

## Low-Rank Matrix Recovery



Yuxin Chen

Princeton University, Spring 2017











# Outline

---

- Low-rank matrix completion and recovery
- Spectral methods
- Nuclear norm minimization
  - RIP and low-rank matrix recovery
  - Phase retrieval / solving random quadratic systems of equations
  - Matrix completion

# Low-rank matrix completion and recovery

# 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?

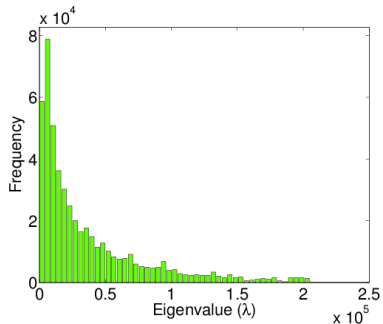
# In general, we cannot infer missing ratings

---

✓	?	?	?	✓	?
?	?	✓	✓	?	?
✓	?	?	✓	?	?
?	?	✓	?	?	✓
✓	?	?	?	?	?
?	✓	?	?	✓	?
?	?	✓	✓	?	?

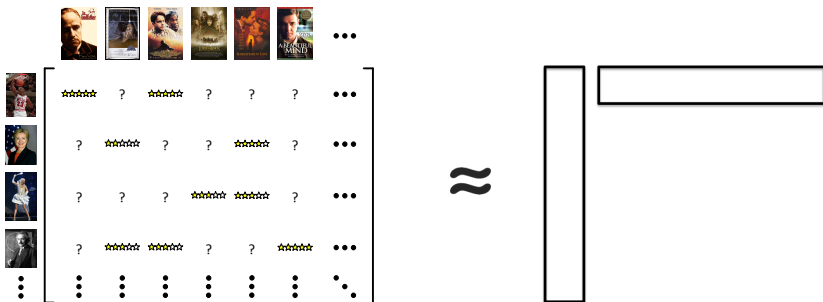
Underdetermined system (more unknowns than observations)

## ... 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?

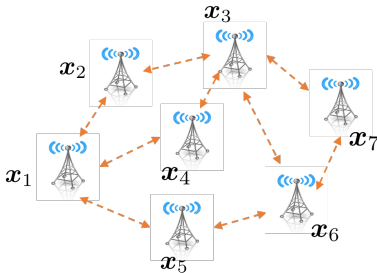
## Motivation 2: sensor localization

---

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

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

- Want to infer distance between every pair of nodes





## Motivation 2: sensor localization

---

Introduce

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

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

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

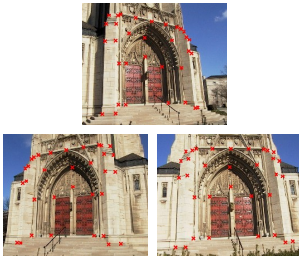
where  $\mathbf{d}_2 := [\|\mathbf{x}_1\|^2, \dots, \|\mathbf{x}_n\|^2]^\top$

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

# Motivation 3: structure from motion

---

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



Snaveley, Seitz, & Szeliski

**Structure from motion:** reconstruct  $\underbrace{\text{3D scene geometry and camera poses}}_{\text{structure}}$  from multiple images  
 $\underbrace{\hspace{10em}}_{\text{motion}}$

## Motivation 3: structure from motion

---

### Tomasi and Kanade's factorization:

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

$$\mathbf{x}_{i,j} = \underbrace{\mathbf{M}_i}_{\text{projection matrix } \in \mathbb{R}^{2 \times 3}} \underbrace{\mathbf{p}_j}_{\text{3D position } \in \mathbb{R}^3}$$

- Matrix of all 2D locations

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

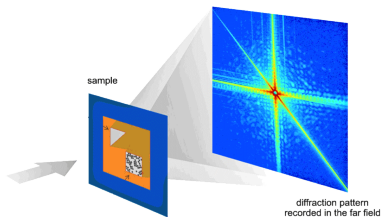
**Goal:** fill in missing entries of  $\mathbf{X}$  given a small number of entries

## Motivation 4: missing phase problem

Detectors record **intensities** of diffracted rays

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

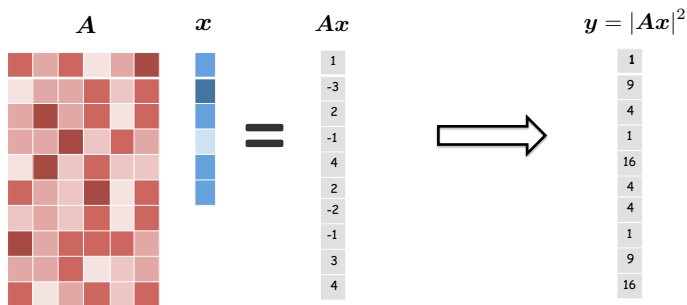
*Fig credit: Stanford SLAC*



intensity of electrical field:  $|\hat{x}(f_1, f_2)|^2 = \left| \int x(t_1, t_2) e^{-i2\pi(f_1 t_1 + f_2 t_2)} dt_1 dt_2 \right|^2$

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

# A discrete-time model: solving quadratic systems



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

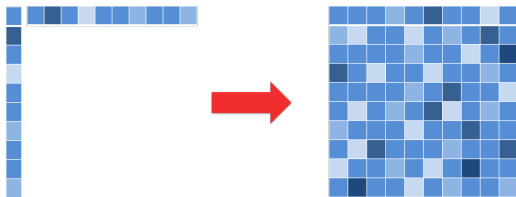
$$y_k = |\langle a_k, x \rangle|^2, \quad k = 1, \dots, m$$

or  $y = |Ax|^2$  where  $|z|^2 := \{|z_1|^2, \dots, |z_m|^2\}$

# An equivalent view: low-rank factorization

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

$$y_k = |\mathbf{a}_k^* \mathbf{x}|^2 = \mathbf{a}_k^* (\mathbf{x}\mathbf{x}^*) \mathbf{a}_k \quad \implies \quad y_k = \mathbf{a}_k^* \mathbf{X} \mathbf{a}_k \quad (8.1)$$



