ECE 18-898G: Special Topics in Signal Processing: Sparsity, Structure, and Inference

Sparse Representations

Yuejie Chi

Department of Electrical and Computer Engineering

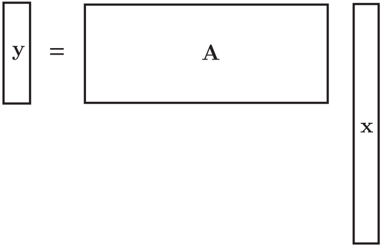
Carnegie Mellon University

Spring 2018

Outline

- Sparse and compressible signals
- Sparse representation in pairs of bases
- Uncertainty principles for basis pairs
  - Uncertainty principles for time-frequency bases
  - Uncertainty principles for general basis pairs
- Sparse representation via ℓ_1 minimization
- Sparse representation for general dictionaries

Basic problem



$$\text{Find } \boldsymbol{x} \in \mathbb{C}^n \quad \text{s.t. } \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}$$

where $\boldsymbol{A} = [\boldsymbol{a}_1, \cdots, \boldsymbol{a}_n] \in \mathbb{C}^{m \times n}$ obeys

- underdetermined system: $m < n$
- full-rank: $\text{rank}(\boldsymbol{A}) = m$

$\boldsymbol{A}$: an *over-complete basis / dictionary*; $\boldsymbol{a}_i$: atom;
$\boldsymbol{x}$: representation in this basis / dictionary

Sparse representation

Clearly, there exist infinitely many feasible solutions to $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}$ ...

- Solution set: $\underbrace{\boldsymbol{A}^*(\boldsymbol{A}\boldsymbol{A}^*)^{-1}}_{\boldsymbol{A}^\dagger}\boldsymbol{y} + \mathsf{null}(\boldsymbol{A})$
- $\boldsymbol{A}^\dagger$ is the pseudo-inverse of $\boldsymbol{A}$; $\mathsf{null}(\boldsymbol{A})$ is the null space of $\boldsymbol{A}$

How many "sparse" solutions are there?

What is sparsity?

Consider a signal $\boldsymbol{x} \in \mathbb{C}^n$.

Definition 2.1 (Support)

The *support* of a vector $\boldsymbol{x} \in \mathbb{C}^n$ is the *index set* of its nonzero entries, i.e.

$$\text{supp}(\boldsymbol{x}) := \{j \in [n] : |x_j| \neq 0\}$$

where $[p] = \{1, \ldots, n\}$.

Definition 2.2 (k-sparse signal)

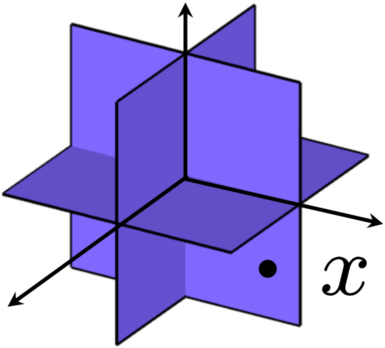
The signal $\boldsymbol{x}$ is called k-sparse, if

$$\|\boldsymbol{x}\|_0 := |\text{supp}(\boldsymbol{x})| \leq k.$$

$\|\boldsymbol{x}\|_0$ is called the sparsity level of $\boldsymbol{x}$. (Note: It is a "pseudo-norm").

Sparse signals belong to a union-of-subspace

- For a fixed sparsity pattern (support), it defines a subspace of dimension k in $\mathbb{R}^p$.
- There're $\binom{p}{k}$ subspaces of dimension k.



Best k-term approximations

We're also interested in signals that are *approximately* sparse (because a lot real-world signals are not exactly sparse). This is measured by how well they can be approximated by sparse signals.

Definition 2.3 (Best k-term approximation)

Denote the index set of the k-largest entries of $|\boldsymbol{x}|$ as S_k. The best k-term approximation $\boldsymbol{x}_k$ of $\boldsymbol{x}$ is defined as

$$\boldsymbol{x}_k(i) = \begin{cases} x_i, & i \in S_k \\ 0, & i \notin S_k \end{cases}$$

The (best) k-term approximation error in ℓ_p norm is then given as

$$\|\boldsymbol{x} - \boldsymbol{x}_k\|_p = \left(\sum_{i \notin S_k} |x_i|^p \right)^{1/p}.$$

Compressible signals

Compressibility: A signal is called *compressible* if

$$R(k) = \|\boldsymbol{x} - \boldsymbol{x}_k\|_p$$

decays "fast" in k.

Compressible signals

Compressibility: A signal is called *compressible* if

$$R(k) = \|\boldsymbol{x} - \boldsymbol{x}_k\|_p$$

decays "fast" in k.

Lemma 2.4 (Compressibility)

For any $q > p > 0$ *and* $\boldsymbol{x} \in \mathbb{R}^n$,

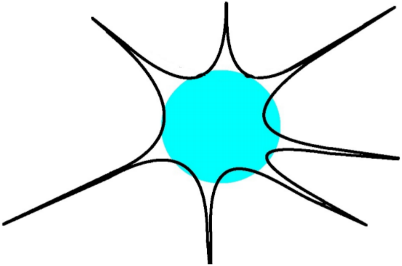
$$\|\boldsymbol{x} - \boldsymbol{x}_k\|_q \leq \frac{1}{k^{1/p-1/q}} \|\boldsymbol{x}\|_p.$$

Signals in ℓ_1 Ball

Example: Set $q = 2$ and $p = 1$, we have

$$\|\boldsymbol{x} - \boldsymbol{x}_k\|_2 \leq \frac{1}{\sqrt{k}} \|\boldsymbol{x}\|_1.$$

Consider a signal $\boldsymbol{x} \in B_1^n := \{\boldsymbol{z} \in \mathbb{R}^n : \|\boldsymbol{z}\|_1 \leq 1\}$. Then $\boldsymbol{x}$ is compressible when $p = 1$.



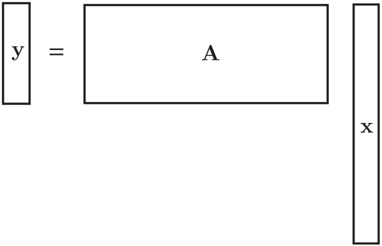
Geometrically, the ℓ_p-ball is pointy when $0 < p < 1$ in high dimension.

