Abstract

Least squares with $\ell_1$-penalty, also known as the Lasso [23], refers to the minimization problem

$$\hat{\beta} := \arg \min_{\beta \in \mathbb{R}^p} \left\{ \| Y - X\beta \|_2^2 / n + \lambda \| \beta \|_1 \right\},$$

where $Y \in \mathbb{R}^n$ is a given $n$-vector, and $X$ is a given $(n \times p)$-matrix. Moreover, $\lambda > 0$ is a tuning parameter, larger values inducing more regularization. Of special interest is the high-dimensional case, which is the case where $p \gg n$. The Lasso is a very useful tool for obtaining good predictions $X\hat{\beta}$ of the regression function, i.e., of mean $f^0 := \mathbb{E}Y$ of $Y$ when $X$ is given. In literature, this is formalized in terms of an oracle inequality, which says that the Lasso predicts almost as well as the $\ell_0$-penalized approximation of $f^0$. We will discuss the conditions for such a result, and extend it to general loss functions. For the selection of variables however, the Lasso needs very strong conditions on the Gram matrix $X^TX/n$. These can be avoided by applying a two-stage procedure. We will show this for the adaptive Lasso. Finally, we discuss a modification that takes into account a group structure in the variables, where both the number of groups as well as the group sizes are large.

Mathematics Subject Classification (2000). Primary 62G05; Secondary 62J07.

Keywords. high-dimensional model, $\ell_1$-penalty, oracle inequality, restricted eigenvalue, sparsity, variable selection

1. Introduction

Estimation with $\ell_1$-penalty, also known as the Lasso [23], is a popular tool for prediction, estimation and variable selection in high-dimensional regression.
problems. It is frequently used in the linear model
\[ Y = X\beta + \epsilon, \]
where \( Y \) is an \( n \)-vector of observations, \( X = (X_1, \ldots, X_p) \) is the \((n \times p)\)-design matrix and \( \epsilon \) is a noise vector. For the case of least squares error loss, the Lasso is then
\[ \hat{\beta} := \arg\min_{\beta \in \mathbb{R}^p} \left\{ \|Y - X\beta\|^2_2/n + \lambda \|\beta\|_1 \right\}, \tag{1} \]
where \( \lambda > 0 \) is a tuning parameter.

A vector \( \beta \) is called sparse if it has only a few non-zero entries. Oracle inequalities are results of the form: with high probability
\[ \|X(\hat{\beta} - \beta_0)\|^2_2/n \leq \text{constant} \times \lambda^2 s_0, \tag{2} \]
where \( \beta_0 \) is the unknown true regression coefficient, or some sparse approximation thereof, and \( s_0 \) is the sparsity index, i.e., the number of non-zero coefficients of \( \beta_0 \).

The terminology oracle inequality is based on the idea of mimicking an oracle that knows beforehand which coefficients \( \beta_0 \) are non-zero. Indeed, suppose that \( EY = X\beta_0 \), and that the noise \( \epsilon = Y - X\beta_0 \) has independent components with variance \( \sigma^2 \). Let \( S_0 := \{ j : \beta_{j,0} \neq 0 \} \), say \( S_0 = \{1, \ldots, s_0\} \) is the set of indices of the first \( s_0 \) variables. Let \( X(S_0) := (X_1, \ldots, X_{s_0}) \) be the design matrix containing these first \( s_0 \) variables, and let \( \beta_0(S_0) \) be the \( s_0 \) non-zero entries of \( \beta_0 \). Suppose that \( X(S_0) \) has full rank \( s_0 \) \((s_0 \leq n)\). If \( S_0 \) were known, we can apply the least squares least squares estimator based on the variables in \( S_0 \)
\[ \hat{\beta}(S_0) := \left( X^T(S_0)X(S_0) \right)^{-1} X^T(S_0)Y. \]
From standard least squares theory, we have
\[ E\|X(S_0)(\hat{\beta}(S_0) - \beta_0(S_0))\|^2_2 = \sigma^2 s_0. \]

Under general conditions, the prediction error of the Lasso behaves as if it knew \( S_0 \), e.g., for i.i.d. centered Gaussian errors with variance \( \sigma^2 \), the inequality (2) holds with large probability, with \( \lambda^2 \) up to a logarithmic factor \( \log p \), of order \( \sigma^2/n \).

In fact, what we will show in Section 2, is an oracle inequality of the form (2), where \( \beta_0 \) is not necessarily the “true” \( \beta \), but may be a sparse approximation of the truth. The “optimal” sparse approximation will be called the oracle. To make the distinction, we denote the truth (if there is any) as \( \beta_{\text{true}} \), and the oracle by \( \beta_{\text{oracle}} \). As we will see, \( \beta_{\text{oracle}} \) will be at least as sparse as \( \beta_{\text{truth}} \), and is possibly much sparser.

Apart from oracle inequalities, one may also consider estimation results, which are bounds on the \( \ell_q \) error \( \|\hat{\beta} - \beta_0\|_q \), for some \( 1 \leq q \leq \infty \). Variable selection refers to estimating the support \( S_0 \) of \( \beta_0 \).
From a numerical point of view, the Lasso is attractive as it is easy to compute and the $\ell_1$-penalty ensures that a number of the estimated coefficients $\hat{\beta}_j$ are exactly zero. Its active set $\hat{S} := \{ j : \hat{\beta}_j \neq 0 \}$ will generally contain less than $n$ variables, even when originally dealing with $p \gg n$ variables. In theory however, there is in general no guarantee that $\hat{S}$ coincides with $S_0$. Indeed, this would be too good to be true, because then we would have a very accurate procedure that in addition can correctly asses its own accuracy. This is somehow in contradiction with statistical uncertainty principles.

