Matching Pursuit LASSO Part I: Sparse Recovery Over Big Dictionary

Mingkui Tan, Ivor W. Tsang, and Li Wang

***Abstract*—Large-scale sparse recovery (SR) by solving ℓ_1-norm relaxations over *Big Dictionary* is a very challenging task. Plenty of greedy methods have therefore been proposed to address big SR problems, but most of them require restricted conditions for the convergence. Moreover, it is non-trivial for them to incorporate the ℓ_1-norm regularization that is required for robust signal recovery. We address these issues in this paper by proposing a Matching Pursuit LASSO (MPL) algorithm, based on a novel quadratically constrained linear program (QCLP) formulation, which has several advantages over existing methods. Firstly, it is guaranteed to converge to a global solution. Secondly, it greatly reduces the computation cost of the ℓ_1-norm methods over *Big Dictionaries*. Lastly, the exact sparse recovery condition of MPL is also investigated.**

***Index Terms*—Sparse recovery, compressive sensing, LASSO, matching pursuit, big dictionary, convex programming.**

I. Introduction

Sparse recovery (SR), also known as sparse representation or sparse reconstruction, has been widely required in many applications, such as signal processing, data mining and machine learning [1]–[4]. Sparse recovery is a fundamental element of the recently developed compressive sensing theory [1], [5]–[7]. Sparse recovery has also been widely applied to many image processing tasks, such as image restoration [8], [9], image super-resolution [10], and so on. In the machine learning area, SR has been successfully applied in robust face recognition [2], human gait recognition [11], subspace clustering [12], [13], dictionary learning [14], [15] and feature learning [16], [17].

Mathematically, given the observation $\mathbf{b} \in \mathbb{R}^n$ of an unknown k-sparse signal $\mathbf{x} \in \mathbb{R}^m$ from an underdetermined linear measurement system $\mathbf{b} = \mathbf{A}\mathbf{x} + \mathbf{e}$, SR seeks to recover $\mathbf{x}$ from $\mathbf{b}$ by solving an ℓ_0-norm minimization problem:

$$\min_{\mathbf{x}} \|\mathbf{x}\|_0, \text{ s.t. } \mathbf{b} = \mathbf{A}\mathbf{x}, \tag{1}$$

where $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_m] \in \mathbb{R}^{n\times m} (n \ll m)$ denotes the measurement matrix or dictionary with m atoms and $\mathbf{a}_j \in \mathbb{R}^n$ denotes the jth atom, $\mathbf{e} \in \mathbb{R}^n$ denotes the additive noise, and $\|\cdot\|_0$ denotes the ℓ_0-norm of a vector. Problem (1) in general is NP-complete [18], [5], [19]. Therefore, practitioners usually seek to solve its ℓ_1-norm convex relaxations [20]–[22]:

$$\min_{\mathbf{x}} \|\mathbf{x}\|_1, \text{ s.t. } \mathbf{b} - \mathbf{A}\mathbf{x} \in \mathcal{B}^p, \tag{2}$$

where the set $\mathcal{B}^p$ is determined by the noise structure with $p \in \{0, 2, \infty\}$ [23]. Here, $\mathcal{B}^0 = \{\mathbf{0}\}$ represents the noiseless constraint set, $\mathcal{B}^2 = \{\boldsymbol{\xi} : \|\boldsymbol{\xi}\| \leq \varepsilon\}$ denotes the ℓ_2-norm noise constraint set, and the set $\mathcal{B}^\infty = \{\boldsymbol{\xi} : \|\mathbf{A}^\top\boldsymbol{\xi}\|_\infty \leq \lambda\}$ corresponds to the Dantzig selector [24]. Studies in [5], [25] have shown that a k-sparse signal $\mathbf{x}$ can be exactly recovered by solving (2) provided that $\mathbf{A}$ satisfies the *restricted isometry property* (RIP) conditions: For a dictionary $\mathbf{A}$ and an integer $k \in [1, m]$, we define the *k-restricted isometry constant* $\sigma_k \in [0, 1)$ to be the smallest number such that

$$(1 - \sigma_k)\|\mathbf{x}\|^2 \leq \|\mathbf{A}\mathbf{x}\|^2 \leq (1 + \sigma_k)\|\mathbf{x}\|^2 \tag{3}$$

for all $\mathbf{x}$ with $\|\mathbf{x}\|_0 \leq k$ [5]. The RIP recovery conditions of ℓ_1-norm relaxations have been thoroughly studied by many researchers [1], [5], [25], [23]. For instance, studies in [23] show that if σ_k satisfies $\sigma_k < 0.307$, the solution to (2) can recover the k-sparse signals exactly.

A. Existing ℓ_1-Norm Methods for Sparse Recovery

During the last decade, many efforts have been made to efficiently solve ℓ_1-norm convex relaxed problems. Basically, the alternating direction method (ADM) or augmented Lagrange multiplier (ALM) method is often used to solve problem (2) with $\mathcal{B}^2$ noise constraint set [31]. In practice, the following LASSO problem is more widely studied in many applications [26], [21], [27]–[32]:

$$\min_{\mathbf{x}} \lambda\|\mathbf{x}\|_1 + \frac{1}{2}\|\mathbf{b} - \mathbf{A}\mathbf{x}\|^2, \tag{4}$$

where λ denotes the regularization parameter.

Many algorithms have been proposed to address the nonsmooth LASSO problem. An interior-point method is proposed in [33], and a gradient projection (GPSR) method was subsequently proposed by transforming LASSO into a quadratic programming problem [27]. A proximal gradient (PG) method is proposed to solve (4) based on a shrinkage-threshold operator [34]. To speed up the PG, a fast iterative shrinkage-threshold algorithm (FISTA) (also known as the accelerated proximal gradient method) is proposed in [29]. Recently, to tackle large-scale

Manuscript received October 22, 2013; revised February 19, 2014, June 25, 2014, and October 11, 2014; accepted November 28, 2014. Date of publication December 22, 2014; date of current version January 08, 2015. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Rui Zhang. This work was supported in part by the Australian Research Council Future Fellowship FT130100746, and Australian Research Council Grants DE120101161 and DP140102270.

M. Tan is with the School of Computer Science, University of Adelaide, Adelaide, SA 5005, Australia (e-mail: mingkui.tan@adelaide.edu.au).

I. W. Tsang is with the Centre for Quantum Computation and Intelligent Systems (QCIS), University of Technology, Sydney (UTS), Ultimo, NSW 2007, Australia (e-mail: Ivor.Tsang@uts.edu.au).

L. Wang is with the Institute for Computational and Experimental Research in Mathematics (ICERM), Brown University, Providence, RI 02912 USA (e-mail: liwangucsd@gmail.com).

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

Digital Object Identifier 10.1109/TSP.2014.2385036

SR problems, coordinate descent methods have been well exploited [35], [36]. For instance, an elegantly designed parallel coordinate descent method, referred to as *Shotgun*, has been developed to improve the efficiency through parallel computation [37]. Another recent work, S-L1, uses a screening test to predict the atoms with zero weights and adopts random projections to reduce the computational cost [38]. Homotopy algorithms that try to find a full path of solutions have been well studied for solving the LASSO problem [21], [30], [32]. For example, in [21], a solution-path based algorithm, which is referred to as *least-angle regression* (LARS), is proposed. More recently, a novel proximal gradient homotopy (PGH) method has been proposed in [32], [39]. By gradually decreasing the regularization parameter from an initial guess λ_0, homotopy methods solves a series of strongly convex subproblems, thus a geometric convergence rate can be maintained the optimization of each subproblem [32]. PGH has shown state-of-the-art performance compared to the other methods above. Recently, active-set methods have also been proposed to solve (4) [40]–[43]. Active-set methods iteratively include one active variable that violates the Lagrangian duality condition into an active set [40], and then solve a subproblem w.r.t. active variables only.

The major computational burden of the above methods occurs as a result of repetitive calculations of $\mathbf{Ax}$ and $\mathbf{A}^\top\boldsymbol{\xi}$, where $\boldsymbol{\xi} = \mathbf{b} - \mathbf{Ax} \in \mathbb{R}^n$ [27], [32]. Since both $\mathbf{Ax}$ and $\mathbf{A}^\top\boldsymbol{\xi}$ take $O(mn)$ time, solving the ℓ_1-norm relaxations over big dictionary is very challenging [2], [44], [45]. For example, for active-set methods, at least k times are needed to compute $\mathbf{A}^\top\boldsymbol{\xi}$ [43]. When k is large, solving large-scale problems would be very expensive.

The ℓ_1-norm regularization may also suffer from the solution bias issue [27]. For LASSO, the sparsity of the optimal solution is determined by the regularization λ [21], [34]. In general, a large λ is required to induce more sparse solutions [34],[1] which may, however, incur solution bias [27]. In particular, if $\lambda \geq \|\mathbf{A}^\top\mathbf{b}\|_\infty$, none of the atoms will be selected, i.e., $\mathbf{x}^* = \mathbf{0}$ [33]. To reduce the solution bias, one may prefer to set a small value to λ, but the solution may be no longer sparse. In summary, the sparsity and unbiased solution cannot be achieved simultaneously via ℓ_1-norm regularization.

B. Existing Greedy Methods for Sparse Recovery

In contrast to ℓ_1-norm methods, orthogonal matching pursuit (OMP) is widely applied to solve (1) [47]–[54]. However, it is computationally expensive when k is large. To improve the scalability, many greedy methods with faster convergence speed have been proposed, such as the compressive sampling matching pursuit (CoSaMP) [55], subspace pursuit (SP) [56], iterative hard thresholding (IHT) [57], accelerated iterative hard thresholding (AIHT) [58], [59], orthogonal matching pursuit with replacement (OMPR) [60] and so on. These methods seek to recover a k-sparse signal using greedy iterative procedures, and rely on the knowledge of the ground-truth sparsity k. In general, much fewer calculations of $\mathbf{A}^\top\boldsymbol{\xi}$ are needed in these methods. As a result, they are more attractive when solving large-scale SR problems [60]. The convergence and sparse recovery conditions of these algorithms in terms of the RIP constant have been well studied [55], [56], [60]. For example, CoSaMP, SP, AIHT and OMPR can recover any k-sparse signal provided that $\sigma_{4k} < 0.35, \sigma_{3k} < 0.35, \sigma_{3k} < 1/\sqrt{32}$ and $\sigma_{2k} < 0.499$, respectively [59], [60]. However, there are two major disadvantages of these methods.

Firstly, since the restricted recovery conditions are also their corresponding convergence assertions, the convergence of these methods might not be guaranteed if the restricted conditions are violated. Indeed the RIP conditions might not be satisfied in real-world applications [61]–[64]. For example, in sparse recovery based face recognition [2], [31], images from the same person might be so highly correlated that the RIP condition may not be satisfied, and these methods **may not converge**.[2]

Secondly, for many greedy methods, their performance depends on a proper estimation of the unknown ground-truth sparsity k, denoted by $\widehat{k}$. If $\widehat{k}$ is smaller than k, none of these methods can recover the k-sparse signal. On the contrary, if $\widehat{k}$ is too large, the efficiency and recovery performance of these methods may degrade. Nevertheless, in practice, it is usually non-trivial to determine $\widehat{k}$, which restricts the application of these methods.

Essentially, the first issue is caused by atom replacement involved in the algorithms. For a highly coherent dictionary, the optimal k-sparse solution may not be unique [62], [61], and the active atom set may change frequently due to the atom replacement. As a result, these methods may not converge [55], [56], [60]. By incrementally including a set of new atoms, some greedy variants avoid the atom replacement, such as the regularized OMP (ROMP) [65], s-OMP [66], [67], stagewise weak gradient pursuits (or SWCGP for short) [68]and stagewise OMP (StOMP) [69]. In s-OMP, s denotes the number of atoms selected per iteration. For s-OMP, a tight RIP condition has been investigated in [67], i.e., $\sigma_{sk} < \frac{\sqrt{s}}{(2+\sqrt{2})\sqrt{k}}$. However, this condition will become very restricted when k is large. For StOMP and SWCGP, the number of newly added atoms is determined by a thresholding scheme.

Recently, several works have shown that linear dependencies (or high coherence) in a dictionary are permitted and beneficial [61]–[64], [70]. In [64], the redundancy issue is addressed by an extension of OMP, i.e., ϵ-OMP algorithm. Moreover, some other works extend classical greedy methods, such as CoSaMP, SP and IHT, to handle coherent dictionaries [61]–[63]. Lastly, a new family of pursuit algorithms have been proposed for the cosparse analysis model that is an interesting alternative to the standard SR [71]. However, in these methods, the estimation of k is still required.

C. Our Contributions

An efficient Matching Pursuit LASSO (MPL) algorithm is developed to solve large-scale SR problems based on a quadratically constrained linear program (QCLP) reformulation of LASSO. The core contributions of this paper are summarized as follows:

[1] The selection of the regularization parameter λ (i.e., the model selection problem) depends on the structure of the noises [46], which is beyond the scope of this paper. In [32], $\lambda \geq \eta\|\mathbf{A}^\top\widehat{\mathbf{e}}\|_\infty$ is suggested, where $\widehat{\mathbf{e}}$ denotes the additive noises and $\eta \geq 8$. Since $\widehat{\mathbf{e}}$ is usually unknown, in [27], $\lambda \leq 0.1\|\mathbf{A}^\top\mathbf{b}\|_\infty$ is tested.

[2] An illustration can be found in Section V.A.

- The convergence of MPL on non-RIP problems is justified. Moreover, we show that, under mild RIP conditions, the objective value decreases linearly in MPL.
- Unlike most existing greedy methods that rely on a good estimation of the ground-truth sparsity k, MPL only needs to estimate a conservative parameter ϱ. Compared to k, it is much easier and more flexible to adjust ϱ.
- MPL is guaranteed to recover any k-sparse signals under the RIP condition $\sigma_k \leq 0.307 - \vartheta$ for a small positive ϑ.
- MPL includes OMP and active-set methods as special cases, but significantly reduces their computational cost. We also show that StOMP [69] and SWCGP [68] can be considered as special algorithms that solve a variant of the proposed QCLP problem.

The rest of this paper is organized as follows. In Section II, we present the QCLP formulation for LASSO. We then present the MPL algorithm in Section III. The convergence and sparse recovery condition of MPL are detailed in IV. Some numerical studies are given in Section V. We conclude this work in Section VI.

II. A QCLP Formulation for Sparse Recovery

A. Notations, Definitions, and Preliminaries

Throughout the paper, we denote the transpose of a vector/matrix by the superscript $^\top$, $\mathbf{0}$ as a zero vector and $\text{diag}(\mathbf{v})$ as a diagonal matrix with diagonal entries equal to $\mathbf{v}$. In addition, $||\mathbf{v}||_p$ and $||\mathbf{v}||$ denote the ℓ_p-norm and ℓ_2-norm, respectively. For a function $f(\mathbf{x})$, the gradient and subgradient of $f(\mathbf{x})$ at $\mathbf{x}$ are denoted by $\nabla f(\mathbf{x})$ and $\partial f(\mathbf{x})$, respectively. For a sparse vector $\mathbf{x}$, let the calligraphic letter $\mathcal{T} = \text{support}(\mathbf{x}) = \{i|x_i \neq 0\} \in \{1, \ldots, m\}$ be its support, $\mathbf{x}_\mathcal{T}$ be the subvector indexed by $\mathcal{T}$, and $\mathcal{T}^c$ be the complementary set of $\mathcal{T}$, i.e., $\mathcal{T}^c = \{1, \ldots, m\} \backslash \mathcal{T}$. Given an index set $\mathcal{I} \subseteq \{1, \ldots, m\}$, let $\mathbf{A}_\mathcal{I}$ denote the columns of $\mathbf{A}$ regarding $\mathcal{I}$. Furthermore, let $\mathbf{A} \odot \mathbf{B}$ represent the element-wise product of two matrices $\mathbf{A}$ and $\mathbf{B}$. Lastly, we define the *restricted eigenvalue (RE) condition, restricted condition number* and *restricted set* as follows.

Definition 1: [28], [52], [4] Given an integer $k > 0$, a matrix $\mathbf{A}$ is said to satisfy the RE Condition at sparsity level k, if there exist positive constants $\gamma_-(\mathbf{A}, k)$ and $\gamma_+(\mathbf{A}, k)$ such that

$$\gamma_-(\mathbf{A}, k) = \inf \left\{ \frac{\mathbf{x}^\top \mathbf{A}^\top \mathbf{A} \mathbf{x}}{\mathbf{x}^\top \mathbf{x}}, \mathbf{x} \neq \mathbf{0}, ||\mathbf{x}||_0 \leq k \right\}, \quad (5)$$

$$\gamma_+(\mathbf{A}, k) = \sup \left\{ \frac{\mathbf{x}^\top \mathbf{A}^\top \mathbf{A} \mathbf{x}}{\mathbf{x}^\top \mathbf{x}}, \mathbf{x} \neq \mathbf{0}, ||\mathbf{x}||_0 \leq k \right\}. \quad (6)$$

In addition, the restricted condition number is defined as

$$\kappa(\mathbf{A}, k) = \frac{\gamma_+(\mathbf{A}, k)}{\gamma_-(\mathbf{A}, k)}. \quad (7)$$

There are many types of matrices that satisfy the RE condition [72]. Actually, the RE condition is less restrictive than the RIP condition [28]. For a matrix that satisfies RIP condition, we immediately have $\gamma_-(\mathbf{A}, k) \geq 1 - \sigma_k$ and $\gamma_+(\mathbf{A}, k) \leq 1 + \sigma_k$ at the sparsity level k.

Definition 2: Restricted Set [73]: Given an index set $\mathcal{T}$ and a positive number D, the restricted set is defined as $\Gamma_D = \{\mathbf{h} \in \mathbb{R}^m : ||\mathbf{h}_{\mathcal{T}^c}||_1 \leq D||\mathbf{h}_\mathcal{T}||_1\}$, where $\mathcal{T}^c$ denotes the complementary set of $\mathcal{T}$.

The property of the restricted set is also known as the restricted nullspace property, which is very important in analyzing ℓ_1-norm methods [72].

