Analysis of the Basis Pursuit Via the Capacity Sets

Joseph Shtok and Michael Elad

The Computer Science Department,
the Technion - Israel Institute of Technology, Haifa 32000 Israel,
email: [shtok,elad]@cs.technion.ac.il.

ABSTRACT. Finding the sparsest solution $\alpha$ for an under-determined linear system of equations $D\alpha = s$ is of interest in many applications. This problem is known to be NP-hard. Recent work studied conditions on the support size of $\alpha$ that allow its recovery using $\ell_1$-minimization, via the Basis Pursuit algorithm. These conditions are often relying on a scalar property of $D$ called the mutual-coherence. In this work we introduce an alternative set of features of an arbitrarily given $D$, called the capacity sets. We show how those could be used to analyze the performance of the basis pursuit, leading to improved bounds and predictions of performance. Both theoretical and numerical methods are presented, all using the capacity values, and shown to lead to improved assessments of the basis pursuit success in finding the sparsest solution of $D\alpha = s$.

1. Introduction

A powerful trend in signal processing that has evolved in recent years is the use of redundant dictionaries, rather than just bases, for a sparse representation of signals (images, sound tracks, and more). In such
a setting, we consider a linear equation $\mathbf{s} = \mathbf{D}\alpha$, where $\mathbf{s}$ is a given signal, $\mathbf{D}$ is the representation dictionary, and $\alpha$ is the signal’s representation. The matrix $\mathbf{D}$ is a general full rank $N \times L$ matrix, where $L > N$, assumed to have $\ell_2$ normalized columns. The number of non-zero elements in the coefficient vector $\alpha$ is measured by the $\ell_0$-norm, $\| \cdot \|_0$, on $\mathbb{R}^L$. The goal is to find, within the $(L - N)$-dimensional affine space of the solutions for this equation, the sparsest representation for $\mathbf{s}$, i.e. one which has the least number of non-zero entries. This goal is formalized by the following optimization problem:

\[
(P_0) : \text{Arg min}_{\alpha \in \mathbb{R}^L} \| \alpha \|_0 \text{ s.t. } \mathbf{D}\alpha = \mathbf{s}.
\]

In this paper, we consider the signals for which the solution of $(P_0)$ is unique, and we define $\mathcal{S}(\mathbf{D})$ as the family of such signals. We denote $\Omega = \{1, \ldots, L\}$, and refer to the support of the vector $\alpha = (\alpha_1, \ldots, \alpha_L)^T$ as the set $\Gamma = \text{supp}(\alpha) = \{n \in \Omega \mid \alpha_n \neq 0\}$.

The problem $(P_0)$ is NP-hard, demanding an exhaustive search over all the subsets of columns of $\mathbf{D}$ [12]. One of the most effective techniques to approximate its solution is the convex relaxation of the $\ell_0$-norm. It uses the $\ell_1$-norm, the closest convex norm on $\mathbb{R}^L$:

\[
(P_1) : \text{Arg min}_{\alpha \in \mathbb{R}^L} \| \alpha \|_1 \text{ s.t. } \mathbf{D}\alpha = \mathbf{s}.
\]

The solution of $(P_1)$ is carried out by linear programming. We are interested in signals $\mathbf{s} \in \mathcal{S}(\mathbf{D})$ for which the solutions of $(P_0)$ and $(P_1)$ coincide. The idea of using $(P_1)$ to find the sparsest solution is called Basis Pursuit (BP), as coined by Chen, Donoho and Saunders [1, 2].

Let $\alpha$ be a representation of $\mathbf{s}$, with support $\Gamma = \text{supp}(\alpha) \subset \Omega$. The matrix $\mathbf{D}_\Gamma$ is a matrix of size $N \times |\Gamma|$ containing the columns (also
referred to as atoms) of $D$ used for the construction of $s$. This matrix is necessarily full-rank (with rank equals $|\Gamma|$). Knowing the support $\Gamma$ suffices to enable perfect recovery of $\alpha$, and thus our interest is confined to the ability to recover the support $\Gamma$.

**Definition 1.1.** A subset $\Gamma \subset \Omega$ is called $\ell_1$-reconstructible with respect to the dictionary $D$ if the solution of $(P_1)$ coincides with the solution of $(P_0)$ for any signal $s \in S(D)$ that admits a representation with the support $\Gamma$.

The main task of the paper is to obtain conditions on support sizes which imply that they are $\ell_1$-reconstructible. For any specific support $\Gamma \subset \Omega$ there exists a straightforward (yet exhaustive) test whether it admits recovery by BP – simply apply BP to the finite family of signals $s = D\alpha$ generated from coefficient vectors $\alpha$ with the support $\Gamma$ covering all possible sign patterns (i.e. $2^{|\Gamma|}$ such tests$^1$). If the recovery succeeds for all these choices of $\alpha$, it will also succeed for any other representation with support $\Gamma$ [15, 8].

Clearly, such a testing approach is impractical in most cases. If we aim to find the prospects of success of the BP for a fixed cardinality $|\Gamma|$, this requires a set of tests as described above per each possible support $\Gamma$ having such a cardinality, and this implies a need for approximately $L^{|\Gamma|}$ groups of tests. Thus, the exhaustive approach should be replaced either by a random set of tests with empirical claims, or a theoretical study.

Within the theoretical attempts to estimate the power of the BP, two approaches are distinguished in the existing literature. Earlier

---

$^1$In fact, half of this amount is required because if $\alpha$ is reconstructible, then so is $-\alpha$. 
work carried out the worst case analysis for a given dictionary, providing conditions on the support cardinality that guarantee that any support satisfying them is $\ell_1$-reconstructible [15, 7, 5, 13, 14, 11]. These conditions are often very restrictive and far from empirical evidence. Another, more recent, approach presents a probabilistic analysis, providing conditions for special families of dictionaries under which most signals of a given cardinality are $\ell_1$-reconstructible [3, 4, 6]. The results depict a general asymptotic behavior with regard to the sparse support recovery.

In both worst-case and probabilistic-analysis branches of work, the analysis relies heavily on a scalar feature of the dictionary, known as the mutual-coherence. In this work we set to improve the existing worst case results for a given general dictionary $\mathbf{D}$, as reported in [5, 13, 14, 11]. We achieve this progress by replacing the mutual-coherence with a set of alternative features that we refer to as the capacity sets of the dictionary. A thorough computational analysis of $\mathbf{D}$ and probabilistic tools are applied to the problem, leading to improved probabilistic bounds.

In the next section we recall the existing theoretical results concerning $\ell_1$-recovery as a function of the support cardinality. In section 3 we define two versions of the capacity set and present the main theoretical results of this paper using these features. Section 4 expands on the above results by providing two numerical algorithms using the capacity sets. Section 5 provides an overall comparison of the various methods presented in this work to assess the performance of BP for several test-cases.
2. Background

Most known results on sparsity rely on the \textit{mutual-coherence}, denoted as $\mu$, of the dictionary. This is the maximum of the inner products between the columns: $\mu = \max_{i \neq j \in \Omega} |<d_i, d_j>|$. This correlation between the columns, reflected in its worst value by $\mu$, helps establishing the "safe zone" for the support sizes, where both the uniqueness of sparsest representation and its $\ell_1$-recovery can be guaranteed.

