Sparse Regularization via Convex Analysis

Ivan Selesnick, *Fellow, IEEE*

***Abstract*—Sparse approximate solutions to linear equations are classically obtained via L1 norm regularized least squares, but this method often underestimates the true solution. As an alternative to the L1 norm, this paper proposes a class of nonconvex penalty functions that maintain the convexity of the least squares cost function to be minimized, and avoids the systematic underestimation characteristic of L1 norm regularization. The proposed penalty function is a multivariate generalization of the minimax-concave penalty. It is defined in terms of a new multivariate generalization of the Huber function, which in turn is defined via infimal convolution. The proposed sparse-regularized least squares cost function can be minimized by proximal algorithms comprising simple computations.**

***Index Terms*—Sparse regularization, sparse approximation, convex function, optimization, denoising.**

I. Introduction

NUMEROUS signal and image processing techniques build upon sparse approximation [59]. A sparse approximate solution to a system of linear equations ($y = Ax$) can often be obtained via convex optimization. The usual technique is to minimize the regularized linear least squares cost function $J : \mathbb{R}^N \to \mathbb{R}$,

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

The ℓ_1 norm is classically used as a regularizer here, since among convex regularizers it induces sparsity most effectively [9]. But this formulation tends to underestimate high-amplitude components of $x \in \mathbb{R}^N$. Non-convex sparsity-inducing regularizers are also widely used (leading to more accurate estimation of high-amplitude components), but then the cost function is generally non-convex and has extraneous suboptimal local minimizers [43].

This paper proposes a class of non-convex penalties for sparse-regularized linear least squares that generalizes the ℓ_1 norm and maintains the convexity of the least squares cost function to be minimized. That is, we consider the cost function

$F : \mathbb{R}^N \to \mathbb{R}$

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

and we propose a new non-convex penalty $\psi_B : \mathbb{R}^N \to \mathbb{R}$ that makes F convex. The penalty ψ_B is parameterized by a matrix B, and the convexity of F depends on B being suitably prescribed. In fact, the choice of B will depend on A.

The matrix (linear operator) A may be arbitrary (i.e., injective, surjective, both, or neither). In contrast to the ℓ_1 norm, the new approach does not systematically underestimate high-amplitude components of sparse vectors. Since the proposed formulation is convex, the cost function has no suboptimal local minimizers.

The new class of non-convex penalties is defined using tools of convex analysis. In particular, *infimal convolution* is used to define a new multivariate generalization of the Huber function. In turn, the generalized Huber function is used to define the proposed non-convex penalty, which can be considered a multivariate generalization of the minimax-concave (MC) penalty. Even though the generalized MC (GMC) penalty is non-convex, it is easy to prescribe this penalty so as to maintain the convexity of the cost function to be minimized.

The proposed convex cost functions can be minimized using proximal algorithms, comprising simple computations. In particular, the minimization problem can be cast as a kind of saddle-point problem for which the forward-backward splitting algorithm is applicable. The main computational steps of the algorithm are the operators A, A^{T}, and soft thresholding. The implementation is thus 'matrix-free' in that it involves the operators A and A^{T}, but does not access or modify the entries of A. Hence, the algorithm can leverage efficient implementations of A and its transpose.

We remark that while the proposed GMC penalty is non-separable, we do not advocate non-separability in and of itself as a desirable property of a sparsity-inducing penalty. But in many cases (depending on A), non-separability is simply a requirement of a non-convex penalty designed so as to maintain convexity of the cost function F to be minimized. If $A^{\mathsf{T}}A$ is singular (and none of its eigenvectors are standard basis vectors), then a separable penalty that maintains the convexity of the cost function F must, in fact, be a convex penalty [55]. This leads us back to the ℓ_1 norm. Thus, to improve upon the ℓ_1 norm, the penalty must be non-separable.

This paper is organized as follows. Section II sets notation and recalls definitions of convex analysis. Section III recalls the (scalar) Huber function, the (scalar) MC penalty, and how they arise in the formulation of threshold functions (instances of proximity operators). The subsequent sections generalize these

Manuscript received February 16, 2017; revised May 13, 2017; accepted May 18, 2017. Date of publication June 2, 2017; date of current version June 28, 2017. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Laura Cottatellucci. This work was supported by National Science Foundation under Grant CCF-1525398 and ONR under Grant N00014-15-1-2314.

The author is with the Department of Electrical and Computer Engineering, Tandon School of Engineering, New York University, Brooklyn, NY 10003 USA (e-mail: selesi@nyu.edu).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org., provided by the author. The material includes computer software (Matlab and Python) implementing the method. Contact the author for further questions about this work.

Digital Object Identifier 10.1109/TSP.2017.2711501

concepts to the multivariate case. In Section IV, we define a multivariate version of the Huber function. In Section V, we define a multivariate version of the MC penalty. In Section VI, we show how to set the GMC penalty to maintain convexity of the least squares cost function. Section VII presents a proximal algorithm to minimize this type of cost function. Section VIII presents examples wherein the GMC penalty is used for signal denoising and approximation.

Elements of this work were presented in Ref. [49].

A. Related Work

Many prior works have proposed non-convex penalties that strongly promote sparsity or describe algorithms for solving the sparse-regularized linear least squares problem, e.g., [11], [13], [15], [16], [19], [25], [30], [31], [38], [39], [43], [47], [58], [64], [66]. However, most of these papers (*i*) use separable (additive) penalties or (*ii*) do not seek to maintain convexity of the cost function. Non-separable non-convex penalties are proposed in Refs. [60], [63], but they are not designed to maintain cost function convexity. The development of convexity-preserving non-convex penalties was pioneered by Blake, Zisserman, and Nikolova [7], [41]–[44], and further developed in [6], [17], [23], [32], [36], [37], [45], [54], [56]. But these are separable penalties, and as such they are fundamentally limited. Specifically, if $A^{\mathsf{T}}A$ is singular, then a separable penalty constrained to maintain cost function convexity can only improve on the ℓ_1 norm to a very limited extent [55]. Non-convex regularization that maintains cost function convexity was used in [35] in an iterative manner for non-convex optimization, to reduce the likelihood that an algorithm converges to suboptimal local minima.

To overcome the fundamental limitation of separable non-convex penalties, we proposed a bivariate non-separable non-convex penalty that maintains the convexity of the cost function to be minimized [55]. But that penalty is useful for only a narrow class of linear inverse problems. To handle more general problems, we subsequently proposed a multivariate penalty formed by subtracting from the ℓ_1 norm a function comprising the composition of a linear operator and a separable nonlinear function [52]. Technically, this type of multivariate penalty is non-separable, but it still constitutes a rather narrow class of non-separable functions.

Convex analysis tools (especially the Moreau envelope and the Fenchel conjugate) have recently been used in novel ways for sparse regularized least squares [12], [57]. Among other aims, these papers seek the convex envelope of the ℓ_0 pseudo-norm regularized least squares cost function, and derive alternate cost functions that share the same global minimizers but have fewer local minima. In these approaches, algorithms are less likely to converge to suboptimal local minima (the global minimizer might still be difficult to calculate).

For the special case where $A^{\mathsf{T}}A$ is diagonal, the proposed GMC penalty is closely related to the 'continuous exact ℓ_0' (CEL0) penalty introduced in [57]. In [57] it is observed that if $A^{\mathsf{T}}A$ is diagonal, then the global minimizers of the ℓ_0 regularized problem coincides with that of a convex function defined using the CEL0 penalty. Even though the diagonal case is simpler than the non-diagonal case and a convex cost function can be readily constructed with a non-convex penalty (e.g., [54]), the connection to the ℓ_0 problem is enlightening.

In other related work, we use convex analysis concepts (specifically, the Moreau envelope) for the problem of total variation (TV) denoising [51]. In particular, we prescribe a non-convex TV penalty that preserves the convexity of the TV denoising cost function to be minimized. The approach of Ref. [51] generalizes standard TV denoising so as to more accurately estimate jump discontinuities.

II. Notation

The ℓ_1, ℓ_2, and ℓ_∞ norms of $x \in \mathbb{R}^N$ are defined $\|x\|_1 = \sum_n |x_n|$, $\|x\|_2 = \left(\sum_n |x_n|^2\right)^{1/2}$, and $\|x\|_\infty = \max_n |x_n|$. If $A \in \mathbb{R}^{M \times N}$, then component n of Ax is denoted $[Ax]_n$. If the matrix $A - B$ is positive semidefinite, we write $B \preccurlyeq A$. The matrix 2-norm of matrix A is denoted $\|A\|_2$ and its value is the square root of the maximum eigenvalue of $A^{\mathsf{T}}A$. We have $\|Ax\|_2 \leqslant \|A\|_2 \|x\|_2$ for all $x \in \mathbb{R}^N$. If A has full row-rank (i.e., AA^{T} is invertible), then the pseudo-inverse of A is given by $A^+ := A^{\mathsf{T}}(AA^{\mathsf{T}})^{-1}$. We denote the transpose of the pseudo-inverse of A as $A^{+\mathsf{T}}$, i.e., $A^{+\mathsf{T}} := (A^+)^{\mathsf{T}}$. If A has full row-rank, then $A^{+\mathsf{T}} = (AA^{\mathsf{T}})^{-1}A$.

This work uses definitions and notation of convex analysis [4]. The infimal convolution of two functions f and g from $\mathbb{R}^N$ to $\mathbb{R} \cup \{+\infty\}$ is given by

$$(f \,\square\, g)(x) = \inf_{v \in \mathbb{R}^N} \big\{ f(v) + g(x - v) \big\}. \tag{3}$$

The Moreau envelope of the function $f : \mathbb{R}^N \to \mathbb{R}$ is given by

$$f^{\mathsf{M}}(x) = \inf_{v \in \mathbb{R}^N} \left\{ f(v) + \tfrac{1}{2}\|x - v\|_2^2 \right\}. \tag{4}$$

In the notation of infimal convolution, we have

$$f^{\mathsf{M}} = f \,\square\, \tfrac{1}{2}\|\cdot\|_2^2. \tag{5}$$

The set of proper lower semicontinuous (lsc) convex functions from $\mathbb{R}^N$ to $\mathbb{R} \cup \{+\infty\}$ is denoted $\Gamma_0(\mathbb{R}^N)$.

If the function f is defined as the composition $f(x) = h(g(x))$, then we write $f = h \circ g$.

The soft threshold function soft : $\mathbb{R} \to \mathbb{R}$ with threshold parameter $\lambda \geqslant 0$ is defined as

$$\operatorname{soft}(y; \lambda) := \begin{cases} 0, & |y| \leqslant \lambda \\ (|y| - \lambda)\operatorname{sign}(y), & |y| \geqslant \lambda. \end{cases} \tag{6}$$

III. Scalar Penalties

We recall the definition of the Huber function [33].

Definition 1: The Huber function $s : \mathbb{R} \to \mathbb{R}$ is defined as

$$s(x) := \begin{cases} \frac{1}{2}x^2, & |x| \leqslant 1 \\ |x| - \frac{1}{2}, & |x| \geqslant 1, \end{cases} \tag{7}$$

as illustrated in Fig. 1.

Proposition 1: The Huber function can be written as

$$s(x) = \min_{v \in \mathbb{R}} \left\{ |v| + \tfrac{1}{2}(x - v)^2 \right\}. \tag{8}$$

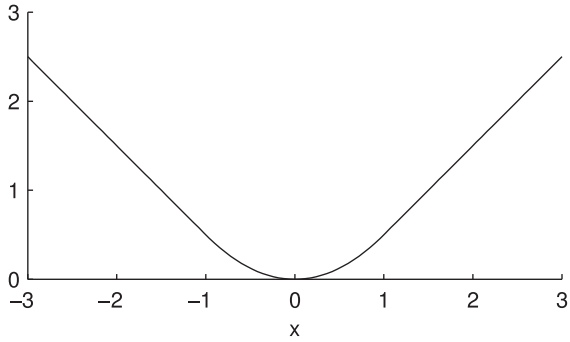

Fig. 1. The Huber function.

