On the Use of *A Priori* Information for Sparse Signal Approximations

Oscar Divorra Escoda, Lorenzo Granai and Pierre Vandergheynst
Signal Processing Institute (ITS)
Ecole Polytechnique Fédérale de Lausanne (EPFL)
LTS2-ITS-STI-EPFL, 1015 Lausanne, Switzerland

Technical Report No. 23/2004

Abstract

This report is the extension to the case of sparse approximations of our previous study on the effects of introducing *a priori* knowledge to solve the recovery of sparse representations when overcomplete dictionaries are used [1]. Greedy algorithms and Basis Pursuit Denoising are considered in this work. Theoretical results show how the use of "reliable" *a priori* information (which in this work appears under the form of weights) can improve the performances of these methods. In particular, we generalize the sufficient conditions established by *Tropp* [2], [3] and *Gribonval and Vandergheynst* [4], that guarantee the retrieval of the sparsest solution, to the case where *a priori* information is used. We prove how the use of *prior* models at the signal decomposition stage influences these sufficient conditions. The results found in this work reduce to the classical case of [4] and [3] when no *a priori* information about the signal is available. Finally, examples validate and illustrate theoretical results.

Index Terms

Sparse Approximations, Sparse Representations, Basis Pursuit Denoising, Matching Pursuit, Relaxation Algorithms, Greedy Algorithms, *A Priori* Knowledge, Redundant Dictionaries, Weighted Basis Pursuit Denoising, Weighted Matching Pursuit.

Contents

I Introduction 2

II Recovery of General Signals: Sparse Approximations 2
II-A Greedy Algorithms: *Weak*-MP 3
II-A.1 Robustness 3
II-A.2 Rate of Convergence 4
II-B Convex Relaxation of the Subset Selection Problem 4

III Including *A Priori* Information on Greedy algorithms 5
III-A Influence on Sparse Approximations 5
III-B Rate of Convergence of Weighted-MP/OMP on Sparse Approximations 8
III-C Example: Use of Footprints for ϵ-Sparse Approximations. 12

IV Approximations with Weighted Basis Pursuit Denoising 13
IV-A A Bayesian Approach to Weighted Basis Pursuit Denoising 14
IV-B Preliminary Propositions 15
IV-C Weighted Relaxed Subset Selection 17
IV-D Relation with the Weighted Cumulative Coherence 18
IV-E Example 19

V Examples: A Natural Signal Approximation with Coherent Dictionaries and an *A Priori* Model 19
V-A A Noisy Piecewise-smooth Natural Signal 20
V-B Modeling the Relation Signal-Dictionary 21
V-C Signal Approximation 21
V-D Results 21
V-D.1 Approximation Results with OMP 21

Web page: http://lts2www.epfl.ch
The work of Oscar Divorra Escoda is partly sponsored by the IM.2 NCCR
The work of Lorenzo Granai is supported by the SNF grant 2100-066912.01/1

V-D.2 Approximation Results with BPDN 21
V-D.3 Capturing the Piecewise-smooth Component with Footprints Basis 23
V-D.4 Parameter Search 23

VI Conclusions 24

References 26

I. Introduction

In many applications, such as compression, denoising or source separation, one seeks an efficient representation or approximation of the signal by means of a linear expansion into a possibly overcomplete family of functions. In this setting, efficiency is often characterized by sparseness of the associated series of coefficients. The criterion of sparseness has been studied for a long time and in the last few years has become popular in the signal processing community [5], [6], [2], [3]. Natural signals are very unlikely to be exact sparse superpositions of vectors. In fact the set of such signals is of measure zero in $\mathbb{C}^N$ [2]. We thus extend our previous work on sparse exact representations [1] to the more useful case of sparse approximations:

$$\min_{\mathbf{c}} \|f - D\mathbf{c}\|_2^2 \quad \text{s.t.} \quad \|\mathbf{c}\|_0 \leq m. \tag{1}$$

In general, the problem of recovering the sparsest signal approximation (or representation) over a redundant dictionary is a NP-hard problem. However this does not impair the possibility of finding this efficiently when particular classes of dictionaries are used [7]. As demonstrated in [2], [3], [6], [4], in order to ensure the good behavior of algorithms like General *Weak*(α) Matching Pursuit (*Weak*-MP) and Basis Pursuit Denoising (BPDN), dictionaries need to be incoherent enough. Under this main hypothesis, sufficient conditions have been stated so that both methods are able to recover the atoms from the sparsest m-terms expansion of a signal.

However, experience and intuition dictate that good dictionaries for sparse approximations of natural signals can be very dense and, depending on the kind of signal structures to exploit, they may in most cases be highly coherent. For example, consider a natural image from which an efficient approximation of edges is desired. Several approaches have been proposed where the functions in use have a very strong geometrical meaning [8], [9], [10], [11]. Indeed, they represent local orientation of edges. In order to accurately represent all orientations, the set of functions to use need to finely sample the direction parameter. This yields that the atoms of the dictionary may have a strong correlation. Moreover, if further families of functions are considered in the dictionary in order to model other components like textures, or smooth areas, the coherence between the whole set of functions can become even higher. As a further motivation, one can also observe that the set of visual primitives obtained by Olshausen and Field while studying the spatial receptive fields of simple cells in mammalian striate cortex [12] is redundant and with a high coherence.

Concerning the case of exact sparse signal representations, we introduced in [1] a way of using more coherent dictionaries with *Weak*-MP and Basis Pursuit (BP), while keeping the possibility of recovering the optimal solution. Two methods were proposed based on the use of *a priori* information about the signal to decompose: we called them Weighted-MP and Weighted-BP. In this report we address the case of sparse signal approximations, discussing the potentiality of using *a priori* knowledge in the atom selection procedure. We do not face here the issue of how to find a reliable and useful *a priori* knowledge about a signal. This problem strongly depends on the nature of the signal and on the kind of dictionary used. The aim of this paper is the theoretical study of the weighted algorithms in the prospective of achieving sparseness. Main results are:

- The definition of Weighted-BPDN and Weighted-MP/OMP algorithms for approximation purposes. We reformulate classic BPDN and *Weak*-MP in order to take *a priori* information into account when decomposing the signal.
- A sufficient condition under which Weighted Basis Pursuit Denoising and Weighted-MP/OMP find the best m-term signal approximation.
- A study of how adapting the decomposition algorithm depending on *a priori* information may help in the recovery of sparse approximations.
- An analysis of the effects of adding the *a priori* weights on the rate of convergence of *Weak*-MP.
- An empirical analysis, on natural signals, of the effect of using *prior* models at the decomposition stage when coherent overcomplete dictionaries are used.

II. Recovery of General Signals: Sparse Approximations

Exact sparse representations are mostly useless in practice, since the set of such signals is of measure zero in $\mathbb{C}^N$ [2]. There are very few cases where, for a given signal, one can hope to find an m-sparse representation.

Sparse approximations, on the other hand, have found numerous applications, e.g.. in compression, restoration, denoising or sources separation.

The use of overcomplete dictionaries, although in a sense advantageous, has the drawback that finding a solution to (1) may be rather difficult or even practically infeasible. Suboptimal algorithms that are used instead (like Weak General Matching Pursuit algorithms - *Weak*-MP- [4] or ℓ_1-norm Relaxation Algorithms -BPDN- [13]) do not necessarily supply the same solution as the problem formulated in (1). However, there exist particular situations in which they succeed in recovering the "correct" solution. Very important results have been found in the case of using incoherent dictionaries to find sparse approximations of signals. Indeed, sufficient conditions have been found such that for a dictionary *Weak*-MP and BPDN algorithms can be guaranteed to find the set of atoms that form the sparsest m-terms approximant f_m^{opt} of a signal f. Moreover, for the case where Orthogonal Matching Pursuit (OMP) is used the set of coefficients found by the algorithm will correspond also to the optimal one.

Prior to reviewing the results that state the sufficient conditions introduced above, let us define and recall a series of elements that will be used in the remaining of the paper.

- $f \in \mathcal{H}$ is the function to be approximated, where $\mathcal{H}$ is a Hilbert space. Unless otherwise stated, in this report we assume $f \in \mathbb{R}^n$.
- $\mathcal{D}$ and D define respectively the set of atoms included in the dictionary and the dictionary synthesis matrix where each one of the columns corresponds to an atom ($\mathcal{D} = \{g_i : \ i \in \Omega\}$).
- f_m^{opt} is the best approximant of f such that $f_m^{opt} = D \cdot \mathbf{c}_{opt}$ where the support of $\mathbf{c}_{opt}$ is smaller than or equal to a positive integer m.
- Given $n \geq 0$, r_n and f_n are the residual and approximant generated by a greedy algorithm at its nth iteration.
- Γ_m is the optimal set of m atoms that generate f_m^{opt}. Often, in the text, this will be referred to as Γ for simplicity.
- The best approximation of f over the atoms in Λ is called $a_\Lambda = DD_\Lambda^+ f$.
- $\alpha \in (0,1]$, is defined as the weakness factor associated to the atom selection procedure of *Weak*-MP algorithms [14].
- $\mu_1(m, \mathcal{D})$ is a measure of the internal cumulative coherence of $\mathcal{D}$ [2]:
$$\mu_1(m, \mathcal{D}) \triangleq \max_{|\Lambda|=m} \max_{i \in \Omega \setminus \Lambda} \sum_{\lambda \in \Lambda} |\langle g_i, g_\lambda \rangle| \,, \tag{2}$$
where $\Lambda \subset \Omega$ has size m. Remark that the measure known as coherence of a dictionary (μ) and often used to characterize redundant dictionaries corresponds to the particular case of $\mu = \mu_1(1, \mathcal{D})$. Furthermore $\mu_1(m, \mathcal{D}) \leq m\mu$.
- Let η ($\eta \geq 0$) be a suboptimality factor associated to the case where the best m-terms approximation can not be reached by the algorithm in use. In such a case, the residual error energy after approximation is $(1+\eta)^2 \, \|r_m^{opt}\|_2^2$ instead of $\|r_m^{opt}\|_2^2$.

A. Greedy Algorithms: Weak-*MP*

Gribonval and Vandergheynst extended the results *Tropp* found for the particular case of OMP to the general *Weak*-MP. Akin to the case of signal representation, the main results consist in the sufficient conditions that guarantee that *Weak*-MP will recover the optimal set of atoms that generate the best m-terms approximant f_m^{opt}. Moreover, a result establishes as well an upper bound on the decay of the residual energy in the approximation of a signal that depends on the internal coherence of $\mathcal{D}$, and a bound on how many "correct" iterations can be performed by the greedy algorithm depending on the dictionary and the energy of f_m^{opt}.

1) Robustness: The sufficient conditions found in [4] that ensure that *Weak*-MP will recover the set of atoms that compose the best m-terms approximant are enounced in Theorem 1. First of all, it is necessary that the optimal set Γ_m satisfies the Stability Condition [4]. If in addition some conditions are satisfied concerning the remaining residual energy at the *nth* iteration ($\|r_n\|_2^2$) and the optimal residual energy $\|r_m^{opt}\|_2^2$, then an additional atom belonging to Γ_m will be recovered. This condition, called originally the General Recovery Condition in [2], was named, for the case of general *Weak*-MP, the Robustness Condition in [4].

***Theorem** 1:* (*Gribonval & Vandergheynst* [4]) Let $\{r_n\}_{n \geq 0}$ be a sequence of residuals computed by General MP to approximate some $f \in \mathcal{H}$. For any integer m such that $\mu_1(m-1) + \mu_1(m) \leq 1$, let $f_m^{opt} = \sum_{\gamma \in \Gamma_m} c_\gamma g_\gamma$ be a best m-terms approximation to f, and let $N_m = N_m(f)$ be the smallest integer such that
$$\|r_{N_m}\|_2^2 \leq \|r_m^{opt}\|_2^2 \cdot \left(1 + \frac{m \cdot (1 - \mu_1(m-1))}{(1 - \mu_1(m-1) - \mu_1(m))^2}\right). \tag{3}$$
Then, for $1 \leq n < N_m$, General MP picks up a "correct" atom. If no best m-terms approximant exists, the same results are valid provided that $\|r_m^{opt}\|_2 = \|f - f_m^{opt}\|_2$ is replaced with $\|f - f_m^{opt}\|_2 = (1+\eta) \, \|r_m^{opt}\|_2$ in (3).

2) Rate of Convergence: In the following the main result concerning the exponential decay of the error energy bound, as well as the bound on how many "correct" iterations can be performed by the greedy algorithm, is reviewed.

Theorem 2: (*Gribonval & Vandergheynst* [4]) Let $\{r_n\}_{n\geq 0}$ be a sequence of residuals computed by General MP to approximate some $f \in \mathcal{H}$. For any integer m such that $\mu_1(m-1)+\mu_1(m) \leq 1$, we have that

$$\|r_n\|_2^2 - \left\|r_m^{opt}\right\|_2^2 \leq \left(1 - \frac{1-\mu_1(m-1)}{m}\right)^{n-l} \left(\|r_l\|_2^2 - \left\|r_m^{opt}\right\|_2^2\right). \quad (4)$$

Moreover, $N_1 \leq 1$, and for $m \geq 2$:

- if $\|r_m^{opt}\|_2^2 \leq 3\,\|r_1\|_2^2/m$, then

$$2 \leq N_m < 2 + \frac{m}{1-\mu_1(m-1)} \cdot \ln \frac{3 \cdot \|r_1\|_2^2}{m \cdot \|r_m\|_2^2} \quad (5)$$

- else $N_m \leq 1$.

B. Convex Relaxation of the Subset Selection Problem

Another instance of problem (1) is given by

$$\min_{\mathbf{c}} \|f - D\mathbf{c}\|_2^2 + \tau^2 \|\mathbf{c}\|_0. \quad (6)$$

Unfortunately the function that has to be minimized is not convex. One can define a p-norm of a vector $\mathbf{c}$ for any positive real p:

$$\|\mathbf{c}\|_p = \left(\sum_{i\in\Omega} |c_i|^p\right)^{1/p}. \quad (7)$$

It is well known that the smallest p for which Eq. (7) is convex is 1. For this reason the convex relaxation of the subset selection problem was introduced in [13] under the name of Basis Pursuit Denoising:

$$\min_{\mathbf{b}} \frac{1}{2} \|f - D\mathbf{b}\|_2^2 + \gamma \|\mathbf{b}\|_1. \quad (8)$$

This problem can be solved recurring to Quadratic Programming techniques (see also Section IV). In [3], the author studies the relation between the subset selection problem (6) and its convex relaxation (8). Next theorem shows that any coefficient vector which minimizes Eq. (8) is supported inside the optimal set of indexes.

Theorem 3: (Correlation Condition, *Tropp* [3]) Suppose that the maximum inner product between the residual signal and any atom satisfies the condition

$$\|D^*(f - \mathbf{a}_\Lambda)\|_\infty < \gamma(1 - \sup_{i\notin\Lambda} \|D_\Lambda^+ g_i\|_1).$$

Then any coefficient vector $\mathbf{b}_*$ that minimizes the function (8) must satisfy *support*$(\mathbf{b}_*) \subset \Lambda$.

In particular, the following theoretical result show how the trade off parameters τ and γ are related.

Theorem 4: (*Tropp* [3]) Suppose that the coefficient vector $\mathbf{b}_*$ minimizes the function (8) with threshold $\gamma = \tau/(1 - \sup_{i\notin\Gamma} \left\|D_\Gamma^+ g_i\right\|_1)$. Then we have that:

1) the relaxation never selects a non optimal atom since *support*$(\mathbf{b}_*) \subset$ *support*$(\mathbf{c}_{opt})$.
2) The solution of the convex relaxation is unique.
3) The following upper bound is valid:

$$\|\mathbf{c}_{opt} - \mathbf{b}_*\|_\infty \leq \frac{\tau \cdot \left\|\left(D_\Gamma^* D_\Gamma\right)^{-1}\right\|_{\infty,\infty}}{1 - \sup_{i\notin\Gamma} \|D_\Gamma^+ g_i\|_1}. \quad (9)$$

4) The support of $\mathbf{b}_*$ contains every index j for which

$$|\mathbf{c}_{opt}(j)| > \frac{\tau \cdot \left\|\left(D_\Gamma^* D_\Gamma\right)^{-1}\right\|_{\infty,\infty}}{1 - \sup_{i\notin\Gamma} \|D_\Gamma^+ g_i\|_1}. \quad (10)$$

If the dictionary we are working with is orthonormal it follows that

$$\sup_{i\notin\Gamma} \|D_\Gamma^+ g_i\|_1 = 0 \text{ and } \left\|\left(D_\Gamma^* D_\Gamma\right)^{-1}\right\|_{\infty,\infty} = 1$$

and the previous theorem becomes much stronger. In particular we obtain that $\|\mathbf{c}_{opt} - \mathbf{b}_*\|_\infty \leq \tau$ and $|\mathbf{c}_{opt}(j)| > \tau$ [3], [15].

