Sparse Signal Approximation via Non-separable Regularization

Ivan Selesnick and Masoud Farshchian

***Abstract*—The calculation of a sparse approximate solution to a linear system of equations is often performed using either (1) L1-norm regularization and convex optimization or (2) non-convex regularization and non-convex optimization. Combining these principles, this paper describes a type of non-convex regularization that maintains the convexity of the objective function, thereby allowing the calculation of a sparse approximate solution via convex optimization. The preservation of convexity is viable in the proposed approach because it uses a regularizer that is non-separable. The proposed method is motivated and demonstrated by the calculation of sparse signal approximation using tight frames. Examples of denoising demonstrate improvement relative to L1 norm regularization.**

***Index Terms*—sparse signal model, sparse approximation, denoising, convex function, optimization**

I. Introduction

Sparse representations are used in applications such as noise reduction, deblurring, filling in missing data, tomography, and compressed sensing [79]. A basic step in many algorithms for these applications is the calculation of a sparse solution or sparse *approximate* solution to an ill-conditioned or highly under-determined system of linear equations $y = Ax$ [11]. A widely used approach to find a sparse approximate solution is to minimize the objective function $J\colon \mathbb{R}^N \to \mathbb{R}$,

$$J(x) = \frac{1}{2}\|y - Ax\|_2^2 + \lambda \|x\|_1, \quad \lambda > 0 \tag{1}$$

comprising a quadratic fidelity term and an ℓ_1 norm regularization (or 'penalty') term. In particular, basis pursuit denoising (BPD) [20] performs noise reduction this way. In BPD, y represents a signal in zero-mean noise. The BPD approach is effective precisely when the signal to be estimated admits a sparse approximation with respect to A; i.e., the signal can be expressed or approximated as a linear combination of relatively few columns of A. The objective function J is convex; hence, efficient algorithms to calculate a minimizer are available [24]. Problem (1) is also referred to as a lasso (least absolute shrinkage and selection operator) [80].

It has been shown that improved results can be obtained by replacing the ℓ_1 norm in (1) by a suitably chosen non-convex function [2], [12], [17], [18], [39], [53], [58], [59], [88]. That is, sparser solutions can be obtained with the same approximation error, or similarly, equally sparse solutions can be obtained with reduced approximation error. This leads to improved denoising, etc. However, replacing the ℓ_1 norm by a non-convex function leads generally (but not necessarily) to the objective function J being non-convex, thereby complicating the process in general: (*i*) Algorithms may sometimes fail to converge to a global minimizer. (*ii*) The global minimizer (if unique) may vary abruptly as λ is varied.

In this paper, we propose a family of non-convex multivariate penalty functions that preserve the convexity of the objective function to be minimized. Our goal is to improve upon ℓ_1-norm regularization while preserving a convex formulation. We consider the objective function $F\colon \mathbb{R}^N \to \mathbb{R}$,

$$F(x) = \frac{1}{2}\|y - Ax\|_2^2 + \lambda\, \psi(x), \quad \lambda > 0 \tag{2}$$

where the non-convex penalty function $\psi\colon \mathbb{R}^N \to \mathbb{R}$ is to be chosen so that F is convex.

The proposed multivariate penalty is constructed by subtracting a smooth convex function from the ℓ_1 norm. The properties of the penalty therefore depend on the properties of this convex function. The type of penalty function we propose is *non-separable*, meaning it can not be written as $\psi(x) = \sum_n \phi(x_n)$. The penalty we propose is given in (63) in Theorem 2 which is the main result.

For the proposed multivariate sparse regularization (MUSR) approach, the objective function F can be minimized using the same efficient proximal algorithms used for ℓ_1-norm minimization. Specifically, the forward-backward splitting (FBS) algorithm can be used to derive a matrix-free algorithm to minimize the objective function F.

In this work, we consider A to be an arbitrary matrix. It need not be injective nor surjective. In particular, A can be a wide matrix, i.e., $A^{\mathsf{T}}A$ is highly rank deficient. (This is in contrast to our earlier work.) In our numerical examples, we consider primarily wide matrices A for which $AA^{\mathsf{T}} = pI$ for some $p > 0$, i.e., the columns of A form a tight frame. We illustrate the proposed method for signal denoising via sparse signal approximation (SSA).

A. Related work

This work is related to recent papers on the formulation of *convex* objective functions for various linear inverse problems using *non-convex* sparsity-inducing penalties [4], [7], [19], [27], [40], [46], [47], [49], [62], [63], [70], [73]. However, these papers are of limited applicability when $A^{\mathsf{T}}A$ is highly rank deficient in problem (2). These papers use *separable* (additive) penalty functions, i.e., $\psi(x) = \sum_n \phi(x_n)$, which are fundamentally limited in this context. We recently proposed a bivariate non-separable penalty to overcome this limitation

Ivan Selesnick is with the Department of Electrical and Computer Engineering, Tandon School of Engineering, New York University, NY, USA. Masoud Farschian is with Empyreal Waves LLC, VA, USA. Email: selesi@nyu.edu and mfarsh@gmail.com.

This work was supported by NSF under grant CCF-1525398 and ONR under grant N00014-15-1-2314.

TABLE I
UNIVARIATE PENALTIES SATISFYING PROPERTY 1.

Log	$\phi(t) = \log(1 + \lvert t\rvert)$
Rat	$\phi(t) = \dfrac{\lvert t\rvert}{1 + \lvert t\rvert/2}$
Atan	$\phi(t) = \frac{2}{\sqrt{3}}\left(\tan^{-1}\left(\frac{1+2\lvert t\rvert}{\sqrt{3}}\right) - \frac{\pi}{6}\right)$
Exp	$\phi(t) = 1 - \mathrm{e}^{-\lvert t\rvert}$
MC	$\phi(t) = \begin{cases} \lvert t\rvert - \frac{1}{2}t^2, & \lvert t\rvert \leqslant 1 \\ \frac{1}{2}, & \lvert t\rvert \geqslant 1 \end{cases}$

[71], but its effectiveness for $N > 2$ variables is limited to a narrow class of problems.

This work is related to several other prior papers. The formulation of convex objective functions with non-convex penalties for signal processing was pioneered by Blake, Zisserman, and Nikolova who used non-convex separable penalties in the graduated non-convexity (GNC) technique [9], [57], [60], [61] and binary image estimation [56]. Additionally, non-convex non-separable penalties have been proposed by Tipping [81] and Wipf [86] to strongly induce sparsity. On the other hand, we are interested here in problems where both the objective function is convex and the penalty is non-separable.

This work is related more generally to the literature on techniques designed to outperform ℓ_1 norm regularization for sparse approximation. Methods based on the ℓ_p pseudo-norm ($0 \leqslant p < 1$) and other penalty functions have been developed [2], [15]–[17], [21], [28], [29], [33]–[35], [48], [52], [54], [55], [90]. Algorithms that seek directly to obtain (approximate) sparse solutions have also been developed: matching pursuit [51], greedy ℓ_1 [45], iterative thresholding [10], [43], [50], [66], [83], [84], [87], single best replacement [77], [78], smoothed ℓ_0 [54], and smoothed ℓ_1/ℓ_2 [68]. The continuous exact ℓ_0 (CEL0) penalty [76] and the work of Ref. [13] aim to approximate the convex hull of the ℓ_0 pseudo-norm regularized least squares objective function, so as to reduce the number of extraneous non-optimal local minimizers.

In addition to these methods, a novel approach for the calculation of a global minimizer of a non-convex sparse deconvolution problem was recently proposed using a hierarchy of semidefinite programming relaxations [14].

B. Notation

The vector $x \in \mathbb{R}^N$ is written $x = (x_1,\ x_2,\ \ldots,\ x_N)$. The ℓ_1 and ℓ_2 norms of $x \in \mathbb{R}^N$ are defined as $\|x\|_1 = \sum_n |x_n|$ and $\|x\|_2 = \left(\sum_n |x_n|^2\right)^{1/2}$, respectively. If the matrix $A - B$ is positive semidefinite, we write $A \succcurlyeq B$. The matrix norm $\|A\|_1$ is defined as

$$\|A\|_1 = \max_j \sum_i |A_{i,j}|. \tag{3}$$

Also, $\|A\|_2^2$ is defined as the maximum eigenvalue of $A^{\mathsf{T}}A$. We also use the notation $\mathbb{R}_+ = \{x \in \mathbb{R} : x \geqslant 0\}$ and $\mathbb{R}_+^* = \{x \in \mathbb{R} : x > 0\}$.

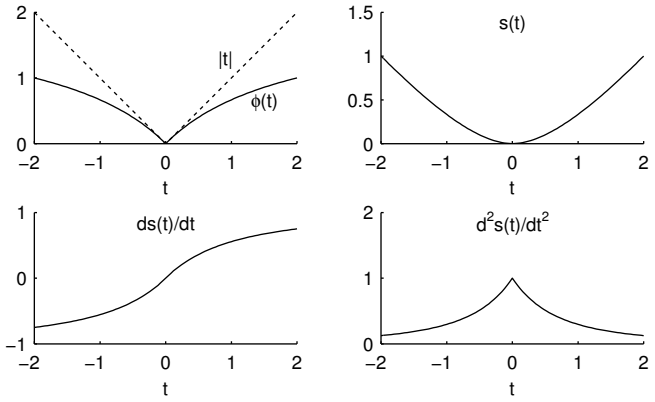

Fig. 1. The rational penalty ϕ, the corresponding function s, and its derivatives.

II. UNIVARIATE PENALTIES

We define a class of non-convex univariate penalty functions $\phi\colon \mathbb{R} \to \mathbb{R}$ as follows.

Property 1. *The penalty function $\phi\colon \mathbb{R} \to \mathbb{R}$ satisfies the following properties.*

1) *ϕ is continuous on $\mathbb{R}$*
2) *ϕ is continuously differentiable, non-decreasing, and concave on $\mathbb{R}_+$*
3) *ϕ' is convex on $\mathbb{R}_+$.*
4) $\phi(0) = 0$
5) $\phi(-t) = \phi(t)$
6) $\phi'(0^+) = 1$
7) *$t \to t^2/2 + \phi(t)$ is convex on $\mathbb{R}$*
8) $\phi''(t) \to 0$ *as* $t \to \infty$
9) $\phi(t) \geqslant |t| - t^2/2$ *for all* $t \in \mathbb{R}$

Table I lists several penalty functions (penalties) satisfying Property 1: the logarithmic [12], [59], rational [34], [59], arctangent [70], and exponential [47], [49], [54] penalties, and the minimax-concave (MC) penalty [6], [64], [89]. The rational and MC penalties are illustrated in Figs. 1 and 2, respectively. We note that point 7 in Property 1 implies that ϕ is 'weakly' convex [6], [18].

Definition 1. *Let penalty $\phi\colon \mathbb{R} \to \mathbb{R}$ satisfy Property 1. We define $s\colon \mathbb{R} \to \mathbb{R}$, as*

$$s(t) = |t| - \phi(t). \tag{4}$$

It will be useful later to write $\phi(t)$ as $|t| - s(t)$. Figures 1 and 2 illustrate the function s corresponding to ϕ. The first two derivatives of s are also illustrated. We note that when ϕ is the MC penalty, then s is the Huber function [41],

$$s_0(t) := \begin{cases} \frac{1}{2}t^2, & |t| \leqslant 1 \\ |t| - \frac{1}{2}, & |t| \geqslant 1. \end{cases} \tag{5}$$

For the representative penalties in Table I, the function s satisfies

$$0 \leqslant s(t) \leqslant |t| \tag{6}$$

and

$$s(t) \approx \tfrac{1}{2}t^2 \text{ as } t \to 0. \tag{7}$$

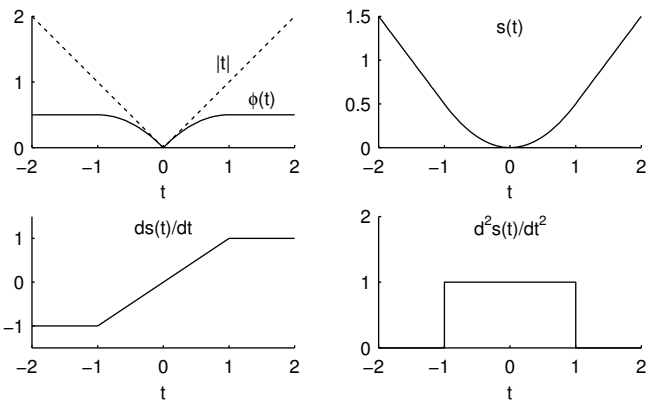

Fig. 2. The minimax concave (MC) penalty ϕ, the corresponding function s, and its derivatives. The function s here is the Huber function.

