{(1 - \delta)\lambda_0}.$$

Then on the set where $\hat{R} \leq R$ and $\|\epsilon\|_n \geq \underline{\sigma}$, we have

$$\begin{aligned} & 2\delta\lambda_0\|\hat{\beta} - \beta^0\|_1\|\epsilon\|_n + \|X(\hat{\beta} - \beta^0)\|_n^2 \\ \leq & \min_{S \in \{1, \dots, p\}} \min_{\beta \in \mathbb{R}^p} \left\{ 2\delta\lambda_0\|\beta - \beta^0\|_1\|\epsilon\|_n + \|X(\beta - \beta^0)\|_n^2 \right. \\ & \left. + \frac{\bar{\lambda}_0^2|S|\|\epsilon\|_n^2}{\hat{\phi}^2(L, S)} + 4\lambda_0(1 + \eta)\|\epsilon\|_n\|\beta_{-S}\|_1 \right\}. \end{aligned} \quad (2.5)$$

Proof of Theorem 2.5.1. This follows from the same arguments as those used for Theorem 1.8.1, and inserting Lemma 2.4.1. \square

The minimizer (β^*, S_*) in (2.5) is again called the oracle and (2.5) is called an oracle inequality. The paper Sun and Zhang [2013] contains (among other things) similar results as Theorem 2.5.1, although with different constants and the oracle inequality shown there is not a sharp one.

2.6 A bound for the mean ℓ_1 -error

It is of interest to have bounds for the mean ℓ_1 -estimation error $\mathbf{E}\|\hat{\beta} - \beta^0\|_1$ (or even for higher moments $\mathbf{E}\|\hat{\beta} - \beta^0\|_1^m$ with $m > 1$). Such bounds are will be important when aiming at proving so-called strong asymptotic unbiasedness of certain (de-sparsified) estimators, which in turn is invoked for deriving asymptotic lower bounds for the variance of such estimators. We refer to

Lemma 2.6.1 *Suppose the conditions of Theorem 2.5.1. Let moreover for some constant $\underline{\phi}(L, S) > 0$, \mathcal{T} be the set*

$$\mathcal{T} := \{\hat{R} \leq R, \|\epsilon\|_n \geq \bar{\sigma}, \hat{\phi}(L, S) \geq \underline{\phi}(L, S)\}.$$

Let (for the case of random design)

$$\|X\beta\|^2 := \mathbf{E}\|X\beta\|_n^2, \quad \beta \in \mathbb{R}^p.$$

Define (as in (2.5))

$$\begin{aligned} \eta_n := & \min_{S \in \{1, \dots, p\}} \min_{\beta \in \mathbb{R}^p} \left\{ \|\beta - \beta^0\|_1 + \frac{\|X(\beta - \beta^0)\|^2}{2\delta\bar{\sigma}\lambda_0} \right. \\ & \left. + \frac{\bar{\lambda}_0|S|\sigma_0}{2\delta\underline{\phi}^2(L, S)} + \frac{4\lambda_0(1 + \eta)\|\beta_{-S}\|_1}{2\delta\lambda_0} \right\}. \end{aligned}$$

Define moreover

$$\zeta_n := \frac{\sigma_0}{\lambda_0} \mathbf{P}^{1/2}(\mathcal{T}^c) + \frac{2\left(\bar{\sigma}\sqrt{1 + (\eta/2)^2} - 1\right) + 1}{\lambda_0} \mathbf{P}(\mathcal{T}^c).$$

Then

$$\mathbf{E}\|\hat{\beta} - \beta^0\|_1 \leq \eta_n + \zeta_n.$$

In an asymptotic formulation and with fixed design (where $\hat{\phi}(L, S)$ is fixed), one can choose R and $\bar{\sigma}$ large such that $\mathbf{P}(\mathcal{T}^c) = \mathcal{O}(p^{-\tau})$ for some $\tau > 0$, but such that the bound η_n for $\|\hat{\beta} - \beta^0\|_1$ is only effected by this in terms of constants. For p large the leading term in the bound $\eta_n + \zeta_n$ for $\mathbf{E}\|\hat{\beta} - \beta^0\|_1$ is then η_n . In other words, the bound in probability for $\|\hat{\beta} - \beta^0\|_1$ is of the same order as the bound in expectation.

To bound $\mathbf{P}(\mathcal{T}^c)$ for the case of fixed design we refer to Lemma 4.2.2 in Section 4.2. Then, when for example $s_0 = o(\delta_n \sqrt{n/\log p})$ (say) the overall conclusion is

$$\mathbf{E}\|\hat{\beta} - \beta^0\|_1 = o(\delta_n).$$

Similar conclusions hold under weak sparsity assumptions.

Proof of Lemma 2.6.1. Let $\mathcal{T} := \{\hat{R} \leq R, \|\epsilon\|_n \geq \bar{\sigma}, \hat{\phi}(L, S) \geq \underline{\phi}(L, S)\}$. Then by Theorem 2.5.1

$$\mathbf{E}\|\hat{\beta} - \beta^0\|_1 1_{\mathcal{T}} \leq \eta_n.$$

Moreover, by the definition of $\hat{\beta}$

$$\|\hat{\beta}\|_1 \leq \|\epsilon\|_n/\lambda_0 + \|\beta^0\|_1 \leq \|\epsilon\|_n/\lambda_0 + 2\left(\bar{\sigma}\sqrt{1 + (\eta/2)^2} - 1\right)/\lambda_0.$$

It follows that

$$\|\hat{\beta} - \beta^0\|_1 \leq \frac{\|\epsilon\|_n}{\lambda_0} + \frac{2\left(\bar{\sigma}\sqrt{1 + (\eta/2)^2} - 1\right) + 1}{\lambda_0}.$$

Therefore

$$\mathbf{E}\|\hat{\beta} - \beta^0\|_1 1_{\mathcal{T}^c} \leq \frac{\sigma_0}{\lambda_0} \mathbf{P}^{1/2}(\mathcal{T}^c) + \frac{2\left(\bar{\sigma}\sqrt{1 + (\eta/2)^2} - 1\right) + 1}{\lambda_0} \mathbf{P}(\mathcal{T}^c) = \zeta_n.$$

□

2.7 Comparison with scaled Lasso

Fix a tuning parameter $\lambda_0 > 0$. Consider the Lasso with scale parameter σ

$$\hat{\beta}(\sigma) := \arg \min_{\beta} \left\{ \|Y - X\beta\|_n^2 + 2\lambda_0\sigma\|\beta\|_1 \right\},$$

the (scale free) square-root Lasso

$$\hat{\beta}_{\sharp} := \arg \min_{\beta} \left\{ \|Y - X\beta\|_n + \lambda_0 \|\beta\|_1 \right\}$$

and the *scaled Lasso* (Sun and Zhang [2012])

$$(\hat{\beta}_b, \tilde{\sigma}_b^2) := \arg \min_{\beta, \sigma} \left\{ \frac{\|Y - X\beta\|_n^2}{\sigma^2} + \log \sigma^2 + \frac{2\lambda_0 \|\beta\|_1}{\sigma} \right\}.$$

Then one easily verifies that

$$\tilde{\sigma}_b^2 = \|Y - X\hat{\beta}_b\|_n^2 + \lambda_0 \tilde{\sigma}_b \|\hat{\beta}_b\|_1$$

and that $\hat{\beta}_b = \hat{\beta}(\tilde{\sigma}_b)$. Moreover, if we define

$$\hat{\sigma}_{\sharp}^2 := \|Y - X\hat{\beta}_{\sharp}\|_n^2$$

we see that $\hat{\beta}_{\sharp} = \hat{\beta}(\hat{\sigma}_{\sharp})$.

Let us write the residual sum of squares (normalized by n^{-1}) when using σ as scale parameter as

$$\hat{\sigma}^2(\sigma) := \|Y - X\hat{\beta}(\sigma)\|_n^2.$$

Moreover, write the penalized (and normalized) residual sum of squares plus penalty when using σ as scale parameter as

$$\tilde{\sigma}^2(\sigma) := \|Y - X\hat{\beta}(\sigma)\|_n^2 + \lambda_0 \sigma \|\hat{\beta}(\sigma)\|_1.$$

Let furthermore

$$\tilde{\sigma}_{\sharp}^2 := \|Y - X\hat{\beta}_{\sharp}\|_n^2 + \lambda_0 \tilde{\sigma}_{\sharp} \|\hat{\beta}_{\sharp}\|_1$$

and

$$\hat{\sigma}_b^2 := \|Y - X\hat{\beta}_b\|_n^2.$$

The scaled Lasso includes the penalty in its estimator $\tilde{\sigma}_b^2$ of the noise variance $\sigma_0^2 := \mathbb{E}\|\epsilon\|_n^2$ (assuming the latter exists). The square-root Lasso does not include the penalty in its estimator $\hat{\sigma}_{\sharp}^2$ of σ_0^2 . It obtains $\hat{\sigma}_{\sharp}^2$ as a stable point of the equation $\hat{\sigma}_{\sharp}^2 = \hat{\sigma}^2(\hat{\sigma}_{\sharp})$ and the scaled Lasso obtains $\tilde{\sigma}_b^2$ as a stable point of the equation $\tilde{\sigma}_b^2 = \tilde{\sigma}^2(\tilde{\sigma}_b)$. By the mere definition of $\tilde{\sigma}^2(\sigma)$ and $\hat{\sigma}^2(\sigma)$ we also have $\tilde{\sigma}_{\sharp}^2 = \tilde{\sigma}^2(\hat{\sigma}_{\sharp})$ and $\hat{\sigma}_b^2 = \hat{\sigma}^2(\tilde{\sigma}_b)$.

We end this section with a lemma showing the relation between the penalized residual sum of squares and the inner product between response and residuals.

Lemma 2.7.1 *It holds that*

$$\tilde{\sigma}^2(\sigma) = Y^T(Y - X\hat{\beta}(\sigma))/n.$$

Proof of Lemma 2.7.1. We have

$$Y^T(Y - X\hat{\beta}(\sigma))/n = \|Y - X\hat{\beta}(\sigma)\|_n^2 + \hat{\beta}^T(\sigma)X^T(Y - X\hat{\beta}(\sigma))/n$$

and by the KKT-conditions (see (1.4))

$$\hat{\beta}^T(\sigma)X^T(Y - X\hat{\beta}(\sigma))/n = \lambda_0 \sigma \|\hat{\beta}(\sigma)\|_1.$$

□

2.8 The multivariate square-root Lasso

For bounds for the bias of the Lasso and also for the construction of confidence sets we will consider the regression of X_J on X_{-J} (J being some subset of $\{1, \dots, p\}$) invoking a multivariate version of the square-root Lasso. Here, we use here a standard notation with X being the input and Y being the response. We will then later replace X by X_{-J} and Y by X_J .

The matrix X is as before an $n \times p$ input matrix and the response Y is now an $n \times q$ matrix for some $q \geq 1$. For a matrix A we write

$$\|A\|_1 := \sum_{j,k} |A_{j,k}|$$

and we denote its nuclear norm by

$$\|A\|_{\text{nuclear}} := \text{trace}((A^T A)^{1/2}).$$

We define the *multivariate square-root Lasso*

$$\hat{B} := \arg \min_B \left\{ \|Y - XB\|_{\text{nuclear}} / \sqrt{n} + \lambda_0 \|B\|_1 \right\} \quad (2.6)$$

with $\lambda_0 > 0$ again a tuning parameter. The minimization is over all $p \times q$ matrices B . We consider $\hat{\Sigma} := (Y - X\hat{B})^T(Y - X\hat{B})/n^1$ as estimator of the noise co-variance matrix.

The KKT-conditions for the multivariate square-root Lasso will be a major ingredient of later results. We present these KKT-conditions in the following lemma in equation (2.7).

Lemma 2.8.1 *We have*

$$(\hat{B}, \hat{\Sigma}) = \arg \min_{B, \Sigma > 0} \left\{ \text{trace} \left((Y - XB)^T (Y - XB) \Sigma^{-1/2} \right) / n \right. \\ \left. + \text{trace}(\Sigma^{1/2}) + 2\lambda_0 \|B\|_1 \right\}$$

where the minimization is over all symmetric positive definite matrix Σ (this being denoted by $\Sigma > 0$) and where it is assumed that the minimum is indeed attained at some $\hat{\Sigma} > 0$. The multivariate Lasso satisfies the KKT-conditions

$$X^T (Y - X\hat{B}) \hat{\Sigma}^{-1/2} / n = \lambda_0 \hat{Z}, \quad (2.7)$$

where \hat{Z} is a $p \times q$ matrix with $\|\hat{Z}\|_\infty \leq 1$ and with $\hat{Z}_{k,j} = \text{sign}(\hat{B}_{k,j})$ if $\hat{B}_{k,j} \neq 0$ ($k = 1, \dots, p, j = 1, \dots, q$).

¹In this subsection $\hat{\Sigma}$ is not the Gram matrix $X^T X/n$

Proof of Lemma 2.8.1. Let us write, for a $p \times q$ matrix B , the residuals as $\Sigma_B := (Y - XB)^T(Y - XB)/n$. Let $\Sigma_{\min}(B)$ be the minimizer of

$$\text{trace}(\Sigma_B \Sigma^{-1/2}) + \text{trace}(\Sigma^{1/2}) \quad (2.8)$$

over Σ . Then $\Sigma_{\min}(B)$ equals Σ_B . To see this we invoke the reparametrization $\Omega := \Sigma^{-1/2}$ so that $\Sigma^{1/2} = \Omega^{-1}$. We now minimize

$$\text{trace}(\Sigma_B \Omega) + \text{trace}(\Omega^{-1})$$

over $\Omega > 0$. The matrix derivative with respect to Ω of $\text{trace}(\Sigma_B \Omega)$ is Σ_B . The matrix derivative of $\text{trace}(\Omega^{-1})$ with respect to Ω is equal to $-\Omega^{-2}$. Hence the minimizer $\Omega_{\min}(B)$ satisfies the equation

$$\Sigma_B - \Omega_{\min}^{-2}(B) = 0,$$

giving

$$\Omega_{\min}(B) = \Sigma_B^{-1/2}.$$

so that

$$\Sigma_{\min}(B) = \Omega_{\min}^{-2}(B) = \Sigma_B.$$

Inserting this solution back in (2.8) gives $2\text{trace}(\Sigma_B^{1/2})$ which is equal to $2\|Y - XB\|_{\text{nuclear}}/\sqrt{n}$. This proves the first part of the lemma.

Let now for each $\Sigma > 0$, B_Σ be the minimizer of

$$\text{trace}(\Sigma_B \Sigma^{-1/2}) + 2\lambda_0 \|B\|_1.$$

By sub-differential calculus we have

$$X^T(Y - XB_\Sigma)\Sigma^{-1/2}/n = \lambda_0 Z_\Sigma$$

where $\|Z_\Sigma\|_\infty \leq 1$ and $(Z_\Sigma)_{k,j} = \text{sign}((B_\Sigma)_{k,j})$ if $(B_\Sigma)_{k,j} \neq 0$ ($k = 1, \dots, p$, $j = 1, \dots, q$). The KKT-conditions (2.7) follow from $\hat{B} = B_{\hat{\Sigma}}$. \square

Chapter 3

Structured sparsity

3.1 The Ω -structured sparsity estimator

Like Chapter 1 this chapter studies the linear model with fixed design

$$Y = X\beta^0 + \epsilon$$

where $Y \in \mathbb{R}^n$ is an observed response variable, X is a $n \times p$ input matrix, $\beta^0 \in \mathbb{R}^p$ is a vector of unknown coefficients and $\epsilon \in \mathbb{R}^n$ is unobservable noise. The Ω -structured sparsity estimator is

$$\hat{\beta} := \arg \min_{\beta \in \mathbb{R}^p} \left\{ \|Y - X\beta\|_n^2 + 2\lambda\Omega(\beta) \right\},$$

with Ω a given norm on \mathbb{R}^p . The reason for applying some other norm than the ℓ_1 -norm depends on the particular application. In this chapter, we have in mind the situation of a sparsity inducing norm, which means roughly that it favours solutions $\hat{\beta}$ with many zeroes structured in a particular way¹. Such generalizations of the Lasso are examined in Jenatton et al. [2011], Micchelli et al. [2010], Bach [2010], Bach et al. [2012], Maurer and Pontil [2012], van de Geer [2014] for example. The norm Ω is constructed in such a way that the sparsity pattern in $\hat{\beta}$ follow a suitable structure. This may for example facilitate interpretation.

This chapter largely follows Stucky and van de Geer [2015].

The question is now: can we prove oracle inequalities (as given in for example Theorem 1.7.1) for more general norms Ω than the ℓ_1 -norm? To answer this question we first recall the ingredients of the proof Theorem 1.7.1.

- the two point margin
- the two point inequality

¹For example the least-squares estimator with so-called *nuclear norm* penalization is formally also a structured sparsity estimator. This will be considered in Section 6.4. The topic of this chapter is rather norms which are weakly decomposable as defined in Definition 3.4.1.

- the dual norm inequality
- the (ℓ_1) -triangle property
- (ℓ_1) -compatibility
- the convex conjugate inequality.

We will also need empirical process theory to bound certain functions of ϵ . This will be done in Chapter 4.

The convex conjugate inequality and two point margin have to do with the loss function and not with the penalty. Since our loss function is still least squares loss we can use these two ingredients as before. The other ingredients: two point inequality, dual norm inequality, Ω -triangle property and Ω -compatibility will be discussed in what follows.

3.2 Dual norms and KKT-conditions for structured sparsity

The dual norm of Ω . is defined as

$$\Omega_*(w) := \max_{\Omega(\beta) \leq 1} |w^T \beta|, \quad w \in \mathbb{R}^p.$$

Therefore the dual norm inequality holds by definition: for any two vectors w and β

$$|w^T \beta| \leq \Omega_*(w) \Omega(\beta).$$

The sub-differential of Ω is given by

$$\partial\Omega(\beta) = \begin{cases} \{w \in \mathbb{R}^p : \Omega_*(w) \leq 1\} & \text{if } \beta = 0 \\ \{w \in \mathbb{R}^p : \Omega_*(w) = 1, w^T \beta = \Omega(\beta)\} & \text{if } \beta \neq 0 \end{cases}$$

(Bach et al. [2012], Proposition 1.2). It follows that the Ω -structured sparsity estimator $\hat{\beta}$ satisfies

$$\Omega_*(X^T(Y - X\hat{\beta}))/n \leq \lambda,$$

and if $\hat{\beta} \neq 0$,

$$\Omega_*(X^T(Y - X\hat{\beta}))/n = \lambda, \quad \hat{\beta}^T X^T(Y - X\hat{\beta})/n = \Omega(\hat{\beta}).$$

The KKT-conditions are

$$X^T(Y - X\hat{\beta})/n = \lambda \hat{z},$$

where $\Omega_*(\hat{z}) \leq 1$ and $\hat{z}^T \hat{\beta} = \Omega(\hat{\beta})$.

3.3 Two point inequality

We call the result (3.1) below in Lemma 3.3.1 a *two point inequality*. See also Lemma 5.2.1 which treats more general loss functions.

Lemma 3.3.1 *Let $\hat{\beta}$ be the estimator*

$$\hat{\beta} := \arg \min_{\beta \in \mathbb{R}^p} \left\{ \|Y - X\beta\|_n^2 + 2\text{pen}(\beta) \right\},$$

where $\text{pen} : \mathbb{R}^p \rightarrow \mathbb{R}$ is a convex penalty. Then for any $\beta \in \mathbb{R}^p$

$$(\beta - \hat{\beta})^T X^T (Y - X\hat{\beta})/n \leq \text{pen}(\beta) - \text{pen}(\hat{\beta}) \quad (3.1)$$

Proof of Lemma 3.3.1. Fix $\beta \in \mathbb{R}^p$ and define for $0 < t \leq 1$,

$$\hat{\beta}_t := (1 - t)\hat{\beta} + t\beta.$$

We have

$$\begin{aligned} \|Y - X\hat{\beta}\|_n^2 + 2\text{pen}(\hat{\beta}) &\leq \|Y - X\hat{\beta}_t\|_n^2 + 2\text{pen}(\hat{\beta}_t) \\ &\leq \|Y - X\hat{\beta}_t\|_n^2 + 2(1 - t)\text{pen}(\hat{\beta}) + 2t\text{pen}(\beta) \end{aligned}$$

where we used the convexity of the penalty. It follows that

$$\frac{\|Y - X\hat{\beta}\|_n^2 - \|Y - X\hat{\beta}_t\|_n^2}{t} + 2\text{pen}(\hat{\beta}) \leq 2\text{pen}(\beta).$$

But clearly

$$\lim_{t \downarrow 0} \frac{\|Y - X\hat{\beta}\|_n^2 - \|Y - X\hat{\beta}_t\|_n^2}{t} = 2(Y - X\hat{\beta})^T X(\beta - \hat{\beta})/n.$$

□

Note that the two point inequality (3.1) can be written in the form

$$(\hat{\beta} - \beta)^T \hat{\Sigma}(\hat{\beta} - \beta^0) \leq \epsilon^T X(\hat{\beta} - \beta)/n + \text{pen}(\beta) - \text{pen}(\hat{\beta}).$$

For the case $\text{pen} = \lambda\Omega$ an alternative proof can be formulated from the KKT-conditions.

