Square-Root Lasso With Nonconvex Regularization: An ADMM Approach

Xinyue Shen, Student Member, IEEE, Laming Chen, Student Member, IEEE, Yuantao Gu, Member, IEEE, and H. C. So, Fellow, IEEE

Abstract—Square-root least absolute shrinkage and selection operator (Lasso), a variant of Lasso, has recently been proposed with a key advantage that the optimal regularization parameter is independent of the noise level in the measurements. In this letter, we introduce a class of nonconvex sparsity-inducing penalties to the square-root Lasso to achieve better sparse recovery performance over the convex counterpart. The resultant formulation is converted to a nonconvex but multiconvex optimization problem, i.e., it is convex in each block of variables. Alternating direction method of multipliers is applied as the solver, according to which two efficient algorithms are devised for row-orthonormal sensing matrix and general sensing matrix, respectively. Numerical experiments are conducted to evaluate the performance of the proposed methods.

Index Terms—Alternating direction method of multipliers (ADMM), linearized ADMM, nonconvex regularization, sparse recovery, square-root penalty.

I. INTRODUCTION

S
PARSE recovery refers to extracting a sparse vector from a small number of noisy linear measurements [1]–[4]. Mathematically, the observation vector \( \mathbf{y} \in \mathbb{R}^N \) is modeled as

\[
\mathbf{y} = \mathbf{A}\mathbf{x}^* + \mathbf{e}
\]

(1)

where \( \mathbf{A} \in \mathbb{R}^{M \times N} \) with \( M < N \) is the sensing matrix, \( \mathbf{x}^* \in \mathbb{R}^N \) is the sparse vector with at most \( k \) (\( k \ll N \)) nonzero elements, and \( \mathbf{e} \) contains the additive zero-mean random noise components with unknown variance \( \sigma^2 \). The aim is to find \( \mathbf{x}^* \) given \( \mathbf{y} \) and \( \mathbf{A} \).

The least absolute shrinkage and selection operator (Lasso) [5] is a well-studied approach for sparse recovery with fast algorithms [6]–[8]. Its hinge lies in using \( \ell_1 \)-norm to induce sparsity of the vector. However, the noise level should be a prior for Lasso, in that the optimal regularization parameter, which is crucial to balance the sparsity-inducing and noise penalties, is directly determined by \( \sigma^2 \). Recently, the square-root Lasso is proposed with advantages that the optimal regularization parameter is independent of \( \sigma^2 \) [9], and that when the entries of \( \mathbf{A} \) and \( \mathbf{e} \) are i.i.d. Gaussian, the recovery error is precisely characterized, and the optimal regularization parameter can be analytically determined [10]. Efficient algorithmic methods have been developed [9] for the square-root Lasso via a conic programming problem formulation. Furthermore, other fast implementations are suggested [11]–[14] to improve the efficiency as well as scalability. It has been unveiled that, compared to convex relaxation with \( \ell_1 \) norm, a proper nonconvex regularization is able to achieve sparse recovery with fewer measurements and faster convergence, and is more robust against noise [15]–[20]. In this paper, we propose to combine the nonconvex sparsity-inducing penalty with the square-root error penalty to find the sparse vector. Despite the fact that the resultant formulation is nonconvex, we add a slack variable and an extra quadratic term so that the equivalent optimization problem is convex in each block of variables. Two algorithms based on alternating direction method of multipliers (ADMM) [21] are then devised as the solvers for row-orthonormal and general sensing matrices.

II. NONCONVEX REGULARIZED SQUARE-ROOT LASSO

Consider the square-root minimization problem regularized by a nonconvex function \( J(\cdot) \):

\[
\min_{\mathbf{x}} \lambda J(\mathbf{x}) + \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2
\]

(2)

in which \( \lambda > 0 \) is the regularization parameter while the sparsity-inducing penalty is defined as

\[
J(\mathbf{x}) = \sum_{i=1}^{N} F(x_i)
\]

(3)

where \( F(\cdot) \) satisfies the following definition [18].