Proof of Lemma 2.4

Without loss of generality we assume the coefficients of $\boldsymbol{x}$ is ordered in descending order of magnitudes. We then have

$$
\begin{aligned}
\|\boldsymbol{x}-\boldsymbol{x}_k\|_q^q &= \sum_{j=k+1}^{n} |x_j|^q \quad \text{(by definition)} \\
&= |x_k|^{q-p} \sum_{j=k+1}^{n} |x_j|^p (|x_j|/|x_k|)^{q-p} \\
&\le |x_k|^{q-p} \sum_{j=k+1}^{n} |x_j|^p \quad (|x_j|/|x_k| \le 1) \\
&\le \left(\frac{1}{k}\sum_{j=1}^{k} |x_j|^p\right)^{\frac{q-p}{p}} \left(\sum_{j=k+1}^{n} |x_j|^p\right) \\
&\le \left(\frac{1}{k}\|\boldsymbol{x}\|_p^p\right)^{\frac{q-p}{p}} \|\boldsymbol{x}\|_p^p = \frac{1}{k^{q/p-1}}\|\boldsymbol{x}\|_p^q.
\end{aligned}
$$

Sparse representation in pairs of bases



A special type of dictionary: two-ortho case

Motivation for over-complete dictionary: many signals are mixtures of diverse phenomena; no single basis can describe them well

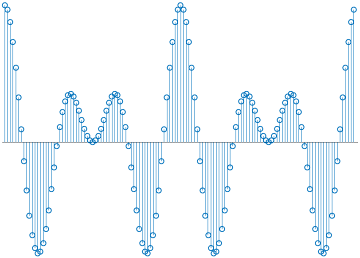
Two-ortho case: $\boldsymbol{A}$ is a concatenation of 2 orthonormal matrices

$$\boldsymbol{A} = [\boldsymbol{\Psi}, \boldsymbol{\Phi}] \qquad \text{where } \boldsymbol{\Psi}\boldsymbol{\Psi}^* = \boldsymbol{\Psi}^*\boldsymbol{\Psi} = \boldsymbol{\Phi}\boldsymbol{\Phi}^* = \boldsymbol{\Phi}^*\boldsymbol{\Phi} = \boldsymbol{I}$$

- A classical example: $\boldsymbol{A} = [\boldsymbol{I}, \boldsymbol{F}]$ $\quad$ ($\boldsymbol{F}$: Fourier matrix)
    - representing a signal $\boldsymbol{y}$ as a superposition of spikes and sinusoids

Example 1

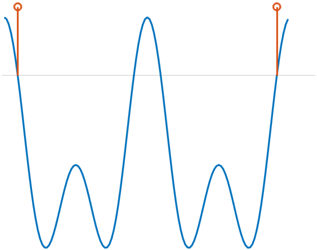
The following signal $\boldsymbol{y}_1$ is dense in the time domain, but sparse in the frequency domain


time-representation of $\boldsymbol{y}_1$


frequency-representation of $\boldsymbol{y}_1$

Example 2

The following signal $\boldsymbol{y}_2$ is dense in both time domain and frequency domain, but sparse in the overcomplete basis $[\boldsymbol{I}, \boldsymbol{F}]$

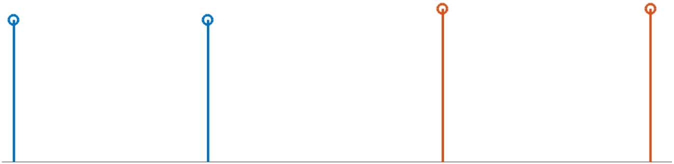


time representation of $\boldsymbol{y}_2$

frequency representation of $\boldsymbol{y}_2$

Example 2

The following signal $\boldsymbol{y}_2$ is dense in both time domain and frequency domain, but sparse in the overcomplete basis $[\boldsymbol{I}, \boldsymbol{F}]$



representation of $\boldsymbol{y}_2$ in overcomplete basis (time + frequency)

Uniqueness of sparse representation

A natural strategy to promote sparsity:

— seek the *sparsest* solution to the linear system

$$(P_0) \quad \text{minimize}_{\boldsymbol{x} \in \mathbb{C}^p} \ \|\boldsymbol{x}\|_0 \quad \text{s.t. } \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}$$

- When is the solution unique?
- How to test whether a candidate solution is the sparsest possible?

Application: multiuser detection

- 2 (or more) users wish to communicate to the same receiver over a shared wireless medium

- The jth user transmits $\boldsymbol{a}_j$; the receiver sees

$$\boldsymbol{y} = \sum_{j \text{ is active}} \boldsymbol{a}_j$$

- Let $\boldsymbol{A} = [\boldsymbol{a}_1; \cdots, \boldsymbol{a}_n]$ be the codebook containing all users of messages; then

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$$

  where the location of the non-zero entries of $\boldsymbol{x}$ indicates active users.

Unique representation $\mapsto$ unambiguous user identification

Connection to null space of A

Suppose $\boldsymbol{x}$ and $\boldsymbol{x}+\boldsymbol{h}$ are both solutions to the linear system, then

$$\boldsymbol{A}\boldsymbol{h} = \boldsymbol{A}(\boldsymbol{x}+\boldsymbol{h}) - \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y} - \boldsymbol{y} = \boldsymbol{0}$$

Write $\boldsymbol{h} = \begin{bmatrix} \boldsymbol{h}_{\boldsymbol{\Psi}} \\ \boldsymbol{h}_{\boldsymbol{\Phi}} \end{bmatrix}$ with $\boldsymbol{h}_{\boldsymbol{\Psi}}, \boldsymbol{h}_{\boldsymbol{\Phi}} \in \mathbb{C}^n$, then

$$\boldsymbol{\Psi}\boldsymbol{h}_{\boldsymbol{\Psi}} = -\boldsymbol{\Phi}\boldsymbol{h}_{\boldsymbol{\Phi}}$$

- $\boldsymbol{h}_{\boldsymbol{\Psi}}$ and $-\boldsymbol{h}_{\boldsymbol{\Phi}}$ are representations of the same vector in different bases
- (Non-rigorously) In order for $\boldsymbol{x}$ to be the sparsest solution, we hope $\boldsymbol{h}$ is much denser, i.e. we don't want $\boldsymbol{h}_{\boldsymbol{\Psi}}$ and $-\boldsymbol{h}_{\boldsymbol{\Phi}}$ to be simultaneously sparse

Detour: uncertainty principles for basis pairs

Heisenberg's uncertainty principle

A pair of ***complementary variables*** *cannot both be highly concentrated*

- Quantum mechanics

$$\underbrace{\mathsf{Var}[x]}_{\text{position}} \cdot \underbrace{\mathsf{Var}[p]}_{\text{momentum}} \geq \hbar^2/4$$

  - $\hbar$: Planck constant

Heisenberg's uncertainty principle

*A pair of **complementary variables** cannot both be highly concentrated*

- Quantum mechanics

$$\underbrace{\mathsf{Var}[x]}_{\text{position}} \cdot \underbrace{\mathsf{Var}[p]}_{\text{momentum}} \;\geq\; \hbar^2/4$$

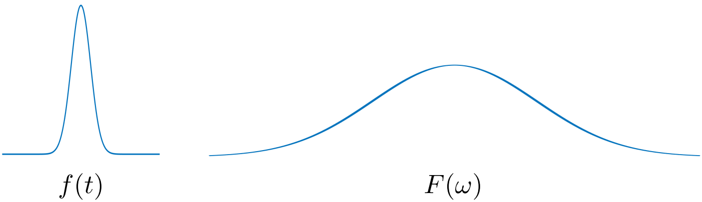
  - $\hbar$: Planck constant

- Signal processing

$$\underbrace{\int_{-\infty}^{\infty} t^2 |f(t)|^2 \mathrm{d}t}_{\text{concentration level of } f(t)} \quad \int_{-\infty}^{\infty} \omega^2 |F(\omega)|^2 \mathrm{d}\omega \geq 1/4$$

