Guaranteed bounds for L1-recovery

Anatoli B. Juditsky, Fatma Kilinc Karzan, Arkadii S. Nemirovski

Accuracy guarantees for ℓ_1-recovery *

Anatoli Juditsky
LJK, Université J. Fourier, B.P. 53, 38041 Grenoble Cedex 9, France
Anatoli.Juditsky@imag.fr

Fatma Kılınç Karzan
Georgia Institute of Technology, Atlanta, Georgia 30332, USA
fkilinc@isye.gatech.edu

Arkadi Nemirovski
Georgia Institute of Technology, Atlanta, Georgia 30332, USA
nemirovs@isye.gatech.edu

August 21, 2010

Abstract

We propose new methods of recovery of sparse signals from noisy observation based on ℓ_1-minimization. They are closely related to the well-known techniques such as Lasso and Dantzig Selector. However, these estimators come with *efficiently verifiable guaranties of performance*. By optimizing these bounds with respect to the method parameters we are able to construct the estimators which possess better statistical properties than the commonly used ones.

We also provide an oracle inequality to justify the proposed algorithms and show how the estimates can be computed using the Basis Pursuit algorithm.

Key words : sparse recovery, linear estimation, oracle inequalities, nonparametric estimation by convex optimization

AMS Subject Classification : 62G08, 90C25

1 Introduction

Recently a several methods of estimation and selection which refer to the ℓ_1-minimization received much attention in the statistical literature. For instance, *Lasso estimator*, which is the ℓ_1-penalized least-squares method is probably the most studied (a theoretical analysis of the Lasso estimator is provided in, e.g., [2, 3, 4, 19, 20, 21, 17, 18], see also the references cited therein). Another, closely related to the Lasso, statistical estimator is the *Dantzig Selector* [7, 2, 16, 17]. To be more precise, let us consider the estimation problem as follows: Assume that an observation $y \in \mathbf{R}^m$ is available where

$$y = Ax + \sigma\xi, \tag{1}$$

where $x \in \mathbf{R}^n$ is an unknown signal and $A \in \mathbf{R}^{m\times n}$ is *the sensing matrix*. We suppose that $\sigma\xi$ is a Gaussian disturbance, where $\xi \sim N(0, I_m)$ (i.e., $\xi = (\xi_1, ..., \xi_n)^T$, where ξ_i are independent normal r.v. with zero mean and unit variance), $\sigma > 0$ being known.

*Research of the second and the third authors was supported by the Office of Naval Research grant # N000140811104.

The Dantzig Selector estimator $\widehat{x}_{\rm DS}$ of the signal x is defined as follows [7]:

$$\widehat{x}_{\rm DS}(y) \in \operatorname*{Argmin}_{v\in\mathbf{R}^n}\{\|v\|_1 \mid \|A^T(Av-y)\|_\infty \le \delta\}$$

where $\delta = O\left(\sigma\sqrt{\ln n}\right)$ is the algorithm parameter. Since $\widehat{x}_{\rm DS}$ is obtained as a solution of linear programm, it is very attractive by its low computational cost. Accuracy bounds for this estimator are readily available. For instance, a well known result about this estimator (cf. [7, Theorem 1.1]) is that if $\delta = O\left(\sigma\sqrt{\ln(n\epsilon^{-1})}\right)$ then

$$\|\widehat{x}_{\rm DS}(y) - x\|_2 \le K\sigma\sqrt{s\log(n\epsilon^{-1})}$$

with probability $1-\epsilon$ if a) the signal x is s-sparse, i.e. has at most s non-vanishing components, and b) the sensing matrix A with unit columns possesses the *Restricted Isometry Property* RIP(λ, k) with parameters $0 < \lambda < \frac{1}{1+\sqrt{2}}$ and $k \ge 3s$. [1] Further, in this case the constant $K = C(1-\lambda)^{-1}$, where C is a moderate absolute constant. This result is quite impressive, due to the established fact (see, e.g. [5, 6]) that there exist $m \times n$ random matrices, with $m < n$, which possess the RIP with probability close to 1, λ close to zero and the value k as large as $O\left(m\ln^{-1}(n/m)\right)$.

On the other hand, Dantzig Selector will become suboptimal when the λ parameter of the RIP is close to 1. Indeed, consider an example of the problem (1) with a 2×2 matrix A with the singular values 1 and ϵ, and the RIP holds with $\delta = 1-\varepsilon^2$. It can be easily seen that if x is aligned with the second right singular vector of A (corresponding to the singular value ε) the error of the Dantzig Selector may be as large as $O(\varepsilon^{-2}\sigma)$, or $O(1-\lambda)^{-1}\sigma)$, while one would expect it to be $O(\varepsilon^{-1}\sigma)$ (up to the logarithmic terms in ϵ). While in our toy example the method can be easily modified to get rid of this drawback, there is no evident way to improve Dantzig Selector in the case of a problem of nontrivial size. The reason to this is that Dantzig Selector (and, to some extent, Lasso) algorithm is really "tailored" to comply with the Resticted Isometry Property, and that property cannot be efficiently verified. For instance, given a matrix A of any "reasonable size", we will not even be able to answer the question if for a given k the corresponding value λ is close to 0 or to 1 in a foreseeable future. New accuracy bounds for Lasso and Dantzig Selector have been proposed recently, which rely upon less restrictive assumptions about the sensing matrix, such as Restricted Eigenvalue [2] or Compatibility [3] conditions (a complete overview of those and several other assumptions with description of how they relate to each other is provided in [19]). However, these assumptions share with the RIP the above drawback: given a problem instance they cannot be efficiently verified. The latter implies that there is no way to provide any guaranties (e.g., confidence sets) of the performance of the proposed procedures. A notable exception from the rule is the *Mutual Incoherence* assumption (see, e.g. [10, 11, 12]) which can be used to computes the accuracy bounds for recovery algorithms: a matrix A with columns of unit ℓ_2-norm and *mutual incoherence* $\mu(A)$ possesses RIP(λ, k) with $\lambda = (m-1)\mu(A)$.[2] Unfortunately, the latter relation implies that $\mu(A)$ should be very small to certify the possibility of accurate ℓ_1-recovery of non-trivial sparse signals, so that the estimates of a "goodness" of sensing for ℓ_1-recovery based on mutual incoherence are very conservative. This "theoretical observation" is supported by numerical experiments – the practical guarantees which may be obtained using the mutual incoherence are generally quite poor even for the problems with nice theoretical properties (cf. [14, 15]).

[1] Recall that RIP(λ, k), same as *uniform uncertainty principle*, means that for any $v \in \mathbf{R}^n$ with at most k non-vanishing components,

$$(1-\lambda)\|v\|_2 \le \|Av\|_2^2 \le (1+\lambda)\|v\|_2$$

This property essentially requires that every set of columns of A with cardinality less than k approximately behaves like an orthonormal system.

[2] The *mutual incoherence* $\mu(A)$ of a sensing matrix $A = [A_1, ..., A_n]$ is computed according to

$$\mu(A) = \max_{i\ne j}\frac{|A_i^TA_j|}{A_i^TA_i}.$$

Obviously, the mutual incoherence can be easily computed even for large matrices.

Recently the authors have proposed a new approach for computing the guaranteed bounds for the "s-goodness" of a sensing matrix A, i.e. the maximal s such that the ℓ_1-recovery of *all* signals with no more than s non-vanishing components is accurate in the case without measurement noise (see [14]). In the present paper we aim to use those *verifiable sufficient conditions* of "goodness" of the matrix A to provide efficiently computable bounds for the error of ℓ_1 recovery procedures. Namely, we consider two following estimation routines:

- *regular recovery*:
$$\widehat{x}_{\rm reg}(y) \in \mathop{\rm Argmin}_{v\in\mathbf{R}^n}\{\|v\|_1 \mid \|H^T(Av-y)|_\infty \le \delta\},$$
where $H \in \mathbf{R}^{m\times n}$ is the *contrast matrix*;

- *penalized recovery*:
$$\widehat{x}_{\rm pen}(y) \in \mathop{\rm Argmin}_{v\in\mathbf{R}^n}\{\|v\|_1 | + \theta\|H^T(Av-y)|_\infty\},$$
where $H \in \mathbf{R}^{m\times n}$ is the contrast matrix and $\theta > 0$ is the parameter of penalization.[3]

We provide the confidence intervals for the error $\|\widehat{x}_{\rm reg} - x\|$ (and $\|\widehat{x}_{\rm pen} - x\|$ of the regular and penalized recovery which involve *efficiently computable* characteristics of the sensing matrix A. Further, we show how the optimal with respect to these bounds contrast matrices can be computed. We also justify our approach with two oracle inequalities and show how the approximation to $\widehat{x}_{\rm reg}$ may be computed using a specific version of the Basis Pursuit algorithm.

2 Accuracy bounds for ℓ_1-Recovery Routines

2.1 Problem statement and notations

We consider an observation $y \in \mathbf{R}^m$

$$y = Ax + u + \sigma\xi, \tag{2}$$

where $x \in \mathbf{R}^n$ is an unknown signal and $A \in \mathbf{R}^{m\times n}$ is *the sensing matrix*. We suppose that $\sigma\xi$ is a Gaussian disturbance, where $\xi \sim N(0, I_m)$ (i.e., $\xi = (\xi_1, ..., \xi_n)^T$, where ξ_i are independent normal r.v. with zero mean and unit variance), $\sigma > 0$ being known, and u is a nuisance parameter known to belong to a given set $\mathcal{U} \subset \mathbf{R}^m$ which we will suppose to be convex, compact and symmetric w.r.t. the origin.

For a vector $x \in \mathbf{R}^n$ and $1 \le s \le n$ we denote x^s the vector obtained from x by setting to 0 all but the s largest in magnitude components of x. We use the notation $\|x\|_{s,p}$ for the usual ℓ_p-norm of x^s (obviously, $\|x\|_{s,\infty} = \|x\|_\infty$). We say that a vector z is *s-sparse* if it has at most a given number s of nonzero entries. Finally, for a set $I \subset \{1, ..., n\}$ we denote by $\bar{I}$ its complement $\{1, ..., n\}\backslash I$; and given $x \in \mathbf{R}^n$, we denote x_I the vector obtained from x by zeroing the entries with indices outside of I, so that $x = x_I + x_{\bar{I}}$.

We say that a vector z is *s-sparse* if it has at most a given number s of nonzero entries.

Our goal is to recover x from y, provided that x is "nearly s-sparse". Specifically, we consider the set

$$X(s, \upsilon) = \{x \in \mathbf{R}^n : \|x - x^s\|_1 \le \upsilon\}.$$

Further, given ϵ and $\sigma > 0$ let us denote

$$\nu(v) = \sup_{u\in\mathcal{U}} u^T v + \sigma\sqrt{2\ln(n\epsilon^{-1})}\|v\|_2, \tag{3}$$

[3] Observe that regular and penalized recoveries can be seen as appropriate modifications of Dantzig Selector and Lasso methods.

and let ν_* be the norm on $\mathbf{R}^n$ conjugate to ν:

$$\nu_*(u) = \max_v \{v^T u |\ \nu(v) \leq 1\}.$$

Since $\mathcal{U}$ is convex, closed and symmetric with respect to the origin, $\nu(\cdot)$, indeed, is a norm.

Let $\widehat{x}(y)$ be an estimate of x (a Borel function $\widehat{x}(\cdot) :\ \mathbf{R}^m \to \mathbf{R}^n$) given the observation y in (2). Give the tolerance level $\epsilon \in (0,1)$ we quantify the risk of $\widehat{x}$ by its worst-case, over $x \in X(s,\upsilon)$, confidence set. Specifically, we define the associated *ϵ-risk of $\widehat{x}$* as

$$\mathrm{Risk}_p(\widehat{x}(\cdot)|\epsilon,\sigma,s,\upsilon) = \inf\left\{\delta : \sup_{x\in X(s,\upsilon),\, u\in\mathcal{U}} \mathrm{Prob}\left\{\|\widehat{x}(y)-x\|_p > \delta\right) \leq \epsilon\right\}.$$

Consider the following condition on the matrices A and $H \in \mathbf{R}^{m\times n}$:

$\mathbf{H}(\gamma,\rho)$: *there are $\gamma_i > 0$ and $\rho_i > 0$, $i = 1,...,n$ such that for all $x \in \mathbf{R}^n$, and $1 \leq i \leq n$,*

$$|x_i| \leq |h_i^T Ax| + \gamma_i \|x\|_1, \tag{4}$$

and

$$\rho_i(\epsilon,\sigma,\mathcal{U};h_i) = \nu(h_i), \tag{5}$$

where $\nu(\cdot)$ is defined in (3).

