Orthogonal Matching Pursuit: Recursive Function Approximation with Applications to Wavelet Decomposition

Ramin Rezaiifar
Qualcomm

/ To appear in Proc. of the 27^{th} Annual Asilomar Conference on Signals Systems and Computers, Nov. 1–3, 1993 /

Orthogonal Matching Pursuit: Recursive Function Approximation with Applications to Wavelet Decomposition

Y. C. Pati
Information Systems Laboratory
Dept. of Electrical Engineering
Stanford University, Stanford, CA 94305

R. Rezaiifar and P. S. Krishnaprasad
Institute for Systems Research
Dept. of Electrical Engineering
University of Maryland, College Park, MD 20742

Abstract

In this paper we describe a recursive algorithm to compute representations of functions with respect to nonorthogonal and possibly overcomplete *dictionaries* of elementary building blocks *e.g.* affine (wavelet) frames. We propose a modification to the Matching Pursuit algorithm of Mallat and Zhang (1992) that maintains full backward orthogonality of the residual (error) at every step and thereby leads to improved convergence. We refer to this modified algorithm as Orthogonal Matching Pursuit (OMP). It is shown that all additional computation required for the OMP algorithm may be performed recursively.

1 Introduction and Background

Given a collection of vectors $D = \{x_i\}$ in a Hilbert space $\mathcal{H}$, let us define

$$\mathbf{V} = \overline{\mathrm{Span}\{x_n\}}, \quad \text{and} \quad \mathbf{W} = \mathbf{V}^{\perp} \quad (\text{in } \mathcal{H}).$$

We shall refer to D as a *dictionary*, and will assume the vectors x_n, are normalized ($\|x_n\| = 1$). In [3] Mallat and Zhang proposed an iterative algorithm that they termed Matching Pursuit (MP) to construct representations of the form

$$\mathbf{P}_V f = \sum_n a_n x_n, \tag{1}$$

where $\mathbf{P}_V$ is the orthogonal projection operator onto V. Each iteration of the MP algorithm results in an intermediate representation of the form

$$f = \sum_{i=1}^{k} a_i x_{n_i} + \mathbf{R}_k f = f_k + \mathbf{R}_k f,$$

where f_k is the current approximation, and $\mathbf{R}_k f$ the current residual (error). Using initial values of $\mathbf{R}_0 f = f$, $f_0 = 0$, and $k = 1$, the MP algorithm is comprised of the following steps,

(I) Compute the inner-products $\{\langle \mathbf{R}_k f, x_n \rangle\}_n$.

(II) Find n_{k+1} such that

$$\left|\left\langle \mathbf{R}_k f, x_{n_{k+1}} \right\rangle\right| \geq \alpha \sup_j \left|\langle \mathbf{R}_k f, x_j \rangle\right|,$$

where $0 < \alpha \leq 1$.

(III) Set,

$$\begin{aligned} f_{k+1} &= f_k + \left\langle \mathbf{R}_k f, x_{n_{k+1}} \right\rangle x_{n_{k+1}} \\ \mathbf{R}_{k+1} f &= \mathbf{R}_k f - \left\langle \mathbf{R}_k f, x_{n_{k+1}} \right\rangle x_{n_{k+1}} \end{aligned}$$

(IV) Increment k, ($k \leftarrow k+1$), and repeat steps (I)–(IV) until some convergence criterion has been satisfied.

The proof of convergence [3] of MP relies essentially on the fact that $\left\langle \mathbf{R}_{k+1} f, x_{n_{k+1}} \right\rangle = 0$. This orthogonality of the residual to the last vector selected leads to the following "energy conservation" equation.

$$\|\mathbf{R}_k f\|^2 = \|\mathbf{R}_{k+1} f\|^2 + \left|\left\langle \mathbf{R}_k f, x_{n_{k+1}} \right\rangle\right|^2. \tag{2}$$

It has been noted that the MP algorithm may be derived as a special case of a technique known as Projection Pursuit (*c.f.* [2]) in the statistics literature.

