ECE 8201: Low-dimensional Signal Models for High-dimensional Data Analysis

Lecture 7: Matrix completion

Yuejie Chi
The Ohio State University

THE OHIO STATE UNIVERSITY

Reference

- "Guaranteed Minimum-Rank Solutions of Linear Matrix Equations via Nuclear Norm Minimization", Recht, Fazel, and Parrilo, 2007.

- "The power of convex relaxation: Near-optimal matrix completion", E. J. Candès and T. Tao, 2007.

Outline

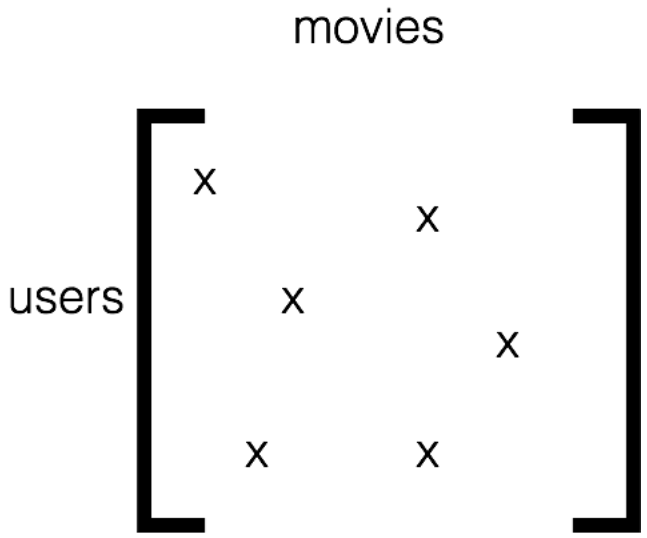
Matrix completion

- Motivation
- Theoretical aspects:
  - nuclear norm
  - low-rank matrix sensing
  - low-rank matrix completion
- efficient algorithm

The Netflix problem

The Netflix problem, or collaborative filtering

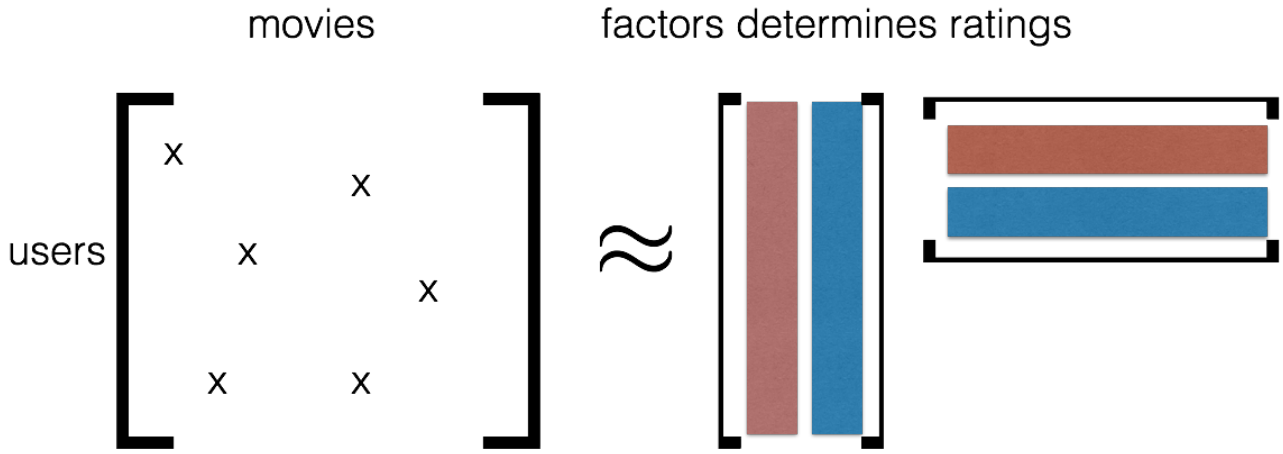
- How to estimate the missing ratings?



- About a million users, and 25,000 movies, with sparsely sampled ratings

Solution: low-rank matrix completion

- Matrix completion problem: consider $\boldsymbol{M} \in \mathbb{R}^{n_1 \times n_2}$ to represent the Netflix data set, we may model it through factorization:



- The rank r of $\boldsymbol{M}$ is much smaller than its dimension $r \ll \min\{n_1, n_2\}$.

