Minimization of Transformed $L_1$ Penalty: Theory, Difference of Convex Function Algorithm, and Robust Application in Compressed Sensing

Shuai Zhang, and Jack Xin *

November 26, 2014

Abstract

We study the minimization problem of a non-convex sparsity promoting penalty function, the transformed $l_1$ (TL1), and its application in compressed sensing (CS). The TL1 penalty interpolates $l_0$ and $l_1$ norms through a nonnegative parameter $a \in (0, +\infty)$, similar to $l_p$ with $p \in (0, 1]$. TL1 is known in the statistics literature to enjoy three desired properties: unbiasedness, sparsity and Lipschitz continuity. We first consider the constrained minimization problem and prove the uniqueness of global minimizer and its equivalence to $l_0$ norm minimization if the sensing matrix $A$ satisfies a restricted isometry property (RIP) and if $a > a^*$, where $a^*$ depends only on $A$. This result contains the well-known equivalence of $l_1$ norm and $l_0$ norm, in the limit $a \to +\infty$. The solution is stable under noisy measurement. For general sensing matrix $A$, we show that the support set of a local minimizer corresponds to linearly independent columns of $A$, and recall sufficient conditions for a critical point to be a local minimum. Next, we present difference of convex algorithms for TL1 (DCATL1) in computing TL1-regularized constrained and unconstrained problems in CS. The DCATL1 algorithm involves outer and inner loops of iterations, one time matrix inversion, repeated shrinkage operations and matrix-vector multiplications. For the unconstrained problem, we prove convergence of DCATL1 to a stationary point satisfying the first order optimality condition. Finally in numerical experiments, we identify the optimal value $a = 1$, and compare DCATL1 with other CS algorithms on three classes of sensing matrices: Gaussian random matrices, over-sampled discrete cosine transform matrices (ODCT), and uniformly distributed M-sphere matrices (whose columns are uniformly distributed unit vectors on the M-dimensional sphere). Among existing algorithms, the iterated reweighted least squares method based on $L_{1/2}$ norm is the best in sparse recovery for Gaussian matrices, and the DCA algorithm

* S. Zhang and J. Xin were partially supported by NSF grants DMS-0928427, and DMS-1222507. They are with the Department of Mathematics, University of California, Irvine, CA, 92697, USA. E-mail: szhang3@uci.edu; jxin@math.uci.edu. Phone: (949)-824-5309. Fax: (949)-824-7993.
based on $L_1 - L_2$ penalty is the best for ODCT matrices (the most coherent among the three classes). We find that for all three classes of sensing matrices, the performance of DCATL1 algorithm (initiated with $L_1$ minimization) always ranks near the top (if not the top), and is the most robust choice insensitive to RIP (incoherence) of the underlying CS problems.

**Index terms**— Transformed $l_1$ penalty, sparse signal recovery theory, difference of convex function algorithm, convergence analysis, coherent random matrices, compressed sensing, robust recovery.

1 Introduction

Compressed sensing \[4, 8\] has generated enormous interest and research activities in mathematics, statistics, signal processing, imaging and information sciences, among numerous other areas. One of the basic problems is to reconstruct a sparse signal under a few linear measurements (linear constraints) far less than the dimension of the ambient space of the signal. Consider a sparse signal $\beta \in \mathbb{R}^N$, an $M \times N$ sensing matrix $A$ and an observation $y \in \mathbb{R}^M$, $M \ll N$, such that:

$$y = A\beta + \epsilon,$$

where $\epsilon$ is an $N$-dimensional observation error. If $\beta$ is sparse enough, it can be reconstructed exactly in the noise-free case and in stable manner in the noisy case provided that the sensing matrix $A$ satisfies certain incoherence or the restricted isometry property (RIP) \[4, 8\].

The direct approach is $l_0$ optimization. The constrained formulation is:

$$\min_{\beta \in \mathbb{R}^N} ||\beta||_0, \text{ s.t. } y = A\beta,$$  \hspace{1cm} (1.1)

and the unconstrained $l_0$ regularized version is:

$$\min_{\beta \in \mathbb{R}^N} \{||y - A\beta||_2^2 + \lambda||\beta||_0\}$$ \hspace{1cm} (1.2)

for some positive parameter $\lambda$. Since minimizing $L_0$ norm is NP-hard \[20\], many viable alternatives are available. Greedy methods (matching pursuit \[19\], orthogonal matching pursuits (OMP) \[27\], and regularized OMP (ROMP) \[21\]) work well if the dimension $N$ is not too large. For the unconstrained problem \[1\], the penalty decomposition method \[17\] replaces the term $\lambda ||\beta||_0$ by $\rho_k ||\beta - y||_2^2 + \lambda ||y||_0$, and minimizes over $(\beta, y)$ for a diverging sequence $\rho_k$. The variable $y$ allows the iterative hard thresholding procedure.

The relaxation methods replace $l_0$ norm by a continuous sparsity promoting penalty functions $p(\cdot)$. The minimization takes the form:

$$\min \ P(\beta), \text{ s.t. } y = A\beta.$$ \hspace{1cm} (1.3)
for the constrained problem and

$$\min_{\beta \in \mathbb{R}^N} \left\{ \|y - A\beta\|_2^2 + \lambda P(\beta) \right\}$$  \hspace{1cm} (1.4)$$

for the unconstrained problem. Convex relaxation uniquely selects $P(\cdot)$ as the $l_1$ norm. The resulting problems are known as basis pursuit (LASSO in the over-determined regime \[26\]). The $l_1$ algorithms include $l_1$-magic \[4\], Bregman and split Bregman methods \[34, 14\] and yall1 \[32\]. Theoretically, Candès and Tao introduced RIP condition and used it to establish the equivalent and unique global solution to $l_0$ minimization via $l_1$ relaxation among other stable recovery results \[2, 4, 1\].

There are many choices of $P$ for non-convex relaxation. One is the $l_p$ norm ($p \in (0, 1)$) with $l_0$ equivalence under RIP \[6\]. The $l_{1/2}$ norm is representative of this class of functions, with the reweighted least squares and half-thresholding algorithms for computation \[12, 30, 29\]. Near the RIP regime, $l_{1/2}$ penalty tends to have higher success rate of sparse reconstruction than $l_1$, however, it is not as good as $l_1$ if the sensing matrix is far away from RIP \[15, 33\] as we shall see later as well. In the highly non-RIP (coherent) regime, it is recently found that the difference of $L_1$ and $L_2$ norm minimization gives the best sparse recovery results \[33, 15\]. It is therefore of both theoretical and practical interest to find a non-convex penalty that is consistently better than $l_1$ and always ranks among the top in sparse recovery whether the sensing matrix is RIP or non-RIP (as an all-around champion).

In the statistics literature of variable selection, Fan and Li \[11\] advocated for classes of penalty functions with three desired properties: unbiasedness, sparsity and continuity. To help identify such a penalty function denoted by $\rho$, Fan and Lv \[18\] proposed the following condition for characterizing unbiasedness and sparsity promoting properties.

**Condition 1.** The penalty function $\rho(\cdot)$ satisfies:

1. $\rho(t)$ is increasing and concave in $t \in [0, \infty)$;
2. $\rho'(t)$ is continuous with $\rho'(0+) \in (0, \infty)$;
3. if $\rho(t)$ depends on a positive parameter $\lambda$, then $\rho'(t; \lambda)$ is increasing in $\lambda \in (0, \infty)$ and $\rho'(0+)$ is independent of $\lambda$.

It follows that $\rho'(t)$ is positive and decreasing, and $\rho'(0+)$ is the upper bound of $\rho'(t)$. It is shown in \[11\] that penalties satisfying Condition 1 and $\lim_{t \to \infty} \rho'(t) = 0$ enjoy both unbiasedness and sparsity. Though continuity does not generally hold for this class of penalty functions, a special one parameter family of functions, the so called transformed $l_1$ functions (TL1) $\rho_a(|t|)$, where $\rho_a(t) = \frac{(a + 1)t}{a + t}$, $a \in (0, +\infty)$, satisfies all three desired properties \[11\]. We shall study the minimization of TL1 functions for CS problems, in terms of theory, algorithms and computation. We shall show via numerical results that
a difference of convex algorithm of TL1 (DCATL1) is the all around champion among the existing representative CS algorithms based on three classes of coherent random sensing matrices. Same as $L_{1/2}$ regularization \[30\,31\], there also exists thresholding algorithm for TL1, which is being studied in the companion paper \[16\].

The rest of the paper is organized as follows. In section 2, we study the elementary inequalities of TL1. In section 3, a RIP condition is given for finding the unique global minimizer of the constrained TL1 model, which is also proven to be stable under noisy measurement. The local minimizers of both the constrained and unconstrained models share the property that they extract independent columns from the sensing matrix $A$. In section 4, we present two DC algorithms for TL1 optimization (DCATL1), and establish the relevant convergence theory. In section 5, we compare the performance of DCATL1 with $l_1$, reweighted least squares, $l_1-l_2$ algorithms in constrained and unconstrained test problems for three classes of matrices: the Gaussian, the oversampled discrete cosine transform (DCT), and the uniformly distributed M-sphere matrices of varying degrees of incoherence. Numerical experiments indicate that our algorithm — DCATL1 is most robust and consistently top ranked while maintaining high sparse recovery rates across all sensing matrices. Concluding remarks are in section 6.

