Sparse Representations are Most Likely to be the Sparsest Possible

Michael Elad
Department of Computer Science
The Technion - Israel Institute of Technology, Haifa 32000 Israel
email: elad@cs.technion.ac.il

August 1st, 2004

Abstract

Given a signal $\underline{S} \in \mathcal{R}^N$ and a full rank matrix $\mathbf{D} \in \mathcal{R}^{\mathbf{N}\times\mathbf{L}}$ with $N < L$, we define the signal's overcomplete representations as all $\underline{\alpha} \in \mathcal{R}^L$ satisfying $\underline{S} = \mathbf{D}\underline{\alpha}$. Among all the possible solutions, we have special interest in the sparsest one – the one minimizing $\|\underline{\alpha}\|_0$. Previous work has established that a representation is unique if it is sparse enough, requiring $\|\underline{\alpha}\|_0 < \text{Spark}(\mathbf{D})/2$. The measure Spark($\mathbf{D}$) stands for the minimal number of columns from $\mathbf{D}$ that are linearly dependent. This bound is tight – examples can be constructed to show that with Spark($\mathbf{D}$)/2 or more non-zero entries, uniqueness is violated. In this paper we study the behavior of overcomplete representations beyond the above bound. While tight from a worst-case standpoint, a probabilistic point-of-view leads to uniqueness of representations satisfying $\|\underline{\alpha}\|_0 < \text{Spark}(\mathbf{D})$. Furthermore, we show that even beyond this point, uniqueness can still be claimed with high confidence. This new result is important for the study of the average performance of pursuit algorithms – when trying to show an equivalence between the pursuit result and the ideal solution, one must also guarantee that the ideal result is indeed the sparsest.

Keywords: Sparse representation, Overcomplete transforms, Spark, Signature of matrices, Uniqueness.

1 Introduction

1.1 General - Sparse Representations

In signal processing we are often interested in a replacement of the representation, seeking some simplification for an obvious gain. This is the rational behind the so many transforms proposed over the past several centuries, such as the Fourier, Cosine, Wavelets, and many others. The basic idea is to "change language", and describe the signal differently, in the hope that the new description is better for the application in mind. A natural justification for a transform is that given a signal, a representation has already been imposed due to the use of the trivial basis (e.g., samples as a function of time/space), and there is no reason to believe that this representation is the most appropriate one for our needs.

The ease with which linear transforms are operated and analyzed keeps those as the first priority candidates in defining alternative representations. It is therefore not surprising to find that linear transforms are the more popular ones in theory and practice in signal processing. A linear transform is defined through the use of a full rank matrix $\mathbf{D} \in \mathcal{R}^{\mathbf{N}\times\mathbf{L}}$, where $L \geq N$. Given the signal $\underline{S} \in \mathcal{R}^N$,

its representation is defined by

$$\underline{S} = \mathbf{D}\underline{\alpha}, \tag{1}$$

where $\underline{\alpha} \in \mathcal{R}^L$. For the case of $L = N$ (and a non-singular matrix $\mathbf{D}$ due to the full-rank property), the above relationship implies a linear operation both for the forward transform (from $\underline{S}$ to $\underline{\alpha}$) and its inverse. Many of the practical transforms are of this type, and many of them go further and simplify the matrix $\mathbf{D}$ to be structured and unitary, so that its inverse is easier to operate and both directions can be computed with nearly $O(N)$ operations. Such is the case with the DFT, DCT, the Hadamard, orthonormal wavelet and other transforms.

In this paper we are interested in the case of $L > N$, referred to as the overcomplete transforms. When $L > N$, the relationship in (1) is an under-determined linear set of equations, and thus in general it leads to an infinite number of possible solutions. Further information is therefore needed in order to uniquely define the transform, and this is typically achieved by defining the representation as the solution of

$$(P_p) \quad \min_{\underline{\alpha}} \ \|\underline{\alpha}\|_p \ \text{ subject to } \ \underline{S} = \mathbf{D}\underline{\alpha}. \tag{2}$$

For $p = 2$ it is easy to show that again we obtain linearity in both directions (forward and inverse transforms). This case, typically referred to as "Frame Theory", has drawn a lot of attention because of this obvious simplicity. However, it is clear that two-way linearity poses a hard restriction on the space of possibilities, and may cost in performance.

A different and far more complicated approach advocated strongly in recent years is to consider $p = 0$. The ℓ^0 notation is an abused ℓ^p-norm with $p \to 0$, effectively counting the number of non-zeros in the vector $\underline{\alpha}$. In such an approach we seek among all feasible representations (satisfying the constraint in (2)) the one with the fewest non-zero entries, this way achieving an ultimate simplicity in representation. Referring to the matrix $\mathbf{D}$ as a dictionary of signal–prototypes as its columns, we build $\underline{s}$ as a linear combination of only few of these columns, typically referred to as atoms. Thus, we can think of our signal as a molecule, and the forward transform decomposes it to its building atoms, where we try to use the fewest in this construction [1].

From the numerical standpoint, the forward transform, defined as (P_0), is a non-convex and highly non-smooth optimization problem, with many possible local minimum points. Prior work has established that this problem is an NP-Hard one, implying that its complexity grows exponentially with the number of columns in the dictionary [2, 3]. Recent study of this problem and methods to approximate its solution give promising new results, indicating that even though complicated, means exist to solve it at least in some cases using either greedy [4, 5, 6, 7, 8, 9, 10, 11, 12, 13] or convex programming approaches [1, 14, 15, 16, 17, 18, 19, 10, 20]. One aspect of these recent work is the result of uniqueness which will be the focus of this paper.

1.2 The Uniqueness Result - Worst-Case Analysis

We consider the problem

$$(P_0) \quad \min_{\underline{\alpha}} \ \|\underline{\alpha}\|_0 \ \text{ subject to } \ \underline{S} = \mathbf{D}\underline{\alpha}. \tag{3}$$

Previous work has shown that if a feasible solution $\underline{\alpha}_0$ is sparse enough, it can be guaranteed to be the solution of (P_0) [18, 19]. The argument is surprisingly simple and has the following reasoning: For a given dictionary $\mathbf{D}$, its Spark is defined as the smallest number of columns from $\mathbf{D}$ that are

linearly dependent. This scalar characterizes the dictionary with respect to sparse representations from a worst-case standpoint. By definition, the vectors in the null-space of the dictionary $\mathbf{D}\underline{\delta} = \mathbf{0}$ must satisfy $\|\underline{\delta}\|_0 \geq \text{Spark}(\mathbf{D})$, since they combine linearly columns from $\mathbf{D}$ to give the zero vector, and at least Spark such columns are necessary.

If $\underline{\alpha}_0$ represents $\underline{S}$, i.e., $\underline{S} = \mathbf{D}\underline{\alpha}_0$, it implies that all the alternative representations of the same signal are characterized as $\underline{\alpha}_0 + \underline{\delta}$, for $\underline{\delta} \in \mathcal{N}ull(\mathbf{D})$. If $\underline{\alpha}_0$ satisfies $\|\underline{\alpha}\|_0 < \text{Spark}(\mathbf{D})/2$, no vector $\underline{\delta}$ from this null-space exists such that it could be added to $\underline{\alpha}_0$, nulling more entries than the newly introduced ones. Thus, this representation must be the sparsest one possible.

This result is very elementary and yet quite surprising, baring in mind that (P_0) is a highly complicated optimization task of combinatorial flavor. In general, one cannot expect to successfully solve it unless a brute-force search is used. The above uniqueness result, while not constructive, implies that if a sparse enough solution is found via some approximation method, it can be guaranteed to be the desired global optimizer. In general, when solving such complicated optimization problems, even if a solution is proposed, one can at best guarantee that locally it is optimal, by searching a feasible descent direction and finding none. Here we are able to guarantee globally that this is the best solution, and hence the surprise.

Clearly, we can show that the above result is tight by definition. Suppose that we have a null–space vector $\underline{\delta}$ such that $\mathbf{D}\underline{\delta} = \mathbf{0}$ with $\|\underline{\delta}\|_0 = \text{Spark}(\mathbf{D})$ – this vector is realizing the Spark, and thus it existence is guaranteed. Then, taking its first $\text{Spark}(\mathbf{D})/2$ non-zero entries and zeroing the rest, we get a new vector $\tilde{\underline{\alpha}}$ with $\|\tilde{\underline{\alpha}}\|_0 = \text{Spark}(\mathbf{D})/2$ being the representation of a necessarily non-zero signal $\mathbf{D}\tilde{\underline{\alpha}} = \tilde{\underline{S}}$. For this specific signal we have yet another representation of the same cardinality, because $\tilde{\underline{\alpha}} - \underline{\delta}$ has $\text{Spark}(\mathbf{D})/2$ non-zeros too, representing the same signal. Similarly, when choosing more than half of the non-zeros in $\underline{\delta}$ to $\tilde{\underline{\alpha}}$, the remaining entries will form an alternative sparser representation. This way we have constructed a special signal for which uniqueness cannot be guaranteed.