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

# Setup

---

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

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

where  $\Sigma = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{bmatrix}$  contain all singular values  $\{\sigma_i\}$ ;

$U := [\mathbf{u}_1, \dots, \mathbf{u}_r]$ ,  $V := [\mathbf{v}_1, \dots, \mathbf{v}_r]$  consist of singular vectors

# Low-rank matrix completion

---

Observed entries

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

**Completion via rank minimization**

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



# Low-rank matrix completion

---

## Observed entries

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

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

## Completion via rank minimization

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

# More general: low-rank matrix recovery

---

## Linear measurements

$$y_i = \langle \mathbf{A}_i, \mathbf{M} \rangle = \text{Tr}(\mathbf{A}_i^\top \mathbf{M}), \quad i = 1, \dots, m$$

- An operator form

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

## Recovery via rank minimization

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

# Spectral methods

# Signal + noise

---

$(i, j) \in \Omega$  independently with prob.  $p$

One can write observation  $\mathcal{P}_\Omega(M)$  as

$$\frac{1}{p}\mathcal{P}_\Omega(M) = \underbrace{M}_{\text{signal}} + \underbrace{\frac{1}{p}\mathcal{P}_\Omega(M) - M}_{\text{noise}}$$

- Noise has mean zero:  $\mathbb{E}\left[\frac{1}{p}\mathcal{P}_\Omega(M)\right] = M$

# Low-rank denoising

---

$$\frac{1}{p}\mathcal{P}_{\Omega}(M) = \underbrace{M}_{\text{low-rank signal}} + \underbrace{\frac{1}{p}\mathcal{P}_{\Omega}(M) - M}_{:=E \text{ (zero-mean noise)}}$$

---

## Algorithm 8.1 Spectral method

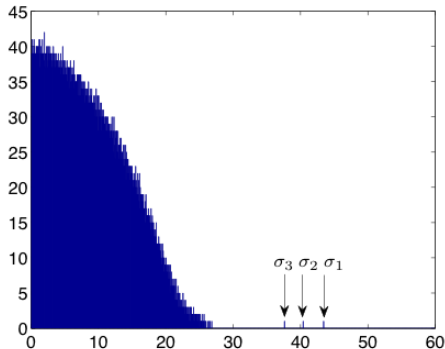
---

$\hat{M} \leftarrow$  best rank- $r$  approximation of  $\frac{1}{p}\mathcal{P}_{\Omega}(M)$

---

# Histograms of singular values of $\mathcal{P}_\Omega(M)$

---



A  $10^4 \times 10^4$  random rank-3 matrix  $M$  with  $p = 0.003$

Fig. credit: Keshavan, Montanari, Oh '10

# Performance of spectral methods

## Theorem 8.1 (Keshavan, Montanari, Oh '10)

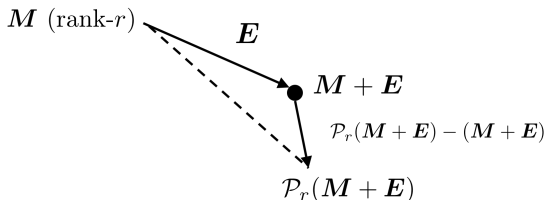
Suppose number of observed entries  $m$  obeys  $m \gtrsim n \log n$ . Then

$$\frac{\|\hat{\mathbf{M}} - \mathbf{M}\|_F}{\|\mathbf{M}\|_F} \lesssim \underbrace{\frac{\max_{i,j} |M_{i,j}|}{\frac{1}{n} \|\mathbf{M}\|_F}}_{:=\nu} \cdot \sqrt{\frac{nr \log^2 n}{m}},$$

- $\nu$  reflects whether energy of  $\mathbf{M}$  is spread out
- When  $m \gg \nu^2 n \log^2 n$ , estimate  $\hat{\mathbf{M}}$  is very close to truth<sup>1</sup>
- Degrees of freedom  $\asymp rn$   
→ nearly-optimal sample complexity for **incoherent** matrices

<sup>1</sup>The logarithmic factor can be improved.

# Perturbation bounds



To ensure  $\hat{M}$  is good estimate, it suffices to control noise  $E$

## Lemma 8.2

Suppose  $\text{rank}(M) = r$ . For any perturbation  $E$ ,

$$\|\mathcal{P}_r(M + E) - M\| \leq 2\|E\|$$

$$\|\mathcal{P}_r(M + E) - M\|_F \leq 2\sqrt{2r}\|E\|$$

where  $\mathcal{P}_r(X)$  is best rank- $r$  approximation of  $X$



## Proof of Lemma 8.2

---

By matrix perturbation theory,

$$\begin{aligned} & \|\mathcal{P}_r(\mathbf{M} + \mathbf{E}) - \mathbf{M}\| \\ & \stackrel{\text{triangle inequality}}{\leq} \|\mathcal{P}_r(\mathbf{M} + \mathbf{E}) - (\mathbf{M} + \mathbf{E})\| + \|(\mathbf{M} + \mathbf{E}) - \mathbf{M}\| \\ & \leq \sigma_{r+1}(\mathbf{M} + \mathbf{E}) + \|\mathbf{E}\| \\ & \stackrel{\text{Weyl's inequality}}{\leq} \underbrace{\sigma_{r+1}(\mathbf{M})}_{=0} + \|\mathbf{E}\| + \|\mathbf{E}\| = 2\|\mathbf{E}\| \end{aligned}$$

The 2nd inequality of Lemma 8.2 follows since both  $\mathcal{P}_r(\mathbf{M} + \mathbf{E})$  and  $\mathbf{M}$  are rank- $r$ , and hence

$$\|\mathcal{P}_r(\mathbf{M} + \mathbf{E}) - \mathbf{M}\|_{\text{F}} \leq \sqrt{2r} \|\mathcal{P}_r(\mathbf{M} + \mathbf{E}) - \mathbf{M}\|$$

# Controlling the noise

---

Recall that entries of  $\mathbf{E} = \frac{1}{p}\mathcal{P}_\Omega(\mathbf{M}) - \mathbf{M}$  are zero-mean and independent

A bit of random matrix theory ...

