Constructing New Weighted ℓ_1-Algorithms for the Sparsest Points of Polyhedral Sets

YUN-BIN ZHAO* and ZHI-QUAN LUO†

(First version: June 2013, Final version: December 2015)

Abstract. The ℓ_0-minimization problem that seeks the sparsest point of a polyhedral set is a long-standing challenging problem in the fields of signal and image processing, numerical linear algebra and mathematical optimization. The weighted ℓ_1-method is one of the most plausible methods for solving this problem. In this paper, we develop a new weighted ℓ_1-method through the strict complementarity theory of linear programs. More specifically, we show that locating the sparsest point of a polyhedral set can be achieved by seeking the densest possible slack variable of the dual problem of weighted ℓ_1-minimization. As a result, ℓ_0-minimization can be transformed, in theory, to ℓ_0-maximization in dual space through some weight. This theoretical result provides a basis and an incentive to develop a new weighted ℓ_1-algorithm, which is remarkably distinct from existing sparsity-seeking methods. The weight used in our algorithm is computed via a certain convex optimization instead of being determined locally at an iterate. The guaranteed performance of this algorithm is shown under some conditions, and the numerical performance of the algorithm has been demonstrated by empirical simulations.

Key words. Polyhedral set, sparsest point, weighted ℓ_1-algorithm, convex optimization, sparsity recovery, strict complementarity, duality theory, bilevel programming

AMS subject classifications: 90C05, 90C25, 90C26, 90C59, 52A41, 49M29, 65K05

*School of Mathematics, University of Birmingham, Edgbaston, Birmingham, West Midlands B15 2TT, UK (e-mail: y.zhao.2@bham.ac.uk). The research of this author was supported by the Engineering and Physical Sciences Research Council (EPSRC) under grant #EP/K00946X/1.

†Department of Electrical and Computer Engineering, University of Minnesota, Twin Cites, Minnesota, USA (e-mail: luozq@ece.umn.edu).

1 Introduction

A polyhedral set P in finite-dimensional spaces can be represented as

$$P = \{x \in \mathbb{R}^n : \ Ax = b, \ x \geq 0\}, \tag{1}$$

where $A \in \mathbb{R}^{m\times n}$ is a given matrix and $b \in \mathbb{R}^m$ is a given vector. We assume $m < n$, $b \neq 0$ and $P \neq \emptyset$ throughout, and consider the so-called ℓ_0-minimization problem

$$\min\{\|x\|_0 : \ x \in P\}, \tag{2}$$

where $\|x\|_0$ denotes the number of nonzero components of x. This problem arises in many practical scenarios, including nonnegative signal and image processing [19, 3, 9, 42, 55, 56, 31], machine and statistical learning [35, 6, 7, 36, 20, 47], computer vision and pattern recognition [55, 50], and nonnegative sparse coding and compressed sampling [29, 58, 19, 9, 28, 60]. It is also closely relevant to nonnegative matrix factorization [44, 43]. In fact, seeking sparsity has become a common request in various data processing tasks such as compression, reconstruction, transmission, denoising and separation [8, 20, 49, 21]. Thus developing reliable computational methods for (2) is fundamentally important from both mathematical and practical viewpoints, despite the fact that the problem (2) is NP-hard in general [39, 1].

One avenue for solving ℓ_0-minimization is to use ℓ_1-minimization [12], i.e., min $\{\|x\|_1 : x \in P\}$, which is also called basis pursuit (BP). Heuristic methods such as orthogonal matching pursuit (OMP) [34, 17, 52, 54, 41] and the thresholding algorithm [16, 4, 20, 5] have also been widely exploited in recent years. The guaranteed performance of these algorithms has been analyzed under various conditions. For example, the performance of ℓ_1-minimization can be guaranteed if one of the following conditions (or properties) holds: mutual coherence [18, 20], restricted isometry property (RIP) [10], null space property (NSP) [14, 58, 59] or the so-called verifiable condition in [30, 51]. Moreover, the performance of OMPs can be guaranteed under the mutual coherence condition [18, 20] or the ERC assumption [24, 53, 54]. The mutual-coherence and the RIP-type assumptions can also ensure the success of thresholding methods [20, 5]. Most existing conditions often imply the uniqueness of solutions to ℓ_0- and ℓ_1-minimization. Recently, the guaranteed performance of ℓ_1-minimization has been more broadly interpreted in [61, 60] through a range-space-property-based analysis and by distinguishing the equivalence and strong equivalence between ℓ_0- and ℓ_1-minimization. This analysis shows that ℓ_0-minimization can be solved by ℓ_1-minimization when the range space property (RSP) of A^T holds at a sparsest point of P, irrespective of whether P possesses a unique sparsest point or not.

To find a method that may outperform ℓ_1-minimization, Candès, Wakin and Boyd [11] have proposed a weighted ℓ_1-method and have empirically demonstrated that such a method may outperform ℓ_1-minimization in many situations. It is known that the weighted ℓ_1-method can be developed through the first-order estimation of a continuously differentiable approximation of $\|x\|_0$ (see, e.g., [57, 62]). In particular, based on the first-order approximation of the ℓ_p-quasinorm ($0 < p < 1$), Foucart and Lai [22] proposed another weighted ℓ_1-method, for which a further investigation was made in [32, 13]. Recently, some unified analysis for a family of weighted ℓ_1-algorithms has been provided by Zhao and Li [62]. Other recent studies for weighted ℓ_1-algorithms and their applications can be found, for example, in [40, 45, 38, 2, 48, 63].

Defining a weight by the iterate is a common feature of existing weighted ℓ_1-methods. Although empirical results indicate that the algorithms within this framework may perform very well in many situations [11, 22, 32, 15, 62, 13, 48], some important issues associated with these algorithms, from algorithmic structure to theoretical analysis, have still not been properly addressed. For example, if the sparsity pattern of the current iterate is very different from that of the sparsest points of P, existing weighted ℓ_1-methods might continue to generate the next iterate which admits the same sparsity pattern, causing the algorithm to fail. Also, these methods use the first-order estimation of a concave approximation of $\|x\|_0$. As a result, restrictive conditions (including the RIP, NSP or their variants [22, 32, 62]) are often imposed on the problem in order to ensure the theoretical efficiency of these methods. Most of these conditions imply that ℓ_0- and ℓ_1-minimization are strongly equivalent in the sense that both problems have the same unique solution. This is not always the case, however, since either ℓ_0- or ℓ_1-minimization may have multiple optimal solutions, and the solution of ℓ_1-minimization may not be the sparsest one at all (in which case, ℓ_1-minimization completely fails to solve ℓ_0-minimization). So an immediate question arises: Can we develop a new weighted ℓ_1-algorithm so that it goes beyond the current algorithmic framework, and can its theoretical performance be rigorously shown under some conditions that allow a polyhedral set to possess multiple sparsest points?

In this paper, we work toward addressing the above questions. By using the classic strict complementarity theory of linear programs, we first prove that finding a sparsest point of P amounts to searching for the densest possible slack variable of the dual problem of certain weighted ℓ_1-minimization. In other words, ℓ_0-minimization can be translated, in theory, into ℓ_0-maximization with an implicitly given weight. Thus ℓ_0-minimization can be cast as a bilevel programming problem with a special structure (see Theorem 3.1 for details). This provides a theoretical basis for the development of a new weighted ℓ_1-algorithm, going beyond the existing algorithmic framework. The weight used in this method is computed through convex optimization instead of being defined directly at an iterate generated in the course of the algorithm. As desired, the guaranteed performance of the proposed method for locating the sparsest point of P can be shown under certain conditions which allow a polyhedral set to admit multiple sparsest points. The perturbation theory of linear programs [37] is also employed to carry out such a theoretical analysis. Numerical experiments indicate that in many situations the proposed algorithm remarkably outperforms the normal ℓ_1-minimization when locating the sparsest points of polyhedral sets.

This paper is organized as follows. The connection between the sparsest point and weighted ℓ_1-minimization and a class of merit functions for sparsity are given in Section 2. A new weighted ℓ_1-algorithm for the sparsest points of polyhedral sets is developed in Section 3, and the guaranteed performance of the proposed algorithm is shown in Section 4. Finally, numerical results are reported in Section 5.

2 Preliminary

2.1 Notation

In this paper, $\mathbb{R}^n$ denotes the n-dimensional Euclidean space, $\mathbb{R}^n_+$ the set of nonnegative vectors in $\mathbb{R}^n$, and $\mathbb{R}^n_{++}$ the set of positive vectors in $\mathbb{R}^n$. $x \in \mathbb{R}^n_+$ ($x \in \mathbb{R}^n_{++}$) is also written as $x \geq 0$ ($x > 0$). $e = (1, 1, \ldots, 1)^T \in \mathbb{R}^n$ is the vector of ones, P is the polyhedral set defined by (1), and $W = \text{diag}(w)$ is the diagonal matrix with diagonal entries equal to the components of the vector w. For a set $J \subseteq \{1, 2, \ldots, n\}$, we use $\overline{J}$ to denote the complement of J, i.e., $\overline{J} = \{1, 2, \ldots, n\} \backslash J$. For an index set J and a matrix $A \in \mathbb{R}^{m \times n}$ with columns a_i ($i = 1, \ldots, n$), A_J denotes the submatrix of A with columns a_i, $i \in J$, and A_J^T denotes the transpose of A_J. A similar notation is applied to the subvector of a vector. For vectors $x, y \in \mathbb{R}^n$, $x \geq y$ ($x > y$) means $x - y \in \mathbb{R}^n_+ (x - y \in \mathbb{R}^n_{++})$. Given $x \in \mathbb{R}^n$, let $J_+(x) = \{i : \ x_i > 0\}$. Clearly, $J_+(x)$ coincides with the support of $x \in \mathbb{R}^n_+$.

2.2 Basic properties of the sparsest point of P

Note that for any $x \in P$, $b = Ax = A_{J_+(x)} x_{J_+(x)}$ where $J_+(x) = \{i : \ x_i > 0\}$. Lemma 3.1 in [60] claims that at any sparsest point x of P, the associated matrix $\begin{bmatrix} A_{J_+(x)} \\ e^T_{J_+(x)} \end{bmatrix}$ has full column rank. This result can be further improved, as shown by the following lemma.

Lemma 2.1. *At any sparsest point x of P, the matrix $A_{J_+(x)}$ has full column rank.*

Proof. Let x be a sparsest point in P. Assume that the columns of $A_{J_+(x)}$ are linearly dependent. Then there exists a vector v such that $A_{J_+(x)} v = 0$ and $v \neq 0$. Consider the point $x_{J_+(x)} + \lambda v$ where $\lambda \in \mathbb{R}$. Note that $b = A_{J_+(x)} x_{J_+(x)}$ and $x_{J_+(x)} > 0$. By the definition of v, we see that $A_{J_+(x)} \left(x_{J_+(x)} + \lambda v \right) = b$ holds for any number $\lambda \in \mathbb{R}$. Since $x_{J_+(x)} > 0$ and $v \neq 0$, there exists a number λ^* such that $x_{J_+(x)} + \lambda^* v \geq 0$ and at least one of the components of $x_{J_+(x)} + \lambda^* v$ is equal to zero. Let $\widetilde{x} \in \mathbb{R}^n_+$ be given by $\widetilde{x}_{J_+(x)} = x_{J_+(x)} + \lambda^* v \geq 0$ and $\widetilde{x}_{\overline{J_+(x)}} = x_{\overline{J_+(x)}} = 0$. Then $\widetilde{x} \in P$ and $\|\widetilde{x}\|_0 < \|x\|_0$ (i.e., $\widetilde{x}$ is sparser than x), leading to a contradiction. Thus when x is a sparsest point of P, $A_{J_+(x)}$ must have full column rank. $\square$

Clearly, the sparsest point of P may not be an optimal solution to the ℓ_1-minimization problem $\min\{\|x\|_1 : x \in P\}$. Even if the sparsest point of P is an optimal solution to ℓ_1-minimization, it may not be the only optimal solution to this problem. Thus we consider the weighted ℓ_1-minimization problem

$$\gamma^* = \min_x \{\|Wx\|_1 : \ x \in P\},$$

where $W = \text{diag}(w)$, $w \in \mathbb{R}^n_+$, and γ^* is the optimal value of the problem. Note that replacing w by αw for any $\alpha > 0$ does not affect the optimal solution of the above problem. Particularly, when $\gamma^* \neq 1$, replacing W by W/γ^* yields $1 = \min\{\|(W/\gamma^*)x\|_1 : x \in P\}$. Thus without loss of generality, we consider the following weighted ℓ_1-minimization with $\gamma^* = 1$:

$$1 = \min_x \{\|Wx\|_1 : \ x \in P\}. \tag{3}$$

Throughout the paper, we use $\mathcal{W}$ to denote the set of such weights, i.e.,

$$\mathcal{W} = \left\{ w \in \mathbb{R}^n_+ : \ 1 = \min_x \{\|Wx\|_1 : x \in P\} \right\}, \tag{4}$$

which is not necessarily bounded. It is worth noting that the uniqueness of solutions to weighted ℓ_1-minimization plays a vital role in sparse signal recovery and in solving an individual ℓ_0-minimization problem [8, 20, 21]. If x is a sparsest point of P, then there exists a weight $w \in \mathbb{R}^n_+$ such that x is the unique optimal solution to (3), as indicated by the next result.

Theorem 2.2. *Let x be a point in P such that $A_{J_+(x)}$ has full column rank. Then for any weight $w \in \mathcal{W}$ satisfying*

$$w_{\overline{J_+(x)}} > A^T_{\overline{J_+(x)}} A_{J_+(x)} \left(A^T_{J_+(x)} A_{J_+(x)} \right)^{-1} w_{J_+(x)}, \tag{5}$$

x is the unique optimal solution of (3).

