Dual-density-based reweighted ℓ_1-algorithms for a class of ℓ_0-minimization problems

Jialiang Xu[a] and Yun-Bin Zhao[b]

[a,b] School of Mathematics, University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom

ARTICLE HISTORY
Compiled October 2, 2019

ABSTRACT
The optimization problem with sparsity arises in many areas of science and engineering such as compressed sensing, image processing, statistical learning and data sparse approximation. In this paper, we study the dual-density-based reweighted ℓ_1-algorithms for a class of ℓ_0-minimization models which can be used to model a wide range of practical problems. This class of algorithms is based on certain convex relaxations of the reformulation of the underlying ℓ_0-minimization model. Such a reformulation is a special bilevel optimization problem which, in theory, is equivalent to the underlying ℓ_0-minimization problem under the assumption of strict complementarity. Some basic properties of these algorithms are discussed, and numerical experiments have been carried out to demonstrate the efficiency of the proposed algorithms. Comparison of numerical performances of the proposed methods and the classic reweighted ℓ_1-algorithms has also been made in this paper.

KEYWORDS
Merit functions for sparsity, ℓ_0-minimization, dual-density-based algorithm, strict complementarity, bilevel optimization, convex relaxation.

1. Introduction

Let $\|x\|_0$ denote the number of nonzero components of the vector x. We consider the ℓ_0-minimization problem

$$\begin{aligned} \min_{x \in R^n} \quad & \|x\|_0 \\ \text{s.t.} \quad & \|y - Ax\|_2 \leqslant \epsilon, \ Bx \leqslant b, \end{aligned} \tag{1}$$

where $A \in R^{m \times n}$ and $B \in R^{l \times n}$ are two matrices with $m \ll n$ and $l \leq n$, $y \in R^m$ and $b \in R^l$ are two given vectors, and $\epsilon \geq 0$ is a given parameter, and $\|x\|_2 = (\sum_{i=1}^n |x_i|^2)^{1/2}$ is the ℓ_2-norm of x. In compressed sensing (CS), the parameter ϵ denotes the level of the measurement error $\eta = y - Ax$. Clearly, the problem (1) is to find the sparsest point in the convex set

$$T = \{x : \|y - Ax\|_2 \leqslant \epsilon, Bx \leqslant b\}. \tag{2}$$

CONTACT Yun-Bin Zhao (y.zhao.2@bham.ac.uk), and Jialiang Xu (JXX168@alumni.bham.ac.uk)

The constraint $Bx \le b$ is motivated by some practical applications. For instance, many signal recovery models might include extra constraints reflecting certain special structures or prior information of the target signals. The model (1) is general enough to cover several important applications in compressed sensing [4, 5, 11, 12], 1-bit compressed sensing [18, 22, 32] and statistical regression [20, 23, 27, 30]. The following two models are clearly the special cases of (1):

$$\text{(C1)} \ \min_x \{\|x\|_0 : \ y = Ax\}; \quad \text{(C2)} \ \min_x \{\|x\|_0 : \ \|y - Ax\|_2 \le \varepsilon\}.$$

The problem (C1) is often called the standard ℓ_0-minimization problem [6, 15, 32]. Some structured sparsity models, including the nonnegative sparsity model [5, 6, 15, 32] and the monotonic sparsity model (isotonic regression) [31], are also the special cases of the model (1).

Clearly, directly solving the problem (1) is generally very difficult since the ℓ_0-norm is a nonlinear, nonconvex and discrete function. Some algorithms have been developed for some special cases of the problem such as (C1) and (C2) over the past decade, including convex optimization and heuristic methods [11, 13, 15, 32]. For instance, by replacing the ℓ_0-norm in problem (1) with the ℓ_1-norm, we immediately obtain the ℓ_1-minimization problem

$$\min_x \quad \{\|x\|_1 : x \in T\}, \tag{3}$$

where T is given by (2). A more efficient class of models than (3) is the so-called weighted ℓ_1-minimization model [7, 16, 32, 35]. For (C1) and (C2), the reweighted ℓ_1-minimization model can be stated respectively as

$$\text{(E1)} \ \min_x \{\|Wx\|_1 : \ y = Ax\}; \quad \text{(E2)} \ \min_x \{\|Wx\|_1 : \ \|y - Ax\|_2 \le \varepsilon\},$$

where $W = \text{diag}(w)$ is a diagonal matrix with $w \in R^n_+$ being a weight vector. A single weighted ℓ_1-minimization is not efficient enough to outperform the standard ℓ_1-minimization. As a result, the reweighted ℓ_1-algorithm has been developed, which consists of solving a series of individual weighted ℓ_1-minimization problems [1, 2, 7, 16, 32, 35]. Taking the (C1) as example, this method solves a series of the following reweighted ℓ_1-problems:

$$\min_x \{(w^k)^T |x| : y = Ax\},$$

where k denotes the kth iteration and the weight w^k is updated by a certain rule. For example, the first-order method would yield a good updating scheme for w^k. The convergence of some reweighted algorithms was shown under certain conditions [8, 21, 32, 35]. The reweighted ℓ_1-minimization may perform better than ℓ_1-minimization on sparse signal recovery when the initial point is suitably chosen (see, e.g., [7, 8, 14, 21, 32, 35]). Although this paper focuses on the study of reweighted algorithms, it is worth mentioning that there exist other types of algorithms for ℓ_0-minimization problems, which have also been widely studied in the CS literature, such as orthogonal matching pursuits [13, 24, 29], compressed sampling matching pursuits [15, 25], subspace pursuits [9, 15], thresholding algorithms [3, 10, 13, 15], and the newly developed optimal k-thresholding algorithms [33].

Recently, a new framework of reweighted algorithms for sparse optimization problems was proposed in [34, 36] which is derived from the perspective of the dual density. The key idea is to use the complementarity between the solutions of the ℓ_0-minimization and theoretically equivalent weighted ℓ_1-minimization problem. Such complementarity property makes it possible to reformulate the ℓ_0-minimization problem as an equivalent bilevel optimization which seeks the densest solution of the dual problem of a weighted ℓ_1-problem. In this paper, we generalize this idea to the ℓ_0-minimization problem (1) and develop new dual-density-based algorithms through convex relaxation of the bilevel optimization. More specifically, to possibly solve the model (1), we consider the problem

$$\min_x \quad \{\|Wx\|_1 = w^T|x| : x \in T\}, \tag{4}$$

which is the weighted ℓ_1-minimization problem associated with the problem (1) for a given weight $w \in R^n_+$. The dual-density-based reweighted ℓ_1-algorithms for (1) are directly derived from the relaxation of the bilevel-optimization reformulation of the problem (1). To this goal, we develop some sufficient condition for the strict complementarity of the solutions of weighted ℓ_1-minimization problem associated with the problem (1) and the solution of its dual problem. We propose three types of convex relaxations of the bilevel optimization problem in order to develop our dual-density-based ℓ_1-algorithms for the problem (1).

The paper is organized as follows. In Sections 2, we recall the merit functions for sparsity and give a few examples of such functions, and we introduce the classic reweighted ℓ_1-algorithms. Section 3 is denoted to the development of a sufficient condition for the strict complementarity property to hold. In Section 4, we show that the ℓ_0-problem (1) can be reformulated equivalently as a bilevel optimization problem which, in theory, can generate an optimal weight for weighted ℓ_1-minimization problems. In Section 5, we discuss several new relaxation strategies for such a bilevel optimization problem, based on which we develop the dual-density-based reweighted ℓ_1-algorithms for the problem (1). Finally, we demonstrate some numerical results for the proposed algorithms.

Notation : The ℓ_p-norm on R^n is defined as $\|x\|_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$, where $p \geq 1$. The identity matrix of a suitable size is denoted by I. The n-dimensional Euclidean space is denoted by R^n. R^n_+ and R^n_{++} are the sets of nonnegative and positive vectors respectively, and R^n_- be the set of the nonpositive vectors. The complementary set of $S \subseteq \{1, ..., n\}$ is denoted by $\bar{S}$, i.e., $\bar{S} = \{1, ..., n\} \setminus S$. For a given vector $x \in R^n$ and $S \subseteq \{1, ..., n\}$, x_S is the subvector of x supported on S.

2. Preliminary

In this section, we recall the notion of merit functions for sparsity and list a few such examples. We also briefly outline the classic reweighted ℓ_1-methods for the problem (1). A function is called a merit function for sparsity if it can approximate the ℓ_0-norm in some sense [32, 35]. Some concave functions are shown to be the good candidates for the merit functions for sparsity [7, 19, 32, 34, 35]. As pointed out in [35, 36], we may choose a family of merit functions in the form

$$\Psi_\varepsilon(s) = \sum_{i=1}^{n} \varphi_\varepsilon(s_i),\ \ s \in R_+^n,$$

where φ_ε is a function from R_+ to R_+. $\Psi_\varepsilon(s)$ satisfies the following properties:

- $(P1)$ for any given $s \in R_+^n$, $\Psi_\varepsilon(s)$ tends to $\|s\|_0$ as ε tends to 0;
- $(P2)$ $\Psi_\varepsilon(s)$ is twice continuously differentiable with respect to $s \in R_+^n$ in the open neighborhood of R_+^n;
- $(P3)$ $\varphi_\varepsilon(s_i)$ is concave and strictly increasing with respect to every $s_i \in R_+$.

We denote the set of such merit functions by

$$\mathbf{F} = \{\Psi_\varepsilon :\ \Psi_\varepsilon \text{ satisfies } (P1) - (P3)\}.$$

The following merit functions satisfying $(P1)$-$(P3)$ have been used in [35, 36]:

$$\Psi_\varepsilon(s) = n - \frac{\sum_{i=1}^{n} \log(s_i + \varepsilon)}{\log \varepsilon},\ \ s \in R_+^n, \tag{5}$$

$$\Psi_\varepsilon(s) = \sum_{i=1}^{n} \frac{s_i}{s_i + \varepsilon},\ \ s \in R_+^n, \tag{6}$$

$$\Psi_\varepsilon(s) = \sum_{i=1}^{n} (s_i + \varepsilon^{1/\varepsilon})^\varepsilon,\ \ s \in R_+^n \tag{7}$$

where $\varepsilon \in (0, 1)$. In this paper, we also use the following merit function:

$$\Psi_\varepsilon(s) = \frac{2}{\pi} \sum_{i=1}^{n} \arctan(\frac{s_i}{\varepsilon}),\ \ s \in R_+^n, \tag{8}$$

where $\varepsilon > 0$. It is easy to show that (8) belongs to the set $\mathbf{F}$.

