Iterative Thresholding for Sparse Approximations

The Journal of Fourier Analysis and Applications

Iterative Thresholding for Sparse Approximations

Thomas Blumensath and Mike E. Davies

ABSTRACT. Sparse signal expansions represent or approximate a signal using a small number of elements from a large collection of elementary waveforms. Finding the optimal sparse expansion is known to be NP hard in general and non-optimal strategies such as Matching Pursuit, Orthogonal Matching Pursuit, Basis Pursuit and Basis Pursuit De-noising are often called upon. These methods show good performance in practical situations, however, they do not operate on the ℓ_0 penalised cost functions that are often at the heart of the problem. In this paper we study two iterative algorithms that are minimising the cost functions of interest. Furthermore, each iteration of these strategies has computational complexity similar to a Matching Pursuit iteration, making the methods applicable to many real world problems. However, the optimisation problem is non-convex and the strategies are only guaranteed to find local solutions, so good initialisation becomes paramount. We here study two approaches. The first approach uses the proposed algorithms to refine the solutions found with other methods, replacing the typically used conjugate gradient solver. The second strategy adapts the algorithms and we show on one example that this adaptation can be used to achieve results that lie between those obtained with Matching Pursuit and those found with Orthogonal Matching Pursuit, while retaining the computational complexity of the Matching Pursuit algorithm.

1. Introduction

Sparse signal approximations have over the last decade gained in popularity in several areas of signal processing. For example, a wide range of signal processing applications such as source coding [1], [2], denoising [3], source separation [4] and pattern analysis [5] have benefited from progress made in this area. A sparse signal approximation models a signal $\mathbf{x}$ from a Hilbert

Keywords and Phrases. Sparse Approximations, Iterative Thresholding, ℓ_0 Regularisation, Subset Selection.

ISSN 1069-5869

space $\mathcal{H}$ as

$$\mathbf{x} = \sum_i \phi_i y_i + \mathbf{e},$$

where $\{\phi_i\}$ is a set of elements from $\mathcal{H}$, commonly called the dictionary, which span $\mathcal{H}$, i.e. which contain a basis for $\mathcal{H}$ and where $\mathbf{e} \in \mathcal{H}$ is an approximation error. In this paper we are dealing primarily with sparse approximations as opposed to sparse representations, i.e. we allow for a non-zero error $\mathbf{e}$ in the above signal model. We commonly use a matrix like notation, using the operator Φ and coefficient sequence $\mathbf{y}$, and write $\Phi\mathbf{y} = \sum_i \phi_i y_i$.

1.1 Useful Dictionary Properties

Before proceeding, we introduce some useful concepts and properties of dictionaries, which will help us in the development below. The *spark* of a dictionary is defined in [6] as

Definition 1 : Spark. *spark*(Φ) is the size of the smallest subset of elements from $\{\phi_i\}$ such that the elements in this subset are linearly dependent.

For example if all M dimensional subsets of elements from $\{\phi_i\}$ are linearly independent, but there exist a subset of size $M+1$ in which the elements are dependent, then the spark of Φ is $M+1$.

We also need the following definitions

Definition 2 : Cumulative Coherence. The Cumulative Coherence or Babel function [7] is defined as

$$\mu_1(m) := \sup_{|\Gamma|=m} \sup_{\omega\notin\Gamma} \sum_{\gamma\in\Gamma} |\langle\phi_\omega, \phi_\gamma\rangle|. \tag{1.1}$$

A useful bound on the cumulative coherence is given in terms of the coherence

Definition 3 : Coherence.

$$\mu_1(1) = \mu_0 := \sup_{\omega\neq\gamma} |\langle\phi_\omega, \phi_\gamma\rangle|. \tag{1.2}$$

A bound on the Cumulative Coherence is then

$$\mu_1(m) \leq m\mu_0. \tag{1.3}$$

See for example [7]. Note, that contrary to common practice, in the above definitions and throughout this paper we do not assume that $\|\phi_i\|_2 = 1$. However, on occasion we use the weaker condition that $\|\phi_i\|_2 = c$ for all i.

In the proof of [8, Theorem 9.10] Mallat also introduced a useful property of Φ. Under the condition that $\{\phi_i\}$ contains a basis for the signal

space and with the assumption that $\|\phi_i\|_2 > c > 0$, there exists a constant $\beta(\Phi) > 0$ such that

$$\sup_i |\langle \phi_i, \mathbf{r} \rangle| \geq \beta \|\mathbf{r}\|. \tag{1.4}$$

The proof of this property in [8] assumes that $\|\phi_i\|_2 = 1$, but the assumption $\|\phi_i\|_2 > c > 0$ and a simple re-normalisation argument allow the proof to carry forward to the more general case used here.

1.2 Problem Formulation

In this paper we look at two incarnations of the sparse approximation problem. The ℓ_0 regularised optimisation problem is defined as follows: for given $\mathbf{x}$ and Φ, find coefficients $\mathbf{y}$ minimising the cost function

$$C_{\ell_0}(\mathbf{y}) = \|\mathbf{x} - \Phi\mathbf{y}\|_2^2 + \lambda \|\mathbf{y}\|_0, \tag{1.5}$$

where $\|\mathbf{y}\|_0$ is defined as $|\Gamma_1(\mathbf{y})|$ and throughout this paper we use $\Gamma_1(\mathbf{y}) = \{y_i : y_i \neq 0\}$ as the set of non-zero coefficients. $|\Gamma_1(\mathbf{y})|$ is the size of this set so that $\|\mathbf{y}\|_0$ counts the number of non-zero coefficients.

The M-sparse problem is a constrained optimisation problem of the form

$$\min_{\mathbf{y}} \|\mathbf{x} - \Phi\mathbf{y}\|_2^2 \text{ subject to } \|\mathbf{y}\|_0 \leq M, \tag{1.6}$$

i.e now we constrain the number of non-zero coefficients to be below a certain value[1].

Solving (1.6) is known to be NP-hard in general [9, 10]. Therefore, two common themes have been adopted to approximately solve the problem, greedy optimisation strategies and relaxation of the cost function. Greedy strategies, such as Matching Pursuit (MP) type algorithms [11], are relatively fast iterative procedures that have been used extensively in practical applications. The performance of these methods is however not guaranteed in general and only under very strict conditions can they be shown to optimise the above cost function [12] [13]. Relaxation methods replace the $\|\mathbf{y}\|_0$ constraint by an almost everywhere differentiable and often convex cost function as in the the Basis Pursuit De-noising method [15] or the more general models optimised by the FOCUSS algorithm [14]. These approaches offer better performance in many cases, but can also be computationally more demanding.

[1] It is worth pointing out that the two problems are related in that there exist a λ, which depends on $\mathbf{x}$ and M, such that the solution to the ℓ_0 regularised problem is the same as that to the M-sparse problem. However, it is also important to realise that the algorithms derived here can have quite different performance, even though they have a very similar structure.

Basis Pursuit De-noising [15] relaxes the ℓ_0 penalty and replaces it with the convex ℓ_1 penalty. This leads to the convex optimisation problem

$$\min_{\mathbf{y}} \|\mathbf{x} - \Phi\mathbf{y}\|_2^2 + \lambda\|\mathbf{y}\|_1. \tag{1.7}$$

Recently, iterative thresholding algorithms have been proposed to solve this problem [16], [17], [18], [19], [20] and [21]. A similar algorithm to directly solve the ℓ_0 regularised optimisation problem had previously been put forward by Kingsbury in [22] and more recently, a slight variation of this was used in [23]. All of these methods are particular instances of a more general class of iterative thresholding algorithms [24], [25] and [26]. A good general overview over iterative thresholding methods can be found in [27]. Related convergence results can also be found in [28].

1.3 Paper Overview

In section 2 we derive the algorithm used in [23] using ideas from [16] and in section 3 a novel variation of the algorithm is derived to solve the M-sparse problem. Importantly, we present the following novel results with regard to both algorithms.

- We show that the algorithms are guaranteed not to increase the cost functions (1.5) and (1.6) respectively.
- We give conditions specifying the fixed points of the algorithms.
- We give a simple condition guaranteeing the convergence of the methods to local optima of the cost functions (1.5) and (1.6).
- We analyse the convergence speed of the methods.
- We give bounds on the error and the number of non-zero elements of the fixed points.

These two algorithms work directly on the cost functions (1.5) and (1.6). As these functions are non-convex, we find that the algorithms only converge to local optima. Even worse, we find that the fixed points of the first algorithm are not guaranteed to be sparse. Numerical studies confirm that the initialisation of the methods is important. For example, when initialising the coefficients with zero, the algorithms were often found to perform worse than Matching Pursuit. We therefore suggest two strategies for a successful application of the methods. The first strategy is to use the methods in conjunction with other methods such as Matching Pursuit or Basis Pursuit De-noising. The solutions found with these algorithms are in general not even local optima of (1.5) and (1.6). In fact, the solutions can always be improved by orthogonally projecting the signal onto the space defined by the non-zero components. This projection is typically done using a conjugate gradient algorithm. By replacing the conjugate gradient algorithm with the methods proposed here one does not only calculate such a

projection. More importantly, the proposed algorithms can also change the support set of the solution, while at the same time guaranteeing to improve it. This is shown numerically in subsections 4.1 and 4.2.

The other suggested approach is a slight modification of the algorithms. This method varies the number of retained coefficients in each iteration, starting with a single coefficient and adding additional ones as the algorithm progresses. On one particular example, we show that if the number of non-zero coefficients is increased in each iteration, the performance is comparable to Matching pursuit, whilst by increasing the number of non-zero coefficients more slowly allows us to improve the performance. This approach is studied in subsection 4.3.

This paper can be read on three different levels: the casual reader, interested in the algorithms and their properties, but less curious about the more formal statements of these properties nor in the exact derivation of the algorithms can read the digest subsections given at the beginning of the next two sections; more formal derivations of the algorithms and statements of the main theorems and lemmata comprise the rest of the next two sections; the keen reader, interested also in the proofs, is referred to the appendices.

