Information and Inference: A Journal of the IMA (2013) Page 1 of 27

doi:10.1093/imaiai/iarxxx

Damping Noise-Folding and Enhanced Support Recovery in Compressed Sensing

MARCO ARTINA,
Faculty of Mathematics, Technische Universität München, Boltzmannstrasse 3, 85748 Garching, Germany
marco.artina@ma.tum.de
MASSIMO FORNASIER
Faculty of Mathematics, Technische Universität München, Boltzmannstrasse 3, 85748 Garching, Germany
massimo.fornasier@ma.tum.de

AND

STEFFEN PETER
Faculty of Mathematics, Technische Universität München, Boltzmannstrasse 3, 85748 Garching, Germany
steffen.peter@ma.tum.de

[Received on XX XX XXXX; revised on XX XX XXXX; accepted on XX XX XXXX]

The practice of compressive sensing suffers importantly in terms of the efficiency/accuracy trade-off when acquiring noisy signals prior to measurement. It is rather common to find results treating the noise affecting the measurements, avoiding in this way to face the so-called *noise-folding* phenomenon, related to the noise in the signal, eventually amplified by the measurement procedure. In this paper we present a new decoding procedure, combining ℓ_1-minimization followed by a selective least p-powers, which not only is able to reduce this component of the original noise, but also has enhanced properties in terms of support identification with respect to the sole ℓ_1-minimization. We prove such features, providing relatively simple and precise theoretical guarantees. We additionally confirm and support the theoretical estimates by extensive numerical simulations, which give a statistics of the robustness of the new decoding procedure with respect to more classical ℓ_1-minimization.

Keywords: Noise folding in compressed sensing, support identification, ℓ_1-minimization, selective least p-powers, iterative thresholding, phase transitions.

2000 Math Subject Classification: 94A12, 94A20, 65F22, 90C05, 90C30

1. Introduction

Compressive sensing focuses on the robust recovery of nearly sparse vectors from the minimal amount of measurements obtained by a linear process. So far, a vast literature appeared considering problems where deterministic or random noise is added after the measurement process, while it is not strictly related to the signal. One typically considers model problems of the type

$$y = Ax + w \tag{1.1}$$

where $x \in \mathbb{R}^N$ is a nearly sparse vector, $A \in \mathbb{R}^{m\times N}$ is the linear measurement matrix, $y \in \mathbb{R}^m$ is the result of the measurement, and w is a white noise vector affecting the measurements. However, in practice it is very uncommon to have a signal detected by a certain device, totally free from some external noise. Therefore, it is reasonable to consider the more realistic model

$$y = A(x+n)+w,$$

instead of (1.1) where $n \in \mathbb{R}^N$ is the noise on the original signal.

The recent work [1] shows how the measurement process actually causes the *noise-folding phenomenon*, which implies that the variance of the noise on the original signal is amplified by a factor of $\frac{N}{m}$, additionally contributing to the measurement noise, playing to our disadvantage in the recovery phase. Indeed this phenomenon may significantly reduce in practice the potential advantages of compressed sensing in terms of the trade-off between robustness and efficient compression, with respect to other more traditional subsampling encoding methods [9]. To control this effect, as proposed in [1], one may tune the linear measurement process in order to a priori filter the noise. However, this strategy requires to have a precise knowledge of the noise statistics and to design proper filters. In this paper we shall follow a *blind-to-statistics* approach, which does not modify the measurements. In particular, we propose a new decoding procedure, combining ℓ_1-minimization and selective least p-powers, to reduce the noise component affecting the signal, but also to enhance the support identification. In fact, for certain applications, such as radar [15], the support recovery can be even more relevant than an accurate estimate of the signal values.

The paper is organized as follows. In the next section, we concisely recall the pertinent features of the theory of compressive sensing. In Section 3, we shall describe the limitations of ℓ_1-minimization when noise on the signal is present. Afterwards, as an alternative, we propose the linearly constrained minimization of the selective p-potential functional, and show that, under certain conditions, it performs significantly better than ℓ_1-minimization. Section 4 recalls the main properties of a very robust and efficient algorithm to perform selective least p-powers. Finally, in Section 5, we report the results of extensive numerical experiments, which we made to illustrate and support our theoretical guarantees.

2. Compressive sensing

We are able to uniquely and robustly identify the solution $x \in \mathbb{R}^N$ of the linear system $Ax = y$ for an arbitrary given measurement vector $y \in \mathbb{R}^m$, if $A \in \mathbb{R}^{m\times N}$ has rank m, and $m = N$. However, in many applications, we either are not able to take enough measurements, or we are interested in taking much fewer measurements to save costs or time, i.e., $m \ll N$. The theory of compressive sensing studies this scenario under some restrictions, and assumes that the original signal x is nearly sparse. In this section we recall concisely terms and principles of this theory, and we refer to some of the known tutorials for more details [3, 5, 11, 13].

In compressive sensing, we call the matrix A the *encoder* which transforms the N-

dimensional signal x into to the *measurement vector* $y \in \mathbb{R}^m$ of dimension $m \ll N$. Further, we assume A to have rank m from now on in this article. In practice we do not know x and wonder if it is possible to recover it somehow robustly by an efficient nonlinear *decoder* $\Delta\colon \mathbb{R}^m \to \mathbb{R}^N$. As already mentioned, the theory only works if we assume the signal x to be sparse or at least compressible.

DEFINITION 2.1 (k-sparse vector) Let $k \in \mathbb{N}^+$, $k \leqslant N$. We call the vector $x \in \mathbb{R}^N$ *k-sparse* if

$$x \in \Sigma_k := \left\{ z \in \mathbb{R}^N | \#\mathrm{supp}(z) \leqslant k \right\}$$

where $\mathrm{supp}(z) := \{i \in \{1,\dots,N\} | z_i \neq 0\}$ denotes the *support* of z.

In applications, signals are often not exactly sparse but at least compressible. We refer to [16] for more details. We define compressibility in terms of the *best k-term approximation error* with respect to the ℓ_p-norm, given by

$$\|x\|_{\ell_p} = \left(\sum_{i=1}^{N} |x_i|^p \right)^{1/p}, \quad 1 \leqslant p < \infty.$$

DEFINITION 2.2 (Best k-term approximation) Let x be an arbitrary vector in $\mathbb{R}^N$. We denote the *best k-term approximation* of x by

$$x_{[k]} := \underset{z \in \Sigma_k}{\arg\min} \|x - z\|_{\ell_p}, \qquad 1 \leqslant p < \infty,$$

and the respective *best k-term approximation error* of x by

$$\sigma_k(x)_{\ell_p} := \min_{z \in \Sigma_k} \|x - z\|_{\ell_p} = \left\| x - x_{[k]} \right\|_{\ell_p}.$$

REMARK 2.1 The best k-term approximation error is the minimal distance of x to a k-sparse vector. Vectors having a relatively small best k-term approximation error are considered to be *compressible*.

REMARK 2.2 If we define the *nonincreasing rearrangement* of x by

$$r(x) = (|x_{i_1}|, \dots, |x_{i_N}|)^T, \text{ and } |x_{i_j}| \geqslant |x_{i_{j+1}}|,\ j = 1,\dots,N-1,$$

then

$$\sigma_k(x)_{\ell_p} = \left(\sum_{j=k+1}^{N} r_j(x)^p \right)^{\frac{1}{p}}, \quad 1 \leqslant p < \infty.$$

Alternatively, we can describe the best-k-term approximation error by

$$\sigma_k(x)_{\ell_p} = \left(\sum_{j \in \Lambda^c} |x_j|^p \right)^{\frac{1}{p}},$$

where $\Lambda := \mathrm{supp}(x_{[k]})$, and Λ^c is its complement in $\{1,\dots,N\}$.

A desirable property of an encoder/decoder pair (A,Δ) is given by the following stability estimate, called *instance optimality*

$$\|x - \Delta(Ax)\|_{\ell_p} \leqslant C\sigma_k(x)_{\ell_p}, \tag{2.1}$$

for all $x \in \mathbb{R}^N$, with a positive constant C independent of x, and k the closest possible to m [7]. This would in particular imply that we are able to recover a k-sparse signal exactly. It turns out that the existence of such a pair restricts the range of k to be maximally of the order of $\frac{m}{\log \frac{m}{N}+1}$. We refer to [4, 7, 10] for more details.

Actually, the above mentioned condition (2.1) can be realized in practice, at least for $p = 1$, by pairing the ℓ_1-*minimization* as the decoder with the choice of an encoder which has the so-called *Null Space Property* with optimal order k. (For realizations of the instance optimality in other ℓ_p-norms, for instance for $p = 2$, one needs more restrictive requirements, see [21]. For the analysis within this paper, we shall use (2.1) just for $p = 1$.)

DEFINITION 2.3 (Null Space Property) A matrix $A \in \mathbb{R}^{m\times N}$ has the *Null Space Property of order k and for positive constant $\gamma_k > 0$* if

$$\|z|_\Lambda\|_{\ell_1} \leqslant \gamma_k \|z|_{\Lambda^c}\|_{\ell_1},$$

for all $z \in \ker A$ and all $\Lambda \subset \{1,\dots,N\}$ such that $\#\Lambda \leqslant k$. We abbreviate this property with the writing (γ_k,k)-*NSP*.

