A SCALE INVARIANT APPROACH FOR SPARSE SIGNAL RECOVERY*

YAGHOUB RAHIMI†, CHAO WANG†, HONGBO DONG‡, AND YIFEI LOU†

Abstract. In this paper, we study the ratio of the L_1 and L_2 norms, denoted as L_1/L_2, to promote sparsity. Due to the non-convexity and non-linearity, there has been little attention to this scale-invariant metric. Compared to popular models in the literature such as the L_p model for $p \in (0, 1)$ and the transformed L_1 (TL1), this ratio model is parameter free. Theoretically, we present a weak null space property (wNSP) and prove that any sparse vector is a local minimizer of the L_1/L_2 model provided with this wNSP condition. Computationally, we focus on a constrained formulation that can be solved via the alternating direction method of multipliers (ADMM). Experiments show that the proposed approach is comparable to the state-of-the-art methods in sparse recovery. In addition, a variant of the L_1/L_2 model to apply on the gradient is also discussed with a proof-of-concept example of MRI reconstruction.

Key words. Sparsity, L_0, L_1, null space property, alternating direction method of multipliers, MRI reconstruction

AMS subject classifications. 90C90, 65K10, 49N45, 49M20

1. Introduction.

Sparse signal recovery is to find the sparsest solution of $A\mathbf{x} = \mathbf{b}$ where $\mathbf{x} \in \mathbb{R}^n, \mathbf{b} \in \mathbb{R}^m$, and $A \in \mathbb{R}^{m\times n}$ for $m \ll n$. This problem is often referred to as *compressed sensing* (CS) in the sense that the sparse signal $\mathbf{x}$ is compressible. Mathematically, this fundamental problem in CS can be formulated as

$$\min_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{x}\|_0 \quad \text{s.t.} \quad A\mathbf{x} = \mathbf{b}, \tag{1.1}$$

where $\|\mathbf{x}\|_0$ is the number of nonzero entries in $\mathbf{x}$. Unfortunately, (1.1) is NP-hard [29] to solve. A popular approach in CS is to replace L_0 by the convex L_1 norm, *i.e.*,

$$\min_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{x}\|_1 \quad \text{s.t.} \quad A\mathbf{x} = \mathbf{b}. \tag{1.2}$$

Computationally, there are various L_1 minimization algorithms such as primal dual [8], forward-backward splitting [34], and alternating direction method of multipliers (ADMM) [4].

A major breakthrough in CS was the *restricted isometry property* (RIP) [6], which provides a sufficient condition of minimizing the L_1 norm to recover the sparse signal. Another sufficient condition is given in terms of null space of the matrix A, thus referred to as *null space property* (NSP). In particular, Zhang [48] proved that if a vector $\mathbf{x}^*$ satisfies $A\mathbf{x}^* = \mathbf{b}$ and

$$\sqrt{\|\mathbf{x}^*\|_0} < \frac{1}{2} \min_{\mathbf{v}} \left\{ \frac{\|\mathbf{v}\|_1}{\|\mathbf{v}\|_2} : \mathbf{v} \in \ker(A)\backslash\{\mathbf{0}\} \right\}, \tag{1.3}$$

then $\mathbf{x}^*$ is an optimal solution to both (1.1) and (1.2). Unfortunately, neither RIP nor NSP can be computed to verify for a given matrix [1, 39].

*Submitted to the journal's Methods and Algorithms for Scientific Computing section December 18, 2018.

Funding: YR and YL were partially supported by NSF grant DMS-1522786.

†Department of Mathematical Sciences, University of Texas at Dallas, Richardson, TX 75080 (yxr160430@utdallas.edu, chaowang.hk@gmail.com, yifei.lou@utdallas.edu).

‡Department of Mathematics and Statistics, Washington State University, Pullman, WA 99164 (hongbo.dong@wsu.edu).

This manuscript is for review purposes only.

Alternatively, a computable condition for L_1's exact recovery is based on *coherence*, which is defined as

$$\mu(A) := \max_{i \neq j} \frac{|\mathbf{a}_i^T \mathbf{a}_j|}{\|\mathbf{a}_i\| \|\mathbf{a}_j\|}, \tag{1.4}$$

for a matrix $A = [\mathbf{a}_1, \cdots, \mathbf{a}_N]$. Donoho-Elad [12] and Gribonval [16] proved independently that if

$$\|\mathbf{x}\|_0 < \frac{1}{2}\left(1 + \frac{2}{\mu(A)}\right), \tag{1.5}$$

then the L_1 minimization (1.2) yields the sparsest solution of (1.1). Clearly, the coherence $\mu(A)$ is bounded by $[0, 1]$. The inequality (1.5) implies that L_1 may not perform well for highly coherent matrices, *i.e.*, $\mu(A) \sim 1$, as $\|\mathbf{x}\|_0$ is then at most one (often not feasible).

Other than the popular L_1 norm, there are a variety of regularization functionals to promote sparsity, such as L_p [9, 42, 21], L_1-L_2 [43, 23], capped L_1 (CL1) [47, 37], and transformed L_1 (TL1) [27, 45, 46]. Most of these models are nonconvex, leading to difficulties in proving exact recovery guarantees and algorithmic convergence, but they tend to give better empirical results compared to the convex L_1 approach. For example, it was reported in [43, 23] that L_p gives superior results for incoherent matrices (*i.e.*, $\mu(A)$ is small), while L_1-L_2 is the best for the coherent scenario. In addition, TL1 is always the second best no matter whether the matrix is coherent or not [45, 46].

In this paper, we study the ratio of L_1 and L_2 as a scale-invariant metric to approximate the desired L_0, which is scale-invariant itself. In one dimensional (1D) case (*i.e.*, $n = 1$), the L_1/L_2 model is exactly the same as the L_0 model if we use the convention $\frac{0}{0} = 0$. The ratio of L_1 and L_2 was first proposed by Hoyer [18] as a sparseness measure and later highlighted in [19] as a scale-invariant metric. However, there has been little attention on it due to its computational difficulties arisen from being non-convex and non-linear. There are some theorems that establish the equivalence between the L_1/L_2 and the L_0 models, but only restricted to nonnegative signals [14, 43]. We aim to apply this ratio model to arbitrary signals. On the other hand, the L_1/L_2 metric has an intrinsic drawback that it tends to produce one erroneously large coefficient while suppressing the other non-zero elements, under which case the ratio is reduced. To compensate for this drawback, it is helpful to incorporate a box constraint, which will also be addressed in this paper.

Now we turn to a sparsity-related assumption that signal is sparse after a given transform, as opposed to signal itself being sparse. This assumption is widely used in image processing. For example, a natural image, denoted by u, is mostly sparse after taking gradient, and hence it is reasonable to minimize the L_0 norm of the gradient, *i.e.*, $\|\nabla u\|_0$. To bypass the NP-hard L_0 norm, the convex relaxation replaces L_0 by L_1, where the L_1 norm of the gradient is the well-known total variation (TV) [36] of an image. A weighted L_1-αL_2 model (for $\alpha > 0$) on the gradient was proposed in [24], which suggested that $\alpha = 0.5$ yields better results than $\alpha = 1$ for image denoising, deblurring, and MRI reconstruction. The ratio of L_1 and L_2 on the image gradient was used in deconvolution and blind deconvolution [20, 35]. We further adapt the proposed ratio model from sparse signal recovery to imaging applications, specifically focusing on MRI reconstruction.

The rest of the paper is organized as follows. In Section 2, we give a brief review of related works in CS including models and corresponding algorithms. Section 3 is devoted to theoretical analysis of the L_1/L_2 model. In Section 4, we apply the ADMM to minimize the ratio of L_1 and L_2 with two variants of incorporating a box constraint as well as applying on the image gradient. We conduct extensive experiments in Section 5 to demonstrate the performance of the proposed

approaches over the state-of-the-art in sparse recovery and MRI reconstruction. Section 6 is a fun exercise, where we use the L_1/L_2 minimization to compute the right-hand-side of the NSP condition (1.3), leading to an empirical upper bound of the exact L_1 recovery guarantee. Finally, conclusions and future works are given in Section 7.

2. Previous Methods. We review the ADMM approach [4] for minimizing the constrained L_1 problem (1.2) and a difference of convex algorithm (DCA) [32, 33] that is applicable for both L_1-L_2 and TL1 models.

2.1. L_1 via ADMM. We rewrite the L_1 minimization (1.2) using an indicator function,

$$\min_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{x}\|_1 + I(A\mathbf{x}-\mathbf{b}), \tag{2.1}$$

where $I(\cdot)$ denotes an indicator function of the feasible set, *i.e.*,

$$I(A\mathbf{x}-\mathbf{b}) = \begin{cases} 0 & A\mathbf{x}=\mathbf{b}, \\ +\infty & A\mathbf{x}\neq\mathbf{b}. \end{cases} \tag{2.2}$$

By introducing an auxiliary variable $\mathbf{y}$, we further express (2.1) as

$$\min_{\mathbf{y}\in\mathbb{R}^n} \|\mathbf{x}\|_1 + I(A\mathbf{y}-\mathbf{b}) \quad \text{s.t.} \quad \mathbf{x}=\mathbf{y}. \tag{2.3}$$

The new formulation makes the objective function separable with respect to two variables $\mathbf{x}$ and $\mathbf{y}$. As a result, we form the augmented Lagrangian corresponding to (2.3) as

$$L_\rho(\mathbf{x},\mathbf{y};\mathbf{u}) = \|\mathbf{x}\|_1 + I(A\mathbf{y}-\mathbf{b}) + \mathbf{u}^T(\mathbf{x}-\mathbf{y}) + \frac{\rho}{2}\|\mathbf{x}-\mathbf{y}\|_2^2, \tag{2.4}$$

for a positive parameter ρ and a dual variable $\mathbf{u}$. We aim to update $\mathbf{x},\mathbf{y}$ in an alternating or sequential fashion, with a hope that each subproblem has a closed-form solution or can be calculated efficiently. This is a basic idea of the ADMM [4] that consists of three steps:

$$\begin{cases} \mathbf{x}^{(k+1)} = \arg\min_{\mathbf{x}} L_\rho(\mathbf{x},\mathbf{y}^{(k)};\mathbf{u}^{(k)}) = \arg\min_{\mathbf{x}} \|\mathbf{x}\|_1 + \frac{\rho}{2}\|\mathbf{x}-\mathbf{y}^{(k)}+\frac{\mathbf{u}^{(k)}}{\rho}\|_2^2, \\ \mathbf{y}^{(k+1)} = \arg\min_{\mathbf{y}} L_\rho(\mathbf{x}^{(k+1)},\mathbf{y};\mathbf{u}^{(k)}) = \arg\min_{\mathbf{y}} I(A\mathbf{y}-\mathbf{b}) + \frac{\rho}{2}\|\mathbf{x}^{(k+1)}-\mathbf{y}+\frac{\mathbf{u}^{(k)}}{\rho}\|_2^2, \\ \mathbf{u}^{(k+1)} = \mathbf{u}^{(k)} + \rho(\mathbf{x}^{(k+1)}-\mathbf{y}^{(k+1)}). \end{cases} \tag{2.5}$$

Both $\mathbf{x}$ and $\mathbf{y}$ subproblems have closed-form solutions. Specifically, the update for $\mathbf{x}$ is given by

$$\mathbf{x}^{(k+1)} = \mathbf{shrink}\left(\mathbf{y}^{(k)} - \frac{\mathbf{u}^{(k)}}{\rho}, \frac{1}{\rho}\right), \tag{2.6}$$

where **shrink** is often referred to as *soft shrinkage*,

$$\mathbf{shrink}(\mathbf{v},\mu)_i = \operatorname{sign}(v_i)\max\left(|v_i|-\mu, 0\right), \quad i=1,2,\cdots,n. \tag{2.7}$$

The update for $\mathbf{y}$ is a projection to the affine space of $A\mathbf{y}=\mathbf{b}$, *i.e.*,

$$\mathbf{y}^{(k+1)} = \left(I - A^T(AA^T)^{-1}A\right)(\mathbf{x}^{(k+1)} + \frac{\mathbf{u}^{(k)}}{\rho}) + A^T(AA^T)^{-1}\mathbf{b}. \tag{2.8}$$

2.2. DCA Minimization. We describe a general framework of the difference of convex algorithm (DCA) [32, 33]. The DCA is applicable when the objective function has the form of difference of convex (DC) functions, *i.e.*,

$$\min_{\mathbf{x}\in\mathbb{R}^n} g(\mathbf{x}) - h(\mathbf{x}). \tag{2.9}$$

By linearizing the second term $h(\cdot)$, the DCA iterates as follows,

$$\mathbf{x}^{(k+1)} = \arg\min_{\mathbf{x}\in\mathbb{R}^n} g(\mathbf{x}) - \left\langle \mathbf{x}, \nabla h(\mathbf{x}^{(k)}) \right\rangle. \tag{2.10}$$

If $h(\cdot)$ is not differentiable at $\mathbf{x}^k$, then the gradient $\nabla h(\mathbf{x}^{(k)})$ can be replaced by any subgradient at $\mathbf{x}^k$. Due to the flexibility of DC decomposition, g and h can be assumed to be strongly convex without loss of generality, in this case any limit point of the generated sequence is a *critical point* of (2.9) [32, 33]. On the other hand, there may still exist directions at critical points where the objective functions can be improved locally (*e.g.*, one can find an univariate convex example [11, Example 2]). Some recent works focus on a sharper notion of d(irectional)-stationarity where no local improvement can be made [31, 25, 38], which is beyond the scope of the current paper.