1.3 Behavior Beyond the Bound

Does the above example implies that beyond the $\text{Spark}(\mathbf{D})/2$ bound we are destined to non-uniqueness? The answer is yes, if we think in terms of worst-case. Does the above example imply that a randomly chosen representation with cardinality beyond the bound is necessarily not unique? Definitely not! In fact, given an arbitrary sparse representation with number of non-zeros beyond this bound but still relatively low, chances are that this is the sparesest possible representation for the signal it forms. Said differently, examples showing the tightness of the uniqueness theorem, as constructed above, are very few and rare, and when a probabilistic point of view is adopted, their relative weight is expected to diminish entirely.

Thus, the question we pose here is more general and addresses the uniqueness property of candidate solutions to (P_0), hoping to enable some guarantees even beyond the worst-case bound mentioned above.

1.4 This Work and Prior Art

In this paper we study the behavior of overcomplete representations beyond the known uniqueness bound. When adopting a probabilistic point-of-view, we show both empirically and theoretically that uniqueness can be guaranteed with high confidence with $\text{Spark}(\mathbf{D})/2$ and more non-zero entries. We show that the above–mentioned counter examples to uniqueness are of zero measure for

representations satisfying $\|\underline{\alpha}\|_0 < \text{Spark}(\mathbf{D})$. Furthermore, we show that even beyond this point, uniqueness can be still claimed with reasonable probability.

In order to build these results, we propose to characterize the dictionary in a way that extends the Spark, forming the *Signature* of the matrix. Whereas the Spark "thinks" worst-case, the signature gets the more general picture by gathering all the subsets of columns from $\mathbf{D}$ that are linearly dependent. This signature is used to analyze the uniqueness and compose the results presented.

At the heart of the analysis proposed here is a specific Probability Density Function (PDF) assumed on the signal space. However, instead of specifying this PDF directly with respect to the signals, it is driven via the representations. A commonly used regularization in inverse problem forces sparsity of the representation of the unknown signal, and assumes independence in its coefficients [1]. Such regularization is essentially the manifestation of a PDF on the unknown signal. Thus, generating representations following these rules, signals emerge with a PDF being a mixture of Gaussians. Thus, motivated by how sparsity is used in inverse problems, we propose a simple signal source model. The bottom line to this work is the claim that when a sparse representation is given from this source, it is most likely to be the sparsest one possible. The bound for measuring how sparse is sparse enough for this claim to be true is less restrictive then previously believed.

One word of caution is necessary here: we use here a probabilistic model to describe how signals and their originating representations are constructed. For those signals once generated, the rules of uniqueness apply as indicated here. However, when a candidate representation describing a signal $\underline{S}$ is given from a different source, we cannot apply the given analysis. The reason is that the proposed representation may be drawn from a different distribution, with more emphasis on the "dark-zone" of non-uniqueness. This, for example, explains why we cannot use the uniqueness results we obtain and impose them on the output of the pursuit algorithms as a simple test of success. Pursuit algorithms may (and will) tend to generate non-unique representations, which explains why a separate analysis for them is required. Still, for such analysis to take place, we must start with representations known to be unique, at least in probability, in order to carry out the study. Such analysis will benefit from the results given here.

Indeed, in a very recent pioneering work by Candes, Romberg and Tao (see [21]), the average performance of the Basis Pursuit has been studied, using the same signal source model as described above. A vital part of their analysis is the uniqueness claim: When trying to show an equivalence between the pursuit result and the ideal solution, one must also guarantee that the ideal result is indeed the sparsest one possible. In their work, the authors considered a special dictionary structure built of two unitary matrices, and focused on asymptotic results. Here we discuss the uniqueness for general dictionaries of arbitrary finite sizes, and take a completely different route.

1.5 This Paper

In the next section we start with a simple experiment that explains what is the goal of this work. We show that the empirical probability for obtaining uniqueness is far better than theoretically suggested so far. In Section 3 we propose an analysis to explain this behavior. Section 4 summarizes some of our thoughts about the role of the pursuit algorithms in seeking approximate solutions to (P_0), and our expectations regarding their average performance, compared to the existing worst-case analysis. Section 5 summarizes and concludes this paper.

2 Empirical Evidence

Before proving new properties on sparse representations, let us start by simple yet illustrative experiments that will demonstrate the results we expect to theoretically document later on. We start with the construction of the dictionary: Assume that $\mathbf{D} \in \mathcal{R}^{8\times 20}$ is built by a concatenation of random white and zero mean multivariate Gaussian vectors as its columns. We obtain a full rank matrix and its Spark is 9, i.e. no 8 columns in this matrix can be found to be linearly dependent. This is a general property for such random matrices, stemming from the fact that square random matrices are non-singular with probability 1 (see the seminal work by Edelman and later by Shen [22, 23] on the probabilistic behavior of the extreme singular values of such matrices). Based on the known uniqueness result, every representation with less than $\text{Spark}(\mathbf{D})/2 = 4.5$ non-zeros must be unique. Thus, we are interested in studying the representations with $\|\underline{\alpha}\|_0 > 4$. Clearly, there is no point in considering representations with $\|\underline{\alpha}\|_0 > 8$, since those cannot be unique by definition.

Studying representations with $\|\underline{\alpha}\|_0 = 8$ is also expected to give non-uniqueness, although of a weaker form. Any such representation could be at least replaced by equally good representations coming from all $\binom{L}{8}$ column combinations from $\mathbf{D}$. Nevertheless, it is interesting to see whether better representations (with cardinality strictly smaller than 8) could be found in such cases. Thus, the range $\|\underline{\alpha}\|_0 \in [5, 8]$ is to be experimented on. In our experiment we cover the interval $[1, 8]$ disregarding the lower part being theoretically guaranteed by a known result. We can generate many such representations by first choosing the k non-zero locations at random with uniform probability, and then assigning values to these k locations independently, using some scalar distribution rule. In our experiment we assume that these values are drawn from a zero mean, unit variance, and independent Gaussian distribution.

In the above process we actually induce a probability density function on the representations of cardinality k, and through it a distribution on the signal family that has representations of that cardinality. This is a key feature that will be repeated in our theoretical analysis – rather than starting from the signal PDF, we choose to embark from the representation vectors. As sparsity of representations has been shown to be a powerful signal prior used in inverse problems, generating signals this way makes a lot of sense, not just due to the ease of analysis it brings along here.

Given such a random representation, $\underline{\alpha}$, the signal it represents is given by $\mathbf{D}\underline{\alpha} = \underline{s}$. We can now search exhaustively through all other combinations of k or less columns from $\mathbf{D}$ to seek an alternative representation. For the size chosen, $L = 20$, we have at the most $\sum_{k=1}^{8}\binom{20}{k} \approx 264,000$ tests per each representation, and such a sweep of tests is still doable (though time demanding). This explains the sizes chosen for $\mathbf{D}$ in this experiment. Per each such candidate set of columns, the Least-Squares (LS) solution with the sub-matrix containing the chosen columns dictates the most suitable coefficients, and if the LS error is below the arithmetic accuracy threshold, a candidate representation is assumed found.

We have conducted this experiment as described, and Figure 1 documents its results. Per every cardinality in the range $[1, 8]$ we performed 100 experiments and we present the relative number of experiments that ended with a perfect success (no sparser solution is found) and also the relative number of experiments that ended with a partial success (representations with the same cardinality could also be considered as acceptable). We see that uniqueness can be empirically guaranteed for all representations with cardinality smaller or equal to 7. For representations with 8 non-zero entries, while there are other equivalently sparse representations, there were no better (sparser) ones found.

