Average Performance Analysis for Thresholding

Karin Schnass* and Pierre Vandergheynst
Signal Processing Institute
Swiss Federal Institute of Technology
Lausanne, Switzerland
{karin.schnass, pierre.vandergheynst}@epfl.ch
EPFL-STI-ITS-LTS2
CH-1015 Lausanne
Tel: +41 21 693 6874
Fax: +41 21 693 7600
EDICS: DSP-TFSR, DSP-FAST

***Abstract*— In this article is shown that with high probability the thresholding algorithm can recover signals that are sparse in a redundant dictionary as long as the *2-Babel function* is growing slowly. This implies that it can succeed for sparsity levels up to the order of the ambient dimension. The theoretical bounds are illustrated with numerical simulations. As an application of the theory *sensing dictionaries* for optimal average performance are characterised and their performance is tested numerically.**

I. Introduction

By now the possibilities of sparsely approximating signals in redundant dictionaries have been widely recognised in the signal processing community. Consequently there exist a lot of algorithms, like thresholding, Matching Pursuits or the Basis Pursuit Principle, see [6], [1], [3], [2], to name just the most popular, which are successfully employed to find sparse approximations. However so far the theoretical analysis of these algorithms was reduced to studying their worst case performance, see [4], [7]. The resulting worst case bounds for recoverable sparsity levels turned out to be overly pessimistic and quite in contrast to the much better performance in practice. So if we define the coherence of the dictionary as the maximal absolute inner product between two different normalised atoms, i.e. $\mu = \max_{i\neq j} |\langle \varphi_i, \varphi_j \rangle|$, the worst case analysis tells us that we can recover superpositions of S atoms as long as:

$$S \lesssim \mu^{-1} \approx \sqrt{d},$$

where d is the ambient dimension, while in practice it is usually possible to recover supports of sizes proportional to d. Motivated by the desire to better understand and capture the performance of an algorithm together with a dictionary people have started to analyse the average case performance. In a recent paper Tropp was able to show that random sub-dictionaries of a general dictionary are very likely to be well conditioned as long as their size is of the order of $\mu^{-2} \approx d$, [8, Theorem B]. As an application of this result it is shown that a signal constructed from a random superposition of S atoms with coefficients drawn from a continuous distribution has almost surely no sparser representation, [8, Theorem 12]. If additionally the signs of the coefficients are drawn from a uniform distribution then this representation is with high probability recoverable via Basis Pursuit.

Theorem 1 ([8], Theorem 13)**.** *Assume that $\mathbf{\Phi}_\Lambda$, the matrix we get by extracting the S atoms in Λ from all K atoms in $\mathbf{\Phi}$, has least singular value $\sigma_{\min}(\mathbf{\Phi}_\Lambda) \geq \sqrt{1/2}$. Assume also that the signal $y = \mathbf{\Phi}_\Lambda x_\Lambda$ is synthesised from a coefficient sequence x_Λ whose signs form a Steinhaus sequence, i.e. $\varepsilon_i = x_i/|x_i|$, $i \in \Lambda$, are independent realisations of the random variable e^{iX} with X uniformly distributed on $(0, 2\pi)$. Then the probability that Basis Pursuit fails to recover x_Λ from y is less than*

$$\mathbb{P}(\textit{BP fails}) \leq 2K \exp\Big(-\frac{1}{8\mu^2 S}\Big) \qquad (1)$$

One of the conclusions of the above results is that Basis Pursuit is able to recover sparse signal representations even when the sparsity level is higher than the worst case barrier of $\sqrt{d}$.

However the problem is that in practice Basis Pursuit is simply too complex. Consider for instance image compression, a small picture of size 64×64 already results in $d = 4096$. Taking a dictionary with reasonable redundancy 2 means that we have to solve a convex optimisation problem in $\mathbb{R}^{8192}$. On the other hand one would typically be happy to recover the 100 most important components of the signal. Unfortunately this is still more than $64 = \sqrt{d}$ signifying the worst case performance bottleneck for simpler algorithms like thresholding or the Matching Pursuits, see [6], [7]. In the following we will therefore analyse the average behaviour of thresholding to find out that also here the recoverable sparsity scales with the ambient dimension. Again the result will be in terms of the coherence μ or rather the 2-Babel function μ_2, defined as

$$\begin{aligned} \mu_2(\Lambda, k) &= \big(\sum_{i\in\Lambda} |\langle \varphi_i, \varphi_k \rangle|^2\big)^{\frac{1}{2}}, \\ \mu_2(\Lambda) &= \max_{k\notin\Lambda} \mu_2(\Lambda, k), \\ \mu_2(S) &= \max_{|\Lambda|=S} \mu_2(\Lambda). \end{aligned}$$

From the estimate $\mu_2(S) \leq \sqrt{S}\mu$ we see that the 2-Babel function grows much slower than the 1-Babel function, which is of the order $S\mu$.

