Enhanced Sparsity by Non-Separable Regularization

Ivan Selesnick and İlker Bayram

***Abstract*—This paper develops a convex approach for sparse one-dimensional deconvolution that improves upon L1-norm regularization, the standard convex approach. We propose a sparsity-inducing non-separable non-convex bivariate penalty function for this purpose. It is designed to enable the convex formulation of ill-conditioned linear inverse problems with quadratic data fidelity terms. The new penalty overcomes limitations of separable regularization. We show how the penalty parameters should be set to ensure that the objective function is convex, and provide an explicit condition to verify the optimality of a prospective solution. We present an algorithm (an instance of forward-backward splitting) for sparse deconvolution using the new penalty.**

I. Introduction

Methods for sparse regularization can be broadly categorized as convex or non-convex. In the standard convex approach, the regularization terms (penalty functions) are convex; and the objective function, consisting of both data fidelity and regularization terms, is convex [4], [60]. The convex approach has several benefits: the objective function is free of extraneous local minima, and globally convergent optimization algorithms can be leveraged [11].

Despite the attractive properties of convex regularization, non-convex regularization often performs better [12], [13], [57]. Classical and recent examples are in edge preserving tomography [14], [39], [56], [58] and compressed sensing [15], [18], [77], respectively. In the non-convex approach, penalty functions are non-convex as they can be designed to induce sparsity more effectively than convex ones. Therefore, the convexity of the objective function is generally sacrificed. Consequently, non-convex regularization is hampered by complications: the objective function will generally possess many sub-optimal local minima in which optimization algorithms can become entrapped.

It turns out, non-convex penalties can be utilized without giving up the convexity of the objective function and corresponding benefits. This is achieved by carefully specifying the penalty in accordance with the data fidelity term, as described by Blake, Zimmerman, and Nikolova [9], [55], [56], [58]. In recent work, a class of sparsity-inducing non-convex penalties has been developed to formulate convex objective functions and applied to several signal estimation problems [6], [7], [20], [29], [46], [61], [62], [69], [70]. This approach maintains the benefits of the convex framework (absence of spurious local minima, etc.), yet estimates sparse signals more accurately than convex regularization (e.g., the ℓ_1 norm) due to the sparsity-inducing properties of non-convex regularization. However, this previous work considers only *separable* (additive) penalties, which have fundamental limitations.

In this paper, we introduce a parameterized sparsity-inducing non-separable non-convex bivariate penalty function. The penalty is designed to enable the convex formulation of ill-conditioned linear inverse problems with quadratic data fidelity terms. The new penalty overcomes limitations of separable non-convex regularization. We show how the penalty parameters should be set to ensure the objective function is convex. We also show how this bivariate penalty can be incorporated into linear inverse problems of N variables ($N > 2$), and we provide an explicit condition to verify the optimality of a prospective solution. We present an iterative algorithm (an instance of forward-backward splitting) for sparse signal reconstruction using the new penalty; and we demonstrate its effectiveness for one-dimensional sparse deconvolution.

A. Basic problem statement

We consider the problem of bivariate sparse regularization (BISR) with a quadratic data fidelity term:

$$\hat{x} = \arg\min_{x\in\mathbb{R}^2} \left\{ f(x) = \frac{1}{2}\|y - Hx\|_2^2 + \lambda\psi(x) \right\} \tag{1}$$

where $\lambda > 0$, H is a 2×2 matrix, and $\psi\colon \mathbb{R}^2 \to \mathbb{R}$ is a bivariate penalty. [In Sec. V, it will be shown how to extend this bivariate problem to an N-point linear inverse problem.] In this paper, we suppose $H^\mathsf{T} H$ is Toeplitz, as this naturally arises in deconvolution problems. Correspondingly, we write $H^\mathsf{T} H = K(\gamma)$ where $\gamma = (\gamma_1, \gamma_2) \in \mathbb{R}^2$ and

$$K(\gamma) := \frac{1}{2}\begin{bmatrix} \gamma_1+\gamma_2 & \gamma_1-\gamma_2 \\ \gamma_1-\gamma_2 & \gamma_1+\gamma_2 \end{bmatrix} = Q\,\Gamma\,Q^\mathsf{T} \tag{2}$$

where

$$Q = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, \quad \Gamma = \begin{bmatrix} \gamma_1 & 0 \\ 0 & \gamma_2 \end{bmatrix}. \tag{3}$$

This is an eigenvalue decomposition of $H^\mathsf{T} H$. The parameters γ_1 and γ_2 are the eigenvalues of the positive semidefinite matrix $H^\mathsf{T} H$; hence, they are nonnegative.

First, suppose ψ is a separable convex penalty, e.g., $\psi(x) = |x_1| + |x_2|$ corresponding to the ℓ_1 norm (Fig. 1(a)). Then the objective function f in (1) is convex, but it does not induce sparsity as effectively as non-convex penalties can. In particular, ℓ_1 norm regularization tends to underestimate the true signal values.

Second, suppose ψ is a separable non-convex penalty, i.e., $\psi(x) = \phi(x_1) + \phi(x_2)$, as illustrated in Fig. 1(b). Then the objective function f is convex only if ϕ is suitably chosen. In

I. Selesnick is with the Dept. of Electrical and Computer Engineering, Tandon School of Engineering, New York University, 6 Metrotech Center, Brooklyn, NY 11201. İ Bayram is with the Department of Electronics and Telecommunications Engineering, Istanbul Technical University, Istanbul, Turkey. Email: selesi@nyu.edu, ilker.bayram@itu.edu.tr

This research was supported by NSF under grant CCF-1525398.

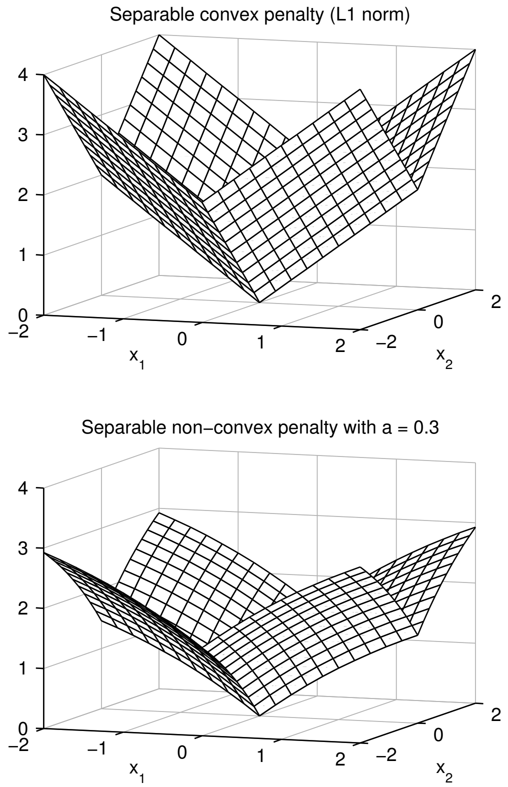

Fig. 1. Separable bivariate penalties, convex and non-convex.

particular, for the class of penalties we consider, the objective function is convex only if $\phi''(t) \geqslant -\min\{\gamma_1, \gamma_2\}/\lambda$ for all nonzero t where γ_1 and γ_2 are the eigenvalues of $H^\mathsf{T} H$. (See Lemma 3, [69], and [9].) When H is singular, the minimum eigenvalue of $H^\mathsf{T} H$ is zero and ϕ must be convex (i.e., ϕ induces sparsity relatively weakly). Hence, if H is singular and we restrict the penalty ψ to be separable, then we can not use sparsity-inducing non-convex regularization without sacrificing the convexity of the objective function f. Indeed, when separable non-convex penalties are utilized for their strong sparsity-inducing properties, the convexity of f is generally sacrificed.

Our aim is to prescribe a non-separable penalty ψ so that the objective function is guaranteed to be convex even though the penalty ψ itself is not. Such a penalty will be given in Sec. IV. It turns out, when we utilize a *non-separable* non-convex penalty to strongly induce sparsity, we need *not* sacrifice the convexity of the objective function f, even when H is *singular*.

B. Related work

The design of non-convex regularizers ensuring convexity of an objective function was proposed as part of the Graduated Non-Convexity (GNC) approach [9], [56], [58] and for binary image estimation [55]. Most methods for non-convex sparse regularization do not aim to maintain convexity of the objective function. The ℓ_p pseudo-norm ($0 \leqslant p < 1$) is widely used, but other regularizers have also been advocated [15], [17], [38]–[40], [48], [49], [51], [52], [78], [83]. Several algorithms have been developed specifically for the ℓ_0 pseudo-norm: matching pursuit [50], greedy ℓ_1 [44], iterative hard thresholding and its variations [10], [35], [43], [63], [65], [79], smoothed ℓ_0, [52], and single best replacement [73], [74].

Non-separable penalties are usually used to model statistical relationships among the significant values of sparse signals (possibly linearly transformed[1]) or to induce structured sparsity. For example, mixed norms [4], [51], [71] are often used for this purpose, rather than to induce 'pure' unstructured sparsity. But non-convex non-separable penalties have also been demonstrated to strongly promote unstructured sparsity [76], [81]. In particular, non-convex non-separable (*nonfactorial*) "Type II" penalties can approximate the ℓ_0 pseudo-norm more effectively than any separable penalty can, while producing many fewer local minima than other non-convex penalties (e.g., the ℓ_p pseudo-norm) [81].

The non-separable penalty developed here is different from sparsity-inducing penalties described in the above literature. Moreover, the new penalty is parameterized and utilized differently: namely, to induce sparsity subject to the constraint that the objective function is convex.

As the proposed method leads by design to a convex problem, convex sparse optimization methods may be used or adapted for its solution. Representative algorithms are the iterative shrinkage/thresholding algorithm (ISTA/FISTA) [8], [33], proximal methods [24], [25], alternating direction method of multipliers (ADMM) [2], [41], and majorization-minimization (MM) [32], [45].

Several algorithms are suitable for non-convex sparse regularization problems, such as iteratively reweighted least squares (IRLS) [28], [42], iteratively reweighted ℓ_1 (IRL1) [3], [13], [80], FOCUSS [66], related algorithms [53], [54], [75], non-convex MM [22], [40], [49], [78], extensions of GNC [52], [56], [58], and other methods [14], [15], [38] including new proximal algorithms [23]. Non-convex regularization has also been used for blind deconvolution [67] and low-rank plus sparse matrix decompositions [16]. Convergence of these algorithms are generally to local optima only. However, conditions for convergence to a global minimizer, or to guarantee that all local minimizers are near a global minimizer, have been recently reported [19], [47]. Whereas Refs. [19], [47] focuses on convergence guarantees for given penalties, here we focus on the design of penalties.

C. Notation

We write the vector $x \in \mathbb{R}^N$ as $x = (x_1,\ x_2,\ \ldots,\ x_N)$. Given $x \in \mathbb{R}^N$, we define $x_n = 0$ for $n \notin \{1, 2, \ldots, N\}$. (This simplifies expressions involving summations over n.) The ℓ_1 norm of $x \in \mathbb{R}^N$ is defined as $\|x\|_1 = \sum_n |x_n|$. If the matrix A is positive semidefinite, we write $A \succcurlyeq 0$. If the matrix $A - B$ is positive semidefinite, we write $A \succcurlyeq B$.

II. Univariate Penalties

The bivariate penalty to be given in Sec. IV will be based on a parameterized non-convex univariate penalty function

[1]Isotropic two-dimensional total variation, which induces joint sparsity of horizontal and vertical gradients, exemplifies this type of regularization.

TABLE I
UNIVARIATE PENALTIES

$\phi(t;a) = \dfrac{\lvert t\rvert}{1 + a\lvert t\rvert/2}, \quad a \geqslant 0$
$\phi(t;a) = \begin{cases} \frac{1}{a}\log(1 + a\lvert t\rvert), & a > 0 \\ \lvert t\rvert, & a = 0 \end{cases}$
$\phi(t;a) = \begin{cases} \frac{2}{a\sqrt{3}}\left(\tan^{-1}\left(\frac{1+2a\lvert t\rvert}{\sqrt{3}}\right) - \frac{\pi}{6}\right), & a > 0 \\ \lvert t\rvert, & a = 0 \end{cases}$

$\phi(\,\cdot\,;a)\colon \mathbb{R} \to \mathbb{R}$ with parameter $a \geqslant 0$. We shall assume ϕ has the following properties:

P1) $\phi(\,\cdot\,;a)$ is continuous on $\mathbb{R}$
P2) $\phi(\,\cdot\,;a)$ is twice continuously differentiable, increasing, and concave on $\mathbb{R}_+$
P3) $\phi(0;a) = 0$
P4) $\phi(t;0) = |t|$
P5) $\phi(-t;a) = \phi(t;a)$
P6) $\phi'(0^+;a) = 1$
P7) $\phi''(0^+;a) = -a$
P8) $\phi''(t;a) \geqslant -a$ for all $t \neq 0$
P9) $\phi(t;a)$ is decreasing and convex in a.
P10) $\phi(t;a) = (b/a)\,\phi\big(at/b;b\big)$ for $a,b > 0$ [Scaling].

