SIAM J. SCI. COMPUT.
Vol. 32, No. 5, pp. 2832–2852

LOWER BOUND THEORY OF NONZERO ENTRIES IN SOLUTIONS OF ℓ_2-ℓ_p MINIMIZATION*

XIAOJUN CHEN†, FENGMIN XU‡, AND YINYU YE§

Abstract. Recently, variable selection and sparse reconstruction are solved by finding an optimal solution of a minimization model, where the objective function is the sum of a data-fitting term in ℓ_2 norm and a regularization term in ℓ_p norm ($0 < p < 1$). Since it is a nonconvex model, most algorithms for solving the problem can provide only an approximate local optimal solution, where nonzero entries in the solution cannot be identified theoretically. In this paper, we establish lower bounds for the absolute value of nonzero entries in every local optimal solution of the model, which can be used to indentify zero entries precisely in any numerical solution. Therefore, we have developed a lower bound theorem to classify zero and nonzero entries in every local solution. These lower bounds clearly show the relationship between the sparsity of the solution and the choice of the regularization parameter and norm so that our theorem can be used for selecting desired model parameters and norms. Furthermore, we also develop error bounds for verifying the accuracy of numerical solutions of the ℓ_2-ℓ_p minimization model. To demonstrate applications of our theory, we propose a hybrid orthogonal matching pursuit-smoothing gradient (OMP-SG) method for solving the nonconvex, non-Lipschitz continuous ℓ_2-ℓ_p minimization problem. Computational results show the effectiveness of the lower bounds for identifying nonzero entries in numerical solutions and the OMP-SG method for finding a high quality numerical solution.

Key words. variable selection, sparse solution, linear least-squares problem, ℓ_p regularization, smoothing approximation, first order condition, second order condition

AMS subject classifications. 90C26, 90C46, 90C90

DOI. 10.1137/090761471

1. Introduction.

We consider the following minimization problem

$$\min_{x\in R^n} \|Ax-b\|_2^2 + \lambda\|x\|_p^p, \tag{1.1}$$

where $A \in R^{m\times n}$, $b \in R^m$, $\lambda \in (0,\infty)$, $p \in (0,1)$. Recently, minimization problem (1.1) attracted great attention in variable selection and sparse reconstruction [5, 7, 8, 9, 28]. The objective function of (1.1),

$$f(x) := \|Ax-b\|_2^2 + \lambda\|x\|_p^p,$$

consists of a data-fitting term $\|Ax-b\|_2^2$ and a regularization term $\lambda\|x\|_p^p$. Problem (1.1) is intermediate between the ℓ_2-ℓ_0 minimization problem

$$\min_{x\in R^n} \|Ax-b\|_2^2 + \lambda\|x\|_0 \tag{1.2}$$

*Received by the editors June 8, 2008; accepted for publication (in revised form) September 10, 2010; published electronically November 4, 2010.
http://www.siam.org/journals/sisc/32-5/76147.html

†Department of Applied Mathematics, The Hong Kong Polytechnic University, Kowloon, Hong Kong (maxjchen@polyu.edu.hk). This author's work is supported partly by the Hong Kong Research Grants Council PolyU5003/08P.

‡School of Science and Department of Computer Science and Technology, Xi'an Jiaotong University, People's Republic of China (fengminxu@mail.xjtu.edu.cn). This author's work is supported partly by the national 973 project of China 2007CB311002.

§Department of Management Science and Engineering, School of Engineering, Stanford University, Stanford, CA 94305, and Visiting Professor of The Hong Kong Polytechnic University, Kowloon, Hong Kong (yyye@stanford.edu). This author is partially supported by DOE DE-SC0002009.

and the ℓ_2-ℓ_1 minimization problem

$$\min_{x\in R^n} \|Ax-b\|_2^2 + \lambda\|x\|_1 \tag{1.3}$$

in the sense that

$$\|x\|_0 = \sum_{\substack{i=1\\ x_i\neq 0}}^n |x_i|^0, \qquad \|x\|_p^p = \sum_{i=1}^n |x_i|^p, \qquad \text{and} \qquad \|x\|_1 = \sum_{i=1}^n |x_i|. \tag{1.4}$$

Naturally, one expects that using the ℓ_p norm[1] in the regularization term can find a sparser solution than using the ℓ_1 norm, which was evidenced in extensive computational studies [5, 7, 8, 9, 17, 28]. However, some major theoretical issues remain open. Is there any theoretical justification for solving minimization problem (1.1) with $p < 1$? What are the solution characteristics of (1.1)? Is there theory to dictate the choice of the regularization parameter λ and norm p? Our first main contribution of this paper is to answer these questions. We establish lower bounds for the absolute value of nonzero entries in every local optimal solution of (1.1) only when $p < 1$. Therefore, we have developed a lower bound theorem to classify zero and nonzero entries in every local solution of (1.1). These lower bounds clearly show the relationship between the sparsity of the solution and the choice of the regularization parameter and norm so that the theorem can be used to guide the selection of desired model parameters and norms in (1.1). It also can be used to identify zero and nonzero entries in the numerical optimal solution.

More specifically, using the second order necessary condition for a local minimizer, we present a componentwise lower bound

$$L_i = \left(\frac{\lambda p(1-p)}{2\|a_i\|^2}\right)^{\frac{1}{2-p}} \tag{1.5}$$

for each nonzero entry x_i^* of any local optimal solution x^* of (1.1), that is,

$$\text{for any} \quad i \in \mathcal{N}, \qquad L_i \leq |x_i^*| \qquad \text{if } x_i^* \neq 0,$$

which is equivalent to the following statement:

$$\text{for any} \quad i \in \mathcal{N}, \qquad x_i^* \in (-L_i, L_i) \quad \Rightarrow \quad x_i^* = 0.$$

Here $\mathcal{N} = \{1, \ldots, n\}$, and a_i is the ith column of the matrix A. We show that the columns $\{a_i \,|\, i \in \text{support}(x^*)\}$ are linearly independent, which implies that $\|x^*\|_0 \leq m$ and the ℓ_2-ℓ_p minimization problem (1.1) has a finite number of local minimizers.

Most minimization algorithms are descent-iterative in nature, that is, starting from an initial point x^0, they generate a sequence of points x^k, $k = 0, 1, \ldots$, such that the objective values $f(x^k)$ are strictly decreasing along the sequence. Thus, any local minimizer, including the global minimizer, that a descent algorithm may find must be in the level set $\{x : \ f(x) \leq f(x^0)\}$, and the set must contain at least one local minimizer. Therefore, in both theory and practice, one may be interested in only the minimizers satisfying $f(x) \leq f(x^0)$. Specifically, for our problem, the zero vector $x^0 = 0$ would be a trivial initial point for (1.1) with $f(0) = \|b\|^2$, the least-squares

[1] $\|x\|_p$ $(0 < p < 1)$ is a quasinorm which satisfies the norm axioms except the triangle inequality. We call $\|x\|_p$ a norm for simplicity.

solution of $\min_x \|Ax-b\|$ is another choice, and so is a point generated by any heuristic procedure such as the orthogonal matching pursuit (OMP) method. Based on this observation, we use the first necessary condition for a local minimizer to present a lower bound

$$L = \left(\frac{\lambda p}{2\|A\|\sqrt{f(x^0)}}\right)^{\frac{1}{1-p}} \tag{1.6}$$

for the absolute value of nonzero entries in a local optimal solution x^* of (1.1), which satisfies $f(x^*) \le f(x^0)$, that is,

$$\text{for any} \quad i \in \mathcal{N}, \qquad L \le |x_i^*| \qquad \text{if } x_i^* \neq 0.$$

Moreover, we show that the number of nonzero entries in every local optimal solution x^* satisfying $f(x^*) \le f(x^0)$ is bounded by

$$\|x^*\|_0 \le \min\left(m, \quad \frac{f(x^0)}{\lambda L^p}\right).$$

The lower bounds in (1.5) and (1.6) are not only useful for identification of zero entries in local optimal solutions from approximation ones, but also for selection of the regularization parameter λ and norm $\|\cdot\|_p$. In particular, for a given norm $\|\cdot\|_p$, the lower bounds can help us to choose the regularization parameter λ for controlling the sparsity level of the solution. On the other hand, for a given λ, the lower bounds can also help us to understand the ℓ_2-ℓ_p problem with different values $p \in (0,1)$.

We need to mention that Nikolova [22] proved a similar bound based on the second order condition in a different context. Nikolova's result is important: it shows that using nonconvex potential functions is good for piecewise constant image restoration. However, the result has not been used in practical algorithms because one needs to solve an optimization problem to construct the bound. On the other hand, our first and second order bounds have explicit closed forms and are easily checked. Moreover, we have found that the bound based on the first order condition seems more effective in practice, especially when a good initial point x^0 is chosen.

Our second main contribution is on some numerical issues for solving (1.1). The ℓ_p norm $\|\cdot\|_p$ for $0 < p < 1$ is neither convex nor Lipschitz continuous. Solving the nonconvex, non-Lipschitz continuous minimization problem (1.1) is difficult.

Most optimization algorithms are efficient only for smooth and convex problems. Nevertheless, some algorithms for nonsmooth and nonconvex optimization problems have been developed recently [4, 12, 20, 29]. However, the Lipschitz continuity remains a necessary condition to define the Clarke subgradient [13] in these algorithms. To overcome the non-Lipschitz continuity, some approximation methods have been considered for solving (1.1). For example, at the kth iteration, replace $\|x\|_p^p$ by the following terms [5, 7, 23]:

$$\sum_{i=1}^n \frac{x_i^2}{((x_i^{k-1})^2+\varepsilon_i)^{1-p/2}}, \qquad \sum_{i=1}^n (|x_i|+\varepsilon_i)^p, \qquad \text{or} \qquad \sum_{i=1}^n \frac{|x_i|}{(|x_i^{k-1}|+\varepsilon_i)^{1-p}}.$$

Here $\varepsilon \in R^n$ is a small positive vector. The question is: are there error bounds for verifying accuracy of numerical solutions of these approximation methods? We have resolved this question by developing several error bounds.

