

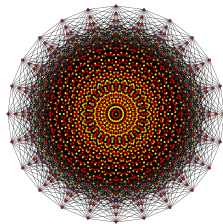
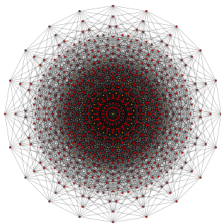
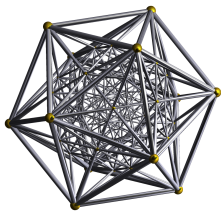
# Computational Principles for High-dim Data Analysis

## (Lecture Four)

**Yi Ma**

EECS Department, UC Berkeley

September 7, 2021



# Convex Methods for Sparse Signal Recovery

## ① Geometric Intuition

## ② A First Correctness Result via Incoherence

Coherence of a Matrix

Correctness of  $\ell^1$  Minimization

Constructing an Incoherent Matrix

Limitations of Incoherence

*“Algebra is but written geometry; geometry is but drawn algebra.”*  
– Sophie Germain

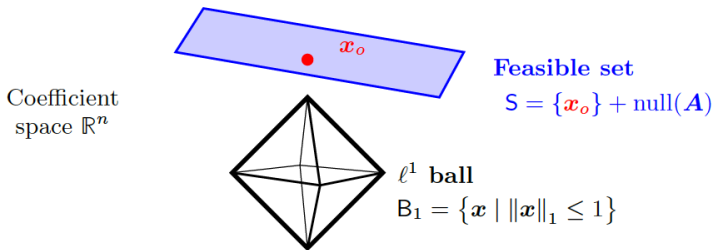
# Geometric Intuition: Coefficient Space

Given  $\mathbf{y} = \mathbf{A}\mathbf{x}_o \in \mathbb{R}^m$  with  $\mathbf{x}_o \in \mathbb{R}^n$  sparse:

$$\min \|\mathbf{x}\|_1 \quad \text{subject to} \quad \mathbf{A}\mathbf{x} = \mathbf{y}. \quad (1)$$

The space of all feasible solutions is an affine subspace:

$$\mathcal{S} = \{\mathbf{x} \mid \mathbf{A}\mathbf{x} = \mathbf{y}\} = \{\mathbf{x}_o\} + \text{null}(\mathbf{A}) \subset \mathbb{R}^n. \quad (2)$$

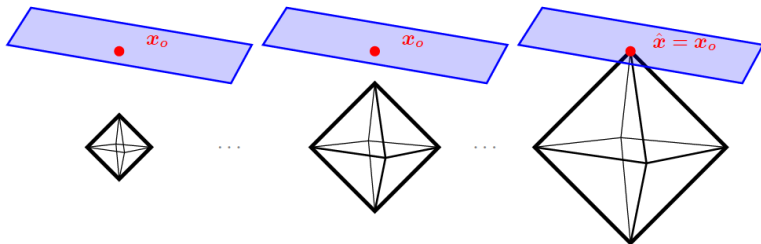


# $\ell^1$ Minimization in the Coefficient Space

Gradually expand a  $\ell^1$  ball of radius  $t$  from the origin  $\mathbf{0}$ :

$$t \cdot B_1 = \{\mathbf{x} \mid \|\mathbf{x}\|_1 \leq t\} \subset \mathbb{R}^n, \quad (3)$$

till its boundary first touches the feasible set  $S$ :



