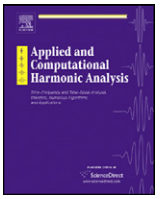
Appl. Comput. Harmon. Anal. 27 (2009) 265–274

Contents lists available at ScienceDirect

Applied and Computational Harmonic Analysis

www.elsevier.com/locate/acha



Iterative hard thresholding for compressed sensing

Thomas Blumensath *, Mike E. Davies

Institute for Digital Communications & the Joint Research Institute for Signal and Image Processing, The University of Edinburgh, King's Buildings, Mayfield Road, Edinburgh EH9 3JL, UK

ARTICLE INFO

Article history:
Received 1 October 2008
Revised 18 March 2009
Accepted 25 April 2009
Available online 3 May 2009
Communicated by Thomas Strohmer

Keywords:
Algorithms
Compressed sensing
Sparse inverse problem
Signal recovery
Iterative hard thresholding

ABSTRACT

Compressed sensing is a technique to sample *compressible* signals below the Nyquist rate, whilst still allowing near optimal reconstruction of the signal. In this paper we present a theoretical analysis of the iterative hard thresholding algorithm when applied to the compressed sensing recovery problem. We show that the algorithm has the following properties (made more precise in the main text of the paper)

- It gives near-optimal error guarantees.
- It is robust to observation noise.
- It succeeds with a minimum number of observations.
- It can be used with any sampling operator for which the operator and its adjoint can be computed.
- The memory requirement is linear in the problem size.
- Its computational complexity per iteration is of the same order as the application of the measurement operator or its adjoint.
- It requires a fixed number of iterations depending only on the logarithm of a form of signal to noise ratio of the signal.
- Its performance guarantees are uniform in that they only depend on properties of the sampling operator and signal sparsity.

1. Introduction

For more than fifty years, the Nyquist–Shannon [16,18] sampling theorem was generally used as the foundation of signal acquisition systems. Using this theory, it was a commonly held belief that signals have to be sampled at twice the signal bandwidth. Whilst this is true for general band-limited signals, this theory does not account for additional signal structures that might be known a priori. The recently emerging field of compressed sensing [3–5,7,8], and the related theory of signals with a finite rate of innovations [20], start from another premise. In compressed sensing, signals are assumed to be (approximately) sparse in some transform domain. This sparsity constraint significantly reduces the size of the set of possible signals compared to the signal space dimension.

Instead of taking samples at the Nyquist rate, compressed sensing uses linear sampling operators that map the signal into a small (compared to the Nyquist rate) dimensional space, whilst reconstruction of the signal is highly non-linear. One of the important contributions of the seminal work by Candes, Romberg, Tao [3–5] and Donoho [7], was to show that linear

* Corresponding author. Fax: +44 (0) 131 6506554.
E-mail addresses: thomas.blumensath@ed.ac.uk (T. Blumensath), mike.davies@ed.ac.uk (M.E. Davies).

1063-5203/$ – see front matter
doi:10.1016/j.acha.2009.04.002

programming algorithms can be used under certain conditions on the sampling operator to reconstruct the original signal with high accuracy.

Another set of algorithms, which could be shown to efficiently reconstruct signals from compressed sensing observations are greedy methods. A by now traditional approach is Orthogonal Matching Pursuit [10], which was analysed as a reconstruction algorithm for compressed sensing in [19]. Better theoretical properties were recently proven for a regularised Orthogonal Matching Pursuit algorithm [14,15]. Even more recently, the Compressive Sampling Matching Pursuit (CoSaMP) [13] and the nearly identical Subspace Pursuit [6] algorithms were introduced and analysed for compressed sensing signal reconstruction. Of all of these methods, CoSaMP currently offers the most comprehensive set of theoretic performance guarantees. It works for general sampling operators, is robust against noise and the performance is uniform in that it only depends on a property of the sampling operator and the sparsity of the signal, but not on the size of the non-zero signal coefficients. Furthermore, it requires minimal storage and computations and works with (up to a constant) a minimal number of observations.

In a previous paper [1], iterative hard thresholding algorithms were studied. In particular, their convergence to fixed points of ℓ_0 regularised (or constrained) cost functions could be proven. In this paper, it is shown that one of these algorithms (termed from now on IHT_s) has similar performance guarantees to those of CoSaMP.

1.1. Paper overview

Section 2 starts out with a definition of sparse signal models and a statement of the compressed sensing problem. In Section 3, we then discuss an iterative hard thresholding algorithm. The rest of this paper shows that this algorithm is able to recover, with high accuracy, signals from compressed sensing observations. This result is formally stated in the theorems of the first subsection of Section 4. The rest of Section 4 is devoted to the proof of the theorems. In fact, the derived result is near-optimal as shown in Section 5. Section 6 takes a closer look at a stopping criterion for the algorithm, which guarantees a certain estimation accuracy. The results of this paper are similar to those for the CoSaMP algorithm of [13] and a more detailed comparison is given in Section 7.

1.2. Notation

The following notation will be used in this paper. $\mathbf{y}$ is an M-dimensional real or complex vector. $\mathbf{x}$ is an N dimensional real or complex vector. Φ will denote an $M \times N$ real or complex matrix, whose transpose (Hermitian transpose) will be denoted by Φ^T. Many of the arguments in this paper will use sub-matrices and sub-vectors. The letters Γ, B and Λ will denote sets of indices that enumerate the columns in Φ and the elements in the vectors $\mathbf{x}$. Using these sets as subscripts, e.g., Φ_Γ or $\mathbf{x}_\Gamma$, we mean matrices (or vectors) formed by removing all but those columns (elements) from the matrix (vector) other than those in the set. Occasionally we also refer to quantities in a given iteration. Iterations are counted using n or k. For sets, we use the simplified notation Γ^n whilst for vectors, the iteration count is given in square brackets $\mathbf{x}^{[n]}$.

The following norms are used repeatedly. $\|\cdot\|_2$ is the Euclidean vector norm or, for matrices, the operator norm from ℓ_2 to ℓ_2. We will also need the vector ℓ_1 norm $\|\cdot\|_1$. The notation $\|\cdot\|_0$ will denote the number of non-zero elements in a vector. For a general vector $\mathbf{x}$ we use $\mathbf{x}^s$ to be any of the best s term approximations to $\mathbf{x}$. The difference between the two will be $\mathbf{x}_r = \mathbf{x} - \mathbf{x}^s$. The support, that is the index set labelling the non-zero elements in $\mathbf{x}^s$, is defined as $\Gamma_s^\star = \operatorname{supp}\{\mathbf{x}^s\}$ and similarly, $\Gamma^n = \operatorname{supp}\{\mathbf{x}^{[n]}\}$, where $\mathbf{x}^{[n]}$ is the estimation of $\mathbf{x}$ in iteration n. Finally, the set $B^n = \Gamma_s^\star \cup \Gamma^n$ is a superset of the support of the error $\mathbf{r}^{[n]} = \mathbf{x}^s - \mathbf{x}^{[n]}$. Set difference is denoted using $\cdot \setminus \cdot$.

