Compressed Sensing and Redundant Dictionaries

Holger Rauhut, Karin Schnass and Pierre Vandergheynst

***Abstract*— This article extends the concept of *compressed sensing* to signals that are not sparse in an orthonormal basis but rather in a redundant dictionary. It is shown that a matrix, which is a composition of a random matrix of certain type and a deterministic dictionary, has small restricted isometry constants. Thus, signals that are sparse with respect to the dictionary can be recovered via Basis Pursuit from a small number of random measurements. Further, thresholding is investigated as recovery algorithm for compressed sensing and conditions are provided that guarantee reconstruction with high probability. The different schemes are compared by numerical experiments.**

Key words: compressed sensing, redundant dictionary, sparse approximation, random matrix, restricted isometry constants, Basis Pursuit, thresholding, Orthogonal Matching Pursuit

I. Introduction

Recently there has been a growing interest in recovering sparse signals from their projection onto a small number of random vectors [5], [6], [9], [14], [20], [21]. The word most often used in this context is *compressed sensing*. It originates from the idea that it is not necessary to invest a lot of power into observing the entries of a sparse signal in all coordinates when most of them are zero anyway. Rather it should be possible to collect only a small number of measurements that still allow for reconstruction. This is potentially useful in applications where one cannot afford to collect or transmit a lot of measurements but has rich resources at the decoder.

Until now the theory of compressed sensing has only been developed for classes of signals that have a very sparse representation in an orthonormal basis (ONB). This is a rather stringent restriction. Indeed, allowing the signal to be sparse with respect to a redundant dictionary adds a lot of flexibility and significantly extends the range of applicability. Already the use of two ONBs instead of just one dramatically increases the class of signals that can be modelled in this way. A more practical example would be a dictionary made up of damped sinusoids which is used for NMR spectroscopy, see [13].

Before we can go into further explanations about the scope of this paper it is necessary to provide some background information. The basic problem in compressed sensing is to determine the minimal number n of linear non-adaptive measurements that allows for (stable) reconstruction of a signal $x \in \mathbb{R}^d$ that has at most S non-zero components. Additionally, one requires that this task can be performed reasonably fast. Each of the n measurements can be written as an inner product of the sparse signal $x \in \mathbb{R}^d$ with a vector in $\mathbb{R}^d$. To simplify the notation we store all the vectors as rows in a matrix $\mathbf{\Psi} \in \mathbb{R}^{n\times d}$ and all the measurements in the n-dimensional vector $s = \mathbf{\Psi} x$.

A naive approach to the problem of recovering x from s consists in solving the ℓ_0 minimization problem

$$(P_0) \qquad \min \|x\|_0 \text{ subject to } \|s - \mathbf{\Psi} x\|_2 \leq \eta,$$

where η is the expected noise on the measurements, $\|\cdot\|_0$ counts the number of non-zero entries

[1] H. Rauhut is with the Numerical Harmonic Analysis Group, Faculty of Mathematics, University of Vienna, Nordbergstrasse 15, A-1090 Vienna, Austria, (email: holger.rauhut@univie.ac.at)

HR acknowledges the financial support provided by the European Union's Human Potential Programme under contract MEIF-CT-2006-022811.

[2] K. Schnass and P. Vandergheynst are with the Signal Processing Institute, Swiss Federal Institute of Technology (EPFL), CH-1015 Lausanne, Switzerland, name.surname@epfl.ch

of x and $\|\cdot\|_2$ denotes the standard Euclidean norm. Although there are simple recovery conditions available, the above approach is not reasonable in practice because its solution is NP-hard [8], [19].

In order to avoid this severe drawback there have been basically two approaches proposed in the signal recovery community. The first is using greedy algorithms like *Thresholding* [15] or *(Orthogonal) Matching Pursuit* (OMP) [17], [22], see Table I for sketches of both algorithms.

TABLE I

GREEDY ALGORITHMS

```
Goal: reconstruct $x$ from $s = \mathbf{\Psi} x$
    columns of $\mathbf{\Psi}$ denoted by $\psi_j$,
    $\mathbf{\Psi}_\Lambda^\dagger$: pseudo-inverse of $\mathbf{\Psi}_\Lambda$

OMP
initialise: $z = 0$, $r = s$, $\Lambda = \emptyset$
find: $i = \arg\max_j |\langle r, \psi_j \rangle|$
update: $\Lambda = \Lambda \cup \{i\}$, $r = s - \mathbf{\Psi}_\Lambda \mathbf{\Psi}_\Lambda^\dagger s$
iterate until stopping criterion is attained
output: $x = \mathbf{\Psi}_\Lambda^\dagger s$

Thresholding
find: $\Lambda$ that contains the indices
      corresponding to the $S$ largest
      values of $|\langle s, \psi_j \rangle|$
output: $x = \mathbf{\Psi}_\Lambda^\dagger s$
```

The second approach is the *Basis Pursuit* (BP) principle. Instead of considering (P_0) one solves its convex relaxation

$$(P_1) \qquad \min \|x\|_1 \text{ subject to } \|s - \mathbf{\Psi} x\|_2 < \eta,$$

where $\|x\|_1 = \sum |x_i|$ denotes the ℓ_1-norm. This can be done via linear programming in the real case and via cone programming in the complex case. Clearly, one hopes that the solutions of (P_0) and (P_1) coincide, see [7], [10] for details.

Both approaches pose certain requirements on the matrix $\mathbf{\Psi}$ in order to ensure recovery success. Recently, Candès, Romberg and Tao [5], [6] observed that successful recovery by BP is guaranteed whenever $\mathbf{\Psi}$ obeys a uniform uncertainty principle. Essentially this means that every submatrix of $\mathbf{\Psi}$ of a certain size has to be well-conditioned. More precisely, let $\Lambda \subset \{1, \ldots, d\}$ and $\mathbf{\Psi}_\Lambda$ be the submatrix of $\mathbf{\Psi}$ consisting of the columns indexed by Λ. The local isometry constant $\delta_\Lambda = \delta_\Lambda(\mathbf{\Psi})$ is the smallest number satisfying

$$(1 - \delta_\Lambda)\|x\|_2^2 \le \|\mathbf{\Psi}_\Lambda x\|_2^2 \le (1 + \delta_\Lambda)\|x\|_2^2, \quad \text{(I.1)}$$

for all coefficient vectors x supported on Λ. The (global) restricted isometry constant is then defined as

$$\delta_S = \delta_S(\mathbf{\Psi}) := \sup_{|\Lambda| = S} \delta_\Lambda(\mathbf{\Psi}), \quad S \in \mathbb{N}.$$

The matrix $\mathbf{\Psi}$ is said to satisfy a uniform uncertainty principle if it has small restricted isometry constants, say $\delta_S(\mathbf{\Psi}) \le 1/2$. Based on this concept, Candès, Romberg and Tao proved the following recovery theorem for BP in [5, Theorem 1].

Theorem I.1. *Assume that $\mathbf{\Psi}$ satisfies*

$$\delta_{3S}(\mathbf{\Psi}) + 3\delta_{4S}(\mathbf{\Psi}) < 2$$

for some $S \in \mathbb{N}$. Let x be an S-sparse vector and assume we are given noisy data $y = \mathbf{\Psi} x + \xi$ with $\|\xi\|_2 \le \eta$. Then the solution $x^\#$ to the problem (P_1) satisfies

$$\|x^\# - x\|_2 \le C\eta. \quad \text{(I.2)}$$

The constant C depends only on δ_{3S} and δ_{4S}. If $\delta_{4S} \le 1/3$ then $C \le 15.41$.

In particular, if no noise is present, i.e., $\eta = 0$, then under the stated condition BP recovers x exactly. Note that a slight variation of the above theorem holds also in the case that x is not sparse in a strict sense, but can be well-approximated by an S-sparse vector [5, Theorem 2].