## Lemma 8.3 (Chapter 2.3, Tao '12)

*Suppose  $\mathbf{X} \in \mathbb{R}^{n \times n}$  is a random symmetric matrix obeying*

- $\{X_{i,j} : i < j\}$  are independent*
- $\mathbb{E}[X_{i,j}] = 0$  and  $\text{Var}[X_{i,j}] \lesssim 1$*
- $\max_{i,j} |X_{i,j}| \lesssim \sqrt{n}$*

*Then  $\|\mathbf{X}\| \lesssim \sqrt{n} \log n$*

# Proof of Theorem 8.1

---

Let  $M_{\max} := \max_{i,j} |M_{i,j}|$ . The zero-mean matrix  $\tilde{\mathbf{E}} = \frac{\sqrt{p}}{M_{\max}} \mathbf{E}$  obeys

$$\begin{aligned}\mathrm{Var} [\tilde{E}_{i,j}] &= p(1-p) \cdot \left( \frac{\sqrt{p}}{M_{\max}} \cdot \frac{1}{p} M_{i,j} \right)^2 \leq \left( \frac{M_{i,j}}{M_{\max}} \right)^2 \lesssim 1, \\ |\tilde{E}_{i,j}| &\leq \frac{|M_{i,j}|}{\sqrt{p} M_{\max}} \lesssim \frac{1}{\sqrt{p}}.\end{aligned}$$

Lemma 8.3 tells us that if  $p \gtrsim \frac{1}{n}$ , then

$$\|\tilde{\mathbf{E}}\| \lesssim \sqrt{n} \log n \iff \|\mathbf{E}\| \lesssim \frac{M_{\max} \sqrt{n}}{\sqrt{p}} \log n$$

This together with Lemma 8.2 and the fact  $m \asymp pn^2$  establishes Theorem 8.1

# Nuclear norm minimization

# Convex relaxation

---

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

$$\begin{array}{ll} \text{minimize}_{\mathbf{X}} & \underbrace{\text{rank}(\mathbf{X})}_{\text{nonconvex}} \\ \text{s.t.} & \mathcal{A}(\mathbf{X}) = \mathcal{A}(\mathbf{M}) \end{array}$$

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

# Nuclear norm

## Definition 8.4

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

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

- Nuclear norm is a counterpart of  $\ell_1$  norm for rank function
- Equivalence among different norms

$$\|\mathbf{X}\| \leq \|\mathbf{X}\|_F \leq \|\mathbf{X}\|_* \leq \sqrt{r} \|\mathbf{X}\|_F \leq r \|\mathbf{X}\|.$$

- **(Tightness)**  $\{\mathbf{X} : \|\mathbf{X}\|_* \leq 1\}$  is convex hull of rank-1 matrices obeying  $\|\mathbf{u}\mathbf{v}^\top\| \leq 1$  (Fazel '02)

# Additivity of nuclear norm

## Fact 8.5

*Let  $A$  and  $B$  be matrices of the same dimensions. If  $AB^T = 0$  and  $A^T B = 0$ , then  $\|A + B\|_* = \|A\|_* + \|B\|_*$ .*

- If row (resp. column) spaces of  $A$  and  $B$  are orthogonal, then  $\|A + B\|_* = \|A\|_* + \|B\|_*$
- Similar to  $\ell_1$  norm: when  $x$  and  $y$  have disjoint support,

$$\|x + y\|_1 = \|x\|_1 + \|y\|_1 \quad (\text{a key to study } \ell_1\text{-min under RIP})$$

## Proof of Fact 8.5

---

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

$$\begin{array}{lcl} \mathbf{A} \mathbf{B}^\top & = & \mathbf{0} \\ \mathbf{A}^\top \mathbf{B} & = & \mathbf{0} \end{array} \iff \begin{array}{lcl} \mathbf{V}_A^\top \mathbf{V}_B & = & \mathbf{0} \\ \mathbf{U}_A^\top \mathbf{U}_B & = & \mathbf{0} \end{array}$$

Thus, one can write

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

and hence

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



# Dual norm

## Definition 8.6 (Dual norm)

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

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

- $\ell_1$  norm  $\overset{\text{dual}}{\longleftrightarrow}$   $\ell_{\infty}$  norm
- nuclear norm  $\overset{\text{dual}}{\longleftrightarrow}$  spectral norm
- $\ell_2$  norm  $\overset{\text{dual}}{\longleftrightarrow}$   $\ell_2$  norm
- Frobenius norm  $\overset{\text{dual}}{\longleftrightarrow}$  Frobenius norm

# Representing nuclear norm via SDP

Since spectral norm is dual norm of nuclear norm,

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

The constraint is equivalent to

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

## Fact 8.7

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

# Representing nuclear norm via SDP

Since spectral norm is dual norm of nuclear norm,

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

The constraint is equivalent to

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

## Fact 8.8 (Dual characterization)

$$\|X\|_* = \min_{W_1, W_2} \left\{ \frac{1}{2} \text{Tr}(W_1) + \frac{1}{2} \text{Tr}(W_2) \mid \begin{bmatrix} W_1 & X \\ X^\top & W_2 \end{bmatrix} \succeq 0 \right\}$$