More precisely, we consider smoothing methods for nonconvex, nonsmooth optimization problems, for example, the smoothing gradient (SG) method [29]. We choose a smoothing function $s_\mu(t)$ of $|t|$ such that s_μ^p is continuously differentiable for any fixed scalar $\mu > 0$ and satisfies

$$0 \le (s_\mu(t))^p - |t|^p \le \left(\frac{\mu}{2}\right)^p.$$

See section 3. Let the smoothing objective function of f be

$$f_\mu(x) := \|Ax - b\|_2^2 + \sum_{i=1}^n (s_\mu(x_i))^p.$$

We can show that the solution x_μ^* of the smoothing nonconvex minimization problem

$$\min_{x \in R^n} f_\mu(x) \tag{1.7}$$

converges to a solution of (1.1) as $\mu \to 0$. For some small $\mu < \frac{L}{2}$, let

$$(\bar{x}_\mu^*)_i = \begin{cases} 0 & \text{if } |(x_\mu^*)_i| \le \mu, \\ (x_\mu^*)_i & \text{otherwise.} \end{cases}$$

We show that there is x^* in the solution set of (1.1) such that

$$(\bar{x}_\mu^*)_i = 0 \quad \text{if and only if} \quad x_i^* = 0, \quad i \in \mathcal{N},$$

and

$$\|\bar{x}_\mu^* - x^*\| \le \kappa \|\nabla f_\mu(\bar{x}_\mu^*)\|, \tag{1.8}$$

where κ is a computable constant.

To demonstrate the significance of the absolute lower bounds (1.5) and (1.6) and the error bound (1.8), we propose a hybrid orthogonal matching pursuit-smoothing gradient (OMP-SG) method for the nonconvex, non-Lipschitz ℓ_2-ℓ_p minimization problem (1.1). We first use the OMP method to select candidates of nonzero entries in the solution. Next we use the SG method in [29] to find an approximate solution of (1.1). Both before and after the SG method, we use the lower bound theory to identify zero entries in the solution.

Our preliminary numerical results show that using OMP-SG with elimination of small entries in the numerical solution by the lower bounds for the ℓ_2-ℓ_p minimization problem (1.1) can provide more sparse solutions with smaller predictor errors compared with several well-known approaches for variable selection.

This paper is organized as follows. In section 2, we present absolute lower bounds (1.5) and (1.6) for nonzero entries in local solutions of ℓ_2-ℓ_p minimization problem (1.1). In section 3, we present the computable error bound (1.8) for numerical solutions. In section 4, we give the hybrid OMP-SG method for solving the ℓ_2-ℓ_p minimization problem (1.1). Numerical results are given to demonstrate the effectiveness of the lower bounds, the error bounds, and the OMP-SG method.

Notations. Throughout the paper, $\|\cdot\|$ denotes the ℓ_2 norm, and $|\cdot|$ denotes the vector of the componentwise absolute value. For any $x, y \in R^n$, $x \cdot y$ represents the vector $(x_1y_1, \ldots, x_ny_n)^T$, and x^Ty denotes the inner product. Let $\mathcal{X}_p^*$ denote the set of local minimizers of (1.1). For a vector $x \in R^n$, $\text{support}(x) = \{\, i \in \mathcal{N} \mid x_i \neq 0 \,\}$ denotes the support set of x.

2. Lower bounds for nonzero entries in solutions. In this section we present two lower bounds for nonzero entries in local solutions of the ℓ_2-ℓ_p minimization problem (1.1).

Since $f(x) \geq \lambda \|x\|_p^p$, the objective function $f(x)$ is bounded below and $f(x) \to \infty$ if $\|x\| \to \infty$. Moreover, the set $\mathcal{X}_p^*$ of local minimizers of (1.1) is nonempty and bounded.

THEOREM 2.1 (the second order bound). *Let $L_i = (\frac{\lambda p(1-p)}{2\|a_i\|^2})^{\frac{1}{2-p}}, i \in \mathcal{N}$. Then, for any $x^* \in \mathcal{X}_p^*$, the following statements hold:*

(1)

$$\text{For any} \quad i \in \mathcal{N}, \qquad x_i^* \in (-L_i, L_i) \quad \Rightarrow \quad x_i^* = 0.$$

(2) *The columns of the submatrix $B := A_\Lambda \in R^{m \times |\Lambda|}$ of A are linearly independent, where $\Lambda = support(x^*)$ and $|\Lambda| = \|x^*\|_0$ is the cardinality of the set Λ.*

(3)

$$\|B^T A(x^* - b)\| \leq \frac{\lambda p}{2} \cdot \sqrt{\|x^*\|_0} \left(\min_{1 \leq i \leq \|x^*\|_0} L_i \right)^{p-1}.$$

In particular, if $\|a_i\| = 1$ for all $i \in \mathcal{N}$ (that is, A is columnwise normalized), then

$$\|B^T A(x^* - b)\| \leq \sqrt{\|x^*\|_0} \left(\frac{\lambda p}{2} \right)^{\frac{1}{2-p}} \left(\frac{1}{1-p} \right)^{\frac{1-p}{2-p}}.$$

(4)

$$\|x^*\| \leq \|(B^T B)^{-1} B^T b\| + \frac{\lambda p}{2} \|(B^T B)^{-1}\| \left(\min_{1 \leq i \leq |\Lambda|} L_i \right)^{p-1}.$$

If $\|a_i\| = 1$ for all $i \in \mathcal{N}$, then

$$\|x^*\| \leq \|(B^T B)^{-1} B^T b\| + \|(B^T B)^{-1}\| \left(\frac{\lambda p}{2} \right)^{\frac{1}{2-p}} \left(\frac{1}{1-p} \right)^{\frac{1-p}{2-p}}.$$

Proof. For $x^* \in \mathcal{X}_p^*$, with $\|x^*\|_0 = k$, without loss of generality, we assume

$$x^* = (x_1^*, \ldots, x_k^*, 0, \ldots, 0)^T.$$

Let $z^* = (x_1^*, \ldots, x_k^*)^T$ and $B \in R^{m \times k}$ be the submatrix of A, whose columns are the first k columns of A. Define a function $g : R^k \to R$ by

$$g(z) = \|Bz - b\|^2 + \lambda \|z\|_p^p.$$

We have

$$f(x^*) = \|Ax^* - b\|^2 + \lambda \|x^*\|_p^p = \|Bz^* - b\|^2 + \lambda \|z^*\|_p^p = g(z^*).$$

Since $|z_i^*| > 0, i = 1, \ldots, k$, g is continuously differentiable at z^*. Moreover, in a neighborhood of x^*,

$$\begin{aligned} g(z^*) = f(x^*) &\leq \min\{f(x) \mid x_i = 0, i = k+1, \ldots, n\} \\ &= \min\{g(z) \mid z \in R^k\}, \end{aligned}$$

which implies that z^* is a local minimizer of the function g. Hence the second order necessary condition for

$$\min_{z\in R^k} g(z) \tag{2.1}$$

holds at z^*.

(1) The second order necessary condition at z^* gives that the matrix

$$2B^TB + \lambda p(p-1)\text{diag}(|z^*|^{p-2})$$

is positive semidefinite. Therefore, we obtain

$$2e_i^TB^TBe_i + \lambda p(p-1)|z^*|_i^{p-2} \geq 0, \quad i = 1,\dots,k,$$

where e_i is the ith column of the identity matrix of $R^{k\times k}$.

Note that $\|a_i\|^2 = e_i^TB^TBe_i$. We find that

$$|z^*|_i^{p-2} \leq \frac{2\|a_i\|^2}{\lambda p(1-p)}, \quad i = 1,\dots,k,$$

which implies that

$$|z^*|_i \geq \left(\frac{\lambda p(1-p)}{2\|a_i\|^2}\right)^{\frac{1}{2-p}} = L_i, \quad i = 1,\dots,k.$$

Hence for any $x^* \in \mathcal{X}_p^*$, if $x_i^* \neq 0$, $i \in \mathcal{N}$, then $|x_i^*| \geq L_i$. This is equivalent to the statement that, if $x_i^* \in (-L_i, L_i)$, $i \in \mathcal{N}$, then $x_i^* = 0$.

(2) Since the matrix $2B^TB + \lambda p(p-1)\text{diag}(|z^*|^{p-2})$ is positive semidefinite and $\lambda p(p-1)\text{diag}(|z^*|^{p-2})$ is negative definite, the matrix B^TB must be positive definite. Hence the columns of B must be linearly independent.

(3) Since z^* is a local minimizer of g, the first order necessary condition must hold at z^*. Hence, we find, with $Bz^* = Ax^*$,

$$\|B^T(Ax^*-b)\| = \|B^T(Bz^*-b)\| = \frac{\lambda p}{2}\||z^*|^{p-1}\| \leq \frac{\lambda p}{2}\cdot\sqrt{|\Lambda|}\left(\min_{1\leq i\leq|\Lambda|} L_i\right)^{p-1}.$$

If $\|a_i\| = 1$ for all $i \in \mathcal{N}$, then $L_i = (\frac{\lambda p(1-p)}{2})^{\frac{1}{2-p}}$ for all $i \in \Lambda$, which implies (3).