3.4 Weak decomposability and Ω -triangle property

What we need is a more general version of the ℓ_1 -triangle property: the Ω -triangle property

$$\Omega(\beta) - \Omega(\beta') \leq \Omega(\beta'_S - \beta_S) + \Omega(\beta_{-S}) - \Omega^{-S}(\beta'_{-S}), \quad \forall \beta, \beta'.$$

Here Ω^{-S} is a norm defined on $\mathbb{R}^{p-|S|}$. This property holds if S is a *allowed* set which is defined as follows

Definition 3.4.1 *The set S is called (Ω) -allowed if for a norm Ω^{-S} on $\mathbb{R}^{p-|S|}$ it holds that*

$$\Omega(\beta) \geq \Omega(\beta_S) + \Omega^{-S}(\beta_{-S}), \quad \forall \beta \in \mathbb{R}^p. \quad (3.2)$$

We call Ω weakly decomposable for the set S .

Clearly for the ℓ_1 -norm $\|\cdot\|_1$ any subset S is allowed, Ω^{-S} is again the ℓ_1 -norm and one has in fact equality: $\|\beta\|_1 = \|\beta_S\|_1 + \|\beta_{-S}\|_1$. More examples are in Section 3.9 and Subsection 3.10.2. Observe also that by the triangle inequality

$$\Omega(\beta) \leq \Omega(\beta_S) + \Omega(\beta_{-S}), \quad \forall \beta \in \mathbb{R}^p.$$

For allowed sets, one thus in a sense also has the reverse inequality, albeit that $\Omega(\beta_{-S})$ is now replaced by some other norm.

We introduce some further notation. If Ω and $\underline{\Omega}$ are two norms on Euclidean space, say \mathbb{R}^p , we write

$$\Omega \geq \underline{\Omega} \Leftrightarrow \Omega(\beta) \geq \underline{\Omega}(\beta) \quad \forall \beta \in \mathbb{R}^p$$

and then say that Ω is a stronger norm than $\underline{\Omega}$. Note that

$$\Omega \geq \underline{\Omega} \Rightarrow \Omega_* \leq \underline{\Omega}_*.$$

For an allowed set S , write (3.2) shorthand as

$$\Omega \geq \Omega(\cdot|S) + \Omega^{-S}$$

where for any set J and $\beta \in \mathbb{R}^p$, the notation $\Omega(\beta|J) = \Omega(\beta_J)$ is used. Define Ω^{-S} as the largest norm among the norms $\underline{\Omega}^{-S}$ for which

$$\Omega \geq \Omega(\cdot|S) + \underline{\Omega}^{-S}.$$

If $\Omega^{-S} = \Omega(\cdot| - S)$ we call Ω decomposable for the set S .

Let us compare the various norms.

Lemma 3.4.1 *Let S be an allowed set so that*

$$\Omega \geq \Omega(\cdot|S) + \Omega^{-S} =: \underline{\Omega}.$$

Then

$$\Omega_* \leq \underline{\Omega}_* = \max\{\Omega_*(\cdot|S), \Omega_*^{-S}\}$$

and

$$\Omega^{-S} \leq \Omega(\cdot| - S), \quad \Omega_*^{-S} \geq \Omega_*(\cdot| - S).$$

We see that the original norm Ω is stronger than the decomposed version $\underline{\Omega} = \Omega(\cdot|S) + \Omega^{-S}$. As we will experience this means we will lose a certain amount of this strength by replacing Ω by $\underline{\Omega}$ at places.

Proof of Lemma 3.4.1. We first prove $\underline{\Omega}_* \leq \max\{\Omega_*(\cdot|S), \Omega_*^{-S}\}$. Clearly

$$\underline{\Omega}_*(w_S) = \Omega_*(w_S) = \max\{\Omega_*(w_S), \Omega_*^{-S}(0)\}$$

and

$$\underline{\Omega}_*(w_{-S}) = \Omega_*^{-S}(w_{-S}) = \max\{\Omega_*(0), \Omega_*^{-S}(w_{-S})\}.$$

So it suffices to consider vectors w with both $w_S \neq 0$ and $w_{-S} \neq 0$. By the definition of the dual norm $\underline{\Omega}_*$

$$\begin{aligned} \underline{\Omega}_*(w) &= \max_{\underline{\Omega}(\beta) \leq 1} w^T \beta \\ &= \max_{\Omega(\beta_S) + \Omega^{-S}(\beta_{-S}) \leq 1} \left\{ \frac{w_S^T \beta_S}{\Omega(\beta_S)} \Omega(\beta_S) + \frac{w_{-S}^T \beta_{-S}}{\Omega^{-S}(\beta_{-S})} \Omega^{-S}(\beta_{-S}) \right\} \\ &\leq \max_{\Omega(\beta_S) + \Omega^{-S}(\beta_{-S}) \leq 1} \left\{ \Omega_*(w_S) \Omega(\beta_S) + \Omega_*^{-S}(w_{-S}) \Omega^{-S}(\beta_{-S}) \right\} \\ &\leq \max\{\Omega_*(w_S), \Omega_*^{-S}(w_{-S})\}. \end{aligned}$$

The reverse inequality $\underline{\Omega}_* \geq \max\{\Omega_*(\cdot|S), \Omega_*^{-S}\}$. follows from

$$\underline{\Omega}_*(w) = \max_{\underline{\Omega}(\beta) \leq 1} w^T \beta \geq \max_{\underline{\Omega}(\beta) \leq 1, \beta_{-S}=0} w^T \beta = \Omega_*(w_S)$$

and similarly $\underline{\Omega}(w) \geq \Omega_*^{-S}(w_{-S})$.

For the second result of the lemma, we use the triangle inequality

$$\Omega \leq \Omega(\cdot|S) + \Omega(\cdot - S)$$

Since S is assumed to be allowed we also have

$$\Omega \geq \Omega(\cdot|S) + \Omega^{-S}.$$

So it must hold that $\Omega^{-S} \leq \Omega(\cdot - S)$. Hence also $\Omega_*^{-S} \geq \Omega(\cdot - S)$. \square

3.5 Ω -compatibility

As for the Lasso the results will depend on compatibility constants, which in the present setup are defined as follows.

Definition 3.5.1 (*van de Geer [2014]*) *Suppose S is an allowed set. Let $L > 0$ be some stretching factor. The Ω -compatibility constant (for S) is*

$$\hat{\phi}_\Omega^2(L, S) := \min \left\{ |S| \|X\beta_S - X\beta_{-S}\|_n^2 : \Omega(\beta_S) = 1, \Omega^{-S}(\beta_{-S}) \leq L \right\}.$$

A comparison of this definition with the compatibility condition for the Lasso (Definition 1.6.1 in Section 1.6) we merely see that the ℓ_1 -norm is replaced by a more general norm. The geometric interpretation is however less evident. On top of that the various Ω -compatibility constants do not allow a clear ordering, i.e., generally one cannot say that one norm gives smaller compatibility

constants than another. Suppose for example that both $\Omega(\cdot|S)$ and Ω^{-S} are stronger than ℓ_1 (see Section 3.7 for this terminology). Then clearly

$$(i) := \left\{ \beta : \Omega(\beta_S) = 1, \Omega^{-S}(\beta_{-S})\|_1 \leq L \right\} \subset \left\{ \beta : \|\beta_S\|_1 \leq 1, \|\beta_{-S}\|_1 \leq L \right\} =: (ii).$$

But the latter set (ii) is *not* a subset of

$$(iii) := \left\{ \beta : \|\beta_S\|_1 = 1, \|\beta_{-S}\|_1 \leq L \right\}.$$

Hence we cannot say whether the minimum over the first set (i) is larger (or smaller) than the minimum over the third set (iii) .

3.6 A sharp oracle inequality with structured sparsity

Let Ω be a norm on \mathbb{R}^p . We have in mind a sparsity inducing norm for which the collection of allowed sets (see Definition 3.4.1) does not only consist of the trivial sets \emptyset and \mathbb{R}^p . Recall that the Ω -structured sparsity estimator $\hat{\beta}$ is defined as

$$\hat{\beta} := \arg \min_{\beta \in \mathbb{R}^p} \left\{ \|Y - X\beta\|_n^2 + 2\lambda\Omega(\beta) \right\}.$$

Theorem 3.6.1 *Consider an allowed set S . Let λ_S and λ^{-S} be constants such that*

$$\lambda_S \geq \Omega_*(X_S^T \epsilon)/n, \quad \lambda^{-S} \geq \Omega_*^{-S}(X_{-S}^T \epsilon)/n.$$

Let $\delta_1 \geq 0$ and $0 \leq \delta_2 < 1$ be arbitrary. Take $\lambda > \lambda^{-S}$ and define

$$\underline{\lambda} := \lambda - \lambda^{-S}, \quad \bar{\lambda} := \lambda + \lambda_S + \delta_1 \underline{\lambda}$$

and

$$L := \frac{\bar{\lambda}}{(1 - \delta_2)\underline{\lambda}}.$$

Then for any β it holds that

$$\begin{aligned} & 2\delta_1 \underline{\lambda} \Omega(\hat{\beta}_S - \beta_S) + 2\delta_2 \underline{\lambda} \Omega^{-S}(\hat{\beta}_{-S} - \beta_{-S}) + \|X(\hat{\beta} - \beta^0)\|_n^2 \\ & \leq \|X(\beta - \beta^0)\|_n^2 + \frac{\bar{\lambda}^2 |S|}{\hat{\phi}_\Omega^2(L, S)} + 4\lambda \Omega(\beta_{-S}). \end{aligned}$$

Theorem 3.6.1 is a special case of Theorem 5.5.1 in Section 5.5. We do however provide a direct proof in Subsection 3.10.3. The direct proof moreover facilitates the verification of the claims made in Proposition 3.8.1 which treats the case of square-root loss with sparsity inducing norms.

One may minimize the result of Theorem 3.6.1 over all candidate oracles (β, S) with β a vector in \mathbb{R}^p and S an allowed set S . This then gives oracle values

(β^*, S_*) . Theorem 3.6.1 is a generalization of Theorem 1.8.1 in Section 1.8. As there, with the choice $\delta_1 = \delta_2 = 0$ it has no result for the $\Omega(|S|)$ or Ω^{-S} estimation error. If we take these values strictly positive say $\delta_1 = \delta_2 = \delta > 0$ one obtains the following corollary.

Corollary 3.6.1 *Let S be an Ω -allowed set and define*

$$\underline{\Omega} = \Omega(\cdot|S) + \Omega^{-S}.$$

Then, using the notation of Theorem 3.6.1 with $\delta_1 = \delta_2 := \delta$, we have for any β

$$2\delta\underline{\lambda} \underline{\Omega}(\hat{\beta} - \beta^0) \leq 2\delta\underline{\lambda} \underline{\Omega}(\beta - \beta^0) + \|X(\beta - \beta^0)\|_n^2 + \frac{\bar{\lambda}^2|S|}{\hat{\phi}_\Omega^2(L, S)} + 4\lambda\underline{\Omega}(\beta_{-S}). \quad (3.3)$$

Remark 3.6.1 *The good news is that the oracle inequalities thus hold for general norms. The bad news is that by the definition of an allowed set S*

$$\Omega \geq \underline{\Omega},$$

where

$$\underline{\Omega} := \Omega(\cdot|S) + \Omega^{-S}.$$

Hence in general the bounds for $\underline{\Omega}$ -estimation error (as given in Corollary 3.6.1) do not imply bounds for the Ω -estimation error of $\hat{\beta}$. As an illustration, we see in Example 3.9.2 ahead (Section 3.9) that Ω^{-S} can be very small when $|S|$ is large. Lemma 3.4.1 moreover shows that $\Omega_^{-S} \geq \Omega_*(\cdot - S)$, leading by the condition $\lambda > \Omega_*^{-S}(X_{-S}^T \epsilon)/n$ to a perhaps large tuning parameter.*

3.7 Norms stronger than ℓ_1

We say that the norm Ω is stronger than $\underline{\Omega}$ if $\Omega \geq \underline{\Omega}$. For such two norms the dual norm of Ω is weaker: $\Omega_* \leq \underline{\Omega}_*$. Thus, when Ω is stronger than the ℓ_1 -norm $\|\cdot\|_1$, Theorem 3.6.1 gives stronger results than Theorem 1.8.1 and its bounds on the tuning parameter λ are weaker. (This is modulo the behaviour of the compatibility constants: $\hat{\phi}^2$ and $\hat{\phi}_\Omega^2$ are generally not directly comparable.) Section 3.9 considers a general class of norms that are stronger than the ℓ_1 -norm.

With norms stronger than ℓ_1 one can apply the “conservative” ℓ_1 -based choice of the tuning parameter. This is important for the following reason. In view of Corollary 3.6.1, one would like to choose S in (3.3) in an optimal “oracle” way trading off the terms involved. But such a choice depends on the unknown β^0 . Hence we need to prove a bound for $\Omega_*^{-S}(X_{-S}^T \epsilon)/n$ which holds for all S which are allowed and which we want to include in our collection of candidate oracles. If the norm is stronger than the ℓ_1 -norm a value $\lambda > \lambda_\epsilon$ with λ_ϵ at least $\|X^T \epsilon\|_\infty/n$ works. This “conservative” choice is of course a bit too severe overruling of the noise and in that sense not optimal. There may be cases

where one still can use smaller values. Perhaps by using cross-validation one can escape from this dilemma. On the other hand, it is clear that the only gain when using some “optimal” tuning parameter is in the logarithmic terms and constants. Chapter 4 further examines the situation for a general class of norms (see in particular Corollaries 6.2.2 and 6.2.3 in Section 6.2).

3.8 Structured sparsity and square-root loss

Let Ω be a norm on \mathbb{R}^p . The topic of this section is the *square-root* Ω -structured sparsity estimator

$$\hat{\beta} := \arg \min_{\beta \in \mathbb{R}^p} \left\{ \|Y - X\beta\|_n + \lambda_0 \Omega(\beta) \right\}.$$

Let the residuals be

$$\hat{\epsilon} := Y - X\hat{\beta}.$$

The motivation for studying square-root quadratic loss is as before: it allows one to have a tuning parameter that does not depend on the scale of the noise. This motivation is perhaps less strong though, as we have seen in the previous sections (see also the discussion in Section 3.7) that the “good” (i.e. minimal yet effective) choice for the tuning parameter is more subtle as it may depend on the oracle. On the other hand, for certain examples (for instance the group-Lasso example (Example 3.9.1 in the next section) this is not an issue and square-root quadratic loss gives a universal choice *and* not overly conservative choice for the tuning parameter.

The idea in this section is as in Chapter 2 to first present an oracle inequality under the assumption that $\hat{\epsilon} \neq 0$, i.e. no overfitting. This is done in Subsection 3.8.1. Then Subsection 3.8.2 shows that indeed $\hat{\epsilon} \neq 0$ with high probability. Finally Subsection 3.8.3 combines the results. The arguments are throughout completely parallel to those for the square-root Lasso as presented in Chapter 2.

3.8.1 Assuming there is no overfitting

In this subsection we assume $\hat{\epsilon} \neq 0$. In the next section we show this is the case with high probability if β^0 is Ω -sparse.

We define

$$\hat{R} := \frac{\Omega_*(X^T \epsilon)}{n \|\epsilon\|_n}.$$

Moreover, for allowed sets S we define

$$\hat{R}_S := \frac{\Omega_*(X_S^T \epsilon)}{n \|\epsilon\|_n}, \quad \hat{R}^{-S} := \frac{\Omega_*^{-S}(X_{-S}^T \epsilon)}{n \|\epsilon\|_n}.$$

Proposition 3.8.1 *Suppose $\|\hat{\epsilon}\|_n > 0$. Consider an allowed set S . Let*

$$R_S \geq \hat{R}_S, \quad R^{-S} \geq \hat{R}^{-S},$$

Let $\delta_1 \geq 0$ and $0 \leq \delta_2 < 1$ be arbitrary. Take $\lambda_0 \|\hat{\epsilon}\|_n > R^{-S} \|\epsilon\|_n$ and define $\hat{\lambda}_L \|\epsilon\|_n := \lambda_0 \|\hat{\epsilon}\|_n - R^{-S} \|\epsilon\|_n$, $\hat{\lambda}_U \|\epsilon\|_n := \lambda_0 \|\hat{\epsilon}\|_n + R_S \|\epsilon\|_n + \delta_1 \hat{\lambda}_L \|\epsilon\|_n$ and

$$\hat{L} := \frac{\hat{\lambda}_U}{(1 - \delta_2) \hat{\lambda}_L}.$$

Then for any β we have

$$\begin{aligned} & 2\delta_1 \hat{\lambda}_L \|\epsilon\|_n \Omega(\hat{\beta}_S - \beta) + 2\delta_2 \hat{\lambda}_L \|\epsilon\|_n \Omega^{-S}(\hat{\beta}_{-S}) + \|X(\hat{\beta} - \beta^0)\|_n^2 \\ & \leq \|X(\beta - \beta^0)\|_n^2 + \frac{\hat{\lambda}_U^2 \|\epsilon\|_n^2 |S|}{\hat{\phi}_\Omega^2(\hat{L}, S)} + 4\lambda_0 \|\hat{\epsilon}\|_n \Omega(\beta_{-S}). \end{aligned}$$

Proof of Proposition 3.8.1. This follows by the same arguments as for Theorem 3.6.1 (see Subsection 3.10.3) and using the two point inequality (??). \square

3.8.2 Showing there is no overfitting

Conditions that ensure that $\hat{\epsilon} \neq 0$, and in fact $\|\hat{\epsilon}\|_n$ is close to $\|\epsilon\|_n$, are of the same flavour as for the square-root Lasso in Lemma 2.4.1.

Lemma 3.8.1 *Suppose that for some $0 < \eta < 1$, some $R > 0$ and some $\underline{\sigma} > 0$, we have*

$$\lambda_0(1 - \eta) \geq R$$

and

$$\lambda_0 \Omega(\beta^0) / \underline{\sigma} \leq 2 \left(\sqrt{1 + (\eta/2)^2} - 1 \right). \quad (3.4)$$

Then on the set where $\hat{R} \leq R$ and $\|\epsilon\|_n \geq \underline{\sigma}$ we have $\left| \|\hat{\epsilon}\|_n / \|\epsilon\|_n - 1 \right| \leq \eta$.

Proof of Lemma 3.8.1. This follows by exactly the same arguments as those used for Lemma 2.4.1. \square

3.8.3 A sharp oracle inequality

Putting the previous two subsections together yields the following oracle result.

Theorem 3.8.1 *Let S be an allowed set. Let for some positive constants R , R_S , R^{-S} , $0 < \eta < 1$ and $\underline{\sigma}$, the Ω -sparsity (3.4) hold, and*

$$R_S \geq \hat{R}_S, \quad R^{-S} \geq \hat{R}^{-S},$$

$$R \geq \hat{R}, \quad \|\epsilon\|_n \geq \underline{\sigma}$$

and

$$\lambda_0(1 - \eta) \geq \max\{R, R^{-S}\}$$

Define

$$\underline{\lambda}_0 := \lambda_0(1 - \eta) - R^{-S}, \quad \bar{\lambda}_0 := \lambda_0(1 + \eta)R_S + \delta_1 \underline{\lambda}_0$$

and

$$L := \frac{\bar{\lambda}_0}{(1 - \delta_2)\underline{\lambda}_0}.$$

Then for any β we have

$$\begin{aligned} & 2\delta_1 \underline{\lambda}_0 \|\epsilon\|_n \Omega(\hat{\beta}_S - \beta) + 2\underline{\lambda}_0 \|\epsilon\|_n \Omega_2(\hat{\beta}_{-S}) + \|X(\hat{\beta} - \beta^0)\|_n^2 \\ & \leq \|X(\beta - \beta^0)\|_n^2 + \frac{\bar{\lambda}_0^2 \|\epsilon\|_n^2 |S|}{\hat{\phi}_\Omega^2(L, S)} + 4\lambda_0(1 + \eta) \|\epsilon\|_n \Omega(\beta_{-S}). \end{aligned}$$

Proof of Theorem 3.8.1. This follows from Proposition 3.8.1 combined with Lemma 3.8.1. \square

3.9 Norms generated from cones

This section introduces a general class of norms for which the weak decomposability property, as presented in Definition 3.4.1, holds. The corresponding allowed sets are the sets which one believes to be candidate active sets.

Let \mathcal{A} be a convex cone in $\mathbb{R}_+^p =: [0, \infty)^p$. This cone is given beforehand and will describe the sparsity structure one believes is (approximately) valid for the underlying target β^0 .

Definition 3.9.1 *The norm Ω generated by the convex cone \mathcal{A} is*

$$\Omega(\beta) := \min_{a \in \mathcal{A}} \frac{1}{2} \sum_{j=1}^p \left[\frac{|\beta_j|^2}{a_j} + a_j \right], \quad \beta \in \mathbb{R}^p.$$

Here we use the convention $0/0 = 0$. If $\beta_j \neq 0$ one is forced to take $a_j \neq 0$ in the above minimum. It is shown in Micchelli et al. [2010] that Ω is indeed a norm. We present a proof for completeness.

Lemma 3.9.1 *The function Ω defined in Definition 3.9.1 above is a norm.*

Proof of Lemma 3.9.1. It is clear that $\Omega(\beta) \geq 0$ for all β and that it can only be zero when $\beta \equiv 0$. It is also immediate that the scaling property

$$\Omega(\lambda\beta) = \lambda\Omega(\beta), \quad \forall \lambda > 0, \beta \in \mathbb{R}^p,$$

holds, where we use that \mathcal{A} is a cone. The function $\beta \mapsto \Omega(\beta)$ is convex because $(a, b) \mapsto b^2/a$ and $a \mapsto a$ are convex functions and \mathcal{A} is convex. The triangle inequality follows from this and from the scaling property. \square

We call Ω the norm generated by the cone \mathcal{A} . One may verify that penalty proportional to the norm Ω generated by the convex cone \mathcal{A} favours sparse vectors which lie in \mathcal{A} . It is easy to see that the ℓ_1 -norm is a special case with $\mathcal{A} = \mathbb{R}_+^p$.