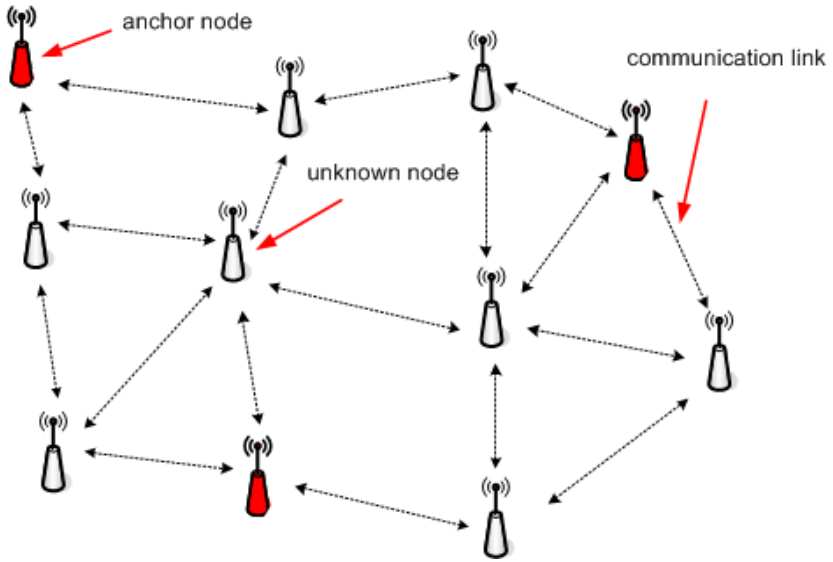
Sensor localization

- Given n points $\{\boldsymbol{x}_j\}_{j=1}^n \in \mathbb{R}^3$

- Observe partial information about distances:

$$M_{i,j} = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2$$

e.g. in wireless sensor network, each sensor can measure the distance to its neighbors, would like to globally locate all sensors.



Solution: low-rank matrix completion

- Write the matrix

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

  then

$$M_{i,j} = \boldsymbol{x}_i^T \boldsymbol{x}_i + \boldsymbol{x}_j^T \boldsymbol{x}_j - 2\boldsymbol{x}_i^T \boldsymbol{x}_j$$

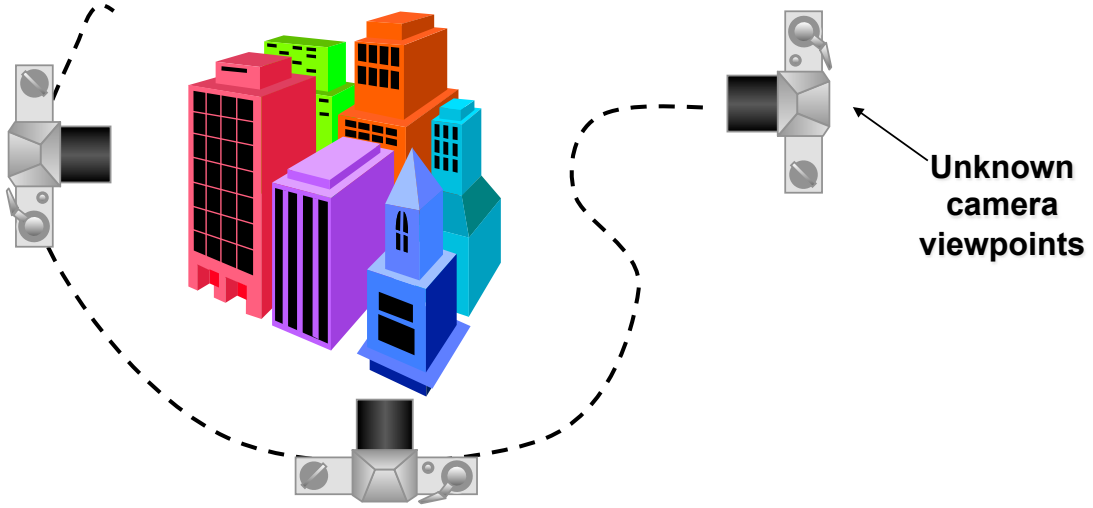
- Matrix completion problem: Let $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{X}^T$. The distance matrix $\boldsymbol{M} \in \mathbb{R}^{n \times n}$ between points can be written as

$$\boldsymbol{M} = \text{diag}(\boldsymbol{Y})\boldsymbol{e}^T + \boldsymbol{e}\text{diag}(\boldsymbol{Y})^T - 2\boldsymbol{Y}$$

- The rank r of $\boldsymbol{M}$ is much smaller than its dimension $r \ll n$.

Structure from motion

Structure from motion: reconstruct scene geometry and camera motion from multiple images.



In the pipeline of performing SFM, assume we've found a set of good feature points with their corresponding 2D locations in the images.

