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)
EPFL-STI-ITS-LTS2, station 11, 1015 Lausanne, Switzerland
e-mail: oscar.divorra@ieee.org,{lorenzo.granai,pierre.vandergheynst}@epfl.ch
phone: +41 21 693 56 45, fax: +41 21 693 76 00
EDICS: 2-NLSP

***Abstract*— Recent results have underlined the importance of incoherence in redundant dictionaries for a good behavior of decomposition algorithms like Matching and Basis Pursuit. However, appropriate dictionaries for a given application may not be able to meet the incoherence condition. In such case, decomposition algorithms may completely fail in the retrieval of the sparsest approximation. This paper studies the effect of introducing *a priori* knowledge when recovering sparse approximations over redundant dictionaries. 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 standard approaches such as greedy algorithms and relaxation methods. Our results reduce to the classical case when no *prior* information is available. Examples validate and illustrate our theoretical statements.**

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

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 [1], [2], [3], [4]. Natural signals, though, are very unlikely to be exact sparse superpositions of vectors. In fact the set of such signals is of measure zero in $\mathbb{C}^N$ [1]. In the present work, we thus analyze the case of sparse approximations:

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

where $f \in \mathcal{H}$ is the function to be approximated and $\mathcal{H}$ is a Hilbert space (unless otherwise stated, it is assumed that $\mathcal{H} \equiv \mathbb{R}^n$). D is the $n \times d$ synthesis matrix of the dictionary $(\mathcal{D})$, where each one of the columns corresponds to an atom (g_i). $\mathcal{D}$ is defined such that $\mathcal{D} = \{g_i : \ i \in \Omega\}$ where $\forall i \ \ \|g_i\|_2 = 1$, Ω is the index set for the whole dictionary and $|\Omega| = d$. Finally, $\mathbf{c}$ is the vector of coefficients to be recovered.

In general, the problem of recovering the sparsest signal approximation (or representation) over a redundant dictionary is a NP-hard problem. However, when *particular* classes of dictionaries are used, this does not impair the possibility of solving this problem using faster sub-optimal methods.

As demonstrated in [1], [2], [4], [5], in order to ensure the good behavior of algorithms like General *Weak* Matching Pursuit (Weak-MP) and Basis Pursuit Denoising (BPDN), dictionaries need to be incoherent enough. Under this main hypothesis, sufficient conditions have been stated such that both methods are able to recover the atoms from the sparsest m-term expansion of a signal.

However, experience and intuition dictate that good dictionaries for approximations of natural signals can be very redundant and, depending on the kind of signal structures to describe, they may be highly coherent. This is a strong discrepancy between theory and practice. In order to tackle this problem, in this paper we discuss the potentiality of using *a priori* knowledge in the atom selection procedure for sparse approximations. In this work, *prior* information is inserted by means of a weighting procedure. We do not treat 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 a given signal and on the kind of dictionary used. Nevertheless, we give an insight through a realistic example in Section V. The question of achieving sparseness in some class of coherent dictionaries have already been faced in [6], but for the particular case of *Vandermonde* and *Fourier* dictionaries, while the use of weights into some highly nonlinear approximation algorithms was previously suggested in [7] and [8].

The aim of this paper is the theoretical study of weighted algorithms in the prospective of achieving sparseness. Our main results are:

- The definition of the Weighted Basis Pursuit Denoising (Weighted-BPDN) and Weighted Matching

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

Pursuit (Weighted-MP) algorithms for approximation purposes. We reformulate classic BPDN and Weak-MP in order to take *a priori* information into account when decomposing a signal.
- A theoretical study of the approximation properties of Weighted-BPDN and Weighted-MP/OMP.
- An example, on natural signals, of the effect of using *prior* models at the decomposition stage, when coherent overcomplete dictionaries are used. Models exploit the relationship between the class of signals and the internal structure of the selected dictionary.

This paper is structured as follows. Section II recalls some aspects of the state of the art in sparse signal approximations. In Section III and IV we study the effect of introducing *a priori* information, respectively, in greedy and relaxation algorithms. Examples on natural signal decompositions are presented in Section V, while conclusions are drawn in Section VI.

II. Sparse Approximations: An Overview

In this section, the basic principles of Weak-MP and BPDN are reviewed together with very recent theoretical results concerning their ability to succeed in recovering best m-term signal approximations.

A. Greedy Algorithms

General Matching Pursuits [5], [9], [10] iteratively build m-term approximants by selecting at each step the most appropriate term from $\mathcal{D}$ according to a certain rule. Each iteration k (where $k \geq 0$) can be seen as a two step procedure:

1) A selection step where an atom $g_{i_k} \in \mathcal{D}$ is chosen.
2) A projection step where an approximant $f_{k+1} \in span(g_{i_l} : l \in \{0, ..., k\})$ and a residual $r_{k+1} = f - f_{k+1}$ are generated.

The selection step, at iteration k, can be generally formulated as the maximization of a similarity measure $C(r_k, g_i)$ between the signal to approximate (the residual at the kth iteration: $r_k = f - f_k$) and the dictionary atoms:

$$g_{i_k} = \arg\max_{g_i \in \mathcal{D}} C(r_k, g_i). \tag{2}$$

Pure Matching Pursuit uses the modulus of the scalar product as similarity measure, i.e. $C(r_k, g_i) = |\langle r_k, g_i \rangle|$. More generally, Weak-MP allows an additional flexibility factor $\alpha \in (0, 1]$ allowing the selected atom g_{i_k} to be such that $|\langle r_k, g_{i_k} \rangle| \geq \alpha \max_{i \in \Omega} |\langle r_k, g_i \rangle|$. The sub-optimality factor α, as demonstrated in [9], does not necessarily prevent the greedy algorithm from converging to a solution (i.e. $\lim_{k \to \infty} \|r_k\|_2^2 = 0$), though it may affect negatively the convergence speed.

The projection step determines whether Matching Pursuit (MP) or Orthogonal Matching Pursuit (OMP) is in use. The former just guarantees that the atom selected at iteration k is orthogonal to the residual r_{k+1} [10]. The latter, constructs the approximant f_{k+1} by finding an orthogonal projection of f over the space spanned by all selected atoms until iteration k [11].

Even though greedy algorithms do not directly solve the problem of Eq. (1), in some cases, the solution they supply coincides with that of (1). Depending on the characteristics of the dictionary, one can establish some guaranties that Weak-MP is able to recover the atoms of a m-term sparsest approximation of a signal f [1]. This is performed by means of a set of sufficient conditions obtained by Tropp in [1], for the particular case of OMP, and further extended to the case of Weak-MP by Gribonval and Vandergheynst in [5]. All sufficient conditions rely on the fact that $\mathcal{D}$ must be incoherent enough. Indeed, given $\mathcal{D}$ and m, if a cumulative coherence measure [1] of the dictionary, defined as

$$\mu_1(m, \mathcal{D}) \triangleq \max_{|\Lambda|=m} \max_{i \in \Omega \setminus \Lambda} \sum_{\lambda \in \Lambda} |\langle g_i, g_\lambda \rangle|, \tag{3}$$

where $\Lambda \subset \Omega$ has size m, is small enough such that

$$\frac{\mu_1(m)}{1 - \mu_1(m-1)} < \alpha,$$

then Weak-MP(OMP) works as desired. Remark that the measure known as coherence of a dictionary μ, and often used to characterize redundant dictionaries corresponds to the particular case $\mu = \mu_1(1, \mathcal{D})$. Furthermore $\mu_1(m, \mathcal{D}) \leq m\mu$.

Based on the coherence measure of Eq. (3), some results concerning the behavior of Weak-MP (or OMP) algorithms have been also obtained. For example, two of them are: an upper bound on the decay of the residual energy ($\|r_k\|_2^2$) as a function of k and $\mu_1(m, \mathcal{D})$, and a condition on the recoverability of a "good" atom at iteration k depending on the residual energy $\|r_k\|_2^2$. Indeed, if the remaining residual energy at the kth iteration ($\|r_k\|_2^2$) is bigger than a certain factor, which depends on the optimal residual energy for an m-term approximation ($\|r_m^{opt}\|_2^2$) and the cumulative coherence for m terms, then an additional atom belonging to Γ_m (Γ of size m) can be recovered. This last condition is called General Recovery Condition in [1], and Robustness Condition in [5].

B. Convex Relaxation of the Subset Selection Problem

Another instance of problem (1) is given by

$$(P_0) \quad \min_{\mathbf{c}} \|f - D\mathbf{c}\|_2^2 + \tau^2 \|\mathbf{c}\|_0. \tag{4}$$

In statistics this problem is also known as Subset Selection. It searches for a sparse approximation of f, considering a trade-off between the error and the number of elements that participate into the expansion.

Unfortunately, solving P_0 is NP-hard. A possible way of simplifying the computation is to substitute the ℓ_0 quasi-norm with the convex ℓ_1 norm. This relaxation leads to problem P_1:

$$(P_1) \quad \min_{\mathbf{b}} \frac{1}{2} \|f - D\mathbf{b}\|_2^2 + \gamma \|\mathbf{b}\|_1. \tag{5}$$

P_1 is the minimization of a convex functional that can be solved by classical Quadratic Programming methods. This

relaxation is similar to the one leading to the definition of the Basis Pursuit (BP) principle for the case of exact signal representation [12]. Note that, if $\mathcal{D}$ is orthonormal, the solution of P_1 can be found by a soft shrinkage of the coefficients [13], [12], while, if $\mathcal{D}$ is a union of orthonormal sub-dictionaries, the problem can be solved recurring to the Block Coordinate Relaxation method [14], faster than Quadratic Programming.

