Available online at www.sciencedirect.com
ScienceDirect
Journal of Approximation Theory 163 (2011) 1134–1145
Journal of Approximation Theory
www.elsevier.com/locate/jat

Full length article

On performance of greedy algorithms☆

Vladimir N. Temlyakov, Pavel Zheltov*

Department of Mathematics, University of South Carolina, Columbia, SC 29208, USA

Received 28 January 2010; received in revised form 19 March 2011; accepted 28 March 2011
Available online 3 April 2011

Communicated by Martin Buhmann

Abstract

We show that the Orthogonal Greedy Algorithm (OGA) for dictionaries in a Hilbert space with small coherence M performs almost as well as the best m-term approximation for all signals with sparsity close to the best theoretically possible threshold $m = \frac{1}{2}(M^{-1} + 1)$ by proving a Lebesgue-type inequality for arbitrary signals. Additionally, we present a dictionary with coherence M and a $\frac{1}{2}(M^{-1} + 1)$-sparse signal for which OGA fails to pick up any atoms from the support, showing that the above threshold is sharp. We also show that the Pure Greedy Algorithm (PGA) matches the rate of convergence of the best m-term approximation beyond the saturation limit of $m^{-\frac{1}{2}}$.

Keywords: Additive-type Lebesgue inequality; Greedy algorithm; Orthogonal matching pursuit; m-term approximation; Incoherent dictionary; Coherence; Orthogonal greedy algorithm; Sparse representation

1. Introduction

In this paper, we study the efficiency of the Orthogonal Greedy Algorithm (OGA). OGA is also known as the Orthogonal Matching Pursuit (OMP) in the field of compressed sensing community concerned mostly with finite-dimensional spaces. We retain a traditional term from approximation theory to preserve the theoretical flavor of our result.

OGA is a simple yet powerful algorithm for highly nonlinear sparse approximation that has seen a large amount of research over its history. See [14,1,9,10,16,17,2,5,15] for a survey. Since its inception, the performance of OGA has served as a baseline of comparison for other algorithms like Regularized OMP [13], Stagewise OMP [6] and others [12].

☆ This research was supported by the National Science Foundation Grant DMS-0906260.
* Corresponding author.
E-mail addresses: temlyak@math.sc.edu (V.N. Temlyakov), zheltovp@sc.edu, pzheltov@gmail.com (P. Zheltov).

0021-9045/$ - see front matter
doi:10.1016/j.jat.2011.03.009

Previous work [3,8,16,4] has shown that both OGA and a convex relaxation algorithm known as Basis Pursuit recover sparse signals exactly if their support size does not exceed a critical threshold $m = \frac{1}{2}(M^{-1} + 1)$. In particular, OGA does so in exactly m steps, recovering one atom from the support of sparse f at every step.

In Section 2, we show that OGA performs as well as the best m-term approximation (up to a factor of $\exp(\sqrt{\log m})$) for all signals with sparsity close to the best theoretically possible threshold $m = O\left(\frac{1}{M}\right)$ (up to the same factor). Next, in Section 3, we explore OGA's sibling Pure Greedy Algorithm (PGA), known as Projection Pursuit [10,9] among statisticians. While even easier to implement, PGA has been shown in [2,11] to approximate elements from special classes less efficiently than OGA. We show that PGA matches the rate of convergence of the best m-term approximation for signals with sparsity close to the same threshold $m = \frac{1}{2}(M^{-1} + 1)$.

In Section 4, we show that this threshold is sharp by explicitly constructing a dictionary with a small coherence and a signal with sparsity $m = \frac{1}{2}(M^{-1} + 1)$, for which OGA fails to find its sparse approximation within m steps. The proofs are relegated to Section 5.

We work in a Hilbert space H. The *dictionary* $\mathcal{D}$ is an arbitrary collection of elements $\{\varphi_i, i \in \mathbb{N}\} \subset H$ such that $\operatorname{span} \mathcal{D}$ is dense in H. We assume that all elements (or atoms) are normalized ($\|\varphi\| = 1$). We define *coherence* M of a dictionary as

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

We say that $\mathcal{D}$ is M-coherent if its coherence is exactly M.

For finite collections of atoms from the dictionary we use a notation that is common in the compressed sensing literature:

$$\Phi_s := \{\varphi_1, \varphi_2, \dots, \varphi_s\} \subset \mathcal{D},$$

and conveniently abuse it to mean a linear operator on its elements

$$\Phi_s x = \sum_{i=1}^{s} x_i \varphi_i.$$

It allows us to see scalar products of f with Φ_s as actions of an adjoint operator of Φ_s

$$\Phi_s^* f = [\langle \varphi_i, f \rangle]_{i=1}^{s},$$

and coefficients of projection of f onto $\operatorname{span} \Phi_s$ as an output of a pseudoinverse of Φ_s:

$$\Phi_s^\dagger f = \arg\min_{x \in \mathbb{R}^s} \|f - \Phi_s x\|, \qquad \Phi_s \Phi_s^\dagger f = \operatorname{proj}_{\Phi_s} f.$$

Throughout the paper, log stands for the base-2 logarithm.

