

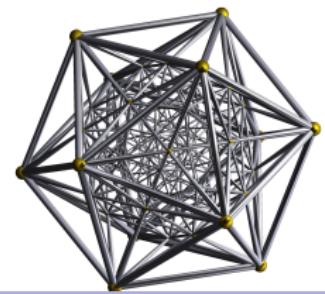
ICASSP 2023 Short Course

# Learning Nonlinear and Deep Representations from High-Dimensional Data From Theory to Practice

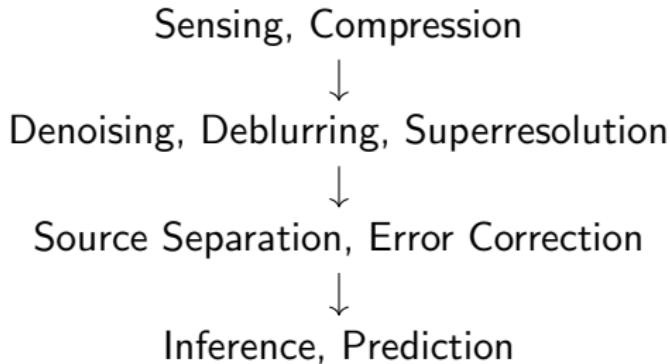
## Lecture 1: Introduction to Low-Dimensional Models

Sam Buchanan, Yi Ma, Qing Qu, Atlas Wang  
John Wright, Yuqian Zhang, Zhihui Zhu

June 6, 2023



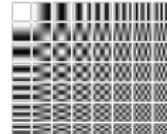
# The Signal Processing Pipeline



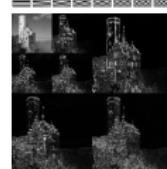
The **pursuit of low-dimensional structure** is a universal task!

# Historical Context: Quest for Low-Dimensionality

Fourier



Wavelets



X-lets: Curvelets, Contourlets, Bandelets, ...



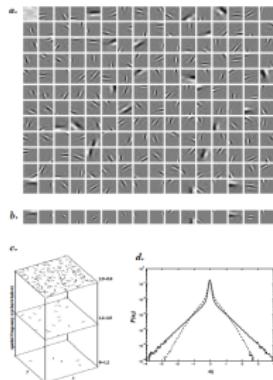
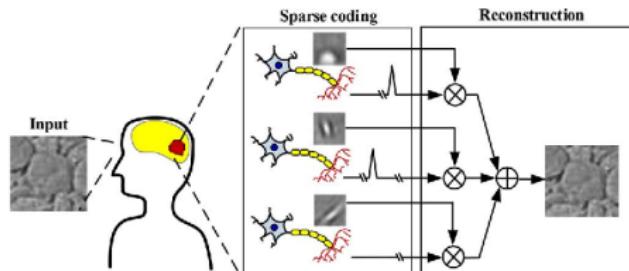
Learned Dictionaries

Learned Reconstruction Procedures

A continuing quest for **sparse signal representations**  
leveraging mathematics + massive data and computation!

# Historical Context: Sparsity in Neuroscience

**Dogma for natural vision** [Barlow 1972]: “... to represent the input as completely as possible by activity in as few neurons as possible.”



Find sparse  $\{x_i\}$  such that

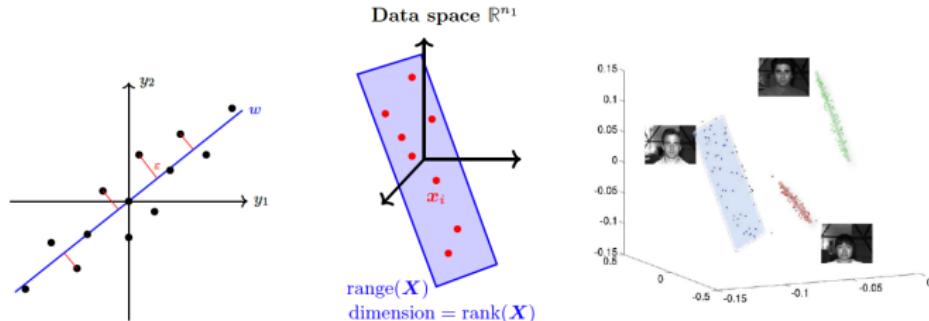
$$\mathbf{y} = \sum_{i=1}^n x_i \mathbf{a}_i + \epsilon \quad \in \mathbb{R}^m, \quad (1)$$

[Nature, Olshausen and Field 1996.]

# Historical Context: Sparse and Low-d in Statistics

## Principal Component Analysis

Linear correlations in data (**low-rank model!**)



[Pearson 1901], [Hotelling 1933], [Eckart and Young 1936]

## Best Subset Selection

Select a few relevant predictors (**sparse model!**)

[Hocking, Leslie, and Beale 1967], Stagewise pursuit [Efroymson 1966],

Lasso [Tibshirani 1996], Basis pursuit [Chen, Donoho, and Saunders 1998]

# Historical Context: Estimation, Errors, Missing Data

A **long and rich history** of robust estimation with error correction and missing data imputation:



R. J. Boscovich. *De calculo probabilitatum que respondent diversis valoribus summe errorum post plures observationes ...*, before 1756



A. Legendre. *Nouvelles methodes pour la determination des orbites des cometes*, 1806



C. Gauss. *Theory of motion of heavenly bodies*, 1809



A. Beurling. *Sur les integrales de Fourier absolument convergentes et leur application a une transformation fonctionnelle*, 1938

⋮

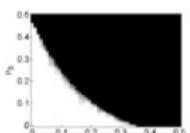
$$\boxed{A} \quad x + \circled{n}$$

over-determined  
+ dense, Gaussian

$$\boxed{A} \quad x + \diamond e$$

underdetermined  
+ sparse, Laplacian

# The Modern Era: Massive Data and Computation



(a) Robust PCA, Random Signs

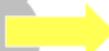
**BIG DATA**  
(images, videos,  
voices, texts,  
biomedical, geospatial,  
consumer data...)



**Mathematical Theory**  
(high-dimensional statistics, convex geometry,  
measure concentration, combinatorics...)



**Cloud Computing**  
(parallel, distributed,  
scalable platforms)



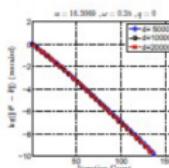
**Applications  
& Services**

(data processing,  
analysis, compression,  
knowledge discovery,  
search, recognition...)



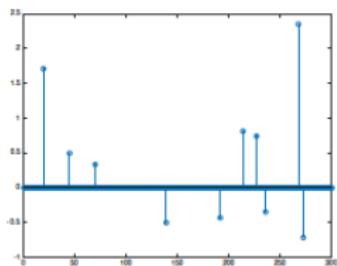
**Computational Methods**

(convex optimization, first-order algorithms,  
random sampling, deep networks...)

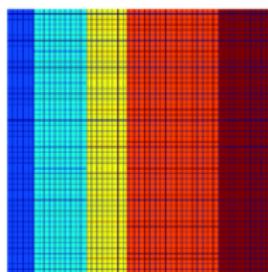


# Motivating Issues I: Correctness?

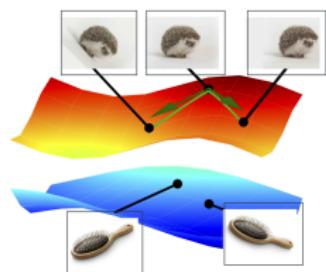
How can we **correctly** compute with **low-dimensional structure**?



Sparse Vectors



Low-rank Matrices

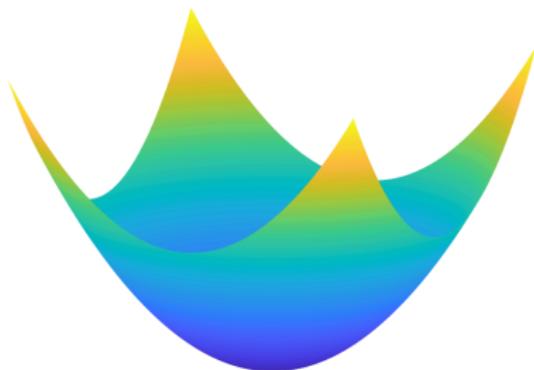


Nonlinear Manifolds

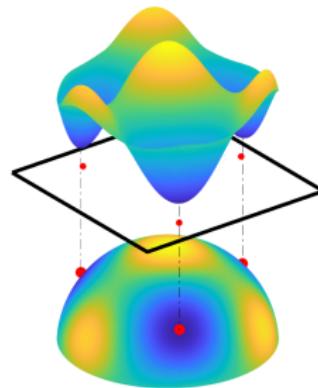
**Low-d. structure** leads to principled answers *and* practical methods!

