Journal of Scientific Computing (2021) 87:31
https://doi.org/10.1007/s10915-021-01443-w

A Novel Regularization Based on the Error Function for Sparse Recovery

Weihong Guo[1] · Yifei Lou[2] · Jing Qin[3] · Ming Yan[4]

Received: 28 April 2020 / Revised: 11 January 2021 / Accepted: 23 February 2021 / Published online: 6 March 2021

Abstract
Regularization plays an important role in solving ill-posed problems by adding extra information about the desired solution, such as sparsity. Many regularization terms usually involve some vector norms. This paper proposes a novel regularization framework that uses the error function to approximate the unit step function. It can be considered as a surrogate function for the L_0 norm. The asymptotic behavior of the error function with respect to its intrinsic parameter indicates that the proposed regularization can approximate the standard L_0, L_1 norms as the parameter approaches to 0 and ∞, respectively. Statistically, it is also less biased than the L_1 approach. Incorporating the error function, we consider both constrained and unconstrained formulations to reconstruct a sparse signal from an under-determined linear system. Computationally, both problems can be solved via an iterative reweighted L_1 (IRL1) algorithm with guaranteed convergence. A large number of experimental results demonstrate that the proposed approach outperforms the state-of-the-art methods in various sparse recovery scenarios.

WG was partially supported by NSF DMS-1521582. YL was partially supported by NSF CAREER 1846690. JQ was partially supported by NSF DMS-1941197. MY was partially supported by NSF DMS-2012439. The MATLAB codes of this manuscript will be available under https://sites.google.com/site/louyifei/Software after publication.

✉ Yifei Lou
yifei.lou@utdallas.edu

Weihong Guo
wxg49@case.edu

Jing Qin
jing.qin@uky.edu

Ming Yan
yanm@math.msu.edu

[1] Department of Mathematics, Applied Mathematics and Statistics, Case Western Reserve University, Cleveland, OH 44106, USA

[2] Department of Mathematical Sciences, The University of Texas at Dallas, Richardson, TX 75080, USA

[3] Department of Mathematics, University of Kentucky, Lexington, KY 40506, USA

[4] Department of Computational Mathematics, Science and Engineering, Department of Mathematics, Michigan State University, East Lansing, MI 48824, USA

Keywords Error function · Iterative reweighted L_1 · Compressed sensing · Sparsity · Biaseness

Mathematics Subject Classification 49N45 · 90C05 · 90C26

1 Introduction

These days, "big data" is ubiquitous due to the developments and advancements in science and technologies. On the other hand, one often faces "small data," when technical or economic restrictions limit the amount of data that can be collected and/or transmitted. For example, when a patient undergoes the computed tomography (CT) scanning, the number of measurements can not exceed the maximum safe radiation dosage. Mathematically speaking, a small data problem corresponds to an under-determined (linear) system, where the number of measurements is considerably smaller than the ambient dimension. In this case, reasonable assumptions shall be taken into account and formulated as regularizations to obtain the desired solution.

In data science, a signal of interest is often assumed to be sparse (i.e., having a few nonzero elements) either by itself or after a linear transformation. One popular signal recovery technique based on sparsity is *compressed sensing* (CS), coined by David Donoho [11]. CS enables data compression to facilitate data storage and transmission, as only a small portion of data, which are nonzero, is being processed. In order to find the sparsest signal, it is natural to minimize the L_0 norm[1], i.e., the number of nonzero entries in a vector. Unfortunately, the L_0 minimization problem is NP-hard [24], as it involves combinatorial search that is time-consuming, especially in high-dimensional spaces. One of the most popular approaches in CS is to replace the L_0 norm by the convex L_1 norm, which often gives satisfactory results. This L_1 convex relaxation technique has been applied in many different fields such as geology and geophysics [30], Fourier transform spectroscopy [23], and ultrasound imaging [26]. One major tool for analyzing CS algorithms is the restricted isometry property (RIP) [6], which provides a sufficient condition for the exact recovery of a sparse signal by minimizing the L_1 norm.

The L_1 minimization in CS is closely related to the least absolute shrinkage and selection operator (LASSO) [32] in statistical learning. Assume that the data is generated by a linear regression model polluted by Gaussian noise, $\mathbf{b} = \mathbf{A}\mathbf{x} + \varepsilon$, where each row of $\mathbf{A}$ is a sample of feature vectors, and $\mathbf{b}$, ε are the response and noise, respectively. In this setting, one aims to find a sparse vector $\mathbf{x}$ consisting of model coefficients. It is a reasonable assumption since only a few features contribute to the response. However, Fan and Li [13] pointed out that LASSO (or L_1) produces biased estimates for large coefficients. Note that there is no common ground for various definitions and interpretations of biasedness. Here, we follow a sufficient condition from [13] to characterize unbiasedness: $y = f(x)$ is unbiased if its derivative is zero for large magnitude values of x. In addition, unbiasedness can be characterized by the closeness of the proximal operator to the straight line $y = x$ when $|x|$ increases. To mitigate the estimation bias, Fan and Li [13] proposed a nonconvex regularization, called smoothly clipped absolute deviation (SCAD). Later, many other nonconvex regularizations have emerged in statistics, such as capped L_1 (CL1) [31,44], transformed L_1 (TL1) [22], and minimax concave penalty (MCP) [41]. Some of these models have been adopted for sparse signal recovery [21,42,43].

[1] Note that $\|\cdot\|_0$ is a pseudo-norm, but is often called as the L_0 norm.

Nonconvex regularization terms can be further categorized into two groups: smooth and nonsmooth. In particular, CL1, SCAD, and MCP are nonsmooth. Specifically, their proximal functions are not differentiable, which leads to numerical instability from the algorithmic point of view. Transformed L_1 is smooth except at zero, and its proximal function is also smooth. However, it causes more bias than CL1, SCAD, and MCP, as its proximal operator does not coincide with $y = x$ for large values of $|x|$. We aim to propose a nonconvex regularization, which is smooth and less biased compared to L_1 and TL1; see Fig. 1 for comparison among these regularization terms.

In this paper, we propose a novel nonconvex regularization based on the ERror Function (ERF) to promote sparsity. It is motivated from a graph-based approach [3] to enforce a bi-modal weight distribution when reconstructing a skeleton image. We discover that their numerical scheme is equivalent to minimizing the error function via the iterative reweighted L_1 (IRL1) algorithm [7]. As a good approximation to the Heaviside step function (or the unit step function), the error function can serve as a surrogate function for the L_0 norm, which has not been considered in the CS literature to the best of our knowledge. The major contributions of this paper are three-fold:

(a) We propose a novel regularization based on the error function for sparse signal recovery and establish its connections to the standard L_0, L_1 regularizations;
(b) We adapt the IRL1 algorithms to solve the proposed model in both constrained and unconstrained formulations with guaranteed convergence;
(c) We conduct extensive experiments to demonstrate the superior performance of the proposed approaches.

The rest of the paper is organized as follows. We review some existing models and related algorithms for sparse recovery in Sect. 2. The proposed regularization is described in Sect. 3, followed by numerical schemes in Sect. 4. We present experimental results in Sect. 5, showing that the proposed approaches outperform the state-of-the-art methods in sparse recovery. Finally, conclusions and future works are given in Sect. 6.

2 Preliminaries

Throughout the paper, we use bold uppercase letters to denote matrices, bold lowercase letters to denote vectors, and lowercase letters to denote vector or matrix entries, e.g., a vector $\mathbf{x}$ with its j-th component x_j. The set of all n-dimensional real vectors is denoted by $\mathbb{R}^n$, and the set of all nonnegative vectors is denoted by $\mathbb{R}^n_+ := [0, \infty)^n$. The L_p norm of a vector $\mathbf{x} \in \mathbb{R}^n$ is defined as $\|\mathbf{x}\|_p = (\sum_{j=1}^n |x_j|^p)^{1/p}$ for $0 < p \leq \infty$. The sign function applied to $\mathbf{x} \in \mathbb{R}^n$ returns a vector, denoted by $\text{sign}(\mathbf{x})$, whose j-th component is $x_j/|x_j|$ if $x_j \neq 0$ and zero otherwise. Inequalities involving vectors are defined component-wise, e.g., $\mathbf{x} \leq \mathbf{y}$ means that each component of $\mathbf{x}$ is less than or equal to the corresponding component of $\mathbf{y}$. We use $\odot$ to denote the component-wise multiplication of two vectors. The set of all $m \times n$ real matrices is denoted by $\mathbb{R}^{m\times n}$. The kernel of a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ is defined as $\ker(\mathbf{A}) := \{\mathbf{x} \in \mathbb{R}^n : \mathbf{A}\mathbf{x} = \mathbf{0}\}$. $\mathbf{A}_S$ is the submatrix with columns selected from the index set S, $\mathbf{x}_S$ is the subvector of $\mathbf{x}$ with components selected from the index set S, and $\overline{\mathbf{x}_S}$ is the vector obtained by zeroing out the components of $\mathbf{x}$ in S^c – the complement of the index set S.

2.1 Sparsity-Promoting Models

There are a variety of regularizers that can approximate the L_0 norm, including L_1, L_p with $0 < p < 1$ [9,36,37], capped L_1 [21,31,44], transformed L_1 [22,42,43], L_1-L_2 [20,28,39], and L_1/L_2 [29,35]. In this paper, we focus on developing a separable regularization, which allows a component-wise implementation to enhance computational efficiency. Besides L_1 and the pth power of L_p with $0 < p < 1$, there are other popular separable regularizations, some of which are listed as follows: for any $\mathbf{x} \in \mathbb{R}^n$ and $a > 0,\ \lambda > 0,\ \gamma > 1$,

– Capped L_1 (CL1):

$$J_a^{\mathrm{CL1}}(\mathbf{x}) := \sum_{j=1}^{n} \min\{|x_j|, a\};$$

– Transformed L_1 (TL1):

$$J_a^{\mathrm{TL1}}(\mathbf{x}) := \sum_{j=1}^{n} \frac{(a+1)|x_j|}{a+|x_j|};$$

– Smoothly clipped absolute deviation (SCAD):

$$J_{\lambda,\gamma}^{\mathrm{SCAD}}(\mathbf{x}) := \sum_{j=1}^{n} \Phi_{\lambda,\gamma}^{\mathrm{SCAD}}(x_j)$$

with

$$\Phi_{\lambda,\gamma}^{\mathrm{SCAD}}(x_j) = \begin{cases} \lambda|x_j|, & |x_j| \le \lambda; \\ \frac{2\gamma\lambda|x_j| - |x_j|^2 - \lambda^2}{2(\gamma-1)}, & \lambda < |x_j| \le \gamma\lambda; \\ \frac{(\gamma+1)\lambda^2}{2}, & |x_j| > \gamma\lambda; \end{cases}$$

– Minimax concave penalty (MCP):

$$J_{\lambda,\gamma}^{\mathrm{MCP}}(\mathbf{x}) := \sum_{j=1}^{n} \Phi_{\lambda,\gamma}^{\mathrm{MCP}}(x_j)$$

with

$$\Phi_{\lambda,\gamma}^{\mathrm{MCP}}(x_j) = \begin{cases} \lambda|x_j| - \frac{|x_j|^2}{2\gamma}, & |x_j| \le \gamma\lambda; \\ \frac{1}{2}\gamma\lambda^2, & |x_j| > \gamma\lambda; \end{cases}$$

It is straightforward that the pth power of L_p converges to L_0 and L_1 as p goes to 0 and 1, respectively. By letting $a = 0$ in TL1 and using the standard assumption of $\frac{0}{0} = 0$, J_0^{TL1} is equivalent to the L_0 norm. On the other hand, we have

$$\lim_{a\to\infty} J_a^{\mathrm{TL1}}(\mathbf{x}) = \sum_{j=1}^{n} \lim_{a\to\infty} \frac{(1+\frac{1}{a})|x_j|}{1+\frac{1}{a}|x_j|} = \sum_{j=1}^{n} |x_j| = \|\mathbf{x}\|_1.$$

Therefore, J_a^{TL1} approaches to the L_0 and L_1 norms by letting $a \to 0$ and ∞, respectively.