A problem similar to the subset selection is given by the retrieval of a sparse approximation given an error constraint:

$$\min_{\mathbf{c}} \|\mathbf{c}\|_0 \quad s.t. \quad \|f - D\mathbf{c}\|_2 \leq \epsilon_0, \tag{11}$$

whose natural convex relaxation is given by

$$\min_{\mathbf{b}} \|\mathbf{b}\|_1 \quad s.t. \quad \|f - D\mathbf{b}\|_2 \leq \epsilon_1. \tag{12}$$

In this paper, we are not going to explore this problem, but let us just recall that if the dictionary is incoherent, then the solution to (12) for a given ϵ_1 is at least as sparse as the solution to (11), with a tolerance ϵ_0 somewhat smaller than ϵ_1 [3].

From all these results, one can infer that the use of incoherent dictionaries is very important for the good behavior of greedy and ℓ_1-norm relaxation algorithms. However, as discussed in the introduction, experience seems to teach us that the overcomplete dictionaries which are likely to be powerful for natural signals approximation would be very redundant and with significant internal coherence. Hence, this inconsistent and contradictory situation claims for a solution to be found. In the following sections, we introduce a general approach that intends to tackle this problem. In our opinion, a more careful analysis and modeling of the signal to approximate is necessary. Dictionary waveforms are not enough good modeling elements to be exploited in the signal decomposition stage. Further analysis is required to better drive the decomposition algorithm.

As in our previous work concerning the case of exact signal representations [1], in this report, *a priori* knowledge that relates the signal f and the dictionary $\mathcal{D}$ are considered for signal approximations.

III. Including *A Priori* Information on Greedy algorithms

As seen in our previous work on the *exact* sparse representation of signals [1], the use of *a priori* information on greedy algorithms may make the difference between recovering the optimal set of components for a given approximation or not. In this section we explore the effect of using *a priori* knowledge on greedy algorithms on the recovery of the best m-terms approximant (f_m^{opt}) of a signal f. First, sufficient conditions for the recovery of a “correct” atom from the sparsest m-terms approximant are established for the case where *a priories* are taken into account. Later, we study how *prior* knowledge affects the rate of convergence of greedy algorithms. Finally, a practical example is presented.

A. Influence on Sparse Approximations

An important result concerning sparse approximations is the feasibility of recovering the sparsest m-terms approximation f_m^{opt}. Akin to the statements established for the exact representation case, sufficient conditions have been determined such that, given a *Weak*-MP and the associated series of atoms g_{γ_n} and residuals r_n ($n \geq 0$) up to the nth step, a “correct” atom at the $(n+1)$th step can be guaranteed (see Sec. II). In Theorem 5, sufficient conditions are presented for the case when some *a priori* knowledge is available. The main interest of this result is to show that if an appropriate *a priori* (concerning f and $\mathcal{D}$) is in use, a better approximation result can be achieved.

First of all, let us expose the elements that take part into the following results. In order to enhance the clarity of the explanation, we first recall the definition of the main concepts that will be used. Let us consider first the diagonal matrix $W(f, \mathcal{D})$ introduced in [1] to represent the *a priori* knowledge taken into account in the atom selection procedure.

***Definition** 1:* A weighting matrix $W = W(f, \mathcal{D})$ is a square diagonal matrix of size $d \times d$. Each of the entries $w_i \in (0, 1]$ from the diagonal corresponds to the *a priori* likelihood of a particular atom $g_i \in \mathcal{D}$ to be part of the sparsest decomposition of f.

We define also $w_{\overline{\Gamma}}^{max}$ as the biggest of the weights corresponding to the subset of atoms belonging to $\overline{\Gamma} = \Omega \setminus \Gamma$, hence:

$$w_{\overline{\Gamma}}^{max} \triangleq \sup_{\gamma\in\overline{\Gamma}} |w_\gamma| \,. \tag{13}$$

Moreover, an additional quantity is required in the results depicted below:

$$\epsilon_{max} \triangleq \sup_{\gamma\in\Gamma} \left|1 - w_{\gamma}^2\right|. \tag{14}$$

Eqs. (13) and (14) give information about how good the *a priori* information is. The reader will notice that these quantities depend on the optimal set of atoms Γ, making not possible to establish a rule to compute them in advance. The role of these magnitudes is to represent the influence of the *prior* knowledge quality in the results obtained below. Notice that $0 \leq \epsilon_{max} \leq 1$ and $0 \leq w_{\overline{\Gamma}}^{max} \leq 1$. ϵ_{max} is close to zero if "good" atoms (the ones belonging to Γ) are not penalized by the *a priori*. If the supplied *a priori* is a good enough model of the relation between the signal and the dictionary in use, we state that the *a priori* knowledge is reliable. $w_{\overline{\Gamma}}^{max}$ becomes small if "bad" atoms are strongly penalized by the *prior* knowledge.As we will see in the following, if the *a priori* is reliable and $w_{\overline{\Gamma}}^{max}$ is small, then the *prior* knowledge can have a relevant positive influence in the behavior of the greedy algorithm.

The consequence of taking into account the *a priori* matrix W, is to allow a new definition of the *Babel Function* introduced by *Tropp* [2]. In [1] the fact that not all the atoms of $\mathcal{D}$ are equiprobabile is taken into account. In effect, the availability of some *prior* should be considered when judging whether a greedy algorithm is going to be able to recover the m-sparsest approximation of a signal f or not. As seen in the Sec. II, the conditions that ensure the recoverability of the best m term approximant relay on the internal coherence measure of a dictionary $\mu_1(m)$. Using the *a priori* information, some atom interactions can be penalized or even dismissed in the cumulative coherence measure. Hence, a new signal dependent cumulative coherence measure was introduced in [1]:

***Definition** 2:* The *Weighted Cumulative Coherence* function of $\mathcal{D}$ is defined as the following data dependent measure:

$$\mu_1^w(m,\mathcal{D},f) \triangleq \max_{|\Lambda|=m} \max_{i\in\Omega\backslash\Lambda} \sum_{\lambda\in\Lambda} |< g_\lambda, g_i >| \cdot w_\lambda \cdot w_i. \tag{15}$$

Once all necessary elements have been defined, we can finally state the result that shows the behavior of greedy algorithms on the use of *a priori* information for the recovery of m-sparse approximants. As proved later in this section, the use of such knowledge implies an improvement with respect to the classic *Weak*-MP also for the case of signal approximation.

***Theorem** 5:* Let $\{r_n\} \ : n \geq 0$, be the set of residuals generated by Weighted-MP/OMP in the approximation of a signal f, and let f_m^{opt} be the best m-terms approximant of f over $\mathcal{D}$. Then, for any positive integer m such that, for a reliable *a priori* information $W(f,\mathcal{D})$, $\mu_1^w(m-1) + \mu_1^w(m) < 1 - \epsilon_{max}$, $\eta \geq 0$ and

$$\|r_n\|_2^2 > \|f - f_m^{opt}\|_2^2 (1+\eta)^2 \left(1 + \frac{m\left(1 - (\mu_1^w(m-1) + \epsilon_{max})\right)\left(w_{\overline{\Gamma}}^{max}\right)^2}{\left(1 - (\mu_1^w(m-1) + \mu_1^w(m) + \epsilon_{max})\right)^2}\right), \tag{16}$$

Weighted-MP/OMP will recover an atom that belongs to the optimal set Γ_m that expand the best m-sparse approximant of f. If the best m-terms approximant f_m^{opt} exist and can be reached, then $\eta = 0$.

This means that if the approximation error at the nth iteration is still bigger than a certain quantity which depends on the optimal error ($\|f - f_m^{opt}\|_2^2$), the internal dictionary cumulative coherence and the reliability of the *a priori* information, then still another term of the best $m - term$ approximant can still be recovered. The use of reliable *a priori* information makes the bound easier to satisfy compared to when no *prior* knowledge is used [4]. Thus, a higher number of terms from the best $m - term$ approximant may be recovered.

Proof: To demonstrate the result of Theorem 5, we follow the steps of the original proofs by *Tropp* [2] and *Gribonval and Vandergheynst* [4]. This time however, *a priori* knowledge is taken into account. First of all, let us remind the following statements:

- $f_m^{opt} \in span(\Gamma_m)$
- $r_n = f - f_n$
- $r_m^{opt} = f - f_m^{opt}$ is such that $r_m^{opt} \perp (f_m^{opt} - f_n) \ \forall\ 0 \leq n < m$, hence $\|r_n\|_2^2 = \|f_m^{opt} - f_n\|_2^2 + \|f - f_m^{opt}\|_2^2$.

In order to ensure the recovery of any atom belonging to the optimal set $\Gamma = \Gamma_m$, the following needs to be satisfied:

$$\rho^w(r_n) = \frac{\|D_{\overline{\Gamma}} \cdot W_{\overline{\Gamma}} \cdot r_n\|_\infty}{\|D_\Gamma \cdot W_\Gamma \cdot r_n\|_\infty} < \alpha, \tag{17}$$

where $\alpha \in (0,1]$ is the weakness factor [14]. To establish (16), the previous expression has to be put in terms of f_m^{opt} and f_n. Hence,

$$\begin{aligned}
\rho^w(r_n) &= \frac{\|D_{\overline{\Gamma}}\cdot W_{\overline{\Gamma}}\cdot r_n\|_\infty}{\|D_\Gamma\cdot W_\Gamma\cdot r_n\|_\infty} = \frac{\|D_{\overline{\Gamma}}\cdot W_{\overline{\Gamma}}\cdot (f-f_n)\|_\infty}{\|D_\Gamma\cdot W_\Gamma\cdot (f-f_n)\|_\infty} \\
&= \frac{\left\|D_{\overline{\Gamma}}\cdot W_{\overline{\Gamma}}\cdot\left(f-f_m^{opt}\right)+D_{\overline{\Gamma}}\cdot W_{\overline{\Gamma}}\cdot\left(f_m^{opt}-f_n\right)\right\|_\infty}{\left\|D_\Gamma\cdot W_\Gamma\cdot\left(f-f_m^{opt}\right)+D_\Gamma\cdot W_\Gamma\cdot\left(f_m^{opt}-f_n\right)\right\|_\infty} \\
&= \frac{\left\|D_{\overline{\Gamma}}\cdot W_{\overline{\Gamma}}\cdot\left(f-f_m^{opt}\right)+D_{\overline{\Gamma}}\cdot W_{\overline{\Gamma}}\cdot\left(f_m^{opt}-f_n\right)\right\|_\infty}{\left\|D_\Gamma\cdot W_\Gamma\cdot\left(f_m^{opt}-f_n\right)\right\|_\infty} \\
&\le \frac{\left\|D_{\overline{\Gamma}}\cdot W_{\overline{\Gamma}}\cdot\left(f-f_m^{opt}\right)\right\|_\infty}{\left\|D_\Gamma\cdot W_\Gamma\cdot\left(f_m^{opt}-f_n\right)\right\|_\infty} + \frac{\left\|D_{\overline{\Gamma}}\cdot W_{\overline{\Gamma}}\cdot\left(f_m^{opt}-f_n\right)\right\|_\infty}{\left\|D_\Gamma\cdot W_\Gamma\cdot\left(f_m^{opt}-f_n\right)\right\|_\infty} \\
&= \frac{\left\|D_{\overline{\Gamma}}\cdot W_{\overline{\Gamma}}\cdot\left(f-f_m^{opt}\right)\right\|_\infty}{\left\|D_\Gamma\cdot W_\Gamma\cdot\left(f_m^{opt}-f_n\right)\right\|_\infty} + \rho^w\left(f_m^{opt}-f_n\right),
\end{aligned} \tag{18}$$

where the second term can be upper bounded since $(f_m^{opt}-f_n)\in span(\Gamma)$ [1],

$$\rho^w\left(f_m^{opt}-f_n\right) \le \frac{\mu_1^w(m)}{1-(\mu_1^w(m-1)+\epsilon_{max})}. \tag{19}$$

The first term of the last equality in (18) can be upper bounded in the following way:

$$\frac{\left\|D_{\overline{\Gamma}}\cdot W_{\overline{\Gamma}}\cdot\left(f-f_m^{opt}\right)\right\|_\infty}{\left\|D_\Gamma\cdot W_\Gamma\cdot\left(f_m^{opt}-f_n\right)\right\|_\infty} = \frac{\sup\limits_{\gamma\in\overline{\Gamma}}\left|\left\langle g_\gamma\cdot w_\gamma,\left(f-f_m^{opt}\right)\right\rangle\right|}{\left\|D_\Gamma\cdot W_\Gamma\cdot\left(f_m^{opt}-f_n\right)\right\|_\infty}, \tag{20}$$

and by the Cauchy-Schwarz inequality,

$$\begin{aligned}
\frac{\sup\limits_{\gamma\in\overline{\Gamma}}\left|\left\langle g_\gamma\cdot w_\gamma,\left(f-f_m^{opt}\right)\right\rangle\right|}{\left\|D_\Gamma\cdot W_\Gamma\cdot\left(f_m^{opt}-f_n\right)\right\|_\infty} &\le \frac{\sup\limits_{\gamma\in\overline{\Gamma}}\|g_\gamma\cdot w_\gamma\|_2\cdot\left\|f-f_m^{opt}\right\|_2}{\left\|D_\Gamma\cdot W_\Gamma\cdot\left(f_m^{opt}-f_n\right)\right\|_\infty} \\
= \frac{\sup\limits_{\gamma\in\overline{\Gamma}}|w_\gamma|\cdot\left\|f-f_m^{opt}\right\|_2}{\left\|D_\Gamma\cdot W_\Gamma\cdot\left(f_m^{opt}-f_n\right)\right\|_\infty} &= \frac{\sup\limits_{\gamma\in\overline{\Gamma}}|w_\gamma|\cdot\left\|f-f_m^{opt}\right\|_2}{\sup\limits_{\gamma\in\Gamma}\left|\left\langle g_\gamma\cdot w_\gamma,\left(f_m^{opt}-f_n\right)\right\rangle\right|}.
\end{aligned} \tag{21}$$

In order to further upper bound the expression above, the denominator can be lower bounded, as shown in [1]. Indeed, by the singular value decomposition:

$$\sup_{\gamma\in\Gamma}\left|\left\langle g_\gamma\cdot w_\gamma,\left(f_m^{opt}-f_n\right)\right\rangle\right| \ge \sqrt{\frac{\sigma^2_{min_w}}{m}}\left\|f_m^{opt}-f_n\right\|_2, \tag{22}$$

where $\sigma^2_{min_w}$ is the minimum of the squared singular values of $G \triangleq (D_\Gamma W_\Gamma)^T(D_\Gamma W_\Gamma)$, and can be bounded as $\sigma^2_{min_w} \ge 1-\epsilon_{max}-\mu_1^w(m-1)$. Moreover, in (21), $\|f-f_m^{opt}\|_2$ can be defined as $\|f-f_m^{opt}\|_2 = (1+\eta)\cdot\|r_m^{opt}\|_2$, where $\eta\ge 0$ stands for a sub-optimality factor which indicates whether f_m^{opt} can be reached and, if not possible (i.e. $\eta\ne 0$), sets the best possible reachable approximation error. Hence, (21) can be rewritten as:

$$\frac{\sup\limits_{\gamma\in\overline{\Gamma}}|w_\gamma|\cdot\left\|f-f_m^{opt}\right\|_2}{\sup\limits_{\gamma\in\overline{\Gamma}}\left|\left\langle g_\gamma\cdot w_\gamma,\left(f_m^{opt}-f_n\right)\right\rangle\right|} \le \frac{\sup\limits_{\gamma\in\overline{\Gamma}}|w_\gamma|\cdot(1+\eta)\cdot\left\|r_m^{opt}\right\|_2}{\sqrt{\frac{1-\mu_1^w(m-1)-\epsilon_{max}}{m}}\left\|f_m^{opt}-f_n\right\|_2}. \tag{23}$$

Thus, from (23) and (19), a sufficient condition for the recovery of a correct atom can be expressed as:

$$\begin{aligned}
\rho^w(r_n) &\le \frac{\sup\limits_{\gamma\in\overline{\Gamma}}|w_\Gamma|\cdot(1+\eta)\cdot\left\|r_m^{opt}\right\|_2}{\sqrt{\frac{1-\mu_1^w(m-1)-\epsilon_{max}}{m}}\left\|f_m^{opt}-f_n\right\|_2} + \frac{\mu_1^w(m)}{1-\mu_1^w(m-1)-\epsilon_{max}} \\
&= \frac{w_{\overline{\Gamma}}^{max}\sqrt{(1-\mu_1^w(m-1)-\epsilon_{max})\,m}\cdot(1+\eta)\cdot\left\|r_m^{opt}\right\|_2+\left\|f_m^{opt}-f_n\right\|_2\mu_1^w(m)}{(1-\mu_1^w(m-1)-\epsilon_{max})\left\|f_m^{opt}-f_n\right\|_2} < \alpha.
\end{aligned} \tag{24}$$

Considering that $\|f_m^{opt}-f_n\|_2^2=\|r_n\|_2^2-\|r_m^{opt}\|_2^2$, it easily follows that

$$\frac{w_{\Gamma}^{max}\sqrt{(1-\mu_1^w(m-1)-\epsilon_{max})\,m}\cdot(1+\eta)\cdot\left\|r_m^{opt}\right\|_2+\sqrt{\|r_n\|_2^2-(1+\eta)^2\left\|r_m^{opt}\right\|_2^2}\,\mu_1^w(m)}{(1-\mu_1^w(m-1)-\epsilon_{max})\sqrt{\|r_n\|_2^2-(1+\eta)^2\left\|r_m^{opt}\right\|_2^2}}<\alpha. \tag{25}$$

Then, we solve for $\|r_n\|_2^2$:

$$\|r_n\|_2^2>(1+\eta)^2\left\|r_m^{opt}\right\|_2^2\left(1+\frac{\left(w_{\Gamma}^{max}\right)^2(1-\mu_1^w(m-1)-\epsilon_{max})}{\left(\alpha\left(1-\mu_1^w(m-1)-\epsilon_{max}\right)-\mu_1^w(m)\right)^2}\right). \tag{26}$$

For simplicity, let us consider the case where a full search atom selection algorithm is available. Thus, replacing $\alpha=1$ in (26) proves Theorem 5.

