I. INTRODUCTION

The Lasso and related techniques based on convex relaxations became extremely successful because of:

- their computational efficiency
- strong theoretical results on their statistical accuracy

This mini-course will focus on this second aspect.

1. Sparsity in high dimensional modeling

A wide class of problems of interest in statistics/machine learning can be written as follows. We observe the labels $Y_1, \ldots, Y_n$ on a sample of $n$ individuals (randomly) drawn from a large population. Each of these individuals is non-exhaustively described by a vector of features (denoted by $x_i \in \mathbb{R}^p$ for the individual number $i$).

It is assumed then that for a parametric family $\{P_\eta\}$ of distributions we have - for some $\eta^* \in \mathbb{R}^p$ -

- for each $i$, $L(Y_i | x_i) = P_{x_i}^{\eta^*}$
- $(x_i, Y_i), i=1, \ldots, n$ are independent.
Main examples include:

- Gaussian linear model
  \[ P_\eta = \mathcal{N}( \eta, \sigma^2 ) \]
  where \( \sigma^2 > 0 \) is a nuisance parameter
  \[ Y_i = x_i^T \beta^* + \xi_i \quad \text{and} \quad \xi_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2) \]

- Logistic regression
  \[ P_\eta = \text{Bernoulli} \left( \frac{1}{1 + e^{\eta \mathbf{x}_i}} \right) \quad \eta \in \mathbb{R} \]

\[ \Pr(Y_i = 0 \mid \mathbf{x}_i) = \frac{1}{1 + \exp(\eta \mathbf{x}_i^T \beta^*)} \quad i = 1, \ldots, n \]

- Poisson regression
  \[ P_\eta = \text{Poisson}(e^{\eta \mathbf{x}_i}) \quad \eta \in \mathbb{R} \]

**REMARK** The aforementioned models are members of the family of generalized linear models (GLM). If the dimension \( p \) of the feature vector is small compared to \( n \), then these models are parametric and, therefore, are unable to describe complex phenomena.

The standard approach in such a situation is to embed the original feature vectors into a space of much larger dimension. More precisely, we choose a non-linear mapping \( \Phi : \mathbb{R}^p \rightarrow \mathbb{R}^\tilde{p} \), with \( \tilde{p} \) much larger than \( p \), and replace the feature vector \( \mathbf{x}_i \) by \( \tilde{\mathbf{x}}_i = \Phi(\mathbf{x}_i) \in \mathbb{R}^\tilde{p} \).

From now on, we assume that \( p \) is large, possibly much larger than \( n \). In such a high-dimensional situation, it is impossible to develop consistent statistical inference without further assumptions on \( \beta^* \).

**Exact Sparsity Assumption:** Only a few coordinates of \( \beta^* \) are non-zero.

\[ S = \# \{ i : \beta_i^* \neq 0 \} = \| \beta^* \|_0 \ll n. \]
APPROXIMATE SPARSITY: There is a sparse vector $\hat{\beta}$ such that the distribution $\bigotimes_{i=1}^{n} P_{\hat{x}_i \hat{\beta}}$ is close to the true distribution $\bigotimes_{i=1}^{n} P_{x_i \beta^*}$.

In a sparsity scenario with a high-dimensional parameter, three main problems are of interest:

- **variable selection**: given $\{(x_i, y_i)\}_{i=1}^{n}$, provide an estimator $\hat{\beta} \in \{1, \ldots, p\}$ of the sparsity pattern $\beta^* = \{j \in \{1, \ldots, p\} : \beta^*_j \neq 0\}$.
- **estimation**: construct $\hat{\beta}$ based on $\{(x_i, y_i)\}_{i=1}^{n}$ such that for a given norm $\| \cdot \|$, the error of estimation $\|\hat{\beta} - \beta^*\|$ is as small as possible.
- **prediction**: construct $\hat{\beta}$ such that $x_i^T \hat{\beta}$ is close to $x_i^T \beta^*$ for every $i \in \{1, \ldots, n\}$ or, more generally, the distribution $\bigotimes_{i=1}^{n} P_{x_i \hat{\beta}}$ is close to $\bigotimes_{i=1}^{n} P_{x_i \beta^*}$.

Remark: In general, prediction is the task requiring the less assumptions, whereas variable selection requires the strongest assumptions. The assumptions will be imposed on the design matrix:

$$X = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix} \in \mathbb{R}^{n \times p}$$

Remark: In the case of a Gaussian linear model, the distance between the distributions $\bigotimes_{i=1}^{n} P_{x_i \hat{\beta}}$ and $\bigotimes_{i=1}^{n} P_{x_i \beta^*}$ measured by the Kullback-Leibler divergence is given by:

$$KL(\bigotimes_{i=1}^{n} P_{x_i \hat{\beta}}, \bigotimes_{i=1}^{n} P_{x_i \beta^*}) = \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i \beta - x_i \beta^*)^2 = \frac{1}{2\sigma^2} \|X(\beta - \beta^*)\|^2_c.$$
II The Lasso and the Dantzig Selector

From now on, we deal with the Gaussian linear model.