Proof. Suppose that $x \in P$ and $A_{J_+(x)}$ has full column rank. Let $\widetilde{x}$ be an arbitrary point in P. Then $A\widetilde{x} = b$ and $\widetilde{x} \geq 0$, which can be written as $A_{J_+(x)}\widetilde{x}_{J_+(x)} + A_{\overline{J_+(x)}}\widetilde{x}_{\overline{J_+(x)}} = b,\ \widetilde{x}_{J_+(x)} \geq 0$ and $\widetilde{x}_{\overline{J_+(x)}} \geq 0$. This, together with $A_{J_+(x)}x_{J_+(x)} = b$, implies that

$$A_{J_+(x)}(\widetilde{x}_{J_+(x)} - x_{J_+(x)}) + A_{\overline{J_+(x)}}\widetilde{x}_{\overline{J_+(x)}} = 0,\ \widetilde{x}_{J_+(x)} \geq 0,\ \widetilde{x}_{\overline{J_+(x)}} \geq 0.$$

Since $A_{J_+(x)}$ has full column rank, we have

$$\widetilde{x}_{J_+(x)} = x_{J_+(x)} - \left(A_{J_+(x)}^T A_{J_+(x)}\right)^{-1} A_{J_+(x)}^T A_{\overline{J_+(x)}}\widetilde{x}_{\overline{J_+(x)}},\ \widetilde{x}_{J_+(x)} \geq 0,\ \widetilde{x}_{\overline{J_+(x)}} \geq 0.$$

For any $\widetilde{x} \in P$ and $\widetilde{x} \neq x$, the above relation implies that $\widetilde{x}_{\overline{J_+(x)}} \neq 0$. We now verify that under (5), x must be the unique optimal solution to (3). Indeed, for any $\widetilde{x} \in P$ and $\widetilde{x} \neq x$, we have

$$\begin{aligned}
\|Wx\|_1 - \|W\widetilde{x}\|_1 &= w^T x - w^T \widetilde{x} \\
&= w_{J_+(x)}^T x_{J_+(x)} - (w_{J_+(x)}^T \widetilde{x}_{J_+(x)} + w_{\overline{J_+(x)}}^T \widetilde{x}_{\overline{J_+(x)}}) \\
&= \left(w_{J_+(x)}^T \left(A_{J_+(x)}^T A_{J_+(x)}\right)^{-1} A_{J_+(x)}^T A_{\overline{J_+(x)}} - w_{\overline{J_+(x)}}^T\right) \widetilde{x}_{\overline{J_+(x)}} \\
&= \left(A_{\overline{J_+(x)}}^T A_{J_+(x)} \left(A_{J_+(x)}^T A_{J_+(x)}\right)^{-1} w_{J_+(x)} - w_{\overline{J_+(x)}}\right)^T \widetilde{x}_{\overline{J_+(x)}} \\
&< 0,
\end{aligned}$$

where the inequality follows from (5) and the fact that $\widetilde{x}_{\overline{J_+(x)}} \neq 0$. Thus $\|Wx\|_1 < \|W\widetilde{x}\|_1$ for any $\widetilde{x} \in P$ and $\widetilde{x} \neq x$, and hence x is the unique optimal solution to (3). □

To suit the convenience of later discussions, we introduce the following definition.

Definition 2.3. A weight $w \in \mathcal{W}$ given as (4) is called an optimal weight if any solution of (3) with this weight is a sparsest point in P.

Clearly, there exist infinitely many weights satisfying (5). Merging Lemma 2.1 and Theorem 2.2 yields the following corollary.

Corollary 2.4. *For any sparsest point x of P, there exist infinitely many optimal weights in $\mathcal{W}$ such that x is the unique optimal solution of (3).*

Although Corollary 2.4 seems intuitive and some informal discussions might be found in the literature such as [26], we state this property with a rigorous analysis given in Theorem 2.2. Corollary 2.4 indicates that an optimal weight $w \in \mathcal{W}$ always exists and that locating the sparsest point of a polyhedral set can be transformed, in theory, to solving a weighted ℓ_1-problem (a linear program) through an optimal weight. Depending on the sparsest point x, however, the optimal weight satisfying (5) is not given explicitly. How to identify or estimate an optimal weight becomes fundamentally important for tackling ℓ_0-minimization.

2.3 Merit functions for sparsity

It is well known that the function $\|s\|_0$, where $s \in \mathbb{R}^n_+$, can be approximated by a concave function $\mathcal{F}_\varepsilon(s)$, where $\varepsilon \in (0, 1)$ is a given parameter (see, e.g., [35, 6, 36, 45, 62]). For example, $\|s\|_0$ can be approximated by each of the following functions:

$$\mathcal{F}_\varepsilon(s) := n - \frac{1}{\log \varepsilon}\left(\sum_{i=1}^n \log(s_i + \varepsilon)\right), \tag{6}$$

$$\mathcal{F}_\varepsilon(s) := \sum_{i=1}^n \frac{s_i}{s_i+\varepsilon}, \tag{7}$$

where $s_i > -\varepsilon$ for all $i = 1, \ldots, n$, and

$$\mathcal{F}_\varepsilon(s) := \sum_{i=1}^n (s_i + \varepsilon^{1/\varepsilon})^\varepsilon = \left\| s + \varepsilon^{1/\varepsilon} e \right\|_\varepsilon^\varepsilon \tag{8}$$

where $s_i > -\varepsilon^{1/\varepsilon}$ for all $i = 1, \ldots, n$. It is evident that every function above satisfies the following properties: (i) $\mathcal{F}_\varepsilon(s)$ is concave in $s \in \mathbb{R}^n_+$; (ii) for every given $s \in \mathbb{R}^n_+$, $\mathcal{F}_\varepsilon(s) \to \|s\|_0$ as $\varepsilon \to 0$; (iii) $\mathcal{F}_\varepsilon(s)$ is continuously differentiable in s over an open neighborhood of $\mathbb{R}^n_+$. Similar to the concept in [62], the function with the above properties is called a merit function for sparsity in the sense that minimizing such a function promotes the sparsity of its variable. While the above merit functions are defined over the neighborhood of $\mathbb{R}^n_+$, they can be extended to the whole space by replacing s_i with $|s_i|$ in (6)–(8). Thus when restricted to the first orthant, the definition of merit functions for sparsity in [62] is the same as the one discussed here.

In this paper, we focus on a class of merit functions for sparsity satisfying the following assumption.

Assumption 2.5. Let $\mathcal{F}_\varepsilon(s)$, where $\varepsilon > 0$ is a parameter, be concave and continuously differentiable with respect to s over an open neighborhood of $\mathbb{R}^n_+$. Let $0 < \delta_1 < \delta_2$ be two arbitrarily given constants and $\mathfrak{S}(\delta_1, \delta_2) = \{s \in \mathbb{R}^n_+ : \delta_1 \le s_i \le \delta_2 \text{ if } s_i \neq 0\}$. For any fixed number $\delta^* \in (0,1)$, there exists a small number $\varepsilon^* > 0$ such that

$$\mathcal{F}_\varepsilon(s) - \mathcal{F}_\varepsilon(s') \ge 1 - \delta^* > 0 \tag{9}$$

holds for any $\varepsilon \in (0, \varepsilon^*]$, $s \in \mathfrak{S}(\delta_1, \delta_2)$, and any s' with $0 \le s' \le \delta_2 e$ and $\|s'\|_0 < \|s\|_0$.

Roughly speaking, a function satisfying Assumption 2.5 has some monotonicity property in the sense that when $\|s\|_0$ decreases, so does the value of $\mathcal{F}_\varepsilon(s)$. Thus minimizing such a function yields a sparse vector, and maximizing yields a dense vector. It is not difficult to construct a function satisfying Assumption 2.5. In fact, (6), (7) and (8) are such functions, as shown by the lemma below.

Lemma 2.6. *For any $\varepsilon \in (0,1)$, the functions (6)–(8) are concave and continuously differentiable with respect to s over an open neighborhood of $\mathbb{R}^n_+$, and these functions satisfy Assumption 2.5.*

Proof. Clearly, (6)–(8) are concave and continuously differentiable functions over the neighborhood of $\mathbb{R}^n_+$. We now prove that these functions satisfy Assumption 2.5. Let $\delta^* \in (0,1)$ be a fixed constant and let $0 < \delta_1 < \delta_2$ be two arbitrarily given constants. Define the set $\mathfrak{S}(\delta_1, \delta_2) = \{s \in \mathbb{R}^n_+ : \delta_1 \le s_i \le \delta_2 \text{ if } s_i \neq 0\}$. We now show that for each of the functions (6)–(8), there exists a small $\varepsilon^* > 0$ such that (9) holds for any $\varepsilon \in (0, \varepsilon^*]$ and any nonnegative vectors s, s' satisfying that $s \in \mathfrak{S}(\delta_1, \delta_2)$, $0 \le s' \le \delta_2 e$ and $\|s'\|_0 < \|s\|_0$.

(i) Consider the function (6). Note that $\log \varepsilon < 0$ for any $\varepsilon \in (0,1)$. Then for any $t \in [\delta_1, \delta_2]$ and $\varepsilon \in (0,1)$, we have

$$\log(\delta_2 + \varepsilon)/\log \varepsilon \le \log(t + \varepsilon)/\log \varepsilon \le \log(\delta_1 + \varepsilon)/\log \varepsilon, \tag{10}$$

and for any $t \in [0, \delta_1)$ and $\varepsilon \in (0,1)$ we have

$$\log(\delta_1+\varepsilon)/\log\varepsilon \leq \log(t+\varepsilon)/\log\varepsilon \leq 1. \tag{11}$$

Note that $\log(\delta_1+\varepsilon)/\log\varepsilon \to 0$ and $\log(\delta_2+\varepsilon)/\log\varepsilon \to 0$ as $\varepsilon \to 0^+$. There exists $\varepsilon^* \in (0,1)$ such that for any $\varepsilon \in (0, \varepsilon^*]$,

$$|\log(\delta_1+\varepsilon)/\log\varepsilon| \leq \delta^*/3n, \quad |\log(\delta_2+\varepsilon)/\log\varepsilon| \leq \delta^*/3n. \tag{12}$$

Thus for any $s \in \mathfrak{S}(\delta_1, \delta_2)$ and $\varepsilon \in (0, \varepsilon^*]$, it follows from (10) and (12) that

$$\left|\left(\sum_{s_i>0} \log(s_i+\varepsilon)\right)/\log\varepsilon\right| \leq \|s\|_0 (\frac{\delta^*}{3n}) \leq \frac{\delta^*}{3},$$

which implies that for any given $\varepsilon \in (0, \varepsilon^*]$ and $s \in \mathfrak{S}(\delta_1, \delta_2)$,

$$\begin{aligned} \mathcal{F}_\varepsilon(s) &= n - \left[\left(\sum_{s_i>0} \log(s_i+\varepsilon)\right)/\log\varepsilon + \left(\sum_{s_i=0} \log(s_i+\varepsilon)\right)/\log\varepsilon\right] \\ &= \|s\|_0 - \left(\sum_{s_i>0} \log(s_i+\varepsilon)\right)/\log\varepsilon \geq \|s\|_0 - \frac{\delta^*}{3}. \end{aligned} \tag{13}$$

We now consider the value of $\mathcal{F}_\varepsilon$ at s' where $0 \leq s' \leq \delta_2 e$. Then

$$\begin{aligned} \mathcal{F}_\varepsilon(s') &= n - \sum_{s'_i=0} \frac{\log(s'_i+\varepsilon)}{\log\varepsilon} - \sum_{s'_i\in(0,\delta_1)} \frac{\log(s'_i+\varepsilon)}{\log\varepsilon} - \sum_{s'_i\in[\delta_1,\delta_2]} \frac{\log(s'_i+\varepsilon)}{\log\varepsilon} \\ &= \|s'\|_0 - \sum_{s'_i\in(0,\delta_1)} \frac{\log(s'_i+\varepsilon)}{\log\varepsilon} - \sum_{s'_i\in[\delta_1,\delta_2]} \frac{\log(s'_i+\varepsilon)}{\log\varepsilon}. \end{aligned}$$

By (10), (11) and (12), we can see that

$$\left|\sum_{s'_i\in[\delta_1,\delta_2]} \frac{\log(s'_i+\varepsilon)}{\log\varepsilon}\right| \leq \frac{\delta^*}{3}, \quad \sum_{s'_i\in(0,\delta_1)} \frac{\log(s'_i+\varepsilon)}{\log\varepsilon} \geq -\frac{\delta^*}{3}.$$

Therefore, $\mathcal{F}_\varepsilon(s') \leq \|s'\|_0 + \delta^*/3 + \delta^*/3 = \|s'\|_0 + 2\delta^*/3$. This, combined with (13), yields

$$\mathcal{F}_\varepsilon(s) - \mathcal{F}_\varepsilon(s') \geq (\|s\|_0 - \delta^*/3) - (\|s'\|_0 + 2\delta^*/3) = (\|s\|_0 - \|s'\|_0) - \delta^*,$$

which implies (9) as long as $\|s\|_0 > \|s'\|_0$ (i.e., $\|s\|_0 - \|s'\|_0 \geq 1$).

(ii) Consider the function (7). Clearly, there exists $\varepsilon^* \in (0,1)$ such that for any $\varepsilon \in (0, \varepsilon^*]$ and for any $s_i \in [\delta_1, \delta_2]$, we have $1 - \frac{\delta^*}{n} \leq \frac{\delta_1}{\delta_1+\varepsilon} \leq \frac{s_i}{s_i+\varepsilon} \leq \frac{\delta_2}{\delta_2+\varepsilon} \leq 1$. Thus for any $\varepsilon \in (0, \varepsilon^*]$, $s \in \mathfrak{S}(\delta_1, \delta_2)$ and $0 \leq s' \leq \delta_2 e$, we have

$$\mathcal{F}_\varepsilon(s) = \sum_{i=1}^n \frac{s_i}{s_i+\varepsilon} = \sum_{s_i\in[\delta_1,\delta_2]} \frac{s_i}{s_i+\varepsilon} \geq \|s\|_0 \left(1 - \frac{\delta^*}{n}\right), \ \mathcal{F}_\varepsilon(s') = \sum_{s'_i>0} \frac{s'_i}{s'_i+\varepsilon} \leq \|s'\|_0.$$

Merging these inequalities yields

$$\mathcal{F}_\varepsilon(s) - \mathcal{F}_\varepsilon(s') \geq (1 - \frac{\delta^*}{n})\|s\|_0 - \|s'\|_0 = \|s\|_0 - \|s'\|_0 - \frac{\delta^*}{n}\|s\|_0 \geq (\|s\|_0 - \|s'\|_0) - \delta^*,$$

which implies (9) provided that $\|s\|_0 > \|s'\|_0$ (i.e., $\|s\|_0 - \|s'\|_0 \geq 1$).