## Motivating Issues II: Computational Efficiency?

**Computational Tractability:** easy vs./ hard problems:

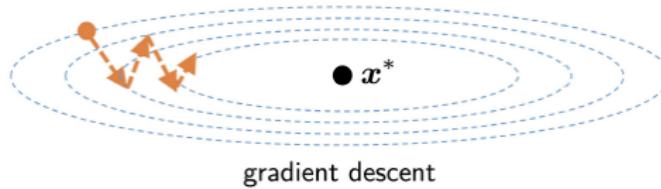


Convexity



Benign Nonconvexity

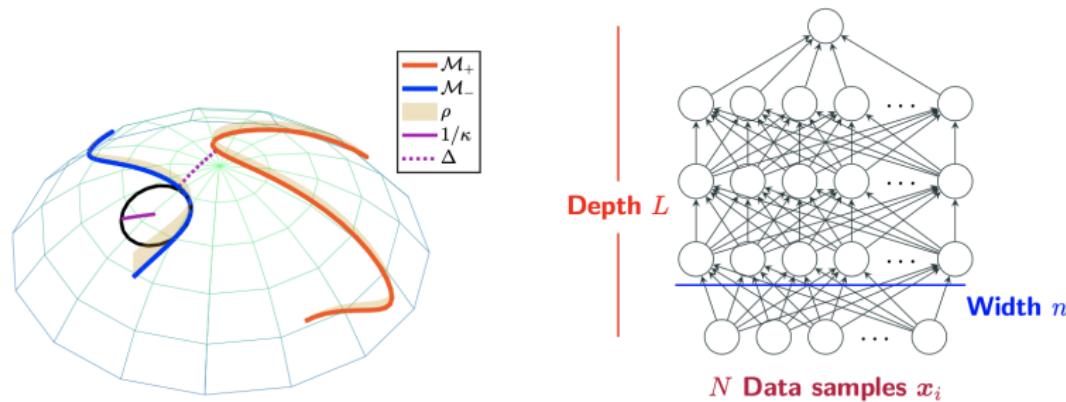
**Efficient, scalable methods** leveraging problem geometry:



# Motivating Issues III: Resource Efficiency?

**Data Efficiency:** How many samples? How many labels?

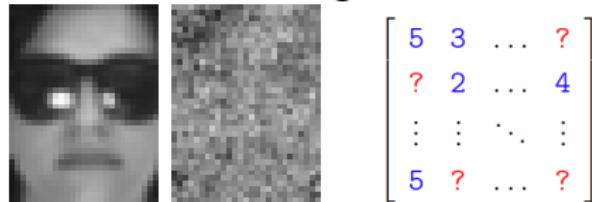
**Architecture Efficiency:** How deep? How wide? What operations?



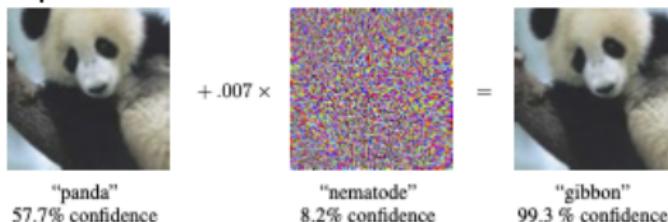
**Low-d. structure** of data sets fundamental resource requirements  
for **sensing** and **learning**.

# Motivating Issues IV – Robustness?

**Robustness:** to errors, outliers, missing data:



Robustness and deep networks?

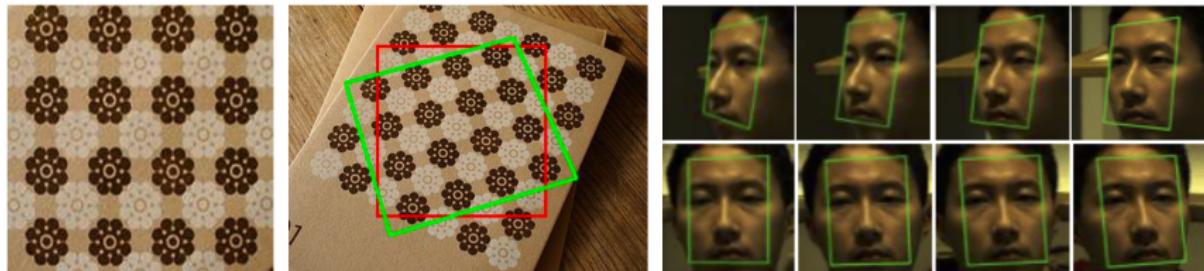


From [Goodfellow, Shlens and Szegedy, 2015]

**Low-d structure of signal and error can lead to principled approaches to robustness.**

# Motivating Issues V: Invariance?

Transformations of the signal domain:



can cause still lead to disturbing failures:



From [Azulay and Weiss, 2019]

**Low-d. structure** in texture / appearance and transformation!

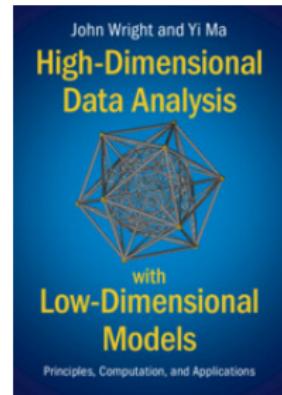
# This Course: The Plan

- Lecture 1: Introduction to Low-D Models
- Lecture 2 (today): Low-D in Neural Networks: Practice and Theory
- Lecture 3 (6/7): *Designing* Deep Networks for Low-D Structure
- Lecture 4 (6/7): Nonconvex Optimization for Low-D Structure
- Lecture 5-7 (6/8-9): *Learning* Deep Networks for Low-D Structure

# This Tutorial: Resources

## **High-Dimensional Data Analysis with Low-Dimensional Models Principles, Computation, and Applications**

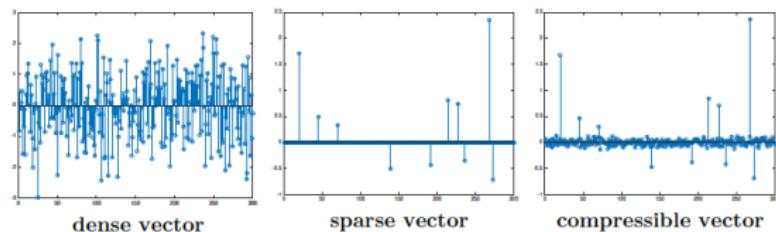
John Wright and Yi Ma  
Cambridge University Press, 2022.



Preproduction Copy from Website: <https://book-wright-ma.github.io>  
Slides, Code, etc: <https://book-wright-ma.github.io/Lecture-Slides/>

# Sparse Signal Models

**Sparse Signals:** Call  $x_o \in \mathbb{R}^n$  sparse if it has only a few nonzero entries:



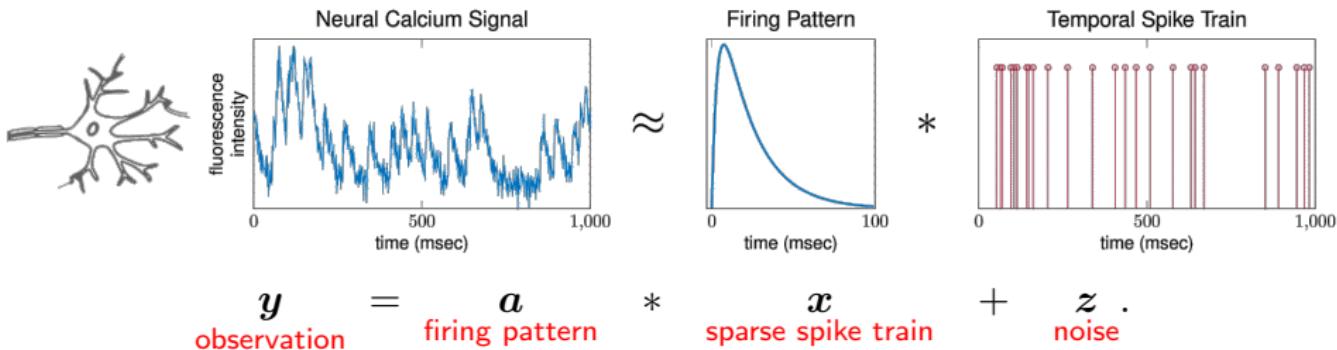
**Sparse Recovery:** Given *linear measurements*  $y \in \mathbb{R}^m$  of a sparse signal  $x_o$ :

$$\begin{matrix} \left[ \begin{array}{c} ? \\ ? \\ ? \\ ? \\ ? \end{array} \right] = \left[ \begin{array}{ccccc} ? & ? & ? & ? & ? \\ ? & ? & ? & ? & ? \\ ? & ? & ? & ? & ? \\ ? & ? & ? & ? & ? \\ ? & ? & ? & ? & ? \end{array} \right] \left[ \begin{array}{c} ? \\ ? \\ ? \\ ? \\ ? \end{array} \right] \end{matrix}$$

$y$       =       $A$       <sub>observation</sub>      <sub>measurement matrix</sub>       $x_o$       <sub>unknown</sub>

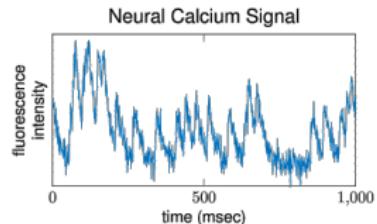
recover  $x_o$ .