Lemma 2.1. *The function* (8) *satisfies* $(P1)$-$(P3)$ *on* R_+^n.

Proof. Obviously, the function (8) satisfies $(P1)$ and $(P2)$. We now prove that it also satisfies $(P3)$. In R_+^n, note that

$$\nabla \Psi_\varepsilon(s) = (\nabla \varphi_\varepsilon(s_1), \ldots, \nabla \varphi_\varepsilon(s_n))^T = \frac{2}{\pi} \left(\frac{\varepsilon}{s_1^2 + \varepsilon^2}, \ldots, \frac{\varepsilon}{s_n^2 + \varepsilon^2} \right)^T,$$

and

$$\nabla^2 \Psi_\varepsilon(s) = \frac{4}{\pi} \text{diag}\left(-\frac{\varepsilon s_1}{(s_1^2 + \varepsilon^2)^2}, \ldots, -\frac{\varepsilon s_n}{(s_n^2 + \varepsilon^2)^2} \right).$$

Due to $s_i \geqslant 0$ and $\varepsilon > 0$, we have $\nabla \varphi_\varepsilon(s_i) > 0$ and $\nabla^2 \varphi_\varepsilon(s_i) \leqslant 0$ for $i = 1, ..., n$ which implies that $\Psi_\varepsilon(s)$ is concave and strictly increasing with respect to every entry of $s \in R_+^n$. Thus (8) satisfies $(P1), (P2)$ and $(P3)$. □

In order to compare the algorithms proposed in later sections, we briefly introduce the classic reweighted ℓ_1-method. Following the idea in [35] and [32], replacing $\|x\|_0$ with $\Psi_\varepsilon(|x|) \in \mathbf{F}$ leads to the following approximation of the problem (1):

$$\min_{(x,t)} \{\Psi_\varepsilon(t) : \ x \in T, \ |x| \leq t\}. \tag{9}$$

By using the first order approximation of $\Psi_\varepsilon(t) \in \mathbf{F}$ at the point t^k, the problem (9) can be approximated by the linear optimization

$$\min_{(x,t)} \{\nabla \Psi_\varepsilon^T(t^k)t : \ x \in T, \ |x| \leq t\}, \tag{10}$$

which is used to generate the new iterate (x^{k+1}, t^{k+1}). Due to the fact that $\Psi_\varepsilon(t)$ is strictly increasing with respect to each $t_i \in R_+$, it is evident that the iterate (x^k, t^k) must satisfy $t^k = |x^k|$, which implies that

$$x^{k+1} \in \underset{x}{\operatorname{argmin}} \{\nabla \Psi_\varepsilon^T(|x^k|)|x| : x \in T\}.$$

This is the classic reweighted ℓ_1-minimization method described in [32].

Algorithm 1 Reweighted ℓ_1-algorithm (**RA**)

Input:

merit function $\Psi_\varepsilon \in \mathbf{F}$, matrices $A \in R^{m \times n}$ and $B \in R^{l \times n}$;
vectors $y \in R^m$, $b \in R^l$ and $\epsilon \in R_+$ and parameters $\varepsilon \in R_{++}$;
initial weight w^0, the iteration index k and the largest number of iterations $k_{\max}$.

Main step: At the current iterate x^{k-1}, solve the weighted ℓ_1-minimization

$$x^k \in \operatorname{argmin} \left\{ \sum_{i=1}^n w_i^k |x_i| : \quad x \in T \right\},$$

where $w_i^k = \nabla \Psi_\varepsilon(|x^{k-1}|)_i = \nabla \varphi_\varepsilon(|x_i^{k-1}|), \ i = 1, ..., n$.

Update: $w_i^{k+1} := (\nabla \Psi_\varepsilon(|x^k|))_i = \nabla \varphi_\varepsilon(|x_i^k|)$, $i = 1, ..., n$; Repeat the above main step until $k = k_{\max}$ (or certain other stopping criterion is met).

Based on the generic convergence of revised Frank-Wolfe algorithms (*FW-RD*) for a class of concave functions in [26], the generic convergence of the algorithm RA can be obtained (see details in [26]), that is, there exists a family of merit functions $\Psi_\varepsilon \in \mathbf{F}$ such that RA converges to a stationary point of the problem. The convergence of RA to a sparse point in the case of linear-system constraints can be found in [32].

3. Duality, strict complementarity and optimality condition

To develop the dual-density-based reweighted ℓ_1-algorithms, we first discuss the duality and the optimality condition of the model (4), and we give a sufficient condition for the strict complementarity to satisfy for the model (4).

3.1. Duality and complementary condition

By introducing two variables $t \in R^n$ and $\gamma \in R^m$ such that

$$|x| \leqslant t \quad \text{and} \quad \gamma = y - Ax,$$

we can rewrite (4) as the following problem:

$$\begin{array}{ll} \min\limits_{(x,\gamma,t)} & w^T t \\ \text{s.t.} & \|\gamma\|_2 \leqslant \epsilon, \ Bx \leqslant b, \\ & \gamma = y - Ax, \ |x| \leqslant t, \ t \geqslant 0. \end{array} \tag{11}$$

Obviously, (11) is equivalent to (4). Additionally, if $w \in R^n_{++}$, then the solution (x^*, t^*, γ^*) to (11) must satisfy that $|x^*| = t^*$ and $\gamma^* = y - Ax^*$, and the following relation of the solutions of (4) and (11) is obvious.

Lemma 3.1. *If x^* is optimal to the problem* (4)*, then all vectors (x^*, t^*, γ^*) satisfying*

$$|x^*_{\text{supp}(w)}| = t^*_{\text{supp}(w)}, \ |x^*_{\overline{\text{supp}(w)}}| \leqslant t^*_{\overline{\text{supp}(w)}} \quad \text{and} \quad \gamma^* = y - Ax^*$$

are optimal to the problem (11)*. Moreover, if $(\bar{x}, \bar{t}, \bar{\gamma})$ is optimal to the problem* (11)*, then $\bar{x}$ is optimal to the problem* (4)*.*

Let $\lambda = (\lambda_1, ..., \lambda_6)$ be the dual variable, then the dual problem of (11) can be stated as follows:

$$\begin{array}{ll} \max\limits_{\lambda} & -\lambda_1 \epsilon - \lambda_2^T b + \lambda_3^T y \\ \text{s.t.} & B^T \lambda_2 - A^T \lambda_3 + \lambda_4 - \lambda_5 = 0, \\ & w = \lambda_4 + \lambda_5 + \lambda_6, \ \|\lambda_3\|_2 \leq \lambda_1, \\ & \lambda_i \geqslant 0, \ i = 1, 2, 4, 5, 6, . \end{array} \tag{12}$$

The strong duality between (11) and (12) can be guaranteed under suitable condition. Thus the following results follows from the classic optimization theory [28].

Lemma 3.2. *Let the Slater condition hold for the convex problem* (11)*, i.e., there exists $(x^*, \gamma^*, t^*) \in ri(T)$ such that*

$$\|\gamma^*\|_2 < \epsilon, \ Bx^* \leqslant b, \ |x^*| \leqslant t^*, \ y = Ax^* - \gamma^*, \ t^* \geqslant 0,$$

where $ri(T)$ is the relative interior of T. Then there is no duality gap between (11) *and its dual problem* (12)*. Moreover, if the optimal value of* (11) *is finite, then there exists at least one optimal Lagrangian multiplier such that the dual optimal value can be attained.*

In this paper, we assume that the Slater condition holds for (11). Clearly, the optimal value of (11) is finite when w is a given vector, and hence the strong duality holds for (11) and (12) and the dual optimal value can be attained. Actually, the set $\Omega = \{x : Ax = y, \ Bx \leqslant b\}$ is in practice not empty due to the fact that y and b are the measurements of the signals. Thus the Slater condition is a very mild sufficient condition for strong duality to hold for the problems (11) and (12).

3.2. Optimality condition for (11) and (12)

It is well-known that for any convex minimization problem with differentiable objective and constraint functions for which the strong duality holds, Karush-Kuhn-Tucker (KKT) condition is the necessary and sufficient optimality condition for the problem and its dual problem [28]. Since the Slater condition holds for (11), by Lemma 3.2, the optimality condition for (11) is stated as follows.

Theorem 3.3. *If Slater condition holds for* (11)*, then* (x^*, γ^*, t^*) *is optimal to* (11) *and* $\lambda_i^*, i = 1, ..., 6$ *is optimal to* (12) *if and only if* $(x^*, \gamma^*, t^*, \lambda^*)$ *satisfy the KKT conditions for* (11)*, i.e.,*

$$\begin{cases} \gamma^* = y - Ax^*, \ \ \|\gamma^*\|_2 \leqslant \epsilon, \ x^* \leqslant t^*, \ \ -t^* \leqslant x^*, \\ Bx^* \leqslant b, \ t^* \geqslant 0, \ \lambda_i^* \geq 0, \ i = 1, 2, 4, 5, 6, \\ \lambda_1^*(\epsilon - \|\gamma^*\|_2) = 0, \ \lambda_2^{*T}(b - Bx^*) = 0, \\ \lambda_4^{*T}(t^* - x^*) = 0, \ \lambda_6^{*T} t^* = 0, \ \lambda_5^{*T}(x^* + t^*) = 0, \\ L(x, \gamma, t, \lambda^*) = w^T t - \lambda_1^*(\epsilon - \|\gamma\|_2) - \lambda_2^{*T}(b - Bx) - \lambda_3^{*T}(Ax + \gamma - y) \\ \qquad\qquad - \lambda_4^{*T}(t^* - x^*) - \lambda_5^{*T}(x^* + t^*) - \lambda_6^{*T} t^*, \\ \partial_x L(x^*, \gamma^*, t^*, \lambda^*) = B^T \lambda_2^* - A^T \lambda_3^* + \lambda_4^* - \lambda_5^* = 0, \\ \partial_\gamma L(x^*, \gamma^*, t^*, \lambda^*) = (\lambda_1^*) \nabla(\|\gamma^*\|_2) - \lambda_3^* = 0, \\ \partial_t L(x^*, \gamma^*, t^*, \lambda^*) = w - \lambda_4^* - \lambda_5^* - \lambda_6^* = 0. \end{cases} \tag{13}$$

From the optimality condition in (13), we see that t^* and λ_6^* satisfy the complementary condition.

