WEIGHTED SPARSE SIGNAL DECOMPOSITION

Massoud Babaie-Zadeh[1] *Behzad Mehrdad*[2] *Georgios B. Giannakis*[3]

[1]Electrical Engineering Department, Sharif University of Technology, Tehran, IRAN.
[2]Courant Institute of Mathematical Sciences, New York University, USA.
[3]Digital Technology Center (DTC), University of Minnesota, USA.
mbzadeh@yahoo.com, mehrdad@cims.nyu.edu, georgios@umn.edu

ABSTRACT

Standard sparse decomposition (with applications in many different areas including compressive sampling) amounts to finding the minimum ℓ^0-norm solution of an underdetermined system of linear equations. In this decomposition, all atoms are treated 'uniformly' for being included or not in the decomposition. However, one may wish to weigh more or less certain atoms, or, assign higher costs to some other atoms to be included in the decomposition. This can happen for example when there is prior information available on each atom. This motivates generalizing the notion of minimal ℓ^0-norm solution to that of minimal *weighted* ℓ^0-norm solution. On the other hand, relaxing weighted ℓ^0-norm via the weighted ℓ^1-norm is challenging. This paper deals with minimal weighted ℓ^0-norm solutions of underdetermined linear systems, provides conditions for their uniqueness, and develops an algorithm for their estimation.

***Index Terms*—** Weighted sparse decomposition, Sparse decomposition, Compressive sampling, weighted compressive sampling.

1. INTRODUCTION

Sparse solutions of underdetermined systems of linear equations attract growing interest for their potential application in areas as diverse as compressive sampling (CS) [1, 2], classification and recognition [3], underdetermined sparse component analysis (SCA) for source separation [4], real-field coding [5], and denoising [6], to name a few. Let $\mathbf{A} \triangleq [\mathbf{a}_1, \ldots, \mathbf{a}_m]$ be an $n \times m$ matrix with $m > n$, where $\{\mathbf{a}_i\}_{i=1}^m$ denote its columns; $\mathbf{x}$ be an $m \times 1$ vector, and consider the underdetermined linear system

$$\mathbf{A}\mathbf{s} = \mathbf{x} \tag{1}$$

with $\mathbf{A}$ assumed to have full row rank, so that (1) has no redundant or contradictory equations. Being underdetermined, this system has infinitely many solutions, but the sparse solution is a solution $\mathbf{s} \triangleq (s_1, \ldots, s_m)^T$ having as few non-zero entries as possible. In other words, it is the solution of the problem

$$(P_0)\colon \quad \min_{\mathbf{s}} \|\mathbf{s}\|_0 \triangleq \sum_{i=1}^m |s_i|_0 \ \text{ subject to } \mathbf{A}\mathbf{s} = \mathbf{x} \tag{2}$$

where the ℓ^0-norm $\|\cdot\|_0$ stands for the number of non-zero entries of its vector argument, and $|\cdot|_0$ stands for the indicator function:

$$|x|_0 \triangleq \begin{cases} 0 & \text{if } x = 0 \\ 1 & \text{otherwise} \end{cases} . \tag{3}$$

This work has been partially funded by Iran Telecom Research Center (ITRC). Part of this work was done when the first author was in sabbatical at the University of Minnesota.

In the signal (or atomic) decomposition parlance [7], $\mathbf{x}$ is a vector expressed as a linear superposition of atoms $\{\mathbf{a}_i\}_{i=1}^m$ comprising the so-termed dictionary $\mathbf{A}$ over which the signal $\mathbf{x}$ is to be decomposed. When $m > n$, the decomposition is not unique, but by a sparse decomposition, one means a decomposition which uses a minimum number of atoms to decompose $\mathbf{x}$.