B. A QCLP Reformulation for LASSO

A very high dimensional sparse signal over a big dictionary can easily be recovered by solving a small-scale optimization problem w.r.t. the detected support atoms, if we can detect its supports [60]. Motivated by this computational advantage, we introduce a **support selection** vector $\boldsymbol{\tau} \in \{0, 1\}^m$ to detect the active atoms by $(\boldsymbol{\tau} \odot \mathbf{x})$, where the ith atom will be selected if and only if $\tau_i = 1$. However, there are two remaining difficulties: Firstly, $\boldsymbol{\tau}$ does not necessarily induce sparse solutions; Secondly, in general, the target sparsity k is unknown in advance. To address these difficulties, we impose a sparsity constraint $||\boldsymbol{\tau}||_0 \leq \varrho$. Here, ϱ is assumed to be several times smaller than k, and we do not have to estimate it accurately. Rather than estimating k directly, we estimate ϱ instead, which would be much easier than setting k.

With the introduction of $\boldsymbol{\tau}$, we present a new model for SR. For simplicity, let $\Lambda = \{\boldsymbol{\tau} : ||\boldsymbol{\tau}||_0 \leq \varrho, \boldsymbol{\tau} \in \{0, 1\}^m\}$ be the domain of $\boldsymbol{\tau}$. Let $\boldsymbol{\xi} = \mathbf{b} - \mathbf{A}(\boldsymbol{\tau} \odot \mathbf{x})$ be the regression error, we consider solving an alternative model to formulation (4):

$$M1: \quad \min_{\boldsymbol{\tau} \in \Lambda} \min_{\mathbf{x}, \boldsymbol{\xi}} \ \lambda||\mathbf{x}||_1 + \frac{1}{2}||\boldsymbol{\xi}||^2 \quad \text{s.t.} \quad \boldsymbol{\xi} = \mathbf{b} - \mathbf{A}(\mathbf{x} \odot \boldsymbol{\tau}). \quad (8)$$

In problem (8), even though we have explicitly imposed the sparsity constraint $||\boldsymbol{\tau}||_0 \leq \varrho$, the sparsity-induced regularization (i.e., the ℓ_1-norm regularization) is important for the robust signal recovery, specifically for the protection from noise [74]. Equally, although LASSO itself can induce sparse solutions, the explicit sparsity constraint $||\boldsymbol{\tau}||_0 \leq \varrho$ is still necessary and important for the purpose of de-biasing with a small λ [27]. Note that a small λ is preferable for de-biasing, though it may incur non-sparse solutions. In the extreme case, if $\lambda = 0$, we achieve a special case of $M1$:

$$M2: \quad \min_{\boldsymbol{\tau} \in \Lambda} \min_{\mathbf{x}, \boldsymbol{\xi}} \ \frac{1}{2}||\boldsymbol{\xi}||^2 \quad \text{s.t.} \quad \boldsymbol{\xi} = \mathbf{b} - \mathbf{A}(\mathbf{x} \odot \boldsymbol{\tau}). \quad (9)$$

In problems (8) and (9), there are $\sum_{i=0}^{\varrho} \binom{m}{i}$ feasible $\boldsymbol{\tau}$'s in Λ, and the task of the optimization is to find the **BEST** $\boldsymbol{\tau}$ from the feasible set that minimizes the regression loss $||\boldsymbol{\xi}||^2$. The optimal solutions to (8) and (9) may not be unique. In fact, since $\varrho < k$, two different $\boldsymbol{\tau}$'s may exist that produce the same objective values. Nevertheless, due to the many $\boldsymbol{\tau}$'s, solving the two problems is very difficult. To make them tractable, we make the following transformations.

Proposition 1: By introducing dual variables $\boldsymbol{\alpha} \in \mathbb{R}^n$ to the constraint $\boldsymbol{\xi} = \mathbf{b} - \mathbf{A}(\mathbf{x} \odot \boldsymbol{\tau})$ regarding the inner minimization problem, problem (8) can be reformulated as

$$\min_{\boldsymbol{\tau} \in \Lambda} \max_{\boldsymbol{\alpha} \in \mathbb{R}^n} \ -\frac{1}{2}||\boldsymbol{\alpha}||^2 + \boldsymbol{\alpha}^\top \mathbf{b} \quad \text{s.t.} \quad ||\boldsymbol{\alpha}^\top \mathbf{A} \text{diag}(\boldsymbol{\tau})||_\infty \leq \lambda. \quad (10)$$

Let $\mathbf{x}^*$ be the optimal solution to (8), then $x_j^* = 0$ if $\tau_j = 0$.

The proof can be found in Appendix A.

Due to the constraint $\|\boldsymbol{\alpha}^\top \mathbf{A}\mathrm{diag}(\boldsymbol{\tau})\|_\infty \leq \lambda, \boldsymbol{\alpha}$ is bounded. Without loss of generality, we define a compact set $\mathcal{A} = [-l, l]^n$ as the domain of $\boldsymbol{\alpha}$, where $l > 0$ is a large number such that the optimal solution $\boldsymbol{\alpha}^*$ always exists in $\mathcal{A}$. For convenience, we define a feasible compact domain for $\boldsymbol{\alpha}$ w.r.t. each $\boldsymbol{\tau}$ as $\mathcal{A}_{\boldsymbol{\tau}}^{\lambda} = \{\boldsymbol{\alpha} : \|\boldsymbol{\alpha}^\top \mathbf{A}\mathrm{diag}(\boldsymbol{\tau})\|_\infty \leq \lambda, \boldsymbol{\alpha} \in [-l, l]^n\}$ and

$$f(\boldsymbol{\alpha}, \boldsymbol{\tau}) = \frac{1}{2}\|\boldsymbol{\alpha}\|^2 - \boldsymbol{\alpha}^\top \mathbf{b}, \boldsymbol{\alpha} \in \mathcal{A}_{\boldsymbol{\tau}}^{\lambda}. \tag{11}$$

Recall that $\mathcal{A}$ is convex and compact. According to the minimax inequality (5.46) in [75], we have

$$\min_{\boldsymbol{\tau}\in\Lambda} \max_{\boldsymbol{\alpha}\in\mathcal{A}} -f(\boldsymbol{\alpha}, \boldsymbol{\tau}) \geq \max_{\boldsymbol{\alpha}\in\mathcal{A}} \min_{\boldsymbol{\tau}\in\Lambda} -f(\boldsymbol{\alpha}, \boldsymbol{\tau}). \tag{12}$$

According to (12), problem $\max_{\boldsymbol{\alpha}\in\mathcal{A}} \min_{\boldsymbol{\tau}\in\Lambda} -f(\boldsymbol{\alpha}, \boldsymbol{\tau})$ is a lower bound to the non-convex problem in (10). Moreover, it is a **convex relaxation** to (10). To see the convexity, we can formulate it as a QCLP problem by introducing a new variable $\theta \in \mathbb{R}$ [76].

$$\min_{\boldsymbol{\alpha},\theta} \theta, \quad \text{s.t.} \quad f(\boldsymbol{\alpha}, \boldsymbol{\tau}) \leq \theta, \quad \forall \boldsymbol{\tau} \in \Lambda. \tag{13}$$

For each $\boldsymbol{\tau} \in \Lambda$, the inequality $f(\boldsymbol{\alpha}, \boldsymbol{\tau}) \leq \theta$ in (13) defines a quadratic constraint w.r.t. $\boldsymbol{\alpha}$. Since Λ contains $T = \sum_{i=0}^{\varrho} \binom{m}{i}$ elements, there are T constraints involved in problem (13), making it intractable even for medium size m and ϱ.

III. Matching Pursuit for LASSO

Problem (13) can be considered as a special case of the semi-infinite programming (SIP) problem that has infinite number of constraints [77]. To solve the SIP problem, a central cutting plane (CCP) algorithm has been developed [78], [77], and has been shown to have great computational advantages when solving problems of many constraints. Therefore, in this paper, we seek to address problem (13) by adapting the CCP algorithm in [77], which is presented in Algorithm 1.

Algorithm 1 Central cutting plane method for solving (13).

```
1: Initialize $\boldsymbol{\alpha}^0 = \mathbf{b}$, and find the most-violated $\boldsymbol{\tau}_0$.
2: Let $z^0 = -f(\boldsymbol{\alpha}^0, \boldsymbol{\tau}_0), \Lambda_0 = \{\boldsymbol{\tau}_0\}$ and $t = 1$.
3: Solve the following master problem:

   $\max_{\boldsymbol{\alpha},\theta,\delta} \delta, \text{s.t.} \theta + \delta \leq z^{t-1}, f(\boldsymbol{\alpha}, \boldsymbol{\tau}_i) \leq \theta - \delta, \forall \boldsymbol{\tau}_i \in \Lambda_{t-1}.$   (14)

4: Let $(\boldsymbol{\alpha}^t, \theta^t, \delta^t)$ be the solution of (14). If the stopping
   condition is achieved, stop.
5: Find a new most-violated constraint $\boldsymbol{\tau}_t$.
6: Set $\Lambda_t = \Lambda_{t-1} \cup \{\boldsymbol{\tau}_t\}$. If $f(\boldsymbol{\alpha}^t, \boldsymbol{\tau}_t) > \theta^t$, that is, $\boldsymbol{\alpha}^t$ is
   an infeasible solution to (13). Let $z^t = \min(z^{t-1}, \theta^t)$.
7: Let $t = t + 1$ and go to Step 3.
```

In Algorithm 1, instead of solving (13) with all constraints, we iteratively find the **most-violated constraint** and add it into an active constraint set Λ_t, and then solve a much reduced master problem in (14) with active constraints only. Finding the most-violated constraint is referred to as the **worst-case analysis**, and the initial active constraint set Λ_0 is set to be an empty set $\emptyset$. In problem (14), z^{t-1} is strictly greater than the objective value of (13). Moreover, the number of constraints in (14) monotonically increases as t increases, thus δ^t will monotonically decrease. In the following subsections, we will present the details of the worst-case analysis, master problem optimization, and stopping conditions.

A. Worst-Case Analysis

The worst-case analysis is to find the most active constraint w.r.t. $\boldsymbol{\tau}$ from a huge number of candidates. Let $\mathbf{g} = \mathbf{A}^\top \boldsymbol{\alpha}$, at the optima of (13), the following condition must hold:

$$\left\|\boldsymbol{\alpha}^\top \mathbf{A}\mathrm{diag}(\boldsymbol{\tau})\right\|_\infty = \|\mathbf{g} \odot \boldsymbol{\tau}\|_\infty \leq \lambda. \tag{15}$$

Apparently, any atom $\mathbf{a}_j$ with $|g_j| > \lambda$ will **violate** the above optimality condition, and the atoms with the largest $|g_j|$ violate the condition the most. Since $\|\boldsymbol{\tau}\|_0 \leq \varrho$, we can choose the ϱ atoms with the largest $|g_j|$ to construct the most active constraint. To obtain the most active $\boldsymbol{\tau}_t$, we can set the ϱ entries of $\boldsymbol{\tau}_t$ w.r.t. the largest $|g_j|$ to 1, and the rests to 0. Essentially, we only need to record the indices of the ϱ atoms into a set $\mathcal{J}_t$, i.e., $\mathcal{J}_t = \mathrm{support}(\boldsymbol{\tau}_t)$.

Let $\mathcal{I}_t$ record the indices of atoms selected up to the tth iteration, i.e., $\mathcal{I}_t = \cup_i \mathcal{J}_i, i = 1, \ldots, t$. In general, once an atom is added into $\mathcal{I}_t$, it is unlikely to be selected in the following steps. However, if we do not solve the master problem accurately, some of the selected atoms may have large value of $|g_j|$, thus they might be chosen again. To avoid this, we choose the atoms from $\{1, \ldots, m\} \backslash \mathcal{I}_t$ to form $\mathcal{J}_{t+1}$. In this way, there will be no overlapping element among $\mathcal{J}_i$'s, where $i = 1, \ldots, t$.

The worst-case analysis above is akin to the matching step in greedy methods [55], [60], [52]. Moreover, we will show that Algorithm 1 actually addresses the original LASSO problem (see Lemma 2). In this sense, hereafter we refer to Algorithm 1 as Matching Pursuit LASSO or MPL for short.

B. MPL in Primal

After updating Λ_t, we tend to solve (14), which is referred to as the master problem optimization. Let $\mathcal{J}_i$ be the index set of atoms selected by $\boldsymbol{\tau}_i$ and $T_t = |\Lambda_t|$ be the number of active constraints. Although the number of constraints is greatly reduced, it is still not easy to solve problem (14) w.r.t. $\boldsymbol{\alpha}$ directly, especially when n is very large. However, since problem (14) involves only a small set of active atoms, faster optimization might be achieved w.r.t. the primal variable $\mathbf{x}_{\mathcal{I}_t}$.

Proposition 2: Suppose there are no overlapping elements among $\mathcal{J}_i$'s and $\mathcal{I}_t = \cup_i \mathcal{J}_i$, problem (14) can be addressed by solving a LASSO problem w.r.t. atoms in $\mathcal{I}_t$:

$$\min_{\mathbf{x},\boldsymbol{\xi}} \lambda \left\|\mathbf{x}_{\mathcal{I}_t}\right\|_1 + \frac{1}{2}\|\boldsymbol{\xi}\|^2 : \text{ s.t. } \boldsymbol{\xi} = \mathbf{b} - \mathbf{A}\mathbf{x}, \mathbf{x}_{\mathcal{I}_t^c} = \mathbf{0}. \tag{16}$$

Furthermore, the optimal dual variable $\boldsymbol{\alpha}^*$ of problem (14) can be recovered by $\boldsymbol{\alpha}^* = \boldsymbol{\xi}^*$.

The proof can be found in Appendix B.

Based on Proposition 2, MPL can immediately be implemented in the primal form, and is depicted in Algorithm 2. Due to the relation $\boldsymbol{\alpha}^* = \boldsymbol{\xi}^*$, we can still conduct the worst-case analysis even if we address the master problem in primal form. Note that since no $\boldsymbol{\tau}$ involved at the initial stage, we initialize $\mathbf{x}^0 = \mathbf{0}$. Correspondingly, we have $\boldsymbol{\alpha}^* = \mathbf{b}$.

Algorithm 2 MPL in primal.

```
Initialize $\mathbf{x}^0 = \mathbf{0}, \boldsymbol{\alpha}^0 = \mathbf{b}, \mathcal{I}_0 = \emptyset$. Let $t = 1$.
while (The stopping condition is not achieved) do
    Conduct the worst-case analysis: Let $\mathbf{g} = \mathbf{A}^\top \boldsymbol{\alpha}^{t-1}$;
    choose the $\varrho$ largest $|g_j|$ (where $j \in \{1, \ldots, m\} \backslash \mathcal{I}_{t-1}$)
    and record their indices in $\mathcal{J}_t$; let $\mathcal{I}_t = \mathcal{I}_{t-1} \cup \mathcal{J}_t$.
    Initialize $\mathbf{u}^0_{\mathcal{I}_t} = \mathbf{x}^{t-1}_{\mathcal{I}_t}$ (warm-start) and $\mathbf{u}^0_{\mathcal{I}^c_t} = \mathbf{0}$.
    for $s = 1, \ldots,$ do
        Update $\mathbf{u}^s_{\mathcal{I}_t}$ using PG ($\lambda > 0$) or CGD ($\lambda = 0$) rules.
        Break if the stopping conditions are achieved.
    end for.
    Set $\mathbf{x}^t_{\mathcal{I}_t} = \mathbf{u}^k_{\mathcal{I}_t}, \mathbf{x}^t_{\mathcal{I}^c} = \mathbf{0}$ and $\boldsymbol{\alpha}^t = \boldsymbol{\xi}^t$. Let $t = t + 1$.
end while
```

MPL involves two layers of loops. The outer loop is corresponding to that in Algorithm 1, and the inner loop is w.r.t. the master problem optimization. To distinguish the two kinds of loops, we use t as the outer loop index and s as the **inner loop index**. Correspondingly, we denote the outer iteration variable by $\mathbf{x}^t$ and the inner iteration variable by $\mathbf{u}^s$. To accelerate the convergence, we initialize $\mathbf{u}^0_{\mathcal{I}_t} = \mathbf{x}^{t-1}_{\mathcal{I}_t}$ for **warm-start** for each inner loop.

Any existing ℓ_1-norm solver can be adopted to solve problem (16). When $\lambda > 0$, we use the PG algorithm [34]. When $\lambda = 0$, we apply the conjugate gradient descent (CGD) method. For convenience, let us define

$$\varphi(\mathbf{x}) = \frac{1}{2}\|\mathbf{b} - \mathbf{A}\mathbf{x}\|^2, \text{ and} \tag{17}$$

$$f(\mathbf{x}) = \lambda\|\mathbf{x}\|_1 + \varphi(\mathbf{x}). \tag{18}$$

Basically, PG iteratively minimizes the following local quadratic approximation of $f(\mathbf{x})$ at a fixed point $\mathbf{u}$:

$$\phi(\mathbf{x}, \mathbf{u}) = \lambda\|\mathbf{x}\|_1 + \varphi(\mathbf{u}) + \nabla\varphi(\mathbf{u})^\top(\mathbf{x} - \mathbf{u}) + \frac{L}{2}\|\mathbf{x} - \mathbf{u}\|^2,$$

where L is a positive number. A close solution for this problem exists which relies on a soft-thresholding operator: $S_{L,\lambda}(\mathbf{o})_i = \text{sign}(o_i)\max\{|o_i| - \frac{\lambda}{L}, 0\}$. Let $\mathbf{u}^s$ be the point at the sth iteration and $\mathbf{g} = \nabla\varphi(\mathbf{u}^s) = \mathbf{A}^\top(\mathbf{A}\mathbf{u}^s - \mathbf{b})$, the minimizer of $\phi(\mathbf{x}, \mathbf{u}^s)$ can be calculated by $S_{L,\lambda}(\mathbf{u}^s - \mathbf{g}/L)$ [34]. Therefore, the basic updating rule of PG is:

$$\mathbf{u}^{s+1} = S_{L,\lambda}(\mathbf{u}^s - \mathbf{g}/L) = \mathbf{u}^s - 1/LG(\mathbf{u}^s), \tag{19}$$

where $G(\mathbf{u}^s) = L(\mathbf{u}^k - S_{L,\lambda}(\mathbf{u}^s - \mathbf{g}/L))$ is referred to as the generalized gradient, and L is adjusted by a line search [34]. By applying the line search, we have the following bound w.r.t. the objective improvement of each outer iteration.