It follows from symmetry that $\phi'(-t) = -\phi'(t)$ and $\phi''(-t) = \phi''(t)$. The scaling property of ϕ also induces a scaling property of ϕ' and ϕ''. Namely,

$$\phi'(t;a) = \phi'(at/b;b), \tag{4}$$
$$\phi''(t;a) = (a/b)\,\phi''(at/b;b). \tag{5}$$

Table I lists several penalty functions that satisfy the above properties: the rational [39], logarithmic, and arctangent functions [13], [57], [69] (when suitably normalized). The arctangent penalty is illustrated in Fig. 2. A comparison of these three penalties is illustrated in Fig. 1 of Ref. [20] for a fixed value of a. Of the three penalties, the arctangent penalty increases the slowest for a fixed value of a. The penalties have the property that as a increases, the penalty function $\phi(t;a)$ increases more slowly as a function of t, becomes more concave at $t = 0^+$, and induces sparsity more strongly (i.e., by mildly penalizing large values).

We mention that we do not use the simpler form of the arctangent penalty, $\phi(t;a) = (1/a)\tan^{-1}(a|t|)$, as it does not satisfy $\phi''(0^+;a) = -a$ which is property P7 listed above.

Corresponding to a penalty ϕ having the above properties, we define a smooth concave function.

Definition 1. *Let $\phi\colon \mathbb{R} \to \mathbb{R}$ be a penalty function satisfying the properties listed above. For $a \geqslant 0$, we define $s\colon \mathbb{R} \to \mathbb{R}$,*

$$s(t;a) = \phi(t;a) - |t|. \tag{6}$$

Figure 2 illustrates the function s corresponding to the arctangent penalty for $a = 0.3$. The following proposition follows straightforwardly [61].

Proposition 1. *Let $a \geqslant 0$. Let $\phi\colon \mathbb{R} \to \mathbb{R}$ be a penalty function satisfying the properties listed above. The function $s\colon \mathbb{R} \to \mathbb{R}$ in Definition 1 is twice continuously differentiable, concave, and satisfies*

$$-a \leqslant s''(t;a) \leqslant 0. \tag{7}$$

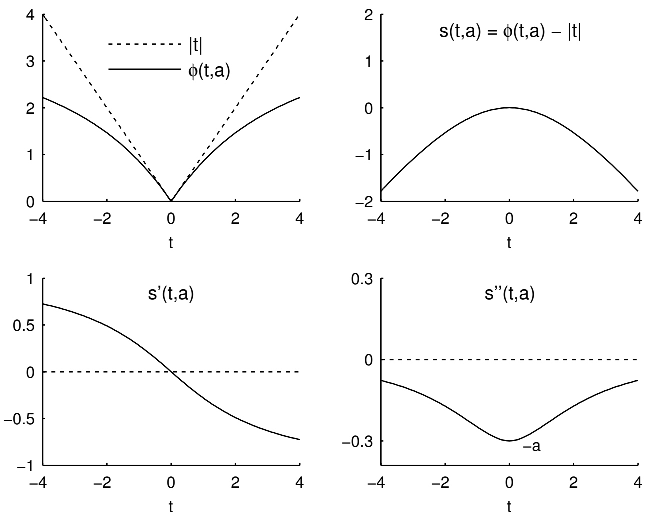

Fig. 2. A univariate penalty ϕ, its corresponding function s, and the first and second-order derivatives of s. The function s is twice continuously differentiable and concave.

This property will be of particular importance. Note that the value $\phi''(0)$ is not defined since ϕ is not differentiable at zero. But the value $s''(0)$ is defined (and is equal to $-a$).

$$s''(t;a) = \begin{cases} -a, & t = 0 \\ \phi''(t;a) & t \neq 0. \end{cases} \tag{8}$$

Also, although $\phi'(0)$ is not defined [because $\phi'(0^+) = 1$ and $\phi'(0^-) = -1$], the value $s'(0)$ is defined [$s'(0) = 0$]. Many of the properties listed above for ϕ are inherited by s, such as the symmetry and scaling properties:

$$s(-t;a) = s(t;a) \tag{9}$$
$$s(t;a) = (b/a)\,s\big(at/b;b\big) \tag{10}$$
$$s'(t;a) = s'\big(at/b;b\big) \tag{11}$$
$$s''(t;a) = (a/b)\,s''\big(at/b;b\big) \tag{12}$$

The following proposition follows straightforwardly.

Proposition 2. *Let $\phi\colon \mathbb{R} \to \mathbb{R}$ satisfy the properties listed above. Let $s\colon \mathbb{R} \to \mathbb{R}$ be given by Definition 1. Let $\lambda > 0$. If $0 \leqslant a \leqslant 1/\lambda$, then the functions $g\colon \mathbb{R} \to \mathbb{R}$*

$$g(t) = t^2/2 + \lambda s(t;a) \tag{13}$$

and $f\colon \mathbb{R} \to \mathbb{R}$

$$f(t) = t^2/2 + \lambda\phi(t;a) \tag{14}$$

are convex functions.

III. Bivariate Concave Function

The bivariate penalty to be given in Sec. IV will be defined in terms of a concave bivariate function S. The role of S, in describing the bivariate penalty, will be analogous to the role of s in describing the univariate penalty ϕ.

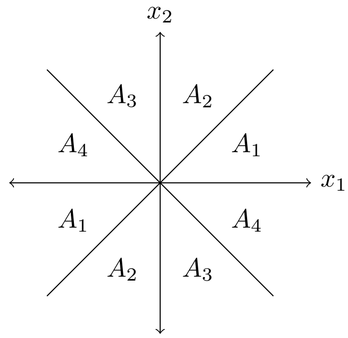

Fig. 3. Regions A_1 through A_4 in Definition 2.

Definition 2. *Let $a = (a_1, a_2)$ with $a_i \geqslant 0$. Let $\phi\colon \mathbb{R} \to \mathbb{R}$ be a univariate penalty function having the properties listed in Sec. II. Let $s\colon \mathbb{R} \to \mathbb{R}$ be given by Definition 1. If at least one of $\{a_1, a_2\}$ is non-zero, we define the function $S\colon \mathbb{R}^2 \to \mathbb{R}$ as*

$$S(x; a) = \begin{cases} s(x_1 + rx_2; \alpha) + (1 - r)\, s(x_2; a_1), & x \in A_1 \\ s(rx_1 + x_2; \alpha) + (1 - r)\, s(x_1; a_1), & x \in A_2 \\ s(rx_1 + x_2; \alpha) + (1 + r)\, s(x_1; a_2), & x \in A_3 \\ s(x_1 + rx_2; \alpha) + (1 + r)\, s(x_2; a_2), & x \in A_4 \end{cases}$$

where

$$\alpha = \frac{a_1 + a_2}{2}, \qquad r = \frac{a_1 - a_2}{a_1 + a_2} \tag{15}$$

and sets $A_i \subset \mathbb{R}^2$ are

$$A_1 = \{x \in \mathbb{R}^2 \mid x_2(x_1 - x_2) \geqslant 0\} \tag{16}$$
$$A_2 = \{x \in \mathbb{R}^2 \mid x_1(x_1 - x_2) \leqslant 0\} \tag{17}$$
$$A_3 = \{x \in \mathbb{R}^2 \mid x_1(x_1 + x_2) \leqslant 0\} \tag{18}$$
$$A_4 = \{x \in \mathbb{R}^2 \mid x_2(x_1 + x_2) \leqslant 0\} \tag{19}$$

as shown in Fig. 3. If both $a_i = 0$, we define $S(x; 0) = 0$.

The non-negative parameters a_i characterize how strongly concave S is. Figure 4 illustrates the function S where we set $a_1 = 1.5$ and $a_2 = 0.3$ and we let ϕ be the arctangent penalty in Table I.

The function S has three symmetries:

$$S((x_1, x_2); a) = S((x_2, x_1); a) \tag{20a}$$
$$= S((-x_1, -x_2); a) \tag{20b}$$
$$= S((-x_2, -x_1); a) \tag{20c}$$

(any two of which imply the remaining one). Equivalently, S is symmetric with respect to the origin and the two lines $x_1 = x_2$ and $x_1 = -x_2$. These symmetries follow directly from Definition 2 and from the symmetry of the univariate function s.

The following lemmas are proven in the Appendix. It will be useful in the proofs to note some identities. First, note that $S(0; a) = 0$. From the definitions of α and r in (15), we have

$$|r| \leqslant 1, \qquad (1 + r)\, \alpha = a_1 \tag{21}$$
$$r\alpha = (a_1 - a_2)/2, \qquad (1 - r)\, \alpha = a_2. \tag{22}$$

Lemma 1. *The bivariate function $S\colon \mathbb{R}^2 \to \mathbb{R}$ in Definition 2 is twice continuously differentiable and concave on $\mathbb{R}^2$.*

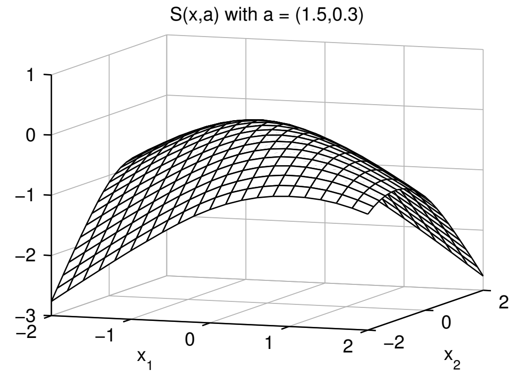

Fig. 4. Non-separable concave function S in Definition 2.

Lemma 2. *Let $a = (a_1, a_2)$ with $a_i \geqslant 0$. The Hessian of the bivariate function S in Definition 2 satisfies*

$$-K(a) \preccurlyeq \nabla^2 S(x; a) \preccurlyeq 0, \quad \textit{for all } x \in \mathbb{R}^2 \tag{23}$$

where $K(a)$ is defined by its eigenvalue decomposition

$$K(a) := Q^{\mathsf{T}} \begin{bmatrix} a_1 & 0 \\ 0 & a_2 \end{bmatrix} Q = \frac{1}{2} \begin{bmatrix} a_1 + a_2 & a_1 - a_2 \\ a_1 - a_2 & a_1 + a_2 \end{bmatrix} \tag{24}$$

where Q is the orthonormal matrix defined in (3). Furthermore, the lower bound is attained at $x = 0$, i.e.,

$$\nabla^2 S(0; a) = -K(a). \tag{25}$$

Lemma 2 is a key result for the subsequent results. The lemma can be illustrated in terms of ellipses. If M is a positive semidefinite matrix, then the set $\mathcal{E}[M] = \{x : x^{\mathsf{T}} M^{-1} x \leqslant 1\}$ is an ellipsoid [11]. In addition, $M_1 \preccurlyeq M_2$ if and only if $\mathcal{E}[M_1] \subseteq \mathcal{E}[M_2]$. In Fig. 5, we let ϕ be the logarithmic penalty in Table I and we set $a_1 = 1.5$ and $a_2 = 0.3$. The ellipses corresponding to $K(a)$ and $-\nabla^2 S(x; a)$ are shown in gray and black, respectively. For each x, the black ellipse is contained within the gray ellipse, illustrating (23). At $x = 0$, the black and gray ellipses coincide, reflecting the fact that the Hessian of S is equal to $-K(a)$ at the origin (25). For large x, the black ellipse shrinks, indicating that the function S becomes less concave away from the origin. (This is true for each penalty listed in Table I because the second derivative of each penalty goes monotonically to zero.)