■

The general effect of using *prior* knowledge can thus be summarized by the following two Corollaries.

***Corollary** 1:* Let $W(f,\mathcal{D})$ be a reliable *a priori* knowledge and assuming $\alpha=1$, then for any positive integer m such that $\mu_1(m-1)+\mu_1(m)\geq 1$ and $\mu_1^w(m-1)+\mu_1^w(m)<1-\epsilon_{max}$, Weighted-MP/OMP (unlike *Weak*(α)-MP/OMP) will be sure to recover the atoms belonging to the best m-terms approximation f_m^{opt}.

***Corollary** 2:* Let $W(f,\mathcal{D})$ be a reliable a priori knowledge and assuming $\alpha=1$, then for any positive integer m such that $\mu_1(m-1)+\mu_1(m)<1$ and $\mu_1^w(m-1)+\mu_1^w(m)+\epsilon_{max}<\mu_1(m-1)+\mu_1(m)<1$, Weighted-MP/OMP has a weaker sufficient condition than MP/OMP for the recovery of correct atoms from the best m-terms approximant. Hence, the correction factor of the right hand side of expression (16) is smaller for the Weighted-MP/OMP than for the pure greedy algorithm case:

$$\left(1+\frac{m\left(1-(\mu_1^w(m-1)+\epsilon_{max})\left(w_{\Gamma}^{max}\right)^2\right)}{\left(1-(\mu_1^w(m-1)+\mu_1^w(m)+\epsilon_{max})\right)^2}\right)\leq\left(1+\frac{m\,(1-\mu_1(m-1))}{\left(1-(\mu_1(m-1)+\mu_1(m))\right)^2}\right).$$

Therefore, Weighted-MP/OMP is guaranteed to recover better approximants than classic MP/OMP when reliable and good enough *a priori* information is in use.

B. Rate of Convergence of Weighted-MP/OMP on Sparse Approximations

The energy of the series of residuals r_n $(n\geq 0)$ generated by the greedy algorithm progressively converges toward zero as n increases. In the same way, Weighted-MP/OMP with reliable *a priori* information is expected to have a better behavior and a faster convergence rate than the *Weak*-MP for the approximation case. A more accurate estimate of the dictionary coherence conditioned to the signal to be analyzed is available: $\mu_1^w(m)$ (where $\mu_1^w(m)\leq\mu_1(m)$). Then a better bound for the rate of convergence can be found for the case of Weighted-MP/OMP. We follow the path suggested in [2] for OMP and in [4] for the case of general *Weak*-MP algorithm to prove this. As before, we introduce the consideration of the *a priori* information in the formulation. The results formally show how Weighted-MP/OMP can outperform *Weak*-MP when the *prior* knowledge is reliable enough.

***Theorem** 6:* Let $W(f,\mathcal{D})$ be a reliable *a priori* information matrix and $\{r_n\}\ :n\geq 0$ a sequence of residuals produced by Weighted-MP/OMP, then as long as $\|r_n\|_2^2$ satisfies Eq. (16) (Theorem 5), Weighted-MP/OMP picks up a correct atom and

$$\left(\|r_n\|_2^2-\left\|r_m^{opt}\right\|_2^2(1+\eta)^2\right)\leq\left(1-\alpha^2\frac{(1-\mu_1^w(m-1)-\epsilon_{max})}{m}\right)^{n-l}\left(\|r_l\|_2^2-\left\|r_m^{opt}\right\|_2^2(1+\eta)^2\right), \tag{27}$$

where $n\geq l$.

This implies that Weighted-MP/OMP, in the same way as weak-MP, has a exponentially decaying upper bound on the rate of convergence. Moreover, in the case where reliable *a priori* information is used, the bound appears to be lower than in the case where *priors* are not used. This result suggests that the convergence of Weighted greedy algorithms may be faster than for the case of classic pure greedy algorithms.

Proof: Let us consider n such that $\|r_n\|_2^2$ satisfies Eq. (16) of Theorem 5. Then, it is known that for *Weak*-MP:

$$\|r_{n-1}\|_2^2-\|r_n\|_2^2\geq\left|\langle r_n,g_{k_n}\rangle\right|^2, \tag{28}$$

where the inequality applies for OMP, while in the case of MP the equality holds. Moreover, considering the weighted selection, then

$$\|r_{n-1}\|_2^2 - \|r_n\|_2^2 \geq \alpha \sup_{\gamma\in\Gamma} |\langle r_n, g_\gamma \cdot w_\gamma\rangle|^2 \frac{1}{w_\gamma^2} = \alpha \sup_{\gamma\in\Gamma} \left|\left\langle f_m^{opt} - f_n, g_\gamma \cdot w_\gamma\right\rangle\right|^2 \frac{1}{w_\gamma^2}, \tag{29}$$

where the last equality follows from the assumption that Eq. (16) of Theorem 5 is satisfied and because $(f - f_m^{opt}) \perp span(\Gamma)$. And by (22),

$$\|r_{n-1}\|_2^2 - \|r_n\|_2^2 \geq \frac{\alpha}{w_\gamma^2} \frac{\sigma_{min_w}^2}{m} \left\|f_m^{opt} - f_n\right\|_2^2. \tag{30}$$

As stated before, $\|f_m^{opt} - f_n\|_2^2 = \|r_n\|_2^2 - \|r_m^{opt}\|_2^2$, hence $\|f_m^{opt} - f_{n-1}\|_2^2 - \|f_m^{opt} - f_n\|_2^2 = \|r_{n-1}\|_2^2 - \|r_n\|_2^2$, which together with (30) gives:

$$\left\|f_m^{opt} - f_n\right\|_2^2 \leq \left\|f_m^{opt} - f_{n-1}\right\|_2^2 \left(1 - \frac{\alpha}{w_\gamma^2}\frac{\sigma_{min_w}^2}{m}\right) \leq \left\|f_m^{opt} - f_{n-1}\right\|_2^2 \left(1 - \alpha\frac{\sigma_{min_w}^2}{m}\right). \tag{31}$$

Finally, by simply considering $0 \leq l \leq n$ by recursion it follows:

$$\|r_n\|_2^2 - \left\|r_m^{opt}\right\|_2^2 (1+\eta)^2 \leq \left(1 - \alpha\frac{\sigma_{min_w}^2}{m}\right)^{n-l} \left(\|r_l\|_2^2 - \left\|r_m^{opt}\right\|_2^2 (1+\eta)^2\right), \tag{32}$$

and the Theorem is proved.

■

Depending on the sufficient conditions specified in Sec. III-A, the recovery of the optimal set Γ will be guaranteed. However, it is not yet clear how long a non-orthogonalized greedy algorithm (Weighted-MP in our case) will last iterating over the optimal set of atoms in the approximation case. Let us define the number of correct iterations as follows:

***Definition** 3:* Consider a Weighted-MP/OMP algorithm used for the approximation of signals. We define the number of provably correct steps N_m as the smallest positive integer such that

$$\|r_{N_m}\|_2^2 \leq \left\|f - f_m^{opt}\right\|_2^2 (1+\eta)^2 \left(1 + \frac{m\left(1 - (\mu_1^w(m-1) + \epsilon_{max})\left(w_\Gamma^{max}\right)^2\right)}{\left(1 - (\mu_1^w(m-1) + \mu_1^w(m) + \epsilon_{max})\right)^2}\right),$$

which corresponds to the number of atoms belonging to the optimal set that is possible to recover given a signal f, a dictionary and an *a priori* information matrix $W(f, \mathcal{D})$.

In the case of OMP and Weighted-OMP N_m will be always smaller or equal to the cardinality of Γ. For *Weak*-MP and Weighted-MP, provided that $\mu_1(m-1) + \mu_1(m) + \epsilon_{max} < 1$, the probable number of correct iterations will depend on the final error that remains after the best m-term approximation has been found. In the following Theorem, some bounds on the quantity N_m are given for Weighted-MP/OMP. To obtain the results we follow [4].

Before stating the following theorem, the reader should note that from now on, $w_{\Gamma_l}^{max}$ defines the same concept of (13) for an optimal set of atoms Γ of size l, i.e. for Γ_l.

***Theorem** 7:* Let $W(f, \mathcal{D})$ be a reliable *a priori* information and $\{r_n\}\ : n \geq 0$ a sequence of residuals produced by Weighted-MP/OMP when approximating f. Then, for any integer m such that $\mu_1(m-1) + \mu_1(m) + \epsilon_{max} < 1$, we have $N_1 \leq 1$ and for $m \geq 2$:

- if $3\left\|r_1^{opt}\right\|_2^2 \geq m \cdot \left\|r_m^{opt}\right\|_2^2 (1 - \epsilon_{max_m}) \cdot \left(w_{\Gamma_m}^{max}\right)^2$, then

$$2 \leq N_m < 2 + \frac{2 \cdot m}{1 - \epsilon_{max}} \log\left(\frac{3\left\|r_1^{opt}\right\|_2^2}{m \cdot \left\|r_m^{opt}\right\|_2^2 (1 - \epsilon_{max_m}) \cdot \left(w_{\Gamma_m}^{max}\right)^2}\right). \tag{33}$$

- else $N_m \leq 1$.

From (33) we can draw that the upper bound on the number of correct steps N_m is higher for the case of using reliable *a priori* information. This implies that a better behavior of the algorithm is possible with respect to [4]. The term $w_{\Gamma_m}^{max}$ (that depends on the *a priori* information used for Weighted-MP/OMP, $0 \leq w_{\Gamma_m}^{max} \leq 1$ and $w_{\Gamma_m}^{max} = 1$ for the case of classic *Weak*-MP) helps increasing the value of this bound, allowing Weighted-MP to recover a higher number of correct iterations. $w_{\Gamma_m}^{max}$ represents the capacity to discriminate between "good" and "bad" atoms of the *a* *priori* model.

In order to prove Theorem 7, several intermediate results are necessary.

***Lemma** 1:* Let $W(f, \mathcal{D})$ be a reliable *a priori* information and $\{r_n\}\ : n \geq 0$ a sequence of residuals produced by Weighted-MP/OMP, then as long as $\|r_n\|_2^2$ satisfies Eq. (16) of Theorem 5, for any $1 \leq k < m$ such that $N_k < N_m$,

$$N_m - N_k < 1 + \frac{m}{1 - \mu_1^w\,(m-1) - \epsilon_{max}} \left[\log\left(\frac{\left\|r_k^{opt}\right\|_2^2}{\left\|r_m^{opt}\right\|_2^2} \right) + \log\left(\frac{1+\lambda_k^w}{1+\lambda_m^w} \right) \right], \tag{34}$$

where

$$\lambda_l^w \triangleq \frac{l\,(1 - (\mu_1^w(l-1) + \epsilon_{max})) \cdot \left(w_{\overline{\Gamma}_l}^{max} \right)^2}{\left[1 - (\mu_1^w(l-1) + \mu_1^w(l) + \epsilon_{max}) \right]^2}, \tag{35}$$

in which l corresponds to the size of a particular optimal set of atoms ($l = |\Gamma_l|$).

Proof: From Theorem 6, it follows that for $l = N_k$, $n = N_m - 1$, defining

$$\beta_l^w \triangleq 1 - \frac{1 - \mu_1^w(l-1) - \epsilon_{max}}{l}, \tag{36}$$

where l is defined as in (35), and starting from the condition in the residual $\left\|r_{N_{m-1}}\right\|_2^2$ as defined in the Definition 3, the following is accomplished if $\alpha = 1$:

$$\begin{aligned}
\lambda_m^w\,(1+\eta)^2 \left\|r_m^{opt}\right\|_2^2 &< \left\|r_{N_{m-1}}\right\|_2^2 - \left\|r_m^{opt}\right\|_2^2 (1+\eta)^2 \\
&\leq (\beta_m^w)^{N_m - 1 - N_k} \cdot \left(\|r_{N_k}\|_2^2 - \left\|r_m^{opt}\right\|_2^2 (1+\eta)^2 \right) \\
&\leq (\beta_m^w)^{N_m - 1 - N_k} \cdot \left((1+\lambda_k) \left\|r_k^{opt}\right\|_2^2 (1+\eta)^2 - \left\|r_m^{opt}\right\|_2^2 (1+\eta)^2 \right).
\end{aligned} \tag{37}$$

Operating on (37) as in [4], it easily follows that:

$$\left(\frac{1}{\beta_m^w} \right)^{N_m - 1 - N_k} < \frac{\left\|r_k^{opt}\right\|_2^2}{\left\|r_m^{opt}\right\|_2^2} \frac{1+\lambda_k^w}{1+\lambda_m^w},$$

thus,

$$N_m - 1 - N_k \log\left(\frac{1}{\beta_m^w} \right) < \log\left[\frac{\left\|r_k^{opt}\right\|_2^2}{\left\|r_m^{opt}\right\|_2^2} \frac{1+\lambda_k^w}{1+\lambda_m^w} \right].$$

If $t \geq 0$ then $\log(1-t) \leq -t$ and so

$$\frac{1}{\log\left(\frac{1}{\beta_m^w}\right)} \leq \frac{m}{1 - \mu_1(m-1) - \epsilon_{max}}.$$

This proves the result presented in (34) and so the Lemma.

■

In order to use Lemma 1 in Theorem 7, an estimate of the argument of the second logarithm in (34) is necessary. This can be found in the following Lemma.

