



**HAL**  
open science

# Sharp Support Recovery from Noisy Random Measurements by L1 minimization

Charles H Dossal, Marie-Line Chabanol, Gabriel Peyré, Jalal M. Fadili

► **To cite this version:**

Charles H Dossal, Marie-Line Chabanol, Gabriel Peyré, Jalal M. Fadili. Sharp Support Recovery from Noisy Random Measurements by L1 minimization. 2010. hal-00553670v1

**HAL Id: hal-00553670**

**<https://hal.science/hal-00553670v1>**

Preprint submitted on 7 Jan 2011 (v1), last revised 11 Sep 2011 (v2)

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Sharp Support Recovery from Noisy Random Measurements by $\ell_1$ minimization

Charles Dossal<sup>a</sup>, Marie-Line Chabanol<sup>b</sup>, Gabriel Peyré<sup>c</sup>, Jalal Fadili<sup>d</sup>

<sup>a</sup>IMB Université Bordeaux 1,  
351, cours de la Libération F-33405 Talence cedex, France

<sup>b</sup>IMB Université Bordeaux 1,  
351, cours de la Libération F-33405 Talence cedex, France

<sup>c</sup>CNRS and CEREMADE, Université Paris-Dauphine,  
Place du Maréchal De Lattre De Tassigny, 75775 Paris Cedex 16, France

<sup>d</sup>GREYC, CNRS-ENSICAEN-Université Caen,  
6 Bd du Maréchal Juin 14050 Caen Cedex, France

---

## Abstract

In this paper, we investigate the theoretical guarantees of penalized  $\ell^1$  minimization (also called Basis Pursuit Denoising or Lasso) in terms of sparsity pattern recovery (support and sign consistency) from noisy measurements with non-necessarily random noise, when the sensing operator belongs to the Gaussian ensemble (i.e. random design matrix with i.i.d. Gaussian entries). More precisely, we derive sharp non-asymptotic bounds on the sparsity level and (minimal) signal-to-noise ratio that ensure support identification for most signals and most Gaussian sensing matrices by solving the Lasso problem with an appropriately chosen regularization parameter.

Our first purpose is to establish conditions allowing exact sparsity pattern recovery when the signal is strictly sparse. Then, these conditions are extended to cover the compressible or nearly sparse case. In these two results, the role of the minimal signal-to-noise ratio is crucial. Our third main result gets rid of this assumption in the strictly sparse case, but this time, the Lasso allows only partial recovery of the support. We also provide in this case a sharp  $\ell_2$ -consistency result on the coefficient vector.

The results of the present work have several distinctive features compared to previous ones. One of them is that the leading constants involved in all the bounds are sharp and explicit. This is illustrated by some numerical experiments where it is indeed shown that the sharp sparsity level threshold identified by our theoretical results below which sparsistency of the Lasso is guaranteed meets that empirically observed. <sup>1</sup>

*Key words:* Compressed sensing,  $\ell^1$  minimization, sparsistency, consistency.

---

## 1. Introduction

### 1.1. Problem setup

The conventional wisdom in digital signal processing is the Shannon sampling theorem valid for bandlimited signals. However, such a sampling scheme excludes many signals

---

*Email addresses:* [charles.dossal@math.u-bordeaux1.fr](mailto:charles.dossal@math.u-bordeaux1.fr) (Charles Dossal),  
[Marie-Line.Chabanol@math.u-bordeaux1.fr](mailto:Marie-Line.Chabanol@math.u-bordeaux1.fr) (Marie-Line Chabanol),  
[gabriel.peyre@ceremade.dauphine.fr](mailto:gabriel.peyre@ceremade.dauphine.fr) (Gabriel Peyré), [jalal.fadili@greyc.ensicaen.fr](mailto:jalal.fadili@greyc.ensicaen.fr) (Jalal Fadili)

<sup>1</sup>This work is supported by ANR grant NatImages ANR-08-EMER-009.

of interest that are not necessarily bandlimited but can still be explained either exactly or accurately by a small number of degrees of freedom. Such signals are termed sparse signals.

In fact we distinguish two types of sparsity: strict and weak sparsity (the latter is also termed compressibility). A signal  $x$ , considered as a vector in a finite dimensional subspace of  $\mathbb{R}^p$ , is strictly or exactly sparse if most of its entries vanish; i.e. if its support  $I(x) = \text{supp}(x) = \{1 \leq i \leq p \mid x(i) \neq 0\}$  is of cardinality  $k \ll p$ . A  $k$ -sparse signal is a signal where exactly  $k$  samples have a non-zero value. Signals and images of practical interest may be *compressible* or *weakly sparse* in the sense that the sorted magnitudes  $|x^{\text{sorted}}(i)|$  decay quickly. Thus  $x$  can be well-approximated as  $k$ -sparse up to an error term (this property will be used when we will tackle compressible signals). If a signal is not sparse in its original domain, it may be *sparsified* in an appropriate orthobasis  $\Phi$  (hence the importance of the point of view of computational harmonic analysis and approximation theory). Without loss of generality, we assume throughout that  $\Phi$  is the standard basis.

The compressed/sensing/sampling [1, 2, 3] asserts that sparse or compressible signals can be reconstructed with theoretical guarantees from far few measurements than the ambient dimension of the signal. Furthermore, the reconstruction is stable if the measurements are corrupted by an additive bounded noise. The encoding (or sampling) step is very fast since it gathers  $n$  non-adaptive linear measurements that preserve the structure of the signal  $x_0$ :

$$y = Ax_0 + w \in \mathbb{R}^n, \quad (1)$$

where  $A \in \mathbb{R}^{n \times p}$  is a rectangular measurement matrix, i.e.  $n < p$ , and  $w$  accounts for possible noise with bounded  $\ell_2$  norm. In this work, there is no need that  $w$  is random and we consider that  $A$  is drawn from the Gaussian matrix ensemble<sup>2</sup>, i.e. the entries of  $A$  are independent and identically distributed (i.i.d.)  $\mathcal{N}(0, 1/n)$ . The columns of  $A$  are denoted  $a_i$ , for  $i = 1, \dots, p$ . In the sequel, the sub-matrix  $A_I$  is the restriction of  $A$  to the columns indexed by  $I(x_0)$ . The latter is denoted  $I$  unless understood otherwise from the context.

The signal is reconstructed from this underdetermined system of linear equations by solving a convex program of the form:

$$x(\gamma) \in \underset{x \in \mathbb{R}^p}{\text{argmin}} \quad \|x\|_1 \quad \text{such that } Ax - y \in \mathcal{C}, \quad (2)$$

where  $\mathcal{C}$  is a closed convex set, and  $\|x\|_p := (\sum_i |x(i)|^p)^{1/p}$ ,  $p \geq 1$  is the  $\ell_p$ -norm of a vector with the usual adaptation for  $p = \infty$ :  $\|x\|_\infty = \max_i |x(i)|$ . We also denote  $\|x\|_0$  as the  $\ell_0$  pseudo-norm which counts the number of non-zero entries of  $x$ . Obviously,  $\|x\|_0 = |I(x)|$ . For all  $x \in \mathbb{R}^p$ , the notation  $\bar{x} \in \mathbb{R}^{|I(x)|}$  means the restriction of  $x$  to its support  $I(x)$ .

Typically, if  $\mathcal{C} = \{0\}$  (no noise), we end up with the so-called Basis Pursuit [4] problem

$$\min_{x \in \mathbb{R}^p} \|x\|_1 \quad \text{such that } y = Ax. \quad (\text{BP})$$

Taking  $\mathcal{C}$  as the  $\ell_2$  ball of radius  $\epsilon$ , we have a noise-aware variant of BP

$$\min_{x \in \mathbb{R}^p} \|x\|_1 \quad \text{such that } \|Ax - y\|_2 \leq \epsilon \quad (\ell_1\text{-constrained})$$

where the parameter  $\epsilon > 0$  depends on the noise level  $\|w\|_2$ . This constrained form can also be shown to be equivalent to the penalized form, which goes by the name of Basis

---

<sup>2</sup>In a statistical linear regression setting, we would speak of a random Gaussian design.

Pursuit DeNoising [4] or Lasso in the statistics community after [5]:

$$\min_{x \in \mathbb{R}^p} \frac{1}{2} \|y - Ax\|_2^2 + \gamma \|x\|_1, \quad (\text{Lasso})$$

where  $\gamma$  is the regularization parameter. ( $\ell_1$ -constrained) and (Lasso) are equivalent in the sense that there is a bijection between  $\gamma$  and  $\epsilon$  such that both problems share the same set of solutions. However, this bijection is unknown explicitly and depends on  $y$  and  $A$ , so that in practice, one needs to use different algorithms to solve each problem, and theoretical results are stated using one formulation or the other. In this paper, we focus on the Lasso formulation. It is worth noting that the Dantzig selector [6, 7] is also a special instance of (2) when  $\mathcal{C} = \{z \in \mathbb{R}^p \mid \|A^T z\|_\infty \leq \gamma\}$ .

The convex problems of the form ( $\ell_1$ -constrained) and (Lasso) are computationally tractable and many algorithms have been developed to solve them, and we only mention here a few representatives. Homotopy continuation algorithms [8, 9, 10] track the whole regularization path. Many first-order algorithms originating from convex non-smooth optimization theory have been proposed to solve (Lasso). These include one-step iterative thresholding algorithms [11, 12, 13, 14], or accelerated variants [15, 16], multi-step schemes such as [17] or [18]. The Douglas-Rachford algorithm [19, 20] is a first order scheme that can be used to solve ( $\ell_1$ -constrained). A more comprehensive account can be found in [21, Chapter 7].

### 1.2. Theoretical performance measures of the Lasso

These last years, we have witnessed a flurry of research activity where efforts have been made to investigate the theoretical guarantees of the Lasso as a sparse recovery procedure from noisy measurements in the underdetermined case  $n < p$ . Overall, the derived conditions hinge on strong assumptions on the structure and interaction between the variables in  $A$  as indexed by  $x_0$ . An overview of the literature pertaining to our work will be covered in Section 1.3 after notions are introduced so that the discussions are clearer.

Let  $x_0$  be the original vector as defined in (1),  $f_0 = Ax_0$  the noiseless measurements,  $x(\gamma)$  a minimizer of the Lasso and  $f(\gamma) = Ax(\gamma)$ .

*Consistency.*  $\ell_p$ -consistency on the signal  $x$  means that the  $\ell_p$ -error  $\|x_0 - x(\gamma)\|_p$ , for typically  $p = 1, 2$  or  $\infty$ , between the unknown vector  $x_0$  and a solution  $x(\gamma)$  of either (Lasso) or ( $\ell_1$ -constrained) comes within a factor of the noise level.

*Sparsistency.* Sparsity pattern recovery (also dubbed sparsistency for short or variable selection in the statistical language) requires that the indices and signs of the solutions  $x(\gamma)$  are equal to those of  $x_0$  for a well chosen value of  $\gamma$ . Partial support recovery occurs when the recovered support is included (strictly) in that of  $x_0$  with the correct sign pattern.

In general, it is not clear which of these performance measures is better to characterize the Lasso. Nevertheless, in the noisy case, consistency does not tell the whole story and there are many applications where bounds on the  $\ell_p$ -error are insufficient to characterize the accuracy of the Lasso estimates. In this case, exact or partial recovery of the support, hence of the correct model variables, is the desirable property to have. Among other advantages, this allows for instance to circumvent the bias of the Lasso and thus enhance the estimation of  $x_0$  and  $Ax_0$  using a debiasing procedure: recover the support  $I$  by solving the Lasso, followed by least-squares regression on the selected variables  $(a_i)_{i \in I}$ ; see e.g. [6, 22]. Our work falls within this scope and focuses on exact and partial support

identification for both strictly sparse and compressible signals in the presence of noise on Gaussian random measurements.

### 1.3. Literature overview

The properties of the Lasso have been extensively studied, including consistency and distribution of the Lasso estimates. There is of course a huge literature on the subject, and covering it fairly is beyond the scope of this paper. In this section, we restrict our overview to those works pertaining to ours, i.e. sparsity pattern recovery in presence of noise.