Corollary 3.4. *Let the Slater condition hold for* (11)*. Then, for any optimal solution pair* $((x^*, t^*, \gamma^*), \lambda^*)$*, where* (x^*, t^*, γ^*) *is optimal to* (11) *and* $\lambda^* = (\lambda_1^*, ..., \lambda_6^*)$ *is optimal to* (12)*,* t^* *and* λ_6^* *are complementary in the sense that*

$$(t^*)^T \lambda_6^* = 0, \ t^* \geqslant 0 \text{ and } \lambda_6^* \geqslant 0.$$

Clearly, if (x^*, t^*, γ^*) is optimal to (11) and w is positive, it must hold $|x^*| = t^*$. Hence by Corollary 3.4, for $i = 1, ..., n$, we have

$$|x_i^*|(\lambda_6^*)_i = 0, \ (\lambda_6^*)_i \geq 0. \tag{14}$$

When w is nonnegative, and if (x^*, t^*, γ^*) is optimal to (11), we have

$$|x_i^*| = t_i^*, \ i \in \text{supp}(w); \ |x_i^*| \leqslant t_i^*, \ i \in \overline{\text{supp}(w)}.$$

For $i \in \text{supp}(w)$, (14) is valid. For $i \in \overline{\text{supp}(w)}$, due to the constraints $w = \lambda_4 + \lambda_5 + \lambda_6$ and $\lambda_4, \lambda_5, \lambda_6 \geqslant 0$, $w_i = 0$ implies that $(\lambda_6^*)_i = 0$. This means (14) is also valid for $i \in \overline{\text{supp}(w)}$. Therefore, we have the following result:

Theorem 3.5. *Let* w *be a nonnegative given vector, and let the Slater condition hold for* (11)*. Then, for any optimal solution pair* $((x^*, t^*, \gamma^*), \lambda^*)$*, where* (x^*, t^*, γ^*) *is optimal to* (11) *and* $\lambda^* = (\lambda_1^*, ..., \lambda_6^*)$ *is optimal to* (12)*,* $|x_i^*|$ *and* $(\lambda_6^*)_i$ *are complementary in the sense that*

$$|x_i^*|(\lambda_6^*)_i = 0 \text{ and } (\lambda_6^*)_i \geqslant 0, \ i = 1, ..., n. \tag{15}$$

The relation (15) implies that

$$\|x^*\|_0 + \|\lambda_6^*\|_0 \leq n,$$

where n is the dimension of x^* or λ_6^*. Suppose $|x^*|$ and λ_6^* are strictly complementary, i.e.,

$$|x^*|^T \lambda_6^* = 0, \ \lambda_6^* \geq 0 \text{ and } |x^*| + \lambda_6^* > 0.$$

Then

$$\|x^*\|_0 + \|\lambda_6^*\|_0 = n.$$

3.3. Strict complementarity

For nonlinear optimization models, the strictly complementary property might not hold. However, it might be possible to develop a condition such that the strict complementarity holds for the model (4) or (11). We now develop such a condition for the problems (11) and (12) under the following assumption.

Assumption 3.6. Let $W = \text{diag}(w)$ satisfy the following properties:

- $\langle G1 \rangle$ The problem (4) with w has an optimal solution which is a relative interior point in the feasible set T, denoted by $x^* \in ri(T)$, such that

$$\|y - Ax^*\|_2 < \epsilon, \ Bx^* \leq b,$$

- $\langle G2 \rangle$ the optimal value Z^* of (4) is finite and positive, i.e., $Z^* \in (0, \infty)$,

- $\langle G3 \rangle$ $w_j \in (0, \infty]$ for all $1 \leqslant j \leqslant n$.

Example 3.7. Consider the system $\|y - Ax\|_2 \leqslant \epsilon, Bx \leqslant b$ with $\epsilon = 10^{-1}$, where

$$A = \begin{bmatrix} 1 & 0 & -2 & 5 \\ 0 & 1 & 4 & -9 \\ 1 & 0 & -2 & 5 \end{bmatrix}, B = \begin{bmatrix} -0.5 & 0 & 1 & -2.5 \\ 0.5 & -0.5 & -1 & 2 \\ -3 & -3 & -2 & 3 \end{bmatrix}, y = \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}, b = \begin{bmatrix} -0.5 \\ 1 \\ -1 \end{bmatrix}.$$

We can see that the problem (4) with $w = (1, 100, 1, 100)^T$ has an optimal solution $(1/2, 0, -1/4, 0)^T$ which satisfies Assumption 3.6.

Next we prove the following theorem concerning the strict complementarity for (11) and (12) under Assumption 3.6.

Theorem 3.8. *Let y and b be two given vectors, $A \in R^{m \times n}$ and $B \in R^{l \times n}$ be two given matrices, and w be a given weight which satisfies Assumption 3.6. Then there exists a pair $((x^*, t^*, \gamma^*), \lambda^*)$, where (x^*, t^*, γ^*) is an optimal solution to (11) and $\lambda^* = (\lambda_1^*, ..., \lambda_6^*)$ is an optimal solution to (12), such that t^* and λ_6^* are strictly complementary, i.e.,*

$$(t^*)^T \lambda_6^* = 0, \ t^* + \lambda_6^* > 0, \ (t^*, \lambda_6^*) \geq 0.$$

Proof. Note that $(G1)$ in Assumption 3.6 implies the Slater condition for (11). This, combined with $(G2)$, indicates from Lemma 3.2 that the duality gap is 0, and the optimal value Z^* for (12) can be attained. For any given index $j : 1 \le j \le n$, we consider a series of minimization problems:

$$\begin{array}{ll} \min\limits_{(x,t,\gamma)} & -t_j \\ \text{s.t.} & \|\gamma\|_2 \leqslant \epsilon,\ Bx \leqslant b,\ \gamma = y - Ax, \\ & |x| \leqslant t,\ \ -w^T t \ge -Z^*,\ t \geqslant 0. \end{array} \tag{16}$$

The dual problem of (16) can be obtained by using the same method for developing the dual problem of (11), which is stated as follows:

$$\begin{array}{ll} \max\limits_{(\mu,\tau)} & -\mu_1\epsilon - \mu_2^T b + \mu_3^T y - \tau Z^* \\ \text{s.t.} & B^T\mu_2 - A^T\mu_3 + \mu_4 - \mu_5 = 0,\ \ \|\mu_3\|_2 \le \mu_1, \\ & \tau w = \mu_4 + \mu_5 + \mu_6 + e^j,\ \mu_i \geqslant 0,\ i = 1,2,4,5,6,\ \tau \ge 0, \end{array} \tag{17}$$

where e^j is a vector whose jth component is 1 and the remains are 0, i.e.,

$$e_i^j = 1,\ i = j;\ e_i^j = 0,\ i \neq j.$$

Next we show that (16) and (17) satisfy the strong duality property under Assumption 3.6. It can be seen that (x,t,γ) is a feasible solution to (16) if and only if (x,t,γ) is an optimal solution of (11), or if x is optimal to (4). If w satisfies the conditions in Assumption 3.6, then there exists an optimal solution $\bar{x}$ of (4) such that $\|y - A\bar{x}\|_2 < \epsilon,\ B\bar{x} \leqslant b,\ w^T|\bar{x}| = Z^*$ which means there is a relative interior point $(\bar{x},\bar{t},\bar{\gamma})$ of the feasible set of (16) satisfying

$$\|\bar{\gamma}\|_2 < \epsilon,\ B\bar{x} \leqslant b,\ \bar{\gamma} = y - A\bar{x},\ |\bar{x}| \le \bar{t},\ w^T\bar{t} \le Z^*,\ \bar{t} \ge 0.$$

As a result, the strong duality holds for (16) and (17) for all j. Moreover, due to $(G2)$ and $(G3)$, w is positive and Z^* is finite, so t_j cannot be ∞. Thus the optimal value of all jth minimization problems (16) is finite. It follows from Lemma 3.2 that for each jth optimization (16) and (17), the duality gap is 0, and each jth dual problem (17) can achieve their optimal value.

We use ξ_j^* to denote the optimal value of the jth primal problem in (16). Clearly, ξ_j^* is nonpositive, i.e.,

$$\xi_j^* < 0 \ \text{ or } \ \xi_j^* = 0.$$

Case 1: $\xi_j^* < 0$. Then (11) has an optimal solution (x', t', γ') where the jth component in t' is positive since $t'_j = -\xi_j^*$ and admits the largest value amongst all the optimal solutions of (11). By Theorem 3.4, the complementary condition implies that (12) has an optimal solution $\lambda' = (\lambda'_1, ..., \lambda'_6)$ where jth component in λ'_6 is 0. Then we have an optimal solution pair $((x', t', \gamma'), \lambda')$ for (11) and (12) such that $t'_j > 0$ and $(\lambda'_6)_j = 0$. It means that

$$t'_j = -\xi_j^* > 0 \quad \text{implies} \quad (\lambda'_6)_j = 0.$$

Case 2: $\xi_j^* = 0$. Following from the strong duality between (16) and (17), we have an optimal solution (μ, τ) of the jth optimization problem (17) such that

$$-\mu_1 \epsilon - \mu_2^T b + \mu_3^T y = \tau Z^*.$$

First, we consider $\tau \neq 0$. The above equality can be reduced to

$$-\frac{\mu_1}{\tau} - \frac{\mu_2^T}{\tau} b + \frac{\mu_3^T}{\tau} y = Z^*,$$

and we also have

$$B^T \frac{\mu_2}{\tau} - A^T \frac{\mu_3}{\tau} + \frac{\mu_4}{\tau} - \frac{\mu_5}{\tau} = 0, \ \left\| \frac{\mu_3}{\tau} \right\|_2 \leq \frac{\mu_1}{\tau}, \ w = \frac{\mu_4}{\tau} + \frac{\mu_5}{\tau} + \frac{\mu_6}{\tau} + \frac{e^j}{\tau}.$$

We set

$$\lambda_1^{'} = \frac{\mu_1}{\tau}, \ \lambda_2^{'} = \frac{\mu_2}{\tau}, \ \lambda_3^{'} = \frac{\mu_3}{\tau}, \ \lambda_4^{'} = \frac{\mu_4}{\tau}, \ \lambda_5^{'} = \frac{\mu_5}{\tau}, \ \lambda_6^{'} = \frac{\mu_6}{\tau} + \frac{e^j}{\tau}.$$

Due to strong duality of (11) and (12) again, $\lambda^{'} = (\lambda_1^{'}, ..., \lambda_6^{'})$ is optimal to (12). Note that

$$(\lambda_6)_j^{'} = \frac{(\mu_6)_j + 1}{\tau}.$$

Thus $(\lambda_6)_j^{'} > 0$, which follows from $\mu_6 \geq 0$ and $\tau > 0$. Thus

$$t_j' = -\xi_j^* = 0 \ \text{ implies } \ (\lambda_6)_j^{'} > 0.$$

Note that the third constraint in jth optimization of (17) requires $\tau \neq 0$ since w, μ_4, μ_5, μ_6 are all non-negative and $e_j^j = 1$ so that the jth component in τw must be greater or equal than 1. Therefore, all jth optimization problems in (17) are infeasible if $\tau = 0$. As a result, the optimal solution (μ, τ) of (17) with $\tau = 0$ is impossible to occur. Combining the cases 1 and 2 implies that for each $1 \leq j \leq n$, we have an optimal solution pair $((x^j, t^j, \gamma^j), \lambda^j)$ such that $t_j^j > 0$ or $(\lambda_6^j)_j > 0$. For all jth solution pairs, they all satisfy the following properties:

- (1) (x^j, t^j, γ^j) is optimal to (11), and $(\lambda_1^j, \lambda_2^j, \lambda_3^j, \lambda_4^j, \lambda_5^j, \lambda_6^j)$ is optimal to (12);
- (2) the jth component of t^j and the jth component of λ_6^j are strictly complementary, such that $t_j^j (\lambda_6^j)_j = 0, \ t_j^j + (\lambda_6^j)_j > 0$.