Both SCAD and MCP are proposed to correct the estimation bias for large coefficients caused by the L_1 approach. As $\Phi_{\lambda,\gamma}^{\mathrm{SCAD}}$ and $\Phi_{\lambda,\gamma}^{\mathrm{MCP}}$ become constant for $|x| > \gamma\lambda$, both SCAD and MCP estimates are unbiased. However, one major drawback of SCAD and MCP is that there are two model parameters involved, which causes difficulties in the parameter

selection. The parameter-free models include L_1, L_1-L_2, and L_1/L_2, the last of which also has a scale-invariant property to mimic the L_0 norm.

2.2 Optimization Techniques

A fundamental problem in CS is to find a sparse vector subject to an under-determined linear system,

$$\min_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{x}\|_0 \quad \text{s.t.} \quad \mathbf{A}\mathbf{x} = \mathbf{b}, \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{m\times n}(m \ll n)$ is called a *sensing matrix* and $\mathbf{b} \in \mathbb{R}^m$ denotes a measurement vector. Candès *et al.* [7] proposed the following reweighted L_1 minimization (IRL1),

$$\begin{cases} w_j^k = \frac{1}{|x_j^k|+\epsilon}, \text{ for } j = 1, \ldots, n, \\ \mathbf{x}^{k+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \sum_{j=1}^n w_j^k |x_j|, \end{cases} \tag{2}$$

where ϵ is a positive parameter for the sake of stability and $\mathcal{X} = \{\mathbf{x} : \mathbf{A}\mathbf{x} = \mathbf{b}\}$ denotes the feasible set. From the perspective of a majorization-minimization (MM) framework [19], the iteration (2) is in fact to minimize the following problem

$$\min_{\mathbf{x}\in\mathcal{X}} \sum_{j=1}^n \log(|x_j| + \epsilon). \tag{3}$$

The objective function in (3) is often called a *log-sum penalty function*, denoted by

$$J_\epsilon^{\log}(\mathbf{x}) := \sum_{j=1}^n \Phi_\epsilon^{\log}(|x_j|) \quad \text{where} \quad \Phi_\epsilon^{\log}(x) = \log(x + \epsilon).$$

Since $\Phi_\epsilon^{\log}$ is an increasing concave function on $[0, +\infty)$, we have

$$J_\epsilon^{\log}(\mathbf{x}) \le J_\epsilon^{\log}(\mathbf{x}^k) + \langle \nabla J_\epsilon^{\log}(\mathbf{x}^k), |\mathbf{x}| - |\mathbf{x}^k| \rangle.$$

Note note that $|\cdot|$ means to take the absolute value for each component of a vector. Instead of directly minimizing $J_\epsilon^{\log}$, the MM framework considers the following iteration scheme

$$\mathbf{x}^{k+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} J_\epsilon^{\log}(\mathbf{x}^k) + \langle \nabla J_\epsilon^{\log}(\mathbf{x}^k), |\mathbf{x}| - |\mathbf{x}^k| \rangle,$$

which is equivalent to (2).

The iterative reweighed algorithm is generalized in [25], where the authors considered a certain class of nonsmooth and nonconvex functions of the form

$$\min_{\mathbf{x}\in\mathcal{X}} F_1(\mathbf{x}) + F_2(G(\mathbf{x})), \tag{4}$$

with a convex function F_1, a coordinate-wise convex function G, a concave function F_2, and a feasible set $\mathcal{X} \subseteq \mathbb{R}^n$. The iterative reweighted algorithm can be expressed as

$$\begin{cases} \mathbf{w}^k \in \partial F_2(\mathbf{y}) \quad \text{with} \quad \mathbf{y} = G(\mathbf{x}^k) \\ \mathbf{x}^{k+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} F_1(\mathbf{x}) + \langle \mathbf{w}^k, G(\mathbf{x}) \rangle, \end{cases} \tag{5}$$

where ∂F_2 denotes the subdifferential of F_2. If $F_1 = 0$, $F_2(\mathbf{y}) = \sum_{j=1}^n \Phi_\epsilon^{\log}(y_j)$, and $G(\mathbf{x}) = |\mathbf{x}|$, then (5) reduces to (2).

3 Proposed Regularization

We propose the following novel regularization to promote sparsity:

$$J_\sigma^{\mathrm{ERF}}(\mathbf{x}) := \sum_{j=1}^{n} \Phi_\sigma^{\mathrm{ERF}}(|x_j|) \quad \text{with} \quad \Phi_\sigma^{\mathrm{ERF}}(x) = \int_0^x e^{-\tau^2/\sigma^2} d\tau, \tag{6}$$

with $\sigma > 0$. Note that the standard error function is defined as

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-\tau^2} d\tau,$$

and hence $\Phi_\sigma^{\mathrm{ERF}}$ is a scaled error function in that

$$\Phi_\sigma^{\mathrm{ERF}}(x) = \sigma \int_0^{\frac{x}{\sigma}} e^{-\tau^2} d\tau = \frac{\sigma\sqrt{\pi}}{2} \mathrm{erf}\left(\frac{x}{\sigma}\right). \tag{7}$$

We refer our model (6) as the "ERF" regularization. In what follows, we omit the superscript "ERF" in J_σ and Φ_σ when the context clearly refers to the proposed regularization.

3.1 Properties

We list some useful properties about Φ_σ and J_σ, especially the asymptotical behaviors of J_σ as characterized in Theorem 1.

(P1) The derivative of Φ_σ at x is given by

$$\frac{d}{dx}\Phi_\sigma(x) = \exp(-\frac{x^2}{\sigma^2}). \tag{8}$$

(P2) The upper/lower bounds of Φ_σ are given by

$$c\sqrt{1 - e^{-ax^2}} \le \Phi_\sigma(x) \le c\sqrt{1 - e^{-bx^2}}, \quad \forall\, x \in \mathbb{R}, \tag{9}$$

where $a = 1/\sigma^2$, $b = \pi/4\sigma^2$, and $c = \frac{\sigma\sqrt{\pi}}{2}$ based on the lower and upper bounds of the standard error function [10].

(P3) J_σ is concave on $\mathbb{R}^n_+$, i.e., for any $t \in [0, 1]$ and any $\mathbf{x},\ \mathbf{y} \in \mathbb{R}^n_+$

$$J_\sigma(t\mathbf{x} + (1-t)\mathbf{y}) \ge tJ_\sigma(\mathbf{x}) + (1-t)J_\sigma(\mathbf{y}). \tag{10}$$

(P4) J_σ is subadditive or satisfies the triangle inequality on $\mathbb{R}^n$, i.e.,

$$J_\sigma(\mathbf{x} + \mathbf{y}) \le J_\sigma(\mathbf{x}) + J_\sigma(\mathbf{y}), \quad \forall\, \mathbf{x}, \mathbf{y} \in \mathbb{R}^n. \tag{11}$$

In addition, if $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ have disjoint supports, then

$$J_\sigma(\mathbf{x} + \mathbf{y}) = J_\sigma(\mathbf{x}) + J_\sigma(\mathbf{y}).$$

The proofs of (P3)-(P4) can be found in the Appendix.

Theorem 1 *For any nonzero vector $\boldsymbol{x} \in \mathbb{R}^n$, we have*

(a) $J_\sigma(\boldsymbol{x}) \to \|\boldsymbol{x}\|_1$, *as* $\sigma \to +\infty$;
(b) $J_\sigma(\boldsymbol{x})/\sigma \to \frac{\sqrt{\pi}}{2}\|\boldsymbol{x}\|_0$, *as* $\sigma \to 0^+$.

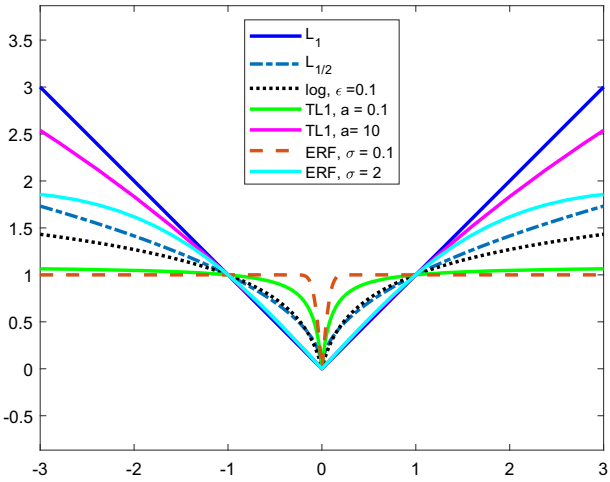

Fig. 1 The objective functions of various sparsity promoting models, all of which are scaled to attain the point (1, 1). The proposed ERF gives the best approximation to the L_0 norm for a small value of σ and is also less "biased" compared to other models

Proof Since $J_\sigma(\cdot)$ is separable with respect to each component of $\mathbf{x}$, it suffices to discuss the limits for a scalar. For any component $x_j \neq 0$, we let $t = |x_j|/\sigma$ which approaches zero as $\sigma \to +\infty$ and hence we have

$$\lim_{\sigma \to +\infty} \frac{\Phi_\sigma(|x_j|)}{|x_j|} = \lim_{t \to 0} \frac{\int_0^t e^{-\tau^2} d\tau}{t} = 1. \tag{12}$$

The last equality is obtained by applying the l'Hospital's Rule. When $x_j = 0$, it is obvious that $\Phi_\sigma(|x_j|) = 0$, and thereby $J_\sigma(\mathbf{x}) \to \|\mathbf{x}\|_1$ as $\sigma \to +\infty$.