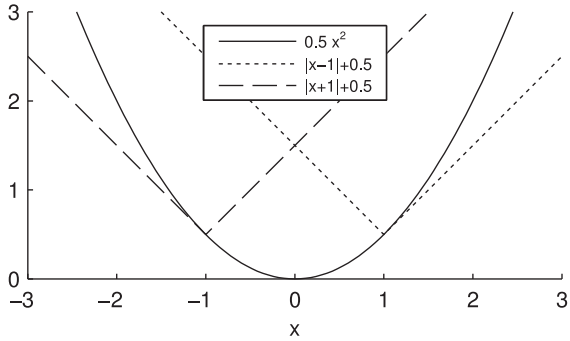

Fig. 2. The Huber function as the pointwise minimum of three functions.

In the notation of infimal convolution, we have equivalently

$$s = |\cdot| \,\square\, \tfrac{1}{2}(\cdot)^2. \tag{9}$$

And in the notation of the Moreau envelope, we have equivalently $s = |\cdot|^{\mathsf{M}}$.

The Huber function is a standard example of the Moreau envelope. For example, see Sec. 3.1 of Ref. [46] and [20]. We note here that, given $x \in \mathbb{R}$, the minimum in (8) is achieved for v equal to 0, $x-1$, or $x+1$, i.e.,

$$s(x) = \min_{v \in \{0,\, x-1,\, x+1\}} \left\{ |v| + \tfrac{1}{2}(x-v)^2 \right\}. \tag{10}$$

Consequently, the Huber function can be expressed as

$$s(x) = \min\left\{ \tfrac{1}{2}x^2,\ |x-1| + \tfrac{1}{2},\ |x+1| + \tfrac{1}{2} \right\} \tag{11}$$

as illustrated in Fig. 2.

We now consider the scalar penalty function illustrated in Fig. 3. This is the minimax-concave (MC) penalty [65]; see also [5], [6], [28].

Definition 2: The minimax-concave (MC) penalty function $\phi : \mathbb{R} \to \mathbb{R}$ is defined as

$$\phi(x) := \begin{cases} |x| - \tfrac{1}{2}x^2, & |x| \leqslant 1 \\ \tfrac{1}{2}, & |x| \geqslant 1, \end{cases} \tag{12}$$

as illustrated in Fig. 3.

The MC penalty can be expressed as

$$\phi(x) = |x| - s(x) \tag{13}$$

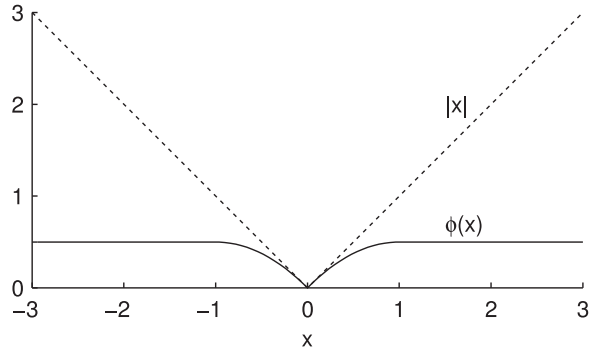

Fig. 3. The MC penalty function.

where s is the Huber function. This representation of the MC penalty will be used in Section V to generalize the MC penalty to the multivariate case.

A. Scaled functions

It will be convenient to define scaled versions of the Huber function and MC penalty.

Definition 3: Let $b \in \mathbb{R}$. The scaled Huber function $s_b : \mathbb{R} \to \mathbb{R}$ is defined as

$$s_b(x) := s(b^2 x)/b^2, \quad b \neq 0. \tag{14}$$

For $b = 0$, the function is defined as

$$s_0(x) := 0. \tag{15}$$

Hence, for $b \neq 0$, the scaled Huber function is given by

$$s_b(x) = \begin{cases} \tfrac{1}{2}b^2x^2, & |x| \leqslant 1/b^2 \\ |x| - \tfrac{1}{2b^2}, & |x| \geqslant 1/b^2. \end{cases} \tag{16}$$

The scaled Huber function s_b is shown in Fig. 4 for several values of the scaling parameter b. Note that

$$0 \leqslant s_b(x) \leqslant |x|, \quad \forall x \in \mathbb{R}, \tag{17}$$

and

$$\lim_{b \to \infty} s_b(x) = |x| \tag{18}$$

$$\lim_{b \to 0} s_b(x) = 0. \tag{19}$$

Incidentally, we use b^2 in definition (14) rather than b, so as to parallel the generalized Huber function to be defined in Sec. IV.

Proposition 2: Let $b \in \mathbb{R}$. The scaled Huber function can be written as

$$s_b(x) = \min_{v \in \mathbb{R}} \left\{ |v| + \tfrac{1}{2}b^2(x-v)^2 \right\}. \tag{20}$$

In terms of infimal convolution, we have equivalently

$$s_b = |\cdot| \,\square\, \tfrac{1}{2}b^2(\cdot)^2. \tag{21}$$

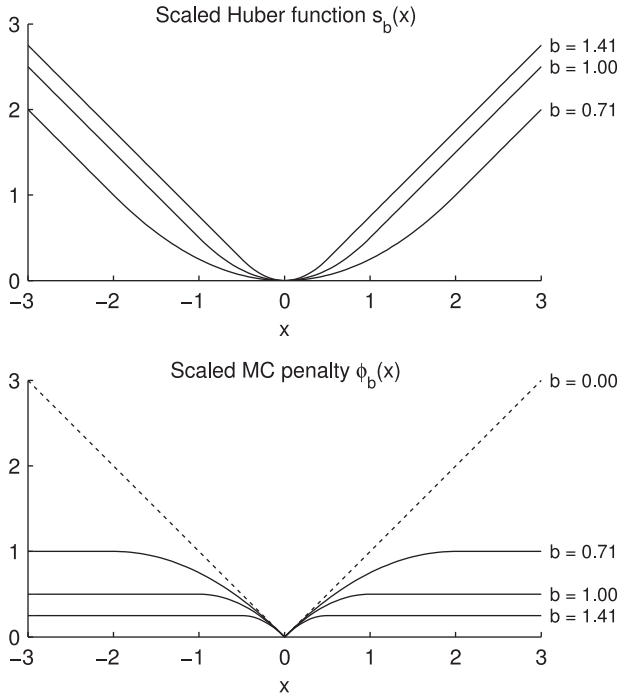

Fig. 4. Scaled Huber function and MC penalty for several values of the scaling parameter.

Proof: For $b \neq 0$, we have from (14) that

$$\begin{aligned} s_b(x) &= \min_{v\in\mathbb{R}} \left\{ |v| + \tfrac{1}{2}(b^2 x - v)^2 \right\} / b^2 \\ &= \min_{v\in\mathbb{R}} \left\{ |b^2 v| + \tfrac{1}{2}(b^2 x - b^2 v)^2 \right\} / b^2 \\ &= \min_{v\in\mathbb{R}} \left\{ |v| + \tfrac{1}{2} b^2 (x - v)^2 \right\}. \end{aligned}$$

It follows from $|\cdot| \square\, 0 = 0$ that (20) holds for $b = 0$. ■

Definition 4: Let $b \in \mathbb{R}$. The scaled MC penalty function $\phi_b : \mathbb{R} \to \mathbb{R}$ is defined as

$$\phi_b(x) := |x| - s_b(x) \tag{22}$$

where s_b is the scaled Huber function.

The scaled MC penalty ϕ_b is shown in Fig. 4 for several values of b. Note that $\phi_0(x) = |x|$. For $b \neq 0$,

$$\phi_b(x) = \begin{cases} |x| - \frac{1}{2} b^2 x^2, & |x| \leqslant 1/b^2 \\ \frac{1}{2b^2}, & |x| \geqslant 1/b^2. \end{cases} \tag{23}$$

B. Convexity Condition

In the scalar case, the MC penalty corresponds to a type of threshold function. Specifically, the *firm* threshold function is the *proximity operator* of the MC penalty, provided that a particular convexity condition is satisfied. Here, we give the convexity condition for the scalar case. We will generalize this condition to the multivariate case in Sec. VI.

Proposition 3: Let $\lambda > 0$ and $a \in \mathbb{R}$. Define $f : \mathbb{R} \to \mathbb{R}$,

$$f(x) = \frac{1}{2}(y - ax)^2 + \lambda\, \phi_b(x) \tag{24}$$

where ϕ_b is the scaled MC penalty (22). If

$$b^2 \leqslant a^2/\lambda, \tag{25}$$

then f is convex.

There are several ways to prove Proposition 3. In anticipation of the multivariate case, we use a technique in the following proof that we later use in the proof of Theorem 1 in Sec. VI.

Proof of Proposition 3: Using (22), we write f as

$$\begin{aligned} f(x) &= \frac{1}{2}(y - ax)^2 + \lambda|x| - \lambda s_b(x) \\ &= g(x) + \lambda|x| \end{aligned}$$

where s_b is the scaled Huber function and $g : \mathbb{R} \to \mathbb{R}$ is given by

$$g(x) = \frac{1}{2}(y - ax)^2 - \lambda s_b(x). \tag{26}$$

Since the sum of two convex functions is convex, it is sufficient to show g is convex. Using (20), we have

$$\begin{aligned} g(x) &= \frac{1}{2}(y - ax)^2 - \lambda \min_{v\in\mathbb{R}} \left\{ |v| + \tfrac{1}{2} b^2 (x - v)^2 \right\} \\ &= \max_{v\in\mathbb{R}} \left\{ \tfrac{1}{2}(y - ax)^2 - \lambda|v| - \tfrac{1}{2}\lambda b^2 (x - v)^2 \right\} \\ &= \tfrac{1}{2}(a^2 - \lambda b^2) x^2 \\ &\quad + \max_{v\in\mathbb{R}} \left\{ \tfrac{1}{2}(y^2 - 2axy) - \lambda|v| - \tfrac{1}{2}\lambda b^2 (v^2 - 2xv) \right\}. \end{aligned}$$

Note that the expression in the curly braces is affine (hence convex) in x. Since the pointwise maximum of a set of convex functions is itself convex, the second term is convex in x. Hence, g is convex if $a^2 - \lambda b^2 \geqslant 0$. ■

The firm threshold function was defined by Gao and Bruce [29] as a generalization of hard and soft thresholding.

Definition 5: Let $\lambda > 0$ and $\mu > \lambda$. The threshold function $\text{firm} : \mathbb{R} \to \mathbb{R}$ is defined as

$$\text{firm}(y; \lambda, \mu) := \begin{cases} 0, & |y| \leqslant \lambda \\ \mu(|y| - \lambda)/(\mu - \lambda)\, \text{sign}(y), & \lambda \leqslant |y| \leqslant \mu \\ y, & |y| \geqslant \mu \end{cases} \tag{27}$$

as illustrated in Fig. 5 .

In contrast to the soft threshold function, the firm threshold function does not underestimate large amplitude values, since it equals the identity for large values of its argument. As $\mu \to \lambda$ or $\mu \to \infty$, the firm threshold function approaches the hard or soft threshold function, respectively.

We now state the correspondence between the MC penalty and the firm threshold function. When f in (24) is convex (i.e., $b^2 \leqslant a^2/\lambda$), the minimizer of f is given by firm thresholding. This is noted in Refs. [5], [28], [64], [65].