What is so special about the $\ell_1$-penalty? The theoretically ideal penalty (at least, in the linear model) for sparse situations is actually the $\ell_0$-penalty $\lambda \| \beta \|_0$, where $\| \beta \|_0 := \sum_{j=1}^p |\beta_j|^0 = \# \{ \beta_j \neq 0 \}$. But with this, the minimization problem is computationally intractable. The $\ell_1$-penalty has the advantage of being convex. Minimization with $\ell_1$-penalty can be done using e.g. interior point methods or path-following algorithms. Convexity is important from the computational point of view (as well as from the theoretical point of view as soon as we leave the linear model context). For theoretical analysis, it is important that the $\ell_1$-penalty satisfies the triangle inequality

$$\| \beta + \tilde{\beta} \|_1 \leq \| \beta \|_1 + \| \tilde{\beta} \|_1,$$

and is separable:

$$\| \beta \|_1 = \| \beta_S \|_1 + \| \beta_{S^c} \|_1,$$

for any $S \subset \{1, \ldots, p\}$. Here $\beta_S$ denotes the vector $\beta$ with the entries in $S^c$ set to zero, and $\beta_{S^c} = \beta - \beta_S$ has the entries in $S$ set to zero. Note for example that among the $\ell_q$-penalties $\lambda \| \beta \|_q^q$ (or $\lambda \| \beta \|_q$, $q \geq 1$), the $\ell_1$-penalty is the only one which unites these three properties.

There has been an explosion of papers on the topic. The theoretical properties - and limitations - of the standard Lasso are by now quite well understood. We mention some of the key papers. Consistency was obtained in [9]. Its prediction error and estimation error is derived in [12], [13] and [1], where also the so-called restricted eigenvalue conditions are introduced. The slightly weaker compatibility condition is given in [25]. In [8] an alternative to the Lasso is introduced, which is called the Dantzig selector. The papers [3], [4] and [5] also present oracle and estimation bounds, and treat incoherence assumptions.

Variable selection with the Lasso is studied in [21] and [32], [16] presents conditions for convergence sup-norm, and [31] for convergence in $\ell_q$, $1 \leq q \leq \infty$. Modifications of the Lasso procedure have also been developed, for example, the group Lasso [30], the fused Lasso [24], and the elastic net [34]. Moreover, two-stage procedures have been proposed and studied, such as the adaptive Lasso [33, 10], and the relaxed Lasso [20]. Extension to density estimation is in [6], and to generalized-linear models in [15] (for the case of orthonormal design) and [26].

The present paper puts some of our theoretical results in a single framework. This will reveal the common aspects of various versions of the Lasso (and some links with decoding). We will mainly refer to own work, but stress here that
this work in turn builds upon results and ideas from literature. In Section 2, we present an oracle inequality in the context of the linear model. This is extended to general convex loss in Section 3. Section 4 discusses the restricted eigenvalue condition and the related compatibility condition. We turn to estimation results and variable selection in Section 5. First, we give a bound for the $\ell_2$-error (Subsection 5.1). We then show in Subsection 5.2 that the Lasso needs strong conditions for correctly estimating the support set of the coefficients. We show in Subsection 5.3 that the adaptive Lasso has a limited number of false positive selections but may have less good prediction error than the Lasso. In Section 6, we consider an extension, where the variables are divided into groups, with within each group a certain ordering of the coefficients. We provide an oracle inequality involving sparsity in the number of groups. Section 7 concludes.

2. An Oracle Inequality in the Linear Model

In this section, we present a version of the oracle inequality, which is along the lines of results in [25].

Suppose that the observations $Y$ are of the form
$$Y = f^0 + \epsilon,$$
where $f^0$ is some unknown vector in $\mathbb{R}^n$, and $\epsilon$ is a noise vector. Let $X = \{X_1, \ldots, X_p\}$ be the design matrix. We assume that $X$ is normalized, i.e., that
$$\hat{\sigma}_{j,j} = 1, \ \forall \ j,$$
where $\{\hat{\sigma}_{j,j}\}$ are the diagonal elements of the Gram matrix
$$\hat{\Sigma} := X^T X/n := (\hat{\sigma}_{j,k}).$$
The empirical correlation between the noise $\epsilon$ and the $j$-th variable $X_j$ is controlled by introducing the set
$$T(\lambda) := \left\{ \max_{1 \leq j \leq p} 4|\epsilon^T X_j|/n \leq \lambda \right\}.$$
The tuning parameter $\lambda$ is to be chosen in such a way that the probability of $T(\lambda)$ is large.

For any index set $S \subset \{1, \ldots, p\}$, and any $\beta \in \mathbb{R}^p$, we let
$$\beta_{j,S} := \beta_j 1\{j \in S\}, \ j = 1, \ldots, p.$$ We sometimes identify $\beta_S \in \mathbb{R}^p$ with the vector in $\mathbb{R}^{|S|}$ containing only the entries in $S$.

We write the projection of $f^0$ on the space spanned by the variables $\{X_j\}_{j \in S}$ as
$$f_S := Xb^S := \arg \min_{f = X\beta} \|f - f^0\|^2_2.$$
When $p > n$, the Gram matrix $\hat{\Sigma}$ is obviously singular: it has at least $p - n$ eigenvalues equal to zero. We do however need some kind of compatibility of norms, namely the $\ell_1$-norm $\|\beta_S\|_1$ should be compatible with $\|X\beta\|_2$. Observe that $\|X\beta\|_2^2/n = \beta^T \hat{\Sigma} \beta$.

**Definition compatibility condition** Let $L > 0$ be a given constant and $S$ be an index set. We say that the $(L,S)$-compatibility condition holds if

$$\phi^2_{\text{comp}}(L,S) := \min \left\{ \frac{|S|\beta^T \hat{\Sigma} \beta}{\|\beta_S\|_1^2} : \|\beta_S\|_1 \leq L\|\beta_S\|_1 \right\} > 0.$$  

Section 4 will briefly discuss this condition.

**Theorem 2.1.** Let $\hat{f} = X\hat{\beta}$, where $\hat{\beta}$ is the Lasso estimator defined in (1). Then on $T(\lambda)$, and for all $S$, it holds that

$$\|\hat{f} - f^0\|_2^2/n + \lambda\|\hat{\beta} - b_S\|_1 \leq 7\|f_S - f^0\|_2^2/n + \frac{(7\lambda)^2|S|}{\phi^2_{\text{comp}}(6,S)}.$$  

The constants in the above theorem can be refined. We have chosen some explicit values for definiteness. Moreover, the idea is to apply the result to sets $S$ with $\phi_{\text{comp}}(6,S)$ not too small (say bounded from below by a constant not depending on $n$ or $p$, if possible).

