A Generalization of Weighted Sparse Decomposition to Negative Weights

Ghazaleh Delfi*, Shayan Aziznejad*, Sana Amani* , Massoud Babaie-Zadeh*, and Christian Jutten†,
*Department of Electrical Engineering, Sharif University of Technology, Tehran, Iran.
Emails: {ghazalehdelfi, sh.aziznejad, sana.amani94}@gmail.com and mbzadeh@yahoo.com
†University Grenoble Alpes, CNRS, GIPSA-lab, Grenoble, France
Email: christian.jutten@gipsa-lab.grenoble-inp.fr

***Abstract*—Sparse solutions of underdetermined linear systems of equations are widely used in different fields of signal processing. This problem can also be seen as a sparse decomposition problem. Traditional sparse decomposition gives the same priority to all atoms for being included in the decomposition or not. However, in some applications, one may want to assign different priorities to different atoms for being included in the decomposition. This results to the so called "weighted sparse decomposition" problem [Babaie-Zadeh et al. 2012]. However, Babaie-Zadeh et al. studied this problem only for positive weights; but in some applications (e.g. classification) better performance can be obtained if some weights become negative. In this paper, we consider "weighted sparse decomposition" problem in its general form (positive and negative weights). A tight uniqueness condition and some applications for the general case will be presented.**

***Index Terms*—Sparse signal processing, Weighted sparse decomposition, Weighted ℓ_0 norm minimization, Negative weights decomposition, Weighted Sparse Representation for Classification**

I. Introduction

Solving an under-determined system of linear equations for sparse solutions has attracted lots of attention during the last decade [1]. Application examples include blind source separation [2], compressed sensing [3] and classification [4], to name a few. Consider the linear system of equations

$$\mathbf{As} = \mathbf{b}, \tag{1}$$

where $\mathbf{A}$ is an n by m matrix and $\mathbf{b}$ is an n by 1 vector. When $n < m$ and $\mathbf{A}$ is a full rank matrix, (1) has infinitely many solutions, but the sparsest solution is achieved by solving

$$P_0 : \min_{\mathbf{s}} \|\mathbf{s}\|_0 \;\; s.t. \;\; \mathbf{As} = \mathbf{b}, \tag{2}$$

where the ℓ_0 norm $\|s\|_0$ stands for the number of nonzero components of $\mathbf{s}$. Note that P_0 can be interpreted as a "sparse decomposition problem" where the aim is to decompose $\mathbf{b}$ as a linear combination of the minimum number of columns of $\mathbf{A}$ (called "atoms" after [5]).

It is seen that P_0 gives the same priority to all the atoms for being included in the decomposition or not. However, in some applications, one may want to assign different priorities to different atoms for being included in the decomposition. This problem is studied in [6] under the title "weighted sparse decomposition". It defines the weighted ℓ_0-norm as

$$\|\mathbf{s}\|_{0,\mathbf{w}} \triangleq \sum_{i=1}^{m} w_i |s_i|_0, \tag{3}$$

where $\mathbf{w} = [w_1, w_2, \ldots, w_m]^T$ is a known weight vector and $|s_i|_0$ is defined as

$$|s_i|_0 \triangleq \begin{cases} 0 & s_i = 0, \\ 1 & \text{otherwise.} \end{cases} \tag{4}$$

Therefore the weighted ℓ_0 norm minimization problem is expressed as

$$P_{0,\mathbf{w}} : \min_{\mathbf{s}} \|\mathbf{s}\|_{0,\mathbf{w}} \;\; s.t. \;\; \mathbf{As} = \mathbf{b}. \tag{5}$$

Note that the weight vector $\mathbf{w}$ is known and it is derived from the application.