Definition 1. The scalar function \( F : \mathbb{R} \rightarrow \mathbb{R}^+ \) satisfies

(a) \( F(0) = 0 \), \( F(\cdot) \) is even and not identically zero;
(b) \( F(\cdot) \) is nondecreasing on \( (0, +\infty) \);
(c) The function \( x \rightarrow F(x)/x \) is nonincreasing on \( (0, +\infty) \);
(d) \( F(\cdot) \) is weakly convex on \( (0, +\infty) \).

The concept of weak convexity is introduced in [22]. Basically, it allows us to define \( \beta < 0 \) as the largest quantity such that \( H(x) = F(x) - \beta x^2 \) is convex. According to [18, Lemma 1.1], there exists \( \alpha > 0 \) such that \( F(x)/x \rightarrow \alpha \) as \( x \rightarrow 0^+ \). With Definition 1 and (3), the nonconvexity of \( F(\cdot) \) and \( J(\cdot) \) can be defined as \( \zeta := -\beta/\alpha \) [18].

Functions satisfying Definition 1 are quite common in the literature, and concrete examples can be found in [18, Table I]. A specific example is

\[
F(x) = (|x| - \zeta^2)^21_{|x| \leq \zeta}(x) + \frac{1}{4\zeta^2}1_{|x| > \zeta}(x)
\]

(4)

where the indicator function \( 1_P(\cdot) \) equals 1 if the argument satisfies \( P \), and equals 0 otherwise. \( F(\cdot) \) in (4) is a continuous
piecewise quadratic function, and it is easy to check that $F(\cdot)$ satisfies Definition 1 with $\beta = -\zeta$ and $\alpha = 1$.

When $J(\cdot)$ is replaced by the $\ell_1$ norm, problem (2) becomes the square-root Lasso [9]. Its key advantage over the standard Lasso is that the optimal value of $\lambda$ does not depend on $\sigma^2$, so we do not have to estimate the noise level prior to solving the problem. Similarly, it is expected that the optimal $\lambda$ in (2) is also independent of $\sigma^2$. Moreover, with a proper nonconvex penalty $J(\cdot)$, the sparsity pattern can be better induced than the $\ell_1$ norm penalty counterpart.

To solve (2), we rewrite it as

$$
\min_{\mathbf{x}, \mathbf{z}} \lambda J(\mathbf{x}) + \|\mathbf{z}\|_2 \quad \text{s.t.} \; \mathbf{A} \mathbf{x} - \mathbf{z} = \mathbf{y}. \tag{5}
$$

Nevertheless, the objective function in problem (5) is nonconvex with respect to $\mathbf{x}$. To address this issue, by introducing a slack variable $\mathbf{w} \in \mathbb{R}^{M+N}$ and adding a quadratic term, we devise an equivalent form of (5) as

$$
\min_{\mathbf{x}, \mathbf{z}, \mathbf{w}} \lambda J(\mathbf{x}) + \|\mathbf{z}\|_2 + \frac{\mu}{2} \|\mathbf{x}^T \mathbf{z}^T - \mathbf{w}\|_2^2 \quad \text{s.t.} \; \|\mathbf{A} - \mathbf{I}\| \mathbf{w} = \mathbf{y}, \; \|\mathbf{x}^T \mathbf{z}^T\| = \mathbf{w}. \tag{6}
$$

Since $J(\cdot)$ is weakly convex, problem (6) is convex with respect to $\mathbf{x}, \mathbf{z},$ and $\mathbf{w}$ separately when $\zeta \leq \mu/(2\lambda \alpha)$. In doing so, it can be solved by ADMM where each iterative step corresponds to a convex optimization.

### III. Algorithm Development

In this section, two algorithms based on ADMM are developed to solve (6), where the first one assumes that $\mathbf{A}$ is a row-orthonormal sensing matrix, and the second one considers a general sensing matrix.

#### A. Row-Orthonormal Sensing Matrix

Concrete examples of $\mathbf{A}$ with orthonormal rows include partial Fourier sensing matrix, partial two-dimensional (2D) DFT matrix, and partial Haar wavelet transform, which have applications in magnetic resonance imaging [1] and 2D tomographic reconstruction [23].