# Comparison between $\ell^1$ and $\ell^2$ Minimization

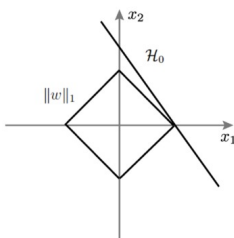
Given  $y = Ax_o$  with  $x_o$  sparse:

$$\mathbf{A} : \min \|x\|_1 \quad \text{subject to} \quad Ax = y. \quad (4)$$

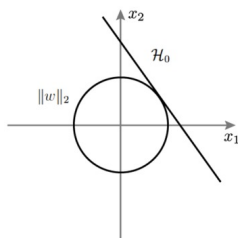
versus

$$\mathbf{B} : \min \|x\|_2 \quad \text{subject to} \quad Ax = y \quad (5)$$

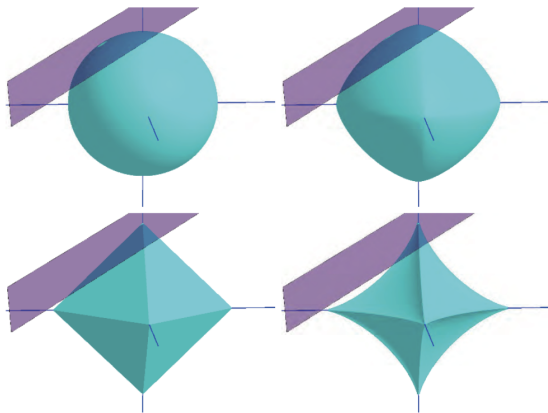
**A** L1 regularization



**B** L2 regularization



# Sparsity Promoting with Different $\ell^p$ Norms



**Figure:** Intersection between the  $\ell^p$ -ball and the feasible set  $S$ , for  $p = 2, 1.5, 1$  and  $0.7$ , respectively. (Some argue  $p = 0.5$  is somewhat special.)

Figure from *Sparse and Redundant Representations*, Michael Elad, Springer, 2010.

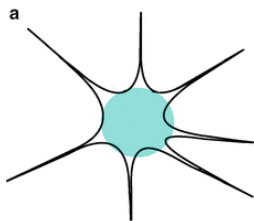
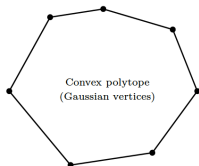
# Geometric Intuition: High-dimensional Polytopes

## Nearby Polytopes

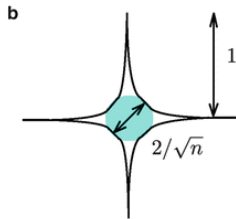
(vertices from a Gaussian matrix):

$$A = [a_1, a_2, \dots, a_n] \in \mathbb{R}^{m \times n}.$$

The “**correct**” visualization of high-dimensional convex polytopes,<sup>1</sup> including the  $\ell^1$  ball:



A general convex set



The  $\ell_1$  ball

<sup>1</sup>Lectures on Discrete Geometry, Jiri Matousek, Springer 2002.

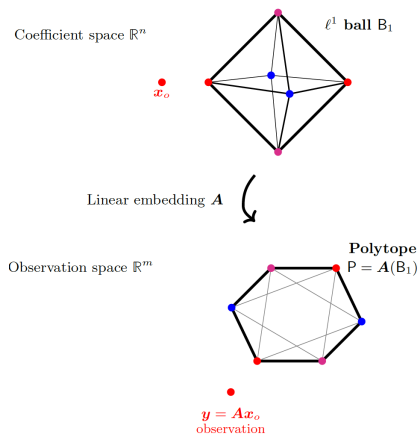
# Geometric Intuition: Observation Space

The matrix  $\mathbf{A} \in \mathbb{R}^{m \times n}$  can be viewed as a linear projection from  $\mathbb{R}^n$  to  $\mathbb{R}^m$ :

$$\mathbf{A} : B_1 \rightarrow P = \mathbf{A}(B_1), \quad (6)$$

which maps a convex polytope to a convex polytope. Similarly,  $\forall t \geq 0$ :

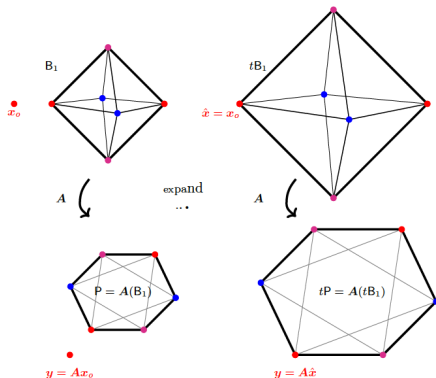
$$t \cdot B_1 \rightarrow t \cdot \mathbf{A}(B_1).$$