2. An iterative algorithm for the ℓ_0 regularised problem

2.1 Digest: the Iterative Hard-Thresholding Algorithm

To solve the optimisation problem

$$\min_{\mathbf{y}} \|\mathbf{x} - \Phi\mathbf{y}\|_2^2 + \lambda\|\mathbf{y}\|_0, \tag{2.1}$$

we derive the following iterative algorithm

$$\mathbf{y}^{n+1} = H_{\lambda^{0.5}}(\mathbf{y}^n + \Phi^H(\mathbf{x} - \Phi\mathbf{y}^n)), \tag{2.2}$$

where $H_{\lambda^{0.5}}$ is the element wise hard thresholding operator

$$H_{\lambda^{0.5}}(y_i) = \begin{cases} 0 & \text{if } |y_i| \leq \lambda^{0.5} \\ y_i & \text{if } |y_i| > \lambda^{0.5}. \end{cases} \tag{2.3}$$

This algorithm will be called the *iterative hard-thresholding algorithm*. We show that under the assumption that $\|\Phi\|_2 < 1$ the algorithm is guaranteed not to increase (1.5) and in fact converges to a local minimum of (1.5). Furthermore, the asymptotic convergence rate is linear and, assuming the set $\{\phi_i\}$ contains a basis for the signal space and $\|\phi_i\|_2 > c > 0$, then at any fixed point $\mathbf{y}^\star$ the error satisfies the bound

$$\|\mathbf{x} - \Phi\mathbf{y}^\star\|_2 \leq \frac{\lambda^{0.5}}{\beta(\Phi)}, \tag{2.4}$$

where $\beta(\Phi)$ is a constant depending only on Φ.

2.2 Optimisation Transfer

Instead of optimising (1.5), let us introduce a surrogate objective function, as proposed in [29].

$$C^S_{\ell_0}(\mathbf{y},\mathbf{z}) = \|\mathbf{x}-\Phi\mathbf{y}\|_2^2 + \lambda\|\mathbf{y}\|_0 - \|\Phi\mathbf{y}-\Phi\mathbf{z}\|_2^2 + \|\mathbf{y}-\mathbf{z}\|_2^2. \tag{2.5}$$

If $\|\Phi\|_2 < 1$, then this surrogate objective function is a majorisation of the objective function and minimisation of the surrogate function leads to a majorisation minimisation (MM) algorithm [30]. Because $C_{\ell_0}(\mathbf{y}) = C^S_{\ell_0}(\mathbf{y},\mathbf{y})$, optimising (2.5) with respect to $\mathbf{y}$ will then decrease the original cost function (1.5). The surrogate objective function (2.5) can be rewritten as

$$C^S_{\ell_0}(\mathbf{y},\mathbf{z}) = \sum_i [y_i^2 - 2y_i(z_i + \phi_i^H\mathbf{x} - \phi_i^H\Phi\mathbf{z}) + \lambda|y_i|^0] + \|\mathbf{x}\|_2^2 + \|\mathbf{z}\|_2^2 - \|\Phi\mathbf{z}\|_2^2,$$

where $|y_i|^0$ is one if $y_i \neq 0$ and zero otherwise. Now the y_i are decoupled. Therefore, the minimum of (2.5) can be calculated by minimising with respect to each y_i individually. To derive the minimum, we distinguish two cases, $y_i = 0$ and $y_i \neq 0$. In the first case, the element wise cost is (ignoring the constant terms) 0. In the second case the cost is (again ignoring the constant terms) $y_i^2 - 2y_i(z_i + \phi_i^H\mathbf{x} - \phi_i^H\Phi\mathbf{z}) + \lambda$, the minimum of which is achieved at $y_i^\star = z_i + \phi_i^H\mathbf{x} - \phi_i^H\Phi\mathbf{z}$.

Comparing the cost for both cases, i.e

$$\begin{aligned} 0 \quad &\text{if} \quad y_i = 0 \\ -(z_i + \phi_i^H\mathbf{x} - \phi_i^H\Phi\mathbf{z})^2 + \lambda \quad &\text{if} \quad y_i = z_i + \phi_i^H\mathbf{x} - \phi_i^H\Phi\mathbf{z}, \end{aligned}$$

we see that the minimum of (2.5) is attained at

$$\mathbf{y} = H_{\lambda^{0.5}}(\mathbf{z} + \Phi^H(\mathbf{x} - \Phi\mathbf{z})),$$

where we use the element-wise hard thresholding operator given in (2.3). Note that the minimum need not be unique whenever $z_i + \phi_i^H\mathbf{x} - \phi_i^H\Phi\mathbf{z} = \lambda^{0.5}$. However, using a strict inequality in the definition of the thresholding operator as done here guarantees a unique update.

The iterative hard thresholding algorithm is now defined as

$$\mathbf{y}^{n+1} = H_{\lambda^{0.5}}(\mathbf{y}^n + \Phi^H(\mathbf{x} - \Phi\mathbf{y}^n)). \tag{2.6}$$

In the rest of this section we will often simplify the notation and introduce the non-linear operator $T\mathbf{y} = H_\theta(\mathbf{y} + \Phi^H(\mathbf{x} - \Phi\mathbf{y}))$. This algorithm, previously used in [23], is a thresholded version of the well known Landweber iteration [31]. However, we found that this algorithm is not stable in general. In this paper we show that a sufficient requirement for the above algorithm

to converge is that the eigenvalues of the linear operator $(\mathbf{I} - \Phi^H\Phi)$ are $0 < eig(\mathbf{I} - \Phi^H\Phi) \leq 1$. As pointed out by one of the reviewers to an earlier version of this manuscript, the bound form above is satisfies because $\Phi^H\Phi$ is positive semi-definite. To satisfy the lower bound, we then require that $\|\Phi\|_2 < 1$.

2.3 Relationship Between Optimisation of the Surrogate Function and the Original Cost Function

In this subsection we give an important lemma

Lemma 1. *Assume* $\|\Phi\|_2 < 1$ *and let* $\mathbf{y}^{n+1} = H_\theta(\mathbf{y}^n + \Phi^H(\mathbf{x} - \Phi\mathbf{y}^n))$*, then the sequences* $(C_{\ell_0}(\mathbf{y}^n))_n$ *and* $(C^S_{\ell_0}(\mathbf{y}^{n+1}, \mathbf{y}^n))_n$ *are non-increasing.*

This lemma is a trivial consequence following from the majorisation minimisation (MM) framework [30]. A proof is included in appendix A for completeness.

This lemma states that the cost function (1.5) does not increase from iteration to iteration, or, more bluntly, using the algorithm cannot lead to worse results than not using the algorithm.

2.4 Specifying the Fixed Points

As the algorithm has multiple fixed points, it is important to analyse these in more detail.

Lemma 2. *Define the sets* $\Gamma_0 = \{i : y^\star_i = 0\}$ *and* $\Gamma_1 = \{i : |y^\star_i| > \lambda^{0.5}\}$*. A necessary and sufficient condition for a point* $\mathbf{y}^\star$ *to be a fixed point of algorithm (2.2) is that for each* ϕ_i*,*

$$|\phi_i^H(\mathbf{x} - \Phi\mathbf{y}^\star)| \begin{cases} = 0 & \text{if } i \in \Gamma_1 \\ \leq \lambda^{0.5} & \text{if } i \in \Gamma_0. \end{cases}$$

The proof of this result is straightforward and we give it here as it has some merit in itself.

Proof. A fixed point is any $\mathbf{y}^\star$ such that $\mathbf{y}^\star = T(\mathbf{y}^\star)$. Looking at this equality element wise and inserting the algorithm we have

$$y^\star_i = H_{\lambda^{0.5}}(y^\star_i + \phi_i^H(\mathbf{x} - \Phi\mathbf{y}^\star)).$$

If $y^\star_i = 0$, this equality holds if and only if $|\phi_i^H(\mathbf{x} - \Phi\mathbf{y}^\star)| \leq \lambda^{0.5}$. Similarly for $i \in \Gamma_1$ we have

$$y^\star_i = y^\star_i + \phi_i^H(\mathbf{x} - \Phi\mathbf{y}^\star),$$

where we have dropped the thresholding operator, as $y^\star_i \neq 0$. Again, this holds if and only if $\phi_i^H(\mathbf{x} - \Phi\mathbf{y}^\star) = 0$. □

One of the main results in this section relates the fixed points with the cost function (1.5)

Lemma 3. *Assume $\|\Phi\|_2 < 1$, then a fixed point $\mathbf{y}^\star = T\mathbf{y}^\star$ is a local minimum of (1.5).*

The proof of this lemma is a bit more involved and can be found in appendix B. By a local minimum we mean that perturbing $\mathbf{y}^\star$ by an infinitesimal small amount (in any direction) will not decrease the cost function.

Before proceeding, we answer another important question, whether the set of fixed points includes the optimal solution. To show that this is in fact true we appeal to theorem 12 in [7], which we enhance here by adding an additional property.

Theorem 1 (Tropp [7]: Theorem 12). *For an input signal $\mathbf{x}$ and a threshold $\lambda^{0.5}$, denote by $\mathbf{y}^{opt}$ the global minimum of the optimisation problem (1.5). Define $\Gamma_0 = \{i : y_i^{opt} = 0\}$ and $\Gamma_1 = \{i : y_i^{opt} \neq 0\}$.*

- $\forall i \in \Gamma_1,\ |y_i^{opt}| \geq \lambda^{0.5}$
- $\forall i \in \Gamma_0,\ \phi_i^H(\mathbf{x} - \Phi\mathbf{y}^{opt}) \leq \lambda^{0.5}$
- $\forall i \in \Gamma_1,\ \phi_i^H(\mathbf{x} - \Phi\mathbf{y}^{opt}) = 0,$

The third condition implies that the error is orthogonal to the atoms ϕ_i^H when $i \in \Gamma_1$. In other words, the signal is projected orthogonally onto the space spanned by these atoms. This condition is not given in [7] but can be proven easily by the argument in appendix C.

Comparing the conditions in theorem 1 to the fixed point conditions of the algorithm in lemma 2 we have

Theorem 2. *Assume $\|\Phi\|_2 < 1$, then the optimal solution to the optimisation problem (1.5) belongs to the fixed points of the iterative algorithm defined by (2.2).*

2.5 Convergence

We have shown in the previous section that the iterative hard thresholding algorithm is guaranteed not to increase the cost function in (1.5). In this subsection we state an even more important property of the algorithm, namely, the algorithm converges to a local minimum of (1.5).

More formally, we have the following theorem

Theorem 3. *If $C_{\ell_0}(\mathbf{y}^0) < \infty$ and if $\|\Phi\|_2 < 1$, then the sequence $(\mathbf{y}^n)_n$ defined by the iterative procedure in (2.2) converges to a local minimum of (1.5).*

Note that the condition $C_{\ell_0}(\mathbf{y}^0) < \infty$, which we use in lemma D.1, is only importance in infinite dimensional spaces, where it implies that only a

finite number of y_i^0 are non-zero at any one time. The proof of the above theorem is given in appendix D.

