

Low Rank Matrix Completion

Ref

- https://users.ece.cmu.edu/~yuejiec/papers/SPM_lrmc_final.pdf
- <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8759045>
- <https://arxiv.org/pdf/0805.4471.pdf>
- http://math.oit.edu/~watermang/math_341/341_ch8/F13_341_book_sec_8-4.pdf

Rank and Low Rank

In linear algebra, the **rank** of a matrix A is the dimension of the vector space generated (or spanned) by its columns.

This corresponds to the maximal number of linearly independent columns of A .

The matrix

$$\begin{bmatrix} 1 & 0 & 1 \\ -2 & -3 & 1 \\ 3 & 3 & 0 \end{bmatrix}$$

has rank 2: the first two columns are linearly independent, so the rank is at least 2, but since the third is a linear combination of the first two (the second subtracted from the first), the three columns are linearly dependent so the rank must be less than 3.

Why Low Rank?

- **Imagine one observes a small subset of entries in a large matrix and aims to recover the entire matrix. Without a priori knowledge of the matrix, this problem is highly ill-posed.**
- Fortunately, data matrices often exhibit low- dimensional structures that can be used effectively to regularize the solution space.
- Correspondingly, the data matrix can be modeled as a low-rank matrix, at least approximately. *Is it possible to complete a partially observed matrix if its rank, i.e., its maximum number of linearly-independent row or column vectors, is small?*

How useful?

A low rank approximation can be used to make filtering and statistics either computationally feasible or more efficient.


In machine learning, low rank approximations to data tables are often employed to:

- **impute missing data**
- **denoise noisy data**
- **perform feature extraction**
- **develop algorithms in recommender systems**

Example of Low Rank Matrix: Rating matrix

Rating matrix in the recommendation systems:

- users expressing similar ratings on multiple products tend to have the same interest for the new product
- **columns associated with users sharing the same interest are highly likely to be the same, resulting in the low rank structure.**
- users are recommended to submit the feedback in a form of rating number, e.g., 1 to 5 for the purchased product.
- However, users often do not want to leave a feedback and thus the rating matrix will have many missing entries.

						...
Alice	1			4		
Bob		2	5			
Carol			4	5		
Dave	5				4	
⋮						

Example of Low Rank Matrix: Phase Retrieval

Phase retrieval: The problem to recover a signal not necessarily sparse from the magnitude of its observation is referred to as the phase retrieval. Phase retrieval is an important problem in X-ray crystallography and quantum mechanics since only the magnitude of

the Fourier transform is measured in these applications [5]. Suppose the unknown time-domain signal $\mathbf{m} = [m_0 \cdots m_{n-1}]$ is acquired in a form of the measured magnitude of the Fourier transform. That is,

$$|z_\omega| = \frac{1}{\sqrt{n}} \left| \sum_{t=0}^{n-1} m_t e^{-j2\pi\omega t/n} \right|, \quad \omega \in \Omega,$$

where Ω is the set of sampled frequencies. Further, let

$$\mathbf{f}_\omega = \frac{1}{\sqrt{n}} [1 \ e^{-j2\pi\omega/n} \ \dots \ e^{-j2\pi\omega(n-1)/n}]^H, \quad (3)$$

$\mathbf{M} = \mathbf{m}\mathbf{m}^H$ where \mathbf{m}^H is the conjugate transpose of \mathbf{m} . Then, (2) can be rewritten as

$$|z_\omega|^2 = |\langle \mathbf{f}_\omega, \mathbf{m} \rangle|^2 \quad (4)$$

$$= \text{tr}(\mathbf{f}_\omega^H \mathbf{m}\mathbf{m}^H \mathbf{f}_\omega) \quad (5)$$

$$= \text{tr}(\mathbf{m}\mathbf{m}^H \mathbf{f}_\omega \mathbf{f}_\omega^H) \quad (6)$$

$$= \langle \mathbf{M}, \mathbf{F}_\omega \rangle, \quad (7)$$

where $\mathbf{F}_\omega = \mathbf{f}_\omega \mathbf{f}_\omega^H$ is the rank-1 matrix of the waveform \mathbf{f}_ω . Using this simple transform, we can express the quadratic magnitude $|z_\omega|^2$ as linear measurement of \mathbf{M} . In essence,

the phase retrieval problem can be converted to the problem to reconstruct the rank-1 matrix \mathbf{M} in the positive **semi-definite** (PSD) cone³ [5]:

$$\begin{aligned} & \min_{\mathbf{X}} \quad \text{rank}(\mathbf{X}) \\ & \text{subject to} \quad \langle \mathbf{M}, \mathbf{F}_\omega \rangle = |z_\omega|^2, \ \omega \in \Omega \\ & \quad \mathbf{X} \succeq 0. \end{aligned} \quad (8)$$

Positive semi-definite*

$$\mathbf{X} \succeq 0.$$

- a **symmetric matrix** with **real** entries is **positive-definite** if the real number λ is positive for every nonzero real **column vector** \mathbf{v} .
- **Positive semi-definite** matrices are defined similarly, except that the scalars λ and μ are required to be positive *or zero* (that is, nonnegative). **Negative-definite** and **negative semi-definite** matrices are defined analogously. A matrix that is not positive semi-definite and not negative semi-definite is sometimes called **indefinite**.
- if M, N are positive semidefinite, then $\alpha M + \beta N$ is also positive semidefinite for positive α, β . Hence, the set of positive semidefinite matrices is a convex cone in $\mathbb{R}^{\frac{n(n+1)}{2}}$.

