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

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

$$\mathbf{y} = \mathbf{A}\mathbf{x}^* + \mathbf{e} \tag{1}$$

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 σ^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 ℓ_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 σ^2. Recently, the square-root Lasso is proposed with advantages that the optimal regularization parameter is independent of σ^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 ℓ_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 $J(\cdot)$:

$$\min_{\mathbf{x}} \ \lambda J(\mathbf{x}) + \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2 \tag{2}$$

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

$$J(\mathbf{x}) = \sum_{i=1}^{N} F(x_i) \tag{3}$$

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

Definition 1. The scalar function $F : \mathbb{R} \to \mathbb{R}^+$ satisfies

(a) $F(0) = 0$, $F(\cdot)$ is even and not identically zero;
(b) $F(\cdot)$ is nondecreasing on $[0, +\infty)$;
(c) The function $x \to F(x)/x$ 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 $F(x)/x \to \alpha$ as $x \to 0^+$. 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

$$F(x) = (|x| - \zeta x^2)\mathbf{1}_{|x|\le \frac{1}{2\zeta}}(x) + \frac{1}{4\zeta}\mathbf{1}_{|x|> \frac{1}{2\zeta}}(x) \tag{4}$$

where the indicator function $\mathbf{1}_P(\cdot)$ equals 1 if the argument satisfies P, and equals 0 otherwise. $F(\cdot)$ in (4) is a continuous

Manuscript received February 13, 2016; revised May 10, 2016; accepted May 10, 2016. Date of publication May 12, 2016; date of current version June 6, 2016. This work was supported by a grant from the NSFC/RGC Joint Research Scheme sponsored by the National Natural Science Foundation of China and the Research Grants Council of Hong Kong (Project 61531166005, N_CityU 104/15). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Gitta Kutyniok. *(Corresponding author: Yuantao Gu).*

X. Shen, L. Chen, and Y. Gu are with the Department of Electronic Engineering and Tsinghua National Laboratory for Information Science and Technology, Tsinghua University, Beijing 100084, China (e-mail: shenxy13@mails.tsinghua.edu.cn; chen-lm06@mails.tsinghua.edu.cn; gyt@tsinghua.edu.cn).

H. C. So is with the Department of Electronic Engineering, City University of Hong Kong, Hong Kong (e-mail: hcso@ee.cityu.edu.hk).

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2016.2567482

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 ℓ_1 norm, problem (2) becomes the square-root Lasso [9]. Its key advantage over the standard Lasso is that the optimal value of λ does not depend on σ^2, so we do not have to estimate the noise level prior to solving the problem. Similarly, it is expected that the optimal λ in (2) is also independent of σ^2. Moreover, with a proper nonconvex penalty $J(\cdot)$, the sparsity pattern can be better induced than the ℓ_1 norm penalty counterpart.

To solve (2), we rewrite it as

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

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

$$\begin{aligned} &\min_{\mathbf{x},\mathbf{z},\mathbf{w}} \ \lambda J(\mathbf{x}) + \|\mathbf{z}\|_2 + \frac{\mu}{2}\left\| \left[\mathbf{x}^{\mathrm{T}} \ \mathbf{z}^{\mathrm{T}}\right]^{\mathrm{T}} - \mathbf{w} \right\|_2^2 \\ &\quad \text{s.t. } [\mathbf{A} \ -\mathbf{I}]\mathbf{w} = \mathbf{y}, \ \left[\mathbf{x}^{\mathrm{T}} \ \mathbf{z}^{\mathrm{T}}\right]^{\mathrm{T}} = \mathbf{w}. \end{aligned} \tag{6}$$

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:

$$\begin{aligned} &\min_{\mathbf{x},\mathbf{z},\mathbf{w}} \ \lambda J(\mathbf{x}) + \|\mathbf{z}\|_2 + \frac{\mu}{2}\left\| \left[\mathbf{x}^{\mathrm{T}} \ \mathbf{z}^{\mathrm{T}}\right]^{\mathrm{T}} - \mathbf{w} \right\|_2^2 + g(\mathbf{w}) \\ &\quad \text{s.t. } \left[\mathbf{x}^{\mathrm{T}} \ \mathbf{z}^{\mathrm{T}}\right]^{\mathrm{T}} = \mathbf{w} \end{aligned} \tag{7}$$

