Error estimates for orthogonal matching pursuit and random dictionaries *

Paweł Bechler
Warsaw University
pbechler@mimuw.edu.pl

Przemysław Wojtaszczyk
Warsaw University
wojtaszczyk@mimuw.edu.pl

August 5, 2009

Abstract

In this paper we investigate the efficiency of the Orthogonal Matching Pursuit (OMP) for random dictionaries. We concentrate on dictionaries satisfying the Restricted Isometry Property. We also introduce a stronger Homogenous Restricted Isometry Property which we show is satisfied with overwhelming probability for random dictionaries used in compressed sensing. For these dictionaries we obtain upper estimates for the error of approximation by OMP in terms of the error of the best n-term approximation (Lebesgue-type inequalities). We also present and discuss some open problems about OMP. This is a development of recent results obtained by D.L. Donoho, M. Elad and V.N. Temlyakov.

Keywords: Orthogonal matching pursuit, restricted isometry property, random dictionaries, Lebesgue inequalities, n-term approximation.

AMS classification: 41A25 (15A52, 41A17, 41A46)

1 Introduction

In this paper we investigate the efficiency of the Orthogonal Matching Pursuit algorithm (OMP), also known in literature as Orthogonal Greedy Algorithm,

*This research was partially supported by the Polish Ministry of Science and Higher Education grant no. N N201 269335.

for random dictionaries. OMP (cf. [8, 9]) is a well known greedy algorithm widely used in approximation theory, statistical estimations and compressed sensing (for a general review of greedy algorithms see [12]). One of its main features is that it can be applied for arbitrary dictionary. However the efficiency of the algorithm seems to depend very strongly on properties of the dictionary.

In this paper we work in the context of a Hilbert space $\mathcal{H}$ (which may be assumed to be finite dimensional) with the scalar product $\langle\, ,\,\rangle$ and the norm $\|\;\|$. The dictionary is a subset $\mathbf{\Phi} = \{\phi_j : j \in J\} \subset \mathcal{H}$ such that $\overline{\operatorname{span}\mathbf{\Phi}} = \mathcal{H}$. We usually assume that $\|x\|$ is close to 1 for $x \in \mathbf{\Phi}$. Usually in the literature it is assumed that $\|x\| = 1$ for $x \in \mathbf{\Phi}$ (see e.g. [12]). However for random dictionaries it is very rarely satisfied. On the other hand for such dictionary $\|x\|$ is close to 1 with great probability.

In the space $\mathcal{H}$ we consider the Orthogonal Matching Pursuit algorithm with respect to the dictionary $\mathbf{\Phi}$. This algorithm obtains iteratively a sequence $\mathrm{OMP}_n f \in \mathcal{H}$ of approximants of a given element $f \in \mathcal{H}$ in the following way:

- Define $\mathrm{OMP}_0 f = 0$.
- Given $\mathrm{OMP}_{n-1} f$ choose $j_n \in J$ such that

  $$|\langle f - \mathrm{OMP}_{n-1} f, \phi_{j_n}\rangle| = \sup\{|\langle f - \mathrm{OMP}_{n-1} f, \phi_j\rangle| : j \in J\}$$

  and define $\mathrm{OMP}_n f$ as the orthogonal projection of f onto the subspace $\operatorname{span}\{\phi_{j_1}, \dots, \phi_{j_n}\}$.

For a fixed $f \in H$ we denote $f_n = f - \mathrm{OMP}_n f$.

The standard measure of approximation power of a dictionary is the error of the best m–term approximation. We define the set of m-sparse vectors (with respect to the dictionary $\mathbf{\Phi}$) as

$$\Sigma_m(\mathbf{\Phi}) = \Sigma_m = \left\{ \sum_{j=1}^{m} a_j \phi_j \;:\; \{\phi_j\}_{j=1}^m \subset \mathbf{\Phi} \right\}. \tag{1.1}$$

For a given $f \in \mathcal{H}$ we define its best error of m–term approximation (cf. [12]) as

$$\sigma_m(f, \mathbf{\Phi}) = \inf\{\|f - z\| \;:\; z \in \Sigma_m\}. \tag{1.2}$$

Clearly, we always have $\sigma_m(f) \le \|f - \mathrm{OMP}_m(f)\| = \|f_m\|$.

When our dictionary is an orthonormal basis then, obviously, $\sigma_m(f) = \|f - \mathrm{OMP}_m(f)\|$ for each $f \in \mathcal{H}$. Unfortunately, this is the only case when it is so. The fundamental, and still largely unanswered question is how close

$\mathrm{OMP}_m(f)$ can get to this optimal rate of approximation given by $\sigma_m(f)$. It is to be expected that the answer to the above question must depend on some extra properties of the dictionary. We will discuss it in more detail in the last Section of the paper.

In this paper we concentrate on a random dictionary in $\mathbb{R}^n$ of the following form: $\mathbf{\Phi} = \{\phi_1, \dots, \phi_N\}$, with $\phi_j = \frac{1}{\sqrt{n}}(\eta_{1,j}, \dots, \eta_{n,j})$ where $(\eta_{i,j})_{i=1}^{n}{}_{j=1}^{N}$ are independent, identically distributed, mean zero subgaussian random variables with $\mathbb{E}\eta_{i,j}^2 = 1$. It is a natural class of dictionaries which recently gained prominence due to its importance in compressed sensing (see e.g. [2, 5, 4]). In compressed sensing we think about such a dictionary as a matrix whose columns are ϕ_j's. Then any approximation scheme for such a dictionary provides a decoder for a measurement matrix $\mathbf{\Phi}$. For such random dictionaries we prove that there exist positive constants c, c_1, c_2 such that for $K = cn/\log_2 N$ and $0 \le k < S \le K$ we have

$$\|f_S\|^2 \le c_1\big(\sigma_{S-k}(f_k) + c_2\sqrt{S/K}\lceil \log_2(2S-k)\rceil \|f_k\|\big). \tag{1.3}$$

As a main application we derive the estimate

$$\big\|f_{\lceil m(4\log_2 m - 1)\rceil}\big\| \le c\sigma_m(f) \tag{1.4}$$

valid for $m \le c\sqrt{K}$. These results improve for random dictionaries the results from [6]. Technically speaking, the results in [6] are for dictionaries having small coherence while we introduce a different assumption: *homogenous restricted isometry property.*

2 Dictionaries

Despite the fact that we are mostly interested in random dictionaries, our main results are formally deterministic. We isolate the properties of a dictionary which a random dictionary has with overwhelming probability and prove our results under the assumption that our dictionary has this property. A widely used characteristic of a dictionary is its coherence.

