AVERAGE CASE ANALYSIS OF SPARSE RECOVERY WITH THRESHOLDING : NEW BOUNDS BASED ON AVERAGE DICTIONARY COHERENCE

Mohammad Golbabaee and Pierre Vandergheynst

Signal Processing Institute, Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland
E-mail:{mohammad.golbabaei, pierre.vandergheynst}@epfl.ch

ABSTRACT

This paper analyzes the performance of the simple thresholding algorithm for sparse signal representations. In particular, in order to be more realistic we introduce a new probabilistic signal model which assumes randomness for both the amplitude and also the location of nonzero entries. Based on this model we show that thresholding in average can correctly recover signals for much higher sparsity levels than was previously reported. The bounds we obtain in this paper are based on a new concept of average dictionary coherence and are shown to be much sharper than in former works [1, 2].

Index Terms— Sparse representation, redundant dictionary, cumulative and average coherence, Thresholding

1. INTRODUCTION

Sparse approximation is a fairly new branch of applied mathematics and signal processing that is mostly concerned with the problem of finding a sparse solution to an undetermined system of equations:

$$y = \Phi x, \quad s.t. \quad ||x||_0 = s. \tag{1}$$

The notation $\|x\|_0$ counts the number of nonzero entries in the d-dimensional vector x and Φ is a $d \times m$ matrix, often called dictionary, where $m >> d >> s$. It is often assumed that the columns (atoms) of the dictionary have unit ℓ_2 norm. There has been an impressive research activity recently in using sparse signal models such as (1) for problems such as denoising, compression, missing data estimation (inpainting) [3, 4, 5, 6]. The rationale behind these applications is to extend a key property of wavelet-like decompositions : finding a basis where signals are sparsely approximated which somehow means compacting information in few coefficients such that it becomes very robust to noise or can be efficiently encoded. Using redundant dictionaries offers the possibility to model complex behaviors that a single basis cannot catch.

This increase in flexibility comes at the expense of higher algorithmic complexity. In particular, problem (1) is known to be NP-hard and cannot thus be tackled directly. A common solution is to relax the sparsity constraint by penalizing $\|x\|_1$ instead of $\|x\|_0$. That modification leads to a simpler problem, known as Basis Pursuit (BP) [7] that can be easily solved with linear programming methods. Another common approach is to attack the problem using greedy heuristics such as Matching Pursuit (MP) or Orthogonal Matching Pursuit (OMP) [8, 9, 10]. The good behavior of these algorithms in practice has fueled an intense activity for understanding their theoretical properties and limitations. A detailed account is far beyond the scope of this paper, however so far it is schematically known that the solution of (1) can be recovered by BP or greedy algorithms provided the dictionary is not too redundant and the signal is sparse enough. More precisely, let φ_k be the k-th column of Φ and define the Cumulative Coherence of order l as follows :

$$\mu_l(s) = \max_{\Lambda \text{ s.t. } |\Lambda|=s} \max_j \mu_l(j, \Lambda). \tag{2}$$

where, $\mu_l(j, \Lambda) = (\sum_{i \in \Lambda} |\langle \varphi_j, \varphi_i \rangle|^l)^{1/l}$. Worst case analysis guarantees that those algorithms can recover the correct sparse vector x provided its sparsity is at most of the order square root of the ambient dimension d, i.e

$$\mu_1(s) \approx \frac{s}{\sqrt{d}} \lessapprox 1. \tag{3}$$

It is important to realize that this bound is very lose, which should come as no surprise since it represents the worst possible configuration for the coefficient vector x. In reality, most signals will be substantially milder than this extreme case and indeed simulations show that *in average*, algorithms can recover signals that are much less sparse than (3). It is therefore of primal importance to understand the behavior of sparse approximation algorithms in the average sense and not just in the worst case. So far, there have been very few attempts at characterizing these kind of average case behavior, and it is the main goal of this short paper to actually provide new results in that direction.

