Aggregation for Linear Inverse Problems

Mohamed Hebiri, Jean-Michel Loubes, Paul Rochet

Aggregation for Linear Inverse Problems

M. Hebiri[*], J.M. Loubes[†], P. Rochet[‡]

May 22, 2014

Abstract

In the framework of inverse problems, we consider the question of aggregating estimators taken from a given collection. Extending usual results for the direct case, we propose a new penalty to achieve the best aggregation. An oracle inequality provides the asymptotic behavior of this estimator. We investigate here the price for considering indirect observations.

Introduction

In this article we are interested in recovering an unobservable signal x^* based on observations

$$y(t_i) = F(x^*)(t_i) + \delta_i, \tag{1}$$

where $F : \mathcal{X} \to \mathcal{Y}$ is a linear functional, with $\mathcal{X}$, $\mathcal{Y}$ Hilbert spaces and $t_i, i = 1, \ldots, n$ is a fixed observation scheme. Moreover, $x^* : \mathbf{R} \to \mathbf{R}$ is the unknown function to be recovered from the data $y(t_i)$, $i = 1, \ldots, n$. The regularity condition over the unknown parameter of interest is expressed through the assumption $x^* \in \mathcal{X}$ and will be made precise later in Section 3. We assume that the observations $y(t_i) \in \mathbf{R}$ and that the observation noise δ_i are i.i.d. realizations of a certain random variable δ. We assume F is Fréchet differentiable and ill-posed in the sense that our noise corrupted observations might lead to large deviations when trying to estimate x^*. In a deterministic framework, the statistical model (1) is formulated as the problem of approximating the solution in x of

$$F(x) = y,$$

when y is not known, and is only available through an approximation y^δ,

$$\|y - y^\delta\| \leqslant \delta.$$

[*]Université Paris-Est – Marne-la-Vallée
[†]Institut de Mathématiques de Toulouse
[‡]Université de Nantes

It is important to remark that whereas in this case consistency of the estimators depends on the approximation parameter δ, in (1) the noise level depends on the number of observations n, hence we will consider asymptotic results when the number of observations increases.

In the linear case, the best L^2 approximation of x^* is $x^\dagger = F^\dagger y$, where $F^\dagger$ is the Moore-Penrose (generalized) inverse of F. We will say the problem is ill-posed if $F^\dagger$ is unbounded. This might entail, and is generally the case, that $F^\dagger(y^\delta)$ is not close to $x^\dagger$. Hence, the inverse operator needs to be, in some sense, regularized.

Regularization methods replace an ill-posed problem by a family of well-posed problems. Their solution, called regularized solutions, are used as approximations of the desired solution of the inverse problem. These methods always involve some parameter measuring the closeness of the regularized and the original (unregularized) inverse problem. Rules (and algorithms) for the choice of these regularization parameters as well as convergence properties of the regularized solutions are central points in the theory of these methods, since they allow to find the right balance between stability and accuracy.
Hence there exist a wide range of possible estimators for inverse problems, each method with their advantages and their inconvenience. For a complete review on regularization methods for inverse problems, we refer to [8] and references therein. A natural idea is thus to look for a new, improved estimator constructed by combining such suitable estimators in a proper way. Such an estimator is called an aggregate and its construction is called aggregation. Aggregation of estimators have been studied within a large number of frameworks (we refer for instance to [3], [22] [5] and references therein). Here we study linear aggregation for inverse problems.

Actually one of the main differences between direct and indirect problems comes from the fact that two spaces are at hand: the space of the observations $\mathcal{Y}$ and the space where the function will be estimated, namely $\mathcal{X}$, the operator mapping one space into another, $F : \mathcal{X} \to \mathcal{Y}$. Hence to build a statistical procedure, a choice must be made which will determine the whole methodology. This question is at the core of the inverse problem structure and is encountered in many cases. When trying to build basis well adapted to the operator, two strategies can be chosen, either expanding the function onto a wavelet basis of the space $\mathcal{X}$ and taking the image of the basis by the operator as stated in [11], or expanding the image of the function onto a wavelet basis of $\mathcal{Y}$ and looking at the image of the basis by the inverse of the operator, studied in [1]. For the estimation problem with model selection theory, estimators can be obtained either by considering sieves on $(Y_m)_m \subset \mathcal{Y}$ with their counterpart $X_m := F^\star Y_m \subset \mathcal{X}$ or sieves on $(X_m)_m \subset \mathcal{X}$ and their image $Y_m := FX_m \subset \mathcal{Y}$ (see for instance in [17, 16]) where $F^\star$ states for the adjoint of F. In this paper we provide an aggregation procedure based on an ℓ^1 penalty which aggregates functions in the functional space $\mathcal{X}$. We prove that the choice of a penalty taking into account the ill-posedness of the inverse problem enables to recover an oracle inequality which warrants the good behavior of the estimate.

The paper falls into the following parts. Section 1 describes the inverse problem model we are dealing with. The main result concerning the behavior of the aggregation procedure is stated in Section 2 while all the proofs and auxiliary results are postponed to the Appendix **B**. The gap between the functional model and the functional observation model is tackled in the Appendix **A**.

