Orthogonal Matching Pursuit with random dictionaries

P. Bechler, and P. Wojtaszczyk

Institut of Applied Mathematics, University of Warsaw
P.Bechler@mimuw.edu.pl, wojtaszczyk@mimuw.edu.pl

Abstract:

In this paper we investigatet the efficiency of the Orthogonal Matching Pursuit for random dictionaries. We concentrate on dictionaries satisfying Restricted Isometry Property. We introduce a stronger Homogenous Restricted Isometry Property which is satisfied with overwhelming probability for random dictionaries used in compressed sensing. We also present and discuss some open problems about OMP.

1. Introduction

In this paper we investigate the efficiency of the Orthogonal Matchin$T = U\sqrt{T^*T}$g Pursuit for random dictionaries. Orthogonal Matching Pursuit is a well known greedy algorithm widely used in approximation theory, statistical estimations and compressed sensing (for the review of greedy algorithms see [6]). One of its main features is that it can be applied for arbitrary dictionary. However the efficiency of the algorithm depend very strongly on properties of the dictionary. We work in the context of a Hilbert space $\mathcal{H}$ (which may be assumed to be finite dimensional). The dictionary is a subset $\mathcal{D} \subset \mathcal{H}$ such that $\overline{\operatorname{span}\mathcal{D}} = \mathcal{H}$. We usually assume that $\|x\|$ is close to 1 for $x \in \mathcal{D}$. Generally it is assumed that $\|x\| = 1$ for $x \in \mathcal{D}$ (see e.g. [6]). However for random dictionaries it is very rarely satisfied. On the other hand for such dictionary $\|x\|$ is close to 1 with great probability.
The Orthogonal Matching Pursuit algorithm with respect to the dictionary $\mathcal{D}$ obtains iteratively a sequence $\mathrm{OMP}_n f \in \mathcal{H}$ of approximants of a given element $f \in \mathcal{H}$ and a sequence $d_1, \ldots, d_n \in \mathcal{D}$ in the following way:

- Define $\mathrm{OMP}_0 f = 0$.
- Given $\mathrm{OMP}_{n-1} f$ and $d_1, \ldots, d_{n-1} \in \mathcal{D}$ choose $d_n \in \mathcal{D}$ such that

  $$|\langle f - \underset{n-1}{\mathrm{OMP}} f, d_n\rangle| = \sup\left\{|\langle f - \underset{n-1}{\mathrm{OMP}} f, d\rangle \ : d \in \mathcal{D}\right\}$$

  and define $\mathrm{OMP}_n f$ as the orthogonal projection of f onto $\operatorname{span}\{d_1, \ldots, d_n\}$.

Generally we will write $f - \mathrm{OMP}_s f := f_s$.
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 $\mathcal{D}$) as

$$\Sigma_m^{\mathcal{D}} = \Sigma_m = \{\sum_{j=1}^m a_j d_j \ : \ \{d_j\}_{j=1}^m \subset \mathcal{D}\}. \tag{1}$$

For a given $f \in \mathcal{H}$ we define its best error of m–term approximation as

$$\sigma_m(f) = \inf\{\|f - z\| \ : \ z \in \Sigma_m\}. \tag{2}$$

Clearly we always have $\sigma_m(f) \le \|f - \mathrm{OMP}_m(f)\| = \|f_m\|$.
Obviously when our dictionary is an orthonormal basis then $\sigma_m(f) = \|f - \mathrm{OMP}_m(f)\|$ 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.

2. Dictionaries

One of the commonly used quantitative parameters of the dictionary is its mutual coherence. It is defined as

$$\eta = \sup_{d_1 \neq d_2 \in \mathcal{D}} |\langle d_1, d_2\rangle|. \tag{3}$$

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. [1, 2]).

Definition 1 *The dictionary $\mathbf{\Phi} = \{\phi_j\}_{j=1}^N$ has $\mathrm{RIP}(k, \delta)$, $0 < \delta < 1$ if for any set $T \subset \{1, \ldots, N\}$ with $\#T = k$ and any sequence of numbers x_j we have*

$$(1 - \delta)\sqrt{\sum_{j \in T} |x_j|^2} \le \|\sum_{j \in T} x_j \phi_j\| \le (1 + \delta)\sqrt{\sum_{j \in T} |x_j|^2}. \tag{4}$$