Definition 1. The *coherence* of a dictionary $\mathbf{\Phi}$ is defined as

$$\eta = \eta(\mathbf{\Phi}) = \sup\{|\langle \phi_1, \phi_2\rangle| : \phi_1, \phi_2 \in \mathbf{\Phi}, \phi_1 \ne \phi_2\}.$$

Recently, especially in the context of compressed sensing, a restricted isometry property (RIP for short) became very useful. Let us recall the following well known definition (c.f. [2]) phrased in terms of dictionary not a measurement matrix.

Definition 2. The dictionary $\mathbf{\Phi}$ satisfies the *Restricted Isometry Property* $\mathrm{RIP}(K,\varepsilon)$, with $0<\varepsilon<1$, if for any subset $I\subset J$ with $\#I\le K$ and any scalars a_j, $j\in I$, the following inequality holds:

$$(1-\varepsilon)\left(\sum_{j\in I}|a_j|^2\right)^{1/2}\le\left\|\sum_{j\in I}a_j\phi_j\right\|\le(1+\varepsilon)\left(\sum_{j\in I}|a_j|^2\right)^{1/2}. \tag{2.1}$$

This definition in particular means that $\{\phi_j\}_{j\in I}$ is a Riesz basis in its linear span. From [3, Prop. 3.6.4] we get the following

Proposition 2.1. *If the dictionary* Φ *satisfies* $\mathrm{RIP}(K,\varepsilon)$ *with* $I\subset J$ *such that* $\#I\le K$ *and* $f\in\mathrm{span}\{\phi_i : i\in I\}$, *then*

$$(1-\varepsilon)\|f\|\le\left(\sum_{i=1}^{n}|\langle f,\phi_i\rangle|^2\right)^{1/2}\le(1+\varepsilon)\|f\|.$$

The following is true:

Proposition 2.2. *(i) If the dictionary* $\mathbf{\Phi}$ *has coherence* η *then it satisfies* $\mathrm{RIP}(K,\eta(K-1))$ *for* $K\le\eta^{-1}+1$.

(ii) If the dictionary $\mathbf{\Phi}$ *satisfies* $\mathrm{RIP}(K,\varepsilon)$, *then* $\eta(\mathbf{\Phi})\le\varepsilon(2+\varepsilon)$.

Proof. (i) is shown in [6, Lemma 2.1]. (ii) is obtained by straightforward calculation. □

In this paper we concentrate on a random dictionary in $\mathbb{R}^n$ of the following form: $\mathbf{\Phi}=\{\phi_1,\dots,\phi_N\}$ where $\phi_j=\frac{1}{\sqrt{n}}(\eta_{1,j},\dots,\eta_{n,j})$ where $(\eta_{i,j})_{i=1}^{n}{}_{j=1}^{N}$ are independent, identically distributed, mean zero subgaussian random variables with $\mathbb{E}\eta_{i,j}^2=1$. In compressed sensing we think about such a dictionary as a random matrix whose columns are ϕ_j's.

Let us introduce the following

Definition 3. The dictionary $\mathbf{\Phi}$ has homogenous restricted isometry property $\mathrm{HRIP}(k,\delta)$, $0<\delta<1$ if for any set $T\subset\{1,\dots,N\}$ with $\#T=l\le k$ and any sequence of numbers a_j we have

$$\left(1-\delta\sqrt{\tfrac{l}{k}}\right)\left(\sum_{j\in T}|a_j|^2\right)^{1/2}\le\left\|\sum_{j\in T}a_j\phi_j\right\|\le\left(1+\delta\sqrt{\tfrac{l}{k}}\right)\left(\sum_{j\in T}|a_j|^2\right)^{1/2}. \tag{2.2}$$

The following theorem whose proof uses standard arguments justifies this definition.

Theorem 2.3. *Suppose that integers n, N and numbers $0 < \delta < 1$ and $a > 0$ are given and suppose that the dictionary $\mathbf{\Phi} = \{\phi_1, \dots, \phi_N\} \subset \mathbb{R}^n$ is as described above. Then there exist $c > 0$ which depend only on the subgaussian distribution involved, δ and a such that dictionary $\mathbf{\Phi}$ satisfies* HRIP(k, δ) *for $k = \lfloor cn/\log N \rfloor$ with probability $\geq 1 - 3N^{-a}$*

Proof. It is known, see e.g. [11], that such matrices (dictionaries) satisfy the concentration of measure property of the form: there is $c_0 > 0$ such that for each $1 \geq \epsilon > 0$ for any $x \in \mathbb{R}^N$ we have

$$\mathbb{P}\Big(\big|\|\sum_{j=1}^{N} x_j\phi_j\|^2 - \|x\|^2\big| > \epsilon\|x\|^2\Big) \leq 2e^{-nc_0\epsilon^2}. \tag{2.3}$$

Then Lemma 5.1 from [1] says that for any fixed set $T \subset \{1, \dots, N\}$ with $\#T = l$ the inequality (2.1) fails with probability $\leq 2(12/\delta)^l e^{-c_0(\delta/2)^2 n}$. Since there are $\binom{N}{l} < \big(en/l\big)^l$ such subsets we see that (2.1) fails for *all* sets T with $\#T = l$ with probability

$$\leq 2\left(\frac{eN}{l}\right)^l \left(\frac{12}{\delta}\right)^l e^{-c_0\delta^2 n/4} \tag{2.4}$$

so (2.2) fails for all sets T with $\#T = l$ with probability

$$\begin{aligned}
&\leq\ 2\big(\frac{eN}{l}\big)^l\big(\frac{12\sqrt{k}}{\delta\sqrt{l}}\big)^l e^{-c_0\delta^2 ln/(4k)}\\
&=\ 2\exp\Big[\big(l(\ln eN + \ln 12 + \ln(1/\delta) + \tfrac{1}{2}\ln(k/l)\big) - l\ln l - c_0\delta^2\tfrac{ln}{4k}\Big]\\
&\leq\ 2\exp\big(\gamma l\ln N - c_0\delta^2\tfrac{ln}{4k}\big)
\end{aligned}$$

where $\gamma > 0$ is a constant depending on δ. Now we set

$$k = \left\lfloor \tfrac{c_0\delta^2}{\gamma\mu} \cdot \frac{n}{\ln N} \right\rfloor \tag{2.5}$$

where $\mu = 4(1 + a/\gamma)$. We continue our estimates to get

$$\leq 2\exp\big(\gamma(1 - \tfrac{\mu}{4})l\ln N\big) = 2\exp - al\ln N = 2N^{-al}. \tag{2.6}$$