- Optimal point:  $W_1 = U\Sigma U^\top$ ,  $W_2 = V\Sigma V^\top$  (where  $X = U\Sigma V^\top$ )

## Aside: dual of semidefinite program

---

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



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

Exercise: use this to verify Fact 8.8

# Nuclear norm minimization via SDP

---

## Convex relaxation of rank minimization

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

This is solvable via SDP

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

# **RIP and low-rank matrix recovery**

# RIP for low-rank matrices

---

*Almost parallel results to compressed sensing ...<sup>2</sup>*

## Definition 8.9

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

$$(1 - \delta_r^{\text{lb}})\|\mathbf{X}\|_{\text{F}} \leq \|\mathcal{A}(\mathbf{X})\|_{\text{F}} \leq (1 + \delta_r^{\text{ub}})\|\mathbf{X}\|_{\text{F}}, \quad \forall \mathbf{X} : \text{rank}(\mathbf{X}) \leq r$$

---

<sup>2</sup>One can also define RIP w.r.t.  $\|\cdot\|_{\text{F}}^2$  rather than  $\|\cdot\|_{\text{F}}$ .

# RIP and low-rank matrix recovery

---

## Theorem 8.10 (Recht, Fazel, Parrilo '10, Candes, Plan '11)

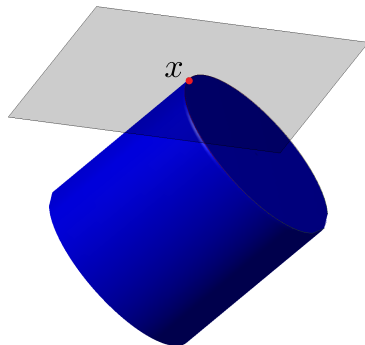
*Suppose  $\text{rank}(\mathbf{M}) = r$ . For any fixed integer  $K > 0$ , if  $\frac{1+\delta_{Kr}^{\text{ub}}}{1-\delta_{(2+K)r}^{\text{lb}}} < \sqrt{\frac{K}{2}}$ , then nuclear norm minimization is exact*

- 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  $M = U\Sigma V^\top$

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

$$T = \{UX^\top + YV^\top : X, Y \in \mathbb{R}^{n \times r}\}$$

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

$$\mathcal{P}_T(X) = UU^\top X + XVV^\top - UU^\top XVV^\top$$

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

$$\mathcal{P}_{T^\perp}(X) = (I - UU^\top)X(I - VV^\top)$$

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

## Proof of Theorem 8.10

---

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

The key is to decompose  $\mathbf{H}$  into  $\mathbf{H}_0 + \underbrace{\mathbf{H}_1 + \mathbf{H}_2 + \dots}_{\mathbf{H}_c}$

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

## Proof of Theorem 8.10

---

Informally, the proof proceeds by showing that

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

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

## Proof of Theorem 8.10

---

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

$$\sum_{j \geq 2} \|\mathbf{H}_j\|_F \leq \|\mathbf{H}_0\|_* / \sqrt{Kr}. \quad (8.2)$$

This follows immediately by combining the following 2 observations:

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

$$\begin{aligned} \|\mathbf{M}\|_* &\geq \|\mathbf{M} + \mathbf{H}\|_* \geq \|\mathbf{M} + \mathbf{H}_c\|_* - \|\mathbf{H}_0\|_* \\ &\geq \underbrace{\|\mathbf{M}\|_* + \|\mathbf{H}_c\|_*}_{\text{Fact 8.5 } (\mathbf{M}\mathbf{H}_c^\top = \mathbf{0} \text{ and } \mathbf{M}^\top \mathbf{H}_c = \mathbf{0})} - \|\mathbf{H}_0\|_* \end{aligned} \quad (8.3)$$

$$\implies \|\mathbf{H}_c\|_* \leq \|\mathbf{H}_0\|_* \quad (8.4)$$

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

$$\begin{aligned} \|\mathbf{H}_j\|_F &\leq \sqrt{Kr} \|\mathbf{H}_j\| \leq \sqrt{Kr} [\|\mathbf{H}_{j-1}\|_* / (Kr)] \leq \|\mathbf{H}_{j-1}\|_* / \sqrt{Kr} \\ \implies \sum_{j \geq 2} \|\mathbf{H}_j\|_F &\leq \frac{1}{\sqrt{Kr}} \sum_{j \geq 2} \|\mathbf{H}_{j-1}\|_* = \frac{1}{\sqrt{Kr}} \|\mathbf{H}_c\|_* \end{aligned} \quad (8.5)$$

## Proof of Theorem 8.10

---

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

$$\|\mathbf{H}_0 + \mathbf{H}_1\|_F \leq \rho \sum_{j \geq 2} \|\mathbf{H}_j\|_F \quad (8.6)$$

If this claim holds, then

$$\begin{aligned} \|\mathbf{H}_0 + \mathbf{H}_1\|_F &\leq \rho \sum_{j \geq 2} \|\mathbf{H}_j\|_F \stackrel{(8.2)}{\leq} \rho \frac{1}{\sqrt{Kr}} \|\mathbf{H}_0\|_* \\ &\leq \rho \frac{1}{\sqrt{Kr}} \left( \sqrt{2r} \|\mathbf{H}_0\|_F \right) = \rho \sqrt{\frac{2}{K}} \|\mathbf{H}_0\|_F \\ &\leq \rho \sqrt{\frac{2}{K}} \|\mathbf{H}_0 + \mathbf{H}_1\|_F. \end{aligned} \quad (8.7)$$

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

## Proof of Theorem 8.10

---

We now prove (8.6). To connect  $\mathbf{H}_0 + \mathbf{H}_1$  with  $\sum_{j \geq 2} \mathbf{H}_j$ , we use feasibility:

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

which taken collectively with RIP yields

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

This establishes (8.6) as long as  $\rho := \frac{1 + \delta_{Kr}^{\text{ub}}}{1 - \delta_{(2+K)r}^{\text{lb}}} < \sqrt{\frac{K}{2}}$ .