Lemma 1: Let $\mathbf{g} = \nabla\varphi(\mathbf{x}^t)$, $\mathbf{x}^{t+1}$ be the point at the $(t+1)$th iteration, and $\mathcal{J}_{t+1}$ be the index set obtained by the worst-case analysis, where $|g_i| > \lambda$ for $\forall i \in \mathcal{J}_{t+1}$. With proper line search, we have:

$$f(\mathbf{x}^t) - f(\mathbf{x}^{t+1}) \geq \frac{1}{2L}\sum_{i \in \mathcal{J}_{t+1}} (|g_i| - \lambda)^2.$$

The proof can be found in Appendix C.

When $\lambda = 0$, problem (16) is reduced to a least square regression problem. In this case, we can adopt the CGD method instead, where the step size can be cheaply calculated by an exact line search [79]. Note that, if $\gamma_-(\mathbf{A}, t\varrho) > 0$ (or $1 - \sigma_{t\varrho} > 0$), problem (16) is strongly convex w.r.t. $\mathbf{x}_{\mathcal{I}_t}$. The following theorem indicates that, both PG and CGD converge linearly under this condition.

Theorem 1: ([79], [34]) At the tth iteration of MPL, if $\gamma_-(\mathbf{A}, t\varrho) > 0$ or $1 - \sigma_{t\varrho} > 0$, then both PG and CGD converge linearly. Specifically, let $\eta_u > 1$ be the adjustment parameter in the line search of PG, $\{\mathbf{u}^s\}$ be the sequence generated by PG or CGD, $\mathbf{u}^*$ be the minimizer of (16) and $f(\mathbf{u}^*)$ be the optimal function value, then $\{\mathbf{u}^k\}$ satisfies

$$\frac{f(\mathbf{u}^s) - f(\mathbf{u}^*)}{f(\mathbf{u}^0) - f(\mathbf{u}^*)} \leq (\chi_t)^s, \tag{20}$$

where s is the inner iteration index, $\chi_t = (\sqrt{\kappa(\mathbf{A}, t\varrho)} - 1)/(\sqrt{\kappa(\mathbf{A}, t\varrho)} + 1)$ is for CGD and $\chi_t = (1 - \frac{1}{4\eta_u\kappa(\mathbf{A}, t\varrho)})$ is for PG.

Since the condition number $\kappa(\mathbf{A}, t\varrho)$ increases w.r.t. t, the factor χ_t increases w.r.t. t for both PG and CGD. Therefore, the **warm-start** in Algorithm 2 is particularly important to accelerate the convergence speed when t is large. Moreover, χ_t for CGD is smaller than that of PG, thus it is a better choice when $\lambda = 0$. If the inner loop stops after s iterations, we have $f(\mathbf{x}_t) = f(\mathbf{u}^s)$ and $f(\mathbf{x}_{t-1}) = f(\mathbf{u}^0)$, which implies

$$\frac{f(\mathbf{x}_t) - f(\mathbf{u}^*)}{f(\mathbf{x}_{t-1}) - f(\mathbf{u}^*)} \leq (\chi_t)^s. \tag{21}$$

However, this relation does not tell us how much the objective will decrease since $f(\mathbf{u}^*)$ is unknown.

In practice, it might be expensive to achieve an exact solution to the master problem. To avoid this, we can stop the inner layer loops once the master problem is sufficiently minimized. We use the following stopping criterion for the inner layer loop:

$$\frac{f(\mathbf{u}^{s-1}) - f(\mathbf{u}^s)}{f(\mathbf{u}^0) - f(\mathbf{u}^s)} \leq \varepsilon_{\text{in}}, \tag{22}$$

where ε_{in} denotes the tolerance. In Theorem 3, we will demonstrate that a solution satisfying this condition is sufficient with proper ε_{in}.

C. Parameter Setting of ϱ

In general, MPL with a larger ϱ needs fewer calculations $\mathbf{A}^\top\boldsymbol{\xi}$, which is essential for reducing the overall complexity. However, a large ϱ may incur the problem of adding non-support atoms and a large condition number of the master problem. To avoid these issues, ϱ should be sufficiently small. Although we can set $\varrho = 1$, it is computationally very expensive when k is large. In principle, if k is known, one can set $\varrho = k/r$, with $r \geq 5$ being suggested. When k is unknown, we propose the following two strategies.

Recall that SR might recover a k-sparse signal with $n = O(k \log(m))$ non-adaptive measurements for dictionaries that satisfy the RIP conditions [1], [5], [25]. Motivated by this observation, we can set ϱ for RIP dictionaries by

$$\varrho = \lceil n/(r\log(m))\rceil, \tag{23}$$

where $\lceil \cdot \rceil$ is the ceiling function, and $r \geq 5$ is suggested. This setting of ϱ is simple. However, it is not adaptive to general dictionaries.

Another method is motivated by the thresholding strategy used in SWCGP [68]. Let $\boldsymbol{\beta} = \mathbf{A}^\top \mathbf{b}$ and $\lambda_{\max} = ||\boldsymbol{\beta}||_\infty$, we count the number of atoms satisfying

$$\boldsymbol{\beta}_i \geq \eta \lambda_{\max}, \tag{24}$$

where $0 < \eta \leq 1$, and then let ϱ be this number. Obviously, if η is very small, ϱ can be very large. In practice, we suggest setting $\eta \geq 0.6$. If $\eta = 1$, we have $\varrho = 1$, and MPL is reduced to the active-set method. Note that once ϱ is set at the initialization stage in MPL, it will be fixed, which is different from that in SWCGP [68] and StOMP [69].

D. Early Stopping and Debiasing

Recall that, in practice, we may use a small λ to reduce the solution bias. For instance, one can set $\lambda = 0$, and then MPL is reduced to the s-OMP algorithm [67]. On the other hand, for any $\lambda \geq 0$, a natural stopping condition for MPL is

$$||\boldsymbol{\alpha}^\top \mathbf{A}||_\infty \leq \lambda. \tag{25}$$

Since $\boldsymbol{\alpha} = \boldsymbol{\xi}$, when λ is arbitrarily small, MPL will stop at a point where $||\boldsymbol{\xi}|| = ||\boldsymbol{\alpha}|| \ll ||\widehat{\mathbf{e}}||$ (where $\widehat{\mathbf{e}}$ denotes the ground-truth noise). In this case, the solution is no longer sparse, and the over-fitting problem will also arise.

To address these issues, we need to stop MPL earlier, thus we stop it if the following conditions are achieved:

$$||\boldsymbol{\alpha}^\top \mathbf{A}||_\infty \leq r_\infty \text{ or } ||\boldsymbol{\alpha}|| \leq r_2, \tag{26}$$

where r_∞ and r_2 are pre-determined parameters. We also use the relative function value difference to stop MPL if

$$\frac{\delta^t}{|\varrho\theta^0|} \leq \varepsilon, \tag{27}$$

where δ^t is the function value difference between the $(t-1)$th and tth iteration, ε is a small tolerance and θ^0 denotes the initial objective value. Since $\theta^0 = -||\mathbf{b}||^2/2$, the stopping condition becomes $2|\delta^t|/(\varrho||\mathbf{b}||^2) \leq \varepsilon$.

The above early stopping criteria are not applicable to many standard LASSO methods, such as GPSR [27], FISTA [29] and PGH [32]. Since a small λ cannot induce sparse solutions, the debiasing cannot be directly achieved through these methods. The active-set method in [40], [42] adopts similar optimization strategies to MPL, thus the above early stopping criteria can be also applied. For many greedy methods, such as CoSaMP [55] and SP [56], the criterion in (27) is not applicable since these methods may not converge for general problems.

IV. Performance Analysis

In this section, we study the complexity and convergence of MPL, and its performance guarantees on the sparse recovery.

A. Complexity of MPL

The complexity of MPL mainly includes two parts:

1) The worst-case analysis that needs to calculate $\mathbf{A}^\top \boldsymbol{\xi}$ of each outer iteration takes $O(mn)$ time complexity;
2) The inner master problem optimization w.r.t. $|\mathcal{I}|$ variables takes $O(|\mathcal{I}|n)$ complexity only.

On big dictionaries, the calculation of $\mathbf{A}^\top \boldsymbol{\xi}$ dominates the overall complexity. As previously mentioned, most of the existing ℓ_1-methods need to calculate $\mathbf{A}\mathbf{x}$ and $\mathbf{A}^\top \boldsymbol{\xi}$ many times. Therefore, in general they are expensive over big dictionaries. In contrast, MPL only needs to compute $\mathbf{A}^\top \boldsymbol{\xi}$ $O(k/\varrho)$ times. It is worth mentioning that, when tackling many signals simultaneously, the overall complexity of MPL can be further reduced to $O(|\mathcal{I}|(m+n))$ using a batch mode scheme, which can be found in Part II [80].

Remark 1: By fixing $\varrho = 1$ and solving the master problem exactly, MPL includes the active methods (e.g., [40], [41], [43]) and OMP [47] as special cases for $\lambda > 0$ and $\lambda = 0$, respectively.

Note that both OMP and active-set methods need to compute $\mathbf{A}^\top \boldsymbol{\xi}$ at least k times, which takes $O(mnk)$ cost. Therefore, they are expensive when dealing with big dictionaries. In MPL, we have used a fixed ϱ for each constraint in the QCLP problem (13); however, a changeable ϱ is also applicable. In other words, we can adaptively set different ϱ for different outer iterations. If we adopt the thresholding strategies used in SWCGP and StOMP, this leads to the following remark.

Remark 2: SWCGP [68] and StOMP [69] address the QCLP problem with an adaptive ϱ, where $\lambda = 0$.

The thresholding rule in StOMP is not applicable to general dictionaries [69], [68]. Moreover, it is expensive for StOMP to deal with large-scale problems since it solves the master problem exactly. SWCGP performs only one iteration w.r.t. the inner loop [68]. Accordingly, the master problem may not be sufficiently optimized. As a result, many non-active atoms may be mistakenly included, and more iterations may be required to converge. Consequently, the cost will increase.

B. Convergence of MPL

In this subsection, we discuss the convergence of MPL. Firstly, MPL converges to an optimal solution of problem (13).

Theorem 2: (Lemma 3.1 and Theorem 3.1 in [77]) The sequence $\{\delta^t\}(\delta^t \geq 0)$ generated in Algorithm 1 converges to 0. Moreover, there exists $\widehat{t}$, such that $\boldsymbol{\alpha}^{\widehat{t}-1}$ is feasible for (13) and Algorithm 1 stops in the $\widehat{t}$th iteration with $\boldsymbol{\alpha}^{\widehat{t}}$ to be an optimal solution of (13).

The proof can be adapted from [77]. It is worth mentioning that the above convergence property does not rely on any restricted conditions, that is, MPL converges on non-RIP dictionaries.

Now we demonstrate that the accumulation point of $\{\mathbf{x}^t\}$ satisfies the optimality conditions of LASSO if the early stopping is not applied.

Lemma 2: Let $\mathbf{x}^*$ be the accumulation point of $\{\mathbf{x}^t\}$ generated by MPL, then $\mathbf{x}^*$ is also an optimal solution to (4). Moreover, let $\mathcal{T}$ denote the support of the ground-truth $\widehat{\mathbf{x}}$ and $\mathbf{h} = \widehat{\mathbf{x}} - \mathbf{x}^*$, if $||\nabla\varphi(\widehat{\mathbf{x}})||_\infty \leq \epsilon\lambda$, where $0 < \epsilon \leq \frac{1}{2}$, then $\mathbf{h}$ satisfies the restricted set condition, namely $\mathbf{h} \in \Gamma_C = \{\mathbf{h} \in \mathbb{R}^m : ||\mathbf{h}_{\mathcal{T}^c}||_1 \leq D||\mathbf{h}_{\mathcal{T}}||_1\}$, where $D = \frac{1+\epsilon}{1-\epsilon}$.

The proof can be found in Appendix D.

In the following, we further show that MPL decreases the objective value exponentially under restricted conditions. Without loss of generality, we assume MPL stops when $||\boldsymbol{\alpha}^\top \mathbf{A}||_\infty \leq \lambda$, where λ is properly chosen. Let $f(\mathbf{x}^*)$ and $\mathbf{e}^*$ be the optimal function value and regression error of the LASSO problem, respectively, $\mathcal{I}^*$ be the support of $\mathbf{x}^*$. Assuming $|\mathcal{I}^*| \leq k$, we have the following theorem.

Theorem 3: Suppose $\lambda > 0$ is properly chosen. If $f(\mathbf{x}^t) \geq Cf(\mathbf{x}^*) = C(\frac{1}{2}\|\mathbf{e}^*\|^2 + \lambda\|\mathbf{x}^*\|_1)$, where $C > 1$, and there exists an $\iota > 1$ such that $\sigma_{k+\iota\varrho} < 1/2$, and the inner loop of MPL stops when (22) is satisfied, MPL linearly decreases in objective values if $t < \iota$, namely $f(\mathbf{x}^{t+1}) \leq \nu f(\mathbf{x}^t)$, where $0 < \nu < 1$ is a constant.

The proof can be found in Appendix E.

When $\lambda > 0$ is arbitrarily small or $\lambda = 0$, without early stopping, MPL will stop at a point which is over-fitted. To address this, we can stop MPL early based on condition (26) or (27), thus MPL will never reach to the optimal solution $\mathbf{x}^*$ of LASSO. Instead, it will approach to $\widehat{\mathbf{x}}$, where $\widehat{\mathbf{x}}$ and $\widehat{\mathbf{e}}$ denote the ground-truth k-sparse vector and additive noise, respectively.

Theorem 4: Suppose $\lambda \geq 0$ and $f(\mathbf{x}^t) \geq Cf(\widehat{\mathbf{x}}) = C(\frac{1}{2}\|\widehat{\mathbf{e}}\|^2 + \lambda\|\widehat{\mathbf{x}}\|_1)$ (where $C > 1$). Suppose the inner loop of MPL stops when (22) is satisfied, and there exists an integer $\iota > 0$ such that $\sigma_{k+\iota\varrho} < 1/2$, MPL linearly decreases in objective values for $t < \iota$, namely $f(\mathbf{x}^{t+1}) \leq \nu f(\mathbf{x}^t)$, where $0 < \nu < 1$ is a constant.

The proof can be found in Appendix F.

When $\lambda = 0$ and $\varrho = 1$, Theorem 4 improves the results of OMP in [81], where the exponential convergence of OMP has been revealed under more restricted conditions.

C. Sparse Signal Recovery Guarantees

The condition for exact sparse recovery is important for signal processing, and is usually expressed in terms of RIP constant. Based on Lemma 2, we obtain the following recovery condition for MPL without early stopping, i.e., MPL converges to the LASSO solution.

Theorem 5: Given $\lambda > 0$, suppose the ground-truth noise $\widehat{\mathbf{e}}$ satisfies $\|\mathbf{A}^\top\widehat{\mathbf{e}}\|_\infty \leq \epsilon\lambda$, where $\epsilon < 1$, and MPL stops without early stopping. Let $\mathbf{h} = \widehat{\mathbf{x}} - \mathbf{x}^*$, where $\widehat{\mathbf{x}}$ denotes a ground-truth k-sparse signal and $\mathbf{x}^*$ denotes the accumulation point of MPL. Let $g(e) = 1 + (\frac{63}{\sqrt{24}}\frac{1}{1-8e} + \frac{\sqrt{24}}{4})\frac{\sqrt{1+e}}{\sqrt{39-24e}}$, where $e = \frac{(D-1)}{(8D+1)} \in (0, 0.125)$ and $D = \frac{1+\epsilon}{1-\epsilon}$. If the restricted isometry constant satisfies $\sigma_k < 1/g(e) = 1/g(0) - \vartheta$, where $\vartheta = 1/g(0) - 1/g(e)$, we have

$$\|\mathbf{h}\| \leq \frac{2\sqrt{2}\sqrt{k}\lambda}{1 - g(e)\sigma_k}.$$

A simplified bound follows that $\sigma_k < 0.307 - \vartheta < 1/g(0) - \vartheta$.

The proof can be found in Appendix G.

According to Theorem 5, to recover a k-sparse signal, we must select λ such that $\|\mathbf{A}^\top\widehat{\mathbf{e}}\|_\infty \leq \epsilon\lambda$ (where $\epsilon < 1$). When $\epsilon = 0$, we must have $\|\mathbf{A}^\top\widehat{\mathbf{e}}\|_\infty = 0$ and $\widehat{\mathbf{e}} = \mathbf{0}$, which corresponds to the noiseless case. By the definition of ϑ in Theorem 5, when $\epsilon = 0$, we shall have $\vartheta = 0$ and $\sigma_k < 0.307$. That is, for the noiseless case, MPL will recover a k-sparse signal if $\sigma_k < 0.307$ and $\lambda > 0$.

On the contrary, if $\epsilon \to 1$, we will have $D = \frac{1+\epsilon}{1-\epsilon} \to +\infty$ and $e = \frac{(D-1)}{(8D+1)} \to 0.125$, which implies $\vartheta \to \frac{1}{g(0)}$ and $\sigma_k \to 0$. In this case, the recovery of $\widehat{\mathbf{x}}$ tends to be impossible in theory. In particular, if setting $\lambda = 0$, even for the noiseless case, MPL cannot guarantee to recover a k-sparse signal.

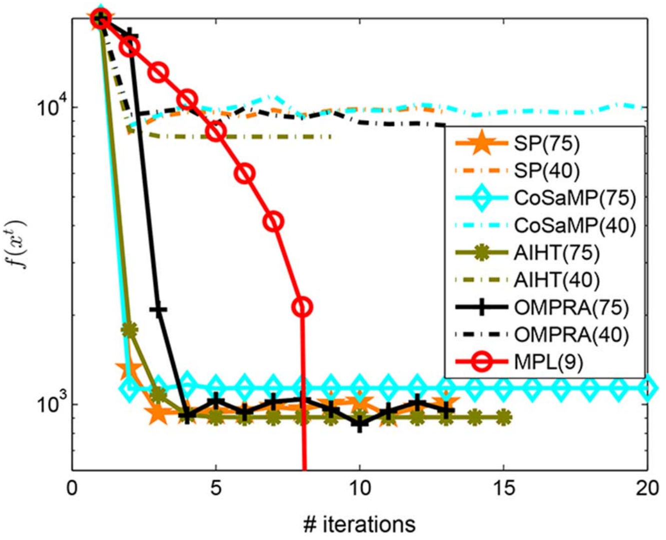


Fig. 1. Objective value evolutions of various greedy methods on a non-RIP problem. For MPL, the numbers in brackets denote the value of ϱ. The objective value at the 9th iteration for MPL is $f(\mathbf{x}^9) = 4.10\text{E} - 5$.