1 Inverse problem model

Consider the following inverse model:

$$y_i = F(x^*)(t_i) + \delta_i,\ i = 1, \ldots, n. \tag{2}$$

$F : \mathcal{X} \to \mathcal{Y}$ is a known operator. Set Q_n the empirical measure of the covariates. The $L_2(Q_n)$-norm of a function $y \in \mathcal{Y}$ is then given by

$$\|y\|_n = \left(\int y^2 dQ_n \right)^{1/2} = \left(\frac{1}{n} \sum_{i=1}^{n} y^2(t_i) \right)^{\frac{1}{2}},$$

and the empirical scalar product by $< y, y' >_n = \frac{1}{n} \sum_{i=1}^{n} y'(t_i) y(t_i)$ for any $y' \in \mathcal{Y}$. This model is the discretized version of the inverse continuous model

$$Y(g) = (Fx^*, g) + \delta(g),\ g \in \mathcal{Y} \tag{3}$$

where $\delta(g)$ is a centered Gaussian variable with variance $\|g\|^2 := (g, g)$.

In the following, if F is a linear operator, we will denote $F^\star$ its adjoint. As often F is not of full rank, so the singular value decomposition (SVD) of the operator is then a useful tool. Let $(b_j; \varphi_j, \psi_j)_{j \geqslant 1}$ be a singular system for a linear operator F, that is, $F\psi_j = b_j \varphi_j$ and $F^\star \varphi_j = b_j \psi_j$; where $\{b_j^2\}_{j \geqslant 1}$ are the non zero eigenvalues of the selfadjoint operator $F^\star F$ (and also of $FF^\star$), considered in decreasing order. Furthermore, $\{\psi_j\}_{j=1,\ldots,n}$ and $\{\varphi_j\}_{j=1,\ldots,n}$ are a corresponding complete orthonormal system with respect to $\|.\|_n$ of eigenvectors of $F^\star F$ and $FF^\star$, respectively. For general linear operators with an SVD decomposition, we can write for any $x \in \mathcal{X}$ and $y \in \mathcal{Y}$

$$Fx = \sum_{j=1}^{n} b_j (x, \psi_j) \varphi_j \tag{4}$$

$$F^\star y = \sum_{j=1}^{n} b_j (y, \varphi_j) \psi_j. \tag{5}$$

For y in the domain of $F^\dagger$, $\mathcal{D}(F^\dagger)$, the best L^2-approximate solution has the expression

$$F^\dagger y = \sum_{j=1}^{n} \frac{(y, \varphi_j)}{b_j} \psi_j = \sum_{j=1}^{n} \frac{(F^\star y, \psi_j)}{b_j^2} \psi_j.$$

Note that for large j, the term $1/b_j$ grows to infinity. Thus, the *high frequency errors* are strongly amplified. This amplification measures the difficulty of the inverse problem, the faster the decay of the eigenvalues, the more difficult is the inverse problem. In this paper we will tackle the problem of polynomial decay of eigenvalues. So we assume that there exists an index t such that $b_j = \mathcal{O}(j^{-t/2})$ for some $t > 0$. The parameter t is called the index of ill-posedness of the operator F, following notations in [12].

2 Aggregation with ℓ^1 penalty for inverse problems

Let $\mathcal{C} = \{x_1, ..., x_M\}$, with $2 \leqslant M \leqslant n$, be a collection of functions in $\mathcal{X}$, independent from the observations. The x_m's can be viewed as preliminary estimators of x^*, constructed from some training sample. Aggregation procedures aim to build an estimator of x^* by combining in a suitable way the functions $x_1, ..., x_M$ (we refer to [19, 21, 6, 22] for relevant references in aggregation). The purpose is to filter out irrelevant elements in the collection $x_1, ..., x_M$ as well as to combine several possibly competing estimators. Thus, an estimator is sought as a linear combination of the x_m's, called *aggregate*, and noted

$$x_{\boldsymbol{\lambda}} = \sum_{m=1}^{M} \lambda_m x_m,$$

for $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_M)^\top$ lying in some subset Λ of $\mathbf{R}^M$.

As in many inverse problems, two points of view can be considered : either finding an approximation of the data in the operator space, or looking at the best possible aggregation in $\mathcal{X}$.

Let $\gamma(.)$ denote the following loss function

$$\gamma(v) = \sum_{j=1}^{n} | < y - Fv, \frac{\varphi_j}{b_j} >_n |^2, \ v \in \mathcal{X}.$$

This criterion corresponds to a quadratic loss between the image by the operator of a candidate function v and the observed data. Note that, this corresponds to *inverting* the operator F in the sense that

$$b_j^{-1} < y, \varphi_j >_n = < x, \psi_j >_n + b_j^{-1} \delta(\varphi_j). \tag{6}$$

Viewing the preliminary estimators $x_1, ..., x_M$ as a collection of regressors, a natural solution to the aggregation problem would be to consider the least square estimator, obtained by minimizing $\boldsymbol{\lambda} \mapsto \gamma(x_{\boldsymbol{\lambda}})$ over $\mathbb{R}^M$. However, this solution is known to be inefficient if the number of regressors is too large. For this reason, penalized procedures, favoring low-dimensional values of $\boldsymbol{\lambda}$ are often preferred to classical least square.
For a given penalty $\text{pen}(\boldsymbol{\lambda})$, the penalized aggregation estimator $\tilde{x} = x_{\tilde{\boldsymbol{\lambda}}}$ is built by minimizing over $\mathbb{R}^M$

$$\boldsymbol{\lambda} \mapsto \gamma(x_{\boldsymbol{\lambda}}) + \text{pen}(\boldsymbol{\lambda}) := L_n(\boldsymbol{\lambda}). \tag{7}$$

