REWEIGHTED $\ell_1$-MINIMIZATION FOR SPARSE SOLUTIONS TO UNDERDETERMINED LINEAR SYSTEMS

YUN-BIN ZHAO† AND DUAN LI‡

Abstract. Numerical experiments have indicated that the reweighted $\ell_1$-minimization performs exceptionally well in locating sparse solutions of underdetermined linear systems of equations. We show that reweighted $\ell_1$-methods are intrinsically associated with the minimization of the so-called merit functions for sparsity, which are essentially concave approximations to the cardinality function. Based on this observation, we further show that a family of reweighted $\ell_1$-algorithms can be systematically derived from the perspective of concave optimization through the linearization technique. In order to conduct a unified convergence analysis for this family of algorithms, we introduce the concept of range space property (RSP) of a matrix, and prove that if its adjoint has this property, the reweighted $\ell_1$-algorithm can find a sparse solution to the underdetermined linear system provided that the merit function for sparsity is properly chosen. In particular, some convergence conditions for the Candès-Wakin-Boyd method and the recent $\ell_p$-quasi-norm-based reweighted $\ell_1$-method can be obtained as special cases of the general framework.

Key words. Reweighted $\ell_1$-minimization, sparse solution, underdetermined linear system, concave minimization, merit function for sparsity, compressed sensing.

AMS subject classifications. 15A06, 15A29, 65K05, 90C25, 90C26, 90C59

1. Introduction. Given an $m \times n$ matrix $A$ with $m \leq n$ and a nonzero vector $b \in \mathbb{R}^m$, the linear system $Ax = b$ has infinitely many solutions when the system is underdetermined. Depending on the nature of source problems, we are often interested in finding a particular solution, and thus optimization methods come into a play through certain merit functions that measure the desired special structure of the solution.

One of the recent interests is to find the sparsest solution of an underdetermined linear system, which has found many applications in signal and image processing [16, 3, 2]. To find a sparsest solution of $Ax = b$, perhaps the ideal merit function is the cardinality of a vector, denoted by $\|x\|_0$, i.e., the number of nonzero components of $x$. Clearly, the set of the sparsest solutions of $Ax = b$ coincides with the set of solutions to the cardinality minimization problem

$$(P_0) \quad \text{minimize} \{\|x\|_0 : Ax = b\},$$

which is an NP-hard discrete optimization problem [35]. The recent study in the field of compressed sensing nevertheless shows that not all cardinality minimization problems are equally hard, and there does exist a class of matrices $A$ such that the problem $(P_0)$ is computationally tractable. These matrices can be characterized by such concepts as the spark which was formally defined by Donoho and Elad [17], restricted isometry property (RIP) introduced by Candès and Tao [8], mutual coherence (MC) [34, 18, 17], and null space property (NSP) [17, 27, 13, 43].

---

*This work was partially supported by the Research Grants Council of Hong Kong under grant CUHK 414610.
†School of Mathematics, The University of Birmingham, Edgbaston, Birmingham B15 2TT, United Kingdom (y.zhao.2@bham.ac.uk).
‡Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, NT, Hong Kong (dli@se.cuhk.edu.hk).
Problem \((P_0)\) is not easy to solve in general. From a convex analysis point of view, a natural methodology is to minimize the convex envelope of \(\|x\|_0\). It is well-known that \(\ell_1\)-norm is the convex envelope of \(\|x\|_0\) over the region \(\{x : \|x\|_\infty \leq 1\}\). One of the main approaches to attack \((P_0)\) is through \(\ell_1\)-minimization

\[
(P_1) \quad \text{minimize } \{\|x\|_1 : Ax = b\},
\]

which is identical to a linear program (LP) and hence can be solved very efficiently.

Using \(\ell_1\)-norm as a merit function for sparsity can be traced back several decades in a wide range of areas from seismic traces [40], sparse-signal recovery [19, 11], sparse-model selection (LASSO algorithm) in statistics [39] to image processing [1], and continues its growth in other areas. A brief history of using \(\ell_1\)-minimization can be found in [9]. The \(\ell_1\)-minimization approach has been generalized to low-rank matrix recovery/matrix completion (see e.g., [4, 7]), and to the so-called matrix rank minimization as well (see e.g., [20, 21, 22, 37, 38, 44]). Thanks to the intensive study in the field of compressed sensing (see e.g., [16, 3]), both theoretical properties and numerical performances of \(\ell_1\)-minimization have been well-established over the past few years. Various conditions (including the above-mentioned MC, RIP, NSP, and others) for the relationship

\[
\arg \min \{\|x\|_0 : Ax = b\} = \{x^*\} = \arg \min \{\|x\|_1 : Ax = b\}
\]

have been developed (see e.g., [18, 17, 27, 5, 6, 41, 2, 13, 43]). In terms of sparse signal recovery, this relationship implies that \(\ell_1\)-minimization allows recovery of sparse signals from a small number of measurements [17, 16, 5, 6]. A comprehensive discussion and survey of recent results in this field can be found in [16, 3, 2, 23].

Inspired by the efficiency of \(\ell_1\)-minimization, it is natural to ask whether there are other alternatives which can either be comparable to or even outperform \(\ell_1\)-minimization in finding sparse solutions of linear systems. Numerical experiments indicate that the reweighted \(\ell_1\)-minimization does outperform unweighted \(\ell_1\)-minimization in many situations [9, 24, 31, 12, 14]. The key feature of reweighted \(\ell_1\)-minimization is the solution of a series of weighted \(\ell_1\)-problems

\[
(P_{W_k}) \quad \text{min } \{\|W_k x\|_1 : Ax = b\},
\]

where \(W_k = \text{diag}(w^k)\) and \(w_k = (w^k_1, \ldots, w^k_n)^T \in \mathbb{R}^n_+\) is the vector of weights determined by the current iterate \(x^k = (x^k_1, \ldots, x^k_n)^T \in \mathbb{R}^n\). The solution to \((P_{W_k})\) is set to be \(x^{k+1}\), based on which the new weight \(W^{k+1}\) is computed. Some theoretical analysis has been made for reweighted \(\ell_1\)-algorithms since 2008, when Candès, Wakin and Boyd [9] proposed the reweighted method with

\[
w^k_i = \frac{1}{|x^k_i| + \varepsilon}, \quad i = 1, \ldots, n, \quad \text{for } \varepsilon > 0. \tag{1.1}
\]

We refer to this as the CWB method in this paper. Needell [36] showed that the error bounds for noisy signal recovery via the CWB method can be tighter than those of standard \(\ell_1\)-minimization. Foucart and Lai [24] proved that under the assumption of RIP, the reweighted \(\ell_1\)-method with weights

\[
w^k_i = \frac{1}{(|x^k_i| + \varepsilon)^{1-p}}, \quad i = 1, \ldots, n, \tag{1.2}
\]
where \( p \in (0, 1) \) is a given parameter, can exactly recover the sparse signal. Lai and Wang [31], and Chen and Zhou [12] further prove that under RIP/NSP-type conditions, the accumulation points of the sequence generated by the reweighted \( \ell_1 \)-algorithm with weights (1.2) can converge to a stationary point of certain \( \ell_2 - \ell_p \) minimization problem that is an approximation to \((P_0)\). Note that the objective of \((P_{W_k})\) is separable in \( x \). It is worth mentioning that some nonseparable iterative reweighted methods were also proposed recently by Wipf and Nagarajan [42]. However, as pointed out in [9], the understanding of reweighted \( \ell_1 \)-minimization remains very incomplete so far. Even the convergence property of the CWB algorithm remains unclear at present.

We note that while the major study of reweighted \( \ell_1 \)-minimization is carried out recently, the reweighted least square (RLS) method has a relatively long history. RLS was proposed by Lawson [32] in 1960s, and was extended to \( \ell_p \)-minimization later. The idea of RLS methods was also used in the algorithm for robust statistical estimation [29], and in FOCUSS methods [25] for the sparse solution of linear systems. The interplay of null space property (NSP), \( \ell_1 \)-minimization, and RLS method has been clarified recently in [14].

The main contributions of this paper are as follows. First, we provide a unified derivation of the reweighted \( \ell_1 \)-minimization, which can be viewed as the first-order method for concave programming with an objective called the merit function for sparsity that is certain approximation of \( \| x \|_0 \). Second, we provide a unified theoretical analysis for a large family of reweighted \( \ell_1 \)-algorithms for the sparse solution of underdetermined linear systems. Interestingly, various new reweighted \( \ell_1 \)-methods can be systematically constructed/extracted from this family. To show the generic convergence of this family of algorithms, we introduce the new concept of range space property (RSP) of a matrix, which is different from (but has some link to) RIP and NSP. One of the results in this paper claims that if \( A^T \) has the RSP of order \( K \) with constant \( \rho > 0 \) satisfying \((1 + \rho)K < n\), then there exist a large number of merit functions for sparsity, associated with which reweighted \( \ell_1 \)-algorithms can generate a \([((1 + \rho)K)]\)-sparse solution to the linear system. Based on optimization theory merely, the analysis in this paper is remarkably different from the existing RIP/NSP-based analysis. It should be stressed that the CWB method, and the algorithm with weights (1.2) are special cases of our general framework of reweighted \( \ell_1 \)-algorithms, and hence a new convergence result for these existing algorithms under RSP assumption has been established for the first time in this paper. Finally, we carry out some numerical experiments to demonstrate the performance of several new and existing reweighted \( \ell_1 \)-algorithms in locating the sparsest solution of linear systems. Our numerical results show that in many situations the NW1-NW4 algorithms proposed in this paper, the CWB method, and the algorithm with weights (1.2) do remarkably outperform the standard \( \ell_1 \)-minimization (See section 4 for details).

This paper is organized as follows. In section 2, a unified approach for deriving a large family of reweighted \( \ell_1 \)-algorithms based on concave merit functions for sparsity is proposed. In section 3, some convergence properties of this family of algorithms are proved via the range space property. Numerical results are given in section 4, and conclusions are given in the last section.

Notation: Let \( \mathbb{R}^n \) be the \( n \)-dimensional Euclidean space, and \( \mathbb{R}_+^n \) and \( \mathbb{R}_++^n \) the sets of nonnegative and positive vectors, respectively. For \( x \in \mathbb{R}^n \), \( \| x \|_p = (\sum_{i=1}^n |x_i|^p)^{1/p} \) denotes the \( \ell_p \)- (quasi-)norm where \( p \in (0, \infty) \). Given a set \( S \subseteq \{1, 2, ..., n\} \), the symbol \(|S|\) denotes the cardinality of \( S \), and \( S_c = \{1, 2, ..., n\} \setminus S \) is the complement.
of $S$. For a matrix $A$ and a vector $x \in \mathbb{R}^n$, we use the notation $A_S$ to denote the submatrix extracted from $A$ with column indices in $S$, and $x_S$ the subvector extracted from $x$ with component indices in $S$. Let $T(x) = \{i : x_i \neq 0\}$ denote the support of $x$. Let $\sigma(x)$ be the vector which is the non-increasing rearrangement of the absolute values of the entries of $x \in \mathbb{R}^n$, and let $|x| = ([x_1], ..., [x_n])^T \in \mathbb{R}^n$. Clearly, we have $\sigma(x) = \sigma(|x|)$, and $\sigma(x)_i$ is the $i$th largest component of $|x|$. In this paper, $x \in \mathbb{R}^n$ is said to be $K$-sparse if $x$ contains at most $K$ nonzero components. Thus $x$ is $K$-sparse if and only if $\sigma(x)_K+1 = 0$. For any $x$ and $y$ in $\mathbb{R}^n$, the inequality $x \leq y$ means $x_i \leq y_i$ for all $i = 1, ..., n$.