Assuming that $f^0 := X\beta_{\text{true}}$ is linear, the above theorem tells us that

$$\|\hat{f} - f^0\|_2^2/n + \lambda\|\hat{\beta} - \beta_{\text{true}}\|_1 \leq \frac{(7\lambda)^2|S_{\text{true}}|}{\phi^2_{\text{comp}}(6,S_{\text{true}})},$$  

where $S_{\text{true}} := \{j : \beta_{j,\text{true}} \neq 0\}$. This is an inequality of the form (2), with $\beta_0$ taken to be $\beta_{\text{true}}$. We admit that the constant $\phi^2_{\text{comp}}(6,S_{\text{true}})$ is hiding in the unspecified “constant” of (2). The improvement which replaces $\beta_{\text{true}}$ by a sparse approximation is based on the oracle set

$$S_{\text{oracle}} := \arg \min_S \left\{ \frac{|S|\beta^T \hat{\Sigma} \beta}{\phi^2_{\text{comp}}(6,S)} + \frac{7\lambda^2|S|}{\phi^2_{\text{comp}}(6,S)} \right\},$$  

and the oracle predictor

$$f_{\text{oracle}} := f_{S_{\text{oracle}}} = X\beta_{\text{oracle}},$$

where

$$\beta_{\text{oracle}} := b_{S_{\text{oracle}}}.$$  

By the above theorem

$$\|\hat{f} - f^0\|_2^2/n + \lambda\|\hat{\beta} - \beta_{\text{oracle}}\|_1 \leq 7\|f_{\text{oracle}} - f^0\|_2^2/n + \frac{(7\lambda)^2|S_{\text{oracle}}|}{\phi^2_{\text{comp}}(6,S_{\text{oracle}})},$$
which is a - possibly substantial - improvement of (4). We think of this oracle as the $\ell_0$-penalized sparse approximation of the truth. Nevertheless, the constant $\phi_{\text{comp}}(6, S_{\text{oracle}})$ can still be quite small and spoil this interpretation.

We end this section with a simple bound for the probability of the set $T(\lambda)$ for the case of normally distributed errors. It is clear that appropriate probability inequalities can also be derived for other distributions. A good common practice is not to rely on distributional assumptions, and to choose the tuning parameter $\lambda$ using cross-validation.

**Lemma 2.1.** Suppose that $\epsilon$ is $N(0, \sigma^2 I)$-distributed. Then we have for all $x > 0$, and for

$$\lambda := 4\sigma \sqrt{\frac{2x + 2\log p}{n}},$$

$$\Pr\left( T(\lambda) \right) \geq 1 - 2\exp[-x].$$

### 3. An Oracle Inequality for General Convex Loss

As in [25, 26] one can extend the framework for squared error loss with fixed design to the following scenario. Consider data $\{Z_i\}_{i=1}^n \subset \mathcal{Z}$, where $\mathcal{Z}$ is some measurable space. We denote, for a function $g : \mathcal{Z} \to \mathbb{R}$, the empirical average by

$$P_n g := \frac{1}{n} \sum_{i=1}^n g(Z_i),$$

and the theoretical mean by

$$P g := \frac{1}{n} \sum_{i=1}^n \mathbb{E} g(Z_i).$$

Thus, $P_n$ is the “empirical” measure, that puts mass $1/n$ at each observation $Z_i$ ($i = 1, \ldots, n$), and $P$ is the “theoretical” measure.

Let $\mathbf{F}$ be a (rich) parameter space of real-valued functions on $\mathcal{Z}$, and, for each $f \in \mathbf{F}$, $\rho_f : \mathcal{Z} \to \mathbb{R}$ be a loss function. We assume that the map $f \mapsto \rho_f$ is convex. For example, in a density estimation problem, one can consider the loss

$$\rho_f(\cdot) := -f(\cdot) + \log \int e^f d\mu,$$

where $\mu$ is a given dominating measure. In a regression setup, one has (for $i = 1, \ldots, n$) response variables $Y_i \in \mathcal{Y} \subset \mathbb{R}$ and co-variables $X_i \in \mathcal{X}$ i.e., $Z_i = (X_i, Y_i)$. The parameter $f$ is a regression function. Examples are quadratic loss

$$\rho_f(\cdot, y) = (y - f(\cdot))^2,$$
\(\rho_f(\cdot, y) = -y f(\cdot) + \log(1 + \exp[f(\cdot)])\),

etc.

The empirical risk, and theoretical risk, at \(f\), is defined as \(P_n \rho_f\), and \(P \rho_f\), respectively. We furthermore define the target - or truth - as the minimizer of the theoretical risk

\[ f^0 := \arg\min_{f \in F} P \rho_f. \]

Consider a linear subspace

\[ \mathcal{F} := \left\{ f_\beta(\cdot) = \sum_{j=1}^{p} \beta_j \psi_j(\cdot) : \beta \in \mathbb{R}^p \right\} \subset F. \]

Here, \(\{\psi_j\}_{j=1}^{p}\) is a collection of functions on \(\mathcal{Z}\), often referred to as the dictionary. The Lasso is

\[ \hat{\beta} = \arg\min_{\beta} \left\{ P_n \rho_{f_\beta} + \lambda \|\beta\|_1 \right\}. \] (6)

We write \(\hat{f} = f_{\hat{\beta}}\).

For \(f \in \mathcal{F}\), the excess risk is

\[ \mathcal{E}(f) := P(\rho_f - \rho_{f^0}). \]

Note that by definition, \(\mathcal{E}(f) \geq 0\) for all \(f \in \mathcal{F}\).

Before presenting an oracle result of the same spirit as for the linear model, we need three definitions, and in additional some further notation. Let the parameter space \(\mathcal{F} := (F, \| \cdot \|)\) be a normed space. Recall the notation

\[ \beta_{j,S} := \beta_j \mathbb{1}\{j \in S\}, \ j = 1, \ldots, p. \]

Our first definition is as in the previous section, but now with a general norm \(\| \cdot \|\).

