Matching Pursuit with Block Incoherent Dictionaries

Lorenzo Peotta and Pierre Vandergheynst
Signal Processing Institute
École Polytechnique Féderal de Lausanne
EPFL-STI-ITS-LTS2, ELD 241, Station 11, 1015 Lausanne, Switzerland
phone: +41 21 6935651, fax: +41 21 6937600
{lorenzo.peotta,pierre.vandergheynst}@epfl.ch

***Abstract*— There has been an intense activity recently in the field of sparse approximations with redundant dictionaries, largely motivated by the practical performances of algorithms such as Matching Pursuit and Basis Pursuit. However, most of the theoretical results obtained so far are valid only for the restricted class of incoherent dictionaries. This paper investigates a new class of overcomplete dictionaries, called block incoherent dictionaries, where coherence can be arbitrarily big. We show that a simple greedy algorithm can correctly identify stable subdictionaries (called blocks) and demonstrate how one can use the extra coherence freedom for approximation purposes.**

I. Introduction

The problem of building good sparse approximations of signals or functions has received tremendous attention recently. From a practical point of view, sparse expansions allow to replace complicated signals by few elementary building blocks that essentially synthesize all the information at hand. The very strong links between approximation theory and computational harmonic analysis on one hand and data processing on the other hand, resulted in fruitful cross-fertilizations over the last decade, from fundamental results (near optimal rate of non-linear approximations for wavelets and other basis [1]) to practical ones (like the JPEG2000 image compression standard).

Natural signals however do not generally lend themselves to simple models, for which orthonormal basis are generally near optimal. Images for example do contain smooth parts and regular contours that could be efficiently represented by a curvelet tight frame [2], but they also contain various kind of irregular edges together with a plethora of textures. Audio signals contain sharp transients and smooth parts that are suitable for wavelet basis, but they also contain stationary oscillatory parts that are better suited for local trigonometric basis [3]. Bearing in mind the multiple components of natural data, one is tempted to approximate them with mixtures of basis functions. Approximating data with general dictionaries seemed a daunting task, and raised many questions concerning the unicity and optimality of sparse representations or approximations. There has been recently an intense activity in this field, showing that constructive results can be obtained on all fronts. The possibility of recovering optimal sparse representations using Basis Pursuit (BP) opened the way [4]–[7]. When an exact sparse representation is not needed, approximation results become more useful, and recent results have shown that variations around greedy algorithms such as Matching Pursuit (MP) and Orthogonal Matching Pursuit (OMP) are promising [8], [9].

One of the key properties in the above-mentioned results lies in the characteristics of the dictionary, and one could roughly say that in most cases the latter is required to be sufficiently incoherent, i.e *close enough* to an orthogonal basis. Putting strong restrictions on the dictionary though may damage the original goal in the sense that we loose flexibility in designing it. Our main contribution in this paper is to relax some of these strong hypotheses by allowing more redundancy in the dictionaries. We introduce the concept of block incoherence, which basically describes a dictionary that can be represented as an "incoherent" union of coherent blocks. Each of these blocks could model particular characteristics of input signals taking advantage of the high redundancy of each block. We show that even pure greedy algorithms can strongly benefit from such design by proving a recovery condition under which Matching Pursuit will always pick up atoms from correct blocks during the signal expansion.

The outline of our paper is as follows. After some basic definitions in Section II, we provide exact recovery and rate of approximation results in Sections III and IV respectively. In Section VI we explore the links between block incoherent dictionaries and Grassmanian packings. We modify an algorithm proposed by Tropp [10] and use it to numerically construct block incoherent dictionaries. We also provide simulations aimed at illustrating the performances of these dictionaries. We conclude by discussing potential applications and future work.

II. Block Incoherent Dictionaries

In this paper, we will exclusively deal with finite dimensional signals modeled as d-dimensional real or complex vectors. We will call dictionary a large collection $\mathcal{D}$ of vectors $g_k \in \mathbb{R}^d$ (or $\mathbb{C}^d$), suitably normalized such that $\|g_k\|_2 = 1$. Equivalently, we will sometimes arrange these vectors as the columns of a large $d \times L$ matrix. The cardinality L of the dictionary is usually very large, $L \gg d$, and that is what we mean by $\mathcal{D}$ being redundant. We also assume, unless otherwise stated, that the dictionary is complete, i.e it spans $\mathbb{R}^d$ (or $\mathbb{C}^d$).

Given a redundant dictionary $\mathcal{D}$, we consider the following P-subset decomposition $\mathcal{D} = \bigcup_{k=1}^{P} B_k$ with $B_i \bigcap B_j = \emptyset$ for $i \neq j$, and we call *blocks* the P subsets of atoms B_k, $n = 1, \cdots, P$. The *block coherence* is defined as the maximum coherence between any two atoms, taken from different blocks.

Definition 1: The block coherence μ_B, given a block decomposition $\mathcal{D} = \bigcup_{k=1}^{P} B_k$, is

$$\mu_B \triangleq \max_{i \neq j} \max_{k,l} | < g_k^i, g_l^j > |, \tag{1}$$

where g_k^i is the k^{th} atom from the block B_i.

Definition 2: A dictionary is called *block incoherent* if there exists a decomposition such that the *block coherence* μ_B is small.

The block coherence considers similarities between atoms from two different blocks. In order to refine the analysis of the coherence, we introduce another function, called the *cumulative block coherence*, that represents the coherence between sets of m blocks $B_I = \bigcup_{i \in I} B_i$, with $|I| = m$, the cardinality of the set I.

Definition 3: Let $\mathcal{D} = \bigcup_{k=1}^{P} B_k$ denote a decomposition as above, and $B_I = \bigcup_{i \in I} B_i$ represent a set of m blocks. The *cumulative block coherence* is

$$\mu_{1_B}(m) \triangleq \max_{I,s.t.|I|=m} \max_{j \notin I, l} \sum_{i \in I} \max_{k_i} | < g_{k_i}^i, g_l^j > |. \tag{2}$$

Definition 4: A given dictionary $\mathcal{D}$ is said to be *block quasi-incoherent*, if we can find a block decomposition such that $\mu_{1_B}(m)$ grows slowly with m.