# Gaussian sampling operators satisfy RIP

---

If entries of  $\{\mathbf{A}_i\}_{i=1}^m$  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 8.10 with  $K = 3$



## Precise phase transition

---

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

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

where

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

and

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

## Aside: subgradient of nuclear norm

---

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

$$\partial\|\mathbf{M}\|_* = \left\{ \mathbf{UV}^\top + \mathbf{W} : \mathcal{P}_T(\mathbf{W}) = 0, \|\mathbf{W}\| \leq 1 \right\}$$

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

$$\mathcal{P}_T(\mathbf{Z}) = \mathbf{UV}^\top, \quad \|\mathcal{P}_{T^\perp}(\mathbf{Z})\| \leq 1.$$

# Derivation of statistical dimension

---

WLOG, suppose  $\mathbf{X} = \begin{bmatrix} \mathbf{I}_r & \\ & \mathbf{0} \end{bmatrix}$ , then  $\partial \|\mathbf{X}\|_* = \left\{ \begin{bmatrix} \mathbf{I}_r & \\ & \mathbf{W} \end{bmatrix} \mid \|\mathbf{W}\| \leq 1 \right\}$ .

Let  $\mathbf{G} = \begin{bmatrix} \mathbf{G}_{11} & \mathbf{G}_{12} \\ \mathbf{G}_{21} & \mathbf{G}_{22} \end{bmatrix}$  be i.i.d. standard Gaussian.

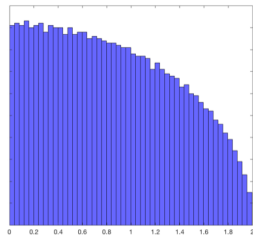
From last lecture, we know that

$$\begin{aligned} \text{stat-dim}(\mathcal{D}(\|\cdot\|_*, \mathbf{X})) &\approx \inf_{\tau \geq 0} \mathbb{E} \left[ \inf_{\mathbf{Z} \in \partial \|\mathbf{X}\|_*} \|\mathbf{G} - \tau \mathbf{Z}\|_{\text{F}}^2 \right] \\ &= \inf_{\tau \geq 0} \mathbb{E} \left[ \inf_{\mathbf{W}: \|\mathbf{W}\| \leq 1} \left\| \begin{bmatrix} \mathbf{G}_{11} & \mathbf{G}_{12} \\ \mathbf{G}_{21} & \mathbf{G}_{22} \end{bmatrix} - \tau \begin{bmatrix} \mathbf{I}_r & \\ & \mathbf{W} \end{bmatrix} \right\|_{\text{F}}^2 \right] \end{aligned}$$

# Derivation of statistical dimension

Observe that

$$\begin{aligned} & \mathbb{E} \left[ \inf_{\mathbf{W}: \|\mathbf{W}\| \leq 1} \left\| \begin{bmatrix} \mathbf{G}_{11} & \mathbf{G}_{12} \\ \mathbf{G}_{21} & \mathbf{G}_{22} \end{bmatrix} - \tau \begin{bmatrix} \mathbf{I}_r & \\ & \mathbf{W} \end{bmatrix} \right\|_{\text{F}}^2 \right] \\ &= \mathbb{E} \left[ \|\mathbf{G}_{11} - \tau \mathbf{I}_r\|_{\text{F}}^2 + \|\mathbf{G}_{21}\|_{\text{F}}^2 + \|\mathbf{G}_{12}\|_{\text{F}}^2 + \inf_{\|\mathbf{W}\| \leq 1} \|\mathbf{G}_{22} - \tau \mathbf{W}\|_{\text{F}}^2 \right] \\ &= r(2n - r + \tau^2) + \mathbb{E} \left[ \sum_{i=1}^{n-r} (\sigma_i(\mathbf{G}_{22}) - \tau)_+^2 \right]. \end{aligned}$$



empirical distributions of  $\{\sigma_i(\mathbf{G}_{22})/\sqrt{n-r}\}$

# Derivation of statistical dimension

---

Observe that

$$\begin{aligned} & \mathbb{E} \left[ \inf_{\mathbf{W}: \|\mathbf{W}\| \leq 1} \left\| \begin{bmatrix} \mathbf{G}_{11} & \mathbf{G}_{12} \\ \mathbf{G}_{21} & \mathbf{G}_{22} \end{bmatrix} - \tau \begin{bmatrix} \mathbf{I}_r & \\ & \mathbf{W} \end{bmatrix} \right\|_{\text{F}}^2 \right] \\ &= \mathbb{E} \left[ \|\mathbf{G}_{11} - \tau \mathbf{I}_r\|_{\text{F}}^2 + \|\mathbf{G}_{21}\|_{\text{F}}^2 + \|\mathbf{G}_{12}\|_{\text{F}}^2 + \inf_{\|\mathbf{W}\| \leq 1} \|\mathbf{G}_{22} - \tau \mathbf{W}\|_{\text{F}}^2 \right] \\ &= r(2n - r + \tau^2) + \mathbb{E} \left[ \sum_{i=1}^{n-r} (\sigma_i(\mathbf{G}_{22}) - \tau)_+^2 \right]. \end{aligned}$$

Recall from random matrix theory ([Marchenko-Pastur law](#))

$$\frac{1}{n-r} \mathbb{E} \left[ \sum_{i=1}^{n-r} (\sigma_i(\tilde{\mathbf{G}}_{22}) - \tau)_+^2 \right] \rightarrow \int_0^2 (u - \tau)_+^2 \frac{\sqrt{4 - u^2}}{\pi} du,$$

where  $\tilde{\mathbf{G}}_{22} \sim \mathcal{N}(\mathbf{0}, \frac{1}{n-r} \mathbf{I})$ . Taking  $\rho = r/n$  and minimizing over  $\tau$  lead to closed-form expression for phase transition boundary.