Finally, Pure and Orthogonal Greedy Algorithms construct sequences of approximations of a given signal $f \in H$ according to the following theoretical procedure:

Initialize residual $f_0 := f$ and the set $\Phi_0 := \emptyset$
Repeat for $s = 1, 2, \dots$:

Find the best atom in $\mathcal{D}$:	$\varphi_s = \arg\max_{\varphi \in \mathcal{D}} \lvert\langle f_{s-1}, \varphi \rangle\rvert$
Add it to the list:	$\Phi_s = \Phi_{s-1} \cup \{\varphi_s\}$
Subtract the projection of the residual...	
PGA: ... onto φ_s:	$f_s = f_{s-1} - \langle f_{s-1}, \varphi_s \rangle \varphi_s$
OGA: ... onto Φ_s:	$f_s = f - \operatorname{proj}_{\Phi_s} f$

We assume that the maximizer φ_s exists at every step. Further modifications are necessary otherwise.

2. Lebesgue-type inequalities for OGA

To reduce visual clutter, we use f_k to denote the residuals of both OGA and PGA throughout the paper. It will always be clear from context, which algorithm is being used. We define the best m-term approximation error for a function f from H

$$\sigma_m(f) := \inf_{\varphi_{i_1},\dots,\varphi_{i_m}} \inf_{x\in\mathbb{R}^m} \left\| f - \sum_{k=1}^{m} x_k \varphi_{i_k} \right\|.$$

Following [5], we say inequalities are of *Lebesgue-type* if they compare the norm of the error of a particular method of approximation by elements of a special form or from a special class, with the best possible approximation of f by elements of the same form. For example, a classical inequality of Lebesgue for the error (in the uniform norm) of the mth partial sum S_m of the trigonometric Fourier series of the periodic function f is

$$\|f - S_m f\|_\infty \le C\sigma_m(f)_\infty,$$

where the factor $C = O(\log m)$. Generally, we let both C and the amount of steps η that the particular method of approximation is allowed to make, to depend on m:

$$\left\| f_{\eta(m)} \right\| \le C(m)\sigma_m(f).$$

Clearly, the closer $\eta(m)$ to m, the better, and in the ideal case C is constant.

The first result of this kind was proven by Gilbert et al. in [7].

Theorem 1 (*[7, Corollary 2.2]*)**.** *For every M-coherent dictionary $\mathcal{D}$ and any signal $f \in H$ the application of OGA guarantees*

$$\|f_m\| \le 8\sqrt{m}\sigma_m(f), \quad \text{if } m+1 \le \frac{1}{8\sqrt{2}M}.$$

The constants were refined by Tropp in [16].

Theorem 2 (*[16, Corollary 4.4]*)**.** *For every M-coherent dictionary $\mathcal{D}$ and any signal $f \in H$ the application of OGA guarantees*

$$\|f_m\| \le \sqrt{1+6m}\,\sigma_m(f), \quad \text{if } m \le \frac{1}{3M}.$$

This provides a guarantee that OGA will recover the support of any m-sparse signal after at most m iterations. However, the factor beside σ is very large. This problem was solved by Donoho et al. in [5], who showed the following theorem.

Theorem 3 (*[5, Corollary 1.5]*)**.** *For every M-coherent dictionary $\mathcal{D}$ and any signal $f \in H$ application of OGA guarantees*

$$\left\| f_{\lfloor m \log m \rfloor} \right\| \le 24\sigma_m(f), \quad \text{if } m \le \frac{1}{20M^{\frac{2}{3}}}.$$

While this does provide an absolute constant as a factor, it forces us to sacrifice the amount of steps ($m \log m$ now) and the critical sparsity (only $M^{-\frac{2}{3}}$) to kill the square root in Theorem 2. The method used in [5] was based on the following theorem.

Theorem 4 (*[5, Theorem 1.3]*)**.** *For every M-coherent dictionary $\mathcal{D}$, any signal $f \in H$, and any k, s the application of OGA guarantees*

$$\|f_{k+s}\|^2 \le 2\,\|f_k\|\,(3M(k+s)\,\|f_k\| + \sigma_s(f_k)), \quad \text{if } k+s \le \frac{1}{2M}.$$

In other words, it is possible to estimate f_{k+s} in terms of the best s-term approximation of f_k. A clever recursive argument primed with Theorem 2 then establishes Theorem 3.

The approach used in this paper is a modification of their argument that replaces a crude triangle inequality in [5, (2.6)] with Parseval identity. This essentially closes the gap between $\frac{2}{3}$ and 1. We believe that Theorem 5 is more natural for the construction that was used in [5]. We suggest calling it an *additive-type Lebesgue inequality*.