The Null Space Property states that the kernel of the encoding matrix A contains no vectors where some entries have a significantly larger magnitude with respect to the others. In particular, no compressible vector is contained in the kernel. This is a natural requirement since otherwise no decoder would be able to distinguish a sparse vector from zero.

LEMMA 2.1 Let $A \in \mathbb{R}^{m\times N}$ have the (γ_k,k)-NSP, with $\gamma_k < 1$, and define

$$\mathcal{F}(y) := \{z \in \mathbb{R}^N | Az = y\},$$

the set of feasible vectors for the measurement vector $y \in \mathbb{R}^m$. Then the decoder

$$\Delta_1(y) := \underset{z \in \mathcal{F}(y)}{\arg\min} \|z\|_{\ell_1}, \tag{2.2}$$

which we call ℓ_1-*minimization*, performs

$$\|x - \Delta_1(y)\|_{\ell_1} \leqslant C\sigma_k(x)_{\ell_1}, \tag{2.3}$$

for all $x \in \mathcal{F}(y)$ and the constant $C := \frac{2(1+\gamma_k)}{1-\gamma_k}$.

Although the result we stated here is by now well-known, we report its short proof for the sake of completeness, and for comparison with the enhanced guarantees given in Theorem 3.4 below.

Proof. Let us denote $x^* = \Delta_1(y)$ and $z = x^* - x$. Then $z \in \ker A$ and

$$\|x^*\|_{\ell_1} \leqslant \|x\|_{\ell_1},$$

because x^* is a solution of the ℓ_1-minimization problem (2.2). Let Λ be the set of the k-largest entries of x in absolute value. One has

$$\|x^*|_\Lambda\|_{\ell_1} + \|x^*|_{\Lambda^c}\|_{\ell_1} \leqslant \|x|_\Lambda\|_{\ell_1} + \|x|_{\Lambda^c}\|_{\ell_1}.$$

It follows immediately from the triangle inequality that

$$\|x|_\Lambda\|_{\ell_1} - \|z|_\Lambda\|_{\ell_1} + \|z|_{\Lambda^c}\|_{\ell_1} - \|x|_{\Lambda^c}\|_{\ell_1} \leqslant \|x|_\Lambda\|_{\ell_1} + \|x|_{\Lambda^c}\|_{\ell_1}.$$

Hence,

$$\|z|_{\Lambda^c}\|_{\ell_1} \leqslant \|z|_\Lambda\|_{\ell_1} + 2\|x|_{\Lambda^c}\|_{\ell_1} \leqslant \gamma_k\|z|_{\Lambda^c}\|_{\ell_1} + 2\sigma_k(x)_{\ell_1},$$

or, equivalently,

$$\|z|_{\Lambda^c}\|_{\ell_1} \leqslant \frac{2}{1-\gamma_k}\sigma_k(x)_{\ell_1}. \tag{2.4}$$

Finally,

$$\|x - x^*\|_{\ell_1} = \|z|_\Lambda\|_{\ell_1} + \|z|_{\Lambda^c}\|_{\ell_1} \leqslant (\gamma_k + 1)\|z|_{\Lambda^c}\|_{\ell_1} \leqslant \frac{2(1+\gamma_k)}{1-\gamma_k}\sigma_k(x)_{\ell_1},$$

and the proof is completed. □

Unfortunately, the NSP is hard to verify in practice. Therefore one can introduce another property which is called the *Restricted Isometry Property* and implies the NSP.

Being a spectral concentration property, the Restricted Isometry Property is particularly suited to be verified with high probability by certain random matrices; we mention some instances of such classes of matrices below.

DEFINITION 2.4 (Restricted Isometry Property) A matrix $A \in \mathbb{R}^{m\times N}$ has the *Restricted Isometry Property (RIP) of order K with constant* $0 < \delta_K < 1$ if

$$(1-\delta_K)\,\|z\|_{\ell_2} \leqslant \|Az\|_{\ell_2} \leqslant (1+\delta_K)\,\|z\|_{\ell_2}\,,$$

for all $z \in \Sigma_K$. We refer to this property by (δ_K, K)-*RIP*.

LEMMA 2.2 Let $k, h \in \mathbb{N}^+$ and $K = k+h$. Assume that $A \in \mathbb{R}^{m\times N}$ has (δ_K, K)-RIP. Then A has (γ_k, k)-NSP, where

$$\gamma_k := \sqrt{\frac{k}{h}}\frac{1+\delta_K}{1-\delta_K}.$$

The proof of this result can be found, for instance, in [11], and not being of specific relevance for this paper we do not include it here. Encoders which have the RIP with optimal constants, i.e., with k in the order of $\frac{m}{\log\frac{m}{N}+1}$ exist, but, so far, as mentioned above, can be realized exclusively by randomization. By now, classical examples of such stochastic encoders are i.i.d. Gaussian matrices [10] or discrete Fourier matrices with randomly chosen rows [6]. Further details and generalizations are provided in [13, 20]. In the rest of the paper we will use as prototypical cases mainly such stochastic encoders.

What we recalled up to this point of the theory of compressive sensing tells us that we are able to recover by ℓ_1-minimization compressible vectors within a certain accuracy (2.3). If we re-interpret compressible vectors as sparse vectors which are corrupted by noise, we immediately see that the accuracy of the recovered solution is basically driven by the noise level affecting the vector. Nevertheless, neither inequality (2.3) tells us immediately if the recovered support of the k largest entries of the decoded vector is the same as the one of the original signal nor are we able to identify the large entries exceeding a given threshold. Section 3 addresses these issues in detail and investigates the limitations of ℓ_1-minimization. Furthermore, a new decoder is proposed which is able to outperform ℓ_1-minimization in terms of the ability of simultaneously identifying the exact position of the large entries and reducing the noise level on the signal.

3. Support identification

We have seen in Lemma 2.1 that sparse vectors can be recovered exactly by the ℓ_1-minimization decoder Δ_1 if the matrix has the NSP. Moreover, a sparse signal which is disturbed by noise is recovered within a certain accuracy depending on the best k-term approximation error. In this section, we investigate in detail the noise level we can tolerate without loosing the ability of ℓ_1-minimization to recover the support of the undisturbed sparse signal. Let us stress that many other decoders can be used, for instance orthogonal matching pursuit [18, 19] or CoSaMP [17], but the evidence is that such greedy methods tend to be less robust in presence of *noise on the signal*. Hence, we focus on ℓ_1-minimization.

3.1 *The ℓ_1-minimization result*

For later use, let us denote, for $1 \leqslant p \leqslant 2$ and q such that $\frac{1}{p} + \frac{1}{q} = 1$,

$$\kappa_p := \kappa_p(N,k) := \begin{cases} 1, & p = 1, \\ \sqrt[q]{N-k}, & 1 < p \leqslant 2. \end{cases} \tag{3.1}$$

The following proposition shows how we can exactly identify the support of the original signal if we know the ℓ_1-minimizer. It turns out that the large entries of the original signal in absolute value need to exceed a certain threshold, which depends on the noise level.

THEOREM 3.1 Let $x \in \mathbb{R}^N$ be a noisy signal with k relevant entries and the noise level $\eta \in \mathbb{R}$, $\eta \geqslant 0$, i.e., for $\Lambda = \operatorname{supp}(x_{[k]})$,

$$\sum_{j \in \Lambda^c} |x_j|^p \leqslant \eta^p, \tag{3.2}$$

for a *fixed* $1 \leqslant p \leqslant 2$. Consider further an encoder $A \in \mathbb{R}^{m \times N}$ which has the (k, γ_k)-NSP, with $\gamma_k < 1$, the respective measurement vector $y = Ax \in \mathbb{R}^m$, and the ℓ_1-minimizer

$$x^* := \operatorname*{arg\,min}_{z \in \mathcal{F}(y)} \|z\|_{\ell_1}.$$

If the i-th component of the original signal x is such that

$$|x_i| > \frac{2(1+\gamma_k)}{1-\gamma_k} \kappa_p \, \eta,$$

then $i \in \operatorname{supp}(x^*)$.

Proof. We know by (2.3) that

$$\|x^* - x\|_{\ell_1} \leqslant \frac{2(1+\gamma_k)}{1-\gamma_k} \sigma_k(x)_{\ell_1}. \tag{3.3}$$

Thus, by Hölder's inequality and assumption (3.2), we obtain the estimate

$$\|x^* - x\|_{\ell_1} \leqslant \frac{2(1+\gamma_k)}{1-\gamma_k} \sigma_k(x)_{\ell_1} \leqslant \frac{2(1+\gamma_k)}{1-\gamma_k} \kappa_p \, \eta. \tag{3.4}$$

We now choose a component $i \in \{1,\dots,N\}$ such that

$$|x_i| > \frac{2(1+\gamma_k)}{1-\gamma_k}\kappa_p\,\eta,$$

and assume $i \notin \operatorname{supp}(x^*)$. This leads to the contradiction:

$$|x_i| = |x_i - x_i^*| \leqslant \|x - x^*\|_{\ell_1} \leqslant \frac{2(1+\gamma_k)}{1-\gamma_k}\kappa_p\,\eta < |x_i|. \tag{3.5}$$

Hence, necessarily $i \in \operatorname{supp}(x^*)$. □

The noise level substantially influences the ability of support identification. Here, the noisy signal must have the k largest entries in absolute value above

$$r_p := \frac{2(1+\gamma_k)}{1-\gamma_k}\kappa_p\,\eta.$$

In the following, we shall introduce a certain class of vectors for which a smaller threshold is required for support identification.