2.6 Bounds on error and number of non-zero coefficients

Lemma 2 gives a necessary and sufficient condition. We can, for example, choose a subset of $\{\phi_i\}$ which constitute a basis for the signal space. In this case, we can choose $\mathbf{y}_\Gamma$ such that $\mathbf{x} - \Phi_\Gamma \mathbf{y}_\Gamma = 0$. This solution satisfies the necessary condition of the lemma which means that the fixed points of the algorithm are not guaranteed to be sparse.

Note also that the algorithm is guaranteed to find a local minimum of the cost function. This implies that the local minima of the cost function are not required to be sparse either.

Clearly, having no guarantee on the sparsity of the solution is not in general a desirable property of an algorithm to find sparse representations. One would therefore want to modify the algorithm to impose constraints on the number of non-zero coefficients of the solution. Such an approach is studied in the next section. In the rest of this section we will present a bound on the reconstruction error achieved with the proposed method.

A bound on the approximation error is stated in

Lemma 4. *Assume that $\|\Phi\|_2 \leq 1$, that $\{\phi_i\}$ contains a basis for the signal space and that $\|\phi_i\|_2 > c > 0$, then a tight bound for the approximation error $\|\mathbf{x} - \Phi\mathbf{y}^\star\|_2$ achieved at a fixed point $\mathbf{y}^\star$ is*

$$\|\mathbf{x} - \Phi\mathbf{y}^\star\|_2 \leq \frac{\lambda^{0.5}}{\beta(\Phi)}, \tag{2.7}$$

where $\beta(\Phi) > 0$ is such that $\sup_i |\phi_i^H \mathbf{x}| \geq \beta(\Phi)\|\mathbf{x}\|_2$ holds for all $\mathbf{x}$.

The proof can be found in appendix E.

2.7 Speed of convergence

The convergence proof of the algorithm relied on the fact that after a finite number of iterations, the selected subset remains fixed, in which case the algorithm simplifies to the standard Landweber iteration [31]. Therefore the asymptotic convergence speed is the linear convergence of the Landweber algorithm [31] given by

$$\|\mathbf{y}^n - \mathbf{y}^\star\|_2 \leq \|\mathbf{I} - \Phi_{\Gamma_1}^H \Phi_{\Gamma_1}\|_2^{(n-m)} \|\mathbf{y}^m - \mathbf{y}^\star\|_2 \tag{2.8}$$

Note that we have expressed this result in terms of the matrices $\Phi_{\Gamma_1}^H \Phi_{\Gamma_1}$ containing the inner products between the $\{\phi_i\}$ associated with non-zero elements. Assuming that $\|\phi_i\|_2 = c$ for all i and that $c^2 > \mu_1(M-1)$,

where M is the size of the set Γ_1, we can use results form [12] to bound the eigenvalues of $\mathbf{I} - \Phi_{\Gamma_1}^H \Phi_{\Gamma_1}$ with the cumulative coherence, leading to

$$\|\mathbf{y}^n - \mathbf{y}^\star\|_2 \leq [1 - (c^2 - \mu_1(M-1))]^{\frac{n-m}{2}} \|\mathbf{y}^m - \mathbf{y}^\star\|_2. \tag{2.9}$$

Because the cumulative coherence is an increasing function of M, it can be seen that the bound decreases, the smaller the selected sub-dictionary.

3. An iterative algorithm for the M-sparse problem

3.1 Digest: the M-Sparse Algorithm

In this section we turn to the M-sparse problem

$$\min_{\mathbf{y}} \|\mathbf{x} - \Phi\mathbf{y}\|_2^2 \text{ subject to } \|\mathbf{y}\|_0 \leq M, \tag{3.1}$$

and derive the following iterative algorithm

$$\mathbf{y}^{n+1} = H_M(\mathbf{y}^n + \Phi^H(\mathbf{x} - \Phi\mathbf{y}^n)), \tag{3.2}$$

where H_M is now a non-linear operator that only retains the M coefficients with the largest magnitude

$$H_M(y_i) \begin{cases} 0 & \text{if } |y_i| < \lambda_M^{0.5}(\mathbf{y}) \\ y_i & \text{if } |y_i| \geq \lambda_M^{0.5}(\mathbf{y}). \end{cases} \tag{3.3}$$

The threshold $\lambda_M^{0.5}(\mathbf{y})$ is set to the M^{th} largest absolute value of $\mathbf{y}^n + \Phi^H(\mathbf{x} - \Phi\mathbf{y}^n)$, if less than M values are non-zero we define $\lambda_M^{0.5}(\mathbf{y})$ to be the smallest absolute value of the non-zero coefficients. We call this algorithm the *M-sparse algorithm*.

If $\|\Phi\|_2 < 1$ and assume $\{\phi_i\}$ contains a basis for the signal space and $\|\phi_i\|_2 > c > 0$, then the algorithm is guaranteed not to increase (1.6) and converges to a local minimum of (1.6). As before, the asymptotic convergence rate is linear. If $\|\phi_i\|_2 = c$ for all i and if $c^2 > \mu_1(M-1)$, then at the fixed point $\mathbf{y}^\star$ the error satisfies the bound

$$\|\mathbf{x} - \Phi\mathbf{y}^\star\|_2 \leq \frac{c\|\mathbf{x}\|_2}{c^2 - \mu_1(M-1)}. \tag{3.4}$$

3.2 Optimisation Transfer

We again use optimisation transfer to derive the iterative algorithm. The surrogate objective function is then

$$C_M^S(\mathbf{y}, \mathbf{z}) = \|\mathbf{x} - \Phi\mathbf{y}\|_2^2 - \|\Phi\mathbf{y} - \Phi\mathbf{z}\|_2^2 + \|\mathbf{y} - \mathbf{z}\|_2^2. \tag{3.5}$$

In order for the surrogate cost to majorise the cost function we again require that $\|\Phi\|_2 < 1$. Note that we do not use a regularisation term here, however, in the minimisation of the surrogate cost function we now require that the constraint $\|\mathbf{y}\|_0 \leq M$ is satisfied. We again have $C_M(\mathbf{y}) = C_M^S(\mathbf{y}, \mathbf{y})$. As in the previous section we write equation (3.5) as

$$C_M^S(\mathbf{y}, \mathbf{z}) \propto \sum_i [y_i^2 - 2y_i(z_i + \phi_i^H \mathbf{x} - \phi_i^H \Phi \mathbf{z})].$$

This again de-couples the y_i. If we ignore the constraint on the number of non-zero coefficients we would get the standard Landweber minimum of

$$y_i^\star = z_i + \phi_i^H \mathbf{x} - \phi_i^H \Phi \mathbf{z}.$$

At this minimum, the cost function would be

$$C_M^S(\mathbf{y}^\star, \mathbf{z}) \propto \sum_i [y_i^{\star 2} - 2y_i^\star(z_i + \phi_i^H \mathbf{x} - \phi_i^H \Phi \mathbf{z})] = \sum_i -y_i^{\star 2}.$$

The *constrained* minimum of the surrogate cost function is then achieved by choosing the M largest (in absolute value) coefficients $y_i^\star$.

The minimum of (2.5) is thus attained at

$$\mathbf{y} = H_M(\mathbf{z} + \Phi^H(\mathbf{x} - \Phi \mathbf{z})),$$

where now the thresholding operator H_M chooses the threshold depending on its argument.

The iterative M-sparse algorithm is therefore

$$\mathbf{y}^{n+1} = H_M(\mathbf{y}^n + \Phi^H(\mathbf{x} - \Phi \mathbf{y}^n)). \tag{3.6}$$

Again, a sufficient requirement for the above algorithm to converge is that the eigenvalues of the linear operator $(\mathbf{I} - \Phi^H \Phi)$ are $0 < eig(\mathbf{I} - \Phi^H \Phi) \leq 1$, that is $\|\Phi\|_2 < 1$.

3.3 Relationship Between Optimisation of the Surrogate Function and the Original Cost Function

We show that the algorithm reduces the cost function $C_M(\mathbf{y}^n)$. It is important to stress that we here require $\|\mathbf{y}^0\|_0 \leq M$, which is guaranteed by the fact that we choose only the largest M coefficients in each iteration. (In cases where there are more than one coefficient with equal magnitude, such that the M largest coefficients are not uniquely defined, we assume that the algorithm selects from the offending coefficients using a predefined order.) The equivalent to lemma 1 then holds

Lemma 5. *Assume that* $\|\Phi\|_2 < 1$ *and let* $\mathbf{y}^{n+1} = H_M(\mathbf{y}^n + \Phi^H(\mathbf{x} - \Phi \mathbf{y}^n))$, *then the sequences* $(C_M(\mathbf{y}^n))_n$ *and* $(C_M^S(\mathbf{y}^{n+1}, \mathbf{y}^n))_n$ *are non-increasing.*

The proof to this lemma is exactly the same as that for lemma 1 given in appendix A, with the cost functions chosen appropriately.

3.4 Specifying the Fixed Points

It is again important to analyse the fixed points of this algorithm, which can be done using a similar approach to that taken above. Note, however, we have now a slightly different lemma.

Lemma 6. *Assume that* $\|\Phi\|_2 < 1$ *and define the sets* $\Gamma_0 = \{i : y_i^\star = 0\}$ *and* $\Gamma_1 = \{i : |y_i^\star| > \lambda_M^{0.5}(\mathbf{y}^\star)\}$*, then* $\mathbf{y}^\star$ *is a fixed point of algorithm (3.2) if and only if*

$$|\phi_i^H(\mathbf{x} - \Phi\mathbf{y}^\star)| \begin{cases} = 0 & \text{if } i \in \Gamma_1 \\ \leq \lambda_M^{0.5}(\mathbf{y}^\star) & \text{if } i \in \Gamma_0. \end{cases}$$

Furthermore, if $\{\phi_i\}$ *contains a basis for the signal space and if* $\|\phi_i\|_2 > c > 0$*, then* $\|\mathbf{y}^\star\|_0 = M$ *unless* $\mathbf{x} - \Phi\mathbf{y}^\star = 0$*, in which case* $C_M(\mathbf{y}^\star) = 0$*.*

Proof. A fixed point is defined as $\mathbf{y}^\star = T(\mathbf{y}^\star)$ and after insertion of the algorithm we get $y_i^\star = H_M(y_i^\star + \phi_i^H(\mathbf{x} - \Phi\mathbf{y}^\star))$, where the threshold $\lambda_M^{0.5}(\mathbf{y}) > 0$ depends on $\phi_i^H(\mathbf{x} - \Phi\mathbf{y}^\star)$. For $i \in \Gamma_0$, $y_i^\star = 0$, which holds if and only if $|\phi_i^H(\mathbf{x} - \Phi\mathbf{y}^\star)| < \lambda_M^{0.5}(\mathbf{y})$. For $i \in \Gamma_1$ we have $y_i^\star = y_i^\star + \phi_i^H(\mathbf{x} - \Phi\mathbf{y}^\star)$. Again, this holds if and only if $\phi_i^H(\mathbf{x} - \Phi\mathbf{y}^\star) = 0$.