Summing over $l = 1, 2, \dots$ we get that HRIP(k, δ) fails with probability at most $2\sum_{l=1}^{\infty} N^{-al} \leq \frac{2}{N^a - 1}$ which implies the Theorem. $\square$

3 Main results

We prove the following theorem, which is a RIP analogue of Theorem 1.3 from [6]:

Theorem 3.1. *Assume that the dictionary* $\mathbf{\Phi}$ *satisfies* RIP$(2S, \varepsilon)$ *and* $0 \leq k < S$. *Then*

$$\|f_S\|^2 \leq 2\|f_k\| \Big(\sigma_{S-k}(f_k) + 4\varepsilon\big(2 + \lceil \log_2 S \rceil\big)\|f_k\|\Big). \tag{3.1}$$

Note that in particular seting $k = 0$ we get

$$\|f_S\|^2 \leq C\|f\|(\sigma_S(f) + A\epsilon\|f\|). \tag{3.2}$$

To prove this theorem we require the following proposition.

Proposition 3.2. *Let* $0 < \varepsilon < 1$ *and* $A = [a_{i,j}]$ *be an* $n \times n$ *upper triangular matrix such that for any* $x \in \mathbb{R}^n$

$$(1 - \varepsilon)\|x\| \leq \|Ax\| \leq (1 + \varepsilon)\|x\| \tag{3.3}$$

and $|a_{i,i}| \geq 1 - \varepsilon$ *for* $i = 1, \dots, n$. *Let* $i_1, i_2, \dots, i_n \in \{0, 1, \dots, n\}$ *be such that*

$$i_{j+1} \geq i_j > j \text{ for } j = 1, 2, \dots, n-1 \qquad \text{and} \qquad i_n < n.$$

Let $B = [b_{i,j}]$ *be another* $n \times n$ *matrix, with*

$$b_{i,j} = \begin{cases} a_{i,j} & \text{if } 1 \leq i \leq i_j \\ 0 & \text{otherwise.} \end{cases}.$$

Then $\|B\| \leq 4\varepsilon\lceil \log_2 n \rceil$.

The idea of the proof is to cut matrix B into rectangular pieces. In this we follow [10]. The heart of the proof of Proposition 3.2 is the following Lemma

Lemma 3.3. *Let* A *be an* $n \times n$ *matrix as in Proposition 3.2. Let* $1 < r < n$ *and* A_1 *and* A_2 *be respectively* $r \times r$ *and* $(n - r) \times (n - r)$ *upper diagonal matrices such that*

$$A = \begin{bmatrix} A_1 & C \\ 0 & A_2 \end{bmatrix}. \tag{3.4}$$

Then A_1 *and* A_2 *satisfy* (3.3) *and* $\|C\| \leq 4\varepsilon$.

Proof. For $y \in \mathbb{R}^r$ and $x = \begin{bmatrix} y \\ 0 \end{bmatrix} \in \mathbb{R}^n$ we have $\|Ax\| = \|A_1 y\|$. Hence, for any $y \in \mathbb{R}^r$ the matrix A_1 satisfies:

$$(1-\varepsilon)\|y\| \le \|A_1 y\| \le (1+\varepsilon)\|y\|. \tag{3.5}$$

Because the inequality (3.3) is also satisfied if A is replaced by A^H, analogous argument gives that the same estimates hold for A_2.

We now estimate $\|C\|$. Clearly $\|C\| \le \|A\| < 2$ so we need to consider only $\varepsilon < \frac{1}{2}$. Let $x \in \mathbb{R}^{n-r}$ be such that $\|Cx\| = \|C\|$ and $\|x\| = 1$. From (3.5) it follows that A_1 is onto, so there exists $y \in \mathbb{R}^r$ such that $\|y\| = 1$ and $A_1 y = \lambda Cx$ for some $\lambda > 0$. Therefore $\|A_1 y + Cx\| = \|A_1 y\| + \|Cx\|$. Let $z = \begin{bmatrix} y \\ x \end{bmatrix} \in \mathbb{R}^n$. Then $\|z\|^2 = 2$ and $Az = \begin{bmatrix} A_1 y + Cx \\ A_2 x \end{bmatrix}$. Hence

$$\begin{aligned}
2(1+\varepsilon)^2 \ge \|Az\|^2 &= \|A_1 y + Cx\|^2 + \|A_2 x\|^2 \\
&= (\|A_1 y\| + \|Cx\|)^2 + \|A_2 x\|^2 \\
&\ge (1-\varepsilon)^2 + ((1-\varepsilon) + \|C\|)^2 \\
&= 2(1-\varepsilon)^2 + 2(1-\varepsilon)\|C\| + \|C\|^2.
\end{aligned}$$

Solving this inequality for $\|C\|$ we obtain $\|C\| \le 4\varepsilon$. □

Proof of Proposition 3.2. We first prove the proposition for $n = 2^m$. For $k = 1, 2, \ldots, n-1$ we fix $r = 0, 1, \ldots, m-1$ such that $2^r \le k < 2^{r+1}$ and define

$$j_k = 2^{m-r-1}(2(k-2^r)+1)+1.$$

Let C_k be the matrix obtained from A by setting to 0 all the coefficients except those at the intersections of columns $j_k, j_k+1, \ldots, j_k + 2^{m-r-1}$ with rows $1, 2, \ldots, i_{j_k}$. We have $\|C_k\| \le 4\varepsilon$.