2. Sparsity and compressed sensing

A vector is called s-spare if no more than s of its elements have non-zero values. However, most signals in the real world are not exactly sparse, instead, they are often well approximated by an s-sparse vector. Such signals are called approximately sparse. To capture this notion, we let $\mathbf{x}^s$ be the best s-sparse approximation to a vector $\mathbf{x}$, that is, any one of the s-sparse vectors that minimise $\|\mathbf{x}^s - \mathbf{x}\|_2$.

2.1. Compressed sensing

In compressed sensing the signal of interest is an $N < \infty$ dimensional real or complex vector $\mathbf{x}$ assumed to be approximately sparse. Samples are then taken using a linear mapping Φ into an M dimensional real or complex observation space. In matrix notation, the observed samples $\mathbf{y}$ are

$$\mathbf{y} = \Phi\mathbf{x} + \mathbf{e}. \tag{1}$$

Here, Φ is the linear sampling operator and $\mathbf{e}$ models possible observation noise due to, for example, sensor noise or quantisation errors in digital systems, whose norm is written as $\|\mathbf{e}\|_2 = \epsilon$.

Here, two main problems have to be addressed. The first problem, not studied in detail in this paper, is to design measurement systems Φ that possess certain desirable properties (such as the restricted isometry property of the next subsection), which allow for an efficient estimation of $\mathbf{x}$. The interested reader is referred to the extensive literature, such

as [3–5,7,11–13,17] and references therein. The second problem, which is the focus of this paper, is the study of concrete algorithms for the efficient estimation of $\mathbf{x}$ given only $\mathbf{y}$ and Φ.

2.2. The restricted isometry property

The analysis of algorithms for compressed sensing relies heavily on the following property of the observation matrix Φ. A matrix Φ satisfies the Restricted Isometry Property (RIP) [4] if

$$(1-\delta_s)\|\mathbf{x}\|_2^2 \leqslant \|\Phi\mathbf{x}\|_2^2 \leqslant (1+\delta_s)\|\mathbf{x}\|_2^2 \tag{2}$$

for all s-sparse $\mathbf{x}$ and some $\delta_s < 1$. The *restricted isometry constant* δ_s is defined as the smallest constant for which this property holds for all s-sparse vectors $\mathbf{x}$.

The following properties of the restricted isometry constant are at the heart of this paper. Proofs can be found in, for example, [13].

Lemma 1. *For all index sets Γ and all Φ for which the RIP holds with $s = |\Gamma|$*

$$\left\|\Phi_\Gamma^T\mathbf{y}\right\|_2 \leqslant \sqrt{1+\delta_{|\Gamma|}}\|\mathbf{y}\|_2, \tag{3}$$

$$(1-\delta_{|\Gamma|})\|\mathbf{x}_\Gamma\|_2 \leqslant \left\|\Phi_\Gamma^T\Phi_\Gamma\mathbf{x}_\Gamma\right\|_2 \leqslant (1+\delta_{|\Gamma|})\|\mathbf{x}_\Gamma\|_2 \tag{4}$$

and

$$\left\|\left(\mathbf{I}-\Phi_\Gamma^T\Phi_\Gamma\right)\mathbf{x}_\Gamma\right\|_2 \leqslant \delta_{|\Gamma|}\|\mathbf{x}_\Gamma\|_2. \tag{5}$$

Furthermore, for two disjoint *sets Γ and Λ (i.e. $\Gamma\cap\Lambda=\emptyset$) and all Φ for which the RIP holds with $s = |\Gamma\cup\Lambda|$*

$$\left\|\Phi_\Gamma^T\Phi_\Lambda\mathbf{x}_\Lambda\right\|_2 \leqslant \delta_s\|\mathbf{x}_\Lambda\|_2. \tag{6}$$

Lemma 2 *(Needell and Tropp, Proposition 3.5 in [13]). Suppose the matrix Φ satisfies the RIP $\|\Phi\mathbf{x}^s\|_2 \leqslant \sqrt{1+\delta_s}\|\mathbf{x}^s\|_2$ for all $\mathbf{x}^s$: $\|\mathbf{x}^s\|_0 \leqslant s$, then for all vectors $\mathbf{x}$, the following bound holds*

$$\|\Phi\mathbf{x}\|_2 \leqslant \sqrt{1+\delta_s}\|\mathbf{x}\|_2 + \sqrt{1+\delta_s}\frac{\|\mathbf{x}\|_1}{\sqrt{s}}. \tag{7}$$

Lemma 3 *(Needell and Tropp, Lemma 6.1 in [13]). For any $\mathbf{x}$, let $\mathbf{x}^s$ be (the/any) best s-term approximation to $\mathbf{x}$. Let $\mathbf{x}_r = \mathbf{x}-\mathbf{x}^s$. Let*

$$\mathbf{y} = \Phi\mathbf{x}+\mathbf{e} = \Phi\mathbf{x}^s + \Phi\mathbf{x}_r + \mathbf{e} = \Phi\mathbf{x}^s + \tilde{\mathbf{e}}. \tag{8}$$

If the RIP holds for sparsity s, then the norm of the error $\tilde{\mathbf{e}}$ can be bounded by

$$\|\tilde{\mathbf{e}}\|_2 \leqslant \sqrt{1+\delta_s}\left\|\mathbf{x}-\mathbf{x}^s\right\|_2 + \sqrt{1+\delta_s}\frac{\|\mathbf{x}-\mathbf{x}^s\|_1}{\sqrt{s}} + \|\mathbf{e}\|_2. \tag{9}$$

3. Iterative hard thresholding

3.1. Definition of the algorithm

In [1], we discussed the following Iterative Hard Thresholding algorithm (IHT_s) previously used in [9]. Let $\mathbf{x}^{[0]} = \mathbf{0}$ and use the iteration

$$\mathbf{x}^{[n+1]} = H_s\left(\mathbf{x}^{[n]} + \Phi^T\left(\mathbf{y}-\Phi\mathbf{x}^{[n]}\right)\right), \tag{10}$$

where $H_s(\mathbf{a})$ is the non-linear operator that sets all but the largest (in magnitude) s elements of $\mathbf{a}$ to zero. If there is no unique such set, a set can be selected either randomly or based on a predefined ordering of the elements. The convergence of this algorithm was proven in [1] under the condition that $\|\Phi\|_2 < 1$. In this case, the above algorithm converges to a local minimum of the optimisation problem

$$\min_{\mathbf{x}} \|\mathbf{y}-\Phi\mathbf{x}\|_2^2 \quad \text{subject to } \|\mathbf{x}\|_0 \leqslant s. \tag{11}$$

Note that a more general form of the above algorithm could include an additional step size $\mu > 0$, that is, we could use

$$\mathbf{x}^{[n+1]} = H_s\left(\mathbf{x}^{[n]} + \mu\Phi^T\left(\mathbf{y}-\Phi\mathbf{x}^{[n]}\right)\right). \tag{12}$$

If we use $\mu \neq 1$, the conditions on Φ derived below will change. For notational convenience, we decided to use $\mu = 1$, however, it is easy to repeat our arguments for the more general setting $\mu \neq 1$.