In [6] all the assigned weights are assumed to be positive. However, in some applications there might be a need to consider negative weights as well. As a motivation for considering negative weights, suppose that in (1), the aim is to estimate $\mathbf{s}$ based on the observation vector $\mathbf{b}$. Moreover, assume that the $p_i = Pr(s_i \neq 0)$ has a Bernoulli distribution, and the probability of its activity, p_i, is known a priori. As [6] shows, the MAP estimation for this problem can be achieved by solving a $P_{0,\mathbf{w}}$ problem with the weights $w_i = \ln[(1-p_i)/p_i]$. Then, if the probability of a component being non-zero is higher than 0.5, w_i will have a negative value. Note that this sounds heuristically: if an entry of $\mathbf{s}$ is more likely to be active ($p_i > 0.5$), it is less expensive that it has a non-zero value ($w_i < 0$). Therefore, in order to solve this problem generally, it is needed to consider $P_{0,\mathbf{w}}$ with both negative and positive weights.

In this paper, we study $P_{0,\mathbf{w}}$ for the general case where there exist positive, zero and negative weights. We present a tight condition on the weights (w_i's) that guarantees the uniqueness of the solution of $P_{0,\mathbf{w}}$. Moreover, we present an application example for $P_{0,\mathbf{w}}$ with negative weights.

The paper is organized as follows: In the next section, we study the uniqueness condition of our problem and, using WSL0 algorithm [6], we will verify the condition numerically. In Section III, we provide an application example for our problem. Section IV concludes the paper.

II. Uniqueness and Algorithm

A. Theory

An important question is whether or not problem $P_{0,\mathbf{w}}$ has a unique solution. Note that $P_{0,\mathbf{w}}$ has at least one solution because the weighted ℓ_0 norm is bounded from below[1].

To explain our conditions which guarantee uniqueness, let Spark($\mathbf{A}$) denote the smallest number of columns of $\mathbf{A}$ which are linearly dependent [7]. It is proven that a solution $\mathbf{s}$ of (1) which satisfies $\|\mathbf{s}\|_0 < \frac{\text{Spark}(\mathbf{A})}{2}$ is the unique solution of P_0 [7]. Moreover [6] defines $S_{\mathbf{w}}(k)$ as the sum of k smallest weights. It is proved in [6] that for the problem $P_{0,\mathbf{w}}$ with positive weights, if there exists some $\mathbf{s}$ satisfying (1) and

$$\|\mathbf{s}\|_{0,\mathbf{w}} < \frac{S_{\mathbf{w}}(\text{Spark}(\mathbf{A}))}{2}, \tag{6}$$

then it is the unique solution of $P_{0,\mathbf{w}}$.

To generalize the above uniqueness conditions to negative weights, let us first make the following assumptions:

- H1: The weight vector $\mathbf{w}$ has at most n negative weights.
- H2: The solution of $P_{0,\mathbf{w}}$ has at most $n-1$ nonzero components, i.e. it has a "suitable" sparsity.

Remark 1: If H1 is not satisfied, i.e. if there exist more than n negative weights one can assign 0 to all other components and reach a system with at least two different solutions (see Lemma 1 in the appendix for a detailed proof).

The following theorem states a condition for the uniqueness of the solution of $P_{0,\mathbf{w}}$.

Theorem 1: Assume that H1 and H2 are satisfied and $\mathbf{s}$ is a solution of (1) such that

$$\|\mathbf{s}\|_{0,\mathbf{w}} < \frac{S_{\mathbf{w}}\left(Spark\left(\mathbf{A}\right)\right)}{2} + |w_{min}|\left(\frac{Spark\left(\mathbf{A}\right)}{2} - n + 1\right), \tag{7}$$

where $w_{min} \triangleq \min_i w_i$, then $\mathbf{s}$ is the unique solution of $P_{0,\mathbf{w}}$.

Proof is left to appendix.

