ECE 18-898G: Special Topics in Signal Processing: Sparsity, Structure, and Inference

Low-rank matrix recovery via convex relaxations

Yuejie Chi

Department of Electrical and Computer Engineering

Carnegie Mellon University

Spring 2018

Outline

- Low-rank matrix completion and recovery
- Nuclear norm minimization (this lecture)
  - RIP and low-rank matrix recovery
  - Matrix completion
  - Algorithms for nuclear norm minization
- Non-convex methods (next lecture)
  - Spectral methods
  - (Projected) gradient descent

Low-rank matrix completion and recovery: motivation

Motivation 1: recommendation systems

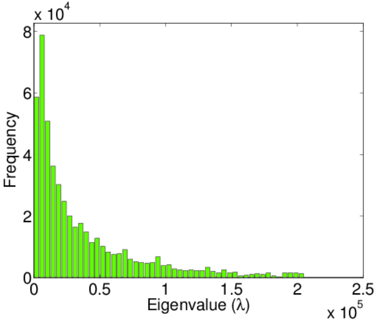
- Netflix challenge: Netflix provides highly incomplete ratings from 0.5 million users for & 17,770 movies
- How to predict unseen user ratings for movies?

In general, we cannot infer missing ratings

$$\begin{bmatrix} \checkmark & ? & ? & ? & \checkmark & ? \\ ? & ? & \checkmark & \checkmark & ? & ? \\ \checkmark & ? & ? & \checkmark & ? & ? \\ ? & ? & \checkmark & ? & ? & \checkmark \\ \checkmark & ? & ? & ? & ? & ? \\ ? & \checkmark & ? & ? & \checkmark & ? \\ ? & ? & \checkmark & \checkmark & ? & ? \end{bmatrix}$$

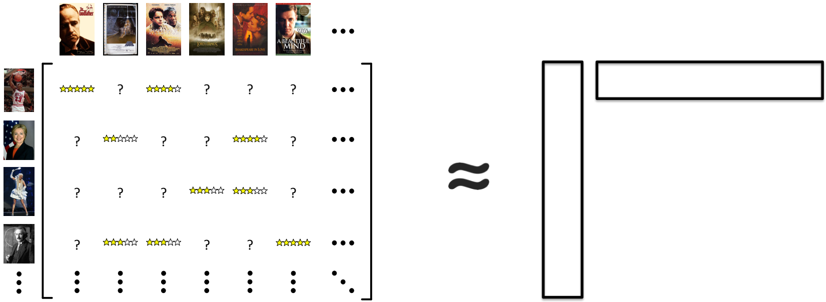
Underdetermined system (more unknowns than observations)

… unless rating matrix has other structure



A few factors explain most of the data

… unless rating matrix has other structure



A few factors explain most of the data $\longrightarrow$ low-rank approximation

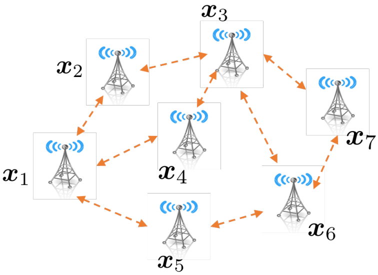
How to exploit (approx.) low-rank structure in prediction?

Motivation 2: sensor localization

- n sensors / points $\boldsymbol{x}_j \in \mathbb{R}^3$, $j = 1, \cdots, n$
- Observe partial information about pairwise distances

$$D_{i,j} = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2 = \|\boldsymbol{x}_i\|^2 + \|\boldsymbol{x}_j\|^2 - 2\boldsymbol{x}_i^\top \boldsymbol{x}_j$$

- Want to infer distance between every pair of nodes



Motivation 2: sensor localization

Introduce

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^\top \\ \boldsymbol{x}_2^\top \\ \vdots \\ \boldsymbol{x}_n^\top \end{bmatrix} \in \mathbb{R}^{n\times 3}$$

then distance matrix $\boldsymbol{D} = [D_{i,j}]_{1\le i,j\le n}$ can be written as

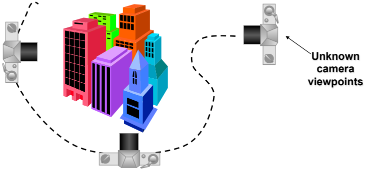
$$\boldsymbol{D} = \underbrace{\boldsymbol{d}_2\boldsymbol{e}^\top + \boldsymbol{e}\boldsymbol{d}_2^\top - 2\boldsymbol{X}\boldsymbol{X}^\top}_{\text{low rank}}$$

where $\boldsymbol{d}_2 := [\,\|\boldsymbol{x}_1\|^2, \cdots, \|\boldsymbol{x}_n\|^2\,]^\top$

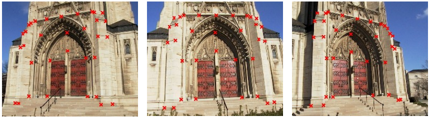
rank$(\boldsymbol{D}) \ll n \quad \longrightarrow \quad$ low-rank matrix completion

Motivation 3: structure from motion

Structure from motion: reconstruct $\underbrace{\text{3D scene geometry}}_{\text{structure}}$ and $\underbrace{\text{camera poses}}_{\text{motion}}$ from multiple images



Given multiple images and a few correspondences between image features, how to estimate locations of 3D points?



Motivation 3: structure from motion

Tomasi and Kanade's factorization:

- Consider n 3D points in m different 2D frames
- $\boldsymbol{x}_{i,j} \in \mathbb{R}^{2\times 1}$: locations of j^{th} point in i^{th} frame

$$\boldsymbol{x}_{i,j} = \underbrace{\boldsymbol{P}_i}_{\text{projection matrix } \in \mathbb{R}^{2\times 3}} \quad \underbrace{\boldsymbol{s}_j}_{\text{3D position } \in \mathbb{R}^3}$$

- Matrix of all 2D locations $\text{rank}(\boldsymbol{X}) = 3$.

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_{1,1} & \cdots & \boldsymbol{x}_{1,n} \\ \vdots & \ddots & \vdots \\ \boldsymbol{x}_{m,1} & \cdots & \boldsymbol{x}_{m,n} \end{bmatrix} = \underbrace{\begin{bmatrix} \boldsymbol{P}_1 \\ \vdots \\ \boldsymbol{P}_m \end{bmatrix} \begin{bmatrix} \boldsymbol{s}_1 & \cdots & \boldsymbol{s}_n \end{bmatrix}}_{\text{low-rank factorization}} \in \mathbb{R}^{2m\times n}$$

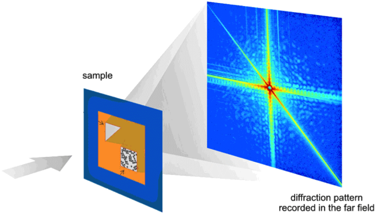
Due to occlusions, there are many missing entries in the matrix $\boldsymbol{X}$.
Goal: Can we complete the missing entries?