3.2. Computational complexity per iteration

The iterative hard thresholding algorithm is very simple. It involves the application of the operators Φ and Φ^T once in each iteration as well as two vector additions. The operator H_s involves a partial ordering of the elements of $\mathbf{a}^{[n]} = \mathbf{x}^{[n]} + \Phi^T(\mathbf{y} - \Phi\mathbf{x}^{[n]})$ in magnitude. The storage requirements are small. Apart from storage of $\mathbf{y}$, we only require the storage of the vector $\mathbf{a}$, which is of length N. Storage of $\mathbf{x}^{[n]}$, which has only s-non-zero elements, requires $2s$ numbers to be stored.

The computational bottle neck, both in terms of storage and computation time, is due to the operators Φ and Φ^T. If these are general matrices, the computational complexity and memory requirement is $O(MN)$. For large problems, it is common to use structured operators, based for example on fast Fourier transforms or wavelet transforms, which require substantially less memory and can often be applied with $O(N \log M)$ or even $O(N)$ operations. In this case, the above algorithm has minimal computational requirements per iteration. If $\mathcal{L}$ is the complexity of applying the operators Φ and Φ^T, then the computational complexity of the algorithm is $O(k^\star\mathcal{L})$, where $k^\star$ is the total number of iterations.

4. Iterative hard thresholding for compressed sensing

In this section we derive the main result of this paper. We show that if $\delta_{3s} < 1/\sqrt{32}$, then the iterative hard thresholding algorithm reduces the estimation error in each iteration and is guaranteed to come within a constant factor of the best attainable estimation error. In fact, the algorithm needs a fixed number of iterations, depending only on the logarithm of a form of signal to noise ratio.

4.1. Digest: The main result

The main result of this paper can be formally stated in the following theorems.

Theorem 4. *Given a noisy observation $\mathbf{y} = \Phi\mathbf{x} + \mathbf{e}$, where $\mathbf{x}$ is an arbitrary vector. Let $\mathbf{x}^s$ be an approximation to $\mathbf{x}$ with no more than s non-zero elements for which $\|\mathbf{x} - \mathbf{x}^s\|_2$ is minimal. If Φ has restricted isometry property with $\delta_{3s} < 1/\sqrt{32}$, then, at iteration k, IHT$_s$ will recover an approximation $\mathbf{x}^{[k]}$ satisfying*

$$\left\|\mathbf{x} - \mathbf{x}^{[k]}\right\|_2 \leqslant 2^{-k}\left\|\mathbf{x}^s\right\|_2 + 6\tilde{\epsilon}_s, \tag{13}$$

where

$$\tilde{\epsilon}_s = \left\|\mathbf{x} - \mathbf{x}^s\right\|_2 + \frac{1}{\sqrt{s}}\left\|\mathbf{x} - \mathbf{x}^s\right\|_1 + \|\mathbf{e}\|_2. \tag{14}$$

Furthermore, after at most

$$k^\star = \left\lceil \log_2\left(\frac{\|\mathbf{x}^s\|_2}{\tilde{\epsilon}_s}\right)\right\rceil \tag{15}$$

iterations, IHT$_s$ estimates $\mathbf{x}$ with accuracy

$$\left\|\mathbf{x} - \mathbf{x}^{[k^\star]}\right\|_2 \leqslant 7\left[\left\|\mathbf{x} - \mathbf{x}^s\right\|_2 + \frac{1}{\sqrt{s}}\left\|\mathbf{x} - \mathbf{x}^s\right\|_1 + \|\mathbf{e}\|_2\right]. \tag{16}$$

For exact sparse signals, we have a slightly better result.

Theorem 5. *Given a noisy observation $\mathbf{y} = \Phi\mathbf{x}^s + \mathbf{e}$, where $\mathbf{x}^s$ is s-sparse. If Φ has the restricted isometry property with $\delta_{3s} < 1/\sqrt{32}$, then, at iteration k, IHT$_s$ will recover an approximation $\mathbf{x}^{[k]}$ satisfying*

$$\left\|\mathbf{x}^s - \mathbf{x}^{[k]}\right\|_2 \leqslant 2^{-k}\left\|\mathbf{x}^s\right\|_2 + 5\|\mathbf{e}\|_2. \tag{17}$$

Furthermore, after at most

$$k^\star = \left\lceil \log_2\left(\frac{\|\mathbf{x}^s\|_2}{\|\mathbf{e}\|_2}\right)\right\rceil \tag{18}$$

iterations, IHT$_s$ estimates $\mathbf{x}$ with accuracy

$$\left\|\mathbf{x}^s - \mathbf{x}^{[k^\star]}\right\|_2 \leqslant 6\|\mathbf{e}\|_2. \tag{19}$$

4.2. Discussion of the main results

The main theorem states that the algorithm will find an approximation that comes close to the true vector. However, there is a limit to this. Asymptotically, we are only guaranteed to get as close as a multiple of

$$\tilde{\epsilon}_s = \left\|\mathbf{x}-\mathbf{x}^s\right\|_2 + \frac{1}{\sqrt{s}}\left\|\mathbf{x}-\mathbf{x}^s\right\|_1 + \|\mathbf{e}\|_2. \tag{20}$$

The quantity $\tilde{\epsilon}_s$ can be understood as an error term. This error term is composed of two components, the observation error $\mathbf{e}$ and the difference between the signal $\mathbf{x}$ and its best s term approximation $\mathbf{x}^s$. This makes intuitive sense. Assume, the observation error is zero and $\mathbf{x}$ is s-sparse. In this case, the algorithm is guaranteed (under the conditions of the theorem) to find $\mathbf{x}$ exactly. For exact sparse signals, but with noisy observation, our success in recovering $\mathbf{x}$ is naturally limited by the size of the error. Assuming that $\mathbf{x}$ is not s-sparse, there will be an error between any s-term approximation and $\mathbf{x}$. The closest we can get to $\mathbf{x}$ with any s-sparse approximation is therefore limited by how well $\mathbf{x}$ can be approximated with s-sparse signals.

The overall number of iterations required to achieve a desired accuracy depends on the logarithm of $\frac{\|\mathbf{x}^s\|_2}{\tilde{\epsilon}_s}$. We can think of the quantity $\frac{\|\mathbf{x}^s\|_2}{\tilde{\epsilon}_s}$ as a signal to noise ratio appropriate for sparse signal estimates. This term is large, whenever the observation noise is small and the signal $\mathbf{x}$ is well approximated by an s-sparse vector.

4.3. Derivation of the error bound

We now turn to the derivation of the main result which is heavily inspired by the arguments in [13].

