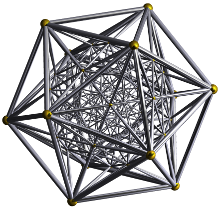
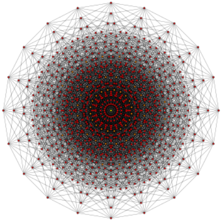
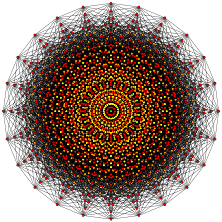
Computational Principles for High-dim Data Analysis

(Lecture Four)

Yi Ma

EECS Department, UC Berkeley

September 7, 2021

Convex Methods for Sparse Signal Recovery

"Algebra is but written geometry; geometry is but drawn algebra."
– Sophie Germain

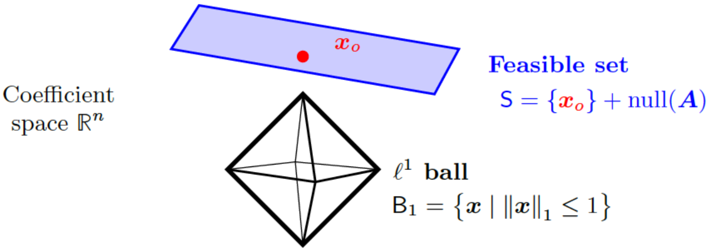
Geometric Intuition: Coefficient Space

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

$$\min \|\boldsymbol{x}\|_1 \quad \text{subject to} \quad \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}. \tag{1}$$

The space of all feasible solutions is an affine subspace:

$$\mathsf{S} = \{\boldsymbol{x} \mid \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}\} = \{\boldsymbol{x}_o\} + \operatorname{null}(\boldsymbol{A}) \quad \subset \mathbb{R}^n. \tag{2}$$

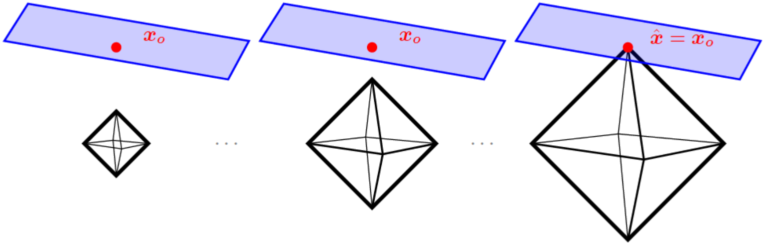


ℓ^1 Minimization in the Coefficient Space

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

$$t \cdot \mathsf{B}_1 = \{\boldsymbol{x} \mid \|\boldsymbol{x}\|_1 \leq t\} \quad \subset \mathbb{R}^n, \tag{3}$$

till its boundary first touches the feasible set S:



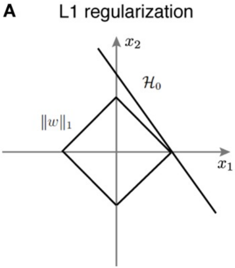
Comparison between ℓ^1 and ℓ^2 Minimization

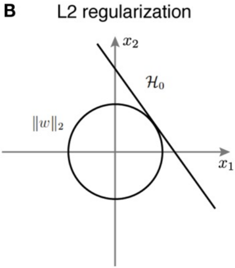
Given $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_o$ with $\boldsymbol{x}_o$ sparse:

$$\textbf{A}: \quad \min \|\boldsymbol{x}\|_1 \quad \text{subject to} \quad \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}. \tag{4}$$

versus

$$\textbf{B}: \quad \min \|\boldsymbol{x}\|_2 \quad \text{subject to} \quad \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y} \tag{5}$$