Having sparsity in mind, a minimal requirement seems to be that when coordinates are put to zero this does not increase the norm. This is indeed the case for a norm generated by a cone, as the following lemma shows.

Lemma 3.9.2 *For $J \subset \bar{J}$ we have*

$$\Omega(\cdot|J) \leq \Omega(\cdot|\bar{J}), \quad \Omega_*(\cdot|J) \geq \Omega_*(\cdot|\bar{J}).$$

Proof of Lemma 3.9.2. Let $\beta \in \mathbb{R}^p$ be arbitrary. For all $a \in \mathcal{A}$

$$\frac{1}{2} \sum_{j \in J} \left[\frac{\beta_j^2}{a_j} + a_j \right] \leq \frac{1}{2} \sum_{j=1}^p \left[\frac{\beta_j^2}{a_j} + a_j \right].$$

Hence also

$$\min_{a \in \mathcal{A}} \frac{1}{2} \sum_{j \in J} \left[\frac{\beta_j^2}{a_j} + a_j \right] \leq \min_{a \in \mathcal{A}} \frac{1}{2} \sum_{j=1}^p \left[\frac{\beta_j^2}{a_j} + a_j \right].$$

□

The rest of this section is organized as follows. First in Lemma 3.9.3 an alternative representation of the norm Ω generated by a cone is presented, and also the dual norm. Then Lemma 3.9.4 shows which sets S are allowed and the corresponding weak decomposability into $\Omega(\cdot|S)$ and Ω^{-S} . Then in Lemma 3.9.5 a bound for $\Omega(\cdot| - J)$ in terms of Ω^{-J} is given, for general sets J and hence in particular for allowed sets $J = S$. Lemma 3.9.6 states that $\underline{\Omega} := \Omega(\cdot|J) + \Omega^{-J}$ is stronger than the ℓ_1 -norm. We end the section with some examples.

Lemma 3.9.3 *We have*

$$\Omega(\beta) = \min_{a \in \mathcal{A}, \|a\|_1=1} \sqrt{\sum_{j=1}^p \frac{\beta_j^2}{a_j}} = \min_{a \in \mathcal{A}, \|a\|_1 \leq 1} \sqrt{\sum_{j=1}^p \frac{\beta_j^2}{a_j}} \quad (3.5)$$

and

$$\Omega_*(w) = \max_{a \in \mathcal{A}, \|a\|_1=1} \sqrt{\sum_{j=1}^p a_j w_j^2} = \max_{a \in \mathcal{A}, \|a\|_1 \leq 1} \sqrt{\sum_{j=1}^p a_j w_j^2}. \quad (3.6)$$

Proof. Exercise. □

For $J \subset \{1, \dots, p\}$ we set

$$\mathcal{A}_J = \{a_J : a \in \mathcal{A}\}.$$

Note that \mathcal{A}_J is a convex cone in $\mathbb{R}_+^{|J|}$ (whenever \mathcal{A} is one in \mathbb{R}_+^p). Denote the norm on $\mathbb{R}_+^{|J|}$ generated by \mathcal{A}_J as

$$\Omega^J(\beta_J) := \min_{a_J \in \mathcal{A}_J} \frac{1}{2} \sum_{j \in J} \left[\frac{|\beta_j|^2}{a_j} + a_j \right], \quad \beta \in \mathbb{R}^p.$$

Recall that a set S is called *allowed* if Ω is weakly decomposable for the set S .

Lemma 3.9.4 *If \mathcal{A}_S considered as subset of \mathbb{R}^p is a subset of \mathcal{A} we have the weak decomposability*

$$\Omega \geq \Omega(|S) + \Omega^{-S}$$

so that S is allowed.

Proof. Observe first that $\mathcal{A}_S \subset \mathcal{A}$ implies $\Omega(\beta_S) = \Omega^S(\beta_S)$. Moreover

$$\begin{aligned} \Omega(\beta) &\geq \min_{a \in \mathcal{A}} \frac{1}{2} \sum_{j \in S} \left[\frac{\beta_j^2}{a_j} + a_j \right] + \min_{a \in \mathcal{A}} \frac{1}{2} \sum_{j \notin S} \left[\frac{\beta_j^2}{a_j} + a_j \right] \\ &\geq \min_{a_S \in \mathcal{A}_S} \frac{1}{2} \sum_{j \in S} \left[\frac{\beta_j^2}{a_j} + a_j \right] + \min_{a_{-S} \in \mathcal{A}_{-S}} \frac{1}{2} \sum_{j \notin S} \left[\frac{\beta_j^2}{a_j} + a_j \right] \\ &= \Omega_S(\beta_S) + \Omega^{-S}(\beta_{-S}). \end{aligned}$$

□

Lemma 3.4.1 pointed out that in the case of an allowed set the Ω^{-S} -norm may be quite small. We now examine this for the special case of a norm generated by a cone.

Lemma 3.9.5 *Let \mathcal{E}^{-J} be the extreme points of the Ω^{-J} -unit ball. Then*

$$\Omega(\cdot | -J) \leq \omega^{-J} \Omega^{-J}$$

where $\omega^{-J} = \max\{\Omega(e^{-J} | -J) : e^{-J} \in \mathcal{E}^{-J}\}$.

Proof. Define $\omega := \max\{\Omega(\beta_{-J} | -J) : \Omega^{-J}(\beta_{-J}) = 1\}$. The maximum is attained in the extreme points of the Ω^{-J} -unit ball. □

Recall the bad news in Remark 3.6.1 that the oracle results of Theorem 3.6.1 and its relatives in general do not imply bounds for the Ω -estimation error. However, there is some good news too: they do imply bounds for the ℓ_1 -estimation error. This is clear from the next lemma.

Lemma 3.9.6 *For any set J ,*

$$\|\cdot\|_1 \leq \min\{\Omega(\cdot | J), \Omega^J\}.$$

Proof of Lemma 3.9.6. We clearly have

$$\Omega(\beta) = \min_{a \in \mathcal{A}} \frac{1}{2} \sum_{j=1}^p \left[\frac{|\beta_j|^2}{a_j} + a_j \right] \geq \min_{a \in \mathbb{R}_+^p} \frac{1}{2} \sum_{j=1}^p \left[\frac{|\beta_j|^2}{a_j} + a_j \right].$$

But for each j the minimum of

$$\frac{1}{2} \sum_{j=1}^p \left[\frac{|\beta_j|^2}{a_j} + a_j \right]$$

over $a_j \geq 0$ is equal to $|\beta_j|$. We apply this argument with Ω respectively replaced by $\Omega(\cdot|J)$ and Ω^J . \square

We give four examples from Micchelli et al. [2010].

Example 3.9.1 (Group Lasso penalty) Let $\{G_j\}_{j=1}^m$ be a partition of $\{1, \dots, p\}$ into m groups. The set \mathcal{A} consists of all non-negative vectors which are constant within groups. This gives

$$\Omega(\beta) := \sum_{j=1}^m \sqrt{|G_j|} \|\beta_{G_j}\|_2.$$

With squared error loss a penalty proportional to this choice of Ω is called the Group Lasso. It is introduced in Yuan and Lin [2006]. Oracle inequalities for the group Lasso have been derived in Lounici et al. [2011] for example. For the square-root version we refer to Bunea et al. [2014]. The dual norm is

$$\Omega_*(w) = \max_{1 \leq j \leq m} \|w_{G_j}\|_2 / \sqrt{|G_j|}.$$

Any union of groups is an allowed set and we moreover have for any allowed set S

$$\Omega^{-S} = \Omega(\cdot| - S)$$

and

$$\Omega = \Omega(\cdot|S) + \Omega^{-S}.$$

In other words, this norm is decomposable which frees it from the concerns expressed in Remark 3.6.1.

Example 3.9.2 (Wedge penalty) Consider the norm corresponding to the wedge penalty:

$$\mathcal{A} := \{a_1 \geq a_2 \geq \dots\}.$$

Let for some $s \in \mathbb{N}$, the set $S := \{1, \dots, s\}$ be the first s indices. Then S is an allowed set. To see that Ω^{-S} can be much smaller than $\Omega(\cdot| - S)$, take the vector $\beta \in \mathbb{R}^p$ to be one in its $s+1$ -th entry and zero elsewhere. Then $\Omega^{-S}(\beta) = 1$ but $\Omega(\beta| - S) = \sqrt{s+1}$.

Example 3.9.3 (DAG penalty) Let $\mathcal{A} = \{Aa \geq 0\}$ where A is the incidence matrix of a directed acyclic graph (DAG) with nodes $\{1, \dots, p\}$. Then removing orphans is allowed, i.e., successively removing nodes with only outgoing edges the remaining set is allowed at each stage.

Example 3.9.4 (Convexity inducing penalty) Let $\mathcal{A} := \{a_{k+2} - 2a_{k-1} + a_k \geq 0\}$.

3.10 Complements

3.10.1 The case where some coefficients are not penalized

Suppose the coefficients with index set $U \subset \{1, \dots, p\}$ are not penalized. The Ω -structured sparsity estimator is then

$$\hat{\beta} := \arg \min_{\beta \in \mathbb{R}^p} \left\{ \|Y - X\beta\|_n^2 + 2\lambda\Omega(\beta| - U) \right\}$$

where $\Omega(\beta| - U) := \Omega(\beta_{-U})$, $\beta \in \mathbb{R}^p$. We need the following result.

Lemma 3.10.1 *Suppose that $\Omega(\cdot| - U) \leq \Omega$. Then for all $z_{-U} \in \mathbb{R}^p$*

$$\Omega_*(z_{-U}| - U) = \Omega_*(z_{-U}).$$

Proof. By the definition of Ω_*

$$\Omega_*(z_{-U}) = \max_{\Omega(\beta) \leq 1} \beta^T z_{-U}.$$

Hence

$$\Omega_*(z_{-U}) \geq \max_{\Omega(\beta) \leq 1, \beta = \beta_{-U}} \beta^T z_{-U} = \max_{\Omega(\beta_{-U}) \leq 1} \beta_{-U}^T z_{-U} = \Omega_*(z_{-U}| - U).$$

On the other hand, the condition $\Omega(\cdot| - U) \leq \Omega$ implies

$$\Omega(\beta) \leq 1 \Rightarrow \Omega(\beta_{-U}) \leq 1$$

and therefore

$$\Omega_*(z_{-U}) \leq \max_{\Omega(\beta_{-U}) \leq 1} \beta_{-U}^T z_{-U} = \Omega_*(z_{-U}| - U).$$

□

When $\Omega(\cdot| - U) \leq \Omega$ the KKT-conditions are

$$X^T(Y - X\hat{\beta})/n + \lambda\hat{z}_{-U} = 0, \Omega_*(\hat{z}_{-U}) \leq 1, \hat{z}_{-U}\hat{\beta}_{-U} = \Omega(\hat{\beta}_{-U}).$$

3.10.2 The sorted ℓ_1 -norm

Let $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$ be a given increasing sequence. For $\beta = (\beta_1, \dots, \beta_p)^T \in \mathbb{R}^p$ we define the vector of absolute values in increasing order $|\beta|_{(1)} \geq |\beta|_{(2)} \geq \dots \geq |\beta|_{(p)}$. The sorted ℓ_1 -norm is

$$\Omega(\beta) = \sum_{j=1}^p \lambda_j |\beta|_{(j)}.$$

It was introduced in Bogdan et al. [2013]. In Zeng and Mario [2014] it is shown that this is indeed a norm and they provide its dual norm. We now show that this norm is weakly decomposable.

Lemma 3.10.2 *Let*

$$\Omega(\beta) = \sum_{j=1}^p \lambda_j |\beta|_{(j)},$$

and

$$\Omega^{-S}(\beta_{-S}) = \sum_{l=1}^r \lambda_{p-r+l} |\beta|_{(l,-S)},$$

where $r = p - s$ and $|\beta|_{(1,-S)} \geq \dots \geq |\beta|_{(r,-S)}$ is the ordered sequence in β_{-S} . Then $\Omega(\beta) \geq \Omega(\beta_S) + \Omega^{-S}(\beta_{-S})$. Moreover Ω^{-S} is the strongest norm among all $\underline{\Omega}^{-S}$ for which $\Omega(\beta) \geq \Omega(\beta_S) + \underline{\Omega}^{-S}(\beta_{-S})$

Proof of Lemma 3.10.2 . Without loss of generality assume $\beta_1 \geq \dots \geq \beta_p \geq 0$. We have

$$\Omega(\beta^S) + \Omega^{-S}(\beta_{-S}) = \sum_{j=1}^p \lambda_j \beta_{\pi(j)}$$

for a suitable permutation π . It follows that (Problem ??)

$$\Omega(\beta_S) + \Omega^{-S}(\beta_{-S}) \leq \Omega(\beta).$$

To show Ω^{-S} is the strongest norm it is clear we need only to search among candidates of the form

$$\underline{\Omega}^{-S}(\beta_{-S}) = \sum_{l=1}^r \lambda_{p-r+l} \beta_{\pi^{-S}(l)}$$

where $\{\lambda_{p-r+l}\}$ is a decreasing positive sequence and where $\pi^{-S}(1), \dots, \pi^{-S}(r)$ is a permutation of indices in S^c . This is then maximized by ordering the indices in S^c in decreasing order. But then it follows that the largest norm is obtained by taking $\lambda_{p-r+l} = \lambda_{p-r+l}$ for all $l = 1, \dots, r$. \square

3.10.3 A direct proof of Theorem 3.6.1

In stead of checking the conditions of the more general Theorem 5.5.1 we give here a direct proof. This also helps to follow the assertion of Proposition 3.8.1. We simplify the notation somewhat by writing $\Omega_2 := \Omega^{-S}$, $\lambda_1 := \lambda_S$ and $\lambda_2 := \lambda^{-S}$.

• If

$$\delta_1 \underline{\lambda} \Omega(\hat{\beta}_S - \beta_S) + \delta_2 \underline{\lambda} \Omega_2(\hat{\beta}_{-S} - \beta_{-S}) + (\hat{\beta} - \beta)^T \hat{\Sigma}(\hat{\beta} - \beta^0) \leq 2\lambda \Omega(\beta_{-S})$$

we know from the two point margin that

$$\begin{aligned} 2\delta_1 \underline{\lambda} \Omega(\hat{\beta}_S - \beta_S) &+ 2\delta_2 \underline{\lambda} \Omega_2(\hat{\beta}_{-S} - \beta_{-S}) + \|X(\hat{\beta} - \beta^0)\|_n^2 \\ &\leq \|X(\beta - \beta^0)\|_n^2 + 4\lambda \Omega(\beta_{-S}). \end{aligned}$$

• Suppose now that

$$\delta_1 \underline{\lambda} \Omega(\hat{\beta}_S - \beta_S) + \delta_2 \underline{\lambda} \Omega_2(\hat{\beta}_{-S} - \beta_{-S}) + (\hat{\beta} - \beta)^T \hat{\Sigma}(\hat{\beta} - \beta^0) \geq 2\lambda \Omega(\beta_{-S}). \quad (3.7)$$

By Lemma 3.3.1

$$\begin{aligned}
(\hat{\beta} - \beta)^T \hat{\Sigma}(\hat{\beta} - \beta^0) &\leq (\hat{\beta} - \beta)^T X^T \epsilon/n + \lambda \Omega(\beta) - \lambda \Omega(\hat{\beta}) \\
&\leq \lambda_1 \Omega(\hat{\beta}_S - \beta_S) + \lambda_2 \Omega_2(\hat{\beta}_{-S}) + (\lambda + \lambda_2) \Omega(\beta_{-S}) + \lambda \Omega(\beta_S) - \lambda \Omega(\hat{\beta}) \\
&\leq (\lambda + \lambda_1) \Omega(\hat{\beta}_S - \beta_S) - \underline{\lambda} \Omega_2(\hat{\beta}_{-S} - \beta_{-S}) + 2\lambda \Omega(\beta_{-S}).
\end{aligned}$$

We summarize this and give the inequality a number for reference:

$$(\hat{\beta} - \beta)^T \hat{\Sigma}(\hat{\beta} - \beta^0) \leq (\lambda + \lambda_1) \Omega(\hat{\beta}_S - \beta_S) - \underline{\lambda} \Omega_2(\hat{\beta}_{-S} - \beta_{-S}) + 2\lambda \Omega(\beta_{-S}). \quad (3.8)$$

From (3.7) we see that

$$(1 - \delta_2) \underline{\lambda} \Omega_2(\hat{\beta}_{-S} - \beta_{-S}) \leq \bar{\lambda} \Omega(\hat{\beta}_S - \beta_S)$$

or

$$\Omega_2(\hat{\beta}_{-S} - \beta_{-S}) \leq L \Omega(\hat{\beta}_S - \beta_S).$$

It follows that

$$\Omega(\hat{\beta}_S - \beta_S) \leq \sqrt{|S|} \|X(\hat{\beta} - \beta)\|_n / \hat{\phi}_\Omega(L, S).$$

But then, inserting (3.8),

$$\begin{aligned}
(\hat{\beta} - \beta)^T \hat{\Sigma}(\hat{\beta} - \beta^0) &+ \delta_1 \underline{\lambda} \Omega(\hat{\beta}_S - \beta_S) + \delta_2 \underline{\lambda} \Omega_2(\hat{\beta}_{-S} - \beta_{-S}) \\
&\leq \bar{\lambda} \Omega(\hat{\beta}_S - \beta_S) + 2\lambda \Omega(\beta_{-S}) \\
&\leq \bar{\lambda} \sqrt{|S|} \|X(\hat{\beta} - \beta)\|_n / \hat{\phi}_\Omega(L, S) + 2\lambda \Omega(\beta_{-S}) \\
&\leq \frac{1}{2} \frac{\bar{\lambda}^2 |S|}{\hat{\phi}_\Omega^2(L, S)} + \frac{1}{2} \|X(\hat{\beta} - \beta)\|_n^2 + 2\lambda \Omega(\beta_{-S}).
\end{aligned}$$

By the two point margin this gives

$$\begin{aligned}
&\|X(\hat{\beta} - \beta^0)\|_n^2 + 2\delta_1 \underline{\lambda} \Omega(\hat{\beta}_S - \beta_S) + 2\delta_2 \underline{\lambda} \Omega_2(\hat{\beta}_{-S} - \beta_{-S}) \\
&\leq \|X(\beta - \beta^0)\|_n^2 + \frac{\bar{\lambda}^2 |S|}{\hat{\phi}^2(L, S)} + 4\lambda \Omega(\beta_{-S}).
\end{aligned}$$

□

Chapter 4

Empirical process theory for dual norms

4.1 Introduction

Consider a vector $\epsilon \in \mathbb{R}^n$ with independent entries mean zero and variance σ_0^2 . We let X be a given $n \times p$ matrix. We are interested in the behaviour of $\Omega_*(X^T \epsilon)$ where Ω_* is the dual norm of Ω . Note that $X^T \epsilon$ is a p -dimensional random vector with components $X_j^T \epsilon$ where X_j is the j -th column of X ($j = 1, \dots, p$). For each j the random variable $W_j := X_j^T \epsilon / n$ is an average of n independent random variables with mean zero and variance $\sigma_0^2 \|X_j\|_n^2 / n$. Under suitable conditions, W_j has “Gaussian-type” behaviour. In this chapter, we assume for simplicity throughout that ϵ is Gaussian:

Condition 4.1.1 *The vector $\epsilon \in \mathbb{R}^n$ has a $\mathcal{N}_n(0, \sigma_0^2)$ -distribution.*

Then $X_j^T \epsilon$ is Gaussian as well and derivations are then simpler than for more general distributions. Although, the Gaussianity assumption is not crucial for the general picture, it does make a difference.

4.2 The dual norm of ℓ_1 and the scaled version

The dual norm of $\|\cdot\|_1$ is $\|\cdot\|_\infty$. We will derive the following corollary.

Corollary 4.2.1 *Let $\epsilon \sim \mathcal{N}_n(0, \sigma_0^2 I)$ and let X be a fixed $n \times p$ matrix with $\text{diag}(X^T X) / n = I$. Let $0 < \alpha < 1$ be a given error level. Then for*

$$\lambda_\epsilon := \sigma_0 \sqrt{\frac{2 \log(2p/\alpha)}{n}},$$

we have

$$\mathbb{P}\left(\|X^T \epsilon\|_\infty / n \geq \lambda_\epsilon\right) \leq \alpha.$$

The scaled version is

$$\hat{R} := \frac{\|X^T \epsilon\|_\infty / n}{\|\epsilon\|_n}.$$

We first present a probability inequality for the angle between a fixed and a random vector on the sphere in \mathbb{R}^n .