Observe that $\mathbf{H}(\gamma,\rho)$ implies, in particular, that

$$\mathrm{Prob}\left\{\exists(i \leq n, u \in \mathcal{U}) : |h_i^T(u+\sigma\xi)| \geq \rho_i\right\} \leq \epsilon.$$

2.2 Regular ℓ_1 Recovery

In this section we discuss the properties of the *regular ℓ_1-recovery* $\widehat{x}_{\mathrm{reg}}$ given by:

$$\widehat{x}_{\mathrm{reg}} = \widehat{x}_{\mathrm{reg}}(y) \in \mathop{\mathrm{Argmin}}_{v\in\mathbf{R}^n}\{\|v\|_1 |\ |h_i^T(Av-y)| \leq \delta_i,\ \ i = 1,...,n\}, \tag{6}$$

where y is as in (2), h_i, $i = 1,...,n$ are some vectors in $\mathbf{R}^m$ and $\delta_i > 0$, $i = 1,...,n$.

The starting point of our developments is the following

Proposition 1 *Given an $m \times n$ sensing matrix A, noise intensity σ, uncertainty set $\mathcal{U}$ and a tolerance $\epsilon \in (0,1)$, let the matrix $H = [h_1,...,h_n] \in \mathbf{R}^{m\times n}$ satisfy the condition $\mathbf{H}(\gamma,\rho)$ and let $\delta_i \geq \rho_i = \rho_i(\epsilon,\sigma,\mathcal{U})$, $i \leq n$.*

Then there exists a set Ξ, $\mathrm{Prob}\{\xi \in \Xi\} \geq 1-\epsilon$, of "good" realizations of ξ such that whenever $\xi \in \Xi$, for every $x \in \mathbf{R}^n$, every $u \in \mathcal{U}$ and every subset $I \subset \{1,...,n\}$ such that

$$\gamma_I := \sum_{i\in I} \gamma_i < \tfrac{1}{2}, \tag{7}$$

the regular ℓ_1-recovery $\widehat{x}_{\mathrm{reg}}$, associated with H and $\{\delta_i\}_{i\leq n}$ satisfies:

$$\begin{array}{ll} (a) & \|\widehat{x}_{\mathrm{reg}}(y)-x\|_1 \leq \frac{2\|x_{\bar{I}}\|_1+2\delta_I+2\rho_I}{1-2\gamma_I}; \\ (b) & |\left[\widehat{x}_{\mathrm{reg}}(y)-x\right]_i| \leq \delta_i+\rho_i+\gamma_i\|\widehat{x}_{\mathrm{reg}}(y)-x\|_1 \leq \delta_i+\rho_i+\gamma_i\frac{2\|x_{\bar{I}}\|_1+2\delta_I+2\rho_I}{1-2\gamma_I},\ i = 1,...,n, \end{array} \tag{8}$$

where $\delta_I = \sum_{i\in I}\delta_i$ and $\rho_I = \sum_{i\in I}\rho_i$.

The proof of the proposition is put into the appendix.

Let for $1 \le s \le n$

$$\widehat{\delta}_s = \|[\delta_1, ..., \delta_n]\|_{s,1}, \quad \bar{\rho}_s = \|[\rho_1, ..., \rho_n]\|_{s,1}, \quad \bar{\gamma}_s := \|[\gamma_1, ..., \gamma_n]\|_{s,1},$$
$$\delta = \bar{\delta}_1 = \max_i \delta_i, \qquad \rho = \bar{\rho}_1 = \max_i \rho_i, \qquad \gamma = \bar{\gamma}_1 = \max_i \gamma_i.$$

Corollary 1 *Assume that $\bar{\gamma}_s < \frac{1}{2}$ and $\delta_i \ge \rho_i$, $1 \le i \le n$. Then for all $1 \le p \le \infty$ and $\upsilon \ge 0$:*

$$\mathrm{Risk}_p(\widehat{x}_{\mathrm{reg}}(\cdot)|\epsilon, \sigma, s, \upsilon) \le \phi_s[= \phi(\epsilon, \sigma, \upsilon, s, \bar{\rho}; \bar{\gamma}, \delta)],$$

where

$$\phi_s = \frac{2}{1 - 2\bar{\gamma}_s}\left[\upsilon + \bar{\delta}_s + \bar{\rho}_s\right]^{\frac{1}{p}}\left[\gamma\upsilon + \tfrac{1}{2}[\delta + \rho] + \gamma[\bar{\delta}_s + \bar{\rho}_s] - \bar{\gamma}_s[\delta + \rho]\right]^{\frac{p-1}{p}}. \tag{9}$$

Further, if $s\gamma < 1/2$, we have also

$$\phi_s \le \frac{2s^{\frac{1}{p}}}{1 - 2s\gamma}\left[\gamma\upsilon + \tfrac{1}{2}[\delta + \rho]\right]^{\frac{p-1}{p}}\left[s^{-1}\upsilon + \delta + \rho\right]^{\frac{1}{p}} \le \frac{(2s)^{\frac{1}{p}}}{1 - 2s\gamma}(s^{-1}\upsilon + \delta + \rho). \tag{10}$$

2.3 Penalized ℓ_1 Recovery

Now consider the *penalized ℓ_1-recovery* $\widehat{x}_{\mathrm{pen}}$ as follows:

$$\widehat{x}_{\mathrm{pen}}(y) \in \operatorname*{Argmin}_{v \in \mathbf{R}^n}\{\|v\|_1 + \theta s\|H^T(Av - y)\|_\infty\}, \tag{11}$$

where y is as in (2), an integer $s \le n$ and a positive θ, same as matrix H, are parameters of the construction.

Proposition 2 *Given an $m \times n$ sensing matrix A, an integer $s \le n$, a matrix $H = [h_1, ..., h_n] \in \mathbf{R}^{m \times n}$, positive reals γ_i, ρ_i, $1 \le i \le n$, satisfying the contrast condition $\mathbf{H}(\gamma, \rho)$, and a $\theta > 0$, assume that*

$$\rho = \max_i \rho_i; \quad \gamma = \max_i \gamma_i < (2s)^{-1} \tag{12}$$

and

$$(1 - s\gamma)^{-1} < \theta < (s\gamma)^{-1} \tag{13}$$

and consider the associated estimate $\widehat{x}_{\mathrm{pen}}(\cdot)$.

(i) *For every $\epsilon \in (0, 1)$, there exists a set Ξ, $\mathrm{Prob}\{\xi \in \Xi\} \ge 1 - \epsilon$, of "good" realizations of ξ such that whenever $\xi \in \Xi$, for every $\sigma \ge 0$, every signal $x \in \mathbf{R}^n$ and every $u \in \mathcal{U}$ one has*

$$\begin{array}{ll} (a) & \|\widehat{x}_{\mathrm{pen}}(y) - x\|_1 \le \frac{2\|x - x^s\|_1 + 2s\theta\rho}{\min[\theta(1-s\gamma)-1, 1-\theta s\gamma]} \\ (b) & \|\widehat{x}_{\mathrm{pen}}(y) - x\|_\infty \le \left(\frac{1}{s\theta} + \gamma\right)\|\widehat{x}_{\mathrm{pen}}(y) - x\|_1 + 2\rho \le \frac{2\left(\frac{1}{s\theta} + \gamma\right)\|x - x^s\|_1 + 2(1 + \min[\theta - 1, 1])\rho}{\min[\theta(1-s\gamma)-1, 1-\theta s\gamma]} \end{array} \tag{14}$$

here, same as in Proposition 1, $\rho = \max_i \rho_i(\epsilon, \sigma, \mathcal{U})$.

(ii) *When $\theta = 2$, one has for every $\epsilon \in (0, 1), \sigma \ge 0$, $\upsilon \ge 0$ and $1 \le p \le \infty$:*

$$\mathrm{Risk}_p(\widehat{x}_{\mathrm{pen}}(\cdot)|\epsilon, \sigma, s, \upsilon) \quad \le \quad \frac{1}{1 - 2s\gamma}[2\upsilon + 4s\rho]^{\frac{1}{p}}[(s^{-1} + 2\gamma)\upsilon + 4\rho]^{\frac{p-1}{p}} \le \frac{2s^{\frac{1}{p}}}{1 - 2s\gamma}(s^{-1}\upsilon + 2\rho). \tag{15}$$

2.4 Goodness certificate of A and the origin of condition $\mathbf{H}(\cdot,\cdot)$

Condition $\mathbf{H}(\gamma,\rho)$ and the results of this section merit some comments. Let A be an $m\times n$ matrix and let $s\leq n$ be a positive integer. We say that A is s-good if for all s-sparse $x\in\mathbf{R}^n$ the ℓ_1-recovery $\widehat{x}$,

$$\widehat{x}\in\mathrm{argmin}\,\{\|v\|_1|\ Av=y\}$$

is exact in the case of noiseless observation $y=Ax$. Let us consider the condition:

$$\mathbf{H}_{\mathrm{s}}(\gamma)\qquad \|x\|_{s,1}\leq\|H^TAx\|_\infty+\gamma\|x\|_1\ \text{ for all } x\in\mathbf{R}^n.$$

The existence of a matrix $H\in\mathbf{R}^{M\times n}$ which satisfies $\mathbf{H}_s(\gamma,\rho)$ is intimately related to the necessary and sufficient conditions of s-goodness of A. Namely, whenever the sensing matrix A is s-good, a matrix H with $\gamma<\frac{1}{2}$ always exists. We have the following simple lemma:

Lemma 1 *Suppose that the sensing matrix A is such that the ℓ_1-recovery $\widehat{x}$ is exact for any x which is s-sparse. Then there exist a matrix H which satisfies $\mathbf{H}_s(\gamma)$ with $\gamma<\frac{1}{2}$.*

We refer to H which satisfy $\mathbf{H}_s(\gamma)$ as certificate of s-goodness of A.

In fact, $\mathbf{H}_s(\gamma)$ is not only necessary but also a sufficient condition of s-goodness of the matrix A. As such, condition $\mathbf{H}_s(\gamma)$ may be also used to provide bounds for the accuracy of ℓ_1-recovery. For instance, we have the following analogue of Proposition 1 (the reader may easily reproduce its proof):

Proposition 3 *Given an $m\times n$ sensing matrix A, let the matrix $H=[h_1,...,h_M]\in\mathbf{R}^{m\times M}$, $\nu(h_i)\leq\rho_i$, $i=1,...,M$, satisfy the condition $\mathbf{H}(\gamma)$ and let $\delta_i\geq\rho_i=\rho_i(\epsilon,\sigma,\mathcal{U})$, $i\leq M$. Then the regular recovery $\widehat{x}_{\mathrm{reg}}$ in (6) satisfies*

$$\mathrm{Risk}_1(\widehat{x}_{\mathrm{reg}}(\cdot)|\epsilon,\sigma,s,\upsilon)\leq 2\frac{\delta+\rho+\upsilon}{1-2\gamma},\tag{16}$$

where $\delta=\max_i\delta_i$ and $\rho=\max_i\rho_i$.

Further, a reader familiar with the results of [2, 3, 19] will recognize in $\mathbf{H}_s(\gamma)$ an extended version of the *Compatibility* condition. Indeed, the latter condition on the sensing matrix A means exactly (cf. [3, Section 2.1]) that for some "compatible" norm $\|\cdot\|$ and $\phi>0$ one has

$$\|x^s\|_1\leq\frac{\sqrt{s}}{\phi}\|Ax\|,\ \text{ for all }\ x\in\{x\in\mathbf{R}^n|\ 3\|x^s\|_1\geq\|x-x^s\|_1\}.$$

Obviously, the Compatibility conditions implies that for all $x\in\mathbf{R}^n$:

$$\|x\|_{s,1}\leq\frac{\sqrt{s}}{\phi}\|Ax\|+\tfrac{1}{4}\|x\|_1.\tag{17}$$

One can easily show (we leave the proof as an exercise for the reader) that (17) implies the existence of a matrix $H\in\mathbf{R}^{m\times M}$ such that

$$\|x\|_{s,1}\leq\|H^TAx\|_\infty+\tfrac{1}{4}\|x\|_1,$$

where $H=[h_1,....,h_M]$ is such that $\|h_i\|_*\leq\frac{\sqrt{s}}{\phi}$ (here $\|u\|_*=\max_{\|v\|\leq1}v^Tu$ is the norm conjugate to $\|\cdot\|$).