2. A unified framework of reweighted $\ell_1$-minimization. The central idea of reweighted $\ell_1$-algorithms is to define a weight based on the current iterate $x^k$, solve the weighted $\ell_1$-minimization for this weight, and then use its solution to define a new weight. The weight is used to penalize the components which are small, in order to drive them to be as small as possible via minimizing the weighted $\ell_1$-norm. In this section, we introduce a unified framework for reweighted $\ell_1$-minimization. To this end, we need to specify a family of merit functions for sparsity. A function from $\mathbb{R}^n$ to $\mathbb{R}$ is said to be a merit function for sparsity if it is an approximation to $\|x\|_0$ in some sense. For instance, $\ell_1$-norm is a convex relaxation of $\|x\|_0$ over $\|x\|_\infty \leq 1$, and $\|x\|_p$, $p \in (0, 1)$, is also a merit function for sparsity since $\|x\|_p \to \|x\|_0$ as $p \to 0$. Clearly, there exist a vast number of merit functions for sparsity. Minimizing such a function may drive the variable $x$ to become sparse when a sparse solution exists.

From a computational point of view, a merit function for sparsity should admit certain desired properties such as convexity or concavity. Due to the NP-hardness of $(P_0)$, it seems that there is no hope to approximate $\|x\|_0$ in $\mathbb{R}^n$ to any level of accuracy by a convex function. On the contrary, there exist various concave functions that can approximate $\|x\|_0$ to any level of accuracy. Concave merit functions appear more natural than convex ones when finding the sparsest solution of linear systems, since the ‘bulged’ feature of convex merit functions might prohibit locating the sparsest solution in some situations. This phenomenon was observed by Harikumar and Bresler [28]. So, throughout the remainder of this paper, we focus on concave merit functions for sparsity. For simplicity, we consider separable concave merit functions of the form

$$F(x) = \sum_{i=1}^n \phi_i(|x_i|),$$

where $\phi_i : \mathbb{R}_+ \to \mathbb{R}$ is called the kernel function. Given a separable function $F(x)$ as above and a set $S \subseteq \{1, ..., n\}$, we use $F(x_S)$ to denote the reduced separable function, i.e., $F(x_S) := \sum_{i \in S} \phi_i(|x_i|)$. To avoid the division by zero when computing the gradient of a merit function, we perturb the function by replacing $|x_i|$ by $|x_i| + \varepsilon$, where $\varepsilon > 0$ is a given parameter. This leads to the approximation problem of $(P_0)$:

$$\min \left\{ F_\varepsilon(x) = \sum_{i=1}^n \phi_i(|x_i| + \varepsilon) : Ax = b \right\}. \quad (2.1)$$

For example, if all $\phi_i(t) = t^p$ with $p \in (0, 1)$, problem (2.1) is the $\ell_p$-quasi-norm minimization (see e.g., [10, 15, 24]). By setting $\phi_i(t) = \log(t)$, the function $F_\varepsilon(x) = \sum_{x_i \neq 0} \log(|x_i| + \varepsilon)$ was used by Gorodnitsky and Rao [25] to design the FOCUSS algorithm for sparse signal reconstruction.

We use $e = (1, 1, ..., 1)^T \in \mathbb{R}^n$ throughout to denote the vector of ones. Recall that a function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be coercive in the region $D \subseteq \mathbb{R}^n$ if $f(x) \to \infty$
as \( \|x\| \to \infty \) and \( x \in D \). We now specify a class of merit functions satisfying the following assumption.

**Assumption 2.1.** For any given \( \varepsilon > 0 \), the merit function \( F_\varepsilon : \mathbb{R}^n \to \mathbb{R} \) satisfies the following properties:

(a) \( F_\varepsilon(x) = F_\varepsilon(|x|) \) for any \( x \in \mathbb{R}^n \), and \( F_\varepsilon(x) \) is separable in \( x \), and twice continuously differentiable with respect to \( x \in \mathbb{R}^n \).

(b) In \( \mathbb{R}^n_+ \), \( F_\varepsilon(x) \) is strictly increasing with respect to every component \( x_i \) and \( \varepsilon \), and for any given \( \gamma > 0 \) there exists a finite number \( Q(\gamma) \) such that for any \( S \subseteq \{1, \ldots, n\} \), \( g(x_S) := \inf_{x_S \in \mathbb{R}^n} F_\varepsilon(x_S) \geq Q(\gamma) \) provided \( x_S \geq \gamma e_S \), and \( g(x_S) \) is coercive in the set \( \{x_S : x_S \geq \gamma e_S\} \).

(c) In \( \mathbb{R}^n_+ \), the gradient satisfies that \( \nabla F_\varepsilon(x) \in \mathbb{R}^n_+ \) for any \( (x, \varepsilon) \in \mathbb{R}^n_+ \times \mathbb{R}^n_+ \), \( [\nabla F_\varepsilon(x)]_i \to \infty \) as \( (x_i, \varepsilon) \to 0 \), and for any given \( x_i > 0 \) the component \( \nabla F_\varepsilon(x)_i \) is continuous in \( \varepsilon \) and tends to a finite positive number (dependent only on \( x_i \)) as \( \varepsilon \to 0 \).

(d) In \( \mathbb{R}^n_+ \), the Hessian satisfies that \( y^T \nabla^2 F_\varepsilon(x) y \leq -C(\varepsilon, r) \|y\|^2 \) for any \( y \in \mathbb{R}^n \) and \( x \in \mathbb{R}^n_+ \) with \( \|x\| \leq r \), where \( r > 0 \) and \( C(\varepsilon, r) > 0 \) are constants, and \( C(\varepsilon, r) \) is continuous in \( \varepsilon \) and bounded away from zero as \( \varepsilon \to 0 \). (Thus \( F_\varepsilon(x) \) is strictly concave with respect to \( x \).)

There exists a large family of functions satisfying the above assumption, and we show that it is very easy to construct examples in this family. Note that any nonnegative combination of a finite number of functions satisfying Assumption 2.1 still satisfies this assumption. So the set of such functions is a convex cone, denoted by \( \mathcal{M} \), i.e.,

\[
\mathcal{M} = \{ F_\varepsilon : F_\varepsilon \text{ satisfies Assumption 2.1 for } \varepsilon > 0 \}.
\]

Based on the merit functions in \( \mathcal{M} \), we can derive a family of reweighted \( \ell_1 \)-minimization algorithms. Note that for any \( F_\varepsilon \in \mathcal{M} \), problem (2.1) can be rewritten as

\[
\min \{ F_\varepsilon(v) : Ax = b, |x| \leq v \} = \min_{(x,v) \in \mathcal{F}} F_\varepsilon(v), \tag{2.2}
\]

where \( |x| = (|x_1|, \ldots, |x_n|)^T \) and

\[
\mathcal{F} = \{(x,v) : Ax = b, |x| \leq v \}.
\]

Throughout this paper, we assume that the system \( Ax = b \) has a solution, i.e., \( \mathcal{F} \neq \emptyset \). This can be always guaranteed when the system is underdetermined.

At the current point \( v^k \), since \( F_\varepsilon(v) = F_\varepsilon(v^k) + \nabla F_\varepsilon(v^k)^T(v - v^k) + o(\|v - v^k\|) \), the simplest tractable approximation to the concave minimization (2.2) is the problem of minimizing the linear approximation of \( F_\varepsilon(v) \) over the same feasible set. Thus it makes sense to solve the following problem to generate the next point \( (x^{k+1}, v^{k+1}) : \)

\[
(x^{k+1}, v^{k+1}) = \arg \min_{(x,v) \in \mathcal{F}} \{ F_\varepsilon(v^k) + \nabla F_\varepsilon(v^k)^T(v - v^k) \}
\]

\[
= \arg \min_{(x,v) \in \mathcal{F}} \nabla F_\varepsilon(v^k)^T v, \tag{2.3}
\]

which is an LP. Note that \( v^k \in \mathbb{R}^n_+ \) and \( \nabla F_\varepsilon(v^k) \in \mathbb{R}^n_+ \) (by Assumption 2.1(c)). It is easy to see that the optimal solution \( (x^{k+1}, v^{k+1}) \) of (2.3) always satisfies

\[
v^{k+1} = |x^{k+1}| \text{ for all } k \geq 0. \tag{2.4}
\]
Hence it follows from (2.4) and the positiveness of $\nabla F_\varepsilon(v^k)$ that
\[
\nabla F_\varepsilon(v^k)^T v^{k+1} = \nabla F_\varepsilon(|x|^k)T |x|^{k+1} = \left\| \text{diag} \left( \nabla F_\varepsilon(|x|^k) \right) \right\|_1.
\]

Therefore, the iterative scheme (2.3) is nothing but
\[
x^{k+1} = \arg \min \left\{ \| \text{diag} \left( \nabla F_\varepsilon(|x|^k) \right) \|_1 : Ax = b \right\}, \tag{2.5}
\]

which is the reweighted $\ell_1$-minimization with weight $w^k = \nabla F_\varepsilon(|x|^k) \in R_{++}^n$. By Assumption 2.1(c), we have that $[\nabla F_\varepsilon(v^k)]_i \to \infty$ as $(v_i, \varepsilon) \to 0$, so for small $v^k_i = |x^k_i|$ and $\varepsilon > 0$, the corresponding weight $[\nabla F_\varepsilon(v^k)]_i$ is large. The iterative scheme (2.3) (equivalently, (2.5)) provides a unified approach to derive reweighted $\ell_1$-algorithms for sparse solutions of linear systems, which can be described as follows.

**Algorithm 2.2 (Reweighted $\ell_1$-Minimization)**

S1. Choose $\alpha, \varepsilon_0 \in (0, 1)$, and let $(x^0, v^0) \in R^n \times R^n_+$ be an initial point.

S2. At the current iterate $(x^k, v^k)$ with $\varepsilon_k > 0$, compute
\[
(x^{k+1}, v^{k+1}) = \arg \min_{(x, v) \in F} \left( \nabla F_{\varepsilon_k}(v^k) \right)^T v
\]

(i.e., $x^{k+1} = \arg \min \{|\text{diag}(\nabla F_{\varepsilon_k}(|x|^k)|x|)_1 : Ax = b\}$).

S3. Set $\varepsilon_{k+1} = \alpha \varepsilon_k$. Replace $(x^k, v^k, \varepsilon_k)$ by $(x^{k+1}, v^{k+1}, \varepsilon_{k+1})$ and repeat S2.

Note that $\varepsilon$ is reduced by a factor after every iteration, and several stopping criteria can be used in Algorithm 2.2, for instance, $\varepsilon_k < \varepsilon^*$ or $v^{k+1} - v^k < \varepsilon^*$ ($\varepsilon^*$ is a prescribed tolerance). We note that the general concave optimization theory is a long-lasting research topic in the field of global optimization (see for instance [30]). From a concave programming point of view, Algorithm 2.2 is not new, which is essentially a linearization method for concave minimization. However, it is the objective function (i.e., the merit function for sparsity) that makes the algorithm unique.

Algorithm 2.2 is a general framework for reweighted $\ell_1$-algorithms, from which many specific algorithms, including some existing ones, can be immediately obtained by choosing various merit functions. In this paper, we focus on the merit functions in $\mathcal{M}$. A few examples of such functions together with their gradients are given as follows.

**Example 2.3.** (i) Note that \((n + \sum_{i=1}^n \log(|x_i| + \varepsilon)) / \log \varepsilon \to \|x\|_0 \) as $\varepsilon \to 0$. This motivates the merit function
\[
F_\varepsilon(x) = \sum_{i=1}^n \log(|x_i| + \varepsilon)
\]
in $\mathcal{M}$. The gradient of this function at $x \in R^n_+$ is given by $\nabla F_\varepsilon(x) = \left( \frac{1}{x_1 + \varepsilon}, \ldots, \frac{1}{x_n + \varepsilon} \right)^T \in R^n_+$, and the Hessian $\nabla^2 F_\varepsilon(x) = -\text{diag} \left( \frac{1}{(x_1 + \varepsilon)^2}, \ldots, \frac{1}{(x_n + \varepsilon)^2} \right)$. Using this merit function, (2.3) is exactly the CWB reweighted $\ell_1$-method with weights given by (1.1), which has been studied by several authors [21, 22, 33, 9, 36]. A convergence result for this method can be obtained as a special case of our general framework.

