Aggregation for Linear Inverse Problems

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

July 10, 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.

Keywords: Heteroscedastic linear model; ℓ^1 penalization; oracle inequalities

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 is a linear operator, $\mathcal{X}$, $\mathcal{Y}$ functional Hilbert spaces, $t_i, i = 1, \ldots, n$ is a fixed observation scheme and $\delta = (\delta_1, ..., \delta_n)$ a noise vector.

Due to indirect nature of the observation, nothing is known of the component of x^* lying in the kernel of F. This means that the best L^2 approximation one can get from the available information is the orthogonal projection of x^*

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

onto $\ker(F)^{\perp}$. This function, usually noted $x^{\dagger}$, can be expressed as the image of $F(x^*)$ through the Moore-Penrose (generalized) inverse of F, i.e., the reverse image of $F(x^*)$ with minimal norm on $\mathcal{X}$. The inverse problem is said to be *ill-posed* if the Moore-Penrose inverse $F^{\dagger}$ is unbounded. This might entail, and is generally the case, that $F^{\dagger}(y)$ is not close to $x^{\dagger}$. Hence, the inverse operator needs to be, in some sense, regularized.

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 [13], 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 [18, 19]) where $F^{\star}$ states for the adjoint of F.

Regularization methods replace an ill-posed problem by a family of well-posed problems. Numerous regularization methods have been proposed such as Tikhonov regularization , iterative Landweber's method or truncated singular value decomposition to cite a few. In each case, the regularized solutions are used as approximations of the desired solution of the inverse problem. These methods 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 also refer to [8] and references therein.

A natural idea is thus to look for a new, improved estimator called *aggregate*, constructed by combining the existing estimators in a proper way. Aggregation of estimators have been studied within a large number of frameworks (we refer for instance to [3], [25] [5] and references therein). Here, we study linear aggregation in the context of inverse problems. More precisely, we assume that a collection $x_1, ..., x_M$ of preliminary estimators of x^* are available, and we search for the best linear combination of them. We provide an aggregation procedure based on an empirical risk minimization with ℓ^1 penalty, which aggregates functions in the space $\mathcal{X}$. The advantage of a ℓ^1 penalty is to promote the sparsity of the solution while preserving the convexity of the minimizing criterion. In the frame of aggregation, sparsity is a crucial issue as it allows one to select only the relevant estimators in the collection $x_1, ..., x_M$. 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.

1 Inverse problem model

Consider the following inverse problem

$$y(t_i) = F(x^*)(t_i) + \delta_i, \tag{2}$$

where F is a linear operator, $\mathcal{X}$, $\mathcal{Y}$ functional Hilbert spaces and $t_i, i = 1, \ldots, n$ is a fixed observation scheme. In this framework, it is important to remark that, while the image $F(x^*)$ lies in $\mathcal{Y}$, the available information regarding x^*

is more faithfully described by the discretized operator $F_n : \mathcal{X} \to \mathbb{R}^n$ defined as

$$F_n(x) = (F(x)(t_1), ..., F(x)(t_n))^\top, \ x \in \mathcal{X},$$

with $a^\top$ the transpose of a. In this setting, the observation vector $y := (y(t_1), ..., y(t_n))^\top \in \mathbb{R}^n$ can be defined as the noisy image of x^* through the operator F_n,

$$y = F_n x^* + \delta.$$

The operator F_n is assumed one-to-one and continuous for $\mathbb{R}^n$ endowed with inner product $\langle a, b \rangle_n = \frac{1}{n} \sum_{i=1}^n a_i b_i$ and associated norm $\|.\|_n$. The inner product and norm on $\mathcal{X}$ are simply noted $\langle ., . \rangle$ and $\|.\|$ respectively. Moreover, we assume for simplicity that δ is a standard Gaussian vector of $\mathbb{R}^n$. This particular situation has been extensively studied in the literature in inverse problems, see for instance [9, 10, 20, 22].