For $D = [\Phi_1, \Phi_2]$ a pair of orthonormal bases, the following sufficient condition for $\Gamma$ to be $\ell_1$-reconstructible is proven in [7]:

$$|\Gamma| \leq \frac{\sqrt{2} - 0.5}{\mu}.$$ 

Donoho and Elad in [5] treat a general dictionary $D$. They define the problem \((C_\Gamma) : \max_{\delta \in \text{Null}(D)} \sum_{k \in \Gamma} |\delta_k| \text{ s.t. } \|\delta\|_1 = 1 \), \hspace{1cm} (2.1)

and show that its solution is intimately tied to the ability to recover the support $\Gamma$, by the following lemma:

\textbf{Lemma 2.1.} ([5], Lemma 2) A sufficient condition on the support $\Gamma$ to be $\ell_1$-reconstructible is

$$\text{val}(C_\Gamma) < \frac{1}{2}.$$ \hspace{1cm} (2.2)

This criteria is used to prove the following theorem:

\textbf{Theorem 2.2.} ([5], Theorem 7) A sufficient condition on a support $\Gamma \subset \Omega$ to be $\ell_1$-reconstructible is

$$|\Gamma| < \frac{1}{2} \left(1 + \frac{1}{\mu} \right).$$ \hspace{1cm} (2.3)
Typically, the coherence behaves at best like $O(\frac{1}{\sqrt{N}})$, hence the results stated above predict quite weak $\ell_1$-recovery, which is refuted by the empirical evidence: usually BP recovers supports of size proportional to $N$ (and not its squared-root).

A generalization of the coherence is introduced in [5] and later used by J. Tropp in [11]: for any $0 \leq m \leq L$, the Babel function $\mu_1(m)$ is defined by

$$\mu_1(m) = \max_{|\Lambda|=m} \max_{\eta \in \Omega \setminus \Lambda} \sum_{\lambda \in \Lambda} |<\phi_\lambda, \phi_\eta> |.$$ 

In terms of this function, a support of size $m$ is proven to be $\ell_1$-reconstructible provided the following inequality holds [11]:

$$\mu_1(m - 1) + \mu_1(m) < 1.$$ 

Unfortunately, in cases where the coherence $\mu$ tends towards 1 (implying an existence of at least one problematic pair of atoms), the growth of $\mu_1(m)$ is too fast to provide any improvement.

Average case analysis improves the asymptotic bounds of reconstructible support sizes. Emmanuel Candès and Justin Romberg show in [4] that for the dictionary $D = [I, F^*]$, where $F$ is the Fourier transform, random uniformly sampled support admits $\ell_1$-recovery with high probability if (the expectation of) its cardinality is $O(N/\log N)$, which improves the $O(\sqrt{N})$ estimation of the worst case approach. For a general orthonormal pair, it is shown that most random supports which cardinality behaving like $O(1/(\mu^2 \log^6 N))$ admit recovery by BP. The $\log N$ appearing in these expressions is suspected by the authors of [4] to be unnecessary, which in effect turns this expression into $O(N)$.

This conjecture is supported by the results of David Donoho [6], who considers dictionaries $D$ of size $N \times L$ constructed by concatenating
random vectors of unit $l_2$-norm, independently drawn from the uniform distribution. This work shows that there exists a constant $\rho(A) > 0$ such that all but negligible proportion of such matrices $D$ with $L = AN$, have the following property: for every system $s = D\alpha$ allowing a solution with fewer than $\rho(A)N$ non-zeros, $\ell_1$-minimization uniquely finds it. In effect, this means that the upper bound on recoverable support sizes by BP is asymptotically linear in $N$.

As good as these results sound, they do not provide an explicit information about the ability of $\ell_1$-reconstruction applied to a specifically given dictionary $D$, which is a practical and central question in the application of BP. Thus, such information can only be obtained today by the worst case results, and those involve the coherence $\mu$ or its descendants (like $\mu_1(m)$). In this work we introduce new features of the dictionary $D$, the capacity sets. These features are obtained as the solutions to specific linear programming problems that probe the dictionary $D$. We consider two such options: a vector of capacities $q$ and a matrix $Q$, as we shall explain in details in the next section. These features are used to develop novel analysis of BP performance as a function of the support’s cardinality.

One interesting benefit of the proposed analysis is a better treatment of dictionaries which are not “uniformly coherent”. In cases where there exists a small set of columns in $D$ with strong linear dependency, the coherence and the babel function behave badly, tending to lead to overly pessimistic bounds. As we show, the use of the capacities leads in these cases to much better results. Besides that, the capacities are shown to be more delicate indicators of the dictionary, as reflected in a better prediction of the BP performance.
Our results cannot be compared directly with the above-mentioned work of Candès and Romberg [4]. However, the fact their results contain the parameter \( \mu \) in their formulae leads us to believe that the capacities introduced here could be used to extend their results and provide further improvement on the average performance analysis, along the lines they provide.

3. The Capacity Sets and Their Use

In this section we define two versions of the capacity sets, and state the main theoretical results that employ them for the analysis of the BP.

3.1 The Capacity Vector \( q \)

The capacity vector consists of elements related to an intermediate tool used in the proof of Theorem 2.2 in [5]:

**Definition 3.1.** The capacity vector \( q = (q_1, ..., q_L)^T \) of a dictionary \( D \in \mathbb{R}^{N \times L} \) is defined for all \( k \in \Omega \) by

\[
q_k = \max_{\delta \in \text{Null}(D)} \delta_k \quad \text{s.t.} \quad \|\delta\|_1 = 1.
\]

Computing the elements of \( q \) is relatively easy, and amounts to a simple set of \( L \) independent linear programming problems of the form

\[
\hat{x}_k = \arg \min_x \|x\|_1 \quad \text{subject to} \quad Dx = 0 \quad \text{and} \quad x_k = 1,
\]

and then assigning \( q_k = 1/\|\hat{x}_k\|_1 \).

Via Lemma 2.1, the definition of \( q \) provides a sufficient condition \( \sum_{k \in \Gamma} q_k < \frac{1}{2} \) on a given support \( \Gamma \) to ensure its recovery by \( \ell_1 \)-minimization. Furthermore, by gathering the \( |\Gamma| \) largest entries from \( q \), a simple generalization of Theorem 2.2 can be proposed. However,
in this work we seek a better bound that takes into account the variety of possible supports, rather than the worst one. One such numerical technique is suggested in section 4, proposing a special quantization of the values in \( q \) to obtain a lower bound on the fraction of support sizes which admit recovery by BP.

In this section we aim to obtain a more theoretically flavored result that uses \( q \). Let \( z_q : \Omega \to [0, 1] \) be the random variable defined by \( z_q(k) = q_k \), with the uniform distribution over \( \Omega \). The following theorem uses its expectation \( E_q = \mathbb{E}(z_q) \) and variance \( \sigma^2_q = \text{var}(z_q) \) to evaluate the probability of \( \ell_1 \)-reconstruction for a given support size:

**Theorem A.** For any \( 1 \leq \ell < \frac{1}{2E_q} \), a support \( \Gamma \) of size \( \ell \), sampled uniformly at random from \( \Omega \), admits \( \ell_1 \)-recovery with probability

\[
P(\ell) \geq \frac{\left( \frac{1}{2} - \ell E_q \right)^2}{\ell \sigma^2_q + \left( \frac{1}{2} - \ell E_q \right)^2},
\]

where \( E_q = \mathbb{E}(z_q) \) and \( \sigma^2_q = \mathbb{E}((z_q - E_q)^2) \).

**Proof:** We define the random variable \( x_\ell \), being the sum of \( \ell \) randomly chosen entries from \( q \) with uniform probability. For any \( \Gamma \subset \Omega \) such that \( |\Gamma| = \ell \), \( x_\ell = \sum_{k \in \Gamma} q_k \). Thus the probability \( P(\ell) \), defined in the statement of the theorem, is simply \( P(x_\ell \leq \frac{1}{2}) \).

In order to analyze this random variable, we construct a similar probability model, considering another random variable \( y_\ell \) which is the sum of \( \ell \) independent instances of the variable \( z_q \) defined before the statement of the theorem. Thus, an instance of \( y_\ell \) is the sum of \( \ell \) randomly (and independently) chosen elements from \( q \), which indices are not necessarily distinct anymore.