  - $f(t)$: a signal obeying $\int_{-\infty}^{\infty} |f(t)|^2 \mathrm{d}t = 1$
  - $F(\omega)$: Fourier transform of $f(t)$

Heisenberg's uncertainty principle



Roughly speaking, if $f(t)$ vanishes outside an interval of length Δt, and its Fourier transform vanishes outside an interval of length $\Delta\omega$, then

$$\Delta t \cdot \Delta\omega \geq \text{const}$$

Proof of Heisenberg's uncertainty principle

(assuming f is real-valued and $tf^2(t) \to 0$ as $|t| \to \infty$)

1. **Rewrite $\int \omega^2 |F(\omega)|^2 \mathrm{d}\omega$ in terms of f.** Since $f'(t) \xrightarrow{\mathcal{F}} i\omega F(\omega)$, Parseval's theorem yields

$$\int \omega^2 |F(\omega)|^2 \mathrm{d}\omega = \int |i\omega F(\omega)|^2 \mathrm{d}\omega = \int |f'(t)|^2 \mathrm{d}t$$

2. **Invoke Cauchy-Schwarz:**

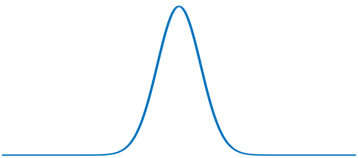
$$\left(\int t^2 |f(t)|^2 \mathrm{d}t\right)^{1/2} \left(\int |f'(t)|^2 \mathrm{d}t\right)^{1/2} \geq -\int t f(t) f'(t) \mathrm{d}t$$

$$= -0.5 \int t \frac{\mathrm{d}f^2(t)}{\mathrm{d}t} \mathrm{d}t$$

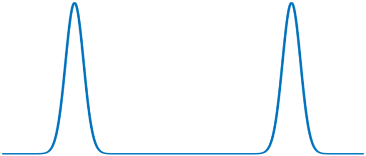
$$= -0.5 t f^2(t) \Big|_{-\infty}^{\infty} + 0.5 \int f^2(t) \mathrm{d}t \qquad \text{(integration by part)}$$

$$= 0.5 \qquad \text{(by our assumptions)}$$

Uncertainty principle for time-frequency bases



concentrated signal



sparse but non-concentrated signal

More general case: concentrated signals $\rightarrow$ sparse signals

- $f(t)$ and $F(\omega)$ are not necessarily concentrated on intervals

Question: is there a signal that can be sparsely represented both in time and in frequency?

- *Formally,* for an arbitrary $\boldsymbol{x}$, suppose $\hat{\boldsymbol{x}} = \boldsymbol{F}\boldsymbol{x}$.

How small can $\|\hat{\boldsymbol{x}}\|_0 + \|\boldsymbol{x}\|_0$ *be* ?

Uncertainty principle for time-frequency bases

Theorem 2.5 (Donoho & Stark '89)

Consider any nonzero $\boldsymbol{x} \in \mathbb{C}^n$, and let $\hat{\boldsymbol{x}} := \boldsymbol{F}\boldsymbol{x}$. Then

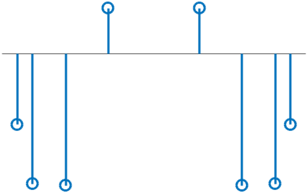
$$\underbrace{\|\boldsymbol{x}\|_0 \cdot \|\hat{\boldsymbol{x}}\|_0}_{\textit{time-bandwidth product}} \quad \geq n$$

- $\boldsymbol{x}$ and $\hat{\boldsymbol{x}}$ cannot be highly sparse simultaneously
- Does not rely on the support of $\boldsymbol{x}$ and $\hat{\boldsymbol{x}}$
- *Sanity check:* if $\boldsymbol{x} = [1, 0, \cdots, 0]^\top$ with $\|\boldsymbol{x}\|_0 = 1$, then $\|\hat{\boldsymbol{x}}\|_0 = n$ and hence $\|\boldsymbol{x}\|_0 \cdot \|\hat{\boldsymbol{x}}\|_0 = n$

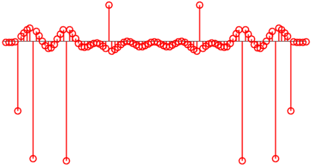
Corollary 2.6 (Donoho & Stark '89)

$$\|\boldsymbol{x}\|_0 + \|\hat{\boldsymbol{x}}\|_0 \geq 2\sqrt{n} \qquad \textit{(by AM-GM inequality)}$$

Application: super-resolution



wideband sparse signal $\boldsymbol{x}$



its low-pass version $\boldsymbol{x}_{\text{LP}}$

Consider a sparse wideband (i.e. $\|\boldsymbol{x}\|_0 \ll n$) signal $\boldsymbol{x} \in \mathbb{C}^n$, and project it onto a baseband B (of bandwidth $|B| < n$) to obtain its low-pass version $\boldsymbol{x}_{\text{LP}} = \text{Proj}_B(\boldsymbol{x})$. Then we can recover $\boldsymbol{x}$ from $\boldsymbol{x}_{\text{LP}}$ if

$$2\|\boldsymbol{x}\|_0 \cdot \underbrace{(n - |B|)}_{\text{size of unobserved band}} < n. \tag{2.1}$$

Application: super-resolution

Examples:

- If $\|\boldsymbol{x}\|_0 = 1$, then it's recoverable if $|B| > \frac{1}{2}n$
- If $\|\boldsymbol{x}\|_0 = 2$, then it's recoverable if $|B| > \frac{3}{4}n$
- $\cdots$

- First nontrivial performance guarantee for super-resolution
- *Somewhat pessimistic: we need to measure half of the bandwidth in order to recover just 1 spike*
- As will be seen later, we can do much better if nonzero entries of $\boldsymbol{x}$ are scattered

Application: super-resolution

Proof: If $\exists$ another solution $\boldsymbol{z} = \boldsymbol{x} + \boldsymbol{h}$ with $\|\boldsymbol{z}\|_0 \leq \|\boldsymbol{x}\|_0$, then

- $\mathsf{Proj}_B(\boldsymbol{h}) = \boldsymbol{0} \quad \Longrightarrow \quad \|\boldsymbol{F}\boldsymbol{h}\|_0 \leq n - |B|$
- $\|\boldsymbol{h}\|_0 \leq \|\boldsymbol{x}\|_0 + \|\boldsymbol{z}\|_0 \leq 2\|\boldsymbol{x}\|_0$

This together with the assumption (2.1) gives

$$\|\boldsymbol{h}\|_0 \cdot \|\boldsymbol{F}\boldsymbol{h}\|_0 \leq 2\|\boldsymbol{x}\|_0 \cdot (n - |B|) < n,$$

which violates Theorem 2.5 unless $\boldsymbol{h} = \boldsymbol{0}$.

