Sparse Representation

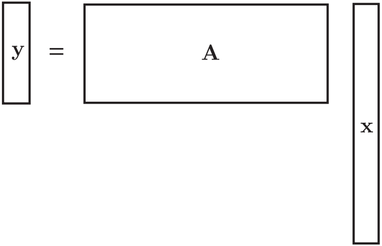
Yuxin Chen

Wharton Statistics & Data Science, Spring 2022

Outline

- 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



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

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

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

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

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

Sparse representation

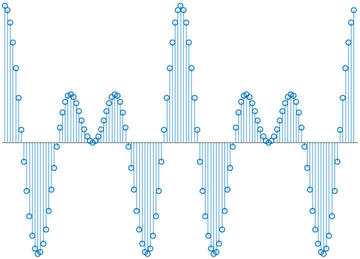
Clearly, there exist infinitely many feasible solutions to $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}$ *...*

- Solution set: $\boldsymbol{A}^*(\boldsymbol{A}\boldsymbol{A}^*)^{-1}\boldsymbol{y} + \mathsf{null}(\boldsymbol{A})$

How many "sparse" solutions are there?

Example 1

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

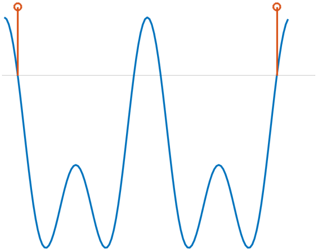


time-representation of $\boldsymbol{y}_1$

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

Example 2

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

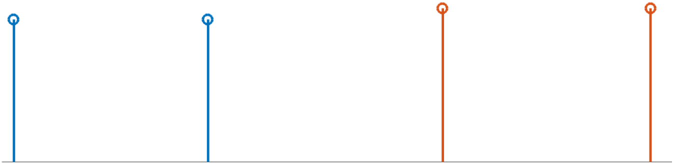


time representation of $\boldsymbol{y}_2$

frequency representation of $\boldsymbol{y}_2$

Example 2

The following signal $\boldsymbol{y}_2$ is dense in both the time and the frequency domains, 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 *sparsest* solution to a 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 private communications

- 2 (or more) users wish to communicate to the same receiver over a shared wireless medium
- the jth user transmits $\boldsymbol{s}_j$; the receiver sees $\boldsymbol{s} = \sum_j \boldsymbol{s}_j$
- for the sake of privacy, the jth user adopts its own codebook

$$\boldsymbol{s}_j = \boldsymbol{A}_j \boldsymbol{x}_j$$

  where $\boldsymbol{x}_j$ is the message (typically sparse), and $\boldsymbol{A}_j$ is the dictionary (known to the receiver; unknown to other users)

It comes down to whether the receiver can recover all messages unambiguously

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

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

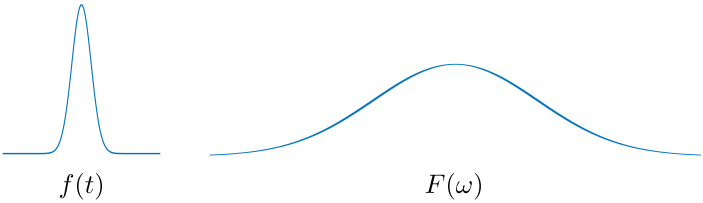
  - $\hbar$: Planck constant

- 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 $|t| \to \infty$)

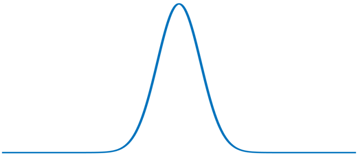
1. **Rewrite $\int \omega^2 |F(\omega)|^2 \mathrm{d}\omega$ in terms of f.** Since $f'(t) \xrightarrow{\mathcal{F}} i\omega F(\omega)$, Parseval's theorem yields

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

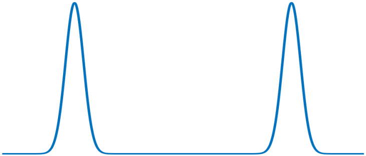
2. **Invoke Cauchy-Schwarz:**

$$\begin{aligned}
\left(\int t^2 |f(t)|^2 \mathrm{d}t \right)^{1/2} \left(\int |f'(t)|^2 \mathrm{d}t \right)^{1/2} &\geq -\int t f(t) f'(t) \mathrm{d}t \\
&= -0.5 \int t \frac{\mathrm{d}f^2(t)}{\mathrm{d}t} \mathrm{d}t \\
&= -0.5 t f^2(t) \Big|_{-\infty}^{\infty} + 0.5 \int f^2(t) \mathrm{d}t \qquad \text{(integration by part)} \\
&= 0.5 \qquad \text{(by our assumptions)}
\end{aligned}$$

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 $\hat{\boldsymbol{x}} = \boldsymbol{F}\boldsymbol{x}$.