Putting the first constraint in (6) into the cost function yields the following equivalent problem:

$$
\min_{\mathbf{x}, \mathbf{z}, \mathbf{w}} \lambda J(\mathbf{x}) + \|\mathbf{z}\|_2 + \frac{\mu}{2} \|\mathbf{x}^T \mathbf{z}^T\| - \mathbf{w}\|_2^2 + g(\mathbf{w}) \\
\text{s.t.} \; \|\mathbf{x}^T \mathbf{z}^T\| = \mathbf{w} \tag{7}
$$

where $g(\mathbf{w})$ equals $0$ if $\|\mathbf{A} - \mathbf{I}\| \mathbf{w} = \mathbf{y}$ holds, and equals positive infinity otherwise.

Now we proceed to unveil each step of the ADMM for problem (7). The augmented Lagrangian of (7) is

$$
L(\mathbf{x}, \mathbf{z}, \mathbf{w}, \gamma) = \lambda J(\mathbf{x}) + \|\mathbf{z}\|_2 + \frac{\mu}{2} \|\mathbf{x}^T \mathbf{z}^T\| - \mathbf{w}\|_2^2 + g(\mathbf{w}) + \gamma^T \left(\|\mathbf{x}^T \mathbf{z}^T\| - \mathbf{w}\|_2^2\right)
$$

where $\gamma \in \mathbb{R}^{M+N}$ is the dual variable vector and $\rho > 0$ is the penalty parameter. Denote $\mathbf{w}_l^T = [\mathbf{w}_1^T \mathbf{w}_2^T]$ and $\gamma^T = [\gamma_1^T \gamma_2^T]$, where $\mathbf{w}_1, \gamma_1 \in \mathbb{R}^N$ and $\mathbf{w}_2, \gamma_2 \in \mathbb{R}^M$.

In the $(t+1)$th iteration, the update of $\mathbf{x}$ is

$$
\mathbf{x}^{t+1} = \arg \min_{\mathbf{x}} L(\mathbf{x}, \mathbf{z}^t, \mathbf{w}^t, \gamma^t) = \text{prox}_{\frac{1}{\mu + \rho} J(\cdot)} \left(\mathbf{w}_1^t - \frac{\gamma_1^t}{\mu + \rho}\right). \tag{8}
$$

Since $J(\cdot)$ admits a coordinatewise decomposition as (3), the proximal operator [24] in (8) can be evaluated in parallel for each coordinate. Furthermore, for some specific $J(\cdot)$, their proximal operators have closed-form solutions. For instance, for $F(\cdot)$ in (4), when $\zeta < 1/(2\lambda)$, its proximal operator is [20]

$$
\text{prox}_{\frac{1}{\mu + \rho} J(\cdot)}(v) = \frac{v - \text{sign}(v)}{1 - 2\zeta} \mathbf{1}_{\|v\|_2 \leq \frac{\|v\|_2}{\lambda \alpha}}(v) + v \mathbf{1}_{\|v\|_2 > \frac{\|v\|_2}{\lambda \alpha}}(v). \tag{9}
$$

The update of $\mathbf{w}$ is

$$
\mathbf{w}^{t+1} = \arg \min_{\mathbf{w}} L(\mathbf{x}^{t+1}, \mathbf{z}^t, \mathbf{w}, \gamma^t) = \Pi_{\mathcal{C}^\perp}(\mathbf{v}^t) = \left(\|\mathbf{A}^T \mathbf{y}\| + \frac{\|\mathbf{v}^t\|_2}{\zeta}\right)^{-1} \mathbf{A}^T \mathbf{y}.
$$

#### B. General Sensing Matrix

When the rows of $\mathbf{A}$ are not orthonormal, the pseudoinverse of $[\mathbf{A} - \mathbf{I}]$ does not result in a computationally efficient calculation. In this section, we propose to efficiently solve (6) with a general $\mathbf{A}$ using the linearized ADMM [25], [26].
**Algorithm 1:** Nonconvex regularized square-root Lasso with row-orthonormal sensing matrix