Theorem 5 (*Additive-Type Lebesgue Inequality*)**.** *For every M-coherent dictionary $\mathcal{D}$, any signal $f \in H$, and any $k \le s$ the application of OGA guarantees*

$$\|f_{k+s}\|^2 \le 7Ms\,\|f_k\|^2 + \sigma_s(f_k)^2, \quad \text{if } k+s \le \frac{1}{2M}.$$

Corollary 6. *For every M-coherent dictionary $\mathcal{D}$ and any signal $f \in H$ the application of OGA guarantees*

$$\left\| f_{m\lfloor 2^{\sqrt{\log m}}\rfloor} \right\| \le 3\sigma_m(f), \quad \text{if } m2^{\sqrt{2\log m}} \le \frac{1}{26M}.$$

Note that the expression $m2^{\sqrt{\log m}}$ grows slower than any $m^{1+\varepsilon}$. If we compare this to $m \log m$ in Theorem 3, we observe that some sacrifices in the amount of iterations had to be made, but these losses do not offset the gains on the sparsity front. This is evident from another corollary.

Corollary 7. *For every M-coherent dictionary $\mathcal{D}$, any signal $f \in H$, and any fixed $\delta > 0$ the application of OGA guarantees*

$$\left\| f_{m2^{\lceil \frac{1}{\delta} \rceil}} \right\| \le 3\sigma_m(f), \quad \text{if } m \le \left(\frac{1}{14M} 2^{-\lceil \frac{1}{\delta} \rceil} \right)^{\frac{1}{1+\delta}}.$$

3. Lebesgue inequality for PGA

In a PGA setting we lack Theorem 2. The possibility of the previously chosen atoms reappearing in the expansion prevents application of Theorem 5. However, this is an advantage over OGA in a sense that PGA outputs a greedy expansion, i.e. it is completely sequential, allowing us to use $f_{m+n} = (f_m)_n$. This causes Theorem 8, an analogue of Theorem 5 with $k = 0$, to provide a surprising pair of corollaries. We will show that if the best m-term approximation rate is $O(m^{-r})$ for some fixed r, then PGA matches this rate up to a constant factor. This is the first result that breaks the saturation barrier, albeit at a cost to the number of iterations. While a similar result was proven in [5] for a general class of dictionaries called λ-quasiorthogonal (which includes M-coherent dictionaries), it suffered from what is known as *saturation property*. Even by imposing very stringent restrictions on $\sigma_m(f)$, we still cannot achieve a rate of approximation better than $m^{-\frac{1}{2}}$ using PGA. In fact, DeVore and Temlyakov construct a signal and a dictionary in [2] such that $\sigma_2(f) = 0$, but $\|f - f_m\| \geq m^{-\frac{1}{2}}$ for $m \geq 4$. Note that the coherence of their dictionary is $M = \sqrt{33/89} = 0.61\ldots$. This entails $m < \frac{1}{2M} < 1$, which prevents the following theorems from applying to that particular situation.

Theorem 8 (*Additive-Type Lebesgue Inequality for PGA*)**.** *For every M-coherent dictionary $\mathcal{D}$ and any signal $f \in H$ the application of PGA guarantees*

$$\|f_s\|^2 \leq 9Ms \|f\|^2 + \sigma_s(f)^2, \quad \textit{if } s \leq \frac{1}{2M}.$$

Corollary 9. *Let $\mathcal{D}$ have coherence M and signal $f \in H$ be such that for some fixed $r > 0$ and for all*

$$m 2^{\sqrt{10r \log m}} \leq \frac{1}{18M}$$

it is true that

$$\sigma_m(f) \leq m^{-r} \|f\|.$$

Then for all such m the application of PGA guarantees

$$\left\| f_{m 2^{\sqrt{10r \log m}}} \right\| \leq 2m^{-r} \|f\|.$$

Just as Corollary 7 is a "hard" realization of a "soft" Corollary 6 as far as the power of m is concerned, Corollary 10 is the power of m version of the previous Corollary 9.

Corollary 10. *Let $\mathcal{D}$ have coherence M and signal $f \in H$ be such that for some fixed $\delta > 0$, $r > 0$ and for all*

$$m \leq N(\delta, r) := \left(\frac{4^{r+1}}{9M} \right)^{\frac{1}{1+\delta}}$$

it is true that

$$\sigma_m(f) \leq m^{-r} \|f\|.$$

Then there exists a constant $C(\delta, r)$ such that the mth PGA residual is suboptimal:

$$\|f_m\| \leq C(\delta, r) m^{-r} \|f\| \quad \textit{for all } m \leq N(\delta, r).$$

In other words, we get PGA residuals matching the rate of best approximation on almost the entire interval $m \in \left[0, O\left(\frac{1}{M}\right)\right]$. The proofs for PGA mostly correspond to the proofs for OGA, so we will only describe the necessary changes.

4. A dictionary with small coherence that is difficult for OGA

From [16,4] we know that OGA recovers an m-sparse signal over M-coherent dictionary $\mathcal{D}$ exactly in no more than m steps if

$$m < \frac{1}{2}\left(\frac{1}{M} + 1\right).$$

We will show that this estimate is sharp.