In [2], the author studies the relation between the subset selection problem (4) and its convex relaxation (5), and shows that, given an index subset $\Lambda \subset \Omega$, any coefficient vector which minimizes Eq. (5) is supported inside Λ if the following condition is satisfied:

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

where $a_\Lambda = D_\Lambda D_\Lambda^+ f$, D_Λ is the sub-matrix containing only the atoms indexed by Λ, and $+$ denotes the Moore-Penrose matrix inverse. In particular, the relationship between the trade-off parameters τ and γ is studied, proving that if the coefficient vector $\mathbf{b}_*$ minimizes the function (5) with threshold $\gamma = \tau/(1 - \sup_{i \notin \Gamma} \left\|D_\Gamma^+ g_i\right\|_1)$, where Γ indicates the optimal set of atoms of a m-term sparsest approximation, then the relaxation never selects a non-optimal atom and the solution of the convex relaxation is unique.

C. Lessons Learned

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 relaxed algorithms. However, as discussed in the introduction, experience seems to teach us that highly redundant and, often, coherent dictionaries are more powerful for natural signals approximation. In the following sections, we solve this problem by carefully studying the relationship between the signal and the dictionary, through the introduction of *a priori* information in the decomposition process.

III. Including *A Priori* Information in Greedy algorithms: Weighted-MP

In this section we explore the effect of using *a priori* knowledge in greedy algorithms for the recovery of the best m-term approximant ($f_m^{opt} = D \cdot \mathbf{c}_{opt}$) of a signal f. For this purpose we define a Bayesian formulation of Matching Pursuit that we call Weighted-MP.

In Sec. II, we recall how MP and OMP use the scalar product as similarity measure for the selection of the most appropriate atom. This bears some similarity with searching the atom g_{i_k} with "Maximum Likelihood" for a given residual r_k. Indeed, the selection procedure in MP may be seen as a maximization of the probability $p(g_i|r_k)$, that is like to consider $C(r_k, g_i) \sim p(g_i|r_k)$ in Eq. (2). $|\langle r_k, g_i \rangle|$ may be intuitively seen as a measure of the conditional probability $p(r_k|g_i)$, assuming that each residual r_k is composed by the superposition of an atom g_{i_k} and gaussian noise, i.e. $r_k = g_{i_k} + n_k$. In the event that all atoms are *a priori* equiprobable, maximizing $p(r_k|g_i)$ is equivalent to maximize $p(g_i|r_k)$. Let us now consider the case where the atoms do not have the same *a priori* probability to appear in the optimal set Γ_m, and let us assume that we have at our disposal a *prior* knowledge about the likelihood of each g_i. By means of the Bayes' Rule, when some *a priori* $p(g_i)$ is available, the probability to maximize becomes

$$p(g_i|r_k) = \frac{p(r_k|g_i)\, p(g_i)}{p(r_k)}, \tag{6}$$

where the denominator is a constant for each signal r_k. Emulating this, the selection rule of MP can, thus, be modified multiplying the modulus of the scalar product by a weighting factor $w_i \in (0, 1]$, which depends on the atom index i. This is done in order to represent the insertion of some heuristic measure of *prior* information. Hence, now $C(r_k, g_i)$ in Eq. (2) can be considered such that:

$$C(r_k, g_i) = |\langle r_k, g_i \rangle| \cdot w_i.$$

We call this family of weighted greedy algorithms Weighted-MP. The Weighted-MP approach does not modify the projection step of the algorithm, allowing to freely select the MP or OMP projection strategy. For the sake of simplicity, Weighted-MP is used in the remaining of the paper as a general term to refer to both projection approaches. The kind of projection will not be specified unless judged relevant. In this work, we assume for simplicity that the *a priori* knowledge (appearing under the form of weights w_i) is independent of the iteration of the algorithm[1].

One may interpret Weighted-MP as a greedy algorithm where the use of non-unit norm atoms within the dictionary is allowed. Unit norm atoms are re-weighted according to some heuristic measure of *prior* information, which gives some hint about their likelihood to belong to the optimal set Γ. Based on that, in the following, sufficient conditions for the recovery of a "correct" atom of the sparsest m-term approximant are established. Later, we study how *a priori* knowledge affects the rate of convergence of greedy algorithms, and finally, an example is presented. Even if Weighted-MP can be considered a kind of weak greedy algorithm, we will see that it can perform better than Pure MP.

A. Influence on Sparse Approximations

In this work, the *a priori* knowledge is expressed by the diagonal matrix $W(f, \mathcal{D})$.

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

Notice that all the atoms in the dictionary are assumed to have some *a priori* non-zero probability. In effect,

[1]However, one could decide to update the atom weights at every iteration. This would introduce more flexibility in the formulation of Weighted-MP

those that would have a zero weight are considered to be excluded from the dictionary. We define also $w_{\overline{\Gamma}}^{max}$ as the biggest weight corresponding to the subset of atoms indexed in $\overline{\Gamma} = \Omega \setminus \Gamma$, hence:

$$w_{\overline{\Gamma}}^{max} \triangleq \sup_{\gamma \in \overline{\Gamma}} w_\gamma. \tag{7}$$

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

$$\epsilon_{max} \triangleq \sup_{\gamma \in \Gamma} \left(1 - w_\gamma^2\right). \tag{8}$$

Eqs. (7) and (8) concern the goodness of the *a priori* information. The reader will notice that these quantities depend on the optimal set of atoms Γ, preventing from establishing a rule to compute them in advance. The role of these magnitudes is to represent the influence of the *prior* in the results obtained below. Notice that $0 \leq \epsilon_{max} < 1$ and $0 < w_{\overline{\Gamma}}^{max} \leq 1$.

***Definition** 2: ϵ_{max} is close to zero if "good" atoms (the ones belonging to Γ) are not penalized by the* a priori *information. In such a case we state that the* a priori *knowledge is "reliable".*

The quantity $w_{\overline{\Gamma}}^{max}$ becomes small if all "bad" atoms are strongly penalized by the *a priori* knowledge. Notice that the "reliability" of a *prior* model does not impose any condition on $w_{\overline{\Gamma}}^{max}$.

The weights are not arbitrary and are not supposed to be independently and blindly optimized by the algorithm during the subset selection procedure. These values alone are not meant to determine whether an atom shall be included in the selection or not. The weights introduce a fuzzy likelihood that could be derived from a good parametric model[2] on the interaction between signals and the dictionary.

The *a priori* matrix W allows a new signal dependent definition of the cumulative coherence $\mu_1(m)$ introduced in [1]. Indeed, the conditions that ensure the recoverability of the best m-term approximant relay on this quantity. Using *a priori* information, some atom interactions can be penalized or even dismissed in the cumulative coherence measure:

***Definition** 3: The* Weighted Cumulative Coherence *function of $\mathcal{D}$ is defined as the following data dependent measure:*

$$\mu_1^w(m, \mathcal{D}, W) \triangleq \max_{|\Lambda|=m} \max_{i \in \Omega \setminus \Lambda} \sum_{\lambda \in \Lambda} |\langle g_\lambda, g_i \rangle| \cdot w_\lambda \cdot w_i. \tag{9}$$

Note that if $W = I$, then $\mu_1^w(m, \mathcal{D}, I) = \mu_1(m, \mathcal{D})$. Moreover, $\forall m, \mathcal{D}, W$ we have that $\mu_1^w(m, \mathcal{D}, W) \leq \mu_1(m, \mathcal{D})$.

We can now observe the behavior of greedy algorithms when *a priori* information is used.

***Theorem** 1: Let $\{r_k\} : 0 \leq k \leq m$, be the set of residuals generated by Weighted-MP in the approximation of a signal f, and let f_m^{opt} be its best m-term approximant. Then, for any positive integer m such that $\mu_1^w(m-1) + \mu_1^w(m) < 1 - \epsilon_{max}$ and*

$$\|r_k\|_2^2 > \left\|f - f_m^{opt}\right\|_2^2 \cdot \left(1 + \frac{m\left(1 - (\mu_1^w(m-1) + \epsilon_{max})\right)\left(w_{\overline{\Gamma}}^{max}\right)^2}{\left(1 - (\mu_1^w(m-1) + \mu_1^w(m) + \epsilon_{max})\right)^2}\right), \tag{10}$$

Weighted-MP will recover an atom that belongs to the optimal set Γ_m (in the sense of (1)*).*

In the case that f_m^{opt} can not be reached or just an approximate solution exists, a sub-optimality factor $\eta \geq 0$ can be introduced by substituting $\|f - f_m^{opt}\|_2^2$ by $\|f - f_m^{opt}\|_2^2 (1+\eta)^2$ in Eq. (10).

Flavor of the Proof: To prove Theorem 1, we follow [1] and [5]. In order to ensure the recovery of any atom belonging to the optimal set $\Gamma = \Gamma_m$, the following *a priori* dependent condition needs to be satisfied:

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

where $\alpha \in (0, 1]$ is a weakness factor [9]. We compute an upper bound on $\rho^w(r_k)$. This must be smaller than α so that the recovery condition holds. As detailed in the proof appearing in our report [15], this yields

$$\rho^w(r_k) \leq \frac{w_{\overline{\Gamma}}^{max}\sqrt{\left(1 - \mu_1^w(m-1) - \epsilon_{max}\right)m} \cdot (1+\eta) \cdot \left\|r_m^{opt}\right\|_2}{\left(1 - \mu_1^w(m-1) - \epsilon_{max}\right)\left\|f_m^{opt} - f_k\right\|_2} + \frac{\left\|f_m^{opt} - f_k\right\|_2 \mu_1^w(m)}{\left(1 - \mu_1^w(m-1) - \epsilon_{max}\right)\left\|f_m^{opt} - f_k\right\|_2} < \alpha, \tag{11}$$

where $r_m^{opt} = f - f_m^{opt}$. Considering that $\|f_m^{opt} - f_k\|_2^2 = \|r_k\|_2^2 - \|r_m^{opt}\|_2^2$ and solving (11) for $\|r_k\|_2^2$, in the particular case of $\alpha = 1$, yields the result of Theorem 1.