On the other hand, we have $\Phi_\sigma(0) = 0$. When $x_j \neq 0$, we have

$$\lim_{\sigma \to 0^+} \frac{\Phi_\sigma(|x_j|)}{\sigma} = \frac{\sqrt{\pi}}{2}. \tag{13}$$

Therefore, $J_\sigma(\mathbf{x})/\sigma \to \frac{\sqrt{\pi}}{2}\|\mathbf{x}\|_0$ as $\sigma \to 0^+$. □

Figure 1 shows the objective functions of various sparsity promoting models. We scale them to attain the point (1, 1) in order to have a better visual comparison. It indicates that the proposed ERF gives the best approximation to the L_0 norm for a small value of σ and is also less "biased" compared to other models.

3.2 Proximal Operator

Given a function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, the proximal operator [27] $\mathbf{prox}_\mu^f : \mathbb{R}^n \to \mathbb{R}^n$ of f with a parameter $\mu > 0$ is defined by

$$\mathbf{prox}_\mu^f(\mathbf{v}) = \arg\ \min_{\mathbf{x}} \left(\mu f(\mathbf{x}) + \frac{1}{2}\|\mathbf{x} - \mathbf{v}\|_2^2 \right). \tag{14}$$

If f is the L_1 norm, then the corresponding proximal operator is the *soft shrinkage* operator, defined by

$$\text{shrink}_\mu(\mathbf{v}) = \begin{cases} \mathbf{v} - \mu, & \mathbf{v} > \mu, \\ \mathbf{0}, & |\mathbf{v}| \le \mu, \\ \mathbf{v} + \mu, & \mathbf{v} < -\mu. \end{cases}$$

Due to its component-wise calculation, this operator is a key to make many L_1 minimization algorithms efficient. The proximal operator of the L_0 norm with parameter μ is given by the *hard thresholding*

$$\text{thresh}_\mu(\mathbf{v}) = \begin{cases} \mathbf{v}, & |\mathbf{v}| > \mu, \\ 0, & |\mathbf{v}| \le \mu. \end{cases}$$

As for TL1, its proximal operator [42] can be expressed as

$$\mathbf{prox}_\mu^{\text{TL1}}(\mathbf{v}) = \begin{cases} \text{sign}(\mathbf{v})\left[\frac{2}{3}(a + |\mathbf{v}|)\cos(\frac{\varphi(\mathbf{v})}{3}) - \frac{2a}{3} + \frac{|\mathbf{v}|}{3}\right], & \mathbf{v} > \mu, \\ 0, & |\mathbf{v}| \le \mu, \end{cases}$$

where $\varphi(\mathbf{v}) = \arccos\left(1 - \frac{27\mu a(a+1)}{2(a+|\mathbf{v}|^3)}\right)$.

Next we derive the proximal operator of the proposed ERF model, which is defined by

$$\mathbf{prox}_\mu^{\text{ERF}}(\mathbf{v}) = \arg\ \min_{\mathbf{x}} \left(\mu J_\sigma^{\text{ERF}}(\mathbf{x}) + \frac{1}{2}\|\mathbf{x} - \mathbf{v}\|_2^2\right). \tag{15}$$

Note that there is no closed-form solution for this problem. We can find the solution via the Newton's method. In particular, the optimality condition of (15) reads as

$$\mathbf{v} \in \mu \partial J_\sigma^{\text{ERF}}(\mathbf{x}) + \mathbf{x} = \mu \exp\left(-\frac{\mathbf{x}^2}{\sigma^2}\right) \odot \partial\,|\mathbf{x}| + \mathbf{x}.$$

When $|v_i| \le \mu$, we have $x_i = 0$. Otherwise the optimality condition becomes

$$v_i = \mu \exp\left(-\frac{x_i^2}{\sigma^2}\right)\text{sign}(v_i) + x_i,$$

which can be found by the Newton's method.

We plot the proximal operators for L_0, L_1, TL1 with $a = 0.1, 10$, and ERF with $\sigma = 0.1, 2$ in Fig. 2. Both TL1 and ERF provide the asymptotic approximations to L_0 and L_1 when varying their intrinsic parameter. The bias issue can be illustrated by whether the proximal operator approaches the line $y = x$ when the magnitude of x increases. In this sense, the plots indicate that ERF causes less bias than L_1 and TL1.

3.3 Exact Recovery Guarantee

Based on the subadditive property, we adopt a generalized null space property (gNSP) [34] to prove that the proposed ERF model is guaranteed to exactly find the desired sparse solution.

Definition 1 A matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ is said to satisfy a generalized null space property (gNSP) relative to a function $J : \mathbb{R}^n \to [0, \infty)$ and $S \subseteq \{1, \cdots, n\}$ if

$$J(\overline{\mathbf{v}_S}) < J(\overline{\mathbf{v}_{S^c}}) \tag{16}$$

for all $\mathbf{v} \in \ker(\mathbf{A})\backslash\{\mathbf{0}\}$. Here $\overline{\mathbf{v}_S}$ and $\overline{\mathbf{v}_{S^c}}$ are obtained by zeroing out the components of $\mathbf{v}$ indexed by S^c and S, respectively. The matrix $\mathbf{A}$ is said to satisfy gNSP of order $s \le n$ relative to J if it satisfies gNSP relative to J and any set $S \subseteq \{1, \cdots, n\}$ with $|S| \le s$.

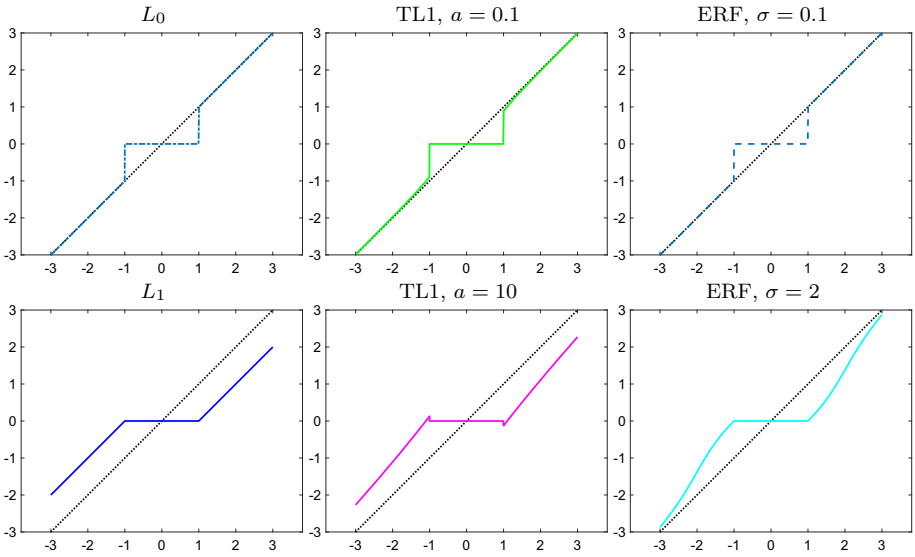

Fig. 2 Proximal operators of various sparsity promoting models with $\mu = 1$

If we replace J in (16) by the L_1 norm, then Definition 1 becomes the standard NSP for the exact L_1 recovery [12]. We establish in Theorem 3 an exact recovery of the proposed ERF regularization based on gNSP. To make the paper self-contained, we include a brief proof, which follows the NSP discussion of the L_1 norm [15, Theorem 4.4]. In addition, since the proposed J_σ is separable on $\mathbb{R}^n$, concave on $\mathbb{R}^n_+$, and $J_\sigma(\mathbf{0}) = 0$, Theorem 3 can be obtained in a similar way as [34, Theorem 4.3].

Theorem 2 *Given a matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and $\sigma > 0$, every vector $\boldsymbol{x} \in \mathbb{R}^n$ supported on a set $S \subseteq \{1, \cdots, n\}$ is the unique solution of the problem*

$$\min_z J_\sigma(z) \quad s.t. \quad \boldsymbol{Az} = \boldsymbol{b},$$

if and only if A satisfies gNSP relative to J_σ and S.

Proof Given a fixed index set S, let us first assume that every vector $\mathbf{x} \in \mathbb{R}^n$ supported on S is the unique minimizer of $\min_{\mathbf{z}} J_\sigma(\mathbf{z})$ subject to $\mathbf{Az} = \mathbf{Ax}$. Thus for any $\mathbf{v} \in \ker(\mathbf{A})\backslash\{\mathbf{0}\}$, the vector $\overline{\mathbf{v}_S}$ is the unique minimizer of $J_\sigma(\mathbf{z})$ subject to $\mathbf{Az} = \mathbf{A}\overline{\mathbf{v}_S}$. Since $\mathbf{Av} = \mathbf{0}$ with $\mathbf{v} \neq \mathbf{0}$, we have

$$\mathbf{A}(-\overline{\mathbf{v}_{S^c}}) = \mathbf{A}\overline{\mathbf{v}_S}.$$

Since $-\overline{\mathbf{v}_{S^c}} \neq \overline{\mathbf{v}_S}$, we can get the inequality (16), which establishes gNSP relative to J_σ and S.

Conversely, we assume that (16) holds for J_σ and S. For $\mathbf{x} \in \mathbb{R}^n$ supported on S and a vector $\mathbf{z} \in \mathbb{R}^n$ with $\mathbf{z} \neq \mathbf{x}$ and $\mathbf{Az} = \mathbf{Ax}$, we have $\mathrm{supp}(\mathbf{x} - \overline{\mathbf{z}_S}) = S$. Then the vector $\mathbf{v} = \mathbf{x} - \mathbf{z} \in \ker(\mathbf{A})\backslash\{\mathbf{0}\}$ satisfies that $\overline{\mathbf{v}_S} = \mathbf{x} - \overline{\mathbf{z}_S}$. Furthermore, due to the subadditive property, we obtain

$$\begin{aligned} J_\sigma(\mathbf{x}) &\leq J_\sigma(\mathbf{x} - \overline{\mathbf{z}_S}) + J_\sigma(\overline{\mathbf{z}_S}) = J_\sigma(\overline{\mathbf{v}_S}) + J_\sigma(\overline{\mathbf{z}_S}) \\ &< J_\sigma(\overline{\mathbf{v}_{S^c}}) + J_\sigma(\overline{\mathbf{z}_S}) = J_\sigma(-\overline{\mathbf{z}_{S^c}}) + J_\sigma(\overline{\mathbf{z}_S}) \end{aligned}$$

$$= J_\sigma(\overline{\mathbf{z}_{S^c}}) + J_\sigma(\overline{\mathbf{z}_S}) = J_\sigma(\mathbf{z}),$$

which implies that $\mathbf{x}$ is the unique minimizer of $J_\sigma(\mathbf{z})$ subject to $\mathbf{Az} = \mathbf{Ax}$. □

By varying the support sets S, we can get the following theorem for necessary and sufficient conditions of the exact sparse recovery.