(ii) Let $p \in (0, 1)$. It is easy to verify that the function
\[
F_\varepsilon(x) = \sum_{i=1}^n \log(|x_i| + \varepsilon + (|x_i| + \varepsilon)^p)
\]
REWEIGHTED $\ell_1$-MINIMIZATION

is also in $\mathcal{M}$. At $x \in \mathbb{R}_+^n$, the gradient is given by $[\nabla F_\infty(x)]_i = \frac{p+(x_i+\varepsilon)^{1-p}}{(x_i+\varepsilon)^{1-p}[x_i+\varepsilon+(x_i+\varepsilon)p]}$, $i = 1, \ldots, n$, and the Hessian $\nabla^2 F_\infty(x)$ is a diagonal matrix with diagonal entries

$$
-\frac{1}{[x_i+\varepsilon+(x_i+\varepsilon)p]^2} \left(1 + \frac{p(3-p)}{(x_i+\varepsilon)^{1-p}} + \frac{p}{(x_i+\varepsilon)^{2(1-p)}}\right), \; i = 1, \ldots, n.
$$

This merit function yields a reweighted $\ell_1$-algorithm with the weights

$$
w_i = [\nabla F_\infty(|x|)]_i = \frac{p+(|x_i|+\varepsilon)^{1-p}}{(|x_i|+\varepsilon)^{1-p}[|x_i|+\varepsilon+(|x_i|+\varepsilon)p]} , \; i = 1, \ldots, n,
$$

(termed the ‘NW1’ algorithm) which has not been studied in the literature so far.

**Example 2.4.** Note that $\|x\|_p^p \to \|x\|_0$ as $p \to 0$. For a given $p \in (0,1)$, we define $F_\varepsilon(x) = \frac{1}{p} \sum_{i=1}^n (|x_i|+\varepsilon)^p$, which is in $\mathcal{M}$. At $x \in \mathbb{R}_+^n$, the gradient of this function is given by

$$
\nabla F_\varepsilon(x) = \begin{pmatrix} \frac{1}{(x_1+\varepsilon)^{1-p}}, \ldots, \frac{1}{(x_n+\varepsilon)^{1-p}} \end{pmatrix}^T \in \mathbb{R}_+^n,
$$

and the Hessian $\nabla^2 F_\varepsilon(x) = -\text{diag} \left(\frac{1}{(x_1+\varepsilon)^{1-p}}, \ldots, \frac{1}{(x_n+\varepsilon)^{1-p}}\right)$. By this merit function, (2.3) is exactly the reweighted $\ell_1$-method with weights $w_i = \frac{1}{(|x_i|+\varepsilon)^p}$, $i = 1, \ldots, n$, (termed the ‘W1p’ algorithm) which was recently studied in [24, 31, 12]. Some new properties of this method can be extracted from our general results in this paper.

**Example 2.5.** Let $p$ and $q \in (0,1)$ be given. Define

$$
F_\varepsilon(x) = \frac{1}{p} \sum_{i=1}^n (|x_i|+\varepsilon+(|x_i|+\varepsilon)^q)^p,
$$

which is in $\mathcal{M}$. The gradient of this function at $x \in \mathbb{R}_+^n$ is given by

$$
[\nabla F_\varepsilon(x)]_i = \frac{q+(x_i+\varepsilon)^{1-q}}{(x_i+\varepsilon)^{1-q}[x_i+\varepsilon+(x_i+\varepsilon)^q]^{1-p}}, \; i = 1, \ldots, n,
$$

and the Hessian of this function is a diagonal matrix with diagonal entries

$$
-\frac{1}{[x_i+\varepsilon+(x_i+\varepsilon)^q]^{2(1-p)}} \left((1-p) + \frac{2q(1-p) + q(1-q)}{(x_i+\varepsilon)^{1-q}} + \frac{q^2(1-p) + q(1-q)}{(x_i+\varepsilon)^{2(1-q)}}\right),
$$

Based on this merit function, the associated reweighted $\ell_1$-minimization (termed the ‘NW2’ algorithm) remains new. It is interesting to note that the parameter $q$ above can be chosen to be greater than 1 if we restrict the range of $p$. For instance, we can let $q = 2$ and $p \leq \frac{1}{2}$. Then the function

$$
F_\varepsilon(x) = \frac{1}{p} \sum_{i=1}^n (|x_i|+\varepsilon+(|x_i|+\varepsilon)^2)^p
$$

is in $\mathcal{M}$. The associated reweighted $\ell_1$-minimization (termed the ‘NW3’ algorithm) with

$$
w_i = [\nabla F_\varepsilon(|x|)]_i = \frac{1 + 2(|x_i|+\varepsilon)}{(|x_i|+\varepsilon+(|x_i|+\varepsilon)^2)^{1-p}}, \; i = 1, \ldots, n$$
is also a new algorithm. Numerical experiments show that NW2 and NW3 are quite strong in finding the sparsest solution of linear systems in many situations (see section 4 for details).

**Example 2.6.** Let \( p \in (0, \infty) \) be fixed. Note that \( \varphi_\varepsilon(x) = \sum_{i=1}^{n} \frac{(|x_i| + \varepsilon)^p - \varepsilon^p}{(|x_i| + \varepsilon)^p} \rightarrow \|x\|_0 \) as \( \varepsilon \rightarrow 0 \) (in particular, for \( p = 1 \), \( \varphi_\varepsilon(x) = \sum_{i=1}^{n} \frac{|x_i|}{|x_i| + \varepsilon} \rightarrow \|x\|_0 \) as \( \varepsilon \rightarrow 0 \)). This motivates the merit function \( g_\varepsilon(x) = -\sum_{i=1}^{n} \frac{1}{(|x_i| + \varepsilon)^p} \) which satisfies almost all conditions of Assumption 2.1 except for the coercivity. However, we can make this function coercive as follows: Adding a log function to \( g_\varepsilon(x) \) yields the merit function (in \( M \))

\[
F_\varepsilon(x) = \sum_{i=1}^{n} \left[ \log(|x_i| + \varepsilon) - \frac{1}{(|x_i| + \varepsilon)^p} \right],
\]

(2.6)

for which the gradient at \( x \in \mathbb{R}^n_+ \) is given by \( \nabla F_\varepsilon(x) = \left( \frac{1+((x_1+\varepsilon)^p)}{(x_1+\varepsilon)^{p+1}}, ..., \frac{1+((x_n+\varepsilon)^p)}{(x_n+\varepsilon)^{p+1}} \right) \), and the Hessian \( \nabla^2 F_\varepsilon(v) = -\text{diag} \left( \frac{p+1+(v_1+\varepsilon)^p}{(v_1+\varepsilon)^{p+2}}, ..., \frac{p+1+(v_n+\varepsilon)^p}{(v_n+\varepsilon)^{p+2}} \right) \). The associated reweighted \( \ell_1 \)-minimization (termed the ‘NW4’ algorithm) uses the weights \( w_i = \frac{1+((x_i+\varepsilon)^p)}{(x_i+\varepsilon)^{p+1}}, \) \( i = 1, ..., n \). Another way to make \( g_\varepsilon \) coercive is to add the \( \ell_1 \)-norm, leading to the merit function

\[
F_\varepsilon(x) = \sum_{i=1}^{n} \left( |x_i| - \frac{1}{(|x_i| + \varepsilon)^p} \right),
\]

which is in \( M \) and yields a new reweighted \( \ell_1 \)-method with weights \( w_i = [\nabla F_\varepsilon(|x|)]_i = \frac{1+((x_i+\varepsilon)^p)}{(x_i+\varepsilon)^{p+1}}, \) \( i = 1, ..., n \).

Note that when \( p \rightarrow 0 \), the function in Example 2.4 reduces to that of Example 2.3 (i), and the function in Example 2.5 reduces to that of Example 2.3(ii). Thus the CWB method can be viewed as an extreme case of the \( W_l_p \) method as \( p \rightarrow 0 \), and it can be also viewed as an extreme case of (2.6) as \( p \rightarrow 0 \).

**Remark.** Examples 2.3-2.6 show that it is not difficult to construct merit functions for sparsity in \( M \), and the well-known CWB and \( W_{l_p} \) methods fall into this family. To construct a function in \( M \), we may start with a choice of the kernel function \( \phi_i(t) \), which should be twice continuously differentiable, strictly concave and increasing in \( \mathbb{R}^n_+ \). Every merit function in \( M \) yields a reweighted \( \ell_1 \)-minimization algorithm for the sparse solution of linear systems. The family of such algorithms is large. Because \( M \) is a convex cone, any positive combination of functions in \( M \) is in this family. Thus, the combination can be used to generate new merit functions from known ones. For instance, a simple combination of Example 2.3(i) and Example 2.4 yields the function

\[
F_\varepsilon(v) = \left( \sum_{i=1}^{n} \log(v_i + \varepsilon) \right) + \sum_{i=1}^{n} (v_i + \varepsilon)^p
\]

in \( M \), where \( p \in (0, 1) \). Its kernel function is \( \phi_i(t) = \log(t) + t^p \) \( (i = 1, ..., n) \). In many cases, applying the log operation to a nonnegative convex function a finite number of times, we may reverse the convexity to concavity (see Zhao et al [45]). For instance, if \( \phi : R_+ \rightarrow R_+ \) is a twice differentiable, strictly increasing and convex function obeying \( \phi''(t)(1+\phi(t)) < (\phi'(t))^2 \), then \( \log(1+\phi(t)) \) is concave, so is \( \log(1+\log(1+\phi(t))) \). Since the log operation maintains the coercivity and monotonicity of the original function, this strategy can be used to construct a concave merit function for sparsity from a convex function.
3. Convergence analysis of reweighted $\ell_1$-minimization. From the above section, we see that there exist infinitely many reweighted algorithms. Therefore it is necessary to study these algorithms in a unified approach, in order to identify their common properties. The remainder of this paper is devoted to this task. We carry out a unified convergence analysis for Algorithm 2.2 based on the merit functions for sparsity in $M$. We note that Algorithm 2.2 can start at any initial point $(x^0,v^0)$ with $v^0 = |x^0|$ where $x^0$ is not necessarily a solution to $Ax = b$. After the first step, the algorithm will generate an iterate $(x^1,v^1)$ satisfying $v^1 = |x^1|$ and $Ax^1 = b$, from which all subsequent iterates satisfy $Ax = b$ and (2.4), i.e., $v^k = |x^k|$ for all $k \geq 1$.

The first result below shows that if the algorithm generates at the $k$th step a new point $v^{k+1} \neq v^k$, then the value of the merit function for sparsity strictly decreases.

**Lemma 3.1.** Suppose that $F_\varepsilon$ satisfies Assumption 2.1, i.e., $F_\varepsilon \in M$. For a given $(\bar{x},\bar{v}) \in F$ with $\bar{v} = |\bar{x}|$ and a parameter $\varepsilon > 0$, let

$$ (x^+,v^+) = \arg \min_{(x,v) \in F} \nabla F_\varepsilon(v)^T v. \quad (3.1) $$

If $v^+ \neq \bar{v}$, then $F_\varepsilon(v^+) < F_\varepsilon(\bar{v})$.

**Proof.** Note that $(\bar{x},\bar{v})$ is feasible to the problem (3.1). Thus the minimizer $v^+$ satisfies that $\nabla F_\varepsilon(v^+)^T v^+ \leq \nabla F_\varepsilon(\bar{v})^T \bar{v}$. When $v^+ \neq \bar{v}$, we have only two cases.

Case 1. $\nabla F_\varepsilon(v^+)^T (v^+ - \bar{v}) < 0$. In this case, by concavity of $F_\varepsilon$, we have $F_\varepsilon(v^+) \leq F_\varepsilon(\bar{v}) + \nabla F_\varepsilon(v^+)^T (v^+ - \bar{v}) < F_\varepsilon(\bar{v})$.