Lemma 4.2.1 *Let $\epsilon \sim \mathcal{N}_n(0, \sigma_0^2)$ where $n \geq 2$. Then for any $u \in \mathbb{R}^n$ with $\|u\|_n = 1$ and for all $0 < t < (n-1)/2$ we have*

$$\mathbb{P}\left(\frac{|u^T \epsilon|}{n\|\epsilon\|_n} > \sqrt{\frac{2t}{n-1}}\right) \leq 2 \exp[-t].$$

Proof of Lemma 4.2.1. Without loss of generality we may assume $\sigma_0 = 1$. Because $\epsilon/\|\epsilon\|_n$ is uniformly distributed on the sphere with radius \sqrt{n} in \mathbb{R}^n , we may without loss of generality assume that $u = \sqrt{n}e_1$, the first unit vector scaled with \sqrt{n} . Then $u^T \epsilon / (n\|\epsilon\|_n) = \epsilon_1 / (\sqrt{n}\|\epsilon\|_n) = \epsilon_1 / \sqrt{\sum_{i=1}^n \epsilon_i^2}$. It follows that for $0 < t < n/2$

$$\begin{aligned} \mathbb{P}\left(\frac{|u^T \epsilon|}{n\|\epsilon\|_n} \geq \sqrt{2t/n}\right) &= \mathbb{P}\left(\epsilon_1^2 \geq \frac{2t}{n} \sum_{i=1}^n \epsilon_i^2\right) \\ &= \mathbb{P}\left(\left(1 - \frac{2t}{n}\right) \epsilon_1^2 \geq \frac{2t}{n} \sum_{i=2}^n \epsilon_i^2\right) = \mathbb{P}\left(\epsilon_1^2 \geq \left(\frac{2t}{n-2t}\right) \sum_{i=2}^n \epsilon_i^2\right). \end{aligned}$$

The random variable $Z := \sum_{i=2}^n \epsilon_i^2$ has a χ^2 -distribution with $n-1$ degrees of freedom. It follows that for $v > 0$

$$\mathbb{E}e^{-vZ/2} = \left(\frac{1}{1+v}\right)^{\frac{n-1}{2}}.$$

We moreover have that for all $a > 0$,

$$\mathbb{P}(\epsilon_1^2 \geq 2a) \leq 2 \exp[-a].$$

So we find, with f_Z being the density of Z

$$\begin{aligned} \mathbb{P}\left(\epsilon_1^2 \geq \left(\frac{2t}{n-2t}\right) \sum_{i=2}^n \epsilon_i^2\right) &= \int_0^\infty \mathbb{P}\left(\epsilon_1^2 > \left(\frac{2tz}{n-2t}\right)\right) f_Z(z) dz \\ &= 2 \int_0^\infty \exp\left[-\frac{tz}{n-2t}\right] f_Z(z) dz \\ &= 2 \left(\frac{1}{1 + \frac{2t}{n-2t}}\right)^{\frac{n-1}{2}} = 2 \left(\frac{n-2t}{n}\right)^{\frac{n-1}{2}} \\ &\leq 2 \exp\left[-t \left(\frac{n-1}{n}\right)\right]. \end{aligned}$$

Finalize the proof by replacing t by $tn/(n-1)$. □

Lemma 4.2.2 *Let $\epsilon \sim \mathcal{N}_n(0, \sigma_0^2 I)$ and let X be a fixed $n \times p$ matrix with $\text{diag}(X^T X)/n = I$. Let α , $\underline{\alpha}$ and $\bar{\alpha}$ be given positive error levels. Define*

$$\underline{\sigma}^2 := \sigma_0^2 \left(1 - 2\sqrt{\frac{\log(1/\underline{\alpha})}{n}} \right),$$

$$\bar{\sigma}^2 := \sigma_0^2 \left(1 + 2\sqrt{\frac{\log(1/\bar{\alpha})}{n}} + \frac{2\log(1/\bar{\alpha})}{n} \right)$$

and

$$R := \sqrt{\frac{2\log(2p/\alpha)}{n-1}}.$$

We have

$$\mathbb{P}(\|\epsilon\|_n \leq \underline{\sigma}) \leq \alpha, \quad \mathbb{P}(\|\epsilon\|_n \geq \bar{\sigma}) \leq \bar{\alpha}$$

and

$$\mathbb{P}(\hat{R} \geq R) \leq \alpha.$$

Proof of Lemma 4.2.2. Without loss of generality we can assume $\sigma_0^2 = 1$. From Laurent and Massart [2000] we know that for all $t > 0$

$$\mathbb{P}\left(\|\epsilon\|_n^2 \leq 1 - 2\sqrt{t/n}\right) \leq \exp[-t]$$

and

$$\mathbb{P}\left(\|\epsilon\|_n^2 \geq 1 + 2\sqrt{t/n} + 2t/n\right) \leq \exp[-t].$$

A proof of the latter can also be found in Lemma 4.6.1

Apply this with $t = \log(1/\underline{\alpha})$ and $t = \log(1/\bar{\alpha})$ respectively. The bound for \hat{R} follows from Lemma 4.2.1 and the union bound. \square

4.3 Dual norms generated from cones

In Maurer and Pontil [2012] one can find first moment inequalities for a general class of dual norms. Here, we consider only a special case and we establish probability inequalities directly (i.e. not via concentration inequalities).

Let Ω be the norm generated by a given convex cone \mathcal{A} :

$$\Omega(\beta) := \min_{a \in \mathcal{A}} \frac{1}{2} \sum_{j=1}^p \left[\frac{\beta_j^2}{a_j} + a_j \right], \beta \in \mathbb{R}^p.$$

(see Section 3.9). Lemma 3.9.3 expresses the dual norm as

$$\Omega_*(w) = \max_{a \in \mathcal{A}, \|a\|_1=1} \sqrt{\sum_{j=1}^p a_j w_j^2}, \quad w \in \mathbb{R}^p.$$

Aim of the rest of this chapter is to bound $\Omega_*(W)$, with W_1, \dots, W_p random variables (in our setup, $W_j = X_j^T \epsilon/n$, $j = 1, \dots, p$). Recall that in order to simplify the exposition it is assumed that these are Gaussian random variables. The results can be extended to sub-Gaussian ones.

It is easy to see that $\Omega \geq \|\cdot\|_1$ and hence we have $\Omega_* \leq \|\cdot\|_\infty$. However, in some instances this bound can be improved. This is for example the case for the group Lasso, as we show below.

4.4 A generalized Bernstein inequality

In this section it is shown that under a condition on the moment generating function of a non-negative random variable Z one has a Bernstein-like inequality involving a sub-Gaussian part and a sub-exponential part. We apply this in the next section to squared Gaussians.

The following result can be deduced from in [Birgé and Massart [1998], Lemma 8 and its proof] or [Bühlmann and van de Geer [2011], Lemma 14.9 and its proof].

Lemma 4.4.1 *Let $Z \in \mathbb{R}$ be a random variable that satisfies for some K and c and for all $L > K$*

$$\mathbb{E} \exp[Z/L] \leq \exp \left[\frac{c}{(L^2 - LK)} \right].$$

Then for all $t > 0$

$$\mathbb{P} \left(Z \geq 2\sqrt{tc} + Kt \right) \leq \exp[-t].$$

Proof of Lemma 4.4.1. Let $a > 0$ be arbitrary and take

$$K/L = 1 - (1 + aK/c)^{-1/2},$$

apply Chebyshev's inequality to obtain

$$\mathbb{P} (Z \geq a) \leq \exp \left[-\frac{a^2}{aK + 2c + 2\sqrt{acK} + c^2} \right].$$

Now, choose $a = Kt + 2\sqrt{tc}$ to get

$$\mathbb{P} \left(Z \geq 2\sqrt{tc} + Kt \right) \leq \exp[-t].$$

□

Lemma 4.4.2 *Let $Z \in \mathbb{R}$ be a random variable that satisfies for a constant L_0*

$$C_0^2 := \mathbb{E} \exp[|Z|/L_0] < \infty.$$

Then for $L > 2L_0$

$$\mathbb{E} \exp[(Z - \mathbb{E}Z)/L] \leq \exp \left[\frac{2L_0^2 C_0^2}{L^2 - 2LL_0} \right].$$

Proof of Lemma 4.4.2. We have for $m \in \{1, 2, \dots\}$

$$\mathbb{E}|Z|^m \leq m!L_0^m C_0^2.$$

Hence

$$\mathbb{E}|Z - \mathbb{E}Z|^m \leq m!(2L_0)^m C_0^2.$$

So for $L < 2L_0$

$$\begin{aligned} \mathbb{E} \exp[(Z - \mathbb{E}Z)/L] &\leq 1 + \sum_{m=2}^{\infty} \frac{1}{m!L^m} \mathbb{E}|Z - \mathbb{E}Z|^m \leq 1 + \sum_{m=2}^{\infty} \left(\frac{2L_0}{L}\right)^m C_0^2 \\ &= 1 + \frac{2L_0^2 C_0^2}{L^2 - 2LL_0} \leq \exp\left[\frac{2L_0^2 C_0^2}{L^2 - 2LL_0}\right]. \end{aligned}$$

□

Combining Lemma 4.4.1 with Lemma 4.4.2 gives us back the following form of Bernstein's inequality.

Corollary 4.4.1 *Let Z_1, \dots, Z_n be independent random variables in \mathbb{R} that satisfy for some constant L_0*

$$C_0^2 := \max_{1 \leq i \leq n} \mathbb{E} \exp[|Z_i|/L_0] < \infty.$$

Then we can apply Lemma 4.4.1 with $K = 2L_0$ and $c = 2nL_0^2 C_0^2$ to find that for all $t > 0$

$$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n (Z_i - \mathbb{E}Z_i) \geq 2L_0 \left(C_0 \sqrt{2t/n} + t/n\right)\right) \leq \exp[-t].$$

4.5 Bounds for weighted sums of squared Gaussians

Consider p normally distributed random variables W_1, \dots, W_p , with mean zero and variance σ_0^2/n . Let $W := (W_1, \dots, W_p)^T$ be the p -dimensional vector collecting the W_j , $j = 1, \dots, p$. Let a_1, \dots, a_m be m given vectors in \mathbb{R}^p , with $\|a_l\|_1 = 1$ for $l = 1, \dots, m$.

Key ingredient of the proof of the next lemma is that for a $\mathcal{N}(0, 1)$ -distributed random variable V , the conditions of Lemma 4.4.1 hold with $K = 2$ if we take $Z = V^2 - 1$, see [Laurent and Massart [2000], Lemma 1 and its proof].

Lemma 4.5.1 *Let $0 < \alpha < 1$ be a given error level. Then for*

$$\lambda_\epsilon^2 := \frac{\sigma_0^2}{n} \left(1 + 2\sqrt{\log(m/\alpha)} + 2\log(m/\alpha)\right)$$

we have

$$\mathbb{P}\left(\max_{1 \leq l \leq m} \sum_{j=1}^p a_{j,l} W_j^2 \geq \lambda_\epsilon^2\right) \leq \alpha.$$

Lemma 4.5.1 is somewhat a quick and dirty lemma, although the bound is “reasonable”. As a special case, suppose that $a_j = e_j$, the j -th unit vector, $j = 1, \dots, m$, and $m = p$. Then we see that the bound of Corollary 4.2.1 in Section 4.2 is generally better than the one of the above lemma. Thus, since we know that the dual norm of a norm Ω generated by a convex cone is weaker than the $\|\cdot\|_\infty$ -norm, Lemma 4.5.1 is in general somewhat too rough.

Proof of Lemma 4.5.1. Write $V_j := \sqrt{n}W_j/\sigma_0$. First check that for all $L > 2$

$$\mathbf{E} \exp \left[(V_j^2 - 1)/L \right] \leq \exp \left[\frac{1}{L^2 - 2L} \right],$$

see also [Laurent and Massart [2000], Lemma 1 and its proof]. We moreover have for all l

$$\mathbf{E} \exp \left[\sum_{j=1}^p a_{j,l}(V_j^2 - 1)/L \right] = \mathbf{E} \left(\prod_{j=1}^p \exp \left[a_{j,l}(V_j^2 - 1)/L \right] \right).$$

We now use Hölder’s inequality, which says that for two random variables X and Y in \mathbb{R} , and for $0 < \alpha < 1$

$$\mathbf{E}|X|^\alpha|Y|^{1-\alpha} \leq (\mathbf{E}|X|)^\alpha(\mathbf{E}|Y|)^{1-\alpha}.$$

Hence also

$$\begin{aligned} \mathbf{E} \left(\prod_{j=1}^p \exp \left[a_{j,l}(V_j^2 - 1)/L \right] \right) &\leq \prod_{j=1}^p \left(\mathbf{E} \exp \left[(V_j^2 - 1)/L \right] \right)^{a_{j,l}} \\ &\leq \prod_{j=1}^p \left(\exp \left[\frac{1}{L^2 - 2L} \right] \right)^{a_{j,l}} = \exp \left[\frac{1}{L^2 - 2L} \right]. \end{aligned}$$

Therefore by Lemma 4.4.1, for all $t > 0$

$$\mathbf{P} \left(\sum_{j=1}^p a_{j,l}(V_j^2 - 1) > 2t + 2\sqrt{t} \right) \leq \exp[-t].$$

Apply the union bound to find that for all $t > 0$

$$\mathbf{P} \left(\max_{1 \leq l \leq m} \sum_{j=1}^p a_{j,l}(V_j^2 - 1) \geq 2\sqrt{t + \log(m)} + 2(t + \log m) \right) \leq \exp[-t].$$

Finally, take $t = \log(1/\alpha)$. □

4.6 The special case of χ^2 -random variables

We now reprove part of Lemma 1 in Laurent and Massart [2000]. This allows us a comparison with the results of the previous section.

Lemma 4.6.1 *Let χ_T^2 be a chi-squared distributed with m degrees of freedom. Then for all $t > 0$*

$$\mathbb{P}\left(\chi_T^2 \geq T + 2\sqrt{tT} + 2t\right) \leq \exp[-t].$$

Proof of Lemma 4.6.1. Let V_1, \dots, V_T be i.i.d. $\mathcal{N}(0, 1)$. Then (see the proof of Lemma 4.5.1)

$$\mathbb{E} \exp\left[(V_j^2 - 1)/L\right] \leq \exp\left[\frac{1}{L^2 - 2L}\right]$$

Hence, by the independence of the V_j ,

$$\mathbb{E} \exp\left[\sum_{j=1}^T (V_j^2 - 1)/L\right] \leq \exp\left[\frac{T}{L^2 - 2L}\right].$$

The result now follows from Lemma 4.4.1 (with $K = 2$ and $c = T$). \square

As a consequence, when one considers the maximum of a collection of chi-squared random variables, each with a relatively large number of degrees of freedom, one finds that the log-term in the bound becomes negligible.

Corollary 4.6.1 *Let, for $j = 1, \dots, m$, the random variables $\chi_{T_j}^2$ be chi-square distributed with T_j degrees of freedom. Define $T_{\min} := \min\{T_j : j = 1, \dots, m\}$. Let $0 < \alpha < 1$ be a given error level. Then for*

$$\lambda_0^2 := \frac{1}{n} \left(1 + 2\sqrt{\frac{\log(m/\alpha)}{T_{\min}}} + \frac{2\log(m/\alpha)}{T_{\min}}\right),$$

we have

$$\mathbb{P}\left(\max_{1 \leq j \leq m} \chi_{T_j}^2 / T_j \geq n\lambda_0^2\right) \leq \alpha.$$

4.7 The wedge dual norm

The wedge penalty is proportional to the norm

$$\Omega(\beta) = \min_{a \in \mathcal{A}, \|a\|_1=1} \sqrt{\sum_{j=1}^p \frac{\beta_j^2}{a_j}}, \quad \beta \in \mathbb{R}^p$$

with $\mathcal{A} := \{a_1 \geq \dots \geq a_p\}$ (see Example 3.9.2 in Section 3.9). Its dual norm is

$$\Omega_*(w) = \max_{a \in \mathcal{A}, \|a\|_1=1} \sqrt{\sum_{j=1}^p a_j w_j^2}, \quad w \in \mathbb{R}^p.$$

The maximum is attained in the extreme points of $\mathcal{A} \cap \{\|a\|_1 = 1\}$ so

$$\Omega_*(w) = \max_{1 \leq k \leq p} \sqrt{\sum_{j=1}^k \frac{w_j^2}{k}}, \quad w \in \mathbb{R}^p.$$

Lemma 4.7.1 *Let W_1, \dots, W_p be i.i.d. $\mathcal{N}(0, 1)$. Then for all $t > 0$*

$$\mathbb{P}\left(\max_{1 \leq k \leq p} \frac{1}{k} \sum_{j=1}^k W_j^2 \geq 1 + 2\sqrt{t} + 2t\right) \leq \frac{e^{-t}}{1 - e^{-t}}.$$

Proof of Lemma 4.7.1. By Lemma 4.6.1 we have for all k

$$\mathbb{P}\left(\frac{1}{k} \sum_{j=1}^k W_j^2 \geq 1 + 2\sqrt{t} + 2t\right) \leq \exp[-kt].$$

Hence

$$\mathbb{P}\left(\max_{1 \leq k \leq p} \frac{1}{k} \sum_{j=1}^k W_j^2 \geq 1 + 2\sqrt{t} + 2t\right) \leq \sum_{k=1}^p \exp[-kt] \leq \frac{e^{-t}}{1 - e^{-t}}.$$

□

Chapter 5

General loss with norm-penalty

5.1 Introduction

Let X_1, \dots, X_n be independent observations with values in some observation space \mathcal{X} and let for β in a space $\bar{\mathcal{B}} \subset \mathbb{R}^p$ be given a loss function $\rho_\beta : \mathcal{X} \rightarrow \mathbb{R}$. The parameter space \mathcal{B} is some given subset of $\bar{\mathcal{B}}$. The parameter space \mathcal{B} is potentially high-dimensional, so that possibly $p \gg n$. We require throughout convexity of parameter space and loss function. That is, we require Condition 5.1.1 without further explicit mentioning.

Condition 5.1.1 *The parameter space $\mathcal{B} \subset \bar{\mathcal{B}}$ is convex and the map*

$$\beta \mapsto \rho_\beta, \beta \in \mathcal{B}$$

is convex.

Define for all β in the extended space $\bar{\mathcal{B}}$ the empirical risk

$$R_n(\beta) := P_n \rho_\beta := \frac{1}{n} \sum_{i=1}^n \rho_\beta(X_i)$$

and the theoretical risk

$$R(\beta) := P \rho_\beta := \mathbb{E} R_n(\beta).$$

Let Ω be a norm on \mathbb{R}^p . This chapter studies the Ω -structured sparsity M -estimator

$$\hat{\beta} := \arg \min_{\beta \in \mathcal{B}} \left\{ R_n(\beta) + \lambda \Omega(\beta) \right\}.$$

with $\lambda > 0$ a tuning parameter.

The “true” parameter or “target” is defined as the minimizer of the theoretical risk over the extended space $\tilde{\mathcal{B}}$

$$\beta^0 := \arg \min_{\beta \in \tilde{\mathcal{B}}} R(\beta)$$

(where uniqueness is not required without expressing this in the notation). In many cases one simply is interested in the target with $\mathcal{B} = \tilde{\mathcal{B}}$.¹ On the other hand β^0 may be some more general reference value. As a look-ahead, the main result, Theorem 5.5.1 in Section 5.5.1 makes no explicit mention of any target β^0 (as it should be from a learning point of view). However, there is a mention of a local set $\mathcal{B}_{\text{local}}$. This generally points to a neighbourhood of some target β^0 .

5.2 Two point inequality, convex conjugate and two point margin

We first need to introduce a “local” set $\mathcal{B}_{\text{local}}$. Without further explicit mentioning, we require:

Condition 5.2.1 *The set $\mathcal{B}_{\text{local}}$ is a convex subset of \mathcal{B} .*

The set $\mathcal{B}_{\text{local}}$ is typically a neighbourhood of β^0 (for some suitable topology). The reason is that typically the conditions we will impose (to be precise, Condition 5.2.2) only hold locally. One then needs to prove that the estimator is in the local neighbourhood. Here one may exploit the assumed convexity of the loss. Section 5.6 illustrates how this works. There $\mathcal{B}_{\text{local}}$ is the set of $\beta' \in \mathcal{B}$ which are in a suitable Ω -norm close to β^0 . In the case of quadratic loss, one generally does not need to localize, i.e., then one can take $\mathcal{B}_{\text{local}} = \mathcal{B}$. For the moment we leave the form of the local set unspecified (but we do require its convexity).

In what follows we will use parameter values β and β' . The value β will represent a “candidate oracle”, that is, one should think of it as some fixed vector. The assumption $\beta \in \mathcal{B}_{\text{local}}$ is thus reasonable: candidate oracles are supposed to know how to get close to the target β^0 . The value β' typically represents the estimator $\hat{\beta}$. Thus the assumption $\beta' \in \mathcal{B}_{\text{local}}$ may mean that some work is to be done here.

Definition 5.2.1 *We call R_n right-differentiable if for all $\beta', \beta \in \mathcal{B}_{\text{local}}$*

$$\lim_{t \downarrow 0} \frac{R_n((1-t)\beta' + t\beta) - R_n(\beta')}{t} \leq \dot{R}_n(\beta')^T (\beta - \beta')$$

where $\dot{R}_n(\beta') \in \mathbb{R}^p$. We call $\dot{R}_n(\beta')$ the right-derivative of R_n at β' .

¹An example where this is not the case is where \mathcal{B} is a lower-dimensional subspace of $\tilde{\mathcal{B}}$. This is comparable to the situation where one approximates a function (an ∞ -dimensional object) by a p -dimensional linear function (with p large). Formally (since we take $\tilde{\mathcal{B}}$ finite-dimensional) we do not cover the latter case. This latter case does not really lead to additional theoretical complications, but seems to need cumbersome notations.