2 Transformed $l_1$ (TL1) and Preliminaries

The transformed $l_1$ functions (TL1) are $\rho_a(|t|)$ \[18\] with:

$$\rho_a(t) = \frac{(a+1)t}{a+t}, \quad t \geq 0,$$

(2.1)

where the parameter $a \in (0, +\infty)$. It interpolates the $l_0$ and $l_1$ norms as

$$\lim_{a \to 0^+} \rho_a(|t|) = \chi_{\{t \neq 0\}}$$

and

$$\lim_{a \to \infty} \rho_a(|t|) = |t|.$$

In Fig. 1, we draw level lines of $l_1$ and TL1. With the adjustment of parameter $a$, the TL1 can approximate both $l_1$ and $l_0$ well. The TL1 function is Lipschitz continuous and satisfies Condition 1, thus enjoying the unbiasedness, sparsity and continuity properties \[18\].

Let us define:

$$P_a(x) = \sum_{i=1,\ldots,N} \rho_a(|x_i|),$$

(2.2)

and focus on the constrained TL1 minimization model:

$$\min_{x \in \mathbb{R}^N} f(x) = \min_{x \in \mathbb{R}^N} P_a(x) \quad s.t. \quad Ax = y,$$

(2.3)
Figure 1: Level lines of two sparsity promoting measures. Compared with $l_1$, the level line of TL1 is closer to the axes or those of $l_0$ when parameter $a$ is small ($a = 0.01$). When $a$ becomes larger ($a = 100$), the level lines converge to those of $l_1$. 
and the unconstrained TL1-regularized model:

$$\min_{x \in \mathbb{R}^N} f(x) = \min_{x \in \mathbb{R}^N} \frac{1}{2} \|Ax - y\|_2^2 + \lambda P_a(x).$$

(2.4)

First, we prove elementary inequalities of $\rho_a$ for later use.

**Lemma 2.1.** For $a \geq 0$, any $x_i$ and $x_j$ in $\mathbb{R}$, the following inequalities hold:

$$\rho_a(|x_i + x_j|) \leq \rho_a(|x_i| + |x_j|) \leq \rho_a(|x_i|) + \rho_a(|x_j|) \leq 2\rho_a\left(\frac{|x_i + x_j|}{2}\right)$$

(2.5)

**Proof.** Let us prove these inequalities one by one, starting from the left.

1.) According to Condition 1, we know that $\rho_a(|t|)$ is increasing in the variable $|t|$. By triangle inequality $|x_i + x_j| \leq |x_i| + |x_j|$, we have:

$$\rho_a(|x_i + x_j|) \leq \rho_a(|x_i| + |x_j|)$$

2.) Since $\rho_a(|t|) = \frac{(a + 1)|t|}{a + |t|}$,

$$\rho_a(|x_i|) + \rho_a(|x_j|) = \frac{(a + 1)|x_i|}{a + |x_i|} + \frac{(a + 1)|x_j|}{a + |x_j|}
= \frac{a(a + 1)(|x_i| + |x_j| + 2|x_i x_j|/a)}{a(|x_i| + |x_j| + |x_i x_j|/a)}
\geq \frac{(a + 1)(|x_i| + |x_j| + |x_i x_j|/a)}{(a + |x_i| + |x_j| + |x_i x_j|/a)}
= \rho_a(|x_i| + |x_j| + |x_i x_j|/a)
\geq \rho_a(|x_i| + |x_j|)$$

3.) By concavity of the function $\rho_a$,

$$\frac{\rho_a(|x_i|) + \rho_a(|x_j|)}{2} \leq \rho_a\left(\frac{|x_i| + |x_j|}{2}\right).$$

**Remark 2.1.** It follows from Lemma [2.7] that the triangular inequality holds for the function $\rho(x) \equiv \rho_a(|x|) : \rho(x_i + x_j) = \rho_a(|x_i| + |x_j|) \leq \rho_a(|x_i|) + \rho_a(|x_j|) = \rho(x_i) + \rho(x_j)$.

Also we have: $\rho(x) \geq 0$, and $\rho(x) = 0 \Leftrightarrow x = 0$. Our penalty function $\rho$ acts almost like a norm. However, it lacks absolute scalability, or $\rho(cx) \neq |c|\rho(x)$ in general. The next lemma further analyses this inequality.
Lemma 2.2.

\[ \rho_a(|cx|) = \begin{cases} \leq |c|\rho_a(|x|) & \text{if } |c| > 1; \\ \geq |c|\rho_a(|x|) & \text{if } |c| \leq 1. \end{cases} \]  

(2.6)

Proof.

\[ \rho_a(|cx|) = \frac{(a+1)|c|x}{a + |c||x|} = |c|\rho_a(|x|) \frac{a + |x|}{a + |cx|} \]

So if \(|c| \leq 1\), the factor \(\frac{a+|x|}{a+|cx|} \geq 1\). Then \(\rho_a(|cx|) \geq |c|\rho_a(|x|)\). Similarly when \(|c| > 1\), we have \(\rho_a(|cx|) \leq |c|\rho_a(|x|)\).

\[\qed\]

3 Theory of TL1 Minimization

3.1 RIP Condition for Constrained Model

For the constrained TL1 model (2.3), we present a theory on sparse recovery based on RIP [2].

Suppose \(\beta^0\) is a sparsest solution for \(l_0\) minimization s.t. \(A\beta^0 = y\), while another vector \(\beta\) is defined as

\[ \beta = \arg\min_{\beta \in \mathbb{R}^N} \{ P_a(\beta) \mid A\beta = y \}. \]  

(3.1)

We addressed the question whether the two vectors \(\beta\) and \(\beta^0\) are equal to each other. That is to say, under what condition we can recover the sparsest solution \(\beta^0\) via solving the relaxation problem (2.3).

For an \(M \times N\) matrix \(A\) and set \(T \subset \{1, ..., N\}\), let \(A_T\) be the matrix consisting of the column \(a_j\) of \(A\) for \(j \in T\). Similarly for vector \(x\), \(x_T\) is a sub-vector, consisting of components indexed from the set \(T\).

Definition 3.1. (Restricted Isometry Constant)

For each number \(s\), define the \(s\)-restricted isometry constant of matrix \(A\) as the smallest number \(\delta_s\) such that for all subset \(T\) with \(|T| \leq s\) and all \(x \in \mathbb{R}_{|T|}\), the inequality

\[ (1 - \delta_s)||x||^2_2 \leq ||A_Tx||^2_2 \leq (1 + \delta_s)||x||^2_2, \]

holds.

For a fixed \(y\), the under-determined linear system has infinitely many solutions. Let \(x\) be one solution of \(Ax = y\). It does not need to be the \(l_0\) or \(\rho_a\) minimizer. If \(P_a(x) > 1\), we scale \(y\) by the positive scalar \(C\) as:

\[ y_C = \frac{y}{C}; \quad x_C = \frac{x}{C}. \]  

(3.2)

Now \(x_C\) is a solution to the modified problem: \(Ax_C = y_C\). When \(C\) becomes larger, the number \(P_a(x_C)\) is smaller and tends to 0 in the limit \(C \to \infty\). Thus,
we can find a constant \( C \geq 1 \), such that \( P_a(x_C) \leq 1 \). That is to say, for scaled vector \( x_C \), we always have: \( P_a(x_C) \leq 1 \).

Since the penalty \( \rho_a(t) \) is increasing in positive variable \( t \), we have the inequality:

\[
P_a(x_C) \leq \frac{|T|\rho_a(|x_C|_\infty)}{aC + |x|_\infty},
\]

where \( |T| \) is the cardinality of the support set of vector \( x \). For \( P_a(x_C) \leq 1 \), it suffices to impose:

\[
\frac{|T|(a + 1)|x|_\infty}{aC + |x|_\infty} \leq 1,
\]

or:

\[
C \geq \frac{|x|_\infty}{a} (a|T| + |T| - 1). \tag{3.3}
\]

Let \( \beta^0 \) be the \( l_0 \) minimizer for the constrained \( l_0 \) optimization problem (1.1) with support set \( T \). Due to the scale-invariance of \( l_0 \), \( \beta^0_C \) (defined similarly as above) is a global \( l_0 \) minimizer for the modified problem:

\[
\min \| \beta \|_0, \text{ s.t. } y_C = A\beta. \tag{3.4}
\]

with the same support set \( T \). Then for the modified \( \rho_a \) optimization:

\[
\min P_a(\beta), \text{ s.t. } y_C = A\beta. \tag{3.5}
\]

we have the following:

**Theorem 3.1.** (Exact Sparse Recovery) For a given sensing matrix \( A \), if there is a number \( R > |T| \), such that

\[
\delta_R + \frac{R}{|T|} \delta_{R + |T|} < \frac{R}{|T|} - 1, \tag{3.6}
\]

then there exists \( a^* > 0 \), depending only on matrix \( A \), such that for any \( a > a^* \), the minimizer \( \beta_C \) for (3.3) is unique and equal to the minimizer \( \beta^0_C \) in (3.4) for any \( C \) satisfying (3.3).

**Proof.** The proof generally follows the lines of arguments in [2] and [6], while using special properties of the penalty function \( \rho_a \).

For simplicity, we denote \( \beta_C \) by \( \beta \) and \( \beta^0_C \) by \( \beta^0 \).

Define the function:

\[
f(a) = \frac{a^2}{(a + 1)^2} \frac{R}{|T|} (1 - \delta_{R + |T|}) - 1 - \delta_R
\]
It is continuous and increasing in the parameter $a$. Note that at $a = 0$, $f(0) = -1 - \delta_M < 0$, and as $a \to \infty$, $f(a) \to \frac{R}{|T|}(1 - \delta_{R+|T|}) - 1 - \delta_R > 0$ by (3.6).

There exists a constant $a^*$, such that $f(a^*) = 0$. The number $a^*$ depends on the RIP of matrix $A$ only, and so it is independent of the scalar $C$.

