Sparse & Redundant Representations and Their Applications in Signal and Image Processing

Theoretical Analysis of the Two-Ortho Case

Michael Elad
The Computer Science Department
The Technion – Israel Institute of technology
Haifa 32000, Israel
The Two-Ortho Case
Our Starting Point

This is our main goal (for now):

\[
(P_0) \quad \min_{x} \|x\|_0 \quad \text{s.t.} \quad Ax = b
\]

We are well-aware of its flaws:

- The equality $Ax=b$ is too demanding, and slight variations in $A$ or $b$ should be permitted
- $L_0$ is too strict – negligible non-zeros are counted as part of the support

WE WILL TREAT THESE MATTERS AT A LATER STAGE
Our Goals

Here is What we Know

- There is a unique solution to this problem &
- This solution is easily computed (closed-form)

Our Questions:

- Is there a unique solution? Under which conditions?
- Given a candidate solution, could we test its optimality easily?
- How can we get this solution in reasonable time?

\[(P_2) \quad \min_x \|x\|_2^2 \quad \text{s.t.} \quad Ax = b\]

\[(P_0) \quad \min_x \|x\|_0 \quad \text{s.t.} \quad Ax = b\]
The Two-Ortho Case

\[
(P_0) \min_{x} \|x\|_0 \quad \text{s.t.} \quad Ax = b
\]

As our goals seem daunting, let's start with a special case

**The Two-Ortho Case:**

\[
A = \begin{bmatrix} \Psi, \Phi \end{bmatrix}
\]

Orthogonal Matrices
Recall Orthogonal Matrices

- A square matrix is called Orthogonal (unitary) if it satisfies the conditions:
  \[ \Psi^T \Psi = \Psi \Psi^T = I \]

- This means that its columns (or rows) are orthogonal to each other and L₂-normalized.

- Implications:
  - Given the linear system, \( \Psi x = b \), its solution is readily given by \( x = \Psi^T b \).
  - Parseval: \[ \| \Psi x \|_2^2 = x^T \Psi^T \Psi x = x^T x = \| x \|_2^2 \]
  - Implications: Given the linear system, \( \Psi x = b \), its solution is readily given by \( x = \Psi^T b \).
  - Parseval: \[ \| \Psi x \|_2^2 = x^T \Psi^T \Psi x = x^T x = \| x \|_2^2 \]
Why Two-Orhto Interesting?

- Consider the case: $\mathbf{A} = [\mathbf{I}, \mathbf{F}]$ – Identity & Fourier
- Given a signal $\mathbf{b}$ with few harmonies and few spikes, it gets a dense description in each of these bases
- When solving $(P_0)$ with the two-ortho matrix $\mathbf{A}$, the solution becomes sparse

$$(P_0) \quad \min_x \|x\|_0$$

s.t. $[\mathbf{I}, \mathbf{F}] x = \mathbf{b}$

These four atoms are merged to create the above signal
OK. What Are We After?

So, we are considering for now the two-ortho case \((P_0)\) problem

\[
(P_0) \min_x \|x\|_0 \quad \text{s.t.} \quad \begin{bmatrix} \Psi & \Phi \end{bmatrix} x = b
\]

and would like to characterize somehow the solutions of this problem and their uniqueness
Mutual Coherence

- Just before we proceed, we need to quantify how different $\Psi$ and $\Phi$ are.

**Definition:** For an arbitrary pair of orthogonal matrices of size $n \times n$, $\Psi$ and $\Phi$, the mutual-coherence is the maximal absolute inner-product between their columns

$$\mu(\Psi, \Phi) = \max_{1 \leq i, j \leq n} \left| \frac{\Psi_i^T \Phi_j}{\max_{1 \leq i \leq n} \| \Psi_i \|_{\infty}} \right|$$

- The higher $\mu$ is, the closer these two matrices are. The maximal value ($\mu=1$) is obtained when at least one column is the same in the two matrices.
Mutual Coherence

\[ \mu(\Psi, \Phi) = \max_{1 \leq i, j \leq n} \left| \psi_i^T \phi_j \right| = \max_{1 \leq i, j \leq n} \left| \Psi^T \Phi \right|_{i,j} \]

- How small can \( \mu \) be?
  
  **Claim:** \( \frac{1}{\sqrt{n}} \leq \mu \leq 1 \)

- Proof:
  
  - \( \Psi^T \Phi \) is an orthogonal matrix of size \( n \times n \) (why?)
  
  - Assume that \( \mu < 1/\sqrt{n} \)
  
  - Thus, all \( |\Psi^T \Phi| \)'s elements are strictly smaller than \( 1/\sqrt{n} \)
  
  - Contradiction: the column-norms are smaller than 1
Mutual Coherence

- Two orthogonal matrices are maximally incoherent if
  \[ \mu(\Psi, \Phi) = \frac{1}{\sqrt{n}} \]

- Examples: Identity and Fourier (or Hadamard)

- In these cases: \( \forall 1 \leq i, j \leq n, |\Psi_i^T \Phi_j| = \frac{1}{\sqrt{n}} \)

This is a constant matrix
\[ |\Psi^T \Phi| = \frac{1}{\sqrt{n}} \mathbf{1} \]
Sparse & Redundant Representations and Their Applications in Signal and Image Processing

Theoretical Analysis of the Two-Ortho Case

Michael Elad
The Computer Science Department
The Technion – Israel Institute of technology
Haifa 32000, Israel
An Uncertainty Principle
Defining Our Goal

Armed with the mutual-coherence \( \mu(\Psi, \Phi) \)
we return to our prime goal:

Characterize the solutions of the problem \((P_0)\)

\[
(P_0) \quad \min_x \|x\|_0 \quad \text{s.t.} \quad \begin{bmatrix} \Psi & \Phi \end{bmatrix} x = b
\]

and their Uniqueness
A (Small) Diversion in Our Story

We are given a signal

\[ \mathbf{b} = \mathbf{b}_0 \]

\( \mathbf{b} \in \mathbb{R}^n \)
\( \mathbf{b} \neq 0 \)

Represent with \( \Psi \)

\[ \mathbf{b} = \Psi \alpha \]

Represent with \( \Phi \)

\[ \mathbf{b} = \Phi \beta \]

Clearly, \( \alpha \) & \( \beta \) are uniquely defined by the relations

\[ \alpha = \Psi^T \mathbf{b} \]
\[ \beta = \Phi^T \mathbf{b} \]

Assume \( \|\mathbf{b}\|_2 = 1 \) for simplicity

Question: Could \( \alpha \) & \( \beta \) be arbitrarily jointly sparse?

Answer: No, there is an ‘uncertainty’ law

\[ \|\alpha\|_0 + \|\beta\|_0 \geq T \]
This relationship looks familiar – it should remind you of the Heisenberg's uncertainty theorem.

Both statements tie the supports of the representations in two domains (e.g., time and frequency).

The new relation thinks “discretely” in terms of non-zeros, regardless of their locations.
Theorem: Given two orthogonal matrices of size $n \times n$, $\Psi$ and $\Phi$, with mutual-coherence $\mu$, and given a non-trivial $b \in \mathbb{R}^n$, its representations, $b = \Psi \alpha = \Phi \beta$, cannot be jointly sparse:

$$\|\alpha\|_0 + \|\beta\|_0 \geq \frac{2}{\mu}$$

Comment: In the trivial case in which $\mu = 1$, the above inequality is meaningless, since the bound is 2, and since $b \neq 0$, the number of non-zeros in $\alpha$ & $\beta$ is at least 1 in each.
Uncertainty Law #1: Proof

- Since \( b = \Psi \alpha = \Phi \beta \neq 0 \), assume w.l.o.g.
  \[
  \|b\|_2 = \|\alpha\|_2 = \|\beta\|_2 = 1
  \]

- Let's describe \( b \) as a weighted sum of the columns of \( \Psi \) over \( S_\alpha \) - the support of \( \alpha \):
  \[
  b = \Psi \alpha = \sum_{i \in S_\alpha} \alpha_i \psi_i
  \]

- This leads to \( (\beta = \Phi^T b) \):
  \[
  |\beta_j|^2 = |\Phi^T b\|_2 = \left| \sum_{i \in S_\alpha} \alpha_i \psi_i^\top \phi_j \right|^2
  \]
Uncertainty Law #1: Proof

- So far we got that
  \[ |\beta_j|^2 = |b^T \phi_j|^2 = \left| \sum_{i \in S_a} \alpha_i \psi_i^T \phi_j \right|^2 \]

- Using the Cauchy-Schwarz Inequality
  \[ \left| x^T y \right|^2 \leq \|x\|_2^2 \cdot \|y\|_2^2 \]
  \[ \leq \left( \sum_{i \in S_a} \alpha_i^2 \right) \cdot \left( \sum_{i \in S_a} (\psi_i^T \phi_j)^2 \right) \leq \|\alpha\|_2^2 \cdot \|\alpha\|_0 \cdot \mu^2 \]
  \[ = \|\alpha\|_0 \cdot \mu^2 \]
Uncertainty Law #1: Proof

- We got the inequality
  \[ |\beta_j|^2 \leq \|\alpha\|_0 \cdot \mu^2 \]

- Summing this over the support of \( \beta \), we get
  \[ 1 = \sum_{j \in S_\beta} |\beta_j|^2 \leq \sum_{j \in S_\beta} \|\alpha\|_0 \cdot \mu^2 = \|\alpha\|_0 \cdot \|\beta\|_0 \cdot \mu^2 \]

\[ 1 \leq \|\alpha\|_0 \cdot \|\beta\|_0 \cdot \mu^2 \]

This is a worthy uncertainly result by itself but we do not stop here.
Relying on the relation between geometric and arithmetic means

\[ \|\alpha\|_0 \cdot \|\beta\|_0 \geq \frac{1}{\mu^2} \]

- \[ \frac{\|\alpha\|_0 + \|\beta\|_0}{2} \geq \sqrt{\|\alpha\|_0 \cdot \|\beta\|_0} \geq \frac{1}{\mu} \]

\[ \|\alpha\|_0 + \|\beta\|_0 \geq \frac{2}{\mu} \]

Uncertainty Law #1: Proof
We already mentioned that the $[I,F]$ case is maximally incoherent, i.e.:
\[ \mu(I,F) = \frac{1}{\sqrt{n}} \]

The uncertainty theorem implies that a signal and its Fourier Transform cannot be jointly sparse

\[ b = I\alpha = F\beta \Rightarrow \|\alpha\|_0 + \|\beta\|_0 \geq \frac{2}{\mu} = 2\sqrt{n} \]

Is this inequality tight? The answer is Yes!