Lemma 5.2.1 (*Two point inequality*) Suppose R_n is right-differentiable and that $\hat{\beta} \in \mathcal{B}_{\text{local}}$. Then for all $\beta \in \mathcal{B}_{\text{local}}$

$$-\dot{R}_n(\hat{\beta})^T(\beta - \hat{\beta}) \leq \lambda\Omega(\beta) - \lambda\Omega(\hat{\beta}).$$

Proof of Lemma 5.2.1 . Let $\beta \in \mathcal{B}$ and define for $0 < t < 1$,

$$\hat{\beta}_t := (1 - t)\hat{\beta} + t\beta.$$

Recall that we require $\mathcal{B}_{\text{local}}$ to be convex, so $\hat{\beta}_t \in \mathcal{B}_{\text{local}}$ for all $0 < t < 1$. We have for $\text{pen} := \lambda\Omega$

$$R_n(\hat{\beta}) + \text{pen}(\hat{\beta}) \leq R_n(\hat{\beta}_t) + \text{pen}(\hat{\beta}_t) \leq R_n(\hat{\beta}_t) + (1 - t)\text{pen}(\hat{\beta}) + t\text{pen}(\beta).$$

Hence

$$\frac{R_n(\hat{\beta}) - R_n(\hat{\beta}_t)}{t} \leq \text{pen}(\beta) - \text{pen}(\hat{\beta}).$$

The results now follows by sending $t \downarrow 0$. □

Definition 5.2.2 (*Convex conjugate*) Let G be an increasing strictly convex non-negative function on $[0, \infty)$ with $G(0) = 0$. The convex conjugate of G is

$$H(v) := \sup_{u \geq 0} \left\{ uv - G(u) \right\}, \quad v \geq 0.$$

For example, the convex conjugate of the function $u \mapsto u^2/2$ is $v \mapsto v^2/2$.

Clearly, if H is the convex conjugate of G one has for all positive u and v

$$uv \leq G(u) + H(v).$$

This is the one-dimensional version of the so-called Fenchel-Young inequality.

We assume that R is differentiable with derivative \dot{R} at all $\beta \in \mathcal{B}_{\text{local}} \subset \mathcal{B}$.

Condition 5.2.2 (*Two point margin condition*) There is an increasing strictly convex non-negative function G with $G(0) = 0$ and a semi-norm τ on \mathcal{B} such that for all β and β' in $\mathcal{B}_{\text{local}}$ we have

$$R(\beta) - R(\beta') \geq \dot{R}(\beta')^T(\beta - \beta') + G(\tau(\beta - \beta')).$$

Note that $R(\cdot)$ is in view of our assumptions a convex function. One calls

$$B_R(\beta, \beta') := R(\beta) - R(\beta') - \dot{R}(\beta')^T(\beta - \beta'), \quad \beta, \beta' \in \mathcal{B}_{\text{local}}$$

the *Bregman divergence*. Convexity implies that

$$B_R(\beta, \beta') \geq 0, \quad \forall \beta, \beta' \in \mathcal{B}_{\text{local}}.$$

But the Bregman divergence is not symmetric in β and β' (nor does it satisfy the triangle inequality). The two point margin assumption thus assumes the

the Bregman divergence is lower bounded by a symmetric convex function. We present examples in Chapter ??.

We have in mind applying the two point margin condition at $\beta' = \hat{\beta}$ and $\beta = \beta^*$ where β^* is some “oracle” which trades off approximation error, effective sparsity and part of the vector β^* where the Ω -norm is smallish. Important to realize here is that the oracle β^* is a fixed vector. We note now that in the two point margin condition we assume the margin function G and the semi-norm τ not to depend on β' and β . The first (no dependence on β') is important, the last (no dependence on β) can be omitted (because we only need our conditions at a fixed value β^*). For ease of interpretation we refrain from the more general formulation.

5.3 Triangle property and effective sparsity

In this section we introduce the *triangle property* for general norms Ω . The triangle property is a major ingredient for proving sharp oracle inequalities, see Theorem 5.5.1 in Section 5.5. Section 5.4 shows that the triangle property holds for certain vectors which are either *allowed* or *allowed** (or both). Examples can be found in Chapter 6.

Definition 5.3.1 *Let Ω^+ and Ω^- be two semi-norms. We call them a complete pair if $\Omega^+ + \Omega^-$ is a norm.*

Definition 5.3.2 *We say that the triangle property holds at β if for a complete pair of semi-norms Ω_β^+ and Ω_β^- and $\Omega_\beta^- \neq 0$ one has*

$$\Omega(\beta) - \Omega(\beta') \leq \Omega_\beta^+(\beta' - \beta) - \Omega_\beta^-(\beta'), \quad \forall \beta' \in \mathbb{R}^p.$$

Note that in this definition one may choose for Ω_β^+ a very strong norm. This has its advantages (Theorem 5.5.1 then gives bounds for estimation error in a strong norm) but also a major disadvantage as for stronger norms Ω_β^+ the effective sparsity defined below will generally be larger.

In the next lemma, a vector β is written as the sum of two terms:

$$\beta = \beta^+ + \beta^-.$$

The situation we have in mind is the following. The vector β represents a candidate oracle. It may have a “good” sparsity-like part β^+ and a “bad” smallish-like part β^- . For the “good” part, the triangle property is assumed. The “bad” part of a candidate oracle better have small Ω -norm, otherwise this candidate oracle fails, i.e., it will not pass the test of being oracle. So we think of the situation where $\Omega(\beta^-)$ is small. The term $\Omega(\beta^-)$ is carried around in all the calculations: it is simply there without playing a very active role in the derivations.

Lemma 5.3.1 *Let $\beta = \beta^+ + \beta^-$ where β^+ has the triangle property and where $\Omega_{\beta^+}^+(\beta^-) = 0$. Then for any $\beta' \in \mathbb{R}^p$*

$$\Omega(\beta) - \Omega(\beta') \leq \Omega^+(\beta' - \beta) - \Omega^-(\beta' - \beta) + 2\Omega(\beta^-)$$

with $\Omega^+ = \Omega_{\beta^+}^+$ and $\Omega^- = \Omega_{\beta^+}^-$.

Proof of Lemma 5.3.1. We will first show that $\Omega^-(\beta^-) \leq \Omega(\beta^-)$. By applying the triangle property at $\beta' := \beta^+$ we obtain $0 \leq -\Omega^-(\beta^+)$. Hence $\Omega^-(\beta^+) = 0$. We next apply the triangle property at $\beta' := \beta^+ + \beta^-$. This gives

$$\Omega(\beta^+) - \Omega(\beta^+ + \beta^-) \leq \Omega^+(\beta^-) - \Omega^-(\beta^+ + \beta^-) = -\Omega^-(\beta^+ + \beta^-)$$

since by assumption $\Omega^+(\beta^-) = 0$. By the triangle inequality

$$\Omega^-(\beta^+ + \beta^-) \geq \Omega^-(\beta^-) - \Omega^-(\beta^+) = \Omega^-(\beta^-)$$

since we just showed that $\Omega^-(\beta^+) = 0$. Thus we have

$$\Omega(\beta^+) - \Omega(\beta^+ + \beta^-) \leq -\Omega^-(\beta^-).$$

On the other hand, by the triangle inequality

$$\Omega(\beta^+) - \Omega(\beta^+ + \beta^-) \geq -\Omega(\beta^-).$$

Combining the two gives indeed $\Omega^-(\beta^-) \leq \Omega(\beta^-)$.

Let now β' be arbitrary. By the triangle inequality

$$\Omega(\beta) - \Omega(\beta') \leq \Omega(\beta^+) + \Omega(\beta^-) - \Omega(\beta').$$

Apply the triangle property to find

$$\Omega(\beta) - \Omega(\beta') \leq \Omega^+(\beta^+ - \beta') - \Omega^-(\beta') + \Omega(\beta^-).$$

Then apply twice the triangle inequality to get

$$\begin{aligned} \Omega(\beta) - \Omega(\beta') &\leq \Omega^+(\beta - \beta') + \Omega^+(\beta^-) - \Omega^-(\beta - \beta') + \Omega^-(\beta) + \Omega(\beta^-) \\ &\leq \Omega^+(\beta - \beta') - \Omega^-(\beta - \beta') + 2\Omega(\beta^-), \end{aligned}$$

where in the last step we used that $\Omega^+(\beta^-) = 0$ and $\Omega^-(\beta) \leq \Omega^-(\beta^-) \leq \Omega(\beta^-)$. □

Definition 5.3.3 *Let β have the triangle property. For τ a semi-norm on \mathbb{R}^p and for a stretching factor $L > 0$, we define*

$$\Gamma_{\Omega}(L, \beta, \tau) := \left(\min \left\{ \tau(\tilde{\beta}) : \tilde{\beta} \in \mathbb{R}^p, \Omega_{\tilde{\beta}}^+(\tilde{\beta}) = 1, \Omega_{\tilde{\beta}}^-(\tilde{\beta}) \leq L \right\} \right)^{-1}.$$

We call $\Gamma_{\Omega}^2(L, \beta, \tau)$ the effective sparsity (for the norm Ω , the vector β , the stretching factor L and the semi-norm τ).

Effective sparsity is a generalization of compatibility. The reason for the (somewhat) new terminology is because the scaling by the size of some active set is no longer defined in this general context.

5.4 Two versions of weak decomposability

Definition 5.4.1 We call a vector β allowed if for a complete pair of seminorms Ω_β^+ and Ω_β^- with $\Omega_\beta^+(\beta) = \Omega(\beta)$, $\Omega_\beta^- \not\equiv 0$ and $\Omega_\beta^-(\beta) = 0$, one has

$$\Omega \geq \Omega_\beta^+ + \Omega_\beta^-.$$

We then call Ω weakly decomposable at β . If in fact we have equality: $\Omega = \Omega_\beta^+ + \Omega_\beta^-$, we call Ω decomposable at β .

Recall that for $\beta \neq 0$

$$\partial\Omega(\beta) = \{z \in \mathbb{R}^p : \Omega_*(z) = 1, z^T \beta = \Omega(\beta)\}.$$

Definition 5.4.2 We call a vector β allowed* if for a complete pair of seminorms Ω_β^+ and Ω_β^- with $\Omega_\beta^- \not\equiv 0$ one has for all $\beta' \in \mathbb{R}^p$

$$\min_{z \in \partial\Omega(\beta)} z^T (\beta - \beta') \leq \Omega_\beta^+(\beta' - \beta) - \Omega_\beta^-(\beta').$$

We then call Ω weakly decomposable* at β .

Lemma 5.4.1 Suppose β is an allowed or an allowed* vector. Then the triangle property holds at β :

$$\Omega(\beta) - \Omega(\beta') \leq \Omega_\beta^+(\beta' - \beta) - \Omega_\beta^-(\beta').$$

Proof of Lemma 5.4.1.

• If β is an allowed vector we have for any β' the inequality

$$\Omega(\beta) - \Omega(\beta') \leq \Omega(\beta) - \Omega_\beta^+(\beta') - \Omega_\beta^-(\beta') \leq \Omega_\beta^+(\beta' - \beta) - \Omega_\beta^-(\beta').$$

• If β is an allowed* vector we have for any $z \in \partial\Omega(\beta)$

$$\Omega(\beta) - \Omega(\beta') \leq z^T (\beta - \beta').$$

Hence

$$\Omega(\beta) - \Omega(\beta') \leq \min_{z \in \partial\Omega(\beta)} z^T (\beta - \beta') \leq \Omega_\beta^+(\beta' - \beta) - \Omega_\beta^-(\beta').$$

□

If we allow for a "good" and a "bad" part in the vector β we get:

Corollary 5.4.1 Let $\beta = \beta^+ + \beta^-$ where β^+ is allowed or allowed* and where $\Omega_{\beta^+}^+(\beta^-) = 0$. Then by Lemma 5.3.1 combined with Lemma 5.4.1 we have for any $\beta' \in \mathbb{R}^p$

$$\Omega(\beta) - \Omega(\beta') \leq \Omega^+(\beta' - \beta) - \Omega^-(\beta' - \beta) + 2\Omega(\beta^-)$$

with $\Omega^+ = \Omega_{\beta^+}^+$ and $\Omega_{\beta^+}^- = \Omega^-$.

We note that β allowed* does not imply β allowed (nor the other way around). In fact there are norms Ω where for all allowed* β

$$\Omega \leq \Omega_{\beta}^{+} + \Omega_{\beta}^{-}$$

i.e. \leq instead of \geq as is per definition the case for allowed vectors. Lemma 6.4.2 in Subsection 6.4.2 shows an example. Here Ω is the nuclear norm as defined there (Section 6.4).

5.5 A sharp oracle inequality

Notation for the candidate oracle In the next theorem we fix some $\beta \in \mathcal{B}_{\text{local}}$, a “candidate oracle”. We assume β to be the sum of two vectors $\beta = \beta^{+} + \beta^{-}$ where Ω has the triangle property at β^{+} and where $\Omega_{\beta^{+}}^{+}(\beta^{-}) = 0$. Write then $\Omega^{+} := \Omega_{\beta^{+}}^{+}$ and $\Omega^{-} := \Omega_{\beta^{+}}^{-}$. We let

$$\underline{\Omega} := \gamma\Omega_{\beta}^{+} + (1 - \gamma)\Omega_{\beta}^{-} =: \underline{\Omega}_{\beta^{+}}$$

be the strongest norm among all convex combinations $\gamma\Omega_{\beta}^{+} + (1 - \gamma)\Omega_{\beta}^{-}$, $\gamma \in [0, 1]$.

Theorem 5.5.1 *Assume R_n is right-differentiable and that Condition 5.2.2 (the two point margin condition) holds. Let H be the convex conjugate of G . Let*

$$\lambda_{\epsilon} \geq \underline{\Omega}_{*} \left(\dot{R}_n(\hat{\beta}) - \dot{R}(\hat{\beta}) \right). \quad (5.1)$$

Set $\lambda_1 := \lambda_{\epsilon}\gamma_{\beta^{+}}$ and $\lambda_2 := \lambda_{\epsilon}(1 - \gamma_{\beta^{+}})$. Take the tuning parameter λ large enough, so that $\lambda > \lambda_2$. Let $\delta_1 \geq 0$ and $0 \leq \delta_2 < 1$ be arbitrary and define

$$\underline{\lambda} := \lambda - \lambda_2, \quad \bar{\lambda} := \lambda + \lambda_1 + \delta_1 \underline{\lambda}$$

and stretching factor

$$L := \frac{\bar{\lambda}}{(1 - \delta_2)\underline{\lambda}}.$$

Then, when $\hat{\beta} \in \mathcal{B}_{\text{local}}$,

$$\begin{aligned} & \delta_1 \underline{\lambda} \Omega^{+}(\hat{\beta} - \beta) + \delta_2 \underline{\lambda} \Omega^{-}(\hat{\beta} - \beta) + R(\hat{\beta}) \\ & \leq R(\beta) + H \left(\bar{\lambda} \Gamma_{\Omega}(L, \beta^{+}, \tau) \right) + 2\lambda \Omega(\beta^{-}). \end{aligned}$$

Note that it is assumed that $\hat{\beta} \in \mathcal{B}_{\text{local}}$. Theorem 5.6.1 gives an illustration how this can be established. Note also that no reference is made to the target β^0 . However, in Theorem 5.6.1 $\mathcal{B}_{\text{local}}$ as some local neighbourhood of β^0 , so in the end the target *does* play a prominent role.

We need inequalities for $\underline{\Omega}_*(\dot{R}_n(\hat{\beta}) - \dot{R}(\hat{\beta}))$. This term occurs because in the proof of the theorem the dual norm inequality is applied:

$$(\dot{R}_n(\hat{\beta}) - \dot{R}(\hat{\beta}))^T(\hat{\beta} - \beta) \leq \underline{\Omega}_*(\dot{R}_n(\hat{\beta}) - \dot{R}(\hat{\beta}))\underline{\Omega}(\hat{\beta} - \beta).$$

This is in some cases too rough. An alternative route is possible.

We refer the a vector $\beta^* = \beta^{*+} + \beta^{*-}$ which trades off approximation error, estimation error (the term involving $H(\cdot)$ in Theorem 5.5.1) and Ω -smallish coefficients as the oracle.

Typically, the margin function G is quadratic, say $G(u) = u^2/2$, $u \geq 0$. Then its convex conjugate $H(v) = v^2/2$, $v \geq 0$ is quadratic as well. The estimation error is then

$$H\left(\bar{\lambda}\Gamma_{\Omega}(L, \beta^+, \tau)\right) = \bar{\lambda}^2\Gamma_{\Omega}^2(L, \beta^+, \tau).$$

Proof of Theorem 5.5.1. Define

$$\text{Rem}(\hat{\beta}, \beta) := R(\beta) - R(\hat{\beta}) - \dot{R}(\hat{\beta})^T(\beta - \hat{\beta}).$$

Then we have

$$R(\hat{\beta}) - R(\beta) + \text{Rem}(\hat{\beta}, \beta) = -\dot{R}(\hat{\beta})^T(\beta - \hat{\beta}).$$

• So if

$$\dot{R}(\hat{\beta})^T(\beta - \hat{\beta}) \geq \delta_1\lambda\Omega^+(\hat{\beta} - \beta) + \delta_2\lambda\Omega^-(\hat{\beta} - \beta) - 2\lambda\Omega(\beta^-)$$

we find from Condition 5.2.2

$$\delta_1\lambda\Omega^+(\hat{\beta} - \beta) + \delta_2\lambda\Omega^-(\hat{\beta} - \beta) + R(\hat{\beta}) \leq R(\beta) + 2\lambda\Omega(\beta^-)$$

(as $\text{Rem}(\hat{\beta}, \beta) \geq 0$). So then we are done.

• Assume now in the rest of the proof that

$$\dot{R}(\hat{\beta})^T(\beta - \hat{\beta}) \leq \delta_1\lambda\Omega^+(\hat{\beta} - \beta) + \delta_2\lambda\Omega^-(\hat{\beta} - \beta) - 2\lambda\Omega(\beta^-).$$

From Lemma 5.2.1

$$-\dot{R}_n(\hat{\beta})^T(\beta - \hat{\beta}) \leq \lambda\Omega(\beta) - \lambda\Omega(\hat{\beta}).$$

Hence by the dual norm inequality

$$\begin{aligned} & -\dot{R}(\hat{\beta})^T(\beta - \hat{\beta}) + \delta_1\lambda\Omega^+(\hat{\beta} - \beta) + \delta_2\lambda\Omega^-(\hat{\beta} - \beta) \\ & \leq (\dot{R}_n(\hat{\beta}) - \dot{R}(\hat{\beta}))^T(\beta - \hat{\beta}) + \delta_1\lambda\Omega^+(\hat{\beta} - \beta) + \delta_2\lambda\Omega^-(\hat{\beta} - \beta) \\ & \quad + \lambda\Omega(\beta) - \lambda\Omega(\hat{\beta}) \\ & \leq \lambda\epsilon\Omega(\hat{\beta} - \beta) + \delta_1\lambda\Omega^+(\hat{\beta} - \beta) + \delta_2\lambda\Omega^-(\hat{\beta} - \beta) + \lambda\Omega(\beta) - \lambda\Omega(\hat{\beta}) \\ & \leq \lambda_1\gamma_{\beta^+}\Omega^+(\hat{\beta} - \beta) + \lambda_2(1 - \gamma_{\beta^+})\Omega^-(\hat{\beta} - \beta) + \delta_1\lambda\Omega^+(\hat{\beta} - \beta) \\ & \quad + \delta_2\lambda\Omega^-(\hat{\beta} - \beta) + \lambda\Omega^+(\hat{\beta} - \beta) - \lambda\Omega^-(\hat{\beta} - \beta) + 2\lambda\Omega(\beta^-) \\ & = \bar{\lambda}\Omega^+(\hat{\beta} - \beta) - (1 - \delta_2)\lambda\Omega^-(\hat{\beta} - \beta) + 2\lambda\Omega(\beta^-) \end{aligned}$$

(here we applied Corollary 5.4.1). In summary

$$\begin{aligned} & -\dot{R}(\hat{\beta})^T(\beta - \hat{\beta}) + \delta_1 \underline{\lambda} \Omega^+(\hat{\beta} - \beta) + \delta_2 \underline{\lambda} \Omega^-(\hat{\beta} - \beta) \\ & \leq \bar{\lambda} \Omega^+(\hat{\beta} - \beta) - (1 - \delta_2) \underline{\lambda} \Omega^-(\hat{\beta} - \beta) + 2\lambda \Omega(\beta^-) \end{aligned} \quad (5.2)$$

But then

$$(1 - \delta_2) \underline{\lambda} \Omega^-(\beta - \beta) \leq \bar{\lambda} \Omega^+(\hat{\beta} - \beta)$$

or

$$\Omega^-(\hat{\beta} - \beta) \leq L \Omega^+(\hat{\beta} - \beta).$$

This implies by the definition of the effective sparsity $\Gamma_\Omega(L, \beta^+, \tau)$

$$\Omega^+(\hat{\beta} - \beta) \leq \tau(\hat{\beta} - \beta) \Gamma_\Omega(L, \beta^+, \tau).$$

Continuing with (5.2), we find

$$\begin{aligned} -\dot{R}(\hat{\beta})^T(\beta - \hat{\beta}) & + \underline{\lambda} \Omega^-(\hat{\beta} - \beta) + \delta_1 \underline{\lambda} \Omega^+(\hat{\beta} - \beta) \\ & \leq \bar{\lambda} \Omega^+(\hat{\beta} - \beta) + 2\lambda \Omega(\beta^-) \\ & \leq \bar{\lambda} \Gamma_\Omega(L, \beta^+, \tau) \tau(\hat{\beta} - \beta) + 2\lambda \Omega(\beta^-) \end{aligned}$$

or

$$\begin{aligned} R(\hat{\beta}) - R(\beta) & + \text{Rem}(\hat{\beta}, \beta) + \underline{\lambda} \Omega^-(\hat{\beta} - \beta) + \delta_1 \underline{\lambda} \Omega^+(\hat{\beta} - \beta) \\ & \leq \bar{\lambda} \Gamma_\Omega(L, \beta^+, \tau) \tau(\hat{\beta} - \beta) + 2\lambda \Omega(\beta^-) \\ & \leq H \left(\bar{\lambda} \Gamma_\Omega(L, \beta^+, \tau) \right) + G(\tau(\hat{\beta} - \beta)) + 2\lambda \Omega(\beta^-) \\ & \leq H \left(\bar{\lambda} \Gamma_\Omega(L, \beta^+, \tau) \right) + \text{Rem}(\hat{\beta}, \beta) + 2\lambda \Omega(\beta^-). \end{aligned}$$

□

5.6 Localizing (or a non-sharp oracle inequality)

This section considers the situation where one settles for showing that $\hat{\beta}$ is consistent in $\underline{\Omega}$ -norm. The local set $\mathcal{B}_{\text{local}}$ is taken in the set where $\hat{\beta}$ is $\underline{\Omega}$ -close to the candidate oracle β .