(iii) Finally, we consider (8). Note that $(\delta_1 + \varepsilon^{1/\varepsilon})^\varepsilon \to 1$ and $(\delta_2 + \varepsilon^{1/\varepsilon})^\varepsilon \to 1$ as $\varepsilon \to 0$. Thus there exists a $\varepsilon^* \in (0,1)$ satisfying $\varepsilon^* < \delta^*/(2n+1)$ such that for any $\varepsilon \in (0, \varepsilon^*]$ and $s_i \in [\delta_1, \delta_2]$, the following inequalities hold:

$$1 - \frac{\delta^*}{2n+1} \leq (\delta_1 + \varepsilon^{1/\varepsilon})^\varepsilon \leq (s_i + \varepsilon^{1/\varepsilon})^\varepsilon \leq (\delta_2 + \varepsilon^{1/\varepsilon})^\varepsilon \leq 1 + \frac{\delta^*}{2n+1}. \tag{14}$$

Hence, for any $s \in \mathfrak{S}(\delta_1, \delta_2)$, we have

$$\mathcal{F}_\varepsilon(s) = \sum_{s_i > 0} (s_i + \varepsilon^{1/\varepsilon})^\varepsilon + \sum_{s_i = 0} (s_i + \varepsilon^{1/\varepsilon})^\varepsilon \geq (1 - \frac{\delta^*}{2n+1})\|s\|_0 + (n - \|s\|_0)\varepsilon.$$

For $0 \leq s' \leq \delta_2 e$, it follows from (14) that

$$\mathcal{F}_\varepsilon(s') = \sum_{s'_i > 0} (s'_i + \varepsilon^{1/\varepsilon})^\varepsilon + \sum_{s'_i = 0} (s'_i + \varepsilon^{1/\varepsilon})^\varepsilon \leq \left(1 + \frac{\delta^*}{2n+1}\right) \|s'\|_0 + (n - \|s'\|_0)\varepsilon.$$

Note that $\varepsilon^* \leq \frac{\delta^*}{2n+1} < \delta^*$ and $\|s\|_0 + \|s'\|_0 \leq 2n$. For any $\varepsilon \in (0, \varepsilon^*]$ and s, s' satisfying $s \in \mathfrak{S}(\delta_1, \delta_2)$, $0 \leq s' \leq \delta_2 e$ and $\|s'\|_0 < \|s\|_0$, we have

$$\begin{aligned} \mathcal{F}_\varepsilon(s) - \mathcal{F}_\varepsilon(s') &\geq \left(1 - \frac{\delta^*}{2n+1}\right) \|s\|_0 + (n - \|s\|_0)\varepsilon - \left[\left(1 + \frac{\delta^*}{2n+1}\right) \|s'\|_0 + (n - \|s'\|_0)\varepsilon\right] \\ &= (\|s\|_0 - \|s'\|_0)(1 - \varepsilon) - \frac{\delta^*}{2n+1}(\|s\|_0 + \|s'\|_0) \\ &\geq \left(1 - \frac{\delta^*}{2n+1}\right) - \frac{\delta^*}{2n+1}(2n) = 1 - \delta^* > 0, \end{aligned}$$

as desired. $\square$

3 New construction of weighted ℓ_1-algorithms

A central (and open) question concerning weighted ℓ_1-algorithms is how the weight should be chosen so that the algorithm can locate a sparsest point of P. In existing reweighted ℓ_1-methods (e.g., [11, 22, 32, 13, 15, 62]), the weight w^k is chosen according to the current iterate x^k. Consider the following example in [11]:

$$w_i^k = \frac{1}{|x_i^k| + \varepsilon} \text{ for } i = 1, \ldots, n,$$

where ε is a small positive parameter. We see that w_i^k is very large when $|x_i^k|$ and ε are very small. In this case, weighted ℓ_1-minimization penalizes the ith component of x, yielding the next iterate x^{k+1} with $x_i^{k+1} \approx 0$. Thus x^{k+1} might have a sparsity pattern same as or very similar to that of x^k. This might cause the algorithm to fail to locate a sparsest point if the sparsity pattern of the initial iterate is far from being the sparsest. The need to investigate other approaches for the choice of weight is apparent.

In this section, we will develop a new construction of the weighted ℓ_1-algorithm. Recall that the dual problem of (3) is given by

$$\max_{(y,s)} \{b^T y : \ A^T y + s = w, \ s \geq 0\} \tag{15}$$

where s is called the dual slack variable. It is well known that any linear programming problem with an optimal solution must have a strictly complementary optimal solution (see Goldman and Tucker [25]) in the sense that there exists a pair $(x,(y,s))$, where x is an optimal solution to (3) and (y,s) is an optimal solution to (15), satisfying that $(x,s)\geq 0$, $x^Ts=0$ and $x+s>0$. Using this property, we first prove that solving an ℓ_0-minimization problem is equivalent to finding a weight $w\in\mathcal{W}$ such that the dual problem (15) has the densest nonnegative vector $s=w-A^Ty$ among all possible choices of $w\in\mathcal{W}$.

Theorem 3.1. *Let (w^*,y^*,s^*) be an optimal solution to the bilevel programming problem*

$$\begin{array}{ll}\max_{(w,y,s)} & \|s\|_0 \\ s.t. & b^Ty=1,\ s=w-A^Ty\geq 0,\ w\geq 0,\ 1=\min_x\{\|Wx\|_1:\ x\in P\},\end{array}\tag{16}$$

where $W=diag(w)$. Then any optimal solution x^* to the weighted ℓ_1-minimization

$$\min_x\{\|W^*x\|_1:\ x\in P\},\tag{17}$$

where $W^*=diag(w^*)$, is a sparsest point of P. In addition, x^* and s^* satisfy that $\|x^*\|_0+\|s^*\|_0=n$. Conversely, let $w^*\in\mathcal{W}$ given by (4) satisfy that any optimal solution to (17) is a sparsest point of P. Then there is a vector (y^*,s^*) such that (w^*,y^*,s^*) is an optimal solution to (16).

Proof. Let (w^*,y^*,s^*) be an optimal solution to (16). We now prove that any optimal solution to (17) is a sparsest point of P. Let $\widetilde{x}$ be a sparsest point of P. By Corollary 2.4, there exists a weight $\widetilde{w}\in\mathcal{W}$ such that $\widetilde{x}$ is the unique optimal solution to the problem $1=\min_x\{\|\widetilde{W}x\|_1:\ x\in P\}$, where $\widetilde{W}=\mathrm{diag}(\widetilde{w})$, to which the dual problem is given by

$$\max_{(y,s)}\{b^Ty:\quad s=\widetilde{w}-A^Ty,\ s\geq 0\}.$$

There is a solution $(\widetilde{y},\widetilde{s})$ to this problem so that $\widetilde{x}$ and $\widetilde{s}$ are strictly complementary, and hence

$$\|\widetilde{x}\|_0+\|\widetilde{s}\|_0=n.\tag{18}$$

By the strong duality, we have $b^T\widetilde{y}=\|\widetilde{W}\widetilde{x}\|_1=\min\{\|\widetilde{W}x\|_1:\ x\in P\}=1$, which implies that $(\widetilde{w},\widetilde{y},\widetilde{s})$ is a feasible solution to (16). Since (w^*,y^*,s^*) is an optimal solution to (16), we have

$$\|\widetilde{s}\|_0\leq\|s^*\|_0.\tag{19}$$

Let x^* be an arbitrary optimal solution of (17), to which the dual problem is given by

$$\max_{(y,s)}\{b^Ty:\quad s=w^*-A^Ty,\ s\geq 0\}.\tag{20}$$

By the strong duality again, the constraints of (16) imply that (y^*,s^*) is an optimal solution to (20). Thus x^* and s^* are complementary, and hence

$$\|x^*\|_0+\|s^*\|_0\leq n.\tag{21}$$

Merging (18), (19) and (21) yields

$$\|x^*\|_0\leq n-\|s^*\|_0\leq n-\|\widetilde{s}\|_0=\|\widetilde{x}\|_0$$

which implies that x^* is a sparsest point of P. Therefore, $\|x^*\|_0 = \|\widetilde{x}\|_0$. This in turn implies from the above inequalities that $\|x^*\|_0 = n - \|s^*\|_0$ (and hence x^* and s^* are strictly complementary).

Conversely, for a given weight $w^* \in \mathcal{W}$, we assume that any optimal solution of (17) is a sparsest point of P. We prove that there exists a vector (y^*, s^*) such that (w^*, y^*, s^*) is an optimal solution to (16). Let (w, y, s) be an arbitrary feasible point of (16). For this w, let x be an optimal solution to the weighted ℓ_1-problem $\min\{\|Wx\|_1 : x \in P\}$ where $W = \mathrm{diag}(w)$. By linear programming duality theory, the constraints of (16) imply that (y, s) is an optimal solution to the dual problem of this weighted ℓ_1-problem, and hence x and s are complementary. Thus the following holds:

$$\|s\|_0 \le n - \|x\|_0 \le n - z^*, \tag{22}$$

where z^* is the optimal value of (2), i.e., $z^* = \min\{\|x\|_0 : x \in P\}$. So $n - z^*$ is the upper bound for the optimal value of (16). We now consider the problem (17) with weight w^* and its dual problem. Let $(x^*, (y^*, s^*))$ be a pair of strictly complementary solutions to (17) and its dual problem (such a pair always exists by linear programming theory). Thus

$$\|s^*\|_0 = n - \|x^*\|_0 = n - z^*, \tag{23}$$

where the second equality follows from the assumption that any optimal solution to (17) is a sparsest point of P. Since (w^*, y^*, s^*) is a feasible solution to (16), it follows from (22) and (23) that $\|s^*\|_0$ is the optimal value of (16). Thus (w^*, y^*, s^*) is an optimal solution to (16). $\square$

Remark 3.2. For any given $w \in \mathcal{W}$, let $x(w)$ denote an optimal solution to the problem (3), and let $(y(w), s(w))$ denote an optimal solution to its dual problem (15). By linear programming theory, $x(w)$ and $s(w)$ are complementary. This implies that

$$\|x(w)\|_0 + \|s(w)\|_0 \le n \ \text{ for any } w \in \mathcal{W}, \tag{24}$$

where the equality can be achieved when $x(w)$ and $s(w)$ are strictly complementary. From (24), we see that an increase in $\|s(w)\|_0$ leads to a decrease of $\|x(w)\|_0$. Thus the densest possible vector in $\{s(w) : w \in \mathcal{W}\}$, denoted by $s(w^*)$, yields the sparsest vector $x(w^*)$ which must be a sparsest point of P. Clearly, the vector $(w^*, y(w^*), s(w^*))$ is an optimal solution to the bilevel programming problem (16). By Definition 2.3, we see that the weight

$$w^* \in \arg\max\{\|s(w)\|_0 : \ w \in \mathcal{W}\}$$

is an optimal weight by which the sparsest point of P can be immediately obtained via weighted ℓ_1-minimization. Theorem 3.1 shows that such an optimal weight can be obtained by solving the bilevel program (16).

Although it is generally difficult to solve a bilevel program precisely (see, e.g., [33]), the special structure of (16) makes it possible to find an approximate optimal weight via a certain approximation or relaxation of (16). Let $\mathcal{F}_\varepsilon(s)$ be a merit function for sparsity defined in Section 2.3. Replacing the objective of (16) by $\mathcal{F}_\varepsilon(s)$ leads to the following continuous approximation of (16):

$$\begin{array}{ll} \max_{(w,y,s)} & \mathcal{F}_\varepsilon(s) \\ \text{s.t.} & b^T y = 1, \ s = w - A^T y \ge 0, \ 1 = \min_x\{\|Wx\|_1 : x \in P\}, \ w \ge 0. \end{array} \tag{25}$$

By duality theory, the constraints of (25) imply that for any given $w \in \mathcal{W}$, the vector (y, s) satisfying the constraints is an optimal solution to the problem (15) which maximizes $b^T y$ subject to the constraint $s = w - A^T y \geq 0$. Namely, (25) can be written as

$$\begin{array}{ll} \max_{(w,s)} & \mathcal{F}_\varepsilon(s) \\ \text{s.t.} & s = w - A^T y, \ w \in \mathcal{W}, \text{ where } y \text{ is an optimal solution to} \\ & \max_y \{b^T y : \ w - A^T y \geq 0\}. \end{array} \tag{26}$$

In this bilevel program, both $b^T y$ and $\mathcal{F}_\varepsilon(s)$ are required to be maximized, subject to the constraint $s = w - A^T y \geq 0$ with all possible choices of $w \in \mathcal{W}$. Thus we consider the following model in order to possibly maximize both objectives:

$$\max_{(w,y,s)} \left\{ b^T y + \alpha \mathcal{F}_\varepsilon(s) : \ s = w - A^T y \geq 0, \ w \in \mathcal{W} \right\}, \tag{27}$$

where $\alpha > 0$ is a given parameter. From (26), we see that $\mathcal{F}_\varepsilon(s)$ should be maximized based on the condition that $b^T y$ is maximized over the feasible set of (27). This indicates that a small parameter α in (27) should be chosen in order to meet this condition.

Note that (27) might not have a finite optimal value since $\mathcal{W}$ is not necessarily bounded. Another difficulty of (27) is that the constraint $w \in \mathcal{W}$ is not given explicitly. Since scaling of weight does not affect the solution of (3), we may confine w to a bounded convex set $\mathcal{B}$ in $\mathbb{R}^n_+$ so that the problem (27) has a finite optimal value. The constraint $w \in \mathcal{W}$ guarantees $1 = \min_x\{\|Wx\|_1 : x \in P\}$, which by strong duality is equivalent to $1 = \max\{b^T y : s = w - A^T y \geq 0\}$. Thus $b^T y \leq 1$ is a certain relaxation of the constraint $w \in \mathcal{W}$. Based on these observations, we propose the following convex relaxation model of (27):

$$\max_{(w,y,s)} \{b^T y + \alpha \mathcal{F}_\varepsilon(s) : \ s = w - A^T y \geq 0, \ b^T y \leq 1, \ w \in \mathcal{B}\}, \tag{28}$$

where $\mathcal{B} \subseteq \mathbb{R}^n_+$ is a bounded convex set. From the above discussion, we see that ε is a parameter controlling the proximity of the ℓ_0-norm and merit functions, and α is a parameter controlling the proximity of an optimization problem and its perturbation. Thus small values should be chosen for (α, ε) in order to ensure that (28) is a certain approximation of (16), and thus the vector w resulting from (28) can be seen as an approximate optimal weight.