We observe \((x_i, Y_i), i=1, \ldots, n\) such that

\[
Y_i = x_i^T \beta^* + \xi_i; \quad \xi_i \sim \mathcal{N}(0, \sigma^2).
\]

This can be written in a vector form as

\[
Y = X \beta^* + \xi; \quad \xi \sim \mathcal{N}(0, \sigma^2 I_n).
\]  

We assume from now on that \(X\) is deterministic.

**DEF.** For a tuning parameter \(\lambda > 0\), we call the Lasso a solution of the optimisation problem:

\[
\hat{\beta}_{\text{lasso}} \in \arg \min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2n} \| Y - X \beta \|^2_2 + \lambda \| \beta \|_1 \right\} \quad \text{(1).}
\]

**Remark.** The term Lasso has been proposed by Tibshirani in 1996. The \(l_1\)-penalty term can be seen as a convex relaxation of the more natural \(l_0\)-penalty.

**Proposition 1**

For any \(\lambda > 0\), any solution \(\hat{\beta}\) of (1) satisfies

\[
\frac{1}{n} \| X^T (Y - X \hat{\beta}_{\text{lasso}}) \|_{\infty} \leq \lambda.
\]

**Proof.** We set \(L(\beta) = \frac{1}{2n} \| Y - X \beta \|^2_2\) and \(P(\beta) = \lambda \| \beta \|_1\).

It is clear that \(L(\beta) + P(\beta)\) is convex, \(L(\beta)\) is \(C^\infty\) while \(P(\beta)\) is non-differentiable. According to the first-order conditions for a convex function, \(\hat{\beta}\) is a minimizer of \(L(\beta) + P(\beta)\) iff the zero vector belongs to
to the sub-differential of $L(\beta) + P(\beta)$.

$$0 \in \partial (L(\beta) + P(\beta)) \iff 0 \in \nabla L(\beta) + \partial P(\beta) \iff -\nabla L(\beta) \in \partial P(\beta) \iff \frac{1}{n} X^T (Y - X \beta) \in \partial P(\beta).$$

On the other hand, the sub-differential of the mapping $\beta \mapsto \|\beta\|_1$ is composed of all vectors $v \in \mathbb{R}^p$ such that

$$v_j = \begin{cases} 1 & \text{if } \beta_j > 0 \\ -1 & \text{if } \beta_j < 0 \\ [1, 1] & \text{if } \beta_j = 0 \end{cases}$$

Therefore, $v \in \partial \|\beta\|_1$ implies that $\|v\|_\infty \leq 1$ and hence

$$\frac{1}{n} X^T (Y - X \beta) \in \partial P(\beta) = \lambda \partial \|\beta\|_1 \Rightarrow \frac{1}{n} \|X^T (Y - X \beta)\|_\infty \leq \lambda.$$

DEF For a $\lambda > 0$, we call the Dantzig Selector (DS)

any solution of the optimisation problem

$$\hat{\beta}_D \in \arg \min_{\beta \in \mathbb{R}^p} \{ \|\beta\|_1 \}$$

where $B_\lambda = \{ \beta \in \mathbb{R}^p : \frac{1}{n} \|X^T (Y - X \beta)\|_\infty \leq \lambda \}.$

Remark DS has been proposed by Candès & Tao (2007).

(2) can be cast into a linear program whereas (1) can be cast into a second-order cone program (SOCP).

Remark If $X$ is orthogonal, i.e. if $\frac{1}{n} X^T X = I_p$ , then the Lasso and the DS are equal and coincide with the Soft-Thresholding estimator

$$\hat{\beta}_j^{\text{ST}} = \text{sign}(\tilde{Y}_j) (|\tilde{Y}_j| - \lambda)_+ ; j = 1, \ldots, p \quad \tilde{Y} = \frac{1}{n} X^T Y.$$
Proposition 2 Let model (D) be true and \( J^* = \{ j : \beta_j^* \neq 0 \} \).