Proof of Theorem 2.5: a key lemma

The key to prove Theorem 2.5 is to establish the following lemma

Lemma 2.7 (Donoho & Stark '89)

If $\boldsymbol{x} \in \mathbb{C}^n$ has k nonzero entries, then $\hat{\boldsymbol{x}} := \boldsymbol{F}\boldsymbol{x}$ cannot have k consecutive 0's.

Proof: Suppose $x_{\tau_1}, \cdots, x_{\tau_k}$ are the nonzero entries, and let $z = e^{-\frac{2\pi i}{n}}$.

1. For any consecutive frequency interval $(s, \cdots, s+k-1)$, the $(s+l)^{\text{th}}$ frequency component is

$$\hat{x}_{s+l} = \frac{1}{\sqrt{n}} \sum_{j=1}^{k} x_{\tau_j} z^{\tau_j (s+l)}, \quad l = 0, \cdots, k-1$$

Proof of Lemma 2.7

Proof (continued): One can thus write

$$\boldsymbol{g} := [\hat{x}_{s+l}]_{0\le l<k} = \frac{1}{\sqrt{n}}\boldsymbol{Z}\boldsymbol{x}_\tau,$$

where $\boldsymbol{x}_\tau := \begin{bmatrix} x_{\tau_1} z^{\tau_1 s} \\ x_{\tau_2} z^{\tau_2 s} \\ \vdots \\ x_{\tau_k} z^{\tau_k s} \end{bmatrix}$, $\boldsymbol{Z} := \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ z^{\tau_1} & \cdots & \cdots & \cdots & z^{\tau_k} \\ z^{2\tau_1} & \cdots & \cdots & \cdots & z^{2\tau_k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ z^{(k-1)\tau_1} & \cdots & \cdots & \cdots & z^{(k-1)\tau_k} \end{bmatrix}$

2. Recognizing that $\boldsymbol{Z}$ is a Vandermonde matrix yields

$$\det(\boldsymbol{Z}^\top) = \prod_{1\le i<j\le k} \left(z^{\tau_j} - z^{\tau_i}\right) \neq 0,$$

and hence $\boldsymbol{Z}$ is invertible. Therefore, $\boldsymbol{x}_\tau \neq \boldsymbol{0} \;\Rightarrow\; \boldsymbol{g} \neq \boldsymbol{0}$ as claimed.

Proof of Theorem 2.5

Suppose $\boldsymbol{x}$ is k-sparse, and suppose $n/k \in \mathbb{Z}$.

1. Partition $\{1, \cdots, n\}$ into n/k intervals of length k each.
2. By Lemma 2.7, none of these intervals of $\hat{\boldsymbol{x}}$ can vanish. Since each interval contains at least 1 non-zero entry, one has

$$\|\hat{\boldsymbol{x}}\|_0 \geq \frac{n}{k}$$

$$\iff \quad \|\boldsymbol{x}\|_0 \cdot \|\hat{\boldsymbol{x}}\|_0 \geq n$$

Exercise: fill in the proof for the case where k does not divide n.

Tightness of uncertainty principle

The lower bounds in Theorem 2.5 and Corollary 2.6 are achieved by the picket-fence signal $\boldsymbol{x}$ (a signal with uniform spacing $\sqrt{n}$).

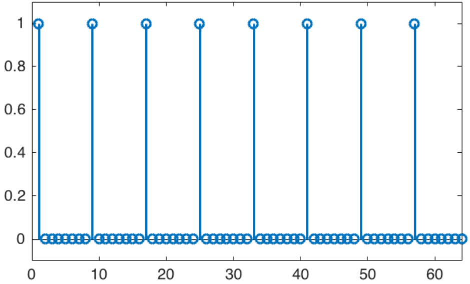


Figure 2.1: The picket-fence signal for $n = 64$, which obeys $\boldsymbol{F}\boldsymbol{x} = \boldsymbol{x}$

Uncertainty principle for general basis pairs

There are many other bases beyond time-frequency pairs

- Wavelets
- Ridgelets
- Hadamard
- ...

Generally, for an arbitrary $\boldsymbol{y} \in \mathbb{C}^n$ and arbitrary bases $\boldsymbol{\Psi}$ and $\boldsymbol{\Phi}$, suppose $\boldsymbol{y} = \boldsymbol{\Psi}\boldsymbol{\alpha} = \boldsymbol{\Phi}\boldsymbol{\beta}$:

How small can $\|\boldsymbol{\alpha}\|_0 + \|\boldsymbol{\beta}\|_0$ *be* ?

Uncertainty principle for general basis pairs

The degree of "uncertainty" depends on the basis pair.

- **Example:** suppose $\boldsymbol{\phi}_1, \boldsymbol{\phi}_2 \in \boldsymbol{\Psi}$ and $\frac{1}{\sqrt{2}}(\boldsymbol{\phi}_1 + \boldsymbol{\phi}_2)$, $\frac{1}{\sqrt{2}}(\boldsymbol{\phi}_1 - \boldsymbol{\phi}_2) \in \boldsymbol{\Psi}$. Then $\boldsymbol{y} = \boldsymbol{\phi}_1 + 0.5\boldsymbol{\phi}_2$ can be sparsely represented in both $\boldsymbol{\Psi}$ and $\boldsymbol{\Phi}$.

Message: uncertainty principle depends on how "different" $\boldsymbol{\Psi}$ and $\boldsymbol{\Phi}$ are.

Mutual coherence

A rough way to characterize how "similar" $\mathbf{\Psi}$ and $\mathbf{\Phi}$ are:

Definition 2.8 (Mutual coherence)

For any pair of orthonormal bases $\mathbf{\Psi} = [\psi_1, \cdots, \psi_n]$ and $\mathbf{\Phi} = [\phi_1, \cdots, \phi_n]$, the mutual coherence of these two bases is defined by

$$\mu(\mathbf{\Psi}, \mathbf{\Phi}) = \max_{1 \le i,j \le n} |\langle \psi_i, \phi_j \rangle| = \max_{1 \le i,j \le n} |\psi_i^* \phi_j|$$

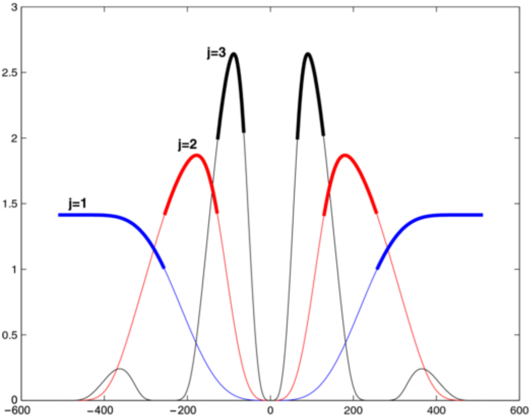
- $1/\sqrt{n} \le \mu(\mathbf{\Psi}, \mathbf{\Phi}) \le 1$ (homework)
- For $\mu(\mathbf{\Psi}, \mathbf{\Phi})$ to be small, each ψ_i needs to be "spread out" in the $\mathbf{\Phi}$ domain

