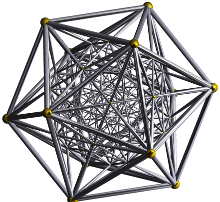
ICASSP 2023 Short Course

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**

Sensing, Compression

$\downarrow$

Denoising, Deblurring, Superresolution

$\downarrow$

Source Separation, Error Correction

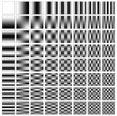
$\downarrow$

Inference, Prediction

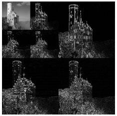
The **pursuit of low-dimensional structure** is a universal task!

Historical Context: **Quest for Low-Dimensionality**

Fourier

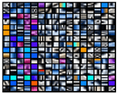


Wavelets



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

Learned Dictionaries

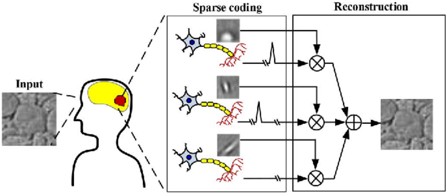


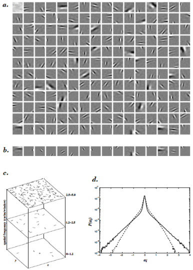
Learned Reconstruction Procedures

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

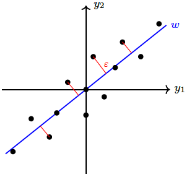
$$\boldsymbol{y} = \sum_{i=1}^{n} x_i \boldsymbol{a}_i + \boldsymbol{\epsilon} \quad \in \mathbb{R}^m, \qquad (1)$$

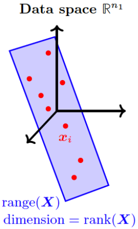
[Nature, Olshausen and Field 1996.]

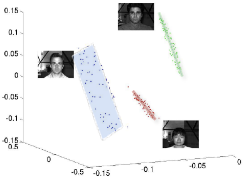
Historical Context: **Sparse and Low-d in Statistics**

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]

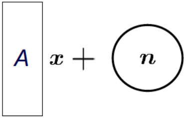
Historical Context: **Estimation, Errors, Missing Data**

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

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

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

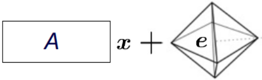
C. Gauss. *Theory of motion of heavenly bodies, 1809*



over-determined
+ dense, Gaussian

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

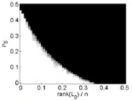
B. Logan. *Properties of High-Pass Signals*, 1965



underdetermined
+ sparse, Laplacian

⋮

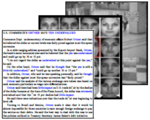
The Modern Era: **Massive Data and Computation**



(a) Robust PCA, Random Signs

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

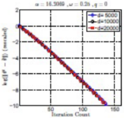
BIG DATA
(images, videos, voices, texts, biomedical, geospatial, consumer data…)



Cloud Computing
(parallel, distributed, scalable platforms)

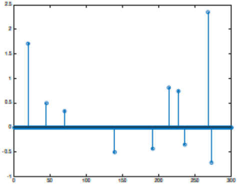
Applications & Services
(data processing, analysis, compression, knowledge discovery, search, recognition…)

Computational Methods
(convex optimization, first-order algorithms, random sampling, deep networks…)

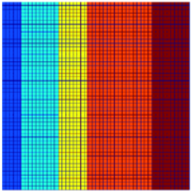


Motivating Issues I: **Correctness?**

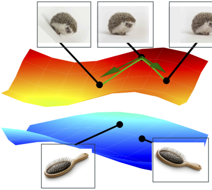
How can we **correctly** compute with **low-dimensional structure**?



Sparse Vectors



Low-rank Matrices

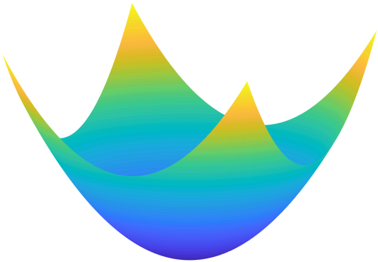


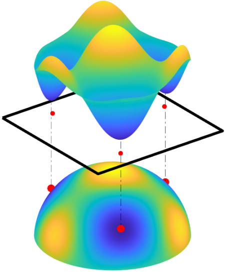
Nonlinear Manifolds

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

Motivating Issues II: **Computational Efficiency?**

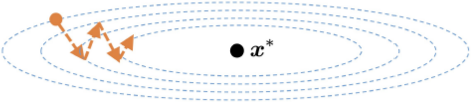
Computational Tractability: easy vs./ hard problems:





Convexity **Benign Nonconvexity**

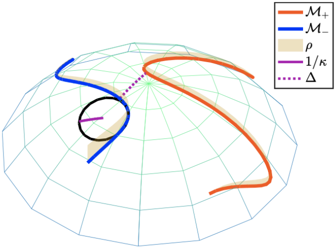
Efficient, scalable methods leveraging problem geometry:

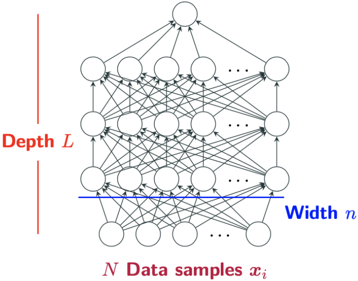


gradient descent

Motivating Issues III: **Resource Efficiency?**

Data Efficiency: How many samples? How many labels?
Architecture Efficiency: How deep? How wide? What operations?

