Journal of Machine Learning Research 21 (2020) 1-38 Submitted 3/19; Revised 6/20; Published 12/20

A Sparse Semismooth Newton Based Proximal Majorization-Minimization Algorithm for Nonconvex Square-Root-Loss Regression Problems

Peipei Tang TANGPP@ZUCC.EDU.CN
School of Computer and Computing Science
Zhejiang University City College
Hangzhou 310015, China

Chengjing Wang RENASCENCEWANG@HOTMAIL.COM
School of Mathematics, and National Engineering Laboratory of Integrated Transportation Big Data Application Technology
Southwest Jiaotong University
No. 999, Xian Road, West Park, High-tech Zone
Chengdu 611756, China

Defeng Sun DEFENG.SUN@POLYU.EDU.HK
Department of Applied Mathematics
The Hong Kong Polytechnic University
Hung Hom, Hong Kong

Kim-Chuan Toh MATTOHKC@NUS.EDU.SG
Department of Mathematics, and Institute of Operations Research and Analytics
National University of Singapore
10 Lower Kent Ridge Road, Singapore

Editor: David Wipf

Abstract

In this paper, we consider high-dimensional nonconvex square-root-loss regression problems and introduce a proximal majorization-minimization (PMM) algorithm for solving these problems. Our key idea for making the proposed PMM to be efficient is to develop a sparse semismooth Newton method to solve the corresponding subproblems. By using the Kurdyka-Łojasiewicz property exhibited in the underlining problems, we prove that the PMM algorithm converges to a d-stationary point. We also analyze the oracle property of the initial subproblem used in our algorithm. Extensive numerical experiments are presented to demonstrate the high efficiency of the proposed PMM algorithm.

Keywords: nonconvex square-root regression problems, proximal majorization-minimization, semismooth Newton method

1. Introduction

Sparsity estimation is one of the most important problems in statistics, machine learning and signal processing. One typical example on this aspect is to estimate a vector $\hat{\beta}$ from a high-dimensional linear regression model

$$b = X\ddot{\beta} + \varepsilon,$$

where $X \in \mathbb{R}^{m\times n}$ is the design matrix, $b \in \mathbb{R}^m$ is the response vector, and $\varepsilon \in \mathbb{R}^m$ is the noise vector for which each of its component ε_i has zero-mean and unknown variance ς^2. In many applications, the number of attributes n is much larger than the sample size m and $\ddot{\beta}$ is known to be sparse a priori. Under the assumption of sparsity, a regularizer which controls the overfitting and/or variable selection is often added to the model. One of the most commonly used regularizers in practice is the ℓ_1 norm and the resulting model, first proposed by Tibshirani (1996), is usually referred to as the Lasso model, which is given by

$$\min_{\beta\in\mathbb{R}^n}\left\{\frac{1}{2}\|X\beta - b\|^2 + \lambda\|\beta\|_1\right\}, \tag{1}$$

where $\|\cdot\|$ is the Euclidean norm in $\mathbb{R}^m$. The Lasso estimator produced from (1) is computationally attractive because it minimizes a structured convex function. Moreover, when the error vector ε follows a normal distribution and suitable design conditions hold, this estimator achieves a near-oracle performance. Despite having these attractive features, the Lasso recovery of $\ddot{\beta}$ relies on knowing the standard deviation of the noise. However, it is non-trivial to estimate the deviation when the feature dimension is large, particularly when it is much larger than the sample size. To overcome the aforementioned defect, Belloni et al. (2011) proposed a new estimator that solves the square-root Lasso (srLasso) model

$$\min_{\beta\in\mathbb{R}^n}\left\{\|X\beta - b\| + \lambda\|\beta\|_1\right\}, \tag{2}$$

which eliminates the need to know or to pre-estimate the deviation. It has been shown (see e.g., Bellec et al., 2018; Derumigny, 2018) that the srLasso estimator can achieve the minimax optimal rate of convergence under some suitable conditions, even though the noise level ς is unknown. It is worth mentioning that the scaled Lasso proposed by Sun and Zhang (2012) is essentially equivalent to the srLasso model (2). The solution approach proposed by Sun and Zhang (2012) is to iteratively solve a sequence of Lasso problems, which can be expensive numerically. Moreover, Xu et al. (2010) proved that the srLasso model (2) is equivalent to a robust linear regression problem subject to an uncertainty set that bounds the norm of the disturbance to each feature, which itself is an ideal approach of reducing sensitivity of linear regression.

The Lasso problem and the srLasso problem are both convex and computationally attractive. A number of algorithms, such as the accelerated proximal gradient (APG) method (Beck and Teboulle, 2009), interior-point method (IPM) (Kim et al., 2007), and least angle regression (LARS) (Efron et al., 2004) have been proposed to solve the Lasso problem (1). In a very recent work, Li et al. (2018) proposed a highly efficient semismooth Newton augmented Lagrangian method to solve the Lasso problem (1). In contrast to the Lasso problem (1), there are currently no efficient algorithms for solving the more challenging srLasso problem (2) due to the fact that the square-root loss function in the objective is nonsmooth. Notably the alternating direction method of multipliers (ADMM) was applied to solve the srLasso problem (2) by Li et al. (2015). However, as can be seen from the numerical experiments conducted later in this paper, the ADMM approach is not very efficient in solving many large-scale problems.

Going beyond the ℓ_1 norm regularizer, other regularization functions for variables selection are often used to avoid overfitting in the area of support vector machines and other

statistical applications. It has also been shown that, instead of a convex relaxation with the ℓ_1 norm, a proper nonconvex regularization can achieve a sparse estimation with fewer measurements, and is more robust against noises (Chartrand, 2007; Chen and Gu, 2014). After the pioneering work of Fan and Li (2001), various nonconvex sparsity functions have been proposed as surrogates of the ℓ_0 function in the last decade and they have been used as regularizers to avoid model overfitting (see e.g., Hastie et al., 2015) in high-dimensional statistical learning. It has been proven that each of these nonconvex surrogate sparsity functions can be expressed as the difference of two convex functions (Ahn et al., 2017; Le Thi et al., 2015). Given the d.c. (difference of convex functions) property of these nonconvex regularizers, it is natural for one to design a majorization-minimization algorithm to solve the nonconvex problem. Such an exploitation of the d.c. property of the regularization function had been considered in the majorized penalty approach proposed by Gao and Sun (2010) for solving a rank constrained correlation matrix estimation problem.

In this paper, we aim to develop an efficient and robust algorithm for solving the following square-root regression problem

$$\min_{\beta\in\mathbb{R}^n} \Big\{ g(\beta) := \|X\beta - b\| + \lambda p(\beta) - q(\beta) \Big\}, \tag{3}$$

where the first part of the regularization function $p : \mathbb{R}^n \to \mathbb{R}_+$ is a norm function whose proximal mapping is strongly semismooth and the second part $q : \mathbb{R}^n \to \mathbb{R}$ is a convex smooth function (the dependence of q on λ has been dropped here). We should note that the assumption made on p is rather mild as the proximal mappings of many interesting functions, such as the l_1-norm function, are strongly semismooth (Meng et al., 2005). For the case when $q \equiv 0$, the oracle property of the model has been established by Stucky and van de Geer (2017) when p is a weakly decomposable norm. For the need of efficient computations, here we shall extend the analysis to the same model but with the proximal terms $\frac{\sigma}{2}\|\beta\|^2 + \frac{\tau}{2}\|X\beta - b\|^2$ added, where $\sigma \geq 0$ and $\tau \geq 0$ are given parameters.

Based on the d.c. structure of the nonconvex regularizer in (3), we design a two stage proximal majorization-minimization (PMM) algorithm to solve the problem (3). In both stages of the PMM algorithm, the key step in each iteration is to solve a convex subproblem whose objective contains the sum of two nonsmooth functions, namely, the square-root-loss regression function and $p(\cdot)$. One of the main contributions of this paper is in proposing a novel proximal majorization approach to solve the said subproblem via its dual by a highly efficient semismooth Newton method. We also analyze the convergence properties of our algorithm. By using the Kurdyka-Łojasiewicz (KL) property exhibited in the underlying problem, we prove that the PMM algorithm converges to a d-stationary point. In the last part of the paper, we present comprehensive numerical results to demonstrate the efficiency of our semismooth Newton based PMM algorithm. Based on the performance of our algorithm against two natural first-order methods, namely the primal based and dual based ADMM (for the convex case), we can safely conclude that our algorithmic framework is far superior for solving the square-root regression problem (3).

2. Preliminary

Let $f : \mathbb{R}^n \to (-\infty, +\infty]$ be a proper closed convex function. The Fenchel conjugate of f is defined by $f^*(x) := \sup_{y\in\mathbb{R}^n}\{\langle y, x\rangle - f(y)\}$, the proximal mapping and the Moreau envelope

function of f with parameter $t > 0$ are defined, respectively, as

$$\begin{aligned} \mathrm{Prox}_f(x) &:= \operatorname*{argmin}_{y\in\mathbb{R}^n}\Big\{f(y)+\frac{1}{2}\|y-x\|^2\Big\}, \quad \forall\, x\in\mathbb{R}^n, \\ \Phi_{tf}(x) &:= \min_{y\in\mathbb{R}^n}\Big\{f(y)+\frac{1}{2t}\|y-x\|^2\Big\}, \quad \forall\, x\in\mathbb{R}^n. \end{aligned}$$

Let $t > 0$ be a given parameter. Then by Moreau's identity theorem (see e.g., Rockafellar, 1970, Theorem 31.5), we have that

$$\mathrm{Prox}_{tf}(x) + t\mathrm{Prox}_{f^*/t}(x/t) = x, \quad \forall\, x\in\mathbb{R}^n. \tag{4}$$

We also know from (Rockafellar and Wets, 1998, Proposition 13.37) that Φ_{tf} is continuously differentiable with

$$\nabla\Phi_{tf}(x) = t^{-1}(x - \mathrm{Prox}_{tf}(x)), \quad \forall\, x\in\mathbb{R}^n.$$

Given a set $C \subseteq \mathbb{R}^n$ and an arbitrary collection of functions $\{f_i \mid i \in I\}$ on $\mathbb{R}^n$, we denote $\delta_C(\cdot)$ as the indicator function of C such that $\delta_C(\beta) = 0$ if $\beta \in C$ and $\delta_C(\beta) = +\infty$ if $\beta \notin C$, $\mathrm{conv}(C)$ as the convex hull of C and $\mathrm{conv}\{f_i \mid i \in I\}$ as the convex hull of the pointwise infimum of the collection. This means that $\mathrm{conv}\{f_i \mid i \in I\}$ is the greatest convex function f on $\mathbb{R}^n$ such that $f(\beta) \le f_i(\beta)$ for any $\beta \in \mathbb{R}^n$ and $i \in I$.

Next we present some results in variational analysis from Rockafellar and Wets (1998). Let $\Psi : \mathcal{O} \to \mathbb{R}^m$ be a locally Lipschitz continuous vector-valued function defined on an open set $\mathcal{O} \subseteq \mathbb{R}^n$. It follows from (Rockafellar and Wets, 1998, Theorem 9.60) that Ψ is F(réchet)-differentiable almost everywhere on $\mathcal{O}$. Let $\mathcal{D}_\Psi$ be the set of all points where Ψ is F-differentiable and $J\Psi(x) \in \mathbb{R}^{m\times n}$ be the Jacobian of Ψ at $x \in \mathcal{D}_\Psi$. For any $x \in \mathcal{O}$, the B-subdifferential of Ψ at x is defined by

$$\partial_B\Psi(x) := \left\{ V\in\mathbb{R}^{m\times n} \,\middle|\, \exists\,\{x^k\}\subseteq\mathcal{D}_\Psi \text{ such that } \lim_{k\to\infty} x^k = x \text{ and } \lim_{k\to\infty} J\Psi(x^k) = V \right\}.$$

The Clarke subdifferential of Ψ at x is defined as the convex hull of the B-subdifferential of Ψ at x, that is $\partial\Psi(x) := \mathrm{conv}(\partial_B\Psi(x))$.

Let ϕ be defined from $\mathbb{R}^n$ to $\mathbb{R}$. The Clarke subdifferential of ϕ at $x \in \mathbb{R}^n$ is defined by

$$\partial_C\phi(x) := \left\{ h\in\mathbb{R}^n \,\middle|\, \limsup_{x'\to x, t\downarrow 0} \frac{\phi(x'+ty)-\phi(x')-th^Ty}{t} \ge 0\ ,\ \forall\, y\in\mathbb{R}^n \right\}.$$

The regular subdifferential of ϕ at $x \in \mathbb{R}^n$ is defined as

$$\hat\partial\phi(x) := \left\{ h\in\mathbb{R}^n \,\middle|\, \liminf_{x\neq y\to x} \frac{\phi(y)-\phi(x)-h^T(y-x)}{\|y-x\|} \ge 0 \right\}$$

and the limiting subdifferential of ϕ at x is defined as

$$\partial\phi(x) := \left\{ h\in\mathbb{R}^n \,\middle|\, \exists\ \{x^k\}\to x \text{ and } \{h^k\}\to h \text{ satisfying } h^k\in\hat\partial\phi(x^k),\ \forall\, k \right\}.$$

If ϕ is a convex function, then the Clarke subdifferential, the regular subdifferential and the limiting subdifferential of ϕ at x coincide with the set of (transposed) subgradients of ϕ at x in the sense of convex analysis.

We know from (Rockafellar and Wets, 1998, Theorem 10.1) that $0 \in \hat{\partial}\phi(\bar{x})$ is a necessary condition for $\bar{x} \in \mathbb{R}^n$ to be a local minimizer of ϕ. If the function ϕ (may not be convex) is locally Lipschitz continuous near $\bar{x}$ and directionally differentiable at $\bar{x}$, then $0 \in \hat{\partial}\phi(\bar{x})$ is equivalent to the directional-stationarity (d-stationarity) of $\bar{x}$, that is

$$\phi'(\bar{x}; h) := \lim_{\lambda \downarrow 0} \frac{\phi(\bar{x} + \lambda h) - \phi(\bar{x})}{\lambda} \geq 0, \quad \forall\, h \in \mathbb{R}^n.$$

In this paper, we will prove that the sequence generated by our algorithm converges to a d-stationary point of the problem.

For further discussions, we recall the concept of semismoothness originated from (Mifflin, 1977; Qi and Sun, 1993) and other two definitions used in (van de Geer, 2014).

Definition 1 *Let $F : \mathcal{O} \subseteq \mathbb{R}^n \to \mathbb{R}^m$ be a locally Lipschitz continuous function and $\mathcal{K} : \mathcal{O} \rightrightarrows \mathbb{R}^{m\times n}$ be a nonempty and compact valued, upper-semicontinuous set-valued mapping on the open set $\mathcal{O}$. F is said to be semismooth at $v \in \mathcal{O}$ with respect to the set-valued mapping $\mathcal{K}$ if F is directionally differentiable at v and for any $\Gamma \in \mathcal{K}(v+\Delta v)$ with $\Delta v \to 0$,*

$$F(v + \Delta v) - F(v) - \Gamma \Delta v = o(\|\Delta v\|).$$

F is said to be γ-order ($\gamma > 0$) (strongly, if $\gamma = 1$) semismooth at v with respect to $\mathcal{K}$ if F is semismooth at v and for any $\Gamma \in \mathcal{K}(v + \Delta v)$,

$$F(v + \Delta v) - F(v) - \Gamma \Delta v = O(\|\Delta v\|^{1+\gamma}).$$

F is called a semismooth (γ-order semismooth, strongly semismooth) function on $\mathcal{O}$ with respect to $\mathcal{K}$ if it is semismooth (γ-order semismooth, strongly semismooth) at every $v \in \mathcal{O}$ with respect to $\mathcal{K}$.

Definition 2 *The norm function p defined in $\mathbb{R}^n$ is said to be weakly decomposable for an index set $S \subset \{1, 2, \ldots, n\}$ if there exists a norm $p^{\bar{S}}$ defined on $\mathbb{R}^{|\bar{S}|}$ such that*

$$p(\beta) \geq p(\beta_S) + p^{\bar{S}}(\beta^{\bar{S}}), \quad \forall\, \beta \in \mathbb{R}^n,$$

where $\bar{S} = \{1, 2, \ldots, n\} \backslash S$, $\beta_S = \beta \circ \mathbf{1}_S$ and $\beta^{\bar{S}} := (\beta_j)_{j\in\bar{S}} \in \mathbb{R}^{|\bar{S}|}$. Here $\mathbf{1}_S \in \mathbb{R}^n$ denotes the indicator vector of S and "$\circ$" denotes the elementwise product.

The weakly decomposable property of a norm is a relaxation of the decomposability property of the ℓ_1 norm. It has been proved by Stucky and van de Geer (2017) that many interesting regularizers such as the sparse group Lasso and SLOPE are weakly decomposable. A set S is said to be an allowed set if p is a weakly decomposable norm for this set. We say that a vector $\beta \in \mathbb{R}^n$ satisfies the (L, S)-cone condition for a norm p if $p^{\bar{S}}(\beta^{\bar{S}}) \leq Lp(\beta_S)$ with $L > 0$ and S being an allowed set.

Definition 3 *Given $X \in \mathbb{R}^{m\times n}$. Let S be an allowed set of a weakly decomposable norm p and $L > 0$ be a constant. Then the p-eigenvalue is defined as*

$$\delta_p(L,S) := \min\Big\{\|X\beta_S - X\beta_{\bar{S}}\| \;\Big|\; p(\beta_S) = 1,\ p^{\bar{S}}(\beta^{\bar{S}}) \le L,\ \beta \in \mathbb{R}^n\Big\}.$$

The p-effective sparsity is defined as

$$\Gamma_p(L,S) := \frac{1}{\delta_p(L,S)}.$$

Note that the p-eigenvalue defined above is a generalization of the compatibility constant defined by van de Geer (2007).