Now let $D = [d_{i,j}]$ and $E = [e_{i,j}]$ be two matrices obtained from A by setting some of the coefficients to 0. We define $D \setminus E = [f_{i,j}]$ as the matrix obtained from A by setting to 0 all coefficients except those which are non-zero in D and equal to zero in E, i.e.

$$f_{i,j} = \begin{cases} a_{i,j} & \text{if } d_{i,j} \ne 0 \text{ and } e_{i,j} = 0 \\ 0 & \text{otherwise.} \end{cases}$$

For $r = 0, 1, \ldots, m-1$ we now define

$$B_r = \left(\sum_{k=1}^{2^{r+1}-1} C_k \right) \setminus \left(\sum_{k=1}^{2^r-1} C_k \right).$$

We show that $\|B_r\| \le 4\varepsilon$. Let $D_l = C_l \setminus \left(\sum_{k=1}^{l-1} C_l\right)$. Because $\|C_l\| \le 4\varepsilon$ and D_l is obtained from C_l by setting some rows to 0, we have $\|D_l\| \le 4\varepsilon$. Observe that $B_r = \sum_{l=2^r}^{2^{r+1}-1} B_l$ and each of the matrices $D_{2^r}, D_{2^r+1}, \dots, D_{2^{r+1}-1}$ has non-zero coefficients in different rows and columns. Hence

$$\|B_r\| \le \max(\|D_{2^r}\|, \|D_{2^r+1}\|, \dots, \|D_{2^{r+1}-1}\|) \le 4\varepsilon.$$

Because $B = B_0 + B_1 + \cdots + B_{m-1}$ we get $\|B\| \le m \cdot 4\varepsilon = 4\varepsilon \cdot \log_2 n$.

We deal with the situation when $n \neq 2^m$ in the following way: let $m = \lceil \log_2 n \rceil$. We extend the matrix A to a $2^m \times 2^m$ matrix $A' = [a_{i,j}]_{i,j=1}^{2^m}$ by defining

$$a_{i,j} = \begin{cases} 1 & \text{for } n+1 \le i = j \le 2^m \\ 0 & \text{for } n+1 \le i \le 2^m \text{ or } n+1 \le j \le 2^m. \end{cases}$$

For $j = n+1, \dots, 2^m$ we define $i_j = j-1$. The matrix A' satisfies the assumptions of the lemma and the matrix B' obtained from A' satisfies $\|B'\| \le 4\varepsilon \cdot m$. Because B is a sub-matrix of B', we have $\|B\| \le \|B'\| \le 4\varepsilon \cdot \lceil \log_2 n \rceil$. The proof of the lemma is complete. □

Proof of Theorem 3.1. We assume that $f_k \neq 0$. Otherwise $f_S = 0$ as well and the inequality (3.1) is trivially satisfied.

For a given closed subspace $U \subset \mathcal{H}$ let P_U be the orthogonal projection onto U. Let $\phi_1, \phi_2, \dots, \phi_S \in \mathbf{\Phi}$ be the distinct elements returned by the first S iterations of the OMP when applied to f. For $U_\nu = \operatorname{span}(\phi_1, \dots, \phi_\nu)$ and $k \le \nu \le S$ we have

$$f_\nu = f - P_{U_\nu} f = f_k - P_{U_\nu} f_k \tag{3.6}$$

as well as $\langle f_k, \phi_j \rangle = 0$ for $j \in \{1, \dots, k\}$.

For $f \in \mathcal{H}$ let

$$d(f) = \sup_{g \in \mathbf{\Phi}} |\langle f, g \rangle|.$$

Let us fix $\psi \in U_\nu$ with $\|\psi\| = 1$ and $\psi \perp U_{\nu-1}$. Then $\|f_{\nu-1}\|^2 = \|f_\nu\|^2 + \langle f_{\nu-1}, \psi \rangle^2$. Since $d(f_{\nu-1}) = |\langle f_{\nu-1}, \phi_\nu \rangle|$, $\|\phi_\nu\| \le 1 + \varepsilon$ and $|\langle f_{\nu-1}, \psi \rangle| \ge |\langle f_{\nu-1}, \|\phi_\nu\|^{-1} \phi_\nu \rangle|$ we get

$$\|f_\nu\|^2 \le \|f_{\nu-1}\|^2 - (1+\varepsilon)^{-2} d(f_{\nu-1})^2.$$

Repeating this we obtain

$$\|f_S\|^2 \le \|f_k\|^2 - (1+\varepsilon)^{-2} \sum_{\nu=k+1}^{S} d(f_\nu)^2$$

This implies

$$\|f_S\|^2 \leq 2\,\|f_k\| \left(\|f_k\| - (1+\varepsilon)^{-1} \left(\sum_{\nu=k+1}^{S} d(f_\nu)^2 \right)^{1/2} \right). \tag{3.7}$$

We will now provide a lower estimate for $\left(\sum_{\nu=k+1}^{S} d(f_\nu)^2\right)^{1/2}$.

Let $g_1, \dots, g_{S-k} \in \mathbf{\Phi}$ be distinct elements which have the biggest scalar products with f_k, i.e.

$$|\langle f_k, g_1\rangle| \geq |\langle f_k, g_2\rangle| \geq \dots \geq |\langle f_k, g_{S-k}\rangle| \geq \sup\{|\langle f_k, g\rangle| : \phi \in \mathbf{\Phi}, \phi \neq g_i\}.$$

and each g_i, $i \in \{1, \dots, S-k\}$, is different from all ϕ_j, $j \in \{1, \dots, k\}$. Because $f_k \neq 0$, we have $d(f_k) = |\langle f_k, g_1\rangle| > 0$. Observe also that $g_1 = \phi_{k+1}$. We will need also another enumeration of g_i's that will allow us to apply roposition 3.2. To do this we show that there exists a bijective mapping $\pi : \{k+1, \dots, S\} \to \{1, \dots, S-k\}$ such that

$$\text{if } g_{\pi(\nu)} = \phi_j \text{ then } j > \nu \qquad \text{for } \nu = k, k+1, \dots, S-1. \tag{3.8}$$

Let $A = \{g_1, \dots, g_{S-k}\} \cap \{\phi_{k+1}, \dots, \phi_{S-1}\} = \{\phi_{j_1}, \dots, \phi_{j_r}\}$. We assume that

$$k+1 = j_1 < j_2 < \dots < j_r.$$

Define $\pi(k+\mu) = j_{\mu+1}$ for $\mu = 0, \dots, r-1$. The set $\{g_1, \dots, g_{S-k}\} \setminus A$ is exhausted in an arbitrary way by $g_{\pi(k+r)}, \dots, g_{\pi(S-1)}$. Now the property (3.8) follows from the fact that $g_{\pi(k)} = \phi_{k+1}$ and the ordering of $j_1, \dots, j_r$.