A second experiment was performed following the same structure, but with a modified dictionary. After creating the random matrix as before, we replaced the first column with a linear combination

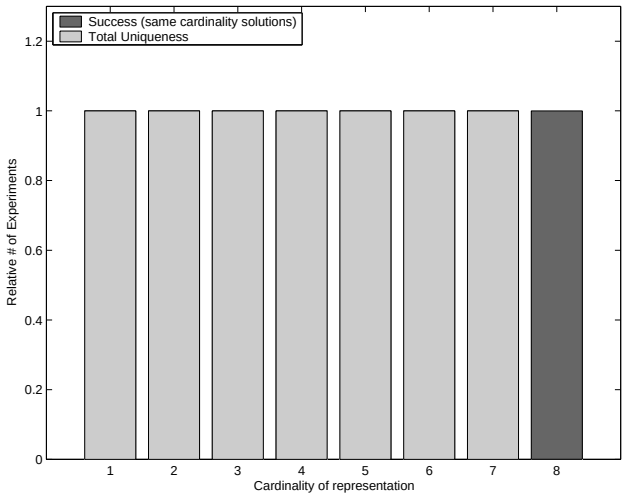

Figure 1: Results of the first experiment: A random dictionary of size 8×20 with Spark=9. Dark gray: Relative number of examples giving equally sparse alternatives, Light gray: relative number of examples giving no alternatives with equal or lower cardinality.

of the last five. This way we have changed the dictionary's Spark to 6 (or below, if we are really unlucky – again, probabilistic results on random matrices suggest that in this case finding a group of less than 6 columns being linearly dependent is of probability zero). Figure 2 presents the results obtained this time. We see several interesting effects:

1. The existing uniqueness bound suggests that uniqueness can be guaranteed for 2 non-zero entries and below. We see that up to $5 = \text{Spark}(\mathbf{D}) - 1$ entries, we empirically get that the representations are unique.

2. Even beyond this cardinality we can still get uniqueness with high probability - a phenomenon we will explain later on. For representations with 6 and 7 non-zero entries, even in cases of violated uniqueness, this violation is due to other representations of the same cardinality and not ones with a strictly sparser form.

3. As in the previous experiment, for representations with 8 non-zero entries we can find equivalently sparse representations but not better ones, so this is a weaker uniqueness success.

The obvious question raised by the above two experiments is whether we can theoretically explain such results, and this way discuss average uniqueness performance, rather than letting extreme cases dictate the analysis of uniqueness. The next section provides this theoretic explanation, and as we shall see, it is almost straight forward.

3 Theoretical Study of Uniqueness Beyond the Bound

3.1 Some Ugly Preliminaries

Since our analysis is of a probabilistic flavor, it is clear that it has to be built on a specific random distribution of the signals, and the results will be different for different signals' source models. As we have already indicated in the previous section, instead of specifying this signal's PDF directly, it

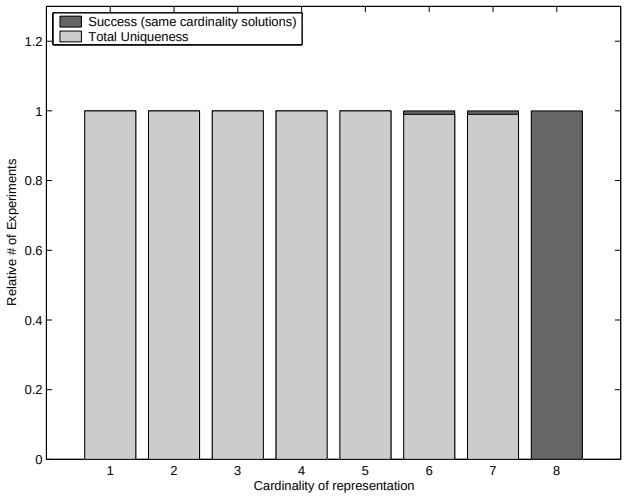

Figure 2: Results of the second experiment: A random dictionary of size 8×20 with Spark=6. Dark gray: Relative number of examples giving equally sparse alternatives, Light gray: relative number of examples giving no alternatives with equal or lower cardinality.

will be implied indirectly by the representation's PDF we effectively use. We assume the following on the PDF of the representation vectors:

1. The probability for each cardinality is fixed and known:

$$p_k = \text{Prob}\left\{\|\underline{\alpha}\|_0 = K\right\} \quad \text{for} \quad K \in 1, 2, \ \ldots \ , L. \tag{4}$$

   For example, p_k could be inversely proportional to K in some way, to indicate that signals tend to have sparse representations, or it could be be uniform in the range $[1, N]$ and zero elsewhere.

2. The non-zeros in the representation vector are uniformly spread with no preference to one zone over another:

$$\begin{aligned} \forall\ j_1 \neq j_2, \quad \text{Prob}\left\{\alpha(j_1) \neq 0 \ / \ \|\underline{\alpha}\|_0 = K\right\} \quad &= \quad \text{Prob}\left\{\alpha(j_2) \neq 0 \ / \ \|\underline{\alpha}\|_0 = K\right\} \\ \text{for} \quad K, j_1, j_2 = 1, 2, \ \ldots \ , L. & \end{aligned} \tag{5}$$

3. The locations of the different non-zero entries in a given representation are statistically independent:

$$\begin{aligned} \forall\ j_1 \neq j_2, \qquad \text{Prob}\left\{\alpha(j_1), \alpha(j_2) \neq 0 \ / \ \|\underline{\alpha}\|_0 = K\right\} \quad &= \\ = \text{Prob}\left\{\alpha(j_1) \neq 0 \ / \ \|\underline{\alpha}\|_0 = K\right\} \cdot \text{Prob}\left\{\alpha(j_2) \neq 0 \ / \ \|\underline{\alpha}\|_0 = K\right\} & \\ \text{for} \quad K, j_1, j_2 = 1, 2, \ \ldots \ , L. & \end{aligned} \tag{6}$$

   Thus, the K locations are chosen at random with equal and independent probability.

4. The K non-zero entries in an representation are randomly generated from a Gaussian distribution with zero mean and unit variance:

$$\begin{aligned} \text{Prob}\left\{\alpha(j) \ / \ \|\underline{\alpha}\|_0 = K \ , \ \alpha(j) \neq 0\right\} \quad &= \quad \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{\alpha(j)^2}{2}\right\} \\ \text{for} \quad K, j = 1, 2, \ \ldots \ , L. & \end{aligned} \tag{7}$$

There is no special reason for the Gaussianity, and any reasonable non-degenerate alternative distribution can be assumed in replacement.

5. The non-zero entries in a representation are mutually independent:

$$\begin{aligned} \forall\ j_1 \neq j_2,\ \ \text{Prob}\left\{\alpha(j_1), \alpha(j_2)\ /\ \|\underline{\alpha}\|_0 = K\ ,\ \alpha(j_1) \neq 0,\ \alpha(j_2) \neq 0\right\} \ &= \\ = \text{Prob}\left\{\alpha(j_1)\ /\ \|\underline{\alpha}\|_0 = K\ ,\ \alpha(j_1) \neq 0\right\} & \\ \cdot\text{Prob}\left\{\alpha(j_2)\ /\ \|\underline{\alpha}\|_0 = K\ ,\ \alpha(j_2) \neq 0\right\} & \end{aligned} \tag{8}$$

$$\text{for} \quad K, j_1, j_2 = 1, 2,\ \ \ldots\ , L.$$

We refer hereafter to representations coming from the above distribution as the output of the machine $\mathcal{M}$. Those are generated first by randomly choosing the cardinality based on p_k, then choosing the involved columns, and finally choosing the non-zero coefficients' values.

When considering representations of cardinality K, there are K specific columns from $\mathbf{D}$ multiplied by a random vector of length K being a white multivariate normalized Gaussian. Thus, referring to the dictionary as deterministic, the PDF $\text{Prob}\left\{\underline{s}\ /\ \underline{\alpha}\right\}$ is of a multivariate Gaussian distribution of dimensionality dictated by both the rank of the sub-matrix of chosen columns from $\mathbf{D}$, and $\|\underline{\alpha}\|_0$. For example, for $\|\underline{\alpha}\|_0 = \text{Spark}(\mathbf{D}) - 1 = K$, the rank of the sub-dictionary is necessarily full (otherwise we contradict the definition of the Spark), and thus we get an ellipsoid in the K^{th} dimensions describing the spread of the signals related to such representations. For more than Spark columns, the rank of the sub-matrix used could be smaller, and then the dimensionality of the signal space is degenerate.

Due to the above, the distribution of the signals $\text{Prob}\left\{\underline{s}\ /\ \|\underline{\alpha}\| = K\right\}$ is expected to be a mixture of $\binom{L}{K}$ equally probable Gaussians of the above form, each referring to a different choice of K columns from the dictionary. Similarly, the signal source model $\text{Prob}\left\{\underline{s}\right\}$ is also a mixture of Gaussians, this time with mixtures of different cardinalities and different weights p_k.

In the coming sub-sections we shall study the uniqueness behavior for different cardinalities. We start with the easier case where $\text{Spark}(\mathbf{D}) = N + 1$ and then turn to discuss the more general case of $\text{Spark}(\mathbf{D}) \leq N$. Throughout our analysis we assume that L, the number of columns in $\mathbf{D}$ is finite.