These observations highlight the importance of ℓ_1-norm regularization, though a large λ may incur biased solution. In practice, with the early stopping, MPL with a small λ (such as OMP which is a special case of MPL when $\lambda = 0$) still has promising sparse recovery performance. In the future, we will investigate the recovery conditions of MPL with a small value of λ (including $\lambda = 0$) for general cases when early stopping is applied.

V. Numerical Studies

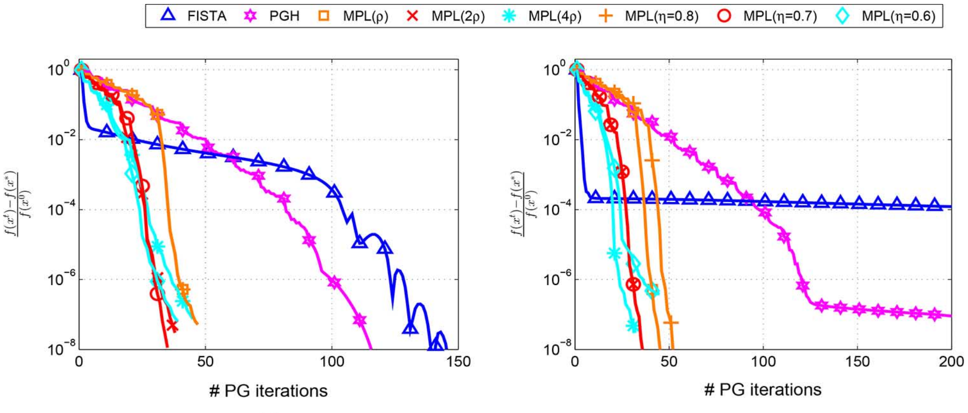
A. Convergence on Non-RIP Dictionaries

To demonstrate this property, we design a synthetic non-RIP problem and compare the performance of SP, CoSaMP, AIHT, OMPR and the proposed **MPL** method.[3] We first generate a *Gaussian* random matrix $\mathbf{A} \in \mathbb{R}^{2^{10}\times 2^{13}}$, and then simply set $\mathbf{A}(:, 41:80) = \mathbf{A}(:, 1:40)$ (in Matlab notation). Since there are repeated columns, the RIP condition of $\mathbf{A}$ does not hold [63]. Moreover, we generate a 40-sparse ground truth $\mathbf{x}$ by letting $\mathbf{x}(1:40) = \mathbf{1}$ and $\mathbf{x}(41:8092) = \mathbf{0}$. Subsequently, $\mathbf{b}$ is produced by $\mathbf{b} = \mathbf{A}\mathbf{x}$ without noise. The objective values and the objective gap between iterations of several greedy methods are reported in Fig. 1, where we calculate $f(\mathbf{x}^t) = \|\mathbf{b} - \mathbf{A}\mathbf{x}^t\|^2$. For the baseline methods, the numbers in the brackets of the legend record the value of $\widehat{k}$. Since $\|\mathbf{x}\|_0 = 40$, $\widehat{k} = 40$ is the best estimation of k; while $\widehat{k} = 75$ overestimates k. From Fig. 1, only the proposed MPL method achieves a global solution (where $f(\mathbf{x}^9) = 4.10\text{E}-5$) after 9 iterations. As shown in Fig. 1, AIHT and CoSaMP(75) achieve a local minimum only; while the rest of the baselines cannot converge on this synthetic problem.

B. Convergence to LASSO Solution

According to Lemma 2, MPL essentially addresses LASSO through a greedy procedure. To demonstrate this, following

[3] The C++ source codes of MPL and the compared methods are available at: http://www.tanmingkui.com/mpl.html.



(a) (b)

Fig. 2. Convergence of different methods on a ±1 sparse signal $\mathbf{x}$, where $\|\mathbf{x}\|_0 = 140$. According to (23), we set a base $\varrho = \lceil n/(8\log(m))\rceil = 14$, and study MPL with $\varrho, 2\varrho$ and 4ϱ, respectively. According to (24), we study MPL with $\eta = 0.8, \eta = 0.7$ and $\eta = 0.6$, respectively. (a) Objective values for $\lambda = 0.005\|\mathbf{A}^\top\mathbf{b}\|_\infty$. (b) Objective values for $\lambda = 0.00005\|\mathbf{A}^\top\mathbf{b}\|_\infty$.

TABLE I
COMPUTATION TIME (IN SECONDS) DIFFERENT METHODS. THE VALUES OF ϱ FOR $\eta = 0.8, 0.7$ AND 0.6 ARE 12, 27, 60, RESPECTIVELY

λ	FISTA	PGH	MPL					
			ϱ	2ϱ	4ϱ	η = 0.8	η = 0.7	η = 0.6
λ_1	5.02	1.31	0.12	0.09	0.08	0.14	0.07	0.09
λ_2	70.48	13.93	0.13	0.08	0.09	0.11	0.09	0.11

[32], we conduct a sparse recovery experiment from an observation $\mathbf{b} = \mathbf{A}\mathbf{x} + \mathbf{e}$, where $\mathbf{A} \in \mathbb{R}^{2^{10}\times 2^{13}}$ is a Gaussian random matrix, $\mathbf{x}$ is a 140-sparse ±1 signal and $\mathbf{e}$ denotes the noise uniformly sampled from $[-0.01, 0.01]$. PGH [32] and FISTA [29] are adopted as the baselines. Detailed experimental settings can be found in the caption of Fig. 2, which records the relative objective values w.r.t. iterations w.r.t. $\lambda_1 = 0.005\|\mathbf{A}^\top\mathbf{b}\|_\infty$ and $\lambda_2 = 0.00005\|\mathbf{A}^\top\mathbf{b}\|_\infty$, respectively. Table II records the time taken by different methods. From Fig. 2 and Table II, MPL with different ϱ's converges to an optimal solution of LASSO. Mover, compared with PGH and FISTA, MPL is much faster with different λ's (particularly when λ is very small).

VI. CONCLUSION

In this paper, we have proposed a Matching Pursuit LASSO (MPL) algorithm to address large-scale SR problems based on a QCLP reformulation for LASSO. MPL iteratively adds ϱ new atoms per iteration, thus avoiding the atom replacement or atom deletion that is required in many greedy algorithms, such as CoSaMP [55] and SP [56]. As a result, the convergence of MPL over non-RIP dictionaries is guaranteed. Our convergence analysis shows that, MPL can even converge linearly with a properly chosen ϱ under mild RIP conditions. Unlike many greedy algorithms, such as CoSaMP, MPL does not need to specify an exact estimation of the target sparsity. Typically, setting ϱ is much simpler than the estimation of sparsity.

Due to the ℓ_1-norm regularization, MPL can recover k-sparse signals if the restricted isometry constant σ_k satisfies $\sigma_k < 0.307 - \vartheta$. More importantly, MPL greatly improves the efficiency of existing ℓ_1-norm methods by reducing the calculations of matrix-vector products. Last but not least, due to the optimization scheme of MPL, a batch-mode MPL can be developed to vastly speed up sparse recovery with many signals. The batch-mode MPL as well as more numerical comparisons with existing methods will be presented in Part II [80].

APPENDIX A
PROOF OF PROPOSITION 1

Note that $\boldsymbol{\xi} = \mathbf{b} - \mathbf{A}(\mathbf{x} \odot \boldsymbol{\tau})$. For any $\boldsymbol{\tau} \neq \mathbf{0}$, we have $\min_{\mathbf{x}} \lambda\|\mathbf{x}\|_1\rangle + \frac{1}{2}\|\boldsymbol{\xi}\|^2 = \min_{\mathbf{x}} \lambda\|\mathbf{x}_{\mathcal{I}}\|_1 + \frac{1}{2}\|\mathbf{b} - \mathbf{A}_{\mathcal{I}}\mathbf{x}_{\mathcal{I}}\|^2$, where $\mathcal{I} = \text{support}(\boldsymbol{\tau})$. By introducing Lagrangian multipliers $\boldsymbol{\alpha} \in \mathbb{R}^n$ to the equality constraint $\boldsymbol{\xi} = \mathbf{b} - \mathbf{A}(\mathbf{x} \odot \boldsymbol{\tau})$, the Lagrangian function of the inner problem is:

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\xi}, \boldsymbol{\alpha}) = \lambda\|\mathbf{x}\|_1 + \frac{1}{2}\|\boldsymbol{\xi}\|^2 - \boldsymbol{\alpha}^\top(\boldsymbol{\xi} - \mathbf{b} + \mathbf{A}\text{diag}(\boldsymbol{\tau})\mathbf{x}).$$

Given a fixed $\boldsymbol{\tau}$, $\lambda\|\mathbf{x}_{\mathcal{I}}\|_1 + \frac{1}{2}\|\mathbf{b} - \mathbf{A}\mathbf{x}_{\mathcal{I}}\|^2$ is a convex function w.r.t. $\mathbf{x}_{\mathcal{I}}$, and the strong duality for this problem holds [26]. We minimize $\mathcal{L}(\mathbf{x}, \boldsymbol{\xi}, \boldsymbol{\alpha})$ w.r.t. $\mathbf{x}$ and $\boldsymbol{\xi}$, respectively:

$$\begin{aligned}
&\min_{\mathbf{x}} \lambda\|\mathbf{x}\|_1 - \boldsymbol{\alpha}^\top\mathbf{A}\text{diag}(\boldsymbol{\tau})\mathbf{x}\\
&\quad = \begin{cases} \mathbf{0}, & \text{if } \|(\mathbf{A}\text{diag}(\boldsymbol{\tau}))^\top\boldsymbol{\alpha}\|_\infty \leq \lambda,\\ -\infty, & \text{otherwise.}\end{cases}\\
&\min_{\boldsymbol{\xi}} \frac{1}{2}\|\boldsymbol{\xi}\|^2 - \boldsymbol{\alpha}^\top\boldsymbol{\xi} + \boldsymbol{\alpha}^\top\mathbf{b} = -\frac{1}{2}\|\boldsymbol{\alpha}\|^2 + \boldsymbol{\alpha}^\top\mathbf{b}.
\end{aligned}$$

To obtain the minimum of $\mathcal{L}(\mathbf{x}, \boldsymbol{\xi}, \boldsymbol{\alpha})$ w.r.t. $\boldsymbol{\xi}$, we set the derivative of $\mathcal{L}(\mathbf{x}, \boldsymbol{\xi}, \boldsymbol{\alpha})$ w.r.t. $\boldsymbol{\xi}$ to $\mathbf{0}$, where we obtain $\boldsymbol{\alpha} = \boldsymbol{\xi} \in \mathbb{R}^n$ at the optimality. Substituting these equations into $\mathcal{L}(\mathbf{x}, \boldsymbol{\xi}, \boldsymbol{\alpha})$, we arrive at (10). This completes the proof.

APPENDIX B
PROOF OF PROPOSITION 2

To complete the proof, we introduce the following lemma.

Lemma 3: In the tth iteration of Algorithm 1, let $\boldsymbol{\mu} \in \Pi = \{\boldsymbol{\mu} | \boldsymbol{\mu} \succeq 0, \sum_{i=1}^{t} \mu_i = 1\}$, problem (14) can be solved by the following minimax problem:

$$\min_{\boldsymbol{\mu}\in\Pi} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \frac{1}{2} z^{t-1} - \frac{1}{2} \sum_{\boldsymbol{\tau}_i \in \Lambda_t} \mu_i f(\boldsymbol{\alpha}, \boldsymbol{\tau}_i). \tag{28}$$

Proof: By introducing the dual variable ν and $\{\hat{\mu}_i\}$ to the constraint, the Lagrangian function of problem (14) is $\mathcal{L}(\boldsymbol{\alpha}, \theta, \delta, \nu, \hat{\mu}) = \delta - \nu(\theta - z^{t-1} + \delta) - \sum_i \hat{\mu}_i (f(\boldsymbol{\alpha}, \boldsymbol{\tau}_i) - \theta + \delta)$. Setting the gradient of $\mathcal{L}(\cdot)$ with respect to θ and δ to zeros, we get the optimal dual variables $\{\nu\}$ and $\{\hat{\mu}_i\}$ satisfying $\sum_i \hat{\mu}_i + \nu = 1 \text{and} \sum_i \hat{\mu}_i - \nu = 0$. Therefore, we have $\sum_i \hat{\mu}_i = \frac{1}{2}$ and $\nu = \frac{1}{2} > 0$. From the complementary condition, $\theta + \delta = z^{t-1}$ holds. Moreover, we can exchange the order of max and min operators using minimax Theorem [82]. Finally, let $\mu_i = 2\hat{\mu}_i$, and we complete the proof. □

Now we complete the proof by showing that the dual problem of (16) is (28). Recall that there is no overlapping element among $\mathcal{J}_i$'s. Since z^{t-1} is a constant and $\sum_i \mu_i = 1$, problem (28) can be simplified as:

$$\begin{aligned} &\max_{\boldsymbol{\alpha}} -\frac{1}{2}\|\boldsymbol{\alpha}\|^2 + \boldsymbol{\alpha}^\top \mathbf{b}. \\ &\text{s.t. } \|\boldsymbol{\alpha}^\top \mathbf{A} \text{diag}(\boldsymbol{\tau}_i)\|_\infty \le \lambda, \quad \forall \boldsymbol{\tau}_i \in \Lambda_t. \end{aligned} \tag{29}$$

Let $\mathbf{x}_{\mathcal{J}_i}$ be the regressor regarding $\mathbf{A}_{\mathcal{J}_i}$. By assumption, we have $\|\mathbf{x}_{\mathcal{I}_t}\|_1 = \sum_{i=1}^{T} \|\mathbf{x}_{\mathcal{J}_i}\|_1$ and $\boldsymbol{\xi} = \mathbf{b} - \sum_{i=1}^{T} \mathbf{A}_{\mathcal{J}_i} \mathbf{x}_{\mathcal{J}_i}$. Therefore, problem (16) can be rewritten as:

$$\min_{\mathbf{x}_{\mathcal{J}_i}, \boldsymbol{\xi}} \lambda \sum_{i=1}^{T} \|\mathbf{x}_{\mathcal{J}_i}\|_1 + \frac{1}{2}\|\boldsymbol{\xi}\|^2, \text{s.t. } \boldsymbol{\xi} = \mathbf{b} - \sum_{i=1}^{T} \mathbf{A}_{\mathcal{J}_i} \mathbf{x}_{\mathcal{J}_i}. \tag{30}$$

By introducing the dual variable $\boldsymbol{\alpha}$ to the constraint $\boldsymbol{\xi} = \mathbf{b} - \sum_{i=1}^{T} \mathbf{A}_{\mathcal{J}_i} \mathbf{x}_{\mathcal{J}_i}$, the Lagrangian function becomes

$$\begin{aligned} \mathcal{L}(\mathbf{x}, \boldsymbol{\xi}, \boldsymbol{\alpha}) = \lambda \sum_{i=1}^{T} \|\mathbf{x}_{\mathcal{J}_i}\|_1 \\ + \frac{1}{2}\|\boldsymbol{\xi}\|^2 - \boldsymbol{\alpha}^\top \left(\boldsymbol{\xi} - \mathbf{b} + \sum_{i=1}^{T} \mathbf{A}_{\mathcal{J}_i} \mathbf{x}_{\mathcal{J}_i} \right). \end{aligned}$$

To derive the dual form, we minimize $\mathcal{L}(\mathbf{x}, \boldsymbol{\xi}, \boldsymbol{\alpha})$ w.r.t. $\mathbf{x}_{\mathcal{J}_i}$ and $\boldsymbol{\xi}$:

$$\begin{aligned} &\min_{\mathbf{x}_{\mathcal{J}_i}} \lambda \sum_{i=1}^{T} \|\mathbf{x}_{\mathcal{J}_i}\|_1 - \boldsymbol{\alpha}^\top \left(\sum_{i=1}^{T} \mathbf{A}_{\mathcal{J}_i} \mathbf{x}_{\mathcal{J}_i} \right) \\ &= \begin{cases} \mathbf{0}, & \text{if } \|\mathbf{A}_{\mathcal{J}_i}^\top \boldsymbol{\alpha}\|_\infty \le \lambda, \forall i, \\ -\infty, & \text{Otherwise.} \end{cases} \\ &\min_{\boldsymbol{\xi}} \frac{1}{2}\|\boldsymbol{\xi}\|^2 - \boldsymbol{\alpha}^\top \boldsymbol{\xi} + \boldsymbol{\alpha}^\top \mathbf{b} = -\frac{1}{2}\|\boldsymbol{\alpha}\|^2 + \boldsymbol{\alpha}^\top \mathbf{b}. \end{aligned}$$

We then can verify that (29) is the Lagrangian dual of (30). In particular, we have $\boldsymbol{\alpha}^* = \boldsymbol{\xi}^*$ by $\nabla_{\boldsymbol{\xi}} \mathcal{L}(\mathbf{x}, \boldsymbol{\xi}, \boldsymbol{\alpha}) = \mathbf{0}$ at the optimality. This completes the proof.

Appendix C
Proof of Lemma 1

Proof: From the worst-case analysis, $|g_i| > \lambda$ is guaranteed; otherwise, the ith atom will not be selected.