□

Theorem 1 means that, if the approximation error at the kth iteration is still bigger than a certain quantity, then another term of the best m-term approximant can be recovered. This is similar to the result of [5], but here the use of *a priori* information results in a smaller bound. More terms may, thus, be recovered.

Finally, the general effect of using *a priori* knowledge can be summarized by the following Corollary.

***Corollary** 1: Let $W(f, \mathcal{D})$ express a reliable* a priori *knowledge and assume $\alpha = 1$, then for any positive integer m such that $\mu_1(m-1) + \mu_1(m) \geq 1$ but $\mu_1^w(m-1) + \mu_1^w(m) < 1 - \epsilon_{max}$, Weighted-MP (unlike MP) will recover the atoms belonging to the best m-term approximant f_m^{opt}. Moreover, for any positive integer m such that $\mu_1^w(m-1) + \mu_1^w(m) + \epsilon_{max} \leq \mu_1(m-1) + \mu_1(m) < 1$, Weighted-MP has a weaker sufficient condition than MP for the recovery of correct atoms from the best m-term approximant. Hence, the correction factor of the right-hand side of expression* (10) *is equal or smaller in the weighted case for any value*

[2]Hereby, fuzzy is used in the sense that the *prior* information is not precise, i.e the *a priori* model is not able to totally determine good and bad atoms by itself.

of $w_{\Gamma}^{max} \in (0,1]$*:*

$$\left(1+\frac{m\left(1-\left(\mu_1^w(m-1)+\epsilon_{max}\right)\left(w_{\Gamma}^{max}\right)^2\right)}{\left(1-\left(\mu_1^w(m-1)+\mu_1^w(m)+\epsilon_{max}\right)\right)^2}\right) \leq \left(1+\frac{m\left(1-\mu_1(m-1)\right)}{\left(1-\left(\mu_1(m-1)+\mu_1(m)\right)\right)^2}\right). \tag{12}$$

See the Appendix for a proof. Therefore, Weighted-MP is guaranteed to recover equally good or better approximants than classic MP when reliable *a priori* information is used (if for some case $\mu_1^w(m)+\mu_1^w(m+1)+\epsilon_{max} < \mu_1(m-1)+\mu_1(m) < 1$, then the better behavior is guaranteed).

B. Rate of Convergence of Weighted-MP

The energy of the series of residuals r_k generated by the greedy algorithm progressively converges toward zero as k increases. In the same way, Weighted-MP with reliable *a priori* information is expected to have a better behavior and a faster convergence rate than Weak-MP. A tighter measure 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. To prove this, we follow the path suggested in [1] and [5], introducing as before the consideration of the *a priori* information in the formulation. The results formally show how Weighted-MP can outperform Weak-MP when the *a priori* knowledge is reliable.

***Theorem** 2: Let $W(f,\mathcal{D})$ be a reliable* a priori *information matrix and $\{r_k\} : k \geq 0$ a sequence of residuals produced by Weighted-MP, then as long as $\|r_k\|_2^2$ satisfies Eq.* (10)*, Weighted-MP picks up a correct atom and*

$$\left(\|r_k\|_2^2 - \|r_m^{opt}\|_2^2\right) \leq \left(1-\alpha^2\frac{\left(1-\mu_1^w(m-1)-\epsilon_{max}\right)}{m}\right)^{k-l} \cdot \left(\|r_l\|_2^2 - \|r_m^{opt}\|_2^2\right), \tag{13}$$

where $k \geq l$.

As observed for Theorem 1, if f_m^{opt} can not be reached or just an approximate solution exist, $\|r_m^{opt}\|_2^2$ is substituted by $\|r_m^{opt}\|_2^2 (1+\eta)^2$ in Eq. (13).

Flavor of the Proof: The proof is technical but it is a simple adaptation of the proof of Theorem 7 in [5]. We just sketch it here and refer the interested reader to [15] for further details.

Consider k such that $\|r_k\|_2^2$ satisfies Eq. (10). Then, it is known that for Weak-MP:

$$\|r_{k-1}\|_2^2 - \|r_k\|_2^2 \geq |\langle r_k, g_{i_k}\rangle|^2,$$

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

$$\begin{aligned}\|r_{k-1}\|_2^2 - \|r_k\|_2^2 &\geq \alpha \sup_{\gamma\in\Gamma} |\langle r_k, g_\gamma \cdot w_\gamma\rangle|^2 \frac{1}{w_\gamma^2} \\ &= \alpha \sup_{\gamma\in\Gamma} |\langle f_m^{opt} - f_k, g_\gamma \cdot w_\gamma\rangle|^2 \frac{1}{w_\gamma^2},\end{aligned}$$

where the last equality follows from the assumption that Eq. (10) is satisfied and because $(f - f_m^{opt}) \perp span(\Gamma)$. As described in [15], $\sup_{\gamma\in\Gamma} |\langle f_m^{opt} - f_k, g_\gamma \cdot w_\gamma\rangle|^2$ may be lower bounded as a function of the smallest square singular value $(\sigma^2_{min_w})$ of $G \triangleq (D_\Gamma W_\Gamma)^T (D_\Gamma W_\Gamma)$, the size of the subset Γ_m, and a residual energy:

$$\|r_{k-1}\|_2^2 - \|r_k\|_2^2 \geq \alpha \frac{\sigma^2_{min_w}}{m} \left\|f_m^{opt} - f_k\right\|_2^2. \tag{14}$$

Finally, combining this with the fact that $\|f_m^{opt} - f_{k-1}\|_2^2 - \|f_m^{opt} - f_k\|_2^2 = \|r_{k-1}\|_2^2 - \|r_k\|_2^2$ allows to easily derive, as stated in [15], that, for $0 \leq l \leq k$,

$$\|r_k\|_2^2 - \|r_m^{opt}\|_2^2 (1+\eta)^2 \leq \left(1-\alpha\frac{\sigma^2_{min_w}}{m}\right)^{k-l} \cdot \left(\|r_l\|_2^2 - \|r_m^{opt}\|_2^2 (1+\eta)^2\right),$$

and the Theorem is proved since $\sigma^2_{min_w} \geq 1-\mu_1^w(m-1)-\epsilon_{max}$ (see Lemma 1 in [15]).

□

Theorem 2 implies that the rate of convergence of Weighted-MP has an upper bound with exponential decay, as well as Weak-MP. Moreover, when reliable *a priori* information is used, the bound is lower. This result suggests that the convergence of suitably weighted greedy algorithms is faster than in the case of pure greedy algorithms. Of course, this is subject to the use of a model that puts in relation both the signal and dictionary. Some fuzzy (i.e. $w_{\Gamma}^{max} = 1$) and non-penalizing indication about the appropriate atoms may be of great help for the convergence of the algorithm.

Depending on the sufficient conditions specified in Sec. III-A, it will be possible to recover the optimal set Γ. 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** 4: Consider a Weighted-MP algorithm used for the approximation of signals. We define the number of provably correct steps N_m as the smallest positive integer such that*

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

which corresponds to the number of atoms belonging to the optimal set that can be recovered given a signal f, a dictionary $\mathcal{D}$ 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 $\mu_1^w(m-1)+\mu_1^w(m)+\epsilon_{max} < 1$, the provable number of correct iterations will depend on the final error of the best m-term approximation. In the following theorem, bounds on the quantity N_m

are given for Weighted-MP. To obtain the results we follow [5].

Before stating the theorem, the reader should note that from now on, $w_{\overline{\Gamma}_l}^{max}$ defines the same concept as in (7) for an optimal set of atoms Γ of size l, i.e. for Γ_l.

***Theorem** 3: Let $W(f,\mathcal{D})$ be a reliable* a priori *information and $\{r_k\} : k \geq 0$ a sequence of residuals produced by Weighted-MP when approximating f. Then, for any integer m such that $\mu_1^w(m-1)+\mu_1^w(m)+\epsilon_{max} < 1$, we have $N_1 \leq 1$ and for $m \geq 2$:*

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

$$2 \leq N_m < 2 + \frac{m}{1-\mu_1^w(m-1)-\epsilon_{max}}.$$

$$\log\left(\frac{3\left\|r_1^{opt}\right\|_2^2}{m \cdot \left\|r_m^{opt}\right\|_2^2 (1-\epsilon_{max_m}) \cdot \left(w_{\overline{\Gamma}_m}^{max}\right)^2}\right). \quad (15)$$

- *else* $N_m \leq 1$.

In order to prove Theorem 3, several intermediate lemmas are necessary. We omit these, as well as the general proof of the Theorem since they follow easily by taking into account the use of *a priori* information in the proofs of Theorem 7 in [5]. We refer again to [15] for a detailed description of the proof.