3.2 Part 1: Spark(D) = N + 1

We start our analysis by treating the special case where the Spark of the dictionary is at its peak, being $\text{Spark}(\mathbf{D}) = N + 1$. In this case we can guarantee uniqueness for representations satisfying $\|\underline{\alpha}\|_0 < (N+1)/2$. This case could refer to dictionaries generated as Grassmanian frames [24, 25], random dictionaries as discussed before, and possibly other constructions. This is the most optimistic scenario, paralleling the first experiment form Section 2. In the coming analysis we consider the following ranges of interest: $\|\underline{\alpha}\|_0 > N$, $\|\underline{\alpha}\|_0 = N$, and $N/2 < \|\underline{\alpha}\|_0 < N$.

3.2.1 Top Interval: $\|\underline{\alpha}\|_0 > N$

Assume a representation $\underline{\alpha}$ is drawn from the above-described random source $\mathcal{M}$, with $\|\underline{\alpha}\|_0 > N$. Clearly, we cannot claim it is the sparsest one describing the signal $\underline{s} = \mathbf{D}\underline{\alpha}$. Since $\text{Spark}(\mathbf{D}) = N + 1$, every subset of N columns from $\mathbf{D}$ is linearly independent. This implies that $\underline{s}$ can be represented by alternative representations with each of those N column-combinations, having only N non-zeros only. Thus we have the following result:

Theorem 1 (Spark($\mathbf{D}$) $= N+1$, $\|\underline{\alpha}\|_0 > N$): *Assume a dictionary $\mathbf{D} \in \mathcal{R}^{N\times L}$ is fixed with Spark($\mathbf{D}$) $= N+1$, and a representation $\underline{\alpha}$ generated from $\mathcal{M}$ with cardinality $\|\underline{\alpha}\|_0 > N$. Then, this representation is necessarily non-unique, implying that for the signal $\underline{s} = \mathbf{D}\underline{\alpha}$ an alternative representation can be found with cardinality N at the most.*

3.2.2 Middle Interval: $\|\underline{\alpha}\|_0 = N$

We now treat the case where the candidate representation drawn from $\mathcal{M}$ is of cardinality N. As before, all the sub-groups of N columns in $\mathbf{D}$ must be linearly independent and thus, whatever signal we get, it can be generated alternatively by all $\binom{L}{N} - 1$ other combinations of N columns from the dictionary, leading to alternative representations with the same cardinality N.

Could we do better? Can we find a group of $N-1$ columns realizing the signal $\underline{s} = \mathbf{D}\underline{\alpha}$? The answer to this question is the essence of this paper, and its rational will be used repeatedly in later cases as well. We shall therefore try to motivate our reply from both algebraic and geometric considerations.

The signal in mind is originally generated as a linear combination of N linear independent columns ($\|\underline{\alpha}\|_0 = N$). Let us fix those columns and denote them as the sub-matrix $\tilde{\mathbf{D}}_N \in \mathcal{R}^{N\times N}$. The N non-zero coefficients in $\underline{\alpha}$ are of random white normalized Gaussian values, implying that as far as those coefficients are involved, the spread of the representation vectors is spherical in this N-dimensional space.

Multiplying the "cloud" of possible representations sharing the same support by the matrix $\tilde{\mathbf{D}}_N$, we get that the signal is also Gaussian random vector with zero mean and a full rank autocorrelation matrix being $\tilde{\mathbf{D}}_N\tilde{\mathbf{D}}_N^H$. This signal occupies the N-dimensional space with non-degenerate ellipsoidal density. By non-degenerate we mean that the volume of this ellipsoid is non-zero, and this is an immediate consequence of the positive definiteness of the autocorrelation matrix we have formed.

Due to the value of the Spark, every sub-group of $N-1$ (or smaller) columns from $\mathbf{D}$ is linearly independent, and as such, spans a subspace of dimension $N-1$ (and below, respectively). Multiplied by normalized Gaussian random vectors representing the non-zero part in candidate representations, random signals are generated in an $(N-1)$-dimensional space. Since there are finite number of such subspaces to consider, being all the combinations of $1, 2, 3, \ \ldots\ , N-1$ columns from $\mathbf{D}$, their addition cannot cover the entire N-dimensional space. Actually, this amalgam of subspaces has a zero volume in the N dimensional space. Thus, chances that the signal we started with will be covered by one of those subspaces is zero. This leads to the conclusion that a representation sparser than N for the discussed case cannot be found. We conclude with the following result:

Theorem 2 (Spark($\mathbf{D}$) $= N+1$, $\|\underline{\alpha}\|_0 = N$): *Assume a dictionary $\mathbf{D} \in \mathcal{R}^{N\times L}$ is fixed with Spark($\mathbf{D}$) $= N+1$, and a representation $\underline{\alpha}$ generated from $\mathcal{M}$ with cardinality $\|\underline{\alpha}\|_0 = N$. Then, considering the signal $\underline{s} = \mathbf{D}\underline{\alpha}$,*

1. *there are $\binom{L}{N} - 1$ alternative representations for $\underline{s}$ with the same cardinality N, and thus the probability to find such alternative is 1.*
2. *the probability to find an alternative representation for $\underline{s}$ with cardinality smaller than N is zero.*

This is a weak form of uniqueness, but nevertheless one of interest, saying that we could find similarly sparse alternatives and not better ones. Let us give a very simple and intuitive example to better explain our result. Suppose that the dictionary $\mathbf{D}$ is of size 2×5, implying that our signals

are points in 2D. We further assume that the Spark of $\mathbf{D}$ is 3, meaning that every 2×2 sub-matrix from $\mathbf{D}$ is full rank. Suppose that a specific signal is constructed by linear combination of the 2 first columns. Since the 2 coefficients used by the linear combination are random normalized Gaussian ones, the signals we can possibly generate are also Gaussian and occupying the 2D space, although with a distorted spread from spherical to ellipsoidal distribution. The shape of the 2D ellipsoid is dictated by the two eigenvalues of the 2×2 matrix formed by the first two columns used to build the signal.

In our analysis we concentrate on the 2D space of signals, and we have just found out that every point in the plain is a possible signal (with varying non-zero probability).

Now let us try building the same signal using only one column, in an attempt to find a sparser representation. Consider a specific column, and by randomly choosing the representation coefficient, consider the signals this can generate. We get a set of 2D Gaussian vectors all on a specific line passing through the origin, in a direction dictated by the column used. By considering all $\binom{5}{1} = 5$ columns, we have 5 such lines where the signals with representation cardinality 1 could reside. Any finite number of lines cannot cover more than zero-volume in the 2D plane, and thus chances are that our signal can never find a sparser representation with cardinality 1.

Note that this analysis suggests that if we are to approximate a signal with some inaccuracies, rather than exactly represent it, the same approach could be used. This time, however, every such line should be replaced by a thickened version of it, increasing chances of failure (i.e., finding sparser representation alternatives). We leave such analysis for future work.

3.2.3 Bottom Interval: $N/2 < \|\underline{\alpha}\|_0 < N$

The analysis required for $N/2 < \|\underline{\alpha}\|_0 < N$ case is quite similar to the one we presented earlier, with one major difference - whereas the previous case led to weak version of uniqueness, here will get more conclusive, strong uniqueness.

Suppose that a representation drawn from $\mathcal{M}$ is of cardinality $\|\underline{\alpha}\|_0 = K$, in the range $(N/2, N)$. This implies that the signal in mind is originally generated as a linear combination of K linear independent columns. As before, fixing those chosen columns and denoting them as the sub-matrix $\tilde{\mathbf{D}}_K \in \mathcal{R}^{N \times K}$, due to the normalized Gaussianity of the K non-zero coefficients in $\underline{\alpha}$, the signal obtained is also a Gaussian random vector with zero mean and a rank K autocorrelation matrix being $\tilde{\mathbf{D}}_K \tilde{\mathbf{D}}_K^H$. These signals reside in the N-dimensional space, but fill only $K < N$ dimensions of it.

Due to the value of the Spark, every sub-group of $K-1$ (or smaller) columns from $\mathbf{D}$ is linearly independent, and as such, spans a subspace of dimension $K-1$ (and below, respectively). As before, those finite number of subgroups of columns define signal subspaces of dimensionality $K-1$ and below, and their addition has zero volume in the overlap with the subspace the original signals can be found in.

What about similarly sparse alternative representations? There are $\binom{L}{K}$ combinations of K columns from $\mathbf{D}$ that could built a competing representation. Choosing one such candidate group, it creates a "cloud" of signals of the same dimensionality K in the N dimensional space. How overlapping are the original and the newly formed subspaces? We shall show that this overlap could either be complete or empty (in measure of volume).

The complete overlap implies that the two different groups of K columns are spanning the same subspace. Thus, a group of $K+1$ linear dependent columns can be built by taking the first K-column group, and adding any of the columns from the second group. Since these $K+1$ columns span a K-dimensional space, they must be linearly dependent, and this contradicts the Spark.

Thus, complete overlap is impossible.