More precisely, the main contribution of this paper resides in the analysis of randomness in sparse approximation. Here, in addition to drawing nonzero coefficients at random, the

This research was supported by Swiss National Science Foundation through grant 200021-117884.

sparse support (Λ) of x, i.e the location of those nonzero entries, is also randomly chosen. We apply this idea to study the performance of a recent greedy-based algorithm which is called Thresholding due to its less complexity. This paper is organized as following. In the next section we start by defining precisely our sparse signal model, which is inspired from [2, 1]. In Section 3 we quickly survey similar approaches and highlight the main differences with the current contribution. Our main result is then stated and proved in Section 4 and Section 5 illustrates our findings with numerical simulations. Finally, we give conclusions and highlight interesting future research alleys in Section 6.

2. SIGNAL MODEL AND RECOVERY ALGORITHM

Let us first rewrite equation (1) in a form that highlights the support of the sparse vector in a cleaner way :

$$y = \Phi_\Lambda x_\Lambda. \tag{4}$$

where x_Λ is a s-dimensional vector which contains nonzero entries of x and Φ_Λ is the restriction of the dictionary matrix to columns listed by Λ (support set of cardinality s). In order to avoid considering only the worst case, we draw x_Λ from i.i.d Gaussian entries with zero mean and unit variance and we will assume that the locations of nonzero elements (Λ) have a uniform distribution among all $\binom{s}{m}$ possible choices.

Moreover the focus of this paper is on the simple Thresholding algorithm. Basically this algorithm simplifies MP by just looking in a single pass for the s most correlated atoms of the dictionary with the signal to determine the support set (Λ) i.e. indices of the s biggest entries of $\Phi^* y$. Eventually, synthesis coefficients are computed by projecting the signal onto this recovered subspace using $\mathbf{x}_\Lambda = \Phi_\Lambda^\dagger y$, where $\dagger$ denotes the Moore-Penrose pseudo-inverse. Note though that in this paper we will only focus on recovering the correct support Λ.

Choosing Thresholding as the sparse recovery algorithm makes things easier. However this come at the expense of very sensitive behavior to the dynamic range of nonzero coefficients, $R = \min |x_i|/||x||_\infty$. In comparison to what we have seen for BP or MP in (3), here the worst case analysis indicates a recovery constraint as $s \lessapprox R/\mu^{-1}$, where R can be very small in our signal model, see [2].

3. PRIOR ART

Regarding average case performances, only a handful of research papers seems relevant. In [11], Tropp studied random subdictionaries Φ_Λ of a dictionary and showed this randomness on the support in average, leads to a better recovery condition. The early model which incorporates randomness for the coefficients was introduced in [1], where average case of single channel thresholding was first studied. However that paper focuses on a case that only coefficients' signs are chosen at random, while their amplitude is kept constant, quite tough constrain regarding real world signals. What we have found the closest to ours is [2] where coefficients drawn from a Gaussian distribution and also some average case results were obtained for thresholding and OMP. However, this model is only valid in the case of multichannel signals and does not scale down to a single signal as we are studying here. Moreover in both [1] and [2], the probability of recovering Λ is proportional to the number of elements m in the dictionary, which can be very large for redundant dictionaries. As a consequence, the associated bounds are very loose. In the next section, we study the average performances of thresholding on the signal model depicted in Section 2 where both Λ and x_Λ are chosen at random. As we will see below, we obtain bounds that are much closer to practice than those obtained in the aforementioned papers. Moreover, our results are expressed in terms of the average coherence of the dictionary instead of the cumulative coherence in former results.

4. MAIN RESULTS

Before getting any further, let us define a new quantity that plays a central role in this paper.

Definition 1 *The Average Coherence of order s is given by*

$$\rho^2(s) = \underset{s.t.\ |\Lambda|=s}{\mathbb{E}_\Lambda} \{\beta(\Lambda)\}, \tag{5}$$

where $\beta(\Lambda) = \max_{j \notin \Lambda} \mu_2^2(j, \Lambda)$.