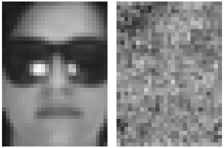




Low-d. structure of data sets fundamental resource requirements for **sensing** and **learning**.

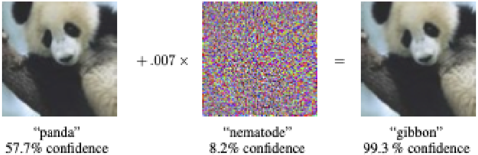
Motivating Issues IV – **Robustness?**

Robustness: to errors, outliers, missing data:



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

Robustness and deep networks?

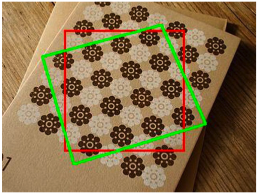


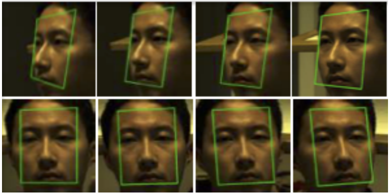
From [Goodfellow, Shlens and Szegedy, 2015]

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

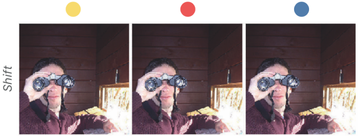
Motivating Issues V: **Invariance?**

Transformations of the signal domain:

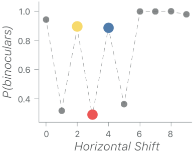




can cause still lead to disturbing failures:



From [Azulay and Weiss, 2019]



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

This Course: **The Plan**

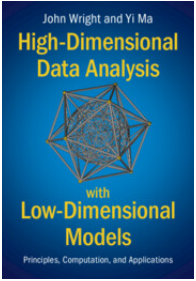
- Lecture 1: Introduction to Low-D Models
- Lecture 2 (today): Low-D in Neural Networks: Practice and Theory
- Lecture 3 (6/7): *Designing* Deep Networks for Low-D Structure
- Lecture 4 (6/7): Nonconvex Optimization for Low-D Structure
- Lecture 5-7 (6/8-9): *Learning* Deep Networks for Low-D Structure

This Tutorial: **Resources**

High-Dimensional Data Analysis
with Low-Dimensional Models

Principles, Computation, and Applications

John Wright and Yi Ma
Cambridge University Press, 2022.

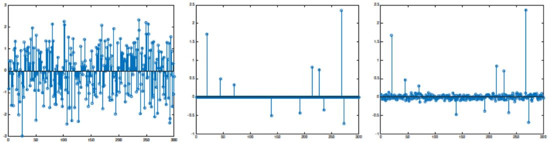


Preproduction Copy from Website: https://book-wright-ma.github.io
Slides, Code, etc: https://book-wright-ma.github.io/Lecture-Slides/

Tutorial Website: tutorial slides, code, etc.:
https://highdimdata-lowdimmodels-tutorial.github.io

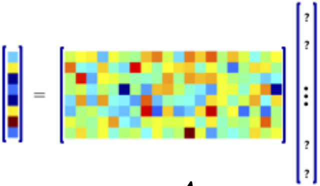
Sparse Signal Models

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



dense vector   sparse vector   compressible vector

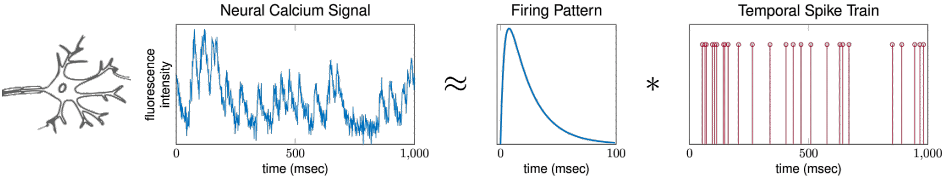
Sparse Recovery: Given *linear measurements* $\boldsymbol{y} \in \mathbb{R}^m$ of a sparse signal $\boldsymbol{x}_o$:



$$\underset{\text{observation}}{\boldsymbol{y}} = \underset{\text{measurement matrix}}{\boldsymbol{A}} \; \underset{\text{unknown}}{\boldsymbol{x}_o}$$

recover $\boldsymbol{x}_o$.

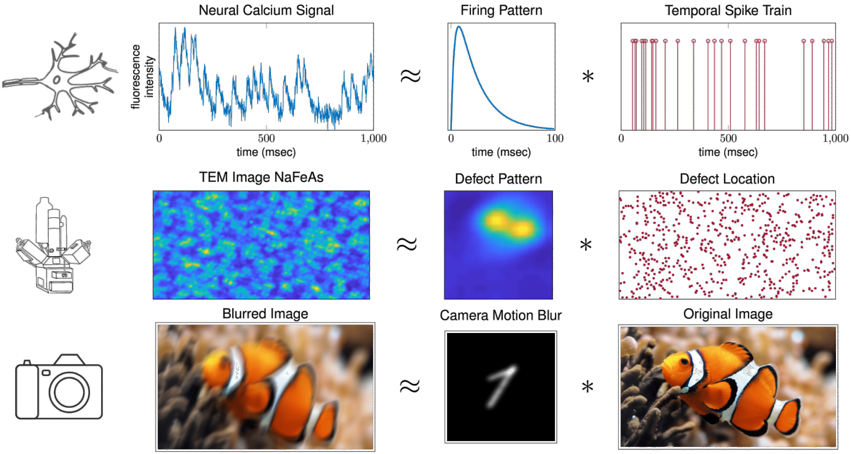
Sparsity I: **Neural Spikes**



$$\underset{\text{observation}}{\boldsymbol{y}} = \underset{\text{firing pattern}}{\boldsymbol{a}} * \underset{\text{sparse spike train}}{\boldsymbol{x}} + \underset{\text{noise}}{\boldsymbol{z}}.$$