(4) The first order necessary condition for (2.1) yields

$$2B^TBz^* = 2B^Tb - \lambda p|z^*|^{p-1}\text{sign}(z^*).$$

From (1) and (2) of this theorem, we know that $|z_i^*| \geq L_i$ and that B^TB is nonsingular. Hence, we obtain the desired results. For the case where $\|a_i\| = 1$, $i \in \mathcal{N}$, we have

$$\begin{aligned}\|x^*\| = \|z^*\| &\leq \|(B^TB)^{-1}B^Tb\| + \|(B^TB)^{-1}\|\frac{1}{2}\lambda p\||z^*|^{p-1}\| \\ &\leq \|(B^TB)^{-1}B^Tb\| + \|(B^TB)^{-1}\|\left(\frac{\lambda p}{2}\right)^{\frac{1}{2-p}}\left(\frac{1}{1-p}\right)^{\frac{1-p}{2-p}}. \qquad \square\end{aligned}$$

COROLLARY 2.1. *The set $\mathcal{X}_p^*$ of local minimizers of problem* (1.1) *has a finite number of elements. Moreover, we have*

$$\mathcal{X}_p^* \subseteq \left\{x \mid \|x\| \leq \sigma\|A^Tb\| + \sigma\frac{\lambda p}{2}\left(\min_{1\leq i\leq|\Lambda|} L_i\right)^{p-1}\right\},$$

where

$$\sigma = \max\{ \|(B^T B)^{-1}\| \mid B \in R^{m\times k},\ \operatorname{rank}(B) = k, \quad B \textit{ lies in the columns of } A\}$$

and $k = \operatorname{rank}(A) \le \min(m, n)$.

Proof. From (2) of Theorem 2.1, we find that $\mathcal{X}_p^*$ has a finite number of elements as there are at most $\binom{n}{m}$ possible matrices B, and the linear independence of the columns guarantees that at most one local minimizer exists for each matrix.

It is known that, for any two sets $\{\hat{a}_i, i = 1,\dots,\ell\} \subseteq \{a_i, i = 1,\dots,k\}$, the matrices $\hat{B} \in R^{m\times\ell}$ and $B \in R^{m\times k}$ whose columns lie on the two sets, respectively, satisfy

$$\lambda_{\min}(B^T B) \le \lambda_{\min}(\hat{B}^T \hat{B}),$$

where $\lambda_{\min}$ denotes the smallest eigenvalue. Thus $\|(B^T B)^{-1}\| \le \sigma$ for any matrix B arising from an $x^* \in \mathcal{X}_p^*$.

From (4) of Theorem 2.1 and from $\|B^T b\| \le \|A^T b\|$, we find the closed ball containing $\mathcal{X}_p^*$ in this corollary. □

Remark 2.1. Note that Theorem 2.1 holds for all local minimizers of (1.1). Result (1) of Theorem 2.1 presents a lower bound theory of nonzero entries in local minimizers of (1.1). Result (2) implies that columns of A corresponding to nonzero entries of x^* must form a basis as long as $0 < p < 1$, while bound (3) shows that x^* approaches the least-squares solution of $\min_x \|Ax - b\|$ (restricted to the support of x^*) as $\lambda \to 0$. Corollary 2.1 points out that (1.1) has a finite number of local minimizers and presents a closed ball which contains all local minimizers of (1.1) and an upper bound for all nonzero entries in any local minimizer.

As we mentioned before, most minimization algorithms are descent-iterative in nature, that is, they generate a sequence of points x^k, $k = 0, 1, \dots$, such that the objective values $f(x^k)$ are strictly decreasing along the sequence. Thus, any local minimizer, including the global minimizer, that a descent algorithm may find must be in the level set $\{x : f(x) \le f(x^0)\}$, where x^0 is any given initial point. Therefore, in both theory and practice, one may be interested in only the minimizers satisfying $f(x) \le f(x^0)$. Indeed, our next theorem presents a lower bound theory of nonzero entries for any local minimizer x^* of (1.1) in $\{x : f(x) \le f(x^0)\}$ and derives an upper bound on $\|x^*\|_0$. The upper bound indicates that, for λ sufficiently large but finite, $\|x^*\|_0$ reduces to 0 for $0 < p < 1$, which means that $x^* = 0$ is the only global minimizer.

THEOREM 2.2 (the first order bound). *Let* x^* *be any local minimizer of* (1.1) *satisfying* $f(x^*) \le f(x^0)$ *for an arbitrarily given initial point* x^0. *Let* $L = \left(\frac{\lambda p}{2\|A\|\sqrt{f(x^0)}}\right)^{\frac{1}{1-p}}$. *Then we have,*

$$\textit{for any} \quad i \in \mathcal{N}, \qquad x_i^* \in (-L, L) \quad \Rightarrow \quad x_i^* = 0.$$

Moreover, the number of nonzero entries in x^* *is bounded by*

$$\|x^*\|_0 \le \min\left(m, \frac{f(x^0)}{\lambda L^p}\right). \tag{2.2}$$

Proof. Suppose $f(x^*) \le f(x^0)$, $x^* \in \mathcal{X}_p^*$. Then we have

$$\begin{aligned} \|A^T(Ax^*-b)\|^2 \le \|A^T\|^2\|Ax^*-b\|^2 &\le \|A^T\|^2(\|Ax^*-b\|^2+\lambda\|x^*\|_p^p) \\ &= \|A^T\|^2 f(x^*) \le \|A^T\|^2 f(x^0). \end{aligned} \tag{2.3}$$

Recall the function g in the proof of Theorem 2.1. The first order necessary condition for

$$\min_{z\in R^k} g(z)$$

at z^* gives

$$2B^T(Bz^*-b)+\lambda p(|z^*|^{p-1}\cdot \operatorname{sign}(z^*))=0.$$

This, together with (2.3), implies

$$\lambda p\big\||z^*|^{p-1}\big\| = 2\big\|B^T(Bz^*-b)\big\| = 2\big\|B^T(Ax^*-b)\big\| \le 2\big\|A^T(Ax^*-b)\big\| \le 2\|A\|\sqrt{f(x^0)}.$$

Therefore, we obtain

$$2\|A\|\sqrt{f(x^0)} \ge \lambda p\big\||z^*|^{p-1}\big\| \ge \lambda p(\min_{1\le i\le k}|z_i^*|)^{p-1}.$$

Note that $p-1<0$. We find

$$\min_{1\le i\le k}|z_i^*| \ge \left(\frac{\lambda p}{2\|A\|\sqrt{f(x^0)}}\right)^{\frac{1}{1-p}} = L.$$

Hence, all nonzero components of x^* are no less than L. In other words, for $i\in\mathcal{N}$, if $x_i^*\in(-L,L)$, then $x_i^*=0$.

Now we show the second part of the theorem. Again

$$\lambda\|x^*\|_p^p \le \|Ax^*-b\|+\lambda\|x^*\|_p^p = f(x^*) \le f(x^0).$$

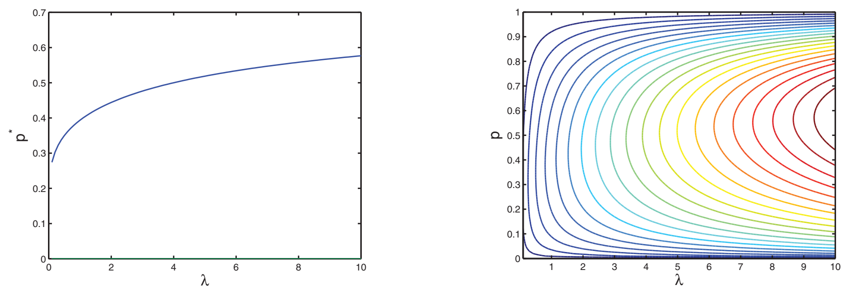
From the first part of this theorem, any nonzero entry of x^* is bounded from below by L. Thus, together with (2) of Theorem 2.1, they imply the desired bound in (2.2). ∎

The lower bound in Theorem 2.1 depends on the parameters λ, p, and the matrix A, while the lower bound in Theorem 2.2 depends on λ, p, A, and the initial objective value $f(x^0)$. In practice, we can take the maximum value of the two bounds to get a new bound. Moreover, in Theorem 2.2, one may simply set $x^0=0$, the trivial local minimizer of (1.1) (so that $f(x)\le f(x^0)=\|b\|^2$), the minimizer of $\|Ax-b\|$, or a point generated by any heuristic procedure such as the OMP method. It is worth noting that x^0 can be replaced by x^* in Theorem 2.2, and the theorem remains true.

The lower bound theory can be extended to the following problem:

$$\min_{x\in R^n}\|Ax-b\|^2+\lambda\sum_{i=1}^{r}\varphi(d_i^Tx_i), \tag{2.4}$$

where $D\in R^{r\times n}$ is the first or second order difference matrix with rows d_i and φ is a non-Lipschitz potential function; see Table 4.6. In fact, as we mentioned earlier,



(a) (b)

FIG. 1. (a) $p^*(\lambda) = \arg\max_{0\le p\le 1} L(\lambda, p)$. (b) $L(\lambda, p) = (\lambda p(1-p))^{\frac{1}{2-p}}$.

Nikolova [22] proved that there is $\theta > 0$ such that every local minimizer x^* of (2.4) satisfies

$$\text{either} \quad |d_i^T x^*| = 0 \quad \text{or} \quad |d_i^T x^*| \ge \theta$$

by using the second order necessary condition for (2.4). However, the result has not been used in practical algorithms because one needs to solve an optimization problem to construct θ. Nikolova [22] also stated that it is difficult to get an explicit solution from the optimization problem for constructing θ.

The lower bounds (1.5) and (1.6) clearly show the relationship between the sparsity of the solution and the choice of the regularization parameter λ and norm $\|\cdot\|_p$. Hence our lower bound theory can be used for selecting model parameters λ and p. In Figure 1, we show some properties of the function $L(\lambda, p) = (\lambda p(1-p))^{\frac{1}{2-p}}$ for $\lambda = (0, 10]$ and $p \in [0, 1]$.