By the definition of $d(f_\nu)$ we have $d(f_\nu) \geq |\langle f_\nu, g_{\pi(\nu)}\rangle|$ and by (3.6) $\langle f_\nu, g_{\pi(\nu)}\rangle = \langle f_k, g_{\pi(\nu)}\rangle - \langle P_{U_\nu} f_k, g_{\pi(\nu)}\rangle$.

Let us define

$$a_\nu = \overline{\langle f_k, g_{\pi(\nu)}\rangle} \cdot \left(\sum_{\nu=k+1}^{S} |\langle f_k, g_{\pi(\nu)}\rangle|^2 \right)^{-1/2}. \tag{3.9}$$

(Note that because $d(f_k) > 0$, the sum $\sum_{\nu=k+1}^{S} |\langle f_k, g_{\pi(\nu)}\rangle|^2$ is positive.) Then $\sum_{\nu=k+1}^{S} |a_\nu|^2 = 1$ and

$$\begin{aligned} \left(\sum_{\nu=k}^{S-1} d(f_\nu)^2 \right)^{1/2} &\geq \left(\sum_{\nu=k+1}^{S} |\langle f_\nu, g_{\pi(\nu)}\rangle|^2 \right)^{1/2} \geq \left| \sum_{\nu=k+1}^{S} a_\nu \langle f_\nu, g_{\pi(\nu)}\rangle \right| \\ &\geq \left| \sum_{\nu=k+1}^{S} a_\nu \langle f_k, g_{\pi(\nu)}\rangle \right| - \left| \sum_{\nu=k+1}^{S} a_\nu \langle P_{U_\nu} f_k, g_{\pi(\nu)}\rangle \right| \\ &= \left(\sum_{i=1}^{S-k} |\langle f_k, g_i\rangle|^2 \right)^{1/2} - \left| \langle f_k, \sum_{\nu=k+1}^{S} a_\nu P_{U_\nu} g_{\pi(\nu)}\rangle \right|. \end{aligned} \tag{3.10}$$

We now estimate

$$\left| \langle f_k, \sum_{\nu=k+1}^{S} a_\nu P_{U_\nu} g_{\pi(\nu)} \rangle \right| \leq \|f_k\| \left\| \sum_{\nu=k+1}^{S} a_\nu P_{U_\nu} g_{\pi(\nu)} \right\|.$$

Now let us consider the system

$$\left\{ \phi_1, \ldots, \phi_S, g_{\pi(r+1)}, \ldots, g_{\pi(S-k)} \right\} \tag{3.11}$$

in this particular order. Since this system consists of elements from $\mathbf{\Phi}$ we will denote it as $\{\phi_j\}_{j=1}^{R}$ with $R = 2S - k - r < 2S$. Let $\rho(\nu)$ be such that $g_{\pi(\nu)} = \phi_{\rho(\nu)}$ for $\nu = l+1, \ldots, S$. Observe that the mapping $\nu \mapsto \rho(\nu)$ is increasing and $\rho(\nu) > \nu$.

Let now $\psi_1, \ldots, \psi_R$ be the Gram-Schmidt orthonormalization of the system (3.11). Then

$$\phi_j = \sum_{i=1}^{j} t_{i,j} \psi_i \tag{3.12}$$

and the upper-triangular $R \times R$ matrix $T = [t_{i,j}]$ satisfies the assumptions of Proposition 3.2, which follows from the RIP property of the dictionary $\mathbf{\Phi}$.

Note that we have

$$P_{U_\nu} g_{\pi(\nu)} = P_{U_\nu} \phi_{\rho(\nu)} = \sum_{i=1}^{\nu} t_{i,\rho(\nu)} \psi_i.$$

For each column index $j \in \{1, 2, \ldots, R\}$ we define a row index i_j so that $i_{\rho(\nu)} = \nu$ and for $j \notin \{\rho(k+1), \ldots \rho(S)\}$ we choose i_j so that the sequence $(i_j)_{j=1}^{R}$ is non-decreasing and $i_j > i$. Let the matrix $\tilde{B} = [b_{i,j}]$ with $i, j = 1, \ldots, R$ be defined as

$$b_{i,j} = \begin{cases} t_{i,j} & \text{if } 1 \leq i \leq i_j \\ 0 & \text{otherwise.} \end{cases}.$$

By Proposition 3.2

$$\left\| \tilde{B} \right\| \leq 4\varepsilon \cdot \lceil \log_2 R \rceil.$$

Let B_j denote the i-th column of the matrix $\tilde{B}$. Let

$$B = [B_{\rho(k+1)}, \ldots, B_{\rho(S)}].$$

Observe that

$$\|B\| \leq \left\| \tilde{B} \right\| \leq 4\varepsilon \cdot \lceil \log_2 R \rceil \leq 4\varepsilon \cdot \lceil 1 + \log_2 S \rceil.$$

For the vector $a = [a_{k+1}, \dots, a_S]^T$ (defined in (3.9)) we have $\|a\| = 1$ and

$$\left\| \sum_{\nu=k+1}^{S} a_\nu P_{U_\nu} g_{\pi(\nu)} \right\| = \|Ba\| \le \|B\| \, \|a\| \le 4\varepsilon \cdot \lceil \log_2(2S-k) \rceil. \tag{3.13}$$

Next we estimate the term $\left(\sum_{i=1}^{S-k} |\langle f_k, g_i \rangle|^2 \right)^{1/2}$. Let $\eta_1, \dots, \eta_{S-k} \in \mathbf{\Phi}$ be distinct elements such that for $V = \mathrm{span}(\eta_1, \dots, \eta_{S-k})$ we have

$$\sigma_{S-k}(f_k) = \|f_k - P_V f_k\|.$$

Let the scalars $b_1, \dots, b_{S-k}$ be such, that

$$P_V f_k = \sum_{j=1}^{S-k} b_j \eta_j.$$

Observe, that $\|P_V f_k\| \ge \|f_k\| - \sigma_{S-k}(f_k)$, which combined with the RIP gives us

$$\left(\sum_{j=1}^{S-k} |b_j|^2 \right)^{1/2} \ge \frac{1}{1+\varepsilon} (\|f_k\| - \sigma_{S-k}(f_k)). \tag{3.14}$$