Theorem 11. *For any $0 < M < 1$ there exists M-coherent dictionary $\mathcal{D}$ and an m-sparse signal f with $m = \frac{1}{2}\left(\frac{1}{M}+1\right)$ but OGA will never recover f exactly.*

Proof. Let $\{e_j\}_{j=1}^{\infty}$ be the standard basis for $H = \ell^2$ and a signal $f = \sum_{i=1}^{m} e_i$ with norm $\|f\| = \sqrt{m}$. Let the dictionary $\mathcal{D}$ be a basis of H comprised of the following two kinds of atoms:

$$D_{\text{good}} = \{\varphi_i = \alpha e_i - \beta f,\ i = 1, \dots, m\}, \quad \text{and}$$
$$D_{\text{bad}} = \{\varphi_j = \eta e_j + \gamma f,\ j = m+1, \dots\}.$$

It is enough to consider $\alpha, \beta, \gamma > 0$. Let η be such that $\eta^2 + m\gamma^2 = 1$ so that $\varphi_j,\ j = m+1, \dots$ are normalized, and

$$(\alpha - \beta)^2 + (m-1)\beta^2 = 1 \tag{1}$$

to normalize $\varphi_i,\ i = 1, \dots, m$. The following are the scalar products of f with the dictionary:

$$\text{For } \varphi_i \in \mathcal{D}_{\text{good}} \quad \langle \varphi_i, f\rangle = \langle \alpha e_i - \beta f, f\rangle = \alpha - m\beta$$
$$\text{For } \varphi_j \in \mathcal{D}_{\text{bad}} \quad \langle \varphi_j, f\rangle = \langle \eta e_j + \gamma f, f\rangle = m\gamma.$$

Let us then require the above dot products to be equal ($R := m\gamma = \alpha - m\beta$). This will allow some realization of OGA to select φ_{m+1} on the first step. Now the scalar products of pairs of distinct elements in $\mathcal{D}$ are as follows:

$$\langle \varphi_i, \varphi_{i'}\rangle = \langle \alpha e_i - \beta f, \alpha e_{i'} - \beta f\rangle = m\beta^2 - 2\alpha\beta = -m\beta(2\gamma + \beta), \quad i, i' \le m,$$
$$\langle \varphi_j, \varphi_{j'}\rangle = \langle \eta e_j + \gamma f, \eta e_{j'} + \gamma f\rangle = m\gamma^2, \quad j, j' > m,$$
$$\langle \varphi_i, \varphi_j\rangle = \langle \alpha e_i - \beta f, \eta e_j + \gamma f\rangle = \gamma(-m\beta + \alpha) = \gamma R = m\gamma^2, \quad i \le m < j.$$

Then coherence of such a dictionary is

$$M := \max(m\gamma^2, m\beta(2\gamma + \beta)),$$

and it makes sense to require $\gamma^2 = \beta(2\gamma + \beta)$. Solving this quadratic equation, we get $\gamma = (1+\sqrt{2})\beta$. Now we can find $\alpha = m\gamma + m\beta = m(2+\sqrt{2})\beta$, and by plugging in (1) we can find β:

$$(m(2+\sqrt{2}) - 1)^2\beta^2 + (m-1)\beta^2 = 1 \qquad \beta^2 = \frac{1}{m^2(2+\sqrt{2})^2 - m(3+2\sqrt{2})}.$$

Plugging everything back in, we see that

$$M = m\gamma^2 = m(1+\sqrt{2})^2\beta^2 = \frac{(1+\sqrt{2})^2}{m(2+\sqrt{2})^2 - (3+2\sqrt{2})}.$$

Denote $A := 1+\sqrt{2}$ and notice that $2+\sqrt{2} = \sqrt{2}A, \quad 3+2\sqrt{2} = A^2$. Now simplify:

$$M = \frac{A^2}{2A^2 \cdot m - A^2} = \frac{1}{2m-1}, \quad \text{or} \quad m = \frac{1}{2}\left(\frac{1}{M}+1\right).$$

Now remember OGA picked a wrong atom ψ from $\mathcal{D}_{\text{bad}}$ on the first step. By induction, suppose that by the nth step OGA has selected n atoms $\psi_1, \psi_2, \dots, \psi_n$ from $\mathcal{D}_{\text{bad}}$. Due to projection, OGA will never select an atom twice, so let us see what happens for $\varphi \in \mathcal{D} \setminus \Psi$:

$$\left\langle \varphi, f - \sum_{j=1}^{n} c_j \psi_j \right\rangle = \langle \varphi, f \rangle - \sum_{j=1}^{n} c_j \langle \varphi, \psi_j \rangle = R - M \sum_{j=1}^{n} c_j.$$

Since all the scalar products are still the same (they do not depend on φ), some realization of OGA will select another atom from $\mathcal{D}_{\text{bad}}$. This completes the induction. In fact, OGA may never select a correct atom from $\mathcal{D}_{\text{good}}$. $\square$

5. Proofs

We need the following simple lemmas.