In the standard sparse recovery problem (2), all entries of $\mathbf{s}$ are treated "uniformly" as far as being zero or non-zero, and the goal is only to minimize the total number of non-zero entries of $\mathbf{s}$. However, in certain applications, one may opt to have different entries of $\mathbf{s}$ affect differently the cost of being (non)zero. For example, in atomic decomposition, one may wish to give a priority to some of the atoms, or assign a higher cost to some other atoms for being included in the decomposition. This can happen for example when there is prior information available on the probabilities of different entries of $\mathbf{s}$ being non-zero, which prompts looking for the sparse solution of (1) by assigning different weights across entries of $\mathbf{s}$. Analytically, the weighted counterpart of (2) is

$$(P_{0,\mathbf{w}})\colon \quad \min_{\mathbf{s}} \|\mathbf{s}\|_{0,\mathbf{w}} \triangleq \sum_{i=1}^m w_i |s_i|_0 \ \text{ s.t. } \mathbf{A}\mathbf{s} = \mathbf{x} \tag{4}$$

where $w_i \geq 0$, $i = 1, \ldots, m$, captures the "cost" of having the i-th entry of $\mathbf{s}$ being nonzero. Note that in (4), $\mathbf{A}$, $\mathbf{x}$ and $\mathbf{w} \triangleq (w_1, \ldots, w_m)^T$ are given, and the minimization is carried over $\mathbf{s}$. Similar to the ℓ^0-norm $\|\mathbf{s}\|_0$, we call $\|\mathbf{s}\|_{0,\mathbf{w}}$ the *weighted* ℓ^0*-norm* of $\mathbf{s}$ (with weight vector $\mathbf{w}$), although, neither $\|\mathbf{s}\|_0$ nor $\|\mathbf{s}\|_{0,\mathbf{w}}$ are mathematical 'norms' (they do not satisfy the scaling property).

A special case of this problem, where w_i's are binary taking values zero or one, has been considered under the term sparse recovery from *partially known support* [8, 9]. However, the main objective behind those methods is to find the sparse solution of (1), meaning to solve P_0, when part of its support is known a priori, whereas in this paper, the problem under study is $P_{0,\mathbf{w}}$, whose solution is generally different from that of P_0 (see Section 2.2).

The paper is organized as follows. The next section motivates the problem by expressing $P_{0,\mathbf{w}}$ as a maximum a posteriori (MAP) estimation of the sparse solution when priors on the activity of each atom are available. Section 3 deals with the uniqueness issue of solving $P_{0,\mathbf{w}}$, while Section 4 presents an algorithm for approximating the solution of $P_{0,\mathbf{w}}$. Finally, simulations are presented in Section 5.

2. MOTIVATION AND CHALLENGES

2.1. MAP sparse recovery with priors on activity

Suppose that in (1), $\mathbf{s}$ is modeled as a vector with statistically independent entries, and the i-th entry is active (non-zero) with probabil-

ity p_i and inactive (zero) with probability $1-p_i$ (sparsity translates to small p_i's). Suppose that p_i's are known a priori, and the goal is to estimate $\mathbf{s}$. To this end, [10] assumes that $\mathbf{s}$ is to be estimated by a weighted ℓ^1-norm minimization, and *within this class of estimators*, the weights are selected to maximize the probability of recovery. In contrast, it is shown here that a MAP-like estimate of $\mathbf{s}$ requires in fact solving a problem of the form $P_{0,\mathbf{w}}$ with certain weights.

To see this, consider expressing the i-th entry of $\mathbf{s}$ as $s_i = b_i r_i$, where $b_i \in \{0,1\}$ is a binary variable denoting the 'activity' of s_i (that is $b_i = 1$ if s_i is active, and $b_i = 0$ otherwise), and r_i is the magnitude of s_i assumed to be independent of b_i. Let also $\mathbf{b} \triangleq (b_1, \ldots, b_m)^T$ and $\mathbf{r} \triangleq (r_1, \ldots, r_m)^T$. Under these notational conventions, the following result is proved in the Appendix.