***Lemma** 2:* For all m such that $\mu_1^w(m-1) + \mu_1^w(m) + \epsilon_{max} < 1$ and $1 \leq k < m$, we have:

$$\lambda_m^w \geq m \cdot \left(w_{\overline{\Gamma}}^{max} \right)^2 \tag{38}$$

$$\frac{\lambda_k^w}{\lambda_m^w} \leq \frac{k}{m} \cdot \frac{(1 - \mu_1^w(k-1) - \epsilon_{max_k}) \cdot \left(w_{\overline{\Gamma}_k}^{max} \right)^2}{(1 - \mu_1^w(m-1) - \epsilon_{max_m}) \cdot \left(w_{\overline{\Gamma}_m}^{max} \right)^2} \tag{39}$$

Proof: Consider the definition of λ_m^w of (35). Then since $\mu_1^w(l-2) + \mu_1^w(l-1) + \epsilon_{max} \leq \mu_1^w(l-1) + \mu_1^w(l) + \epsilon_{max}$ for $2 \leq l \leq m$, the following can be stated:

$$\frac{\lambda_{l-1}^w}{\lambda_l^w} \leq \frac{l-1}{l} \cdot \frac{(1 - \mu_1^w(l-2) - \epsilon_{max_{l-1}}) \cdot \left(w_{\overline{\Gamma}_{l-1}}^{max} \right)^2}{(1 - \mu_1^w(l-1) - \epsilon_{max_l}) \cdot \left(w_{\overline{\Gamma}_l}^{max} \right)^2}. \tag{40}$$

By assuming $k+1 \leq l \leq m$ the Lemma is proved.

$\blacksquare$

Finally, building on the results obtained from Theorem 6 and Lemmas 1 and 2, Theorem 7 can be proved.

Proof: To prove Theorem 7 we need to upper bound the factor $\frac{1+\lambda_k^w}{1+\lambda_m^w}$ in Eq. (34). For this purpose let us consider the following:

$$\frac{(1+\lambda_k^w)}{(1+\lambda_m^w)} \leq \frac{(1+\lambda_k^w)}{(\lambda_m^w)} \leq \frac{1}{\lambda_m^w} + \frac{\lambda_k^w}{\lambda_m^w}. \tag{41}$$

Together with the results of Lemma 2, it gives:

$$\log\left(\frac{1+\lambda_k^w}{1+\lambda_m^w}\right) \leq \log\left(\frac{1}{m} + \frac{k}{m}\cdot\frac{(1-\mu_1^w(k-1)-\epsilon_{max_k})\cdot\left(w_{\Gamma_k}^{max}\right)^2}{(1-\mu_1^w(m-1)-\epsilon_{max_m})\cdot\left(w_{\Gamma_m}^{max}\right)^2}\right). \tag{42}$$

Hence, using Eq. (34) we obtain

$$\begin{aligned} N_m - N_k < 1 + \frac{m}{1-\mu_1^w\,(m-1)-\epsilon_{max}} &\left[\log\left(\frac{\left\|r_k^{opt}\right\|_2^2}{\left\|r_m^{opt}\right\|_2^2}\right) + \cdots\right. \\ &\left.\log\left(\frac{1}{m} + \frac{k}{m}\cdot\frac{(1-\mu_1^w(k-1)-\epsilon_{max_k})\cdot\left(w_{\Gamma_k}^{max}\right)^2}{(1-\mu_1^w(m-1)-\epsilon_{max_m})\cdot\left(w_{\Gamma_m}^{max}\right)^2}\right)\right]. \end{aligned} \tag{43}$$

Theorem 7 is thus proved by particularizing the previous expression for the case where $k=1$. For the case of $N_m \geq N_1 + 1 = 2$, this yields that

$$\begin{aligned} N_m - N_1 &< 1 + \frac{m}{1-\mu_1^w\,(m-1)-\epsilon_{max}}\cdot \\ &\quad \left[\log\left(\frac{\left\|r_1^{opt}\right\|_2^2}{\left\|r_m^{opt}\right\|_2^2}\right) + \log\left(\frac{1}{m} + \frac{(1-\epsilon_{max_1})\cdot\left(w_{\Gamma_1}^{max}\right)^2}{m\,(1-\mu_1^w(m-1)-\epsilon_{max_m})\cdot\left(w_{\Gamma_m}^{max}\right)^2}\right)\right]. \\ &< 1 + \frac{m}{1-\mu_1^w\,(m-1)-\epsilon_{max}}\cdot \\ &\quad \left[\log\left(\frac{\left\|r_1^{opt}\right\|_2^2}{\left\|r_m^{opt}\right\|_2^2}\right) + \log\left(\frac{1}{m} + \frac{1}{m\,(1-\mu_1^w(m-1)-\epsilon_{max_m})\cdot\left(w_{\Gamma_m}^{max}\right)^2}\right)\right]. \end{aligned} \tag{44}$$

Which, since $\mu_1^w(m-1) < \frac{1-\epsilon_{max}}{2}$ and $1-\mu_1^w(m-1)-\epsilon_{max} > \frac{1-\epsilon_{max}}{2}$, then

$$\begin{aligned} N_m &< 2 + \frac{m}{1-\mu_1^w\,(m-1)-\epsilon_{max}}\left[\log\left(\frac{\left\|r_1^{opt}\right\|_2^2}{\left\|r_m^{opt}\right\|_2^2}\right) + \log\left(\frac{2+(1-\epsilon_{max_m})\cdot\left(w_{\Gamma_m}^{max}\right)^2}{m\,(1-\epsilon_{max_m})\cdot\left(w_{\Gamma^m}^{max}\right)^2}\right)\right] \\ &< 2 + \frac{m}{1-\mu_1^w\,(m-1)-\epsilon_{max}}\left[\log\left(\frac{\left\|r_1^{opt}\right\|_2^2}{\left\|r_m^{opt}\right\|_2^2}\right) + \log\left(\frac{3}{m\,(1-\epsilon_{max_m})\cdot\left(w_{\Gamma_m}^{max}\right)^2}\right)\right] \\ &< 2 + \frac{m}{1-\mu_1^w\,(m-1)-\epsilon_{max}}\log\left(\frac{3\left\|r_1^{opt}\right\|_2^2}{m\cdot\left\|r_m^{opt}\right\|_2^2(1-\epsilon_{max_m})\cdot\left(w_{\Gamma_m}^{max}\right)^2}\right) \\ &< 2 + \frac{2\cdot m}{1-\epsilon_{max}}\log\left(\frac{3\left\|r_1^{opt}\right\|_2^2}{m\cdot\left\|r_m^{opt}\right\|_2^2(1-\epsilon_{max_m})\cdot\left(w_{\Gamma_m}^{max}\right)^2}\right). \end{aligned} \tag{45}$$

This is only possible if $3\left\|r_1^{opt}\right\|_2^2 \geq m\cdot\left\|r_m^{opt}\right\|_2^2(1-\epsilon_{max_m})\cdot\left(w_{\Gamma_m}^{max}\right)^2$.

$\blacksquare$

Compared to the case where no *a priori* information is available [4], the condition for the validity of bound (33) is softened in our case. Moreover, the upper bound on N_m is increased, which means that there is room for an improvement on the number of correct iterations Indeed, under the assumption of having a reliable *a priori* information, the smaller $w_{\overline{\Gamma}_m}^{max}$ (which implies a good discrimination between Γ and $\overline{\Gamma}$) the easier it is to fulfill the condition stated in Theorem 7.

C. Example: Use of Footprints for ϵ-Sparse Approximations.

To give an example of the approximation of signals using *a priori* information, we consider again the case presented in [1] where a piecewise-smooth signal is represented by means of an overcomplete dictionary composed by the mixture of an orthonormal wavelet basis and a family of wavelet footprints (see [16]).

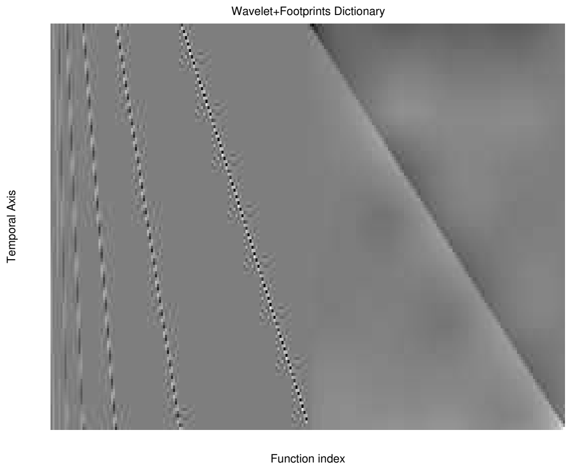

Fig. 1. Dictionary formed by the Symmlet-4 [17] (left half) and its respective footprints for piecewise constant singularities (right half).

Let us remind the considerations for the dictionary. The dictionary is built by the union of an orthonormal basis defined by the *Symmlet*-4 family of wavelets [17] and the respective family of footprints for all the possible translations of the Heaviside function. The latter is used to model the discontinuities. The graphical representation of the dictionary matrix can be seen in Fig. 1 where the columns are the waveforms that compose the dictionary.

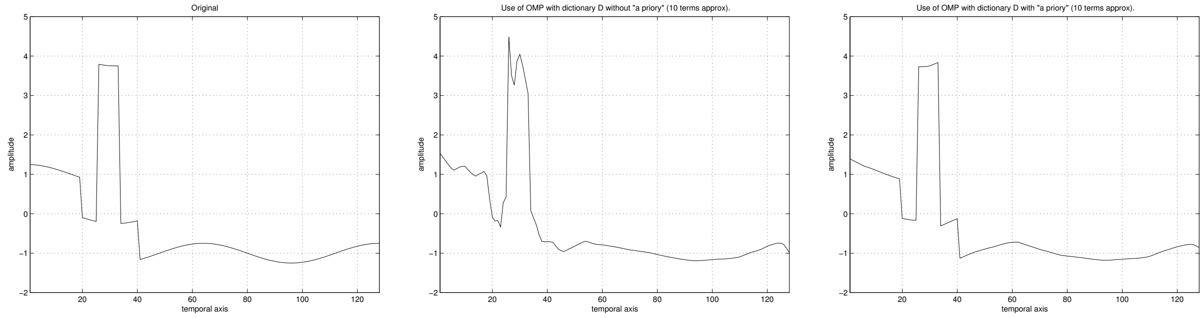

Fig. 2. Comparison of OMP based approximation with 10 terms using the footprints dictionary (Fig. 1). Left: Original signal. Middle: "blind" OMP approximation. Right: OMP with prior knowledge of the footprints location.

The use of such a dictionary, indeed, does not satisfy at all the sufficient condition required to ensure the recovery of an optimal approximant with more than one term. Moreover, even if the best *a priori* was available, it is also far from satisfying the sufficient condition based on the weighted Babel function. Nevertheless, such an example is considered in this section because of two main reasons. The first concerns the fact that sufficient theoretical conditions exposed in the literature are very pessimistic and reflect the worst possible case. The second reason is that, as previously discussed, experience seems to teach us that good dictionaries for efficient approximation of signals, are likely to be highly coherent. This fact conflicts with the requirement of incoherence for the good behavior of greedy algorithms. Hence, we find this example of special interest to underline the benefits of using *a priori* information and additional signal modeling for non-linear expansions. Indeed, with this example it is shown that, by using reliable *a priori* knowledge, better approximations are possible, not only with incoherent dictionaries (where theoretical evidences of the improvement have been shown in this paper) but also with highly coherent ones.

We repeat the procedure used in [1] to estimate the *a priori* information based on the dictionary and the input data. We also refer the reader to Sec. V for a more detailed explanation.

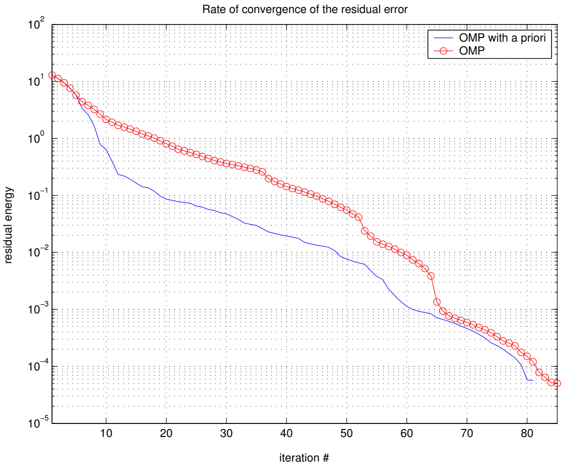

Fig. 3. Rate of convergence of the error with respect to the iteration number in the experiment of Fig. 2

Fig. 2 presents the original signal (left) together with the two approximations obtained in this example: without *a priori* in the middle and with *a priori* at the right. The signal to be approximated has a higher number of polynomial degrees than the number of vanishing moments of the *Symmlet*-4. The figures depict clearly the positive effect of the reliable *a priori* information inserted in the Weighted-OMP algorithm. Indeed, with very few components, the algorithm benefits from the *a priori* information estimated from the signal, and gives a much better approximation. A more global view of this behavioral enhancement can be seen in Fig. 3 where the rate of convergence of the approximation error is presented. The use of weights is definitively helpful and a considerable gain in the reduction of the approximation error is achieved for a small number of terms.

IV. Approximations with Weighted Basis Pursuit Denoising

In this section, the problem of finding a sparse approximation of a signal f is addressed considering a trade-off between the error and the number of elements that participate to the approximation. In statistics this problem is also called Subset Selection and we will refer to it as P_0:

$$(P_0) \qquad \min_{\mathbf{c}} \|f - D\mathbf{c}\|_2^2 + \tau^2 \|\mathbf{c}\|_0. \tag{46}$$

Solving P_0 is NP complex, so a possible way of simplifying the computation can be to substitute the ℓ_0 quasi-norm with the convex ℓ_1 norm. This relaxation leads to the following problem that, from now on, is called P_1:

$$(P_1) \qquad \min_{\mathbf{b}} \frac{1}{2} \|f - D\mathbf{b}\|_2^2 + \gamma \|\mathbf{b}\|_1. \tag{47}$$

This new problem corresponds to the minimization of a convex function that can be solved with classical Quadratic Programming methods. This relaxation is similar to the one that leads to the definition of the Basis Pursuit principle for the case of exact signal representation. The fact that this paradigm is also called Basis Pursuit Denoising can be explained because it was introduced to adapt BP to the case of noisy data [13]. Note that if $\mathcal{D}$ is orthonormal the solution of P_1 can be found by a soft shrinkage of the coefficients [15], [13], while, if $\mathcal{D}$ is a union of orthonormal subdictionaries, the problem can be solved recurring to the Block Coordinate Relaxation method [18], faster than Quadratic Programming.

In [1] we introduced a theoretical framework for sparse representation over redundant dictionaries taking into account some *a priori* information about the signal. In this optic we proposed the Weighted Basis Pursuit method that minimizes a cost function that includes weights expressing the *a priori* information:

$$\min_{\mathbf{b}} \|W^{-1}\mathbf{b}\|_1 \quad \text{s.t.} \quad D\mathbf{b} = f. \tag{48}$$

The main results in [1] concerning WBP are contained in Proposition 1:

Definition 4: Given a dictionary $\mathcal{D}$ indexed in Ω and an index subset $\Lambda \subset \Omega$, we define the Weighted Recovery Factor (WRF) as:

$$WRF(\Lambda) = \sup_{i \notin \Lambda} \left\| (D_\Lambda W_\Lambda)^+ g_i \cdot w_i \right\|_1. \tag{49}$$

***Proposition** 1:* Given a dictionary $\mathcal{D}$ and an *a priori* matrix $W(f, \mathcal{D})$, Weighted Basis Pursuit is able to recover the optimal representation of a sparse signal $f = D_\Gamma \mathbf{b}_\Gamma$ if the following Exact Recovery Condition is respected:

$$WRF(\Gamma) < 1. \tag{50}$$

Moreover, a bound for the WRF is:

$$WRF(\Gamma) < \frac{\mu_1^w(m)}{1 - \epsilon_{max} - \mu_1^w(m-1)}, \tag{51}$$

where $\epsilon_{max} = \sup_{\gamma \in \Gamma} \left|1 - w_\gamma^2\right|$ and w_γ are the elements of the diagonal matrix W_Γ. Therefore, the Exact Recovery Condition for WBP (50) holds for any index set Γ of size at most m such that

$$\mu_1^w(m) + \mu_1^w(m-1) < 1 - \epsilon_{max}. \tag{52}$$

A. A Bayesian Approach to Weighted Basis Pursuit Denoising

In this subsection the problem of signal approximation is studied from a Bayesian point of view. We also examine under which hypotheses BPDN finds the optimal solution. This leads us to generalize the BPDN principle thorugh the definition of Weighted Basis Pursuit Denoising (WBPDN). Let us write again the model of our data approximation, where $\hat{f}$ is the approximant and r is the residual:

$$f = \hat{f} + r = D\mathbf{b} + r. \tag{53}$$

Assuming r to be an iid Gaussian set of variables, the data likelihood is

$$p(f|D, \mathbf{b}) = \frac{1}{\sqrt{2\pi\sigma_r^2}} \cdot \exp\left(-\frac{\|f - D\mathbf{b}\|_2^2}{2\sigma_r^2}\right), \tag{54}$$

where σ_r^2 is the variance of the residual. In the approximation problem, one aims at maximizing the likelihood of $p(\mathbf{b}|f, D)$. Formally, by the Bayes rule, we have

$$p(\mathbf{b}|f, D) = \frac{p(f|D, \mathbf{b}) \cdot p(\mathbf{b})}{p(f, D)},$$

and thus, assuming $p(f, D)$ to be uniform, it follows that the most probable signal representation is:

$$\mathbf{b}_P = \arg\max_{\mathbf{b}} p(f|D, \mathbf{b}) \cdot p(\mathbf{b}). \tag{55}$$

Let us now assume the coefficients b_i are independent and have a Laplacian distribution with standard deviation σ_i:

$$p(b_i) = \frac{1}{\sqrt{2}\sigma_i} \cdot \exp\left(-\frac{\sqrt{2}|b_i|}{\sigma_i}\right)$$

From (55), by computing the logarithm, it follows that

$$\mathbf{b}_P = \arg\max_{\mathbf{b}} \left(\ln(p(f|D, \mathbf{b})) + \sum_i \ln p(b_i)\right) = \arg\min_{\mathbf{b}} \left(\frac{\|f - D\mathbf{b}\|_2^2}{2\sigma_r^2} + \sum_i \frac{\sqrt{2}|b_i|}{\sigma_i}\right).$$

Making the hypothesis that σ_i is constant for every index i, the previous equation means that the most probable $\mathbf{b}$ is the one found by the BPDN algorithm [19]. In fact, this hypothesis does not often correspond to reality. On the contrary, if the variances of the coefficients are not forced to be all the same, it turns out that the most probable signal representation can be found by solving the following problem:

$$(P_1^w) \qquad \min_{\mathbf{b}} \frac{1}{2} \|f - D\mathbf{b}\|_2^2 + \gamma \|W^{-1}\mathbf{b}\|_1, \tag{56}$$

where the diagonal matrix with entries in $(0, 1]$ is defined in Section III. One can notice that in Eq. (56), the introduction of weights allows to individually model the components of $\mathbf{b}$. This approach is analogous to the one introduced in [20] and [1] and, from now on, we will refer to P_1^w as Weighted Basis Pursuit Denoising or WBPDN.