Theorem 5.6.1 below does not require differentiability of R_n and only needs Condition 5.2.2 at β' equal to β^0 . We call this the *one point margin condition*.

Condition 5.6.1 (*One point margin condition*) *There is an increasing strictly convex function G with $G(0) = 0$ and a semi-norm τ on \mathcal{B} such that for all $\beta \in \mathcal{B}_{\text{local}}$*

$$R(\beta) - R(\beta^0) \geq G(\tau(\beta - \beta^0)).$$

Notation for the candidate oracle We again fix some candidate oracle $\beta \in \mathcal{B}_{\text{local}}$ which we assume to be the sum $\beta = \beta^+ + \beta^-$ of two vectors β^+ and β^- with β^+ having the triangle property and with $\Omega_{\beta^+}^+(\beta^-) = 0$. Write then $\Omega^+ := \Omega_{\beta^+}^+$, $\Omega^- := \Omega_{\beta^+}^-$ and (for simplicity) $\underline{\Omega} := \Omega^+ + \Omega^-$.

Theorem 5.6.1 *Assume Condition 5.6.1 and let H be the convex conjugate of G . Suppose that for some constant $0 < M_{\max} \leq \infty$ and λ_ϵ and for all $0 < M \leq M_{\max}$*

$$\sup_{\beta' \in \mathcal{B}: \underline{\Omega}(\beta' - \beta) \leq M} \left| [R_n(\beta') - R(\beta')] - [R_n(\beta) - R(\beta)] \right| \leq \lambda_\epsilon M. \quad (5.3)$$

Let $0 < \delta < 1$, take $\lambda \geq 8\lambda_\epsilon/\delta$ and define M_β by

$$\delta\lambda M_\beta := 4H \left(\lambda(1 + \delta)\Gamma_\Omega \left(\frac{1}{1 - \delta}, \beta^+, \tau \right) \right) + 8 \left(R(\beta) - R(\beta^0) \right) + 16\lambda\Omega(\beta^-).$$

Assume that $M_\beta \leq M_{\max}$ and that $\{\beta' \in \mathcal{B} : \underline{\Omega}(\beta' - \beta) \leq M_\beta\} \subset \mathcal{B}_{\text{local}}$. Then $\underline{\Omega}(\hat{\beta} - \beta) \leq M_\beta$ and hence $\hat{\beta} \in \mathcal{B}_{\text{local}}$. Moreover, it holds that

$$R(\hat{\beta}) - R(\beta) \leq (\lambda_\epsilon + \lambda)M_\beta + \lambda\Omega^-(\beta).$$

Probability inequalities for the empirical process

$$\left\{ [R_n(\beta') - R(\beta')] - [R_n(\beta) - R(\beta)] : \underline{\Omega}(\beta' - \beta) \leq M, \beta' \in \mathcal{B} \right\}$$

(with $\beta \in \mathcal{B}$ and $M > 0$ fixed but arbitrary) will be provided. We note that - unlike Theorem 5.5.1 - Theorem 5.6.1 involves the approximation error $R(\beta) - R(\beta^0)$ and hence it only gives “good” results if the approximation error $R(\beta) - R(\beta^0)$ is “small”. Perhaps in contrast to general learning contexts, this is not too much of a restriction in certain cases. For example in linear regression with fixed design we have seen in Section 1.2 that high-dimensionality implies that the model is not misspecified.

Note that if $\mathcal{B} = \bar{\mathcal{B}}$, then the target $\beta^0 = \arg \min_{\beta \in \mathcal{B}} R(\beta)$ is by definition in the class \mathcal{B} . If one is actually interested in a target $\beta_0 = \min_{\beta \in \bar{\mathcal{B}}} R(\beta)$ outside the class \mathcal{B} , this target will generally have margin behaviour different from the minimizer within \mathcal{B} .

We remark here that we did not try to optimize the constants in Theorem 5.6.1.

Some explanation of the oracle we are trying to mimic here is in place. The oracle is some fixed vector $\beta^* = \beta^{*+} + \beta^{*-}$ satisfying the conditions as stated with $\Omega^+ := \Omega_{\beta^{*+}}^+$ and $\Omega^- := \Omega_{\beta^{*-}}^-$. We take β^* in such a way that $M_* := M_{\beta^*}$ is the smallest value among all β^* 's satisfying the conditions as stated and such that in addition $\underline{\Omega}(\beta^* - \beta^0) \leq M_*$ where $\underline{\Omega} = \Omega^+ + \Omega^-$, i.e. the oracle is in a suitable $\underline{\Omega}$ -neighbourhood of the target (note that $\underline{\Omega}$ depends on β^*). We define $\mathcal{B}_{\text{local}}$ as $\mathcal{B}_{\text{local}} := \mathcal{B} \cap \{\beta' : \underline{\Omega}(\beta' - \beta^0) \leq 2M_*\}$. Then obviously $\beta^* \in \mathcal{B}_{\text{local}}$ and by the triangle inequality $\{\beta' \in \mathcal{B} : \underline{\Omega}(\beta' - \beta^*) \leq M_*\} \subset \mathcal{B}_{\text{local}}$. Hence, then we may apply the above theorem with $\beta = \beta^*$. The situation simplifies drastically if one can choose β^0 itself as candidate oracle. See for example Subsection 6.3.1 for an illustration how Theorem 5.6.1 can be applied.

Proof of Theorem 5.6.1. To simplify the notation somewhat we write $M := M_\beta$. Define $\tilde{\beta} := t\hat{\beta} + (1-t)\beta$, where

$$t := \frac{M}{M + \underline{\Omega}(\hat{\beta} - \beta)}.$$

Then

$$\underline{\Omega}(\tilde{\beta} - \beta) = t\underline{\Omega}(\hat{\beta} - \beta) = \frac{M\underline{\Omega}(\hat{\beta} - \beta)}{M + \underline{\Omega}(\hat{\beta} - \beta)} \leq M.$$

Therefore $\tilde{\beta} \in \mathcal{B}_{\text{local}}$. Moreover, by the convexity of $R_n + \lambda\Omega$

$$\begin{aligned} R_n(\tilde{\beta}) + \lambda\Omega(\tilde{\beta}) &\leq tR_n(\hat{\beta}) + t\lambda\Omega(\hat{\beta}) + (1-t)R_n(\beta) + (1-t)\lambda\Omega(\beta) \\ &\leq R_n(\beta) + \lambda\Omega(\beta). \end{aligned}$$

Rewrite this and apply the assumption (5.3):

$$\begin{aligned} R(\tilde{\beta}) - R(\beta) &\leq -\left[[R_n(\tilde{\beta}) - R(\tilde{\beta})] - [R_n(\beta) - R(\beta)] \right] + \lambda\Omega(\beta) - \lambda\Omega(\tilde{\beta}) \\ &\leq \lambda_\epsilon M + \lambda\Omega(\beta) - \lambda\Omega(\tilde{\beta}) \\ &\leq \lambda_\epsilon M + \lambda\Omega^+(\tilde{\beta} - \beta) - \lambda\Omega^-(\tilde{\beta} - \beta) + 2\lambda\Omega^-(\beta), \end{aligned}$$

where we invoked Lemma 5.3.1.

- If $\lambda\Omega^+(\tilde{\beta} - \beta) \leq (1-\delta)[\lambda_\epsilon M + R(\beta) - R(\beta^0) + 2\lambda\Omega(\beta^-)]/\delta$, we obtain

$$\delta\lambda\Omega^+(\tilde{\beta} - \beta) \leq \lambda_\epsilon M + [R(\beta) - R(\beta^0)] + 2\lambda\Omega(\beta^-)$$

as well as

$$\delta\lambda\Omega^-(\tilde{\beta} - \beta) \leq \lambda_\epsilon M + [R(\beta) - R(\beta^0)] + 2\lambda\Omega(\beta^-).$$

So then

$$\delta\lambda(\Omega^+ + \Omega^-)(\tilde{\beta} - \beta) \leq 2\lambda_\epsilon M + 2[R(\beta) - R(\beta^0)] + 4\lambda\Omega(\beta^-).$$

- If $\lambda\Omega^+(\tilde{\beta} - \beta) \geq (1-\delta)[\lambda_\epsilon M + R(\beta) - R(\beta^0) + 2\lambda\Omega(\beta^-)]/\delta$ we obtain

$$[R(\tilde{\beta}) - R(\beta^0)] + \lambda\Omega^-(\tilde{\beta} - \beta) \leq \lambda\Omega^+(\tilde{\beta} - \beta)/(1-\delta).$$

So then we may apply effective sparsity with stretching factor $L = 1/(1-\delta)$. Hence

$$\begin{aligned} &[R(\tilde{\beta}) - R(\beta^0)] + \lambda\Omega^-(\tilde{\beta} - \beta) + \delta\lambda\Omega^+(\tilde{\beta} - \beta) \\ &\leq \lambda(1+\delta)\Omega^+(\tilde{\beta} - \beta) + \lambda_\epsilon M + [R(\beta) - R(\beta^0)] + 2\lambda\Omega(\beta^-) \\ &\leq \lambda(1+\delta)\tau(\tilde{\beta} - \beta)\Gamma_\Omega(1/(1-\delta), \beta^+, \tau) + \lambda_\epsilon M + [R(\beta) - R(\beta^0)] + 2\lambda\Omega(\beta^-) \\ &\leq H(\lambda(1+\delta)\Gamma_\Omega(1/(1-\delta), \beta^+, \tau)) + \lambda_\epsilon M + 2[R(\beta) - R(\beta^0)] + 2\lambda\Omega(\beta^-). \end{aligned}$$

It follows that

$$\delta\lambda(\Omega^+ + \Omega^-)(\tilde{\beta} - \beta) \leq \lambda\Omega^-(\tilde{\beta} - \beta) + \delta\lambda\Omega^+(\tilde{\beta} - \beta)$$

$$\leq H(\lambda\Gamma_\Omega(1/(1-\delta), \beta^+, \tau)) + \lambda_\epsilon M + 2[R(\beta) - R(\beta^0)] + 2\lambda\Omega(\beta^-).$$

Hence, we have shown in both cases that

$$\begin{aligned} \delta\lambda(\Omega^+ + \Omega^-)(\tilde{\beta} - \beta) &\leq H(\lambda(1+\delta)\Gamma_\Omega(1/(1-\delta), \beta^+, \tau)) \\ &\quad + 2[R(\beta) - R(\beta^0)] + 2\lambda_\epsilon M + 4\lambda\Omega(\beta^-) \\ &= \delta\lambda M/4 + 2\lambda_\epsilon M \leq \delta\lambda M/2 \end{aligned}$$

where we used the definition of M and that $\lambda \geq 8\lambda_\epsilon/\delta$. In turn, this implies that

$$(\Omega^+ + \Omega^-)(\hat{\beta} - \beta) \leq M.$$

For the second result of the theorem we apply the formula

$$\begin{aligned} R(\hat{\beta}) - R(\beta) &\leq -\left[[R_n(\hat{\beta}) - R(\hat{\beta})] - [R_n(\beta) - R(\beta)] \right] + \lambda\Omega(\beta) - \lambda\Omega(\hat{\beta}) \\ &\leq \lambda_\epsilon M + \lambda\Omega^+(\hat{\beta} - \beta) + 2\lambda\Omega^-(\beta) \\ &\leq (\lambda_\epsilon + \lambda)M + 2\lambda\Omega^-(\beta). \end{aligned}$$

□

Chapter 6

Some worked-out examples

6.1 The Lasso and square-root Lasso completed

We use the notation of Chapters 1 and 2. Recall the linear model

$$Y = X\beta^0 + \epsilon$$

with $\epsilon \sim \mathcal{N}_n(0, \sigma_0^2 I)$ and X a given $p \times n$ matrix. We assume $\text{diag}(X^T X)/n = I$. Define $W := X^T \epsilon/n = (W_1, \dots, W_p)^T$. Note that $W_j \sim \mathcal{N}(0, \sigma_0^2/n)$ for all j .

Combining Theorem 1.8.1 with Corollary 4.2.1 completes the result for the Lasso.

Corollary 6.1.1 *Let for some $0 < \alpha < 1$*

$$\lambda_\epsilon := \sigma_0 \sqrt{\frac{2 \log(2p/\alpha)}{n}}.$$

Let $0 \leq \delta < 1$ be arbitrary and define for $\lambda > \lambda_\epsilon$

$$\underline{\lambda} := \lambda - \lambda_\epsilon, \quad \bar{\lambda} := \lambda + \lambda_\epsilon + \delta \underline{\lambda}$$

and

$$L := \frac{\bar{\lambda}}{(1 - \delta)\underline{\lambda}}.$$

Then for all $\beta \in \mathbb{R}^p$ and all S we have with probability at least $1 - \alpha$

$$2\delta \underline{\lambda} \|\hat{\beta} - \beta\|_1 + \|X(\hat{\beta} - \beta^0)\|_n^2 \leq \|X(\beta - \beta^0)\|_n^2 + \frac{\bar{\lambda}^2 |S|}{\hat{\phi}^2(L, S)} + 4\lambda \|\beta_{-S}\|_1.$$

We now combine Theorem 2.5.1 with Lemma 4.2.2 to complete the result for the square-root Lasso.

Corollary 6.1.2 *Define for some positive α and $\underline{\alpha}$ satisfying $\alpha + \underline{\alpha} < 1$ the quantities*

$$R = \sqrt{\frac{2 \log(2p/\alpha)}{n-1}}, \quad \underline{\sigma}^2 := \sigma_0^2 \left(1 - 2\sqrt{\frac{\log(1/\underline{\alpha})}{n}}\right).$$

Assume for some $\eta > 0$

$$\lambda_0 \|\beta^0\|_1 \leq 2\sigma \left(\sqrt{1 + (\eta/2)^2} - 1 \right), \quad \lambda_0(1 - \eta) > R.$$

For arbitrary $0 \leq \delta < 1$ define

$$\begin{aligned} \underline{\lambda}_0 &:= \lambda_0(1 - \eta) - R, \\ \bar{\lambda}_0 &:= \lambda_0(1 + \eta) + R + \delta \underline{\lambda}_0 \end{aligned}$$

and

$$L := \frac{\bar{\lambda}_0}{(1 - \delta)\underline{\lambda}_0}.$$

Then for all β and S , with probability at least $1 - \alpha - \underline{\alpha}$ we have

$$\begin{aligned} 2\delta \underline{\lambda}_0 \|\hat{\beta} - \beta\|_1 \|\epsilon\|_n + \|X(\hat{\beta} - \beta^0)\|_n^2 \\ \leq \|X(\beta - \beta^0)\|_n^2 + \frac{\bar{\lambda}_0^2 |S| \|\epsilon\|_n^2}{\hat{\phi}^2(L, S)} + 4\lambda_0(1 + \eta) \|\epsilon\|_n \|\beta_{-S}\|_1. \end{aligned}$$

6.2 Least squares loss with Ω -structured sparsity completed

We use the notation of Chapter 3. Again the linear model is examined:

$$Y = X\beta^0 + \epsilon$$

with $\epsilon \sim \mathcal{N}_n(0, \sigma_0^2 I)$ and X and a given $p \times n$ matrix with $\text{diag}(X^T X)/n = I$. We set $W := X^T \epsilon/n = (W_1, \dots, W_p)^T$. As in Section 3.9 we let \mathcal{A} be a convex cone in $\mathbb{R}_+^p =: [0, \infty)^p$ and define

$$\Omega(\beta) := \min_{a \in \mathcal{A}} \frac{1}{2} \sum_{j=1}^p \left[\frac{|\beta_j|^2}{a_j} + a_j \right], \quad \beta \in \mathbb{R}^p.$$

For an allowed set S such that $\mathcal{A}_S \subset \mathcal{A}$ (see Lemma 3.9.4) we define $\mathcal{E}_S(\mathcal{A})$ as the set of extreme points of $\mathcal{A}_S \cap \{\|a_S\|_1 \leq 1\}$ and $\mathcal{E}^{-S}(\mathcal{A})$ as the set of extreme points of $\mathcal{A}_{-S} \cap \{\|a_{-S}\|_1 \leq 1\}$. We now assume both $\mathcal{E}_S(\mathcal{A})$ and $\mathcal{E}^{-S}(\mathcal{A})$ are finite and define for positive error levels α_1 and α_2 such that $\alpha_1 + \alpha_2 \leq 1$

$$\frac{n\lambda_S^2}{\sigma_0^2} := \min \left\{ \left(1 + 2\sqrt{\log \left(\frac{|\mathcal{E}_S(\mathcal{A})|}{\alpha_1} \right)} + 2 \log \left(\frac{|\mathcal{E}_S(\mathcal{A})|}{\alpha_1} \right) \right), 2 \log \left(\frac{2|S|}{\alpha_1} \right) \right\}$$

and

$$\begin{aligned} \frac{n(\lambda^{-S})^2}{\sigma_0^2} := \\ \min \left\{ \left(1 + 2\sqrt{\log \left(\frac{|\mathcal{E}^{-S}(\mathcal{A})|}{\alpha_2} \right)} + 2 \log \left(\frac{|\mathcal{E}^{-S}(\mathcal{A})|}{\alpha_2} \right) \right), 2 \log \left(\frac{2(p - |S|)}{\alpha_2} \right) \right\}. \end{aligned}$$

6.2. LEAST SQUARES LOSS WITH Ω -STRUCTURED SPARSITY COMPLETED 69

We obtain from Lemma 4.5.1 that with probability at least $1 - \alpha_1 - \alpha_2$,

$$\Omega_*(X_S^T \epsilon)/n \leq \lambda_S, \quad \Omega_*^{-S}(X_{-S}^T \epsilon)/n \leq \lambda^{-S}.$$

(Recall that $\Omega_*(X^T \epsilon)/n \leq \max\{\Omega_*(X_S^T \epsilon)/n, \Omega_*^{-S}(X_{-S}^T \epsilon)/n\}$, see Lemma 3.4.1.)

Theorem 3.6.1 then leads to the following corollary.

Corollary 6.2.1 *Let Ω be the norm generated from the convex cone \mathcal{A} and consider the Ω -structured sparsity estimator*

$$\hat{\beta} := \arg \min \left\{ \|Y - X\beta\|_n^2 + 2\lambda\Omega(\beta) \right\}.$$

Assume for all allowed sets S that $\mathcal{E}_S(\mathcal{A})$ and $\mathcal{E}^{-S}(\mathcal{A})$ are finite. Let, for allowed sets S , the constants λ_S and λ^{-S} be defined as above. Let $\delta_1 \geq 0$ and $0 \leq \delta_2 < 1$ be arbitrary. Take $\lambda > \max\{\lambda^{-S} : S \text{ allowed}\}$ and define

$$\underline{\lambda} := \lambda - \lambda^{-S}, \quad \bar{\lambda} := \lambda + \lambda_S + \delta_1 \lambda$$

and

$$L := \frac{\bar{\lambda}}{(1 - \delta_2)\underline{\lambda}}.$$

Then for any allowed set S and any β , with probability at least $1 - \alpha_1 - \alpha_2$ it holds that

$$\begin{aligned} 2\delta_1 \underline{\lambda} \Omega(\hat{\beta}_S - \beta) &+ 2\delta_2 \underline{\lambda} \Omega^{-S}(\hat{\beta}_{-S}) + \|X(\hat{\beta} - \beta^0)\|_n^2 \\ &\leq \|X(\beta - \beta^0)\|_n^2 + \frac{\bar{\lambda}^2 |S|}{\hat{\phi}_\Omega^2(L, S)} + 4\lambda \Omega(\beta_{-S}). \end{aligned}$$

For the group Lasso (see Example 3.9.1) we may improve the lower bound on the tuning parameter. We assume orthogonal design within groups $X_{G_t}^T X_{G_t}/n = I$. Equivalently, one may define the penalty as

$$\Omega_{\text{group}}(\beta) := \sum_{t=1}^T \sqrt{|G_t|} \|X\beta_{G_t}\|_n, \quad \beta \in \mathbb{R}^p.$$

Combining Corollary 4.6.1 with Theorem 3.6.1 we arrive at the following.