Motivation 4: missing phase problem

Detectors record intensities of diffracted rays

- electric field $x(t_1, t_2) \longrightarrow$ Fourier transform $\hat{x}(f_1, f_2)$

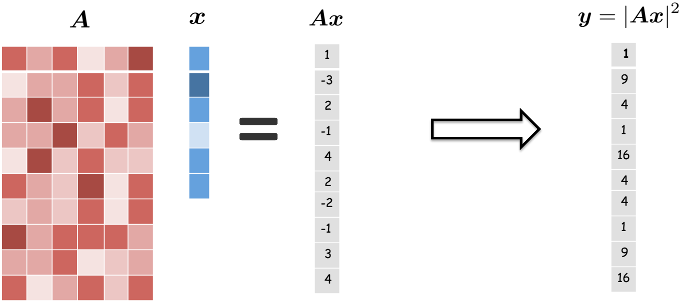
Fig credit: Stanford SLAC



intensity of electrical field: $\left|\hat{x}(f_1, f_2)\right|^2 = \left|\int x(t_1, t_2) e^{-i2\pi(f_1 t_1 + f_2 t_2)} \mathrm{d}t_1 \mathrm{d}t_2\right|^2$

Phase retrieval: recover signal $x(t_1, t_2)$ from intensity $\left|\hat{x}(f_1, f_2)\right|^2$

A discrete-time model: solving quadratic systems



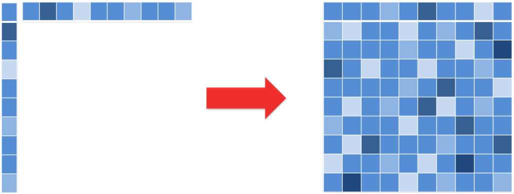
Solve for $\boldsymbol{x} \in \mathbb{C}^n$ in m quadratic equations

$$
\begin{aligned}
y_k &= |\langle \boldsymbol{a}_k, \boldsymbol{x} \rangle|^2, \qquad k = 1, \ldots, m \\
\text{or} \qquad \boldsymbol{y} &= |\boldsymbol{A}\boldsymbol{x}|^2 \qquad \text{where } |\boldsymbol{z}|^2 := \{|z_1|^2, \cdots, |z_m|^2\}
\end{aligned}
$$

An equivalent view: low-rank factorization

Lifting: introduce $\boldsymbol{X} = \boldsymbol{x}\boldsymbol{x}^*$ to linearize constraints

$$y_k = |\boldsymbol{a}_k^* \boldsymbol{x}|^2 = \boldsymbol{a}_k^*(\boldsymbol{x}\boldsymbol{x}^*)\boldsymbol{a}_k \quad \Longrightarrow \quad y_k = \boldsymbol{a}_k^* \boldsymbol{X} \boldsymbol{a}_k \tag{6.1}$$



$$\begin{aligned} \text{find} \quad & \boldsymbol{X} \succeq \boldsymbol{0} \\ \text{s.t.} \quad & y_k = \langle \boldsymbol{a}_k \boldsymbol{a}_k^*, \boldsymbol{X} \rangle, \qquad k = 1, \cdots, m \\ & \text{rank}(\boldsymbol{X}) = 1 \end{aligned}$$

The list continues

- system identification and time series analysis;
- spatial-temporal data: low-rank due to correlations, e.g. MRI video, network traffic, ..
- face recognition;
- quantum state tomography;
- community detection;
- ....

Low-rank matrix completion and recovery: setup and algorithms

Setup

- Consider $\boldsymbol{M} \in \mathbb{R}^{n \times n}$ (square case for simplicity)
- rank$(\boldsymbol{M}) = r \ll n$
- The thin Singular value decomposition (SVD) of $\boldsymbol{M}$:

$$\boldsymbol{M} = \underbrace{\boldsymbol{U\Sigma V}^\top}_{(2n-r)r \text{ degrees of freedom}} = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T$$

where $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{bmatrix}$ contain all singular values $\{\sigma_i\}$;
$\boldsymbol{U} := [\boldsymbol{u}_1, \cdots, \boldsymbol{u}_r]$, $\boldsymbol{V} := [\boldsymbol{v}_1, \cdots, \boldsymbol{v}_r]$ consist of singular vectors

Low-rank matrix completion

Observed entries

$$M_{i,j}, \qquad (i,j) \in \underbrace{\Omega}_{\text{sampling set}}$$

Completion via rank minimization

$$\text{minimize}_{\boldsymbol{X}} \quad \text{rank}(\boldsymbol{X}) \qquad \text{s.t.} \quad X_{i,j} = M_{i,j}, \quad (i,j) \in \Omega$$

Low-rank matrix completion

Observed entries

$$M_{i,j}, \qquad (i,j) \in \underbrace{\Omega}_{\text{sampling set}}$$

- An operator $\mathcal{P}_\Omega$: orthogonal projection onto subspace of matrices supported on Ω

$$[\mathcal{P}_\Omega(\boldsymbol{X})]_{i,j} = \begin{cases} X_{i,j} & \text{if} \quad (i,j) \in \Omega \\ 0 & \text{otherwise} \end{cases}$$

Completion via rank minimization

$$\text{minimize}_{\boldsymbol{X}} \quad \text{rank}(\boldsymbol{X}) \qquad \text{s.t.} \quad \mathcal{P}_\Omega(\boldsymbol{X}) = \mathcal{P}_\Omega(\boldsymbol{M})$$

More general: low-rank matrix recovery

Linear measurements

$$y_i = \langle \boldsymbol{A}_i, \boldsymbol{M} \rangle = \mathsf{Tr}(\boldsymbol{A}_i^\top \boldsymbol{M}), \qquad i = 1, \ldots m$$

- An operator form, with $\mathcal{A} : \mathbb{R}^{n \times n} \mapsto \mathbb{R}^m$:

$$\boldsymbol{y} = \mathcal{A}(\boldsymbol{M}) := \begin{bmatrix} \langle \boldsymbol{A}_1, \boldsymbol{M} \rangle \\ \vdots \\ \langle \boldsymbol{A}_m, \boldsymbol{M} \rangle \end{bmatrix}$$

Recovery via rank minimization

$$\text{minimize}_{\boldsymbol{X}} \quad \text{rank}(\boldsymbol{X}) \qquad \text{s.t.} \quad \boldsymbol{y} = \mathcal{A}(\boldsymbol{X})$$

Nuclear norm minimization

Convex relaxation

Low-rank matrix completion:

$$\begin{aligned} \text{minimize}_{\boldsymbol{X}} \quad & \underbrace{\text{rank}(\boldsymbol{X})}_{\text{nonconvex}} \\ \text{s.t.} \quad & \mathcal{P}_{\Omega}(\boldsymbol{X}) = \mathcal{P}_{\Omega}(\boldsymbol{M}) \end{aligned}$$

Low-rank matrix recovery:

$$\begin{aligned} \text{minimize}_{\boldsymbol{X}} \quad & \underbrace{\text{rank}(\boldsymbol{X})}_{\text{nonconvex}} \\ \text{s.t.} \quad & \mathcal{A}(\boldsymbol{X}) = \mathcal{A}(\boldsymbol{M}) \end{aligned}$$

Question: what is convex surrogate for $\text{rank}(\cdot)$?

Analogy with sparse recovery

For a given matrix $\boldsymbol{X} \in \mathbb{R}^{n \times n}$, take the full SVD of $\boldsymbol{X}$:

$$\begin{aligned}
\boldsymbol{X} &= \widehat{\boldsymbol{U}}\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{V}}^\top \\
&= \Big[\boldsymbol{u}_1, \cdots, \boldsymbol{u}_n\Big] \underbrace{\begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \end{bmatrix}}_{\mathrm{diag}(\boldsymbol{\sigma})} \Big[\boldsymbol{v}_1, \cdots, \boldsymbol{v}_n\Big]^\top .
\end{aligned}$$

Then

$$\mathsf{rank}(\boldsymbol{X}) = \sum_{i=1}^{n} \mathbf{1}\{\sigma_i \neq 0\} = \|\boldsymbol{\sigma}\|_0$$

Convex relaxation: from $\|\boldsymbol{\sigma}\|_0$ to $\|\boldsymbol{\sigma}\|_1$?