**Definition compatibility condition** We say that the \((L, S)\)-compatibility condition is met if

\[ \phi_{comp}^2(L, S) := \min \left\{ \frac{|S| \|f_\beta\|_1}{\|\beta_S\|_1} : \|\beta_S\|_1 \leq L \|\beta_S\|_1 \right\} > 0. \]

**Definition margin condition** Let \(\mathcal{F}_{local} \subset \mathcal{F}\) be some “local neighborhood” of \(f^0\). We say that the margin condition holds with strictly convex function \(G\), if for all \(f \in \mathcal{F}_{local}\), we have

\[ \mathcal{E}(f) \geq G(\|f - f^0\|). \]
**Definition convex conjugate** Let $G$ be a strictly convex function on $[0, \infty)$, with $G(0) = 0$. The convex conjugate $H$ of $G$ is defined as

$$H(v) = \sup_u \{uv - G(u)\}, \; v \geq 0.$$  

The best approximation of $f^0$ using only the variables in $S$ is

$$f^0_S := f_{b_S} := \arg \min_{f=I_{b_S}} \mathcal{E}(f).$$

The function $f^0_S$ plays here the same role as the projection $f_S$ of the previous section.

For $H$ being the convex conjugate of the function $G$ appearing in the margin condition, set

$$2\varepsilon(\lambda, S) = 3\mathcal{E}(f^0_S) + 2H\left(\frac{4\lambda\sqrt{|S|}}{\phi_{\text{comp}}(3, S)}\right). \tag{7}$$

For any $M > 0$, we let $Z_M(S)$ be given by

$$Z_M(S) := \sup_{\beta: \|\beta - b^S\|_1 \leq M} \left| (P_n - P)(\rho_{f_\beta} - \rho_{f_\beta}) \right|. \tag{8}$$

**Theorem 3.1.** Suppose that $S$ is an index set for which $f_\beta \in \mathcal{F}_{\text{local}}$ for all $\|\beta - b^S\|_1 \leq M(\lambda, S)$, where $M(\lambda, S) := \varepsilon(\lambda, S)/(16\lambda)$. Then on the set

$$T(\lambda, S) := \{Z_M(S) \leq \lambda M(\lambda, S)/8\},$$

we have

$$\mathcal{E}(\hat{f}) + \lambda \|\hat{\beta} - b^S\|_1 \leq 4\varepsilon(\lambda, S).$$

The typical case is the case of quadratic margin function $G$, say $G(u) = u^2/2$. Then also the convex conjugate $H(v) = v^2/2$ is quadratic. This shows that Theorem 3.1 is in fact an extension of Theorem 2.1, albeit that the constants do not carry over exactly (the latter due to human inconsistencies). We furthermore remark that - in contrast to the $\ell_0$-penalty - the $\ell_1$-penalty adapts to the margin behavior. In other words, having left the framework of a linear model, the $\ell_1$-penalty exhibits an important theoretical advantage.

One may object that by assuming one is on the set $T(\lambda, S)$, Theorem 3.1 neglects all difficulties coming from the random nature of our statistical problem. However, contraction and concentration inequalities actually make it possible to derive bounds for the probability of $T(\lambda, S)$ in a rather elegant way. Indeed, in the case of Lipschitz loss, one may invoke the contraction inequality of [14], which gives the following lemma.

**Lemma 3.1.** Suppose that $f \mapsto \rho_f$ is Lipschitz:

$$|\rho_f - \rho_{\hat{f}}| \leq |f - \hat{f}|.$$
Then one has
\[ \mathbb{E} Z_M(S) \leq 4\lambda_{\text{noise}} M, \]
where
\[ \lambda_{\text{noise}} := \mathbb{E} \left( \max_{1 \leq j \leq p} \left| \sum_{i=1}^{n} \varepsilon_i \psi_j(Z_i)/n \right| \right), \]
and where \( \varepsilon_1, \ldots, \varepsilon_n \) is a Rademacher sequence independent of \( Z_1, \ldots, Z_n \).

Concentration inequalities [17, 18, 2], which say that \( Z_M(S) \) is with large probability concentrated near its expectation, will then allow one to show that for appropriate \( \lambda \), the set \( T(\lambda, S) \) has large probability.

4. Compatibility and Restricted Eigenvalues

Let \( Q \) be a probability measure on \( Z \), and for \( \beta \in \mathbb{R}^p \), let \( f_\beta = \sum_{j=1}^{p} \beta_j \psi_j \), where \( \{ \psi_j \}_{j=1}^{p} \subset L^2(Q) \) is a given dictionary. Write the Gram matrix as
\[ \Sigma := \int \psi^T \psi dQ, \quad \psi := (\psi_1, \ldots, \psi_p). \]
Moreover, let \( \| \cdot \| \) be the \( L^2(Q) \)-norm induced by the inner product
\[ (f, \tilde{f}) := \int f \tilde{f} dQ. \]
Note thus that
\[ \| f_\beta \|^2 = \beta^T \Sigma \beta. \]

**Definition compatibility and restricted eigenvalue** Let \( L > 0 \) be a given constant and \( S \) be an index set. We say that the \((\Sigma, L, S)\)-compatibility condition holds if
\[ \phi_{\text{comp}}^2(\Sigma, L, S) := \min \left\{ \frac{\| f_\beta \|^2}{\| \beta \|_1^2} : \| \beta_{S^c} \|_1 \leq L \| \beta_S \|_1 \right\} \]
is strictly positive. We say that the \((\Sigma, L, S)\)-restricted eigenvalue condition holds if the restricted eigenvalue
\[ \phi_{\text{RE}}^2(\Sigma, L, S) := \min \left\{ \frac{\| f_\beta \|^2}{\| \beta_{S^c} \|_2^2} : \| \beta_{S^c} \|_1 \leq L \| \beta_S \|_1 \right\} \]
is strictly positive.

The compatibility condition was introduced in [25], and the restricted eigenvalue condition in [1]. It is clear that
\[ \phi_{\text{RE}}^2(\Sigma, L, S) \leq \phi_{\text{comp}}^2(\Sigma, L, S). \]
On the other hand, results involving the set $S_{\text{true}}$, for the $\ell_2$-error $\|\hat{\beta} - \beta_{\text{true}}\|_2$ of the Lasso rely on $\phi_{\text{RE}}(\Sigma, L, S_{\text{true}})$ rather than $\phi_{\text{comp}}(\Sigma, L, S_{\text{true}})$ (and improved results, involving the oracle set $S_{\text{oracle}}$, in fact depend on the so-called adaptive restricted eigenvalue $\phi_{\text{adap}}(\Sigma, L, S_{\text{oracle}})$, see Subsection 5.1).