Corollary 6.2.2 *Consider the group Lasso as in Example 3.9.1:*

$$\hat{\beta} := \arg \min \left\{ \|Y - X\beta\|_n^2 + 2\lambda \Omega_{\text{group}}(\beta) \right\}.$$

Assume within-group orthogonal design. Let $0 \leq \delta < 1$ be arbitrary. Take

$$\lambda > \lambda_\epsilon := \frac{\sigma_0}{\sqrt{n}} \left(1 + 2\sqrt{\frac{\log(m/\alpha)}{T_{\min}}} + \frac{2\log(m/\alpha)}{T_{\min}} \right)^{1/2},$$

where T is the number of groups and $T_{\min} := \min\{|G_j| : j = 1, \dots, m\}$ is the minimal group size. Define

$$\underline{\lambda} := \lambda - \lambda_\epsilon, \quad \bar{\lambda} := \lambda + \lambda_\epsilon + \delta\underline{\lambda}$$

and

$$L := \frac{\bar{\lambda}}{(1 - \delta)\underline{\lambda}}.$$

Then for any allowed set S and any β , with probability at least $1 - \alpha$ it holds that

$$\begin{aligned} & 2\delta\underline{\lambda}\Omega_{\text{group}}(\hat{\beta} - \beta) + \|X(\hat{\beta} - \beta^0)\|_n^2 \\ & \leq \|X(\beta - \beta^0)\|_n^2 + \frac{\bar{\lambda}^2|S|}{\hat{\phi}_{\Omega_{\text{group}}}^2(L, S)} + 4\lambda\Omega_{\text{group}}(\beta_{-S}). \end{aligned}$$

The wedge penalty (see Example 3.9.2) corresponds to taking

$$\Omega_{\text{wedge}}(\beta) = \arg \min_{a_1 \geq \dots \geq a_p \geq 0, \|a\|_1=1} \sqrt{\sum_{j=1}^p \frac{\beta_j^2}{a_j}}, \quad \beta \in \mathbb{R}^p.$$

In the case of orthogonal design we have an improved version of the generic Corollary 6.2.1. For simplicity we take $\delta_1 = \delta_2 =: \delta$ and $\alpha_1 = \alpha_2 =: \alpha$ in this case.

Corollary 6.2.3 *Consider the wedge estimator from Example 3.9.2:*

$$\hat{\beta} := \arg \min \left\{ \|Y - X\beta\|_n^2 + 2\lambda\Omega_{\text{wedge}}(\beta) \right\}.$$

Let $0 \leq \delta < 1$ be arbitrary. Suppose orthogonal design: $\hat{\Sigma} = I$ (and hence $p \leq n$). Let $0 < \alpha < 1/2$. Take

$$\lambda > \lambda_\epsilon := \frac{\sigma_0}{\sqrt{n}} \left(1 + 2\sqrt{\log\left(\frac{1+\alpha}{\alpha}\right)} + 2\log\left(\frac{1+\alpha}{\alpha}\right) \right)^{1/2}.$$

Define

$$\underline{\lambda} := \lambda - \lambda_\epsilon, \quad \bar{\lambda} := \lambda + \lambda_\epsilon + \delta\underline{\lambda}$$

and

$$L := \frac{\bar{\lambda}}{(1 - \delta)\underline{\lambda}}.$$

Apply Lemma 4.7.1 to find that for any allowed set S and any β , with probability at least $1 - 2\alpha$ it holds that

$$\begin{aligned} & 2\delta\underline{\lambda}\Omega_{\text{wedge}}(\hat{\beta} - \beta) + \|X(\hat{\beta} - \beta^0)\|_n^2 \\ & \leq \|X(\beta - \beta^0)\|_n^2 + \frac{\bar{\lambda}^2|S|}{\hat{\phi}_{\Omega_{\text{wedge}}}^2(L, S)} + 4\lambda\Omega_{\text{wedge}}(\beta_{-S}) \end{aligned}$$

where $\underline{\Omega}_{\text{wedge}} = \Omega_{\text{wedge}}(\cdot|S) + \Omega_{\text{wedge}}^{-S}$.

6.3 Logistic regression

Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be independent observations, with $Y_i \in \{0, 1\}$ the response variable and $X_i \in \mathcal{X} \subset \mathbb{R}^p$ a co-variable ($i = 1, \dots, n$). The loss for logistic regression is

$$\rho_\beta(x, y) := -yx\beta + d(x\beta), \quad \beta \in \mathbb{R}^p$$

where

$$d(\xi) = \log(1 + e^\xi), \quad \xi \in \mathbb{R}.$$

We take the norm Ω in the penalty to be the ℓ_1 -norm. Furthermore, we impose no restrictions on β , i.e. $\mathcal{B} := \mathbb{R}^p$. The ℓ_1 -regularized logistic regression estimator is then

$$\hat{\beta} := \arg \min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{n} \sum_{i=1}^n \left[-Y_i X_i \beta + d(X_i \beta) \right] + \lambda \|\beta\|_1 \right\}.$$

Define for $i = 1, \dots, n$ and $x \in \mathcal{X}$

$$\mu_i^0(x) := \mathbb{E}(Y_i | X_i = x), \quad f_i^0(x) = \log\left(\frac{\mu_i^0(x)}{1 - \mu_i^0(x)}\right).$$

We assume the generalized linear model is well-specified: for some β^0

$$f^0(x) = x\beta^0, \quad \forall x \in \mathcal{X}.$$

In the high-dimensional situation with $\text{rank}(X) = n \leq p$, and with fixed design, we can take here $\mathcal{X} = \{X_1, \dots, X_n\}$ and then there always is a solution β^0 of the equation $f^0(x) = x\beta^0$, $x \in \mathcal{X}$. In what follows we consider fixed and random design, but in both cases we take the risk for fixed design, which we write as

$$R(\beta | X) := \frac{1}{n} \sum_{i=1}^n \left[-\dot{d}(X_i \beta^0) X_i \beta + d(X_i \beta) \right], \quad \beta \in \mathbb{R}^p.$$

We have

$$\ddot{d}(\xi) = \frac{e^\xi}{(1 + e^\xi)^2}.$$

It follows that for $\|\beta - \beta^0\|_1 \leq M$

$$\ddot{d}(x\beta) \geq 1/C_M^2(x),$$

where

$$\frac{1}{C_M^2(x)} = \left(\frac{1}{1 + e^{\beta^0 x + \|x^T\|_\infty M}} \right) \left(1 - \frac{1}{1 + e^{\beta^0 x - \|x^T\|_\infty M}} \right).$$

6.3.1 Logistic regression with fixed, bounded design

We assume that X is fixed and that for $\hat{\Sigma} := X^T X/n$, $\text{diag}(\hat{\Sigma}) = I$, i.e., the design is normalized. We write $K_1 := \max_{1 \leq i \leq n} |X_i|$ and $K_0 := \max_{1 \leq i \leq n} |f^0(X_i)|$ and

$$\frac{1}{C_M^2} := \left(\frac{1}{1 + e^{K_0 + K_1}} \right) \left(1 - \frac{1}{1 + e^{-K_0 - K_1}} \right).$$

Theorem 6.3.1 *Let $\lambda_\epsilon := \sqrt{2 \log(2p/\alpha)}$. Let further, for some $0 < \delta < 1$, $\lambda \geq 8\lambda_\epsilon/(1 - \delta)$. Define*

$$\underline{\lambda} := \lambda - \lambda_\epsilon, \quad \bar{\lambda} := \lambda + \lambda_\epsilon + \delta \underline{\lambda}$$

and

$$L := \frac{\bar{\lambda}}{(1 - \delta)\underline{\lambda}}.$$

Furthermore, define for any vector $\beta \in \mathbb{R}^p$ and set $S \subset \{1, \dots, p\}$.

$$\delta \lambda M_{\beta, S} := \frac{2C_1^2 \lambda^2 (1 + \delta^2) |S|}{\hat{\phi}^2(L, S)} + 8(R(\beta|X) - R(\beta^0|X)) + 16\lambda \|\beta_{-S}\|_1.$$

For those β and S such that $M_{\beta, S} \leq 1/2$ we have with probability at least $1 - \alpha$

$$\delta \underline{\lambda} \|\hat{\beta} - \beta\|_1 + R(\hat{\beta}|X) \leq R(\beta|X) + \frac{C^2 \bar{\lambda}^2 |S|}{2\hat{\phi}^2(L, S)} + 2\lambda \|\beta_{-S}\|_1.$$

6.4 Trace regression with nuclear norm penalization

Suppose

$$Y_i = \text{trace}(X_i B^0) + \epsilon_i, \quad i = 1, \dots, n,$$

where B^0 is a $p \times q$ matrix and X_i ($i = 1, \dots, n$) is a $q \times p$ matrix with $q \leq p$.

Writing

$$\tilde{X}_i \beta^0 := \text{trace}(X_i B^0),$$

where $\tilde{X}_i^T := \text{vec}(X_i^T)$, $\beta^0 := \text{vec}(B^0)$, we see that this is the linear model:

$$Y_1 = \tilde{X}_i \beta^0 + \epsilon_i, \quad i = 1, \dots, n.$$

The reason it is written in trace form is because actually the structure in β^0 is now not assumed to be in the sparsity of the coefficients, but rather in the sparsity of the singular values of B^0 . The norm induces this sparsity structure is the nuclear norm

$$\Omega(\beta) := \|B\|_{\text{nuclear}}, \quad B = \text{vec}^{-1}(\beta),$$

where $\|\cdot\|_{\text{nuclear}}$ is the nuclear norm. In what follows, we will identify matrices B with their vectorization $\text{vec}(B)$ and simply write $\Omega(B) = \|B\|_{\text{nuclear}}$. Recall that $\|\cdot\|_2$ is used as notation for the Frobenius norm when matrices are concerned. For a matrix A we let $\Lambda_{\max}^2(A)$ being the largest eigenvalue of $A^T A$.

6.4.1 Some useful matrix inequalities

Lemma 6.4.1 *Let A be a $p \times q$ matrix. Then*

$$\|A\|_{\text{nuclear}} \leq \sqrt{\text{rank}(A)} \|A\|_2.$$

Let P be a $p \times s$ matrix with $P^T P = I$ and $s \leq p$. Then

$$\|PP^T A\|_2 \leq \sqrt{s} \Lambda_{\max}(A)$$

and

$$\|PP^T A\|_2 \leq \|A\|_2.$$

Proof of Lemma 6.4.1. Let $r := \text{rank}(A)$. Write the singular value decomposition of A as

$$A = P_A \Lambda_A Q_A^T$$

with $P_A^T P_A = I$, $Q_A^T Q_A = I$ and $\Lambda_A = \Lambda_{A,1}, \dots, \Lambda_{A,r}$. Then $\|A\|_{\text{nuclear}} = \sum_{k=1}^r \Lambda_{A,k}$ and $\|A\|_2^2 = \text{trace}(A^T A) = \sum_{k=1}^r \Lambda_{A,k}^2$. The first result thus follows from $\|u\|_1 \leq \sqrt{r} \|u\|_2$ for a vector $u \in \mathbb{R}^r$.

For the second result we introduce the p -dimensional j -th unit vector e_j , ($j = 1, \dots, p$). Then

$$e_j^T PP^T AA^T PP^T e_j \leq \Lambda_{\max}^2(A) \|PP^T e_j\|_2^2$$

and hence

$$\begin{aligned} \|PP^T A\|_2^2 &= \text{trace}(PP^T AA^T PP^T) = \sum_{j=1}^p e_j^T PP^T AA^T PP^T e_j \\ &\leq \Lambda_{\max}^2(A) \sum_{j=1}^p \|PP^T e_j\|_2^2 = \Lambda_{\max}^2(A) \text{trace}(PP^T) \\ &= s \Lambda_{\max}^2(A). \end{aligned}$$

For the last result we write

$$\begin{aligned} \|A\|_2^2 &= \text{trace}(A^T A) = \text{trace}((PP^T A + (I - PP^T)^T A)^T (PP^T A + (I - PP^T) A)) \\ &= \text{trace}((PP^T A)^T (PP^T A)) + \text{trace}(((I - PP^T) A)^T (I - PP^T) A) \\ &\geq \text{trace}((PP^T A)^T (PP^T A)) = \|PP^T A\|_2^2. \end{aligned}$$

□

6.4.2 Dual norm of the nuclear norm and its triangle property

The dual norm of $\Omega = \|\cdot\|_{\text{nuclear}}$ is

$$\Omega_* = \Lambda_{\max}.$$

Moreover (see Watson [1992])

$$\partial\|B\|_{\text{nuclear}} = \{Z = PQ^T + (I - PP^T)W(I - QQ^T) : \Lambda_{\max}(W) = 1\}.$$

Let the $p \times q$ matrix B have rank s and singular value decomposition

$$B = P\Lambda Q^T,$$

with P a $p \times s$ matrix, Q a $q \times s$ matrix, $P^T P = I$, $Q^T Q = I$, and $\Lambda = \text{diag}(\Lambda_1, \dots, \Lambda_s)$ the diagonal matrix of non-zero singular values, where $\Lambda_1 \geq \dots \geq \Lambda_s > 0$.

Lemma 6.4.2 *The norm $\Omega = \|\cdot\|_{\text{nuclear}}$ has the triangle property at B , with*

$$\Omega_B^+(B') = \sqrt{s}(\|PP^T B'\|_2 + \|B'QQ^T\|_2 + \|PP^T B'QQ^T\|_2)$$

and

$$\Omega_B^-(B') = \|(I - PP^T)B'(I - QQ^T)^T\|_{\text{nuclear}}.$$

Moreover

$$\|\cdot\|_{\text{nuclear}} \leq \Omega_B^+ + \Omega_B^-.$$

Remark 6.4.1 *As for the last result, note the contrast with weakly decomposable norms as defined in Section 3.4, which have $\Omega \geq \Omega^+ + \Omega^-$.*

Proof of Lemma 6.4.2. Write for $Z \in \partial\|B\|_{\text{nuclear}}$

$$Z := Z_1 + Z_2, \quad Z_1 = PQ^T, \quad Z_2 = (I - PP^T)W(I - QQ^T).$$

We have

$$\begin{aligned} \text{trace}(Z_1^T B') &= \text{trace}(QP^T B') = \text{trace}(P^T B'Q) \\ &= \text{trace}(P^T PP^T B'QQ^T Q) = \text{trace}(QP^T PP^T B'QQ^T) \\ &\leq \|PP^T B'QQ^T\|_{\text{nuclear}} \end{aligned}$$

since $\Lambda_{\max}(PQ^T) = 1$. Moreover

$$\begin{aligned} \text{trace}(Z_2^T B') &= \text{trace}((I - QQ^T)W^T(I - PP^T)B') \\ &= \text{trace}(W^T(I - PP^T)B'(I - QQ^T)). \end{aligned}$$

Hence, there exists a W with $\Lambda_{\max}(W) = 1$ such that

$$\text{trace}(W^T B') = \|(I - PP^T)B'(I - QQ^T)\|_{\text{nuclear}}.$$

We thus see that (replacing B' by $B' - B$)

$$\begin{aligned}
 \max_{Z \in \partial \|B\|_{\text{nuclear}}} \text{trace}(Z^T (B' - B)) &= \max_{\Lambda_{\max}(W)=1} \text{trace}((I - QQ^T)W^T(I - PP^T)(B' - B)) \\
 &+ \text{trace}(QP^T(B' - B)) \\
 &\geq \|(I - PP^T)B'(I - QQ^T)\|_{\text{nuclear}} - \|PP^T(B' - B)QQ^T\|_{\text{nuclear}}.
 \end{aligned}$$

Now use Lemma 6.4.1 to get

$$\|PP^T(B' - B)QQ^T\|_{\text{nuclear}} \leq \sqrt{s}\|PP^T(B' - B)QQ^T\|_2 \leq \Omega^+(B' - B).$$

Obtaining the second result of the lemma is almost trivial: for all B'

$$\begin{aligned}
 \|B'\|_{\text{nuclear}} &= \|PP^TB' + B'QQ^T - PP^TB'QQ^T + (I - PP^T)B'(I - QQ^T)\|_{\text{nuclear}} \\
 &\leq \|PP^TB'\|_{\text{nuclear}} + \|B'QQ^T\|_{\text{nuclear}} \\
 &+ \|PP^TB'QQ^T\|_{\text{nuclear}} + \|(I - PP^T)B'(I - QQ^T)\|_{\text{nuclear}} \\
 &\leq \sqrt{s}(\|PP^TB'\|_2 + \|B'QQ^T\|_2 + \|PP^TB'QQ^T\|_2) \\
 &+ \|(I - PP^T)B'(I - QQ^T)\|_{\text{nuclear}}
 \end{aligned}$$

where we invoked Lemma 6.4.1. \square

Lemma 6.4.3 *Let*

$$\underline{\Omega} := \Omega_B^+ + \Omega_B^-$$

with Ω_B^+ and Ω_B^- as in Lemma 6.4.2 Then

$$\underline{\Omega}_*(\cdot) \leq \Lambda_{\max}(\cdot).$$

Proof of Lemma 6.4.3. This follows from $\|\cdot\|_{\text{nuclear}} \leq \underline{\Omega}$ (see Lemma 6.4.2) and the fact that the nuclear norm has dual norm Λ_{\max} . \square

Notation for the candidate oracle We will next provide the notation for the candidate oracle B which we might aim at mimicking. Recall that $q \leq p$. Let

$$B = P\Lambda Q^T$$

with P a $p \times q$ matrix, Q a $q \times q$ matrix, $P^TP = I$, $Q^TQ = I$, and $\Lambda = \text{diag}(\Lambda_1, \dots, \Lambda_q)$ where $\Lambda_1 \geq \dots \geq \Lambda_q$.

Write

$$B = B^+ + B^-, \quad B^+ = \sum_{k=1}^s \Lambda_k P_k Q_k^T, \quad B^- = \sum_{k=s+1}^q \Lambda_k P_k Q_k^T. \quad (6.1)$$

We see that

$$\|B^-\|_{\text{nuclear}} = \sum_{k=s+1}^q \Lambda_k, \quad \Omega_{B^+}^+(B^-) = 0.$$

Define $\underline{\Omega} := \Omega_{B^+}^+ + \Omega_{B^+}^-$.

6.4.3 An oracle result for trace regression with least squares loss

We consider the nuclear norm regularized estimator

$$\hat{B} := \arg \min_B \left\{ \sum_{i=1}^n (Y_i - \text{trace}(X_i B))^2 / n + 2\lambda \|B\|_{\text{nuclear}} \right\}.$$

Definition 6.4.1 *Let $L > 0$ be some stretching factor. Suppose B has singular value decomposition $P\Lambda Q^T$. Let $s := \text{rank}(B)$. We define the $\|\cdot\|_{\text{nuclear}}$ -compatibility constant at B as*

$$\hat{\phi}_{\text{nuclear}}^2(L, B) := \min \left\{ \begin{array}{l} \frac{s}{n} \sum_{i=1}^n \text{trace}^2(X_i B') : \\ \sqrt{s}(\|PP^T B'\|_2 + \|B'QQ^T\|_2 + \|PP^T B'QQ^T\|_2) = 1, \\ \|(I - PP)^T B'(I - QQ)^T\|_{\text{nuclear}} \leq L \end{array} \right\}.$$

Corollary 6.4.1 *Application of Theorem 5.5.1 to the nuclear norm penalty gives the following. Let $B = B^+ + B^-$ where B^+ and B^- are given in (6.1). Let now*

$$\lambda_\epsilon \geq \Lambda_{\max} \left(\sum_{i=1}^n \epsilon_i X_i \right) / n.$$

For $\lambda > \lambda_\epsilon$, $\underline{\lambda} := \lambda - \lambda_\epsilon$, $\bar{\lambda} := \lambda + \lambda_\epsilon + \delta \underline{\lambda}$, $L := \bar{\lambda} / ((1 - \delta)\underline{\lambda})$, we have

$$\begin{aligned} \delta \underline{\lambda} \Omega (\hat{B} - B)_{\text{nuclear}} &+ \frac{1}{n} \sum_{i=1}^n \text{trace}^2(X_i (\hat{B} - B^0)) / n \\ &\leq \frac{1}{n} \sum_{i=1}^n \text{trace}^2(X_i (B - B^0)) + \frac{s \bar{\lambda}^2}{\hat{\phi}_{\text{nuclear}}^2(L, B^+)} + 4\lambda \|B^-\|_{\text{nuclear}}. \end{aligned}$$

We refer to Section ?? for a probability inequality for the maximal eigenvalue $\Lambda_{\max}(\sum_{i=1}^n \epsilon_i X_i) / n$ in the context of matrix completion.

Recall that (see Lemma 6.4.2) $\|\cdot\|_{\text{nuclear}} \leq \underline{\Omega}$. Hence from Corollary 6.4.1 one may also establish a bound for the nuclear norm estimation error.

6.4.4 Robust matrix completion

Let \mathcal{B} be the collection of $p \times q$ matrices with all entries bounded by some constant $\eta > 0$:

$$\mathcal{B} := \{B : \|B\|_\infty \leq \eta\}.$$

The bounded parameter space \mathcal{B} allows one to take $\mathcal{B}_{\text{local}} = \mathcal{B}$ when applying Theorem 5.6.1. We will not prove a sharp oracle inequality in this subsection because the loss is not twice differentiable. We conjecture though that lack of

differentiability per se is not a reason for impossibility of sharp oracle inequalities.