3. The Oracle Property of the Square-Root Regression Problem with a Generalized Elastic-Net Regularization

We first consider the following convex problem without q in (3), that is

$$\min_{\beta\in\mathbb{R}^n}\Big\{\|X\beta - b\| + \lambda p(\beta)\Big\}. \tag{5}$$

By adding proximal terms, we shall analyze the oracle property of the square-root regression problem with a generalized elastic-net regularization. For given $\sigma \ge 0$ and $\tau \ge 0$, it takes the following form

$$\min_{\beta\in\mathbb{R}^n}\Big\{\|X\beta - b\| + \lambda p(\beta) + \frac{\sigma}{2}\|\beta\|^2 + \frac{\tau}{2}\|X\beta - b\|^2\Big\}, \tag{6}$$

whose optimal solution set is denoted as $\Omega(\sigma,\tau)$. The purpose of this section is to study the oracle property of an estimator $\hat{\beta} \in \Omega(\sigma,\tau)$ (whose residual is given by $\hat{\varepsilon} := b - X\hat{\beta}$) to evaluate how good the estimator is in estimating the true vector $\ddot{\beta}$.

For the given norm p, the dual norm of p is given by

$$p_*(\beta) := \max_{z\in\mathbb{R}^n}\Big\{\langle z,\beta\rangle \mid p(z) \le 1\Big\},\ \forall\,\beta\in\mathbb{R}^n.$$

For a weakly decomposable norm p with the allowed set S, we let

$$n_p = \tfrac{\lambda p(\ddot{\beta})}{\|\varepsilon\|},\ \lambda_0 = \tfrac{p_*(\varepsilon^T X)}{\|\varepsilon\|},\ \lambda_m = \max\Big\{\tfrac{p_*^{\bar{S}}((\varepsilon^T X)^{\bar{S}})}{\|\varepsilon\|}, \tfrac{p_*((\varepsilon^T X)_S)}{\|\varepsilon\|}, p_*^{\bar{S}}(\ddot{\beta}^{\bar{S}}), p_*(\ddot{\beta}_S)\Big\}, \tag{7}$$

$$t_1 = 1 + \tfrac{\tau}{2}\|\varepsilon\| + \tfrac{\sigma p_*(\ddot{\beta})p(\ddot{\beta})}{2\|\varepsilon\|},\quad t_2 = 2 + \tau + \tfrac{\sigma p_*(\ddot{\beta})p(\ddot{\beta})}{\|\varepsilon\|},$$

$$c_u = t_1 + n_p,\quad a = \Big(\lambda_0 + \sigma p_*(\ddot{\beta})c_u\Big)\tfrac{t_1}{\lambda}, \tag{8}$$

where $p_*^{\bar{S}}$ denotes the dual norm of $p^{\bar{S}}$.

Next we state two basic assumptions which are similar to those in (Stucky and van de Geer, 2017). The first assumption is about non-overfitting in the sense that if the optimal solution $\hat{\beta}$ of (3) satisfies $\|\hat{\varepsilon}\| = 0$, then it overfits. Furthermore, it has been proved by Di Pillo and Grippo (1988) that there exists a scalar λ^* such that the solution of the problem (6) satisfies $X\hat{\beta} - b \ne 0$ if $\lambda > \lambda^*$. In other words, one can find the parameter λ to avoid overfitting.

Assumption 1 *We assume that* $\|\hat{\varepsilon}\| \neq 0$ *and* $a + \frac{2\lambda_0 n_p}{\lambda} < 1$, *where the constant* a *is defined in (8).*

Assumption 2 *The function* p *is a norm function and weakly decomposable in* $\mathbb{R}^n$ *for an index set* $S \subset \{1, \ldots, n\}$, *i.e., there exists a norm* $p^{\bar{S}}$ *defined on* $\mathbb{R}^{|\bar{S}|}$ *such that*

$$p(\beta) \geq p(\beta_S) + p^{\bar{S}}(\beta^{\bar{S}}), \quad \forall\, \beta \in \mathbb{R}^n.$$

Remark 4 *If* $\sigma = 0$ *and* $\tau = 0$, *then Assumption 1 goes back to Assumption I given by Stucky and van de Geer (2017). Let* $s > 0$ *and* $P(\|\varepsilon\| \leq s)$ *be the probability of* $\|\varepsilon\| \leq s$. *It follows from the comment to (Laurent and Massart, 2000, Lemma 1) that* $P\left(\|\varepsilon\| \leq \varsigma\sqrt{n - 2\sqrt{ns}}\right) \leq e^{-s}$ *and* $P\left(\|\varepsilon\| \geq \varsigma\sqrt{n + 2\sqrt{ns} + 2s}\right) \leq e^{-s}$. *Noticing that for* $\alpha_1 > e^{-n/4}$, $\alpha_2 > 0$ *and* $\alpha_1 + \alpha_2 < 1$, $P(\bar{\chi}_1 \leq \|\varepsilon\| \leq \bar{\chi}_2) \geq 1 - \alpha_1 - \alpha_2$ *with* $\bar{\chi}_1 := \varsigma\sqrt{n - 2\sqrt{-n\ln(\alpha_1)}}$ *and* $\bar{\chi}_2 := \varsigma\sqrt{n + 2\sqrt{-n\ln(\alpha_2)} - 2\ln(\alpha_2)}$, *the inequalities* $n_p \leq \lambda p(\ddot{\beta})/\bar{\chi}_1$, $t_1 \leq 1 + \tau\bar{\chi}_2/2 + \sigma p_*(\ddot{\beta})p(\ddot{\beta})/(2\bar{\chi}_1) := \bar{t}_1$, $c_u \leq \bar{t}_1 + \lambda p(\ddot{\beta})/\bar{\chi}_1$, $a \leq \left(\bar{\lambda}_0 + \sigma p_*(\ddot{\beta})\bar{t}_1 + \lambda\sigma p_*(\ddot{\beta})p(\ddot{\beta})/\bar{\chi}_1\right)\bar{t}_1/\lambda$ *hold with probability at least* $1 - \alpha_1 - \alpha_2$. *It has been proven in (Stucky and van de Geer, 2017, Section 3.4) that*

$$\max\left\{p_*^{\bar{S}}((\varepsilon^T X)^{\bar{S}})/\|\varepsilon\|, p_*((\varepsilon^T X)_S)/\|\varepsilon\|\right\} \leq \bar{\lambda}_0$$

holds with high probability for some $\bar{\lambda}_0$ *(may depend on* $\sqrt{m}$ *and* $\sqrt{ln(n)}$*). Combining with the result* $\lambda_0 \leq \bar{\lambda}_0$ *by Lemma 5 (see below),* $\bar{\chi}_1 > 2\bar{\lambda}_0 p(\ddot{\beta})$ *is valid for high dimensional problems with high probability. We can choose* τ, σ *and* λ *with*

$$\begin{aligned} \bar{\chi}_1 &> \sigma p_*(\ddot{\beta})p(\ddot{\beta})\bar{t}_1 + 2\bar{\lambda}_0 p(\ddot{\beta}), \\ \lambda &> \max\left\{\frac{\bar{\chi}_1(\bar{\lambda}_0\bar{t}_1 + \sigma p_*(\ddot{\beta})\bar{t}_1^2)}{\bar{\chi}_1 - \sigma p_*(\ddot{\beta})p(\ddot{\beta})\bar{t}_1 - 2\bar{\lambda}_0 p(\ddot{\beta})}, \lambda^*\right\} \end{aligned}$$

such that Assumption 1 holds with high probability with respect to the random noise vector ε. *Assumption 2 was also used in (Bach et al., 2012; van de Geer, 2014).*

We also need some lemmas in order to prove the main theorem. First we introduce the following basic relationship between λ_0 and λ_m.[1]

Lemma 5 *Let* S *be an allowed set of a weakly decomposable norm* p. *For the parameters* λ_0 *and* λ_m *defined by (7), we have* $\lambda_0 \leq \lambda_m$ *and* $p_*(\ddot{\beta}) \leq \lambda_m$.

The following lemma, which is from (van de Geer, 2014), shows that $p(\beta_S)$ is bounded by $\|X\beta\|$.

1. Actually in (Stucky and van de Geer, 2017) Stucky and van de Geer once employed this relationship without a proof. For the sake of clarity, we present this fact in the form of a lemma here and its proof in the appendix.

Lemma 6 *Given $X \in \mathbb{R}^{m\times n}$. Let S be an allowed set of a weakly decomposable norm p and $L > 0$ be a constant. Then the p-eigenvalue can be expressed in the following form:*

$$\delta_p(L,S) = \min\left\{ \frac{\|X\beta\|}{p(\beta_S)} \;\middle|\; \beta \in \mathbb{R}^n \text{ satisfies the } (L,S)\text{-cone condition and } \beta_S \neq 0 \right\}.$$

That is $p(\beta_S) \le \Gamma_p(L,S)\|X\beta\|$.

An upper bound of $\hat{\varepsilon}^T X(\ddot{\beta} - \hat{\beta})$, a lower bound and an upper bound of $\|\hat{\varepsilon}\|$ are also important. They are presented in the following two lemmas.

Lemma 7 *Suppose that Assumption 1 holds. For the estimator $\hat{\beta}$ of the generalized elastic-net square-root regression problem (6), we have*

$$\hat{\varepsilon}^T X(\ddot{\beta} - \hat{\beta}) \le \left(\tau + \frac{1}{\|\hat{\varepsilon}\|}\right)^{-1} \left(\lambda p(\ddot{\beta}) + \sigma p_*(\ddot{\beta}) p(\hat{\beta})\right).$$

Lemma 8 *Suppose that Assumption 1 holds. We have*

$$c_l := \frac{1 - a - \frac{2\lambda_0 n_p}{\lambda}}{2 + \left(1 + \frac{\sigma p_*(\ddot{\beta})}{\lambda}\right) n_p} \;\le\; \frac{\|\hat{\varepsilon}\|}{\|\varepsilon\|} \;\le\; c_u,$$

where the constants c_u and a are defined in (8).

Based on the above lemmas, we can present the following sharp oracle inequality on the prediction error.

Theorem 9 *Let $\delta \in [0,1)$. Under Assumptions 1 and 2, assume that $\frac{s_2 - \sqrt{s_2^2 - 4s_1 s_3}}{2s_1} < \lambda < \frac{s_2 + \sqrt{s_2^2 - 4s_1 s_3}}{2s_1}$ with $s_1 = \frac{\sigma\lambda_m p^2(\ddot{\beta})}{\|\varepsilon\|^2}$, $s_2 = 1 - \frac{\lambda_m(3 + 2\sigma t_1 + \sigma t_2)p(\ddot{\beta})}{\|\varepsilon\|} > 0$ and $s_3 = \lambda_m(t_1 + t_2 + \sigma t_1 t_2 + \sigma t_1^2)$. For any $\hat{\beta} \in \Omega(\sigma,\tau)$, and any $\beta \in \mathbb{R}^n$ such that supp(β) is a subset of S, we have that*

$$\begin{aligned}
&\|X(\hat{\beta} - \ddot{\beta})\|^2 + 2\delta\left((\hat{\lambda} - \tilde{\lambda}_m)p^{\bar{S}}(\hat{\beta}^{\bar{S}}) + (\tilde{\lambda} + \tilde{\lambda}_m)p(\hat{\beta}_S - \beta)\right)\|\varepsilon\| \\
&\le \|X(\beta - \ddot{\beta})\|^2 + \left((1+\delta)(\tilde{\lambda} + \tilde{\lambda}_m)\Gamma_p(L_S,S)\|\varepsilon\|\right)^2 + 2\sigma c_u\|\hat{\beta} - \ddot{\beta}\|\|\beta - \ddot{\beta}\|\|\varepsilon\|,
\end{aligned}$$

where

$$\hat{\lambda} := \frac{\lambda c_l}{1 + \tau c_l},\ \tilde{\lambda}_m := \lambda_m(1 + \sigma c_u),\ \tilde{\lambda} := \lambda c_u,\ L_S := \frac{\tilde{\lambda} + \tilde{\lambda}_m}{\hat{\lambda} - \tilde{\lambda}_m} \cdot \frac{1+\delta}{1-\delta}.$$

An important special case of Theorem 9 is to choose $\beta = \ddot{\beta}$ with supp$(\ddot{\beta}) \subseteq S$ allowed. Then only the p-effective sparsity term $\Gamma_p(L_S,S)$ appears in the upper bound.

Remark 10 *Since $\lim\limits_{\sigma\downarrow 0}\left(s_2^2 - 4s_1 s_3\right) = \left(1 - \frac{3\lambda_m p(\ddot{\beta})}{\|\varepsilon\|}\right)^2 > 0$, we can find some $\tilde{\sigma} > 0$ such that $s_2^2 - 4s_1 s_3 > 0$ holds if $\sigma < \tilde{\sigma}$. Theorem 9 is nearly the same as that in (Stucky and van de Geer, 2017) due to $\lim\limits_{\sigma\downarrow 0,\tau\downarrow 0} \frac{s_2 - \sqrt{s_2^2 - 4s_1 s_3}}{2s_1} = \frac{3\lambda_m\|\varepsilon\|}{\|\varepsilon\| - 3\lambda_m p(\ddot{\beta})}$ and $\lim\limits_{\sigma\downarrow 0,\tau\downarrow 0} \frac{s_2 + \sqrt{s_2^2 - 4s_1 s_3}}{2s_1} = +\infty$ with a different definition of λ_m.*

Remark 11 *From Theorem 9 we can see that the upper bound is related to some random parts λ_m and $\|\varepsilon\|$. If we have Gaussian errors $\varepsilon \sim \mathcal{N}(0, \varsigma^2 I)$, then we know from (Stucky and van de Geer, 2017, Proposition 11) that there exists an upper bound λ_m^u of λ_m such that $\lambda_m \leq \lambda_m^u$ is valid with probability $1-\alpha$ for a given constant α. Furthermore, it follows from Laurent and Massart (2000) that we can find an upper bound $c_1\varsigma$ and a lower bound $c_2\varsigma$ for $\|\epsilon\|$ with a high probability. That is, if λ_m is replaced by λ_m^u and $\|\epsilon\|$ is replaced by $c_1\varsigma$ or $c_2\varsigma$, then the sharp oracle bound with the Gaussian errors holds with a high probability.*

4. The Proximal Majorization-Minimization Algorithm

To deal with the nonconvexity of the regularization function in the square-root regression problem (3), we design a two stage proximal majorization-minimization (PMM) algorithm and solve a series of convex subproblems. In stage I, we first solve a problem by removing q from the original problem and adding appropriate proximal terms, to obtain an initial point to warm-start the algorithm in the second stage. In stage II, a series of majorized subproblems are solved to obtain a solution point.

The basic idea of the PMM algorithm is to linearize the concave term $-q(\beta)$ in the objective function of (3) at each iteration with respect to the current iterate, say $\tilde{\beta}$. By doing so, the subproblem in each iteration is a convex minimization problem, which must be solved efficiently in order for the overall algorithm to be efficient. However, the objective function of the resulting subproblem contains the sum of two nonsmooth terms ($\|X\beta - b\|$ and $p(\beta)$), and it is not obvious how such a problem can be solved efficiently. One important step we take in this paper is to add the proximal term $\frac{\tau}{2}\|X\beta - X\tilde{\beta}\|^2$ to the objective function of the subproblem. Through this novel proximal term, the dual of the majorized subproblem can then be written explicitly as an unconstrained optimization problem. Moreover, its structure is highly conducive for one to apply the semismooth Newton (SSN) method to compute an approximate solution via solving a nonlinear system of equations.