It is easy to see that the cumulative coherence is linked to $\beta(\Lambda)$ since ${\mu_2}^2(s) = \max_\Lambda \beta(\Lambda)$ but while μ_2 measures the highest coherence among s atoms, $\rho^2(s)$ tries to determine this value in average by choosing Λ from a uniform distribution. It is now time to go through the main theorem of the paper in order to upper bound the failure probability of Thresholding.

Theorem 2 *The probability that Thresholding fails to identify the correct support set Λ of a s sparse signal is bounded by:*

$$P_f \leq s\left(1 - 2Q(\gamma)\right) + 2(m-s)Q(\frac{\gamma}{\rho(s)}). \tag{6}$$

where,

$$\gamma^2 = \frac{2\rho^2(s)}{1-\rho^2(s)} \ln\big(\frac{m-s}{s\,\rho(s)}\big), \tag{7}$$

and $Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{r^2/2} dr$.

Proof. We start the proof by assuming an arbitrary support set Λ and defining the failure event as the fact that the correlation between the signal and any of the out-support atoms becomes larger than the weakest in-support correlation i.e :

$$P_f(\Lambda) = P(\min_{i \in \Lambda} |\langle \varphi_i, y \rangle| < \max_{j \notin \Lambda} |\langle \varphi_j, y \rangle|) \tag{8}$$

$$= P(\min_{i \in \Lambda} |\varphi_i^* \Phi_\Lambda x_\Lambda| < \max_{j \notin \Lambda} |\varphi_j^* \Phi_\Lambda x_\Lambda|). \tag{9}$$

Further, we upper bound this probability by the probability that the minimum correct component is smaller than a fixed threshold η, while the projections on the wrong components are larger than this threshold :

$$\begin{aligned} P_f(\Lambda) &\leq P(\min_{i\in\Lambda} |\varphi_i^* \Phi_\Lambda x_\Lambda| < \eta) + P(\max_{j\notin\Lambda} |\varphi_j^* \Phi_\Lambda x_\Lambda| > \eta) \\ &\leq P(\bigcup_{\forall i\in\Lambda} |\varphi_i^* \Phi_\Lambda x_\Lambda| < \eta) + P(\bigcup_{\forall j\notin\Lambda} |\varphi_j^* \Phi_\Lambda x_\Lambda| > \eta) \qquad (10) \\ &\leq \sum_{i\in\Lambda} P(|\varphi_i^* \Phi_\Lambda x_\Lambda| < \eta) + \sum_{j\notin\Lambda} P(|\varphi_j^* \Phi_\Lambda x_\Lambda| > \eta). \qquad (11) \end{aligned}$$

Note that in (10) we use a union bound in both summations. On the other hand, as we assumed the elements of x_Λ to be i.i.d Gaussian with zero mean and unit variance, we can rewrite (11) as follows:

$$P_f(\Lambda) \leq \sum_{i\in\Lambda} 1 - 2Q\big(\frac{\eta}{||\varphi_i^* \Phi_\Lambda||_2}\big) + \sum_{j\notin\Lambda} 2Q\big(\frac{\eta}{||\varphi_j^* \Phi_\Lambda||_2}\big). \qquad (12)$$

According to Definition 1, for the off support atoms we have $||\varphi_j^* \Phi_\Lambda||_2^2 \leq \beta(\Lambda)$. Moreover, assuming a normalized dictionary, for the components on the support we have $||\varphi_i^* \Phi_\Lambda||_2 \geq 1$. Inserting these bounds in (12), we get:

$$P_f(\Lambda) \leq s\,\big(1 - 2Q(\eta)\big) + 2(m-s)Q\big(\frac{\eta}{\sqrt{\beta(\Lambda)}}\big). \qquad (13)$$

As previously mentioned, this inequality holds for all values of η. Therefore, to have the tightest bound, we take the optimal choice of η which minimizes (13) as follows:

$$\eta_{opt}^2 = \frac{2\beta(\Lambda)}{1-\beta(\Lambda)} \ln\big(\frac{m-s}{s\sqrt{\beta(\Lambda)}}\big). \qquad (14)$$