Lemma 12. *Let $\Phi = \{\varphi_1, \varphi_2, \dots, \varphi_n\}$ be a finite collection of atoms from an M-coherent dictionary $\mathcal{D}$. Then for any $x \in \mathbb{R}^n$*

$$(1 - Mn) \|x\|_2^2 \le \|\Phi x\|^2 \le (1 + Mn) \|x\|_2^2 .$$

Lemma 13. *Let $\Phi = \{\varphi_1, \varphi_2, \dots, \varphi_n\}$ be a finite collection of atoms from an M-coherent dictionary $\mathcal{D}$. Then for any $f \in H$*

$$\|\Phi^* f\|_2 \ge \frac{1 - Mn}{\sqrt{1 + Mn}} \|\mathrm{proj}_\Phi f\| .$$

Proof. See [5, Lemmas 2.1, 2.2] for the proofs of the above lemmas with stronger factors $1 \pm M(n-1)$ in place of $1 \pm Mn$. We make this small sacrifice out of convenience. Lemma 13 follows from [5, Lemma 2.2] via Lemma 12 from the following observation:

$$\left\| \Phi^\dagger f \right\|^2 \ge \frac{1}{1 + Mn} \left\| \Phi \Phi^\dagger f \right\|^2 = \frac{1}{1 + Mn} \|\mathrm{proj}_\Phi f\|^2 . \quad \square \tag{2}$$

Proof of Theorem 5. Recall that $\varphi_{i+1} := \arg\max_{\varphi \in \mathcal{D}} |\langle \varphi, f_i \rangle|$ and $\Phi_i := \{\varphi_1, \varphi_2, \dots, \varphi_i\}$. By Pythagorean theorem,

$$\|f_{k+i}\|^2 = \|f_{k+i-1}\|^2 - \left\| \mathrm{proj}_{\Phi_{k+i}} f_{k+i-1} \right\|^2 \le \|f_{k+i-1}\|^2 - |\langle f_{k+i-1}, \varphi_{k+i} \rangle|^2 , \tag{3}$$

and chaining for $i = 1, \dots, s$,

$$\|f_{k+s}\|^2 \le \|f_k\|^2 - \sum_{i=1}^{s} |\langle f_{k+i-1}, \varphi_{k+i} \rangle|^2 . \tag{4}$$

Let $G \subset \mathcal{D}$ be the collection of s distinct elements that have the biggest scalar products with f_k, i.e.

$$g \in G \Leftrightarrow |\langle g, f_k \rangle| \ge |\langle \varphi, f_k \rangle| \quad \forall \varphi \notin G.$$

If for some $g \in G$ $|\langle f_k, g \rangle| = 0$, we can replace G with a smaller $G' \subset G$, $\#(G') = s' < s$ such that $|\langle f_k, g \rangle| \ne 0$ for all $g \in G'$, and $|\langle f_k, g \rangle| = 0$ for all $g \in \mathcal{D} \setminus G'$. We will then proceed to prove the theorem, and the final statement will follow from the following string of inequalities:

$$\|f_{k+s}\|^2 \le \left\| f_{k+s'} \right\|^2 \le 7Ms' \|f_k\|^2 + \sigma_{s'}(f_k)^2 < 7Ms \|f_k\|^2 + \sigma_s(f_k)^2,$$

where we have used the fact that $\sigma_{s'}(f_k) = 0$, since we know that s' elements is enough:

$$f_k \in \operatorname{span}\{g : g \in G'\}.$$

Thus, without loss of generality, G is such that $|\langle f_k, g\rangle| \neq 0$ for all $g \in G$. It is easy to see that we can order G in a such a way that

$$g_i \notin \Phi_{k+i-1} \quad \forall\, i = 1, \ldots, s \tag{5}$$

if we iteratively ($i = 1, \ldots, s$) select

$$g_i := \begin{cases} \varphi_{k+i}, & \text{if } \varphi_{k+i} \in G, \\ \text{an arbitrary element from } G \setminus (\Phi_{k+i} \cup \{g_1, \ldots, g_{i-1}\}), & \text{otherwise.} \end{cases}$$

By the maximality of φ_{k+i}, we can bound the inner product in (3) by

$$|\langle f_{k+i-1}, g_i\rangle| \le |\langle f_{k+i-1}, \varphi_{k+i}\rangle|,$$

and by construction of f_{k+i-1},

$$\langle f_{k+i-1}, g_i\rangle = \langle f_k - \operatorname{proj}_{\Phi_{k+i-1}} f_k, g_i\rangle = \langle f_k, g_i\rangle - \langle p_i, g_i\rangle, \tag{6}$$

where we denote

$$p_i := \operatorname{proj}_{\Phi_{k+i-1}} f_k = \Phi_{k+i-1} c_i.$$

Returning to (4), we find that

$$\|f_{k+s}\|^2 \le \|f_k\|^2 - \sum_{i=1}^{s} |\langle f_{k+i-1}, g_i\rangle|^2 = \|f_k\|^2 - \sum_{i=1}^{s} |\langle f_k, g_i\rangle - \langle p_i, g_i\rangle|^2. \tag{7}$$

