A simple test to check the optimality of sparse signal approximations

Rémi Gribonval, Rosa Maria Figueras I Ventura, Pierre Vandergheynst

A SIMPLE TEST TO CHECK THE OPTIMALITY OF A SPARSE SIGNAL APPROXIMATION

R. Gribonval

IRISA-INRIA
Campus de Beaulieu
F-35042 Rennes Cedex, France

*R. M. Figueras i Ventura *, P. Vandergheynst*

Signal Processing Institute
Swiss Federal Institute of Technology (EPFL)
CH-1015 Lausanne, Switzerland

ABSTRACT

Approximating a signal or an image with a sparse linear expansion from an overcomplete dictionary of atoms is an extremely useful tool to solve many signal processing problems. Finding the sparsest approximation of a signal from an arbitrary dictionary is a NP-hard problem. Despite of this, several algorithms have been proposed that provide sub-optimal solutions. However, it is generally difficult to know how close the computed solution is to being "optimal", and whether another algorithm could provide a better result. In this paper we provide a simple test to check whether the output of a sparse approximation algorithm is nearly optimal, in the sense that no significantly different linear expansion from the dictionary can provide both a smaller approximation error and a better sparsity. As a by-product of our theorems, we obtain results on the identifiability of sparse overcomplete models in the presence of noise, for a fairly large class of sparse priors.

1. INTRODUCTION

Recovering a sparse approximation of a signal is of great interest in many applications, such as coding [1], source separation [2] or denoising [3]. Several algorithms exist (Matching Pursuits [4, 5], Basis Pursuit [6], FOCUSS [7], etc.) that try to decompose a signal in a dictionary in a sparse way, but once the decomposition has been found, it is generally difficult to prove that the computed solution is the sparsest approximation we could obtain given a certain sparsity measure (which can be the number of terms or ℓ^0 "norm", the ℓ^1 norm, or any other metric that may lie "in between", which may be related to the bitrate needed to represent the coefficients). In this paper, we provide a general tool for checking that the solution computed by some algorithm is nearly optimal, in the sense that no significantly different sparse linear expansion from the dictionary can provide both a smaller approximation error and a better sparsity. The test quite naturally consists in checking that the residual (the difference between the signal and its sparse approximation) is "small enough" given the sparsity of the approximant and the magnitude of its smallest non-zero coefficient. When the test is satisfied, the computed solution is so close to the optimally sparse approximation –in the sense of the ℓ^0 norm– that there is an easy way to explicitly compute the latter.

The results in this paper have several implications with different levels of signification. From a numerical optimization point of view, when the test is satisfied, one knows for sure that the algorithm used to decompose the analyzed signal indeed "solved" a NP-hard problem. Since any reasonable person would use a polynomial time algorithm, this might seem contradictory at first sight (if NP $\neq$ P), but it is not: the algorithm solved *a particular instance* of the NP-hard problem, but it will fail on at least one other instance of the problem. From a modeling

*This work, which was completed while the second author was visiting IRISA, is supported in part by the European Union's Human Potential Programme, under contract HPRN-CT-2002-00285 (HASSIP).

point of view, it is often reasonable to assume that most signals in a class of interest (audio signals, natural images, etc.) belong to a "good set" of instances where the NP-hard problem can be solved in polynomial time. Indeed, as a by-product of our results, if the analyzed signal can be modeled as the superposition of a "sufficiently sparse" component and a "sufficiently small" noise, then the "sufficiently sparse" component is close to the solutions of both an (*a priori* NP-hard) ℓ^0-sparse approximation problem and a (convex) ℓ^1-sparse approximation problem, and the three of them can therefore be estimated in polynomial time by solving a convex optimization problem. This corollary of our results is in the spirit of recent work by Donoho, Elad and Temlyakov [8], Tropp [9] and Fuchs [10] on the topic of recoverability of sparse approximate overcomplete representations. However, in this paper our emphasis is on testing the near optimality of a computed sparse approximation rather than predicting the recovery of an ideal sparse model with additive noise. Several other features distinguish our contribution from the previous ones:

- previous results on recovery of sparse expansions in the noisy setting [11, 12, 8, 9, 10] make assumptions on the *ideal* sparse approximation which do not seem easy to check in practice. We provide a test that can be implemented in practice since it only depends on the *observed* sparse approximation to determine its optimality. When the test is satisfied we provide a way to recover the ideal sparse approximation (best M-term approximation).
- the test is independent of the particular algorithm used to get the sparse approximation: there is no need to make a new proof or find new optimality conditions when one introduces a new algorithm. Our emphasis is indeed on the optimality *of a decomposition* rather than on the optimality *of an algorithm*, as in the work of Wohlberg [13].
- in the case where the error is measured with the mean square error (MSE) and the dictionary is incoherent, our test is close to being sharp (see Sections 6.4-6.5). Moreover, the test is satisfied in some cases where the residual seems "too large" for the previous contributions [11, 12, 8, 9, 10] to provide conclusive results. Indeed, one of the key contributions of this paper is a new measure of the "size" of a residual which is less pessimistic than the worst-case measures based on the energy or the maximum correlation with the incoherent dictionary.
- besides the MSE, we can deal with non-quadratic distortion measures, so one could imagine to insert visual criteria if one is dealing with images, or auditive criteria if one is dealing with sounds, or any other criteria more appropriate to the data than the MSE.
- not only do we deal with the ℓ^0 and ℓ^1 sparsity measures but also with all the ℓ^τ sparsity measures[1] $\|\cdot\|_\tau^\tau$, $0 \leq \tau \leq 1$, as well as a much larger class of "admissible" measures, as discussed in Section 2.

Reading guide

In Section 2 we state the sparse approximation problem and introduce the main concepts and results. We explain the meaning of the results and discuss how they can help make connections between sparse *models*, sparse *optimization problems* and sparse approximation *algorithms*. At the end of the section we provide explicit examples to illustrate how one can use our results to build a numerical test of optimality of a sparse approximation.

The rest of the paper is devoted to the proof of the results and can be skipped by readers more interested in the test itself than in the underlying mathematics. In Section 3 we give some useful definitions and properties which are at the core of the proofs of the results. Section 4 contains proofs for the case of the canonical basis in an ℓ^p space. Section 5 provides some abstract results for arbitrary dictionary and general sub-additive distortion measures. Examples at the end of the section show how these abstract results can be used to recover results from

[1] Throughout this paper we use the notation $||x||_0^0$ to denote the ℓ^0 "norm" which counts the number of nonzero coefficients in x.

[8]. Finally, the main results with incoherent dictionaries in Hilbert spaces are proved in Section 6, where their sharpness is also discussed.

2. MAIN CONCEPTS AND RESULTS

In a finite or infinite dimensional real or complex vector space $\mathcal{H}$ (which may be a Hilbert space or more generally a Banach space) we consider $\mathbf{\Phi}$ a dictionary of atoms $\{\mathbf{g}_k\}$, which will be assumed to be normalized ($\|\mathbf{g}_k\|_{\mathcal{H}} = 1$). Using various sparse approximation algorithms (Matching Pursuits [4, 5], Basis Pursuit [6], FOCUSS [7], etc.) one can decompose a signal $\mathbf{y} \in \mathcal{H}$ as

$$\mathbf{y} = \sum_k x_k \mathbf{g}_k + \mathbf{e} \tag{1}$$

where the sequence $x = (x_k)$ is "sparse" and the residual $\mathbf{e}$ is "small". Throughout this paper, Eq. (1) will be written $\mathbf{y} = \mathbf{\Phi}x + \mathbf{e}$, where we use the same notation for the dictionary $\mathbf{\Phi}$ and the corresponding synthesis operator which maps representation coefficients to signals. In other words, we will consider the representation coefficients x and the signal $\mathbf{y}$ as column vectors and the dictionary $\mathbf{\Phi}$ as a matrix. We will use bold characters to denote signals (vectors in the space $\mathcal{H}$) and plain characters to denote coefficient sequences.

The goodness of the approximation (1) can be measured by some distortion measure $d(\mathbf{e})$ (such as a norm on $\mathcal{H}$) which only depends on the residual $\mathbf{e}$. The sparsity[2] of a representation x can be measured by an ℓ^τ norm ($0 \leq \tau \leq 1$) or more generally by an f-norm

$$\|x\|_f := \sum_k f(|x_k|), \tag{2}$$

where $f : [0, \infty) \to [0, \infty)$ is non-decreasing, not identically zero, and $f(0) = 0$. The smaller $\|x\|_f$, the sparser the representation x. The most popular sparsity measures are the ℓ^τ "norms" $\|\cdot\|_\tau^\tau = \|\cdot\|_{f_\tau}$ where $f_\tau(t) := t^\tau$ for $0 \leq \tau \leq 1$ (with the convention $0^0 := 0$ and $t^0 = 1, t > 0$) but one can imagine many other more exotic sparsity measures, see Appendix A.1. Of particular interest will be the class $\mathcal{S}$ of **sub-additive** sparsity measures which, in addition to the above properties, satisfy

$$f(t + u) \leq f(t) + f(u) \text{ for all } t, u \geq 0,$$

and the class $\mathcal{M}$ of **admissible** sparsity measures where

$$t \mapsto f(t)/t \text{ is non-increasing.}$$

It is easy to check that $\mathcal{M} \subset \mathcal{S}$, (see [14] and Appendix A.1). One can define a partial order [14] on $\mathcal{S}$ by letting $f \ll g$ if, and only if, there is some $h \in \mathcal{M}$ such that $f = h \circ g$ ($\mathcal{S}$ is stable by composition, see Appendix A.1). With respect to this partial order, the ℓ^0 and ℓ^1 "norms" are respectively the smallest and the largest admissible sparsity measures, in that $f_0 \ll f \ll f_1$ for each $f \in \mathcal{M}$.