Theorem 3 *Given a matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and $\sigma > 0$, every s-sparse vector $\boldsymbol{x} \in \mathbb{R}^n$ is the unique solution of the problem*

$$\min_{z} J_\sigma(z) \quad s.t. \quad \boldsymbol{Az} = \boldsymbol{b},$$

if and only if $\boldsymbol{A}$ satisfies gNSP of order s relative to J_σ.

It is NP-hard to determine whether a matrix satisfies NSP or not [33]. However, the NP-hardness of gNSP for a non-L_1 regularization still remains an open question. On the other hand, if we relax "every s-sparse vector," then gNSP is no longer necessary in Theorem 3 for exact sparse recovery.

4 Algorithms

We follow the reweighted L_1 approach in [7] to minimize the proposed regularization J_σ. We shall discuss two optimization formulations: constrained and unconstrained, separately.

4.1 Constrained Formulation

Consider an ERF-regularized minimization problem with a linear constraint

$$\min_{\mathbf{x}\in\mathbb{R}^n} J_\sigma(\mathbf{x}) \quad \text{s.t.} \quad \mathbf{Ax} = \mathbf{b}. \tag{17}$$

According to the general iterative reweighted framework (5), the objective function in (17) can be expressed as $J_\sigma(\mathbf{x}) = F_2(G(\mathbf{x}))$, where $F_2(\mathbf{y}) = \sum_{j=1}^n \Phi_\sigma(y_j)$ and $G(\mathbf{x}) = |\mathbf{x}|$. By calculating the derivative of the error function (8), we obtain the following iterative scheme

$$\begin{cases} w_j^k = \exp\left(-\left(\frac{x_j^k}{\sigma}\right)^2\right), & \text{for } j = 1,\ldots,n, \\ \mathbf{x}^{k+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \sum_{j=1}^n w_j^k |x_j|, \end{cases} \tag{18}$$

where $\mathcal{X}$ is the same feasible set as in (2). The $\mathbf{x}$-subproblem in (18) can be cast as a linear programming problem. We use the commercial Gurobi solver (https://www.gurobi.com/) to solve this subproblem. Other popular methods include alternating direction method of multipliers (ADMM) [4,16,17] and primal-dual algorithms [8,38]. The convergence of the scheme (18) is presented in Theorem 4.

Theorem 4 *The sequence $\{\boldsymbol{x}^k\}_{k=1}^\infty$ generated by the reweighted L_1 iteration (18) is bounded. It has a convergent subsequence and any accumulation point of $\{\boldsymbol{x}^k\}_{k=1}^\infty$ is a stationary point of (17). In addition, we have*

$$\min_{k=1,\cdots,K} \sum_{j=1}^n \left[\Phi_\sigma(|x_j^k|) - \Phi_\sigma(|x_j^{k+1}|) - w_j^k\big(|x_j^k| - |x_j^{k+1}|\big)\right] \le \frac{J_\sigma(\boldsymbol{x}^1)}{K}.$$

Proof We start by showing that the sequence $\{\mathbf{x}^k\}_{k=1}^{\infty}$ is bounded. In particular, we aim to show that the sequence of $\{\mathbf{x}^k\}_{k=1}^{\infty}$ is in the convex hull constructed by the set $\mathcal{V} = \{\mathbf{x} : \exists\, S$ such that $\mathbf{A}_S$ has full column rank, $\mathbf{A}_S\mathbf{x}_S = \mathbf{b}$, $\mathbf{x}_{S^c} = \mathbf{0}\}$. Since the number of submatrices of $\mathbf{A}$ with full column rank is finite, $\mathcal{V}$ is bounded, leading to boundedness of the convex hull and the sequence $\{\mathbf{x}^k\}$.

Let $\bar{\mathbf{x}} \in \mathbb{R}^n$ be an optimal solution for the $\mathbf{x}$-subproblem in (18) and S the set of corresponding indices of nonzero components in $\bar{\mathbf{x}}$. Given an arbitrary set of nonnegative weights, denoted by w_j, we consider the following problem restricted to the index set S,

$$\min_{\mathbf{x}\in\mathbb{R}^n} \sum_{j\in S} w_j|x_j|, \quad \text{s.t.} \quad \mathbf{A}_S\mathbf{x}_S = \mathbf{b},\ \mathbf{x}_{S^c} = \mathbf{0}. \tag{19}$$

The optimality condition of (19) is

$$\begin{bmatrix} \mathbf{M} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_S^\top \end{bmatrix} \begin{bmatrix} \bar{\mathbf{x}} \\ \mathbf{y} \end{bmatrix} := \begin{bmatrix} \mathbf{A}_S & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{A}_S^\top \end{bmatrix} \begin{bmatrix} \bar{\mathbf{x}}_S \\ \bar{\mathbf{x}}_{S^c} \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} \mathbf{b} \\ \mathbf{0} \\ \hat{\mathbf{w}}_S \end{bmatrix}, \tag{20}$$

where $\mathbf{y}$ is the dual variable and $\hat{\mathbf{w}}_S = \mathbf{w}_S \odot \text{sign}(\bar{\mathbf{x}}_S)$. If $\ker(\mathbf{M}) \neq \{\mathbf{0}\}$, then we choose a nonzero vector $\triangle\mathbf{x}$ from the kernel space of $\mathbf{M}$ (note $(\triangle\mathbf{x})_{S^c} = \mathbf{0}$). Thus, from the optimality condition (20), we can find $t_1 < 0$ and $t_2 > 0$ such that $\bar{\mathbf{x}} + t \triangle \mathbf{x}$ is also an optimal solution for any $t \in [t_1, t_2]$. Both t_1 and t_2 exist; otherwise the optimal objective value will be $+\infty$. Furthermore, we can choose t_1 (and t_2) such that at least one component of $\hat{\mathbf{x}}_S := \bar{\mathbf{x}}_S + t_1(\triangle\mathbf{x})_S$ (and $\check{\mathbf{x}}_S := \bar{\mathbf{x}}_S + t_2(\triangle\mathbf{x})_S$) is zero. It implies that $\bar{\mathbf{x}}$ is a weighted average of $\check{\mathbf{x}}$ and $\hat{\mathbf{x}}$, both of which have fewer nonzero components. Then we can apply the same technique on $\hat{\mathbf{x}}$ and $\check{\mathbf{x}}$ until we end up with solutions such that $\ker(\mathbf{M}) = \{\mathbf{0}\}$. Therefore, we have $\bar{\mathbf{x}}$ is within the convex hull constructed by some solutions of $\mathbf{A}_S\mathbf{x}_S = \mathbf{b}$ and $\mathbf{x}_{S^c} = \mathbf{0}$ with full column rank sub-matrix $\mathbf{A}_S$. Since the number of such submatrices is finite, the whole sequence is bounded.

Now that $\{\mathbf{x}^k\}_{k=1}^{\infty}$ is bounded, then the Bolzano–Weierstrass Theorem guarantees the existence of a convergent subsequence, denoted by $\{\mathbf{x}^{n_k}\}_{k=1}^{\infty}$. Assume that it converges to $\mathbf{x}^*$. Since $\mathbf{x}^{n_k} \to \mathbf{x}^*$ and $\mathbf{A}\mathbf{x}^{n_k} = \mathbf{b}$, we have $\mathbf{A}\mathbf{x}^* = \mathbf{b}$.

Due to the $\mathbf{x}$-subproblem given in (18), it is straightforward that

$$\sum_{j=1}^{n} w_j^k|x_j^{k+1}| \le \sum_{j=1}^{n} w_j^k|x_j^k|.$$

As a result, we have the following inequality

$$\begin{aligned} J_\sigma(\mathbf{x}^k) - J_\sigma(\mathbf{x}^{k+1}) &\ge \sum_{j=1}^{n}\left[\Phi_\sigma(|x_j^k|) - \Phi_\sigma(|x_j^{k+1}|)\right] - \sum_{j=1}^{n} w_j^k\left(|x_j^k| - |x_j^{k+1}|\right) \\ &= \sum_{j=1}^{n}\left[\Phi_\sigma(|x_j^k|) - \Phi_\sigma(|x_j^{k+1}|) - w_j^k\left(|x_j^k| - |x_j^{k+1}|\right)\right] \ge 0, \end{aligned}$$

as $\Phi_\sigma(\cdot)$ is a concave function and w_j^k is the derivative of Φ_σ evaluated at $|x_i^k|$. Therefore, we can obtain that

$$\min_{k=1,\cdots,K} \sum_{j=1}^{n}\left[\Phi_\sigma(|x_j^k|) - \Phi_\sigma(|x_j^{k+1}|) - w_j^k\left(|x_j^k| - |x_j^{k+1}|\right)\right]$$

$$\leq \frac{1}{K}\sum_{k=1}^{K}\sum_{j=1}^{n}\Big[\Phi_\sigma(|x_j^k|)-\Phi_\sigma(|x_j^{k+1}|)-w_j^k\big(|x_j^k|-|x_j^{k+1}|\big)\Big] \leq \frac{J_\sigma(\mathbf{x}^1)}{K}.$$

In addition, we have $|\mathbf{x}^k|-|\mathbf{x}^{k+1}|\to\mathbf{0}$. When k is large enough, we have $\mathbf{x}^k-\mathbf{x}^{k+1}\to\mathbf{0}$; otherwise $J_\sigma((\mathbf{x}^k+\mathbf{x}^{k+1})/2) < J_\sigma(\mathbf{x}^{k+1})$. Therefore, $\mathbf{x}^{n_k+1}$ converges to $\mathbf{x}^*$. According to the optimality condition of (18), we have $p_j^{n_k+1}\in\partial|x_j^{n_k+1}|$ and $\mathbf{w}^{n_k+1}\odot\mathbf{p}^{n_k+1}$ is in the range of $\mathbf{A}^\top$. Since the sequence $\{p_j^{n_k+1}\}$ is bounded by ±1, it has a convergent subsequence. Without loss of generality, we assume that the sequence $\{p_j^{n_k+1}\}$ itself converges. Therefore, we have

$$\exp\left(-\left(\frac{x_j^*}{\sigma}\right)^2\right)p_j^* = \lim_{k\to\infty} w_j^{n_k}p_j^{n_k+1},$$

which is in the range of $\mathbf{A}^\top$ and $p_j^*\in\partial|x_j^*|$. These relationships show that $\mathbf{x}^*$ is a stationary point of (17). □