So far, we have found an upper bound of the failure probability which depends on the choice of the support set. The last step to complete our proof is to average this value over all choices of Λ. Observe that minimizing the failure probability by taking η_{opt}, P_f becomes a concave function of $\beta(\Lambda)$. Using Jensen's inequality, we thus get:

$$\begin{aligned} \mathbb{E}_\Lambda\{P_f(\beta(\Lambda))\} &\leq P_f(\mathbb{E}_\Lambda\{\beta(\Lambda)\}), \quad s.t. \quad |\Lambda| = s \qquad (15) \\ &= P_f(\rho^2(s)). \qquad (16) \end{aligned}$$

Here, P_f is the minimal function based on η_{opt} and (16) is based on our previous definition. Substituting $\rho^2(s)$ instead of $\beta(\Lambda)$ in both (13) and (14) concludes our proof.

Noteworthy to add that for the dictionaries of moderate size, the contribution of the second term in (6) is no more than 8%. Moreover, increasing the dictionary size (d and m) this value becomes less and less. In next section we will verify as d goes to infinity this term becomes zero while the first term stays between zero and one. Therefore for the same γ we can approximate previous bound as:

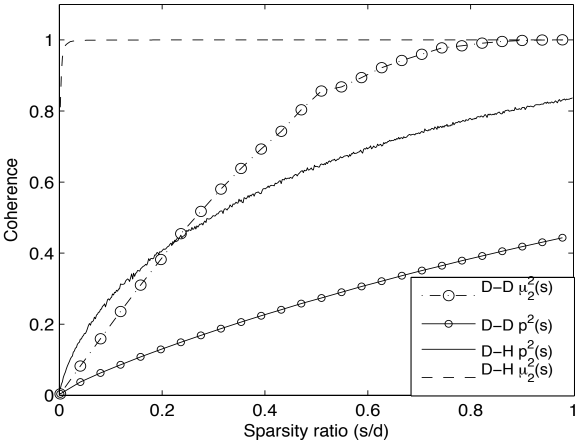

Fig. 1. Comparison between $\rho^2(s)$ and ${\mu_2}^2(s)$ in DCT-Delta (D-D) and DCT-Haar (D-H) of the same size $d = 512$.

$$P_f \lesssim s\,\big(1 - 2Q(\gamma)\big). \qquad (17)$$

5. SIMULATION RESULTS AND ANALYSIS

In this section we would like to highlight what motivates our main theorem in comparison to the former works. In the last section we saw the failure probability can be bounded by the average coherence instead of the maximal one. In order to be more familiar with this new object and also compare it to $\mu_2(s)$ we consider two different dictionaries made by the concatenation of two orthonormal basis : first the $DCT - Delta$ dictionary ($\mu_2(s) = \sqrt{2s/d}$) and the other one is $DCT - Haar$ which has very coherent atoms in low frequencies. As we expected, figure 1 points out the huge difference between ${\mu_2}^2(s)$ in both dictionaries. However, what is more interesting is that in both cases $\rho^2(s)$ stays far below ${\mu_2}^2(s)$. As we can see in this figure, the slope of $\rho^2(s)$ is almost half of ${\mu_2}^2(s)$ in $DCT - Delta$. In $DCT - Haar$, specially for small sparsity levels, there is a significant difference between these two values and moreover while ${\mu_2}^2(s)$ saturates very fast $\rho^2(s)$ stays always less than one. Although based on our calculations $\rho(s)$ scales with the ambient dimension by the same order as $\mu_2(s) \propto \sqrt(s/d)$, the failure probability is very sensitive to the coherence and replacing μ by the average coherence leads to a considerable decrease in P_f and consequently a much tighter bound. Moreover, comparing to the former bounds which are proportional to number of atoms m, we have a slightly better bound as it is proportional to the sparsity level $s << m$.