II. Theoretical Analysis

Before the theoretical analysis we give a quick reminder of how the thresholding algorithm works in the table below and introduce the probabilistic model we assume for our signals y.

Thresholding
find: Λ that contains the indices corresponding to the S largest values of $\lvert\langle y, \varphi_i\rangle\rvert$
reconstruct: $x_\Lambda = (\mathbf{\Phi}_\Lambda)^\dagger y$, $\tilde{y} = \mathbf{\Phi}_\Lambda x_\Lambda$

Signal Model:

$$y = \mathbf{\Phi}_\Lambda x_\Lambda = \sum_{i\in\Lambda} x_i \varphi_i, \qquad x_i = \varepsilon_i |x_i|, \quad \forall i \in \Lambda,$$

where $\mathbf{\Phi}$ is a dictionary of K normalised atoms and $\mathbf{\Phi}_\Lambda$ a subdictionary of all atoms with indices in Λ and $|\Lambda| = S$. While the support Λ and the absolute magnitude of the coefficients are considered to be arbitrary, the signs ε_i form either a Steinhaus sequence or a Rademacher sequence, i.e. $\varepsilon_i = \pm 1$ with equal probability.

Theorem 2. *Let's abbreviate the event "Thresholding fails to recover the component φ_i" as "$\circledS_i$" and "Thresholding fails to recover all components" as "$\circledS$". Under the above signal model*

$$a) \qquad \mathbb{P}(\circledS_i) < 2(K-S+1)\exp\Big(-\frac{|x_i|^2}{\|x\|_\infty^2}\frac{c}{8\mu_2^2(S)}\Big)$$

$$b) \qquad \mathbb{P}(\circledS) < 2K\exp\Big(-\frac{|x_{\min}|^2}{\|x\|_\infty^2}\frac{c}{8\mu_2^2(S)}\Big)$$

where $c = 1$ for Steinhaus and $c = 1/16$ for Rademacher sequences and $x_{\min}$ denotes the coefficient with smallest absolute value.

The proof is a straightforward application of the following large deviation inequalities.

Theorem 3. *Let α be an arbitrary real/complex vector and ε a Rademacher/Steinhaus sequence. Then for all $t > 0$*

$$\mathbb{P}(|\sum_i \varepsilon_i \alpha_i| > t) \leq 2e^{-c_0 t^2/\|\alpha\|_2^2}$$

where $c_0 = 1/32$ for Rademacher and $c_0 = 1/2$ for Steinhaus sequences.

For a proof for Steinhaus sequences see [8] and references therein. The proof for Rademacher sequences can be found in [5, Section 4]. We now turn to the proof of Theorem 2.

Proof: [Theorem 2] We can bound the probability of not recovering φ_i by the probability that its inner product with the signal is lower than a threshold p while the inner product of an atom not in the support is higher than the threshold.

$$\begin{aligned}
\mathbb{P}(\circledS_i) &= \mathbb{P}\big(|\langle y,\varphi_i\rangle| < \max_{k\in\overline{\Lambda}} |\langle y,\varphi_k\rangle|\big) \\
&\leq \mathbb{P}\big(|\langle y,\varphi_i\rangle| < p\big) + \mathbb{P}\big(\max_{k\in\overline{\Lambda}} |\langle y,\varphi_k\rangle| > p\big) \\
&\leq \mathbb{P}\big(|\langle y,\varphi_i\rangle| < p\big) + \mathbb{P}\big(\bigcup_{k\in\overline{\Lambda}} |\langle y,\varphi_k\rangle| > p\big) \\
&\leq \mathbb{P}\big(|\langle y,\varphi_i\rangle| < p\big) + \sum_{k\in\overline{\Lambda}} \mathbb{P}\big(|\langle y,\varphi_k\rangle| > p\big)
\end{aligned}$$