We shall bound the probability \( P(x_\ell \leq \frac{1}{2}) \) by means of the Tchebychev inequality, which involves the mean and the variance of \( x_\ell \). These
parameters are easily computable for \( y_\ell \); our result is based on the following connection between the variables \( x_\ell \) and \( y_\ell \):

\[
\mathbb{E}(x_\ell) = \mathbb{E}(y_\ell) \quad \text{and} \quad \text{var}(x_\ell) \leq \text{var}(y_\ell) \tag{3.2}
\]

See Appendix A for a proof of this claim. By definition of \( y_\ell \) we have \( \mathbb{E}(y_\ell) = \ell E_q \), \( \text{var}(y_\ell) = \ell \sigma^2_q \). These parameters for \( z_q \) are readily computable:

\[
E_q = \mathbb{E}(z_q) = \frac{1}{L} \sum_{k \in \Omega} q_k \quad \text{and} \quad \sigma^2_q = \text{var}(z_q) = \frac{1}{L} \sum_{k \in \Omega} (q_k - E_q)^2.
\]

Given any real scalar \( a > 0 \), the one-tailed version of the Tchebychev inequality [16] for \( x_\ell \) reads

\[
P(x_\ell - E_X \geq a \sigma_X) = P(x_\ell \geq E_X + a \sigma_X) \leq \frac{1}{1 + a^2},
\]

where \( E_X = \mathbb{E}(x_\ell) \), \( \sigma^2_X = \text{var}(x_\ell) \).

By (3.2), we substitute \( E_X = \ell E_q \). Also, since a larger variance implies a lower probability, we put \( \sqrt{\ell} \sigma_q \) instead of \( \sigma_X \) and obtain

\[
P(x_\ell \geq \ell E_q + a \sqrt{\ell} \sigma_q) \leq P(x_\ell \geq E_X + a \sigma_X) \leq \frac{1}{1 + a^2}.
\]

The parameter \( a \) is chosen such that \( \ell E_q + a \sqrt{\ell} \sigma_q = \frac{1}{2} \), leading to \( a = (\frac{1}{2} - \ell E_q)/(\sqrt{\ell} \sigma_q) \). Note that the condition \( a > 0 \) translates to the requirement \( \ell < \frac{1}{2 E_q} \) as claimed in the theorem. In case it holds, we have

\[
P\left(x_\ell \geq \frac{1}{2}\right) \leq \frac{1}{1 + \left(\frac{\frac{1}{2} - \ell E_q}{\ell \sigma^2_q}\right)^2},
\]

or put differently,

\[
P(x_\ell \leq \frac{1}{2}) \geq 1 - \frac{1}{1 + \left(\frac{\frac{1}{2} - \ell E_q}{\ell \sigma^2_q}\right)^2} = \frac{(\frac{1}{2} - \ell E_q)^2}{\ell \sigma^2_q + (\frac{1}{2} - \ell E_q)^2},
\]

as stated by the theorem. \( \square \)
3.2 The Weak Capacity Vector

We mentioned earlier that previous work often uses the mutual coherence to derive performance bounds on $\ell_1$ reconstructible supports. What is the relation between the capacities in $q$ and the inner products between the dictionary atoms, $| \langle d_i, d_j \rangle |$, described earlier? Apparently, this has been already answered in [5].

Given a dictionary $D$, construct its Gram matrix as $G = D^T D$. Define the sequence

$$\mu_k = \max_{i \neq k} |G_{i,k}| \text{ for } k \in \Omega.$$  \hfill (3.3)

Namely, $\mu_k$ is the maximal value on the $k$-th column of $|G|$, disregarding the main diagonal entry. As [5] shows, this sequence of values satisfies

$$q_k \leq \frac{\mu_k}{\mu_k + 1}.$$  

Thus the condition $\sum_{k \in \Gamma} q_k \leq \frac{1}{2}$ can be replaced with $\sum_{k \in \Gamma} \frac{\mu_k}{\mu_k + 1} \leq \frac{1}{2}$, leading of-course, to weaker bounds.

3.3 Using the Capacity Matrix $Q$

One problem with the capacity vector $q$ is the independence with which its entries $q_k$ are computed. This implies that one (or more) of the entries in $q$ may become unnecessarily large, compared to the values obtained in Equation (2.1), causing a weaker bound. By working with pairs of such entries, one could in principle improve the obtained bounds. This leads us to the following definition:

**Definition 3.2.** Denote by $\Omega_2$ the set of indices $\Omega_2 = \{(i,j) | i, j \in \Omega, i < j\}$. The upper triangular capacity matrix $Q = \{Q_{i,j}\}$ is the
matrix with non-zero elements indexed by \((i, j) \in \Omega_2\), defined as follows:

\[
Q_{i,j} = \max_{\delta \in \text{Null}(D)} (\delta_i + \delta_j, \delta_i - \delta_j) \quad \text{s.t.} \quad \|\delta\|_1 = 1.
\]

Each of these entries can be computed by two independent linear programming problems of the form

\[
\begin{align*}
\left\{ \begin{array}{l}
x^+_{(i,j)} = \arg \min_x \|x\|_1 \quad \text{subject to} \quad Dx = 0 \quad \text{and} \quad x_i + x_j = 1 \\
x^-_{(i,j)} = \arg \min_x \|x\|_1 \quad \text{subject to} \quad Dx = 0 \quad \text{and} \quad x_i - x_j = 1
\end{array} \right.
\]

and then assigning \(Q_{i,j} = 1/\min(\|\hat{x}^+_{(i,j)}\|_1, \|\hat{x}^-_{(i,j)}\|_1)\).

As in section 3.1, the obtained values \(Q_{i,j}\) could be used to form an improved worst-case bounds for Lemma 2.1 and consequently for Theorem 2.2: Let \(\Gamma \subset \Omega\) be a randomly chosen support of size\(^2\ell = 2n\).

By definition, the non-zero elements of \(Q\) satisfy

\[
\max_{\delta \in \text{Null}(D)} \|\delta\|_1 = \max_{\|\delta\|_1 = 1} \max_{\delta \in \text{Null}(D)} (|\delta_i| + |\delta_j|) = q_i + q_j.
\]

Thus the values \(Q_{i,j}\) can be used in the evaluation of an upper bound on \(C_\Gamma\). To any partition \(\mathcal{I}\) of \(\Gamma\) into disjoint pairs there corresponds the sum \(\sum_{(k_1,k_2) \in \mathcal{I}} Q_{k_1,k_2}\) that bounds the value of \(C_\Gamma\) from above. Therefore, \(\Gamma\) is \(\ell_1\)-reconstructible if there exists such a partition satisfying \(\sum_{(k_1,k_2) \in \mathcal{I}} Q_{k_1,k_2} \leq \frac{1}{2}\). Naturally, among all such possible partitions, we are interested in the one that leads to the smallest sum.

Working along the lines described above, one could propose a numerical algorithm to evaluate the probability of a given support cardinality to get BP’s success. Such an algorithm is described in section 4.

\(^2\text{We consider hereafter even support sizes. Generalization to odd ones is relatively simple, requiring the use of one entry from } q. \text{ We omit this discussion for simplicity.}\)
Here we concentrate again on a theoretical bound that uses $Q$, similar to the one proposed in Theorem A with few necessary modifications.

Let $w_Q$ be the random variable $w_Q : \Omega_2 \rightarrow [0,1]$, defined by $w_Q(i,j) = Q_{i,j}$, with uniform distribution over the upper triangular part of $Q$. The following result based on $Q$ is similar to the one in Theorem A:

**Theorem B.** For any $1 \leq \ell < \frac{1}{E_Q}$, a support $\Gamma$ of even size $\ell$, sampled uniformly at random from $\Omega$, admits $\ell_1$-recovery with probability

$$P(\ell) \geq \frac{\left(\frac{1}{2} - \frac{\ell}{2}E_Q\right)^2}{\frac{\ell}{2} \sigma_Q^2 + \left(\frac{1}{2} - \frac{\ell}{2}E_Q\right)^2},$$

where $E_Q = \mathbb{E}(w_Q)$, $\sigma_Q^2 = \mathbb{E}((w_Q - E_Q)^2)$.