It is easy to see that

$$\phi_{\text{RE}}^2(\Sigma, L, S) \leq \Lambda_{\min}^2(S),$$

where $\Lambda_{\min}^2(S)$ is the smallest eigenvalue of the Gram matrix corresponding to the variables in $S$, i.e.,

$$\Lambda_{\min}^2(S) := \min_{\beta} \frac{\|f_{\beta_S}\|_2^2}{\|\beta_S\|_2^2}.$$

Conversely, denoting the canonical correlation by

$$\theta(S) := \sup_{\beta} \frac{|(f_{\beta_S}, f_{\beta_{S_c}})|}{\|f_{\beta_S}\| \|f_{\beta_{S_c}}\|},$$

one has the following bound.

**Lemma 4.1.** Suppose that $\theta(S) < 1$. Then

$$\phi_{\text{RE}}^2(\Sigma, L, S) \geq (1 - \theta(S))^2 \Lambda_{\min}^2(S).$$

Lemma 4.1 does not exploit the fact that in the definition of the restricted eigenvalue, we restrict the coefficients $\beta$ to $\|\beta_{S_c}\|_1 \leq L \|\beta_S\|_1$. Using this restriction, the restricted eigenvalue condition can for instance be derived from the restricted isometry property introduced in [7]. The latter paper studies the exact recovery of the true coefficients $\beta_{\text{true}}$ of $f^0 := f_{\beta_{\text{true}}}$, using the linear program

$$\beta_{\text{LP}} := \arg \min \{\|\beta\|_1 : \|f_{\beta} - f^0\| = 0\}. \quad (9)$$

The restrictions on the coefficients also allows one to derive bounds for restricted eigenvalues based on those computed with respect to an approximating (potentially non-singular) matrix. For two symmetric $(p \times p)$-matrices $\Sigma_0$ and $\Sigma_1$, we define

$$\|\Sigma_0 - \Sigma_1\|_{\infty} := \max_{1 \leq j \leq k \leq p} |\Sigma_{0,j,k} - \Sigma_{1,j,k}|.$$

The following lemma is proved in [28].

**Lemma 4.2.** We have

$$\phi_{\text{comp}}(\Sigma_1, L, S) \geq \phi_{\text{comp}}(\Sigma_0, L, S) - (L + 1)\sqrt{\|\Sigma_0 - \Sigma_1\|_{\infty}|S|}.$$

Similarly,

$$\phi_{\text{RE}}(\Sigma_1, L, S) \geq \phi_{\text{RE}}(\Sigma_0, L, S) - (L + 1)\sqrt{\|\Sigma_0 - \Sigma_1\|_{\infty}|S|}.$$
5. Estimation and Variable Selection

We present results for the linear model only.

5.1. Estimation. Consider the model

$$Y = f^0 + \epsilon.$$ 

For estimation in $\ell_2$ of the coefficients, we introduce the adaptive restricted eigenvalue. For a given $S$, our adaptive restricted eigenvalue conditions are stronger than in [1], but the result we give is also stronger, as we consider $S_{\text{oracle}} \subset S_{\text{true}}$ instead of $S_{\text{true}}$.

**Definition adaptive restricted eigenvalue.** We say that the $(L, S)$-adaptive restricted eigenvalue condition holds if

$$\phi^2_{\text{adap}}(L, S) := \min \left\{ \|X\beta\|_2^2 : \|\beta_{S^c}\|_1 \leq L \sqrt{|S|}\|\beta_S\|_2 \right\} > 0.$$ 

Thus,

$$\phi^2_{\text{adap}}(L, S) \leq \phi^2_{\text{RE}}(L, S) \leq \phi^2_{\text{comp}}(L, S).$$

In addition, we consider supersets $N$ of $S$, with size $(1 + \text{constant}) \times |S|$. For definiteness, we take the constant to be equal to 1. The minimal adaptive restricted eigenvalue is

$$\phi_{\text{adap}}(L, S, 2|S|) := \min\{\phi_{\text{adap}}(L, N) : N \supset S, |N| = 2|S|\}.$$

**Lemma 5.1.** Let $\hat{\beta}$ be the Lasso given in (1). Let

$$T(\lambda) := \left\{ \max_{1 \leq j \leq p} 4|\epsilon^T X_j|/n \leq \lambda \right\}.$$ 

Then on $T(\lambda)$, and for $\beta_{\text{oracle}} := b^{S_{\text{oracle}}}$, and $f_{\text{oracle}} := f^{S_{\text{oracle}}}$, with $S_{\text{oracle}}$ given in (5), we have

$$\|\hat{\beta} - \beta_{\text{oracle}}\|_2 \leq \frac{10}{\lambda\sqrt{|S_{\text{oracle}}|}} \left\{ \|f_{\text{oracle}} - f^0\|_2^2/n + \frac{(7\lambda)^2|S_{\text{oracle}}|}{\phi^2_{\text{adap}}(6, S_{\text{oracle}}, 2|S_{\text{oracle}}|)} \right\}.$$ 

This lemma was obtained in [29].

5.2. Variable selection. We now show that the Lasso is not very good in variable selection, unless rather strong conditions on the Gram matrix are met. To simplify the exposition, we assume in this subsection that there is no noise. We let $\{\psi_j\}_{j=1}^p$ be a given dictionary in $L_2(Q)$, with Gram matrix $\Sigma := \int \psi^T \psi dQ := (\sigma_{j,k})$. Furthermore, for an index set $S$, we consider the
submatrices

\[ \Sigma_{1,1}(S) := (\sigma_{j,k})_{j \in S, k \in S}, \quad \Sigma_{2,2}(S) := (\sigma_{j,k})_{j \notin S, k \notin S}, \]

and

\[ \Sigma_{2,1}(S) = (\sigma_{j,k})_{j \notin S, k \in S}, \quad \Sigma_{1,2}(S) := (\sigma_{j,k})_{j \in S, k \notin S}. \]

We let, as before, \( \Lambda_{\min}^2(S) \) be the smallest eigenvalue of \( \Sigma_{1,1}(S) \).