This is the picket-fence signal – it has $\sqrt{n}$ non-zeros ... AND ... its Fourier transform is the very same signal.
Sparse & Redundant Representations and Their Applications in Signal and Image Processing

Theoretical Analysis of the Two-Ortho Case

Michael Elad
The Computer Science Department
The Technion – Israel Institute of technology
Haifa 32000, Israel
From Uncertainty to Uniqueness
Recalling Our Goals

\[(P_0) \min_x \|x\|_0 \text{ s.t. } Ax = b\]

Our Questions:

- Is there a unique solution? Under which conditions?
- Given a candidate solution, could we test its optimality easily?
- How can we get this solution in reasonable time?
We are given the ortho-pair $\Psi$ and $\Phi$

We can compute $\mu(\Psi, \Phi)$

If $\Psi \alpha = \Phi \beta \neq 0$ then necessarily $\frac{\|\alpha\|_0 + \|\beta\|_0}{2} \geq \frac{1}{\mu}$

Assume that $b = [\Psi, \Phi]x$ is known to have two different representations (solutions) $x_1 \neq x_2$

\[ b = [\Psi, \Phi]x_1 = [\Psi, \Phi]x_2 \neq 0 \]

We aim to show that in this case another uncertainly law emerges, this time w.r.t. $x_1$ & $x_2$
Deriving a New Uncertainty Law

\[ b = \begin{bmatrix} \Psi, \Phi \end{bmatrix} x_1 = \begin{bmatrix} \Psi, \Phi \end{bmatrix} x_2 \neq 0 \]

\[ [\Psi, \Phi](x_1 - x_2) = 0 \]

\[ \Psi \alpha = -\Phi \beta \neq 0 \]

\[ \|\alpha\|_0 + \|\beta\|_0 \geq \frac{2}{\mu} \]

Invoke the uncertainly law we have just developed
Deriving a New Uncertainty Law

\[ \|\alpha\|_0 + \|\beta\|_0 \geq \frac{2}{\mu} \]

Observe that

\[ \|\alpha\|_0 + \|\beta\|_0 = \|x_1 - x_2\|_0 \]

Invoking the L_0 Triangle Inequality

\[ \|x_1 - x_2\|_0 \leq \|x_1\|_0 + \|x_2\|_0 \]

Just as promised, we got an uncertainty theorem with respect to the possible solutions of \( Ax = b \)
Theorem: Given two orthogonal matrices of size $n \times n$, $\Psi$ and $\Phi$, with mutual-coherence $\mu$, and given an arbitrary non-trivial vector $b \in \mathbb{R}^n$, any two different representations of $b$ w.r.t. $[\Psi, \Phi]$ cannot be jointly too sparse, namely

$$\|x_1\|_0 + \|x_2\|_0 \geq \frac{2}{\mu}$$

Proof? It is given in the two previous slides.
From Uncertainty to Uniqueness

- We have seen that two different representations cannot be jointly sparse \( \vec{x}_1 \neq \vec{x}_2 \)
  
  \[
  \vec{b} = \begin{bmatrix} \Psi, & \Phi \end{bmatrix} \vec{x}_1
  = \begin{bmatrix} \Psi, & \Phi \end{bmatrix} \vec{x}_2
  \Rightarrow \| \vec{x}_1 \|_0 + \| \vec{x}_2 \|_0 \geq \frac{2}{\mu}
  \]

- Therefore, given a candidate solution that is very sparse,
  
  \[
  \vec{b} = \begin{bmatrix} \Psi, & \Phi \end{bmatrix} \vec{x} \quad & \| \vec{x} \|_0 < \frac{1}{\mu}
  \]

  any other solution is necessarily denser, implying that this is the solution to

  \[
  (P_0) \min_{\vec{x}} \| \vec{x} \|_0 \quad \text{s.t.} \quad \begin{bmatrix} \Psi, & \Phi \end{bmatrix} \vec{x} = \vec{b}
  \]
**Theorem:** Given two orthogonal matrices of size $n \times n$, $\Psi$ and $\Phi$, with mutual-coherence $\mu$, and given an arbitrary non-trivial vector $b \in \mathbb{R}^n$, if a sparse solution $x$ is found such that

$$b = \begin{bmatrix} \Psi \end{bmatrix} x \ & \ |x|_0 < \frac{1}{\mu}$$

then it is necessarily the sparsest possible solution, i.e., the globally optimal solution to the $(P_0)$ problem

$$(P_0) \min_x |x|_0 \ \text{s.t.} \ \begin{bmatrix} \Psi \end{bmatrix} x = b$$
Uniqueness: Implications

- At least for the case where $A = [\Psi, \Phi]$, we now possess a very strong result:

  Given a sparse solution to $Ax = b$, we can check easily for its global optimality for $(P_0)$ by simply counting the number of non-zeros in $x$

  If this count is below $1/\mu$ - this is the best possible (sparsest) solution

- Weaknesses in this statement:
  - If a solution $x$ is given and $\|x\|_0 > 1/\mu$, nothing can be claimed
  - Must there exist a solution with $< 1/\mu$ non-zeros? No
  - The above is limited to the two-ortho case
Our Answers for Now

\[(P_0) \min_x \|x\|_0 \text{ s.t. } [\Psi, \Phi]x = b\]

Our Questions:
- Is there a unique solution? Under which conditions?
- Given a candidate solution, can we test its optimality easily?
- How can we get this solution in reasonable time?

Our Answers:
- If a sparse enough solution found – it is unique
- Simply count the number of non-zeros – if it is below \(1/\mu\) you get uniqueness
- Wait and see ...
Sparse & Redundant Representations and Their Applications in Signal and Image Processing

Theoretical Analysis of the General Case

Michael Elad
The Computer Science Department
The Technion – Israel Institute of technology
Haifa 32000, Israel
Introducing the Spark
We are considering again the general \((P_0)\) problem

\[
(P_0) \quad \min_{x} \|x\|_0 \quad \text{s.t.} \quad Ax = b
\]

and would like to characterize somehow the solutions of this problem and their uniqueness.
Lessons from the Two-Ortho Case?

- Could we leverage the treatment of the two-ortho case in order to suggest a similar treatment for the case of a general matrix $A$?