Let \( \xi \in (0, 1) \) be a tolerance level, and define the event
\[
\mathcal{E} = \{ \| X^T \xi \|_\infty \leq \sigma \sqrt{2n \log (p/\xi)} \}. \]
If the columns of \( X \) are such that \( \frac{1}{n} \| X \|_2^2 = 1 \) then

i) \( \Pr(\mathcal{E}) \geq 1 - \xi \)

ii) if \( \lambda > \sigma \sqrt{\frac{2 \log (p/\xi)}{n}} \) then on \( \mathcal{E} \) we have \([CT \ '07]\)
\[
\| (\hat{\beta}^{DS} - \beta^*)_{J^*} \|_1 \leq \| (\hat{\beta}^{DS} - \beta^*)_{J^*} \|_1
\]

iii) if \( \lambda > 2 \sigma \sqrt{\frac{2 \log (p/\xi)}{n}} \), then on \( \mathcal{E} \) we have \([BRT '09]\)
\[
\| (\beta^* - \beta^*)_{J^*} \|_1 \leq 3 \| (\hat{\beta}^{DS} - \beta^*)_{J^*} \|_1
\]

**Remark** Roughly speaking, this result tells us that the accuracy of estimating the large vector \( \beta^* \) is of the same order as the accuracy of estimating the much smaller sub-vector \( \beta_{J^*}^* \).

**Proof** We set \( \xi_j = (X^T \xi, j=1,...,n) \), since \( X \) is deterministic and \( \xi \sim \mathcal{N}(0, \sigma^2 I_n) \), \( \xi_j \) is a zero-mean Gaussian random variable with variance \( \sigma^2 \| X \|_2^2 = \sigma^2 n \). Using Gaussian tail bound, we get
\[
\Pr(\| \xi \| > \sigma \sqrt{2n \log (p/\xi)}) \leq \exp \left( - \frac{\sigma^2 2n \log (p/\xi)}{2 \sigma^2 n} \right) = \frac{\xi}{p}
\]

In view of the union bound, this yields
\[
\Pr(\mathcal{E}^c) = \Pr(\exists j \text{ s.t. } \| (X^T \xi) \|_1 > \sigma \sqrt{2n \log (p/\xi)})
\]
\[
= 6
\]

**Journées de Statistique de la SFdS**
\[ \sum_{i=1}^{P} \mathbb{P} \left( |S_i| > \sigma \sqrt{2n \log(p/\delta)} \right) \leq \sum_{i=1}^{P} \delta/p = \delta. \]

To prove the second assertion we assume that \( E \) is realised. Then according to 1) we have

\[ \frac{1}{n} \|X^T(Y-X\hat{\beta})\|_\infty = \frac{1}{n} \|X^T\xi\|_\infty \leq \lambda. \]

which implies that \( \hat{\beta} \) is a feasible solution. Thus

\[ \|\hat{\beta}^{DS}\|_1 \leq \|\beta^*\|_1 \]

Using the fact that \( \beta^*_c \neq 0 \) and the triangle inequality, we get

\[ \| (\hat{\beta}^{DS} - \beta^*)_c \|_1 = \| \hat{\beta}^{DS} \|_1 - \| \beta^* \|_1 \leq \| \beta^*_c \|_1 - \| \hat{\beta}^{DS} \|_1 \]

Finally, to prove iii), we note that

\[ \frac{1}{2n} \|Y - X\hat{\beta}_c\|_2 + \lambda \|\hat{\beta}_c\|_1 \leq \frac{1}{2n} \|Y - X\beta^*_c\|_2 + \lambda \|\beta^*_c\|_1. \]

Replacing \( Y \) by \( X\beta^* + \xi \) we get

\[ \frac{1}{2n} \|X\beta^* - X\hat{\beta}_c\|_2 + \frac{1}{n} \xi^T X (\beta^* - \hat{\beta}_c) + \lambda \|\hat{\beta}_c\|_1 \leq \lambda \|\beta^*_c\|_1. \]

Since the first term is \( \geq 0 \) we get

\[ \lambda \|\hat{\beta}_c\|_1 \leq \lambda \|\beta^*_c\|_1 + \frac{1}{n} \|\xi^T X (\beta^* - \hat{\beta}_c)\|_1 \leq \lambda \|\beta^*_c\|_1 + \frac{1}{n} \|\xi^T X (\beta^* - \hat{\beta}_c)\|_1 \quad \text{(duality inequality)} \]

\[ \leq \lambda \|\beta^*_c\|_1 + \frac{1}{2} \|\beta^* - \hat{\beta}_c\|_1 \quad \text{on } E. \]

---

Journées de Statistique de la SFdS
Dividing everything by $\lambda$, we get
$$\| \hat{\beta}^L \|_1 \leq \| \beta^* \|_1 + \frac{1}{2} \| \hat{\beta}^L - \beta^* \|_1 \quad \text{on } \mathcal{E}.$$ 