Proposition 4: Let $\lambda > 0$, $a > 0$, $b > 0$, and $b^2 \leqslant a^2/\lambda$. Let $y \in \mathbb{R}$. Then the minimizer of f in (24) is given by firm thresholding, i.e.,

$$x^{\text{opt}} = \text{firm}(y/a; \lambda/a^2, 1/b^2). \tag{28}$$

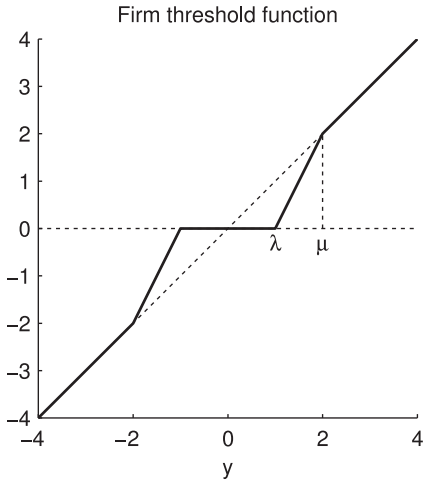

Fig. 5. Firm threshold function.

Hence, the minimizer of the scalar function f in (24) is easily obtained via firm thresholding. However, the situation in the multivariate case is more complicated. The aim of this paper is to generalize this process to the multivariate case: to define a multivariate MC penalty generalizing (12), to define a regularized least squares cost function generalizing (24), to generalize the convexity condition (25), and to provide a method to calculate a minimizer.

IV. Generalized Huber Function

In this section, we introduce a multivariate generalization of the Huber function. The basic idea is to generalize (20) which expresses the scalar Huber function as an infimal convolution.

Definition 6: Let $B \in \mathbb{R}^{M\times N}$. We define the generalized Huber function $S_B : \mathbb{R}^N \to \mathbb{R}$ as

$$S_B(x) := \inf_{v\in\mathbb{R}^N} \left\{ \|v\|_1 + \tfrac{1}{2}\|B(x-v)\|_2^2 \right\}. \tag{29}$$

In the notation of infimal convolution, we have

$$S_B = \|\cdot\|_1 \,\square\, \tfrac{1}{2}\|B\cdot\|_2^2. \tag{30}$$

Proposition 5: The generalized Huber function S_B is a proper lower semicontinuous convex function, and the infimal convolution is exact, i.e.,

$$S_B(x) = \min_{v\in\mathbb{R}^N} \left\{ \|v\|_1 + \tfrac{1}{2}\|B(x-v)\|_2^2 \right\}. \tag{31}$$

Proof: Set $f = \|\cdot\|_1$ and $g = \|B\cdot\|_2^2$. Both f and g are convex; hence $f \,\square\, g$ is convex by proposition 12.11 in [4]. Since f is coercive and g is bounded below, and $f, g \in \Gamma_0(\mathbb{R}^N)$, it follows that $f \,\square\, g \in \Gamma_0(\mathbb{R}^N)$ and the infimal convolution is exact (i.e., the infimum is achieved for some v) by Proposition 12.14 in [4]. ■

Note that if $C^{\mathsf{T}}C = B^{\mathsf{T}}B$, then $S_B(x) = S_C(x)$ for all x. That is, the generalized Huber function S_B depends only on $B^{\mathsf{T}}B$, not on B itself. Therefore, without loss of generality, we may assume B has full row-rank. (If a given matrix B does not have full row-rank, then there is another matrix C with full row-rank such that $C^{\mathsf{T}}C = B^{\mathsf{T}}B$, yielding the same function S_B.)

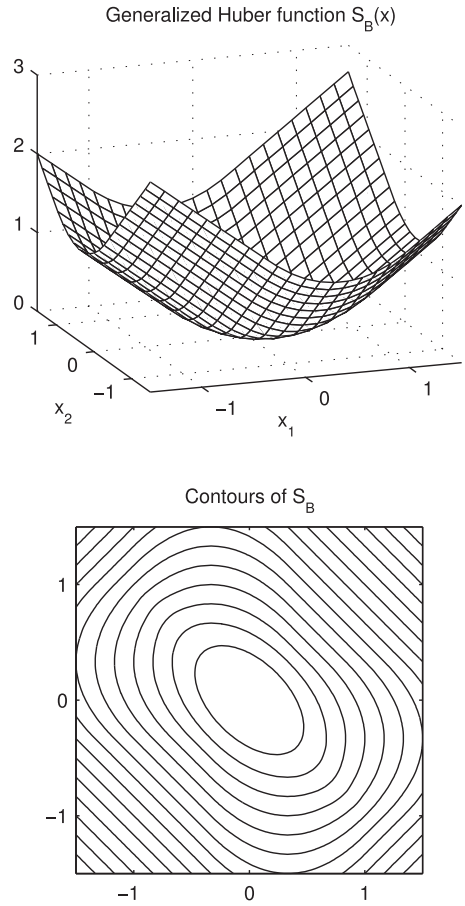

Fig. 6. The generalized Huber function for the matrix B in (32).

As expected, the generalized Huber function reduces to the scalar Huber function.

Proposition 6: If B is a scalar, i.e., $B = b \in \mathbb{R}$, then the generalized Huber function reduces to the scalar Huber function, $S_b(x) = s_b(x)$ for all $x \in \mathbb{R}$.

The generalized Huber function is separable (additive) when $B^{\mathsf{T}}B$ is diagonal.

Proposition 7: Let $B \in \mathbb{R}^{M\times N}$. If $B^{\mathsf{T}}B$ is diagonal, then the generalized Huber function is separable (additive), comprising a sum of scalar Huber functions. Specifically,

$$B^{\mathsf{T}}B = \operatorname{diag}(\alpha_1^2, \ldots, \alpha_N^2) \implies S_B(x) = \sum_n s_{\alpha_n}(x_n).$$

The utility of the generalized Huber function will be most apparent when $B^{\mathsf{T}}B$ is a non-diagonal matrix. In this case, the generalized Huber function is non-separable, as illustrated in the following two examples.

Example 1: For the matrix

$$B = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \end{bmatrix}, \tag{32}$$

the generalized Huber function S_B is shown in Fig. 6. As shown in the contour plot, the level sets of S_B near the origin are ellipses.

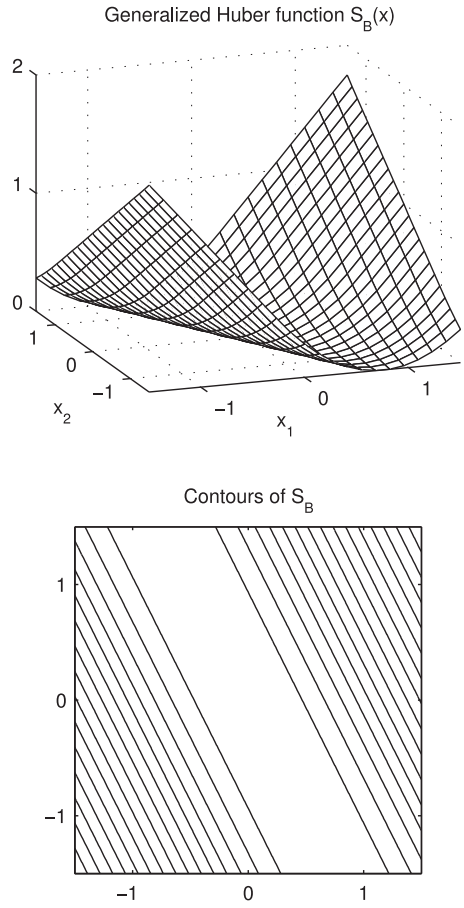

Fig. 7. The generalized Huber function for the matrix B in (33).

Example 2: For the matrix

$$B = [1 \;\; 0.5] \tag{33}$$

the generalized Huber function S_B is shown in Fig. 7 . The level sets of S_B are parallel lines because B is of rank 1.

There is not a simple explicit formula for the generalized Huber function. But, using (31) we can derive several properties regarding the function.

Proposition 8: Let $B \in \mathbb{R}^{M\times N}$. The generalized Huber function satisfies

$$0 \leqslant S_B(x) \leqslant \|x\|_1, \quad \forall x \in \mathbb{R}^N. \tag{34}$$

Proof: Using (31), we have

$$\begin{aligned} S_B(x) &= \min_{v\in\mathbb{R}^N} \left\{\|v\|_1 + \tfrac{1}{2}\|B(x-v)\|_2^2\right\} \\ &\leqslant \left[\|v\|_1 + \tfrac{1}{2}\|B(x-v)\|_2^2\right]_{v=x} \\ &= \|x\|_1. \end{aligned}$$

Since S_B is the minimum of a non-negative function, it also follows that $S_B(x) \geqslant 0$ for all x. ■

The following proposition accounts for the ellipses near the origin in the contour plot of the generalized Huber function in Fig. 6. (Further from the origin, the contours are not ellipsoidal.)

Proposition 9: Let $B \in \mathbb{R}^{M\times N}$. The generalized Huber function satisfies

$$S_B(x) = \tfrac{1}{2}\|Bx\|_2^2 \;\text{ if and only if }\; \|B^{\mathsf{T}}Bx\|_\infty \leqslant 1. \tag{35}$$

Proof: From (31), we have that $S_B(x)$ is the minimum value of g where $g : \mathbb{R}^N \to \mathbb{R}$ is given by

$$g(v) = \|v\|_1 + \tfrac{1}{2}\|B(x-v)\|_2^2.$$

Note that

$$g(0) = \tfrac{1}{2}\|Bx\|_2^2.$$

Hence, it suffices to show that 0 minimizes g if and only if $\|B^{\mathsf{T}}Bx\|_\infty \leqslant 1$. Since g is convex, 0 minimizes g if and only if $0 \in \partial g(0)$ where ∂g is the subdifferential of g given by

$$\partial g(v) = \operatorname{sign}(v) + B^{\mathsf{T}}B(v-x)$$

where sign is the set-valued signum function,

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

It follows that 0 minimizes g if and only if

$$\begin{aligned} & 0 \in \operatorname{sign}(0) - B^{\mathsf{T}}Bx \\ \Leftrightarrow \quad & B^{\mathsf{T}}Bx \in [-1,\ 1]^N \\ \Leftrightarrow \quad & \left[B^{\mathsf{T}}Bx\right]_n \in [-1,\ 1] \;\text{ for } n = 1,\ldots,N \\ \Leftrightarrow \quad & \|B^{\mathsf{T}}Bx\|_\infty \leqslant 1. \end{aligned}$$

Hence, the function S_B coincides with $\frac{1}{2}\|B\cdot\|_2^2$ on a subset of its domain. ■

Proposition 10: Let $B \in \mathbb{R}^{M\times N}$ and set $\alpha = \|B\|_2$. The generalized Huber function satisfies

$$S_B(x) \leqslant S_{\alpha I}(x), \quad \forall x \in \mathbb{R}^N \tag{36}$$

$$= \sum_n s_\alpha(x_n). \tag{37}$$

Proof: Using (31), we have

$$\begin{aligned} S_B(x) &= \min_{v\in\mathbb{R}^N} \left\{\|v\|_1 + \tfrac{1}{2}\|B(x-v)\|_2^2\right\} \\ &\leqslant \min_{v\in\mathbb{R}^N} \left\{\|v\|_1 + \tfrac{1}{2}\|B\|_2^2\,\|(x-v)\|_2^2\right\} \\ &= \min_{v\in\mathbb{R}^N} \left\{\|v\|_1 + \tfrac{1}{2}\alpha^2\|(x-v)\|_2^2\right\} \\ &= \min_{v\in\mathbb{R}^N} \left\{\|v\|_1 + \tfrac{1}{2}\|\alpha\,(x-v)\|_2^2\right\} \\ &= S_{\alpha I}(x). \end{aligned}$$

From Proposition 7 we have (37). ■

The Moreau envelope is well studied in convex analysis [4]. Hence, it is useful to express the generalized Huber function S_B in terms of a Moreau envelope, so we can draw on results in convex analysis to derive further properties of the generalized Huber function.