Structure from motion - continued

Tomasi and Kanade's factorization: Given n points $\boldsymbol{x}_{i,j}^T \in \mathbb{R}^2$ corresponding to the location of the ith point in the jth frame, define the matrix

$$\boldsymbol{M} = \begin{bmatrix} \boldsymbol{x}_{1,1} & \cdots & \boldsymbol{x}_{1,m} \\ \vdots & \ddots & \vdots \\ \boldsymbol{x}_{n,1} & \cdots & \boldsymbol{x}_{n,m} \end{bmatrix} \in \mathbb{R}^{n \times 2m}$$

In the absence of noise, this matrix admits a low-rank factorization:

$$\boldsymbol{M} = \underbrace{\begin{bmatrix} \boldsymbol{s}_1^T \\ \vdots \\ \boldsymbol{s}_n^T \end{bmatrix}}_{\text{3D structure matrix}} \underbrace{\begin{bmatrix} \boldsymbol{w}_1 & \boldsymbol{w}_2 & \cdots & \boldsymbol{w}_{2m} \end{bmatrix}}_{\text{camera motion matrix}}$$

where $\boldsymbol{s}_i \in \mathbb{R}^3$, which gives $\text{rank}(\boldsymbol{M}) = 3$.

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

Many more...

Many more applications:

- spatial-temporal data: low-rank due to correlations, e.g. MRI video, network traffic, etc..
- quantum space tomography
- linear system identification

Problem of interest: Can we recover the matrices of interest from "incomplete" observations, using efficient algorithms?

- the problem is ill-posed without additional constraints

Low-rank matrices

- Let $\boldsymbol{M} \in \mathbb{R}^{n \times n}$ (square case for simplicity) be a matrix of rank $r \ll n$.
- The Singular Value Decomposition (SVD) of $\boldsymbol{M}$ is given as

$$\boldsymbol{M} = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T$$

  where $\{\sigma_i\}_{i=1}^r$ are the singular values; and $\{\boldsymbol{u}_i\}_{i=1}^r$, $\{\boldsymbol{v}_i\}_{i=1}^r$ are the singular vectors.

- $\boldsymbol{M}$ has $(2n - r)r$ degrees of freedom.

Linear measurements of low-rank matrices

We make linear measurements of $\boldsymbol{M}$:

$$y_i = \langle \boldsymbol{A}_i, \boldsymbol{M} \rangle = \mathrm{Tr}(\boldsymbol{A}_i^T \boldsymbol{M}), \qquad i = 1, \dots m,$$

which can be written more concisely in an operator form:

$$\boldsymbol{y} = \mathcal{A}(\boldsymbol{M})$$

where $\mathcal{A} : \mathbb{R}^{n \times n} \mapsto \mathbb{R}^m$ denotes the measurement process. Its adjoint operator $\mathcal{A}^* : \mathbb{R}^m \mapsto \mathbb{R}^{n \times n}$ is defined as

$$\mathcal{A}^*(\boldsymbol{y}) = \sum_{i=1}^{m} y_i \boldsymbol{A}_i.$$

The problem of rank minimization:

$$\hat{\boldsymbol{M}} = \underset{\boldsymbol{X}}{\mathrm{argmin}}\, \mathsf{rank}(\boldsymbol{X}) \quad \text{s.t.} \quad \boldsymbol{y} = \mathcal{A}(\boldsymbol{X}).$$

Nuclear norm

Just as ℓ_1 norm provides a convex relaxation to cardinality minimization, we use the nuclear norm which give a convex relaxation to rank minimization.

Definition 1. *The nuclear norm of $\boldsymbol{X}$ is defined as*

$$\|\boldsymbol{X}\|_* = \sum_{i=1}^{n} \sigma_i(\boldsymbol{X})$$

where $\sigma_i(\boldsymbol{X})$ is the ith largest singular value of $\boldsymbol{X}$.

- Since the rank is $\sum_{i=1}^{n} 1(\sigma_i(\boldsymbol{X}) \neq 0)$, the nuclear norm can be thought as an ℓ_1 norm relaxation of the vector of singular values.

- This is a norm. Relationships between different norms:

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

- Tightest convex relaxation: $\{\boldsymbol{X} : \|\boldsymbol{X}\|_* \leq 1\}$ is the convex hull of rank-1 matrices obeying $\|\boldsymbol{x}\boldsymbol{y}^T\| \leq 1$.

Additivity of the nuclear norm

Lemma 1. *Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be matrices of the same dimensions. If $\boldsymbol{A}\boldsymbol{B}^T = 0$ and $\boldsymbol{A}^T\boldsymbol{B} = 0$, then $\|\boldsymbol{A} + \boldsymbol{B}\|_* = \|\boldsymbol{A}\|_* + \|\boldsymbol{B}\|_*$.*

Remark: this implies that, if the row and column spaces of $\boldsymbol{A}$ and $\boldsymbol{B}$ are orthogonal, then $\|\boldsymbol{A} + \boldsymbol{B}\|_* = \|\boldsymbol{A}\|_* + \|\boldsymbol{B}\|_*$.