The above theorem is indeed useful, as an $n \times d$ random matrix with entries drawn from a standard Gaussian distribution (or some other distribution showing certain concentration properties, see below) will have small restricted isometry constants δ_S with 'overwhelming probability' as long as

$$n = \mathcal{O}(S \log(d/S)), \quad \text{(I.3)}$$

see [3], [5], [6], [21] for details. We note, however, that so far no deterministic construction of measurement matrices obeying the uniform uncertainty principle for reasonably small n (i.e. comparable to (I.3) is known.

In [14] it was shown that also OMP is able to reconstruct a sparse signal from Gaussian random measurements with high probability provided $n \geq CS\log(d)$, although the corresponding statement is slightly weaker than the one for BP.

As already announced we want to address the question whether the techniques described above can be extended to signals y that are not sparse in an ONB but rather in a redundant dictionary $\mathbf{\Phi} \in \mathbb{R}^{d\times K}$ with $K > d$. So now $y = \mathbf{\Phi}x$, where x has only few non-zero components. Again the goal is to reconstruct y from few measurements. More formally, given a suitable measurement matrix $A \in \mathbb{R}^{n\times d}$ we want to recover y from $s = Ay = A\mathbf{\Phi}x$. The key idea then is to use the sparse representation in $\mathbf{\Phi}$ to drive the reconstruction procedure, i.e., try to identify the sparse coefficient sequence x and from that reconstruct y. Clearly, we may represent $s = \mathbf{\Psi}x$ with

$$\mathbf{\Psi} = A\mathbf{\Phi} \in \mathbb{R}^{n\times K}.$$

In particular, we can apply all of the reconstruction methods described above by using this particular matrix $\mathbf{\Psi}$. Of course, the remaining question is whether for a fixed dictionary $\mathbf{\Phi} \in \mathbb{R}^{d\times K}$ one can find a suitable matrix $A \in \mathbb{R}^{n\times d}$ such that the composed matrix $\mathbf{\Psi} = A\mathbf{\Phi}$ allows for reconstruction of vectors having only a small number of non-zero entries. Again the strategy is to choose a random matrix A, for instance with independent standard Gaussian entries, and investigate under which conditions on $\mathbf{\Phi}$, n and S recovery is successful with high probability.

Note that already Donoho considered extensions from orthonormal bases to (redundant) tight frames $\mathbf{\Phi}$ in [9]. There it is assumed that the analysis coefficients $x' = \mathbf{\Phi}^\star y = \mathbf{\Phi}^\star\mathbf{\Phi}x$ are sparse. For redundant frames, however, this assumption does not seem very realistic as even for sparse vectors x the coefficient vector $x' = \mathbf{\Phi}^\star\mathbf{\Phi}x$ is usually fully populated.

Another motivation for investigating the applicability of Compressed Sensing for signals sparse in a dictionary is computational efficiency. If we compare the original problem of finding x from y to the new one of finding x from s we see that instead of the $d \times K$ matrix $\mathbf{\Phi}$ we now have the much smaller $n \times K$ matrix $\mathbf{\Psi}$. Considering that OMP and thresholding, as well as iterative solvers for BP, rely on inner products between the signal and the dictionary elements, we can thus reduce the number of flops per iteration from $\mathcal{O}(dK)$ to $\mathcal{O}(nK)$, where typically $n = \mathcal{O}(S\log(K/S))$, cf. Corollary II.4. Of course this does not make sense when the dictionary has a special structure that allows for fast computation of inner products, e.g. a Gabor dictionary, as the random projections will destroy this structure. However, it has great potential when using for instance a learned and thus unstructured dictionary, cp. [2].

In the following section we will investigate under which conditions on the deterministic dictionary $\mathbf{\Phi}$ its combination with a random measurement matrix will have small isometry constants. By Theorem I.1 this determines how many measurements n will be typically required for BP to succeed in reconstructing all signals of sparsity S with respect to the given dictionary. In Section III we will analyse the performance of thresholding, which actually has not yet been considered as a reconstruction algorithm in compressed sensing because of its simplicity and hence resulting limitations. The last section is dedicated to numerical simulations showing the performance of compressed sensing for dictionaries in practice and comparing it to the situation where sparsity is induced by an ONB. Even though we have only been able to conduct a partial analysis of OMP so far (see Appendix B) we will do simulations for all three approaches.

II. Isometry Constants for $A\mathbf{\Phi}$

In order to determine the isometry constants for a matrix of the type $\mathbf{\Psi} = A\mathbf{\Phi}$, where A is an $n \times d$ measurement matrix and $\mathbf{\Phi}$ is a $d \times K$ dictionary, we will follow the approach taken in [3], which was inspired by proofs for the Johnson-Lindenstrauss lemma [1]. We will not discuss this connection further but use as starting point concentration of measure for random variables. This describes the phenomenon that in high dimensions the probability mass of certain random variables concentrates strongly around their expectation.

In the following we will assume that A is an $n\times d$

random matrix that satisfies

$$\mathbb{P}\left(\left|\|Av\|^2-\|v\|^2\right|\geq\varepsilon\|v\|^2\right)\leq 2e^{-c\frac{n}{2}\varepsilon^2},\quad \varepsilon\in(0,1/3)\quad \text{(II.1)}$$

for all $v\in\mathbb{R}^d$ and some constant $c>0$. Let us list some examples of random matrices that satisfy the above condition.

- **Gaussian ensemble:** If the entries of A are independent normal variables with mean zero and variance n^{-1} then

  $$\mathbb{P}(\left|\|Av\|^2-\|v\|^2\right|\geq\varepsilon\|v\|^2)\leq 2e^{-\frac{n}{2}(\frac{\varepsilon^2}{2}-\frac{\varepsilon^3}{3})},\quad \varepsilon\in(0,1),\quad \text{(II.2)}$$

  see e.g. [1], [3]. In particular, (II.1) holds with $c=1/2-1/9=7/18$.
- **Bernoulli ensemble:** Choose the entries of A as independent realizations of $\pm1/\sqrt{n}$ random variables. Then again (II.2) is valid, see [1], [3]. In particular (II.1) holds with $c=7/18$.
- **Isotropic subgaussian ensembles:** In generalization of the two examples above, we can choose the rows of A as $\frac{1}{\sqrt{n}}$-scaled independent copies of a random vector $Y\in\mathbb{R}^d$ that satisfies $\mathbb{E}|\langle Y,v\rangle|^2=\|v\|^2$ for all $v\in\mathbb{R}^d$ and has subgaussian tail behaviour. See [18, eq. (3.2)] for details.
- **Basis transformation:** If we take any valid random matrix A and a (deterministic) orthogonal $d\times d$ matrix U then it is easy to see that also AU satisfies the concentration inequality (II.1). In particular, this applies to the Bernoulli ensemble although in general AU and A have different probability distributions.

Using the concentration inequality (II.1) we can now investigate the local and subsequently the global restricted isometry constants of the $n\times K$ matrix $A\mathbf{\Phi}$.