For $a > a^*$:

$$\delta_R + \frac{a^2}{(a + 1)^2} \frac{R}{|T|} \delta_{R+|T|} < \frac{a^2}{(a + 1)^2} \frac{R}{|T|} - 1. \quad (3.7)$$

Let $e = \beta - \beta^0$, and we want to prove that the vector $e = 0$. It is clear that, $e_{T^c} = \beta_{T^c}$, since $T$ is the support set of $\beta^0$. By the triangular inequality of $\rho_a$, we have:

$$P_a(\beta^0) - P_a(e_T) = P_a(\beta^0) - P_a(-e_T) \leq P_a(\beta_T).$$

Then

$$P_a(\beta^0) - P_a(e_T) + P_a(e_{T^c}) \leq P_a(\beta_T) + P_a(\beta_{T^c}) = P_a(\beta) \leq P_a(\beta^0)$$

It follows that:

$$P_a(\beta_{T^c}) = P_a(e_{T^c}) \leq P_a(e_T). \quad (3.8)$$

Now let us arrange the components at $T^c$ in the order of decreasing magnitude of $|e|$ and partition into $L$ parts: $T^c = T_1 \cup T_2 \cup \ldots \cup T_L$, where each $T_j$ has $R$ elements (except possibly $T_L$ with less). Also denote $T = T_0$ and $T_01 = T \cup T_1$. Since $Ae = A(\beta - \beta^0) = 0$, it follows that

$$0 = \|Ae\|_2 = \|A_{T_01}e_{T_01} + \sum_{j=2}^L A_{T_j}e_{T_j}\|_2$$

$$\geq \|A_{T_01}e_{T_01}\|_2 - \sum_{j=2}^L \|A_{T_j}e_{T_j}\|_2$$

$$\geq \sqrt{1 - \delta_{|T|} + R}\|e_{T_01}\|_2 - \sqrt{1 + \delta_R} \sum_{j=2}^L \|e_{T_j}\|_2 \quad (3.9)$$

At the next step, we derive two inequalities between the $l_2$ norm and function $P_a$, in order to use the inequality (3.8). Since

$$\rho_a(|t|) = \frac{(a + 1)|t|}{a + |t|} \leq (1 + \frac{1}{a})|t|$$

we have:

$$P_a(e_{T_0}) = \sum_{i \in T_0} \rho_a(|e_i|) \leq (1 + \frac{1}{a})\|e_{T_0}\|_1$$

$$\leq (1 + \frac{1}{a})\sqrt{|T|} \|e_{T_0}\|_2$$

$$\leq (1 + \frac{1}{a})\sqrt{|T_0|} \|e_{T_01}\|_2. \quad (3.10)$$
Now we estimate the $l_2$ norm of $e_{T}$ from above in terms of $P_a$. It follows from $\beta$ being the minimizer of the problem (3.5) and the definition of $x_C$ (3.2) that

$$P_a(\beta_{T^c}) \leq P_a(\beta) \leq P_a(x_C) \leq 1.$$  

For each $i \in T^c$, $\rho_a(\beta_i) \leq P_a(\beta_{T^c}) \leq 1$. Also since

$$\frac{(a + 1)|\beta_i|}{a + |\beta_i|} \leq 1 \iff (a + 1)|\beta_i| \leq a + |\beta_i| \iff |\beta_i| \leq 1$$  \hfill (3.11)

we have

$$|e_i| = |\beta_i| \leq \frac{(a + 1)|\beta_i|}{a + |\beta_i|} = \rho_a(|\beta_i|) \quad \text{for every } i \in T^c.$$  

Using the property that $\rho_a(t)$ is increasing for non-negative variable $t \geq 0$, and that $|e_i| \leq |e_k|$ for each $i \in T_j$ and $k \in T_{j-1}$, $j = 2, 3, \ldots, L$, we have

$$|e_i| \leq \rho_a(|e_i|) \leq P_a(e_{T_{j-1}})/R \Rightarrow \|e_{T_j}\|^2 \leq \frac{P_a(e_{T_{j-1}})^2}{R}$$

\Rightarrow  \|e_{T_j}\| \leq \frac{P_a(e_{T_{j-1}})}{R^{1/2}} \Rightarrow \sum_{j=2}^L \|e_{T_j}\| \leq \sum_{j=1}^L \frac{P_a(e_{T_j})}{R^{1/2}}$$  \hfill (3.12)

Finally, plug (3.10) and (3.12) into inequality (3.9) to get:

$$0 \geq \sqrt{1 - \delta_{|T| + R}} \frac{a}{(a + 1)|T|^{1/2}} P_a(e_T) - \sqrt{1 + \delta_R} \frac{1}{R^{1/2}} P_a(e_T)$$

$$\geq \frac{P_a(e_T)}{R^{1/2}} \left(\sqrt{1 - \delta_{|T| + R}} \frac{a}{a + 1} \sqrt{\frac{R}{|T|}} - \sqrt{1 + \delta_R}\right)$$  \hfill (3.13)

By (3.7), the factor $\sqrt{1 - \delta_{|T| + R}} \frac{a}{a + 1} \sqrt{\frac{R}{|T|}} - \sqrt{1 + \delta_R}$ is strictly positive, hence $P_a(e_T) = 0$, and $e_T = 0$. Also by inequality (3.8), $e_{T^c} = 0$. We have proved that $\beta_C = \beta_0^C$. The equivalence of (3.5) and (3.4) holds. \hfill \square

**Remark 3.1.** Theorem 3.1 contains a sufficient condition for $\beta$ to be the unique global minimizer of $l_0$ optimization problem (1.4). On the other hand, with a choice of $R = 3|T|$, our condition (3.6) becomes:

$$\delta_{3|T|} + 3\delta_{|T|} < 2$$  \hfill (3.14)

which is exactly the condition (1.6) of Theorem 1.1 in [2]. This is consistent with the fact that when parameter $a$ goes to $+\infty$, our penalty function $\rho_a$ recovers the $l_1$ norm.
Next, we prove that TL1 recovery is stable under noisy measurements, i.e.,
\[
\min P_a(\beta), \ s.t. \ |y_C - A\beta|_2 \leq \tau. \tag{3.15}
\]

**Theorem 3.2. (Stable Recovery Theory)** Under the same RIP condition and \(a^*\) in theorem 3.1, for \(a \geq a^*\), the solution \(\beta^n_C\) for optimization \((3.15)\) satisfies
\[
\|\beta^n_C - \beta^0_C\|_2 \leq D\tau,
\]
for some constant \(D\) depending only on the RIP condition.

**Proof.** Set \(n = A\beta - y_C\). In the proof, we use three related notations listed below for clarity:

- \(\beta^n_C\Rightarrow\) optimal solution for the noisy constrained problem \((3.15)\);
- \(\beta^0_C\Rightarrow\) optimal solution for the noiseless constrained problem \((3.5)\);
- \(\beta^0_C\Rightarrow\) optimal solution for the \(l_0\) problem \((3.4)\).

Let \(T\) be the support set of \(\beta^0_C\), i.e., \(T = \text{supp}(\beta^0_C)\), and vector \(e = \beta^n_C - \beta^0_C\). Following the proof of theorem 3.1, we obtain:
\[
\sum_{j=2}^{L} \|e_{T_j}\|_2 \leq \sum_{j=1}^{L} \frac{P_a(e_{T_j})}{R^{1/2}} = \frac{P_a(e_{T^c})}{R^{1/2}}
\]
and
\[
\|e_{T^c}\|_2 \geq \frac{a}{(a+1)\sqrt{|T|}} P_a(e_T).
\]

Further, due to the inequality \(P_a(\beta^n_C) = P_a(e_{T^c}) \leq P_a(e_T)\) and inequality \((3.9)\), we get
\[
\|Ae\|_2 \geq \frac{P_a(e_T)}{R^{1/2}} C_\delta,
\]
where \(C_\delta = \sqrt{1 - \delta R - |T|} \frac{a}{a + 1} \sqrt{\frac{R}{|T|} - \sqrt{1 + \delta R}}\).

By the initial assumption on the size of observation noise, we have
\[
\|Ae\|_2 = \|A\beta^n_C - A\beta^0_C\|_2 = \|n\|_2 \leq \tau, \tag{3.16}
\]
so we have: \(P_a(e_T) \leq \frac{\tau R^{1/2}}{C_\delta}\).

On the other hand, we know that \(P_a(\beta_C) \leq 1\) and \(\beta_C\) is in the feasible set of the noisy problem \((3.15)\). Thus we have the inequality: \(P_a(\beta^n_C) \leq P_a(\beta_C) \leq 1\). By \((3.11)\), \(\beta^n_{C,i} \leq 1\) for each \(i\). So, we have
\[
|\beta_{C,i}^n| \leq \rho_a(|\beta_{C,i}^0|). \tag{3.17}
\]
It follows that
\[
\|e\|_2 \leq \|e_T\|_2 + \|e_{T^*}\|_2 = \|e_T\|_2 + \|\beta_{C,T^*}\|_2
\]
\[
\leq \frac{\|A_T e_T\|_2}{\sqrt{1 - \delta_T}} + \|\beta_{C,T^*}\|_1
\]
\[
\leq \frac{\|A_T e_T\|_2}{\sqrt{1 - \delta_T}} + P_a(\beta_{C,T^*}) = \frac{\|A_T e_T\|_2}{\sqrt{1 - \delta_T}} + P_a(e_{T^*})
\]
\[
\leq \frac{\tau}{\sqrt{1 - \delta_R}} + P_a(e_T) \leq D_r.
\]

where constant number $D$ depends on $\delta_R$ and $\delta_{R+|T|}$. The second inequality uses the definition of RIP, while the first inequality in the last row comes from (3.10).

\[\square\]

### 3.2 Sparsity of Local Minimizer

We study properties of local minimizers of both the constrained problem (3.3) and the unconstrained model (2.4). As in $l_p$ and $l_{1-2}$ minimization \[33, 15\], a local minimizer of TL1 minimization extracts linearly independent columns from the sensing matrix $A$, with no requirement for $A$ to satisfy RIP. Reversely, we state additional conditions on $A$ for a stationary point to be a local minimizer besides the linear independence of the corresponding column vectors.

