Compressed Sensing

LECTURE #1-2
Motivation & geometric insights

Prof. Dr. Volkan Cevher
volkan.cevher@epfl.ch
LIONS/Laboratory for Information and Inference Systems
**Motivation:**

- solve bigger / more important problems
- decrease acquisition times / costs
- entertainment / new consumer products...

**Major trends**

- higher resolution / denser sampling
- large numbers of sensors
- increasing # of modalities / mobility

160MP
Problems of the current paradigm

• **Sampling at Nyquist rate**
  - expensive / difficult

• **Data deluge**
  - communications / storage

• **Sample then compress**
  - inefficient / impossible / not future proof
Recommended for you: A more familiar example

- Recommender systems
  - observe partial information
    - “ratings”
    - “clicks”
    - “purchases”
    - “compatibilities”
• Recommender systems
  – observe partial information
    “ratings”
    “clicks”
    “purchases”
    “compatibilities”

• The Netflix problem
  – from approx. 100,000,000 ratings predict 3,000,000 ratings
  – 17770 movies x 480189 users
  – how would you automatically predict?
• Recommender systems
  – observe partial information
    “ratings”
    “clicks”
    “purchases”
    “compatibilities”

• The Netflix problem
  – from approx. 100,000,000 ratings predict 3,000,000 ratings
  – 17770 movies x 480189 users
  – how would you automatically predict?
  – what is it worth?
Theoretical set-up

- Matrix completion for Netflix

$$X = \begin{array}{c}
\text{users} \\
\end{array} \rightarrow \begin{array}{c}
\text{movies} \\
\end{array}$$
Theoretical set-up

- Matrix completion for Netflix

\[ X = \Phi(X) + n \]

- Mathematical underpinnings: \textit{compressive sensing}

\textit{CS: when we have less samples than the ambient dimension}
Linear Inverse Problems

\[ u \begin{bmatrix} M \times 1 \end{bmatrix} = \Phi \begin{bmatrix} M \times N \ (M < N) \end{bmatrix} x \begin{bmatrix} N \times 1 \end{bmatrix} \]

Myriad applications involve linear dimensionality reduction
deconvolution to data mining
compression to compressive sensing
geophysics to medical imaging

[Baraniuk, C, Wakin 2010; Carin et al. 2011]
Linear Inverse Problems

- **Challenge:** Null space of $\Phi$: $\mathcal{N}(\Phi)$

  $$\Phi x' = \Phi(x + v) = u, \ \forall v \in \mathcal{N}(\Phi)$$
## Linear Inverse Problems

<table>
<thead>
<tr>
<th></th>
<th>Deterministic</th>
<th>Probabilistic</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Prior</strong></td>
<td>$\superscript{\text{parsity}}$</td>
<td>distribution</td>
</tr>
<tr>
<td><strong>Metric</strong></td>
<td>$\ell_p$-norm*</td>
<td>likelihood/posterior</td>
</tr>
</tbody>
</table>

* $: \|x\|_p = \left(\sum_i |x_i|^p\right)^{1/p}$
Deterministic Low-Dimensional Models
Sparse representations

- **Sparse** signal $\alpha$
  
  only $K$ out of $N$ coordinates nonzero
  
  $K \ll N$

$X = \begin{bmatrix} x_1 & \cdots & x_N \end{bmatrix} \in \mathbb{R}^{N \times 1}$

$S = \{ i : x_i \neq 0 \}$

$\|\alpha\|_0 = |S| = K$

$K = 2$

$\mathbb{R}^3$

$\alpha \in \Sigma_2$

$|\alpha_i|$
Sparse representations

- **Sparse** signal $\mathbf{x}$
  - only $K$ out of $N$ coordinates nonzero in an *appropriate representation*

- Sparse representations
  - *Sparse* transform coefficients $\alpha$
    - Basis representations
      - $\Psi \in \mathbb{R}^{N \times N}$
        - Wavelets, DCT...
    - Frame representations
      - $\Psi \in \mathbb{R}^{N \times L}$, $L > N$
        - Gabor, curvelets, shearlets...
    - Other *dictionary* representations...
Sparse representations

- Sparse signal:
  - only $K$ out of $N$ coordinates nonzero
  
  $$K \ll N$$

- Sparse representations:
  - *sparse* transform coefficients

- A fundamental impact:
Sparse representations

- Sparse signal:
  only $K$ out of $N$ coordinates nonzero

\[ K \ll N \]

- Sparse representations:
  *sparse* transform coefficients

- A fundamental impact:
Sparse representations

• Sparse signal:
  
  only $K$ out of $N$
  coordinates nonzero

  \[ K \ll N \]

• Sparse representations:

  *sparse transform coefficients*

• A fundamental impact:
Sparse representations

• Sparse signal:
  only $K$ out of $N$ coordinates nonzero

  \[ K \ll N \]

• Sparse representations:
  \textit{sparse} transform coefficients

• A fundamental impact:
  \[ \Phi \]

  becomes effectively low dimensional*

  \[ M \times K \]

\[ x = \Psi \times \alpha \]

\[ u = \Phi' \alpha \]

\[ M > K \]