Though result of Lemma 3 is a good news – it states that when the matrix A allows for exact recovery of s-sparse signals, one can conceive a regular recovery of such signals from noisy observation (2) with the error which satisfies (16), it is of very limited practical interest. In fact, it cannot be used to construct contrast matrices simply because the condition $\mathbf{H}_s(\gamma)$, same as Compatibility condition, cannot be verified in any reasonable time.

On the other hand, the condition $\mathbf{H}(\gamma,\rho)$ which is sufficient for $\mathbf{H}_s(\gamma)$ to be satisfied, can be, as we shall see in an instant, efficiently verified. Furthermore, to the best of our knowledge, the contrast matrix H which satisfying $\mathbf{H}(\gamma,\rho)$ with $\gamma<\frac{1}{2s}$ and some $\rho<\infty$ is nearly the best efficiently computable certificate of s-goodeness of A [14].

3 Efficient construction of the contrast matrix H

The accuracy bounds for ℓ_1-recovery established in Section 2 (e.g. Corollary 1 and Proposition 2) are completely defined by values of two characteristics of the matrix H satisfying condition $\mathbf{H}(\gamma,\rho)$: $\gamma = \max_i \gamma_i$ in (4) and $\rho = \max_i \rho_i$ in 5. One way to design a "good contrast matrix" for a given problem estimation problem (parameterized with A, $\mathcal{U}$, σ and s) may be to choose among the matrices which satisfy $\mathbf{H}(\gamma,\cdot)$ with the desirable value $\gamma < \frac{1}{2s}$ the matrix with the smallest value of ρ.

Namely, consider for $i = 1,...,n$ the system linear inequalities:

$$(I_i): \quad g^T[A]_i \geq 1-\mu,\ |g^T[A]_j| \leq \mu,\ j \neq i. \tag{18}$$

We start with the following observation:

Lemma 2 *Let us fix ρ, $\gamma > 0$ and $1 \leq i \leq n$. Then the following statements are equivalent:*

(i) *there exists a vector h with $\nu(h) \leq \rho$ which satisfies for any $x \in \mathbf{R}^n$:*

$$|x_i| \leq |h^T Ax| + \gamma\|x\|_1; \tag{19}$$

(ii) *there exists a feasible solution g to the inequality (I_i) with $\mu = \gamma$ such that $\nu(g) \leq \rho$. Further, it holds for all $x \in \mathbf{R}^n$:*

$$|x_i| \leq |g^T Ax| + \gamma\|x\|_1;$$

(iii) It holds for all $x \in \mathbf{R}^n$:

$$|x_i| \leq \rho\nu_*(Ax) + \gamma\|x\|_1,$$

where ν_* is the conjugate to ν norm.

Now consider for $i = 1,...,n$ the series of optimization problems

$$(P_i): \quad Opt(i) = \min_{g\in\mathbf{R}^m} \left\{\nu(g) \,\middle|\, g^T[A]_i \geq 1-\mu,\ |g^T[A]_j| \leq \mu,\ j \neq i\right\}. \tag{20}$$

The following result is an immediate consequence of Lemma 2:

Proposition 4 *Let all problems (P_i), $1 \leq i \leq n$, be feasible (and thus solvable), h_i be an optimal solution to (P_i), $i = 1,...,n$ and let*

$$H = [h_1,...,h_n], \quad \varrho_i = \varrho_i(\epsilon,\sigma,\mathcal{U},A;\mu) = Opt(i).$$

(i) *For $i = 1,...,n$ ϱ_i is the minimum of $\rho_i(\epsilon,\sigma,\mathcal{U};h)$ (cf the definition (5)) over all vectors h which satisfy (19) with $\gamma_i \leq \mu$. In particular, $\varrho = \max_i \varrho_i$ is the minimal value of $\rho = \max_i \rho_i$ over the matrices H which satisfy for all $x \in \mathbf{R}^n$:*

$$\|x\|_\infty \leq \|H^T Ax\|_\infty + \mu\|x\|_1.$$

(ii) *One has*

$$\varrho = \min_r \left\{r \,|\, \|x\|_\infty \leq r\nu_*(Ax) + \mu\|x\|_1 \text{ for all } x \in \mathbf{R}^n\right\}. \tag{21}$$

When putting together our findings (Corollary 1, Proposition 1 and Proposition 4) we obtain

Theorem 1 *Let $\mu < (2s)^{-1}$ and all problems (P_i), $1 \leq i \leq n$, be feasible (and thus solvable), h_i be optimal solution to (P_i) and let*

$$H = [h_1,...,h_n], \quad \varrho_i = Opt(i), \quad \varrho = \max_{i=1,...,n} \varrho_i.$$

Then

(i) *the regular* ℓ_1 *recovery* $\widehat{x}_{\rm reg}$*, as in (6) with* $\delta_i \geq \varrho_i$*,satisfies for* $1 \leq p \leq \infty$

$$\mathrm{Risk}_p(\widehat{x}_{\rm reg}, \epsilon) \leq (2s)^{\frac{1}{p}} \frac{s^{-1}\upsilon + (\delta + \varrho)}{1 - 2s\mu}$$

(here $\delta = \max_i \delta_i$*).*

(ii) *the penalized* ℓ_1 *recovery* $\widehat{x}_{\rm pen}$*, as in (11) with* $\theta = 2$*, satisfies for* $1 \leq p \leq \infty$

$$\mathrm{Risk}_p(\widehat{x}_p, \epsilon) \leq 2s^{\frac{1}{p}} \frac{s^{-1}\upsilon + 2\varrho}{1 - 2s\mu};$$

4 Numerical examples

To illustrate the result of the previous section we present here simulation results for the observation model with "input nuisance":

$$y = A(x + v) + \sigma\xi,$$

where $x \in \mathbf{R}^n$ is an unknown sparse signal, the nuisance $v \in \mathcal{V}$ with known $V \subset \mathbf{R}^n$, σ is known and $\xi \in \mathbf{R}^m$ is standard normal $\xi \sim N(0, I_m)$ (in other words, $u = Av$ in the model (2)). We compare the performance of the proposed algorithms (regular and penalized recovery) to that of the "classical" Lasso and Dantzig Selector procedures. To deal with our problem (recovery of the signal x in the presence of the nuisance) those methods were modified as follows: instead of the Lasso estimator we use the estimator

$$\widehat{x}_{\rm las}(y) \in \mathrm{argmin}_{x \in \mathbf{R}^n,\, v \in \mathcal{V}} \left\{ \|x\|_1 + \kappa \|A(x + v) - y\|_2^2 \right\},$$

where the penalization coefficient κ is chosen according to [2, Theorem 4.1]; in its turn, the Dantzig Selector is substituted with

$$\widehat{x}_{\rm DS}(y) \in \mathrm{argmin}_{x \in \mathbf{R}^n,\, v \in \mathcal{V}} \left\{ \|x\|_1 \mid |[A^T(A(x + v) - y)]_i| \leq \delta_i, \;\; i = 1, ..., m \right\}$$

with $\lambda_i = \sigma\sqrt{2\ln(n\epsilon^{-1})}\|[A]_i\|_2$, where ϵ is given (e.g., in what follows $\epsilon = 0.01$).

We present below the simulation results for two setups with $n = 256$:

1. *Gaussian setup*: a 161×256 sensing matrix $A_{\rm Gauss}$ with independent $N(0, 1)$ entries is generated, then its columns are normalized. The nuisance set $\mathcal{V} = \mathcal{V}(L) \subset \mathbf{R}^{256}$ is as follows:

$$\mathcal{V}(L) = \{v \in \mathbf{R}^{256},\; |v_{i+1} - 2v_i + v_{i-1}| \leq L, \text{ for } i = 2, ..., 255,\; v_2 = v_1 = 0\},$$

   where L is a known parameter.

2. *Convolution setup*: a 240×256 sensing matrix $A_{\rm conv}$ is constructed as follows: consider a signal x "living" on $\mathbf{Z}^2$ and supported on the 16×16 grid $\Gamma = \{(i, j) \in \mathbf{Z}^2 :\; 0 \leq i, j \leq 15\}$. We subject such a signal to discrete time convolution with a kernel supported on the set $\{(i, j) \in \mathbf{Z}^2 : -7 \leq i, j \leq 7\}$, and then restrict the result on the 16×15 grid $\Gamma_+ = \{(i, j) \in \Gamma : 1 \leq j \leq 15\}$. This way we obtain a linear mapping $x \mapsto A_{\rm conv}x :\; \mathbb{R}^{256} \to \mathbb{R}^{240}$. The nuisance set $\mathcal{V} = \mathcal{V}(L) \subset \mathbf{R}^{256}$ is composed of zero-mean signals u on Γ which satisfy

$$|[D^2u]_{i,j}| \leq L,$$

   where D is the discrete (periodic) homogeneous Laplace operator:

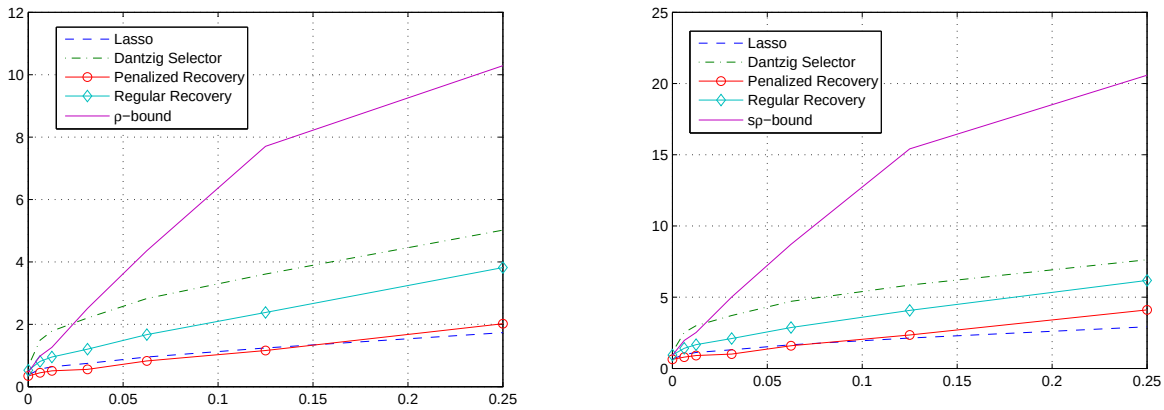
$$[Du]_{i,j} = \frac{1}{4}\left(u_{i,\overline{j-1}} + u_{\overline{i-1},j} + u_{i,\overline{j+1}} + u_{\overline{i+1},j} - 4u_{i,j}\right), \;\; i, j = 1, ..., 16,$$

   with $\overline{i} = i \mod 16$, $\overline{j} = j \mod 16$.

In our simulations we follow the following protocol: given the sensing matrix A, the nuisance set $\mathcal{V}$ and the values of s and σ we compute that contrast matrix H according to (20) with a prescribed $0 < \mu < \frac{1}{2}$. Then $k = 100$ samples of *random* signal x, random nuisance $v \in \mathcal{V}$ and random perturbation ξ are generated, and the mean values ℓ_∞ and ℓ_1 error of recovery are presented on the below plots.[4]

Recovery procedure.are implemented using `Mosek` optimization software [1].

We start with Gaussian setup in which the signal x has $s = 2$ non-vanishing components, randomly drawn, with $\|x\|_1 = 10$. For the penalized and regular recovery algorithms the contrast matrix H is computed according to (20) with $\mu = 0.1$. Given a matrix A, we run $N = 100$ independent simulations of the signal and the noise ξ. On Figure 1 we plot the average error of recovery as a function of the value of the parameter L of the nuisance set $\mathcal{V}$ for fixed $\sigma = 0.1$. For reference, we also trace the corresponding values of ϱ and $s\varrho$, where $\varrho = \max_i \varrho_i$, the optimal values $Opt(i)$ of the problems (20) (solid line on the below plots).