Lemma II.1. *Let A be a random matrix of size $n\times d$ drawn from a distribution that satisfies the concentration inequality* (II.1)*. Extract from the $d\times K$ dictionary $\mathbf{\Phi}$ any sub-dictionary $\mathbf{\Phi}_\Lambda$ of size S, i.e., $|\Lambda|=S$ with (local) isometry constant $\delta_\Lambda=\delta_\Lambda(\mathbf{\Phi})$. For $0<\delta<1$ we set*

$$\nu:=\delta_\Lambda+\delta+\delta_\Lambda\delta. \quad \text{(II.3)}$$

Then

$$(1-\nu)\|x\|^2\leq\|A\mathbf{\Phi}_\Lambda x\|^2\leq\|x\|^2(1+\nu)\quad \text{(II.4)}$$

with probability exceeding

$$1-2\left(1+\frac{12}{\delta}\right)^S e^{-\frac{c}{9}\delta^2 n}. \quad \text{(II.5)}$$

<u>**Proof:**</u> First we choose a finite ε_1-covering of the unit sphere in $\mathbb{R}^S$, i.e., a set of points Q, with $\|q\|=1$ for all $q\in Q$, such that for all $\|x\|=1$

$$\min_{q\in Q}\|x-q\|\leq\varepsilon_1$$

for some $\varepsilon_1\in(0,1)$. According to Lemma 2.2 in [18] there exists such a Q with $|Q|\leq(1+2/\varepsilon_1)^S$. Applying the measure concentration in (II.1) with $\varepsilon_2<1/3$ to all the points $\mathbf{\Phi}_\Lambda q$ and taking the union bound we get

$$(1-\varepsilon_2)\|\mathbf{\Phi}_\Lambda q\|^2\leq\|A\mathbf{\Phi}_\Lambda q\|^2\leq(1+\varepsilon_2)\|\mathbf{\Phi}_\Lambda q\|^2$$

for all $q\in Q$ with probability larger than

$$1-2\left(1+\frac{2}{\varepsilon_1}\right)^S e^{-cn\varepsilon_2^2}.$$

Define ν as the smallest number such that

$$\|A\mathbf{\Phi}_\Lambda x\|^2\leq(1+\nu)\|x\|^2, \quad \text{(II.6)}$$

for all x supported on Λ.

Now we estimate ν in terms of $\varepsilon_1,\varepsilon_2$. We know that for all x with $\|x\|=1$ we can choose a q such that $\|x-q\|\leq\varepsilon_1$ and get

$$\begin{aligned}\|A\mathbf{\Phi}_\Lambda x\| &\leq \|A\mathbf{\Phi}_\Lambda q\|+\|A\mathbf{\Phi}_\Lambda(x-q)\|\\ &\leq(1+\varepsilon_2)^{\frac12}\|\mathbf{\Phi}_\Lambda q\|+\|A\mathbf{\Phi}_\Lambda(x-q)\|\\ &\leq(1+\varepsilon_2)^{\frac12}(1+\delta_\Lambda)^{\frac12}+(1+\nu)^{\frac12}\varepsilon_1.\end{aligned}$$

Since ν is the smallest possible constant for which (II.6) holds it also has to satisfy

$$\sqrt{1+\nu}\quad\leq\quad\sqrt{1+\varepsilon_2}\sqrt{1+\delta_\Lambda}+\varepsilon_1\sqrt{1+\nu}.$$

Simplifying the above equation yields

$$(1+\nu)\leq\frac{1+\varepsilon_2}{(1-\varepsilon_1)^2}(1+\delta_\Lambda).$$

Now we choose $\varepsilon_1 = \delta/6$ and $\varepsilon_2 = \delta/3 < 1/3$. Then

$$\frac{1+\varepsilon_2}{(1-\varepsilon_1)^2} = \frac{1+\delta/3}{(1-\delta/6)^2} = \frac{1+\delta/3}{1-\delta/3+\delta^2/36} < \frac{1+\delta/3}{1-\delta/3} = 1 + \frac{2\delta/3}{1-\delta/3} < 1+\delta.$$

Thus,

$$\nu \quad < \quad \delta + \delta_\Lambda(1+\delta).$$

To get the lower bound we operate in a similar fashion.

$$\begin{aligned} \|A\mathbf{\Phi}_\Lambda x\| &\geq \|A\mathbf{\Phi}_\Lambda q\| - \|A\mathbf{\Phi}_\Lambda(x-q)\| \\ &\geq (1-\varepsilon_2)^{\frac{1}{2}}(1-\delta_\Lambda)^{\frac{1}{2}} - (1+\nu)^{\frac{1}{2}}\varepsilon_1. \end{aligned}$$

Now square both sides and observe that $\nu < 1$ (otherwise we have nothing to show). Then we finally arrive at

$$\begin{aligned} \|A\mathbf{\Phi}_\Lambda x\|^2 &\geq \left((1-\varepsilon_2)^{\frac{1}{2}}(1-\delta_\Lambda)^{1/2} - \varepsilon_1\sqrt{2}\right)^2 \\ &\geq \cdots \geq 1 - \delta_\Lambda - \varepsilon_2 - 2\varepsilon_1\sqrt{2} \\ &\geq 1 - \delta_\Lambda - \delta \geq 1 - \nu. \end{aligned}$$

This completes the proof. □

Note that the choice of ε_1 and ε_2 in the previous proof is not the only one possible. While our choice has the advantage of resulting in an appealing form of ν in (II.3), others might actually yield better constants.

Based on the previous theorem it is easy to derive an estimation of the global restricted isometry constants of the composed matrix $\mathbf{\Psi} = A\mathbf{\Phi}$.

Theorem II.2. *Let $\mathbf{\Phi} \in \mathbb{R}^{d\times K}$ be a dictionary of size K in $\mathbb{R}^d$ with restricted isometry constant $\delta_S(\mathbf{\Phi})$, $S \in \mathbb{N}$. Let $A \in \mathbb{R}^{n\times d}$ be a random matrix satisfying (II.1) and assume*

$$n \geq C\delta^{-2}\Big(S\log(K/S) + \log(2e(1+12/\delta)) + t\Big) \qquad \text{(II.7)}$$

for some $\delta \in (0,1)$ and $t > 0$. Then with probability at least $1 - e^{-t}$ the composed matrix $\mathbf{\Psi} = A\mathbf{\Phi}$ has restricted isometry constant

$$\delta_S(A\mathbf{\Phi}) \leq \delta_S(\mathbf{\Phi}) + \delta(1 + \delta_S(\mathbf{\Phi})). \qquad \text{(II.8)}$$

The constant satisfies $C \leq 9/c$.

<u>Proof:</u> By Lemma II.1 we can estimate the probability that a sub-dictionary $\mathbf{\Psi}_\Lambda = (A\mathbf{\Phi})_\Lambda = A\mathbf{\Phi}_\Lambda$, $\Lambda \subset \{1,\dots,K\}$ fails to have (local) isometry constants $\delta_\Lambda(\mathbf{\Psi}) \leq \delta_\Lambda(\mathbf{\Phi}) + \delta + \delta_\Lambda(\mathbf{\Phi})\delta$ by

$$\begin{aligned} &\mathbb{P}\big(\delta_\Lambda(\mathbf{\Psi}) > \delta_\Lambda(\mathbf{\Phi}) + \delta + \delta_\Lambda(\mathbf{\Phi})\delta\big) \\ &\qquad\qquad \leq 2\big(1 + \frac{12}{\delta}\big)^S e^{-\frac{c}{9}\delta^2 n}. \end{aligned}$$

By taking the union bound over all $\binom{K}{S}$ possible sub-dictionaries of size S we can estimate the probability of $\delta_S(\mathbf{\Psi}) = \sup_{\Lambda\subset\{1,\dots,K\},|\Lambda|=S} \delta_\Lambda(\mathbf{\Psi})$ *not* satisfying (II.8) by

$$\begin{aligned} &\mathbb{P}\big(\delta_S(\mathbf{\Psi}) > \delta_S(\mathbf{\Phi}) + \delta(1+\delta_S(\mathbf{\Phi}))\big) \\ &\qquad\qquad \leq 2\binom{K}{S}\left(1+\frac{12}{\delta}\right)^S e^{-\frac{c}{9}\delta^2 n}. \end{aligned}$$

Using $\binom{K}{S} \leq (eK/S)^S$ (Stirling's formula) and requiring that the above term is less than e^{-t} shows the claim. □

Note that for fixed δ and t condition (II.7) can be expressed in the more compact form

$$n \geq CS\log(K/S).$$

Moreover, if the dictionary $\mathbf{\Phi}$ is an orthonormal basis then $\delta(\mathbf{\Phi}) = 0$ and we recover essentially the previously known estimates of the isometry constants for a random matrix A, see e.g. [3, Theorem 5.2].