Remark 2: Note that (6) is not a special case of (7) where all the weights are positive. In other words, (6) is stronger than the special case of (7) for positive weights. However, (7) is tight. To show this tightness, consider the problem (5) with $\mathbf{A}$, $\mathbf{b}$ and $\mathbf{w}$ be equal to

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}, \mathbf{b} = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}, \mathbf{w} = \begin{bmatrix} -4 \\ -4 \\ 6 \\ 6 \end{bmatrix}.$$

It is seen that $\mathbf{A}$ is a fullrank matrix and Spark($\mathbf{A}$) = 4. Moreover,

$$\frac{S_{\mathbf{w}}(\text{Spark}(\mathbf{A}))}{2} + |w_{min}|\left(\frac{\text{Spark}(\mathbf{A})}{2} - n + 1\right) = 2.$$

It is seen that $\mathbf{s}_1 = (1, 0, 1, 0)^T$ and $\mathbf{s}_2 = (0, 1, 0, 2)^T$ are both solutions to $\mathbf{As} = \mathbf{b}$, and for both of them $\|\mathbf{s}_1\|_{0,\mathbf{w}} = \|\mathbf{s}_2\|_{0,\mathbf{w}} = 2$. Moreover, no other solutions of (1) can have a weighted ℓ_0 norm smaller than 2, because a solution $\mathbf{s}$ of (1) is in the form of $\mathbf{s} = (\alpha, 1-\alpha, \alpha, 2-2\alpha)^T$ for any $\alpha \in \mathbb{R}$. So if $\alpha \notin \{0, 1\}$, all the components of $\mathbf{s}$ are different from zero. Then $\|\mathbf{s}_0\|_0 = 4$ and $\|\mathbf{s}\|_{0,\mathbf{w}} = 4 > 2$ i.e. (7) is not satisfied. The cases $\alpha = 1$ and $\alpha = 0$ correspond to $\mathbf{s}_1$ and $\mathbf{s}_2$, respectively. Both are solutions of $P_{0,\mathbf{w}}$ and do not satisfy the condition (7), although very marginally since (7) is then $2 < 2$. So this example illustrates Theorem 1 by contradiction.

B. Algorithm

In order to solve $P_{0,\mathbf{w}}$, we use the WSL0 algorithm of [6]. Careful examination of the equations of WSL0 in [6] reveals that the method does not use the assumption of positivity of w_i's and hence it is directly applicable to the general case of having both positive and negative weights. However, the initialization of that algorithm (which is the minimizer of the weighted ℓ_2 norm) heavily uses the assumption of positivity of all weights. Actually, the weighted ℓ_2 norm with negative weights is not necessarily bounded from below. So, in this paper, we heuristically use the minimizer of the unweighted ℓ_2 norm problem [8] (that is, $\mathbf{A}^{\dagger}\mathbf{y}$ where $\mathbf{A}^{\dagger}$ is the Moore-Penrose pseudo-inverse of $\mathbf{A}$) as initialization of the algorithm.

C. Simulation

We have designed an experiment in which we show that if the uniqueness condition of Theorem 1 holds then WSL0 algorithm results in a good reconstruction. In order to do that, first we consider a 100 by 300 random matrix $\mathbf{A}$ and two vectors $\mathbf{s}_1$ and $\mathbf{s}_2$ such that $\mathbf{As}_1 = \mathbf{As}_2$ and $\|\mathbf{s}_1\|_0 = 60$ ($\mathbf{s}_1$ is nonzero in its first 60 components) and $\|\mathbf{s}_2\|_0 = 50$ ($\mathbf{s}_2$ is nonzero in its first 5 components and its last 45 components). It is seen that, $\mathbf{s}_2$ is sparser than $\mathbf{s}_1$. Then, we assign a weight vector $\mathbf{w}$ such that $\|\mathbf{s}_1\|_{0,\mathbf{w}} < \|\mathbf{s}_2\|_{0,\mathbf{w}}$. To do this, we assign $w_i = -0.9$ for $i = 1, 2, \ldots, k$ and 1.2 otherwise (k varies from 1 to 120). It can be seen that $\mathbf{s}_1$ satisfies the uniqueness condition of Theorem 1 when $k \in\ [53, 68]$.