- The answer seems to be negative. We cannot split $A$ as in the two-ortho case, the coherence is tightly related to this specific structure.

- However – we are inspired by the ability to make exact claims about uniqueness, and we'll seek an alternative route ...

which brings us to the definition of the Spark
The Spark

- We turn to define a property of $A$ that will be found useful for the uniqueness analysis – the Spark:

**Definition:** For a matrix $A$ of size $n \times m$ we define its Spark as the smallest number of its columns that are linearly dependent.

- Example:

$$A = \begin{bmatrix}
1 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 1 \\
0 & 0 & 1 & 0 & 1 \\
0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 & 0
\end{bmatrix} \quad \Rightarrow \quad \text{Spark} \{A\} = 3$$

By the way, what is the rank of this matrix? $\text{Rank}\{A\} = 4$
We turn to define a property of $A$ that will be found useful for the uniqueness analysis – the Spark:

**Definition:** For a matrix $A$ of size $n \times m$ we define its Spark as the smallest number of its columns that are linearly dependent.

Contrast this with this classic definition of Rank:

**Definition:** For a matrix $A$ of size $n \times m$ we define its rank as the largest number of its columns that are linearly independent.

While the two are so similar, there is a marked difference between them, as Rank is MUCH easier to compute.
Computing the Spark

- Given a matrix $A$, and assuming that $\text{Spark}\{A\}=k$
  this means that
  - There is at least one set of $k$ columns in $A$ that are linearly dependent, &
  - There is no subset of $k-1$ columns that is linearly dependent

- ... Here is an “algorithm” for computing $\text{Spark}\{A\}$:

  1. Set $k=1$
  2. Sweep through all subsets of $k$ columns
  3. Check for Linear Dependency
     - Yes
     - No

   - Possibilities $\binom{m}{k}$
   - Set $k=k+1$
   - Done
The Spark – An Exponential Animal

- Computing the Spark is of exponential complexity w.r.t. m
- This should be contrasted with the polynomial complexity of computing a Rank of a matrix
- We will come back to this matter later on, seeking easier-to-compute alternatives

```
Set k=1
Sweep through all subsets of k columns
Check for Linear Dependency
```

- Set k=k+1
- (m \choose k) Possibilities!!
- Yes
- Done
**Spark – A Key Properties**

**Property 1:** Spark bounds are given by

\[ 1 \leq \text{Spark}\{A\} \leq n+1 \]

corresponding to the case of a zero column in \( A \). If all columns of \( A \) are non-zero, the minimal value of the Spark is 2.

corresponding to the case where every subset of \( n \) columns in \( A \) is linearly independent (e.g. a random matrix or Vandermonde).
Property 2: Relation to Sparsity?

- Assume that we are given a non-trivial solution to the homogeneous system $Ax=0$

- Given the Spark, we can immediately say that

\[ \text{Spark } \{ A \} = k \Rightarrow \|x\|_0 \geq k \]

Explanation: $x$ combines columns from $A$ to create the 0 vector, and thus at least $k$ such columns are needed in this combination.
Thus, if $\mathbf{A}\mathbf{x} = 0$ and $\text{Spark}\{\mathbf{A}\} = k$ then necessarily

$$\|\mathbf{x}\|_0 \geq k$$

This suggests an alternative definition to the Spark, which is more in-line with our objectives.

**Definition:** For a matrix $\mathbf{A}$ of size $n \times m$ we define its Spark as the number of non-zeros in the sparsest possible non-trivial solution to the homogeneous system $\mathbf{A}\mathbf{x} = 0$. 
Little Bit of History

- We (Donoho & Elad) coined the term Spark in 2001, in the context of studying the \((P_0)\) problem.
- Later we found out that this definition has already appeared in the literature in the late 80’s under the name **Kruskal’s-Rank**, and used for high-dimension SVD.
- Gorodnitsky and Rao in 1997 studied the case of full-Spark (=n+1) in order to characterize uniqueness of sparse solutions of linear systems.
- The notion of Spark is well known in its discrete finite-alphabet setting in coding theory, termed “code distance”.
Sparse & Redundant Representations and Their Applications in Signal and Image Processing

Theoretical Analysis of the General Case

Michael Elad
The Computer Science Department
The Technion – Israel Institute of technology
Haifa 32000, Israel
Uniqueness for the General Case via the Spark
Uncertainty via the Spark

Given a system of linear equations $Ax = b$, where $A$ is given, we compute Spark($\{A\}$) = $k$. Now consider the linear system $Ax = b$. We are given two candidate solutions $x_1 \neq x_2$.

We subtract the two solutions:

$$A(x_1 - x_2) = 0 \quad \text{and} \quad (x_1 - x_2) \neq 0$$

By Spark Properties,

$$\|x_1 - x_2\|_0 \geq k$$

By Triangle Inequality,

$$\|x_1\|_0 + \|x_2\|_0 \geq \|x_1 - x_2\|_0 \geq k$$

Michael Elad | The Computer-Science Department | The Technion
Uncertainty via the Spark

We got an uncertainty law for general solutions of linear systems, claiming that

**Theorem:** For a matrix $A$ with $\text{Spark}\{A\} = k$, any two solutions $x_1 \neq x_2$ of the system $Ax = b$ ($b \neq 0$) cannot be jointly too sparse:

$$Ax_1 = b \quad \& \quad Ax_2 = b$$

and $x_1 \neq x_2$

Proof?