Theorem 1. *Let $a = (a_1, a_2)$ with $a_i \geqslant 0$. Let $S\colon \mathbb{R}^2 \to \mathbb{R}$ be the function in Definition 2. Let $K(\gamma) = Q\,\Gamma\,Q^{\mathsf{T}}$ with eigenvalues $\gamma_i \geqslant 0$ be the positive semidefinite matrix defined in (2). The function $g\colon \mathbb{R}^2 \to \mathbb{R}$,*

$$g(x; a, \gamma) = \frac{1}{2} x^{\mathsf{T}} K(\gamma)\, x + \lambda\, S(x; a), \;\; \lambda > 0, \tag{26}$$

is convex if

$$0 \leqslant a_1 \leqslant \gamma_1/\lambda, \qquad 0 \leqslant a_2 \leqslant \gamma_2/\lambda. \tag{27}$$

Proof. Since g is twice continuously differentiable, it is sufficient to show the Hessian of g is positive semidefinite. The Hessian of g is given by

$$\nabla^2 g(x; a, \gamma) = K(\gamma) + \lambda[\nabla^2 S(x; a)] \tag{28}$$

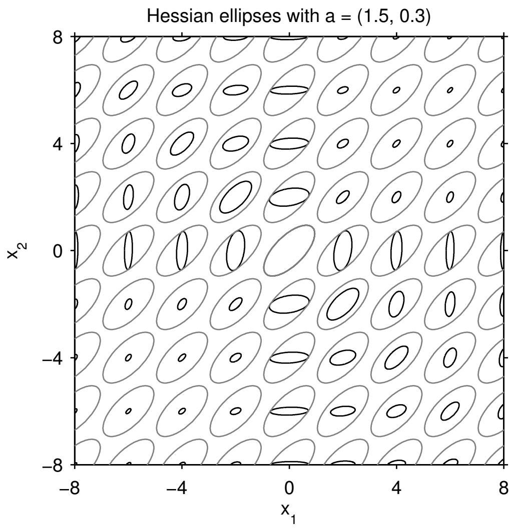

Fig. 5. Illustration of Lemma 2. The black ellipses are inside the gray ellipses, indicating that $0.5x^\mathsf{T}Kx + S(x;a)$ is convex.

$$= Q^\mathsf{T}\Gamma\, Q + \lambda[\nabla^2 S(x;a)]. \tag{29}$$

From Lemma 2 it follows that

$$\nabla^2 g(x;a,\gamma) \succcurlyeq Q^\mathsf{T} \begin{bmatrix} \gamma_1 - \lambda a_1 & 0 \\ 0 & \gamma_2 - \lambda a_2 \end{bmatrix} Q. \tag{30}$$

Hence, $g(x;a,\gamma)$ is convex if $\gamma_i - \lambda a_i \geqslant 0$ for $i = 1, 2$. This proves the result. □

IV. Bivariate Penalties

In this section, we define a non-convex non-separable bivariate penalty. Our intention is to strongly induce sparsity in solutions of problem (1) while maintaining the convexity of the problem. The penalty is parameterized by two non-negative parameters a_1 and a_2, which we restrict so as to ensure convexity of the objective function.

Definition 3. *Let $a = (a_1, a_2)$ with $a_i \geqslant 0$. Let $\phi\colon \mathbb{R} \to \mathbb{R}$ be a univariate penalty function having the properties listed in Sec. II. Let $S\colon \mathbb{R}^2 \to \mathbb{R}$ be the corresponding function in Definition 2. We define the bivariate penalty function $\psi\colon \mathbb{R}^2 \to \mathbb{R}$ as*

$$\psi(x;a) = S(x;a) + \|x\|_1. \tag{31}$$

If $a_1 \neq a_2$, then the penalty ψ is non-separable. Figure 6 illustrates ψ for the parameter values $a_1 = 1.5$ and $a_2 = 0.3$ where ϕ is the arctangent penalty in Table I. The degree of non-convexity differs in different quadrants.

It is informative to consider special cases of the bivariate penalty. If $a_1 > 0$ and $a_2 = 0$, then ψ simplifies to

$$\psi(x) = |x_1| + |x_2| + \phi(x_1 + x_2; a_1/2) - |x_1 + x_2|. \tag{32}$$

If $a_1 = a_2$, then the penalty reduces to a separable function, $\psi(x) = \phi(x_1;a_1) + \phi(x_2;a_1)$ (see Fig. 1(b)). If $a_1 = a_2 = 0$, then it further reduces to the ℓ_1 norm, i.e., $\psi(x) = |x_1| + |x_2|$ (see Fig. 1(a)). In any case, if either a_1 or a_2 is positive, then ψ is non-convex.

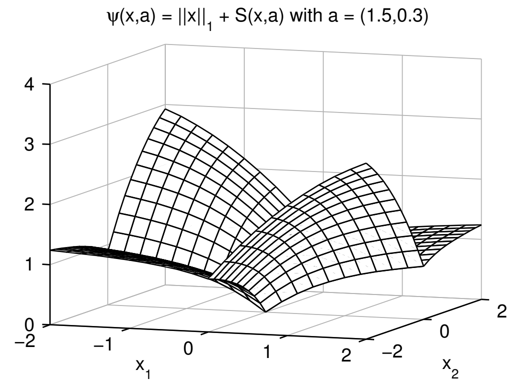

Fig. 6. Non-separable non-convex penalty ψ in Definition 3.

The following theorem, based on Theorem 1, states how to restrict the parameters a_i to ensure problem (1) is convex.

Theorem 2. *Let $\psi\colon \mathbb{R}^2 \to \mathbb{R}$ be the bivariate penalty in Definition 3. Suppose $H^\mathsf{T}H = Q^\mathsf{T}\Gamma Q$ where Q and Γ are given by (3). If $a = (a_1, a_2)$ satisfy $0 \leqslant a_i \leqslant \gamma_i/\lambda$, then the bivariate objective function $f := \mathbb{R}^2 \to \mathbb{R}$,*

$$f(x;a) = \frac{1}{2}\|y - Hx\|_2^2 + \lambda\, \psi(x;a), \tag{33}$$

is convex.

Proof. We write

$$f(x;a) = g(x;a) + \frac{1}{2}y^\mathsf{T}y - y^\mathsf{T}Hx + \lambda\, \|x\|_1 \tag{34}$$

where g is given by (26) with $K = H^\mathsf{T}H$ therein. From Theorem 1, g is convex. Hence, f is convex because it is the sum of convex functions. □

Theorem 2 gives a range for parameters a_1 and a_2 to ensure the objective function f is convex. Precisely, the parameters should be bounded, respectively, by the eigenvalues of $(1/\lambda)H^\mathsf{T}H$. To maximally induce sparsity, the parameters should be set to the maximal (critical) values, $a_i = \gamma_i/\lambda$.

Note that, even when the matrix H is singular (i.e., $\gamma_1 = 0$ or $\gamma_2 = 0$), the bivariate penalty can be non-convex without spoiling the convexity of the objective function f (if at least one of γ_i is positive). In other words, we need *not* sacrifice the convexity of the objective function f in order to use sparsity-inducing non-convex penalties, even when H is *singular*. This is an impossibility when the penalty ψ is a separable function.

Other bivariate penalties can be defined that ensure the objective function is convex; however, the one defined here satisfies a further property we think should be required of a bivariate penalty. Namely, the proposed bivariate penalty lies between the two separable penalties corresponding to the minimum and maximum parameters a_i.

Theorem 3. *Let $a = (a_1, a_2)$ with $a_i \geqslant 0$. Set $a_{\min} = \min\{a_1, a_2\}$ and $a_{\max} = \max\{a_1, a_2\}$. The bivariate penalty ψ in Definition 3 satisfies*

$$\phi(x_1;a_{\max}) + \phi(x_2;a_{\max}) \leqslant \psi(x;a) \leqslant \phi(x_1;a_{\min}) + \phi(x_2;a_{\min}). \tag{35}$$

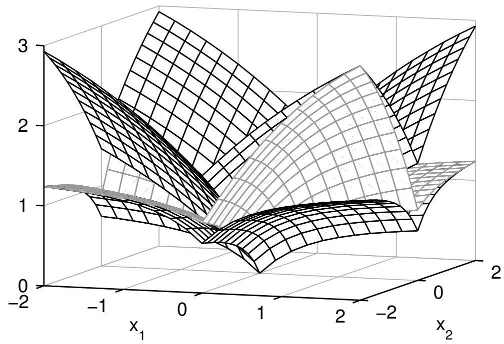

Fig. 7. Illustration of Theorem 3. The non-separable penalty ψ lies between two separable penalties.

The theorem is proven in Appendix C. (The bounds are illustrated by an animation in the supplemental material.)

The inequality in Theorem 3 is tight: the lower and upper bounds are individually satisfied with equality on the lines $\{x = (t, t)\}$ and $\{x = (t, -t)\}$, as illustrated in Fig. 7. Note that when $a_1 = a_2$ (i.e., $a_{\min} = a_{\max}$), the theorem requires the penalty ψ to be separable. Indeed, the penalty (31) is separable when $a_1 = a_2$.

We think a bivariate penalty should satisfy the inequality of Theorem 3 for the following reason. In this work, we aim to induce pure sparsity (i.e., not structured sparsity, etc.). Therefore, when $H^{\mathsf{T}}H$ is a diagonal matrix we should use a separable penalty. (A separable penalty best reflects an iid prior.) It follows that when $H^{\mathsf{T}}H = \gamma_1 I$, the most suitable penalty (maintaining convexity of f) is the separable one: $\phi(x_1; \gamma_1/\lambda) + \phi(x_2; \gamma_1/\lambda)$. A parameterized bivariate penalty should recover this separable penalty as a special case. Moreover, if $H^{\mathsf{T}}H = \gamma_2 I$ with $\gamma_2 < \gamma_1$, then the most suitable penalty is again a separable one: $\phi(x_1; \gamma_2/\lambda) + \phi(x_2; \gamma_2/\lambda)$. But $\gamma_2 < \gamma_1$ means the corresponding data fidelity term is less strongly convex and thus the penalty term must be less strongly non-convex. Consequently, we must have

$$\phi(x_1; \gamma_1/\lambda) + \phi(x_2; \gamma_1/\lambda) < \phi(x_1; \gamma_2/\lambda) + \phi(x_2; \gamma_2/\lambda).$$

When $H^{\mathsf{T}}H$ has distinct eigenvalues γ_1 and γ_2, the most suitable bivariate penalty should lie between these two separable penalties. Theorem 3 assures this.

If not suitably designed and utilized, it is conceivable that a non-separable penalty may lead to correlation or structure in the estimated signal that is not present in the true sparse signal. To avoid unintentionally inducing correlation in the estimated signal, it seems reasonable that the bivariate penalty should exhibit some similarity to the corresponding separable penalties (that reflect unstructured sparsity). Theorem 3 indicates the bivariate penalty (31) conforms to the relevant separable penalties. It is still possible that some erroneous correlation might be introduced, but such correlation is not evident in the experimental results. We attribute this to Theorem 3.

One may question the legitimacy of a method wherein penalty parameters are set according to the data fidelity term. Conventionally, the penalty term should reflect prior knowledge of the signal to be estimated; it should not depend on H, which represents the observation model. (This is formalized in the Bayesian perspective where the objective function corresponds to a likelihood function and the penalty term corresponds to a prior). The approach taken here, wherein the parameters of the penalty term are based on properties of H, appears to violate this principle. However, the common practice of restricting the penalty to be convex violates this principle even more. Probability densities (e.g., generalized Gaussian [50], mixture models [21], [64], Bessel-K [31], and α-stable [1]), that accurately model sparsity, correspond to non-convex penalties. Using the ℓ_1 norm as a penalty corresponds to the Laplace distribution (a relatively weak sparsity model). The proposed bivariate sparse regularization (BISR) approach is simply intended to follow a sparsity prior more closely. We finally note that the sparsity-inducing non-separable "Type II" penalty [81] also depends on H and λ.

A. Separable penalties

To clarify the value of non-separable regularization, we note a limitation of separable penalties.

Lemma 3. *Let the univariate penalty $\phi\colon \mathbb{R} \to \mathbb{R}$ satisfy the properties of Sec. II. The objective function $f\colon \mathbb{R}^2 \to \mathbb{R}$,*

$$f(x; a) = \frac{1}{2}\|y - Hx\|_2^2 + \lambda\, \phi(x_1; a) + \lambda\, \phi(x_2; a), \tag{36}$$

with $\lambda > 0$ and $a \geqslant 0$, is convex only if

$$\phi''(0^+) \geqslant -(1/\lambda) \min\{\gamma_1, \gamma_2\}, \tag{37}$$

or equivalently, $0 \leqslant a \leqslant \min\{\gamma_1, \gamma_2\}/\lambda$, where γ_i are the eigenvalues of $H^{\mathsf{T}}H$.

According to Lemma 3, a *separable* non-convex penalty, that ensures convexity of the objective function, is limited by the minimum eigenvalue of $H^{\mathsf{T}}H$. It cannot exploit the greater eigenvalue. This is unfavorable, because when one of the eigenvalues is close to zero, a separable penalty can be only mildly non-convex and provides negligible improvement relative to the ℓ_1 norm; when $H^{\mathsf{T}}H$ is singular, the penalty reduces to the ℓ_1 norm exactly. In contrast, a non-separable penalty can exploit both eigenvalues independently. Hence, non-separable penalties are most advantageous when the eigenvalues of $H^{\mathsf{T}}H$ are quite different in value.