**Theorem 3.3.** (Local minimizer of constrained model)

Suppose $x^*$ is a local minimizer of the constrained problem (3.3) and $T^* = supp(x^*)$, then $A_{T^*}$ is of full column rank, i.e. columns of $A_{T^*}$ are linearly independent.

**Proof.** Here we argue by contradiction. Suppose that the column vectors of $A_{T^*}$ are not linearly independent, then there exists non-zero vector $v \in ker(A)$, such that $supp(v) \subseteq T^*$. For any neighbourhood of $x^*$, $N(x^*, r)$, we can scale $v$ so that:
\[
\|v\|_2 \leq \min\{r; |x^*_i|, i \in T^*\} \tag{3.18}
\]

Next we define:
\[
\xi_1 = x^* + v;
\xi_2 = x^* - v,
\]

so both $\xi_1$ and $\xi_2, \in N(x^*, r)$, and $x^* = \frac{1}{2}(\xi_1 + \xi_2)$. On the other hand, from $supp(v) \subseteq T^*$, we have that $supp(\xi_1), supp(\xi_2) \subseteq T^*$. Moreover, due to the inequality (3.18), vectors $x^*$, $x_1$, and $x_2$ are located in the same orthant, i.e. $sign(x^*_i) = sign(\xi_1,i) = sign(\xi_2,i)$, for any index $i$. It means that $\frac{1}{2}|\xi_1| + \frac{1}{2}|\xi_2| = \frac{1}{2}|\xi_1 + \xi_2|$. Since the penalty function $P_a(t)$ is strictly concave for non-negative variable $t$,
\[
\frac{1}{2}P_a(\xi_1) + \frac{1}{2}P_a(\xi_2) = \frac{1}{2}P_a(|\xi_1|) + \frac{1}{2}P_a(|\xi_2|)
\]
\[
< P_a\left(\frac{1}{2}|\xi_1| + \frac{1}{2}|\xi_2|\right) = P_a\left(\frac{1}{2}|\xi_1 + \xi_2|\right) = P_a(x^*).
\]
So for any fixed $r$, we can find two vectors $\xi_1$ and $\xi_2$ in the neighbourhood $\mathcal{N}(x^*, r)$, such that $\min\{P_a(\xi_1), P_a(\xi_2)\} \leq \frac{1}{2} P_a(\xi_1) + \frac{1}{2} P_a(\xi_2) < P_a(x^*)$. Both vectors are in the feasible set of the constrained problem (2.3), in contradiction with the assumption that $x^*$ is a local minimizer.

The same property also holds for local minimizers of the unconstrained problem (2.4), because a local minimizer of unconstrained problem is also a local minimizer for a constrained optimization [33] and [2]. We skip the details and state the result below.

**Theorem 3.4. (Local minimizer of unconstrained model)**

Suppose $x^*$ is a local minimizer of the unconstrained problem (2.4) and $T^* = \text{supp}(x^*)$, then columns of $A_{T^*}$ are linearly independent.

From the two theorems above, we conclude the following facts:

**Corollary 3.1.** (a) For any local minimizer of (2.3) or (2.4), e.g. $x^*$, the sparsity of $x^*$ is at most $\text{rank}(A)$;

(b) The number of local minimizers is finite, both for problem (2.3) and (2.4).

In [18], the authors studied sufficient conditions of a strict local minimizer for minimizing any penalty functions satisfying Condition 1. Here we specialize and simplify it for our concave TL1 function $\rho_a$.

For a convex function $h(\cdot)$, the sub-differential $\partial h(x)$ is the closed convex set:

$$\partial h(x) := \{y \in \mathbb{R}^N : h(z) \geq h(z_0) + \langle z - z_0, y \rangle, \forall z \in \mathbb{R}^N\}, \quad (3.19)$$

which generalizes the derivative in the sense that $h$ is differentiable at $x$ if and only if $\partial h(x)$ is a singleton or $\{\nabla h(x)\}$.

The TL1 penalty function $p_a(\cdot)$ can be written as a difference of two convex functions:

$$p_a(t) = \frac{(a+1)t}{a+t},$$

$$= \frac{(a+1)t}{a} - \left( \frac{(a+1)t}{a} - \frac{(a+1)t}{a+t} \right), \quad (3.20)$$

Thus the general derivative of function $P_a(\cdot)$, as a combination of two convex derivatives. Denote: $\partial P_a(x)$: $\partial P_a(x) = \prod_{i=1}^N I_i^a \subset \mathbb{R}_N$, where

$$I_i^a = \begin{cases} 
\text{sgn}(x_i) \frac{(a+1)x_i}{a + |x_i|} - \frac{(a+1)x_i}{(a+|x_i|)^2}, & \text{if } i \in \text{supp}(x), \\
[-1, 1] \frac{(a+1)x_i}{a + |x_i|} - \frac{(a+1)x_i}{(a+|x_i|)^2}, & \text{otherwise.}
\end{cases} \quad (3.21)$$
Also \( \partial \|x\|_1 = \prod_{i=1}^{N} I_i \), where \( I_i \) is defined as:

\[
I_i = \begin{cases} 
\text{sgn}(x_i), & \text{if } i \in \text{supp}(x), \\
[-1, 1], & \text{otherwise.}
\end{cases}
\]  

(3.22)

Thus we can rewrite \( \partial P_a(x) \):

\[
\partial P_a(x) = \prod_{i=1}^{N} \left( I_i \left( \frac{a+1}{a+\|x_i\|} \right) \|x_i\| - \left( \frac{a+1}{a+\|x_i\|} \right) x_i \right) = \partial \|x\|_1. \ast \partial P_a(|x|),
\]

where "\( \ast \)" is the element-wise product of two vectors.

**Definition 3.2.** (Maximum concavity and local concavity of the penalty function)

For a penalty function \( \rho \), we define its maximum concavity as:

\[
\kappa(\rho) = \sup_{t_1, t_2 \in (0, \infty), t_1 < t_2} \frac{\rho'(t_2) - \rho'(t_1)}{t_2 - t_1}
\]  

(3.23)

and its local concavity of \( \rho \) at a point \( b = (b_1, b_2, ..., b_R) \in \mathbb{R}^R \) as:

\[
\kappa(\rho; b) = \lim_{\epsilon \to 0^+} \max_{1 \leq i \leq R} \sup_{t_1, t_2 \in (|b_i| - \epsilon, |b_i| + \epsilon), t_1 < t_2} \frac{\rho'(t_2) - \rho'(t_1)}{t_2 - t_1}
\]  

(3.24)

Let us recapitulate all a set of sufficient conditions on the (strict) local minimizer of (2.4):

**Condition 2.** For vector \( \beta \in \mathbb{R}^N \), \( \lambda > 0 \), and \( T = \text{supp}(\beta) \):

\[\text{C2.1: Matrix } Q = A_T^t A_T \text{ is non-singular, i.e. matrix } A_T^t \text{ is column independent;}\]

\[\text{C2.2: for vector } z = \frac{1}{\lambda} A_T^t (y - A\beta), \|z\|_\infty < \rho_a'(0+) = \frac{a+1}{a};\]

\[\text{C2.3: vector } \beta_T \text{ satisfies the stationary point equation: } \beta_T = Q^{-1} A_T^t y - \lambda Q^{-1} w_T, \quad \text{where } w_T \in \partial P_a(\beta);\]

\[\text{C2.4: } \lambda_{\min}(Q) \geq \lambda \kappa(\rho_a; \beta_T), \quad \text{where } \lambda_{\min}(\cdot) \text{ denotes the smallest eigenvalues of a given symmetric matrix.}\]

Under Condition 2, Lv and Fan [18] showed the following:

**Theorem 3.5.** If a vector \( \beta \in \mathbb{R}^N \) satisfies all four requirements in Condition 2, then \( \beta \) is a local minimizer of problem (2.4).

Furthermore, if the inequality of (C2.4) is strict, then the vector \( \beta \) is a strict local minimizer.
Proof. Here we gave a simplified proof of the theorem to illustrate each of the conditions (C2.1)-(C2.4). The objective function is:

$$\ell(x) = 2^{-1} ||Ax - y||_2^2 + \lambda P_a(x) = 2^{-1} ||Ax - y||_2^2 + \lambda \sum_{j=1}^{N} \rho_a(x_j).$$

Let us define a subspace of $\mathbb{R}^N$ as: $S = \{ \beta \in \mathbb{R}^N | \beta_{T^c} = 0 \}$.

First, by Condition 1, (C2.4), and the definition of $\kappa(\rho_a; \beta_T)$, the objective function $\ell(\cdot)$ is convex in $\mathcal{N}(\beta, r_0) \cap S$, where $r_0$ is a positive number (the radius). By equation (C2.3), $\beta_T$ is a local minimizer of $\ell(\cdot)$ in $S$.