The alternative case where the two subspaces of dimensionality K are different implies that their overlap is of dimensionality $K-1$ at the most (as an example, for a 3D space with two subspaces of dimensionality 2, the complete overlap implies that the two planes passing through the origin are the same, and if they are not so, their intersection is a line). As we have already stated, a subspace of dimensionality $K-1$ has zero volume in the K-dimensional space. Even addition of many such subspaces will not change this fact, if finite number of members participate in this addition. This all leads to the conclusion that even equivalently sparse representations will not be found with probability 1. We thus have the following result:

Theorem 3 (Spark($\mathbf{D}$) $= N+1$, $N/2 < \|\underline{\alpha}\|_0 < N$): *Assume a dictionary $\mathbf{D} \in \mathcal{R}^{N\times L}$ is fixed with Spark($\mathbf{D}$) $= N+1$, and a representation $\underline{\alpha}$ generated from $\mathcal{M}$ with cardinality $\|\underline{\alpha}\|_0$ in the range $(N/2, N)$. Then, considering the signal $\underline{s} = \mathbf{D}\underline{\alpha}$, the probability to find an alternative representation for $\underline{s}$ with cardinality $\|\underline{\alpha}\|_0$ or smaller is zero.*

This is a strong form of uniqueness, but as opposed to the classic result, it leans on probabilistic considerations, meaning that while counter examples to this uniqueness result can be created, their overall weight is negligible in the space of signals we have formed.

3.2.4 Relation to the Empirical Results

In the first experiment in Section 2 we had $N = 8$ and Spark($\mathbf{D}$) $= 9$, matching the case studied here. Due to Theorem 1 it is clear that there is no uniqueness for $\|\underline{\alpha}\|_0 > 8$, and this range was not part of the simulation. Theorem 2 gives us a weak guarantee of uniqueness for $\|\underline{\alpha}\|_0 = 8$, with $\binom{20}{8} - 1$ alternative representations with the same cardinality and no sparser ones. This aligns well with the result documented in Figure 1. Theorem 3 supplies us with the results for $\|\underline{\alpha}\|_0 < 8$, guaranteeing uniqueness, as indeed empirically obtained. Figure 3 presents a graph parallel to 1, as we expect to obtain for general N (assumed for convenience to be even).

3.3 Part 2: Spark($\mathbf{D}$) $\leq N$

We now turn to discuss the more common case where Spark($\mathbf{D}$) $\leq N$. This case refers to dictionaries generated from overcomplete wavelets, ridgelets, curvelets, many other types of frames, and amalgams of them [15, 16, 17, 19, 26, 27]. This is the more realistic scenario, paralleling the second experiment form Section 2.

In this case we can guarantee uniqueness for representations satisfying $\|\underline{\alpha}\|_0 < \text{Spark}(\mathbf{D})/2$. This time the ranges of interest to consider are: $\|\underline{\alpha}\|_0 > N$, $\text{Spark}(\mathbf{D}) \leq \|\underline{\alpha}\|_0 \leq N$, and $\text{Spark}(\mathbf{D})/2 \leq \|\underline{\alpha}\|_0 \leq \text{Spark}(\mathbf{D}) - 1$. As we shall see next, the analysis here is similar but more involved. The range $\text{Spark}(\mathbf{D}) \leq \|\underline{\alpha}\|_0 \leq N$ in particular is problematic and requires a definition of the *Signature* of a dictionary in order to get an evaluation of the uniqueness probability.

Definition 1 (Signature): *For a matrix $\mathbf{D} \in \mathcal{R}^{N\times L}$, its signature is defined as the discrete function $Sig_{\mathbf{D}}(k)$, for $k = 1, 2, 3, \ldots, L$, counting the relative number[1] of k-column-combinations in $\mathbf{D}$ that are linearly dependent.*

Here are some properties of the signature:

[1] By relative we mean that the values $Sig_{\mathbf{D}}(k)$ are in the range $[0, 1]$, due to division by $\binom{L}{k}$, the number of all combinations of k columns from the L existing ones.

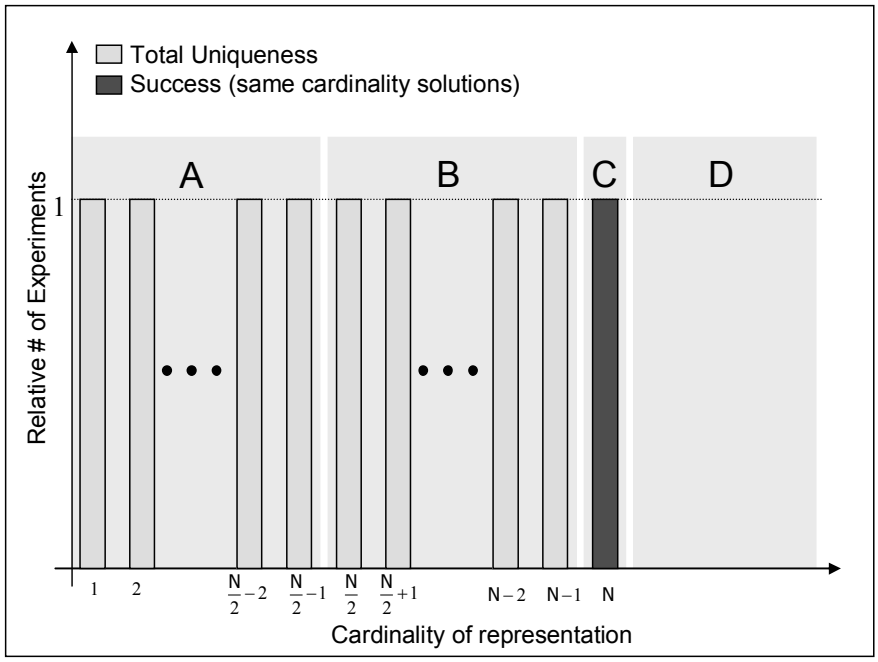

Figure 3: A schematic graph of the probability of uniqueness for Spark($\mathbf{D}$) $= N + 1$. The different zones are A: refers to the classic known uniqueness result; B: Theorem 3; C: Theorem 2; and D: Theorem 1.

- Due to the definition of the Spark, $Sig_{\mathbf{D}}(k)$ is zero for all $k <$ Spark($\mathbf{D}$).
- For $k \geq$ Spark($\mathbf{D}$) there are $\binom{L}{k}$ possible combinations and at least one is linearly dependent, thus leading to strictly positive values, $Sig_{\mathbf{D}}(k) > 0$.
- For $k > N$ we necessarily have $Sig_{\mathbf{D}}(k) = 1$ since all groups of k such columns are linearly dependent.
- We conjecture that the signature is monotonic non-decreasing. This property should be proven, but we leave this as an open problem at the moment, as we do not need to use it in the analysis that follows.
- For the case Spark($\mathbf{D}$) $= N+1$, the signature is necessarily a simple step function, being zero for $k \leq N$, and 1 for $k > N$. This will explain the ease with which the previous analysis was carried out, and the reason for separating the study of this case.
- The signature is NP-hard to compute, just as the Spark. Still, bounds on it can be derived. One such interesting bound based on known Spark is described in the Appendix, based on result due to Björner, related to analysis of Matroids [28].

While for the worst-case analysis the exact value of $Sig_{\mathbf{D}}(k)$ has no consequence, this value becomes of extreme importance in evaluating probabilities of uniqueness under our probabilistic regime of signals. For illustration, the signatures of both dictionaries used in Section 2 are given in Figure 4.

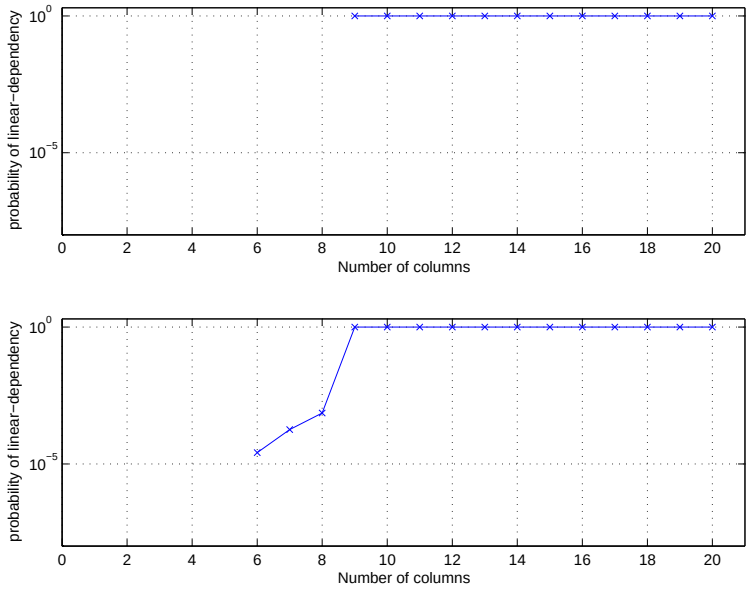

Figure 4: The signatures of the two dictionaries used in Section 2: Top - first experiment, Bottom - second experiment.