We apply the DCA for two sparse promoting metrics: L_1-L_2 [23, 44] and transformed L_1 (TL1) [45, 46]. For practical performance we simply use a natural DC decomposition such that the convex subproblems in iterations are linear programs. In particular, the L_1-L_2 model [23, 44] can be expressed as

$$\min_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{x}\|_1 - \|\mathbf{x}\|_2 \quad \text{s.t.} \quad A\mathbf{x} = \mathbf{b}. \tag{2.11}$$

We take $g(\mathbf{x}) = \|\mathbf{x}\|_1 + I(A\mathbf{x}-\mathbf{b})$ and $h(\mathbf{x}) = \|\mathbf{x}\|_2$. Since $\nabla h(\mathbf{x}) = \frac{\mathbf{x}}{\|\mathbf{x}\|_2}$, the DCA goes as follows,

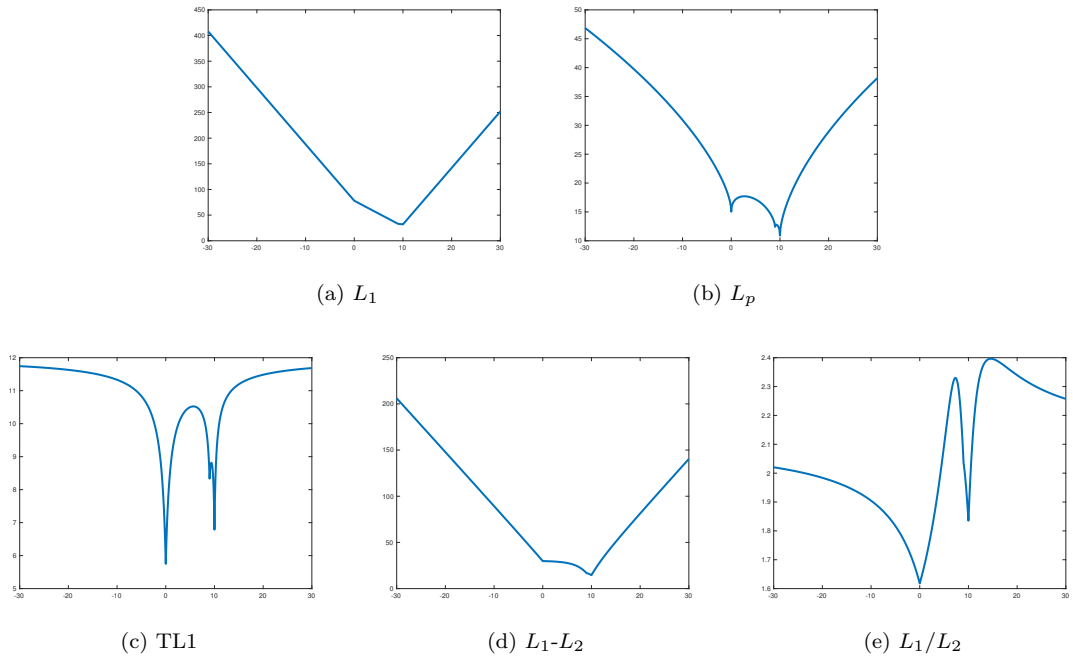
$$\mathbf{x}^{(k+1)} = \arg\min_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{x}\|_1 + I(A\mathbf{x}-\mathbf{b}) - \left\langle \mathbf{x}, \frac{\mathbf{x}^{(k)}}{\|\mathbf{x}^{(k)}\|_2} \right\rangle, \tag{2.12}$$

which is an L_1-type of problem and can be solved by the ADMM. On the other hand, the TL1 model is defined as

$$\min_{\mathbf{x}\in\mathbb{R}^n} \sum_{i=1}^n \rho_a(x_i) + I(A\mathbf{x}-\mathbf{b}), \quad \text{with } \rho_a(t) = \frac{(1+a)|t|}{a+|t|}, \tag{2.13}$$

for $a > 0$. TL1 approaches to L_0 as $a \to 0$ and approaches to L_1 as $a \to \infty$. TL1 can be solved by either DCA [46] or iterative thresholding [45], but DCA tends to give better results. Consequently, we consider the DC decomposition as $g(\mathbf{x}) = \|\mathbf{x}\|_1 + I(A\mathbf{x}-\mathbf{b})$ and $h(\mathbf{x}) = \sum_{i=1}^n \left(\frac{x_i^2}{a+|x_i|}\right)$. By substituting $\nabla h(\mathbf{x}) = \frac{\mathbf{x}(2a+|\mathbf{x}|)}{(a+|\mathbf{x}|)^2}$ into the general DCA iterations (2.10), we obtain a numerical scheme for minimizing the TL1. Due to the presence of the L_1 term in $g(\cdot)$, the DCA scheme for minimizing TL1 consists of a set of L_1-type subproblems that can be solved efficiently. Note that both TL1 and L_1-L_2 do not require strongly convex DC decompositions to guarantee the convergence of DCA [46, 22], thanks to their specific forms.

3. Rationales of the L_1/L_2 metric. We begin with a toy example to illustrate the advantages of L_1/L_2 over other metrics, followed by some theoretical properties of the proposed model.



(a) L_1 (b) L_p

(c) TL1 (d) L_1-L_2 (e) L_1/L_2

FIG. 1. *The objective functions of a toy example illustrate that only L_1/L_2 and TL1 can find $t = 0$ as the global minimizer, but TL1 has a very narrow basin of attraction (thus sensitive to initial guess and difficult to find the global solution.).*

3.1. A toy example. Define a matrix A as

$$A := \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & -1 \end{bmatrix} \in \mathbb{R}^{5\times 6}, \tag{3.1}$$

and $\mathbf{b} = (0, 0, 20, 40, 18)^T \in \mathbb{R}^5$. It is straightforward that any general solutions of $A\mathbf{x} = \mathbf{b}$ have the form of $\mathbf{x} = (t, t, t, 20 - 2t, 40 - 4t, 2(t - 9))^T$ for a scalar $t \in \mathbb{R}$. The sparsest solution occurs at $t = 0$, where the sparsity of $\mathbf{x}$ is 3 and some local solutions include $t = 10$ for sparsity being 4 and $t = 9$ for sparsity being 5. In Figure 1, we plot various objective functions with respect to t, including L_1, L_p (for $p = 1/2$), L_1-L_2, and TL1 (for $a = 1$ as suggested in [46]). Note that all these functions are not differentiable at the values of $t = 0, 9$, and 10, where the sparsity of $\mathbf{x}$ is strictly smaller than 6. The sparsest vector $\mathbf{x}$ corresponding to $t = 0$ can only be found by minimizing TL1 and L_1/L_2, while the other models find $t = 10$ as a global minimum. One drawback of TL1 is that the function is very narrow around the critical points, which implies it is sensitive to initial guess and difficult to find the global solution. In addition, TL1 has more local minimizers than our L_1/L_2 model (3 versus 2).

3.2. Theoretical properties. Our theoretical analysis is based on the null space property (NSP). To make the paper self-contained, we give a review of the NSP as a necessary and sufficient condition for the L_1 exact recovery. Then we present a weaker form that is related to local minimizers of the L_1/L_2 model.

DEFINITION 3.1 (null space property [10]). *For any matrix $A \in \mathbb{R}^{m\times n}$, we say the matrix A satisfies a null space property (NSP) of order s if*

$$\|\mathbf{v}_S\|_1 < \|\mathbf{v}_{\bar{S}}\|_1, \ \mathbf{v} \in \ker(A)\backslash\{\mathbf{0}\}, \ \forall S \subset [n], \ |S| \le s. \tag{3.2}$$

Donoho and Huo [13] proved that every s-sparse signal $\mathbf{x} \in \mathbb{R}^n$ is the unique solution to the L_1 minimization (1.2) if and only if A satisfies the NSP of order s. Note that NSP is no longer necessary if "every s-sparse vector" is relaxed. A sufficient condition for the exact L_1 recovery is given by (1.3) in terms of L_1/L_2, which will be examined numerically in Section 6.

Recently, Tran and Webster [40] generalized the NSP to deal with sparse promoting metrics that are symmetric, separable and concave, which unfortunately does not apply to L_1/L_2 (not separable), but this work motivates us to consider a weaker form of the NSP, as defined in Definition 3.2.

DEFINITION 3.2. *For any matrix $A \in \mathbb{R}^{m\times n}$, we say the matrix A satisfies a weak null space property (wNSP) of order s if*

$$(s+1)\,\|\mathbf{v}_S\|_1 \le \|\mathbf{v}_{\bar{S}}\|_1, \ \mathbf{v} \in \ker(A)\backslash\{\mathbf{0}\}, \ \forall S \subset [n], \ |S| \le s. \tag{3.3}$$

Note that Definition 3.2 is weaker than the original NSP in Definition 3.1 in the sense that if a matrix satisfies wNSP then it also satisfies the NSP. The following theorem says that any s-sparse vector is a local minimizer of L_1/L_2 provided the matrix has the wNSP of order s. The proof is given in Appendix.

THEOREM 3.3. *Assume an $m \times n$ matrix A satisfies the wNSP of order s, then any s-sparse solution of $A\mathbf{x} = \mathbf{b}$ ($\mathbf{b} \neq \mathbf{0}$) is a local minimum for L_1/L_2 in the feasible space of $A\mathbf{x} = \mathbf{b}$.* i.e., *there exists a positive number $t^* > 0$ such that for every $\mathbf{v} \in \ker(A)$ with $0 < \|\mathbf{v}\|_2 \le t^*$ we have*

$$\frac{\|\mathbf{x}\|_1}{\|\mathbf{x}\|_2} \le \frac{\|\mathbf{x}+\mathbf{v}\|_1}{\|\mathbf{x}+\mathbf{v}\|_2}. \tag{3.4}$$

Finally, we show the optimal value of the L_1/L_2 subject to $A\mathbf{x} = \mathbf{b}$ is upper bounded by the same ratio with $\mathbf{b} = \mathbf{0}$; see Proposition 1.

PROPOSITION 1. *For any $A \in \mathbb{R}^{m\times n}, \mathbf{x} \in \mathbb{R}^n$, we have*

$$\inf_{\mathbf{z}\in\mathbb{R}^n}\left\{\frac{\|\mathbf{z}\|_1}{\|\mathbf{z}\|_2}\,\middle|\, A\mathbf{z} = A\mathbf{x}\right\} \le \inf_{\mathbf{z}\in\mathbb{R}^n}\left\{\frac{\|\mathbf{z}\|_1}{\|\mathbf{z}\|_2}\,\middle|\, \mathbf{z} \in \ker(A)\setminus\{\mathbf{0}\}\right\}. \tag{3.5}$$

Proof. Denote

$$\alpha^* = \inf_{\mathbf{z}\in\mathbb{R}^n}\left\{\frac{\|\mathbf{z}\|_1}{\|\mathbf{z}\|_2}\,\middle|\, A\mathbf{z} = A\mathbf{x}\right\}. \tag{3.6}$$

For every $\mathbf{v} \in \ker(A)\setminus\{\mathbf{0}\}$ and $t \in \mathbb{R}$, we have that

$$\alpha^* \le \frac{\|\mathbf{x}+t\mathbf{v}\|_1}{\|\mathbf{x}+t\mathbf{v}\|_2}, \tag{3.7}$$

since $A(\mathbf{x}+t\mathbf{v})=\mathbf{b}$. Then we obtain

$$\lim_{t\to\infty}\frac{\|\mathbf{x}+t\mathbf{v}\|_1}{\|\mathbf{x}+t\mathbf{v}\|_2}=\lim_{t\to\infty}\frac{\|\mathbf{x}/t+\mathbf{v}\|_1}{\|\mathbf{x}/t+\mathbf{v}\|_2}=\frac{\|\mathbf{v}\|_1}{\|\mathbf{v}\|_2}. \tag{3.8}$$

Therefore, for every $\mathbf{v}\in\ker(A)\setminus\{\mathbf{0}\}$,

$$\alpha^*\le\frac{\|\mathbf{v}\|_1}{\|\mathbf{v}\|_2}, \tag{3.9}$$

which directly leads to the inequality between these two infimums in (3.5). ∎

Note that the left-hand-side of the inequality involves both the underlying signal $\mathbf{x}$ and the system matrix A, which can be upper bounded by the minimum ratio that only involves A. This relationship will be numerically checked in Section 6.

4. Numerical schemes.

The proposed model is

$$\min_{\mathbf{x}\in\mathbb{R}^n}\left\{\frac{\|\mathbf{x}\|_1}{\|\mathbf{x}\|_2}+I(A\mathbf{x}-\mathbf{b})\right\}. \tag{4.1}$$

In Subsection 4.1, we detail the ADMM algorithm for minimizing (4.1), followed by a minor change to incorporate additional box constraint in Subsection 4.2. We discuss in Subsection 4.3 another variant of L_1/L_2 on the gradient to deal with imaging applications.