A shortcoming of the Matching Pursuit algorithm in its originally proposed form is that although asymptotic convergence is guaranteed, the resulting approximation after any finite number of iterations will in general be suboptimal in the following sense. Let $N < \infty$, be the number of MP iterations performed. Thus we have

$$f_N = \sum_{k=0}^{N \perp 1} \langle \mathbf{R}_k f, x_{n_{k+1}} \rangle \, x_{n_{k+1}}.$$

Define $\mathbf{V}_N = \overline{\mathrm{Span}\{x_{n_1}, \ldots, x_{n_N}\}}$. We shall refer to f_N as an *optimal N-term approximation* if $f_N = \mathbf{P}_{\mathbf{V}_N} f$, i.e. f_N is the best approximation we can construct using the selected subset $\{x_{n_1}, \ldots, x_{n_N}\}$ of the dictionary D. (Note that this notion of optimality does not involve the problem of selecting an "optimal" N-element subset of the dictionary.) In this sense, f_N is an optimal N-term approximation, if and only if $\mathbf{R}_N f \in \mathbf{V}_N^{\perp}$. As MP only guarantees that $\mathbf{R}_N f - x_{n_N}$, f_N as generated by MP will in general be suboptimal. The difficulty with such suboptimality is easily illustrated by a simple example in $\mathbb{R}^2$. Let x_1, and x_2 be two vectors in $\mathbb{R}^2$, and take $f \in \mathbb{R}^2$, as shown in Figure 1(a). Figure 1(b) is a plot of $\|\mathbf{R}_k f\|^2$

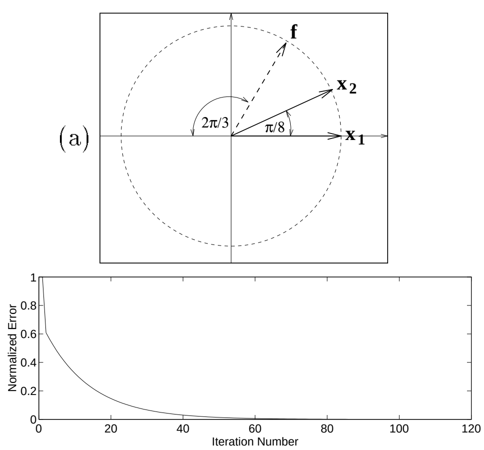

Figure 1: Matching pursuit example in $\mathbb{R}^2$: (a) Dictionary $D = \{x_1, x_2\}$ and a vector $f \in \mathbb{R}^2$

versus k. Hence although asymptotic convergence is guaranteed, after any finite number of steps, the error may still be quite large. [1]

In this paper we propose a refinement of the Matching Pursuit (MP) algorithm that we refer to as Orthogonal Matching Pursuit (OMP). For nonorthogonal dictionaries, OMP will in general converge faster than MP. For any finite size dictionary of N elements, OMP converges to the projection onto the span of the dictionary elements in no more than N steps. Furthermore after any finite number of iterations, OMP gives the optimal approximation with respect to the selected subset of the dictionary. This is achieved by ensuring full backward orthogonality of the error i.e. at each iteration $\mathbf{R}_k f \in \mathbf{V}_k^{\perp}$. For the example in Figure 1, OMP ensures convergence in exactly two iterations. It is also shown that the additional computation required for OMP, takes a simple recursive form.

We demonstrate the utility of OMP by example of applications to representing functions with respect to time-frequency localized affine wavelet dictionaries. We also compare the performance of OMP with that of MP on two numerical examples.