4.2 Unconstrained Formulation

In the noisy case, we consider an unconstrained problem of the form

$$\min_{\mathbf{x}\in\mathbb{R}^n}\lambda J_\sigma(\mathbf{x})+\frac{1}{2}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2, \tag{21}$$

with a positive parameter λ. The reweighted L_1 algorithm requires to solve the following subproblem:

$$\mathbf{x}^{k+1} = \arg\min_{\mathbf{x}\in\mathbb{R}^n}\lambda\sum_{j=1}^{n}w_j^k|x_j|+\frac{1}{2}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2, \tag{22}$$

where the weight vector $\mathbf{w}^k$ is defined in the same way as in (18). Exactly solving the subproblem (22) is impractical due to the lack of closed-form solutions. In practice, we run several iterations of ADMM to estimate the solution to this subproblem. We introduce an auxiliary variable $\mathbf{y}$ and split the objective function as

$$\min_{\mathbf{x},\mathbf{y}\in\mathbb{R}^n}\lambda\sum_{j=1}^{n}w_j|x_j|+\frac{1}{2}\|\mathbf{A}\mathbf{y}-\mathbf{b}\|_2^2 \quad \text{s.t.} \quad \mathbf{x}=\mathbf{y}. \tag{23}$$

We omit the (outer) iteration index k when the context is clear. The corresponding augmented Lagrangian can be expressed by

$$\mathcal{L}(\mathbf{x},\mathbf{y};\mathbf{u}) := \lambda\sum_{j=1}^{n}w_j|x_j|+\frac{1}{2}\|\mathbf{A}\mathbf{y}-\mathbf{b}\|_2^2+\frac{\delta}{2}\|\mathbf{x}-\mathbf{y}+\mathbf{u}\|_2^2, \tag{24}$$

where $\mathbf{u}$ is the dual variable and δ is a positive parameter. The ADMM algorithm involves the following steps:

$$\begin{cases}\mathbf{x}_{l+1} = \arg\min_{\mathbf{x}}\mathcal{L}(\mathbf{x},\mathbf{y}_l;\mathbf{u}_l),\\ \mathbf{y}_{l+1} = \arg\ \min_{\mathbf{y}}\mathcal{L}(\mathbf{x}_{l+1},\mathbf{y};\mathbf{u}_l),\\ \mathbf{u}_{l+1} = \mathbf{u}_l+\mathbf{x}_{l+1}-\mathbf{y}_{l+1},\end{cases} \tag{25}$$

where the inner iteration is indexed by l. There are closed-form solutions for both subproblems of $\mathbf{x}$ and $\mathbf{y}$ given by

$$\mathbf{x}_{l+1} = \text{shrink}(\mathbf{y}_l - \mathbf{u}_l, \frac{\lambda}{\delta}\mathbf{w}) := \text{sign}(\mathbf{y}_l - \mathbf{u}_l) \odot \max\left(|\mathbf{y}_l - \mathbf{u}_l| - \frac{\lambda}{\delta}\mathbf{w}, 0\right),$$
$$\mathbf{y}_{l+1} = (\mathbf{A}^\top\mathbf{A} + \delta\mathbf{I}_d)^{-1}(\mathbf{A}^\top\mathbf{b} + \delta\mathbf{x}_{l+1} + \delta\mathbf{u}_l),$$

where $\mathbf{I}_d$ denotes the identity matrix. The overall algorithm for solving the unconstrained ERF-regularized model is summarized in Algorithm 1.

Algorithm 1 The iterative reweighted L_1 algorithm for solving the unconstrained ERF-regularized model (21).

Input: $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\mathbf{b} \in \mathbb{R}^m$, $\sigma, \lambda, \delta > 0$, and MaxOuter/MaxInner.
Initialization: $k = 1$ and solve for the L_1 minimization to get $\mathbf{x}^1$.
while $k <$ MaxOuter or other stopping criteria **do**
  $\mathbf{w}^k = \exp\{-(\frac{\mathbf{x}^k}{\sigma})^2\}$
  $l = 1, \mathbf{y}_l = \mathbf{x}^k, \mathbf{u}_l = \mathbf{0}$.
  while $l <$ MaxInner or other stopping criteria **do**
    $\mathbf{x}_{l+1} = \text{shrink}(\mathbf{y}_l - \mathbf{u}_l, \frac{\lambda}{\delta}\mathbf{w}^k)$.
    $\mathbf{y}_{l+1} = (\mathbf{A}^\top\mathbf{A} + \delta I_d)^{-1}(\mathbf{A}^\top\mathbf{b} + \delta\mathbf{x} + \delta\mathbf{u})$.
    $\mathbf{u}_{l+1} = \mathbf{u}_l + \mathbf{x}_{l+1} - \mathbf{y}_{l+1}$.
    $l \leftarrow l + 1$.
  end while
  $\mathbf{x}^{k+1} = \mathbf{x}_l$, $k \leftarrow k + 1$.
end while
return $\mathbf{x}^k$

For the convergence, we show in Lemma 1 that $\{\mathbf{x}^{k+1}\}$ is uniformly bounded when (22) is solved exactly. When it is solved inexactly, we shall assume that $\{\mathbf{x}^{k+1}\}$ is also uniformly bounded and satisfies

$$\langle \lambda\mathbf{w}^k \odot \mathbf{p}^{k+1} + \mathbf{A}^\top(\mathbf{A}\mathbf{x}^{k+1} - \mathbf{b}), \mathbf{x}^k - \mathbf{x}^{k+1}\rangle \geq -\|\mathbf{A}(\mathbf{x}^k - \mathbf{x}^{k+1})\|_2^2/4, \tag{26}$$

with $\mathbf{p}^{k+1} \in \partial\|\mathbf{x}^{k+1}\|_1$. This is a reasonable assumption because we have $0 > -\|\mathbf{A}(\mathbf{x}^k - \mathbf{x}^{k+1})\|_2^2/2$ with the exact solution of (22).

Lemma 1 *For any $\{w_j^k\}_{j=1}^n$, solutions of (22) are uniformly bounded.*

Proof Let $\bar{\mathbf{x}}$ be a solution for the $\mathbf{x}$-subproblem in (22) and S the set of corresponding indices of nonzero components in $\bar{\mathbf{x}}$. We consider the following problem restricted to the index set S,

$$\min_{\mathbf{x}} \lambda \sum_{j\in S} w_j|x_j| + \frac{1}{2}\|\mathbf{A}_S\mathbf{x}_S - \mathbf{b}\|_2^2, \quad \text{s.t.} \quad \mathbf{x}_{S^c} = \mathbf{0}, \tag{27}$$

of which $\bar{\mathbf{x}}$ is the optimal solution. The optimality condition is

$$\mathbf{M}\mathbf{x} := \begin{bmatrix} \mathbf{A}_S^\top\mathbf{A}_S & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}\begin{bmatrix} \bar{\mathbf{x}}_S \\ \bar{\mathbf{x}}_{S^c} \end{bmatrix} = \begin{bmatrix} \mathbf{A}_S^\top\mathbf{b} - \lambda\hat{\mathbf{w}}_S \\ \mathbf{0} \end{bmatrix}, \tag{28}$$

where $\hat{\mathbf{w}}_S = (\mathbf{w} \odot \text{sign}(\bar{\mathbf{x}}))_S$. Then using the same technique in the proof of Theorem 4, we can show that $\{\mathbf{x}^k\}_{k=1}^\infty$ is within a convex hull constructed by $\{\mathbf{x} : \exists\, S \text{ and } \hat{\mathbf{w}} \in (0, 1]^n \text{ such that } \mathbf{x}_S = (\mathbf{A}_S^\top\mathbf{A}_S)^{-1}(\mathbf{A}_S^\top\mathbf{b} - \lambda\hat{\mathbf{w}}_S)\}$. This set is bounded, so the optimal solution $\bar{\mathbf{x}}$ is also bounded. □

Theorem 5 *If the subproblem* (22) *is solved in a way that* $\{\boldsymbol{x}^k\}_{k=1}^{\infty}$ *is uniformly bounded and satisfies* (26) *for all* k. *Then* $\{\boldsymbol{x}^k\}_{k=1}^{\infty}$ *generated by* Algorithm 1 *has a convergent subsequence, and any accumulation point of* $\{\boldsymbol{x}^k\}_{k=1}^{\infty}$ *is a stationary point of* (21). *In addition, we have*

$$\begin{aligned}&\min_{k=1,\cdots,K}\sum_{j=1}^{n}\Big[\Phi_\sigma(|x_j^k|)-\Phi_\sigma(|x_j^{k+1}|)-w_j^k\big(|x_j^k|-|x_j^{k+1}|\big)\Big]\\&\le\frac{J_\sigma(\boldsymbol{x}^1)+\frac{1}{2\lambda}\|\boldsymbol{A}\boldsymbol{x}^1-\boldsymbol{b}\|_2^2}{K}.\end{aligned}\tag{29}$$

Proof Because $\{\mathbf{x}^k\}_{k=1}^{\infty}$ is bounded, there exists a subsequence $\{\mathbf{x}^{n_k}\}_{k=1}^{\infty}$ which converges to $\mathbf{x}^*$. It follows from (26) that

$$\begin{aligned}&\frac{1}{2}\|\mathbf{A}\mathbf{x}^k-\mathbf{b}\|_2^2-\frac{1}{2}\|\mathbf{A}\mathbf{x}^{k+1}-\mathbf{b}\|_2^2\\&=\frac{1}{2}\|\mathbf{A}\mathbf{x}^k-\mathbf{A}\mathbf{x}^{k+1}\|_2^2+\langle\mathbf{x}^k-\mathbf{x}^{k+1},\mathbf{A}^\top(\mathbf{A}\mathbf{x}^{k+1}-\mathbf{b})\rangle\\&\ge-\lambda\sum_{j=1}^{n}\langle x_j^k-x_j^{k+1},w_j^kp_j^{k+1}\rangle+\frac{1}{4}\|\mathbf{A}\mathbf{x}^k-\mathbf{A}\mathbf{x}^{k+1}\|_2^2\\&\ge-\lambda\sum_{j=1}^{n}w_j^k(|x_j^k|-|x_j^{k+1}|)+\frac{1}{4}\|\mathbf{A}\mathbf{x}^k-\mathbf{A}\mathbf{x}^{k+1}\|_2^2.\end{aligned}$$