Next, we show that the sparse vector $\beta$ is indeed a local minimizer of $\ell(\cdot)$ in $\mathbb{R}^N$. Because of inequality (C2.2), there exists $\delta \in (0, \infty)$, a positive number $r_1 < \delta$, such that $\rho_a(\delta) \in (0, \rho_a(0^+)]$, and for any vector $x \in \mathcal{N}(\beta, r_1)$,

$$\|w_{T^c}\|_\infty < \rho_a(\delta),$$

where $w = \lambda^{-1}A^T(y - A\beta)$. We can further shrink $r_1$ if necessary so that $r_1 < r_0$, $\mathcal{N}(\beta, r_1) \subseteq \mathcal{N}(\beta, r_0)$. By the mean value theorem, $\forall \beta_1 \in \mathcal{N}(\beta, r_1)$,

$$\ell(\beta_1) = \ell(\beta_2) + \nabla^t \ell(\bar{\beta}_0)(\beta_1 - \beta_2),$$

where $\beta_2$ is the projection of $\beta_1$ onto $S$ and $\bar{\beta}_0$ lies on the line segment joining $\beta_1$ and $\beta_2$. Since,

$$(\beta_1 - \beta_2)^T = 0, \ \beta_0 \in \mathcal{N}(\beta, r_1) \ \text{and} \ \text{sign}(\bar{\beta}_0,T^c) = \text{sign}(\beta_1,T^c),$$

we have the following inequality:

$$\ell(\beta_1) - \ell(\beta_2) = \partial \ell(\bar{\beta}_0,T^c) * \beta_1,T^c = \left[ A_{T^c}^T A_{T^c} \beta_0 - A_{T^c}^T y + \lambda P_a'(\bar{\beta}_0,T^c) \right] * \beta_1,T^c
= -\lambda [A_{T^c}^T(y - A\beta_0)]^t \beta_1,T^c + \lambda \left( \partial \| \beta_0,T^c \|_1 \ast P_a'(\bar{\beta}_0,T^c) \right) * \beta_1,T^c
\geq -\lambda \rho_a'(\delta) \| \beta_1,T^c \|_1 + \lambda P_a'(\| \beta_0,T^c \|_1) \ast | \beta_1,T^c |
\geq -\lambda \rho_a'(\delta) \| \beta_1,T^c \|_1 + \lambda \rho_a'(\delta) \| \beta_1,T^c \|_1 = 0,$$

where we also used the property of generalized derivative $\partial \| \cdot \|_1$, and $\ast$ stands for vector cross product.

So for any $\beta_1 \in \mathcal{N}(\beta, r_1)$, $\ell(\beta_1) > \ell(\beta_2)$. Since $\beta_2$ is a projection on $S$ and it belongs to the ball $\mathcal{N}(\beta, r_1) \subseteq \mathcal{N}(\beta, r_0)$,

$$\ell(\beta_1) > \ell(\beta_2) \geq \ell(\beta).$$

The (C2.4) is only used in the first part of the proof. If we have the strict inequality $\lambda_{\min}(Q) > \lambda \kappa(\rho_a; \beta_T)$, then $\beta_T$ is a strict local minimizer in $S$, as the function $\ell(\cdot)$ is strictly convex in the intersection $\mathcal{N}(\beta, r_0) \cap S$. Further, the same proof shows that $\beta$ is a strict local minimizer in $\mathbb{R}^N$.

$\square$
4 DC Algorithm for Transformed $l_1$ Penalty

4.1 DC Programming

Generally speaking, a DC program is an optimization problem of the form:

$$\alpha = \inf \{ f(x) = g(x) - h(x) : x \in \mathbb{R}^d \} \quad (P_{dc})$$

where $g, h$ are lower semi-continuous proper convex functions on $\mathbb{R}^d$.\[25, 22\]

The difference of convex function algorithm (DCA) approximates $h$, at the current point $x^l$ of iteration, by its affine minorization defined by

$$h_l(x) = h(x^l) + \langle x - x^l, y^l \rangle, \quad y^l \in \partial h(x^l)$$

to produce a convex program in the form:

$$\inf \{ g(x) - h_l(x) : x \in \mathbb{R}^d \} \Leftrightarrow \inf \{ g(x) - \langle x, y^l \rangle : x \in \mathbb{R}^d \}$$

where the optimal solution is denoted as $x^{l+1}$.

4.2 Algorithm for Unconstrained Model — DCATL1

Consider the following unconstrained optimization problem [24]:

$$\min_{x \in \mathbb{R}^N} f(x) = \min_{x \in \mathbb{R}^N} \frac{1}{2} \|Ax - y\|_2^2 + \lambda P_a(x)$$

where

$$P_a(x) = \sum_{i=1, \ldots, N} \rho_a(|x_i|).$$

The DCA for this problem is:

$$\begin{align*}
    f(x) &= g(x) - h(x) \\
    g(x) &= \frac{1}{2} \|Ax - y\|_2^2 + c\|x\|_2^2 + \lambda \frac{(a+1)}{a} \|x\|_1 \\
    h(x) &= \lambda \frac{(a+1)}{a} \|x\|_1 - \lambda P_a(x) + c\|x\|_2^2
\end{align*} \quad (4.1)$$

Define:

$$v^n = \lambda \frac{a+1}{a} \text{sign}(x^n) - \lambda \text{sign}(x^n) \frac{(a+1)}{a + |x^n|} + \lambda \frac{(a+1)x^n}{(a + |x^n|)^2} + 2cx^n, \quad (4.2)$$
then $v^n \in \partial h(x^n)$.

**Algorithm 1:** DCA for unconstrained transformed $l_1$ penalty minimization

Define: $\epsilon_{outer} > 0$, $\epsilon_{inner} > 0$

Initialize: $x^0 = 0$, $n = 0$

while $|x^{n+1} - x^n| > \epsilon_{outer}$ do

$v^n = \lambda \frac{a+1}{a} \text{sign}(x^n) - \lambda \text{sign}(x^n) \frac{a+1}{a} x^n + 2c x^n$

$x^{n+1} = \arg \min_{x \in \mathbb{R}^N} \left\{ \frac{1}{2} \|Ax - y\|_2^2 + c \|x\|^2 + \lambda \frac{a+1}{a} \|x\|_1 - \langle x, v^n \rangle \right\}$

then $n + 1 \rightarrow n$

end while

At each step, we need to solve a strongly convex $l_1$-regularized sub-problem, which is:

$$x^{n+1} = \arg \min_{x \in \mathbb{R}^N} \left\{ \frac{1}{2} \|Ax - y\|_2^2 + c \|x\|^2 + \lambda \frac{a+1}{a} \|x\|_1 - \langle x, v^n \rangle \right\}$$

We now employ the Alternating Direction Method of Multipliers (ADMM). The sub-problem is recast as:

$$x^{n+1} = \arg \min_{x \in \mathbb{R}^N} \left\{ \frac{1}{2} x^t (A^t A + 2c I) x - \langle x, v^n + A^t y \rangle + \lambda \frac{a+1}{a} \|x\|_1 \right\}$$

s.t. $x - z = 0$.

Define the augmented Lagrangian function as:

$$L(x, z, u) = \frac{1}{2} x^t (A^t A + 2c I) x - \langle x, v^n + A^t y \rangle + \lambda \frac{a+1}{a} \|z\|_1 + \frac{\delta}{2} \|x - z\|_2^2 + u^t (x - z),$$

where $u$ is the Lagrange multiplier, and $\delta > 0$ is a penalty parameter. The ADMM consists of three iterations:

$$\begin{align*}
x^{n+1} &= \arg \min_x L(x, z^n, u^n); \\
z^{n+1} &= \arg \min_z L(x^{n+1}, z, u^n); \\
u^{n+1} &= u^n + \frac{x^{n+1} - z^{n+1}}{\delta}.
\end{align*}$$

The first two steps have closed-form solutions and are described in Algorithm 2 where $\text{shrink}(., .)$ is a soft-thresholding operator given by:

$$\text{shrink}(x, r)_i = \text{sign}(x_i) \max\{|x_i| - r, 0\}.$$
Algorithm 2: ADMM for subproblem (4.3)

Initial guess: $x^0, z^0$ and $u^0$

while not converged do
  $x^{n+1} := (A^T A + 2 c I + \delta I)^{-1} (A^T y - v + \delta z^n - u^n)$
  $z^{n+1} := \text{shrink}(x^{n+1} + u^n, \frac{a+1}{a} \lambda)$
  $u^{n+1} := u^n + x^{n+1} - z^{n+1}$
  then $n \rightarrow n + 1$
end while

4.3 Convergence Theory for Unconstrained DCATL1

We present a convergence theory for the Algorithm 1 (DCATL1). We prove that the sequence $\{f(x^n)\}$ is decreasing and convergent, while the sequence $\{x^n\}$ is bounded under some requirement on $\lambda$. Its sub-limit vector $x^*$ is a stationary point satisfying the first order optimality condition. Our proof is based on the convergent theory of DCA for $l_1 - l_2$ penalty function [33] besides the general results [23] [24].

Definition 4.1. (Modulus of strong convexity) For a convex function $f(x)$, the modulus of strong convexity of $f$ on $\mathbb{R}^N$, denoted as $m(f)$, is defined by

$$m(f) := \sup\{\rho > 0 : f - \frac{\rho}{2} \|x\|^2 \text{ is convex on } \mathbb{R}^N\}.$$ 

Let us recall a useful inequality from [24] concerning the sequence $f(x^n)$.

Lemma 4.1. Suppose that $f(x) = g(x) - h(x)$ is a D.C. decomposition, and the sequence $\{x^n\}$ is generated by (4.3), then

$$f(x^n) - f(x^{n+1}) \geq \frac{\rho(g) + \rho(h)}{2} \|x^{n+1} - x^n\|^2.$$ 

Here is the convergence theory for our unconstrained Algorithm 1—DCATL1. The objective function is: $f(x) = \frac{1}{2} \|Ax - y\|^2 + \lambda P_a(x)$.