A useful tool to describe a linear inverse problem is to use the singular value decomposition (SVD). Precisely, let $F_n^\star$ denote the adjoint of F_n and let $b_1^2 \geqslant ... \geqslant b_n^2$ be the ordered eigenvalues of $F_n F_n^\star$ (which are also the non-zero eigenvalues of $F_n^\star F_n$). Now denote by $\varphi_1, ..., \varphi_n \in \mathbb{R}^n$ the corresponding normed eigenvectors of $F_n F_n^\star$. Because $F_n F_n^\star$ is self-adjoint, we know that $\varphi_1, ..., \varphi_n$ form an orthonormal basis of $\mathbb{R}^n$. Similarly, let $\psi_1, ..., \psi_n \in \mathcal{X}$ denote the normed eigenvectors of $F_n^\star F_n$ which form an orthonormal basis of $\ker(F_n)^\perp$. The system $\{b_j; \psi_j, \varphi_j\}_{j=1,...,n}$ is called the singular system of the linear operator F_n. We have in particular, $F_n \psi_j = b_j \varphi_j$ and $F_n^\star \varphi_j = b_j \psi_j$ for all $j = 1, ..., n$. More generally, for any $x \in \mathcal{X}$ and $y \in \mathbb{R}^n$, one can write

$$F_n x = \sum_{j=1}^n b_j \langle x, \psi_j \rangle \ \varphi_j \quad \text{and} \quad F_n^\star y = \sum_{j=1}^n b_j \langle y, \varphi_j \rangle_n \ \psi_j. \tag{3}$$

The Moore-Penrose generalized inverse of F_n, noted $F_n^\dagger$, can be defined as the operator $F_n^\dagger : \mathbb{R}^n \to \mathcal{X}$ with singular system $\{b_j^{-1}; \varphi_j, \psi_j\}_{j=1,...,n}$. The main interest in the Moore-Penrose inversion is that the original inverse problem (2) can be turned into a direct problem, considering $z := F_n^\dagger(y)$, which in the basis $\psi_1, ..., \psi_n$, leads to the following model

$$z_j = x_j^\dagger + \varepsilon_j, \ j = 1, ..., n, \tag{4}$$

where $z_j = \langle F_n^\dagger(y), \psi_j \rangle = b_j^{-1} \langle y, \varphi_j \rangle_n$, $x_j^\dagger = \langle x^\dagger, \psi_j \rangle$ and $\varepsilon_j = \langle F_n^\dagger(\delta), \psi_j \rangle$. In this model, we point out that the noises remain Gaussian, but with unequal variances as we have $\varepsilon_j = b_j^{-1} \langle \delta, \varphi_j \rangle_n \sim \mathcal{N}(0, b_j^{-2})$. Moreover, ε_i and ε_j are independent for $i \neq j$ due to the orthogonality of the φ_j's.

Note that the inverse eigenvalues b_j^{-2} grow with j, resulting in the *high frequency errors* being strongly amplified in the observation z_j. 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 is of the order $j^{-t/2}$ for some $t > 0$. The parameter t is called the index of ill-posedness of the operator F_n, following notations in [14].

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 [21, 24, 6, 25] 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

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

for $\lambda = (\lambda_1, ..., \lambda_M)^\top \in \mathbb{R}^M$. Due to the absence of information on $\ker(F_n)$, we are only interested in the behavior of the x_m's on $\ker(F_n)^\perp$. For convenience, we assume that for all $m = 1, ..., M$, $x_m \in \ker(F_n)^\perp$. This condition is very natural since most regularization methods for inverse problems provide a solution in $\in \ker(F_n)^\perp$. Moreover, if one element x_m does not satisfy this

condition, one can simply replace it by its orthogonal projection onto $\ker(F_n)^\perp$.

The best linear combination x_λ to approximate x^* can be defined naturally as the minimizer of $\lambda \mapsto \|x^* - x_\lambda\|$. With all the preliminary estimators x_m in $\ker(F_n)^\perp$, minimizing $\|x^* - x_\lambda\|$ reduces to minimizing the distance to $x^\dagger$. So, we consider the following loss function

$$\lambda \mapsto \gamma(x_\lambda) := \|z - x_\lambda\|^2.$$

This criterion corresponds to a quadratic loss between the image by the operator of a candidate function x and the observed data. 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 $\lambda \mapsto \gamma(x_\lambda)$ over $\mathbb{R}^M$. Defining the $n \times M$ matrix $\mathbf{X}$ by

$$\mathbf{X}_{i,m} = \langle x_m, \psi_i \rangle, \ i = 1, \dots n, \ m = 1, \dots, M$$

the minimizer of $\lambda \mapsto \gamma(x_\lambda)$ corresponds the ordinary least square solution

$$\hat{\lambda}_{OLS} = \arg\min_{\lambda \in \mathbb{R}^M} \ \|\mathbf{z} - \mathbf{X}\lambda\|_n^2,$$