4.1 A Semismooth Newton Method for the Subproblems

For the purpose of our algorithm developments, given $\sigma > 0$, $\tau > 0$, $\tilde{\beta} \in \mathbb{R}^n$, $\tilde{v} \in \mathbb{R}^n$, and $\tilde{b} \in \mathbb{R}^m$, we consider the following minimization problem:

$$\begin{aligned} \min_{\beta\in\mathbb{R}^n} \Big\{ h(\beta;\sigma,\tau,\tilde{\beta},\tilde{v},\tilde{b}) \;\; := \;\; & \|X\beta - b\| + \lambda p(\beta) - q(\tilde{\beta}) - \langle \tilde{v}, \beta - \tilde{\beta}\rangle \\ & + \frac{\sigma}{2}\|\beta - \tilde{\beta}\|^2 + \frac{\tau}{2}\|X\beta - \tilde{b}\|^2 \Big\}. \end{aligned} \tag{9}$$

The optimization problem (9) is equivalent to

$$\min_{\beta\in\mathbb{R}^n, y\in\mathbb{R}^m} \Big\{ \|y\| + \lambda p(\beta) - \langle \tilde{v}, \beta - \tilde{\beta}\rangle + \frac{\sigma}{2}\|\beta - \tilde{\beta}\|^2 + \frac{\tau}{2}\|y + b - \tilde{b}\|^2 \;\Big|\; X\beta - y = b \Big\}. \tag{10}$$

The dual of problem (10) admits the following equivalent minimization form:

$$\begin{aligned}\min_{u\in\mathbb{R}^m}\Big\{\varphi(u) := \langle u, b\rangle + \frac{\tau}{2}\|\tau^{-1}u+\tilde{b}-b\|^2 - \|\mathrm{Prox}_{\tau^{-1}\|\cdot\|}(\tau^{-1}u+\tilde{b}-b)\| \\ -\frac{1}{2\tau}\|\mathrm{Prox}_{\tau\delta_B}(u+\tau(\tilde{b}-b))\|^2 + \frac{\sigma}{2}\|\tilde{\beta}+\sigma^{-1}(\tilde{v}-X^Tu)\|^2 \\ -\lambda\, p\Big(\mathrm{Prox}_{\sigma^{-1}\lambda p}\big(\tilde{\beta}+\sigma^{-1}(\tilde{v}-X^Tu)\big)\Big) - \frac{1}{2\sigma}\|\mathrm{Prox}_{\sigma(\lambda p)^*}(\sigma\tilde{\beta}+\tilde{v}-X^Tu)\|^2\Big\},\end{aligned} \tag{11}$$

where $B = \{x \mid \|x\| \le 1\}$. Let $\bar{u} := \underset{u\in\mathbb{R}^m}{\mathrm{argmin}}\,\varphi(u)$. Then the optimal solutions $\bar{y}, \bar{\beta}$ to the primal problem (10) can be computed by

$$\bar{y} = \mathrm{Prox}_{\tau^{-1}\|\cdot\|}(\tau^{-1}\bar{u}+\tilde{b}-b), \quad \bar{\beta} = \mathrm{Prox}_{\sigma^{-1}\lambda p}\big(\tilde{\beta}+\sigma^{-1}(\tilde{v}-X^T\bar{u})\big).$$

Here we should emphasize the novelty in adding the proximal term $\frac{\tau}{2}\|X\beta-\tilde{b}\|^2$ in (9). Without this term, the objective function in the dual problem (11) does not admit an analytical expression. As the reader may observe later in the next paragraph, without the analytical expression given in (11), the algorithmic development in the rest of this subsection would break down. As a result, designing the PMM algorithm in the next subsection for solving (3) would also become impossible.

By Moreau's identity (4) and the differentiability of the Moreau envelope functions of $\|\cdot\|$ and λp, we know that the function φ is convex and continuously differentiable and

$$\nabla\varphi(u) = \mathrm{Prox}_{\tau^{-1}\|\cdot\|}(\tau^{-1}u+\tilde{b}-b) - X\mathrm{Prox}_{\sigma^{-1}\lambda p}\big(\tilde{\beta}+\sigma^{-1}(\tilde{v}-X^Tu)\big) + b.$$

Thus $\bar{u}$ can be obtained via solving the following nonlinear system of equations:

$$\nabla\varphi(u) = 0. \tag{12}$$

In the rest of this subsection, we will discuss how we can apply the SSN method to compute an approximate solution of (12) efficiently. Since the mappings $\mathrm{Prox}_{\sigma^{-1}\|\cdot\|}(\cdot)$ and $\mathrm{Prox}_{\tau^{-1}\lambda p}(\cdot)$ are Lipschitz continuous, the following multifunction is well defined:

$$\hat{\partial}^2\varphi(u) := \sigma^{-1}X\partial\mathrm{Prox}_{\sigma^{-1}\lambda p}(\tilde{\beta}+\sigma^{-1}(\tilde{v}-X^Tu))X^T + \tau^{-1}\partial\mathrm{Prox}_{\tau^{-1}\|\cdot\|}(\tau^{-1}u+\tilde{b}-b).$$

Let $U \in \partial\mathrm{Prox}_{\sigma^{-1}\lambda p}(\tilde{\beta}+\sigma^{-1}(\tilde{v}-X^Tu))$ and $V \in \partial\mathrm{Prox}_{\tau^{-1}\|\cdot\|}(\tau^{-1}u+\tilde{b}-b)$. Then we have $H := \sigma^{-1}XUX^T + \tau^{-1}V \in \hat{\partial}^2\varphi(u)$. The following proposition demonstrates that H is nonsingular at the solution point that does not over fit, which, under Assumption 1, holds automatically when it is close to $\hat{\beta}$. This property is crucial to ensure the fast convergence of the SSN method for computing an approximate solution of (11).

Proposition 12 *Suppose that the unique optimal solution $\bar{\beta}$ to the problem (9) satisfies $\|X\bar{\beta}-b\| \neq 0$. Then all the elements of $\hat{\partial}^2\varphi(\bar{u})$ are positive definite.*

Proof By the assumption, $\bar{y} = X\bar{\beta}-b \neq 0$. Furthermore, $\bar{y} = \mathrm{Prox}_{\tau^{-1}\|\cdot\|}(\tilde{u}) = \tilde{u} - \tau^{-1}\Pi_B(\tau\tilde{u})$, where $\tilde{u} = \tau^{-1}\bar{u}+\tilde{b}-b$, Π_B is the Euclidean projection operator onto B. Since $\bar{y} \neq 0$, it follows that $\|\tilde{u}\| > \frac{1}{\tau}$ and $\mathrm{Prox}_{\tau^{-1}\|\cdot\|}(\tilde{u})$ is differentiable with

$$V := \nabla\mathrm{Prox}_{\tau^{-1}\|\cdot\|}(\tilde{u}) = \left(1-\frac{1}{\tau\|\tilde{u}\|}\right)\mathcal{I}_m + \frac{\tilde{u}\tilde{u}^T}{\tau\|\tilde{u}\|^3}.$$

Hence for any $U \in \partial \mathrm{Prox}_{\sigma^{-1}\lambda p}(\tilde{\beta} + \sigma^{-1}(\tilde{v} - X^T\bar{u}))$, $H = \sigma^{-1}XUX^T + \tau^{-1}V \in \hat{\partial}^2\varphi(\bar{u})$. Since V is positive definite and XUX^T is positive semidefinite, H is positive definite. This completes the proof. ■

Now we discuss how to apply the SSN method to solve the nonsmooth equation (12) to obtain an approximate solution efficiently. We first prove that $\nabla\varphi$ is strongly semismooth.

Proposition 13 *The function $\nabla\varphi$ is strongly semismooth.*

Proof Firstly, we have assumed that the proximal operator $\mathrm{Prox}_p(\cdot)$ is strongly semismooth. Secondly, by (Chen et al., 2003, Proposition 4.3), it is known that the projection operator onto the second order cone is strongly semismooth. The strongly semismoothness of the proximal operator $\mathrm{Prox}_{\|\cdot\|}(\cdot)$ then follows from (Meng et al., 2005, Theorem 4), which states that if the projection onto the second order cone is strongly semismooth, then so is the proximal mapping of $\|\cdot\|$. From here, it is easy to prove the required result and we omit the details. ■

Now we can apply the SSN method to solve (12) as follows.

Algorithm SSN($\sigma, \tau, \tilde{\beta}, \tilde{v}, \tilde{b}$) with input $\sigma > 0, \tau > 0, \tilde{\beta}, \tilde{v} \in \mathbb{R}^n, \tilde{b} \in \mathbb{R}^m$**.** Choose $\mu \in (0, \frac{1}{2})$, $\overline{\eta} \in (0, 1)$, $\varrho \in (0, 1]$, $\delta \in (0, 1)$, and $u^0 \in \mathbb{R}^m$. For $j = 0, 1, \ldots,$ iterate the following steps:

Step 1. Choose $U^j \in \partial \mathrm{Prox}_{\sigma^{-1}\lambda p}(\tilde{\beta} + \sigma^{-1}(\tilde{v} - X^T u^j))$ and $V^j \in \partial \mathrm{Prox}_{\tau^{-1}\|\cdot\|}(\tau^{-1}u^j + \tilde{b} - b)$. Let $H^j = \sigma^{-1}XU^jX^T + \tau^{-1}V^j$. Compute an approximate solution Δu^j to the linear system

$$H^j \Delta u = -\nabla\varphi(u^j)$$

such that

$$\|H^j\Delta u^j + \nabla\varphi(u^j)\| \leq \min\{\overline{\eta}, \|\nabla\varphi(u^j)\|^{1+\varrho}\}.$$

Step 2. Set $\alpha_j = \delta^{t_j}$, where t_j is the first nonnegative integer t such that

$$\varphi(u^j + \delta^t \Delta u^j) \leq \varphi(u^j) + \mu\delta^t \langle \nabla\varphi(u^j), \Delta u^j \rangle.$$

Step 3. Set $u^{j+1} = u^j + \alpha_j \Delta u^j$.

With Propositions 12 and 13, the SSN method can be proven to be globally convergent and locally superlinearly convergent. One may see (Li et al., 2018, Theorem 3.6) for the details. The local convergence rate for Algorithm SSN is stated in the next theorem without proof.

Theorem 14 *Suppose that* $\|X\bar{\beta} - b\| \neq 0$ *holds. Then the sequence* $\{u^j\}$ *generated by Algorithm SSN converges to the unique optimal solution* $\overline{u}$ *of the problem (11) and*

$$\|u^{j+1} - \overline{u}\| \;\; = \;\; \mathcal{O}(\|u^j - \overline{u}\|)^{1+\varrho}.$$

4.2 The SSN Based Proximal Majorization-Minimization Algorithm

In this subsection, we describe the details of the PMM algorithm for solving (3) wherein each subproblem is solved by the SSN method. We briefly present the PMM algorithm as follows.

Algorithm PMM. Let $\sigma^{2,0} > 0$, $\tau^{2,0} > 0$ be given parameters.

Step 1. Find $\sigma^1 > 0$, $\tau^1 > 0$ and compute

$$\beta^0 \approx \underset{\beta\in\mathbb{R}^n}{\operatorname{argmin}} \left\{h(\beta; \sigma^1, \tau^1, 0, 0, b)\right\} \tag{13}$$

via solving its dual problem such that the KKT residual of the problem (5) satisfies a prescribed stopping criterion. That is, given $(\sigma, \tau, \tilde{\beta}, \tilde{v}, \tilde{b}) = (\sigma^1, \tau^1, 0, 0, b)$, apply the SSN method to find an approximate solution u^0 of (12) and then set $\beta^0 = \mathrm{Prox}_{\lambda p/\sigma^1}(-X^T u^0/\sigma^1)$. Let $k = 0$ and go to Step 2.1.

Step 2.1 Compute

$$\beta^{k+1} \;\; = \;\; \underset{\beta\in\mathbb{R}^n}{\operatorname{argmin}} \Big\{h(\beta; \sigma^{2,k}, \tau^{2,k}, \beta^k, \nabla q(\beta^k), X\beta^k) + \langle \delta^k, \beta - \beta^k\rangle\Big\}$$

via solving its dual problem. That is, given $(\sigma, \tau, \tilde{\beta}, \tilde{v}, \tilde{b}) = (\sigma^{2,k}, \tau^{2,k}, \beta^k, \nabla q(\beta^k), X\beta^k)$, apply the SSN method to find an approximate solution u^{k+1} of (12) such that the error vector δ^k satisfies the following accuracy condition:

$$\|\delta^k\| \le \frac{\sigma^{2,k}}{4}\|\beta^{k+1} - \beta^k\| + \frac{\tau^{2,k}\|X\beta^{k+1} - X\beta^k\|^2}{2\|\beta^{k+1} - \beta^k\|}, \tag{14}$$

where $\beta^{k+1} = \mathrm{Prox}_{\lambda p/\sigma^{2,k}}(\beta^k + (\nabla q(\beta^k) - X^T u^{k+1})/\sigma^{2,k})$.

Step 2.2. If β^{k+1} satisfies a prescribed stopping criterion, terminate; otherwise update $\sigma^{2,k+1} = \rho_k \sigma^{2,k}$, $\tau^{2,k+1} = \rho_k \tau^{2,k}$ with $\rho_k \in (0, 1)$ and return to Step 2.1 with $k = k+1$.

Since $h(\beta; \sigma^1, \tau^1, 0, 0, b)$ is bounded below, it has been proven in (Flemming, 2011, Proposition 4.19) that the optimal objective value of the problem (13) will converge to the optimal objective value of the problem (5) with a difference of $q(0)$ as $\sigma^1 \to 0$, $\tau^1 \to 0$. We simply describe the convergence result of the algorithm in our first stage as follows and give a similar proof to that in (Flemming, 2011, Proposition 4.19).

Theorem 15 *Let* $\bar{h}(\sigma^1,\tau^1) := \min\limits_{\beta\in\mathbb{R}^n}\big\{h(\beta;\sigma^1,\tau^1,0,0,b)\big\}$. *Then we have*

$$\lim_{\sigma^1,\tau^1\to 0}\bar{h}(\sigma^1,\tau^1) = \min_{\beta\in\mathbb{R}^n}\Big\{\|X\beta-b\|+\lambda p(\beta)-q(0)\Big\}.$$

Proof For any $\sigma^1,\tau^1>0$ and $\beta\in\mathbb{R}^n$, we have that

$$\bar{h}(\sigma^1,\tau^1)\le\|X\beta-b\|+\lambda p(\beta)-q(0)+\frac{\sigma^1}{2}\|\beta\|^2+\frac{\tau^1}{2}\|X\beta-b\|^2.$$

Therefore, $\lim\limits_{\sigma^1,\tau^1\to 0}\bar{h}(\sigma^1,\tau^1)\le\|X\beta-b\|+\lambda p(\beta)-q(0)$. That is

$$\lim_{\sigma^1,\tau^1\to 0}\bar{h}(\sigma^1,\tau^1)\le\min_{\beta\in\mathbb{R}^n}\Big\{\|X\beta-b\|+\lambda p(\beta)-q(0)\Big\}.$$

Furthermore, $\bar{h}(\sigma^1,\tau^1)\ge\min\limits_{\beta\in\mathbb{R}^n}\Big\{\|X\beta-b\|+\lambda p(\beta)-q(0)\Big\}$. The desired result follows. ■

4.3 Convergence Analysis of the PMM Algorithm

In this subsection, we analyze the convergence of the PMM algorithm. First we recall the definition of the KL property of a function (see e.g., Attouch and Bolte, 2009; Bolte and Pauwels, 2016; Bolte et al., 2014, for more details). Let $\eta>0$ and Φ_η be the set of all concave functions $\psi:[0,\eta)\to\mathbb{R}_+$ such that

(1) $\psi(0)=0$;

(2) ψ is continuous at 0 and continuously differentiable on $(0,\eta)$;

(3) $\psi'(x)>0$, for any $x\in(0,\eta)$.

Definition 16 *Let* $f:\mathbb{R}^n\to(-\infty,\infty]$ *be a proper lower semi-continuous function and* $\bar{x}\in dom(\partial f):=\{x\in dom(f)\,|\partial f(x)\neq\emptyset\}$. *The function* f *is said to have the KL property at* $\bar{x}$ *if there exist* $\eta>0$, *a neighbourhood* $\mathcal{U}$ *of* $\bar{x}$ *and a concave function* $\psi\in\Phi_\eta$ *such that*

$$\psi'(f(x)-f(\bar{x}))dist(0,\partial f(x))\ge 1,\quad\forall\, x\in\mathcal{U}\ and\ f(\bar{x})<f(x)<f(\bar{x})+\eta,$$

where $dist(x,C):=\min_{y\in C}\|y-x\|$ *is the distance from a point* x *to a nonempty closed set* C. *Furthermore, a function* f *is called a KL function if it satisfies the KL property at all points in* $dom\partial f$.

Note that a function is said to have the KL property at $\bar{x}$ with an exponent α if the function ψ in the definition of the KL property takes the form of $\psi(x)=\gamma x^{1-\alpha}$ with $\gamma>0$ and $\alpha\in[0,1)$. For the function $f(x)=x$, it has the KL property at any point with the exponent 0.

Now we are ready to conduct the convergence analysis of the PMM algorithm. Denote $h_k(\beta):=h(\beta;\sigma^{2,k},\tau^{2,k},\beta^k,\nabla q(\beta^k),X\beta^k)$. At the k-th iteration of stage II, we have that

$$\beta^{k+1}=\operatorname*{argmin}_{\beta\in\mathbb{R}^n}\Big\{h_k(\beta)+\langle\delta^k,\beta-\beta^k\rangle\Big\}$$

such that condition (14) is satisfied. The following lemma shows the descent property of the function h_k.

Lemma 17 *Let β^{k+1} be an approximate solution of the subproblem in the k-th iteration such that (14) holds. Then we have*

$$h_k(\beta^k) \;\geq\; h_k(\beta^{k+1}) - \frac{\sigma^{2,k}}{4}\|\beta^{k+1}-\beta^k\|^2 - \frac{\tau^{2,k}}{2}\|X\beta^{k+1}-X\beta^k\|^2.$$

Proof Since h_k is a convex function and $-\delta^k \in \partial h_k(\beta^{k+1})$, we obtain

$$h_k(\beta^k) - h_k(\beta^{k+1}) \geq \langle \delta^k, \beta^{k+1}-\beta^k\rangle \geq -\frac{\sigma^{2,k}}{4}\|\beta^{k+1}-\beta^k\|^2 - \frac{\tau^{2,k}}{2}\|X\beta^{k+1}-X\beta^k\|^2.$$

The last inequality is valid since the condition (14) holds. The desired result follows. ■

Next we recall the following lemma which is similar to that in (Cui et al., 2018; Pang et al., 2017).