Sparsity Promoting with Different ℓ^p Norms

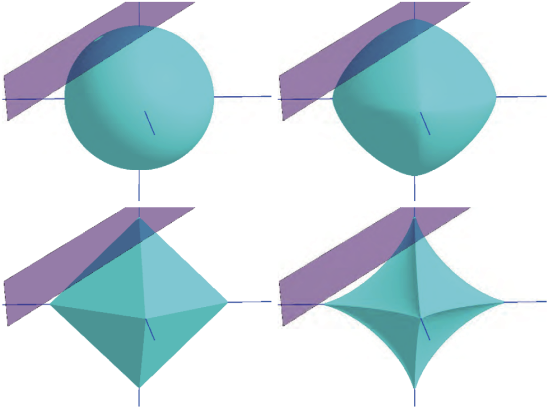


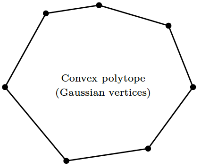
Figure: Intersection between the ℓ^p-ball and the feasible set S, for $p = 2, 1.5, 1$ and 0.7, respectively. (Some argue $p = 0.5$ is somewhat special.)

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

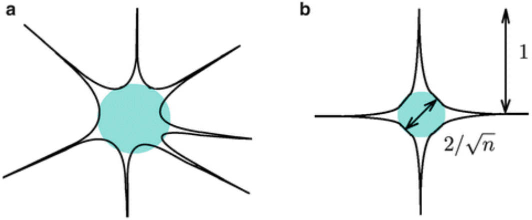
Geometric Intuition: High-dimensional Polytopes

Neighborly Polytopes
(vertices from a Gaussian matrix):

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



The "**correct**" visualization of high-dimensional convex polytopes,[1] including the ℓ^1 ball:



A general convex set

The ℓ_1 ball

[1] Lectures on Discrete Geometry, Jiri Matousek, Springer 2002.

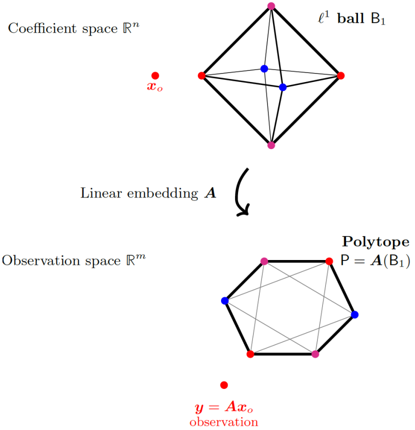
Geometric Intuition: Observation Space

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

$$\boldsymbol{A}: \quad \mathsf{B}_1 \rightarrow \mathsf{P} = \boldsymbol{A}(\mathsf{B}_1), \qquad (6)$$

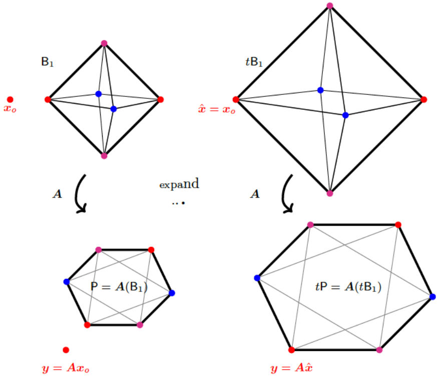
which maps a convex polytope to a convex polytope. Similarly, $\forall t \geq 0$:

$$t \cdot \mathsf{B}_1 \rightarrow t \cdot \boldsymbol{A}(\mathsf{B}_1).$$



Geometric Intuition: Observation Space

All k-faces of B_1 cannot be mapped to the inside of the polytope $\boldsymbol{A}(\mathrm{B}_1)$:



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

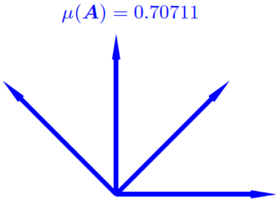
Coherence of a Matrix

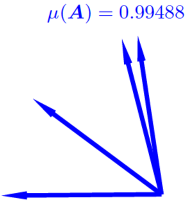
Definition (Mutual Coherence)

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

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

Example:





Uniqueness of Sparse Solution

Proposition (Coherence Controls Kruskal Rank)

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

$$\operatorname{krank}(\boldsymbol{A}) \;\geq\; \frac{1}{\mu(\boldsymbol{A})}. \tag{8}$$