This is similar to the 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 essentially all we need to get the proof of ℓ_1 minimization with RIP...

Compute the nuclear norm via SDP

Lemma 2.

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

This means we can compute the nuclear norm efficient via semidefinite programming (SDP).

Proof: on the blackboard.

Dual norm

Definition 2. *For a given norm $\|\cdot\|_{\mathcal{A}}$ in an inner product space $\langle\cdot,\cdot\rangle$, the dual norm is defined as*

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

By definition, this gives a general version of Cauchy-Schwarz inequality:

$$\langle \boldsymbol{X}, \boldsymbol{Y}\rangle \leq \|\boldsymbol{X}\|_{\mathcal{A}}\|\boldsymbol{Y}\|_{\mathcal{A}}^{\star}.$$

Examples:

- The dual norm of $\|\cdot\|_F$ is $\|\cdot\|_F$;
- The dual norm of $\|\cdot\|_1$ is $\|\cdot\|_\infty$;
- The dual norm of $\|\cdot\|_*$ is $\|\cdot\|$;

Summary

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.

Low-rank matrix recovery via nuclear norm minimization

- The rank minimization problem:

$$\hat{\boldsymbol{M}} = \underset{\boldsymbol{X}}{\operatorname{argmin}} \operatorname{rank}(\boldsymbol{X}) \quad \text{s.t.} \quad \boldsymbol{y} = \mathcal{A}(\boldsymbol{X}).$$

- We pose the following nuclear norm minimization algorithm:

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

which can be solved efficiently via SDP:

$$\hat{\boldsymbol{M}} = \underset{\boldsymbol{X},\boldsymbol{W}_1,\boldsymbol{W}_2}{\operatorname{argmin}} \frac{1}{2}\operatorname{Tr}(\boldsymbol{W}_1) + \frac{1}{2}\operatorname{Tr}(\boldsymbol{W}_2)$$

$$\text{s.t.} \quad \boldsymbol{y} = \mathcal{A}(\boldsymbol{X}), \quad \begin{bmatrix} \boldsymbol{W}_1 & \boldsymbol{X} \\ \boldsymbol{X}^T & \boldsymbol{W}_2 \end{bmatrix} \succeq 0.$$

Low-rank matrix sensing

- If $\mathcal{A}$ satisfies the restricted isometry property for low-rank matrices:

  Definition 3. *The operator $\mathcal{A}$ satisfies the RIP of rank-r, if for any rank-r matrix, we have*

  $$(1-\delta_r)\|\boldsymbol{X}\|_F^2 \leq \|\mathcal{A}(\boldsymbol{X})\|_F^2 \leq (1+\delta_r)\|\boldsymbol{X}\|_F^2$$

  for $0 \leq \delta_r \leq 1$.

- If $\{\boldsymbol{A}_i\}_{i=1}^m$ are composed of i.i.d. Gaussian entries, then it satisfies the matrix RIP of order r with high probability, as soon as $m \gtrsim nr$.

- This allows us to develop almost parallel results to compressed sensing.

Theoretical guarantees

Theorem 3. *If $\boldsymbol{A}$ satisfies the RIP of rank $4r$ with $\delta_{4r} \leq \sqrt{2} - 1$, then for all rank-$r$ matrices, the nuclear norm minimization algorithm recovers $\boldsymbol{M}$ exactly.*

Exact recovery from $O(nr)$ measurements!!

- For the noisy case,

$$\boldsymbol{y} = \mathcal{A}(\boldsymbol{M}) + \boldsymbol{w}$$

where $\boldsymbol{w}$ is composed of i.i.d. $\mathcal{N}(0, \sigma^2)$ entries. We could similarly propose the matrix LASSO algorithm:

$$\hat{\boldsymbol{M}} = \underset{\boldsymbol{X}}{\operatorname{argmin}} \frac{1}{2}\|\boldsymbol{y} - \mathcal{A}(\boldsymbol{X})\|_2^2 + \lambda\|\boldsymbol{X}\|_*,$$

where λ is a regularization parameter.

- If $\|\mathcal{A}^*(\boldsymbol{w})\| \leq \lambda/2$ and $\delta_{4r} < (3\sqrt{2}-1)/17$, then

$$\|\hat{\boldsymbol{M}} - \boldsymbol{M}\|_F \leq C\sqrt{r}\lambda$$

for some constant C. For the Gaussian case,

$$\|\mathcal{A}^*(\boldsymbol{w})\| \leq c_1\sqrt{n}\sigma := \lambda$$

for some large enough constant c_1 with probability at least $1 - 2e^{-cn}$.