Notice that the expression obtained in Equation (3.4) is the same as the one in (3.1), with $\ell$ replaced by $\ell/2$. Since $E_Q$ and $\sigma_Q$ refer to pairs, if $E_Q = 2E_q$ and $\sigma_Q^2 = 2\sigma_q^2$ the two bounds are the same. However, as we shall demonstrate in section 5, $E_Q < 2E_q$ and $\sigma_Q^2 < 2\sigma_q^2$, implying that this bound is indeed stronger. We now turn to the proof of Theorem B.

**Proof:** The proof is very similar to the one of Theorem A. We define the random variable $x_\ell$, being the sum of $\ell/2$ randomly chosen entries from the upper-triangular part of $Q$ with uniform probability, such that no index appears twice. Thus the probability $P(\ell)$, defined in the statement of the theorem, is simply $P(x_\ell \leq \frac{1}{2})$ since we have covered all possible and feasible supports $\Gamma$ of cardinality $\ell$ with uniform probability.

As in Theorem A, we define a simpler random variable, $y_\ell$ which is the sum of $\ell/2$ independent instances of $w_Q$. Thus, an instance of
$y_\ell$ is the sum of $\ell/2$ randomly (and independently) chosen elements from the Capacity Matrix, which indices are not necessarily distinct anymore. We shall assume hereafter that

$$E(x_\ell) = E(y_\ell) = \frac{\ell}{2} E_q \quad \text{and} \quad \text{var}(x_\ell) \leq \text{var}(y_\ell) = \frac{\ell}{2} \sigma_q^2. \quad (3.5)$$

We do not provide a proof of this property and leave it as an open question at this stage. Empirical verification of this property is demonstrated in Appendix B.

Following the steps of Theorem A, given any real $a > 0$, the one-tailed version of the Tchebychev inequality [16] for $x_\ell$ reads

$$P \left( x_\ell \geq \frac{\ell}{2} E_q + a \sqrt{\frac{\ell}{2} \sigma_q} \right) \leq \frac{1}{1 + a^2}.$$  

The parameter $a$ is chosen such that $\frac{\ell}{2} E_q + a \sqrt{\frac{\ell}{2} \sigma_q} = \frac{1}{2}$, leading to $a = (\frac{1}{2} - \frac{\ell}{2} E_q)/\left(\sqrt{\frac{\ell}{2} \sigma_q}\right)$, implying that we should require $\ell < \frac{1}{E_q}$ to get $a > 0$. This leads to

$$P \left( x_\ell \geq \frac{1}{2} \right) \leq \frac{1}{1 + \left(\frac{1}{2} - \frac{\ell}{2} E_q\right)^2},$$

or put differently,

$$P(X_\ell < \frac{1}{2}) > 1 - \frac{1}{1 + \left(\frac{1}{2} - \frac{\ell}{2} E_q\right)^2} = \frac{(\frac{1}{2} - \frac{\ell}{2} E_q)^2}{\frac{\ell}{2} \sigma_q^2 + (\frac{1}{2} - \frac{\ell}{2} E_q)^2},$$

as stated in the theorem. □

4. Numerical Algorithms

Given the capacity vector $\mathbf{q}$ (or its weaker version as described in section 3.2) or matrix $\mathbf{Q}$, we can use Theorems A and B to predict the
\( \ell_1 \)-reconstructible supports, and show lower bounds of the probability for success as a function of the support size \( \ell \). However, we can alternatively evaluate these probabilities numerically, provided that there are shortcuts that avoid the exponential growth in support possibilities. This leads us to the following two algorithms.

### 4.1 A Fast Combinatorial Count Using \( q \)

We would like to establish the fraction of the total number of supports \( \Gamma \) of size \( \ell \) that satisfy \( \text{val}(C_\Gamma) \leq \frac{1}{2} \). Testing the sufficient condition \( \sum_{k \in \Gamma} q_k \leq \frac{1}{2} \) for every single \( \Gamma \) requires \( O(L^\ell) \) flops, which is prohibitive. Instead, we propose to perform a quantization of the entries of \( q \) to \( d \) distinct values, and lead to a more reasonable computational process.

Suppose we are given a partition \( \Lambda = \{ \Lambda_i \}_{i=1}^d \) of \( \Omega \) into \( d \) disjoint clusters, such that \( \Omega = \bigcup_{i=1}^d \Lambda_i \). The corresponding quantized values in \( q \) are denoted by \( \{ q^i_\Lambda \} \), each set to be the maximal in its subset, \( \{ q^i_\Lambda = \max_{k \in \Lambda_i} (q_k) \mid 1 \leq i \leq d \} \).

Given the quantization parameters \( \Lambda = \{ \Lambda_i, q^i_\Lambda \}_{i=1}^d \) every \( \ell \)-sized support \( \Gamma \in \Omega \) can be described as the union \( \bigcup_{i=1}^d \Gamma_i \), where \( \Gamma_i \subseteq \Lambda_i \) is the subset of indices in \( \Gamma \) allocated to the quantized value \( q^i_\Lambda \). Thus, the sum \( \sum_{k \in \Gamma} q_i \) can be replaced by a larger sum, \( \sum_{i=1}^d |\Gamma_i| q^i_\Lambda \).

In order to test all possible supports \( \Gamma \in \Omega \) of size \( \ell \), a combinatorial count of all sequences \( p = (p_1, ..., p_d) \) is performed, such that \( 0 \leq |p_i| \leq |\Lambda_i| \) and \( \sum_{i=1}^d |p_i| = \ell \). For each of these we evaluate \( \sum_{i=1}^d |p_i| q^i_\Lambda \) and count the relative number of those\(^3\) below \( \frac{1}{2} \). The complexity of such computation does not exceed \( O \left( \left( \frac{L}{d} \right)^d \right) \).

\(^3\)Each instance must be weighted by the number of its possible occurrences.
As to the choice of the quantization parameters \( \Lambda = \{ \Lambda_i, q^i_\Lambda \}_{i=1}^d \), as said above, we let
\[
q^i_\Lambda = \max_{k \in \Lambda_i} q_k
\]
to guarantee that the evaluated summations are considering a worst-case scenario. The clustering is done by an attempt to minimize the function
\[
f \left( \{ \Lambda_i, q^i_\Lambda \}_{i=1}^d \right) = \sum_{i=1}^d \left( |\Lambda_i| q^i_\Lambda - \sum_{k \in \Lambda_i} q_k \right).
\]
(4.1)
The difference \( |\Lambda_i| q^i_\Lambda - \sum_{k \in \Lambda_i} q_k \) is the quantization error for the elements in the subset \( \Lambda_i \), and the above error simply sums these values.

The minimization of \( f \left( \{ \Lambda_i, q^i_\Lambda \}_{i=1}^d \right) \) can be done exhaustively in case \( d \) is small – in our experiments we have used \( d = 3 \) implying that the above requires \( O(L^3) \) flops. For larger values of \( d \) a sequential algorithm that chooses \( \Lambda_i \) can be proposed, separating the set \( \Omega \) to two parts, and proceeding in a tree and greedy separation scheme.

Computationally, the results of the combinatorial count are very close to those predicted by Theorem A. Therefore, this method serves as a supporting evidence for the probabilistic approach taken in Theorem A, but its numerical output is omitted from our display of experimental results in section 5.