It is given in the two previous slides

$$\|x_1\|_0 + \|x_2\|_0 \geq \text{Spark}\{A\}$$
Uncertainty Laws Compared

Two-Ortho Case
\[ b = [\Psi, \Phi] x_1 = [\Psi, \Phi] x_2 \]
\[ \|x_1\|_0 + \|x_2\|_0 \geq \frac{2}{\mu} \]

General Case
\[ A x_1 = b \quad \& \quad A x_2 = b \]
\[ \|x_1\|_0 + \|x_2\|_0 \geq \text{Spark}\{A\} \]

Comparison:
- Both have a very similar form of uncertainty law
- Both rely on a basic property of \( A \) (Mutual-Coherence or Spark)
- While Coherence is easy to compute, Spark is impossible to get
- Still, Spark is stronger than coherence: E.g., for an \( n \times 2n \) two-ortho case, the best bound would be \( 2\sqrt{n} \) whereas the Spark bound could lead to \( n+1 \)
Uniqueness via the Spark

**Theorem:** Given a matrix $A$ of size $n \times m$ with $\text{Spark}\{A\} = k$, and given an arbitrary non-trivial $b \in \mathbb{R}^n$, if a sparse solution $x$ is found such that

$$b = Ax$$

and

$$\|x\|_0 < \frac{\text{Spark}\{A\}}{2}$$

then it is necessarily the sparsest possible solution, i.e., the globally optimal solution to the $(P_0)$ problem

$$(P_0) \min_x \|x\|_0 \quad \text{s.t.} \quad Ax = b$$
Uniqueness via the Spark: Proof

- Based on the uncertainty law we have just seen, for any two different solutions of $Ax = b$ we get:
  \[ \|x_1\|_0 + \|x_2\|_0 \geq \text{Spark}\{A\} \]

- Therefore, given a candidate solution that is very sparse,
  \[ b = Ax \quad \& \quad \|x\|_0 < \frac{\text{Spark}\{A\}}{2} \]

any other solution is necessarily denser, implying that this is necessarily the unique solution to

\[ (P_0) \quad \min_x \|x\|_0 \quad \text{s.t. } Ax = b \]
Uniqueness: Implications

- For the most general case of $A$, we now possess a very strong result:
  
  Given a sparse solution to $Ax = b$, we can check easily for its global optimality for $(P_0)$ by simply counting the number of non-zeros in $x$.
  
  If this count is below $0.5 \cdot \text{Spark}\{A\}$ – this is the best possible (sparsest) solution.

- Weaknesses in this statement:
  - What happens above the bound? Nothing can be claimed.
  - Must there exist a solution with $< 0.5 \cdot \text{Spark}\{A\}$ non-zeros? No.
  - **The bound is impossible to compute.**
Our Answers – An Update

\[ (P_0) \quad \min_x \|x\|_0 \quad \text{s.t.} \quad Ax = b \]

Our Questions:

- Is there a unique solution? Under which conditions?
- Given a candidate solution, can we test its optimality easily?
- How can we get this solution in reasonable time?

Our Answers:

- If a sparse enough solution is found – it is unique
- Simply count the number of non-zeros – if it is below \( \text{Spark}\{A\}/2 \) - you get uniqueness
- Wait and see …
Sparse & Redundant Representations and Their Applications in Signal and Image Processing

Theoretical Analysis of the General Case

Michael Elad
The Computer Science Department
The Technion – Israel Institute of technology
Haifa 32000, Israel
Uniqueness via the Mutual-Coherence
Should we be Pleased?

- We have seen that the Spark provides a productive avenue towards studying the uniqueness property of (P₀)'s solutions.

- On the down side, Spark of A is impossible to compute for any reasonable matrix size.

- Is there an alternative? More specifically: Is there a way that generalizes the Mutual-Coherence that was defined for the two-ortho case?

- We now show that such an option exists.
Recall the Mutual Coherence

- Recall the definition: For the pair $\Psi$ and $\Phi$, the mutual-coherence is

$$\mu(\Psi, \Phi) = \max_{1 \leq i, j \leq n} |\psi_i^T \phi_j| = \max_{1 \leq i, j \leq n} |\Psi^T \Phi|_{i,j}$$

- In words: In this definition we are seeking for the two “closest” atoms in $\Psi$ and $\Phi$. The proximity is measured by the absolute inner-product.
Another View of the Mutual Coherence

\[ \mu(\Psi, \Phi) = \max_{1 \leq i, j \leq n} |\psi_i^T \phi_j| = \max_{1 \leq i, j \leq n} |\Psi^T \Phi|_{i,j} \]

- Here is another way to describe this:

\[ A^T A = \begin{bmatrix} \Psi^T & \Phi \\ \Phi^T \end{bmatrix} = \begin{bmatrix} I & \Psi \Phi \\ \Phi^T \Psi & I \end{bmatrix} \]

\[ \mu(A) = \max_{1 \leq i \neq j \leq 2n} |A^T A|_{i,j} \]

How about using this as a definition of mutual coherence for general matrices \(A\)?