Now that we have established how the isometry constants of a deterministic dictionary $\mathbf{\Phi}$ are affected by multiplication with a random measurement matrix, we only need some more initial information about $\mathbf{\Phi}$, before we can finally apply the result to compressed sensing of signals that are sparse in $\mathbf{\Phi}$. The following little lemma gives a very crude estimate of the isometry constants of $\mathbf{\Phi}$ in terms of its coherence μ or Babel function $\mu_1(k)$, which are defined as

$$\mu := \max_{i\neq j} |\langle\varphi_i, \varphi_j\rangle|, \qquad \text{(II.9)}$$

$$\mu_1(k) := \max_{|\Lambda|=k, j\notin\Lambda} \sum_{i\in\Lambda} |\langle\varphi_i, \varphi_j\rangle|. \qquad \text{(II.10)}$$

Lemma II.3. *For a dictionary with coherence μ and Babel function $\mu_1(k)$ we can bound the restricted isometry constants by*

$$\delta_S \leq \mu_1(S-1) \leq (S-1)\mu. \tag{II.11}$$

Proof: Essentially this can be derived from the proof of Lemma 2.3 in [22]. □

Combining this Lemma with Theorem II.2 provides the following estimate of the isometry constants of the composed matrix $\mathbf{\Psi} = A\mathbf{\Phi}$.

Corollary II.4. *Let $\mathbf{\Phi} \in \mathbb{R}^{d\times K}$ be a dictionary with coherence μ. Assume that*

$$S - 1 \leq \frac{1}{16}\mu^{-1}. \tag{II.12}$$

Let $A \in \mathbb{R}^{n\times d}$ be a random matrix satisfying (II.1). Assume that

$$n \geq C_1(S\log(K/S) + C_2 + t).$$

Then with probability at least $1 - e^{-t}$ the composed matrix $A\mathbf{\Phi}$ has restricted isometry constant

$$\delta_S(\mathbf{\Psi}) \leq 1/3. \tag{II.13}$$

The constants satisfy $C_1 \leq 138.51\, c^{-1}$ and $C_2 \leq \log(1250/13) + 1 \approx 5.57$. In particular, for the Gaussian and Bernoulli ensemble $C_1 \leq 356.18$.

Proof: By Lemma II.3 the restricted isometry constant of $\mathbf{\Phi}$ satisfies

$$\delta_S(\mathbf{\Phi}) \leq (S-1)\mu \leq 1/16.$$

Hence, choosing $\delta = 13/(3\cdot 17)$ yields

$$\begin{aligned}\delta(A\mathbf{\Phi}) &\leq \delta_S(\mathbf{\Phi}) + \delta(1 + \delta_S(\mathbf{\Phi}))\\ &\leq \frac{1}{16} + \frac{13}{3\cdot 17}(1 + \frac{1}{16}) = 1/3.\end{aligned}$$

Plugging this particular choice of δ into Theorem II.2 yields the assertion. □

Of course, the numbers $1/16$ and $1/3$ in (II.12) and (II.13) were just arbitrarily chosen. Other choices will only result in different constants C_1, C_2. Combining the previous result with Theorem I.1 yields a result on stable recovery by Basis Pursuit of sparse signals in a redundant dictionary. We leave the straightforward task of formulating the precise statement to the interested reader. We just want to point out that this recovery result is uniform in the sense that a single matrix A can ensure recovery of *all* sparse signals.

The constants C_1 and C_2 of Corollary II.4 are certainly not optimal; however, we did not further pursue the task of improving them. In the case of a Gaussian ensemble A and an orthonormal basis $\mathbf{\Phi}$ recovery conditions for BP with quite small constants were obtained in [21] and precise asymptotic results can be found in [11]. One might raise the objection that the condition $S - 1 \leq \frac{1}{16\mu}$ in Corollary II.4 is too weak for practial applications. A lower bound on the coherence in terms of the dictionary size is

$$\mu > \sqrt{\frac{K-d}{d(K-1)}}$$

and for reasonable dictionaries we can usually expect the coherence to be of the order $\mu \sim C/\sqrt{d}$. The restriction on the sparsity thus is $S < \sqrt{d}/C$. However, compressed sensing is only useful if indeed the sparsity is rather small compared to the dimension d, so this restriction is actually not severe. Moreover, if it is already impossible to recover the support from complete information on the original signal we cannot to expect to do this with even less information.

To illustrate the theorem let us have a look at an example where the dictionary is the union of two ONBs.

Example II.5 (Dirac-DCT)**.** *Assume that our dictionary is the union of the Dirac and the Discrete Cosine Transform bases in $\mathbb{R}^d$ for $d = 2^{2p+1}$. The coherence in this case is $\mu = \sqrt{2/d} = 2^{-p}$ and the number of atoms $K = 2^{2p+2}$. If we assume the sparsity of the signal to be smaller than 2^{p-6} we get the following crude estimate for the number of necessary samples to have $\delta_{4S}(A\mathbf{\Phi}) < 1/3$ as recommended for recovery by BP in Theorem I.1,*

$$n \geq C_1(4S(2p\log 2 - \log S) + C_2 + t)$$

with the constants $C_1 \approx 138.51\, c^{-1}$ and $C_2 \approx 5.57$ from Corollary II.4.

In comparison if the signal is sparse in just the Dirac basis we can estimate the necessary number of samples to have $\delta_{4S}(A) < 1/3$ with Theorem II.2 as

$$n \geq C_1'(4S(2p\log 2 - \log 2S) + C_2' + t)$$

with $C_1' = \left(\frac{13}{17}\right)^2 C_1$ and $C_2' \approx 5.3$, thus implying an improvement of roughly the factor $(\frac{17}{13})^2 \approx 1.71$.

III. Recovery by Thresholding

In this section we investigate recovery from random measurements by thresholding. Since thresholding works by comparing inner products of the signal with the atoms an essential ingredient will be stability of inner products under multiplication with a random matrix A, i.e.,

$$\langle Ax, Ay\rangle \approx \langle x, y\rangle.$$

The exact result that we will use is summarised in the following lemma.

Lemma III.1. *Let $x, y \in \mathbb{R}^d$ with $\|x\|_2, \|y\|_2 \leq 1$. Assume that A is an $n \times d$ random matrix with independent $\mathcal{N}(0, n^{-1})$ entries (independent of x, y). Then for all $t > 0$*

$$\mathbb{P}\big(|\langle Ax, Ay\rangle - \langle x, y\rangle| \geq t\big) \leq 2\exp\left(-n\frac{t^2}{C_1 + C_2 t}\right), \quad \text{(III.1)}$$

with $C_1 = \frac{8e}{\sqrt{6\pi}} \approx 5.0088$ and $C_2 = \sqrt{8}e \approx 7.6885$.

The analogue statement holds for a random matrix A with independent $\pm 1/\sqrt{n}$ Bernoulli entries. In this case the constants are $C_1 = \frac{4e}{\sqrt{6\pi}} \approx 2.5044$ and $C_2 = 2e \approx 5.4366$.

Note that taking $x = y$ in the lemma provides the concentration inequality (II.1) for Gaussian and Bernoulli matrices (with non-optimal constants however).

The proof of the lemma is rather technical and therefore safely locked away in Appendix awaiting inspection by the genuinely interested reader there. However armed with it, we can now investigate the stability of recovery via thresholding.

Theorem III.2. *Let $\mathbf{\Phi}$ be a $d \times K$ dictionary. Assume that the support x of a signal $y = \mathbf{\Phi}x$, normalised to have $\|y\|_2 = 1$, could be recovered by thresholding with a margin ε, i.e.,*

$$\min_{i\in\Lambda} |\langle y, \varphi_i\rangle| > \max_{k\in\overline{\Lambda}} |\langle y, \varphi_k\rangle| + \varepsilon.$$

Let A be an $n \times d$ random matrix satisfying one of the two probability models of the previous lemma. Then with probability exceeding $1 - e^{-t}$ the support and thus the signal can be reconstructed via thresholding from the n-dimensional measurement vector $s = Ay = A\mathbf{\Phi}x$ as long as

$$n \geq C(\varepsilon)(\log(2K) + t).$$

where $C(\varepsilon) = 4C_1\varepsilon^{-2} + 2C_2\varepsilon^{-1}$ and C_1, C_2 are the constants from Lemma III.1. In particular,

$$C(\varepsilon) \leq C_3\varepsilon^{-2}$$

with $C_3 \leq 4C_1 + 2C_2 \leq 35.42$ for the Gaussian case and $C_3 \leq 20.90$ in the Bernoulli case.