ℓ_∞-error $\qquad$ ℓ_1-error

Figure 1: Mean error of recovery of the sparse signal as a function of the nuisance amplitude L. Gaussian setup parameters: $\sigma = 0.1$, $s = 2$, $\mu = 0.1$, $\|x\|_1 = 10$.

On Figure 2 we plot in the same conditions the average error of recovery as a function of σ for fixed value of $L = 0.01$.

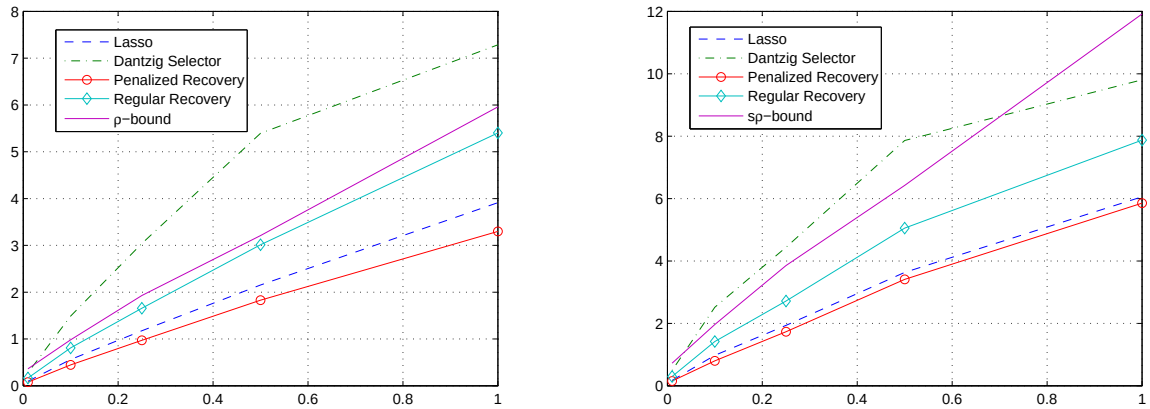
In the next experiment we fix the parameters of estimation and vary the number of non-vanishing components of the signal x. On Figure 3 we present the error of recovery as a function of s. In this experiment $\|x\|_1 = 5s$.

We run the same simulations in the convolution setup. The contrast matrix H for the penalized and regular recovery algorithms is computed according to (20) with $\mu = 0.2$. On Figure 4 we plot the average error of recovery as a function of the value of the parameter L of the nuisance set $\mathcal{V}$ for $\sigma = 0.1$. We provide on Figure 5 the plot of the average error of recovery as a function of σ for fixed value of $L = 0.01$. Finally, on Figure 6 we present the error of recovery as a function of s when parameters of estimators are fixed.

We observe quite different behavior of the recovery procedures in our two setups. In the Gaussian setup the nuisance signal $v \in \mathcal{V}$ does not mask the true signal x, and the performance of the Lasso and Dantzig Selector is quite good in this case. The situation changes dramatically in the convolution setup, where the performance of the Lasso and Dantzig Selector degrades rapidly when the parameter L of the nuisance set increases.[5]

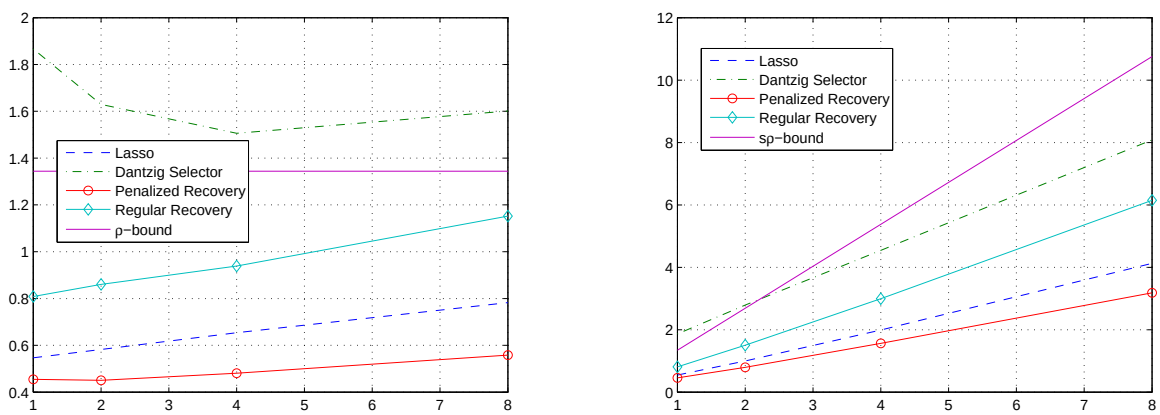
[4] The fact that the sparse signal x is randomly generated is important. Using the techniques of [14] one can verify that in the convolution setup there are signals with only 3 non-vanishing components which cannot be recovered even in the noiseless case (when $\mathcal{V} = \{0\}$ and $\xi = 0$). In other words, the s-goodness characteristic of the corresponding matrix A is equal to 2.

[5] The error plot for these estimators on Figure 4 flatters for higher values of L simply because they always underestimate the signal, and the error of recovery is always less than the corresponding norm of the signal.



ℓ_∞-error $\qquad$ ℓ_1-error

Figure 2: Mean error of recovery of the sparse signal as a function of the noise StD σ. Gaussian setup parameters: $L = 0.01$, $s = 2$, $\mu = 0.1$, $\|x\|_1 = 10$.



ℓ_∞-error $\qquad$ ℓ_1-error

Figure 3: Mean error of recovery of the sparse signal as a function of the number s of non-vanishing components of the signal. Gaussian setup parameters: $L = 0.01$, $\sigma = 0.1$, $\mu = 0.1$, $\|x\|_1 = 5s$.

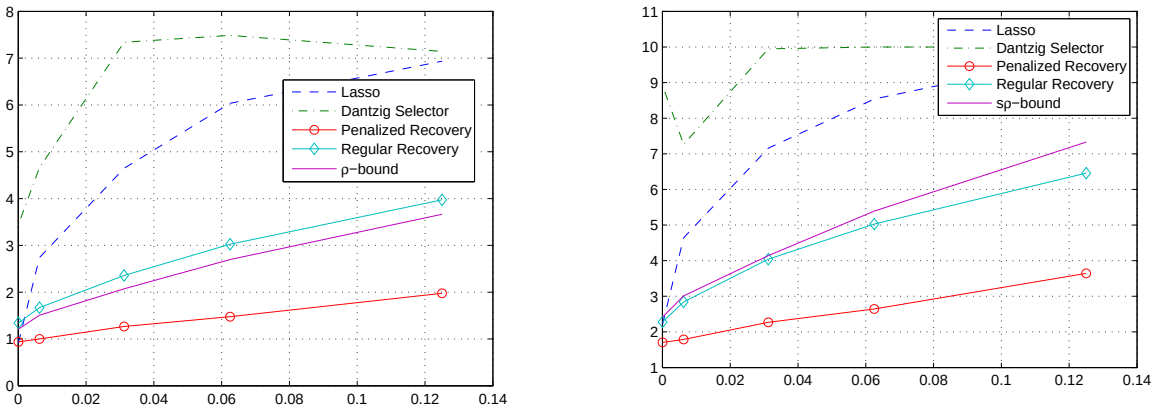
5 Bounding $\varrho(\epsilon, \sigma, \mathcal{U}, A; \mu)$

We address the crucial question of what can be said about the magnitude of the quantity $\varrho(\epsilon, \sigma, \mathcal{U}, A; \mu)$, involved in the risk bounds of Theorem 1. Note that a simple answer to this question is as follows: one can compute the value by solving the corresponding sequence of problems (P_i) in (20). Yet one may be interested to know what may be theoretical guaranties of the bounds of Theorem 1 in certain "reference" situations. In this section we provide two results of this type.

5.0.1 The case of A satisfying the Restricted Isometry Property

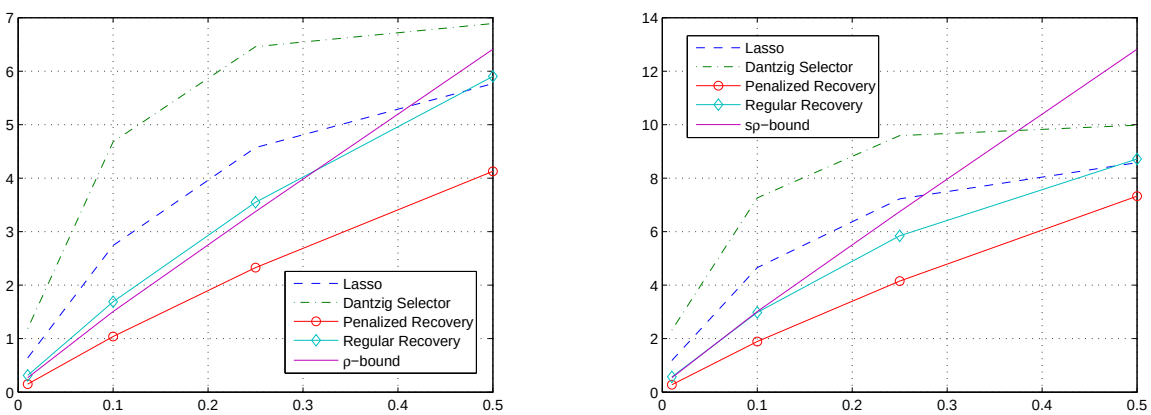
Proposition 5 *Let A satisfy* $\mathrm{RIP}(\delta, k)$ *with some $\delta \in (0, 1)$ and with*

$$k \geq \frac{2\delta^2}{(1-\delta)^2\mu^2} + 1, \tag{22}$$



ℓ_∞-error $\qquad$ ℓ_1-error

Figure 4: Mean error of recovery of the sparse signal as a function of the nuisance amplitude L. Convolution setup parameters: $\sigma = 0.1$, $s = 2$, $\mu = 0.2$, $\|x\|_1 = 10$.



ℓ_∞-error $\qquad$ ℓ_1-error

Figure 5: Mean error of recovery of the sparse signal as a function of the noise StD σ. Convolution setup parameters: $L = 0.01$, $s = 2$, $\mu = 0.2$, $\|x\|_1 = 10$.