V. Sparse Reconstruction

Practical problems in signal processing involve far more than two variables. Therefore, the proposed bivariate penalty (31) and convexity condition (27) are of little practical use on their own. In this section we show how they can be used to solve an N-point linear inverse problem (with $N > 2$). We consider the problem of estimating a signal $x \in \mathbb{R}^N$ given y,

$$y = Hx + w \tag{38}$$

where H is a known linear operator, x is known to be sparse, and w is additive white Gaussian noise (AWGN). We formulate the estimation of x as an optimization problem with bivariate sparse regularization (BISR),

$$\hat{x} = \arg\min_{x \in \mathbb{R}^N} \Big\{ F(x) = \frac{1}{2}\|y - Hx\|_2^2 + \frac{\lambda}{2} \sum_n \psi((x_{n-1}, x_n); a) \Big\}, \quad (39)$$

where $\lambda > 0$, $a = (a_1, a_2)$ and $\psi \colon \mathbb{R}^2 \to \mathbb{R}$ is the proposed bivariate penalty (31). In the penalty term, the first and last signal value pairs, (x_0, x_1) and (x_N, x_{N+1}), straddle the end-points of x. As noted in Sec. I-C, we define $x_n = 0$ for $n \notin \{1, 2, \dots, N\}$, which simplifies subsequent notation.

If $a_1 = a_2$, then the bivariate penalty is separable, i.e., $\psi(u; a) = \phi(u_1; a_1) + \phi(u_2; a_1)$, and the N-point penalty term in (39) reduces to $\lambda \sum_n \phi(x_n; a_1)$. Hence, we recover the standard (separable) formulation of sparse regularization. In particular, if $a_1 = a_2 = 0$, then $\psi(u; 0) = |u_1| + |u_2|$ and the N-point penalty term reduces to $\lambda\|x\|_1$, i.e., the classical sparsity-inducing convex penalty.

In order to induce sparsity more effectively, we allow ψ to be non-separable; i.e., $a_1 \neq a_2$. To that end, the following section addresses the problem of how to set a_1 and a_2 in the bivariate penalty ψ to ensure convexity of the N-variate objective function F in (39).

A. Convexity condition

Lemma 4. *Let $F \colon \mathbb{R}^N \to \mathbb{R}$ be defined in* (39) *where $\psi \colon \mathbb{R}^2 \to \mathbb{R}$ is a parameterized bivariate penalty as defined in Definition 3. Let P be a positive semidefinite symmetric tridiagonal Toeplitz matrix,*

$$P = \begin{bmatrix} p_0 & p_1 & & & \\ p_1 & p_0 & p_1 & & \\ & \ddots & \ddots & \ddots & \\ & & p_1 & p_0 & p_1 \\ & & & p_1 & p_0 \end{bmatrix}, \quad (40)$$

such that $0 \preccurlyeq P \preccurlyeq H^\mathsf{T} H$. If the bivariate function $f \colon \mathbb{R}^2 \to \mathbb{R}$ defined as

$$f(u) = \frac{1}{2} u^\mathsf{T} \begin{bmatrix} p_0 & 2p_1 \\ 2p_1 & p_0 \end{bmatrix} u + \lambda \psi(u; a) \quad (41)$$

is convex, then F is convex.

The lemma is proven in Appendix D. According to the lemma, it is sufficient to restrict ψ so as to ensure convexity of the bivariate function f in (41). Therefore, the allowed penalty parameters a_i can be determined from the tridiagonal matrix P. Using Theorem 1 and Lemma 4, we obtain Theorem 4.

Theorem 4. *Let $F \colon \mathbb{R}^N \to \mathbb{R}$ be defined in* (39) *where $\psi \colon \mathbb{R}^2 \to \mathbb{R}$ is a parameterized bivariate penalty as defined in Definition 3. Let P be a symmetric tridiagonal Toeplitz matrix* (40) *satisfying $0 \preccurlyeq P \preccurlyeq H^\mathsf{T} H$. If*

$$0 \leqslant a_1 \leqslant (p_0 + 2p_1)/\lambda, \quad 0 \leqslant a_2 \leqslant (p_0 - 2p_1)/\lambda, \quad (42)$$

then F in (39) *is convex.*

Proof. Note the eigenvalue value decomposition

$$\begin{bmatrix} p_0 & 2p_1 \\ 2p_1 & p_0 \end{bmatrix} = Q^\mathsf{T} \begin{bmatrix} p_0 + 2p_1 & 0 \\ 0 & p_0 - 2p_1 \end{bmatrix} Q \quad (43)$$

where Q is the orthonormal matrix (3). Hence, f in (41) can be written as

$$f(u) = \frac{1}{2} u^\mathsf{T} Q^\mathsf{T} \begin{bmatrix} p_0 + 2p_1 & 0 \\ 0 & p_0 - 2p_1 \end{bmatrix} Q u + \lambda\, \psi(u; a). \quad (44)$$

By Theorem 1, if $0 \leqslant a_1 \leqslant (p_0 + 2p_1)/\lambda$ and $0 \leqslant a_2 \leqslant (p_0 - 2p_1)/\lambda$, then f is convex. It follows from Lemma 4 that F in (39) is convex. □

B. Optimality condition

In this section, we derive an explicit condition to verify the optimality of a prospective minimizer of the objective function F in (39). The optimality condition is also useful for monitoring the convergence of an optimization algorithm (see the animation in the supplemental material).

The general condition to characterize minimizers of a convex function is expressed in terms of the subdifferential. If F is convex, then $x^{\mathsf{opt}} \in \mathbb{R}^N$ is a minimizer if and only if $0 \in \partial F(x^{\mathsf{opt}})$ where ∂F is the subdifferential of F.

We seek an expression for the subdifferential of the objective function F. The function F in (39) has a regularization term that is non-differentiable, non-convex, and non-separable. But using (31), we may write the regularization term as

$$\frac{1}{2} \sum_n \psi((x_{n-1}, x_n); a) \quad (45)$$

$$= \frac{1}{2} \sum_n \Big[S((x_{n-1}, x_n); a) + \|(x_{n-1}, x_n)\|_1 \Big] \quad (46)$$

$$= \|x\|_1 + \frac{1}{2} \sum_n S((x_{n-1}, x_n); a) \quad (47)$$

where $x_n = 0$ for $n \notin \{1, 2, \dots, N\}$. We define $\Theta \colon \mathbb{R}^N \to \mathbb{R}$ as

$$\Theta(x; a) = \frac{1}{2} \sum_n S((x_{n-1}, x_n); a). \quad (48)$$

The function Θ is twice continuously differentiable because it is the sum of twice differentiable functions. Using (47), we may express the objective function F in (39) as

$$F(x) = \frac{1}{2}\|y - Hx\|_2^2 + \lambda \Theta(x; a) + \lambda \|x\|_1. \quad (49)$$

The benefit of (49) compared to (39) is that the regularization term (which is non-differentiable, non-convex, and non-separable) is separated into a differentiable part and a convex separable part. The Θ term is differentiable and its gradient is easily evaluated. The ℓ_1 norm is separable and convex and its subdifferential is easily evaluated.

The gradient of Θ is given by

$$[\nabla \Theta(x; a)]_n = \frac{1}{2} S_1((x_n, x_{n+1}); a) + \frac{1}{2} S_2((x_{n-1}, x_n); a) \quad (50)$$

where S_i is the partial derivative of $S((x_1, x_2))$ with respect to x_i. The subdifferential of the ℓ_1 norm is separable [11],

$$\partial \|x\|_1 = \mathrm{sign}(x_1) \times \cdots \times \mathrm{sign}(x_N) \quad (51)$$

where sign is the set-valued signum function

$$\operatorname{sign}(t) := \begin{cases} \{1\}, & t > 0 \\ [-1, 1], & t = 0 \\ \{-1\}, & t < 0. \end{cases} \tag{52}$$

Since the first two terms of (49) are differentiable, the subdifferential of F is

$$\partial F(x) = H^{\mathsf{T}}(Hx - y) + \lambda \nabla \Theta(x; a) + \lambda \partial \|x\|_1. \tag{53}$$

Hence the condition $0 \in \partial F(x^{\mathsf{opt}})$ can be expressed as

$$(1/\lambda) H^{\mathsf{T}}(y - Hx^{\mathsf{opt}}) - \nabla \Theta(x^{\mathsf{opt}}; a) \in \partial \|x^{\mathsf{opt}}\|_1. \tag{54}$$

Expressing this condition component-wise, we have the following result.

Theorem 5. *If $a = (a_1, a_2)$ is chosen so that the objective function F in* (39) *is convex, then x^{opt} minimizes F if and only if*

$$\frac{1}{\lambda}[H^{\mathsf{T}}(y - Hx^{\mathsf{opt}})]_n - [\nabla \Theta(x^{\mathsf{opt}}; a)]_n \\ \in \operatorname{sign}(x_n^{\mathsf{opt}}), \quad n = 1, \dots, N. \tag{55}$$

This condition can be depicted using a scatter plot as in Fig. 11 below. The points in the scatter plot show the left-hand-side of (55) versus x_n. A signal x^{opt} is a minimizer of F if and only if the points in the scatter plot lie on the graph of the set-valued signum function (e.g., Fig. 11).

C. Separable penalties

It is informative to consider the case of a separable penalty as in Sec. IV-A. (See also [9].)

Lemma 5. *Let the univariate penalty ϕ satisfy the properties of Sec. II. The objective function $F\colon \mathbb{R}^N \to \mathbb{R}$,*

$$F(x; a) = \frac{1}{2}\|y - Hx\|_2^2 + \lambda \sum_n \phi(x_n; a) \tag{56}$$

with $\lambda > 0$ and $a \geqslant 0$, is convex only if

$$0 \leqslant a \leqslant \gamma_{\min}/\lambda \tag{57}$$

where $\gamma_{\min}$ is the minimum eigenvalue of $H^{\mathsf{T}}H$.

Proof. Let $u \in \mathbb{R}^N$ be an eigenvector of $H^{\mathsf{T}}H$ corresponding to its minimum eigenvalue $\gamma_{\min}$, i.e., $H^{\mathsf{T}}Hu = \gamma_{\min}u$. Consider the restriction of F to a line defined by u. Namely, define $G\colon \mathbb{R} \to \mathbb{R}$ as

$$G(t) = F(tu; a) \tag{58}$$

$$= \frac{1}{2}\|y - tHu\|_2^2 + \lambda \sum_n \phi(tu_n; a). \tag{59}$$

We will show that G is not convex when $a > \gamma_{\min}/\lambda$. It will follow that F is not convex, because the restriction of a multivariate convex function to any line must also be convex.

By the properties of ϕ, the function G is twice continuously differentiable on $\mathbb{R}_+$ and its second derivative is given by

$$G''(t) = uH^{\mathsf{T}}Hu + \lambda \sum_n u_n^2\, \phi''(tu_n; a) \tag{60}$$

$$= \gamma_{\min} u^{\mathsf{T}}u + \lambda \sum_n u_n^2\, \phi''(tu_n; a). \tag{61}$$

Since $\phi''(0^+; a) = -a$ is a defining property of ϕ, we have

$$G''(0^+) = \gamma_{\min} u^{\mathsf{T}}u - a\lambda \sum_n u_n^2 \tag{62}$$

$$= (\gamma_{\min} - \lambda a)\, \|u\|_2^2. \tag{63}$$

Convexity of G requires $\gamma_{\min} - \lambda a \geqslant 0$, hence (57). □

As in Sec. IV-A, a separable penalty that ensures convexity of F is limited by the minimum eigenvalue of $H^{\mathsf{T}}H$. It can not take advantage of the spread of eigenvalues the way a non-separable penalty can.

D. Sparse Deconvolution

We apply Theorem 4 to the sparse deconvolution problem. In this case, the linear operator H represents convolution, i.e.,

$$[Hx]_n = \sum_k h_{n-k}\, x_k. \tag{64}$$

That is, H is a Toeplitz matrix. It represents a linear time-invariant (LTI) system with frequency response given by the Fourier transform of h,

$$H(\omega) = \sum_n h_n\, \mathrm{e}^{-\mathrm{j}\omega n}. \tag{65}$$

Similarly, the matrix P in (40) represents an LTI system with a real-valued frequency response,

$$P(\omega) = p_1 \mathrm{e}^{-\mathrm{j}\omega} + p_0 + p_1 \mathrm{e}^{\mathrm{j}\omega} \tag{66}$$

$$= p_0 + 2p_1 \cos(\omega). \tag{67}$$

Specializing Theorem 4 to the problem of deconvolution, we have the following result.