From Figure 1, we can see clearly that, for any given $\lambda > 0$, $(\lambda p(1-p))^{\frac{1}{2-p}}$ is a nonnegative and concave function of p on $[0, 1]$. It takes the minimum value at $p = 0$ and $p = 1$ for any $\lambda \in (0, 10]$.

3. Error bounds derived from the lower bound theory. Smoothing approximations are widely used in optimization and scientific computing [11]. In the following, we consider a smoothing function of f and give a smooth version of Theorems 2.1 and 2.2.

For $\mu \in (0, \infty)$, let

$$s_\mu(t) = \begin{cases} |t|, & |t| > \mu, \\ \dfrac{t^2}{2\mu} + \dfrac{\mu}{2}, & |t| \le \mu. \end{cases}$$

Then $s_\mu(t)$ is continuously differentiable, and

$$((s_\mu(t))^p)' = \begin{cases} p|t|^{p-1}\text{sign}(t), & |t| > \mu, \\ p\left(\dfrac{t^2}{2\mu} + \dfrac{\mu}{2}\right)^{p-1} \dfrac{t}{\mu}, & |t| \le \mu. \end{cases}$$

However, $s_\mu(t)$ is not twice differentiable at $t = \mu$. For $t \in (-\mu, \mu)$, the second derivative of $(s_\mu(t))^p$ satisfies

$$(3.1)\qquad \begin{aligned}(s_\mu(t)^p)'' &= p(p-1)\left(\frac{t^2}{2\mu}+\frac{\mu}{2}\right)^{p-2}\left(\frac{t}{\mu}\right)^2 + p\left(\frac{t^2}{2\mu}+\frac{\mu}{2}\right)^{p-1}\frac{1}{\mu}\\ &\ge p(p-1)\left(\frac{t^2}{2\mu}+\frac{\mu}{2}\right)^{p-2}\left(\frac{t^2}{2\mu}+\frac{\mu}{2}\right)\frac{1}{\mu} + p\left(\frac{t^2}{2\mu}+\frac{\mu}{2}\right)^{p-1}\frac{1}{\mu}\\ &= p^2\left(\frac{t^2}{2\mu}+\frac{\mu}{2}\right)^{p-1}\frac{1}{\mu} > 0.\end{aligned}$$

Hence $s_\mu^p(t)$ is strictly convex in $(-\mu,\mu)$. Moreover, from $s_\mu(t)=|t|(\frac{t^2+\mu^2}{2\mu|t|})\ge|t|$ and $0=\operatorname{argmax}_{t\in(-\mu,\mu)}(s_\mu(t)-|t|)$, we have that, for any $t\in R$,

$$(3.2)\qquad 0\le (s_\mu(t))^p-|t|^p\le\left(\frac{\mu}{2}\right)^p.$$

Let

$$\psi_\mu(x)=(s_\mu(x_1),\dots,s_\mu(x_n))^T$$

and

$$\Psi_\mu(x)=\left(((s_\mu(x_1))^p)',\dots,((s_\mu(x_n))^p)'\right)^T.$$

We define a smoothing approximation of the objective function $f(x)$

$$f_\mu(x)=\|Ax-b\|^2+\lambda\|\psi_\mu(x)\|_p^p,$$

and we consider the smooth minimization problem (1.7). The smoothing objective function f_μ is continuously differentiable in R^n and strictly convex on the set $\{\,x\,|\,\|x\|_\infty\le\mu\,\}$.

Let $\mathcal{X}^*_{p,\mu}$ denote the set of local minimizers of (1.7). By the definition of ψ_μ and (3.2), for any x, we have

$$\lambda n\left(\frac{\mu}{2}\right)^p\ge f_\mu(x)-f(x)\ge 0.$$

Since $\|x\|\to\infty$ implies $f(x)\to\infty$, we deduce $f_\mu(x)\to\infty$ if $\|x\|\to\infty$. Moreover, for any $x\in R^n$, $\lim_{\mu\downarrow 0}f_\mu(x)=f(x)$. The following theorem presents the smooth version of the first and second order lower bounds.

THEOREM 3.1. *Let* $L=(\frac{\lambda p}{2\|A\|\sqrt{f(x^0)}})^{\frac{1}{1-p}}$ *for an arbitrarily given initial point* x^0 *and* $L_i=(\frac{\lambda p(1-p)}{2\|a_i\|^2})^{\frac{1}{2-p}}, i\in\mathcal{N}$.

(1) *(the second order bound) For any* $\mu>0$ *and any* $x^*_\mu\in\mathcal{X}^*_{p,\mu}$*, we have,*

$$\text{for any}\quad i\in\mathcal{N},\qquad (x^*_\mu)_i\in(-L_i,L_i)\quad\Rightarrow\quad |(x^*_\mu)_i|\le\mu.$$

(2) *(the first order bound) For any* $\mu>0$ *and any* $x^*_\mu\in\mathcal{X}^*_{p,\mu}$ *satisfying* $f(x^*_\mu)\le f(x^0)$*, we have,*

$$\text{for any}\quad i\in\mathcal{N},\qquad (x^*_\mu)_i\in(-L,L)\quad\Rightarrow\quad |(x^*_\mu)_i|\le\mu.$$

Proof. (1) Since $x^*_\mu \in \mathcal{X}^*_{p,\mu}$, the second order necessary condition for (1.7) implies that the matrix

$$\nabla^2 f_\mu(x^*_\mu) = 2A^TA + \lambda\Psi'_\mu(x)$$

is positive semidefinite. Suppose $|(x^*_\mu)_i| > \mu$. Then from

$$e_i^T(2A^TA + \lambda\Psi'_\mu(x))e_i = 2\|a_i\|^2 + \lambda p(p-1)|(x^*_\mu)_i|^{p-2} \geq 0,$$

we can get

$$|(x^*_\mu)_i| \geq \left(\frac{\lambda p(1-p)}{2\|a_i\|^2}\right)^{\frac{1}{2-p}} = L_i.$$

Since $\mu > 0$ and $x^*_\mu \in \mathcal{X}^*_{p,\mu}$ are arbitrarily chosen, we can claim that, for any $\mu > 0$ and $x^*_\mu \in \mathcal{X}^*_{p,\mu}$, if $(x^*_\mu)_i \in (-L_i, L_i)$, $i \in \mathcal{N}$, then $|(x^*_\mu)_i| \leq \mu$.

(2) Since $x^*_\mu \in \mathcal{X}^*_{p,\mu}$, the first order necessary condition for (1.7) gives

$$\nabla f_\mu(x^*_\mu) = 2(A^TAx^*_\mu - A^Tb) + \lambda\Psi_\mu(x^*_\mu) = 0, \tag{3.3}$$

which, together with $f(x^*_\mu) \leq f(x^0)$, implies

$$\|\lambda\Psi_\mu(x^*_\mu)\|^2 \leq 4\|A^T\|^2(\|Ax^*_\mu - b\|^2 + \|x^*_\mu\|_p^p) = 4\|A\|^2 f(x^*_\mu) \leq 4\|A\|^2 f(x^0). \tag{3.4}$$

Suppose $|(x^*_\mu)_i| > \mu$. Then

$$\lambda\|\Psi_\mu(x^*_\mu)\| \geq \lambda|\Psi_\mu(x^*_\mu)_i| = \lambda p|(x^*_\mu)_i|^{p-1}. \tag{3.5}$$

From (3.4) and (3.5), we can get

$$|(x^*_\mu)_i|^{p-1} \leq \frac{2\|A\|\sqrt{f(x^0)}}{\lambda p}.$$

Note that $p - 1 < 0$. We find

$$|(x^*_\mu)_i| \geq \left(\frac{\lambda p}{2\|A\|\sqrt{f(x^0)}}\right)^{\frac{1}{1-p}} = L.$$

Hence we can claim that, for $i \in \mathcal{N}$, if $(x^*_\mu)_i \in (-L, L)$, then $|(x^*_\mu)_i| \leq \mu$. ∎

The function f is not Lipschitz continuous. We define the first order and second order necessary conditions for (1.1) as follows.

DEFINITION 3.1. *For $x \in R^n$, let $X = \mathrm{diag}(x)$.*

(1) *x is said to satisfy the first order necessary condition of* (1.1) *if*

$$2XA^T(Ax - b) + \lambda p|x|^p = 0. \tag{3.6}$$

(2) *x is said to satisfy the second order necessary condition of* (1.1) *if*

$$2XA^TAX + \lambda p(p-1)\mathrm{diag}(|x|^p) \tag{3.7}$$

is positive semidefinite.

Obviously, the zero vector in R^n satisfies the first and second necessary conditions of (1.1).

Let $\{x_{\mu_k}\}$ denote a sequence with $\mu_k > 0$, $k = 1, 2, \ldots,$ and $\mu_k \to 0$ as $k \to \infty$.

THEOREM 3.2.

(1) *Let* $\{x_{\mu_k}\}$ *be a sequence of vectors satisfying the first order necessary condition of* (1.7). *Then any accumulation point of* $\{x_{\mu_k}\}$ *satisfies the first order necessary condition of* (1.1).

(2) *Let* $\{x_{\mu_k}\}$ *be a sequence of vectors satisfying the second order necessary condition of* (1.7). *Then any accumulation point of* $\{x_{\mu_k}\}$ *satisfies the second order necessary condition of* (1.1).

(3) *Let* $\{x_{\mu_k}\}$ *be a sequence of vectors being global minimizers of* (1.7). *Then any accumulation point of* $\{x_{\mu_k}\}$ *is a global minimizer of* (1.1).