where $\mathbf{z} = (z_1, ..., z_n)^\top \in \mathbb{R}^n$. However, this solution is known to be inefficient when the number of regressors is too large. For this reason, penalized procedures, favoring low-dimensional values of λ are often preferred to classical least square. For a given penalty $\text{pen}(\lambda)$, the penalized aggregation estimator $\hat{x} = x_{\hat{\lambda}}$ is built by minimizing over $\mathbb{R}^M$

$$\lambda \mapsto \gamma(x_\lambda) + \text{pen}(\lambda) \tag{5}$$

In order to promote sparsity, we use a ℓ^1-type penalty defined as

$$\text{pen}(\lambda) = \sum_{m=1}^{M} r_{n,m} |\lambda_m|, \tag{6}$$

with the notation $r_{n,m} = r_n \, \sigma_m$ with $r_n = 3\sqrt{2(\log M^2 n)/n}$ and $\sigma_m^2 = \frac{1}{n} \sum_{j=1}^n b_j^{-2} \langle x_m, \psi_j \rangle^2$. 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 $\mathcal{C}$, as it leads to a soft thresholding procedure (see for instance

[20] or [17]). 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 [14].

For any subset S of $\{1,\dots,M\}$ and for a given vector $\lambda \in \mathbb{R}^M$, we introduce the notation λ_S for the vector of size M whose components coincide with λ in S, and equal 0 otherwise. We denote by $\|\lambda\|$ the usual Euclidean norm on $\mathbb{R}^M$ and $|S|$ 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 the set $\Gamma_S = \{\lambda \in \mathbb{R}^M : \sum_{m\in S^c} \sigma_m|\lambda_m| \leqslant 5\sum_{m\in S} \sigma_m|\lambda_m|\}$. We the assume that

$$\phi(s) := \min_{S\subset\{1,\dots,M\}:\,|S|\leqslant s} \quad \min_{\lambda\neq 0:\ \lambda\in\Gamma_S} \frac{\lambda^\top \mathbf{X}^\top \mathbf{X}\lambda}{n\|\lambda_S\|} > 0.$$

This 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 [23, 11] 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 [23, 11]. 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 (5)-(6) in expectation :

Theorem 2.1 *Fix some integer* $1 \leqslant \bar{s} \leqslant M$*. Under the assumption Assump-*

tion RE($\bar{s}$), the penalized estimator $\hat{x} = x_{\hat{\lambda}} = \sum_{m=1}^{M} \hat{\lambda}_m x_m$, *with*

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

satisfies,

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

where $S = \{m \in \{1, \dots, M\} :\ \lambda_m \neq 0\}$ *and* $b_{max}^{-2} = \max_{j=1,\dots,n} b_j^{-2}$.

This theorem provides an oracle inequality that controls the aggregation procedure. The inequality is sharp in the sense that the leading constant in front of the main term is 1. Moreover, several quantities are of interest in the above bound. The main term is given by $\inf_\lambda \left\{ \|x_\lambda - x^*\|^2 + \frac{25}{36} r_n^2\, \phi^{-2}(s) \sum_{m \in S} \sigma_m^2 \right\}$. 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 $\sigma_m^2 = \sigma_m^2 = \frac{1}{n} \sum_{j=1}^{n} b_j^{-2} \langle x_m, \psi_j \rangle^2$ 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 λ^* such that $x_{\lambda^*} = x^*$, 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. For mildly ill-posed problems where the coefficients of the SVD decay at a polynomial rate $b_j = C j^{-t/2}$, this term is of order n^{t-1}. Note that it 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. Actually, the

presence in the oracle inequality of the maximal singular value b_{max}^{-2} (which in direct problems, corresponds to the spectral radius of the covariance operator) is currently a main issue in heteroscedastic aggregation. A similar term, involving the operator norm of the covariance operator, appears for instance in Theorem 1 in [12]. To this date, it seems that aggregation methods for inverse problems can only handle small degrees of ill-posedness in inverse problems.

Appendix

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

$$\mathcal{A} = \bigcap_{m=1}^{M} \{3|V_m| \leqslant r_{n,m}\}, \qquad \text{with } V_m = \frac{1}{n}\sum_{j=1}^{n} \mathbf{X}_{j,m}\varepsilon_j,$$

and which holds with large probability (*cf.* Appendix A for the definition of ε).