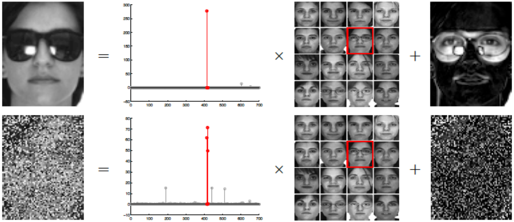
Sparse and low-dimensional models arise naturally from **physical structure** of data!

Sparsity I: **Neural Spikes and Beyond**



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

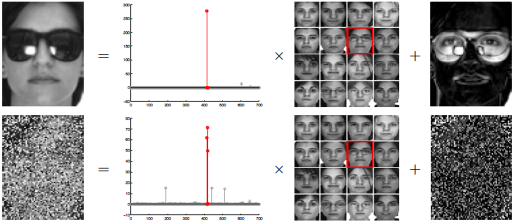
Sparsity II: **Faces and Error Correction**



$$\underset{\text{observation}}{\boldsymbol{y}} = \underset{\text{clean data}}{\boldsymbol{y}_o} + \underset{\text{sparse error}}{\boldsymbol{e}} \in \mathbb{R}^m.$$

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

Sparsity II: **Faces and Error Correction**



$$\underbrace{\boldsymbol{y}}_{\text{observation}} = \underbrace{\boldsymbol{y}_o}_{\text{clean data}} + \underbrace{\boldsymbol{e}}_{\text{sparse error}} \in \mathbb{R}^m.$$

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

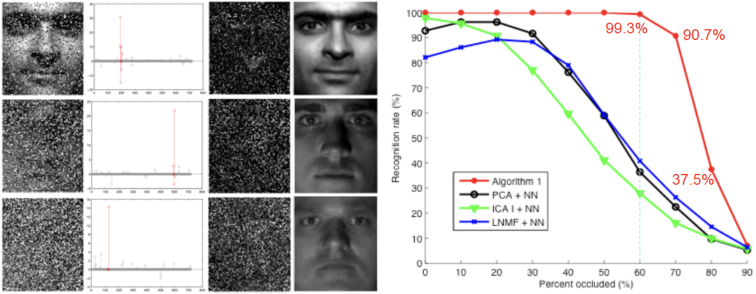
Concatenate gallery images of n subjects into a large "dictionary":

$$\boldsymbol{B} = \underbrace{[\boldsymbol{B}_1 \mid \boldsymbol{B}_2 \mid \cdots \mid \boldsymbol{B}_n]}_{\text{all training images}} \in \mathbb{R}^{m \times n}$$

Sparsity II: **Faces and Error Correction**

Find sparse solutions $(\boldsymbol{x}, \boldsymbol{e})$ to the linear system:

$$\boldsymbol{y} \;=\; \boldsymbol{B}\boldsymbol{x} + \boldsymbol{e} \;=\; [\boldsymbol{B}, \boldsymbol{I}]\left[\begin{smallmatrix}\boldsymbol{x}\\ \boldsymbol{e}\end{smallmatrix}\right].$$



Correcting Gross Errors is also a sparse recovery problem!

Sparsity III: **Magnetic Resonance Imaging**

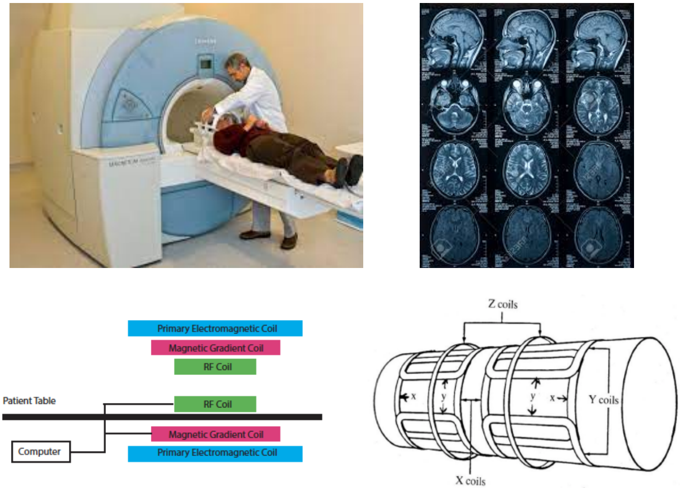


Figure: Left: Key components. Right: The three-axis gradient coils.

Sparsity III: **Magnetic Resonance Imaging**

Simplified mathematical model for MRI:

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

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

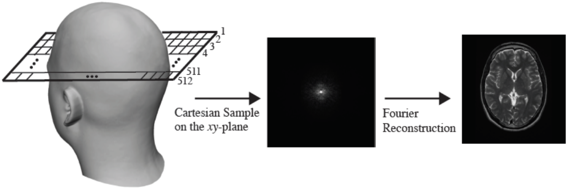


Figure: Recovering MRI image from Fourier measurements.