We further obtain that

$$\begin{aligned}&\left(\lambda J_\sigma(\mathbf{x}^k)+\frac{1}{2}\|\mathbf{A}\mathbf{x}^k-\mathbf{b}\|_2^2\right)-\left(\lambda J_\sigma(\mathbf{x}^{k+1})+\frac{1}{2}\|\mathbf{A}\mathbf{x}^{k+1}-\mathbf{b}\|_2^2\right)\\&=\lambda\sum_{j=1}^{n}\Big[\Phi_\sigma(|x_j^k|)-\Phi_\sigma(|x_j^{k+1}|)-w_j^k(|x_j^k|-|x_j^{k+1}|)\Big]+\frac{1}{4}\|\mathbf{A}\mathbf{x}^k-\mathbf{A}\mathbf{x}^{k+1}\|_2^2\ge0,\end{aligned}$$

which yields the convergence rate (29). In addition, we have $|\mathbf{x}^k|-|\mathbf{x}^{k+1}|\to\mathbf{0}$ and $\mathbf{A}\mathbf{x}^k-\mathbf{A}\mathbf{x}^{k+1}\to\mathbf{0}$ as $k\to\infty$. Then we get $\mathbf{x}^k-\mathbf{x}^{k+1}\to\mathbf{0}$; otherwise $(\mathbf{x}^k+\mathbf{x}^{k+1})/2$ has a smaller function value than $\mathbf{x}^{k+1}$ when k is larger. Therefore, we have $\mathbf{x}^{n_k+1}\to\mathbf{x}^*$ as $k\to\infty$. Since the sequence $\{p_j^{n_k+1}\}$ is bounded by ±1, it has a convergent subsequence. Without loss of generality, we assume that $\{p_j^{n_k+1}\}$ itself converges. Thus

$$\begin{aligned}0&=\lim_{k\to\infty}\lambda w_j^{n_k}p_j^{n_k+1}+\mathbf{A}_j^\top(\mathbf{A}\mathbf{x}^{n_k+1}-\mathbf{b})\\&=\lambda e^{-\frac{(x_j^*)^2}{\sigma^2}}p_j^*+\mathbf{A}_j^\top(\mathbf{A}\mathbf{x}^*-\mathbf{b}),\end{aligned}$$

where $p_j^*\in\partial|x_j^*|$. That is, $\mathbf{x}^*$ is a stationary point of (21). □

5 Experiments

In this section, we demonstrate the performance of the proposed algorithms in comparison to the state-of-the-art methods in sparse signal recovery. All the numerical experiments are conducted on a Windows desktop with CPU (Intel i7-6700, 3.19GHz) and MATLAB (R2019a). The codes, including test data for the experiments, will be available when the paper is published.

5.1 Noise-Free Case

We focus on one type of sparse signal recovery problems that involves highly coherent matrices. The coherence of a matrix $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_n]$ is defined as

$$\mu(\mathbf{A}) := \max_{i \neq j} \frac{|\mathbf{a}_i^\top \mathbf{a}_j|}{\|\mathbf{a}_i\|_2 \|\mathbf{a}_j\|_2}. \tag{30}$$

By the Cauchy-Schwarz inequality, it is straightforward that $0 \leq \mu(\mathbf{A}) \leq 1$. If the coherence of a matrix is close to one, then the matrix is said to be highly coherent, where the standard L_1 model does not work well.

Following the works of [14,20,40], we consider an over-sampled discrete cosine transform (DCT), defined as $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_n] \in \mathbb{R}^{m \times n}$ with

$$\mathbf{a}_j := \frac{1}{\sqrt{m}} \cos\left(\frac{2\pi j \mathbf{w}}{F}\right), \quad j = 1, \cdots, n, \tag{31}$$

where $\mathbf{w}$ is a random vector uniformly distributed in $[0, 1]^m$ and $F \in \mathbb{R}$ is a positive parameter to control the coherence so that a larger value of F yields a more coherent matrix. Throughout the experiments, we consider over-sampled DCT matrices of size 64×1024. The ground truth $\mathbf{x} \in \mathbb{R}^n$ is simulated as an s-sparse signal, where s is the number of nonzero entries. As suggested in [14], we require a minimum separation of $2F$ in the support of $\mathbf{x}$. The values of nonzero elements follow the Gaussian distribution, i.e., $x_{i_s} \sim \mathcal{N}(0, 1)$, $i = 1, 2, \ldots, s$.

We evaluate sparse signal recovery performance in terms of *success rate*, defined as the number of successful trials over the total number of trials, which is 100 for all the experiments. A success is declared if the relative error of the reconstructed solution $\mathbf{x}^*$ to the ground truth $\mathbf{x}$ is less than 10^{-3}, i.e., $\|\mathbf{x}^* - \mathbf{x}\|_2 / \|\mathbf{x}\|_2 \leq 10^{-3}$.

Figure 3 shows the performance of the ERF regularization with respect to different choices of σ, which numerically demonstrates that the proposed regularization approaches to the L_1 norm for a large value of σ. We observe that the smaller the coherence is, the smaller the value of σ is, which serves as the practical guide for the parameter selection.

Following from Fig. 3, we choose $\sigma = 0.1, 0.5, 0.5, 1$ for $F = 1, 5, 10, 20$, respectively, and compare the sparse recovery performance among the state-of-the-art methods in Fig. 4. The competing methods are labeled as L_0 (IRL1 [7]), L_p, TL1, and L_1-L_2 [20,40]. We choose $L_{1/2}$ as a representative regularization in the L_p family [36,37], and $a = 1$ as recommended in [43]. We observe that the proposed approach is always the best or at least the second best under all coherence and sparsity levels.

5.2 Super-Resolution

We also examine the case of super-resolution, in which a coherent sensing matrix is involved. A mathematical model for super-resolution can be expressed as

$$b_k = \frac{1}{\sqrt{N}} \sum_{t=0}^{N-1} x_t e^{-i2\pi kt/N}, \qquad |k| \leq f_c, \tag{32}$$

where i is the imaginary unit, $\mathbf{x} \in \mathbb{R}^N$ is a vector to be recovered and $\mathbf{b} \in \mathbb{C}^n$ consists of the given low frequency measurements with $n = 2f_c + 1 < N$. This is related to super-resolution in the sense that the underlying signal $\mathbf{x}$ is defined on a fine grid with spacing $1/N$, while the frequency data of length n implies that one can only expect to recover the signal on a coarser

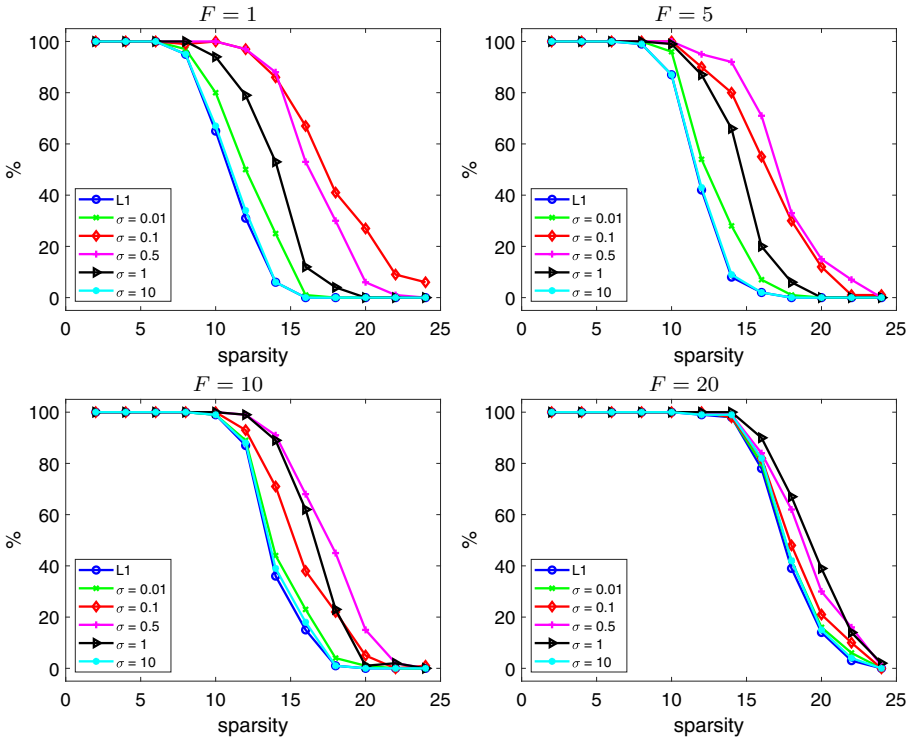

Fig. 3 The performance of the ERF regularization (in terms of success rate) with respect to the choice of σ

grid with spacing $1/n$. For simplicity, we use matrix notation to rewrite (32) as $\mathbf{b} = S_n \mathcal{F}\mathbf{x}$, where S_n is a sampling matrix that indicates what frequency is collected, $\mathcal{F}$ is the Fourier transform matrix, and we denote $\mathcal{F}_n = S_n \mathcal{F}$. The frequency cutoff induces a resolution limit inversely proportional to f_c; below, we set $\lambda_c = 1/f_c$, which is referred to as the Rayleigh length (a classical resolution limit of hardware [18]).

We are interested in reconstructing point sources, i.e., $\mathbf{x} = \sum_{t_j \in T} c_j \delta_{t_j}$, where δ_τ is a Dirac measure at τ, spikes of $\mathbf{x}$ are located at t_j's belonging to a set T, and c_j's are coefficients. Following the work of [5], the sparse spikes are required to be sufficiently separated; please refer to Definition 2 and Theorem 6.

Definition 2 (Minimum Separation [5]) Let $\mathbb{T}$ be the circle obtained by identifying the endpoints on $[0, 1]$ and $\mathbb{T}^d$ the d-dimensional torus. For a family of points $T \subset \mathbb{T}^d$, the minimum separation is defined as the closest wrap-around distance between any two elements from T,

$$\mathrm{MS} := \triangle(T) := \inf_{(t,t') \in T : t \neq t'} |t - t'|, \tag{33}$$

where $|t - t'|$ is the L_∞ distance (maximum deviation along any coordinate axis).