Proof. Let $\bar{x}$ be an accumulation point of $\{x_{\mu_k}\}$. By working on a subsequence, we may assume that $\{x_{\mu_k}\}$ converges to $\bar{x}$. Let $X_{\mu_k} =$diag(x_{μ_k}) and $\bar{X} =$diag$(\bar{x})$.

(1) From the first order necessary condition (3.3) of (1.7), we have

$$X_{\mu_k} \nabla f_{\mu_k}(x_{\mu_k}) = 2X_{\mu_k}(A^T A x_{\mu_k} - A^T b) + \lambda X_{\mu_k} \Psi_\mu(x_{\mu_k}) = 0.$$

By the definition of Ψ_μ, we have

$$(X_{\mu_k} \Psi_{\mu_k}(x_{\mu_k}))_i = p|x_{\mu_k}|_i^p \qquad \text{if} \quad |x_{\mu_k}|_i > \mu_k$$

and

$$\begin{aligned} 0 \le (X_{\mu_k} \Psi_{\mu_k}(x_{\mu_k}))_i &= p\left(\frac{(x_{\mu_k})_i^2}{2\mu_k} + \frac{\mu_k}{2}\right)^{p-1} \frac{(x_{\mu_k})_i^2}{\mu_k} \\ &\le p\left(\frac{(x_{\mu_k})_i^2}{\mu_k}\right)^p \le p|x_{\mu_k}|_i^p \qquad \text{if} \quad |x_{\mu_k}|_i \le \mu_k. \end{aligned}$$

If $|x_{\mu_k}|_i \le \mu_k$ for arbitrarily large k, then $\bar{x}_i = \lim_{k\to\infty}(x_{\mu_k})_i = 0$, and $\lim_{k\to\infty} (X_{\mu_k}\Psi_{\mu_k}(x_{\mu_k}))_i = 0$. Therefore, we have

$$0 = 2 \lim_{k\to\infty} X_{\mu_k}(A^T A x_{\mu_k} - A^T b) + \lambda \lim_{k\to\infty} X_{\mu_k} \Psi_{\mu_k}(x_{\mu_k}) = \bar{X}(A^T A\bar{x} - A^T b) + \lambda p|\bar{x}|^p.$$

Hence $\bar{x}$ satisfies the first order necessary condition of (1.1).

(2) From the second order necessary condition of (1.7), we have

$$X_{\mu_k} \nabla^2 f(x_{\mu_k}) X_{\mu_k} = 2X_{\mu_k} A^T A X_{\mu_k} + \lambda X_{\mu_k} \Psi'_{\mu_k}(x_{\mu_k}) X_{\mu_k}$$

is positive semidefinite. Using the definition of Ψ_μ and (3.1), we have

$$(X_{\mu_k} \Psi'_{\mu_k}(x_{\mu_k}) X_{\mu_k})_{ii} = p(p-1)|x_{\mu_k}|_i^p \qquad \text{if} \quad |x_{\mu_k}|_i > \mu_k$$

and

$$\begin{aligned} 0 &< (X_{\mu_k} \Psi'_{\mu_k}(x_{\mu_k}) X_{\mu_k})_{ii} \\ &= p(p-1)\left(\frac{(x_{\mu_k})_i^2}{2\mu_k} + \frac{\mu_k}{2}\right)^{p-2} \frac{(x_{\mu_k})_i^4}{\mu_k^2} + p\left(\frac{(x_{\mu_k})_i^2}{2\mu_k} + \frac{\mu_k}{2}\right)^{p-1} \frac{(x_{\mu_k})_i^2}{\mu_k} \\ &\le p\left(\frac{(x_{\mu_k})_i^2}{2\mu_k} + \frac{\mu_k}{2}\right)^{p-1} \frac{(x_{\mu_k})_i^2}{\mu_k} \le p\left(\frac{(x_{\mu_k})_i^2}{2\mu_k} + \frac{(x_{\mu_k})_i^2}{2\mu_k}\right)^{p-1} \frac{(x_{\mu_k})_i^2}{\mu_k} \\ &\le p\left(\frac{(x_{\mu_k})_i^2}{\mu_k}\right)^p \le p\mu_k^p \qquad \text{if} \quad |x_{\mu_k}|_i \le \mu_k. \end{aligned}$$

Therefore, for any $y \in R^n$, we have

$$\begin{aligned} 0 &\le \lim_{k\to\infty} y^T \left(2X_{\mu_k}A^TAX_{\mu_k} + \lambda X_{\mu_k}\Psi'_{\mu_k}(x_{\mu_k})X_{\mu_k}\right) y \\ &= y^T \left(2\lim_{k\to\infty} X_{\mu_k}A^TAX_{\mu_k} + \lambda \lim_{k\to\infty} X_{\mu_k}\Psi'_{\mu_k}(x_{\mu_k})X_{\mu_k}\right) y \\ &= y^T \left(2\bar{X}A^TA\bar{X} + \lambda p(p-1)\text{diag}(|\bar{x}|^p)\right) y. \end{aligned}$$

Hence $\bar{x}$ satisfies the second order necessary condition for (1.1).

(3) Let x^* be a global minimizer of (1.1). Then, from the following three inequalities

$$f(x_{\mu_k}) \le f_{\mu_k}(x_{\mu_k}) \le f_{\mu_k}(x^*) \le f(x^*) + \lambda n \left(\frac{\mu_k}{2}\right)^p,$$

we deduce that $\bar{x}$ is a global minimizer of (1.1). ☐

In the following, we present a computable error bound for KKT solutions (satisfying the first order necessary condition) of the smooth minimization problem (1.7) to approximate a KKT solution of the non-Lipschitz optimization problem (1.1).

Let $\mathcal{X}_{p,\mu}$ be the set of KKT solutions of (1.7) and $\mathcal{X}_p$ be the set of KKT solutions of (1.1).

THEOREM 3.3. *Let $\{x_{\mu_k}\}$ be a sequence of vectors satisfying the first order necessary condition of* (1.7) *and $f(x_{\mu_k}) \le f(x^0)$ for an arbitrarily given initial point x^0. Then there is a $K > 0$ such that, for any $k \ge K$, there is $x^* \in \mathcal{X}_p$ such that*

$$\Gamma_{\mu_k} := \{i \in \mathcal{N} \;\;\big|\;\; |(x_{\mu_k})_i| \le \mu_k\} = \{i \in \mathcal{N} \;\;\big|\;\; |x_i^*| = 0\} =: \Gamma. \tag{3.8}$$

Define

$$(\bar{x}^*_{\mu_k})_i = \begin{cases} 0, & i \in \Gamma, \\ (x_{\mu_k})_i, & i \in \mathcal{N}\setminus\Gamma. \end{cases} \tag{3.9}$$

Let B be the submatrix of A whose columns are indicated by $\mathcal{N}\setminus\Gamma$. Suppose $\lambda_{\min}(B^TB) > \frac{\lambda p(1-p)}{2}L^{p-2}$. Then

$$\|\bar{x}^*_{\mu_k} - x^*\| \le \|G^{-1}\|\|\nabla f_{\mu_k}(\bar{x}^*_{\mu_k})\|, \tag{3.10}$$

where $G = 2B^TB + \lambda p(p-1)L^{p-2}I$ and $\lambda_{\min}(B^TB)$ denotes the smallest eigenvalue of the matrix B^TB.

Proof. Since the level set $\{x \,|\, f(x) \le f(x^0)\}$ is bounded, the sequence $\{x_{\mu_k}\}$ is bounded. From (1) of Theorem 3.2, any accumulation point of $\{x_{\mu_k}\}$ is in $\mathcal{X}_p$. Hence we have $\lim_{k\to\infty} \text{dist}(x_{\mu_k}, \mathcal{X}_p) = 0$. This implies that there is $x^* \in \mathcal{X}_p$ such that $\lim_{k\to\infty} x_{\mu_k} = x^*$ and that there is $K > 0$ such that, for $k \ge K$, $\mu_k < \frac{L}{2}$,

$$\text{dist}(x_{\mu_k}, \mathcal{X}_p) = \|x_{\mu_k} - x^*\| < \frac{L}{2},$$

and $f(x^*) \le f(x^0)$ hold. Then

$$|x_i^*| - |(x_{\mu_k})_i| \le |x_i^* - (x_{\mu_k})_i| \le \|x^* - x_{\mu_k}\| < \frac{L}{2}.$$

If $i \in \Gamma_{\mu_k}$, that is, $|(x_{\mu_k})_i| \le \mu_k$, then we have

$$|x_i^*| < |(x_{\mu_k})_i| + \frac{L}{2} < L.$$

Assume that $x_i^* \neq 0$. From (3.6), we derive

$$\lambda p L^{p-1} < \lambda p |x_i^*|^{p-1} = 2|A^T(Ax^*-b)|_i \leq 2\|A^T(Ax^*-b)\| \leq 2\|A\|\sqrt{f(x^0)}$$

which implies $L > (\frac{\lambda p}{2\|A\|\sqrt{f(x^0)}})^{1-p} = L$. This is a contradiction. Hence $|x_i^*| = 0$, that is, $i \in \Gamma$. We obtain that $\Gamma_{\mu_k} \subset \Gamma$.

On the other hand, if $i \in \Gamma$, then $x_i^* = 0$. We have

$$\left|(x_{\mu_k})_i\right| = \left|(x^* - x_{\mu_k})_i\right| \leq \|x^* - x_{\mu_k}\| < \frac{L}{2} < L.$$

From Theorem 3.1, we know that $|(x_{\mu_k})_i| \leq \mu_k$, and thus $i \in \Gamma_{\mu_k}$. Hence $\Gamma \subset \Gamma_\mu$. We obtain (3.8).