**Require:** Row-orthonormal $A, y$, $\mu > 0$, $\rho > 0$, $\varepsilon > 0$, $T_M$;
1: Initialize: $t = 0, w^0, \gamma_1^0, \gamma_2^0, r^0_p = +\infty, r^0_d = +\infty$;
2: while $\max(r^t_p, r^t_d) \geq \varepsilon$ and $t < T_M$ do
3: Update $x^{t+1}$ according to (8);
4: Update $z^{t+1}$ according to (10);
5: Update $w^{t+1}$ according to (12);
6: Update $\gamma_1^{t+1}$ according to (13);
7: Update $r^{t+1}_p$ and $r^{t+1}_d$ according to (14);
8: $t = t + 1$;
9: end while

The augmented Lagrangian of problem (6) is
\[
L(x, z, w, \gamma_1, \gamma_2) = \lambda J(x) + \|z\|_2 + \frac{\mu}{2} \left\| x^T z^T \right\|_2^2 + \frac{\rho}{2} \left\| \gamma_1 - x \right\|_2^2 + \frac{\rho}{2} \left\| \gamma_2 - w \right\|_2^2
\]
where $\gamma_1 \in \mathbb{R}^M$ and $\gamma_2 \in \mathbb{R}^{M-N}$ are dual variable vectors. Denote $w^T = [w_1^T, w_2^T]$ and $\gamma_2 = \gamma_2^1 \gamma_2^2$, where $w_1, \gamma_2^1 \in \mathbb{R}^N$ and $w_2, \gamma_2^2 \in \mathbb{R}^M$.

In the $(t+1)$th iteration, the updates of $x$ and $z$ are
\[
x^{t+1} = \arg \min_x L(x, z^t, w^t, \gamma_1^t, \gamma_2^t)
\]
\[
z^{t+1} = \arg \min_z L(x^{t+1}, z, w^t, \gamma_1^t, \gamma_2^t)
\]

The update of $w$ is
\[
w^{t+1} = \arg \min_w L(x^{t+1}, z^{t+1}, w, \gamma_1^t, \gamma_2^t)
\]
\[
\text{(17)}
\]

**Algorithm 2:** Nonconvex regularized square-root Lasso with general sensing matrix

**Require:** $A, y$, $\mu > 0$, $\rho > 0$, $\delta > 0$, $\varepsilon > 0$, $T_M$;
1: Initialize: $t = 0, w^0, \gamma_1^0, \gamma_2^0, r^0_p = +\infty, r^0_d = +\infty$;
2: while $\max(r^t_p, r^t_d) \geq \varepsilon$ and $t < T_M$ do
3: Update $x^{t+1}$ according to (15);
4: Update $z^{t+1}$ according to (16);
5: Update $w^{t+1}$ according to (18);
6: Update $\gamma_1^{t+1}$ and $\gamma_2^{t+1}$ according to (19);
7: Update $r^{t+1}_p$ and $r^{t+1}_d$ according to (20);
8: $t = t + 1$;
9: end while

**TABLE I**

<table>
<thead>
<tr>
<th>Method</th>
<th>Size</th>
<th>$N = 2^{12}$</th>
<th>$M = 2^{10}$</th>
<th>$N = 2^{16}$</th>
<th>$M = 2^{14}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>CVX [27]</td>
<td>$k = 256$</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>ADMM [14]</td>
<td>3.90s</td>
<td>—</td>
<td>58.41s</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Algorithm 1</td>
<td>0.82s</td>
<td>1.82s</td>
<td>25.98s</td>
<td>66.15s</td>
<td></td>
</tr>
<tr>
<td>Algorithm 2</td>
<td>4.46s</td>
<td>9.62s</td>
<td>130.40s</td>
<td>320.54s</td>
<td></td>
</tr>
</tbody>
</table>

of which the computational complexity is now dominated by matrix-vector multiplications.