Since different sparse approximation algorithms may optimize different sparsity criteria (ℓ^1 norm for Basis Pursuits, various ℓ^τ norms for FOCUSS, etc.), rely on various distortion measures, make a different compromise between sparsity and distortion, or even simply use a heuristic approach such as the greedy approach of Matching Pursuits, it is *a priori* hard to predict how solutions computed through different algorithms are related to one another. Our main theorems provide a simple test to check *a posteriori* if a computed decomposition $\mathbf{y} = \mathbf{\Phi}x + \mathbf{e}$ is nearly optimal, in the sense that x is close to any representation x' which is both sparser and leads to a smaller distortion.

[2] Ironically, the name of the concept is just as ubiquitous as the concept itself : the reader may have noticed that the words "sparseness" and "sparsity" are indifferently used in the literature (and in the English dictionary). We opted for"sparsity" because, as pointed out to us by M.V. Wickerhauser, it is ... sparser!

2.1. Main theorems in a Hilbert space

To state the theorems we need to introduce a few notations first. Let $\mathcal{H}$ be a Hilbert space equipped with the norm $\|\mathbf{y}\|_{\mathcal{H}}^2 = \langle \mathbf{y}, \mathbf{y} \rangle$ where $\langle \cdot, \cdot \rangle$ denotes the inner product. For each integer K we denote

$$\sigma_{\min,K}^2(\mathbf{\Phi}) := \inf_{||\delta||_0^0 \leq K} \frac{\|\mathbf{\Phi}\delta\|_{\mathcal{H}}^2}{\|\delta\|_2^2} \leq 1 \tag{3}$$

and we consider the norm (not to be confused with the ℓ^K norm $\|\cdot\|_K$)

$$|\mathbf{e}|_K := \sqrt{\sum_{k \in I_K(\mathbf{e})} |\langle \mathbf{e}, \mathbf{g}_k \rangle|^2} \tag{4}$$

where $I_K(\mathbf{e})$ indexes the K largest inner products $|\langle \mathbf{e}, \mathbf{g}_k \rangle|$. Notice that even though it is not explicit, $|\mathbf{e}|_K$ also depends on the dictionary $\mathbf{\Phi}$. Since the dictionary is generally fixed, we will indeed often simplify notations by omitting $\mathbf{\Phi}$ in some quantities that depend on it. In infinite dimension, $|\cdot|_K$ is generally not equivalent to the native norm $\|\cdot\|_{\mathcal{H}}$. However, for any integer K we have that

$$\sup_k |\langle \mathbf{e}, \mathbf{g}_k \rangle| = |\mathbf{e}|_1 \leq |\mathbf{e}|_K \leq \sqrt{K} \cdot |\mathbf{e}|_1 \leq \sqrt{K} \cdot \|\mathbf{e}\|_{\mathcal{H}},$$

so the norms $|\cdot|_K$ for different K are equivalent. Based on these definitions we can state our first result.

Theorem 1 *Let x, such that $\mathbf{y} = \mathbf{\Phi}x + \mathbf{e}$, be a sparse approximation of a signal $\mathbf{y}$, which may have been computed with* any *algorithm. Let $M := \|x\|_0^0$ and let x' be any other representation. If $\|\mathbf{y} - \mathbf{\Phi}x'\|_{\mathcal{H}} \leq \|\mathbf{y} - \mathbf{\Phi}x\|_{\mathcal{H}}$ and $\|x'\|_0^0 \leq \|x\|_0^0$, then*

$$\|x' - x\|_\infty \leq \frac{|\mathbf{e}|_1 + |\mathbf{e}|_{2M}}{\sigma_{\min,2M}^2} \tag{5}$$

$$\|x' - x\|_2 \leq \frac{2 \cdot |\mathbf{e}|_{2M}}{\sigma_{\min,2M}^2} \tag{6}$$

$$\|\mathbf{\Phi}x' - \mathbf{\Phi}x\|_{\mathcal{H}} \leq \frac{2 \cdot |\mathbf{e}|_{2M}}{\sqrt{\sigma_{\min,2M}^2}} \tag{7}$$

Generally, the bound (7) is better than the "worst case" one $\|\mathbf{\Phi}x' - \mathbf{\Phi}x\|_{\mathcal{H}} = \|(\mathbf{y} - \mathbf{\Phi}x') - (\mathbf{y} - \mathbf{\Phi}x)\|_{\mathcal{H}} \leq 2\|\mathbf{e}\|_{\mathcal{H}}$. We will comment on this in Section 5.

A few additional definitions are needed to state our second result, which is stronger since it is valid for *any* admissible sparsity measure. We let $\mathbf{\Phi}_I : \ell^2(I) \to \mathcal{H}$ denote the synthesis matrix associated to the sub-dictionary $\{\mathbf{g}_k, k \in I\}$ and $\mathbf{\Phi}_I^+ = (\mathbf{\Phi}_I^\star \mathbf{\Phi}_I)^{-1} \mathbf{\Phi}_I^\star$ be its Moore-Penrose pseudo-inverse, where $(\cdot)^\star$ denotes the adjoint (or in matrix terminology the complex conjugate transpose), that is to say $\mathbf{\Phi}^\star \mathbf{e} := (\langle \mathbf{e}, \mathbf{g}_k \rangle)_k$. Then, much inspired by the Exact Recovery Coefficient introduced in [9] we consider

$$\lambda_M(\mathbf{\Phi}) := 1 - \sqrt{M} \cdot \sup_{\text{card}(I) \leq M} \sup_{k \notin I} \|\mathbf{\Phi}_I^+ \mathbf{g}_k\|_2. \tag{8}$$

Theorem 2 *Let x, such that $\mathbf{y} = \mathbf{\Phi}x + \mathbf{e}$, be a sparse approximation of a signal $\mathbf{y}$, which may have been computed with* any *algorithm. Let $M := \|x\|_0^0$ and assume that $\lambda_M(\mathbf{\Phi}) > 0$. Let x' be any other representation: if*

$\|\mathbf{y}-\mathbf{\Phi}x'\|_{\mathcal{H}} \leq \|\mathbf{y}-\mathbf{\Phi}x\|_{\mathcal{H}}$ *and if there exists* some *admissible sparsity measure* f *such that* $\|x'\|_f \leq \|x\|_f$*, then*

$$\|x'-x\|_\infty \leq \frac{2}{\lambda_M^2}\cdot\frac{|\mathbf{e}|_1+|\mathbf{e}|_M}{\sigma^2_{\min,M}} \tag{9}$$

$$\|x'-x\|_2 \leq \frac{4\cdot\sqrt{1+M}\cdot|\mathbf{e}|_M}{\sigma^2_{\min,M}\cdot\lambda_M^2} \tag{10}$$

$$\|\mathbf{\Phi}x'-\mathbf{\Phi}x\|_{\mathcal{H}} \leq \frac{4\cdot|\mathbf{e}|_M}{\sqrt{\sigma^2_{\min,M}\cdot\lambda_M^2}} \tag{11}$$

Note that, in Eqs. (9)-(11), $2M$ has been replaced with M in the subscripts for $|\mathbf{e}|$ and $\sigma^2_{\min,\cdot}$ compared to Eqs. (5)-(7).

Corollary 1 (Test of ℓ^0 optimality) *Under the hypotheses and notations of Theorem 1, assume that*

$$|\mathbf{e}|_1+|\mathbf{e}|_{2M} < \sigma^2_{\min,2M}\cdot\min_{\{k:|x_k|\neq 0\}}|x_k|. \tag{12}$$

If x' *satisfies* $\|\mathbf{y}-\mathbf{\Phi}x'\|_{\mathcal{H}} \leq \|\mathbf{y}-\mathbf{\Phi}x\|_{\mathcal{H}}$ *and* $\|x'\|_0^0 \leq \|x\|_0^0$*, then* x' *and* x *have the same "support":*

$$\textit{support}(x') := \{k : |x'_k| \neq 0\} = \{k : |x_k| \neq 0\} = \textit{support}(x)$$

and the same sign: $\textit{sign}(x_k) = \textit{sign}(x'_k)$*, for all* k*.*

Corollary 2 (Test of strong optimality) *Under the hypotheses and notations of Theorem 2, assume that*

$$|\mathbf{e}|_1+|\mathbf{e}|_M < \frac{\sigma^2_{\min,M}\cdot\lambda_M^2}{4}\cdot\min_{\{k:|x_k|\neq 0\}}|x_k|. \tag{13}$$

If x' *satisfies* $\|\mathbf{y}-\mathbf{\Phi}x'\|_{\mathcal{H}} \leq \|\mathbf{y}-\mathbf{\Phi}x'\|_{\mathcal{H}}$ *and if there exists* some *admissible sparsity measure* f *such that* $\|x'\|_f \leq \|x\|_f$*, then* x' *and* x *have essentially the same support:*

$$\{k : |x'_k| > \theta\} = \textit{support}(x),\ \textit{with}\ \theta := \frac{1}{2}\min_{\{k:|x_k|\neq 0\}}|x_k|.$$

Moreover for $k \in \textit{support}(x)$ *we have* $\textit{sign}(x_k) = \textit{sign}(x'_k)$*.*

Corollary 3 (Solution of the NP-hard problem) *If* x *and* $\mathbf{e}$ *satisfy either the test* (12) *or the test* (13)*, then the best* M*-term approximation* $\mathbf{\Phi}x_M$ *to* $\mathbf{y} = \mathbf{\Phi}x+\mathbf{e}$ *is exactly the orthogonal projection of* $\mathbf{y}$ *onto span*$\{\mathbf{g}_k, k \in \textit{support}(x)\}$*.*

The first corollary can be proved by combining Eq. (5) of Theorem 1 with Eq. (12): we get that $\|x'-x\|_\infty < \min_{\{k:|x_k|\neq 0\}}|x_k|$, which implies that the two sequences have the same support and sign. The proof for Corollary 2 is done similarly, using Theorem 2.