Examples

- $\mu(\boldsymbol{I}, \boldsymbol{F}) = 1/\sqrt{n}$
  - Spikes and sinusoids are the most mutually incoherent

- Other extreme basis pair obeying $\mu(\boldsymbol{\Phi}, \boldsymbol{\Psi}) = 1/\sqrt{n}$: $\boldsymbol{\Psi} = \boldsymbol{I}$ and $\boldsymbol{\Phi} = \boldsymbol{H}$ (Hadamard matrix)

Fourier basis vs. wavelet basis ($n = 1024$)



Magnitudes of Daubechies-8 wavelets in the Fourier domain (j labels the scales of the wavelet transform with $j = 1$ the finest scale)

Fig. credit: Candes & Romberg '07

Uncertainty principle for general bases

Theorem 2.9 (Donoho & Huo '01, Elad & Bruckstein '02)

Consider any nonzero $\boldsymbol{b} \in \mathbb{C}^n$ *and any pair of orthonormal bases* $\boldsymbol{\Psi}$ *and* $\boldsymbol{\Phi}$. *Suppose* $\boldsymbol{b} = \boldsymbol{\Psi}\boldsymbol{\alpha} = \boldsymbol{\Phi}\boldsymbol{\beta}$. *Then*

$$\|\boldsymbol{\alpha}\|_0 \cdot \|\boldsymbol{\beta}\|_0 \geq \frac{1}{\mu^2(\boldsymbol{\Psi}, \boldsymbol{\Phi})}$$

Corollary 2.10 (Donoho & Huo '01, Elad & Bruckstein '02)

$$\|\boldsymbol{\alpha}\|_0 + \|\boldsymbol{\beta}\|_0 \geq \frac{2}{\mu(\boldsymbol{\Psi}, \boldsymbol{\Phi})}$$

(by AM-GM inequality)

Implications

- If two bases are "mutually incoherent", then we cannot have highly sparse representations in two bases simultaneously

- If $\boldsymbol{\Psi} = \boldsymbol{I}$ and $\boldsymbol{\Phi} = \boldsymbol{F}$, Theorem 2.9 reduces to

$$\|\boldsymbol{\alpha}\|_0 \cdot \|\boldsymbol{\beta}\|_0 \geq n$$

  since $\mu(\boldsymbol{\Psi}, \boldsymbol{\Phi}) = 1/\sqrt{n}$, which coincides with Theorem 2.5.

Proof of Theorem 2.9

1. WLOG, assume $\|\boldsymbol{b}\| = 1$. This gives

$$\begin{aligned} 1 = \boldsymbol{b}^*\boldsymbol{b} &= \boldsymbol{\alpha}^*\boldsymbol{\Psi}^*\Phi\boldsymbol{\beta} \\ &= \sum_{i,j=1}^{p} \alpha_i \langle \boldsymbol{\psi}_i, \boldsymbol{\phi}_j \rangle \beta_j \\ &\le \sum_{i,j=1}^{p} |\alpha_i| \cdot \mu(\boldsymbol{\Psi}, \boldsymbol{\Phi}) \cdot |\beta_j| \\ &\le \mu(\boldsymbol{\Psi}, \boldsymbol{\Phi}) \left(\sum_{i=1}^{p} |\alpha_i|\right) \left(\sum_{j=1}^{p} |\beta_j|\right) \qquad (2.2) \end{aligned}$$

Aside: this shows $\|\boldsymbol{\alpha}\|_1 \cdot \|\boldsymbol{\beta}\|_1 \ge \frac{1}{\mu(\boldsymbol{\Psi}, \boldsymbol{\Phi})}$

Proof of Theorem 2.9 (continued)

2. The assumption $\|\boldsymbol{b}\| = 1$ implies $\|\boldsymbol{\alpha}\| = \|\boldsymbol{\beta}\| = 1$. This together with the elementary inequality $\sum_{i=1}^{k} x_i \leq \sqrt{k \sum_{i=1}^{k} x_i^2}$ yields

$$\sum_{i=1}^{p} |\alpha_i| \leq \sqrt{\|\boldsymbol{\alpha}\|_0 \sum_{i=1}^{p} |\alpha_i|^2} = \sqrt{\|\boldsymbol{\alpha}\|_0}$$

   Similarly, $\sum_{i=1}^{p} |\beta_i| \leq \sqrt{\|\boldsymbol{\beta}\|_0}$.

3. Substitution into (2.2) concludes the proof.

Uniqueness of sparse representation

A natural strategy to promote sparsity:

— seek the *sparsest* solution to the linear system

$$(P_0) \quad \text{minimize}_{\boldsymbol{x} \in \mathbb{C}^p} \ \|\boldsymbol{x}\|_0 \quad \text{s.t. } \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}$$

- When is the solution unique?
- How to test whether a candidate solution is the sparsest possible?

Uniqueness of ℓ_0 minimization

The uncertainty principle leads to the possibility of ideal sparse representation for the system

$$\boldsymbol{y} = [\boldsymbol{\Psi}, \boldsymbol{\Phi}]\boldsymbol{x} \tag{2.3}$$

Theorem 2.11 (Donoho & Huo '01, Elad & Bruckstein '02)

Any two distinct solutions $\boldsymbol{x}^{(1)}$ and $\boldsymbol{x}^{(2)}$ to (2.3) satisfy

$$\|\boldsymbol{x}^{(1)}\|_0 + \|\boldsymbol{x}^{(2)}\|_0 \geq \frac{2}{\mu(\boldsymbol{\Psi}, \boldsymbol{\Phi})}$$

Corollary 2.12 (Donoho & Huo '01, Elad & Bruckstein '02)

If a solution $\boldsymbol{x}$ obeys $\|\boldsymbol{x}\|_0 < \frac{1}{\mu(\boldsymbol{\Psi}, \boldsymbol{\Phi})}$, then it is necessarily the unique sparsest solution.

Proof of Theorem 2.11

Define $\boldsymbol{h} = \boldsymbol{x}^{(1)} - \boldsymbol{x}^{(2)}$, and write $\boldsymbol{h} = \begin{bmatrix} \boldsymbol{h}_\Psi \\ \boldsymbol{h}_\Phi \end{bmatrix}$ with $\boldsymbol{h}_\Psi, \boldsymbol{h}_\Phi \in \mathbb{C}^n$.