Much recent work aims at understanding the Lasso estimates from the point of view of sparsistency. This body of work includes [22, 6, 23, 24, 25, 26, 27, 28, 29]. For the Lasso estimates to be close to the model selection estimates when the data dimensions  $(n, p)$  grow, all the aforementioned papers assumed a sparse model and used various conditions that require the irrelevant variables to be not too correlated with the relevant ones.

*Mutual coherence-based conditions.* Several researchers have studied independently the qualitative performance of the Lasso for either exact or sparsity pattern recovery of sufficiently sparse signals under a mutual coherence condition on the measurement matrix  $A$ ; see for instance [23, 30, 26, 31] when  $A$  is deterministic, and [32] when  $A$  is Gaussian. However, mutual coherence is known to lead to overly pessimistic sparsity bounds.

*Support structure-based conditions.* These sufficient recovery conditions were refined by considering not only the cardinality of the support but also its structure, including the signs of the non-zero elements of  $x_0$ . Such criteria use the interactions between the relevant columns of  $A_I = (a_i)_{i \in I}$  and the irrelevant ones  $(a_i)_{i \notin I}$ . More precisely, we define the following condition developed in [33] to analyse the properties of the Lasso. This condition goes by the name of irrepresentable condition in the statistical literature; see e.g. [28, 22, 27, 34] and [35] for a detailed review.

**Definition 1.** *Let  $I$  be the support of  $x_0$  and  $I^c$  its complement in  $\{1, \dots, p\}$ . The irrepresentable (or Fuchs) condition is fulfilled if*

$$F(x_0) := \|A_{I^c}^T A_I (A_I^T A_I)^{-1} \text{sign}(\overline{x_0})\|_\infty = \max_{i \in I^c} |\langle a_i, d(x_0) \rangle| < 1 \quad (3)$$

$$\text{where } d(x_0) = d_0 = A_I (A_I^T A_I)^{-1} \text{sign}(\overline{x_0}) . \quad (4)$$

Condition (3) will also be the soul of our analysis in this paper.

The criterion (3) is closely related to the exact recovery coefficient (ERC) of Tropp [26]:

$$\text{ERC}(x_0) := 1 - \max_{i \in I^c} \|(A_I^T A_I)^{-1} A_I^T a_i\|_1 . \quad (5)$$

In [26, Corollary 13], it is established that if  $\text{ERC}(x_0) > 0$ , then the support of the Lasso solution of the Lasso with a large enough parameter  $\gamma$  is included in the one of the subset selection  $(\ell_0)$  optimal solution.

In [28], an asymptotic result is reported showing that under (3)<sup>3</sup>, the Lasso selects exactly the set of nonzero coefficients, provided that these coefficients are bounded away

---

<sup>3</sup>In fact, a slightly stronger assumption requiring that all elements in (3) are uniformly bounded away from 1.

from zero at a certain rate, which is sufficient to guarantee exact support recovery and sign consistency in some asymptotic regime. There, it is also shown that (3) is sufficient and essentially necessary for variable selection. The reference [24] develops very similar results and uses similar requirements. [36] and [37] derive asymptotic conditions for sparsistency of the block Lasso [38] by extending (3) and (5) to the group setting.

Reference [22] proposes a non-asymptotic analysis with a sufficient condition ensuring exact support and sign pattern recovery of most sufficiently sparse vectors for matrices satisfying a weak coherence condition (of the order  $(\log p)^{-1}$ ). Their proof relies upon (3) and a bound on norms of random submatrices developed in [39]. The work in [27] considers a condition of the form (3) to ensure sparsity pattern recovery from noisy measurements by solving the Lasso. The analysis in that paper was conducted for both deterministic and standard Gaussian  $A$  in a high-dimensional setting where  $p$  and the sparsity level grow with the number of measurements  $n$ . That author also established that violation of (3) is sufficient for failure of the Lasso in recovering the support set. In [40], the sufficient bound on the number of measurements established in [27] for the standard Gaussian dense ensemble was shown to hold for sparse measurement ensembles. The works of [22] and [27] are certainly the most closely related to ours. We will elaborate more on these connections by highlighting the similarities and differences in Section 2.2.

*Variations on the Lasso.* Other variations of the Lasso have been proposed and their properties investigated from the sparsistency standpoint. For instance, [29] and [41] studied the adaptive<sup>4</sup> Lasso and developed asymptotic results similar to those just above but require either a good initial estimator or a level of coherence on the pessimistic order of  $1/\sqrt{n}$ . In [43], it is shown that a multi-step thresholding procedure can accurately estimate a sparse vector under the restricted eigenvalue condition<sup>5</sup> [7]. The two-stage procedure in [45] applies selective penalization in the second stage and is studied assuming incoherence conditions. A more general framework for multi-stage variable selection was studied by [46] by controlling the probability of false positives at the price of false negatives. [47] quantified the variable selection property of thresholded Lasso using prediction error (closely tied to false negative selections) together with its number of false positive selections, without requiring the irrepresentable or incoherence conditions on the measurement matrix, nor assumptions about the minimal signal-to-noise ratio (SNR). [34] examined the variable selection property and sign consistency of the Lasso followed by hard thresholding (a form of adaptive Lasso), when all non-zero components are large enough. These authors assume that the singular values of  $A_I$  for all  $I$  such that  $|I| = \|x_0\|_0$  are bounded away from zero. They argued that if the irrepresentable condition is relaxed, the Lasso cannot recover the correct sparsity pattern, but under their assumption, they show that the estimator is still consistent in the  $\ell_2$ -norm sense. In addition, they show it is possible to achieve sparsistency. For an overview of other penalized methods that have been proposed for the purpose of variable selection, see [48].

*Information-theoretic bounds.* A recent line of research has developed information-theoretic sufficient and necessary bounds to characterize fundamental limits on *minimal* SNR, the number of measurements  $n$ , and tolerable sparsity level  $k$  required for exact or partial

---

<sup>4</sup>The adaptive Lasso as seen in the statistical literature turns out to be a two-step procedure, where the second step is to solve a reweighted  $\ell_1$  norm problem, with weights given by the Lasso estimate in the first step. In fact, this is a special case of the iterative reweighted  $\ell_1$  minimization [42].

<sup>5</sup>This is a refinement of the well-known restricted isometry property [44], see the review [35] for a thorough discussion on these relationships.

support pattern recovery of exactly sparse signals by any algorithm including the optimal exhaustive  $\ell_0$  decoder [49, 50, 51, 52, 53, 54, 55, 56]. In most of these works, the bounds are asymptotic, i.e. they provide asymptotic scaling and typically require that the sparsity level  $k$  varies at some rate (linearly or sub-linearly) with the signal dimension  $p$  when  $n$  grows to infinity. It is worth mentioning that a careful normalization is needed, for instance of the sampling matrix and noise, when comparing these results in the literature.

The paper [49] was the first to consider the information-theoretic limits of exact sparsity recovery from the Gaussian measurement ensemble, explicitly identifying the minimal SNR (or equivalently  $T = \min_{i \in I(x_0)} |x_0(i)|$ ) as a key parameter. This analysis yielded necessary and sufficient conditions on the tuples  $(n, p, k, T)$  for asymptotically reliable sparsity recovery. This complements the analysis of [27] by showing that in the sub-linear sparsity regime, i.e.  $k = o(p)$ , the number of measurements required by the Lasso  $n \gtrsim k \log(p - k)$ <sup>6</sup> achieves the information-theoretic necessary bound. Subsequent work of [51, 52, 56, 57] has extended this type of analysis to partial support recovery using different support distortion measures. [58] extended some of the results by [51] for the Gaussian ensemble to sub-Gaussian and other random measurement ensembles. [53] consider only exact support recovery, and also provides results for general (non-Gaussian) dense measurement ensembles. The necessary bounds of [49] were strengthened in [52] when  $k$  scales linearly with  $n$  (linear sparsity regime), and in [50] for all scalings. The last authors also give a sufficient condition for sparsistency of a simple maximum correlation procedure that scales similarly but turns out to be conservative, as it does never hold asymptotically, see the counter-example in [55].

[59] found a non-asymptotic upper bound on the probability that the  $\ell_0$  decoder of  $x$  declares a wrong sparsity pattern, given any generic measurement matrix  $A$ . When  $A$  is drawn from the Gaussian ensemble, they obtain asymptotically sufficient conditions on the number of measurements and the minimal SNR for exact recovery, which agree with the known necessary conditions previously established.

In [54], the authors derive sufficient and necessary conditions for exact support recovery by the  $\ell_0$  optimal decoder, with additive white Gaussian noise in either the measurements or the signal. For noise on the observations, they show that asymptotically a number of measurements  $\gtrsim k \log(p/k)$  in conjunction with a noise variance of the order  $(\log p)^{-1}$  are necessary and sufficient for exact support recovery. Furthermore, if a small fraction of support errors can be tolerated, a constant variance turns out to be sufficient in the linear sparsity regime. They argue that their analysis avoids using union bounds that are generally conservative, hence leading to tighter bounds on the number of measurements compared to previously developed.

#### 1.4. Organization of the paper

The rest of the paper is organized as follows. We first state our main results and discuss the connections and novelties with respect to existing work. In Section 3 and 4, we detail the proofs for exact recovery with strictly sparse and compressible signals, before proving the partial support recovery result in Section 5. Numerical experiments are carried out in Section 6. Section 6 includes a final discussion and some concluding remarks.

---

<sup>6</sup>The shorthand notation  $f \gtrsim g$  means that  $g = O(f)$ .

## 2. Contributions

Most of the results developed in the literature on sparsistency of the Lasso exhibit asymptotic scaling results in terms of the triple  $(n, p, k)$ , but this does not tell the whole story. Often, one needs to know the exact constants involved in the bounds, not only their dependence on key quantities such as the SNR and/or other parameters of the signal  $x_0$ , but also the constants appearing in the bounds. As a consequence, the majority of sufficient conditions are more conservative than those suggested by empirical evidence.

In this paper, we investigate the theoretical guarantees of the Lasso in terms of sparsity pattern recovery (support and sign consistency) from noisy measurements –the noise being not necessarily random– when the measurement matrix belongs to the Gaussian ensemble. More precisely, we provide precise *non-asymptotic* bounds, including explicit sharp leading numerical constants, on the key quantities that come into play (sparsity level for a given measurement budget, minimal SNR, regularization parameter) to ensure exact or partial sparsity pattern recovery using the Lasso for both strictly sparse and compressible signals. Our results have several distinctive features compared to previous closely-connected works. This will be discussed in further details in Section 2.2. Numerical evidence reported in Section 6 confirm that the sharp sparsity level threshold identified by our theoretical results below which sparsistency of the Lasso is guaranteed meets that empirically observed.

### 2.1. Statement of main results

Our first result Theorem 1 establishes conditions allowing exact sparsity pattern recovery when the signal is strictly sparse. Then, these conditions are extended to cover the compressible case in Theorem 2. In these two results, the role of the minimal SNR is crucial. Our third main result in Theorem 3 gets rid of this assumption in the strictly sparse case, but this time, the Lasso allows only partial recovery of the support. We also provide in this case a sharp  $\ell_2$ -consistency result on the Lasso estimate.

The three theorems are stated following the same structure: suppose that  $(x_0, w)$  fulfill some requirements formalized by a set  $\mathcal{Y}$ , then with overwhelming probability (w.o.p. for short) on the choice of  $A$ , the Lasso estimate obeys some property  $\mathcal{P}$ . It should be noted that these theorems imply in particular that w.o.p. on the choice of  $A$ , for *most* vectors  $(x_0, w) \in \mathcal{Y}$ , the Lasso estimate satisfies property  $\mathcal{P}$ , whatever the probability measure used on the set  $\mathcal{Y}$ .

The proof of Theorem 1 is given in Section 3. We prove its extension to compressible signals as stated in Theorem 2 in Section 4. Both proofs capitalize on an implicit formula of the Lasso solution. The proof of Theorem 3 given in Section 5 is quite different, since no such implicit formula is used.

#### 2.1.1. Exact Support Recovery with Strictly Sparse Signals

Theorem 1 shows that for most Gaussian matrices and most sparse vectors  $x_0$  and noise vectors  $w$  with bounded  $\ell_2$ -norm, the Lasso estimate recovers exactly the sign and the support of  $x_0$  for a suitable choice of  $\gamma$ .