*: If we knew the locations of the coefficients. \textit{More on this later.}
Low-dimensional signal models

$N$ pixels

$sparse$ signals

$\mathbb{R}^N$

$K \ll N$

large wavelet coefficients (blue = 0)

$\mathbb{R}^N$

low-rank matrices

Information level:

nonlinear models
Low-dimensional signal models

- These lectures

Low-dimensional models based on linear representations

sparse signals

low-rank matrices

nonlinear models
Linear representation of low-dimensional models

• A key notion in sparse representation

  – synthesis of the signal using a few vectors

• A slightly different mathematical formalism for generalization

Synthesis model: \[ x = \sum_{i=1}^{\vert \mathcal{A} \vert} a_i c_i \]

\( a_i \in \mathcal{A}, c_i \geq 0 \)

\( a_i \): atoms
\( \mathcal{A} \): atomic set

i.e., linear (positive) combination of elements from an atomic set

[Chandrasekaran et al. 2010]
Linear representation of low-dimensional models

- A key notion in sparse representation
  - synthesis of the signal using a few vectors

- Sparse representations via the atomic formulation
  - Example:
    \[
    x = \sum_{i=1}^{\|A\|} a_i c_i
    \]
    \[a_i \in A, c_i \geq 0\]
    \[a_i: \text{ atoms}\]
    \[A: \text{ atomic set}\]
    \[
    \Psi = [\psi_1, \ldots, \psi_L]
    \]
    \[\text{rank}(\Psi) = N\]
    \[
    \mathcal{A} = \{\psi_1, \ldots, \psi_L, -\psi_1, \ldots, -\psi_L\}
    \]
    \[c_i = \begin{cases} 
    \alpha_i, & \alpha_i > 0; \\
    0, & \text{otherwise.}
    \end{cases} \quad i = 1, \ldots, L
    \]
    \[c_{i+L} = \begin{cases} 
    -\alpha_i, & \alpha_i < 0; \\
    0, & \text{otherwise.}
    \end{cases}
    \]
Linear representation of low-dimensional models

- Basic definitions on **low-dimensional atomic representations**

\[ x = \sum_{i=1}^{\left| A \right|} a_i c_i \]

\[ a_i \in A, c_i \geq 0 \]

\[ \|c_i\|_0 \leq K \]

\[ K \ll N \]
Linear representation of low-dimensional models

- Basic definitions on low-dimensional atomic representations

\[ x = \sum_{i=1}^{\left| \mathcal{A} \right|} a_i c_i \quad a_i \in \mathcal{A}, c_i \geq 0 \]

\[ \| c_i \|_0 \leq K \quad K \ll N \]

- \( \text{conv}(\mathcal{A}) \): convex hull of atoms in \( \mathcal{A} \)

\[ \mathcal{A} = \left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ -1 \end{bmatrix} \right\} \]

\[ \text{conv}(\mathcal{A}) = \left\{ \sum_{i} a_i \beta_i : a_i \in \mathcal{A}, \beta_i \in \mathbb{R}_+, \sum_{i=1}^{n} \beta_i = 1, n = 1, 2, \ldots, |\mathcal{A}| \right\} \]
Linear representation of low-dimensional models

- Basic definitions on low-dimensional atomic representations

\[ x = \sum_{i=1}^{|A|} a_i c_i \]

- \( a_i \in A, c_i \geq 0 \)
- \( \|c_i\|_0 \leq K \)

\[ K \ll N \]

\( \text{conv}(A) : \) convex hull of atoms in \( A \)

\[ A = \left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ -1 \end{bmatrix} \right\} \]

\[ \text{conv}(A) = \left\{ \sum_i a_i \beta_i : a_i \in A, \beta_i \in \mathbb{R}_+, \sum_{i=1}^{n} \beta_i = 1, n = 1, 2, \ldots, |A| \right\} \]
Linear representation of low-dimensional models

- Basic definitions on low-dimensional atomic representations

$$x = \sum_{i=1}^{|A|} a_i c_i$$

- $a_i \in A$, $c_i \geq 0$
- $\|c_i\|_0 \leq K$

- $\text{conv}(A)$: convex hull of atoms in $A$

$$A = \left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ -1 \end{bmatrix} \right\}$$

- $\|x\|_A$: atomic norm*

$$\|x\|_A = \inf\{t > 0 : x \in t \times \text{conv}(A)\}$$

*: requires $A$ to be centrally symmetric
Linear representation of low-dimensional models

- Basic definitions on low-dimensional atomic representations

\[ x = \sum_{i=1}^{\|A\|} a_i c_i \]

\[ a_i \in A, c_i \geq 0 \]

\[ \|c_i\|_0 \leq K \]

- \( \text{conv}(A) \): convex hull of atoms in \( A \)

\[ A = \left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ -1 \end{bmatrix} \right\} \]