4.1. The L_1/L_2 minimization via ADMM.

In order to apply the ADMM [4] to solve for (4.1), we introduce two auxiliary variables and rewrite (4.1) into an equivalent form,

$$\min_{\mathbf{x},\mathbf{y},\mathbf{z}}\left\{\frac{\|\mathbf{z}\|_1}{\|\mathbf{y}\|_2}+I(A\mathbf{x}-\mathbf{b})\right\}\quad\text{s.t.}\quad\mathbf{x}=\mathbf{y},\quad\mathbf{x}=\mathbf{z}, \tag{4.2}$$

where $I(\cdot)$ is the indicator function defined in (2.2). The augmented Lagrangian for (4.2) is

$$\begin{aligned}L_{\rho_1,\rho_2}(\mathbf{x},\mathbf{y},\mathbf{z};\mathbf{v},\mathbf{w})&=\frac{\|\mathbf{z}\|_1}{\|\mathbf{y}\|_2}+I(A\mathbf{x}-\mathbf{b})+\langle\mathbf{v},\mathbf{x}-\mathbf{y}\rangle+\frac{\rho_1}{2}\|\mathbf{x}-\mathbf{y}\|_2^2+\langle\mathbf{w},\mathbf{x}-\mathbf{z}\rangle+\frac{\rho_2}{2}\|\mathbf{x}-\mathbf{z}\|_2^2\\&=\frac{\|\mathbf{z}\|_1}{\|\mathbf{y}\|_2}+I(A\mathbf{x}-\mathbf{b})+\frac{\rho_1}{2}\left\|\mathbf{x}-\mathbf{y}+\frac{1}{\rho_1}\mathbf{v}\right\|_2^2+\frac{\rho_2}{2}\left\|\mathbf{x}-\mathbf{z}+\frac{1}{\rho_2}\mathbf{w}\right\|_2^2.\end{aligned} \tag{4.3}$$

The ADMM consists of the following five steps:

$$\begin{cases}\mathbf{x}^{(k+1)}:=\arg\min_{\mathbf{x}}L_{\rho_1,\rho_2}(\mathbf{x},\mathbf{y}^{(k)},\mathbf{z}^{(k)};\mathbf{v}^{(k)},\mathbf{w}^{(k)}),\\ \mathbf{y}^{(k+1)}:=\arg\min_{\mathbf{y}}L_{\rho_1,\rho_2}(\mathbf{x}^{(k+1)},\mathbf{y},\mathbf{z}^{(k)};\mathbf{v}^{(k)},\mathbf{w}^{(k)}),\\ \mathbf{z}^{(k+1)}:=\arg\min_{\mathbf{z}}L_{\rho_1,\rho_2}(\mathbf{x}^{(k+1)},\mathbf{y}^{(k+1)},\mathbf{z};\mathbf{v}^{(k)},\mathbf{w}^{(k)}),\\ \mathbf{v}^{(k+1)}:=\mathbf{v}^{(k)}+\rho_1(\mathbf{x}^{(k+1)}-\mathbf{y}^{(k+1)}),\\ \mathbf{w}^{(k+1)}:=\mathbf{w}^{(k)}+\rho_2(\mathbf{x}^{(k+1)}-\mathbf{z}^{(k+1)}).\end{cases} \tag{4.4}$$

The update for $\mathbf{x}$ is a projection to the affine space of $A\mathbf{x}=\mathbf{b}$, similar to (2.8),

$$\begin{aligned}\mathbf{x}^{(k+1)}&:=\arg\min_{\mathbf{x}}L_{\rho_1,\rho_2}(\mathbf{x},\mathbf{y}^{(k)},\mathbf{z}^{(k)};\mathbf{v}^{(k)},\mathbf{w}^{(k)})\\&=\arg\min_{\mathbf{x}}\left\{\frac{\rho_1+\rho_2}{2}\left\|\mathbf{x}-\mathbf{f}^{(k)}\right\|_2^2\quad\text{s.t.}\quad A\mathbf{x}=\mathbf{b}\right\}\\&=\left(I-A^T(AA^T)^{-1}A\right)\mathbf{f}^{(k)}+A^T(AA^T)^{-1}\mathbf{b},\end{aligned}$$

where

$$\mathbf{f}^{(k)} = \frac{\rho_1}{\rho_1+\rho_2}\left(\mathbf{y}^{(k)} - \frac{1}{\rho_1}\mathbf{v}^{(k)}\right) + \frac{\rho_2}{\rho_1+\rho_2}\left(\mathbf{z}^{(k)} - \frac{1}{\rho_2}\mathbf{w}^{(k)}\right). \tag{4.5}$$

As for the $\mathbf{y}$-subproblem, let $c^{(k)} = \|\mathbf{z}^{(k)}\|_1$ and $\mathbf{d}^{(k)} = \mathbf{x}^{(k+1)} + \frac{\mathbf{v}^{(k)}}{\rho_1}$ and the minimization subproblem reduces to

$$\mathbf{y}^{(k+1)} = \arg\min_{\mathbf{y}} \frac{c^{(k)}}{\|\mathbf{y}\|_2} + \frac{\rho_1}{2}\|\mathbf{y}-\mathbf{d}^{(k)}\|_2^2. \tag{4.6}$$

If $\mathbf{d}^{(k)} = 0$ then any vector $\mathbf{y}$ with $\|\mathbf{y}\|_2 = \sqrt[3]{\frac{c^{(k)}}{\rho_1}}$ is a solution to the minimization problem. If $c^{(k)} = 0$ then $\mathbf{y} = \mathbf{d}^{(k)}$ is the solution. Now we consider $\mathbf{d}^{(k)} \neq 0$ and $c^{(k)} \neq 0$. By taking derivative of the objective function with respect to $\mathbf{y}$, we obtain

$$\left(-\frac{c^{(k)}}{\|\mathbf{y}\|_2^3} + \rho_1\right)\mathbf{y} = \rho_1\mathbf{d}^{(k)}.$$

As a result, there exists a positive number $\tau^{(k)} \geq 0$ such that $\mathbf{y} = \tau^{(k)}\mathbf{d}^{(k)}$. Given $\mathbf{d}^{(k)}$, we denote $\eta^{(k)} = \|\mathbf{d}^{(k)}\|_2$. For $\eta^{(k)} > 0$, finding $\mathbf{y}$ becomes a one-dimensional search for the parameter $\tau^{(k)}$. In other words, if we take $D^{(k)} = \frac{c^{(k)}}{\rho_1(\eta^{(k)})^3}$, then $\tau^{(k)}$ is a root of

$$\underbrace{\tau^3 - \tau^2 - D^{(k)}}_{F(\tau)} = 0.$$

The cubic-root formula suggests that $F(\tau) = 0$ has only one real root, which can be found by the following closed-form solution.

$$\tau^{(k)} = \frac{1}{3} + \frac{1}{3}(C^{(k)} + \frac{1}{C^{(k)}}), \text{ with } C^{(k)} = \sqrt[3]{\frac{27D^{(k)} + 2 + \sqrt{(27D^{(k)}+2)^2 - 4}}{2}}. \tag{4.7}$$

In summary, we have

$$\mathbf{y}^{(k+1)} = \begin{cases} \mathbf{e}^{(k)} & \mathbf{d}^{(k)} = 0, \\ \tau^{(k)}\mathbf{d}^{(k)} & \mathbf{d}^{(k)} \neq 0, \end{cases} \tag{4.8}$$

where $\mathbf{e}^{(k)}$ is a random vector with the L_2 norm to be $\sqrt[3]{\frac{c^{(k)}}{\rho_1}}$.

Finally, the ADMM update for $\mathbf{z}$ is given by soft shrinkage operator

$$\mathbf{z}^{(k+1)} = \mathbf{shrink}\left(\mathbf{x}^{(k+1)} + \frac{\mathbf{w}^{(k)}}{\rho_2}, \frac{1}{\rho_2\|\mathbf{y}^{(k+1)}\|_2}\right), \tag{4.9}$$

where **shrink** is defined in (2.7). We summarize the ADMM algorithm for solving the L_1/L_2 minimization problem in Algorithm 4.1.

Algorithm 4.1 The L_1/L_2 minimization via ADMM.

Input: $A \in \mathbb{R}^{m\times n}, \mathbf{b} \in \mathbb{R}^{m\times 1}$, Max and $\epsilon \in \mathbb{R}$
while $k <$ Max or $\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|_2/\|\mathbf{x}^{(k)}\| > \epsilon$ **do**
$\mathbf{x}^{(k+1)} = \left(I - A^T(AA^T)^{-1}A\right)\mathbf{f}^{(k)} + A^T(AA^T)^{-1}\mathbf{b}$
$\mathbf{y}^{(k+1)} = \begin{cases} \mathbf{e}^{(k)} & \mathbf{d}^{(k)} = 0 \\ \tau^{(k)}\mathbf{d}^{(k)} & \mathbf{d}^{(k)} \neq 0 \end{cases}$
$\mathbf{z}^{(k+1)} = \mathbf{shrink}\left(\mathbf{x}^{(k+1)} + \frac{\mathbf{w}^{(k)}}{\rho_2}, \frac{1}{\rho_2\|\mathbf{y}^{(k+1)}\|_2}\right)$
$\mathbf{v}^{(k+1)} = \mathbf{v}^{(k)} + \rho_1(\mathbf{x}^{(k+1)} - \mathbf{y}^{(k+1)})$
$\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} + \rho_2(\mathbf{x}^{(k+1)} - \mathbf{z}^{(k+1)})$
k = k+1
end while
return $\mathbf{x}^{(k)}$

4.2. L_1/L_2 with box constraint. The L_1/L_2 model has an intrinsic drawback that tends to produce one erroneously large coefficient while suppressing the other non-zero elements, under which case the ratio is reduced. To compensate for this drawback, it is helpful to incorporate a box constraint, if we know lower/upper bounds of the underying signal *a priori*. Specifically, we propose

$$\min_{\mathbf{x}\in\mathbb{R}^n} \left\{ \frac{\|\mathbf{x}\|_1}{\|\mathbf{x}\|_2} + I(A\mathbf{x}-\mathbf{b}) \,\middle|\, \mathbf{x} \in [c,d] \right\}, \tag{4.10}$$

which is referred to as L_1/L_2-box. Similar to (4.2), we look at the following form that enforces the box constraint on variable $\mathbf{z}$,

$$\min_{\mathbf{x},\mathbf{y},\mathbf{z}} \left\{ \frac{\|\mathbf{z}\|_1}{\|\mathbf{y}\|_2} + I(A\mathbf{x}-\mathbf{b}) \right\} \quad \text{s.t.} \quad \mathbf{x}=\mathbf{y}, \quad \mathbf{x}=\mathbf{z}, \quad \mathbf{z}\in[c,d]. \tag{4.11}$$

The only change we need to make by adapting Algorithm 4.1 to the L_1/L_2-box is the $\mathbf{z}$ update. The $\mathbf{z}$-subproblem in (4.4) with the box constraint is

$$\min_{\mathbf{z}} \frac{1}{\|\mathbf{y}^{(k+1)}\|_2}\|\mathbf{z}\|_1 + \frac{\rho_2}{2}\|\mathbf{x}^{(k+1)} - \mathbf{z} + \frac{1}{\rho_2}\mathbf{w}^{(k)}\|_2^2 \quad \text{s.t.} \quad \mathbf{z}\in[c,d]. \tag{4.12}$$

For a convex problem (4.12) involving the L_1 norm, it has a closed-form solution given by the soft shrinkage, followed by projection to the interval $[c,d]$. In particular, simple calculations show that

$$z_i^{(k+1)} = \min\left\{\max(\hat{z}_i, c), d\right\}, \quad i = 1, 2, \cdots, n, \tag{4.13}$$

where $\hat{\mathbf{z}} = \mathbf{shrink}\,(\mathbf{r}, \nu)$, $\mathbf{r} = \mathbf{x}^{(k+1)} + \frac{\mathbf{w}^{(k)}}{\rho_2}$ and $\nu = \frac{1}{\rho_2\|\mathbf{y}^{(k+1)}\|}$. If the box constraint $[c,d]$ is symmetric, *i.e.*, $c = -d$ and $d > 0$, it follows from [2] that the update for $\mathbf{z}$ can be expressed as

$$z_i^{(k+1)} = \text{sign}(v_i)\min\left\{\max(|r_i| - \nu, 0), d\right\}, \quad i = 1, 2, \cdots, n. \tag{4.14}$$

4.3. L_1/L_2 on the gradient. We adapt the L_1/L_2 model to apply on the gradient, which enables us to deal with imaging applications. Let $u \in \mathbb{R}^{n\times m}$ be an underlying image of size $n \times m$. Denote A as a linear operator that models a certain degradation process to obtain the measured data f. For example, A can be a subsampling operator in the frequency domain and recovering u from f is called MRI reconstruction. In short, the proposed gradient model is given by

$$\min_{u\in\mathbb{R}^{n\times m}} \frac{\|\nabla u\|_1}{\|\nabla u\|_2} \quad \text{s.t.} \quad Au = f, \tag{4.15}$$

where ∇ denotes discrete gradient operator $\nabla u := \{[u_{ij} - u_{(i+1)j}]_{i=1}^n\}_{j=1}^m, \{[u_{ij} - u_{i(j+1)}]_{j=1}^m\}_{i=1}^n$ with periodic boundary condition; hence the model is referred to as L_1/L_2-grad. To solve for (4.15), we introduce two auxiliary variables $\mathbf{d}$ and $\mathbf{h}$ to hold the value of ∇u, thus leading to

$$\min_{u\in\mathbb{R}^{n\times m}} \frac{\|\mathbf{d}\|_1}{\|\mathbf{h}\|_2} \quad \text{s.t.} \quad Au = f,\ \mathbf{d} = \nabla u,\ \mathbf{h} = \nabla u. \tag{4.16}$$

Note that we denote $\mathbf{d}$ and $\mathbf{h}$ in bold to indicate that they have two components corresponding to both x and y derivatives. The augmented Lagrangian is expressed as

$$\mathcal{L}_{\lambda,\rho_1,\rho_2}(u,\mathbf{d},\mathbf{h};w,\mathbf{b},\mathbf{g}) = \frac{\|\mathbf{d}\|_1}{\|\mathbf{h}\|_2} + \frac{\lambda}{2}\|Au - f - w\|_2^2 + \frac{\rho_1}{2}\|\mathbf{d} - \nabla u - \mathbf{b}\|_2^2 + \frac{\rho_2}{2}\|\mathbf{h} - \nabla u - \mathbf{g}\|_2^2, \tag{4.17}$$

where $w, \mathbf{b}, \mathbf{g}$ are dual variables and λ, ρ_1, ρ_2 are positive parameters. The updates for $\mathbf{d}, \mathbf{h}$ are the same as Algorithm 4.1. Specifically for $\mathbf{h}$, we consider $D^{(k)} = \frac{\|\mathbf{d}\|_1}{\rho_2\|\nabla u^{(k+1)} + \mathbf{g}^{(k)}\|_2^3}$ and hence $\tau^{(k)}$ is the root of the same polynomial as in (4.7). By taking derivative of (4.17) with respect to u, we can obtain the u-update, *i.e.*,

$$u^{(k+1)} = (\lambda A^TA - (\rho_1+\rho_2)\triangle)^{-1}(\lambda A^T(f+w) + \rho_1\nabla^T(\mathbf{d}-\mathbf{b}) + \rho_2\nabla^T(\mathbf{h}-\mathbf{g})). \tag{4.18}$$

Note for certain operator A, the inverse in the u-update (4.18) can be computed efficiently via the fast Fourier transform (FFT). In summary, we present the ADMM algorithm for the L_1/L_2-grad model in Algorithm 4.2.