# Sparsity I: Neural Spikes

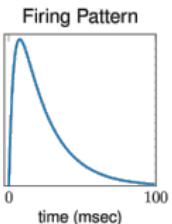


**Sparse and low-dimensional models** arise naturally from **physical structure** of data!

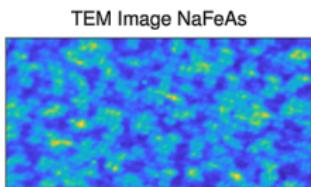
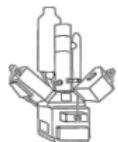
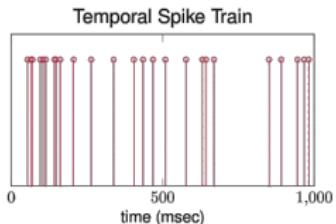
# Sparsity I: Neural Spikes and Beyond



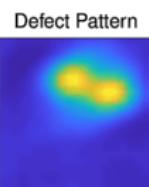
$\approx$



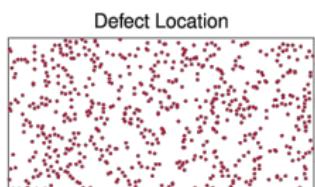
\*



$\approx$



\*



$\approx$

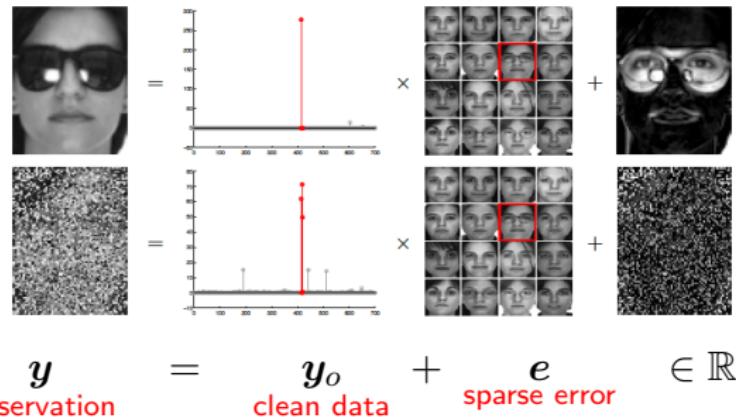


\*



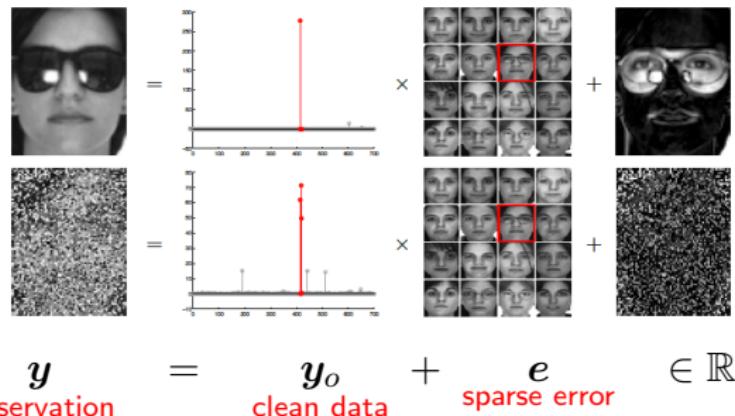
Common Convolutional Model:  $y = a * x + z$ , with  $x$  **sparse**.

## Sparsity II: Faces and Error Correction



Two types of structure: **sparsity of identity** and **sparsity of errors**.

## Sparsity II: Faces and Error Correction



Two types of structure: **sparsity of identity** and **sparsity of errors**.

Concatenate gallery images of  $n$  subjects into a large “dictionary”:

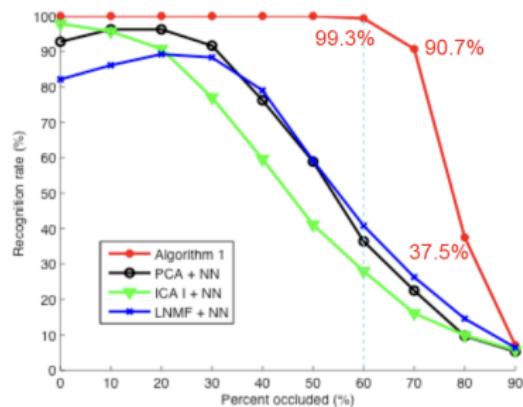
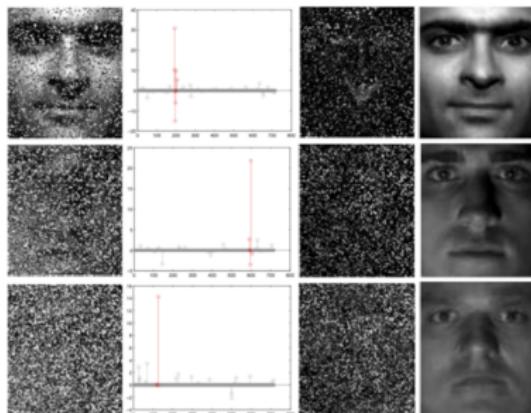
$$\mathbf{B} = [\mathbf{B}_1 \mid \mathbf{B}_2 \mid \cdots \mid \mathbf{B}_n] \in \mathbb{R}^{m \times n}$$

all training images

# Sparsity II: Faces and Error Correction

Find sparse solutions  $(x, e)$  to the linear system:

$$y = Bx + e = [B, I] \begin{bmatrix} x \\ e \end{bmatrix}.$$



**Correcting Gross Errors** is also a sparse recovery problem!

# Sparsity III: Magnetic Resonance Imaging

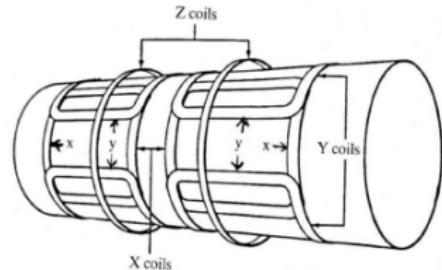
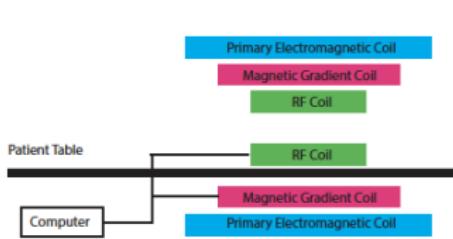
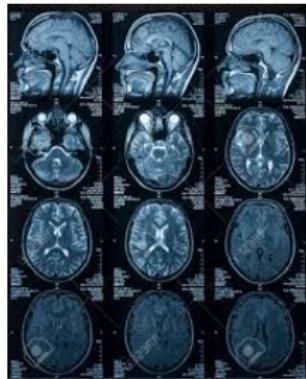


Figure: Left: Key components. Right: The three-axis gradient coils.

# Sparsity III: Magnetic Resonance Imaging

Simplified mathematical model for MRI:

$$y = \mathcal{F}[I](\mathbf{u}) = \int_{\mathbf{v}} I(\mathbf{v}) \exp(-i 2\pi \mathbf{u}^* \mathbf{v}) d\mathbf{v}, \quad \mathbf{u}, \mathbf{v} \in \mathbb{R}^2$$

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} = \begin{bmatrix} \mathcal{F}[I](\mathbf{u}_1) \\ \vdots \\ \mathcal{F}[I](\mathbf{u}_m) \end{bmatrix} \doteq \mathcal{F}_{\mathbf{U}}[I], \quad m \ll N^2.$$

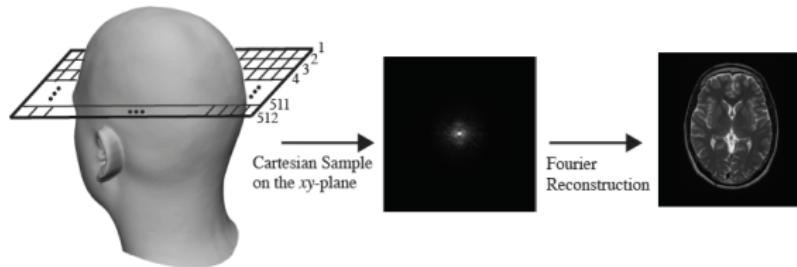


Figure: Recovering MRI image from Fourier measurements.