Assume that $\|\mathbf{y}^\star\|_0 < M$, i.e. that $|\Gamma_1| < M$. The algorithm can only choose a threshold such that less than M elements are zero if less than M elements of $\mathbf{y}^\star + \Phi^H(\mathbf{x} - \Phi\mathbf{y}^\star)$ are non-zero. This implies that $\phi_i^H(\mathbf{x} - \Phi\mathbf{y}^\star) = 0$ for all $i \in \Gamma_0$. By the discussion above, we also require that $\phi_i^H(\mathbf{x} - \Phi\mathbf{y}^\star) = 0$ for all $i \in \Gamma_1$, i.e. $\phi_i^H(\mathbf{x} - \Phi\mathbf{y}^\star) = 0$ for all i. By assumption, the $\{\phi_i\}$ contain a basis for the signal space and $\|\phi_i\|_2 > c > 0$. There must therefore be a $\beta(\Phi) > 0$, such that $\sup_i |\phi_i^H\mathbf{r}| \geq \beta(\Phi)\|\mathbf{r}\|_2$ holds for all $\mathbf{r}$. But this implies that $C_M(\mathbf{y}^\star) = \|\mathbf{x} - \Phi\mathbf{y}^\star\|_2 = 0$. □

We can further show that

Lemma 7. *Assume* $\|\Phi\|_2 < 1$*, then a fixed point* $\mathbf{y}^\star = T\mathbf{y}^\star$ *is a local minimum of the constrained optimisation problem (1.6).*

The proof can be found in appendix F.

3.5 Convergence

We also have a convergence proof for the M-sparse algorithm

Theorem 4. *If* $\|\Phi\|_2 < 1$ *and assume* $\{\phi_i\}$ *contains a basis for the signal space and* $\|\phi_i\|_2 > c > 0$*, then the sequence* $(\mathbf{y}^n)_n$ *defined by the iterative M-sparse algorithm (3.2) converges to a local minimum of (1.6).*

The proof can be found in appendix G.

3.6 Bounds on error and sparsity

The sparsity of the fixed points of the algorithm is naturally given by the parameter M.

A similar bound on the approximation error as given in lemma 4 can be found

Lemma 8. *Assume that $\|\Phi\|_2 < 1$, that $\{\phi_i\}$ contains a basis for the signal space and that $\|\phi_i\|_2 > c > 0$, then a tight bound for the approximation error $\|\mathbf{x} - \Phi\mathbf{y}^\star\|_2$ achieved at a fixed point $\mathbf{y}^\star$ is*

$$\|\mathbf{x} - \Phi\mathbf{y}^\star\| \leq \frac{\lambda_M^{0.5}(\mathbf{y}^\star)}{\beta(\Phi)}, \tag{3.7}$$

where $\beta(\Phi) > 0$ is such that $\sup_i |\phi_i^H \mathbf{x}| \geq \beta(\Phi)\|\mathbf{x}\|_2$ holds for all $\mathbf{x}$.

The proof is identical to the proof of lemma 4. The difference is now that $\lambda^{0.5}(\mathbf{y}^\star)$ is a function of the fixed point itself. We therefore need a lemma bounding $\lambda^{0.5}(\mathbf{y}^\star)$. This is done in

Lemma 9. *Assume that for all i, $\|\phi_i\|_2 = c$. If $\mu_1(M-1) < c^2$, then we have the following bound*

$$\lambda_M^{0.5}(\mathbf{y}^\star) \leq \frac{c\|\mathbf{x}\|_2}{c^2 - \mu_1(M-1)}. \tag{3.8}$$

The proof of this lemma can be found in appendix H.

3.7 Speed of convergence

Again, if the algorithm does not converge to an exact signal representation, then after a certain number of iterations (say m), the algorithm will not change the subset (see proof of theorem 4) and we have the same convergence properties as for the algorithm of section 2

$$\|\mathbf{y}^n - \mathbf{y}^\star\|_2 \leq \|\mathbf{I} - \Phi_{\Gamma_1}^H \Phi_{\Gamma_1}\|_2^{(n-m)} \|\mathbf{y}^m - \mathbf{y}^\star\|_2, \tag{3.9}$$

and, as in subsection 2.7, under the condition that $\|\phi_i\|_2 = c$ for all i and that $c^2 > \mu_1(M-1)$,

$$\|\mathbf{y}^n - \mathbf{y}^\star\|_2 \leq [1 - (c^2 - \mu_1(M-1))]^{\frac{n-m}{2}} \|\mathbf{y}^m - \mathbf{y}^\star\|_2. \tag{3.10}$$

4. Numerical Studies

4.1 Minimising the cost function

In this subsection we study the ability of the algorithms to improve on results calculated with Matching Pursuit [11], which is reviewed in appendix I. We

chose this algorithm for comparison as it is relatively fast and therefore used in many applications. We show that the use of the above algorithms in conjunction with Matching Pursuit often leads to an improvement in the results. Note that Matching Pursuit does not give the minimum squared error solution achievable with the selected subset, which is known to be achieved by an orthogonal projection of the signal onto the selected elements. We therefore compare our results here to those found with Matching Pursuit followed by an orthogonal projection. This projection was calculated using the pseudo-inverse, however, in most situations it would be more efficient to use conjugate gradient type algorithms. The results are shown in Figures 1 (iterative hard-thresholoding) and 2 (M-sparse).

The results in Figure 1 were calculated as follows. We randomly generate 1 000 dictionaries of size 128×256 with elements distributed uniformly on the unit sphere. From each of these we randomly selected 128 elements. The coefficients were generated by drawing i.i.d. zero mean and unit variance Gaussian variables. However, values with a magnitude below $\lambda^{0.5} = \sqrt{2}\ \mathrm{erf}^{-1}(M/128)$ were set to zero. This threshold ensures that on average, only M of the coefficients were non-zero. We choose $M \in \{2, 11, 20, 29, 38, 47, 56, 65, 74, 83, 92, 101, 110, 119, 128\}$. This procedure ensured that we used the same average number of non-zero coefficients as in the experiment below while ensuring that the coefficients are above $\lambda^{0.5}$ as required by theorem 1. We repeated this procedure four times and added different levels of zero mean Gaussian noise giving Signal to Noise Ratios (SNR) of 120 dB, 80 dB, 40 dB and 0 dB.

Matching Pursuit was stopped when the minimum in the cost function $\|\mathbf{x} - \Phi\mathbf{y}\|_2^2 + \lambda\|\mathbf{y}\|_0$ was reached[2]. These coefficients were then used to initialise the iterative hard-thresholding algorithm. In Figure 1, the stars show the ratio between the signal energy and the cost function after orthogonal projection of the Matching Pursuit results (averaged over the 1 000 realisations). The squares are the results for the iterative hard-thresholding algorithm. The lower panels show the difference between the results. We here show the signal to cost ratio (as well as the difference) in dB.

The results in Figure 2 where calculated similarly, the only difference being that the coefficients were generated using M Gaussian coefficients at random locations. This time the coefficient values were not restricted in magnitude. We then run the Matching Pursuit algorithm stopping after M elements had been selected. We used these results to initialise the M-sparse algorithm. In Figure 2 the stars are the ratio between the signal energy and the cost function averages for the projected Matching Pursuit results. The squares are the results for the thresholding algorithm. The lower pan-

[2] We calculated the cost function in each iteration of Matching Pursuit and, once the cost function started to increase, we disregarded the last selected element.

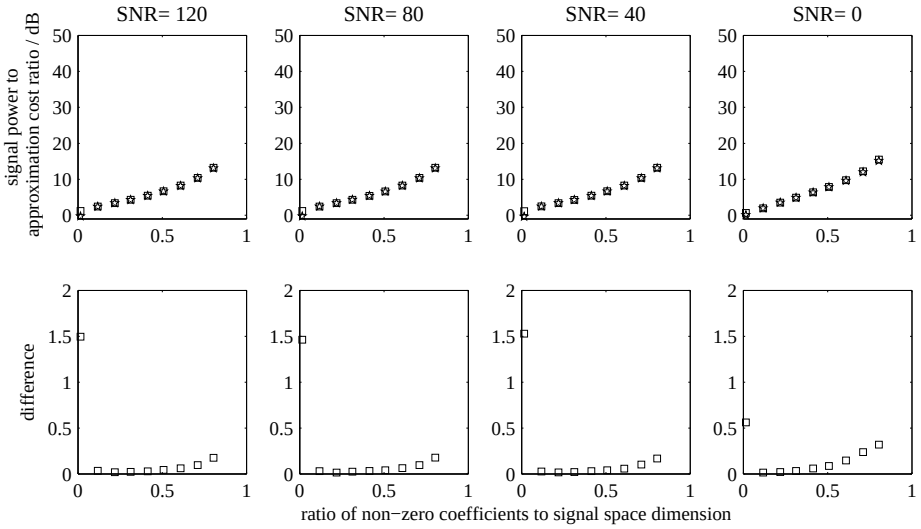

FIGURE 1 The top panels show a comparison between Matching Pursuit followed by orthogonal projection (stars) and additional use of the iterative hard-thresholding algorithm (squares) for different amounts of noise added to the signal. The y-axis shows the ratio between the signal energy and the cost function $\|\mathbf{x} - \Phi\mathbf{y}\|_2^2 + \lambda\|y\|_0$ expressed in dB and the x-axis shows the ratio of non-zero elements used to generate the signal to the signal dimension. The lower panels show the difference between the results in the upper panels.

els again show the difference between the results. We here show the ratio $\|\mathbf{x}\|_2^2/\|\mathbf{x} - \Phi\mathbf{y}\|_2^2$, again expressed in dB. In the left three panels of Figure 2, the results show a common behaviour. This is due to the algorithms being able to find the correct atoms when M is small. For increasing M, however, this happens less often and the average performance decrease until, for M approaching the dimension of $\mathbf{y}$, the performance improves again. This is due to the fact that for large M, arbitrary signals can be approximated well. On the other hand, for low SNR, the signal is dominated by the noise and there is in effect no good sparse approximation of the signal. Therefore, for small M, the approximation performance is poor as shown on the rightmost plot in Figure 2.