Clearly, the solution of (28) relies on the choice of $\mathcal{B}$. For the convenience of convergence analysis, we choose $\mathcal{B}$ as the following bounded polytope:

$$\mathcal{B} = \{w \in \mathbb{R}^n_+ : (x^0)^T w \leq M, \ w \leq M^* e\}, \tag{29}$$

where M and M^* are two given positive numbers, and x^0 is the solution to an initial weighted ℓ_1-minimization problem (e.g. the normal ℓ_1-minimization with weight e). Except for convergence analysis, there is another reason to include the inequality $(x^0)^T w \leq M$ in (29). Note that x^0 (as a solution of initial weighted ℓ_1-minimization) admits a certain level of sparsity and its sparsity pattern might carry some useful information for the selection of weights. In existing framework of reweighted ℓ_1-algorithms, the selection of weights often relies on the magnitude of components of the current iterate, say x^0. If x^0 is not the sparsest one, the existing idea is to generate another iterate that might be sparser than x^0 by solving a weighted ℓ_1-minimization problem with a weight selected according to the following strategy: Assign large weights corresponding to small components of x^0, typically, $w_i = \frac{M}{(x^0)_i}$ for $i = 1, \dots, n$, where M is a constant. This weight-selection idea in existing reweighted ℓ_1-algorithms is reflected in the inequality $(x^0)^T w \leq M$ in (29).

We are now in a position to describe the algorithm.

Algorithm 3.3. Let α, $\varepsilon \in (0, 1)$ be given parameters and let $w^0 \in \mathbb{R}^n_{++}$ be a given vector.

- Initialization: Solve $\min\{\|W^0 x\|_1 : x \in P\}$, where $W^0 = \mathrm{diag}(w^0)$, to obtain a minimizer x^0. Set $\gamma^0 = \|W^0 x^0\|_1$ and choose two constants M and M^* such that $1 \leq M \leq M^*$ and $M\|w^0\|_\infty/\gamma^0 \leq M^*$.

- Step 1. Solve the convex optimization problem

$$\max_{(w,y,s)} \left\{ b^T y + \alpha \mathcal{F}_\varepsilon(s) : \ s = w - A^T y \geq 0, \ b^T y \leq 1, \ (x^0)^T w \leq M, \ 0 \leq w \leq M^* e \right\}. \quad (30)$$

  Let $(\widetilde{w}, \widetilde{y}, \widetilde{s})$ be an optimal solution to this problem.

- Step 2. Let $\widetilde{W} = \mathrm{diag}(\widetilde{w})$ and solve

$$\min\{\|\widetilde{W} x\|_1 : x \in P\} \quad (31)$$

  to obtain a point $\widetilde{x}$.

Theorem 3.1 promotes the following idea: *Seek an (approximate) optimal weight by maximizing the cardinality of the support of dual slack variables with all possible choices of $w \in \mathbb{R}^n_+$.* Algorithm 3.3 can be seen as an initial development in this direction. Clearly, some modification or improvement can be made to Algorithm 3.3, depending on the choice of $\mathcal{B}$ and the method of relaxation of (16). Also, an iterative version of the algorithm can be presented when $\mathcal{B}$, α and ε are iteratively updated in the course of the algorithm according to certain updating schemes. Clearly, a key difference between Algorithm 3.3 and existing weighted ℓ_1-methods lies in the principle for tackling ℓ_0-minimization. Existing weighted ℓ_1-algorithms working in primal space have been derived from minimizing nonlinear merit functions for sparsity via linearization (i.e., first-order approximation) which results in the weight given by the gradient of merit functions at the current iterate (see [22, 57, 62] for details). The theoretical efficiency of these linearization-based algorithms has still not been properly addressed or has been addressed under some restrictive assumptions. The benefit for ℓ_0-norm maximization is that it can be achieved by maximizing a concave merit function for sparsity (i.e., minimizing a convex function) without requiring any first-order estimation of the merit function. The weight $\widehat{w}$ in Algorithm 3.3 can be viewed as an approximate optimal weight generated by the convex optimization problem (30) in dual space.

4 Guaranteed performance

In this section, the theoretical performance of Algorithm 3.3 will be shown under certain assumptions. We first prove a generic guaranteed performance for Algorithm 3.3 under some conditions. The guaranteed performance under a stronger condition will be discussed in Section 4.2.

4.1 Generic performance guarantee

Note that $\mathcal{F}_\varepsilon(s)$ is independent of the polyhedral set P, so Assumption 2.5 is not a condition imposed on P. We make the following assumption on P.

Assumption 4.1. Let $A \in \mathbb{R}^{m\times n}$ $(m < n)$ be a matrix satisfying that $\{y \in \mathbb{R}^m : A^T y \le w\}$ is bounded for some $w \in \mathbb{R}^n_+$.

This assumption is equivalent to saying that $\{d \in \mathbb{R}^m : A^T d \le 0\} = \{0\}$ (see, e.g., Theorem 8.4 in [46]). This is a mild assumption on P. To see this, for a vector $x \in \mathbb{R}^n_+$, let us define $x_+, x_- \in \mathbb{R}^n_+$ as $(x_+)_i = \max\{x_i, 0\}$ and $(x_-)_i = |\min\{x_i, 0\}|$ for $i = 1, \ldots, n$. Then $x = x_+ - x_-$ and $\|x\|_0 = \left\| \begin{bmatrix} x_+ \\ x_- \end{bmatrix} \right\|_0$. Thus for any given full-row-rank matrix $H \in \mathbb{R}^{m\times n}$ $(m < n)$, the ℓ_0-problem $\min_x\{\|x\|_0 : Hx = b\}$ can be reformulated as the following problem seeking the sparsest point of a polyhedral set:

$$\min\left\{ \left\| \begin{bmatrix} x_+ \\ x_- \end{bmatrix} \right\|_0 : \ Hx_+ - Hx_- = b, \ x_+ \ge 0, \ x_- \ge 0 \right\},$$

which clearly satisfies Assumption 4.1 since $\begin{bmatrix} H^T \\ -H^T \end{bmatrix} d \le 0$ implies that $d = 0$ and hence the set $\left\{ y \in \mathbb{R}^m : \begin{bmatrix} H^T \\ -H^T \end{bmatrix} y \le w \right\}$ is bounded for any given $w \in \mathbb{R}^n_+$. We also note that $A^T d \le 0$ means the angles between d and every column of A being greater than or equal to $\pi/2$. Clearly, when the columns of $A \in \mathbb{R}^{m\times n}$, where $m \ll n$, are uniformly distributed on the surface of a m-dimensional sphere, the vector d satisfying $A^T d \le 0$ would be equal to zero almost surely. Thus the polyhedral sets with such matrices satisfy Assumption 4.1 almost surely.

A generic guaranteed performance of Algorithm 3.3 is given as follows.

Theorem 4.2. *Let $\mathcal{F}_\varepsilon(s)$ be a concave and continuously differentiable merit function for sparsity with respect to s over an open set containing $\mathbb{R}^n_+$, and let $\mathcal{F}_\varepsilon(s)$ satisfy Assumption 2.5. Suppose that Assumption 4.1 is satisfied. Let $w^0 \in \mathbb{R}^n_{++}$ be arbitrarily given and $\gamma^0 = \min\{\|W^0 x\|_1 : x \in P\}$. Then there exists a number $\widehat{M} \ge 1$ satisfying the following property: For any $M \ge \widehat{M}$ and $M^* \ge M \max\{1, \|w^0\|_\infty/\gamma^0\}$, there is a number $\varepsilon^* \in (0,1)$ such that for any $\varepsilon \in (0, \varepsilon^*]$, there is $\alpha^* \in (0,1)$ such that for any $\alpha \in (0, \alpha^*]$, the point $\widetilde{x}$ generated by Algorithm 3.3 with $(M, M^*, \alpha, \varepsilon)$ is a sparsest point of P provided that $(\widetilde{w})^T \widetilde{x} = 1$, where $\widetilde{w}$ is the weight generated at Step 1 of Algorithm 3.3.*

Roughly speaking, under Assumptions 2.5 and 4.1, Algorithm 3.3 can find a sparsest point of P if α and ε are chosen to be small enough, M and M^* are chosen to be large enough, and the optimal solution to (31) satisfies $\widetilde{w}^T \widetilde{x} = 1$ (which can be guaranteed under another assumption on P, see Section 4.2 for details). Before proving Theorem 4.2, let us first introduce the following constant that will be used in its proof:

$$\sigma^* = \min_{x \in Z^*} \left\| A^T_{\overline{J_+(x)}} A_{J_+(x)} (A^T_{J_+(x)} A_{J_+(x)})^{-1} \right\|_\infty, \tag{32}$$

where Z^* denotes the set of the sparsest points of P. By Lemma 2.1, for every $x \in Z^*$, the matrix $A_{J_+(x)}$ has full column rank. This implies that no two distinct sparsest points share the same support. Thus Z^* is a finite set (containing only a finite number of elements), and thus σ^* is a finite nonnegative constant (determined by the problem data (A, b)). The following lemma established by Mangasarian and Meyer [37] (see Friedlander and Tseng [23] for a more general setting) will be used in the proof of Theorem 4.2 as well.

Lemma 4.3 [37]. *Let ϕ be a continuously differentiable convex function on some open set containing the feasible set Q of the linear program* $\min\{c^T x : \ x \in Q\}$. *If the solution set $\widehat{S}$ of the linear program is nonempty and bounded, and $c^T x + \widetilde{\alpha}\phi(x)$ is bounded from below on the set Q for some $\widetilde{\alpha} > 0$, then the solution set of the problem* $\min\{c^T x + \alpha\phi(x) : \ x \in Q\}$ *is contained in $\widehat{S}$ for sufficiently small $\alpha > 0$.*

We also need to establish two technical results. The first result below claims that by a suitable choice of (M, M^*), the set $\mathcal{B}$ defined as (29) contains an optimal weight for a sparsest point of P.

Lemma 4.4. *Let x^* be the sparsest point satisfying*

$$\left\| A^T_{\overline{J_+(x^*)}} A_{J_+(x^*)} (A^T_{J_+(x^*)} A_{J_+(x^*)})^{-1} \right\|_\infty = \sigma^*. \tag{33}$$

where σ^* is given by (32). Let $w^0 \in \mathbb{R}^n_{++}$ be any given vector and (γ^0, x^0) be generated at the initial step of Algorithm 3.3. Then there exist a number $\widehat{M} \geq 1$ and an optimal weight $w^* \in \mathcal{W}$ corresponding to x^* such that the following hold:

(i)

$$w^* \in \mathcal{B} = \{w \in \mathbb{R}^n_+ : \ (x^0)^T w \leq M, \ w \leq M^* e\} \tag{34}$$

for any $M \geq \widehat{M}$ and any $M^* \geq M \max\,\{1, \|w^0\|_\infty/\gamma^0\}$.

(ii) *x^* is the unique optimal solution to the problem*

$$\gamma^* = \min_x\{\|W^* x\|_1 : \ x \in P\} \tag{35}$$

where $\gamma^* = 1$.

Proof. Note that γ^0 is the optimal value and x^0 is an optimal solution to the initial problem $\min\{\|W^0 x\|_1 : x \in P\}$. Since $b \neq 0$ and $w^0 \in \mathbb{R}^n_{++}$, we must have that $x^0 \neq 0$ and $\gamma^0 > 0$. Setting $\widehat{w} = w^0/\gamma^0$ and $\widehat{W} = \mathrm{diag}(\widehat{w})$, we see that x^0 is still an optimal solution to the problem

$$\widehat{\gamma} = \min\{\|\widehat{W} x\|_1 : \ x \in P\} \tag{36}$$

where $\widehat{\gamma} = 1$. We now show that there exists a number $\widehat{M} \geq 1$ and an optimal weight $w^* \in \mathcal{W}$ satisfying the desired properties. First, we note that there is a sparsest point $x^* \in P$ satisfying (33). For simplicity, we denote the support of x^* by J (i.e., $J = J_+(x^*)$) and we define $\rho = \|\widehat{W} x^*\|_1 = (\widehat{w}_J)^T x^*_J$. From (36), we see that $\rho \geq \widehat{\gamma} = 1$. Note that σ^* is a finite constant. Let σ be a number satisfying $\sigma > \sigma^*$. By using $(\widehat{w}, \rho, \sigma)$, we define the vector w^* as

$$w^*_J = \widehat{w}_J/\rho, \ w^*_{\overline{J}} = \sigma(\|\widehat{w}_J\|_\infty) e_{\overline{J}} \tag{37}$$

and we define the constant $\widehat{M} := \max\{1, \sigma\beta\}$, where

$$\beta = \frac{\|w^0\|_\infty}{\min_{1 \leq i \leq n} w^0_i} \geq 1$$

is a number determined by w^0. We now show that (34) holds for any $M \geq \widehat{M}$ and $M^* \geq M \max\{1, \|w^0\|_\infty/\gamma^0\}$. Since $\|\widehat{w}_J\|_\infty = \|w^0_J\|_\infty/\gamma^0 \leq \|w^0\|_\infty/\gamma^0 = \beta \left(\min_{1 \leq i \leq n} w^0_i\right)/\gamma^0$, one has

$$(x^0_{\overline{J}})^T (\|\widehat{w}_J\|_\infty e_{\overline{J}}) \leq \beta (x^0_{\overline{J}})^T \left[\left(\min_{1 \leq i \leq n} w^0_i\right) e_{\overline{J}}\right] /\gamma^0 \leq \beta (x^0_{\overline{J}})^T w^0_{\overline{J}}/\gamma^0 = \beta (x^0_{\overline{J}})^T \widehat{w}_{\overline{J}}.$$