Sparsity III: **Structure of MR Images**

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

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

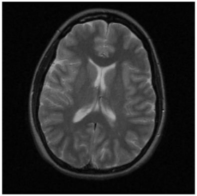
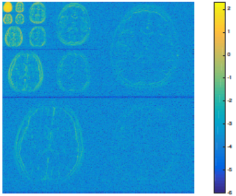
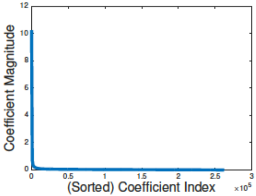


image $I(v)$





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

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

Sparsity III: **Image Reconstruction by Sparse Recovery**

$$\begin{aligned}
\underset{\text{observed Fourier coefficients}}{\boldsymbol{y}} &= \mathcal{F}_{\mathsf{U}}[I], \\
&= \mathcal{F}_{\mathsf{U}}\Big[\ \boldsymbol{\psi}_1 x_1 + \ \cdots \ + \boldsymbol{\psi}_{N^2} x_{N^2}\ \Big], \\
&= \mathcal{F}_{\mathsf{U}}[\boldsymbol{\psi}_1]\, x_1 \ + \ \cdots \ + \ \mathcal{F}_{\mathsf{U}}[\boldsymbol{\psi}_{N^2}]\, x_{N^2}, \\
&= \underbrace{\Big[\mathcal{F}_{\mathsf{U}}[\boldsymbol{\psi}_1] \mid \cdots \mid \mathcal{F}_{\mathsf{U}}[\boldsymbol{\psi}_{N^2}]\Big]}_{\text{matrix } \boldsymbol{A} \in \mathbb{R}^{m \times N^2},\ m \ll N^2.} \boldsymbol{x}, \\
&= \boldsymbol{A}\boldsymbol{x}. \qquad (2)
\end{aligned}$$

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 $I_{\text{noisy}} = I_{\text{clean}} + \boldsymbol{z}$ ... break into patches $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_p$:

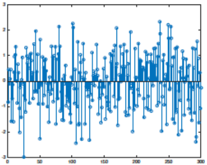
$$\boldsymbol{y}_i \;=\; \boldsymbol{y}_{i\text{clean}} + \boldsymbol{z}_i = \underset{\text{patch dictionary}}{\boldsymbol{A}} \times \underset{\text{sparse coefficient vector}}{\boldsymbol{x}_i} + \boldsymbol{z}_i.$$

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

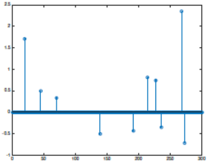
Natural signals are challenging to model analytically $\Longrightarrow$ can **learn the sparse model** from data!

Figure: [Mairal, Elad, Sapiro '08]

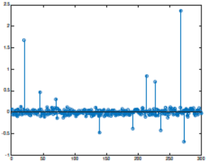
Measuring Sparsity: ℓ^0 Norm



dense vector



sparse vector



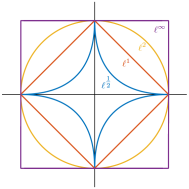
compressible vector

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

Connection to ℓ^p norms

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

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



The ℓ^p balls.

Sparse Recovery: ℓ^0 minimization

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

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

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

$$\boldsymbol{A}_{\mathsf{I}}\boldsymbol{x}_{\mathsf{I}} = \boldsymbol{y}? \quad \forall \mathsf{I} \subseteq \{1, \ldots, n\}, \ |\mathsf{I}| \le k.$$

Sparse Recovery: ℓ^0 minimization

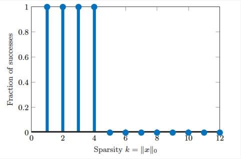
Computational Principle: seek the **sparsest** signal consistent with our observations:

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

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

$$\boldsymbol{A}_{\mathsf{I}}\boldsymbol{x}_{\mathsf{I}} = \boldsymbol{y}? \quad \forall \mathsf{I} \subseteq \{1, \ldots, n\}, \ |\mathsf{I}| \le k.$$

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

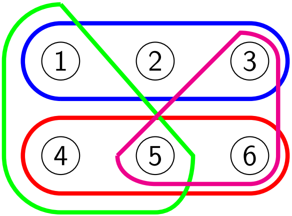


ℓ^0 Minimization is NP-hard

Theorem (Hardness of ℓ^0 Minimization)

The ℓ^0-minimization problem $\min \|\boldsymbol{x}\|_0$ s.t. $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}$ is (strongly) **NP***-hard.*

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



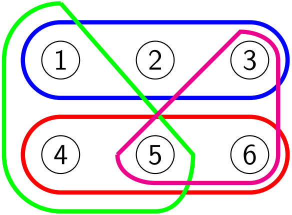
$$\underset{\boldsymbol{y}}{\begin{bmatrix}1\\1\\1\\1\\1\\1\end{bmatrix}} = \underset{\boldsymbol{A}}{\begin{bmatrix}1&1&0&0\\1&0&0&0\\1&0&0&1\\0&1&1&0\\0&1&1&1\\0&0&1&1\end{bmatrix}} \underset{\boldsymbol{x}}{\begin{bmatrix}1\\0\\1\\0\end{bmatrix}}$$

ℓ^0 Minimization is NP-hard

Theorem (Hardness of ℓ^0 Minimization)

The ℓ^0-minimization problem $\min \|\boldsymbol{x}\|_0$ s.t. $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}$ is (strongly) **NP***-hard.*

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



$$\underset{\boldsymbol{y}}{\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}} = \underset{\boldsymbol{A}}{\begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix}} \underset{\boldsymbol{x}}{\begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix}}$$

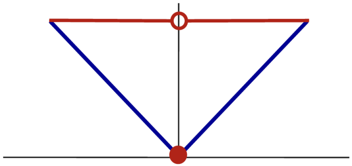
In high dimensions, need to pay attention to *both* **statistical and computational efficiency**!