The *block coherence* μ_B considers coherence between two blocks, and the *cumulative block coherence* $\mu_{1_B}(m)$ measures coherence between m blocks. Notice that the *cumulative block coherence* is bounded by the $block\ coherence$: $\mu_{1_B}(m) \leq m\mu_B$. These definitions are straightforward extensions of the coherence μ and the *cumulative coherence* $\mu_1(m)$ introduced in [5], [6] and [9]. We need now also to consider the coherence within a single block. Generally, a single block B_i has a strong coherence (i.e., the *cumulative coherence* of that block grows quickly). For a more detailed analysis, we are however interested in a measure that represents the coherence of a particular subset of functions in B_i, which leads us to the following concept.

Definition 5: The $disparity$ of a block B_i is

$$\xi(B_i) \triangleq \min_{\iota,s.t.|B_\iota|=rank(B_i)} \max_{k} \sum_{l \neq k} | < g_l^\iota, g_k^\iota > |, \tag{3}$$

where B_ι is a set of linearly independent atoms from B_i such that $span(B_\iota) = span(B_i)$.

The disparity $\xi(B_i)$ indicates how coherent could be a basis of $span(B_i)$ constructed with the least number of atoms from B_i. The set of atoms, i.e. B_ι, where the disparity is minimal is called $B_{i^\star}$. If $\xi(B_i) = 0$, we can find a set $B_{i^\star} \subset B_i$ that is an orthogonal basis for $span(B_i)$. The extension of the $disparity$ to the dictionary $\mathcal{D}$ is simply defined as $\xi(\mathcal{D}) = \max_i \xi(B_i)$.

III. Exact Block Selection

Using the definitions of Sec. II, we now prove in this section that, given a block incoherent dictionary $\mathcal{D}$ and a signal f, the Matching Pursuit (MP) algorithm can recover a block-sparse representation of f.

Matching Pursuit (or MP for short) [11] is a greedy algorithm that iteratively decomposes a given signal over a dictionary of atoms. At the first step the atom g_0 most correlated with the signal is selected. A residual r_1 is constructed by removing this contribution from the signal : $r_1 = f - \langle f, g_0 \rangle g_0$. The whole process is then iterated on this residual. After N steps we get the following decomposition of the original signal :

$$f = \sum_{n=0}^{N-1} \langle r_n, g_n \rangle \, g_n + r_N$$

with $r_0 = f$.

We will now consider the restricted problem $(\mathcal{D}, B_I)$-*SPARSE*, where f is exactly represented as a linear combination of atoms belonging to a subset of m blocks $B_I = \bigcup_{i \in I} B_i$, $|I| = m$. First, we find a single sufficient condition under which MP recovers atoms from a fixed set of incoherent blocks B_I. In this case, we say that MP chooses atoms from *correct* blocks B_i, $i \in I$. In the following theorems it will be useful to identify B_I with the matrix whose columns list all the atoms of the set B_I and we denote B_I^+ its pseudoinverse.

Theorem 1: Let $\mathcal{D}$ be a block incoherent dictionary and $B_I = \bigcup_{i \in I} B_i$. If the signal $f \in \mathcal{V}_I = span(B_I)$, then under the recovery condition

$$\eta(B_I) \triangleq \max_{g \notin B_I} \|B_I^+ g\|_1 < 1 \tag{4}$$

we have that MP:

1) picks up atoms only from correct blocks B_i $i \in I$,
2) converges exponentially to f.

The proof of this theorem follows directly by mimicking Tropp's original proof for incoherent dictionaries and writing everything in terms of the block synthesis matrix B_I, see for example [9], [12]. This result might not look directly useful since the recovery condition is related to a particular set of blocks. The condition on B_I is not very explicit either. The following theorem shows that correct block selection holds whenever f belongs to the $span$ of an arbitrary set of m *sufficiently incoherent* blocks.

Theorem 2: Let $\mathcal{D}$ a block incoherent dictionary and B_I an arbitrary set of m blocks and $K = \max_i \ rank(B_i)$. If the signal $f \in \mathcal{V}_I$ and

$$K \, \mu_{1_B}(m) + \xi(\mathcal{D}) + K \, \mu_{1_B}(m-1) < 1 \tag{5}$$

then we have that MP:

1) picks up atoms only from the correct blocks,
2) converges exponentially to f.

Proof of Theorem 2. Suppose that at step $n-1$ the residual generated by the Matching Pursuit algorithm $r_{n-1} \in \mathcal{V}_I$. If an atom g_{n-1} from B_I is selected by MP, then also $r_n = r_{n-1} - \langle g_{n-1}, r_{n-1} \rangle \, g_{n-1}$ belongs to $\mathcal{V}_I$. The vector $B_I^T r_{n-1}$ lists the inner products between the residual r_{n-1} and all the atoms from the blocks B_i, $i \in I$; taking the ∞ norm of this vector we have that $\|B_I^T r_{n-1}\|_\infty$ is the largest of these inner products in magnitude, where B_I^T represents the complex conjugate of B_I. The number $\|B_I^T r_{n-1}\|_\infty$ corresponds to the largest inner

product in magnitude between r_{n-1} and an atom that does not belong to B_I. An atom is selected from the correct block B_i, $i \in I$, when the following quotient is less than one

$$\rho(r_{n-1}) \triangleq \frac{\|B_{\bar{I}}^T r_{n-1}\|_\infty}{\|B_I^T r_{n-1}\|_\infty} < 1. \quad (6)$$

We indicate with $B'_I = \bigcup_{i \in I} B_{i^\star}$ the union of the m sets associated to the m blocks B_i in Definition 5. Now we define $B_{I^\star}$ to be a set of linearly independent atoms from B'_I such that $|B_{I^\star}| = rank(B'_I)$. It follows that $span(B_{I^\star}) = span(B_I) = \mathcal{V}_I$ and $B_{I^\star}$ is a basis for $\mathcal{V}_I$. Therefore $r_{n-1} = (B_{I^\star}^+)^T B_{I^\star}^T r_{n-1}$ and

$$\begin{aligned} \rho(r_{n-1}) &= \frac{\|B_{\bar{I}}^T r_{n-1}\|_\infty}{\|B_I^T r_{n-1}\|_\infty} \\ &= \frac{\|B_{\bar{I}}^T (B_{I^\star}^+)^T B_{I^\star}^T r_{n-1}\|_\infty}{\|B_I^T r_{n-1}\|_\infty}. \end{aligned}$$