Theorem 6. *Let H in* (39) *represent convolution. Let P in* (67) *satisfy*

$$0 \leqslant P(\omega) \leqslant |H(\omega)|^2, \quad \forall \omega, \tag{68}$$

where $H(\omega)$ is given by (65)*. If $0 \leqslant a_1 \leqslant P(0)/\lambda$ and $0 \leqslant a_2 \leqslant P(\pi)/\lambda$, then F in* (39) *is convex.*

Proof. Using the convolution property of the discrete-time Fourier transform, condition (68) is equivalent to $0 \preccurlyeq P_\infty \preccurlyeq H_\infty^{\mathsf{T}} H_\infty$, where these are doubly-infinite Toeplitz matrices corresponding to discrete-time signals defined on $\mathbb{Z}$. Since any principal sub-matrix of a positive semidefinite matrix is also positive semidefinite, the inequality is also true for finite matrices P and $H^{\mathsf{T}}H$, which can be recognized as principal sub-matrices of the corresponding doubly-infinite matrices. Therefore, $0 \preccurlyeq P \preccurlyeq H^{\mathsf{T}}H$, and by Theorem 4, the objective function F is convex if a_i satisfy (42). Noting that $P(0) = p_0 + 2p_1$ and $P(\pi) = p_0 - 2p_1$ yields the result. □

To induce sparsity as strongly as possible, $P(\omega)$ should be as close as possible to the upper bound $|H(\omega)|^2$. The determination of $P(\omega)$ satisfying such constraints can be efficiently and exactly performed using semi-definite programming (SDP) as described by Dumitrescu [30]. Since P is low-order here, the SDP computation is negligible.

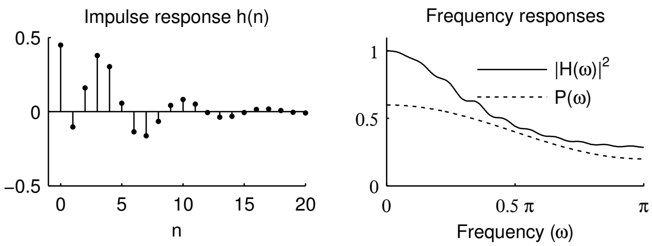

Fig. 8. Filters $H(\omega)$ and $P(\omega)$ for Example 1.

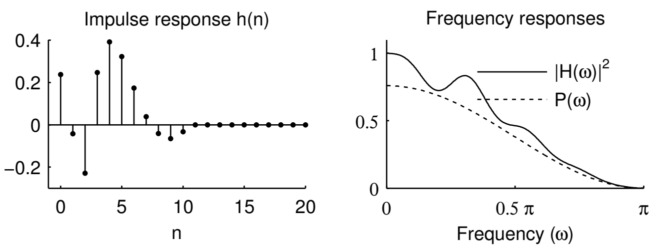

Fig. 9. Filters $H(\omega)$ and $P(\omega)$ for Example 2.

An example is illustrated in Fig. 8. The impulse response h and the square magnitude of the frequency response $|H(\omega)|^2$ are shown in Fig. 8. The frequency response $P(\omega) = 0.4+0.2\cos(\omega)$ is real-valued, non-negative, and approximates $|H(\omega)|^2$ from below. According to Theorem 6, the objective function F is convex if $0 \leqslant a_1 \leqslant 0.6/\lambda$ and $0 \leqslant a_2 \leqslant 0.2/\lambda$. This filter will be used in Example 1 below (Sec. VII-A).

Another example is illustrated in Fig. 9. The frequency response $H(\omega)$ has a null at $\omega = \pi$. Hence, the system H is not invertible. Since $H(\pi) = 0$, any P satisfying (68) also has $P(\pi) = 0$. We find $P(\omega) = 0.38(1+\cos\omega)$ satisfies (68); see Fig. 9(b). Therefore, according to Theorem 6, the objective function F is convex if $0 \leqslant a_1 \leqslant 0.76/\lambda$ and $a_2 = 0$. For $\{a_1 > 0, a_2 = 0\}$, the multivariate penalty is non-convex and non-separable. For $\{a_1 = a_2 = 0\}$, the penalty is simply the ℓ_1 norm (convex and separable). This filter will be used in Example 2 below (Sec. VII-B).

In reference to the filters H illustrated in Figs. 8 and 9, it is informative to consider the case of a separable penalty. When H represents convolution, Lemma 5 indicates that objective function F is convex only if the scalar parameter a satisfies $0 \leqslant a \leqslant \min_\omega |H(\omega)|^2/\lambda$. For the filter of Fig. 8 this leads to the constraint $0 \leqslant a \leqslant 0.26/\lambda$, meaning that the separable penalty may be non-convex (i.e., positive a is allowed). On the other hand, for the filter of Fig. 9 this leads to the constraint $a = 0$, meaning that the separable penalty must be convex (i.e., $\phi(x) = |x|$). In summary, when the filter H is not invertible, the only non-convex penalties maintaining convexity of the objective function F are *non-separable* penalties. Since inverse problems often involve singular or nearly singular operators H, this motivates the investigation of non-separable penalties.

VI. Iterative thresholding algorithm

We present an iterative thresholding algorithm to solve (39). The algorithm is an immediate application of forward-backward splitting (FBS) [25], [26]. The FBS algorithm minimizes a function of the form $f_1 + f_2$ where both f_1 and f_2 are convex and additionally ∇f_1 is Lipschitz continuous. To apply the FBS algorithm to problem (39), we express F using (49) as

$$F(x) = f_1(x) + f_2(x) \tag{69}$$

where

$$f_1(x) = \frac{1}{2}\|y - Hx\|_2^2 + \lambda\Theta(x; a) \tag{70}$$

$$f_2(x) = \lambda\|x\|_1. \tag{71}$$

The function f_1 is smooth and convex. We recall that Θ is twice continuously differentiable. It is also concave because it is a sum of concave functions. Since Θ is concave, the Lipschitz constant of ∇f_1 is bounded by ρ where ρ is the maximum eigenvalue of $H^\mathsf{T} H$.

An FBS algorithm to solve problem (39) is given by

$$z^{(k)} = x^{(k)} - \mu\Big[H^\mathsf{T}(Hx^{(k)} - y) + \lambda\nabla\Theta(x^{(k)}; a)\Big] \tag{72a}$$

$$x^{(k+1)} = \operatorname{soft}(z^{(k)}, \mu\lambda) \tag{72b}$$

where $0 < \mu < 2/\rho$ where ρ is the maximum eigenvalue of $H^\mathsf{T} H$. The parameter μ can be viewed as a step-size. The soft thresholding function

$$\operatorname{soft}(t, T) := \begin{cases} t - T, & t \geqslant T \\ 0, & |t| \leqslant T \\ t + T, & t \leqslant -T \end{cases} \tag{73}$$

is applied element-wise to vector $z^{(k)}$. As an FBS algorithm, it is guaranteed that $x^{(k)}$ converges to a minimizer of F.

This algorithm (72) resembles the classical iterative shrinkage/thresholding algorithm (ISTA) [27], [33]. ISTA is an FBS algorithm for the particular case of minimizing the sum of a quadratic term and an ℓ_1 norm, so that ISTA and FBS have the same convergence properties. (Accordingly, ISTA converges also for $0 < \mu < 2/\rho$.) In practice, larger step-sizes often yield faster convergence of the algorithm. We implement the algorithm with $\mu = 1.9/\rho$.

We further note that the algorithm (72) with $0 < \mu < 2/\rho$ has the property that $F(x^{(k)})$ monotonically decreases [5], [72]. (Accordingly, since ISTA is a special case, it also has the monotonic descent property for $0 < \mu < 2/\rho$.)

Other algorithms can also be used to solve problem (39). For example, the fast ISTA (FISTA) algorithm [8] is directly applicable. For some non-convex settings, proximal algorithms have recently been developed [23] where the convergence of the iterates to a critical point is guaranteed. This is an interesting perspective for future work.

VII. Numerical Examples

A. Example 1

The 100-point sparse signal illustrated in Fig. 10(a) is convolved with the impulse response h shown in Fig. 8. The convolved signal is corrupted by additive white Gaussian noise (AWGN) with standard deviation $\sigma = 4$. The corrupted signal is shown in Fig. 10(b). To perform sparse deconvolution using BISR, we need to define the univariate penalty ϕ and set the regularization parameters λ and $a = (a_1, a_2)$ in the objective

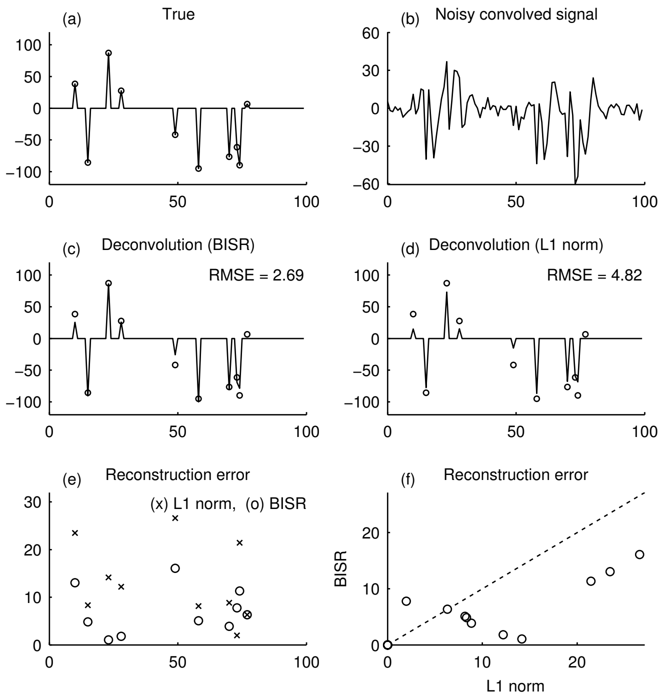

Fig. 10. Example 1 of sparse deconvolution using BISR.

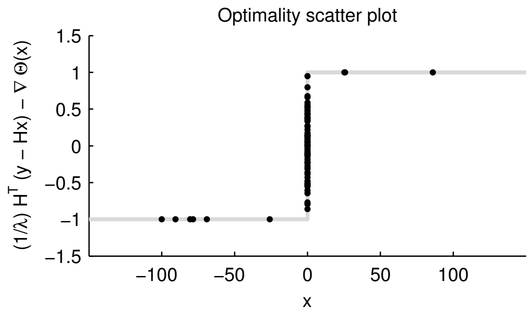

Fig. 11. Illustration of Theorem 5 (optimality) for Example 1.

TABLE II
AVERAGE RMSE AND RUN-TIME FOR EXAMPLE 1

Algorithm	$\sigma=1$	$\sigma=2$	$\sigma=4$	$\sigma=8$	$\sigma=16$	(msec)
L1	1.17	2.32	4.43	8.19	13.47	3.5
L1+debiasing	0.62	1.26	2.57	5.46	11.92	3.6
Lp (p = 0.5)	0.66	1.19	2.39	5.12	12.05	4.3
SBR (L0)	0.65	1.15	2.38	5.33	15.35	2.1
IMSC	0.50	1.00	2.23	5.00	11.02	323.9
IPS	0.51	1.02	2.23	4.89	10.96	5.1
BISR (log)	0.52	1.11	2.53	5.62	11.58	8.2
BISR (rat)	0.51	1.06	2.41	5.42	11.46	8.6
BISR (atan)	0.50	1.03	2.30	5.13	11.22	8.8

function (39). For ϕ, we use the arctangent penalty in Table I. We set λ straightforwardly as

$$\lambda = \beta\sigma\|h\|_2 \tag{74}$$

which follows from an analysis of sparse optimality conditions [36], [37], [69]. We set $\beta = 2.5$, similar to the 'three sigma' rule. This choice of β is not intended to minimize the mean square error, but rather to inhibit false impulses appearing in the estimated sparse signal.