Without loss of generality, we assume that $\mathcal{N}\backslash\Gamma = \{1, 2, \ldots, r\}$. Define the function $g:\ R^r \to R$ by

$$g(z) = \|Bz - b\|_2^2 + \lambda\|z\|_p^p.$$

The first order necessary condition (3.6) at x^* yields

$$\nabla g(z^*) = 2B^T(Bz^* - b) + \lambda p|z^*|^{p-1} \cdot \mathrm{sign}(z^*) = 0$$

at $z^* = (x_1^*, \ldots, x_r^*)^T$. Furthermore, let $z_{\mu_k} = \big((x_{\mu_k})_1, \ldots, (x_{\mu_k})_r\big)^T$. Then

$$\begin{aligned}\nabla g(z_{\mu_k}) &= \nabla g(z_{\mu_k}) - \nabla g(z^*)\\ &= 2B^TB(z_{\mu_k} - z^*) + \lambda p|z_{\mu_k}|^{p-1} \cdot \mathrm{sign}(z_{\mu_k}) - \lambda p|z^*|^{p-1} \cdot \mathrm{sign}(z^*).\end{aligned}$$

Note that $\mathrm{sign}(z_{\mu_k}) = \mathrm{sign}(z^*)$. By using the mean value theorem, we have

$$\begin{aligned}\nabla g(z_{\mu_k}) &= 2B^TB(z_{\mu_k} - z^*) + \lambda p\ \mathrm{sign}(z_{\mu_k}) \cdot \big(|z_{\mu_k}|^{p-1} - |z^*|^{p-1}\big)\\ &= \big(2B^TB + \lambda p(p-1)\ D\ \big)\big(z_{\mu_k} - z^*\big),\end{aligned} \tag{3.11}$$

where $D \in R^{r\times r}$ is a diagonal matrix whose diagonal elements are $|\tilde{z}_{\mu_k}|_i^{p-2}$, where $(\tilde{z}_{\mu_k})_i$ is between $(z_{\mu_k})_i$ and z_i^*, $i = 1, 2, \ldots, r$. Since $|(z_{\mu_k})_i| \geq L$, $|z_i^*| \geq L$, and $\mathrm{sign}((z_{\mu_k})_i) = \mathrm{sign}((z_i^*))$, we have $|(\tilde{z}_{\mu_k})_i| \geq L, i = 1, 2, \ldots, r$.

Since the matrix $2B^TB + \lambda p(p-1)D$ is symmetric, $0 < p < 1$, and $|(\tilde{z}_{\mu_k})_i| \geq L$ for all $i \in \mathcal{N}\backslash\Gamma$, for any $z \in R^r$ with $\|z\| = 1$, we have

$$\begin{aligned}z^T(2B^TB + \lambda p(p-1)D)z &= z^T(2B^TB)z + \lambda p(p-1)z^TDz\\ &\geq 2z^T(B^TB)z + \lambda p(p-1)L^{p-2}\|z\|^2\\ &\geq 2\lambda_{\min}(B^TB) + \lambda p(p-1)L^{p-2}\\ &> 0,\end{aligned}$$

where the last inequality uses the assumption of this theorem. Hence the matrix $2B^TB + \lambda p(p-1)\ D$ is invertible. We conclude from (3.9) and (3.11) that

$$\begin{aligned}\|\bar{x}_{\mu_k}^* - x^*\| = \|z_{\mu_k} - z^*\| &\leq \|\big(2B^TB + \lambda p(p-1)\ D\big)^{-1}\|\|\nabla g(z_{\mu_k})\|\\ &\leq \|\big(2B^TB + \lambda p(p-1)\ L^{p-2}I\big)^{-1}\|\|\nabla g(z_{\mu_k})\|\\ &= \|G^{-1}\|\|\nabla g(z_{\mu_k})\|\\ &\leq \|G^{-1}\|\|\nabla f_{\mu_k}(\bar{x}_{\mu_k}^*)\|,\end{aligned}$$

where the last inequality uses $\|\nabla g(z_{\mu_k})\| \le \|\nabla f_{\mu_k}(\bar{x}^*_{\mu_k})\|$, which can be shown as

$$\begin{aligned}\|\nabla g(z_{\mu_k})\| &= \|2B^T(Bz^*_{\mu_k} - b) + \lambda p|z_{\mu_k}|^{p-1}\cdot \mathrm{sign}(z_{\mu_k})\| \\ &= \|2B^T(A\bar{x}^*_{\mu_k} - b) + \lambda p|z_{\mu_k}|^{p-1}\cdot \mathrm{sign}(z_{\mu_k})\| \\ &= \|2B^T(A\bar{x}^*_{\mu_k} - b) + \lambda \Psi_{\mu_k}(z_{\mu_k})\| \\ &\le \|2A^T(A\bar{x}^*_{\mu_k} - b) + \lambda \Psi_{\mu_k}(\bar{x}^*_{\mu_k})\| \\ &= \|\nabla f_{\mu_k}(\bar{x}^*_{\mu_k})\|,\end{aligned}$$

where the inequality uses $(\bar{x}^*_{\mu_k})_i = 0$ for $i \in \Gamma$ and $(s^p_{\mu_k})'(0) = 0$. ∎

4. Hybrid OMP-SG algorithm using the lower bound theory. The lower bound theory can be applied to improve existing algorithms and to develop new algorithms. To demonstrate the application, we use a hybrid OMP-SG method to solve the ℓ_2-ℓ_p minimization problem (1.1). More specifically, we employ the OMP method to generate an initial point x^0 and its support, to develop an SG method to further reduce the objective value of (1.1), and finally to apply our theoretical result to purify the numerical solution by deleting its entries with small values. Our limited computational experiment in this section does not intend to develop a new algorithm for sparse reconstruction but to show how our theory could improve any existing algorithm to achieve a higher quality performance.

The OMP algorithm is well known in the literature of signal processing [10, 14, 21, 24, 26, 27]. The following algorithm is a standard version of the OMP algorithm [3] but has a different stop criterion.

Algorithm 1 (Orthogonal Matching Pursuit (OMP)).

Parameters: *Given the $m \times n$ matrix A, the vector $b \in R^m$, and the error threshold β_0.*

Initialization: *Initialize $k = 0$, and set*

- *the initial solution $x^0 = 0$.*
- *the initial residual $r^0 = b - Ax^0 = b$.*
- *the initial solution support $\Lambda_0 = \emptyset$.*

Main iteration: *Increment k by 1, and perform the following steps:*

- *Find the index j_k that solves the optimization problem*

$$j_k = \arg\max \frac{\|(Ax^{k-1} - b)^T a_j\|_2^2}{\|a_j\|} \text{ for } j \in \mathcal{N} \setminus \Lambda_{k-1}.$$

- *Let $\Lambda_k = \Lambda_{k-1} \bigcup \{j_k\}$.*
- *Compute x^k, the minimizer of $\|Ax - b\|_2^2$ subject to* $\mathrm{support}(x) = \Lambda^k$.
- *Calculate the new residual $r^k = Ax^k - b$.*
- *If $\|A^T r^k\| < \beta_0$, stop, and let $\Lambda = \Lambda_k$.*

Output: *A point $x_{omp} := x^k$, a set $\Lambda =$* $\mathrm{support}(x_{omp})$, *and a matrix $B = A_\Lambda \in R^{m\times|\Lambda|}$.*

The SG method [29] is a simple method for Lipschitz continuous but nonsmooth nonconvex minimization problems.

Algorithm 2 (Smoothing Gradient (SG)).

Step 1. *Choose constants σ, $\rho \in (0, 1)$, and an initial point x^0. Set $k = 0$.*

Step 2. *Compute the step size ν_k by the Armijo line search, where $\nu_k = \max\{\rho^0, \rho^1, \dots\}$ and ρ^i satisfies*

$$f_{\mu_k}(x^k - \rho^i g_k) \le f_{\mu_k}(x^k) - \sigma \rho^i g_k^T g_k.$$

Set $x^{k+1} = x^k - \nu_k g_k$. Here $g_k = \nabla f_{\mu_k}(x^k)$.

Step 3. *If* $\|\nabla f_{\mu_k}(x^{k+1})\| \geq n\mu_k$, *then set* $\mu_{k+1} = \mu_k$; *otherwise, choose* $\mu_{k+1} = \sigma\mu_k$.

Now we present the hybrid OMP-SG algorithm for solving ℓ_2-ℓ_p minimization problem (1.1) with the lower bound L defined in (1.6).

ALGORITHM 3 (HYBRID OMP-SG).

Step 1. *Use the OMP algorithm to get* x_{omp}, $\Lambda = \text{support}(x_{omp})$ *and* $B = A_\Lambda \in R^{m\times|\Lambda|}$.

Step 2. *Use the SG algorithm with an initial point* $x^0 = x_{omp}$ *to find*

$$y^* = \arg\min g(y) := \|By - b\|_2^2 + \lambda\|y\|_p^p.$$

Step 3. *Output a numerical solution* x^*, *where*

$$x_j^* = \begin{cases} y_j^*, & |y_j^*| \geq L \quad \text{and} \quad j \in \Lambda, \\ 0 & \text{otherwise.} \end{cases}$$

Remark 4.1. We know that from (2.3) at a solution x^* of (1.1), $f(x^*) \leq f(x^0)$ and $\|A^T(Ax^* - b)\| \leq \|A\|\sqrt{f(x^0)}$ for any $x^0 \in R^n$. From Theorems 2.1 and 2.2, the number of nonzero entries of x^* is less than $\kappa = \min(m, \frac{f(x^0)}{\lambda L^p})$, and each nonzero entry satisfies $|x_i^*| \geq L$. In the hybrid OMP-SG method, we first choose candidates of columns of A which correspond to nonzero entries in a solution of (1.1), and we use these candidates of columns to build a submatrix B. Based on Theorems 2.1 and 2.2, the number of columns of B is chosen slightly larger than κ and the error threshold β_0 is slightly larger than $\|A\|\sqrt{f(x^0)}$. Next we use the globally convergent SG method to find an approximate minimizer of the reduced problem $\min g(y)$. According to Theorem 3.3, we set some entries of the approximate solution to zero if their absolute values are less than L. It is worth noting that the lower bound theory is algorithms independent. For instance, we can replace the SG by the smoothing conjugate gradient (SCG) method [12] in Step 2 of the hybrid OMP-SG method to accelerate the algorithm and have a hybrid orthogonal matching pursuit-smoothing conjugate gradient (OMP-SCG) method.