The proof uses the following notation

(1) $\mathbf{y} = \Phi\mathbf{x}^s + \mathbf{e}$,
(2) $\mathbf{r}^{[n]} = \mathbf{x}^s - \mathbf{x}^{[n]}$,
(3) $\mathbf{a}^{[n+1]} = \mathbf{x}^{[n]} + \Phi^T(\mathbf{y} - \Phi\mathbf{x}^{[n]}) = \mathbf{x}^{[n]} + \Phi^T(\Phi\mathbf{x}^s + \mathbf{e} - \Phi\mathbf{x}^{[n]})$,
(4) $\mathbf{x}^{[n+1]} = H_s(\mathbf{a}^{[n+1]})$, where H_s is the hard thresholding operator that keeps the largest s (in magnitude) elements and sets the other elements to zero.
(5) $\Gamma_s^\star = \operatorname{supp}\{\mathbf{x}^s\}$,
(6) $\Gamma^n = \operatorname{supp}\{\mathbf{x}^{[n]}\}$,
(7) $B^{n+1} = \Gamma_s^\star \cup \Gamma^{n+1}$.

We start by proving the error bound in Theorem 5.

Proof of the error bound in Theorem 5. Consider the s-sparse signal $\mathbf{x}^s$. In order to deal with the non-linear operator, we proceed as follows. First we note that the error $\mathbf{x}^s - \mathbf{x}^{[n+1]}$ is supported on the set $B^{n+1} = \Gamma_s^\star \cup \Gamma^{n+1}$. The estimation error in iteration k can then be bounded by considering the effect of the hard thresholding operator, that is, by looking at the error $\|\mathbf{x}^{[n+1]}_{B^{n+1}} - \mathbf{a}^{[n+1]}_{B^{n+1}}\|_2$. This is the difference between the thresholded estimate and the estimate before thresholding (whilst restricting the analysis only to the support B^{n+1} of $\mathbf{x}^s - \mathbf{x}^{[n+1]}$). Using the triangle inequality, we can write

$$\left\|\mathbf{x}^s - \mathbf{x}^{[n+1]}\right\|_2 \leqslant \left\|\mathbf{x}^s_{B^{n+1}} - \mathbf{a}^{[n+1]}_{B^{n+1}}\right\|_2 + \left\|\mathbf{x}^{[n+1]}_{B^{n+1}} - \mathbf{a}^{[n+1]}_{B^{n+1}}\right\|_2. \tag{21}$$

By the thresholding operation, $\mathbf{x}^{[n+1]}$ is the best s-term approximation to $\mathbf{a}^{[n+1]}_{B^{n+1}}$. In particular, it is a better approximation than $\mathbf{x}^s$. This implies that $\|\mathbf{x}^{[n+1]} - \mathbf{a}^{[n+1]}_{B^{n+1}}\|_2 \leqslant \|\mathbf{x}^s - \mathbf{a}^{[n+1]}_{B^{n+1}}\|_2$ and we have

$$\left\|\mathbf{x}^s - \mathbf{x}^{[n+1]}\right\|_2 \leqslant 2\left\|\mathbf{x}^s_{B^{n+1}} - \mathbf{a}^{[n+1]}_{B^{n+1}}\right\|_2. \tag{22}$$

We now expand

$$\mathbf{a}^{[n+1]}_{B^{n+1}} = \mathbf{x}^{[n]}_{B^{n+1}} + \Phi^T_{B^{n+1}}\Phi\mathbf{r}^{[n]} + \Phi^T_{B^{n+1}}\mathbf{e}. \tag{23}$$

We then have

$$\begin{aligned}
\left\|\mathbf{x}^s - \mathbf{x}^{[n+1]}\right\|_2 &\leqslant 2\left\|\mathbf{x}^s_{B^{n+1}} - \mathbf{x}^{[n]}_{B^{n+1}} - \Phi^T_{B^{n+1}}\Phi\mathbf{r}^{[n]} - \Phi^T_{B^{n+1}}\mathbf{e}\right\|_2 \\
&\leqslant 2\left\|\mathbf{r}^{[n]}_{B^{n+1}} - \Phi^T_{B^{n+1}}\Phi\mathbf{r}^{[n]}\right\|_2 + 2\left\|\Phi^T_{B^{n+1}}\mathbf{e}\right\|_2 \\
&= 2\left\|\left(\mathbf{I} - \Phi^T_{B^{n+1}}\Phi_{B^{n+1}}\right)\mathbf{r}^{[n]}_{B^{n+1}} - \Phi^T_{B^{n+1}}\Phi_{B^n\setminus B^{n+1}}\mathbf{r}^{[n]}_{B^n\setminus B^{n+1}}\right\|_2 + 2\left\|\Phi^T_{B^{n+1}}\mathbf{e}\right\|_2 \\
&\leqslant 2\left\|\left(\mathbf{I} - \Phi^T_{B^{n+1}}\Phi_{B^{n+1}}\right)\mathbf{r}^{[n]}_{B^{n+1}}\right\|_2 + 2\left\|\left(\Phi^T_{B^{n+1}}\Phi_{B^n\setminus B^{n+1}}\right)\mathbf{r}^{[n]}_{B^n\setminus B^{n+1}}\right\|_2 + 2\left\|\Phi^T_{B^{n+1}}\mathbf{e}\right\|_2.
\end{aligned}$$

Now $B^n \setminus B^{n+1}$ is disjoint from B^{n+1} and $|B^n \cup B^{n+1}| \leqslant 3s$. We can therefore use the bounds in (3), (6) and (5) and the fact that $\delta_{2s} \leqslant \delta_{3s}$

$$\left\|\mathbf{r}^{[n+1]}\right\|_2 \leqslant 2\delta_{2s}\left\|\mathbf{r}^{[n]}_{B^{n+1}}\right\|_2 + 2\delta_{3s}\left\|\mathbf{r}^{[n]}_{B^n\setminus B^{n+1}}\right\|_2 + 2\sqrt{1+\delta_s}\|\mathbf{e}\|_2 \tag{24}$$

$$\leqslant 2\delta_{3s}\left(\left\|\mathbf{r}^{[n]}_{B^{n+1}}\right\|_2 + \left\|\mathbf{r}^{[n]}_{B^n\setminus B^{n+1}}\right\|_2\right) + 2\sqrt{1+\delta_s}\|\mathbf{e}\|_2. \tag{25}$$

Furthermore $\mathbf{r}^{[n]}_{B^{n+1}}$ and $\mathbf{r}^{[n]}_{B^n\setminus B^{n+1}}$ are orthogonal so that $\|\mathbf{r}^{[n]}_{B^{n+1}}\|_2 + \|\mathbf{r}^{[n]}_{B^n\setminus B^{n+1}}\|_2 \leqslant \sqrt{2}\|\mathbf{r}\|_2$. We therefore have the bound

$$\left\|\mathbf{r}^{[n+1]}\right\|_2 \leqslant \sqrt{8}\delta_{3s}\left\|\mathbf{r}^{[n]}\right\|_2 + 2\sqrt{1+\delta_s}\|\mathbf{e}\|_2. \tag{26}$$

If $\delta_{3s} < \frac{1}{\sqrt{32}}$, then

$$\left\|\mathbf{r}^{[n+1]}\right\|_2 < 0.5\left\|\mathbf{r}^{[n]}\right\|_2 + 2.17\|\mathbf{e}\|_2. \tag{27}$$