3.2 *A class of sparse vectors affected by bounded noise*

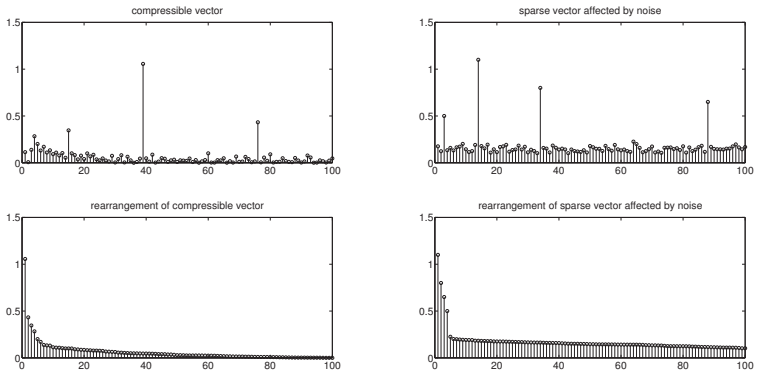
Inspired by the observations of the previous section, we define the class of *sparse vectors affected by bounded noise*,

$$\mathcal{S}^p_{\eta,k,r} := \left\{ x \in \mathbb{R}^N \,\middle|\, \#S_r(x) \leqslant k \text{ and } \sum_{i\in(S_r(x))^c} |x_i|^p \leqslant \eta^p \right\}, \quad 1 \leqslant p \leqslant 2, \tag{3.6}$$

where $S_r(x) := \{i \in \{1,\dots,N\} \,|\, |x_i| > r\}$, $0 \leqslant \eta < r$, and $1 \leqslant k \leqslant m$. It contains all vectors for which at most k large entries exceed the threshold r in absolute value, while the p-norm of the other entries stays below a certain noise level. In Figure 1 we visually exemplify, how vectors in the introduced class are distinguished from compressible vectors. Compressible vectors exhibit a fast decay of the nonincreasing rearrangement. Sparse vectors which are affected by bounded noise have a gap between the relevant entries and the small ones, which represent the noise.

We take two elements of this class and assume that they have the same measurements, i.e., they are both in $\mathcal{F}(y)$ for $y \in \mathbb{R}^m$. In the following result, we provide a precise stability estimate of the discrepancy between the supports of the respective large entries.

THEOREM 3.2 Let $A \in \mathbb{R}^{m\times N}$ which has the $(\gamma_{2k}, 2k)$-NSP, for $\gamma_{2k} < 1$, $1 \leqslant p \leqslant 2$, and



(a) (b)

FIG. 1. Comparison of compressible vector and sparse vector affected by noise.

$x, x' \in \mathcal{S}^p_{\eta,k,r}$ such that $Ax = Ax'$. Then

$$\#(S_r(x) \Delta S_r(x')) \leqslant \frac{(2\gamma_{2k}\kappa_p\eta)^p}{(r-\eta)^p}. \tag{3.7}$$

(Here we denote by "Δ" the set symmetric difference, not to be confused with the symbol of a generic decoder.) If additionally

$$r > \eta(1+2\gamma_{2k}\kappa_p) =: r_S, \tag{3.8}$$

then $S_r(x) = S_r(x')$, i.e., we have unique identification of the large entries in absolute value.

Proof. As $Ax = Ax'$, the difference $(x-x') \in \ker(A)$. By the $(\gamma_{2k}, 2k)$-NSP, Hölder's inequality, and the triangle inequality we have

$$\begin{aligned}
\left\|(x-x')|_{S_r(x)\cup S_r(x')}\right\|_{\ell_p} &\leqslant \left\|(x-x')|_{S_r(x)\cup S_r(x')}\right\|_{\ell_1} \\
&\leqslant \gamma_{2k}\left\|(x-x')|_{(S_r(x)\cup S_r(x'))^c}\right\|_{\ell_1} \\
&\leqslant \gamma_{2k}\kappa_p\left\|(x-x')|_{(S_r(x)\cup S_r(x'))^c}\right\|_{\ell_p} \\
&\leqslant 2\gamma_{2k}\kappa_p\eta.
\end{aligned} \tag{3.9}$$

Further, we investigate the symmetric difference of the supports of the large entries of x

and x' in absolute value. If $i \in S_r(x)\Delta S_r(x')$, then either $|x_i| > r$ and $|x'_i| \leqslant \eta$ or $|x_i| \leqslant \eta$ and $|x'_i| > r$. This implies that $|x'_i - x_i| > (r-\eta)$. Thus we have

$$\left\|(x-x')|_{S_r(x)\Delta S_r(x')}\right\|_{\ell_p}^p \geqslant \left(\#(S_r(x)\Delta S_r(x'))\right)(r-\eta)^p.$$

Together with equation (3.9) and the non-negativity of $\left\|(x-x')|_{S_r(x)\cap S_r(x')}\right\|_{\ell_p}$, we obtain the chain of inequalities

$$\begin{aligned}(2\gamma_{2k}\kappa_p\eta)^p &\geqslant \left\|(x-x')|_{S_r(x)\cup S_r(x')}\right\|_{\ell_p}^p \\ &\geqslant \left\|(x-x')|_{S_r(x)\cap S_r(x')}\right\|_{\ell_p}^p + \left\|(x-x')|_{S_r(x)\Delta S_r(x')}\right\|_{\ell_p}^p \\ &\geqslant \left(\#(S_r(x)\Delta S_r(x'))\right)(r-\eta)^p,\end{aligned}$$

and therefore

$$\#(S_r(x)\Delta S_r(x')) \leqslant \frac{(2\gamma_{2k}\kappa_p\eta)^p}{(r-\eta)^p}. \tag{3.10}$$

For unique support identification, we want the symmetric difference between the sets $S_r(x)$ and $S_r(x')$ to be empty. Thus the left-hand side of inequality (3.10) has to be zero. Since $\#(S_r(x)\Delta S_r(x')) \in \mathbb{N}$, it is sufficient to require that the right-hand side be strictly less than one, and this is equivalent to condition (3.8). □

REMARK 3.1 The gap between the two thresholds r_1, r_S is given by

$$r_1 - r_S = \left(2\left(\frac{1+\gamma_k}{1-\gamma_k} - \gamma_{2k}\right)\kappa_p(N,k) - 1\right)\eta.$$

As $\gamma_{2k} < 1 < \frac{1+\gamma_k}{1-\gamma_k}$ and $\kappa_p(N,k)$ is very large for $N \gg k$, this *positive* gap is actually very large, for $N \gg 1$. Unfortunately, this discrepancy cannot be amended because the ℓ_1-minimization decoder Δ_1 has *not* in general the property

$$x \in \mathcal{S}^p_{\eta,k,r} \Rightarrow x^* = \Delta_1(Ax) \in \mathcal{S}^p_{\eta,k,r},$$

hence, it allows us neither to say that also the ℓ_1-minimizer has a bounded noise component

$$\sum_{i\in(S_r(x^*))^c} |x_i^*|^p \leqslant \eta^p,$$

nor to apply Theorem 3.2 to obtain support stability. We present several examples in Section 5, showing these ineliminable limitations of Δ_1.

3.3 *The selective p-potential functional and its properties*

To overcome the shortcoming of ℓ_1-minimization in 1. damping the noise-folding and consequently in 2. having a stable support recovery, in this section we design a new decoding procedure which allows us to have both these very desirable properties.

Let us first introduce the following functional.

DEFINITION 3.3 (Selective p-potential) We define the *truncated p-power function* $W_r^p\colon \mathbb{R} \to \mathbb{R}_0^+$,

$$W_r^p(t) = \min\{t^p, r^p\}, \qquad r > 0, \qquad 1 \leqslant p \leqslant 2, \tag{3.11}$$

We call the functional $\mathcal{SP}_r^p\colon \mathbb{R}^N \to \mathbb{R}_0^+$,

$$\mathcal{SP}_r^p(x) = \sum_{j=1}^{N} W_r^p(x_j), \qquad r > 0, \qquad 1 \leqslant p \leqslant 2, \tag{3.12}$$

the *selective p-potential (SP) functional*.

On the one hand, the SP functional is neither convex nor differentiable. These two properties may cause the problem of designing an appropriate optimization algorithm for computing a linearly constrained global minimizer. Nevertheless, we address and overcome this difficulty below in Section 4, where we illustrate an efficient and robust algorithm. On the other hand, this potential drawback is additionally compensated by the fact that its minimization has the features of noise-folding damping and support stability which we are seeking.