Theorem 6 *[5, Corollary 1.4] Let $T = \{t_j\}$ be the support of $\mathbf{x}$. If the minimum distance obeys*

$$\triangle(T) \geq 2\lambda_c N, \tag{34}$$

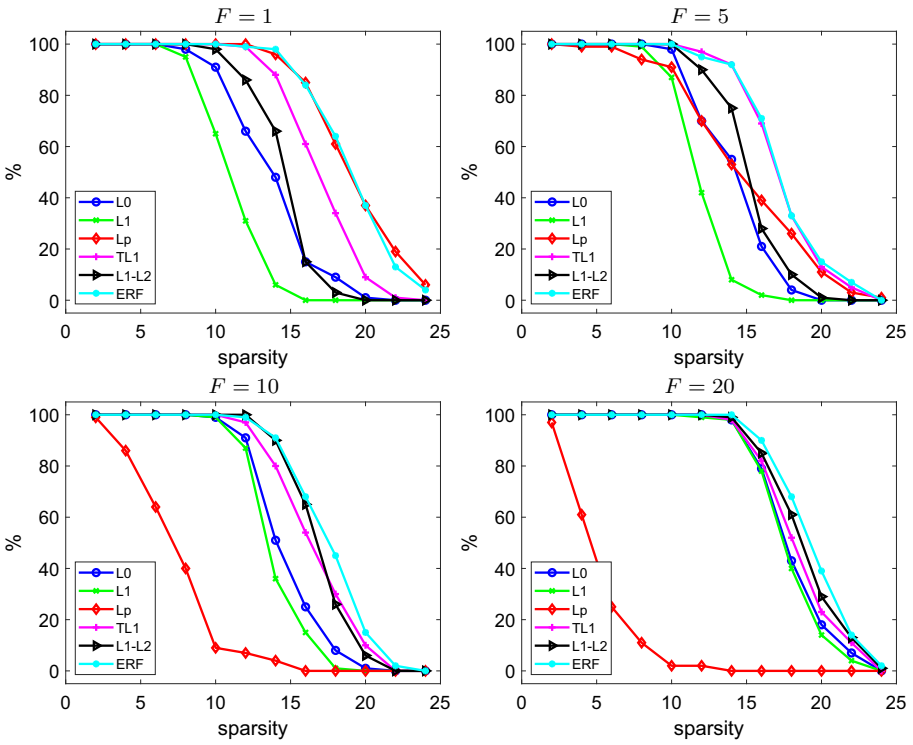

Fig. 4 The comparison with the state-of-the art methods for sparse signal recovery (in terms of success rate)

then $\boldsymbol{x}$ *is the unique solution to the* L_1 *minimization problem:*

$$\min \|\boldsymbol{x}\|_1 \quad s.t. \quad \mathcal{F}_n \boldsymbol{x} = \boldsymbol{b}. \tag{35}$$

If x *is real-valued, then the minimum gap can be lowered to* $1.87\lambda_c N$.

We are interested in the constant in front of $\lambda_c N$ in (34), referred to as minimum separation factor (MSF). Theorem 6 indicates that MSF$\geq$ 2 guarantees exact recovery of L_1 minimization. We want to analyze how different sparse recovery algorithms behave with respect to MSF. For this purpose, we consider a sparse signal (ground truth) $\mathbf{x}_g$ of dimension 1000 with MS = 20. We vary f_c from 31 to 60, thus MSF:= $\triangle(T) \cdot f_c/N$:=MS$\cdot f_c/N$=0.62 : 0.02 : 1.2. Denote $\mathbf{x}^*$ as the reconstructed signal using any of the methods including L_1 via SDP [5], constrained L_1-L_2 minimization via DCA [21], and the proposed ERF model via IRL1. We consider 100 random realizations of the same setting to compute the success rate. Figure 5 shows great advantages of the nonconvex approaches L_1-L_2 and ERF over the convex L_1 approach, while the proposed ERF model is slightly better than L_1-L_2.

5.3 Noisy Case

We provide a series of simulations to demonstrate sparse recovery with noise, following an experimental setup in [36]. We consider a signal $\mathbf{x}$ of length $n = 512$ with $s = 130$ nonzero elements. We try to recover it from m measurements (denoted by $\mathbf{b}$) determined by a

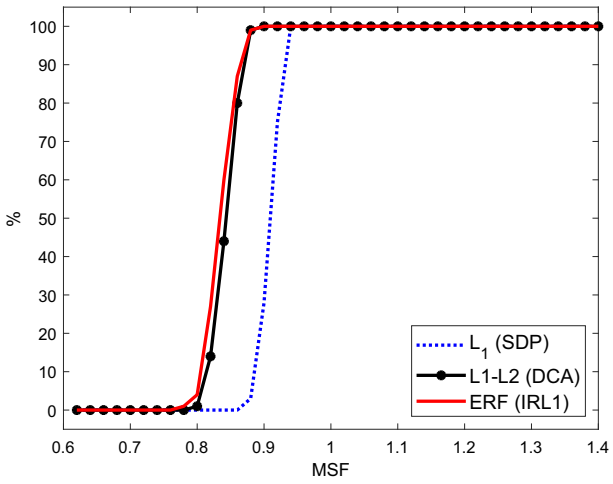

Fig. 5 Success rates (%) of fixed MS= 20 with respect to MSF for the ambient dimension $N = 1000$

Gaussian random matrix $\mathbf{A}$, i.e., a matrix whose columns are normalized with zero-mean and unit Euclidean norm, and Gaussian noise with zero mean and standard deviation $\eta = 0.1$. Taking noise into consideration, we use the mean-square-error (MSE) to quantify the recovery performance. If the support of the ground-truth solution $\mathbf{x}$ is known, denoted as $S = \text{supp}(\mathbf{x})$, we can compute the MSE of an oracle solution, given by the formula $\eta^2 \text{tr}(\mathbf{A}_S^\top \mathbf{A}_S)^{-1}$, as benchmark.

We compare the proposed ERF model with $L_{1/2}$ via the half-thresholding method [36], L_1, and L_1-L_2 (both are solved via ADMM). We use cross validation (CV) [1,2] to determine the value of the parameter σ and λ for ERF. In particular, we use a matrix A of size 300×512 and adopt 10-fold CV by splitting the 300 samples into 10 equal-sized subsets. For $k = 1, \ldots, 10$, we decompose the row vectors of A into the disjoint training and testing sets, where the testing set contains the kth subset and the remaining data constitutes the training set. We consider two candidate parameter sets: $\sigma \in \{0.01, 0.1, 0.5, 1, 10\}$ and $\lambda \in \{0.001, 0.01, 0.1, 1\}$. For each combination of σ and λ, we estimate the sparse coefficient using the training set and calculate the prediction error of the fitted model on the testing set. The optimal values are chosen as $\sigma = 0.5$, $\lambda = 0.1$ which yield the smallest predication error after averaging over 10 folds. The same strategy is used to find the optimal λ value among a candidate set of $\{0.001, 0.01, 0.1, 1\}$ for the other regularized models. We use these optimal parameters selected using CV for all of the following numerical comparison.

In Table 1, we present the mean and the standard deviation of MSE and computation time at the four particular m values: 240, 270, 300, and 340. Each reported value (MSE and time) is the average of 100 random realizations. The proposed method achieves the best results. We present more m values in Fig. 6, which is based on an average of 100 random realizations under the same setup. L_p and ERF are asymptotically approaching the oracle solutions for larger m values. Note that we use 300 samples in CV to find the optimal parameters, and fix them when varying the value of m. The results in Table 1 and Fig. 6 demonstrate that ERF is less sensitive to parameter selection compared to other competing methods.

Table 1 Recovery results of noisy signals (mean and standard deviation over 100 realizations)

Methods	m	MSE	Time (s)	m	MSE	Time (s)
Oracle		*4.55* (0.08)			*3.56* (0.04)	
$L_{1/2}$		5.89 (0.80)	4.03 (2.92)		4.56 (0.79)	7.42 (2.95)
L_1	240	5.77 (0.68)	0.09 (0.11)	270	4.76 (0.63)	0.08 (0.02)
L_1-L_2		5.61 (0.72)	0.25 (0.28)		4.55 (0.63)	0.19 (0.04)
ERF		**5.59** (1.22)	0.57 (0.20)		**3.91** (0.62)	0.48 (0.12)
Oracle		*2.77* (0.02)			*2.37* (0.02)	
$L_{1/2}$		3.22 (0.35)	8.42 (3.27)		2.63 (0.26)	8.94 (3.50)
L_1	310	3.68 (0.36)	0.06 (0.01)	340	3.08 (0.29)	0.06 (0.01)
L_1-L_2		3.49 (0.34)	0.15 (0.03)		2.94 (0.26)	0.13 (0.03)
ERF		**2.91** (0.27)	0.31 (0.05)		**2.49** (0.18)	0.27 (0.05)

The best results are highlighted in boldface and oracle results are in italics

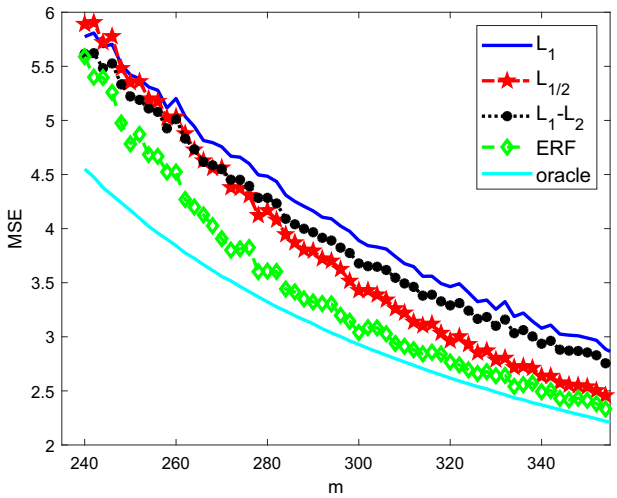

Fig. 6 MSE of sparse recovery under the presence of additive Gaussian white noise. The sensing matrix is of size $m \times n$, where m ranges from 240 to 350 and $n = 512$. The ground-truth sparse vector contains 130 nonzero elements. Each MSE value is the average of 100 random realizations

6 Conclusions and Future Works

We propose a novel regularization based on the error function for sparse signal recovery. The asymptotic behavior of the error function indicates that the proposed regularization can approximate the standard L_0, L_1 norms as the parameter approaches 0 and ∞, respectively. We apply Newton's method to find a solution for the proximal operator corresponding to the proposed regularization. Plots of asymptotic behavior and proximal solutions demonstrate that the proposed regularizer is smoother and less biased than the L_1 counterpart. We also develop the iterative reweighted algorithms for constrained and unconstrained formulations, both with guaranteed convergence. Experiments demonstrate that the proposed model outperforms the state-of-the-art approaches in sparse recovery in various settings.

Our future work will involve theoretical comparisons between gNSP for the proposed regularizer and NSP for L_1. We will also develop alternative numerical schemes to minimize the proposed model, e.g., by using the proximal operator.

