PREPRINT 426

JUNE 2015

Orthogonal Matching Pursuit under the Restricted Isometry Property *

Albert Cohen, Wolfgang Dahmen and Ronald DeVore

Institut für
Geometrie und Praktische Mathematik
Templergraben 55, 52062 Aachen, Germany

* This research was supported by by the ONR Contracts N00014-11-1-0712, N00014-12-1-0561, N00014-15-1-2181; the NSF Grants DMS 1222715, DMS 0915231, DMS 1222390; the Institut Universitaire de France; the ERC Adv grant BREAD; the DFG SFB-Transregio 40; the DFG Research Group 1779; the Excellence Initiative of the German Federal and State Governments, and RWTH Aachen Distinguished Professorship, Graduate School AICES.

Orthogonal Matching Pursuit under the Restricted Isometry Property *

Albert Cohen, Wolfgang Dahmen, and Ronald DeVore

June 11, 2015

Abstract

This paper is concerned with the performance of Orthogonal Matching Pursuit (OMP) algorithms applied to a dictionary $\mathcal{D}$ in a Hilbert space $\mathcal{H}$. Given an element $f \in \mathcal{H}$, OMP generates a sequence of approximations f_n, $n = 1, 2, \dots$, each of which is a linear combination of n dictionary elements chosen by a greedy criterion. It is studied whether the approximations f_n are in some sense comparable to *best n term approximation* from the dictionary. One important result related to this question is a theorem of Zhang [8] in the context of sparse recovery of finite dimensional signals. This theorem shows that OMP exactly recovers n-sparse signal, whenever the dictionary $\mathcal{D}$ satisfies a Restricted Isometry Property (RIP) of order An for some constant A, and that the procedure is also stable in ℓ^2 under measurement noise. The main contribution of the present paper is to give a structurally simpler proof of Zhang's theorem, formulated in the general context of n term approximation from a dictionary in arbitrary Hilbert spaces $\mathcal{H}$. Namely, it is shown that OMP generates near best n term approximations under a similar RIP condition.

AMS Subject Classification: 94A12, 94A15, 68P30, 41A46, 15A52

Key Words: Orthogonal matching pursuit, best n term approximation, instance optimality, restricted isometry property.

1 Introduction

Approximation by sparse linear combinations of elements from a fixed redundant family is a frequently employed technique in signal processing and other application domains. We consider such problems in a separable Hilbert space $\mathcal{H}$ endowed with a norm $\|\cdot\| := \|\cdot\|_{\mathcal{H}}$ induced by the scalar product $\langle\cdot,\cdot\rangle$ on $\mathcal{H} \times \mathcal{H}$. A countable collection $\mathcal{D} = \{\varphi_\gamma\}_{\gamma\in\Gamma} \subset \mathcal{H}$ is called a *dictionary* if it is complete, i.e., the set of finite linear combinations of elements

*This research was supported by by the ONR Contracts N00014-11-1-0712, N00014-12-1-0561, N00014-15-1-2181; the NSF Grants DMS 1222715, DMS 0915231, DMS 1222390; the Institut Universitaire de France; the ERC Adv grant BREAD; the DFG SFB-Transregio 40; the DFG Research Group 1779; the Excellence Initiative of the German Federal and State Governments, and RWTH Aachen Distinguished Professorship, Graduate School AICES.

of the dictionary are dense in $\mathcal{H}$. The simplest example of a dictionary is the set of elements of a fixed basis of $\mathcal{H}$. But our primary interest is in *redundant* families. In such a case, there exists a strict subset of $\mathcal{D}$ that is still a dictionary. A primary example of a redundant dictionary is a frame, e.g., any union of a finite number of bases. Without loss of generality we shall always assume that the dictionary $\mathcal{D}$ is *normalized*, i.e.,

$$\|\varphi_\gamma\| = 1, \quad \gamma \in \Gamma.$$

Given such a dictionary $\mathcal{D}$, we consider the class

$$\Sigma_n = \Sigma_n(\mathcal{D}) := \Big\{ \sum_{\gamma \in S} c_\gamma \varphi_\gamma \ : \ \#(S) \le n \Big\} \subset \mathcal{H}, \quad n \ge 1. \tag{1.1}$$

The elements in Σ_n are said to be sparse with *sparsity* n. We define

$$\sigma_n(f)_{\mathcal{H}} := \inf_{g \in \Sigma_n} \|f - g\|,$$

which is called *the error of best n-term approximation* to f from the dictionary $\mathcal{D}$.