Remark 1 *The tests proposed in Corollaries 1-2 are reminiscent of some results of Tropp [9, Correlation Condition Lemma, Theorem 5.2], but with* $\sup_k|\langle\mathbf{e},\mathbf{g}_k\rangle| = |\mathbf{e}|_1$ *replaced with* $|\mathbf{e}|_1+|\mathbf{e}|_K$ *for* $K \in \{M, 2M\}$*.*

2.2. The meaning of these results

Our results (sometimes) make it possible to clarify the connections between sparse models, sparse optimization problems and sparse approximation algorithms:

- A model is a description of how a signal could have been generated, typically with a probabilistic prior in the Bayesian point of view or with parameters.
- A problem is an optimization problem, independently of how hard it is, what algorithm can solve it, etc. A problem can correspond to a model if e.g. it is the maximum likelihood (ML) or maximum a posteriori (MAP) estimation of model parameters, but it can also be difficult to make a model that fits a given problem.
- An algorithm is a function that takes an input and computes an output, a computer program, independently of which problem it can solve.

If some algorithm has decomposed a signal $\mathbf{y}$ as $\mathbf{y} = \mathbf{\Phi}x + \mathbf{e}$ where $M := \|x\|_0^0$ is such that $\lambda_M > 0$ and $\lambda_M^2 \sigma_{\min,M}^2 > 0$, then Theorem 2 tells us that the computed coefficients x are "not too far" from the solutions of each of the (generally non-convex) optimization problems

$$\min_{x'} \|x'\|_f \text{ subject to } \|\mathbf{y} - \mathbf{\Phi}x'\|_{\mathcal{H}} \leq \epsilon \tag{14}$$

with $\epsilon := \|\mathbf{y} - \mathbf{\Phi}x\|_{\mathcal{H}} = \|\mathbf{e}\|_{\mathcal{H}}$. Therefore:

- for the input signal $\mathbf{y}$, the problems (14) for different sparsity measures f have solutions close to one another;
- on the input signal $\mathbf{y}$, the algorithm which produced the decomposition $\mathbf{y} = \mathbf{\Phi}x + \mathbf{e}$ nearly solved each of these problems.

Corollary 3 shows that if the residual is small enough, one can actually directly "jump" from the computed coefficients x to the solution x_0 of the NP-hard best M-term approximation problem

$$\min_{x'} \|\mathbf{y} - \mathbf{\Phi}x'\|_{\mathcal{H}} \text{ subject to} \|x'\|_0^0 \leq M.$$

Now, assume that an observed signal $\mathbf{y}$ follows the sparse model $\mathbf{y} = \mathbf{\Phi}x + \mathbf{e}$ where (with high probability) $\|x\|_0^0 \leq M$ and $\|\mathbf{e}\|_{\mathcal{H}} = \epsilon$. One can consider the problem of estimating the coefficients x or the signal $\mathbf{\Phi}x$, which is a classical denoising problem. Theorem 2 shows that (with high probability)

- one can robustly estimate x by solving any of the sparse approximation problems (14);
- in particular, one can robustly estimate x by solving the convex ℓ^1-minimization problem

$$\min_{x'} \|x'\|_1 \text{ subject to } \|\mathbf{y} - \mathbf{\Phi}x'\|_{\mathcal{H}} \leq \epsilon \tag{15}$$

  which can be done using any Quadratic Programming algorithm.

Solving the problem (14) is equivalent to solving the Lagrangian problem

$$\min_{x'} \|\mathbf{y} - \mathbf{\Phi}x'\|_{\mathcal{H}}^2 + \alpha\|x'\|_f \tag{16}$$

for an appropriate Lagrange multiplier α. Thus, the solution of (15) is also the Maximum A Posteriori (MAP) estimate of x under a Laplacian model on x and a Gaussian model on $\mathbf{e}$ and we conclude that replacing the original sparse model with a Laplacian+Gaussian model does not significantly change the value of the estimate, yet it simplifies a lot its computation.

In practice, just as in [8, 9] a crucial practical problem is to estimate the noise level ϵ, which is unknown, or equivalently to tune the Lagrange multiplier used in the Quadratic Programming algorithm that solves (16). Further work is needed to investigate what can be said about the accuracy of the estimate when the exact noise level is unknown.

2.3. Explicit tests of optimality in a Hilbert space

To apply these tests in practice, we need to compute explicit (lower) estimates of the numbers $\lambda_M(\mathbf{\Phi})$ and $\sigma^2_{\min,K}(\mathbf{\Phi})$, which –so far– may seem fairly abstract. For M sufficiently small we obtain such estimates using the Babel function $\mu_1(M,\mathbf{\Phi})$, defined in [11, 9] as

$$\mu_1(M,\mathbf{\Phi}) := \sup_{\text{card}(I)\leq M} \sup_{k\notin I} \sum_{i\in I} |\langle \mathbf{g}_k, \mathbf{g}_i\rangle| \tag{17}$$

as well as the 2-Babel function which we define as

$$\mu_2(M,\mathbf{\Phi}) := \sup_{\text{card}(I)\leq M} \sup_{k\notin I} \sqrt{\sum_{i\in I} |\langle \mathbf{g}_k, \mathbf{g}_i\rangle|^2}. \tag{18}$$

Proposition 1 *Let $\mathbf{\Phi}$ be a normalized dictionary in a Hilbert space $\mathcal{H}$. If $\mu_1(2M-1) < 1$ then*

$$\sigma^2_{\min,2M} \geq 1 - \mu_1(2M-1) > 0 \tag{19}$$

If $\sqrt{M}\mu_2(M) + \mu_1(M-1) < 1$ then $\lambda_M > 0$ and

$$\sigma^2_{\min,M} \cdot \lambda^2_M \geq \frac{\left(1 - \sqrt{M}\mu_2(M) - \mu_1(M-1)\right)^2}{1-\mu_1(M-1)} \tag{20}$$

The test can be done by applying the limit for $\sigma^2_{\min,M} \cdot \lambda^2_M$, given by Eq. (20), to Corollary 2. It has to be taken into account that if the test is positive, you are sure you have the sparsest solution. On the other hand, you may have a negative test and still have the sparsest solution. The test is thus sharper than similar tests presented in previous works, and it has the advantage that it is algorithm independent.

2.4. Examples

Orthonormal basis

When $\mathbf{\Phi}$ is an orthonormal basis, we have $\mu_1(M) = 0$ and $\mu_2(M) = 0$ for all M, hence the test of ℓ^0 optimality takes the simple form

$$|\mathbf{e}|_1 + |\mathbf{e}|_{2M} < \min_{\{k:|x_k|\neq 0\}} |x_k|$$

which turns out to be sharp (see Section 4). The test of strong optimality becomes

$$|\mathbf{e}|_1 + |\mathbf{e}|_{M} < \min_{\{k:|x_k|\neq 0\}} |x_k|/4$$

and it it also sharp up to a constant factor (see Sections 6.4-6.5).

Union of incoherent orthonormal bases

When $\mathbf{\Phi}$ is a union of two or more maximally incoherent orthonormal bases in $\mathbb{C}^N$, such as the Dirac basis, the Fourier basis ($\mathbf{g}_k[n] = \exp(2i\pi kn/N)$, $0 \leq k \leq N-1$) and the Chirp basis ($\mathbf{g}_k[n] = \exp(2i\pi kn^2/(2N))$, $0 \leq k \leq N-1$), we have $\mu_1(M) = M/\sqrt{N}$ and $\mu_2(M) = \sqrt{M}/\sqrt{N}$. It follows that, for $M \leq (1+\sqrt{N}/3)/2$, we have $\sigma^2_{\min,2M} \geq 2/3$ and $\sigma^2_{\min,M} \cdot \lambda^2_M \geq 4/9$, so the conclusions of Corollary 1 (resp. Corollary 2) hold if

$$|\mathbf{e}|_1 + |\mathbf{e}|_{2M} < \min_{\{k:|x_k|\neq 0\}} |x_k| \cdot 2/3$$

or

$$|\mathbf{e}|_1 + |\mathbf{e}|_{M} < \min_{\{k:|x_k|\neq 0\}} |x_k| \cdot 1/9,$$

respectively.

More generally, since in this case $\mu_1(2M-1) = \sqrt{M}\mu_2(M)+\mu_1(M-1) = (2M-1)/\sqrt{N}$, one can apply the tests whenever $M < (1+\sqrt{N})/2$. As M gets closer to $(1+\sqrt{N})/2$, the bound on the allowed size of the residual decreases and the tests becomes more restrictive. For three maximally incoherent orthonormal bases in dimension $N = 1024$ and $M \leq 16 < \frac{33}{2}$, both tests can be applied to guarantee the optimality of a sparse approximation. In comparison, without such a test, one would have to compare the quality of the observed approximation with that of $\binom{M}{3N} - 1$ other M-term approximations.