## Sparsity III: Structure of MR Images

Express  $I$  as a superposition of basis functions  $\Psi = \{\psi_1, \dots, \psi_{N^2}\}$ :

$$\underset{\text{image}}{I} = \sum_{i=1}^{N^2} \underset{i\text{-th basis signal}}{\psi_i} \times \underset{i\text{-th coefficient}}{x_i}.$$

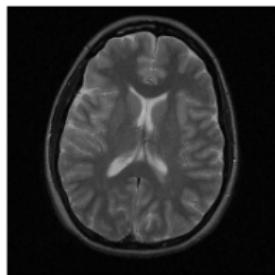
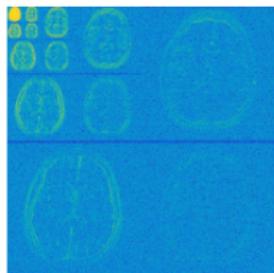
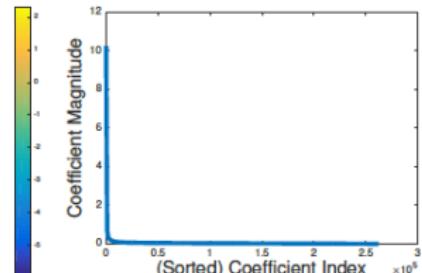


image  $I(v)$



wavelet coefficients  $x$ :  $I = \Psi[x]$ .



Many natural signals become **sparse** or **compressible** in an appropriately designed transform domain!

## Sparsity III: Image Reconstruction by Sparse Recovery

$$\begin{aligned} \mathbf{y} &= \mathcal{F}_{\mathbf{U}}[\mathbf{I}], \\ \text{observed Fourier coefficients} \\ &= \mathcal{F}_{\mathbf{U}} \left[ \boldsymbol{\psi}_1 x_1 + \cdots + \boldsymbol{\psi}_{N^2} x_{N^2} \right], \\ &= \mathcal{F}_{\mathbf{U}}[\boldsymbol{\psi}_1] x_1 + \cdots + \mathcal{F}_{\mathbf{U}}[\boldsymbol{\psi}_{N^2}] x_{N^2}, \\ &= \left[ \mathcal{F}_{\mathbf{U}}[\boldsymbol{\psi}_1] \mid \cdots \mid \mathcal{F}_{\mathbf{U}}[\boldsymbol{\psi}_{N^2}] \right] \mathbf{x}, \\ &\quad \text{matrix } \mathbf{A} \in \mathbb{R}^{m \times N^2}, m \ll N^2. \\ &= \mathbf{Ax}. \end{aligned} \tag{2}$$

$\mathbf{x}$  is sparse or approximately sparse!

**Compressed sensing:** the number of measurements  $m$  for accurate reconstruction should be dictated by signal complexity

## Sparsity IV: Image Patches

Denoising given  $I_{\text{noisy}} = I_{\text{clean}} + z \dots$  break into patches  $y_1, \dots, y_p$ :

$$y_i = y_{i \text{ clean}} + z_i = \underset{\text{patch dictionary}}{A} \times \underset{\text{sparse coefficient vector}}{x_i} + z_i.$$

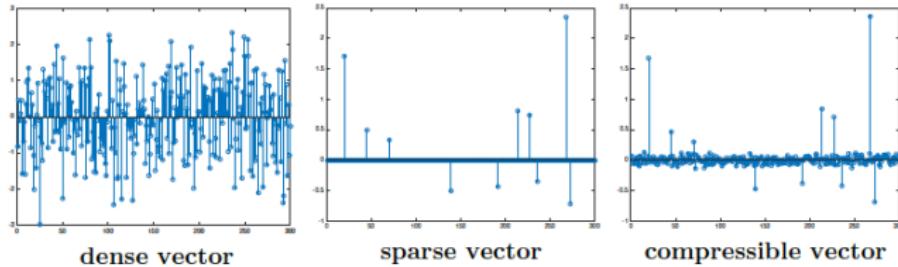


Figure: Left: noisy input; middle: denoised; right: *learned* patch dictionary.

Natural signals are challenging to model analytically  $\implies$  can **learn the sparse model** from data!

Figure: [Mairal, Elad, Sapiro '08]

# Measuring Sparsity: $\ell^0$ Norm

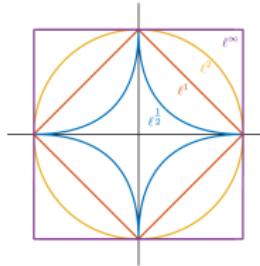


**Def:** the  $\ell^0$  “norm”  $\|x\|_0$  is the **number of nonzero entries** in the vector  $x$ :  $\|x\|_0 = \#\{i \mid x(i) \neq 0\}$ .

Connection to  $\ell^p$  norms

$$\|x\|_p = \left( \sum_i |x_i|^p \right)^{1/p} :$$

$$\|x\|_0 = \lim_{p \searrow} \|x\|_p^p.$$



The  $\ell^p$  balls.

## Sparse Recovery: $\ell^0$ minimization

**Computational Principle:** seek the **sparsest** signal consistent with our observations:

$$\hat{\mathbf{x}} = \arg \min \|\mathbf{x}\|_0 \quad \text{s.t.} \quad \mathbf{A}\mathbf{x} = \mathbf{y}.$$

Brute force exhaustive search: try all possible sets of nonzero entries

$$\mathbf{A}_{\mathcal{I}} \mathbf{x}_{\mathcal{I}} = \mathbf{y} ? \quad \forall \mathcal{I} \subseteq \{1, \dots, n\}, \ |\mathcal{I}| \leq k.$$

# Sparse Recovery: $\ell^0$ minimization

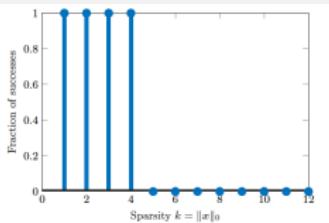
**Computational Principle:** seek the **sparsest** signal consistent with our observations:

$$\hat{\mathbf{x}} = \arg \min \|\mathbf{x}\|_0 \quad \text{s.t.} \quad \mathbf{A}\mathbf{x} = \mathbf{y}.$$

Brute force exhaustive search: try all possible sets of nonzero entries

$$\mathbf{A}_{\mathcal{I}} \mathbf{x}_{\mathcal{I}} = \mathbf{y} ? \quad \forall \mathcal{I} \subseteq \{1, \dots, n\}, \|\mathcal{I}\| \leq k.$$

**Theory:**  $\ell^0$  recovers **any sufficiently sparse signal!** For generic  $\mathbf{A}$ , success when  $\|\mathbf{x}_o\|_0 \leq \frac{m}{2}$ .



# $\ell^0$ Minimization is NP-hard

Theorem (Hardness of  $\ell^0$  Minimization)

The  $\ell^0$ -minimization problem  $\min \|x\|_0$  s.t.  $Ax = y$  is (strongly) **NP-hard**.

**Proof:** Reducible from *Exact 3-Set Cover* (E3C) problem.

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

$y$      $A$      $x$

# $\ell^0$ Minimization is NP-hard

Theorem (Hardness of  $\ell^0$  Minimization)

The  $\ell^0$ -minimization problem  $\min \|x\|_0$  s.t.  $Ax = y$  is (strongly) **NP-hard**.

**Proof:** Reducible from *Exact 3-Set Cover* (E3C) problem.

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

$y \qquad \qquad \qquad A \qquad \qquad \qquad x$

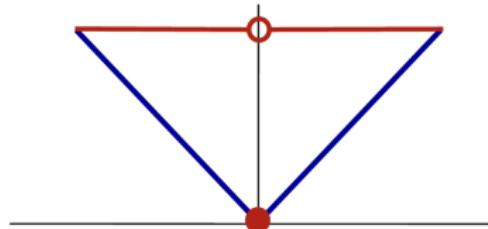
In high dimensions, need to pay attention to *both statistical and computational efficiency*!

## Convex Relaxation: $\ell^1$ Minimization

Intuitive reasons why  $\ell^0$  minimization:

$$\min \|x\|_0 \quad \text{subject to} \quad Ax = y. \quad (3)$$

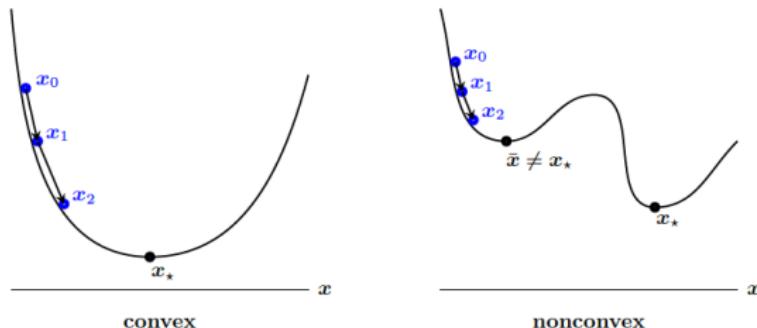
is very challenging:



$\ell^0$  is nonconvex, discontinuous, **not amenable to local search methods such as gradient descent.**

# Convex Relaxation: $\ell^1$ Minimization

For minimizing a generic function:  $\min f(x), x \in C$  (a convex set), **local methods**:  $x_{k+1} = x_k - t\nabla f(x_k)$  succeed *only if*  $f$  has “nice” geometry:



Need to formulate for computational efficiency!