Case 2. $\nabla F_\varepsilon(v^+)^T (v^+ - \bar{v}) = 0$. Let $r > 0$ be a constant such that $\|\bar{v}\| \leq r$. Then for any sufficiently small $t > 0$, we have

$$ F_\varepsilon(\tilde{v} + t(v^+ - \bar{v})) = F_\varepsilon(\bar{v}) + t\nabla F_\varepsilon(\bar{v})^T (v^+ - \bar{v}) + \frac{1}{2} t^2 (v^+ - \bar{v})^T \nabla^2 F_\varepsilon(\bar{v})(v^+ - \bar{v}) + o(t^2) $$

$$ = F_\varepsilon(\bar{v}) + \frac{1}{2} t^2 (v^+ - \bar{v})^T \nabla^2 F_\varepsilon(\bar{v})(v^+ - \bar{v}) + o(t^2) $$

$$ \leq F_\varepsilon(\bar{v}) - \frac{t^2}{2} C(\varepsilon,r)\|v^+ - \bar{v}\|^2 + o(t^2) $$

$$ < F_\varepsilon(\bar{v}). \quad (3.2) $$

The first inequality above follows from the strict concavity of $F_\varepsilon$ (Assumption 2.1(d)). The concavity of $F_\varepsilon$ also implies that, for any sufficiently small $t > 0$, we have

$$ F_\varepsilon(\tilde{v} + t(v^+ - \bar{v})) \geq t F_\varepsilon(v^+) + (1 - t) F_\varepsilon(\bar{v}) = t (F_\varepsilon(v^+) - F_\varepsilon(\bar{v})) + F_\varepsilon(\bar{v}), $$

which, together with (3.2), implies that $F_\varepsilon(v^+) - F_\varepsilon(\bar{v}) < 0$. $lacksquare$

By the structure of Algorithm 2.2, we have the following corollary, showing that the merit function strictly decreases after every iteration.

**Corollary 3.2.** Let $F_\varepsilon \in M$ and $\{(x^k,v^k)\}$ be generated by Algorithm 2.2. Then the sequence $\{F_{\varepsilon_k}(v^k)\}$ is strictly decreasing in the sense that $F_{\varepsilon_{k+1}}(v^{k+1}) < F_{\varepsilon_k}(v^k)$ for all $k \geq 1$.

**Proof.** First, we note that $F_{\varepsilon_k}(v^{k+1}) \leq F_{\varepsilon_k}(v^k)$. In fact, it holds trivially if $v^{k+1} = v^k$; otherwise, it holds strictly by Lemma 3.1. Since $F_\varepsilon$ is strictly increasing in $\varepsilon$ (by Assumption 2.1(b)), we have $F_{\varepsilon_{k+1}}(v^{k+1}) < F_{\varepsilon_k}(v^{k+1}) \leq F_{\varepsilon_k}(v^k)$. $lacksquare$

**Lemma 3.3.** Let $F_\varepsilon \in M$ and $\{(x^k,v^k)\}$ be the sequence generated by Algorithm 2.2. If there is a subsequence, denoted by $\{(x^{j_k},v^{j_k}) : j = 1,2,\ldots\}$, such that $|x^{j_k}| \geq \gamma \varepsilon$ for all $j$ where $\gamma > 0$ is a constant, then there exists a finite constant $C^* > 0$
such that $\sum_{j=1}^{\infty} \| v^{k_j+1} - v^{k_j} \|^2 \leq C^*$. In particular, we have $\| v^{k_j+1} - v^{k_j} \| \to 0$ as $j \to \infty$.

Proof. Note that $v^k = |x^k|$ for all $k \geq 1$. By Assumption 2.1(b), it follows from $v^{k_j} = |x^{k_j}| \geq \gamma e$ that

$$F_{\varepsilon} (v^k) \geq F_{\varepsilon} (\gamma e) \geq \inf_{\varepsilon > 0} F_\varepsilon (\gamma e) \geq Q(\gamma),$$

and hence the sequence $\{ F_{\varepsilon} (v^k) \}$ is bounded below. By Corollary 3.2, the sequence $\{ F_\varepsilon (v^k) \}$ is strictly decreasing, and thus $F_{\varepsilon} (v^k) \leq F_{\varepsilon_0} (v^0)$ for any $j$. This, together with the coercivity of $g = \inf_{\varepsilon > 0} F_\varepsilon$ (Assumption 2.1 (b)), implies that the sequence $\{ v^k \}$ must be bounded. In fact, if $\{ v^k \}$ is unbounded, then there exists a subsequence, denoted still by $\{ v^k \}$, such that $\| v^k \| \to \infty$. Since $v^k \geq \gamma e$, it follows from Assumption 2.1 (b) that $F_{\varepsilon} (v^k) \geq \inf_{\varepsilon > 0} F_\varepsilon (v^k) = g(v^k) \to \infty$ as $\| v^k \| \to \infty$, which contradicts the inequality $F_{\varepsilon} (v^k) \leq F_{\varepsilon_0} (v^0)$ for all $j$. Thus, the sequence $\{ v^k \}$ must be bounded. So there is a positive constant $\gamma' > \gamma$ such that $\gamma e \leq v^k \leq \gamma' e$ for all $j$. For any $\varepsilon > 0$, since $F_\varepsilon (v)$ is separable in $v \in R^n_+$ and $[\nabla^2 F_\varepsilon (v)]_{ii}$ is negative (Assumption 2.1 (d)), it implies that for every $i$, $[\nabla F_\varepsilon (v)]_i$ is decreasing with respect to $v_i$. Therefore,

$$\nabla F_{\varepsilon} (\gamma' e) \leq \nabla F_{\varepsilon} (v^k) \leq \nabla F_{\varepsilon} (\gamma e). \quad (3.3)$$

Assumption 2.1 (c) implies that for every $i$, the components $[\nabla F_\varepsilon (\gamma' e)]_i$ and $[\nabla F_\varepsilon (\gamma e)]_i$ are both positive, and bounded from above and away from zero over the region $\varepsilon \in (0, \varepsilon_0]$. This, together with (3.3), implies that there exist two constants $\beta_1$ and $\beta_2$ ($0 < \beta_1 < \beta_2$) such that

$$\beta_1 e \leq \nabla F_{\varepsilon} (v^k) \leq \beta_2 e \quad \text{for all } j. \quad (3.4)$$

By optimality, we have $\nabla F_{\varepsilon} (v^k)^T v^{k_j+1} \leq \nabla F_{\varepsilon} (v^k)^T v^k$ for any $j$. This, together with (3.4) and $\gamma e \leq v^k \leq \gamma' e$, implies that the sequence $\{ v^{k_j+1} \}$ is bounded.

Since $\{ F_{\varepsilon} (v^k) \}$ is decreasing and bounded below, we have $F_{\varepsilon} (v^k) \to F^*$ as $j \to \infty$, where $F^*$ is a constant. Let $r > 0$ be a constant such that

$$\max \{ \max_{j \geq 1} \| v^k \|, \max_{j \geq 1} \| v^{k_j+1} \| \} \leq r,$$

which is finite since $\{ v^k \}$ and $\{ v^{k_j+1} \}$ are bounded. Note that $\varepsilon(k_{j+1}) \leq \cdots \leq \varepsilon(k_j) < \varepsilon_j$. By the decreasing property of the sequence $\{ F_{\varepsilon} (v^k) \}$ and Assumption 2.1 (b) and (d), we have

$$F_{\varepsilon(j+1)} (v^{k_j+1}) \leq F_{\varepsilon(j+1)} (v^{k_j+1}) < F_{\varepsilon} (v^{k_j+1})$$

$$= F_{\varepsilon} (v^k) + \nabla F_{\varepsilon} (v^k)^T (v^{k_j+1} - v^k) + \frac{1}{2} (v^{k_j+1} - v^k)^T \nabla^2 F_{\varepsilon} (\hat{\varepsilon})(v^{k_j+1} - v^k)$$

$$\leq F_{\varepsilon} (v^k) + \nabla F_{\varepsilon} (v^k)^T (v^{k_j+1} - v^k) - \frac{1}{2} C(\varepsilon_j, r) \| v^{k_j+1} - v^k \|^2,$$

where $\hat{\varepsilon}$ is some point on the line segment between $v^{k_j+1}$ and $v^k$, and hence $\| \hat{\varepsilon} \| \leq r$. The last inequality above follows from Assumption 2.1 (d). Since $\nabla F_{\varepsilon(j+1)} (v^{k_j+1}) = \nabla F_{\varepsilon(j+1)} (v^{k_j+1}) \leq 0$, it follows from the above inequality that

$$\frac{1}{2} C(\varepsilon_j, r) \| v^{k_j+1} - v^k \|^2 \leq F_{\varepsilon} (v^k) - F_{\varepsilon(j+1)} (v^{k_j+1}).$$
By Assumption 2.1 (d), $C(\varepsilon_{k,j}, r) > 0$ is bounded away from zero as $j \to \infty$. Therefore there exists a constant $\gamma^* > 0$ such that $C(\varepsilon_{k,j}, r) \geq \gamma^* > 0$ for all $j$. Thus the above inequality implies that
\[
\lim_{j \to \infty} \|v^{k+1} - v^k\| = 0, \quad \text{and} \quad \sum_{j=0}^{\infty} \|v^{k+1} - v^k\|_2^2 \leq \frac{F_{0\infty}(v^0) - F^*}{\gamma^*} =: C^*.
\]

The proof is complete. □

From Examples 2.3 and 2.6, we see that a merit function in $\mathcal{M}$ is not necessarily bounded below as $|x_i| \to 0$. However, when a merit function is bounded below (such as Examples 2.4 and 2.5), we have the next result which claims that the result of Lemma 3.3 holds for the whole sequence generated by the algorithm, instead of a subsequence.

**Corollary 3.4.** Let $F_\varepsilon \in \mathcal{M}$ and $F_\varepsilon(x)$ be bounded below in the region $(x, \varepsilon) \in R_+^n \times R_+$, and let $g(x) = \inf_{x \in \varepsilon} F_\varepsilon(x)$ be coercive in $R_+^n$. Let $\{(x^k, v^k)\}$ be generated by Algorithm 2.2. Then $\{(x^k, v^k)\}$ is bounded and there exists a finite constant $C^*$ such that $\sum_{k=1}^{\infty} \|v^{k+1} - v^k\| \leq C^*$. In particular, $\|v^{k+1} - v^k\| \to 0$ as $k \to \infty$.

**Proof.** By the assumption and Corollary 3.2, there exists a constant $F^*$ such that $F^* \leq F_\varepsilon(v^k) \leq F_{0\infty}(v^0)$ for all $k$ which, together with the coercivity of $g(x)$, implies that $\{v^k\}$ is bounded. Applying the rest proof of Lemma 3.3 to $\{v^k\}$ instead of $\{v^{k+1}\}$, we can show that $\|v^{k+1} - v^k\| \to 0$ and $\sum_{k=1}^{\infty} \|v^{k+1} - v^k\| \leq C^*$ for some $C^*$. □

To further investigate the properties of Algorithm 2.2, we need some condition on $A$. We introduce the range space property of $A^T$, based on which we can establish some convergence results for Algorithm 2.2.

**Definition 3.5.** (Range Space Property (RSP)) Let $A$ be an $m \times n$ matrix with $m \leq n$. Then $A^T$ is said to satisfy the range space property of order $K$ with a constant $\rho > 0$ if
\[
\|\xi_S\|_1 \leq \rho\|\xi_S\|_1
\]
for all sets $S \subseteq \{1, \ldots, n\}$ with $|S| \geq K$, and for all $\xi \in \mathcal{R}(A^T)$, the range space of $A^T$.