**Theorem 1.** *Let  $A \in \mathbb{R}^{n \times p}$  be a Gaussian matrix, i.e. its entries are i.i.d.  $\mathcal{N}(0, 1/n)$ ,  $w \in \mathbb{R}^n$  is such that  $\|w\|_2 \leq \varepsilon$ ,  $0 \leq \alpha, \beta < 1$  and  $p > e^{\frac{1}{2(1-\sqrt{\beta})}}$ . Suppose that  $x_0 \in \mathbb{R}^p$  obeys*

$$\|x_0\|_0 = k \leq \frac{\alpha\beta n}{2 \log p} \quad (6)$$

and

$$\min_{i \in I} |x_0(i)| = T \geq \frac{5.5\varepsilon}{\sqrt{1-\alpha}} \sqrt{\frac{2 \log p}{n}}. \quad (7)$$



Solve the Lasso problem from the measurements  $y = Ax_0 + w$ . Then with a probability  $P(n, p, \alpha, \beta)$  that tends to 1 as  $n$  grows to infinity, the Lasso solution  $x(\gamma)$  with

$$\gamma = \frac{\varepsilon}{\sqrt{1-\alpha}} \sqrt{\frac{2 \log p}{n}} \quad (8)$$

is unique and satisfies

$$\text{supp}(x(\gamma)) = \text{supp}(x_0) \quad \text{and} \quad \text{sign}(x(\gamma)) = \text{sign}(x_0) .$$

The proof (see Section 3) shows that the probability  $P(n, p, \alpha, \beta)$  is larger than

$$1 - \frac{1}{2} e^{-0.7\sqrt{\log n}} - \frac{1}{2\sqrt{\pi \log p}} - o\left(\frac{1}{\log p}\right) - o(e^{-0.7\sqrt{\log n}}) .$$

Although this bound on the probability is far from optimal.

In plain words, Theorem 1 asserts that for  $(\alpha, \beta) \in [0, 1)$  the support and the sign of most vectors obeying (6) can be recovered using Lasso if the non-zero coefficients of  $x_0$  are large enough compared to noise. This bound on the sparsity of  $x_0$  turns out to be optimal, since for any  $c > 1$ , for most vectors such that  $\|x_0\|_0 \geq \frac{cn}{2 \log p}$ , the support cannot be recovered using a Lasso estimate even with no noise. Indeed, [33] and [60] proved that the Lasso estimate for any  $\gamma$  shares the same sign and the same support as  $x_0$  when  $y = Ax_0$  if and only if

$$\max_{j \notin I} |\langle a_j, A_I(A_I^T A_I)^{-1} \text{sign}(x_0) \rangle| \leq 1 .$$

(Note the difference with the strict inequality in (3)). Hence if  $\|x_0\|_0 \geq \frac{cn}{2 \log p}$  with  $c > 1$ , then w.o.p.  $\|A_I(A_I^T A_I)^{-1} \text{sign}(x_0)\|_2^2 \geq \frac{Cn}{2 \log p}$  for some  $C > 1$  and sufficiently large  $p$ . As a result,  $\max_{j \notin I} |\langle a_j, A_I(A_I^T A_I)^{-1} \text{sign}(x_0) \rangle| \geq \sqrt{C} > 1$ . This optimality discussion is consistent with the information-theoretic bounds of [49], where it was proved that the number of measurements required by the Lasso achieves the (asymptotic) information-theoretic necessary bound that has the scaling (6) when the sparsity regime is sub-linear and  $T^2 \sim 1/\|x_0\|_0$ .

An important feature of Theorem 1 is that all the constants are made explicit and are governed by the two numerical constants  $\alpha$  and  $\beta$ . The role of  $\alpha$  is very instructive since when lowering  $\gamma$  by decreasing  $\alpha$ , the threshold on the minimal SNR is decreased to allow smaller coefficients to be recovered, but simultaneously the probability of success gets lower and the number of measurements required to recover the  $k$ -sparse signal increases. The converse applies when  $\alpha$  is increased. On the other hand, increasing  $\beta$  (in an appropriate range; see Section 3.3 for details) allows a higher threshold on the sparsity level, but again at the price of a smaller probability of success.

### 2.1.2. Support Recovery with Compressible Signals

Theorem 1 can be easily extended to signals that are not exactly sparse in the sense that their energy is concentrated on a few large coefficients. We consider the best  $k$ -term approximation  $x^k$  of  $x_0$  obtained by keeping only the  $k$  largest entries from  $x_0$  and setting the other to zero. This is equivalently defined using a thresholding

$$x^k(i) = \begin{cases} x_0(i) & \text{if } |x_0(i)| \geq T, \\ 0 & \text{otherwise,} \end{cases} \quad \text{and} \quad k = |I(x^k)|. \quad (9)$$

A signal is generally considered as compressible if the residual  $x^k - x_0$  is small. For sparsistency to make sense in this compressible case, stronger constraints are required, namely that the largest components  $x^k$  of the signal are significantly larger than the residual  $x^k - x_0$ . This is detailed in the following theorem.

**Theorem 2.** *Let  $A$ ,  $\alpha$ ,  $\beta$  and  $p$  as in Theorem 1. We measure  $y = Ax_0 + w$ , and let  $x^k$  be the best  $k$ -term approximation of  $x_0$  where  $k$  satisfies (6). Suppose furthermore that*

$$\|w\|_2 + 2\|x_0 - x^k\|_2 \leq \varepsilon$$

and  $T$  as defined in (9) is such that

$$T \geq 5.5 \sqrt{\frac{2 \log p}{n(1-\alpha)}} \varepsilon \quad (10)$$

$$\|x_0 - x^k\|_\infty \leq \frac{4(1-\sqrt{\alpha})}{5} \sqrt{\frac{2 \log p}{n(1+2\sqrt{\alpha}-3\alpha)}} \varepsilon. \quad (11)$$

Then, with a probability  $P(n, p, \alpha, \beta)$  that tends to 1 as  $n$  grows to infinity, the solution  $x(\gamma)$  of the Lasso with measurements  $y$  with

$$\gamma = 2 \sqrt{\frac{2 \log p}{n(1+2\sqrt{\alpha}-3\alpha)}} \varepsilon$$

is unique and satisfies

$$\text{supp}(x(\gamma)) = \text{supp}(x^k) \quad \text{and} \quad \text{sign}(x(\gamma)) = \text{sign}(x^k).$$

Again, all the leading constants are explicit. Conditions (10) and (11) impose compressibility constraints on the signal, namely that the magnitude  $T$  of the largest components are well above the average magnitude  $\varepsilon/\sqrt{n}$  of the residual, and that the residual is “flat” because the ratio of its  $\ell^\infty$  and  $\ell^2$  norms should be small.

Theorem 2 encompasses the strictly sparse result, Theorem 1, which is easily recovered by letting  $x_0 = x^k$ . The parameter  $\alpha$  plays a similar role in both theorems. Furthermore, in Theorem 2, the Lasso becomes more tolerant to compressibility errors  $x_0 - x^k$  as  $\alpha$  decreases. This however comes at the price of a lower probability of success as indicated in our proof.

### 2.1.3. Partial Support Recovery with Strictly Sparse Signals

In both previous theorems, the assumption on  $T$  plays a crucial role: if  $T$  is too small, there is no way to distinguish the small components of  $x_0$  from the noise; see also the discussion and literature review in Section 1.3. Nevertheless, if no assumptions are made on  $T$ , one can nevertheless expect to partly recover the support of  $x_0$ . This is formalized in the following result.

**Theorem 3.** *Let  $A$ ,  $\alpha$  and  $\beta$  as in Theorem 1. Suppose that we measure  $y = Ax_0 + w$ , where  $x_0$  fulfills (6). Solve the Lasso problem from the measurements  $y = Ax_0 + w$ . Then with a probability  $P_2(n, p, \alpha, \beta)$  which tends to 1 as  $n$  goes to infinity, the Lasso solution  $x(\gamma)$  with  $\gamma = \frac{\varepsilon}{\sqrt{1-\alpha}} \sqrt{\frac{2 \log p}{n}}$  is unique and satisfies  $\text{supp}(x(\gamma)) \subset \text{supp}(x_0)$ . Moreover, the Lasso estimate is  $\ell_2$ -consistent*

$$\|x_0 - x(\gamma)\|_2 \leq \varepsilon \left( \sqrt{\frac{\alpha}{1-\alpha}} + 2 \right).$$

If  $\gamma$  is large enough it is clear that  $\text{supp}(x(\gamma)) \subset \text{supp}(x_0)$  since for  $\gamma \geq \|A^T y\|_\infty$ ,  $x(\gamma) = 0$ . Theorem 3 provides a parameter  $\gamma$  proportional to  $\varepsilon$  that ensures a partial support recovery without any assumption on  $T$ . The theorem also gives a sharp upper bound on  $\ell_2$ -error of the Lasso estimate. This  $\ell_2$ -consistency remains valid under the additional hypotheses of Theorem 1 or 2 allowing exact recovery of the support.

## 2.2. Connections to related works

*Sparsistency.* As we mentioned in Section 1.3, our work is closely related to [22, 27], but is different in many important ways that we summarize as follows.

- **Deterministic vs random measurement matrices:** the work of [22] considers deterministic matrices satisfying a weak incoherence condition. Our work focuses on the classical Gaussian ensemble.
- **Asymptotic vs non-asymptotic analysis:** the analysis in [27] applies to high-dimensional setting where even the sparsity level  $k$  grows with the number of measurements  $n$ . As a result,  $k$  appears in the statements of the probabilities, which thus requires that  $k \rightarrow +\infty$ . This is very different from our setting as well as that of [22] where the probabilities depend solely on the dimensions of  $A$ . We believe that this is more natural in many applications.
- **Random vs deterministic noise:** in both previous works, the noise is stochastic (Gaussian in [22] and sub-Gaussian in [27]). In our work, we handle any noise with a finite  $\ell_2$ -norm.
- **Leading numerical constants:** these are not always explicit and sharp in those works. The constant involved in the sparsity level upper-bound in [22, Theorem 1.3] is not given, whereas (6) gives an explicit and sharp bound. The bound (7) and (8) on  $T$  and  $\gamma$  are similar to those given in [22, Theorem 1.3] once specialized for  $\alpha = 3/4$ . In [27, Theorem 2], the constant appearing in the lower-bound on  $T$  is not given, whereas (7) provides an explicit expression that is shown to be reasonably good in Section 6.
- **Compressible signals:** to the best of our knowledge, the compressible case has not been covered in the literature, and Theorem 2 appears then as a distinctly novel result of this paper.
- **$\ell_2$ -consistency:** such a result is not given in those references. A bound on the  $\ell_2$ -prediction error on  $Ax_0 - Ax(\gamma)$  is proved in [22]. An  $\ell_\infty$ -consistency is established in [27], which is an immediate consequence of sparsistency. Our method of proof differs significantly from the one used in [27], and in particular it naturally leads to a result of  $\ell_2$ -consistency.
- **Exact and partial support recovery:** in [22] the partial recovery case was not considered. In [27], exact and partial recovery are somewhat handled simultaneously, while we give two distinct results for each case.

*$\ell_2$ -consistency.* This property of the Lasso estimate has been widely studied by many authors under various sufficient conditions. Theorem 3 may then be compared to this literature, and we here focus on results based on the restricted isometry property (RIP) [44] and more or less similar variants in the literature; see the discussion in [34] and the review in [35].

The RIP results are uniform and ensure  $\ell_2$ -stability of the Lasso for *all* sufficiently sparse vectors from noisy measurements, whereas Theorem 3 guarantees that the Lasso estimate is  $\ell_2$ -consistent for *most* sparse vectors and a given matrix. When  $A$  is Gaussian, the scaling of the sparsity bound is  $O(n/\log(p/n))$  for RIP-based results which is better than  $O(n/\log p)$  in Theorem 3. Note that the scaling  $O(n)$  was derived in [61] when  $A$  belongs to the uniform spherical ensemble to ensure  $\ell_2$ -stability of the Lasso for most matrices  $A$ , although the leading constants are not given explicitly. However, the RIP is a worst-case analysis, and the price is that the leading constants in the sufficient sparsity bounds are overly small. In contrast, the leading numerical constants in our sparsity and  $\ell_2$ -consistency upper-bounds are explicit and solely controlled by  $\alpha$  and  $\beta \in [0, 1)$ . For instance, it can be verified from our proof that the value of the sparsity upper-bound we provide is actually larger than the bounds obtained from the RIP for  $p$  up to  $e^{100}$ . Finally, the RIP is a deterministic property that turns out to be satisfied by many ensembles of random matrices other than the Gaussian. Our Theorem 3 could probably be extended to sub-Gaussian matrices (e.g. using [62, Corollary V.2.1]), but this remains an open question.