Lemma 1: If $B \in \mathbb{R}^{N\times N}$ is invertible, then the generalized Huber function S_B can be expressed in terms of a Moreau envelope as

$$S_B = \big(\|\cdot\|_1 \circ B^{-1}\big)^{\mathsf{M}} \circ B. \tag{38}$$

Proof: Using (29), we have

$$\begin{aligned} S_B &= \|\cdot\|_1 \,\square\, \big(\tfrac{1}{2}\|\cdot\|_2^2 \circ B\big) \\ &= \big(\|\cdot\|_1 \,\square\, \big(\tfrac{1}{2}\|\cdot\|_2^2 \circ B\big)\big) \circ B^{-1} \circ B \\ &= \big(\big(\|\cdot\|_1 \circ B^{-1}\big) \,\square\, \big(\tfrac{1}{2}\|\cdot\|_2^2\big)\big) \circ B \\ &= \big(\|\cdot\|_1 \circ B^{-1}\big)^{\mathsf{M}} \circ B. \end{aligned}$$

■

Lemma 2: If $B \in \mathbb{R}^{M\times N}$ has full row-rank, then the generalized Huber function S_B can be expressed in terms of a Moreau envelope as

$$S_B = \big(d \circ B^{+}\big)^{\mathsf{M}} \circ B \tag{39}$$

where $d : \mathbb{R}^N \to \mathbb{R}$ is the convex distance function

$$d(x) = \min_{w\in \operatorname{null} B} \|x - w\|_1 \tag{40}$$

which represents the distance from the point $x \in \mathbb{R}^N$ to the null space of B as measured by the ℓ_1 norm.

Proof: Using (31), we have

$$\begin{aligned} S_B(x) &= \min_{v\in\mathbb{R}^N} \big\{\|v\|_1 + \tfrac{1}{2}\|B(x-v)\|_2^2\big\} \\ &= f(Bx) \end{aligned}$$

where $f : \mathbb{R}^M \to \mathbb{R}$ is given by

$$\begin{aligned} f(z) &= \min_{v\in\mathbb{R}^N} \big\{\|v\|_1 + \tfrac{1}{2}\|z - Bv\|_2^2\big\} \\ &= \min_{u\in(\operatorname{null} B)^\perp} \min_{w\in\operatorname{null} B} \big\{\|u+w\|_1 + \tfrac{1}{2}\|z - B(u+w)\|_2^2\big\} \\ &= \min_{u\in(\operatorname{null} B)^\perp} \min_{w\in\operatorname{null} B} \big\{\|u+w\|_1 + \tfrac{1}{2}\|z - Bu\|_2^2\big\} \\ &= \min_{u\in(\operatorname{null} B)^\perp} \big\{d(u) + \tfrac{1}{2}\|z - Bu\|_2^2\big\} \end{aligned}$$

where d is the convex function given by (40). The fact that d is convex follows from Proposition 8.26 of [4] and Examples 3.16 and 3.17 of [8]. Since $(\operatorname{null} B)^\perp = \operatorname{range} B^{\mathsf{T}}$,

$$\begin{aligned} f(z) &= \min_{u\in\operatorname{range} B^{\mathsf{T}}} \big\{d(u) + \tfrac{1}{2}\|z - Bu\|_2^2\big\} \\ &= \min_{v\in\mathbb{R}^M} \big\{d(B^{\mathsf{T}}v) + \tfrac{1}{2}\|z - BB^{\mathsf{T}}v\|_2^2\big\} \\ &= \min_{v\in\mathbb{R}^M} \big\{d(B^{\mathsf{T}}(BB^{\mathsf{T}})^{-1}v) + \tfrac{1}{2}\|z - BB^{\mathsf{T}}(BB^{\mathsf{T}})^{-1}v\|_2^2\big\} \\ &= \min_{v\in\mathbb{R}^M} \big\{d(B^{+}v) + \tfrac{1}{2}\|z - v\|_2^2\big\} \\ &= \big(d(B^{+}\cdot)\big)^{\mathsf{M}}(z). \end{aligned}$$

Hence, $S_B(x) = \big(d(B^{+}\cdot)\big)^{\mathsf{M}}(Bx)$ which completes the proof. ■

Note that (39) reduces to (38) when B is invertible. (Suppose B is invertible. Then $\operatorname{null} B = \{0\}$; hence $d(x) = \|x\|_1$ in (40). Additionally, $B^{+} = B^{-1}$.)

Proposition 11: The generalized Huber function is differentiable.

Proof: By Lemma 2, S_B is the composition of a Moreau envelope of a convex function and a linear function. Additionally, by Proposition 5, $S_B \in \Gamma_0(\mathbb{R}^N)$. By Proposition 12.29 in [4], it follows that S_B is differentiable. ■

The following result regards the gradient of the generalized Huber function. This result will be used in Sec. V to show the generalized MC penalty defined therein constitutes a valid penalty.

Lemma 3: The gradient of the generalized Huber function $S_B : \mathbb{R}^N \to \mathbb{R}$ satisfies

$$\|\nabla S_B(x)\|_\infty \leqslant 1 \;\text{ for all } x \in \mathbb{R}^N. \tag{41}$$

Proof: Since S_B is convex and differentiable, we have

$$S_B(v) + \big[\nabla S_B(v)\big]^{\mathsf{T}}(x - v) \leqslant S_B(x), \quad \forall x \in \mathbb{R}^N,\ \forall v \in \mathbb{R}^N.$$

Using Proposition 8, it follows that

$$S_B(v) + \big[\nabla S_B(v)\big]^{\mathsf{T}}(x - v) \leqslant \|x\|_1, \quad \forall x \in \mathbb{R}^N,\ \forall v \in \mathbb{R}^N.$$

Let $x = (0,\ldots,0,t,0,\ldots,0)$ where t is in position n. It follows that

$$c(v) + \big[\nabla S_B(v)\big]_n t \leqslant |t|, \quad \forall t \in \mathbb{R},\ \forall v \in \mathbb{R}^N \tag{42}$$

where $c(v) \in \mathbb{R}$ does not depend on t. It follows from (42) that $|[\nabla S_B(v)]_n| \leqslant 1$. ■

The generalized Huber function can be evaluated by taking the pointwise minimum of numerous simpler functions (comprising quadratics, absolute values, and linear functions). This generalizes the situation for the scalar Huber function, which can be evaluated as the pointwise minimum of three functions, as expressed in (11) and illustrated in Fig. 2. Unfortunately, evaluating the generalized Huber function on $\mathbb{R}^N$ this way requires the evaluation of 3^N simpler functions, which is not practical except for small N. In turn, the evaluation of the GMC penalty is also impractical. However, we do not need to explicitly evaluate these functions to utilize them for sparse regularization, as shown in Sec. VII. For this paper, we compute these functions on $\mathbb{R}^2$ only for the purpose of illustration (Figs. 6 and 7).

V. Generalized MC Penalty

In this section, we propose a multivariate generalization of the MC penalty (12). The basic idea is to generalize (22) using the ℓ_1 norm and the generalized Huber function.

Definition 7: Let $B \in \mathbb{R}^{M\times N}$. We define the generalized MC (GMC) penalty function $\psi_B : \mathbb{R}^N \to \mathbb{R}$ as

$$\psi_B(x) := \|x\|_1 - S_B(x) \tag{43}$$

where S_B is the generalized Huber function (29).

The GMC penalty reduces to a separable penalty when $B^{\mathsf{T}}B$ is diagonal.

Proposition 12: Let $B \in \mathbb{R}^{M\times N}$. If $B^{\mathsf{T}}B$ is a diagonal matrix, then ψ_B is separable (additive), comprising a sum of scalar

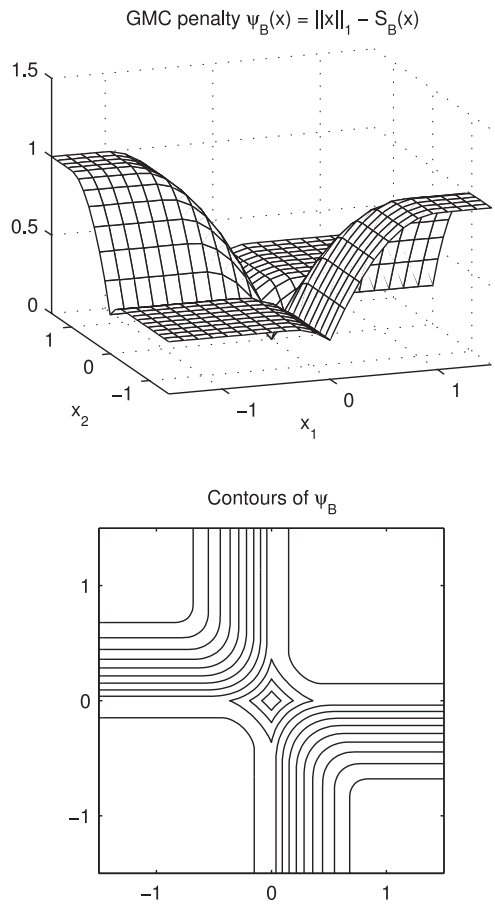

Fig. 8. The GMC penalty for the matrix B in (32).

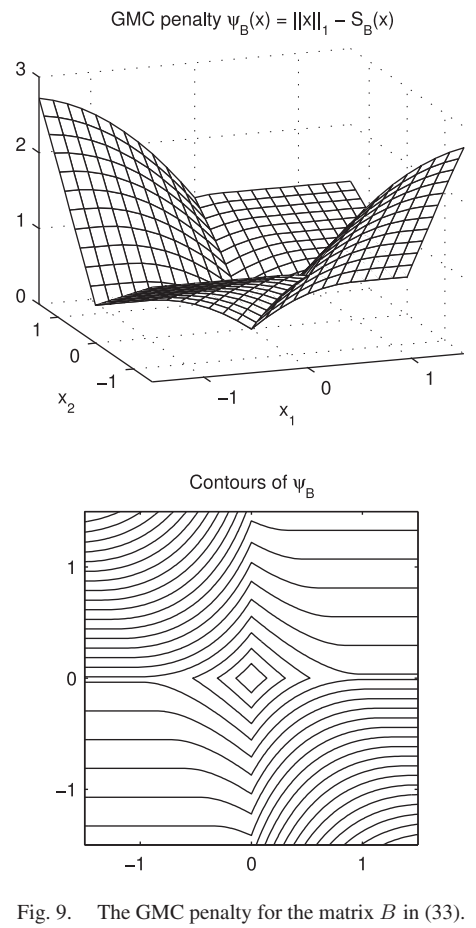

Fig. 9. The GMC penalty for the matrix B in (33).

MC penalties. Specifically,

$$B^{\mathsf{T}}B = \operatorname{diag}(\alpha_1^2, \ldots, \alpha_N^2) \;\Rightarrow\; \psi_B(x) = \sum_n \phi_{\alpha_n}(x_n)$$

where ϕ_b is the scaled MC penalty (22). If $B^{\mathsf{T}}B = 0$, then $\psi_B(x) = \|x\|_1$.