The assumption often made about the Gaussianity of the residual is quite restrictive. However, for another particular problem, one could make the hypothesis that this residual has a Laplacian distribution. It is then possible to prove that the most probable signal representation can be found substituting the ℓ_2 measure of the error with the ℓ_1. This leads to the following minimization problem:

$$\min_{\mathbf{b}} \frac{1}{2} \|f - D\mathbf{b}\|_1 + \gamma \|W^{-1}\mathbf{b}\|_1,$$

where $W = I$ if the variances of the probability density functions of b_i are the same for each i. This problem is faced, for example in [20], where it is also explained that it can be solved by Linear Programming techniques.

B. Preliminary Propositions

Here some preliminary propositions are presented, allowing us to prove the results of the following two subsections. The proofs of most of them follow the arguments given by *Tropp* in [3]. In the following $\mathbf{c}_\Lambda$ and $\mathbf{b}_\Lambda$ lay in $\mathbb{R}^\Lambda$ but sometimes these are extended to $\mathbb{R}^\Omega$ by padding with zeros. The same is valid for the matrix W_Λ.

Next proposition, similar to the Correlation Condition Lemma in [3], establishes a fundamental result for the rest of the report: it basically states that, if the atoms of Λ have a small weighted coherence, expressed by the Weighted Recovery Factor, then the support of any vector that solves P_1^w is a subset of Λ.

***Lemma** 3:* Given an index subset $\Lambda \subset \Omega$, suppose that the following condition is satisfied:

$$\|D^T(f - a_\Lambda)\|_\infty < \frac{\gamma}{w_{\overline{\Gamma}}^{max}} \cdot (1 - WRF(\Lambda)), \tag{57}$$

where $w_{\overline{\Lambda}}^{max} \in (0, 1]$ is the quantity defined by equation (13). Then, any coefficient vector $\mathbf{b}_*$ that minimizes the cost function of problem P_1^w must satisfy

$$support(\mathbf{b}_*) \subset \Lambda. \tag{58}$$

Proof: Assume that $\mathbf{b}_*$ is a vector minimizing (56). Assume also that it uses an index outside Λ. $\mathbf{b}_*$ can be compared with its projection $D_\Lambda^+ D\mathbf{b}_*$, which is supported in Λ and we obtain:

$$\frac{1}{2}\|f - D\mathbf{b}_*\|_2^2 + \gamma\left\|W^{-1}\mathbf{b}_*\right\|_1 \le \frac{1}{2}\left\|f - DD_\Lambda^+ D\mathbf{b}_*\right\|_2^2 + \gamma\left\|W_\Lambda^{-1}(D_\Lambda^+ D\mathbf{b}_*)\right\|_1,$$

which gives

$$2\gamma\left(\left\|W^{-1}\mathbf{b}_*\right\|_1 - \left\|W_\Lambda^{-1}(D_\Lambda^+ D\mathbf{b}_*)\right\|_1\right) \le \left\|f - DD_\Lambda^+ D\mathbf{b}_*\right\|_2^2 - \|f - D\mathbf{b}_*\|_2^2. \tag{59}$$

First we shall provide a lower bound on the left-hand side of the previous inequality. Let us split the vector $\mathbf{b}_*$ into two parts: $\mathbf{b}_* = \mathbf{b}_\Lambda + \mathbf{b}_{\overline{\Lambda}}$, where the former vector contains the components with indexes in Λ, while the latter the remaining components from $\Omega \setminus \Lambda$. This yields, by the upper triangular inequality, that

$$\begin{aligned}\left\|W^{-1}\mathbf{b}_*\right\|_1 - \left\|W_\Lambda^{-1}(D_\Lambda^+ D\mathbf{b}_*)\right\|_1 &= \left\|W^{-1}\mathbf{b}_\Lambda\right\|_1 + \left\|W^{-1}\mathbf{b}_{\overline{\Lambda}}\right\|_1 - \left\|W^{-1}\mathbf{b}_\Lambda + W_\Lambda^{-1}D_\Lambda^+ D\mathbf{b}_{\overline{\Lambda}}\right\|_1 \\ &\ge \left\|W^{-1}\mathbf{b}_{\overline{\Lambda}}\right\|_1 - \left\|W_\Lambda^{-1}D_\Lambda^+ DWW^{-1}\mathbf{b}_{\overline{\Lambda}}\right\|_1.\end{aligned}$$

Since

$$\left\|W_\Lambda^{-1}D_\Lambda^+ DWW^{-1}\mathbf{b}_{\overline{\Lambda}}\right\|_1 \le \sup_{i \notin \Lambda}\left\|(D_\Lambda W_\Lambda)^+ g_i w_i\right\|_1 \cdot \left\|W^{-1}\mathbf{b}_{\overline{\Lambda}}\right\|_1,$$

using (49), one can write that

$$\left\|W^{-1}\mathbf{b}_*\right\|_1 - \left\|W_\Lambda^{-1}(D_\Lambda^+ D\mathbf{b}_*)\right\|_1 \ge (1 - WRF(\Lambda)) \cdot \left\|W^{-1}\mathbf{b}_{\overline{\Lambda}}\right\|_1. \tag{60}$$

We now provide an upper bound for the right-hand side of (59). This quantity does not depend on the weighting matrix, thus, exactly as in [3], it can be stated that:

$$\left\|f - DD_\Lambda^+ D\mathbf{b}_*\right\|_2^2 - \|f - D\mathbf{b}_*\|_2^2 \le 2\|\mathbf{b}_{\overline{\Lambda}}\|_1 \cdot \left\|D^T(f - \mathbf{a}_\Lambda)\right\|_\infty. \tag{61}$$

From (59), (60) and (61) it turns out that:

$$\gamma\,(1 - WRF(\Lambda)) \cdot \left\|W^{-1}\mathbf{b}_{\overline{\Lambda}}\right\|_1 \le \|\mathbf{b}_{\overline{\Lambda}}\|_1 \cdot \left\|D^T(f - \mathbf{a}_\Lambda)\right\|_\infty. \tag{62}$$

Since the weights are in $(0, 1]$, and the vector $\mathbf{b}_{\overline{\Gamma}}$, by assumption, cannot be null, it can be written:

$$\gamma\,(1 - WRF(\Lambda)) \le \frac{\|\mathbf{b}_{\overline{\Lambda}}\|_1}{\left\|W^{-1}\mathbf{b}_{\overline{\Lambda}}\right\|_1} \cdot \left\|D^T(f - a_\Lambda)\right\|_\infty \le w_{\overline{\Lambda}}^{max} \cdot \left\|D^T(f - a_\Lambda)\right\|_\infty. \tag{63}$$

If (57) is valid , then (63) fails and so one must discard the hypothesis that $\mathbf{b}_*$ is non-zero for an index in $\overline{\Lambda} = \Omega \setminus \Lambda$. ■

We now focus on finding a necessary and sufficient condition for the existence and unicity of a minimum of P_1^w. The presence of the ℓ_1 norm imply that the cost function of this problem is non-smooth at zero: for this reason the concept of *subdifferential* is used. Given a real vector variable $\mathbf{x}$, the subdifferential of $\|\mathbf{x}\|_1$ is denoted by $\partial\|\mathbf{x}\|_1$ and defined as:

$$\partial\|\mathbf{x}\|_1 \triangleq \{\mathbf{u}|\mathbf{u}^*\mathbf{x} = \|\mathbf{x}\|_1, \|\mathbf{u}\|_\infty \le 1\}.$$

The vectors $\mathbf{u}$ that compose the subdifferential are called *subgradients* [21].

Lemma 4: A necessary and sufficient condition for $\mathbf{b}_*$ to globally minimize the objective function of P_1^w over all the coefficient vectors with support Λ is that:

$$\mathbf{c}_\Lambda - \mathbf{b}_* = \gamma \left(D_\Lambda^T D_\Lambda\right)^{-1} W_\Lambda^{-1}\mathbf{u}, \tag{64}$$

where $\mathbf{u}$ is a vector from $\partial\|\mathbf{b}_*\|_1$. Moreover, the minimizer is unique.

Proof: One can observe that solving P_1^w is equivalent to minimize the following function over coefficient vectors from $\mathbb{R}^\Lambda$:

$$F(\mathbf{b}) \triangleq \frac{1}{2}\|a_\Lambda - D_\Lambda \mathbf{b}\|_2^2 + \gamma\|W_\Lambda^{-1}\mathbf{b}\|_1.$$

A point $\mathbf{b}_*$ minimizes the second term of $F(\mathbf{b})$ if and only if the following Fermat criterion holds (see [21], [22]):

$$\mathbf{0} \in \partial\|\mathbf{b}_*\|_1.$$

Moreover, $\mathbf{b}_*$ minimizes $F(\mathbf{b})$ if and only if $\mathbf{0} \in \partial F(\mathbf{b}_*)$. In our case this means that

$$\exists\ \mathbf{u} \in \partial\|\mathbf{b}_*\|_1 \quad \text{s.t.} \quad D_\Lambda^T D_\Lambda \mathbf{b}_* - D_\Lambda^T \mathbf{a}_\Lambda + \gamma W_\Lambda^{-1}\mathbf{u} = 0, \tag{65}$$

for some vector $\mathbf{u}$ taken from $\partial\|\mathbf{b}_*\|_1$. Let the atoms in Λ be linearly independent, from (65) it follows:

$$\mathbf{b}_* - \left(D_\Lambda^T D_\Lambda\right)^{-1} D_\Lambda^T \mathbf{a}_\Lambda + \gamma \left(D_\Lambda^T D_\Lambda\right)^{-1} W_\Lambda^{-1}\mathbf{u} = 0,$$

and so

$$D_\Lambda^+ \mathbf{a}_\Lambda - \mathbf{b}_* = \gamma \left(W_\Lambda D_\Lambda^T D_\Lambda\right)^{-1} \mathbf{u}.$$

To conclude the proof it is sufficient to recall that $\mathbf{c}_\Lambda = D_\Lambda^+ \mathbf{a}_\Lambda$.

■

If $W = I$, then this result coincides with the one developed by *Fuchs* in [23] and by *Tropp* in [3] in the complex case.

Lemma 5: Suppose that $\mathbf{b}_*$ minimizes the cost function of problem P_1^w. Then the following bound holds:

$$\|\mathbf{c}_\Lambda - \mathbf{b}_*\|_\infty \leq \frac{\gamma}{w_\Lambda^{min}} \cdot \left\|\left(D_\Lambda^T D_\Lambda\right)^{-1}\right\|_{\infty,\infty}, \tag{66}$$

where w_Λ^{min} is defined as

$$w_\Lambda^{min} \triangleq \inf_{i\in\Lambda} w_i. \tag{67}$$

Proof: Let us consider the necessary and sufficient condition of Lemma 4: taking the ℓ_∞ norm of (64) we obtain:

$$\|\mathbf{c}_\Lambda - \mathbf{b}_*\|_\infty = \gamma \left\|\left(D_\Lambda^T D_\Lambda\right)^{-1} W_\Lambda^{-1}\mathbf{u}\right\|_\infty \leq \gamma \left\|\left(D_\Lambda^T D_\Lambda\right)^{-1} W_\Lambda^{-1}\right\|_{\infty,\infty} \cdot \|\mathbf{u}\|_\infty$$

By definition of subdifferential, $\|\mathbf{u}\|_\infty \leq 1$. Inserting this into the previous equation and using the sub-multiplicative property of matrix norms ($\|AB\|_{p,q} \leq \|A\|_{p,q} \cdot \|B\|_{p,q}$), we can prove that

$$\|\mathbf{c}_\Lambda - \mathbf{b}_*\|_\infty \leq \gamma \left\|\left(D_\Lambda^T D_\Lambda\right)^{-1}\right\|_{\infty,\infty} \cdot \|W_\Lambda^{-1}\|_{\infty,\infty}.$$

Just apply the fact that $\|W_\Lambda^{-1}\|_{\infty,\infty} = \sup_{i\in\Lambda}(1/w_i) = \dfrac{1}{w_\Lambda^{min}}$ to reach the result.

■

The following proposition states a result that will be used in next subsection to prove Theorem 8. Note that here Γ is the optimal index subset of Ω, and thus $\mathbf{c}_\Gamma$ is the sparsest solution to the subset selection problem.

Proposition 2: *(Tropp [3])* Given an input signal f and a threshold τ, suppose that the coefficient vector $\mathbf{c}_\Gamma$, having support Γ of cardinality m, is the sparsest solution of the problem P_0. Set $f_m^{opt} = D\mathbf{c}_\Gamma$. Then:

- $\forall k \in \Gamma$, $|\mathbf{c}_\Gamma(k)| \geq \tau$.
- $\forall i \notin \Gamma$, $|\langle f - f_m^{opt}, g_i\rangle| < \tau$.

These preliminary statements of this subsection will allow us to obtain the main results in the following of the report.

C. Weighted Relaxed Subset Selection

Let us now study the relationship between the results obtained by solving problem P_1^w and P_0. Suppose that $\mathbf{c}_\Gamma$ is the sparsest solution to P_0 and that its support is Γ, with $|\Gamma| = m$. D_Γ will be the matrix containing all the atoms participating to the sparsest approximation of f and f_m^{opt} will be the approximant given by $\mathbf{c}_\Gamma$, i.e $f_m^{opt} = D\mathbf{c}_{opt} = DD_\Gamma^+ f = D_\Gamma D_\Gamma^+ f$. Assuming $WRF(\Gamma) < 1$, we have the following result.