The probability of the correlation of the signal with φ_i being smaller than the threshold can be further bounded as,

$$\begin{aligned}
\mathbb{P}\big(|\langle y,\varphi_i\rangle| < p\big) &= \mathbb{P}\big(|\sum_{j\in\Lambda} x_j\langle\varphi_j,\varphi_i\rangle| < p\big) \\
&= \mathbb{P}\big(|x_i + \sum_{j\neq i} x_j\langle\varphi_j,\varphi_i\rangle| < p\big) \\
&\leq \mathbb{P}\big(|\sum_{j\neq i} x_j\langle\varphi_j,\varphi_i\rangle| > |x_i| - p\big).
\end{aligned}$$

Choosing the threshold as $p = |x_i|/2$ and using Theorem 3 we arrive at,

$$\begin{aligned}
\mathbb{P}\big(|\langle y,\varphi_i\rangle| \leq p\big) &< \mathbb{P}\big(|\sum_{j\neq i} \varepsilon_j |x_j| \langle\varphi_j,\varphi_i\rangle| > \tfrac{1}{2}|x_i|\big) \\
&\leq 2\exp\Big(-\frac{c_0}{4}\frac{|x_i|^2}{\sum_{j\neq i}|x_j|^2|\langle\varphi_j,\varphi_i\rangle|^2}\Big) \\
&\leq 2\exp\Big(-\frac{|x_i|^2}{\|x\|_\infty^2}\frac{c}{8\mu_2^2(S-1)}\Big).
\end{aligned}$$

Similarly we can estimate the probability of the correlation of an atom not in the support being larger than the threshold,

$$\begin{aligned}
\mathbb{P}\big(|\langle y,\varphi_k\rangle| > p\big) &\leq \mathbb{P}\big(|\sum_{j\in\Lambda} \varepsilon_j |x_j| \langle\varphi_j,\varphi_k\rangle| > \tfrac{1}{2}|x_i|\big) \\
&\leq 2\exp\Big(-\frac{c_0}{4}\frac{|x_i|^2}{\sum_{j\in\Lambda}|x_j|^2|\langle\varphi_j,\varphi_k\rangle|^2}\Big) \\
&\leq 2\exp\Big(-\frac{|x_i|^2}{\|x\|_\infty^2}\frac{c}{8\mu_2^2(S)}\Big).
\end{aligned}$$

Putting it all together we finally arrive at,

$$\begin{aligned}
\mathbb{P}(\circledS_i) &\leq 2\exp\Big(-\frac{|x_i|^2}{\|x\|_\infty^2}\frac{c}{8\mu_2^2(S-1)}\Big) \\
&\qquad + |\overline{\Lambda}| 2\exp\Big(-\frac{|x_i|^2}{\|x\|_\infty^2}\frac{c}{8\mu_2^2(S)}\Big) \\
&\leq 2(K-S+1)\exp\Big(-\frac{|x_i|^2}{\|x\|_\infty^2}\frac{c}{8\mu_2^2(S)}\Big).
\end{aligned}$$

To estimate the probability of thresholding failing to recover all components we can proceed in the same fashion. Essentially we just need to adapt the choice of the threshold p.

$$\begin{aligned}
\mathbb{P}(\circledS) &= \mathbb{P}\big(\min_{i\in\Lambda} |\langle y,\varphi_i\rangle| < \max_{k\in\overline{\Lambda}} |\langle y,\varphi_k\rangle|\big) \\
&\leq \mathbb{P}\big(\min_{i\in\Lambda} |\langle y,\varphi_i\rangle| < p\big) + \mathbb{P}\big(\max_{k\in\overline{\Lambda}} |\langle y,\varphi_k\rangle| > p\big).
\end{aligned}$$

The first probability can be expanded as

$$\begin{aligned}\mathbb{P}\big(\min_{i\in\Lambda}|\langle y,\varphi_i\rangle|<p\big) &\le \mathbb{P}\big(\min_{i\in\Lambda}|x_i+\sum_{j\neq i}x_j\langle\varphi_j,\varphi_i\rangle|<p\big)\\ &\le \mathbb{P}\big(\min_{i\in\Lambda}(|x_{\min}|-|\sum_{j\neq i}x_j\langle\varphi_j,\varphi_i\rangle|)<p\big)\\ &\le \mathbb{P}\big(\max_{i\in\Lambda}|\sum_{j\neq i}x_j\langle\varphi_j,\varphi_i\rangle|>|x_{\min}|-p\big)\\ &\le \sum_{i\in\Lambda}\mathbb{P}(|\sum_{j\neq i}x_j\langle\varphi_j,\varphi_i\rangle|>|x_{\min}|-p)\end{aligned}$$