Iterating this relationship, and realising[1] that $2.17(1+0.5+0.25+\cdots) \leqslant 4.34$ and that $\mathbf{x}^{[0]} = \mathbf{0}$, we get

$$\left\|\mathbf{r}^{[k]}\right\|_2 < 2^{-k}\left\|\mathbf{x}^s\right\|_2 + 4.34\|\mathbf{e}\|_2. \tag{28}$$

This proves Theorem 5. □

Proof of the error bound in Theorem 4. To bound the error $\|\mathbf{x}-\mathbf{x}^{[k]}\|_2$ for general $\mathbf{x}$, we first note that

$$\left\|\mathbf{x}-\mathbf{x}^{[k]}\right\|_2 \leqslant \left\|\mathbf{x}^s-\mathbf{x}^{[k]}\right\|_2 + \|\mathbf{x}-\mathbf{x}^s\|_2. \tag{29}$$

The proof of the main theorem then follows by bounding $\|\mathbf{x}^s-\mathbf{x}^{[k]}\|_2 = \|\mathbf{r}^{[k]}\|_2$ using Theorem 5 with $\tilde{\mathbf{e}}$ instead of $\mathbf{e}$ and Lemma 3 to bound $\|\tilde{\mathbf{e}}\|_2$, that is

$$\left\|\mathbf{r}^{[k]}\right\|_2 \leqslant 2^{-k}\left\|\mathbf{x}^s\right\|_2 + 4.71\left[\left\|\left(\mathbf{x}-\mathbf{x}^s\right)\right\|_2 + \frac{1}{\sqrt{s}}\left\|\left(\mathbf{x}-\mathbf{x}^s\right)\right\|_1 + \|\mathbf{e}\|_2\right]. \qquad \square \tag{30}$$

4.4. Derivation of the iteration count

Proof of the second part of Theorem 4. The first part of Theorem 4 shows that

$$\left\|\mathbf{x}-\mathbf{x}^{[k]}\right\|_2 \leqslant 2^{-k}\left\|\mathbf{x}^s\right\|_2 + 6\tilde{\epsilon}_s, \tag{31}$$

where $\tilde{\epsilon}_s = [\|(\mathbf{x}-\mathbf{x}^s)\|_2 + \frac{1}{\sqrt{s}}\|(\mathbf{x}-\mathbf{x}^s)\|_1 + \|\mathbf{e}\|_2]$. We are therefore guaranteed to reduce the error to below any multiple c of $\tilde{\epsilon}_s$, as long as $c > 6$. For example, assume we want to recover $\|\mathbf{x}\|$ with an error of less than $7\tilde{\epsilon}_s$. This implies that we require that

$$2^{-k}\left\|\mathbf{x}^s\right\|_2 \leqslant \tilde{\epsilon}_s \tag{32}$$

i.e. that

$$2^k \geqslant \frac{\|\mathbf{x}^s\|_2}{\tilde{\epsilon}_s}, \tag{33}$$

which in turn implies the second part of the theorem. The proof of the corresponding result in Theorem 5 follows the same argument. □

5. Optimality of the result

The results in this paper are essentially optimal up to the constants. To show this, we show that the error has to depend on the three terms $\|\mathbf{x}-\mathbf{x}^s\|_2$, $\frac{\|\mathbf{x}-\mathbf{x}^s\|_1}{\sqrt{s}}$ and $\|\mathbf{e}\|_2$, even if we used an oracle estimate. Assume an oracle would give us the support set of the s largest coefficients in $\mathbf{x}$. Let Φ_s be the sub-matrix of Φ containing only those columns associated with the s largest coefficients in $\mathbf{x}$ and assume Φ to have a restricted isometry constant of $\delta_s > 0$. We can now write the oracle estimate that would reconstruct the best s-term approximation as $\Phi_s^\dagger\mathbf{y}$, where the † signifies the pseudo-inverse.

The error for such an oracle estimate can be bounded from above as

$$\begin{aligned}\left\|\mathbf{x}-\Phi_s^\dagger\mathbf{y}\right\|_2 &= \left\|\mathbf{x}-\mathbf{x}^s-\Phi_s^\dagger\Phi\left(\mathbf{x}-\mathbf{x}^s\right)-\Phi_s^\dagger\mathbf{e}\right\|_2\\ &\leqslant \left\|\mathbf{x}-\mathbf{x}^s\right\|_2 + \left\|\Phi_s^\dagger\Phi\left(\mathbf{x}-\mathbf{x}^s\right)\right\|_2 + \left\|\Phi_s^\dagger\mathbf{e}\right\|_2 \leqslant \left\|\mathbf{x}-\mathbf{x}^s\right\|_2 + \left\|\Phi_s^\dagger\right\|_2\left\|\Phi\left(\mathbf{x}-\mathbf{x}^s\right)\right\|_2 + \left\|\Phi_s^\dagger\right\|_2\|\mathbf{e}\|_2\\ &\leqslant \left(1+\frac{\sqrt{1+\delta_s}}{\sqrt{1-\delta_s}}\right)\left\|\mathbf{x}-\mathbf{x}^s\right\|_2 + \frac{\sqrt{1+\delta_s}}{\sqrt{1-\delta_s}}\frac{\|\mathbf{x}-\mathbf{x}^s\|_1}{\sqrt{s}} + \frac{1}{\sqrt{1-\delta_s}}\|\mathbf{e}\|_2,\end{aligned} \tag{34}$$

[1] Note that we could have chosen any value for $\delta_{3s} < 1/\sqrt{8}$ here. As long as $\sqrt{8}\delta_{3s} < 1$, the geometric series used to bound the error converges to $1/(1-\sqrt{8}\delta_{3s})$. However, larger δ_{3s} will lead to worse error bounds.

where the bound on $\|\Phi_s^\dagger\|_2$ can be found in, for example, [13] and where the bound on $\|\Phi(\mathbf{x}-\mathbf{x}^s)\|_2$ is due to Lemma 2.

Importantly, the same type of bound can be derived to bound the error from below for some $\mathbf{e}$ and $\mathbf{x}$. Because $\mathbf{x}-\mathbf{x}^s$ is orthogonal to $\Phi_s^\dagger\Phi(\mathbf{x}-\mathbf{x}^s)-\Phi_s^\dagger\mathbf{e}$, we have

$$\begin{aligned}\left\|\mathbf{x}-\Phi_s^\dagger\mathbf{y}\right\|_2 &= \left\|\mathbf{x}-\mathbf{x}^s-\Phi_s^\dagger\Phi\left(\mathbf{x}-\mathbf{x}^s\right)-\Phi_s^\dagger\mathbf{e}\right\|_2 \geqslant \frac{1}{\sqrt{2}}\left\|\mathbf{x}-\mathbf{x}^s\right\|_2+\frac{1}{\sqrt{2}}\left\|\Phi_s^\dagger\left(\Phi\left(\mathbf{x}-\mathbf{x}^s\right)+\mathbf{e}\right)\right\|_2\\ &\geqslant \frac{1}{\sqrt{2}}\left\|\mathbf{x}-\mathbf{x}^s\right\|_2+\frac{\sqrt{1-\delta_s}}{\sqrt{2}}\left\|\Phi\left(\mathbf{x}-\mathbf{x}^s\right)+\mathbf{e}\right\|_2. \end{aligned} \tag{35}$$