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

$$\begin{aligned} &L(\mathbf{x},\mathbf{z},\mathbf{w},\boldsymbol{\gamma}) \\ &= \lambda J(\mathbf{x}) + \|\mathbf{z}\|_2 + \frac{\mu}{2}\left\| \left[\mathbf{x}^{\mathrm{T}} \ \mathbf{z}^{\mathrm{T}}\right]^{\mathrm{T}} - \mathbf{w} \right\|_2^2 + g(\mathbf{w}) \\ &\quad + \boldsymbol{\gamma}^{\mathrm{T}}\left(\left[\mathbf{x}^{\mathrm{T}} \ \mathbf{z}^{\mathrm{T}}\right]^{\mathrm{T}} - \mathbf{w} \right) + \frac{\rho}{2}\left\| \left[\mathbf{x}^{\mathrm{T}} \ \mathbf{z}^{\mathrm{T}}\right]^{\mathrm{T}} - \mathbf{w} \right\|_2^2 \end{aligned}$$

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

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

$$\begin{aligned} \mathbf{x}^{t+1} &= \arg\min_{\mathbf{x}} \ L(\mathbf{x},\mathbf{z}^t,\mathbf{w}^t,\boldsymbol{\gamma}^t) \\ &= \operatorname{prox}_{\frac{\lambda}{\mu+\rho}J(\cdot)}\left(\mathbf{w}_1^t - \frac{\boldsymbol{\gamma}_1^t}{\mu+\rho} \right). \end{aligned} \tag{8}$$

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 $\zeta < 1/(2\epsilon)$, its proximal operator is [20]

$$\operatorname{prox}_{\epsilon F}(v) = \frac{v - \epsilon \operatorname{sign}(v)}{1 - 2\epsilon\zeta}\mathbf{1}_{\epsilon \leq |v| \leq \frac{1}{2\zeta}}(v) + v\mathbf{1}_{|v| > \frac{1}{2\zeta}}(v). \tag{9}$$

The update of $\mathbf{z}$ is

$$\begin{aligned} \mathbf{z}^{t+1} &= \arg\min_{\mathbf{z}} \ L(\mathbf{x}^{t+1},\mathbf{z},\mathbf{w}^t,\boldsymbol{\gamma}^t) \\ &= \operatorname{prox}_{\frac{1}{\mu+\rho}\|\cdot\|_2}\left(\mathbf{w}_2^t - \frac{\boldsymbol{\gamma}_2^t}{\mu+\rho} \right) \end{aligned} \tag{10}$$

in which the proximal operator of ℓ_2 norm is known to be [24]

$$\operatorname{prox}_{\epsilon\|\cdot\|_2}(\mathbf{v}) = \mathbf{1}_{\|\mathbf{v}\|_2 \geq \epsilon}(\mathbf{v})\left(1 - \frac{\epsilon}{\|\mathbf{v}\|_2}\right)\mathbf{v}. \tag{11}$$

The update of $\mathbf{w}$ is

$$\begin{aligned} \mathbf{w}^{t+1} &= \arg\min_{\mathbf{w}} \ L(\mathbf{x}^{t+1},\mathbf{z}^{t+1},\mathbf{w},\boldsymbol{\gamma}^t) \\ &= \Pi_{[\mathbf{A} \ -\mathbf{I}]\mathbf{w}=\mathbf{y}}\left(\left[\left(\mathbf{x}^{t+1}\right)^{\mathrm{T}} \ \left(\mathbf{z}^{t+1}\right)^{\mathrm{T}} \right]^{\mathrm{T}} + \frac{\boldsymbol{\gamma}^t}{\mu+\rho} \right) \end{aligned}$$

where $\Pi_{\mathcal{C}}(\cdot)$ denotes the Euclidean projection onto $\mathcal{C}$. Define

$$\mathbf{v}^{t+1} := \left[\left(\mathbf{x}^{t+1}\right)^{\mathrm{T}} \ \left(\mathbf{z}^{t+1}\right)^{\mathrm{T}} \right]^{\mathrm{T}} + \frac{\boldsymbol{\gamma}^t}{\mu+\rho}$$

then $\mathbf{w}^{t+1}$ is computed as

$$\mathbf{w}^{t+1} = \mathbf{v}^{t+1} - [\mathbf{A} \ -\mathbf{I}]^{\dagger}\left([\mathbf{A} \ -\mathbf{I}]\mathbf{v}^{t+1} - \mathbf{y}\right). \tag{12}$$