Algorithm 4.2 The L_1/L_2-grad minimization via ADMM.

Input: $f \in \mathbb{R}^{n\times m}$, A, Max and $\epsilon \in \mathbb{R}$.
while $k <$ Max or $\|u^{(k)} - u^{(k-1)}\|_2/\|u^{(k)}\| > \epsilon$ **do**

$$u^{(k+1)} = (\lambda A^TA - (\rho_1+\rho_2)\triangle)^{-1}(\lambda A^T(f+w^{(k)}) + \rho_1\nabla^T(\mathbf{d}^{(k)} - \mathbf{b}^{(k)}) + \rho_2\nabla^T(\mathbf{h}^{(k)} - \mathbf{g}^{(k)}))$$

$$\mathbf{h}^{(k+1)} = \begin{cases} \mathbf{e}^{(k)} & \nabla u^{(k+1)} + \mathbf{g}^{(k)} = 0, \\ \tau^{(k)}\left(\nabla u^{(k+1)} + \mathbf{g}^{(k)}\right) & \nabla u^{(k+1)} + \mathbf{g}^{(k)} \neq 0. \end{cases}$$

$$\mathbf{d}^{(k+1)} = \mathbf{shrink}\left(\nabla u^{(k+1)} + \mathbf{b}^{(k)}, \tfrac{1}{\rho_1\|\mathbf{h}^{(k+1)}\|_2}\right)$$

$$\mathbf{b}^{(k+1)} = \mathbf{b}^{(k)} + \nabla u^{(k+1)} - \mathbf{d}^{(k+1)}$$

$$\mathbf{g}^{(k+1)} = \mathbf{g}^{(k)} + \nabla u^{(k+1)} - \mathbf{h}^{(k+1)}$$

$$w^{(k+1)} = w^{(k)} + f - Au^{(k+1)}$$

$k = k + 1$
end while
return $u^{(k)}$

5. Numerical experiments. In this section, we carry out a series of numerical tests to demonstrate the performance of the proposed L_1/L_2 models together with its corresponding algorithms. All the numerical experiments are conducted on a standard desktop with CPU (Intel i7-6700, 3.4GHz) and MATLAB 9.2 (R2017a).

We consider two types of sensing matrices: one is called oversampled discrete cosine transform (DCT) and the other is Gaussian matrix. Specifically for the oversampled DCT, we follow the works of [15, 23, 44] to define $A = [\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_n] \in \mathbb{R}^{m\times n}$ with

$$\mathbf{a}_j := \frac{1}{\sqrt{m}} \cos\left(\frac{2\pi \mathbf{w} j}{F}\right), \quad j = 1, \cdots, n, \tag{5.1}$$

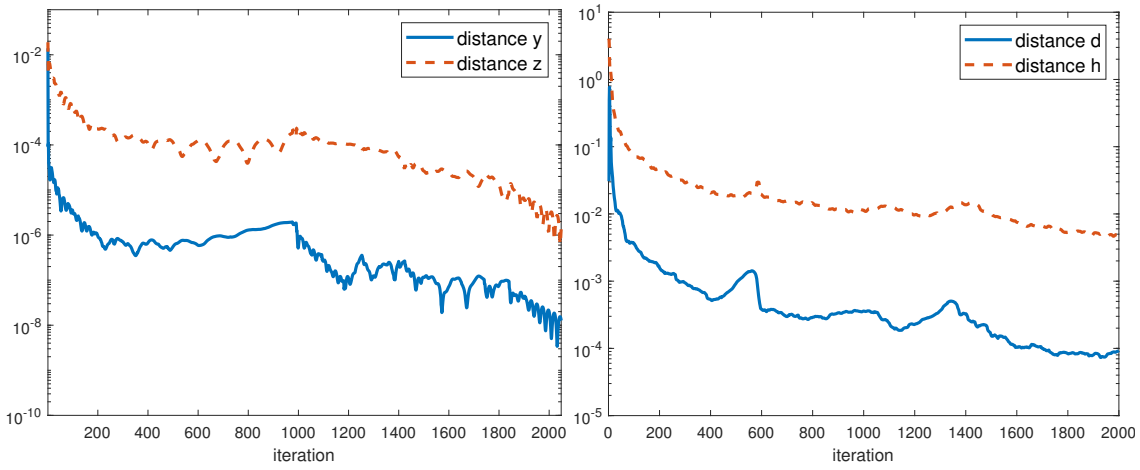
where $\mathbf{w}$ is a random vector uniformly distributed in $[0, 1]^m$ and $F \in \mathbb{R}_+$ controls the coherence in a way that a larger value of F yields a more coherent matrix. In addition, we use $\mathcal{N}(\mathbf{0}, \Sigma)$ (the multi-variable normal distribution) to generate Gaussian matrix, where the covariance matrix is $\Sigma = \{(1 - r) * I(i = j) + r\}_{i,j}$ with a positive parameter r. This type of matrices is used in the TL1 paper [46], which mentioned that a larger r value indicates a more difficult problem in sparse recovery. Throughout the experiments, we consider sensing matrices of size 64×1024. The ground truth $\mathbf{x} \in \mathbb{R}^n$ is simulated as s-sparse signal, where s is the total number of nonzero entries. The support of $\mathbf{x}$ is a random index set and the values of non-zero elements follow Gaussian normal distribution *i.e.*, $(\mathbf{x}_s)_i \sim \mathcal{N}(0, 1), \quad i = 1, 2, \cdots, s$. We then normalize the ground-truth signal to have maximum magnitude as 1 so that we can examine the performance of additional $[-1, 1]$ box constraint.

Due to the non-convex nature of the proposed L_1/L_2 model, the initial guess $\mathbf{x}^{(0)}$ is very important and should be well-chosen. A typical choice is the L_1 solution (1.2), which is used here. Although the L_1 minimization can be solved by the ADMM (as described in Subsection 2.1), we adopt a commercial optimization software called Gurobi [30] to minimize the L_1 norm via linear programming for the sake of efficiency. The stop criterion is when the relative error of $\mathbf{x}^{(k)}$ to $\mathbf{x}^{(k-1)}$ is smaller than 10^{-8} or iterative number exceeds $10n$.

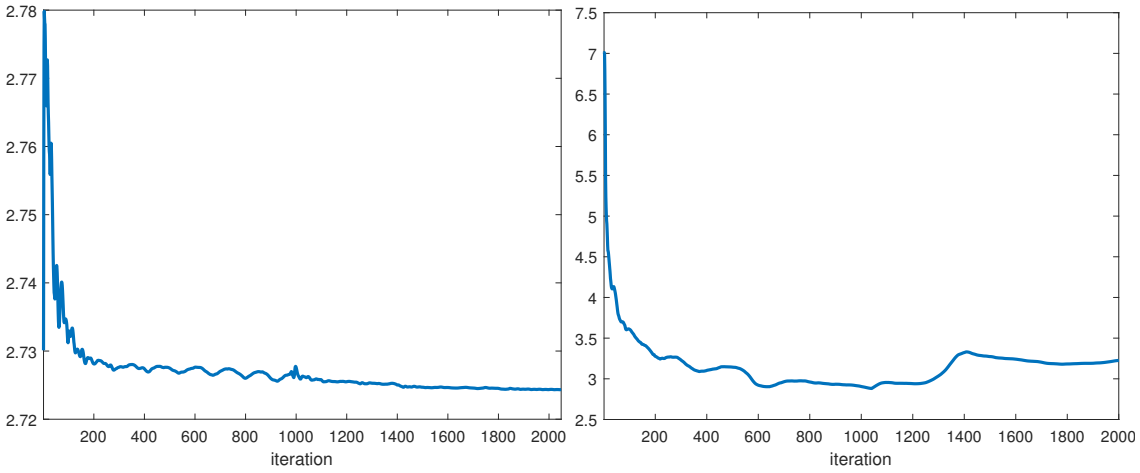
5.1. Algorithmic behaviors. We empirically demonstrate the convergence of the proposed ADMM algorithms in Figure 2. Specifically we examine the L_1/L_2 minimization problem (4.2), where the sensing matrix is an oversampled DCT matrix with $F = 10$ and ground-truth sparse vector has 12 non-zero elements. We also study the MRI reconstruction from 7 radical lines as a particular sparse gradient problem that involves the L_1/L_2-grad minimization of (4.15) by Algorithm 4.2.

There are two auxiliary variables $\mathbf{y}$ and $\mathbf{z}$ in L_1/L_2 such that $\mathbf{x} = \mathbf{y} = \mathbf{z}$, while two auxiliary variables $\mathbf{d}, \mathbf{h}$ are in L_1/L_2-grad for $\nabla u = \mathbf{d} = \mathbf{h}$. We show in the top row of Figure 2 the values of $\left\|\mathbf{x}^{(k)} - \mathbf{y}^{(k)}\right\|_2$ and $\left\|\mathbf{x}^{(k)} - \mathbf{z}^{(k)}\right\|_2$ as well as $\left\|\nabla u^{(k)} - \mathbf{d}^{(k)}\right\|_2$ and $\left\|\nabla u^{(k)} - \mathbf{h}^{(k)}\right\|_2$, all are plotted with respect to the iteration counter k. The bottom row of Figure 2 is for objective functions, *i.e.*, $\left\|\mathbf{x}^{(k)}\right\|_1 / \left\|\mathbf{x}^{(k)}\right\|_2$ and $\left\|\nabla u^{(k)}\right\|_1 / \left\|\nabla u^{(k)}\right\|_2$ for L_1/L_2 and L_1/L_2-grad, respectively. All the plots in Figure 2 decrease rapidly with respect to iteration counters, which serves as heuristic evidence of algorithmic convergence. On the other hand, the objective functions in Figure 2 look oscillatory. This phenomenon implies difficulties in theoretically proving the convergence, as one key step in the convergence proof requires to show that objective function decreases monotonically [3, 41].

5.2. Comparison on various models. We now compare the proposed L_1/L_2 approach with other sparse recovery models: L_1, L_p [9], L_1-L_2 [44, 23], and TL1 [46]. We choose $p = 0.5$ for L_p and $a = 1$ for TL1. The initial guess for all the algorithms is the solution of the L_1 model. Both



(a) Residual errors in $L1/L2$ (b) Residual errors in $L1/L2$-grad



(c) Objective functios of $L1/L2$ (d) Objective functions of $L1/L2$-grad

FIG. 2. *Plots of residual errors and objective functions for empirically demonstrating the convergence of the proposed algorithms.*

L_1-L_2 and TL1 are solved via the DCA, as described in Subsection 2.2, with the same stop criterion as L_1/L_2, *i.e.*, $\frac{\|\mathbf{x}^{(k)}-\mathbf{x}^{(k-1)}\|_2}{\|\mathbf{x}^{(k)}\|_2} \leq 10^{-8}$. As for L_p, we follow the default setting in [9].

We evaluate the performance of sparse recovery in terms of *success rate*, defined as the number of successful trials over the total number of trials. 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.*, $\frac{\|\mathbf{x}^*-\mathbf{x}\|_2}{\|\mathbf{x}\|_2} \leq 10^{-3}$. We further categorize the failure of not recovering the ground-truth as *model failure* and *algorithm failure*. In particular, we compare the objective function $F(\cdot)$ at the ground-truth $\mathbf{x}$ and at the reconstructed

solution $\mathbf{x}^*$. If $F(\mathbf{x}) > F(\mathbf{x}^*)$, it means that $\mathbf{x}$ is not a global minimizer of the model, in which case we call *model failure*. On the other hand, $F(\mathbf{x}) < F(\mathbf{x}^*)$ implies that the algorithm does not reach a global minimizer, which is referred to as *algorithm failure*. Although this type of analysis is not deterministic, it sheds some lights on which direction to improve: model or algorithm. For example, it was reported in [28] that L_1 has the highest model-failure rates, which justifies the need for nonconvex models.

In Figure 3, we examine two coherence levels: $F = 5$ corresponds to relatively low coherence and $F = 20$ for higher coherence. The success rates of various models reveal that L_1/L_2-box performs the best at $F = 5$ and is comparable to L_1-L_2 for the highly coherent case of $F = 20$. We look at Gaussian matrix with $r = 0.1$ and $r = 0.8$ in Figure 4, both of which exhibit very similar performance of various models. In particular, the L_p model gives the best results for the Gaussian case, which is consistent in the literature [43, 23]. The proposed model of L_1/L_2-box is the second best for such incoherent matrices.