The function s defined in (4) is a convex function. Furthermore, since the derivative ϕ' is convex on $\mathbb{R}_+$, it follows that the derivative s' is concave on $\mathbb{R}_+$. Further properties of s based on Property 1 are listed in Proposition 1.

Proposition 1. *Let penalty $\phi\colon \mathbb{R} \to \mathbb{R}$ satisfy Property 1. Then the corresponding function $s\colon \mathbb{R} \to \mathbb{R}$ defined in* (4) *satisfies the following properties.*

1) *s is continuously differentiable and convex on $\mathbb{R}$*
2) *s' is concave on $\mathbb{R}_+$*
3) $s(0) = 0$
4) $s(-t) = s(t)$
5) $s'(0) = 0$
6) *$t \to t^2/2 - s(t)$ is convex on $\mathbb{R}$*
7) *$s''(t) \to 0$ as $t \to \infty$*
8) *$s(t) \leqslant t^2/2$ for all $t \in \mathbb{R}$*

Proof. We prove s is differentiable at zero. We have $s'(0^+) = 1 - \phi'(0^+) = 0$. Due to symmetry of ϕ, we have $\phi'(0^-) = -\phi'(0^+) = -1$. Hence, $s'(0^-) = -1 - \phi'(0^-) = 0$. That gives the equality: $s'(0^-) = s'(0^+) = 0$. The rest of the proposition is straightforward. □

The MC penalty and the Huber function (5) play a particular role in this work. We will use the representation of the Huber function as a Moreau envelope [3], [65].

Proposition 2. *Let $\phi_0\colon \mathbb{R} \to \mathbb{R}$ be the MC penalty (Table I). Let $s_0\colon \mathbb{R} \to \mathbb{R}$ be the convex function correspondingly defined by* (4)*. The function s_0 in* (5) *(the Huber function) can be expressed as the Moreau envelope of the absolute value function, i.e.,*

$$s_0(t) = \min_{\tau \in \mathbb{R}} \Big\{ |\tau| + \tfrac{1}{2}(t-\tau)^2 \Big\}. \tag{8}$$

This is noted, for example, in [23] and Sec. 3.1 of Ref. [65]. The proof comprises a straightforward calculation.

Lemma 1. *Let ϕ be a univariate penalty satisfying Property 1 with the additional property that the corresponding function s defined by* (4) *is three-times continuously differentiable on $\mathbb{R}_+^*$. Let s_0 be the Huber function. Then s can be written as a scale mixture of s_0, i.e.,*

$$s(t) = \int_0^\infty w(a)\, s_0(t/a)\, \mathrm{d}a. \tag{9}$$

The weight function w is given by

$$w(a) = -a^2\, s'''(a), \quad a > 0. \tag{10}$$

Proof. For the Huber function, we have

$$s_0''(t) = \begin{cases} 1, & |t| < 1 \\ 0, & |t| > 1. \end{cases} \tag{11}$$

Following (9), define

$$f(t) = \int_0^\infty w(a)\, s_0(t/a)\, \mathrm{d}a. \tag{12}$$

Then

$$f''(t) = \int_0^{|t|} w(a)\, \frac{1}{a^2}\, s_0''(t/a)\, \mathrm{d}a + \int_{|t|}^\infty w(a)\, \frac{1}{a^2}\, s_0''(t/a)\, \mathrm{d}a. \tag{13}$$

We write the integral in two parts here because s_0 is not twice differentiable at 1. Using (11), we have

$$f''(t) = \int_{|t|}^\infty w(a)\, \frac{1}{a^2}\, \mathrm{d}a. \tag{14}$$

Using (10) we have

$$f''(t) = -\int_{|t|}^\infty s'''(a)\, \mathrm{d}a \tag{15}$$

$$= s''(t) - \lim_{a \to \infty} s''(a) \tag{16}$$

$$= s''(t). \tag{17}$$

where we use the property $s''(t) \to 0$ as $t \to \infty$. Since $s(0) = s'(0) = 0$, it follows that $f = s$, proving (9).

Note that the weight function w in (10) is non-negative because by assumption ϕ' is convex on $\mathbb{R}_+$, hence s' is concave on $\mathbb{R}_+$ and s''' is non-positive on $\mathbb{R}_+$. □

The first four penalties listed in Table I satisfy the hypothesis of Lemma 1. Therefore, the function s corresponding to these penalties can be written as a scale mixture (9) of the Huber function s_0. For example, when ϕ is the rational penalty ('Rat' in Table I), the weight function w in (10) is given by

$$w(a) = \frac{24a^2}{(2+a)^4}. \tag{18}$$

III. Multivariate Penalties

Based on a given univariate penalty ϕ satisfying Property 1, we will define a multivariate penalty $\psi\colon \mathbb{R}^N \to \mathbb{R}$. For this purpose, we first define a multivariate analog of the corresponding function s in (4).

Definition 2. *Let $\phi\colon \mathbb{R} \to \mathbb{R}$ satisfy Property 1. Let $s\colon \mathbb{R} \to \mathbb{R}$ be correspondingly defined in* (4)*. We define $S\colon \mathbb{R}^N \to \mathbb{R}$ as*

$$S(x) = \sum_n s(x_n). \tag{19}$$

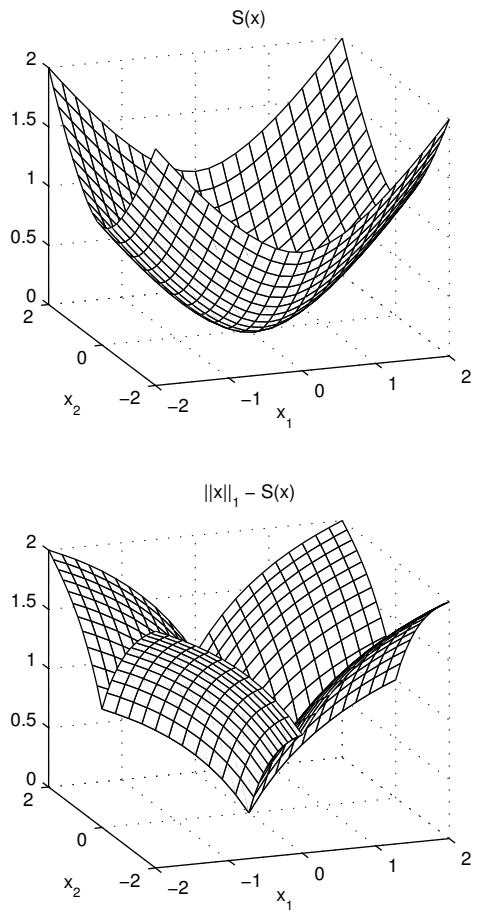

Fig. 3. Separable functions corresponding to the rational penalty. Subtracting the differentiable convex function S from the ℓ_1-norm yields a non-differentiable non-convex penalty that strongly promotes sparsity.

Proposition 3. *Let penalty $\phi\colon \mathbb{R} \to \mathbb{R}$ satisfy Property 1. Then the corresponding function $S\colon \mathbb{R}^N \to \mathbb{R}$ defined in* (19) *has the properties:*

1) *S is continuously differentiable and convex on $\mathbb{R}^N$*
2) *$S(0) = 0$*
3) *$x \to \|x\|_2^2/2 - S(x)$ is convex on $\mathbb{R}^N$*
4) *$S(x) \leqslant \|x\|_2^2/2$ for all $x \in \mathbb{R}^N$*

For the representative penalties in Table I, the function S also satisfies

$$0 \leqslant S(x) \leqslant \|x\|_1 \tag{20}$$

and

$$S(x) \approx \tfrac{1}{2}\|x\|_2^2 \quad \text{as } x \to 0. \tag{21}$$

Separable (additive) penalties can be expressed in terms of S,

$$\sum_n \phi(x_n) = \sum_n \big(|x_n| - s(x_n)\big) \tag{22}$$

$$= \|x\|_1 - S(x). \tag{23}$$

An example in $\mathbb{R}^2$ is shown in Fig. 3, which illustrates the separable function S and associated penalty, when ϕ is the rational penalty.

The non-separable regularizer to be described in Section IV is given in terms of the composition of S and a linear mapping. In order to determine suitable parameter values, we consider how $S(x)$ can be scaled so as to tightly approximate $S(Ax)$ from above. This result is given in Theorem 1, which is the focus of the remainder of this section.

The following follows straightforwardly from Proposition 2.

Proposition 4. *Let $\phi_0\colon \mathbb{R} \to \mathbb{R}$ be the MC penalty (i.e., s_0 is the Huber function). Then the separable function $S_0\colon \mathbb{R}^N \to \mathbb{R}$ defined in* (19) *can be expressed as the Moreau envelope of the ℓ_1 norm, i.e.,*

$$S_0(x) = \min_{v \in \mathbb{R}^N} \Big\{ \|v\|_1 + \tfrac{1}{2}\|x - v\|_2^2 \Big\}. \tag{24}$$

Lemma 2. *Let ϕ_0 be the MC penalty (Table I). Let S_0 be correspondingly defined by* (19)*. Let $A \in \mathbb{R}^{M \times N}$. If $A \neq 0$, then S_0 satisfies*

$$S_0(Ax) \leqslant \frac{\|A\|_1^2}{\|A\|_2^2}\, S_0\bigg(\frac{\|A\|_2^2}{\|A\|_1}\, x \bigg) \tag{25}$$

for all $x \in \mathbb{R}^N$.

Proof. The proof will use the Moreau envelope representation of S_0. Using (24) we have

$$S_0(Ax) = \min_{u \in \mathbb{R}^M} \Big\{ \|u\|_1 + \tfrac{1}{2}\|Ax - u\|_2^2 \Big\} \tag{26}$$

$$\leqslant \min_{v \in \mathbb{R}^N} \Big\{ \|Av\|_1 + \tfrac{1}{2}\|Ax - Av\|_2^2 \Big\} \tag{27}$$

$$\leqslant \min_{v \in \mathbb{R}^N} \Big\{ \|A\|_1 \|v\|_1 + \tfrac{1}{2}\|A\|_2^2\, \|x - v\|_2^2 \Big\}. \tag{28}$$

To obtain (27) we used $\{Av, v \in \mathbb{R}^N\} \subseteq \mathbb{R}^M$. To obtain (28) we used

$$\|Ax\|_1 \leqslant \|A\|_1 \|x\|_1, \tag{29}$$

$$\|Ax\|_2 \leqslant \|A\|_2 \|x\|_2, \tag{30}$$

for all $x \in \mathbb{R}^N$.