Proof: If $B^{\mathsf{T}}B = \operatorname{diag}(\alpha_1^2, \ldots, \alpha_N^2)$, then by Proposition 7 we have

$$\psi_B(x) = \|x\|_1 - \sum_n s_{\alpha_n}(x_n)$$

$$= \sum_n |x_n| - s_{\alpha_n}(x_n)$$

which proves the result in light of definition (22). ■

The most interesting case (the case that motivates the GMC penalty) is the case where $B^{\mathsf{T}}B$ is a non-diagonal matrix. If $B^{\mathsf{T}}B$ is non-diagonal, then the GMC penalty is non-separable.

Example 3: For the matrices B given in (32) and (33), the GMC penalty is illustrated in Fig. 8 and Fig. 9, respectively.

The following corollaries follow directly from Propositions 8 and 9.

Corollary 1: The generalized MC penalty satisfies

$$0 \leqslant \psi_B(x) \leqslant \|x\|_1 \quad \text{for all } x \in \mathbb{R}^N. \tag{44}$$

Corollary 2: Given $B \in \mathbb{R}^{M\times N}$, the generalized MC penalty satisfies

$$\psi_B(x) = \|x\|_1 - \tfrac{1}{2}\|Bx\|_2^2 \text{ if and only if } \|B^{\mathsf{T}}Bx\|_\infty \leqslant 1. \tag{45}$$

The corollaries imply that around zero the generalized MC penalty approximates the ℓ_1 norm (from below), i.e., $\psi_B(x) \approx \|x\|_1$ for $x \approx 0$.

The generalized MC penalty has a basic property expected of a regularization function; namely, that large values are penalized more than (or the same as) small values. Specifically, if $v, x \in \mathbb{R}^N$ with $|v_i| \geqslant |x_i|$ and $\operatorname{sign} v_i = \operatorname{sign} x_i$ for $i = 1, \ldots, N$, then $\psi_B(v) \geqslant \psi_B(x)$. That is, in any given quadrant, the function $\psi_B(x)$ is a non-decreasing function in each $|x_i|$. This is formalized in the following proposition, and illustrated in Figs. 8 and 9. Basically, the gradient of ψ_B points away from the origin.

Proposition 13: Let $x \in \mathbb{R}^N$ with $x_i \neq 0$. The generalized MC penalty ψ_B has the property that $[\nabla\psi_B(x)]_i$ either has the same sign as x_i or is equal to zero.

Proof: Let $x \in \mathbb{R}^N$ with $x_i \neq 0$. Then, from the definition of the MC penalty,

$$\frac{\partial \psi_B}{\partial x_i}(x) = \operatorname{sign}(x_i) - \frac{\partial S_B}{\partial x_i}(x).$$

From Lemma 3, $|\partial S_B(x)/\partial x_i| \leqslant 1$. Hence $\partial\psi_B(x)/\partial x_i \geqslant 0$ when $x_i > 0$, and $\partial\psi_B(x)/\partial x_i \leqslant 0$ when $x_i < 0$. ■

A penalty function not satisfying Proposition 13 would not be considered an effective sparsity-inducing regularizer.

VI. Sparse Regularization

In this section, we consider how to set the GMC penalty to maintain the convexity of the regularized least square cost function. To that end, the condition (47) below generalizes the scalar convexity condition (25).

Theorem 1: Let $y \in \mathbb{R}^M$, $A \in \mathbb{R}^{M\times N}$, and $\lambda > 0$. Define $F : \mathbb{R}^N \to \mathbb{R}$ as

$$F(x) = \frac{1}{2}\|y - Ax\|_2^2 + \lambda\, \psi_B(x) \tag{46}$$

where $\psi_B : \mathbb{R}^N \to \mathbb{R}$ is the generalized MC penalty (43). If

$$B^{\mathsf{T}}B \preccurlyeq \frac{1}{\lambda} A^{\mathsf{T}}A \tag{47}$$

then F is a convex function.

Proof: Write F as

$$\begin{aligned} F(x) &= \tfrac{1}{2}\|y - Ax\|_2^2 + \lambda\left(\|x\|_1 - S_B(x)\right) \\ &= \tfrac{1}{2}\|y - Ax\|_2^2 + \lambda\,\|x\|_1 \\ &\qquad - \min_{v\in\mathbb{R}^N}\left\{\lambda\|v\|_1 + \frac{\lambda}{2}\|B(x-v)\|_2^2\right\} \\ &= \max_{v\in\mathbb{R}^N}\left\{\tfrac{1}{2}\|y - Ax\|_2^2 + \lambda\,\|x\|_1 \right. \\ &\qquad \left. - \lambda\|v\|_1 - \frac{\lambda}{2}\|B(x-v)\|_2^2\right\} \\ &= \max_{v\in\mathbb{R}^N}\left\{\tfrac{1}{2}x^{\mathsf{T}}\big(A^{\mathsf{T}}A - \lambda B^{\mathsf{T}}B\big)x + \lambda\|x\|_1 + g(x,v)\right\} \\ &= \tfrac{1}{2}x^{\mathsf{T}}\big(A^{\mathsf{T}}A - \lambda B^{\mathsf{T}}B\big)x + \lambda\|x\|_1 + \max_{v\in\mathbb{R}^N} g(x,v) \end{aligned}$$

where g is affine in x. The last term is convex as it is the pointwise maximum of a set of convex functions (Proposition 8.14 in [4]). Hence, F is convex if $A^{\mathsf{T}}A - \lambda B^{\mathsf{T}}B$ is positive semidefinite. ■

The convexity condition (47) is easily satisfied. Given A, we may simply set

$$B = \sqrt{\gamma/\lambda}\, A, \quad 0 \leqslant \gamma \leqslant 1. \tag{48}$$

Then $B^{\mathsf{T}}B = (\gamma/\lambda)A^{\mathsf{T}}A$ which satisfies (47) when $\gamma \leqslant 1$. The parameter γ controls the non-convexity of the penalty ψ_B. If $\gamma = 0$, then $B = 0$ and the penalty reduces to the ℓ_1 norm. If $\gamma = 1$, then (47) is satisfied with equality and the penalty is 'maximally' non-convex. In practice, we use a nominal range of $0.5 \leqslant \gamma \leqslant 0.8$.

When $A^{\mathsf{T}}A$ is diagonal, the proposed methodology reduces to element-wise firm thresholding.

Proposition 14: Let $y \in \mathbb{R}^M$, $A \in \mathbb{R}^{M\times N}$, and $\lambda > 0$. If $A^{\mathsf{T}}A$ is diagonal with positive diagonal entries and B is given by (48), then the minimizer of the cost function F in (46) is given by element-wise firm thresholding. Specifically, if

$$A^{\mathsf{T}}A = \operatorname{diag}(\alpha_1^2, \ldots, \alpha_N^2), \tag{49}$$

then

$$x_n^{\mathsf{opt}} = \operatorname{firm}([A^{\mathsf{T}}y]_n/\alpha_n^2; \lambda/\alpha_n^2, \lambda/(\gamma\alpha_n^2)) \tag{50}$$

when $0 < \gamma \leqslant 1$, and

$$x_n^{\mathsf{opt}} = \operatorname{soft}([A^{\mathsf{T}}y]_n/\alpha_n^2; \lambda/\alpha_n^2) \tag{51}$$

when $\gamma = 0$.

Proof: If $A^{\mathsf{T}}A = \operatorname{diag}(\alpha_1^2, \ldots, \alpha_N^2)$, then

$$\begin{aligned} \tfrac{1}{2}\|y - Ax\|_2^2 &= \tfrac{1}{2}y^{\mathsf{T}}y - x^{\mathsf{T}}A^{\mathsf{T}}y + \tfrac{1}{2}x^{\mathsf{T}}A^{\mathsf{T}}Ax \\ &= \tfrac{1}{2}y^{\mathsf{T}}y + \sum_n \left(-x_n[A^{\mathsf{T}}y]_n + \tfrac{1}{2}\alpha_n^2 x_n^2\right) \\ &= \sum_n \tfrac{1}{2}\left([A^{\mathsf{T}}y]_n/\alpha_n - \alpha_n x_n\right)^2 + C \end{aligned}$$

where C does not depend on x. If B is given by (48), then

$$B^{\mathsf{T}}B = (\gamma/\lambda)\operatorname{diag}(\alpha_1^2, \ldots, \alpha_N^2).$$

Using Proposition 12, we have

$$\psi_B(x) = \sum_n \phi_{\alpha_n\sqrt{\gamma/\lambda}}(x_n).$$

Hence, F in (46) is given by

$$F(x) = \sum_n \left[\tfrac{1}{2}\left([A^{\mathsf{T}}y]_n/\alpha_n - \alpha_n x_n\right)^2 + \lambda\phi_{\alpha_n\sqrt{\gamma/\lambda}}(x_n)\right] + C$$

and so (50) follows from (28). ■

VII. Optimization Algorithm

Even though the GMC penalty does not have a simple explicit formula, a global minimizer of the sparse-regularized cost function (46) can be readily calculated using proximal algorithms. It is not necessary to explicitly evaluate the GMC penalty or its gradient.

To use proximal algorithms to minimize the cost function F in (46) when B satisfies (47), we rewrite it as a saddle-point problem:

$$(x^{\mathsf{opt}}, v^{\mathsf{opt}}) = \arg\min_{x\in\mathbb{R}^N}\max_{v\in\mathbb{R}^N} F(x,v) \tag{52}$$

where

$$F(x,v) = \frac{1}{2}\|y - Ax\|_2^2 + \lambda\,\|x\|_1 - \lambda\|v\|_1 - \frac{\lambda}{2}\|B(x-v)\|_2^2 \tag{53}$$

If we use (48) with $0 \leqslant \gamma \leqslant 1$, then the saddle function is given by

$$F(x,v) = \frac{1}{2}\|y - Ax\|_2^2 + \lambda\,\|x\|_1 - \lambda\|v\|_1 - \frac{\gamma}{2}\|A(x-v)\|_2^2. \tag{54}$$

These saddle-point problems are instances of monotone inclusion problems. Hence, the solution can be obtained using the forward-backward (FB) algorithm for such a problems; see Theorem 25.8 of Ref. [4]. The FB algorithm involves only simple computational steps (soft-thresholding and the operators A and A^{T}).

Proposition 15: Let $\lambda > 0$ and $0 \leqslant \gamma < 1$. Let $y \in \mathbb{R}^N$ and $A \in \mathbb{R}^{M\times N}$. Then a saddle-point $(x^{\mathsf{opt}}, v^{\mathsf{opt}})$ of F in (54) can be obtained by the iterative algorithm:

Set $\rho = \max\{1,\ \gamma/(1-\gamma)\}\,\|A^{\mathsf{T}}A\|_2$

Set $\mu : 0 < \mu < 2/\rho$

For $i = 0, 1, 2, \ldots$

$w^{(i)} = x^{(i)} - \mu A^{\mathsf{T}}\big(A(x^{(i)} + \gamma(v^{(i)} - x^{(i)})) - y\big)$

$u^{(i)} = v^{(i)} - \mu\gamma A^{\mathsf{T}}A(v^{(i)} - x^{(i)})$

$x^{(i+1)} = \mathrm{soft}(w^{(i)},\ \mu\lambda)$

$v^{(i+1)} = \mathrm{soft}(u^{(i)},\ \mu\lambda)$

end

where i is the iteration counter.