This definition is oblivious of the inner structure of \(A\).
This leads us naturally to the following general definition of the mutual coherence:

**Definition**: For an arbitrary matrix $A$ of size $n \times m$, the **mutual-coherence** is the maximal absolute inner-product between its normalized columns

$$\mu(A) = \max_{1 \leq i \neq j \leq m} \left| \frac{a_i^T a_j}{\|a_i\|_2 \|a_j\|_2} \right|$$

- If $A$'s columns are normalized: $\mu(A) = \max_{1 \leq i \neq j \leq m} \left| a_i^T a_j \right|$
The Gram Matrix

- Assume that $\mathbf{A}$ has normalized columns

- Compute the Gram matrix $\mathbf{G} = \mathbf{A}^T \mathbf{A}$

- Observations:
  - $\mathbf{G}$ is symmetric
  - $\mathbf{G}$'s main diagonal contains ‘1’-es (due to normalization)
  - The mutual-coherence is the largest (absolute) value in the lower/upper triangular part of $\mathbf{G}$
Definition: For an arbitrary matrix $A$ of size $n \times m$, with normalized columns, the mutual-coherence is the maximal absolute element in the lower-triangle of the Gram matrix $G$:

$$
\mu(A) = \max_{1 \leq j < i \leq m} |G_{i,j}|
$$
**Lemma:** For an arbitrary matrix $A$ the following relation holds:

$$\text{Spark} \{A\} \geq 1 + \frac{1}{\mu(A)}$$

**Theorem:**

(a) Any two solutions of $Ax = b$ must satisfy

$$\|x_1\|_0 + \|x_2\|_0 \geq \text{Spark} \{A\} \geq 1 + \frac{1}{\mu}$$

(b) If a candidate solution satisfies

$$\|x\|_0 < \frac{1}{2} \left(1 + \frac{1}{\mu}\right) \leq \frac{1}{2} \text{Spark} \{A\}$$

it is the globally optimal solution of $(P_0)$.
Sparse & Redundant Representations and Their Applications in Signal and Image Processing

Theoretical Analysis of the General Case

Michael Elad
The Computer Science Department
The Technion – Israel Institute of technology
Haifa 32000, Israel
Spark-Coherence Relation: A Proof
Just Before we Turn to the Proof ...

Gershgorin’s disk Theorem:

Consider a general square matrix $A$, and define the $k$-th disk as

$$
\text{disk}_k = \left\{ x \mid x - a_{kk} \leq \sum_{j \neq k} |a_{kj}| \right\}
$$
Geshgorin’s Disk Theorem

**Theorem**: The eigenvalues of $A$ lie in the union of all the $n$ disks

$$\{\lambda_1, \lambda_2, \ldots, \lambda_n\} \subseteq \bigcup_{k=1}^{n} \text{disk}_k$$
**Theorem:** A diagonally dominant symmetric matrix $A$ is necessarily positive definite.

$$\forall 1 \leq k \leq n \quad a_{kk} > \sum_{j \neq k} |a_{kj}|$$
Relation to the Spark: Proof

- Take an arbitrary set of $S$ atoms from $A$ and form the matrix $A_S$
- Compute the Gram for $A_S$
  \[
  G_S = A_S^T A_S = \begin{bmatrix}
    S & S \\
    S & S \\
  \end{bmatrix}
  \]
- If $G_S$ is Positive-Definite, clearly this implies that these $S$ atoms are linearly independent
- If this is true for every set of $S$ atoms, we can clearly say that $\text{Spark}\{A\} > S$
Relation to the Spark: Proof

- Let’s come up with a condition that would guarantee that $G_s$ is positive definite:
  - Require that it is Diagonally-Dominant

- We require $1 > (S-1)\mu$ since the main diagonal contains ‘1’-es and all the off-diagonals are in the range $[-\mu, \mu]$.

\[
1 > (S - 1) \mu \Rightarrow S < 1 + \frac{1}{\mu}
\]

- Every $S$ atoms are linearly independent
  - $G_s$ is diagonally dominant
  - $G_s$ is positive definite
Uncertainty Laws Compared

**Two-Ortho Case**
\[ b = \begin{bmatrix} \Psi, \Phi \end{bmatrix} x_1 = \begin{bmatrix} \Psi, \Phi \end{bmatrix} x_2 \]
\[ \|x_1\|_0 + \|x_2\|_0 \geq \frac{2}{\mu} \]

**General Case**
\[ A x_1 = b \quad \& \quad A x_2 = b \]
\[ \|x_1\|_0 + \|x_2\|_0 \geq 1 + \frac{1}{\mu} \]

**Comparison:**
- The two provide very similar forms of uncertainly laws, both relying on the mutual coherence.
- The two-ortho bound is (nearly twice) stronger and thus more informative, since it exploits the inner structure of \( A \).
Sparse & Redundant Representations and Their Applications in Signal and Image Processing

Theoretical Analysis of the General Case

Michael Elad
The Computer Science Department
The Technion – Israel Institute of technology
Haifa 32000, Israel
Uniqueness via the Babel-Function
What is Wrong with Mutual Coherence

Mutual Coherence is Too Crude !!

This can be seen in several ways:

1. \( a_i = a_j \) ? \( \mu = 1 \) and all our results collapse, regardless of the rest of the columns in \( \mathbf{A} \)
2. The same happens if there is one pair of close-by atoms
3. There is a large gap between \( 1 + 1/\mu \) and \( \text{Spark}\{\mathbf{A}\} \) (order of \( n^{0.5} \) versus \( n \))

The (partial) remedy: The Babel-Function \( \mu_1(k) \) that generalizes the Mutual Coherence
The Babel-Function: Definition

**Definition:** For an arbitrary matrix $A$ of size $n \times m$ and normalized columns, the Babel-Function is defined by

$$
\mu_1(k) = \max_{|\Lambda|=k} \left[ \max_{j \not\in \Lambda} \sum_{i \in \Lambda} |A|_{i,j} \right]
$$

**Explanation:**
- Choose an arbitrary support $\Lambda$ of cardinality $k$
- Choose $a_j$ outside this set ($j \not\in \Lambda$)
- Repeat the above for ALL possible $|\Lambda|=k$ and $j$, then take the worst possible sum – this is $\mu_1(k)$
All the above may leave the impression that computing $\mu_1(k)$ is a combinatorial process, sweeping through all possible $m$-choose-$k$ supports.

This is not true! Computing $\mu_1(k)$ is in fact easy ...