A numerical example

Let us now give a numerical example to illustrate how the test can be applied in practice. To mimic "musical notes" and transients" of audio signal, consider a dictionary $\mathbf{\Phi} = [\mathrm{C}, \mathrm{E}, \mathrm{G}, \mathrm{B}, I_{25}]$, which is the union of an orthonormal basis of deltas with a set of 5 (normalized) sinusoids, that is to say I_N is the identity matrix of dimension N and $\mathrm{C}, \mathrm{E}, \mathrm{G}, \mathrm{B}$ are unit vectors proportional to the sinusoids

$$\sin\left(\frac{2\pi \cdot \mathrm{freq} \cdot n}{2048}\right), \quad n = 1 \ldots 25. \tag{21}$$

with respective frequencies $\mathrm{freq} = 261, 330, 392$ and 494. These frequencies correspond to the fundamental frequency of the notes C, E, G, B sampled at $2048Hz$ (this is obviously an unrealistic sampling frequency for audio signals but the example is rather a toy for illustration here). For this dictionary, it is easy to compute the first values of the Babel functions. In particular, we can compute $\mu_1(1) \approx 0.2831$ and $\mu_2(2) \approx 0.3998$ and using Eq. (20) we obtain that $\sigma^2_{\mathrm{min},2} \cdot \lambda_2^2 > 0.2113$.

Given a signal of 25 samples $\mathbf{y}$ (which we generated as a superimposition of the 'note' C, with coefficient 15 and the 'note' E, with coefficient 10, together with some additive Gaussian white noise), we can use various sparse approximation algorithms to decompose it in the dictionary $\mathbf{\Phi}$. We performed $M = 2$ steps of Orthogonal Matching Pursuit (OMP) and found that $\mathbf{y} = \mathbf{\Phi}x + \mathbf{e}$, with

$$x \approx [15.0080,\ 10.0276,\ 0, \ldots,\ 0]^T \tag{22}$$

Using the correlations of the residual $\mathbf{e}$ with the atoms of the dictionary (which were already computed as a natural by-product of the two steps of OMP) we computed $|\mathbf{e}|_1 \approx 0.0936$ and $|\mathbf{e}|_M = |\mathbf{e}|_2 \approx 0.1311$. Thus, we were able to check that

$$|\mathbf{e}|_1 + |\mathbf{e}|_M \approx 0.2247 < \frac{0.2213 \cdot 10.0276}{4}$$

which ensures that the strong optimality test is satisfied. Thus, we are sure that the two atoms found by OMP are exactly the two atoms of the best 2 term approximation to the signal. After an additional iteration of OMP, we obtained an $M = 3$ term approximation to the signal, and it turned out that the new residual and the coefficients no longer satisfied the test (because the smallest coefficient in x was of the same order as the noise level). Thus, we were no longer sure that the three term expansion provided by three iterations of OMP was "optimal" and we stopped the iterations. We are investigating how this could be used in more realistic examples to build a good stopping criterion for Matching Pursuit.

Some remarks about the test

To conclude this section, let us illustrate the behavior of our test as a function of the number of terms M in the sparse approximant, using a simple yet perhaps non intuitive example. In $\mathcal{H} = \mathbb{R}^6$, consider a signal $\mathbf{y}$ which has an exact representation $\mathbf{y} = \mathbf{\Phi}x$ on an orthonormal basis $\mathbf{\Phi}$ with $x = [a, b, b, c, d, e]^T$, where $|a| > |b| > |c| > |d| > |e|$. The best one-term approximant to $\mathbf{y}$ from $\mathbf{\Phi}$ – which is obtained, *e.g.*, by one step of Matching Pursuit

– corresponds to $x_1 = [a, 0, 0, 0, 0, 0]^T$. It satisfies the strong test (13) as soon as $2|b| < |a|/4$. For a two-term sparse approximation such as $x_2 = [a, b, 0, 0, 0, 0]$, the strong test is never satisfied since $|b| + \sqrt{b^2 + c^2} \geq |b|/4$: actually, even though x_2 is *an* optimal two-term approximation, it is not *the unique* one. Now, a three-term sparse approximation with $x_3 = [a, b, b, 0, 0, 0]$ will again satisfy the strong test provided that $|c| + \sqrt{c^2 + d^2 + e^2} < |b|/4$. This example shows that it is not possible to define a "breakpoint" of the test: it is simply not true that the test is satisfied for small enough M up to a breakpoint and not satisfied for M larger than the breakpoint. Instead, the test can provide information on the optimality of a given sparse approximation of a given signal with a given number of terms M.

3. CORE ELEMENTS OF THE PROOFS

Now that we have stated the main results and explained how they can be used, let us introduce some technical definitions and lemmas which are at the core of the proof of our theorems. This section is stated in the most general setting, and we will see later on how some quantities can be estimated in specific cases.

Let x, with $\mathbf{y} = \mathbf{\Phi}x + \mathbf{e}$, be a sparse approximation of a signal $\mathbf{y}$. Let $M := \|x\|_0^0$ and assume that, for a fixed f, x' satisfies $d(\mathbf{y} - \mathbf{\Phi}x') \leq d(\mathbf{y} - \mathbf{\Phi}x)$ and $\|x'\|_f \leq \|x\|_f$. Letting $\delta := x' - x$, we see that $\delta \in D_d(\mathbf{e}, \mathbf{\Phi}) \cap C_f(\mathrm{X}_M)$ with

$$D_d(\mathbf{e}, \mathbf{\Phi}) := \left\{ \delta : d(\mathbf{e} - \mathbf{\Phi}\delta) \leq d(\mathbf{e}) \right\} \tag{23}$$

$$C_f(\mathrm{X}) := \bigcup_{z \in \mathrm{X}} \{\delta : \|z + \delta\|_f \leq \|z\|_f\} \tag{24}$$

and $\mathrm{X}_M := \{x : \|x\|_0^0 \leq M\}$. Thus, for $0 < q < \infty$ we have

$$\|x' - x\|_q \leq \sup_{\delta \in D_d(\mathbf{e}, \mathbf{\Phi}) \cap C_f(\mathrm{X}_M)} \|\delta\|_q, \tag{25}$$

and

$$d(\mathbf{\Phi}x' - \mathbf{\Phi}x) \leq \sup_{\delta \in D_d(\mathbf{e}, \mathbf{\Phi}) \cap C_f(\mathrm{X}_M)} d(\mathbf{\Phi}\delta). \tag{26}$$

In the following we will simply denote $D(\mathbf{e})$ since $\mathbf{\Phi}$ and $d(\cdot)$ are generally fixed. The results of this paper follow from upper estimates of the suprema in Eqs. (25)-(26), using the following lemma:

Lemma 1 *If f is a sub-additive sparsity measure then*

$$C_f(X_M) = \left\{ \delta : \sum_{k \in I_M(\delta)} f(|\delta_k|) \geq \frac{\|\delta\|_f}{2} \right\} \tag{27}$$

where $I_M(\delta)$ is the set of the M largest components of $|\delta_k|$.

We postpone the proof to Appendix A.2 to keep the flow of the argument. By [14, Lemma 7], for any admissible sparsity measure $h \in \mathcal{M}$, any sequence $z = (z_k)$ and any integer M, we have

$$\frac{\sum_{k \in I_M(z)} h(|z_k|)}{\|z\|_h} \leq \frac{\sum_{k \in I_M(z)} |z_k|}{\|z\|_1}. \tag{28}$$

Thus, whenever $f \ll g$ we have $C_f(\mathrm{X}_M) \subset C_g(\mathrm{X}_M)$ and it follows that

$$\sup_{\delta \in D(\mathbf{e}) \cap C_f(\mathrm{X}_M)} ||\delta||_q \leq \sup_{\delta \in D(\mathbf{e}) \cap C_g(\mathrm{X}_M)} ||\delta||_q \tag{29}$$

$$\sup_{\delta \in D(\mathbf{e}) \cap C_f(\mathrm{X}_M)} d(\mathbf{\Phi}\delta) \leq \sup_{\delta \in D(\mathbf{e}) \cap C_g(\mathrm{X}_M)} d(\mathbf{\Phi}\delta). \tag{30}$$

Since $f_0 \ll f \ll f_1$ for every admissible sparsity measure f, we will only estimate the suprema in Eqs. (25)-(26) for $f = f_0$ and $f = f_1$.

4. ESTIMATES IN THE CANONICAL BASIS

Estimating the right hand side suprema in Eq. (29) (for $q = \infty$ and $q = 2$) and Eq. (30) with $d(\cdot)$ the Hilbertian norm immediately yields Theorem 1 and Theorem 2. Since these estimates are a bit technical, we postpone them for a while and begin with an estimate for the simple case where $\mathbf{\Phi}$ is the canonical basis in an ℓ^p sequence space. This estimate is both illustrative and technically useful, since it provides the basic tools to obtain the more general estimates for arbitrary dictionaries in Hilbert spaces.