After running WSL0, we expect, when the uniqueness of Theorem 1 holds (i.e. for $k \in [53, 68]$) that the algorithm will estimate $\mathbf{s}_1$ rather than $\mathbf{s}_2$. We compare the estimated solution $\hat{\mathbf{s}}$ to $\mathbf{s}_1$ and $\mathbf{s}_2$ by calculating the signal to noise ratio (SNR) which is defined as

$$\text{SNR}_i = 20\log_{10}\left(\frac{\|\mathbf{s}_i\|_2}{\|\mathbf{s}_i - \hat{\mathbf{s}}\|_2}\right) \quad i = 1, 2. \tag{8}$$

Fig 1 shows that whenever the uniqueness condition holds (i.e. for $k \in [53, 67]$), the WSL0 algorithm provides a good estimation of the solution of $P_{0,\mathbf{w}}$.

III. An Application Example

As an application of weighted ℓ_0 norm minimization with some negative weights, we consider here the classification problem. The aim of this problem is to determine the class of an input data based on a previously labeled training dataset. One of the methods to solve this problem is SRC (Sparse

[1] It is also worth noting that if in $P_{0,\mathbf{w}}$ the weighted ℓ_0 norm is replaced by a weighted ℓ_1 norm, where negative weights exist, the resulting problem may have no solution, because weighted ℓ_1 norm is not bounded below for negative weights.

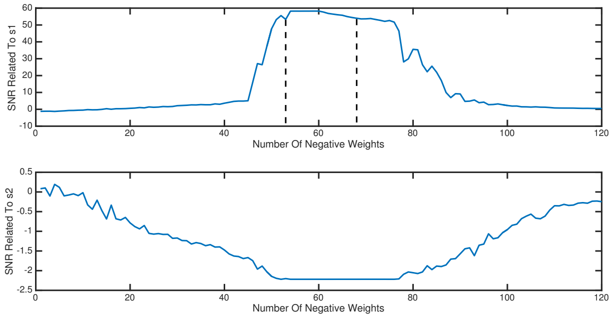

Fig. 1. The figure shows the SNR criterion related to $\mathbf{s}_1$ (top graph) and $\mathbf{s}_2$ (bottom graph) versus k, the number of negative weights. The uniqueness condition holds for $\mathbf{s}_1$ when $k \in [53, 68]$ (the dashed lines depict this interval) and in this interval, the WSL0 algorithm has estimated $\mathbf{s}_1$ properly. Note that the WSL0 algorithm never chooses the sparsest solution $\mathbf{s}_2$. Also note that when $k \notin [53, 68]$, the algorithm may or may not converge to $\mathbf{s}_1$, since the uniqueness condition does not hold anymore.

Representation based Classification) which was introduced in [4]. The SRC method is based on solving a P_0 problem. In order to have better performance [9] suggests to solve the following $P_{1,\mathbf{w}}$ problem (an ℓ_1 relaxation of $P_{0,\mathbf{w}}$)

$$\min_{\mathbf{s}} \sum_{i=1}^{m} w_i |s_i| \quad s.t. \quad \mathbf{A}\mathbf{s} = \mathbf{y}, \tag{9}$$

where $\mathbf{A} = [\mathbf{a}_1|\mathbf{a}_2|\cdots|\mathbf{a}_m]$ is the matrix containing all the training vectors, $\mathbf{y}$ is the input vector and the weight vector $\mathbf{w}$ defined as

$$\mathbf{w} = [w_1, w_2, \cdots, w_m], \quad w_i = \|\mathbf{a}_i - y\|_2^2. \tag{10}$$

This is so called WSRC algorithm. [10] solved a $P_{0,\mathbf{w}}$ problem directly (using WSL0 algorithm) with the same (positive) weight vector as (10) and achieved better results. Since [9] uses the term WSRC for classification based on weighted ℓ_1 norm minimization, we use the term WSRC0 for classification based on weighted ℓ_0 norm minimization.

In this section, we propose a new weight vector (which has some negative components) and we solve the $P_{0,\mathbf{w}}$ problem with this new weight vector in order to show its superiority with respect to the SRC and WSRC0 methods with positive weights. To achieve a better performance, our idea is to use negative weights for the atoms that are very close to the input vector. We propose the following weight vector

$$w_k = \|\mathbf{y} - \mathbf{a}_k\|_2^2 - t. \tag{11}$$

It is seen that if the distance between an atom and the input vector is lower than $\sqrt{t}$, its corresponding weight would be negative. So the presence of that atom would be favorable to the cost function (that is, weighted ℓ_0 norm of the decomposition). We shall denote the WSRC0 method with our proposed weight vector with WSRC0N. As we will show in our examples, with a good choice of the threshold value t, the performance of WSRC0N is better than WSRC0.