- Start by computing the Gram matrix $G = A^T A$
All the above may leave the impression that computing $\mu_1(k)$ is a combinatorial process, sweeping through all possible m-choose-k supports.

This is not true! Computing $\mu_1(k)$ is in fact easy...

- Sort each row of $|G|$ from the largest element down.
All the above may leave the impression that computing $\mu_1(k)$ is a combinatorial process, sweeping through all possible m-choose-k supports.

This is not true! Computing $\mu_1(k)$ is in fact easy ...

- Discard of the first column that contains ‘1’-es
All the above may leave the impression that computing $\mu_1(k)$ is a combinatorial process, sweeping through all possible m-choose-k supports.

This is not true! Computing $\mu_1(k)$ is in fact easy...

- Compute the accumulated values from left to right.
All the above may leave the impression that computing $\mu_1(k)$ is a combinatorial process, sweeping through all possible $m$-choose-$k$ supports.

This is not true! Computing $\mu_1(k)$ is in fact easy ...

- Find the row with the largest value in location $k$ – this is $\mu_1(k)$.
A consequence of the definition is the relation

$$\mu_1(k) = \max_{|\Lambda| = k} \left[ \max_{j \in \Lambda} \sum_{i \in \Lambda} |G_{i,j}| \right] \leq k \cdot \mu$$

obtained by exploiting the fact that for $j \neq i$, $|G_{i,j}| \leq \mu$

Observations:

- $\mu_1(1) = \mu$
- If one pair or atoms is identical, $\mu_1(1) = \mu = 1$
- $\mu_1(k)$ is monotonically-increasing with non-increasing differences

The true strength of the Babel-Function appears when $\mu_1(k)$ grows much slower compared to the linear curve $k \mu$
Using the Babel Function

- If one follows closely the proof that ties the mutual coherence to the Spark, the following observation is obtained:
  - If $\mu_1(k-1) < 1$, this implies that any set of $k$ columns from $A$ are linearly independent.
  - In this case, the Spark necessarily satisfies: $\text{Spark}\{A\} > k$.
Using the Babel Function

- Most results that use the mutual-coherence can be casted in terms of the Babel-Function.
- Here we shall demonstrate this for the uniqueness result and without a proof.

**Theorem:** If a candidate solution of $\mathbf{A}\mathbf{x} = \mathbf{b}$ satisfies

$$
\|\mathbf{x}\|_0 < \frac{1}{2} \left(1 + \frac{1}{\mu} \right)
$$

it is the globally optimal solution of $(P_0)$.

- The new condition becomes: Any solution with $k$ non-zeros is globally optimal if $\mu_1(k-1) + \mu_1(k) < 1$.
Sparse & Redundant Representations and Their Applications in Signal and Image Processing

Theoretical Analysis of the General Case

Michael Elad
The Computer Science Department
The Technion – Israel Institute of technology
Haifa 32000, Israel
Upper-Bounding the Spark
Recall: the Definition of The Spark

**Definition:** For a matrix $A$ of size $n \times m$ we define its Spark as the smallest number of its columns that are linearly dependent.

Or

**Definition:** For a matrix $A$ of size $n \times m$ we define its Spark as the sparsest possible non-trivial solution to the homogeneous system $Ax = 0$.

How can we compute the Spark?
Spark as a Solution of \((P_0)\)

- Suppose that the smallest group of linearly-dependent columns in \(A\) contains the 1\(^{\text{st}}\) column

- Then, the Spark emerges from the solution of this problem

\[
\begin{align*}
\mathbf{x}_{\text{opt}} &= \arg\min_{\mathbf{x}} \|\mathbf{x}\|_0 \quad \text{s.t.} \quad \mathbf{A}\mathbf{x} = \mathbf{0} \quad \& \quad x_1 = 1
\end{align*}
\]

- The constraint \(x_1 = 1\) comes to
  - Assure that we avoid the trivial solution \(x=0\)
  - Force the first column to be included in the solution

- If we would be able to solve this problem and get \(\mathbf{x}_{\text{opt}}\) the number of non-zeroes in it is the Spark
Spark as a Solution of $(P_0)$

- The problem with the above rationale is the need to know that the first column is involved.

- The solution: Define a set of $m$ such problems that sweep over all columns

\[
\left\{ x^{k}_{\text{opt}} = \arg \min_{x} \|x\|_0 \quad \text{s.t.} \quad Ax = 0 \quad \& \quad x_k = 1 \right\}_{k=1}^{m}
\]

and then

\[
\text{Spark}\{A\} = \min_{1 \leq k \leq m} \|x^{k}_{\text{opt}}\|_0
\]

- ... We cannot solve these set of $(P_0)$ tasks.
Upper- Bounding the Spark

True, we cannot solve this set of \((P_0)\) tasks ... but we could approximate their solutions by solving instead:

\[
\begin{align*}
\left\{ x_{opt}^k = \arg\min_{x} \|x\|_0 \text{ s.t. } A x = 0 \text{ & } x_k = 1 \right\} \quad \text{Spark} \{ A \} = \min_{1 \leq k \leq m} \|x_{opt}^k\|_0
\end{align*}
\]

Obviously, Spark \{ A \} = \min_{1 \leq k \leq m} \|x_{opt}^k\|_0 \leq \min_{1 \leq k \leq m} \|z_{opt}^k\|_0

\(z_{opt}^k\) are necessarily denser than their \(x_{opt}^k\) counterparts.
We are solving the following set of $m$ $L_1$ problems (Linear Programming) and getting $m$ solutions

$$\left\{ z_{opt}^k = \arg \min_z \|z\|_1 \text{ s.t. } Az = 0 \ & z_k = 1 \right\}^m_{k=1}$$