Convex Relaxation: ℓ^1 Minimization

Intuitive reasons why ℓ^0 minimization:

$$\min \|\boldsymbol{x}\|_0 \quad \text{subject to} \quad \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}. \tag{3}$$

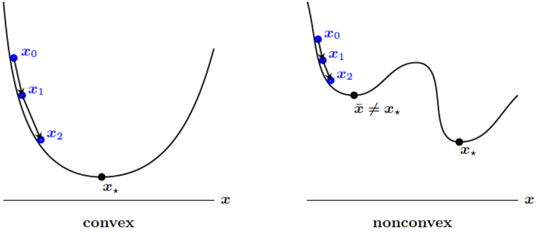
is very challenging:



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

Convex Relaxation: ℓ^1 Minimization

For minimizing a generic function: $\min f(\boldsymbol{x}), \boldsymbol{x} \in \mathsf{C}$ (a convex set), **local methods:** $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - t\nabla f(\boldsymbol{x}_k)$ 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: ℓ^1 Minimization

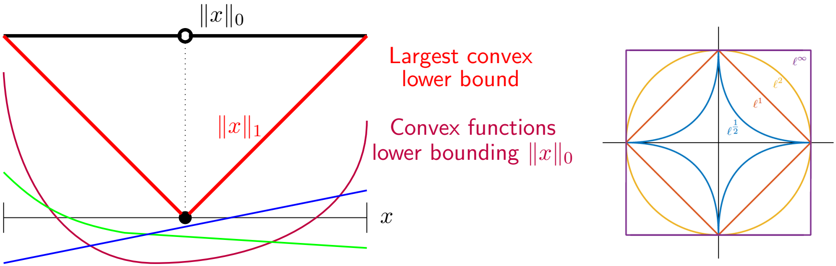


Figure: **Convex surrogates for the ℓ^0 norm.** $\|\boldsymbol{x}\|_1$ is the *convex envelope* of $\|x\|_0$ on B_∞.

Efficient **convex relaxation**:

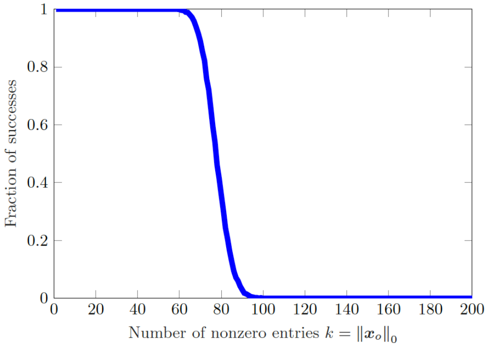
$$\min \|\boldsymbol{x}\|_1 \quad \text{subject to} \quad \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}.$$

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

Minimizing the ℓ^1 Norm: Simulations

$$\textbf{Solve:} \quad \min \|\boldsymbol{x}\|_1 \quad \text{s.t.} \quad \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}. \tag{4}$$

$\boldsymbol{A}$ is of size 200×400. Fraction of success across 50 trials.



Experiment: ℓ^1 minimization recovers *any sufficiently sparse signal?*

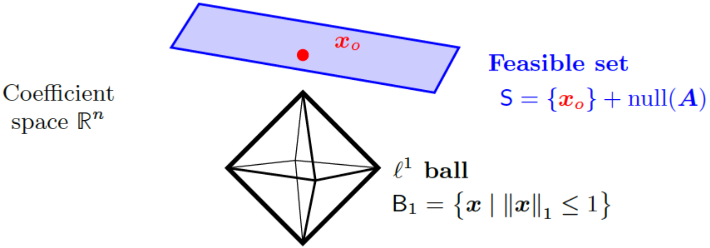
Geometric Intuition: Coefficient Space

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

$$\min \|\boldsymbol{x}\|_1 \quad \text{subject to} \quad \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}. \tag{5}$$

The space of all feasible solutions is an affine subspace:

$$\mathsf{S} = \{\boldsymbol{x} \mid \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}\} = \{\boldsymbol{x}_o\} + \mathrm{null}(\boldsymbol{A}) \quad \subset \mathbb{R}^n. \tag{6}$$

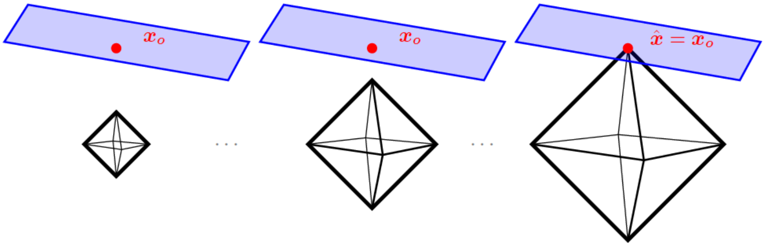


Geometric Intuition: Coefficient Space

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

$$t \cdot \mathsf{B}_1 = \{\boldsymbol{x} \mid \|\boldsymbol{x}\|_1 \leq t\} \quad \subset \mathbb{R}^n, \tag{7}$$

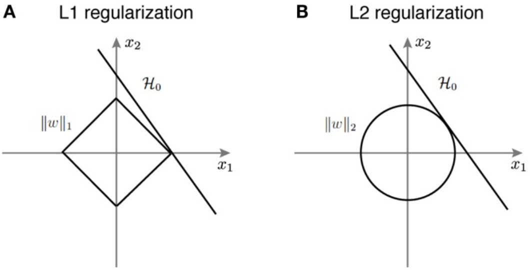
till its boundary first touches the feasible set S:



Geometric Intuition: ℓ^1 vs. ℓ^2?