Using Proposition 2.1 and RIP we next obtain

$$\begin{aligned} \left(\sum_{j=1}^{S-k} \langle f_k, \eta_j \rangle \right)^{1/2} = \left(\sum_{j=1}^{S-k} \langle P_V f_k, \eta_j \rangle \right)^{1/2} &\ge (1-\varepsilon) \|P_V f_k\| \\ &\ge (1-\varepsilon)^2 \left(\sum_{j=1}^{S-k} |b_j|^2 \right)^{1/2}. \end{aligned} \tag{3.15}$$

From (3.14) and (3.15) we get

$$\left(\sum_{i=1}^{S-k} |\langle f_k, g_i \rangle|^2 \right)^{1/2} \ge \left(\sum_{j=1}^{S-k} |\langle f_k, \eta_j \rangle|^2 \right)^{1/2} \ge \frac{(1-\varepsilon)^2}{1+\varepsilon} (\|f_k\| - \sigma_{S-k}(f_k)). \tag{3.16}$$

From (3.7), (3.10), (3.16), (3.12) and (3.13) we obtain

$$\begin{aligned} \|f_S\|^2 &\le 2 \|f_k\| \left(\left(\tfrac{1-\varepsilon}{1+\varepsilon}\right)^2 \sigma_{S-k}(f_k) + \left(1 - \left(\tfrac{1-\varepsilon}{1+\varepsilon}\right)^2 + 4\varepsilon \lceil 1 + \log_2 S \rceil \right) \|f_k\| \right) \\ &\le 2 \|f_k\| \left(\sigma_{S-k}(f_k) + 4\varepsilon \big(2 + \lceil \log_2 S \rceil\big) \|f_k\| \right). \end{aligned}$$

The proof is complete. $\square$

For dictionaries with coherence J. Tropp [13], slightly improving the estimate from [7], showed

Theorem 3.4. *If the dictionary* $\mathbf{\Phi}$ *has coherence* η *then*

$$\|f_m\| \leq \sqrt{1+6m}\sigma_m(f) \tag{3.17}$$

for $m < (2\eta)^{-1}$.

Using the above theorem we obtain

Theorem 3.5. *Assume that the dictionary* $\mathbf{\Phi}$ *satisfies* HRIP(k,δ). *Then there exists a constant* C_δ *such that for* $m \leq \sqrt{k}/(6\delta)$ *we have*

$$\left\| f_{m\lceil 4\log_2 m - 1\rceil} \right\| \leq C_\delta \sigma_m(f). \tag{3.18}$$

Proof. By HRIP and Proposition 2.2 the dictionary $\mathbf{\Phi}$ has coherence

$$\eta \leq \frac{3\delta}{\sqrt{k}}.$$

We take

$$m \leq \frac{1}{6}\delta^{-1}k^{1/2}, \tag{3.19}$$

so that (3.17) holds. We define $m_l := m(2^l - 1)$ for $l = 1, 2, \dots$. Let us fix $S = ak^\gamma$, where $\gamma \in (\frac{1}{2}, \frac{3}{4})$ and $a \in (0,1)$ is chosen so that S is sufficiently large and integer. By HRIP the dictionary $\mathbf{\Phi}$ satisfies RIP$(2S,\varepsilon)$ with

$$\varepsilon = a^{\frac{1}{2}}\delta k^{-\frac{1-\gamma}{2}}. \tag{3.20}$$

Lemma 3.6. *There exists a constant* $B = B(\delta, a, \gamma)$ *such that*

$$B(\delta, a, \gamma) \leq 2^{\frac{5}{4}+\frac{3}{8\gamma}} 3^{-\frac{1}{4}} a^{\frac{1}{2}} e \cdot \left(2 + \frac{8\gamma}{(4\gamma-3)\ln 2}\right)\delta^{\frac{3}{4}}$$

and

$$4\varepsilon(2 + \lceil \log_2 S \rceil) \leq Bm^{-1/4}. \tag{3.21}$$

Proof. By (3.19) we have $m^{-1/4} \geq 6^{1/4}k^{-1/8}\delta^{1/4}$. Because $S = ak^\gamma$ and ε is given by (3.20), we need

$$B \geq 2^{\frac{7}{4}} 3^{-\frac{1}{4}} a^{\frac{1}{2}} \delta^{3/4} k^{1/8}(2 + \lceil \log_2 S \rceil).$$

A routine calculation shows that

$$2 + \lceil \log_2 S \rceil \leq 3 + \gamma \log_2 k.$$

Hence, it suffices that $B = 2^{7/4}3^{-1/4}\delta^{3/4} \cdot \sup_{k>0} h(k)$, with

$$h(k) = k^{-\frac{3}{8}+\frac{\gamma}{2}}(3 + \gamma \log_2 k), \qquad k > 0.$$

The function h has the maximum value of

$$e \cdot 2^{-\frac{1}{2}+\frac{3}{8\gamma}} \left(2 + \frac{8\gamma}{(4\gamma - 3)\ln 2}\right).$$

$\square$

Using Theorem 3.1, inequality (3.21) and the fact that $\sigma_n(f_k) \le \sigma_{n-k}(f)$ for $k \le n$ we get

$$\|f_{m_l}\|^2 \le 2\left\|f_{m_{l-1}}\right\|\left(\sigma_m(f) + Bm^{-1/4}\left\|f_{m_{l-1}}\right\|\right) \tag{3.22}$$

as long as $m_l \le S$.

If we know that $\left\|f_{m_{l-1}}\right\| \le D_{l-1}m^\gamma\sigma_m(f)$ for $\gamma \ge \frac{1}{4}$, from (3.22) using inequality $\sqrt{1+z} \le 2\sqrt{z}$ for $z \ge 1$ we obtain

$$\|f_{m_l}\| \le 2D_{l-1}B^{1/2}m^{\gamma-\frac{1}{8}}\sigma_m(f). \tag{3.23}$$

Let $D_1 = 7$, so that $(1+6m)^{1/2} \le c_1 m^{1/2}$. From (3.17) and (3.23) we obtain (iteratively for $l = 2, 3, 4$)

$$\|f_{m_4}\| \le 8D_1B^{3/2}m^{1/8}\sigma_m(f). \tag{3.24}$$

Denote $D_4 = 8D_1B^{3/2}$.

If $m^{1/8} < 4BD_4$, then

$$\|f_{m_4}\| \le 4BD_4^2\sigma_m(f), \tag{3.25}$$

which ends the proof, yielding $C_\delta \ge 4BD_4^2$.