* https://www.cse.iitk.ac.in/users/rmittal/prev_course/s14/notes/lec12.pdf

Notations

- For a vector $\mathbf{a} \in \mathbb{R}^n$, $\text{diag}(\mathbf{a}) \in \mathbb{R}^{n \times n}$ is the diagonal matrix formed by \mathbf{a} .
- For a matrix $\mathbf{A} \in \mathbb{R}^{n_1 \times n_2}$, $\mathbf{a}_i \in \mathbb{R}^{n_1}$ is the i -th column of \mathbf{A} .
- $\text{rank}(\mathbf{A})$ is the rank of \mathbf{A} .
- $\mathbf{A}^T \in \mathbb{R}^{n_2 \times n_1}$ is the transpose of \mathbf{A} .
- For $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n_1 \times n_2}$, $\langle \mathbf{A}, \mathbf{B} \rangle = \text{tr}(\mathbf{A}^T \mathbf{B})$ and $\mathbf{A} \odot \mathbf{B}$ are the inner product and the Hadamard product (or element-wise multiplication) of two matrices \mathbf{A} and \mathbf{B} , respectively, where $\text{tr}(\cdot)$ denotes the trace operator.
- $\|\mathbf{A}\|$, $\|\mathbf{A}\|_*$, and $\|\mathbf{A}\|_F$ stand for the spectral norm (i.e., the largest singular value), the nuclear norm (i.e., the sum of singular values), and the Frobenius norm of \mathbf{A} , respectively.
- $\sigma_i(\mathbf{A})$ is the i -th largest singular value of \mathbf{A} .
- $\mathbf{0}_{d_1 \times d_2}$ and $\mathbf{1}_{d_1 \times d_2}$ are $(d_1 \times d_2)$ -dimensional matrices with entries being zero and one, respectively.
- \mathbf{I}_d is the d -dimensional identity matrix.
- If \mathbf{A} is a square matrix (i.e., $n_1 = n_2 = n$), $\text{diag}(\mathbf{A}) \in \mathbb{R}^n$ is the vector formed by the diagonal entries of \mathbf{A} .
- $\text{vec}(\mathbf{X})$ is the vectorization of \mathbf{X} .

Example of Low Rank Matrix: IoT network, sensor nodes

- Internet of things (IoT) in healthcare, automatic metering, environmental monitoring (temperature, pressure, moisture), and surveillance.
- Action in IoT networks, such as fire alarm, energy transfer, emergency request, is made primarily on the data center, data center should figure out the location information of whole devices in the networks.
- power outage of a sensor node or the limitation of radio communication range (see Fig. 1), only small number of distance information is available at the data center. Also, in the vehicular networks,
- it is not easy to measure the distance of all adjacent vehicles when a vehicle is located at the dead zone.

Observed Euclidean distance matrix is:

$$\mathbf{M}_o = \begin{bmatrix} 0 & d_{12}^2 & d_{13}^2 & ? & ? \\ d_{21}^2 & 0 & ? & ? & ? \\ d_{31}^2 & ? & 0 & d_{34}^2 & d_{35}^2 \\ ? & ? & d_{43}^2 & 0 & d_{45}^2 \\ ? & ? & d_{53}^2 & d_{54}^2 & 0 \end{bmatrix},$$

where d_{ij} is the pairwise distance between two sensor nodes i and j . Since the rank of Euclidean distance matrix \mathbf{M} is at most $k+2$ in the k -dimensional Euclidean space ($k = 2$ or $k = 3$) [3], [4], the problem to reconstruct \mathbf{M} can be well-modeled as the LRMC problem.

Image compression and restoration:

- When there is dirt or scribble in a two-dimensional image one simple solution is to replace the contaminated pixels with the interpolated version of adjacent pixels.
- Approximate an image to the low-rank matrix without perceptible loss of quality.
- By using clean (uncontaminated) pixels as observed entries, an original image can be recovered via the low-rank matrix completion.

Major benefit

One major benefit of the low-rank matrix is that the essential information, expressed in terms of degree of freedom, in a matrix is much smaller than the total number of entries.

Therefore, even though the number of observed entries is small, we still have a good chance to recover the whole matrix.

When there is no restriction on the rank of a matrix, the problem to recover unknown entries of a matrix from partial observed entries is ill-posed.

This is because any value can be assigned to unknown entries, which in turn means that there are infinite number of matrices that agree with the observed entries.