Denote $(x^*, t^*, \gamma^*, \lambda^*)$ by

$$x^* = \frac{1}{n} \sum_{j=1}^{n} x^j, \ t^* = \frac{1}{n} \sum_{j=1}^{n} t^j, \ \gamma^* = \frac{1}{n} \sum_{j=1}^{n} \gamma^j, \ \lambda_i^* = \frac{1}{n} \sum_{j=1}^{n} \lambda_i^j, \ i = 1, 2, \cdots, 6.$$

Since $(x^j, t^j, \gamma^j), \ j = 1, 2, ..., n$ are all optimal solutions of (11), then for any j, we have

$$\begin{cases} w^T t^j = Z^*, \ \left\| \gamma^j \right\|_2 \leq \epsilon, \ B x^j \leq b, \\ \gamma^j = y - A x^j, \ |x^j| \leq t^j, \ t^j \geq 0. \end{cases} \tag{18}$$

It is easy to see that

$$w^T t^* = Z^*,\ \ Bx^* \le b,\ \ \gamma^* = y - Ax^*,\ \ t^* \ge 0.$$

Moreover,

$$\|\gamma^*\|_2 = \left\| \frac{1}{n}\sum_{j=1}^{n} \gamma^j \right\|_2 \le \sum_{j=1}^{n} \left\| \frac{1}{n}\gamma^j \right\|_2 \le \epsilon,$$

$$|x^*| = \left| \frac{1}{n}\sum_{j=1}^{n} x^j \right| \le \frac{1}{n}\sum_{j=1}^{n} |x^j| \le \frac{1}{n}\sum_{j=1}^{n} t^j = t^*,$$

where the first inequality of each equation above follows from the triangle inequality. Then the vector (x^*, t^*, γ^*) satisfies

$$\begin{cases} w^T t^* = Z^*,\ \ \|\gamma^*\|_2 \le \epsilon,\ \ Bx^* \le b, \\ \gamma^* = y - Ax^*,\ \ |x^*| \le t^*,\ \ t^* \ge 0. \end{cases} \tag{19}$$

Thus (x^*, t^*, γ^*) is optimal to (11), and similarly it can be proven that $\lambda^* = (\lambda_1^*, ..., \lambda_6^*)$ is an optimal solution to (12). By strong duality, t^* and λ_6^* are complementary. Due to the above-mentioned property (2), it is impossible to find a pair (t^*, λ_6^*) such that their jth component are both 0. Thus, (t^*, λ_6^*) is the strictly complementary solution pair for (11) and (12).

□

Remark 1. It can be seen that the following two sets

$$P^* = \{i : t_i^* > 0\} \text{ and } Q^* = \{i : (\lambda_6^*)_i > 0\}$$

are invariant for all pairs of strictly complementary solutions. Suppose there are two distinct optimal pairs of the solutions of (11) and (12), denoted by $(x^k, t^k, \gamma^k, \lambda^k), k = 1, 2$, such that $(t^k, \lambda_6^k), k = 1, 2$ are strictly complementary pairs, where (x^k, t^k, γ^k) are optimal to (11) and (λ^k) are optimal to (12). Due to Theorem 3.4, we know that

$$(\lambda_6^1)^T t^2 = 0 \text{ and } (\lambda_6^2)^T t^1 = 0.$$

It means that the supports of all strictly complementary pairs of (11) and (12) are invariant. Otherwise, there exists an index j such that $t_j^1 > 0$ and $(\lambda_6^2)_j > 0$, leading to a contradiction.

Since the optimal solution (x^*, t^*, γ^*) to (11) must have $t^* = |x^*|$ if $w > 0$, the main results of Theorem 3.8 also imply that $|x^*|$ and λ_6^* are strictly complementary under Assumption 3.6.

4. Bilevel model for optimal weights

For weighted ℓ_1-minimization, how to determine a weight to guarantee the exact recovery, sign recovery or support recovery of sparse signals is an important issue in CS theory. Based on the complementary condition and strict complementarity discussed above, we may develop a bilevel optimization model for such a weight, which is called the optimal weight in [34], [36] and [32].

Definition 4.1 (Optimal Weight)**.** A weight is called an optimal weight if the solution of the weighted ℓ_1-problem with this weight is one of the optimal solution of the ℓ_0-minimization problem.

Let Z^* be the optimal value of (4). Notice that the optimal solution of (4) remains the same when w is replaced by αw for any positive α. When $Z^* \neq 0$, by replacing W by W/Z^*, we can obtain

$$1 = \min_x \{\|(W/Z^*)x\|_1 : \ x \in T\}.$$

We use ζ to denote the set of such weights, i.e.,

$$\zeta = \{w \in R^n_+ : \ 1 = \min_x \{\|Wx\|_1, x \in T\}\}, \tag{20}$$

where $W = \text{diag}(w)$. Clearly, $\bigcup_{\alpha>0} \alpha\zeta$ is the set of weights such that (4) has a finite and positive optimal value, and ζ is not necessarily bounded. Under the Slater condition, Theorem 3.5 implies that given any $w \in \zeta$, any optimal solutions of (11) and (12), denoted by $(x^*(w), t^*(w), \gamma^*(w))$ and $\lambda^*(w) = (\lambda_1^*(w), ..., \lambda_6^*(w))$, satisfy that $|x^*(w)|$ and $\lambda_6^*(w)$ are complementary, i.e.,

$$\|x^*(w)\|_0 + \|\lambda_6^*(w)\|_0 \leqslant n. \tag{21}$$

If w^* satisfies Assumption 3.6, then Slater condition is automatically satisfied for (11) with w^* and (21) is also valid. Moreover, by Theorem 3.8, there exists a strictly complementary pair $(|x^*(w^*)|, \lambda_6^*(w^*))$ such that

$$\|x^*(w^*)\|_0 + \|\lambda_6^*(w^*)\|_0 = n.$$

If w^* is an optimal weight (see Definition 4.1), then $\lambda_6^*(w^*)$ must be the densest slack variable among all $w \in \zeta$, and locating a sparse vector can be converted to

$$\lambda_6^*(w^*) = \text{argmax}\{\|\lambda_6^*(w)\|_0 : \ w \in \zeta\}.$$

Inspired by the above fact, we develop a theorem under Assumption 4.2 which claims that finding a sparsest point in T is equivalent to seeking the proper weight w such that the dual problem (12) has the densest optimal variable λ_6. Such weights are optimal weights and can be determined by certain bilevel optimization. This idea was first introduced by Zhao and Kočvara [34] (and also by Zhao and Luo [36]) to solve the standard ℓ_0-minimization (C1). In this paper, we generalize their idea to solve the model (1) by developing new convex relaxation technique for the underlying bilevel optimization problem. Before that we make the following assumption:

Assumption 4.2. Let ν be an arbitrary sparsest point in T given in (2). There exists a weight $\bar{w} \geqslant 0$ such that

- $\langle H1 \rangle$ the problem (4) with $\bar{w}$ has an optimal solution $\bar{x}$ such that $\|\bar{x}\|_0 = \|\nu\|_0$,
- $\langle H2 \rangle$ there exists an optimal variable in (12) with $\bar{w}$, denoted as $\bar{\lambda}$, such that $\bar{\lambda}_6$ and $\bar{x}$ are strictly complementary,
- $\langle H3 \rangle$ the optimal value of (4) with $\bar{w}$ is finite and positive.

An example for the existence of a weight satisfying Assumption 4.2 is given in the remark following the next theorem.

Theorem 4.3. *Let Slater condition and Assumption 4.2 hold. Consider the bilevel optimization*

$$\begin{array}{lll} (P_b) & \max\limits_{(w,\lambda)} & \|\lambda_6\|_0 \\ & \text{s.t.} & B^T\lambda_2 - A^T\lambda_3 + \lambda_4 - \lambda_5 = 0, \ \|\lambda_3\|_2 \leq \lambda_1, \\ & & -\lambda_1\epsilon - \lambda_2^T b + \lambda_3^T y = \min\limits_x \{\|Wx\|_1 : x \in T\}, \\ & & w = \lambda_4 + \lambda_5 + \lambda_6 \geq 0, \ \lambda_i \geqslant 0, \ i = 1,2,4,5,6, \end{array} \tag{22}$$

where $W = \operatorname{diag}(w)$*, and* T *is given as* (2)*. If* (w^*, λ^*) *is an optimal solution to the above optimization problem* (22)*, then any optimal solution* x^* *to*

$$\min_x \{\|W^*x\|_1 : x \in T\}, \tag{23}$$

is a sparsest point in T*.*

Proof. Let ν be a sparsest point in T. Suppose that (w^*, λ^*) is an optimal solution of (22). We now prove that any optimal solution to (23) is a sparsest point in T under Assumption 4.2. Let w' be a weight satisfying Assumption 4.2, meaning that (4) with $W = \operatorname{diag}(w')$ has an optimal solution x' such that $\|x'\|_0 = \|\nu\|_0$. Moreover, there exists a strictly complementary pair (x', λ'_6) satisfying

$$\|x'\|_0 + \|\lambda'_6\|_0 = n = \|\lambda'_6\|_0 + \|\nu\|_0 . \tag{24}$$

where the vector $\lambda' = (\lambda'_1, ..., \lambda'_6)$ is the dual optimal solution of (12) with $w = w'$, i.e.,

$$\begin{array}{ll} \max\limits_{\lambda} & -\lambda_1\epsilon - \lambda_2^T b + \lambda_3^T y \\ \text{s.t.} & B^T\lambda_2 - A^T\lambda_3 + \lambda_4 - \lambda_5 = 0, \ \|\lambda_3\|_2 \leq \lambda_1, \\ & w' = \lambda_4 + \lambda_5 + \lambda_6, \ \lambda_i \geqslant 0, \ i = 1,2,4,5,6. \end{array} \tag{25}$$