Therefore, by (37) and noting that $(x^0)^T\widehat{w} = \widehat{\gamma} = 1$ and $\rho \ge 1$, we have

$$\begin{aligned}(x^0)^T w^* &= (x^0_J)^T\widehat{w}_J/\rho + (x^0_{\overline{J}})^T(\sigma\|\widehat{w}_J\|_\infty e_{\overline{J}}) \le (x^0_J)^T\widehat{w}_J/\rho + (\sigma\beta)(x^0_{\overline{J}})^T\widehat{w}_{\overline{J}} \\ &\le \max\{1/\rho, \sigma\beta\}[(x^0_J)^T\widehat{w}_J + (x^0_{\overline{J}})^T\widehat{w}_{\overline{J}}] \le \widehat{M} \le M\end{aligned}$$

and

$$\begin{aligned}\|w^*\|_\infty &= \max\{\|w^*_J\|_\infty, \|w^*_{\overline{J}}\|_\infty\} \le \max\{1/\rho, \sigma\}\|\widehat{w}_J\|_\infty \le \max\{1, \sigma\beta\}\|\widehat{w}\|_\infty \\ &\le M\|w^0\|_\infty/\gamma^0 \le M^*.\end{aligned}$$

Therefore, (34) holds. We now show that x^* is the unique solution to (35) with w^* given by (37) and the optimal value of (35) is equal to 1. Indeed, by (33) and (37) and the fact $\rho \ge 1$, we have

$$\left\|A^T_{\overline{J}}A_J(A^T_JA_J)^{-1}w^*_J\right\|_\infty \le \left\|A^T_{\overline{J}}A_J(A^T_JA_J)^{-1}\right\|_\infty \cdot \|w^*_J\|_\infty < \sigma\|\widehat{w}_J\|_\infty/\rho \le \sigma\|\widehat{w}_J\|_\infty,$$

which, together with (37), implies that

$$A^T_{\overline{J}}A_J(A^T_JA_J)^{-1}w^*_J < \sigma(\|\widehat{w}_J\|_\infty)e_{\overline{J}} = w^*_{\overline{J}}.$$

Therefore, by Theorem 2.2, x^* is the unique optimal solution to (35). Moreover, the optimal value of (35) is given by $\gamma^* = \|W^*x^*\|_1 = (w^*_J)^Tx^*_J = (\widehat{w}_J)^Tx^*_J/\rho = 1$. Therefore, $w^* \in \mathcal{W}$ is an optimal weight. $\square$

We now prove the next technical result.

Lemma 4.5. *Let $\mathcal{F}_\varepsilon$ satisfy Assumptions 2.5 and matrix A satisfy Assumption 4.1. Let x^* be the sparsest point satisfying (33), w^* be the optimal weight (corresponding to x^*) satisfying (34) and M^* be the constant given in Lemma 4.4. Let (y^*, s^*) be the optimal solution of the dual problem of (35), i.e., $(y^*, s^*) \in \arg\max\{b^Ty : \ s = w^* - A^Ty, \ s \ge 0\}$, such that s^* and x^* are strictly complementary. Then there exists $\varepsilon^* \in (0, 1)$ such that*

$$\mathcal{F}_\varepsilon(s^*) - \mathcal{F}_\varepsilon(s') \ge 3/4 \tag{38}$$

holds for any $\varepsilon \in (0, \varepsilon^*]$ and any $s' \in S(M^*)$ with $\|s'\|_0 < \|s^*\|_0$, where

$$S(M^*) = \{s \in \mathbb{R}^n_+ : s = w - A^Ty \ge 0, \ y \in \mathbb{R}^m, \ 0 \le w \le M^*e\}. \tag{39}$$

Proof. First, we consider the dual problem of (35)

$$\max_{(y,s)}\{b^Ty : \ s = w^* - A^Ty, \ s \ge 0\}. \tag{40}$$

By the choice of w^*, x^* is the unique optimal solution to (35). Thus by linear programming theory, there is an optimal solution (y^*, s^*) to (40), where $s^* = w^* - A^Ty^* \ge 0$, such that $x^* \in \mathbb{R}^n_+$ and $s^* \in \mathbb{R}^n_+$ are strictly complementary, i.e., $(x^*)^Ts^* = 0$ and $x^* + s^* > 0$. We now further show that $J_+(s^*) \ne \emptyset$. Since x^* is the unique optimal solution to (35), $A_{J_+(x^*)} \in \mathbb{R}^{m\times|J_+(x^*)|}$ has full column rank (see Remark 3.7 in [60]). Thus

$$\|x^*\|_0 = |J_+(x^*)| = \mathrm{rank}(A_{J_+(x^*)}) \le m.$$

Since $m < n$, the strict complementarity of s^* and x^* implies that $|J_+(s^*)| = n - |J_+(x^*)| \geq n - m > 0$. Thus we can define the positive number

$$\delta_1 = \min_{i \in J_+(s^*)} s_i^*. \tag{41}$$

Second, we show that the set (39) is bounded. Note that $\{d \in \mathbb{R}^m : A^T d \leq 0\}$ is the set of directions (i.e., the recession cone) of $\{y \in \mathbb{R}^m : A^T y \leq M^* e\}$. By Theorem 8.4 in [46], a convex set is bounded if and only if its recession cone consists of the zero vector alone. Thus Assumption 4.1 ensures that the recession cone $\{d \in \mathbb{R}^m : A^T d \leq 0\} = \{0\}$, which implies that the set $\{y \in \mathbb{R}^m : A^T y \leq M^* e\}$ is bounded. Thus, it follows from the fact

$$\{y \in \mathbb{R}^m : A^T y \leq w, \ 0 \leq w \leq M^* e\} \subseteq \{y \in \mathbb{R}^m : A^T y \leq M^* e\}$$

that the set $S(M^*)$ given by (39) is bounded. As a result, there exists a number $\delta_2 > 0$ such that $0 \leq s \leq \delta_2 e$ for all $s \in S(M^*)$. In particular, by the choice of w^* which satisfies (34) and by the definition of (y^*, s^*), we see that $s^* = w^* - A^T y^* \geq 0$ and $w^* \leq M^* e$, so $s^* \in S(M^*)$. This, together with (41), implies that $0 < \delta_1 \leq s_i^* \leq \delta_2$ for all $i \in J_+(s^*)$, i.e., $s^* \in \mathfrak{S}(\delta_1, \delta_2)$. Let $\delta^* = 1/4$ be a fixed constant. Since $\mathcal{F}_\varepsilon(s)$ satisfies Assumption 2.5, there exists a number $\varepsilon^* \in (0, 1)$ such that

$$\mathcal{F}_\varepsilon(s^*) - \mathcal{F}_\varepsilon(s') \geq 1 - \delta^* = 3/4$$

holds for any $\varepsilon \in (0, \varepsilon^*]$ and any s' satisfying $\|s'\|_0 < \|s^*\|_0$ and $0 \leq s' \leq \delta_2 e$. In particular, it holds for any $s' \in S(M^*)$ with $\|s'\|_0 < \|s^*\|_0$. $\square$

We are now in a position to prove Theorem 4.2. Note that (30) is solved for given parameters $\alpha, \varepsilon \in (0, 1)$. In the following proof, we write the vector $(\widetilde{w}, \widetilde{y}, \widetilde{s})$ generated at Step 1 of Algorithm 3.3 as $(\widetilde{w}(\alpha, \varepsilon), \widetilde{y}(\alpha, \varepsilon), \widetilde{s}(\alpha, \varepsilon))$, and the optimal solution $\widetilde{x}$ of (31) as $\widetilde{x}(\alpha, \varepsilon)$.

Proof of Theorem 4.2. Given the polyhedral set P, there is a sparsest point $x^* \in P$ satisfying (33). Let (w^0, γ^0, x^0) be given according to the initialization of Algorithm 3.3. By Lemma 4.4, there exist a number $\widehat{M} \geq 1$ and an optimal weight $w^* \in \mathcal{W}$ corresponding to x^* such that

$$w^* \in \mathcal{B} = \{w \in \mathbb{R}^n_+ : \ (x^0)^T w \leq M, \ w \leq M^* e\}$$

for any choice of (M, M^*) satisfying $M \geq \widehat{M}$ and $M^* \geq M \max\{1, \|w^0\|_\infty / \gamma^0\}$. Moreover, x^* is the unique solution to the problem (35). Under Assumptions 2.5 and 4.1, it follows from Lemma 4.5 that there exists a small number $\varepsilon^* \in (0, 1)$ such that (38) holds for any $\varepsilon \in (0, \varepsilon^*]$ and any $s' \in S(M^*)$ with $\|s'\|_0 < \|s^*\|_0$, where $S(M^*)$ is defined by (39).

Let ε be any fixed number in $(0, \varepsilon^*]$. We now show that there exists a positive number $\alpha^* \in (0, 1)$ satisfying the desired property stated in Theorem 4.2. To achieve this goal, we consider the problem (30), which is a perturbed version of the linear program

$$\begin{array}{ll} \max_{(w,y,s)} & b^T y \\ \text{s.t.} & s = w - A^T y \geq 0, \ b^T y \leq 1, \ (x^0)^T w \leq M, \ 0 \leq w \leq M^* e \end{array} \tag{42}$$

in the sense that (30) can be obtained by adding $\alpha \mathcal{F}_\varepsilon(s)$ to the objective of (42). Let $\mathcal{T}^*$ be the set of optimal solutions of (42). As shown in the proof of Lemma 4.5, under Assumption 4.1, the set $\{y \in \mathbb{R}^m : A^T y \leq M^* e\}$ is bounded. This implies that the feasible set of (42) is bounded, and so is $\mathcal{T}^*$. By Assumption 2.5, $\mathcal{F}_\varepsilon(s)$ is a continuously differentiable concave function over an open neighborhood of the set $\{s \in \mathbb{R}^n : s \geq 0\}$. This is equivalent to saying that the function $\widehat{\mathcal{F}}_\varepsilon(w, y, s) := \mathcal{F}_\varepsilon(s) + 0^T w + 0^T y = \mathcal{F}_\varepsilon(s)$ is a continuously differentiable concave function over an open neighborhood of the feasible set of (42). Since the feasible set of (42) is closed and bounded, for any given $\alpha > 0$, the concave function $b^T y + \alpha \mathcal{F}_\varepsilon(s) = b^T y + \alpha \widehat{\mathcal{F}}_\varepsilon(w, y, s)$ is bounded from above over the feasible set of (42). Therefore, by applying Lemma 4.3 to the maximization case, it implies that there exists $\alpha^* \in (0, 1)$ such that for any $\alpha \in (0, \alpha^*]$, the solution set of (30) is a subset of $\mathcal{T}^*$. Thus the optimal solution $(\widetilde{w}(\alpha,\varepsilon), \widetilde{y}(\alpha,\varepsilon), \widetilde{s}(\alpha,\varepsilon))$ of (30) is also an optimal solution to (42) for any $\alpha \in (0, \alpha^*]$.

Because of the constraint $b^T y \leq 1$, the optimal value of (42) is at most 1. We now further prove that the optimal value of (42) is equal to 1. Let $\widehat{w} = w^0/\gamma^0$. As shown at the beginning of the proof of Lemma 4.4, the optimal value $\widehat{\gamma}$ of (36) is equal to 1. By duality theory (or optimality conditions), there exists an optimal solution $(y^{(1)}, s^{(1)})$ to the dual problem

$$\max_{(y,s)} \{b^T y : s = \widehat{w} - A^T y \geq 0\}$$

such that $b^T y^{(1)} = \widehat{\gamma} = 1$. By the definition of $\widehat{w}$, it is easy to verify that $\widehat{w} \in \mathcal{B}$ where $\mathcal{B}$ is given by (34). In fact, since x^0 is an optimal solution to (36), we have that $(x^0)^T \widehat{w} = \widehat{\gamma} = 1 \leq \widehat{M} \leq M$ and

$$\|\widehat{w}\|_\infty = \|w^0\|_\infty/\gamma^0 \leq M\|w^0\|_\infty/\gamma^0 \leq M^*$$

which implies that $\widehat{w} \leq M^* e$. Thus $(\widehat{w}, y^{(1)}, s^{(1)})$ satisfies that

$$s^{(1)} = \widehat{w} - A^T y^{(1)},\ b^T y^{(1)} = 1,\ (x^0)^T \widehat{w} \leq M,\ 0 \leq \widehat{w} \leq M^* e$$

which implies that $(w, y, s) = (\widehat{w}, y^{(1)}, s^{(1)})$ is an optimal solution to (42) and the optimal value of (42) is equal to 1. As shown above, $(\widetilde{w}(\alpha,\varepsilon), \widetilde{y}(\alpha,\varepsilon), \widetilde{s}(\alpha,\varepsilon))$ is an optimal solution to (42) for any $\alpha \in (0, \alpha^*]$. Thus we conclude that

$$b^T \widetilde{y}(\alpha,\varepsilon) = 1 \text{ for any } \alpha \in (0, \alpha^*]. \tag{43}$$

Given the vector $\widetilde{w}(\alpha,\varepsilon)$, let us consider the linear program

$$\widetilde{\gamma}(\alpha,\varepsilon) = \max_{(y,s)} \{b^T y :\ s = \widetilde{w}(\alpha,\varepsilon) - A^T y \geq 0,\ b^T y \leq 1\}, \tag{44}$$

and its dual problem

$$\min_{(x,t)} \{\widetilde{w}(\alpha,\varepsilon)^T x + t :\ Ax + tb = b,\ (x,t) \geq 0\}. \tag{45}$$

Due to the constraint $b^T y \leq 1$, the optimal value of (44) is at most 1. Note that for any $\alpha \in (0, \alpha^*]$, $(\widetilde{y}(\alpha,\varepsilon), \widetilde{s}(\alpha,\varepsilon))$ is a feasible solution of (44) and $\widetilde{y}(\alpha,\varepsilon)$ satisfies (43). This implies that $(\widetilde{y}(\alpha,\varepsilon), \widetilde{s}(\alpha,\varepsilon))$ is an optimal solution to the problem (44) and the optimal value $\widetilde{\gamma}(\alpha,\varepsilon) = 1$. By duality theory, the optimal value of (45) is also equal to 1. Thus the solution set of (45) is given by

$$U^{(\alpha,\varepsilon)} = \{(x,t) :\ \widetilde{w}(\alpha,\varepsilon)^T x + t = 1,\ Ax + bt = b,\ (x,t) \geq 0\}.$$