Proof: The point $(x^{\mathsf{opt}}, v^{\mathsf{opt}})$ is a saddle-point of F if $0 \in \partial F(x^{\mathsf{opt}}, v^{\mathsf{opt}})$ where ∂F is the subdifferential of F. From (54), we have

$$\partial_x F(x,v) = A^{\mathsf{T}}(Ax - y) - \gamma A^{\mathsf{T}}A(x - v) + \lambda\,\mathrm{sign}(x)$$

$$\partial_v F(x,v) = \gamma A^{\mathsf{T}}A(x - v) - \lambda\,\mathrm{sign}(v).$$

Hence, $0 \in \partial F$ if $0 \in P(x,v) + Q(x,v)$ where

$$P(x,v) = \begin{bmatrix} (1-\gamma)A^{\mathsf{T}}A & \gamma A^{\mathsf{T}}A \\ -\gamma A^{\mathsf{T}}A & \gamma A^{\mathsf{T}}A \end{bmatrix}\begin{bmatrix} x \\ v \end{bmatrix} - \begin{bmatrix} A^{\mathsf{T}}y \\ 0 \end{bmatrix}$$

$$Q(x,v) = \begin{bmatrix} \lambda\,\mathrm{sign}(x) \\ \lambda\,\mathrm{sign}(v) \end{bmatrix}.$$

Finding (x, v) such that $0 \in P(x,v) + Q(x,v)$ is the problem of constructing a zero of a sum of operators. The operators P and Q are maximally monotone and P is single-valued and β-cocoercive with $\beta > 0$; hence, the forward-backward algorithm (Theorem 25.8 in [4]) can be used. In the current notation, the forward-backward algorithm is

$$\begin{bmatrix} w^{(i)} \\ u^{(i)} \end{bmatrix} = \begin{bmatrix} x^{(i)} \\ v^{(i)} \end{bmatrix} - \mu P(x^{(i)},\ v^{(i)})$$

$$\begin{bmatrix} x^{(i+1)} \\ v^{(i+1)} \end{bmatrix} = J_{\mu Q}(w^{(i)},\ u^{(i)})$$

where $J_Q = (I + Q)^{-1}$ is the *resolvent* of Q. The resolvent of the sign function is soft thresholding. The constant μ should be chosen $0 < \mu < 2\beta$ where P is β-cocoercive (Definition 4.4 in [4]), i.e., βP is firmly non-expansive. We now address the value β. By Corollary 4.3(v) in [4], this condition is equivalent to

$$\tfrac{1}{2}P + \tfrac{1}{2}P^{\mathsf{T}} - \beta P^{\mathsf{T}}P \succcurlyeq 0. \tag{55}$$

We may write P using a Kronecker product,

$$P = \begin{bmatrix} 1-\gamma & \gamma \\ -\gamma & \gamma \end{bmatrix} \otimes A^{\mathsf{T}}A.$$

Then we have

$$\begin{aligned} &\tfrac{1}{2}P + \tfrac{1}{2}P^{\mathsf{T}} - \beta P^{\mathsf{T}}P \\ &= \begin{bmatrix} 1-\gamma & 0 \\ 0 & \gamma \end{bmatrix} \otimes A^{\mathsf{T}}A \\ &\qquad - \beta\begin{bmatrix} 1-\gamma & -\gamma \\ \gamma & \gamma \end{bmatrix}\begin{bmatrix} 1-\gamma & \gamma \\ -\gamma & \gamma \end{bmatrix} \otimes (A^{\mathsf{T}}A)^2 \\ &= \left(\left(\begin{bmatrix} 1-\gamma & 0 \\ 0 & \gamma \end{bmatrix} - \beta_1\begin{bmatrix} 1-\gamma & -\gamma \\ \gamma & \gamma \end{bmatrix}\begin{bmatrix} 1-\gamma & \gamma \\ -\gamma & \gamma \end{bmatrix}\right) \otimes I_N\right) \\ &\qquad \times \big(I_2 \otimes \big(I_N - \beta_2 A^{\mathsf{T}}A\big)\big)\,\big(I_2 \otimes A^{\mathsf{T}}A\big) \end{aligned}$$

where $\beta = \beta_1\beta_2$. Hence, (55) is satisfied if

$$\begin{bmatrix} 1-\gamma & 0 \\ 0 & \gamma \end{bmatrix} - \beta_1\begin{bmatrix} 1-\gamma & -\gamma \\ \gamma & \gamma \end{bmatrix}\begin{bmatrix} 1-\gamma & \gamma \\ -\gamma & \gamma \end{bmatrix} \succcurlyeq 0$$

and

$$I_N - \beta_2 A^{\mathsf{T}}A \succcurlyeq 0.$$

These conditions are respectively satisfied if

$$\beta_1 \leqslant 1/\max\{1,\ \gamma/(1-\gamma)\}$$

and

$$\beta_2 \leqslant 1/\|A^{\mathsf{T}}A\|_2.$$

The FB algorithm requires that P be β-cocoercive with $\beta > 0$; hence, $\gamma = 1$ is precluded. ■

If $\gamma = 0$ in Proposition 15, then the algorithm reduces to the classic iterative shrinkage/thresholding algorithm (ISTA) [22], [26].

The Douglas-Rachford algorithm (Theorem 25.6 in [4]) may also be used to find a saddle-point of F in (54).

VIII. Numerical Examples

A. Denoising Using Frequency-Domain Sparsity

This example illustrates the use of the GMC penalty for denoising [18]. Specifically, we consider the estimation of the discrete-time signal

$$g(m) = 2\cos(2\pi f_1 m) + \sin(2\pi f_2 m),\quad m = 0, \ldots, M-1$$

of length $M = 100$ with frequencies $f_1 = 0.1$ and $f_2 = 0.22$. This signal is sparse in the frequency domain, so we model the signal as $g = Ax$ where A is an over-sampled inverse discrete Fourier transform and $x \in \mathbb{C}^N$ is a sparse vector of Fourier coefficients with $N \geqslant M$. Specifically, we define the matrix $A \in \mathbb{C}^{M\times N}$ as

$$\begin{aligned} A_{m,n} =\ & \big(1/\sqrt{N}\,\big)\exp(\mathrm{j}(2\pi/N)mn), \\ m =\ & 0, \ldots, M-1,\ n = 0, \ldots, N-1 \end{aligned}$$

with $N = 256$. The columns of A form a normalized tight frame, i.e., $AA^{\mathsf{H}} = I$ where A^{H} is the complex conjugate transpose of A. For the denoising experiment, we corrupt the signal with

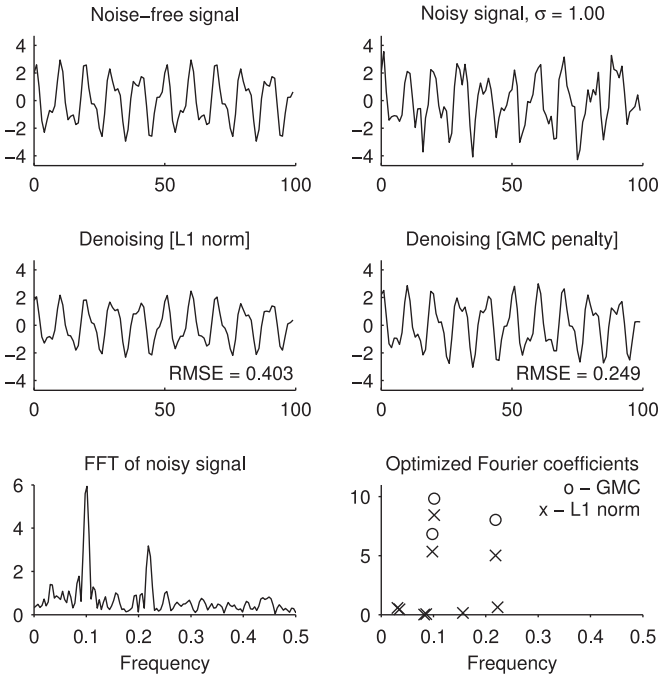

Fig. 10. Denoising using the ℓ_1 norm and the proposed GMC penalty. The plot of optimized coefficients shows only the non-zero values.

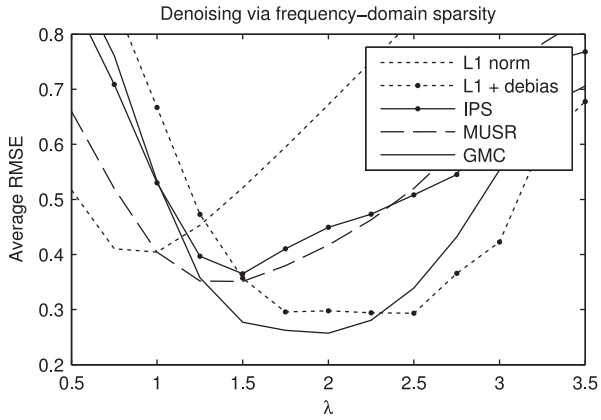

Fig. 11. Average RMSE for three denoising methods.

additive white Gaussian noise (AWGN) with standard deviation $\sigma = 1.0$, as illustrated in Fig. 10.

In addition to the ℓ_1 norm and proposed GMC penalty, we use several other methods: debiasing the ℓ_1 norm solution [27], iterative p-shrinkage (IPS) [62], [64], and multivariate sparse regularization (MUSR) [52]. Debiasing the ℓ_1 norm solution is a two-step approach where the ℓ_1-norm solution is used to estimate the support, then the identified non-zero values are re-estimated by un-regularized least squares. The IPS algorithm is a type of iterative thresholding algorithm that performed particularly well in a detailed comparison of several algorithms [55]. MUSR regularization is a precursor of the GMC penalty, i.e., a non-separable non-convex penalty designed to maintain cost function convexity, but with a simpler functional form.

In this denoising experiment, we use 20 realizations of the noise. Each method calls for a regularization parameter λ to be set. We vary λ from 0.5 to 3.5 (with increment 0.25) and evaluate the RMSE for each method, for each λ, and for each realization. For the GMC method we must also specify the matrix B, which we set using (48) with $\gamma = 0.8$. Since $B^{\mathsf{H}}B$ is not diagonal, the GMC penalty is non-separable. The average RMSE as a function of λ for each method is shown in Fig. 11.

The GMC compares favorably with the other methods, achieving the minimum average RMSE. The next best-performing method is debiasing of the ℓ_1-norm solution, which performs almost as well as GMC. Note that this debiasing method does not minimize an initially prescribed cost function, in contrast to the other methods. The IPS algorithm aims to minimize a (non-convex) cost function.

Figure 10 shows the ℓ_1-norm and GMC solutions for a particular noise realization. The solutions shown in this figure were obtained using the value of λ that minimizes the average RMSE ($\lambda = 1.0$ and $\lambda = 2.0$, respectively). Comparing the ℓ_1 norm and GMC solutions, we observe: the GMC solution is more sparse in the frequency domain; and the ℓ_1 norm solution underestimates the coefficient amplitudes.

Neither increasing nor decreasing the regularization parameter λ helps the ℓ_1-norm solution here. A larger value of λ makes the ℓ_1-norm solution sparser, but reduces the coefficient amplitudes. A smaller value of λ increases the coefficient amplitudes of the ℓ_1-norm solution, but makes the solution less sparse and more noisy.

Note that the purpose of this example is to compare the proposed GMC penalty with other sparse regularizers. We are not advocating it for frequency estimation *per se*.

B. Denoising Using Time-Frequency Sparsity

This example considers the denoising of a bat echolocation pulse, shown in Fig. 12 (sampling period of 7 microseconds).[1] The bat pulse can be modeled as sparse in the time-frequency domain. We use a short-time Fourier transform (STFT) with 75% overlapping segments (the transform is four-times overcomplete). We implement the STFT as a normalized tight frame, i.e., $AA^{\mathsf{H}} = I$. The bat pulse and its spectrogram are illustrated in Fig. 12. For the denoising experiment, we contaminate the pulse with AWGN ($\sigma = 0.05$).