Now we choose as threshold $p=|x_{\min}|/2$ and using Theorem 3 get the bound:

$$\mathbb{P}\big(\min_{i\in\Lambda}|\langle y,\varphi_i\rangle|<p\big)\le 2S\exp\Big(-\frac{|x_{\min}|^2}{\|x\|_\infty^2}\frac{c}{8\mu_2^2(S-1)}\Big).$$

Repeating the steps above we can estimate the probability of an atom not in the support having higher correlation than the threshold as

$$\mathbb{P}\big(\max_{k\in\overline{\Lambda}}|\langle y,\varphi_k\rangle|>p\big)\le 2(K-S)\exp\Big(-\frac{|x_{\min}|^2}{\|x\|_\infty^2}\frac{c}{8\mu_2^2(S)}\Big).$$

In combination this leads to the final bound:

$$\mathbb{P}(\text{☹})<2K\exp\Big(-\frac{|x_{\min}|^2}{\|x\|_\infty^2}\frac{c}{8\mu_2^2(S)}\Big).$$

□

Comparing the above result for Steinhaus sequences to Theorem 1 we see that the essential difference in the failure probability bound for the two algorithms is the additional coefficient $\frac{x_{\min}^2}{\|x\|_\infty^2}$ in the exponent for thresholding. This means that for coefficients of constant absolute magnitude the two algorithms should perform comparably. Also it promises a good behaviour of thresholding as long as the coefficients are reasonably well balanced and in that case makes it an interesting low complexity alternative to BP.

III. Numerical Simulations

A. An Experiment with Dimensions

To show numerically how the recovery rates of thresholding scale with the dimension we conducted the following experiment. In dimensions 2^p, $p=8\dots12$ a dictionary made up the Dirac and the Discrete Cosine Transform bases was constructed. The coherence of these dictionaries is $\mu=\sqrt{2/d}$ and the 2-Babel function behaves approximately like $\mu_2(S)\approx\sqrt{S/d}$. For each dimension and relative sparsity level S/d, 1000 signals were constructed by randomly choosing a support and coefficients with constant absolute value one and random signs, $x_i=\pm1$ with equal probability. Then we counted how often thresholding was able to recover the full support.

From the theorem we know that thresholding will fail with small probability as long as

$$\mu_2^2(S)\lesssim\frac{c'}{\log(2K)}\qquad\Rightarrow\qquad\frac{S}{d}\lesssim\frac{c'}{(p+1)\log 2}.$$

If we compare these theoretical bounds to the simulation results displayed in Figure 1 we see that they reflect the average behaviour quite well. For the bounds as plotted in the figure we chose $c'=0.3$ which is somewhat better than the theorem suggests ($c'\approx\frac{1}{128}$).

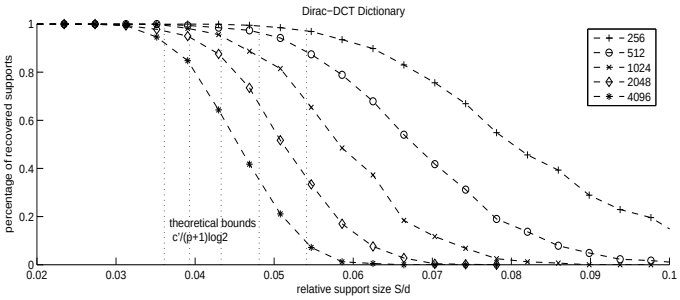

Fig. 1. Comparison of Numerical Recovery Rates and Theoretical Recovery Bounds

B. An Application

As an application of Theorem 2 we will construct a sensing dictionary to improve the average performance of a dictionary for thresholding. The concept of sensing dictionaries for greedy algorithms was introduced in [6], where sensing dictionaries improving the worst case performance were characterised and constructed. The basic idea in the case of thresholding would be to determine which components to pick from the absolute inner products of the signal with atoms in a special sensing dictionary $\psi_i\in\mathbf{\Psi}$ instead of the original dictionary, see table below.