The updates of the dual variable vectors are
\[
\gamma_1^{t+1} = \gamma_1^t + \rho (A - I) z^{t+1} - y
\]
\[
\gamma_2^{t+1} = \gamma_2^t + \rho \left( \left( x^{t+1} \right)^T (z^{t+1})^T - w^{t+1} \right).
\]

The primal and dual residuals are calculated as
\[
r^{t+1}_p = \max \left\{ \|x^{t+1}\|_2, \|z^{t+1}\|_2 \right\}
\]
\[
r^{t+1}_d = \max \left\{ \|\mu z^{t+1} + (\mu + \rho) r^{t+1}_d\|_2 \right\}
\]
\[
\|x^{t+1} - \rho (I/\delta - [A - I] \gamma_1^{t+1}) r^{t+1}_d\|_2
\]
\[
\|x^{t+1} - \rho (I/\delta - [A - I] \gamma_1^{t+1}) r^{t+1}_d\|_2
\]
\[
\|x^{t+1} - \rho (I/\delta - [A - I] \gamma_1^{t+1}) r^{t+1}_d\|_2
\]
\[
\|x^{t+1} - \rho (I/\delta - [A - I] \gamma_1^{t+1}) r^{t+1}_d\|_2
\]
\[
\|x^{t+1} - \rho (I/\delta - [A - I] \gamma_1^{t+1}) r^{t+1}_d\|_2
\]
\[
\|x^{t+1} - \rho (I/\delta - [A - I] \gamma_1^{t+1}) r^{t+1}_d\|_2
\]
\[
\|x^{t+1} - \rho (I/\delta - [A - I] \gamma_1^{t+1}) r^{t+1}_d\|_2
\]

where
\[
r^{t+1}_p = \left[ A - I \right] w^{t+1} - y
\]
\[
r^{t+1}_d = \left[ (x^{t+1})^T (z^{t+1})^T - w^{t+1} \right]
\]
\[
r^{t+1}_d = \left[ (x^{t+1})^T (z^{t+1})^T - w^{t+1} \right]
\]
\[
r^{t+1}_d = \left[ (x^{t+1})^T (z^{t+1})^T - w^{t+1} \right]
\]
\[
r^{t+1}_d = \left[ (x^{t+1})^T (z^{t+1})^T - w^{t+1} \right]
\]

The algorithm is summarized as Algorithm 2.

**IV. NUMERICAL EXAMPLES**

In this section, the nonconvex sparsity-inducing penalty corresponds to (3) with (4), and its convex counterpart refers to the $\ell_1$ norm. The nonzero entries of the sparse vector are uniformly located among all possible choices, and their values follow the standard normal distribution.
It is observed that \( k = 2 \) and \( \lambda \) when \( M = 10 \). The conclusion is that \( 14 = 10 \), and this result is shown in Fig. 2. Similar to square-root Lasso, the error with the nonconvex penalty is smaller than that with the convex one.

In the second experiment, we examine the denoising performance for different choices of \( \lambda \) with convex and nonconvex sparsity-inducing penalties when \( k = 50 \) and \( \zeta = 10^{-0.3} \), and the result is shown in Fig. 2. Similar to square-root Lasso, the optimal choice of \( \lambda \) in problem (2) is also independent of the noise level, namely, \( \sigma = 10^{-2}, 10^{-3} \), and \( 10^{-4} \). In addition, it is evident that the error with the nonconvex penalty is smaller than that with the convex one.

B. General Sensing Matrix

The performance of Algorithm 2 is evaluated here. The sensing matrix is an i.i.d. Gaussian random matrix where \( N = 2^{10} \) and \( M = 2^8 \). The two tests in Section IV-A are repeated with \( \mu = 100 \lambda \zeta, \rho = 4, \), \( \delta = 1/\|A\|_2, \epsilon = 10^{-5} \), and \( T_M = 4 \times 10^4 \). The reason of using a larger \( \mu \) is to speed up its convergence. The results are shown in Figs. 3 and 4, and similar findings are obtained.
REFERENCES