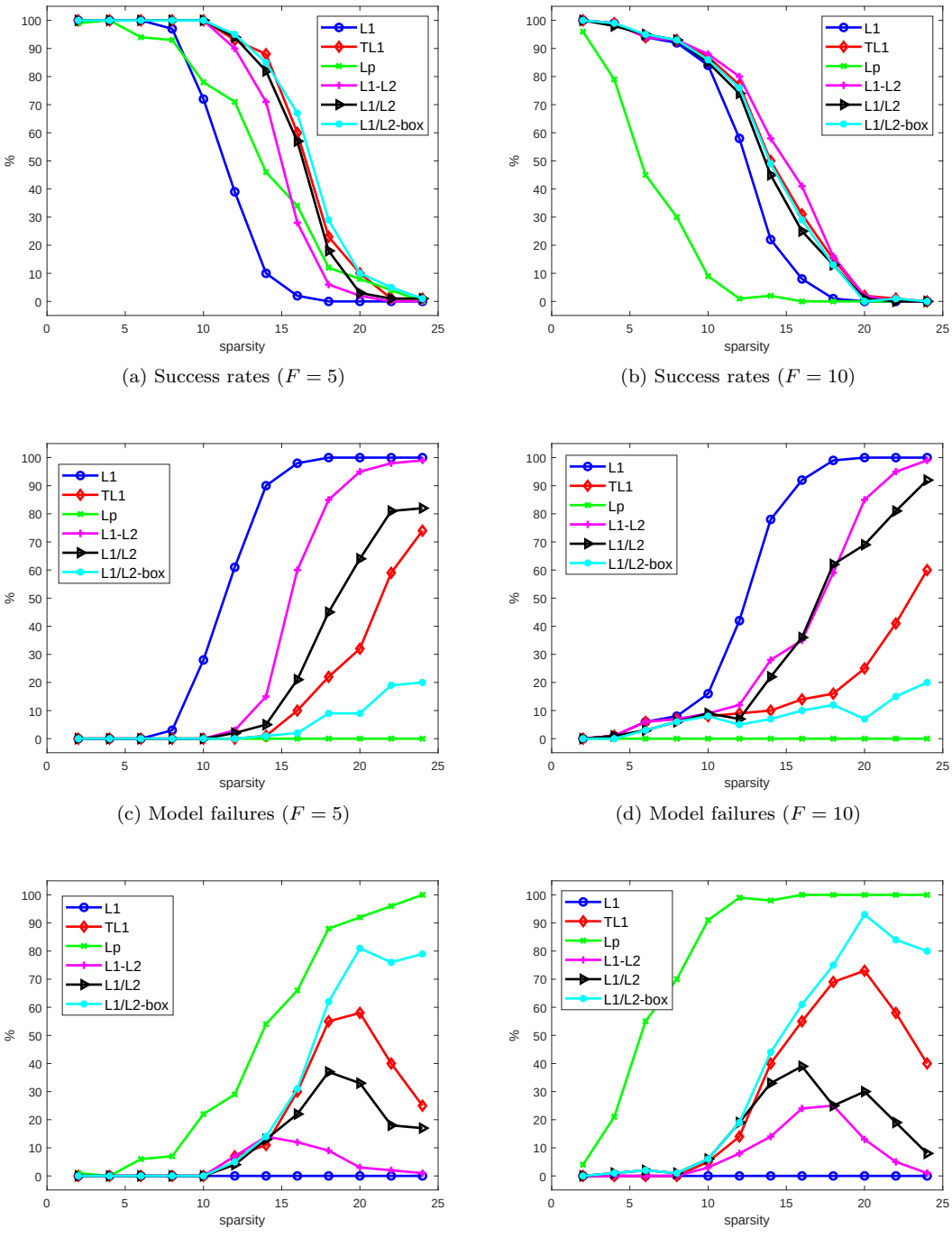
By comparing L_1/L_2 with and without box among the plots for success rates and model failures, we can draw the conclusion that the box constraint can mitigate the inherent drawback of the L_1/L_2 model, thus improving the recovery rates. In addition, L_1/L_2 is the second lowest in terms of model failure rates and simply adding a box constraint also increases the occurrence of algorithm failure compared to the none box version. These two observations suggest a need to further improve upon algorithms of minimizing L_1/L_2.

Finally, we provide the computation time for all the competing algorithms in Table 1 with the shortest time in each case highlighted in bold. The time for L_1 method is not included, as all the other methods use the L_1 solution as initial guess. It is shown that TL1 is the fastest for relatively lower sparsity levels and the proposed L_1/L_2-box is the most efficient at higher sparsity levels. The computational times for all these methods seem consistent with DCT and Gaussian matrices.

5.3. MRI reconstruction. As a proof-of-concept example, we study an MRI reconstruction problem [26] to compare the performance of L_1, L_1-L_2, and L_1/L_2 on the gradient. The L_1 on the gradient is the celebrated TV model [36], while L_1-L_2 on the gradient was recently proposed in [24]. We use a standard Shepp-Logan phantom as a testing image, as shown in Figure 5a. The data is obtained only along several radical lines in the frequency domain (after taking Fourier transform); an example of such sampling scheme using 7 lines is shown in Figure 5b. As this paper focuses on the constrained formulation, we do not consider noise, following the same setting as in the previous works [44, 24]. Since all the competing methods (L_1, L_1-L_2, and L_1/L_2) yield an exact recovery with 8 radical lines, with accuracy in the order of 10^{-8}, we present the reconstructions results of 7 radical lines in Figure 5, which illustrates that the ratio model (L_1/L_2) gives much better results than the difference model (L_1-L_2). Figure 5 also includes quantitative measures of the performance by relative error (RE) between the reconstructed and ground-truth images, which shows significantly improvement of the proposed L_1/L_2-grad over a classic method in MRI reconstruction, called filter-back projection (FBP), and two recent works of using L_1 and L_1-L_2 on the gradient. Note that the state-of-the-art methods in MRI reconstruction are [17, 28] that have reported exact recovery from 7 lines, but both approaches require to tune model parameters, while our model (4.15) is parameter free[1].

6. Empirical validations. A review article [7] indicated that two principles in CS are *sparsity* and *incoherence*, leading an impression that a sensing matrix with smaller coherence is easier for sparse recovery. However, we observe through numerical results [22] (also given in Figure 6b)

[1]There are three algorithmic parameters in (4.17), but they only affect the convergence speed of the algorithm.



(a) Success rates ($F = 5$)

(b) Success rates ($F = 10$)

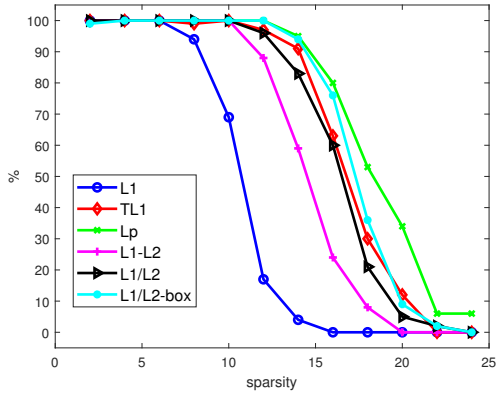
(c) Model failures ($F = 5$)

(d) Model failures ($F = 10$)

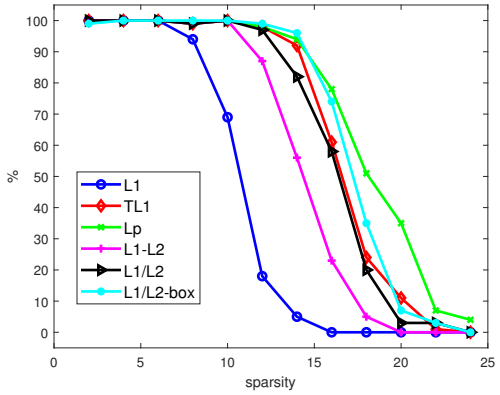
(e) Algorithm failures ($F = 5$)

(f) Algorithm failures ($F = 10$)

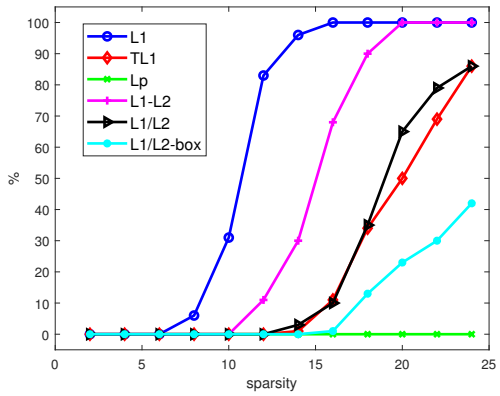
FIG. 3. *Success rates, model failures, algorithm failures for 6 algorithms in the case of oversampled DCT matrices.*



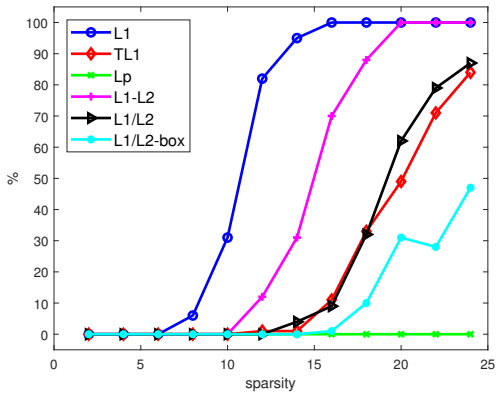
(a) Success rates ($r = 0.1$)



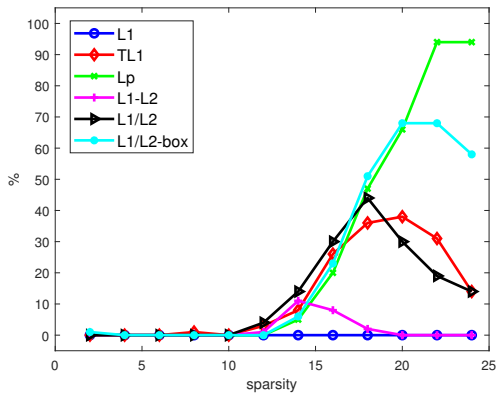
(b) Success rates ($r = 0.8$)



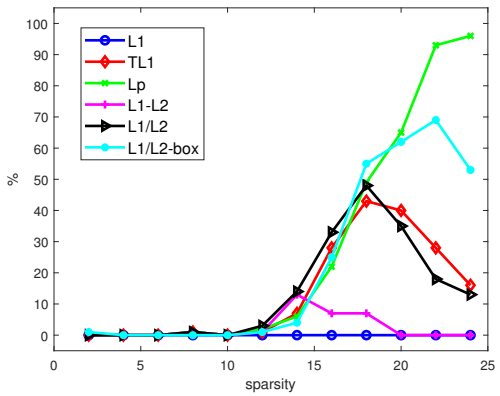
(c) Model failures ($r = 0.1$)



(d) Model failures ($r = 0.8$)



(e) Algorithm failures ($r = 0.1$)



(f) Algorithm failures ($r = 0.8$)

FIG. 4. *Success rates, model failures, algorithm failures for 6 algorithms in the Gaussian matrix case.*

TABLE 1
Computation time (sec.) in 5 algorithms.

(a) DCT matrix

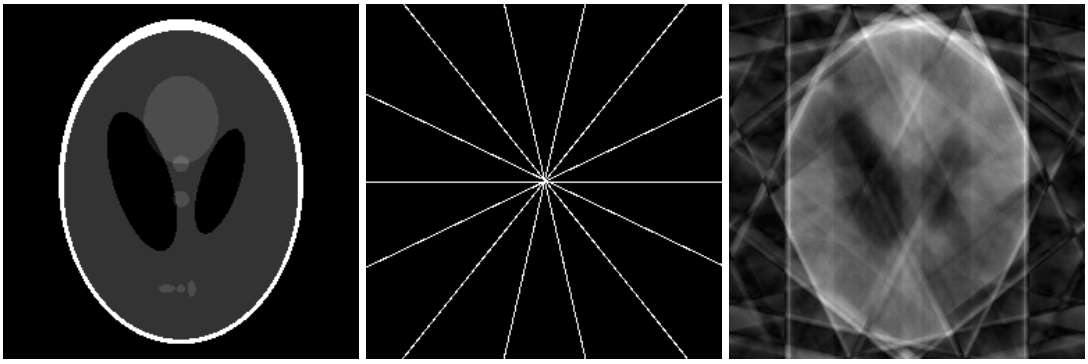
$F = 5$							
sparsity	2	6	10	14	18	22	mean
TL1	**0.049**	**0.050**	**0.066**	**0.207**	0.618	0.795	0.298
L_p	0.061	0.137	0.209	0.355	0.515	0.565	0.307
L_1-L_2	**0.049**	**0.050**	0.071	0.260	0.550	0.625	0.267
L_1/L_2	0.276	0.279	0.311	0.353	0.358	0.366	0.324
L_1/L_2-box	0.102	0.183	0.247	0.313	**0.325**	**0.332**	**0.250**
$F = 10$							
sparsity	2	6	10	14	18	22	mean
TL1	**0.048**	**0.069**	**0.092**	0.330	0.654	0.755	0.325
L_p	0.094	0.254	0.423	0.472	0.530	0.534	0.385
L_1-L_2	0.049	0.070	0.093	**0.272**	0.598	0.677	0.293
L_1/L_2	0.263	0.272	0.295	0.340	0.355	0.356	0.314
L_1/L_2-box	0.090	0.179	0.239	0.301	**0.324**	**0.322**	**0.243**