Since we promote sparsity, the penalty that is used in the present paper is defined as

$$\text{pen}(\boldsymbol{\lambda}) = \sum_{m=1}^{M} r_{n,m}|\lambda_m|, \tag{8}$$

with the notation $r_{n,m} = r_n\,\sigma_m$ with $r_n = 3\sqrt{2(\log M^2 n)/n}$ and $\sigma_m^2 = \frac{1}{n}\sum_{i=1}^{n} \frac{x_m^2(t_i)}{b_i^2}$. This penalty is highly inspired of the ℓ^1-penalty used in [5] and enjoys the property of detection of relevant elements in the collection of functions $\mathcal{C}$. The term r_n plays the role of the usual model selection penalty to prevent the aggregation of a too large number of functions. The term in σ_m is here an extra-term coming from the ill-posedness of the operator since it depends on the regularity of the functions with regards to the decay of the eigenvalues of the operator. In this way, it can be viewed as a source type condition as pointed out in [7] or [12]. Such a penalty, somewhat involving the ℓ^1-norm of the parameter $\boldsymbol{\lambda}$, are closely related to soft-thresholding, as discussed in [18] or [15].
We introduce some more notation : let $\mathbf{X}$ be the $n \times M$ design matrix with (i,m)-th component equals $x_m(t_i)$ for $i = 1, \dots n$ and $m = 1, \dots, M$. Moreover, for $q \in \mathbb{N}^*$, let $|\cdot|_q$ and $|\cdot|_\infty$ denote respectively the ℓ^q-norm and the ℓ^∞-norm in $\mathbb{R}^M$; that is, for a given vector $\boldsymbol{a} \in \mathbf{R}^M$, we write $|\boldsymbol{a}|_q^q = \sum_{m=1}^{M} |a_m|^q$ and $|\boldsymbol{a}|_\infty = \sup_{m=1,\dots M} |a_m|$. We also set the semi-norm $|\boldsymbol{a}|_0 = \sum_{m=1}^{M} \mathbb{1}_{\{a_m \neq 0\}}$, where $\mathbb{1}_{\{.\}}$ stands for the indicator function. Finally, for any subset S of $\{1, \dots, M\}$ and for a given vector $\boldsymbol{a} \in \mathbf{R}^M$, we introduce the notation $\boldsymbol{a}_S$ for the vector of size M whose components coincide with $\boldsymbol{a}$ in S, and equal 0 otherwise. We also denote by $|S|$ the cardinality of the set S.

We now state an assumption required to establish the theoretical result in this part. Fix $s \in \mathbb{N}^*$:

Assumption $RE(s)$**:** Let S be a subset of $\{1, \dots, M\}$, and define Γ_S the set $\Gamma_S = \{\boldsymbol{\delta} \in \mathbf{R}^M : \sum_{m \in S^c} \sigma_m|\delta_m| \leqslant 5 \sum_{m \in S} \sigma_m|\delta_m|\}$. We the assume that

$$\phi(s) := \min_{S \subset \{1,\dots,M\}:\, |S| \leqslant s} \quad \min_{\boldsymbol{\delta} \neq 0:\ \boldsymbol{\delta} \in \Gamma_S} \frac{\|\mathbf{x}_{\boldsymbol{\delta}}\|_n^2}{|\boldsymbol{\delta}_S|_2} > 0.$$

Note that with our notation, we have[1] $\|\mathbf{x}_{\boldsymbol{\delta}}\|_n^2 = \frac{\boldsymbol{\delta}^\top \mathbf{X}^\top \mathbf{X} \boldsymbol{\delta}}{n}$, where $\mathbf{X}$ is the matrix whose

[1] We refer the reader to Appendix **A** for details on this equality

columns are $\mathbf{X}_m = (x_m(t_1), \ldots, x_m(t_n))^\top$ (with $m = 1, \ldots, M$) and then the above assumption can be interpreted as a positive definiteness assumption of square sub-matrices of the Gram matrix $\mathbf{X}^\top \mathbf{X}$ with size smaller than s. This assumption has first been introduced in [2]. Some recent developments [20, 10] introduce other assumptions, weaker than Assumption RE, which also can be used in our framework. We prefer to use the more common Assumption RE to reduce extra technical arguments which would make the paper harder to read. Finally, we point out the book [4] for a complete display of the assumptions needed for ℓ^1-regularized methods.

Most controls on ℓ^1-regularized methods are established with high probability. To the best of our knowledge, the sharpest oracle inequalities for the Lasso (ℓ^1- penalized least squares estimator) available in the literature are the ones presented in [20, 10]. In what follows, we will exploit these results to improve them and develop a control on the error of the ℓ^1-penalized least-square estimator (7)-(8) in expectation :

Theorem 2.1 *Fix some integer $1 \leqslant \bar{s} \leqslant M$. Under the assumption Assumption $RE(\bar{s})$, the penalized estimator $\hat{\boldsymbol{x}}$ obtained $\hat{\boldsymbol{x}} = x_{\hat{\boldsymbol{\lambda}}} = \sum_{m=1}^M \hat{\lambda}_m x_m$, with*

$$\hat{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\lambda} \in \mathbf{R}^M} \left\{ \gamma(x_{\boldsymbol{\lambda}}) + 3\sqrt{2\frac{\log M^2 n}{n}} \sum_{m=1}^M \sigma_m |\lambda_m| \right\}$$

satisfies,

$$\begin{aligned} \mathbb{E}\|\hat{\boldsymbol{x}} - x^*\|_n^2 \quad \leqslant \quad & \inf_{\boldsymbol{\lambda} \in \mathbb{R}^M :\ |S| = s \leqslant \bar{s}} \left\{ \|x_{\boldsymbol{\lambda}} - x^*\|_n^2 + \frac{25}{36} r_n^2 \, \phi^{-2}(s) \sum_{m \in S} \sigma_m^2 \right\} \\ & + \frac{4\|x^*\|_n^2 + 12 b_{max}^{-2}}{Mn} + \frac{6 b_{max}^{-2}(M+1)}{n} \exp\left(-\frac{n}{8}\right), \end{aligned}$$

where the set S is defined as $S = \{m \in \{1, \ldots, M\} :\ \lambda_m \neq 0\}$ and $b_{max}^{-2} = \max_{m=1,\ldots,M} b_m^{-2}$.