There are some easy relations between those notions. If the dictionary $\mathcal{D}$ has mutual coherence η then it satisfies $\mathrm{RIP}(k, 1-\eta)$ for $k < \eta^{-1}$. On the other hand if $\mathcal{D}$ satisfies $\mathrm{RIP}(k, \delta)$ then it has mutual coherence $\sim \delta$.
Usually dictionaries with RIP are exhibited as random dictionaries. To be more precise we define a dictionary in $\mathbb{R}^n$

as $\mathbf{\Phi}(\omega) = \{\phi_j\}_{j=1}^N$ where $\phi_j = (\gamma_{j,1}, \ldots, \gamma_{j,n})$ and $\gamma_{j,i}$ are idependent copies of a fixed subgaussian random variable normalised so that $\mathbb{E}\|\phi_k\|^2 = 1$.
In this context it is known (see e.g. [1]) that for a fixed $0 < \delta < 1$ there exists $c > 0$ such that the dictionary $\mathbf{\Phi}(\omega)$ with overwhelming probability satisfies $\mathrm{RIP}(k, \delta)$ with $k = \lfloor cn/\log N \rfloor$. On the other hand it is also known that such a dictionary with overwhelming probability has mutual coherence of order $k^{-1/2}$. It is clear that when we have two events each of them happening with big probability then they happen simultanously with big probability. This leads to the following definition:

Definition 2 *The dictionary $\mathbf{\Phi}$ has homogenous restricted isometry property* $\mathrm{HRIP}(k, \delta)$*,* $0 < \delta < 1$ *if for any* $l \leq k$ *it satisfies* $\mathrm{RIP}(l, \delta\sqrt{l/k})$*.*

Following standard reasoning we obtain

Theorem 1 *Suppose that integers n, N and numbers $0 < \delta < 1$ and $a > 0$ are given and suppose that the random dictionary $\mathbf{\Phi}(\omega) = \{\phi_1, \ldots, \phi_N\} \subset \mathbb{R}^n$ is as described above. Then there exist $c, c_1 > 0$ which depend only on the subgaussian distribution involved, δ and a such that dictionary $\mathbf{\Phi}(\omega)$ satisfies* $\mathrm{HRIP}(k, \delta)$ *for* $k = \lfloor c_1 n/\log N \rfloor$ *with probability* $\geq 1 - 3N^{-a}$

Basically this tells us that unless we are very unlucky a randomly chosen dictionary satisfies HRIP, which is clearly stronger property than RIP. We believe that HRIP is a useful property. Theorem 4 is some indication of this.

3. Main Result

Now we want to present a result on the approximation power of OMP for dictionaries satisfying RIP. For dictionaries with incoherence analogous results were obtained by D. Donoho, M. Elad and N.V. Temlyakov [3]. If we are interested in random dictionaries results from [3] require $S \leq \sqrt{n/\log N}$ while ours apply for the full range $S \leq n/\log N$.

Theorem 2 *There exist constants C and c depending only on $\epsilon > 0$ such that for the dictionary $\mathbf{\Phi}$ satisfying* $\mathrm{RIP}(2K, \epsilon)$ *and for* $0 \leq k \leq S \leq K$ *we have*

$$\|f_S\|^2 \leq C\|f_k\| \left(\sigma_{S-k}(f_k) + A\epsilon\|f_k\|\right). \qquad (5)$$

with $A = c(1 + \log K))$*.*

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

$$\|f_S\|^2 \leq C\|f\|(\sigma_S(f) + A\epsilon\|f\|). \qquad (6)$$

The proof of Theorem 2 is rather complicated. It uses a lot of geometry of Hilbert space, theory of Riesz bases and ideas from [3] and [5]. The main new technical tool is the following lemma on norm of matrices.

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

$$(1 - \epsilon)\|x\| \leq \|Ax\| \leq (1 + \epsilon)\|x\| \qquad (7)$$

and $|a_{i,i}| \geq 1 - \epsilon$ for $i = 1, 2, \ldots, n$. Let $B = [b_{i,j}]$ be the off diagonal part of A i.e.

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

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

The above inequailties (5) and (6) have some merit only if $\epsilon A < 1$. Generally one would like to avoid the presence of $\|f_k\|$ (or $\|f\|$) inside the brackets in (5), (6). The most desirable would be to have direct estimates of the form $\|f_s\| \leq C\sigma_s(f)$. Unfortunatelly in full generality such estimates are not true even when we replace the constant by a function of s.
Here is an appropriate example. Let $x = \frac{1}{\sqrt{n}}\sum_{j=1}^n e_j \in \mathbb{R}^{2n}$ so $\|x\| = 1$. Let us consider the dictionary consisting of vectors: $e_1, \ldots, e_n$, $\psi_j := \|e_j + \beta n^{-1/2}x\|^{-1}(e_j + \beta n^{-1/2}x)$ for $j = n+1, \ldots, n+s$ plus orthonormal vectors which are orthonormal to all those to make a basis in $\mathbb{R}^{2n}$. We take $\beta = \sqrt[4]{n}$ and $s = \lfloor\epsilon\sqrt{n}\rfloor$. Then the following are easy to check

- The mutual coherence is $\leq n^{-1/2}$.
- The Riesz constant of this basis is $\sqrt{\epsilon}$ so the dictionary has $\mathrm{RIP}(2n, \sqrt{\epsilon})$
- Orthogonal Matching Pursuit for vector x in first s iterations chooses vectors ψ_j and only later chooses vectors e_j.