Recall that the master problem is w.r.t. subvariables indexed by $\mathcal{I}_t$. For any given $\mathcal{I}$, we redefine $\varphi_{\mathcal{I}}(\mathbf{x}_{\mathcal{I}}) = \frac{1}{2}\|\mathbf{b} - \mathbf{A}_{\mathcal{I}} \mathbf{x}_{\mathcal{I}}\|^2$, $f_{\mathcal{I}}(\mathbf{x}_{\mathcal{I}}) = \lambda\|\mathbf{x}_{\mathcal{I}}\|_1 + \varphi(\mathbf{x}_{\mathcal{I}})$, and $\phi_{\mathcal{I}}(\mathbf{x}_{\mathcal{I}}, \mathbf{u}_{\mathcal{I}}) = \lambda\|\mathbf{x}_{\mathcal{I}}\|_1 + \varphi(\mathbf{u}_{\mathcal{I}}) + \nabla\varphi_{\mathcal{I}}(\mathbf{u}_{\mathcal{I}})^\top (\mathbf{x}_{\mathcal{I}} - \mathbf{u}_{\mathcal{I}}) + \frac{L}{2}\|\mathbf{x}_{\mathcal{I}} - \mathbf{u}_{\mathcal{I}}\|^2$. Let $G_{\mathcal{I}}(\mathbf{x}^t)$ be the generalized gradient w.r.t. $\mathcal{I}$. For any point $\mathbf{u}_{\mathcal{I}}$, let $\mathbf{o}_{\mathcal{I}} = \mathbf{u}_{\mathcal{I}} - \mathbf{g}_{\mathcal{I}}/L$ and $\mathbf{o}_{\mathcal{I}^c} = \mathbf{0}$, the following property holds for the generalized gradient [34].

$$\phi_{\mathcal{I}}\left(\mathbf{u}_{\mathcal{I}}, S_{L,\lambda}(\mathbf{o}_{\mathcal{I}})\right) \le f_{\mathcal{I}}(\mathbf{u}_{\mathcal{I}}) - \frac{1}{2L}\|G_{\mathcal{I}}(\mathbf{x}^t)\|^2, \tag{31}$$

Let $\mathbf{x}^t$ be the solution (or approximate solution) to the master problem at the tth iteration. Then $G_{\mathcal{I}_t}(\mathbf{x}^t) = \boldsymbol{\epsilon}_{\mathcal{I}_t}$ holds, where $\boldsymbol{\epsilon}_{\mathcal{I}_t} = \mathbf{0}$ if we solve the master problem exactly. At the $(t+1)$ iteration, by warm start, we have $\mathbf{u}^0 = \mathbf{x}^t$ and $f(\mathbf{x}^t) = f_{\mathcal{I}_t}(\mathbf{x}^t_{\mathcal{I}_t})$, where $\mathbf{x}^t_{\mathcal{I}^c_t} = \mathbf{0}$. Let $\mathbf{o}^0_{\mathcal{I}_{t+1}} = \mathbf{u}^0_{\mathcal{I}_{t+1}} - \mathbf{g}_{\mathcal{I}_{t+1}}/L$ and $\mathbf{o}^0_{\mathcal{I}^c_{t+1}} = \mathbf{0}$, by (31), we have

$$\begin{aligned} &\phi_{\mathcal{I}}\left(\mathbf{u}^0_{\mathcal{I}_{t+1}}, S_{L,\lambda}\left(\mathbf{o}^0_{\mathcal{I}_{t+1}}\right)\right) \\ &\qquad \le f_{\mathcal{I}_{t+1}}\left(\mathbf{x}^t_{\mathcal{I}_{t+1}}\right) - \frac{1}{2L}\left\|G_{\mathcal{J}_{t+1}}(\mathbf{x}^t)\right\|^2 - \frac{1}{2L}\left\|\boldsymbol{\epsilon}_{\mathcal{I}_t}\right\|^2. \end{aligned}$$

With proper line search, in the first iteration of PG, $f(\mathbf{u}^1) = f_{\mathcal{I}_{t+1}}(\mathbf{u}^0_{\mathcal{I}_{t+1}} - 1/LG_{\mathcal{I}_{t+1}}(\mathbf{o}^0_{\mathcal{I}_{t+1}})) \le \phi_{\mathcal{I}}(\mathbf{u}^0_{\mathcal{I}_{t+1}}, S_{L,\lambda}(\mathbf{o}^0_{\mathcal{I}_{t+1}}))$ holds [34]. Since $f_{\mathcal{I}_{t+1}}(\mathbf{x}^t_{\mathcal{I}_{t+1}}) = f(\mathbf{x}^t)$, we have $f(\mathbf{u}^1) \le f(\mathbf{u}^0_{\mathcal{I}_{t+1}} - 1/LG_{\mathcal{I}_{t+1}}(\mathbf{o}^0_{\mathcal{I}_{t+1}})) \le f(\mathbf{x}^t) - \frac{1}{2L}\|G_{\mathcal{J}_{t+1}}(\mathbf{x}^t)\|^2 - \frac{1}{2L}\|\boldsymbol{\epsilon}_{\mathcal{I}_t}\|^2$. Finally, since $G(\mathbf{u}^k) = L(\mathbf{u}^k - S_{L,\lambda}(\mathbf{u}^k - \mathbf{g}/L))$, we have $\|\mathbf{G}_{\mathcal{J}_{t+1}}(\mathbf{x}^t)\|^2 = \|\mathbf{G}_{\mathcal{J}_{t+1}}(\mathbf{u}^0)\|^2 = \|L(-S_{L,\lambda}(0 - \mathbf{g}/L))\|^2 = \sum_{i \in \mathcal{J}_{t+1}} (|g_i| - \lambda)^2$. By eliminating the positive term $\frac{1}{2L}\|\boldsymbol{\epsilon}_{\mathcal{I}_t}\|^2$, we complete the proof. □

Appendix D
Proof of Lemma 2

Proof: First, from the proof of Proposition 1, at the optimality of LASSO, we have $\|\mathbf{A}^\top \boldsymbol{\xi}\|_\infty \le \lambda$ with $\boldsymbol{\xi} = \boldsymbol{\alpha}$. Let $\mathcal{T}$ be the support of $\mathbf{x}^*$, then we have $\|\mathbf{A}^\top_{\mathcal{T}} \boldsymbol{\xi}\|_\infty = \lambda$ and $\|\mathbf{A}^\top_{\mathcal{T}^c} \boldsymbol{\xi}\|_\infty < \lambda$.

For MPL, let $\mathcal{S}$ denote the set of selected atoms and $\mathcal{S}^c$ denote the complementary set regarding $\mathcal{T}$. On these selected atoms, as from Theorem 2, we have $\|\mathbf{A}^\top_{\mathcal{S}_{\mathcal{T}}} \boldsymbol{\xi}\|_\infty = \lambda$ and $\|\mathbf{A}^\top_{\mathcal{S}^c_{\mathcal{T}}} \boldsymbol{\xi}\|_\infty < \lambda$, where $\mathcal{S}_{\mathcal{T}}$ denotes the support of $\mathbf{x}_{\mathcal{S}}$ regarding the master problem, and $\mathcal{S}^c_{\mathcal{T}}$ denotes the complementary set regarding $\mathcal{S}$. Furthermore, since we stop MPL when $\|\mathbf{A}^\top \boldsymbol{\xi}\|_\infty < \lambda$, we have $\|\mathbf{A}^\top_{\mathcal{S}^c} \boldsymbol{\xi}\|_\infty < \lambda$ regarding the non-selected atoms indexed by $\mathcal{S}^c$. Accordingly, the accumulation point of MPL is a KKT point of LASSO.

Let $\widehat{\mathbf{x}}$ be the ground truth and $\mathbf{x}^*$ be the optimal solution to (4) by MPL. At the optimality of (4), we have

$$\varphi(\widehat{\mathbf{x}}) + \lambda\|\widehat{\mathbf{x}}\|_1 \ge \varphi(\mathbf{x}^*) + \lambda\|\mathbf{x}^*\|_1. \tag{32}$$

Due to the convexity of $\varphi(\mathbf{x})$, $\varphi(\mathbf{x}^*) \ge \varphi(\widehat{\mathbf{x}}) + \langle \nabla\varphi(\widehat{\mathbf{x}}), \mathbf{h} \rangle \ge \varphi(\widehat{\mathbf{x}}) - \|\nabla\varphi(\widehat{\mathbf{x}})\|_\infty \cdot \|\mathbf{h}\|_1$. Together with (32), we obtain

$$\|\nabla\varphi(\widehat{\mathbf{x}})\|_\infty \cdot \|\mathbf{h}\|_1 \ge \lambda\|\mathbf{x}^*\|_1 - \lambda\|\widehat{\mathbf{x}}\|_1$$

Suppose $\|\nabla\varphi(\widehat{\mathbf{x}})\|_\infty \le \epsilon\lambda$, where $\epsilon > 0$, we have

$$\begin{aligned}\epsilon\|\mathbf{h}\|_1 + \|\widehat{\mathbf{x}}\|_1 \ge \|\mathbf{x}^*\|_1 &= \|\widehat{\mathbf{x}} + \mathbf{h}\|_1\\ &= \|\widehat{\mathbf{x}} + \mathbf{h}_{\mathcal{T}} + \mathbf{h}_{\mathcal{T}^c}\|_1\\ &= \|\mathbf{h}_{\mathcal{T}^c}\|_1 + \|\widehat{\mathbf{x}} + \mathbf{h}_{\mathcal{T}}\|_1\\ &\ge \|\mathbf{h}_{\mathcal{T}^c}\|_1 + \|\widehat{\mathbf{x}}\|_1 - \|\mathbf{h}_{\mathcal{T}}\|_1.\end{aligned}$$

Thus, we have $\epsilon(\|\mathbf{h}_{\mathcal{T}}\|_1 + \|\mathbf{h}_{\mathcal{T}^c}\|_1) \ge \|\mathbf{h}_{\mathcal{T}^c}\|_1 - \|\mathbf{h}_{\mathcal{T}}\|_1$, thus $\|\mathbf{h}_{\mathcal{T}^c}\|_1 \le \frac{1+\epsilon}{1-\epsilon}\|\mathbf{h}_{\mathcal{T}}\|_1$. Let $D = \frac{1+\epsilon}{1-\epsilon}$. For $0 < \epsilon \le \frac{1}{2}, 1 < D \le 3$. This completes the proof. □

Appendix E
Proof of Theorem 3

To complete the proof, we will use the following notations and definitions. Let $\mathcal{E} = \mathcal{I}_t \backslash \mathcal{I}^*, \mathcal{Q} = \mathcal{I}^* \backslash \mathcal{I}_t, \mathcal{C} = \mathcal{I}_t \backslash \mathcal{I}^*$. Without loss of generality, suppose $|\mathcal{I}^*| \le k$ for a given regularization parameter λ. After t iterations, $\|\mathbf{x}^t\|_0 \le t\varrho$. Thus $\mathbf{x}^t - \mathbf{x}^*$ is at most $(k + t\varrho)$ sparse. We also define two internal variables $\mathbf{y}^{t+1} \in \mathbb{R}^m$ and $\mathbf{z}^{t+1} \in \mathbb{R}^m$ as:

$$\begin{aligned}&\mathbf{y}^{t+1}_{\mathcal{I}_{t+1}} = \mathbf{x}^t_{\mathcal{I}_t+1} - 1/L_t G_{\mathcal{I}_{t+1}}(\mathbf{x}^t) \text{ and } \mathbf{y}^{t+1}_{\mathcal{I}^c_{t+1}} = \mathbf{0},\\ &\mathbf{z}^{t+1} = \mathbf{x}^t - 1/L_t G(\mathbf{x}^t).\end{aligned} \tag{33}$$

Here, L_t is determined by the line search. For the convenience of presentation, we omit the index t from L_t hereafter. Let

$$\chi = \frac{C\left(1 - 2\delta_{(k+t\varrho)}\right)^2}{(\sqrt{C}+1)^2(1 - \delta_{(k+t\varrho)})}. \tag{34}$$

The following lemma holds.

Lemma 4: For MPL, at the tth iteration, suppose $\delta_{(k+t\varrho)} < 1/2$, we have $f(\mathbf{x}^t) \ge \frac{\chi}{2}\|\mathbf{x}^*_{\mathcal{Q}}\|^2$.

Proof: Note that $\sqrt{f(\mathbf{x}^t)} = \sqrt{\frac{1}{2}\|\mathbf{A}\mathbf{x}^t - \mathbf{b}\|^2 + \lambda\|\mathbf{x}^t\|_1}$. As $\|\mathbf{x}^t - \mathbf{x}^*\|_0 \le k + t\varrho$, based on Lemma 16 in [60], we have

$$\begin{aligned}\sqrt{f(\mathbf{x}^t)} &\ge \frac{1}{\sqrt{2}}\left\|\mathbf{A}(\mathbf{x}^t - \mathbf{x}^*) - \mathbf{e}\right\|\\ &\ge \frac{1}{\sqrt{2}}\left(\left\|\mathbf{A}\left(\mathbf{x}^t - \mathbf{x}^*\right)\right\| - \|\mathbf{e}\|\right)\\ &\ge \frac{1}{\sqrt{2}}\left(\min_{\mathbf{x}_{\mathcal{I}^c_t} = \mathbf{0}}\|\mathbf{A}(\mathbf{x} - \mathbf{x}^*)\|\right) - \frac{1}{\sqrt{2}}\|\mathbf{e}\|\\ &\ge \frac{1}{\sqrt{2}}\left(\frac{1 - 2\delta_{(k+t\varrho)}}{\sqrt{(1 - \delta_{(k+t\varrho)})}}\left\|\mathbf{x}^*_{\mathcal{Q}}\right\| - \|\mathbf{e}\|\right).\end{aligned}$$

Since $f(\mathbf{x}^t) \ge Cf(\mathbf{x}^*) \ge \frac{C}{2}\|\mathbf{e}\|^2$, by simple calculation, we have

$$f(\mathbf{x}^t) \ge \frac{\chi}{2}\|\mathbf{x}^*_{\mathcal{Q}}\|^2. \tag{35}$$

□

Lemma 5: Let $\mathcal{J}_{t+1}$ be the index set of the ϱ new atoms obtained from the worst-case analysis at the $(t+1)$th iteration, we have:

$$f(\mathbf{x}^{t+1}) \le f(\mathbf{x}^t) - \frac{L}{2}\left\|\mathbf{y}_{\mathcal{J}_{t+1}}\right\|^2.$$

Proof: Lemma 5 is a direct consequence of Lemma 1. Typically, we have $f(\mathbf{x}^{t+1}) \le f(\mathbf{u}^1) \le f(\mathbf{x}^t) - \frac{1}{2L}\|G_{\mathcal{J}_{t+1}}(\mathbf{x}^t)\|^2$. Since $G_{\mathcal{J}_{t+1}}(\mathbf{x}^t) = L\mathbf{y}_{\mathcal{J}_{t+1}}$, we complete the proof. □

Lemma 6: Let $\mathbf{g} = \mathbf{A}^\top(\mathbf{A}_{\mathcal{I}_t}\mathbf{x}^t_{\mathcal{I}_t} - \mathbf{b})$, where $\mathbf{x}^t_{\mathcal{I}_t}$ is an approximate solution satisfying $G_{\mathcal{I}_t}(\mathbf{x}^t) = \boldsymbol{\epsilon}_{\mathcal{I}_t}$, where $\boldsymbol{\epsilon}_{\mathcal{I}^c_t} = \mathbf{0}$. Let $\varsigma = \partial\|\mathbf{z}^{t+1}\|_1$ be the subgradient of $\|\mathbf{z}^{t+1}\|_1$. Divide $\mathcal{Q}$ into two disjoint sets $\mathcal{Q}_1$ and $\mathcal{Q}_2$ such that $|g_i| \le \lambda, \forall i \in \mathcal{Q}_1$ and $|g_i| > \lambda, \forall i \in \mathcal{Q}_2$, then the following relation holds:

$$\begin{aligned}-\frac{1}{2}\left[\mathbf{x}^{*}_{\mathcal{Q}}{}^\top\right]\left[\mathbf{A}^\top_{\mathcal{Q}}\boldsymbol{\alpha}\right] = \frac{L}{2}\mathbf{x}^{*}_{\mathcal{Q}}{}^\top\mathbf{z}^{t+1}_{\mathcal{Q}} - \frac{1}{2}\mathbf{x}^{*}_{\mathcal{Q}_1}{}^\top\mathbf{A}^\top_{\mathcal{Q}_1}\boldsymbol{\alpha}\\ + \frac{1}{2}\lambda\varsigma^\top_{\mathcal{Q}_2}\mathbf{x}^*_{\mathcal{Q}_2},\end{aligned} \tag{36}$$

$$\left|\frac{1}{2}\mathbf{x}^{*}_{\mathcal{Q}_1}{}^\top\mathbf{A}^\top_{\mathcal{Q}_1}\boldsymbol{\alpha}\right| \le \frac{\lambda}{2}\|\mathbf{x}^*_{\mathcal{Q}_1}\|_1. \tag{37}$$

Proof: By definition, $\mathcal{Q} = \mathcal{I}^* \backslash \mathcal{I}_t, |g_i| < \lambda, \forall i \in \mathcal{Q}_1$, according to the definition of $\mathbf{z}^{t+1}$, it can be expressed as:

$$\mathbf{z}^{t+1}_{\mathcal{I}_t} = \mathbf{x}^t_{\mathcal{I}_t} - 1/LG_{\mathcal{I}_t}(\mathbf{x}^t) \text{ and } \mathbf{z}^{t+1}_{\mathcal{I}^c_t} = -1/LG_{\mathcal{I}^c_t}(\mathbf{x}^t).$$

In other words, we have $z^{t+1}_i = 0, \forall i \in \mathcal{Q}_1$ and $z^{t+1}_i = -\frac{1}{L}\text{sign}(g_i)(|g_i| - \lambda), \forall i \in \mathcal{Q}_2$.