We will treat each of the terms in the last sum separately as follows. (a) To estimate $\sum |\langle f_k, g_i\rangle|^2$, let Ψ represent a collection of elements on which the best s-term approximation of f_k is obtained:

$$\sigma_s(f_k) = \left\| f_k - \operatorname{proj}_{\Psi} f_k \right\|.$$

By maximality of G and Lemma 13,

$$\begin{aligned} \sum_{i=1}^{s} |\langle f_k, g_i\rangle|^2 &\ge \sum_{i=1}^{s} |\langle f_k, \psi_i\rangle|^2 = \left\| \Psi^* f_k \right\|^2 \\ &\ge \frac{(1-Ms)^2}{1+Ms} \left\| \operatorname{proj}_{\Psi} f_k \right\|^2 = \frac{(1-Ms)^2}{1+Ms} (\|f_k\|^2 - \sigma_s(f_k)^2). \end{aligned} \tag{8}$$

(b) For $\sum |\langle p_i, g_i\rangle|^2$, an approach similar but more delicate than [5, (2.4)] works well. We will make use of the assumptions $k \le s$, $M(k+s) \le \frac{1}{2}$.

$$\begin{aligned} \sum_{i=1}^{s} |\langle p_i, g_i\rangle|^2 &= \sum_{i=1}^{s} |\langle \Phi_{k+i-1} c_i, g_i\rangle|^2 \le \sum_{i=1}^{s} \left|\langle c_i, \Phi_{k+i-1}^* g_i\rangle\right|^2 \\ &\le M^2 \sum_{i=1}^{s} \|c_i\|_1^2 \le M^2 \sum_{i=1}^{s} (k+i)\, \|c_i\|_2^2 \\ &\le \sum_{i=1}^{s} \frac{M^2(k+i)}{1 - M(k+i)} \|f_k\|^2 \le M^2 \sum_{i=1}^{s} 2(k+i)\, \|f_k\|^2 \\ &\le 3(Ms)^2 \|f_k\|^2, \end{aligned} \tag{9}$$

where the second inequality is valid since

$$\left\| \Phi_{k+i-1}^{*} g_i \right\|_\infty \leq M,$$

from (5) and coherence of $\mathcal{D}$, and the fourth inequality follows from (2). To combine the two inequalities (8) and (9) back into (7), we use the triangle inequality in a way similar to [5, (2.5)]:

$$\begin{aligned} \left(\sum_{i=1}^{s} |\langle f_k, g_i \rangle - \langle p_i, g_i \rangle|^2 \right)^{\frac{1}{2}} &\geq \left(\sum_{i=1}^{s} |\langle f_k, g_i \rangle|^2 \right)^{\frac{1}{2}} - \left(\sum_{i=1}^{s} |\langle p_i, g_i \rangle|^2 \right)^{\frac{1}{2}} \\ &\geq (1-\alpha x) \left(\|f_k\|^2 - \sigma_s(f_k)^2 \right)^{\frac{1}{2}} - \beta x \|f_k\|, \end{aligned} \tag{10}$$

where we denoted $x := Ms$, $\beta = 2$, and used that on $x \in \left[0, \frac{1}{2}\right]$

$$\frac{1-x}{\sqrt{1+x}} \geq 1 - \alpha x \quad \text{for } \alpha = \frac{3}{2}.$$

For the sake of presentation, $\beta = 2 > \sqrt{3}$ suffices. A more careful treatment of (9) as a right Riemann sum of an increasing function yields the even better $\beta = 4\sqrt{\ln\left(\frac{3}{2}\right) - \frac{1}{4}} = 1.577\ldots$ (see [18]).

The rest is a simple calculus exercise. Observe that the convex quadratic in (10) is above its tangent line at $x = 0$. For simplicity of presentation, let $a = \|f_k\|^2$, $c = \sigma_s(f_k)^2$:

$$\begin{aligned} ((1-\alpha x)\sqrt{a-c} - \beta x \sqrt{a})^2 &\geq a - c - 2x\sqrt{a-c}(\alpha\sqrt{a-c} + \beta\sqrt{a}) \\ &\geq a - c - 2x(\alpha+\beta)a. \end{aligned}$$

Therefore,

$$\|f_{k+s}\|^2 \leq a - (a - c - 7xa) = 7xa + c = 7Ms\|f_k\|^2 + \sigma_s(f_k)^2. \quad \square$$

From here, several analogues to [5, (2.4), (2.5)] can be established. The nature of the estimate allows a purely iterative argument rather than a recursive one. We will need the following trivial lemma about sequences.