Theorem 1. *Let $\hat{\mathbf{s}} \triangleq \hat{\mathbf{b}}_{MAP} \circ \hat{\mathbf{r}}_{MAP}$, where $\circ$ stands for the entry-wise (Hadamard) product, and $(\hat{\mathbf{b}}_{MAP}, \hat{\mathbf{r}}_{MAP}) \triangleq \operatorname{argmax}_{\mathbf{b},\mathbf{r}}\, p(\mathbf{b},\mathbf{r}|\mathbf{x})$. Then $\hat{\mathbf{s}}$ is the solution of $P_{0,\mathbf{w}}$ with weights $w_i = \ln[(1-p_i)/p_i]$.*

2.2. Challenges

A well-known approach to solving P_0 is to relax it by replacing the ℓ^0 by the ℓ^1 norm [11]. However, employing the weighted ℓ^1-norm $\sum_{i=1}^m w_i|s_i|$ for solving $P_{0,\mathbf{w}}$ is not as motivated as in the unweighted case. This is because $\sum_{i=1}^m w_i|s_i| = \sum_{i=1}^m |w_i s_i|$ is a relaxation of $\sum_{i=1}^m |w_i s_i|_0 = \sum_{i=1}^m |s_i|_0$ (assuming that all weights are strictly positive). In other words, both ℓ^1 and weighted ℓ^1 can be seen as relaxations of the *unweighted* ℓ^0-norm. In fact, [12] asserts that since ℓ^1 and weighted ℓ^1 have generally different solutions, one may view the weights in weighted ℓ^1 as free parameters in the convex relaxation of the *unweighted* P_0 problem, whose values can be set judiciously to obtain an improved solution of P_0. Hence, if one adopts weighted ℓ^1 as a relaxation of the 'weighted' problem $P_{0,\mathbf{w}}$, it is not known whether the final solution offers a better estimate of the solution of P_0 or $P_{0,\mathbf{w}}$. We will shed some light on this challenging issue via simulations in Section 5.

Note also that P_0 and $P_{0,\mathbf{w}}$ are two different problems with generally different solutions. For example, for the system

$$\begin{bmatrix} 1 & 1 & -1 & 2 & 1 \\ -1 & 1 & 1 & -1 & 1 \\ 1 & -1 & 1 & 3 & 1 \end{bmatrix} \mathbf{s} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \tag{5}$$

the (unique) solution of P_0 is $\mathbf{s} = (0,0,0,0,1)^T$, but the (again unique) soltion of $P_{0,\mathbf{w}}$ under the weight vector $\mathbf{w} = (1,1,1,2,4)^T$ is $\mathbf{s} = (1,1,1,0,0)^T$. Therefore, when used in applications, *one has to decide first which problem, P_0 or $P_{0,\mathbf{w}}$, is more appropriate for the application under consideration*.

3. UNIQUENESS

An important property of P_0 is that its solution can be unique. More precisely, let *spark*$(\mathbf{A})$ denote the minimum number of columns of $\mathbf{A}$ that are linearly dependent [13]. It is known that if P_0 has a solution $\mathbf{s}_0$ satisfying $\|\mathbf{s}_0\|_0 < \frac{1}{2}$*spark*$(\mathbf{A})$, then it would be its unique solution [14, 13].

This uniqueness theorem can be generalized to problem $P_{0,\mathbf{w}}$ as follows (see the Appendix for the proof).

Theorem 2. *If $\mathbf{A}\mathbf{s} = \mathbf{x}$ has a solution $\mathbf{s}_0$ for which*

$$\|\mathbf{s}_0\|_{0,\mathbf{w}} < \frac{\mathcal{S}_{\mathbf{w}}\big(\textit{spark}(\mathbf{A})\big)}{2} \tag{6}$$

where $\mathcal{S}_{\mathbf{w}}(k)$ stands for the sum of the k smallest weights, then $\mathbf{s}_0$ is the unique solution of $P_{0,\mathbf{w}}$.

As a sanity check, note that the unweighted problem P_0 corresponds to $w_i = 1, \forall i$, for which $\mathcal{S}_{\mathbf{w}}(k) = k$, and hence the bound in (6) becomes the known $\frac{1}{2}$*spark*$(\mathbf{A})$ bound for P_0.