An important distinction between n term dictionary approximation and other forms of approximation, such as approximation from an n dimensional space, is that the set Σ_n is not a linear space since the sum of two elements in Σ_n is generally not in Σ_n, although it is in Σ_{2n}. Thus n-term approximation from a dictionary is an important example of nonlinear approximation [3] that reaches into numerous application areas such as adaptive PDE solvers, image encoding, or statistical learning. It also serves as a performance benchmark in *compressed sensing* that better captures the robustness of compressed sensing than results on exact sparsity recovery [2].

While there are many themes in n term dictionary approximation, our interest here is in analyzing the performance of greedy algorithms for generating n-term approximations to a given target element $f \in \mathcal{H}$. There are numerous papers on this subject. We refer the reader to the survey article [6] as a general reference. Our particular interest is in understanding what properties of the dictionary $\mathcal{D}$ guarantee that these algorithms perform similarly to best n-term approximation.

These algorithms and best n-term approximation have a simple description when the dictionary $\mathcal{D}$ is an orthonormal or, more generally, a Riesz basis of $\mathcal{H}$. In this case, the best n-term approximations to a given $f \in \mathcal{H}$ are realized by expanding f in terms of the basis

$$f = \sum_{\gamma \in \Gamma} c_\gamma \varphi_\gamma \tag{1.2}$$

and retaining n terms from this expansion which correspond to the largest (in absolute value) expansion coefficients. The typcial greedy algorithm will construct the same approximations. The situation is much less clear when dealing with more general dictionaries.

In the case of general dictionaries, algorithms for generating n-term approximations are typically built on some form of greedy selection

$$\varphi_k := \varphi_{\gamma_k}, \ \ k = 1, 2, \dots, \tag{1.3}$$

of elements from $\mathcal{D}$ and then using a linear combination of $\varphi_1, \ldots, \varphi_n$ as the n-term approximation. The standard greedy algorithm (called the Pure Greedy Algorithm) makes the initial selection φ_1 as any element such that

$$\varphi_1 := \operatorname*{Argmax}_{\varphi \in \mathcal{D}} |\langle f, \varphi \rangle|. \tag{1.4}$$

This gives the approximation $f_1 := \langle f, \varphi_1 \rangle \varphi_1$ to f and the residual $r_1 := f - f_1$. Given that $\varphi_1, \ldots, \varphi_{k-1}$ have been selected, and an approximation f_{k-1} from $F_{k-1} := \operatorname{span}\{\varphi_1, \ldots, \varphi_{k-1}\}$ has been constructed, the next dictionary element φ_k is chosen as the best match of the residual

$$r_{k-1} := f - f_{k-1}, \tag{1.5}$$

in the sense that

$$\varphi_k := \operatorname*{Argmax}_{\gamma \in \Gamma} |\langle r_{k-1}, \varphi_\gamma \rangle|. \tag{1.6}$$

There exist different ways of forming the next approximation f_k resulting in different greedy algorithms. We focus our attention on *Orthogonal Matching Pursuit* (OMP), which forms the new approximation as

$$f_k := P_k f, \tag{1.7}$$

where P_k is the orthogonal projector onto F_k. OMP is also called the Orthogonal Greedy Algorithm. More generally, we analyze the *Weak Orthogonal Matching Pursuit* (WOMP) where the choice of φ_k is only required to satisfy

$$|\langle r_{k-1}, \varphi_k \rangle| \geq \kappa \max_{\gamma \in \Gamma} |\langle r_{k-1}, \varphi_\gamma \rangle|, \tag{1.8}$$

where $\kappa \in]0, 1]$ is a fixed parameter, which is a more easily implemented selection rule in practical applications. Once this choice of $\varphi_1, \ldots, \varphi_k$ is made, then f_k is again defined as the orthogonal projection onto F_k.

The main interest of the present paper is to understand what properties of a dictionary $\mathcal{D}$ guarantee that the approximation rate of WOMP after $O(n)$ steps is comparable to the the best n-term approximation error $\sigma_n(f)$, at least for a certain range $n \leq N$. A related question, but less demanding, is to understand when WOMP is guaranteed to exactly recover f whenever $f \in \Sigma_n$ in $O(n)$ steps for a suitable range of n. This is sometimes refered to as *sparse recovery*. Of course, as already mentioned, we know that both of these questions have a positive answer for the entire range of n whenever $\mathcal{D}$ is a Riesz basis for $\mathcal{H}$.

To give a precise formulation of the type of performance we seek, we define the concept of *instance optimality*.

Instance Optimality: *We say that the WOMP algorithm satisfies instance optimality for $n \leq N$, if there are constants $A, C > 0$, with A an integer, such that the outputs f_n of WOMP satisfy*

$$\|f - f_{An}\| \leq C \sigma_n(f)_{\mathcal{H}}, \tag{1.9}$$

for $n \leq N$.