Evaluating the signature of the matrix is harder than finding its Spark, and this in turn is NP-hard. Yet, there are ways to bound the Spark from below and above, and those methods could be used to develop bounds on $Sig_{\mathbf{D}}(k)$. We will not go into this matter here.

3.3.1 Top Interval: $\|\underline{\alpha}\|_0 > N$

Assume a representation $\underline{\alpha}$ is drawn from the above-described random source $\mathcal{M}$, with $\|\underline{\alpha}\|_0 > N$. Just as in Section 3.2.1, we cannot claim it is the sparsest one describing the signal $\underline{s} = \mathbf{D}\underline{\alpha}$. No matter what Spark($\mathbf{D}$) is, a subset of linearly independent N columns from $\mathbf{D}$ can be found, since we assume that $\mathbf{D}$ is full rank. Thus, $\underline{s}$ can be represented by alternative representations with N non-zeros only, leading to the following result:

Theorem 4 (Spark($\mathbf{D}$) $\leq N$, $\|\underline{\alpha}\|_0 > N$): *Assume a dictionary $\mathbf{D} \in \mathcal{R}^{N\times L}$ is fixed with Spark($\mathbf{D}$) $\leq N$, and a representation $\underline{\alpha}$ generated from $\mathcal{M}$ with cardinality $\|\underline{\alpha}\|_0 > N$. Then, this representation is necessarily non-unique, implying that for the signal $\underline{s} = \mathbf{D}\underline{\alpha}$ an alternative representation can be found with cardinality N at the most.*

This case resembles the case of $\|\underline{\alpha}\|_0 > N$ with the maximal Spark, as discussed in Section 3.2.1. There is one major but unimportant difference here - whereas in the maximal Spark case we could have claimed that any group of N columns is linearly independent, here we can just say that one such group exists. Again, this difference has no influence on the outcome, being a complete and certain loss of uniqueness, as expected.

3.3.2 Medium Low Interval: Spark($\mathbf{D}$) $\leq \|\underline{\alpha}\|_0 \leq N$

Given a candidate representation with $\|\underline{\alpha}\|_0 = K$, in the range [Spark($\mathbf{D}$), N], if the K columns chosen are linearly dependent, an immediate reduction can lead to an alternative representation

with a smaller cardinality. By replacing one column in this group with a linear combination of the $K-1$ others, uniqueness is lost. Thus, at least a $Sig_{\mathbf{D}}(K)$ portion of the cases lead to loss of uniqueness this way. Note that all the K-column combinations are equally probable due to prior assumptions, and thus the signature value is applicable directly without weighting.

If the K columns pointed to by $\underline{\alpha}$ are linearly independent, alternative sparser representations cannot be found, leaning on the same rational we have exercised earlier. Any other group of $K-1$ columns or smaller could potentially create a competing representation for the signal in mind. However, addition of all the volumes on these subspaces will not cover substantial portion of the K-dimensional signal space, and thus no sparser solutions will be found. Thus, in the search for sparser representations, the probability that the given representation is the sparsest is $1-Sig_{\mathbf{D}}(K)$, with problems encountered only with linear dependent groups.

When addressing the quest for the same cardinality alternatives, we focus on the linearly independent cases since those have no sparser alternatives. For a given such set of K columns used by the original representation, assume that the first $K-1$ of them together with another column form a linearly dependent set. This implies that $K-1$ alternative representations with the same cardinality are possible by replacing each of the $K-1$ first columns with the external one, and those combinations lead to a weak uniqueness result.

Similarly, if the first $K-2$ columns in the original group can be merged with a different column to give a linear dependency, we get $K-2$ alternative representations of the same cardinality. This could continue with small groups with $K-3,\ K-4,\ \ldots\ ,\text{Spark}(\mathbf{D})-1$ alternatives. Going below this set leads to no other alternatives.

Let us look closely into the first case generating the competing solutions, and count the number of combinations that may lead to this problem. We consider $Sig_{\mathbf{D}}(K)\cdot\binom{L}{K}$ groups of K linearly-dependent columns. Choosing one such group and replacing one of its columns by a different column from the remaining $L-K$ ones, we get $K(L-K)$ such replacements, all leading to the weak uniqueness result.

Similarly, taking the $Sig_{\mathbf{D}}(K-1)\cdot\binom{L}{K-1}$ groups of $K-1$ linearly dependent columns, we can propose per each $(K-1)(L-K+1)$ replacements, and per each of those add an arbitrary column among the remaining $L-K+1$ ones, getting a total of $(K-1)(L-K+1)\binom{L-K+1}{1}Sig_{\mathbf{D}}(K-1)\cdot\binom{L}{K-1}$ combinations of K columns where a competing equally sparse alternatives can be built. This could continue with smaller dependent groups, giving that there are no more than

$$\mathcal{N}=\sum_{j=0}^{K-\text{Spark}(\mathbf{D})}(K-j)(L-K+j)Sig_{\mathbf{D}}(K-j)\cdot\binom{L}{K-j}\cdot\binom{L-K+j}{j}$$

possible combinations of K columns that lead to the existence of alternative representation. The last term in the above expression takes all the groups of j columns from the remaining $L-k+j$ ones in order to finally get K columns. Clearly, in gathering all those, we should count only the linearly independent ones, and discard of repetitions. Thus, the above number is an upper bound on the K-column combinations that lead to the weak uniqueness. Divided by $\binom{L}{K}$ we get a bound on the probability for weak uniqueness. We summarize by the above result:

Theorem 5 ($\mathbf{Spark}(\mathbf{D})\leq N,\ \mathbf{Spark}(\mathbf{D})\leq\|\underline{\alpha}\|_0\leq N$): *Assume a dictionary $\mathbf{D}\in\mathcal{R}^{N\times L}$ is fixed with Spark$(\mathbf{D})\leq N$, and a representation $\underline{\alpha}$ generated from $\mathcal{M}$ with cardinality Spark$(\mathbf{D})\leq\|\underline{\alpha}\|_0=K\leq N$. Then, considering the signal $\underline{s}=\mathbf{D}\underline{\alpha}$,*

1. *The probability that the given representation is the sparsest of all (disregarding equally sparse alternatives) is $1-Sig_{\mathbf{D}}(K)$.*

2. The probability to find an alternative representation of the same cardinality is

$$\sum_{j=0}^{K-Spark(\mathbf{D})} (K-j)(L-K+j)\binom{K}{j} Sig_{\mathbf{D}}(K-j)$$

or lower.

3.3.3 Bottom Interval: $\mathbf{Spark}(\mathbf{D})/2 \leq \|\underline{\alpha}\|_0 < \mathbf{Spark}(\mathbf{D})$

Suppose that a representation drawn from $\mathcal{M}$ is of cardinality $\|\underline{\alpha}\|_0 = K$, in the above range. This implies that the signal in mind is originally generated as a linear combination of K linear independent columns. Same reasoning leads to a K-dimensional Gaussian cloud of signals in the N-dimensional signal space. Every sub-group of $K-1$ (or smaller) columns from $\mathbf{D}$ is linearly independent as well, and as such, all those together span subspaces of dimension $K-1$ (and below, respectively), thus leading to the conclusion that with probability 1 no sparser representation can be found.

As to similarly sparse alternatives, following the same analysis as in Section 3.2.3 give that no such alternatives can be found. Thus we have the following strong uniqueness result:

Theorem 6 ($\mathbf{Spark}(\mathbf{D}) < N+1$, $\mathbf{Spark}(\mathbf{D})/2 \leq \|\underline{\alpha}\|_0 < \mathbf{Spark}(\mathbf{D})$): *Assume a dictionary $\mathbf{D} \in \mathcal{R}^{N\times L}$ is fixed with $Spark(\mathbf{D}) < N+1$, and a representation $\underline{\alpha}$ generated from $\mathcal{M}$ with cardinality $\|\underline{\alpha}\|_0$ in the range $(Spark(\mathbf{D})/2, Spark(\mathbf{D}))$. Then, considering the signal $\underline{s} = \mathbf{D}\underline{\alpha}$, the probability to find an alternative representation for $\underline{s}$ with cardinality $\|\underline{\alpha}\|_0$ or smaller is zero.*

3.3.4 Relation to the Empirical Results

In the second experiment in Section 2 we had $N = 8$ and $\text{Spark}(\mathbf{D}) = 6$, matching the case studied here. Due to Theorem 4 it is clear that there is no uniqueness for $\|\underline{\alpha}\|_0 > 8$.