From now on we assume that

$$4BD_4m^{-1/8} \le 1. \tag{3.26}$$

Then the following is true:

Lemma 3.7. *For $l \ge 4$ we have*

$$\|f_{m_l}\| \le D_l m^{2^{-l+1}}\sigma_m(f), \tag{3.27}$$

and $D_l \le 4D_4$.

Proof. By (3.25) the lemma holds for $l = 4$. We now proceed by induction. Assume that the lemma holds for some $l \geq 4$. From (3.22) and (3.26) we have

$$\begin{aligned}\left\|f_{m_{l+1}}\right\|^2 &\leq 2D_l m^{2^{-l+1}} \left(1 + BD_l m^{-\frac{1}{4}+2^{-l+1}}\right) \sigma_m(f)^2 \\ &\leq 2D_l m^{2^{-l+1}} \left(1 + 4BD_4 m^{-\frac{1}{8}}\right) \sigma_m(f)^2 \\ &\leq 4D_l m^{2^{-l+1}} \sigma_m(f)^2.\end{aligned}$$

Hence $\|f_{m+l+1}\| \leq 2D_l^{1/2} m^{2^{-l}} \sigma_m(f) = D_{l+1} m^{2^{-l}} \sigma_m(f)$ and $D_{l+1} \leq 2D_l^{1/2} \leq 2(4D_4)^{1/2} \leq 4D_4$. □

We now take $l = l^*$ such that $m^{2^{-l+1}} \leq 2$. A routine calculation shows that it suffices to take $l^* = \lceil \log_2 \log_2 m \rceil + 1$. We then have

$$\left\|f_{m\lceil 4\log_2 m - 1\rceil}\right\| \leq \|f_{m_{l^*}}\| \leq 8D_4 \sigma_m(f).$$

Hence, if (3.19) holds, we can take $C_\delta = 8D_4 = 64 \cdot 7 \cdot B(\delta, a, \gamma)^{3/2}$. □

Clearly, the constants we got in the above argument are far from being optimal.

4 Comments and Remarks

Our results are a contribution to the general problem of comparing $\|f_n\| = \|f - \mathrm{OMP}_n\, f\|$ with $\sigma_n(f)$. There are two main types of inequalities one may seek. One is the inequality of the form

$$\|f_m\| \leq C_m \sigma_m(f) \tag{4.1}$$

where we want the constant C_m to be small–preferably independent of m. Another one is the inequality of the form

$$\|f_{\eta(m)}\| \leq C\sigma_m(f) \tag{4.2}$$

where $\eta(m)$ is certain function of m – preferably not much bigger than m. Clearly the combination of both types is possible. Important factor in such inequalities is the range of m's for which it is valid. Our Theorem 3.1 (and Theorem 1.3 from [6]) provide a tool to pass from inequality (4.1) to inequality (4.2) with $\eta(m) \sim \lfloor m \log m \rfloor$.

The main drawback of Theorem 3.5 is the restriction $m \leq c/\sqrt{k}$. The inspection of the proof shows that it is caused by the analogous restriction

in Theorem 3.4. It is rather unlikely that the range of applicability of this theorem can be significantly improved as it uses only coherence of the dictionary. On the other hand the value $\sqrt{1+6m}$ which appears in Theorem 3.4 is not very essential. Replacing it by m to any fixed power would be sufficient for our argument to work. Thus it seems to be an interesting problem to establish an analogon of Theorem 3.4 that for dictionaries with HRIP. So let us state it as a conjecture:
Conjecture Assume that the dictionary satisfies HRIP(k,δ). There exist constants C, c, α and β (possibly depending on δ) such that for every f and for $m\log^\alpha m \le ck$ we have

$$\|f_{\lfloor m\log^\alpha m\rfloor}\| \le Cm^\beta \sigma_m(f).$$

Especially interesting would be to have $\alpha = 0$. This however may require some restrictions on m. We have the following Proposition to support this claim

Proposition 4.1. *For each $0<\epsilon<1$ and $n=1,2,\dots$ there exists a dictionary satisfying* RIP$(2n,\epsilon)$*, having coherence $\le \frac{1}{\sqrt{n}}$ and a vector x such that $\sigma_n(x)=0$ but $x-\mathrm{OMP}_k\, x\neq 0$ for $k<n+\epsilon^2\sqrt{n}$*

Take $x=(\frac{1}{\sqrt{n}},\dots,\frac{1}{\sqrt{n}},0,\dots,0)\in\mathbb{R}^{2n}$ with n square roots, i.e. $\|x\|=1$. Let us consider the dictionary: $e_1,\dots,e_n$ plus $\psi_j = e_j+\frac{\beta}{\sqrt{n}}x$ for $j=n+1,\dots,n+s$ plus orthonormal vectors which are orthonormal to all those to make a basis in $\mathbb{R}^{2n}$. We assume $\beta>1$.

The coherence is $\le \frac{\max(2,\beta^2)}{n}$. We calculate scalar products of different vectors. $\langle\psi_j,\psi_l\rangle = \frac{\beta^2}{n}$ while $\langle e_j,\psi_l\rangle = \frac{2}{n}$. All other scalar products are zero.

For $l\le s$ let us calculate:

$$\begin{aligned}
\|\sum_{j=1}^n a_j e_j + \sum_{j=1}^l b_j\psi_{n+j}\| &= \|\sum_{j=1}^n a_j e_j + \sum_{j=1}^l b_j e_{n+j} + \frac{\beta}{\sqrt{n}}(\sum_{j=1}^l b_j)\cdot x\| \\
&\le \sqrt{\sum a_j^2 + \sum b_j^2} + \frac{\beta}{\sqrt{n}}\sum_{j=1}^l |b_j| \\
&\le \sqrt{\sum a_j^2 + \sum b_j^2} + \sqrt{l}\frac{\beta}{\sqrt{n}}\sqrt{\sum_{j=1}^l |b_j|^2} \\
&\le \left(1+\sqrt{l}\frac{\beta}{\sqrt{n}}\right)\sqrt{\sum a_j^2 + \sum b_j^2}.
\end{aligned}$$