(b) Gaussian matrix

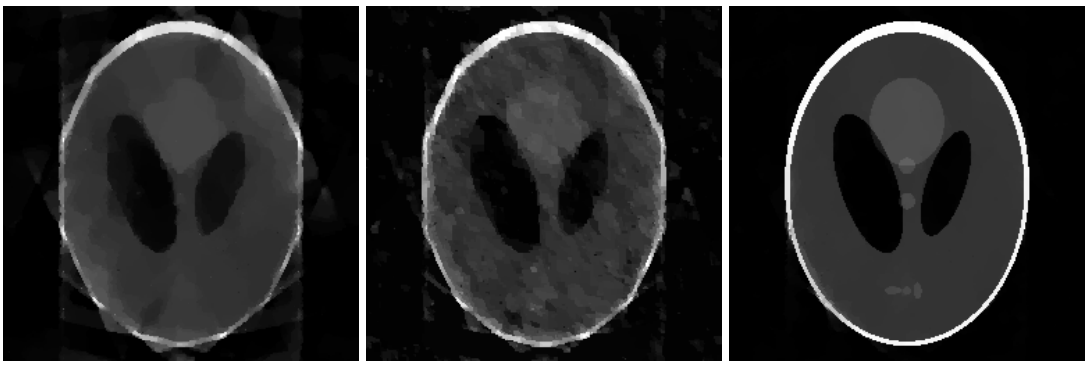
$r = 0.1$							
sparsity	2	6	10	14	18	22	mean
TL1	**0.070**	**0.069**	**0.117**	0.295	1.101	1.633	0.548
L_p	0.079	0.128	0.229	**0.261**	**0.742**	1.218	**0.443**
L_1-L_2	**0.070**	**0.069**	0.122	0.399	0.877	**1.161**	0.450
L_1/L_2	0.864	0.866	1.175	1.130	1.210	1.458	1.117
L_1/L_2-box	0.324	0.625	1.039	1.060	1.146	1.385	0.930
$r = 0.8$							
sparsity	2	6	10	14	18	22	mean
TL1	**0.050**	**0.053**	**0.071**	0.239	0.613	0.750	0.296
L_p	0.061	0.094	0.140	**0.207**	0.426	0.613	0.257
L_1-L_2	0.051	0.054	0.077	0.306	0.497	0.576	0.260
L_1/L_2	0.277	0.277	0.324	0.358	0.364	0.363	0.327
L_1/L_2-box	0.102	0.192	0.265	0.321	**0.332**	**0.327**	**0.256**

that a more coherent matrix gives higher recovery rates. This contradiction motivates us to collect empirical evidence regarding to either prove or refuse whether coherence is relevant to sparse recovery. Here we examine one such evidence by minimizing the ratio of L_1 and L_2, which gives an upper bound for a sufficient condition of L_1 exact recovery, see (1.3). To avoid the trivial solution of $\mathbf{x} = \mathbf{0}$ to the problem of $\min_{\mathbf{x}} \left\{ \frac{\|\mathbf{x}\|_1}{\|\mathbf{x}\|_2} : A\mathbf{x} = \mathbf{0} \right\}$, we incorporate a sum-to-one constraint. In other word, we define an expanded matrix $\tilde{A} = [A; \text{ones}(n, 1)]$ (following Matlab's notation) and an expanded vector $\tilde{\mathbf{b}} = [\mathbf{0}; 1]$. We then adapt the proposed method to solve for $\min_{\mathbf{x}} \left\{ \frac{\|\mathbf{x}\|_1}{\|\mathbf{x}\|_2} : \tilde{A}\mathbf{x} = \tilde{\mathbf{b}} \right\}$. In Figure 6a, we plot the mean value of ratios from 50 random realizations of matrices A at each coherence level (controlled by F), which shows that the ratio actually decreases[2] with respect to F. As the L_0 norm is bounded by the ratio (1.3), smaller ratio indicates it is more difficult to recover the signals. Therefore, Figure 6a is consistent with the common belief in CS.

[2]We also observe that the ratio stagnates for larger F, which is probably because of instability of the proposed method when matrix becomes more coherent.



(a) Original (b) Sampling mask (c) FBP (RE = 0.998)

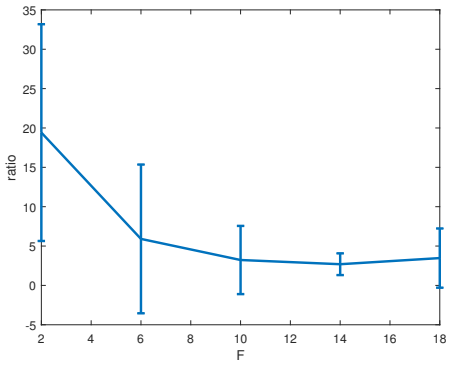


(d) L_1 (RE = 0.482) (e) L_1-L_2 (RE = 0.389) (f) L_1/L_2 (RE = 0.088)

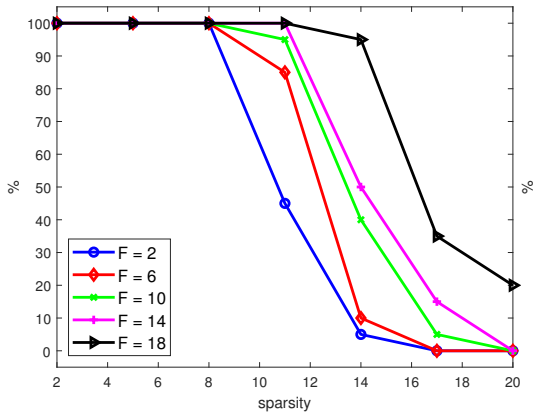
FIG. 5. *MRI reconstruction results from 7 radical lines in the frequency domain (3% measurements). The relative errors (RE) are provided for each method.*

We postulate that an underlying reason of more coherent matrices giving better results is minimum separation (MS), as formally introduced in [5]. In Figure 6b, we enforce the minimum separation of two neighboring spikes to be 40, following the suggestion of $2F$ in [15] (we consider F up to 20). In comparison, we also give the success rates of the L_1 recovery without any restrictions on MS in Figure 6c. Note that we use the exactly same matrices in both cases (with and without MS). Figure 6c does not have a clear pattern regarding how coherence affects the exact recovery, which supports our hypothesis that minimum separation plays an important role in sparse recovery. It will be our future work to analyze it throughly.

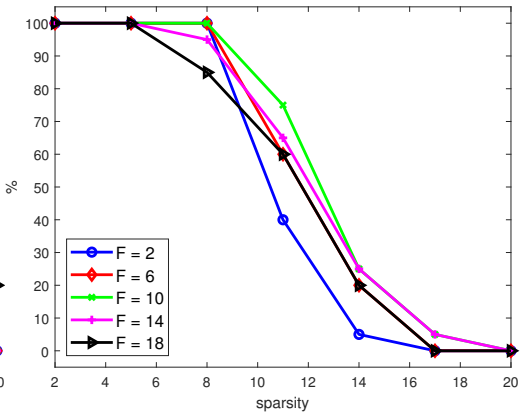
Furthermore, it follows from Proposition 1 that the minimal ratios subject to a linear system $A\mathbf{x} = \mathbf{b}$ is smaller than the one in the kernel space of A. Here, we empirically verify this proposition by calculating the ratio in $\min_{\mathbf{x}} \left\{ \frac{\|\mathbf{x}\|_1}{\|\mathbf{x}\|_2} : A\mathbf{x} = \mathbf{0} \right\}$ with 20 random realizations of the oversampled-DCT matrices A at $F = 2$ and $F = 5$; the ratios are plotted in blue dots in Figure 7. On the



(a) The mean of ratios



(b) Success rates with MS



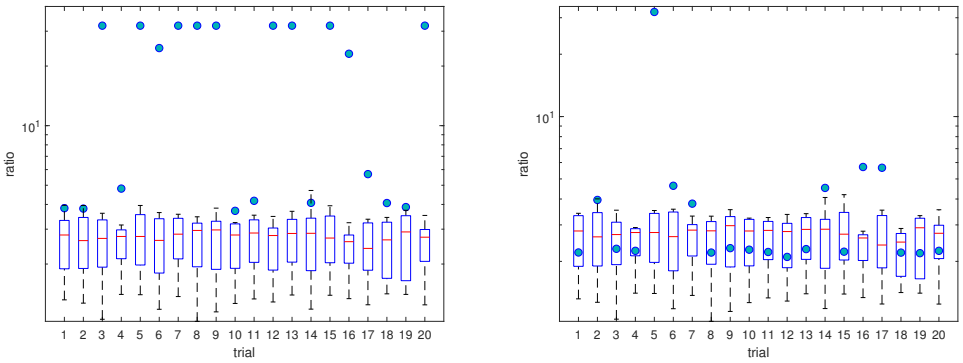
(c) Success rates without MS

FIG. 6. *The use of* $\min_{\mathbf{x}} \left\{ \frac{\|\mathbf{x}\|_1}{\|\mathbf{x}\|_2} : A\mathbf{x} = \mathbf{0} \right\}$ *as an upper bound for the* L_1 *recovery measured by success rates in the cases of with (b) and without (c) minimum separation.*

other hand, we also compute $\min_{\mathbf{x}} \left\{ \frac{\|\mathbf{x}\|_1}{\|\mathbf{x}\|_2} : A\mathbf{x} = \mathbf{b} \right\}$ for each fixed A with different sparsity levels: $s \in \{2, 4, 6, \cdots, 24\}$. We use a box plot[3] to visualize the results. Each box in Figure 7 corresponds to one specific A among 20 trials. We observe that most of blue dots are above the box when $F = 2$, which is consistent with Proposition 1. However, there exist many ratios subject to $A\mathbf{x} = \mathbf{b}$ that appear higher than the ones of $A\mathbf{x} = \mathbf{0}$ at $F = 5$. This is due to our algorithmic limitation that we cannot guarantee the minimizing algorithm actually finds a global minimizer.

7. Conclusions and future works. In this paper, we have studied a novel metric L_1/L_2 to promote sparsity. Two main benefits of L_1/L_2 are scale invariant and parameter free. Two numerical algorithms based on the ADMM are formulated for the assumptions of sparse signals and sparse gradients, together with a variant of incorporating additional box constraint. The

[3]The box plot in Matlab uses a red mark to indicate the median, bottom/top edges of the box to indicate the 25th and 75th percentiles, and the whiskers for the most extreme data points.



(a) $F = 2$ (b) $F = 5$

FIG. 7. *The ratios:* $\min_{\mathbf{x}} \left\{ \frac{\|\mathbf{x}\|_1}{\|\mathbf{x}\|_2} : A\mathbf{x} = \mathbf{b} \right\}$ *(box spot) and* $\min_{\mathbf{x}} \left\{ \frac{\|\mathbf{x}\|_1}{\|\mathbf{x}\|_2} : A\mathbf{x} = \mathbf{0} \right\}$ *(dot). Proposition 1 implies that blue dots should be above all the boxes.*

experimental results demonstrate the performance of the proposed approaches in comparison to the state-of-the-art methods in sparse recovery and MRI reconstruction. As a by-product, minimizing the ratio also gives an empirical upper bound towards L_1's exact recovery, which motivates further investigations on exact recovery theories. Other future works include algorithmic improvement and convergence analysis. In particular, it is shown in Table 1, Figures 3 and 4 that L_1/L_2 is not as fast as competing methods in CS and also has certain algorithmic failures, which calls for a more robust and more efficient algorithm. In addition, we have provided heuristic evidence of the ADMM's convergence and it will be interesting to analyze it theoretically.

Appendix: proof of Theorem 3.3. In order to prove Theorem 3.3, we need to introduce two lemmas:

LEMMA 7.1. *For any* $\mathbf{x}, \mathbf{v} \in \mathbb{R}^n$ *and* $i \in [n]$*, we have*

$$n\|\mathbf{x}\|_2^2 - |x_i|\|\mathbf{x}\|_1 \geq (n-1)\left(\sum_{j\neq i} x_j^2\right), \tag{7.1}$$

$$n\|\mathbf{v}\|_1\|\mathbf{x}\|_2^2 \geq \|\mathbf{x}\|_1 |\langle \mathbf{v}, \mathbf{x}\rangle|. \tag{7.2}$$

Furthermore, if $\|\mathbf{x}\|_0 = s$*, then the constant* n *in the inequalities can be reduced to* s*.*

Proof. Simple calculations show that

$$\begin{aligned} n\|\mathbf{x}\|_2^2 - |x_i|\|\mathbf{x}\|_1 &= n\left(\sum_j x_j^2\right) - |x_i|\left(\sum_j |x_j|\right) \\ &= (n-1)\left(\sum_{j\neq i} x_j^2\right) + \sum_{j\neq i} x_j^2 + (n-1)x_i^2 - \sum_{j\neq i} |x_i||x_j| \\ &= (n-1)\left(\sum_{j\neq i} x_j^2\right) + \sum_{j\neq i}\left((|x_i| - |x_j|)^2 + |x_i||x_j|\right) \\ &\geq (n-1)\left(\sum_{j\neq i} x_j^2\right) \geq 0. \end{aligned} \tag{7.3}$$

Therefore, we have

$$\sum_i \left(n\|\mathbf{x}\|_2^2 - |x_i|\|\mathbf{x}\|_1\right)|v_i| \geq 0, \tag{7.4}$$

which implies that

$$n\|\mathbf{v}\|_1\|\mathbf{x}\|_2^2 \geq \|\mathbf{x}\|_1\left(\sum_i |x_i||v_i|\right) \geq \|\mathbf{x}\|_1 |\langle \mathbf{v}, \mathbf{x}\rangle|. \tag{7.5}$$

∎

LEMMA 7.2. *Suppose that an s-sparse vector $\mathbf{x}$ satisfies $A\mathbf{x} = \mathbf{b}$ ($\mathbf{b} \neq 0$) with its support on an index set S and the matrix A satisfies the wNSP of order s. Define*

$$t_1 := \inf_{v,t}\left\{ \frac{|\sigma_t(\mathbf{v})\|\mathbf{x}\|_2^2 - \langle \mathbf{v}_S, \mathbf{x}\rangle \|\mathbf{x}\|_1|}{|\sigma_t(\mathbf{v})\langle \mathbf{v}_S, \mathbf{x}\rangle - \|\mathbf{x}\|_1\|\mathbf{v}\|_2^2|} \,\middle|\, \mathbf{v} \in \ker(A), \|\mathbf{v}\|_2 = 1, t \neq 0 \right\}, \tag{7.6}$$

where $\sigma_t(\mathbf{v}) = \left(\sum_{i\in S} v_i sign(x_i)\right) + sign(t)\|\mathbf{v}_{\bar{S}}\|_1$. We can show that $t_1 > 0$.