The following theorem provides an oracle inequality that controls the aggregation procedure. The inequality is sharp, that is, the leading constant in front of the main term is 1. Morover, several quantities are of interest in the above bound.
The main term is given by $\inf_{\boldsymbol{\lambda}} \left\{ \|x_{\boldsymbol{\lambda}} - x^*\|_n^2 + \frac{25}{36} r_n^2 \, \phi^{-2}(s) \sum_{m \in S} \sigma_m^2 \right\}$. It is composed of a bias term and an additional term where $\sum_{m \in S} \sigma_m^2$ plays the role of the sparsity index. The rate is penalized on the one hand by r_n^2 and on the other hand by $\sigma_m^2 = \frac{1}{n} \sum_{i=1}^n \frac{x_m^2(t_i)}{b_i^2}$ for all the different functions x_m that are selected in the aggregation set S. This term can be seen as a source condition that links the smoothness of the functions to the decay of the eigenvalues of the inverse operator. It is bounded under the usual source condition assumption. Then if there exists a $\boldsymbol{\lambda}^*$ such that $x_{\boldsymbol{\lambda}^*} = x^*$, and given the definition of r_n^2,

the rate of convergence is $\frac{\log(M^2 n)}{n} \sum_{m \in S^*} \sigma_m^2$, where S^* is the true sparsity index. Compared to the usual rate of convergence, we accepted here to lose a log factor ($\log(M^2 n)$ instead of $\log(M)$) in order to provide a bound in expectation.

The remainder term is made of two parts. An exponential bound which is negligible and a second term of order b_{max}^{-2}/Mn which is the price to pay for using aggregation in an inverse problem settings. Hence when the problem is mildly ill-posed, i.e when the coefficients of the SVD decay at a polynomial rate $b_j = Cj^{-t/2}$ for t the index of ill-posedness, this term is of order n^{t-1}. Note that this term goes to zero when t is smaller than 1, yet hampering the consistency rate. In other cases and in the severely ill-posed setting, this term becomes dominant in the upper bound.

Hence aggregation methods for inverse problems have the same kind of drawbacks than ℓ^1 penalization procedure since they cannot handle too badly ill-posed inverse problems.

Appendix A

The definition of the aggregation estimator given in (7) or (8) is based on a vectorial model while the observation Model (2) is functional. One then needs to establish links between them and to show how the observational model can be written into a sequential one.

We recall that the model is given in (6) by the following equation

$$b_j^{-1} < y, \varphi_j >_n = < x^*, \psi_j >_n + b_j^{-1} \delta(\varphi_j),$$

with $\delta(\varphi_j) = < \delta, \varphi_j >_n$. Hence define $z \in \mathcal{X}$ as $z_j = z(\psi_j) := b_j^{-1} < y, \varphi_j >_n$. This function will be observed with an heteroscedatic observation noise $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$ defined as $\varepsilon_j = b_j^{-1} \delta(\varphi_j)$. Hence, Model (2) can be written as an heteroscedastic model

$$z(\psi_j) = < x^*, \psi_j >_n + \varepsilon_j, \; j = 1, \ldots, n \tag{9}$$

where $\boldsymbol{\varepsilon} \in \mathbf{R}^n$ is a Gaussian noise with heterogeneous variance Σ given by the diagonal terms $\mathrm{Var}(\varepsilon_j) = b_j^{-2}$, growing to infinity while $j \to +\infty$. Moreover, if we denote by $\mathbf{z}$ the vector $\mathbf{z} = (z_1, \ldots, z_n)^\top$ whose components are $z_j := z(\psi_j)$ given above, the model becomes

$$\mathbf{z} = \mathbf{x}^* + \boldsymbol{\varepsilon}, \tag{10}$$

where $\mathbf{x}^*$ is the vector in $\mathbf{R}^n$ with j-th component equal $< x^*, \psi_j >_n$, $j = 1, \ldots, n$. Both above models are not observed and then are unusable. However, their introduction is motived by technical arguments and make calculation more synthetic. It also illustrates how our inverse problem can be summarized into an heteroscedastic linear problem.
Now consider the aggregate estimator considered in this paper and recall its definition

$$x_{\boldsymbol{\lambda}} = \sum_{m=1}^{M} \lambda_m x_m$$

for some $\boldsymbol{\lambda} \in \Lambda$. Then for any function h in $\mathcal{X}$ we can write

$$\langle h, x_{\boldsymbol{\lambda}} \rangle_n = \frac{\mathbf{h}^\top \mathbf{X} \hat{\boldsymbol{\lambda}}}{n},$$

where X is the design matrix whose components are the $x_m(t_i)$ and $\mathbf{h}$ is the vector in $\mathbf{R}^n$ whose j-th coordinate is $h(\psi_j)$.