$$\begin{aligned}&-\frac{1}{2}\left[\mathbf{x}^{*}_{\mathcal{Q}}{}^\top\right]\left[\mathbf{A}^\top_{\mathcal{Q}}\boldsymbol{\alpha}\right]\\ &= -\frac{1}{2}\mathbf{x}^{*}_{\mathcal{Q}_1}{}^\top\mathbf{A}^\top_{\mathcal{Q}_1}\boldsymbol{\alpha} - \frac{1}{2}\mathbf{x}^{*}_{\mathcal{Q}_2}{}^\top\mathbf{A}^\top_{\mathcal{Q}_2}\boldsymbol{\alpha}\\ &= \frac{L}{2}\mathbf{x}^{*}_{\mathcal{Q}}{}^\top\mathbf{z}^{t+1}_{\mathcal{Q}} - \frac{1}{2}\mathbf{x}^{*}_{\mathcal{Q}_1}{}^\top\mathbf{A}^\top_{\mathcal{Q}_1}\boldsymbol{\alpha} - \frac{1}{2}\mathbf{x}^*_{\mathcal{Q}_2}\left(\text{sign}(\mathbf{g}_{\mathcal{Q}_2}) \odot \boldsymbol{\lambda}_{\mathcal{Q}_2}\right)\\ &= \frac{L}{2}\mathbf{x}^{*}_{\mathcal{Q}}{}^\top\mathbf{z}^{t+1}_{\mathcal{Q}} - \frac{1}{2}\mathbf{x}^{*}_{\mathcal{Q}_1}{}^\top\mathbf{A}^\top_{\mathcal{Q}_1}\boldsymbol{\alpha} + \frac{1}{2}\lambda\varsigma^\top_{\mathcal{Q}_2}\mathbf{x}^*_{\mathcal{Q}_2},\end{aligned}$$

where $\boldsymbol{\lambda} = [\lambda, \ldots, \lambda]^\top \in \mathbb{R}^m$ and $\text{sign}(\mathbf{g}_{\mathcal{Q}_2})$ denotes the sign of vector $\mathbf{g}_{\mathcal{Q}_2}$. Finally, since $\|\mathbf{A}^\top_{\mathcal{Q}_1}\boldsymbol{\alpha}\|_\infty < \lambda$, it holds that:

$$\left|\frac{1}{2}\mathbf{x}^{*}_{\mathcal{Q}_1}{}^\top\mathbf{A}^\top_{\mathcal{Q}_1}\boldsymbol{\alpha}\right| \le \frac{1}{2}\left\|\mathbf{x}^*_{\mathcal{Q}_1}\right\|_1\left\|\mathbf{A}^\top_{\mathcal{Q}_1}\boldsymbol{\alpha}\right\|_\infty \le \frac{\lambda}{2}\left\|\mathbf{x}^*_{\mathcal{Q}_1}\right\|_1.$$

This completes the proof. □

Theorem 6: Suppose $\delta_{(k+t\varrho)} < 1/2$, then at the tth iteration, the following inequality holds:

$$\frac{L}{2}\mathbf{x}^{*}_{\mathcal{Q}}{}^\top\mathbf{z}^{t+1}_{\mathcal{Q}} \ge \frac{1}{2}\left(1 - \frac{1}{C}\right)f(\mathbf{x}^t).$$

Proof: Let $\boldsymbol{\zeta} = \partial\|\mathbf{x}^t\|_1$ be the subgradient of $\|\mathbf{x}^t\|_1$ and $\boldsymbol{\alpha} = (\mathbf{A}_{\mathcal{I}_t}\mathbf{x}^t_{\mathcal{I}_t} - \mathbf{b})$. In addition, $\mathbf{x}^t$ must satisfy the following approximate optimality condition:

$$G_{\mathcal{I}_t}(\mathbf{x}^t) = \mathbf{A}^\top_{\mathcal{I}_t}\boldsymbol{\alpha} + \lambda\boldsymbol{\zeta}_{\mathcal{I}_t} = \boldsymbol{\epsilon}_{\mathcal{I}_t}, \text{ and } \boldsymbol{\epsilon}_{\mathcal{I}^c_t} = \mathbf{0}.$$

It follows that $(\mathbf{A}\mathbf{x}^*)^\top\boldsymbol{\alpha} = (\mathbf{A}_{\mathcal{I}^*}\mathbf{x}^*_{\mathcal{I}^*})^\top\boldsymbol{\alpha}$ and

$$\begin{aligned}&-\frac{1}{2}\left(\mathbf{A}_{\mathcal{I}^*}\mathbf{x}^*_{\mathcal{I}^*}\right)^\top\boldsymbol{\alpha}\\ &= -\frac{1}{2}\left[\mathbf{x}^{*}_{\mathcal{Q}}{}^\top\right]\left[\mathbf{A}^\top_{\mathcal{Q}}\boldsymbol{\alpha}\right] + \frac{\lambda}{2}\mathbf{x}^{*}_{\mathcal{I}_t}{}^\top\boldsymbol{\zeta}_{\mathcal{I}_t}\\ &\quad - \frac{1}{2}\mathbf{x}^{*}_{\mathcal{I}_t}{}^\top\boldsymbol{\epsilon}_{\mathcal{I}_t}\\ &= -\frac{1}{2}\mathbf{x}^{*}_{\mathcal{Q}_1}{}^\top\mathbf{A}^\top_{\mathcal{Q}_1}\boldsymbol{\alpha} + \frac{\lambda}{2}\varsigma^\top_{\mathcal{Q}_2}\mathbf{x}^*_{\mathcal{Q}_2}\\ &\quad + \frac{L}{2}\mathbf{x}^{*}_{\mathcal{Q}}{}^\top\mathbf{z}^{t+1}_{\mathcal{Q}} + \frac{\lambda}{2}\mathbf{x}^{*}_{\mathcal{I}_t}{}^\top\boldsymbol{\zeta}_{\mathcal{I}_t}\end{aligned}$$

$$
\begin{aligned}
&- \frac{1}{2}{\mathbf{x}^*_{\mathcal{I}_t}}^\top \boldsymbol{\epsilon}_{\mathcal{I}_t} \ \text{(with (36))}\\
\leq\ & \frac{\lambda}{2}\left\|\mathbf{x}^*_{\mathcal{Q}_1}\right\|_1 + \frac{L}{2}{\mathbf{x}^*_{\mathcal{Q}}}^\top \mathbf{z}^{t+1}_{\mathcal{Q}}\\
&- \frac{1}{2}{\mathbf{x}^*_{\mathcal{I}_t}}^\top \boldsymbol{\epsilon}_{\mathcal{I}_t} + \frac{\lambda}{2}{\mathbf{x}^*_{\mathcal{I}_t}}^\top \boldsymbol{\zeta}_{\mathcal{I}_t}.\\
&+ \frac{\lambda}{2}\varsigma^\top_{\mathcal{Q}_2}\mathbf{x}^*_{\mathcal{Q}_2} \ \text{(with (37))}\\
\leq\ & \frac{\lambda}{2}\|\mathbf{x}^*\|_1 + \frac{L}{2}{\mathbf{x}^*_{\mathcal{Q}}}^\top \mathbf{z}^{t+1}_{\mathcal{Q}} - \frac{1}{2}{\mathbf{x}^*_{\mathcal{I}_t}}^\top \boldsymbol{\epsilon}_{\mathcal{I}_t}.
\end{aligned} \tag{38}
$$

For convenience, we drop the superscript t of $\mathbf{x}^t_{\mathcal{I}_t}$ hereafter. Based on the definition of the subgradient, we have $\lambda\|\mathbf{x}_{\mathcal{I}_t}\|_1 - \lambda\mathbf{x}'_{\mathcal{I}_t}\boldsymbol{\zeta}_{\mathcal{I}_t} = \mathbf{0}$, and it follows that

$$
\begin{aligned}
f(\mathbf{x}^t) &= \frac{1}{2}\left\|\mathbf{A}_{\mathcal{I}_t}\mathbf{x}_{\mathcal{I}_t} - \mathbf{b}\right\|^2 + \lambda\left\|\mathbf{x}_{\mathcal{I}_t}\right\|_1\\
&= \frac{1}{2}\left(\mathbf{A}_{\mathcal{I}_t}\mathbf{x}_{\mathcal{I}_t}\right)^\top \boldsymbol{\alpha} - \frac{1}{2}\mathbf{b}^\top\boldsymbol{\alpha} + \lambda\left\|\mathbf{x}_{\mathcal{I}_t}\right\|_1\\
&= \frac{1}{2}\boldsymbol{\epsilon}^\top_{\mathcal{I}_t}\mathbf{x}_{\mathcal{I}_t} - \frac{1}{2}\left(\mathbf{A}\mathbf{x}^*\right)^\top\boldsymbol{\alpha} - \frac{1}{2}\mathbf{e}^\top\boldsymbol{\alpha} + \frac{\lambda}{2}\left\|\mathbf{x}_{\mathcal{I}_t}\right\|_1\\
&\leq \frac{1}{2}\boldsymbol{\epsilon}^\top_{\mathcal{I}_t}\mathbf{x}_{\mathcal{I}_t} + \frac{\lambda}{2}\|\mathbf{x}^*\|_1 + \frac{L}{2}{\mathbf{x}^*_{\mathcal{Q}}}^\top\mathbf{z}^{t+1}_{\mathcal{Q}} - \frac{1}{2}{\mathbf{x}^*_{\mathcal{I}_t}}^\top\boldsymbol{\epsilon}_{\mathcal{I}_t}\\
&\quad + \frac{\lambda}{2}\left\|\mathbf{x}_{\mathcal{I}_t}\right\|_1 - \frac{1}{2}\mathbf{e}^\top\boldsymbol{\alpha}. \ \text{(with (38))}
\end{aligned}
$$

For convenience, let $q(\boldsymbol{\alpha}) = \frac{1}{2}\mathbf{e}^\top\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\epsilon}^\top_{\mathcal{I}_t}\mathbf{x}_{\mathcal{I}_t} + \frac{1}{2}{\mathbf{x}^*_{\mathcal{I}_t}}^\top\boldsymbol{\epsilon}_{\mathcal{I}_t}$. By assumption, $f(\mathbf{x}^t) \geq Cf(\mathbf{x}^*)$ where $C > 1$, then we have:

$$
\begin{aligned}
&\frac{L}{2}{\mathbf{x}^*_{\mathcal{Q}}}^\top\mathbf{z}^{t+1}_{\mathcal{Q}}\\
&\geq f(\mathbf{x}^t) - \frac{\lambda}{2}\|\mathbf{x}^*\|_1 - \frac{\lambda}{2}\left\|\mathbf{x}_{\mathcal{I}_t}\right\|_1 + q(\boldsymbol{\alpha})\\
&= f(\mathbf{x}^t) - \frac{1}{2}\left(f(\mathbf{x}^*) - \frac{1}{2}\|\mathbf{e}\|^2\right)\\
&\quad - \frac{1}{2}\left(f(\mathbf{x}^t) - \frac{1}{2}\|\boldsymbol{\alpha}\|^2\right) + q(\boldsymbol{\alpha})\\
&\geq f(\mathbf{x}^t) - \frac{1}{2}\left(\frac{f(\mathbf{x}^t)}{C} - \frac{1}{2}\|\mathbf{e}\|^2\right)\\
&\quad - \frac{1}{2}\left(f(\mathbf{x}^t) - \frac{1}{2}\|\boldsymbol{\alpha}\|^2\right) + q(\boldsymbol{\alpha})\\
&= \frac{1}{2}\left(1 - \frac{1}{C}\right)f(\mathbf{x}^t) + \frac{1}{4}\|\mathbf{e}\|^2 + \frac{1}{4}\|\boldsymbol{\alpha}\|^2 + q(\boldsymbol{\alpha})\\
&\geq \frac{1}{2}\left(1 - \frac{1}{C}\right)f(\mathbf{x}^t) - \frac{1}{2}\boldsymbol{\epsilon}^\top_{\mathcal{I}_t}\mathbf{x}_{\mathcal{I}_t} + \frac{1}{2}{\mathbf{x}^*_{\mathcal{I}_t}}^\top\boldsymbol{\epsilon}_{\mathcal{I}_t},
\end{aligned}
$$

where the last inequality holds because $\frac{1}{2}\mathbf{e}^\top\boldsymbol{\alpha} + \frac{1}{4}\|\mathbf{e}\|^2 + \frac{1}{4}\|\boldsymbol{\alpha}\|^2 = (\frac{1}{2}\mathbf{e} + \frac{1}{2}\boldsymbol{\alpha})^\top(\frac{1}{2}\mathbf{e} + \frac{1}{2}\boldsymbol{\alpha}) \geq 0$. Furthermore, suppose $\|\boldsymbol{\epsilon}\|$ is small enough such that $|(\mathbf{x}^*_{\mathcal{I}_t} - \mathbf{x}^t_{\mathcal{I}_t})^\top\boldsymbol{\epsilon}_{\mathcal{I}_t}| \leq \vartheta f(\mathbf{x}^t)$ for a small $\vartheta > 0$, then it follows that:

$$
\frac{L}{2}{\mathbf{x}^*_{\mathcal{Q}}}^\top\mathbf{z}^t_{\mathcal{Q}} \geq \frac{1}{2}\left(1 - \frac{1}{C} - \vartheta\right)f(\mathbf{x}^t).
$$

If $\vartheta \ll \frac{1}{C}$, we can simplify the formulation by absorbing ϑ into C and get

$$
\frac{L}{2}{\mathbf{x}^*_{\mathcal{Q}}}^\top\mathbf{z}^{t+1}_{\mathcal{Q}} \geq \frac{1}{2}\left(1 - \frac{1}{\widehat{C}}\right)f(\mathbf{x}^t).
$$

By letting $C = \widehat{C}$, we can complete the proof. □

Theorem 7: Suppose $\delta_{(k+t\varrho)} < 1/2$, then at the tth iteration, we have $\|\mathbf{z}^t_{\mathcal{Q}}\|^2 \geq \frac{\chi}{2L^2}(1 - \frac{1}{C})^2 f(\mathbf{x}^t)$.

Proof: From Theorem 6, we have

$$
\frac{L}{2}\|\mathbf{x}^*_{\mathcal{Q}}\|\|\mathbf{z}^{t+1}_{\mathcal{Q}}\| \geq \frac{L}{2}{\mathbf{x}^*_{\mathcal{Q}}}^\top\mathbf{z}^{t+1}_{\mathcal{Q}} \geq \frac{1}{2}\left(1 - \frac{1}{C}\right)f(\mathbf{x}^t).
$$

With Lemma 4, it follows that

$$
\begin{aligned}
\|\mathbf{z}^{t+1}_{\mathcal{Q}}\|^2 &\geq \frac{1}{L^2\|\mathbf{x}^*_{\mathcal{Q}}\|^2}\left(1 - \frac{1}{C}\right)^2 f(\mathbf{x}^t)^2\\
&\geq \frac{\chi}{2L^2}\left(1 - \frac{1}{C}\right)^2 f(\mathbf{x}^t).
\end{aligned}
$$

□

Theorem 8: Let $\delta_{(k+t\varrho)} < \frac{1}{2}$, we have $\|\mathbf{y}^{t+1}\|^2 > cf(\mathbf{x}^t)$ where $c = \frac{\chi\varrho}{2kL^2}(1 - \frac{1}{C})^2$.

Proof: By definition, $\mathcal{J}_t = \mathcal{I}_{t+1}\backslash\, \mathcal{I}_t$. If $|\mathcal{J}_t| = \varrho \leq |\mathcal{Q}|$, by definition of $\mathbf{y}^{t+1}$, we have $\frac{\|\mathbf{y}^{t+1}_{\mathcal{J}_t}\|^2}{\varrho} \geq \frac{\|\mathbf{z}^{t+1}_{\mathcal{Q}}\|^2}{|\mathcal{Q}|}$, which means

$$
\|\mathbf{y}^{t+1}_{\mathcal{J}_t}\|^2 \geq \frac{\chi\varrho}{2kL^2}\left(1 - \frac{1}{C}\right)^2 f(\mathbf{x}^t).
$$

Otherwise, if $|\mathcal{J}_t| = \varrho > |\mathcal{Q}|$, and we have $\|\mathbf{y}^{t+1}_{\mathcal{J}_t}\| \geq \|\mathbf{z}^{t+1}_{\mathcal{Q}}\|$. Therefore, the following inequality holds:

$$
\|\mathbf{y}^{t+1}_{\mathcal{J}_t}\|^2 \geq \frac{\chi}{2L^2}\left(1 - \frac{1}{C}\right)^2 f(\mathbf{x}^t).
$$

In summary, since $\varrho < k$, we have $\|\mathbf{y}^{t+1}_{\mathcal{J}_t}\|^2 \geq cf(\mathbf{x}^t)$, where $c = \frac{\chi\varrho}{2kL^2}(1 - \frac{1}{C})^2$. □

With Lemma 5, we have

$$
\begin{aligned}
f(\mathbf{x}^{t+1}) &\leq f(\mathbf{x}^t) - \frac{L}{2}\|\mathbf{y}_{\mathcal{J}_{t+1}}\|^2 = f(\mathbf{x}^t) - \frac{L}{2}\|\mathbf{y}_{\mathcal{F}_t}\|^2\\
&\leq \left(1 - \frac{\chi\varrho}{4kL}\left(1 - \frac{1}{C}\right)^2\right)f(\mathbf{x}^t).
\end{aligned}
$$

Let $1/L_t$ be the step size obtained by the line search method at the tth iteration, there should exist a ζ and $1/\zeta$ $= \min\{1/L_1, \ldots, 1/L_\iota\}$ such that the above relation holds for each $t < \iota$, where $\delta_{(k+\iota\varrho)} < \frac{1}{2}$. Note that $\frac{(1-2\delta_{(k+t\varrho)})^2}{2(1-\delta_{(k+t\varrho)})}$ is decreasing w.r.t. $\delta_{(k+t\varrho)}$ within $\delta_{(k+t\varrho)} \in (0, \frac{1}{2})$. In addition, $\delta_{(k+t\varrho)} \leq \delta_{(k+\iota\varrho)}$ for all $t \leq \iota$ [23]. Therefore, if $\delta_{(k+\iota\varrho)} < 1/2$ and $t < \iota$, the following relation holds:

$$
f(\mathbf{x}^{t+1}) \leq \left(1 - \frac{\chi\varrho}{4k\zeta}\left(1 - \frac{1}{C}\right)^2\right)f(\mathbf{x}^t).
$$

Letting $\nu = (1 - \frac{\chi\varrho}{4k\zeta}(1 - \frac{1}{C})^2)$, we complete the proof.

Appendix F
Proof of Theorem 4

We study the case when $\lambda > 0$ is arbitrarily small or $\lambda = 0$. Let $\widehat{\mathbf{x}}$ and $\widehat{\mathbf{e}}$ denote the ground-truth sparse signal and measurement noise, respectively, that is, $\mathbf{b} = \mathbf{A}\widehat{\mathbf{x}} + \widehat{\mathbf{e}}$. Let $\widehat{\mathcal{I}}$ be the support of $\widehat{\mathbf{x}}$ and $\|\widehat{\mathbf{x}}\|_0 \leq k$. In addition, assume that $f(\mathbf{x}^t) \geq Cf(\widehat{\mathbf{x}}) = C(\frac{1}{2}\|\widehat{\mathbf{e}}\|^2 + \lambda\|\widehat{\mathbf{x}}\|_1)$, where $C > 1$. By replacing

$\mathbf{x}^*$ with $\widehat{\mathbf{x}}$ and $f(\mathbf{x}^*)$ with $f(\widehat{\mathbf{x}})$, we can complete the proof by adapting the proof of Theorem 3.