Theorem 5 supplies us with the results for $6 \leq \|\underline{\alpha}\|_0 \leq 8$, suggesting that the probability to get a strong uniqueness is $1 - Sig_{\mathbf{D}}(\|\underline{\alpha}\|_0)$. Since this number is very close to 1 (e.g., for $\|\underline{\alpha}\|_0 = 8$ it is $1 - 7.2e-4$), the 100 experiments found no such cases, as can be seen in Figure 1.

As to equally sparse alternatives, the probability of finding those for $K = 6$ this probability is $6 \cdot 14/\binom{20}{6} = 0.0022$, and for $K = 7$ it is 0.0316 - in both cases quite low but possible to encounter, as indeed displayed in the found results.

Figure 5 presents a graph parallel to 2, as we expect to obtain for general N, assuming that $\text{Spark}(\mathbf{D})$ is even.

4 Relation to Average Performance of Pursuit Algorithms

Given a signal $\underline{S} \in \mathcal{R}^N$ known to have a sparse representation over the dictionary $\mathbf{D}$, we are interested in finding its representation faithfully, and with a reasonable amount of computations. We assume that the signal is drawn from the presented source model, by generating first a representation $\hat{\underline{\alpha}}$ at random from $\mathcal{M}$ and computing $\underline{S} = \mathbf{D}\hat{\underline{\alpha}}$. This way we have characterized in full how signals are distributed.

Applying pursuit algorithms on $\underline{S}$, could we guarantee successful recovery of $\hat{\underline{\alpha}}$? Clearly, if $\hat{\underline{\alpha}}$ is not the unique (sparsest) representation of $\underline{S}$, there is no point to this question, since we do not

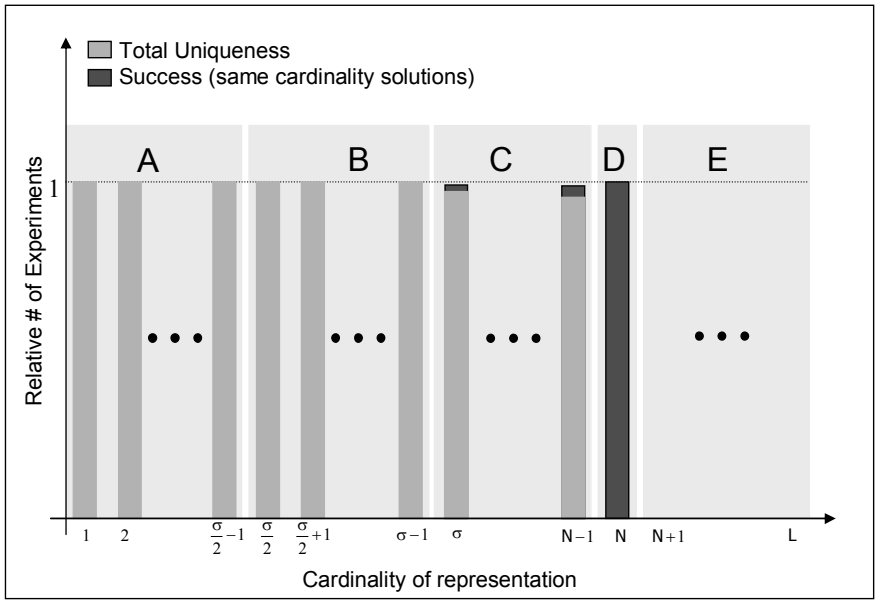

Figure 5: A schematic graph of the probability of uniqueness for Spark($\mathbf{D}$) $< N + 1$. The different zones are A: the classic known uniqueness result; B: Theorem 6; C: Theorem 5; D: Theorem 5; and E: Theorem 4.

want to recover $\hat{\underline{\alpha}}$ in those cases. So, our question focuses on the cases where uniqueness holds true, and ask whether the pursuit algorithms succeed.

Previous work analyzed this question for several variants of the greedy algorithm [6, 7, 8, 9, 10, 11, 12, 13]. Other work studied the Basis Pursuit algorithms [15, 16, 17, 18, 19, 10, 14, 20]. All these work concentrated on the worst-case scenario, just as described above with respect to the uniqueness property, showing that if the signal has a sparse enough representation, the pursuit will succeed. The bound on sparsity is more restrictive compared to the uniqueness one, and its development is far more complicated in general. This bound is built on the definition of the *Mutual Incoherence* ($M(\mathbf{D})$), as the maximum over all absolute off-diagonal entries in the Gram matrix $\mathbf{D}^H\mathbf{D}$. In order to guarantee successful recovery of the representation, it should have smaller than $0.5(1 + M^{-1})$ non-zeros.

Thus, parallel to the results summarized in the previous section, there is great interest in knowing whether the pursuit algorithms are successful beyond this bound in probability. As we have mentioned earlier, the work in [21] is the first to address this question directly, with results for a specific structure of dictionaries, and with an asymptotic formulation. Here we offer some empirical results that will set the stage for a theoretical analysis that will study the behavior of the pursuit algorithms using general dictionaries.

Figures 6 and 7 present the results for two dictionaries of size 30×80. The first, being completely random leads to the maximal Spark, and its mutual incoherence is 0.5575. Thus, success of the pursuit algorithms is guaranteed for representations with one non-zero entry. Similarly, the second graph corresponds to a dictionary of the same size, but with a deliberate reduction of the Spark to 15. This dictionary's mutual incoherence is 0.7075 again implying that only one non-zero representations can be recovered well by the pursuit algorithms. As can be seen from the results,

1. In both cases the success rate is high for $\|\underline{\alpha}\|_0 \leq 5$, and decays gracefully from there.

2. Although the two dictionaries have very different Spark, the results of the pursuit algorithms in both cases are very similar.

3. The two pursuit algorithms perform very similarly with weaker performance of the greedy algorithm for small cardinalities, and slower decay in performance as the cardinality grows.

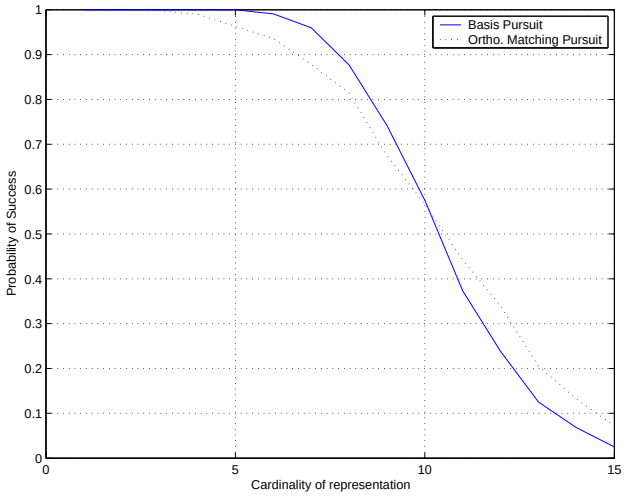

Figure 6: The probability of success (recovering at least equally sparse solution in 1000 random examples) of the Basis Pursuit and the Ortho. Matching Pursuit algorithm. The dictionary is a random matrix of size 30×80, its Spark is 31, and its mutual incoherence is 0.5575.

An analysis of these results from a theoretical point of view is valuable and should be carried out. In particular, it is interesting to ask whether the Spark or the signature have any role in dictating the pursuit results in probability.

5 Conclusions

In this paper we have studied the uniqueness of sparse representations of signals over a given overcomplete dictionary. We saw both empirically and theoretically that such representations are likely to be the sparsest ones for the signal they form if they are sparse enough. Previous work has shown that below half the Spark of the dictionary, the representation is necessarily the sparsest. Here we have extended this result and shown that representations with less than Spark non-zero entries are the sparsest with probability 1, and even beyond this cardinality, uniqueness can be still claimed with high probability.

A very helpful tool in our analysis is the signature of the dictionary. Further work is required in order to find ways to approximate or bound this function. Another promising direction for future research is the analysis of pursuit algorithms using the same probabilistic model drawn here, extending the results in [21] . Simulation results here and in [15] indicate that these algorithms are expected to perform far better than the worst-case bounds suggest. A similar analysis could shed light on this behavior.

Approximate representations rather than exact ones are appealing as well for many applications. A parallel study of the uniqueness of such representations is of great importance as well, extending prior results given in [10].