Appendix B

Proof of Theorem 2.1. This theorem is a control on the prediction error in expectation. To prove it, we establish an intermediate result where we propose a control on the error on the event $\mathcal{A}$ defined by

$$\mathcal{A} = \bigcap_{m=1}^{M} \{3|V_m| \leqslant r_{n,m}\}, \qquad \text{with } V_m = \frac{\mathbf{X}_j^\top \boldsymbol{\varepsilon}}{n},$$

and which holds with large probability (*cf.* Appendix A for the definition of $\boldsymbol{\varepsilon}$).

Proposition 2.2 *Under the assumption of Theorem 2.1, we have on the set* $\mathcal{A}$

$$\|\hat{\boldsymbol{x}} - x^*\|_n^2 \leqslant \inf_{\boldsymbol{\lambda} \in \mathbb{R}^M \ |S| \leqslant s} \left\{ \|x_{\boldsymbol{\lambda}} - x^*\|_n^2 + \frac{25}{36} r_n^2 \, \phi^{-2}(\bar{s}) \sum_{m \in S} \sigma_m^2 \right\},$$

for any $s \leqslant \bar{s}$, *where* $S = \{m \in \{1, \ldots, M\} : \ \lambda_m \neq 0\}$.

Proof of Proposition 2.2. The proof of this result is inspired by the proof of Theorem 2 in [14]. First of all, we notice that if $\langle \hat{\boldsymbol{x}} - x^*, \hat{\boldsymbol{x}} - x_{\boldsymbol{\lambda}} \rangle_n \leqslant 0$, then the identity

$$2 \langle \hat{\boldsymbol{x}} - x^*, \hat{\boldsymbol{x}} - x_{\boldsymbol{\lambda}} \rangle_n = \|\hat{\boldsymbol{x}} - x^*\|_n^2 + \|\hat{\boldsymbol{x}} - x_{\boldsymbol{\lambda}}\|_n^2 - \|x_{\boldsymbol{\lambda}} - x^*\|_n^2 \tag{11}$$

implies $\|\hat{\boldsymbol{x}} - x^*\|_n^2 \leqslant \|x_{\boldsymbol{\lambda}} - x^*\|_n^2$. Then the bound in the proposition is valid.
Then, let us consider the case where $\langle \hat{\boldsymbol{x}} - x^*, \hat{\boldsymbol{x}} - x_{\boldsymbol{\lambda}} \rangle_n > 0$. Recall that we have set $r_n = 3\sqrt{2\frac{\log(M^2 n)}{n}}$ and $\sigma_m^2 = n^{-1} \sum_{i=1}^n \frac{X_{i,m}^2}{b_i^2}$, where $X_{i,m} = x_m(t_i)$ for $i \in \{1, \ldots, n\}$ and $m \in \{1, \ldots, M\}$. In this case, we exploit the optimality condition of the minimization criterion (7)-(8). Since $\hat{\boldsymbol{\lambda}}$ is minimizer of this criterion, the first order optimality conditions imply that

$$2\frac{\mathbf{X}^\top \mathbf{z}}{n} - 2\frac{\mathbf{X}^\top \mathbf{X} \hat{\boldsymbol{\lambda}}}{n} \in r_n \, \partial |\hat{\boldsymbol{\lambda}}|_{1,\sigma}$$

where for any $\boldsymbol{\lambda}$, the quantity $\partial |\boldsymbol{\lambda}|_{1,\sigma}$ denotes the sub-differential of the weighted ℓ^1-norm, defined for any vector $\boldsymbol{a} \in \mathbf{R}^M$ by $|\boldsymbol{a}|_{1,\sigma} = \sum_{m=1}^M \sigma_m |a_m|$. Set $S = \{m : \ \lambda_m \neq 0\}$, the

sparsity pattern of $\boldsymbol{\lambda}$. Thanks to sub-differential of the ℓ^1-norm in $\mathbf{R}^M$, we deduce the set of sub-differential of the above weighted ℓ^1-norm

$$\partial|\boldsymbol{\lambda}|_{1,\sigma} = \{\boldsymbol{\mu} \in \mathbf{R}^M : \quad \mu_m = \sigma_m \mathrm{sgn}(\lambda_m) \text{ if } m \in S \quad \text{and} \quad \mu_m \in [-\sigma_m, \sigma_m] \text{ if } m \in S^c\},$$

where, for a given $\boldsymbol{a} \in \mathbf{R}$, $\mathrm{sgn}(\boldsymbol{a})$ equals ± 1 according to the sign of $\boldsymbol{a}$, and S^c denotes the complementary set of S in $\{1, \ldots, M\}$ (*cf.* [13, page 259] for details on sub-differential tools). Based on the above statement, we can write first

$$2\frac{\hat{\boldsymbol{\lambda}}^\top \mathbf{X}^\top \mathbf{z}}{n} - 2\frac{\hat{\boldsymbol{\lambda}}^\top \mathbf{X}^\top \mathbf{X}\hat{\boldsymbol{\lambda}}}{n} = r_n |\hat{\boldsymbol{\lambda}}|_{1,\sigma} \tag{12}$$

where $\mathbf{z}$ is given by (10) ; second for any $\boldsymbol{\lambda} \in \mathbf{R}^M$ with sparsity pattern $S = \{m : \ \lambda_m \neq 0\}$ we may write

$$2\frac{\boldsymbol{\lambda}^\top \mathbf{X}^\top \mathbf{z}}{n} - 2\frac{\boldsymbol{\lambda}^\top \mathbf{X}^\top \mathbf{X}\hat{\boldsymbol{\lambda}}}{n} \leqslant r_n |\boldsymbol{\lambda}|_{1,\sigma} \tag{13}$$