2 Orthogonal Matching Pursuit

Assume we have the following k^{th}-order model for $f \in \mathcal{H}$,

$$f = \sum_{n=1}^{k} a_n^k x_n + \mathbf{R}_k f, \quad \text{with } \langle \mathbf{R}_k f, x_n \rangle = 0, \quad n = 1, \ldots k. \tag{3}$$

The superscript k, in the coefficients a_n^k, show the dependence of these coefficients on the model-order. We would like to update this k^{th}-order model to a model of order $k+1$,

$$f = \sum_{n=1}^{k+1} a_n^{k+1} x_n + \mathbf{R}_{k+1} f, \quad \text{with } \langle \mathbf{R}_{k+1} f, x_n \rangle = 0, \quad n = 1, \ldots k+1. \tag{4}$$

Since elements of the dictionary D are not required to be orthogonal, to perform such an update, we also require an auxiliary model for the dependence of x_{k+1} on the previous x_n's ($n = 1, \ldots k$). Let,

$$x_{k+1} = \sum_{n=1}^{k} b_n^k x_n + \gamma_k, \quad \text{with } \langle \gamma_k, x_n \rangle = 0, \quad n = 1, \ldots k. \tag{5}$$

Thus $\sum_{n=1}^{k} b_n^k x_n = \mathbf{P}_{\mathbf{V}_k} x_{k+1}$, and $\gamma_k = \mathbf{P}_{\mathbf{V}_k^{\perp}} x_{k+1}$, is the component of x_{k+1} which is unexplained by $\{x_1, \ldots, x_k\}$.

Using the auxiliary model (5), it may be shown that the correct update from the k^{th}-order model to the model of order $k+1$, is given by

$$\begin{aligned} a_n^{k+1} &= a_n^k - \alpha_k b_n^k, \quad n = 1, \ldots, k \qquad (6) \\ \text{and } a_{k+1}^{k+1} &= \alpha_k, \\ \text{where } \alpha_k &= \frac{\langle \mathbf{R}_k f, x_{k+1} \rangle}{\langle \gamma_k, x_{k+1} \rangle} = \frac{\langle \mathbf{R}_k f, x_{k+1} \rangle}{\|\gamma_k\|^2} \\ &= \frac{\langle \mathbf{R}_k f, x_{k+1} \rangle}{\|x_{k+1}\|^2 - \sum_{n=1}^{k} b_n^k \langle x_n, x_{k+1} \rangle}. \end{aligned}$$

[1] A simlar difficulty with the Projection Pursuit algorithm was noted by Donoho *et.al.* [1] who suggested that *backfitting* may be used to improve the convergence of PPR. Although the technique is not fully described in [1] it appears that it is in the same spirit as the technique we present here.

It also follows that the residual $\mathbf{R}_{k+1}f$ satisfies, $\mathbf{R}_k f = \mathbf{R}_{k+1}f + \alpha_k \gamma_k$, and

$$\|\mathbf{R}_k f\|^2 = \|\mathbf{R}_{k+1} f\|^2 + \frac{|\langle \mathbf{R}_k f, x_{k+1}\rangle|^2}{\|\gamma_k\|^2}. \tag{7}$$

2.1 The OMP Algorithm

The results of the previous section may be used to construct the following algorithm that we will refer to as Orthogonal Matching Pursuit (OMP).

Initialization:

$$\begin{array}{lll} f_0 = 0, & \mathbf{R}_0 f = f, & D_0 = \{\ \} \\ x_0 = 0, & a_0^0 = 0, & k = 0 \end{array}$$

(I) Compute $\{\langle \mathbf{R}_k f, x_n\rangle ; x_n \in D \setminus D_k\}$.

(II) Find $x_{n_{k+1}} \in D \setminus D_k$ such that

$$\left|\langle \mathbf{R}_k f, x_{n_{k+1}}\rangle\right| \geq \alpha \sup_j |\langle \mathbf{R}_k f, x_j\rangle|, \ \ 0 < \alpha \leq 1.$$

(III) If $\left|\langle \mathbf{R}_k f, x_{n_{k+1}}\rangle\right| < \delta$, $(\delta > 0)$ then stop.

(IV) Reorder the dictionary D, by applying the permutation $k+1 \leftrightarrow n_{k+1}$.