Hence,
$$\| (\hat{\beta}^L - \beta^*)_{(j\star)c} \|_1 = \| \hat{\beta}^L_{(j\star)c} \|_1 = \| \hat{\beta}^L \|_1 - \| \hat{\beta}^L_{j\star} \|_1$$
\leq \| \beta^* \|_1 - \| \hat{\beta}^L_{j\star} \|_1 + \frac{1}{2} \| \hat{\beta}^L - \beta^* \|_1$$
\leq \| (\beta^* - \hat{\beta}^L)_{j\star} \|_1 + \frac{1}{2} \| \hat{\beta}^L - \beta^* \|_1$$
\leq \frac{3}{2} \| (\beta^* - \hat{\beta}^L)_{j\star} \|_1 + \frac{1}{2} \| (\beta^* - \hat{\beta}^L)_{(j\star)c} \|_1.$$ 

Rearranging the terms, we get the desired result.

\[\square\]

III Mutual coherence and variable selection

We assume that
$$Y = X\beta^* + \xi \quad \xi \sim \mathcal{N}_n \left(0, \sigma^2 I_n\right)$$

$Y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n \times p}$ are known. $\beta^* \in \mathbb{R}^p$ is unknown.

The dimension $p$ is large but $s = |J^*|$ is small.

The columns of $X$ are normalized: $\frac{1}{\nu_i} \| x_i \|_2^2 = 1$.

DEF. We call mutual coherence of $X$ the quantity
$$\mu = \max \frac{1}{n} \langle x_i, x_k \rangle = \max \frac{1}{n} (x_i)^T x_k.$$ 

Clearly, $\mu$ is always between $0$ and $1$ and equals $0$ iff $X$ is orthogonal. Thus, $\mu$ measures how close is $X$ to an orthogonal matrix.

THEOREM1 [Lunici 03] If $\mu \leq \frac{4}{4s}$, then with probability $\geq 1 - \delta$ the DS satisfies $\| \hat{\beta}^{DS} - \beta^* \|_\infty \leq 4 \lambda$.

---

Journées de Statistique de la SFdS
provided that $\lambda \geq \sigma \sqrt{\frac{2 \log(p/s)}{n}}$.

If, furthermore, $p \leq \frac{1}{8s}$ and $\lambda \geq 2\sigma \sqrt{\frac{2 \log(p/s)}{n}}$, then with probability $\geq 1 - 5$ we have

$$\|\hat{\beta}_{\text{lasso}} - \beta^*\|_{\infty} \leq 3\lambda.$$

**Consequence** If the mutual coherence is smaller than $\sqrt{1/8s}$ then both the DS and the Lasso converge to the true vector in the $l_{\infty}$-norm at the rate $(\frac{\log(p)}{n})^{1/2}$. This is very close to the parametric rate $(Vn)^{1/2}$!

**Proof.** We will only prove the inequality for the Lasso, the one for the DS is very similar. We have

$$\|\hat{\beta}^L - \beta^*\|_{\infty} \leq \|\frac{1}{n} X^T X (\hat{\beta}^L - \beta^*)\|_{\infty} + \|\frac{1}{n} X^T X - I_p\| (\hat{\beta}^L - \beta^*)\|_{\infty}.$$

On the one hand,

$$\|\frac{1}{n} X^T X - I_p\| (\hat{\beta}^L - \beta^*)\|_{\infty} = \max_{d} \left| \sum_{k+d} \frac{1}{n} <x_d^k, x^k> (\hat{\beta}^L - \beta^*)_d \right|$$

$$\leq \max_d \sum_{k+d} \frac{1}{n} |<x_d^k, x^k>| \cdot |(\hat{\beta}^L - \beta^*)_d| \leq p \cdot \|\hat{\beta}^L - \beta^*\|_d.$$

According to Prop 2, iii), we have $\|\hat{\beta}^L - \beta^*\|_1 \leq 4 \|\hat{\beta}^L - \beta^*\|_{\infty} \leq 4s \|\hat{\beta}^L - \beta^*\|_{\infty}$. Therefore,

$$\|\hat{\beta}^L - \beta^*\|_{\infty} \leq \|\frac{1}{n} X^T X (\hat{\beta}^L - \beta^*)\|_{\infty} + 4s p \|\hat{\beta}^L - \beta^*\|_{\infty}.$$

On the other hand, using Prop 1. and Prop 2, i) we get

$$\|\frac{1}{n} X^T X (\hat{\beta}^L - \beta^*)\|_{\infty} \leq \frac{1}{n} \|X^T (\hat{\beta}^L - \beta^*)\|_{\infty} + \frac{1}{n} \|X^T \|_{\infty} \leq \lambda + \frac{\lambda}{2}.$$

- 9 -
This leads to
\[ \| \hat{\beta}_L - \beta^* \|_\infty \leq \frac{3\lambda}{2} + 4sp \| \hat{\beta}_L - \beta^* \|_\infty, \]
or, equivalently,
\[ \| \hat{\beta}_L - \beta^* \|_\infty \leq \frac{3\lambda}{2(1 - 4sp)} \leq 3\lambda, \]
provided that \( p \leq 1/8s \).