From these results we can draw the following conclusions

- When working with the regularised cost function, we see that, for the example used here, the iterative hard-thresholding algorithm can improve performance only marginally.
- For the M-sparse problem, we see that, for the example used here, the M-sparse algorithm improves the Matching Pursuit results significantly more than orthogonal projection onto the selected elements alone.
- In figure 4.1, it can be seen that for low M, both, Matching Pursuit and the M-sparse algorithm reach a signal approximation that has

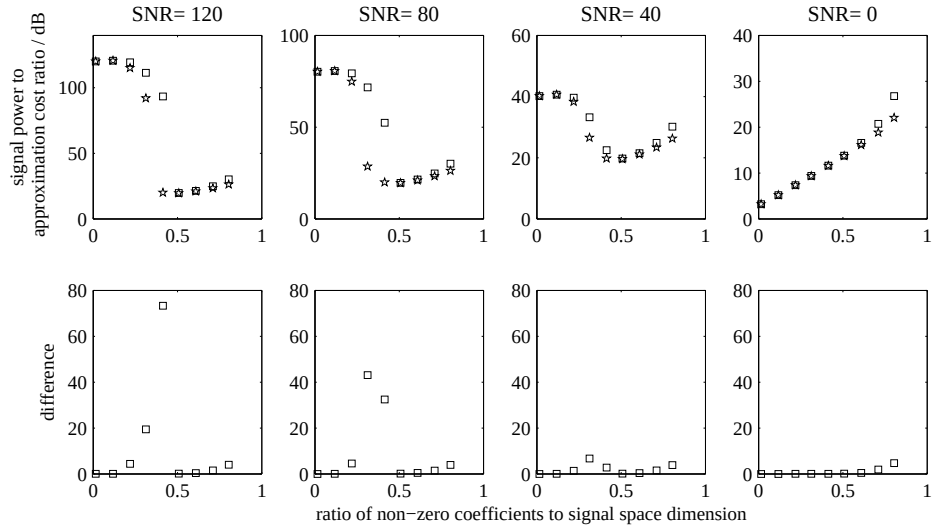

FIGURE 2 The top panels show a comparison between Matching Pursuit followed by orthogonal projection (stars) and additional use of the M-sparse algorithm (squares) for different amounts of noise added to the signal. The y-axis shows the ratio between the signal energy and the cost function $\|\mathbf{x} - \Phi\mathbf{y}\|_2^2$ expressed in dB and the x-axis shows the ratio of non-zero elements used to generate the signal to the signal dimension. The lower panels show the difference between the results in the upper panels.

the same error as the signal to noise ratio. This suggests that in these cases both algorithms often recover exactly the elements used to generate the signal. Importantly, the M-sparse algorithm was found to be able to find the exact representations even in cases in which Matching Pursuit alone failed. This behaviour is studied in more detail in the next subsection.

- Apart from the increased ability to find the exact representation, in the conducted experiment, the M-sparse algorithm also improves the SNR by several dB for only mildly sparse representations.

4.2 Exact Recovery

In some applications, such as compressive sampling [20], it is desirable to exactly recover the elements used to construct the observation $\mathbf{x}$. In order to analyse this we conducted experiments similar to those above. We generated M-sparse signals with $M \in \{30, 40, 50, 60\}$ using Gaussian coefficients, however, this time we did not add noise to the signals.

We then repeated the first experiment reported above for a range of $\lambda^{0.5}$ values. The second experiment was also a repetition of the second experiment in the previous subsection, but using $M \in \{30, 40, 50, 60\}$.

We calculated the number of correctly identified elements (True Positives) as well as the number of elements incorrectly identified as being non-

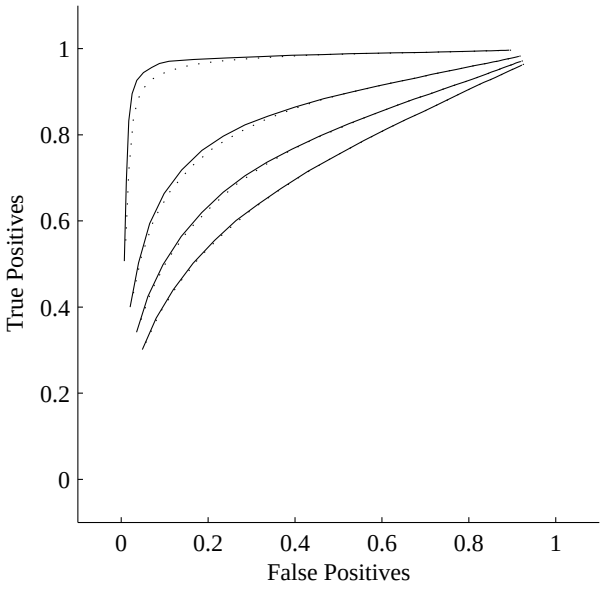

FIGURE 3 Comparison of the iterative hard-thresholding algorithm and matching pursuit in terms of the number of correctly identified elements used to generate the test signal and the number of elements identified not used to generate the test signal. We here show the results for different levels of sparsity, i.e. (from top left to bottom right) $M \in \{30, 40, 50, 60\}$. The solid lines are the results for the iterative hard-thresholding algorithm and the dotted lines are the results for Matching Pursuit.

zero (False Positives). These quantities are here normalised by the number of non-zero and zero elements in the true coefficient vector respectively.

The averaged results for the two algorithms are shown in Figures 3 and 4. The solid lines are the results from the iterative algorithms proposed in this paper, while the dotted lines are the results found with Matching Pursuit alone. The four different lines in each panel correspond to (from top left to bottom right) $M \in \{30, 40, 50, 60\}$.

These results suggest the following conclusions

- Both algorithms improve the ability to correctly identify elements.
- For less sparse signals this advantage becomes smaller.

4.3 Stepping through lambda vs. MP

Instead of calculating the pseudo inverse in each iteration in orthogonal Matching Pursuit, an iterative method could be envisaged to approximate the pseudo inverse solution. This idea is similar to a strategy suggested in [32], which can also be used for the Landweber based algorithms of this paper. The M sparse algorithm can be run for different values of M, starting from $M = 1$. After S iterations M is increased by one. If the algorithm is initialised with a zero vector and if S is large enough such that the algorithm

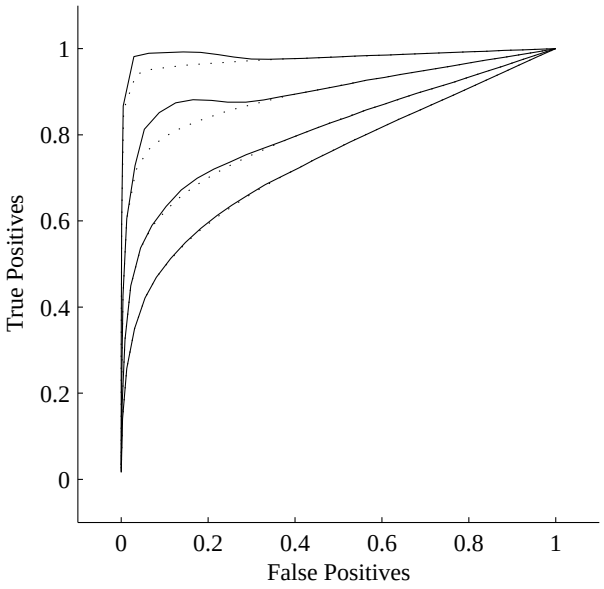

FIGURE 4 Comparison of the M-sparse algorithm and matching pursuit in terms of the number of correctly identified elements used to generate the test signal and the number of elements identified not used to generate the test signal. We here show the results for different levels of sparsity, i.e. (from top left to bottom right) $M \in \{30, 40, 50, 60\}$. The solid lines are the results for the M-sparse algorithm and the dotted lines are the results for Matching Pursuit.

converges to the minimum error solution for each M, then this strategy is very similar to orthogonal Matching Pursuit. The difference with orthogonal Matching Pursuit is that the M sparse algorithm does not necessarily use the same subset of elements at each stage.

The other extreme would be to set $S = 1$. Then the algorithm is similar to Matching Pursuit with a similar computational complexity, however, previously selected atoms are now updated in subsequent iterations. Another difference is that if the columns of Φ are not of unit length, then the value of a newly selected atom is not the same for both algorithms.

One could also use the iterative hard thresholding algorithm with a threshold depending (through some heuristic) on the current residual norm. Such a strategy would be similar to the StOMP algorithm proposed in [33], but again with the difference that we would not necessarily calculate the exact orthogonal projection for each threshold and that we allow the set of selected elements to change from iteration to iteration.

To test these ideas, we generated 1 000 signals $\mathbf{x}$ by randomly choosing 64 elements from Φ generated as above, again without added noise. We then averaged the performance of Matching Pursuit, Orthogonal Matching Pursuit and the strategy in which the number of retained elements is increased by one every $S \in \{1, 2, 5, 10, 50\}$ iterations. The approximation error in dB

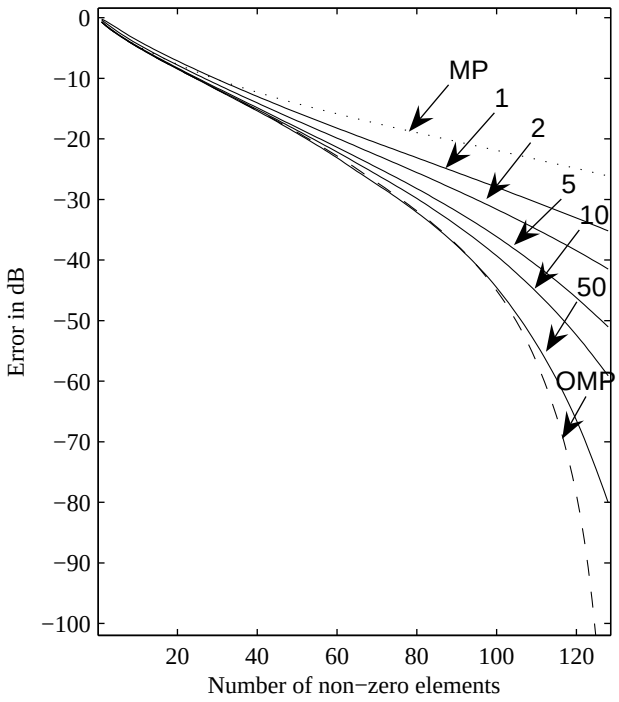

FIGURE 5 Signal to Noise Ratio in dB for different numbers of non-zero elements. The dotted line are the results with the matching pursuit algorithm and the dashed line are the results obtained with orthogonal matching pursuit. The solid lines are the results of the iterative approach in which the number of retained coefficients is increased stepwise any S iterations. The numbers in the figure indicate which S was used for each curve. The results are averaged over 100 randomly generated dictionaries $\Phi \in \mathbb{R}^{128\times 256}$ and signals with 64 non-zero Gaussian coefficients.

is shown in Figure 5 for approximations with a varying number of non-zero coefficients.