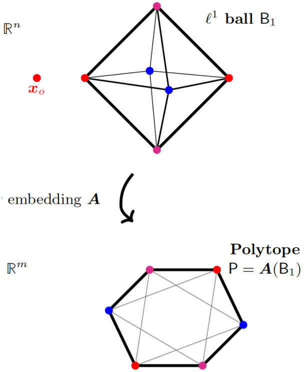
$$\mathbf{A}: \quad \min \|\boldsymbol{x}\|_1 \quad \text{subject to} \quad \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}. \tag{8}$$

$$\mathbf{B}: \quad \min \|\boldsymbol{x}\|_2 \quad \text{subject to} \quad \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y} \tag{9}$$



ℓ^1 picks out sparse signals, because the ℓ^1 ball is pointy!

Theory: Isometry Principles



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

$$(1-\delta)\|\boldsymbol{x}\|_2^2 \leq \|\boldsymbol{A}\boldsymbol{x}\|_2^2 \leq (1+\delta)\|\boldsymbol{x}\|_2^2.$$

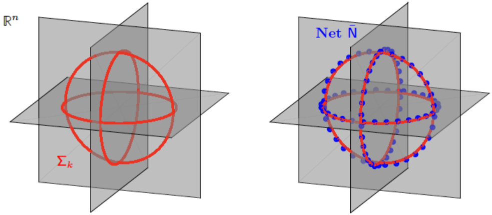
Theorem (RIP $\implies$ ℓ^1 succeeds)

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

Theory: Random Sensing

Theorem (RIP of Gaussian Matrices)

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



$\Longrightarrow$ ℓ^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:

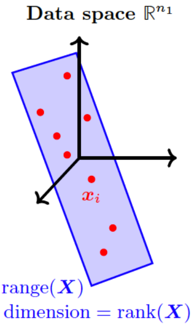
$$\underset{\text{observation}}{\boldsymbol{y}} = \boldsymbol{A} \underset{\text{unknown}}{\boldsymbol{x}_o}$$

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

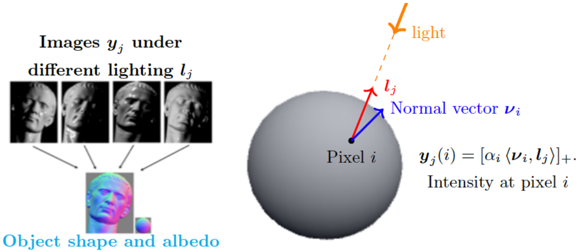
Recovering a low-rank matrix $\boldsymbol{X}_o$:

$$\underset{\text{observation}}{\boldsymbol{y}} = \mathcal{A}\left[\underset{\text{unknown}}{\boldsymbol{X}_o}\right]$$

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



Low-Rank I: **Rank and Geometry**



Multiple images of a Lambertian object with varying light:

$$\boldsymbol{Y} = \mathcal{P}_\Omega[\boldsymbol{N}\boldsymbol{L}], \ \boldsymbol{X} = \boldsymbol{N}\boldsymbol{L} \text{ has rank 3.}$$

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**

Users

$$\begin{bmatrix} 5 & 3 & \dots & ? \\ ? & 2 & \dots & 4 \\ \vdots & \vdots & \ddots & \vdots \\ 5 & ? & \dots & ? \end{bmatrix} = \mathcal{P}_\Omega \left(\underbrace{\begin{bmatrix} 5 & 3 & \dots & 5 \\ 4 & 2 & \dots & 4 \\ \vdots & \vdots & \ddots & \vdots \\ 5 & 5 & \dots & 3 \end{bmatrix}}_{\text{Complete Ratings } \boldsymbol{X}} \right)$$

Items

We observe: Observed (Incomplete) Ratings $\boldsymbol{Y}$

$$\underbrace{\boldsymbol{Y}}_{\text{Observed ratings}} = \mathcal{P}_\Omega \Big[\underbrace{\boldsymbol{X}}_{\text{Complete ratings}} \Big],$$

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 ($\boldsymbol{Y}_{ij} = \boldsymbol{u}_i^T \boldsymbol{v}_j$, with $\boldsymbol{u}_i, \boldsymbol{v}_j \in \mathbb{R}^r$).

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

Rank and Singular Value Decomposition

Theorem (Compact SVD)

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

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

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

Many of the same tools and ideas apply!

Computing SVD: Nice Nonconvex Problem (Lecture 3)

Affine Rank Minimization

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

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

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

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

NP-Hard in general, by reduction from ℓ^0 minimization, using that

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

Let's seek a tractable surrogate...

Convex Relaxation: Nuclear Norm Minimization

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

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

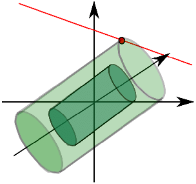
Also known as the *trace norm*, *Schatten* 1*-norm*, and *Ky-Fan* k*-norm*.