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

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

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

Proof of Proposition 2.2. The proof of this result is inspired by the proof of Theorem 2 in [16]. First of all, we notice that if $\left\langle \hat{x} - x^\dagger, \hat{x} - x_\lambda \right\rangle \leqslant 0$, then the identity

$$2\left\langle \hat{x} - x^\dagger, \hat{x} - x_\lambda \right\rangle = \|\hat{x} - x^\dagger\|^2 + \|\hat{x} - x_\lambda\|^2 - \|x_\lambda - x^\dagger\|^2 \tag{7}$$

implies $\|\hat{x} - x^\dagger\|^2 \leqslant \|x_\lambda - x^\dagger\|^2$. Then the bound in the proposition is valid. Then, let us consider the case $\left\langle \hat{x} - x^\dagger, \hat{x} - x_\lambda \right\rangle > 0$. Recall that we have set $r_n = 3\sqrt{2\frac{\log(M^2 n)}{n}}$ and $\sigma_m^2 = \frac{1}{n}\sum_{j=1}^{n} b_j^{-2}\mathbf{X}_{j,m}^2$, where $\mathbf{X}_{j,m} = \langle x_m, \psi_j\rangle$ for $j \in \{1,\dots,n\}$ and $m \in \{1,\dots,M\}$. In this case, we exploit the optimality

condition of the minimization criterion (5)-(6). Since $\hat{\lambda}$ is minimizer of this criterion, the first order optimality conditions imply that

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

where for any λ, the quantity $\partial|\lambda|_{1,\sigma}$ denotes the sub-differential of the weighted ℓ^1-norm, defined for any vector $a \in \mathbb{R}^M$ by $|a|_{1,\sigma} = \sum_{m=1}^M \sigma_m |a_m|$. Set $S = \{m :\ \lambda_m \neq 0\}$, the sparsity pattern of λ. Thanks to sub-differential of the ℓ^1-norm in $\mathbb{R}^M$, we deduce the set of sub-differential $\partial|\lambda|_{1,\sigma}$ of the above weighted ℓ^1-norm, with $\mu = (\mu_1, ..., \mu_M)^\top \in \partial|\lambda|_{1,\sigma}$ if and only if

$$\mu_m = \sigma_m \mathrm{sgn}(\lambda_m) \text{ if } m \in S \quad \text{and} \quad \mu_m \in [-\sigma_m, \sigma_m] \text{ if } m \in S^c,$$

where, for a given $\mathbf{a} \in \mathbb{R}$, $\mathrm{sgn}(\mathbf{a})$ equals ± 1 according to the sign of $\mathbf{a}$, and S^c denotes the complementary set of S in $\{1, \ldots, M\}$ (*cf.* [15, page 259] for details on sub-differential tools). Based on the above statement, we can write

$$2\frac{\hat{\lambda}^\top \mathbf{X}^\top \mathbf{z}}{n} - 2\frac{\hat{\lambda}^\top \mathbf{X}^\top \mathbf{X}\hat{\lambda}}{n} = r_n |\hat{\lambda}|_{1,\sigma} \tag{8}$$

where we recall $\mathbf{z} = (z_1, ..., z_n)^\top \in \mathbb{R}^n$ and for any $\lambda \in \mathbb{R}^M$ with sparsity pattern $S = \{m :\ \lambda_m \neq 0\}$,

$$2\frac{\lambda^\top \mathbf{X}^\top \mathbf{z}}{n} - 2\frac{\lambda^\top \mathbf{X}^\top \mathbf{X}\hat{\lambda}}{n} \leqslant r_n |\lambda|_{1,\sigma} \tag{9}$$

Subtracting (8) from (9), we get for any $\lambda \in \mathbb{R}^M$ with sparsity pattern $S = \{m :\ \lambda_m \neq 0\}$

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

Moreover, according to (4), the above inequality becomes

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

setting $\mathbf{x}^\dagger = (x_1^\dagger, ..., x_n^\dagger)^\top$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_n)^\top$. This inequality states that for any $\lambda \in \mathbb{R}^M$ with sparsity pattern $S = \{m :\ \lambda_m \neq 0\}$

$$2\left\langle \hat{x} - x^\dagger, \hat{x} - x_\lambda \right\rangle \leqslant r_n(|\lambda|_{1,\sigma} - |\hat{\lambda}|_{1,\sigma}) + 2\frac{(\hat{\lambda} - \lambda)^\top \mathbf{X}^\top \boldsymbol{\varepsilon}}{n}. \tag{10}$$

Considering the sparsity pattern of λ, the first term on the rhs of the above inequality can be decomposed as

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

where we set $r_{n,m} = r_n\,\sigma_m$. Let $V = \frac{1}{n}\mathbf{X}^\top\boldsymbol{\varepsilon}$ for short, then (10) becomes