In particular, we can choose $\mathbf{e}$ to be orthogonal to $\Phi(\mathbf{x}-\mathbf{x}^s)$. In this case, we have

$$\left\|\Phi\left(\mathbf{x}-\mathbf{x}^s\right)+\mathbf{e}\right\|_2 \geqslant \frac{1}{\sqrt{2}}\left\|\Phi\left(\mathbf{x}-\mathbf{x}^s\right)\right\|_2+\frac{1}{\sqrt{2}}\|\mathbf{e}\|_2.$$

It remains to show that there are $\mathbf{x}$, such that $\|\Phi(\mathbf{x}-\mathbf{x}^s)\|_2 \geqslant c\|(\mathbf{x}-\mathbf{x}^s)\|_1/\sqrt{s}$. Let us use the following notation. Let Γ_0 be the index set of the largest s elements in $\mathbf{x}$, let Γ_1 be the index set of the next s largest elements and so on. For all those $\mathbf{x}$, for which $\mathbf{x}_{\Gamma_i}=0$ for all $i>1$, we have the following trivial inequality

$$\left\|\sum_{i\geqslant 1}\Phi\mathbf{x}_{\Gamma_i}\right\|_2 \geqslant \sum_{i\geqslant 1}\|\Phi\mathbf{x}_{\Gamma_i}\|_2$$

so that for these $\mathbf{x}$

$$\left\|\Phi(\mathbf{x}-\mathbf{x}_s)\right\|_2=\left\|\sum_{i\geqslant 1}\Phi\mathbf{x}_{\Gamma_i}\right\|_2 \geqslant \sum_{i\geqslant 1}\|\Phi\mathbf{x}_{\Gamma_i}\|_2 \geqslant (1-\delta_s)\sum_{i\geqslant 1}\|\mathbf{x}_{\Gamma_i}\|_2 \geqslant \frac{(1-\delta_s)}{\sqrt{s}}\sum_{i\geqslant 1}\|\mathbf{x}_{\Gamma_i}\|_1.$$

From these arguments we see that there are $\mathbf{x}$ and $\mathbf{e}$ for which the performance bound in this paper is essentially tight up to the constants, that is, we cannot expect to do substantially better, even if we had additional information regarding the location of the largest s elements in $\mathbf{x}$.

6. When to stop

So far, we have given guarantees on the achievable error and a bound on the total number of iterations to achieve this bound. The algorithm will converge to these bounds linearly, however, in practise, it is necessary to monitor quantities of the algorithm and decide to stop the algorithm after a finite number of iterations. From the main result, it is clear that in general, we cannot do any better than to find an estimate with an error of $6\tilde{\epsilon}_s$. The following results can be used as diagnostics that assess when the algorithm has come close to this limit.

A possible stopping criterion is $\|\mathbf{y}-\Phi\mathbf{x}^{[n]}\|_2 \leqslant \epsilon$. For this criterion, it is possible to use the same arguments as in Appendix A of [13], to derive the following result.

Lemma 6. *Assume that Φ satisfies the RIP with $\delta_{3s}<1/\sqrt{32}$. If at any iteration of IHT$_s$ the condition $\|\mathbf{y}-\Phi\mathbf{x}^{[n]}\|_2 \leqslant \epsilon$ holds, then*

$$\left\|\mathbf{x}-\mathbf{x}^{[n]}\right\|_2 \leqslant 1.11\epsilon+2.41\tilde{\epsilon}_s, \tag{36}$$

where

$$\tilde{\epsilon}_s=\left\|\mathbf{x}-\mathbf{x}^s\right\|_2+\frac{1}{\sqrt{s}}\left\|\mathbf{x}-\mathbf{x}^s\right\|_1+\|\mathbf{e}\|_2. \tag{37}$$

Conversely, if at any iteration of IHT$_s$ the condition

$$\left\|\mathbf{x}-\mathbf{x}^{[n]}\right\|_2 \leqslant 1.1\left\|\mathbf{x}-\mathbf{x}^{[n]}\right\|_2+1.11/\sqrt{s}\left\|\mathbf{x}-\mathbf{x}^{[n]}\right\|_1+\|\mathbf{e}\|_2 \tag{38}$$

holds, then $\|\mathbf{y}-\Phi\mathbf{x}^{[n]}\|_2 \leqslant \epsilon$.

This lemma can be used to calculate a stopping criterion for *IHT$_s$*. For example, if we want to estimate $\|\mathbf{x}\|_2$ with accuracy $c\tilde{\epsilon}_s$, we know that we are done as soon as $\|\mathbf{y}-\Phi\mathbf{x}_2^{[n]}\|_2 \leqslant (c/1.11-2)\tilde{\epsilon}_s$. Note that in general, *IHT$_s$* is only guaranteed to work for $c>6$, however, as soon as we observe that $\|\mathbf{y}-\Phi\mathbf{x}^{[n]}\|_2 \leqslant (c/1.11-2)$ holds for arbitrary c, we know that the estimation error must be below $c\tilde{\epsilon}_s$.

Proof. To prove the first part, note that the stopping criterion implies that

$$\epsilon \geqslant \left\|\Phi\left(\mathbf{x}-\mathbf{x}^{[n]}\right)+\mathbf{e}\right\|_2=\left\|\Phi\left(\mathbf{x}^s-\mathbf{x}^{[n]}\right)+\tilde{\mathbf{e}}\right\|_2 \geqslant \sqrt{1-\delta_{2s}}\left\|\mathbf{x}^s-\mathbf{x}^{[n]}\right\|_2-\|\tilde{\mathbf{e}}\|_2,$$

so that

$$\left\|\mathbf{x}^s - \mathbf{x}^{[n]}\right\|_2 \leqslant \frac{\epsilon + \|\tilde{\mathbf{e}}\|_2}{\sqrt{1-\delta_{2s}}}. \tag{39}$$

Furthermore,

$$\left\|\mathbf{x} - \mathbf{x}^{[n]}\right\|_2 \leqslant \left\|\mathbf{x}^s - \mathbf{x}^{[n]}\right\|_2 + \left\|\mathbf{x} - \mathbf{x}^s\right\|_2, \tag{40}$$

so that (using the bound on $\|\tilde{\mathbf{e}}\|_2$ from Lemma 3)

$$\begin{aligned}\left\|\mathbf{x} - \mathbf{x}^{[n]}\right\|_2 &\leqslant \frac{\epsilon + \|\tilde{\mathbf{e}}\|_2}{\sqrt{1-\delta_{2s}}} + \left\|\mathbf{x} - \mathbf{x}^s\right\|_2 \\ &\leqslant \frac{\epsilon + 2\sqrt{1+\delta_{2s}}\|\mathbf{x} - \mathbf{x}^s\|_2 + \sqrt{1+\delta_{2s}}\frac{1}{\sqrt{s}}\|\mathbf{x} - \mathbf{x}^s\|_1 + \|\mathbf{e}\|_2}{\sqrt{1-\delta_{2s}}} \\ &\leqslant 1.11\epsilon + 2.41\tilde{\epsilon}_s.\end{aligned}$$

This proves the first part of the lemma using $\delta_{2s} \leqslant \delta_{3s} < 1/\sqrt{32}$.