Nuclear norm

Definition 6.1

The nuclear norm of $\boldsymbol{X}$ is

$$\|\boldsymbol{X}\|_* := \sum_{i=1}^{n} \underbrace{\sigma_i(\boldsymbol{X})}_{i^{\text{th}} \text{ largest singular value}}$$

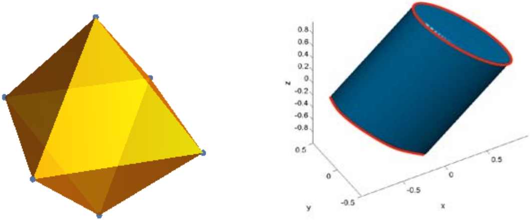
- Nuclear norm is a counterpart of ℓ_1 norm for rank function
- Equivalence among different norms ($r = \text{rank}(\boldsymbol{X})$)

$$\|\boldsymbol{X}\| \leq \|\boldsymbol{X}\|_{\mathsf{F}} \leq \|\boldsymbol{X}\|_* \leq \sqrt{r}\|\boldsymbol{X}\|_{\mathsf{F}} \leq r\|\boldsymbol{X}\|.$$

where $\|\boldsymbol{X}\| = \sigma_1(\boldsymbol{X})$; $\|\boldsymbol{X}\|_{\mathsf{F}} = \left(\sum_{i=1}^{n} \sigma_i^2(\boldsymbol{X})\right)^{1/2}$.

Tightness of relaxation

Recall: the ℓ_1 norm ball is convex hull of 1-sparse, unit-norm vectors.



(a) ℓ_1 norm ball (b) nuclear norm ball

Fact 6.2

The nuclear norm ball $\{\boldsymbol{X} : \|\boldsymbol{X}\|_* \leq 1\}$ is the convex hull of rank-1 matrices, unit-norm matrices obeying $\|\boldsymbol{u}\boldsymbol{v}^\top\| = 1$.

Additivity of nuclear norm

Fact 6.3

Let $\boldsymbol{A}$ *and* $\boldsymbol{B}$ *be matrices of the same dimensions. If* $\boldsymbol{A}\boldsymbol{B}^\top = \boldsymbol{0}$ *and* $\boldsymbol{A}^\top \boldsymbol{B} = \boldsymbol{0}$, *then* $\|\boldsymbol{A}+\boldsymbol{B}\|_* = \|\boldsymbol{A}\|_* + \|\boldsymbol{B}\|_*$.

- If row (resp. column) spaces of $\boldsymbol{A}$ and $\boldsymbol{B}$ are orthogonal, then $\|\boldsymbol{A}+\boldsymbol{B}\|_* = \|\boldsymbol{A}\|_* + \|\boldsymbol{B}\|_*$

- Similar to ℓ_1 norm: when $\boldsymbol{x}$ and $\boldsymbol{y}$ have disjoint support,

$$\|\boldsymbol{x}+\boldsymbol{y}\|_1 = \|\boldsymbol{x}\|_1 + \|\boldsymbol{y}\|_1$$

  which is a key to study ℓ_1-min under RIP.

Proof of Fact 6.3

Suppose $\boldsymbol{A} = \boldsymbol{U}_A \boldsymbol{\Sigma}_A \boldsymbol{V}_A^\top$ and $\boldsymbol{B} = \boldsymbol{U}_B \boldsymbol{\Sigma}_B \boldsymbol{V}_B^\top$, which gives

$$\begin{array}{ll} \boldsymbol{A}\boldsymbol{B}^\top & = \boldsymbol{0} \\ \boldsymbol{A}^\top \boldsymbol{B} & = \boldsymbol{0} \end{array} \quad \Longleftrightarrow \quad \begin{array}{ll} \boldsymbol{V}_A^\top \boldsymbol{V}_B & = \boldsymbol{0} \\ \boldsymbol{U}_A^\top \boldsymbol{U}_B & = \boldsymbol{0} \end{array}$$

Thus, one can write

$$\begin{aligned} \boldsymbol{A} &= [\boldsymbol{U}_A, \boldsymbol{U}_B, \boldsymbol{U}_C] \begin{bmatrix} \boldsymbol{\Sigma}_A & & \\ & \boldsymbol{0} & \\ & & \boldsymbol{0} \end{bmatrix} [\boldsymbol{V}_A, \boldsymbol{V}_B, \boldsymbol{V}_C]^\top \\ \boldsymbol{B} &= [\boldsymbol{U}_A, \boldsymbol{U}_B, \boldsymbol{U}_C] \begin{bmatrix} \boldsymbol{0} & & \\ & \boldsymbol{\Sigma}_B & \\ & & \boldsymbol{0} \end{bmatrix} [\boldsymbol{V}_A, \boldsymbol{V}_B, \boldsymbol{V}_C]^\top \end{aligned}$$

and hence

$$\|\boldsymbol{A} + \boldsymbol{B}\|_* = \left\| [\boldsymbol{U}_A, \boldsymbol{U}_B] \begin{bmatrix} \boldsymbol{\Sigma}_A & \\ & \boldsymbol{\Sigma}_B \end{bmatrix} [\boldsymbol{V}_A, \boldsymbol{V}_B]^\top \right\|_* = \|\boldsymbol{A}\|_* + \|\boldsymbol{B}\|_*$$

Dual norm

Definition 6.4 (Dual norm)

For a given norm $\|\cdot\|_{\mathcal{A}}$, the dual norm is defined as

$$\|\boldsymbol{X}\|_{\mathcal{A}}^{\star} := \max\{\langle \boldsymbol{X}, \boldsymbol{Y}\rangle : \|\boldsymbol{Y}\|_{\mathcal{A}} \leq 1\}$$

- ℓ_1 norm $\quad \stackrel{\text{dual}}{\longleftrightarrow} \quad$ ℓ_∞ norm
- ℓ_2 norm $\quad \stackrel{\text{dual}}{\longleftrightarrow} \quad$ ℓ_2 norm
- Frobenius norm $\quad \stackrel{\text{dual}}{\longleftrightarrow} \quad$ Frobenius norm
- nuclear norm $\quad \stackrel{\text{dual}}{\longleftrightarrow} \quad$ spectral norm

Schur complement

Given a block matrix

$$\boldsymbol{D} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{B}^\top & \boldsymbol{C} \end{bmatrix} \in \mathbb{R}^{(p+q)\times(p+q)},$$

where $\boldsymbol{A} \in \mathbb{R}^{p\times p}$, $\boldsymbol{B} \in \mathbb{R}^{p\times q}$ and $\boldsymbol{C} \in \mathbb{R}^{q\times q}$.