- Lectures 1: **convex relaxations** for sparse, low-rank models
- Lectures 2+: **benign nonconvex formulations** for nonlinear models

# Convex Relaxation: $\ell^1$ Minimization

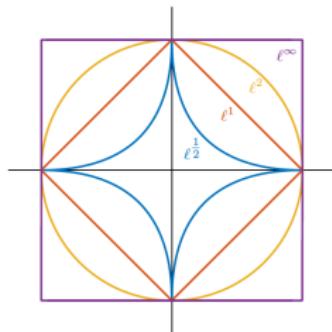
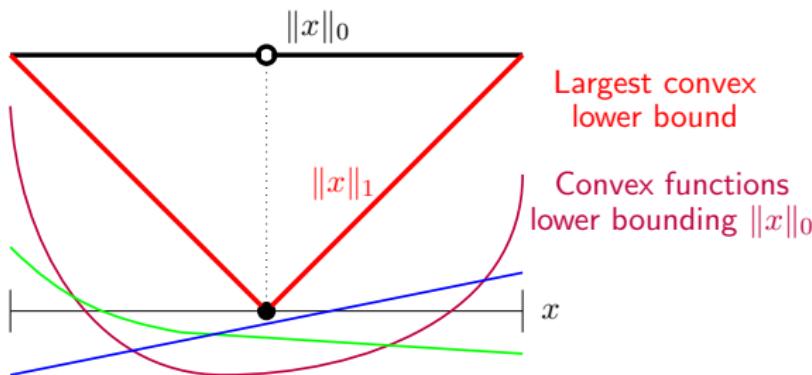


Figure: Convex surrogates for the  $\ell^0$  norm.  $\|x\|_1$  is the *convex envelope* of  $\|x\|_0$  on  $B_\infty$ .

Efficient **convex relaxation**:

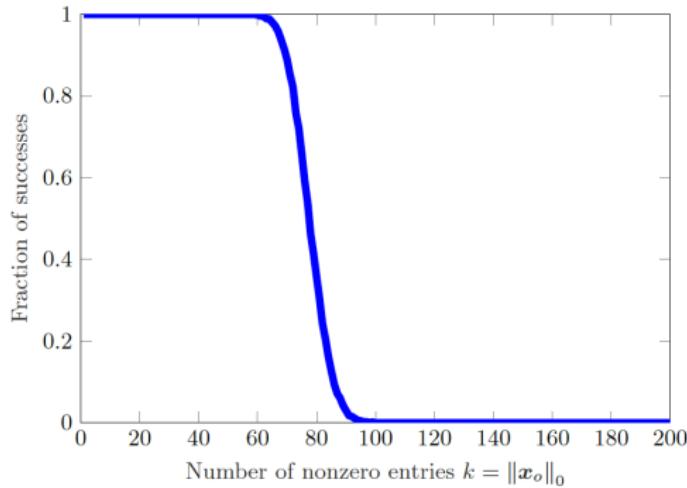
$$\min \|x\|_1 \quad \text{subject to} \quad Ax = y.$$

Solvable *quickly* at *large scale* using dedicated methods (Lecture 2).

# Minimizing the $\ell^1$ Norm: Simulations

**Solve:**  $\min \|x\|_1 \quad \text{s.t.} \quad Ax = y.$  (4)

$A$  is of size  $200 \times 400$ . Fraction of success across 50 trials.



**Experiment:**  $\ell^1$  minimization recovers *any sufficiently sparse signal?*

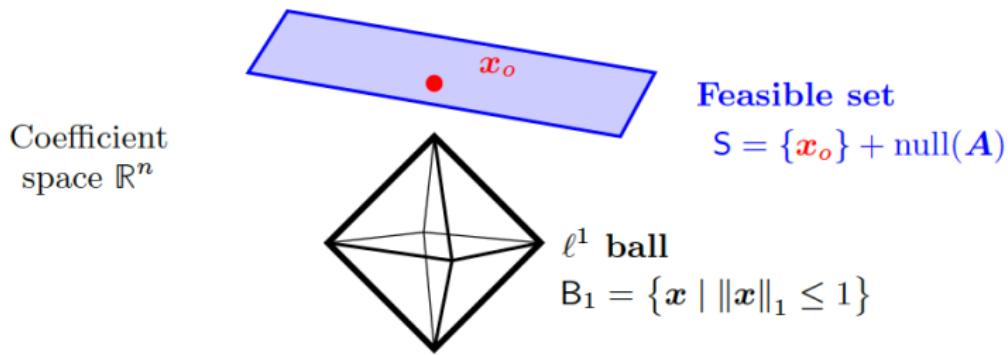
# Geometric Intuition: Coefficient Space

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

$$\min \|\mathbf{x}\|_1 \quad \text{subject to} \quad \mathbf{A}\mathbf{x} = \mathbf{y}. \quad (5)$$

The space of all feasible solutions is an affine subspace:

$$S = \{\mathbf{x} \mid \mathbf{A}\mathbf{x} = \mathbf{y}\} = \{\mathbf{x}_o\} + \text{null}(\mathbf{A}) \subset \mathbb{R}^n. \quad (6)$$

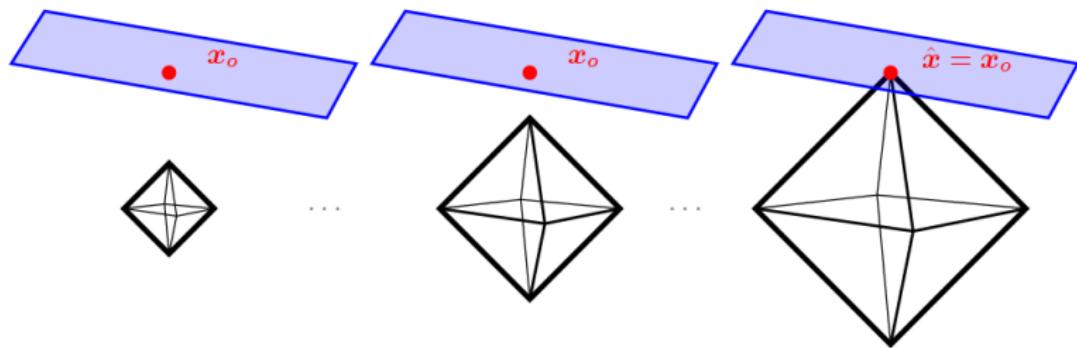


# Geometric Intuition: Coefficient Space

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

$$t \cdot \mathcal{B}_1 = \{\mathbf{x} \mid \|\mathbf{x}\|_1 \leq t\} \subset \mathbb{R}^n, \quad (7)$$

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

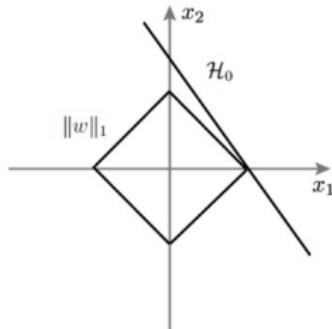


# Geometric Intuition: $\ell^1$ vs. $\ell^2$ ?

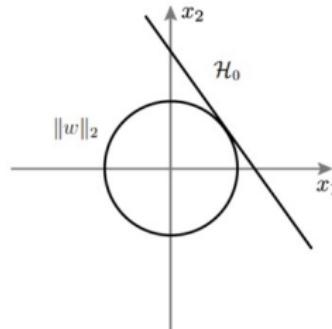
$$\mathbf{A} : \min \|x\|_1 \text{ subject to } Ax = y. \quad (8)$$

$$\mathbf{B} : \min \|x\|_2 \text{ subject to } Ax = y \quad (9)$$

**A** L1 regularization



**B** L2 regularization

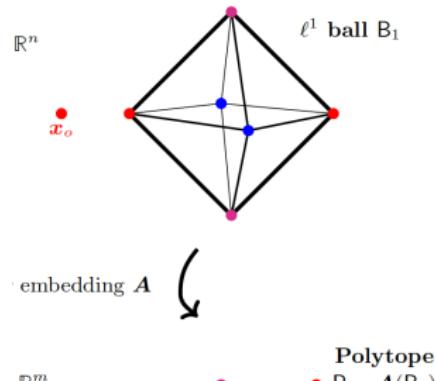


$\ell^1$  picks out **sparse** signals, because the  $\ell^1$  ball is pointy!