Since $B_{I^\star} \subset B_I$ we have that $\|B_I^T r_{n-1}\|_\infty \geq \|B_{I^\star}^T r_{n-1}\|_\infty$ and

$$\begin{aligned} \rho(r_{n-1}) &\leq \frac{\|B_{\bar{I}}^T (B_{I^\star}^+)^T B_{I^\star}^T r_{n-1}\|_\infty}{\|B_{I^\star}^T r_{n-1}\|_\infty} \\ &\leq \|B_{\bar{I}}^T (B_{I^\star}^+)^T\|_{\infty,\infty} \\ &= \max_{g \in B_{\bar{I}}} \|B_{I^\star}^+ g\|_1. \end{aligned}$$

Now we can expand the pseudoinverse and apply the norm bound $\|Ax\|_1 \leq \|A\|_{1,1}\|x\|_1$

$$\begin{aligned} \rho(r_{n-1}) &\leq \max_{g \in B_{\bar{I}}} \|(B_{I^\star}^T B_{I^\star})^{-1} B_{I^\star}^T g\|_1 \\ &\leq \|(B_{I^\star}^T B_{I^\star})^{-1}\|_{1,1} \max_{g \in B_{\bar{I}}} \|B_{I^\star}^T g\|_1. \quad (7) \end{aligned}$$

We can easily bound the second term of the right hand side of (7) using the *cumulative block coherence* and denoting with $K = \max_i \; rank(B_i)$

$$\begin{aligned} \max_{g \in B_{\bar{I}}} \|B_{I^\star}^T g\|_1 &= \max_{g \in B_{\bar{I}}} \sum_{g' \in B_{I^\star}} |< g', g >| \\ &\leq K \, \mu_{1_B}(m) \quad (8) \end{aligned}$$

since $B_{I^\star}$ is in general composed of m incoherent blocks of maximum rank K.

In order to bound the first term of the right hand side of (7), we follow Tropp again [9] and use the Von Neumann series to compute the inverse $(B_{I^\star}^T B_{I^\star})^{-1}$. Writing $B_{I^\star}^T B_{I^\star} = \mathcal{I} + A$, where $\mathcal{I}$ is the identity matrix, and under the condition that $\|A\|_{1,1} < 1$, it follows that :

$$\begin{aligned} \|(B_{I^\star}^T B_{I^\star})^{-1}\|_{1,1} &= \|(\mathcal{I} + A)^{-1}\|_{1,1} = \|\sum_{k=0}^{\infty} (-A)^k\|_{1,1} \\ &\leq \sum_{k=0}^{\infty} \|A\|_{1,1}^k = \frac{1}{1 - \|A\|_{1,1}}. \end{aligned}$$

The matrix A has zero diagonal and the out of diagonal values correspond to the inner product between atoms from $B_{I^\star}$. Taking into account the structure of $B_{I^\star}$ (it is composed of m incoherent blocks) we can bound the norm using the disparity and cumulative block coherence:

$$\begin{aligned} \|A\|_{1,1} &= \max_k \sum_{j \neq k} |< g_j^{I^\star}, g_k^{I^\star} >| \\ &\leq \xi(\mathcal{D}) + K \, \mu_{1_B}(m-1). \quad (9) \end{aligned}$$

Putting together the bounds obtained in (8),(9) into (7) we get

$$\rho(r_{n-1}) \leq \frac{K \, \mu_{1_B}(m)}{1 - (\xi(\mathcal{D}) + K \, \mu_{1_B}(m-1))}.$$

So the condition

$$K \, \mu_{1_B}(m) + \xi(\mathcal{D}) + K \, \mu_{1_B}(m-1) < 1 \quad (10)$$

ensures that $\rho(r_{n-1}) < 1$ and MP selects an atom from the correct block B_i. By induction the first part of the theorem is proved. For the second part, we simply notice that MP loops in a finite dimensional subset and thus converges exponentially.

□

The main point of this proof is the introduction of the set $B_{I^\star}$. It is crucial in order to allow high redundancy inside each block. In fact if in equation (7) we had used B_I instead of $B_{I^\star}$, the factor $K = \max_i rank(B_i)$ in equation (8) would have been $K = \max_i card(B_i)$, the cardinality of each block!

Bounding the cumulative block coherence by the block coherence $\mu_{1_B}(m-1) \leq \mu_{1_B}(m) \leq m \, \mu_B$, we get a upper bound on the number of recoverable blocks :

$$m < \frac{1 - \xi(\mathcal{D})}{2K \, \mu_B}.$$

In Section V we construct two simple dictionaries that satisfy condition (5) for every m. More redundant dictionaries are constructed in Section VI. The recovery condition is again satisfied, but the maximum number m of blocks that MP can recover is limited. In Table I we list the values of μ_B and the number m of blocks that can be recovered for these dictionaries.

IV. Rate of Convergence

An important factor that determines the quality of a signal expansion is the rate of convergence of the approximation. If the exact block selection condition (10) holds, we can bound the energy of the residual sequence generated by MP using the block coherence defined previously.

Theorem 3: If the signal $f \in \mathcal{V}_I$ and $K \, \mu_{1_B}(m) + \xi(\mathcal{D}) + K \, \mu_{1_B}(m-1) < 1$, then MP picks up atoms only from the correct blocks at each step and

$$\|r_n\|_2^2 \leq \|f\|_2^2 \left(1 - \frac{1 - \xi(\mathcal{D}) - K \, \mu_{1_B}(m-1)}{K \, m}\right)^n. \quad (11)$$

Proof of Theorem 3. From theorem 2 we know that under condition (5) MP picks up atoms from the right set of blocks B_I. At each step the residual belongs to the space $\mathcal{V}_I$, and the energy of the residual is

$$\begin{aligned} \|r_n\|_2^2 &= \|r_{n-1}\|_2^2 - \max_{g \in B_I} |\langle r_{n-1}, g \rangle|^2 \\ &= \|r_{n-1}\|_2^2 \left(1 - \frac{\max_{g \in B_I} |\langle r_{n-1}, g \rangle|^2}{\|r_{n-1}\|_2^2}\right). \quad (12) \end{aligned}$$