The **Schur complement** of the block $\boldsymbol{A}$ (assume it's invertible) in $\boldsymbol{D}$ is given as

$$\boldsymbol{C} - \boldsymbol{B}^\top \boldsymbol{A}^{-1} \boldsymbol{B}.$$

Fact 6.5

$$\boldsymbol{D} \succeq \boldsymbol{0} \quad \Longleftrightarrow \quad \boldsymbol{A} \succeq \boldsymbol{0},\ \boldsymbol{C} - \boldsymbol{B}^\top \boldsymbol{A}^{-1} \boldsymbol{B} \succeq \boldsymbol{0}$$

Representing nuclear norm via SDP

Since spectral norm is dual norm of nuclear norm,

$$\|\boldsymbol{X}\|_* = \max\{\langle \boldsymbol{X}, \boldsymbol{Y}\rangle : \|\boldsymbol{Y}\| \le 1\}$$

The constraint is equivalent to

$$\|\boldsymbol{Y}\| \le 1 \quad \Longleftrightarrow \quad \boldsymbol{Y}\boldsymbol{Y}^\top \preceq \boldsymbol{I} \quad \overset{\text{Schur complement}}{\Longleftrightarrow} \quad \begin{bmatrix} \boldsymbol{I} & \boldsymbol{Y} \\ \boldsymbol{Y}^\top & \boldsymbol{I} \end{bmatrix} \succeq \boldsymbol{0}$$

Fact 6.6

$$\|\boldsymbol{X}\|_* = \max_{\boldsymbol{Y}} \left\{ \langle \boldsymbol{X}, \boldsymbol{Y}\rangle \;\middle|\; \begin{bmatrix} \boldsymbol{I} & \boldsymbol{Y} \\ \boldsymbol{Y}^\top & \boldsymbol{I} \end{bmatrix} \succeq \boldsymbol{0} \right\}$$

Representing nuclear norm via SDP

Since spectral norm is dual norm of nuclear norm,

$$\|\boldsymbol{X}\|_* = \max\{\langle \boldsymbol{X}, \boldsymbol{Y}\rangle : \|\boldsymbol{Y}\| \leq 1\}$$

The constraint is equivalent to

$$\|\boldsymbol{Y}\| \leq 1 \quad \Longleftrightarrow \quad \boldsymbol{Y}\boldsymbol{Y}^\top \preceq \boldsymbol{I} \quad \overset{\text{Schur complement}}{\Longleftrightarrow} \quad \begin{bmatrix} \boldsymbol{I} & \boldsymbol{Y} \\ \boldsymbol{Y}^\top & \boldsymbol{I} \end{bmatrix} \succeq \boldsymbol{0}$$

Fact 6.7 (Dual characterization)

$$\|\boldsymbol{X}\|_* = \min_{\boldsymbol{W}_1, \boldsymbol{W}_2} \left\{ \frac{1}{2}\mathrm{Tr}(\boldsymbol{W}_1) + \frac{1}{2}\mathrm{Tr}(\boldsymbol{W}_2) \;\middle|\; \begin{bmatrix} \boldsymbol{W}_1 & \boldsymbol{X} \\ \boldsymbol{X}^\top & \boldsymbol{W}_2 \end{bmatrix} \succeq \boldsymbol{0} \right\}$$

- Optimal point: $\boldsymbol{W}_1 = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^\top$, $\boldsymbol{W}_2 = \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ (where $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$)

Aside: dual of semidefinite program

$$\begin{array}{llll} \text{(primal)} & \text{minimize}_{\boldsymbol{X}} & \langle \boldsymbol{C}, \boldsymbol{X} \rangle & \\ & \text{s.t.} & \langle \boldsymbol{A}_i, \boldsymbol{X} \rangle = b_i, & 1 \le i \le m \\ & & \boldsymbol{X} \succeq \boldsymbol{0} & \end{array}$$

$$\Updownarrow$$

$$\begin{array}{lll} \text{(dual)} & \text{maximize}_{\boldsymbol{y}} & \boldsymbol{b}^\top \boldsymbol{y} \\ & \text{s.t.} & \sum_{i=1}^{m} y_i \boldsymbol{A}_i + \boldsymbol{S} = \boldsymbol{C} \\ & & \boldsymbol{S} \succeq \boldsymbol{0} \end{array}$$

Exercise: use this to verify Fact 6.7

Nuclear norm minimization via SDP

Nuclear norm minization

$$\hat{\boldsymbol{M}} = \mathrm{argmin}_{\boldsymbol{X}} \|\boldsymbol{X}\|_* \quad \text{s.t.} \quad \boldsymbol{y} = \mathcal{A}(\boldsymbol{X})$$

This is solvable via SDP

$$\begin{aligned} \text{minimize}_{\boldsymbol{X},\boldsymbol{W}_1,\boldsymbol{W}_2} \quad & \frac{1}{2}\mathrm{Tr}(\boldsymbol{W}_1) + \frac{1}{2}\mathrm{Tr}(\boldsymbol{W}_2) \\ \text{s.t.} \quad & \boldsymbol{y} = \mathcal{A}(\boldsymbol{X}), \\ & \begin{bmatrix} \boldsymbol{W}_1 & \boldsymbol{X} \\ \boldsymbol{X}^\top & \boldsymbol{W}_2 \end{bmatrix} \succeq \boldsymbol{0} \end{aligned}$$

Rank minimization vs Cardinality minimization

parsimony concept	cardinality	rank
Hilbert Space norm	Euclidean	Frobenius
sparsity inducing norm	ℓ_1	nuclear
dual norm	ℓ_∞	operator
norm additivity	disjoint support	orthogonal row and column spaces
convex optimization	linear programming	semidefinite programming

Table 1: A dictionary relating the concepts of cardinality and rank minimization.

Fig. credit: Fazel et.al. '10

Proximal algorithm

In the presence of noise, one needs to solve

$$\text{minimize}_{\boldsymbol{X}} \quad \frac{1}{2}\|\boldsymbol{y} - \mathcal{A}(\boldsymbol{X})\|_{\mathrm{F}}^2 + \lambda\|\boldsymbol{X}\|_*$$

which can be solved via proximal methods.

Algorithm 6.1 Proximal gradient methods

for $t = 0, 1, \cdots$**:**

$$\boldsymbol{X}^{t+1} = \text{prox}_{\mu_t \lambda \|\cdot\|_*}\Big(\boldsymbol{X}^t - \mu_t \mathcal{A}^*(\mathcal{A}(\boldsymbol{X}^t) - \boldsymbol{y})\Big)$$

where μ_t: step size / learning rate

Proximal operator for nuclear norm

Proximal operator:

$$\begin{aligned}\mathsf{prox}_{\lambda\|\cdot\|_*}(\boldsymbol{X}) &= \arg\min_{\boldsymbol{Z}} \left\{ \frac{1}{2}\|\boldsymbol{Z}-\boldsymbol{X}\|_{\mathsf{F}}^2 + \lambda\|\boldsymbol{Z}\|_* \right\} \\ &= \boldsymbol{U}\mathcal{T}_\lambda(\boldsymbol{\Sigma})\boldsymbol{V}^\top\end{aligned}$$

where SVD of $\boldsymbol{X}$ is $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ with $\boldsymbol{\Sigma} = \mathsf{diag}(\{\sigma_i\})$, and

$$\mathcal{T}_\lambda(\boldsymbol{\Sigma}) = \mathsf{diag}(\{(\sigma_i - \lambda)_+\})$$

Accelerated proximal gradient

Algorithm 6.2 Accelerated proximal gradient methods

for $t = 0, 1, \cdots$**:**

$$\boldsymbol{X}^{t+1} = \mathsf{prox}_{\mu_t \lambda \|\cdot\|_*} \left(\boldsymbol{Z}^t - \mu_t \mathcal{A}^* (\mathcal{A}(\boldsymbol{Z}^t) - \boldsymbol{y}) \right)$$

$$\boldsymbol{Z}^{t+1} = \boldsymbol{X}^{t+1} + \underbrace{\alpha_t \left(\boldsymbol{X}^{t+1} - \boldsymbol{X}^t \right)}_{\text{momentum term}}$$

where μ_t: step size / learning rate, α_t is the momentum.