Lemma 18 *The vector $\bar{\beta} \in \mathbb{R}^n$ is a d-stationary point of (3) if and only if there exist $\sigma, \tau \geq 0$ such that*

$$\bar{\beta} \in \operatorname*{argmin}_{\beta\in\mathbb{R}^n} \Big\{ h(\beta; \sigma, \tau, \bar{\beta}, \nabla q(\bar{\beta}), X\bar{\beta}) \Big\}.$$

Proof Recall the objective function g defined in (3). Since g is directionally differentiable at $\bar{\beta}$, we can see that $\bar{\beta}$ being a d-stationary point of g is equivalent to $0 \in \partial g(\bar{\beta})$. It is easy to show that $\partial g(\bar{\beta}) = \partial_\beta h(\bar{\beta}; \sigma, \tau, \bar{\beta}, \nabla q(\bar{\beta}), X\bar{\beta})$. For given σ, τ and $\bar{\beta}$, the function $h(\cdot; \sigma, \tau, \bar{\beta}, \nabla q(\bar{\beta}), X\bar{\beta})$ is convex. Thus $0 \in \partial_\beta h(\bar{\beta}; \sigma, \tau, \bar{\beta}, \nabla q(\bar{\beta}), X\bar{\beta})$ is equivalent to $\bar{\beta} \in \operatorname{argmin}_{\beta\in\mathbb{R}^n} \Big\{ h(\beta; \sigma, \tau, \bar{\beta}, \nabla q(\bar{\beta}), X\bar{\beta}) \Big\}$. This completes the proof. ■

It has been proven by Cui et al. (2018) that the sequence generated by the PMM algorithm converges to a directional stationary solution if the exact solutions of the subproblems are obtained. The following theorem shows that the result is also true if the subproblems are solved approximately.

Theorem 19 *Suppose that the function g in (3) is bounded below and Assumption 1 holds. Assume that $\{\sigma^{2,k}\}$ and $\{\tau^{2,k}\}$ are convergent sequences. Let $\{\beta^k\}$ be the sequence generated by the PMM algorithm. Then every cluster point of the sequence $\{\beta^k\}$, if exists, is a d-stationary point of (3).*

Proof Combing Lemma 17 and the convexity of q, we have

$$\begin{aligned} g(\beta^k) \;&=\; h_k(\beta^k) \;\geq\; h_k(\beta^{k+1}) - \frac{\sigma^{2,k}}{4}\|\beta^{k+1}-\beta^k\|^2 - \frac{\tau^{2,k}}{2}\|X\beta^{k+1}-X\beta^k\|^2 \\ &\geq\; g(\beta^{k+1}) + \frac{\sigma^{2,k}}{4}\|\beta^{k+1}-\beta^k\|^2. \end{aligned}$$

Therefore the sequence $\{g(\beta^k)\}$ is non-increasing. Since $g(\beta)$ is bounded below, the sequence $\{g(\beta^k)\}$ converges and so is the sequence $\{\|\beta^{k+1}-\beta^k\|\}$ which converges to zero. Next, we prove that the limit of a convergent subsequence of $\{\beta^k\}$ is a d-stationary point of (3). Let

β^∞ be the limit of a convergent subsequence $\{\beta^k\}_{k\in\mathcal{K}}$. We can easily prove that $\{\beta^{k+1}\}_{k\in\mathcal{K}}$ also converges to β^∞. It follows from the definition of β^{k+1} that

$$h_k(\beta) \;\geq\; h_k(\beta^{k+1}) + \langle \delta^k, \beta^{k+1} - \beta\rangle \geq h_k(\beta^{k+1}) - \|\delta^k\|\|\beta^{k+1} - \beta\|, \quad \forall\, \beta \in \mathbb{R}^m.$$

Letting $k(\in \mathcal{K}) \to \infty$, we obtain that

$$\beta^\infty \in \operatorname*{argmin}_{\beta\in\mathbb{R}^n} \Big\{ h(\beta; \sigma^{2,\infty}, \tau^{2,\infty}, \beta^\infty, \nabla q(\beta^\infty), X\beta^\infty) \Big\},$$

where $\sigma^{2,\infty} = \lim\limits_{k\to\infty} \sigma^{2,k} \geq 0$ and $\tau^{2,\infty} = \lim\limits_{k\to\infty} \tau^{2,k} \geq 0$. The desired result follows from Lemma 18. This completes the proof. ■

We can also establish the local convergence rate of the sequence $\{\beta^k\}$ under either an isolation assumption of the accumulation point or the KL property assumption.

Theorem 20 *Suppose that the function g is bounded below and Assumption 1 holds. Let $\{\beta^k\}$ be the sequence generated by the PMM algorithm. Let $\mathcal{B}^\infty$ be the set of cluster points of the sequence $\{\beta^k\}$. If either one of the following two conditions holds,*

(a) $\mathcal{B}^\infty$ contains an isolated element;

(b) The sequence $\{\beta^k\}$ is bounded; for all $\beta \in \mathcal{B}^\infty$, ∇q is locally Lipschitz continuous near β and the function g has the KL property at β;

then the whole sequence $\{\beta^k\}$ converges to an element of $\mathcal{B}^\infty$. Moreover, if condition (b) is satisfied such that $\{\beta^k\}$ converges to $\beta^\infty \in \mathcal{B}^\infty$ and the function g has the KL property at β^∞ with an exponent $\alpha \in [0, 1)$, then we have the following results:

(i) If $\alpha = 0$, then the sequence $\{\beta^k\}$ converges in a finite number of steps;

(ii) If $\alpha \in (0, \frac{1}{2}]$, then the sequence $\{\beta^k\}$ converges R-linearly, that is, for all $k \geq 1$, there exist $\nu > 0$ and $\eta \in [0, 1)$ such that $\|\beta^k - \beta^\infty\| \leq \nu\eta^k$;

(iii) If $\alpha \in (\frac{1}{2}, 1)$, then the sequence $\{\beta^k\}$ converges R-sublinearly, that is, for all $k \geq 1$, there exists $\nu > 0$ such that $\|\beta^k - \beta^\infty\| \leq \nu k^{-\frac{1-\alpha}{2\alpha-1}}$.

Proof We know from Theorem 19 that $\lim\limits_{k\to\infty} \|\beta^{k+1} - \beta^k\| = 0$. Then it follows from (Facchinei and Pang, 2003, Proposition 8.3.10) that the sequence $\{\beta^k\}$ converges to an isolated element of $\mathcal{B}^\infty$ under the condition (a). In order to derive the convergence rate of the sequence $\{\beta^k\}$ under the condition (b), we first establish some properties of the sequence $\{\beta^k\}$, i.e.,

(1) $g(\beta^k) \geq g(\beta^{k+1}) + \frac{\sigma^{2,k}}{4}\|\beta^{k+1} - \beta^k\|^2$;

(2) there exists a subsequence $\{\beta^{k_j}\}$ of $\{\beta^k\}$ such that $\beta^{k_j} \to \beta^\infty$ with $g(\beta^{k_j}) \to g(\beta^\infty)$ as $j \to \infty$;

(3) for k sufficient large, there exist a constant $K > 0$ and $\xi^{k+1} \in \partial g(\beta^{k+1})$ such that $\|\xi^{k+1}\| \leq K\|\beta^{k+1} - \beta^k\|$.

The properties (1) and (2) are already known from Theorem 19. To establish the property (3), we first note that $\mathcal{B}^\infty$ is a nonempty, compact and connected set by (Facchinei and Pang, 2003, Proposition 8.3.9). Furthermore, let $\xi^{k+1} = \nabla q(\beta^k) - \nabla q(\beta^{k+1}) - \sigma^{2,k}(\beta^{k+1} - \beta^k) - \tau^{2,k} X^T X(\beta^{k+1} - \beta^k) - \delta^k$. We have that $\xi^{k+1} \in \partial g(\beta^{k+1})$. Since ∇q is locally Lipschitz continuous near all $\beta \in \mathcal{B}^\infty$ and $\|\delta^k\| \le \frac{\sigma^{2,k}}{4}\|\beta^{k+1} - \beta^k\| + \frac{\tau^{2,k}\|X\beta^{k+1} - X\beta^k\|^2}{2\|\beta^{k+1} - \beta^k\|}$, the property (3) holds for some constant $K > 0$ with $\|\xi^{k+1}\| \le K\|\beta^{k+1} - \beta^k\|$ for k sufficiently large. With the properties (1)-(3), the convergence rate of the sequence $\{\beta^k\}$ can be established similarly to that of (Bolte and Pauwels, 2016, Proposition 4). ■

5. Numerical Experiments

In this section, we use some numerical experiments to demonstrate the efficiency of our PMM algorithm for the square-root regression problems. We implemented the algorithm in MATLAB R2017a. All runs were performed on a PC (Intel Core 2 Duo 2.6 GHz with 4 GB RAM). We tested our algorithm on two types of data sets. The first set consists of synthetic data generated randomly in the high-sample-low-dimension setting. That is,

$$b = X\ddot{\beta} + \varsigma\varepsilon, \quad \varepsilon \sim N(0, I).$$

Each row of the input data $X \in \mathbb{R}^{m\times n}$ is generated randomly from the multivariate normal distribution $N(0, \Sigma)$ with Σ as the covariance matrix. Now we present four examples which are similar to that in (Zou and Hastie, 2005). For each instance, we generate 8000 observations for the training data set and 2000 observations for the validation data set.

(a) In example 1, the problem has 800 predictors. Let $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)$ and $\ddot{\beta} = (\underbrace{\beta, \ldots, \beta}_{100})^T$. The parameter ς is set to 3 and the pairwise correlation between the i-th predictor and the j-th predictor is set to be $\Sigma_{ij} = 0.5^{|i-j|}$.

(b) In example 2, the setting is the same as that in example 1 except that $\ddot{\beta} = (\underbrace{\beta, \ldots, \beta}_{400})^T$ with the vector $\beta = (0, 1)$.

(c) In example 3, we set $\ddot{\beta} = (\underbrace{\beta, \ldots, \beta}_{200})^T$ with the vector $\beta = (0, 1)$, $\varsigma = 15$ and $\Sigma_{ij} = 0.8^{|i-j|}$.

(d) In example 4, the problem has 800 predictors. We choose $\ddot{\beta} = (\underbrace{3, \ldots, 3}_{300}, \underbrace{0, \ldots, 0}_{500})$ and $\varsigma = 3$. Let X_i be the i-th predictor of X. For $i \le 300$, X_i is generated as follows:

$$\begin{aligned} X_i &= Z_1 + \tilde{\varepsilon}_i, \quad Z_1 \sim N(0, I), \quad i = 1, \ldots, 100, \\ X_i &= Z_2 + \tilde{\varepsilon}_i, \quad Z_2 \sim N(0, I), \quad i = 101, \ldots, 200, \\ X_i &= Z_3 + \tilde{\varepsilon}_i, \quad Z_3 \sim N(0, I), \quad i = 201, \ldots, 300, \end{aligned}$$

with $\tilde{\varepsilon}_i \sim N(0, 0.01I)$, $i = 1, \ldots, 300$. For $i > 300$, the predictor X_i is just white noise, i.e., $X_i \sim N(0, I)$.

We also evaluate our algorithm on some large scale LIBSVM data sets (X, b) (Chang and Lin, 2011) which are obtained from the UCI data repository (Lichman, 2013). As in (Li et al., 2018), we use the method by Huang et al. (2010) to expand the features of these data sets by using polynomial basis functions. The last digit in the names of the data sets, abalone7, bodyfat7, housing7, mpg7 and space9, indicate the order of the polynomial used to expand the features. The number of nonzero elements of a vector is defined as the minimal k such that

$$\sum_{i=1}^{k} |\check{\beta}_i| \geq 0.9999\|\beta\|_1,$$

where $\check{\beta}$ is obtained by sorting β such that $|\check{\beta}_1| \geq |\check{\beta}_2| \geq \ldots \geq |\check{\beta}_n|$.

In all the experiments, the parameter λ is set to $\lambda = \lambda_c \Lambda$, where $\Lambda = 1.1\Phi^{-1}(1 - 0.05/(2n))$ with Φ being the cumulative normal distribution function and Λ is the theoretical choice recommended by Belloni et al. (2011) to compute a specific coefficient estimate. For all the tables in the following sections, we use "s sign$(t)|t|$" to denote a number of the form "$s \times 10^t$", e.g., 1.0-2 denotes 1.0×10^{-2}.

5.1 Numerical Experiments for the Convex Square-Root Regression Problems

In this section, we compare the performances of the alternating direction method of multipliers (ADMM) and our stage I algorithm for solving the convex square-root regression problem (5). For comparison purpose, we adopt the widely used ADMM algorithms for both the primal and dual problems of (5). For convenience, we use pADMM to denote the ADMM applied to the primal problem, dADMM to denote the ADMM applied to the dual problem, and PMM to denote our stage I algorithm for solving the convex square-root regression problem (5).

5.1.1 The ADMM Algorithm for the Problem (5)

In this subsection, we describe the implementation details of the ADMM for the problem (5). The convex problem (5) can be written equivalently as

$$\min_{\beta,z\in\mathbb{R}^n, y\in\mathbb{R}^m} \left\{ \|y\| + \lambda p(z) \,\middle|\, X\beta - y = b, \beta - z = 0 \right\}. \tag{15}$$

The dual problem corresponding to (15) has the following form

$$\min_{u,w\in\mathbb{R}^m, v\in\mathbb{R}^n} \left\{ \delta_B(w) + (\lambda p)^*(v) + \langle u, b\rangle \,\middle|\, X^T u + v = 0, -u + w = 0 \right\}. \tag{16}$$

Given $\zeta > 0$, the augmented Lagrangian functions corresponding to (15) and (16) are given by

$$\begin{aligned}
\mathcal{L}_\zeta(\beta, y, z; u, v) &:= \|y\| + \lambda p(z) + \langle u, X\beta - y - b\rangle + \frac{\zeta}{2}\|X\beta - y - b\|^2 \\
&\quad + \langle v, \beta - z\rangle + \frac{\zeta}{2}\|\beta - z\|^2, \\
\tilde{\mathcal{L}}_\zeta(u, v, w; \beta, y) &:= \delta_B(w) + (\lambda p)^*(v) + \langle u, b\rangle - \langle \beta, X^T u + v\rangle + \frac{\zeta}{2}\|X^T u + v\|^2 \\
&\quad - \langle y, -u + w\rangle + \frac{\zeta}{2}\| - u + w\|^2,
\end{aligned}$$

respectively. Based on the above augmented Lagrangian functions, the ADMMs (see e.g., Eckstein and Bertsekas, 1992; Gabay and Mercier, 1976) for solving (15) and (16) are given as below.

In the pADMM and dADMM, we set the parameter $\rho = 1.618$ and solve the linear system in Step 1 of pADMM and dADMM by using the Sherman-Morrison-Woodbury formula (Golub and Van Loan, 1996) if it is necessary, i.e.,

$$\begin{aligned}
\left(I_n + X^T X\right)^{-1} &= I_m - X\left(I_m + XX^T\right)^{-1} X^T, \\
\left(I_m + XX^T\right)^{-1} &= I_n - X^T\left(I_n + X^T X\right)^{-1} X.
\end{aligned}$$

Depending on the dimension n, m of the problem, we either solve the linear system with coefficient matrix $I_m + XX^T$ (or $I_n + X^T X$) by Cholesky factorization or by an iterative solver such as the preconditioned conjugate gradient (PCG) method. We should mention that when the latter approach is used, the linear system only needs to be solved to a sufficient level of accuracy that depend on the progress of the algorithm without sacrificing the convergence of the ADMMs. For the details, we refer the reader to Chen et al. (2017).

Algorithm pADMM for the primal problem (15). Let $\rho \in (0, (1+\sqrt{5})/2)$, $\zeta > 0$ be given parameters. Choose $(y^0, z^0, u^0, v^0) \in \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^n$, set $k = 0$ and iterate.

Step 1. Compute

$$\begin{aligned}
\beta^{k+1} &= \operatorname*{argmin}_{\beta\in\mathbb{R}^n}\left\{\mathcal{L}_\zeta(\beta, y^k, z^k; u^k, v^k)\right\} \\
&= (I_n + X^T X)^{-1}(z^k - \zeta^{-1}v^k + X^T(y^k + b - \zeta^{-1}u^k)), \\
(y^{k+1}, z^{k+1}) &= \operatorname*{argmin}_{y\in\mathbb{R}^m, z\in\mathbb{R}^n}\left\{\mathcal{L}_\zeta(\beta^{k+1}, y, z; u^k, v^k)\right\} \\
&= \left(\mathrm{Prox}_{\zeta^{-1}\|\cdot\|}(X\beta^{k+1} - b + \zeta^{-1}u^k), \mathrm{Prox}_{\zeta^{-1}\lambda p}(\beta^{k+1} + \zeta^{-1}v^k)\right).
\end{aligned}$$

Step 2. Update

$$u^{k+1} = u^k + \rho\zeta(X\beta^{k+1} - y^{k+1} - b),\ v^{k+1} = v^k + \rho\zeta(\beta^{k+1} - z^{k+1}).$$

If the prescribed stopping criterion is satisfied, terminate; otherwise return to Step 1 with $k = k + 1$.

Algorithm dADMM for the dual problem (16). Let $\rho \in (0, (1+\sqrt{5})/2)$, $\zeta > 0$ be given parameters. Choose $(v^0, w^0, \beta^0, y^0) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}^m$, set $k = 0$ and iterate.

Step 1. Compute

$$\begin{aligned} u^{k+1} &= \underset{u\in\mathbb{R}^m}{\operatorname{argmin}} \left\{ \tilde{\mathcal{L}}_\zeta(u, v^k, w^k; \beta^k, y^k) \right\} \\ &= (I_m + XX^T)^{-1}(w^k - \zeta^{-1}y^k + X(-v^k + \zeta^{-1}\beta^k) - b), \\ (v^{k+1}, w^{k+1}) &= \underset{v\in\mathbb{R}^n, w\in\mathbb{R}^m}{\operatorname{argmin}} \left\{ \tilde{\mathcal{L}}_\zeta(u^{k+1}, v, w; \beta^k, y^k) \right\} \\ &= \left(\operatorname{Prox}_{\zeta^{-1}(\lambda p)^*}(\zeta^{-1}\beta^k - X^T u^{k+1}), \operatorname{Prox}_{\zeta^{-1}\delta_B}(\zeta^{-1}y^k + u^{k+1}) \right). \end{aligned}$$

Step 2. Update

$$\beta^{k+1} = \beta^k - \rho\zeta(X^T u^{k+1} + v^{k+1}),\ y^{k+1} = y^k - \rho\zeta(-u^{k+1} + w^{k+1}).$$

If the prescribed stopping criterion is satisfied, terminate; otherwise return to Step 1 with $k = k + 1$.