# Theory: Isometry Principles

Say that  $A$  satisfies the **restricted isometry property** of order  $k$  with coefficient  $\delta$  if for all  $k$ -sparse  $x$ ,

$$(1 - \delta)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1 + \delta)\|x\|_2^2.$$



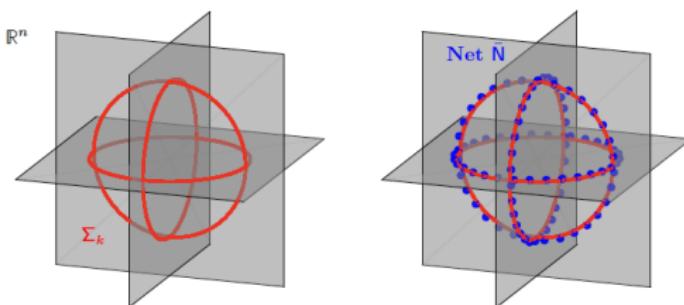
Theorem (RIP  $\implies$   $\ell^1$  succeeds)

Suppose that  $\delta_{2k}(A) < \sqrt{2} - 1$ . Then  $\ell^1$  minimization recovers any  $k$ -sparse signal  $x$ !

# Theory: Random Sensing

Theorem (RIP of Gaussian Matrices)

If  $\mathbf{A} \in \mathbb{R}^{m \times n}$  with entries independent  $\mathcal{N}(0, \frac{1}{m})$  random variables, with high probability,  $\delta_k(\mathbf{A}) < \delta$ , provided  $m \geq Ck \log(n/k)/\delta^2$ .



$\implies \ell^1$ -minimization recovers  $k$ -sparse vectors from about  $k \log(n/k)$  measurements (nearly minimal)!

**Extensions:** other distributions, structured random matrices.

# From Sparse Recovery to Low-Rank Recovery

Recovering a sparse signal  $x_o$ :

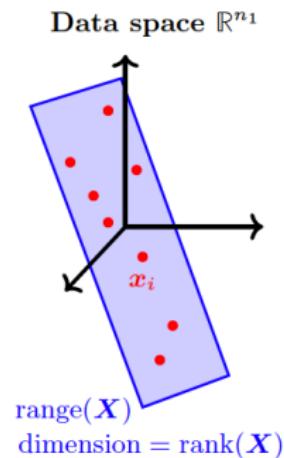
$$\begin{matrix} \mathbf{y} \\ \text{observation} \end{matrix} = \mathbf{A} \begin{matrix} \mathbf{x}_o \\ \text{unknown} \end{matrix}$$

where  $\mathbf{A} \in \mathbb{R}^{m \times n}$  is a linear map.

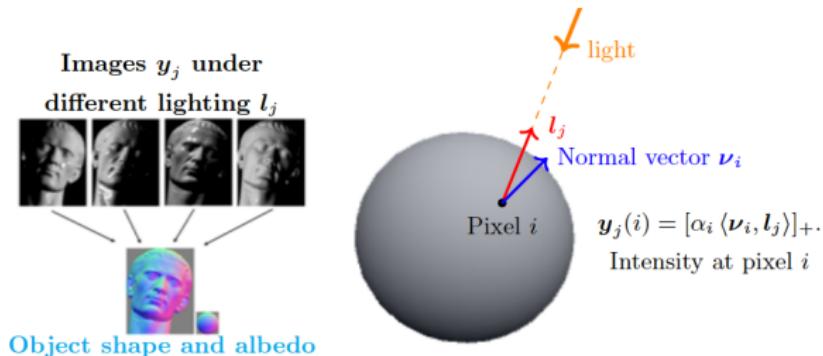
Recovering a low-rank matrix  $X_o$ :

$$\begin{matrix} \mathbf{y} \\ \text{observation} \end{matrix} = \mathcal{A} \begin{bmatrix} \mathbf{X}_o \\ \text{unknown} \end{bmatrix}$$

where  $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \rightarrow \mathbb{R}^m$  is a linear map.



# Low-Rank I: Rank and Geometry



Multiple images of a Lambertian object with varying light:

$$\mathbf{Y} = \mathcal{P}_\Omega[\mathbf{NL}], \quad \mathbf{X} = \mathbf{NL} \text{ has rank 3.}$$

Low-rank model from **physical constraints** (3 degrees of freedom in point illumination)

See also: multiview geometry, system identification, sensor positioning...

## Low-Rank II: Rank and Collaborative Filtering

The diagram illustrates the process of completing user-item rating matrices. On the left, four user icons are aligned vertically, labeled "Users". Below them is a 4x4 matrix labeled "Observed (Incomplete) Ratings  $\mathbf{Y}$ ". The matrix contains numerical values (e.g., 5, 3, ...) and question marks. To the right of the matrix is an equals sign followed by the formula  $\mathcal{P}_{\Omega} \begin{pmatrix} \mathbf{X} \\ \text{Complete ratings } \mathbf{X} \end{pmatrix}$ . This formula shows the matrix  $\mathbf{X}$  being projected onto the set  $\Omega$  to produce the "Complete ratings  $\mathbf{X}$ ". Below the matrix  $\mathbf{Y}$ , the word "Items" is written above three book icons, which are labeled "Books".

$$\underset{\text{Users}}{\begin{matrix} \text{User 1} \\ \text{User 2} \\ \vdots \\ \text{User 4} \end{matrix}} \underset{\text{Items}}{\begin{matrix} \text{Book 1} \\ \text{Book 2} \\ \dots \\ \text{Book 4} \end{matrix}} \underset{\text{Observed (Incomplete) Ratings } \mathbf{Y}}{\begin{bmatrix} 5 & 3 & \dots & ? \\ ? & 2 & \dots & 4 \\ \vdots & \vdots & \ddots & \vdots \\ 5 & ? & \dots & ? \end{bmatrix}} = \mathcal{P}_{\Omega} \begin{pmatrix} \mathbf{X} \\ \text{Complete ratings } \mathbf{X} \end{pmatrix}$$

We observe:

Observed (Incomplete) Ratings  $\mathbf{Y}$

$$\underset{\text{Observed ratings}}{\mathbf{Y}} = \mathcal{P}_{\Omega} \begin{pmatrix} \mathbf{X} \\ \text{Complete ratings} \end{pmatrix},$$

where  $\Omega \doteq \{(i, j) \mid \text{user } i \text{ has rated product } j\}$ .

**Low-rank model:** user preferences are linearly correlated; **a few factors** predict preferences ( $Y_{ij} = u_i^T v_j$ , with  $u_i, v_j \in \mathbb{R}^r$ ).

**See also:** latent semantic analysis, topic modeling...

# Rank and Singular Value Decomposition

Theorem (Compact SVD)

Let  $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2}$  be a matrix, and  $r = \text{rank}(\mathbf{X})$ . Then there exist  $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_r)$  with numbers  $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r > 0$  and matrices  $\mathbf{U} \in \mathbb{R}^{n_1 \times r}$ ,  $\mathbf{V} \in \mathbb{R}^{n_2 \times r}$ , such that  $\mathbf{U}^*\mathbf{U} = \mathbf{I}$ ,  $\mathbf{V}^*\mathbf{V} = \mathbf{I}$  and

$$\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^* = \sum_{i=1}^r \sigma_i \mathbf{u}_i \mathbf{v}_i^*.$$

**Low-rank is sparsity of the singular values:**  $\text{rank}(\mathbf{X}) = \|\boldsymbol{\sigma}(\mathbf{X})\|_0!$

Many of the same tools and ideas apply!

Computing SVD: Nice Nonconvex Problem (Lecture 3)

# Affine Rank Minimization

**Problem:** recover a low-rank matrix  $\mathbf{X}_o$  from linear measurements:

$$\min \text{rank}(\mathbf{X}) \quad \text{subject to} \quad \mathcal{A}[\mathbf{X}] = \mathbf{y}$$

where  $\mathbf{y} \in \mathbb{R}^m$  is an observation and  $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \rightarrow \mathbb{R}^m$  is linear.

General linear map:  $\mathcal{A}[\mathbf{X}] = (\langle \mathbf{A}_1, \mathbf{X} \rangle, \dots, \langle \mathbf{A}_m, \mathbf{X} \rangle)$ ,  $\mathbf{A}_i \in \mathbb{R}^{n_1 \times n_2}$ .

**NP-Hard in general**, by reduction from  $\ell^0$  minimization, using that

$$\text{rank}(\mathbf{X}) = \|\boldsymbol{\sigma}(\mathbf{X})\|_0.$$

Let's seek a tractable surrogate...

# Convex Relaxation: Nuclear Norm Minimization

Replace the rank, which is the  $\ell^0$  norm  $\sigma(\mathbf{X})$  with the  $\ell^1$  norm of  $\sigma(\mathbf{X})$ :

$$\text{Nuclear norm: } \|\mathbf{X}\|_* \doteq \|\sigma(\mathbf{X})\|_1 = \sum_i \sigma_i(\mathbf{X}).$$

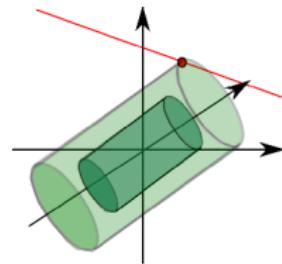
Also known as the *trace norm*, *Schatten 1-norm*, and *Ky-Fan k-norm*.