The following observations can be made for the particular example used here

- For $S = 1$, the proposed method shows marginally worse performance than Matching Pursuit if less than 30 coefficients are non-zero. If more than 30 coefficients are non-zero, the proposed method outperforms Matching Pursuit.
- If S is increased, the performance was also found to increase and for $S = 2$ the method outperforms Matching Pursuit if more than 18 elements are non-zero.
- Using $S = 50$ we see that the proposed algorithm outperforms even orthogonal Matching Pursuit if the number of non-zero values is between 25 and 94. This is a sign that the algorithm is not just an iterative orthogonal Matching Pursuit implementation.

5. Conclusion

In this paper we derived two algorithms that operate directly on the ℓ_0 regularised cost function and the M-sparse constrained optimisation problem, respectively. To our knowledge, these are the only algorithms (apart from exhaustive search), that have this property[3]. We have derived novel theoretical results for the methods. These results reveal that the algorithms have multiple fixed points making a straightforward application difficult. However, we here argued for the use of the algorithms in two contexts. Firstly, the algorithms can be used to improve the results calculated with other methods such as Matching Pursuit. In this case we have shown that the algorithms might offer benefits that cannot be explained by orthogonal projection onto the selected elements alone, i.e. they often also discover better sets of elements to describe the signal. In our experiments, the improved performance was apparent both in terms of the cost function of interest as well as in terms of identification of the elements that were used to generate the signal. Secondly, by running the M-sparse algorithm for increasing values of M, we have shown that the method can be used on its own. In the example used here, the performance was found to range from the performance of Matching Pursuit to that of orthogonal Matching Pursuit and was found to even beat the latter in certain circumstances.

Most importantly, the methods are comparable to Matching Pursuit in the computational cost of each iteration. Furthermore, many of the fast computational techniques suggested for Matching Pursuit [34] can be used also for the proposed algorithms. The proposed algorithms are therefore applicable to very large signals and dictionaries so that they can potentially be used in many real-world applications.

A. Proof of lemma 1 and lemma 5

Proof.

$$\begin{aligned} C(\mathbf{y}^{n+1}) &\leq C(\mathbf{y}^{n+1}) + \|(\mathbf{y}^{n+1} - \mathbf{y}^{n})\|_2^2 - \|\Phi(\mathbf{y}^{n+1} - \mathbf{y}^{n})\|_2^2 \\ &= C^S(\mathbf{y}^{n+1}, \mathbf{y}^{n}) \\ &\leq C^S(\mathbf{y}^{n}, \mathbf{y}^{n}) \\ &= C(\mathbf{y}^{n}) \\ &\leq C(\mathbf{y}^{n}) + \|(\mathbf{y}^{n} - \mathbf{y}^{n-1})\|_2^2 - \|\Phi(\mathbf{y}^{n} - \mathbf{y}^{n-1})\|_2^2 \\ &= C^S(\mathbf{y}^{n}, \mathbf{y}^{n-1}). \end{aligned}$$

[3] Though it is now well known that under certain conditions, other algorithms can also find the optimal solutions to the problems studied here.

The first inequality is a consequence of the condition $\|\Phi\|_2 < 1$, which ensures that the operator $(\mathbf{I} - \Phi^H\Phi)$ is positive definite. The first equality is the definition of the surrogate cost function, which therefore majorises $C(\mathbf{y})$. The second inequality is due the fact that $\mathbf{y}^{n+1}$ is the minimiser of $C^S(\mathbf{y}, \mathbf{y}^n)$. □

B. Proof of lemma 3

Proof. Given a fixed point $\mathbf{y}^\star = T\mathbf{y}^\star$ and any small perturbation $|\partial h_i| < \epsilon$, for some $\epsilon > 0$, we show that $C_{\ell_0}(\mathbf{y}^\star + \partial h) > C_{\ell_0}(\mathbf{y}^\star)$. However, we first show $\exists\ \epsilon > 0 : \forall\ \partial h$ with $|\partial h_i| < \epsilon$ the following inequality holds

$$C^S_{\ell_0}(\mathbf{y}^\star + \partial h, \mathbf{y}^\star) \geq C^S_{\ell_0}(\mathbf{y}^\star, \mathbf{y}^\star) + \|\partial h\|_2^2.$$

$$\begin{aligned} C^S_{\ell_0}(\mathbf{y}^\star + \partial h, \mathbf{y}^\star) - C^S_{\ell_0}(\mathbf{y}^\star, \mathbf{y}^\star) = \\ \sum_i (y_i + \partial h_i)^2 - 2(y_i + \partial h_i)y_i - 2(y_i + \partial h_i)\phi_i^H(\mathbf{x} - \Phi\mathbf{y}^\star) \\ - y_i^2 + 2y_i^2 + 2y_i\phi_i^H(\mathbf{x} - \Phi\mathbf{y}^\star) - \lambda|y_i|^0 + \lambda|y_i + \partial h_i|^0. \end{aligned}$$

After simplification of the above equation, we split the summation into two parts, one for $\Gamma_0 = \{i : y_i = 0\}$ and one for $\Gamma_1 = \{i : y_i \neq 0\}$. We get

$$\begin{aligned} C^S_{\ell_0}(\mathbf{y}^\star + \partial h, \mathbf{y}^\star) - C^S_{\ell_0}(\mathbf{y}^\star, \mathbf{y}^\star) = \\ \|\partial h\|_2^2 + \sum_{\Gamma_0} \lambda|\partial h_i|^0 - 2\partial h_i\phi_i^H(\mathbf{x} - \Phi\mathbf{y}^\star) \\ + \sum_{\Gamma_1} -2\partial h_i\phi_i^H(\mathbf{x} - \Phi\mathbf{y}^\star) \end{aligned}$$

For a fixed point $\mathbf{y}^\star$ the last line is zero as stated in lemma 2. For the summation over Γ_0 we have to consider two cases, if $\partial h_i = 0$, then this term is zero. If $\partial h_i \neq 0$, then choosing $|\partial h_i| \leq |\frac{\lambda}{2(\phi_i^H(\mathbf{x} - \Phi\mathbf{y}))}|$ guarantees the non-negativity of this term. Note that we also need the condition that $|\partial h_i| \leq \mathbf{y}_i$ for all $i \in \Gamma_1$ such that $y_i - \partial h_i \neq 0$. This condition is required when splitting the cost function $|y_i + \partial h_i|^0$. Therefore $\exists\ \epsilon : \forall\ \partial h, |\partial h_i| \leq \epsilon, C^S_{\ell_0}(\mathbf{y}^\star + \partial h, \mathbf{y}^\star) \geq C^S_{\ell_0}(\mathbf{y}^\star, \mathbf{y}^\star) + \|\partial h\|_2^2$. Using this we get

$$\begin{aligned} C_{\ell_0}(\mathbf{y}^\star + \partial h) &= C^S(\mathbf{y}^\star + \partial h, \mathbf{y}^\star) - \|\partial h\|_2^2 + \|\Phi\partial h\|_2^2 \\ &\geq C^S_{\ell_0}(\mathbf{y}^\star + \partial h, \mathbf{y}^\star) - \|\partial h\|_2^2 \geq C^S_{\ell_0}(\mathbf{y}^\star, \mathbf{y}^\star) = C_{\ell_0}(\mathbf{y}^\star) \end{aligned}$$

□

C. Proof of condition 3 in theorem 1

Proof. For the subset of atoms Γ_1 in the theorem, let $\mathbf{r}^{opt} = \mathbf{x} - \Phi\mathbf{y}^{opt}$ and split the error into the error within the subspace spanned by Φ_{Γ_1} and its orthogonal complement, $\mathbf{r}^{opt} = \mathbf{r}^{opt}_{\Gamma_1} + \mathbf{r}^{opt}_{\Gamma_0}$, where $\mathbf{r}^{opt}_{\Gamma_0}$ is orthogonal to all ϕ_i when $i \in \Gamma_1$, i.e $\phi_i^H \mathbf{r}^{opt}_{\Gamma_0} = 0$ for $i \in \Gamma_1$.

We now show that $\mathbf{r}^{opt}_{\Gamma_1} = \mathbf{0}$, which proves the theorem. For any fixed Γ_1 an optimum of the cost function requires that $\|\mathbf{x} - \Phi_{\Gamma_1}\mathbf{y}_{\Gamma_1}\|_2^2 = \|\mathbf{r}\|_2^2$ is minimal. Using Pythagoras $\|\mathbf{r}\|_2^2 = \|\mathbf{r}_{\Gamma_1}\|_2^2 + \|\mathbf{r}_{\Gamma_0}\|_2^2$. For any fixed Γ_1, changing $\mathbf{y}_{\Gamma_1}$ only affects the first term. Also, by definition, $\mathbf{r}_{\Gamma_1}$ lies in the span of ϕ_i, with $i \in \Gamma_1$ and by appropriate choice of $\mathbf{y}_{\Gamma_1}$ can be set to zero, at which point the error is minimal. This holds for all sets Γ_1 and holds therefore also for the set of non-zero elements in $\mathbf{y}^{opt}$, implying that $\mathbf{r}^{opt}_{\Gamma_1} = \mathbf{0}$. □

D. Proof of Theorem 3

The proof of theorem 3 is based on the following lemma.