- Convergence rate: $O\left(\frac{1}{\sqrt{\epsilon}}\right)$ iterations to reach ϵ-accuracy.
- Per-iteration cost is a (partial-)SVD.

Frank-Wolfe for nuclear norm minimization

Consider the constrained problem:

$$\text{minimize}_{\boldsymbol{X}} \quad \frac{1}{2}\|\boldsymbol{y} - \mathcal{A}(\boldsymbol{X})\|_{\mathrm{F}}^2 \quad \text{s.t.} \quad \|\boldsymbol{X}\|_* \leq \tau.$$

which can be solved via conditional gradient method (Frank-Wolfe 1956).

- More generally, consider the problem:

$$\text{minimize}_{\boldsymbol{\beta}} \quad f(\boldsymbol{\beta}) \quad \text{s.t.} \quad \boldsymbol{\beta} \in \mathcal{C},$$

  where $f(\boldsymbol{x})$ is smooth and convex, and $\mathcal{C}$ is a convex set.
- Recall projected gradient descent:

$$\boldsymbol{\beta}^{t+1} = \mathcal{P}_{\mathcal{C}}\left(\boldsymbol{\beta}^t - \mu_t \nabla f(\boldsymbol{\beta}^t)\right)$$

Conditional Gradient Method

Algorithm 6.3 Frank-Wolfe

for $t = 0, 1, \cdots$**:**

$$\boldsymbol{s}^t = \text{argmin}_{\boldsymbol{s} \in \mathcal{C}} \nabla f(\boldsymbol{\beta}^t)^\top \boldsymbol{s}$$

$$\boldsymbol{\beta}^{t+1} = (1 - \gamma_t)\boldsymbol{\beta}^t + \gamma_t \boldsymbol{s}^t$$

where $\gamma_t := 2/(t+1)$ is the (default) step size.

- The first step is a constrained optimization of a linear approximation at $f(\boldsymbol{\beta}^t)$;
- The second step controls how much we move towards $\boldsymbol{s}^t$.

Figure illustration

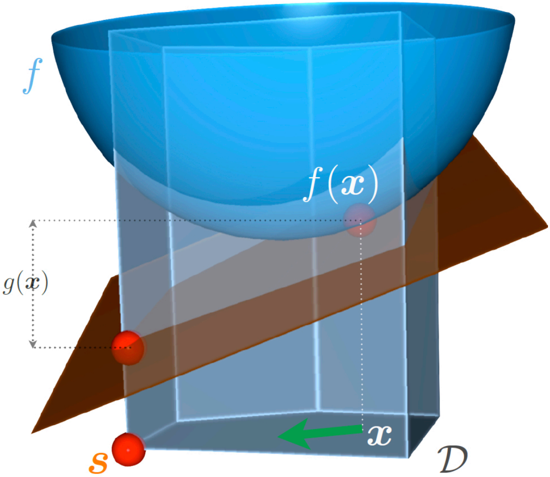


Figure credit: Jaggi 2011

Frank-Wolfe for nuclear norm minimization

Algorithm 6.4 Frank-Wolfe for nuclear norm minimization

for $t = 0, 1, \cdots$**:**

$$\boldsymbol{S}^t = \operatorname{argmin}_{\|\boldsymbol{S}\|_* \leq \tau} \langle \nabla f(\boldsymbol{X}^t), \boldsymbol{S} \rangle,$$

$$\boldsymbol{X}^{t+1} = (1 - \gamma_t)\boldsymbol{X}^t + \gamma_t \boldsymbol{S}^t;$$

where $\gamma_t := 2/(t+1)$ is the (default) step size.

- (Homework) Note that $\nabla f(\boldsymbol{X}^t) = \mathcal{A}^*(\mathcal{A}(\boldsymbol{X}^t) - \boldsymbol{y})$, and

$$\begin{aligned}\boldsymbol{S}^t &= \tau \cdot \operatorname{argmin}_{\|\boldsymbol{S}\|_* \leq 1} \langle \nabla f(\boldsymbol{X}^t), \boldsymbol{S} \rangle \\ &= \tau \boldsymbol{u}\boldsymbol{v}^T,\end{aligned}$$

  where $\boldsymbol{u}$, $\boldsymbol{v}$ are the left and right top singular vector of $-\nabla f(\boldsymbol{X}^t)$.

Further comments on Frank-Wolfe

- Extremely low per-iteration cost (only top singular vectors are needed);
- Every iteration is a rank-1 update;
- Convergence rate: $O(\frac{1}{\epsilon})$ to reach ϵ-accuracy, which can be very slow.
- Various ways to speed up; active research area.

RIP and low-rank matrix recovery

RIP for low-rank matrices

Almost parallel results to compressed sensing ...[1]

Definition 6.8

The r-restricted isometry constants $\delta_r^{\mathrm{ub}}(\mathcal{A})$ and $\delta_r^{\mathrm{lb}}(\mathcal{A})$ are smallest quantities s.t.

$$(1-\delta_r^{\mathrm{lb}})\|\boldsymbol{X}\|_{\mathsf{F}} \le \|\mathcal{A}(\boldsymbol{X})\|_{\mathsf{F}} \le (1+\delta_r^{\mathrm{ub}})\|\boldsymbol{X}\|_{\mathsf{F}}, \qquad \forall \boldsymbol{X} : \mathrm{rank}(\boldsymbol{X}) \le r$$

[1] One can also define RIP w.r.t. $\|\cdot\|_{\mathrm{F}}^2$ rather than $\|\cdot\|_{\mathrm{F}}$.

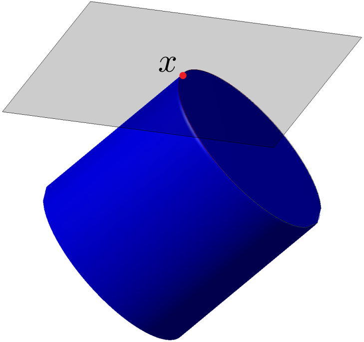
RIP and low-rank matrix recovery

Theorem 6.9 (Recht, Fazel, Parrilo '10, Candes, Plan '11)

Suppose $\mathrm{rank}(\boldsymbol{M}) = r$*. For any fixed integer* $K > 0$*, if* $\frac{1+\delta_{Kr}^{\mathrm{ub}}}{1-\delta_{(2+K)r}^{\mathrm{lb}}} < \sqrt{\frac{K}{2}}$*, then nuclear norm minimization is exact*

- Can be easily extended to account for noise and imperfect structural assumption

Geometry of nuclear norm ball



Level set of nuclear norm ball: $\left\| \begin{bmatrix} x & y \\ y & z \end{bmatrix} \right\|_* \leq 1$

Fig. credit: Candes '14

Some notation

Recall $\boldsymbol{M} = \boldsymbol{U\Sigma V}^\top$

- Let T be span of matrices of the form (called *tangent space*)

$$T = \{\boldsymbol{U}\boldsymbol{X}^\top + \boldsymbol{Y}\boldsymbol{V}^\top : \boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{n\times r}\}$$

- Let $\mathcal{P}_T$ be orthogonal projection onto T:

$$\mathcal{P}_T(\boldsymbol{X}) = \boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{X} + \boldsymbol{X}\boldsymbol{V}\boldsymbol{V}^\top - \boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{X}\boldsymbol{V}\boldsymbol{V}^\top$$

- Its complement $\mathcal{P}_{T^\perp} = \mathcal{I} - \mathcal{P}_T$:

$$\mathcal{P}_{T^\perp}(\boldsymbol{X}) = (\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^\top)\boldsymbol{X}(\boldsymbol{I} - \boldsymbol{V}\boldsymbol{V}^\top)$$

  - $\boldsymbol{M}\mathcal{P}_{T^\perp}^\top(\boldsymbol{X}) = \boldsymbol{0}$ and $\boldsymbol{M}^\top\mathcal{P}_{T^\perp}(\boldsymbol{X}) = \boldsymbol{0}$

Proof of Theorem 6.9

Suppose $\boldsymbol{X} = \boldsymbol{M} + \boldsymbol{H}$ is feasible and obeys $\|\boldsymbol{M} + \boldsymbol{H}\|_* \leq \|\boldsymbol{M}\|_*$. The goal is to show that $\boldsymbol{H} = \boldsymbol{0}$ under RIP.