### 4.2 A Sampling Algorithm Using Q

An alternative to Theorem B is a direct evaluation of \( \ell_1 \)-reconstructible supports \( \Gamma \) of cardinality \( \ell \), by the following stages:

- We draw \( M \gg L \) such supports \( \{ \Gamma_i \}_{i=1}^M \).
- For each \( \Gamma_i \) we evaluate a near-optimal partition \( I_i \) that leads to the smallest value of \( \sum_{(k,l) \in \mathcal{I}} Q_{k,l} \). While finding the best partition is combinatorial in complexity, an approximate algorithm of complexity \( O(\ell^3) \) can be proposed, again leaning on a greedy
accumulation of the smallest possible entries from $Q$, one pair at a time.

- Given the near optimal partition, test $\sum_{(k,l)\in I} Q_{k,l} \leq \frac{1}{2}$ and accumulate the relative number of such occurrences.

The fact that this method relies on capacity values implies that the predicted performance is expected to be weaker compared to the true behavior of BP. Nevertheless, among the various methods discussed thus far, this method is expected to be the most optimistic because it uses $Q$ and not $q$, and also because it does not build the evaluation through the Tchebychev inequality that looses also part of the tightness. However, as opposed to all the other methods described above, this method cannot claim theoretical correctness of its results.

## 5. Experimental Results

### 5.1 Test-Cases to Study

We carry out a number of tests on each of the three following dictionaries:

1. **D – Random** is the dictionary of size $128 \times 256$, which consists of $\ell_2$-normalized random vectors, independently drawn from the Normal distribution on the unit sphere. Such a dictionary is often used in numerical experiments as well as in various applications.

2. **D – Spoiled** is the dictionary $D – Random$, which has undergone an operation designed to create a small set of columns with high linear dependence. More precisely, we re-generate a set of 3 columns as a random linear combination of other 12
columns. This dictionary is used to demonstrate the ability of the capacity-sets methods to better handle dictionaries with a non-uniform distribution of inner products.

3. \( D - DCT \) is the orthonormal pair \([I, C^*]\) of size 128\( \times \)256, where \( C \) is the 1-dimensional Discrete Cosine basis and \( I \) the identity matrix.

5.2 Behavior of \( q \) and \( Q \)

As explained earlier, the passage from the capacity vector \( q \) to the matrix \( Q \) was motivated by the fact that \( Q_{i,j} \) provide a lower bound in this context. To exhibit the numerical behavior of these bounds, we compute the mean and the variance of the family of ratios

\[
R_{k,l} = \frac{Q_{k,l}}{q_k + q_l} \quad \text{for} \quad k \neq l \in \Omega.
\]

(5.1)

The mean and variance of these ratios for the three test cases is given in Table 1.1.

| Dictionary | \( \mathbb{E}(R) \) | \( \sigma(R) \) |
|------------|----------------|--------------|
| \( D - \text{Random} \) | 0.7175 | 0.0008 |
| \( D - \text{Spoiled} \) | 0.7154 | 0.001 |
| \( D - DCT \) | 0.6509 | 0.0109 |

Table 1.1

Behavior of the capacity-sets \( q \) and \( Q \) by evaluating the mean and variance of the ratios.

As these figures show, we earn up to 30\% of the upper bound value by upgrading to Capacity Matrix from the Capacity Vector. This ratio between the two bounds for the corresponding indices is very stable, as seen from the low values of the standard deviation \( \sigma(R) \).
5.3 Compared Methods

We perform a number of computations, applying various methods for the estimation of BP performance on the given dictionaries. The results are expressed via a set of Estimation Functions, $EF : \Omega \rightarrow \mathbb{R}$, which value at $\ell \in \Omega$ is the predicted percentage of $\ell$-sized supports which admit recovery by $\ell_1$-norm optimization. The EFs considered are the following:

1. EF-emp - The standard empirical test on the dictionary. This test is done by drawing 1,000 random supports for each cardinality $\ell$, generating a corresponding signal, and solving the BP per each. EF-emp is obtained by showing the relative number of successes in recovering the support.

2. Classical Bound - the coherence-based upper bound $\frac{1}{2}(1 + \frac{1}{\mu})$, provided by the Theorem 2.2.

3. EF-thmA - The results of the Theorem A, $EF-thmA(\ell) = P(\ell)$ as defined in the statement of the theorem. The values are computed from $q$ of the dictionary.

4. EF-thmB - The results of the Theorem B, computed from the capacity matrix $Q$ of the dictionary.

5. EF-compB - The results of the sampling algorithm based on $Q$, which results support the estimation of Theorem B (see section 4.2).

6. Grassmanian Bound - The Grassmanian upper bound, computed by the formula for the Classical Bound using the ideal coherence $\mu = \sqrt{\frac{L-N}{N(L-1)}}$. 
This last EF deserves more explanation: Among all possible dictionaries of size $N \times L$, the Grassmannian frame is the one leading to the smallest possible coherence $\mu = \sqrt{\frac{L-N}{N(L-1)}}$ \cite{9}. Thus, this leads to the most optimistic worst-case bound. When the dictionary is “un-balanced”, implying a large spread of inner-products in the Gram-matrix, we know that the mutual-coherence-bound deteriorates dramatically. Thus, by using the Grassmannian Bound, we test what is the best achievable coherence-based performance behavior for the same dictionary size.

5.4 Comparison Results

Figures 1, 2, and 3 presents the obtained graphs of the various EF-s functions described above, for the three dictionaries described at the top of this section. As we see from the left-side graphs in the figures, for all the dictionaries the empirically established support size which admits BP recovery is at least 40 columns, which is far superior to all theoretical predictions. Nevertheless, the estimation made by the sampling algorithm based on the Capacity Matrix (EF-compB) is much better than the Classical Bound, established so far in the literature. The difference is especially high for the D-Spoiled dictionary, which reflects the fact that methods based on Capacity Sets manage well the non-uniform distribution of inner products.

On the right side of each figure we display various method developed in this work. Noticeably, the results of Theorem B (EF-thmB) are stronger than those of Theorem A (EF-thmA), which is explained by the benefit of using the Capacity Matrix rather than the Capacity Vector. This benefit is expressed in the ratio values given in the Table 1.1 and explained thereafter. Apparently, Theorem B does not express
Figure 1: Random dictionary – The EF functions compared with the classical and empirical bounds.

Figure 2: Spoiled dictionary – The EF functions compared with the classical and empirical bounds.
the full power of the Capacity Matrix estimation, since the sampling algorithm based on its values (EF-compB) outperforms EF-thmB by 15 – 20%. This algorithm produces values which are quite close to the Grassmanian Bound, the best possible bound one can hope to obtain using coherence-based estimation for the given dictionary size. We do not have enough information to explain the fact that values of EF-compB and of Grassmanian bound nearly coincide for all the dictionaries discussed here (and additional ones examined during the work); Discovering the reason underlying this connection may be a lead to important insights regarding the Basis Pursuit performance.

5.5 Varying dictionary size

We carry out an additional experiment which is designed to trace the increment in the BP-recoverable support size as a function of dictionary size. We use the $D - Random$ dictionary of size $2^n \times 2^{n+1}$, with $n = 2, ..., 8$. For each dictionary, we compute, by the formula provided in theorem A, the support size which admits $l_1$-recovery with (at least)
94% success.

Figure 4: The BP-reconstructible support size as a function of dictionary size

Though the number of dictionaries is too small to speak of general nature of this curve, one can see that the support size $\ell$ increases at square rate relative to $\log(N)$, where $N$ is the dimension of the signal (number of rows in the dictionary). The coherence-based analysis suggests that $\ell$ behaves roughly like $\sqrt{L}$ (say, for Fourier-Identity orthonormal pair, $\mu = N^{-0.5}$), Since $\log^2(N)$ is larger than $\sqrt{N}$ for $N < 5500$, we conclude that the results of Theorem A indeed improve on the existing estimation bounds for the BP-performance on this scale.