Proof. For any $\mathbf{v} \in \ker(A)$ and $\|\mathbf{v}\|_2 = 1$, it is straightforward that

$$\begin{aligned} |\sigma_t(\mathbf{v})\langle \mathbf{v}_S, \mathbf{x}\rangle - \|\mathbf{x}\|_1\|\mathbf{v}\|_2^2| &\leq |\sigma_t(\mathbf{v})|\|\mathbf{v}\|_2\|\mathbf{x}\|_2 + \|\mathbf{x}\|_1\|\mathbf{v}\|_2^2 \\ &\leq \|\mathbf{v}\|_1\|\mathbf{v}\|_2\|\mathbf{x}\|_2 + \|\mathbf{x}\|_1\|\mathbf{v}\|_2^2 \\ &= \|\mathbf{v}\|_1\|\mathbf{x}\|_2 + \|\mathbf{x}\|_1 \\ &\leq \sqrt{n}\|\mathbf{x}\|_2 + \|\mathbf{x}\|_1, \end{aligned} \tag{7.7}$$

and

$$|\sigma_t(\mathbf{v})| \geq |\text{sign}(t)\|\mathbf{v}_{\bar{S}}\|_1| - |\sum_{i\in S} v_i \text{sign}(x_i)| \geq \|\mathbf{v}_{\bar{S}}\|_1 - \sum_{i\in S}|v_i| = \|\mathbf{v}_{\bar{S}}\|_1 - \|\mathbf{v}_S\|_1. \tag{7.8}$$

It follows from the wNSP that $\|\mathbf{v}_{\bar{S}}\|_1 \geq (s+1)\|\mathbf{v}_S\|_1$, thus leading to the following two inequalities,

$$\begin{aligned} |\sigma_t(\mathbf{v})| &\geq \|\mathbf{v}_{\bar{S}}\|_1 - \|\mathbf{v}_S\|_1 \geq s\|\mathbf{v}_S\|_1 \\ |\sigma_t(\mathbf{v})| &\geq \|\mathbf{v}_{\bar{S}}\|_1 - \|\mathbf{v}_S\|_1 \geq (1 - \frac{1}{s+1})\|\mathbf{v}_{\bar{S}}\|_1 = \frac{s}{s+1}\|\mathbf{v}_{\bar{S}}\|_1. \end{aligned} \tag{7.9}$$

Next we will discuss two cases: $s > 1$ and $s = 1$.

(i) For $s > 1$. Since $\|\mathbf{v}\|_2 = 1$, there exists an index $j \in [n]$ such that $|v_j| \geq \frac{1}{\sqrt{n}}$. Suppose such $j \in S$, then we have that

$$\begin{aligned}
\left|\sigma_t(\mathbf{v})\|\mathbf{x}\|_2^2 - \langle \mathbf{v}_S, \mathbf{x}\rangle \|\mathbf{x}\|_1\right| &\geq |\sigma_t(\mathbf{v})|\|\mathbf{x}\|_2^2 - |\langle \mathbf{v}_S, \mathbf{x}\rangle|\|\mathbf{x}\|_1 \\
&\geq s\|\mathbf{v}_S\|_1\|\mathbf{x}\|_2^2 - \left(\sum_{i\in S}|x_i||v_i|\right)\|\mathbf{x}\|_1 \\
&= \sum_{i\in S}\left(s\|\mathbf{x}\|_2^2 - |x_i|\|\mathbf{x}\|_1\right)|v_i| \\
&\geq \left(s\|\mathbf{x}\|_2^2 - |x_j|\|\mathbf{x}\|_1\right)|v_j| \\
&\geq \frac{1}{\sqrt{n}}\left(s\|\mathbf{x}\|_2^2 - |x_j|\|\mathbf{x}\|_1\right) \\
&\geq \frac{1}{\sqrt{n}}\left((s-1)\sum_{i\neq j}x_i^2\right) \\
&\geq \frac{s-1}{\sqrt{n}}\min_{j\in S}\left(\sum_{i\neq j}x_i^2\right) > 0.
\end{aligned} \tag{7.10}$$

Otherwise, we have that $|v_j| < \frac{1}{\sqrt{n}} \forall j \in S$ and hence $\|\mathbf{v}_S\|_1 < \frac{s}{\sqrt{n}}$. Since $\|\mathbf{v}\|_1 \geq \|\mathbf{v}\|_2 = 1$, then $\|\mathbf{v}_{\bar{S}}\|_1 \geq 1 - \frac{s}{\sqrt{n}} = \frac{\sqrt{n}-s}{\sqrt{n}}$. As a result, we get

$$\begin{aligned}
\left|\sigma_t(\mathbf{v})\|\mathbf{x}\|_2^2 - \langle \mathbf{v}_S, \mathbf{x}\rangle \|\mathbf{x}\|_1\right| &\geq |\sigma_t(\mathbf{v})|\|\mathbf{x}\|_2^2 - |\langle \mathbf{v}_S, \mathbf{x}\rangle|\|\mathbf{x}\|_1 \\
&\geq \frac{s}{s+1}\|\mathbf{v}_{\bar{S}}\|_1\|\mathbf{x}\|_2^2 - \left(\sum_{i\in S}|x_i||v_i|\right)\|\mathbf{x}\|_1 \\
&\geq \frac{s}{s+1}\left(\frac{\sqrt{n}-s}{\sqrt{n}}\right)\|\mathbf{x}\|_2^2 - \left(\sum_{i\in S}|x_i|\frac{1}{\sqrt{n}}\right)\|\mathbf{x}\|_1 \\
&= \frac{(\sqrt{n}-s)s}{(s+1)\sqrt{n}}\|\mathbf{x}\|_2^2 - \frac{1}{\sqrt{n}}\|\mathbf{x}\|_1^2 \\
&\geq \frac{(\sqrt{n}-s)s}{(s+1)\sqrt{n}}\|\mathbf{x}\|_2^2 - \frac{1}{\sqrt{n}}s\|\mathbf{x}\|_2^2 \\
&= \frac{(\sqrt{n}+1)s}{(s+1)\sqrt{n}}\|\mathbf{x}\|_2^2 > 0.
\end{aligned}$$

Therefore, we have

$$t_1 = \frac{1}{2}\frac{\min\left\{\frac{s-1}{\sqrt{n}}\min_{j\in S}\left(\sum_{i\neq j}x_i^2\right), \left(\frac{\sqrt{n}+1}{s+1}\right)\frac{s}{\sqrt{n}}\|\mathbf{x}\|_2^2\right\}}{\sqrt{n}\|\mathbf{x}\|_2 + \|\mathbf{x}\|_1} > 0. \tag{7.11}$$

(ii) For $s = 1$. Suppose the only non-zero element is $x_i \neq 0$, and hence we have

$$|\sigma_t(\mathbf{v})\|\mathbf{x}\|_2^2 - \langle \mathbf{v}_S, \mathbf{x}\rangle \|\mathbf{x}\|_1| = |(v_i\text{sign}(x_i) + \text{sign}(t)\|\mathbf{v}_{\bar{S}}\|_1)x_i^2 - (v_i x_i)|x_i|| = \|\mathbf{v}_{\bar{S}}\|_1 x_i^2. \tag{7.12}$$

If $|v_i| \geq \frac{1}{\sqrt{n}}$ then $\|\mathbf{v}_{\bar{S}}\|_1 \geq (s+1)|v_i| \geq \frac{s+1}{\sqrt{n}}$ and hence

$$|\sigma_t(\mathbf{v})\|\mathbf{x}\|_2^2 - \langle \mathbf{v}_S, \mathbf{x}\rangle \|\mathbf{x}\|_1| = \|\mathbf{v}_{\bar{S}}\|_1\|\mathbf{x}\|_2^2 \geq \frac{s+1}{\sqrt{n}}\|\mathbf{x}\|_2^2. \tag{7.13}$$

If $|v_i| < \frac{1}{\sqrt{n}}$, then we have $\|\mathbf{v}_{\bar{S}}\|_1 \geq 1 - |v_i| = 1 - \frac{1}{\sqrt{n}} = \frac{\sqrt{n}-1}{\sqrt{n}}$ and

$$|\sigma_t(v)\|\mathbf{x}\|_2^2 - \langle \mathbf{v}_S, \mathbf{x}\rangle \|\mathbf{x}\|_1| = \|\mathbf{v}_{\bar{S}}\|_1\|\mathbf{x}\|_2^2 \geq \frac{\sqrt{n}-1}{\sqrt{n}}\|\mathbf{x}\|_2^2. \tag{7.14}$$

So

$$t_1 = \frac{1}{2}\frac{\min\left\{\frac{s+1}{\sqrt{n}}\|\mathbf{x}\|_2^2, \frac{\sqrt{n}-1}{\sqrt{n}}\|\mathbf{x}\|_2^2\right\}}{\sqrt{n}\|\mathbf{x}\|_2 + \|\mathbf{x}\|_1} > 0. \tag{7.15}$$

∎

We are ready to prove Theorem 3.3.

Proof. Denote S as the support of $\mathbf{x}$. We take a vector $\mathbf{v} \in \ker(A)\backslash\{0\}$ with $\|\mathbf{v}\|_2 = 1$, then $|v_i| \leq 1$, $\forall i$. Denote $t_0 = \frac{1}{2}\min_{i\in S}|x_i|$, we have that $|tv_i| < |x_i|$ for $|t| < t_0$, thus getting $\text{sign}(x_i + tv_i) = \text{sign}(x_i)$, $\forall i$. Therefore, we have

$$\begin{aligned}
\|\mathbf{x} + t\mathbf{v}\|_1 &= \sum_{i\in S}|x_i + tv_i| + \sum_{\imath\notin S}|t||v_i| \\
&= \sum_{i\in S}(x_i + tv_i)\text{sign}(x_i) + |t|\|\mathbf{v}_{\bar{S}}\|_1 \\
&= \sum_{i\in S}x_i\text{sign}(x_i) + t\sum_{i\in S}v_i\text{sign}(x_i) + |t|\|\mathbf{v}_{\bar{S}}\|_1 \\
&= \|\mathbf{x}\|_1 + t\left(\sum_{i\in S}v_i\text{sign}(x_i)\right) + |t|\|\mathbf{v}_{\bar{S}}\|_1 \\
&= \|\mathbf{x}\|_1 + t\left(\sum_{i\in S}v_i\text{sign}(x_i) + \text{sign}(t)\|v_{\bar{S}}\|_1\right) \\
&= \|\mathbf{x}\|_1 + t\sigma_t(\mathbf{v}),
\end{aligned}$$

where

$$\sigma_t(\mathbf{v}) = \sum_{i\in S}v_i\text{sign}(x_i) + \text{sign}(t)\|\mathbf{v}_{\bar{S}}\|_1. \tag{7.16}$$

Now we consider the following function

$$g(t) = \left(\frac{\|\mathbf{x} + t\mathbf{v}\|_1}{\|\mathbf{x} + t\mathbf{v}\|_2}\right)^2 = \frac{(\|\mathbf{x}\|_1 + t\sigma_t(\mathbf{v}))^2}{\|\mathbf{x}\|_2^2 + 2t\langle \mathbf{v}_S, \mathbf{x}\rangle + t^2\|\mathbf{v}\|_2^2}. \tag{7.17}$$

It is straightforward that g is a continuous function on $|t| < t_1$ as the denominator is non-zero. Notice that $\sigma_t(v)$ only depends on the sign of t, thus constant on $t \in (-t_1, 0)$ and on $(0, t_1)$.