We now consider the case $\widetilde{w}(\alpha,\varepsilon)^T \widetilde{x}(\alpha,\varepsilon) = 1$. In this case, $(x,t) = (\widetilde{x}(\alpha,\varepsilon), 0) \in U^{(\alpha,\varepsilon)}$, and hence $(\widetilde{x}(\alpha,\varepsilon), 0)$ is an optimal solution of (45). At the optimal solution $(\widetilde{y}(\alpha,\varepsilon), \widetilde{s}(\alpha,\varepsilon))$ of (44), the values for the associated dual slack variables of (44) are $(\widetilde{s}(\alpha,\varepsilon),0)$. By the complementary slackness, $(\widetilde{x}(\alpha,\varepsilon),0)$ and $(\widetilde{s}(\alpha,\varepsilon),0)$ are complementary, i.e., $\widetilde{s}(\alpha,\varepsilon)^T\widetilde{x}(\alpha,\varepsilon)=0$. This, together with the nonnegativity of $\widetilde{x}(\alpha,\varepsilon)$ and $\widetilde{s}(\alpha,\varepsilon)$, implies that

$$\|\widetilde{x}(\alpha,\varepsilon)\|_0+\|\widetilde{s}(\alpha,\varepsilon)\|_0\le n. \tag{46}$$

We now prove that the point $\widetilde{x}(\alpha,\varepsilon)$ must be a sparsest point of P. We prove this by contradiction. Assume that $\widetilde{x}(\alpha,\varepsilon)$ is not a sparsest point of P. Then

$$\|x^*\|_0<\|\widetilde{x}(\alpha,\varepsilon)\|_0, \tag{47}$$

where x^* is the sparsest point of P satisfying (33). By linear programming theory, there exists a solution (y^*,s^*) to the dual problem of (35) such that x^* and s^* are strictly complementary. Thus we have $\|x^*\|_0+\|s^*\|_0=n$ which, together with (46) and (47), yields

$$\|s^*\|_0-\|\widetilde{s}(\alpha,\varepsilon)\|_0\ge(n-\|x^*\|_0)-(n-\|\widetilde{x}(\alpha,\varepsilon)\|_0)=\|\widetilde{x}(\alpha,\varepsilon)\|_0-\|x^*\|_0>0.$$

Thus $\|\widetilde{s}(\alpha,\varepsilon)\|_0<\|s^*\|_0$. Clearly, $\widetilde{s}(\alpha,\varepsilon)\in S(M^*)$, where $S(M^*)$ is defined by (39). It follows from Lemma 4.5 that (38) is satisfied, and hence

$$\mathcal{F}_\varepsilon(s^*)-\mathcal{F}_\varepsilon(\widetilde{s}(\alpha,\varepsilon))>3/4. \tag{48}$$

On the other hand, the vector (w^*,y^*,s^*) is a feasible point of (30). Since $(\widetilde{w}(\alpha,\varepsilon),\widetilde{y}(\alpha,\varepsilon),\widetilde{s}(\alpha,\varepsilon))$ is an optimal solution to (30), we must have

$$b^T\widetilde{y}(\alpha,\varepsilon)+\alpha\mathcal{F}_\varepsilon(\widetilde{s}(\alpha,\varepsilon))\ge b^Ty^*+\alpha\mathcal{F}_\varepsilon(s^*).$$

By strong duality, Lemma 4.4 (ii) implies that $b^Ty^*=1$. This, together with (43), implies that the above inequality is equivalent to $\mathcal{F}_\varepsilon(\widetilde{s}(\alpha,\varepsilon))\ge\mathcal{F}_\varepsilon(s^*)$, which contradicts (48). This contradiction shows that $\widetilde{x}=\widetilde{x}(\alpha,\varepsilon)$, generated at Step 2 of Algorithm 3.3, is a sparsest point of P. $\Box$

From the above proof, we see that under Assumptions 2.5 and 4.1, the point $(\widetilde{w},\widetilde{y},\widetilde{s})=(\widetilde{w}(\alpha,\varepsilon),\widetilde{y}(\alpha,\varepsilon),\widetilde{s}(\alpha,\varepsilon))$ generated by Algorithm 3.3 satisfies the property (43) for any given sufficiently small (α,ε). Since $(\widetilde{y}(\alpha,\varepsilon),\widetilde{s}(\alpha,\varepsilon))$ is a feasible solution to the linear program $\max_{(y,s)}\{b^Ty: s=\widetilde{w}(\alpha,\varepsilon)-A^Ty\ge 0\}$, which is the dual problem of (31). Thus by (43), the optimal value of this dual problem is at least 1. By duality theory, this in turn implies that the optimal value of (31) is also at least 1. Thus

$$\widetilde{w}(\alpha,\varepsilon)^T\widetilde{x}(\alpha,\varepsilon)\ge 1 \tag{49}$$

for any given sufficiently small (α,ε). Theorem 4.2 claims that when (49) holds with equality for a sufficiently small pair (α,ε), $\widetilde{x}(\alpha,\varepsilon)$ must be a sparsest point of P. In the next subsection, we develop a condition to ensure that (49) holds with equality.

Note that all functions (6)–(8) satisfy Assumption 2.5. We also note that Assumption 4.1 is a mild condition imposed on polyhedral sets. From a theoretical point of view, such a mild assumption is generally not enough to ensure Algorithm 3.3 to solve ℓ_0-minimization, unless some further conditions are imposed on the problem. This is exactly reflected by the condition $\widehat{w}^T\widehat{x}=1$ (i.e., $\widetilde{w}(\alpha,\varepsilon)^T\widetilde{x}(\alpha,\varepsilon)=1$) in Theorem 4.2. The performance analysis in this paper does not rely on the existing assumptions such as mutual coherence, RIP, NSP, and ERC. The condition $\widehat{w}^T\widehat{x}=1$

in Theorem 4.2 is verifiable in contrast to some existing assumptions. In the next section, we will show that the so-called strict regularity of P (the ERC counterpart for polyhedral sets) implies that $\widehat{w}^T\widehat{x} = 1$ if the parameters in Algorithm 3.3 are properly chosen. We will point out that the mutual coherence condition can also imply our condition. Clearly, the standard ℓ_1-minimization (i.e., BP) and OMP only find one solution of ℓ_0-minimization. By contrast, when ℓ_0-minimization admits multiple solutions, Algorithm 3.3 in this paper can find different or all solutions of the problem by varying the initial weight (see the discussion at the end of Section 4.2). Moreover, numerical experiments demonstrate that our algorithm remarkably outperforms BP for solving ℓ_0-minimization in many situations (see Section 5 for details).

4.2 Performance guarantee under strict regularity

In this subsection, we still denote by $(\widetilde{w}(\alpha,\varepsilon), \widetilde{y}(\alpha,\varepsilon), \widetilde{s}(\alpha,\varepsilon))$ the solution of (30), i.e.,

$$\max_{(w,y,s)} \left\{ b^T y + \alpha \mathcal{F}_\varepsilon(s) : \ s = w - A^T y \geq 0, \ b^T y \leq 1, \ (x^0)^T w \leq M, \ 0 \leq w \leq M^* e \right\}$$

and by $\widetilde{x}(\alpha,\varepsilon)$ the solution of (31), i.e.,

$$\min\{\|\widetilde{W}x\|_1 : x \in P\}.$$

Under Assumption 4.1, we have shown that the vector $(\widetilde{w}(\alpha,\varepsilon)$, $\widetilde{y}(\alpha,\varepsilon)$, $\widetilde{s}(\alpha,\varepsilon))$, generated by Step 1 in Algorithm 3.3, satisfies (49) for any sufficiently small (α,ε). We now impose another condition on P to ensure that (49) holds with equality for any sufficiently small (α,ε). The condition is based on the next definition.

Definition 4.6. P is said to be *strictly regular* if there is a sparsest point $x \in P$ such that

$$\|A^T_{\overline{J_+(x)}} A_{J_+(x)} (A^T_{J_+(x)} A_{J_+(x)})^{-1}\|_\infty < 1.$$

Reduced to the sparsest solution of a system of linear equations, it is easy to see that the strict regularity coincides with the exact recovery condition (ERC) introduced in [53]. In general, however, the sparsest point of P is not a sparsest solution of the underlying system of linear equations and vice versa. The strict regularity of P can be viewed as the generalized concept of ERC to polyhedral sets. Clearly, if P is strictly regular, then $\sigma^* < 1$, where σ^* is the constant determined by (32). Under the strict regularity of P, the performance of Algorithm 3.3 can be guaranteed when the initial weight w^0 and parameters $(M, M^*, \alpha, \varepsilon)$ are properly chosen, as shown by the next result.

Theorem 4.7. *Let $\mathcal{F}_\varepsilon(s)$ be a concave and continuously differentiable merit function for sparsity with respect to s over an open set containing $\mathbb{R}^n_+$, and let $\mathcal{F}_\varepsilon(s)$ satisfy Assumption 2.5. Suppose that P is strictly regular and Assumption 4.1 is satisfied. Then there exists a constant $\beta^* > 1$ such that the following holds: For any given $w^0 \in \mathbb{R}^n_{++}$ with $\beta := \|w^0\|_\infty / \left(\min_{1\leq i\leq n} w^0_i\right) < \beta^*$ and for any $M^* \geq M \max\{1, \|w^0\|_\infty/\gamma^0\}$ where $M = 1$ and $\gamma^0 = \min\{\|W^0 x\|_0 : x \in P\}$, there exists $\varepsilon^* \in (0,1)$ such that for any $\varepsilon \in (0, \varepsilon^*]$, there is $\alpha^* \in (0,1)$ such that for any $\alpha \in (0, \alpha^*]$, the point $\widetilde{x} = \widetilde{x}(\alpha,\varepsilon)$ generated by Algorithm 3.3 with such parameters $(M, M^*, \alpha, \varepsilon)$ is a sparsest point of P.*

Proof. Define

$$\beta^* = \begin{cases} \frac{1}{\sigma^*} & \text{if } \sigma^* > 0 \\ +\infty & \text{otherwise,} \end{cases}$$

where σ^* is given by (32). Since P is strictly regular, we have $\sigma^* < 1$, which implies that if $\beta^* < \infty$ then $\beta^* > 1$. Let $w^0 \in R^n_{++}$ be any vector satisfying $\beta = \|w^0\|_\infty / \left(\min_{1\leq i\leq n} w^0_i\right) < \beta^*$ (for instance, $w^0 = (1, 1, \ldots, 1)^T \in \mathbb{R}^n_{++}$ always satisfies this condition). It is easy to see that $\sigma^*\beta < 1$. Let $\sigma = 1/\beta$, then $\sigma > \sigma^*$. As in the proof of Lemma 4.4, we let $\widehat{M} := \max\{1, \sigma\beta\} = 1$. We then set $M = \widehat{M} = 1$ and let M^* be any given number satisfying $M^* \geq \|w^0\|_\infty/\gamma^0$. Repeating the same proof of Theorem 4.2, we conclude that there exists $\varepsilon^* \in (0, 1)$ such that for every fixed $\varepsilon \in (0, \varepsilon^*]$, there is a sufficiently small number $\alpha^* \in (0, 1)$ such that for any $\alpha \in (0, \alpha^*]$, the point $\widetilde{x}(\alpha, \varepsilon)$ generated by Algorithm 3.3 with such parameters $(M, M^*, \alpha, \varepsilon)$ is a sparsest point of P provided that

$$\widetilde{w}(\alpha, \varepsilon)^T \widetilde{x}(\alpha, \varepsilon) = 1. \tag{50}$$

Thus, to prove the theorem, it is sufficient to show that (50) holds under the assumptions of the theorem. In fact, we have shown in the proof of Theorem 4.2 that for any $\alpha \in (0, \alpha^*]$, the solution $(\widetilde{w}(\alpha, \varepsilon), \widetilde{y}(\alpha, \varepsilon), \widetilde{s}(\alpha, \varepsilon))$ of the problem (30), i.e.,

$$\max_{(w,y,s)} \left\{b^T y + \alpha\mathcal{F}_\varepsilon(s) : \ s = w - A^T y \geq 0, \ b^T y \leq 1, \ (x^0)^T w \leq M, \ 0 \leq w \leq M^* e\right\}$$

satisfies the relation (43), i.e., $b^T\widetilde{y}(\alpha, \varepsilon) = 1$ for any $\alpha \in (0, \alpha^*]$. This implies (49), i.e., $\widetilde{w}(\alpha, \varepsilon)^T$ $\widetilde{x}(\alpha, \varepsilon) \geq 1$. From the constraints of the above problem, we also have $\widetilde{w}(\alpha, \varepsilon)^T x^0 \leq M = 1$. Moreover, since $\widetilde{x}(\alpha, \varepsilon)$ is an optimal solution and x^0 is a feasible solution to the problem $\min\{\|\widetilde{W}x\|_1 : x \in P\}$, it follows from $\widetilde{w}(\alpha, \varepsilon)^T \widetilde{x}(\alpha, \varepsilon) \geq 1$ that

$$1 \leq \widetilde{w}(\alpha, \varepsilon)^T \widetilde{x}(\alpha, \varepsilon) \leq \widetilde{w}(\alpha, \varepsilon)^T x^0 \leq M = 1. \tag{51}$$

Thus (50) holds for any sufficiently small (α, ε), as desired. $\square$

It is well known that the performance of BP, OMP and other greedy pursuits can be guaranteed under a verifiable mutual coherence assumption. Similarly, the guaranteed performance of Algorithm 3.3 can be also established under this assumption. For instance, suppose that there is a point $\widehat{x} \in P$ satisfying

$$\|\widehat{x}\|_0 \leq \left(1 + \frac{1}{\mu(A)}\right)/2, \tag{52}$$

where $\mu(A)$ is the mutual coherence of A. As $\widehat{x} \in P$ is also a solution to the system $Az = b$, (52) implies that $\widehat{x}$ is the unique sparsest solution of this system of linear equations (see, e.g., [8, 20]). This in turn implies that $\widehat{x}$ is the unique sparsest point of P. In this case, the set of sparsest points of P coincides with that of the underlying system of linear equations, and hence the strict regularity of P coincides with the ERC condition. It is known that for systems of linear equations, the mutual coherence condition implies the ERC condition (see [53]). Thus the condition (52) implies the strict regularity of P. However, (52) is very restrictive from a practical viewpoint. In this paper, the performance analysis of Algorithm 3.3 is carried out under the assumptions which are remarkably different from those widely assumed conditions in the literature such as mutual coherence, RIP, NSP, and their variants. These conditions in the literature often imply the uniqueness of the sparsest solutions to a system of linear equations. Assumption 4.1 and the strict regularity in this paper allow a polyhedral set to possess multiple sparsest points. For example, $P = \{x : Ax = b, x \geq 0\}$ with

$$A = \begin{bmatrix} 1 & 0 & -1/5 & 1/3 & 0 \\ 0 & 0 & -1/5 & 1/6 & 1/5 \\ 0 & -1 & -3/5 & 1/3 & 4/5 \end{bmatrix}, \ b = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} \tag{53}$$

contains distinct sparsest points: $x^* = (1,1,0,0,0)^T$ and $\widetilde{x}^* = (0,0,5,6,0,)^T$. It is straightforward to verify that $A^T d \leq 0$ implies $d = 0$. Thus Assumption 4.1 is satisfied for this example. Moreover, it is easy to verify that at $x^* = (1,1,0,0,0)^T$,

$$\left\| A^T_{\overline{J_+(x^*)}} A_{J_+(x^*)} (A^T_{J_+(x^*)} A_{J_+(x^*)})^{-1} \right\|_\infty = \left\| \begin{bmatrix} -1/5 & 3/5 \\ 1/3 & -1/3 \\ 0 & -4/5 \end{bmatrix} \right\|_\infty = \frac{4}{5} < 1,$$

which implies that P is strictly regular.