In order to bound the decay of the residual energy, we need a lower bound for

$$\frac{\max_{g\in B_I} |\langle r,g\rangle|^2}{\|r\|_2^2}, \tag{13}$$

with $r \in \mathcal{V}_I = span(B_I) = span(B_{I^\star})$ and the set $B_{I^\star}$ is defined in the proof of theorem 2. Since we can write r as a combination of elements from $B_{I^\star}$, $r = B_{I^\star}c$ for some sequence of coefficients c, we have

$$\begin{aligned}\|r\|_2^2 &= \langle r,r\rangle = \left\langle \sum g_i c_i, r\right\rangle \\ &\leq \sum_i |\langle g_i, r\rangle|\,|c_i| \\ &\leq \max_{g\in B_{I^\star}} |\langle g,r\rangle|\,\|c\|_1, \end{aligned}\tag{14}$$

and we obtain the following lower bound for (13)

$$\frac{\max_{g\in B_I} |\langle r,g\rangle|^2}{\|r\|_2^2} \geq \frac{\max_{g\in B_{I^\star}} |\langle r,g\rangle|^2}{\|r\|_2^2} \geq \frac{\|r\|_2^2}{\|c\|_1^2}. \tag{15}$$

We wish to change $\|c\|_1$ with $\|c\|_2$ in order to bound (15) with the minimum norm of the operator $B_{I^\star}$. We know that $rank(B_{I^\star}) = p \leq Km$, where $K = \max_i rank(B_i)$, which means that $\|c\|_0 \leq p \leq Km$, and using the Jensen inequality we have

$$\begin{aligned}\|c\|_2^2 &= \sum_{i=1}^{p} c_i^2 = p\sum_{i=1}^{p}\frac{1}{p}|c_i|^2 \\ &\geq p\left(\sum_{i=1}^{p}\frac{1}{p}|c_i|\right)^2 = \frac{1}{p}\|c\|_1^2.\end{aligned}$$

Using the upper bound $\|c\|_1^2 \leq p\,\|c\|_2^2 \leq Km\,\|c\|_2^2$ into (15), we obtain

$$\frac{\max_{g\in B_I} |\langle r,g\rangle|^2}{\|r\|_2^2} \geq \frac{\|r\|_2^2}{Km\,\|c\|_2^2}. \tag{16}$$

Using the Thin Singular Value Decomposition we can write $B_{I^\star} = U\Sigma V^T$, with orthogonal matrices U, V and Σ is diagonal and full rank since $B_{I^\star}$ has full rank. We now write:

$$\begin{aligned}\|r\|_2^2 &= \|B_{I^\star}c\|_2^2 = c^T V\Sigma^2 V^T c \qquad (y = V^T c) \\ &= y^T\Sigma^2 y = \sum_i \sigma_i^2 y_i^2 \\ &\geq \sigma_{min}^2\,\|y\|_2^2 = \sigma_{min}^2\,\|c\|_2^2.\end{aligned}\tag{17}$$

The square singular values of $B_{I^\star}$ coincide with the eigenvalues of the the Gram matrix $G = B_{I^\star}^T B_{I^\star}$, since Σ^2 and G are similar matrices. The smallest eigenvalue $\lambda_{min} = \sigma_{min}^2$ can be bounded using the Geršgorin disc theorem [13]: every eigenvalue of G lies in one of the p discs

$$\text{Disc}_k = \left\{ z : |G_{kk} - z| \leq \sum_{j\neq k} |G_{jk}| \right\}.$$

The matrix G has unit diagonal because of the normalisation of the atoms. Taking into account the block incoherent structure of $B_{I^\star}$ we can bound the sum above with

$$|1-\lambda_{min}| \leq \sum_{j\neq k} |G_{jk}| \leq \xi(\mathcal{D}) + K\mu_{1_B}(m-1),$$

and the square minimum singular value $\sigma_{min}^2 \geq 1 - \xi(\mathcal{D}) - K\mu_{1_B}(m-1)$. Putting this bound into (17) and (16) we obtain

$$\frac{\max_{g\in B_I} |\langle r,g\rangle|^2}{\|r\|_2^2} \geq \frac{1-\xi(\mathcal{D}) - K\mu_{1_B}(m-1)}{Km}.$$

Finally from (12) we end the proof

$$\begin{aligned}\|r_n\|_2^2 &\leq \|r_{n-1}\|_2^2\left(1 - \frac{1-\xi(\mathcal{D}) - K\mu_{1_B}(m-1)}{Km}\right) \\ &\leq \|f\|_2^2\left(1 - \frac{1-\xi(\mathcal{D}) - K\mu_{1_B}(m-1)}{Km}\right)^n.\end{aligned}$$

□

This result is very similar in nature to those already obtained in [9], [12], expressed at the level of blocks, though. However the arbitrarily high redundant structure of the blocks is not taken into account although we know that the energy decay rate of the residual generated by MP is strongly influenced by the redundancy of the dictionary [11], [14]. Indeed, the decay of the residue norm is bounded by an exponential

$$\|r_n\|_2^2 \leq (1-\beta_D^2)^n\,\|f\|_2^2,$$

where the parameter β_D depends on the size/structure of the dictionary. In particular it corresponds to the cosine of the maximum angle between any possible f in the span of the dictionary and the closest atom of the dictionary. As we are dealing with a block incoherent dictionary, the redundancy parameter β_D is affected by the "holes" of the dictionary due to the incoherence between blocks. Consider for example a dictionary with two very dense blocks orthogonal to each other, the redundancy parameter will be $\beta_D \leq \frac{\sqrt{2}}{2}$.

At this point, it is therefore natural to take into account the redundancy parameter β_i of each block B_i. We thus define :

$$\max_{g^i\in B_i} \left|\left\langle f, g^i\right\rangle\right| \geq \beta_i\,\|f\|_2 \qquad \forall f \in span(B_i). \tag{18}$$

We can now analyse the energy decay of the residual using the block redundancy factor which leads to the following result.

Theorem 4: If the signal $f \in \mathcal{V}_I$ and $K\,\mu_{1_B}(m) + \xi(\mathcal{D}) + K\,\mu_{1_B}(m-1) < 1$, then MP picks up atoms only from the correct blocks at each step and

$$\|r_n\|_2^2 \leq \|f\|_2^2\left(1 - \frac{\beta^2}{m}\right)^n, \tag{19}$$

where $\beta = \min_i \beta_i$, and β_i is related to the redundancy and structure of block B_i (18).