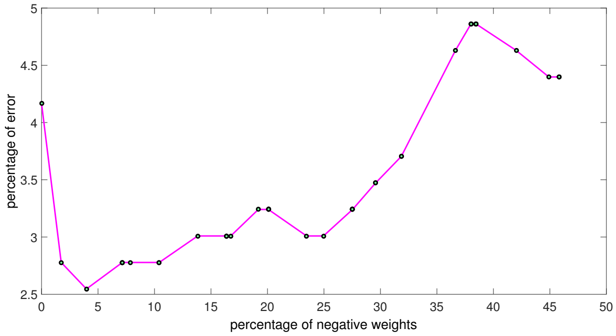

Fig. 2. Error percentage versus the percentage of present negative weights. This figure is obtained for different thresholds used in classifying the monk2 dataset.

TABLE I
COMPARISON OF THE PERCENTAGE OF ERROR OBTAINED BY RUNNING EACH OF THESE THREE ALGORITHMS: SRC, WSRC0 WITH POSITIVE WEIGHTS (WSRC0P) AND WSRC0 WITH OUR PROPOSED WEIGHT VECTOR (WSRC0N)

benchmark	SRC	WSRC0P	WSRC0N
zoo	5.9	4.9	4
monk2	11.8	4.2	2.8
heart	25.9	19.6	18.9

To verify the performance of WSRC0 with this new weight vector, a series of experiments were carried out. This experiments are conducted on the KEEL benchmark datasets [11] and YALE facial recognition dataset [12]. For performing these experiments, the leave-one-out method [13] was used. In this method, one vector is considered to be the test vector and the remaining vectors are used as training data. Then the classification is performed. This step is repeated until the number of repeats reaches the number of sets.

To perform these classifications we needed to see how to choose the value of t so that the algorithm would deliver an acceptable performance. Therefore we have performed the classification on monk2 dataset for different values of t hence changing the percent of negative weights. As it can be seen from Fig 2 the algorithm performs best when around 2-15 percent of the weights are negative. As depicted in Fig 2, small changes in the percentage of negative weights (in the 2-15 interval) results to small changes in the percentage of error. Therefore we can conclude that the WSRC0 method is robust with respect to the percentage of negative weights.

Table 1 shows the results for the 3 different benchmarks of KEEL dataset: zoo, heart and monk2. In each experiment, the value of t is chosen such that around 10 percent of the weights become negative. In the experiment results (Table1 and Fig 3) the WSRC0 algorithm is denoted WSRC0P with only positive weights and WSRC0N with both positive and negative weights. As you see, the performance of WSRC0 with this new weight vector surpasses the performance of SRC and WSRC0 methods.

The YALE facial recognition dataset encompasses 15

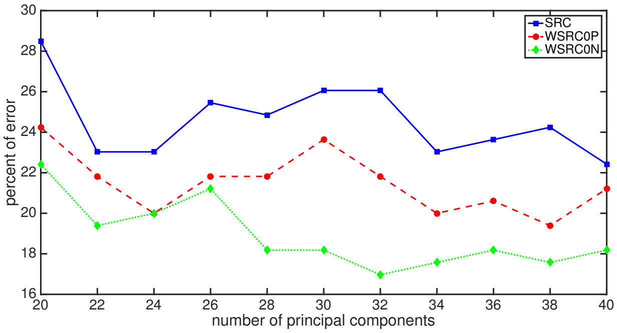

Fig. 3. Error percentage comparison of SRC, WSRC0 with positive weights (WSRC0P) and WSRC0 with our proposed weight vector (WSRC0N) algorithms for the YALE facial recognition dataset.

classes and each class has 11 gray images. We reduced the dimensions of the images before classification. In order to do that, we use Principal Component Analysis (PCA) method [14]. Then we choose the threshold t such that around 10 percent of weights become negative.