From (15) we can draw that the upper bound on the provably correct number of steps N_m is tighter for Weighted-MP if a reliable *a priori* knowledge is used. Indeed, in accordance with Theorem 2, which states a tighter residual error convergence bound for Weighted-MP, one can also have a tighter estimate for Weighted-MP about which is the maximum number of good iterations the algorithm might do. If some *a priori* is available, some atom interactions will not influence $\mu_1^w(m-1)$ in Eq. (15), unlike in the case of Theorem 7 in [5] where $\mu_1(m-1)$ was used.

Moreover, in a situation where the reliable *a priori* model was discriminative enough, we know from (15), that there would be additional room for an improvement on the number of correct iterations recovered by the greedy algorithm with respect to [5]. The term $w_{\overline{\Gamma}_m}^{max}$ helps to increase the value of the bound, describing the fact that Weighted-MP can recover a higher number of correct iterations than MP. In addition, compared to the case when no *a priori* information is available, the condition for the validity of bound (15) is softened.

The assumption of good discrimination capabilities of the *a priori* model is somehow unrealistic in practice, i.e. a small value for $w_{\overline{\Gamma}_m}^{max}$ indicates that the model can already discriminate between Γ and $\overline{\Gamma}$. Nevertheless, the result of Theorem 3, gives a better estimate on the upper bound of N_m thanks to the use of $\mu_1^w(m-1)$ instead of $\mu_1(m-1)$, and furthermore it suggests that using an *a priori* model should have a positive effect on the stability of Weighted-MP. In practice, if the *prior* information is capable to handle some punctual ambiguity (i.e. somewhere where the greedy algorithm may get confused and make a mistake)

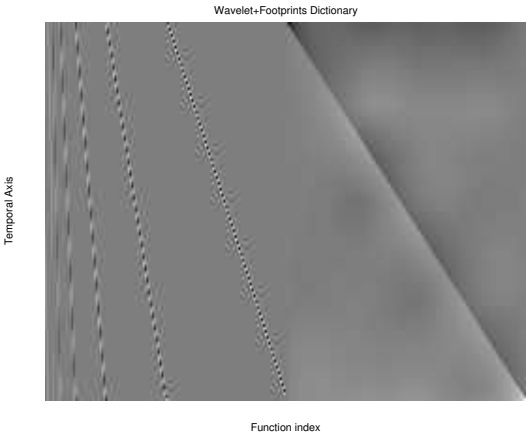

Fig. 1. Dictionary formed by the Symmlet-4 [16] (left half) and its respective footprints for piecewise constant singularities (right half).

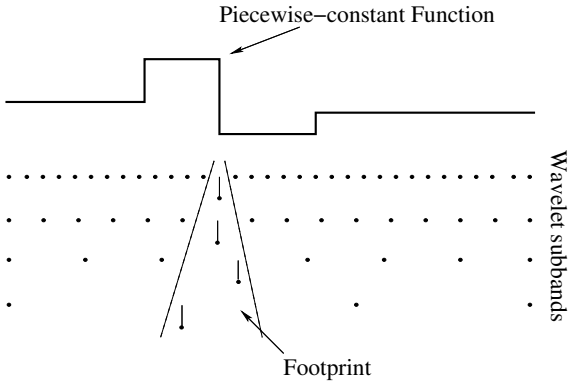

Fig. 2. Wavelet Footprints description scheme for a piecewise-constant signal [17].

that may affect the choice of the appropriate function at a given MP step, then the benefits for the convergence of the algorithm can be of extreme relevance. This can be the case even if the *a priori* model does not supply a good discrimination between Γ and $\overline{\Gamma}$. Examples in Sec. III-C and Sec. V illustrate this situation.

C. Example: Use of Footprints and Weighted-OMP for Sparse Approximations.

To give an example of approximation using *a priori* information, we consider the case where a piecewise-smooth signal is decomposed over an overcomplete dictionary. The dictionary is built by the union of an orthonormal basis defined by the *Symmlet*-4 family of wavelets [16] and the respective family of footprints for all the possible translations of the Heaviside function (see [17]). The former is intended to represent the smooth part of the signal, while the latter is used to model the discontinuities. Footprints are functions composed by the superposition of all wavelet coefficients that a given deterministic singularity model (translations of the Heaviside function in our case) generates on a wavelet basis (see Fig. 2). The graphical representation of the dictionary matrix can be seen in Fig. 1, where the columns are the waveforms that

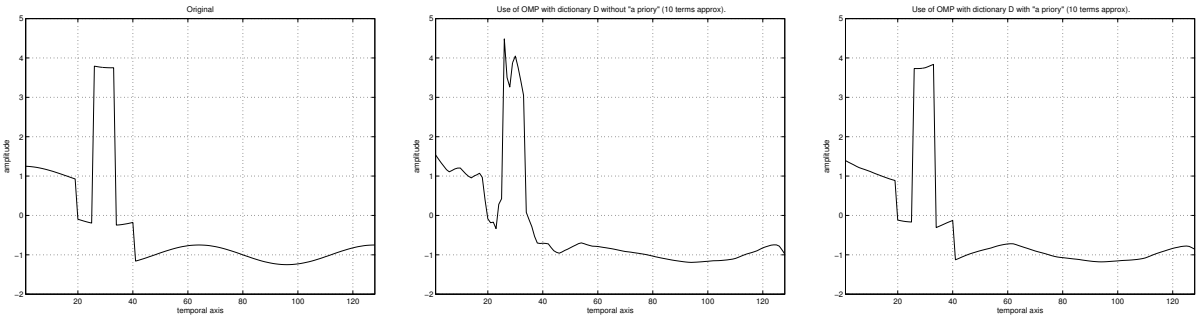

Fig. 3. Comparison of OMP based approximation with 10 terms using the footprints dictionary (Fig. 1). Left: Original signal. Middle: "blind" OMP approximation. Right: OMP with prior knowledge of the footprints location.

compose the dictionary. Such a dictionary is far from satisfying 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 cumulative coherence. Nevertheless, we consider this example because of two main reasons. First, because the 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 (see [18], [19] for some dicussion on the use coherent dictionaries). 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 nonlinear expansions.

The estimation of the *a priori* information is based on a signal adaptive parametric model that establishes a relationship between the dictionary, its internal structure and the input data. Roughly speaking, the *a priori* model used here is composed of two steps: first, an estimate of the location of edges in the signal is generated; then, W is configured so that footprints are favored to describe discontinuities, while wavelets are privileged for smooth regions. For a more detailed explanation on the model configuration as well as for the parameter optimization, we refer to Sec. V, Algorithm 1.

Fig. 3 presents, from left to right, the original signal and the two approximations obtained by OMP without and with *a priori* information. The input signal has a polynomial degree which is higher than the number of vanishing moments of the *Symmlet*-4. 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 enhancement can be seen in Fig. 4 where the convergence of the approximation error is presented. The use of weights is definitively helpful and a considerable reduction of the error is achieved for a small number of terms.

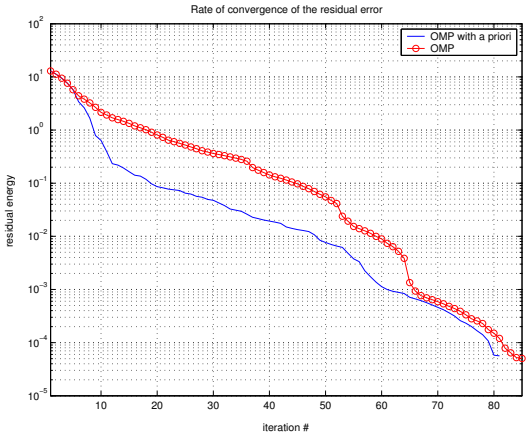

Fig. 4. Rate of convergence of the error with respect to the iteration number in the experiment of Fig. 3

IV. Approximations with Weighted Basis Pursuit Denoising

Another sub-optimal method to solve the problem in Eq. (1) is given by relaxation algorithms, whose recovery capabilities, in presence of *a priori* knowledge, are treated in this section. More precisely, we investigate the effects of inserting *a priori* knowledge in the convex relaxation of the subset selection problem (see Sec. II), i.e. in the approximation case.

A. A Bayesian Approach to Weighted Basis Pursuit Denoising

Let us now study the problem of signal approximation from a Bayesian point of view. We examine under which hypotheses BPDN is an appropriate approach. This leads us to generalize the BPDN principle through the definition of Weighted Basis Pursuit Denoising (WBPDN).

First, we write our model for the data approximation problem, where $\hat{f}$ is the approximant and r is the residual:

$$f = \hat{f} + r = D\mathbf{b} + r. \tag{16}$$

Assuming r to be an iid Gaussian set of variables, the probability that f corresponds to $\hat{f}$, given D and $\mathbf{b}$ is:

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

where σ_r^2 is the variance of the residual. In the approximation problem, one aims at maximizing the likelihood $p(\mathbf{b}|f, D)$. 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 $p(f, D)$ uniform, it follows that the most probable signal representation is:

$$\mathbf{b}_P = \arg\max_{\mathbf{b}} p(f|D, \mathbf{b}) \cdot p(\mathbf{b}). \tag{17}$$

Let us now assume that 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 (17), by computing the logarithm, it follows that

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

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 [20]. 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:

$$(P_1^w) \qquad \min_{\mathbf{b}} \frac{1}{2}\|f - D\mathbf{b}\|_2^2 + \gamma\|W^{-1}\mathbf{b}\|_1, \tag{18}$$

where the diagonal matrix with entries in $(0, 1]$ is defined in Sec. III. One can notice that in Eq. (18), the introduction of weights allows to individually model the components of $\mathbf{b}$. This approach is analogous to the one introduced in [21] 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[3] distribution. This problem is faced, for example, in [21].