- If $\boldsymbol{M}$ is an approximately low-rank matrix, we further have

$$\|\hat{\boldsymbol{M}} - \boldsymbol{M}\|_F \leq C_1\frac{\|\boldsymbol{M} - \boldsymbol{M}_r\|_*}{\sqrt{r}} + C_2\sqrt{nr}\sigma$$

with probability at least $1 - 2e^{-cn}$, in the Gaussian sampling case.

Low-rank matrix completion

In the matrix completion setting, we are given partial observations of the entries of $\boldsymbol{M}$, and wish to recover the missing entries.

- Denote $\Omega = \{(i,j) \in [n] \times [n]\}$ as the index set of observed entries.
- The observation can be written as

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

  where $Y_{ij} = M_{ij}$ if $(i,j) \in \Omega$ and $Y_{ij} = 0$ otherwise.

- Consider the following algorithm:

$$\min \|\boldsymbol{X}\|_* \quad \text{s.t.} \quad \boldsymbol{Y} = \mathcal{P}_\Omega(\boldsymbol{X})$$

- The observation operator doesn't satisfy RIP!

Which sampling pattern?

Consider a rank-one matrix $\boldsymbol{M} = \boldsymbol{x}\boldsymbol{y}^T$ with the following sampling pattern:

If single row (or column) is not sampled, recovery is not possible.

Fix the number of observed entries $m = |\Omega|$, would like to get performance bound that holds for almost all sampling patterns.

$\Longrightarrow$ We'll consider subset of m entries selected uniformly at random.

Which low-rank matrices can we recover?

Compare the following two rank-one matrices:

$$\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix} \quad \text{vs} \quad \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \vdots & & \\ 1 & 1 & 1 & \cdots & 1 \end{bmatrix} \quad \text{vs} \quad \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}$$

The middle one would be "easier" to complete.

Column and row spaces cannot be aligned with basis vectors.

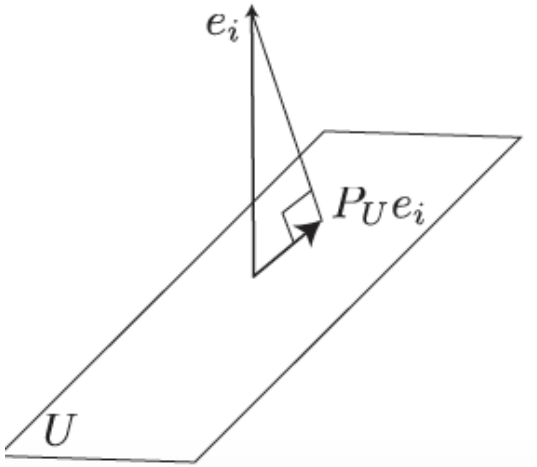
Coherence

Let $\boldsymbol{M}$ be a rank-r matrix with the SVD $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$, where $\boldsymbol{U}, \boldsymbol{V} \in \mathbb{R}^{n \times r}$.

Definition 4. [Coherence] *Smallest scalar μ obeying*

$$\max_{1 \le i \le n} \|\boldsymbol{U}^T \boldsymbol{e}_i\|_2^2 \le \mu \frac{r}{n}, \quad \max_{1 \le i \le n} \|\boldsymbol{V}^T \boldsymbol{e}_i\|_2^2 \le \mu \frac{r}{n},$$

where $\boldsymbol{e}_i$ is the ith standard basis vector.



We would like $\mu = O(1)$.

- Geometric condition: $\boldsymbol{U} = \text{colspan}(\boldsymbol{M})$
- Since $\sum_{i=1}^n \|\boldsymbol{U}^T \boldsymbol{e}_i\|_2^2 = r$, $\mu \ge 1$.
- If $\boldsymbol{e}_i \in \boldsymbol{U}$, $\mu = n/r$;
- If $\frac{1}{\sqrt{n}}\mathbf{1} = \boldsymbol{U}$, $\mu = 1$.

Information-theoretic lower bound

Theorem 4. [Candes and Tao, 2009] *No method can succeed with*

$$m \lesssim \mu \times nr \times \log n \approx \textit{dof} \times \mu \log n$$

Remarks:

- When $\mu = O(1)$, we need $m \lesssim nr \log n$.
- Need at least one observation /row and column – related to the coupon collector's problem: Suppose that there is an urn of n different coupons, from which coupons are being collected, equally likely, with replacement. How many trials do we need to collect all n coupons?
- The adjacency graph needs to be fully connected

Performance Guarantee

Theorem 5. [Chen, Gross, Recht, Candes and Tao, etc..] *There exists universal constant* $c_0, c_1, c_2 > 0$ *such that if*