When $\beta = \min_i \beta_i$ is bigger than $\sqrt{\frac{1-\xi(D)-K\mu_{1_B}(m)}{K}}$, this result is actually better than Theorem 3. The proof requires a simple lemma and some more notation.

Lemma 1: Let $f \in \mathcal{V}_I = span(B_I)$ with $B_I = \bigcup_{i=1}^{m} B_i$. If we indicate with f^i the orthogonal projection of f onto the space $V_i = span(B_i)$, it follows that

$$\|f\|_2^2 \leq \sum_{i=1}^{m} \left\|f^i\right\|_2^2. \tag{20}$$

Notice that this is not a generalised triangular inequality since $\sum_{i=1}^{m} f^i \neq f$. The proof is however simple and omitted for concision.

Proof of Theorem 4. By induction we know that the sequence of residuals $r_n \in \mathcal{V}_I$. The normalisation of the atoms implies

$$\|r_n\|_2^2 = \|r_{n-1}\|_2^2 - \max_{g \in B_I} |\langle r_{n-1}, g\rangle|^2. \tag{21}$$

In order to caracterize the decay of the residual energy, we need a meaningful lower bound for $\max_{g\in B_I} |\langle r, g\rangle|^2$ where $r \in \mathcal{V}_I$. If MP selects an atom from the block B_j, it follows that

$$\begin{aligned} \max_{g\in B_I} |\langle r, g\rangle|^2 &= \max_{g\in B_j} \left|\langle r^j, g\rangle\right|^2 \\ &\geq \beta^2 \left\|r^j\right\|_2^2 \\ &\geq \beta^2 \frac{\|r\|_2^2}{m} \end{aligned} \tag{22}$$

where r^j is the orthogonal projection of r on $V_j = span(B_j)$. Inequality (22) can be derived analysing the case of a residual r with energy uniformly spread over all V_i's. This means $\left\|r^i\right\|_2^2 = C$ for all $i \in I$ and for some positive constant C. Using *lemma 1* it follows that

$$\left\|r^j\right\|_2^2 \geq \frac{\|r\|_2^2}{m}. \tag{23}$$

When the energy is not uniformly spread, it means there is at least one component r^k, $k \in I$, with energy bigger than (23) and MP will thus select an atom from B_k. Putting (22) in (21) we end the proof

$$\begin{aligned} \|r_n\|_2^2 &\leq \|r_{n-1}\|_2^2 - \beta^2 \frac{\|r_{n-1}\|_2^2}{m} \\ &\leq \|f\|_2^2 \left(1 - \frac{\beta^2}{m}\right)^n. \end{aligned}$$

□

The exponential bound in eq. (19) depends on the redundancy factor β and the number of blocks m. The parameter β can be made close to one by increasing the redundancy inside each block. Notice that the exact block recovery condition must remain valid.

The result of Theorem 4 is obtained considering at each iteration the worst possible residual r_n which has equally distributed energy over the subspaces spanned by the blocks $B_i \in B_I$. It is clear that a function f with energy equally spread over the m different subspaces will be approximated by MP with the slowest error energy decay. Analysing carefully this case, we see that at first iteration as in Theorem 4 we can bound the residual energy by

$$\|r_1\|_2^2 \leq \|f\|_2^2 \left(1 - \frac{\beta^2}{m}\right). \tag{24}$$

At the second iteration MP will select an atom from a different block provided the following condition is satisfied :

$$\frac{\sqrt{1-\beta^2}}{\beta^2} + \frac{\mu_B}{\beta} < \frac{1}{\sqrt{m}}. \tag{25}$$

In this case the energy of the residual can be bounded by

$$\|r_2\|_2^2 \leq \|f\|_2^2 \left(1 - \beta^2 \frac{2}{m}(1 - \mu_B)\right). \tag{26}$$

Condition (25) is quite strong, but we can say that if at first iteration a block B_i is selected then a different different block B_j will be selected at the second iteration if

$$\|f^j\|_2 > \|f^i\|_2 \left(\frac{\sqrt{1-\beta^2}}{\beta} + \mu_B\right). \tag{27}$$

Bound (26) is tighter than (19) when $\mu_B < \frac{\beta^2}{2m}$.

With similar arguments it is possible to generalise this bound. After n iteration, with $n < m$ and under appropriate different block selection condition we have that

$$\|r_n\|_2^2 \leq \|f\|_2^2 \left(1 - \beta^2 \frac{n}{m}(1 + \mu_B P_{n-2}(\mu_B))\right), \tag{28}$$

where $P_{n-2}(\mu_B)$ is a polynomial of degree $n-2$. Notice that it becomes more difficult to check when the bound in (28) is better than in (19).

V. Block incoherent dictionary examples

Let us now analyze two simple examples that satisfy the aforementioned constraints. We are interested in building a dictionary whose block coherence we explicitly control. The easiest way is to start by designing the special subdictionary $B_{i^\star}$, introduced in def.(5), and then add redundancy inside each block. As we will now see this allows us to get rid off the disparity ξ in the simple case where $B_{i^\star}$ is an orthogonal basis.

Example 1: The simplest block dictionary is that one with orthogonal blocks,

$$\mathcal{D} = \bigcup_{n=0}^{N-1} B_n \quad \text{with} \quad B_i \perp B_j \quad \text{if} \quad i \neq j.$$

For simplicity let us examine the case of blocks with rank $K = 2$ in the real d dimensional vector space IR^d. We can take any orthonormal basis set for IR^d $\{e_0, e_1, \ldots, e_{d-1}\}$ and supposing d is even, collect $N = d/2$ sets each containing two vectors, $U_n = \{e_{2n}, e_{2n+1}\}$ with $n = 0, \ldots, N-1$. Now we can add redundancy inside each set

$$B_n = \left\{ \bigcup_{p=0}^{P-1} e_{2n} \cos\left(\frac{\pi}{P}p\right) + e_{2n+1} \sin\left(\frac{\pi}{P}p\right) \right\},$$

increasing the redundancy parameter P we obtain more redundant blocks. Obviously if a signal f belongs to the subspace $\mathcal{V}_I$ generated by the union of m blocks, it follows that the inner products between f and all the atoms that are not in B_I are zero, and so MP recovers atoms from correct blocks. We just notice that the hypotheses of Theorem 2 are satisfied since for this trivial dictionary we have $\xi(\mathcal{D}) = 0$ and $\mu_{1_B}(m) = 0,\ \forall m$.