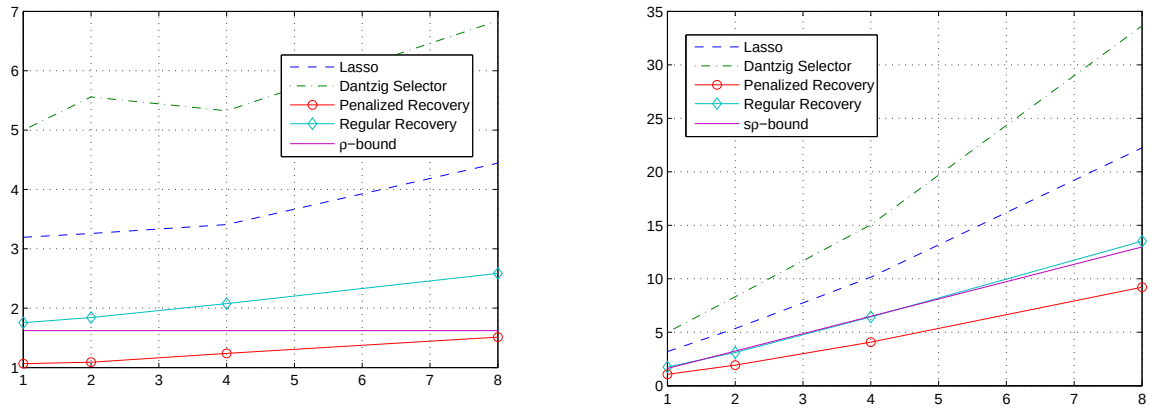
and let $\mathcal{U}$ be contained in the centered at the origin Euclidean ball of radius r. Then

$$\varrho(\epsilon, \sigma, \mathcal{U}, A; \mu) \leq \frac{1}{\sqrt{1-\delta}} \left[r + \sigma \sqrt{2 \ln(n/\epsilon)} \right]. \tag{23}$$

5.0.2 Oracle inequality

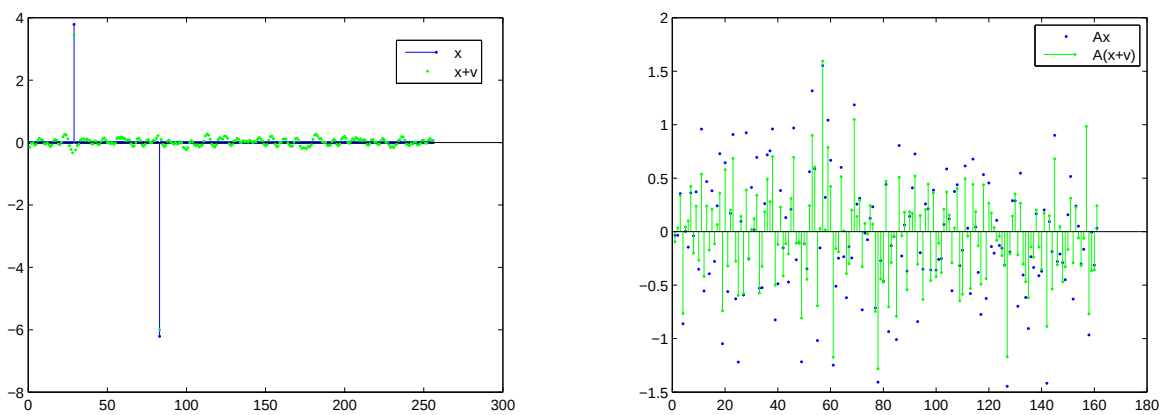
Here we assume that

> $\mathbf{O}(S, \rho^*)$: A is such that for certain $\varrho_* < \infty$ and positive integer S and $\upsilon > 0$ there exists a routine $\mathcal{R}$ with the following property: for every $i \in \{1, ..., n\}$ and every S-element subset $I \subset \{1, ..., n\}$, containing i, the routine, given at input I, i and observation
>
> $$y = Ax + u + \sigma e,$$
>
> associated with *unknown* sparse signal $x \in \mathbf{R}^n$, *known* to be supported on I, produces an



ℓ_∞-error $\qquad$ ℓ_1-error

Figure 6: Mean error of recovery of the sparse signal as a function of the number s of non-vanishing components of the signal. Convolution setup parameters: $L = 0.01$, $\sigma = 0.1$, $\mu = 0.2$, $\|x\|_1 = 5s$.



signal x and nuisance v $\qquad$ image Ax and $A(x+v)$

Figure 7: A typical signal/worst Lasso nuisance. Gaussian setup with parameters: $L = 0.05$, $\sigma = 0.1$, $s = 2$, $\|x\|_1 = 10$, $\mu = 0.1$.

estimate $\mathcal{R}_i(y)$ of x_i such that for any $u \in U$

$$P\left(|\mathcal{R}_i(Ax + u + \sigma e) - x_i| \geq \varrho^*\right) \leq \epsilon.$$

We intend to demonstrate that in this situation for all s in certain range (which depends on S, A and ϱ_*) the quantity ϱ as defined in Theorem 1 is "close" to ϱ_*, so that the performance of the penalized and the regular ℓ_1-recoveries for those s is "close" to the performance of the routine $\mathcal{R}$, postulated in $\mathbf{O}(S, \rho^*)$. This allows us to provide a justification for our approach which is as follows:

> *if there exists a routine which recovers S-sparse signals* with a priori known sparsity pattern *within certain accuracy (measured component-wise), then our recovering routines exhibit "close" performance without any knowledge of the sparsity pattern, albeit in a smaller range of values of the sparsity parameter.*

The precise statement is as follows:

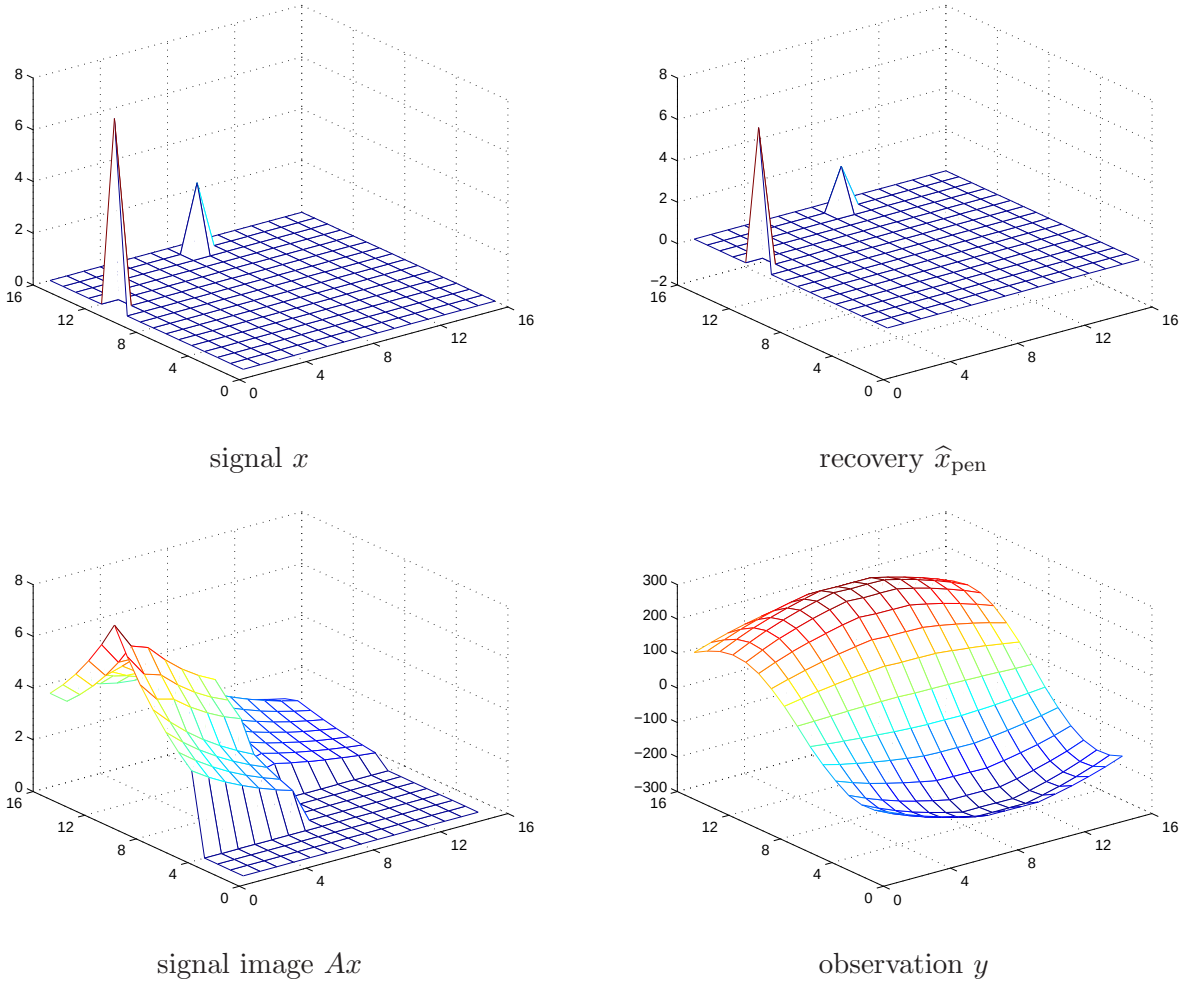


Figure 8: A typical signal/recovery in Convolution setup. Parameters: $L = 0.025$, $\sigma = 0.1$, $s = 2$, $\|x\|_1 = 10$, $\mu = 0.2$.

Proposition 6 *Let $\epsilon \leq 1/16$. Assume that $\mathbf{O}(S, \rho^*)$ takes place, and let μ be a positive real satisfying*

$$\mu \geq \frac{\|A\| \varrho_*}{\sigma \sqrt{2S \ln \epsilon^{-1}}}. \tag{24}$$

Then the quantity ϱ associated with ϵ, σ, $\mathcal{U}$, A and μ according to Theorem 1, admits the bound

$$\varrho \leq 2\varrho_* \sqrt{1 + \frac{\ln n}{\ln(\epsilon^{-1})}}. \tag{25}$$

6 Non-Euclidean matching pursuit algorithm

The Matching Pursuit algorithm for signal recovery is motivated by the desire to provide a reduced complexity alternative to the ℓ_1-recovery problem. Several implementations of Matching Pursuit has been proposed in the Compressive Sensing literature (see, e.g., [11, 10, 12]). They are based on successive Euclidean projections of the signal and the corresponding performance results rely upon the bounds on mutual incoherence parameter $\mu(A)$ of the sensing matrix. We are about to show how the construction of Section 3 can be used to design a specific version of the Matching Pursuit algorithm which we refer to *Non-Euclidean Matching Pursuit (NEMP) algorithm*. The NEMP algorithm can be an interesting option if the ℓ_1-recovery is to be

used repeatedly on the observations obtained with the same sensing matrix A; the numerical complexity of the pursuit algorithm for a given matrix A may only be a fraction of that of the ℓ_1-recovery, especially when used on high-dimensional data.

Suppose that we have in our disposal $\tau \geq 0$ and a matrix $H = [h_1, ..., h_n]$, such that

$$|[I - H^T A]_{ij}| \leq \tau, \quad \text{and} \quad \nu(h_j) \leq \varrho \ \text{ for all } 1 \leq i, j \leq n, \tag{26}$$

where $\nu(\cdot)$ is the norm defined in (3).

Consider a signal $x \in \mathbf{R}^n$ such that $\|x - x^s\|_1 \leq \upsilon$, where, as usual, x^s is the vector obtained from x by replacing all but s largest magnitudes of entries in w with zeros, and let y be an observation as in (2).

Suppose that $s\tau < 1$. Consider the following iterative procedure:

Algorithm 1

1. *<u>Initialization:</u> Set* $v^{(0)} = 0$,
$$\alpha_0 = \frac{\|H^T y\|_{s,1} + s\varrho + \nu}{1 - s\tau}.$$

2. *<u>Step k, $k = 1, 2, ...$:</u> Given* $v^{(k-1)} \in \mathbf{R}^n$ *and* $\alpha_{k-1} \geq 0$, *compute*

   (a) $u = H^T(y - Av^{(k-1)})$ *and* $\Delta \in \mathbf{R}^n$ *with the entries*
$$\Delta_i = \text{sign}(u_i)[|u_i| - \tau\alpha_{k-1} - \varrho]_+, \quad 1 \leq i \leq n$$
   (here $[a]_+ = \max[0, a]$*).*

   (b) *Set* $v^{(k)} = v^{(k-1)} + \Delta$ *and*
$$\alpha_k = 2s\tau\alpha_{k-1} + 2s\varrho + \upsilon. \tag{27}$$
   and loop to step $k + 1$.

3. *The approximate solution found after* k *iterations is* $v^{(k)}$.

Proposition 7 *Assume that* $s\tau < 1$ *and that* $x \in \mathbf{R}^n$ *is such that* $\|x - x^s\|_1 \leq \upsilon$ *with a known in advance value of* υ. *Then there exists a set* Ξ, $\text{Prob}\{\xi \in \Xi\} \geq 1 - \epsilon$, *of "good" realizations of* ξ *such that whenever* $\xi \in \Xi$, *for every* $x \in \mathbf{R}^n$, *every* $u \in \mathcal{U}$, *the approximate solution* $v^{(k)}$ *and the value* α_k *after the* k*-th step of Algorithm 1 satisfy*

$$\begin{array}{ll} (a_k) & \textit{for all } i \ \ v_i^{(k)} \in \text{Conv}\{0; x_i\} \\ (b_k) & \|x - v^{(k)}\|_1 \leq \alpha_k \ \ \textit{and} \ \ \|x - v^{(k+1)}\|_\infty \leq 2\tau\alpha_k + 2\varrho. \end{array}$$

The proof of Proposition 7 is given Appendix A.8.