This result can be used to derive the variable-selection-consistency of the thresholded Lasso estimator.

**Theorem 2.** Assume that \( p \leq 1/8s \) and \( |\beta_j^*| > 6\lambda \) for every \( j \in J^* \). Then, if \( \lambda > 2 \sqrt{2 \log(p/s)} \), the estimator
\[ \hat{\beta}_d^{TL} = \hat{\beta}_d^{L} \times 1( |\hat{\beta}_d^{L} | > 3\lambda ) \]
satisfies
\[ P( \text{sign}(\hat{\beta}_d^{TL}) = \text{sign}(\beta^*)) \geq 1 - \delta. \]

In particular, \( \hat{J} = \{ j : |\hat{\beta}_j^{L} | > 3\lambda \} \) satisfies
\[ P( \hat{J} = J^* ) \geq 1 - \delta. \]

**Proof** is trivial.

**Remark** Other variable-selection-consistency results for the Lasso & the DS available in the literature rely on the (strong) irrepresentability assumption. It is weaker than the mutual coherence assumption but requires in addition to assume that the solution to (1) is unique, while Theorems 1 and 2 hold for any solution of (1).
We still consider the model
\[ Y = X \beta^* + \xi, \quad \xi \sim \mathcal{N}_n(0, \sigma^2 I_n) \] (0)

Recall that the Lasso is any solution of the problem
\[ \hat{\beta} \in \arg \min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2n} \| Y - X \beta \|_2^2 + \lambda \| \beta \|_1 \right\}. \]

The goal of this and the next sections is to study the prediction loss of the Lasso:
\[ l_n(\hat{\beta}, \beta^*) = \frac{1}{n} \|X(\hat{\beta} - \beta^*)\|_2^2. \]

**Theorem 3** [Rigollet & Tsybakov '11]

If \( \lambda > \sigma \sqrt{2 \log(p)/n} \) and \( \frac{1}{n} \|x_i\|_2^2 = 1 \) for every \( i \), then with probability \( \geq 1 - \delta \) we have
\[ l_n(\hat{\beta}, \beta^*) \leq \inf_{\beta \in \mathbb{R}^p} \left\{ l_n(\beta, \beta^*) + 4\lambda \| \beta \|_1 \right\}. \] (3)

**Discussion**

1) Assuming that \( \beta^* \) is of a bounden \( l_2 \)-norm (with a bound independent on \( p \) and \( n \)) we get from (3) and Cauchy-Schwarz that
\[ l_n(\hat{\beta}, \beta^*) \leq 4\lambda \| \beta^* \|_1 \leq 4\lambda \sqrt{s} \times \| \beta^* \|_2 = C_s \sigma \sqrt{\log(p)/n}, \]
provided that \( \lambda \) is of the order of \( \sigma \sqrt{\log(p)/n} \).

This shows that under the exact sparsity condition, the prediction loss of the Lasso is of order \( \left( \frac{s \log(p)/n}{\lambda} \right)^{1/2} \).
which is called "slow rate". This rate is valid without requiring any assumption from the matrix $X$.

2) Inequality (3) is an oracle inequality which remains meaningful in the case of inexact sparsity. If $\beta \in \mathbb{R}^p$ is an $s$-sparse vector, then the prediction loss of $\hat{\beta}^L$ will be as small as the prediction loss of $\bar{\beta}$ up to a remainder term of order $\left( \frac{s \log (p/s)}{n} \right)^{1/2}$.

**Proof of Thm 3**

Recall that using the first-order conditions, we get

$$\frac{1}{n} X^T (Y - X \hat{\beta}^L) \in \lambda \text{sign} (\hat{\beta}^L),$$

Therefore, we have

$$\frac{1}{n} \hat{\beta}^L X^T (Y - X \hat{\beta}^L) = \lambda \| \hat{\beta}^L \|_1$$

$$\frac{1}{n} \hat{\beta}^L X^T (Y - X \hat{\beta}^L) \leq \lambda \| \beta \|_1 \quad \forall \beta \in \mathbb{R}^p$$

These inequalities yield

$$\frac{1}{n} (\beta - \hat{\beta}^L) X^T (Y - X \hat{\beta}^L) \leq \lambda \left( \| \beta \|_1 - \| \hat{\beta}^L \|_1 \right).$$