Let us make things more complicated and design non orthogonal blocks.

Example 2: We start once again from an orthonormal basis set for IR^d and we construct the set $B_{i^\star} = \{a_i, b_i\}$ combining the vectors $\{e_0, \ldots, e_{d-1}\}$ in this way

$$\begin{cases} a_0 = e_0 \\ b_0 = e_1 \cos(\alpha) + e_2 \sin(\alpha) \end{cases}$$

$$\begin{cases} a_1 = e_2 \\ b_1 = e_3 \cos(\alpha) + e_4 \sin(\alpha) \end{cases}$$

$$\vdots$$

$$\begin{cases} a_i = e_{2i} \\ b_i = e_{2i+1} \cos(\alpha) + e_{2i+2} \sin(\alpha) \end{cases}$$

$$\vdots$$

$$\begin{cases} a_{\frac{d}{2}-1} = e_{d-2} \\ b_{\frac{d}{2}-1} = e_{d-1} \cos(\alpha) + e_0 \sin(\alpha) \end{cases}$$

and as before we can add redundancy in order to get the redundant blocks

$$B_n = \left\{ \bigcup_{p=0}^{P-1} a_n \cos\left(\frac{\pi}{P} p\right) + b_n \sin\left(\frac{\pi}{P} p\right) \right\}.$$

For this dictionary we have that the rank of all blocks is $K = 2$ and the disparity is $\xi(\mathcal{D}) = 0$. Without putting any constraint on the redundancy, it is quite easy to bound the *cumulative coherence* function

$$\begin{aligned} \mu_{1_B}(1) &= \mu_B = |\sin(\alpha)| \\ \mu_{1_B}(2) &= \max_{\substack{v \in B_i \\ u_{i-1} \in B_{i-1} \\ u_{i+1} \in B_{i+1}}} |\langle v, u_{i-1} \rangle| + |\langle v, u_{i+1} \rangle| \\ &\leq \left| \left\langle a_i \cos\left(\frac{\pi}{4}\right) + b_i \sin\left(\frac{\pi}{4}\right), b_{i-1} \right\rangle \right| + \left| \left\langle a_i \cos\left(\frac{\pi}{4}\right) + b_i \sin\left(\frac{\pi}{4}\right), a_{i+1} \right\rangle \right| \\ &\leq \left| \cos\left(\frac{\pi}{4}\right) \sin(\alpha) \right| + \left| \cos\left(\frac{\pi}{4}\right) \sin(\alpha) \right| \\ &\leq \sqrt{2} \, |\sin(\alpha)| \\ \mu_{1_B}(m) &= \mu_{1_B}(2) \qquad \text{for} \qquad m > 2. \end{aligned}$$

For α small and positive and $m > 2$ the recovery condition becomes

$$2\mu_{1_B}(2) + 2\mu_{1_B}(2) = 4\sqrt{2} \sin(\alpha) < 1.$$

Therefore we can say that for every m block sparse signal, MP is able to recover atoms from the correct blocks when

$$\alpha < \arcsin\left(\frac{1}{4\sqrt{2}}\right) \simeq 10^o.$$

The two dictionaries designed above are trivial since there is no redundancy of "subspace" :

$$\sum_{n=0}^{N-1} \mathrm{rank}(B_n) = d.$$

We would like to build a dictionary that is the union of N blocks such that $\sum_{n=0}^{N-1} \mathrm{rank}(B_n) > d$. Since this problem is in general non-trivial, we next describe a numerical technique that can design such dictionaries.

VI. Block dictionaries and Grassmannian packings

Suppose we want to build a block dictionary $\mathcal{D} = \bigcup_{n=0}^{N-1} B_n$ to represent signals in the vector space IR^d. If we indicate with K the rank of each block, $K = \mathrm{rank}(B_n)\ \forall n$, the block incoherent dictionary design problem identifies with finding N subspaces of dimension K in IR^d which are as far apart as possible. This is equivalent to the Grassmannian packing problem [15].

The real Grassmannian space $G(K, \mathrm{IR}^d)$ is the set of all K-dimensional subspaces of IR^d. The packing problem is the problem of finding N points of $G(K, \mathrm{IR}^d)$ such that the minimum distance between any two of these points becomes as large as possible. In order to connect the Grassmannian packing problem to the block incoherent dictionary design problem, we shall use an appropriate metric to pack points of $G(K, \mathrm{IR}^d)$. In particular we need a metric that upper bounds the cumulative block coherence. The packing radius of a set E is the size of largest open ball that can be centered on any point and doesn't contain any other point in the set :

$$\mathrm{pack}(E) = \min_{i \neq j} d_E(g_i, g_j),$$

where d_E is a distance in E. Since $\mu_{1_B}(m) \leq m\mu_B$, we are looking for a metric in $G(K, \mathrm{IR}^d)$ such that when the packing radius of a set of N points of $G(K, \mathrm{IR}^d)$ is bigger than some $\rho > 0$, it follows that the block coherence naturally associated to these N points is controlled by the radius : $\mu = h(\rho)$.

The coherence between two blocks B_i, B_j is defined as the largest inner product between any two vectors from the two blocks. Consider the two sets $B_{i^\star}$ and $B_{j^\star}$ whose columns form orthonormal bases for the subspaces $span(B_i)$ and $span(B_j)$. The maximum singular value of the product $B_{ij} = B_i^T B_j$, which coincides with the $(2,2)$ operator norm $\|B_{ij}\|_{2,2}$, bounds the coherence between B_i and B_j

$$\mu_{ij} = \max_{k,l} \left| \left\langle g_k^i, g_l^j \right\rangle \right| \leq \|B_{ij}\|_{2,2}. \qquad (29)$$

Since we do not put any constrain on the way we add redundancy to build the blocks B_i and B_j from respectively $B_{i^\star}$ and $B_{j^\star}$, we can say that eq. (29) is sharp and proceed with the worst case.