***Theorem** 8:* Suppose that $\mathbf{b}_*$ minimizes the cost function of problem P_1^w with threshold

$$\gamma = \frac{\tau \cdot w_\Gamma^{max}}{1 - WRF(\Gamma)},$$

where w_Γ^{max} is defined in (13). Then:

1) WBP never selects a non optimal atom since *support*$(\mathbf{b}_*) \subset \Gamma$.
2) The solution of WBP is unique.
3) The following upper bound is valid:

$$\|\mathbf{c}_\Gamma - \mathbf{b}_*\|_\infty \leq \frac{\tau \cdot \frac{w_\Gamma^{max}}{w_\Gamma^{min}} \cdot \left\| \left(D_\Gamma^T D_\Gamma\right)^{-1} \right\|_{\infty,\infty}}{1 - WRF(\Gamma)}. \tag{68}$$

4) The support of $\mathbf{b}_*$ contains every index j for which

$$|\mathbf{c}_\Gamma(j)| > \frac{\tau \cdot \frac{w_\Gamma^{max}}{w_\Gamma^{min}} \cdot \left\| \left(D_\Gamma^T D_\Gamma\right)^{-1} \right\|_{\infty,\infty}}{1 - WRF(\Gamma)}. \tag{69}$$

The scalar w_Γ^{min} appearing in Eqs. (68) and (69) is defined in Eq. (67).

Proof: Considering the first stated result, note that every atom indexed by Γ has zero inner product with the optimal residual ($r_m^{opt} = f - f_m^{opt}$) since f_m^{opt} is the best approximation of f using the atoms in Γ. Using Proposition 2 and recalling that $\mathcal{D}$ is finite, it can be stated that

$$\left\| D^T (f - f_m^{opt}) \right\|_\infty < \tau. \tag{70}$$

Moreover, Lemma 3 guarantees that

$$\forall \gamma \quad \text{s.t.} \quad \left\| D^T (f - f_m^{opt}) \right\|_\infty < \frac{\gamma}{w_\Gamma^{max}} \cdot (1 - WRF(\Gamma)), \tag{71}$$

and that the solution $\mathbf{b}_*$ to the convex problem P_1^w is supported on Γ. From (70) and (71) it follows that for any γ that satisfies the following condition, it is insured that *support*$(\mathbf{b}_*) \subset \Gamma$:

$$\gamma \geq \frac{\tau \cdot w_\Gamma^{max}}{1 - WRF(\Gamma)}. \tag{72}$$

In the following, the smallest possible value for γ is chosen: in this way, Eq. (72) becomes an equality.

The uniqueness of the solution follows from Lemma 4.

With regard to the third point, the results from Lemma 5 yield

$$\|\mathbf{c}_\Gamma - \mathbf{b}_*\|_\infty \leq \frac{\gamma}{w_\Gamma^{min}} \left\| \left(D_\Gamma^T D_\Gamma\right)^{-1} \right\|_{\infty,\infty} \leq \frac{\tau \cdot \frac{w_\Gamma^{max}}{w_\Gamma^{min}}}{1 - WRF(\Gamma)} \cdot \left\| \left(D_\Gamma^T D_\Gamma\right)^{-1} \right\|_{\infty,\infty}.$$

This proves the point 3.

Concerning the fourth result of the theorem, one can observe that, for every index j for which

$$|\mathbf{c}_\Gamma(j)| > \gamma \left\| \left(D_\Gamma^T D_\Gamma\right)^{-1} W_\Gamma^{-1} \right\|_{\infty,\infty},$$

the corresponding coefficient $\mathbf{b}_*(j)$ must be different form zero. As before, substituting Eq. (72) leads to prove the last result of the Theorem.

■

This theorem states two important concepts. First, if the trade off parameter is correct and the weighted cumulative coherence of the dictionary is small enough, WBPDN is able to select the correct atoms for the signal expansion and it will not pick up the wrong ones. Furthermore, the error achieved by the algorithm on the amplitude of the coefficients related to the selected atoms is bounded. Nevertheless, once the algorithm has recovered the atom subset, the appropriate amplitudes of the coefficients can be computed by the orthogonal projection of the signal onto the space generated by the selected atoms. This method is illustrated in Section IV-E to generate some examples.

The quantities w_Γ^{min} and w_Γ^{max} depend on the reliability and goodness of the *a priori*. In particular, if W tends to be optimal (i.e. its diagonal entries tend to 1 for the elements that should appear in the sparsest approximation and to 0 for the ones that should not), one can observe that $w_\Gamma^{min} \to 1$ and $w_\Gamma^{max} \to 0$.

D. Relation with the Weighted Cumulative Coherence

In this subsection, the previous results are described using the weighted cumulative coherence function defined in (15). In this way a comparison is made between the results achievable by BPDN and WBPDN.

***Theorem** 9:* Assume that the real vector $\mathbf{b}_*$ solves P_1^w with

$$\gamma = \frac{w_\Gamma^{max} \cdot \tau(1 - \epsilon_{max} - \mu_1^w(m-1))}{1 - \epsilon_{max} - \mu_1^w(m) - \mu_1^w(m-1)}.$$

Then *support*$(\mathbf{b}_*) \subset \Gamma$ and

$$\|\mathbf{b}_* - \mathbf{c}_\Gamma\|_\infty \leq \frac{\tau \cdot \frac{w_\Gamma^{max}}{w_\Gamma^{min}}(1 - \epsilon_{max} - \mu_1^w(m-1))}{(1 - \epsilon_{max} - \mu_1^w(m) - \mu_1^w(m-1))(1 - \mu_1(m-1))}. \tag{73}$$

Proof: This result can be obtained from Proposition 1 and Theorem 8, since:

$$\|\mathbf{b}_* - \mathbf{c}_\Gamma\|_\infty \leq \frac{\gamma}{w_\Gamma^{min}} \left\| \left(D_\Gamma^T D_\Gamma\right)^{-1} \right\|_{\infty,\infty} = \frac{\tau \cdot \frac{w_\Gamma^{max}}{w_\Gamma^{min}}(1 - \epsilon_{max} - \mu_1^w(m-1)) \cdot \left\| \left(D_\Gamma^T D_\Gamma\right)^{-1} \right\|_{\infty,\infty}}{(1 - \epsilon_{max} - \mu_1^w(m) - \mu_1^w(m-1))}.$$

The last term of the numerator in the previous expression is the norm of the inverse Gram matrix. Since

$$\| \left(D_\Gamma^T D_\Gamma\right)^{-1} \|_{\infty,\infty} = \| \left(D_\Gamma^T D_\Gamma\right)^{-1} \|_{1,1} \leq \frac{1}{1 - \mu_1(m-1)},$$

(see [3], [23], [1]) this proves equation (73).

■

This result is valid in general and illustrates how the distance between the optimal signal approximation and the solution found by solving P_1^w can be bounded. In case no *a priori* is given, the bound on the coefficient error is obtained from Eq. (73) setting $W = I$. Consequently, $w_\Gamma^{min} = 1, \epsilon_{max} = 0$ and $w_\Gamma^{max} = 1$ (see also [3]):

$$\|\mathbf{b}_* - \mathbf{c}_\Gamma\|_\infty \leq \frac{\tau}{1 - \mu_1(m) - \mu_1(m-1)}. \tag{74}$$

Comparing the two bounds, one can observe how the availability of a reliable *prior* on the signal can help in finding a sparser signal approximation. This concept is emphasized in the following corollary.

***Corollary** 3:* Let $W(f, \mathcal{D})$ be a reliable a priori knowledge, with

$$\frac{w_\Gamma^{max}}{w_\Gamma^{min}} \leq 1,$$

then for any positive integer m such that $\mu_1(m-1)+\mu_1(m) < 1$ and $\mu_1^w(m-1)+\mu_1^w(m)+\epsilon_{max} < \mu_1(m-1)+\mu_1(m) < 1$, the error $\|\mathbf{b}_* - \mathbf{c}_\Gamma\|_\infty$ given by the coefficients found by WBPDN is smaller than the one obtained by BPDN.

Hence, the bound stated by Eq. (73) is lower than the one in Eq. (74), i.e.

$$\frac{\tau \cdot \frac{w_\Gamma^{max}}{w_\Gamma^{min}}(1 - \epsilon_{max} - \mu_1^w(m-1))}{(1 - \epsilon_{max} - \mu_1^w(m) - \mu_1^w(m-1))(1 - \mu_1(m-1))} \leq \frac{\tau}{1 - \mu_1(m) - \mu_1(m-1)}.$$

Note the similarity between Corollaries 2 and 3. The proof of the latter Corollary is reported in the appendix. Here the hypothesis that $\frac{w_\Gamma^{max}}{w_\Gamma^{min}} \leq 1$ is made. Observe that if the *a priori* is particularly good this factor can be much smaller than one and so the improvement with respect to the general case can be really big.

E. Example

We examine again the example presented in section III-C, but this time using the Basis Pursuit Denoising and Weighted Basis Pursuit Denoising algorithms. The dictionary used for the decomposition is illustrated in section III-C, as well as the input signal. For an explanation of the prior model and the extraction of the *a priori* matrix, see [1] or Section V.

The signal f is decomposed first solving the minimization problem illustrated in (47) (BPDN). Then, the *a priori* knowledge is introduced, and we solve the minimization problem illustrated in (56) (WBPDN). Both solutions were numerically found using Quadratic Programming techniques. The trade off parameter γ controls the ℓ_1 norm of the coefficient vector and indirectly its sparseness. The signal representations present many components with negligible values due to the numerical computation: a hard thresholding is performed in order to get rid of this misleading information, paying attention not to eliminate significant elements. In this optic it is possible to measure the ℓ_0 norm of the vector $\mathbf{b}$. The data reported here refer to a threshold value of 0.01. Of course the reconstructions are computed starting from the thresholded coefficients. Figure 4 shows the reconstructions of the input signal given by a 10-terms approximation found by BPDN and WBPDN. Figure 6 (on the left-hand) illustrates the mean square error of the approximations of f with m terms found by BPDN and WBPDN.

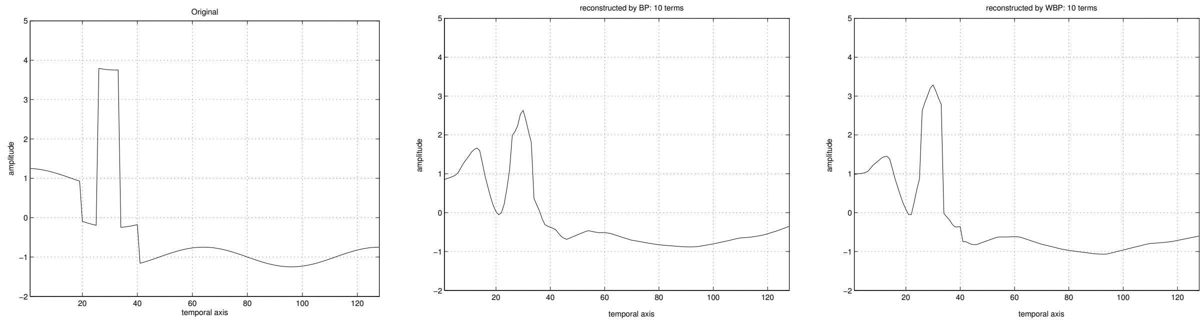

Fig. 4. Comparison of BPDN and WBPDN based approximation with 10 terms using the footprints dictionary (Fig. 1). Left: Original signal. Center: reconstruction using 10 BPDN coefficients. Right: reconstruction using 10 WBPDN coefficients.

Let us call $\mathbf{b}_*$ the approximation found by BPDN and $\mathbf{b}_*^w$ the one found by WBPDN. As just explained, these vectors are thresholded removing the numerically negligible components, and in this way we are able to individuate a sparse support and so a subset of the dictionary. Let us label the subdictionary found by WBPDN with $\mathcal{D}_*^w$ (composed by the atoms corresponding to the non-zero elements of $\mathbf{b}_*^w$). Once this is given, there are no guarantees that the coefficients that represent f are optimal (see Theorems 8 and 4). These are, thus, recomputed projecting the signal onto $\mathcal{D}_*^w$ and a new approximation of f named $\mathbf{b}_{**}^w$ is found. Exactly the same is done for BPDN, ending up with a subdictionary $\mathcal{D}_*$ and a new approximation $\mathbf{b}_{**}$. Of course, $\mathit{support}(\mathbf{b}_*) = \mathit{support}(\mathbf{b}_{**})$ and $\mathit{support}(\mathbf{b}_*^w) = \mathit{support}(\mathbf{b}_{**}^w)$. Formally the approximants found by BPDN and WBPDN are respectively:

$$\begin{aligned} f_{**} = & \quad D_* D_*^+ f = D\mathbf{b}_{**} \text{ and} \\ f_{**}^w = & \quad D_*^w (D_*^w)^+ f = D\mathbf{b}_{**}. \end{aligned} \tag{75}$$

In synthesis, the pursuit algorithm is used only to select a dictionary subset and then the coefficients of the approximation are computed again, by means of a simple projection. Fig. 5 and 6, show how this technique considerably improves the results obtained by solving problems P_1 and P_1^w. Moreover they confirm the advantages of the weighted algorithm with respect to the non weighted one.

V. Examples: A Natural Signal Approximation with Coherent Dictionaries and an *A Priori* Model

In the previous sections, the example of approximating a synthetically generated piecewise-smooth signal has been presented. This successfully illustrates how we can exploit *prior* signal models in order to ameliorate the behavior of sub-optimal algorithms in the retrieval of sparse approximations. In this section we show that the improvement is not limited to artificially generated signal, provide a very simple example where weighted algorithms perform better also with natural signals. Moreover, as we will show, the *a priori* weights can be automatically extracted from the data and optimized such that the performance of the weighted algorithm in use is maximized. Continuing with the class of signals analyzed in previous examples, we would like to approximate a signal that can be considered as piecewise-smooth. This kind of signals can be efficiently approximated by using dictionaries of functions that can optimally represent both features of the signal: discontinuities and smooth parts. For this purpose an overcomplete coherent dictionary given by footprints and wavelets is used, as in [1] and in all the previous examples in the present work.

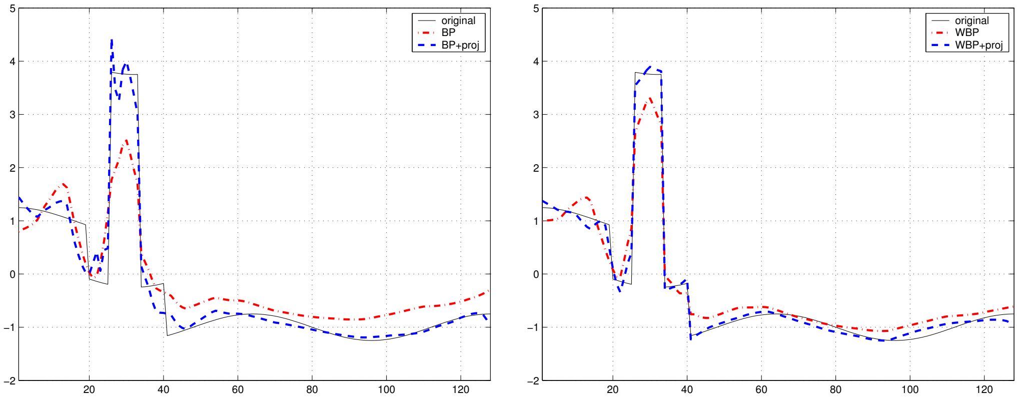

Fig. 5. The original signal reconstructed from a 10-terms approximation computed by BPDN (left) and WBPDN (right). The comparison show the improvement given by recomputing the projections once that the algorithm has selected a subdictionary. For the errors see Figure 6

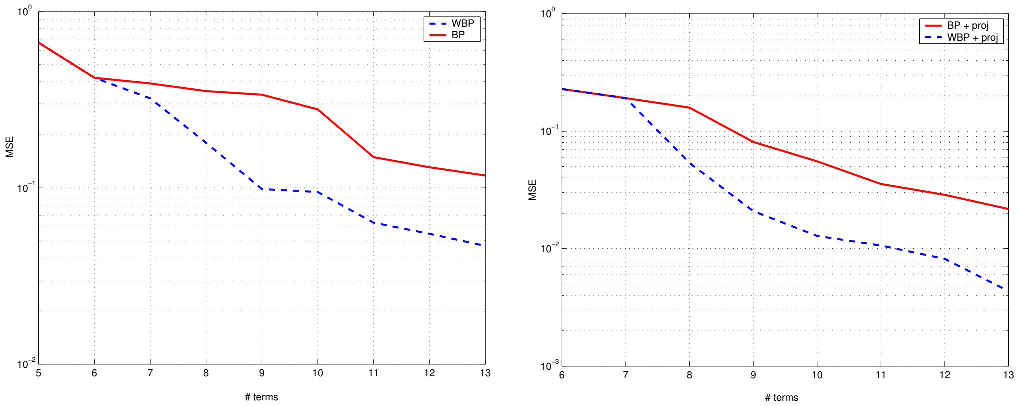

Fig. 6. Errors (in log scale) of the m-terms approximations f with BPDN and WBPDN. In the figure on the right the approximations are computed projecting the signal on the subdictionary selected by the algorithm (see Eq. 75)

A. A Noisy Piecewise-smooth Natural Signal

As often suggested, images may be represented by a piecewise-smooth component and a texture component [24], [25]. A way to get a "natural" 1D piecewise-smooth signal is to extract a column or a row from a natural image. In particular, we study two 1D signals extracted from the image "cameraman", shown in Fig. 7: one corresponds to the 140th column, and the second to the 80th row.