Notice that if (1.9) is satisfied then it implies a positive solution to the sparse recovery problem for the same range of n since $\sigma_n(f) = 0$ when f is in Σ_n. To obtain results on sparse recovery or instance optimality requires structure on the dictionary $\mathcal{D}$. The first results of this type were obtained under assumptions on the *coherence* of a dictionary $\mathcal{D} \subset \mathcal{H}$ defined by

$$\mu = \mu(\mathcal{D}) := \sup\{|\langle \varphi, \psi \rangle| \ : \ \varphi, \psi \in \mathcal{D}, \ \varphi \neq \psi\}.$$

The first results on this general circle of problems centered on sparse recovery. Tropp [7] proved that whenever the dictionary has coherence $\mu < \frac{1}{2n-1}$, then n steps of OMP recover any $f \in \Sigma_n$ exactly.

Concerning instance optimality, we mention that Livschitz [5] proved that whenever $\mu \leq \frac{1}{20n}$, then after $2n$ steps, the OMP algorithm returns $f_{2n} \in \Sigma_{2n}$ such that

$$\|f - f_{2n}\| \leq 3\sigma_n(f)_{\mathcal{H}}. \tag{1.10}$$

A weaker assumption on a dictionary, known as the *Restricted Isometry Property* (RIP), was introduced in the context of compressed sensing [1]. To formulate this property, we introduce the notation

$$\Phi \mathbf{c} = \sum_{\gamma \in \Gamma} c_\gamma \varphi_\gamma, \tag{1.11}$$

whenever $\mathbf{c} = (c_\gamma)_{\gamma \in \Gamma}$ is a finitely supported sequence. The dictionary $\mathcal{D}$ is said to satisfy the RIP of order $n \in \mathbb{N}$ with constant $0 < \delta < 1$ provided

$$(1-\delta)\|\mathbf{c}\|_{\ell^2}^2 \leq \|\Phi \mathbf{c}\|^2 \leq (1+\delta)\|\mathbf{c}\|_{\ell^2}^2, \quad \|\mathbf{c}\|_{\ell^0} := \#(\operatorname{supp} \mathbf{c}) \leq n. \tag{1.12}$$

Hence this property quantifies the deviation of any subset of cardinality at most n from an orthonormal set. We denote by δ_n the minimal value of δ for which this property holds and remark that trivially $\delta_n \leq \delta_{n+1}$.

It is well-known that a coherence bound

$$\mu(\mathcal{D}) < (n-1)^{-1} \tag{1.13}$$

implies the validity of RIP(n) for $\delta_n \leq (n-1)\mu$, but not vice versa [7].

In [8], Tong Zhang proved that OMP exactly recovers finite dimensional n-sparse signals, whenever the dictionary $\mathcal{D}$ satisfies a Restricted Isometry Property (RIP) of order An for some constant A, and that the procedure is also stable in ℓ^2 under measurement noise. The main result of the present paper is the following related theorem on instance optimality for WOMP.

Theorem 1.1 *Given the weakness parameter $\kappa \leq 1$, there exist fixed constants A, C, δ^*, such that the following holds for all $n \geq 0$: if $\mathcal{D}$ is a dictionary in a Hilbert space $\mathcal{H}$ for which* RIP((A + 1)n) *holds with $\delta_{(A+1)n} \leq \delta^*$, then, for any target function $f \in \mathcal{H}$, the WOMP algorithm returns after An steps an approximation f_{An} to f that satisfies*

$$\|f - f_{An}\| \leq C\sigma_n(f)_{\mathcal{H}}. \tag{1.14}$$

The values of A, C, κ, and δ^* for which the above result holds are coupled. For example, it is possible to have a smaller value of A at the price of a larger value of C or of a smaller value of δ^*. Similarly, a smaller weakness parameter κ can be compensated by increasing A.

While the theorem of [8] is not stated in the above form, it can be used to derive Theorem 1.1 by interpreting the error of best n-term approximation as a measurement noise. In this way, one version of the above result can be derived from [8] for OMP ($\kappa = 1$) with $\delta^* = \frac{1}{3}$ and $A = 30$. Let us mention that Zhang's theorem is also established in [4], with the same proof, but with different constants $\delta^* = \frac{1}{6}$ and $A = 12$.

In what follows, we do not focus on improving the constants, but rather our interest is to provide a conceptually more elementary proof for Theorem 1.1. Namely the proof for [8] and [4] is based on an induction argument which involves an auxiliary greedy algorithm (initialized from a non trivial sparse approximation) in an inner loop. Our proof avoids using this auxiliary step. It is also presented in the framework of a possibly infinite dimensional Hilbert space $\mathcal{H}$. We give the new proof in the following section. We then give some observations that can be derived from Theorem 1.1.