### 3. Proof of Support Identification of Exactly Sparse Signals

This section gives the proof of Theorem 1. In the following, for any vector  $x \in \mathbb{R}^p$ , we denote as  $\bar{x} \in \mathbb{R}^{|I(x)|}$  the restriction of  $x$  to its non zero entries  $I(x)$ . We also denote  $A_I = (a_i)_{i \in I} \in \mathbb{R}^{n \times |I|}$  the matrix obtained from  $A$  by selecting the columns indexed by  $I$ , where we dropped the dependence of  $x$  for simplicity. We denote as

$$A_I^+ = (A_I^T A_I)^{-1} A_I^T$$

the pseudo inverse of  $A_I$ .

#### 3.1. Optimality Conditions for Penalized Minimization

First order optimality conditions show that a vector  $x^*$  is a solution of the Lasso if and only if

$$\begin{cases} A_I^T(y - Ax^*) = \gamma \text{sign}(\bar{x}^*) \\ \forall j \notin I, \quad |a_j^T(y - Ax^*)| \leq \gamma, \end{cases} \quad (12)$$

where  $I = I(x^*)$ , see for instance [33].

Hence if the goal pursued is to ensure that  $I(x^*) = I(x_0)$  and  $\text{sign}(x^*) = \text{sign}(x_0)$ , the only candidate solution of the Lasso is

$$\bar{x}^* = \bar{x}_0 - \gamma(A_I^T A_I)^{-1} \text{sign}(\bar{x}_0) + A_I^+ w. \quad (13)$$

This vector  $x^*$  is thus a solution of the Lasso if and only if the two following conditions are in force :

$$\text{sign}(x_0) = \text{sign}(x^*) \quad (C_1)$$

$$\forall j \notin I(x_0), \quad |\langle a_j, \gamma d_0(x) + P_{(A_I)^\perp}(w) \rangle| \leq \gamma \quad (C_2)$$

where  $P_{(A_I)^\perp}$  is the orthogonal projection on the vector space orthogonal to the vectors  $(a_i)_{i \in I}$ .

Sections 3.2 and 3.3 show that under the hypotheses of Theorem 1, conditions  $(C_1)$  and  $(C_2)$  are in force with a probability that tends toward 1 with  $n$  going to infinity. This will thus conclude the proof of Theorem 1.

### 3.2. Condition (C<sub>1</sub>)

One has  $\text{sign}(x_0) = \text{sign}(x^*)$  as soon as

$$\|\gamma(A_I^T A_I)^{-1} \text{sign}(\bar{x}_0) + A_I^+ w\|_\infty \leq T \quad (14)$$

We prove that this is indeed the case with high probability.

Lemma 2, whose proof is given in Appendix A, shows that if  $\gamma = \frac{\epsilon}{\sqrt{1-\alpha}} \sqrt{\frac{2 \log p}{n}} \leq \frac{T}{5.5}$  then

$$\gamma \|(A_I^T A_I)^{-1} \text{sign}(\bar{x}_0)\|_\infty \leq \frac{T(1 + 4\sqrt{\alpha})}{5.5}$$

with a probability greater than  $1 - kp^{-1.28} - 2e^{-\frac{n\alpha(0.75\sqrt{2}-1)^2}{4 \log p}}$

To prove (14), we will now bound  $\|A_I^+ w\|_\infty$ . To this end, we use the following decomposition

$$\|A_I^+ w\|_\infty = D_1 \times D_2 \times D_3$$

where

$$D_1 = \frac{\|A_I^+ w\|_\infty}{\|A_I^+ w\|_2}, \quad D_2 = \frac{\|A_I^+ w\|_2}{\|A_I^T w\|_2}, \quad D_3 = \frac{\|A_I^T w\|_2}{\|w\|_2} \times \|w\|_2.$$

*Bounding  $D_1$ .* Since  $A$  and  $w$  are independent, lemma 3, whose proof is given in Appendix A.2, shows that the law of  $A_I^+ w$  is invariant under orthogonal transform  $\mathbb{R}^k$ .

Hence the random variable

$$\frac{A_I^+ w}{\|A_I^+ w\|_2}$$

follows the uniform law on the unit  $\ell^2$  sphere of  $\mathbb{R}^k$ .

Using the concentration lemma 7 detailed in Appendix B with  $\epsilon = \left(\frac{8 \log n \log k}{k^2}\right)^{\frac{1}{4}}$ , one gets

$$P\left(\frac{\|A_I^+ w\|_\infty}{\|A_I^+ w\|_2} \leq \sqrt{\frac{2}{k}} (2 \log n \log k)^{\frac{1}{4}}\right) \geq 1 - 4ke^{-\sqrt{2 \log n \log k}} \geq 1 - \max(4n^{-\frac{1}{3}}, 8e^{-\sqrt{2 \log(2n)}})$$

(One can notice that  $\frac{\|A_I^+ w\|_\infty}{\|A_I^+ w\|_2} \leq 1$  actually gives a better bound if  $k$  is small compared to  $n$ ; moreover the bound on the probability is  $1 - 4n^{-\frac{1}{3}}$  for  $k$  big.)

*Bounding  $D_2$ .*  $D_2$  is bounded by the maximum of the eigenvalues of  $(A_I^T A_I)^{-1}$ . Using lemma 5 with  $t = 1 - \sqrt{\frac{k}{n}} - \frac{1}{2^4}$  gives

$$P\left(D_2 \leq 2^{\frac{1}{4}}\right) \geq 1 - e^{-\frac{n}{2}\left(1 - \frac{1}{2^4} - \frac{1}{\sqrt{2 \log n}}\right)^2}$$

*Bounding  $D_3$ .* One has

$$D_3 = \frac{1}{\|w\|_2} \sum_{i \in I} |\langle a_i, w \rangle|^2.$$

Since each  $\langle a_i, w \rangle$  is a centered Gaussian variable of variance  $\frac{\|w\|_2^2}{n}$ , the variable

$$\frac{n \|A_I^T w\|_2^2}{\|w\|_2^2}$$

follows a  $\chi^2$  distribution with  $k$  degrees of freedom.

Using lemma 8, given in Appendix B, with

$$1 + \delta = 2\sqrt{\frac{\log n}{\log k}}$$

one has

$$P\left(\|A_I^T w\|_2^2 \leq \frac{2k\sqrt{\log n}\|w\|_2^2}{n\sqrt{\log k}}\right) \geq 1 - \frac{1}{\sqrt{2\pi k}} e^{-k(\sqrt{\frac{\log n}{\log k}} - \frac{1}{2} - \frac{\log 2}{2} - \frac{1}{4} \log(\frac{\log n}{\log k}))} \geq 1 - \frac{1}{2} e^{-0.7\sqrt{\log n}}$$

This last bound is quite pessimistic; when  $k$  is large this probability is actually much bigger.

This shows that with high probability,

$$D_3 \leq \sqrt{\frac{2k}{n}} \left(\frac{\log n}{\log k}\right)^{\frac{1}{4}}.$$

*Putting everything together..* Hence one obtains that with a probability bigger than

$$1 - \frac{1}{2} e^{-0.7\sqrt{\log n}} - e^{-\frac{n}{2}(1 - \frac{1}{2^4} - \frac{1}{\sqrt{2\log n}})^2} - \max(4n^{-\frac{1}{3}}, 8e^{-\sqrt{2\log(2n)}}) - kn^{-\frac{3}{2}} - 2e^{-\frac{n(\sqrt{2}-1)^2}{4\log n}}$$

that tends to 1 when  $n$  tends to  $+\infty$ .

$$\|A_I^+ w\|_\infty \leq 2\varepsilon \sqrt{\frac{2\log n}{n}},$$

which, together with the hypotheses of Theorem 1 on  $T$ , implies

$$\|A_I^+ w\|_\infty \leq \frac{2T\sqrt{1-\alpha}}{5.5}.$$

This shows that condition  $(C_1)$  is in force with a probability that tends to 1 when  $n$  tends to  $+\infty$ .

### 3.3. Condition $(C_2)$

In the following, we denote as

$$u = \gamma d_0(x_0) + P_{(A_I)^\perp}(w) \tag{15}$$

that depends on  $x_0$  and  $w$ .

One needs to bound the  $(\langle a_j, u \rangle)_{j \notin I}$  with high probability. We will start by bounding  $\|u\|_2$ .

*Bounding  $\|u\|_2$ .* Using the orthogonality of  $d_0(x_0)$  and  $P_{A_I^\perp}(w)$ , one has

$$\|u\|_2^2 = \gamma^2 \|d_0(x_0)\|_2^2 + \|P_{(A_I)^\perp}(w)\|_2^2. \tag{16}$$

Denoting  $S = \text{sign}(\bar{x}_0)$ , one has

$$\frac{nk}{\|d_0(x_0)\|_2^2} = \frac{n\|S\|_2^2}{S^T(A_I^T A_I)^{-1}S}.$$

Since  $x_0$  and  $A$  are independent, Lemma 6, detailed in Appendix B, shows that  $\frac{nk}{\|d_0(x_0)\|_2^2}$  follows a  $\chi^2$  distribution with  $n - k + 1$  degrees of freedom.

Using Lemma 9, detailed in Appendix B, one obtains that for all  $\delta > 0$ ,

$$P\left(\frac{nk}{n-k+1} < (1-\delta)\|d_0\|_2^2\right) \leq e^{-\frac{(n-k+1)\log(1-\delta)}{2}}$$

Since  $\frac{k}{n} \leq \frac{1}{2\log p}$ , one gets for  $p \geq e^{\frac{1}{2\delta}}$ ,

$$P\left(k < \|d_0\|_2^2(1-\delta)^2\right) \leq e^{-\frac{n\log(1-\delta)(4-\delta)}{8}}$$

and thus, choosing  $\delta$  such that  $(1-\delta) > \sqrt{\beta}$

$$P\left(\|d_0\|_2^2 \leq \frac{k}{\beta}\right) \geq 1 - e^{-\frac{n(3-\sqrt{\beta})\log b}{16}}$$

This shows that

$$\|d_0(x_0)\|_2^2 \leq \frac{k}{\beta}$$

with a probability that tends to 1 when  $n$  tends to  $+\infty$ .

One should note that the condition  $p > e^{\frac{1}{2(1-\sqrt{\beta})}}$  actually guarantees that one can find a suitable  $\delta$ .

Using the fact that  $\|P_{(A_I)^\perp}(w)\|_2 \leq \|w\|_2 \leq \varepsilon$ , the decomposition (16) then shows that

$$P(\|u\|_2^2 \leq \gamma^2 \frac{k}{\beta} + \varepsilon^2) \geq 1 - e^{-\frac{n(3-\sqrt{\beta})\log \beta}{16}} \quad (17)$$

*Bounding  $\max_{j \notin I} |\langle u, a_j \rangle|$ .* For a fixed  $u$ , the random variables  $\langle a_j, u \rangle$  are Gaussian variables of variance  $\frac{\|u\|_2^2}{n}$ .

Using the bound on  $\|u\|_2^2$  (17), one gets that with a probability larger than  $1 - e^{-\frac{n(3-\sqrt{\beta})\log \beta}{16}} - \frac{1}{2\sqrt{\pi}\log p}$

$$\max_{j \notin I} |\langle a_j, u \rangle| \leq \sqrt{\frac{2\log p}{n} \left(\gamma^2 \frac{k}{\beta} + \varepsilon^2\right)}$$

Condition  $(C_2)$  is thus in force with high probability if

$$\sqrt{\frac{2\log p}{n} \left(\gamma^2 \frac{k}{\beta} + \varepsilon^2\right)} \leq \gamma.$$

One thus sees that

$$\frac{\varepsilon}{\sqrt{1-\alpha}} \sqrt{\frac{2\log p}{n}} \leq \gamma$$

implies that  $(C_2)$  is in force with high probability which concludes proof of Theorem 1.