# Geometric Intuition: Observation Space

All  $k$ -faces of  $B_1$  cannot be mapped to the inside of the polytope  $A(B_1)$  :



**A Million Dollar Question: When  $\hat{x} = x_o$ ?**

# Coherence of a Matrix

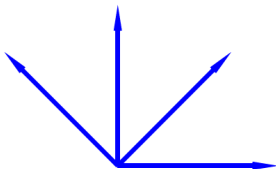
## Definition (Mutual Coherence)

For a matrix  $\mathbf{A} = [\mathbf{a}_1 \mid \mathbf{a}_2 \mid \cdots \mid \mathbf{a}_n] \in \mathbb{R}^{m \times n}$  with nonzero columns, the *mutual coherence*  $\mu(\mathbf{A})$  is the largest normalized inner product between two distinct columns:

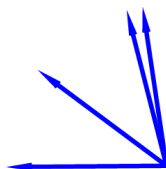
$$\mu(\mathbf{A}) = \max_{i \neq j} \left| \left\langle \frac{\mathbf{a}_i}{\|\mathbf{a}_i\|_2}, \frac{\mathbf{a}_j}{\|\mathbf{a}_j\|_2} \right\rangle \right|. \quad (7)$$

### Example:

$$\mu(\mathbf{A}) = 0.70711$$



$$\mu(\mathbf{A}) = 0.99488$$



# Uniqueness of Sparse Solution

## Proposition (Coherence Controls Kruskal Rank)

For any  $\mathbf{A} \in \mathbb{R}^{m \times n}$ ,

$$\text{krank}(\mathbf{A}) \geq \frac{1}{\mu(\mathbf{A})}. \quad (8)$$

In particular, if  $\mathbf{y} = \mathbf{A}\mathbf{x}_o$  and

$$\|\mathbf{x}_o\|_0 \leq \frac{1}{2\mu(\mathbf{A})}, \quad (9)$$

then  $\mathbf{x}_o$  is the unique optimal solution to the  $\ell^0$  minimization problem

$$\min \|\mathbf{x}\|_0 \quad \text{s.t.} \quad \mathbf{A}\mathbf{x} = \mathbf{y}. \quad (10)$$

**Proof:**

$$1 - k\mu(\mathbf{A}) < \sigma_{\min}(\mathbf{A}_l^* \mathbf{A}_l) \leq \sigma_{\max}(\mathbf{A}_l^* \mathbf{A}_l) < 1 + k\mu(\mathbf{A}). \quad (11)$$

# Correctness of $\ell^1$ Minimization

## Theorem ( $\ell^1$ Succeeds under Incoherence)

Let  $\mathbf{A}$  be a matrix whose columns have unit  $\ell^2$  norm, and let  $\mu(\mathbf{A})$  denote its mutual coherence. Suppose that  $\mathbf{y} = \mathbf{A}\mathbf{x}_o$ , with

$$\|\mathbf{x}_o\|_0 \leq \frac{1}{2\mu(\mathbf{A})}. \quad (12)$$

Then  $\mathbf{x}_o$  is the unique optimal solution to the problem

$$\min \|\mathbf{x}\|_1 \quad \text{s.t.} \quad \mathbf{y} = \mathbf{A}\mathbf{x}. \quad (13)$$

**Tightness:** there exist examples of  $\mathbf{A}$  and  $\mathbf{x}_o$  with  $\|\mathbf{x}_o\|_0 > \frac{1}{2} \left(1 + \frac{1}{\mu(\mathbf{A})}\right)$  for which  $\ell^1$  minimization does not recover  $\mathbf{x}_o$ .