Theorem 4.1. The sequences $\{x^n\}$ and $\{f(x^n)\}$ in Algorithm 1 satisfy:

(I) Sequence $\{f(x^n)\}$ is decreasing and convergent.

(II) $\|x^{n+1} - x^n\|_2 \rightarrow 0$ as $n \rightarrow \infty$. If $\lambda > \frac{\|y\|^2}{a + 1}, \{x^n\}_{n=1}^{\infty}$ is bounded.

(III) Any subsequential limit vector $x^*$ of $\{x^n\}$ satisfies the first order optimality condition:

$$0 \in A^T (Ax^* - y) + \lambda \partial P_a(x^*), \quad (4.5)$$

implying that $x^*$ is a stationary point of (2.4).
Proof. 1. By the definition of \(g(x)\) and \(h(x)\) in equation (4.1), it is easy to see that:

\[
\rho(g) \geq 2c; \\
\rho(h) \geq 2c.
\]

By Lemma 4.1, we have:

\[
f(x^n) - f(x^{n+1}) \geq \rho(g) + \rho(h)2\|x^{n+1} - x^n\|_2^2.
\]

So the sequence \(\{f(x^n)\}\) is decreasing and non-negative, thus convergent.

2. It follows from the convergence of \(\{f(x^n)\}\) that:

\[
\|x^{n+1} - x^n\|_2^2 \leq \frac{f(x^n) - f(x^{n+1})}{2c} \to 0, \text{ as } n \to \infty.
\]

If \(y = 0\), since the initial vector \(x^0 = 0\), and the sequence \(\{f(x^n)\}\) is decreasing, we have \(f(x^n) = 0, \forall n \geq 1\). So \(x^n = 0\), and the boundedness holds.

Consider non-zero vector \(y\). Then

\[
f(x^n) = \frac{1}{2}\|Ax^n - y\|_2^2 + \lambda P_a(x^n) \leq f(x^0) = \|y\|_2^2.
\]

So \(\lambda P_a(x^n) \leq \|y\|_2^2\), implying \(\lambda P_a(\|x^n\|_\infty) \leq \|y\|_2^2\), or:

\[
\frac{\lambda(a + 1)}{a + \|x^n\|_\infty} \leq \|y\|_2^2.
\]

So if \(\lambda > \frac{\|y\|_2^2}{a + 1}\), then

\[
|\mathbf{x}| \leq \frac{a\|y\|_2^2}{\lambda(a + 1) - \|y\|_2^2},
\]

or the sequence \(\{x_n\}_{n=1}^\infty\) is bounded.

3. Let \(\{x^n_k\}\) be a subsequence of \(\{x^n\}\) which converges to \(x^*\). So the optimality condition at the \(n_k\)-th step of Algorithm 1 is expressed as:

\[
0 \in A^T(Ax^{n_k} - y) + \partial\|x^{n_k} - x^{n_k - 1}\|_1 + \lambda(\frac{a + 1}{a})\partial\|x^{n_k} - x^{n_k - 1}\|_1 + \lambda \partial P_a(x^{n_k} - 1).
\]

Since \(\|x^{n+1} - x^n\|_2 \to 0\) as \(n \to \infty\) and \(x^{n_k}\) converges to \(x^*\), as shown in Proposition 3.1 of [33], we have that for sufficiently large index \(n_k\),

\[
\partial\|x^{n_k} - x^*\|_1 \leq \partial\|x^*\|_1 \\
\partial\|x^{n_k - 1} - x^*\|_1 \leq \partial\|x^*\|_1 \\
P_a(x^{n_k} - 1) \leq \partial P_a(x^*).
\]

Letting \(n_k \to \infty\) in (4.6), we have \(0 \in A^T(Ax^* - y) + \lambda \partial P_a(x^*).

\[\square\]
Remark 4.1. The above theorem says that the sub-sequence limit \( x^* \) is a stationary point for (2.4). Let \( T^* = \text{supp}(x^*) \), there exists vector \( w \in \partial P_a(x^*) \), s.t.

\[
0 = A^t(Ax^* - y) + \lambda w \\
\Rightarrow 0 = A^t_{T^*}(AT^*x^* - y) + \lambda w_{T^*} \\
\Rightarrow 0 = Qx^*_{T^*} - A^t_{T^*}y + \lambda w_{T^*} \\
\Rightarrow x^*_{T^*} = Q^{-1}A^t_{T^*}y - \lambda Q^{-1}w_{T^*}.
\]

(4.7)

So (C2.3) is automatically satisfied by \( x^* \). If (C2.1)-(C2.3) are also satisfied, the limit point \( x^* \) is a local minimizer of (2.4).

4.4 Algorithm for Constrained Model

We also use DCA to solve the constrained problem:

\[
\min_{x \in \mathbb{R}^N} P_a(x) \quad \text{s.t.} \quad Ax = y \\
\Leftrightarrow \min_{x \in \mathbb{R}^N} \frac{a+1}{a} \|x\|_1 - \left\{ \frac{a+1}{a} \|x\|_1 - P_a(x) \right\} \quad \text{s.t.} \quad Ax = y
\]

Choose vector \( z \in \frac{a+1}{a} \partial \|x\|_1 - \partial P_a(x) \), then the convex sub-problem is:

\[
\min_{x \in \mathbb{R}^N} \frac{a+1}{a} \|x\|_1 - \langle z, x \rangle \quad \text{s.t.} \quad Ax = y
\]

(4.8)

To solve (4.8), we introduce two Lagrange multipliers \( u, v \) and define an augmented Lagrangian:

\[
L_\delta(x, w, u, v) = \frac{a+1}{a} \|x\|_1 - z^t x + u^t (x - w) + v^t (Ax - y) + \frac{\delta}{2} \|x - w\|^2 + \frac{\delta}{2} \|Ax - y\|^2,
\]

where \( \delta > 0 \). ADMM finds a saddle point \( (x^*, w^*, u^*, v^*) \), such that:

\[
L_\delta(x^*, w^*, u, v) \leq L_\delta(x^*, w^*, u^*, v^*) \leq L_\delta(x, w, u^*, v^*) \quad \forall x, w, u, v
\]

by alternately minimizing \( L_\delta \) with respect to \( x \), minimizing with respect to \( y \) and updating the dual variables \( u \) and \( v \). The saddle point \( x^* \) will be a solution to (4.8). The overall algorithm for solving the constrained TL1 is described in
Algorithm (3). The explicit expressions for $z$ come from (3.21) and (3.22).

**Algorithm 3:** DCA method for constrained TL1 minimization

Define $\epsilon_{outer} > 0$, $\epsilon_{inner} > 0$ and initialize $x^0 = 0$

\[\text{while } \|x^n - x^{n+1}\| \geq \epsilon_{outer} \text{ do} \]

\[z = (z_1, \ldots, z_N), \text{ where } z_i = \frac{a + 1}{\alpha} \text{sign}(x_i) - \frac{(a+1)\text{sign}(x_i)}{\alpha + |x_i|} + \frac{(a+1)x_i}{(\alpha + |x_i|)^2}.\]

\[x^{1}_{in} = x^n, \quad j = 1, \quad w^1 = x^{1}_{in}, \quad v^1 = 0 \text{ and } u^1 = 0.\]

\[\text{while } \|x^j_{in} - x^{j+1}_{in}\| \geq \epsilon_{inner} \text{ do} \]

\[x^{j+1}_{in} := (A^tA + I)^{-1}(w^j + A^ty + z - \frac{u^j - A^tv^j}{\delta})\]

\[w^{j+1} := \text{shrink}(x^{j+1}_{in} + \frac{u^j}{\delta}, \frac{a+1}{\alpha \delta})\]

\[u^{j+1} := u^j + \delta(x^{j+1}_{in} - w^{j+1})\]

\[v^{j+1} := v^j + \delta(Ax^{j+1} - y)\]

end while

\[n = n + 1\]

\[x^n = x^n\]

end while

5 Numerical Results

In this section, we use three classes of randomly generated matrices to illustrate the effectiveness of our Algorithms: DCATL1 (difference convex algorithm for transformed $l_1$ penalty) and its constrained version. We compare them separately with several state-of-the-art solvers on recovering sparse vectors:

- unconstrained algorithms: reweighted $l_{1/2}$ [12], yall1 (an alternating direction $l_1$ algorithm) [32] and DCA $l_{1-2}$ algorithm [33] [15];

- constrained algorithms: Bregman algorithm [34], yall1, and $L_p - RLS$ [7].

All our tests were performed on a Lenovo desktop with 16 GB of RAM and Intel Core processor i7 – 4770 with CPU at 3.40GHz x8 under 64-bit Ubuntu system.

The three classes of random matrices are:

1) Gaussian matrix.

2) Over-sampled DCT with factor $F$.

3) Uniformly distributed M-sphere matrix.

We did not use prior information of the true sparsity level. Also, for all the tests, the computation is initialized with zero vectors. In fact, the DCATL1 does not guarantee a global minimum in general, due to nonconvexity of the problem. Indeed we observe that DCATL1 with random starts often gets stuck at local minima especially when the matrix $A$ is ill-conditioned (e.g. $A$ has a large condition number or is highly coherent). In the numerical experiments, by setting $x_0 = 0$, we find that DCATL1 usually produces a global minimizer. The
intuition behind our choice is that by using zero vector as initial guess, the first step of our algorithm reduces to solving an unconstrained weighted $l_1$ problem. So basically we are minimizing TL1 on the basis of $l_1$, which possibly explains why minimization of TL1 initialized by $x_0 = 0$ always outperforms $l_1$.