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

Uncertainty principle for time-frequency bases

Theorem 7.1 (Donoho & Stark '89)

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

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

- $\boldsymbol{x}$ and $\hat{\boldsymbol{x}}$ cannot be highly sparse simultaneously
- does not rely on the kind of sets where $\boldsymbol{x}$ and $\hat{\boldsymbol{x}}$ are nonzero
- *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 7.2 (Donoho & Stark '89)

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

Proof of Theorem 7.1: a key lemma

The key to proving Theorem 7.1 is to establish the following lemma:

Lemma 7.3 (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 of Theorem 7.1

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 7.3, none of these intervals of $\hat{\boldsymbol{x}}$ can vanish. Since each interval contains at least 1 non-zero entry, one has

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

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

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

Proof of Lemma 7.3

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

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

   One can thus write

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

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

Proof of Lemma 7.3 (cont.)

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

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

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

Tightness of uncertainty principle

Lower bounds in Theorem 7.1 and Corollary 7.2 are achieved by the picket-fence signal $\boldsymbol{x}$ (a signal with uniform spacing $\sqrt{n}$)

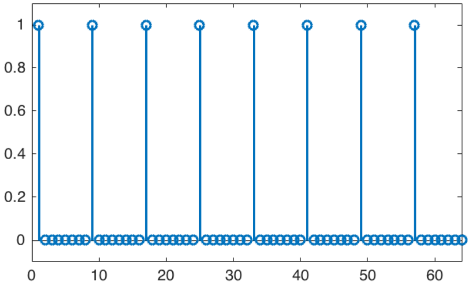


Figure 7.1: The picket-fence signal for $n = 64$, which obeys $\boldsymbol{F}\boldsymbol{x} = \boldsymbol{x}$

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 $\phi_1, \phi_2 \in \mathbf{\Psi}$ and $\frac{1}{\sqrt{2}}(\phi_1 + \phi_2)$, $\frac{1}{\sqrt{2}}(\phi_1 - \phi_2) \in \mathbf{\Psi}$ (so two bases share similarity). Then $\boldsymbol{y} = \phi_1 + 0.5\phi_2$ can be sparsely represented in both $\mathbf{\Psi}$ and $\mathbf{\Phi}$ (i.e. we have multiple sparse representations)

The uncertainty principle depends on how "different" $\mathbf{\Psi}$ and $\mathbf{\Phi}$ are

Mutual coherence

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

Definition 7.4 (Mutual coherence)

For any pair of orthonormal bases $\mathbf{\Psi} = [\boldsymbol{\psi}_1, \cdots, \boldsymbol{\psi}_n]$ and $\mathbf{\Phi} = [\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_n]$, the mutual coherence of these two bases is defined by

$$\mu(\mathbf{\Psi}, \mathbf{\Phi}) = \max_{1 \leq i,j \leq n} |\boldsymbol{\psi}_i^* \boldsymbol{\phi}_j|$$

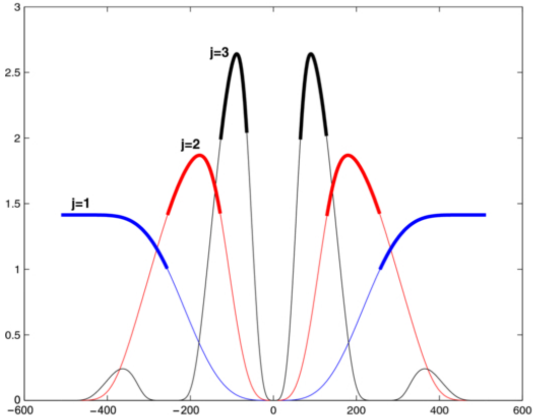
- $1/\sqrt{n} \leq \mu(\mathbf{\Psi}, \mathbf{\Phi}) \leq 1$ (homework)
- For $\mu(\mathbf{\Psi}, \mathbf{\Phi})$ to be small, each $\boldsymbol{\psi}_i$ needs to be "spread out" in the $\mathbf{\Phi}$ domain

Examples

- $\mu(\boldsymbol{I}, \boldsymbol{F}) = 1/\sqrt{n}$
    - Spikes and sinusoids are 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 7.5 (Donoho & Huo '01, Elad & Bruckstein '02)

Consider any nonzero $\boldsymbol{b} \in \mathbb{C}^n$ *and any pair of orthonormal bases* $\boldsymbol{\Psi}, \boldsymbol{\Phi} \in \mathbb{C}^n$*. 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 7.6 (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 7.5 reduces to

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

  since $\mu(\boldsymbol{\Psi}, \boldsymbol{\Phi}) = 1/\sqrt{n}$, which coincides with Theorem 7.1