Furthermore, the bound in (6) is *tight* in the sense that it is impossible to have a tighter bound that works for all linear systems. To show this, it is sufficient to provide an example for which the bound (6) is not marginally satisfied and $P_{0,\mathbf{w}}$ has two different solutions. An example of this kind is offered by the system

$$\begin{bmatrix} 1 & 1 & -1 & 2 & 0 \\ 1 & -1 & 1 & 1 & 1 \\ -1 & 1 & 1 & 3 & 1 \\ 1 & 1 & 1 & 2 & -2 \end{bmatrix} \mathbf{s} = \begin{bmatrix} 2 \\ 0 \\ 2 \\ 4 \end{bmatrix} \tag{7}$$

and the weight vector $\mathbf{w} = (0.5, 1, 1, 1, 1.5)^T$. It can be seen that here *spark*$(\mathbf{A}) = 5$, and $P_{0,\mathbf{w}}$ has two solutions $\mathbf{s}_0 = (0,0,0,1,-1)^T$ and $\mathbf{s}_1 = (1,2,1,0,0)^T$, whose weighted ℓ^0-norms are both equal to $\frac{1}{2}\mathcal{S}_{\mathbf{w}}(5) = 2.5$.

4. A WEIGHTED SPARSE RECOVERY ALGORITHM

Being a generalization of P_0, problem $P_{0,\mathbf{w}}$ is NP-hard to solve. Moreover, we pointed out in Section 2.2 that the use of (weighted) ℓ^1-norm approach for solving $P_{0,\mathbf{w}}$ can be challenging. In this section, we present an approach for approximating the solution of $P_{0,\mathbf{w}}$ when the given weights are strictly positive ($\forall i, w_i > 0$). This approach is the weighted counterpart of the smoothed ℓ^0 (SL0) solver of P_0 [15], so, we call it weighted SL0 (WSL0).

The main idea of SL0 is to use a smooth approximation of the ℓ^0-norm. Specifically, let $f_\sigma(\cdot)$ be a continuous function satisfying $\lim_{\sigma\to 0} f_\sigma(s) = 1 - |s|_0$. An example of such f_σ's is $f_\sigma(s) \triangleq \exp(-s^2/2\sigma^2)$. Then $|s|_0$ can be approximated as $|s|_0 \approx 1 - f_\sigma(s)$, where σ determines the accuracy of the approximation: the smaller σ, the better approximation, and the larger σ, the smoother approximation. So, the ℓ^0-norm can be approximated by $\|\mathbf{s}\|_0 \approx m - \sum_{i=1}^m f_\sigma(s_i)$, and hence SL0 aims to maximize $F_\sigma(\mathbf{s}) \triangleq \sum_{i=1}^m f_\sigma(s_i)$ (subject to $\mathbf{A}\mathbf{s} = \mathbf{x}$) for a small σ. The major challenge here is that F_σ is not concave and especially for small σ's, it has many local maxima, and hence its maximization is not easy. To cope with this challenge, SL0 adopts a graduated non-convexity (GNC[1]) [16] approach for maximizing it. GNC starts with a very large σ, for which F_σ is nearly concave and its maximization is easy, and then gradually decreases σ, and for each σ it *starts the search for the maximizer* of F_σ from the maximizer for the previous (larger) σ. Using such an annealing process, it is expected (but not mathematically guaranteed) to escape from being trapped into local maxima.

The GNC-based SL0 solver is directly applicable to the weighted minimization, too. In this case, $\|\mathbf{s}\|_{0,\mathbf{w}} \approx \sum_{i=1}^m w_i - \sum_{i=1}^m w_i f_\sigma(s_i)$, and so, the idea behind WSL0 is to

$$\max_{\mathbf{s}}\ F_\sigma^{\mathbf{w}}(\mathbf{s}) \triangleq \sum_{i=1}^m w_i f_\sigma(s_i) \ \text{ s.t. } \mathbf{A}\mathbf{s} = \mathbf{x} \tag{8}$$

for a very small σ. The use of GNC to cope with local maxima is exactly as in SL0. By opting to choose a gradient-projection (GP) iteration for a fixed σ, the final algorithm would be as shown in Fig. 1. There are a few more points to be mentioned about this algorithm:

Step size: For smaller σ's, $F_\sigma^{\mathbf{w}}$ is more fluctuating and we should use a smaller step-size in the gradient ascent loop for its

[1] GNC can be seen as a deterministic version of simulated annealing. In this viewpoint, the parameter σ is called 'temperature'.