Appendix G
Proof of Theorem 5

The proof parallels the results in [23]. Since the parameter D in the **Restricted Set** should be taken into consideration, some results and proofs are slightly different from those in [23]. For completeness, we present the detailed proof. First, we list the definitions and lemmas needed in the proof.

Definition 3 (Restricted Orthogonality Constant) [23]: If $k + k' \leq m$, the k, k'-restricted orthogonality constant $\theta_{k,k'}$ is the smallest number such that

$$|\langle \mathbf{Ay}, \mathbf{Ay}' \rangle| \leq \theta_{k,k'} \|\mathbf{y}\| \cdot \|\mathbf{y}'\|$$

for all $\mathbf{y}$ and $\mathbf{y}'$ of $\|\mathbf{y}\|_0 \leq k$ and $\|\mathbf{y}\|_0 \leq k'$ respectively, and have disjoint supports.

Lemma 7: The following monotone relation regarding RIP constant and ROC constant follows [23]:

$$\begin{aligned} &\sigma_k \leq \sigma_{k_1}, \text{ if } k \leq k_1 \leq m \\ &\theta_{k,k_1} \leq \theta_{k',k'_1}, \text{ if } k \leq k', k_1 \leq k'_1 \textit{ and } k + k'_1 \leq m. \end{aligned} \tag{39}$$

Lemma 8 (Square Root Lifting Inequality) [23]: For any $a \geq 1$ and positive integers k, k_1 such that ak_1 is an integer and it follows that $\theta_{k,ak_1} \leq \sqrt{a}\theta_{k,k_1}$.

Notations. For convenience, we will use the following notations. Let $\widehat{\mathbf{x}}$ denote the ground-truth k-sparse signal and $\mathbf{h} = \widehat{\mathbf{x}} - \mathbf{x}^*$. Without loss of generality, assume that $|h_1| \geq |h_2| \geq \ldots, |h_m|$. In addition, given two integers k_1 and k_2 such that $k_1 \geq k$ and $4(1+D)(k_1 - k) \leq k_2$ [23]. Furthermore, partition $\{1, \ldots, m\}$ into the following sets: $\mathcal{S}_0 = \{1, 2, \ldots, k_1\}, \mathcal{S}_1 = \{k_1 + 1, \ldots, k_1 + k_2\}, \mathcal{S}_2 = \{k_1 + k_2 + 1, \ldots, k_1 + 2k_2\}, \ldots$.

Lemma 9 [23]: For any vector $\mathbf{h} \in \mathbb{R}^m$, it satisfies

$$\|\mathbf{h}\|_2 - \frac{\|\mathbf{h}\|_1}{\sqrt{m}} \leq \frac{\sqrt{m}}{4} \left(\max_{1 \leq i \leq m} |h_i| - \min_{1 \leq i \leq m} |h_i| \right).$$

Lemma 10: Let $\mathbf{h}$ be defined above, if $4(1+D)(k_1 - k) \leq k_2$ we have

$$\sum_{i \geq 1} \|\mathbf{h}_{\mathcal{S}_i}\|_2 \leq t \|\mathbf{h}_{\mathcal{S}_0}\|_2,$$

where $t = D\sqrt{\frac{k_1}{k_2}} + \frac{1}{4}\sqrt{\frac{k_2}{k_1}} - \frac{(1+D)(k_1-k)}{\sqrt{k_1 k_2}}$ and $D = \frac{1+\epsilon}{1-\epsilon}$.

Proof: At first, from Lemma 2, we have $\|\mathbf{h}_{\mathcal{T}^c}\|_1 \leq D\|\mathbf{h}_{\mathcal{T}}\|_1$. Since $k_1 \geq k$ and $\mathcal{S}_0 \supseteq \mathcal{T}$, then we have

$$\|\mathbf{h}_{\mathcal{T}}\|_1 \leq \|\mathbf{h}_{\mathcal{S}_0}\|_1 \text{ and } \|\mathbf{h}_{\mathcal{T}}\|_1 \leq \|\mathbf{h}_{\mathcal{S}_0}\|_1 - (k_1 - k)|h_{(k_1+1)}|.$$

and $\|\mathbf{h}_{\mathcal{T}^c}\|_1 \geq \|\mathbf{h}_{\mathcal{S}_0^c}\|_1 = \sum_{i \geq 1} \|\mathbf{h}_{\mathcal{S}_i}\|_1$. By Lemma 9,

$$\begin{aligned}
&\sum_{i \geq 1} \|\mathbf{h}_{\mathcal{S}_i}\|_2 \\
&\quad \leq \sum_{i \geq 1} \frac{\|\mathbf{h}_{\mathcal{S}_i}\|_1}{\sqrt{k_2}} + \frac{\sqrt{k_2}}{4} \big(|h_{(k_1+1)}| - |h_{(k_1+k_2)}| \\
&\qquad + |h_{(k_1+k_2+1)}| - |h_{(k_1+2k_2)}| + \cdots \big) \text{ (by Lemma 9)} \\
&\quad \leq \frac{1}{\sqrt{k_2}} \left(\|\mathbf{h}_{\mathcal{T}^c}\|_1 - (k_1 - k)|h_{(k_1+1)}| \right) \\
&\qquad + \frac{\sqrt{k_2}}{4} (|h_{(k_1+1)}|) \text{ (by } \|\mathbf{h}_{\mathcal{T}^c}\|_1 \geq \|\mathbf{h}_{\mathcal{S}_0^c}\|_1 = \sum_{i \geq 1} \|\mathbf{h}_{\mathcal{S}_i}\|_1) \\
&\quad \leq \frac{1}{\sqrt{k_2}} \left(D\|\mathbf{h}_{\mathcal{T}}\|_1 - (k_1 - k)|h_{(k_1+1)}| \right) \\
&\qquad + \frac{\sqrt{k_2}}{4} (|h_{(k_1+1)}|) \text{ (with } \|\mathbf{h}_{\mathcal{T}^c}\|_1 \leq D\|\mathbf{h}_{\mathcal{T}}\|_1) \\
&\quad \leq \frac{1}{\sqrt{k_2}} \left(D\|\mathbf{h}_{\mathcal{S}_0}\|_1 - (1+D)(k_1 - k)|h_{(k_1+1)}| \right) \\
&\qquad + \frac{\sqrt{k_2}}{4} (|h_{(k_1+1)}|) \\
&\text{(with } \|\mathbf{h}_{\mathcal{T}}\|_1 \leq \|\mathbf{h}_{\mathcal{S}_0}\|_1 - (k_1 - k)|h_{(k_1+1)}|) \\
&\quad \leq \frac{1}{\sqrt{k_2}} \left(\sqrt{k_1} D\|\mathbf{h}_{\mathcal{S}_0}\| - (1+D)(k_1 - k)|h_{(k_1+1)}| \right) \\
&\qquad + \frac{\sqrt{k_2}}{4} (|h_{(k_1+1)}|) \\
&\quad \leq \frac{D\sqrt{k_1}}{\sqrt{k_2}} \|\mathbf{h}_{\mathcal{S}_0}\| - \frac{(1+D)(k_1 - k)|h_{(k_1+1)}|}{\sqrt{k_2}} \\
&\qquad + \frac{\sqrt{k_2}}{4} (|h_{(k_1+1)}|) \\
&\quad \leq \frac{D\sqrt{k_1}}{\sqrt{k_2}} \|\mathbf{h}_{\mathcal{S}_0}\| + \left(\frac{\sqrt{k_2}}{4\sqrt{k_1}} - \frac{(1+D)(k_1 - k)}{\sqrt{k_1 k_2}} \right) \|\mathbf{h}_{\mathcal{S}_0}\| \\
&\qquad \text{(by } 4(1+D)(k_1 - k) \leq k_2) .
\end{aligned}$$

□

Lemma 11: Suppose $\|\mathbf{A}^\top(\mathbf{A}\widehat{\mathbf{x}} - \mathbf{b})\|_\infty = \|\mathbf{A}^\top \widehat{\mathbf{e}}\|_\infty \leq \epsilon\lambda$. Let $\mathbf{h} = \widehat{\mathbf{x}} - \mathbf{x}^*$, and we have

$$\|\mathbf{h}\| \leq \frac{2\sqrt{2}\sqrt{k_1}\lambda}{1 - \sigma_{k_1} - t\theta_{k_1,k_2}}.$$

Proof: At first, we have

$$\begin{aligned}
|\langle \mathbf{Ah}, \mathbf{Ah}_{\mathcal{S}_0} \rangle| &= \left| \langle \mathbf{A} \left(\sum_{i \geq 0} \mathbf{h}_{\mathcal{S}_i} \right), \mathbf{Ah}_{\mathcal{S}_0} \rangle \right| \\
&= \left| \langle \mathbf{Ah}_{\mathcal{S}_0}, \mathbf{Ah}_{\mathcal{S}_0} \rangle + \sum_{i \geq 1} \langle \mathbf{A}(\mathbf{h}_{\mathcal{S}_i}), \mathbf{Ah}_{\mathcal{S}_0} \rangle \right| \\
&\geq \left(1 - \sigma_{k_1} \|\mathbf{h}_{\mathcal{S}_0}\|^2 \right) - \theta_{k_1,k_2} \|\mathbf{h}_{\mathcal{S}_0}\| \sum_{i \geq 1} \|\mathbf{h}_{\mathcal{S}_i}\| \\
&\geq \left(1 - \sigma_{k_1} - t\theta_{k_1,k_2} \right) \|\mathbf{h}_{\mathcal{S}_0}\|^2
\end{aligned}$$

At the optimality, we have $\|\mathbf{A}^\top(\mathbf{Ax}^* - \mathbf{b})\|_\infty \leq \lambda$, which implies $\|\mathbf{A}_{\mathcal{S}_0}^\top(\mathbf{Ax}^* - \mathbf{b})\| \leq \sqrt{k_1}\lambda$. Since $k \leq k_1$, and $\|\mathbf{A}^\top(\mathbf{A}\widehat{\mathbf{x}} - \mathbf{b})\|_\infty = \|\mathbf{A}^\top \widehat{\mathbf{e}}\|_\infty \leq \epsilon\lambda$, we have $\|\mathbf{A}_{\mathcal{S}_0}^\top(\mathbf{A}\widehat{\mathbf{x}} - \mathbf{b})\| \leq \sqrt{k_1}\lambda$. Therefore, we have

$$\begin{aligned}
\|\mathbf{A}_{\mathcal{S}_0}^\top \mathbf{Ah}\| &= \|\mathbf{A}_{\mathcal{S}_0}^\top \left(\mathbf{Ax}^* - \mathbf{A}\widehat{\mathbf{x}} \right)\| \\
&\leq \|\mathbf{A}_{\mathcal{S}_0}^\top \left(\mathbf{Ax}^* - \mathbf{b} + \mathbf{b} - \mathbf{A}\widehat{\mathbf{x}} \right)\| \\
&\leq \|\mathbf{A}_{\mathcal{S}_0}^\top (\mathbf{Ax}^* - \mathbf{b})\| + \|\mathbf{A}_{\mathcal{S}_0}^\top (\mathbf{A}\widehat{\mathbf{x}} - \mathbf{b})\| \\
&\leq 2\sqrt{k_1}\lambda.
\end{aligned}$$

Since $|\mathcal{S}_0| = k_1$, we have

$$\|\mathbf{h}\| \leq \sqrt{2}\|\mathbf{h}_{\mathcal{S}_0}\|$$

$$\begin{aligned}
&\leq \frac{\sqrt{2}|\langle \mathbf{Ah}, \mathbf{Ah}_{\mathcal{S}_0}\rangle|}{(1-\sigma_{k_1}-t\theta_{k_1,k_2})\,\|\mathbf{h}_{\mathcal{S}_0}\|}\\
&= \frac{\sqrt{2}|\langle \mathbf{Ah}, \mathbf{A}_{\mathcal{S}_0}\mathbf{h}_{\mathcal{S}_0}\rangle|}{(1-\sigma_{k_1}-t\theta_{k_1,k_2})\,\|\mathbf{h}_{\mathcal{S}_0}\|}\\
&= \frac{\sqrt{2}|\langle \mathbf{A}_{\mathcal{S}_0}^\top\mathbf{Ah}, \mathbf{h}_{\mathcal{S}_0}\rangle|}{(1-\sigma_{k_1}-t\theta_{k_1,k_2})\,\|\mathbf{h}_{\mathcal{S}_0}\|}\\
&\leq \frac{\sqrt{2}\|\mathbf{A}_{\mathcal{S}_0}^\top\mathbf{Ah}\|\cdot\|\mathbf{h}_{\mathcal{S}_0}\|}{(1-\sigma_{k_1}-t\theta_{k_1,k_2})\,\|\mathbf{h}_{\mathcal{S}_0}\|}\\
&\leq \frac{2\sqrt{2}\sqrt{k_1}\lambda\|\mathbf{h}_{\mathcal{S}_0}\|}{(1-\sigma_{k_1}-t\theta_{k_1,k_2})\,\|\mathbf{h}_{\mathcal{S}_0}\|}\\
&= \frac{2\sqrt{2}\sqrt{k_1}\lambda}{(1-\sigma_{k_1}-t\theta_{k_1,k_2})}.
\end{aligned}$$

This completes the proof. □

Let $k_1 = k, 1 \leq k_2 < k$, and we have

$$\begin{aligned}
t &= D\sqrt{\frac{k_1}{k_2}} + \frac{1}{4}\sqrt{\frac{k_2}{k_1}} - \frac{(1+D)(k_1-k)}{\sqrt{k_1k_2}}.\\
&= D\sqrt{\frac{k}{k_2}} + \frac{1}{4}\sqrt{\frac{k_2}{k}}.
\end{aligned}$$

Note that the square root lifting inequality tells that

$$\sigma_k + t\theta_{k,k_2} \leq \left(1 + t\sqrt{\frac{k}{k-k_2}}\right)\sigma_k.$$

Let $P_k = (1 + t\sqrt{\frac{k}{k-k_2}})$. If $1 > \sigma_{k_1} + t\theta_{k_1,k_2}$, we have

$$\|\mathbf{h}\| \leq \frac{2\sqrt{2}\sqrt{k}\lambda}{(1-\sigma_k - t\theta_{k,k_2})} \leq \frac{2\sqrt{2}\sqrt{k}\lambda}{(1-P_k\sigma_k)}. \tag{40}$$

Define $f(x) = (1 + (\frac{D}{\sqrt{x}} + \frac{\sqrt{x}}{4})\sqrt{\frac{1}{1-x}})$, where $x \in (0,1)$. Then $P_k = f(\frac{k_2}{k})$. Let $x = \frac{k_2}{k}$, then it follows that

$$\|\mathbf{h}\| \leq \frac{2\sqrt{2}\sqrt{k}\lambda}{(1-f(x)\sigma_k)}. \tag{41}$$

Let $e = \frac{(D-1)}{(8D+1)}$. Since D can be arbitrarily close to 1, we have $0 < e < 0.1250$ for $D > 1$. Moreover, it follows that $\frac{4D}{8D+1} = \frac{4}{9} + \frac{4(D-1)}{9(8D+1)} = \frac{4}{9} + \frac{4e}{9}$. Let r_s be an integer such that $r_s = 4k(\text{mod } 9)$. Now, k_2 can be specified by

$$k_2 = \begin{cases} \left\lfloor \dfrac{(4+4e)k}{9} \right\rfloor, & \text{if } r_s \leq 4 \\ \left\lceil \dfrac{(4+4e)k}{9} \right\rceil, & \text{if } r_s > 4. \end{cases}$$

Since $f'(x) = \frac{(8D+1)x-4D}{8(x-x^2)^{\frac{3}{2}}}$, $f(x)$ is increasing when $\frac{4D}{8D+1} \leq x < 1$ and decreasing when $0 < x < \frac{4D}{8D+1}$. By the definition of k_2 in (42), $P_k \leq \max(f(\frac{4+4e}{9} + \frac{4+4e}{9k}), f(\frac{4+4e}{9} - \frac{4+4e}{9k}))$.

Now we show that if $k \geq 7$, $f(\frac{4+4e}{9} + \frac{4+4e}{9k}) < f(\frac{4+4e}{9} - \frac{4+4e}{9k})$ for $e \in (0, 0.1250)$. Actually, note that $\frac{f(\frac{4+4e}{9}+\frac{4+4e}{9k})}{f(\frac{4+4e}{9}-\frac{4+4e}{9k})} = \frac{\sqrt{6}(71-64e)\sqrt{39-24e}}{\sqrt{8}(69-48e)(\sqrt{31-32e})}$. Let $g_1(e) = (\sqrt{6}(71-64e)\sqrt{39-24e})^2$ and $g_2(e) = (\sqrt{8}(69-48e)(\sqrt{31-32e}))^2$. Then we have $g_1(e) - g_2(e) = 4536e - 567$. Easily, $g_1(e) - g_2(e) < 0$ if $e \in (0, 0.1250)$.

Moreover, let $g(e) = f(\frac{4+4e}{9} - \frac{4+4e}{9\times 7}) = 1 + (\frac{63}{\sqrt{24}}\frac{1}{1-8e} + \frac{\sqrt{24}}{4})\frac{\sqrt{1+e}}{\sqrt{39-24e}}$. As $g(e)$ is smooth and increasing in $e \in (0, 0.1250)$, we have $g(e) = g(0) + c$, where $c \in (0, +\infty)$. Since $P_k \leq g(e)$, to make bound (40) valid, we shall have

$$1 > g(e)\sigma_k = (g(0)+c)\sigma_k \geq P_k\sigma_k.$$

Inductively, we have

$$\sigma_k < \frac{1}{g(e)} = \frac{1}{g(0)+c} = \frac{1}{g(0)} - \vartheta. \tag{42}$$

Easily, we have

$$\vartheta = \frac{1}{g(0)} - \frac{1}{1 + \left(\frac{63}{\sqrt{24}}\frac{1}{1-8e} + \frac{\sqrt{24}}{4}\right)\frac{\sqrt{1+e}}{\sqrt{39-24e}}}, \tag{43}$$

where $e = \frac{(D-1)}{(8D+1)}$ and $D = \frac{1+\epsilon}{1-\epsilon}$. Since $g(0) < 3.256$ and $g(e)$ is increasing in $e \in (0, 0.125)$, we can choose $\sigma_k < 0.307 - \vartheta$. Obviously, $(0, 0.307-\vartheta) \subset (0, 1/g(0) - \vartheta)$ for $\vartheta \leq 0.307$.