Lemma 2 *Consider $\mathbf{\Phi} = \mathbf{B}$ the canonical basis in ℓ^p, $0 < p \leq \infty$, $d(\cdot) = \|\cdot\|_p$ and let $K \geq 1$. When $0 < p < \infty$ we have for any $b = (b_k) \in \ell^p$:*

$$\sup_{\substack{||\delta||_0^0 \leq K \\ d(b - \mathbf{B}\delta) \leq d(b)}} ||\delta||_\infty = \sup_k |b_k| + \left(\sum_{k \in I_K(b)} |e_k|^p \right)^{1/p} \tag{31}$$

where $I_K(b)$ indexes the K largest components $|b_k|$. Similarly when $p = \infty$ we have

$$\sup_{\substack{||\delta||_0^0 \leq K \\ d(b - \mathbf{B}\delta) \leq d(b)}} ||\delta||_\infty = 2 \cdot \sup_k |b_k|. \tag{32}$$

Remark 2 *Notice that in the above Lemma, and in all this section, b denotes both a signal and a sequence of coefficients, hence according to our convention we could write it either in bold or in plain letters. We chose the plain notation because when the lemma will be used later on we will rather consider b as a sequence.*

Since $\delta \in C_{f_0}(\mathrm{X}_M)$ if, and only if, $||\delta||_0^0 \leq 2M$, the lemma provides the exact value of $\sup_{\delta \in D(\mathbf{e}) \cap C_{f_0}(\mathrm{X}_M)} ||\delta||_\infty$ by letting $K = 2M$ in Eqs. (31)-(32). For $p = 2$, if $b_k = \langle \mathbf{e}, \mathbf{g}_k \rangle$ for some dictionary $\mathbf{\Phi}$ and some residual $\mathbf{e}$ in a Hilbert space, it is not difficult to check that the value is exactly $|\mathbf{e}|_1 + |\mathbf{e}|_{2M}$ (see Eq. (4)).

Proof of Lemma 2. We begin with the case $0 < p < \infty$. Let δ with $||\delta||_0^0 \leq K$ and consider $I = \mathrm{support}(\delta)$. If $||b - \mathbf{B}\delta||_p \leq ||b||_p$ we have $\sum_{k \in I} |b_k - \delta_k|^p \leq \sum_{k \in I} |b_k|^p$. Thus, for any $j \in I$, we have

$$|b_j - \delta_j| \leq \left(\sum_{k \in I} |b_k|^p \right)^{1/p}$$

which implies

$$|\delta_j| \leq |b_j| + \left(\sum_{k \in I} |b_k|^p \right)^{1/p} \leq \sup_k |b_k| + \left(\sum_{k \in I_K(b)} |b_k|^p \right)^{1/p}. \tag{33}$$

Thus, the left hand side in Eq. (31) is no larger than the right hand side. To get the converse inequality let $I = I_K(b)$ and $j \in I$ with $|b_j| = \sup_k |b_k|$, let $\delta_k := 0, k \notin I, \delta_k := b_k, k \in I, k \neq j$, and

$$\delta_j := \mathrm{sign}(b_j) \cdot \left(|b_j| + \left(\sum_{k \in I} |b_k|^p \right)^{1/p} \right).$$

Obviously $\|\delta\|_0^0 \leq K$, $\|b - \mathbf{B}\delta\|_p \leq \|b\|_p$ and $\|\delta\|_\infty$ is no smaller than the right hand side in Eq. (31). The case $p = \infty$ is even easier: if $\|b - \mathbf{B}\delta\|_\infty \leq \|b\|_\infty$ then

$$\|\delta\|_\infty = \|b - \mathbf{B}\delta - b\|_\infty \leq \|b - \mathbf{B}\delta\|_\infty + \|b\|_\infty \leq 2\|b\|_\infty,$$

thus the left hand side in Eq. (32) is no larger than the right hand side. To get the converse inequality, assume for the sake of simplicity that there is an index l such that $|b_l| = \|b\|_\infty$. Letting $\delta_l = -2b_l$ and $\delta_k = 0, k \neq l$ we get a sequence $\delta = (\delta_k)$ which satisfies $\|b - \mathbf{B}\delta\|_\infty = \|b\|_\infty$ and $\|\delta\|_0^0 = 1 \leq K$. We let the reader check that the argument can be adapted to the case where the ℓ^∞ norm is not attained.

5. SUB-ADDITIVE DISTORTION MEASURES

When $\mathbf{\Phi}$ is not the canonical basis or $d(\cdot)$ is not an ℓ^p norm, it is difficult to get exact estimates. Here, we investigate "worst case" upper estimates for sub-additive distortion measure, *i.e.* when for any $\mathbf{u}$ and $\mathbf{v}$ we have $d(\mathbf{u} + \mathbf{v}) \leq d(\mathbf{u}) + d(\mathbf{v})$. By worst case, we mean that instead of using the full knowledge of the residual we only summarize it with the distortion $d(\mathbf{e})$ to which it corresponds. Thus, the estimates will tell us what happens with the "worst" residual which yields the same distortion. For the quadratic distortion measure, this approach will recover known results [8, 9, 10] on the identification of sparse approximations, since these were obtained using a worst case approach. However, we will see in the Section 6 how to prove Theorems 1-2 which provide much more precise bounds than the worst case ones.

Lemma 3 *Let $d(\cdot)$ be a sub-additive distortion measure. Then*

$$\sup_{\delta \in D(\mathbf{e})} d(\mathbf{\Phi}\delta) \leq 2 \cdot d(\mathbf{e}). \tag{34}$$

Proof. For any $\delta \in D(\mathbf{e})$, we have $d(\mathbf{e} - \mathbf{\Phi}\delta) \leq d(\mathbf{e})$, and by the sub-additivity of $d(\cdot)$ we get

$$d(\mathbf{\Phi}\delta) = d(\mathbf{e} - (\mathbf{e} - \mathbf{\Phi}\delta)) \leq d(\mathbf{e}) + d(\mathbf{e} - \mathbf{\Phi}\delta) \leq 2 \cdot d(\mathbf{e}).$$

□

Building upon this result we get general (but somewhat abstract) upper estimates of the suprema in (25). Denoting

$$A_q(C, \epsilon) := \sup_{\substack{z \in C \\ d(\mathbf{\Phi}z) \leq \epsilon}} \|z\|_q, \qquad \epsilon \geq 0, \tag{35}$$

Lemma 3 can be rewritten as:

$$\sup_{\delta \in D(\mathbf{e}) \cap C_f(\mathrm{X}_M)} \|\delta\|_q \leq A_q \left(C_f(\mathrm{X}_M), 2 \cdot d(\mathbf{e}) \right). \tag{36}$$

This theoretically allows us to express results similar to Theorem 1-2 for general sub-additive distortion measures. However, it is an euphemism to say that the numbers $A_q((C_f(\mathrm{X}_M), \epsilon)$ are not straightforward to compute. Even though estimating them for specific distortion measures (such as those designed to model auditive or visual

distortion criteria) would be quite interesting, it is beyond the scope of this paper and we focus instead on a more restricted case where more can be said. If $d(\cdot)$ is not only sub-additive but it is indeed a norm, then $d(\lambda \mathbf{u}) = |\lambda| \cdot d(\mathbf{u})$ for any λ and $\mathbf{u}$, hence we have

$$A_q(C_f(\mathrm{X}_M), \epsilon) = \epsilon \cdot A_q(C_f(\mathrm{X}_M), 1) =: \epsilon \cdot A_q(C_f(\mathrm{X}_M))$$

and we get an analogue to Theorem 1 and Corollary 1.

Theorem 3 *Let $d(\cdot)$ be a norm. Let x, such that $\mathbf{y} = \mathbf{\Phi} x + \mathbf{e}$, be a sparse approximation of a signal $\mathbf{y}$, which may have been computed with* any *algorithm. Let $M := \|x\|_0^0$ and let x' be any other representation. If $d(\mathbf{y} - \mathbf{\Phi} x') \leq d(\mathbf{e})$ and $\|x'\|_0 \leq \|x\|_0$, then*

$$d(\mathbf{\Phi} x' - \mathbf{\Phi} x) \leq 2 \cdot d(\mathbf{e}) \tag{37}$$

and for any $0 < q \leq \infty$

$$\|x' - x\|_q \leq 2 \cdot A_q(C_{f_0}(X_M)) \cdot d(\mathbf{e}). \tag{38}$$

Corollary 4 (Test of ℓ^0 optimality) *Under the hypotheses and notations of Theorem 3, assume that*

$$d(\mathbf{e}) < \frac{\min_{\{k:|x_k| \neq 0\}} |x_k|}{4 \cdot A_\infty(C_{f_0}(X_M)))}. \tag{39}$$

If x' satisfies $d(\mathbf{y} - \mathbf{\Phi} x') \leq d(\mathbf{y} - \mathbf{\Phi} x)$ and $\|x'\|_0^0 \leq \|x\|_0^0$, then x' and x have the same support and sign.

This corollary is proved from Theorem 3 just as Corollary 1 is proved from Theorem 1. We let the reader express what would be the analogue of Theorem 2 and Corollaries 2-3. Even with $d(\cdot)$ a norm, computing $A_q(C_f(\mathrm{X}_M))$ (or more realistically estimating it from above) seems difficult. Let us consider two illustrative examples to see how one can address the estimation with two particular norms of interest: the quadratic distortion, and the maximum correlation with the dictionary vectors.

Example 1 *When $d(\cdot) = \|\cdot\|_{\mathcal{H}}$ with $\mathcal{H}$ a Hilbert space, the reader can easily check that for $q \geq 2$ we have*

$$A_q(C_{f_0}(X_M)) \leq \sup_{\substack{\|\delta\|_0^0 \leq 2M \\ \|\mathbf{\Phi}\delta\|_{\mathcal{H}} \leq 1}} \|\delta\|_2 = \frac{1}{\sqrt{\sigma^2_{\min,2M}(\mathbf{\Phi})}}.$$

Combining the results obtained so far with Proposition 1, we recover [8, Theorem 2.1]: if $\mathbf{y}$ has a representation $\mathbf{y} = \mathbf{\Phi} x + \mathbf{e}$ with $M := \|x\|_0^0 < (1 + 1/\mu_1(1))/2$, and $\|\mathbf{e}\|_{\mathcal{H}} = \epsilon$, then the solution x_0 of the optimization problem

$$\min_{x'} \|x'\|_0^0 \ \text{ subject to } \ \|\mathbf{y} - \mathbf{\Phi} x'\|_{\mathcal{H}} \leq \epsilon \tag{40}$$

satisfies

$$\|x_0 - x\|_2^2 \leq \frac{4\epsilon^2}{1 - \mu_1(2M-1)} \leq \frac{4\epsilon^2}{1 - \mu_1(1) \cdot (2M-1)} \tag{41}$$

Notice that, compared to our own Theorem 1, $\|x_0 - x\|_2$ is upper estimated by $\|\mathbf{e}\|_{\mathcal{H}}$ instead of $2 \cdot |\mathbf{e}|_{2M}$, and we will see in Sections 6.4-6.5 that our estimate can give a much smaller bound than the worst case estimate. A similar analysis with an estimate of $A_q(C_{f_1}(X_M))$ would recover a result similar to [8, Theorem 3.1].