Clearly, RSP can be equivalently stated in some other ways. First, note that the inequality $|S| \geq K$ in the above definition can be replaced by the equality $|S| = K$. In fact, it is easy to see that if $\|\xi_S\|_1 \leq \rho\|\xi_S\|_1$ holds for any $S \subseteq \{1, \ldots, n\}$ with $|S| = K$, then it holds for all $S$ with $|S| > K$. Second, it is evident that the RSP can be restated as follows: $A^T$ is said to satisfy the range space property of order $K$ with constant $\rho > 0$ if $\|\xi_S\|_1 \leq \rho\|\xi_S\|_1$ for any set $S \subseteq \{1, \ldots, n\}$ with $|S| \leq n - K$, and for all $\xi \in \mathcal{R}(A^T)$. It is interesting to understand the relationship between this property and the restricted isometry property (RIP) and null space property (NSP) of $A$ which have been widely used in the compressed sensing literature. Recall that $A$ has the RIP of order $k$ if there exists a constant $\delta \in (0, 1)$ such that $(1 - \delta)\|z\|_2^2 \leq \|Az\|_2^2 \leq (1 + \delta)\|z\|_2^2$ for any $k$-sparse vector $z \in R^n$, and that $A$ has the NSP of order $k$ if there exists a constant $\tau \in (0, 1)$ such that $\|\eta_S\|_1 \leq \tau\|\eta_S\|_1$ for all $S \subseteq \{1, \ldots, n\}$ with $|S| \leq k$, and any $\eta \in \mathcal{N}(A)$, the null space of $A$. In this paper, we do not make use of RIP and NSP in our analysis. The following proposition and remark shed some light on the relationship between RIP, NSP, and RSP.

**Proposition 3.6.** Let $m < n$, and let $A \in R^{m \times n}$ and $M \in R^{(n-m) \times n}$ be full-rank matrices satisfying $AM^T = 0$. Then the following hold.

(i) $M$ has the NSP of order $k$ with constant $\tau \in (0, 1)$ if and only if $A^T$ has the RSP of order $(n - k)$ with the same constant $\rho = \tau$. 

(ii) Let $\eta \in \mathcal{N}(A)$. Then $\|A\eta\|_2 \leq \|M\eta\|_2$.
(ii) If $M$ has the RIP of order $k$ with constant $\delta \in (0,1)$, then $A^T$ has the RSP of order $\left(n - \left\lceil \frac{\delta k}{1 + \rho} \right\rceil \right)$ with the constant $\rho = \left( \frac{1}{k - \frac{\delta k}{1 + \rho}} \right)^{1/2} \left( \frac{1}{\rho} \right) < 1$ where $\varrho = (1 - \delta)/(1 + \delta)$.

Proof. (i) Since $A$ and $M$ are full-rank matrices and $AM^T = 0$ (and thus $MA^T = 0$), the columns of $A^T$ comprise the basis of the null space of $M$. Thus $\mathcal{R}(A^T) = \mathcal{N}(M)$. Suppose that $M$ has the NSP of order $k$ with constant $\tau \in (0,1)$. Then $\|\eta_{L}\|_1 \leq \tau \|\eta_{Lc}\|_1$ for any $\eta \in \mathcal{N}(M)$ and for any set $L \subseteq \{1, \ldots, n\}$ with $|L| \leq k$. By setting $S = L_c$, this is equivalent to saying that $\|\eta_{S}\|_1 \leq \tau \|\eta_{S}\|_1$ for any $\eta \in \mathcal{R}(A^T)$ and for any set $S \subseteq \{1, \ldots, n\}$ with $|S| \geq n - k$. Thus $A$ has the RIP of order $(n - k)$ with constant $\rho = \tau$. Clearly, the converse can be shown in a similar way.

(ii) Suppose that $M$ has the RIP of order $k$ with constant $\delta \in (0,1)$. Denote by $\varrho = (1 - \delta)/(1 + \delta)$. Let $\nu = \left[ \frac{\delta k}{1 + \varrho} \right]$, which is smaller than $k$. It was shown in [13] that $M$ has the RIP of order $\nu$ with constant $\tau = \sqrt{\frac{\nu}{k - \frac{\delta k}{1 + \varrho}}} \left( \frac{1}{\varrho} \right)$, which is less than 1 by the definition of $\nu$. By (i), this in turn implies that $A^T$ has the RSP of order $n - \nu = n - \left\lceil \frac{\delta k}{1 + \varrho} \right\rceil$ with constant $\rho = \tau$. \qed

From Proposition 3.6, if $M \in \mathcal{R}^{(n-m) \times n}$ has the NSP (or more restrictively, the RIP), we may construct a matrix satisfying the range space property by simply choosing $A = [v_1, \ldots, v_m]^T$ where $v_1, \ldots, v_m \in \mathbb{R}^{n}$ are the basis of the null space of $M$.

Remark. One might be also interested in the direct relationship between the RIP of $A$ and RSP of $A^T$, which remains not quite clear at this stage. However, such a direct relationship might exist as shown by the following observation. Note that for any $\xi \in \mathcal{R}(A^T)$ and $\eta \in \mathcal{N}(A)$, $\xi^T \eta = 0$ implies that for any index set $S \subset \{1, \ldots, n\}$ we have $(\xi_S)^T \eta_S + (\xi_{S_c})^T \eta_{S_c} = 0$, and hence

$$\|\xi_S\|_2 \|\eta_S\|_2 \cos(\theta) = \|\xi_{S_c}\|_2 \|\eta_{S_c}\|_2 \cos(\theta'),$$

(3.5)

where $\theta$ is the angle between $\xi_S$ and $\eta_S$, and $\theta'$ between $\xi_{S_c}$ and $\eta_{S_c}$. Assume that there exists a $k$ such that for any $S \subset \{1, \ldots, n\}$ with $|S| \leq k$, it holds that $|\cos(\theta')| \geq \gamma > 0$ (where $\gamma$ is a constant) for any $\eta \in \mathcal{N}(A)$ and $\xi \in \mathcal{R}(A^T)$ with $\eta_{S_c} \neq 0$ and $\xi_{S_c} \neq 0$. Under this assumption, we see that when $\eta_{S_c} \neq 0$ and $\xi_{S_c} \neq 0$, all other terms in (3.5) are nonzero. In this case, (3.5) can be written as

$$\|\xi_S\|_2 \|\xi_{S_c}\|_2 = \left( \frac{\|\eta_{S_c}\|_2}{\|\eta_S\|_2} \right) (|\cos(\theta')|/|\cos(\theta)|).$$

By the equivalence of $\| \cdot \|_1$ and $\| \cdot \|_2$, it is not difficult to see that there exists a constant $\vartheta > 0$ (dependent on $n$ and $k$) such that

$$\|\xi_S\|_1 \|\xi_{S_c}\|_1 \geq \vartheta (\|\eta_{S_c}\|_1/\|\eta_S\|_1) (|\cos(\theta')|/|\cos(\theta)|).$$

Now, if $A$ has the NSP of order $k$, then there exists a constant $\tau \in (0,1)$ such that $\|\eta_{S_c}\|_1 / \|\eta_S\|_1 \geq 1/\tau$, and thus $\|\xi_S\|_1 / \|\xi_{S_c}\|_1 \geq (\vartheta/\tau) (|\cos(\theta')|/|\cos(\theta)|)$, i.e.,

$$\|\xi_{S_c}\|_1 \leq \left( \frac{\tau}{\vartheta} (|\cos(\theta')|/|\cos(\theta)|) \right) \|\xi_S\|_1 \leq \frac{\tau}{\gamma \vartheta} \|\xi_S\|_1,$$

which holds for all $S \subset \{1, \ldots, n\}$ with $|S| \leq k$. In particular, it holds for all subset $S$ with $|S| = 1$. This implies that $A^T$ has the RSP of order 1, and thus $A^T$ has the RSP of any order $k \geq 1$, which is the strongest RSP. Note that this conclusion is restrictive, since it is drawn from the strong assumption on the angle $\theta'$ as above. So
we believe that this observation has not given a full picture of the true relationship between RSP and NSP yet.

We now study the properties of reweighted $\ell_1$-algorithms under the RSP assumption. The first result below shows that if the RSP is satisfied, Algorithm 2.2 with $F_c \in \mathcal{M}$ generates a sparse solution from any initial point in the sense that at least one component of $x^k$ tends to zero.

**Theorem 3.7.** Let $A \in \mathbb{R}^{m \times n}$ with $m \leq n$, and assume that $A^T$ has the RSP of order $K$ with constant $\rho > 0$ satisfying $1 + \rho < \frac{\pi}{4}$. Let $F_c \in \mathcal{M}$ and $\{(x^k, v^k)\}$ be generated by Algorithm 2.2. Then

$$[\sigma(x^k)]_n = \min_{1 \leq i \leq n} |x^k_i| \to 0 \text{ as } k \to \infty. \quad (3.6)$$

**Proof.** We now assume the contrary that (3.6) does not hold. Then there exist a constant $\gamma > 0$ and a subsequence, denoted by $\{(x^{k_j}, v^{k_j})\}_{j=1}^\infty$, such that $[\sigma(x^{k_j})]_n \geq \gamma > 0$ for all $j$. Thus $v^{k_j}(= |x^{k_j}|) \geq \gamma e$ for all $j$. By Lemma 3.3, we have $\|v^{k_j+1} - v^{k_j}\| \to 0$ as $j \to \infty$. So there exists a $j'$ such that for all $j \geq j'$ the vector $v^{k_j+1}$ is positive, i.e., $v^{k_j+1} \in R^{n+}_+$. Let $j \geq j'$ and consider the $k_j$-th step of the algorithm. Note that $(x^{k_j+1}, v^{k_j+1})$ is an optimal solution to the LP min$(\nabla F_{x^{k_j}}(v^{k_j}))^T v : Ax = b, |x| \leq v$, which can be written as

$$\min \{\nabla F_{x^{k_j}}(v^{k_j})^T v : Ax = b, \ x_i \leq v_i, -x_i \leq v_i, \ i = 1, \ldots, n\}. \quad (3.8)$$

By the optimality condition, there exist $\alpha^{k_j}, \beta^{k_j} \in R^n_+$ and $\lambda^{k_j} \in \mathbb{R}^m$ such that

$$A^T \lambda^{k_j} - \alpha^{k_j} + \beta^{k_j} = 0, \quad (3.7)$$

$$\nabla F_{x^{k_j}}(v^{k_j}) - \alpha^{k_j} - \beta^{k_j} = 0, \quad (3.8)$$

$$-x^{k_j+1}_i - v^{k_j+1}_i \leq 0, \quad \alpha^{k_j}_i \left(-x^{k_j+1}_i - v^{k_j+1}_i\right) = 0, \quad \alpha^{k_j}_i \geq 0, \quad i = 1, \ldots, n, \quad (3.9)$$

$$x^{k_j+1}_i - v^{k_j+1}_i \leq 0, \quad \beta^{k_j}_i \left(x^{k_j+1}_i - v^{k_j+1}_i\right) = 0, \quad \beta^{k_j}_i \geq 0, \quad i = 1, \ldots, n. \quad (3.10)$$