## Correctness of $\ell^1$ Minimization

Given  $\mathbf{y} = \mathbf{A}\mathbf{x}_o$ , try to find  $\mathbf{x}_o$  via  $\ell^1$  minimization:

$$\min \|\mathbf{x}\|_1 \quad \text{s.t.} \quad \mathbf{y} = \mathbf{A}\mathbf{x}. \quad (14)$$

**Lagrangian formulation:**

$$\min \|\mathbf{x}\|_1 + \boldsymbol{\lambda}^*(\mathbf{y} - \mathbf{A}\mathbf{x}), \quad \exists \boldsymbol{\lambda} \in \mathbb{R}^m. \quad (15)$$

**Optimality condition:**  $\mathbf{x}_o$  is a minimum of  $f(\mathbf{x})$  if and only if  $\mathbf{0}$  is in the subgradient  $\partial f(\mathbf{x})$  at  $\mathbf{x}_o$ :

$$f(\mathbf{x}) \geq f(\mathbf{x}_o) + \mathbf{0}^*(\mathbf{x} - \mathbf{x}_o).$$

Optimality condition for  $\ell^1$  Minimization:

$$\mathbf{0} \in \partial \|\mathbf{x}_o\|_1 - \mathbf{A}^* \boldsymbol{\lambda} \quad \Leftrightarrow \quad \mathbf{A}^* \boldsymbol{\lambda} \in \partial \|\mathbf{x}_o\|_1. \quad (16)$$

# Correctness of $\ell^1$ Minimization

## Proof (a sketch of key ideas):

Due to convexity of  $\|\cdot\|_1$ , for any  $\mathbf{v} \in \partial \|\cdot\|_1(\mathbf{x}_o)$  and  $\mathbf{x}' \in \mathbb{R}^n$ ,

$$\|\mathbf{x}'\|_1 \geq \|\mathbf{x}_o\|_1 + \langle \mathbf{v}, \mathbf{x}' - \mathbf{x}_o \rangle \quad (17)$$

For  $\mathbf{v} = \mathbf{A}^* \boldsymbol{\lambda}$ , we have:  $\langle \mathbf{A}^* \boldsymbol{\lambda}, \mathbf{x}' - \mathbf{x}_o \rangle = \langle \boldsymbol{\lambda}, \mathbf{A}(\mathbf{x}' - \mathbf{x}_o) \rangle = 0$ . Therefore

$$\|\mathbf{x}'\|_1 \geq \|\mathbf{x}_o\|_1.$$

To find such an optimality certificate  $\mathbf{A}^* \boldsymbol{\lambda} \in \partial \|\cdot\|_1(\mathbf{x}_o)$ , we need:

$$\mathbf{A}_I^* \boldsymbol{\lambda} = \boldsymbol{\sigma}, \quad \|\mathbf{A}_{I^c}^* \boldsymbol{\lambda}\|_\infty \leq 1. \quad (18)$$

A natural “candidate”:

$$\hat{\boldsymbol{\lambda}}_{\ell^2} \doteq \mathbf{A}_I (\mathbf{A}_I^* \mathbf{A}_I)^{-1} \boldsymbol{\sigma}. \quad (19)$$

The rest is to check this satisfies (18) under the given conditions.

# Correctness of $\ell^1$ Minimization

## Proof (continued):

By construction,  $\mathbf{A}_I^* \hat{\boldsymbol{\lambda}}_{\ell^2} = \boldsymbol{\sigma}$ . We are just left to verify (18), by calculating

$$\|\mathbf{A}_{I^c}^* \hat{\boldsymbol{\lambda}}_{\ell^2}\|_{\infty} = \|\mathbf{A}_{I^c}^* \mathbf{A}_I (\mathbf{A}_I^* \mathbf{A}_I)^{-1} \boldsymbol{\sigma}\|_{\infty}. \quad (20)$$