Thus we see that $\sigma_n(x) = 0$ while $x - \mathrm{OMP}_k(x) \neq 0$ for $k = n + s - 1$.
For dictionaries with mutual coherence η J. Tropp [7], slightly improving estimate from [4], have proved

Theorem 3 *If the dictionary has mutual coherence η then*

$$\|f_m\| \leq 8\sqrt{m}\sigma_m(f) \text{ for } m < (3\eta)^{-1}. \qquad (8)$$

Using this we obtain

Theorem 4 *Let the dictionary $\mathbf{\Phi}$ satisfies* $\mathrm{HRIP}(k, \delta)$*. Then for $m \leq c/\sqrt{k}$ we have*

$$\|f_{\lfloor m\log m\rfloor}\| \leq C\sigma_m(f). \qquad (9)$$

Let us give a sketch of a proof which follows arguments from [3]. We start with $m \leq c'\sqrt{k}$ for which (8) holds. We set $m_l = m(2^l - 1)$ and we fix $K \sim k^{3/4}$. Using HRIP we get that dictionary $\mathbf{\Phi}$ satisfies $\mathrm{RIP}(2K, \epsilon)$ with $A\epsilon \leq \delta k^{-1/8} \leq \beta m^{-1/4}$. From Theorem 2 and (8) we get

$$\begin{aligned} \|f_{m_2}\|^2 &\leq C\|f_{m_1}\|(\sigma_m(f) + A\epsilon\|f_{m_1}\|) \\ &\leq C\|f_{m_1}\|(\sigma_m(f) + 8\beta m^{1/4}\sigma_m(f)) \\ &\leq 8Cm^{1/2}(1 + 8\beta m^{1/4})\sigma_m^2(f) \\ &\leq C'm^{3/4}\sigma_m^2(f). \end{aligned}$$

Thus we get $\|f_{m_2}\| \leq \sqrt{C'}m^{3/8}\sigma_m(f)$. Repeating this argument and carefully tracking constants we see that after at most $\mu \sim \log\log m$ steps we get the claim.

Analogous result from [3] uses only mutual coherence and in our case gives (9) for $m \leq c\sqrt[3]{k}$. The main drawback of Theorem 4 is the limitation on m. It is clear from the above sketch that this restriction is inherited from Theorem 3. It is very unlikely that (8) can be substantially improved using only mutual coherence. We believe however that for dictionaries with RIP or HRIP one can prove more. So let us state the following conjecture

Conjecture Assume that the dictionary satisfies $\mathrm{HRIP}(k, \delta)$. There exist constants C, c, α and β (possibly depending on δ) such that for every f and for $m \log^\alpha m \leq ck$ we have

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

Let us note that it follows from Theorem 3 that there exists a function $\psi(k, \delta)$ and constants C and β such that if the dictionary satisfies $\mathrm{HRIP}(k, \delta)$ then for every $f \in \mathcal{H}$

$$\|f_m\| \leq Cm^\beta \sigma_m(f).$$

for $m \leq \psi(k, \delta)$. (Clearly Theorem 3 gives $\beta = 1/2$ and $\psi(k, \delta) \sim \sqrt{k}$). It would be interesting to know if ψ can grow significantly faster than $\sqrt{k}$.

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** (2008), no. 3, 253–263.

[2] E. Candés, *The restricted isometry property and its implications for compressed sensing*, Compte Rendus de l'Academie des Sciences, Paris, Series I, **346**(2008), 589–592.

[3] D.Donoho, M.Elad, V.N.Temlyakov, *On Lebesgue-type inequalities for greedy approximation* J. Approx. Theory **147** (2007), no. 2, 185–195.

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

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

[6] V.N.Temlyakov, *Greedy approximation*, Acta Numerica **17** (2008) 235–409

[7] J.Tropp, *Greed is good: Algorithmic results for sparse approximation*, IEEE Trans. Inform. Theory, **50** (2004), 2231-2242