5.1.2 Stopping Criteria

In order to measure the accuracy of an approximate optimal solution β, we use the relative duality gap defined by

$$\eta_G := \frac{|\text{pobj} - \text{dobj}|}{1 + |\text{pobj}| + |\text{dobj}|},$$

where $\text{pobj} := \|X\beta - b\| + \lambda p(\beta)$, $\text{dobj} := -\langle u, b\rangle$ are the primal and dual objective values, respectively. We also adopt the relative KKT residual

$$\eta_{\text{kkt}} := \frac{\left\| \beta - \operatorname{Prox}_{\lambda p}\left(\beta - \frac{X^T(X\beta - b)}{\|X\beta - b\|} \right) \right\|}{1 + \|\beta\| + \frac{\|X^T(X\beta - b)\|}{\|X\beta - b\|}}$$

to measure the accuracy of an approximate optimal solution β. For a given tolerance, our stage I algorithm is terminated if

$$\eta_{\text{kkt}} < \epsilon_{\text{kkt}} = 10^{-6}, \tag{17}$$

or the number of iterations reaches the maximum of 200 while the ADMMs are terminated if (17) is satisfied or the number of iterations reaches the maximum of 10000. All the algorithms are stopped if they reach the pre-set maximum running time of 4 hours.

5.1.3 Numerical Results for the srLasso Problem (2)

Here we compare the performance of different methods for solving the convex problem (2). In (Stucky and van de Geer, 2017), it adopted the R package Flare (Li et al., 2015) to

solve the srLasso problem (2). As the algorithm in Flare is in fact the pADMM with unit steplength, we first compare our own implementation of the pADMM with that in the Flare package. For a fair comparison, our pADMM is also written in R. Since the stopping criterion of the Flare package is not stated explicitly, we first run the Flare package to obtain a primal objective value and then run our pADMM, which is terminated as soon as our primal objective value is smaller than that obtained by Flare. We note that since (2) is an unconstrained optimization problem, it is meaningful to compare the objective function values obtained by Flare and our pADMM.

We report the numerical results in Tables 1 and 2. We report the problem name (probname), the number of samples (m) and features (n), λ_c, the primal objective value (pobj), and the computation time (time) in the format of "hours:minutes:seconds". The symbol "–" in Table 2 means that the Flare package fails to solve the problem due to excessive memory requirement. From Tables 1 and 2, we can observe that our pADMM is clearly faster than the Flare package. A possible cause of this difference may lie in the different strategies for dynamically updating the parameter ζ in the practical implementations of the pADMM. As our implementation of the pADMM is much more efficient than that in the Flare package, in the subsequent experiments, we will not compare the performance of our PMM algorithm with the Flare package but with our own pADMM.

Table 1: The performance of the Flare package and our pADMM on synthetic datasets for the srLasso problem.

probname m; n	λ_c	pobj		time	
		Flare	pADMM	Flare	pADMM
exmp1	1.0	3.8876+3	3.5799+3	11:26	12
8000;800	0.5	3.0501+3	1.9174+3	21:09	13
	0.1	1.0487+3	5.8738+2	28:42	16
exmp2	1.0	2.2422+3	2.2419+3	14:09	19
8000;800	0.5	1.8050+3	1.2811+3	27:18	11
	0.1	5.6150+2	4.6013+2	27:37	09
exmp3	1.0	2.4758+3	2.4569+3	10:05	07
8000;400	0.5	1.9819+3	1.9421+3	7:26	07
	0.1	1.4888+3	1.4438+3	7:14	05
exmp4	1.0	1.1210+4	1.1205+4	29:11	20:16
8000;4000	0.5	1.0165+4	1.0165+4	1:43:48	21:48
	0.1	7.6846+3	3.4069+3	3:11:27	5:12

Next we conduct numerical experiments to evaluate the performance of the pADMM, dADMM and PMM. For the numerical results, besides the results reported in Tables 1 and 2, we also report the relative KKT residual ($\eta_{\rm kkt}$), the relative duality gap (η_G), the number of nonzero elements of β (nnz), the mean square error defined by $\|\beta - \ddot{\beta}\|^2/n$ (MSE), and the percentage (P) of the nonzero positions of $\ddot{\beta}$ that are picked up by β. The last three results were obtained from the PMM algorithm. In the implementation of the pADMM and dADMM, we first compute the (sparse) Cholesky decomposition of $I_n + X^T X$ or $I_m + XX^T$ and then solve the linear system of equations in each iteration by using the pre-computed Cholesky factor.

Tables 3 and 4 show the performance of the three algorithms. For the synthetic datasets, the pADMM is more efficient than the dADMM in almost all cases. Furthermore, we can

Table 2: The performance of the Flare package and our pADMM on UCI datasets for the srLasso problem.

probname m; n	λ_c	pobj		time	
		Flare	pADMM	Flare	pADMM
abalone.scale.expanded7	1.0	–	2.3852+2	–	25:57
4177;6435	0.5	–	2.0312+2	–	25:32
	0.1	–	1.5586+2	–	26:29
mpg.scale.expanded7	1.0	2.3550+2	2.3544+2	1:00	04
392;3432	0.5	1.5856+2	1.5831+2	57	03
	0.1	7.8656+1	7.8616+1	1:06	03
space.ga.scale.expanded9	1.0	1.3113+1	1.3113+1	12:59	5:19
3107;5005	0.5	2.2419+1	2.1607+1	9:01	2:00
	0.1	1.2950+1	1.1999+1	6:13	3:00

see that our PMM algorithm can solve all the problems to the required accuracy. The PMM algorithm not only takes much less time than the pADMM or dADMM does but also obtains more accurate solutions (in terms of $\eta_{\rm kkt}$) in almost all cases.

5.2 Numerical Experiments for the Square-Root Regression Problems with Nonconvex Regularizers

In this section, we compare the performance of the ADMM and our PMM algorithm for solving the nonconvex square-root regression problem (3). The relative KKT residual

$$\tilde{\eta}_{\rm kkt} := \frac{\left\|\beta - {\rm Prox}_{\lambda p-q}\left(\beta - \frac{X^T(X\beta-b)}{\|X\beta-b\|}\right)\right\|}{1+\|\beta\|+\frac{\|X^T(X\beta-b)\|}{\|X\beta-b\|}}$$

is adopted to measure the accuracy of an approximate optimal solution β. In our PMM algorithm, stage I is implemented to generate an initial point for stage II and is stopped if $\eta_{\rm kkt} < 10^{-4}$. The tested algorithms will be terminated if $\tilde{\eta}_{\rm kkt} < \tilde{\epsilon}_{\rm kkt} = 10^{-6}$. In addition, the algorithms are also stopped when they reach the pre-set maximum number of iterations (200 for stage II of the PMM and 10000 for the ADMM) or the pre-set maximum running time of 4 hours. For each synthetic data set, the models are fitted on the training data set and the validation data set is used to select the regularization parameter λ_c. For each UCI data set, we adopt a tenfold cross validation to find the regularization parameter. The PMM algorithm is used to perform the cross validation.

5.2.1 The ADMM Algorithm for the Problem (3)

To describe the ADMM implemented (which is not guaranteed to converge though due to the nonconvexity) for solving the nonconvex square-root regression problem (3), we first reformulate it to the following constrained problem:

$$\min_{\beta,z\in\mathbb{R}^n,y\in\mathbb{R}^m}\left\{\|y\|+\lambda p(z)-q(z)\,\middle|\,X\beta-y=b,\beta-z=0\right\}. \tag{18}$$

Table 3: The performance of different algorithms on synthetic datasets for the srLasso problem. In the table, "a"=PMM, "b"=pADMM, "c"=dADMM.

probname m; n	λ_c	nnz	$\eta_{\rm kkt}$			η_G			pobj			time			MSE			P
			a	b	c	a	b	c	a	b	c	a	b	c	a	b	c	
exmp1 8000;800	1	305	8.6-7	1.0-6	9.7-7	8.2-10	3.6-6	1.4-9	3.3895+3	3.3895+3	3.3895+3	12	1:03	5:09	1.4882-2	1.4882-2	1.4882-2	100%
exmp2 8000;800	1	437	4.4-7	1.0-6	1.0-6	1.8-9	1.1-6	2.9-7	2.1067+3	2.1067+3	2.1067+3	17	1:10	5:21	2.9740-2	2.9740-2	2.9740-2	100%
exmp3 8000;400	1	277	4.0-7	1.0-6	1.0-6	6.8-10	1.5-6	1.1-8	2.2169+3	2.2169+3	2.2169+3	07	1:16	1:01	1.2955-1	1.2955-1	1.2955-1	98%
exmp4 8000;800	1	300	7.6-7	9.9-7	1.0-6	3.1-9	1.3-6	4.1-7	4.5283+3	4.5283+3	4.5283+3	12	2:36	4:47	2.5358-1	2.5358-1	2.5358-1	100%

Table 4: The performance of different algorithms on UCI datasets for the srLasso problem. In the table, "a"=PMM, "b"=pADMM, "c"=dADMM.

probname m; n	λ_c	nnz	$\eta_{\rm kkt}$			η_G			pobj			time		
			a	b	c	a	b	c	a	b	c	a	b	c
E2006.test 3308;150358	1	1	9.4-9	9.8-7	9.1-7	1.5-6	1.1-5	5.3-10	2.6706+1	2.6706+1	2.6706+1	10	03	03
log1p.E2006.test 3308;1771946	1	5	9.8-7	*1.2-4*	*5.2-3*	2.7-4	8.8-2	1.3-3	2.6046+1	2.8627+1	2.6046+1	1:06	1:10:39	49:11
E2006.train 16087;150358	1	1	3.8-7	8.0-7	9.6-7	2.2-6	3.1-6	2.1-10	5.4180+1	5.4180+1	5.4180+1	26	35	44
log1p.E2006.train 16087;4272227	1	22	4.5-7	*1.6-1*	*7.6-1*	5.9-4	9.9-1	5.6-1	5.2032+1	6.9860+5	8.1873+2	4:27	4:00:19	4:01:26
abalone.scale.expanded7 4177;6435	1	6	9.7-7	*9.6-6*	*7.5-3*	9.9-8	1.1-3	1.3-2	2.3562+2	2.3589+2	2.3575+2	04	20:10	17:57
housing.scale.expanded7 506;77520	1	22	7.7-7	*2.7-6*	*3.7-4*	4.1-6	1.2-3	1.5-3	2.6957+2	2.6989+2	2.6957+2	07	40:35	22:48
mpg.scale.expanded7 392;3432	1	5	9.6-7	1.0-6	1.0-6	1.6-8	4.0-5	1.2-6	2.1320+2	2.1321+2	2.1320+2	01	18	26
space.ga.scale.expanded9 3107;5005	1	4	5.5-7	1.0-6	9.6-7	8.2-8	2.7-4	1.1-7	1.3111+1	1.3115+1	1.3111+1	03	1:57	56
pyrim.scale.expanded5 74;201376	1	23	8.0-7	*5.6-6*	*8.5-2*	1.8-4	7.4-2	2.1-2	3.3790+0	3.7206+0	3.7958+0	07	15:11	9:39
bodyfat.scale.expanded7 252;116280	1	2	7.0-7	*1.9-6*	1.0-6	1.3-5	2.1-2	2.5-8	4.5326+0	4.6458+0	4.5326+0	07	30:53	6:18

For $\zeta > 0$, the augmented Lagrangian function of (18) can be written as

$$\begin{aligned} L_\zeta(\beta, y, z; u, v) \;:=\; & \|y\| + \lambda p(z) - q(z) + \langle u, X\beta - y - b\rangle + \frac{\zeta}{2}\|X\beta - y - b\|^2 \\ & + \langle v, \beta - z\rangle + \frac{\zeta}{2}\|\beta - z\|^2. \end{aligned}$$

The template of the ADMM for solving the problem (3) is given as the following form.

Algorithm ADMM for the problem (3). Let $\zeta > 0$ be a given parameter. Choose $(y^0, z^0, u^0, v^0) \in \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^n$, set $k = 0$ and iterate.

Step 1. Compute

$$\begin{aligned} \beta^{k+1} &= \underset{\beta\in\mathbb{R}^n}{\operatorname{argmin}} \Big\{ L_\zeta(\beta, y^k, z^k; u^k, v^k) \Big\} \\ &= (I_n + X^T X)^{-1}(z^k - \zeta^{-1}v^k + X^T(y^k + b - \zeta^{-1}u^k)), \\ (y^{k+1}, z^{k+1}) &= \underset{y\in\mathbb{R}^m, z\in\mathbb{R}^n}{\operatorname{argmin}} \Big\{ L_\zeta(\beta^{k+1}, y, z; u^k, v^k) \Big\} \\ &= \Big(\mathrm{Prox}_{\zeta^{-1}\|\cdot\|}(X\beta^{k+1} - b + \zeta^{-1}u^k), \mathrm{Prox}_{\zeta^{-1}(\lambda p - q)}(\beta^{k+1} + \zeta^{-1}v^k) \Big). \end{aligned}$$

Step 2. Update

$$u^{k+1} = u^k + \zeta(X\beta^{k+1} - y^{k+1} - b),\ v^{k+1} = v^k + \zeta(\beta^{k+1} - z^{k+1}).$$

If the prescribed stopping criterion is satisfied, terminate; otherwise return to Step 1 with $k = k + 1$.

5.2.2 Numerical Experiments for the Square-Root Regression Problems with SCAD Regularizations

The SCAD regularization involves a concave function p_λ, proposed in (Fan and Li, 2001), that has the following properties: $p_\lambda(0) = 0$ and for $|t| > 0$,

$$p'_\lambda(|t|) = \begin{cases} \lambda, & \text{if } |t| \le \lambda, \\ \frac{(a_s\lambda - |t|)_+}{a_s - 1}, & \text{otherwise}, \end{cases}$$

for some given parameter $a_s > 2$. In the above, $(a_s\lambda - |t|)_+$ denotes the positive part of $a_s\lambda - |t|$. We can reformulate the expression of the SCAD regularization function as $\lambda p(\beta) - q(\beta)$ with $p(\beta) = \|\beta\|_1$ and

$$q(\beta) = \sum_{i=1}^n q^{\mathrm{scad}}(\beta_i; a_s, \lambda),\ q^{\mathrm{scad}}(t; a_s, \lambda) = \begin{cases} 0, & \text{if} \quad |t| \le \lambda, \\ \frac{(|t| - \lambda)^2}{2(a_s - 1)}, & \text{if} \quad \lambda \le |t| \le a_s\lambda, \\ \lambda|t| - \frac{a_s + 1}{2}\lambda^2, & \text{if} \quad |t| > a_s\lambda. \end{cases}$$

The function $q(\beta)$ is continuously differentiable with

$$\frac{\partial q(\beta)}{\partial \beta_i} = \begin{cases} 0, & \text{if } |\beta_i| \le \lambda, \\ \frac{\text{sign}(\beta_i)(|\beta_i|-\lambda)}{a_s-1}, & \text{if } \lambda < |\beta_i| \le a_s\lambda, \\ \lambda \text{sign}(\beta_i), & \text{if } |\beta_i| > a_s\lambda. \end{cases}$$

We can see that the SCAD regularization function associated with β_i is increasing and concave in $[0, +\infty)$. It has been shown by Fan and Li (2001) that the SCAD regularization usually performs better than the classical ℓ_1 regularization in selecting significant variables without creating excessive biases.

The performance of the PMM algorithm and ADMM for the SCAD regularization with $a_s = 3.7$ are listed in Tables 5 and 6. We can see that in most cases, the PMM algorithm is not only much more efficient than the ADMM, but it can also obtain better objective function values. Although the objective value of the ADMM is less than that of the PMM algorithm in the housing.scale.expanded7 data set, the solution of the PMM algorithm is more sparse than that of the ADMM with nnz being 62 versus 68777. Figure 1 shows the log-log curves of the KKT residuals and the MSEs versus the iteration counts for the SCAD regularized problem on the first two random data sets, while Figure 2 shows the log-log curves of the KKT residuals versus the iteration counts for the SCAD regularized problem on the abalone.scale.expanded7 and housing.scale.expanded7 data sets. We observe that both the PMM and ADMM algorithms achieved about the same level of MSE.

5.2.3 Numerical Experiments for the Square-Root Regression Problems with MCP Regularizations

In this subsection, we consider the regularization by a minimax concave penalty (MCP) function (Zhang, 2010). For two positive parameters $a_m > 2$ and λ, the MCP regularization can be defined as $\lambda p(\beta) - q(\beta)$ with $p(\beta) = 2\|\beta\|_1$ and

$$q(\beta) = \sum_{i=1}^{n} q^{\text{mcp}}(\beta_i; a_m, \lambda),\ q^{\text{mcp}}(t; a_m, \lambda) = \begin{cases} \frac{t^2}{a_m}, & \text{if } |t| \le a_m\lambda, \\ 2\lambda|t| - a_m\lambda^2, & \text{if } |t| > a_m\lambda. \end{cases}$$

The function $q(\beta)$ is continuously differentiable with its derivative given by

$$\frac{\partial q(\beta)}{\partial \beta_i} = \begin{cases} \frac{2\beta_i}{a_m}, & \text{if } |\beta_i| \le a_m\lambda, \\ 2\lambda \text{sign}(\beta_i), & \text{if } |\beta_i| > a_m\lambda. \end{cases}$$

We evaluate the performance of our PMM algorithm on the same set of problems as in the last subsection with the MCP regularization. The numerical results are presented in Tables 7 and 8. In this case, we set the parameter $a_m = 3.7$.