Consider a single element of this vector ( $j \in I^c$ ), which has the form:

$$|a_j^* \mathbf{A}_I (\mathbf{A}_I^* \mathbf{A}_I)^{-1} \boldsymbol{\sigma}| \leq \underbrace{\|\mathbf{A}_I^* a_j\|_2}_{\leq \sqrt{k}\mu} \underbrace{\|(\mathbf{A}_I^* \mathbf{A}_I)^{-1}\|_{2,2}}_{< \frac{1}{1-k\mu(\mathbf{A})}} \underbrace{\|\boldsymbol{\sigma}\|_2}_{=\sqrt{k}} \quad (21)$$

$$< \frac{k\mu(\mathbf{A})}{1 - k\mu(\mathbf{A})} \quad (22)$$

$$\leq \frac{1}{\text{Provided } k\mu(\mathbf{A}) \leq 1/2}. \quad (23)$$



# Constructing Incoherent Matrices

**Example I.** Consider a discrete Fourier transform matrix  $\mathbf{F}$ . Let  $I \subset [n]$  be a random set of  $m$  indices,

$$\mathbf{A} = \mathbf{F}_I^* \in \mathbb{C}^{m \times n}. \quad (24)$$

**Example II.** For two orthogonal matrices  $\Phi$  and  $\Psi$ ,

$$\mathbf{A} = \Phi_I^* \Psi. \quad (25)$$

**Example III.** For two orthogonal matrices, say  $\Phi$  is Fourier  $\mathbf{F}$  and  $\Psi$  is the identify  $\mathbf{I}$  or the Wavelet  $\mathbf{W}$ ,

$$\mathbf{A} = [\Phi \mid \Psi] \in \mathbb{C}^{n \times 2n}. \quad (26)$$



# Incoherence and Uncertainty Principle

**Incoherence between  $I$  and  $F$ :**  $|\langle e_i, f_j \rangle| = \frac{1}{\sqrt{n}}$ .

**Facts:** A signal cannot be sparse in both time  $I$  and frequency  $F$ . Let  $\hat{x} = Fx \in \mathbb{C}^n$  be the discrete Fourier transform of  $x \in \mathbb{C}^n$ . Then the **Heisenberg uncertainty principle** states that:

$$\text{Var}(|x|^2) \text{Var}(|\hat{x}|^2) \geq \frac{1}{16\pi^2}. \quad (27)$$

Or a deterministic uncertainty principle:

$$\|x\|_0 \cdot \|\hat{x}\|_0 \geq n \quad \text{or} \quad \|x\|_0 + \|\hat{x}\|_0 \geq 2\sqrt{n}. \quad (28)$$

# Incoherence and Uncertainty Principle

## Theorem (Uncertainty Principle I<sup>2</sup>)

For  $\mathbf{A} = [\Phi \mid \Psi] \in \mathbb{C}^{n \times 2n}$  with two orthogonal matrices  $\Phi$  and  $\Psi$ . For any  $\mathbf{0} = \Phi \mathbf{e} + \Psi \hat{\mathbf{e}}$  with  $\Phi \mathbf{e} = -\Psi \hat{\mathbf{e}} \neq \mathbf{0}$ , we have

$$\|\mathbf{e}\|_0 + \|\hat{\mathbf{e}}\|_0 \geq \frac{2}{\mu(\mathbf{A})}. \quad (29)$$

## Corollary (Uncertainty Principle II)

For  $\mathbf{A} = [\Phi \mid \Psi] \in \mathbb{C}^{n \times 2n}$  with two orthogonal matrices  $\Phi$  and  $\Psi$ . For any nonzero  $\mathbf{y} = \Phi \mathbf{x} + \Psi \hat{\mathbf{x}}$ , we have

$$\|\mathbf{x}\|_0 + \|\hat{\mathbf{x}}\|_0 \geq \frac{2}{\mu(\mathbf{A})}. \quad (30)$$

**Question:** What can you say about  $\mathbf{y} = \mathbf{A}\mathbf{x}$  with  $\|\mathbf{x}\|_0 < \frac{1}{\mu(\mathbf{A})}$ ?