To set a, we use the non-negative function $P(\omega)$ shown in Fig. 8. This filter has $P(0) = 0.6$ and $P(\pi) = 0.2$. Therefore, according to Theorem 6, the objective function is convex if $0 \leqslant a_1 \leqslant 0.6/\lambda$ and $0 \leqslant a_2 \leqslant 0.2/\lambda$. To maximally induce sparsity, we set a_i to their respective maximal values. To perform deconvolution using BISR (i.e., to minimize the objective function F) we use the iterative thresholding algorithm (Sec. VI) with a step-size of $\mu = 1.9/\rho$ where $\rho = \max_\omega |H(\omega)|^2$ gives $\rho = 1$. We run the algorithm until a stopping condition is satisfied. As a stopping condition, we use $\|x^{(k+1)} - x^{(k)}\|_\infty \leqslant 10^{-4} \times \|x^{(k)}\|_\infty$ where k is the iteration index. The run-time (averaged over 50 signal and noise realizations) is about 8.8 milliseconds on a 2013 MacBook Pro (2.5 GHz Intel Core i5) running Matlab R2011a.

The BISR solution is shown in Fig. 10(c). It has a root-mean-square-error (RMSE) of 2.7, about 56% that of the ℓ_1 norm solution, shown in Fig. 10(d) for comparison. Figure 10(e) shows the reconstruction error of both solutions. The relative accuracy of the BISR solution is further illustrated in the scatter plot of Fig. 10(f), which shows the error of the two solutions plotted against each other. Most of the points lie below the diagonal line, meaning that the BISR solution has less error than the ℓ_1 norm solution.

We verify the optimality of the obtained BISR solution in Fig. 11 using Theorem 5. The obtained solution is validated as a global minimizer of the objective function because the points in the scatter plot lie on the graph of the signum function.

We compare the proposed BISR method with several other algorithms in Table II and Fig. 12. For the comparison, we vary the noise standard deviation σ and generate 200 sparse signals and noise realizations for each value of σ. Each sparse signal consists of 10 randomly located impulses with amplitudes uniformly distributed between -100 and 100. Table II gives the average RMSE for each σ, and the average run-time for $\sigma = 4$ (with all algorithms implemented in Matlab on the same computer).

Included in the comparison are the following methods: ℓ_1 norm regularization, ℓ_1 norm regularization with debiasing, ℓ_p pseudo-norm regularization with p of 0.5, the *single best replacement* (SBR) algorithm [73], *iterative maximally sparse convex* (IMSC) regularization [69], and the *iterative p-shrinkage* (IPS) algorithm [78], [82].

Each method provides an improvement over ℓ_1 norm regularization. The first approach is ℓ_1 norm regularization with debiasing. Debiasing is a post-processing step that applies unbiased least squares approximation to re-estimate the non-zero amplitudes [34]. This method solves the systematic underestimation of non-zero amplitudes from which ℓ_1 norm regularization suffers, but it is still influenced by noise in the observed data. As shown in Table II, several other methods generally perform better than ℓ_1 regularization with debiasing.

Non-convex regularization using the ℓ_p pseudo-norm with $0 < p < 1$ is also a common approach for improving upon ℓ_1 norm regularization. We use $p = 0.5$ with λ set according to (74) with $\beta = 8$, a value we found worked well on average

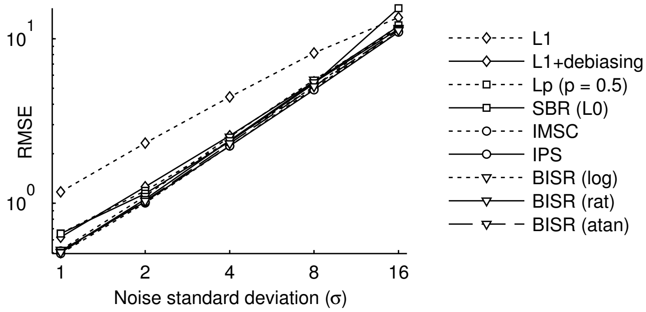

Fig. 12. Average RMSE for Example 1.

for this example. This method performed similarly as ℓ_1 norm regularization with debiasing.

For non-convex regularization using the ℓ_0 pseudo-norm we use the SBR algorithm [73]. A comparison of several methods for ℓ_0 pseudo-norm regularization showed the SBR algorithm to be state-of-the-art [69]. We use SBR with λ set according to (74) with $\beta = 70$, which we found worked well here except for high noise levels. We were unable to find a single β value that worked well over the range of σ considered here. The SBR algorithm performed similarly to ℓ_p pseudo-norm regularization, except for the highest noise level of $\sigma = 8$.

The IMSC algorithm for sparse deconvolution proceeds by solving a sequence of convex sub-problems [69]. The regularization term of each sub-problem is individually designed to be maximally non-convex (i.e., sparsity-inducing). In contrast to the proposed BISR approach, IMSC does not solve a prescribed optimization problem — each iteration yields a new convex problem to be solved. The formulation of each problem in IMSC requires the solution to a semidefinite problem (SDP); therefore, the IMSC approach is very slow. (IMSC is up to 100 times slower than other algorithms; see Table II.) In IMSC, the parameter λ can be set the same as for ℓ_1 deconvolution; hence, we again set λ using (74) with $\beta = 2.5$. From Table II, IMSC performs better than the preceding methods.

The IPS algorithm [78], [82] generalizes the classic iterative shrinkage-thresholding algorithm (ISTA) [27], [33]. The IPS algorithm replaces soft-thresholding in ISTA by a threshold function that does not underestimate large-amplitude signal values. Notably, IPS can be understood as a method that seeks to minimize a prescribed (albeit implicit) non-convex objective function; furthermore, each iteration of IPS can be understood as the exact minimization of a convex problem. Moreover, as IPS is an iterative thresholding algorithm, it is computationally very efficient. For this example, IPS gives excellent results when the parameter λ is set using (74) with $\beta = 2.5$. It performs about the same as IMSC but runs much faster.

To complete the comparison, we compare the proposed BISR method using the three penalties listed in Table I. (The BISR method can be used with any penalty function satisfying the properties listed at the beginning of Sec. II.) Referring to Table II, the three penalties give about the same result at low noise levels; at higher noise levels, the arctangent penalty tends to perform better than the other two penalties. Compared to the other algorithms, BISR with the arctangent penalty does about as well as IPS and IMSC at low noise levels, but not quite as well at higher noise levels. Note that the proposed BISR approach is the only algorithm minimizing a prescribed convex objective function.

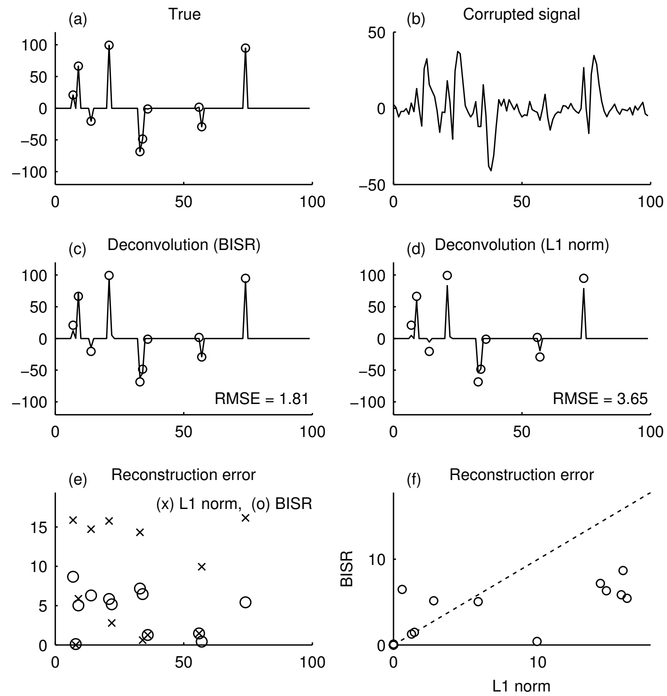

Fig. 13. Example 2 of sparse deconvolution using BISR.

We finally note that setting λ using the rule (74) is an empirical choice. Even though we choose a different β for different algorithms, this rule may favor some algorithms over other algorithms for which another rule would be preferable. For example, an efficient alternate way to set λ for an improved SBR algorithm has been developed [74]. Therefore, the forgoing comparisons indicate only tendencies; they do not lead to a conclusive ranking of methods.

B. Example 2

This is like Example 1, except we use the convolution filter H shown in Fig. 9. This filter is not invertible. The sparse signal (Fig. 13(a)) is convolved with the impulse response h and corrupted by AWGN ($\sigma = 4$). We set λ using (74) with $\beta = 2.5$ as in Example 1. Since the filter P has $P(0) = 0.76$ and $P(\pi) = 0$, Theorem 6 states that the objective function (39) is convex if $0 \leqslant a_1 \leqslant 0.76/\lambda$ and $a_2 = 0$. We run the algorithm of Sec. VI with the same stopping condition to obtain the BISR solution (Fig. 13(c)). The BISR solution has an RMSE of 1.8, which is about 50% that of the ℓ_1 norm solution (Fig. 13(d)). As in Example 1, the optimality of the obtained BISR solution can be readily verified.

It is noteworthy in this example that we obtain a strictly non-convex penalty ensuring convexity of the objective function F, because in this example the convolution filter H is not invertible. It is possible only because the utilized penalty is non-separable.

In Table III we compare the same algorithms as in Example 1. The regularization parameter λ for each algorithm was chosen as in Example 1. The remarks made in Example 1 mostly

TABLE III
AVERAGE RMSE AND RUN-TIME FOR EXAMPLE 2

Algorithm	$\sigma=1$	$\sigma=2$	$\sigma=4$	$\sigma=8$	$\sigma=16$	(msec)
L1	1.37	2.70	5.01	8.96	14.01	5.1
L1+debiasing	0.76	1.55	3.14	6.56	13.31	5.2
Lp (p = 0.5)	1.07	1.57	2.89	6.14	13.38	5.7
SBR (L0)	0.73	1.44	2.73	6.59	17.64	3.4
IMSC	0.54	1.18	2.69	6.26	12.38	305.9
IPS	0.59	1.20	2.66	5.88	12.41	6.9
BISR (log)	0.75	1.60	3.19	6.65	12.45	13.9
BISR (rat)	0.73	1.56	3.08	6.49	12.37	14.6
BISR (atan)	0.75	1.56	3.00	6.29	12.25	15.4

apply here, but with a few exceptions as follows. In contrast to the general tendency, the ℓ_p pseudo-norm performed worse than the debiased ℓ_1 norm solution for the lowest noise level ($\sigma = 1$). It appears that the performance of the ℓ_p pseudo-norm and SBR methods may degrade for low or high noise levels when one attempts to set λ proportional to σ. In addition, in contrast to Example 1, the BISR approach did not outperform ℓ_p pseudo-norm or SBR for several noise levels. IPS performs very well at most noise levels.

VIII. LIMITATIONS OF BIVARIATE PENALTIES

The examples in Sec. VII demonstrate the improvement attainable by non-separable bivariate penalties in comparison with separable penalties; however, the degree of improvement depends on the linear operator H in the data fidelity term. In particular, for some H, a bivariate penalty will offer little or no improvement in comparison with a separable penalty.

In reference to Figs. 8 and 9, the effectiveness of a bivariate penalty for sparse deconvolution depends on how well $|H(\omega)|^2$ can be approximated by a function $P(\omega)$ of the form (67) satisfying condition (68). Some filters H can not be well approximated by such a filter P. For example, if $H(\omega)$ is Gaussian or has a sharp peak, then it can not be well approximated by such a filter P. In this case, a bivariate penalty does not offer significant improvement in comparison with a separable penalty. For a second example, if the frequency response $H(\omega)$ has multiple nulls (e.g., a K-point moving average filter with $K > 2$), then the only filter P of the form (67) satisfying condition (68) is identically zero, $P(\omega) = 0$. In this case, the bivariate penalty reduces to the ℓ_1 norm and the proposed BISR approach offers no improvement.

In both situations, a higher-order filter P is needed to accurately approximate H, and a higher-order non-separable penalty is needed to strongly induce sparsity. Hence, to extend the applicability of non-separable sparse regularization, it will be necessary to generalize the bivariate penalty to non-separable K-variate penalties ($K > 2$). Provided such an extension can be constructed, we expect the proposed BISR approach will be applicable to more general problems.

IX. CONCLUSION

This paper aims to develop a convex approach for sparse signal estimation that improves upon ℓ_1 norm regularization, the standard convex approach. We consider both well-conditioned (Example 1) and ill-conditioned (Example 2) linear inverse problems with a quadratic data fidelity term. We focus in particular on the deconvolution problem. The proposed method is based on a non-convex penalty designed so that the objective function is convex. Our previous work [69] using this idea considered only separable (additive) penalties; this is a fundamental limitation when the observation matrix is singular or near singular.

The proposed bivariate sparse regularization (BISR) approach provides a mechanism by which to improve upon ℓ_1 norm regularization while adhering to a convex framework. The greater generality of non-separable regularization, as compared with separable regularization, allows for the design of regularizers that more effectively induce sparsity while maintaining convexity of the objective function. Both BISR and ℓ_1 norm regularization lead to convex optimization problems which can be solved by similar optimization techniques.