- Initialization:
  1. Set $\hat{\mathbf{s}}_0$ equal the minimum weighted ℓ^2-norm solution of $\mathbf{As} = \mathbf{x}$, given in Theorem 3.
  2. Choose a suitable decreasing sequence for σ: $[\sigma_1 \ldots \sigma_J]$.
- For $j = 1, \ldots, J$:
  1. Let $\sigma = \sigma_j$.
  2. Solve (8) using L Gradient-Projection iterations:
     - Initialization: $\mathbf{s} = \hat{\mathbf{s}}_{j-1}$.
     - For $\ell = 1, 2, \ldots, L$
       (a) Let $\mathbf{s} \leftarrow \mathbf{s} + (\mu\sigma^2)\nabla F_\sigma^{\mathbf{w}}(\mathbf{s})$.
       (b) Project $\mathbf{s}$ back onto the feasible set $\{\mathbf{s}|\mathbf{As} = \mathbf{x}\}$:
       $$\mathbf{s} \leftarrow \mathbf{s} - \mathbf{A}^\dagger(\mathbf{As} - \mathbf{x}).$$
  3. Set $\hat{\mathbf{s}}_j = \mathbf{s}$.
- Final answer is $\hat{\mathbf{s}} = \hat{\mathbf{s}}_J$.

Fig. 1. WSL0 algorithm. $\mathbf{A}^\dagger$ stands for the Moore-Penrose pseudo inverse of $\mathbf{A}$ (*i.e.* $\mathbf{A}^\dagger \triangleq \mathbf{A}^T(\mathbf{AA}^T)^{-1}$).

maximization. For reasons detailed in [15], a good choice is to decrease the step-size proportional to σ^2. So, in Fig. 1, the step-size $\mu\sigma^2$ is used.

Initialization: Bearing in mind the GNC idea, a good initialization is to use the maximizer of $F_\sigma^{\mathbf{w}}(\mathbf{s})$ for $\sigma \to \infty$. This is interestingly the minimum weighted ℓ^2-norm solution as asserted next.

Theorem 3. *Assume that* $\forall i, w_i > 0$, *and let* $f_\sigma(s) \triangleq \exp(-s^2/2\sigma^2)$. *Then, as* $\sigma \to \infty$, *the solution of (8) converges to the minimizer of* $\sum_{i=1}^m w_i s_i^2$ *subject to* $\mathbf{As} = \mathbf{x}$, *which is given by* $\mathbf{W}^{-1}\mathbf{A}^T(\mathbf{AW}^{-1}\mathbf{A}^T)^{-1}\mathbf{x}$, *where* $\mathbf{W} \triangleq \mathrm{diag}(w_1, \ldots, w_m)$.

This result holds also for a large class of smoothing functions f_σ; see also [15, Theorem 2]. Due to lack of space, the general theorem and its proof is delegated to the journal version of this work. For the Gaussian smoothing function however, Theorem 3 can be justified by simply noting that for very large σ's, $\exp(-s_i^2/2\sigma^2) \approx 1 - s_i^2/2\sigma^2$.