Consider the following 2×2 matrix with one unknown entry marked ?

$$\mathbf{M} = \begin{bmatrix} 1 & 5 \\ 2 & ? \end{bmatrix}$$

Fundamental principle to recover a large dimensional matrix using low-rank constraint

If M is a full rank, i.e., the rank of M is two, then any value except 10 can be assigned to $?$.

Whereas, if M is a low-rank matrix (the rank is one in this trivial example), two columns differ by only a constant and hence unknown element $?$ can be easily determined using a linear relationship between two columns ($? = 10$).

Objective

observe a small subset of entries in a large matrix and aims to recover the entire matrix.

Is it possible to complete a partially observed matrix if its rank, i.e., its maximum number of linearly-independent row or column vectors, is small?

Basic Concepts

Let $M \in \mathbb{R}^{n_1 \times n_2}$ be a rank- r matrix, whose thin Singular Value Decomposition (SVD) is given as

$$M = U \Sigma V^\top, \quad (1)$$

where $U \in \mathbb{R}^{n_1 \times r}$, $V \in \mathbb{R}^{n_2 \times r}$ are composed of orthonormal columns, and Σ is an r -dimensional diagonal matrix

with the singular values arranged in a non-increasing order, i.e. $\sigma_1 \geq \dots \geq \sigma_r > 0$.

Goal

Assume we are given partial observations of M over an index set $\Omega \subset \{1, 2, \dots, n_1\} \times \{1, 2, \dots, n_2\}$. To concisely put it, define the observation operator $\mathcal{P}_\Omega : \mathbb{R}^{n_1 \times n_2} \rightarrow \mathbb{R}^{n_1 \times n_2}$ as

$$[\mathcal{P}_\Omega(M)]_{ij} = \begin{cases} M_{ij}, & (i, j) \in \Omega \\ 0, & \text{otherwise} \end{cases}.$$

Our goal is to recover M from $\mathcal{P}_\Omega(M)$, when the number of observation $m = |\Omega| \ll n_1 n_2$ is much smaller than the number of entries in M , under the assumption that M is low-rank, i.e. $r \ll \min\{n_1, n_2\}$. For notational simplicity in the sequel, let $n = \max\{n_1, n_2\}$.

Which low-rank matrices can we complete?

Two main properties of matrix

1) Sparsity

2) Incoherence

→ Sparsity indicates an accurate recovery of undersample matrix even when the observed entries are small.

→ Incoherence indicates nonzero entries should be spread out widely for efficient recovery.

What kind of low-rank matrices can we complete?

1. Sparsity:

$$M_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}, \quad M_2 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

- The matrix M_1 is more difficult to complete, since most of its entries are zero, and we need to collect more measurements to make sure enough mass comes from its nonzero entries.
- In contrast, the mass of M_2 is more uniformly distributed across all entries, making it easier to propagate information from one entry to another.

What kind of low-rank matrices can we complete?

2.Coherence:

A low-rank matrix is easier to complete if its energy spreads evenly across different coordinates.

This property is captured by the notion of coherence, which measures the alignment between the column/row spaces of the low-rank matrix with standard basis vectors

Cohrence

Considers following two matrices in $\mathbb{R}^{n \times n}$:

$$M_1 = \begin{bmatrix} 1 & 1 & \dots & 0 \\ 1 & 1 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix} \rightarrow \text{Rank } 1$$

$$M_2 = \begin{bmatrix} 1 & 1 & \dots & 1 \\ & 1 & 1 & \\ & \vdots & & \\ & 1 & & 1 \end{bmatrix} \text{ Rank } 1$$

for $M_1 \rightarrow$ only 4 entries are non zero.
and observed.

The coherence :

$$\mu(V) = \frac{n}{2} \max_{1 \leq i \leq n} \|P_V e_i\|^2$$

[e_i = standard basis .

P_V = projection onto the range space of V]

[Standard basis : Standard basis / natural basis / canonical basis in \mathbb{R}^n or \mathbb{C}^n are set of vectors whose components are all zero except one .

eg. $e_x = (1, 0)$ or $e_y = (0, 1)$]

• Projection P_V :

A projection on a vector space V is a linear operator $P : V \rightarrow V$, such that $P^2 = P$.

Revisit Coherence definition:

$$\mu(U) = \frac{n}{k} \max_{1 \leq i \leq n} \|P_U e_i\|^2.$$

$$P_U = UU^T \left[\begin{array}{l} \text{if } U \text{ is of orthonormal} \\ \text{columns} \end{array} \right]$$

$P_U = U(U^T U)^{-1} U^T$

 \rightarrow General formula

for orthonormal matrix:

[each column vector has length 1 and orthogonal to all other column vectors]

$$\boxed{U^T = U^{-1}}$$

Then, $P_U = U U^T$

Orthogonal projection :

point $(x, y, z) \in \mathbb{R}^3 \longrightarrow \text{point}(x, y, 0)$
 \downarrow
 $(x, y) \rightarrow \text{plane}$

$$P \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} x \\ y \\ 0 \end{bmatrix}, \text{ where } P = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

- range space : column space of a matrix.
 \rightarrow set of all possible linear combinations of column vectors.