Subtracting (12) from (13), we get for any $\boldsymbol{\lambda} \in \mathbf{R}^M$ with sparsity pattern $S = \{m : \ \lambda_m \neq 0\}$

$$2\frac{(\boldsymbol{\lambda} - \hat{\boldsymbol{\lambda}})^\top \mathbf{X}^\top \mathbf{z}}{n} - 2\frac{(\boldsymbol{\lambda} - \hat{\boldsymbol{\lambda}})^\top \mathbf{X}^\top \mathbf{X}\hat{\boldsymbol{\lambda}}}{n} \leqslant r_n(|\boldsymbol{\lambda}|_{1,\sigma} - |\hat{\boldsymbol{\lambda}}|_{1,\sigma}).$$

Moreover, according to (10), we have $\mathbf{z} = \mathbf{x}^* + \boldsymbol{\varepsilon}$ and then the above inequality becomes

$$2\frac{(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda})^\top \mathbf{X}^\top \mathbf{X}\hat{\boldsymbol{\lambda}}}{n} - 2\frac{(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda})^\top \mathbf{X}^\top \mathbf{x}^*}{n} \leqslant r_n(|\boldsymbol{\lambda}|_{1,\sigma} - |\hat{\boldsymbol{\lambda}}|_{1,\sigma}) + 2\frac{(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda})^\top \mathbf{X}^\top \boldsymbol{\varepsilon}}{n}.$$

Now, using the correspondence between vectorial and functional notations stated in Appendix A 2, this inequality states that for any $\boldsymbol{\lambda} \in \mathbf{R}^M$ with sparsity pattern $S = \{m : \lambda_m \neq 0\}$

$$2\left\langle \hat{\boldsymbol{x}} - x^*, \hat{\boldsymbol{x}} - x_{\boldsymbol{\lambda}} \right\rangle_n \leqslant r_n(|\boldsymbol{\lambda}|_{1,\sigma} - |\hat{\boldsymbol{\lambda}}|_{1,\sigma}) + 2\frac{(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda})^\top \mathbf{X}^\top \boldsymbol{\varepsilon}}{n}. \tag{14}$$

Considering the sparsity pattern of $\boldsymbol{\lambda}$, the first term on the rhs of the above inequality can be decomposed as

$$\begin{aligned} r_n(|\boldsymbol{\lambda}|_{1,\sigma} - |\hat{\boldsymbol{\lambda}}|_{1,\sigma}) &= -\sum_{m \in S^c} r_{n,m} |\hat{\lambda}_m| + \sum_{m \in S} r_{n,m} \left(|\lambda_m| - |\hat{\lambda}_m| \right) \\ &\leqslant -\sum_{m \in S^c} r_{n,m} |\hat{\lambda}_m| + \sum_{m \in S} r_{n,m} |\hat{\lambda}_m - \lambda_m| \end{aligned}$$

where we set $r_{n,m} = r_n \, \sigma_m$. Note $\langle \cdot, \cdot \rangle$ the usual scalar product in $\mathbf{R}^M$ and let $V = \frac{\boldsymbol{\varepsilon}^\top \mathbf{X}}{n}$ for short, then (14) becomes

$$2\left\langle \hat{\boldsymbol{x}} - x^*, \hat{\boldsymbol{x}} - x_{\boldsymbol{\lambda}} \right\rangle_n + \sum_{m \in S^c} r_{n,m} |\hat{\lambda}_m| \leqslant \sum_{m \in S} r_{n,m} |\hat{\lambda}_m - \lambda_m| + 2\left\langle V, \hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda} \right\rangle.$$

Thanks to (11), the above inequality gives us the fundamental results

$$\|\hat{\boldsymbol{x}} - x^*\|_n^2 + \|\hat{\boldsymbol{x}} - x_{\boldsymbol{\lambda}}\|_n^2 + \sum_{m\in S^c} r_{n,m}|\hat{\lambda}_m| \leqslant \|x_{\boldsymbol{\lambda}} - x^*\|_n^2 + \sum_{m\in S} r_{n,m}|\hat{\lambda}_m - \lambda_m| + 2\left\langle V, \hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}\right\rangle. \tag{15}$$

Once we established this last major inequality, we will first use it to show that $\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}$ belongs to the set Γ_S in Assumption $RE(\bar{s})$. Then we will use it again to establish the bound announced in the proposition.
First, since $\lambda_m = 0$ for $m \in S^c$, (15) implies that

$$\begin{aligned} \textstyle\sum_{m\in S^c} r_{n,m}|\hat{\lambda}_m| \quad &\leqslant \sum_{m\in S} r_{n,m}|\hat{\lambda}_m - \lambda_m| + 2\sum_{m\in S}|V_m||\hat{\lambda}_m - \lambda_m| + 2\sum_{m\in S^c}|V_m||\hat{\lambda}_m| \\ \Leftrightarrow \quad \textstyle\sum_{m\in S^c}(r_{n,m} - |V_m|)|\hat{\lambda}_m| \quad &\leqslant \sum_{m\in S}(r_{n,m} + |V_m|)|\hat{\lambda}_m - \lambda_m|. \end{aligned} \tag{16}$$