The key is to decompose $\boldsymbol{H}$ into $\boldsymbol{H}_0 + \underbrace{\boldsymbol{H}_1 + \boldsymbol{H}_2 + \ldots}_{\boldsymbol{H}_{\mathrm{c}}}$

- $\boldsymbol{H}_0 = \mathcal{P}_T(\boldsymbol{H})$ (rank $2r$)
- $\boldsymbol{H}_{\mathrm{c}} = \mathcal{P}_T^{\perp}(\boldsymbol{H})$ (obeying $\boldsymbol{M}\boldsymbol{H}_{\mathrm{c}}^{\top} = \boldsymbol{0}$ and $\boldsymbol{M}^{\top}\boldsymbol{H}_{\mathrm{c}} = \boldsymbol{0}$)
- $\boldsymbol{H}_1$: best rank-(Kr) approximation of $\boldsymbol{H}_{\mathrm{c}}$ (K is const)
- $\boldsymbol{H}_2$: best rank-(Kr) approximation of $\boldsymbol{H}_{\mathrm{c}} - \boldsymbol{H}_1$
- ...

Proof of Theorem 6.9

Informally, the proof proceeds by showing that

1. $\boldsymbol{H}_0$ dominates $\sum_{i\geq 2}\boldsymbol{H}_i$ (by objective function)
2. (converse) $\sum_{i\geq 2}\boldsymbol{H}_i$ dominates $\boldsymbol{H}_0+\boldsymbol{H}_1$ (by RIP + feasibility)

These can happen simultaneously only when $\boldsymbol{H}=\boldsymbol{0}$

Proof of Theorem 6.9

Step 1 (which does not rely on RIP). Show that

$$\sum_{j\geq 2}\|\boldsymbol{H}_j\|_{\mathrm{F}} \leq \|\boldsymbol{H}_0\|_*/\sqrt{Kr}. \tag{6.2}$$

This follows immediately by combining the following 2 observations:

(i) Since $\boldsymbol{M}+\boldsymbol{H}$ is assumed to be a better estimate:

$$\begin{aligned}\|\boldsymbol{M}\|_* \geq \|\boldsymbol{M}+\boldsymbol{H}\|_* &\geq \|\boldsymbol{M}+\boldsymbol{H}_{\mathrm{c}}\|_* - \|\boldsymbol{H}_0\|_* \qquad (6.3)\\ &= \underbrace{\|\boldsymbol{M}\|_* + \|\boldsymbol{H}_{\mathrm{c}}\|_*}_{\text{Fact 6.3 } (\boldsymbol{M}\boldsymbol{H}_{\mathrm{c}}^\top=\boldsymbol{0} \text{ and } \boldsymbol{M}^\top\boldsymbol{H}_{\mathrm{c}}=\boldsymbol{0})} - \|\boldsymbol{H}_0\|_*\end{aligned}$$

$$\Longrightarrow \quad \|\boldsymbol{H}_{\mathrm{c}}\|_* \leq \|\boldsymbol{H}_0\|_* \tag{6.4}$$

(ii) Since nonzero singular values of $\boldsymbol{H}_{j-1}$ dominate those of $\boldsymbol{H}_j$ $(j\geq 2)$:

$$\|\boldsymbol{H}_j\|_{\mathrm{F}} \leq \sqrt{Kr}\|\boldsymbol{H}_j\| \leq \sqrt{Kr}\left[\|\boldsymbol{H}_{j-1}\|_*/(Kr)\right] \leq \|\boldsymbol{H}_{j-1}\|_*/\sqrt{Kr}$$

$$\Longrightarrow \quad \sum_{j\geq 2}\|\boldsymbol{H}_j\|_{\mathrm{F}} \leq \frac{1}{\sqrt{Kr}}\sum_{j\geq 2}\|\boldsymbol{H}_{j-1}\|_* = \frac{1}{\sqrt{Kr}}\|\boldsymbol{H}_{\mathrm{c}}\|_* \tag{6.5}$$

Proof of Theorem 6.9

Step 2 (using feasibility + RIP). Show that $\exists \rho < \sqrt{K/2}$ s.t.

$$\|\boldsymbol{H}_0 + \boldsymbol{H}_1\|_{\mathrm{F}} \le \rho \sum\nolimits_{j \ge 2} \|\boldsymbol{H}_j\|_{\mathrm{F}} \tag{6.6}$$

If this claim holds, then

$$\begin{aligned} \|\boldsymbol{H}_0 + \boldsymbol{H}_1\|_{\mathrm{F}} &\le \rho \sum\nolimits_{j \ge 2} \|\boldsymbol{H}_j\|_{\mathrm{F}} \overset{(6.2)}{\le} \rho \frac{1}{\sqrt{Kr}} \|\boldsymbol{H}_0\|_* \\ &\le \rho \frac{1}{\sqrt{Kr}} \left(\sqrt{2r} \|\boldsymbol{H}_0\|_{\mathrm{F}} \right) = \rho \sqrt{\frac{2}{K}} \|\boldsymbol{H}_0\|_{\mathrm{F}} \\ &\le \rho \sqrt{\frac{2}{K}} \|\boldsymbol{H}_0 + \boldsymbol{H}_1\|_{\mathrm{F}}. \end{aligned} \tag{6.7}$$

This cannot hold with $\rho < \sqrt{K/2}$ unless $\underbrace{\boldsymbol{H}_0 + \boldsymbol{H}_1 = \boldsymbol{0}}_{\text{equivalently, } \boldsymbol{H}_0 = \boldsymbol{H}_1 = \boldsymbol{0}}$

Proof of Theorem 6.9

We now prove (6.6). To connect $\boldsymbol{H}_0 + \boldsymbol{H}_1$ with $\sum_{j\geq 2} \boldsymbol{H}_j$, we use feasibility:

$$\mathcal{A}(\boldsymbol{H}) = \boldsymbol{0} \quad \Longleftrightarrow \quad \mathcal{A}(\boldsymbol{H}_0 + \boldsymbol{H}_1) = -\sum_{j\geq 2} \mathcal{A}(\boldsymbol{H}_j),$$

which taken collectively with RIP yields

$$\begin{aligned}(1-\delta^{\mathrm{lb}}_{(2+K)r})\|\boldsymbol{H}_0 + \boldsymbol{H}_1\|_{\mathrm{F}} &\leq \left\|\mathcal{A}(\boldsymbol{H}_0 + \boldsymbol{H}_1)\right\|_{\mathrm{F}} = \left\|\sum_{j\geq 2} \mathcal{A}(\boldsymbol{H}_j)\right\|_{\mathrm{F}} \\ &\leq \sum_{j\geq 2} \left\|\mathcal{A}(\boldsymbol{H}_j)\right\|_{\mathrm{F}} \\ &\leq \sum_{j\geq 2} (1+\delta^{\mathrm{ub}}_{Kr})\|\boldsymbol{H}_j\|_{\mathrm{F}}\end{aligned}$$

This establishes (6.6) as long as $\rho := \frac{1+\delta^{\mathrm{ub}}_{Kr}}{1-\delta^{\mathrm{lb}}_{(2+K)r}} < \sqrt{\frac{K}{2}}$.