THEOREM 3.4 Let $A \in \mathbb{R}^{m\times N}$ have the $(\gamma_{2k}, 2k) - NSP$, with $\gamma_{2k} < 1$, and $1 \leqslant p \leqslant 2$. Furthermore, we assume $x \in \mathcal{S}_{\eta,k,r}^p$, $0 < \eta < r$, with the property of having the minimal $\#S_r(x)$ within $\mathcal{F}(y)$, where $y = Ax$ is its associated measurement vector, i.e.,

$$\#S_r(x) \leqslant \#S_r(z) \text{ for all } z \in \mathcal{F}(y). \tag{3.13}$$

If x^* is obtained by the decoder, which we call the *selective least p-powers (SLP)*,

$$x^* = \Delta_{\mathcal{SP}}(y) := \operatorname*{arg\,min}_{z\in\mathcal{F}(y)} \mathcal{SP}_r^p(z),$$

then also $x^* \in \mathcal{S}_{\eta,k,r}^p$, implying noise-folding damping. Moreover, we have the support stability property

$$\#(S_r(x)\Delta S_r(x^*)) \leqslant \frac{(2\gamma_{2k}\kappa_p\eta)^p}{(r-\eta)^p}. \tag{3.14}$$

Proof. Notice that we can equally describe the SP functional by

$$\mathcal{SP}_r^p(z) = r^p \#S_r(z) + \sum_{i \in (S_r(z))^c} |z_i|^p.$$

By minimality of x^*, i.e.,

$$\mathcal{SP}_r^p(x^*) \leqslant \mathcal{SP}_r^p(x), \tag{3.15}$$

and the assumption $x \in \mathcal{S}^p_{\eta,k,r}$, we have

$$\begin{aligned} r^p \#S_r(x^*) \leqslant \mathcal{SP}_r^p(x^*) \leqslant \mathcal{SP}_r^p(x) &= r^p \#S_r(x) + \sum_{i \in (S_r(x))^c} |x_i|^p \\ &\leqslant r^p \#S_r(x) + \eta^p, \end{aligned} \tag{3.16}$$

and thus,

$$\#S_r(x^*) \leqslant \left(\frac{\eta}{r}\right)^p + \#S_r(x).$$

As $\frac{\eta}{r} < 1$ by hypothesis, the minimality property (3.13) yields immediately

$$\#S_r(x^*) = \#S_r(x) \leqslant k. \tag{3.17}$$

This latter equivalence and

$$\begin{aligned} r^p \#S_r(x^*) + \sum_{i \in (S_r(x^*))^c} |x_i^*|^p &\leqslant r^p \#S_r(x) + \sum_{i \in (S_r(x))^c} |x_i|^p, \text{ or} \\ \sum_{i \in (S_r(x^*))^c} |x_i^*|^p &\leqslant \sum_{i \in (S_r(x))^c} |x_i|^p \leqslant \eta^p, \end{aligned}$$

imply $x^* \in \mathcal{S}^p_{\eta,k,r}$. By an application of Theorem 3.2, we obtain (3.14). □

REMARK 3.2 Notice that it is actually *not* necessary that x^* is the *global* minimizer of $\mathcal{SP}_r^p$ over $\mathcal{F}(y)$ to guarantee $x^* \in \mathcal{S}^p_{\eta,k,r}$. In fact, a careful reading of the previous proof makes clear that actually condition (3.15) and $x^* \in \mathcal{F}(y)$ are sufficient to imply $x^* \in \mathcal{S}^p_{\eta,k,r}$.

4. Minimization of the selective p-potential functional

In the last section, we introduced the SP functional, which is nonconvex and nonsmooth. Unluckily, this makes its linearly constrained minimization also nontrivial. Here, we recall a novel and very robust algorithm for linearly constrained nonconvex and nonsmooth minimization, introduced and analyzed first in [2]. The algorithm is particularly suited for our purpose, by introducing a very small C^1-regularization of the functional. This distinguishes

it from other well-known methods such as SQP and Newton methods, which require a more restrictive C^2-regularity. All notions and results written in this section are collected more in general in [2]. Nevertheless we consider it as straightforward to report them adapted to our specific case in order to have a simplified and more immediate formulation.

4.1 *A mild regularization*

First, we show how the functional $\mathcal{SP} := \mathcal{SP}_r^p$ can be approximated by the smooth functional $\mathcal{SP}^\varepsilon$, which possesses the regularity that is necessary to apply the method cited above. The smoothing is introduced on the truncated p-power function W_r^p, defined in (3.11), in order to obtain a C^1 function. We shall then substitute in (3.12) the term $W_r^p(t)$ by the function

$$W_r^{p,\varepsilon}(t) = \begin{cases} t^p & 0 \leqslant t < r-\varepsilon, \\ \pi_p(t) & r-\varepsilon \leqslant t \leqslant r+\varepsilon, \qquad t \geqslant 0, \\ r^p & t > r+\varepsilon, \end{cases} \tag{4.1}$$

where $0 < \varepsilon < r$, and $\pi_p(t)$ is the third degree interpolating polynomial

$$\pi_p(t) := A(t-s_2)^3 + B(t-s_2)^2 + C,$$

with

$$\begin{cases} C = \gamma_3, \\ B = \frac{\gamma_1}{s_2-s_1} - \frac{3(\gamma_3-\gamma_2)}{(s_2-s_1)^2}, \\ A = \frac{\gamma_1}{3(s_2-s_1)^2} + \frac{2B}{3(s_2-s_1)}. \end{cases} \tag{4.2}$$

The parameters which appeared in the definition of the interpolating function are defined as: $s_1 = (r-\varepsilon)$, $s_2 = (r+\varepsilon)$, $\gamma_1 = p(r-\varepsilon)^{p-1}$, $\gamma_2 = (r-\varepsilon)^p$, and $\gamma_3 = r^p$. For example, for the relevant case $p = 2$, this function has the analytic form:

$$\pi_2(t) := \frac{[t+(r-\varepsilon)][\varepsilon(r+t)-(r-t)^2]}{4\varepsilon}.$$

The newly smoothed selective p-power function has been defined so far only for $t \geqslant 0$. We define it for $t < 0$ as $W_r^{p,\varepsilon}(t) = W_r^{p,\varepsilon}(-t)$. Thanks to this definition, we can explicitly write the new decoder:

$$\Delta_{\mathcal{SP}}^\varepsilon(y) = \underset{Az=y}{\arg\min} \left(\mathcal{SP}^{p,\varepsilon}(z) := \sum_{i=1}^m W_r^{p,\varepsilon}(z_i) \right). \tag{4.3}$$

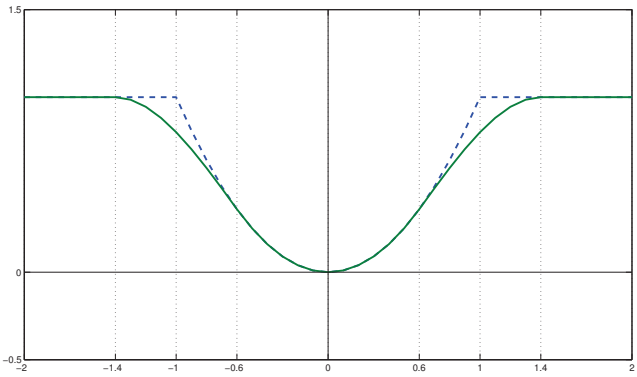


FIG. 2. Truncated quadratic potential W_r^p and its regularization $W_r^{p,\varepsilon}$, for $p=2$, $r=1$, and $\varepsilon=0.4$.

As shown in [2, Corollary 3.7]

$$\Delta_{\mathcal{SP}}(y) = \lim_{\varepsilon\to 0} \Delta^{\varepsilon}_{\mathcal{SP}}(y),$$

hence we will be content with finding an appropriate algorithm to perform (4.3).

REMARK 4.1 For the sake of the analysis below, let us note that the functional $\mathcal{SP}^{p,\varepsilon}$ is semi-convex, that means that there exists a constant $\omega > 0$ such that $\mathcal{SP}^{p,\varepsilon}(\cdot) + \omega \|\cdot\|^2$ is convex.

4.2 *The algorithm*

In this section we present the algorithm to perfom (4.3). Before describing it, it is necessary to introduce the concept of ν-convexity, which plays a key-role in the minimization process. In fact, to achieve the minimization of the functional $\mathcal{SP}^{p,\varepsilon}$, we use a Bregman-like inner loop, which requires this property to converge.

DEFINITION 4.1 (ν-convexity) A convex function $f : \mathbb{R}^N \to \mathbb{R}$ is ν-convex if there exists a constant $\nu > 0$ such that for all $x, x' \in \mathbb{R}^N$ and $\phi \in \partial f(x), \psi \in \partial f(x')$

$$\langle \phi - \psi, x - x' \rangle \geqslant \nu \left\| x - x' \right\|^2_{\ell_2}, \tag{4.4}$$

where ∂f is the subdifferential of the function f.