Note that if $2s\tau\lambda < 1$ then also $s\tau < 1$ and Proposition 7 holds true. Furthermore, by (27) the sequence α_k converges exponentially fast to the limit $\alpha_\infty := \frac{2s\varrho + \upsilon}{1 - 2s\tau}$:

$$\|v^{(k)} - x\|_1 \leq \alpha_k = (2s\tau)^k[\alpha_0 - \alpha_\infty] + \alpha_\infty.$$

Along with the second inequality of (b_k) this implies the bounds:

$$\|v^{(k)} - x\|_\infty \leq 2\tau\alpha_{k-1} + 2\rho \leq \frac{\alpha_k}{s},$$

and, since $\|a\|_p \leq \|a\|_1^{\frac{1}{p}} \|a\|_\infty^{\frac{p-1}{p}}$ for $1 \leq p \leq \infty$,

$$\|v^{(k)} - x\|_p \leq s^{\frac{1-p}{p}} \left((2s\tau)^k[\alpha_0 - \alpha_\infty] + \alpha_\infty \right).$$

The bottom line here is as follows:

Corollary 2 *Let $G = [g_1, ..., g_n]$ be a solution to the series (P_i), $i = 1, ..., n$ of optimization problems (20) with the optimal values $\varrho_i \leq \varrho$. Let $x \in X(s, \upsilon)$, and let y be defined in (2). Then the approximate solution $v^{(t)}$ found by Algorithm 1 after t iterations satisfies*

$$\mathrm{Risk}_p(v^{(t)}|\epsilon, \sigma, s, \upsilon) \leq s^{\frac{1}{p}} \left(\frac{2\varrho + s^{-1}\upsilon}{1 - 2s\mu} + (2s\mu)^t \left[\frac{\varrho + s^{-1}(\|G^T y\|_{s,1} + \upsilon)}{1 - s\mu} - \frac{2\varrho + s^{-1}\upsilon}{1 - 2s\mu} \right] \right).$$

Our concluding remark is on the condition

$$\frac{\mu(A)}{1 + \mu(A)} < \frac{1}{2s}, \tag{28}$$

where $\mu(A)$ is the mutual incoherence of A.This condition is usually used in order to establish convergence results for the Matching Pursuit algorithms (see, e.g. [11, 10, 12]). As it is immediately seen, when $\mu(A)$ is well defined (i.e., all columns in A are nonzero), the matrix $H = [h_1, ..., h_n]$ with the columns

$$h_i = \frac{A_i}{(1 + \mu(A))A_i^T A_i}$$

satisfies for all $i = 1, ..., m$ and $j = 1, ..., n$ the relations

$$|[I - H^T A]_{ij}| \leq \frac{\mu(A)}{1 + \mu(A)}.$$

It follows that H certifies the validity of the condition $\overline{\mathbf{H}}_s(\xi, \sigma, 1)$ with the outlined ξ and with all

$$\varrho \geq \max_i \frac{\nu(A_i)}{(1 + \mu(A))\|A_i\|_2^2}$$

, and thus the above H can be readily used in Matching Pursuit. Note that in the situation in question Corollary 2 recovers some results from [10, 11, 12].

References

[1] Andersen, E. D., Andersen, K. D. *The MOSEK optimization tools manual. Version 5.0* http://www.mosek.com/fileadmin/products/5_0/tools/doc/html/tools/index.html

[2] Bickel, P., Ritov, Y. and Tsybakov, A. B. Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.*, **37**, 17051732 (2009).

[3] Bühlmann, P., van de Geer, S. On the conditions used to prove oracle results for the Lasso *Electron. J. Statist.*, **3**, 1360-1392 (2009).

[4] Bunea, F., Tsybakov, A.B. and Wegkamp, M.H. Sparsity oracle inequalities for the Lasso. *Electron. J. Stat.*, **1**, 169194 (2007).

[5] Candès, E., Romberg, J., Tao T. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inform. Theory,* **52** 489-509 (2006).

[6] Candès, E., Romberg, J., Tao T. Stable signal recovery from incomplete and inaccurate measurements. Comm. Pure Appl. Math., **59** 1207-1223 (2006).

[7] Candès, E. and Tao, T. The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, **35**, 23132351 (2007).

[8] Candès, E. J. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus de l'Acad. des Sci.*, Serie I, 346, 589592 (2008).

[9] Donoho, D., Statistical estimation and optimal recovery. *The Annals of Statistics* **22**, 1, 238-270 (1995).

[10] Donoho, D., Elad, M. Optimally sparse representation in general (non-orthogonal) dictionaries via ℓ_1 minimization. *Proc. Natl. Acad. Sci. USA*, **100**, 21972202 (2003).

[11] Elad, E., Bruckstein, A.M. A generalized uncertainty principle and sparse representation in pairs of $\mathbf{R}^n$ bases. *IEEE Trans. Inform. Theory*, **48**, 25582567 (2002).

[12] Gribonval, R., Nielsen, R. Sparse representations in unions of bases. *IEEE Trans. Inform. Theory*, **49**, 33203325 (2003).

[13] Juditsky, A. B., Nemirovski A.S. Nonparametric estimation by convex programming. Anatoli B. Juditsky and Arkadi S. Nemirovski. *Ann. Statist.*, **37**, 5a, 2278-2300 (2009).

[14] Juditsky, A., Nemirovski, A. On verifiable sufficient conditions for sparse signal recovery via ℓ_1 minimization. To appear *Math. Prog. Ser. B*, see also http://hal.archives-ouvertes.fr/docs/00/32/17/75/PDF/CSNote-Submitted.pdf (2008).

[15] Juditsky, A., Kilinc Karzan, F., Nemirovski, A. Verifiable conditions of ℓ_1-recovery of sparse signals with sign restrictions. To appear *Math. Prog. Ser. B*, see also http://hal.archives-ouvertes.fr/docs/00/37/21/41/PDF/CS_positiv_submitted.pdf (2009).

[16] Koltchinskii, V. The Dantzig selector and sparsity oracle inequalities. *Bernoulli*, **15** 799828 (2009).

[17] Lounici, K. Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electronic Journal of Statistics*, **2**, 90102 (2008).

[18] Meinshausen, N. and Yu, B. Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics*, **37**, 246270 (2009).

[19] van de Geer, S. The deterministic Lasso. In JSM proceedings, (see also http://stat.ethz.ch/research/research_reports/2007/140). American Statistical Association (2007).

[20] Zhang, C.-H., Huang, J. The sparsity and biais of the Lasso selection in highdimensional linear regression. *Ann. Statist.*, bf 36, 15671594 (2008).

[21] Zhang, T. Some sharp performance bounds for least squares regression with L1 regularization. *Ann. Statist.*, **37**, 21092144 (2009).

A Proofs

A.1 Proof of proposition 1

Let

$$\Xi = \{\xi : \ |h_i^T y| \leq \sqrt{2\ln(n\epsilon^{-1})}\|h_i\|_2 \ 1 \leq i \leq n\},$$

so that $\mathrm{Prob}\{\xi \in \Xi\} \geq 1-\epsilon$. Let us fix $\xi \in \Xi$, a set $I \subset \{1,...,n\}$ satisfying (7), a signal $x \in \mathbf{R}^n$ and a realization $u \in \mathcal{U}$ of the nuisance, and let $\widehat{x}$ be the value of the estimate (6) at the observation $y = Ax+u+\sigma\xi$. We are about to verify that $\widehat{x}$ satisfies (8), which, of course, will complete the proof.

Observe that because $\xi \in \Xi$, we have

$$|h_i^T(u+\sigma\xi)| \leq \max_{u' \in \mathcal{U}} |h_i^T u'| + \sigma\sqrt{2\ln(n\epsilon^{-1})}\|h_i\|_2 = \rho_i, \ \ 1 \leq i \leq n.$$

Now, since by assumption $\delta_i \geq \rho_i$, we conclude that $|h_i^T(y - Ax)| \leq \delta_i$ for all i, and thus x is a feasible solution to the optimization problem in (6). It follows that $\|\widehat{x}\|_1 \leq \|x\|_1$, whence

$$\|\widehat{x}_{\bar{I}}\|_1 \leq \|x\|_1 - \|\widehat{x}_I\|_1 = \|x_{\bar{I}}\|_1 + \|x_I\|_1 - \|\widehat{x}_I\|_1 \leq \|x_{\bar{I}}\|_1 + \|x_I - \widehat{x}_I\|_1.$$

Setting $z = \widehat{x} - x$, the resulting inequality reads $\|\widehat{x}_{\bar{I}}\|_1 \leq \|x_{\bar{I}}\|_1 + \|z_I\|_1$, and thus

$$\|z_{\bar{I}}\|_1 \leq \|\widehat{x}_{\bar{I}}\|_1 + \|x_{\bar{I}}\|_1 \leq \|z_I\|_1 + 2\|x_{\bar{I}}\|_1. \tag{29}$$

Further, $|h_i^T A(\widehat{x} - x)| \leq |h_i^T(A\widehat{x} - y)| + |h_i^T(Ax - y)|$. Since $\widehat{x}$ is feasible for the optimization problem in (6), we have $|h_i^T(A\widehat{x} - y)| \leq \delta_i$, and we have already seen that $|h_i^T(Ax - y)| \leq \rho_i$, hence

$$|h_i^T Az| \leq \delta_i + |h_i^T \eta| \leq \delta_i + \rho_i \tag{30}$$

for all $1 \leq i \leq n$. Applying (4) we now get

$$\begin{array}{rcl} \|z_I\|_1 & = & \sum_{i\in I} |z_i| \leq \sum_{i\in I} [|h_i^T Az| + \gamma_i \|z\|_1] \leq \sum_{i\in I} (\delta_i + \rho_i) + \left[\sum_{i\in I} \gamma_i\right] [\|z_I\|_1 + \|z_{\bar{I}}\|_1] \\ & = & \delta_I + \rho_I + \gamma_I [\|z_I\|_1 + \|z_{\bar{I}}\|_1] \leq \delta_I + \rho_I + 2\gamma_I [\|z_I\|_1 + \|x_{\bar{I}}\|_1], \end{array}$$

where the concluding $\leq$ is given by (29). Taking into account that $\gamma_I < \frac{1}{2}$, we get

$$\|z_I\|_1 \leq \frac{\delta_I + \rho_I + 2\gamma_I \|x_{\bar{I}}\|_1}{1 - 2\gamma_I}.$$

Invoking (29) once again, we finally get

$$\|z\|_1 = \|z_I\|_1 + \|z_{\bar{I}}\|_1 \leq 2\|z_I\|_1 + 2\|x_{\bar{I}}\|_1 \leq 2\frac{\delta_I + \rho_I + 2\gamma_I \|x_{\bar{I}}\|_1}{1 - 2\gamma_I} + 2\|x_{\bar{I}}\|_1,$$

and we arrive at (8.a).