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

$$\|\boldsymbol{x}_o\|_0 \;\leq\; \frac{1}{2\mu(\boldsymbol{A})}, \tag{9}$$

then $\boldsymbol{x}_o$ *is the unique optimal solution to the* ℓ^0 *minimization problem*

$$\min \|\boldsymbol{x}\|_0 \quad \text{s.t.} \quad \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}. \tag{10}$$

Proof:

$$1 - k\mu(\boldsymbol{A}) \;<\; \sigma_{\min}(\boldsymbol{A}_{\mathsf{I}}^* \boldsymbol{A}_{\mathsf{I}}) \;\leq\; \sigma_{\max}(\boldsymbol{A}_{\mathsf{I}}^* \boldsymbol{A}_{\mathsf{I}}) \;<\; 1 + k\mu(\boldsymbol{A}). \tag{11}$$

Correctness of ℓ^1 Minimization

Theorem (ℓ^1 Succeeds under Incoherence)

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

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

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

$$\min \|\boldsymbol{x}\|_1 \quad \text{s.t.} \quad \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}. \tag{13}$$

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

Correctness of ℓ^1 Minimization

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

$$\min \|\boldsymbol{x}\|_1 \quad \text{s.t.} \quad \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}. \tag{14}$$

Lagrangian formulation:

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

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

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

Optimality condition for ℓ^1 Minimization:

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

Correctness of ℓ^1 Minimization

Proof (a sketch of key ideas):

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

$$\left\|\boldsymbol{x}'\right\|_1 \geq \left\|\boldsymbol{x}_o\right\|_1 + \langle \boldsymbol{v}, \boldsymbol{x}' - \boldsymbol{x}_o \rangle \tag{17}$$

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

$$\left\|\boldsymbol{x}'\right\|_1 \geq \left\|\boldsymbol{x}_o\right\|_1 .$$

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

$$\boldsymbol{A}_{\mathsf{I}}^* \boldsymbol{\lambda} = \boldsymbol{\sigma}, \quad \left\|\boldsymbol{A}_{\mathsf{I}^c}^* \boldsymbol{\lambda}\right\|_\infty \leq 1. \tag{18}$$

A natural "candidate":

$$\hat{\boldsymbol{\lambda}}_{\ell^2} \doteq \boldsymbol{A}_{\mathsf{I}} (\boldsymbol{A}_{\mathsf{I}}^* \boldsymbol{A}_{\mathsf{I}})^{-1} \boldsymbol{\sigma}. \tag{19}$$

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

Correctness of ℓ^1 Minimization

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

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

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

$$\begin{aligned} |\boldsymbol{a}_j^* \boldsymbol{A}_{\mathsf{I}} (\boldsymbol{A}_{\mathsf{I}}^* \boldsymbol{A}_{\mathsf{I}})^{-1} \boldsymbol{\sigma}| &\leq \underbrace{\|\boldsymbol{A}_{\mathsf{I}}^* \boldsymbol{a}_j\|_2}_{\leq \sqrt{k}\mu} \underbrace{\left\|(\boldsymbol{A}_{\mathsf{I}}^* \boldsymbol{A}_{\mathsf{I}})^{-1}\right\|_{2,2}}_{< \frac{1}{1-k\mu(\boldsymbol{A})}} \underbrace{\|\boldsymbol{\sigma}\|_2}_{=\sqrt{k}} && (21) \\ &< \frac{k\mu(\boldsymbol{A})}{1 - k\mu(\boldsymbol{A})} && (22) \\ &\leq \underbrace{1}_{\text{Provided } k\mu(\boldsymbol{A}) \leq 1/2.}. && (23) \end{aligned}$$

□

Constructing Incoherent Matrices

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

$$\boldsymbol{A} = \boldsymbol{F}_{\mathsf{I}}^{*} \in \mathbb{C}^{m \times n}. \tag{24}$$