Gaussian sampling operators satisfy RIP

If entries of $\{\boldsymbol{A}_i\}_{i=1}^m$ are i.i.d. $\mathcal{N}(0, 1/m)$, then

$$\delta_{5r}(\mathcal{A}) < \frac{\sqrt{3}-\sqrt{2}}{\sqrt{3}+\sqrt{2}}$$

with high prob., provided that

$$m \gtrsim nr \quad \text{(near-optimal sample size)}$$

This satisfies assumption of Theorem 6.9 with $K = 3$

Precise phase transition

Using statistical dimension machienry, we can locate precise phase transition (Amelunxen, Lotz, McCoy & Tropp '13)

$$\text{nuclear norm min} \begin{cases} \text{works if} & m > \text{stat-dim}\big(\mathcal{D}\left(\|\cdot\|_*, \boldsymbol{X}\right)\big) \\ \text{fails if} & m < \text{stat-dim}\big(\mathcal{D}\left(\|\cdot\|_*, \boldsymbol{X}\right)\big) \end{cases}$$

where

$$\text{stat-dim}\big(\mathcal{D}\left(\|\cdot\|_*, \boldsymbol{X}\right)\big) \;\approx\; n^2 \psi\left(\frac{r}{n}\right)$$

and

$$\psi\left(\rho\right) = \inf_{\tau \geq 0} \left\{ \rho + (1-\rho)\left[\rho(1+\tau^2) + (1-\rho)\int_\tau^2 \left(u-\tau\right)^2 \frac{\sqrt{4-u^2}}{\pi} \mathrm{d}u \right] \right\}$$

Numerical phase transition ($n = 30$)

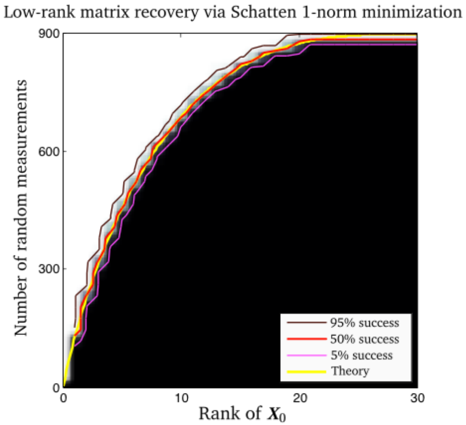


Figure credit: Amelunxen, Lotz, McCoy, & Tropp '13

Sampling operators that do NOT satisfy RIP

Unfortunately, many sampling operators fail to satisfy RIP
(e.g. none of the 4 motivating examples in this lecture satisfies RIP)

Matrix completion

Sampling operators for matrix completion

Observation operator (projection onto matrices supported on Ω)

$$\boldsymbol{Y} = \mathcal{P}_{\Omega}(\boldsymbol{M})$$

where $(i,j) \in \Omega$ with prob. p (random sampling)

- $\mathcal{P}_{\Omega}$ does NOT satisfy RIP when $p \ll 1$!
- For example,

$$\underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}}_{\boldsymbol{M}} \qquad \underbrace{\begin{bmatrix} ? & \checkmark & ? & \checkmark & \checkmark \\ \checkmark & ? & \checkmark & ? & \checkmark \\ ? & \checkmark & \checkmark & ? & ? \\ \checkmark & ? & ? & \checkmark & ? \\ \checkmark & ? & \checkmark & ? & \checkmark \end{bmatrix}}_{\Omega}$$

$\|\mathcal{P}_{\Omega}(\boldsymbol{M})\|_{\mathsf{F}} = 0$, or equivalently, $\frac{1+\delta_K^{\mathrm{ub}}}{1-\delta_{2+K}^{\mathrm{lb}}} = \infty$

Which sampling pattern?

Consider the following sampling pattern

$$\begin{bmatrix} \checkmark & \checkmark & \checkmark & \checkmark & \checkmark \\ ? & ? & ? & ? & ? \\ \checkmark & \checkmark & \checkmark & \checkmark & \checkmark \\ \checkmark & \checkmark & \checkmark & \checkmark & \checkmark \\ \checkmark & \checkmark & \checkmark & \checkmark & \checkmark \end{bmatrix}$$

- If some rows/columns are not sampled, recovery is impossible.

Which low-rank matrices can we recover?

Compare following rank-1 matrices:

$$\underbrace{\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}}_{\text{hard}} \quad \text{vs.} \quad \underbrace{\begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \vdots & & \\ 1 & 1 & 1 & \cdots & 1 \end{bmatrix}}_{\text{easy}}$$

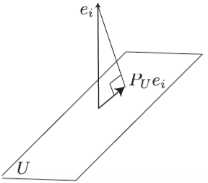
Column / row spaces cannot be aligned with canonical basis vectors

Coherence

Definition 6.10

Coherence parameter μ of $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ is smallest quantity s.t.

$$\max_i \|\boldsymbol{U}^\top \boldsymbol{e}_i\|^2 \le \frac{\mu r}{n} \quad \text{and} \quad \max_i \|\boldsymbol{V}^\top \boldsymbol{e}_i\|^2 \le \frac{\mu r}{n}$$



- $\mu \ge 1$ (since $\sum_{i=1}^n \|\boldsymbol{U}^\top \boldsymbol{e}_i\|^2 = \|\boldsymbol{U}\|_F^2 = r$)
- $\mu = 1$ if $\frac{1}{\sqrt{n}}\mathbf{1} = \boldsymbol{U} = \boldsymbol{V}$ (most incoherent)
- $\mu = \frac{n}{r}$ if $\boldsymbol{e}_i \in \boldsymbol{U}$ (most coherent)

Performance guarantee

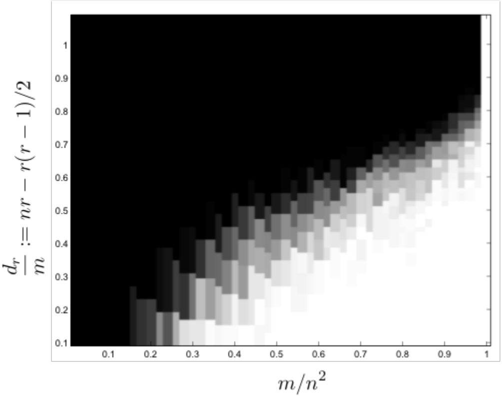
Theorem 6.11 (Candes & Recht '09, Candes & Tao '10, Gross '11, ...)