(V) Compute $\{b_n^k\}_{n=1}^k$, such that,

$$x_{k+1} = \sum_{n=1}^k b_n^k x_n + \gamma_k$$
$$\text{and } \ \langle \gamma_k, x_n\rangle = 0, \quad n = 1, \ldots, k.$$

(VI) Set, $a_{k+1}^{k+1} = \alpha_k = \|\gamma_k\|^{-2}\langle \mathbf{R}_k f, x_{k+1}\rangle$,

$$a_n^{k+1} = a_n^k - \alpha_k b_n^k, \quad n = 1, \ldots, k,$$

and update the model,

$$\begin{aligned} f_{k+1} &= \sum_{n=1}^{k+1} a_n^{k+1} x_n \\ \mathbf{R}_{k+1} f &= f - f_{k+1} \\ D_{k+1} &= D_k \bigcup \{x_{k+1}\}. \end{aligned}$$

(VII) Set $k \leftarrow k+1$, and repeat (I)–(VII).

2.2 Some Properties of OMP

As in the case of MP, convergence of OMP relies on an energy conservation equation that now takes the form (7). The following theorem summarizes the convergence properties of OMP.

Theorem 2.1 *For $f \in \mathcal{H}$, let $\mathbf{R}_k f$ be the residuals generated by OMP. Then*

(i) $\lim_{k\to\infty} \|\mathbf{R}_k f - \mathbf{P}_{V^\perp} f\| = 0.$

(ii) $f_N = \mathbf{P}_{\mathbf{V}_N} f, \quad N = 0, 1, 2, \ldots.$

Proof: The proof of convergence parallels the proof of Theorem 1 in [3]. The proof of the the second property follows immediately from the orthogonality conditions of Equation (3).

Remarks:

The key difference between MP and OMP lies in Property (iii) of Theorem 2.1. Property (iii) implies that at the N^{th} step we have the best approximation we can get using the N vectors we have selected from the dictionary. Therefore in the case of finite dictionaries of size M, OMP converges in no more than M iterations to the projection of f onto the span of the dictionary elements. As mentioned earlier, Matching Pursuit does not possess this property.

2.3 Some Computational Details

As in the case of MP, the inner products $\{\langle \mathbf{R}_k f, x_j\rangle\}$ may be computed recursively. For OMP we may express these recursions implicitly in the formula

$$\langle \mathbf{R}_k f, x_j\rangle = \langle f - f_k, x_j\rangle = \langle f, x_j\rangle - \sum_{n=1}^k a_n^k \langle x_n, x_j\rangle. \tag{8}$$

The only additional computation required for OMP, arises in determining the b_n^k's of the auxiliary model (5). To compute the b_n^k's we rewrite the normal equations associated with (5) as a system of k linear equations,

$$\mathbf{v}_k = \mathbf{A}_k \mathbf{b}_k, \tag{9}$$

where

$$\begin{aligned} \mathbf{v}_k &= [\langle x_{k+1}, x_1\rangle, \langle x_{k+1}, x_2\rangle \ldots, \langle x_{k+1}, x_k\rangle]^T \\ \mathbf{b}_k &= [b_1^k, b_2^k, \ldots, b_k^k]^T \end{aligned}$$

and

$$\mathbf{A}_k = \begin{bmatrix} \langle x_1, x_1\rangle & \langle x_2, x_1\rangle & \cdots & \langle x_k, x_1\rangle \\ \langle x_1, x_2\rangle & \langle x_2, x_2\rangle & \cdots & \langle x_k, x_2\rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle x_1, x_k\rangle & \langle x_2, x_k\rangle & \cdots & \langle x_k, x_k\rangle \end{bmatrix}.$$

Note that the positive constant δ used in Step (III) of OMP guarantees nonsingularity of the matrix $\mathbf{A}_k$, hence we may write

$$\mathbf{b}_k = \mathbf{A}_k^{\perp 1}\mathbf{v}_k. \tag{10}$$

However, since $\mathbf{A}_{k+1}$ may be written as

$$\mathbf{A}_{k+1} = \left[\begin{array}{c|c} \mathbf{A}_k & \mathbf{v}_k \\ \hline \mathbf{v}_k^* & 1 \end{array}\right], \tag{11}$$