To prove the second part, note that if

$$\left\|\mathbf{x} - \mathbf{x}^{[n]}\right\|_2 \leqslant \frac{\epsilon - \frac{\sqrt{1+\delta_s}}{\sqrt{s}}\|\mathbf{x} - \mathbf{x}^{[n]}\|_1 - \|\mathbf{e}\|_2}{\sqrt{1+\delta_s}} \tag{41}$$

holds, then

$$\epsilon \geqslant \sqrt{1+\delta_s}\left\|\mathbf{x} - \mathbf{x}^{[n]}\right\|_2 + \frac{\sqrt{1+\delta_s}}{\sqrt{s}}\left\|\mathbf{x} - \mathbf{x}^{[n]}\right\|_1 + \|\mathbf{e}\|_2 \geqslant \left\|\Phi\left(\mathbf{x} - \mathbf{x}^{[n]}\right)\right\|_2 + \|\mathbf{e}\|_2 \geqslant \left\|\Phi\left(\mathbf{x} - \mathbf{x}^{[n]}\right) + \mathbf{e}\right\|_2.$$

We have here used Lemma 2 to bound $\|\Phi(\mathbf{x} - \mathbf{x}^{[n]})\|_2 \leqslant \sqrt{1+\delta_s}\|(\mathbf{x} - \mathbf{x}^{[n]})\|_2 + \sqrt{1+\delta_s}\frac{1}{\sqrt{s}}\|(\mathbf{x} - \mathbf{x}^{[n]})\|_1$. □

7. Comparison to CoSaMP

In [13] the authors introduced the Compressed Sensing Matching Pursuit (CoSaMP) algorithm, which offers similar guarantees to the iterative hard thresholding approach of this paper. The result for CoSaMP is as follows:

Theorem 7 *(Needell and Tropp [13]). If Φ has the restricted isometry property with $\delta_{4s} \leqslant 0.1$, then, at iteration k, CoSaMP will recover an approximation $\mathbf{x}^{[k]}$ satisfying*

$$\left\|\mathbf{x}^s - \mathbf{x}^{[k]}\right\|_2 \leqslant 2^{-k}\left\|\mathbf{x}^s\right\|_2 + 15\|\mathbf{e}\|_2 \tag{42}$$

if $\mathbf{y} = \Phi\mathbf{x}^s + \mathbf{e}$ for $\mathbf{x}^s$ s-sparse and

$$\left\|\mathbf{x} - \mathbf{x}^{[k]}\right\|_2 \leqslant 2^{-k}\|\mathbf{x}\|_2 + 20\tilde{\epsilon}_s, \tag{43}$$

if $\mathbf{y} = \Phi\mathbf{x} + \mathbf{e}$ for all $\mathbf{x}$.

Two remarks are in order. Firstly, for *IHT*$_s$, we require $\delta_{3s} \leqslant 0.175$, whilst for CoSaMP, $\delta_{4s} \leqslant 0.1$ is required. Note that the *IHT*$_s$ requirement is for δ_{3s}, whilst CoSaMP has a requirement on δ_{4s}. To further compare these requirements, we use Corollary 3.4 from [13], which states that for integers a and s, $\delta_{as} \leqslant a\delta_{2s}$. Therefore, if $\delta_{2s} \leqslant 0.025$, the condition for CoSaMP is satisfied, whilst for *IHT*$_s$, we have a the condition requiring that $\delta_{2s} \leqslant 0.058$.

The theorems derived for CoSaMP have therefore somewhat stricter conditions. Furthermore, the theorem derived here for *IHT*$_s$ guarantees a several times lower approximation error. For example, in the exact sparse case, *IHT*$_s$ is guaranteed to calculate an error approaching $5\|\mathbf{e}\|_2$, which should be compared to the guarantee of $15\|\mathbf{e}\|_2$ for CoSaMP. For general signals, the guarantees are $6\tilde{\epsilon}_s$ for *IHT*$_s$ and $20\tilde{\epsilon}_s$ for CoSaMP.

The number of iterations required for *IHT*$_s$ is logarithmic in the signal to noise ratio. This means that for noiseless observations, *IHT*$_s$ would require an infinite number of iterations to reduce the error to zero. This is a well known property of algorithms that use updates of the form $\mathbf{x}^{[n]} + \Phi^T(\mathbf{y} - \Phi\mathbf{x}^{[n]})$. CoSaMP on the other hand is guaranteed to estimate $\mathbf{x}$ with precision $20\tilde{\epsilon}_s$ in at most $6(s+1)$ iterations, however, to achieve this, CoSaMP requires the solution to an inverse problem in each iteration, which is costly. *IHT*$_s$ does not require the exact solution to an inverse problem. If CoSaMP is implemented using fast partial solutions to the inverse problems, the iteration count guarantees become similar to the ones derived here for *IHT*$_s$.

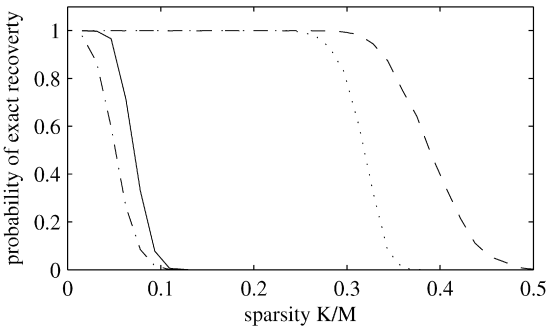


Fig. 1. Comparison in terms of exact support recovery between the Iterative Hard Thresholding algorithm (solid), the CoSaMP algorithm using three conjugate gradient steps (dash-dotted), the CoSaMP algorithm using exact solutions to the inverse problem in each iteration (dotted) and the ℓ_1 based method (dashed).

8. What's in a theorem

A word of caution is in order. We have here shown that the Iterative Hard Thresholding algorithm has theoretical properties, which are comparable to those of other state of the art algorithms such as CoSaMP and has recovery guarantees of the same order as ℓ_1 based approaches. At a first glance, this seems to be at odds with previously reported numerical results [1].

To understand this disparity, it has to be realised that the uniform performance guarantees derived here for IHT_s and elsewhere for CoSaMP and ℓ_1 based methods are worst case bounds, that is, they guarantee the performance of the algorithms in the worst possible scenario. Numerical studies on the other hand cannot in general test this worst case behaviour. This is because we do not in general know what particular signals would be the most difficult to recover. Numerical experiments instead analyse average behaviour, that is, they study the recovery of *typical* signals.

We demonstrate this in Fig. 1 where we show the average recovery performance of different methods, that is, the empirical probability with which the methods recover the support of an exact sparse signal. The results were derived by generating $\Phi \in \mathbb{R}^{128\times 256}$ by drawing its column vectors uniformly from the unit sphere and by generating exact s-sparse signals where the non-zero elements were i.i.d. Gaussian. We then used the Iterative Hard Thresholding algorithm, two versions of the CoSaMP algorithm and an ℓ_1 minimisation ($\min_{\mathbf{x}} \|\mathbf{x}\|_1$: $\Phi\mathbf{x} = \mathbf{y}$). The two versions of CoSaMP differed in that in one version we used an exact solution to the inverse problem in each iteration whilst in the other version we replaced this exact solution with three conjugate gradient iterations as suggested in [13]. Note that Theorem 7 holds for both of these methods.