The experimental section demonstrated the improvement of BISR in comparison with ℓ_1 norm regularization. However, methods based on minimizing non-convex objective functions (e.g., IPS) outperformed BISR. Hence, to improve upon BISR, it will be of interest as future work to 1) develop multivariate generalizations of the bivariate penalty, and 2) develop methods that combine BISR with non-convex algorithms.

While we have focused on deconvolution, we note that several filtering methods can be formulated as sparse deconvolution problems, e.g., peak detection [59] and denoising [68]. Hence, the proposed approach (and extensions thereof) is not limited to deconvolution.

ACKNOWLEDGMENT

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

APPENDIX

A. Proof of Lemma 1

In order to show that S in Definition 2 is twice continuously differentiable, we only need to verify that S, its gradient, and its Hessian are continuous on the boundaries of A_i. This is because S is defined as a linear combination of twice continuously differentiable concave functions in the interior of each A_i. We will use the symmetries (20). The symmetries imply that S on $A_1 \cup A_4$ determine S on $\mathbb{R}^2$.

To show S is continuous, we show Definition 2 is consistent on the boundaries of A_i. The boundary of A_1 and A_4 is the line $\{(x_1, 0) : x_1 \in \mathbb{R}\}$. Setting $x_2 = 0$ in Definition 2 gives

$$\begin{aligned} S((x_1,0);a) &= s(x_1;\alpha) + (1-r)s(0;a_1), \quad x \in A_1 \qquad (75)\\ &= s(x_1;\alpha) + (1+r)s(0;a_2), \quad x \in A_4. \qquad (76)\end{aligned}$$

Both expressions reduce to $s(x_1;\alpha)$ because $s(0,c) = 0$ for any c. Hence, S is continuous on this boundary. By symmetry of S, it follows that S is also continuous on the boundary of A_2 and A_3, i.e., the line $\{(0, x_2) : x_2 \in \mathbb{R}\}$.

Consider now the boundary of A_1 and A_2, i.e., the line $\{(x_1, x_1) : x_1 \in \mathbb{R}\}$. Setting $x_2 = x_1$ in Definition 2 gives

$$S((x_1,x_1);a) = s((1+r)x_1;\alpha) + (1-r)s(x_1;a_1) \qquad (77)$$

for both $x \in A_1$ and $x \in A_2$. Hence, S is continuous on this boundary, and also on the boundary of A_3 and A_4 by symmetry. Therefore, S is continuous on $\mathbb{R}^2$.

We now consider continuity of the gradient of S. We note that for $x = (x_1, x_2) \in A_1 \cup A_4$, the gradient is given by

$$\nabla S(x;a) = \begin{bmatrix} s'(x_1 + r x_2; \alpha) \\ r s'(x_1 + r x_2; \alpha) + (1-r) s'(x_2; a_1) \end{bmatrix}, \quad x \in A_1, \tag{78a}$$

$$\begin{bmatrix} s'(x_1 + r x_2; \alpha) \\ r s'(x_1 + r x_2; \alpha) + (1+r) s'(x_2; a_2) \end{bmatrix}, \quad x \in A_4. \tag{78b}$$

To verify ∇S is continuous on the boundary between A_1 and A_4, we set $x_2 = 0$ in these expressions. Since $s'(0;c) = 0$ for any c, we find that the expressions coincide:

$$\nabla S(x;a) = s'(x_1;\alpha) \begin{bmatrix} 1 \\ r \end{bmatrix}, \quad x_2 = 0. \tag{79}$$

Therefore ∇S is continuous on the interior of $A_1 \cup A_4$. By symmetry, it follows that ∇S is continuous on the interior of $A_2 \cup A_3$ also.

It remains to verify continuity of ∇S on the boundary between $A_1 \cup A_4$ and $A_2 \cup A_3$. For $x_1 = x_2$, the expression in (78a) reduces to

$$\nabla S(x;a) = \begin{bmatrix} s'\big((1+r)x_1;\alpha\big) \\ r s'\big((1+r)x_1;\alpha\big) + (1-r) s'(x_1;a_1) \end{bmatrix} \tag{80}$$

$$= s'\big((1+r)x_1;\alpha\big) \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad x_1 = x_2 \tag{81}$$

where we have used identities (11) and (21). For $x_2 = -x_1$, the expression in (78b) reduces to

$$\nabla S(x;a) = \begin{bmatrix} s'\big((1-r)x_1;\alpha\big) \\ r s'\big((1-r)x_1;\alpha\big) - (1+r) s'(x_1;a_2) \end{bmatrix} \tag{82}$$

$$= s'\big((1-r)x_1;\alpha\big) \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \quad x_1 = -x_2 \tag{83}$$

where we have used identities (11) and (22). The gradients point in the directions $[1, \ 1]^\mathsf{T}$ and $[1, \ -1]^\mathsf{T}$ respectively. Since S satisfies the symmetries (20), it follows that ∇S is continuous.

We now consider continuity of the Hessian of S. For $x \in A_1 \cup A_4$, the Hessian is given by

$$\nabla^2 S(x;a) = $$

$$s''(x_1 + r x_2;\alpha) \begin{bmatrix} 1 & r \\ r & r^2 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & (1-r)s''(x_2;a_1) \end{bmatrix}, x \in A_1, \tag{84a}$$

$$s''(x_1 + r x_2;\alpha) \begin{bmatrix} 1 & r \\ r & r^2 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & (1+r)s''(x_2;a_2) \end{bmatrix}, x \in A_4. \tag{84b}$$

On the boundary between A_1 and A_4 (i.e., the line $x_2 = 0$), we find that both expressions reduce to

$$\nabla^2 S(x;a) = s''(x_1;\alpha) \begin{bmatrix} 1 & r \\ r & r^2 \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ 0 & a_1 a_2/\alpha \end{bmatrix}, \ x_2 = 0,$$

where we have used the equalities

$$(1-r)s''(0;a_1) = -(1-r)a_1 = -a_1 a_2/\alpha,$$

$$(1+r)s''(0;a_2) = -(1+r)a_2 = -a_1 a_2/\alpha.$$

Therefore $\nabla^2 S$ is continuous on the interior of $A_1 \cup A_4$. By symmetry, it follows that $\nabla^2 S$ is continuous on the interior of $A_2 \cup A_3$ also.

We now consider the boundary between $A_1 \cup A_4$ and $A_2 \cup A_3$. For $x_1 = x_2$, the expression in (84a) reduces to

$$\nabla^2 S(x;a) = s''((1+r)x_1;\alpha) \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}, \text{ for } x_1 = x_2$$

where we have used identities (12), (21), (22). For $x_2 = -x_1$, the expression in (84b) likewise reduces to

$$\nabla^2 S(x;a) = s''((1-r)x_1;\alpha) \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}, \text{ for } x_1 = -x_2.$$

Both matrices have eigenvectors in the directions $[1, \ 1]^\mathsf{T}$ and $[1, \ -1]^\mathsf{T}$. Since S satisfies the symmetries (20), it follows that $\nabla^2 S$ is continuous on the boundary between A_1 and A_2 and on the boundary between A_3 and A_4. Hence, $\nabla^2 S$ is continuous on $\mathbb{R}^2$.

We now show that S is concave. First note in Definition 2 that S is defined on each A_i as a sum of concave functions (the univariate function s is concave). Since S is twice differentiable, this is equivalent to $\nabla^2 S(x) \preccurlyeq 0$ for x in the interior of each A_i. Since S is twice continuously differentiable, it furthermore follows by continuity that $\nabla S^2(x) \preccurlyeq 0$ everywhere. Therefore S is concave on $\mathbb{R}^2$.

B. Proof of Lemma 2

Using (12), (21) and (22), the Hessian of S is given by

$$\nabla^2 S(x;a) = \tag{85}$$

$$\begin{cases} s''(x_1 + r x_2;\alpha) \left[\begin{smallmatrix} 1 & r \\ r & r^2 \end{smallmatrix}\right] + s''((1+r)x_2;\alpha) \left[\begin{smallmatrix} 0 & 0 \\ 0 & 1-r^2 \end{smallmatrix}\right], & x \in A_1 \\ s''(r x_1 + x_2;\alpha) \left[\begin{smallmatrix} r^2 & r \\ r & 1 \end{smallmatrix}\right] + s''((1+r)x_1;\alpha) \left[\begin{smallmatrix} 1-r^2 & 0 \\ 0 & 0 \end{smallmatrix}\right], & x \in A_2 \\ s''(r x_1 + x_2;\alpha) \left[\begin{smallmatrix} r^2 & r \\ r & 1 \end{smallmatrix}\right] + s''((1-r)x_1;\alpha) \left[\begin{smallmatrix} 1-r^2 & 0 \\ 0 & 0 \end{smallmatrix}\right], & x \in A_3 \\ s''(x_1 + r x_2;\alpha) \left[\begin{smallmatrix} 1 & r \\ r & r^2 \end{smallmatrix}\right] + s''((1-r)x_2;\alpha) \left[\begin{smallmatrix} 0 & 0 \\ 0 & 1-r^2 \end{smallmatrix}\right], & x \in A_4. \end{cases}$$

Note that the matrices in (85) are positive semidefinite because $|r| \leqslant 1$. We also have $-\alpha \leqslant s''(t;\alpha) < 0$ for all t. Hence,

$$\nabla^2 S(x;a) \succcurlyeq -\alpha \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}, \quad x \in \mathbb{R}^2. \tag{86}$$

Using (15) and (22), we write (86) as

$$\nabla^2 S(x;a) \succcurlyeq -\frac{1}{2} \begin{bmatrix} a_1 + a_2 & a_1 - a_2 \\ a_1 - a_2 & a_1 + a_2 \end{bmatrix}. \tag{87}$$

Since $s''(0;\alpha) = -\alpha$ [see (8)], we similarly have

$$\nabla^2 S(0;a) = -\frac{1}{2} \begin{bmatrix} a_1 + a_2 & a_1 - a_2 \\ a_1 - a_2 & a_1 + a_2 \end{bmatrix}. \tag{88}$$

C. Proof of Theorem 3

We prove the claim for $x \in A_1$ and then for other A_i by a symmetry argument. We first consider the case $a_1 \geqslant a_2$. From (31) and (6), the inequality (35) can be written as

$$s(x_1;a_1)+s(x_2;a_1) \leqslant S(x;a) \leqslant s(x_1;a_2)+s(x_2;a_2) \quad (89)$$

where S is given by Definition 2. Since $a_1 \geqslant a_2$, we have $a_1 \geqslant \alpha \geqslant a_2 \geqslant 0$ and $0 \leqslant r \leqslant 1$. We seek to prove the lower bound of (89), given by

$$s(x_1+rx_2;\alpha)+(1-r)\,s(x_2;a_1) \geqslant s(x_1;a_1)+s(x_2;a_1) \quad (90)$$

for $x \in A_1$. Using (10) and (21), we have

$$\begin{aligned} s(x_1+rx_2;\alpha) &= (1+r)\,s\Big(\frac{x_1+rx_2}{1+r};a_1\Big) \\ &\geqslant (1+r)\left[\frac{1}{1+r}s(x_1;a_1)+\frac{r}{1+r}s(x_2;a_1)\right] \\ &= s(x_1;a_1)+r\,s(x_2;a_1) \end{aligned} \quad (91)$$

where the inequality is due to s being a concave function. Adding $(1-r)\,s(x_2;a_1)$ to both sides of (91) gives (90).

We now seek to prove the upper bound of (89), given by

$$s(x_1;a_2)+s(x_2;a_2) \geqslant s(x_1+rx_2;\alpha)+(1-r)\,s(x_2;a_1) \quad (92)$$

for $x \in A_1$. Using (10) and (21), we have

$$s(x_2;a_1) = \frac{1}{1+r}\,s((1+r)x_2;\alpha). \quad (93)$$

Suppose $0 \leqslant x_2 \leqslant x_1$. As s is concave and $s(0;\cdot)=0$,

$$s(x_1;\alpha) \geqslant \frac{x_1}{x_1+rx_2}\,s(x_1+rx_2;\alpha) \quad (94)$$

because $0 \leqslant x_1 \leqslant x_1+rx_2$. Similarly,

$$s((1+r)x_2;\alpha) \geqslant \frac{(1+r)x_2}{x_1+rx_2}\,s(x_1+rx_2;\alpha) \quad (95)$$

as $0 \leqslant (1+r)x_2 \leqslant x_1+rx_2$. From (93) and (95),

$$rs(x_2;a_1) \geqslant \frac{rx_2}{x_1+rx_2}\,s(x_1+rx_2;\alpha). \quad (96)$$

Adding (94) and (96) gives

$$s(x_1;\alpha)+rs(x_2;a_1) \geqslant s(x_1+rx_2;\alpha). \quad (97)$$

Adding $(1-r)\,s(x_2;a_1)$ to both sides of (97) gives

$$s(x_1;\alpha)+s(x_2;a_1) \geqslant s(x_1+rx_2;\alpha)+(1-r)\,s(x_2;a_1). \quad (98)$$

Since $\alpha \geqslant a_2$, we have $s(x_1;a_2) \geqslant s(x_1;\alpha)$. Similarly, since $a_1 \geqslant a_2$, we have $s(x_2;a_2) \geqslant s(x_2;a_1)$. It follows that (92) holds. It also holds for $x_1 \leqslant x_2 \leqslant 0$ as s is an even function.