Proof: Thresholding will succeed if we have

$$\min_{i\in\Lambda} |\langle Ay, A\varphi_i\rangle| > \max_{k\in\overline{\Lambda}} |\langle Ay, A\varphi_k\rangle|.$$

So let us estimate the probability that the above inequality does *not* hold,

$$\begin{aligned}
&\mathbb{P}(\min_{i\in\Lambda} |\langle Ay, A\varphi_i\rangle| \leq \max_{k\in\overline{\Lambda}} |\langle Ay, A\varphi_k\rangle|) \\
&\leq \mathbb{P}(\min_{i\in\Lambda} |\langle Ay, A\varphi_i\rangle| \leq \min_{i\in\Lambda} |\langle y, \varphi_i\rangle| - \frac{\varepsilon}{2}) \\
&\quad + \mathbb{P}(\max_{k\in\overline{\Lambda}} |\langle Ay, A\varphi_k\rangle| \geq \max_{k\in\overline{\Lambda}} |\langle y, \varphi_k\rangle| + \frac{\varepsilon}{2})
\end{aligned}$$

The probability of the good components having responses lower than the threshold can be further estimated as

$$\begin{aligned}
&\mathbb{P}(\min_{i\in\Lambda} |\langle Ay, A\varphi_i\rangle| \leq \min_{i\in\Lambda} |\langle y, \varphi_i\rangle| - \frac{\varepsilon}{2}) \\
&\quad\leq \mathbb{P}\left(\bigcup_{i\in\Lambda} \{|\langle Ay, A\varphi_i\rangle| \leq |\langle y, \varphi_i\rangle| - \frac{\varepsilon}{2}\}\right) \\
&\quad\leq \sum_{i\in\Lambda} \mathbb{P}\left(|\langle y, \varphi_i\rangle - \langle Ay, A\varphi_i\rangle| \geq \frac{\varepsilon}{2}\right) \\
&\quad\leq 2|\Lambda| \exp\left(-n\frac{\varepsilon^2/4}{C_1 + C_2\varepsilon/2}\right).
\end{aligned}$$

Similarly we can bound the probability of the bad components being higher than the threshold,

$$\begin{aligned}\mathbb{P}(\max_{k\in\overline{\Lambda}}|\langle Ay, A\varphi_k\rangle| \geq \max_{k\in\overline{\Lambda}}|\langle y,\varphi_k\rangle| + \frac{\varepsilon}{2})\\ \leq \mathbb{P}(\bigcup_{k\in\overline{\Lambda}}\{|\langle Ay, A\varphi_k\rangle| \geq |\langle y,\varphi_k\rangle| + \frac{\varepsilon}{2}\})\\ \leq \sum_{k\in\overline{\Lambda}}\mathbb{P}(|\langle Ay, A\varphi_k\rangle - \langle y,\varphi_k\rangle| \geq \frac{\varepsilon}{2})\\ \leq 2|\overline{\Lambda}|\exp\left(-n\frac{\varepsilon^2/4}{C_1+C_2\varepsilon/2}\right).\end{aligned}$$

Combining these two estimates we see that the probability of success for thresholding is exceeding

$$1-2K\exp\left(-n\frac{\varepsilon^2/4}{C_1+C_2\varepsilon/2}\right).$$

The lemma finally follows from requiring this probability to be higher than $1-e^{-t}$ and solving for n. □

The result above may appear surprising because the number of measurements seems to be independent of the sparsity. The dependence, however, is quite well hidden in the margin ε and the normalization $\|y\|_2 = 1$. For clarification we will estimate ε given the coefficients and the coherence of the dictionary.

Corollary III.3. *Let $\mathbf{\Phi}$ be an $d\times K$ dictionary with Babel function μ_1 defined in (II.10). Assume a signal $y=\mathbf{\Phi}_\Lambda x$ with $|\Lambda| = S$ satisfies the sufficient recovery condition for thresholding,*

$$\frac{|x_{\min}|}{\|x\|_\infty} > \mu_1(S)+\mu_1(S-1), \tag{III.2}$$

where $|x_{\min}| = \min_{i\in\Lambda}|x_i|$. If A is an $n\times d$ random matrix according to one of the probability models in Lemma III.1 then with probability at least $1-e^{-t}$ thresholding can recover x (and hence y) from $s = Ay = A\mathbf{\Phi}x$ as long as

$$\begin{aligned}n \geq & C_3S(1+\mu_1(S-1))(\log(2K)+t)\\ &\cdot\left(\frac{|x_{\min}|}{\|x\|_\infty}-\mu_1(S)-\mu_1(S-1)\right)^{-2}.\end{aligned} \tag{III.3}$$

Here, C_3 is the constant from Theorem III.2.

In the special case that the dictionary is an ONB the signal always satisfies the recovery condition and the bound for the necessary number of samples reduces to

$$n > C_3S\left(\frac{\|x\|_\infty}{|x_{\min}|}\right)^2(\log(2K)+t). \tag{III.4}$$

Proof: The best possible value for ε in Theorem III.2 is quite obviously

$$\begin{aligned}\varepsilon &= \min_{i\in\Lambda}|\langle y/\|y\|_2,\varphi_i\rangle| - \max_{k\in\overline{\Lambda}}|\langle y/\|y\|_2,\varphi_k\rangle|\\ &= \frac{1}{\|y\|_2}\big(|\min_{i\in\Lambda}\sum_{j\in\Lambda}x_j\langle\varphi_j,\varphi_i\rangle|\\ &\qquad - \max_{k\in\overline{\Lambda}}|\sum_{j\in\Lambda}x_j\langle\varphi_j,\varphi_k\rangle|\big)\\ &\geq \frac{1}{\|y\|_2}\left(|x_{\min}| - \|x\|_\infty\mu_1(S-1) - \|x\|_\infty\mu_1(S)\right).\end{aligned}$$

Therefore, we can bound the factor $C(\varepsilon)$ in Theorem III.2 as

$$\begin{aligned}C(\varepsilon) &\leq C_3\varepsilon^{-2}\\ &\leq C_3\frac{\|y\|_2^2}{\|x\|_\infty^2}\cdot\big(\frac{|x_{\min}|}{\|x\|_\infty}-\mu_1(S)-\mu_1(S-1)\big)^{-2}.\end{aligned}$$

To get to the final estimate observe that by Lemma II.3

$$\begin{aligned}\frac{\|y\|_2^2}{\|x\|_\infty^2} = \frac{\|\mathbf{\Phi}_\Lambda x\|_2^2}{\|x\|_\infty^2} &\leq (1+\mu_1(S-1))\frac{\|x\|_2^2}{\|x\|_\infty^2}\\ &\leq (1+\mu_1(S-1))S.\end{aligned}$$

The case of an ONB simply follows from $\mu_1(S)=0$. □

The previous results tell us that as for BP we can choose the number n of samples linear in the sparsity S. However, for thresholding successful recovery additionally depends on the ratio of the largest to the smallest coefficient. Also, in contrast to BP the result is no longer uniform, meaning that the stated success probability is only valid for the given signal x. It does not imply that a single matrix A can ensure recovery for all sparse signals. Indeed, in the case of a Gaussian matrix A and an orthonormal basis $\mathbf{\Phi}$ it is known that once A is randomly chosen then with high probability there exists a sparse signal x (depending on A) such that thresholding fails on x unless the number of samples n is quadratic in the sparsity S, see e.g. [12, Section 7]. This fact seems to generalise to redundant $\mathbf{\Phi}$.