We evaluate the Spark by counting the number of non-zeros in the sparsest of these solutions

$$\text{Spark} \{ A \} = \min_{1 \leq k \leq m} \| x_{opt}^k \|_0 \leq \min_{1 \leq k \leq m} \| z_{opt}^k \|_0$$

This is necessarily an upper bound to the true value of the Spark
Sparse & Redundant Representations and Their Applications in Signal and Image Processing

Theoretical Analysis of the General Case

Michael Elad
The Computer Science Department
The Technion – Israel Institute of technology
Haifa 32000, Israel
Constructing Grassmanian Matrices
Recall the Definition: Mutual-Coherence

**Definition:** For a matrix $A$ of size $n \times m$, the **mutual-coherence** is the maximal absolute element in the lower-triangle of the Gram matrix $G$:

$$\mu(A) = \max_{1 \leq j < i \leq m} |G_{i,j}| = \max_{1 \leq j < i \leq m} |A^T A|_{i,j}$$
Theorem: The mutual-coherence bounds for a matrix $A$ of size $n \times m$ ($m \geq n$) are given by

$$\sqrt{\frac{m-n}{n(m-1)}} \leq \mu(A) \leq 1$$

This is known as the Welch bound: The minimal possible $\mu$, representing the optimal (maximally-spread) packing of $m$ vectors of length $n$ on the sphere (see proof next) corresponds to the case where two columns in $A$ are the same (up to a sign).
Minimal Coherence: Proof

- Since the $m$ diagonal elements of $G$ are ‘1’-es, $\text{tr}^2 \{G\} = m^2$
- Let $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ be the $n$ non-zero eigenvalues of the Gram matrix $G$ (which is of rank $n$ since $A$ is of rank $n$). Thus
  $$m^2 = \text{tr}^2 \{G\} = \left( \sum_{k=1}^{n} \lambda_k \right)^2$$
- Using the Cauchy-Swartz inequality
  $$\left( \sum_{k=1}^{n} \lambda_k \right)^2 = \left( \sum_{k=1}^{n} 1 \cdot \lambda_k \right)^2 \leq \left( \sum_{k=1}^{n} 1^2 \right) \cdot \left( \sum_{k=1}^{n} \lambda_k^2 \right) = n \cdot \sum_{k=1}^{n} \lambda_k^2$$
- Observe that $\sum_{k=1}^{n} \lambda_k^2 = \text{tr} \{G^2\}$

All this leads to the relation $\frac{m^2}{n} \leq \text{tr} \{G^2\}$
Minimal Coherence: Proof \[ \frac{m^2}{n} \leq \text{tr} \{ G^2 \} \]

- On the other hand, observe that

\[
\text{tr} \{ G^2 \} = \sum_{k=1}^{m} \sum_{j=1}^{m} (g_{k,j})^2 = \sum_{k=1}^{m} \sum_{j=1}^{m} (a_k^T a_j)^2 = m + \sum_{k \neq j} (a_k^T a_j)^2
\]

- Exploiting the relation \((k \neq j)\)

\[ m + \sum_{k \neq j} (a_k^T a_j)^2 \leq m + m(m - 1)\mu^2 \]

Thus \[ \frac{m^2}{n} \leq \text{tr} \{ G^2 \} \leq m + m(m - 1)\mu^2 \]

\[ \frac{m}{n} \leq 1 + (m - 1)\mu^2 \]

\[ \mu \geq \sqrt{\frac{m - n}{n(m - 1)}} \]
Grassmanian Frames

- Who are those special matrices $A$ of size $n \times m$ that have this minimal mutual coherence?

$$\mu_0 = \sqrt{\frac{m-n}{n(m-1)}}$$

The special case of $m=n$ leads to unitary matrices of coherence 0

- The answer is Grassmanian Frames

- Given $n$ and $m$, such a matrix does not necessarily exist

- If such a matrix does exist, the minimal coherence is in fact the same for all the pairs of atoms, implying that

$$\forall \ 1 \leq k < j \leq m, \ |g_{k,j}| = \mu_0 = \sqrt{\frac{m-n}{n(m-1)}}$$
Constructing Grassmanian Frames

- In 2005, Tropp et al. proposed a numerical algorithm for constructing Grassmanian matrices.
- Key property: Rather than operate on $A$, generate the corresponding Gram matrix $G$ and eventually take its square-root.
- Their method applies series of projections onto the constraints that such matrices obey:
  - Each column in $A$ is $L_2$-normalized (to be forced on $G$ by column+row normalization).
  - Given $A$ and its corresponding $G$, the off-diagonal values of $G$ should get closer to the value $\mu_0$.
  - When computing the Gram matrix $G$, its rank must be $n$. 
The Algorithm

Initialize \( A \) as random \( n \times m \) matrix with normalized columns

Compute the Gram matrix \( G = A^T A \)

Shrink off-diagonal Elements by \( \alpha < 1 \)

\[
\begin{align*}
g_{k,j} &> \mu_0 \quad \& \quad k \neq j \\
g_{k,j} &\Rightarrow g_{k,j} = \alpha \cdot g_{k,j}
\end{align*}
\]

Force rank(\( G \)) = \( n \) by SVD

\[
G = U \Sigma U^T \\
\Rightarrow G = U \Sigma_n U^T
\]

Obtain \( A \) by SVD

\[
G = U \Sigma U^T \Rightarrow A = \sqrt{\Sigma_n} U_n^T
\]

Force normalized Columns in \( A \)

\[
D = \sqrt{\text{diag} \{ G \}} \\
\Rightarrow G = D^{-1} G D^{-1}
\]
Run a Demo