Using (24) we similarly have

$$S_0\bigg(\frac{\|A\|_2^2}{\|A\|_1}\, x \bigg) \tag{31}$$

$$= \min_{v \in \mathbb{R}^N} \left\{ \|v\|_1 + \frac{1}{2} \left\| \frac{\|A\|_2^2}{\|A\|_1}\, x - v \right\|_2^2 \right\} \tag{32}$$

$$= \min_{v \in \mathbb{R}^N} \left\{ \left\| \frac{\|A\|_2^2}{\|A\|_1}\, v \right\|_1 + \frac{1}{2} \left\| \frac{\|A\|_2^2}{\|A\|_1}\, x - \frac{\|A\|_2^2}{\|A\|_1}\, v \right\|_2^2 \right\} \tag{33}$$

$$= \min_{v \in \mathbb{R}^N} \left\{ \frac{\|A\|_2^2}{\|A\|_1} \|v\|_1 + \frac{1}{2} \frac{\|A\|_2^4}{\|A\|_1^2} \|x - v\|_2^2 \right\} \tag{34}$$

$$= \frac{\|A\|_2^2}{\|A\|_1^2} \cdot \min_{v \in \mathbb{R}^N} \Big\{ \|A\|_1 \|v\|_1 + \tfrac{1}{2}\|A\|_2^2 \|x - v\|_2^2 \Big\} \tag{35}$$

$$\geqslant \frac{\|A\|_2^2}{\|A\|_1^2}\, S_0(Ax) \tag{36}$$

where we use (28) in the last line. □

Lemma 3. *Let $S_0\colon \mathbb{R}^N \to \mathbb{R}$ satisfy* (25)*. Let $w\colon \mathbb{R}_+ \to \mathbb{R}_+$. Then a new function $S\colon \mathbb{R}^N \to \mathbb{R}$ defined as*

$$S(x) := \int_0^\infty w(a)\, S_0(x/a)\, \mathrm{d}a \tag{37}$$

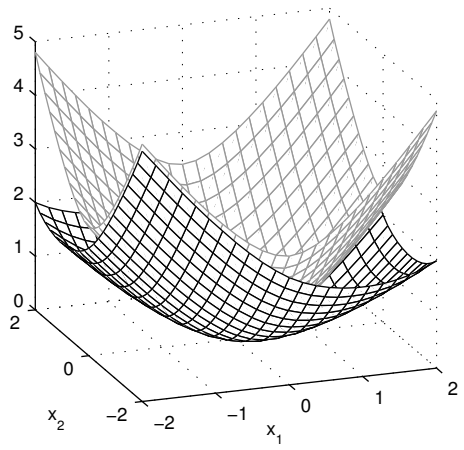

Fig. 4. Illustration of inequality (41). The upper and lower bounds are separable and non-separable, respectively. The inequality seeks an isotropic scaling of the separable function S (upper bound) so as to tightly approximate the non-separable scaling of S by matrix A (lower bound).

also satisfies (25).

Proof. Using (37) we have

$$S(Ax) = \int_0^\infty w(a)\, S_0(Ax/a)\,\mathrm{d}a \tag{38}$$

$$\leqslant \frac{\|A\|_1^2}{\|A\|_2^2} \int_0^\infty w(a)\, S_0\Big(\frac{\|A\|_2^2}{\|A\|_1}\, x/a\Big)\,\mathrm{d}a \tag{39}$$

$$= \frac{\|A\|_1^2}{\|A\|_2^2}\, S\Big(\frac{\|A\|_2^2}{\|A\|_1}\, x\Big) \tag{40}$$

where we use (25) to obtain the inequality. □

Theorem 1. *Let ϕ be a univariate penalty satisfying Property 1 with the additional property that the corresponding function s defined in* (4) *can be written as a scale mixture* (9) *of the Huber function s_0. Let $S\colon \mathbb{R}^N \to \mathbb{R}$ be correspondingly defined in* (19)*. Let $A \in \mathbb{R}^{M\times N}$. Then S satisfies*

$$S(Ax) \leqslant \frac{\|A\|_1^2}{\|A\|_2^2}\, S\Big(\frac{\|A\|_2^2}{\|A\|_1}\, x\Big) \tag{41}$$

for all $x \in \mathbb{R}^N$.

Proof. From (9) and (19), we have

$$S(x) = \int_0^\infty w(a)\, S_0(x/a)\,\mathrm{d}a. \tag{42}$$

By Lemma 2, S_0 satisfies (25). Hence, by Lemma 3, S satisfies (41). □

By Lemma 1, the function S corresponding to each penalty listed in Table I satisfies (41). For example, Fig. 4 illustrates (41) where

$$A = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \end{bmatrix} \tag{43}$$

and ϕ is the rational penalty.

IV. Sparse Regularization

Based on a given univariate penalty ϕ satisfying Property 1, we will define a multivariate penalty $\psi\colon \mathbb{R}^N \to \mathbb{R}$ as

$$\psi(x) = \|x\|_1 - \alpha\, S(\beta B x) \tag{44}$$

where S is given by (19), α and β are appropriate scalars, and B is an appropriate matrix. In the following, we address how to set α, β, and B.

Lemma 4. *Let ϕ be a univariate penalty satisfying Property 1. Let s and S be correspondingly defined by* (4) *and* (19)*. Let $A \in \mathbb{R}^{M\times N}$. Let $F\colon \mathbb{R}^N \to \mathbb{R}$ be the objective function*

$$F(x) = \frac{1}{2}\|y - Ax\|_2^2 + \lambda\, \psi(x) \tag{45}$$

where $\lambda > 0$ and the multivariate penalty $\psi\colon \mathbb{R}^N \to \mathbb{R}$ is given by

$$\psi(x) = \|x\|_1 - \alpha\, S(\beta B x) \tag{46}$$

where $\alpha \geqslant 0$, $\beta \geqslant 0$, and $B \in \mathbb{R}^{L\times N}$ is a matrix such that $B^{\mathsf{T}}B \preccurlyeq A^{\mathsf{T}}A$. Then F is a convex function if

$$\alpha\beta^2 \leqslant 1/\lambda. \tag{47}$$

Proof. Define $G\colon \mathbb{R}^N \to \mathbb{R}$ as

$$G(x) = \tfrac{1}{2}\|Ax\|_2^2 - \lambda\alpha\, S(\beta B x). \tag{48}$$

Then $F(x) = G(x) + \lambda\|x\|_1 - y^{\mathsf{T}}Ax + \|y\|_2^2/2$, i.e., F is the sum of G and a convex function. Hence, F is convex if G is convex. Hence, it is sufficient to show G is convex. We write

$$G(x) = G_1(x) + G_2(x) \tag{49}$$

where

$$G_1(x) = \tfrac{1}{2}\|Ax\|_2^2 - \tfrac{1}{2}\lambda\alpha\|\beta B x\|_2^2 \tag{50}$$

$$G_2(x) = \tfrac{1}{2}\lambda\alpha\|\beta B x\|_2^2 - \lambda\alpha\, S(\beta B x). \tag{51}$$

The function G_1 is convex if $A^{\mathsf{T}}A \succcurlyeq \lambda\alpha\beta^2 B^{\mathsf{T}}B$. Since it is given that $A^{\mathsf{T}}A \succcurlyeq B^{\mathsf{T}}B$ and $1 \geqslant \lambda\alpha\beta^2$, it follows that G_1 is convex.

We now show G_2 is convex. Since $t \to t^2/2 - s(t)$ is convex by Proposition 1, it follows that $f\colon \mathbb{R}^N \to \mathbb{R}$ defined by $f(x) = \|x\|_2^2/2 - S(x)$ is convex. Hence $G_1 = \lambda\alpha f \circ \beta B$ is convex(because the composition of a convex function with a linear functional is convex, and the multiplication of a convex function by a positive number is convex).

Since G_1 and G_2 are convex, G is convex. □

Condition (47) by itself is not sufficient to properly set α and β. Figure 5 illustrates a function ψ of the form (44) where α and β satisfy (47); but ψ is decreasing over some of its domain and even becomes negative. Such a function is generally not considered a suitable penalty. To avoid this behavior, we will prescribe a non-negative function and we will see to it that the penalty is greater than or equal to the prescribed function. (See Property 2 below.) The lower bound will be derived by considering the case where the data fidelity term is separable.

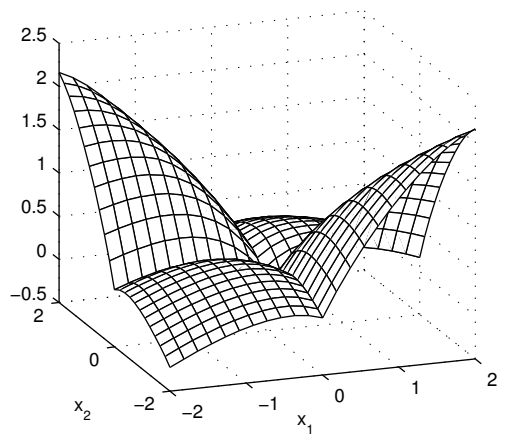

Fig. 5. Badly formed penalty (44). Inappropriately chosen parameter values yield a function that is not a useful sparsity-inducing penalty. (The function is negative on part of its domain.)

A. Separable case

As a baseline, it is useful to consider the special case where A is a scaled identity matrix, i.e., $A = \rho I$ where $\rho > 0$. We will use this case to guide the choice of parameters for the general case. The following lemma addresses how to set a separable penalty ψ so that F in (45) is convex when $A = \rho I$.

Lemma 5. *Let ϕ be a univariate penalty satisfying Property 1. Let s and S be correspondingly defined by* (4) *and* (19)*. Let $F\colon \mathbb{R}^N \to \mathbb{R}$ be the objective function*

$$F(x) = \frac{1}{2}\|y - \rho\, x\|_2^2 + \lambda\, \psi(x) \tag{52}$$

where $\lambda > 0$ and $\rho > 0$, and the separable multivariate penalty $\psi\colon \mathbb{R}^N \to \mathbb{R}$ is given by

$$\psi(x) = \frac{\lambda}{\gamma\rho^2} \sum_n \phi\Big(\frac{\gamma\rho^2}{\lambda} x_n\Big) \tag{53}$$

$$= \|x\|_1 - \frac{\lambda}{\gamma\rho^2}\, S\Big(\frac{\gamma\rho^2}{\lambda} x\Big). \tag{54}$$

Then F is a convex function if

$$0 < \gamma \leqslant 1. \tag{55}$$

Proof. The proof uses Lemma 4 with $A = \rho I$ in (45). The penalty (54) is given by (46) with $\alpha = \lambda/(\gamma\rho^2)$, $\beta = \gamma/(\lambda\rho)$, and $B = \rho I$. Hence, the convexity condition (47) is given by $\gamma \leqslant 1$. □

B. Lower bound function

This section prescribes a lower bound function that will be used to guide the setting of parameters of the proposed multivariate penalty. To obtain sparse approximate solutions to $y = Ax$, we minimize the objective function $F\colon \mathbb{R}^N \to \mathbb{R}$

$$F(x) = \frac{1}{2}\|y - Ax\|_2^2 + \lambda\, \psi(x) \tag{56}$$

where ψ is chosen such that F is convex. To induce sparsity more strongly than the ℓ_1 norm, our approach is to make ψ non-convex (specifically, weakly convex). The convexity of the quadratic term $\|Ax\|_2^2$ determines the allowed negative curvature of ψ. We consider only penalties that are tangent to the ℓ_1 norm at the origin (i.e., $\psi(x) \to \|x\|_1$ as $x \to 0$). Therefore, the negative curvature of ψ determines how slowly ψ may increase away from zero. We write

$$\psi_A^{\mathsf{LB}}(x) \leqslant \psi(x) \tag{57}$$

for some hypothetical lower bound function that depends on A. The more convex the quadratic term, the smaller the lower bound function, i.e.,

$$A_1^{\mathsf{T}} A_1 \preccurlyeq A_2^{\mathsf{T}} A_2 \quad \iff \quad \psi_{A_2}^{\mathsf{LB}}(x) \leqslant \psi_{A_1}^{\mathsf{LB}}(x). \tag{58}$$

If the quadratic term has no positive curvature, then F is convex only if ψ is also convex. Hence, as $A \to 0$ we have $\psi_A^{\mathsf{LB}}(x) \to \|x\|_1$.

Since $A^{\mathsf{T}}A \preccurlyeq \|A\|_2^2\, I$, it follows that a lower bound for ψ is given in turn by the lower bound $\psi_{\rho I}^{\mathsf{LB}}$, i.e.,

$$\psi_{\rho I}^{\mathsf{LB}}(x) \leqslant \psi_A^{\mathsf{LB}}(x) \leqslant \psi(x), \quad \rho = \|A\|_2. \tag{59}$$

When A is a scaled identity matrix, we obtain a specific lower bound using Lemma 5 with $\rho = \|A\|_2$ and $\gamma = 1$. This motivates defining the following property.

Property 2. *Let ϕ be a univariate penalty satisfying Property 1. Let S be correspondingly defined by* (19)*. Let $\lambda > 0$ and let A be a matrix of size $M \times N$. We consider the penalty $\psi\colon \mathbb{R}^N \to \mathbb{R}$ to be well formed if it satisfies*

$$\psi(x) \geqslant \frac{\lambda}{\|A\|_2^2} \sum_n \phi\Big(\frac{\|A\|_2^2}{\lambda} x_n\Big) \tag{60}$$

$$= \|x\|_1 - \frac{\lambda}{\|A\|_2^2}\, S\Big(\frac{\|A\|_2^2}{\lambda} x\Big) \tag{61}$$

for all $x \in \mathbb{R}^N$.