Thresholding with Sensing Dictionary $\mathbf{\Psi}$
find: Λ that contains the indices corresponding to the S largest values of $\lvert\langle y,\psi_i\rangle\rvert$
reconstruct: $x_\Lambda=(\mathbf{\Phi}_\Lambda)^\dagger y$, $\tilde{y}=\mathbf{\Phi}_\Lambda x_\Lambda$

The only a priori requirements we pose on the sensing matrix are that it has the same size as the original dictionary and that the inner products between corresponding atoms, i.e. with the same index, are one, $\langle\varphi_i,\psi_i\rangle=1$. The average performance of thresholding with a sensing dictionary can be analysed as before. We only need to adjust the definition of the 2-Babel function to describe the pseudo Gram matrix $\mathbf{\Psi}^\star\mathbf{\Phi}$ instead of the Gram matrix.

$$\begin{aligned}\tilde{\mu}_2(\Lambda,k) &= \big(\sum_{i\in\Lambda}|\langle\varphi_i,\psi_k\rangle|^2\big)^{\frac{1}{2}}\\ \tilde{\mu}_2(\Lambda) &= \max_{k\notin\Lambda}\tilde{\mu}_2(\Lambda,k)\\ \tilde{\mu}_2(S) &= \max_{|\Lambda|=S}\tilde{\mu}_2(\Lambda).\end{aligned}$$

The analogue of part b) of Theorem 2 now reads:

Theorem 4. *Under the same assumptions on the signal model as in the previous section we can bound the probability that thresholding with the sensing matrix $\mathbf{\Psi}$ fails as*

$$\mathbb{P}(\text{☹})<2K\exp\Big(-\frac{|x_{\min}|^2}{\|x\|_\infty^2}\frac{c}{8\tilde{\mu}_2^2(S)}\Big).$$

Proof: Follow the proof of Theorem 2 mutatis mutandis. □

One deduction from the Theorem is that a sensing matrix for good average performance should minimise the 2-Babel function. Let us consider what this would mean for the distribution of the off-diagonal entries of the pseudo-Gram matrix. Since to calculate the 2-Babel function we always sum over the squared maximal entries in one row we would want the distribution as flat as possible, see [6] for more details in that direction. However this might not be optimal in the sense that dampening the (absolutely) maximal off side entries comes at the price of rising otherwise small entries. Taking a look back at the proof of Theorem 2 we see that this could have devastating results since more of the relevant off diagonal entries $\Psi^\star\mathbf{\Phi}_\Lambda$ could have been raised than lowered, increasing the failure probability in the end,

$$\mathbb{P}(\circledcirc) \leq 2(\sum_{k\in\Lambda} e^{-c'/\mu_2^2(\Lambda/k,k)} + \sum_{k\notin\Lambda} e^{-c'/\mu_2^2(\Lambda,k)}).$$

Another consideration suggesting a different approach is that the support Λ might be chosen at random as well. Then the optimal sensing dictionary would need to minimise

$$\mathbb{E}_\Lambda(\sum_{k\in\Lambda} e^{-c'/\mu_2^2(\Lambda/k,k)} + \sum_{k\notin\Lambda} e^{-c'/\mu_2^2(\Lambda,k)}),$$

and thus should not have a pseudo gram matrix with a flat distribution of the entries. The problem of how the off-diagonal entries should be distributed is quite intricate and a definite topic of further study. For now we only observe that all off-diagonal entries are equally likely to contribute to the final bound. So as simplified but feasible approach we will reduce their cumulative destructive power by finding the sensing dictionary that minimises the Frobenius norm of the pseudo-Gram matrix.

$$\begin{aligned}\mathbf{\Psi}_0 &= \arg\min_{\langle\psi_i,\varphi_i\rangle=1} \|\mathbf{\Psi}^\star\mathbf{\Phi}\|_F \\ &= \arg\min_{\langle\psi_i,\varphi_i\rangle=1} \big(\sum_i\sum_j |\langle\varphi_i,\psi_j\rangle|^2\big)^{\frac{1}{2}}.\end{aligned}$$