The starting values $x_0 = x_{(0,0)} \in \mathbb{R}^N$ and $q_{(0,0)} \in \mathbb{R}^m$ are taken arbitrarily. For a fixed scaling parameter $\lambda > 0$ and an adaptively chosen sequence of integers $(L_\ell)_{\ell\in\mathbb{N}}$, we set

Algorithm 4.2

while $\|x_{\ell-1}-x_\ell\|_{\ell_2} \leqslant TOL$ **do**
  $x_{(\ell,0)} = x_{\ell-1} := x_{(\ell-1,L_{\ell-1})}$
  $q_{(\ell,0)} = q_{\ell-1} := q_{(\ell-1,L_{\ell-1})}$
  for $k = 1,\dots,L_\ell$ **do**
    $x_{(\ell,k)} = \arg\min_{x\in\mathbb{R}^N} \left(\mathcal{SP}^{p,\varepsilon}_{\omega,x_{\ell-1}}(x) - \langle q_{(\ell,k-1)}, Ax\rangle + \lambda \|Ax-y\|^2_{\ell_2}\right)$
    $q_{(\ell,k)} = q_{(\ell,k-1)} + 2\lambda(y - Ax_{(\ell,k)})$
  end for
end while

The reader can notice that the functional $\mathcal{SP}^{p,\varepsilon}_{\omega,x_{\ell-1}}$, which appears in the algorithm, has not been yet introduced. Indeed, a further modification to the original functional $\mathcal{SP}$ must be introduced in order to have $\nu-$convexity, which is necessary for the convergence of the algorithm. It is defined as

$$\mathcal{SP}^{p,\varepsilon}_{\omega,x'}(x) := \mathcal{SP}^{p,\varepsilon}(x) + \omega \left\|x-x'\right\|^2_{\ell_2},$$

where ω is chosen such that the new functional has the $\nu-$convexity, which we require. The finite adaptively chosen number of inner loop iterates L_ℓ is defined by the condition

$$(1+\|q_{\ell-1}\|_{\ell_2}) \left\|Ax_{(\ell,L_\ell)} - y\right\|_{\ell_2} \leqslant \frac{1}{\ell^\alpha},$$

for a given parameter α, which in our numerical experiments will be set $\alpha = 1.1$. We refer to [2, Section 2.2] for details on the finiteness of L_ℓ and for the proof of convergence of Algorithm 4.2 to critical points of $\mathcal{SP}^{p,\varepsilon}$ in $\mathcal{F}(y)$. According to Remark 3.2, and as we will empirically show in Section 5, this algorithm can find critical points (likely close to a global minimizer) with the desired properties illustrated in Theorem 3.4, as soon as we select the starting point x_0 by an appropriate warm-up procedure.

Algorithm 4.2 does not yet specify how to minimize the convex functional

$$\left(\mathcal{SP}^{p,\varepsilon}_{\omega,x_{\ell-1}}(x) - \langle q_{(\ell,k-1)}, Ax\rangle + \lambda \|Ax-y\|^2_{\ell_2}\right),$$

in the inner loop. For that we can use an iterative thresholding algorithm introduced in [2, Section 3.7], inspired by the previous work [12] for the corresponding *unconstrained* optimization of selective p-potentials. This method ensures the convergence to a minimizer and is extremely agile to be implemented, as it is based on matrix-vector multiplications and very simple componentwise nonlinear thresholdings.

By the iterative thresholding algorithm we actually equivalently minimize the functional

$$\mathcal{SP}^{p,\varepsilon}_{\omega,x'}(x,q) = \mathcal{SP}^{p,\varepsilon}_{\omega,x'}(x) + \lambda \|Ax-(y+q)\|^2_{\ell_2},$$

where we set $\lambda = \frac{1}{2}$ only for simplicity. The thresholding functions $\mathbb{S}_p^\mu$ we use are defined in [2, Lemma 3.15] and, for the relevant case $p = 2$, it has the analytic form

$$\mathbb{S}_2^\mu(\xi) := \begin{cases} \dfrac{\xi}{1+\mu} & |\xi| < (\bar{r}-\varepsilon)(1+\mu), \\ \dfrac{4\varepsilon}{3\mu}\left(1+\dfrac{\mu}{4\varepsilon}(2\varepsilon+\bar{r}) - \sqrt{\dfrac{\Gamma(\xi)}{4}}\right) & (\bar{r}-\varepsilon)(1+\mu) \leqslant |\xi| \leqslant r+\varepsilon, \\ \xi & |\xi| > \bar{r}+\varepsilon, \end{cases} \tag{4.5}$$

where

$$\Gamma(\xi) := 4\left(1+\left(\frac{\mu}{4\varepsilon}\right)^2(2\bar{r}+\varepsilon)^2 + \frac{\mu}{2\varepsilon}(2\varepsilon+\bar{r}) - \frac{3\mu}{2\varepsilon}\xi\right).$$

Note that in case $p \neq 2$ the only part that varies is the one for $|\xi| < (\bar{r}-\varepsilon)(1+\mu)$ because the remaining ones do not depend on p. We show in Figure 3 the typical shapes of these thresholding functions for different choices of $p \in \{1, 3/2, 2\}$. With the help of these thresholding functions, the algorithm is given by the componentwise fixed point iteration, for $n \geqslant 0$,

$$x_i^{n+1} = \mathbb{S}_p^\mu\left(\left\{\frac{1}{2}\left[(I-\frac{1}{2}A^*A)+(1-\omega)I\right]x^n + \frac{1}{2}A^*(y+q) + \omega x'\right\}_i\right), \quad i = 1,\dots,N. \tag{4.6}$$

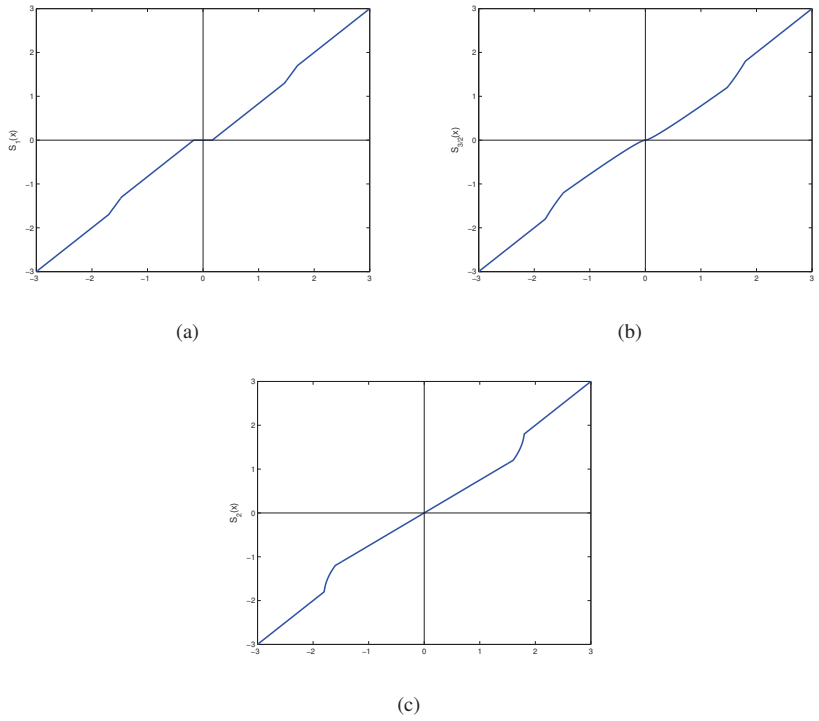
We refer to [2, Theorem 3.17] for the convergence properties of this algorithm.

To summarize, Algorithm 4.2 can be realized in practice by nesting three loops. One external loop makes vanishing the quadratic convexification, the second external loop updates the Lagrange multipliers $q_{(\ell,k)}$ for a fixed quadratic convexification, and the final inner loop implements (4.6).

5. Numerical results

The following numerical simulations provide empirical confirmation that, as predicted by Theorem 3.4, the decoder $\Delta_{\mathcal{SP}}$ recovers a signal in the class $\mathcal{S}^p_{\eta,k,r}$, defined in (3.6). For the sake of simplicity, we perform our numerical experiments only for the case $p = 2$.

We shall show that the selective least p-powers (SLP-minimization), equipped with a proper starting point, has much better support identification properties than ℓ_1-minimization because the signal is in the *same* compressibility class of the original signal, and so we know a priori that all the entries of x^* larger than r are part of the support of the large entries



(a) (b)

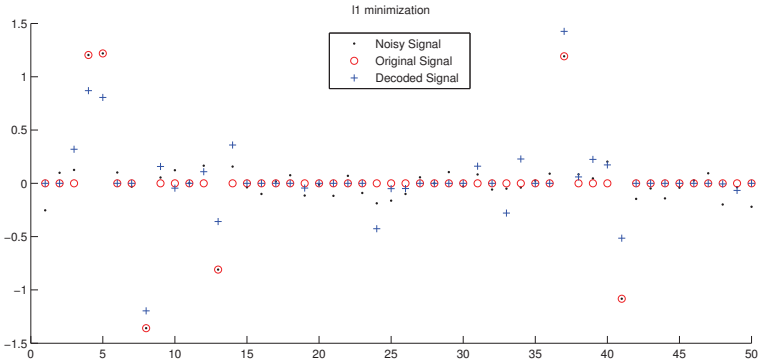
(c)

FIG. 3. The Lipschitz continuous thresholding functions $\mathbb{S}_1^\mu$, $\mathbb{S}_{3/2}^\mu$, and $\mathbb{S}_2^\mu$, with parameters $p = 1, 3/2$, and 2, respectively, and $r = 1.5$, $\mu = 5$, $\varepsilon = 0.3$.