Nuclear norm minimization problem:

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

Geometry of nuclear norm minimization:

Nuclear norm ball $\mathrm{B}_* = \{\boldsymbol{X} \mid \|\boldsymbol{X}\|_* \le 1\}$



Low-Rank Recovery with Generic Measurements

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

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

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

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

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

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

$y_i = \Big\langle$  $, \boldsymbol{X}_o \Big\rangle$

$\boldsymbol{A}_i$ random

Matrix Sensing

$y_i = \Big\langle$  $, \boldsymbol{X}_o \Big\rangle$

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

Matrix Completion

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

Rank-RIP: no low-rank $\boldsymbol{X}$ in $\mathrm{null}(\mathcal{A})$.

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

$\Longrightarrow$ **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**

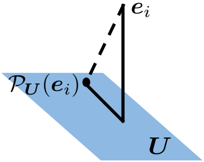
Theorem

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

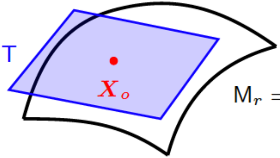
$$m \geq C n r \nu \log^2 n.$$

Restrict to incoherent $\boldsymbol{X}_o$
(not concentrated on a few entries!)

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



$$\nu = \max\{\tfrac{n}{r}\max_i \|\mathcal{P}_{\boldsymbol{U}} e_i\|_2^2, \tfrac{n}{r}\max_j \|\mathcal{P}_{\boldsymbol{V}} e_j\|_2^2\}$$

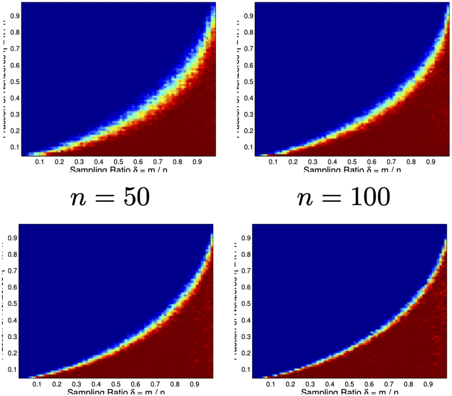


$\mathrm{M}_r = \{\boldsymbol{X} \mid \mathrm{rank}(\boldsymbol{X}) = r\}$

Parallelism between **Rank** and **Sparsity**

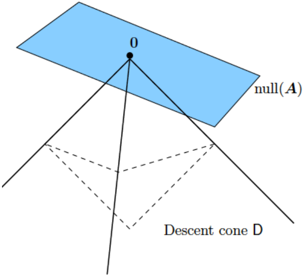
	Sparse Vector	**Low-rank Matrix**
Low-dimensionality of	individual signal $\boldsymbol{x}$	a set of signals $\boldsymbol{X}$
Compressive sensing	$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$	$\boldsymbol{Y} = \mathcal{A}(\boldsymbol{X})$
Low-dim measure	ℓ^0 norm $\Vert\boldsymbol{x}\Vert_0$	rank$(\boldsymbol{X})$
Convex surrogate	ℓ^1 norm $\Vert\boldsymbol{x}\Vert_1$	nuclear norm $\Vert\boldsymbol{X}\Vert_*$
Success conditions (RIP)	$\delta_{2k}(\boldsymbol{A}) \geq \sqrt{2} - 1$	$\delta_{4r}(\boldsymbol{A}) \geq \sqrt{2} - 1$
Random measurements	$m = O\big(k \log(n/k)\big)$	$m = O(nr)$
Stable/Inexact recovery	$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{z}$	$\boldsymbol{Y} = \mathcal{A}(\boldsymbol{X}) + \boldsymbol{Z}$
Phase transition at	Stat. dim. of descent cone: $m^* = \delta(\mathrm{D})$	

Sharp Phase Transitions with Gaussian Measurements



$n = 50$ $\quad$ $n = 100$

$n = 200$ $\quad$ $n = 400$



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

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

Noise and Inexact Structure

Observation: $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_o + \boldsymbol{z}$, with $\boldsymbol{x}_o$ structured, and $\boldsymbol{z}$ noise.

Goal: produce $\widehat{\boldsymbol{x}}$ as close to $\boldsymbol{x}_o$ as possible! Relax:

- **Lasso** for stable sparse recovery

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

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

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

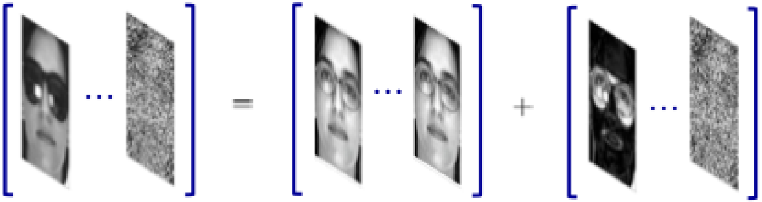
Wealth of statistical results: if $\boldsymbol{A}$ "nice" (say, RIP or RSC) ...

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

Parallelism between **Rank** and **Sparsity**

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

Combining Rank and Sparsity: Robust PCA?