5. SIMULATIONS

We conducted a simulation to compare weighted ℓ^1 and WSL0. In this simulation, we randomly created a system $\mathbf{As} = \mathbf{x}$ of $n = 40$ equations in $m = 100$ unknowns having two different sparse solutions $\mathbf{s}_1$ and $\mathbf{s}_2$ with $\|\mathbf{s}_1\|_0 = 6$ and $\|\mathbf{s}_2\|_0 = 13$[2]. Then, we chose the weighting vector $\mathbf{w}$ such that $\|\mathbf{s}_1\|_{0,\mathbf{w}} = \alpha$ and $\|\mathbf{s}_2\|_{0,\mathbf{w}} = 1 - \alpha$, where $0 < \alpha < 1$. So, for $\alpha < 0.5$, $\mathbf{s}_1$ is the solution of both P_0 and $P_{0,\mathbf{w}}$ problems, whereas for $\alpha > 0.5$ the solution of P_0 is $\mathbf{s}_1$, and the solution of $P_{0,\mathbf{w}}$ is $\mathbf{s}_2$. Then three algorithms, namely ℓ^1-norm minimization, weighted ℓ^1-norm minimization, and WSL0 were run on the system $\mathbf{As} = \mathbf{x}$. The output ($\hat{\mathbf{s}}$) was then compared to $\mathbf{s}_1$ and $\mathbf{s}_2$ on the basis of the signal-to-noise-ratio (SNR) defined as $\mathrm{SNR}_i = 10\log_{10}(\|\mathbf{s}_i\|_2^2/\|\hat{\mathbf{s}} - \mathbf{s}_i\|_2^2)$, $i = 1, 2$. For $\mathrm{SNR}_1 > 10$dB, it was declared that the algorithm has estimated $\mathbf{s}_1$, while for $\mathrm{SNR}_2 > 10$dB it was declared that the algorithm has estimated $\mathbf{s}_2$; otherwise, it was declared that the algorithm has failed to estimate either one of these solutions. This experiment was repeated 1,000 times with different randomly generated systems. Fig. 2 depicts the percentages of estimating $\mathbf{s}_1$, $\mathbf{s}_2$ and failure rates versus α.

[2]This was done as follows: First, $\mathbf{A}_1$ of size $n = 40$ by $m/2 = 50$ was randomly created with entries drawn independently from a standardized Gaussian distribution. Then $\mathbf{s}_1'$ of length $m/2 = 50$ was created to have only 6 non-zero entries (whose locations and magnitudes were chosen randomly), and $\mathbf{x} \triangleq \mathbf{A}_1\mathbf{s}_1'$ was calculated. Similarly, $\mathbf{A}_2$ of size $n = 40$ by $m/2 - 1 = 49$, and $\mathbf{s}_2'$ of length $m/2 - 1 = 49$ with $13 - 1 = 12$ non-zero entries were created. Matrix $\mathbf{A}$ and vectors $\mathbf{s}_1$ and $\mathbf{s}_2$ were then formed as $\mathbf{A} \triangleq [\mathbf{A}_1, \mathbf{A}_2, \mathbf{x} - \mathbf{A}_2\mathbf{s}_2']$, $\mathbf{s}_1 \triangleq ({\mathbf{s}_1'}^T, \mathbf{0}_{1\times 50})^T$, and $\mathbf{s}_2 \triangleq (\mathbf{0}_{1\times 50}, {\mathbf{s}_2'}^T, 1)^T$. Finally, the columns of $\mathbf{A}$ were normalized to have unit ℓ^2-norm, and the entries of $\mathbf{s}_1$ and $\mathbf{s}_2$ were scaled accordingly.