Condition (60) prevents the penalty ψ from straying too far from the ℓ_1 norm. If $\psi(x)$ violates the condition, then it can become negative for large x as illustrated in Fig. 5, which we wish to avoid (if large x were penalized less than $x = 0$, then totally non-sparse solutions would be more preferred than sparse solutions).

C. Sparsity-inducing non-separable penalty

Theorem 2 specifies the proposed multivariate penalty.

Theorem 2. *Let ϕ be a univariate penalty satisfying Property 1 with the additional property that the corresponding function s can be written as a scale mixture* (9) *of the Huber function s_0. Let S be correspondingly defined by* (19)*. Let $F\colon \mathbb{R}^N \to \mathbb{R}$ be the objective function*

$$F(x) = \frac{1}{2}\|y - Ax\|_2^2 + \lambda\, \psi(x) \tag{62}$$

where $\lambda > 0$ and the penalty $\psi\colon \mathbb{R}^N \to \mathbb{R}$ is given by

$$\psi(x) = \|x\|_1 - \frac{\lambda}{\gamma\|B\|_1^2}\, S\Big(\frac{\gamma\|B\|_1}{\lambda} Bx\Big) \tag{63}$$

where the non-zero matrix B satisfies $B^{\mathsf{T}}B \preccurlyeq A^{\mathsf{T}}A$. If $0 < \gamma \leqslant 1$, then F is convex and the penalty satisfies Property 2.

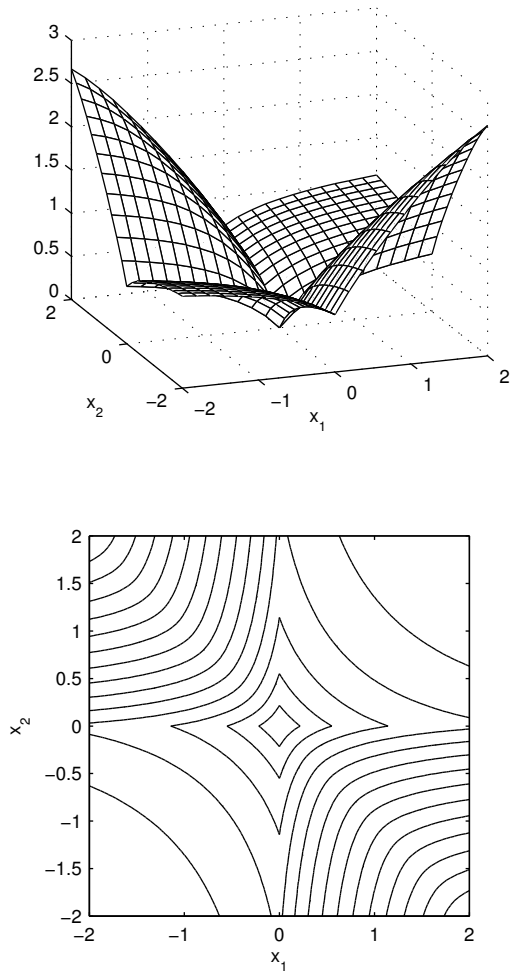

Fig. 6. Penalty ψ in (63) and its contour plot. The star-shaped contour plot is characteristic of non-convex sparsity-inducing penalties. The design of the penalty so as to preserve convexity of the objective function results in distinct behavior in different quadrants.

Proof. The penalty ψ in (63) has the form (46) where $\alpha = \lambda/(\gamma\|B\|_1^2)$ and $\beta = \gamma\|B\|_1/\lambda$. Thus $\alpha\beta^2 = \gamma/\lambda$. Hence, by Lemma 4, F is convex for $0 < \gamma \leqslant 1$.

The penalty ψ in (63) satisfies (60) for all $x \in \mathbb{R}^N$, if S satisfies

$$\frac{\lambda}{\gamma\|B\|_1^2}\, S\Big(\frac{\gamma\|B\|_1}{\lambda} Bx\Big) \leqslant \frac{\lambda}{\|A\|_2^2}\, S\Big(\frac{\|A\|_2^2}{\lambda}\, x\Big) \qquad (64)$$

for all $x \in \mathbb{R}^N$, where we have cancelled the ℓ_1 norm common to (60) and (63). The function S satisfies (64) for all $x \in \mathbb{R}^N$ if

$$S\Big(\frac{\gamma\|B\|_1}{\lambda} Bx\Big) \leqslant \gamma\, \frac{\|B\|_1^2}{\|A\|_2^2}\, S\Big(\frac{\|A\|_2^2}{\lambda}\, x\Big) \qquad (65)$$

for all $x \in \mathbb{R}^N$. The function S satisfies (65) for all $x \in \mathbb{R}^N$ if

$$S(Bx) \leqslant \gamma\, \frac{\|B\|_1^2}{\|A\|_2^2}\, S\Big(\frac{1}{\gamma}\, \frac{\|A\|_2^2}{\|B\|_1}\, x\Big) \qquad (66)$$

for all $x \in \mathbb{R}^N$. Since S is convex and $S(0) = 0$ and $0 < \gamma \leqslant 1$, we have $\gamma\, S(x) \geqslant S(\gamma\, x)$ for all $x \in \mathbb{R}^N$. Hence, S satisfies (66) for all $x \in \mathbb{R}^N$ if it satisfies the tighter condition

$$S(Bx) \leqslant \frac{\|B\|_1^2}{\|B\|_2^2}\, S\Big(\frac{\|A\|_2^2}{\|B\|_1}\, x\Big) \qquad (67)$$

for all $x \in \mathbb{R}^N$. Since $B^{\mathsf{T}}B \preccurlyeq A^{\mathsf{T}}A$, we have $\|B\|_2 \leqslant \|A\|_2$. Hence, S satisfies (67) for all $x \in \mathbb{R}^N$ if it satisfies the tighter condition

$$S(Bx) \leqslant \frac{\|B\|_1^2}{\|B\|_2^2}\, S\Big(\frac{\|B\|_2^2}{\|B\|_1}\, x\Big) \qquad (68)$$

for all $x \in \mathbb{R}^N$; just as (67) implies (66). By Theorem 1, S satisfies (68) for all $x \in \mathbb{R}^N$. □

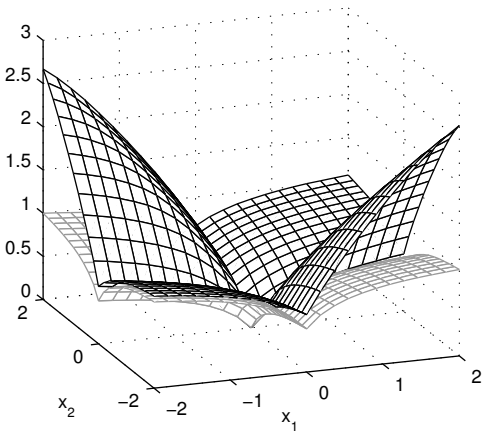

Fig. 7. Illustration that the penalty ψ in (63) satisfies the lower-bound condition (60). This avoids the undesirable behavior shown in Fig. 5.

Figure 6 illustrates the penalty ψ in (63) for the matrix A in (43), $\lambda = 1$, $\gamma = 1$, and $B = A$. The contours of ψ are star-shaped, which is characteristic of non-convex penalties. But the curvature of the contours is less pronounced in some quadrants relative to other quadrants. This behavior is because ψ is designed to maintain the convexity of the objective function F. Figure 7 illustrates that ψ satisfies inequality (60).

The matrix B determines the shape of ψ. When B is a diagonal matrix, then ψ is a separable (additive) function. When $B = \rho I$, the separable penalty (54) is retrieved as a special case of (63). (If $B = \rho I$, then $\|B\|_1 = |\rho|$.) In this case, B satisfies $B^{\mathsf{T}}B \preccurlyeq A^{\mathsf{T}}A$ only if ρ^2 is less than the minimum eigenvalue of $A^{\mathsf{T}}A$. If $B = \rho I$ and $A^{\mathsf{T}}A$ is singular, then $B^{\mathsf{T}}B \preccurlyeq A^{\mathsf{T}}A$ only if $B = 0$, which precludes non-convex regularization while maintaining convexity of the objective function F. Hence, non-diagonal B, i.e., non-separable regularization, is required in this case in order to induce sparsity more strongly than the ℓ_1 norm.

The parameter γ adjusts the degree of non-convexity of ψ. If γ is close to zero, then ψ is nearly convex. We have $\psi(x) \to \|x\|_1$ as $\gamma \to 0$.

We note that the proposed regularizer (63) depends on the linear operator A. Customarily, the regularizer is chosen independently of A. However, the dependence of the regularizer on A is a property of certain optimal estimators, as noted in works discussing connections between regularization-based and Bayesian-based estimation approaches [37], [38], [67].

D. Remark

The motivation for the proposed class of penalties is two-fold. First, many methods for obtaining sparse solutions use non-convex penalties of the separable form $\sum_n \phi(x_n)$ which can be written $\|x\|_1 - \sum_n s(x_n)$ where s is correspondingly defined. Second, it is of some interest to formulate problems as convex when possible, which often precludes penalties of the form $\|x\|_1 - \sum_n s(x_n)$. In particular, penalties of this form are precluded if the forward mapping A is non-injective which is the usual case for ill-conditioned linear inverse problems. The proposed penalty is aimed to capture the strongly sparsity-inducing behavior of penalties of the form $\sum_n \phi(x_n)$ while at the same time maintaining convexity of the objective function to be minimized.

We also remark on the question of how to prescribe the linear operator B. Our view is that B should in some sense approximate a scaled identity matrix. For if $B = cI$, then the proposed penalty is separable, as appropriate to induce pure (non-structured) sparsity. On the other hand, we aim that $B^{\mathsf{T}}B$ be close to $A^{\mathsf{T}}A$ while satisfying $B^{\mathsf{T}}B \preccurlyeq A^{\mathsf{T}}A$, so the penalty is 'as non-convex as it can be' while preserving convexity of the objective function to be minimized. However, given an arbitrary A, the form B should take to achieve these properties, remains an open question.

E. Additional properties

The following lemmas regard the differentiable part of the objective function F. These lemmas will be used in Section V to prove convergence of the iterative thresholding algorithm.

Lemma 6. *Let ϕ be a univariate penalty satisfying Property 1. Let S be correspondingly defined by* (19)*. Let $A \in \mathbb{R}^{M\times N}$ and $B^{\mathsf{T}}B \preccurlyeq A^{\mathsf{T}}A$. Let $\lambda > 0$ and $0 < \gamma < 1$. Then*

$$\frac{1}{2}\|y - Ax\|_2^2 - \frac{\lambda^2}{\gamma\|B\|_1^2}\, S\Big(\frac{\gamma\|B\|_1}{\lambda}Bx\Big) \geqslant \frac{(1-\gamma)}{2}\,\Big\|\frac{1}{(1-\gamma)}\,y - Ax\Big\|_2^2 - \frac{\gamma}{2(1-\gamma)}\,\|y\|_2^2 \quad (69)$$

for all $x \in \mathbb{R}^N$.

Proof. From Proposition 3, $S(x) \leqslant \|x\|_2^2/2$ for all $x \in \mathbb{R}^N$. Hence,

$$\frac{\lambda^2}{\gamma\|B\|_1^2}\, S\Big(\frac{\gamma\|B\|_1}{\lambda}Bx\Big) \leqslant \frac{\gamma}{2}\|Bx\|_2^2 \quad (70)$$

$$\leqslant \frac{\gamma}{2}\|Ax\|_2^2 \quad (71)$$

for all $x \in \mathbb{R}^N$, where we have used $B^{\mathsf{T}}B \preccurlyeq A^{\mathsf{T}}A$. It follows that

$$\frac{1}{2}\|Ax\|_2^2 - \frac{\lambda^2}{\gamma\|B\|_1^2}\, S\Big(\frac{\gamma\|B\|_1}{\lambda}Bx\Big) \geqslant \frac{(1-\gamma)}{2}\|Ax\|_2^2. \quad (72)$$

A completion of the square leads straightforwardly to inequality (69). □

Lemma 6 leads straightforward to the following corollary.