The noiseless Lasso is

\[ \beta_{\text{Lasso}} := \arg \min_{\beta} \{ \| f_{\beta} - f^0 \|^2 + \lambda \| \beta \|_1 \}. \]

Here,

\[ f^0 = f_{\beta_{\text{true}}}, \]

is assumed to be linear, with a sparse vector of coefficients \( \beta_{\text{true}} \). Our aim is to estimate \( S_{\text{true}} := \{ j : \beta_{j,\text{true}} \neq 0 \} \) using the Lasso \( S_{\text{Lasso}} = \{ j : \beta_{j,\text{Lasso}} \neq 0 \} \).

The irrepresentable condition can be found in [32]. We use a slightly modified version.

**Definition**

**Part 1** We say that the irrepresentable condition is met for the set \( S \), if for all vectors \( \tau_S \in \mathbb{R}^{|S|} \) satisfying \( \| \tau_S \|_\infty \leq 1 \), we have

\[ \| \Sigma_{2,1}(S) \Sigma_{1,1}^{-1}(S) \tau_S \|_\infty < 1. \] (10)

**Part 2** Moreover, for a fixed \( \tau_S \in \mathbb{R}^{|S|} \) with \( \| \tau_S \|_\infty \leq 1 \), the weak irrepresentable condition holds for \( \tau_S \), if

\[ \| \Sigma_{2,1}(S) \Sigma_{1,1}^{-1}(S) \tau_S \|_\infty \leq 1. \]

**Part 3** Finally, for some \( 0 < \theta < 1 \), the \( \theta \)-uniform irrepresentable condition is met for the set \( S \), if

\[ \max_{\| \tau_S \|_\infty \leq 1} \| \Sigma_{2,1}(S) \Sigma_{1,1}^{-1}(S) \tau_S \|_\infty \leq \theta. \]

The next theorem summarizes some results of [28].

**Theorem 5.1.**

**Part 1** Suppose the irrepresentable condition is met for \( S_{\text{true}} \). Then \( S_{\text{Lasso}} \subseteq S_{\text{true}} \).

**Part 2** Conversely, suppose that \( S_{\text{Lasso}} \subseteq S_{\text{true}} \), and that

\[ |\beta_{j,\text{true}}| > \lambda \sup_{\| \tau_{S_{\text{true}}} \|_\infty \leq 1} \| \Sigma_{1,1}^{-1}(S_{\text{true}}) \tau_{S_{\text{true}}} \|_\infty / 2. \]

Then the weak irrepresentable condition holds for the sign-vector

\[ \tau_{\text{true}} := \text{sign}( (\beta_{\text{true}})_{S_{\text{true}}}). \]
Part 3 Suppose that for some $\theta < 1/L$, the $\theta$-uniform irrepresentable condition is met for $S$. Then the compatibility condition holds with $\phi^2(\Sigma, L, S) \geq (1 - L\theta)^2 \Lambda_{\min}^2(S)$.

One may also verify that the irrepresentable condition implies exact recovery:

$$\beta_{LP} = \beta_{\text{true}},$$

where $\beta_{LP}$ is given in (9).

5.3. The adaptive Lasso. The adaptive Lasso introduced by [33] is

$$\hat{\beta}_{\text{adap}} := \arg\min_{\beta} \left\{ \|Y - X\beta\|_2^2/n + \lambda\init \lambda_{\text{adap}} \sum_{j=1}^p |\hat{\beta}_j| \right\},$$

(11)

Here, $\hat{\beta}_{\init}$ is the one-stage Lasso defined in (1), with initial tuning parameter $\lambda := \lambda\init$, and $\lambda_{\text{adap}} > 0$ is the tuning parameter for the second stage. Note that when $|\hat{\beta}_{\init,j}| = 0$, we exclude variable $j$ in the second stage.

We write $\hat{f}_{\init} := X\hat{\beta}_{\init}$ and $\hat{f}_{\text{adap}} := X\hat{\beta}_{\text{adap}}$, with active sets $\hat{S}_{\init} := \{j : \hat{\beta}_{\init,j} \neq 0\}$ and $\hat{S}_{\text{adap}} := \{j : \hat{\beta}_{\text{adap},j} \neq 0\}$, respectively.

Let

$$\hat{\delta}_{\init}^2 := \|X\hat{\beta}_{\init} - f^0\|_2^2/n,$$

be the prediction error of the initial Lasso, and and, for $q > 1$,

$$\hat{\delta}_q := \|\hat{\beta}_{\init} - \beta_{\text{oracle}}\|_q$$

be its $\ell_q$-error. Denote the prediction error of the adaptive Lasso as

$$\hat{\delta}_{\text{adap}}^2 := \|X\hat{\beta}_{\text{adap}} - f^0\|_2^2/n.$$

The next theorem was obtained in [29]. The first two parts actually repeat the statements of Theorem 2.1 and Lemma 5.1, albeit that we everywhere invoke the smaller minimal adaptive restricted eigenvalue $\phi_{\text{adap}}(6, S_{\text{oracle}}, 2|S_{\text{oracle}}|)$ instead of $\phi_{\text{comp}}(6, S_{\text{oracle}})$, which is not necessary for the bounds on $\hat{\delta}_{\init}^2$ and $\hat{\delta}_1$. This is only to simplify the exposition.

**Theorem 5.2.** Consider the oracle set $S_0 := S_{\text{oracle}}$ given in (5), with cardinality $s_0 := |S_{\text{oracle}}|$. Let $\phi_0 := \phi_{\text{adap}}(6, S_0, 2s_0)$. Let

$$T(\lambda\init) := \left\{ \max_{1 \leq j \leq p} 4|\epsilon^TX_j|/n \leq \lambda\init \right\}.$$

Then on $T(\lambda\init)$, the following statements hold.