Replacing $Y$ by $X \beta^* + \xi$, we get

$$\frac{1}{n} \left[ X (\beta - \hat{\beta}^L) \right]^T X (\beta^* - \hat{\beta}^L) \leq \frac{1}{n} \left( \hat{\beta}^L - \beta \right)^T X^T \xi + \lambda \left( \| \beta \|_1 - \| \hat{\beta}^L \|_1 \right)$$

We use now the identity $u^T v = \frac{1}{2} \| u \|_2^2 + \frac{1}{2} \| v \|_2^2 - \frac{1}{2} \| u - v \|_2^2$ which yields

$$-12-$$

Journées de Statistique de la SFdS
\[ \ln(\hat{\beta}, \beta^*) \leq \ln(\beta, \beta^*) + \frac{2}{n} (\hat{\beta} - \beta)^T X^T \xi + 2 \lambda (||\beta||_1 - ||\hat{\beta}||_1) \]

According to Prop 2, i), we know that with prob > 1-\delta we have \( ||X^T \xi||_\infty \leq \ln \). Using the duality inequality, this implies that

\[ \ln(\hat{\beta}, \beta^*) \leq \ln(\beta, \beta^*) + \frac{2}{n} ||\hat{\beta} - \beta||_1 \times ||X^T \xi||_\infty + 2 \lambda (||\beta||_1 - ||\hat{\beta}||_1) \]

\[ \leq \ln(\beta, \beta^*) + 2 \lambda ||\hat{\beta} - \beta||_1 + 2 \lambda (||\beta||_1 - ||\hat{\beta}||_1) \]

\[ \leq \ln(\beta, \beta^*) + 4 \lambda ||\beta||_1 \quad \forall \beta \in \mathbb{R}^p. \]

This completes the proof.

\[ \frac{\nabla \quad \text{FAST RATES UNDER THE RE ASSUMPTION}}{\nabla} \]

When \( X \) is close to an orthogonal matrix the "slow rate" of Thm. 3 can be significantly improved.

**DEF** Let \( J \subset \{1, \ldots, p\} \) and \( c_0 > 0 \). We say that \( X \) fulfills the assumption \( \text{RE}(c_0, J) \) \([\text{RE} = \text{restricted eigenvalues}]\)

if for some constant \( \varepsilon > 0 \) we have

\[ \forall \epsilon \in \mathbb{R}^p \quad ||v||_1 < c_0 ||v||_1 \implies \frac{1}{n} ||Xv||^2_2 > \varepsilon ||v||^2_2. \]

**Proposition 3** If \( X \) satisfies the mutual coherence condition

\[ p < \frac{1}{2 \varepsilon^2 (1 + c_0)^2} \]

then it also satisfies \( \text{RE}(c_0, J) \) with any \( J \) of cardinality \( \leq s \) and with \( \varepsilon = \frac{1}{2} \).
THEOREM 4

Assume that $X$ satisfies $RE(3, J)$ for some $J \subset \{1, \ldots, p\}$.

Choose $\lambda = 2\sqrt{\frac{2\log(p/s)}{n}} \times \delta$ for some $\delta \in (0, 1)$. Then, with probability $\geq 1 - \delta$, we have

$$
\ln(\hat{\beta}, \beta^*) \leq \inf_{\beta: \beta^*_s = 0} \left\{ \ln(\beta, \beta^*) + \frac{9}{4\delta} \times X \cdot |J| \right\}
$$

(5)

Proof. Applying the duality inequality to (4), cf. page 13, we get

$$
\ln(\hat{\beta}^L, \beta^*) \leq \ln(\beta, \beta^*) + \frac{2}{n} \|X^T \xi\|_{\infty} \|\hat{\beta}^L - \beta\|_1 + 2\lambda (\|\hat{\beta}\|_1 - \|\hat{\beta}^L\|_1) - \frac{4}{n} \|X(\beta - \hat{\beta}^L)\|_1.
$$