<sup>2</sup>*Sparse and Redundant Representations*, Michael Elad, Springer, 2010. 

# Constructing Incoherent Matrices

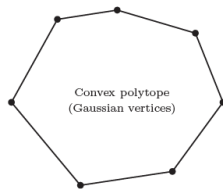
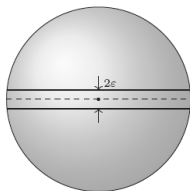
Recall phenomena associated with random matrices:

- **Measure Concentration** ( $\epsilon \sim O(n^{-1/2})$ )

$$\text{Area}\{x \in \mathbb{S}^{n-1} : -\epsilon \leq x_n \leq \epsilon\} = 0.99 \cdot \text{Area}(\mathbb{S}^{n-1}), \quad (31)$$

- **Neighborly Polytopes** (vertices from a Gaussian matrix):

$$\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n] \in \mathbb{R}^{m \times n}.$$



# Constructing Incoherent Matrices

## Theorem (Spherical Measure Concentration<sup>3</sup>)


Let  $\mathbf{u} \sim \text{uni}(\mathbb{S}^{m-1})$  be distributed according to the uniform distribution on the sphere. Let  $f : \mathbb{S}^{m-1} \rightarrow \mathbb{R}$  be an 1-Lipschitz function:

$$\forall \mathbf{u}, \mathbf{u}', \quad |f(\mathbf{u}) - f(\mathbf{u}')| \leq 1 \cdot \|\mathbf{u} - \mathbf{u}'\|_2, \quad (32)$$

and let  $\text{med}(f)$  denote any median of the random variable  $Z = f(\mathbf{u})$ . Then

$$\mathbb{P}[f(\mathbf{u}) > \text{med}(f) + t] \leq 2 \exp\left(-\frac{mt^2}{2}\right), \quad (33)$$

$$\mathbb{P}[f(\mathbf{u}) < \text{med}(f) - t] \leq 2 \exp\left(-\frac{mt^2}{2}\right). \quad (34)$$

<sup>3</sup>Lectures on Discrete Geometry, Jiri Matousek, Springer 2002. 

# Constructing Incoherent Matrices

## Theorem

Let  $\mathbf{A} = [\mathbf{a}_1 \mid \cdots \mid \mathbf{a}_n]$  with columns  $\mathbf{a}_i \sim \text{uni}(\mathbb{S}^{m-1})$  chosen independently according to the uniform distribution on the sphere. Then with probability at least  $3/4$ ,

$$\mu(\mathbf{A}) \leq C \sqrt{\frac{\log n}{m}}, \quad (35)$$

where  $C > 0$  is a numerical constant.

**Proof (a sketch):** For any  $\mathbf{v} \in \mathbb{S}^{m-1}$ ,  $\mathbb{E}[|\mathbf{v}^* \mathbf{a}|]^2 \leq [(\mathbf{v}^* \mathbf{a})^2] \leq \frac{1}{m}$  implies

$$\text{med}(|\mathbf{v}^* \mathbf{a}|) \leq 2\mathbb{E}[|\mathbf{v}^* \mathbf{a}|] \leq \frac{2}{\sqrt{m}}.$$

$$\mathbb{P}\left[|\mathbf{v}^* \mathbf{a}| > \frac{2+t}{\sqrt{m}}\right] \leq 2 \exp\left(-\frac{t^2}{2}\right). \quad (36)$$

# Constructing Incoherent Matrices

## Proof (continued):

As all the  $n$  columns  $\{\mathbf{a}_i\}$  are independent:

$$\mathbb{P} \left[ |\mathbf{a}_i^* \mathbf{a}_j| > \frac{2+t}{\sqrt{m}} \right] \leq 2 \exp \left( -\frac{t^2}{2} \right). \quad (37)$$