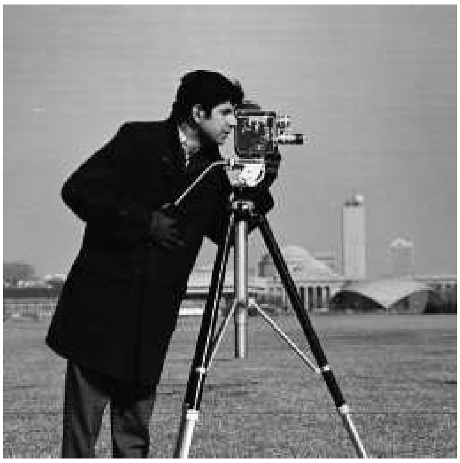

Fig. 7. Cameraman picture used to extract the example of a "real world" 1D signal with piecewise-smooth characteristics

B. Modeling the Relation Signal-Dictionary

The dictionary in use is composed by the union of the *Symmlet*-4 orthonormal base and the set of piecewise-constant footprints (see Sec. III-C and Fig. 1). Since the input signals are constituted by 256 samples, the dictionary is expressed by a matrix of size 256×512. The modeling of the interaction between the signal and the dictionary is performed using the simple approach described in [1]. The weighting matrix $W(f, D)$ is generated by means of a pre-estimation of the locations where footprints are likely to be used and the assumption that in such locations wavelets have less probability to be required. This discrimination is reflected in the diagonal matrix $W(f, D)$ in the following way: locations where a footprint is likely to be placed are not penalized (thus the weighting factor remains 1). On the contrary, wavelets that overlap the footprint and footprints considered unlikely to be used get a penalizing factor $\beta \in (0, 1]$. To be more explicit, the edge detection based model used in [1] is recalled:

Algorithm 1 $W(f, D)$ estimation

```
Require: $\mathcal{D} = \mathcal{D}_{Symmlet} \cup \mathcal{D}_{Footprints}$, define a threshold $\lambda$ , define a penalty factor $\beta$
1: $f_{diff} = D^{+}_{Footprints} \cdot f$ {Footprints location estimation (edge detection)}
2: Threshold $f_{diff}$ by $\lambda$ putting greater values to 1 or $\beta$ otherwise.
3: $W^{diag}_{footprints} = f_{diff}$ {Diagonal of the sub-matrix of $W(f, D)$ corresponding to footprints.}
4: Create $W^{diag}_{wave}$ s.t. all wavelets intersecting the found footprints locations equal $\beta$, set to 1 otherwise.
5: $W(f, D) = diag\left(\left[W^{diag}_{wave} \quad W^{diag}_{footprints}\right]\right)$;
```

As one can observe, two parameters configure the model that generates $W(f, D)$: a threshold λ and a penalty weight β. In practice, as shown later in this section, these can be selected by an optimization procedure such that the energy of the approximation error is minimized.

C. Signal Approximation

An additional phase is inserted before the approximation. It consists of an estimation of the main features of the particular signal to approximate and their relation with the dictionary in use. We thus resume the general procedure by these two steps:

1) Estimation of the *a priori* information from the “real world” signal using a *prior* model.
2) Use of a weighted algorithm (greedy or relaxed) based on the estimated *a priori* knowledge to find the appropriate atoms subset to approximate the “real world” signal.

Furthermore, an iterative version of this two phase algorithm can be considered in order to optimize the parameters that configure the *prior* model used in the first step: the Expectation Maximization (EM) algorithm. A first approach for the parameters tuning can be a grid search, or a multi-scale grid search. More sophisticated search techniques could be used, in particular, an overview can be found in [26].

D. Results

The results obtained from the framework introduced above are illustrated in the following. First, we show the quantitative impact of using *Weighted*-MP/OMP and WBPDN in terms of the residual error energy. Right after, the use of atoms of the dictionary to represent the main features of the signal is analyzed. Finally, we explore the influence of tuning in an appropriate way the two parameters that configure our penalty model.

1) Approximation Results with OMP: Two approximation examples, where OMP and Weighted-OMP are used, are presented. Following the two steps procedure, described above, we look for approximants of the signals appearing at the left of Figs. 8 and 9. The former corresponds to the 140th column of the cameraman picture and the latter to the 80th row. The improvement in performance of Weighted-OMP in the case of sparse approximations is assessed by the rate of convergence of the residual energy. On the right hand of Figs. 8 and 9, the graphs show that after a certain number of iterations, Weighted-MP selects better atoms than classic *Weak*-MP. Hence the convergence of the error improves and this yields a gain of up to 2 dBs for the first example and up to 2.5 dBs in the second one (depending on the iteration).

2) Approximation Results with BPDN: The 1D signal extracted from the 140th column of cameraman, illustrated on the left side of Fig. 8, is approximated by using BPDN and WBPDN. As explained in section IV-E, the pursuit algorithm is used only to select a dictionary subset and then, the coefficients of the approximation are computed again by means of a simple projection. Fig. 10 shows the decay of the energy of the error while the number of atoms selected increases. It is clear how the use of the *a priori* helps the algorithm in finding a better approximation of the signal. The results concerning WBPDN are obtained by adopting a weighting matrix that corresponds to $\lambda = 90$ and $\beta = 0.2$. Notice that these values are not optimal for all the numbers of non-zero coefficients. Better results can be achieved by tuning appropriately β and γ for any desired m.

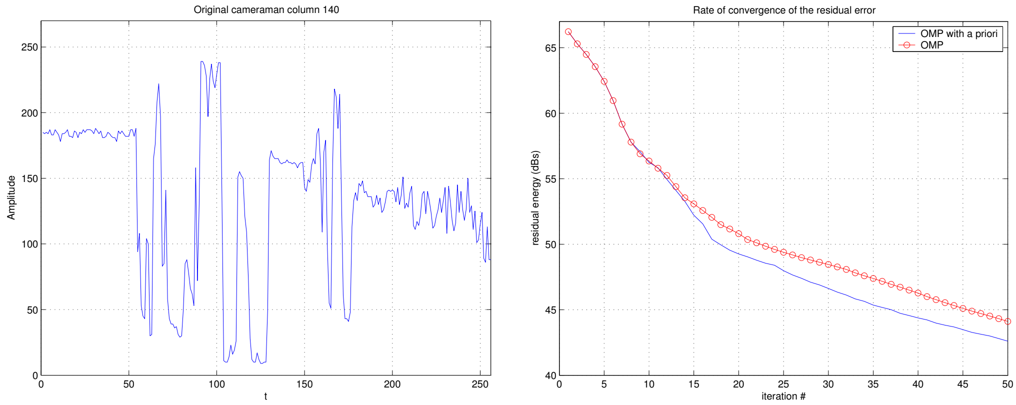

Fig. 8. Experiment of approximating the 1D signal extracted from the 140th column of "cameraman". Left, 1D signal used in the experience can be seen. Right, the rate of convergence of the residual error. In red can be observed the OMP result. In blue the Weighted-OMP result.

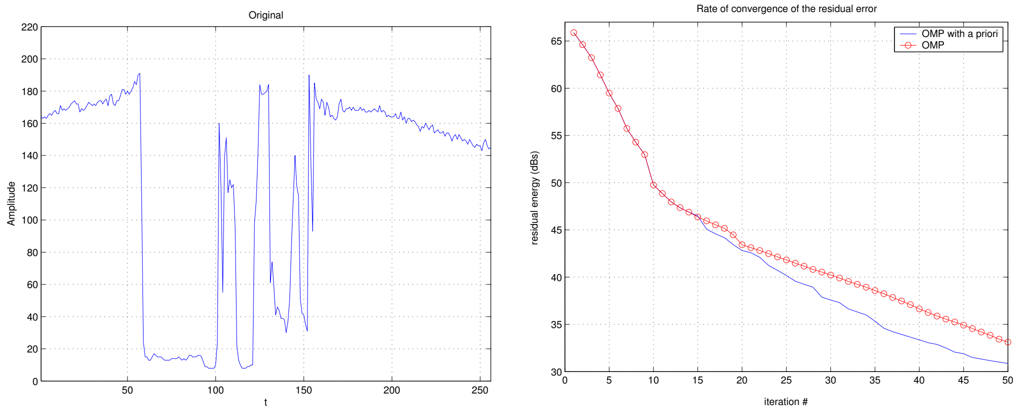

Fig. 9. Experiment of approximating the 1D signal extracted from the 80th row of the 256x256 cameraman picture from Fig. 7. Left, 1D signal used in the experience can be seen. Right, the rate of convergence of the residual error. In red can be observed the OMP result. In blue the Weighted-OMP result.

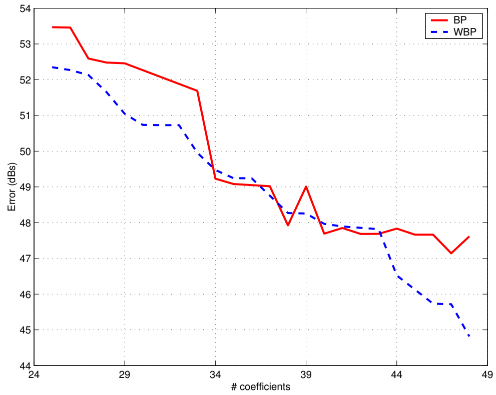

Fig. 10. Error (in dB) obtained by BPDN and WBPDN approximating the 1D signal extracted from the 140th column of the image "cameraman" (see Fig. 7) using different numbers of atoms. Both results are obtained by using quadratic programming for selecting a dictionary subset and then recomputing the coefficients by re-projecting the signal onto the span of the subdictionary. The procedure is illustrated in section IV-E.

3) Capturing the Piecewise-smooth Component with Footprints Basis: Here, the results intend to underline the importance of selecting the appropriate atom to represent a particular signal feature. In Fig. 11 we can see in the upper row the resulting approximants of original signal obtained from the 140th column (Fig. 8) after 50 iterations of OMP (left) and Weighted-OMP (right). The result which considers the *a priori* is a 1.51 dBs better than the approximant obtained by OMP. At this point, it is important to notice the result depicted in the lower row of Fig. 11. These waveforms represent the signal components that are captured exclusively by the footprints atoms and *Symmlet*-4 scaling functions. These signal components should correspond to the piecewise-smooth parts of the signal. However, and as depicted by the lower row of Fig. 11, in the case of OMP (bottom left) the piecewise-smooth component captured by footprints and low-pass functions is far from what one could expect. Intuitively one can understand that the OMP algorithm is failing in the selection of atoms. On the other hand, the result obtained by Weighted-OMP (bottom right) clearly shows that footprints and *Symmlet*-4 scaling functions are capturing a much more accurate approximant of the piecewise-smooth component of the signal. Hence, a better approximation is achieved by using the *a priori* information. This leads to a sparser approximation too.

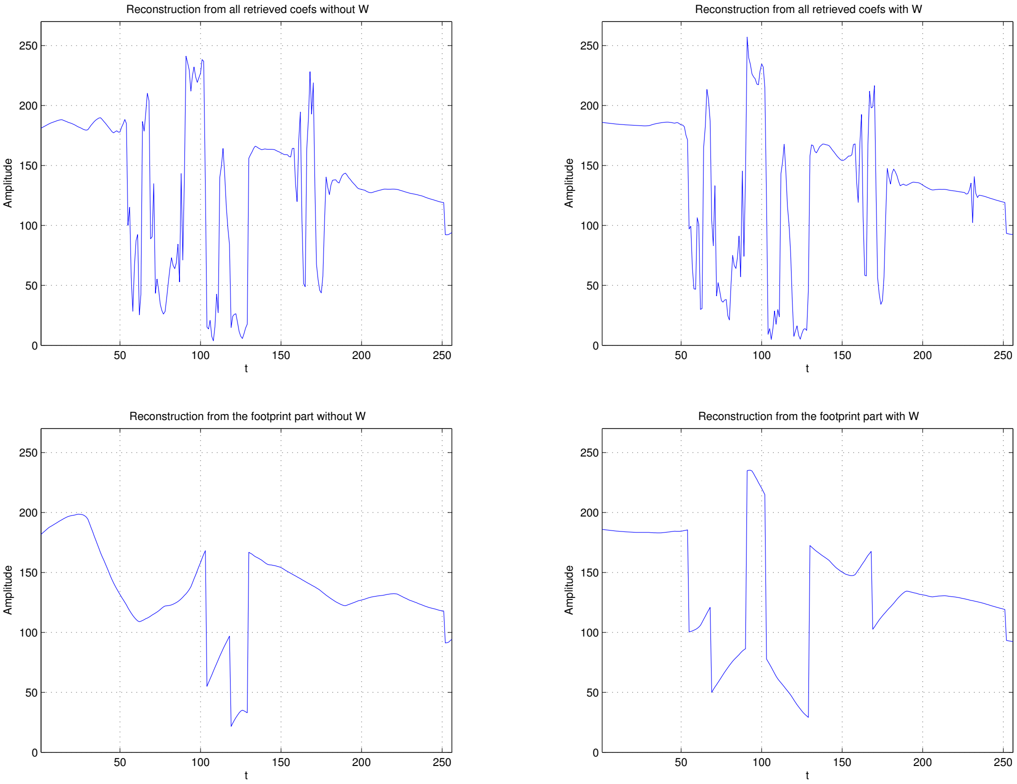

Fig. 11. Upper left: Approximation after 50 iterations with OMP without using *a priori* information. Upper right: Approximation after 50 iterations with Weighted-OMP(+1.51 dBs). Bottom left: Signal components captured by *Symmlet*-4 scaling functions and Footprints when simple OMP is used. Bottom right: Signal components captured by *Symmlet*-4 scaling functions and Footprints using Weighted-OMP.

4) Parameter Search: Finally, we show the influence of the parameters λ and β in the average quadratic error of the residues obtained by Weighted-MP, i.e.

$$E\left\{r_n|\lambda',\beta'\right\} = \frac{\sum_{n=0}^{N-1}\|r_n\|^2}{N} \quad \text{s.t.} \quad r_n \text{ has been obtained fixing } \lambda=\lambda' \text{ and } \beta=\beta'. \tag{76}$$

In Figs. 12 and 13 the magnitude of Eq. (76) is shown as a function of λ (model threshold) and β (penalty weight) for the two natural examples exposed in this work. Fig. 12 corresponds to the case of the 140th column of cameraman, and Fig. 13 concerns the 80th row.

In the figures, a mapping of the meaning of the colors is available. The lower the value of $E\left\{r_n|\lambda',\beta'\right\}$ is, the more probably the associated model parameters are the good ones. In this case, it can be easily observed how the optimal configuration of parameters concentrates in both cases in a unique global optimum. Hence, the set of optimal parameters that fit the data model can be easily found by some iterative procedure.

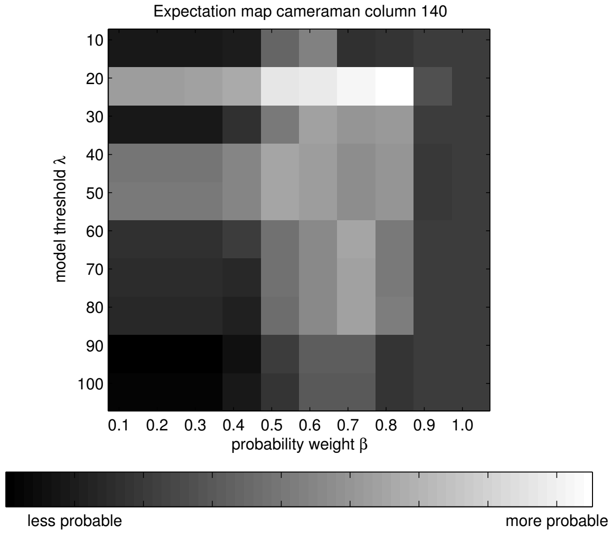

Fig. 12. Representation of the expectation map depending on the parameters that configure the *a priori* model in the experiment set up in Fig. 8. The expectation corresponds to the energy of the residual error.