- \( \|x\|_A \): atomic norm*

\[ \|x\|_A = \inf\{ t > 0 : x \in t \times \text{conv}(A) \} \]

\[ \|x\|_A = \frac{6}{5} \]

*: requires \( A \) to be centrally symmetric
Linear representation of low-dimensional models

- Basic definitions on low-dimensional atomic representations

\[ x = \sum_{i=1}^{\lvert A \rvert} a_i c_i \]

\[ a_i \in A, c_i \geq 0 \]

\[ \lVert c_i \rVert_0 \leq K \]

- \( \text{conv}(A) \): convex hull of atoms in \( A \)

\[ A = \left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ -1 \end{bmatrix} \right\} \]

- \( \lVert x \rVert_A \): atomic norm*

\[ \lVert x \rVert_A = \inf \{ t > 0 : x \in t \times \text{conv}(A) \} \]

\[ x = \begin{bmatrix} -1/5 \\ 1 \end{bmatrix} \]

\[ \lVert x \rVert_A = \frac{6}{5} \]

Alternative:

\[ \lVert x \rVert_A = \inf \left\{ \sum_{i=1}^{\lvert A \rvert} c_i : x = \sum_{i=1}^{\lvert A \rvert} a_i c_i, c_i \geq 0, \forall a_i \in A \right\} \]

*: requires \( A \) to be centrally symmetric
Linear representation of low-dimensional models

Examples with easy forms:

- **sparse vectors**
  \[ A = \{ \pm e_i \}_{i=1}^{N} \]
  \[ \text{conv}(A) = \text{cross-polytope} \]
  \[ \|x\|_A = \|x\|_1 \]

- **low-rank matrices**
  \[ A = \{ A : \text{rank}(A) = 1, \|A\|_F = 1 \} \]
  \[ \text{conv}(A) = \text{nuclear norm ball} \]
  \[ \|x\|_A = \|x\|_* \]

- **binary vectors**
  \[ A = \{ \pm 1 \}^N \]
  \[ \text{conv}(A) = \text{hypercube} \]
  \[ \|x\|_A = \|x\|_\infty \]
Linear representation of low-dimensional models

Examples with easy forms:

- **sparse vectors**
  \[ A = \{ \pm e_i \}_{i=1}^{N} \]

Examples with no-so-easy forms:

- A: infinite set of unit-norm rank-one tensors
- A: finite (but large) set of permutation matrices
- A: infinite set of orthogonal matrices
- A: infinite set of matrices constrained by eigenvalues
- A: infinite set of measures
- A: finite (but large) set of cut matrices

\[ \text{conv}(A) = \text{hypercube} \]
\[ \| x \|_A = \| x \|_\infty \]

[Chandrasekaran et al. 2010]
Linear representation of low-dimensional models

Pop-quiz:

What is $\|x\|_A$?

$$\|x\|_A = \inf\{t > 0 : x \in t \times \text{conv}(A)\}$$
Linear representation of low-dimensional models

Pop-quiz:

HINT:

\[ \mathcal{A} = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \\ 0 \end{bmatrix}, \|x_G\|_2 = 1 \right\} \]

\[ G = \{2, 3\} \]

What is \( \|x\|_{\mathcal{A}} \)?

\[ \|x\|_{\mathcal{A}} = \inf\{t > 0 : x \in t \times \text{conv}(\mathcal{A})\} \]
Linear representation of low-dimensional models

Pop-answer:

\[ \mathcal{A} = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \\ 0 \end{bmatrix} \mid \|x_G\|_2 = 1 \right\} \]

What is \( \|x\|_\mathcal{A} \)?

\[ \|x\|_\mathcal{A} = |x_1| + \|x_G\|_2 \]

\( G = \{2, 3\} \)
Towards algorithms: a geometric perspective

Other key concepts:

- Cone $C$: $x, y \in C \Rightarrow tx + \omega y \in C, \forall t, \omega \in \mathbb{R}_+$
Towards algorithms: a geometric perspective

Other key concepts:

• Cone $\mathcal{C}$: $x, y \in \mathcal{C} \Rightarrow tx + \omega y \in \mathcal{C}, \forall t, \omega \in \mathbb{R}_+$

• Tangent cone of $x^*$ with respect to $\|x^*\|_{\mathcal{A}\text{conv}(\mathcal{A})}$:

$$T_{\mathcal{A}}(x^*) = \text{cone}\{z - x^* : \|z\|_{\mathcal{A}} \leq \|x^*\|_{\mathcal{A}}\}$$

$\|x^*\|_{\mathcal{A}\text{conv}(\mathcal{A})}$
Towards algorithms: a geometric perspective

Other key concepts:

- Cone $\mathcal{C}$: $x, y \in \mathcal{C} \Rightarrow tx + \omega y \in \mathcal{C}, \forall t, \omega \in \mathbb{R}_+$

- Tangent cone of $x^*$ with respect to $\|x^*\|_{\mathcal{A}\text{conv}(\mathcal{A})}$:

$$T_{\mathcal{A}}(x^*) = \text{cone}\{z - x^* : \|z\|_{\mathcal{A}} \leq \|x^*\|_{\mathcal{A}}\}$$