Corollary 1. *In the setting of Lemma 6, let $f\colon \mathbb{R}^N \to \mathbb{R}$ be defined*

$$f(x) = \frac{1}{2}\|y - Ax\|_2^2 - \frac{\lambda^2}{\gamma\|B\|_1^2}\, S\Big(\frac{\gamma\|B\|_1}{\lambda}Bx\Big). \quad (73)$$

Then f is bounded below, i.e., $f(x) > c$ for all $x \in \mathbb{R}^N$ for some $c \in \mathbb{R}$ that does not depend on x.

Following [5], we will use the following result which is the equivalence (*i*) $\Leftrightarrow$ (*vi*) of Theorem 18.15 in [3].

Lemma 7. *Let $f\colon \mathbb{R}^N \to \mathbb{R}$ be convex and differentiable. Then the gradient ∇f is ρ-Lipschitz continuous if and only if $(\rho/2)\|\cdot\|_2^2 - f$ is convex.*

Lemma 8. *In the setting of Lemma 6, let $f\colon \mathbb{R}^N \to \mathbb{R}$ be defined*

$$f(x) = \frac{1}{2}\|y - Ax\|_2^2 - \frac{\lambda^2}{\gamma\|B\|_1^2}\, S\Big(\frac{\gamma\|B\|_1}{\lambda}Bx\Big). \quad (74)$$

Then the gradient ∇f is ρ-Lipschitz continuous where $\rho = \|A\|_2^2$. (That is, ρ is the maximum eigenvalue of $A^{\mathsf{T}}A$.)

Proof. The proof will use Lemma 7. Since both terms in (74) are differentiable, f is differentiable. Proceeding as in the proof of Lemma 4, it follows that f is convex. Note also that $q\colon \mathbb{R}^2 \to \mathbb{R}$ defined as

$$q(x) = \frac{\rho}{2}\|x\|_2^2 - \frac{1}{2}\|y - Ax\|_2^2 \quad (75)$$

is convex. We now show $(\rho/2)\|\cdot\|_2^2 - f$ is convex. Define $g\colon \mathbb{R}^N \to \mathbb{R}$ as

$$g(x) = \frac{\rho}{2}\|x\|_2^2 - f(x) \quad (76)$$

$$= \frac{\rho}{2}\|x\|_2^2 - \frac{1}{2}\|y - Ax\|_2^2 + \frac{\lambda^2}{\gamma\|B\|_1^2}\, S\Big(\frac{\gamma\|B\|_1}{\lambda}Bx\Big)$$

$$= q(x) + \frac{\lambda^2}{\gamma\|B\|_1^2}\, S\Big(\frac{\gamma\|B\|_1}{\lambda}Bx\Big). \quad (77)$$

Then g is is convex because both terms in (77) are convex. By Lemma 7, it follows that ∇f is ρ-Lipschitz continuous. □

V. Iterative Thresholding Algorithm

The forward-backward splitting (FBS) algorithm [24], [25] can be used to obtain a minimizer of the objective function F in (62). The resulting iterative thresholding algorithm uses the soft-threshold function, which is defined as

$$\operatorname{soft}(t, T) := \begin{cases} t - T, & t \geqslant T \\ 0, & |t| \leqslant T \\ t + T, & t \leqslant -T. \end{cases} \quad (78)$$

Lemma 9. *In the setting of Theorem 2, with $0 < \gamma < 1$, let $\rho = \|A\|_2^2$ and $0 < \mu < 2/\rho$. Then the sequence $x^{(k)}$, $k \in \mathbb{N}$, generated by the iteration,*

$$z^{(k)} = x^{(k)} - \mu\Big[A^{\mathsf{T}}(Ax^{(k)} - y) - \frac{\lambda}{\|B\|_1}B^{\mathsf{T}}\nabla S\Big(\frac{\gamma\|B\|_1}{\lambda}Bx^{(k)}\Big)\Big] \quad (79a)$$

$$x^{(k+1)} = \operatorname{soft}(z^{(k)}, \mu\lambda) \quad (79b)$$

converges to a minimizer of the objective function F defined by (62)*. The soft thresholding function is applied element-wise to vector $z^{(k)}$.*

Proof. As described in Proposition 1.3.4 in [24], the FBS algorithm can be used to minimize a function of the form

$$F(x) = f_1(x) + f_2(x) \tag{80}$$

where f_1 is convex and differentiable with ρ-Lipschitz continuous gradient ∇f_1, f_2 is lower semicontinuous convex, and $F(x) \to \infty$ as $\|x\| \to \infty$. To apply FBS to F in (62), we set

$$f_1(x) = \frac{1}{2}\|y - Ax\|_2^2 - \frac{\lambda^2}{\gamma\|B\|_1^2}\, S\Big(\frac{\gamma\|B\|_1}{\lambda} Bx\Big) \tag{81}$$

$$f_2(x) = \lambda\|x\|_1. \tag{82}$$

By Lemma 8, ∇f_1 is ρ-Lipschitz continuous. Since f_2 is a norm and any norm is continuous and convex, f_2 is a lower semicontinuous convex function. By Corollary 1, f_1 is bounded below. Hence, $F(x) \to \infty$ as $\|x\| \to \infty$. Hence the hypothesis of Proposition 1.3.4 in [24] is satisfied, i.e., the iterates $x^{(k)}$ produced by the FBS algorithm converge to a minimizer of F.

A basic form of FBS comprises the iteration

$$z^{(k)} = x^{(k)} - \mu\big[\nabla f_1(x^{(k)})\big] \tag{83a}$$

$$x^{(k+1)} = \min_x \Big\{\frac{1}{2}\|z^{(k)} - x\|_2^2 + \mu f_2(x)\Big\} \tag{83b}$$

where $0 < \mu < 2/\rho$, which leads to (79). □

The parameter μ in (79) can be interpreted as a step size. We usually implement the algorithm with $\mu = 1.9/\rho$ (near the upper allowed value) because larger step sizes often yield faster convergence in practice. The algorithm has the property that $F(x^{(k)})$ monotonically decreases [6], [74].

This algorithm (79) is like the classical iterative shrinkage/thresholding algorithm (ISTA) [26], [30] which can be viewed as a special case of the FBS algorithm. Other algorithms are also applicable [24], including accelerated versions of ISTA such as fast ISTA (FISTA) [8] and FASTA [36]. Additionally, new extensions and generalizations of FBS further extends its applicability [22].

A. Optimality condition

Since F is convex, the optimality of a prospective minimizer of F can be validated using the optimality condition $0 \in \partial F(x^{\mathsf{opt}})$ where ∂F is the subdifferential of F. For (80), a vector x^{opt} is optimal if

$$-\nabla f_1(x^{\mathsf{opt}}) \in \partial f_2(x^{\mathsf{opt}}) \tag{84}$$

where ∂f_2 is the subdifferential of f_2. For f_2 in (82), this leads to

$$-\Big[\frac{1}{\lambda}\nabla f_1(x^{\mathsf{opt}})\Big]_n \in \mathrm{SGN}(x_n^{\mathsf{opt}}) \tag{85}$$

for $n = 1, \dots, N$ where SGN is the set-valued signum function,

$$\mathrm{SGN}(t) := \begin{cases} \{1\}, & t > 0 \\ [-1, 1], & t = 0 \\ \{-1\}, & t < 0. \end{cases} \tag{86}$$

B. Complex-valued case

In many problems, the matrix A in problem (62) is complex-valued and the minimization of F is performed over $\mathbb{C}^N$. For example, A and x may comprise Fourier transforms and Fourier coefficients, respectively. In this case, the real and imaginary parts of x can be embedded into a real optimization problem of greater size. The forward-backward splitting algorithm leads again to the iteration (79) except the transpose is replaced by the complex-conjugate transpose and the soft-threshold rule (78) is replaced by its generalization,

$$\mathrm{soft}(u, T) := \begin{cases} 0, & |u| \leqslant T \\ (|u| - T)\, u/|u|, & |u| > T \end{cases} \tag{87}$$

for $u \in \mathbb{C}$ and $T \geqslant 0$.

VI. Numerical Examples

A. Denoising by sparse signal approximation (SSA)

In this section, we apply multivariate sparse regularization (MUSR) as described in Theorem 2 to denoising using sparse signal approximation (SSA), i.e., by obtaining a sparse approximate solution to the linear system $y = Ax$ for the purpose of estimating a signal s in zero-mean noise wherein the signal admits a sparse representation with respect to the columns of matrix A. In particular, we consider the case where the columns of A form a tight frame for $\mathbb{R}^N$, i.e., A satisfies

$$AA^{\mathsf{T}} = pI \tag{88}$$

for some $p > 0$ [44]. The matrix A may be square or *wide*. Numerous transforms designed for the sparse representation of signals can be implemented as tight frames, including Fourier transforms, short-time Fourier transforms, filter banks, and multiscale transforms [42]. In these cases, A is a wide matrix satisfying (88) or $AA^{\mathsf{H}} = pI$ in the complex case. Note that in this context, A represents the 'inverse' transform (i.e., mapping transform coefficients to the signal domain), and A^{T} represents the 'forward' transform.

If $y = v + w$ is an observation of v corrupted by additive Gaussian noise, then an estimates $\hat{v}$ is given by $\hat{v} = Ax^{\mathsf{opt}}$ where x^{opt} is the solution to the SSA denoising problem, i.e., the minimizer of F.

Given a sparsifying transform A and univariate penalty ϕ, the use of MUSR for SSA means the minimization of F in (62). We must specify the matrix B and the parameters λ and γ. First, we set the matrix B such that $B^{\mathsf{T}}B \preccurlyeq A^{\mathsf{T}}A$. We suggest setting

$$B = \frac{1}{\sqrt{p}} A^{\mathsf{T}}A, \tag{89}$$

because then B satisfies

$$B^{\mathsf{T}}B = \Big(\frac{1}{\sqrt{p}}A^{\mathsf{T}}A\Big)^{\mathsf{T}}\Big(\frac{1}{\sqrt{p}}A^{\mathsf{T}}A\Big) \tag{90}$$

$$= \frac{1}{p}A^{\mathsf{T}}AA^{\mathsf{T}}A \tag{91}$$

$$= \frac{1}{p}A^{\mathsf{T}}(pI)A \tag{92}$$

$$= A^{\mathsf{T}}A. \tag{93}$$

The choice of B in (89) is motivated by the fact that if A is orthonormal, then this B is the identity matrix and the proposed regularizer ψ reduces to a *separable* non-convex regularizer, i.e., the natural sparsity-inducing regularizer. With B given by (89), the update equation (79a) can be written as

$$z^{(k)} = x^{(k)} - \mu A^{\mathsf{T}} \times \left(A \left[x^{(k)} - \frac{\lambda}{\sqrt{p}\,\|B\|_1} \nabla S\Big(\frac{\gamma \|B\|_1}{\lambda \sqrt{p}} A^{\mathsf{T}} A x^{(k)} \Big) \right] - y \right).$$

This expression of the update equation is computationally more efficient because B^{T} is adsorbed into $A^{\mathsf{T}}A$ which reduces the instances the transform A must be applied per iteration.

To perform denoising using SSA we must also select the positive regularization parameter λ. We remark first that a λ value that works well for the ℓ_1-norm form of SSA (i.e., BPD) serves as a reasonable value for the proposed MUSR-SSA problem (62). This is because the proposed regularizer ψ is designed to approximate the ℓ_1 norm around zero and to preserve the convexity of the objective function F. Hence, the primary effect of the MUSR-SSA formulation in relation to L1-SSA is to relax the penalization of large magnitude components of x. We remark further that a reasonable value of λ may be simply obtained by the 'three sigma' rule, i.e., a pure zero-mean noise signal lies mostly within three standard deviations of zero, hence setting signal values below three sigma to zero effectively attenuates the noise. The view of λ as a quasi-threshold value in the context of BPD follows from using the optimality condition (85) and by considering the output of BPD as applied to a pure white noise signal; see [32], [72]. Assuming each column of A has the same ℓ_2 norm η, we hence suggest setting $\lambda = \beta\eta\sigma$, where $2.5 \leqslant \beta \leqslant 3.0$, η is the common column norm, and σ is the standard deviation of the noise in the signal domain. Tight frame transforms A such as oversampled Fourier transforms, short-time Fourier transforms, and some filter banks have the property that each column of A has the same ℓ_2 norm.