# Numerical phase transition ( $n = 30$ )

---

Low-rank matrix recovery via Schatten 1-norm minimization

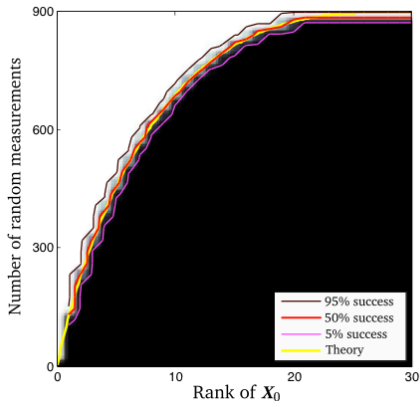


Figure credit: Amelunxen, Lotz, McCoy, & Tropp '13

# 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)

## Two important examples:

- Phase retrieval / solving random quadratic systems of equations
- Matrix completion

# **Phase retrieval / solving random quadratic systems of equations**



# Rank-one measurements

---

Measurements: see (8.1)

$$y_i = \mathbf{a}_i^\top \underbrace{\mathbf{x}\mathbf{x}^\top}_{:=\mathbf{M}} \mathbf{a}_i = \langle \underbrace{\mathbf{a}_i\mathbf{a}_i^\top}_{:=\mathbf{A}_i}, \mathbf{M} \rangle, \quad 1 \leq i \leq m$$

$$\mathcal{A}(\mathbf{X}) = \begin{bmatrix} \langle \mathbf{A}_1, \mathbf{X} \rangle \\ \langle \mathbf{A}_2, \mathbf{X} \rangle \\ \vdots \\ \langle \mathbf{A}_m, \mathbf{X} \rangle \end{bmatrix} = \begin{bmatrix} \langle \mathbf{a}_1\mathbf{a}_1^\top, \mathbf{X} \rangle \\ \langle \mathbf{a}_2\mathbf{a}_2^\top, \mathbf{X} \rangle \\ \vdots \\ \langle \mathbf{a}_m\mathbf{a}_m^\top, \mathbf{X} \rangle \end{bmatrix}$$

# Rank-one measurements

---

Suppose  $\mathbf{a}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$

- If  $\mathbf{x}$  is independent of  $\{\mathbf{a}_i\}$ , then

$$\langle \mathbf{a}_i \mathbf{a}_i^\top, \mathbf{x} \mathbf{x}^\top \rangle = |\mathbf{a}_i^\top \mathbf{x}|^2 \asymp \|\mathbf{x}\|^2 \Rightarrow \|\mathcal{A}(\mathbf{x} \mathbf{x}^\top)\|_{\text{F}} \asymp \sqrt{m} \|\mathbf{x} \mathbf{x}^\top\|_{\text{F}}$$

- Consider  $\mathbf{A}_i = \mathbf{a}_i \mathbf{a}_i^\top$ :

$$\langle \mathbf{a}_i \mathbf{a}_i^\top, \mathbf{A}_i \rangle = \|\mathbf{a}_i\|^4 \approx n \|\mathbf{a}_i \mathbf{a}_i^\top\|_{\text{F}}$$

$$\implies \|\mathcal{A}(\mathbf{A}_i)\|_{\text{F}} \geq |\langle \mathbf{a}_i \mathbf{a}_i^\top, \mathbf{A}_i \rangle| \approx n \|\mathbf{A}_i\|_{\text{F}}$$

# Rank-one measurements

---

Suppose  $\mathbf{a}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$

- If sample size  $m \asymp n$  (information limit), then

$$\frac{\max_{\mathbf{X}: \text{rank}(\mathbf{X})=1} \frac{\|\mathcal{A}(\mathbf{X})\|_F}{\|\mathbf{X}\|_F}}{\min_{\mathbf{X}: \text{rank}(\mathbf{X})=1} \frac{\|\mathcal{A}(\mathbf{X})\|_F}{\|\mathbf{X}\|_F}} \gtrsim \frac{n}{\sqrt{m}} \gtrsim \sqrt{n}$$

$$\Rightarrow \frac{1 + \delta_K^{\text{ub}}}{1 - \delta_{2+K}^{\text{lb}}} \geq \frac{\max_{\mathbf{X}: \text{rank}(\mathbf{X})=1} \frac{\|\mathcal{A}(\mathbf{X})\|_F}{\|\mathbf{X}\|_F}}{\min_{\mathbf{X}: \text{rank}(\mathbf{X})=1} \frac{\|\mathcal{A}(\mathbf{X})\|_F}{\|\mathbf{X}\|_F}} \gtrsim \sqrt{n} \gg \sqrt{K}$$

- Violate RIP condition in Theorem 8.10

# Why do we lose RIP?

---

## Problem:

- Low-rank matrices  $\mathbf{X}$  (e.g.  $\mathbf{a}_i \mathbf{a}_i^\top$ ) might be too aligned with some rank-one measurements
  - loss of incoherence in some measurements
- Some measurements  $\langle \mathbf{A}_i, \mathbf{X} \rangle$  might have too high of a leverage on  $\mathcal{A}(\mathbf{X})$  when measured in  $\|\cdot\|_{\mathbf{F}}$ 
  - Change  $\|\cdot\|_{\mathbf{F}}$  to other norms!