Tangent cone is the set of descent directions where you do not increase the atomic norm.
Towards algorithms: a geometric perspective

Other key concepts:

- Cone $\mathcal{C}$: $x, y \in \mathcal{C} \Rightarrow tx + \omega y \in \mathcal{C}, \forall t, \omega \in \mathbb{R}_+$

- Tangent cone of $x^*$ with respect to $\|x^*\|_{\mathcal{A}\text{conv}(\mathcal{A})}$:

$$T_{\mathcal{A}}(x^*) = \text{cone}\{z - x^* : \|z\|_{\mathcal{A}} \leq \|x^*\|_{\mathcal{A}}\}$$

Tangent cone is the set of descent directions where you do not increase the atomic norm.
Towards algorithms: a geometric perspective

Null space of $\Phi$: $\mathcal{N}(\Phi)$

$\Phi v = 0, \forall v \in \mathcal{N}(\Phi)$
Towards algorithms: a geometric perspective

\[ u \quad \Phi \quad x^* \]

\[ M \times 1 \quad M \times N \quad M < N \quad N \times 1 \]

\[ \mathcal{N}(\Phi) \]

\[ x^* \]
Towards algorithms: a geometric perspective

Consider the criteria:

\[ \hat{x} = \arg \min_{x: u = \Phi x} \|x\|_A \]
Towards algorithms: a geometric perspective

Consider the criteria:

\[
\hat{x} = \arg \min_{x: u = \Phi x} \|x\|_A
\]
Towards algorithms: a geometric perspective

Consider the criteria:

\[
\hat{x} = \arg \min_{x:u=\Phi x} \|x\|_A
\]
Towards algorithms: a geometric perspective

Consider the criteria:

\[
\hat{x} = \arg \min_{x : u = \Phi x} \| x \|_A
\]
Towards algorithms: a geometric perspective

Key observation:

\[ \mathcal{N}(\Phi) \cap T_A(x^*) = \{0\} \Rightarrow x^* = \arg \min_{x : u = \Phi x} \|x\|_A \]
Towards algorithms: a geometric perspective

How about noise?

\[ \| u - \Phi x \| \leq \sigma \]

\[ \| x^* \|_{\mathcal{A}\text{conv}(\mathcal{A})} \]

\[ T_{\mathcal{A}}(x^*) \]

\[ \mathcal{N}(\Phi) \]

\[ \hat{x} = \arg \min_{x : \| u - \Phi x \| \leq \sigma} \| x \|_{\mathcal{A}} \]

\[ u \]

\[ \Phi \]

\[ x^* \]

\[ n \]

\[ \| u \|_{M \times 1} \]

\[ \| \Phi \|_{M \times N (M < N)} \]

\[ \| x^* \|_{N \times 1} \]

\[ \| n \| \leq \sigma \]
Towards algorithms: a geometric perspective

How about noise?

Stability assumption:

\[ \| \Phi v \| \geq \epsilon \| v \|, \forall v \in T_{A}(x^*) \]
Towards algorithms: a geometric perspective

How about noise?

$u \Phi x^* + n$

$M \times 1 \quad M \times N \quad M \times 1$

$N \times 1$

R$^N$

$x^*$

$T_A(x^*)$

$N(\Phi)$

Stability assumption:

$\|\Phi v\| \geq \epsilon \|v\|, \forall v \in T_A(x^*)$

Note that if $N(\Phi) \cap T_A(x^*) = \{0\}$

$\Rightarrow \|\Phi v\| > 0, \forall v \neq 0$
Towards algorithms: a geometric perspective

How about noise?

Stability assumption:
\[ \|\Phi v\| \geq \epsilon \|v\|, \forall v \in T_A(x^*) \]

want epsilon large to minimize overlap between \(\|x^*\|_A\) and \(\|u - \Phi x\| \leq \sigma\)

For this 2D example: \(\|\Phi v\| \geq \|v\| \sin(\varphi) \min_i \|\Phi(i,:)\|\)

Matlab notation

\[ u = \Phi x^* + n \]

\(M \times 1\) \(M \times N (M < N)\) \(M \times 1\)

\(N \times 1\)
Towards algorithms: a geometric perspective

How about noise?

Stability assumption:
\[ \|\Phi v\| \geq \epsilon\|v\|, \forall v \in T_A(x^*) \]

\[ \hat{x} = \arg\min_{x:\|u - \Phi x\| \leq \sigma} \|x\|_A \]

\[ \Rightarrow \|x^* - \hat{x}\| \leq \frac{2\sigma}{\epsilon} \]
Towards algorithms: a geometric perspective

Can we guarantee the following?*

\[ \mathcal{N}(\Phi) \cap T_A(x^*) = \{0\} \]

*without knowing \( x^* \)
Towards algorithms: a geometric perspective

Can we guarantee the following?*

\[ \mathcal{N}(\Phi) \cap T_A(x^*) = \{0\} \]

YES: with randomized measurements!