It is worth mentioning that when P admits multiple sparsest points, the point $\widehat{x}$ found by Algorithm 3.3 has the least weighted ℓ_1-norm with respect to the weight $\widehat{W} = \mathrm{diag}(\widehat{w})$, i.e., $\|\widehat{W}\widehat{x}\|_1 \leq \|\widehat{W}x\|_1$ for all $x \in P$, where $\widehat{w}$ is generated at Step 1 of the algorithm. Clearly, a change of $\widehat{w}$ may result in a different sparsest point. By the structure of Algorithm 3.3, $\widehat{w}$ depends on the choice of w^0. This indicates that every sparsest point of P might be located by the algorithm if different initial weights are used. As an example, all sparsest points of P with (53) can be found by Algorithm 3.3 with different initial weights. Indeed, using function (6), $\alpha = \varepsilon = 10^{-8}$, $M = 10$ and $M^* = M(\max\{1, \|w^0\|_\infty/\gamma^0\} + 1)$ in Algorithm 3.3, we immediately obtain the following results for (53) (see details to how to do so in Section 5): If $w^0 = (1,1,1,1,1)^T$, then $\widehat{w} = (0.5096, 0.4904, 5.2493, 5.1408, 4.8934)^T$ is generated by the algorithm which yields the sparsest point $x^* = (1,1,0,0,0)^T$; if $w^0 = (1, 0.5, 0.1, 0.1, 0.1)^T$, then $\widehat{w} = (5.9349, 5.7046, 0.1047, 0.0794, 5.2895)^T$ is found by the algorithm which yields the sparsest point $\widetilde{x}^* = (0,0,5,6,0)^T$.

5 Numerical experiments

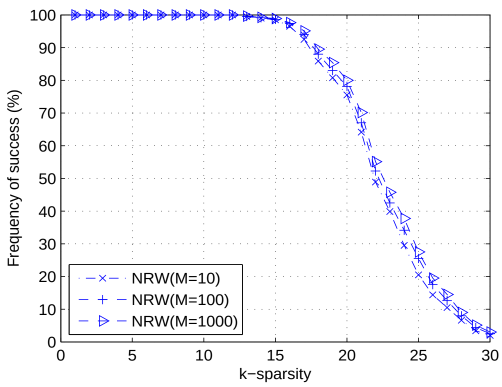
To evaluate the efficiency of Algorithm 3.3, we consider random examples of polyhedral sets. The entries of A and the sparse vector x^* are generated as i.i.d Gaussian random variables with zero mean and unit variance. For every generated matrix $A \in \mathbb{R}^{m \times n}$ $(m < n)$ and k-sparse vector $x^* \in \mathbb{R}^n_+$, we set $b = Ax^*$. Then the polyhedral set $P = \{x : Ax = b, x \geq 0\}$ contains the known k-sparse vector x^*. All examples of polyhedral sets in our experiments are generated this way. We perform a large number of random trials of polyhedral sets in order to demonstrate the performance of our algorithm in locating the sparsest points of these polyhedral sets. When x^* is the unique sparsest point of P, we say that x^* is exactly recovered by an algorithm if x^* is the solution found by the algorithm. However, this is not the only case for the success of an algorithm in solving ℓ_0-minimization. When P admits multiple sparsest points, an algorithm can still succeed in solving ℓ_0-minimization provided that the solution generated by the algorithm is one of the sparsest points of P. Thus the criterion *"exact recovery of x^*"* does not properly measure the success rate of an algorithm when P has multiple sparsest points. Thus we count one "success" if the found solution $\widetilde{x}$ satisfies that $\|\widetilde{x}\|_0 \leq \|x^*\|_0$. We use such a criterion to measure the success rate of our algorithm.

In what follows, Algorithm 3.3 is referred to as the "NRW" algorithm. We use (6) as the default merit function (unless otherwise stated) and we take $w^0 = e$ as the initial weight. With this w^0 and a prescribed number $M \geq 1$, we set

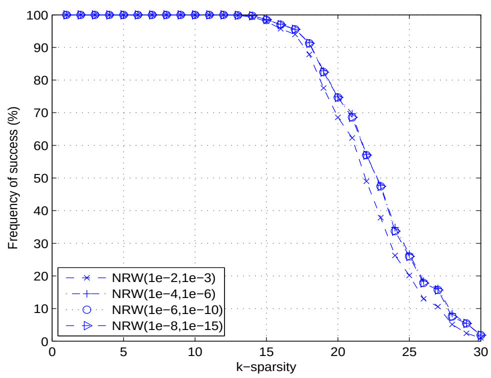
$$M^* = M(\max(1, 1/\gamma^0) + 1) \tag{54}$$

which satisfies that $1 \leq M \leq M^*$ and $M^* \geq M\|w^0\|_\infty/\gamma^0$. We implement the algorithm in MATLAB and use the CVX [27] with solver "sedumi" to solve all convex optimization problems (30) and (31).

From the structure of the NRW and theoretical analysis in Section 4, we see that (M, M^*) should be taken to be relatively large. In the above-mentioned environment, we have carried out an experiment to illustrate how the choice of M might affect the performance of the algorithm. We fix $(\alpha, \varepsilon) = (10^{-8}, 10^{-15})$ and compare the performance of our algorithm with different M. For every sparsity level of $x^* \in \mathbb{R}^{200}_+$ (ranged from 1 to 30), we performed 800 trials of (A, x^*) where $A \in \mathbb{R}^{50\times 200}$. The success frequencies of the NRW with $M = 10, 100$ and 1000, respectively, are shown in Fig. 1(i), from which it can be seen that the NRW with a larger M performs slightly better than the algorithm with a smaller M. However, such a difference between their performances is not remarkable, especially when M is relatively large. Thus we set $M = 100$ as a default in our algorithm.



(i) Different choices of M



(ii) Different choices of (α, ε)

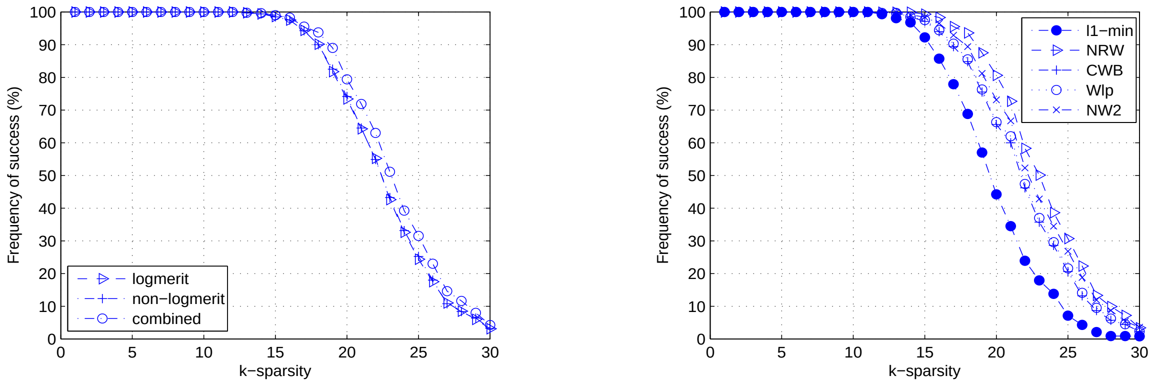
Figure 1: Comparison of the performance of the NRW algorithm with different M and (α, ε).

From the analysis in Section 4, we see that (α, ε) should be taken to be small. To see how the choice of (α, ε) might influence the performance of the NRW method, we compare the algorithms with different choices of (α, ε). The success rates of the NRW with $(\alpha, \varepsilon) = (10^{-2}, 10^{-3})$, $(10^{-4}, 10^{-6})$, $(10^{-6}, 10^{-10})$ and $(10^{-8}, 10^{-15})$ are given in Fig. 1(ii). In this experiment, a total of 800 trials of (A, x^*), where $A \in \mathbb{R}^{50\times 200}$, were run for every given sparsity level of $x^* \in \mathbb{R}^{200}_+$. Fig. 1(ii) demonstrates that the performance of the NRW method is insensitive to the choice of (α, ε) provided that they are small enough. Thus we set

$$(\alpha, \varepsilon) = (10^{-8},\ 10^{-15}) \tag{55}$$

as the default parameters in our algorithm.

Under (55) and the choice $M = 100$, we can also compare the performance of our algorithm with the following merit functions: (6) (termed "logmerit"), (7) (termed "non-logmerit"), and the sum of (6) and (7) (termed "combined"). Note that any positive combination of merit functions yields a new merit function. Thus the sum of (6) and (7) is also a merit function for sparsity. For every sparsity level of $x^* \in \mathbb{R}^{200}_+$, we performed 1000 trials of (A, x^*), where $A \in \mathbb{R}^{50\times 200}$. The result is given in Fig. 2(i) which indicates that there is no remarkable difference between the performance of the NRW algorithm with these merit functions.



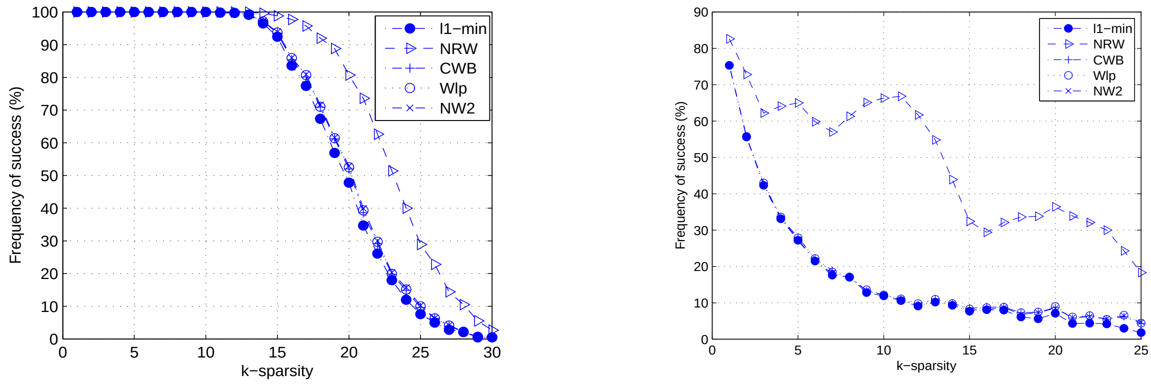
(i) Different merit functions (ii) $\mu = 10^{-2}$ in (56)

Figure 2: (i) Performance comparison of the NRW algorithm with different choices of merit functions. (ii) Comparison of algorithms when $\mu = 10^{-2}$ is taken in CBW, Wlp, and NW2 algorithms.

We now compare the performance of the NRW method and ℓ_1-minimization (basis pursuit) and several other iteratively reweighted ℓ_1-methods such as the one proposed in [11] (termed "CWB"), the one in [22] (termed "Wlp"), and the "NW2" method presented in [62]. In this experiment, we use $M = 100$, (54), (55) and (6) in the NRW method. The following weights

$$w^k = \frac{1}{|x^k| + \mu}, \quad w^k = \frac{1}{(|x^k| + \mu)^{1-p}}, \quad w^k = \frac{q + (|x_i^k| + \mu)^{1-q}}{(|x_i^k| + \mu)^{1-q} \left[|x_i^k| + \mu + (|x_i^k| + \mu)^q \right]^{1-p}}, \tag{56}$$

where $p, q, \mu \in (0, 1)$ are given parameters, are used in CBW, Wlp and NW2 algorithms, respectively. The parameters $p = q = 0.05$ are set in Wlp and NW2 algorithms. The standard ℓ_1-minimization is used as the initial step for all these algorithms. Simulations have indicated that the performance of these iteratively reweighted methods is sensitive to the choice of μ. For small μ, their performances are very similar to that of ℓ_1-minimization. In this case, if the initial step fails to find a sparsest point, the remaining iterations of algorithms are also likely to fail. This suggests that μ in (56) should not be chosen to be too small. Empirical results have also indicated that if a reweighted ℓ_1-algorithm can solve an ℓ_0-minimization problem, it usually solves the problem within a few iterations, and if the algorithm fails to locate the sparsest point within the first few iterations, it is usually difficult for the algorithm to find a sparsest point even if more iterations are carried out. In addition, ℓ_0-minimization is a global optimization problem, for which no optimality condition has been developed. Thus, in all our experiments, the CWB, Wlp and the NW2 were performed for a total of 5 iterations on every generated problem. It is worth noting that the NRW algorithm only consists of a single iteration. To compare these algorithms, we performed 1000 trials of (A, x^*) for every sparsity level of x^*, where $A \in \mathbb{R}^{50 \times 200}$. The success frequencies of these algorithms when applied to these polyhedral sets are given in Fig. 2(ii) and Fig. 3(i), in which $\mu = 10^{-2}$ and $\mu = 10^{-3}$ are used in CWB, Wlp and NW2 algorithms, respectively. From these experiments, we see that the NRW method remarkably outperforms the standard ℓ_1-minimization and it performs better than CWB, Wlp and NW2 as well, especially as μ in (56) is relatively small. Experiments indicate that when $\mu \leq 10^{-4}$, the performance of CWB, Wlp and NW2 is almost identical to that of ℓ_1-minimization (this has already been observed from Fig. 3(i) when $\mu = 10^{-3}$).