Acknowledgements The authors would like to acknowledge Dr. Chao Wang for providing sparse recovery codes and the anonymous reviewers for their comments and suggestions. This research was initialized at the American Institute of Mathematics Structured Quartet Research Ensembles (SQuaREs), July 22–26, 2019.

Appendix: Proofs of (P3)–(P4) in Section 3.1

Proof To show the concavity of $J_\sigma = \sum_{j=1}^n \Phi_\sigma(|x_j|)$ on $\mathbb{R}^n_+$, we have for $x > 0$ that

$$\frac{d^2}{dx^2}\Phi_\sigma(x) = -\frac{2x}{\sigma^2}\exp\Big(-\frac{x^2}{\sigma^2}\Big) < 0,$$

which implies that $\Phi_\sigma(x)$ is concave on $[0, \infty)$. Consequently, for any $t \in [0, 1]$, $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n_+$, we have

$$\Phi_\sigma\big(tx_j + (1-t)y_j\big) \geq t\Phi_\sigma(x_j) + (1-t)\Phi_\sigma(y_j), \qquad \text{for } j = 1, \ldots, n.$$

Therefore, (P3) holds, i.e.,

$$\begin{aligned} J_\sigma\big(t\mathbf{x} + (1-t)\mathbf{y}\big) &= \sum_{j=1}^n \Phi_\sigma\big(tx_j + (1-t)y_j\big) \\ &\geq t\sum_{j=1}^n \Phi_\sigma(x_j) + (1-t)\sum_{j=1}^n \Phi_\sigma(y_j) \\ &= tJ_\sigma(\mathbf{x}) + (1-t)J_\sigma(\mathbf{y}). \end{aligned}$$

As for (P4), we recall that

$$J_\sigma(\mathbf{x}) = \sum_{j=1}^n \int_0^{|x_j|} e^{-\pi^2/\sigma^2} d\tau.$$

Then for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, we have

$$\begin{aligned} J_\sigma(\mathbf{x}+\mathbf{y}) &= \sum_{j=1}^n \int_0^{|x_j+y_j|} e^{-\tau^2/\sigma^2} d\tau \\ &= \sum_{j=1}^n \int_0^{|x_j|} e^{-\tau^2/\sigma^2} d\tau + \sum_{j=1}^n \int_{|x_j|}^{|x_j+y_j|} e^{-\tau^2/\sigma^2} d\tau \\ &\leq \sum_{j=1}^n \int_0^{|x_j|} e^{-\tau^2/\sigma^2} d\tau + \sum_{j=1}^n \int_{|x_j|}^{|x_j|+|y_j|} e^{-\tau^2/\sigma^2} d\tau \\ &\leq \sum_{j=1}^n \int_0^{|x_j|} e^{-\tau^2/\sigma^2} d\tau + \sum_{j=1}^n \int_0^{|y_j|} e^{-\tau^2/\sigma^2} d\tau. \end{aligned}$$

The last inequality is guaranteed by the fact that $e^{-\tau/\sigma^2}$ is nonnegative and decreasing for $\tau \in [0, \infty)$. Therefore, we have $J_\sigma(\mathbf{x}+\mathbf{y}) \leq J_\sigma(\mathbf{x}) + J_\sigma(\mathbf{y})$. □

References

1. Adcock, B., Bao, A., Brugiapaglia, S.: Correcting for unknown errors in sparse high-dimensional function approximation. Numer. Math. **142**(3), 667–711 (2019)
2. Arlot, S., Celisse, A., et al.: A survey of cross-validation procedures for model selection. Stat. Surv. **4**, 40–79 (2010)
3. Bai, Y., Cheung, G., Liu, X., Gao, W.: Graph-based blind image deblurring from a single photograph. IEEE Trans. Image Process. **28**(3), 1404–1418 (2018)
4. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., et al.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends Mach. Learn. **3**(1), 1–122 (2011)
5. Candès, E.J., Fernandez-Granda, C.: Towards a mathematical theory of super-resolution. Commun. Pure Appl. Math. **67**(6), 906–956 (2014)
6. Candès, E.J., Romberg, J.K., Tao, T.: Stable signal recovery from incomplete and inaccurate measurements. Commun. Pure Appl. Math. **59**(8), 1207–1223 (2006)
7. Candès, E.J., Wakin, M.B., Boyd, S.P.: Enhancing sparsity by reweighted l1 minimization. J. Fourier Anal. Appl. **14**(5–6), 877–905 (2008)
8. Chambolle, A., Pock, T.: A first-order primal-dual algorithm for convex problems with applications to imaging. J. Math. Imaging Vis. **40**(1), 120–145 (2011)
9. Chartrand, R.: Exact reconstruction of sparse signals via nonconvex minimization. IEEE Signal Process Lett. **14**(10), 707–710 (2007)
10. Chu, J.T.: On bounds for the normal integral. Biometrika **42**(1/2), 263–265 (1955)
11. Donoho, D.L.: Compressed sensing. IEEE Trans. Inf. Theory **52**(4), 1289–1306 (2006)
12. Donoho, D.L., Huo, X.: Uncertainty principles and ideal atomic decomposition. IEEE Trans. Inf. Theory **47**(7), 2845–2862 (2001)
13. Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Stat. Assoc. **96**(456), 1348–1360 (2001)
14. Fannjiang, A., Liao, W.: Coherence pattern-guided compressive sensing with unresolved grids. SIAM J. Imag. Sci. **5**(1), 179–202 (2012)
15. Foucart, S., Rauhut, H.: A Mathematical Introduction to Compressive Sensing. Birkhäuser, New York (2013)
16. Gabay, D., Mercier, B.: A dual algorithm for the solution of nonlinear variational problems via finite element approximation. Comput. Math. Appl. **2**(1), 17–40 (1976)
17. Glowinski, R., Marroco, A.: Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de dirichlet non linéaires. ESAIM Math. Model. Numer. Anal.-Modélisation Mathématique et Analyse Numérique **9**(R2), 41–76 (1975)
18. Goodman, J.W.: Introduction to Fourier Optics. Roberts and Company Publishers, Greenwood Village (2005)
19. Lange, K., Hunter, D., Yang, I.: Optimization transfer using surrogate objective functions. J. Comput. Graph. Stat. **9**(1), 1–20 (2000)
20. Lou, Y., Yin, P., He, Q., Xin, J.: Computing sparse representation in a highly coherent dictionary based on difference of L_1 and L_2. J. Sci. Comput. **64**(1), 178–196 (2015)
21. Lou, Y., Yin, P., Xin, J.: Point source super-resolution via non-convex l1 based methods. J. Sci. Comput. **68**, 1082–1100 (2016)
22. Lv, J., Fan, Y., et al.: A unified approach to model selection and sparse recovery using regularized least squares. Ann. Stat. **37**(6A), 3498–3528 (2009)
23. Mammone, R.J.: Spectral extrapolation of constrained signals. J. Opt. Soc. Am. **73**(11), 1476–1480 (1983)
24. Natarajan, B.K.: Sparse approximate solutions to linear systems. SIAM J. Comput. **24**(2), 227–234 (1995)
25. Ochs, P., Dosovitskiy, A., Brox, T., Pock, T.: On iteratively reweighted algorithms for nonsmooth nonconvex optimization in computer vision. SIAM J. Imaging Sci. **8**(1), 331–372 (2015)
26. Papoulis, A., Chamzas, C.: Improvement of range resolution by spectral extrapolation. Ultra. Imag. **1**(2), 121–135 (1979)
27. Parikh, N., Boyd, S., et al.: Proximal algorithms. Found. Trends Optim. **1**(3), 127–239 (2014)
28. Qin, J., Lou, Y.: l_{1-2} regularized logistic regression. In: 2019 53rd Asilomar Conference on Signals, Systems, and Computers, pp. 779–783. IEEE (2019)
29. Rahimi, Y., Wang, C., Dong, H., Lou, Y.: A scale invariant approach for sparse signal recovery. SIAM J. Sci. Comput. **41**(6), A3649–A3672 (2019)
30. Santosa, F., Symes, W.W.: Linear inversion of band-limited reflection seismograms. SIAM J. Sci. Stat. Comp. **7**(4), 1307–1330 (1986)

31. Shen, X., Pan, W., Zhu, Y.: Likelihood-based selection and sharp parameter estimation. J. Am. Stat. Assoc. **107**(497), 223–232 (2012)
32. Tibshirani, R.: Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B **58**(1), 267–288 (1996)
33. Tillmann, A.M., Pfetsch, M.E.: The computational complexity of the restricted isometry property, the nullspace property, and related concepts in compressed sensing. IEEE Trans. Inf. Theory **60**(2), 1248–1259 (2013)
34. Tran, H., Webster, C.: A class of null space conditions for sparse recovery via nonconvex, non-separable minimizations. Results Appl. Math. **3**, 100011 (2019)
35. Wang, C., Yan, M., Rahimi, Y., Lou, Y.: Accelerated schemes for the l_1/l_2 minimization. IEEE Trans. Signal Process. **68**, 2660–2669 (2020)
36. Xu, Z., Chang, X., Xu, F., Zhang, H.: $l_{1/2}$ regularization: A thresholding representation theory and a fast solver. IEEE Trans. Neural Netw. Learn. Syst. **23**, 1013–1027 (2012)
37. Xu, Z., H., G., Yao, W., Zhang, H.: Representative of $l_{1/2}$ regularization among l_q $(0 < q < 1)$ regularizations: an experimental study based on phase diagram. Acta Automatica Sinica **38**(7), 1225–1228 (2012)
38. Yan, M.: A new primal-dual algorithm for minimizing the sum of three functions with a linear operator. J. Sci. Comput. **76**(3), 1698–1717 (2018)
39. Yin, P., Esser, E., Xin, J.: Ratio and difference of l_1 and l_2 norms and sparse representation with coherent dictionaries. Commun. Inf. Syst. **14**(2), 87–109 (2014)
40. Yin, P., Lou, Y., He, Q., Xin, J.: Minimization of ℓ_{1-2} for compressed sensing. SIAM J. Sci. Comput. **37**(1), A536–A563 (2015)
41. Zhang, C.: Nearly unbiased variable selection under minimax concave penalty. Ann. Stat. pp. 894–942 (2010)
42. Zhang, S., Xin, J.: Minimization of transformed L_1 penalty: Closed form representation and iterative thresholding algorithms. Commun. Math. Sci. **15**, 511–537 (2017)
43. Zhang, S., Xin, J.: Minimization of transformed L_1 penalty: theory, difference of convex function algorithm, and robust application in compressed sensing. Math. Program. **169**(1), 307–336 (2018)
44. Zhang, T.: Multi-stage convex relaxation for learning with sparse regularization. In: Adv. Neural Inf. Proces. Syst., pp. 1929–1936 (2009)