Choice of Parameter: a

In DCATL1, parameter a is also very important. When a tends to zero, the penalty function approaches the $l_0$ norm. If a goes to $\infty$, objective function will be more convex and act like the $l_1$ optimization. So choosing a better ‘a’ will improve the effectiveness and success rate for our algorithm.

We tested DCATL1 on recovering sparse vectors with different parameter a, varying among $\{0.1 \ 0.3 \ 1 \ 2 \ 10\}$. In this test, $A$ is a $64 \times 256$ random matrix generated by normal Gaussian distribution. The true vector $x^*$ is also a randomly generated sparse vector with sparsity $k$ from the set $\{8 \ 10 \ 12 \ \ldots \ 32\}$. Here the parameter $\lambda$ was set to be $10^{-5}$ for all tests. Although the best $\lambda$ should be dependent on a in general, we considered the noiseless case, and $\lambda = 10^{-5}$ is small enough to approximately enforce $Ax = Ax^*$. For each a, we sampled 100 times with different $A$ and $x^*$. The recovered vector $x_r$ is regarded as successful if the relative error: $\frac{\|x_r - x^*\|_2}{\|x^*\|_2} \leq 10^{-3}$.

Fig. 2 shows the success rate using DCATL1 over 100 independent trials for various parameter a and sparsity k. From the figure, we see that DCATL1 with $a = 1$ is the best among all tested values. Also numerical results for $a = 0.3$ and $a = 2$ (near 1), are better than those with 0.1 and 10. This is because the objective function is more non-convex at a smaller ’a’ and thus more difficult to solve. On the other hand, the iterations more likely stop at a local $l_1$ minima far from $l_0$ solution if $a$ is too large. Thus for all the following tests, we set the parameter $a = 1$.

5.1 Numerical Experiment for Unconstrained Algorithm

5.1.1 Gaussian matrix

We use $\mathcal{N}(0, \Sigma)$, the multi-variable normal distribution to generate Gaussian matrix $A$. Here covariance matrix is $\Sigma = \{(1 - r) * \chi_{(i=j)} + r\}_{i,j}$, where the value of ‘r’ varies from 0 to 0.8. In theory, the larger the $r$ is, the more difficult it is to recovery true sparse vector. For matrix $A$, the row number and column number are set to be $M = 64$ and $N = 1024$. The sparsity $k$ varies among $\{5 \ 7 \ 9 \ \ldots \ 25\}$.

We compare four algorithms in terms of success rate. Denote $x_r$ as a reconstructed solution by a certain algorithm. We consider one algorithm to be successful, if the relative error of $x_r$ to the truth solution $x$ is less than 0.001, i.e., $\frac{\|x_r - x\|}{\|x\|} < 1.0e - 3$. In order to improve success rates for all compared algorithms, we set tolerance parameter to be smaller or maximum cycle number
Figure 2: Numerical tests on parameter $a$ with $M = 64$, $N = 256$ by the unconstrained DCATL1 method.
to be higher inside each algorithm. As a result, it takes a long time to run one realization using all algorithms separately.

The success rate of each algorithm is plotted in Figure 3 with parameter $r$ from the set: \{0.0 0.2 0.6 0.8\}. For all cases, DCATL1 and reweighted $l_{1/2}$ algorithms (IRucL$q$-v) performed almost the same and both were much better than the other two, while the $l_1$ algorithm (yall1) has the lowest success rate.

5.1.2 Over-sampled DCT

The over-sampled DCT matrices $A$ [13] [15] are:

$$A = [a_1, ..., a_N] \in \mathbb{R}^{M \times N}$$

where $a_j = \frac{1}{\sqrt{M}} \cos \left( \frac{2\pi \omega(j - 1)}{F} \right)$, $j = 1, ..., N$, \hspace{1cm} (5.1)

and $\omega$ is a random vector, drawn uniformly from $(0, 1)^M$.
Table 1: The success rates (%) of DCATL1 for different combination of sparsity and minimum separation lengths.

| sparsity | 5  | 8  | 11 | 14 | 17 | 20 |
|----------|----|----|----|----|----|----|
| 1RL      | 100| 100| 95 | 70 | 22 | 0  |
| 2RL      | 100| 100| 98 | 74 | 19 | 5  |
| 3RL      | 100| 100| 97 | 71 | 19 | 3  |
| 4RL      | 100| 100| 100| 71 | 20 | 1  |
| 5RL      | 100| 100| 96 | 70 | 28 | 1  |

Such matrices appear as the real part of the complex discrete Fourier matrices in spectral estimation [13]. An important property is their high coherence: for a $100 \times 1000$ matrix with $F = 10$, the coherence is 0.9981, while the coherence of the same size matrix with $F = 20$, is typically 0.9999.

The sparse recovery under such matrices is possible only if the non-zero elements of solution $x$ are sufficiently separated. This phenomenon is characterized as minimum separation in [5], and this minimum length is referred as the Rayleigh length (RL). The value of RL for matrix $A$ is equal to the factor $F$. It is closely related to the coherence in the sense that larger $F$ corresponds to larger coherence of a matrix. We find empirically that at least $2\text{RL}$ is necessary to ensure optimal sparse recovery with spikes further apart for more coherent matrices.

Under the assumption of sparse signal with $2\text{RL}$ separated spikes, we compare those four algorithms in terms of success rate. Denote $x_r$ as a reconstructed solution by a certain algorithm. We consider one algorithm successful, if the relative error of $x_r$ to the truth solution $x$ is less that 0.001, i.e., $\|x_r - x\|/\|x\| < 0.001$.

The success rate is averaged over 50 random realizations.

Fig. 4 shows success rates for those algorithms with increasing factor $F$ from 2 to 20. The sensing matrix is of size $100 \times 1500$. It is interesting to see that along with the increasing of value $F$, DCA of $l_1 - l_2$ algorithm performs better and better, especially after $F \geq 10$, and it has the highest success rate among all. Meanwhile, reweighted $l_{1/2}$ is better for low coherent matrices. When $F \geq 10$, it is almost impossible for it to recover sparse solution for the high coherent matrix. Our DCATL1, however, is more robust and consistently performed near the top, sometimes even the best. So it is a valuable choice for solving sparse optimization problems where coherence of sensing matrix is unknown.

We further look at the success rates of DCATL1 with different combinations of sparsity and separation lengths for the over-sampled DCT matrix $A$. The rates are recorded in Table 1 which shows that when the separation is above with the minimum length, the sparsity relative to $M$ plays more important role in determining the success rates of recovery.
Figure 4: Numerical test for unconstrained algorithms under over-sampled DCT matrices: \(M = 100, N = 1500\) with different \(F\), and peaks of solutions separated by \(2RL = 2F\).
5.1.3 Uniformly distributed $M$-sphere matrix

The column vectors of the $M \times N$ matrix $A$ are sampled from uniform distribution on the surface of unit $M$-hypersphere. We use standard normal distribution $\mathcal{N}(0, I_M)$ to generate a random matrix $B$, then normalize each column vector of matrix $B$ to unit $l_2$ norm. The density function of multi-variate normal random variables is:

$$f_{\mathbf{x}}(x_1, \ldots, x_M) = \frac{1}{(2\pi)^{M/2}} \exp \left( - \sum_{i=1}^{M} \frac{x_i^2}{2} \right)$$

and the column vectors of matrix $A$ are independent. After normalization, distribution of column vectors in matrix $A$ is uniform on the $M$-sphere surface.

In our numerical tests, we vary three parameters: column number $N$, row number $M$, and sparsity $k$. We fixed two parameters, and changed the value of the remaining one. The results are based on 100 trials. In the top-left plot of Fig. 5 we chose $M = 50$ and $N = 500$, the sparsity $k$ varied from 3 to 13. We see that the success rate curves for reweighted $l_{1/2}$, DCA $l_1 - l_2$ and DCATL1 are almost the same, with a little higher value for TL1. In the other three plots, we fixed parameters: $M$ and $k$, with row number $N$ changing from 100 to 1900. The mean values of the coherence $\mu$ of $A$ (maximum absolute value of pairwise column vector inner product) for different combinations of $M$ and $N$, are shown in Table 2. It is interesting to see that for all three cases, curves for DCA $l_1 - l_2$ and DCATL1 are almost identical, and better than the other two (rewighted $l_{1/2}$ and $l_1$). The DCA algorithm for both $l_1 - l_2$ and TL1 may have helped too.

Table 2: Mean coherence of $M$-sphere random matrices for different values of $(M, N)$ over 100 samples.

| Column number $N$: | 100 | 300 | 500 | 700 | 900 |
|-------------------|-----|-----|-----|-----|-----|
| Mu: $M = 10$     | 0.9065 | 0.9431 | 0.9537 | 0.9605 | 0.9651 |
| Mu: $M = 20$     | 0.7418 | 0.8024 | 0.8198 | 0.8363 | 0.8416 |
| Mu: $M = 50$     | 0.5105 | 0.5680 | 0.5943 | 0.6031 | 0.6119 |

| Column number $N$: | 1100 | 1300 | 1500 | 1700 | 1900 |
|-------------------|------|------|------|------|------|
| Mu: $M = 10$     | 0.9674 | 0.9685 | 0.9708 | 0.9742 | 0.9733 |
| Mu: $M = 20$     | 0.8526 | 0.8573 | 0.8643 | 0.8634 | 0.8663 |
| Mu: $M = 50$     | 0.6209 | 0.6309 | 0.6364 | 0.6378 | 0.6454 |
$\{M, N\} = (50, 500)$, mean coherence $\mu = 0.5939$.