Gordon’s Minimum Restricted Singular Values Theorem has a probabilistic characterization.

\[ \text{Prob}(\min_v \|\Phi v\| \geq \epsilon) \]

∀v ∈ T_A(x^*), \|v\| = 1

*without knowing x^*
Towards algorithms: a geometric perspective

Can we guarantee the following?*

\[ \mathcal{N}(\Phi) \cap T_A(x^*) = \{0\} \]

\[ u \quad \Phi \quad x^* \]

Gordon’s Minimum Restricted Singular Values Theorem has a probabilistic characterization.

Key concept: **width of the tangent cone!**

*without knowing \( x^* \)
Towards algorithms: a geometric perspective

Can we guarantee the following?*

\[ \mathcal{N}(\Phi) \cap T_{A}(x^*) = \{0\} \]

Gordon’s Minimum Restricted Singular Values Theorem has a probabilistic characterization.

Gaussian width of \( S \subseteq \mathbb{R}^M \)

\[ w(S) = \mathbb{E} \left[ \sup_{z \in S} g^T z \right] ; \ g \sim \mathcal{N}(0, I) \]

\( \lambda_k \) expected norm of a \( k \)-dimensional Gaussian random vector:

\[ \lambda_k = \sqrt{\mathbb{E} \left[ \sum_{i=1}^{k} g_i^2 \right]} = \frac{\sqrt{2} \Gamma((k + 1)/2)}{k/2} \]

*without knowing \( x^* \)
Can we guarantee the following?*

\[ \mathcal{N}(\Phi) \cap T_A(x^*) = \{0\} \]

Gordon’s Minimum Restricted Singular Values Theorem has a probabilistic characterization.

Let \( \Omega \) be a closed subset of the unit sphere and \( A \) be an \( M \times N \) matrix with iid \( \mathcal{N}(0, 1) \) entries. Then, if \( \lambda_k \geq w(\Omega) + \epsilon \):

\[
P \left[ \min_{z \in \Omega} \|Az\|_2 \geq \epsilon \right] \geq 1 - \frac{1}{2} \frac{1}{\epsilon} \exp \left( \frac{-1}{8} (\lambda_k - w(\Omega) \epsilon)^2 \right)
\]

*without knowing \( x^* \)
Towards algorithms: a geometric perspective

Can we guarantee the following?*

\[ \mathcal{N}(\Phi) \cap T_A(x^*) = \{0\} \]

\[ \Phi \sim \text{iid } \mathcal{N}(0, 1/M), \Omega = T_A(x^*) \cap \mathbb{S}^{N-1} \]

Let \( \Omega \) be a closed subset of the unit sphere and \( A \) be an \( M \times N \) matrix with iid \( \mathcal{N}(0, 1) \) entries. Then, if \( \lambda_k \geq w(\Omega) + \epsilon \):

\[ P \left[ \min_{z \in \Omega} \|Az\|_2 \geq \epsilon \right] \geq 1 - \frac{1}{2} e^{-\frac{1}{18} (\lambda_k - w(\Omega) - \epsilon)^2} \]

*without knowing \( x^* \)

Gordon’s Minimum Restricted Singular Values Theorem has a probabilistic characterization.
Towards algorithms: a geometric perspective

Key observation:

\[ \mathcal{N}(\Phi) \cap T_\mathcal{A}(x^*) = \{0\} \implies x^* = \arg\min_{x : u = \Phi x} \|x\|_\mathcal{A} \]

\[ M \geq w(\Omega)^2 + O(1) \]
Towards algorithms: a geometric perspective

How about noise?

\[ u \Phi + x^* + n \]

\[ M \times 1 \quad M \times N \quad M \times 1 \]

\[ M \times 1 \quad M < N \quad N \times 1 \]

Stability assumption:

\[ \| \Phi v \| \geq \epsilon \| v \|, \forall v \in T_A(x^*) \]

\[ \hat{x} = \arg \min_{x : \| u - \Phi x \| \leq \sigma} \| x \|_A \]

\[ \Rightarrow \| x^* - \hat{x} \| \leq \frac{2\sigma}{\epsilon} \]

\[ M \geq \frac{w(\Omega)^2}{(1-\epsilon)^2} + O(1) \]
Towards algorithms: a geometric perspective

Can we guarantee the following?*

\[ \mathcal{N}(\Phi) \cap T_A(x^*) = \{0\} \]

Gordon’s Minimum Restricted Singular Values Theorem has a probabilistic characterization.

\[ g \sim_{\text{iid}} \mathcal{N}(0, 1) \]

\[ \Phi \sim_{\text{iid}} \mathcal{N}(0, 1/M), \Omega = T_A(x^*) \cap S^{N-1} \]

\[ w(T_A(x^*) \cap S^{N-1}) \leq E_g [\text{dist} (g, T_A(x^*))] \]

\[ w^2(T_A(x^*) \cap S^{N-1}) + w^2(T_A(x^*) \cap S^{N-1}) \leq N \]

\[ N \geq 9 \]

\[ w(T_A(x^*) \cap S^{N-1}) \leq \sqrt{\log \left( \frac{4}{\text{vol}(T_A(x^*) \cap S^{N-1})} \right)} \]