Since $v^{k_j+1} \in R^{n+}_+$, for every $i$ one of the inequalities $x^{k_j+1}_i - v^{k_j+1}_i \leq 0$ and $-x^{k_j+1}_i - v^{k_j+1}_i \leq 0$ holds strictly. Thus by the complementarity conditions (3.9) and (3.10), we see that for every $i$, either $\alpha^{k_j}_i$ or $\beta^{k_j}_i$ must be zero. On the other hand, since $v^{k_j} \in R^{n+}_+$, $\pi^{k_j} > 0$ and $F_{x^{k_j}} \in \mathcal{M}$, it implies that $\nabla F_{x^{k_j}}(v^{k_j}) \in R^{n+}_+$ (Assumption 2.1(c)). Thus it follows from (3.8) that for every $i$, $\alpha^{k_j}_i$ and $\beta^{k_j}_i$ cannot vanish at the same time. So we conclude that for every $i$, one and only one of $\alpha^{k_j}_i$ and $\beta^{k_j}_i$ is equal to zero, and hence $\alpha^{k_j}_i + \beta^{k_j}_i = |\alpha^{k}_i - \beta^{k}_i|$ which, together with (3.7) and (3.8), implies that

$$\nabla F_{x^{k_j}}(v^{k_j}) = \alpha^{k_j} + \beta^{k_j} = |\alpha^{k}_i - \beta^{k}_i| = |A^T \lambda^{k_j}| = |\xi^{k_j}|, \quad (3.11)$$

where $\xi^{k_j} = A^T \lambda^{k_j} \in \mathcal{R}(A^T)$. Thus $\sigma(\nabla F_{x^{k_j}}(v^{k_j})) = \sigma(\xi^{k_j})$. Since $A^T$ has the RSP of order $K$ with $\rho > 0$, we have

$$\|\xi^{k_j}\|_1 \leq \rho \|\xi^{k_j}\|_1. \quad (3.11)$$
for any $S \subseteq \{1, \ldots, n\}$ with $|S| \geq K$. In particular, we consider the index set $S$ which is the set of indices $i$ such that $|\xi^{k_j}_i| = |\sigma(\xi^{k_j}_i)|$, where $n - K + 1 \leq j \leq n$, i.e., $S$ denotes the set of indices corresponding to the $K$ smallest components of $|\xi^{k_j}|$. Thus,

$$\left\| \xi^{k_j}_{S_i} \right\|_i = \sum_{i=1}^{n-K} |\sigma(\xi^{k_j}_i)| = \sum_{i=1}^{n-K} \left[ \sigma(\nabla F_{\epsilon \xi_j}(v^{k_j})) \right]_i,$$

and

$$\left\| \xi^{k_j}_{S_i} \right\|_i = \sum_{i=n-K+1}^{n} |\sigma(\xi^{k_j}_i)| = \sum_{i=n-K+1}^{n} \left[ \sigma(\nabla F_{\epsilon \xi_j}(v^{k_j})) \right]_i.$$

Therefore, by (3.11), we have

$$\sum_{i=1}^{n-K} \left[ \sigma(\nabla F_{\epsilon \xi_j}(v^{k_j})) \right]_i \leq \rho \sum_{i=n-K+1}^{n} \left[ \sigma(\nabla F_{\epsilon \xi_j}(v^{k_j})) \right]_i. \quad (3.12)$$

However, we have

$$\sum_{i=1}^{n-K} \left[ \sigma(\nabla F_{\epsilon \xi_j}(v^{k_j})) \right]_i \geq (n - K) \left[ \sigma(\nabla F_{\epsilon \xi_j}(v^{k_j})) \right]_{n-K+1} \geq \rho K \left[ \sigma(\nabla F_{\epsilon \xi_j}(v^{k_j})) \right]_{n-K+1}, \quad (3.13)$$

where the second inequality follows from $n - K > \rho K$ and $\left[ \sigma(\nabla F_{\epsilon \xi_j}(v^{k_j})) \right]_{n-K+1} > 0$ (since $\nabla F_{\epsilon \xi_j}(v^{k_j}) \in R^n_{++}$). The right-hand side of (3.13) is greater than or equal to

$$\rho \sum_{i=n-K+1}^{n} \left[ \sigma(\nabla F_{\epsilon \xi_j}(v^{k_j})) \right]_i,$$

contradicting with (3.12). Thus (3.6) holds. \[\square\]

The above result holds for the whole family of merit functions for sparsity in $M$. Roughly speaking, Algorithm 2.2 guarantees to generate a sparse solution if the RSP is satisfied. However, this result does not confirm how sparse the solution is. So Theorem 3.7 can be called a weak convergence theorem, based on which some stronger convergence result can be further proved. To this end, we need to establish the next technical result. Recall that for a given vector $x$, $T(x) = \{ i : x_i \neq 0 \}$ denotes the support of $x$.

**Lemma 3.8.** Let $A \in R^{m \times n}$ with $m \leq n$, and assume that $A^T$ has the RSP of order $K$ with constant $\rho > 0$ satisfying $1 + \rho < \frac{1}{K}$. Let $F_x \in M$ and $\{ (x^k, v^k) \}$ be generated by Algorithm 2.2 such that $\|v^{k+1} - v^k\| \to 0$ as $k \to \infty$. If there exists a positive constant $\mu > 0$ such that $|I_\mu(x^k)| \geq K$ for all sufficiently large $k$, where $I_\mu(x^k) = \{ i : |x_i^k| \geq \mu \}$, then there exists a $k'$ such that $\|x^{k'}\|_0 < n$ and $T(x^{k'}) \subseteq T(x^{k})$ for all $k \geq k'$.

**Proof.** At the $k$-th step, since $(x^{k+1}, v^{k+1})$ is an optimal solution to the LP problem $\min \{ \nabla F_{\epsilon \xi_j}(v^{k})^T v : Ax = b, |x| \leq v \}$, by the optimality condition there exist
\[\alpha^k, \beta^k \in R_+^n\] and \(\lambda^k \in R^m\) such that
\[
A^T \lambda^k - \alpha^k + \beta^k = 0, \quad (3.14)
\]
\[
\nabla F_{x_k}(v^k) - \alpha^k - \beta^k = 0, \quad (3.15)
\]
\[
-x_i^{k+1} - v_i^{k+1} \leq 0, \quad \alpha_i^k (x_i^{k+1} - v_i^{k+1}) = 0, \quad \alpha_i^k \geq 0, \quad i = 1, \ldots, n, \quad (3.16)
\]
\[
x_i^{k+1} - v_i^{k+1} \leq 0, \quad \beta_i^k (x_i^{k+1} - v_i^{k+1}) = 0, \quad \beta_i^k \geq 0, \quad i = 1, \ldots, n, \quad (3.17)
\]
\[
\sum_{i \in I^k(x^k)} \alpha_i^k = 0 \quad \text{for every} \quad i \in I^k(x^k). \quad (3.20)
\]

By (3.19) and (3.20), for all sufficiently large \(k\) we have
\[
\sum_{i \in I^k(x^k)} |\alpha_i^k - \beta_i^k| = \sum_{i \in I^k(x^k)} \alpha_i^k + \beta_i^k = \sum_{i \in I^k(x^k)} \nabla F_{x_k}(v^k)_i \leq \sum_{i \in I^k(x^k)} \nabla F_{x_k}(\mu e)_i. \quad (3.21)
\]

The right-hand side of the above is bounded. In fact, by Assumption 2.1(c), \(\nabla F_{x_k}(\mu e)_i\) is continuous in \(\varepsilon\) and there exists constant \(\gamma_i^* > 0\) such that \(\nabla F_{x_k}(\mu e)_i \to \gamma_i^*\) as \(\varepsilon \to 0\). Thus there exists a constant \(W^*\) such that \(\sum_{i=1}^n |\nabla F_{x_k}(\mu e)_i| \leq W^*\) for all \(\varepsilon \in (0, \varepsilon_0]\). Note that \(\varepsilon_k \in (0, \varepsilon_0]\). It follows from (3.21) that for all sufficiently large \(k\)
\[
\sum_{i \in I^k(x^k)} |\alpha_i^k - \beta_i^k| \leq \sum_{i \in I^k(x^k)} |\nabla F_{x_k}(\mu e)_i| \leq \sum_{i=1}^n |\nabla F_{x_k}(\mu e)_i| \leq W^*. \quad (3.22)
\]

We now prove that \(\|x^k\|=|T(x^k)| < n\) for all sufficiently large \(k\). From Assumption 2.1(c), \(|F_{x_k}(v)| \to \infty\) as \((v_i, \varepsilon) \to 0\). Thus there exists a small constant \(0 < \epsilon^* < \mu\) such that
\[
|\nabla F_{x_k}(v)| > \rho W^* \quad \text{for any} \quad |v_i| + \varepsilon \leq \epsilon^*. \quad (3.23)
\]

By Theorem 3.7, we have that \(|x^{k+1}| \to 0\) as \(k \to \infty\). Thus there exists a sufficiently large number \(k'\) such that \(|x^{k+1}| + \varepsilon < \epsilon^*\) for all \(k \geq k'\). Let \(i_0\) be the index such that \(v_{i_0}^{k'} = |x_{i_0}^{k'}| = |\sigma(x^{k'})|_{n_i}. \) Since \(\epsilon^* < \mu,\) we see that \(i_0 \notin I_{x^{k'}}\). Consider the next point \((x^{k'+1}, v^{k'+1})\), which satisfies the optimality condition (3.14)-(3.18) where \((\lambda^k, v^k, x^{k+1}, v^{k+1}, \alpha^k, \beta^k, \varepsilon_k)\) is replaced by \((\lambda^{k'}, v^{k'}, x^{k'+1}, v^{k'+1}, \alpha^{k'}, \beta^{k'}, \varepsilon_{k'})\). We
now prove that \( x_{i_0}^{k'+1} = 0 \). In fact, if it is not, by (3.15)-(3.17), one and only one of \( \alpha_{i_0}^{k'} \) and \( \beta_{i_0}^{k'} \) is zero. This, together with (3.15) and (3.23), implies that

\[
|\alpha_{i_0}^{k'} - \beta_{i_0}^{k'}| = \alpha_{i_0}^{k'} + \beta_{i_0}^{k'} = \left[ \nabla F_{\varepsilon_0}^{*}(x^{k'}) \right]_{i_0} > \rho W^*,
\]

and thus

\[
\sum_{i \notin I_\rho(x^{k'})} |\alpha_i^{k'} - \beta_i^{k'}| \geq |\alpha_{i_0}^{k'} - \beta_{i_0}^{k'}| > \rho W^*. \tag{3.24}
\]

However, since \( |I_\mu(x^k)| \geq K \) and \( A^T \) has the RSP of order \( K \) with constant \( \rho > 0 \), by (3.14) and (3.22), we have

\[
\sum_{i \notin I_\rho(x^{k'})} |\alpha_i^{k'} - \beta_i^{k'}| = \sum_{i \notin I_\rho(x^{k'})} |A^T \lambda_i^{k'}| \leq \rho \sum_{i \in I_\rho(x^{k'})} |A^T \lambda_i^{k'}| = \rho \sum_{i \in I_\rho(x^{k'})} |\alpha_i^{k'} - \beta_i^{k'}| \leq \rho W^*.
\]  

This contradicts (3.24). So we conclude that \( x_{i_0}^{k'+1} = 0 \) and thus \( |T(x^{k'+1})| < n \). This also indicates that \( |\sigma(x^{k'+1})| \leq m \) where \( m = x_{i_0}^{k'+1} = 0 \). Replacing \( x^{k'} \) by \( x^{k'+1} \), considering the point \( (x^{k'+2}, v^{k'+2}) \) and repeating the same proof above, we can show that \( x_{i_0}^{k'+2} = 0 \). Thus, by induction, we conclude that for all \( k > k' \) the iterates will keep this component being zero. This proof can be applied to any other component \( x_i^k = 0 \) from which we can show that \( x_i^k = 0 \) for all \( k > k' \). Thus there exists a \( k' \) such that \( T(x^k) \subset T(x^{k'}) \) for all \( k \geq k' \).

The requirement that \( \|v^{k+1} - v^k\| \to 0 \) as \( k \to \infty \) used in Lemma 3.8 is mild, and it can hold trivially when the merit function is suitably chosen. For instance, by Corollary 3.4, the merit functions in Examples 2.4 and 2.5 can ensure this condition.

In what follows, we prove the next main result in this section. Let \( I_\rho(x) \) be defined as in Lemma 3.8. Since \( b \neq 0 \), there exists a small number \( \mu_0 > 0 \) such that for any given \( \mu \in (0, \mu_0) \) the set \( I_\mu(x) \neq \emptyset \) for any solution of the system \( Ax = b \). Clearly, we have \( I_\mu(x) \subseteq T(x) \) (and thus \( |I_\mu(x)| \leq |T(x)| \)) for any \( x \in R^n \). The next result is stronger than Theorem 3.7.