Example III.4 (Dirac-DCT)**.** *Assume again that our dictionary is the union of the Dirac and the Discrete Cosine Transform bases in $\mathbb{R}^d$ for $d = 2^{2p+1}$. The coherence is again $\mu = 2^{-p}$ and the number of atoms $K = 2^{2p+1}$. If we assume the sparsity $S \leq 2^{p-2}$ and balanced coefficients, i.e., $|x_i| = 1$, we get the following crude estimate for the number of necessary samples*

$$n \geq 6C_3\, S(\log(2)(2p+2)+t).$$

If we just allow the use of one of the two ONBs to build the signal, the number of necessary samples reduces to

$$n \geq C_3\, S(\log(2)(2p+1)+t).$$

Again we see that whenever the sparsity $S \lesssim \sqrt{d}$ the results for ONBs and general dictionaries are comparable. At this point it would be nice to have a similar result for OMP. This task seems rather difficult due to stochastic dependency issues and so, unfortunately, we have not been able to do this analysis yet.

IV. Numerical Simulations

In order to give a quantitative illustration of the results in Theorem II.2 and Theorem III.2 we will run numerical simulations using the same dictionary as for the examples, i.e., the combination of the Dirac and the Discrete Cosine Transform bases in $\mathbb{R}^d$, $d = 256$, with coherence $\mu = \sqrt{1/128} \approx 0.0884$, cp. Lemma II.3 for the resulting bound on the isometry constants.

We drew six measurement matrices of size $n \times d$, with n varying between 64 and 224 in steps of 32, by choosing each entry as independent realisation of a centered Gaussian random variable with variance $\sigma^2 = n^{-1}$. Then for every sparsity level S, varying between 4 and 64 in steps of 4, respectively between 2 and 32 in steps of 2 for thresholding, we constructed 100 signals. The support Λ was chosen uniformly at random among all $\binom{K}{S}$ possible supports of the given sparsity S. For BP and OMP the coefficients $(x_i)_{i\in\Lambda}$ of the corresponding entries were drawn from a normalised standard Gaussian distribution while for thresholding we chose them of absolute value one with random signs. Then for each of the algorithms we counted how often the correct support could be recovered. For comparison the same setup was repeated replacing the dictionary with the canonical (Dirac) basis. The results are displayed in Figures 1, 2 and 3.

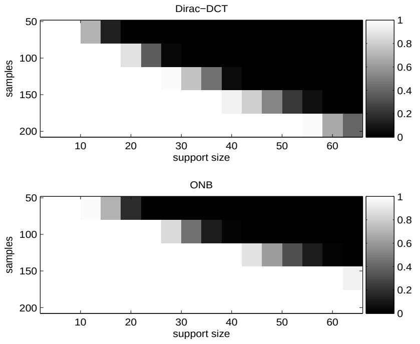

Fig. 1. Recovery Rates for BP as a Function of the Support and Sample Sizes

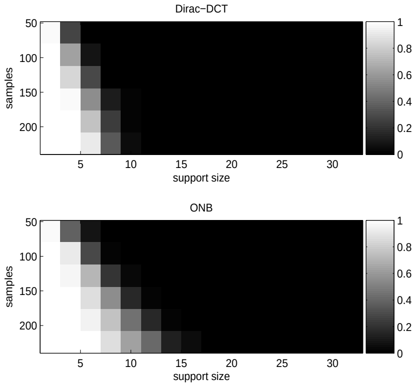

Fig. 2. Recovery Rates for Thresholding as a Function of the Support and Sample Sizes

As predicted by the theorems the necessary number of measurements is higher if the sparsity inducing dictionary is not an ONB. If we compare the three recovery schemes we see that thresholding

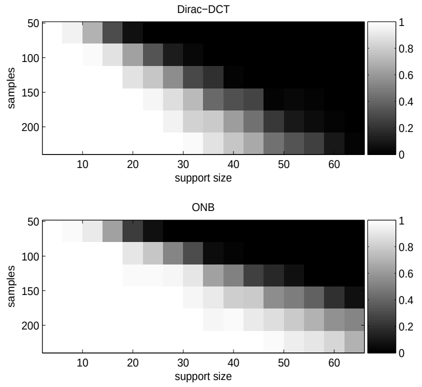

Fig. 3. Recovery Rates for OMP as a Function of the Support and Sample Sizes

gives the weakest results as expected. However, the improvement in performance of BP over OMP is not that significant. This is especially interesting considering that in practice BP is a lot more computationally intensive than OMP.

V. Conclusions & Future Work

We have shown that compressed sensing can also be applied to signals that are sparse in a redundant dictionary. The spirit is that whenever the support can be reconstructed from the signal itself it can also be reconstructed from a small number of random samples with high probability. We have shown that this kind of stability is valid for reconstruction by Basis Pursuit as well as for the simple thresholding algorithm. Thresholding has the advantage of being much faster and easier to implement than BP. However, it has the slight drawback that the number of required samples depends on the ratio of the largest to the smallest coefficient, and recovery is only guaranteed with high probability for a given signal and not uniformly for all signals in contrast to BP. Concerning OMP, we unfortunately could only derive an analysis of its first step, see Appendix B. A complete analysis is still open and seems difficult. However, there is numerical evidence that Orthogonal Matching Pursuit indeed works well. In particular, it is still faster than BP and the required number of samples does not seem to depend on the ratio of the largest to the smallest coefficient (as suggested as well by the partial result Lemma .2).

For the future there remains plenty of work to do. First of all we would like to have a recovery theorem for the full application of OMP comparable to Theorem III.2. However, since in the course of iterating the updated residuals become stochastically dependent on the random matrix A this task does not seem to be straightforward (see also the comments in Appendix B). Then we would like to investigate for which dictionaries it is possible to replace the random Gaussian/Bernoulli matrix by a random Fourier matrix, see also [20]. This would have the advantage that the Fast Fourier Transform can be used in the algorithms in order to speed up the reconstruction. Finally, it would be interesting to relax the incoherence assumption on the dictionary.

Appendix

Our proof uses the following inequality due to Bennett (also refered to as Bernstein's inequality) [4, eq. (7)], see also [23, Lemma 2.2.11].

Theorem .1. *Let $X_1, \ldots, X_n$ be independent random variables with zero mean such that*

$$\mathbb{E}|X_i|^q \leq q! M^{q-2} v_i / 2 \qquad (.1)$$

for every $m \geq 2$ and some constants M and v_i, $i = 1, \ldots, n$. Then for $x > 0$

$$\mathbb{P}\left(|\sum_{i=1}^{n} X_i| \geq x \right) \leq 2e^{-\frac{1}{2}\frac{x^2}{v+Mx}}$$

with $v = \sum_{i=1}^{n} v_i$.

Now let us prove Lemma III.1. Observe that

$$\langle Ax, Ay \rangle = \frac{1}{n} \sum_{\ell=1}^{n} \sum_{k=1}^{d} \sum_{j=1}^{d} g_{\ell k} g_{\ell j} x_k y_j$$

where $g_{\ell k}$, $\ell = 1, \ldots, n, k = 1, \ldots, d$ are independent standard Gaussians. We define the random variable

$$Y := \sum_{k,j=1}^{d} g_k g_j x_k y_j$$

where again the g_k, $k = 1, \dots, d$ are independent standard Gaussians. Then we can write

$$\langle Ax, Ay\rangle = \frac{1}{n}\sum_{\ell=1}^{n} Y_\ell$$

where the Y_ℓ are independent copies of Y.

Let us investigate Y. The expectation of Y is easily calculated as

$$\mathbb{E}Y = \sum_{k=1}^{d} x_k y_k = \langle x, y\rangle.$$

Hence, also $\mathbb{E}\left[\langle Ax, Ay\rangle\right] = \langle x, y\rangle$. Now let

$$Z := Y - \mathbb{E}Y = \sum_{k\neq j} g_j g_k x_j x_k + \sum_k (g_k^2 - 1) x_k y_k.$$

The random variable Z is known as Gaussian chaos of order 2.