In this paper, we shall sometimes use the notation $\Phi^* v = (\langle v, \varphi_\gamma\rangle)_{\gamma\in\Gamma}$ for any $v \in \mathcal{H}$, and $\mathbf{c}_T$ to denote, for any $\mathbf{c} = (c_\gamma)_{\gamma\in\Gamma}$ and $T \subset \Gamma$, the sequence whose entries coincides with those of $\mathbf{c}$ on T and are 0 otherwise.

2 Proof of Theorem 1.1

In this section, we give a proof of Theorem 1.1. We begin with the following elementary lemma which guarantees the existence of near best n term approximations from a dictionary.

Lemma 2.1 *Let $\mathcal{D}$ be a dictionary in a Hilbert space $\mathcal{H}$ that satisfies RIP$(2n)$. Then,*
(i) *the set Σ_n of all n-term linear combinations from $\mathcal{D}$ is closed in $\mathcal{H}$.*
(ii) *For each $f \in \mathcal{H}$, $\varepsilon > 0$, and $n \geq 1$, there exists a $g \in \Sigma_n$ such that*

$$\|f - g\| \leq (1+\varepsilon)\sigma_n(f)_{\mathcal{H}}. \tag{2.1}$$

Proof: To prove (i), we let $(g^k)_{k\geq 0}$ be a sequence of elements from Σ_n that converges in $\mathcal{H}$ towards some $g \in \mathcal{H}$. We may write

$$g^k = \Phi \mathbf{c}^k = \sum_{\gamma\in\Gamma} c_\gamma^k \varphi_\gamma, \tag{2.2}$$

with $\|\mathbf{c}^k\|_{\ell^0} \leq n$. For any $\varepsilon > 0$, there exists K such that

$$\|g^k - g^l\| \leq \varepsilon, \quad k, l \geq K. \tag{2.3}$$

From RIP$(2n)$, it follows that

$$\|\mathbf{c}^k - \mathbf{c}^l\|_{\ell^2} \leq \frac{\varepsilon}{\sqrt{1-\delta_{2n}}}, \tag{2.4}$$

which shows that the sequence $(\mathbf{c}^k)_{k\geq 0}$ converges in ℓ^2 to some $\mathbf{c} \in \ell^2$. In particular, we find that

$$\lim_{k\to+\infty} c_\gamma^k = c_\gamma, \quad \gamma \in \Gamma. \tag{2.5}$$

If $c_\gamma \neq 0$ for more than n values of γ, we find that $\|\mathbf{c}^k\|_{\ell^0} > n$ for k sufficiently large which is a contradiction. It follows that $g = \sum_{\gamma\in\Gamma} c_\gamma\varphi_\gamma \in \Sigma_n$.

To prove (ii), let $g^k \in \Sigma_n$ be such that $\|g_k - f\| \to \sigma_n(f)_{\mathcal{H}}$. If $\sigma_n(f) > 0$, then $g = g_k$ will satisfy (ii) if k is sufficiently large. On the other hand, if $\sigma_n(f) = 0$, then $g_k \to f$, $k \to \infty$. By (i) $f \in \Sigma_n$ and so we can take $g = f$. □

2.1 Reduction of the residual

Our starting point in proving Theorem 1.1 is the following lemma from [8] which quantifies the reduction of the residuals generated by the WOMP algorithm under the RIP condition. In what follows, we denote by

$$S_k := \{\gamma_1, \dots, \gamma_k\}, \tag{2.6}$$

the set of indices selected after k steps of WOMP applied to the given target element $f \in \mathcal{H}$, and denote as before the residual by $r_k = f - f_k$.

Lemma 2.2 *Let $(f_k)_{k\geq 0}$ be the sequence of approximations generated by the WOMP algorithm applied to f, and let $g = \Phi\mathbf{z}$ with $\mathbf{z}$ supported on a finite set T. Then, if T is not contained in S_k, one has*

$$\|r_{k+1}\|^2 \leq \|r_k\|^2 - \frac{\kappa^2(1-\delta)}{\#(T \setminus S_k)} \max\{0, \|r_k\|^2 - \|f-g\|^2\}, \tag{2.7}$$

where $\delta := \delta_{\#(T\cup S_k)}$ is the corresponding RIP-constant and $\kappa \in]0,1[$ is the weakness parameter in the WOMP algorithm.

For completeness, we recall the proof at the end of this section. It is at this point, we depart from the arguments in [8] with the goal of providing a simpler more transparent argument. An immediate consequence of Lemma 2.2 is the following.