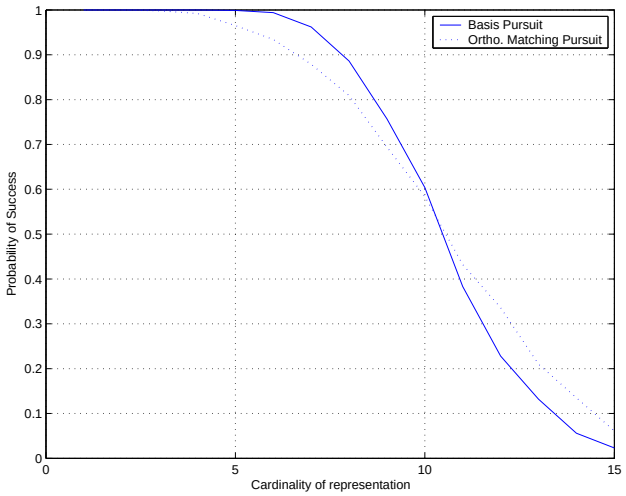

Figure 7: The probability of success (recovering at least equally sparse solution in 1000 random examples) of the Basis Pursuit and the Ortho. Matching Pursuit algorithm. The dictionary is a random matrix of size 30×80, its Spark is 15, and its mutual incoherence is 0.7075.

6 Acknowledgements

The author would like to thank Prof. Alfred M. Bruckstein (CS department – the Technion), Joel Tropp (Math department, University of Michigan), and Justin Romberg (Applied Math - California Inst. of Technology) for long and helpful discussions. Thanks are also in order to Felix Goldberg from the Mathematics department at the Technion for the helpful reference to the work on Matroids and its relevance to the signature.

References

[1] Chen, S.S., Donoho, D.L. & Saunders, M.A. (2001) Atomic decomposition by basis pursuit, *SIAM Review*, Volume 43, number 1, pages 129–59.

[2] Natarajan, B.K. (1995) Sparse approximate solutions to linear systems. *SIAM J. Comput.*, 24:227234.

[3] Davis, G., Mallat, S., & Avellaneda, M. (1997) Greedy adaptive approximation. *J. Constr. Approx.*, 13:5798.

[4] Mallat, S. & Zhang, Z. (1993) Matching Pursuit in a time-frequency dictionary, *IEEE Transactions on Signal Processing*, Volume 41, pages 3397–3415.

[5] Pati, Y.C., Rezaiifar, R. & Krishnaprasad, P.S. (1993) Orthogonal Matching Pursuit: Recursive Function Approximation with Applications to Wavelet Decomposition, Proceedings of the 27 th Annual Asilomar Conference on Signals, Systems, and Computers.

[6] Couvreur, C. & Bresler, Y. (2000) On the optimality of the Backward Greedy Algorithm for the subset selection problem. *SIAM J. Matrix Anal. Appl.*, 21(3):797808.

[7] Gilbert, A.C. Muthukrishnan, S. & Strauss, M.J. (2003) Approximation of functions over redundant dictionaries using coherence. 14th Ann. ACM-SIAM Symposium Discrete Algorithms.

[8] Tropp, J.A., Gilbert, A.C., Muthukrishnan, S., & Strauss M.J. (2003) Improved sparse approximation over quasi-incoherent dictionaries. IEEE International Conference on Image Processing, Barcelona, September.

[9] Tropp, J.A. (2003) Greed is Good: Algorithmic results for sparse approximation. To appear in *IEEE Trans IT*.

[10] Donoho, D.L., Elad, M., & Temlyakov, V. (2004) Stable Recovery of Sparse Overcomplete Representations in the Presence of Noise, submitted to the *IEEE Trans. IT* on February.

[11] Tropp, J.A. (2004) Just relax: Convex programming methods for subset selection and sparse approximation. ICES Report 04-04, UT-Austin, February.

[12] Temlyakov, V.N. (1999) Greedy algorithms and m-term approximation, *J. Approx. Theory* **98** 117-145.

[13] Temlyakov, V.N. (2000) Weak greedy algorithms , *Adv. Comput. Math.* **5** 173-187.

[14] Gorodnitsky I.F. & Rao B.D. (1997) Sparse Signal Reconstruction from Limited Data Using FOCUSS: A Re-Weighted Norm Minimization Algorithm, *IEEE Trans. On Signal Processing*, pp. 600-616, Vol. 45, no. 3, March.

[15] Donoho, D.L. & Huo, X. (2001) Uncertainty principles and ideal atomic decomposition, *IEEE Trans. on Inf. Theory*, volume 47, number 7, pages 2845–62.

[16] Huo, X. (1999) Sparse Image representation Via Combined Transforms, PhD thesis, Stanford.

[17] Elad, M. & Bruckstein, A.M. (2002) A generalized uncertainty principle and sparse representation in pairs of $\Re^N$ bases, *IEEE Trans. on Inf. Theory*, Volume 48, pages 2558–2567.

[18] Donoho, D.L. & Elad, M. (2002) Optimally sparse representation in general (non-orthogonal) dictionaries via ℓ^1 minimization, *Proc. Nat. Aca. Sci.* **100** 2197-2202.

[19] Gribonval, R. & Nielsen, M. (2002) Sparse representations in unions of bases, submitted to the *IEEE Trans. on Inf. Theory*.

[20] Fuchs, J.-J. (2002) On sparse representations in arbitrary redundant bases. IRISA Technical Report, Univ. de Rennes I, Dec.. Submitted to *IEEE Trans. IT*.

[21] Candès E., Romberg J. & Tao T. (2004) Robust uncertanty principles: Exact signal reconstruction from highly incomplete frequency information, *Paper Draft – Personal Commpunication*.

[22] Edelman, A. (1988) Eigenvalues and condition numbers of random matrices, *SIAM Journal on Matrix Analysis and Applications*, 543–560.

[23] Shen, J. (2001), On the singular values of Gaussian random matrices, *Linear Algebra and its Applications*, 326(1-3), 1-14.

[24] Strohmer, T. & Heath Jr, R. (2003) Grassmannian frames with applications to coding and communcations, to appear in *Appl.Comp.Harm.Anal.*.

[25] Tropp, J.A., Dhillon, I.S., Heath, R.W., & Strohmer, T. (2004) Constructing Grassmannian packings via alternating projection, working draft.

[26] Candès, E.J. & Donoho, D.L. (2002) New tight frames of curvelets and the problem of approximating piecewise C^2 images with piecewise C^2 edges. to appear *Comm. Pure Appl. Math.*

[27] Starck, J.-L., Candès, E. & Donoho, D.L. (2002) The curvelet transform for image denoising *IEEE Transactions on Image Processing* 11(6), pp. 131–141.

[28] A. Björner (1980) Some matroid inequalities, *Discrete Math.*, Vol. 31, pp. 101–103.

[29] F. Goldberg (2004) Overcomplete Bases, Spark, and Matroids, *personal communication.*

Appendix - An Upper bound on the Signature

We have given the following definition:

Definition 2 (Signature): *For a matrix* $\mathbf{D} \in \mathcal{R}^{N\times L}$*, its signature is defined as the discrete function* $Sig_{\mathbf{D}}(k)$*, for* $k = 1, 2, 3, \ \dots \ , L$*, counting the relative number of* k*-column-combinations in* $\mathbf{D}$ *that are linearly dependent.*

The signature is NP-hard to compute, just as the Spark (and actually harder). Still, bounds on it can be derived. One such interesting bound that we will show here is based on the assumption that the Spark is known. This result is due to Björner, who analyzed and bounded Matroids properties [28]. This was adapted to the bounding of the signature by Felix Goldberg [29]. We will state the result here without proof or discussion. Further work is required in order to bound the signature better, taking into account known interactions between the dictionary's columns, and more.

Theorem 7 (Upper bound on the Signature)*: For a full rank dictionary* $\mathbf{D} \in \mathcal{R}^{N\times L}$ *(rank* N*) with Spark*$(\mathbf{D}) = \sigma \leq N + 1$*, the signature of the dictionary is upper bounded by*

$$Sig_{\mathbf{D}}(k) \leq 1 - \frac{\sum_{i=0}^{\sigma-1} \left(\begin{array}{c} L-N+i-1 \\ i \end{array} \right) \left(\begin{array}{c} N-i \\ k-i \end{array} \right)}{\left(\begin{array}{c} L \\ k \end{array} \right)}, \qquad k = 1, 2, \ \dots \ , N. \tag{A-1}$$

The rational behind this result is that if Spark$(\mathbf{D}) = \sigma$, it does not necessarily mean that **every** σ combination of columns from $\mathbf{D}$ is linearly dependent. In fact, only a limited number of such sets can be dependent without violating the rank of the matrix. As a simple example, for a full rank dictionary of size 3×4 and Spark$(\mathbf{D}) = 3$, if every triplet is linearly dependent, than the rank must be 2. This can be proven by construction of the column space representation: we drop the first column as it is spanned by the next two. The remaining triplet is also dependent and thus we can drop one more column without affecting the columns space spanned. Thus the rank is 2, violating the initial assumption about $\mathbf{D}$ being full rank. In such case, only one of the 4 possible triplets can be linearly dependent. The above theorem generalizes this idea.