B. Weighted Relaxed Subset Selection

In this subsection, the relationship between the results obtained by solving problem P_1^w and P_0 are studied. Note that 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. First, let us introduce the Weighted Recovery Factor:

***Definition** 5: Given a dictionary $\mathcal{D}$ indexed in Ω and an index subset $\Lambda \subset \Omega$, we define the Weighted Recovery Factor (WRF) as:*

$$WRF(\Lambda) = \sup_{i \notin \Lambda} \left\|(D_\Lambda W_\Lambda)^+ g_i \cdot w_i\right\|_1. \tag{19}$$

The best approximation of the input signal over the atoms indexed in Λ is given by $a_\Lambda = DD_\Lambda^+ f$. The next lemma, similarly to the Correlation Condition Lemma in [2], basically states that, if the atoms of Λ have a small correlation with the residual $(f - a_\Lambda)$, then the support of any vector that solves P_1^w is a subset of Λ. This result will be used to prove Theorem 4.

***Lemma** 1: Given an index subset $\Lambda \subset \Omega$, suppose that the following condition is satisfied:*

$$\|D^T(f - a_\Lambda)\|_\infty < \frac{\gamma}{w_{\overline{\Lambda}}^{max}} \cdot (1 - WRF(\Lambda)), \tag{20}$$

where $w_{\overline{\Lambda}}^{max} \in (0, 1]$ is the quantity defined by equation (7). Then, any coefficient vector $\mathbf{b}_*$ that minimizes the cost function of problem P_1^w must have a support contained in Λ.

Proof: Assume that $\mathbf{b}_*$ minimizes (18), but it uses an index outside Λ. One can compare $\mathbf{b}_*$ with its projection $D_\Lambda^+ D\mathbf{b}_*$, which is supported in Λ, obtaining:

$$\begin{aligned}&2\gamma\left(\left\|W^{-1}\mathbf{b}_*\right\|_1 - \left\|W_\Lambda^{-1}(D_\Lambda^+ D\mathbf{b}_*)\right\|_1\right) \leq \\ &\left\|f - DD_\Lambda^+ D\mathbf{b}_*\right\|_2^2 - \|f - D\mathbf{b}_*\|_2^2.\end{aligned} \tag{21}$$

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 $\overline{\Lambda} = \Omega \setminus \Lambda$. Acting as in the proof of the Correlation Condition Lemma in [2] it follows that:

$$\begin{aligned}&\left\|W^{-1}\mathbf{b}_*\right\|_1 - \left\|W_\Lambda^{-1}(D_\Lambda^+ D\mathbf{b}_*)\right\|_1 \geq \\ &(1 - WRF(\Lambda)) \cdot \left\|W^{-1}\mathbf{b}_{\overline{\Lambda}}\right\|_1.\end{aligned} \tag{22}$$

For more details, see [15]. The quantity appearing on the right-hand side of (21) does not depend on the weighting matrix, thus, exactly as in [2], it can be upper bounded by $2\|\mathbf{b}_{\overline{\Lambda}}\|_1 \cdot \left\|D^T(f - \mathbf{a}_\Lambda)\right\|_\infty$. This, together with (21) and (22), gives:

$$\begin{aligned}&\gamma\,(1 - WRF(\Lambda)) \cdot \left\|W^{-1}\mathbf{b}_{\overline{\Lambda}}\right\|_1 \leq \\ &\|\mathbf{b}_{\overline{\Lambda}}\|_1 \cdot \left\|D^T(f - \mathbf{a}_\Lambda)\right\|_\infty.\end{aligned} \tag{23}$$

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

$$\begin{aligned}\gamma\,(1 - WRF(\Lambda)) &\leq \frac{\|\mathbf{b}_{\overline{\Lambda}}\|_1}{\left\|W^{-1}\mathbf{b}_{\overline{\Lambda}}\right\|_1} \cdot \left\|D^T(f - a_\Lambda)\right\|_\infty \\ &\leq w_{\overline{\Lambda}}^{max} \cdot \left\|D^T(f - a_\Lambda)\right\|_\infty.\end{aligned} \tag{24}$$

[3]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 in Eq. (18), leading to the following minimization problem: $\min_{\mathbf{b}} \frac{1}{2}\|f - D\mathbf{b}\|_1 + \gamma\|W^{-1}\mathbf{b}\|_1$.

If (20) is valid , then (24) fails and so one must discard the hypothesis that $\mathbf{b}_*$ is non-zero for an index in $\overline{\Lambda}$.

■

Suppose now 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 $f_m^{opt} = D\mathbf{c}_\Gamma = DD_\Gamma^+ f = D_\Gamma D_\Gamma^+ f$. Assuming $WRF(\Gamma) < 1$, we have the following result.

***Theorem** 4: Given $\tau > 0$, trade-off parameter of the problem P_0, suppose that $\mathbf{b}_*$ minimizes the cost function of problem P_1^w with threshold*

$$\gamma = \frac{\tau \cdot w_{\overline{\Gamma}}^{max}}{1 - WRF(\Gamma)}, \tag{25}$$

where $w_{\overline{\Gamma}}^{max}$ is defined in (7). Then:

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

$$\|\mathbf{c}_\Gamma - \mathbf{b}_*\|_\infty \leq \frac{\tau \cdot \frac{w_{\overline{\Gamma}}^{max}}{w_\Gamma^{min}} \cdot \left\| \left(D_\Gamma^T D_\Gamma\right)^{-1} \right\|_{\infty,\infty}}{1 - WRF(\Gamma)}. \tag{26}$$

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

$$|\mathbf{c}_\Gamma(j)| > \frac{\tau \cdot \frac{w_{\overline{\Gamma}}^{max}}{w_\Gamma^{min}} \cdot \left\| \left(D_\Gamma^T D_\Gamma\right)^{-1} \right\|_{\infty,\infty}}{1 - WRF(\Gamma)}. \tag{27}$$

The scalar w_Γ^{min} appearing in Eqs. (26) and (27) is defined as

$$w_\Gamma^{min} \triangleq \inf_{i \in \Gamma} w_i. \tag{28}$$

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 5.1 in [2] and recalling that $\mathcal{D}$ is finite, it can be stated that

$$\left\|D^T(f - f_m^{opt})\right\|_\infty < \tau. \tag{29}$$

Moreover, Lemma 1 guarantees that for any γ satisfying

$$\left\|D^T(f - f_m^{opt})\right\|_\infty < \frac{\gamma}{w_{\overline{\Gamma}}^{max}} \cdot (1 - WRF(\Gamma)), \tag{30}$$

the solution $\mathbf{b}_*$ to the convex problem P_1^w is supported on Γ. From (29) and (30) it follows that for any γ that satisfies the following condition, it is insured that $support(\mathbf{b}_*) \subset \Gamma$:

$$\gamma \geq \frac{\tau \cdot w_{\overline{\Gamma}}^{max}}{1 - WRF(\Gamma)}. \tag{31}$$

In the following, the smallest possible value for γ is chosen, so that, Eq. (31) becomes an equality. The uniqueness of the solution follows from the use of *a priori* weights on the optimality conditions for BPDN found by Fuchs in [22]. With regard to the third point, it can be proved that (see [15]) if $\mathbf{b}_*$ minimizes the cost function of problem P_1^w, then the following bound holds:

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

This yields

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

Using equation (31), the fourth result of the theorem can be proved exactly as in [2].

■

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 to obtain the sparsest signal approximation. Furthermore, the error made by the algorithm to compute the coefficients with respect to the optimal ones is bounded. The quantities w_Γ^{min} and $w_{\overline{\Gamma}}^{max}$ depend on the reliability and goodness of the *prior* knowledge respectively. 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: $w_\Gamma^{min} \to 1$ and $w_{\overline{\Gamma}}^{max} \to 0$), then this results in an improved bound for the error of the coefficients and a condition for γ in Eq. (25) that is easier to respect. The reader will notice that such an "optimal" W is quite improbable to exist in practice. Indeed, the typical information supplied by an *a priori* model will be quite imprecise (this, however, does not prevent an *a priori* model of being reliable and helpful). This aspect is discussed and justified at the end of Sec. IV-C.

Note that, 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. Hence, the error made by the algorithm in the coefficient computation is avoided (see Eq. (26)). This method is used in Sec. IV-D to generate some examples.

C. Relation with the Weighted Cumulative Coherence

In this subsection, the previous results are described using the weighted cumulative coherence function defined in (9). In this way a comparison is made between the results achievable by BPDN and WBPDN.

***Theorem** 5: Assume that the real vector $\mathbf{b}_*$ solves P_1^w with*

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

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

$$\|\mathbf{b}_* - \mathbf{c}_\Gamma\|_\infty \leq$$

$$\frac{\tau \cdot \frac{w_{\overline{\Gamma}}^{max}}{w_\Gamma^{min}}(1 - \epsilon_{max} - \mu_1^w(m-1))}{(1 - \epsilon_{max} - \mu_1^w(m) - \mu_1^w(m-1))(1 - \mu_1(m-1))}. \tag{32}$$

Proof: This result can be obtained from [2] and Theorem 4, since:

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

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

Considering that

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

(see [2] and [22]) proves equation (32). See [15] for a more detailed version of the proof.