When $\mathbf{A}$ is row-orthonormal, $[\mathbf{A} \ -\mathbf{I}]^{\dagger} = \frac{1}{2}[\mathbf{A} \ -\mathbf{I}]^{\mathrm{T}}$, and thus (12) can be evaluated efficiently.

The update of the dual variable is

$$\boldsymbol{\gamma}^{t+1} = \boldsymbol{\gamma}^t + \rho\left(\left[\left(\mathbf{x}^{t+1}\right)^{\mathrm{T}} \ \left(\mathbf{z}^{t+1}\right)^{\mathrm{T}} \right]^{\mathrm{T}} - \mathbf{w}^{t+1} \right). \tag{13}$$

The stopping criterion of ADMM is that the primal and dual residuals must be small, and for problem (7) the quantities are:

$$r_{\mathrm{p}}^{t+1} = \left\|\mathbf{r}_{\mathrm{p}}^{t+1}\right\|_2, \ \ r_{\mathrm{d}}^{t+1} = \left\|\mu\mathbf{r}_{\mathrm{p}}^{t+1} + (\mu+\rho)\mathbf{r}_{\mathrm{d}}^{t+1}\right\|_2 \tag{14}$$

where

$$\mathbf{r}_{\mathrm{p}}^{t+1} = \left[\left(\mathbf{x}^{t+1}\right)^{\mathrm{T}} \ \left(\mathbf{z}^{t+1}\right)^{\mathrm{T}} \right]^{\mathrm{T}} - \mathbf{w}^{t+1}, \ \ \mathbf{r}_{\mathrm{d}}^{t+1} = \mathbf{w}^{t+1} - \mathbf{w}^t.$$

The algorithm is summarized as Algorithm 1.

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 $\mathbf{A}$, $\mathbf{y}$, $\mu > 0$, $\rho > 0$, $\varepsilon > 0$, $T_M$;
Initialize: $t = 0$, $\mathbf{w}^0$, $\boldsymbol{\gamma}^0$, $r_\mathrm{p}^0 = +\infty$, $r_\mathrm{d}^0 = +\infty$;
while $\max(r_\mathrm{p}^t, r_\mathrm{d}^t) \geq \varepsilon$ and $t < T_M$ do
    Update $\mathbf{x}^{t+1}$ according to (8);
    Update $\mathbf{z}^{t+1}$ according to (10);
    Update $\mathbf{w}^{t+1}$ according to (12);
    Update $\boldsymbol{\gamma}^{t+1}$ according to (13);
    Update $r_\mathrm{p}^{t+1}$ and $r_\mathrm{d}^{t+1}$ according to (14);
    $t = t + 1$;
end while
```

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