# Mixed-norm RIP

---

**Solution:** modify RIP appropriately ...

## Definition 8.11 (RIP- $\ell_2/\ell_1$ )

Let  $\xi_r^{\text{ub}}(\mathcal{A})$  and  $\xi_r^{\text{lb}}(\mathcal{A})$  be smallest quantities s.t.

$$(1 - \xi_r^{\text{lb}})\|\mathbf{X}\|_{\text{F}} \leq \|\mathcal{A}(\mathbf{X})\|_1 \leq (1 + \xi_r^{\text{ub}})\|\mathbf{X}\|_{\text{F}}, \quad \forall \mathbf{X} : \text{rank}(\mathbf{X}) \leq r$$

# Analyzing phase retrieval via RIP- $\ell_2/\ell_1$

---

## Theorem 8.12 (Chen, Chi, Goldsmith '15)

*Theorem 8.10 continues to hold if we replace  $\delta_r^{\text{ub}}$  and  $\delta_r^{\text{lb}}$  with  $\xi_r^{\text{ub}}$  and  $\xi_r^{\text{lb}}$ , respectively.*

- Follows same proof as for Theorem 8.10, except that  $\|\cdot\|_{\text{F}}$  (highlighted in red) is replaced by  $\|\cdot\|_1$  in Slide 8-44

# Analyzing phase retrieval via RIP- $\ell_2/\ell_1$

## Theorem 8.12 (Chen, Chi, Goldsmith '15)

*Theorem 8.10 continues to hold if we replace  $\delta_r^{\text{ub}}$  and  $\delta_r^{\text{lb}}$  with  $\xi_r^{\text{ub}}$  and  $\xi_r^{\text{lb}}$ , respectively.*

- Back to example in Slide 8-54:
  - If  $\mathbf{x}$  is independent of  $\{\mathbf{a}_i\}$ , then

$$\langle \mathbf{a}_i \mathbf{a}_i^\top, \mathbf{x} \mathbf{x}^\top \rangle = |\mathbf{a}_i^\top \mathbf{x}|^2 \asymp \|\mathbf{x}\|^2 \Rightarrow \|\mathcal{A}(\mathbf{x} \mathbf{x}^\top)\|_1 \asymp m \|\mathbf{x} \mathbf{x}^\top\|_F$$

- $\|\mathcal{A}(\mathbf{A}_i)\|_1 = |\langle \mathbf{a}_i \mathbf{a}_i^\top, \mathbf{A}_i \rangle| + \sum_{j:j \neq i} |\langle \mathbf{a}_i \mathbf{a}_i^\top, \mathbf{A}_j \rangle| \approx (n+m) \|\mathbf{A}_i\|_F$
- For both cases,  $\frac{\|\mathcal{A}(\mathbf{X})\|_1}{\|\mathbf{X}\|_F}$  are of same order

# Analyzing phase retrieval via RIP- $\ell_2/\ell_1$

---

A **debiased** operator satisfies RIP condition of Theorem 8.12 when  $m \gtrsim nr$

$$\mathcal{B}(\mathbf{X}) := \begin{bmatrix} \langle \mathbf{A}_1 - \mathbf{A}_2, \mathbf{X} \rangle \\ \langle \mathbf{A}_3 - \mathbf{A}_4, \mathbf{X} \rangle \\ \vdots \end{bmatrix} \in \mathbb{R}^{m/2}$$

- Debiasing is crucial when  $r \gg 1$
- A consequence of Hanson-Wright inequality for quadratic form (Hanson & Wright '71, Rudelson & Vershynin '03)



# Theoretical guarantee for phase retrieval

$$\begin{aligned} (\text{PhaseLift}) \quad & \underset{\mathbf{X} \in \mathbb{R}^{n \times n}}{\text{minimize}} \quad \underbrace{\text{Tr}(\mathbf{X})}_{\|\cdot\|_* \text{ for PSD matrices}} \\ & \text{s.t.} \quad y_i = \mathbf{a}_i^\top \mathbf{X} \mathbf{a}_i, \quad 1 \leq i \leq m \\ & \quad \mathbf{X} \succeq \mathbf{0} \quad (\text{since } \mathbf{X} = \mathbf{x}\mathbf{x}^\top) \end{aligned}$$

**Theorem 8.13 (Candes, Strohmer, Voroninski '13, Candes, Li '14, ...)**

*Suppose  $\mathbf{a}_i \stackrel{\text{ind.}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I})$ . With high prob., PhaseLift recovers  $\mathbf{x}\mathbf{x}^\top$  exactly as soon as  $m \gtrsim n$*

# Extension of phase retrieval

$$\begin{aligned} \text{(PhaseLift)} \quad & \underset{\mathbf{X} \in \mathbb{R}^{n \times n}}{\text{minimize}} \quad \underbrace{\text{Tr}(\mathbf{X})}_{\|\cdot\|_* \text{ for PSD matrices}} \\ & \text{s.t.} \quad \mathbf{a}_i^\top \mathbf{X} \mathbf{a}_i = \mathbf{a}_i^\top \mathbf{M} \mathbf{a}_i, \quad 1 \leq i \leq m \\ & \quad \mathbf{X} \succeq \mathbf{0} \end{aligned}$$

**Theorem 8.14 (Chen, Chi, Goldsmith '15, Cai, Zhang '15, Kueng, Rauhut, Terstiege '17)**

*Suppose  $\mathbf{M} \succeq \mathbf{0}$ ,  $\text{rank}(\mathbf{M}) = r$ , and  $\mathbf{a}_i \stackrel{\text{ind.}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I})$ . With high prob., PhaseLift recovers  $\mathbf{M}$  exactly as soon as  $m \gtrsim nr$*

# Matrix completion

# Sampling operators for matrix completion

Observation operator (projection onto matrices supported on  $\Omega$ )

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

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

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

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

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

# 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:

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

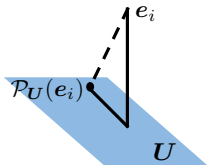
Column / row spaces cannot be aligned with canonical basis vectors

# Coherence

## Definition 8.15

Coherence parameter  $\mu$  of  $M = U\Sigma V^\top$  is smallest quantity s.t.