1. Since $\boldsymbol{y} = [\boldsymbol{\Psi}, \boldsymbol{\Phi}]\boldsymbol{x}^{(1)} = [\boldsymbol{\Psi}, \boldsymbol{\Phi}]\boldsymbol{x}^{(2)}$, one has

$$[\boldsymbol{\Psi}, \boldsymbol{\Phi}]\boldsymbol{h} = \boldsymbol{0} \quad \Longleftrightarrow \quad \boldsymbol{\Psi}\boldsymbol{h}_\Psi = -\boldsymbol{\Phi}\boldsymbol{h}_\Phi$$

2. By Corollary 2.10,

$$\|\boldsymbol{h}\|_0 = \|\boldsymbol{h}_\Psi\|_0 + \|\boldsymbol{h}_\Phi\|_0 \geq \frac{2}{\mu(\boldsymbol{\Psi}, \boldsymbol{\Phi})}$$

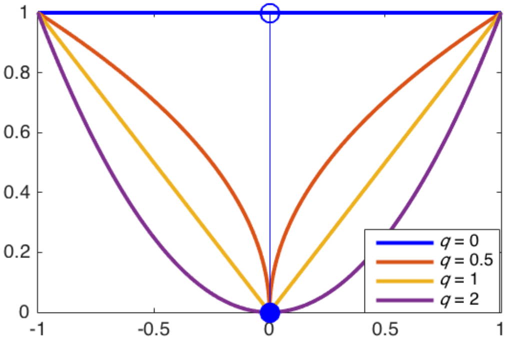
3. $\|\boldsymbol{x}^{(1)}\|_0 + \|\boldsymbol{x}^{(2)}\|_0 \geq \|\boldsymbol{h}\|_0 \geq \frac{2}{\mu(\boldsymbol{\Psi}, \boldsymbol{\Phi})}$ as claimed.

Sparse representation via ℓ_1 minimization

Relaxation of the highly discontinuous ℓ_0 norm

Unfortunately, ℓ_0 minimization is computationally intractable ...

Simple heuristic: replacing ℓ_0 norm with continuous (or even smooth) approximation



$|x|^q$ vs. x

Convexification: ℓ_1 minimization (basis pursuit)

$$\text{minimize}_{\boldsymbol{x}\in\mathbb{C}^p} \ \|\boldsymbol{x}\|_0 \quad \text{s.t. } \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}$$

$$\Downarrow$$

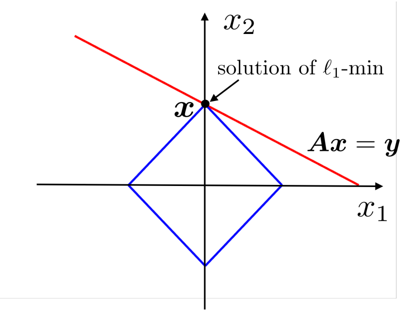
Convexifying $\|\boldsymbol{x}\|_0$ with $\|\boldsymbol{x}\|_1$

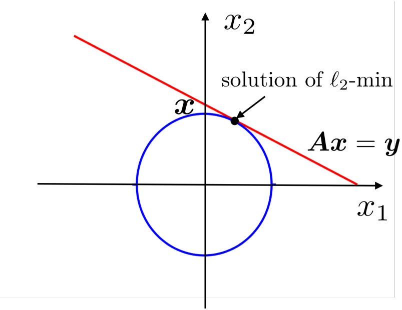
$$\Downarrow$$

$$\text{minimize}_{\boldsymbol{x}\in\mathbb{C}^p} \ \|\boldsymbol{x}\|_1 \quad \text{s.t. } \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y} \tag{2.4}$$

- $|x|$ is the largest convex function less than $\mathbf{1}\{x \neq 0\}$ over $\{x : |x| \leq 1\}$
- ℓ_1 minimization is a linear program (homework)
- ℓ_1 minimization is non-smooth optimization (since $\|\cdot\|_1$ is non-smooth)
- ℓ_1 minimization does not rely on prior knowledge on sparsity level

Geometry





$\min_{\boldsymbol{x}} \ \|\boldsymbol{x}\|_1$ s.t. $\boldsymbol{Ax} = \boldsymbol{y}$ $\quad\quad$ $\min_{\boldsymbol{x}} \ \|\boldsymbol{x}\|_2$ s.t. $\boldsymbol{Ax} = \boldsymbol{y}$

Even pointier in the high dimension

- Level sets of $\|\cdot\|_1$ are pointed, enabling it to promote sparsity
- Level sets of $\|\cdot\|_2$ are smooth, often leading to dense solutions

Effectiveness of ℓ_1 minimization

Theorem 2.13 (Donoho & Huo '01, Elad & Bruckstein '02)

$\boldsymbol{x} \in \mathbb{C}^p$ *is the unique solution to* ℓ_1 *minimization (2.4) if*

$$\|\boldsymbol{x}\|_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(\boldsymbol{\Psi}, \boldsymbol{\Phi})}\right) \tag{2.5}$$

- ℓ_1 minimization yields the sparse solution too!
- The recovery condition (2.5) can be improved to, e.g.,

$$\|\boldsymbol{x}\|_0 < \frac{0.914}{\mu(\boldsymbol{\Psi}, \boldsymbol{\Phi})} \quad \text{[Elad \& Bruckstein '02]}$$

Effectiveness of ℓ_1 minimization

$$\|\boldsymbol{x}\|_0 < \frac{1}{\mu(\boldsymbol{\Psi},\boldsymbol{\Phi})} \quad \Longrightarrow \quad \ell_0 \text{ minimization works}$$

$$\|\boldsymbol{x}\|_0 < \frac{0.914}{\mu(\boldsymbol{\Psi},\boldsymbol{\Phi})} \quad \Longrightarrow \quad \ell_1 \text{ minimization works}$$

The recovery condition for ℓ_1 miniization is within a factor of $1/0.914 \approx 1.094$ of the condition derived for ℓ_0 minimization

Proof of Theorem 2.13

We need to show that $\|\boldsymbol{x}+\boldsymbol{h}\|_1 > \|\boldsymbol{x}\|_1$ holds for any other feasible solution $\boldsymbol{x}+\boldsymbol{h}$. To this end, we proceed as follows

$$\begin{aligned}
& \|\boldsymbol{x}+\boldsymbol{h}\|_1 > \|\boldsymbol{x}\|_1 \\
\Longleftarrow \quad & \sum_{i \notin \mathsf{supp}(\boldsymbol{x})} |h_i| + \sum_{i \in \mathsf{supp}(\boldsymbol{x})} (|x_i + h_i| - |x_i|) > 0 \\
\Longleftarrow \quad & \sum_{i \notin \mathsf{supp}(\boldsymbol{x})} |h_i| - \sum_{i \in \mathsf{supp}(\boldsymbol{x})} |h_i| > 0 \qquad (\text{since } |a+b| - |a| \geq -|b|) \\
\Longleftarrow \quad & \|\boldsymbol{h}\|_1 > 2 \sum_{i \in \mathsf{supp}(\boldsymbol{x})} |h_i| \\
\Longleftarrow \quad & \sum_{i \in \mathsf{supp}(\boldsymbol{x})} \frac{|h_i|}{\|\boldsymbol{h}\|_1} < \frac{1}{2} \\
\Longleftarrow \quad & \|\boldsymbol{x}\|_0 \frac{\|\boldsymbol{h}\|_\infty}{\|\boldsymbol{h}\|_1} < \frac{1}{2} \qquad (2.6)
\end{aligned}$$