■

This result is valid in general and illustrates how the distance between the optimal coefficients and the solution found by solving P_1^w can be bounded. In case no *prior* knowledge is given, the bound on the coefficient error is obtained from Eq. (32) setting $W = I$. Consequently, $w_\Gamma^{min} = 1, \epsilon_{max} = 0$ and $w_\Gamma^{max} = 1$ (see also [2]):

$$\|\mathbf{b}_* - \mathbf{c}_\Gamma\|_\infty \leq \frac{\tau}{1 - \mu_1(m) - \mu_1(m-1)}. \qquad (33)$$

Comparing the two bounds, one can observe how the availability of reliable *prior* information can help in finding a sparser signal approximation. Let $W(f, \mathcal{D})$ be a reliable *a priori* knowledge, with $w_\Gamma^{max}/w_\Gamma^{min} \leq 1$. Then for any positive integer m such that $\mu_1^w(m-1) + \mu_1^w(m) + \epsilon_{max} < \mu_1(m-1) + \mu_1(m) < 1$, 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. (32) is lower than the one in Eq. (33), i.e.

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

$$\frac{\tau}{1 - \mu_1(m) - \mu_1(m-1)}. \qquad (34)$$

This result can be proved in the same way as Corollary 1, whose proof is reported in the Appendix. See [15] for more details.

The reader may notice that if $\frac{w_\Gamma^{max}}{w_\Gamma^{min}} < 1$ the *a priori* information already tells which is the right support of the solution. Indeed, a simple threshold on the weights would find the appropriate set of atoms. This is an unrealistic situation in practice. However, provided that the *a priori* information is reliable, we do not need $\frac{w_\Gamma^{max}}{w_\Gamma^{min}} < 1$ to justify an improvement on the behavior of the algorithm. Suppose that the weights do not penalize the optimal atoms, but only some (not all) of the "wrong" ones: in this case $\frac{w_\Gamma^{max}}{w_\Gamma^{min}} = 1$. In such a situation, if $\mu_1^w(m-1) + \mu_1^w(m) + \epsilon_{max} \leq \mu_1(m-1) + \mu_1(m) < 1$, Eq. (34) is still valid. This means that, even if the *a priori* knowledge is imprecise (but reliable), WBPDN can behave significantly better than BPDN. The same consideration applies to Eqs. (26) and (27).

D. Example: Use of Footprints and WBPDN for Sparse Approximation

We examine again the example presented in section III-C, this time using the Basis Pursuit Denoising and Weighted Basis Pursuit Denoising methods. For an explanation of the *prior* model and the extraction of the *a priori* matrix, see Sec. V. The signal f is decomposed by solving BPDN (problem P_1) and WBPDN (problem P_1^w), where the *a priori* knowledge is introduced. Both solutions are 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 resulting signal approximations still present many non-zero components with negligible values due to the numerical computation: a hard thresholding is, thus, performed in order to get rid of these insignificant elements. In this way, it is possible to measure the ℓ_0 norm of the vector $\mathbf{b}$. The data reported here refer to a threshold value of 10^{-9}. However, in general, the threshold value should depend on the algorithm used to solve the minimization problem, on the machine precision and on the "typical" amplitude of primitives composing the signal to approximate. Of course, the reconstructions are computed starting from the thresholded coefficients. Fig. 5 shows the reconstructions of the input signal given by a 10-terms approximation found by BPDN and WBPDN. The left-hand side of Fig. 6 illustrates the mean square error of the approximations.

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 thus a subset of the dictionary. Let us label the sub-dictionary 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 Theorem 4 and [2]). These are, thus, recomputed projecting the signal onto the subspace spanned by $\mathcal{D}_*^w$ and a new approximation of f named $\mathbf{b}_{**}^w$ is found. Exactly the same is done for BPDN, ending up with a sub-dictionary $\mathcal{D}_*$ and a new approximation $\mathbf{b}_{**}$. Of course, $support(\mathbf{b}_*) = support(\mathbf{b}_{**})$ and $support(\mathbf{b}_*^w) = support(\mathbf{b}_{**}^w)$. Formally the approximants found by BPDN and WBPDN after the projection step are respectively:

$$\begin{aligned} f_{**} &= D_* D_*^+ f = D\mathbf{b}_{**} \text{ and} \\ f_{**}^w &= D_*^w (D_*^w)^+ f = D\mathbf{b}_{**}^w. \end{aligned} \qquad (35)$$

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.

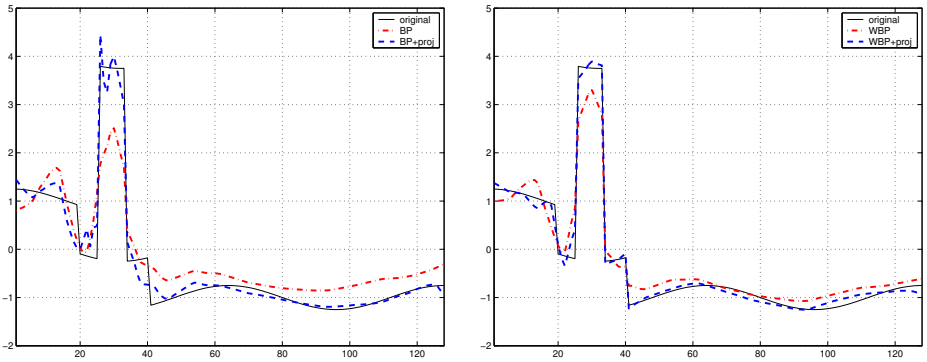

Fig. 5. The original signal reconstructed from a 10-terms approximation computed by BPDN (left) and WBPDN (right). The comparison shows the improvement given by recomputing the projections once the algorithm has selected a sub-dictionary. For the errors see Fig. 6.

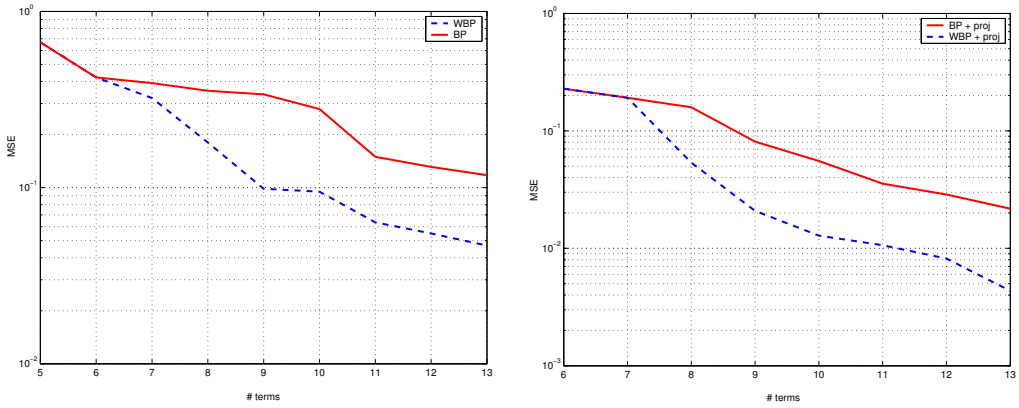

Fig. 6. Errors (in log scale) of the m-term approximations with BPDN and WBPDN. On the right-hand, the approximations are computed projecting the signal onto the sub-dictionary selected by the algorithm (see Eq. (35)).

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

In this section, we apply the methodology introduced in Sections III and IV to natural signals. We also discuss the problem of finding reliable *a priori* information on a concrete example. Moreover, we show how the *a priori* weights can be automatically extracted from the data and optimized in order to maximize the performance of weighted algorithms. We approximate several 1D signals, extracted from a variety of columns of *cameraman* and *Lena* images, considered to be piecewise-smooth, by using an overcomplete coherent dictionary.

A. Modeling the Relation Signal-Dictionary

The dictionary is composed by the union of the *Symmlet*-4 orthonormal basis, used to model smooth parts of the signal, and the set of piecewise-constant footprints meant to model discontinuities (see Sec. III-C and Fig. 1). Since our input signals have 256 samples, D is a matrix of size 256×512. The weighting matrix $W(f, D)$ is generated by means of a pre-estimation of the locations where footprints are likely to be used (there w_i are set to 1). In such locations wavelets are assumed to have a lower probability to be necessary, hence a certain penalty ($\beta \leq 1$ is introduced by means of the corresponding w_i). Wavelets are favored for areas where no appreciable edge is detected (thus, setting the concerned w_i to 1). In such smooth areas, footprints are, thus, penalized. More formally, the modeling of the interaction between signals and the dictionary is performed using the following simple algorithm:

Algorithm 1 $W(f, D)$ estimation

Require: $\mathcal{D} = \mathcal{D}_{Symmlet} \cup \mathcal{D}_{Footprints}$, define a threshold λ , define a penalty factor β

1: $f_{diff} = D^{+}_{Footprints} \cdot f$ {Footprints location estimation (edge detection)}
2: Threshold f_{diff} by λ putting greater values to 1 or β 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 footprint locations equal β, set to 1 otherwise.
5: $W(f, D) = diag\left(\begin{bmatrix} W^{diag}_{wave} & W^{diag}_{footprints} \end{bmatrix}\right)$;

As one can observe, two parameters configure the model generating $W(f, D)$: a threshold λ and a penalty weight β. We will show later that these can be selected by an optimization procedure minimizing the average energy of the approximation error.

B. Signal Approximation

We resume the general procedure for signal approximation by these two steps:

1) Estimation of the *a priori* information from the "real world" signal using an *a priori* model.
2) Use of a weighted algorithm (greedy or relaxed) based on the estimated *a priori* knowledge to find the appropriate atoms subset. Optionally, once these have been selected, their coefficients can be computed again, by means of a simple projection.