Notice that the maximum singular value of B_{ij} is just the cosine of the smallest principal angle θ_1, the minimum angle formed by any pair of vectors from $span(B_i)$ and $span(B_j)$. Suppose that $\mathcal{E}_i$, $\mathcal{E}_j$ are two subspaces or points of $G(K, \mathrm{IR}^d)$ and E_i, E_j are orthonormal bases for the respective subspaces, we shall use the following spectral distance measure as metric for the Grassmannian packing problem :

$$\mathrm{dist}(\mathcal{E}_i, \mathcal{E}_j) = \sin(\theta_1) = \sqrt{1 - \left\| E_i^T E_j \right\|_{2,2}^2}.$$

An optimal packing of N subspaces of $G(K, \mathrm{IR}^d)$ is a set $\mathcal{E}$ that maximizes the packing radius

$$\mathrm{pack}(\mathcal{E}) = \min_{i \neq j} \mathrm{dist}(\mathcal{E}_i, \mathcal{E}_j).$$

If we are able to pack N points with a packing radius $\mathrm{pack}(\mathcal{E}) \geq \rho$ using the spectral distance, at the same time we

obtain a block dictionary with coherence $\mu_B \leq \sqrt{1-\rho^2}$. An elegant method for solving packing problems in Grassmannian spaces equipped with various metrics has been developed by Tropp [10], [16]. The method consists in constructing a Gram matrix that has certain structural and spectral properties. The structural constrains control the packing radius, while the spectral properties are needed in order to be able to associate to the matrix a set of N points in $G(K, \mathrm{IR}^d)$. It turns out to be a difficult issue to impose both kind of properties simultaneously, and [16] proposes an iterative algorithm that alternatively enforces the two properties. Details about the alternating projection algorithm can be found in [10], [16]. We just recall here the matrix representation of a set of N points in $G(K, \mathrm{IR}^d)$. Such a set is represented by a set of N matrices of dimension $d \times K$, $\mathcal{E} = \{E_n\}$. The columns of each matrix will be an orthogonal basis for each subspaces or points of $G(K, \mathrm{IR}^d)$. The N matrices are concatenated to form a $d \times KN$ matrix

$$E = [E_1 E_2 \cdots E_N],$$

and the structural properties can be easily imposed to the Gram matrix $G = E^T E$ (see [10]) .

We implemented the alternating projection algorithm for the block dictionary construction. The algorithm returns a block incoherent dictionary composed of N blocks with *block coherence* $\mu_B \leq \mu = \sqrt{1-\rho^2}$ that bounds the cumulative block coherence :

$$\mu_B(m) \leq m\mu_B \leq m\mu.$$

The recovery condition of Theorem 2 gives the condition so that the block labels of any sparse signal drawn from such a dictionary can be perfectly recovered using MP. The disparity $\xi(\mathcal{D})$ is zero since each block contains an orthogonal basis. Indeed we are able to recover a m block sparse signal when

$$mK\mu + (m-1)K\mu < 1.$$

It worth noticing that the bound in equation (9) can be improved. In fact we need a bound for the (1,1)-norm of the matrix $A = B_{I^\star}^T B_{I^\star} - I$, where $B_{I^\star}^T B_{I^\star}$ is just a m blocks sub-matrix of the Gram matrix $G = E^T E$ (remember that $B_{i^\star} = E_i$). Therefore it follows that

$$\|A\|_{1,1} \leq (m-1) \max_{i \neq j} \|G_{ij}\|_{1,1},$$

where $G_{ij} = E_i^T E_j$. It is rather difficult to impose another structural condition to the matrix G such as $\max_{i \neq j} \|G_{ij}\|_{1,1} < \mu_K$, so when implementing the alternating projection algorithm we just compute the maximum value μ_K and in general we have that

$$\mu_K \triangleq \max_{i \neq j} \|G_{ij}\|_{1,1} < K\mu. \tag{30}$$

The recovery condition becomes $mK\mu + (m-1)\mu_K < 1$ which means

$$m < \frac{1+\mu_K}{K\mu + \mu_K}. \tag{31}$$

Using the alternating projection algorithm we built dictionaries for IR^d with $d = 100, 500, 800, 1000$ and different redundancy varying the numbers of block N and the rank K. The resulting upper bound μ for the block coherence and the number of block m for which the recovery condition is satisfied, are given in Table I . Between brakets is the upper bound for m from eq. (31).

K	N	d	μ	μ_K	m
3	50	100	0.104	0.179	2 (2.39)
4	50	100	0.144	0.289	1 (1.48)
10	60	500	0.060	0.189	1 (1.50)
6	100	500	0.046	0.114	2 (2.83)
7	100	500	0.064	0.196	1 (1.89)
8	100	500	0.078	0.219	1 (1.44)
3	200	500	0.033	0.058	6 (6.63)
4	200	500	0.055	0.111	3 (3.33)
5	200	500	0.071	0.160	2 (2.24)
2	300	500	0.027	0.039	10 (10.9)
3	300	500	0.053	0.092	4 (4.33)
4	300	800	0.041	0.083	4 (4.32)
5	300	800	0.054	0.121	2 (2.84)
5	300	1000	0.041	0.093	3 (3.63)

TABLE I

BLOCK INCOHERENT DICTIONARIES OBTAINED WITH THE ALTERNATING PROJECTION ALGORITHM OF [10]. K IS THE RANK OF THE BLOCKS, N THE NUMBER OF BLOCKS, d THE DIMENSION OF THE REAL VECTOR SPACE, μ THE UPPER BOUND FOR THE BLOCK COHERENCE μ_B. THE CONSTANT μ_K IS DEFINED IN EQ. (30) AND m IS THE NUMBER OF BLOCKS WITH ITS UPPER BOUND GIVEN BY (31) IN BETWEEN THE PARENTHESIS.

It is interesting to notice that it is easier to designe a block incoherent dictionary with low rank blocks. For example consider the cases with $d = 500$ and the "dimension redundancy" $\frac{K \cdot N}{d} = \frac{600}{500}$ that corresponds to $K = 10, 6, 3$ and $N = 60, 100, 200$ respectively. If we relax the integer constrain on m, we can say that for $K = 10$ MP is able to correctly "recover" $K \cdot m \simeq 15$ dimensions, while for $K = 6$ we have $K \cdot m \simeq 17$ and for $K = 3$, $K \cdot m \simeq 20$. In practice MP can recover less block sparse signals in comparison to the prediction of the theorem presented here, which are based on worst case analysis. To illustrate this we ran simulations using a dictionary for IR^d with $d = 100$, $K = 3$ and $N = 50$ with a redundancy factor of each block $\frac{|B_i|}{K} = 10$. Random signals f, generated from random sets of m blocks (m-block sparse signals) with $m = 1, \cdots, d/K$ were decomposed with MP and the frequency of correct block selection successes is plotted in Fig. 1. For example 9948 over 10000 random 11-block sparse signals are recovered by MP, where Theorem 2 garanties the possibility to recover only 2-block-sparse signals, as also noted in Table I.