As reported before, we observe that the Iterative Hard Thresholding algorithm (solid line) only recovers very sparse signals, whilst the ℓ_1 optimisation based approach (dashed line) can recover much less sparse signals. Interestingly, CoSaMP works relatively well when using the implementation based on exact solutions of the inverse problem, whilst it performs slightly worse than the Iterative Hard Thresholding approach when using the faster implementation based on partial solutions of this inverse problem.

Whilst the theoretical results tell us that if the restricted isometry constant is below some value, all algorithms will be able to recover exact sparse signals, the above experiment shows that, in the regime in which the restricted isometry constant becomes too large, some algorithms still perform well on average, whilst others do not. The difference in numerically observed performance between methods with similar uniform recovery guarantees therefore indicates that uniform guarantees are not necessarily a good measure to indicate good average performance.

9. Conclusions

The abstract made eight claims regarding the performance of the iterative hard thresholding algorithm. Let us here summarise these in somewhat more detail.

- Error guarantee: We could show that an estimation error of $7\|\tilde{\mathbf{e}}\|_2$ can be achieved within a finite number of iterations.
- Robustness to noise: We could show that the achievable estimation error depends linearly on the size of the observation error. Performance therefore degrades linearly if the noise is increased.
- Minimum number of observations: The requirement on the isometry constant dictates that the number of observations grows linearly with s and logarithmically with N. Up to a constant, this relation is known to be the best attainable.
- Sampling operator: The algorithm is simple and requires only the application of Φ and Φ^T.
- Memory requirement: We could show this to be linear in the problem size, if we can ignore the storage requirement for Φ and Φ^T.

- Computational complexity: We could shown that the computational complexity is of the same order as the application of the measurement operator or its adjoint per iteration. The total number of iterations is bounded due to the linear convergence of the algorithm and depends logarithmically on the signal to noise ratio $\|\mathbf{x}^s\|_2/\|\tilde{\mathbf{e}}\|_2$.
- Number of iterations: We could show that after at most $\lceil \log_2(\frac{\|\mathbf{x}^s\|_2}{\tilde{\epsilon}_s}) \rceil$ iterations, the estimation error is smaller than $7\tilde{\epsilon}_s$.
- Uniform performance guarantees: The results presented here only depend on δ_{3s} and do not depend on the size and distribution of the largest s elements in $\mathbf{x}$.

This is an impressive list of properties for such a relatively simple algorithm. To our knowledge, only the CoSaMP algorithm shares similar guarantees. However, as discussed in Section 8, uniform guarantees are not the only consideration and in practise marked differences in the average performance of different methods are apparent. For many small problems, the restricted isometry property of random matrices is often too large to explain the behaviour of the different methods in these studies. Furthermore, it has long been observed that the distribution of the magnitude of the non-zero coefficients also has an important influence on the performance of different methods. Whilst the theoretical guarantees derived in this and similar papers are important to understand the behaviour of an algorithm, it is also clear that other facts have to be taken into account in order to predict the *typical* performance of algorithms in many practical situations.

Acknowledgments

This research was supported by EPSRC grants D000246/2 and EP/F039697/1. M.E.D. 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. An abridged earlier version of this manuscript has been submitted to the International Conference on Acoustics, Speech and Signal Processing (ICASSP) [2].

References

[1] T. Blumensath, M. Davies, Iterative thresholding for sparse approximations, J. Fourier Anal. Appl. 14 (5) (2008) 629–654.
[2] T. Blumensath, M. Davies, A simple, efficient and near optimal algorithm for compressed sensing, in: Proceedings of the Int. Conf. on Acoustics, Speech and Signal Processing, 2009.
[3] E. Candès, J. Romberg, Quantitative robust uncertainty principles and optimally sparse decompositions, Found. Comput. Math. 6 (2) (2006) 227–254.
[4] E. Candès, J. Romberg, T. Tao, Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information, IEEE Trans. Inform. Theory 52 (2) (2006) 489–509.
[5] E. Candès, J. Romberg, T. Tao, Stable signal recovery from incomplete and inaccurate measurements, Comm. Pure Appl. Math. 59 (8) (2006) 1207–1223.
[6] W. Dai, O. Milenkovic, Subspace pursuit for compressive sensing signal reconstruction, IEEE Trans. Inform. Theory 55 (5) (2009) 2230–2249.
[7] D. Donoho, Compressed sensing, IEEE Trans. Inform. Theory 52 (4) (2006) 1289–1306.
[8] I.F. Gorodnitsky, J.S. George, B.D. Rao, Neuromagnetic source imaging with focuss: A recursive weighted minimum norm algorithm, Neurophysiology 95 (4) (1995) 231–251.
[9] N.G. Kingsbury, T.H. Reeves, Iterative image coding with overcomplete complex wavelet transforms, in: Proc. Conf. on Visual Communications and Image Processing, 2003.
[10] S. Mallat, G. Davis, Z. Zhang, Adaptive time–frequency decompositions, SPIE J. Opt. Eng. 33 (7) (1994) 2183–2191.
[11] S. Mendelson, A. Pajor, N. Tomczak-Jaegermann, Reconstruction and subgaussian operators in asymptotic geometric analysis, Geom. Funct. Anal. 17 (2007) 1248–1282.
[12] S. Mendelson, A. Pajor, N. Tomczak-Jaegermann, Uniform uncertainty principle for Bernoulli and subgaussian ensembles, Constr. Approx. 28 (3) (2008) 277–289.
[13] D. Needell, J.A. Tropp, CoSaMP: Iterative signal recovery from incomplete and inaccurate samples, Appl. Comput. Harmon. Anal. 26 (3) (2009) 301–321.
[14] D. Needell, R. Vershynin, Signal recovery from incomplete and inacurate measurements via regularized orthogonal matching pursuit, arXiv:0712.1360v1.
[15] D. Needell, R. Vershynin, Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit, Found. Comput. Math. 9 (2009) 317–334.
[16] H. Nyquist, Certain topics in telegraph transmission theory, Trans. A.I.E.E. (1928) 617–644.
[17] M. Rudelson, R. Vershynin, Sparse reconstruction by convex relaxation: Fourier and gaussian measurements, in: 40th Annual Conference on Information Sciences and Systems, 2006.
[18] C.A. Shannon, W. Weaver, The Mathematical Theory of Communication, University of Illinois Press, 1949.
[19] J.A. Tropp, A.C. Gilbert, Signal recovery from partial information via orthogonal matching pursuit, IEEE Trans. Inform. Theory 53 (12) (2006) 4655–4666.
[20] M. Vetterli, P. Marziliano, T. Blu, Sampling signals with finite rate of innovation, IEEE Trans. Signal Process. 50 (6) (2002) 1417–1428.