Proposition 2.3 *Assume that for a given $A > 0$ and $\delta^* < 1$, $\mathrm{RIP}((\mathrm{A}+1)\mathrm{n})$ holds with $\delta_{(A+1)n} \leq \delta^*$. If $g = \Phi\mathbf{z}$, where $\mathbf{z}$ is supported on a set T such that $\#(T) \leq n$, then for any non-negative integers (j, m, L) such that $\#(T \setminus S_j) \leq m$ and $j + mL \leq An$, one has*

$$\|r_{j+mL}\|^2 \leq e^{-\kappa^2(1-\delta^*)L}\|r_j\|^2 + \|f-g\|^2. \tag{2.8}$$

Proof: By Lemma 2.2, if $g = \Phi\mathbf{z}$ where $\mathbf{z}$ is supported on a set T such that $\#(T) \leq n$, then for any non-negative integers (j, m, L) such that $\#(T \setminus S_j) \leq m$ and $j + mL \leq An$, one has

$$\begin{aligned}\max\{0, \|r_{j+mL}\|^2 - \|f-g\|^2\} &\leq \Big(1 - \kappa^2(1-\delta^*)/m\Big)^{mL} \max\{0, \|r_j\|^2 - \|f-g\|^2\} \\ &\leq e^{-\kappa^2(1-\delta^*)L} \max\{0, \|r_j\|^2 - \|f-g\|^2\},\end{aligned}$$

where we have used the fact that $\#(T \setminus S_l) \leq m$ for all $l \geq j$, This gives (2.8) and completes the proof of Proposition 2.3. □

Proof of Theorem 1.1: We fix f and use the abbreviated notation

$$\sigma_n := \sigma_n(f)_{\mathcal{H}}, \quad n \geq 0. \tag{2.9}$$

We first observe that the assertion of the theorem follows from the following.

Claim: *If $0 \leq k < n$ satisfies*

$$\|r_{Ak}\| \leq 2\sigma_k, \tag{2.10}$$

and is such that $\sigma_n < \frac{\sigma_k}{4}$, then there exists $k < k' \leq n$ such that

$$\|r_{Ak'}\| \leq 2\sigma_{k'}. \tag{2.11}$$

Indeed, assuming that this claim holds, we complete the proof of the Theorem as follows. We let k be the largest integer in $\{0, \ldots, n\}$ for which $\|r_{Ak}\| \leq 2\sigma_k$. Since $\|r_0\| = \sigma_0 = \|f\|$, such a k exists. If $k < n$, then we must have $\sigma_k \leq 4\sigma_n$ and therefore

$$\|r_{An}\| \leq \|r_{Ak}\| \leq 2\sigma_k \leq 8\sigma_n, \tag{2.12}$$

so that (1.14) holds with $C = 8$.

We are therefore left with proving the claim. For this, we fix

$$\delta^* = \frac{1}{6}, \tag{2.13}$$

and $0 \leq k < n$ such that (2.10) holds and such that $\sigma_n < \frac{\sigma_k}{4}$. Let $k < K \leq n$ be the first integer such that $\sigma_K < \frac{\sigma_k}{4}$. By (ii) of Lemma 2.1 we know that for any $B > 1$ there is a $g \in \Sigma_K$ with $\|f - g\| \leq B\sigma_K(f)$. Therefore, g has the form

$$g = \Phi \mathbf{z} = \sum_{\gamma \in T} z_\gamma \varphi_\gamma, \quad \#(T) = K. \tag{2.14}$$

The significance of K is that on the one hand

$$\|f - g\| \leq B\sigma_K < \frac{B}{4}\sigma_k, \tag{2.15}$$

while on the other hand

$$\sigma_k \leq 4\sigma_{K-1}. \tag{2.16}$$

To eventually apply Proposition 2.3 for the above g and $j = Ak$, we need to bound $\#(T \setminus S_{Ak})$ with A yet to be specified. To this end, we write $K = k + M$, with $M > 0$, and observe that if $S \subset T$ is any set with $\#(S) = M$ and $g_S := \sum_{\gamma \in S} z_\gamma \varphi_\gamma$, then

$$\|g_S\| \geq \|f - (g - g_S)\| - \|f - g\| \geq \sigma_k - B\sigma_K \geq \Big(1 - \frac{B}{4}\Big)\sigma_k, \tag{2.17}$$

where we have used the fact that $g - g_S \in \Sigma_k$. Using RIP, we obtain the following lower bound for the coefficients of g: for any set $S \subset T$ of cardinality M

$$\Big(1 - \frac{B}{4}\Big)^2 \sigma_k^2 \leq \|g_S\|^2 \leq (1 + \delta^*) \sum_{\gamma \in S} |z_\gamma|^2 = \frac{7}{6} \sum_{\gamma \in S} |z_\gamma|^2. \tag{2.18}$$