## Nuclear norm minimization problem:

$$\min \|\mathbf{X}\|_* \quad \text{subject to} \quad \mathcal{A}[\mathbf{X}] = \mathbf{y}.$$

Geometry of nuclear norm minimization:

$$\text{Nuclear norm ball } \mathcal{B}_* = \{\mathbf{X} \mid \|\mathbf{X}\|_* \leq 1\}$$



# Low-Rank Recovery with Generic Measurements

- **Rank Restricted Isometry Property:** for all rank- $r$   $\mathbf{X}$ ,

$$(1 - \delta) \|\mathbf{X}\|_F \leq \|\mathcal{A}[\mathbf{X}]\| \leq (1 + \delta) \|\mathbf{X}\|_F$$

- **Rank RIP  $\implies$  accurate recovery:** if  $\delta_{4r}(\mathcal{A}) \leq \sqrt{2} - 1$ , nuclear norm minimization recovers any rank- $r$   $\mathbf{X}_o$ .
- **Random linear maps have rank-RIP if**

$$\mathcal{A}[\mathbf{X}] = (\langle \mathbf{A}_1, \mathbf{X} \rangle, \dots, \langle \mathbf{A}_m, \mathbf{X} \rangle)$$

with  $\mathbf{A}_1, \dots, \mathbf{A}_m$  independent Gaussian matrices,  $\mathcal{A}$  has rank-RIP with high probability when  $m \geq C(n_1 + n_2)r/\delta^2$ .

Nuclear norm minimization recovers low-rank matrices from **near minimal** number  $m \sim r(n_1 + n_2 - r)$  of **generic measurements**.

## Generic vs. Structured Measurements

$$y_i = \left\langle \begin{bmatrix} \text{[Colorful Grid]} \end{bmatrix}, \mathbf{X}_o \right\rangle$$

$\mathbf{A}_i$  random

Matrix Sensing

$$y_i = \left\langle \begin{bmatrix} \text{[Black Box with White Square]} \end{bmatrix}, \mathbf{X}_o \right\rangle$$

$\mathbf{A}_i = \mathbf{E}_{u_i, v_i}$

Matrix Completion

$$\begin{bmatrix} 5 & 3 & \dots & ? \\ ? & 2 & \dots & 4 \\ \vdots & \vdots & \ddots & \vdots \\ 5 & ? & \dots & ? \end{bmatrix}$$

**Rank-RIP:** no low-rank  $\mathbf{X}$  in  $\text{null}(\mathcal{A})$ .

**Matrix completion:**  $\exists$  rank-1  $\mathbf{X}$  in  $\text{null}(\mathcal{A})$ . E.g.,  $\mathbf{X} = \mathbf{E}_{ij}$ ,  $(i, j) \notin \Omega$ .

⇒ **Matrix completion** does not have restricted isometry property!

**Analogous instances:** superresolution of point sources, sparse spike deconvolution, analysis of dictionary learning methods.

# Theory for Matrix Completion

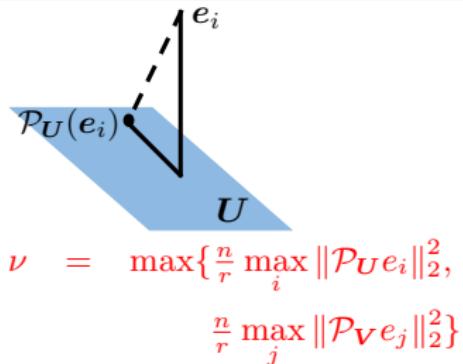
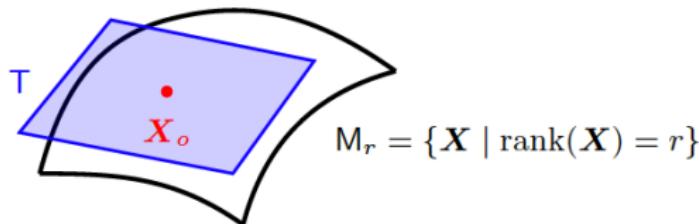
## Theorem

With high probability, nuclear norm minimization recovers an  $n \times n$ ,  $\nu$ -incoherent, rank- $r$  matrix from a random subset of entries, of size

$$m \geq Cnr\nu \log^2 n.$$

Restrict to **incoherent**  $X_o$   
(not concentrated on a few entries!)

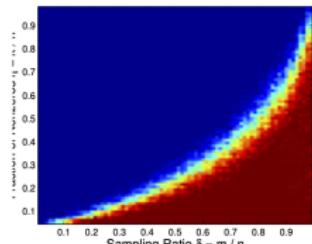
Proof ideas: **local isometry** plus clever  
use of **convexity and probability**.



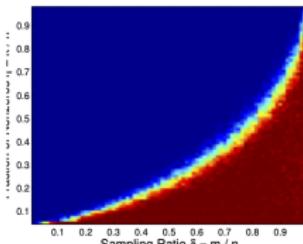
# Parallelism between Rank and Sparsity

|                          | Sparse Vector                                    | Low-rank Matrix                                     |
|--------------------------|--|---|
| Low-dimensionality of    | individual signal $\mathbf{x}$                   | a set of signals $\mathbf{X}$                       |
| Compressive sensing      | $\mathbf{y} = \mathbf{A}\mathbf{x}$              | $\mathbf{Y} = \mathcal{A}(\mathbf{X})$              |
| Low-dim measure          | $\ell^0$ norm $\ \mathbf{x}\ _0$                 | $\text{rank}(\mathbf{X})$                           |
| Convex surrogate         | $\ell^1$ norm $\ \mathbf{x}\ _1$                 | nuclear norm $\ \mathbf{X}\ _*$                     |
| Success conditions (RIP) | $\delta_{2k}(\mathbf{A}) \geq \sqrt{2} - 1$      | $\delta_{4r}(\mathbf{A}) \geq \sqrt{2} - 1$         |
| Random measurements      | $m = O(k \log(n/k))$                             | $m = O(nr)$   |
| Stable/Inexact recovery  | $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{z}$ | $\mathbf{Y} = \mathcal{A}(\mathbf{X}) + \mathbf{Z}$ |
| Phase transition at      | Stat. dim. of descent cone: $m^* = \delta(D)$    |   |

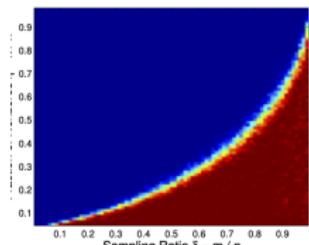
# Sharp Phase Transitions with Gaussian Measurements



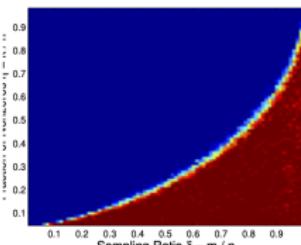
$n = 50$



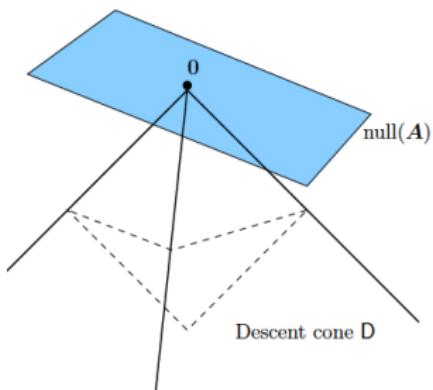
$n = 100$



$n = 200$



$n = 400$



**High dimensions** (large  $n$ ): sharp line between success and failure!

Beautiful math: convex polytopes, conic geometry, high-D probability.

# Noise and Inexact Structure

**Observation:**  $y = Ax_o + z$ , with  $x_o$  structured, and  $z$  noise.

**Goal:** produce  $\hat{x}$  as close to  $x_o$  as possible! Relax:

- **Lasso** for stable sparse recovery

$$\min_{\mathbf{x}} \frac{1}{2} \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 + \mu \|\mathbf{x}\|_1$$

- **Matrix Lasso** for stable low-rank recovery

$$\min_{\mathbf{X}} \frac{1}{2} \|\mathcal{A}[\mathbf{X}] - \mathbf{y}\|_2^2 + \mu \|\mathbf{X}\|_*$$

**Wealth of statistical results:** if  $A$  “nice” (say, RIP or RSC) ...

- (i) Deterministic noise:  $\|\hat{x} - x_o\| \leq C\|z\|_2$
- (ii) Stochastic noise:  $\|\hat{x} - x_o\| \leq C\sigma\sqrt{k \log n/m}$ .
- (iii) Inexact structure:  $\|\hat{x} - x_o\| \leq C\|x_o - [x_o]_k\|$ .