#### 4. Proof of Support Identification of Compressible Signals

Since  $y = Ax^k + A(x_0 - x^k) + w$ , results of section 3.1 are still valid, provided one replaces  $w$  by  $w_2 = Ah + w$  where  $h = x_0 - x^k$  and  $x_0$  by  $x^k$ . Hence we just have to check conditions  $C_1$  and  $C_2$ ; the main difference with the previous section is that the noise  $w_2$  is not independent of  $A$  anymore. More precisely,  $w_2$  is independent of  $(a_i)_{i \in I}$  but not of  $(a_j)_{j \notin I}$ .

*Condition C<sub>1</sub>*. Since this condition only depends on  $A_I$ , it is verified with a probability which tends to 1 as  $n$  tends to infinity, as already seen before, provided  $\|w_2\|_2 \leq \frac{T}{5.5} \sqrt{\frac{(1-\alpha)n}{2 \log p}}$ . But  $\|w_2\|_2 \leq \|w\|_2 + \|Ah\|_2$ , where  $Ah$  is a Gaussian vector, whose coordinates are independent and of variance  $\frac{\|h\|_2^2}{n}$ . Hence  $\frac{n\|Ah\|_2^2}{\|h\|_2^2}$  follows a  $\chi_2$  distribution with  $n$  degrees of freedom. Therefore using lemma 8, one gets

$$P(\|Ah\|_2 \leq 2\|h\|_2) \geq 1 - \frac{1}{3\sqrt{2\pi n}} e^{-0.8n}$$

Using the hypotheses of Theorem 2 one then has with a probability that tends to 1 as  $n$  tends to infinity

$$\|w_2\|_2 \leq \|w\|_2 + 2\|h\|_2 \leq \frac{T}{5.5} \sqrt{\frac{(1-\alpha)n}{2 \log p}}$$

Condition  $C_1$  is thus in force with high probability.

*Condition C<sub>2</sub>*. Let  $j \notin I$ . We denote  $v_j = w_2 - h_j a_j$ . In particular,  $v_j$  is independent of  $a_j$ . Condition  $C_2$  can now be written :

$$\forall j \notin I, |\langle a_j, \gamma d_0(x^k) + P_{(A_I)^\perp}(v_j) + h_j P_{(A_I)^\perp}(a_j) \rangle| \leq \gamma$$

The proof of Theorem 1 shows that with high probability

$$\left\| \gamma d_0(x^k) + P_{(A_I)^\perp}(v_j) \right\|_2^2 \leq \gamma^2 \frac{k}{\beta} + \|v_j\|_2^2$$

But  $\|v_j\|_2 \leq \|w_2\|_2 + \|h\|_\infty \|a_j\|_2$ , where  $n \|a_j\|_2^2$  follows a  $\chi_2$  distribution with  $n$  degrees of freedom; as before one can apply lemma 8 and obtain as in the proof of Theorem 1 that with a probability that tends to 1 as  $n$  tends to infinity,

$$\max_{j \notin I} |\langle a_j, \gamma d_0(x^k) + P_{(A_I)^\perp}(v_j) \rangle| \leq \sqrt{\frac{2 \log p}{n} \left( \gamma^2 \frac{k}{\beta} + (\|w\|_2 + 2\|h\|_2)^2 \right)}$$

Hence the hypotheses of Theorem 2 ensure that with a high probability

$$\max_{j \notin I} |\langle a_j, \gamma d_0(x^k) + P_{(A_I)^\perp}(v_j) \rangle| \leq \frac{\gamma}{2} (1 + \sqrt{\alpha}) \quad (18)$$

Moreover for all  $j \leq p$ ,

$$|h_j \langle a_j, P_{(A_I)^\perp}(a_j) \rangle| \leq \|h\|_\infty \left\| P_{(A_I)^\perp}(a_j) \right\|_2^2$$

where  $\left\| P_{(A_I)^\perp}(a_j) \right\|_2^2$  is the square of the norm of the projection of a Gaussian vector on a vector space of dimension  $n - k$ . The vector space  $(A_I)^\perp$  is independent of  $a_j$ , therefore  $(n - k) \left\| P_{(A_I)^\perp}(a_j) \right\|_2^2$  follows a  $\chi_2$  distribution with  $n - k$  degrees of freedom. Hence one has with high probability

$$\max_{j \notin j} |h_j \langle a_j, P_{(A_I)^\perp}(a_j) \rangle| \leq \frac{5}{4} \|h\|_\infty \leq \frac{\gamma}{2} (1 - \sqrt{\alpha}) \quad (19)$$

Equations (18) and (19) show that  $C_2$  holds with high probability.



## 5. Proof of Partial Support Recovery

To prove the first part of Theorem 3, we need to show that with w.o.p, the extension  $x_1(\gamma)$  on  $\mathbb{R}^p$  of the solution of

$$\min_{x \in \mathbb{R}^{|I|}} \frac{1}{2} \|y_1 - Ax\|_2^2 + \gamma \|x\|_1 \quad (20)$$

where  $y_1 = P_{A_I}(y)$ , is the solution of the Lasso. By definition, the support  $J$  of this extension is included in  $I$ .

To prove this, one needs to prove that

$$\begin{cases} A_J^T(y - Ax_1(\gamma)) = \gamma \text{sign}(\overline{x_1(\gamma)}) \\ \forall l \notin J, \quad |a_l^T(y - Ax_1(\gamma))| \leq \gamma, \end{cases} \quad (21)$$

By definition, the support  $J$  of  $x_1(\gamma)$  is included in  $I$ . Since  $y_1 = P_{A_I}(y)$ , one has  $A_J^T(y - Ax_1(\gamma)) = A_J^T(y_1 - Ax_1(\gamma))$ ; moreover, since  $x_1(\gamma)$  is the extension of the solution of (20), one has :

$$A_J^T(y - Ax_1(\gamma)) = \gamma \text{sign}(\overline{x_1(\gamma)})$$

and

$$\forall l \in (I \cap J^c), \quad |a_l^T(y - Ax_1(\gamma))| \leq \gamma.$$

To prove Theorem 3, it is thus sufficient to prove that w.o.p

$$\forall l \notin I, \quad |a_l^T(y - Ax_1(\gamma))| \leq \gamma. \quad (22)$$

Following the proofs of Theorems 1 and 2, we will use the independency between vectors  $(a_l)_{l \notin I}$  and vector  $y - Ax_1(\gamma)$  to bound these scalar products.

**Lemma 1.** *Let  $A \in \mathbb{R}^{n \times k}$  such that  $(A^T A)$  is invertible, let  $y \in \mathbb{R}^n$  and  $x(\gamma)$  be a solution of the Lasso. The application  $f$  from  $\mathbb{R}^{+*}$  to  $\mathbb{R}^+$  defined by  $f(\gamma) = \frac{\|y - Ax(\gamma)\|_2}{\gamma}$  is well-defined and non-increasing.*

**Proof:** The authors in [8] and [60] independently proved that:

- the solution  $x(\gamma)$  of the Lasso is unique;
- there is a finite increasing sequence  $(\gamma_k)_{k \leq K}$  with  $\gamma_0 = 0$  and  $\gamma_K = \|A^T y\|_\infty$  such that for all  $k < K$ , the sign and the support of  $x(\gamma)$  are constant on each interval  $(\gamma_k, \gamma_{k+1})$ .
- $x(\gamma)$  is a continuous function of  $\gamma$ .

Moreover  $x(\gamma)$  satisfies

$$\overline{x(\gamma)} = (A_J^T A_J)^{-1} A_J^T y - \gamma (A_J^T A_J)^{-1} \text{sign}(\overline{x(\gamma)}) \quad (23)$$

which implies

$$r(\gamma) := y - Ax(\gamma) = P_{A_J^\perp}(y) - \gamma A_J (A_J^T A_J)^{-1} \text{sign}(\overline{x(\gamma)})$$

Hence on each interval  $(\gamma_k, \gamma_{k+1})$ ,  $r(\gamma)$  is an affine function of  $\gamma$  which can be written

$$r(\gamma) = z - \gamma v, \quad \text{with } v \in V := \text{Span}((a_j)_{j \in J}) \text{ and } z \in V^\perp$$

For  $\gamma \in (\gamma_k, \gamma_{k+1})$ , one has

$$\frac{\|r(\gamma)\|_2^2}{\gamma^2} = \frac{\|z\|_2^2}{\gamma^2} + \|v\|_2^2 \quad (24)$$

hence  $f(\gamma) = \frac{\|r(\gamma)\|_2}{\gamma}$  is a non increasing function of  $\gamma$  on each interval  $(\gamma_k, \gamma_{k+1})$ . Since  $f$  is continuous, it follows that  $f$  is non increasing on  $\mathbb{R}^{+*}$ .

**Remark 1.** *If  $(A_I^T A_I)$  is not invertible, the Lasso may have several solutions. Nevertheless  $r(\gamma)$  is always uniquely defined and the lemma should also apply.*

From Lemma 1, we deduce that  $\frac{\|y_1 - Ax_1(\gamma)\|_2}{\gamma}$  is a non increasing function of  $\gamma$ . Since  $y_1 \in \text{Im}(A_I)$  and  $A_I$  has full column-rank,

$$\lim_{\gamma \rightarrow 0} x_1(\gamma) = x_1 = x_0 + (A_I^T A_I)^{-1} A_I^T w$$

is the unique vector of  $\mathbb{R}^{|I|}$  such that  $A_I x = y_1$ . Since  $A_I$  is Gaussian and independant from  $x_0$  and  $w$ , the support of  $x_1$  is almost surely equal to  $I$ . Hence there exists  $\gamma_1 > 0$  such that if  $\gamma < \gamma_1$ , the support and the sign of  $x_1(\gamma)$  are equal those of  $x_1$ . More precisely, if  $\gamma < \gamma_1$ , one has

$$x(\gamma) = x_1 - \gamma(A_I^T A_I)^{-1} \text{sign}(\bar{x}_1) \quad \text{and} \quad r(\gamma) = y_1 - Ax_1(\gamma) = \gamma A_I (A_I^T A_I)^{-1} \text{sign}(\bar{x}_1)$$

Hence for  $\gamma \in (0, \gamma_1)$ ,  $\frac{\|y_1 - Ax_1(\gamma)\|_2}{\gamma} = \|A_I (A_I^T A_I)^{-1} \text{sign}(\bar{x}_1)\|_2$ . Since

$$\|A_I (A_I^T A_I)^{-1} \text{sign}(\bar{x}_1)\|_2^2 = \langle (A_I^T A_I)^{-1} \text{sign}(\bar{x}_1), \text{sign}(\bar{x}_1) \rangle$$

we deduce that for all  $\gamma > 0$ ,  $\frac{\|y_1 - Ax_1(\gamma)\|_2}{\gamma} \leq \sqrt{|I| \rho((A_I^T A_I)^{-1})}$ ; hence for any  $\beta < (1 - \sqrt{\frac{k}{n}})^2$ ,

$$P\left(\frac{\|y_1 - Ax_1(\gamma)\|_2}{\gamma} \leq \sqrt{\frac{k}{\beta}}\right) \geq 1 - e^{-\frac{n(1-\sqrt{\beta}-\sqrt{\frac{k}{n}})^2}{2}} \quad (25)$$

Since  $\|y - Ax_1(\gamma)\|_2^2 \leq \varepsilon^2 + \|y_1 - Ax_1(\gamma)\|_2^2$  we deduce that with a probability larger than  $1 - e^{-\frac{n(1-\sqrt{\beta}-\sqrt{\frac{k}{n}})^2}{2}} - \frac{1}{2\sqrt{\pi \log p}}$ ,

$$\max_{l \notin I} |a_l^T (y - Ax_1(\gamma))| \leq \sqrt{\frac{2 \log p (\varepsilon^2 + \frac{\gamma^2 k}{\beta})}{n}}. \quad (26)$$

If  $k \leq \frac{\alpha \beta n}{2 \log p}$  and  $\gamma \geq \frac{\varepsilon}{\sqrt{1-\alpha}} \sqrt{\frac{2 \log p}{n}}$  then  $\sqrt{\frac{2 \log p (\varepsilon^2 + \frac{\gamma^2 k}{\beta})}{n}} \leq \gamma$  and then inequality (22) is satisfied w.o.p. which ends the proof of the first part of the theorem.