(where $*$ denotes conjugate transpose) it may be shown using the block matrix inversion formula that

$$\mathbf{A}_{k+1}^{\perp 1} = \left[\begin{array}{c|c} \mathbf{A}_k^{\perp 1} + \beta\mathbf{b}_k\mathbf{b}_k^* & -\beta\mathbf{b}_k \\ \hline -\beta\mathbf{b}_k^* & \beta \end{array}\right], \tag{12}$$

where $\beta = 1/(1-\mathbf{v}_k^*\mathbf{b}_k)$. Hence $\mathbf{A}_{k+1}^{\perp 1}$, and therefore $\mathbf{b}_{k+1}$, may be computed recursively using $\mathbf{A}_k^{\perp 1}$, and $\mathbf{b}_k$ from the previous step.

3 Examples

In the following examples we consider representations with repect to an affine wavelet frame constructed from dilates and translates of the second derivate of a Gaussian, *i.e.* $D = \{\psi_{m,n},\ m,n \in \mathbb{Z}\}$ where,

$$\psi_{m,n}(x) = 2^{m/2}\psi(2^m x - n),$$

and the *analyzing wavelet* ψ is given by,

$$\psi(x) = \left(\frac{4}{3\sqrt{\pi}}\right)^{1/2}\left(x^2-1\right)e^{\perp x^2/2}.$$

Note that for wavelet dictionaries, the initial set of inner products $\{\langle f, \psi_{m,n}\rangle\}$, are readily computed by one convolution followed by sampling at each dilation level m. The dictionary used in these examples consists of a total of 351 vectors.

In our first example, both OMP and MP were applied to the signal shown in Figure 2(a). We see from Figure 2(b) that OMP clearly converges in far fewer iterations than MP. The squared magnitude of the coefficients a_k, of the resulting representation is shown in Figure 3. We could also compare the two algorithms on the basis of required computational effort to compute representations of signals to within a prespecified error. However such a comparison can only be made for a given signal and dictionary, as the number of iterations required for each algorithm depends on both the signal and the dictionary. For example, for the signal of Example I, we see from Figure 4 that it is 3

(a)

(a)

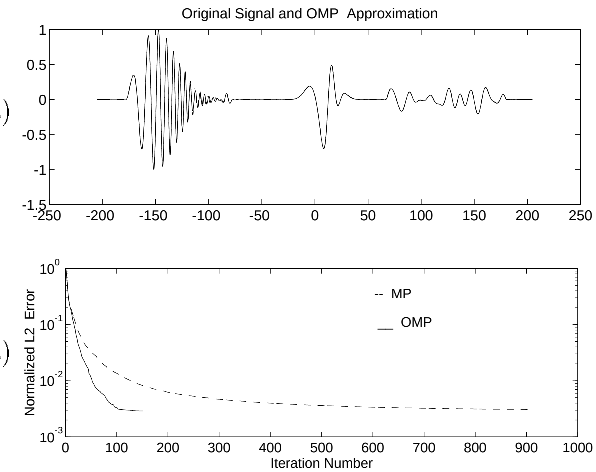

Figure 2: Example I : (a) Original signal f, with OMP approximation superimposed, (b) Squared L^2 norm of residual $\mathbf{R}_k f$ versus iteration number k, for both OMP (solid line) and MP (dashed line).

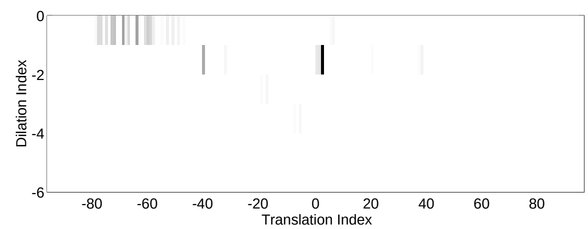

Figure 3: Distribution of coefficients obtained by applying OMP in Example I. Shading is proportional to squared magnitude of the coefficients a_k, with dark colors indicating large magnitudes.

to 8 times more expensive to achieve a prespecified error using OMP even though OMP converges in fewer iterations. On the other hand for the signal shown in Figure 5, which lies in the span of three dictionary vectors, it is approximately 20 times more expensive to apply MP. In this case OMP converges in exactly three iterations.