Let \mathcal{X} be the space of all $p \times q$ matrices X consisting of zeroes at all entries except for a single entry at which the value equal to one:

$$X = \begin{pmatrix} 0 & \cdots & 0 & \cdots & 0 & 0 \\ \vdots & & \vdots & & \vdots & \vdots \\ 0 & \cdots & 1 & \cdots & 0 & 0 \\ 0 & \cdots & 0 & \cdots & 0 & 0 \\ \vdots & & \vdots & & \vdots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & 0 \end{pmatrix}.$$

Such matrices - called masks - have also been studied in Section ???. There are $q \times p$ such matrices. We let $\{X_1, \dots, X_n\}$ be i.i.d. with values in \mathcal{X} . Consider the least absolute deviations estimator

$$\hat{B} := \arg \min_{B \in \mathcal{B}} \left\{ \frac{1}{n} \sum_{i=1}^n |Y_i - \text{trace}(X_i B)| + \lambda \|B\|_{\text{nuclear}} \right\}.$$

Theorem 6.4.1 *Let B be given in (6.1). Suppose that $\epsilon_1, \dots, \epsilon_n$ are i.i.d. with median zero and with density f_ϵ with respect to Lebesgue measure. Assume that for some positive constant C and some $\eta > 0$.*

$$f_\epsilon(u) \geq 1/C^2 \quad \forall |u| \leq 2\eta.$$

Define for C_0 is suitable universal constant

$$\begin{aligned} \lambda_\epsilon &:= 4C_0 \sqrt{\frac{1}{q}} \sqrt{\frac{\log(p+q)}{n}} \\ &+ 4C_0 \sqrt{\log(1+q)} \left(\frac{\log(p+q)}{n} \right) + \sqrt{\frac{8 \log(1/\alpha)}{n}}. \end{aligned}$$

Take for some $0 < \delta < 1$ $\lambda \geq 8\lambda_\epsilon/\delta$ and define M_B by

$$\delta \lambda M_B = 6C^2 \lambda^2 (1+\delta)^2 pqs + 8 \left(R(B) - R(B^0) \right) + 16\lambda \|B^-\|_{\text{nuclear}}.$$

Then with probability at least $1 - \alpha$ we have $\underline{\Omega}(\hat{B} - B) \leq M_B$ and

$$R(\hat{B}) - R(B) \leq (\lambda_\epsilon + \lambda) M_B + 2\lambda \|B^-\|_{\text{nuclear}}.$$

Asymptotics and weak sparsity Suppose that $q \log(1+q)$ is of small order $n/\log p$. Theorem 6.4.1 shows that for a suitable value for the tuning parameter λ of order $\lambda \asymp \sqrt{\log p/nq}$ one has

$$R(\hat{B}) - R(B^0) = \mathcal{O}_{\mathbf{P}} \left(\frac{ps \log p}{n} + R(B) - R(B^0) + \sqrt{\frac{\log p}{nq}} \|B^-\|_{\text{nuclear}} \right).$$

This implies

$$\|\hat{B} - B_0\|_2^2 = \mathcal{O}_{\mathbf{P}}\left(\frac{p^2qs \log p}{n} + pq(R(B) - R(B^0)) + p\sqrt{q}\sqrt{\frac{\log p}{n}}\|B^-\|_{\text{nuclear}}\right).$$

For example, taking $B = B^0$ and letting s_0 be the rank of B^0 , we get

$$\|\hat{B} - B_0\|_2^2 = \mathcal{O}_{\mathbf{P}}\left(\frac{p^2qs \log p}{n}\right).$$

Admittedly, this is a slow rate, but this is as it should be. For each parameter, the rate of estimation is $\sqrt{pq/n}$ because we have only about $n/(pq)$ noisy observations of this parameter. Without penalization, the rate in squares Frobenius norm would thus be

$$pq \times \frac{pq}{n} = \frac{p^2q^2}{n}.$$

With penalization, the estimator mimicks an oracle that only has to estimate ps_0 (instead of pq) parameters, with a $\log p$ -prize to be paid.

Instead of assuming B^0 itself is of low rank, one may assume it is only weakly sparse. Let B^0 have singular values $\{\Lambda_k^0\}_{k=1}^q$. Fix some $0 < r < 1$ and let

$$\rho_r^r := \sum_{k=1}^q |\Lambda_k^0|^r.$$

Then we obtain (Problem ??) (take $B = B^0$ and use the same arguments as in Lemma 1.10.1 in Section 1.10)

$$\|\hat{B} - B^0\|_2^2 = \mathcal{O}_{\mathbf{P}}\left(\frac{p^2q \log p}{n}\right)^{1-r} \rho_r^{2r}. \quad (6.2)$$

6.5 Sparse principal components

Consider an $n \times p$ matrix X with i.i.d. rows $\{X_i\}_{i=1}^n$. Let $\hat{\Sigma} := X^T X/n$ and $\Sigma_0 := \mathbf{E}\hat{\Sigma}$. In this section the estimation of the first principal component $q^0 \in \mathbb{R}^p$ corresponding to the largest eigenvalue $\phi_{\max}^2 := \Lambda_{\max}(\Sigma_0)$ of Σ_0 is studied. The parameter of interest is $\beta^0 := q^0 \phi_{\max}$, so that $\|\beta^0\|_2^2 = \phi_{\max}^2$ (since the eigenvector q^0 is normalized to have $\|\cdot\|_2$ -length one). It is assumed that β^0 is sparse.

Denote the Frobenius norm of a matrix A by $\|A\|_2$:

$$\|A\|_2^2 := \sum_j \sum_k A_{j,k}^2.$$

We use the ℓ_1 -penalized estimator

$$\hat{\beta} := \arg \min_{\beta \in \mathcal{B}} \left\{ \frac{1}{4} \|\hat{\Sigma} - \beta\beta^T\|_2^2 + \lambda \|\beta\|_1 \right\},$$

with $\lambda > 0$ a tuning parameter. The estimator is termed a *sparse PCA estimator*.

For the set \mathcal{B} we take an “ ℓ_2 -local” set:

$$\mathcal{B} := \{\tilde{\beta} \in \mathbb{R}^p : \|\tilde{\beta} - \beta^0\|_2 \leq \eta\}$$

with $\eta > 0$ a suitable constant. To get into such a local set, one may have to use another algorithm, with perhaps a slower rate than the one we obtain in Theorem 6.5.1 below. This caveat is as it should be, see Berthet and Rigollet [2013]: the fast rate of Theorem 6.5.1 cannot be achieved by any polynomial time algorithm unless e.g. one assumes a priori bounds. In an asymptotic setting, the constant η is not required to tend to zero. We will need 3η to be smaller than the gap between the square-root largest and square-root second largest eigenvalue of Σ_0 .

In the risk notation: the empirical risk is

$$R_n(\beta) := \|\hat{\Sigma} - \beta\beta^T\|_2^2 = -\frac{1}{2}\beta^T\hat{\Sigma}\beta + \frac{1}{4}\|\beta\|_2^4.$$

Here, it may be useful to note that for a symmetric matrix A

$$\|A\|_2^2 = \text{trace}(A^2).$$

Hence

$$\|\beta\beta^T\|_2^2 = \text{trace}(\beta\beta^T\beta\beta^T) = \|\beta\|_2^2\text{trace}(\beta\beta^T) = \|\beta\|_2^4.$$

The theoretical risk is

$$R(\beta) = -\frac{1}{2}\beta^T\Sigma_0\beta + \frac{1}{4}\|\beta\|_2^4.$$

6.5.1 Two-point margin and two point inequality for sparse PCA

By straightforward differentiation

$$\dot{R}(\beta) = -\Sigma_0\beta + \|\beta\|_2^2\beta.$$

The minimizer β^0 of $R(\beta)$ satisfies $\dot{R}(\beta^0) = 0$, i.e.,

$$\Sigma_0\beta^0 = \|\beta^0\|_2^2\beta^0.$$

Indeed, with $\beta^0 = \phi_{\max}q^0$

$$\begin{aligned} \Sigma_0\beta^0 &= \phi_{\max}\Sigma_0q^0 = \phi_{\max}^3q^0 \\ &= \|\phi_{\max}q^0\|_2^2\phi_{\max}q^0 = \|\beta^0\|_2^2\beta^0. \end{aligned}$$

We moreover have

$$\ddot{R}(\beta) = -\Sigma_0 + \|\beta\|_2^2I + 2\beta\beta^T,$$

with I denoting the $p \times p$ identity matrix.

Let now the spectral decomposition of Σ_0 be

$$\Sigma_0 := Q\Phi^2Q^T,$$

with $\Phi = \text{diag}(\phi_1 \cdots \phi_p)$, $\phi_1 \geq \cdots \geq \phi_p \geq 0$, and with $Q = (q_1, \dots, q_p)$, $QQ^T = Q^TQ = I$. Thus $\phi_{\max} = \phi_1$ and $q^0 = q_1$. We assume the following spikiness condition.

Condition 6.5.1 For some $\rho > 0$,

$$\phi_{\max} \geq \phi_j + \rho, \quad \forall j \neq 1.$$

Let, for $\tilde{\beta} \in \mathbb{R}^p$, $\Lambda_{\min}(\ddot{R}(\tilde{\beta}))$ be the smallest eigenvalue of the matrix $\ddot{R}(\tilde{\beta})$.

Lemma 6.5.1 Assume Condition 6.5.1 and suppose that $3\eta < \rho$. Then for all $\tilde{\beta} \in \mathbb{R}^p$ satisfying $\|\tilde{\beta} - \beta^0\|_2 \leq \eta$ we have

$$\Lambda_{\min}(\ddot{R}(\tilde{\beta})) \geq 2(\rho - 3\eta).$$

Proof of Lemma 6.5.1 . Let $\tilde{\beta} \in \mathbb{R}^p$ satisfy $\|\tilde{\beta} - \beta^0\|_2 \leq \eta$. The second derivative matrix at $\tilde{\beta}$ is

$$\begin{aligned} \ddot{R}(\tilde{\beta}) &= -\Sigma_0 + \|\tilde{\beta}\|_2^2 I + 2\tilde{\beta}\tilde{\beta}^T \\ &= \|\tilde{\beta}\|_2^2 \sum_{j=1}^p q_j q_j^T - \sum_{j=1}^p \phi_j^2 q_j q_j^T + 2\tilde{\beta}\tilde{\beta}^T \\ &= (\|\tilde{\beta}\|_2^2 - \phi_{\max}^2) q_1 q_1^T + \sum_{j=2}^p (\|\tilde{\beta}\|_2^2 - \phi_j^2) q_j q_j^T + 2\tilde{\beta}\tilde{\beta}^T. \end{aligned}$$

Since by assumption $\|\tilde{\beta} - \beta^0\|_2 \leq \eta$, it holds that

$$\|\tilde{\beta}\|_2 \geq \|\beta^0\|_2 - \eta = \phi_{\max} - \eta.$$

It follows that

$$\|\tilde{\beta}\|_2^2 \geq \phi_{\max}^2 - 2\eta\phi_{\max}$$

and hence for all $j \geq 2$

$$\|\tilde{\beta}\|_2^2 - \phi_j^2 \geq 2\rho\phi_{\max} - 2\eta\phi_{\max} = 2(\rho - \eta)\phi_{\max}.$$

Moreover, for all $x \in \mathbb{R}^p$

$$\begin{aligned} (x^T \tilde{\beta})^2 &= (x^T(\tilde{\beta} - \beta^0) + x^T \beta^0)^2 \\ &= (x^T(\tilde{\beta} - \beta^0))^2 + 2(x^T \beta^0)(x^T(\tilde{\beta} - \beta^0)) + (x^T \beta^0)^2 \\ &\geq (x^T \beta^0)^2 - 2\phi_{\max}\eta\|x\|_2^2 \end{aligned}$$

and

$$x^T(\|\tilde{\beta}\|_2^2 - \phi_{\max}^2) q_1 q_1^T x \geq -2\eta\phi_{\max}\|x\|_2^2.$$

We thus see that

$$\begin{aligned}
x^T \ddot{R}(\tilde{\beta})x &\geq 2(x^T \beta^0)^2 - 4\eta\phi_{\max}\|x\|_2^2 + 2(\rho - \eta)\phi_{\max} \sum_{j=2}^p (x^T q_j)^2 \\
&\geq 2(\rho - \eta)\phi_{\max} \sum_{j=1}^p (x^T q_j)^2 - 4\eta\phi_{\max}\|x\|_2^2 \\
&= 2(\rho - 3\eta)\phi_{\max}\|x\|_2^2.
\end{aligned}$$

□

By a two term Taylor expansion we have

$$R(\beta) - R(\beta') = \dot{R}(\beta')^T(\beta - \beta') + \frac{1}{2}(\beta - \beta')^T \ddot{R}(\tilde{\beta})(\beta - \beta')$$

with $\tilde{\beta}$ an intermediate point. Hence the two point margin condition holds with $G(u) = 2(\rho - 3\eta)\phi_{\max}u^2$, $u > 0$, $\tau = \|\cdot\|_2$, and $\mathcal{B}_{\text{local}} = \mathcal{B} = \{\beta' \in \mathbb{R}^p : \|\beta' - \beta^0\|_2 \leq \eta\}$.

6.5.2 Effective sparsity and dual-norm inequality for sparse PCA

We have seen in Subsection 6.5.1 that the (two-point) margin condition holds with norm $\tau = \|\cdot\|_2$. Clearly for all S

$$\|\tilde{\beta}_S\|_1 \leq \sqrt{s}\|\tilde{\beta}\|_2.$$

The effective sparsity depends only on β via its active set $S := S_\beta$ and does not depend on L :

$$\Gamma_{\|\cdot\|_1}^2(L, \beta, \|\cdot\|_2) = |S|.$$

The empirical process is

$$[R_n(\beta') - R(\beta')] - [R_n(\beta) - R(\beta)] = \frac{1}{2}\beta'^T W \beta' - \frac{1}{2}\beta^T W \beta,$$

where $W := \hat{\Sigma} - \Sigma_0$. Thus

$$\begin{aligned}
\left| [R_n(\beta') - R(\beta')] - [R_n(\beta) - R(\beta)] \right| &\leq 2 \left| \beta'^T W (\beta' - \beta) \right| + (\beta' - \beta)^T W (\beta' - \beta) \\
&\leq 2\|\beta' - \beta\|_1 \|W\beta\|_\infty + \|\beta' - \beta\|_1^2 \|W\|_\infty.
\end{aligned}$$

6.5.3 A sharp oracle inequality for sparse PCA

Theorem 6.5.1 (Sketch) *Suppose the spikiness condition (Condition 6.5.1). Let $\mathcal{B} := \{\tilde{\beta} \in \mathbb{R}^p : \|\tilde{\beta} - \beta^0\|_2 \leq \eta\}$ where $3\eta \leq \rho$. Fix some $\beta \in \mathcal{B}$. Let for $W = \hat{\Sigma} - \Sigma_0$*

$$\lambda_\epsilon \geq 2\|W\beta\|_\infty + \|W\|_\infty.$$

Let $\lambda \geq 8\lambda_\epsilon/\delta$. Define $\underline{\lambda} := \lambda - \lambda_\epsilon$ and $\bar{\lambda} := \lambda + \lambda_\epsilon + \delta\lambda$. Furthermore, define for $S \subset \{1, \dots, p\}$

$$\delta\lambda M_{\beta,S} := \frac{\lambda^2(1+\delta)^2|S|}{2(\rho-3\eta)\phi_{\max}} + 8(R(\beta) - R(\beta^0)) + 16\lambda\|\beta_{-S}\|_1.$$

Assume that $M_\beta \leq 1$. Then - under some additional assumptions (bounded data) -

$$\delta\lambda\|\hat{\beta} - \beta\|_1 + R(\hat{\beta}) \leq R(\beta) + \bar{\lambda}^2|S|/8 + 2\lambda\|\beta_{-S}\|_1.$$

Note that we did not provide a high probability bound for $2\|W\beta\|_\infty + \|W\|_\infty$. This can be done assuming for example a bound for $\|X_1^T\|_\infty$. The variable $X_1\beta$, $\beta \in \mathcal{B}$, has a bounded second moment: $\mathbb{E}(X_1\beta)^2 \leq \phi_{\max}^2(\phi_{\max} + \eta)^2$. One can then apply Dümbgen et al. [2010]. One then establishes the following asymptotics.

Asymptotics For simplicity we take $\beta = \beta^0$ and $S = S_0$. Suppose $p \log p/n = o(1)$, $\|X_1\|_\infty = \mathcal{O}(1)$, $\Lambda_{\max} = \mathcal{O}(1)$ and $1/(\rho - 3\eta) = \mathcal{O}(1)$. Then one may take $\lambda \asymp \sqrt{\log p/n}$. Assuming $s_0\sqrt{\log p/n}$ is sufficiently small (to ensure $M_{\beta_0, S_0} \leq 1$) one obtains $\|\hat{\beta} - \beta^0\|_2^2 = \mathcal{O}_{\mathbf{P}}(s_0 \log p/n)$ and $\|\hat{\beta} - \beta^0\|_1 = \mathcal{O}_{\mathbf{P}}(s_0\sqrt{\log p/n})$.

Bibliography

- F. Bach. Structured sparsity-inducing norms through submodular functions. In *Advances in Neural Information Processing Systems (NIPS)*, volume 23, pages 118–126, 2010.
- F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Optimization with sparsity-inducing penalties. In *Foundations and Trends in Machine Learning*, volume 4, pages 1–106, 2012.
- A. Belloni, V. Chernozhukov, and L. Wang. Square-root Lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.
- Q. Berthet and P. Rigollet. Complexity theoretic lower bounds for sparse principal component detection. In *Conference on Learning Theory*, pages 1046–1066, 2013.
- L. Birgé and P. Massart. Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4(3):329–375, 1998.
- M. Bogdan, E. van den Berg, W. Su, and E. Candes. Statistical estimation and testing via the sorted l_1 norm, 2013. arXiv:1310.1969.
- P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, 2011.
- F. Bunea, J. Lederer, and Y. She. The group square-root lasso: Theoretical properties and fast algorithms. *IEEE Transactions on Information Theory*, 2:1313–1325, 2014.
- G. Chen and M. Teboulle. Convergence analysis of a proximal-like minimization algorithm using Bregman functions. *SIAM Journal on Optimization*, 3(3):538–543, 1993.
- S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1998.
- L. Dümbgen, S.A. van de Geer, M.C. Veraar, and J.A. Wellner. Nemirovski’s inequalities revisited. *The American Mathematical Monthly*, 117:138–160, 2010.

- O. Güler. On the convergence of the proximal point algorithm for convex minimization. *SIAM Journal on Control and Optimization*, 29(2):403–419, 1991.
- R. Jenatton, J.-Y. Audibert, and F. Bach. Structured variable selection with sparsity inducing norms. *Journal of Machine Learning Research*, 12:2777–2824, 2011.
- V. Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: Ecole d'Été de Probabilités de Saint-Flour XXXVIII-2008*, volume 38. Springer Science & Business Media, 2011.
- V. Koltchinskii, K. Lounici, and A.B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Annals of Statistics*, 39(5):2302–2329, 2011.
- B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000.
- K. Lounici, M. Pontil, S. van de Geer, and A.B. Tsybakov. Oracle inequalities and optimal inference under group sparsity. *Annals of Statistics*, 39:2164–2204, 2011.
- A. Maurer and M. Pontil. Structured sparsity and generalization. *Journal of Machine Learning Research*, 13:671–690, 2012.
- C.A. Micchelli, J.M. Morales, and M. Pontil. A family of penalty functions for structured sparsity. In *Advances in Neural Information Processing Systems, NIPS 2010*, volume 23, pages 1612–1623, 2010.
- N. Städler, P. Bühlmann, and S. van de Geer. ℓ_1 -penalization for mixture regression models. *Test*, 19(2):209–256, 2010.
- N. Städler, P. Bühlmann, and S. van de Geer. Rejoinder ℓ_1 -penalization in mixture regression models. *Test*, 19(2):280–285, 2010.
- B. Stucky and S. van de Geer. Sharp oracle inequalities for square root regularization, 2015. arXiv:1509.04093.
- T. Sun and C.-H. Zhang. Comments on: ℓ_1 -penalization in mixture regression models. *Test*, 19(2):270–275, 2010.
- T. Sun and C.-H. Zhang. Scaled sparse linear regression. *Biometrika*, 99:879–898, 2012.
- T. Sun and C.-H. Zhang. Sparse matrix inversion with scaled lasso. *The Journal of Machine Learning Research*, 14(1):3385–3418, 2013.
- R. Tibshirani. Regression analysis and selection via the Lasso. *Journal of the Royal Statistical Society Series B*, 58:267–288, 1996.
- S. van de Geer. Weakly decomposable regularization penalties and structured sparsity. *Scandinavian Journal of Statistics*, 41(1):72–86, 2014.

- S.A. van de Geer. Least squares estimation with complexity penalties. *Mathematical Methods of Statistics*, 10:355–374, 2001.
- S.A. van de Geer. The deterministic Lasso. In *JSM proceedings, 2007, 140*. American Statistical Association, 2007.
- G.A. Watson. Characterization of the subdifferential of some matrix norms. *Linear Algebra and its Applications*, 170:33–45, 1992.
- M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B*, 68:49, 2006.
- X. Zeng and A.T.F. Mario. The ordered weighted l1 norm: Atomic formulation, dual norm, and projections, 2014. arXiv:1409.4271.