VII. CONCLUSIONS AND OUTCOME

We discussed the formal advantages of using block incoherent dictionaries in terms of block recovery properties and approximation results for redundant dictionaries. We showed that, if one is willing to drop the exact recovery of atoms and exchange it with a weaker block recovery property, there are several advantages with this new construction. The block incoherent structure allows one to design dictionaries that are very redundant, yet maintaining the stability of the decomposition at the level of the blocks. There are at least two reasons

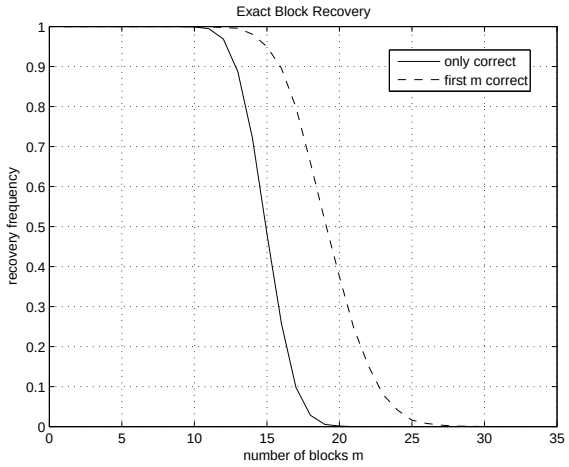

Fig. 1. MP correct block selection for random m-block sparse signals. Normalized frequency of right block selection versus the number of blocks from which the signal is randomly generated. Solid line: only correct blocks are selected. Dashed line: all the correct blocks are selected before a wrong block had been selected.

why this could be important in practice. First, redundant dictionaries are more easily designed when one doesn't have to handle the incoherence constraint. In this paper we have proposed a numerical algorithm to construct block incoherent dictionaries through Grassmannian packings. Interestingly the link with packings opens the door to applications in coding theory as already noticed in [15] and [17] and the influence of redundancy inside blocks could be worth investigating as a way to overcome channel errors. Redundancy has been frequently advocated as an important property when designing efficient dictionaries for higher dimensional signals, most notably images. We have shown how redundant blocks affect the approximation properties of dictionaries, confirming that, in the stable regime of the pursuit, redundancy increases the approximation rate. Second, the block incoherent constraint studied here imposes a particular *structure* on the dictionary. This structure was used in [18] to derive a fast tree-based algorithm for implementing MP and in [19] for multiple description coding of images. These early practical applications of the construction presented in this paper are encouraging but also bring several interesting questions. For example, the Grassmannian packing construction we presented here cannot be used for big high-dimensional dictionaries (typically the ones encountered in image processing). It would be interesting to come up with provably correct design techniques that would build a block incoherent dictionary starting from an initial very redundant one. On the theoretical side, the worst case analysis performed in this paper provide bounds that are still far what one experiments in practice. Clearly these results should be refined.

REFERENCES

[1] R. DeVore, "Nonlinear approximation," *Acta Numerica*, vol. 7, pp. 51–150, 1998.

[2] E. J. Candès and D. L. Donoho, "Curvelets - a surprisingly effective nonadaptive representation for objects with edges." *Curves and Surfaces, L. L. S. et al., ed., Nashville, TN, (Vanderbilt University Press)*, pp. 123–143, 1999.

[3] L. Daudet and B. Torrésani, "Hybrid representations for audiophonic signal encoding," *Signal Processing*, vol. 82, no. 11, pp. 1595–1617, November 2002.

[4] S. Chen and D. Donoho, "Atomic Decomposition by Basis Pursuit," in *SPIE International Conference on Wavelets*, Sant Diego, July 1995.

[5] M. Elad and A. M. Bruckstein, "Generalized uncertainty principles and sparse representation in pairs of bases," *IEEE Trans. Inform. Theory*, vol. 48, no. 9, pp. 2558–2567, September 2002.

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

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

[8] A. C. Gilbert, S. Muthukrishnan, and M. J. Strauss, "Approximation of functions over redundant dictionaries using coherence," in *Proc. 14th Annual ACM-SIAM Symposium on Discrete Algorithms*, 2003.

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

[10] J. Tropp, I. Dhillon, R. Heath, and T. Strohmer, "Designing structured tight frames via alternating projection," *IEEE Trans. Inform. Theory*, vol. 51, no. 1, pp. 1568–1570, January 2005.

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

[12] R. Gribonval and P. Vandergheynst, "On the exponential convergence of matching pursuits in quasi-incoherent dictionaries," *IEEE Transactions on Information Theory*, vol. 52, no. 1, pp. 255–261, January 2006.

[13] R. Horn and C. Johnson, *Matrix Analysis*. Cambridge University Press, 1985.

[14] P. Frossard and P. Vandergheynst, "Redundancy in Non-Orthogonal Transform," in *Proceedings of IEEE International Symposium on Information Theory*, June 2001.

[15] J. Conway, R. Hardin, and N. Sloane, "Packing lines, planes, etc.: Packings in grassmannian spaces," *Experimental Mathematics*, vol. 51, no. 2, pp. 139–159, April 1996.

[16] J. Tropp, "Topics in sparse approximation," Ph.D. dissertation, Computational and Applied Mathematics, UT-Austin, Tech. Rep., August 2004.

[17] T. Strohmer and R. Heath, "Grassmannian frames with applications to coding and communications," *Applied and Computational Harmonic analysis*, vol. 14, no. 3, pp. 257–275, May 2003.

[18] P. Jost, P. Vandergheynst, and P. Frossard, "Tree-based pursuit: Algorithm and properties," *to appear in IEEE Transactions on Signal Processing*, 2006.

[19] I. Radulovic and P. Frossard, "Multiple description image coding of with block-coherent redundant dictionaries," in *Proc. Picture Coding Symposium*, Beijin, China, 2006.