Measure the concentration :

→ Matrix has two dimensional structure
↳ both row and column concentration

→ Check concentration of left and right singular vectors

Left singular vector of $A \in \mathbb{R}^n$ (Eigen vectors of $A A^T$) [matrix A]
Right singular vector of $A \in \mathbb{R}^n$ (Eigen vectors of $A^T A$)

From SVD of a matrix M :

$$\begin{aligned} M &= U \Sigma V^T \\ &= \sum_{i=1}^k \sigma_i u_i v_i^T \end{aligned}$$

$$M = U \Sigma V^T$$

$$= \sum_{i=1}^k \sigma_i u_i v_i^T$$

$$U = \begin{bmatrix} u_1 & \dots & u_k \end{bmatrix}$$

$$V = \begin{bmatrix} v_1 & \dots & v_k \end{bmatrix}$$

→ constructed by
left and
right vectors.

Σ : diagonal matrix whose entries are σ_i

Let's say: a standard basis vectors
 e_i is formed as $e_1 = [1 \ 0 \ 0 \ \dots \ 0]^T$.
 spanned by $[u_1 \ \dots \ u_k]$.

⇒ Non-zero value is only concentrated
 in first row. Hence, the first row can't
 be inferred just by sampling other rows.

Example:

Case 1

$$M_1 = \begin{bmatrix} 2 & 1 & 0 \\ 2 & 1 & 0 \\ 2 & 1 & 0 \end{bmatrix}$$

$$M_2 = \begin{bmatrix} 2 & 2 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

SVD of M : $M = \sigma_1 u_1 v_1^T$

$$M_1 = 3.8730 \begin{bmatrix} -0.5774 \\ -0.5774 \\ -0.5774 \end{bmatrix} \begin{bmatrix} -0.8944 & -0.4472 \end{bmatrix}$$

$$M_2 = 3.8417 \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} 0.8018 & 0.5345 & 0.2673 \end{bmatrix}$$

for M_2 :

$$U = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}^T \Rightarrow$$

[Standard basis e_1 on the space spanned by U , others are orthogonal]

$$M_1 = 3.8730 \begin{bmatrix} -0.5774 \\ -0.5774 \\ -0.5774 \end{bmatrix} \begin{bmatrix} -0.8944 & -0.4472 \end{bmatrix}$$

$$M_2 = 3.8417 \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} 0.8018 & 0.5345 & 0.2673 \end{bmatrix}$$

for M_2 :

$$U = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}^T \Rightarrow \begin{array}{l} \text{Standard basis } e_1 \text{ on the} \\ \text{space spanned by } U, \text{ others} \\ \text{are orthogonal} \end{array}$$

$$\|P_U e_1\|_2 = 1 \quad \|P_U e_2\|_2 = 0 \quad \|P_U e_3\|_2 = 0$$

$$\mu(U) = 3. \rightarrow \underline{\text{maximum coherence}}$$

$$\text{for } M_1: P_U = U \cdot U^T = \frac{1}{3} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

$$\text{Thus, } \|P_U e_1\|_2^2 = \|P_U e_2\|_2^2 = \|P_U e_3\|_2^2 = \frac{1}{3}$$

→ Both $\mu(u)$ & $\mu(v)$ need to be computed to check the concentration in horizontal and vertical directions.

What are minimum and maximum values of $\mu(u)$?

$$1 \leq \mu(v) \leq \frac{n}{r}$$

Coherence

Measures the alignment between the column/row spaces of the low-rank matrix with standard basis vectors.

For a matrix $U \in \mathbb{R}^{n_1 \times r}$ with orthonormal columns, let P_U be the orthogonal projection onto the column space of U . The coherence parameter of U is defined as

$$\mu(U) = \frac{n_1}{r} \max_{1 \leq i \leq n_1} \|P_U e_i\|_2^2 = \frac{n_1}{r} \max_{1 \leq i \leq n_1} \|U^\top e_i\|_2^2, \quad (2)$$

where e_i is the i th standard basis vector. Fig. 1 provides a geometric illustration of the coherence parameter $\mu(U)$.

For a low-matrix M whose SVD is given in (1), the coherence of M is defined as

$$\mu = \max\{\mu(U), \mu(V)\}. \quad (3)$$

Notably, the coherence μ is determined by the the singular vectors of M and independent of its singular values. Since $1 \leq \mu(U) \leq n_1/r$ and $1 \leq \mu(V) \leq n_2/r$, we have $1 \leq \mu \leq n/r$. In the earlier example, the coherence of M_1 matches the upper bound n/r , while the coherence of M_2 matches the lower bound 1. The smaller μ is, the easier it is to complete the matrix.