To perform MUSR-SSA we must also set γ in (62). Values of γ close to 1.0 induce sparsity more strongly. As γ goes to zero, the minimizer x^{opt} goes to the ℓ_1 norm solution. We usually set γ between 0.5 and 0.9.

1) Example 1: We use SSA to estimate the signal

$$v_n = 2\cos(2\pi f_1 n) + \sin(2\pi f_2 n), \quad n = 0, \ldots, N-1 \quad (94)$$

of length $N = 100$ with frequencies $f_1 = 0.1$ and $f_2 = 0.22$. The noisy signal is $y_n = v_n + w_n$ where w is additive white Gaussian noise (AWGN) with $\sigma = 1.0$. The matrix A is an overcomplete discrete Fourier transform (DFT) matrix of size 100×256 normalized so that $AA^{\mathsf{H}} = I$. The operator A is implemented as a truncated inverse FFT; the operator A^{H} is implemented as a zero-padded FFT. Each vector comprising A has an ℓ_2 norm of $5/8$. According to the discussion above, we set $\lambda = 2.5 \times (5/8) \times \sigma = 1.5625$. For MUSR-SSA, we use the MC penalty and we set $B = A^{\mathsf{H}}A$ and $\gamma = 0.9$. Figure 8 illustrates the signal v, the noisy signal y, and the L1-SSA and MUSR-SSA solutions obtained using the forward-backward splitting (FBS) algorithm. MUSR-SSA reduces the root-mean-square error (RMSE) by more than 25% relative to L1-SSA.

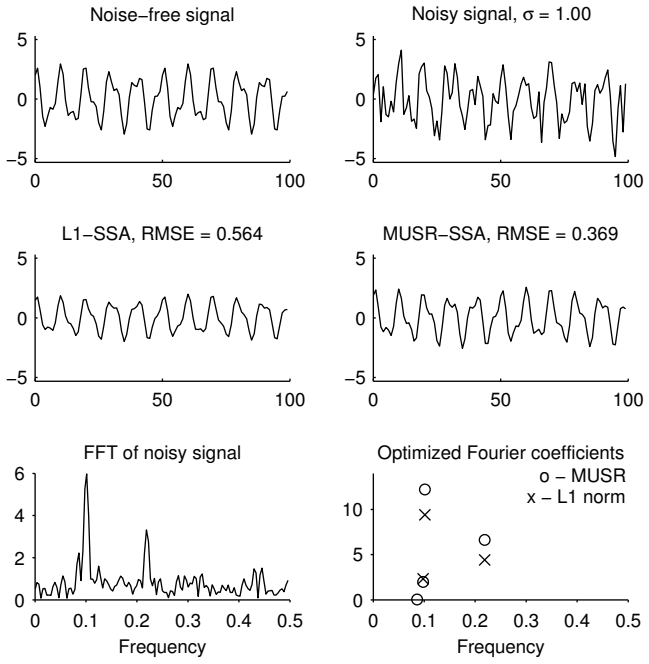

Fig. 8. Denoising via sparse signal approximation (SSA) using the ℓ_1-norm and MUSR (proposed). The plot of the optimized coefficients shows only the non-zero values.

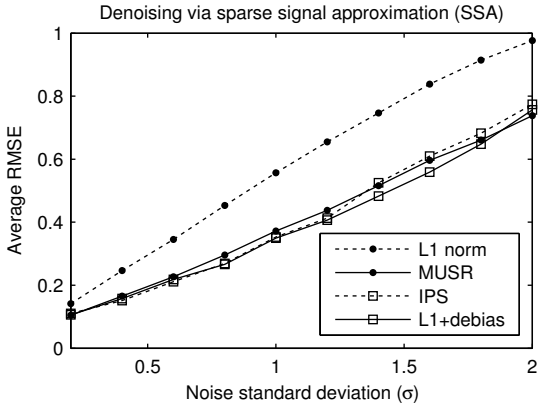

Fig. 9. Average RMSE corresponding to Fig. 8.

Figure 8 shows that the ℓ_1 norm optimized Fourier coefficients underestimate the true coefficients. The MUSR optimized coefficients estimate the true coefficients more accurately. This experiment is repeated for $0.2 \leqslant \sigma \leqslant 2.0$ (with 50 noise realization for each σ) and the average RMSE as a function of σ is shown in Fig. 9. In this experiment, MUSR-SSA reduces the average RMSE by more than 20% relative to L1-SSA.

We compare with the *iterative p-shrinkage* (IPS) algorithm [82], [87], an iterative thresholding algorithm designed to locally minimize a non-convex objective function. The IPS algorithm was found to be particularly effective in comparison to several other algorithms [71]. As shown in Fig. 9, the average RMSE of MUSR is similar to that of IPS.

We also compare with the two-step approach wherein the ℓ_1-norm solution is followed by a debiasing step [31]. First, the ℓ_1-norm solution is used to estimate the support (the indices

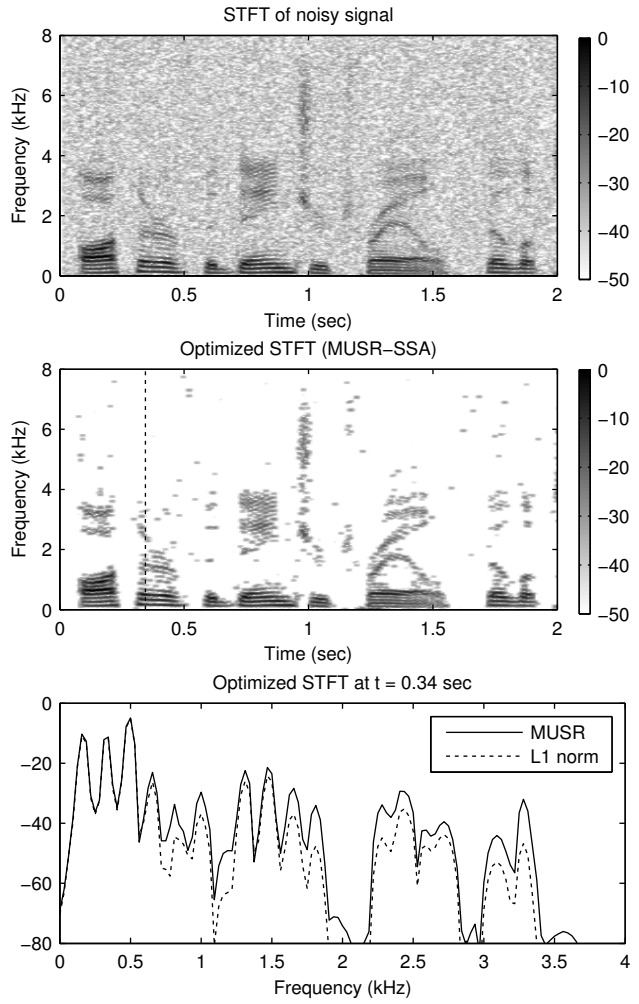

Fig. 10. Speech denoising via sparse signal approximation (SSA) using the ℓ_1-norm and MUSR (proposed).

of the non-zero values of x). Second, the identified non-zero values are re-estimated by unregularized least squares approximation. The debiasing post-processing step avoids the systematic underestimation of non-zero amplitudes, yet it is nevertheless influenced by noise in the observed data. As shown in Fig 9, this method yields average RMSE values the same or slightly better than the other considered methods. While ℓ_1-norm with debiasing is effective in this example, it is not wholly based on a variational principle (does not minimize a prescribed objective function) as the other considered methods do.

2) Example 2: This example uses SSA to estimate a speech signal in AWGN. The signal has a sampling rate of 16,000 samples/second and the noise standard deviation is $\sigma = 0.025$. For the sparsifying transform A we use a short-time Fourier transform (STFT) implemented so that $AA^{\mathsf{H}} = I$. We use an STFT window of 512 samples (32 msec) with 75% overlapping. Figure 10 illustrates the spectrogram (STFT magnitude) in dB of the noisy speech signal. Each vector comprising A has an ℓ_2 norm of 0.5, hence we set $\lambda = 3 \times 0.5 \times \sigma = 0.0375$ as discussed above. For MUSR-SSA, we use the MC penalty and

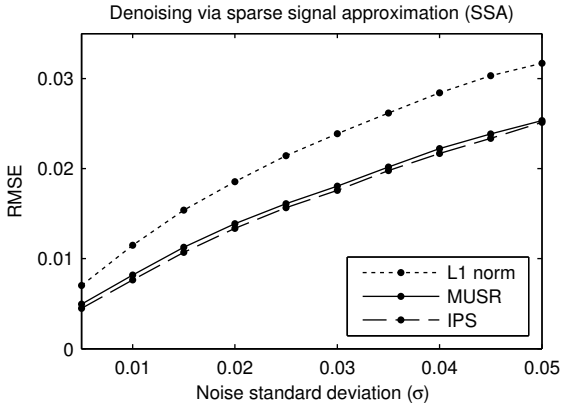

Fig. 11. RMSE corresponding to Fig. 10.

we set $B = A^{\mathsf{H}}A$ and $\gamma = 0.9$ as in Example 1. The denoised signals using L1-SSA and MUSR-SSA are calculated using the FBS algorithm. The spectrogram of the denoised signal shows the noise is well suppressed. A slice of the STFT at time $t = 0.34$ seconds (i.e., the Fourier transform of a 32 msec segment) shows the ℓ_1 norm solution tends to underestimate the true spectrum in comparison to the MUSR solution at frequencies around 1 kHz and above. This experiment is repeated for $0.005 \leqslant \sigma \leqslant 0.05$ and the RMSE as a function of σ is shown in Fig. 11 for the three methods: L1-SSA, MUSR-SSA, and IPS. As shown, MUSR-SSA consistently reduces the RMSE by about 20% relative to L1-SSA. The IPS algorithm performs similarly (slightly better).

Improved speech denoising can be achieved using overlapping group sparsity (OGS) which better models the behavior of a speech spectrogram [19]. In future work, we hope to generalize the MUSR approach to utilize OGS.

B. Simultaneous denoising and missing data estimation

Here we consider the problem of estimating a signal from partial, noisy data [1], [75], [85]. We assume the unknown signal v can be well-approximated as a linear combination of relatively few columns of a known matrix A. We denote by P the operator that selects partial data. We model the observed data g as

$$g = PAx + w \tag{95}$$

where x is the vector of sparse coefficients and w is AWGN. Specifically, the matrix P is obtained by deleting rows from the identity matrix, where the deleted rows correspond to the indices of the missing data. Note that $PP^{\mathsf{T}} = I$. We assume that $AA^{\mathsf{H}} = pI$ for some $p > 0$ as in Sec. VI-A.

The problem of estimating the unknown signal v can be expressed as

$$x^{\mathsf{opt}} = \arg\min_x \left\{ \frac{1}{2}\|g - PAx\|_2^2 + \lambda\psi(x) \right\} \tag{96}$$

where ψ is a sparsity-inducing penalty. The estimated signal is then given by $\hat{v} = Ax^{\mathsf{opt}}$. We define $A_2 = PA$. Hence, $A_2 A_2^{\mathsf{H}} = pI$, i.e., the columns of A_2 form a tight frame. Therefore, (96) can be solved as in Section VI-A.

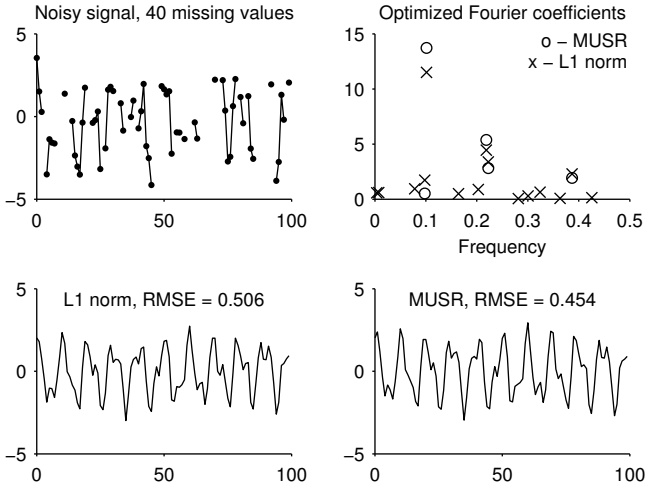

Fig. 12. Estimation of missing data and denoising using the ℓ_1-norm and MUSR (proposed). The plot of the optimized coefficients shows only the non-zero values.