Appendix A

Let $z_q : \Omega \rightarrow [0, 1]$ be the random variable defined by $z_q(k) = q_k$, with the uniform distribution over $q$, as defined in section 3. We define
the random variable $x_\ell$, being the sum of $\ell$ randomly chosen entries from $q$ with uniform probability. For any $\Gamma \subset \Omega$ such that $|\Gamma| = \ell$, $x_\ell = \sum_{k \in \Gamma} q_k$. Similarly, we construct another random variable $y_\ell$ which is the sum of $\ell$ independent instances of the variable $z_q$. Thus, an instance of $y_\ell$ is the sum of $\ell$ randomly (and independently) chosen elements from $q$, which indices are not necessarily distinct anymore.

**Theorem C.** For the two random variables, $x_\ell$ and $y_\ell$, the following relations between the first and second moments are true:

$$E(x_\ell) = E(y_\ell) \quad \text{and} \quad \text{var}(x_\ell) \leq \text{var}(y_\ell). \quad (A-1)$$

**Proof:** We begin by introducing some notation. Fix a natural number $1 \leq \ell \leq L$ which stands for the size of the support. For any $1 \leq k \leq \ell$, we denote by $C^k_\ell$ the collection of all $\ell$-sized sets of indices from $\Omega$ (with repetitions), where each such set has precisely $k$ distinct indices. Also, we let $D^\ell_n = C^\ell_\ell \cup C^{\ell-1}_\ell \cup \ldots \cup C^{\ell-n}_\ell$, i.e. $D^\ell_n$ is the collection of all $\ell$-sized sets of indices from $\Omega$ where each set has at least $\ell-n$ distinct elements.

In this notation, the domain of the random variable $x_\ell$ is the collection $D^\ell_0$, and the domain of the variable $y_\ell$ is $D^{\ell-1}_\ell$. Both $x_\ell$ and $y_\ell$ can be considered as the same random variable (sum of elements from capacity vector) defined on different domains. Therefore, we write $x_\ell = x_{|D^\ell_0}$, $y_\ell = x_{|D^{\ell-1}_\ell}$. In the proof we use the following basic property of the variance:

**Proposition 5.1.** Let $z$ be a random variable defined on a domain given as the disjoint union $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2 \cup \ldots \cup \mathcal{D}_n$. Denote $v = \text{var}(z|\mathcal{D})$, $v_i = \text{var}(z|\mathcal{D}_i), s_i = |\mathcal{D}_i|$. Then $v = \frac{\sum_{i=1}^n s_i v_i}{\sum_{i=1}^n s_i}$.

**Part 1.** The expectation of the random variable $x$ restricted to $\mathcal{D}^\ell_0$ is
computed by
\[ \mathbb{E}(x_{|D^0_\ell}) = \frac{1}{|D^0_\ell|} \sum_{\Lambda \in D^0_\ell} \sum_{k \in \Lambda} q_k. \]
This sum contains \(|D^0_\ell| \cdot \ell\) elements, and for each \(j \in \Omega\), \(q_j\) appears in it same number of times. Therefore, each \(q_j\) appears \(|D^0_\ell| \cdot \ell\) times, and we have \(\mathbb{E}(x_{|D^0_\ell}) = \frac{\ell}{L} \sum_{k \in \Omega} q_k = \ell E_q\). The mean of \(x_{|D^{\ell-1}_\ell}\) is computed similarly:
\[ \mathbb{E}(x_{|D^{\ell-1}_\ell}) = \frac{1}{|D^{\ell-1}_\ell|} \sum_{\Lambda \in D^{\ell-1}_\ell} \sum_{k \in \Lambda} q_k. \]
Here each \(q_j\) appears \(|D^{\ell-1}_\ell| \cdot \ell\) times, and we have \(\mathbb{E}(x_{|D^{\ell-1}_\ell}) = \frac{\ell}{L} \sum_{k \in \Omega} q_k = \ell E_q\).

In the light of this result, for the proof of variance inequality we can assume that the expectation of \(x\) on both domains equals zero. Notice this is equivalent to assumption \(E_q = 0\), since \(\mathbb{E}(x_{|D^{\ell-1}_\ell}) = \ell E_q\).

**Part 2.** We consider the extension of the random variable \(x\) to the domain which is the collection of \(\ell\)-sized sets of elements from \(\Omega\), where every set may appear any (finite) number of times. For any \(0 \leq n < \ell\), we define two disjoint unions

\[ A_n = \bigcup_{\Gamma \in D^{\ell-1}_n} \{ \Gamma \cup \{j\} \mid j \in \Gamma \}, \]

\[ B_n = \bigcup_{\Gamma \in D^{\ell-1}_n} \{ \Gamma \cup \{j\} \mid j \in \Omega \} \]

(In the definition of \(A_n\), the set \(\Gamma \cup \{j\}\) is added to the collection one time for each appearance of \(j\) in \(\Gamma\).)

Let \(\Lambda \in \mathcal{C}_\ell^k\) be a set which contains distinct indices \(j_1, \ldots, j_k\) with multiplicities \(m_1, \ldots, m_k\) (so that \(\sum_{i=1}^k m_i = \ell\)). For each \(1 \leq i \leq k\), \(\Lambda\) is obtained in \(A_n\) \(m_i - 1\) times in the form \(\Gamma \cup \{j_i\}\) for an appropriate \(\Gamma = \Gamma_i \in \mathcal{C}_{\ell-1}^k\) (this claim also holds vacuously for \(m_i = 1\)). Therefore,
the number of copies of \( \Lambda \) in \( \mathcal{A}_n \) equals \( \sum_{i=1}^{k} (m_i - 1) = \ell - k \). Also, \( \Lambda \) appears in \( \mathcal{B}_n \) precisely once for each \( j_1, \ldots, j_k \), in the form \( \Gamma \cup \{ j_i \} \) (for an appropriate \( \Gamma = \Gamma_i \) each time). Therefore, \( \mathcal{B}_n \) contains \( k \) copies of \( \Lambda \).

Denote disjoint union of \( a \) distinct copies of some collection \( \mathcal{C} \) by \( a \cdot \mathcal{C} \). Then we can write \( \mathcal{A}_n, \mathcal{B}_n \) as

\[
\mathcal{A}_n = 0 \cdot \mathcal{C}_\ell^\ell \cup 1 \cdot \mathcal{C}_\ell^{\ell-1} \cup \ldots \cup n \cdot \mathcal{C}_\ell^{\ell-n} \quad (A-2)
\]

\[
\mathcal{B}_n = \ell \cdot \mathcal{C}_\ell^\ell \cup (\ell - 1) \cdot \mathcal{C}_\ell^{\ell-1} \cup \ldots \cup (\ell - n) \cdot \mathcal{C}_\ell^{\ell-n} \quad (A-3)
\]

We prove the following inequality:

\[
\text{var}(x|\mathcal{B}_n) \leq \text{var}(x|\mathcal{A}_n).
\]

Since \( \mathbb{E}_q = 0 \) by our assumption, the expectations of \( x|\mathcal{A}_n \) and \( x|\mathcal{B}_n \) also equal zero: by the argument similar to one presented in the first part of the proof, \( \mathbb{E}(x|\mathcal{A}_n) = \mathbb{E}(x|\mathcal{B}_n) = \ell \cdot \mathbb{E}_q \). Thus we have

\[
\text{var}(x|\mathcal{A}_n) = \frac{1}{|\mathcal{D}_\ell^{n}|} \sum_{\mathcal{D}_\ell^{n-1}} \frac{1}{\ell - 1} \sum_{\Gamma \in \mathcal{D}_\ell^{n-1}} \big( \sum_{k \in \Gamma} q_k + q_j \big)^2.
\]

For the brevity of the argument we introduce the notation \( q_\Gamma = \sum_{k \in \Gamma} q_k \).