Lemma D.1. *$\forall\ \epsilon > 0,\ \exists\ N$ such that $\forall\ n > N,\ \|\mathbf{y}^{n+1} - \mathbf{y}^n\|_2^2 \leq \epsilon$.*

Proof. We show that $\sum_{n=1}^{N} \|\mathbf{y}^{n+1} - \mathbf{y}^n\|_2^2$ converges, which implies the lemma [35, Theorem 3.23]. This is done by showing that $\sum_{n=1}^{N} \|\mathbf{y}^{n+1} - \mathbf{y}^n\|_2^2$ is monotonically increasing and bounded. We have monotonicity by

$$\sum_{n=1}^{N-1} \|\mathbf{y}^{n+1} - \mathbf{y}^n\|_2^2 + \|\mathbf{y}^{N+1} - \mathbf{y}^N\|_2^2 \geq \sum_{n=1}^{N-1} \|\mathbf{y}^{n+1} - \mathbf{y}^n\|_2^2.$$

and boundedness follows from

$$\begin{aligned}
\sum_{n=0}^{N} \|\mathbf{y}^{n+1} - \mathbf{y}^n\|_2^2 &\leq \frac{1}{c}\sum_{n=0}^{N}(\|(\mathbf{y}^{n+1} - \mathbf{y}^n)\|_2^2 - \|\Phi(\mathbf{y}^{n+1} - \mathbf{y}^n)\|_2^2) \\
&\leq \frac{1}{c}\sum_{n=0}^{N}[C_{\ell_0}(\mathbf{y}^n) - C_{\ell_0}(\mathbf{y}^{n+1})] \\
&= \frac{1}{c}(C_{\ell_0}(\mathbf{y}^0) - C_{\ell_0}(\mathbf{y}^{N+1})) \\
&\leq \frac{1}{c}C_{\ell_0}(\mathbf{y}^0),
\end{aligned} \tag{D.1}$$

where c is a lower bound on the spectrum of the linear operator $(\mathbf{I} - \Phi^H\Phi)$, which by assumption is strictly greater than zero. The second inequality is taken from the proof of Lemma 1. □

Proof of theorem 3. In lemma D.1 take $\epsilon < \lambda$. If $|y_i^n| > \lambda^{0.5}$ and $y_i^{n+1} = 0$, then $\|\mathbf{y}^{n+1} - \mathbf{y}^n\|_2^2 \geq \lambda$, which by lemma D.1 is impossible for $n > N$ for some N. Therefore, for large N, the set of zero and non-zero coefficients will not change and $|y_i^n| > \lambda^{0.5}, \forall i \in \Gamma_1, n > N$. For $y_i^n, i \in \Gamma_1$, the algorithm then reduces to the standard Landweber algorithm with guaranteed convergence [31]. Note that the largest (smallest) eigenvalue of $(\mathbf{I} - \Phi^H\Phi)$ will not increase (decrease) if we delete columns from Φ (see for example Theorem 4.3.15 in [36]) ensuring that the eigenvalue constraint required for the Landweber convergence is satisfied.

Also, by lemma 3 the fixed point is a local minimum of (1.5). □

E. Proof of lemma 4

Proof. From lemma 2 we have

$$\|\Phi^H(\mathbf{x} - \Phi\mathbf{y}^\star)\|_\infty \leq \lambda^{0.5}. \tag{E.1}$$

Define $\beta(\Phi) > 0$ such that $\sup_i |\phi_i^H\mathbf{x}| \geq \beta(\Phi)\|\mathbf{x}\|_2$ holds for all $\mathbf{x}$. Then $\beta(\Phi)\|(\mathbf{x}-\Phi\mathbf{y}^\star)\|_2 \leq \|\Phi^H(\mathbf{x}-\Phi\mathbf{y}^\star)\|_\infty \leq \lambda^{0.5}$, from which the lemma follows. □

F. Proof of lemma 7

Proof. Again $\|\mathbf{y}^\star\|_0 \leq M$ due to the constraint. We want to show that

$$C_M(\mathbf{y}^\star + \partial h) \geq C_M(\mathbf{y}^\star),$$

for any small perturbation $|\partial h_i| < \epsilon$, for some $\epsilon > 0$. If we restrict the solution to the support of $\mathbf{y}^\star$, then $\mathbf{y}^\star$ is the minimum squared error solution [31]. Now the support of ∂h might contain elements that are not in the support of $\mathbf{y}^\star$. If $\|\mathbf{y}^\star\|_0 < M$, then by lemma 6 $C(\mathbf{y}^\star) = \|\mathbf{x} - \Phi\mathbf{y}^\star\|_2^2 = 0$, which is a fixed point of the algorithm. Otherwise $\|\mathbf{y}^\star\|_0 = M$, therefore, to enforce the constraint $\|\mathbf{y}^\star + \partial h\|_0 \leq M$, if ∂h contains an element not in the support of $\mathbf{y}^\star$ it also needs to include an element within the support of $\mathbf{y}^\star$ with opposite value to that in $\mathbf{y}^\star$, say $y_i^\star$. This value is strictly larger than zero, so we can chose $y_i^\star > 2\epsilon > 0$. Therefore, we can always chose ϵ small enough such that a perturbation of a fixed point with radius smaller than ϵ does not satisfy the constraint on $\|\mathbf{y}\|_0$. □

G. Proof of Theorem 4

Proof. We distinguish two possible cases.

Case 1: The support of $\mathbf{y}^n$ is the same for all $n > N$, for some N. In this case, the set of non-zero coefficients does not change and the convergence follows from the convergence of the Landweber algorithm as in the proof of theorem 3.

Case 2: There exist infinitely many n such that $\mathbf{y}^{n+1}$ and $\mathbf{y}^n$ have different support. Now lemma D.1 also holds for the M-sparse algorithm. Therefore, for any $\epsilon > 0$, there is an N such that $\forall\, n > N$, $\|\mathbf{y}^{n+1} - \mathbf{y}^n\|_2 \leq \epsilon$, which implies that $|y_i^{n+1} - y_i^n| \leq \epsilon$. With the thresholding operator $H_{\lambda_M^{0.5}}$ and using the abbreviation $\mathbf{r}^n = \mathbf{x} - \Phi\mathbf{y}^n$, we have

$$|y_i^{n+1} - y_i^n| = |H_{\lambda_M^{0.5}}(y_i^n + \phi_i^H \mathbf{r}^n) - y_i^n| \leq \epsilon.$$

For any n, we can group the indices into four disjoint sets, depending on whether y_i^{n+1} and y_i^n are zero or not. If both y_i^{n+1} and y_i^n are nonzero, then we can drop the thresholding operator such that $|\phi_i^H \mathbf{r}^n| = |y_i^{n+1} - y_i^n| \leq \epsilon$.

For the other three cases, we distinguish two possibilities. If there is any $y_\gamma^n = 0$, but $y_\gamma^{n+1} \neq 0$, then we must have $|y_\gamma^{n+1}| \leq \epsilon$, which implies that $\lambda_M^{0.5} \leq \epsilon$. In this case, for all other y_i^n and y_i^{n+1}, assume both are zero, then $|\phi_i^H \mathbf{r}^n| \leq \lambda_M^{0.5} \leq \epsilon$, while for $y_i^n \neq 0$ and $y_i^{n+1} = 0$, we have $|y_i^n| \leq \epsilon$ and $|y_i^n + \phi_i^H \mathbf{r}^n| \leq \lambda_M^{0.5} \leq \epsilon$, which implies that $|\phi_i^H \mathbf{r}^n| \leq 2\epsilon$.

It now remains to look at the other possibility in which there is no γ such that $y_\gamma^n = 0$ and $y_\gamma^{n+1} \neq 0$. This can only happen if the size of the support set decreases. By the definition of the algorithm, this implies that $y_i^n + \phi_i \mathbf{r}^n = 0$ for all $i \in \Gamma_0$, where Γ_0 is the set of zero elements in $\mathbf{y}^{n+1}$. There are now two remaining possibilities. If $y_i^n = 0$ and $y_i^{n+1} = 0$, then $\phi_i \mathbf{r}^n = 0$. Otherwise, if $y_i^n \neq 0$ but $y_i^{n+1} = 0$, then $y_i^n = -\phi_i \mathbf{r}^n$ and $|\phi_i \mathbf{r}^n| = |y_i^n| \leq \epsilon$.

We have therefore shown that for all $\epsilon > 0$ there is an N such that for all $n > N$ and for all i

$$|\phi_i^H \mathbf{r}^n| \leq 2\epsilon.$$

Furthermore, if the $\{\phi_i\}$ contain a basis for the signal space and if $\|\phi_i\|_2 > c > 0$, then there is a $\beta > 0$ such that $\sup_i \langle \phi_i, \mathbf{r}\rangle \geq \beta \|\mathbf{r}\|_2$. We can now choose a decreasing sequence of ϵ approaching zero, which means $|\phi_i^H \mathbf{r}^n|$ approaches zero. But this then implies that $\|\mathbf{r}^n\|_2$ approaches zero and any point $\mathbf{y}^\star$, such that $\mathbf{r}^\star = \mathbf{x} - \Phi\mathbf{y}^\star = \mathbf{0}$ is a fixed point of the algorithm as well as a minimum of (1.6). □

H. Proof of lemma 9

Let us first prove

Lemma H.1. *Assume that for all i, $\|\phi_i\|_2 = c$. Let $G = \Phi_\Gamma^H \Phi_\Gamma$ be a matrix with the inner products between the element ϕ_i with $i \in \Gamma$ for $|\Gamma| \leq M$ and*

suppose that $\mu_1(M-1) < c^2$. *Then*

$$\|G^{-1}\|_\infty \leq \frac{1}{c^2 - \mu_1(M-1)}. \tag{H.1}$$

Proof. Write $\hat{G} = \frac{1}{c^2}G$. Note that G is hermitian and therefore $\|G\|_1 = \|G\|_\infty$. Using the same reasoning as that in [12, equation (12)] then gives

$$\begin{aligned}\|G^{-1}\|_\infty = \frac{1}{c^2}\|\hat{G}^{-1}\|_\infty = \frac{1}{c^2}\|\sum_{k=0}^{\infty}(\mathbf{I}-\hat{G})^k\|_\infty \\ \leq \frac{1}{c^2}\sum_{k=0}^{\infty}\|(\hat{G}-\mathbf{I})\|_\infty^k = \frac{1}{c^2 - c^2\|(\hat{G}-\mathbf{I})\|_\infty} \\ \leq \frac{1}{c^2-\mu_1(M-1)}.\end{aligned} \tag{H.2}$$