Taking for S the set S_g of the M smallest coefficients of g and noting that then for any more general $S \subset T$ with $\#(S) \geq M$, one has $\Big(\sum_{\gamma\in S} |z_\gamma|^2\Big)/\Big(\sum_{\gamma\in S_g} |z_\gamma|^2\Big) \geq \#(S)/M$, and hence

$$\frac{6}{7}\Big(1-\frac{B}{4}\Big)^2\frac{\#(S)}{M}\sigma_k^2 \leq \sum_{\gamma\in S}|z_\gamma|^2. \tag{2.19}$$

For the particular set $S := T \setminus S_{Ak}$, if $\#(S) \geq M$, the above bound combined with the RIP implies

$$\begin{aligned}(1-\delta^*)\tfrac{6}{7}\Big(1-\tfrac{B}{4}\Big)^2\tfrac{\#(S)}{M}\sigma_k^2 &\leq \|g_S\|^2 \leq \|g-f_{Ak}\|^2 \leq (\|g-f\|+\|r_{Ak}\|)^2\\ &\leq (B\sigma_K+2\sigma_k)^2 \leq \Big(\tfrac{B}{4}+2\Big)^2\sigma_k^2.\end{aligned}$$

Since $\delta^* = 1/6$ this gives the bound

$$\#(T\setminus S_{Ak}) \leq \frac{7}{5}\frac{\Big(\frac{B}{4}+2\Big)^2}{\Big(1-\frac{B}{4}\Big)^2}M \leq 13M, \tag{2.20}$$

where the second inequality is obtained by taking B sufficiently close to 1.

We proceed now verifying the claim with $k' = K-1$ when $K-1 > k$ and with $k' = k+1$ otherwise. In the first case we can use the reduction estimate provided by Proposition 2.3 with $j = Ak$ in combination with (2.16) to deal with the term $\|r_{Ak}\|$ in (2.8). When $K = k+1$, however, we cannot bound $\|r_{Ak}\|$ directly in terms of a σ_l for some $l > k$. Accordingly, we use Proposition 2.3 in different ways for the two cases.

In the case where $M \geq 2$, i.e., $K-1 > k$, we apply (2.8) with $j = Ak$, $m = 13M$ and $L = \lceil 4\kappa^{-2}\rceil$. Indeed $Ak + Lm = Ak + 52M \leq An$ holds for $k + M \leq n$ whenever $A \geq 52\kappa^{-2}$. Moreover, notice that for such an A

$$A(K-1) = Ak + A(M-1) \geq Ak + \frac{1}{2}AM = Ak + \frac{Am}{26} = Ak + Lm, \tag{2.21}$$

whenever

$$A \geq 26\lceil 4\kappa^{-2}\rceil. \tag{2.22}$$

This gives

$$\begin{aligned}\|r_{A(K-1)}\|^2 &\leq \|r_{Ak+Lm}\|^2\\ &\leq e^{-10/3}\|r_{Ak}\|^2 + \|f-g\|^2\\ &\leq e^{-10/3}4\sigma_k^2 + B^2\sigma_K^2\\ &\leq e^{-10/3}64\sigma_{K-1}^2 + B^2\sigma_{K-1}^2\\ &\leq 4\sigma_{K-1}^2,\end{aligned}$$

where we have used (2.16) in the fourth inequality, and the last inequality follows by taking B sufficiently close to 1. We thus obtain (2.11) for the value $k' = K-1 > k$.

In the case $M = 1$, i.e., $K = k+1$, we apply (2.8) with $j = Ak$, $m = 13$ and

$$L = \lceil 6\kappa^{-2}\rceil.$$

In fact, from (2.20) we know that $\#(T \setminus S_{Ak}) \leq 13$ and $An \geq A(k+1) \geq Ak + mL$ for A satisfying (2.22). This yields

$$\begin{aligned} \|r_{A(k+1)}\|^2 &\leq \|r_{Ak+mL}\|^2 \\ &\leq e^{-5}\|r_{Ak}\|^2 + \|f-g\|^2 \\ &\leq 4e^{-5}\sigma_k^2 + B^2\sigma_{k+1}^2 \\ &\leq \Big(4e^{-5} + \tfrac{B^2}{16}\Big)\sigma_k^2. \end{aligned}$$

This implies that $S_{A(k+1)}$ contains T. Indeed, if it missed one of the indices $\gamma \in T$, then we infer from the RIP,

$$\begin{aligned} (1-\delta^*)|z_\gamma|^2 &\leq \|g - f_{A(k+1)}\|^2 \\ &\leq (\|f-g\| + \|r_{A(k+1)}\|)^2 \\ &\leq \Big(B\sigma_K + \sqrt{4e^{-5} + \tfrac{B^2}{16}}\sigma_k\Big)^2 \\ &\leq \Big(\tfrac{B}{4} + \sqrt{4e^{-5} + \tfrac{B^2}{16}}\Big)^2\sigma_k^2. \end{aligned}$$