We now prove that w.o.p

$$\|x_0 - x(\gamma)\|_2 \leq \varepsilon \left( \sqrt{\frac{\alpha}{1-\alpha}} + 2 \right).$$

To prove this inequality we notice that for large  $\gamma$ ,  $x(\gamma)$  is also the extension of the solution of (20) w.o.p and we use the Lipschitz property of the function mapping  $\gamma$  to  $x_1(\gamma)$ .

Indeed, one has

$$\|x_0 - x_1(\gamma)\|_2 \leq \|x_0 - x_1\|_2 + \|x_1 - x_1(\gamma)\|_2 . \quad (27)$$

Noticing  $x_0 - x_1 = (A_I^T A_I)^{-1} A_I^T w$ , it follows that  $\|x_0 - x_1\|_2 \leq \varepsilon \sqrt{\rho((A_I^T A_I)^{-1})}$ .

We deduce that  $\|x_0 - x_1\|_2 \leq 2\varepsilon$  with probability larger than  $1 - e^{-\frac{n}{2}(0.29 - \sqrt{\frac{kn}{n}})^2}$ . Since for all  $\gamma > 0$ ,  $x_1(\gamma)$  is defined by (23) and  $\lim_{\gamma \rightarrow 0} x_1(\gamma) = x_1$  one has

$$\|x_1 - x_1(\gamma)\|_2 \leq \gamma \max_{J \subset I, S \in \{-1, 1\}^{|J|}} \|(A_J^T A_J)^{-1} S\|_2 . \quad (28)$$

For all  $J \subset I$ , one has  $\rho((A_J^T A_J)^{-1}) \leq \rho((A_I^T A_I)^{-1})$  and for all  $S \in \{-1, 1\}^{|J|}$ ,  $\|S\|_2 \leq \sqrt{k}$ ; thus

$$P \left( \|x_1 - x_1(\gamma)\|_2 \leq \gamma \sqrt{\frac{k}{\beta}} \right) \geq 1 - e^{-\frac{n}{2} \left( 0.29 - \sqrt{\frac{k}{n}} \right)^2} .$$

If  $\gamma = \frac{\varepsilon}{\sqrt{1-\alpha}} \sqrt{\frac{2 \log p}{n}}$  and  $k \leq \frac{\alpha \beta n}{2 \log p}$

$$\|x_1 - x_1(\gamma)\|_2 \leq \varepsilon \sqrt{\frac{\alpha}{1-\alpha}} ,$$

which concludes the proof.

## 6. Numerical Illustrations

This section aims at providing empirical support of the sharpness of our bounds by assessing experimentally the quality of the constants involved in Theorem 1. More specifically, we perform a probabilistic analysis of support and sign recovery, to show that the bounds (6), (8) and (7) are quite tight.

In all the numerical tests, we use problems of size  $(n, p) = (8000, 32000)$ , which is a realistic high-dimensional scenario in agreement with signal and image processing applications. We perform a randomized analysis, where the probability of exact recovery of supports and signs (sparsistency) are computed by Monte-Carlo sampling with respect to a probability distribution on the measurement matrix,  $k$ -sparse signals and on the noise  $w$ . As detailed in Section 1.1, the matrix  $A$  is a realization drawn from the Gaussian ensemble. We assume that the non-zero entries  $x(i)$  for  $i \in I(x)$  of a signal  $x$  are independent realizations of a Bernoulli variable taking equiprobable values  $\{+T, -T\}$ . We also assume that the noise  $w$  is drawn from the uniform distribution on the sphere  $\{w \in \mathbb{R}^n \mid \|w\| = \varepsilon\}$ . Since only the signal-to-noise ratio matters in the bounds, we fix  $\varepsilon = 1$  and only vary the value of  $T$ .

*Challenging the sparsity bound (6).* We first evaluate, for  $\alpha = 0.8$ , and for a varying value of  $k$ , the probability of sparsistency given that

$$T = \frac{5.5\varepsilon}{\sqrt{1-\alpha}} \sqrt{\frac{2 \log p}{n}} \quad \text{and} \quad \gamma = \frac{T}{5.5} \quad (29)$$

which are values in accordance with the bounds (7) and (8).

In order to compute numerically this probability, for each  $k$ , we generate 1000 sparse signals  $x_0$  with  $\|x_0\|_0 = k$ , and check whether conditions  $(C_1)$  and  $(C_2)$  defined in Section 3.1

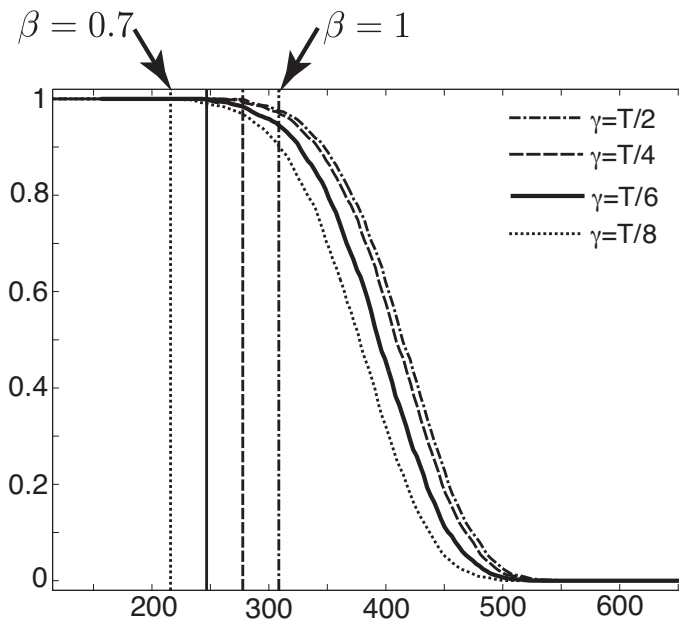


Figure 1: Probability of sparsistency as a function of  $k$  for  $(n, p) = (8000, 32000)$  and  $\alpha = 0.8$ . The vertical lines corresponds to our sparsistency bound  $k_\beta$ , from left to right, for  $\beta = 0.7, 0.8, 0.9, 1$ .

are satisfied. Figure 1 shows how this probability decays when  $k$  increases. The vertical lines corresponds to the critical sparsity thresholds

$$k_\beta = \frac{\alpha\beta n}{2 \log p} \quad (30)$$

as identified by the bounds (6). The estimated probability exhibits a typical phase transition that is located precisely around the critical value  $k_\beta$  for  $\beta$  close to one. This shows that our bound is quite sharp. We have also displayed the same probability curve for other, less conservative, values of  $\gamma \in \{T/4, T/2\}$ , which improves slightly the probability with respect to  $\gamma = T/5.5$ .

*Challenging the regularization parameter value* (8). We evaluate, for  $(\alpha, \beta) = (0.8, 0.8)$ , the probability of sparsistency using a value of  $\gamma$  that differs from

$$\gamma_0 = \frac{\varepsilon}{\sqrt{1-\alpha}} \sqrt{\frac{2 \log p}{n}} \quad (31)$$

given in (8), for which Theorem 1 is valid. We use the critical sparsity level  $k = k_\beta$  defined in (30). To study only the influence of  $\gamma$ , we use a signal-to-noise ratio that is infinite, meaning that  $\varepsilon$  is small in comparison to  $T$ . This implies in particular that in this regime, only condition  $(C_1)$  is needed to be checked to estimate the probability of sparsistency.

Figure 2 shows the increase in this probability as the ratio  $\gamma/\gamma_0$  increases. This makes sense because the signal is large with respect to the noise so that a large threshold should be preferred. One can see that at the critical value  $\gamma = \gamma_0$  suggested by Theorem 1, this probability is close to 1. This shows that the value (8) of  $\gamma$  is quite sharp.

*Challenging the signal-to-noise ratio* (7). Lastly, we estimate, for  $(\alpha, \beta) = (0.8, 0.8)$ , the minimal signal level  $T$  that is required to ensure the inclusion of the support, meaning  $I(x(\gamma)) \subset I(x_0)$ . We use the critical sparsity  $k = k_\beta$  defined in (30), and  $\gamma = \gamma_0$  where  $\gamma_0$

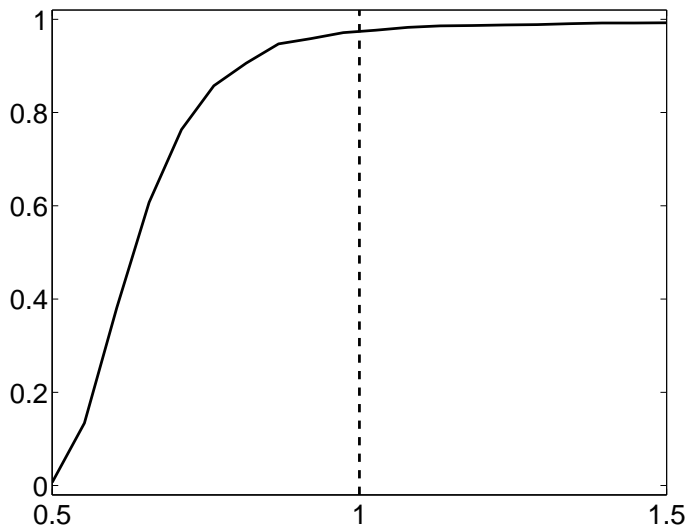


Figure 2: Probability of support recovery for large  $T$  as a function of  $\gamma/\gamma_0$  for  $k = k_\beta$ ,  $(n, p) = (8000, 32000)$  and  $(\alpha, \beta) = (0.8, 0.8)$ .

is defined in (31). Since we are only interested in support inclusion, we only need to check condition  $(C_2)$ .

The bound in (7) suggests that  $T \geq 5.5\gamma_0$  is enough. Figure 3 however suggests that this bound is pessimistic, and that  $T \geq 2\gamma_0$  appears to be enough to guarantee the support inclusion with high probability. A few reasons may explain this sub-optimality.

- There is no guarantee that the concentration lemmas we use are optimal.
- The limit ratio  $\frac{T}{\epsilon}$  relies mainly on Lemma 2 and especially on the bound  $1 + 4\sqrt{b}$  in it. This bound can be improved by at least three ways.
  - Using the same proof, the bound can be slightly enhanced by decaying the probability of success.
  - The result in the lemma is non asymptotic. The bound and the probability were computed to be available for all  $\alpha \leq 1, \beta \leq 1$  and for all  $p \geq 1212$ . With the values used in the numerical experiments, and decaying a bit the probability of success, the bound can turn into  $1 + 2.7\sqrt{b}$ , yielding a better bound  $T \geq 4.37\gamma_0$ .
  - In the proof of Lemma 2, the majorization  $\|B_i\|_2 \leq \rho(B)$  is used. This bound is available for any matrix, but one might perhaps do better by exploiting Gaussianness of the measurement matrix.

## Conclusion

This paper has presented a novel analysis of the sparsistency of the Lasso from noisy Gaussian measurements. We derived sharp bounds on the sparsity of the signal to guarantee sparsistency with high probability. This result is extended to handle compressible signals and to establish sharp  $\ell^2$ -consistency. A distinctive feature of our analysis is that it provides explicit constants for the three key parameters of the problem: the sparsity of the signal, the minimal signal-to-noise ratio and the Lasso regularization parameter. Numerical results support the claim that these constants are either sharp or at least reasonably well behaved.