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

Using (7), the above inequality gives us the fundamental results

$$\begin{aligned} \|\hat{x} - x^\dagger\|^2 + \|\hat{x} - x_\lambda\|^2 + \sum_{m\in S^c} r_{n,m}|\hat{\lambda}_m| \leqslant \|x_\lambda - x^\dagger\|^2 + \sum_{m\in S} r_{n,m}|\hat{\lambda}_m - \lambda_m| \\ + 2V^\top(\hat{\lambda} - \lambda). \end{aligned} \tag{11}$$

Once we established this last major inequality, we will first use it to show that $\hat{\lambda} - \lambda$ 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 $m \in S^c$, (11) combined with $\left\langle \hat{x} - x^\dagger, \hat{x} - x_\lambda \right\rangle > 0$ implies that

$$\sum_{m\in S^c} r_{n,m}|\hat{\lambda}_m| \leqslant \sum_{m\in S} r_{n,m}|\hat{\lambda}_m - \lambda_m| + 2\sum_{m\in S}|V_m||\hat{\lambda}_m - \lambda_m| + 2\sum_{m\in S^c}|V_m||\hat{\lambda}_m|$$

yielding

$$\sum_{m\in S^c}(r_{n,m} - |V_m|)|\hat{\lambda}_m| \leqslant \sum_{m\in S}(r_{n,m} + |V_m|)|\hat{\lambda}_m - \lambda_m|. \tag{12}$$

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 $\hat{\lambda} - \lambda$ 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

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

Combining this last inequality, with (11) 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

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

where we used, for the first inequality, similar reasoning as those exploited to get (12). 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 $\|\hat{x} - x_\lambda\|^2$ to both sides leads to the result in the proposition

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

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, $\gamma(\hat{x}) + \text{pen}(\hat{\lambda}) \leqslant \gamma(x_\lambda) + \text{pen}(\lambda)$ for all $\lambda \in \mathbb{R}^M$. Taking $\lambda = 0$, we deduce that $\gamma(\hat{x}) \leqslant 0$. Moreover, using the definition of γ we find

$$\begin{aligned}
\|\hat{x} - x^\dagger\|^2 &\leqslant \|x^\dagger\|^2 + 2|\langle \hat{x}, \varepsilon \rangle| \leqslant \|x^\dagger\|^2 + 2\|\hat{x}\|\|\varepsilon\| \\
&\leqslant \|x^\dagger\|^2 + 2\|\varepsilon\|\big(\|\hat{x} - x^\dagger\| + \|x^\dagger\|\big) \\
&\leqslant \|x^\dagger\|^2 + \frac{\|\hat{x} - x^\dagger\|^2}{2} + \|x^\dagger\|^2 + 3\|\varepsilon\|^2
\end{aligned}$$

using the inequality $2ab \leqslant \theta a^2 + \theta^{-1} b^2$ successively for $\theta = 1/2$ and $\theta = 1$. Now, we use that $\|\varepsilon\|^2 \leqslant n b_{\max}^{-2} \|\delta\|_n^2 = b_{\max}^{-2} W$, where $W := n\|\delta\|_n^2$ has $\chi^2(n)$

distribution and $b_{max}^{-2} = \max_{m=1,\dots,M} b_m^{-2}$. Thus,

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

As in 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 $\chi^2(n)$ distribution, it satisfies in particular $\mathbb{E}(W^2) \leqslant 4n^2$ and $\mathbb{P}(W > 2n) \leqslant \exp(-n/8)$ (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

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

yielding

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

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. J. Bickel, Y. Ritov, and A. B. 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. H. 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. Tsybakov. Oracle inequalities for inverse problems. *Ann. Statist.*, 30(3):843–874, 2000.

[10] L. Cavalier, M. Reiß, et al. Sparse model selection under heterogeneous noise: Exact penalisation and data-driven thresholding. *Electronic Journal of Statistics*, 8:432–455, 2014.

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

[12] Arnak S. Dalalyan and Joseph Salmon. Sharp oracle inequalities for aggregation of affine estimators. *Ann. Statist.*, 40(4):2327–2355, 2012.

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

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

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

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

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

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

[19] J-M. Loubes and C. Ludeña. Model selection for non linear inverse problems. *ESAIM PS*, 2:661–677, 2009.

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

[21] 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.

[22] P. Rochet. Adaptive hard-thresholding for linear inverse problems. *ESAIM: Probability and Statistics*, 17:485–499, 2013.

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

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

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