3^0. To prove (8.b), we apply (4) to $z = \widehat{x} - x$, thus getting

$$|z_i| \leq |h_i^T Az| + \gamma_i \|z\|_1.$$

As we have already seen, $|h_i^T Az| \leq \delta_i + \rho_i$, and the first $\leq$ in (8.b) follows; the second $\leq$ in (8.b) is then readily given by (8.a). ∎

A.2 Proof of Corollary 1

For $x \in X(s, \upsilon)$, when denoting by I the support of x^s, we have

$$\|x_{\bar{I}}\|_1 \leq \upsilon,\ \delta_I \leq \bar{\delta}_s \leq s\delta,\ \rho_I \leq \bar{\rho}_s \leq s\rho,\ \gamma_I \leq \bar{\gamma}_s \leq s\gamma.$$

Assuming $\bar{\gamma}_s < \frac{1}{2}$, for $\xi \in \Xi$ (which happens with probability $\geq 1 - \epsilon$), (8) implies that for all $u \in \mathcal{U}$ it holds

$$\|\bar{x}_{\rm reg}(y) - x\|_1 \leq \underbrace{\frac{2}{1 - 2\bar{\gamma}_s}[\upsilon + \bar{\delta}_s + \bar{\rho}_s]}_{P}, \quad \text{and} \quad \|\bar{x}_{\rm reg}(y) - x\|_\infty \leq \underbrace{\delta + \rho + 2\gamma \frac{\upsilon + \bar{\delta}_s + \bar{\rho}_s}{1 - 2\bar{\gamma}_s}}_{Q},$$

which combines with the standard bound $\|z\|_p \leq \|z\|_1^{\frac{1}{p}} \|z\|_\infty^{\frac{p-1}{p}}$ to imply (9). When $s\gamma < \frac{1}{2}$, we clearly have

$$\begin{array}{rcl} P & \leq & \frac{2}{1 - 2s\gamma}[\upsilon + s(\delta + \rho)], \\ Q & \leq & \delta + \rho + \frac{2\gamma}{1 - 2s\gamma}[\upsilon + s(\delta + \rho)] = \frac{2}{1 - 2s\gamma}[\gamma\upsilon + \frac{1}{2}[\delta + \rho]], \end{array}$$

and (10) follows due to $\phi_s = P^{\frac{1}{p}} Q^{\frac{p-1}{p}}$. ∎

A.3 Proof of Proposition 2

. ***(i)***: Let us fix $\epsilon \in (0,1)$, and let, same as in the proof of Proposition 1,

$$\Xi = \{\xi : |h_i^T \xi| \le \sqrt{2\ln(n\epsilon^{-1})}\|h_i\|_2,\ 1 \le i \le n\},$$

so that $\mathrm{Prob}\{\xi \in \Xi\} \ge 1-\epsilon$. Let us fix $\xi \in \Xi$, $\sigma \ge 0$, $u \in \mathcal{U}$ and a signal $x \in \mathbf{R}^n$, and let us prove that for these data (14) takes place, this clearly will prove (i). Let us set $\widehat{x} = \widehat{x}(y)$, $z = \widehat{x} - x$, $\eta = u + \sigma\xi$. Let also I be the support of x^s.

Observe that by the origin of $\widehat{x}$, we have

$$\|\widehat{x}\|_1 + s\theta\|H^T(A\widehat{x} - y)\|_\infty \le \|x\|_1 + s\theta\|H^T(Ax - y)\|_\infty = \|x\|_1 + s\theta\|H^T\eta\|_\infty, \tag{31}$$

and

$$\|H^T(A\widehat{x} - y)\|_\infty = \|H^T(Az + Ax - y)\|_\infty \ge \|H^T Az\|_\infty - \|H^T(Ax - y)\|_\infty = \|H^T Az\|_\infty - \|H^T\eta\|_\infty.$$

Combining the resulting inequality with (31), we get

$$\|\widehat{x}\|_1 + s\theta\|H^T Az\|_\infty \le \|x\|_1 + 2s\theta\|H^T\eta\|_\infty \le \|x\|_1 + 2s\theta\rho, \tag{32}$$

where the concluding $\le$ is due to $\xi \in \Xi$ combined with the origin of ρ. Further,

$$\|\widehat{x}\|_1 = \|x + z\|_1 = \|x_I + z_I\|_1 + \|x\bar{I} + z_{\bar{I}}\|_1 \ge \|x_I\|_1 - \|z_I\|_1 + \|z_{\bar{I}}\|_1 - \|x_{\bar{I}}\|_1,$$

which combines with (32) to imply that

$$\|x_I\|_1 - \|z_I\|_1 + \|z_{\bar{I}}\|_1 - \|x_{\bar{I}}\|_1 + s\theta\|H^T Az\|_\infty \le \|x\|_1 + 2s\theta\rho,$$

or, which is the same,

$$\|z_{\bar{I}}\|_1 - \|z_I\|_1 + s\theta\|H^T Az\|_\infty \le 2\|x_{\bar{I}}\|_1 + 2s\theta\rho. \tag{33}$$

Since, by (4),

$$\|z\|_\infty \le \|H^T Az\|_\infty + \gamma\|z\|_1, \tag{34}$$

we now get due to $\|z_I\|_1 \le s\|z\|_\infty$:

$$(1 - s\gamma)\|z_I\|_1 - s\gamma\|z_{\bar{I}}\|_1 - s\|H^T Az\|_\infty \le 0.$$

Multiplying the latter inequality by θ and summing up with (33), we get

$$[\theta(1 - s\gamma) - 1]\|z_I\|_1 + (1 - \theta s\gamma)\|z_{\bar{I}}\|_1 \le 2\|x_{\bar{I}}\|_1 + 2s\theta\rho.$$

In view of condition (13), the coefficients in the left hand side are positive, and (14.a) follows.

To prove (14.b), note that from (32) it follows that

$$\|H^T Az\|_\infty \le \frac{1}{s\theta}[\|x\|_1 - \|\widehat{x}\|_1] + 2\rho \le \frac{1}{s\theta}\|z\|_1 + 2\rho$$

which combines with (34) to imply that

$$\|z\|_\infty \le \frac{1}{s\theta}\|z\|_1 + 2\rho + \gamma\|z\|_1.$$

Recalling that $z = \widehat{x} - x$ and invoking (14.a), (14.b) follows.

(ii): Assuming that $x \in X(s,\upsilon)$ and $\theta = 2$ we obtain from (14) that uniformly on $\xi \in \Xi$ and $u \in \mathcal{U}$

$$\|\bar{x}_{\mathrm{reg}}(y) - x\|_1 \le \underbrace{\frac{2\upsilon + 4s\rho}{1 - 2s\gamma}}_{P},\quad \|\bar{x}_{\mathrm{reg}}(y) - x\|_\infty \le \underbrace{\frac{(s^{-1} + 2\gamma)\upsilon + 4\rho}{1 - 2s\gamma}}_{Q}.$$

Using, as in the proof of Corollary 1, the standard bound

$$\|z\|_p \le \|z\|_1^{\frac{1}{p}}\|z\|_\infty^{\frac{p-1}{p}} \le P^{\frac{1}{p}}Q^{\frac{p-1}{p}}$$

we come to (15), where the second $\le$ is due to the fact that in the premise of the theorem $\gamma \le \frac{1}{2s}$. ∎

A.4 Proof of Lemma 1

Recall (cf, e.g., Theorem 2.1 in [14]) that a necessary and sufficient condition for an $m \times n$ matrix A to be s-good, i.e. to allow for exact ℓ_1-recovery of s-sparse vectors from the noiseless observation, is the existence of $\gamma < \frac{1}{2}$ such that for any set $I \subset \{1, ..., n\}$ of cardinality $\leq s$ and any collection of signs $\chi^{(k)} = \{\chi_i^{(k)} \in \{\pm 1\},\ i \in I\}$, $k = 1, ..., M[= \binom{n}{s} 2^s]$, there is a vector $h_k \in \mathbf{R}^m$ such that

$$\chi_i^{(k)} h_k^T [A]_i \geq 1 - \gamma \ \text{ and } \ \chi_i^{(k)} h^T [A]_j \leq \gamma \ \text{ for all } \ j \neq i. \tag{35}$$

Let $H = [h_1, ..., h_M] \in \mathbf{R}^{m \times M}$. It is obvious that H satisfies $\mathbf{H}_s(\gamma)$ with $\gamma < \frac{1}{2}$. Indeed, for any $x \in \mathbf{R}^n$ let I be the support set of s largest in magnitude components of x and let $\chi^{(x)} \in \{\chi_i^{(x)}, i \in I\}$ the collection of signs of x_i, $i \in I$. We have for the corresponding h_x:

$$\|x\|_{s,1} - x^T A^T h_x = \sum_{i \in I} \chi_i^{(x)} x_i - \sum_{i=1}^{n} (A^T h_x)_i x_i \leq \gamma \|x\|_1,$$

thus

$$\|x\|_{s,1} \leq h_x^T A x + \gamma \|x\|_1 \leq \|H^T A x\|_\infty + \gamma \|x\|_1.$$

∎

A.5 Proof of Lemma 2

(i)⇔(ii) Because of the homogeneity of (19), it is obviously equivalent to

$$|x_i| \leq |h^T A x| + \gamma, \ \text{ for all } x \text{ such that } \|x\|_1 \leq 1.$$

Then we can write:

$$\begin{aligned}
\gamma &= \max_x \left\{ e_i^T x - |h^T A x| \mid \|x\|_1 \leq 1 \right\} = \max_x \min_g \left\{ e_i^T x - g^T A x \mid \|x\|_1 \leq 1,\ g \in [-h, h] \right\} \\
&= \min_g \max_x \left\{ (e_i - A^T g)^T x \mid \|x\|_1 \leq 1,\ g \in [-h, h] \right\} = \min_g \left\{ \|e_i - A^T g\|_\infty \mid g \in [-h, h] \right\} \\
&= \min_g \left\{ [A^T g]_i - 1,\ |[A^T g]_j|,\ j \neq i \mid g \in [-h, h] \right\}.
\end{aligned}$$

The latter exactly means that there is $g \in [-h, h]$ (and thus $\nu(g) \leq \nu(h) \leq \rho$) which is feasible to (I_i) with $\mu = \gamma$. By inverting the above chain of equalities we conclude that $h = g$ will satisfy (19) if g is feasible to (I_i) with $\mu = \gamma$.

(ii)⇔(iii) This is immediate too: if (I_i) is feasible for some g, $\nu(g) \leq \rho$ and $\mu = \gamma$, then, due to the compactness of the ball $\{g \in \mathbf{R}^m | \ \nu(g) \leq \rho\}$,

$$\begin{aligned}
\gamma &\geq \min_g \left\{ \|e_i - A^T g\|_\infty \mid \nu(g) \leq \rho \right\} = \min_g \max_x \left\{ (e_i - A^T g) x \mid \nu(g) \leq \rho,\ \|x\|_1 \leq 1 \right\} \\
&= \max_x \min_g \left\{ e_i^T x - g^T A x \mid \nu(g) \leq \rho,\ \|x\|_1 \leq 1 \right\} = \max_x \left\{ e_i^T x - \rho \nu_*(Ax) \mid \|x\|_1 \leq 1 \right\},
\end{aligned}$$

and we come to $|x_i| \leq \rho \nu_*(Ax) + \gamma$ for all x such that $\|x\|_1 \leq 1$. ∎

A.6 Proof of Proposition 5

By (ii) of Proposition 4, all we need to prove is the bound

$$\|x\|_\infty \leq \varrho\nu_*(Ax) + \gamma\|x\|_1$$

for all $x \in \mathbf{R}^n$. We are in the situation where the unit ball $\mathcal{U} + \sigma\sqrt{2\ln(n/\epsilon)}B$ of the norm $\nu_*(\cdot)$, where B is the unit Euclidean ball, is contained in the Euclidean ball of radius $\lambda = r + \sigma\sqrt{2\ln(n\epsilon^{-1})}$, whence $\nu_*(z) \geq \|z\|_2/\lambda$ for all z. In other words, it suffices to prove that for all $x \in \mathbf{R}^n$,

$$\|x\|_\infty \leq \varrho\lambda^{-1}\|Ax\|_2 + \gamma\|x\|_1. \tag{36}$$

When proving this relation, we can assume w.l.o.g. that $|x_1| \geq |x_2| \geq ... \geq |x_n|$. Setting $t = \text{floor}(k/2)$, let x^0 be the vector obtained from x by zeroing all entries except for the first one, x^1 be the vector obtained from x by zeroing all entries with indices outside of $2, 3, ..., t$, x^2 be obtained from x by zeroing all entries not in $t+1, ..., 2t$, x^3 be obtained from x by zeroing all entries not in $2t+1, 2+2, ..., .3t$, and so on. We clearly have $\|x^2\|_2 \leq t^{-1/2}\|x^0 + x^1\|_1$ and $\|x^j\|_\infty \leq t^{-1/2}\|x^{j-1}\|_1$, $j = 3, ..., q$, where q is such that $\sum_{i=0}^q x^i = x$. Setting $\|Ax\|_2 = \alpha$ and $\|A(x_0 + x_1)\|_2 = \beta$, we have

$$\begin{aligned}
\alpha\beta &= \|Ax\|_2\|A(x^0 + x^1)\|_2 \geq (Ax)^T A(x^0 + x^1) \\
&= (x^0 + x^1)^T A^T A(x^0 + x^1) + \sum_{j=2}^{1}(x^0 + x^1)^T A^T Ax^j \geq \beta^2 - \sum_{j=2}^{q}\delta\|x^0 + x^1\|_2\|x^j\|_2,
\end{aligned}$$

where the last $\geq$ is given by the following well known fact [8]: if A is RIP(δ, k) and u, v are supported on a common set of indices I of cardinality k and are orthogonal, we have $|u^T A^T Av| \leq \delta\|u\|_2\|v\|_2$. It follows that

$$\begin{aligned}
\alpha\beta &\geq \beta^2 - \delta\|x^0 + x^1\|_2\sum_{j=2}^{q}\|x^j\|_2 \geq \beta^2 - \delta t^{-1/2}\|x^0 + x^1\|_2\sum_{j=2}^{q}\|x^{j-1}\|_1 \\
&\geq \beta^2 - \delta t^{-1/2}\|x^0 + x^1\|_2\sum_{j=1}^{q}\|x^j\|_1 \geq \beta^2 - \delta t^{-1/2}\|x^0 + x^1\|_2\|x\|_1.
\end{aligned}$$