On the set $\mathcal{A} := \bigcap_{m=1}^M \{3|V_m| \leqslant r_{n,m}\}$, we easily obtain $\sum_{m\in S^c}\sigma_m|\hat{\lambda}_m| \leqslant 5\sum_{m\in S}\sigma_m|\hat{\lambda}_m - \lambda_m|$ and then the vector $\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}$ belongs to Γ_S as announced above. Since $s \leqslant \bar{s}$, Assumption $RE(\bar{s})$ implies Assumption $RE(s)$, and as a consequence (thanks to Assumption $RE(s)$), we can write

$$|(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda})_S|_2 \leqslant \phi^{-1}(s)\,\|\hat{\boldsymbol{x}} - x_{\boldsymbol{\lambda}}\|_n.$$

Combining this last inequality, with (15) and the fact that on the set $\mathcal{A}$, $r_{n,m} \geqslant 3|V_m|$ (and then $r_{n,m} - 2|V_m| \geqslant r_{n,m}/3$) for all $m \in \{1,\dots,M\}$, we have

$$\begin{aligned} \|\hat{\boldsymbol{x}} - x^*\|_n^2 + \|\hat{\boldsymbol{x}} - x_{\boldsymbol{\lambda}}\|_n^2 + \sum_{m\in S^c}\frac{r_{n,m}}{3}|\hat{\lambda}_m| &\leqslant \|x_{\boldsymbol{\lambda}} - x^*\|_n^2 + \left(1 + \frac{2}{3}\right)\sum_{m\in S} r_{n,m}|\hat{\lambda}_m - \lambda_m| \\ &\leqslant \|x_{\boldsymbol{\lambda}} - x^*\|_n^2 + \frac{5}{3}r_n\sqrt{\sum_{m\in S}\sigma_m^2}\,|(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda})_S|_2 \\ &\leqslant \|x_{\boldsymbol{\lambda}} - x^*\|_n^2 + \frac{5}{3}r_n\sqrt{\sum_{m\in S}\sigma_m^2}\,\phi^{-1}(s)\,\|\hat{\boldsymbol{x}} - x_{\boldsymbol{\lambda}}\|_n \\ &\leqslant \|x_{\boldsymbol{\lambda}} - x^*\|_n^2 + \frac{25}{36}r_n^2\,\phi^{-2}(s)\sum_{m\in S}\sigma_m^2 + \|\hat{\boldsymbol{x}} - x_{\boldsymbol{\lambda}}\|_n^2. \end{aligned}$$

where we used, for the first inequality, similar reasoning as those exploited to get (16). We also used Cauchy-Schwarz Inequality and the fact that $r_{n,m} = r_n\,\sigma_m,\ \forall m \in \{1,\dots,M\}$ for the second inequality, and the relation $2ab \leqslant a^2 + b^2$ (for any positive reals a and b) in the last one. Subtracting $\|\hat{\boldsymbol{x}} - x_{\boldsymbol{\lambda}}\|_n^2$ to both sides leads to the result in the proposition

$$\mathbb{E}\|\hat{\boldsymbol{x}} - x^*\|_n^2\mathbb{1}_{\mathcal{A}} \leqslant \inf_{\lambda\in\mathbb{R}^M\ |S|\leqslant s}\left\{\|x_{\boldsymbol{\lambda}} - x^*\|_n^2 + \frac{25}{36}r_n^2\,\phi^{-2}(\bar{s})\sum_{m\in S}\sigma_m^2\right\}.$$

since $\phi^{-2}(s) \leqslant \phi^{-2}(\bar{s})$ for all $s \leqslant \bar{s}$. This finishes the proof of Proposition 2.2.

Now, let's go back to the proof of the theorem. It remains to deal with error when the event $\mathcal{A}^c$ occurs. By definition, $\gamma(\hat{\boldsymbol{x}}) + \text{pen}(\hat{\boldsymbol{\lambda}}) \leqslant \gamma(x_{\boldsymbol{\lambda}}) + \text{pen}(\boldsymbol{\lambda})$ for all $\boldsymbol{\lambda} \in \mathbb{R}^M$. Taking $\boldsymbol{\lambda} = 0$, we deduce that $\gamma(\hat{\boldsymbol{x}}) \leqslant 0$.Moreover, using the definition of γ we find

$$\begin{aligned}
\|\hat{\boldsymbol{x}} - x^*\|_n^2 &\leqslant \|x^*\|_n^2 + 2|\langle \hat{\boldsymbol{x}}, \varepsilon \rangle_n| \leqslant \|x^*\|_n^2 + 2\|\hat{\boldsymbol{x}}\|_n \|\varepsilon\|_n \\
&\leqslant \|x^*\|_n^2 + 2\|\varepsilon\|_n (\|\hat{\boldsymbol{x}} - x^*\|_n + \|x^*\|_n) \\
&\leqslant \|x^*\|_n^2 + \frac{\|\hat{\boldsymbol{x}} - x^*\|_n^2}{2} + \|x^*\|_n^2 + 3\|\varepsilon\|_n^2
\end{aligned}$$

using the inequality $2ab \leqslant \theta a^2 + \theta^{-1} b^2$ successively for $\theta = 1/2$ and $\theta = 1$.
The random variable $W = n\|\Sigma^{-1/2}\varepsilon\|_n^2$ has Chi-square distribution with M degrees of freedom and satisfies $b_{max}^{-2} W/n \geqslant \|\varepsilon\|_n^2$ where $b_{max}^{-2} = \max_{m=1,\dots,M} b_m^2$. Thus,

$$\|\hat{\boldsymbol{x}} - x^*\|_n^2 \leqslant 4\|x^*\|_n^2 + \frac{6b_{max}^{-2}}{n} W.$$