1) There exists a bound $\delta_{\init}^{\text{upper}} = O(\lambda\init \sqrt{s_0}/\phi_0)$ such that

$$\hat{\delta}_{\init} \leq \delta_{\init}^{\text{upper}}.$$
2) For $q \in \{1, 2, \infty\}$, there exists bounds $\delta_q^{\text{upper}}$ satisfying
\[
\delta_1^{\text{upper}} = O(\lambda_{\text{init}} s_0/\phi_0^2), \quad \delta_2^{\text{upper}} = O(\lambda_{\text{init}} \sqrt{s_0}/\phi_0^2), \quad \delta_\infty^{\text{upper}} = O(\lambda_{\text{init}} \sqrt{s_0}/\phi_0^2),
\]
such that
\[
\hat{\delta}_q \leq \delta_q^{\text{upper}}, \quad q \in \{1, 2, \infty\}.
\]

3) Let $\delta_2^{\text{upper}}$ and $\delta_\infty^{\text{upper}}$ be such bounds, satisfying $\delta_\infty^{\text{upper}} \geq \delta_2^{\text{upper}}/\sqrt{s_0}$, and $\delta_2^{\text{upper}} = O(\lambda_{\text{init}} \sqrt{s_0}/\phi_0^2)$. Let $|\beta_{\text{oracle}}|_{\text{harm}}^2$ be the trimmed harmonic mean
\[
|\beta_{\text{oracle}}|_{\text{harm}}^2 := \left( \sum_{|\beta_j, \text{oracle}| > 2\delta_\infty^{\text{upper}}} \frac{1}{|\beta_j, \text{oracle}|^2} \right)^{-1}.
\]

Suppose that
\[
\lambda_{\text{adap}}^2 \asymp \left( \frac{1}{n} \left\| f_{S_0}^{\text{thres}} - f^0 \right\|_2^2 + \frac{\lambda_{\text{init}}^2 s_0}{\phi_0^2} \right) |\beta_{\text{oracle}}|_{\text{harm}}^2 \lambda_{\text{init}}^2/\phi_0^2,
\]
where $S_0^{\text{thres}} := \{ j : |\beta_j, \text{oracle}| > 4\delta_\infty^{\text{upper}} \}$. Then
\[
\hat{\delta}_{\text{adap}}^2 = O\left( \frac{1}{n} \left\| f_{S_0}^{\text{thres}} - f^0 \right\|_2^2 + \frac{\lambda_{\text{init}}^2 s_0}{\phi_0^2} \right),
\]
and
\[
|\hat{S}_{\text{adap}} \setminus S_0| = O\left( \frac{\lambda_{\text{init}}^2 s_0}{\phi_0^2} |\beta_{\text{oracle}}|_{\text{harm}}^2 \right).
\]

The value (12) for the tuning parameter seems complicated, but it generally means we take it in such a way that the the adaptive Lasso has its prediction error optimized. The message of the theorem is that when using cross-validation for choosing the tuning parameters, the adaptive Lasso will - when the minimal adaptive restricted eigenvalues are under control - have $O(s_0)$ false positives, and possibly less, e.g., when the trimmed harmonic mean of the oracle coefficients is large. As far as we know, the cross-validated initial Lasso can have $O(s_0)$ false positives only when strong conditions on the Gram matrix $\hat{\Sigma}$ are met, for instance the condition that the maximal eigenvalue of $\hat{\Sigma}$ is $O(1)$ (and in that case the adaptive Lasso wins again by having $O(\sqrt{s_0})$ false positives). On the other hand, the prediction error of the adaptive Lasso is possibly less good than that of the initial Lasso.

6. The Lasso with Within Group Structure

Finally, we study a procedure for regression with group structure. The co-
variables are divided into $p$ given groups. The parameters within a group are
assumed to either all zero, or all non-zero.
We consider in the linear model
\[ Y = X\beta + \epsilon. \]
As before, \( \epsilon \) is a vector of noise, which, for definiteness, we assume to be \( \mathcal{N}(0, I) \)-distributed. Furthermore, \( X \) a now an \((n \times M)\)-matrix of co-variables. There are \( p \) groups of co-variables, each of size \( T \) (i.e., \( M = pT \)), where both \( p \) and \( T \) can be large. We rewrite the model as
\[ Y = \sum_{j=1}^{p} X_j \beta_j + \epsilon, \]
where \( X_j = \{X_{j,t}\}_{t=1}^{T} \) is an \((n \times T)\)-matrix and \( \beta_j = (\beta_{j,1}, \ldots, \beta_{j,T})^T \) is a vector in \( \mathbb{R}^T \). To simplify the exposition, we consider the case where \( T \leq n \) and where the Gram matrix within groups is normalized, i.e., \( X_j^T X_j/n = I \) for all \( j \). The number of groups \( p \) can be very large. The group Lasso was introduced by [30]. With large \( T \) (say \( T = n \)), the standard group Lasso will generally not have good prediction properties, even when \( p \) is small (say \( p = 1 \)). Therefore, one needs to impose a certain structure within groups. Such an approach has been considered by [19], [22], and [11].

We present results from [27], which are similar to those in [19]. We assume that for all \( j \), there is an ordering in the variables of group \( j \): the larger \( t \), the less important variable \( X_{j,t} \) is likely to be. Given positive weights \( \{w_t\}_{t=1}^{T} \) (which we for simplicity assume to be the same for all groups \( j \)), satisfying \( 0 < w_1 \leq \cdots \leq w_T \), we express the structure in group \( j \) as the weighted sum
\[ \|W \beta_j\|_2^2 := \sum_{t=1}^{T} w_t^2 \beta_{j,t}^2, \quad \beta_j \in \mathbb{R}^p. \]
The structured group Lasso estimator is defined as
\[
\hat{\beta}_{\text{SGL}} := \arg\min_{\beta \in \mathbb{R}^p} \left\{ \|Y - X\beta\|_2^2/n + \lambda \sum_{j=1}^{p} \|\beta_j\|_2 + \lambda \mu \sum_{j=1}^{p} \|W \beta_j\|_2 \right\},
\]
where \( \lambda \) and \( \mu \) are tuning parameters. The idea here is that the variables \( X_{j,t} \) with \( t \) large are thought of as being less important. For example \( X_{j,t} \) could to the \( t \)th resolution level in a Fourier expansion, or the \( t \)th order interaction term for categorical variables, etc.