```
Require: $\mathbf{A}$, $\mathbf{y}$, $\mu > 0$, $\rho > 0$, $\delta > 0$, $\varepsilon > 0$, $T_M$;
Initialize: $t = 0$, $\mathbf{w}^0$, $\boldsymbol{\gamma}_1^0$, $\boldsymbol{\gamma}_2^0$, $r_\mathrm{p}^0 = +\infty$, $r_\mathrm{d}^0 = +\infty$;
while $\max(r_\mathrm{p}^t, r_\mathrm{d}^t) \geq \varepsilon$ and $t < T_M$ do
    Update $\mathbf{x}^{t+1}$ according to (15);
    Update $\mathbf{z}^{t+1}$ according to (16);
    Update $\mathbf{w}^{t+1}$ according to (18);
    Update $\boldsymbol{\gamma}_1^{t+1}$ and $\boldsymbol{\gamma}_2^{t+1}$ according to (19);
    Update $r_\mathrm{p}^{t+1}$ and $r_\mathrm{d}^{t+1}$ according to (20);
    $t = t + 1$;
end while
```

TABLE I
COMPARISON OF CPU RUNNING TIME OF DIFFERENT ALGORITHMS

Size / Method	$N = 2^{12}$, $M = 2^{10}$		$N = 2^{16}$, $M = 2^{14}$	
	$k = 256$	$k = 384$	$k = 4096$	$k = 5632$
CVX [27]	272.10s	—		
ADMM [14]	3.90s	—	58.41s	—
Algorithm 1	0.82s	1.82s	25.98s	66.15s
Algorithm 2	4.46s	9.62s	130.40s	320.54s

The augmented Lagrangian of problem (6) is

$$\begin{aligned} L(\mathbf{x}, \mathbf{z}, \mathbf{w}, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2) &= \lambda J(\mathbf{x}) + \|\mathbf{z}\|_2 + \frac{\mu}{2}\left\|\left[\mathbf{x}^\mathrm{T}\ \mathbf{z}^\mathrm{T}\right]^\mathrm{T} - \mathbf{w}\right\|_2^2 \\ &+ \boldsymbol{\gamma}_1^\mathrm{T}\left([\mathbf{A}\ \ -\mathbf{I}]\mathbf{w} - \mathbf{y}\right) + \frac{\rho}{2}\|[\mathbf{A}\ \ -\mathbf{I}]\mathbf{w} - \mathbf{y}\|_2^2 \\ &+ \boldsymbol{\gamma}_2^\mathrm{T}\left(\left[\mathbf{x}^\mathrm{T}\ \mathbf{z}^\mathrm{T}\right]^\mathrm{T} - \mathbf{w}\right) + \frac{\rho}{2}\left\|\left[\mathbf{x}^\mathrm{T}\ \mathbf{z}^\mathrm{T}\right]^\mathrm{T} - \mathbf{w}\right\|_2^2 \end{aligned}$$

in which $\boldsymbol{\gamma}_1 \in \mathbb{R}^M$ and $\boldsymbol{\gamma}_2 \in \mathbb{R}^{M+N}$ are dual variable vectors and $\rho > 0$ is the penalty parameter, respectively. Denote $\mathbf{w}^\mathrm{T} = [\mathbf{w}_1^\mathrm{T}\ \ \mathbf{w}_2^\mathrm{T}]$ and $\boldsymbol{\gamma}_2^\mathrm{T} = [\boldsymbol{\gamma}_{21}^\mathrm{T}\ \ \boldsymbol{\gamma}_{22}^\mathrm{T}]$, where $\mathbf{w}_1, \boldsymbol{\gamma}_{21} \in \mathbb{R}^N$ and $\mathbf{w}_2, \boldsymbol{\gamma}_{22} \in \mathbb{R}^M$.

In the $(t+1)$th iteration, the updates of $\mathbf{x}$ and $\mathbf{z}$ are

$$\begin{aligned} \mathbf{x}^{t+1} &= \arg\min_{\mathbf{x}}\ L(\mathbf{x}, \mathbf{z}^t, \mathbf{w}^t, \boldsymbol{\gamma}_1^t, \boldsymbol{\gamma}_2^t) \\ &= \operatorname{prox}_{\frac{\lambda}{\mu+\rho}J(\cdot)}\left(\mathbf{w}_1^t - \frac{\boldsymbol{\gamma}_{21}^t}{\mu+\rho}\right) \end{aligned} \tag{15}$$

and

$$\begin{aligned} \mathbf{z}^{t+1} &= \arg\min_{\mathbf{z}}\ L(\mathbf{x}^{t+1}, \mathbf{z}, \mathbf{w}^t, \boldsymbol{\gamma}_1^t, \boldsymbol{\gamma}_2^t) \\ &= \operatorname{prox}_{\frac{1}{\mu+\rho}\|\cdot\|_2}\left(\mathbf{w}_2^t - \frac{\boldsymbol{\gamma}_{22}^t}{\mu+\rho}\right). \end{aligned} \tag{16}$$