*without knowing \( x^* \)
Towards algorithms: a geometric perspective

Can we guarantee the following?*

\[ \mathcal{N}(\Phi) \cap T_A(x^*) = \{0\} \]

\[ A = \left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ -1 \end{bmatrix} \right\} \]

\[ \mathcal{N}(\Phi) \cap T_A(x^*) = \{0\} \text{ w.p. } 1/2 \]

\[ \Rightarrow x^* = \arg \min_{x:u=\Phi x} \|x\|_1 \]

*without knowing 1-sparse \( x^* \) and 1-random measurement
Towards algorithms: a geometric perspective

Can we guarantee the following?*

\[ \mathcal{N}(\Phi) \cap T_{\mathcal{A}}(x^*) = \{0\} \]

\[ \mathcal{A} = \left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ -1 \end{bmatrix} \right\} \]

\[ \mathcal{N}(\Phi) \cap T_{\mathcal{A}}(x^*) = \{0\} \text{ w.p. } 1/2 \]

\[ \Rightarrow x^* = \arg \min_{x:u=\Phi x} \|x\|_1 \]

\[ \bar{\mathcal{A}} = \left\{ \begin{bmatrix} \sqrt{3}/2 \\ 1/2 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} -\sqrt{3}/2 \\ 1/2 \end{bmatrix}, \begin{bmatrix} -\sqrt{3}/2 \\ -1/2 \end{bmatrix}, \begin{bmatrix} 0 \\ -1 \end{bmatrix}, \begin{bmatrix} \sqrt{3}/2 \\ -1/2 \end{bmatrix} \right\} \]

\[ \mathcal{N}(\Phi) \cap T_{\bar{\mathcal{A}}}(x^*) = \{0\} \text{ w.p. } 1/3 \]

\[ \Rightarrow x^* = \arg \min_{x:u=\Phi x} \|x\|_{\bar{\mathcal{A}}} \]

*without knowing 1-sparse \(x^*\) and 1-random measurement
Towards algorithms: a geometric perspective

Can we guarantee the following?*

\[ \mathcal{N}(\Phi) \cap T_{A}(x^*) = \{0\} \]

\[ A = \left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ -1 \end{bmatrix} \right\} \]

\[ \mathcal{N}(\Phi) \cap T_{A}(x^*) = \{0\} \text{ w.p. } 1/2 \]

\[ \Rightarrow x^* = \arg \min_{x:u=\Phi x} \|x\|_1 \]

\[ \tilde{A} = \left\{ \begin{bmatrix} \sqrt{3}/2 \\ 1/2 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} -\sqrt{3}/2 \\ 1/2 \end{bmatrix}, \begin{bmatrix} -\sqrt{3}/2 \\ -1/2 \end{bmatrix}, \begin{bmatrix} 0 \\ -1 \end{bmatrix}, \begin{bmatrix} \sqrt{3}/2 \\ -1/2 \end{bmatrix} \right\} \]

\[ \mathcal{N}(\Phi) \cap T_{\tilde{A}}(x^*) = \{0\} \text{ w.p. } 1/3 \]

\[ \Rightarrow x^* = \arg \min_{x:u=\Phi x} \|x\|_{\tilde{A}} \]

\[ \tilde{\mathcal{A}} = \{\|x\|_2 = 1\} \]

\[ \mathcal{N}(\Phi) \cap T_{\tilde{\mathcal{A}}}(x^*) = \{0\} \text{ w.p. } 0 \]

\[ \Rightarrow x^* = \arg \min_{x:u=\Phi x} \|x\|_2 \]

*without knowing 1-sparse \( x^* \) and 1-random measurement
Towards algorithms: a geometric perspective

Can we guarantee the following?*

\[ \mathcal{N}(\Phi) \cap T_A(x^*) = \{0\} \]

A projected 6D hypercube with 64 vertices

Blessing-of-dimensionality!

http://www.agrell.info/erik/chalmers/hypercubes/
Towards algorithms: a geometric perspective

Pop-quiz:

$$\mathcal{N}(\Phi) \cap T_A(x^*) = \{0\}$$

$$\bar{A} = \left\{ \begin{bmatrix} \sqrt{3}/2 \\ 1/2 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} -\sqrt{3}/2 \\ 1/2 \end{bmatrix}, \begin{bmatrix} -\sqrt{3}/2 \\ -1/2 \end{bmatrix}, \begin{bmatrix} 0 \\ -1 \end{bmatrix}, \begin{bmatrix} \sqrt{3}/2 \\ -1/2 \end{bmatrix} \right\}$$

$$\mathcal{N}(\Phi) \cap T_{\bar{A}}(x^*) = \{0\} \text{ w.p. } ???$$

$$\Rightarrow x^* = \arg \min_{x:u=\Phi x} \|x\|_{\bar{A}}$$

What is the probability that we can determine a 2-sparse $x^*$ with 1-random measurement?
Towards algorithms: a geometric perspective