For $k = 4, 6$, we can choose $k_2 = 2, 3$, respectively. Thus $P_4 = P_6 = f(0.5) = 1.25 + 2D$, which implies $\sigma_k < \frac{1}{1.25+2D}$. With the same $D > 1$, we have $\{\sigma_k | \sigma_k < \frac{1}{1.25+2D}\} \supset \{\sigma_k | \sigma_k < 0.307 - \vartheta\}$.

For $k = 5$, we can choose $k_2 = 2$, which implies $P_5 = f(0.4) = 1.2041 + 2.0412D$. Similarly, with the same $D > 1$, we have $\{\sigma_k | \sigma_k < \frac{1}{1.2041+2.0412D}\} \supset \{\sigma_k | \sigma_k < 0.307 - \vartheta\}$. For $k = 2, 3$, we can choose $k_2 = 1$, and $t = \sqrt{k}$ [23], which implies $P_2 = 3$ and $P_3 < 3.1214$. Apparently, we can still choose $\sigma_k \in \{\sigma_k | \sigma_k < 0.307 - \vartheta\}$.

This completes the proof.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their insightful comments and suggestions which have greatly improved the paper.

REFERENCES

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

[2] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach.*, vol. 31, no. 2, pp. 210–227, Apr. 2009.

[3] T. Peleg, Y. Eldar, and M. Elad, "Exploiting statistical dependencies in sparse representations for signal recovery," *IEEE Trans. Signal Process.*, vol. 60, no. 5, pp. 2286–2303, May 2012.

[4] , Y. C. Eldar and G. Kutyniok, Eds., *Compressed Sensing: Theory and Applications*. Cambridge, U.K.: Cambridge Univ. Press, 2012.

[5] E. J. Candès and T. Tao, "Decoding by linear programming," *IEEE Trans. Inf. Theory*, vol. 51, no. 12, pp. 4203–4215, Dec. 2005.

[6] M. F. Duarte and Y. C. Eldar, "Structured compressed sensing: From theory to applications," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4053–4085, Sep. 2011.

[7] T. T. Do, L. Gan, N. H. Nguyen, and T. D. Tran, "Fast and efficient compressive sensing using structurally random matrices," *IEEE Trans. Signal Process.*, vol. 60, no. 1, pp. 139–154, Jan. 2012.

[8] J. Mairal, M. Elad, and G. Sapiro, "Sparse representation for color image restoration," *IEEE Trans. Image Process.*, vol. 17, no. 1, pp. 53–69, Jan. 2008.

[9] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Non-local sparse models for image restoration," presented at the ICCV, 2009.

[10] J. Yang, J. Wright, T. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861–2873, Nov. 2010.

[11] D. Xu, Y. Huang, Z. Zeng, and X. Xu, "Human gait recognition using patch distribution feature and locality-constrained group sparse representation," *IEEE Trans. Image Process.*, vol. 21, no. 1, pp. 316–326, Jan. 2012.

[12] E. Elhamifar and R. Vidal, "Sparse subspace clustering," presented at the CVPR, 2009.

[13] A. Adler, M. Elad, and Y. Hel-Or, "Probabilistic subspace clustering via sparse representations," *IEEE Signal Process. Lett.*, vol. 20, no. 1, pp. 63–66, Jan. 2013.

[14] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," presented at the ICML, 2009.

[15] L. Bo, X. Ren, and D. Fox, "Hierarchical matching pursuit for image classification: Architecture and fast algorithms," presented at the NIPS, 2011.

[16] A. Coates and A. Ng, "The importance of encoding versus training with sparse coding and vector quantization," presented at the ICML, 2011.

[17] K. Yu, Y. Lin, and J. Lafferty, "Learning image representations from the pixel level via hierarchical sparse coding," presented at the CVPR, 2011.

[18] G. Davis, S. Mallat, and M. Avellaneda, "Adaptive greedy approximations," *Constr. Approx.*, vol. 13, no. 1, pp. 57–98, 1997.

[19] D. Ge, X. Jiang, and Y. Ye, "A note on the complexity of lp minimization," *Math. Program.*, vol. 129, no. 2, pp. 285–299, 2011.

[20] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization," in *Proc. Nat. cad. Sci.*, 2003, vol. 100, no. 5, pp. 2197–2202.

[21] B. Efron, T. Hastie, L. Johnstone, and R. Tibshirani, "Least angle regression," *Ann. Statist.*, vol. 32, no. 2, pp. 407–499, 2004.

[22] H. Lee, A. Battle, R. Raina, and A. Y. Ng, "Efficient sparse coding algorithms," presented at the NIPS, 2006.

[23] T. T. Cai, L. Wang, and G. Xu, "New bounds for restricted isometry constants," *IEEE Trans. Inf. Theory*, vol. 56, no. 9, pp. 4388–4394, Sep. 2010.

[24] E. J. Candès and T. Tao, "The dantzig selector: Statistical estimation when p is much larger than n," *Ann. Statist.*, vol. 35, no. 6, pp. 2313–2351, 2007.

[25] E. J. Candès, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Commun. Pure Appl. Math*, vol. 59, no. 8, pp. 1207–1223, 2006.

[26] M. R. Osborne, B. Presnell, and B. A. Turlach, "On the lasso and its dual," *J. Comput. Graph. Statist.*, vol. 9, no. 2, pp. 319–337, 2000.

[27] M. A. T. Figueiredo, R. D. Nowak, and S. J. Wright, "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems," *IEEE J. Sel. Top. Signal Proces.: Special Issue Convex Optimiz. Methods Signal Process.*, 2007.

[28] P. J. Bickel, Y. Ritov, and A. B. Tsybakov, "Simultaneous analysis of lasso and dantzig selector," *Ann. Statist.*, vol. 37, no. 4, pp. 1705–1732, 2009.

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

[30] Z. Wen, W. Yin, D. Goldfarb, and Y. Zhang, "A fast algorithm for sparse reconstruction based on shrinkage, subspace optimization, and continuation," *SIAM J. Sci. Comput.*, vol. 32, no. 4, pp. 1832–1857, 2010.

[31] A. Yang, A. Ganesh, Y. Ma, and S. Sastry, "Fast l1-minimization algorithms and an application in robust face recognition: A review," presented at the ICIP, 2010.

[32] L. Xiao and T. Zhang, "A proximal-gradient homotopy method for the l1-regularized least-squares problem," presented at the ICML, 2012.

[33] 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 J. Sel. Topics Signal Process.*, vol. 1, no. 4, pp. 606–617, Aug. 2007.

[34] Y. Nesterov, "Gradient Methods for Minimizing Composite Objective Function," Center for Operations Research and Econometrics (CORE), Catholic Univ. of Louvain (UCL), 2007 [Online]. Available: http://www.ecore.be/DPs/dp_1191313936.pdf

[35] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *J. Statist. Software*, vol. 33, no. 1, p. 1, 2010.

[36] S. Yun and K.-C. Toh, "A coordinate gradient descent method for ℓ_1-regularized convex minimization," *Comput. Optimiz. Appl.*, vol. 48, no. 2, pp. 273–307, 2011.

[37] J. K. Bradley, A. Kyrola, D. Bickson, and C. Guestrin, "Parallel coordinate descent for l_1-regularized loss minimization," presented at the ICML, 2011.

[38] Z. J. Xiang, H. Xu, and P. J. Ramadge, "Learning sparse representations of high dimensional data on large scale dictionaries," in *NIPS*, 2012.

[39] L. Xiao and T. Zhang, "A proximal-gradient homotopy method for the sparse least-squares problem," *SIAM J. Optimiz.*, vol. 23, no. 2, pp. 1062–1091, 2013.

[40] V. Roth and B. Fischer, "The group-lasso for generalized linear models: Uniqueness of solutions and efficient algorithms," presented at the ICML, 2008.

[41] J. Kim and H. Park, "Fast active-set-type algorithms for l1-regularized linear regression," presented at the AISTAT, 2010.

[42] S. Shalev-Shwartz, N. Srebro, and T. Zhang, "Trading accuracy for sparsity in optimization problems with sparsity constraint," *SIAM J. Optimiz.*, vol. 20, pp. 2807–2832, 2010.

[43] M. Kowalski, P. Weiss, A. Gramfort, and S. Anthoine *et al.*, ""Accelerating ISTA With An Active Set Strategy,"," presented at the OPT 2011: 4th Int. Workshop Optimiz Mach. Learn., 2011.

[44] R. Rigamonti, M. A. Brown, and V. Lepetit, "Are sparse representations really relevant for image classification?," presented at the CVPR, 2011.

[45] Q. Shi, A. Eriksson, A. V. D. Hengel, and C. Shen, "Is face recognition really a compressive sensing problem?," presented at the CVPR, 2011.

[46] E. J. Candès and Y. Plan, "Near-ideal model selection by ell_1 minimization," *Ann. Statist.*, vol. 37, no. 5A, pp. 2145–2177, 2007.

[47] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3397–3415, Dec. 1993.

[48] J. A. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Trans. Inf. Theory*, vol. 50, no. 10, pp. 2231–2242, Oct. 2004.

[49] Y. C. Pati, R. Rezaiifar, and P. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Proc. 27th Asilomar Conf. Signals, Syst., Comput.*, 1993, pp. 40–44.

[50] J. Tropp and A. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Inf. Theory*, vol. 53, no. 12, pp. 4655–4666, Dec. 2007.

[51] T. Blumensath and M. E. Davies, "Gradient pursuits," *IEEE Trans. Signal Process.*, vol. 56, no. 6, pp. 2370–2382, Jun. 2008.

[52] T. Zhang, "Sparse recovery with orthogonal matching pursuit under RIP," *IEEE Trans. Inf. Theory*, vol. 57, no. 9, pp. 6215–6221, Sep. 2011.

[53] A. Tewari, P. Ravikumar, and I. S. Dhillon, "Greedy algorithms for structurally constrained high dimensional problems," presented at the NIPS, 2011.

[54] X. Yuan and S. Yan, "Forward basis selection for sparse approximation over dictionary," presented at the AISTATS, 2012.

[55] D. Needell and J. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples," *Appl. Comput. Harmon. Anal.*, vol. 26, no. 3, pp. 301–321, 2009.

[56] W. Dai and O. Milenkovic, "Subspace pursuit for compressive sensing signal reconstruction," *IEEE Trans. Inf. Theory*, vol. 55, no. 5, pp. 2230–2249, May 2009.

[57] T. Blumensath and M. E. Davies, "Iterative hard thresholding for compressed sensing," *Appl. Comput. Harmon. Anal.*, vol. 27, no. 3, pp. 265–274, 2009.

[58] T. Blumensath, "Accelerated iterative hard thresholding," *Signal Process.*, vol. 92, no. 3, pp. 752–756, 2011.

[59] R. Giryes and M. Elad, "RIP-based near-oracle performance guarantees for subspace-pursuit, CoSaMP, and iterative hard-thresholding," *IEEE Trans. Signal Process.*, vol. 60, no. 3, pp. 1465–1468, Mar. 2012.

[60] P. Jain, A. Tewari, and I. S. Dhillon, "Orthogonal matching pursuit with replacement," presented at the NIPS, 2011.

[61] M. Davenport, D. Needell, and M. Wakin, "Signal space CoSaMP for sparse recovery with redundant dictionaries," *IEEE Trans. Inf. Theory*, vol. 59, no. 10, pp. 6820–6829, Oct. 2013.

[62] R. Giryes and M. Elad, "Can we allow linear dependencies in the dictionary in the sparse synthesis framework?," presented at the ICASSP, 2013.

[63] R. Giryes and D. Needell, "Greedy Signal Space Methods for Incoherence and Beyond," arXiv:1309.2676, 2013.

[64] R. Giryes and M. Elad, "OMP with highly coherent dictionaries," presented at the Int. Conf. Sampling Theory Appl. (SAMPTA), 2013.

[65] D. Needell and R. Vershynin, "Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit," *J. Found. Comput. Math.*, vol. 9, no. 3, pp. 317–334, 2009.

[66] R. Maleh, "Efficient Sparse Approximation Methods for Medical Imaging," Ph.D. dissertation, Univ. Michigan, Ann Arbor, MI, USA, 2009.

[67] E. Liu and V. N. Temlyakov, "The orthogonal super greedy algorithm and applications in compressed sensing," *IEEE Trans. Inf. Theory*, vol. 58, no. 4, pp. 2040–2047, Apr. 2012.

[68] T. Blumensath and M. E. Davies, "Stagewise weak gradient pursuits," *IEEE Trans. Signal Process.*, vol. 57, no. 11, pp. 4333–4346, Nov. 2009.

[69] D. L. Donoho, Y. Tsaig, I. Drori, and J. L. Starck, "Sparse solution of underdetermined systems of linear equations by stagewise orthogonal matching pursuit," *IEEE Trans. Inf. Theory*, vol. 58, no. 2, pp. 1094–1121, Feb. 2012.

[70] C. Hegde, P. Indyk, and L. Schmidt, "Approximation-tolerant model-based compressive sensing," presented at the ACM Symp. Discrete Algorithms (SODA), 2014.

[71] R. Giryes, S. Nam, M. Elad, R. Gribonval, and M. E. Davies, "Greedy-like algorithms for the cosparse analysis model," *Linear Algebra Appl.*, vol. 441, pp. 22–60, 2014.

[72] G. Raskutti, M. J. Wainwright, and B. Yu, "Restricted eigenvalue properties for correlated gaussian designs," *JMLR*, vol. 99, pp. 2241–2259, 2010.

[73] L. Wang, "L1 Penalized LAD Estimator for High Dimensional Linear Regression," MIT, 2012 [Online]. Available: http://math.mit.edu/~liewang/L1PLAD_Wang.pdf

[74] H. Xu, C. Caramanis, and S. Mannor, "Robust regression and lasso," *IEEE Trans. Inf. Theory*, vol. 56, no. 7, pp. 3561–3574, Jul. 2010.

[75] S. P. Boyd and L. Vandenberghe*, Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[76] E. Y. Pee and J. O. Royset, "On solving large-scale finite minimax problems using exponential smoothing," *J. Optimiz. Theory App.*, vol. 148, no. 2, pp. 390–421, 2011.

[77] K. O. Kortanek and H. No, "A central cutting plane algorithm for convex semi-infinite programming problems," *SIAM J. Optimiz.*, vol. 3, no. 4, pp. 901–918, 1993.

[78] J. Elzinga and T. G. Moore, "A central cutting plane algorithm for the convex programming problem," *Math. Program.*, vol. 8, no. 1, pp. 134–145, 1975.

[79] B. Beckermann and A. B. J. Kuijlaars, "Superlinear convergence of conjugate gradients," *SIAM J. Numer. Anal.*, vol. 39, no. 1, pp. 300–329, 2002.

[80] M. Tan, I. Tsang, and L. Wang, "Matching Pursuit LASSO Part II: Applications and Sparse Recovery Over Batch Signals," *IEEE Trans. Signal Processing*, vol. 63, no. 3, pp. 742–753, Feb. 1, 2015.

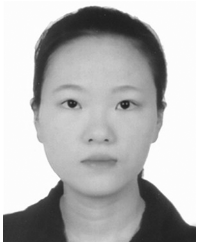
[81] R. Gribonval and P. Vandergheynst, "On the exponential convergence of matching pursuits in quasi-incoherent dictionaries," *IEEE Trans. Inf. Theory*, vol. 52, no. 1, pp. 255–261, Jan. 2006.

[82] M. Sion, "On general minimax theorems," *Pacific J. Math.*, vol. 8, no. 1, pp. 171–176, 1958.

Mingkui Tan received the Master's degree in control science and engineering in 2009 and his Bachelor's degree in environmental science and engineering in 2006, both from Hunan University in Changsha, China. He is currently working as a senior research associate with the School of Computer Science at The University of Adelaide in Australia. He received his Ph.D. degree in computer science from Nanyang Technological University, Singapore, in 2014. His research interests include compressive sensing, big data learning, and large-scale optimization.

Ivor W. Tsang is an Australian Future Fellow and Associate Professor with the Centre for Quantum Computation Intelligent Systems (QCIS), at the University of Technology, Sydney (UTS). Before joining UTS, he was the Deputy Director of the Centre for Computational Intelligence, Nanyang Technological University, Singapore. He was awarded his Ph.D. in computer science from the Hong Kong University of Science and Technology in 2007. He has published more than 100 research papers in refereed international journals and conference proceedings, including JMLR, TPAMI, TNN/TNNLS, NIPS, ICML, UAI, SIGKDD, IJCAI, AAAI, ACL, ICCV and CVPR.

In 2009, Dr. Tsang was conferred the 2008 Natural Science Award (Class II) by the Ministry of Education, China, which recognized his contributions to kernel methods. In 2013, Dr. Tsang received the prestigious Australian Research Council Future Fellowship for his research regarding Machine Learning on Big Data. In addition, he received the prestigious IEEE Transactions on Neural Networks Outstanding 2004 Paper Award in 2006, the 2014 IEEE Transactions on Multimedia Prized Paper Award, and a number of best paper awards and honors from reputable international conferences, including the Best Student Paper Award at CVPR 2010, the Best Paper Award at ICTAI 2011 and the Best Poster Award Honorable Mention at ACML 2012, etc. He was also awarded the Microsoft Fellowship 2005, and the ECCV 2012 Outstanding Reviewer Award.



Li Wang received her Ph.D. degree with Department of Mathematics in University of California, San Diego, USA. She received the masters degree in computational mathematics from Xi'an Jiaotong University, Shaanxi, China, in 2009 and the Bachelors degree in information and computing science from China University of Mining and Technology, Jiangsu, China in 2006. She is currently a postdoctoral Fellow at ICERM, Brown University, USA. Her research interests include large scale polynomial optimization, semi-infinite polynomial programming and machine learning.