We plot the percentage of error versus the number of principal components for SRC, WSRC0 with positive weights and WSRC0 with our proposed weight vector. Again, WSRC0 with our proposed vector surpasses SRC and WSRC0 with positive weights.

IV. Conclusion

In this paper, a generalization of weighted sparse signal decomposition to negative weights was studied, and a condition and tight bound for the uniqueness of negative weighted ℓ_0 norm minimization was presented. Then, the uniqueness condition was evaluated numerically. Finally by proposing a weight vector with some negative weights, we showed its superiority in the classification problem comparing to SRC and WSRC0 algorithms.

V. Acknowledgement

This work has been supported by the Center for International Scientific Studies and Collaboration (CISSC), and also by European project ERC-2012-AdG-320684-CHESS.

Appendix

Lemma 1: If $\mathbf{A}$ is an n by m full rank matrix and $\mathbf{w}$ is a weight vector which has more than n negative weights, then the problem $P_{0,\mathbf{w}}$ has at least two different solutions.

Proof: Suppose that we have t negative weights and $t > n$. Without any loss of generality we assume that these are $[w_1, \ldots, w_t]$. We need only to find two solutions $\mathbf{s}_1$ and $\mathbf{s}_2$ of $\mathbf{As} = \mathbf{b}$ such that none of the first t entries of these two solutions are equal to zero, and all of their other entries are equal to zero (obviously, such vectors are both solutions of $P_{0,\mathbf{w}}$, because the minimum value of $\|.\|_{0,\mathbf{w}}$ is $\sum_{i=1}^{t} w_i$ which is equal to $\|\mathbf{s}_1\|_{0,\mathbf{w}}$ and $\|\mathbf{s}_2\|_{0,\mathbf{w}}$). Assume that there is at most one vector which satisfies the above condition. We define

$$\mathcal{A} = \{\mathbf{s} \in \mathbb{R}^m | \mathbf{As} = \mathbf{b}, s_i = 0\, \forall i \in \{t+1, \ldots, m\}\},$$

$$\mathcal{A}_j = \{\mathbf{s} \in \mathcal{A} | s_j = 0\} \qquad j = 1, 2, \ldots, t.$$

It is obvious that $\mathcal{A}_j$'s ($j = 1, 2, \ldots, t$) are all proper affine subspaces of the affine space $\mathcal{A}$. We consider two cases separately:

1) $\mathrm{Card}(\mathcal{A} \setminus \bigcup_{j=0}^{t} \mathcal{A}_j) \leq 1$, where $\mathrm{Card}(\cdot)$ stands for the cardinality of a set. If $\mathrm{Card}(\mathcal{A} \setminus \bigcup_{j=0}^{t} \mathcal{A}_j) = 1$ then there exists a vector $\mathbf{s}$ such that $\mathcal{A} \setminus \bigcup_{j=0}^{t} \mathcal{A}_j = \{\mathbf{s}\}$. Therefore $\mathcal{A} = \{\mathbf{s}\} \cup (\bigcup_{j=0}^{t} \mathcal{A}_j)$ (Note that $\{\mathbf{s}\}$ is also a proper affine subspace of $\mathcal{A}$). Now if $\mathrm{Card}(\mathcal{A} \setminus \bigcup_{j=0}^{t} \mathcal{A}_j) = 0$, then $\mathcal{A} \setminus \bigcup_{j=0}^{t} \mathcal{A}_j = \emptyset$. Therefore we have $\mathcal{A} = \bigcup_{j=0}^{t} \mathcal{A}_j$. So in both cases we have written $\mathcal{A}$ as a finite union of proper affine subspaces of itself which is a contradiction [15]. So this case can never happen.
2) $\mathrm{Card}(\mathcal{A} \setminus \bigcup_{j=0}^{t} \mathcal{A}_j) \geq 2$. That means there exist two different vectors $\mathbf{s}_1$, $\mathbf{s}_2$ in $\mathcal{A} \setminus \bigcup_{j=0}^{t} \mathcal{A}_j$. Therefore by the definition of $\mathcal{A}$ and $\mathcal{A}_j$'s, $\mathbf{s}_1$ and $\mathbf{s}_2$ are nonzero in their first t components and zero everywhere else. Therefore $\mathbf{s}_1$, $\mathbf{s}_2$ are two different solutions of $P_{0,\mathbf{w}}$.