Example 2 *In a Hilbert space $\mathcal{H}$, when $d(\cdot) = \sup_k |\langle \cdot, \mathbf{g}_k \rangle| = \|\mathbf{\Phi}^\star \cdot\|_\infty$ we have, for $2 \leq q \leq \infty$,*

$$A_q(C_{f_0}(X_M)) \leq \sup_{\substack{\|\delta\|_0^0 \leq 2M \\ d(\mathbf{\Phi}\delta) \leq 1}} \|\delta\|_2 \leq \frac{\sqrt{2M}}{\sigma^2_{\min,2M}(\mathbf{\Phi})}$$

which is sharp for $q = 2$ when $\mathbf{\Phi}$ is a basis (we leave the proof of sharpness to the reader). The estimate is proved as follows: for any δ with $\|\delta\|_0^0 \leq 2M$, we have

$$\sigma^2_{\min,2M} \cdot \|\delta\|_2^2 \leq \|\mathbf{\Phi}\delta\|_{\mathcal{H}}^2 = \langle \delta, \mathbf{\Phi}^\star \mathbf{\Phi}\delta \rangle \leq \|\delta\|_1 \cdot d(\mathbf{\Phi}\delta),$$

hence

$$\|\delta\|_2 \leq \frac{d(\mathbf{\Phi}\delta)}{\sigma^2_{\min,2M}} \cdot \frac{\|\delta\|_1}{\|\delta\|_2} \leq \frac{d(\mathbf{\Phi}\delta)}{\sigma^2_{\min,2M}} \cdot \sqrt{2M}.$$

As a result, if $\mathbf{y}$ has a representation $\mathbf{y} = \mathbf{\Phi}x + \mathbf{e}$ with $M := \|x\|_0^0 < (1 + 1/\mu_1(1))/2$, and $\sup_k |\langle \mathbf{e}, \mathbf{g}_k \rangle| = \epsilon$, then the solution x_0 of the optimization problem

$$\min_{x'} \|x'\|_0^0 \text{ subject to } \sup_k |\langle \mathbf{y} - \mathbf{\Phi}x', \mathbf{g}_k \rangle| \leq \epsilon \tag{42}$$

satisfies

$$\|x_0 - x\|_2^2 \leq \frac{8M\epsilon^2}{1 - \mu_1(2M-1)}. \tag{43}$$

6. PROOF OF THE MAIN RESULTS

When the distortion $d(\cdot)$ is the MSE, the worst case analysis carried out with general sub-additive measures can be drastically improved. The key observation is that $\delta \in D(\mathbf{e})$ if, and only if, $|\langle \mathbf{\Phi}^\star \mathbf{e}, \delta \rangle| \geq \|\mathbf{\Phi}\delta\|_{\mathcal{H}}^2/2$. This allows us to recast a few problems to a ℓ^2 sparse approximation problem in the canonical basis, for which we can use the results of Section 4. As a result, the obtained bounds will be close to sharp, as discussed below in Sections 6.4-6.5.

6.1. Proof of Theorem 1.

In all the following we assume that $\delta \in D(\mathbf{e}) \cap C_{f_0}(\mathrm{X}_M)$, that is to say $\|\delta\|_0^0 \leq 2M$ and $|\langle \mathbf{\Phi}^\star \mathbf{e}, \delta \rangle| \geq \|\mathbf{\Phi}\delta\|_{\mathcal{H}}^2/2$. Denoting $\delta' = \sigma^2_{\min,2M}(\mathbf{\Phi}) \cdot \delta$, we have $\delta' \in C_{f_0}(\mathrm{X}_M)$. We will check that $\|b - \mathbf{B}\delta'\|_2 \leq \|b\|_2$ with $\mathbf{B}$ the canonical basis in ℓ^2 and b the restriction of $\mathbf{\Phi}^\star \mathbf{e}$ to the finite support of δ', which is of size at most $2M$. Thus, Lemma 2 for $p = 2$ and $K = 2M$ will tell us exactly that

$$\sigma^2_{\min,2M}(\mathbf{\Phi}) \cdot \|\delta\|_\infty = \|\delta'\|_\infty \leq |\mathbf{e}|_1 + |\mathbf{e}|_{2M}$$

and we will get the first inequality (5).

Indeed, using the assumptions on δ and the definition of $\sigma^2_{\min,2M}(\mathbf{\Phi})$ we have

$$\begin{aligned} \frac{|\langle b, \delta' \rangle|}{\|\delta'\|_2^2} &= \frac{|\langle b, \delta \rangle|}{\sigma^2_{\min,2M}(\mathbf{\Phi}) \cdot \|\delta\|_2^2} \\ &= \frac{|\langle \mathbf{\Phi}^\star \mathbf{e}, \delta \rangle|}{\|\mathbf{\Phi}\delta\|_{\mathcal{H}}^2} \cdot \frac{\|\mathbf{\Phi}\delta\|_{\mathcal{H}}^2}{\sigma^2_{\min,2M}(\mathbf{\Phi}) \cdot \|\delta\|_2^2} \geq \frac{1}{2}. \end{aligned}$$

It follows that $\|b - \mathbf{B}\delta'\|_2 \leq \|b\|_2$ as claimed. To get the second inequality (6), simply notice that

$$\|\delta\|_2^2 \leq \frac{2 \cdot |\langle b, \delta \rangle|}{\sigma^2_{\min,2M}(\mathbf{\Phi})} \leq \frac{2 \cdot |\mathbf{e}|_{2M} \cdot \|\delta\|_2}{\sigma^2_{\min,2M}(\mathbf{\Phi})}$$

where we used the Cauchy-Schwarz inequality and the fact that $\|b\|_2 = |\mathbf{e}|_{2M}$. To get the third inequality (7) we observe that

$$\|\mathbf{\Phi}\delta\|_{\mathcal{H}}^2 \leq 2 \cdot |\langle \mathbf{\Phi}^\star \mathbf{e}, \delta \rangle| \leq 2 \cdot |\mathbf{e}|_{2M} \cdot \|\delta\|_2 \leq \frac{4 \cdot |\mathbf{e}|_{2M}^2}{\sigma^2_{\min,2M}(\mathbf{\Phi})}$$

where we used (6) for the right hand side inequality.

6.2. Proof of Theorem 2.

In all the following we assume that $\delta \in D(\mathbf{e}) \cap C_{f_1}(\mathrm{X}_M)$. Let $(|\langle \mathbf{e}, \mathbf{g}_{k_m}\rangle|)_{m\geq 1}$ and $(|\delta_{l_m}|)_{m\geq 1}$ be decreasing rearrangements of $(|\langle \mathbf{e}, \mathbf{g}_k\rangle|)_k$ and $(|\delta_k|)_k$. We let $c := \sigma^2_{\min,M}(\mathbf{\Phi}) \cdot \lambda^2_M(\mathbf{\Phi})$ and we define $\delta' = (\delta'_m)$ by $\delta'_m = c \cdot |\delta_{l_m}|$ for $1 \leq m \leq M$ and $\delta'_m = 0$ for $m > M$. Similarly, we let $b = (b_m)$ with $b_m = 2 \cdot |\langle \mathbf{e}, \mathbf{g}_{k_m}\rangle|$ for $1 \leq m \leq M$ and $b_m = 0$ for $m > M$.

Since $\|\delta'\|_0^0 \leq M$ and $|\langle \mathbf{\Phi}^\star \mathbf{e}, \delta\rangle|/\|\mathbf{\Phi}\delta\|_2^2 \geq 1/2$, if we can prove that

$$\frac{|\langle b, \delta'\rangle|}{\|\delta'\|_2^2} \geq \frac{|\langle \mathbf{\Phi}^\star \mathbf{e}, \delta\rangle|}{\|\mathbf{\Phi}\delta\|_{\mathcal{H}}^2} \tag{44}$$

then we will get that $\|b - \mathbf{B}\delta'\|_2 \leq \|b\|_2$, Lemma 2 with $p = 2$ for $K = M$ will tell us exactly that

$$c \cdot \|\delta\|_\infty = \|\delta'\|_\infty \leq 2 \cdot (|\mathbf{e}|_1 + |\mathbf{e}|_M)$$

and we will get the inequality (9).