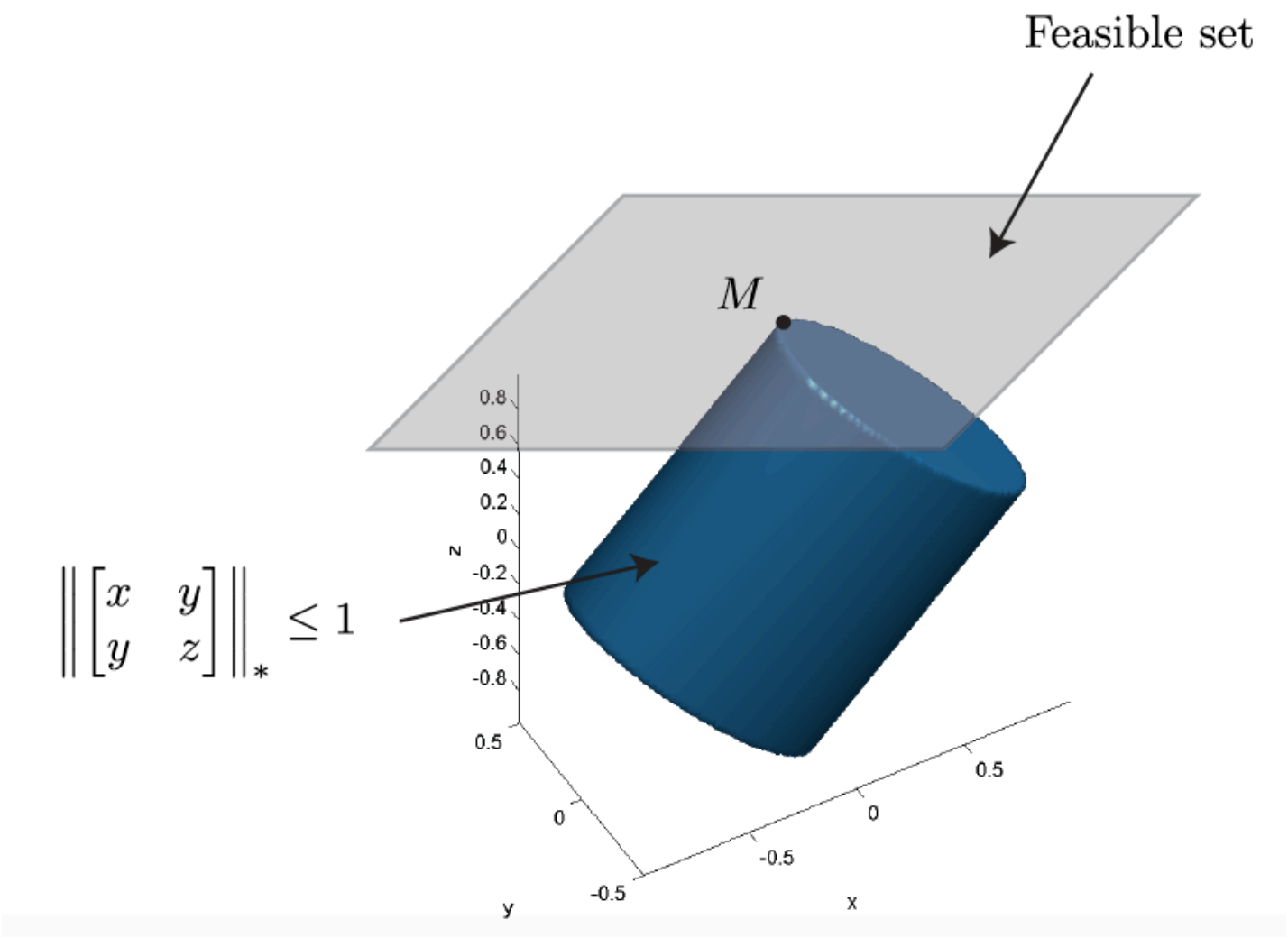
$$m \geq c_0 \mu n r \log^2 n,$$

then $\boldsymbol{M}$ *is the unique optimal solution to the nuclear norm minimization problem with probability at least* $1 - c_1 n^{-c_2}$.

Remark:

- This result is optimal up to a logarithmic factor in n. See [Chen, Incoherence-Optimal Matrix Completion].