Pop-answer:

\[ \mathcal{N}(\Phi) \cap T_A(x^*) = \{0\} \]

\[ \tilde{A} = \left\{ \begin{bmatrix} \sqrt{3}/2 \\ 1/2 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} -\sqrt{3}/2 \\ 1/2 \end{bmatrix}, \begin{bmatrix} -\sqrt{3}/2 \\ -1/2 \end{bmatrix}, \begin{bmatrix} 0 \\ -1 \end{bmatrix}, \begin{bmatrix} \sqrt{3}/2 \\ -1/2 \end{bmatrix} \right\} \]

\[ \mathcal{N}(\Phi) \cap T_{\tilde{A}}(x^*) = \{0\} \text{ w.p. 0} \]

\[ \Rightarrow x^* = \arg \min_{x: u = \Phi x} \|x\|_{\tilde{A}} \]

Tangent cone is too wide!
Need at least 2 measurements!
Take home messages

<table>
<thead>
<tr>
<th>Underlying Model</th>
<th>Atomic Norm</th>
<th>Gaussian Measurements</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K$-sparse vector in $\mathbb{R}^N$</td>
<td>$\ell_1$-norm</td>
<td>$(2K + 1) \log(N - K)$</td>
</tr>
<tr>
<td>$N \times N$ rank-$R$ matrix</td>
<td>nuclear norm</td>
<td>$3R(2N - R) + 2(N - R - R^2)$</td>
</tr>
<tr>
<td>sign vector ${\pm 1}^N$</td>
<td>$\ell_\infty$-norm</td>
<td>$N/2$</td>
</tr>
<tr>
<td>$N \times N$-perm. matrix</td>
<td>Birkhoff polytope norm</td>
<td>$9N \log(N)$</td>
</tr>
<tr>
<td>$N \times N$ orth. matrix</td>
<td>spectral norm</td>
<td>$(3N^2 - N)/4$</td>
</tr>
</tbody>
</table>

[Chandrasekaran et al. 2010]

convex polytope $\leftrightarrow$ atomic norm

- geometry (and algebra) of representations in high dimensions

geometric perspective $\leftrightarrow$ convex criteria

- convex optimization algorithms in high dimensions

tangent cone width $\leftrightarrow$ # of randomized samples

- probabilistic concentration-of-measures in high dimensions
Back to the initial example

- Matrix completion for Netflix
  
  $X = \begin{pmatrix}
  \text{users} & \text{movies}
  \end{pmatrix}$

  $M \times N$

- What is low-rank?

  $X = \begin{pmatrix}
  \text{users} & \text{movies}
  \end{pmatrix}$

  $R \times N$

  $M \times R$

  $R \ll \min\{M, N\}$
Back to the initial example

- Matrix completion for Netflix: 17,770 movies x 480,189 users

\[ X = \begin{pmatrix} \ldots \end{pmatrix} \]

\[ M \times N \]

- What does the simple low-rank assumption buy?

**Leaderboard**

<table>
<thead>
<tr>
<th>Rank</th>
<th>Team Name</th>
<th>Best Score</th>
<th>Improvement</th>
<th>Last Submit Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>The Ensemble</td>
<td>0.8553</td>
<td>10.10</td>
<td>2009-07-26 18:38:22</td>
</tr>
<tr>
<td>2</td>
<td>BellKor’s Pragmatic Chaos</td>
<td>0.8554</td>
<td>10.09</td>
<td>2009-07-26 18:18:28</td>
</tr>
<tr>
<td>3</td>
<td>Grand Prize Team</td>
<td>0.8571</td>
<td>9.91</td>
<td>2009-07-24 13:07:49</td>
</tr>
<tr>
<td>4</td>
<td>Opera Solutions and Vandelau United</td>
<td>0.8573</td>
<td>9.99</td>
<td>2009-07-25 20:05:52</td>
</tr>
<tr>
<td>5</td>
<td>Vandelau Industries</td>
<td>0.8576</td>
<td>9.83</td>
<td>2009-07-25 02:49:53</td>
</tr>
<tr>
<td>6</td>
<td>Pragmatic Theory</td>
<td>0.8582</td>
<td>9.80</td>
<td>2009-07-12 15:09:53</td>
</tr>
<tr>
<td>7</td>
<td>BellKor in BigChaos</td>
<td>0.8590</td>
<td>9.71</td>
<td>2009-07-26 12:57:25</td>
</tr>
<tr>
<td>8</td>
<td>Dace</td>
<td>0.8603</td>
<td>9.58</td>
<td>2009-07-24 17:18:43</td>
</tr>
<tr>
<td>9</td>
<td>Opera Solutions</td>
<td>0.8611</td>
<td>9.49</td>
<td>2009-07-20 18:02:08</td>
</tr>
<tr>
<td>10</td>
<td>BellKor</td>
<td>0.8612</td>
<td>9.48</td>
<td>2009-07-26 17:11:11</td>
</tr>
<tr>
<td>11</td>
<td>BigChaos</td>
<td>0.8613</td>
<td>9.47</td>
<td>2009-06-23 23:05:52</td>
</tr>
<tr>
<td>12</td>
<td>Feeds2</td>
<td>0.8613</td>
<td>9.47</td>
<td>2009-07-24 20:08:46</td>
</tr>
</tbody>
</table>