Let
\[ R^2(t) := \sum_{s > t} \frac{1}{w_s^2}, \quad t = 1, \ldots, T. \]
Let \( T_0 \in \{1, \ldots, T\} \) be the smallest value such that
\[ T_0 \geq R(T_0)\sqrt{n}. \]
Take $T_0 = T$ if such a value does not exist. We call $T_0$ the hidden truncation level. The faster the $w_j$ increase, the smaller $T_0$ will be, and the more structure we have within groups. The choice of $T_0$ is in a sense inspired by a bias-variance trade-off.

We will throughout take the tuning parameters $\lambda$ and $\mu$ such that $\lambda \geq \sqrt{\frac{T_0}{n}}$ and $\lambda \mu \geq \frac{T_0}{n}$.

Let, for $x > 0$,

$$
\nu_0^2 := \nu_0^2(x) = (2x + 2 \log(pT)),
$$

and

$$
\xi_0^2 := \xi_0^2(x) = 1 + \sqrt{\frac{4x + 4 \log p}{T_0}} + \frac{4x + 4 \log p}{T_0}.
$$

Define the set

$$
T := \left\{ \max_{1 \leq j \leq p} \|V_j\|_\infty \leq \nu_0, \max_{1 \leq j \leq p} \frac{\xi_j^2}{T_0} \leq \xi_0^2 \right\}.
$$

Here, $V_j := \epsilon_j^T X_j / \sqrt{n}$, and $\xi_j^2 = \sum_{t=1}^{T_0} V_{j,t}^2$, $j = 1, \ldots, p$.

Define

$$
\hat{\Sigma} := X^T X / n,
$$

and write

$$
\|\beta\|^2_\Sigma := \beta^T \hat{\Sigma} \beta.
$$

When $M = pT$ is larger than $n$, it is clear that $\hat{\Sigma}$ is singular. To deal with this, we will (as in Lemma 4.2) approximate $\hat{\Sigma}$ by a matrix $\Sigma$, which potentially is non-singular. We let $\Sigma_j$ be the $(T \times T)$-submatrix of $\Sigma$ corresponding to the variables in the $j$th group (as $\hat{\Sigma}_j = I$, we typically take $\Sigma_j = I$ as well), and we write

$$
\|\beta\|^2_\Sigma := \beta^T \Sigma \beta, \|\beta_j\|^2_{\Sigma_j} := \beta_j^T \Sigma_j \beta_j, j = 1, \ldots, p.
$$

We invoke the notation

$$
\text{pen}_1(\beta) := \lambda \sum_j \|\beta_j\|_2, \text{pen}_2(\beta) := \lambda \mu \sum_j \|W_j\|_2,
$$

and

$$
\text{pen}(\beta) := \text{pen}_1(\beta) + \text{pen}_2(\beta).
$$

For an index set $S \subset \{1, \ldots, p\}$, we let

$$
\beta_j, S = \beta_j 1\{ j \in S \}, j = 1, \ldots, p
$$

(recall that $\beta_j$ is now a $T$-vector).
Definition The structured group Lasso \((\Sigma, L, S)\)-compatibility condition holds if
\[
\phi_{\text{struc}}^2(\Sigma, L, S) := \min \left\{ \frac{|S|\|\beta\|_2^2}{(\sum_{j \in S} \|\beta_j\|_2)^2} : \text{pen}_1(\beta_S) + \text{pen}_2(\beta) \leq L\text{pen}_1(\beta_S) \right\}
\]
is strictly positive.

Let
\[
\mathcal{S}(\Sigma) := \left\{ S : \frac{64n\lambda^2 \|\hat{\Sigma} - \Sigma\|_2 |S|}{\phi_{\text{struc}}^2(\Sigma, 3, S)} \leq \frac{1}{2} \right\}.
\]

By considering only sets \(S \in \mathcal{S}(\Sigma)\), we actually put a bound on the sparsity we allow, i.e., we cannot handle very non-sparse problems very well. Mathematically, it is allowed to take \(\Sigma = \hat{\Sigma}\), having \(\mathcal{S}(\hat{\Sigma})\) being every set \(S\) with strictly positive \(\phi_{\text{struc}}(\hat{\Sigma}, 3, S)\). The reason we generalize to approximating matrices \(\Sigma\) is that this helps to check the structured group Lasso \((\Sigma, L, S)\)-compatibility condition.

**Theorem 6.1.** Let
\[
Y = f^0 + \epsilon,
\]
where \(\epsilon\) is \(\mathcal{N}(0, I)\)-distributed. We have \(P(T) \geq 1 - 3 \exp[-x]\). Consider the structured group Lasso \(\hat{\beta}_{\text{SGL}}\) given in (13), and define \(\hat{f}_{\text{SGL}} := X\hat{\beta}_{\text{SGL}}\). Assume
\[
\lambda \geq 8\xi_0 \sqrt{T_0/n}, \quad \lambda \mu \geq 8\nu_0 T_0/n.
\]
On \(T\), we have for all \(S \in \mathcal{S}(\Sigma)\) and all \(\beta_S\),
\[
\|\hat{f}_{\text{SGL}} - f^0\|_2^2/n + \text{pen}(\hat{\beta}_{\text{struc}} - \beta_S) \leq 4\|f_{\beta_S} - f^0\|_2^2/n + \frac{(4\lambda)^2 |S|}{\phi_{\text{struc}}^2(\Sigma, 3, S)} + 8\text{pen}_2(\beta_S).
\]

In other words, the structured group Lasso mimics an oracle that selects groups of variables in a sparse way. Note that the tuning parameter \(\lambda\) is now generally of larger order than in the standard Lasso setup (1). This is the price to pay for having large groups. As an extreme case, one may consider the situation with weights \(w_t = 1\) for all \(t\). Then \(T_0 = T\), and the oracle bound is up to \(\log p\)-factors the same as the one obtained by the standard Lasso.

### 7. Conclusion

The Lasso is an effective method for obtaining oracle optimal prediction error or excess risk. For variable selection, the adaptive Lasso or other two-stage procedures can be applied, generally leading to less false positives at the price of reduced predictive power (or a larger number of false negatives). A priori structure in the variables can be dealt with by using a group Lasso, possibly with an additional within group penalty.

Future work concerns modifications that try to cope with large correlations between variables. Moreover, it will be of interest to go beyond generalized linear modeling.
References