While the update of $\mathbf{w}$ is

$$\begin{aligned} \mathbf{w}^{t+1} =&\ \arg\min_{\mathbf{w}}\ L(\mathbf{x}^{t+1}, \mathbf{z}^{t+1}, \mathbf{w}, \boldsymbol{\gamma}_1^t, \boldsymbol{\gamma}_2^t) \\ =&\ \arg\min_{\mathbf{w}}\ \frac{\mu+\rho}{2}\left\|\mathbf{w} - \left[\left(\mathbf{x}^{t+1}\right)^\mathrm{T}\ \left(\mathbf{z}^{t+1}\right)^\mathrm{T}\right]^\mathrm{T}\right\|_2^2 \\ &- \mathbf{w}^\mathrm{T}\boldsymbol{\gamma}_2^t + \mathbf{w}^\mathrm{T}[\mathbf{A}\ \ -\mathbf{I}]^\mathrm{T}\left(\boldsymbol{\gamma}_1^t - \rho\mathbf{y}\right) \\ &+ \frac{\rho}{2}\|[\mathbf{A}\ \ -\mathbf{I}]\mathbf{w}\|_2^2. \end{aligned} \tag{17}$$

Problem (17) does admit a closed-form solution, but it is not computationally efficient for large-scale problems. Thus, we linearize the last term $\|[\mathbf{A}\ \ -\mathbf{I}]\mathbf{w}\|_2^2$ at point $\mathbf{w}^t$ with $0 < \delta \leq 1/\|\mathbf{A}\mathbf{A}^\mathrm{T} + \mathbf{I}\|_2$, and (17) is modified as

$$\begin{aligned} \mathbf{w}^{t+1} =&\ \arg\min_{\mathbf{w}}\ \frac{\mu+\rho}{2}\left\|\mathbf{w} - \left[\left(\mathbf{x}^{t+1}\right)^\mathrm{T}\ \left(\mathbf{z}^{t+1}\right)^\mathrm{T}\right]^\mathrm{T}\right\|_2^2 \\ &- \mathbf{w}^\mathrm{T}\boldsymbol{\gamma}_2^t + \mathbf{w}^\mathrm{T}[\mathbf{A}\ \ -\mathbf{I}]^\mathrm{T}\left(\boldsymbol{\gamma}_1^t - \rho\mathbf{y}\right) \\ &+ \frac{\rho}{2\delta}\left\|\mathbf{w} - \mathbf{w}^t\right\|_2^2 + \rho\mathbf{w}^\mathrm{T}[\mathbf{A}\ \ -\mathbf{I}]^\mathrm{T}[\mathbf{A}\ \ -\mathbf{I}]\mathbf{w}^t \\ =&\ \frac{1}{\mu+\rho+\rho/\delta}\left[(\mu+\rho)\left[\left(\mathbf{x}^{t+1}\right)^\mathrm{T}\ \left(\mathbf{z}^{t+1}\right)^\mathrm{T}\right]^\mathrm{T} + \boldsymbol{\gamma}_2^t\right. \\ &\left. + \frac{\rho}{\delta}\mathbf{w}^t - [\mathbf{A}\ \ -\mathbf{I}]^\mathrm{T}\left(\boldsymbol{\gamma}_1^t - \rho\mathbf{y} + \rho[\mathbf{A}\ \ -\mathbf{I}]\mathbf{w}^t\right)\right] \end{aligned} \tag{18}$$

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

The updates of the dual variable vectors are

$$\begin{aligned} \boldsymbol{\gamma}_1^{t+1} =&\ \boldsymbol{\gamma}_1^t + \rho\left([\mathbf{A}\ \ -\mathbf{I}]\mathbf{w}^{t+1} - \mathbf{y}\right) \\ \boldsymbol{\gamma}_2^{t+1} =&\ \boldsymbol{\gamma}_2^t + \rho\left(\left[\left(\mathbf{x}^{t+1}\right)^\mathrm{T}\ \left(\mathbf{z}^{t+1}\right)^\mathrm{T}\right]^\mathrm{T} - \mathbf{w}^{t+1}\right). \end{aligned} \tag{19}$$

The primal and dual residuals are calculated as

$$\begin{aligned} r_\mathrm{p}^{t+1} &= \max\left\{\left\|\mathbf{r}_{\mathrm{p}1}^{t+1}\right\|_2, \left\|\mathbf{r}_{\mathrm{p}2}^{t+1}\right\|_2\right\} \\ r_\mathrm{d}^{t+1} &= \max\left\{\ \left\|\mu\mathbf{r}_{\mathrm{p}2}^{t+1} + (\mu+\rho)\mathbf{r}_\mathrm{d}^{t+1}\right\|_2\right. \\ &\qquad \left.\left\|\mu\mathbf{r}_{\mathrm{p}2}^{t+1} - \rho\left(\mathbf{I}/\delta - [\mathbf{A}\ \ -\mathbf{I}]^\mathrm{T}[\mathbf{A}\ \ -\mathbf{I}]\right)\mathbf{r}_\mathrm{d}^{t+1}\right\|_2\right\} \end{aligned} \tag{20}$$