Proof of Theorem 7.5

1. WLOG, assume $\|\boldsymbol{b}\|_2 = 1$. This gives

$$\begin{aligned} 1 = \boldsymbol{b}^*\boldsymbol{b} &= \boldsymbol{\alpha}^*\boldsymbol{\Psi}^*\boldsymbol{\Phi}\boldsymbol{\beta} \\ &= \sum\nolimits_{i,j=1}^{p} \alpha_i \langle \boldsymbol{\psi}_i, \boldsymbol{\phi}_j \rangle \beta_j \\ &\le \sum\nolimits_{i,j=1}^{p} |\alpha_i| \cdot \mu(\boldsymbol{\Psi}, \boldsymbol{\Phi}) \cdot |\beta_j| \\ &\le \mu(\boldsymbol{\Psi}, \boldsymbol{\Phi}) \left(\sum\nolimits_{i=1}^{p} |\alpha_i|\right) \left(\sum\nolimits_{j=1}^{p} |\beta_j|\right) \qquad (7.1) \end{aligned}$$

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

Proof of Theorem 7.5 (cont.)

2. The assumption $\|\boldsymbol{b}\|_2 = 1$ implies $\|\boldsymbol{\alpha}\|_2 = \|\boldsymbol{\beta}\|_2 = 1$. This together with elementary inequality $\sum_{i=1}^k x_i \le \sqrt{k \sum_{i=1}^k x_i^2}$ yields

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

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

3. Substitution into (7.1) concludes the proof

Back to the uniqueness of ℓ_0 minimization

Uncertainty principle suggests the possibility of ideal sparse representation

$$\boldsymbol{y} = [\boldsymbol{\Psi}, \boldsymbol{\Phi}]\boldsymbol{x} \tag{7.2}$$

Theorem 7.7 (Donoho & Huo '01, Elad & Bruckstein '02)

Any two distinct solutions $\boldsymbol{x}^{(1)}$ and $\boldsymbol{x}^{(2)}$ to (7.2) must satisfy

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

Corollary 7.8 (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 7.7

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 7.6,

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

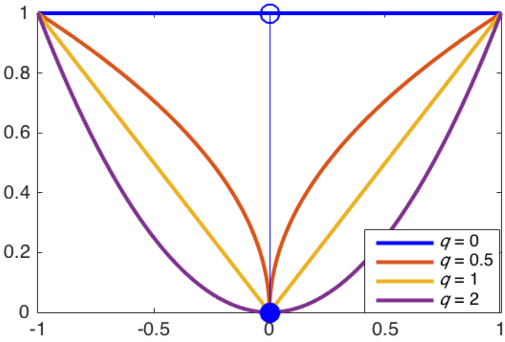
3. $\|\boldsymbol{x}^{(1)}\|_0 + \|\boldsymbol{x}^{(2)}\|_0 \geq \|\boldsymbol{h}\|_0 \geq \frac{2}{\mu(\boldsymbol{\Psi}, \boldsymbol{\Phi})}$ as claimed

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)

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

$$\Downarrow$$

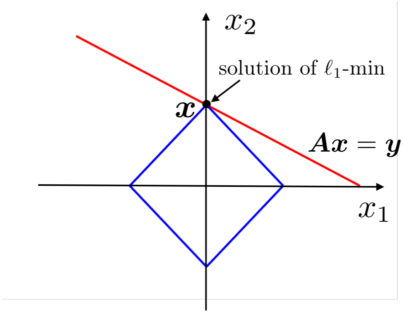
$$\text{convexifying } \|\boldsymbol{x}\|_0 \text{ with } \|\boldsymbol{x}\|_1$$

$$\Downarrow$$

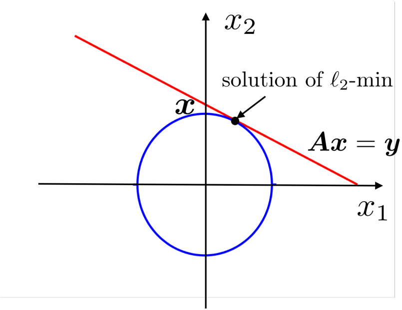
$$\text{minimize}_{\boldsymbol{x}\in\mathbb{C}^p}\ \|\boldsymbol{x}\|_1 \quad \text{s.t. } \boldsymbol{A}\boldsymbol{x}=\boldsymbol{y} \tag{7.3}$$

- $|x|$ 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. $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}$



$\min_{\boldsymbol{x}} \ \|\boldsymbol{x}\|_2$ s.t. $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}$

- 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 7.9 (Donoho & Huo '01, Elad & Bruckstein '02)

$\boldsymbol{x} \in \mathbb{C}^p$ *is the unique solution to* ℓ_1 *minimization (7.3) if*

$$\|\boldsymbol{x}\|_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(\boldsymbol{\Psi}, \boldsymbol{\Phi})}\right) \tag{7.4}$$