Nuclear norm minimization is exact and unique with high probability, provided that

$$m \gtrsim \mu n r \log^2 n$$

- This result is optimal up to a logarithmic factor
- Established via a RIPless theory

Numerical performance of nuclear-norm minimization



$n = 50$

Fig. credit: Candes, Recht '09

Subgradient of nuclear norm

Subdifferential (set of subgradients) of $\|\cdot\|_*$ at $\boldsymbol{M}$ is

$$\partial\|\boldsymbol{M}\|_* = \Big\{\boldsymbol{U}\boldsymbol{V}^\top + \boldsymbol{W} : \quad \mathcal{P}_T(\boldsymbol{W}) = 0,\ \|\boldsymbol{W}\| \le 1\Big\}$$

- Does not depend on singular values of $\boldsymbol{M}$
- $\boldsymbol{Z} \in \partial\|\boldsymbol{M}\|_*$ iff

$$\mathcal{P}_T(\boldsymbol{Z}) = \boldsymbol{U}\boldsymbol{V}^\top, \quad \|\mathcal{P}_{T^\perp}(\boldsymbol{Z})\| \le 1.$$

KKT condition

Lagrangian:

$$\mathcal{L}\left(\boldsymbol{X}, \boldsymbol{\Lambda}\right) = \|\boldsymbol{X}\|_* + \langle \boldsymbol{\Lambda}, \mathcal{P}_\Omega(\boldsymbol{X}) - \mathcal{P}_\Omega(\boldsymbol{M}) \rangle = \|\boldsymbol{X}\|_* + \langle \mathcal{P}_\Omega(\boldsymbol{\Lambda}), \boldsymbol{X} - \boldsymbol{M} \rangle$$

When $\boldsymbol{M}$ is minimizer, KKT condition reads

$$\boldsymbol{0} \in \partial_{\boldsymbol{X}} \mathcal{L}(\boldsymbol{X}, \boldsymbol{\Lambda}) \,\Big|_{\,\boldsymbol{X}=\boldsymbol{M}} \Longleftrightarrow \quad \exists \boldsymbol{\Lambda} \text{ s.t. } -\mathcal{P}_\Omega(\boldsymbol{\Lambda}) \in \partial \|\boldsymbol{M}\|_*$$

$$\Longleftrightarrow \quad \exists \boldsymbol{W} \text{ s.t. } \qquad \begin{array}{l} \boldsymbol{U}\boldsymbol{V}^\top + \boldsymbol{W} \text{ is supported on } \Omega, \\ \mathcal{P}_T(\boldsymbol{W}) = \boldsymbol{0}, \text{ and } \|\boldsymbol{W}\| \le 1 \end{array}$$

Optimality condition via dual certificate

Slightly stronger condition than KKT guarantees uniqueness:

Lemma 6.12

$\boldsymbol{M}$ is unique minimizer of nuclear norm minimization if

- *sampling operator $\mathcal{P}_\Omega$ restricted to T is injective, i.e.*

$$\mathcal{P}_\Omega(\boldsymbol{H}) \neq \boldsymbol{0} \quad \forall \text{ nonzero } \boldsymbol{H} \in T$$

- *$\exists \boldsymbol{W}$ s.t.*

$$\begin{aligned} &\boldsymbol{U}\boldsymbol{V}^\top + \boldsymbol{W} \text{ is supported on } \Omega, \\ &\mathcal{P}_T(\boldsymbol{W}) = \boldsymbol{0}, \text{ and } \|\boldsymbol{W}\| < 1 \end{aligned}$$

Proof of Lemma 6.12

For any $\boldsymbol{W}_0$ obeying $\|\boldsymbol{W}_0\| \leq 1$ and $\mathcal{P}_T(\boldsymbol{W}) = \boldsymbol{0}$, one has

$$\begin{aligned}
\|\boldsymbol{M}+\boldsymbol{H}\|_* &\geq \|\boldsymbol{M}\|_* + \left\langle \boldsymbol{U}\boldsymbol{V}^\top + \boldsymbol{W}_0, \boldsymbol{H} \right\rangle \\
&= \|\boldsymbol{M}\|_* + \left\langle \boldsymbol{U}\boldsymbol{V}^\top + \boldsymbol{W}, \boldsymbol{H} \right\rangle + \langle \boldsymbol{W}_0 - \boldsymbol{W}, \boldsymbol{H} \rangle \\
&= \|\boldsymbol{M}\|_* + \left\langle \mathcal{P}_\Omega\left(\boldsymbol{U}\boldsymbol{V}^\top + \boldsymbol{W}\right), \boldsymbol{H} \right\rangle + \langle \mathcal{P}_{T^\perp}(\boldsymbol{W}_0 - \boldsymbol{W}), \boldsymbol{H} \rangle \\
&= \|\boldsymbol{M}\|_* + \left\langle \boldsymbol{U}\boldsymbol{V}^\top + \boldsymbol{W}, \mathcal{P}_\Omega(\boldsymbol{H}) \right\rangle + \langle \boldsymbol{W}_0 - \boldsymbol{W}, \mathcal{P}_{T^\perp}(\boldsymbol{H}) \rangle \\
&\qquad\qquad \text{if we take } \boldsymbol{W}_0 \text{ s.t. } \langle \boldsymbol{W}_0, \mathcal{P}_{T^\perp}(\boldsymbol{H}) \rangle = \|\mathcal{P}_{T^\perp}(\boldsymbol{H})\|_* \\
&\geq \|\boldsymbol{M}\|_* + \|\mathcal{P}_{T^\perp}(\boldsymbol{H})\|_* - \|\boldsymbol{W}\| \cdot \|\mathcal{P}_{T^\perp}(\boldsymbol{H})\|_* \\
&= \|\boldsymbol{M}\|_* + (1 - \|\boldsymbol{W}\|)\, \|\mathcal{P}_{T^\perp}(\boldsymbol{H})\|_* > \|\boldsymbol{M}\|_*
\end{aligned}$$

unless $\mathcal{P}_{T^\perp}(\boldsymbol{H}) = \boldsymbol{0}$.

But if $\mathcal{P}_{T^\perp}(\boldsymbol{H}) = \boldsymbol{0}$, then $\boldsymbol{H} = \boldsymbol{0}$ by injectivity. Thus, $\|\boldsymbol{M}+\boldsymbol{H}\|_* > \|\boldsymbol{M}\|_*$ for any $\boldsymbol{H} \neq \boldsymbol{0}$, concluding the proof.

Constructing dual certificates

Use "golfing scheme" to produce approximate dual certificate (Gross '11)

- Think of it as an iterative algorithm (with sample splitting) to solve KKT

Reference

[1] "*Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization*," B. Recht, M. Fazel, P. Parrilo, *SIAM Review*, 2010.

[2] "*Exact matrix completion via convex optimization*," E. Candes, and B. Recht, *Foundations of Computational Mathematics*, 2009

[3] "*Matrix rank minimization with applications*," M. Fazel, *Ph.D. Thesis*, 2002.

[4] "*The power of convex relaxation: Near-optimal matrix completion*," E. Candes, and T. Tao, 2010.

[5] "*Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements*," E. Candes, and Y. Plan, *IEEE Transactions on Information Theory*, 2011.

[6] "*Shape and motion from image streams under orthography: a factorization method*," C. Tomasi and T. Kanade, *International Journal of Computer Vision*, 1992.

Reference

[7] "*An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems*," K. C. Toh and S. Yun, S, *Pacific Journal of optimization*, 2010.

[8] "*Topics in random matrix theory*," T. Tao, *American mathematical society*, 2012.

[9] "*A singular value thresholding algorithm for matrix completion*," J. Cai, E. Candes, Z. Shen, *SIAM Journal on Optimization*, 2010.

[10] "*Recovering low-rank matrices from few coefficients in any basis*," D. Gross, *IEEE Transactions on Information Theory*, 2011.

[11] "*Incoherence-optimal matrix completion*," Y. Chen, *IEEE Transactions on Information Theory*, 2015.

[12] "*Revisiting Frank-Wolfe: Projection-Free Sparse Convex Optimization*," M. Jaggi, *ICML*, 2013.