■

Proof of Theorem 1: Arguing by contradiction, assume that there exist two vectors $\mathbf{s}_1$ and $\mathbf{s}_2$ that satisfy $\|\mathbf{s}_1\|_{0,\mathbf{w}} = \|\mathbf{s}_2\|_{0,\mathbf{w}} < \frac{S_{\mathbf{w}}(\mathrm{Spark}(\mathbf{A}))}{2} + |w_{min}|(\frac{\mathrm{Spark}(\mathbf{A})}{2} - n + 1)$. Without loss of generality, assume that $\mathbf{w}$ is sorted such that

$$w_1 \leq w_2 \leq \ldots \leq w_t < 0 \leq w_{t+1} \leq \ldots \leq w_m.$$

Assuming H1 and H2, we have $t < n$ and $\|\mathbf{s}_1\|_0, \|\mathbf{s}_2\|_0 \leq n-1$. Now we define a new weight vector (called $\mathbf{w}'$) as $w_i{}' = w_i + |w_1|$ for $i = 1, 2, \ldots,\ m$. It is obvious that $\mathbf{w}'$ is a non-negative weight vector. $\|.\|_{0,\mathbf{w}}$ and $\|.\|_{0,\mathbf{w}'}$ are related as

$$\begin{aligned} \|\mathbf{s}\|_{0,\mathbf{w}'} = \sum_{i=1}^{m} |x_i|_0 w'{}_i &= \sum_{i=1}^{m} |x_i|_0 (w_i + |w_1|) \\ &= \|\mathbf{s}\|_{0,\mathbf{w}} + \|\mathbf{s}\|_0 |w_1|. \end{aligned}$$

Therefore,

$$\|\mathbf{s}_1\|_{0,\mathbf{w}'} < \frac{S_{\mathbf{w}}(\mathrm{Spark}(\mathbf{A}))}{2} + |w_1|(\frac{\mathrm{Spark}(\mathbf{A})}{2} - n + 1 + \|\mathbf{s}_1\|_0).$$

So, it can be seen that

$$\|\mathbf{s}_1\|_{0,\mathbf{w}'} < \frac{S_{\mathbf{w}}(\mathrm{Spark}(\mathbf{A})) + |w_1|\mathrm{Spark}(\mathbf{A})}{2} + |w_1|(\|\mathbf{s}_1\|_0 - (n-1)).$$

Using H2, we must have $|w_1|(\|\mathbf{s}_1\|_0 - (n-1)) \leq 0$ which yields

$$\|\mathbf{s}_1\|_{0,\mathbf{w}'} < \frac{S_{\mathbf{w}}(\mathrm{Spark}(\mathbf{A})) + |w_1|\mathrm{Spark}(\mathbf{A})}{2}.$$

Using the definition of $\mathbf{w}'$, the previous inequality becomes:

$$\|\mathbf{s}_1\|_{0,\mathbf{w}'} < \frac{S_{\mathbf{w}'}(\mathrm{Spark}(\mathbf{A}))}{2}. \tag{12}$$

Similarly

$$\|\mathbf{s}_2\|_{0,\mathbf{w}'} < \frac{S_{\mathbf{w}'}\ (\mathrm{Spark}\ (\mathbf{A}))}{2} \tag{13}$$

Then according to positivity of $\mathbf{w}'$ and triangle inequality for ℓ_0 norm (note that although the so called ℓ_0 "norm" is not a mathematical "norm", it satisfies the triangle inequality)

$$\|\mathbf{s}_1 - \mathbf{s}_2\|_{0,\mathbf{w}'} = \sum_{i=1}^{m} |\mathbf{s}_{1i} - \mathbf{s}_{2i}|_0 w'_i \leq \sum_{i=1}^{m} (|\mathbf{s}_{1i}| + |\mathbf{s}_{2i}|_0) w'_i.$$