**Grand Prize** - RMSE <= 0.8563

Quite a lot of extrapolation power!
There are two types of people in this world:

Those who can extrapolate from incomplete data
There are two types of people in this world:

Those who can extrapolate from incomplete data and do this fast with theoretical guarantees
Sampling/sketching design

scene

single photon detector

random pattern on DMD array

- Structured random matrices
- 1-bit CS $u = \text{sign}(\Phi x)$
- expanders & extractors

+ Coding theory
+ Theoretical computer science
+ Learning theory
+ Databases
Structured recovery

- **Sparsity**

Sparse vector

only $K$ out of $N$ coordinates nonzero

$$K \ll N$$

$K = 2$

$\mathbb{R}^3$

$x \in \Sigma_2$
Structured recovery

- **Sparsity**

\[
\left| x_i \right| \\
K \quad \text{sorted index} \quad N
\]

**Structured sparse** vector

only certain K out of N coordinates nonzero

\[
K \ll N
\]
Structured recovery

- **Structured sparsity**

  + requires smaller sketches
  + enhanced recovery
  + faster recovery

  \[ P_{\Sigma_M}(u; K) \in \arg \min_x \{ \|x - u\| : x \in \Sigma_{MK} \} \]

  support of the solution \( <> \) modular approximation problem

  integer linear program

*matroid structured sparse models*

*clustered /diversified sparsity models*

tightly connected with max-cover, binpacking, knapsack problems

- Recovery with low-dimensional models, including low-rank...
Quantum tomography

• Quantum state estimation
  a state of $n$ possibly-entangled qubits takes $\sim 2^n$ bits to specify, even approximately

• Recovery with rank and trace constraints
  \[ \text{with } M = O(N) \]
  1. Create Pauli measurements (semi-random)
  2. Estimate $\text{Tr}(\Phi_i \rho)$ for each $1 \leq i \leq M$
  3. Find any “hypothesis state” $\sigma$ st $\text{Tr}(\Phi_i \sigma) \approx \text{Tr}(\Phi_i \rho)$ for all $1 \leq i \leq M$

• Huge dimensional problem!
  – (desperately) need scalable algorithms
  – also need theory for perfect density estimation

+Theoretical computer science
+Databases
+Information theory
+Optimization
A fundamental problem: given \((y_i, x_i): \mathbb{R} \times \mathbb{R}^d, i = 1, \ldots, m\), learn a mapping \(f: x \rightarrow y\).

Our interest: non-parametric functions, graphs (e.g., social networks), dictionary learning...

Rigorous foundations: sample complexity, approximation guarantees, tractability.

Key tools: sparsity/low-rankness, submodularity, smoothness.
Compressible priors

+ Learning theory
+ Statistics
+ Information theory

- **Goal:** seek distributions whose iid realizations \( x_i \sim p(x) \) can be well-approximated as *sparse*

**Definition:**

The PDF \( p(x) \) is a *q-compressible prior* with parameters \((\epsilon, \kappa)\), when

\[
\lim_{N \to \infty} \bar{\sigma}_{k_N}(x)_q^{a.s.} \leq \epsilon, \text{ (a.s.: almost surely)};
\]

for any sequence \( k_N \) such that \( \lim_{N \to \infty} \inf \frac{k_N}{N} \geq \kappa \), where \( \epsilon \ll 1 \) and \( \kappa \ll 1 \).

**relative k-term approximation:**

\[
\bar{\sigma}_k(x)_q = \frac{\sigma_k(x)_q}{\|x\|_q}
\]

\[
\sigma_k(x)_q := \inf_{\|u\|_0 \leq k} \|x - u\|_q
\]
Compressible priors

+Learning theory
+Information theory

• **Goal:** seek distributions whose iid realizations can be well-approximated as *sparse*

\[
\|x\|_{\ell^p} := \sup_i \left\{ |x(i)| \cdot i^{1/p} \right\} \leq R
\]

\[
\frac{\sigma_{\kappa N}(x)_q}{\|x\|_q} \leq \epsilon
\]

Classical:

New:
Compressible priors

• **Goal:** seek distributions whose iid realizations can be well-approximated as *sparse*

• **Motivations:**
  - deterministic embedding scaffold for the probabilistic view
  - analytical proxies for sparse signals
    - learning (e.g., dim. reduced data)
    - algorithms (e.g., structured sparse)
  - information theoretic (e.g., coding)

lots of applications in vision, image understanding / analysis

+Learning theory
+Statistics
+Information theory
References


References


References


References