Summing the failure probability over all  $n(n-1)/2$  pairs of  $(\mathbf{a}_i, \mathbf{a}_j)$ :

$$\mathbb{P} \left[ \exists (i, j) : |\mathbf{a}_i^* \mathbf{a}_j| > \frac{2+t}{\sqrt{m}} \right] \leq n(n-1) \exp \left( -\frac{t^2}{2} \right). \quad (38)$$

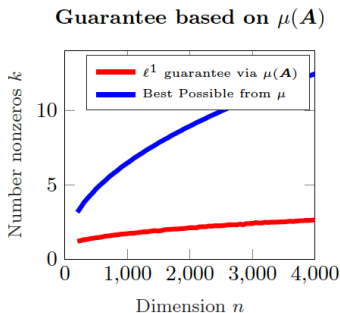
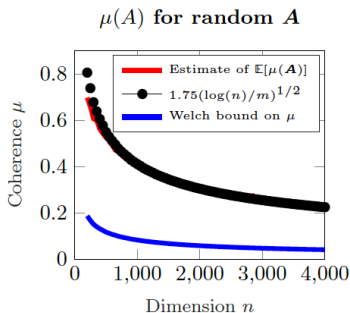
Setting  $t = 2\sqrt{\log 2n}$ , the RHS probability is less than  $1/4$ . □

# Limitations of Incoherence

## Theorem (Welch Bound)

For any matrix  $\mathbf{A} = [\mathbf{a}_1 \mid \cdots \mid \mathbf{a}_n] \in \mathbb{R}^{m \times n}$ ,  $m \leq n$ , and suppose that the columns  $\mathbf{a}_i$  have unit  $\ell^2$  norm. Then

$$\mu(\mathbf{A}) = \max_{i \neq j} |\langle \mathbf{a}_i, \mathbf{a}_j \rangle| \geq \sqrt{\frac{n-m}{m(n-1)}} = \Omega\left(\frac{1}{\sqrt{m}}\right). \quad (39)$$



# Limitations of Incoherence

## Proof of the Welch bound.

Let  $\mathbf{G} = \mathbf{A}^* \mathbf{A} \in \mathbb{R}^{n \times n}$  and its eigenvalues satisfy:  $\sum_{i=1}^m \lambda_i(\mathbf{G}) = \text{trace}(\mathbf{G}) = \sum_{i=1}^n \|\mathbf{a}_i\|_2^2 = n$ . Using this fact, we have:

$$\frac{n^2}{m} \leq \frac{n^2}{m} + \sum_{i=1}^m \left( \lambda_i(\mathbf{G}) - \frac{n}{m} \right)^2 \quad (40)$$

$$= \frac{n^2}{m} + \sum_{i=1}^m \left\{ \lambda_i^2(\mathbf{G}) + \frac{n^2}{m^2} - 2 \frac{n}{m} \lambda_i(\mathbf{G}) \right\} \quad (41)$$

$$= \sum_{i=1}^m \lambda_i^2(\mathbf{G}) = \|\mathbf{G}\|_F^2 = \sum_{i,j} |\mathbf{a}_i^* \mathbf{a}_j|^2 = n + \sum_{i \neq j} |\mathbf{a}_i^* \mathbf{a}_j|^2 \quad (42)$$

$$\leq n + n(n-1) \left( \max_{i \neq j} |\mathbf{a}_i^* \mathbf{a}_j| \right)^2. \quad (43)$$





# Limitations of Incoherence

**Incoherence** ensures to recover  $k$ -sparse solution from

$$m \geq \tilde{O}(k^2)$$

measurements.

**Experimental results suggest**  $m = O(k)$  :

*In a proportional growth setting  $m \propto n$ ,  $k \propto m$ ,  $\ell^1$  minimization succeeds with very high probability whenever the constants of proportionality  $n/m$  and  $k/m$  are small enough.*

**Next: how to sharpen the bound?**

# Assignments

- Reading: Section 3.1 & 3.2 of Chapter 3.
- Programming Homework # 1.