Then \( \text{var}(x|\mathcal{A}_n) \) reads as

\[
\text{var}(x|\mathcal{A}_n) = \frac{1}{|\mathcal{D}_\ell^{n}|} \sum_{\mathcal{D}_\ell^{n-1}} \frac{1}{\ell - 1} \sum_{j \in \Gamma} \big( q_\Gamma^2 + q_j^2 + 2 q_\Gamma q_j \big) = \frac{1}{|\mathcal{D}_\ell^{n}|} \sum_{\mathcal{D}_\ell^{n-1}} q_\Gamma^2 + \frac{1}{\ell - 1} \sum_{j \in \Gamma} \big( q_j^2 + 2 q_\Gamma q_j \big).
\]

Similarly, we have

\[
\text{var}(x|\mathcal{B}_n) = \frac{1}{|\mathcal{D}_\ell^{n}|} \sum_{\mathcal{D}_\ell^{n-1}} \frac{1}{L} \sum_{j \in \Omega} \big( \sum_{k \in \Gamma} q_k + q_j \big)^2 = \frac{1}{|\mathcal{D}_\ell^{n}|} \sum_{\mathcal{D}_\ell^{n-1}} q_\Omega^2 + \frac{1}{L} \sum_{j \in \Omega} \big( q_j^2 + 2 q_\Omega q_j \big).
\]
The summand \( \frac{1}{|\mathcal{D}_{\ell-1}^n|} \sum_{\Gamma \in \mathcal{D}_{\ell-1}^n} q_i^2 \) appears in both expressions hence cancels out. We consider the term \( \frac{1}{|\mathcal{D}_{\ell-1}^n|} \sum_{\Gamma \in \mathcal{D}_{\ell-1}^n} \frac{1}{\ell-1} \sum_{j \in \Gamma} q_j^2 \) in \( \text{var}(x|_{\mathcal{A}_n}) \).

The element \( q_a^2 \) appears in it same number of times for every \( a \in \Omega \). Hence \( \frac{1}{|\mathcal{D}_{\ell-1}^n|} \sum_{\Gamma \in \mathcal{D}_{\ell-1}^n} \frac{1}{\ell-1} \sum_{j \in \Gamma} q_j^2 = \frac{1}{L} \sum_{a \in \Omega} q_a^2 \). By same argument, in the expression of \( \text{var}(x|_{\mathcal{B}_n}) \) we have \( \frac{1}{|\mathcal{D}_{\ell-1}^n|} \sum_{\Gamma \in \mathcal{D}_{\ell-1}^n} \frac{1}{L} \sum_{j \in \Omega} q_j^2 = \frac{1}{L} \sum_{a \in \Omega} q_a^2 \), hence this quadratic term also cancels out. In the light of these observations, we obtain

\[
\text{var}(x|_{\mathcal{A}_n}) - \text{var}(x|_{\mathcal{B}_n}) = \frac{2}{|\mathcal{D}_{\ell-1}^n|} \sum_{\Gamma \in \mathcal{D}_{\ell-1}^n} q_r \left( \frac{1}{\ell-1} \sum_{i \in \Gamma} q_i - \frac{1}{L} \sum_{j \in \Omega} q_j \right).
\]

Here we substitute again \( q_r \) for \( \sum_{i \in \Gamma} q_i \) and recall \( \frac{1}{L} \sum_{j \in \Omega} q_j = E_q = 0 \). Thus, we have

\[
\text{var}(x|_{\mathcal{A}_n}) - \text{var}(x|_{\mathcal{B}_n}) = \frac{2}{(\ell-1)|\mathcal{D}_{\ell-1}^n|} \sum_{\Gamma \in \mathcal{D}_{\ell-1}^n} q_r^2 \geq 0.
\]

In order to use this result for the proof of the theorem, we make the following observations: Denote \( v_n = \text{var}(x|_{\mathcal{C}_n^\ell}) \) and \( s_n = |\mathcal{C}_n^\ell| \). By virtue of the decomposition A-2, \( \text{var}(x|_{\mathcal{A}_n}) \) can be written as \( \text{var}(x|_{\mathcal{A}_n}) = \sum_{i=0}^n \ell \cdot s_{\ell-i} v_{\ell-i} \) (see Proposition 5.1). Similarly, we have \( \text{var}(x|_{\mathcal{B}_n}) = \sum_{i=0}^n (\ell - i) \cdot s_{\ell-i} v_{\ell-i} \). We compute the coefficients of \( v_i \) in the expression

\[
\text{var}(x|_{\mathcal{A}_n}) - \text{var}(x|_{\mathcal{B}_n}) = \sum_{i=0}^n \ell \cdot s_{\ell-i} v_{\ell-i} - \sum_{i=0}^n (\ell - i) \cdot s_{\ell-i} v_{\ell-i}.
\]

For any \( 0 \leq k \leq n \), the coefficient of \( v_{\ell-k} \) is

\[
s_{\ell-k} \left( k \sum_{i=1}^n (\ell - i) \cdot s_{\ell-i} - (\ell - k) \sum_{i=1}^n i \cdot s_{\ell-i} \right) = \ell s_{\ell-k} \sum_{i=0}^n (k - i) s_{\ell-i}.
\]
We denote $\alpha_{\ell-k} = \ell \sum_{i=0}^{n} (k - i) s_{\ell-i}$, for $1 \leq k \leq n$, in order to write the above difference as

$$0 \leq \text{var}(x|A_n) - \text{var}(x|B_n) = \sum_{k=0}^{n} \alpha_{\ell-k} s_{\ell-k} v_{\ell-k}. \quad (A-4)$$

The coefficients in this expression have two following properties:

1. $\sum_{k=0}^{n} s_{\ell-k} \alpha_{\ell-k} = 0$.
2. $\forall j, \alpha_{j-1} - \alpha_j = \ell \sum_{i=0}^{n} s_{\ell-i}$.

To show the first equality, we consider the sum in (1) as the linear combination of the elements $s_{\ell-i}s_{\ell-j}$, $i, j = 0, ..., n$. The coefficient of $s_{\ell-i}s_{\ell-j}$ is zero for any $i$. For any $i \neq j$, $s_{\ell-i}s_{\ell-j}$ appears just in two components of the sum above, namely, $s_{\ell-i}\alpha_{\ell-i}$ and $s_{\ell-j}\alpha_{\ell-j}$. Specifically, $\alpha_{\ell-i}$ contains the summand $\ell(i - j) s_{\ell-j}$, and $\alpha_{\ell-j}$ contains the summand $\ell(j - i) s_{\ell-i}$, therefore in the sum $s_{\ell-i}\alpha_{\ell-i} + s_{\ell-j}\alpha_{\ell-j}$ the coefficient of $s_{\ell-i}s_{\ell-j}$ is zero. The second property follows from the definition of $\alpha_i$. In the light of the first property, A-4 can be written as

$$\left( \sum_{k=1}^{n} \alpha_{\ell-k} s_{\ell-k} \right) v_{\ell} \leq \sum_{k=1}^{n} \alpha_{\ell-k} s_{\ell-k} v_{\ell-k}. \quad (A-5)$$

Equipped with these observations, we prove, by induction on $n$, the inequality

$$\text{var}(x|D_0^n) \leq \text{var}(x|D^n_\ell),$$

for any $n = 1, ..., \ell - 1$. The theorem follows for $n = \ell - 1$. By Proposition 5.1, $\text{var}(x|D_0^n) = \frac{\sum_{i=0}^{n} s_{\ell-i} v_{\ell-i}}{\sum_{i=0}^{n} s_{\ell-i}}$, and $\text{var}(x|D^n_\ell)$ is just $v_{\ell}$. Thus we need to prove

$$v_{\ell} \leq \frac{\sum_{i=0}^{n} s_{\ell-i} v_{\ell-i}}{\sum_{i=0}^{n} s_{\ell-i}},$$
or
\[
\left( \sum_{i=1}^{n} s_{\ell-i} \right) v_\ell \leq \sum_{i=1}^{n} s_{\ell-i} v_{\ell-1}. \tag{A-6}
\]

For \( n = 1 \), A-5 reads as
\[
\alpha_{\ell-1} s_{\ell-1} v_\ell \leq \alpha_{\ell-1} s_{\ell-1} v_{\ell-1}.
\]

Here \( \alpha_{\ell-1} = \ell s_\ell > 0 \), thus we obtain the inequality \( v_\ell \leq v_{\ell-1} \). It implies
\[
(s_\ell + s_{\ell-1}) v_\ell \leq s_\ell v_\ell + s_{\ell-1} v_{\ell-1},
\]
as required. Now, we assume by induction that inequality A-6 holds up to \( n - 1 \) and prove for \( n \). We use (A-5):
\[
(E1) : \left( \sum_{k=1}^{n} \alpha_{\ell-k} s_{\ell-k} \right) v_\ell \leq \sum_{k=1}^{n} \alpha_{\ell-k} s_{\ell-k} v_{\ell-k}.
\]

This inequality undergoes a series of transformations designed to bring it to the form of A-6.