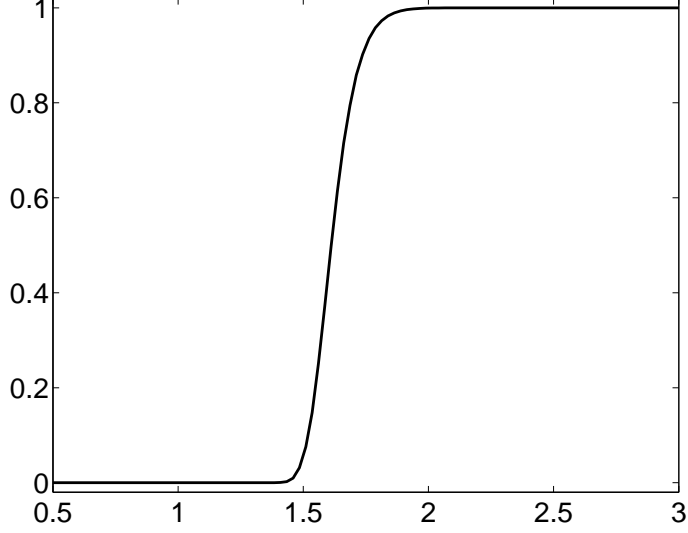


Figure 3: Probability of support inclusion as a function of  $T/\gamma_0$  for  $k = k_\beta$ ,  $(n, p) = (8000, 32000)$  and  $(\alpha, \beta) = (0.8, 0.8)$ .

## A. Properties of Wishart Matrices

### A.1. Sup-norm of a projected Rademacher sequence

**Lemma 2.** If  $C \in \mathbb{R}^{n \times k}$  is a Gaussian matrix, with  $k \leq \frac{nb}{2 \log p}$  with  $0 < b \leq 1$  and if  $S \in \{-1, 1\}^k$  is drawn independently from  $C$ , then if  $p \geq 1212$ ,

$$P\left(\|(C^T C)^{-1} S\|_\infty \leq 1 + 4\sqrt{b}\right) \geq 1 - kp^{-1.28} - 2e^{-\frac{nb(0.75\sqrt{2}-1)^2}{4 \log p}}$$

**Proof:** We use the following splitting

$$(C^T C)^{-1} = I + ((C^T C)^{-1} - I) = I + B.$$

This shows that

$$\|(C^T C)^{-1} S\|_\infty \leq \|S\|_\infty + \|BS\|_\infty = 1 + \|BS\|_\infty.$$

One can then observe that  $(BS)(i) = \sum_{j \leq k} |B_{i,j}| S(j) \text{sign}(B_{i,j})$ ; one has  $B_{i,i} > 0$ , and according to lemma 4, for given  $i$ , the variables  $\text{sign}(B_{i,j})_{j \neq i}$  form a Rademacher sequence (this means that they are independent and uniformly distributed on  $\{-1, 1\}$ ), and this sequence is independent of  $B_{i,i}$  and of  $(|B_{i,j}|)_{j \neq i}$ . Hence one can apply Hoeffding's lemma 10 (multiplying the line by an independent variable uniform on  $\{-1, 1\}$  to take care of the fact that  $\text{sign}(B_{i,i})$  is not uniformly distributed), thus getting for any  $i \leq k$  and any  $t > 0$ ,

$$P\left(\left|\sum_{j=1}^k B_{i,j} S(j)\right| \geq t \|B_i\|_2\right) \leq e^{-\frac{t^2}{2}} \quad (32)$$

Now, for all  $i \leq k$ ,  $\|B_i\|_2 \leq \rho(B)$ , where  $\rho(B)$  is the spectral radius of  $B$ . Using lemma 5 with  $t = (0.75 - \frac{1}{\sqrt{2}}) \sqrt{\frac{b}{\log p}}$  and using  $\frac{k}{n} \leq \frac{b}{2 \log p}$ , one gets :

$$P\left(\lambda_{\min}(C^T C) \leq \left(1 - 0.75 \sqrt{\frac{b}{\log p}}\right)^2\right) \leq e^{-\frac{nb(0.75\sqrt{2}-1)^2}{4 \log p}}$$

Hence

$$P \left( \lambda_{\max}((C^T C)^{-1}) \geq \left( 1 - 0.75 \sqrt{\frac{b}{\log p}} \right)^{-2} \right) \leq e^{-\frac{(0.75\sqrt{2}-1)^2 bn}{4 \log p}}$$

Similarly, one gets

$$P \left( \lambda_{\min}((C^T C)^{-1}) \leq \left( 1 + 0.75 \sqrt{\frac{b}{\log p}} \right)^{-2} \right) \leq e^{-\frac{(0.75\sqrt{2}-1)^2 bn}{4 \log p}}$$

Finally, one has with a probability bigger than  $1 - 2e^{-\frac{nb(0.75\sqrt{2}-1)^2}{4 \log p}}$

$$\rho(B) \leq \max \left( \left| \left( 1 + 0.75 \sqrt{\frac{b}{\log p}} \right)^{-2} - 1 \right|, \left| \left( 1 - 0.75 \sqrt{\frac{b}{\log p}} \right)^{-2} - 1 \right| \right)$$

In particular, if  $\frac{\log(p)}{b} \geq \frac{15^2}{(17-\sqrt{129})^2} \simeq 7.07$ , one gets  $\rho(B) \leq 2.5 \sqrt{\frac{b}{\log p}}$  with a probability larger than  $1 - 2e^{-\frac{nb(0.75\sqrt{2}-1)^2}{4 \log p}}$ .

Using this bound in (32) with  $t = 1.6\sqrt{\log(p)}$  yields

$$\begin{aligned} P \left( \|BS\|_{\infty} \geq 4\sqrt{b} \right) &\leq P \left( \|BS\|_{\infty} \geq t \|B_i\|_2 \text{ and } \rho(B) \leq 2.5 \sqrt{\frac{b}{\log p}} \right) \\ &\quad + P \left( \rho(B) \geq 2.5 \sqrt{\frac{b}{\log p}} \right) \\ &\leq kp^{-1.28} + 2e^{-\frac{nb(0.75\sqrt{2}-1)^2}{4 \log p}} \end{aligned}$$

Finally if  $\frac{\log(p)}{b} \geq 7.08$

$$P \left( \|(C^T C)^{-1} S\|_{\infty} \leq 1 + 4\sqrt{b} \right) \geq 1 - kp^{-1.28} - 2e^{-\frac{nb(0.75\sqrt{2}-1)^2}{4 \log p}}$$

One can notice that if one actually has  $\frac{\log p}{b} \geq 16.2$  as in the numerical experiments ( $b = 0.64, p = 32000$ ), one can adapt this proof and, by losing a bit on the probability (applying lemmas with smaller values of  $t$ ), one can get  $\|(C^T C)^{-1} S\|_{\infty} \leq 1 + 2.7\sqrt{b}$  w.o.p. ■

### A.2. Rotation invariance

**Lemma 3.** *If  $C \in \mathbb{R}^{n \times k}$  is a Gaussian matrix, and  $w \in \mathbb{R}^n$  is independent of  $C$ , the law of  $C^+ w$  is invariant under orthogonal transformation of  $\mathbb{R}^k$ , where  $C^+ = (C^T C)^{-1} C^T$ .*

**Proof:** If  $C \in \mathbb{R}^{n \times k}$  is a Gaussian matrix, then for any orthogonal matrix  $U \in \mathbb{R}^{k \times k}$ ,  $D = CU$  and  $C$  have the same law. The law of  $D^+ w$  and  $C^+ w$  are thus the same. Since for all  $w$ , one has

$$D^+ w = U^{-1} C^+ w$$

the law of  $U^{-1} C^+ w$  is the same as the law of  $C^+ w$ .

### A.3. Signs of non-diagonal entries of an inverse Wishart matrix

**Lemma 4.** *If  $B \in \mathbb{R}^{k \times k}$  is the inverse of a Wishart matrix, then for all  $i \leq k$ , the variables  $(\text{sign}(B_{i,j}), j \neq i)$  form a Rademacher sequence, that is they are independent and uniformly distributed on  $\{-1, 1\}$ . Moreover this sequence is independent of  $B_{i,i}$ , and of  $(|B_{i,j}|)_{j \neq i}$ .*

**Proof:** If  $B = (b_{ij})_{i \leq k, j \leq k} \in \mathbb{R}^{k \times k}$  is the inverse of a Wishart matrix, then  $B = (A^T A)^{-1}$  where  $A \in \mathcal{M}_{n,k}(\mathbb{R})$  is a Gaussian matrix. Let  $E \in \mathcal{M}_{k,k}(\mathbb{R})$  be diagonal such that for all  $1 \leq i \leq k, |E_{i,i}| = 1$ . Then  $(AE)^T AE = EA^T AE$ , hence  $((AE)^T (AE))^{-1} = E(A^T A)^{-1} E$ . Therefore the coefficients of  $C = ((AE)^T AE)^{-1}$  are  $(c_{ij})_{i \leq k, j \leq k} = (e_{ii} e_{jj} b_{ij})_{i \leq k, j \leq k}$ . But  $A$  and  $AE$  have the same law, hence  $B$  and  $C$  also have the same law. Hence for all  $(\epsilon_j)_{j \leq k, j \neq i} \in \{-1, 1\}^{k-1}$ , the laws of  $(B_{i,1}, \dots, B_{i,k})$  and  $(\epsilon_1 B_{i,1}, \dots, B_{i,i}, \dots, \epsilon_k B_{i,k})$  are the same. This implies that the variables  $(\text{sign}(B_{i,j}), j \neq i)$  form a Rademacher sequence, and this sequence is independent of  $B_{i,i}$ , and of  $(|B_{i,j}|)_{j \neq i}$ . ■

### A.4. Extreme eigenvalues of a Wishart matrix

The proof of the following lemma can be found in [63, page 42].

**Lemma 5.** *If  $A \in \mathbb{R}^{n \times k}$  is a Gaussian matrix whose coefficients are centered of variance  $\frac{1}{n}$ , then the maximal and minimal eigenvalues of the Wishart matrix  $B = A^T A$  satisfy for all  $t > 0$*

$$P \left( \lambda_{\max}(B) \geq \left(1 + \sqrt{\frac{k}{n}} + t\right)^2 \right) \leq e^{-\frac{nt^2}{2}}$$

and

$$P \left( \lambda_{\min}(B) \leq \left(1 - \sqrt{\frac{k}{n}} - t\right)^2 \right) \leq e^{-\frac{nt^2}{2}}$$

### A.5. Distribution of a quadratic form

The following lemma is a consequence of [64, Theorem 3.2.12].

**Lemma 6.** *If  $B$  is a Wishart matrix as described in Lemma 5, then for all  $X \in \mathbb{R}^k$  independent of  $B$ , the random variable  $\frac{n\|X\|_2^2}{X^T B^{-1} X}$  follows a  $\chi^2$  distribution with  $n - k + 1$  degrees of freedom.*

## B. Concentration inequalities

The following lemma is well known; a proof can be found in [65].

**Lemma 7.** *Let  $\mu_k$  denote the uniform probability on the unit sphere  $S^{k-1}$  in  $\mathbb{R}^k$ , and let  $A \subset S^{k-1}$  such that  $\mu_k(A) \geq \frac{1}{2}$ . Then  $\mu_k(\{x \in S^{k-1}, d(x, A) \leq \epsilon\}) \geq 1 - 2e^{-\frac{k\epsilon^2}{2}}$ . As a corollary,  $\mu_k(x \in S^{k-1}, |x_1| \leq \epsilon) \geq 1 - 4e^{-\frac{k\epsilon^2}{2}}$ .*

The following lemma is due to Cai et Silverman, see [66].

**Lemma 8.** *If  $X$  follows a  $\chi^2$  distribution with  $k$  degrees of freedom, then for all  $\delta > 0$ ,*

$$P(X > (1 + \delta)k) \leq \frac{1}{\sqrt{2\pi k \delta}} e^{-\frac{k}{2}(\delta - \log(1 + \delta))}$$



The following lemma is due to Hoeffding, see [67].

**Lemma 9.** *If  $X$  follows a  $\chi^2$  distribution with  $k$  degrees of freedom, then for all  $\delta > 0$ ,*

$$P(X < (1 - \delta)k) \leq e^{-\frac{k \log(1-\delta)}{2}}$$

The following lemma can be obtained by applying the Chernoff-Hoeffding inequality.