By Lemma 3.2, the Slater condition implies that strong duality holds for the problems (25) and (11) with w'. Note that the optimal values of (11) and (4) with w' are equal and finite so that (w', λ') is feasible to (22). Let x^* be an arbitrary solution to (23). Note that (11) with w^* is equivalent to (23), to which the dual problem is

$$\begin{array}{ll} \max\limits_{\lambda} & -\lambda_1\epsilon - \lambda_2^T b + \lambda_3^T y \\ \text{s.t.} & B^T\lambda_2 - A^T\lambda_3 + \lambda_4 - \lambda_5 = 0, \ \|\lambda_3\|_2 \leq \lambda_1, \\ & w^* = \lambda_4 + \lambda_5 + \lambda_6, \ \lambda_i \geqslant 0, \ i = 1,2,4,5,6. \end{array} \tag{26}$$

Moreover, $\lambda^* = (\lambda_1^*, ..., \lambda_6^*)$ is feasible to (26) and the fourth constraint of (22) implies that there is no duality gap between (11) with w^* and (26). Thus, by strong duality, $\lambda^* = (\lambda_1^*, ..., \lambda_6^*)$ is an optimal solution to (26). Therefore, by Theorem 3.5, $|x^*|$ and λ_6^* are complementary. Hence, we have

$$\|x^*\|_0 \leqslant n - \|\lambda_6^*\|_0 . \tag{27}$$

Since (w^*, λ^*) is optimal to (22), we have

$$\left\|\lambda_6'\right\|_0 \leqslant \|\lambda_6^*\|_0 . \tag{28}$$

Plugging (24) and (28) into (27) yields

$$\|x^*\|_0 \leqslant n - \|\lambda_6^*\|_0 \leqslant n - \left\|\lambda_6'\right\|_0 = \left\|x'\right\|_0 = \|\nu\|_0 ,$$

which implies $\|x^*\|_0 = \|\nu\|_0$, due to the assumption that ν is the sparsest point in T. Then any optimal solution to (24) is a sparsest point in T.

□

Given Assumption 4.2 and Slater condition, finding a sparsest point in T is tantamountly equal to look for the densest dual solution via the bilevel model (22).

By the definition of optimal weights, Theorem 4.3 implies that w^* is an optimal weight by which a sparsest point can be obtained via (4). If there is no weight satisfying the properties in Assumption 4.2, a heuristic method for finding a sparse point in T can be also developed from (21) since the increase in $\|\lambda_6(w)\|_0$ leads to the decrease of $\|x(w)\|_0$ to a certain level. Before we close this section, we make some remarks for Assumption 4.2.

Remark 2. Consider Example 3.7. It can be seen that $(0, 0, 2, 1)^T$ is a sparsest point in the feasible set T of this example. If we choose weights $w = (100, 100, 1, 1)^T$, then we can see that $(0, 0, 2, 1)^T$ is the unique optimal solution of (4) which satisfies $\langle H1\rangle$ and $\langle H3\rangle$ in Assumption 4.2. In addition, $(0, 0, 2, 1)^T$ is a relative interior point in the feasible set T. This, combined with the fact that weights are positive, implies that Assumption 3.6 is satisfied, and hence the strict complementarity is satisfied which means that $\langle H2\rangle$ in Assumption 4.2 is satisfied. Specifically, we can find an optimal dual solution $\bar{\lambda} = (\bar{\lambda}_1, ..., \bar{\lambda}_6)$ with $\bar{\lambda}_6 = (32.27, 31.71, 0, 0)^T$. Therefore, the weight $w = (100, 100, 1, 1)^T$ satisfies Assumption 4.2.

5. Dual-density-based algorithms

Note that it is difficult to solve a bilevel optimization. We now develop three types of relaxation models for solving the bilevel optimization (22).

5.1. Relaxation models

Zhao and Luo [36] presented a method to relax a bilevel problem similar to (22). Motivated by their idea, we now relax our bilevel model. We focus on relaxing the difficult constraint $-\lambda_1\epsilon - \lambda_2^T b + \lambda_3^T y = \min_x\{\|Wx\|_1 : x \in T\}$ in (22). By replacing the objective function $\|\lambda_6\|_0$ in (22) by $\Psi_\varepsilon(\lambda_6) \in \mathbf{F}$, where $\lambda_6 \geqslant 0$, we obtain an approximation problem of (22), i.e.,

$$\begin{array}{ll} \max\limits_{(w,\lambda)} & \Psi_\varepsilon(\lambda_6) \\ \text{s.t.} & B^T\lambda_2 - A^T\lambda_3 + \lambda_4 - \lambda_5 = 0, \ \|\lambda_3\|_2 \leq \lambda_1 \\ & -\lambda_1\epsilon - \lambda_2^T b + \lambda_3^T y = \min_x\{\|Wx\|_1 : x \in T\}, \\ & w = \lambda_4 + \lambda_5 + \lambda_6 \geq 0, \ \lambda_i \geqslant 0, \ i = 1,2,4,5,6. \end{array} \tag{29}$$

We recall the set of the weights ζ given in (20). It can be seen that w being feasible to (29) implies that (11) and (12) satisfy the strong duality and have the same finite optimal value, which is equivalent to the fact that $w \in \zeta$ when Slater condition holds for (11). Moreover, note that the constraints of (29) indicate that for any given $w \in \zeta$, λ satisfying the constraints of (29) is optimal to (12). Therefore the purpose of (29) is to find the densest dual optimal variable λ_6 for all $w \in \zeta$. Thus (29) can be rewritten as

$$\begin{array}{ll} \max\limits_{(w,\lambda_6)} & \Psi_\varepsilon(\lambda_6) \\ \text{s.t.} & w \in \zeta, \ B^T\lambda_2 - A^T\lambda_3 + \lambda_4 - \lambda_5 = 0, \ \|\lambda_3\|_2 \leq \lambda_1, \\ & w = \lambda_4 + \lambda_5 + \lambda_6 \geq 0, \ \lambda_i \geq 0, \ i = 1,2,4,5,6, \\ & \text{where } \lambda_i, i = 1,2,...,5 \text{ is optimal to} \\ & \max_\lambda\{-\lambda_1\epsilon - \lambda_2^T b + \lambda_3^T y : \|\lambda_3\|_2 \leq \lambda_1, \ w = \lambda_4 + \lambda_5 + \lambda_6, \\ & B^T\lambda_2 - A^T\lambda_3 + \lambda_4 - \lambda_5 = 0, \ \lambda_i \geqslant 0, \ i = 1,2,4,5,6\}. \end{array} \tag{30}$$

Denote the feasible set of (12) by

$$\begin{aligned} D(w) :=& \{\lambda : B^T\lambda_2 - A^T\lambda_3 + \lambda_4 - \lambda_5 = 0, \ \|\lambda_3\|_2 \leq \lambda_1, \ w = \lambda_4 + \lambda_5 + \lambda_6 \geq 0, \\ & \lambda_i \geq 0, \ i = 1,2,4,5,6\}. \end{aligned} \tag{31}$$

Clearly, the problem (30) can be presented as

$$\begin{array}{ll} \max\limits_{(w,\lambda)} & \Psi_\varepsilon(\lambda_6) \\ \text{s.t.} & w \in \zeta, \ \lambda \in D(w), \ \text{where } \lambda \text{ is optimal to} \\ & \max_\lambda\{-\lambda_1\epsilon - \lambda_2^T b + \lambda_3^T y : \lambda \in D(w)\}. \end{array} \tag{32}$$

An optimal solution of (32) can be obtained by maximizing $\Psi_\varepsilon(\lambda_6)$ which is based on maximizing $-\lambda_1\epsilon - \lambda_2^T b + \lambda_3^T y$ over the feasible set of (32). Therefore, $\Psi_\varepsilon(\lambda_6)$ and $-\lambda_1\epsilon - \lambda_2^T b + \lambda_3^T y$ are required to be maximized over the dual constraints $\lambda \in D(w)$ for all $w \in \zeta$. To maximize both the objective functions, we consider the following model as the first relaxation of (30):

$$\begin{array}{ll} \max\limits_{(w,\lambda)} & -\lambda_1\epsilon - \lambda_2^T b + \lambda_3^T y + \alpha\Psi_\varepsilon(\lambda_6) \\ \text{s.t.} & w \in \zeta, \ \lambda \in D(w). \end{array} \tag{33}$$

where $\alpha > 0$ is a given small parameter.

Now we develop the second type of relaxation of the bilevel optimization (22). Note that under Slater condition, for all $w \in \zeta$, the dual objective $-\lambda_1\epsilon - \lambda_2^T b + \lambda_3^T y$ must be nonnegative and is homogeneous in $\lambda = (\lambda_1, ..., \lambda_6)$. Moreover, if $w \in \zeta$, then $-\lambda_1\epsilon - \lambda_2^T b + \lambda_3^T y$ has a nonnegative upper bound due to the weak duality. Inspired by this observation, in order to maximize both $\Psi_\varepsilon(\lambda_6)$ and $-\lambda_1\epsilon - \lambda_2^T b + \lambda_3^T y$, we may introduce a small positive α and consider the following approximation:

$$\begin{array}{ll} \max\limits_{(w,\lambda)} & -\lambda_1\epsilon - \lambda_2^T b + \lambda_3^T y \\ \text{s.t.} & w \in \zeta,\ \lambda \in D(w),\ \ -\lambda_1\epsilon - \lambda_2^T b + \lambda_3^T y \le \alpha\Psi_\varepsilon(\lambda_6). \end{array} \tag{34}$$

The constraint

$$-\lambda_1\epsilon - \lambda_2^T b + \lambda_3^T y \le \alpha\Psi_\varepsilon(\lambda_6) \tag{35}$$

implies that $\Psi_\varepsilon(\lambda_6)$ might be maximized when $-\lambda_1\epsilon - \lambda_2^T b + \lambda_3^T y$ is maximized if α is small and suitably chosen.