The advantage of the problem as formulated above is that there exists an analytic solution, that can be easily derived using Lagrange multipliers. To make our lives easier we consider the square of the objective function $\|\mathbf{\Psi}^\star\mathbf{\Phi}\|_F^2$. We then derive both the objective and the constraint function along ψ_i,

$$\begin{aligned}\frac{d}{d\psi_j}\|\mathbf{\Psi}^\star\mathbf{\Phi}\|_F^2 &= \sum_i 2\langle\varphi_i,\psi_j\rangle\varphi_i = 2\mathbf{\Phi}\mathbf{\Phi}^\star\psi_j \\ \frac{d}{d\psi_j}\langle\varphi_j,\psi_j\rangle &= \varphi_j\end{aligned}$$

Since at the minimum the derivatives need to be parallel we set $2\mathbf{\Phi}\mathbf{\Phi}^\star\psi_j = c_j\varphi_j$ which leads to $\psi_j = \frac{c_j}{2}(\mathbf{\Phi}\mathbf{\Phi}^\star)^{-1}\varphi_j$. If we choose the constants c_j appropriately to ensure $\langle\varphi_j,\psi_j\rangle = 1$ and collect them in the diagonal matrix D, we see that the optimal sensing matrix is just the rescaled transpose of the Moore Penrose pseudo inverse,

$$\mathbf{\Psi}_0 = (\mathbf{\Phi}\mathbf{\Phi}^\star)^{-1}\mathbf{\Phi}D = (\mathbf{\Phi}^\dagger)^\star D.$$

To test the performance of an average sensing matrix we did the following small experiment. We built a dictionary of 256 atoms that are randomly distributed on the sphere in $\mathbb{R}^{128}$. For each support size between 1 and 20 we constructed 1000 signals by choosing the support set uniformely at random and coefficients of absolute value one but with random signs, i.e. $x_i = \pm 1$ with equal probability. We then compared how often thresholding could recover the full support when using the original dictionary, the worst case sensing matrix, see [6], and the average case sensing matrix. The results are displayed in Figure 2

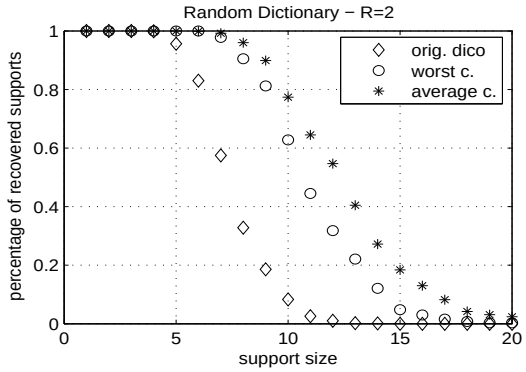

Fig. 2. Recovery Rates for Different Sensing Dictionaries

The improvement already gained by using the worst case sensing matrix is further increased by using the average case sensing matrix. The performance differences are also well reflected by the Frobenius norms of the (pseudo-) Gram matrices.

dictionary	original	worst case	average case
$\|\mathbf{\Psi}^\star\mathbf{\Phi}\|_F$	27.7217	23.8902	22.6743

So there is a large decrease in norm between the original dictionary and the worst case sensing matrix accounting for the large performance gap and a smaller decrease between the worst case and the average case sensing matrix reflecting a smaller improvement.

REFERENCES

[1] G. Davis, S. Mallat, and M. Avellaneda. Adaptive greedy approximations. *Constructive Approximation*, 13:57–98, 1997. Springer-Verlag New York Inc.

[2] D. Donoho and M. Elad. Optimally sparse representation in general (non-orthogonal) dictionaries via ℓ_1 minimization. *Proc. Nat. Aca. Sci.*,, 100(5):2197–2202, March 2003.

[3] J. J. Fuchs. Extension of the Pisarenko method to sparse linear arrays. *IEEE Transactions on Signal Processing*, 45(2413-2421), October 1997.

[4] J. J. Fuchs. Detection and estimation of superimposed signals. In *Proc. IEEE ICASSP98*, volume 3, pages 1649–1652, 1998.

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

[6] K. Schnass and P. Vandergheynst. Dictionary preconditioning for greedy algorithms. *submitted to IEEE Trans. Signal Processing*, 2007.

[7] J. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2242, October 2004.

[8] J. Tropp. Random subdictionaries of general dictionaries. *preprint*, 2006.