where

$$\begin{aligned} \mathbf{r}_{\mathrm{p}1}^{t+1} &= [\mathbf{A}\ \ -\mathbf{I}]\mathbf{w}^{t+1} - \mathbf{y} \\ \mathbf{r}_{\mathrm{p}2}^{t+1} &= \left[\left(\mathbf{x}^{t+1}\right)^\mathrm{T}\ \left(\mathbf{z}^{t+1}\right)^\mathrm{T}\right]^\mathrm{T} - \mathbf{w}^{t+1} \\ \mathbf{r}_\mathrm{d}^{t+1} &= \mathbf{w}^{t+1} - \mathbf{w}^t. \end{aligned} \tag{21}$$

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 ℓ_1 norm. The nonzero entries of the sparse vector are uniformly located among all possible choices, and their values follow the standard normal distribution.

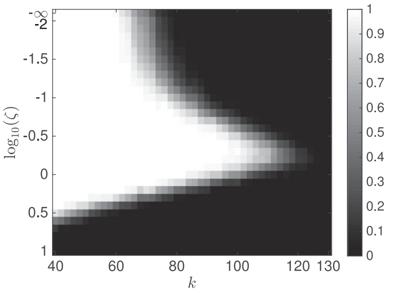

Fig. 1. Recovery probability of Algorithm 1 versus k and ζ when $N = 2^{10}$, $M = 2^8$, and $\lambda = 0.01$. A total of 200 trials are repeated for each point.

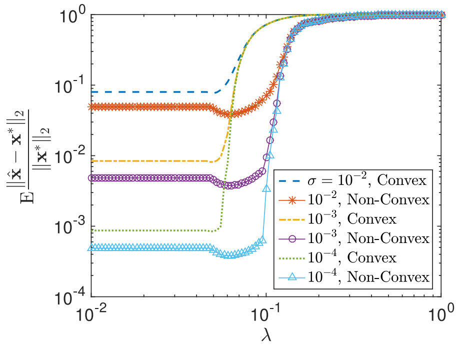

Fig. 2. Mean relative error versus λ when $k = 50$ and $\zeta = 10^{-0.3}$. The convex square-root Lasso is solved by CVX [27], and the nonconvex problem (2) is solved by Algorithm 1. A total of 100 trials are repeated for each point.

A. Row-Orthonormal Sensing Matrix

Now we evaluate the performance of Algorithm 1. The sensing matrix is generated by orthonormalizing rows of a Gaussian random matrix where $N = 2^{10}$ and $M = 2^8$.

In the first experiment, we test the recovery probability of Algorithm 1 versus different choices of k and nonconvexity $\zeta = -\beta/\alpha$ in the noise-free case. We set $\lambda = 0.01$, $\mu = 2\lambda\zeta$, $\rho = 4$, $\varepsilon = 10^{-5}$, and $T_M = 10^4$. If the relative error is less than 10^{-2}, the recovery is regarded as a success. The result is shown in Fig. 1, revealing that ζ determines whether any k-sparse signal can be recovered. When $0 < \zeta < 10^{0.3}$, the proposed algorithm can recover signals with more nonzero entries than that with $\zeta = 0$, which corresponds to square-root Lasso. The optimal choice of ζ is about $\zeta = 10^{-0.3}$.

In the second experiment, we examine the denoising performance for different choices of λ 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 λ 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 $M = 2^8$, and each entry follows $\mathcal{N}(0, 1/N)$. The two tests in Section IV-A are repeated with $\mu = 100\lambda\zeta$, $\rho = 4$, $\delta = 1/\|\mathbf{A}\mathbf{A}^{\mathrm{T}} + \mathbf{I}\|_2$, $\varepsilon = 10^{-5}$, and $T_M = 4 \times 10^4$. The reason of using a larger μ is to speed up its convergence. The results are shown in Figs. 3 and 4, and similar findings are obtained.

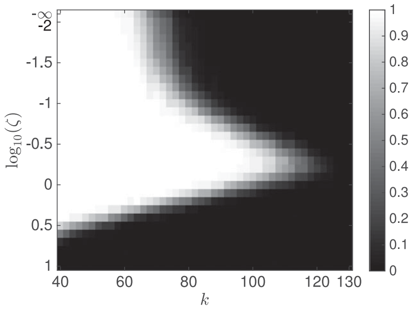

Fig. 3. Recovery probability of Algorithm 2 versus k and ζ when $N = 2^{10}$, $M = 2^8$, and $\lambda = 0.01$. A total of 200 trials are repeated for each point.

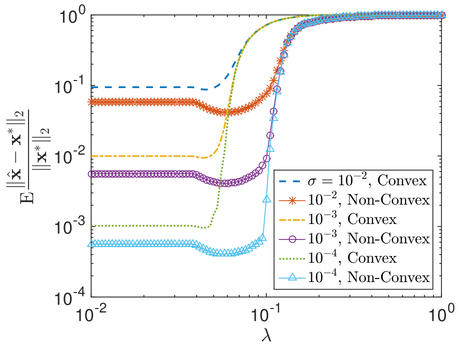