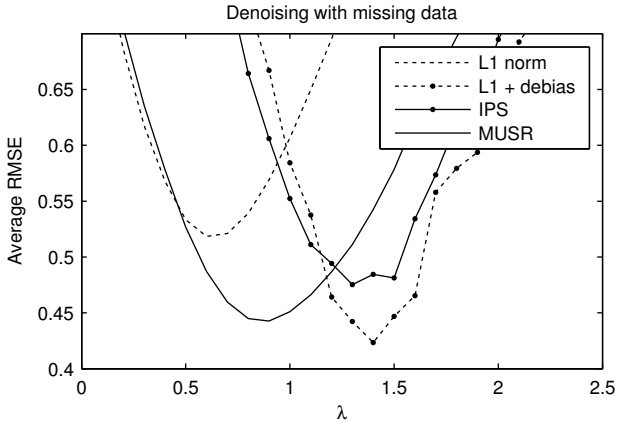

Fig. 13. Average RMSE corresponding to Fig. 12.

In this example, we use the signal (94) again, but with 40 randomly located values missing. We use AWGN with $\sigma = 1$. A realization is illustrated in Fig. 12. The goal is to estimate the signal from the noisy partial data. We set A as in Example 1; hence, x represents Fourier coefficients to be optimized.

As above, we use four methods: ℓ_1 norm and MUSR sparse regularization, debiasing of the ℓ_1 norm solution, and the IPS algorithm [87]. Each method calls for a regularization parameter λ to be set. We vary λ from 0.1 to 2.5 and evaluate the RMSE for each method. For MUSR, we set the parameter $\gamma = 0.8$ and use the MC penalty. Furthermore, we repeat this for 20 realizations of the noise. The average RMSE as a function of λ is shown in Fig. 13. The solution obtained by debiasing the ℓ_1 norm solution achieves the minimum average RMSE. The MUSR solution reduces the average RMSE by about 10% compared to ℓ_1 norm solution.

Figure 12 shows a particular realization, and the ℓ_1 norm and MUSR solutions, where in each case, the value of λ was taken to be the value that minimizes the average RMSE for the respective methods.

C. Sparse Deconvolution

This example illustrates MUSR as applied to the sparse deconvolution problem where the unknown sparse signal x is to be determined from data $y = Ax + w$ where A represents convolution and w is AWGN. In contrast to the previous examples, the columns of A do not comprise a tight frame, i.e., $AA^{\mathsf{T}} \neq pI$. Therefore, the choice of matrix B according to (89) is no longer justified. The proposed MUSR approach is still applicable; however, the matrix B appearing in the penalty ψ in (63) must be specified otherwise. In this example, we set $B = A$ which trivially satisfies the condition $BB^{\mathsf{T}} \preccurlyeq AA^{\mathsf{T}}$ required by Theorem 2.

We generate sparse signals of length $N = 200$ with 10 non-zero values (uniformly distributed in value between 0 and 100, located at random positions). Figure 14 illustrates one realization. We set the convolution operator A to be a 10-point moving average filter, i.e., $y = h * x + w$ where $h(n) = 0.1$ for $n = 0, \ldots, 9$ and $h(n) = 0$ otherwise. We set the AWGN standard deviation to $\sigma = 2$. The observed signal y is shown in Fig. 14. To perform sparse deconvolution using ℓ_1 norm regularization and MUSR, we set λ straightforwardly as $\lambda = \beta\sigma\|h\|_2$ with $\beta = 2.5$. For MUSR we set $\gamma = 0.6$ and use the MC penalty. The sparse signal estimated using MUSR is illustrated in Fig. 14. With 200 realizations, the average RMSE values are 4.87 and 4.32 for ℓ_1 norm regularization and MUSR, respectively. MUSR reduces the average RMSE by about 10% relative to ℓ_1 norm regularization. For each realization, the RMSE of the MUSR solution versus the RMSE of the ℓ_1 norm solution is designated as a single point in the scatter plot in Fig. 15. Points below the diagonal line represent realizations where MUSR improves upon ℓ_1 norm regularization. Even though the average RMSE of MUSR is less (better) than ℓ_1 norm regularization, for a few realizations the RMSE of MUSR is worse.

For comparison, we also perform sparse deconvolution using the IPS algorithm for each realization, with λ chosen to minimize the average RMSE. We initialize the IPS algorithm with the ℓ_1 norm solution. The IPS average RMSE is 3.89, about 20% better than ℓ_1 norm regularization. For each realization, the RMSE of the IPS solution versus the RMSE of the ℓ_1 norm solution is illustrated in Fig. 15. The scatter plot shows that IPS reduces the RMSE relative to both MUSR and ℓ_1 norm regularization on average. But as above, for a few realizations IPS performs worse than ℓ_1 norm regularization. The scatter plots shows the IPS RMSE values are substantially more spread than the MUSR RMSE values. Compared to IPS, MUSR more often performs better than the ℓ_1 norm, but on average IPS provides twice the improvement of MUSR relative to ℓ_1 norm regularization.

This example suggest that, while the MUSR approach can yield an improvement relative to ℓ_1 norm regularization for general A, the improvement is more significant when the columns of A form a tight frame.

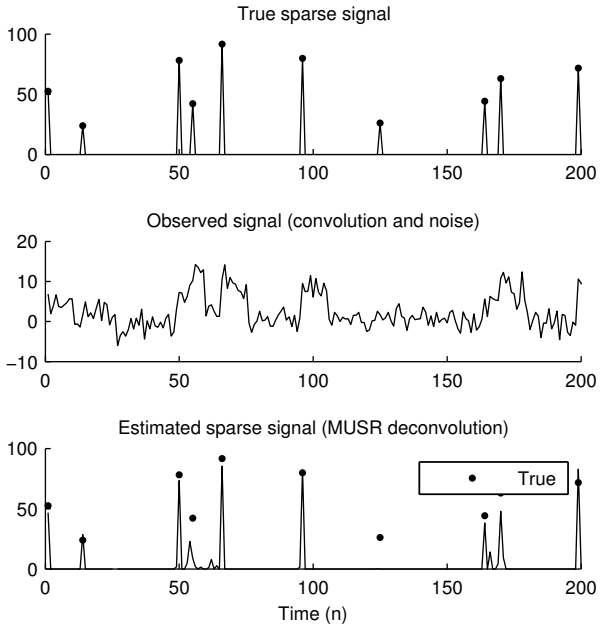

Fig. 14. Sparse deconvolution example.

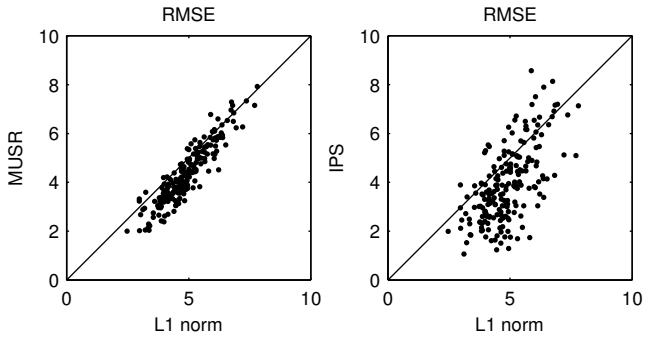

Fig. 15. RMSE values of 200 realizations of sparse deconvolution.

VII. Conclusion

This paper has considered the question of how to formulate, as a convex optimization problem, the calculation of a sparse approximate solution to a system of linear equations. To this end, we proposed a class of non-convex (specifically, weakly convex) penalty functions obtained by subtracting a smooth convex function from the ℓ_1 norm. The proposed approach compares favorably to ℓ_1 norm regularization, especially for sparse signal approximation using tight-frames.

The functional form of the proposed penalty (44) [comprising the difference of the ℓ_1 norm and a separable function composed with a linear operator] is a rather constrained class of non-separable functions. To prescribe non-convex penalties that preserve objective function convexity, there may be other classes of non-separable functions (yet to be determined) that are even more effective. For example, for total variation denoising, we have recently found that a quite different kind of non-separable penalty is particularly effective [69]. Hence, further research on this topic will be of interest.

Acknowledgment

The authors gratefully acknowledge suggestions and corrections from the anonymous reviewers.

References

[1] A. Adler, V. Emiya, M. G. Jafari, M. Elad, R. Gribonval, and M. D. Plumbley. Audio inpainting. *IEEE Trans. on Audio, Speech, and Lang. Proc.*, 20(3):922–932, March 2012.

[2] A. Antoniadis and J. Fan. Regularization of wavelet approximations. *J. Amer. Statist. Assoc.*, 96(455):939–955, 2001.

[3] H. H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, 2011.

[4] İ. Bayram. Penalty functions derived from monotone mappings. *IEEE Signal Processing Letters*, 22(3):265–269, March 2015.

[5] I. Bayram. Correction for On the convergence of the iterative shrinkage/thresholding algorithm with a weakly convex penalty. *IEEE Trans. Signal Process.*, 64(14):3822–3822, July 2016.

[6] I. Bayram. On the convergence of the iterative shrinkage/thresholding algorithm with a weakly convex penalty. *IEEE Trans. Signal Process.*, 64(6):1597–1608, March 2016.

[7] I. Bayram, P.-Y. Chen, and I. Selesnick. Fused lasso with a non-convex sparsity inducing penalty. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, May 2014.

[8] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imag. Sci.*, 2(1):183–202, 2009.

[9] A. Blake and A. Zisserman. *Visual Reconstruction*. MIT Press, 1987.

[10] T. Blumensath and M. E. Davies. Iterative hard thresholding for compressed sensing. *J. of Appl. and Comp. Harm. Analysis*, 27(3):265–274, 2009.

[11] A. Bruckstein, D. Donoho, and M. Elad. From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM Review*, 51(1):34–81, 2009.

[12] E. J. Candès, M. B. Wakin, and S. Boyd. Enhancing sparsity by reweighted l1 minimization. *J. Fourier Anal. Appl.*, 14(5):877–905, December 2008.

[13] M. Carlsson. On convexification/optimization of functionals including an l2-misfit term. https://arxiv.org/abs/1609.09378, September 2016.

[14] M. Castella and J.-C. Pesquet. Optimization of a Geman-McClure like criterion for sparse signal deconvolution. In *IEEE Int. Workshop Comput. Adv. Multi-Sensor Adaptive Proc. (CAMSAP)*, pages 309–312, December 2015.

[15] P. Charbonnier, L. Blanc-Feraud, G. Aubert, and M. Barlaud. Deterministic edge-preserving regularization in computed imaging. *IEEE Trans. Image Process.*, 6(2):298–311, February 1997.

[16] R. Chartrand. Nonconvex splitting for regularized low-rank + sparse decomposition. *IEEE Trans. Signal Process.*, 60(11):5810–5819, November 2012.

[17] R. Chartrand. Shrinkage mappings and their induced penalty functions. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, pages 1026–1029, May 2014.

[18] L. Chen and Y. Gu. The convergence guarantees of a non-convex approach for sparse recovery. *IEEE Trans. Signal Process.*, 62(15):3754–3767, August 2014.

[19] P.-Y. Chen and I. W. Selesnick. Group-sparse signal denoising: Non-convex regularization, convex optimization. *IEEE Trans. Signal Process.*, 62(13):3464–3478, July 2014.

[20] S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, 20(1):33–61, 1998.

[21] E. Chouzenoux, A. Jezierska, J. Pesquet, and H. Talbot. A majorize-minimize subspace approach for $\ell_2 - \ell_0$ image regularization. *SIAM J. Imag. Sci.*, 6(1):563–591, 2013.

[22] E. Chouzenoux, J.-C. Pesquet, and A. Repetti. Variable metric forward–backward algorithm for minimizing the sum of a differentiable function and a convex function. *J. Optim. Theory Appl.*, 162(1):107–132, 2014.