4 Summary and Conclusions

In this paper we have described a recursive algorithm, which we refer to as Orthogonal Matching Pursuit (OMP), to compute representations of signals with respect to arbitrary dictionaries of elementary functions. The algorithm we have described is a modification of the Matching Pursuit (MP) algorithm of Mallat and Zhang [3] that improves convergence us-

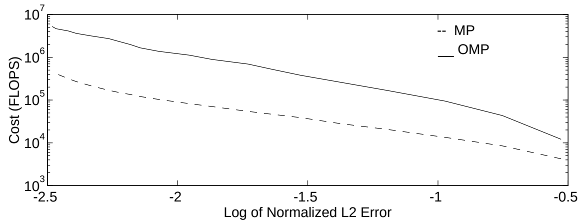

Figure 4: Computational cost (FLOPS) versus approximation error for both OMP (solid line) and MP (dashed line) applied to the signal in Example I.

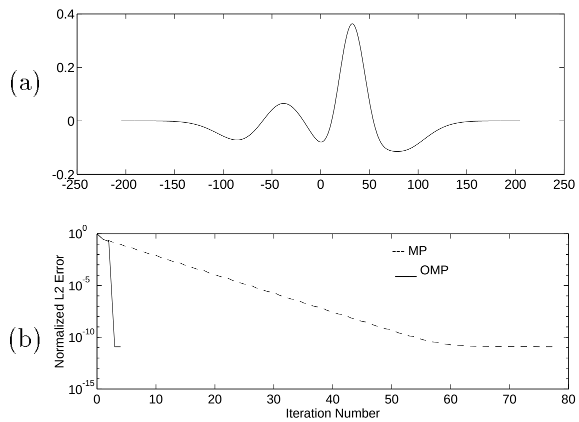

Figure 5: Example II: (a) Original signal f, (b) Squared L^2 norm of residual $\mathbf{R}_k f$ versus iteration number k, for both OMP (solid line) and MP (dashed line).

ing an additional orthogonalization step. The main benefit of OMP over MP is the fact that it is guaranteed to converge in a finite number of steps for a finite dictionary. We also demonstrated that all additional computation that is required for OMP may be performed recursively.

The two algorithms, MP and OMP, were compared on two simple examples of decomposition with respect to a wavelet dictionary. It was noted that although OMP converges in fewer iterations than MP, the computational effort required for each algorithm depends on both the class of signals and choice of dictionary. Although we do not provide a rigorous argument here, it seems reasonable to conjecture that OMP will be computationally cheaper than MP for very redundant dictionaries, as knowledge of the redundancy is exploited in OMP to reduce the error as much as possible at each step.

Acknowledgements

This research of Y.C.P. was supported in part by NASA Headquarters, Center for Aeronautics and Space Information Sciences (CASIS) under Grant NAGW419,S6, and in part by the Advanced Research Projects Agency of the Department of Defense monitored by the Air Force Office of Scientific Research under Contract F49620-93-1-0085.

This research of R.R. and P.S.K was supported in part by the Air Force Office of Scientific Research under contract F49620-92-J-0500, the AFOSR University Research Initiative Program under Grant AFOSR-90-0105, by the Army Research Office under Smart Structures URI Contract no. DAAL03-92-G-0121, and by the National Science Foundation's Engineering Research Centers Program, NSFD CDR 8803012.

References

[1] D. Donoho, I. Johnstone, P. Rousseeuw, and W. Stahel. *The Annals of Statistics*, 13(2):496–500, 1985. Discussion following article by P. Huber.

[2] P. J. Huber. Projection pursuit. *The Annals of Statistics*, 13(2):435–475, 1985.

[3] S. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. Preprint. Submitted to *IEEE Transactions on Signal Processing*, 1992.