Proof of Theorem 2.13 (continued)

It remains to control $\frac{\|\boldsymbol{h}\|_\infty}{\|\boldsymbol{h}\|_1}$. As usual, due to feasibility constraint we have $[\boldsymbol{\Psi}, \boldsymbol{\Phi}]\boldsymbol{h} = \boldsymbol{0}$, or

$$\boldsymbol{\Psi}\boldsymbol{h}_\psi = -\boldsymbol{\Phi}\boldsymbol{h}_\phi \quad \Longleftrightarrow \quad \boldsymbol{h}_\psi = -\boldsymbol{\Psi}^*\boldsymbol{\Phi}\boldsymbol{h}_\phi \qquad \text{where } \boldsymbol{h} = \begin{bmatrix} \boldsymbol{h}_\psi \\ \boldsymbol{h}_\phi \end{bmatrix}.$$

For any i, the inequality $|\boldsymbol{a}^*\boldsymbol{b}| \leq \|\boldsymbol{a}\|_\infty \|\boldsymbol{b}\|_1$ gives

$$|(\boldsymbol{h}_\psi)_i| = |(\boldsymbol{\Psi}^*\boldsymbol{\Phi})_{\text{row } i} \cdot \boldsymbol{h}_\phi| \leq \|\boldsymbol{\Psi}^*\boldsymbol{\Phi}\|_\infty \cdot \|\boldsymbol{h}_\phi\|_1 = \mu(\boldsymbol{\Psi}, \boldsymbol{\Phi}) \cdot \|\boldsymbol{h}_\phi\|_1$$

On the other hand, $\|\boldsymbol{h}_\psi\|_1 \geq |(\boldsymbol{h}_\psi)_i|$. Putting them together yields

$$\|\boldsymbol{h}\|_1 = \|\boldsymbol{h}_\phi\|_1 + \|\boldsymbol{h}_\psi\|_1 \geq |(\boldsymbol{h}_\psi)_i| \left(1 + \frac{1}{\mu(\boldsymbol{\Psi}, \boldsymbol{\Phi})}\right) \tag{2.7}$$

Proof of Theorem 2.13 (continued)

In fact, this inequality (2.7) holds for any entry of $\boldsymbol{h}$, giving that

$$\frac{\|\boldsymbol{h}\|_\infty}{\|\boldsymbol{h}\|_1} \leq \frac{1}{1 + \frac{1}{\mu(\boldsymbol{\Psi}, \boldsymbol{\Phi})}}$$

Finally, if $\|\boldsymbol{x}\|_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(\boldsymbol{\Psi}, \boldsymbol{\Phi})}\right)$, then

$$\|\boldsymbol{x}\|_0 \cdot \frac{\|\boldsymbol{h}\|_\infty}{\|\boldsymbol{h}\|_1} < \frac{1}{2}$$

as claimed in (2.6), thus concluding the proof.

Sparse representation for general dictionaries

Beyond two-ortho case

$$\text{minimize}_{\boldsymbol{x}} \ \|\boldsymbol{x}\|_0 \quad \text{s.t. } \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$$

What if $\boldsymbol{A} \in \mathbb{C}^{n \times p}$ is a general overcomplete dictionary?

We will study this general case through 2 metrics

1. Mutual coherence

2. Spark

Mutual coherence for arbitrary dictionaries

Definition 2.14 (Mutual coherence)

For any $\boldsymbol{A} = [\boldsymbol{a}_1, \cdots, \boldsymbol{a}_p] \in \mathbb{C}^{n \times p}$, the mutual coherence of $\boldsymbol{A}$ is defined by

$$\mu(\boldsymbol{A}) = \max_{1 \leq i,j \leq p,\ i \neq j} \frac{|\boldsymbol{a}_i^* \boldsymbol{a}_j|}{\|\boldsymbol{a}_i\| \|\boldsymbol{a}_j\|}$$

- If $\|\boldsymbol{a}_i\| = 1$ for all i, then $\mu(\boldsymbol{A})$ is the maximum off-diagonal entry (in absolute value) of the Gram matrix $\boldsymbol{G} = \boldsymbol{A}^* \boldsymbol{A}$
- $\mu(\boldsymbol{A})$ characterizes "second-order" dependency across the atoms $\{\boldsymbol{a}_i\}$
- (Welch bound) $\mu(\boldsymbol{A}) \geq \sqrt{\frac{p-n}{n(p-1)}}$, with equality attained by a family called *Grassmannian frames*

Uniqueness of sparse representation via $\mu(\boldsymbol{A})$

A theoretical guarantee similar to the two-ortho case

Theorem 2.15 (Donoho & Elad '03, Gribonval & Nielsen '03, Fuchs '04)

If $\boldsymbol{x}$ is a feasible solution that obeys $\|\boldsymbol{x}\|_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(\boldsymbol{A})}\right)$, then $\boldsymbol{x}$ is the unique solution to both ℓ_0 and ℓ_1 minimization.

Tightness?

Suppose $p = cn$ for some constant $c > 2$, then Welch bound gives

$$\mu(\boldsymbol{A}) \geq 1/\sqrt{2n}.$$

$\Longrightarrow$ for the "most incoherent" (and hence best possible) dictionary, the recovery condition reads

$$\|\boldsymbol{x}\|_0 = O(\sqrt{n})$$

This says: to recover a $\sqrt{n}$-sparse signal (and hence $\sqrt{n}$ degrees of freedom), we need an order of n samples

- The measurement burden is way too high!
- *Mutual coherence might not capture the information bottleneck!*

Another metric: Spark

Definition 2.16 (Spark, Donoho & Elad '03)

spark($\boldsymbol{A}$) is the size of the smallest linearly dependent column subset of $\boldsymbol{A}$, i.e.

$$\text{spark}(\boldsymbol{A}) = \min_{\boldsymbol{z}} \|\boldsymbol{z}\|_0 \quad \text{s.t.} \quad \boldsymbol{A}\boldsymbol{z} = \boldsymbol{0}$$

- A way of characterizing null-space of $\boldsymbol{A}$ using ℓ_0 norm
- Comparison to rank
  - rank($\boldsymbol{A}$): largest number of columns from $\boldsymbol{A}$ that are linearly independent
  - spark($\boldsymbol{A}$) is far more difficult to compute than rank($\boldsymbol{A}$)
- $2 \leq \text{spark}(\boldsymbol{A}) \leq \text{rank}(\boldsymbol{A}) + 1$ for nontrivial $\boldsymbol{A}$

Examples

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

- spark$(\boldsymbol{A}) = 3$
- rank$(\boldsymbol{A}) = 4$

Examples

Suppose $\sqrt{n} \in \mathbb{Z}$. Then $\boldsymbol{A} = [\boldsymbol{I}, \boldsymbol{F}] \in \mathbb{C}^{n \times 2n}$ obeys

$$\mathsf{spark}(\boldsymbol{A}) = 2\sqrt{n}$$

- Hint: consider the concatenation of two picket-fence signals each with $\sqrt{n}$ peaks

Examples

Suppose the entries of $\boldsymbol{A}$ are i.i.d. standard Gaussian, then

$$\mathsf{spark}(\boldsymbol{A}) = n + 1$$

with probability 1, since no n columns are linearly dependent.