- 14 -
To ease notation, we set \( v = \beta - \hat{\beta}^L \). Since \( \beta_{j,c} = 0 \), we have
\[
\| \beta \|_1 = \| \hat{\beta}^L \|_1 = \| \beta_{j}^L \|_1 + \| \beta_{j,c}^L \|_1 \leq \| v_j \|_1 - \| v_{j,c} \|_1.
\]
Therefore,
\[
\ln (\hat{\beta}^L, \beta^*) \leq \ln (\beta, \beta^*) + \frac{2}{n} \| X_i \|_\infty \| v \|_1 + 2 \lambda \left( \| v_j \|_1 - \| v_{j,c} \|_1 \right).
\]
In view of 1), Prop 2, with prob \( 1 - \delta \) we have \( \| X_i \|_\infty \leq \frac{\lambda n}{2} \).
Therefore,
\[
\ln (\hat{\beta}^L, \beta^*) \leq \ln (\beta, \beta^*) + \lambda \left( 3 \| v_j \|_1 - \| v_{j,c} \|_1 \right) - \frac{4}{n} \| v \|_2^2. \quad (6)
\]
First case: \( \| v_{j,c} \|_1 \geq 3 \| v_j \|_1 \)
In this case, (6) implies that \( \ln (\hat{\beta}^L, \beta^*) \leq \ln (\beta, \beta^*) \)
and the desired inequality follows.
Second case: \( \| v_{j,c} \|_1 < 3 \| v_j \|_1 \)
In view of the RE(3,J) assumption, we have
\[
\lambda \left( 3 \| v_j \|_1 - \| v_{j,c} \|_1 \right) \leq 3 \lambda \| v_j \|_1
\]
\[
\leq 3 \lambda \sqrt{\| v_j \|_2^2} \leq \frac{9}{4} \| v_j \|_2^2 + \| v \|_2 \| X \|_2 \| e \|
\leq \frac{9}{4} \| v \|_2 + \frac{1}{n} \| v \|_2^2.
\]
Combining this inequality with (6), we get the desired result. \( \square \)

**Comments**

1) The first result in the spirit of (5) has been proved in Bickel - Ritov - Tsybakov '09. The version with constant 1 in front of \( \ln (\beta, \beta^*) \) at the RHS of (5) appeared in - 15 -
Koltchinskii, Lunici & Tsybakov '11. The proof we presented here is inspired by the one in Sun & Zhang '12.

2) Choosing \( \lambda = 3\sigma \sqrt{\frac{\log(p/s)}{n}} \), we get that with prob. \( > 1 - \delta \)

\[
\ln(\hat{\beta}^*, \beta^*) \leq \inf_{\beta^*_j = 0} \left\{ \ln(\beta, \beta^*) + C \cdot \frac{\sigma^2 s \log(p/s)}{n} \right\}
\]

where \( C \) is a constant and \( s = |J| \).

The rate \( \frac{s \log(p/s)}{n} \) is called "fast rate".

3) It is known that the best possible rate over the set of all possible estimators is of order \( \frac{s \log(p/s)}{n} \) [Lunici 07].

So Thm 4 shows that the Lasso is rate optimal up to log-factors when \( X \) satisfies the RE condition.

4) Even for moderate values of \( p \), the computational complexity of checking that \( \text{RE}(c_0, J) \) holds for every \( J \) such that \(|J| \leq s \) is prohibitively large. An alternative condition has been proposed by Juditsky & Nemirovski (2011) which has the advantage of being verifiable even for large \( p \).

\[\text{VI} \quad \text{RECENT RESULTS ON THE PREDICTION LOSS}\]

We start this section, which contains some results from Dalalyan, Hebiri and Lederer (2014), by stating a negative result.
Let us consider the following example. Let \( n \geq 2 \), \( m = \lceil \sqrt{2n} \rceil \)
\( p = 2m \) and

\[
X = \begin{bmatrix}
1 & 1 & 1 & \ldots & 1 & 1 \\
1 & -1 & 0 & \ldots & 0 & 0 \\
0 & 0 & 1 & -1 & \ldots & 0 & 0 \\
0 & 0 & 0 & \ldots & 1 & -1 \\
\end{bmatrix} \times \sqrt{\frac{n}{2}} \quad \beta^* = \begin{bmatrix}
1 \\
1 \\
0 \\
\vdots \\
0 \\
\end{bmatrix}
\]

It is clear that \( \frac{1}{n} \| X^j \|_2^2 = 1 \). For simplicity, we replace the assumption \( \xi_i \overset{iid}{\sim} N(0, \sigma^2) \) by \( \xi_i \overset{iid}{\sim} \text{Rademacher} \) which means that \( \mathbb{P}(\xi_i = 1) = \mathbb{P}(\xi_i = -1) = \frac{1}{2} \).

**Proposition 4.** For every \( \lambda > 0 \), we have

\[
\mathbb{P}\left( \frac{1}{n} \| X(\hat{\beta}^\lambda - \beta^*) \|_2^2 \geq \frac{1}{\sqrt{8n}} \right) \geq \frac{1}{2}.
\]

**Comment.** This result shows that for some matrices \( X \) it is impossible to improve on the "slow rate."

One can note that in this example

- sparsity is very small : \( s = 2 \)
- the correlation is not very large : \( p = \frac{1}{2} \)
- the \( l_1 \)-norm of \( \beta^* \) is independent on \( n \) and \( p \)
- the negative result holds for any \( \lambda > 0 \). Furthermore, one can even show a stronger result of the form

\[
\mathbb{P}\left( \min_{\lambda > 0} \frac{1}{n} \| X(\hat{\beta}^\lambda - \beta^*) \|_2^2 \geq \frac{1}{\sqrt{8n}} \right) \geq \frac{1}{2}.
\]

- the proof is based on the first-order conditions which enable us to write down \( \hat{\beta}^\lambda \) explicitly.
**DEF** Let us define the weights

\[ w_j = \frac{1}{\sqrt{n}} \| (I_n - \Pi_J) X^j \|_2. \]

where \( \Pi_J \) is the orthogonal projector onto \( \text{Span}(X^j) \).