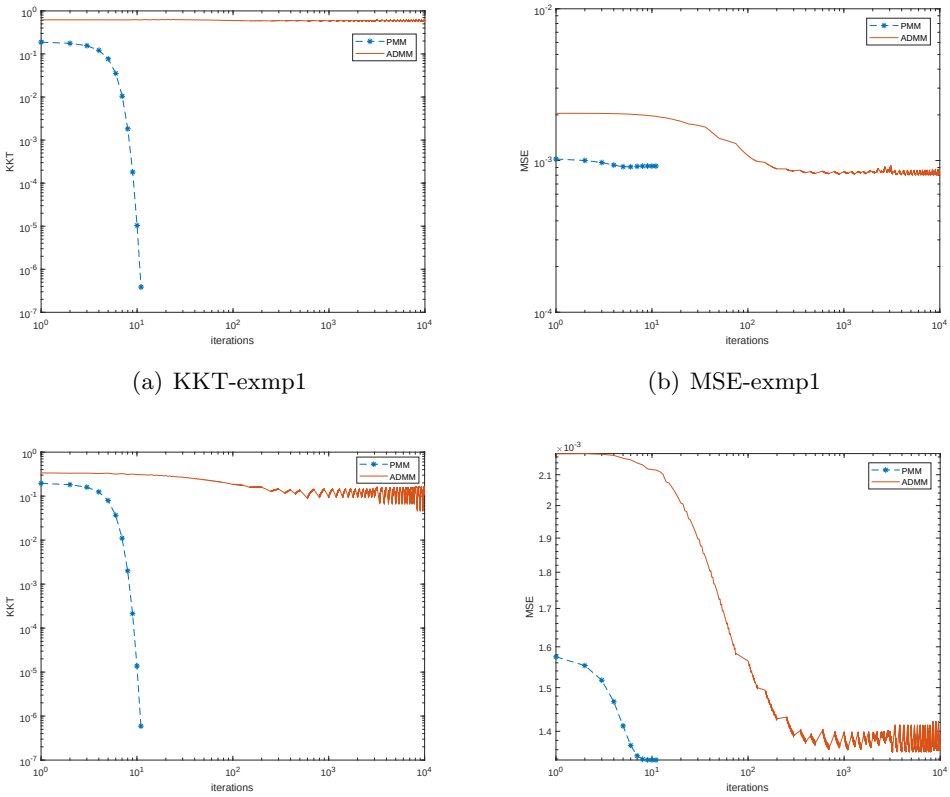
From the numerical results, one can see the efficiency and power of our SSN method based PMM algorithm. Note that though for the abalone.scale.expanded7 data set the objective value obtained by the ADMM is less than that obtained the PMM algorithm, the solution obtained by the PMM algorithm is more sparse than that by the ADMM with nnz being 55 versus 1039. Overall, our PMM algorithm is clearly more efficient and accurate than the ADMM on the tested datasets. Figures 3 and 4 are the same as Figures 1 and 2

Table 5: The performance of the ADMM and PMM on synthetic datasets for the SCAD regularization. In the table, "a"=PMM, "b"=ADMM.

probname m; n	λ_c	nnz	η_{kkt}		pobj		time		MSE		P
			a	b	a	b	a	b	a	b	
exmp1 8000;800	0.145	460	3.9-7	*5.9-1*	5.9368+2	5.9392+2	20	3:39	9.1948-4	9.4117-4	100%
exmp2 8000;800	0.087	616	5.9-7	*1.0-1*	4.0760+2	4.0777+2	28	3:33	1.3380-3	1.3737-3	100%
exmp3 8000;400	0.230	293	7.8-7	*2.7-1*	1.5486+3	1.5529+3	10	2:02	9.8282-2	1.0018-1	100%
exmp4 800;4000	0.184	451	6.4-7	*7.0-1*	8.4087+2	8.4154+2	15	4:45	5.7208-4	6.0617-4	100%

Table 6: The performance of the ADMM and PMM on UCI datasets for the SCAD regularization. In the table,"a"=PMM, "b"=ADMM.

probname m; n	λ_c	nnz	η_{kkt}		pobj		time	
			a	b	a	b	a	b
E2006.test 3308;150358	0.071	1	2.2-8	9.0-7	2.2165+1	2.2165+1	08	12:51
log1p.E2006.test 3308;1771946	0.257	207	2.1-7	*5.9-3*	2.1613+1	2.1366+2	3:50	2:36:14
E2006.train 16087;150358	0.021	1	5.1-7	*9.8-1*	4.8922+1	4.9442+1	12	3:10:28
log1p.E2006.train 16087;4272227	0.562	65	2.7-7	*9.6-1*	4.9516+1	2.8862+2	6:49	4:02:03
abalone.scale.expanded7 4177;6435	0.011	49	9.9-7	*6.9-1*	1.3292+2	1.3864+2	12	21:41
housing.scale.expanded7 506;77520	0.070	62	2.1-7	*5.0-1*	6.1449+1	5.7203+1	25	30:37
mpg.scale.expanded7 392;3432	0.107	27	3.1-9	*4.9-1*	5.5558+1	5.9918+1	01	1:38
space.ga.scale.expanded9 3107;5005	0.043	16	1.9-7	*4.4-1*	6.9072+0	9.0447+0	03	12:50
pyrim.scale.expanded5 74;201376	0.109	70	1.4-7	*4.3-3*	6.8301-1	7.2608-1	13	21:26
bodyfat.scale.expanded7 252;116280	0.201	2	3.9-8	*7.6-2*	9.4125-1	9.5136-1	06	25:51



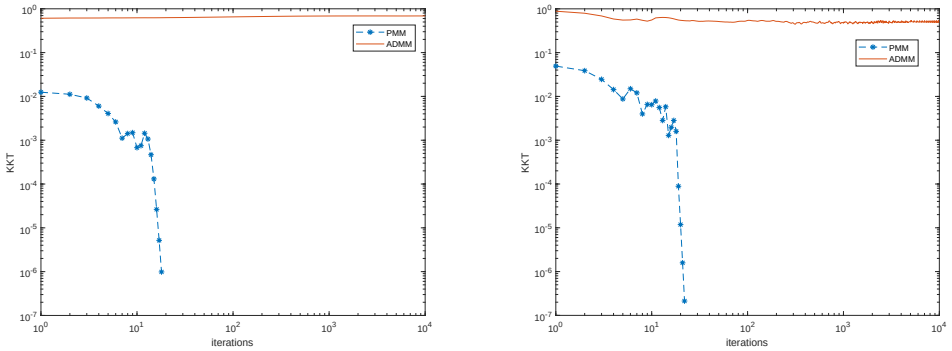
(a) KKT-exmp1 (b) MSE-exmp1

(c) KKT-exmp2 (d) MSE-exmp2

Figure 1: The KKT residuals and the MSEs of the PMM and ADMM algorithms for solving the SCAD regularized problem with the first two random data sets.

Table 7: The performance of the ADMM and PMM on synthetic datasets for the MCP regularization. In the table, "a"=PMM, "b"=ADMM.

probname m; n	λ_c	nnz	$\eta_{\rm kkt}$		pobj		time		MSE		P
			a	b	a	b	a	b	a	b	
exmp1 8000;800	0.209	380	5.2-8	*1.9-2*	5.5695+2	5.6091+2	29	3:33	7.0382-4	1.1677-3	100%
exmp2 8000;800	0.151	535	2.7-7	*1.3-1*	4.5225+2	4.5414+2	38	3:30	9.8362-4	1.6202-3	100%
exmp3 8000;400	0.081	267	9.3-7	*1.5-1*	1.3590+3	1.3617+3	1:11	2:01	1.5072-1	1.7776-1	98%
exmp4 800;4000	0.321	334	7.7-7	*3.5-2*	9.6711+2	9.7146+2	14	4:45	4.4676-4	8.4202-4	100%

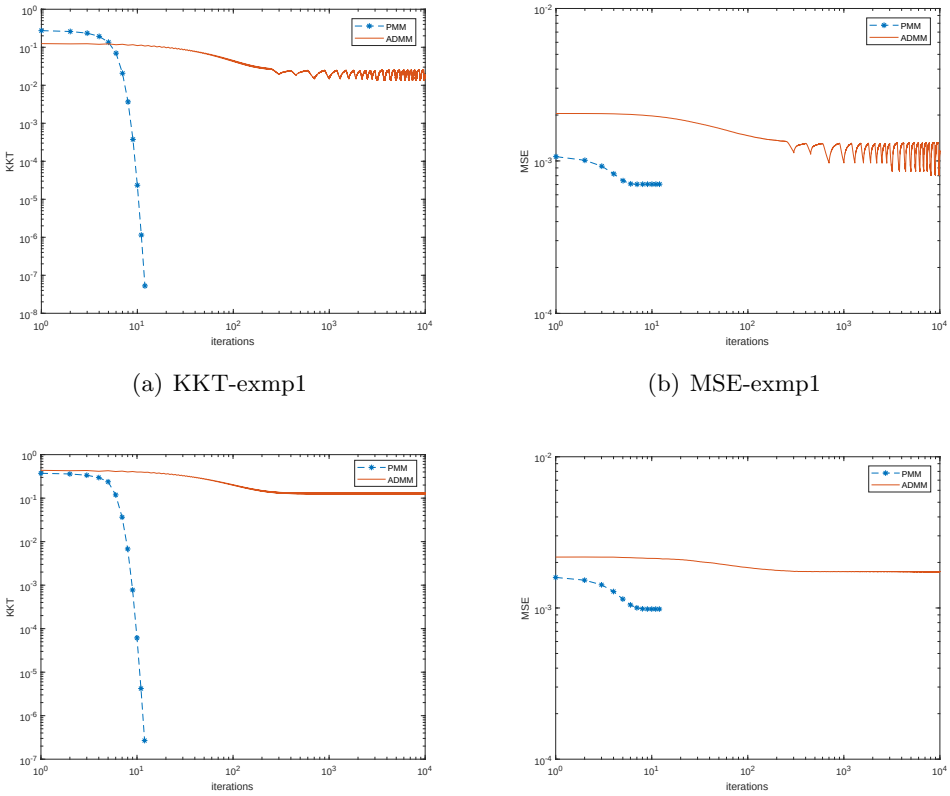


(a) (b)

Figure 2: The KKT residuals of the PMM and ADMM algorithms for solving the SCAD regularized problem with the UCI data. (a) The abalone.scale.expanded7 data; (b) The housing.scale.expanded7 data.

Table 8: The performance of the ADMM and PMM on UCI datasets for the MCP regularization. In the table, "a"=PMM, "b"=ADMM.

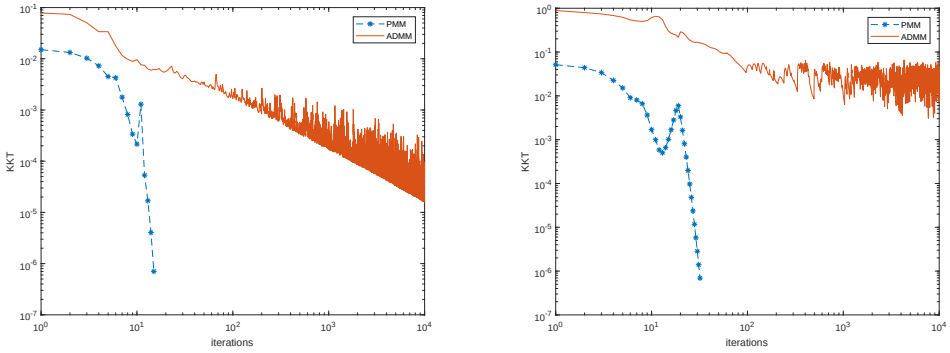
probname m; n	λ_c	nnz	$\eta_{\rm kkt}$		pobj		time	
			a	b	a	b	a	b
E2006.test 3308;150358	0.090	1	2.4-8	9.3-7	2.2077+1	2.2077+1	07	07
log1p.E2006.test 3308;1771946	0.261	187	8.2-7	*2.2-3*	2.1455+1	3.6500+1	4:09	2:20:46
E2006.train 16087;150358	0.028	1	6.8-7	*1.1-6*	4.8914+1	4.9256+1	20	3:12:39
log1p.E2006.train 16087;4272227	0.541	67	1.2-7	*1.4-2*	4.9200+1	1.6911+2	6:49	4:02:05
abalone.scale.expanded7 4177;6435	0.012	55	7.1-7	*1.6-5*	1.3271+2	1.2693+2	09	21:32
housing.scale.expanded7 506;77520	0.282	22	6.9-7	*4.1-2*	1.1022+2	7.2573+2	26	30:09
mpg.scale.expanded7 392;3432	0.102	23	7.3-7	*2.1-2*	5.0964+1	5.9492+1	01	1:36
space.ga.scale.expanded9 3107;5005	0.046	15	5.9-7	*3.3-2*	6.6399+0	7.8456+0	05	12:46
pyrim.scale.expanded5 74;201376	0.221	43	9.9-7	*7.0-3*	1.1428+0	4.6112+0	18	19:36
bodyfat.scale.expanded7 252;116280	0.183	2	2.8-7	*5.3-6*	5.7347-1	5.8278-1	06	25:11



(a) KKT-exmp1 (b) MSE-exmp1

(c) KKT-exmp2 (d) MSE-exmp2

Figure 3: The KKT residuals and the MSEs of the PMM and ADMM algorithms for solving the MCP regularized problem with the first two random data sets.



(a) (b)

Figure 4: The KKT residuals of the PMM and ADMM algorithms for solving the MCP regularized problem with the UCI data. (a) The abalone.scale.expanded7 data; (b) The housing.scale.expanded7 data.

but for the MCP regularized problem. Observe that the MSEs achieved by the PMM are better than those attained by the ADMM.

We have mentioned in the introduction that the scaled Lasso problem is equivalent to the srLasso problem (2). However, in order to solve the scaled Lasso problem, we have to call an algorithm several times to solve the standard Lasso subproblems. However, by handling the srLasso problem (2) directly, our algorithm is as fast as the highly efficient algorithm, LassoNAL (Li et al., 2018), for solving a single standard Lasso problem.

6. Conclusion

In this paper, we proposed a two stage PMM algorithm to solve the square-root regression problems with nonconvex regularizations. We are able to achieve impressive computational efficiency for our algorithm by designing an innovative proximal majorization framework for the convex subproblem arising in each PMM iteration so that it can be solved via its dual by the SSN method. We presented the oracle property of the problem in stage I and analyzed the convergence of the PMM algorithm with its subproblems solved inexactly. Extensive numerical experiments have demonstrated the efficiency of our PMM algorithm when compared to other natural alternative algorithms such as the ADMM based algorithms in solving the problem of interest.

From the superior performance of our algorithm, it is natural for us to consider applying a similar proximal majorization-minimization algorithmic framework to design efficient algorithms to solve other square-root regression problems with structured sparsity requirements such a group sparsity in the regression coefficients (Bunea et al., 2014). We leave such an investigation as a future research topic.

Acknowledgments

We would like to thank Prof. Jian Huang and Dr. Ying Cui for bringing the references (Sun and Zhang, 2012; Xu et al., 2010), respectively, to our attention. We also thank Prof. Sara van de Geer at ETH Zürich and Dr. Benjamin Stucky at University of Zürich for the fruitful discussions. Last but not least, we thank the action editor Dr. David Wipf and the anonymous referees for their helpful suggestions to improve the manuscript.

The work of Peipei Tang is supported by the Natural Science Foundation of Zhejiang Province of China under Grant No. LY19A010028 and the Zhejiang Science and Technology Plan Project of China (No. 2020C03091, No. 2021C01164). The work of Defeng Sun is supported by Hong Kong Research Grant Council under Grant PolyU 153014/18P and Shenzhen Research Institute of Big Data, Shenzhen 518000 under Grant 2019ORF01002. The work of Kim-Chuan Toh is supported in part by the Academic Research Fund of the Ministry of Education of Singapore under Grant No. R-146-000-257-112. Part of this research was done while Kim-Chuan Toh was visiting the Shenzhen Research Institute of Big Data at the Chinese University of Hong Kong at Shenzhen.

Appendix A. Proofs for Section 3

In this appendix, we first provide the proofs for Lemma 5, Lemma 7 and Lemma 8. Based on these results, we then give the proof for Theorem 9.