Furthermore, an iterative version of this algorithm can be considered in order to optimize the parameters that configure the *a priori* model used in the first step (λ and β in our examples). This can be seen as a kind of Expectation Maximization algorithm. The simplest approach for parameter tuning can be a grid search, or a multi-scale grid search. Nevertheless, much more sophisticated and efficient search techniques may be used to optimize the *a priori* models. See [23] for some global optimization techniques.

C. Results

In this subsection we show the quantitative impact of using weighted algorithms in terms of the residual error energy. Then, we describe how atoms can represent the main features of a signal. We also explore the influence of tuning the two parameters that configure our penalty model, and finally, an empirical consistency analysis is performed on Weighted-MP.

1) Approximation Results with OMP: The improvement of Weighted-OMP is assessed by the rate of convergence of the residual energy, on the right-hand side of Fig. 7: the graph shows that after few iterations, Weighted-OMP selects better atoms than classic OMP. Hence the convergence of the error improves and this yields a gain of up to 2 dB.

We stress again that, extracting relevant footprints and wavelets by simply selecting those with higher *a priori* weights does not yield good sparse approximations. The *a priori* model is just supposed to give rough hints about which functions are useful for every particular signal feature. For instance, the weights computed in our example equal 1 for more than 200 functions, making thus impossible to use thresholding on W as a self-standing selection criterion. Indeed, the use of simple thresholding would imply that $\beta = 0$ in the model. As one can see in Fig. 10 ("probability weight" axis), $\beta = 0$ does not supply the best approximation error average. The model must be, thus, used in conjunction with the atom selection procedure of an appropriate nonlinear subset selection algorithm.

2) Approximation Results with BPDN: The same signal is now approximated by BPDN and WBPDN. As explained in section IV-D, 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. 8 shows the decay of the error versus the number of atoms. 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, as can be seen in the area between 34th and 43rd coefficients in the graph of Fig. 8. Better results can be achieved by tuning appropriately β and γ for any desired m.

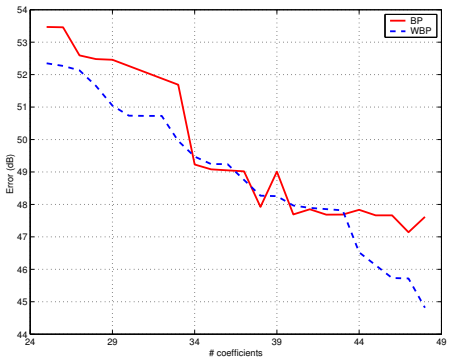

Fig. 8. Error (in dB) obtained by BPDN and WBPDN. Both results are obtained by using quadratic programming for selecting a dictionary subset and then recomputing the coefficients by projecting the signal onto the span of the sub-dictionary. The procedure is illustrated in Sec. IV-D.

3) Capturing the Piecewise-smooth Component with Footprints: Here, the results intend to underline the importance of selecting the appropriate atom to represent a particular signal feature. In the top row of Fig. 9 we can see the resulting approximants after 50 iterations of OMP (left) and Weighted-OMP (right). The result obtained by including the *a priori* is 1.5 dB better than the one obtained by OMP. At this point, it is important to observe the bottom row of Fig. 9. These waveforms represent the signal components captured exclusively by the footprints and wavelet scaling functions. These components should correspond to the piecewise-smooth parts of the signal. However, 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 OMP 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 capture a much more accurate approximant of the piecewise-smooth component of the signal. We can thus argue that a better approximation is achieved by using the *a priori* information, and this leads to a sparser approximation too.

4) Parameter Search: Let us now consider the influence of the parameters λ and β in the average quadratic error of the residues obtained by Weighted-OMP, i.e.

$$E\left\{r_k|\lambda',\beta'\right\} = \frac{\sum_{k=0}^{K-1} \|r_k\|^2}{K}, \tag{36}$$

such that r_k is obtained fixing $\lambda = \lambda'$ and $\beta = \beta'$.

In Fig. 10, the magnitude of Eq. (36) is shown as a function of λ (model threshold) and β (penalty weight).

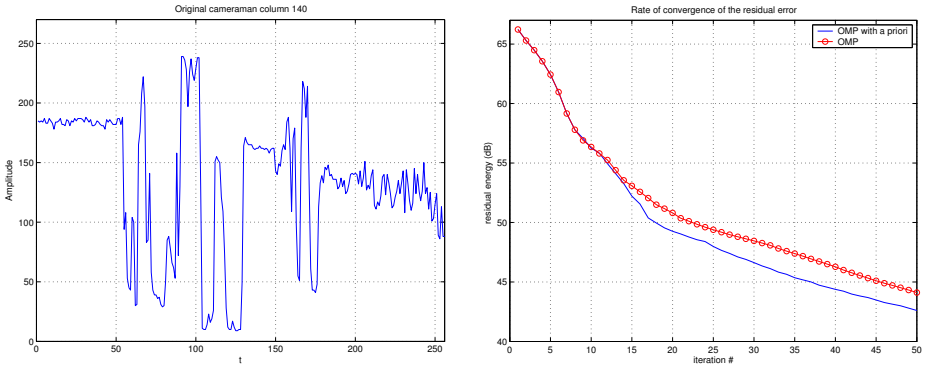

Fig. 7. Experiment of approximating the 1D signal extracted from the 140th column of *cameraman* (On the left). On the right, the rate of convergence of the residual error for OMP and Weighted-OMP.

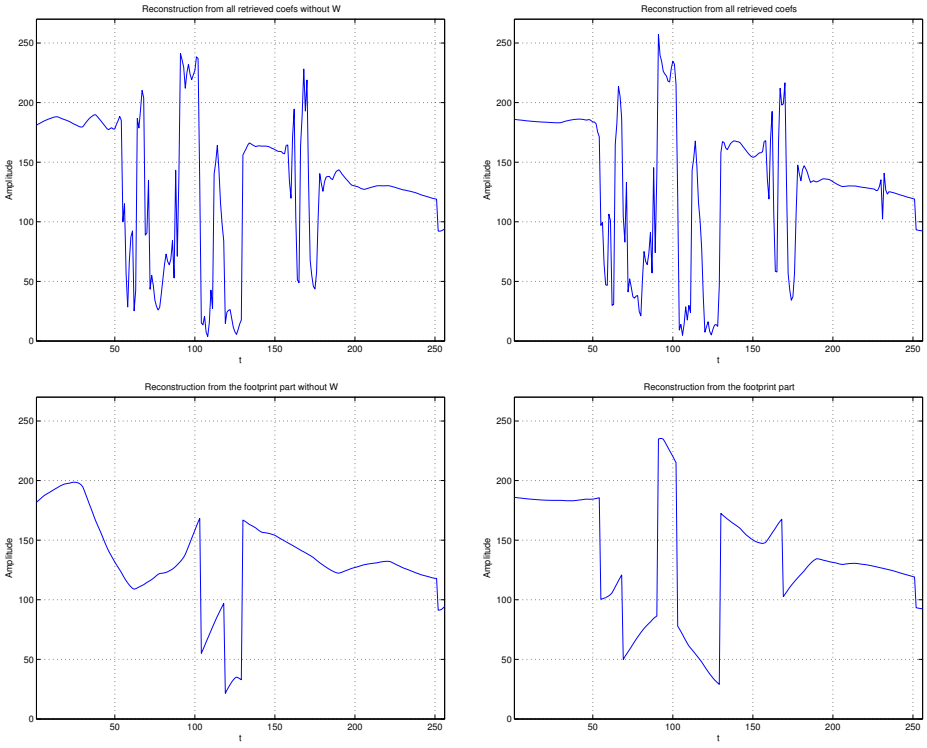

Fig. 9. Top: Approximation after 50 iterations of OMP with (right) and without (left) *a priori* information. Bottom left: Signal components captured by *Symmlet* scaling functions and Footprints using OMP. Bottom right: Signal components captured by *Symmlet* scaling functions and Footprints using Weighted-OMP.

The lower the value of $E\{r_k|\lambda', \beta'\}$, the higher the probability of the parameters to be the good ones. Hence, it can be easily observed that a unique global optimum in the parameter space exists. In this example, it looks like the set of parameters, which best fit the data model, could be easily found by some iterative procedure based on a nonlinear optimization method. Additional experimental results may be found in [15].

5) Behavior of Weighted-OMP on a Large Set of Piecewise-Smooth Signals: Finally, a larger data set has been tested in order to give a better overview on how our algorithm behaves in average. For this particular experiment, we selected 86 columns of *cameraman* (one out of three) and 46 columns from *Lena* (one out of six). Then, the average residual error for all signals is compared, as a function of the greedy iteration, for Weighted-OMP and OMP in Fig. 11 (left). On the right of Fig. 11, we also give a representation of the approximation gains supplied by the weighted algorithm for each of the selected columns of *cameraman*. This shows that, depending on the particular structure of the signal, Weighted-MP may supply very significant improvements (up to 4.5 dB better in approximation error). It also reflects the fact that, if the *prior* model is properly defined, one can not get worse results than those of the pure greedy approach.