Fig. 4. Mean relative error versus λ when $k = 50$ and $\zeta = 10^{-0.3}$. The convex square-root Lasso is solved by CVX, and the nonconvex problem (2) is solved by Algorithm 2. A total of 100 trials are repeated for each point.

C. Running Time for Large-Scale Problems

Finally, we compare the CPU running time for larger data dimensions. The measurements are partial DCT data. The parameters in each algorithm are the same as in the previous experiments. The benchmark algorithms are CVX [27] and ADMM [14] for the convex square-root Lasso, and the results averaged over 10 trials are tabulated in Table I. When the sparse signal cannot be successfully recovered, "—" is marked. Since CVX needs explicit sensing matrix, it is not simulated when $N = 2^{16}$ and $M = 2^{14}$ due to memory restriction. It is observed that the proposed algorithms enjoy comparable running time with ADMM for the convex square-root Lasso.

V. CONCLUSION

We have presented a class of nonconvex sparsity-inducing penalties for the square-root Lasso. Our motivation is twofold, namely, to achieve improved sparse recovery performance over the convex counterpart, and to eliminate the need of the noise level information. To tackle the resultant nonconvex formulation, we equivalently transform it to a multiconvex optimization problem, which is then solved by the ADMM approach. Two efficient algorithms are devised where the first one considers the special case of a row-orthonormal sensing matrix. Numerical results show that the proposed methods have the advantages of higher recoverable sparsity level, lower recovery error against noise, optimal regularization parameter independent of noise level, and comparable running time with the state-of-the-art approach.

REFERENCES

[1] E. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inform. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[2] D. Donoho, "Compressed sensing," *IEEE Trans. Inform. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[3] E. Candès and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?" *IEEE Trans. Inform. Theory*, vol. 52, no. 12, pp. 5406–5425, Dec. 2006.

[4] M. Figueiredo, R. Nowak, and S. Wright, "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems," *IEEE J. Sel. Topics Signal Process.*, vol. 1, no. 4, pp. 586–597, Dec. 2007.

[5] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statist. Soci. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[6] E. T. Hale, W. Yin, and Y. Zhang, "Fixed-point continuation for ℓ_1-minimization: Methodology and convergence," *SIAM J. Optim.*, vol. 19, no. 3, pp. 1107–1130, 2008.

[7] W. Yin, S. Osher, D. Goldfarb, and J. Darbon, "Bregman iterative algorithms for ℓ_1-minimization with applications to compressed sensing," *SIAM J. Imag. Sci.*, vol. 1, no. 1, pp. 143–168, 2008.

[8] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imag. Sci.*, vol. 2, no. 1, pp. 183–202, 2009.