We now consider the case $a_1 \leqslant a_2$ and $x \in A_1$. We now have $a_2 \geqslant \alpha \geqslant a_1 \geqslant 0$ and $-1 \leqslant r \leqslant 0$. We seek to prove the upper bound, given by

$$s(x_1+rx_2;\alpha)+(1-r)\,s(x_2;a_1) \leqslant s(x_1;a_1)+s(x_2;a_1) \quad (99)$$

for $x \in A_1$. Since $-1 \leqslant r \leqslant 0$, we have

$$\begin{aligned} s(x_1+rx_2;\alpha) &= (1+r)\,s\Big(\frac{x_1+rx_2}{1+r};a_1\Big) \\ &\leqslant (1+r)\left[\frac{1}{1+r}s(x_1;a_1)+\frac{r}{1+r}s(x_2;a_1)\right] \\ &= s(x_1;a_1)+r\,s(x_2;a_1). \end{aligned} \quad (100)$$

Adding $(1-r)\,s(x_2;a_1)$ to both sides of (100) gives (99).

We now seek to prove the lower bound, given by

$$s(x_1;a_2)+s(x_2;a_2) \leqslant s(x_1+rx_2;\alpha)+(1-r)\,s(x_2;a_1) \quad (101)$$

for $x \in A_1$. Using (10) and (21), we have

$$s(x_2;a_1) = \frac{1}{1+r}\,s((1+r)x_2;\alpha). \quad (102)$$

Suppose $0 \leqslant x_2 \leqslant x_1$. As s is concave, $-1 \leqslant r \leqslant 0$, and $s(0;\cdot)=0$, we have

$$s(x_1;\alpha) \leqslant \frac{x_1}{x_1+rx_2}\,s(x_1+rx_2;\alpha) \quad (103)$$

because $0 \leqslant x_1+rx_2 \leqslant x_1$. Similarly,

$$s((1+r)x_2;\alpha) \geqslant \frac{(1+r)x_2}{x_1+rx_2}\,s(x_1+rx_2;\alpha) \quad (104)$$

as $0 \leqslant (1+r)x_2 \leqslant x_1+rx_2$. From (102) and (104) and $-1 \leqslant r \leqslant 0$, we obtain (reversing the sign of the inequality in (104) due to multiplication with r),

$$rs(x_2;a_1) \leqslant \frac{rx_2}{x_1+rx_2}\,s(x_1+rx_2;\alpha). \quad (105)$$

Adding (103) and (105) gives

$$s(x_1;\alpha)+rs(x_2;a_1) \leqslant s(x_1+rx_2;\alpha). \quad (106)$$

Adding $(1-r)\,s(x_2;a_1)$ to both sides of (106) gives

$$s(x_1;\alpha)+s(x_2;a_1) \leqslant s(x_1+rx_2;\alpha)+(1-r)\,s(x_2;a_1). \quad (107)$$

Since $\alpha \leqslant a_2$, we have $s(x_1;a_2) \leqslant s(x_1;\alpha)$. Similarly, since $a_1 \leqslant a_2$, we have $s(x_2;a_2) \leqslant s(x_2;a_1)$. It follows that (101) holds. It also holds for $x_1 \leqslant x_2 \leqslant 0$ as s is an even function.

Let us now consider the other A_i. The result for A_4 deduces from the result for A_1 since $S((x_1,x_2);(a_1,a_2)) = S((x_1,-x_2);(a_2,a_1))$. Given that the claim is established for $A_1 \cup A_4$, the claim for $A_2 \cup A_3$ follows because $A_2 \cup A_3$ is obtained by reflecting $A_1 \cup A_4$ with respect to the line $x_1 = x_2$ and the functions appearing in (89), namely $s(x_1;a_1)+s(x_2;a_1)$, $S(x;a)$, $s(x_1;a_2)+s(x_2;a_2)$, are symmetric with respect to same the line $x_2 = x_1$.

D. Proof of Lemma 4

We express F in (39) as

$$F(x) = \frac{1}{2}y^\mathsf{T}y - y^\mathsf{T}Hx + \frac{1}{2}x^\mathsf{T}\left(H^\mathsf{T}H-P\right)x + g(x) \quad (108)$$

where $g\colon \mathbb{R}^N \to \mathbb{R}$ is defined as

$$g(x) = \frac{1}{2}x^\mathsf{T}Px + \frac{\lambda}{2}\sum_n \psi((x_{n-1},x_n);a). \quad (109)$$

The first three terms of (108) are convex. Hence, F is convex if g is convex. The function g may be written as

$$g(x) = \frac{1}{2}\sum_n \Big\{ \big(x_{n-1},\, x_n\big) \begin{bmatrix} \frac{1}{2}p_0 & p_1 \\ p_1 & \frac{1}{2}p_0 \end{bmatrix} \begin{pmatrix} x_{n-1} \\ x_n \end{pmatrix} + \lambda\,\psi((x_{n-1},x_n);a) \Big\} \quad (110)$$

$$= \frac{1}{2} \sum_n f((x_{n-1}, x_n)) \tag{111}$$

where $f\colon \mathbb{R}^2 \to \mathbb{R}$ is given by (41). If f is convex, then g is the sum of convex functions and is thus convex itself.

REFERENCES

[1] A. Achim and E. E. Kuruoğlu. Image denoising using bivariate α-stable distributions in the complex wavelet domain. *IEEE Signal Processing Letters*, 12(1):17–20, January 2005.

[2] M. V. Afonso, J. M. Bioucas-Dias, and M. A. T. Figueiredo. Fast image recovery using variable splitting and constrained optimization. *IEEE Trans. Image Process.*, 19(9):2345–2356, September 2010.

[3] M. S. Asif and J. Romberg. Fast and accurate algorithms for re-weighted l_1-norm minimization. *IEEE Trans. Signal Process.*, 61(23):5905–5916, December 2013.

[4] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends in Machine Learning*, 4(1):1–106, 2012.

[5] İ. Bayram. On the convergence of the iterative shrinkage/thresholding algorithm with a weakly convex penalty. http://arxiv.org/abs/1510.07821, October 2015.

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

[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] A. Blumensath. Accelerated iterative hard thresholding. *Signal Processing*, 92(3):752–756, 2012.

[11] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[12] 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.

[13] 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.

[14] 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.

[15] R. Chartrand. Fast algorithms for nonconvex compressive sensing: MRI reconstruction from very few data. In *IEEE Int. Symp. Biomed. Imag. (ISBI)*, pages 262–265, July 2009.

[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] R. Chartrand, E. Y. Sidky, and P. Xiaochuan. Nonconvex compressive sensing for X-ray CT: an algorithm comparison. In *Asilomar Conf. on Signals, Systems and Computers*, pages 665–669, November 2013.

[19] 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.

[20] 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.

[21] H. Chipman, E. Kolaczyk, and R. McCulloch. Adaptive Bayesian wavelet shrinkage. *J. Am. Stat. Assoc.*, 92(440):1413–1421, December 1997.

[22] 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.

[23] 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.

[24] P. L. Combettes and J.-C. Pesquet. Proximal thresholding algorithm for minimization over orthonormal bases. *SIAM J. Optim.*, 18(4):1351–1376, 2008.

[25] 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.

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

[27] 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.

[28] I. Daubechies, R. DeVore, M. Fornasier, and C. Gunturk. Iteratively reweighted least squares minimization for sparse recovery. *Comm. Pure App. Math.*, 63(1):1–38, January 2010.

[29] 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.

[30] B. Dumitrescu. *Positive trigonometric polynomials and signal processing applications*. Springer, 2007.

[31] J. M. Fadili and L. Boubchir. Analytical form for a Bayesian wavelet estimator of images using the Bessel K form densities. *IEEE Trans. Image Process.*, 14(2):231–240, February 2005.

[32] M. Figueiredo, J. Bioucas-Dias, and R. Nowak. Majorization-minimization algorithms for wavelet-based image restoration. *IEEE Trans. Image Process.*, 16(12):2980–2991, December 2007.

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

[34] 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.

[35] S. Foucart. Hard thresholding pursuit: an algorithm for compressive sensing. *SIAM J. Numer. Anal.*, 49(6):2543–2563, 2010.

[36] 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.

[37] J. J. Fuchs. Identification of real sinusoids in noise, the global matched filter approach. In *15th IFAC Symp. on System Identification*, pages 1127–1132, Saint-Malo, France, July 2009.

[38] 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.

[39] 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.

[40] 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.

[41] T. Goldstein and S. Osher. The split Bregman method for L1-regularized problems. *SIAM J. Imag. Sci.*, 2(2):323–343, 2009.

[42] 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.

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

[45] K. Lange, E. C. Chi, and H. Zhou. A brief survey of modern optimization for statisticians. *Int. Stat. Rev.*, 82(1):46–70, 2014.

[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] P.-L. Loh and M. J. Wainwright. Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *J. Machine Learning Research*, 16:559–616, 2015.

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

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

[51] 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.

[52] 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.

[53] 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.

[54] N. Mourad and J. P. Reilly. Minimizing nonconvex functions for sparse vector reconstruction. *IEEE Trans. Signal Process.*, 58(7):3485–3496, July 2010.

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

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

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

[58] 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.

[59] X. Ning, I. W. Selesnick, and L. Duval. Chromatogram baseline estimation and denoising using sparsity (BEADS). *Chemometrics and Intelligent Laboratory Systems*, 139:156–167, December 2014.

[60] D. P. Palomar and Y. C. Eldar, editors. *Convex Optimization in Signal Processing and Communications*. Cambridge University Press, 2010.

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

[62] A. Parekh and I. W. Selesnick. Convex fused lasso denoising with non-convex regularization and its use for pulse detection. http://arxiv.org/abs/1509.02811, October 2015.

[63] 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.

[64] J. Portilla, V. Strela, M. J. Wainwright, and E. P. Simoncelli. Image denoising using scale mixtures of Gaussians in the wavelet domain. *IEEE Trans. Image Process.*, 12(11):1338–1351, November 2003.

[65] K. Qiu and A. Dogandzic. Sparse signal reconstruction via ECME hard thresholding. *IEEE Trans. Signal Process.*, 60(9):4551–4569, September 2012.

[66] B. D. Rao, K. Engan, S. F. Cotter, J. Palmer, and K. Kreutz-Delgado. Subset selection in noise based on diversity measure minimization. *IEEE Trans. Signal Process.*, 51(3):760–770, March 2003.

[67] 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.

[68] I. W. Selesnick. Sparsity-assisted signal smoothing. In R. Balan et al., editors, *Excursions in Harmonic Analysis, Volume 4*. Birkhäuser Basel, 2015.

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

[70] 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.

[71] L. Sendur and I. W. Selesnick. Bivariate shrinkage functions for wavelet-based denoising exploiting interscale dependency. *IEEE Trans. Signal Process.*, 50(11):2744–2756, November 2002.

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

[73] 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.

[74] 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.

[75] X. Tan, W. Roberts, J. Li, and P. Stoica. Sparse learning via iterative minimization with application to MIMO radar imaging. *IEEE Trans. Signal Process.*, 59(3):1088–1101, March 2011.

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

[77] J. Trzasko and A. Manduca. Highly undersampled magnetic resonance image reconstruction via homotopic L0-minimization. *IEEE Trans. Medical Imaging*, 28(1):106–121, January 2009.

[78] 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.

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

[80] D. Wipf and S. Nagarajan. Iterative reweighted ℓ_1 and ℓ_2 methods for finding sparse solutions. *IEEE. J. Sel. Top. Signal Processing*, 4(2):317–329, April 2010.

[81] 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.

[82] J. Woodworth and R. Chartrand. Compressed sensing recovery via nonconvex shrinkage penalties. http://arxiv.org/abs/1504.02923, April 2015.

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