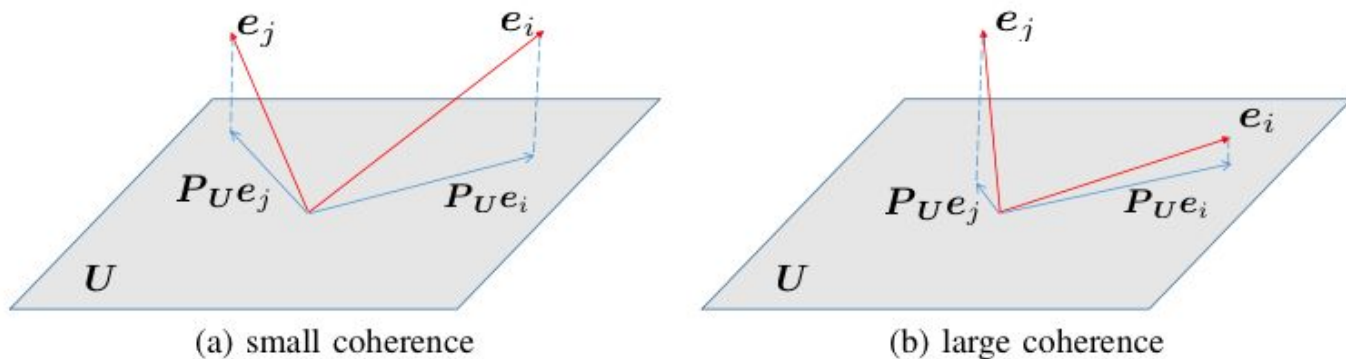


Fig. 1. Illustration of the coherence parameter $\mu(U)$. $\mu(U)$ is small when all the standard basis vectors e_i have approximately the same projections onto the subspace U , as shown in (a); and $\mu(U)$ is large if U is too aligned with certain standard basis vector, as shown in (b).

What kind of low-rank matrices can we complete?

How many minimum observations are required?

- Is Degree of Freedom sufficient?

Degrees of freedom

Degrees of freedom refers to the maximum number of logically independent values, which are values that have the freedom to vary.

Example 1: Consider a data sample consisting of five positive integers. The values of the five integers must have an average of six. If four of the items within the data set are {3, 8, 5, and 4}, the fifth number must be 10. Because the first four numbers can be chosen at random, the degrees of freedom is four.

The “degrees of freedom” of M is $(n_1 + n_2 - r)r$, which is the total number of parameters we need to uniquely specify M .

For $n \times n$ matrix, $\text{DOF} = 2nr - r^2$

Sparsity

eg:

$$M = \begin{bmatrix} 1 & 3 & 5 & 7 \\ 2 & 6 & 10 & 14 \\ 3 & 9 & 15 & 21 \\ 4 & 12 & 20 & 28 \end{bmatrix} \rightarrow \text{Rank} - 1$$

→ Observe one column and one row,
then determine the rest.

Degree of freedom (DOF)

- Number of freely chosen variables in the matrix.
- DOF of $M = 4+4-1 = 7$

Lemma : DOF of a $n \times n$ matrix with rank r is $2nr - r^2$. DOF of $n_1 \times n_2$ matrix is $(n_1 + n_2)r - r^2$.

Proof : rank of matrix $= r$.
 \rightarrow freely choose values of all entries of r columns
 \Rightarrow nr values to construct $m_1 \dots m_r$ independent columns.

Now, $(n-r)$ columns is expressed as linear combinations of r columns:

$$m_{r+1} = \alpha_1 m_1 + \alpha_2 m_2 - \dots + \alpha_r m_r$$

\vdots
 $m_n = \dots$
 $(\alpha_1, \alpha_2, \dots, \alpha_r)$ can be freely chosen for each of them.

$$\therefore \text{Total DOF} = (n-r)r + nr \\ = 2nr - r^2$$

\Rightarrow If n is very large and r is small, essential information in matrix is in $O(n)$.