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

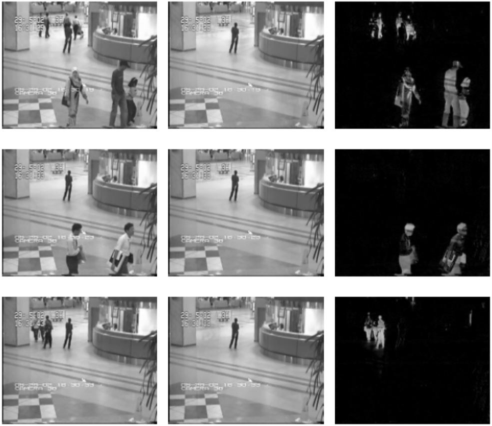
Given $\boldsymbol{Y} = \boldsymbol{L}_o + \boldsymbol{S}_o$, with $\boldsymbol{L}_o$ low-rank, $\boldsymbol{S}_o$ sparse, recover $(\boldsymbol{L}_o, \boldsymbol{S}_o)$.

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**

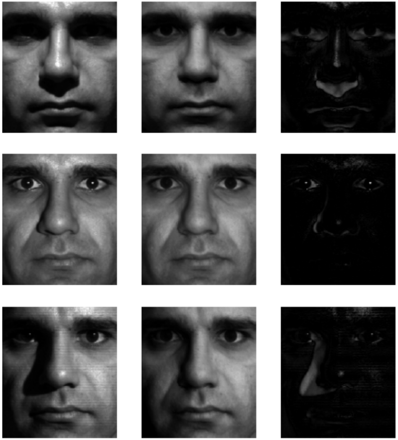
A sequence of video frames can be modeled as a static background (low-rank) and moving foreground (sparse).



(a) Original frames (b) Low-rank $\hat{L}$ (c) Sparse $\hat{S}$

Low-rank + Sparse II: **Faces**

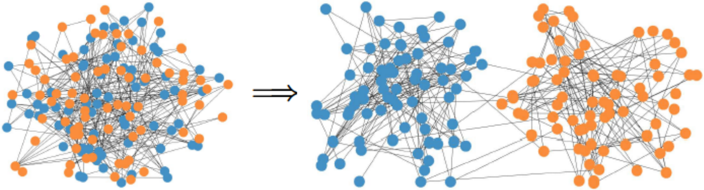
A set of face images of the same person under different lightings can be modeled as a low-dimensional, $3 \sim 9$d, subspace and sparse occlusions and corruptions (specularities).



(a) $\boldsymbol{Y}$ (b) $\hat{\boldsymbol{L}}$ (c) $\hat{\boldsymbol{S}}$

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 $\boldsymbol{M}$. Hence $\boldsymbol{M}$ is a low-rank matrix and some sparse connections across communities.



Low-rank + Sparse: **Convex Relaxations**

Optimization formulation:

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

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

$$\|\boldsymbol{S}\|_0 \to \|\boldsymbol{S}\|_1, \qquad \mathrm{rank}(\boldsymbol{L}) = \|\boldsymbol{\sigma}(\boldsymbol{L})\|_0 \to \|\boldsymbol{L}\|_*$$

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

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

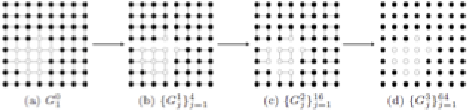
General Low-Dimensional Models

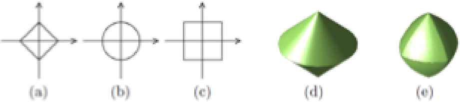
Atomic Norms and Structured Sparsity

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

- **Sparsity**: $\mathcal{D} = \{\boldsymbol{e}_i\}$,
- **Low-rank**: $\mathcal{D} = \{\boldsymbol{u}\boldsymbol{v}^T\}$,
- **Column sparse matrices**: $\mathcal{D} = \{\boldsymbol{u}\boldsymbol{e}_j^T\}$,
- **Sinusoids**: $\mathcal{D} = \{\exp(\mathrm{i}(2\pi f t + \xi))\}$,
- **Tensors**: $\mathcal{D} = \{\boldsymbol{u}_1 \otimes \boldsymbol{u}_2 \otimes \boldsymbol{u}_N\}$, ...

Structured Sparsity: capture relationship between nonzeros

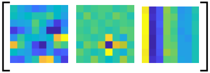




Learned Low-Dimensional Models:

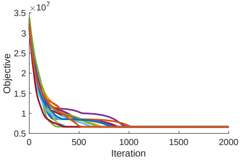
Dictionary Learning, Deconvolution

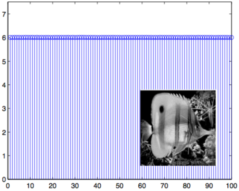
Data $\boldsymbol{Y}$



Dictionary $\boldsymbol{A}$

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





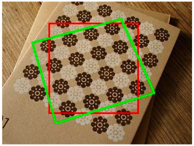
The same **modeling toolkit**, but optimization formulations become **nonconvex**! (see Lecture 4)

Nonlinear Low-Dimensional Models

Nonlinear Observations: Transformed low-rank texture

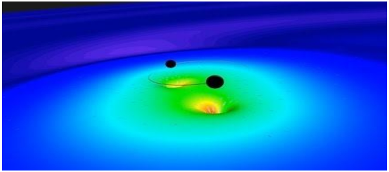
(a) Low-rank texture $\boldsymbol{I}_o$

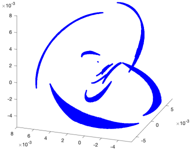
$\xleftarrow{\tau}$



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

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 $m \sim \#\mathrm{dof}(\boldsymbol{x})$.
- **Tractable Computation**: Convex relaxations ℓ^1, nuclear norm
- **Extensions**: Combinations, learned dictionaries, nonlinear structures.

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

Thank You! Questions?