Considering a very large dictionary size, Theorem 2 asymptotically guaranties a small failure probability as long as

$$s^3 \lesssim \frac{d}{\ln m}. \tag{18}$$

The key point for the proof is using an approximation to simplify (17) as $1 - 2Q(\gamma) \simeq \sqrt{2/\pi}\,\gamma$ for small γ. This is along with experimental results in dictionaries of moderate size and for reasonable s/d. Therefore, we rewrite (17) by using expression (7) to have P_f less than a constant $c << 1$:

$$\begin{aligned} P_f^2 &\lessapprox \frac{s^2\rho^2(s)}{1-\rho^2(s)} \ln\big(\frac{m-s}{s\,\rho(s)}\big) \\ &\approx \frac{s^3}{d}\ln m + \frac{s^3}{d}\ln\frac{d}{s^3} \lesssim c^2. \end{aligned} \tag{19}$$

There, we assume $\rho^2(s)$ is proportional to s/d and also d scales with order higher than s^3. This makes the second term in (19) vanish as d tends to infinity and (18) follows. The resulting scaling law confirms the validity of our assumption and approximation. Moreover, it explains why we can neglect the second term in (6) to reach the simpler upper bound in (17). We can observe that the sparsity ratio (s/d) which is required for perfect recovery is decreasing as the dictionary size grows. However, the algorithm preserves its applicability since for larger dictionaries less nonzero coefficients are sufficient to express signals.

Finally an overall comparison shows: the worst case analysis is pointless in our Gaussian model since the signal range R could be very close to zero. Moreover, the related average case results in [2] are not applicable in the single signal case e.g. it indicates certain failure even in orthonormal basis. However, considering the asymptotic behavior, our theorem ensures almost reliable recovery even in huge dictionaries up to a certain sparsity levels.

6. CONCLUSION AND FUTURE PLANS

In this paper we have considered a new model for sparse signals where both the coefficients and support set are drawn at random. Thanks to the average case study for Thresholding, we have developed a new upper bound on its failure probability which is not anymore related to maximal coherence but to the average one. This fact together with some optimizations results in a significantly tighter bound in comparison to prior works.

As a future plan we would like to extend this idea to the multichannel case where several nodes (channels) observe signals that are jointly sparse in a given dictionary. Since exploiting these inter signal correlations leads to an enormous improvement by increasing number of the channels, we expect that the techniques developed in this paper will indicate recovery almost surely with utilizing very few sensor nodes.

7. REFERENCES

[1] Karin Schnass and Pierre Vandergheynst, "Average Performance Analysis for Thresholding," *to appear in IEEE Signal Processing Letters*.

[2] Rémi Gribonval, Holger Rauhut, Karin Schnass, and Pierre Vandergheynst, "Atoms of all channels, unite! average case analysis of multi-channel sparse recovery using greedy algorithms," *Journal of Fourier Analysis and Applications*, 2007, Submitted to the special issue on Sparsity.

[3] J.L. Starck, M. Elad, and D.L. Donoho, "Image decomposition via the combination of sparse representations and a variational approach," Tech. Rep., CEA-Saclay, DAPNIA/SEDI-SAP, 2004.

[4] M. Elad, J.L. Starck, P. Querre, and D.L. Donoho, "Simultaneous cartoon and texture image inpainting using morphological component analysis (mca)," *Journal of Applied and Computational Harmonic Analysis*, pp. 19:340–358, March 2005.

[5] O.K. Al-Shaykh, E. Miloslavsky, T. Nomura, R. Neff, and A. Zakhor, "Video compression using matching pursuits," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 9, no. 1, pp. 123–143, February 1999.

[6] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Transaction on Image Processing*, vol. 15, pp. 3736–3745, December 2006.

[7] S.S. Chen, D.L. Donoho, and M.A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comp.*, vol. 20, no. 1, pp. 33–61, 1999.

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

[9] Y. C. Pati, R. Rezaifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decompositions," in *Proc. 27th Asilomar Conf. on Signals, Systems and Comput.*, November 1993, vol. 1, pp. 40–44.

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

[11] J.A. Tropp, "On the conditioning of random subdictionaries," *Appl. Comput. Harmonic Anal.*, 2007 (to appear).