Lemma 14. *Let $\{a_l\}_{l=1}^{\infty}$, $\{b_l\}_{l=1}^{\infty}$ be two nonnegative sequences of real numbers such that $b_l < \frac{1}{2}$ for all l, and c be a nonnegative real number. Also, let*

$$a_{l+1} \leq a_l b_l + c \quad \textit{for all } l \in \mathbb{N}. \tag{11}$$

Then for all natural L

$$a_{L+1} \leq a_1 \prod_{l=1}^{L} b_l + 2c.$$

Proof. For $L = 1$, the statement is obvious. Suppose the desired inequality holds for some $L-1$. Then by (11) and by induction hypothesis,

$$a_{L+1} \leq a_L b_L + c \leq \left(a_1 \prod_{l=1}^{L-1} b_l + 2c \right) b_L + c \leq a_1 \prod_{l=1}^{L} b_l + 2c. \quad \square$$

Proof of Corollaries 6 and 7. Fix $m \geq 1$. Let $k_l := m(2^l - 1)$ be a sequence of indices and $\{a_l\}_{l=1}^{\infty}$ be a sequence of squared norms $a_l := \|f_{k_l}\|^2$. Then by Theorem 5, while $Mm2^l \leq \frac{1}{2}$, we have

$$a_{l+1} \leq 7Mm2^l a_l + \sigma_{k_l+m}(f_{k_l})^2 \leq 7Mm2^l a_l + \sigma_m(f)^2, \tag{12}$$

where we use the degrees-of-freedom argument to estimate

$$\sigma_{k_l+m}(f_{k_l}) \leq \sigma_m(f) =: \sigma.$$

By Lemma 14 until $7Mm2^l > \frac{1}{2}$ we get

$$a_L \leq a_1 \prod_{l=1}^{L-1} 7Mm2^l + 2\sigma^2.$$

From Theorem 2 we can initialize $a_1 = \|f_m\|^2 \leq (6m+1)\sigma^2$ for the final estimate

$$a_L \leq \left(2 + (6m+1)\prod_{l=1}^{L-1} 7Mm2^l\right)\sigma^2. \tag{13}$$

If we require $7Mm2^{L-1} \leq \frac{1}{2}m^{-\delta}$ for some fixed $\delta \geq 0$, then

$$7Mm2^l = (7Mm2^{L-1})2^{l-L+1} \leq m^{-\delta}2^{l-L},$$

and therefore

$$\prod_{l=1}^{L-1} 7Mm2^l \leq m^{-(L-1)\delta} \prod_{l=1}^{L-1} 2^{l-L} \leq m^{-(L-1)\delta}2^{-\frac{1}{2}(L-1)^2}.$$

Because each of the factors will suffice to overpower $6m+1$ in (13), we obtain two corollaries. Both conditions will then provide us with

$$\|f_{m2^L}\| \leq 3\sigma_m(f). \tag{14}$$

Proof of Corollary 6 ($\delta = 0$)**.** We need $7Mm2^{L-1} \leq \frac{1}{2}$ and $m2^{-\frac{1}{2}(L-1)^2} < 1$. By stipulating $L = \left\lceil \sqrt{2\log m} \right\rceil + 1$, if

$$m2^{\sqrt{2\log m}} \leq \frac{1}{26M},$$

then after $m(2^{\left\lceil \sqrt{\log m} \right\rceil + 1} - 1)$ iterations we get (14). □

Proof of Corollary 7 ($\delta > 0$)**.** Stipulate $L - 1 = \left\lceil \frac{1}{\delta} \right\rceil$. If there exists $\delta > 0$ such that

$$7Mm2^{\left\lceil \frac{1}{\delta} \right\rceil} \leq \frac{1}{2}m^{-\delta}, \quad \text{or rewriting,} \quad m \leq \left(\frac{1}{14M}2^{-\left\lceil \frac{1}{\delta} \right\rceil}\right)^{\frac{1}{1+\delta}},$$

the work is finished after at most $m2^{\left\lceil \frac{1}{\delta} \right\rceil + 1}$ iterations. □

Proof of Theorem 8. In (7), we have an expansion $f_s = f - \Phi_s c_s$ instead. If we use $\|\Phi_s c_s\| \le \|f\| + \|f_s\| \le 2\|f\|$, the following estimate holds:

$$|\langle \Phi_s c_s, g_i\rangle| \le M\sqrt{s}\,\|c_s\|_2 \le \frac{M\sqrt{s}}{\sqrt{1-Ms}}\,\|\Phi_s c_s\| \le \frac{2M\sqrt{s}\,\|f\|}{\sqrt{1-Ms}},$$

and then (10) holds with $\alpha = \frac{3}{2}$, $\beta = 3$. Therefore, for all $s \le \frac{1}{2M}$

$$\|f_s\|^2 \le 9Ms\,\|f\|^2 + \sigma_s(f)^2. \quad \square$$

Proof of Corollary 9. By using the same argument found in the above proof of Corollaries 6 and 7, we obtain (if $9Mm2^{L-1} \le \frac{1}{2}$)

$$\left\|f_{m2^L}\right\|^2 \le 2^{-\frac{1}{2}(L-1)^2}\,\|f\|^2 + 2\sigma_m(f)^2.$$