Following the proof in [5], we now introduce the event $\mathcal{B} = \{W \leqslant 2n\}$. Remark that $\mathbb{E}(W \mathbb{1}_{\mathcal{A}^c}) \leqslant 2n\mathbb{P}(\mathcal{A}^c) + \mathbb{E}(W \mathbb{1}_{\mathcal{B}^c})$, where the second term can be bounded by

$$\mathbb{E}(W \mathbb{1}_{\mathcal{B}^c}) \leqslant \sqrt{\mathbb{E}(W^2)} \sqrt{\mathbb{P}(\mathcal{B}^c)},$$

by Cauchy-Schwarz's inequality. Since W has $\chi^2(M)$ distribution (with $M \leqslant n$, it satisfies in particular $\mathbb{E}(W^2) \leqslant (M+1)^2$ and $\mathbb{P}(W > 2n) \leqslant \mathbb{P}(\chi^2(n) > 2n) \leqslant \exp(-n/8)$ (for the second statement, see [9], page 857). Moreover, since $V_m \sim \mathcal{N}(0, n^{-1}\sigma_m^2)$, a standard tail bound for Gaussian distributions gives

$$\begin{aligned}
\mathbb{P}(\mathcal{A}^c) &\leqslant \mathbb{P}\left(\bigcup_{m=1}^M \{3|V_m| > r_{n,m}\} \right) \leqslant \sum_{m=1}^M \mathbb{P}\left(|V_m| > \frac{r_{n,m}}{3} \right) \\
&\leqslant \sum_{m=1}^M \exp\left\{ -\frac{(r_{n,m}/3)^2}{2n^{-1}\sigma_m^2} \right\} = \sum_{m=1}^M \frac{1}{M^2 n} = (nM)^{-1},
\end{aligned}$$

yielding

$$\mathbb{E}(\|\hat{\boldsymbol{x}} - x^*\|_n^2 \mathbb{1}_{\mathcal{A}^c}) \leqslant \frac{4\|x^*\|_n^2 + 12b_{max}^{-2}}{Mn} + \frac{6b_{max}^{-2}(M+1)}{n} \exp\left(-\frac{n}{8}\right),$$

which completes the proof.

References

[1] F. Abramovich and B. W. Silverman. Wavelet decomposition approaches to statistical inverse problems. *Biometrika*, 85(1):115–129, 1998.

[2] P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 2009.

[3] L. Birgé. Model selection via testing: an alternative to (penalized) maximum likelihood estimators. *Ann. Inst. H. Poincaré Probab. Statist.*, 42(3):273–325, 2006.

[4] P. Buehlmann and S. van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, 2011.

[5] F. Bunea, A. Tsybakov, and M. Wegkamp. Aggregation for Gaussian regression. *Ann. Statist.*, 35(4):1674–1697, 2007.

[6] O. Catoni. *Statistical Learning Theory and Stochastic Optimization*. Lecture Notes in Mathematics 1851. Springer-Verlag, Berlin. Lecture notes from the 31st Summer School on Probability Theory held in Saint-Flour 2001, 2004.

[7] L. Cavalier. Nonparametric statistical inverse problems. *Inverse Problems*, 24(3):034004, 19, 2008.

[8] L. Cavalier. Inverse problems in statistics. In *Inverse problems and high-dimensional estimation*, volume 203 of *Lect. Notes Stat. Proc.*, pages 3–96. Springer, Heidelberg, 2011.

[9] L. Cavalier, G. K. Golubev, D. Picard, and A. B. Tsybakov. Oracle inequalities for inverse problems. *Ann. Statist.*, 30(3):843–874, 2000.

[10] A. Dalalyan, M. Hebiri, and J. Lederer. On the prediction performance of the Lasso. *Submitted*, 2014.

[11] D.L. Donoho. Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition. *Appl. Comput. Harmon. Anal.*, 2(2):101–126, 1995.

[12] H. Engl, M. Hanke, and A. Neubauer. *Regularization of inverse problems*, volume 375 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht, 1996.

[13] J.B. Hiriart-Urruty and C. Lemarechal. *Convex Analysis and Minimization Algorithms: Part 1: Fundamentals*. Grundlehren der mathematischen Wissenschaften Series. Springer, 2011.

[14] V. Koltchinskii, A. Tsybakov, and K. Lounici. Nuclear norm penalization and optimal rates for noisy low rank matrix completion. 2011.

[15] J-M. Loubes. l^1 penalty for ill-posed inverse problems. *Comm. Statist. Theory Methods*, 37(8-10):1399–1411, 2008.

[16] J-M. Loubes and C. Ludeña. Model selection for non linear inverse problems. *ESAIM PS*.

[17] J-M. Loubes and C. Ludeña. Adaptive complexity regularization for linear inverse problems. *Electron. J. Stat.*, 2:661–677, 2008.

[18] J-M Loubes and S. van de Geer. Adaptive estimation with soft thresholding penalties. *Statist. Neerlandica*, 56(4):454–479, 2002.

[19] A. Nemirovski. *Topics in non-parametric statistics*. Lecture Notes in Mathematics 1738. Springer, New York. Lecture notes from the 28th Summer School on Probability Theory held in Saint-Flour 1998, 2000.

[20] T. Sun and C-H. Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012.

[21] A. Tsybakov. Optimal rates of aggregation. *COLT, Lecture Notes in Computer Science . Springer*, pages 303–313, 2003.

[22] Y. Yang. Aggregating regression procedures to improve performance. *Bernoulli*, 10(1):25–47, 2004.