THEOREM 3.9. Let \( A \in R^{m \times n} \) with \( m \leq n \). Assume that \( A^T \) has the RSP of order \( K \) with constant \( \rho > 0 \) satisfying \( 1 + \rho < \frac{\alpha}{\beta} \). Let \( F_\varepsilon \in \mathcal{M} \) and the sequence \( \{(x^k, v^k)\} \) be generated by Algorithm 2.2. If \( \|v^{k+1} - v^k\| \to 0 \) as \( k \to \infty \), then there is a subsequence \( \{x^j\} \) that converges to a \( \{(1 + \rho)K\}\)-sparse solution of \( Ax = b \) in the sense that \( |\sigma(x^j)|_{(1 + \rho)K+1} \to 0 \) as \( j \to \infty \).

Proof. Consider the sequence \( \{(x^k, v^k)\} \) generated by Algorithm 2.2. Under the condition of the theorem, we prove that it has a subsequence \( \{x^{j_i}\} \) convergent to a \( \{(1 + \rho)K\}\)-sparse solution in the sense that \( |\sigma(x^{j_i})|_{(1 + \rho)K+1} \to 0 \) as \( j \to \infty \). We prove this by contradiction. Assume the contrary that there is no such subsequence. Then \( |\sigma(x^j)|_{(1 + \rho)K+1} \) must be bounded away from zero, i.e., there exists a number \( \mu^* > 0 \) such that \( |\sigma(x^j)|_{(1 + \rho)K+1} \geq \mu^* \) for all sufficiently large \( k \). In other words

\[
|I_{\mu^*}(x^j)| \geq \lfloor (1 + \rho)K + 1 \rfloor > (1 + \rho)K
\]
for all large $k$. By Lemma 3.8, there exists a $k'$ such that
\[ \|x^k\|_0 = |\mathcal{T}(x^k)| < n, \quad \mathcal{T}(x^k) \subseteq \mathcal{T}(x^{k'}) \] for all $k \geq k'$. 

Note that $I_{\mu^*}(x^k) \subseteq \mathcal{T}(x^k)$. So
\[ n > |\mathcal{T}(x^k)| \geq |I_{\mu^*}(x^k)| > (1 + \rho)K > K \] for any $k \geq k'$.

If $|\mathcal{T}(x^k)| > |I_{\mu^*}(x^k)|$ at the $k$-step $(k \geq k')$, we can prove that the algorithm will continue to reduce the value of $|\mathcal{T}(x^k)|$ until for some $k'' > k'$ we have $|\mathcal{T}(x^k)| = |I_{\mu^*}(x^k)|$ for all $k \geq k''$. Since $F_c(v)$ is separable in $v$, it can be represented as $F_c(v) = \sum_{i=1}^n \phi_i(\varepsilon, v_i)$, where $\phi_i$'s are some kernel functions. From (3.26), we see that if $x_i^k = 0$ then $x_i^k = 0$ for all $k \geq k'$. Thus for all $k \geq k'$ the problem (2.5), i.e.,
\[ \min \left\{ \langle \nabla F_c(v^k)^T v : Ax = b, |x| \leq v \rangle \right\}, \] is exactly equivalent to the reduced problem
\[ \min \left\{ \langle \nabla F_c(v^k)^T v_S : A_S x_S = b, |x_S| \leq v_S \rangle \right\}, \] where $S = \mathcal{T}(x^k)$. In other words, for all $k \geq k'$ the solution $(x^{k+1}, v^{k+1})$ to (2.5) can be partition into $(x^{k+1}, v^{k+1}) = (x_S^{k+1}, 0), v^{k+1} = (v_S^{k+1}, 0)$ where $(x_S^{k+1}, v_S^{k+1})$ is the solution to the reduced problem (3.28). The merit function for sparsity associated with (3.28) is given by
\[ F_c(v_S) := \sum_{i \in S} \phi_i(\varepsilon, v_i), \] which still satisfies the Assumption 2.1 in space $\ell_2^{|S|}$, where $|S| = |\mathcal{T}(x^k)| > K$ by (3.27). The reduced function above is obtained from $F_c(v)$ by simply dropping the components $\phi_i(\varepsilon, v_i)$ with $i \in S_c \neq \{1, ..., n\} \setminus S$, and retaining the ones with indices in $S$ only. We now show that the reduced matrix $A_S^T$ has the same RSP as that of $A_T$. Indeed, let $\eta \in \mathcal{R}(A_S^T)$. Then there exists a $\lambda$ such that $\eta = A_S^T \lambda$. Setting $\eta' = A_S^T \lambda$ and rearranging the components of $(\eta, \eta')$ if necessary, we have $(\eta, \eta') \in \mathcal{R}(A_T^T)$. For any $L \subseteq S$ with $|L| \geq K$, by the RSP of $A_T^T$, we have
\[ \|(\eta_L, \eta')\|_1 \leq \rho \|\eta_L\|, \] where $L_c = S \setminus L$. Thus $\|\eta_L\|_1 \leq \rho \|\eta_L\|$, which implies that $A_S^T$ satisfies the RSP of order $K$ with the same constant $\rho$. Similarly, it is evident that removing rows from the matrix $A$, the transpose of the resulting submatrix still satisfies the RSP with the same constant $\rho$ as that of $A_T^T$. Note that $A_S \in \mathbb{R}^{n \times |S|}$. If $m > |S|$, the rows of $A_S$ are linearly dependent, and hence some equations of $A_S x_S = b$ are redundant, and can be removed from the system without any change to the solution of (3.28).

Therefore without loss of generality, the size of $A_S$ can be assume to be $m \times |S|$ with $m \leq |S|$, and $A_S^T$ has the RSP of order $K$ with the same constant as that of $A_T^T$, as shown above. By (3.27), we have that $1 + \rho < \frac{|S|}{K}$ where $|S| = |\mathcal{T}(x^k)|$. Therefore, applying Lemma 3.8 to the $m \times |S|$ matrix $A_S$ and the reduced merit function for sparsity $F_c(v_S) = \sum_{i \in S} \phi_i(\varepsilon, v_i)$ where $v_S \in \mathbb{R}^{|S|}$, we conclude that there exists a $k'' > k'$ such that $|\mathcal{T}(x^{k''})| < |S|$, $\mathcal{T}(x^{k''}) \subseteq \mathcal{T}(x_S^{k''})$ for all $k \geq k''$. Notice that $x^{k''}$ and $x^k$ are partitioned, respectively, into $(x_S^{k''}, 0)$ and $(x_S^{k''}, 0)$ for all $k \geq k''$. This is equivalent to
\[ |\mathcal{T}(x^{k''})| < |S| = |\mathcal{T}(x^k)|, \quad \mathcal{T}(x^k) \subseteq \mathcal{T}(x^{k''}) \] for all $k \geq k''$. 

REWEIGHTED $\ell_1$-MINIMIZATION

17
If $|T(x^{k''})|$ remains larger than $|I_{\mu^*}(x^{k''})|$, which is larger than $|(1+\rho)K|$ by (3.27), then replace $x^{k''}$ by $x^{k'''}$ and repeat the same proof above, we can conclude that there exists $k''' > k''$ such that $|T(x^{k''})|$ is strictly smaller than $|T(x^k)|$. Therefore, by induction, there must exist an integer number, denoted still by $k''$, such that

$$T(x^k) = I_{\mu^*}(x^k), \quad T(x^k) \subseteq T(x^{k''}) \text{ for all } k \geq k''.$$ 

Let $S = I_{\mu^*}(x^k)$, which is larger than $(1+\rho)K$ by (3.27). The above relation implies that for all $k \geq k''$ the vector $x^k$ is $|S|$-sparse vector where $|S| = |I_{\mu^*}(x^k)| > (1+\rho)K$, and all nonzero components of $x^k$ are bounded below by $\mu^* > 0$. All the rest iterations are equivalent to solving the reduced minimization problem (3.28) with $S = I_{\mu^*}(x^k) = T(x^k)$. Note that $A_S$ is a submatrix of $A$, so $A_S$ satisfies the RSP with the same order and constant. Thus applying to the reduced merit function $F_3(x_S)$, Theorem 3.7 implies that $[\sigma(x^k_{S})]|_{|S|} \to 0$ as $k \to \infty$, i.e., the smallest component of $x_{|S|}^k$ tends to zero, which contradicts with $x^k_{|S|} \geq \mu^*e > 0$. This contradiction shows that there must exist a subsequence $\{x^{k_j}\}$ convergent to a $(1+\rho)K$-sparse solution in the sense that $[\sigma(x^{k_j})]|_{[(1+\rho)K]+1} \to 0$ as $j \to \infty$. 

An immediate result is given as follows.

**Corollary 3.10.** Under the same condition of Theorem 3.7. Let $\{(x^j, v^j)\}$ be the sequence generated by Algorithm 2.2, and let $\{(x^j, v^j)\}$ be the subsequence such that $[\sigma(x^j)]|_{[(1+\rho)K]+1} \to 0$ as $j \to \infty$. Then the following hold.

(i) For any given integer number $t \geq 1$, the subsequence $\{x^{j+t}\}$ converges also to a $(1+\rho)K$-sparse solution in the sense that $[\sigma(x^{j+t})]|_{[(1+\rho)K]+1} \to 0$ as $j \to \infty$.

(ii) If $v^k \to v^*$, then any accumulation point of $\{x^k\}$ is a $(1+\rho)K$-sparse solution of $Ax = b$. In particular, if $x^k \to x^*$ or $v^k \to v^*$, then $x^k$ converges to the sparsest solution.

**Proof.** Clearly, we have

$$|\sigma(x^{j+t})|_{[(1+\rho)K]+1} - |\sigma(x^j)|_{[(1+\rho)K]+1}| \leq \left| \sigma(x^{j+t}) - \sigma(x^j) \right| = \left| \sigma(v^{j+t}) - \sigma(v^j) \right| \leq \left| v^{j+t} - v^j \right|.$$ 

Note that $\left| v^{j+t} - v^j \right| \leq \sum_{i=1}^{t} \left| v^{j+i} - v^{j+i-1} \right| \to 0$ as $j \to \infty$, which follows from $\left| v^{j+i} - v^j \right| \to 0$ as $k \to \infty$. Combining the two relations above leads to the result (i). The results (ii) and (iii) are evident. 

As shown by Corollary 3.4, the merit function in $M$ can be chosen to ensure that the sequence $\{x^k\}$ is bounded and $\|x^{k+1} - v^k\| \to 0$. So the requirement $\|x^{k+1} - v^k\| \to 0$ used in Theorem 3.9 and Corollary 3.10 can be removed when the merit functions are suitably chosen (e.g., Examples 2.4 and 2.5). We summarize this result as follows.

**Theorem 3.11.** Assume that $A^T$ has the RSP of order $K$ with constant $\rho > 0$ satisfying that $1+\rho < \frac{K}{\mu^*}$. Let $F_c \in M$ and $F_e(v)$ be bounded below in $(x, v) \in R^n \times R_+$, and $g(x) = \inf_{c \in C} F_c(x)$ be coercive in $R^n$. Let $\{(x^k, v^k)\}$ be generated by Algorithm 2.2. Then there is a subsequence $\{x^{k_j}\}$ that converges to a $(1+\rho)K$-sparse solution of $Ax = b$ in the sense that $[\sigma(x^{k_j})]|_{[(1+\rho)K]+1} \to 0$ as $j \to \infty$.