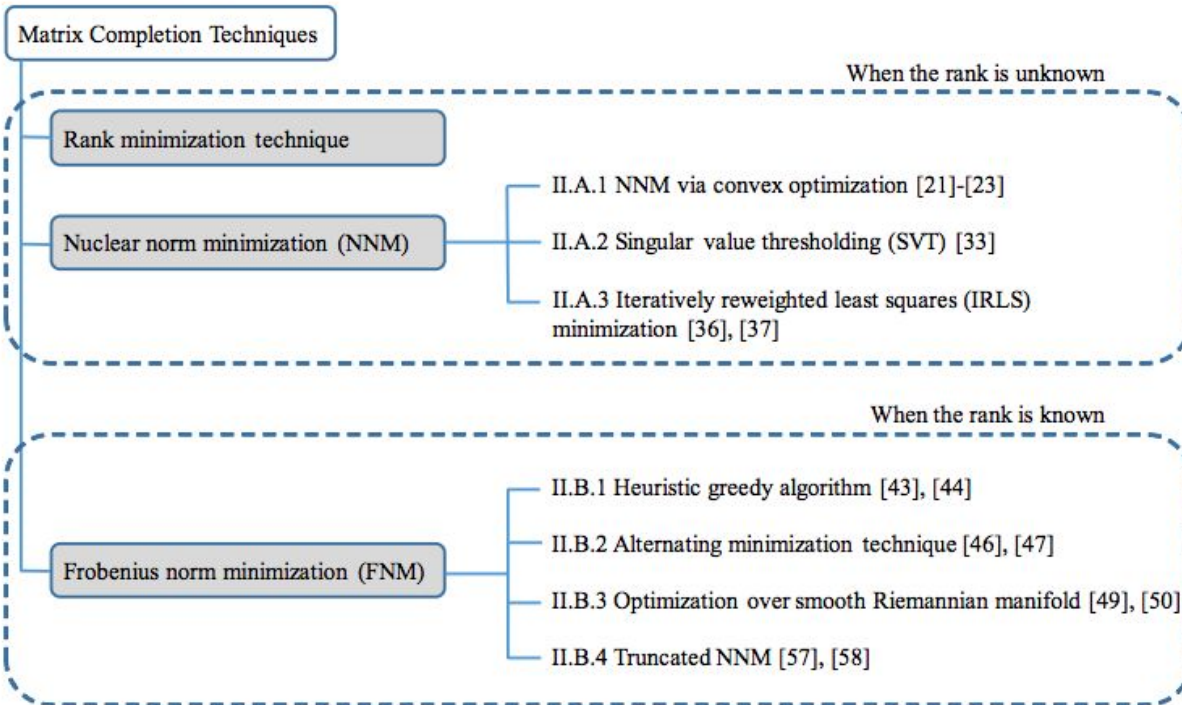
Which observation patterns can we handle?

Low-rank matrix completion can still be hopeless even when most of the entries are observed.

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

- The last column of the matrix cannot be recovered since it can lie anywhere in the column space of the low-rank matrix.
- Therefore, we require at least r observations per column/row.

Low Rank Matrix Completion techniques



How to recover a low-rank matrix from partial observations?

The desired low-rank matrix M can be recovered by solving the rank minimization problem:

$$\begin{array}{ll} \min_{\mathbf{X}} & \text{rank}(\mathbf{X}) \\ \text{subject to} & x_{ij} = m_{ij}, \quad (i, j) \in \Omega, \end{array}$$

where Ω is the index set of observed entries (e.g., $\Omega = \{(1, 1), (1, 2), (2, 1)\}$)

Alternative representation

The sampling operation

$P_\Omega(\mathbf{A})$ of a matrix \mathbf{A} is defined as

$$[P_\Omega(\mathbf{A})]_{ij} = \begin{cases} a_{ij} & \text{if } (i, j) \in \Omega \\ 0 & \text{otherwise.} \end{cases}$$

Using this sampling operator,

$$\begin{aligned} & \min_{\mathbf{X}} \quad \text{rank}(\mathbf{X}) \\ & \text{subject to} \quad P_\Omega(\mathbf{X}) = P_\Omega(\mathbf{M}). \end{aligned}$$

Rank minimization problem is the combinatorial search

we first assume that $\text{rank}(\mathbf{M}) = 1$. Then, any two columns of \mathbf{M} are linearly dependent and thus we have the system of expressions $\mathbf{m}_i = \alpha_{i,j}\mathbf{m}_j$ for some $\alpha_{i,j} \in \mathbb{R}$. If the system has no solution for the rank-one assumption, then we move to the next assumption of $\text{rank}(\mathbf{M}) = 2$. In this case, we solve the new system of expressions $\mathbf{m}_i = \alpha_{i,j}\mathbf{m}_j + \alpha_{i,k}\mathbf{m}_k$. This procedure is repeated until the solution is found. Clearly, the combinatorial search strategy would not be feasible for most practical scenarios since it has an exponential complexity in the problem size [76]. For example, when \mathbf{M} is an $n \times n$ matrix, it can be shown that the number of the system expressions to be solved is $\mathcal{O}(n2^n)$.

Matrix completion via convex optimization

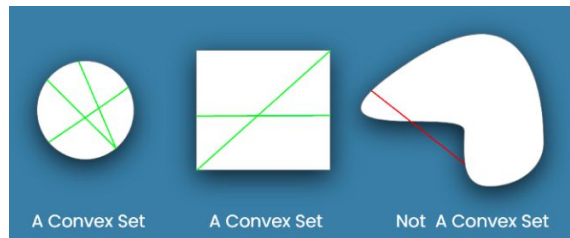
- LRMC Algorithms Without the Rank Information

Convex optimization

Convex optimization is a subfield of **mathematical optimization** that studies the problem of minimizing **convex functions** over convex sets.

- A convex set is a collection of points in which the line AB connecting any two points A, B in the set lies completely within the set.