Geometry


Feasible set
M
$\left\| \begin{bmatrix} x & y \\ y & z \end{bmatrix} \right\|_* \leq 1$
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
z
0.5
0
-0.5
y
0.5
0
-0.5
x

A few notations

- $\boldsymbol{P}_U$ is the orthogonal projection to the column space of $\boldsymbol{M}$;
- $\boldsymbol{P}_V$ is the orthogonal projection to the row space of $\boldsymbol{M}$;
- Let T be the span of matrices of the form:

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

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

$$\mathcal{P}_T(\boldsymbol{X}) = \boldsymbol{P}_U\boldsymbol{X} + \boldsymbol{X}\boldsymbol{P}_V - \boldsymbol{P}_U\boldsymbol{X}\boldsymbol{P}_V$$

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

$$\mathcal{P}_{T^\perp}(\boldsymbol{X}) = (\boldsymbol{I} - \boldsymbol{P}_U)\boldsymbol{X}(\boldsymbol{I} - \boldsymbol{P}_V)$$

Subgradient of the nuclear norm

The subgradient of $\|\cdot\|_*$ at $\boldsymbol{M}$ can be written as

$$\partial\|\boldsymbol{M}\|_* = \left\{\boldsymbol{U}\boldsymbol{V}^T + \boldsymbol{W} : \quad \mathcal{P}_T(\boldsymbol{W}) = 0,\ \|\boldsymbol{W}\| \leq 1\right\}$$

$\boldsymbol{Z} \in \partial\|\boldsymbol{M}\|_*$ if and only if

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

The subgradient doesn't depend on the singular values of $\boldsymbol{M}$.

Basic consequence of incoherence

For any $(i, j) \in [n] \times [n]$:

$$\left\| \mathcal{P}_T(\boldsymbol{e}_i \boldsymbol{e}_j^T) \right\|_F^2 \leq \frac{2\mu r}{n}.$$

The sampling basis is incoherent to the tangent space T.

Sampling with replacement

It turns out it is easier to use a sampling with replacement model, where we assume each observed entry (i_k, j_k), $k = 1, \ldots, m$ is **i.i.d.** observed uniformly at random from $[n] \times [n]$.

This is much easier to analyze, however it is different from the sampling without replacement model stated earlier because we may sample the same entry several times.

Proposition 1. *The probability that the nuclear norm heuristic fails when the set of observed entries is sampled uniformly from the collection of sets of size m is less than or equal to the probability that the heuristic fails when m entries are sampled independently with replacement.*

Bound the number of repetitions

Proposition 2. *With probability at least* $1 - n^{2-2\beta}$*, the maximum number of repetitions of any entry in* Ω *is less than* $\frac{8}{3}\beta \log(n)$ *for* $n \geq 9$ *and* $\beta > 1$*.*

Define the operator

$$\mathcal{R}_\Omega(\boldsymbol{X}) = \sum_{k=1}^{m} \left\langle \boldsymbol{X}, \boldsymbol{e}_{i_k} \boldsymbol{e}_{j_k}^T \right\rangle \boldsymbol{e}_{i_k} \boldsymbol{e}_{j_k}^T = \sum_{k=1}^{m} X_{i_k, j_k} \boldsymbol{e}_{i_k} \boldsymbol{e}_{j_k}^T$$

where (i_k, j_k) is uniformly drawn from $[n] \times [n]$. From the above proposition, we have

$$\|\mathcal{R}_\Omega\| \leq \frac{8}{3}\beta \log(n)$$

with probability at least $1 - n^{2-2\beta}$.

Optimality condition

Proposition 3. [Exact Dual Certificate] *M is the unique minimizer of the nuclear norm minimization problem if the following holds:*

- *the sampling operator $\mathcal{P}_\Omega$ restricted to elements in T is injective;*
- *there exists $\boldsymbol{Y}$ supported on Ω such that $\boldsymbol{Y} \in \partial\|\boldsymbol{M}\|_*$, i.e.*

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

The first equality constraint is not easy to satisfy, see [Candes and Tao, 2009].

Dual certificate

Under a stronger injectivity requirement, we can relax the second requirement a bit, which much simplifies the analysis.

Proposition 4. [Inexact Dual Certificate] *Suppose that*

$$\frac{n^2}{m}\left\|\mathcal{P}_T\mathcal{R}_\Omega\mathcal{P}_T - \frac{m}{n^2}\mathcal{P}_T\right\| \le \frac{1}{2},$$

and there exists $\boldsymbol{Y}$ *supported on* Ω *such that*

$$\left\|\mathcal{P}_T(\boldsymbol{Y}) - \boldsymbol{U}\boldsymbol{V}^T\right\|_F \le \sqrt{\frac{r}{2n}}, \quad \|\mathcal{P}_{T^\perp}(\boldsymbol{Y})\| < \frac{1}{2},$$

then $\boldsymbol{M}$ *is the unique minimizer of the nuclear norm minimization problem if the following holds:*

Injectivity of $\mathcal{R}_\Omega$ on T

Proposition 5. *For all* $\beta > 1$,

$$\frac{n^2}{m}\left\|\mathcal{P}_T\mathcal{R}_\Omega\mathcal{P}_T - \frac{m}{n^2}\mathcal{P}_T\right\| \le \sqrt{\frac{32\beta\mu_0 nr\log n}{3m}}$$

with probability at least $1-2n^{2-2\beta}$ *provided* $m \ge \frac{32}{3}\beta\mu_0 nr\log n$.