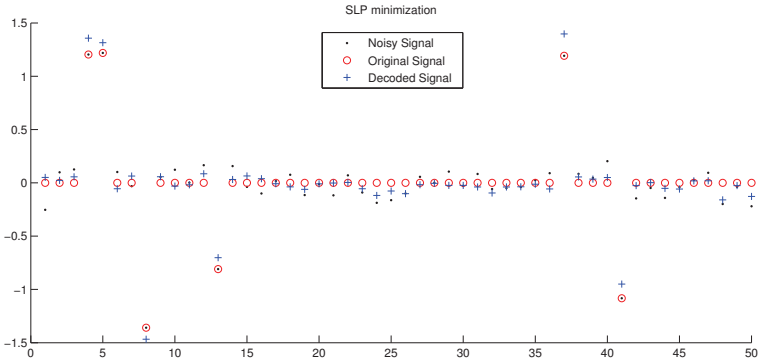
of the original signal and that $\sum_{i\in(S_r(x^*))^c} |x_i^*|^2 \leqslant \eta^2$. This fact is of particular importance when we do not know how many large entries the original vector has.

We start the discussion of the results with a comparison between the ℓ_1-minimization and the SLP-minimization reported in Figure 4. Although the setting of the two minimizations is the same, the results are very different: SLP-minimization can recover the signal with a very good approximation of its large entries in absolute value and a significant reduction of the noise level, while ℓ_1-minimization not only approximates the signal in a bad way, but also looses some information for some entries, due to an amplification of the noise. It is evident that in Subfigure 4(a) $|x_{\ell_1}(13)| < |x_{\ell_1}(24)|$ gives a wrong information about the location of the relevant entries.

In Section 4 we mentioned that the algorithm, which minimizes the functional $\mathcal{SP}_r^p$, finds only a critical point, so the condition $\mathcal{SP}_r^2(x^*) \leqslant \mathcal{SP}_r^2(x)$ (3.15) used in the proof of Theorem 3.4 may not be always valid. In order to enhance the chance of validity of this condition, the choice of an appropriate starting point is crucial. As we know that the



(a)

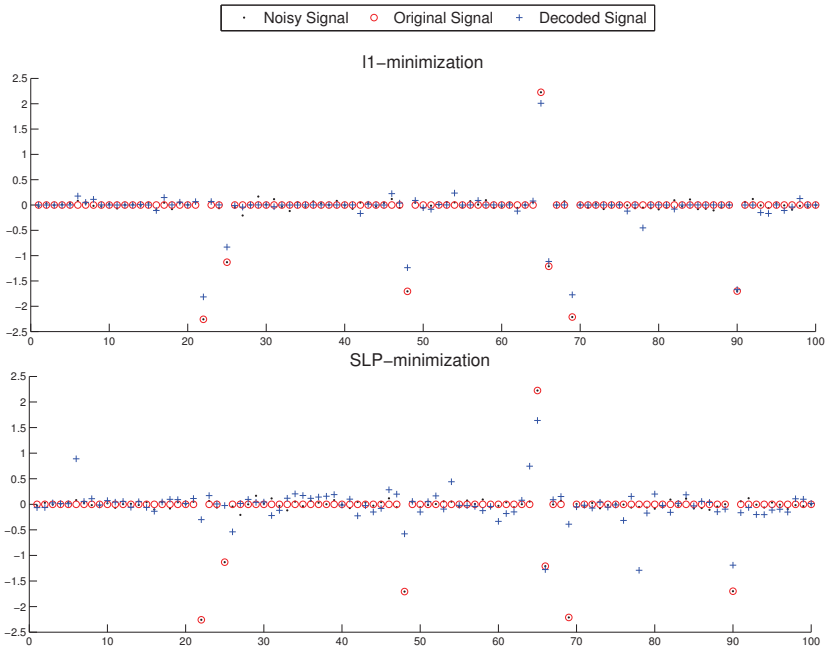


(b)

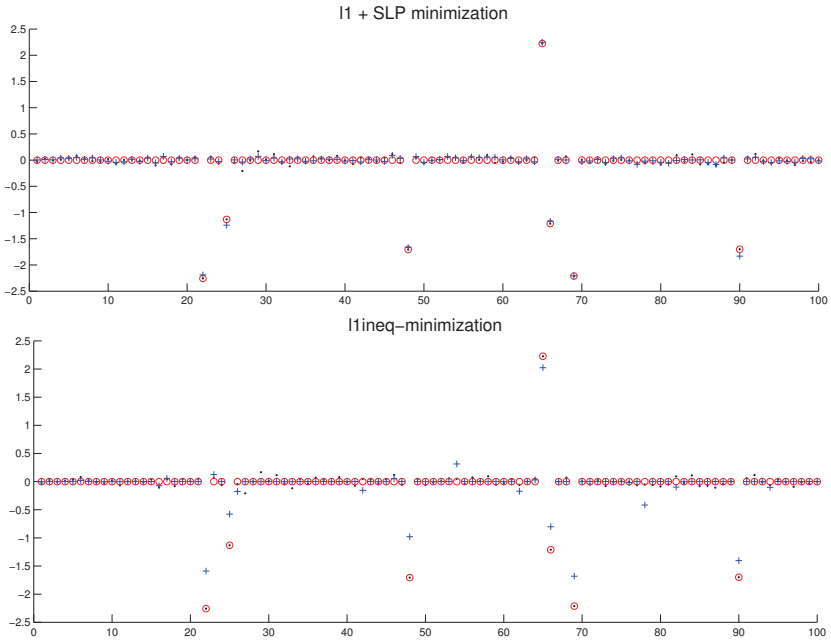
FIG. 4. The results of the minimization processes ℓ_1 and SLP are shown in Subfigure 4(a) and 4(b) respectively. The two decoders are intended to recover the original signal (o), starting from the measurement of the noisy signal ($\cdot$). The output of the two processes is represented by ($+$). In this case the starting value for both minimizations is $x_0 = 0$.

ℓ_1-minimization converges to its global minimizer with at least some guarantees given by Theorem 3.1, we use the result of this minimization process as a warm up to select the starting point of Algorithm 4.2.

In Figure 5 we illustrate the robust effect of this relevant combination ($\ell_1 \Rightarrow$ SLP). The bottom subfigure of Figure 5(a) shows the result obtained with the SLP-minimization starting from the value $x_0 = 0$; the algorithm converged to a feasible critical point, but it is quite evident that the decoding process failed. If we look at the ℓ_1-minimization result, as shown on the upper subfigure of Figure 5(a), the minimization process brings us close to the solution, but the result is still quite bad: for instance, the entry around position 80



(a)



(b)

FIG. 5. The figure reports the results of four different decoding processes (+) of the same problem where the circles (o) represent the original signal and the points (·) represent the original signal corrupted by the noise.

is not much smaller than the one around position 25, but the first is part of the noise and the second is the signal. If we then apply on the result of the ℓ_1-minimization the SLP-minimization, as shown on the upper subfigure of Figure 5(b), we obtain as a result a good approximation of the original signal and we get a significant correction and an improved recovery; all the relevant entries are approximated better and the noise is confined within a very small stripe around zero, whose amplitude is even less than the one of the original noise! The last subfigure represents another result obtained by the ℓ_1-minimization to which we have substituted the equality constraint $Ax = y$ with another inequality constraint that take into account the noise level, folded from the noise on the signal though; thus

$$\|Ax - y\|_{\ell_2} \leqslant \|An\|_{\ell_2}.$$

We can see that in this case the small entries are closer to zero with respect to the other ℓ_1-minimization, but the price to pay for this reduction is high: the relevant entries are approximated much worse and it is easier to confuse them with some noise. For example, notice that the entries 80 and 25, which were already mentioned before, are now even closer and it is very hard to decide which of those represents part of the noise or the signal.

These specific examples in support of our new decoding strategy are actually typical. In order to support this work with even more impressive and convincing evidences, we present some statistical data obtained by solving series of problems. We decided to fix most of the parameters in order to have the most coherent data to be analyzed; in particular, we set $N = 100$, $m = 40$, $r = 1$, $k = 1, \ldots, 7$, and $\eta = 0.7$. The vector n is composed of random entries with normal distribution and then it is rescaled in order to have $\|n\|_{\ell_2} = \eta$. Figures 6, 7, and 8 report the results obtained considering 30 different i.i.d.Gaussian encoding matrices while Figures 9, 10, and 11 used 30 random subsampled Fourier encoding matrices. For a more immediate notation, we will call the results of the different minimization processes with the following intuitive notation: $x^*_{\ell_1}$, $x^*_{\ell_1+\mathcal{SP}}$, $x^*_{\mathcal{SP}^p_r}$, and $x^*_{\ell_1 ineq}$. In case where we will use x^*, we will not refer to a specific minimization technique.