[9] A. Belloni, V. Chernozhukov, and L. Wang, "Square-root lasso: Pivotal recovery of sparse signals via conic programming," *Biometrika*, vol. 98, no. 4, pp. 791–806, 2011.

[10] C. Thrampoulidis, A. Panahi, D. Guo, and B. Hassibi, "Precise error analysis of the lasso," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 3467–3471.

[11] C. Rojas, D. Katselis, and H. Hjalmarsson, "A note on the SPICE method," *IEEE Trans. Signal Process.*, vol. 61, no. 18, pp. 4545–4551, Sep. 2013.

[12] P. Babu and P. Stoica, "Connection between SPICE and square-root lasso for sparse parameter estimation," *Signal Process.*, vol. 95, pp. 10–14, 2014.

[13] F. Bunea, J. Lederer, and Y. She, "The group square-root lasso: Theoretical properties and fast algorithms," *IEEE Trans. Inform. Theory*, vol. 60, no. 2, pp. 1313–1325, Feb. 2014.

[14] X. Li, T. Zhao, X. Yuan, and H. Liu, "The flare package for high dimensional linear regression and precision matrix estimation in R," *J. Mach. Learn. Res.*, vol. 16, pp. 553–557, 2015.

[15] R. Chartrand, "Exact reconstruction of sparse signals via nonconvex minimization," *IEEE Signal Process. Lett.*, vol. 14, no. 10, pp. 707–710, Oct. 2007.

[16] S. Foucart and M.-J. Lai, "Sparsest solutions of underdetermined linear systems via ℓ_q-minimization for $0 < q \leq 1$," *Appl. Comput. Harmonic Anal.*, vol. 26, no. 3, pp. 395–407, 2009.

[17] R. Gribonval and M. Nielsen, "Highly sparse representations from dictionaries are unique and independent of the sparseness measure," *Appl. Comput. Harmonic Anal.*, vol. 22, no. 3, pp. 335–355, 2007.

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

[19] L. Chen and Y. Gu, "The convergence guarantees of a non-convex approach for sparse recovery using regularized least squares," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 3350–3354.

[20] L. Chen and Y. Gu, "Fast sparse recovery via non-convex optimization," in *Proc. IEEE Global Conf. Signal Inform. Process.*, 2015, pp. 1275–1279.

[21] 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.*, vol. 3, no. 1, pp. 1–122, 2011.

[22] J.-P. Vial, "Strong and weak convexity of sets and functions," *Math. Operat. Res.*, vol. 8, no. 2, pp. 231–259, 1983.

[23] A. Dogandžić and K. Qiu, "Automatic hard thresholding for sparse signal reconstruction from NDE measurements," in *Proc. AIP Conf. Proc.*, vol. 1211, 2010, pp. 806–813.

[24] N. Parikh and S. Boyd, "Proximal algorithms," *Found. Trends Optim.*, vol. 1, no. 3, pp. 127–239, 2014.

[25] X. Zhang, M. Burger, X. Bresson, and S. Osher, "Bregmanized nonlocal regularization for deconvolution and sparse reconstruction," *SIAM J. Imag. Sci.*, vol. 3, no. 3, pp. 253–276, 2010.

[26] B. He and X. Yuan, "On the $o(1/n)$ convergence rate of the Douglas-Rachford alternating direction method," *SIAM J. Numerical Anal.*, vol. 50, no. 2, pp. 700–709, 2012.

[27] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," (Mar. 2014). [Online]. Available: http://cvxr.com/cvx