Example II. For two orthogonal matrices $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$,

$$\boldsymbol{A} = \boldsymbol{\Phi}_{\mathsf{I}}^{*} \boldsymbol{\Psi}. \tag{25}$$

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

$$\boldsymbol{A} = [\boldsymbol{\Phi} \mid \boldsymbol{\Psi}] \in \mathbb{C}^{n \times 2n}. \tag{26}$$

Incoherence and Uncertainty Principle

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

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

$$\mathrm{Var}(|\boldsymbol{x}|^2)\mathrm{Var}(|\hat{\boldsymbol{x}}|^2) \geq \frac{1}{16\pi^2}. \tag{27}$$

Or a deterministic uncertainty principle:

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

Incoherence and Uncertainty Principle

Theorem (Uncertainty Principle I[2])

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

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

Corollary (Uncertainty Principle II)

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

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

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

[2] *Sparse and Redundant Representations*, Michael Elad, Springer, 2010.

Constructing Incoherent Matrices

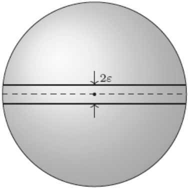
Recall phenomena associated with random matrices:

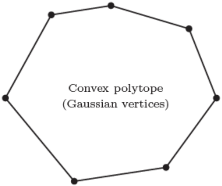
- **Measure Concentration** ($\epsilon \sim O(n^{-1/2})$)

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

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

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





Constructing Incoherent Matrices

Theorem (Spherical Measure Concentration[3])

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

$$\forall\, \boldsymbol{u}, \boldsymbol{u}', \quad \left| f(\boldsymbol{u}) - f(\boldsymbol{u}') \right| \;\le\; 1 \cdot \left\| \boldsymbol{u} - \boldsymbol{u}' \right\|_2, \tag{32}$$

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

$$\mathbb{P}\left[\, f(\boldsymbol{u}) > \mathrm{med}(f) + t \,\right] \quad \le \quad 2 \exp\left(-\frac{mt^2}{2} \right), \tag{33}$$

$$\mathbb{P}\left[\, f(\boldsymbol{u}) < \mathrm{med}(f) - t \,\right] \quad \le \quad 2 \exp\left(-\frac{mt^2}{2} \right). \tag{34}$$

[3] Lectures on Discrete Geometry, Jiri Matousek, Springer 2002.

Constructing Incoherent Matrices

Theorem

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

$$\mu(\boldsymbol{A}) \;\le\; C\sqrt{\frac{\log n}{m}}, \tag{35}$$

where $C > 0$ is a numerical constant.

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

$$\mathrm{med}(|\boldsymbol{v}^*\boldsymbol{a}|) \;\le\; 2\mathbb{E}[|\boldsymbol{v}^*\boldsymbol{a}|] \;\le\; \frac{2}{\sqrt{m}}.$$

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

Constructing Incoherent Matrices

Proof (continued):
As all the n columns $\{\boldsymbol{a}_i\}$ are independent:

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

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

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

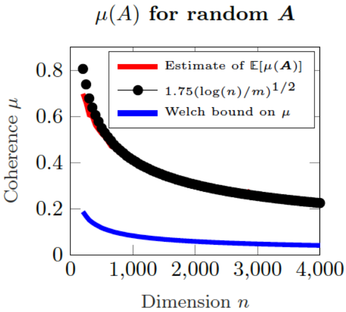
Setting $t = 2\sqrt{\log 2n}$, the RHS probability is less than $1/4$. $\square$

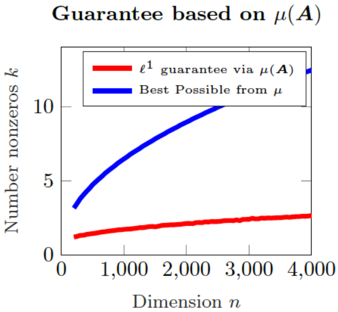
Limitations of Incoherence

Theorem (Welch Bound)

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

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





Limitations of Incoherence

Proof of the Welch bound.

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

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

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

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

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

□

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 $m \propto n$, $k \propto m$, ℓ^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.