Many classes of convex optimization problems admit polynomial-time algorithms, whereas mathematical optimization is in general **NP-hard**



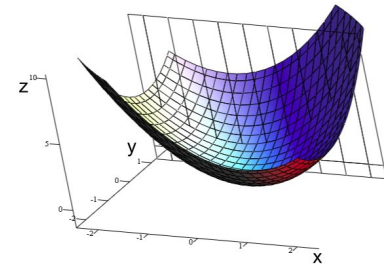
A function f is said to be a convex function if the second-order derivative of that function is greater than or equal to 0.

$$f''(x) \geq 0$$

Condition for convex functions.

Examples of convex functions: $y=e^x$, $y=x^2$. Both of these functions are differentiable twice.

If $-f(x)$ (minus $f(x)$) is a convex function, then the function is called a concave function.



A Convex function. Source Wikipedia.

$$f''(x) \geq 0$$

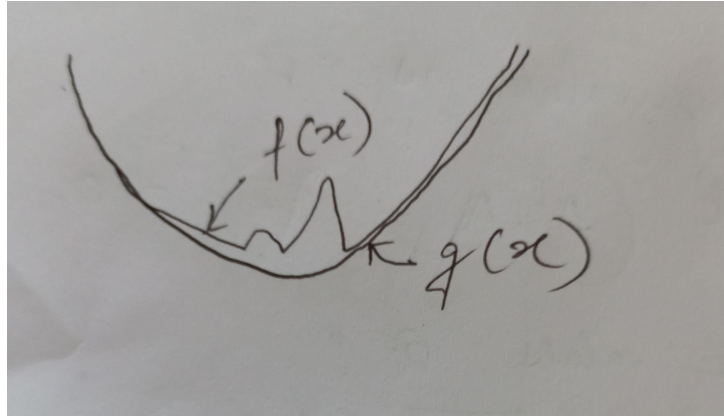
Condition for convex functions.

Examples of convex functions: $y=e^x$, $y=x^2$. Both of these functions are differentiable twice.

If $-f(x)$ (minus $f(x)$) is a convex function, then the function is called a concave function.

Convex Hull

Given a set of points in the plane. the convex hull of the set is the smallest convex polygon that contains all the points of it.



Matrix completion via convex optimization

Natural heuristic:

- is to find the matrix with the minimum rank that is consistent with the observations

$$\min_{\Phi \in \mathbb{R}^{n_1 \times n_2}} \text{rank}(\Phi) \quad \text{s.t.} \quad \mathcal{P}_{\Omega}(\Phi) = \mathcal{P}_{\Omega}(M).$$

Rank minimization is NP-hard, the above formulation is intractable.

Apply Convex Relaxation

What is convex relaxation?

Because the problem is NP hard,, one of the possible ways to solve a non-convex optimization is to solve a similar convex optimization problem. This idea is known as convex relaxation.

We replace $\text{rank}(\Phi)$ by the sum of its singular values, denoted as the nuclear norm:

$$\|\Phi\|_* \triangleq \sum_{i=1}^{\min\{n_1, n_2\}} \sigma_i(\Phi),$$

- The singular values are the diagonal entries of the matrix and are arranged in descending order.
- The singular values are always real numbers. If the matrix A is a real matrix, then U and V are also real.

Assignment [Complete it]

Find the nuclear norm of matrix:

1	-1	3
3	1	1

Calculate singular values of a matrix:

- Let A be square or non-square matrix.

Compute AA^T [always square]

- Find eigenvalues as:

$$\det(A^T A - \lambda I) = 0.$$

- Singular values σ_i :

$$\sigma_i = \sqrt{\lambda_i} \quad [\lambda_i = \text{non-zero eigenvalues}]$$

Example:

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$$

$$A^T A = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$$

- Calculate $\det(A^T A - \lambda I) = 0$

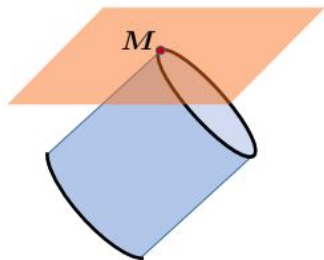
$$\lambda = 0, 0.597, 90.4.$$

$$\sigma_i = \sqrt{0.597}, \sqrt{90.4}.$$

Ref:

https://web.mit.edu/be.400/www/SVD/Singular_Value_Decomposition.htm#:~:text=Also%2C%20the%20singular%20values%20in%20and%20V%20are%20also%20real

Geometric illustration of nuclear norm minimization:



- the cylinder represents level sets of the nuclear norm;
- the hyperplane represents the measurement constraint.
- The two sets intersect at the thickened edges, which correspond to low-rank solutions.

We replace $\text{rank}(\Phi)$ by the sum of its singular values, denoted as the nuclear norm:

$$\|\Phi\|_* \triangleq \sum_{i=1}^{\min\{n_1, n_2\}} \sigma_i(\Phi),$$

$$\min_{\Phi \in \mathbb{R}^{n_1 \times n_2}} \|\Phi\|_* \quad \text{s.t.} \quad \mathcal{P}_\Omega(\Phi) = \mathcal{P}_\Omega(M).$$

- The nuclear norm is **equal to the sum of the singular values of a matrix**.
- The best convex lower bound of the rank function on the set of matrices whose singular values are all bounded by 1.