To estimate from below we get

$$\begin{aligned}
\|\sum_{j=1}^{n} a_j e_j + \sum_{j=1}^{l} b_j \psi_{n+j}\| &= \|\sum_{j=1}^{n} a_j e_j + \sum_{j=1}^{l} b_j e_{n+j} + \frac{\beta}{\sqrt{n}}(\sum_{j=1}^{l} b_j)\cdot x\| \\
&\geq \sqrt{\sum a_j^2 + \sum b_j^2} - \frac{\beta}{\sqrt{n}} \sum_{j=1}^{s} |b_j| \\
&\geq \sqrt{\sum a_j^2 + \sum b_j^2} - \sqrt{l}\frac{\beta}{\sqrt{n}} \sqrt{\sum_{j=1}^{s} |b_j|^2 + \sum_{j} |a_j|^2} \\
&\geq \left(1 - \sqrt{l}\frac{\beta}{\sqrt{n}}\right) \sqrt{\sum a_j^2 + \sum b_j^2}.
\end{aligned}$$

This shows that for any $\mu \leq 2n$ our dictionary has $\mathrm{RIP}(\mu, \beta\sqrt{\min(s,\mu)/n})$.

Now let us see how OMP acts for vector x. Clearly $\langle x, e_j\rangle = \frac{1}{\sqrt{n}}$ and $\langle x, \psi_j\rangle = \frac{\beta}{\sqrt{n}}$. Note that $\|\psi_j\| > 1$ and other elements from the dictionary have norm one. To avoid undue preference for ψ_j's we may normalise them. If we not do this we will be choosing ψ_j's longer. This normalisation introduces the factor $\sqrt{\frac{n}{n+\beta^2}}$ into the second scalar product. But

$$\frac{\beta}{\sqrt{n}}\sqrt{\frac{n}{n+\beta^2}} > \frac{1}{\sqrt{n}}$$

for $\beta > \sqrt{\frac{n}{n-1}}$ so for such β we choose ψ_{j_1} first. After the first step of OMP we get

$$\begin{aligned}
x - \langle x, \psi_{j_1}\rangle \psi_{j_1} \frac{1}{\|\psi_{j_1}\|^2} &= x - \frac{\beta}{\sqrt{n}(1+\beta^2 n^{-1})}(e_{j_1} + \frac{\beta}{\sqrt{n}}) \\
&= -\frac{\beta}{\sqrt{n}(1+\beta^2 n^{-1})} e_{j_1} + (\frac{n}{n+\beta^2})x
\end{aligned}$$

Note that if in the second step we get ψ_{j_2} in the corresponding sum vector x will appear with multiple $\left(\frac{n}{n+\beta^2}\right)^2$ etc. This means that

$$x - \mathrm{OMP}_l\, x = \sum_{\mu=1}^{l} a_\mu e_{j_\mu} + (\frac{n}{n+\beta^2})^l x. \tag{4.3}$$

From this we infer that if we look at next scalar products e_j's will give $\frac{1}{\sqrt{n}}$ while ψ_j's after normalisation will give

$$\frac{n}{n+\beta^2}\sqrt{\frac{n}{n+\beta^2}}\frac{\beta}{\sqrt{n}}$$

So we will be getting ψ_j's as long as

$$\left(\frac{n}{n+\beta^2}\right)^l \sqrt{\frac{n}{n+\beta^2}} \frac{\beta}{\sqrt{n}} > \frac{1}{\sqrt{n}} \tag{4.4}$$

Proof of Proposition 4.1. Let us fix $\beta = \sqrt[4]{n}$. This gives coherence $\leq \frac{1}{\sqrt{n}}$ and RIP$(2n, \epsilon)$ as long as $s \leq \epsilon^2 \sqrt{n}$. Substituting β into (4.4) we infer that we will be getting ψ_j's for first l steps of OMP as long as

$$\left(\frac{n}{n+\sqrt{n}}\right)^{l+1/2} > \frac{1}{\sqrt[4]{n}}.$$

Inverting and taking ln we get

$$\frac{1}{4}\ln n > (l+1/2)\ln(1+\frac{1}{\sqrt{n}})$$

Since $\ln(1+\frac{1}{\sqrt{n}}) \leq \frac{1}{\sqrt{n}}$ we get $l \leq \frac{1}{4}\sqrt{n}\ln n$. Since $s \leq \epsilon^2\sqrt{n}$ we infer that first we choose all ψ_j's, and only then we start picking e_j's. $\square$

References

[1] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. A simple proof of the restricted isometry property for random matrices. *Constr. Approx.*, 28(3):253–263, 2008.

[2] E. J. Candés and T. Tao. Decoding by linear programming. *IEEE Trans. Inform. Theory*, 51(12):4203–4215, 2005.

[3] O. Christensen. *An introduction to frames and Riesz bases.* Applied and Numerical Harmonic Analysis. Birkhäuser Boston Inc., Boston, MA, 2003.

[4] A. Cohen, W. Dahmen, and R. DeVore. Compressed sensing and best k-term approximation. *J. Amer. Math. Soc.*, 22(1):211–231, 2009.

[5] D. L. Donoho. Compressed sensing. *IEEE Trans. Inform. Theory*, 52(4):1289–1306, 2006.

[6] D. L. Donoho, M. Elad, and V. N. Temlyakov. On Lebesgue-type inequalities for greedy approximation. *J. Approx. Theory*, 147(2):185–195, 2007.

[7] A. C. Gilbert, S. Muthukrishnan, and M. J. Strauss. Approximation of functions over redundant dictionaries using coherence. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms (Baltimore, MD, 2003)*, pages 243–252, New York, 2003. ACM.

[8] P. J. Huber. Projection pursuit. *Ann. Statist.*, 13(2):435–525, 1985. With discussion.

[9] L. K. Jones. On a conjecture of Huber concerning the convergence of projection pursuit regression. *Ann. Statist.*, 15(2):880–882, 1987.

[10] S. Kwapień and A. Pełczyński. The main triangle projection in matrix spaces and its applications. *Studia Math.*, 34:43–68, 1970.

[11] S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann. Uniform uncertainty principle for Bernoulli and subgaussian ensembles. *Constr. Approx.*, 28(3):277–289, 2008.

[12] V. N. Temlyakov. Greedy approximation. *Acta Numer.*, 17:235–409, 2008.

[13] J. A. Tropp. Greed is good: algorithmic results for sparse approximation. *IEEE Trans. Inform. Theory*, 50(10):2231–2242, 2004.