- ℓ_1 minimization yields the sparse solution too!
- recovery condition (7.4) 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

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

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

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 7.9

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

$$
\begin{aligned}
& \|\boldsymbol{x}+\boldsymbol{h}\|_1 > \|\boldsymbol{x}\|_1 \\
\Longleftarrow \quad & \sum_{i \notin \mathsf{supp}(\boldsymbol{x})} |h_i| + \sum_{i \in \mathsf{supp}(\boldsymbol{x})} \left(|x_i+h_i| - |x_i|\right) > 0 \\
\Longleftarrow \quad & \sum_{i \notin \mathsf{supp}(\boldsymbol{x})} |h_i| - \sum_{i \in \mathsf{supp}(\boldsymbol{x})} |h_i| > 0 \qquad (\text{since } |a+b|-|a| \geq -|b|) \\
\Longleftarrow \quad & \|\boldsymbol{h}\|_1 > 2 \sum_{i \in \mathsf{supp}(\boldsymbol{x})} |h_i| \\
\Longleftarrow \quad & \sum_{i \in \mathsf{supp}(\boldsymbol{x})} \frac{|h_i|}{\|\boldsymbol{h}\|_1} < \frac{1}{2} \\
\Longleftarrow \quad & \|\boldsymbol{x}\|_0 \frac{\|\boldsymbol{h}\|_\infty}{\|\boldsymbol{h}\|_1} < \frac{1}{2} \qquad (7.5)
\end{aligned}
$$

Proof of Theorem 7.9 (cont.)

It remains to control $\frac{\|\boldsymbol{h}\|_\infty}{\|\boldsymbol{h}\|_1}$. As usual, due to the 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

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

In addition, $\|\boldsymbol{h}_\psi\|_1 \geq |(\boldsymbol{h}_\psi)_i|$. Putting them together yields

$$\|\boldsymbol{h}\|_1 = \|\boldsymbol{h}_\phi\|_1 + \|\boldsymbol{h}_\psi\|_1 \geq |(\boldsymbol{h}_\psi)_i| \left(1 + \frac{1}{\mu(\boldsymbol{\Psi}, \boldsymbol{\Phi})}\right) \qquad (7.6)$$

Proof of Theorem 7.9 (cont.)

Similarly, this inequality (7.6) holds if we replace $(\boldsymbol{h}_\psi)_i$ by $(\boldsymbol{h}_\phi)_i$. As a consequence,

$$\frac{\|\boldsymbol{h}\|_\infty}{\|\boldsymbol{h}\|_1} \leq \frac{1}{1 + \frac{1}{\mu(\boldsymbol{\Psi}, \boldsymbol{\Phi})}} \tag{7.7}$$

Finally, if $\|\boldsymbol{x}\|_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(\boldsymbol{\Psi}, \boldsymbol{\Phi})}\right)$, then this together with (7.7) yields

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

as claimed in (7.5), 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?

Mutual coherence for arbitrary dictionaries

Definition 7.10 (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

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

- If $\|\boldsymbol{a}_i\|_2 = 1$ 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) $\mu(\boldsymbol{A}) \ge \sqrt{\frac{p-n}{n(p-1)}}$, 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 7.11 (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 the 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 information bottleneck!*

Summary

- For many dictionaries, if a signal is representable in a highly sparse manner, then it is often guaranteed to be unique sparse solution

- Seeking a sparse solution often becomes a well-posed question with interesting properties

Reference

- "*Sparse and redundant representations: from theory to applications in signal and image processing*," M. Elad, *Springer*, 2010.
- "*Uncertainty principles and signal recovery*," D. Donoho and P. Stark, *SIAM Journal on Applied Mathematics*, 1989.
- "*Uncertainty principles and ideal atomic decomposition*," D. Donoho and X. Huo, *IEEE Trans. on Info. Theory,* 2001.
- "*A generalized uncertainty principle and sparse representation in pairs of bases*," M. Elad and A. Bruckstein, *IEEE Trans. on Info. Theory,* 2002.
- "*Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization*," D. Donoho, and M. Elad, *Proceedings of the National Academy of Sciences,* 2003.

Reference

- "*High-dimensional data analysis with sparse models: Theory, algorithms, and applications*," J. Wright, Y. Ma, and A. Yang, 2018.
- "*Sparsity and incoherence in compressive sampling*," E. Candes, and J. Romberg, *Inverse Problems,* 2007.
- "*Atomic decomposition by basis pursuit*," S. Chen, D. Donoho, M. A. Saunders, *SIAM review,* 2001.
- "*On sparse representations in arbitrary redundant bases*," J. Fuchs, *IEEE Trans. on Info. Theory,* 2004.
- "*Sparse representations in unions of bases*," R. Gribonval, and M. Nielsen, *IEEE Trans. on Info. Theory,* 2003.