Lemma 5 *Let S be an allowed set of a weakly decomposable norm p. For the parameters λ_0 and λ_m defined by (7), we have $\lambda_0 \le \lambda_m$ and $p_*(\ddot{\beta}) \le \lambda_m$.*

Proof For a given allowed set S of a weakly decomposable norm p, denote

$$C_1 = \Big\{ z \in \mathbb{R}^n \Big| p(z_S) \le 1,\ z^{\bar{S}} = 0 \Big\}, \quad C_2 = \Big\{ z \in \mathbb{R}^n \Big| p^{\bar{S}}(z^{\bar{S}}) \le 1,\ z^S = 0 \Big\},$$
$$C = \Big\{ z \in \mathbb{R}^n \Big| p(z_S) + p^{\bar{S}}(z^{\bar{S}}) \le 1 \Big\}.$$

Then we have that

$$\begin{aligned} \delta^*_{C_1}(\beta) &= \max_{z\in\mathbb{R}^n} \Big\{ \langle z, \beta\rangle \Big| p(z_S) \le 1, z^{\bar{S}} = 0 \Big\} = \max_{z\in\mathbb{R}^n} \Big\{ \langle z, \beta_S\rangle \Big| p(z_S) \le 1, z^{\bar{S}} = 0 \Big\} \\ &\le \max_{z\in\mathbb{R}^n} \Big\{ \langle z, \beta_S\rangle \Big| p(z) \le 1 \Big\} = p_*(\beta_S), \qquad (19) \\ \delta^*_{C_2}(\beta) &= \max_{z\in\mathbb{R}^n} \Big\{ \langle z, \beta\rangle \Big| p^{\bar{S}}(z^{\bar{S}}) \le 1, z^S = 0 \Big\} = \max_{z\in\mathbb{R}^n} \Big\{ \langle z^{\bar{S}}, \beta^{\bar{S}}\rangle \Big| p^{\bar{S}}(z^{\bar{S}}) \le 1 \Big\} \\ &= p^{\bar{S}}_*(\beta^{\bar{S}}). \qquad (20) \end{aligned}$$

Furthermore, on one hand, for any $x \in C_1$, $y \in C_2$ and $0 \le t \le 1$, it is easy to prove that $tx+(1-t)y \in C$. That is $\mathrm{conv}(C_1 \cup C_2) \subseteq C$. On the other hand, for any $z \in C$ with $z^S = 0$ or $z^{\bar{S}} = 0$, it is clear that $z \in \mathrm{conv}(C_1 \cup C_2)$; and for any $z \in C$ with $z^S \ne 0$ and $z^{\bar{S}} \ne 0$, we can find $x = \frac{z_S}{p(z_S)} \in C_1$ and $y = \frac{z_{\bar{S}}}{1-p(z_S)} \in C_2$ such that $z = p(z_S)x + (1-p(z_S))y$. Therefore we have shown that $C = \mathrm{conv}(C_1 \cup C_2)$.

Due to (Rockafellar, 1970, Theorem 5.6), we can prove the following fact easily.

$$\mathrm{conv}(\delta_{C_1}, \delta_{C_2})(\beta) = \delta_{\mathrm{conv}(C_1\cup C_2)}(\beta), \quad \forall\, \beta \in \mathbb{R}^n, \qquad (21)$$

where $\mathrm{conv}(\delta_{C_1}, \delta_{C_2})$ denotes the greatest convex function that is less than or equal to δ_{C_1} and δ_{C_2} pointwise over the entire $\mathbb{R}^n$. Based on the above basic results (19)-(21), $C = \mathrm{conv}(C_1 \cup C_2)$ and (Rockafellar, 1970, Theorem 16.5), we have that

$$\begin{aligned} p_*(\beta) &= \max_{z\in\mathbb{R}^n}\left\{\langle\beta, z\rangle\middle| p(z) \le 1\right\} \le \max_{z\in\mathbb{R}^n}\left\{\langle\beta, z\rangle\middle| p(z_S) + p^{\bar{S}}(z^{\bar{S}}) \le 1\right\} \\ &= \delta_C^*(\beta) = \delta^*_{\mathrm{conv}(C_1\cup C_2)}(\beta) = (\mathrm{conv}(\delta_{C_1}, \delta_{C_2}))^*(\beta) = \max\left\{\delta^*_{C_1}(\beta), \delta^*_{C_2}(\beta)\right\} \\ &\le \max\left\{p_*(\beta_S), p^{\bar{S}}_*(\beta^{\bar{S}})\right\}. \end{aligned}$$

The desired results follow by taking $\beta = \ddot{\beta}$ and dividing both sides of the above inequality by $\|\varepsilon\|$ with $\beta = \varepsilon^T X$, respectively. ■

Lemma 7 *Suppose that Assumption 1 holds. For the estimator $\hat{\beta}$ of the generalized elastic-net square-root regression problem (6), we have*

$$\hat{\varepsilon}^T X(\ddot{\beta} - \hat{\beta}) \le \left(\tau + \frac{1}{\|\hat{\varepsilon}\|}\right)^{-1}\left(\lambda p(\ddot{\beta}) + \sigma p_*(\ddot{\beta}) p(\hat{\beta})\right).$$

Proof Since $\hat{\beta} \in \underset{\beta\in\mathbb{R}^n}{\mathrm{argmin}}\{h(\beta; \sigma, \tau, 0, 0, b)\}$ and p is a convex function, we have

$$-\frac{X^T(X\hat{\beta} - b)}{\|X\hat{\beta} - b\|} - \sigma\hat{\beta} - \tau X^T(X\hat{\beta} - b) \in \lambda\partial p(\hat{\beta}).$$

Hence

$$\lambda p(\beta) \ge \lambda p(\hat{\beta}) + \left\langle \frac{X^T\hat{\varepsilon}}{\|\hat{\varepsilon}\|} - \sigma\hat{\beta} + \tau X^T\hat{\varepsilon}, \beta - \hat{\beta}\right\rangle. \tag{22}$$

Let $\beta = \ddot{\beta}$. Then the inequality (22) can be rearranged to

$$\begin{aligned} \left(\tau + \frac{1}{\|\hat{\varepsilon}\|}\right)\hat{\varepsilon}^T X(\ddot{\beta} - \hat{\beta}) &\le \lambda p(\ddot{\beta}) - \lambda p(\hat{\beta}) + \sigma\hat{\beta}^T(\ddot{\beta} - \hat{\beta}) \le \lambda p(\ddot{\beta}) + \sigma\hat{\beta}^T\ddot{\beta} \\ &\le \lambda p(\ddot{\beta}) + \sigma p_*(\ddot{\beta}) p(\hat{\beta}). \end{aligned}$$

Note that the last inequality is obtained by the definition of p_*. The desired result now follows readily. ■

Lemma 8 *Suppose that Assumption 1 holds. We have*

$$c_l := \frac{1 - a - \frac{2\lambda_0 n_p}{\lambda}}{2 + \left(1 + \frac{\sigma p_*(\ddot{\beta})}{\lambda}\right) n_p} \le \frac{\|\hat{\varepsilon}\|}{\|\varepsilon\|} \le c_u,$$

where the constants c_u and a are defined in (8).

Proof Since $\hat{\beta} \in \underset{\beta\in\mathbb{R}^n}{\operatorname{argmin}}\{h(\beta;\sigma,\tau,0,0,b)\}$, we have $h(\hat{\beta};\sigma,\tau,0,0,b) \le h(\ddot{\beta};\sigma,\tau,0,0,b)$. Thus, by the definition of the dual norm, we get

$$\|\hat{\varepsilon}\| \le \|\varepsilon\| + \frac{\tau}{2}\|\varepsilon\|^2 + \frac{\sigma}{2}\|\ddot{\beta}\|^2 + \lambda p(\ddot{\beta}) \le \|\varepsilon\| + \frac{\tau}{2}\|\varepsilon\|^2 + \left(\lambda + \frac{\sigma p_*(\ddot{\beta})}{2}\right) p(\ddot{\beta}), \tag{23}$$

$$\lambda p(\hat{\beta}) \le \|\varepsilon\| + \frac{\tau}{2}\|\varepsilon\|^2 + \frac{\sigma}{2}\|\ddot{\beta}\|^2 + \lambda p(\ddot{\beta}) \le \|\varepsilon\| + \frac{\tau}{2}\|\varepsilon\|^2 + \left(\lambda + \frac{\sigma p_*(\ddot{\beta})}{2}\right) p(\ddot{\beta}). \tag{24}$$

Dividing both sides of (23) by $\|\varepsilon\|$, we obtain

$$\frac{\|\hat{\varepsilon}\|}{\|\varepsilon\|} \le 1 + \frac{\tau}{2}\|\varepsilon\| + n_p + \frac{\sigma p_*(\ddot{\beta})p(\ddot{\beta})}{2\|\varepsilon\|} = c_u,$$

where c_u is defined in (8). In order to obtain the lower bound of $\frac{\|\hat{\varepsilon}\|}{\|\varepsilon\|}$, we first use the triangle inequality $\|\hat{\varepsilon}\| = \|\varepsilon - X(\hat{\beta}-\ddot{\beta})\| \ge \|\varepsilon\| - \|X(\hat{\beta}-\ddot{\beta})\|$, and then the upper bound of $\|X(\hat{\beta}-\ddot{\beta})\|$. By Lemma 7 and the definition of the dual norm, we have

$$\begin{aligned}
\|X(\hat{\beta}-\ddot{\beta})\|^2 &= \varepsilon^T X(\hat{\beta}-\ddot{\beta}) + \hat{\varepsilon}^T X(\ddot{\beta}-\hat{\beta}) \\
&\le \varepsilon^T X(\hat{\beta}-\ddot{\beta}) + \kappa\left(\lambda p(\ddot{\beta}) + \sigma p_*(\ddot{\beta})p(\hat{\beta})\right) \\
&\le \lambda_0 p(\hat{\beta}-\ddot{\beta})\|\varepsilon\| + \kappa\left(\lambda p(\ddot{\beta}) + \sigma p_*(\ddot{\beta})p(\hat{\beta})\right) \\
&\le \lambda_0\left(p(\hat{\beta}) + p(\ddot{\beta})\right)\|\varepsilon\| + \kappa\left(\lambda p(\ddot{\beta}) + \sigma p_*(\ddot{\beta})p(\hat{\beta})\right) \\
&= \left(\lambda_0\|\varepsilon\| + \lambda\kappa\right) p(\ddot{\beta}) + \left(\lambda_0\|\varepsilon\| + \sigma\kappa p_*(\ddot{\beta})\right) p(\hat{\beta}),
\end{aligned}$$

where $\kappa = \left(\tau + \frac{1}{\|\hat{\varepsilon}\|}\right)^{-1}$. Substituting the inequality (24) into the above formula, we can obtain

$$\begin{aligned}
\|X(\hat{\beta}-\ddot{\beta})\|^2 \le\ & \frac{\|\varepsilon\| + \frac{\tau}{2}\|\varepsilon\|^2}{\lambda}\left(\lambda_0\|\varepsilon\| + \sigma\kappa p_*(\ddot{\beta})\right) \\
& + \Big(2\lambda_0\|\varepsilon\| + (\lambda\kappa + \sigma\kappa p_*(\ddot{\beta})) + \frac{\sigma p_*(\ddot{\beta})}{2\lambda}(\lambda_0\|\varepsilon\| + \sigma\kappa p_*(\ddot{\beta}))\Big) p(\ddot{\beta}).
\end{aligned}$$

Rearranging the above inequality, we have

$$\begin{aligned}
\|X(\hat{\beta}-\ddot{\beta})\| &\le \|\varepsilon\|\sqrt{\hat{a} + \frac{2\lambda_0 p(\ddot{\beta})}{\|\varepsilon\|} + \frac{\|\hat{\varepsilon}\|}{\|\varepsilon\|}\frac{\left(\lambda + \sigma p_*(\ddot{\beta})\right) p(\ddot{\beta})}{(1+\tau\|\hat{\varepsilon}\|)\|\varepsilon\|}} \\
&\le \|\varepsilon\|\sqrt{\hat{a} + \frac{2\lambda_0 n_p}{\lambda} + \frac{\|\hat{\varepsilon}\|}{\|\varepsilon\|}\left(1 + \frac{\sigma p_*(\ddot{\beta})}{\lambda}\right) n_p},
\end{aligned}$$

where

$$\begin{aligned}
\hat{a} &= \left(\frac{\left(1 + \frac{\tau}{2}\|\varepsilon\|\right)}{\lambda} + \frac{\sigma p_*(\ddot{\beta})p(\ddot{\beta})}{2\lambda\|\varepsilon\|}\right)\left(\lambda_0 + \frac{\sigma\kappa p_*(\ddot{\beta})}{\|\varepsilon\|}\right) \\
&\le \left(\frac{\left(1 + \frac{\tau}{2}\|\varepsilon\|\right)}{\lambda} + \frac{\sigma p_*(\ddot{\beta})p(\ddot{\beta})}{2\lambda\|\varepsilon\|}\right)\left(\lambda_0 + \sigma p_*(\ddot{\beta})c_u\right) = a.
\end{aligned}$$

Therefore, by noting that $\|X(\hat{\beta}-\ddot{\beta})\| = \|\varepsilon - \hat{\varepsilon}\|$ and triangle inequality, we have

$$\frac{\|\hat{\varepsilon}\|}{\|\varepsilon\|} \geq 1 - \sqrt{a + \frac{2\lambda_0 n_p}{\lambda} + \frac{\|\hat{\varepsilon}\|}{\|\varepsilon\|}\left(1 + \frac{\sigma p_*(\ddot{\beta})}{\lambda}\right) n_p}.$$

By rearranging the above inequality, in the case when $\frac{\|\hat{\varepsilon}\|}{\|\varepsilon\|} < 1$, we further derive that

$$a + \frac{2\lambda_0 n_p}{\lambda} + \frac{\|\hat{\varepsilon}\|}{\|\varepsilon\|}\left(1 + \frac{\sigma p_*(\ddot{\beta})}{\lambda}\right) n_p \geq \left(1 - \frac{\|\hat{\varepsilon}\|}{\|\varepsilon\|}\right)^2 \geq 1 - \frac{2\|\hat{\varepsilon}\|}{\|\varepsilon\|}.$$

Then we can obtain that

$$\frac{\|\hat{\varepsilon}\|}{\|\varepsilon\|} \geq \frac{1 - a - \frac{2\lambda_0 n_p}{\lambda}}{2 + \left(1 + \frac{\sigma p_*(\ddot{\beta})}{\lambda}\right) n_p} := c_l > 0.$$

In the other case, if $\frac{\|\hat{\varepsilon}\|}{\|\varepsilon\|} \geq 1$, we have already obtain a lower bound that is larger than c_l. ■

Theorem 9 *Let $\delta \in [0,1)$. Under Assumptions 1 and 2, assume that $\frac{s_2 - \sqrt{s_2^2 - 4s_1 s_3}}{2s_1} < \lambda < \frac{s_2 + \sqrt{s_2^2 - 4s_1 s_3}}{2s_1}$ with $s_1 = \frac{\sigma\lambda_m p^2(\ddot{\beta})}{\|\varepsilon\|^2}$, $s_2 = 1 - \frac{\lambda_m(3 + 2\sigma t_1 + \sigma t_2)p(\ddot{\beta})}{\|\varepsilon\|} > 0$ and $s_3 = \lambda_m(t_1 + t_2 + \sigma t_1 t_2 + \sigma t_1^2)$. For any $\hat{\beta} \in \Omega(\sigma,\tau)$, and any $\beta \in \mathbb{R}^n$ such that supp(β) is a subset of S, we have that*

$$\begin{aligned}
&\|X(\hat{\beta} - \ddot{\beta})\|^2 + 2\delta\left((\hat{\lambda} - \tilde{\lambda}_m)p^{\bar{S}}(\hat{\beta}^{\bar{S}}) + (\tilde{\lambda} + \tilde{\lambda}_m)p(\hat{\beta}_S - \beta)\right)\|\varepsilon\| \\
&\leq \|X(\beta - \ddot{\beta})\|^2 + \left((1+\delta)(\tilde{\lambda} + \tilde{\lambda}_m)\Gamma_p(L_S, S)\|\varepsilon\|\right)^2 + 2\sigma c_u\|\hat{\beta} - \ddot{\beta}\|\|\beta - \ddot{\beta}\|\|\varepsilon\|,
\end{aligned}$$

where

$$\hat{\lambda} := \frac{\lambda c_l}{1 + \tau c_l},\ \tilde{\lambda}_m := \lambda_m(1 + \sigma c_u),\ \tilde{\lambda} := \lambda c_u,\ L_S := \frac{\tilde{\lambda} + \tilde{\lambda}_m}{\hat{\lambda} - \tilde{\lambda}_m} \cdot \frac{1+\delta}{1-\delta}.$$

Proof First we note that if the following inequality holds

$$\begin{aligned}
\langle X(\hat{\beta} - \ddot{\beta}), X(\hat{\beta} - \beta)\rangle \ \leq\ & -\delta\left((\tilde{\lambda} + \tilde{\lambda}_m)p(\hat{\beta}_S - \beta) + (\hat{\lambda} - \tilde{\lambda}_m)p^{\bar{S}}(\hat{\beta}^{\bar{S}})\right)\|\varepsilon\| \\
& + \sigma c_u\|\hat{\beta} - \ddot{\beta}\|\|\beta - \ddot{\beta}\|\|\varepsilon\|,
\end{aligned}$$

then we can verify that the theorem is valid by the following simple calculations:

$$\begin{aligned}
&\|X(\hat{\beta} - \ddot{\beta})\|^2 - \|X(\beta - \ddot{\beta})\|^2 + 2\delta\left((\tilde{\lambda} + \tilde{\lambda}_m)p(\hat{\beta}_S - \beta) + (\hat{\lambda} - \tilde{\lambda}_m)p^{\bar{S}}(\hat{\beta}^{\bar{S}})\right)\|\varepsilon\| \\
&= \ 2\langle X(\hat{\beta} - \ddot{\beta}), X(\hat{\beta} - \beta)\rangle - \|X(\beta - \hat{\beta})\|^2 + 2\delta\left((\tilde{\lambda} + \tilde{\lambda}_m)p(\hat{\beta}_S - \beta) + (\hat{\lambda} - \tilde{\lambda}_m)p^{\bar{S}}(\hat{\beta}^{\bar{S}})\right)\|\varepsilon\| \\
&\leq \ -\|X(\beta - \hat{\beta})\|^2 + 2\sigma c_u\|\hat{\beta} - \ddot{\beta}\|\|\beta - \ddot{\beta}\|\|\varepsilon\| \\
&\leq \ 2\sigma c_u\|\hat{\beta} - \ddot{\beta}\|\|\beta - \ddot{\beta}\|\|\varepsilon\|.
\end{aligned}$$