Uniqueness via spark

Spark provides a simple criterion for uniqueness:

Theorem 2.17

If $\boldsymbol{x}$ is a solution to $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}$ and obeys $\|\boldsymbol{x}\|_0 < \mathrm{spark}(\boldsymbol{A})/2$, then $\boldsymbol{x}$ is necessarily the unique sparsest possible solution.

- If $\boldsymbol{A}$ is an i.i.d. Gaussian matrix (and hence $\mathrm{spark}(\boldsymbol{A}) = n + 1$), then this condition reads

$$\|\boldsymbol{x}\|_0 < (n + 1)/2$$

i.e., n samples enable us to recover $n/2$ units of information!

  - *much better than the condition based on $\mu(\boldsymbol{A})$*

Proof of Theorem 2.17

Consider any other feasible solution $\boldsymbol{z} \neq \boldsymbol{x}$.

1. Since $\boldsymbol{Az} = \boldsymbol{Ax} = \boldsymbol{y}$, one has

$$\boldsymbol{A}(\boldsymbol{x} - \boldsymbol{z}) = \boldsymbol{0},$$

   i.e. the columns of $\boldsymbol{A}$ at indices coming from the support of $\boldsymbol{x} - \boldsymbol{z}$ are linearly dependent

2. By definition,

$$\mathsf{spark}(\boldsymbol{A}) \leq \|\boldsymbol{x} - \boldsymbol{z}\|_0$$

3. The fact $\|\boldsymbol{x}\|_0 + \|\boldsymbol{z}\|_0 \geq \|\boldsymbol{x} - \boldsymbol{z}\|_0$ then gives

$$\|\boldsymbol{x}\|_0 + \|\boldsymbol{z}\|_0 \geq \mathsf{spark}(\boldsymbol{A})$$

4. If $\|\boldsymbol{x}\|_0 < \mathsf{spark}(\boldsymbol{A})/2$, then

$$\|\boldsymbol{z}\|_0 \geq \mathsf{spark}(\boldsymbol{A})/2 > \|\boldsymbol{x}\|_0$$

Connecting Spark with mutual coherence

Theorem 2.18 (Donoho & Elad '03)

$$\text{spark}(\boldsymbol{A}) \geq 1 + 1/\mu(\boldsymbol{A})$$

Connecting Spark with mutual coherence

Corollary 2.19 (Donoho & Elad '03)

If a solution $\boldsymbol{x}$ *obeys* $\|\boldsymbol{x}\|_0 < 0.5(1 + 1/\mu(\boldsymbol{A}))$*, then it is the sparsest possible solution.*

- Corollary 2.19 is, however, much weaker than Theorem 2.17
- **Example (2-ortho case):**
    - Corollary 2.19 gives $\|\boldsymbol{x}\|_0 = O(\sqrt{n})$ at best, since $\mu(\boldsymbol{A}) \geq 1/\sqrt{n}$
    - Theorem 2.17 may give a bound as large as $\|\boldsymbol{x}\|_0 = O(n)$ since spark$(\boldsymbol{A})$ may be as large as n

Proof of Theorem 2.18

WLOG, assume $\|\boldsymbol{a}_i\| = 1$, $\forall i$, then the Gram matrix $\boldsymbol{G} := \boldsymbol{A}^* \boldsymbol{A}$ obeys

$$G_{i,i} = 1 \quad \forall i \qquad \text{and} \qquad |G_{i,j}| \le \mu(\boldsymbol{A}) \quad \forall i \neq j \tag{2.8}$$

1. Consider any $k \times k$ principal submatrix $\boldsymbol{G}_{J,J}$ of $\boldsymbol{G}$ with J an index subset. If $\boldsymbol{G}_{J,J} \succ \boldsymbol{0}$, then the k columns of $\boldsymbol{A}$ at indices in J are linearly independent

2. If this holds for all $k \times k$ principal submatrices, then by definition spark$(\boldsymbol{A}) > k$

3. Finally, by Gershgorin circle theorem, one would have $\boldsymbol{G}_{J,J} \succ \boldsymbol{0}$ if $|G_{i,i}| > \sum_{j \in J,\ j \neq i} |G_{i,j}|$, which would follow if (by (2.8))

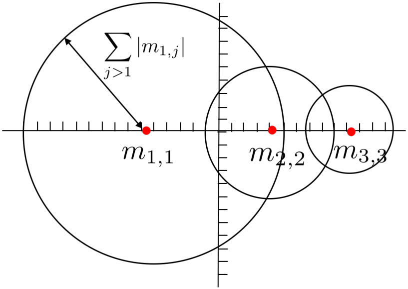
$$1 > (k-1)\mu(\boldsymbol{A})$$

   i.e. k can be as large as $1 + \lfloor 1/\mu(\boldsymbol{A}) \rfloor$

Gershgorin circle theorem

Lemma 2.20 (Gershgorin circle theorem)

The eigenvalues of $\boldsymbol{M} = [m_{i,j}]_{1 \le i,j \le n}$ *lie in the union of* n *discs* $\text{disc}(c_i, r_i)$, $1 \le i \le n$, *centered at* $c_i = m_{ii}$ *and with radius* $r_i = \sum_{j:j \ne i} |m_{ij}|$.



Summary

- For many dictionaries, if a signal is representable in a highly sparse manner, then it is often guaranteed to be the unique sparse solution.
- Seeking a sparse solution often becomes a well-posed question with interesting properties

Reference

[1] "*Sparse and redundant representations: from theory to applications in signal and image processing*," M. Elad, *Springer*, 2010.

[2] "*Uncertainty principles and signal recovery*," D. Donoho and P. Stark, *SIAM Journal on Applied Mathematics*, 1989.

[3] "*Uncertainty principles and ideal atomic decomposition*," D. Donoho and X. Huo, *IEEE Trans. on Info. Theory,* 2001.

[4] "*A generalized uncertainty principle and sparse representation in pairs of bases*," M. Elad and A. Bruckstein, *IEEE Trans. on Info. Theory,* 2002.

[5] "*Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization*," D. Donoho, and M. Elad, *Proceedings of the National Academy of Sciences,* 2003.

[6] "*Sparsity and incoherence in compressive sampling*," E. Candes, and J. Romberg, *Inverse Problems,* 2007.

Reference

[7] "*Atomic decomposition by basis pursuit*," S. Chen, D. Donoho, M. A. Saunders, *SIAM review,* 2001.

[8] "*On sparse representations in arbitrary redundant bases*," J. Fuchs, *IEEE Trans. on Info. Theory,* 2004.

[9] "*Sparse representations in unions of bases*," R. Gribonval, and M. Nielsen, *IEEE Trans. on Info. Theory,* 2003.