On the other hand, we know from (2.19) that

$$\frac{6}{7}\Big(1 - \frac{B}{4}\Big)^2\sigma_k^2 \leq |z_\gamma|^2, \tag{2.23}$$

which for B sufficiently close to 1 is a contradiction since $\frac{6}{7}\Big(1-\frac{B}{4}\Big)^2 > \frac{6}{5}\Big(\frac{B}{4}+\sqrt{4e^{-5} + \frac{B^2}{16}}\Big)^2$. This implies that $\|r_{A(k+1)}\| \leq \sigma_{k+1}$, and therefore (2.11) holds for the value $k' = k+1$. This verifies the claim and hence completes the proof of Theorem 1.1. □

Let us observe that Theorem 1.1 does not give that f_n is a near-best n-term approximation in the form

$$\|f - f_n\| \leq C_0\sigma_n(f)_{\mathcal{H}}. \tag{2.24}$$

However a simple postprocessing of f_{An} by retaining its n largest components does satisfy (2.24).

Corollary 2.4 *Under the assumptions of Theorem 1.1, let $f_{An} = \Phi\mathbf{c}^{An}$ be the output of WOMP after An steps. Let $T \subset \Gamma$, $\#(T) = n$, be a set of indices corresponding to n largest entries of $\mathbf{c}^{An}$. Define $f_n^* \in \Sigma_n$ to be the element obtained by retaining from f_{An} only the n terms corresponding to the indices in T. Then,*

$$\|f - f_n^*\| \leq C^*\sigma_n(f)_{\mathcal{H}}, \tag{2.25}$$

where the constant C^* depends on the constant C in Theorem 1.1 and on the RIP-constant $\delta_{(A+1)n}$.

Proof: By Lemma 2.1, there exists a $\mathbf{c}$ with $\|\mathbf{c}\|_{\ell^0} \leq n$, such that

$$\|f - \Phi\mathbf{c}\| \leq 2\sigma_n(f)_{\mathcal{H}}. \tag{2.26}$$

It follows that

$$\|\mathbf{c}-\mathbf{c}^{An}\|_{\ell^2} \le \frac{1}{\sqrt{1-\delta_{(A+1)n}}}\|\Phi\mathbf{c}-\Phi\mathbf{c}^{An}\|_{\ell^2} \le \frac{C+2}{\sqrt{1-\delta_{(A+1)n}}}\sigma_n(f)_{\mathcal{H}}. \tag{2.27}$$

If $S = \mathrm{supp}(\mathbf{c})$, we obtain

$$\begin{aligned}\|\mathbf{c}-\mathbf{c}_T^{An}\|_{\ell^2} &\le \|\mathbf{c}_T-\mathbf{c}_T^{An}\|_{\ell^2} + \|\mathbf{c}_{T^c}-\mathbf{c}_{T^c}^{An}\|_{\ell^2} + \|\mathbf{c}_{T^c}^{An}\|_{\ell^2} \\ &\le 2\|\mathbf{c}-\mathbf{c}^{An}\|_{\ell^2} + \|\mathbf{c}_{S^c}^{An}\|_{\ell^2} \\ &\le 3\|\mathbf{c}-\mathbf{c}^{An}\|_{\ell^2},\end{aligned} \tag{2.28}$$

which, by (2.27), provides

$$\|\mathbf{c}-\mathbf{c}_T^{An}\|_{\ell^2} \le 3\|\mathbf{c}-\mathbf{c}^{An}\|_{\ell^2} \le \frac{3(C+2)}{\sqrt{1-\delta_{(A+1)n}}}\sigma_n(f)_{\mathcal{H}}. \tag{2.29}$$

The approximation $\Phi\mathbf{c}_T^{An}$ is in Σ_n and satisfies

$$\|f-\Phi\mathbf{c}_T^{An}\| \le 2\sigma_n(f)_{\mathcal{H}} + \|\Phi(\mathbf{c}_T^{An}-\mathbf{c})\| \le \Big(2+\frac{3\sqrt{1+\delta_{(A+1)n}}(C+2)}{\sqrt{1-\delta_{(A+1)n}}}\Big)\sigma_n(f)_{\mathcal{H}}, \tag{2.30}$$

which proves (2.25). □

Proof of Lemma 2.2: We may assume that $\|r_k\| \ge \|f-g\|$ otherwise there is nothing to prove. First observe now that

$$\begin{aligned}\|r_{k+1}\|^2 &= \|f-P_{k+1}f\|^2 \\ &= \|f-P_kf\|^2 - \|(P_k-P_{k+1})f\|^2 \\ &\le \|r_k\|^2 - |\langle r_k, \varphi_{\gamma_{k+1}}\rangle|^2.\end{aligned}$$