Finally, we consider the following inequality in order to develop third type of convex relaxation.

$$-\lambda_1\epsilon - \lambda_2^T b + \lambda_3^T y + f(\lambda_6) \leqslant \gamma, \tag{36}$$

where γ is a given positive number, $f(\lambda_6)$ is a certain function depending on $\varphi_\varepsilon((\lambda_6)_i)$, which satisfies the following properties:

- $(I1)$ $f(\lambda_6)$ is convex and continuous with respect to $\lambda_6 \in R^n_+$;
- $(I2)$ maximizing $\Psi_\varepsilon(\lambda_6)$ over the feasible set can be equivalently or approximately achieved by minimizing $f(\lambda_6)$.

There are many functions satisfying the properties $(I1)$ and $(I2)$. For instance, we may consider the following functions:

$(J1)$ $e^{-\Psi_\varepsilon(\lambda_6)}$; $(J2)$ $-\log(\Psi_\varepsilon(\lambda_6) + \sigma_2)$; $(J3)$ $\frac{1}{\Psi_\varepsilon(\lambda_6)+\sigma_2}$; $(J4)$ $\frac{1}{n}\sum_{i=1}^n \frac{1}{\varphi_\varepsilon((\lambda_6)_i)+\sigma_2}$,

where σ_2 is a small positive number. Now we claim that the functions $(J1)$-$(J4)$ satisfy $(I1)$ and $(I2)$. Clearly, the functions $(J1), (J2)$ and $(J3)$ satisfy $(I2)$. Note that

$$\frac{1}{\Psi_\varepsilon(\lambda_6) + \sigma_2} \leqslant \frac{1}{n}\sum_{i=1}^n \frac{1}{\varphi_\varepsilon((\lambda_6)_i) + \sigma_2}.$$

Thus the minimization of $\frac{1}{n}\sum_{i=1}^n \frac{1}{\varphi_\varepsilon((\lambda_6)_i)+\sigma_2}$ is likely to imply the minimization of $\frac{1}{\Psi_\varepsilon(\lambda_6)}$, which means the maximization of $\Psi_\varepsilon(\lambda_6)$. It is easy to check that the functions $(J1)$-$(J4)$ are continuous in $\lambda_6 \geqslant 0$. It is also easy to check that $(J1)$-$(J3)$ are convex for $\lambda_6 \geqslant 0$. Note that for any $\varphi_\varepsilon((\lambda_6)_i) > -\sigma_2 > 0,\ i = 1, ..., n$, all functions $\frac{1}{\varphi_\varepsilon((\lambda_6)_i)+\sigma_2}$ are convex. Therefore their sum is convex for $\lambda_6 \geqslant 0$ as well. Thus all functions $(J1)$-$(J4)$ satisfy the two properties $(I1)$ and $(I2)$. Moreover, the functions $(J1), (J3), (J4)$ have finite values even when $(\lambda_6)_i \to \infty$.

Replacing $-\lambda_1\epsilon - \lambda_2^T b + \lambda_3^T y \le \alpha\Psi_\varepsilon(\lambda_6)$ in (34) by (36) leads to the model

$$\begin{array}{ll} \max\limits_{(w,\lambda)} & -\lambda_1\epsilon - \lambda_2^T b + \lambda_3^T y \\ \text{s.t.} & w \in \zeta,\ \lambda \in D(w),\ \ -\lambda_1\epsilon - \lambda_2^T b + \lambda_3^T y + f(\lambda_6) \leqslant \gamma. \end{array} \tag{37}$$

Clearly, the convexity of $f(\lambda_6)$ guarantees that (37) is a convex optimization. Moreover,

(36) and the property $(I2)$ of $f(\lambda_6)$ imply that maximizing $-\lambda_1\epsilon-\lambda_2^T b+\lambda_3^T y$ is roughly equivalent to minimizing $f(\lambda_6)$ over the feasible set, and thus maximizing $\Psi_\varepsilon(\lambda_6)$. The properties $(I1)$ and $(I2)$ ensure that the problem (37) is computationally tractable and is a certain relaxation of (32).

5.2. One-step dual-density-based algorithm

Note that the set ζ has no explicit form, and we need to deal with the set ζ to solve three relaxation problems (33), (34) and (37). First we relax $w \in \zeta$ to $w \in R^n_+$ and obtain three convex minimization models. In this case, the difficulty for solving the problems (33) and (34) is that $\Psi_\varepsilon(\lambda_6)$ might attain an infinite value when $w_i \to \infty$. We may introduce a bounded merit function $\Psi \in \mathrm{F}$ into (33) and (34) so that the value of $\Psi_\varepsilon(\lambda_6)$ is finite. Moreover, to avoid the infinite optimal value in the model (33), $w \in \zeta$ can be relaxed to $-\lambda_1\epsilon - \lambda_2^T b + \lambda_3^T y \leq 1$. Based on the above observation, we obtain a solvable relaxation for (33) and (34) respectively as follows:

$$\begin{array}{cl} \max\limits_{(w,\lambda)} & -\lambda_1\epsilon - \lambda_2^T b + \lambda_3^T y + \alpha\Psi_\varepsilon(\lambda_6) \\ \text{s.t.} & w \in R^n_+,\ \lambda \in D(w),\ -\lambda_1\epsilon - \lambda_2^T b + \lambda_3^T y \leq 1. \end{array} \tag{38}$$

and

$$\begin{array}{cl} \max\limits_{(w,\lambda)} & -\lambda_1\epsilon - \lambda_2^T b + \lambda_3^T y \\ \text{s.t.} & w \in R^n_+,\ \lambda \in D(w),\ -\lambda_1\epsilon - \lambda_2^T b + \lambda_3^T y \leq \alpha\Psi_\varepsilon(\lambda_6). \end{array} \tag{39}$$

Due to the constraints (36), the optimal value of the problem (37) is finite if it is feasible. By replacing ζ by R^n_+ in (37) , we also obtain a new relaxation of (22):

$$\begin{array}{cl} \max\limits_{(w,\lambda)} & -\lambda_1\epsilon - \lambda_2^T b + \lambda_3^T y \\ \text{s.t.} & w \in R^n_+,\ \lambda \in D(w),\ -\lambda_1\epsilon - \lambda_2^T b + \lambda_3^T y + f(\lambda_6) \leqslant \gamma. \end{array} \tag{40}$$

Thus, a new weighted ℓ_1-algorithm for the model (1) is developed:

Algorithm 2 One-step dual-density-based algorithm [**DDA** for short]

Input:
merit function $\Psi_\varepsilon \in \mathbf{F}$, matrices $A \in R^{m\times n}$ and $B \in R^{l\times n}$;
vectors $y \in R^m$ and $b \in R^l$, small positive parameters $(\varepsilon, \epsilon) \in R^2_{++}$;
Step:
1. Solve the problem (38) or (39) or (40) to obtain (w^0, λ_6^0),
2. Let $x^0 \in \operatorname{argmin}\{(w^0)^T|x| : x \in T\}$.

The performance of this algorithm is demonstrated in Section 6.

5.3. Dual-density-based reweighted ℓ_1-algorithm

Now we develop reweighted ℓ_1-algorithms for (1) based on (34). To this need, we introduce a bounded convex set $\mathcal{W}$ for w to approximate the set ζ. By replacing ζ

with $\mathcal{W}$ in the models (33), (34) and (37), we obtain the following three types of convex relaxation models of (32):

$$\begin{array}{ll} \max\limits_{(w,\lambda)} & -\lambda_1\epsilon - \lambda_2^T b + \lambda_3^T y + \alpha\Psi_\varepsilon(\lambda_6) \\ \text{s.t.} & w \in \mathcal{W},\ \lambda \in D(w),\ -\lambda_1\epsilon - \lambda_2^T b + \lambda_3^T y \leq 1, \end{array} \tag{41}$$

$$\begin{array}{ll} \max\limits_{(w,\lambda)} & -\lambda_1\epsilon - \lambda_2^T b + \lambda_3^T y \\ \text{s.t.} & w \in \mathcal{W},\ \lambda \in D(w),\ -\lambda_1\epsilon - \lambda_2^T b + \lambda_3^T y \leq \alpha\Psi_\varepsilon(\lambda_6), \end{array} \tag{42}$$

$$\begin{array}{ll} \max\limits_{(w,\lambda)} & -\lambda_1\epsilon - \lambda_2^T b + \lambda_3^T y \\ \text{s.t.} & w \in \mathcal{W}, \lambda \in D(w),\ -\lambda_1\epsilon - \lambda_2^T b + \lambda_3^T y + f(\lambda_6) \leqslant \gamma. \end{array} \tag{43}$$

Inspired by [34] and [36], we can choose the following bounded convex set:

$$\mathcal{W} = \left\{ w \in R_+^n : (x^0)^T w \leq M, 0 \leq w \leq M^* e \right\}, \tag{44}$$

where x^0 is the initial point, which can be the solution of the ℓ_1-minimization (3), and M, M^* are two given numbers such that $1 \leqslant M \leqslant M^*$. We also consider the set

$$\mathcal{W} = \left\{ w \in R_+^n :\ w_i \leqslant \frac{M}{|x_i^0| + \sigma_1} \right\}, \tag{45}$$

where both M and σ_1 are two given positive numbers. $(x^0)^T w \leq M$ in (44) and $w_i \leqslant \frac{M}{|x_i^0|+\sigma_1}$ in (45) are motivated by the idea of existing reweighted algorithm in [7, 34, 36]. The set $\mathcal{W}$ can be seen as not only a relaxation of ζ, but also being used to ensure the boundedness of $\Psi_\varepsilon(\lambda_6)$. Based on (44) and (45), we update $\mathcal{W}$ in the algorithms either as:

$$\mathcal{W}^k = \left\{ w \in R_+^n :\ (x^{k-1})^T w \leq M,\ 0 \leq w \leq M^* e \right\}, \tag{46}$$

or

$$\mathcal{W}^k = \left\{ w \in R_+^n :\ w_i \leqslant \frac{M}{|x_i^{k-1}| + \sigma_1} \right\}. \tag{47}$$

This yields the following algorithm (DRA for short).

Algorithm 3 Dual-density-based reweighted ℓ_1-algorithm [**DRA**] for short

Input:

merit function $\Psi_\varepsilon \in \mathbf{F}$, matrices $A \in R^{m\times n}$ and $B \in R^{l\times n}$;
vectors $y \in R^m$ and $b \in R^l$, small positive parameters $(\varepsilon, \epsilon) \in R^2_{++}$;
the iteration index k, the largest number of iteration $k_{\max}$;

Initialization:

1. Solve the problem (43) or (39) or (40) to get w^0;
2. Solve the weighted ℓ_1-minimization $\min\{(w^0)^T|x| : x \in T\}$ to get x^0 and $\mathcal{W}^1$.