Figure 6 reports the first part of the statistics which we have collected for the Gaussian matrices. We start commenting the subfigures clockwise. The first subfigure, on the upper-left, represents the mean value of the error between the exact signal and the decoded one $\|x - x^*\|_{\ell_2}$. The second column refers to the SLP-minimization result using as the starting value $x^*_{\ell_1}$. In most of the experiments the statistics is independent of the number of large entries. Moreover, it often produces the smallest error we had in the signal recovery. The second subfigure is the mean value of the noise level and we can see exactly what we inferred looking at Figure 5: $x^*_{\ell_1}$ has a larger noise level with respect to $x^*_{\ell_1+\mathcal{SP}}$, and $x^*_{\ell_1 ineq}$ has the best noise reduction property. The third is the mean computational times for the four methods. The algorithm minimizing the SLP functional takes longer with respect to *CVX*, a package for specifying and solving convex programs [8, 14], which we used to perform the ℓ_1-minimization. However, a good starting point for SLP can also give an advantage in terms of computational time. The fourth plot reports the mean value of the separation between noise level and the large entries of the signal. We see that the second

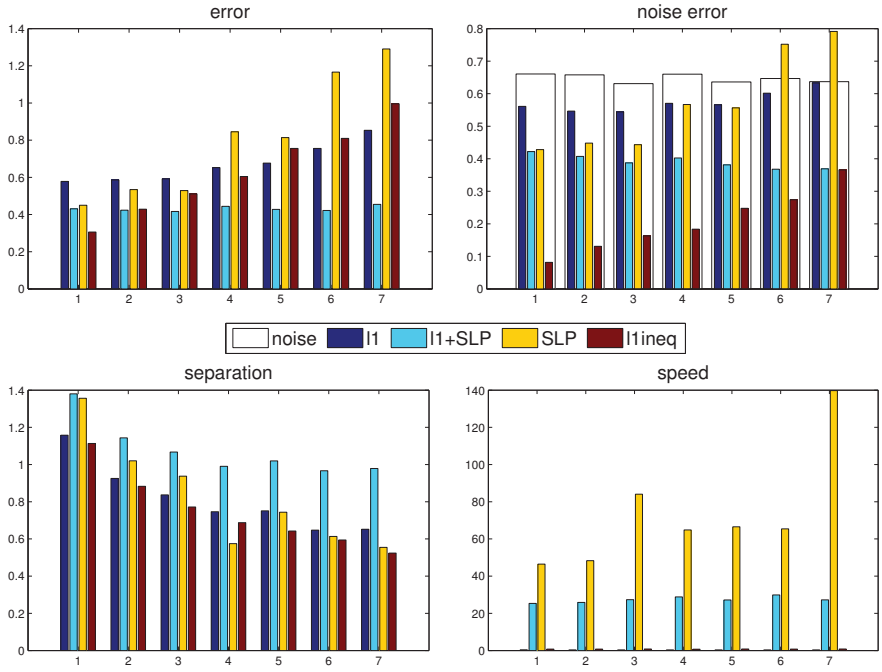
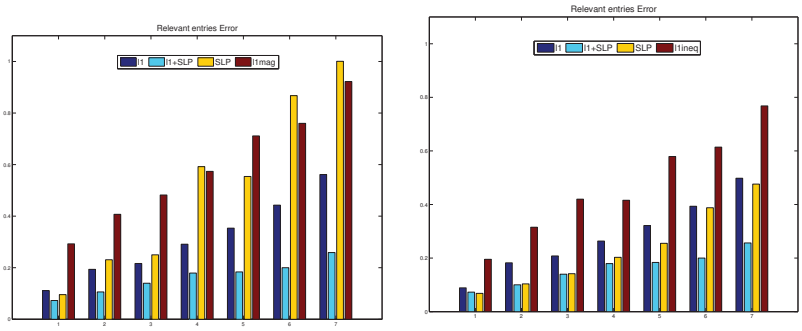


FIG. 6. The blue column refers to the solution x^* obtained with ℓ_1-minimization with $x_0 = 0$, the cyan the one given by the SLP that used the ℓ_1 solution as starting point, the yellow refers to the SLP-minimizer with $x_0 = 0$, and the brown column reports the data related to the ℓ_1-minimization with the inequality constraint. In the *noise error* subfigure the white column in the background represents the noise level. On the x-axis the different values of k are displayed and each column is the mean of the results given by 30 trials.

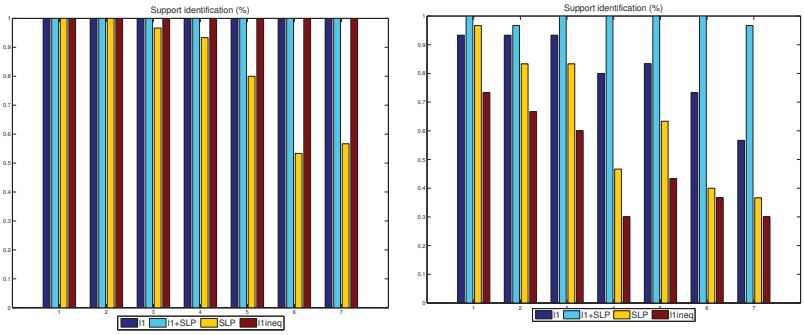
column is always larger than the others: the SLP minimizer can better discriminate the signal from the noise.

In Figure 7 we report the histogram of the mean-value of the errors on the relevant entries: the quantities on the left are $\frac{1}{ntr}\sum_{i=1}^{ntr}\left\|x_{[k],i}-x^*_{[k],i}\right\|_{\ell_2}$ where k is supposed to be known and ntr is the number of the problems which we solved, in our case $ntr = 30$; the right ones are $\frac{1}{ntr}\sum_{i=1}^{ntr}\left\|x_i|_{S_r(x_i^*)}-x_i^*|_{S_r(x_i^*)}\right\|_{\ell_2}$ in which we only consider the entries larger than r in the decoded signal. We see that the SLP combined with the ℓ_1-starting point has always the smallest error while, as predicted, the largest error is computed by $x^*_{\ell_1 ineq}$. Notice that the difference between $x^*_{\ell_1}$ and $x^*_{\ell_1+\mathcal{SP}}$ is relevant for large k.



(a) (b)

FIG. 7. The subfigures represent the error on the relevant entries. For more details on the displayed data we refer to the caption of Figure 6.



(a) (b)

FIG. 8. Support identification statistics. For more details on the displayed data we refer to the caption of Figure 6.

Figure 8 reports the analysis of the capability of the different methods to detect the support. The left subfigure shows the percentage of support recovery in case that k is known. There are practically no differences between the three methods, while SLP, with a bad starting vector, is not so efficient, showing us that this minimization really depends on the choice of the starting point. More important is the subfigure on the right: it represents the percentage of support identification on 30 trials assuming that we do not know the quantity of the relevant entries. We attribute a positive match in case $S_r(x^*) = S_r(x)$ so that the relevant entries of x^* coincide with the ones of the original signal. Thanks to Theorem 3.2 and Theorem 3.4 we expect that the method which combines ℓ_1 with SLP-minimization has better results than $x^*_{\ell_1 ineq}$. Actually this is confirmed by the experiments: if we know nothing related to the original signal, and we only take the entries located in the position j such that $|x^*_j| > r$, then ℓ_1 combined with SLP-minimization does a very accurate recovery,

as it gives us almost always 100% of the correct result while the other percentages are significantly lower.

Figures 9, 10, and 11 represent the same statistical data reported respectively in the Figures 6, 7, and 8 but using random subsampled Fourier encoding matrices. Without describing the result in detail, we state that they are very similar for these problems as well. The $x^*_{\ell_1+\mathcal{SP}}$ definitively approximates better the original signal than the other three methods.

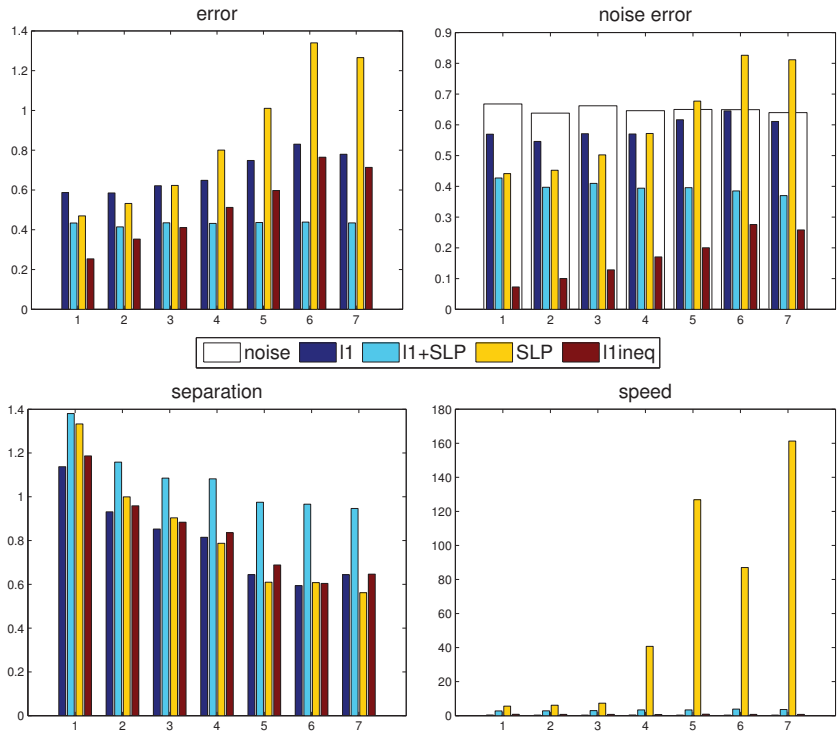
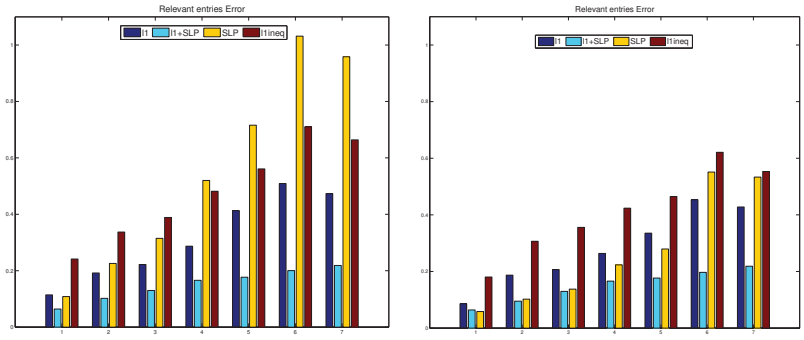