**Lemma 10.** *If  $(\varepsilon_i)_{i \leq k}$  is a Rademacher sequence, then for all  $a = (a_i)_{i \leq k} \in \mathbb{R}^k$  and for all  $t > 0$ ,*

$$P\left(\left|\sum_{i=1}^k \varepsilon_i a_i\right| \geq t \|a\|_2\right) \leq e^{-\frac{t^2}{2}}.$$

## References

- [1] E. Candès, J. Romberg, T. Tao, Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information, *IEEE Trans. Info. Theory* 52 (2) (2006) 489–509.
- [2] E. Candès, T. Tao, Near-optimal signal recovery from random projections: Universal encoding strategies?, *IEEE Trans. Info. Theory* 52 (12) (2006) 5406–5425.
- [3] D. Donoho, Compressed sensing, *IEEE Trans. Info. Theory* 52 (4) (2006) 1289–1306.
- [4] S. S. Chen, D. Donoho, M. Saunders, Atomic decomposition by basis pursuit, *SIAM Journal on Scientific Computing* 20 (1) (1998) 33–61.
- [5] R. Tibshirani, Regression shrinkage and selection via the Lasso, *Journal of the Royal Statistical Society* 58 (1), 267–288.
- [6] E. J. Candès, T. Tao, Rejoinder: the Dantzig selector: statistical estimation when  $p$  is much larger than  $n$ , *Annals of Statistics* 35 (6) (2007) 2392–2404.
- [7] P. J. Bickel, Y. Ritov, A. Tsybakov, Simultaneous analysis of lasso and Dantzig selector, *Annals of Statistics* 37, 1705–1732.
- [8] M. R. Osborne, B. Presnell, B. A. Turlach, On the lasso and its dual, *Journal of Computational and Graphical Statistics* 9 (2) (2000) 319–337.
- [9] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, Least angle regression, *Annals of Statistics* 32 (2), 407–499.
- [10] D. L. Donoho, Y. Tsaig, Fast solution of  $\ell_1$ -norm minimization problems when the solution may be sparse, *IEEE Trans. Info. Theory* 54 (11) (2008) 4789–4812.
- [11] M. Figueiredo, R. Nowak, An EM Algorithm for Wavelet-Based Image Restoration, *IEEE Trans. Image Proc.* 12 (8) (2003) 906–916.
- [12] I. Daubechies, M. Defrise, C. D. Mol, An iterative thresholding algorithm for linear inverse problems with a sparsity constraint, *Commun. on Pure and Appl. Math.* 57 (2004) 1413–1541.
- [13] J. Bect, L. Blanc Féraud, G. Aubert, A. Chambolle, A  $\ell_1$ -unified variational framework for image restoration, in: *Proc. of ECCV04*, Springer-Verlag, 2004, pp. Vol IV: 1–13.

- [14] P. L. Combettes, V. R. Wajs, Signal recovery by proximal forward-backward splitting, *SIAM Journal on Multiscale Modeling and Simulation* 4 (4) (2005) 1168–1200.
- [15] M. A. T. Figueiredo, R. D. Nowak, S. J. Wright, Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems, *IEEE Journal of Selected Topics in Signal Processing* 1 (4) (2007) 586–598.
- [16] J. M. B. Dias, M. A. T. Figueiredo, A new twIST: Two-step iterative shrinkage/thresholding algorithms for image restoration, *IEEE Trans. Image Proc.* 16 (12) (2007) 2992–3004.
- [17] Y. Nesterov, Gradient methods for minimizing composite objective function, *CORE Discussion Papers 2007076*, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE) (Sep. 2007).
- [18] A. Beck, M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems, *Journal on Imaging Sciences* 2 (1) (2009) 183–202.
- [19] P. L. Combettes, J.-C. Pesquet, A Douglas-Rachford splitting approach to nonsmooth convex variational signal recovery, *IEEE Journal of Selected Topics in Signal Processing* 1 (4) (2007) 564–574.
- [20] M. Fadili, J.-L. Starck, Monotone operator splitting for fast sparse solutions of inverse problems, in: *Proc. of IEEE ICIP*, Cairo, Egypt, 2009.
- [21] J.-L. Starck, F. Murtagh, M. Fadili, *Sparse Signal and Image Processing: Wavelets, Curvelets and Morphological Diversity*, Cambridge University Press, Cambridge, UK, 2010.
- [22] E. J. Candès, Y. Plan, Near-ideal model selection by  $\ell_1$  minimization, *Annals of Statistics* 37 (5A) (2009) 2145–2177.
- [23] D. L. Donoho, M. Elad, V. N. Temlyakov, Stable recovery of sparse overcomplete representations in the presence of noise, *IEEE Trans. Info. Theory* 52 (1) (2006) 6–18.
- [24] N. Meinshausen, P. Bühlmann, High-dimensional graphs and variable selection with the lasso, *Ann. Statist.* 34 (3) (2006) 1436–1462.
- [25] E. Greenshtein, Best subset selection, persistence in high-dimensional statistical learning and optimization under  $\ell_1$  constraint, *Annals of Statistics* 34 (2006) 2367–2386.
- [26] J. A. Tropp, Just relax: convex programming methods for identifying sparse signals in noise, *IEEE Trans. Info. Theory* 52 (3) (2006) 1030–1051.
- [27] M. J. Wainwright, Sharp thresholds for high-dimensional and noisy sparsity recovery using  $\ell_1$ -constrained quadratic programming (lasso), *IEEE Trans. Info. Theory* 55 (5) (2009) 2183–2202.
- [28] P. Zhao, B. Yu, On model selection consistency of lasso, *J. Mach. Learn. Res.* 7 (2006) 2541–2563.
- [29] H. Zou, The adaptive lasso and its oracle properties, *Journal of the American Statistical Association* 101 (476) (2006) 1418–1429.

- [30] J. Fuchs, Recovery of exact sparse representations in the presence of bounded noise, *IEEE Trans. Info. Theory* 51 (10) (2005) 3601–3608.
- [31] F. Bunea, Consistent selection via the lasso for high dimensional approximating regression models, in: *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh*, Vol. 3, Institute of Mathematical Statistics, 2008, pp. 122–137.
- [32] S. Zhou, J. D. Lafferty, L. A. Wasserman, Compressed and privacy-sensitive sparse regression, *IEEE Trans. Info. Theory* 55 (2) (2009) 846–866.
- [33] J.-J. Fuchs, On sparse representations in arbitrary redundant bases, *IEEE Trans. Info. Theory* 50 (6) (2004) 1341–1344.
- [34] N. Meinshausen, B. Yu, Lasso-type recovery of sparse representations for high-dimensional data, *Ann. Statist.* 37 (1) (2009) 246–270.
- [35] S. A. van de Geer, P. Bühlmann, On the conditions used to prove oracle results for the lasso, *Electron. J. Statist.* 3 (2009) 1360–1392.
- [36] F. R. Bach, Consistency of the group lasso and multiple kernel learning, *J. Mach. Learn. Res.* 9 (2008) 1179–1225.
- [37] Y. Nardi, A. Rinaldo, On the asymptotic properties of the group lasso estimator for linear models, *Electron. J. Statist.* 2 (2008) 605–633.
- [38] M. Yuan, Y. Lin, Model selection and estimation in regression with grouped variables, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68 (1) (2006) 49–67.
- [39] J. A. Tropp, Norms of random submatrices and sparse approximation, *C. R. Math. Acad. Sci.* 346 (2008) 1271–1274.
- [40] D. Omidiran, M. J. Wainwright, High-dimensional subset recovery in noise: Sparsified measurements without loss of statistical efficiency, UC Berkeley Technical Report 753.
- [41] J. Huang, S. Ma, C.-H. Zhang, Adaptive lasso for sparse high dimensional regression models, Tech. rep., Univ. of Iowa (2006).
- [42] E. J. Candes, M. B. Wakin, S. P. Boyd, Enhancing sparsity by reweighted L1 minimization, *J. Fourier Anal. Appl.* 14 (5) (2008) 877–905.
- [43] S. Zhou, Thresholded lasso for high dimensional variable selection and statistical estimation, Tech. Rep. arXiv:1002.1583v1 (2009).
- [44] E. Candès, T. Tao, Decoding by linear programming, *IEEE Trans. Info. Theory* 51 (12) (2005) 4203–4215.
- [45] T. Zhang, Some sharp performance bounds for least squares regression with  $\ell_1$  regularization, *Annals of Statistics* 37 (2009) 2109–2144.
- [46] L. Wasserman, K. Roeder, High dimensional variable selection, *Annals of statistics* 37 (2009) 2178–2201.
- [47] S. A. van de Geer, P. Bühlmann, S. Zhou, Prediction and variable selection with the adaptive lasso, Tech. Rep. arXiv:1001.5176v2 (February 2010).

- [48] J. Fan, J. Lv, A selective overview of variable selection in high dimensional feature space (invited review article), To appear in *Statistica Sinica*.
- [49] M. J. Wainwright, Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting, *IEEE Trans. Info. Theory* 55 (12) (2009) 5728–5741.
- [50] A. K. Fletcher, S. Rangan, V. K. Goyal, Necessary and sufficient conditions on sparsity pattern recovery, *IEEE Trans. on Information Theory* 55 (12) (2009) 5758–5772.
- [51] M. Akçakaya, V. Tarokh, Shannon theoretic limits on noisy compressive sampling, *IEEE Trans. on Information Theory* 56 (1) (2010) 492–504.
- [52] G. Reeves, M. Gastpar, Sampling bounds for sparse support recovery in the presence of noise, in: *Proceedings IEEE Int. Symp. on Inform. Theory*, 2008, pp. 2187–2191.
- [53] W. Wang, M. J. Wainwright, K. Ramchandran, Information-theoretic limits on sparse support recovery: Dense versus sparse measurements, *IEEE Trans. on Information Theory* 56 (6) (2010) 2967–2979.
- [54] S. Aeron, V. Saligrama, M. Zhao, Information theoretic bounds for compressed sensing, *IEEE Trans. on Information Theory* 56 (10) (2010) 5111–5130.
- [55] V. Saligrama, M. Zhao, Thresholded basis pursuit: An lp algorithm for achieving optimal support recovery for sparse and approximately sparse signals from noisy random measurements, Tech. Rep. arxiv 0809.4883v3 (2010).
- [56] G. Reeves, M. Gastpar, Approximate sparsity pattern recovery: Information-theoretic lower bounds, Tech. Rep. arXiv:1002.4458v1 (February 2010).
- [57] A. Hormati, A. Karbasi, S. Mohajer, M. Vetterli, An estimation theoretic approach for sparsity pattern recovery in the noisy setting, Preprint LCAV-ARTICLE-2009-014.
- [58] P. Tune, S. R. Bhaskaran, S. Hanly, Number of measurements in sparse signal recovery, in: *Proceedings IEEE Int. Symp. on Inform. Theory*, 2009, pp. 16–20.
- [59] K. R. Rad, Sharp sufficient conditions on exact sparsity pattern recovery, Preprint arXiv:0910.0456v3.
- [60] C. Dossal, A necessary and sufficient condition for exact recovery by  $\ell_1$  minimization, Preprint Hal-00164738 (2007) .
- [61] D. Donoho, For most large underdetermined systems of linear equations, the minimal  $\ell_1$  norm near-solution approximates the sparsest near-solution, *Commun. on Pure and Appl. Math.* 59 (7) (2006) 797–829.
- [62] O. N. Feldheim, S. Sodin, A universality result for the smallest eigenvalues of certain sample covariance matrices, *Geometric and Functional Analysis* 20 (1) (2010) 88–123.
- [63] K. Davidson, S. Szarek, Local operator theory, random matrices and Banach spaces, Vol. I, North-Holland, Amsterdam, ed. W.B. Johnson and J. Lindenstrauss, 2001, Ch. 8, pp. 317–366.
- [64] R. J. Muirhead, *Aspects of Multivariate Statistical Theory*, Wiley, New York, 1982.
- [65] J. Matousek, *Lectures on discrete geometry*, Springer Verlag, New York, 2002.

- [66] T. Cai, B. W. Silverman, Incorporating information on neighboring coefficients into wavelet estimation, *Sankhya* 63 (2001) 127–148.
- [67] W. Hoeffding, Probability inequalities for sums of bounded random variables, *Journal of the American Statistical Association* 58 (301) (1963) 1330.