We say that \( X \) satisfies the weighted compatibility assumption \( WC(J) \) if for some \( \varepsilon > 0 \) we have

\[
\inf_{V \in C(J)} \frac{||J|| \times \|X \|_2^2}{n(||V||_1 - ||(1 - \frac{1}{2} \omega)_{\mathcal{C}} \otimes V \|_1)^2} > \varepsilon
\]

where \( C(J) = \{ V : \| (1 - \frac{1}{2} \omega)_{\mathcal{C}} \otimes V \|_1 \leq \| V \|_1 \} \).

**THEOREM 5**

Assume that \( X \) satisfies \( WC(J) \) for some \( J \subset \{1, \ldots, p\} \).

Choose \( \lambda = 2 \sigma \sqrt{\frac{2 \log(p/\delta)}{n}} \) for some \( \delta \in (0, 1) \). Then, with probability at least \( 1 - 2 \delta \) we have

\[
\ln(\hat{\beta}^L, \beta^*) \leq \inf_{\beta : \beta_{\mathcal{C}} = 0} \left\{ \ln(\beta, \beta^*) + \frac{4 \sigma^2 ||J|| \log(p/\delta)}{n} \Gamma_{n, p, l} \right\}
\]

where \( \Gamma_{n, p, l} = \frac{1}{\log(p/\delta)} + \frac{2}{||J||} + \frac{4 \sigma^2}{\varepsilon} \).

**Proof.** Cf. Dalalyan et al. (2014).

**REMARK** If \( X \) satisfies \( RE(3, J) \) it also satisfies \( WC(J) \) with the same constant \( \varepsilon \). The converse is not true, so Thm 5 is strictly stronger than Thm 4.
Indeed, on the one hand, on $C(J)$

$$
\left( \| V_3 \|_1 - \| (1 - \frac{1}{2} \omega)_J \odot V_3 \|_1 \right)^2 \leq \| V_3 \|_1^2 \leq |J| \| V_3 \|_2^2
\$$

and, therefore,

$$
\frac{|J| \cdot \| X V \|_2^2}{n \left( \| V_3 \|_1 - \| (1 - \frac{1}{2} \omega)_J \odot V_3 \|_1 \right)^2} \geq \frac{\| X V \|_2^2}{n \| V_3 \|_2^2}
\$$

On the other hand, $V \in C(J)$ implies that

$$
\| V_3 \|_1 \leq 2 \| (1 - \frac{1}{2} \omega)_J \odot V_3 \|_1 \leq 2 \| V_3 \|_1 \leq 3 \| V_3 \|_1
\$$

which implies that

$$
\inf_{V \in C(J)} \frac{|J| \cdot \| X V \|_2^2}{n \left( \| V_3 \|_1 - \| (1 - \frac{1}{2} \omega)_J \odot V_3 \|_1 \right)^2} \geq \inf_{V: \| V_3 \|_1 \leq 3 \| V_3 \|_1} \frac{\| X V \|_2^2}{n \| V_3 \|_2^2} = \infty_R
\$$

**Remark**

In the case of TV-penalized least-squares estimator

$$
\hat{f}^{TV} \in \arg \min_{f \in \mathbb{R}^n} \left( \frac{1}{2n} \| Y - f \|_2^2 + \lambda \sum_{i=0}^{n-1} | f_{i+1} - f_i | \right)
\$$

computing $\hat{f}^{TV}$ is equivalent to computing $X \hat{\beta}^L$ with

$$
X = \left[ \begin{array}{ccc}
1 & \cdots & 0 \\
1 & 1 & \cdots \\
\end{array} \right] \in \mathbb{R}^{n \times n}
\$$

Using the last theorem, we managed to prove that this matrix $X$ satisfies $WC(J)$ for every fixed $J$ with a constant $\varepsilon$ of order $\frac{1}{\log n}$.

This leads to

$$
\frac{1}{n} \| \hat{f}^{TV} - f^* \|_2^2 \leq C \times \sigma^2 |J| \log(\frac{\varepsilon}{8}) \log(n)
\$$

for any piece-wise constant signal $f^*$.