Now we report numerical results to compare the performance of the hybrid OMP-SG and OMP-SCG methods for solving (1.1) with several other approaches to find sparse solutions. Our preliminary computational results indicate that the variable elimination according to our theory makes a significant difference. The computational test was conducted on a Philips PC (2.36 GHz, 1.96 GB of RAM) using MATLAB 7.4.

We consider the following four approaches.

- LASSO: Solve the ℓ_2-ℓ_1 problem (1.3) by the least-squares algorithm (Lars) proposed in [16].
- ConApp: Solve the ℓ_2-ℓ_p problem (1.1) with $p = \frac{1}{2}$ by using the ℓ_2-ℓ_1 convex approximation [5]

$$\min \|Ax - b\|_2^2 + \lambda \sum_{i=1}^n \frac{|x_i|}{\sqrt{|x_i^{k-1}| + \varepsilon}} \tag{4.1}$$

  at the kth iteration, where $\varepsilon > 0$ is a parameter. We use the Lars to solve (4.1).
- OMP-SG: Solve the ℓ_2-ℓ_p problem (1.1) by the hybrid OMP-SG.
- OMP-SCG: Solve the ℓ_2-ℓ_p problem (1.1) by the hybrid OMP-SCG.

4.1. Variable selection. This example is artificially generated and was first used in Tibshirani [25] to test the effectiveness of LASSO. The true solution is

Table 4.1
Results for variable selection.

m	σ	Approach	MSE	ANZ	NANZ
		LASSO	0.4730	4.77	0.23
40	3	ConApp	0.4688	4.83	0.17
		OMP-SCG	0.4755	4.88	0.12
		LASSO	0.1595	4.77	0.23
40	1	ConApp	0.1541	4.86	0.14
		OMP-SCG	0.1511	4.91	0.09
		LASSO	0.3582	4.92	0.08
60	1	ConApp	0.3503	4.93	0.07
		OMP-SCG	0.3464	4.95	0.05

$x^* = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$. We simulated 100 data sets consisting of m observations from the model

$$Ax = b + \sigma\eta,$$

where η is a noise vector generated by the standard normal distribution. We select three cases to discuss the performance of the three approaches LASSO, ConApp, and OMP-SCG. The first case is $m = 40$, $\sigma = 3$; the second case is $m = 40$, $\sigma = 1$; and the last case is $m = 60$, $\sigma = 1$. We used 80 of the 100 data sets to select the variables and then tested the performance on the remaining 20. The mean squared errors (MSE) over the test set are summarized in Table 4.1. The average number of correctly identified zero coefficients (ANZ) and the average number of the coefficients erroneously set to zero (NANZ) over the test set are also presented in Table 4.1. In our numerical experiment, we used $p = 0.5$ and $\lambda \approx 1.1$ in the ℓ_2-ℓ_p problem (1.1). From Table 4.1, we observe that OMP-SCG performs the best, followed by LASSO and ConApp.

4.2. Signal reconstruction. Signal reconstruction has been studied extensively in the past decades [6, 15]. According to Donoho [15], signal reconstruction can be solved by the ℓ_2-ℓ_1 model (1.3). In this subsection, we apply the ℓ_2-ℓ_p model with $p = 0.5$ to solve signal reconstruction problems.

Consider a real-valued, finite-length signal $x \in R^n$. Suppose x is T-sparse, that is, only T of the signal coefficients are nonzero and the others are zero. We use the following MATLAB code to generate the original signal, a matrix A, and a vector b.

```
x_or =zeros(n,1); q = randperm(n); x_or(q(1:T)) = 2*randn(T,1);
A = randn(m,n); A = orth(A')'; b= A*x_or ; .
```

Our aim is to obtain good reconstructions of x with fewer nonzero entries. We applied OMP-SCG, LASSO, and ConApp to reconstruct the signal. The error between the reconstructed signal and the original one is computed by 2-norm.

In Table 4.2, we present numerical results of three sets of signal examples with different values of L and λ. The CPU time is given in seconds. From Table 4.2, we observe that the three approaches can reconstruct the original signal with $n = 512, T = 60, m = 184$, while OMP-SCG has the highest accuracy. Moreover, LASSO cannot reconstruct the original signal with $n = 512, T = 60, m = 182$, but OMP-SCG and ConApp can reconstruct the original signal, while OMP-SCG has small error. Furthermore, if the original signal has $n = 512, T = 130, m = 225$, the LASSO and ConApp algorithms cannot reconstruct this signal, but OMP-SCG can reconstruct this signal with error = 0.41. OMP-SG gives similar results as OMP-SCG but uses more time.

TABLE 4.2
Results for signal reconstruction without noise.

Problem	LASSO	ConApp	OMP-SCG			
	(Error,Time)	(Error,Time)	L	λ	Error	Time
$n = 512$ $T = 60$ $m = 184$	$(5.33 \times 10^{-4}$, 0.653)	$(1.29 \times 10^{-5}$, 6.82)	0.8	0.002	1.12×10^{-16}	1.02
$n = 512$ $T = 60$ $m = 182$	(38.64, 0.43)	$(2.41 \times 10^{-5}$, 7.84)	0.7	0.001	1.03×10^{-16}	1.34
$n = 512$ $T = 130$ $m = 225$	(122.25, 0.69)	(119.43, 19.99)	0.00001	0.00006	0.41	4.03

TABLE 4.3
Results for signal reconstruction with noise ($n = 512, T = 130, \sigma = 0.1$).

m	Method(s)	MSE	Ratio	CPU	m	Method(s)	MSE	Ratio	CPU
	LASSO	3.71	1.46	3.2541		LASSO	5.34	1.77	1.7519
330	ConApp	3.58	1.41	63.01	310	ConApp	4.10	1.36	60.83
	OMP-SCG	3.42	1.34	5.23		OMP-SCG	4.05	1.34	22.45
	Oracle	*2.5434*				*Oracle*	*3.0180*		
	LASSO	5.30	1.77	2.3011		LASSO	6.1	1.75	2.01
300	ConApp	4.04	1.35	69.12	275	ConApp	5.05	1.45	78.15
	OMP-SCG	3.97	1.33	23.42		OMP-SCG	4.94	1.41	18.83
	Oracle	*2.9845*				*Oracle*	*3.4877*		

Now we add noisy signals to the problem

$$\texttt{b= A*x}_{or}\texttt{ - w ; },$$

where $w = \sigma\eta$ is independent identically distributed Gaussian noise with zero mean and variance σ^2. We measure the quality of a reconstructed signal $\hat{x}$ using the MSE, defined as $E[\|\hat{x} - x_{or}\|^2]$. To compare the capability of algorithms in recovering the original signals under noisy circumstances, we use the oracle estimator, defined as

$$x_{oracle} = \sigma^2 \text{tr}(A_\Lambda^T A_\Lambda)^{-1},$$

where $\Lambda = \text{support}(x_{or})$.

For each algorithm, we calculated the ratio of the MSE of a reconstructed signal generated from the algorithm and the MSE of the oracle estimator, and we listed the results as "Ratio" in Table 4.3. The closer the ratio is to 1, the more robust is the algorithm. From Table 4.3, we can see that the ratio of OMP-SCG is always closer to 1 than the ratios of LASSO and ConApp.

4.3. Prostate cancer. The data set in this subsection is downloaded from the University of California at Irvine (UCI) Standard database [1] for the study of prostate cancer. The data set consists of the medical records of 97 patients who were about to receive a radical prostatectomy. The predictors are eight clinical measures: lcavol, lweight, lage, lbph, svi, lcp, gleason, and pgg45. Detailed explanation can be found in the UCI Standard database. This is a variable selection problem with $A \in R^{97\times 8}$.

TABLE 4.4
Results for prostate cancer.

Parameter	LASSO	Ridge	Best Subset	ConApp	OMP-SG	OMP-SCG
x_1(lcavol)	0.545	0.389	0.740	0.6187	0.6436	0.6436
x_2(lweight)	0.237	0.238	0.367	0.2362	0.2804	0.2804
x_3(lage)	0	-0.029	0	0	0	0
x_4(lbph)	0.098	0.159	0	0.1003	0	0
x_5(svi)	0.165	0.217	0	0.1858	0.1856	0.1857
x_6(lcp)	0	0.026	0	0	0	0
x_7(gleason)	0	0.042	0	0	0	0
x_8(pgg45)	0.059	0.123	0	0	0	0
Number of nonzero	5	8	2	4	3	3
Prediction error	0.478	0.5395	0.5723	0.468	0.4418	0.4419

TABLE 4.5
Error bounds for $\|\bar{x}^*_\mu - x^*\|$.

μ	L	λ	error bound
0.001	0.015	0.1304	1.5793×10^{-5}
0.0001	0.0119	0.1164	5.7310×10^{-6}
0.00001	0.0119	0.1164	5.5721×10^{-6}

One of our main aims is to identify which predictors are most significant in predicting the response.

The prostate cancer data were divided into two parts: a training set with 67 observations and a test set with 30 observations. The prediction error is the MSE over the test set. The numerical results of Ridge regression [19] and Best Subset [2] were derived from [18]. In this example, we also select $p = 0.5$ in the ℓ_2-ℓ_p model (1.1).