Therefore, $g(t)$ is differentiable for $t \neq 0$. Simple calculations show that

(7.18)
$$\begin{aligned}\frac{\partial}{\partial t} g(t) &= \frac{\partial}{\partial t}\left(\frac{\left(\|\mathbf{x}\|_1 + t\sigma_t(\mathbf{v})\right)^2}{\|\mathbf{x}\|_2^2 + 2t\left\langle \mathbf{v}_S, \mathbf{x}\right\rangle + t^2\|\mathbf{v}\|_2^2}\right) \\ &= \frac{2\sigma_t(\mathbf{v})\left(\|\mathbf{x}\|_1 + t\sigma_t(\mathbf{v})\right)\left(\|\mathbf{x}\|_2^2 + 2t\left\langle \mathbf{v}_S, \mathbf{x}\right\rangle + t^2\|\mathbf{v}\|_2^2\right) - \left(2\left\langle \mathbf{v}_S, \mathbf{x}\right\rangle + 2t\|\mathbf{v}\|_2^2\right)\left(\|\mathbf{x}\|_1 + t\sigma_t(\mathbf{v})\right)^2}{\left(\|\mathbf{x}\|_2^2 + 2t\left\langle \mathbf{v}_S, \mathbf{x}\right\rangle + t^2\|\mathbf{v}\|_2^2\right)^2} \\ &= \frac{2\left(\|\mathbf{x}\|_1 + t\sigma_t(\mathbf{v})\right)\left[\sigma_t(\mathbf{v})\left(\|\mathbf{x}\|_2^2 + 2t\left\langle \mathbf{v}_S, \mathbf{x}\right\rangle + t^2\|\mathbf{v}\|_2^2\right) - \left(\left\langle \mathbf{v}_S, \mathbf{x}\right\rangle + t\|\mathbf{v}\|_2^2\right)\left(\|\mathbf{x}\|_1 + t\sigma_t(\mathbf{v})\right)\right]}{\left(\|\mathbf{x}\|_2^2 + 2t\left\langle \mathbf{v}_S, \mathbf{x}\right\rangle + t^2\|\mathbf{v}\|_2^2\right)^2} \\ &= \frac{2\left(\|\mathbf{x}\|_1 + t\sigma_t(\mathbf{v})\right)\left[\left(\sigma_t(\mathbf{v})\|\mathbf{x}\|_2^2 - \left\langle \mathbf{v}_S, \mathbf{x}\right\rangle\|\mathbf{x}\|_1\right) + \left(\sigma_t(\mathbf{v})\left\langle \mathbf{v}_S, \mathbf{x}\right\rangle - \|\mathbf{x}\|_1\|\mathbf{v}\|_2^2\right)t\right]}{\left(\|\mathbf{x}\|_2^2 + 2t\left\langle \mathbf{v}_S, \mathbf{x}\right\rangle + t^2\|\mathbf{v}\|_2^2\right)^2}.\end{aligned}$$

It follows from (7.16) that $\|\mathbf{x}\|_1 + t\sigma_t(\mathbf{v}) = \|\mathbf{x} + t\mathbf{v}\| > 0$ for $|t| < t_0$. Also there exists a positive number t_1 defined in lemma 2 and hence we have for $|t| < t_1$

$$\begin{aligned}&\operatorname{sign}\left[\left(\sigma_t(\mathbf{v})\|\mathbf{x}\|_2^2 - \left\langle \mathbf{v}_S, \mathbf{x}\right\rangle\|\mathbf{x}\|_1 + \left(\sigma_t(\mathbf{v})\left\langle \mathbf{v}_S, \mathbf{x}\right\rangle - \|\mathbf{x}\|_1\|\mathbf{v}\|_2^2\right)t\right]\right. \\ &= \operatorname{sign}\left(\sigma_t(\mathbf{v})\|\mathbf{x}\|_2^2 - \left\langle \mathbf{v}_S, \mathbf{x}\right\rangle\|\mathbf{x}\|_1\right).\end{aligned}$$

Letting $t^* = \min\{t_0, t_1\}$, we have for any $t \in (0, t^*)$ that

(7.19)
$$\begin{aligned}\sigma_t(\mathbf{v}) &= \left(\sum_{i\in S} v_i \operatorname{sign}(x_i)\right) + \operatorname{sign}(t)\|\mathbf{v}_{\bar{S}}\|_1 \\ &= \left(\sum_{i\in S} v_i \operatorname{sign}(x_i)\right) + \|\mathbf{v}_{\bar{S}}\|_1 \\ &\geq \|\mathbf{v}_{\bar{S}}\|_1 - \|\mathbf{v}_S\|_1 \\ &\geq \max\left\{s\|\mathbf{v}_S\|_1, \frac{s}{s+1}\|\mathbf{v}_{\bar{S}}\|_1\right\} > 0.\end{aligned}$$

Therefore, we have

(7.20)
$$\sigma_t(\mathbf{v})\|\mathbf{x}\|_2^2 = |\sigma_t(\mathbf{v})|\|\mathbf{x}\|_2^2 \geq |\left\langle \mathbf{v}_S, \mathbf{x}\right\rangle|\|\mathbf{x}\|_1 \geq \left\langle \mathbf{v}_S, \mathbf{x}\right\rangle\|\mathbf{x}\|_1,$$

where the first inequality is followed by Lemma 7.1. This implies that

(7.21)
$$\sigma_t(\mathbf{v})\|\mathbf{x}\|_2^2 - \left\langle \mathbf{v}_S, \mathbf{x}\right\rangle\|\mathbf{x}\|_1 \geq 0.$$

As a result, we have $g'(t) \geq 0$ if $0 < t < t^*$. The function $g(t)$ is not differentiable at zero, but we can compute the sub-derivative as follows,

(7.22)
$$g'(0^+) = \lim_{t\longrightarrow 0^+}\frac{g(t) - g(0)}{t - 0} = \frac{2\|\mathbf{x}\|_1\left(\sigma_{+1}(\mathbf{v})\|\mathbf{x}\|_2^2 - \left\langle \mathbf{v}_S, \mathbf{x}\right\rangle\|\mathbf{x}\|_1\right)}{\|\mathbf{x}\|_2^4} \geq 0.$$

Similarly, we can get $g'(t) \leq 0$ if $-t^* < t < 0$ and $g'(0^-) \leq 0$. Therefore for any $|t| < t^*$ we have $g(0) \leq g(t)$, which implies that

(7.23)
$$\frac{\|\mathbf{x} + t\mathbf{v}\|_1}{\|\mathbf{x} + t\mathbf{v}\|_2} \geq \frac{\|\mathbf{x}\|_1}{\|\mathbf{x}\|_2}, \qquad \forall |t| < t^*.$$

∎

REFERENCES

[1] A. S. Bandeira, E. Dobriban, D. G. Mixon, and W. F. Sawin, *Certifying the restricted isometry property is hard*, IEEE Trans. Inf. Theory, 59 (2013), pp. 3448–3450.

[2] A. Beck, *First-Order Methods in Optimization*, vol. 25, SIAM, 2017.

[3] J. Bolte, S. Sabach, and M. Teboulle, *Proximal alternating linearized minimization for nonconvex and nonsmooth problems*, Math. Program., 146 (2014), pp. 459–494.

[4] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, *Distributed optimization and statistical learning via the alternating direction method of multipliers*, Found. Trends Mach. Learn., 3 (2011), pp. 1–122.

[5] E. J. Candès and C. Fernandez-Granda, *Towards a mathematical theory of super-resolution*, Comm. Pure Appl. Math., 67 (2014), pp. 906–956.

[6] E. J. Candès, J. Romberg, and T. Tao, *Stable signal recovery from incomplete and inaccurate measurements*, Comm. Pure Appl. Math., 59 (2006), pp. 1207–1223.

[7] E. J. Candès and M. B. Wakin, *An introduction to compressive sampling*, IEEE Signal Process. Mag., 25 (2008), pp. 21–30.

[8] A. Chambolle and T. Pock, *A first-order primal-dual algorithm for convex problems with applications to imaging*, J. Math. Imaging and Vision, 40 (2011), pp. 120–145.

[9] R. Chartrand, *Exact reconstruction of sparse signals via nonconvex minimization*, IEEE Signal Process. Lett., 10 (2007), pp. 707–710.

[10] A. Cohen, W. Dahmen, and R. DeVore, *Compressed sensing and the best k-term approximation*, J. Am. Math. Soc., 22 (2009), pp. 211–231.

[11] Y. Cui, J. S. Pang, and B. Sen, *Composite difference-max programs for modern statistical estimation problems*, arXiv preprint arXiv:1803.00205, (2018).

[12] D. Donoho and M. Elad, *Optimally sparse representation in general (nonorthogonl) dictionaries via l1 minimization*, Proc. Nat. Acad. Scien. USA, 100 (2003), pp. 2197–2202.

[13] D. L. Donoho and X. Huo, *Uncertainty principles and ideal atomic decomposition*, IEEE Trans. Inf. Theory, 47 (2001), pp. 2845–2862.

[14] E. Esser, Y. Lou, and J. Xin, *A method for finding structured sparse solutions to non-negative least squares problems with applications*, SIAM J. Imaging Sci., 6 (2013), pp. 2010–2046.

[15] A. Fannjiang and W. Liao, *Coherence pattern–guided compressive sensing with unresolved grids*, SIAM J. Imaging Sci., 5 (2012), pp. 179–202.

[16] R. Gribonval and M. Nielsen, *Sparse representations in unions of bases*, IEEE Trans. Inf. Theory, 49 (2003), pp. 3320–3325.

[17] W. Guo and W. Yin, *Edge guided reconstruction for compressive imaging*, SIAM J. Sci. Imaging, 5 (2012), pp. 809–834.

[18] P. O. Hoyer, *Non-negative sparse coding*, in Proc. IEEE Workshop on Neural Networks for Signal Processing, 2002, pp. 557–565.

[19] N. Hurley and S. Rickard, *Comparing measures of sparsity*, IEEE Trans. on Inform. Theory, 55 (2009), pp. 4723–4741.

[20] D. Krishnan, T. Tay, and R. Fergus, *Blind deconvolution using a normalized sparsity measure*, in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2011, pp. 233–240.

[21] M. J. Lai, Y. Xu, and W. Yin, *Improved iteratively reweighted least squares for unconstrained smoothed lq minimization*, SIAM J. Numer. Anal., 5 (2013), pp. 927–957.

[22] Y. Lou, S. Osher, and J. Xin, *Computational aspects of l1-l2 minimization for compressive sensing*, in Model. Comput. & Optim. in Inf. Syst. & Manage. Sci., Advances in Intelligent Systems and Computing, vol. 359, 2015, pp. 169–180.

[23] Y. Lou, P. Yin, Q. He, and J. Xin, *Computing sparse representation in a highly coherent dictionary based on difference of l1 and l2*, J. Sci. Comput., 64 (2015), pp. 178–196.

[24] Y. Lou, T. Zeng, S. Osher, and J. Xin, *A weighted difference of anisotropic and isotropic total variation model for image processing*, SIAM J. Imaging Sci., 8 (2015), pp. 1798–1823.

[25] Z. Lu, Z. Zhou, and Z. Sun, *Enhanced proximal DC algorithms with extrapolation for a class of structured nonsmooth DC minimization*, Math. Program. (to appear), (2018).

[26] M. Lustig, D. L. Donoho, and J. M. Pauly, *Sparse MRI: The application of compressed sensing for rapid MR imaging*, Magnet. Reson. Med., 58 (2007), pp. 1182–1195.

[27] J. Lv and Y. Fan, *A unified approach to model selection and sparse recovery using regularized least squares*, Ann. Appl. Stat., (2009), pp. 3498–3528.

[28] T. Ma, Y. Lou, and T. Huang, *Truncated l1-l2 models for sparse recovery and rank minimization*, SIAM J. Imaging Sci., 10 (2017), pp. 1346–1380.

[29] B. K. Natarajan, *Sparse approximate solutions to linear systems*, SIAM J. Comput., (1995), pp. 227–234.
[30] G. Optimization, *Gurobi optimizer reference manual*, 2015.
[31] J. S. Pang, M. Razaviyayn, and A. Alvarado, *Computing B-stationary points of nonsmooth dc programs*, Math. Oper. Res., 42 (2016), pp. 95–118.
[32] T. Pham-Dinh and H. A. Le-Thi, *A D.C. optimization algorithm for solving the trust-region subproblem*, SIAM J. Optim., 8 (1998), pp. 476–505.
[33] T. Pham-Dinh and H. A. Le-Thi, *The DC (difference of convex functions) programming and DCA revisited with DC models of real world nonconvex optimization problems*, Ann. Oper. Res., 133 (2005), pp. 23–46.
[34] H. Raguet, J. Fadili, and G. Peyré, *A generalized forward-backward splitting*, SIAM J. Imaging Sci., 6 (2013), pp. 1199–1226.
[35] A. Repetti, M. Q. Pham, L. Duval, E. Chouzenouxe, and J.-C. Pesquet, *Euclid in a taxicab: Sparse blind deconvolution with smoothed ℓ_1/ℓ_2 regularization*, IEEE Signal Process. Lett., 22 (2015), pp. 539–543.
[36] L. Rudin, S. Osher, and E. Fatemi, *Nonlinear total variation based noise removal algorithms*, Physica D, 60 (1992), pp. 259–268.
[37] X. Shen, W. Pan, and Y. Zhu, *Likelihood-based selection and sharp parameter estimation*, J. Am. Stat. Assoc., 107 (2012), pp. 223–232.
[38] M. Tao and H. Dong, *On the linear convergence of difference-of-convex algorithms for nonsmooth dc programming*. http://www.optimization-online.org/DB_HTML/2018/08/6766.html, 2018 (submitted).
[39] A. M. Tillmann and M. E. Pfetsch, *The computational complexity of the restricted isometry property, the nullspace property, and related concepts in compressed sensing*, IEEE Trans. Inf. Theory, 60 (2014), pp. 1248–1259.
[40] H. Tran and C. Webster, *Unified sufficient conditions for uniform recovery of sparse signals via nonconvex minimizations*, arXiv preprint arXiv:1710.07348, (2017).
[41] Y. Wang, W. Yin, and J. Zeng, *Global convergence of ADMM in nonconvex nonsmooth optimization*, J. Sci. Comput., (2015), pp. 1–35.
[42] Z. Xu, X. Chang, F. Xu, and H. Zhang, *$l_{1/2}$ regularization: A thresholding representation theory and a fast solver*, IEEE Trans. Neural Netw., 23 (2012), pp. 1013–1027.
[43] P. Yin, E. Esser, and J. Xin, *Ratio and difference of l1 and l2 norms and sparse representation with coherent dictionaries*, Comm. Info. Systems, 14 (2014), pp. 87–109.
[44] P. Yin, Y. Lou, Q. He, and J. Xin, *Minimization of ℓ_{1-2} for compressed sensing*, SIAM J. Sci. Comput., 37 (2015), pp. A536–A563.
[45] S. Zhang and J. Xin, *Minimization of transformed l_1 penalty: Closed form representation and iterative thresholding algorithms*, Comm. Math. Sci., 15 (2017), pp. 511–537.
[46] S. Zhang and J. Xin, *Minimization of transformed l_1 penalty: Theory, difference of convex function algorithm, and robust application in compressed sensing*, Math. Program., Ser. B, 169 (2018), pp. 307–336.
[47] T. Zhang, *Multi-stage convex relaxation for learning with sparse regularization*, in Adv. Neural. Inf. Process. Syst., 2009, pp. 1929–1936.
[48] Y. Zhang, *Theory of compressive sensing via L1-minimization: a non-RIP analysis and extensions*, J. Oper. Res. Soc. China, 1 (2013), pp. 79–105.