If we require a certain rate of convergence on $\sigma_m(f)$, we can match it for the PGA residual at some cost. Suppose $\sigma_m(f) \le m^{-r}\,\|f\|$. By selecting $L = \sqrt{10r\log m} \ge \left\lceil \sqrt{4r\log m - 2} + 1\right\rceil$, we find that

$$\left\|f_{m2^L}\right\| \le 2m^{-r}\,\|f\|$$

for all m such that

$$m2^L \le \frac{1}{18M}. \quad \square$$

Proof of Corollary 10. If we require a stronger condition $9Mm2^{L-1} \le \frac{1}{2}m^{-\delta}$, we find

$$\left\|f_{m2^L}\right\|^2 \le m^{-(L-1)\delta}\,\|f\|^2 + 2\sigma_m(f)^2. \tag{15}$$

Suppose now we have $L = \left\lceil \frac{2r}{\delta} \right\rceil + 1$, and $m \le N'(\delta, r) := \left(\frac{1}{9M}2^{-L}\right)^{\frac{1}{1+\delta}}$. Then

$$\left\|f_{m2^L}\right\|^2 \le m^{-2r}\,\|f\|^2 + 2\sigma_m(f)^2. \tag{16}$$

Denoting $n = m2^L$ and using that $\sigma_m(f) \le m^{-r}\,\|f\|$, from (16) we find

$$\|f_n\|^2 \le C(\delta, r)n^{-2r}\,\|f\|^2$$

for all

$$n \le N'(\delta, r)2^L = \left(\frac{1}{9M}\right)^{\frac{1}{1+\delta}} 2^{L\left(1-\frac{1}{1+\delta}\right)} \le \left(\frac{1}{9M}\right)^{\frac{1}{1+\delta}} 2^{\frac{2(r+1)}{\delta+1}} =: N(\delta, r). \quad \square$$

References

[1] G. Davis, S. Mallat, M. Avellaneda, Adaptive greedy approximations, Constr. Approx. 13 (1) (1997) 57–98. doi:10.1007/s003659900033.

[2] R.A. DeVore, V.N. Temlyakov, Some remarks on greedy algorithms, Adv. Comput. Math. 5 (2–3) (1996) 173–187. doi:10.1007/BF02124742.

[3] D.L. Donoho, M. Elad, Optimally sparse representation in general (nonorthogonal) dictionaries via l^1 minimization, Proc. Natl. Acad. Sci. USA 100 (5) (2003) 2197–2202 (electronic) doi:10.1073/pnas.0437847100.

[4] D.L. Donoho, M. Elad, V.N. Temlyakov, Stable recovery of sparse overcomplete representations in the presence of noise, IEEE Trans. Inform. Theory 52 (1) (2006) 6–18. doi:10.1109/TIT.2005.860430.

[5] D.L. Donoho, M. Elad, V.N. Temlyakov, On Lebesgue-type inequalities for greedy approximation, J. Approx. Theory 147 (2007) 185–195.
[6] D.L. Donoho, Y. Tsaig, I. Drori, J. Luc Starck, Sparse solution of underdetermined linear equations by stagewise orthogonal matching pursuit, Tech. Rep., 2006.
[7] A.C. Gilbert, S. Muthukrishnan, 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, ACM, New York, 2003, pp. 243–252.
[8] R. Gribonval, M. Nielsen, Sparse representations in unions of bases, IEEE Trans. Inform. Theory 49 (12) (2003) 3320–3325. doi:10.1109/TIT.2003.820031.
[9] P.J. Huber, Projection pursuit, Ann. Statist. 13 (2) (1985) 435–525. doi:10.1214/aos/1176349519.
[10] L.K. Jones, On a conjecture of Huber concerning the convergence of projection pursuit regression, Ann. Statist. 15 (2) (1987) 880–882.
[11] E.D. Livshitz, V.N. Temlyakov, Two lower estimates in greedy approximation, Constr. Approx. 19 (4) (2003) 509–523. doi:10.1007/s00365-003-0533-6.
[12] D. Needell, J.A. Tropp, CoSaMP: iterative signal recovery from incomplete and inaccurate samples, Appl. Comput. Harmon. Anal. 26 (3) (2009) 301–321. doi:10.1016/j.acha.2008.07.002.
[13] D. Needell, R. Vershynin, Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit, Found. Comput. Math. 9 (3) (2009) 317–334. doi:10.1007/s10208-008-9031-3.
[14] Y.C. Pati, R. Rezaiifar, P.S. Krishnaprasad, Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition, in: Proc. 27th Annu. Asilomar Conf. Signals, Systems and Computers, 1993.
[15] V.N. Temlyakov, Greedy approximation, Acta Numer. 17 (2008) 235–409. doi:10.1017/S0962492906380014.
[16] J.A. Tropp, Greed is good: algorithmic results for sparse approximation, IEEE Trans. Inform. Theory 50 (10) (2004) 2231–2242. doi:10.1109/TIT.2004.834793.
[17] J.A. Tropp, A.C. Gilbert, Signal recovery from random measurements via orthogonal matching pursuit, IEEE Trans. Inform. Theory 53 (12) (2007) 4655–4666. doi:10.1109/TIT.2007.909108.
[18] P. Zheltov, Lebesgue-type inequalities for greedy approximation, Ph.D. Thesis, USC, 2010.