From Table 4.4 we find that OMP-SG and OMP-SCG succeed in finding three main factors and have smaller prediction error than ConApp and LASSO. This implies that OMP-SG and OMP-SCG can find more sparse solutions with smaller prediction errors than LASSO.

Now we apply Theorem 3.3 to compute the error bound $\|G^{-1}\|\|\nabla f_\mu(\bar{x}^*_\mu)\|$ of $\bar{x}^*_\mu$ to $x^* \in \mathcal{X}^*_p$ for a given $\mu > 0$. We set $\mu < 0.01$ and $p = 0.5$. The numerical results are listed in Table 4.5.

It is worth noting that the lower bound theory, the error bounds, and the hybrid OMP-SCG method can be extended to

$$\min_{x\in R^n} \|Ax - b\|_2^2 + \sum_{i=1}^n \varphi(x_i), \tag{4.2}$$

where $\varphi : R_+ \to R$ is a potential function, e.g., [23], which includes (1.1) as a special case. Table 4.6 lists some well-used potential functions (left) and their extensions (right).

The numerical results with different potential functions and $\alpha = 0.1699$ are listed in Table 4.7. We observe that choosing $p \leq 0.5$ seems good for this example since using $p \leq 0.5$ can find three main factors with smaller prediction errors than $p > 0.5$.

Table 4.6
Potential functions, where $\alpha \in (0,1)$ is a parameter.

	Convex	Non-Lipschitz
f_1	$\varphi(t) = \|t\|$	$\varphi(t) = \|t\|^p$
	Nonconvex	Non-Lipschitz
f_2	$\varphi(t) = \|t\|^p$	$\varphi(t) = (\|t\|^p)^\alpha$
f_3	$\varphi(t) = \frac{\alpha\|t\|}{1+\alpha\|t\|}$	$\varphi(t) = \frac{\alpha\|t\|^p}{1+\alpha\|t\|^p}$
f_4	$\varphi(t) = \log(\alpha\|t\|+1)$	$\varphi(t) = \log(\alpha\|t\|^p+1)$

Table 4.7
Comparisons of different p with different potential functions.

p	(L, *Number of nonzero, Prediction error*)			
	f_1	f_2	f_3	f_4
0.9	(0.0001, 4, 0.4754)	(0.011, 4, 0.473)	(2.500, 4, 0.475)	(2.040, 4, 0.474)
0.8	(0.0015, 4, 0.4740)	(0.013, 4, 0.468)	(1.990, 4, 0.474)	(1.851, 4, 0.474)
0.7	(0.0050, 4, 0.4741)	(0.012, 4, 0.465)	(1.755, 4, 0.474)	(1.550, 4, 0.474)
0.6	(0.0084, 4, 0.4661)	(0.015, 3, 0.446)	(1.545, 4, 0.475)	(1.344, 4, 0.475)
0.5	(0.0119, 3, 0.4419)	(0.016, 3, 0.445)	(1.420, 3, 0.477)	(1.200, 3, 0.483)
0.4	(0.0148, 3, 0.4456)	(0.014, 3, 0.445)	(1.480, 3, 0.477)	(1.114, 3, 0.484)
0.3	(0.0176, 3, 0.4429)	(0.012, 3, 0.443)	(1.590, 3, 0.484)	(1.190, 3, 0.483)
0.2	(0.0196, 3, 0.4359)	(0.018, 3, 0.443)	(1.955, 3, 0.483)	(1.240, 3, 0.482)

5. Final remarks. Using the first and second order necessary condition for a local minimizer, we establish lower bounds for nonzero entries in any local optimal solution of a minimization model where the objective function is the sum of a data-fitting term in the ℓ_2 norm and a regularization term in the ℓ_p norm $(0 < p < 1)$. This establishes a theoretical justification by "zeroing" those entries in an approximate solution whose values are small enough and provides an explanation of why the model generates more sparse solutions when the norm parameter $p < 1$.

Moreover, the lower bounds clearly show the relationship between the sparsity of the solution and the choice of the regularization parameter and norm. These provide a systematic mechanism for selecting the model parameters, such as regularization weight λ and norm p. Based on these results, we propose a hybrid OMP-SG method for the nonconvex, non-Lipschitz continuous ℓ_2-ℓ_p minimization problem. Numerical results show that using the OMP-SG method to solve the ℓ_2-ℓ_p minimization problem (1.1) can provide more sparse solutions with smaller predictor errors compared with several well-known approaches for variable selection.

Acknowledgments. We would like to thank Professor Misha Kilmer and two referees for their valuable comments and suggestions for improving the presentation of this paper. We also would like to thank Rick Chartrand, the late Paul Tseng, and Weijun Zhou for their helpful comments.

REFERENCES

[1] C. Blake and C. Merz, *Repository of Machine Learning Databases [DB/OL]*, Department of Information and Computer Science, University of California, Irvine, CA, 1998, http://www.ics.uci.edu/~mlearn/MLRepository.html.

[2] L. Breiman, *Better subset regression using the nonnegative garrote*, Technometrics, 37 (1995), pp. 373–384.

[3] A. M. Bruckstein, D. L. Donoho, and M. Elad, *From sparse solutions of systems of equations to sparse modeling of signals and images*, SIAM Rev., 51 (2009), pp. 34–81.
[4] J. V. Burke, A. S. Lewis, and M. L. Overton, *A robust gradient sampling algorithm for nonsmooth, nonconvex optimization*, SIAM J. Optim., 15 (2005), pp. 751–779.
[5] E. J. Candes, M. B. Wakin, and S. P. Boyd, *Enhancing sparsity by reweighted ℓ_1 minimization*, J. Fourier Anal. Appl., 14 (2008), pp. 877–905.
[6] E. Candes, J. Romberg, and T. Tao, *Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information*, IEEE Trans. Inform. Theory, 52 (2006), pp. 489–509.
[7] R. Chartrand, *Exact reconstructions of sparse signals via nonconvex minimization*, IEEE Signal Process. Lett., 14 (2007), pp. 707–710.
[8] R. Chartrand, *Nonconvex regularization for shape preservation*, in 2007 IEEE International Conference on Image Processing, San Antonio, TX, 2007, pp. I293–I296.
[9] R. Chartrand and W. Yin, *Iteratively reweighted algorithms for compressive sensing*, in 2008 IEEE International Conference on Acoustics, Speech, and Signal Processing, Las Vegas, NV, 2008, pp. 3869–3872.
[10] S. S. Chen, D. L. Donoho, and M. A. Saunders, *Atomic decomposition by basis pursuit*, SIAM J. Sci. Comput., 43 (2001), pp. 129–159.
[11] X. Chen and Y. Ye, *On homotopy-smoothing methods for box-constrained variational inequalities*, SIAM J. Control Optim., 37 (1999), pp. 589–616.
[12] X. Chen and W. Zhou, *Smoothing nonlinear conjugate gradient method for image restoration using nonsmooth nonconvex minimization*, SIAM J. Imaging Sci., 3 (2010), pp. 765–790.
[13] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Classics Appl. Math. 5, SIAM, Philadelphia, 1990.
[14] G. Davis, S. Mallat, and M. Avellaneda, *Adaptive greedy approximations*, Const. Approx., 13 (1997), pp. 57–98.
[15] D. L. Donoho, *Compressed sensing*, IEEE Trans. Inform. Theory, 52 (2006), pp. 1289–1306.
[16] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, *Least angle regreeion*, Ann. Statist., 23 (2004), pp. 407–499.
[17] J. Fan and R. Li, *Variable selection via nonconcave penalized likelihood and its oracle properties*, J. Amer. Statist. Assoc., 96 (2001), pp. 1348–1360.
[18] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: Data mining*, in Inference and Prediction, Springer-Verlag, New York, 2001.
[19] A. E. Hoerl and R. Kennard, *Ridge regression: Biased estimation for nonorthogonal problems*, Technometrics, 12 (1970), pp. 55–67.
[20] K. C. Kiwiel, *Convergence of the gradient sampling algorithm for nonsmooth nonconvex optimization*, SIAM J. Optim., 18 (2007), pp. 379–388.
[21] S. Mallat and Z. Zhang, *Matching pursuits with time-frequency dictionaries*, IEEE Trans. Signal Process., 41 (1993), pp. 3397–3415.
[22] M. Nikolova, *Analysis of the recovery of edges in images and signals by minimizing nonconvex regularized least-squares*, Multiscale Model. Simul., 4 (2005), pp. 960–991.
[23] M. Nikolova, M. K. Ng, S. Zhang, and W. Ching, *Efficient reconstruction of piecewise constant images using nonsmooth nonconvex minimization*, SIAM J. Imaging Sci., 1 (2008), pp. 2–25.
[24] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, *Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition*, in Proceedings of the 27th Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, IEEE, 1993, pp. 40–44.
[25] R. Tibshirani, *Regression shrinkage and selection via the Lasso*, J. R. Stat. Soc. Ser. B Stat. Methodol., 58 (1996), pp. 267–288.
[26] J. A. Tropp and A. C. Gilbert, *Signal recovery from random measurements via orthogonal matching pursuit*, IEEE Trans. Inform. Theory, 53 (2007), pp. 4655–4667.
[27] J. A. Tropp, A. C. Gilbert, S. Muthukrishnan, and M. J. Strauss, *Improved sparse approximation over quasi-incoherent dictionaries*, in 2003 International Conference on Image Processing, Barcelona, Spain, 2003, pp. 37–40.
[28] Z. Xu, H. Zhang, Y. Wang, and X. Chang, *$L_{\frac{1}{2}}$ regularizer*, Sci. China Ser. F, 52 (2009), pp. 1–9.
[29] C. Zhang and X. Chen, *Smoothing projected gradient method and its application to stochastic linear complementarity problems*, SIAM J. Optim., 20 (2009), pp. 627–649.