$$\max_i \|U^\top e_i\|^2 \leq \frac{\mu r}{n} \quad \text{and} \quad \max_i \|V^\top e_i\|^2 \leq \frac{\mu r}{n}$$



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

# Performance guarantee

---

## Theorem 8.16 (Candes & Recht '09, Candes & Tao '10, Gross '11, ...)

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

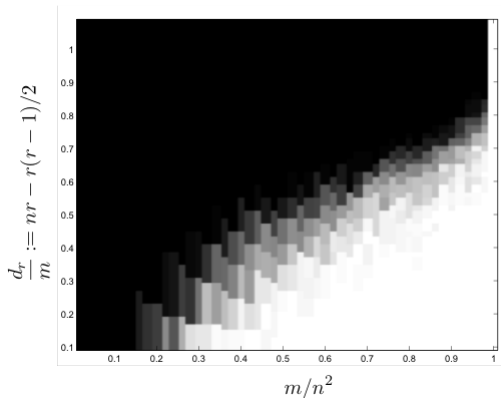
$$m \gtrsim \mu n r \log^2 n$$

- 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

# KKT condition

---

Lagrangian:

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

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

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

$$\iff \exists \mathbf{W} \text{ s.t. } \quad \mathbf{U}\mathbf{V}^\top + \mathbf{W} \text{ is supported on } \Omega, \\ \mathcal{P}_T(\mathbf{W}) = \mathbf{0}, \text{ and } \|\mathbf{W}\| \leq 1$$

# Optimality condition via dual certificate

---

Slightly stronger condition than KKT guarantees uniqueness:

## Lemma 8.17

*$M$  is unique minimizer of nuclear norm minimization if*

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

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

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

$$\mathbf{UV}^\top + \mathbf{W} \text{ is supported on } \Omega,$$

$$\mathcal{P}_T(\mathbf{W}) = \mathbf{0}, \text{ and } \|\mathbf{W}\| < 1$$

## Proof of Lemma 8.17

---

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

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

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

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

# Proximal algorithm

---

In the presence of noise, one needs to solve

$$\hat{\mathbf{X}} = \operatorname{argmin}_{\mathbf{X}} \frac{1}{2} \|\mathbf{y} - \mathcal{A}(\mathbf{X})\|_{\text{F}}^2 + \lambda \|\mathbf{X}\|_*$$

which can be solved via proximal methods

**Proximal operator:**

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

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

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

# Proximal algorithm

---

---

## Algorithm 8.2 Proximal gradient methods

---

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

$$\mathbf{X}^{t+1} = \mathcal{T}_{\mu_t} \left( \mathbf{X}^t - \mu_t \mathcal{A}^* \mathcal{A}(\mathbf{X}^t) \right)$$

where  $\mu_t$ : step size / learning rate

---

# Reference

---

- [1] "*Lecture notes, Advanced topics in signal processing (ECE 8201)*," Y. Chi, 2015.
- [2] "*Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization*," B. Recht, M. Fazel, P. Parrilo, *SIAM Review*, 2010.
- [3] "*Exact matrix completion via convex optimization*," E. Candes, and B. Recht, *Foundations of Computational Mathematics*, 2009
- [4] "*Matrix completion from a few entries*," R. Keshavan, A. Montanari, S. Oh, *IEEE Transactions on Information Theory*, 2010.
- [5] "*Matrix rank minimization with applications*," M. Fazel, *Ph.D. Thesis*, 2002.
- [6] "*Modeling the world from internet photo collections*," N. Snavely, S. Seitz, and R. Szeliski, *International Journal of Computer Vision*, 2008.



# Reference

---

- [7] "*Computer vision: algorithms and applications*," R. Szeliski, *Springer*, 2011.
- [8] "*Shape and motion from image streams under orthography: a factorization method*," C. Tomasi and T. Kanade, *International Journal of Computer Vision*, 1992.
- [9] "*Topics in random matrix theory*," T. Tao, *American mathematical society*, 2012.
- [10] "*A singular value thresholding algorithm for matrix completion*," J. Cai, E. Candes, Z. Shen, *SIAM Journal on Optimization*, 2010.
- [11] "*Recovering low-rank matrices from few coefficients in any basis*," D. Gross, *IEEE Transactions on Information Theory*, 2011.
- [12] "*Incoherence-optimal matrix completion*," Y. Chen, *IEEE Transactions on Information Theory*, 2015.

# Reference

---

- [13] "*The power of convex relaxation: Near-optimal matrix completion*," E. Candes, and T. Tao, 2010.
- [14] "*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.
- [15] "*PhaseLift: Exact and stable signal recovery from magnitude measurements via convex programming*," E. Candes, T. Strohmer, and V. Voroninski, *Communications on Pure and Applied Mathematics*, 2013.
- [16] "*Solving quadratic equations via PhaseLift when there are about as many equations as unknowns*," E. Candes, X. Li, *Foundations of Computational Mathematics*, 2014.
- [17] "*A bound on tail probabilities for quadratic forms in independent random variables*," D. Hanson, F. Wright, *Annals of Mathematical Statistics*, 1971.

# Reference

---

- [18] "*Hanson-Wright inequality and sub-Gaussian concentration*," M. Rudelson, R. Vershynin, *Electronic Communications in Probability*, 2013.
- [19] "*Exact and stable covariance estimation from quadratic sampling via convex programming* ," Y. Chen, Y. Chi, and A. Goldsmith, *IEEE Transactions on Information Theory*, 2015.
- [20] "*ROP: Matrix recovery via rank-one projections*," T. Cai, and A. Zhang, *Annals of Statistics*, 2015.
- [21] "*Low rank matrix recovery from rank one measurements*," R. Kueng, H. Rauhut, and U. Terstiege, *Applied and Computational Harmonic Analysis*, 2017.
- [22] "*Mathematics of sparsity (and a few other things)*," E. Candes, *International Congress of Mathematicians*, 2014.