**Remark.** Except for Corollary 3.4 and Theorem 3.11, all other results established in this section can be viewed as the common properties shared among the reweighted $\ell_1$-algorithms based on the merit function in $M$. Theorem 3.7 claims that under RSP, any reweighted $\ell_1$-minimization algorithm associated with a merit function in $M$ can
find a sparse solution of the underdetermined linear system. If the sequence generated by the algorithm satisfies \( \|v^{k+1} - v^k\| \to 0 \), then Theorem 3.9 further claims that, under RSP of order \( K \) with \( (1+\rho)K < n \), the algorithm can find at least a \([ \lfloor (1+\rho)K \rfloor \)-sparse solution. Since the CWB method falls into the framework of Algorithm 2.2 and it is based on a merit function in \( \mathcal{M} \) (see Example 2.3), a convergence result for the CWB method can be extracted from Theorem 3.7 and 3.9 and their corollaries immediately. The statement of this special result is omitted here. Theorem 3.11 that is stronger than Theorem 3.9 has further identified a subclass of merit functions in \( \mathcal{M} \), including \( \text{WL}_p \), \( \text{NW2} \), and \( \text{NW3} \) methods, which can ensure that the generated sequence is bounded and satisfies \( \|v^{k+1} - v^k\| \to 0 \). Thus, the convergence results for \( \text{WL}_p \), \( \text{NW2} \), and \( \text{NW3} \) methods can be immediately obtained from Theorem 3.11 as well.

4. Numerical Experiments. As seen in section 3, \( \mathcal{M} \) is a large family of merit functions for sparsity, based on which various reweighted \( \ell_1 \)-methods can be constructed. It is interesting to compare these algorithms through numerical experiments. Since it is impossible to test all algorithms of this family, we single out a few of them, and compare their performances in finding sparse solution of underdetermined linear systems. Let us first list a few of these specific methods as follows.

(a) Candès-Wakin-Boyd (CWB) method

\[
x^{k+1} = \arg\min \left\{ \sum_{i=1}^{n} \left( \frac{1}{|x_i^k| + \varepsilon_k} \right) |x_i| : Ax = b \right\}.
\]

(b) \( \text{WL}_p \) method

\[
x^{k+1} = \arg\min \left\{ \sum_{i=1}^{n} \left( \frac{1}{(|x_i^k| + \varepsilon_k)^{1-p}} \right) |x_i| : Ax = b \right\}, \quad p \in (0, 1).
\]

(c) \( \text{NW1} \) algorithm derived from Example 2.3(ii)

\[
x^{k+1} = \arg\min \left\{ \sum_{i=1}^{n} \left( \frac{p + (|x_i^k| + \varepsilon_k)^{1-p}}{(|x_i^k| + \varepsilon_k)^{1-p} [x_i^k] + \varepsilon_k + (|x_i^k| + \varepsilon_k)^p} \right) |x_i| : Ax = b \right\},
\]

where \( p \in (0, 1) \).

(d) \( \text{NW2} \) algorithm derived from Example 2.5

\[
x^{k+1} = \arg\min \left\{ \sum_{i=1}^{n} \left[ \frac{q + (|x_i^k| + \varepsilon_k)^{1-q}}{(|x_i^k| + \varepsilon_k)^{1-q} [x_i^k] + \varepsilon_k + (|x_i^k| + \varepsilon_k)^q} \right] |x_i| : Ax = b \right\},
\]

where \( p, q \in (0, 1) \).

(e) \( \text{NW3} \) algorithm derived from Example 2.5

\[
x^{k+1} = \arg\min \left\{ \sum_{i=1}^{n} \left( \frac{1 + 2(|x_i^k| + \varepsilon_k)}{(|x_i^k| + \varepsilon_k + (|x_i^k| + \varepsilon_k)^2)^{1-p}} \right) |x_i| : Ax = b \right\},
\]

where \( p \in (0, 1/2) \).

(f) \( \text{NW4} \) algorithm based on Example 2.6

\[
x^{k+1} = \arg\min \left\{ \sum_{i=1}^{n} \left( \frac{1 + (|x_i^k| + \varepsilon_k)^p}{(|x_i^k| + \varepsilon_k)^{1+p}} \right) |x_i| : Ax = b \right\}, \quad p \in (0, \infty).
\]
To compare these methods, we randomly generate the pair \((A, x)\), where \(A \in \mathbb{R}^{50 \times 250}\) and \(x\) is a \(k\)-sparse vector in \(\mathbb{R}^{250}\) with \(k = 1, 2, \ldots, 30\). Throughout this section, all random pairs \((A, x)\) are generated based on the following assumption: The entries of \(A\) and \(x\) on its support are i.i.d Gaussian random variables with zero mean and unit variances. Once \((A, x)\) is generated, we set \(b = Ax\) and test the algorithms on this system. For every given sparsity \(k\), 500 pairs of \((A, x)\) were generated, and we compare the success probability of all the above-mentioned algorithms in locating \(k\)-sparse solutions. For all tested instances of \(Ax = b\), every reweighted algorithm was executed only 4 iterations, and the same parameters \(\alpha = 0.5, \varepsilon_0 = 0.01\), and the initial point \(x^0 = e \in \mathbb{R}^{250}\) were used in Algorithm 2.2. Note that \(x^0 = e\) implies that the first iteration of the algorithm is actually the \(\ell_1\)-minimization. Given a \(k\)-sparse solution \(x\) of \(Ax = b\), the algorithm claims to be successful in finding (or exact reconstruction of) the \(k\)-sparse solution \(x^k\) if the found solution \(x^k\) satisfies that \(\|x^k\|_0 \leq k\) and \(\|x^k - x\| \leq 10^{-5}\) where \(\|x^k\|_0\) is defined, in our experiments, to be the number of components of \(x\) satisfying \(|x^k_i| \geq 10^{-5}\). Clearly, the main computational cost in Algorithm 2.2 is solving weighted \(\ell_1\)-minimization problems. To solve these problems, we use CVX, a package for specifying and solving convex programs [26].

Since NW2 has two parameters \((p, q)\), we set \(q = p\) for this algorithm in all tests for simplicity. Our experiments show that no matter what value of \(p \in (0, 1)\) is taken (for NW3, \(p\) is restricted in \((0, 1/2)\)), all six reweighted \(\ell_1\)-algorithms defined above remarkably outperform \(\ell_1\)-minimization in recovering the desired sparse solutions. The new algorithms NW2 and NW3 proposed in this paper and the existing \(W_{l_p}\) method are particularly strong. The results for \(p = 0.3\) and 0.5 are summarized in Figure 4.1, in which the success probability of \(\ell_1\)-minimization is also included.

![Figure 4.1](image-url)

**Fig. 4.1.** Comparison of success rates of finding the \(k\)-sparse solution \(x\) of \(b = Ax\), where \(A \in \mathbb{R}^{50 \times 250}\) and \(x \in \mathbb{R}^{250}\). For each \(k\)-sparsity, 500 attempts were made. All six reweighted \(\ell_1\)-methods outperform \(\ell_1\)-minimization.

The success probability of these algorithms are different. For example, for the sparsity \(k = 15\), Figure 4.1 (a) shows that the success probability of \(\ell_1\)-minimization to find the desired sparse solution is about 17\%, NW4 is about 28\%, NW1 and CWB are about 38\%, NW3 is about 46\%, and NW2 and \(W_{l_p}\) are about 52\%. A similar result can be seen from Figure 4.1 (b). Figure 4.1 demonstrates that NW2 and \(W_{l_p}\) methods perform particularly well in finding the sparse solutions of linear systems. Thus it is interesting to single out these two algorithms, and further compare their performances (also with the \(\ell_1\) - and CWB method). The experiments show that when
$p$ is relatively small (e.g. $p < 0.2$), NW2 outperforms $W_1^p$ method. For $0.3 \leq p \leq 0.6$, NW2 and $W_1^p$ are quite comparable, and in many cases their numerical performances are almost identical. When $p$ is relatively large (e.g. $p \geq 0.7$), $W_1^p$ method can outperform NW2 in many situations. The results for $p = 0.01, 0.1, 0.3$ and 0.7 were summarized in Figure 4.2.

![Graphs showing performance comparison](image)

**Fig. 4.2.** Comparison of NW2, $W_1^p$, CWB, and $\ell_1$-minimization through the success frequency of finding the $k$-sparse solution of $b = Ax$, where $A \in \mathbb{R}^{50 \times 250}$. 500 attempts were made for each $k$-sparsity ($k = 1, \ldots, 30$), and different values of $p$ were tested.

The above numerical experiments were carried out by using the updating rule $\varepsilon_{k+1} = \alpha \varepsilon_k$ where $\alpha = 0.5$. These experiments have demonstrated that all tested reweighted $\ell_1$-algorithms (NW1-NW4, CWB, and $W_1^p$) do outperform the standard $\ell_1$-minimization in many situations. We also observed that in the aforementioned testing environment, NW2, NW3 and $W_1^p$ perform better than CWB, NW1, and NW4 in many situations, and NW2 and $W_1^p$ methods are quite comparable to each other.

However, these numerical results cannot imply that the overall performance of NW2, NW3, and $W_1^p$ is always better than CWB, NW1, and NW4. It is interesting to test algorithms using a different parameter updating rule. Candès, Wakin and Boyd [9] proposed the following rule:

$$\varepsilon_k = \max \left\{ |\sigma(x^k)|_{i_0}, 10^{-3} \right\},$$

(4.1)

where $i_0$ denotes the nearest integer to $m/[4 \log(n/m)]$. Let us replace the updating scheme $\varepsilon_{k+1} = \varepsilon_k/2$ in Algorithm 2.2 by (4.1), and redo experiments. The results
for $p = 0.5$ and $0.3$ were summarized in Figure 4.3, from which we see that all tested reweighted algorithms still remarkably outperform the standard $\ell_1$-minimization, but this time CWB, NW1, and NW4 perform better than NW2, NW3, and Wlp, and these three methods (CWB, NW1, and NW4) are quite comparable and the recovery by these three methods is robust with respect to the choice of the parameter $p$.

![Comparison of success rates of reweighted algorithms with (4.1) in finding the $k$-sparse solution of $b = Ax$, where $A \in \mathbb{R}^{50 \times 250}$. For each $k$-sparsity, $500$ attempts were made.](image)

(i) $p = 0.5$, $\varepsilon_k$ is updated by (4.1) (ii) $p = 0.3$, $\varepsilon_k$ is updated by (4.1)

**Fig. 4.3.** Comparison of success rates of reweighted algorithms with (4.1) in finding the $k$-sparse solution of $b = Ax$, where $A \in \mathbb{R}^{50 \times 250}$. For each $k$-sparsity, 500 attempts were made.

In summary, all the tested reweighted methods can outperform the standard $\ell_1$-method in finding sparse solutions of linear systems. From Figures 4.1 and 4.3, the numerical performance of a reweighted algorithm may depend on the updating rule of the parameter $\varepsilon_k$, and we observe that CWB, NW1, and NW4 perform best when using (4.1), compared with the remaining algorithms using either the rule (4.1) or $\varepsilon_{k+1} = \varepsilon_k / 2$.

5. Conclusions. Via a merit function for sparsity which is certain concave approximation of the cardinality function, the concave minimization plays an important role in locating sparse solutions of underdetermined linear systems of equations. Through a linearization technique, minimizing concave merit functions for sparsity yields a unified approach for reweighted $\ell_1$-minimization algorithms. This unified approach not only makes it easy to construct various new specific reweighted $\ell_1$-algorithms for the sparse solution of linear systems, but also enables us to develop a new and unified convergence theory for a large family of such algorithms. The analysis in this paper is based on the so-called range space property, which is different from the existing RIP/NSP-based analysis. As special cases of our general framework, a convergence result for the well-known $\ell_p$-quasi-norm-based reweighted algorithm and Candès-Wakin-Boyd method can be obtained, respectively, from Theorems 3.11 and 3.9 in this paper. Moreover, several specific reweighted $\ell_1$-algorithms have been constructed, and their efficiency of finding sparse solutions of linear systems has been demonstrated by numerical experiments. Although the simulation shows that reweighted $\ell_1$-algorithms outperform the standard $\ell_1$-method in many situations, a rigorous mathematical proof for this phenomena has not been carried out so far. This remains an open question in this field. What we have actually proved in this paper is that, under suitable conditions, a large family of reweighted $\ell_1$-algorithms can generate a solution with a certain level of sparsity to the linear system.
REFERENCES


[26] R. Horst, P.M. Pardalos and N. V. Thoai, *Introduction to Global Optimization*, Kluwer Acad-