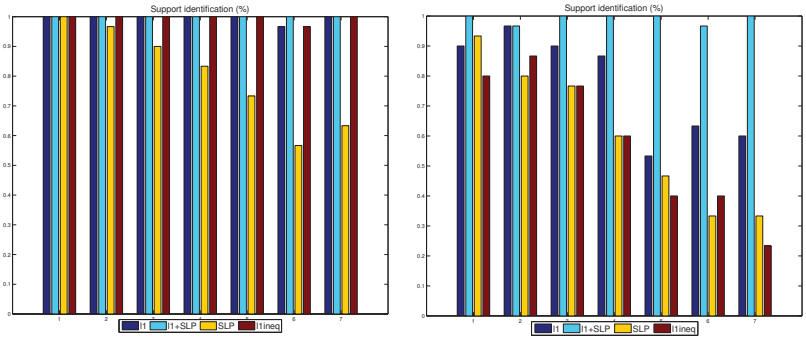
FIG. 9. The subfigures represent the errors, the speed of convergence, and the separation of the component. For more details on the displayed data we refer to the caption of Figure 6.

In Figure 12 we present phase transition diagrams of success rates in support recovery for ℓ_1 and ℓ_1+SLP in presence of nearly maximally allowed noise, i.e., $0.8 = r > \delta = 0.75$. In this test we varied the dimension of the measurement vector $m = 1,\dots,N$ with $N = 100$, and solved 20 different problems for all the admissible $k = \#S_r(x) = 1,\dots,m$. The graph on the left hand side reports the results obtained by ℓ_1-minimization, while the one on the right hand side refers to ℓ_1+SLP-minimization. We colored black all the points (m,k),



(a) (b)

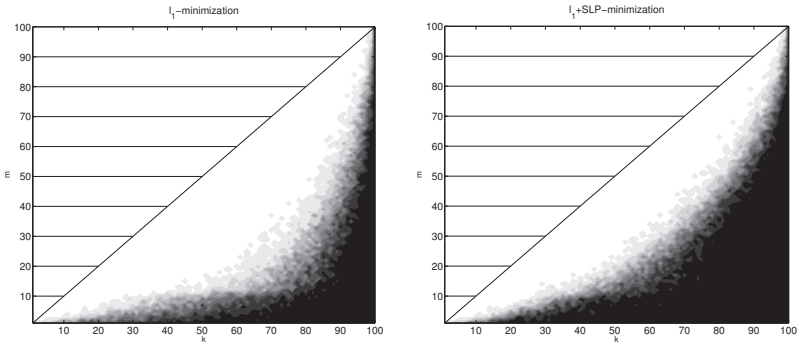
FIG. 10. The subfigures represent the error on the relevant entries. For more details on the displayed data we refer to the caption of Figure 6.



(a) (b)

FIG. 11. Support identification statistics. For more details on the displayed data we refer to the caption of Figure 6.

with $k \leqslant m$, which reported 100% of correct support identification, and gradually we reduce the tone up to white for the 0% result. A visual comparison of the corresponding phase transitions confirms our previous expectations. In particular, ℓ_1+SLP very significantly outperforms ℓ_1 in terms of correct support recovery.



(a) (b)

FIG. 12. Phase transition diagrams. The black area represents the couple (m,k) for which we had 100% of support recovery. The graph on the left hand side reports the results of ℓ_1-minimization and on the right hand side it is displayed the ones of ℓ_1+SLP. Note that the area for $k > m$ is not admissible.

Funding

Marco Artina acknowledges the support of the International Research Training Group IGDK 1754 "Optimization and Numerical Analysis for Partial Differential Equations with Nonsmooth Structures" of the German Science Foundation. Massimo Fornasier acknowledges the support of the ERC-Starting Grant HDSPCONTR "High-Dimensional Sparse Optimal Control". Steffen Peter acknowledges the support of the Project "SparsEO: Exploiting the Sparsity in Remote Sensing for Earth Observation" funded by Munich Aerospace.

References

[1] ARIAS-CASTRO, E. & ELDAR, Y. C. (2011) Noise folding in compressed sensing. *IEEE Signal Process. Lett.*, pp. 478–481.

[2] ARTINA, M., FORNASIER, M. & SOLOMBRINO, F. (2012) Linearly constrained nonsmooth and nonconvex minimization. *Preprint*.

[3] BARANIUK, R. (2007) Compressive sensing. *IEEE Signal Process. Magazine*, **24**(4), 118–121.

[4] BARANIUK, R. G., DAVENPORT, M., DEVORE, R. A. & WAKIN, M. (2008) A simple proof of the restricted isometry property for random matrices. *Constr. Approx.*, **28**(3), 253–263.

[5] CANDÈS, E. & WAKIN, M. (2008) An introduction to compressive sampling. *IEEE Signal Process. Magazine*, **25**(2), 21–30.

[6] CANDÈS, E. J., ROMBERG, J. & TAO, T. (2006) Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.*, **59**(8), 1207–1223.

[7] COHEN, A., DAHMEN, W. & DEVORE, R. A. (2009) Compressed sensing and best k-term approximation. *J. Amer. Math. Soc.*, **22**(1), 211–231.

[8] CVX RESEARCH, I. (2012) CVX: Matlab Software for Disciplined Convex Programming, version 2.0. http://cvxr.com/cvx.

[9] DAVENPORT, M. A., LASKA, J. N., TREICHLER, J. R. & BARANIUK, R. G. (2012) The Pros and Cons of Compressive Sensing for Wideband Signal Acquisition: Noise Folding vs. Dynamic Range. *IEEE Transactions on Signal Processing*, **60**(9), 4628–4642.

[10] DONOHO, D. L. (2006) Compressed sensing. *IEEE Transactions on Information Theory*, **52**(4), 1289–1306.

[11] FORNASIER, M. & RAUHUT, H. (2011) Compressive Sensing. in *Handbook of Mathematical Methods in Imaging*, ed. by O. Scherzer, pp. 187–228. Springer.

[12] FORNASIER, M. & WARD, R. (2010) Iterative thresholding meets free-discontinuity problems. *Found. Comput. Math.*, **10**(5), 527–567.

[13] FOUCART, S. & RAUHUT, H. (in preparation) *A Mathematical Introduction to Compressive Sensing*, Appl. Numer. Harmon. Anal. Birkhäuser, Boston.

[14] GRANT, M. & BOYD, S. (2008) Graph implementations for nonsmooth convex programs. in *Recent Advances in Learning and Control*, ed. by V. Blondel, S. Boyd, & H. Kimura, Lecture Notes in Control and Information Sciences, pp. 95–110. Springer-Verlag Limited, http://stanford.edu/∼boyd/graph_dcp.html.

[15] HÜGEL, M., RAUHUT, H. & STROHMER, T. (2012) Remote sensing via ℓ_1-minimization. *Preprint*.

[16] MALLAT, S. (2008) *A Wavelet Tour of Signal Processing: The Sparse Way*. Academic Press, 3rd edn.

[17] NEEDELL, D. & TROPP, J. A. (2008) CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Appl. Comput. Harmon. Anal.*, **26**(3), 301–321.

[18] NEEDELL, D. & VERSHYNIN, R. (2009) Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit. *Found. Comput. Math.*, **9**(3), 317–334.

[19] NEEDELL, D. & VERSHYNIN, R. (2010) Signal recovery from incomplete and inaccurate measurements via regularized orthogonal matching pursuit. *IEEE J. Sel. Topics Sig. Process.*, **4**(2), 310 – 316.

[20] RAUHUT, H. (2010) Compressive sensing and structured random matrices. in *Theoretical Foundations and Numerical Methods for Sparse Recovery*, ed. by M. Fornasier, vol. 9 of *Radon Series Comp. Appl. Math.*, pp. 1–92. deGruyter.

[21] WOJTASZCZYK, P. (2010) Stability and instance optimality for Gaussian measurements in compressed sensing. *Found. Comput. Math.*, **10**(1), 1–13.

List of Figures