VI. Conclusions

Sparse approximation requires the use of dictionaries capable to efficiently catch the main features and salient structures of signals. Particular applications often focus

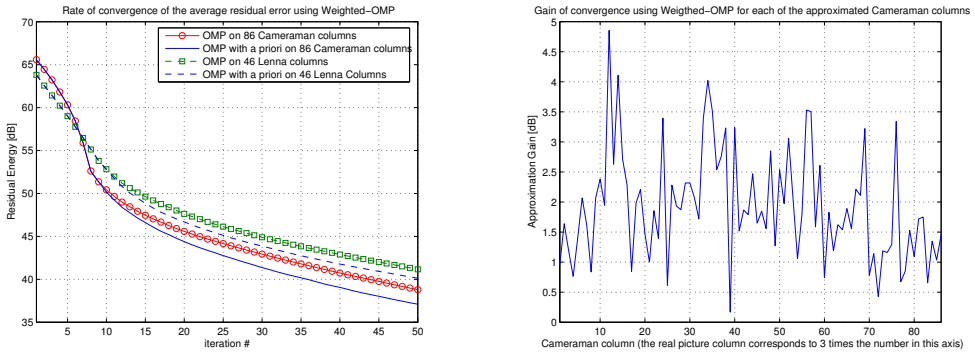

Fig. 11. Left: Average residual error convergence for Weighted-OMP and OMP for 86 columns sampled from image *cameraman* and 46 columns sampled from image Lenna. Right: Approximation gain when using Weighted-OMP depending on the column sampled from image *cameraman*.

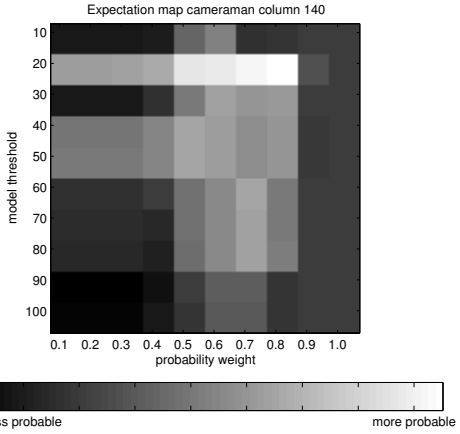

Fig. 10. Representation of the expectation map depending on the parameters that configure the *a priori* model in the experiment set up in Fig. 7. The expectation corresponds to the average energy of the residual error.

on a certain class of signals, thus a wise strategy is to use dictionaries adapted to this class. Such dictionaries often have high internal coherence, while practical algorithms for the retrieval of sparse approximations, like MP or BPDN, have been proved to work well with quasi-incoherent dictionaries. In order to overcome this conflict, adaptive subset selection algorithms are of key importance to obtain optimal m-term signal approximations. Hence, we have introduced weighted variants of the MP and BPDN algorithms called: Weighted-MP and WBPDN. Theoretical results show that these may supply much better approximations than classic approaches.

In order to guarantee this, sufficiently reliable *prior* knowledge must be used. Our practical examples show how appropriate *a priori* models may be able to characterize the interaction between signal and dictionaries, giving very good results even for highly coherent dictionaries. A possible direction to explore is to determine some bound on the quantity ϵ_{max} (i.e. the reliability factor), depending on the class of signals to approximate, the selected dictionary and the practical estimators in use (those that generate the *a priori* weights). The knowledge of some bound on ϵ_{max} for certain model may be of great help in determining in advance whether a model can be suitable for a particular application. In general, very good feature estimators exist and a lot of experience about them is available in literature. For every particular application, models exploiting particular signal features may be found in order to marry them with most common algorithms used for sparse approximations/representations. In fact, the examples presented in this work may be subject to improvement if a higher order model was used.

Some practical applications of the concepts discussed in this work are: the eventual use of edge estimators (e.g. edginess measurements) to better approximate images using multi-component edge/smooth adapted dictionaries [21], or the use of QRST point estimators for the separation of ventricular and atrial activities in electrocardiogram signals through sparse decompositions [24].

Appendix

Proof of Corollary 1 of Section III-A:

Proof: For simplicity, let us use an upper bound on the left hand side of Eq. (12). Indeed, the factor $0 < \left(w_{\overline{\Gamma}^{max}}\right)^2 \leq 1$ is removed:

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

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

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

$$\begin{aligned}\mu_1^w(m-1)+\mu_1^w(m)+\epsilon_{max} &\leq \mu_1(m-1)+\mu_1(m)<1,\\ \mu_1^w(m-1)+\epsilon_{max} &\leq \mu_1(m-1). \qquad (37)\end{aligned}$$

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

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

$$\frac{\left(1-\mu_1(m-1)\right)}{\left(1-(\mu_1(m-1)+\mu_1(m))\right)^2}.$$

Then,

$$1 \leq \frac{(1-\mu_1(m-1))}{(1-(\mu_1^w(m-1)+\epsilon_{max}))}\cdot \frac{(1-(\mu_1^w(m-1)+\mu_1^w(m)+\epsilon_{max}))^2}{(1-(\mu_1(m-1)+\mu_1(m)))^2}. \quad (38)$$

According to the relations in (37), the following can be considered:

$$\tau_1 \triangleq \mu_1(m-1)-(\mu_1^w(m-1)+\epsilon_{max}) \quad (39)$$

$$\tau_2 \triangleq \mu_1(m-1)+\mu_1(m)-(\mu_1^w(m)+\mu_1^w(m-1)+\epsilon_{max}) \quad (40)$$

where $0 \leq \tau_1 \ll \mu_1(m)$ and $0 \leq \tau_2 \ll \mu_1(m-1)+\mu_1(m)$.

Hence, being $\delta \triangleq \tau_2/(1-(\mu_1(m-1)+\mu_1(m)))$ the second fraction in (38) can be substituted:

$$1 \leq \frac{(1-\mu_1(m-1))}{(1-(\mu_1^w(m-1)+\epsilon_{max}))}\cdot(1+\delta)^2. \quad (41)$$

Moreover, being $\delta' \triangleq \tau_1/(1-\mu_1(m-1))$, the remaining fractional term of (41) may be considered such that

$$1 \leq \frac{1}{1+\delta'}\cdot(1+\delta)^2 = (1+\delta)\cdot\frac{1+\delta}{1+\delta'}. \quad (42)$$

From this, clearly $(1+\delta) \geq 1$. So, if $(1+\delta) \geq (1+\delta')$, then Corollary 1 is proved. Hence, let us check, finally, if this last condition holds. Inserting in $(1+\delta) \geq (1+\delta')$ the definitions of δ and δ' we find:

$$\frac{\tau_2}{(1-(\mu_1(m-1)+\mu_1(m)))} \geq \frac{\tau_1}{(1-\mu_1(m-1))},$$

which will be always true if $\tau_2/\tau_1 \geq 1$. Let us assume, then, that $\tau_2 \geq \tau_1$. This, together with (39) and (40), yields $(\mu_1(m)-\mu_1^w(m)) \geq 0$, which asserts all hypothesis and concludes the whole proof.

■

Acknowledgments

The authors would like to thank Rosa M. Figueras i Ventura for her help, Lorenzo Peotta for fruitful discussions and the anonymous reviewers for their valuable comments which contributed to improve the present paper.

References

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

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

[3] 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.

[4] D. L. 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.

[5] R. Gribonval and P. Vandergheynst, "On the exponential convergence of matching pursuits in quasi-incoherent dictionaries," *IEEE Trans. Inform. Theory*, to appear.

[6] J. J. Fuchs, "Sparsity and uniqueness for some specific underdetermined linear systems," in *Proc. of IEEE ICASSP '05*, vol. 5, March 2005, pp. 729–732.

[7] I. F. Gorodnitsky and B. D. Rao, "Sparse signal reconstruction from limited data using FOCUSS: A re-weighted minimum norm algorithm," *IEEE Trans. Signal Processing*, vol. 45, no. 3, pp. 600–616, Mar. 1997.

[8] C. Canton, supervised by: R. M. Figueras i Ventura, and P. Vandergheynst, "Texture and edge adaptive weak matching pursuit," Master's thesis, Ecole Polytechnique Fédérale de Lausanne, Signal Processing Institute, August 2003. [Online]. Available: http://lts2www.epfl.ch/~figueras/publications.html

[9] V. N. Temlyakov, "Weak greedy algorithms," Department of Mathematics, University of South Carolina, Columbia, Tech. Rep., 1999. [Online]. Available: http://www.math.sc.edu/~imip/99papers/9903.ps

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

[11] Y. Pati, R. Reziifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Conference Record of The Twenty-Seventh Asilomar Conference on Signals, Systems and Computers*, 1993.

[12] 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.

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

[14] 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.

[15] O. Divorra Escoda, L. Granai, and P. Vandergheynst, "On the use of *a Priori* information for sparse signal approximations," ITS/LTS-2 EPFL, Tech. Rep. 2004/23, November 2004. [Online]. Available: http://lts2www.epfl.ch/publications.html

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

[17] 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.

[18] D. L. Donoho, M. Elad, and V. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Trans. Inform. Theory*, to appear.

[19] B. Wohlberg, "Noise sensitivity of sparse signal representations: Reconstruction error bounds for the inverse problem," *IEEE Trans. Signal Processing*, vol. 51, no. 12, pp. 3053–3060, December 2003.

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

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

[22] J. J. Fuchs, "On sparse representations in arbitrary redundant bases," *IEEE Trans. Inform. Theory*, vol. 50, no. 6, pp. 1341–1344, 2004.

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

[24] O. Divorra Escoda, L. Granai, M. Lemay, J. Molinero Hernandez, P. Vandergheynst, and J.-M. Vesin, "Ventricular and atrial activity estimation through sparse ECG signal decompositions," Ecole Polytechnique Fédérale de Lausanne (EPFL), ITS/STI EPFL CH-1015 Lausanne, Switzerland, Tech. Rep. TR-ITS-2005.28, October 2005. [Online]. Available: http://lts2www.epfl.ch/publications.html