[23] P. L. Combettes. Perspective functions: Properties, constructions, and examples. http://arxiv.org/abs/1610.01552, October 2016.

[24] P. L. Combettes and J.-C. Pesquet. Proximal splitting methods in signal processing. In H. H. Bauschke et al., editors, *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, pages 185–212. Springer-Verlag, 2011.

[25] P. L. Combettes and V. R. Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Modeling & Simulation*, 4(4):1168–1200, 2005.

[26] I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Commun. Pure Appl. Math*, 57(11):1413–1457, 2004.

[27] Y. Ding and I. W. Selesnick. Artifact-free wavelet denoising: Non-convex sparse regularization, convex optimization. *IEEE Signal Processing Letters*, 22(9):1364–1368, September 2015.

[28] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, 96(456):1348–1360, 2001.

[29] M. Figueiredo and R. Nowak. Wavelet-based image estimation: An empirical Bayes approach using Jeffrey's noninformative prior. *IEEE Trans. Image Process.*, 10(9):1322–1331, September 2001.

[30] M. Figueiredo and R. Nowak. An EM algorithm for wavelet-based image restoration. *IEEE Trans. Image Process.*, 12(8):906–916, August 2003.

[31] M. A. T. Figueiredo, R. D. Nowak, and S. J. Wright. Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. *IEEE J. Sel. Top. Signal Process.*, 1(4):586–598, December 2007.

[32] J.-J. Fuchs. Convergence of a sparse representations algorithm applicable to real or complex data. *IEEE. J. Sel. Top. Signal Processing*, 1(4):598–605, December 2007.

[33] G. Gasso, A. Rakotomamonjy, and S. Canu. Recovering sparse signals with a certain family of nonconvex penalties and DC programming. *IEEE Trans. Signal Process.*, 57(12):4686–4698, December 2009.

[34] D. Geman and G. Reynolds. Constrained restoration and the recovery of discontinuities. *IEEE Trans. Pattern Anal. and Machine Intel.*, 14(3):367–383, March 1992.

[35] A. Gholami and S. M. Hosseini. A general framework for sparsity-based denoising and inversion. *IEEE Trans. Signal Process.*, 59(11):5202–5211, November 2011.

[36] T. Goldstein, C. Studer, and R. Baraniuk. A field guide to forward-backward splitting with a FASTA implementation. http://arxiv.org/abs/1411.3406, 2014.

[37] R. Gribonval. Should penalized least squares regression be interpreted as maximum a posteriori estimation? *IEEE Trans. Signal Process.*, 59(5):2405–2410, 2011.

[38] R. Gribonval and P. Machart. Reconciling 'priors' & 'priors' without prejudice? In *Adv. in Neural Info. Process. Syst. (NIPS)*, pages 2193–2201, 2013.

[39] G. Harikumar and Y. Bresler. A new algorithm for computing sparse solutions to linear inverse problems. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, volume 3, pages 1331–1334, May 1996.

[40] W. He, Y. Ding, Y. Zi, and I. W. Selesnick. Sparsity-based algorithm for detecting faults in rotating machines. *Mechanical Systems and Signal Processing*, 72-73:46–64, May 2016.

[41] P. J. Huber. Robust regression: asymptotics, conjectures and Monte Carlo. *The Annals of Statistics*, pages 799–821, 1973.

[42] L. Jacques, L. Duval, C. Chaux, and G. Peyré. A panorama on multiscale geometric representations, intertwining spatial, directional and frequency selectivity. *Signal Processing*, 91(12):2699–2730, December 2011.

[43] N. Kingsbury and T. Reeves. Redundant representation with complex wavelets: how to achieve sparsity. In *Proc. IEEE Int. Conf. Image Processing (ICIP)*, volume 1, pages 45–48, 2003.

[44] J. Kovačević and A. Chebira. *An Introduction to Frames*. Now Publishers, 2008.

[45] I. Kozlov and A. Petukhov. Sparse solutions of underdetermined linear systems. In W. Freeden et al., editor, *Handbook of Geomathematics*. Springer, 2010.

[46] A. Lanza, S. Morigi, and F. Sgallari. Convex image denoising via non-convex regularization. In J.-F. Aujol, M. Nikolova, and N. Papadakis, editors, *Scale Space and Variational Methods in Computer Vision*, volume 9087 of *Lecture Notes in Computer Science*, pages 666–677. Springer, 2015.

[47] A. Lanza, S. Morigi, and F. Sgallari. Convex image denoising via non-convex regularization with parameter selection. *J. Math. Imaging and Vision*, 56(2):195–220, 2016.

[48] D. A. Lorenz. Non-convex variational denoising of images: Interpolation between hard and soft wavelet shrinkage. *Current Development in Theory and Application of Wavelets*, 1(1):31–56, 2007.

[49] M. Malek-Mohammadi, C. R. Rojas, and B. Wahlberg. A class of nonconvex penalties preserving overall convexity in optimization-based mean filtering. *IEEE Trans. Signal Process.*, 64(24):6650–6664, December 2016.

[50] D. Malioutov and A. Aravkin. Iterative log thresholding. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, pages 7198–7202, May 2014.

[51] S. Mallat. *A wavelet tour of signal processing*. Academic Press, 1998.

[52] Y. Marnissi, A. Benazza-Benyahia, E. Chouzenoux, and J.-C. Pesquet. Generalized multivariate exponential power prior for wavelet-based multichannel image restoration. In *Proc. IEEE Int. Conf. Image Processing (ICIP)*, pages 2402–2406, September 2013.

[53] R. Mazumder, J. H. Friedman, and T. Hastie. SparseNet: Coordinate descent with non-convex penalties. *J. Amer. Statist. Assoc.*, 106(495):1125–1138, 2011.

[54] H. Mohimani, M. Babaie-Zadeh, and C. Jutten. A fast approach for overcomplete sparse decomposition based on smoothed l0 norm. *IEEE Trans. Signal Process.*, 57(1):289–301, January 2009.

[55] L. B. Montefusco, D. Lazzaro, and S. Papi. A fast algorithm for nonconvex approaches to sparse recovery problems. *Signal Processing*, 93(9):2636–2647, 2013.

[56] M. Nikolova. Estimation of binary images by minimizing convex criteria. In *Proc. IEEE Int. Conf. Image Processing (ICIP)*, pages 108–112 vol. 2, 1998.

[57] M. Nikolova. Markovian reconstruction using a GNC approach. *IEEE Trans. Image Process.*, 8(9):1204–1220, 1999.

[58] M. Nikolova. Local strong homogeneity of a regularized estimator. *SIAM J. Appl. Math.*, 61(2):633–658, 2000.

[59] M. Nikolova. Energy minimization methods. In O. Scherzer, editor, *Handbook of Mathematical Methods in Imaging*, chapter 5, pages 138–186. Springer, 2011.

[60] M. Nikolova, J. Idier, and A. Mohammad-Djafari. Inversion of large-support ill-posed linear operators using a piecewise Gaussian MRF. *IEEE Trans. Image Process.*, 7(4):571–585, 1998.

[61] M. Nikolova, M. K. Ng, and C.-P. Tam. Fast nonconvex nonsmooth minimization methods for image restoration and reconstruction. *IEEE Trans. Image Process.*, 19(12):3073–3088, December 2010.

[62] A. Parekh and I. W. Selesnick. Convex denoising using non-convex tight frame regularization. *IEEE Signal Processing Letters*, 22(10):1786–1790, October 2015.

[63] A. Parekh and I. W. Selesnick. Convex fused lasso denoising with non-convex regularization and its use for pulse detection. In *IEEE Signal Processing in Medicine & Biology Symp. (SPMB)*, December 2015.

[64] A. Parekh and I. W. Selesnick. Enhanced low-rank matrix approximation. *IEEE Signal Processing Letters*, 23(4):493–497, April 2016.

[65] N. Parikh and S. Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):123–231, 2014.

[66] J. Portilla and L. Mancera. L0-based sparse approximation: two alternative methods and some applications. In *Proceedings of SPIE*, volume 6701 (Wavelets XII), San Diego, CA, USA, 2007.

[67] M. Raphan and E. P. Simoncelli. Learning to be Bayesian without supervision. In *Adv. in Neural Info. Process. Syst. (NIPS)*, pages 1145–1152, 2006.

[68] A. Repetti, M. Q. Pham, L. Duval, E. Chouzenoux, and J.-C. Pesquet. Euclid in a taxicab: Sparse blind deconvolution with smoothed l1/l2 regularization. *IEEE Signal Processing Letters*, 22(5):539–543, May 2015.

[69] I. Selesnick. Total variation denoising via the Moreau envelope. *IEEE Signal Processing Letters*, 24(2):216–220, February 2017.

[70] I. W. Selesnick and I. Bayram. Sparse signal estimation by maximally sparse convex optimization. *IEEE Trans. Signal Process.*, 62(5):1078–1092, March 2014.

[71] I. W. Selesnick and I. Bayram. Enhanced sparsity by non-separable regularization. *IEEE Trans. Signal Process.*, 64(9):2298–2313, May 2016.

[72] I. W. Selesnick, H. L. Graber, Y. Ding, T. Zhang, and R. L. Barbour. Transient artifact reduction algorithm (TARA) based on sparse optimization. *IEEE Trans. Signal Process.*, 62(24):6596–6611, December 2014.

[73] I. W. Selesnick, A. Parekh, and I. Bayram. Convex 1-D total variation denoising with non-convex regularization. *IEEE Signal Processing Letters*, 22(2):141–144, February 2015.

[74] Y. She. Thresholding-based iterative selection procedures for model selection and shrinkage. *Electronic Journal of Statistics*, 3:384–415, 2009.

[75] L. Shen, Y. Xu, and N. Zhang. An approximate sparsity model for inpainting. *J. of Appl. and Comp. Harm. Analysis*, 37(1):171–184, 2014.

[76] E. Soubies, L. Blanc-Féraud, and G. Aubert. A continuous exact ℓ_0 penalty (CEL0) for least squares regularized problem. *SIAM J. Imag. Sci.*, 8(3):1607–1639, 2015.

[77] C. Soussen, J. Idier, D. Brie, and J. Duan. From Bernoulli-Gaussian deconvolution to sparse signal restoration. *IEEE Trans. Signal Process.*, 59(10):4572–4584, October 2011.

[78] C. Soussen, J. Idier, J. Duan, and D. Brie. Homotopy based algorithms for ℓ_0-regularized least-squares. *IEEE Trans. Signal Process.*, 63(13):3301–3316, July 2015.

[79] J.-L. Starck, F. Murtagh, and J. Fadili. *Sparse image and signal processing: Wavelets and related geometric multiscale analysis*. Cambridge University Press, 2015.

[80] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc., Ser. B*, 58(1):267–288, 1996.

[81] M. E. Tipping. Sparse Bayesian learning and the relevance vector machine. *J. Machine Learning Research*, 1:211–244, 2001.

[82] S. Voronin and R. Chartrand. A new generalized thresholding algorithm for inverse problems with sparsity constraints. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, pages 1636–1640, May 2013.

[83] S. Voronin and H. J. Woerdeman. A new iterative firm-thresholding algorithm for inverse problems with sparsity constraints. *J. of Appl. and Comp. Harm. Analysis*, 35(1):151–164, 2013.

[84] Y. Wang and W. Yin. Sparse signal reconstruction via iterative support detection. *SIAM J. Imag. Sci.*, 3(3):462–491, 2010.

[85] Y. Wang and Q. Zhu. Error control and concealment for video communication: A review. *Proc. IEEE*, 86(5):974–997, May 1998.

[86] D. P. Wipf, B. D. Rao, and S. Nagarajan. Latent variable Bayesian models for promoting sparsity. *IEEE Trans. Inform. Theory*, 57(9):6236–6255, September 2011.

[87] J. Woodworth and R. Chartrand. Compressed sensing recovery via nonconvex shrinkage penalties. *Inverse Problems*, 32(7):75004–75028, July 2016.

[88] P. Yin, Y. Lou, Q. He, and J. Xin. Minimization of ℓ_{1-2} for compressed sensing. *SIAM J. Scientific Computing*, 37(1):A536–A563, 2015.

[89] C.-H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, pages 894–942, 2010.

[90] H. Zou and R. Li. One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.*, 36(4):1509–1533, 2008.