(i) $\mu = 10^{-3}$ in (56) (ii) $A = [G\ B\ G\ D] \in R^{50\times 200}$

Figure 3: (i) Comparison of algorithms when $\mu = 10^{-3}$ is taken in CBW, Wlp, and NW2 algorithms. (ii) Comparison of algorithms when P admits multiple sparsest points.

Finally, we compare the performance of algorithms when applied to polyhedral sets which are very likely to possess multiple sparsest points. A simple way of constructing such examples is to generate A at random but allow A to have some repeated columns. For instance, let $A = [G\ B\ G\ D] \in \mathbb{R}^{50\times 200}$ where $G \in \mathbb{R}^{50\times 25}, B, D \in \mathbb{R}^{50\times 75}$ are random matrices, and let $x^* \in \mathbb{R}^{200}_+$ be a randomly generated k-sparse vector. Setting $b = Ax^*$, we see that the generated P is very likely to possess multiple sparsest points and multiple ℓ_1-minimizers thanks to repeated columns in A. We ran 1000 such trials for each given sparsity level $k = 1, 2, \ldots, 25$. The performance of algorithms is shown in Fig. 3(ii). It can be seen that the NRW method has performed remarkably better than ℓ_1-minimization, CWB, Wlp and NW2 in this situation.

6 Conclusions

Based on the strict complementarity theory of linear programs, we have shown that seeking the sparsest point of a polyhedral set is equivalent to solving a bilevel programming problem with ℓ_0-maximization as its outer layer and with weighted ℓ_1-minimization as its inner problem. As a result, locating the sparsest point of a polyhedral set can be transformed to searching for the densest possible slack variable of the dual problem of weighted ℓ_1-minimization. This property provides a new basis to understand ℓ_0-minimization, leading to a new development of weighted ℓ_1-algorithms. This new method computes the weight through a certain convex optimization instead of defining the weight directly by the iterate. The efficiency of the proposed method has been demonstrated by empirical results and has been shown under some assumptions, including the monotonicity of the merit function for sparsity and boundedness of the feasible set of the dual problem of weighted ℓ_1-minimization. These assumptions do not require the uniqueness of the sparsest points of polyhedral sets.

References

[1] E. Amaldi and V. Kann, On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems, *Theoret. Comput. Sci.*, 209 (1998), pp. 237–260.

[2] M.S. Asif and J. Romberg, Fast and accurate algorithms for re-weighted ℓ_1-norm minimization, *IEEE Trans. Signal Process.*, 61 (2013), pp. 5905–5916.

[3] J. Bardsley and J. Nagy, Covariance-preconditioned iterative methods for nonnegativity constrainted astronomical imaging, *SIAM J. Matrix Anal. Appl.*, 27 (2006), pp. 1184–1198.

[4] A. Beck and M. Teboulle, A fast iterative shrinkage-thresholding algorihm for linear inverse problems, *SIAM J. Imaging Sci.*, 2 (2009), pp. 183–202.

[5] T. Blumensath, M. Davies and G. Rilling, *Greedy algorithms for compressed sensing*, in Compressed Sensing: Theory and Applications (Y. Eldar and G. Kutyniok Eds.), Cambridge University Press, 2012

[6] P.S. Bradley, O.L. Mangasarian and J.B. Rosen, Parsimonious least norm approximation, *Comput. Optim. Appl.*, 11 (1998), pp. 5–21.

[7] P.S. Bradley, U.M. Fayyad and O.L. Mangasarian, Mathematical programming for data mining: formulations and challenges, *INFORMS J. Computing*, 11 (1999), pp. 217–238.

[8] A.M. Bruckstein, D. Donoho and M. Elad, From sparse solutions of systems of equations to sparse modeling of signals and images, *SIAM Rev.*, 51 (2009), pp. 34–81.

[9] A.M. Bruckstein, M. Elad and M. Zibulevsky, On the uniqueness of nonnegative sparse solutions to underdetermined systems of equations, *IEEE Trans. Inform. Theory*, 54 (2008), pp. 4813–4820.

[10] E. Candès and T. Tao, Decoding by linear programming, *IEEE Trans. Inform. Theory*, 51 (2005), pp. 4203–4215.

[11] E. Candès, M. Wakin and S. Boyd, Enhancing sparsity by reweighted ℓ_1 minimization, *J. Fourier Anal. Appl.*, 14 (2008), pp. 877–905.

[12] S. Chen, D. Donoho, and M. Saunders, Atomic decomposition by basis pursuit, *SIAM J. Sci. Comput.*, 20 (1998), pp. 33–61.

[13] X. Chen and W. Zhou, Convergence of reweighted ℓ_1 minimization algorithm for $\ell_2 - \ell_p$ minimization, *Comput. Optim. Appl.,* 59 (2014), pp. 47–61.

[14] A. Cohen, W. Dahmen and R. Devore, Compressed sensing and best k-term aproximation, *J. Amer. Math. Soc.*, 22 (2009), pp. 211–231.

[15] I. Daubechies, R. DeVore, M. Fornasier and C.S. Güntürk, Iteratively reweighted least squares minimization for sparse recovery, *Comm. Pure Appl. Math.*, 63 (2010), pp. 1–38.

[16] I. Daubechies, M. Defrise and C. D. Mol, An iterative thresholding algorithm for linear inverse problems with a sparsity constraint, *Comm. Pure Appl. Math.*, 57 (2004), pp. 1413–1457.

[17] G. Davis, S. Mallat and Z. Zhang, Adaptive time-frquency decompositions, *Optical Egineering*, 33 (1994), pp. 2183–2191.

[18] D. Donoho and X. Huo, Uncertainty principles and ideal atomic decomposition, *IEEE Trans. Inform. Theory,* 47 (2001), pp. 2845–2862.

[19] D. Donoho and J. Tanner, Sparse nonnegative solutions of underdetermined linear equations by linear programming, *Proc. Natl. Acad. Sci.*, 102 (2005), pp. 9446–9451.

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

[21] Y. Eldar and G. Kutyniok, *Compressed Sensing: Theory and Applications*, Cambridge University Press, 2012.

[22] S. Foucart and M. Lai, Sparsest solutions of undertermined linear systems via ℓ_p-minimization for $0 < q \leq 1$, *Appl. Comput. Harmon. Anal.*, 26 (2009), pp. 395–407.

[23] M.P. Friedlander and P. Tseng, Exact regularization of convex programs, *SIAM J. Optim.*, 18 (2007), pp. 1326–1350.

[24] J.J. Fuchs, On sparse representations in arbitrary redundant bases, *IEEE Trans. Inform. Theory,* 50 (2004), pp.1341–1344.

[25] A.J. Goldman and A.W. Tucker, *Theory of linear programming*, in Linear Inequalities and Related Systems (edited by H.W. Kuhn and A.W. Tucker), 1956, pp. 53–97.

[26] I. Gorodnitsky and B. Rao, Sparse signal reconstruction from limited data using FOCUSS: A weighted minimum norm algorithm, *IEEE Trans. Signal Process.*, 45 (1997), pp. 600–616.

[27] M. Grant and S. Boyd, *CVX: Matlab software for disciplined convex programming*, Version 1.21, April 2011.

[28] R. He, W. Zheng, B. Hu and X. Kong, *Nonnegative sparse coding for discriminative semi-supervised learning*, in Proceedings of IEEE conference on Computer Vision and Pattern Recognition (CVPR), 2011, pp. 2849–2856.

[29] P.O. Hoyer, *Nonnegative sparse coding*, in Proc. of the 12th IEEE Workshop, 2002, pp. 557–565.

[30] A. Juditski and A. Nemirovski, On verifiable sufficient conditions for sparse signal recovery via ℓ_1 minimization, *Math. Program. Ser. B*, 127 (2011), pp. 57–88.

[31] M. Khajehnejad, A. Dimakis, W. Xu and B. Hassibi, Sparse recovery of nonnegative signals with minima expansion, *IEEE Trans. Signal Process.*, 59 (2011), pp. 196–208.

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

[33] Z.Q. Luo, J.S. Pang and D. Ralph, *Mathematical Programs with Equilibrium Constraints*, Cambridge University Press, 1996.

[34] S. Mallat and Z. Zhang, Matching pursuits with time-frequency dictionaries, *IEEE Trans. Signal Process.*, 41 (1993), pp. 3397–3415.

[35] O.L. Mangasarian, *Machine learning via polydedral concave minimization*, in Applied Mathematics and Parallel Computing-Festschrift for Klaus Ritter (H. Fischer, B. Riedmueller and S. Schaeffler eds.), Springer, Heidelberg, 1996, pp. 175–188.

[36] O.L. Mangasarian, Minimum-support solutions of polyhedral concave programs, *Optimization*, 45 (1999), pp. 149–162.

[37] O.L. Mangasarian and R.R. Meyer, Nonlinear perturbation of linear programs, *SIAM J. Control & Optim.*, 17 (1979), pp. 745–752.

[38] H. Mansour and O. Yilmaz, Support driven reweighted ℓ_1-minimization, ICASSP 2012, IEEE, pp. 3309–3312.

[39] B.K. Natarajan, Sparse approximate solutions to linear systems, *SIAM J. Comput.*, 24 (1995), pp. 227–234.

[40] D. Needell, *Noisy signal recovery via iterative reweighted ℓ_1-minimization*, In Proceedings of the 43rd Asilomar conference on Signals, Systems and Computers, Asilomar'09, 2009, pp. 113–117.

[41] D. Needell and R. Vershynin, Uniform uncertainty principle and signal recovery via regularized orthorgonal matching pursuit, *Found. Comput. Math.*, 9 (2009), pp. 317–334.

[42] P.D. O'Grady and S.T. Rickard, *Recovery of nonnegative signals from compressively sampled observations via nonnegative quadratic programming*, SPASR'09, 6 April, 2009.

[43] H. Park and H. Kim, Nonnegative matrix factorization based on alternating nonnegativity constrainted least squares and the active set method, *SIAM J. Matrix Anal. Appl.*, 30 (2008), pp. 713–730.

[44] P. Pauca, J. Piper and R. Plemmons, Nonnegative matrix factorization for spectral data analysis, *Linear Algebra Appl.*, 416 (2006), pp. 29–47.

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

[46] R.T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, New Jersey, 1972.

[47] S.-S. Shai and B.-D. Shai, *Understanding Machine Learning from Theory to Algorithms*, Cambridge University Press, New York, 2014.

[48] L. Shen and B.W. Suter, *Blind one-bit compressive sampling*, arXiv:1302.1419vl, 6 Feb 2013.

[49] J. Starch, F. Murtagh, and J. Fadili, *Sparse Image and Signal Processing: Wavelets, Curvelets, Morphological Diversity*, Cambridge University Press, 2010.

[50] A. Szlam, Z. Guo and S. Osher, *A split Bregman method for nonnegative sparsity penalized least squares with applications to hyperspectral demixing*, in IEEE Inter. Conf. on Imaging Processing, 2010, pp. 1917–1920.

[51] G. Tang and A. Nehorai, Verfiable and computational performance analysis of sparsity recovery, Technical Report, arXiv:1102.4868v2, 5 Oct, 2011.

[52] V. TemlyaKov, Nonlinear methods of approximation, *Found. Comput. Math.*, 3 (2003), pp. 33–107.

[53] J.A. Tropp, Greed is good: Algorithmic results for sparse approximation, *IEEE Trans. Inform. Theory*, 50 (2004), pp. 2231–2242.

[54] J.A. Tropp and A.C. Gilbert, Signal recovery from random measurements via orthogonal mathcing pursuit, *IEEE Trans. Inform. Theory*, 53 (2007), pp. 4655–4666.

[55] N. Vo, B. Moran and S. Challa, *Nonnegative-least-square classifier for face recognition*, in Proc. of the 6th Inter. Symposium on Neural Networks: Advances in Neural Networks, 2009, pp. 449–456.

[56] M. Wang, W. Xu and A.Tang, A unique nonnegative solution to an underdetermined system: from vectors to matrices, *IEEE Trans. Signal Proc.*, 59 (2011), pp. 1107–1016.

[57] D. Wipf and S. Nagarajan, Iterative reweighted ℓ_1 and ℓ_2 methods for finding sparse solutions, *IEEE J. Sel. Topics Signal Process.*, 4 (2010), pp. 317–329.

[58] Y. Zhang, *A simple proof for recoverability of ℓ_1-minimization (II): the nonnegative case*, Technical Report, Rice University, 2005.

[59] Y. Zhang, Theory of compressive sensing via ℓ_1-mimimization: A non-RIP analysis and extensions, *J. Oper. Res. Soc. China*, 1 (2013), pp. 79–105.

[60] Y.-B. Zhao, Equivalence and strong equivalence between the sparsest and least ℓ_1-norm nonnegative solutions of linear systems and their applications, *J. Oper. Res. Soc. China*, 2 (2014), pp. 171–193

[61] Y.-B. Zhao, RSP-based analysis for sparsest and least ℓ_1-norm solutions to underdetermined linear systems, *IEEE Trans. Signal Process.*, 61 (2013), no. 22, pp. 5777–5788.

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

[63] S. Zhou, N. Xiu, Y. Wang and L. Kong, Exact recovery for sparse signal via weighted ℓ_1-minimizaton, Technical Report, arXiv:1312.2358, 2013.