Main step:

At the current iterate x^{k-1},
1. solve the problem (41) or (42) or (43) with $\mathcal{W}^k$ to obtain (w^k, λ_6^k),
2. solve the ℓ_1-minimization $\min\{(w^k)^T|x| : x \in T\}$ to get the vector x^k;
3. Update $\mathcal{W}^{k+1}$ and repeat the above iteration until $k = k_{\max}$ (or certain other stopping criterion is met).

The initial step of DDA is to solve DRA and to get the initial weight w^0 and the set $\mathcal{W}^1$. Different choice of the dual weighted and reweighted ℓ_1-minimization problem and the set $\mathcal{W}$ yields different forms of DRA. In this paper, we consider the following forms of DRA(I)-DRA(VI). The corresponding constants, $\mathcal{W}$, DDA and the dual-density-based reweighted ℓ_1-minimization for these algorithms are listed in the following table.

Table 1.: DRA(I)-DRA(VI)

Name	Constants	DDA	$\mathcal{W}$	dual-density-based reweighted problem
DRA(I)	α, M, M^*	DDA(I)	(46)	(41)
DRA(II)	α, σ_1, M	DDA(I)	(47)	(41)
DRA(III)	α, M, M^*	DDA(II)	(46)	(42)
DRA(IV)	α, σ_1, M	DDA(II)	(47)	(42)
DRA(V)	γ, M, M^*	DDA(III)	(46)	(43)
DRA(VI)	γ, σ_1, M	DDA(III)	(47)	(43)

Notice that w is restricted in the bounded set $\mathcal{W}$ so that the optimal value of (41) cannot be infinite. Therefore, we can use the bounded or unbounded merit functions in $\Psi \in \mathbf{F}$, for example, (5), (6), (7) and (8). In addition, M can not be too small. If M is a sufficiently small positive number, there might be a gap between the maximum of $-\lambda_1\epsilon - \lambda_2^T b + \lambda_3^T y$ and the maximum of $\Psi_\varepsilon(\lambda_6)$ over the feasible set.

The existing reweighted ℓ_1-algorithm, PRA, always need an initial iterate, which is often obtained by solving a simple ℓ_1-minimization. Unlike these existing methods, DRA(I)-DRA(VI) can create an initial iterate by themselves. All developed algorithms are based on the relaxation of the set ζ and the choice of merit functions.

6. Numerical experiments

In this section, by choosing proper parameters and merit functions, the performance of the dual-density-based reweighted ℓ_1-algorithms DRA(I)-DRA(VI) will be demonstrated. We use the random examples of convex sets T in our experiments. We first set the noise level ϵ and the parameter ε of merit functions. The sparse vector x^* and the entries of A and B (if B is not deterministic) are generated from Gaussian random variables with zero mean and unit variance. For each generated (x^*, A, B), we set y and b as follows:

$$y = Ax^* + \frac{c_1 \epsilon}{\|c\|_2} c, \;\; Bx^* + d = b, \tag{48}$$

where $d \in R^l_+$ is generated as absolute Gaussian random variables with zero mean and unit variance, and $c_1 \in R$ and $c \in R^m$ are generated as Gaussian random variables with zero mean and unit variance. Then the convex set T is generated, and all examples of T are generated this way. We use

$$\left\|x' - x^*\right\| / \|x^*\| \leqslant 10^{-5} \tag{49}$$

as our default stopping criterion where x' is the solution found by the algorithm, and one success is counted as long as (49) is satisfied. In our experiments, we make 200 random examples for each sparsity level. All the algorithms are implemented in Matlab 2018a, and all the convex problems are solved by CVX (Grant and Boyd [17]).

To demonstrate the performance of the dual-density-based reweighted ℓ_1-algorithms listed in Table 1, we mainly consider the two cases in our experiments

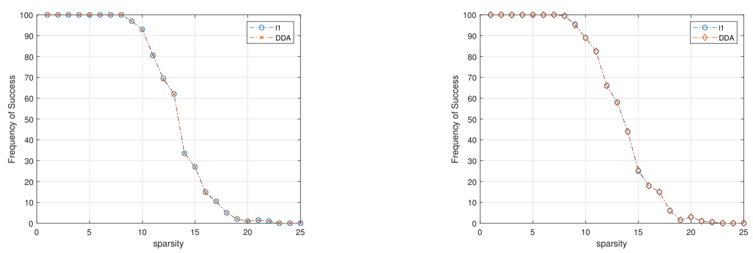
(N1) $B = 0$ and $b = 0$ (that is the model (C2));
(N2) $B \in R^{50\times 200}$.

For all cases, we implement the algorithms DRA(I)- DRA(VI), and compare their performance in finding the sparse vectors in T with ℓ_1-minimization and the algorithm PRA with different merit functions. Before that we test the performance of one-step Dual-density-based algorithm and compare with the ℓ_1-minimization.

We choose (5) and (6) for DDA(II) and $\varphi_\varepsilon((\lambda_6)_i) = \frac{(\lambda_6)_i}{(\lambda_6)_i+\varepsilon}$, $(\lambda_6)_i \in R_+$ in $f(\lambda_6)$ for DDA(III). By setting the parameters

$$(m, n, \epsilon, \varepsilon, \alpha, \gamma, \sigma_2) = (50, 200, 10^{-4}, 10^{-5}, 10^{-5}, 1, 1)$$

and performing 200 random examples for each sparsity level (ranged from 1 to 25), we carry out the experiments for DDA(II) with (6), and DDA(III) with $(J3)$, and compare their performances with ℓ_1-minimization, which is shown in Figure 1:



(i) $\Psi_\varepsilon(\lambda_6) = \sum_{i=1}^{n} \frac{(\lambda_6)_i}{(\lambda_6)_i+\varepsilon}$ (v) $f(\lambda_6) = \frac{1}{\Psi_\varepsilon(\lambda_6)+\sigma_2}$

Figure 1.: The performance of DDA(I) and DDA(II) in finding the sparsest points

Clearly, in this case, the performance of these algorithms are quit similar to that of ℓ_1-minimization (3).

6.1. Merit functions and parameters

The default parameters and merit functions in DRA(I) and DRA(II) are set as that of the algorithms in [36]. We set (6) as the default merit function for DRA(III) and DRA(IV), and set $(J2)$ with

$$f(\lambda_6) = \frac{1}{\Psi_\varepsilon(\lambda_6) + \sigma_2}, \ \Psi_\varepsilon(\lambda_6) = \sum_{i=1}^{n} \frac{(\lambda_6)_i}{(\lambda_6)_i + \varepsilon}, \ \lambda_6 \in R_+^n \tag{50}$$

as the default function for DRA(V) and DRA(VI). We also set $\sigma_2 = 10^{-1}$ as a default parameter. The default parameters for each dual-density-based reweighted ℓ_1-algorithm are summarized in the following table:

Table 2.: Default parameters in each dual-density-based reweighted ℓ_1-algorithm

Algorithm/Parameter	α	γ	M	M^*	σ_1	ε
DRA(I)	10^{-8}		10^2	10^3		10^{-15}
DRA(II)	10^{-8}		10^2		10^{-1}	10^{-15}
DRA(III)	10^{-5}		10	10		10^{-5}
DRA(IV)	10^{-5}		10		10^{-1}	10^{-5}
DRA(V)		1	10	10		10^{-5}
DRA(VI)		10^3	10		10^{-1}	10^{-5}

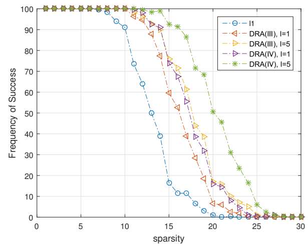
The algorithms in the following table will be compared to DRA(I)-DRA(VI).

Table 3.: Algorithms to be compared

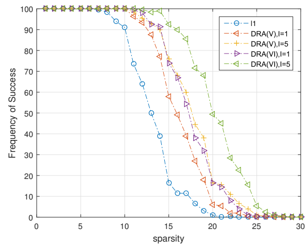
Name	Merit Function	reweighted Methods
ℓ_1	$\lVert x \rVert_1$	ℓ_1-minimization
CWB	$\sum_{i=1}^n \log(\lvert x_i \rvert + \varepsilon)$	PRA
ARCTAN	(8)	PRA

ε in the above PRA algorithms is set to 10^{-1}, and the remaining parameters are the same as DRA. We choose the noisy level $\epsilon = 10^{-4}$ for both cases.

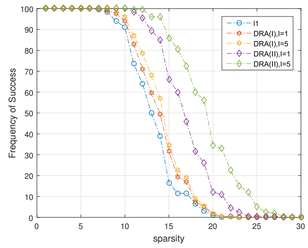
6.2. Case (N1)*:* $B = 0$ *and* $b = 0$



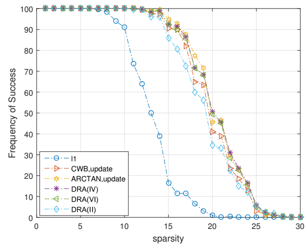
(i) DRA(III) and DRA(IV)



(ii) DRA(V) and DRA(VI)



(iii) DRA(I) and DRA(II)



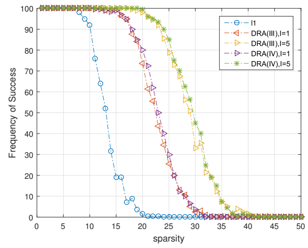
(iv) CWB, ARCTAN with $\varepsilon = 0.1$

Figure 2.: (i)-(iii) Comparison of the performance of the dual-density-based reweighted algorithms by performing 1 iteration and 5 iterations respectively. (iv) Comparison of DRA and PRA.

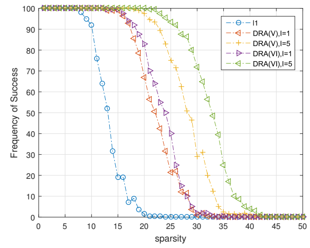
Now we perform numerical experiments to show the behaviors of the dual-density-based re-weighted ℓ_1-algorithms in two cases (N1) and (N2). Note that in the case of (N1), the model (1) is reduced to the sparse model (C2). The numerical results are given in Figure 2 (i)-(iii), Note that there are five legends in each figure (i)-(iii), corresponding to ℓ_1-minimization, the dual-density-based reweighted ℓ_1-algorithms with one iteration or five iterations. For instance, in (i), we compare DRA(III) and DRA(IV) which all perform either one iteration or five iterations. For example, (DRA(III),1) and (DRA(III),5) represent DRA(III) with one iteration and five iterations, respectively.