We will prove (44) by getting an upper bound on the numerator and a lower bound on the denominator of the right hand side. For the numerator, we have

$$\begin{aligned}
|\langle \mathbf{\Phi}^\star \mathbf{e}, \delta\rangle| &\leq \sum_{m=1}^{M} |\langle \mathbf{e}, \mathbf{g}_{k_m}\rangle| \cdot |\delta_{l_m}| \\
&\quad + \sum_{m\geq M+1} |\langle \mathbf{e}, \mathbf{g}_{k_m}\rangle| \cdot |\delta_{l_m}| \\
&\leq \sum_{m=1}^{M} |\langle \mathbf{e}, \mathbf{g}_{k_m}\rangle| \cdot |\delta_{l_m}| \\
&\quad + |\langle \mathbf{e}, \mathbf{g}_{k_{M+1}}\rangle| \cdot \sum_{m\geq M+1} |\delta_{l_m}| \\
&\leq \sum_{m=1}^{M} |\langle \mathbf{e}, \mathbf{g}_{k_m}\rangle| \cdot |\delta_{l_m}| \\
&\quad + |\langle \mathbf{e}, \mathbf{g}_{k_{M+1}}\rangle| \cdot \sum_{m=1}^{M} |\delta_{l_m}| \\
&= \sum_{m=1}^{M} \left(|\langle \mathbf{e}, \mathbf{g}_{k_m}\rangle| + |\langle \mathbf{e}, \mathbf{g}_{k_{M+1}}\rangle|\right) \cdot |\delta_{l_m}| \\
&\leq |\langle b, \delta'/c\rangle|
\end{aligned} \tag{45}$$

For the denominator we proceed as follows. Let $I = I_M(\delta)$ and consider δ_I the sequence with zeros everywhere except on I where it coincides with δ. Since $\mathbf{\Phi}_I \mathbf{\Phi}_I^+$ is an orthonormal projector and I is of cardinal M, we have $\|\mathbf{\Phi}\delta\|_{\mathcal{H}}^2 \geq \|\mathbf{\Phi}_I \mathbf{\Phi}_I^+ \mathbf{\Phi}\delta\|_{\mathcal{H}}^2 = \|\mathbf{\Phi}_I z\|_{\mathcal{H}}^2 \geq \sigma^2_{\min,M}(\mathbf{\Phi}) \cdot \|z\|_2^2$ with $z := \mathbf{\Phi}_I^+ \mathbf{\Phi}\delta = \delta_I + \sum_{k\notin I} \delta_k \cdot \mathbf{\Phi}_I^+ \mathbf{g}_k$. Now, we have

$$\begin{aligned}
\|z\|_2 &\geq \|\delta_I\|_2 - \sum_{k\notin I} |\delta_k| \cdot \sup_{k\notin I} \|\mathbf{\Phi}_I^+ \mathbf{g}_k\|_2 \\
&\geq \|\delta_I\|_2 - \|\delta_I\|_1 \cdot \sup_{k\notin I} \|\mathbf{\Phi}_I^+ \mathbf{g}_k\|_2 \\
&\geq \|\delta_I\|_2 - \sqrt{M}\|\delta_I\|_2 \cdot \sup_{k\notin I} \|\mathbf{\Phi}_I^+ \mathbf{g}_k\|_2 \\
&\geq \lambda_M(\mathbf{\Phi}) \cdot \|\delta_I\|_2
\end{aligned}$$

where we used the fact that $\delta \in C_{f_1}(\mathrm{X}_M)$ to get the second inequality. If $\lambda_M(\mathbf{\Phi}) > 0$, we obtain

$$||\mathbf{\Phi}\delta||_{\mathcal{H}}^2 \geq \sigma_{\min,M}^2(\mathbf{\Phi}) \cdot \lambda_M^2(\mathbf{\Phi}) \cdot ||\delta_I||_2^2 = ||\delta'||_2^2/c. \tag{46}$$

Combining (45) and (46) yields the desired inequality

$$\frac{|\langle \mathbf{\Phi}^\star \mathbf{e}, \delta\rangle|}{||\mathbf{\Phi}\delta||_{\mathcal{H}}^2} \leq \frac{|\langle b, \delta'/c\rangle|}{||\delta'||_2^2/c} = \frac{|\langle b, \delta'\rangle|}{||\delta'||_2^2}$$

To get the third inequality (11), we write

$$\begin{aligned} ||\mathbf{\Phi}\delta||_{\mathcal{H}}^2 &\leq 2 \cdot |\langle \mathbf{\Phi}^\star \mathbf{e}, \delta\rangle| \leq 2 \cdot |\langle b, \delta'/c\rangle| \\ &\leq 4 \cdot |\mathbf{e}|_M \cdot ||\delta_I||_2 \\ &\leq 4 \cdot |\mathbf{e}|_M \cdot ||\mathbf{\Phi}\delta||_{\mathcal{H}}/\sqrt{c} \end{aligned}$$

To get the second inequality (10), noticing that (46) and (11) yield

$$||\delta_I||_2^2 \leq ||\mathbf{\Phi}\delta||_{\mathcal{H}}^2/c \leq 4 \cdot |\mathbf{e}|_M \cdot ||\delta_I||_2/c,$$

we obtain $||\delta_I||_2 \leq 4 \cdot |\mathbf{e}|_M/c$, and we conclude using the fact that

$$\begin{aligned} ||\delta||_2^2 &= ||\delta_I||_2^2 + \sum_{k \notin I} |\delta_k|^2 \leq ||\delta_I||_2^2 + \left(\sum_{k \notin I} |\delta_k|\right)^2 \\ &\leq ||\delta_I||_2^2 + ||\delta_I||_1^2 \leq (1+M) \cdot ||\delta_I||_2^2. \end{aligned}$$

□

Notice that in the proof, instead of $b_m = 2 \cdot |\langle \mathbf{e}, \mathbf{g}_{k_m}\rangle|$, we could have used $b'_m = \left(|\langle \mathbf{e}, \mathbf{g}_{k_m}\rangle| + |\langle \mathbf{e}, \mathbf{g}_{k_{M+1}}\rangle|\right)/2$. This would have lead to slightly sharper estimates, however they would have been more cumbersome to express and, as discussed below in Sections 6.4-6.5, the simplified estimate with $b_m = 2 \cdot |\langle \mathbf{e}, \mathbf{g}_{k_m}\rangle|$ is almost sharp.

6.3. Proof of Proposition 1

Proposition 1 immediately follows from the lemmas below.

Lemma 4 *Let $\mathbf{\Phi}$ be a normalized dictionary in a Hilbert space $\mathcal{H}$. For every integer K we have*

$$\sigma_{\min,K}^2(\mathbf{\Phi}) \geq 1 - \mu_1(K-1, \mathbf{\Phi}). \tag{47}$$

The proof is based on Geršgorin Disc Theorem and can be found in [9].

Lemma 5 *Let $\mathbf{\Phi}$ be a normalized dictionary in a Hilbert space $\mathcal{H}$. We have*

$$\lambda_M(\mathbf{\Phi}) \geq 1 - \frac{\sqrt{M} \cdot \mu_2(M, \mathbf{\Phi})}{1 - \mu_1(M-1, \mathbf{\Phi})}. \tag{48}$$

Proof. For any index set I with $\mathrm{card}(I) \leq M$ we have

$$|||(\mathbf{\Phi}_I^\star \mathbf{\Phi}_I)^{-1}|||_{2,2} \leq \frac{1}{\sigma_{\min,M}^2(\mathbf{\Phi})} \leq \frac{1}{1 - \mu_1(M-1)}.$$

Since $\mathbf{\Phi}_I^+ = (\mathbf{\Phi}_I^\star \mathbf{\Phi}_I)^{-1}\mathbf{\Phi}_I^\star$ and, for $k \notin I$, we have

$$||\mathbf{\Phi}_I^\star \mathbf{g}_k||_2 = \sqrt{\sum_{i\in I} |\langle \mathbf{g}_k, \mathbf{g}_i\rangle|^2} \leq \mu_2(M),$$

we get

$$||\mathbf{\Phi}_I^+ \mathbf{g}_k||_2 = ||(\mathbf{\Phi}_I^\star \mathbf{\Phi}_I)^{-1}\mathbf{\Phi}_I^\star \mathbf{g}_k||_2 \leq \frac{\mu_2(M)}{1-\mu_1(M-1)}.$$

We conclude using the definition of $\lambda_M(\mathbf{\Phi})$ (see Eq. (8)). □

6.4. Sharpness of Theorem 1

When $\mathbf{\Phi}$ is an orthonormal basis, the estimate (5) in Theorem 1 is sharp since $\sigma^2_{\min,2M} = 1$ and we have an exact estimate given by Lemma 2. Similarly, one can check the sharpness of the estimates (6)-(7). For a general $\mathbf{\Phi}$, a slight modification of the proof of Theorem 1 leads to

$$\sup_{\delta \in D(\mathbf{e})\cap C_{f_0}(\mathrm{X}_M)} ||\delta||_\infty \geq \frac{|\mathbf{e}|_1 + |\mathbf{e}|_{2M}}{\sigma^2_{\max,2M}} \tag{49}$$

with

$$\sigma^2_{\max,K}(\mathbf{\Phi}) := \sup_{||\delta||_0^0 \leq K} \frac{||\mathbf{\Phi}\delta||^2_{\mathcal{H}}}{||\delta||_2^2} \leq 1 + \mu_1(K-1).$$

Thus, if $\mu_1(2M-1)$ is small enough, the estimate (5) is almost sharp in the sense that it cannot be significantly improved.