Thus it is sufficient to show that the result is true if

$$\begin{aligned}\langle X(\hat{\beta}-\ddot{\beta}), X(\hat{\beta}-\beta)\rangle \;\; \geq \;\; & -\delta\left((\tilde{\lambda}+\tilde{\lambda}_m)p(\hat{\beta}_S-\beta)+(\hat{\lambda}-\tilde{\lambda}_m)p^{\bar{S}}(\hat{\beta}^{\bar{S}})\right)\|\varepsilon\| \\ & +\sigma c_u\|\hat{\beta}-\ddot{\beta}\|\|\beta-\ddot{\beta}\|\|\varepsilon\|. \end{aligned} \tag{25}$$

By the inequality (22) and the fact that $\hat{\varepsilon} = X(\ddot{\beta}-\hat{\beta}+\varepsilon)$, we can get

$$\langle X(\hat{\beta}-\ddot{\beta}), X(\hat{\beta}-\beta)\rangle + \lambda\kappa p(\hat{\beta}) \;\; \leq \;\; \langle \varepsilon, X(\hat{\beta}-\beta)\rangle + \sigma\kappa\langle\hat{\beta}, \beta-\hat{\beta}\rangle + \lambda\kappa p(\beta), \tag{26}$$

where $\kappa := \left(\tau + \frac{1}{\|\bar{\varepsilon}\|}\right)^{-1}$. Since $\mathrm{supp}(\beta) \subseteq S$, it follows from the definition of the dual norm and the generalized Cauchy-Schwartz inequality that

$$\begin{aligned}\langle \varepsilon, X(\hat{\beta}-\beta)\rangle \;\; &= \;\; \langle \varepsilon, X(\hat{\beta}_S-\beta)\rangle + \langle \varepsilon, X(\hat{\beta}_{\bar{S}}-\beta)\rangle \\ &\leq \;\; \left(p_*((\varepsilon^T X)_S)p(\hat{\beta}_S-\beta)+p_*^{\bar{S}}((\varepsilon^T X)^{\bar{S}})p^{\bar{S}}(\hat{\beta}^{\bar{S}})\right) \\ &\leq \;\; \lambda_m\left(p(\hat{\beta}_S-\beta)+p^{\bar{S}}(\hat{\beta}^{\bar{S}})\right)\|\varepsilon\|. \end{aligned} \tag{27}$$

By substituting (27) into (26), we obtain

$$\begin{aligned}&\langle X(\hat{\beta}-\ddot{\beta}), X(\hat{\beta}-\beta)\rangle + \lambda\kappa p(\hat{\beta}) \\ &\leq \;\; \lambda_m\left(p(\hat{\beta}_S-\beta)+p^{\bar{S}}(\hat{\beta}^{\bar{S}})\right)\|\varepsilon\| + \sigma\kappa\langle\hat{\beta}, \beta-\hat{\beta}\rangle + \lambda\kappa p(\beta). \end{aligned} \tag{28}$$

Furthermore, by using the weak decomposability and the triangle inequality in (28) we derive

$$\begin{aligned}&\langle X(\hat{\beta}-\ddot{\beta}), X(\hat{\beta}-\beta)\rangle + \lambda\kappa\left(p^{\bar{S}}(\hat{\beta}^{\bar{S}})+p(\hat{\beta}_S)\right) \\ &\leq \;\; \lambda_m\left(p(\hat{\beta}_S-\beta)+p^{\bar{S}}(\hat{\beta}^{\bar{S}})\right)\|\varepsilon\| + \sigma\kappa\langle\hat{\beta}, \beta-\hat{\beta}\rangle + \lambda\kappa\left(p(\hat{\beta}_S)+p(\hat{\beta}_S-\beta)\right). \end{aligned} \tag{29}$$

Then by eliminating $\lambda\kappa p(\hat{\beta}_S)$ on both sides of (29) and using the weak decomposability, we get

$$\begin{aligned}&\langle X(\hat{\beta}-\ddot{\beta}), X(\hat{\beta}-\beta)\rangle + \lambda\kappa p^{\bar{S}}(\hat{\beta}^{\bar{S}}) \\ &\leq \;\; \lambda_m\left(p(\hat{\beta}_S-\beta)+p^{\bar{S}}(\hat{\beta}^{\bar{S}})\right)\|\varepsilon\| + \sigma\kappa\langle\hat{\beta}, \beta-\hat{\beta}\rangle + \lambda\kappa p(\hat{\beta}_S-\beta) \\ &\leq \;\; \lambda_m\left(p(\hat{\beta}_S-\beta)+p^{\bar{S}}(\hat{\beta}^{\bar{S}})\right)\|\varepsilon\| + \sigma\kappa\langle\ddot{\beta}, \beta-\hat{\beta}\rangle + \sigma\kappa\langle\hat{\beta}-\ddot{\beta}, \beta-\ddot{\beta}\rangle + \lambda\kappa p(\hat{\beta}_S-\beta) \\ &\leq \;\; \lambda_m\left(p(\hat{\beta}_S-\beta)+p^{\bar{S}}(\hat{\beta}^{\bar{S}})\right)\|\varepsilon\| + \lambda\kappa p(\hat{\beta}_S-\beta) + \lambda_m\sigma\kappa\left(p(\hat{\beta}_S-\beta)+p^{\bar{S}}(\hat{\beta}^{\bar{S}})\right) \\ &\quad +\sigma\kappa\|\hat{\beta}-\ddot{\beta}\|\|\beta-\ddot{\beta}\|. \end{aligned} \tag{30}$$

By using the result of Lemma 8, the inequality (30) becomes

$$\begin{aligned}&\langle X(\hat{\beta}-\ddot{\beta}), X(\hat{\beta}-\beta)\rangle + \left(\hat{\lambda}-\tilde{\lambda}_m\right)p^{\bar{S}}(\hat{\beta}^{\bar{S}})\|\varepsilon\| \\ &\leq \;\; \left(\tilde{\lambda}+\tilde{\lambda}_m\right)p(\hat{\beta}_S-\beta)\|\varepsilon\| + \sigma c_u\|\hat{\beta}-\ddot{\beta}\|\|\beta-\ddot{\beta}\|\|\varepsilon\|. \end{aligned} \tag{31}$$

From the condition (25) in (31) and simple rearrangement, we have that

$$\left(\hat{\lambda}-\tilde{\lambda}_m\right)(1-\delta)p^{\bar{S}}(\hat{\beta}^{\bar{S}}) \le \left(\tilde{\lambda}+\tilde{\lambda}_m\right)(1+\delta)p(\hat{\beta}_S-\beta).$$

By Lemma 5 we have $\lambda_0 \le \lambda_m$ and $p_*(\ddot{\beta}) \le \lambda_m$. Since

$$\frac{\lambda-\lambda_m t_1(1+\sigma t_1+\sigma n_p)-2\lambda_m n_p}{\lambda\left(2+n_p+\frac{\sigma p_*(\ddot{\beta})p(\ddot{\beta})}{\|\varepsilon\|}\right)} < c_l < \frac{1}{2+n_p+\frac{\sigma p_*(\ddot{\beta})p(\ddot{\beta})}{\|\varepsilon\|}},$$

we can see that

$$\hat{\lambda} > \frac{\lambda-\lambda_m t_1(1+\sigma t_1+\sigma n_p)-2\lambda_m n_p}{t_2+n_p}.$$

Then it is easy to find that if

$$\frac{s_2-\sqrt{s_2^2-4s_1s_3}}{2s_1} < \lambda < \frac{s_2+\sqrt{s_2^2-4s_1s_3}}{2s_1},$$

then we have $\hat{\lambda}=\lambda c_l/(1+\tau c_l) > \tilde{\lambda}_m = \lambda_m(1+\sigma c_u)$.

Furthermore,

$$p^{\bar{S}}(\hat{\beta}^{\bar{S}}) \le \left(\frac{\tilde{\lambda}+\tilde{\lambda}_m}{\hat{\lambda}-\tilde{\lambda}_m}\right)\cdot\frac{1+\delta}{1-\delta}\cdot p(\hat{\beta}_S-\beta).$$

From the definition of L_S and Lemma 6 with the assumption $\mathrm{supp}(\beta)\subseteq S$, it follows that

$$p^{\bar{S}}(\hat{\beta}^{\bar{S}}) \le L_S p(\hat{\beta}_S-\beta), \quad p(\hat{\beta}_S-\beta) \le \Gamma_p(L_S,S)\|X(\hat{\beta}-\beta)\|.$$

By using the inequality (31), we can derive that

$$\begin{aligned}
&\langle X(\hat{\beta}-\ddot{\beta}), X(\hat{\beta}-\beta)\rangle + \delta\|\varepsilon\|(\hat{\lambda}-\tilde{\lambda}_m)p^{\bar{S}}(\hat{\beta}^{\bar{S}})\\
&\le (\tilde{\lambda}+\tilde{\lambda}_m)p(\hat{\beta}_S-\beta)\|\varepsilon\| + \sigma c_u\|\hat{\beta}-\ddot{\beta}\|\|\beta-\ddot{\beta}\|\|\varepsilon\|\\
&\le (1+\delta)(\tilde{\lambda}+\tilde{\lambda}_m)\Gamma_p(L_S,S)\|X(\hat{\beta}-\beta)\|\|\varepsilon\| - \delta(\tilde{\lambda}+\tilde{\lambda}_m)p(\hat{\beta}_S-\beta)\|\varepsilon\|\\
&\quad +\sigma c_u\|\hat{\beta}-\ddot{\beta}\|\|\beta-\ddot{\beta}\|\|\varepsilon\|.
\end{aligned}$$

Noticing that

$$2\langle X(\hat{\beta}-\ddot{\beta}), X(\hat{\beta}-\beta)\rangle = \|X(\hat{\beta}-\ddot{\beta})\|^2 - \|X(\beta-\ddot{\beta})\|^2 + \|X(\hat{\beta}-\beta)\|^2,$$

$$2(1+\delta)|(\tilde{\lambda}+\tilde{\lambda}_m)\Gamma_p(L_S,S)\|X(\hat{\beta}-\beta)\|\|\varepsilon\| \le \left((1+\delta)(\tilde{\lambda}+\tilde{\lambda}_m)\Gamma_p(L_S,S)\right)^2\|\varepsilon\| + \|X(\hat{\beta}-\beta)\|^2,$$

we get

$$\begin{aligned}
&\|X(\hat{\beta}-\ddot{\beta})\|^2 + 2\delta\left((\hat{\lambda}-\tilde{\lambda}_m)p^{\bar{S}}(\hat{\beta}^{\bar{S}}) + (\tilde{\lambda}+\tilde{\lambda}_m)p(\hat{\beta}_S-\beta)\right)\|\varepsilon\|\\
&\le \|X(\beta-\ddot{\beta})\|^2 + (1+\delta)^2(\tilde{\lambda}+\tilde{\lambda}_m)^2\Gamma_p^2(L_S,S)\|\varepsilon\|^2 + 2\sigma c_u\|\hat{\beta}-\ddot{\beta}\|\|\beta-\ddot{\beta}\|\|\varepsilon\|.
\end{aligned}$$

Therefore the oracle inequality holds and this completes the proof. ■

References

M. Ahn, J.-S. Pang, and J. Xin. Difference-of-convex learning: directional stationarity, optimality, and sparsity. *SIAM Journal on Optimization*, 27(3):1637–1665, 2017.

H. Attouch and J. Bolte. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Mathematical Programming*, 116(1-2):5–16, 2009.

F.R. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends in Machine Learning*, 4(1):1–106, 2012.

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

P.C. Bellec, G. Lecué, and A.B. Tsybakov. SLOPE meets LASSO: improved oracle bounds and optimality. *Annals of Statistics*, 46(6B):3603–3642, 2018.

A. Belloni, V. Chernozhukov, and L. Wang. Square-root LASSO: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.

J. Bolte and E. Pauwels. Majorization-minimization procedures and convergence of SQP methods for semi-algebraic and tame programs. *Mathematics of Operations Research*, 41 (2):442–465, 2016.

J. Bolte, S. Sabach, and M. Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1-2):459–494, 2014.

F. Bunea, J. Lederer, and Y. She. The group square-root LASSO: theoretical properties and fast algorithms. *IEEE Transactions on Information Theory*, 60(2):1313–1325, 2014.

C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):Article 27, 2011.

R. Chartrand. Exact reconstruction of sparse signals via nonconvex minimization. *IEEE Signal Processing Letters*, 14(10):707–710, 2007.

L. Chen and Y. Gu. The convergence guarantees of a non-convex approach for sparse recovery. *IEEE Transactions on Signal Processing*, 62(15):3754–3767, 2014.

X.D. Chen, D.F. Sun, and J. Sun. Complementarity functions and numerical experiments for second-order-cone complementarity problems. *Computational Optimization and Applications*, 25(1):39–56, 2003.

L. Chen, D.F. Sun, and K.-C. Toh. An efficient inexact symmetric Gauss-Seidel based majorized ADMM for high-dimensional convex composite conic programming. *Mathematical Programming*, 161(1-2):237–270, 2017.

Y. Cui, J.-S. Pang, and B. Sen. Composite difference-max programs for modern statistical estimation problems. *SIAM Journal on Optimization*, 28(4):3344–3374, 2018.

A. Derumigny. Improved bounds for square-root LASSO and square-root SLOPE. *Electronic Journal of Statistics*, 12(1):741–766, 2018.

G. Di Pillo and L. Grippo. On the exactness of a class of nondifferentiable penalty functions. *Journal of Optimization Theory and Applications*, 57(3):399–410, 1988.

J. Eckstein and D.P. Bertsekas. On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55(1): 293–318, 1992.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004.

F. Facchinei and J.-S. Pang. *Finite-Dimensional Variational Inequalities and Complementarity Problems-Volume I.* Springer New York, 2003.

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.

J. Flemming. *Generalized Tikhonov regularization, basic theory and comprehensive results on convergence rates.* PhD thesis, Fakultat fur Mathematik Technische Universitat Chemnitz, 2011.

D. Gabay and B. Mercier. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & Mathematics with Applications*, 2(1):17–40, 1976.

Y. Gao and D.F. Sun. A majorized penalty approach for calibrating rank constrained correlation matrix problems. *Manuscript, Department of Mathematics, National University of Singapore, Singapore,* revised May, 2010.

G. Golub and C.F. Van Loan. *Matrix Computations.* 3nd Edition, Johns Hopkins University Press, Baltimore, MD, 1996.

T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity: the LASSO and Generalizations.* CRC Press, 2015.

L. Huang, J. Jia, B. Yu, B.G. Chun, P. Maniatis, and M. Naik. Predicting execution time of computer programs using sparse polynomial regression. In *Advances in Neural Information Processing Systems 23: Conference on Neural Information Processing Systems A Meeting Held December*, 2010.

S.-J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky. An interior-point method for large-scale ℓ_1-regularized least squares. *IEEE Journal of Selected Topics in Signal Processing*, 1(4):606–617, 2007.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000.

H.A. Le Thi, D.T. Pham, and X.T. Vo. DC approximation approaches for sparse optimization. *European Journal of Operations Research*, 244(1):26–46, 2015.

X.D. Li, D.F. Sun, and K.-C. Toh. A highly efficient semismooth Newton augmented Lagrangian method for solving LASSO problems. *SIAM Journal on Optimization*, 28(1): 433–458, 2018.

X.G. Li, T. Zhao, X.M. Yuan, and H. Liu. The flare package for high dimensional linear regression and precision matrix estimation in R. *Journal of Machine Learning Research*, 16:553–557, 2015.

M. Lichman. UCI machine learning repository. 2013.

F.W. Meng, D.F. Sun, and G.Y. Zhao. Semismoothness of solutions to generalized equations and the Moreau-Yosida regularization. *Mathematical Programming*, 104:561–581, 2005.

R. Mifflin. Semismooth and semiconvex functions in constrained optimization. *SIAM Journal on Control and Optimization*, 15(6):959–972, 1977.

J.-S. Pang, M. Razaviyayn, and A. Alvarado. Computing B-stationary points of nonsmooth DC programs. *Mathematics of Operations Research*, 42(1):95–118, 2017.

L. Qi and J. Sun. A nonsmooth version of Newton's method. *Mathematical Programming*, 58:353–367, 1993.

R.T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, NJ, 1970.

R.T. Rockafellar and R.J.-B. Wets. *Variational Analysis*. Springer, New York, 1998.

B. Stucky and S. van de Geer. Sharp oracle inequalities for square root regularization. *Journal of Machine Learning Research*, 18:1–29, 2017.

T. Sun and C. Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012.

R. Tibshirani. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, Series B (Methodological)*, 58(1):267–288, 1996.

S. van de Geer. The deterministic LASSO. In *Proceeding of the Joint Statistical Meeting*, 2007.

S. van de Geer. Weakly decomposable regularization penalties and structured sparsity. *Scandinavian Journal of Statistics*, 41(1):72–86, 2014.

H. Xu, C. Caramanis, and S. Mannor. Robust regression and LASSO. *IEEE Transactions on Information Theory*, 56(7):3561–3573, 2010.

C. Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 2010.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.