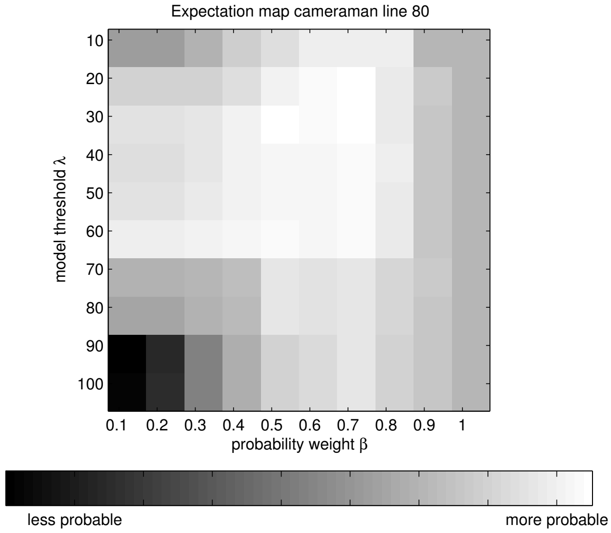

Fig. 13. Representation of the expectation map depending on the parameters that configure the *a priori* model in the experiment set up in Fig. 9. The expectation corresponds tot the energy of the residual error.

VI. Conclusions

This work presents theoretical results on the performance of including *a priori* knowledge in algorithms for signal approximation like *Weak*-MP or BPDN. We introduce weighted variants called Weighted-MP/OMP and WBPDN. Theoretical results show that these algorithms may supply much better results than classic approaches for highly non-linear signal approximations, if sufficiently reliable *prior* models are used. This reliability is theoretically represented in our results by ϵ_{max} while the discriminative ability of the *a priori* is given by w_{Γ}^{max} and w_{Γ}^{min}. The fact that these quantities are normally unavailable in practice, disables us to use numerically the sufficient conditions found unlike in the case where no *prior* is used. Nevertheless, the results found are able to explain how weighted Greedy and BPDN algorithms behave compared to non-weighted ones. A field to explore may be to try to determine some bounds on this quantity depending on the class of signals to approximate and the practical estimators (those that generate the *a priori* weights) in use. Practical examples concerning synthetic and natural signals have been presented where we used a dictionary with high internal cumulative coherence. Theoretically, with such a dictionary, greedy or relaxation algorithms are not guaranteed to recover the set of atoms of the best m-terms approximation. Sufficient conditions for the recovery of all the correct atoms of the best m-terms approximation are far from being satisfied.

Nevertheless, these are too restrictive and pessimistic. Even if they are not satisfied, in some cases, the recovery is still possible. In our examples, classic algorithms like OMP or BP do not seem to succeed in selecting what can be considered, intuitively, as the appropriate atoms to represent efficiently the different signal events. To the contrary, the use of an appropriate joint *prior* model of the interaction between the signal and the dictionary is able to significantly enhance the final subset selection result. This fact confirms our theoretical results. Weighted-MP/OMP and WBPDN outperform their classic counterpart when a reliable *a priori* knowledge is used. Sparse signals approximation requires the use of dictionaries capable to catch efficiently the main features of a certain signal. Particular applications often focus on a certain class of signals to be treated. Thus, it is a wise strategy to use adapted dictionaries for such signals. On top of that, and in order to overcome the possible internal coherence of the dictionary in use an appropriate strategy is also to adapt the subset selection algorithm to the class of signals to be approximated. In effect, the lesson drawn from our study is that signal adapted algorithms are of key importance in optimal atoms selection for best m-terms signal approximations with dictionaries that are not highly incoherent. This can be done from a Bayesian point of view by introducing appropriate *prior* models in known algorithms, resulting, for example, in weighted versions of those: Weighted-MP/OMP or WBPDN. Moreover, we show also how, in the Bayesian scope, *Weak*-MP and BPDN are nothing but particular cases where a specific non-discriminative *a priori* is used. Finally, we conclude that a suitable framework for non-linear signal approximation turns to be the use of optimized dictionaries as pool of adapted building pieces in the approximation synthesis step and the use of signal adapted algorithms by means of *a priories* in the analysis (subset selection) step. Very good feature estimators exist and a lot of experience on them is available in the literature. For every particular application models exploiting certain features may be found. These can be married with most common algorithms used for sparse approximations/representations (*Weak*-MP, BPDN, BP) to obtain much better performances. In fact, the examples presented in this work may be subject to improvement if more robust estimators were used.

Appendix

Proof of Corollary 2 of Section III-A:

Proof: For simplicity, let us use an upper bound on the left hand side of the above inequality. Indeed, the factor $0 < \left(w_{\overline{\Gamma}^{max}}\right)^2 \leq 1$ is removed:

$$\left(1 + \frac{m\left(1 - \left(\mu_1^w(m-1) + \epsilon_{max}\right)\right)\left(w_{\overline{\Gamma}^{max}}\right)^2}{\left(1 - \left(\mu_1^w(m-1) + \mu_1^w(m) + \epsilon_{max}\right)\right)^2}\right) \leq \left(1 + \frac{m\left(1 - \left(\mu_1^w(m-1) + \epsilon_{max}\right)\right)}{\left(1 - \left(\mu_1^w(m-1) + \mu_1^w(m) + \epsilon_{max}\right)\right)^2}\right).$$

Let us suppose the *a priori* knowledge in use is reliable. Then the following relations can be assumed:

$$\mu_1^w(m-1) + \mu_1^w(m) + \epsilon_{max} \quad \leq \quad \mu_1(m-1) + \mu_1(m) < 1 \tag{77}$$

and

$$\mu_1^w(m-1) + \epsilon_{max} \quad \leq \quad \mu_1(m-1). \tag{78}$$

Now we can proof the inequality. Let us make the hypothesis that the following is true:

$$\begin{aligned} \left(1 + \frac{m\left(1 - \left(\mu_1^w(m-1) + \epsilon_{max}\right)\right)}{\left(1 - \left(\mu_1^w(m-1) + \mu_1^w(m) + \epsilon_{max}\right)\right)^2}\right) &\leq \left(1 + \frac{m\left(1 - \mu_1(m-1)\right)}{\left(1 - \left(\mu_1(m-1) + \mu_1(m)\right)\right)^2}\right) \\ \frac{\left(1 - \left(\mu_1^w(m-1) + \epsilon_{max}\right)\right)}{\left(1 - \left(\mu_1^w(m-1) + \mu_1^w(m) + \epsilon_{max}\right)\right)^2} &\leq \frac{\left(1 - \mu_1(m-1)\right)}{\left(1 - \left(\mu_1(m-1) + \mu_1(m)\right)\right)^2}. \end{aligned}$$

Then,

$$1 \leq \frac{\left(1 - \mu_1(m-1)\right)}{\left(1 - \left(\mu_1^w(m-1) + \epsilon_{max}\right)\right)} \cdot \frac{\left(1 - \left(\mu_1^w(m-1) + \mu_1^w(m) + \epsilon_{max}\right)\right)^2}{\left(1 - \left(\mu_1(m-1) + \mu_1(m)\right)\right)^2}. \tag{79}$$

According to (77) and (78), the following can be considered:

$$\begin{aligned} \tau_1 &\triangleq \mu_1(m-1) - \left(\mu_1^w(m-1) + \epsilon_{max}\right) \\ \tau_2 &\triangleq \mu_1(m-1) + \mu_1(m) - \left(\mu_1^w(m) + \mu_1^w(m-1) + \epsilon_{max}\right) \end{aligned}$$

where $0 \leq \tau_1 \ll \mu_1(m)$ and $0 \leq \tau_2 \ll \mu_1(m-1) + \mu_1(m)$. Hence, being $\delta \triangleq \tau_2 / \left(1 - \left(\mu_1(m-1) + \mu_1(m)\right)\right)$ the second fraction in (79) can be substituted:

$$1 \leq \frac{\left(1 - \mu_1(m-1)\right)}{\left(1 - \left(\mu_1^w(m-1) + \epsilon_{max}\right)\right)} \cdot (1 + \delta)^2. \tag{80}$$

Moreover, considering for the first fractional term of (80) that $\delta' \triangleq \tau_1 / (1-\mu_1(m-1))$, then

$$1 \leq \frac{1}{1+\delta'} \cdot (1+\delta)^2 = (1+\delta) \cdot \frac{1+\delta}{1+\delta'}. \tag{81}$$

From this, clearly $(1+\delta) \geq 1$. So, if $(1+\delta) \geq (1+\delta')$ then Corollary 2 is proved.

Hence, let us check finally if this last condition holds:

$$\begin{aligned}
(1+\delta) &\geq (1+\delta') && (82)\\
\delta &\geq \delta' && (83)\\
\frac{\tau_2}{(1-(\mu_1(m-1)+\mu_1(m)))} &\geq \frac{\tau_1}{(1-\mu_1(m-1))} && (84)\\
\frac{\tau_2}{\tau_1}\frac{(1-\mu_1(m-1))}{(1-(\mu_1(m-1)+\mu_1(m)))} &\geq 1 && (85)\\
\frac{\tau_2}{\tau_1} &\geq 1. && (86)
\end{aligned}$$

Which is indeed true, since:

$$\begin{aligned}
\tau_2 &\geq \tau_1 && (87)\\
(\mu_1(m-1)-\mu_1^w(m-1))+(\mu_1(m)-\mu_1^w(m))-\epsilon_{max} &\geq (\mu_1(m-1)-\mu_1^w(m-1))-\epsilon_{max} && (88)\\
(\mu_1(m)-\mu_1^w(m)) &\geq 0, && (89)
\end{aligned}$$

and this concludes the whole proof.

■

The use of reliable *a priori* information ensures that Weighted-MP/OMP will be able to recover an approximant as good as *Weak*-MP or better.

Proof of corollary 3 of section IV-D:

Proof: We want to prove that

$$\frac{\tau}{1-\mu_1(m)-\mu_1(m-1)} \geq \frac{\tau \cdot \frac{w_{\overline{\Gamma}}^{max}}{w_{\Gamma}^{min}} \cdot (1-\mu_1^w(m-1)-\epsilon_{max})}{(1-\epsilon_{max}-\mu_1^w(m)-\mu_1^w(m-1))(1-\mu_1(m-1))}, \tag{90}$$

or equivalently

$$1 \leq \frac{(1-\epsilon_{max}-\mu_1^w(m)-\mu_1^w(m-1))\,(1-\mu_1(m-1))}{(1-\mu_1^w(m-1)-\epsilon_{max})\,(1-\mu_1(m)-\mu_1(m-1))} \cdot \frac{w_{\Gamma}^{min}}{w_{\overline{\Gamma}}^{max}}.$$

If the *a priori* is reliable the weights corresponding to the atoms indexed by Γ are big and the ones corresponding to the atoms in $\overline{\Gamma}$ are small. That is, $\frac{w_{\overline{\Gamma}}^{max}}{w_{\Gamma}^{min}} \leq 1$. Moreover we suppose that $\mu_1(m-1)+\mu_1(m) < 1$ and $\mu_1^w(m-1)+\mu_1^w(m)+\epsilon_{max} < \mu_1(m-1)+\mu_1(m) < 1$. Under these hypotheses one can prove equation (90) following the same procedure of the previous demonstration.

■

Acknowledgments

We would like to thank Rosa M. Figueras i Ventura and Lorenzo Peotta for fruitful discussions.

References

[1] O. Divorra Escoda, L. Granai, and P. Vandergheynst, "On the use of *a Priori* information for sparse signal representations," ITS/LTS-2 EPFL, Tech. Rep. 18.2004, September 2004. [Online]. Available: http://lts2www.epfl.ch/publications.html

[2] J. A. Tropp, "Greed is good : Algorithmic results for sparse approximation," *IEEE Trans. Inform. Theory*, vol. 50, no. 10, pp. 2231–2242, Oct 2004.

[3] ——, "Just relax: Convex programming methods for subset selection and sparse approximation," ICES, University of Texas at Austin, Austin, USA, Tech. Rep., 2004.

[4] R. Gribonval and P. Vandergheynst, "On the exponential convergence of matching pursuits in quasi-incoherent dictionaries," IRISA, Rennes, France, Tech. Rep., 2004.

[5] D. L. Donoho and X. Huo, "Uncertainty principles and ideal atom decomposition," *IEEE Trans. Inform. Theory*, vol. 47, no. 7, pp. 2845–2862, Nov 2001.

[6] D. Donoho and M. Elad, "Optimally sparse representation in general (non-orthogonal) dictionaries via ℓ_1 minimization," *Proc. Nat. Aca. Sci.*,, vol. 100, no. 5, pp. 2197–2202, March 2003.

[7] R. Coifman and M. Wickerhauser, "Entropy-based algortihms for best basis selection," *IEEE Transactions in Information Theory*, vol. 38, pp. 713–718, 1992.

[8] R. M. Figueras i Ventura, P. Vandergheynst, and P. Frossard, "Highly flexible image coding using non-linear representations," ITS, Tech. Rep., 2003.

[9] L. Peotta, L. Granai, and P. Vandergheynst, “Very low bit rate image coding using redundant dictionaries,” in *Proc. of 48th SPIE Annual Meeting*, SPIE. San Diego, CA: SPIE, August 2003.

[10] M. Wakin, J. Romberg, H. Choi, and R. Baraniuk, “Wavelet-domain Approximation and Compression of Piecewise Smooth Images,” *IEEE Transactions on Signal Processing*, April 2004, submitted.

[11] R. Shukla, P. Dragotti, M. Do, and M. Vetterli, “Rate distortion optimized tree structured compression algorithms for piecewise smooth images,” *IEEE Trans. Image Processing*, April 2004.

[12] B. Olshausen and D. Field, “Sparse coding with an overcomplete basis set: A strategy employed by v1?” *Vision Research*, vol. 37, pp. 3311–3325, 1997.

[13] S. S. Chen, D. L. Donoho, and M. A. Saunders, “Atomic decomposition by basis pursuit,” *SIAM J. Sci. Comp.*, vol. 20, no. 1, pp. 33–61, 1999.

[14] V. N. Temlyakov, “Weak greedy algorithms,” Department of Mathematics, University of South Carolina, Columbia, Tech. Rep., 1999.

[15] D. L. Donoho and I. M. Johnstone, “Minimax estimation via wavelet shrinkage,” *Annals of Statistics*, no. 26, pp. 879–921, 1998.

[16] P. Dragotti and M. Vetterli, “Wavelet footprints: Theory, algorithms and applications,” *IEEE Transactions on Signal Processing*, vol. 51, no. 5, pp. 1306–1323, May 2003.

[17] S. Mallat, *A Wavelet Tour of Signal Processing.* Academic Press, 1998.

[18] S. Sardy, A. Bruce, and P. Tseng, “Block coordinate relaxation methods for nonparametric wavelet denoising,” *Journal of Computational and Graphical Statistics*, vol. 9, no. 2, pp. 361–379, 2000.

[19] M. Lewicki and T. Sejnowski, “Learning overcomplete representations,” *Neural Comp.*, vol. 12, pp. 337–365, 2000.

[20] L. Granai and P. Vandergheynst, “Sparse decomposition over multi-component redundant dictionaries,” in *Proc. of Multimedia Signal Processing, Workshop on. MMSP04*, September 2004, pp. 494–497.

[21] A. Kusraev and S. Kutateladze, *Subdifferentials: Theory and Applications.* Kluwer Academic Publishers, 1995.

[22] R. Fletcher, *Practical Methods of Optimization.* Wiley-Interscience, 1997.

[23] J. J. Fuchs, “On sparse representations in arbitrary redundant bases,” *IEEE Trans. Inform. Theory*, vol. 50, no. 6, 2004.

[24] Y. Meyer, “Oscillating patterns in image processing and in some nonlinear evolution equations,” in *AMS*, ser. University Lecture Series, 2002, vol. 22.

[25] L. A. Vese and S. J. Osher, “Modeling textures with total variation minimization and oscillating patterns in image processing,” *Journal of Scientific Computing*, vol. 19, no. 1-3, pp. 553–572, December 2003.

[26] *Global Optimization in Action*, ser. Nonconvex Optimization and its Applications. Dordrecht, The Netherlands: Kluwer Academic Publishers, 1996, vol. 6.