6.5. Sharpness of Theorem 2

Since

$$\sup_{\delta \in D(\mathbf{e})\cap C_{f_0}(\mathrm{X}_M)} ||\delta||_\infty \leq \sup_{\delta \in D(\mathbf{e})\cap C_{f_1}(\mathrm{X}_M)} ||\delta||_\infty,$$

the (almost) sharpness of the results in Theorem 2 is a consequence of that of Theorem 1: combining Eq. (49) with the estimate (9) we have

$$\frac{|\mathbf{e}|_1 + |\mathbf{e}|_{2M}}{\sigma^2_{\max,2M}} \leq \sup_{\delta \in D(\mathbf{e})\cap C_{f_1}(\mathrm{X}_M)} ||\delta||_\infty \leq \frac{2\cdot(|\mathbf{e}|_1 + |\mathbf{e}|_M)}{\sigma^2_{\min,M}\cdot\lambda^2_M}.$$

When $\mu_1(2M-1)$ is small enough, the upper bounds become approximately $|\mathbf{e}|_1 + |\mathbf{e}|_{2M}$ and $2\cdot(|\mathbf{e}|_1 + |\mathbf{e}|_M)$ which differ at most by a factor two since $|\mathbf{e}|^2_M \leq |\mathbf{e}|^2_{2M} \leq 2|\mathbf{e}|^2_M$.

7. DISCUSSION AND CONCLUSION

We provided tools to check if a given sparse approximation of an input signal –which may have been computed using any algorithm– is nearly optimal, in the sense that no other significantly different representation can at the same time be as sparse and provide as good an approximation. In particular we proposed a test to check if the atoms used in a sparse approximation are "the good ones" corresponding to the ideal sparse approximation for a fairly large class of admissible sparsity measures. The test is easy to implement, it does not depend on which algorithm was used to obtain the decomposition and does not rely on any prior knowledge on the ideal sparse approximation.

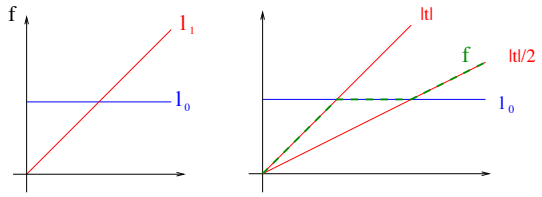

Fig. 1. (a) The "nice" functions $f_0(t)$ and $f_1(t)$ corresponding to the ℓ^0 and the ℓ^1 norm (b) An "exotic" sparsity measure, based on a mix of the ℓ^0 and ℓ^1 measures plotted in (a).

Eventually, we provided extended results of the same flavor including the case of some non quadratic distortion measures, and we discussed some implications of our results in terms of Bayesian estimation and signal denoising with a fairly large class of sparse priors and random noise.

We are currently trying to investigate how this work could also be extended to obtain results on the optimality of simultaneous sparse approximation of several signals, in order to apply the results to blind source separation. In addition, we are investigating the use of the optimality tests to design provably good sparse approximation algorithms.

A. APPENDIX

A.1. Sparsity measures

For the sake of completeness we include some properties of sub-additive (resp. admissible) sparsity measures.

Lemma 6 ([14, Prop. 1]) *Every admissible sparsity measure is sub-additive, but the reciprocal is false.*

Proof. For $t, u \geq 0$, since $f(t+u)/(t+u) \leq f(u)/u$ and $f(t+u)/(t+u) \leq f(t)/t$, we have

$$f(t) + f(u) \geq f(t+u) \cdot \frac{t}{t+u} + f(t+u) \cdot \frac{u}{t+u} = f(t+u).$$

Now, denoting $\lfloor t \rfloor$ the largest integer such that $\lfloor t \rfloor < t \leq \lfloor t \rfloor$, and f the function such that $f(0) = 0$ and $f(t) = 1 + \lfloor t \rfloor$ for $t > 0$, we let the reader check that f is sub-additive but not admissible.

Lemma 7 *If f, g are sub-additive sparsity measures, then $f \circ g$ is also a sub-additive sparsity measure.*

Proof. First, since $f(0) = g(0) = 0$ and f, g are non decreasing, $f \circ g$ has the same properties. Now, for any $t, u \geq 0$ we have

$$f\big[g(t+u)\big] \leq f\big[g(t) + g(u)\big] \leq f\big[g(t)\big] + f\big[g(u)\big]$$

Figure 1 illustrates the fact that admissible sparsity measures are not necessarily as "nice" as one could imagine from their most natural examples the ℓ^τ measures. Using the fact that the $\min$ or the $\max$ of two admissible sparsity measures yields another admissible sparsity measure [14], we can combine the ℓ^0 and the ℓ^1 norms into the "exotic" measure $f(t) := \min(f_1(t), \max(f_0(t), f_1(t)/2))$.

A.2. Proof of Lemma 1

For any index set I, we let $\mathrm{X}(I) = \{x : \mathrm{support}(x) = I\}$. Since $\mathrm{X}_M = \bigcup_{\mathrm{card}(I)=M} \mathrm{X}(I)$ we have

$$C_f(\mathrm{X}_M) = \bigcup_{\mathrm{card}(I) \leq M} C_f(\mathrm{X}(I)). \tag{50}$$

By definition, $\delta \in C_f(\mathrm{X}(I))$ if, and only if, for all $z \in \mathrm{X}(I)$

$$\begin{aligned}\sum_{k\in I} f(|z_k|) &= \|z\|_f \geq \|z+\delta\|_f \\ &= \sum_{k\in I} f(|z_k+\delta_k|) + \sum_{k\notin I} f(|\delta_k|),\end{aligned}$$

or in other words if, and only if, for all $z \in \mathrm{X}(I)$

$$\sum_{k\in I} [f(|\delta_k|) + f(|z_k|) - f(|z_k+\delta_k|)] \geq \|\delta\|_f. \tag{51}$$

Letting $G_f(t) := \sup_{u\in\mathbb{R}} f(|u|) - f(|u+t|)$ we have

- if $\sum_{k\in I} [f(|\delta_k|) + G_f(\delta_k)] > \|\delta\|_f$ then $\delta \in C_f(\mathrm{X}(I))$;
- if $\sum_{k\in I} [f(|\delta_k|) + G_f(\delta_k)] < \|\delta\|_f$ then $\delta \notin C_f(\mathrm{X}(I))$.

and the case $\sum_{k\in I} [f(|\delta_k|) + G_f(\delta_k)] = \|\delta\|_f$ depends whether the supremum is achieved in the definition of G_f. For any f one can see that $G_f(t) \geq f(|t|)$ by letting $u = -t$ in the definition of G_f. When f is non-decreasing and sub-additive, we have for all u and t

$$\begin{aligned}f(|u|) - f(|u+t|) &= f(|u+t-t|) - f(|u+t|) \\ &\leq f(|u+t| + |t|) - f(|u+t|) \\ &\leq f(|u+t|) + f(|t|) - f(|u+t|) \\ &= f(|t|)\end{aligned}$$

which shows that we have indeed $G_f(t) = f(|t|)$, and the supremum is achieved in the definition of G_f. Thus, if f is non-decreasing and sub-additive, we have $\delta \in C_f(\mathrm{X}(I))$ if, and only if, $\sum_{k\in I} f(|\delta_k|) \geq \|\delta\|_f/2$. From the characterization (50) we conclude that $\delta \in C_f(\mathrm{X}_M)$ if and only if

$$\sum_{k\in I_M(\delta)} f(|\delta|) \geq \frac{\|\delta\|_f}{2}$$

where $I_M(\delta)$ indexes the M largest components $|\delta_k|$.

B. REFERENCES

[1] E. Le Pennec and S. Mallat, "Sparse geometrical image approximation with bandelets," *IEEE Transaction on Image Processing*, vol. (to appear), 2005.

[2] M. Zibulevsky and B.A. Pearlmutter, "Blind source separation by sparse decomposition in a signal dictionary," *Neural Computations*, vol. 13, no. 4, pp. 863–882, 2001.

[3] J.-L. Starck, E. J. Candès, and D. L. Donoho, "The curvelet transform for image denoising," *IEEE Transactions on Image Processing*, vol. 2, no. 6, pp. 670–684, 2002.

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

[5] Y.C. Pati, R. Rezaiifar, and P.S. Krishnaprasad, "Orthonormal matching pursuit : recursive function approximation with applications to wavelet decomposition," in *Proceedings of the* 27^{th} *Annual Asilomar Conf. on Signals, Systems and Computers*, Nov. 1993.

[6] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing*, vol. 20, no. 1, pp. 30–61, 1998.

[7] I. F. Gorodnitsky and B. D. Rao, "Sparse signal reconstruction from limited data using FOCUSS : a reweighted norm minimization algorithm," *IEEE Trans. Signal Proc.*, vol. 45, no. 3, pp. 600–616, mar 1997.

[8] D. Donoho, M. Elad, and V. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," Tech. Rep. 2004-01, Department of Statistics, Stanford University, 2004, submitted February 2004.

[9] Joel A. Tropp, "Just relax: Convex programming methods for subset selectin and sparse approximation," Tech. Rep. 04-04, Institute for Computational Engineering and Sciences (ICES), The University of Texas at Austin, Feb. 2004.

[10] J.-J. Fuchs, "Recovery of exact sparse representations in the presence of bounded noise," Tech. Rep. 1618, IRISA, Rennes, France, Apr. 2004, submitted.

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

[12] R. Gribonval and P. Vandergheynst, "On the exponential convergence of Matching Pursuit in quasi-incoherent dictionaries," Tech. Rep., IRISA, 2004.

[13] Brendt Wohlberg, "Noise sensitivity of sparse signal representations: Reconstruction error bounds for the inverse problem," *IEEE Transactions on Signal Processing*, vol. 51, no. 12, pp. 3053–3060, Dec. 2003.

[14] R. Gribonval and M. Nielsen, "Highly sparse representations from dictionaries are unique and independent of the sparseness measure," Tech. Rep. R-2003-16, Dept of Math. Sciences, Aalborg University, Oct. 2003.