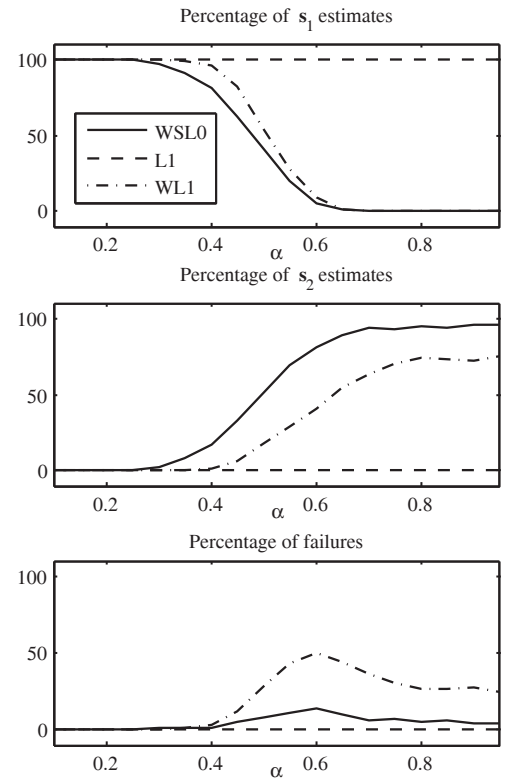

Fig. 2. The percentages of $\mathbf{s}_1$ and $\mathbf{s}_2$ estimates, and the percentages of failure versus α. For $\alpha < 0.5$, $\mathbf{s}_1$ is the solution of both P_0 and $P_{0,\mathbf{w}}$ problems, whereas for $\alpha > 0.5$ the solution of P_0 is $\mathbf{s}_1$, and the solution of $P_{0,\mathbf{w}}$ is $\mathbf{s}_2$.

It is seen from Fig. 2 that the ℓ^1-norm minimization always estimates the solution of P_0, as expected. Although it can be considered as a relaxation of both P_0 and $P_{0,\mathbf{w}}$ problems, especially for large α's (that is when $\|\mathbf{s}_1\|_{0,\mathbf{w}} \gg \|\mathbf{s}_2\|_{0,\mathbf{w}}$), the weighted ℓ^1 recovers $\mathbf{s}_2$ most of the times, but it has also a large failure rate (around 20% for α near 1). WSL0 does a better job, and for $\alpha \approx 1$ it recovers $\mathbf{s}_2$ about 98% of the times.

6. CONCLUSIONS

In this paper, the weighted sparse signal decomposition problem $P_{0,\mathbf{w}}$ was studied. It was shown that $P_{0,\mathbf{w}}$ has applications in recovering the sparse solution of an underdetermined linear system where prior information on the activity of different atoms is available. Then, a condition for the uniqueness of the solution of $P_{0,\mathbf{w}}$

was presented. The challenges of using the weighted ℓ^1 relaxation for solving $P_{0,\mathbf{w}}$ were also pointed out, and an algorithm based on generalizing SL0 to the weighted case was developed to approximate the solution of $P_{0,\mathbf{w}}$. Finally, a simulation was presented to compare the results of weighted ℓ^1 and WSL0 approaches.

Future work will include generalization of [15, Theorem 1] to the weighted case, and studying the stability of $P_{0,\mathbf{w}}$.

7. APPENDIX

Proof of Theorem 1. Since there is no prior information on $\mathbf{r}$, the probability density function $p(\mathbf{r})$ is treated as uniform; hence:

$$\begin{aligned}(\hat{\mathbf{b}}_{\text{MAP}}, \hat{\mathbf{r}}_{\text{MAP}}) &= \underset{\mathbf{b},\mathbf{r}}{\text{argmax}}\ p(\mathbf{x}|\mathbf{b},\mathbf{r})p(\mathbf{b}) = \underset{\mathbf{b},\mathbf{r}}{\text{argmax}}\ p(\mathbf{x}|\mathbf{s})p(\mathbf{b}) \\ &= \underset{\mathbf{b},\mathbf{r}}{\text{argmax}}\ p(\mathbf{b}) \text{ s.t } \mathbf{x} = \mathbf{A}\mathbf{s} \\ &= \underset{\mathbf{b},\mathbf{r}}{\text{argmin}}\ -\ln p(\mathbf{b}) \text{ s.t } \mathbf{x} = \mathbf{A}\mathbf{s} \end{aligned} \tag{9}$$

The prior probability of b_i is:

$$\begin{aligned} p(b_i) &= \begin{cases} p_i & \text{if } b_i = 0 \\ 1 - p_i & \text{if } b_i = 1 \end{cases} \\ &= p_i^{|b_i|_0}(1-p_i)^{1-|b_i|_0} = (1-p_i)\left(\frac{p_i}{1-p_i}\right)^{|b_i|_0}. \end{aligned}$$

Hence, the prior probability of $\mathbf{b}$ is

$$p(\mathbf{b}) = \prod_{i=1}^{m} p(b_i) = \left[\prod_{i=1}^{m}(1-p_i)\right] / \left[\prod_{i=1}^{m}\left(\frac{1-p_i}{p_i}\right)^{|b_i|_0}\right] \tag{10}$$

and so

$$-\ln p(\mathbf{b}) = \text{constant} + \sum_{i=1}^{m} w_i |b_i|_0 \tag{11}$$

where w_i is as given in the theorem. Substituting (11) into (9), and noting that $|b_i|_0 = |s_i|_0$ proves the result. □