Remark: Provided $\sqrt{\frac{32\beta\mu_0 nr\log n}{3m}} \le \frac{1}{2}$, i.e.

$$m \ge \frac{128\beta\mu_0 nr\log n}{3}$$

we have with probability at least $1-2n^{2-2\beta}$,

$$\frac{n^2}{m}\left\|\mathcal{P}_T\mathcal{R}_\Omega\mathcal{P}_T - \frac{m}{n^2}\mathcal{P}_T\right\| \le \frac{1}{2}.$$

Constructing the dual certificate

We introduce the clever golfing scheme proposed by David Gross.

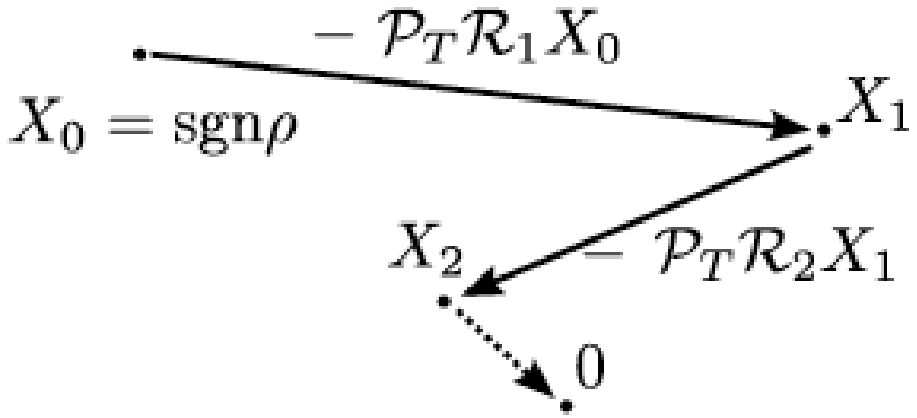


Fig. 3. Caricature of the "golfing scheme" used to construct the certificate. In the ith step, X_{i-1} designates the vector we aim to represent. The approximation of X_{i-1} actually obtained is $\mathcal{P}_T\mathcal{R}_iX_{i-1}$. The distance of the new goal $X_i = X_{i-1} - \mathcal{P}_T\mathcal{R}_iX_{i-1}$ to the origin is guaranteed to be only half the previous one. The sequence X_i thus converges exponentially fast to the origin.

Computational aspect using FISTA

Recall the FISTA algorithm we discussed to solve

$$\hat{\boldsymbol{M}} = \underset{\boldsymbol{X}}{\operatorname{argmin}} \frac{1}{2}\|\boldsymbol{y} - \mathcal{A}(\boldsymbol{X})\|_2^2 + \lambda\|\boldsymbol{X}\|_*$$

- Initialization: $\boldsymbol{x}_0 = \boldsymbol{x}_{-1} \in \mathbb{R}^n$, $\theta_0 = 1$,
- For $k = 1, 2, \ldots,$

$$\theta_k = \frac{1 + \sqrt{1 + 4\theta_{k-1}^2}}{2}$$

$$\boldsymbol{Y}_k = \boldsymbol{X}_{k-1} + \left(\frac{\theta_{k-1} - 1}{\theta_k}\right)(\boldsymbol{X}_{k-1} - \boldsymbol{X}_{k-2})$$

$$\boldsymbol{X}_k = \operatorname{prox}_{t_k\lambda\|\boldsymbol{X}\|_*}(\boldsymbol{Y}_k - t_k\mathcal{A}^*(\mathcal{A}(\boldsymbol{Y}_k) - \boldsymbol{y}))$$

- What is the proximal operator for $\|\cdot\|_*$?

Proximal operator for $\|\cdot\|_*$

Proposition 6.

$$\text{prox}_{t_k\lambda\|\cdot\|_*}(\boldsymbol{X}) = \underset{\boldsymbol{Z}}{\text{argmin}} \left\{ \frac{1}{2}\|\boldsymbol{Z} - \boldsymbol{X}\|_2^2 + t_k\lambda\|\boldsymbol{Z}\|_* \right\} = \mathcal{T}_{t_k\lambda}(\boldsymbol{X})$$

where

$$\mathcal{T}_\tau(\boldsymbol{X}) = \boldsymbol{U}\mathcal{T}_\tau(\boldsymbol{\Sigma})\boldsymbol{V}^T,$$

where the SVD of $\boldsymbol{X}$ *is given as* $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$, $\boldsymbol{\Sigma} = \mathit{diag}(\{\sigma_i\})$, *and*

$$\mathcal{T}_\tau(\boldsymbol{\Sigma}) = \mathit{diag}(\{(\sigma_i - \tau)_+\}).$$