# Parallelism between Rank and Sparsity

|                          | Sparse Vector                                    | Low-rank Matrix                                     |
|--------------------------|--|---|
| Low-dimensionality of    | individual signal $\mathbf{x}$                   | a set of signals $\mathbf{X}$                       |
| Compressive sensing      | $\mathbf{y} = \mathbf{A}\mathbf{x}$              | $\mathbf{Y} = \mathcal{A}(\mathbf{X})$              |
| Low-dim measure          | $\ell^0$ norm $\ \mathbf{x}\ _0$                 | $\text{rank}(\mathbf{X})$                           |
| Convex surrogate         | $\ell^1$ norm $\ \mathbf{x}\ _1$                 | nuclear norm $\ \mathbf{X}\ _*$                     |
| Success conditions (RIP) | $\delta_{2k}(\mathbf{A}) \geq \sqrt{2} - 1$      | $\delta_{4r}(\mathbf{A}) \geq \sqrt{2} - 1$         |
| Random measurements      | $m = O(k \log(n/k))$                             | $m = O(nr)$   |
| Stable/Inexact recovery  | $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{z}$ | $\mathbf{Y} = \mathcal{A}(\mathbf{X}) + \mathbf{Z}$ |
| Phase transition at      | Stat. dim. of descent cone: $m^* = \delta(D)$    |   |

## Combining Rank and Sparsity: Robust PCA?

$$\begin{bmatrix} \text{Image 1} & \dots & \text{Image n} \end{bmatrix} = \begin{bmatrix} \text{Image 1} & \dots & \text{Image n} \end{bmatrix} + \begin{bmatrix} \text{Image 1} & \dots & \text{Image n} \end{bmatrix}$$

Observation  $\mathbf{Y}$       Low-rank Matrix  $\mathbf{L}_o$       Sparse Error  $\mathbf{S}_o$

Given  $\mathbf{Y} = \mathbf{L}_o + \mathbf{S}_o$ , with  $\mathbf{L}_o$  low-rank,  $\mathbf{S}_o$  sparse, recover  $(\mathbf{L}_o, \mathbf{S}_o)$ .

A robust counterpart to classical principal component analysis:

**Classical PCA:** Low-rank + small noise

**Matrix Completion:** Low-rank from a subset of entries

**Low-rank and Sparse:** Low-rank + gross errors

# Low-rank + Sparse I: Video

A sequence of video frames can be modeled as a static background (low-rank) and moving foreground (sparse).



(a) Original frames

(b) Low-rank  $\hat{L}$

(c) Sparse  $\hat{S}$

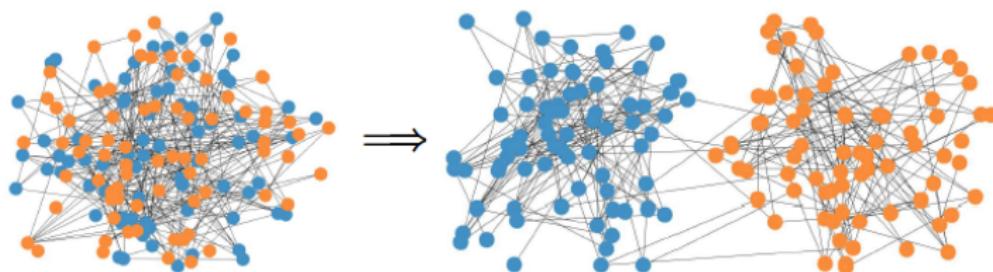
## Low-rank + Sparse II: Faces

A set of face images of the same person under different lightings can be modeled as a low-dimensional,  $3 \sim 9d$ , subspace and sparse occlusions and corruptions (specularities).



## Low-rank + Sparse III: Communities

Finding communities in a large social networks. Each community can be modeled as a clique of the social graph  $\mathcal{G}$ , hence a rank-1 block in the connectivity matrix  $M$ . Hence  $M$  is a low-rank matrix and some sparse connections across communities.



## Low-rank + Sparse: Convex Relaxations

Optimization formulation:

$$\text{minimize} \quad \text{rank}(\mathbf{L}) + \lambda \|\mathbf{S}\|_0 \quad \text{subject to} \quad \mathbf{L} + \mathbf{S} = \mathbf{Y},$$

which is intractable. Consider **convex relaxation**:

$$\|\mathbf{S}\|_0 \rightarrow \|\mathbf{S}\|_1, \quad \text{rank}(\mathbf{L}) = \|\boldsymbol{\sigma}(\mathbf{L})\|_0 \rightarrow \|\mathbf{L}\|_*$$

$$\text{minimize} \quad \|\mathbf{L}\|_* + \lambda \|\mathbf{S}\|_1 \quad \text{subject to} \quad \mathbf{L} + \mathbf{S} = \mathbf{Y}.$$

- **Theory:** recovery, e.g., when  $\mathbf{L}_o$  incoherent,  $\mathbf{S}_o$  random sparse.
- **Efficient, scalable methods:** see Lecture 2 and course resources.

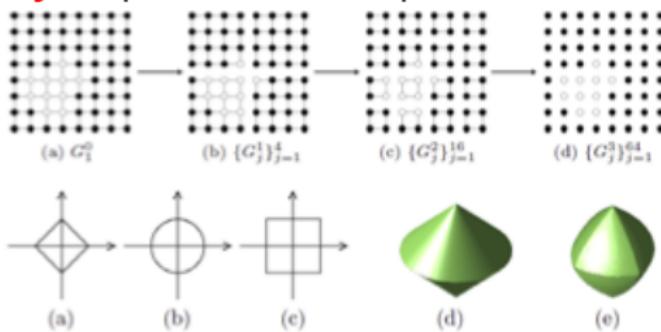
# General Low-Dimensional Models

## Atomic Norms and Structured Sparsity

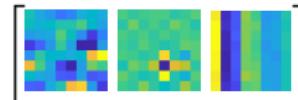
**Atomic Norm:** for a set of atoms  $\mathcal{D}$ ,  $\|x\|_{\diamond} = \inf\{\sum_i c_i \mid \sum_i c_i d_i = x\}$

- **Sparsity:**  $\mathcal{D} = \{e_i\}$ ,
- **Low-rank:**  $\mathcal{D} = \{uv^T\}$ ,
- **Column sparse matrices:**  $\mathcal{D} = \{ue_j^T\}$ ,
- **Sinusoids:**  $\mathcal{D} = \{\exp(i(2\pi ft + \xi))\}$ ,
- **Tensors:**  $\mathcal{D} = \{u_1 \otimes u_2 \otimes \dots \otimes u_N\}$ ,

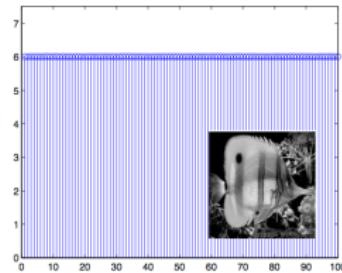
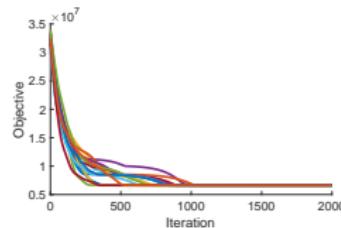
**Structured Sparsity:** capture relationship between nonzeros



# Learned Low-Dimensional Models: Dictionary Learning, Deconvolution



$$\min \quad f(\mathbf{A}, \mathbf{X}) \doteq \frac{1}{2} \|\mathbf{Y} - \mathbf{AX}\|_F^2 + \lambda \|\mathbf{X}\|_1, \quad \text{s.t. } \mathbf{A} \in O_n$$



The same **modeling toolkit**, but optimization formulations become **nonconvex**! (see Lecture 4)

# Nonlinear Low-Dimensional Models

**Nonlinear Observations:** Transformed low-rank texture



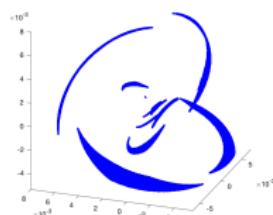
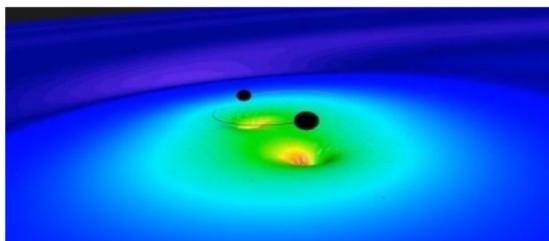
(a) Low-rank texture  $\mathbf{I}_o$



(b) Its image  $\mathbf{I}$  under a different viewpoint

$\xrightarrow{\tau}$

**Nonlinear (Manifold) Structure:** Gravitational wave astronomy



Nonconvex optimization + deep networks as tools for **Linearizing Nonlinear Low-d Structure!** (see Lectures 3,5-7)

# Conclusion and Coming Attractions

- **Models:** Sparse and Low-rank provide a flexible toolkit for modeling high-dimensional signals
- **Sample Complexity:** Structured signals can be recovered from near-minimal measurements  $m \sim \#\text{dof}(\mathbf{x})$ .
- **Tractable Computation:** Convex relaxations  $\ell^1$ , nuclear norm
- **Extensions:** Combinations, learned dictionaries, nonlinear structures.

**Next lecture:** low-dimensionality meets deep networks [Atlas Wang].

Thank You! Questions?