Proof of Theorem 2. Arguing by contradiction, suppose that besides $\mathbf{s}_0$ there is another solution $\mathbf{s}_1$ satisfying (6). It is easy to see that the pseudo-norm $\|\cdot\|_{0,\mathbf{w}}$ satisfies the triangle inequality, and hence

$$\|\mathbf{s}_0 - \mathbf{s}_1\|_{0,\mathbf{w}} \le \|\mathbf{s}_0\|_{0,\mathbf{w}} + \|\mathbf{s}_1\|_{0,\mathbf{w}} < \mathcal{S}_{\mathbf{w}}\big(\textit{spark}(\mathbf{A})\big). \tag{12}$$

Note also that if for a vector $\mathbf{y}$, $\|\mathbf{y}\|_{0,\mathbf{w}} < \mathcal{S}_{\mathbf{w}}(k)$, then the number of non-zero entries in $\mathbf{y}$ should be smaller than k (otherwise, the sum of the corresponding weights could not be smaller than the sum of k smallest weights, that is, $\mathcal{S}_{\mathbf{w}}(k)$). In other words, $\|\mathbf{y}\|_{0,\mathbf{w}} < \mathcal{S}_{\mathbf{w}}(k)$ implies $\|\mathbf{y}\|_0 < k$. Therefore, (12) implies $\|\mathbf{s}_0 - \mathbf{s}_1\|_0 < \textit{spark}(\mathbf{A})$. On the other hand, $\mathbf{A}\mathbf{s}_0 = \mathbf{A}\mathbf{s}_1 = \mathbf{x} \Rightarrow \mathbf{A}(\mathbf{s}_0 - \mathbf{s}_1) = \mathbf{0}$. But this contradicts the fact that $\|\mathbf{s}_0 - \mathbf{s}_1\|_0 < \textit{spark}(\mathbf{A})$, because a linear combination of less than $\textit{spark}(\mathbf{A})$ columns of $\mathbf{A}$ (corresponding to non-zero entries of $\mathbf{s}_0 - \mathbf{s}_1$) is null. □

8. REFERENCES

[1] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency info.," *IEEE Trans. on Info. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

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

[3] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

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

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

[6] M. Elad, "Why simple shrinkage is still relevant for redundant representations?," *IEEE Trans. on Image Processing*, vol. 52, no. 12, pp. 5559–5569, 2006.

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

[8] N. Vaswani and W. Lu, "Modified-CS: Modifying compressive sensing for problems with partially known support," *IEEE Trans. on Signal Processing*, vol. 58, no. 9, pp. 4595–4607, Sep. 2010.

[9] C. J. Miosso, R. von Borries, M. Argàez, L. Velazquez, C. Quintero, and C.M. Potes, "Compressive sensing reconstruction with prior information by iteratively reweighted least-squares," *IEEE Trans. on Signal Processing*, vol. 57, no. 6, pp. 2424–2431, June 2009.

[10] M. A. Khajehnejad, W. Xu, A. S. Avestimehr, and B. Hassibi, "Weighted ℓ_1 minimization for sparse recovery with prior information," in *IEEE Intl. Symp. on Info. Theory*, Seoul, South Korea, 2009, pp. 483–487.

[11] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing*, vol. 20, no. 1, pp. 33–61, 1999.

[12] E. Candès, M. B. Wakin, and S.P. Boyd, "Enhancing sparsity by reweighted l_1 minimization," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5, pp. 877–905, Dec. 2008.

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

[14] R. Gribonval and M. Nielsen, "Sparse decompositions in unions of bases," *IEEE Trans. Inform. Theory*, vol. 49, no. 12, pp. 3320–3325, Dec. 2003.

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

[16] A. Blake and A. Zisserman, *Visual Reconstruction*, MIT Press, 1987.