□

We can now give

Proof of lemma 9. Let $\mathbf{y}^\star$ be supported on Γ. The same argument used in the proof of the last condition of theorem 1 shows that $\Phi_\Gamma \mathbf{y}^\star_\Gamma$ has to be an orthogonal projection onto the span of ϕ_i, $i \in \Gamma$. An argument similar to that in [6] shows that $\mu(M-1) < c^2$ implies that $M < \text{spark}(\Phi)$. Therefore, the unique orthogonal projection can be written using matrix notation $\mathbf{y}^\star_\Gamma = (\Phi_\Gamma^H \Phi_\Gamma)^{-1}\Phi_\Gamma^H \mathbf{x}$. By the definition of $\lambda_M^{0.5}(\mathbf{y}^\star)$, we then have

$$\begin{aligned}\lambda_M^{0.5}(\mathbf{y}^\star) &\leq \|(\Phi_\Gamma^H\Phi_\Gamma)^{-1}\Phi_\Gamma^H\mathbf{x}\|_\infty \\ &\leq \|(\Phi_\Gamma^H\Phi_\Gamma)^{-1}\|_\infty\|\Phi_\Gamma^H\mathbf{x}\|_\infty \\ &\leq \|(\Phi_\Gamma^H\Phi_\Gamma)^{-1}\|_\infty c\|\mathbf{x}\|_2.\end{aligned} \tag{H.3}$$

Proposition H.1 then bounds the first term on the right proving the lemma. □

I. Matching Pursuit and Orthogonal Matching Pursuit.

Matching Pursuit (MP) [11] is a greedy iterative algorithm that calculates a sparse approximation using the following steps

1. Initialise $\mathbf{r}^0 = \mathbf{x}, \mathbf{y}^0 = 0$
2. $\alpha_i = \frac{\langle \mathbf{r}^n, \phi_i\rangle}{\|\phi_i\|_2^2}$
3. $i_{max} = \arg_i \max |\alpha_i|$
4. $y^n_{i_{max}} = y^{n-1}_{i_{max}} + \alpha_{i_{max}}$

5. $\mathbf{r}^n = \mathbf{r}^{n-1} - \phi_{i_{max}} \alpha_{i_{max}}$
6. iterate from 2 until stopping criterion is fulfilled.

Orthogonal Matching Pursuit (OMP) [8] is a variation of MP in which in each iteration the coefficient vector is the orthogonal projection of the signal onto the dictionary elements selected up to this iteration.

1. Initialise $\mathbf{r}^0 = \mathbf{x}, \mathbf{y}^0 = 0, \Gamma_1^0 = \emptyset$
2. $\alpha_i = \frac{\langle \mathbf{r}^n, \phi_i \rangle}{\|\phi_i\|_2^2}$
3. $i_{max} = \arg_i \max |\alpha_i|$
4. $\Gamma_1^n = \Gamma_1^{n-1} \cup i_{max}$
5. $\mathbf{y}^n = \Phi_{\Gamma_1^n}^{\dagger} \mathbf{x}$
6. $\mathbf{r}^n = \mathbf{x} - \Phi \mathbf{y}^n$
7. iterate from 2 until stopping criterion is fulfilled.

Here $\Phi_{\Gamma_1^n}^{\dagger}$ is the pseudo-inverse of the sub-dictionary $\Phi_{\Gamma_1^n}$.

Orthogonal matching pursuit is normally implemented using QR or Cholesky factorisation and is computationally more demanding than Matching Pursuit. However, the projection ensures that the algorithm selects a new element in each iteration and that the error is minimal for the currently selected set of elements.

Acknowledgments

We like to thank the anonymous reviewers for their detailed and insightful comments, which helped to significantly strengthen this manuscript. This research was supported by EPSRC grant D000246/1. MED acknowledges support of his position from the Scottish Funding Council and their support of the Joint Research Institute with the Heriot-Watt University as a component part of the Edinburgh Research Partnership.

References

[1] D. L. Donoho, M. Vetterli, R. A. DeVore, and I. Daubechies, "Data compression and harmonic analysis," *IEEE Transactions on Information Theory*, vol. 44, pp. 2435–2476, Oct. 1998.

[2] S. Mallat and F. Falzon, "Analysis of low bit rate image transform coding," *IEEE Transactions on Signal Processing*, vol. 46, pp. 1027–1042, Apr. 1998.

[3] D. L. Donoho, "De-noising by soft-thresholding," *IEEE Transactions on Information Theory*, vol. 41, no. 3, pp. 613–627, 1995.

[4] M. Davies and N. Mitianoudis, "A simple mixture model for sparse overcomplete ICA," *IEE Proc.-Vision, Image and Signal Processing*, vol. 151, pp. 35–43, August 2004.

[5] T. Blumensath and M. Davies, "Sparse and shift-invariant representations of music,"

IEEE Transactions on Audio, Speech and Language Processing, vol. 14, pp. 50–57, Jan 2006.

[6] D. L. Donoho and M. Elad, "Optimally-sparse representation in general (non-orthogonal) dictionaries via l1 minimization," *Proc. Nat. Aca. Sci.*, vol. 100, pp. 2197–2202, 2003.

[7] J. Tropp, "Just relax: Convex programming methods for identifying sparse signals in noise," *IEEE Transactions on Information Theory*, vol. 51, no. 3, pp. 1031–1051, 2006.

[8] S. Mallat, *A Wavelet Tour of Signal Processing*. Academic Press, 1999.

[9] G. Davis, *Adaptive Nonlinear Approximations*. PhD thesis, New York University, 1994.

[10] B. K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM Journal on Computing*, vol. 24, pp. 227–234, Apr 1995.

[11] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3397–3415, 1993.

[12] J. A. Tropp, "Greed is good: algorithmic results for sparse approximation," *IEEE Transactions on Information Theory*, vol. 50, no. 10, pp. 2231–2242, 2004.

[13] R. Gribonval and P. Vandergheynst, "On the exponential convergence of matching pursuits in quasi-incoherent dictionaries," *IEEE Transactions on Information Theory*, vol. 52, no. 1, pp. 255–261, 2006.

[14] J. F. Murray and K. Kreutz-Delgado, "An improved FOCUSS-based learning algorithm for solving sparse linear inverse problems," in *Conf. Record of the Thirty-Fifth Asilomar Conf. on Signals, Systems and Computers*, pp. 347–351, 2001.

[15] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal of Scientific Computing*, vol. 20, no. 1, pp. 33–61, 1998.

[16] I. Daubechies, M. Defries, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Communications on Pure and Applied Mathematics*, vol. 57, pp. 1413–1457, 2004.

[17] J. Bect, L. Blanc Féraud, G. Aubert, and A. Chambolle, *Lecture Notes in Computer Sciences 3024*, ch. A l1-unified variational framework for image restoration, pp. 1–13. Springer Verlag, 2004.

[18] M. Elad, "Why simple shrinkage is still relevant for redundant representation," *IEEE Transactions on Information Theory*, vol. 52, no. 12, pp. 5559–5569, 2006.

[19] M. Elad, B. Matalon, and M. Zibulevsky, "Coordinate and subspace optimization methods for linear least squares with non-quadratic regularization," *Applied and Computational Harmonic Analysis*, vol. 23, pp. 346–367, November 2007.

[20] E. Candès and J. Romberg, "Practical signal recovery from random projections," in *Proc. SPIE Conf, Wavelet Applications in Signal and Image Processing XI*, Jan. 2005.

[21] K. Bredies and D. A. Lorenz, "Iterated hard shrinkage for minimization problems with sparsity constraints." 2006.

[22] N. G. Kingsbury, "Complex wavelets for shift invariant analysis and filtering of signals," *Journal of Applied and Computational Harmonic Analysis*, vol. 10, pp. 234–253, May 2001.

[23] K. K. Herrity, A. C. Gilbert, and J. A. Tropp, "Sparse approximation via iterative thresholding.," in *Proceedings of the Int. Conf. on Acoustics, Speech and Signal Processing*, 2006.

[24] R. Nowak and M. Figueiredo, "Fast wavelet-based image deconvolution using the EM algorithm," in *Proceedings of the 35th Asilomar Conference on Signals, Systems, and Computers*, (Monterey), November 2001.

[25] M. Figueiredo and R. Nowak, "An EM algorithm for wavelet-based image restoration," *IEEE Transactions on Image Processing*, vol. 12, no. 8, pp. 906–916, 2003.

[26] J. Starck, M. Nguyen, and F. Murtagh, "Wavelet and curvelet for image deconvolution: a combined approach," *Journal of Signal Processing,*, vol. 83, no. 10, pp. 2279–

2283, 2003.

[27] M. Elad, B. Matalon, J. Shtok, and M. Zibulevsky, "A wide-angle view at iterated shrinkage algorithms," in *SPIE (Wavelet XII)*, (San-Diego, CA, USA), August 2007.

[28] P. L. Combettes and V. R. Wajs, "Signal recovery by proximal forward-backward splitting," *SIAM Journal on Multiscale Modeling and Simulation*, vol. 4, pp. 1168–1200, November 2005.

[29] K. Lange, D. R. Hunter, and I. Yang, "Optimization transfer using surrogate objective functions," *Journal of Computational and Graphical Statistics*, vol. 9, pp. 1–20, Mar. 2006.

[30] K. Lange, *Optimization*. Springer Verlag, 2004.

[31] L. Landweber, "An iterative formula for Fredholm integrals of the first kind," *American Journal of Mathematics*, vol. 73, pp. 615–624, Jul 1951.

[32] N. G. Kingsbury and T. H. Reeves, "Iterative image coding with overcomplete complex wavelet transforms," in *Proc. Conf. on Visual Communications and Image Processing*, 2003.

[33] D. Donoho, Y. Tsaig, I. Drori, and J. Starck, "Sparse solutions of underdetermined linear equations by stagewise orthogonal matching pursuit," 2006.

[34] S. Krustulovic and R. Gribonval, "MPTK: Matching pursuit made tractable," in *Proc. Int. Conf. on Acoustic Speech and Signal Processing*, (Toulouse, France), May 2006.

[35] W. Rudin, *Principles of Mathematical Analysis*. McGraw-Hill, 1976.

[36] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge University Press, 1985.

Received ...

Revision received ...

IDCOM & Joint Research Institute for Signal and Image Processing
Edinburgh University, King's Buildings, Mayfield Road
Edinburgh EH9 3JL, UK
e-mail: thomas.blumensath@ed.ac.uk

IDCOM & Joint Research Institute for Signal and Image Processing
Edinburgh University, King's Buildings, Mayfield Road
Edinburgh EH9 3JL, UK
e-mail: mike.davies@ed.ac.uk