We perform denoising by estimating the STFT coefficients by minimizing the cost function F in (46) where A represents the inverse STFT operator. We set λ so as to minimize the root-mean-square error (RMSE). This leads to the values $\lambda = 0.030$ and $\lambda = 0.51$ for the ℓ_1-norm and GMC penalties, respectively. For the GMC penalty, we set B as in (48) with $\gamma = 0.7$. Since $B^{\mathsf{H}}B$ is not diagonal, the GMC penalty is non-separable. We then estimate the bat pulse by computing the inverse STFT of the optimized coefficients. With λ individually set for each method, the resulting RMSE is about the same (0.026). The optimized STFT coefficients (time-frequency representation) for each solution is shown in Fig. 12. We observe that the GMC solution has substantially fewer extraneous noise artifacts in the

[1]The bat echolocation pulse data is curtesy of Curtis Condon, Ken White, and Al Feng of the Beckman Center at the University of Illinois. Available online at http://dsp.rice.edu/software/bat-echolocation-chirp.

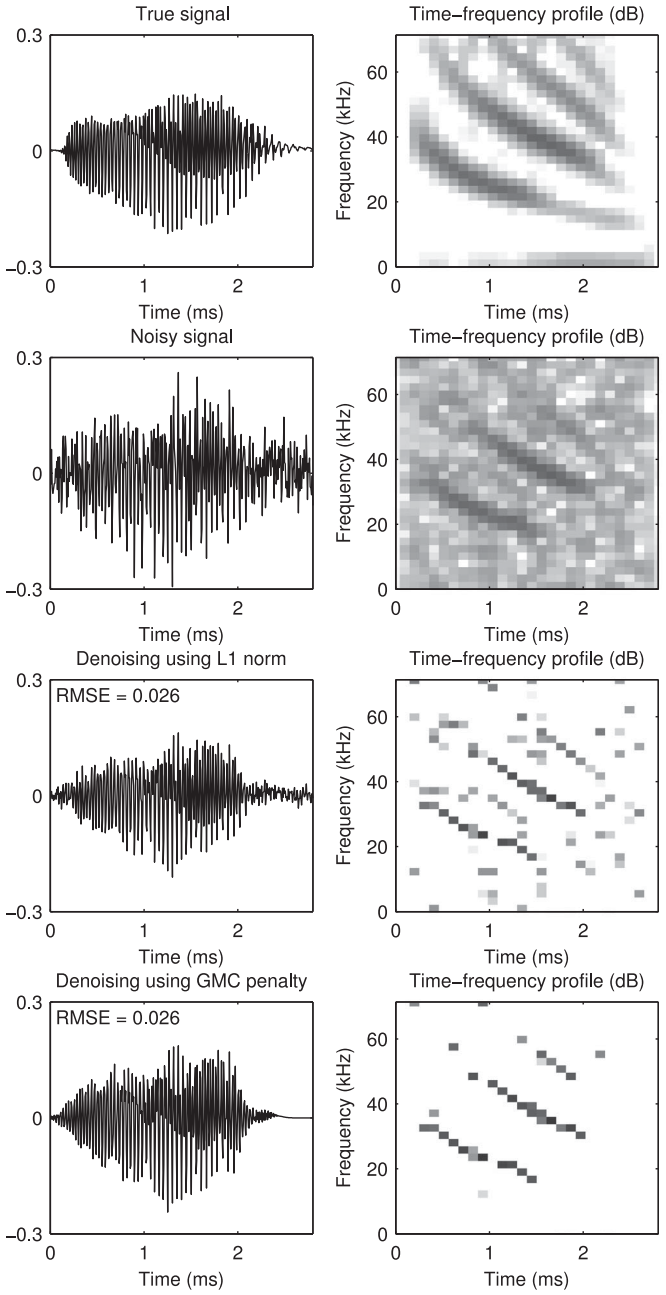

Fig. 12. Denoising a bat echolocation pulse using the ℓ_1 norm and GMC penalty. The GMC penalty results in fewer extraneous noise artifacts in the time-frequency representation.

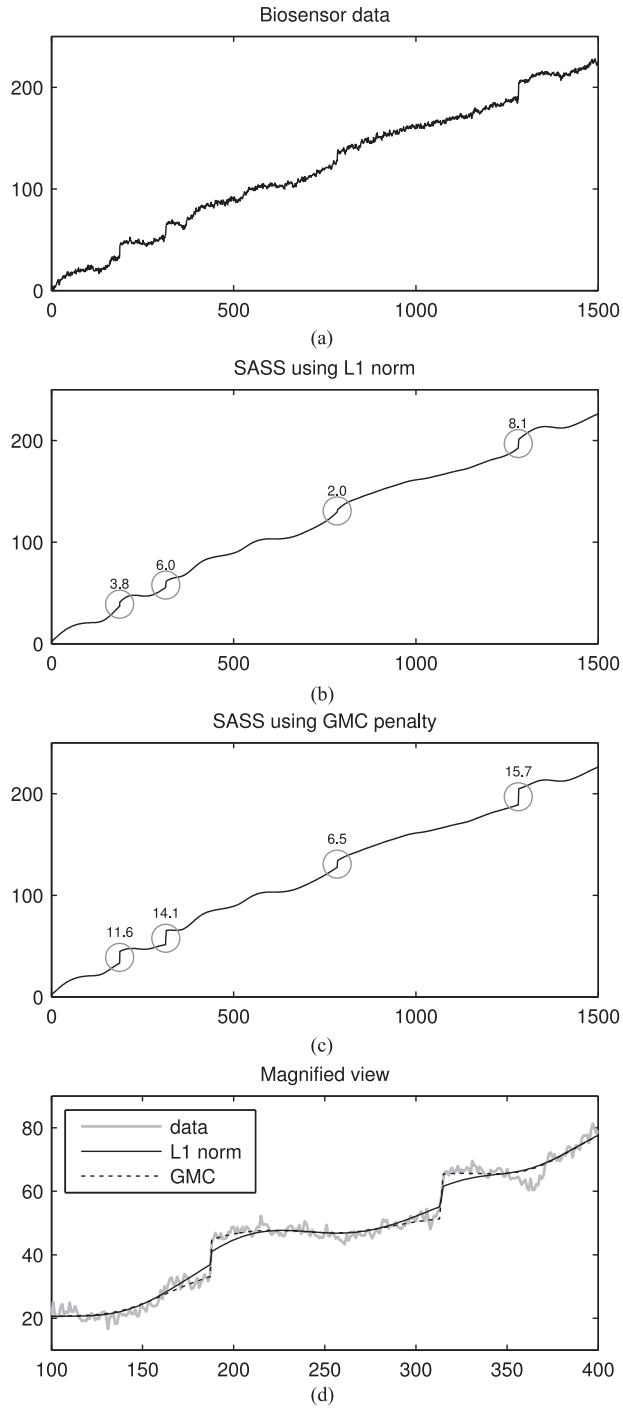

Fig. 13. Sparsity-assisted signal smoothing (SASS) using ℓ_1-norm and GMC regularization, as applied to biosensor data. The GMC method more accurately estimates jump discontinuities.

time-frequency representation, compared to the ℓ_1 norm solution. (The time-frequency representations in Fig. 12 are shown in decibels with 0 dB being black and -50 dB being white.)

C. Sparsity-Assisted Signal Smoothing

This example uses the GMC penalty for sparsity-assisted signal smoothing (SASS) [50], [53]. The SASS method is suitable for the denoising of signals that are smooth for the exception of singularities. Here, we use SASS to denoise the biosensor data illustrated in Fig. 13(a), which exhibits jump discontinuities. This data was acquired using a whispering gallery mode (WGM) sensor designed to detect nano-particles with high sensitivity [3], [21]. Nano-particles show up as jump discontinuities in the data.

The SASS technique formulates the denoising problem as a sparse deconvolution problem. The cost function to be minimized has the form (46). The exact cost function, given by

equation (42) in Ref. [50], depends on a prescribed low-pass filter and the order of the singularities the signal is assumed to possess. For the biosensor data shown in Fig. 13, the singularities are of order $K = 1$ since the first-order derivative of the signal exhibits impulses. In this example, we use a low-pass filter of order $d = 2$ and cut-off frequency $f_c = 0.01$ (these parameter values designate a low-pass filter as described in [50]). We set $\lambda = 32$ and, for the GMC penalty, we set $\gamma = 0.7$. Solving the SASS problem using the ℓ_1 norm and GMC penalty yields the denoised signals shown in Figs. 13(b) and 13(c), respectively. The amplitudes of the jump discontinuities are indicated in the figure.

It can be seen, especially in Fig. 13(d), that the GMC solution estimates the jump discontinuities more accurately than the ℓ_1 norm solution. The ℓ_1 norm solution tends to underestimate the amplitudes of the jump discontinuities. To reduce this tendency, a smaller value of λ could be used, but that tends to produce false discontinuities (false detections).

IX. Conclusion

In regards to the sparse-regularized linear least squares problem, this work bridges the convex (i.e., ℓ_1 norm) and the non-convex (e.g., ℓ_p norm with $p < 1$) approaches, which are usually mutually exclusive and incompatible. Specifically, this work formulates the sparse-regularized linear least squares problem using a non-convex generalization of the ℓ_1 norm that preserves the convexity of the cost function to be minimized. The proposed method leads to optimization problems with no extraneous suboptimal local minima and allows the leveraging of globally convergent, computationally efficient, scalable convex optimization algorithms. The advantage compared to ℓ_1 norm regularization is (*i*) more accurate estimation of high-amplitude components of sparse solutions or (*ii*) a higher level of sparsity in a sparse approximation problem. The sparse regularizer is expressed as the ℓ_1 norm minus a smooth convex function defined via infimal convolution. In the scalar case, the method reduces to firm thresholding (a generalization of soft thresholding).

Several extensions of this method are of interest. For example, the idea may admit extension to more general convex regularizers such as total variation [48], nuclear norm [10], mixed norms [34], composite regularizers [1], [2], co-sparse regularization [40], and more generally, atomic norms [14], and partly smooth regularizers [61]. Another extension of interest is to problems where the data fidelity term is not quadratic (e.g., Poisson denoising [24]).

References

[1] M. V. Afonso, J. M. Bioucas-Dias, and M. A. T. Figueiredo, "An augmented Lagrangian approach to linear inverse problems with compound regularization," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2010, pp. 4169–4172.

[2] R. Ahmad and P. Schniter, "Iteratively reweighted L1 approaches to sparse composite regularization," *IEEE Trans. Comput. Imag.*, vol. 1, no. 4, pp. 220–235, Dec. 2015.

[3] S. Arnold, M. Khoshima, I. Teraoka, S. Holler, and F. Vollmer, "Shift of whispering-gallery modes in microspheres by protein adsorption," *Opt. Lett*, vol. 28, no. 4, pp. 272–274, Feb. 2003.

[4] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. New York, NY, USA: Springer, 2011.

[5] İ. Bayram, "Penalty functions derived from monotone mappings," *IEEE Signal Process. Lett.*, vol. 22, no. 3, pp. 265–269, Mar. 2015.

[6] I. Bayram, "On the convergence of the iterative shrinkage/thresholding algorithm with a weakly convex penalty," *IEEE Trans. Signal Process.*, vol. 64, no. 6, pp. 1597–1608, Mar. 2016.

[7] A. Blake and A. Zisserman," *Visual Reconstruction*. Cambridge, MA, USA: MIT Press, 1987.

[8] S. Boyd and L. Vandenberghe," *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[9] A. Bruckstein, D. Donoho, and M. Elad, "From sparse solutions of systems of equations to sparse modeling of signals and images," *SIAM Rev.*, vol. 51, no. 1, pp. 34–81, 2009.