First, we have \( \alpha_{\ell-1} < \alpha_{\ell-2} \). Since \( v_\ell \leq v_{\ell-1} \) by the proof for \( n = 1 \), we have an inequality
\[
(d1) : (\alpha_{\ell-2} - \alpha_{\ell-1}) s_{\ell-1} v_\ell \leq (\alpha_{\ell-2} - \alpha_{\ell-1}) s_{\ell-1} v_{\ell-1}
\]
Adding (d1) to the inequality (E1), we arrive at
\[
(E2) : \left( \alpha_{\ell-2} (s_{\ell-1} + s_{\ell-2}) + \sum_{k=3}^{n} \alpha_{\ell-k} s_{\ell-k} \right) v_\ell \leq \alpha_{\ell-2} (s_{\ell-1} v_{\ell-1} + s_{\ell-2} v_{\ell-2}) + \sum_{k=3}^{n} \alpha_{\ell-k} s_{\ell-k} v_{\ell-k}.
\]

Second, by induction assumption for \( n = 2 \) we have the inequality
\[
(s_{\ell-1} + s_{\ell-2}) v_\ell \leq s_{\ell-1} v_{\ell-1} + s_{\ell-2} v_{\ell-2}.
\]
Also, $\alpha_{\ell-2} \leq \alpha_{\ell-3}$ as noticed earlier. Then we can construct the next inequality in order to add it to $(E2)$:

$$(d1) : (\alpha_{\ell-3} - \alpha_{\ell-2})(s_{\ell-1} + s_{\ell-2})v_\ell \leq (\alpha_{\ell-3} - \alpha_{\ell-2})(s_{\ell-1}v_{\ell-1} + s_{\ell-2}v_{\ell-2})$$

This results in the following expression:

$$(E3) : \left( \alpha_{\ell-3} \sum_{i=1}^{3} s_{\ell-i} + \sum_{k=4}^{n} \alpha_{\ell-k}s_{\ell-k} \right) v_\ell \leq \alpha_{\ell-3} \sum_{i=1}^{3} (s_{\ell-i}v_{\ell-i}) + \sum_{k=4}^{n} \alpha_{\ell-k}s_{\ell-k}v_{\ell-k}.$$ 

In this fashion we make $n - 1$ steps resulting in the inequality

$$(E(n - 1)) : (\alpha_{\ell-n} \sum_{i=1}^{n} s_{\ell-i})v_\ell \leq \alpha_{\ell-n} \sum_{i=1}^{n} s_{\ell-i}v_{\ell-i}$$

Notice that $\alpha_{\ell-n}$ is positive: $\alpha_{\ell-n} = s_{\ell-n} \ell(ns_{\ell} + (n-1)s_{\ell-1} + ... + s_{\ell-n+1}).$

Thus, we obtain the desired result. As mentioned, the theorem follows for $n = \ell - 1$. □

Appendix B

We provide an empirical evidence to the claims (3.2), (3.5), relating to random variables $x_\ell$ and $y_\ell$ based on the Capacity sets. The first claim, corresponding to the $q$-based variables, was proven in the Appendix A, and the second one relates to quite similar situation with Capacity Matrix. Here we provide statistical data displaying the following inequality:

$$\text{var}(x_\ell) \leq \text{var}(y_\ell) \quad (B-1)$$

To that end, we approximate the variance of both variables $x_\ell$, $y_\ell$ by sampling a large number of supports from domains of both variables.
The computation is carried out both for the case of Capacity vector (claim 3.2) and the case of Capacity Matrix (claim 3.5). We use D – Random dictionary of size $32 \times 64$ and compute the results for support sizes up to half os the column dimension of the dictionary. The results are displayed in figure 5.

![Comparison of variances for Capacity Vector](image1)

![Comparison of variances for Capacity Matrix](image2)

Figure 5 : The graph of variances for random variables $x_\ell, y_\ell$.

References

[1] S.S. Chen, Basis pursuit, Ph.D. dissertation, Stanford Univ., Stanford, CA, Nov. 1995
[2] S. S. Chen, D. L. Donoho, and M. A. Saunders, Atomic decomposition by basis pursuit, *SIAM J. Scientific Computing*, vol. 20, no. 1, pp. 3361, 1999.
[3] E.J. Cand`es, J. Romberg and T. Tao., Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. on Information Theory*, vol. 52(2), pp. 489-509, 2006.
[4] E. Cand`es, J. Romberg., Quantitative robust uncertainty principles and optimally sparse decompositions, *Foundations of Computational Mathematics*, Vol. 6, Issue 2, April 2006.
[5] Donoho, D.L. and Elad, M., Optimally sparse representation in general (non-orthogonal) dictionaries via $l^1$ minimization, *Proceedings of the National Academy of Sciences*, vol 100(5), pp. 2197-2202, 2003.
[6] Donoho D.L., For most large underdetermined systems of linear equations the minimal $l^1$-norm solution is also the sparsest solution, Technical Report n.2004-10, July 2004.
[7] M. Elad and A.M. Bruckstein, A generalized uncertainty principle and sparse representations in pairs of bases, *IEEE Trans. Info. Thry*, vol. 49,
pp. 2558-2567, 2002

[8] D.M. Malioutov, M. Cetin, and A.S. Willsky, Optimal sparse representations in general overcomplete bases, *IEEE International Conference on Acoustics, Speech, and Signal Processing - ICASSP*, May 2004, Montreal, Canada.

[9] T. Strohmer and R. W. Heath, Jr., Grassmannian Frames with applications to coding and communications, *Applied and Computational Harmonic Analysis*, vol. 14, Issue 3, pp. 257-275, May 2003.

[10] T. Tao, An uncertainty principle for cyclic groups of prime order, *Math. Res. Lett.*, vol. 12, no. 1, pp 121-127, 2005.

[11] J. Tropp, Greed is good: algorithmic results for sparse approximation, *IEEE Trans. Inform. Theory*, vol. 50, no. 10, pp. 2231-2242, October 2004.

[12] B.K. Natarajan, Sparse approximate solutions to linear systems, *SIAM Journal on Computing*, 24:227–234, 1995.

[13] R. Gribonval and M. Nielsen, Sparse decompositions in unions of bases, *IEEE Trans. on Information Theory*, vol 49(12), pp. 3320-3325, 2003.

[14] J.J. Fuchs, On sparse representations in arbitrary redundant bases, *IEEE Trans. on Information Theory*, vol. 50(6), pp. 1341-1344, 2004.

[15] D.L. Donoho and X. Huo, Uncertainty principles and ideal atomic decomposition, *IEEE Trans. On Information Theory*, 47(7):2845–2862, 1999.

[16] G. H. Hardy, J.E. Littlewood, and G. Pólya, *Inequalities*, 2nd ed. Cambridge, England: Cambridge University Press, pp. 43–45 and 123, 1988.