Thus, we have to show the moment bound (.1) for the random variable Z. Note that $\mathbb{E}Z = 0$. A general bound for Gaussian chaos (see [16, p. 65]) gives

$$\mathbb{E}|Z|^q \leq (q-1)^q \left(\mathbb{E}|Z|^2\right)^{q/2} \qquad (.2)$$

for all $q \geq 2$. Using Stirling's formula, $q! = \sqrt{2\pi q}\, q^q e^{-q} e^{R_q}$, $\frac{1}{12q+1} \leq R_q \leq \frac{1}{12q}$, we further obtain, for all $q \geq 3$:

$$\begin{aligned}
\mathbb{E}|Z|^q &= q! \frac{(q-1)^q}{e^{R_q}\sqrt{2\pi q}\, e^{-q} q^q} \left(\mathbb{E}|Z|^2\right)^{q/2} \\
&= \left(1 - \frac{1}{q}\right)^q \frac{e^2 q!}{e^{R_q}\sqrt{2\pi q}} \left(e^2 \mathbb{E}|Z|^2\right)^{(q-2)/2} \mathbb{E}|Z|^2 \\
&\leq \frac{e}{e^{R_q}\sqrt{2\pi q}} q! \left(e^2 \mathbb{E}|Z|^2\right)^{(q-2)/2} \mathbb{E}|Z|^2 \\
&\leq q! \left(e (\mathbb{E}|Z|^2)^{1/2}\right)^{q-2} \frac{e}{\sqrt{6\pi}} \mathbb{E}|Z|^2.
\end{aligned}$$

Hence, the moment bound (.1) holds for all $q \geq 3$ with

$$M = e\left(\mathbb{E}|Z|^2\right)^{1/2}, \qquad v = \frac{2e}{\sqrt{6\pi}} \mathbb{E}|Z|^2,$$

and by direct inspection it then also holds for $q = 2$. So let us determine $\mathbb{E}|Z|^2$. Using independence of the g_k we obtain

$$\begin{aligned}
\mathbb{E}|Z|^2 = \mathbb{E}\Bigg[&\sum_{j\neq k}\sum_{j'\neq k'} g_j g_k g_{j'} g_{k'} x_j y_k x_{j'} y_{k'} \\
&+ 2\sum_{j\neq k}\sum_{k'} g_j g_k (g_{k'}^2 - 1) x_j y_k x_{k'} y_{k'} \\
&+ \sum_{k}\sum_{k'} (g_k^2 - 1)(g_{k'}^2 - 1) x_k y_k x_{k'} y_{k'}\Bigg] \\
&= \sum_{k\neq j} \mathbb{E}[g_j^2]\mathbb{E}[g_k^2] x_j^2 y_k^2 + \sum_k \mathbb{E}[(g_k^2 - 1)^2] x_k^2 y_k^2 \\
&= \sum_{k\neq j} x_j^2 y_k^2 + 2\sum_k x_k^2 y_k^2 \\
&= \|x\|_2^2 \|y\|_2^2 + \langle x, y\rangle^2 \ \leq\ 2 \qquad (.3)
\end{aligned}$$

since by assumption $\|x\|_2, \|y\|_2 \leq 1$. Denoting by Z_ℓ, $\ell = 1, \dots, n$ independent copies of Z, Theorem .1 yields

$$\begin{aligned}
&\mathbb{P}\big(|\langle Ax, Ay\rangle - \langle x, y\rangle| \geq t\big) \\
&\qquad = \mathbb{P}\left(|\sum_{\ell=1}^{n} Z_\ell| \geq nt\right) \\
&\qquad \leq 2e^{-\frac{1}{2}\frac{n^2 t^2}{nv + nMt}} = 2e^{-n\frac{t^2}{C_1 + C_2 t}},
\end{aligned}$$

with $C_1 = \frac{4e}{\sqrt{6\pi}} E|Z|^2 \leq \frac{8e}{\sqrt{6\pi}} \approx 5.0088$ and $C_2 = 2e\sqrt{2} \approx 7.6885$.

For the case of Bernoulli random matrices the proof is completely analogue. We just have to replace the standard Gaussians g_k by ±1 Bernoulli variables. In particular, the estimate (.2) for the chaos variable Z is still valid, see [16, p. 105]. Furthermore, for Bernoulli variables g_k we clearly have $E[g_k^2] = 1$ and $\mathbb{E}[(g_k^2 - 1)] = 0$. Hence, the corresponding estimate in (.3) yields $\mathbb{E}|Z|^2 \leq 1$, and we end up with the constants $C_1 = \frac{4e}{\sqrt{6\pi}} \approx 2.5044$ and $C_2 = 2e = 5.4366$.

Although we have not yet been able to conduct a theoretical analysis of the full OMP algorithm in our setting, the following result at least analyses the first step of OMP.

Lemma .2. *Let $\mathbf{\Phi}$ be a $d \times K$ dictionary and set $y = \mathbf{\Phi}_\Lambda x$ with $|\Lambda| = S$. Further, choose $A \in \mathbb{R}^{n\times d}$ at random according to the Gaussian or Bernoulli distribution, and take the n-dimensional* *measurement vector* $s = Ay = A\boldsymbol{\Phi}_\Lambda x$. *Suppose that the sparsity* S *satisfies*

$$S \le \frac{1}{4\mu}. \qquad (.4)$$

and that the number of measurements exceeds

$$n \ge CS(\log(K) + t).$$

Then the probability that OMP fails to recover an element of the support Λ *in the first step is smaller than* e^{-t}.

Proof: OMP selects an element of Λ in the first step if

$$\max_{i\in\Lambda} |\langle A\varphi_i, A\boldsymbol{\Phi}_\Lambda x\rangle| > \max_{k\notin\Lambda} |\langle A\varphi_k, A\boldsymbol{\Phi}_\Lambda x\rangle| \quad (.5)$$

To bound the probability of not satisfying the above condition we will use the same trick as for thresholding, i.e.,

$$\begin{aligned}\mathbb{P}\big(\max_{i\in\Lambda} |\langle A\varphi_i, A\boldsymbol{\Phi}_\Lambda x\rangle| &> \max_{k\notin\Lambda} |\langle A\varphi_k, A\boldsymbol{\Phi}_\Lambda x\rangle|\big)\\ &< \mathbb{P}\big(\max_{i\in\Lambda} |\langle A\varphi_i, A\boldsymbol{\Phi}_\Lambda x\rangle| \le p\big)\\ &\quad + \mathbb{P}\big(\max_{k\notin\Lambda} |\langle A\varphi_k, A\boldsymbol{\Phi}_\Lambda x\rangle| \ge p\big), \quad (.6)\end{aligned}$$

and then make a suitable choice for p. Let's start by estimating the first term in the expression above. Without loss of generality we may assume $\|x\|_2 = 1$. Using the Cauchy Schwarz inequality we see that

$$\begin{aligned}\max_{i\in\Lambda} |\langle A\varphi_i, A\boldsymbol{\Phi}_\Lambda x\rangle| &= \|(A\boldsymbol{\Phi}_\Lambda)^\star(A\boldsymbol{\Phi}_\Lambda)x\|_\infty\\ &\ge S^{-1/2}\|(A\boldsymbol{\Phi}_\Lambda)^\star(A\boldsymbol{\Phi}_\Lambda)x\|_2\\ &\ge \frac{1-\delta_\Lambda(A\boldsymbol{\Phi})}{\sqrt{S}}.\end{aligned}$$

Thus we can bound the first probability as

$$\begin{aligned}\mathbb{P}\big(\max_{i\in\Lambda} |\langle A\varphi_i, &A\boldsymbol{\Phi}_\Lambda x\rangle| \le p\big)\\ &\le \mathbb{P}\big(\delta_\Lambda(A\boldsymbol{\Phi}) \ge 1 - p\sqrt{S}\big).\end{aligned}$$