$M = 10, k = 2$

$M = 20, k = 3$

$M = 50, k = 8$

Figure 5: Comparison of success rates of four unconstrained algorithms for $M$-sphere random matrices. In the top left plot, $(M, N) = (50, 500)$, sparsity $k$ varies from 3 to 15. In the other three plots, $(M, k)$ are fixed, and $N$ is varied.
5.2 Numerical Experiment for Constrained Algorithm

For constrained algorithms, we performed similar numerical experiments. An algorithm is considered successful if the relative error of the numerical result $x_r$ from the ground truth $x$ is less than 0.001, or $\frac{|x_r - x|}{\|x\|} < 0.001$. We did 50 trials to compute average success rates for all the numerical experiments as for the unconstrained algorithms.

5.2.1 Gaussian Random Matrices

We fix parameters $(M, N) = (64, 1024)$, while covariance parameter $r$ is varied from 0 to 0.8. Comparison is with the reweighted $l_{1/2}$ and two $l_1$ algorithms (Bregman and yall1). In Fig. 6, we see that $Lp - RLS$ is the best among the four algorithms with DCATL1 trailing not much behind.
5.2.2 Over-sampled DCT

We fix \((M, N) = (100, 1500)\), and vary parameter \(F\) from 2 to 20, so the coherence of these matrices has a wider range and almost reaches 1 at the high end. In Fig. 7, when \(F\) is small, say \(F = 2, 4\), \(L_p - RLS\) still performs the best, similar to the case of Gaussian matrices. However, with increasing \(F\), the success rates for \(L_p - RLS\) decline quickly, worse than the Bregman \(l_1\) algorithm at \(F = 6, 10\). The performance for DCATL1 is very stable and maintains a high level consistently even at the very high end of coherence (\(F = 20\)).
Varying sparsity $k$ at $(M, N) = (50, 500)$.

Varying $N$ at $(M, k) = (10, 2)$.

Varying $N$ at $(M, k) = (20, 3)$.

Varying $N$ at $(M, k) = (50, 8)$.

Figure 8: Comparison of success rates of algorithms for $M$-sphere random matrices.

5.2.3 Uniformly Distributed M-sphere Random Matrices

We conducted two types of experiments for the $M$-sphere random matrices. In one, we fixed $(M, N)$ and vary sparsity $k$. In the other, we fixed $(M, k)$, and varied $N$. The results are shown in Figure 8. DCATL1 is consistently at the top.
6 Concluding Remarks

We have studied compressed sensing problem with the transformed $l_1$ penalty function for both the unconstrained and constrained models. We established a theory for the uniqueness and $l_0$ equivalence of global minimizer of the unconstrained model under RIP and analyzed properties of local minimizers. We presented two DC algorithms along with a convergence theory.

In numerical experiments, we observed that for incoherent Gaussian matrices, DCATl1 is on par with the best method reweighted $l_{1/2}$ $(L_p - RLS)$ in the unconstrained (constrained) model. For highly coherent over-sampled DCT matrices, DCATl1 is comparable to the best method DCA $l_1 - l_2$ algorithm. For random matrices of varied degree of coherence we tested (Gaussian, over-sampled DCT, uniform M-sphere), the DCATL1 algorithm is the most robust for constrained and unconstrained models alike.

In future work, we plan to develop TL1 algorithms for imaging processing applications such as deconvolution and deblurring.

Acknowledgment

The authors would like to thank Professor Wenjiang Fu of Michigan State University for suggesting reference [18] and helpful discussion.

References

[1] E. Candès, T. Tao, Decoding by linear programming, IEEE Trans. Info. Theory, 51(12):4203-4215, 2005.
[2] E. Candès, M. Rudelson, T. Tao, R. Vershynin, Error correction via linear programming, in 46th Annual IEEE Symposium on Foundations of Computer Science, pp. 668-681, 2005.
[3] E. Candès, J. Romberg, T. Tao, Robust uncertainty principles: Exact signal reconstruction from highly incomplete Fourier information, IEEE Trans. Info. Theory, 52(2), 489-509, 2006.
[4] E. Candès, J. Romberg, T. Tao, Stable signal recovery from incomplete and inaccurate measurements, Comm. Pure Applied Mathematics, 59(8):1207-1223, 2006.
[5] E. Candès, C. Fernandez-Granda, Super-resolution from noisy data, Journal of Fourier Analysis and Applications, 19(6):1229-1254, 2013.
[6] R. Chartrand, Nonconvex compressed sensing and error correction, ICASSP 2007, vol. 3, p. III 889.
[7] R. Chartrand, W. Yin, Iteratively reweighted algorithms for compressive sensing, ICASSP 2008, pp. 3869-3872.
[8] D. Donoho, *Compressed sensing*, IEEE Trans. Info. Theory, 52(4), 1289-1306, 2006.

[9] D. Donoho, M. Elad, *Optimally sparse representation in general (nonorthogonal) dictionaries via $l_1$ minimization*, Proc. Nat. Acad. Sci. USA, vol. 100, pp. 2197-2202, Mar. 2003.

[10] 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 Sciences, 6(2013), pp. 2010-2046.

[11] J. Fan, and R. Li, *Variable selection via nonconcave penalized likelihood and its oracle properties*, Journal of the American Statistical Association, 96(456):1348-1360, 2001.

[12] M-J Lai, Y. Xu, and W. Yin, *Improved Iteratively Reweighted Least Squares for Unconstrained Smoothed $L_q$ Minimization*, SIAM Journal on Numerical Analysis, 51(2):927-957, 2013.

[13] A. Fannjiang, W. Liao, *Coherence Pattern-Guided Compressive Sensing with Unresolved Grids*, SIAM J. Imaging Sciences, Vol. 5, No. 1, pp. 179–202, 2012.

[14] T. Goldstein and S. Osher, *The Split Bregman Method for $l_1$-regularized Problems*, SIAM Journal on Imaging Sciences, 2(1):323-343, 2009.

[15] Y. Lou, P. Yin, Q. He, and J. Xin, *Computing Sparse Representation in a Highly Coherent Dictionary Based on Difference of L1 and L2*, CAM Report 14-02, UCLA, 2014; J. Sci. Computing, to appear.

[16] S. Zhang and J. Xin, *Minimization of Transformed $L_1$ Penalty: A Thresholding Representation Theory and Fast Algorithms*, in preparation.

[17] Z. Lu and Y. Zhang, *Sparse approximation via penalty decomposition methods*, SIAM J. Optimization, 23(4):2448-2478, 2013.

[18] J. Lv, and Y. Fan, *A unified approach to model selection and sparse recovery using regularized least squares*, Annals of Statistics, 37(6A), pp. 3498-3528, September 2009.

[19] S. Mallat and Z. Zhang, *Matching pursuits with time-frequency dictionaries*, IEEE Trans. Signal Processing, 41(12):3397-3415, 1993.

[20] B. Natarajan, *Sparse approximate solutions to linear systems*, SIAM Journal on Computing, 24(2):227-234, 1995.

[21] D. Needell and R. Vershynin, *Signal recovery from incomplete and inaccurate measurements via regularized orthogonal matching pursuit*, IEEE Journal of Selected Topics in Signal Processing, 4(2):310-316, 2010.
[22] C.S. Ong, L.T.H. An, Learning sparse classifiers with difference of convex functions algorithms, Optimization Methods and Software, 28(4):830-854, 2013.

[23] P.D. Tao and L.T.H. An, Convex analysis approach to d.c. programming: Theory, algorithms and applications, Acta Mathematica Vietnamica, vol. 22, no. 1, pp. 289-355, 1997.

[24] P.D. Tao and L.T.H. An, A DC optimization algorithm for solving the trust-region subproblem, SIAM Journal on Optimization, 8(2), pp. 476–505, 1998.

[25] H.A.L. Thi, B.T.A. Thi, and H.M. Le, Sparse signal recovery by difference of convex functions algorithms, in Intelligent Information and Database Systems, pp. 387-397. Springer, 2013.

[26] R. Tibshirani, Regression shrinkage and selection via the lasso, J. Royal. Statist. Soc, 58(1):267-288, 1996.

[27] J. Tropp and A. Gilbert, Signal recovery from partial information via orthogonal matching pursuit, IEEE Trans. Inform. Theory, 53(12):4655-4666, 2007.

[28] J. Zeng, S. Lin, Y. Wang, and Z. Xu, $L_{1/2}$ regularization: Convergence of iterative half thresholding algorithm, Signal Processing, IEEE Transactions on, 62(9):2317-2329, 2014.

[29] F. Xu and S. Wang, A hybrid simulated annealing thresholding algorithm for compressed sensing, Signal Processing, 93:1577-1585, 2013.

[30] Z. Xu, X. Chang, F. Xu, and H. Zhang, $L_{1/2}$ regularization: A thresholding representation theory and a fast solver, Neural Networks and Learning Systems, IEEE Transactions on, 23(7):1013-1027, 2012.

[31] W. Cao, J. Sun, and Z. Xu, Fast image deconvolution using closed-form thresholding formulas of regularization, Journal of Visual Communication and Image Representation, 24(1):31-41, 2013.

[32] J. Yang and Y. Zhang, Alternating direction algorithms for $l_1$ problems in compressive sensing, SIAM Journal on Scientific Computing, 33(1):250-278, 2011.

[33] P. Yin, Y. Lou, Q. He, and J. Xin, Minimization of $L_1 - L_2$ for compressed sensing, CAM Report 14-01, UCLA, 2014.

[34] W. Yin, S. Osher, D. Goldfarb, and J. Darbon, Bregman iterative algorithms for $l_1$-minimization with applications to compressed sensing, SIAM Journal on Imaging Sciences, 1(1):143-168, 2008.