Therefore, it suffices to prove that $\|r_k\|^2 - |\langle r_k, \varphi_{\gamma_{k+1}}\rangle|^2$ is bounded by the right hand side of (2.7) which amounts to showing that

$$(1-\delta)(\|r_k\|^2 - \|f-g\|^2) \le \kappa^{-2}\#(T\setminus S_k)|\langle r_k, \varphi_{\gamma_{k+1}}\rangle|^2. \tag{2.31}$$

To prove this, we first note that

$$\begin{aligned}2\|g-f_k\|\sqrt{\|r_k\|^2-\|f-g\|^2} &\le \|g-f_k\|^2 + \|r_k\|^2 - \|f-g\|^2 \\ &= \|g-f_k\|^2 + \|r_k\|^2 - \|g-f_k-r_k\|^2 \\ &\le 2|\langle g-f_k, r_k\rangle| = 2|\langle g, r_k\rangle|.\end{aligned}$$

This is the same as

$$\|r_k\|^2 - \|f-g\|^2 \le \frac{|\langle g, r_k\rangle|^2}{\|g-f_k\|^2}. \tag{2.32}$$

If we write $f_k = \Phi \mathbf{c}^k$, with $\mathbf{c}^k$ supported on S_k, then the numerator of the right side satisfies

$$\begin{aligned} |\langle g, r_k\rangle| &= |\langle \Phi \mathbf{z}, r_k\rangle| \\ &= |\langle \mathbf{z}_{S_k^c}, \Phi^* r_k\rangle_{\ell^2}| \\ &\leq \|\mathbf{z}_{S_k^c}\|_{\ell^1} \|\Phi^* r_k\|_{\ell^\infty} \\ &\leq \kappa^{-1} \|\mathbf{z}_{S_k^c}\|_{\ell^1} |\langle r_k, \varphi_{\gamma_{k+1}}\rangle| \\ &\leq \kappa^{-1} \sqrt{\#(T \setminus S_k)} \|\mathbf{z}_{S_k^c}\|_{\ell^2} |\langle r_k, \varphi_{\gamma_{k+1}}\rangle| \\ &\leq \kappa^{-1} \sqrt{\#(T \setminus S_k)} \|\mathbf{z} - \mathbf{c}^k\|_{\ell^2} |\langle r_k, \varphi_{\gamma_{k+1}}\rangle|. \end{aligned}$$

On the other hand, recalling that $\delta = \delta_{\#(S_k \cup T)}$, the denominator satisfies by the RIP,

$$\|g - f_k\|^2 = \|\Phi(\mathbf{z} - \mathbf{c}^k)\|^2 \geq (1 - \delta) \|\mathbf{z} - \mathbf{c}^k\|_{\ell^2}^2. \tag{2.33}$$

Therefore we have obtained

$$\|r_k\|^2 - \|f - g\|^2 \leq \frac{\#(T \setminus S_k) |\langle r_k, \varphi_{\gamma_{k+1}}\rangle|^2}{\kappa^2 (1 - \delta)}, \tag{2.34}$$

which is (2.31). □

References

[1] E. Candes, J. Romberg, T. Tao, *Stable signal recovery from incomplete and inaccurate measurements*, Comm. Pure and Appl. Math., **59**(2006), 1207–1223.

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

[3] R.A. DeVore, *Nonlinear approximation*, Acta Numerica, **7** (1998), 51–150.

[4] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing* , Birkhäuser Verlag, Basel, 2013.

[5] E.D. Livshitz, *On the optimality of the Orthogonal Greedy Algorithm for μ-coherent dictionaries*, Journal of Approximation Theory 164-5, 668-681, 2012.

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

[7] J.A. Tropp, *Greed is good: algorithmic results for sparse approximation*, IEEE Trans. Inf. Theory, 50, 2231-2242, 2004.

[8] T. Zhang, *Sparse recovery with orthogonal matching pursuit under RIP*, IEEE Trans. Inform. Theory, 57(9) (2011), 6215–6221.

Albert Cohen
Laboratoire Jacques-Louis Lions, Univerisité Pierre et Marie Curie, 75005, Paris, France
cohen@ann.jussieu.fr

Wolfgang Dahmen
Institut für Geometrie und Praktische Mathematik, RWTH Aachen, Templergraben 55, D-52056 Aachen Germany
dahmen@igpm.rwth-aachen.de

Ronald DeVore
Department of Mathematics, Texas A&M University, College Station, TH77840, USA
rdevore@math.tamu.edu