If we now set $p = \frac{1}{2\sqrt{S}}$ what we need to do is check when $\delta_\Lambda(A\boldsymbol{\Phi}) \le 1/2$ holds. Condition (.4) implies by (II.11) that $\delta_\Lambda(\boldsymbol{\Phi}) \le 1/4$. Setting $\delta := 1/5$ we obtain $\nu = \delta_\Lambda(\boldsymbol{\Phi}) + \delta + \delta_\Lambda(\boldsymbol{\Phi})\delta \le 1/2$ (see condition (II.3)), and by Lemma II.1 we have $\delta_\Lambda(A\boldsymbol{\Phi}) \le 1/2$ fails only with probability smaller than

$$2\big(1 + \frac{12}{\delta}\big)^S e^{-\frac{c}{9}\delta^2 n} = 2\cdot 61^S e^{-\frac{c}{225}n}.$$

Next we bound the second probability in (.6) for our choice $p = \frac{1}{2\sqrt{S}}$.

$$\begin{aligned}\mathbb{P}\Big(\max_{k\notin\Lambda} |\langle A\varphi_k, A\boldsymbol{\Phi}_\Lambda x\rangle| &\ge \frac{1}{2\sqrt{S}}\Big)\\ \le \sum_{k\notin\Lambda} \mathbb{P}\Big(&|\langle A\varphi_k, A\boldsymbol{\Phi}_\Lambda x\rangle| \ge \frac{1}{2\sqrt{S}}\Big)\\ \le \sum_{k\notin\Lambda} \mathbb{P}\Big(&|\langle A\varphi_k, A\boldsymbol{\Phi}_\Lambda x\rangle - \langle\varphi_k, \boldsymbol{\Phi}_\Lambda x\rangle|\\ &\qquad \ge \frac{1}{2\sqrt{S}} - \langle\varphi_k, \boldsymbol{\Phi}_\Lambda x\rangle\Big).\end{aligned}$$

By Cauchy Schwarz, $\|x\|_2 = 1$, and condition (.4) we have

$$\begin{aligned}|\langle\varphi_k, \boldsymbol{\Phi}_\Lambda x\rangle| \le \|\boldsymbol{\Phi}_\Lambda^\star\varphi_k\|_2 &= \Big(\sum_{i\in\Lambda} |\langle\varphi_i, \varphi_k\rangle|^2\Big)^{1/2}\\ &\le \sqrt{S}\mu \le \frac{1}{4\sqrt{S}} \qquad (.7)\end{aligned}$$

Hence, using Lemma III.1 we can further estimate the second probability by

$$\begin{aligned}\sum_{j\notin\Lambda} \mathbb{P}\Big(|\langle A\varphi_j, A\boldsymbol{\Phi}_\Lambda x\rangle - \langle\varphi_j, \boldsymbol{\Phi}_\Lambda x\rangle| &\ge \frac{1}{4\sqrt{S}}\Big)\\ &\qquad (.8)\\ \le 2(K-S)\exp\Big(-\frac{n}{16S}\,\frac{1}{C_1 + C_2/(4\sqrt{S})}\Big)&.\end{aligned}$$

Combining the two estimates we can bound the probability of OMP failing in the first step by

$$\begin{aligned}&2{\cdot}61^S \exp\Big(-n\,\frac{c}{225}\Big)\\ &\quad + 2(K-S)\exp\Big(-\frac{n}{16S}\,\frac{1}{C_1 + C_2/(4\sqrt{S})}\Big),\end{aligned}$$

which can (with an easy but boring calculation) be shown to be smaller than e^{-t}, whenever

$$n \ge CS(\log(4K) + t) \qquad (.9)$$

for $C < \max\{225/c, 16C_1 + 4C_2\}$.

Of course, we would like to analyse also the further steps of OMP rather than only the first one.

Indeed, we conjecture that Lemma .2 holds literally for the full application of OMP (with possibly a different constant C). However, starting with the second step the coefficients $x^{(r)}$ of the current residual as well as the selected subsets become stochastically dependent on the random matrix A, and in this case Lemma III.1 does not apply any more in (.8). These subtle stochastic dependencies can be resolved when the columns of $A\mathbf{\Phi}$ are stochastically independent [14]. However, this happens only when $\mathbf{\Phi}$ is the identity matrix (or A Gaussian and $\mathbf{\Phi}$ orthonormal), and in the general case of redundant dictionaries it seems rather difficult to analyse the full OMP algorithm.

References

[1] D. Achlioptas. Database-friendly random projections. In *Proc. 20th Annual ACM SIGACT-SIGMOD-SIGART Symp. on Principles of Database Systems*, pages 274–281, 2001.

[2] M. Aharon, M. Elad, and A. Bruckstein. K-SVD: An algorithm for designing of overcomplete dictionaries for sparse representation. *IEEE Trans. on Signal Processing.*, 54(11):4311–4322, November 2006.

[3] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. A simple proof of the restricted isometry property for random matrices. *Constr. Approx.*, to appear.

[4] G. Bennett. Probability inequalities for the sum of independent random variables. *J. Am. Stat. Assoc.*, 57:33–45, 1962.

[5] E. Candès, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.*, 59(8):1207–1223, 2006.

[6] E. Candès and T. Tao. Near optimal signal recovery from random projections: universal encoding strategies? *IEEE Trans. Inf. Theory*, 52(12):5406–5425, 2006.

[7] S. Chen, D. Donoho, and M. Saunders. Atomic decomposition by Basis Pursuit. *SIAM J. Sci. Comput.*, 20(1):33–61, 1999.

[8] G. Davis, S. Mallat, and M. Avellaneda. Adaptive greedy approximations. *Constr. Approx.*, 13(1):57–98, 1997.

[9] D. Donoho. Compressed Sensing. *IEEE Trans. Inf. Theory*, 52(4):1289–1306, 2006.

[10] D. Donoho, M. Elad, and V. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. Inf. Theory*, 52(1):6–18, 2006.

[11] D. Donoho and J. Tanner. Counting faces of randomly-projected polytopes when the projection radically lowers dimension. *Preprint arXiv:math.MG/0607364*, 2006.

[12] D. L. Donoho. For most large underdetermined systems of linear equations the minimal ℓ^1-norm solution is also the sparsest solution. *Comm. Pure Appl. Math.*, 59(6):797–829, 2006.

[13] I. Drori. Fast ℓ_1 minimization by iterative thresholding for multidimensional NMR spectroscopy. *Preprint*, 2006.

[14] A. C. Gilbert and J. A. Tropp. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Trans. Inform. Theory*, to appear.

[15] R. Gribonval, B. Mailhe, H. Rauhut, K. Schnass, and P. Vandergheynst. Average case analysis of multichannel thresholding. In *Proc. IEEE ICASSP07, Honolulu*, 2007.

[16] M. Ledoux and M. Talagrand. *Probability in Banach spaces. Isoperimetry and processes.* Springer-Verlag, Berlin, Heidelberg, NewYork, 1991.

[17] S. G. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Trans. Signal Process.*, 41(12):3397–3415, 1993.

[18] S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann. Uniform uncertainty principle for Bernoulli and subgaussian ensembles. *Preprint*, 2006.

[19] B. Natarajan. Sparse approximate solutions to linear systems. *SIAM J. Comput.*, 24:227–234, 1995.

[20] H. Rauhut. Random sampling of sparse trigonometric polynomials. *Appl. Comput. Harm. Anal.*, 22(1):16–42, 2007.

[21] M. Rudelson and R. Vershynin. Sparse reconstruction by convex relaxation: Fourier and Gaussian measurements. In *Proc. CISS 2006 (40th Annual Conference on Information Sciences and Systems)*, 2006.

[22] J. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Trans. Inf. Theory*, 50(10):2231–2242, 2004.

[23] A. Van der Vaart and J. Wellner. *Weak convergence and empirical processes*. Springer-Verlag, 1996.