[10] E. J. Candes and Y. Plan, "Matrix completion with noise," *Proc. IEEE*, vol. 98, no. 6, pp. 925–936, Jun. 2010.

[11] E. J. Candès, M. B. Wakin, and S. Boyd, "Enhancing sparsity by reweighted l1 minimization," *J. Fourier Anal. Appl.*, vol. 14, no. 5, pp. 877–905, Dec. 2008.

[12] M. Carlsson, "On convexification/optimization of functionals including an l2-misfit term," Sep. 2016. [Online]. Available: https://arxiv.org/abs/1609.09378

[13] M. Castella and J.-C. Pesquet, "Optimization of a Geman-McClure like criterion for sparse signal deconvolution," in *Proc. IEEE Int. Workshop Comput. Adv. Multi-Sensor Adaptive Process.*, Dec. 2015, pp. 309–312.

[14] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky, "The convex geometry of linear inverse problems," *Found. Comput. Math.*, vol. 12, no. 6, pp. 805–849, 2012.

[15] R. Chartrand, "Shrinkage mappings and their induced penalty functions," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2014, pp. 1026–1029.

[16] L. Chen and Y. Gu, "The convergence guarantees of a non-convex approach for sparse recovery," *IEEE Trans. Signal Process.*, vol. 62, no. 15, pp. 3754–3767, Aug. 2014.

[17] P.-Y. Chen and I. W. Selesnick, "Group-sparse signal denoising: Non-convex regularization, convex optimization," *IEEE Trans. Signal Process.*, vol. 62, no. 13, pp. 3464–3478, Jul. 2014.

[18] S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comput.*, vol. 20, no. 1, pp. 33–61, 1998.

[19] 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.*, vol. 6, no. 1, pp. 563–591, 2013.

[20] P. L. Combettes, "Perspective functions: Properties, constructions, and examples," *Set-Valued Variational Anal.*, pp. 1–18, 2017, doi: 10.1007/s11228-017-0407-x.

[21] V. R. Dantham, S. Holler, V. Kolchenko, Z. Wan, and S. Arnold, "Taking whispering gallery-mode single virus detection and sizing to the limit," *Appl. Phys. Lett.*, vol. 101, no. 4, 2012, Art. no. 043704.

[22] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Commun. Pure Appl. Math*, vol. 57, no. 11, pp. 1413–1457, 2004.

[23] Y. Ding and I. W. Selesnick, "Artifact-free wavelet denoising: Non-convex sparse regularization, convex optimization," *IEEE Signal Process. Lett.*, vol. 22, no. 9, pp. 1364–1368, Sep. 2015.

[24] F.-X. Dupé, J. M. Fadili, and J.-L. Starck, "A proximal iteration for deconvolving Poisson noisy images using sparse representations," *IEEE Trans. Image Process.*, vol. 18, no. 2, pp. 310–321, Feb. 2009.

[25] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *J. Amer. Statist. Assoc.*, vol. 96, no. 456, pp. 1348–1360, 2001.

[26] M. Figueiredo and R. Nowak, "An EM algorithm for wavelet-based image restoration," *IEEE Trans. Image Process.*, vol. 12, no. 8, pp. 906–916, Aug. 2003.

[27] 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. Topics Signal Process.*, vol. 1, no. 4, pp. 586–598, Dec. 2007.

[28] M. Fornasier and H. Rauhut, "Iterative thresholding algorithms," *J. Appl. Component Harmonic Anal.*, vol. 25, no. 2, pp. 187–208, 2008.

[29] H.-Y. Gao and A. G. Bruce, "Waveshrink with firm shrinkage," *Statistica Sinica*, vol. 7, pp. 855–874, 1997.

[30] G. Gasso, A. Rakotomamonjy, and S. Canu, "Recovering sparse signals with a certain family of nonconvex penalties and DC programming," *IEEE Trans. Signal Process.*, vol. 57, no. 12, pp. 4686–4698, Dec. 2009.

[31] A. Gholami and S. M. Hosseini, "A general framework for sparsity-based denoising and inversion," *IEEE Trans. Signal Process.*, vol. 59, no. 11, pp. 5202–5211, Nov. 2011.

[32] W. He, Y. Ding, Y. Zi, and I. W. Selesnick, "Sparsity-based algorithm for detecting faults in rotating machines," *Mech. Syst. Signal Process.*, vol. 72–73, pp. 46–64, May 2016.
[33] P. J. Huber, "Robust estimation of a location parameter," *Ann. Math. Statist.*, vol. 35, no. 1, pp. 73–101, 1964.
[34] M. Kowalski and B. Torrésani, "Sparsity and persistence: mixed norms provide simple signal models with dependent coefficients," *Signal, Image Video Process.*, vol. 3, no. 3, pp. 251–264, 2009.
[35] A. Lanza, S. Morigi, I. Selesnick, and F. Sgallari, "Nonconvex nonsmooth optimization via convex–nonconvex majorization–minimization," *Numerische Mathematik*, vol. 136, pp. 1–39, 2016.
[36] A. Lanza, S. Morigi, and F. Sgallari, "Convex image denoising via nonconvex regularization with parameter selection," *J. Math. Imag. Vis.*, vol. 56, no. 2, pp. 195–220, 2016.
[37] 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.*, vol. 64, no. 24, pp. 6650–6664, Dec. 2016.
[38] 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 20th Int. Conf. Image Process.*, 2013, pp. 2402–2406.
[39] H. Mohimani, M. Babaie-Zadeh, and C. Jutten, "A fast approach for overcomplete sparse decomposition based on smoothed l0 norm," *IEEE Trans. Signal Process.*, vol. 57, no. 1, pp. 289–301, Jan. 2009.
[40] S. Nam, M. E. Davies, M. Elad, and R. Gribonval, "The cosparse analysis model and algorithms," *J. Appl. Comput. Harmonic Anal.*, vol. 34, no. 1, pp. 30–56, 2013.
[41] M. Nikolova, "Estimation of binary images by minimizing convex criteria," in *Proc. IEEE Int. Conf. Image Process.*, vol. 2, 1998, pp. 108–112.
[42] M. Nikolova, "Markovian reconstruction using a GNC approach," *IEEE Trans. Image Process.*, vol. 8, no. 9, pp. 1204–1220, Sep. 1999.
[43] M. Nikolova, "Energy minimization methods," in *Handbook of Mathematical Methods in Imaging*, O. Scherzer, Ed. New York, NY, USA: Springer, 2011, pp. 138–186, ch. 5.
[44] M. Nikolova, M. K. Ng, and C.-P. Tam, "Fast nonconvex nonsmooth minimization methods for image restoration and reconstruction," *IEEE Trans. Image Process.*, vol. 19, no. 12, pp. 3073–3088, Dec. 2010.
[45] A. Parekh and I. W. Selesnick, "Enhanced low-rank matrix approximation," *IEEE Signal Process. Lett.*, vol. 23, no. 4, pp. 493–497, Apr. 2016.
[46] N. Parikh and S. Boyd, "Proximal algorithms," *Foundations Trends Optim.*, vol. 1, no. 3, pp. 123–231, 2014.
[47] J. Portilla and L. Mancera, "L0-based sparse approximation: two alternative methods and some applications," *Proc. SPIE*, vol. 6701, 2007, Art. no. 67011Z.
[48] L. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Phys. D*, vol. 60, pp. 259–268, 1992.
[49] I. Selesnick, "Sparsity amplified," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, New Orleans, LA, USA, Mar. 2017, pp. 4356–4360.
[50] I. Selesnick, "Sparsity-assisted signal smoothing (revisited)," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, New Orleans, LA, USA, Mar. 2017, pp. 4546–4550.
[51] I. Selesnick, "Total variation denoising via the Moreau envelope," *IEEE Signal Process. Lett.*, vol. 24, no. 2, pp. 216–220, Feb. 2017.
[52] I. Selesnick and M. Farshchian, "Sparse signal approximation via nonseparable regularization," *IEEE Trans. Signal Process.*, vol. 65, no. 10, pp. 2561–2575, May 2017.
[53] I. W. Selesnick, "Sparsity-assisted signal smoothing," in *Excursions in Harmonic Analysis*, vol. 4, R. Balan, Ed. Basel, Switzerland: Birkhäuser, 2015, pp. 149–176.
[54] I. W. Selesnick and I. Bayram, "Sparse signal estimation by maximally sparse convex optimization," *IEEE Trans. Signal Process.*, vol. 62, no. 5, pp. 1078–1092, Mar. 2014.
[55] I. W. Selesnick and I. Bayram, "Enhanced sparsity by non-separable regularization," *IEEE Trans. Signal Process.*, vol. 64, no. 9, pp. 2298–2313, May 2016.
[56] I. W. Selesnick, A. Parekh, and I. Bayram, "Convex 1-D total variation denoising with non-convex regularization," *IEEE Signal Process. Lett.*, vol. 22, no. 2, pp. 141–144, Feb. 2015.
[57] E. Soubies, L. Blanc-Féraud, and G. Aubert, "A continuous exact ℓ_0 penalty (CEL0) for least squares regularized problem," *SIAM J. Imag. Sci.*, vol. 8, no. 3, pp. 1607–1639, 2015.
[58] C. Soussen, J. Idier, J. Duan, and D. Brie, "Homotopy based algorithms for ℓ_0-regularized least-squares," *IEEE Trans. Signal Process.*, vol. 63, no. 13, pp. 3301–3316, Jul. 2015.
[59] J.-L. Starck, F. Murtagh, and J. Fadili, *Sparse Image and Signal Processing: Wavelets and Related Geometric Multiscale Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 2015.
[60] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Mach. Learning Res.*, vol. 1, pp. 211–244, 2001.
[61] S. Vaiter, C. Deledalle, J. Fadili, G. Peyré, and C. Dossal, "The degrees of freedom of partly smooth regularizers," *Ann. Inst. Stat. Math.*, pp. 1–42, 2016, doi: 10.1007/s10463-016-0563-z.
[62] S. Voronin and R. Chartrand, "A new generalized thresholding algorithm for inverse problems with sparsity constraints," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2013, pp. 1636–1640.
[63] D. P. Wipf, B. D. Rao, and S. Nagarajan, "Latent variable Bayesian models for promoting sparsity," *IEEE Trans. Inform. Theory*, vol. 57, no. 9, pp. 6236–6255, Sep. 2011.
[64] J. Woodworth and R. Chartrand, "Compressed sensing recovery via nonconvex shrinkage penalties," *Inverse Problems*, vol. 32, no. 7, pp. 75004–75028, Jul. 2016.
[65] C.-H. Zhang, "Nearly unbiased variable selection under minimax concave penalty," *Ann. Statist.*, vol. 38, pp. 894–942, 2010.
[66] H. Zou and R. Li, "One-step sparse estimates in nonconcave penalized likelihood models," *Ann. Statist.*, vol. 36, no. 4, pp. 1509–1533, 2008.

Ivan Selesnick (S'91–M'98–SM'08–F'16) received the B.S., M.E.E., and Ph.D. degrees in electrical engineering from Rice University, Houston, TX, USA, in 1990, 1991, and 1996, respectively. In 1997, he was a Visiting Professor at the University of Erlangen-Nurnberg, Germany. He then joined the Department of Electrical and Computer Engineering, Polytechnic University, Brooklyn, New York, NY, USA (now the Tandon School of Engineering, New York University), where he is currently a Professor. His research interests include signal and image processing, wavelet-based signal processing, sparsity techniques, and biomedical signal processing. He has been an Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE SIGNAL PROCESSING LETTERS, and IEEE TRANSACTIONS ON SIGNAL PROCESSING.