It can be seen that the dual-density-based reweighted algorithms are performing better when the number of iteration is increased and all of them outperform ℓ_1-minimization, while the performance of DRA(I) with one or five iterations is similar to the performance of ℓ_1-minimization. (i)-(iii) indicate the same phenomena: the algorithms based on (47) might achieve more improvement than the ones based on (46) when the number of iteration is increased. For example, in (ii), the success rate of DRA(VI) with five iterations has improved by nearly 25% compared with those with one iteration for each sparsity from 14 to 20, while DRA(V) has only improved its performance by 10% after increasing the number of iterations. We filter the algorithms with the best performance from (i)-(iii) in Figure 2 and merge them into Figure (iv) in 2. It can be seen that DRA(II), DRA(IV) and DRA(VI) slightly outperform CWB with $\varepsilon = 0.1$ and ARCTAN with $\varepsilon = 0.1$. The CWB is one of the efficient choices for the existing reweighted algorithms.

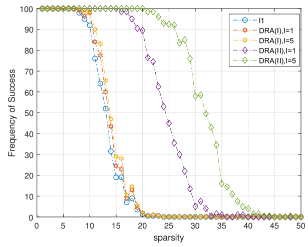
6.3. Case (N2)*:* $B \in R^{50\times 200}$



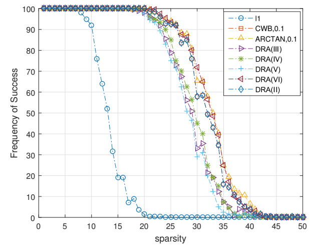
(i) DRA(III) and DRA(IV)



(ii) DRA(V) and DRA(VI)



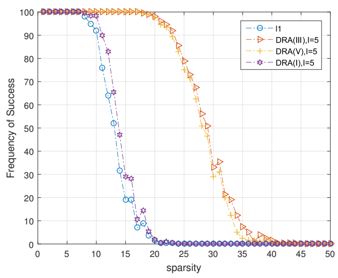
(iii) DRA(I) and DRA(II)



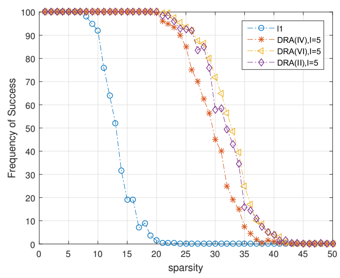
(iv) CWB, ARCTAN with $\varepsilon = 0.1$

Figure 3.: (i)-(iii) Comparison of the performance of DRA with one iteration and five iterations. (iv) Comparison of the performance of the DRA and PRA.

We compare the reweighted ℓ_1-algorithms with updating rule (46) and (47), which are shown in (i) and (ii) in Figure 4, respectively. For the algorithms using (46), when executing 5 iterations, Figure 4 (i) shows that DRA(III) and DRA(V) perform much better than DRA(I). For the algorithms using (47), when executing 5 iterations, Figure 4 (ii) indicates that the success rates of finding the sparse vectors in T by DRA(II) and DRA(VI) are very similar. The other behaviors are similar to the case of $B = 0$ and $b = 0$.



(i) Algorithms with rule (46)



(ii) Algorithms with rule (47)

Figure 4.: Comparison of the performance of DRA with (46) or (47)

Finally, we carry out experiment to show how the parameter ε of merit functions affect the performance of locating the sparse vectors in T by dual-density-based reweighted ℓ_1-algorithms. Some numerical results for PRA-typed algorithms and dual-density-based reweighted algorithms with different ε indicate that the performance of the DRA-typed algorithm is relatively insensitive to the choice of small ε compared to the PRA-typed algorithms.

7. Conclusions

In this paper, we have studied a class of algorithms for the ℓ_0-minimization problem (1). The one-step dual-density-based algorithms (DDA) and the dual-density-based reweighted ℓ_1-algorithms (DRA) are developed. These algorithms are developed based on the new relaxation of the equivalent bilevel optimization of the underlying ℓ_0-minimization problem. Unlike PRA, the DRA can automatically generate an initial iterate instead of obtaining the initial iterate by solving ℓ_1-minimization. Numerical experiments show that in some cases such as (N1) and (N2), the dual-density-based methods proposed in this paper can perform better than ℓ_1-minimization in solving the sparse optimization problem (1), and can be comparable to some existing reweighted ℓ_1-methods.

References

[1] M. S. Asif and J. Romberg, *Fast and accurate algorithms for re-Weighted ℓ_1-norm minimization*, IEEE Transaction on Signal Processing. 61 (2013), pp. 5905–5916.

[2] M. S. Asif and J. Romberg, *Sparse recovery of streaming signals using ℓ_1-homotopy*, IEEE Transaction on Signal Processing. 62 (2014), pp. 4209–4223.

[3] T. Blumensath and M. Davies and G. Rilling, *Greedy algorithms for compressed sensing*, Compressed Sensing: Theory and Applications 2012, pp. 348–393.

[4] E. Candès, *compressive sampling*, Proceedings of the International Congress of Mathematicians 3 (2006), pp. 1433–1452.

[5] E. Candès, J. Romberg and T. Tao, *Stable signal recovery from incomplete and inaccurate measurements*, Communications on Pure and Applied Mathematics 59 (2006), pp. 1207–1223.

[6] E. Candès and T. Tao, *Decoding by linear programming*, IEEE Transactions on Information Theory 51 (2005), pp. 4203–4215.

[7] E. Candès and M. Wakin and S. Boyd, *Enhancing sparsity by reweighted ℓ_1 minimization*, Journal of Fourier Analysis and Applications 14 (2008), pp. 877–905.

[8] X. Chen and W. Zhou, *Convergence of reweighted l1 minimization algorithms and unique solution of truncated lp minimization*, Department of Applied Mathematics, The Hong Kong Polytechnic University, 2010.

[9] W. Dai, and O. Milenkovic, *Subspace pursuit for compressive sensing signal reconstruction*, IEEE Transactions on Information Theory 55 (2009), pp. 2230–2249.

[10] I. Daubechies and M. Defrise and C. D. Mol, *An iterative thresholding algorithm for linear inverse problems with a sparsity constraint*, Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences 57 (2004), pp. 1413–1457.

[11] M. A. Davenport and M. F. Duarte and Y. C. Eldar and G. Kutyniok, *Compressed Sensing: Theory and Applications*, Cambridge University Press, 2011.

[12] D. Donoho, *Compressed sensing*, IEEE Transactions on Information Theory 52 (2006), pp. 1289–1306.

[13] M. Elad, *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*, Springer, NY, 2010.

[14] S. Foucart and M. Lai, *Sparsest solutions of underdetermined linear systems via ℓ_q-minimization for* $0 < q < 1$, Applied and Computational Harmonic Analysis 26 (2009), pp. 395–407.

[15] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*, Springer, NY, 2013.

[16] I. F. Gorodnitsky and J. George and B. Rao, *Neuromagnetic source imaging with FOCUSS: a recursive weighted minimum norm algorithm*, Electroencephalography and Clinical Neurophysiology 95 (1995), pp. 231–251.

[17] M. Grant and S. Boyd, *CVX: Matlab software for disciplined convex programming*, Version 2.1, 2017.

[18] A. Gupta and R. Nowak and B. Recht, *Sample complexity for 1-bit compressed sensing and sparse classification*, IEEE International Symposium on Information Theory 2010, pp. 1553–1557.

[19] G. Harikumar and Y. Bresler, *A new algorithm for computing sparse solutions to linear inverse problems*, Acoustics, Speech, and Signal Processing, 1996 (ICASSP-96) 3 (1996), pp. 1331–1334.

[20] H. Hoefling, *A path algorithm for the fused lasso signal approximator*, Journal of Computational and Graphical Statistics 19 (2010), pp. 984–1006.

[21] M. Lai and J. Wang, *An unconstrained ℓ_q minimization with $0 < q \leqslant 1$ for sparse solution of underdetermined linear systems*, SIAM Journal on Optimization 21 (2010), pp. 82–101.

[22] J. Laska and Z. Wen and W. Yin and R. Baraniuk, *Trust, but verify: Fast and accurate signal recovery from 1-bit compressive measurements*, IEEE Transactions on Signal Processing 59 (2011), pp. 5289–5301.

[23] J. Liu and L. Yuan and J. Ye, *An efficient algorithm for a class of fused lasso problems*, Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2010, pp. 323–332.

[24] S. Mallat and Z. Zhang, *Matching pursuit with time-frequency dictionaries*, Courant Institute of Mathematical Sciences New York United States,1993.

[25] D. Needell and J. A. Tropp, *CoSaMP: Iterative signal recovery from incomplete and inaccurate samples*, Applied and Computational Harmonic Analysis 26 (2009), pp. 301–321.

[26] F. Rinaldi, *Concave programming for finding sparse solutions to problems with convex constraints*, Optimization Methods and Software 26 (2011), pp. 971–992.

[27] A. Rinaldo and others, *Properties and refinements of the fused lasso*, The Annals of Statistics 37 (2009), pp. 2922–2952.

[28] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press (2004).

[29] V. N. Temlyakov, *Nonlinear methods of approximation*, Foundations of Computational Mathematics 3 (2003), pp 33–107.

[30] R. Tibshirani and P. Wang, *Spatial smoothing and hot spot detection for CGH data using the fused lasso*, Biostatistics 9 (2007), pp. 18–29.

[31] R. Tibshirani, M. Wainwright, and T. Hastie, *Statistical Learning with Sparsity: The Lasso and Generalizations*, Chapman and Hall/CRC, 2015.

[32] Y.B. Zhao, *Sparse Optimization Theory and Methods*, CRC Press, Taylor & Francis Group, 2018.

[33] Y.B. Zhao, Optimal k-thresholding algirithms for sparse optimization problems, Technical report, https://arxiv.org/abs/1909.00717.

[34] Y.B. Zhao and M. Kočvara, *A new computational method for the sparsest solutions to systems of linear equations*, SIAM Journal on Optimization 25 (2015), pp. 1110–1134.

[35] Y.B. Zhao and D. Li, *Reweighted ℓ_1-minimization for sparse solutions to underdetermined linear systems*, SIAM Journal on Optimization 22 (2012), pp. 1065–1088.

[36] Y.B. Zhao and Z. Q. Luo, *Constructing new weighted ℓ_1-algorithms for the sparsest points of polyhedral sets*, Mathematics of Operations Research 42 (2017), pp. 57–76.