Note that $\sum_{i=1}^{m}(|\mathbf{s}_{1i}| + |\mathbf{s}_{2i}|_0)w'_i = \|\mathbf{s}_1\|_{0,\mathbf{w}'} + \|\mathbf{s}_2\|_{0,\mathbf{w}'}$. Using (12) and (13),

$$\|\mathbf{s}_1 - \mathbf{s}_2\|_{0,\mathbf{w}'} < S_{\mathbf{w}'}\left(\text{Spark}\left(\mathbf{A}\right)\right).$$

Therefore $\mathbf{s}_1 - \mathbf{s}_2$ has at most Spark$(\mathbf{A}) - 1$ nonzero components because its $\ell_{0,\mathbf{w}}$ norm is less than the sum of Spark$(\mathbf{A})$ number of smallest weighs. So $\|\mathbf{s}_1 - \mathbf{s}_2\|_0 <$ Spark $(\mathbf{A})$. But according to the fact that $\mathbf{s}_1$ and $\mathbf{s}_2$ are the solutions of the linear system $\mathbf{As} = \mathbf{b}$, we have $\mathbf{As}_1 = \mathbf{As}_2 = \mathbf{b}$. Therefore $\mathbf{A}(\mathbf{s}_1 - \mathbf{s}_2) = 0$. So we have found $\|\mathbf{s}_1 - \mathbf{s}_2\|_0$ columns of $\mathbf{A}$ that are linearly independent and this contradicts the definition of Spark$(\mathbf{A})$.

■

References

[1] M. Elad, *Sparse and Redundant Representations*. Springer New York, 2010.

[2] R. Gribonval and S. Lesage, "A survey of sparse component analysis for blind source separation: principles, perspectives, and new challenges," in *Proceedings of ESANN'06*, April 2006, pp. 323–330.

[3] D. L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, April 2006.

[4] K. Huang and S. Aviyente, "Sparse representation for signal classification," in *Advances in neural information processing systems*, 2006, pp. 609–616.

[5] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. on Signal Proc.*, vol. 41, no. 12, pp. 3397–3415, 1993. [Online]. Available: citeseer.ist.psu.edu/mallat93matching.html

[6] M. Babaie-Zadeh, B. Mehrdad, and G. B. Giannakis, "Weighted sparse signal decomposition," in *Proceedings of* ICASSP2012, Kyoto, Japan, March 25-30 2012, pp. 3425–3428.

[7] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ^1 minimization," *Proc. Nat. Aca. Sci.*, vol. 100, no. 5, pp. 2197–2202, March 2003.

[8] H. Mohimani, M. Babaie-Zadeh, and C. Jutten, "A fast approach for overcomplete sparse decomposition based on smoothed ℓ^0 norm," *IEEE Transactions on Signal Processing*, vol. 57, no. 1, pp. 289–301, January 2009.

[9] C.-Y. Lu, H. Min, J. Gui, L. Zhu, and Y.-K. Lei, "Face recognition via weighted sparse representation," *Journal of Visual Communication and Image Representation*, vol. 24, no. 2, pp. 111–116, 2013.

[10] M. Nazari, "Sparse representation-based classification and application to image and speech processing," Master's thesis, Sharif University of Technology, Iran, 2015, in persian.

[11] J. Alcalá, A. Fernández, J. Luengo, J. Derrac, S. García, L. Sánchez, and F. Herrera, "Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework," *Journal of Multiple-Valued Logic and Soft Computing*, vol. 17, no. 2-3, pp. 255–287, 2010.

[12] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 19, no. 7, pp. 711–720, 1997.

[13] S. Arlot, A. Celisse *et al.*, "A survey of cross-validation procedures for model selection," *Statistics surveys*, vol. 4, pp. 40–79, 2010.

[14] I. Jolliffe, *Principal component analysis*. Wiley Online Library, 2002.

[15] P. L. Clark, "Covering numbers in linear algebra," *The American Mathematical Monthly*, vol. 119, no. 1, pp. 65–67, 2012.