Hence

$$\beta \leq \alpha + \frac{\delta\|x\|_1\|x^0 + x^1\|_2}{\sqrt{t}\beta} \leq \alpha + \frac{\delta\|x\|_1}{\sqrt{t(1-\delta)}},$$

and

$$\|x\|_\infty \leq \|x^0 + x^1\|_2 \leq \frac{\beta}{\sqrt{1-\delta}} \leq \frac{\alpha}{\sqrt{1-\delta}} + \frac{\delta\|x\|_1}{(1-\delta)\sqrt{t}}.$$

Recalling that $\alpha = \|Ax\|_2$ and $t = \text{floor}(k/2)$, (36) follows. ∎

A.7 Proof of Proposition 6

Proof. We start with an appropriate translation of $\mathbf{O}(S, \rho^*)$. Let $i \in \{1, ..., n\}$ and let $I \ni i$ be a subset of $\{1, ..., n\}$ of cardinality S. Let $\mathbf{R}^{(S)}$ be the linear space of all vectors supported on I, and let $X_R = \{x \in \mathbf{R}^{(S)} | \ \|x\|_2 \leq R\}$. Assume that we are given a noisy observation y of the signal $z = (x, u) \in (X_R, \mathcal{U})$, and that we want to recover from this observation the linear form x_i of the signal. From $\mathbf{O}(S, \rho^*)$ it follows that there exists a recovering routine such that for every $x \in X_R$ and $u \in \mathcal{U}$ the probability of recovering error to be $\geq \varrho_*$ is $\leq \epsilon$. Assuming $\epsilon \leq 1/16$ and applying the celebrated result of Donoho [9] there exists a *linear*

estimate $\phi_R^T y$ such that for every $x \in X_R$ and $u \in \mathcal{U}$ the probability for the error of this estimate to be $\geq 1.22\varrho_*$ is $\leq \epsilon$. Moreover (cf Proposition 4.2 of [13]), one can pick ϕ_R such that

$$\forall x \in X_R,\ u \in \mathcal{U},\ \left\{ \begin{array}{rcl} P(\phi_R^T[u+\sigma e + A_I x] - x_i > 1.22\varrho_*) & \leq & \epsilon/2, \\ P(\phi_R^T[u+\sigma e + A_I x] - x_i < -1.22\varrho_*) & \leq & \epsilon/2, \end{array} \right.$$

where A_I is the $m \times S$ submatrix of A comprised of columns with indexes from I. Setting $p(R) = \max_{u\in\mathcal{U}} |\phi_R^T u|$ and $r(R) = \|A_I^T \phi_R - e_i\|_2$, where e_i is the i-th basic orth (so that $x_i = e_i^T x$), we conclude that

$$\begin{array}{rcl} P(\sigma\phi_R^T e > 1.22\varrho_* - Rr(R) - p(R)) & \leq & \epsilon/2, \\ P(\sigma\phi_R^T e < -1.22\varrho_* + Rr(R) + p(R)) & \leq & \epsilon/2. \end{array}$$

Hence, denoting by $\mathrm{erfinv}(\epsilon)$ the value of the inverse error function at ϵ, we obtain

$$\mathrm{erfinv}\left(\frac{\epsilon}{2}\right)\sigma\|\phi_R\|_2 \leq 1.22\varrho_* - Rr(R) - p(R).$$

It follows that as $R \to \infty$ ϕ_R remains bounded and $r(R) = \|e_i - A_I^T a_R\|_2 \to 0$. Thus, there exists a sequence $R_k \to +\infty$, of values of R such that ϕ_{R_t} goes to a limit ϕ as $k \to \infty$, and this limit satisfies the relations

$$\mathrm{erfinv}\left(\frac{\epsilon}{2}\right)\sigma\|\phi\|_2 \leq 1.22\varrho_*, \ \text{ and } A_I^T\phi = e_i.$$

Taking into account that $\mathrm{erfinv}\left(\frac{\epsilon}{2}\right) \geq 0.92\sqrt{\ln(1/\epsilon)}$ when $\epsilon \leq 1/16$, we arrive at the following result:

Lemma 3 *In the premises of $\mathbf{O}(S,\rho^*)$, for every $i \leq n$ and every S-element subset $I \ni i$ of $\{1,...,n\}$ there exists $\phi \in \mathbf{R}^m$ such that $\phi^T A_i = 1$, $\phi^T A_j = 0$ for all $j \in I$, $j \neq i$, and*

$$\max_{u\in\mathcal{U}} |u^T\phi| + \sigma\sqrt{\ln(\epsilon^{-1})}\|\phi\|_2 \leq \sqrt{2}\varrho_*.$$

We claim that in this case for all $x \in \mathbf{R}^n$ it holds:

$$\|x\|_\infty \leq 2\sqrt{1+\frac{\ln n}{\ln(\epsilon^{-1})}}\,\varrho_*\nu_*(Ax) + \frac{\varrho_*\|A\|}{\sigma\sqrt{2S\ln(\epsilon^{-1})}}\|x\|_1. \tag{37}$$

Note that this claim combines with the result of Proposition 4 to imply (25). To prove (37), let us fix x; w.l.o.g. we may assume that $x_1 = |x_1| \geq |x_2| \geq ... \geq |x_n|$. Let us set $x^0 = [x_1;...;x_S;0;...;0]$ and $x^1 = [0;...;0;x_{S+1};...;x_n]$. Observe that

$$\|x^1\|_2 \leq \|x^1\|_\infty^{1/2}\|x^1\|_1^{1/2} \leq S^{-1/2}\|x^0\|_1^{1/2}\|x^1\|_1^{1/2} \leq \frac{1}{2}S^{-1/2}\|x\|_1.$$

By Lemma 3 there exists $\phi \in \mathbf{R}^m$ such that

$$\|\phi\|_2 \leq \frac{\sqrt{2}\varrho_*}{\sigma\sqrt{\ln(\epsilon^{-1})}} \ \text{ and } \ \nu(\phi) \leq 2\varrho_*\sqrt{1+\frac{\ln(n)}{\ln(\epsilon^{-1})}},$$

with $\phi^T A_1 = 1$ and $\phi^T A_i = 0$ for $2 \leq i \leq S$. We have

$$\begin{array}{rcl} \nu(\phi)\nu_*(Ax) & \geq & \phi^T Ax = \phi^T Ax^0 + \phi^T Ax^1 = x_1 + \phi^T Ax^1 = \|x\|_\infty + \phi^T Ax^1 \\ & \geq & \|x\|_\infty - \|\phi\|_2\|Ax^1\|_2 \geq \|x\|_\infty - \frac{1}{2}\|A\|S^{-1/2}\frac{\sqrt{2}\varrho_*}{\sigma\sqrt{\ln(\epsilon^{-1})}}, \end{array}$$

as required in (37). ∎

A.8 Proof of Proposition 7

Let for $\epsilon \in (0,1)$, same as in the proofs of Proposition 1 and 2, $\Xi = \{\xi : |h_i^T\xi| \le \sqrt{2\ln(n\epsilon^{-1})}\|h_i\|_2,\ 1 \le i \le n\}$, so that, indeed, $\mathrm{Prob}\{\xi \in \Xi\} \ge 1-\epsilon$. Then for $\eta = y - Ax = u + \sigma\xi$, by the definition (3) of the norm ν and because $\nu(h_i) \le \varrho$, we have $\|H^T\eta\|_\infty \le \varrho$ for all $u \in \mathcal{U}$ and $\xi \in \Xi$.

Now, let us proceed by induction. First, let us show that (a_{k-1}, b_{k-1}) implies (a_k, b_k). Thus, assume that (a_{k-1}, b_{k-1}) holds true. Let $z^{(k-1)} = x - v^{(k-1)}$. By (a_{k-1}), $z^{(k-1)}$ is supported on the support of x. Note that

$$\begin{array}{rcl} z^{(k-1)} - u & = & x - v^{(k-1)} - H^T(y - Av^{(k-1)}) = (I - H^TA)(x - v^{(k-1)}) - H^T\eta \\ & = & (I - H^TA)z^{(k-1)} - H^T\eta, \end{array}$$

Then by (26) for any $1 \le i \le n$,

$$-\tau\sum_j |z_j^{(k-1)}| - \varrho \le z_i^{(k-1)} - u_i \le \tau\sum_j z_j^{(k-1)} + \varrho,$$

consequently,

$$-\gamma := -\tau\alpha_{k-1} - \varrho \le z_i^{(k-1)} - u_i \le \tau\alpha_{k-1} + \varrho := \gamma, \tag{38}$$

so that the segment $S_i = [u_i - \gamma,\ u_i + \gamma]$ of the width $\ell = 2\tau\alpha_{k-1} + 2\varrho$, covers $z_i^{(k-1)}$, and the closest to zero point of this interval is

$$\widetilde{\Delta}_i = \left\{ \begin{array}{ll} [u_i - \gamma]_+, & u_i \ge 0, \\ -[|u_i| - \gamma]_+, & u_i < 0, \end{array} \right.$$

that is, $\widetilde{\Delta}_i = \Delta_i$ for all i. Since the segment S_i covers $z_i^{(k-1)}$ and Δ_i is the closest to 0 point in S_i, while the width of S_i is at most ℓ, we clearly have

$$\begin{array}{ll} (a) & \Delta_i \in \mathrm{Conv}\left\{0, z_i^{(k-1)}\right\}, \\ (b) & |z_i^{(k-1)} - \Delta_i| \le \ell. \end{array} \tag{39}$$

Since (a_{k-1}) is valid, (39.a) implies that

$$v_i^{(k)} = v_i^{(k-1)} + \Delta_i \in \left[v_i^{(k-1)} + \mathrm{Conv}\left\{0, x_i - v_i^{(k-1)}\right\}\right] \subseteq \mathrm{Conv}\{0, x_i\},$$

and (a_k) holds. Further, let I be the support of x^s. Relation (a_k) clearly implies that $|z_i^{(k)}| \le |x_i|$, and we can write due to (39.b):

$$\begin{array}{rcl} \|x - v^{(k)}\|_1 & = & \sum_{i\in I} |x_i - [v_i^{(k-1)} + \Delta_i]| + \sum_{i\notin I} |z_i^{(k)}| \\ & \le & \sum_{i\in I} |z_i^{(k-1)} - \Delta_i| + \sum_{i\notin I} |x_i| \le s\ell + \mu = \alpha_k. \end{array}$$

Since by (39.b)

$$\|x - v^{(k)}\|_\infty \;=\; \|x - v^{(k-1)} - \Delta\|_\infty \le \ell = 2\tau\alpha_{k-1} + 2\varrho,$$

we conclude (b_k). The induction step is justified.

It remains to show that (a_0, b_0) holds true. Since (a_0) is evident, all we need is to justify (b_0). Let

$$\alpha_* = \|x\|_1,$$

and let $u = H^T y$. Same as above (cf. (38)), we have for all i:

$$|x_i - u_i| \leq \tau\alpha_* + \varrho.$$

Then

$$\alpha_* = \sum_{i\in I} |x_i| + \sum_{i\notin I} |x_i| \leq \sum_{i\in I} [|u_i| + \tau\alpha_* + \varrho] + \upsilon \leq \|u\|_{s,1} + s\tau\alpha_* + s\varrho + \upsilon.$$

Hence

$$\alpha_* \leq \alpha_0 = \frac{\|u\|_{s,1} + s\varrho + \upsilon}{1 - s\tau},$$

which implies (b_0). ∎