Nuclear Norm and Semidefinite Program

The symmetric matrix A is said positive semidefinite ($A \succeq 0$) **if all its eigenvalues are non negative**.

the nuclear norm can be represented as the solution to a semidefinite program:

$$\begin{aligned} \|\Phi\|_* = \min_{W_1, W_2} \quad & \frac{1}{2} (\text{Tr}(W_1) + \text{Tr}(W_2)) \\ \text{s. t.} \quad & \begin{bmatrix} W_1 & \Phi \\ \Phi^\top & W_2 \end{bmatrix} \succeq 0. \end{aligned}$$

W_1 and W_2 : Other two symmetric matrices

What is Semidefinite Program?

In semidefinite programming we minimize a linear function subject to the constraint that an affine combination of symmetric matrices is positive semidefinite.

The computational and memory complexities of nuclear norm minimization can be quite expensive for large-scale problems, even with first-order methods, due to optimizing over and storing the matrix variable

Trace?

The **trace** of a **square matrix** \mathbf{A} , denoted $\text{tr}(\mathbf{A})$, is defined to be the sum of elements on the **main diagonal** (from the upper left to the lower right) of \mathbf{A} .

The trace is only defined for a square matrix ($n \times n$).

It can be proved that the trace of a matrix is the sum of its (complex) **eigenvalues**

Let \mathbf{A} be a matrix, with

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 3 \\ 11 & 5 & 2 \\ 6 & 12 & -5 \end{pmatrix}$$

Then

$$\text{tr}(\mathbf{A}) = \sum_{i=1}^3 a_{ii} = a_{11} + a_{22} + a_{33} = 1 + 5 + (-5) = 1$$

Final Goal

We solve nuclear norm minimization instead of Rank minimization problem, which searches for a matrix with the minimum nuclear norm that satisfies all the measurements:

$$\min_{\Phi \in \mathbb{R}^{n_1 \times n_2}} \|\Phi\|_* \quad \text{s.t.} \quad \mathcal{P}_\Omega(\Phi) = \mathcal{P}_\Omega(M).$$

Advantages:

- This gives a convex program that can be solved efficiently in polynomial time.
- It doesn't require knowledge of the rank a priori.

Final number of measurements?

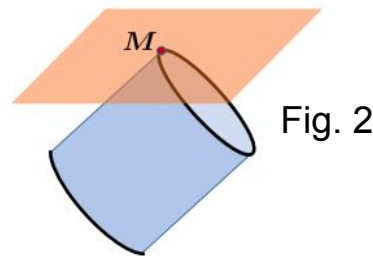


Fig. 2

$$\min_{\Phi \in \mathbb{R}^{n_1 \times n_2}} \|\Phi\|_* \quad \text{s.t.} \quad \mathcal{P}_\Omega(\Phi) = \mathcal{P}_\Omega(M). \quad (5)$$

it can exactly recover a low-rank matrix as soon as the number of measurements is slightly larger than the information-theoretic lower bound by a logarithmic factor. Suppose that each entry of M is observed independently with probability $p \in (0, 1)$. If p satisfies

$$p \geq C \frac{\mu r \log^2 n}{n},$$

16

for some large enough constant $C > 0$, then with high probability, the nuclear norm minimization algorithm (5) exactly recovers M as the unique optimal solution of (5). Fig. 2 illustrates the geometry of nuclear norm minimization when the number of measurements is sufficiently large. When both μ and r are much smaller than n , this means we can recover a low-rank matrix even when the proportion of observations is vanishingly small.

L1 Norm and Nuclear norm

The notation for L1 norm of a vector x is $\|x\|_1$.

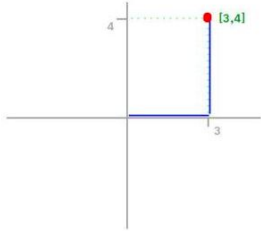
To calculate the norm, you need to **take the sum of the absolute vector values**.

$$\|v\|_1 = |a_1| + |a_2| + |a_3|$$

The nuclear norm (sometimes called Schatten 1-norm or trace norm) of a matrix A , denoted $\|A\|_*$, is defined as **the sum of its singular values**. $\|A\|_* = \sum_i \sigma_i(A)$.

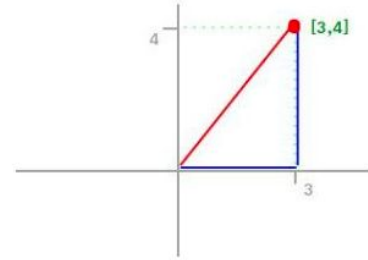
<https://www.youtube.com/watch?v=SXEYIGqXSxk&t=312s>

Having, for example, the vector $X = [3, 4]$:



The L1 norm is calculated by

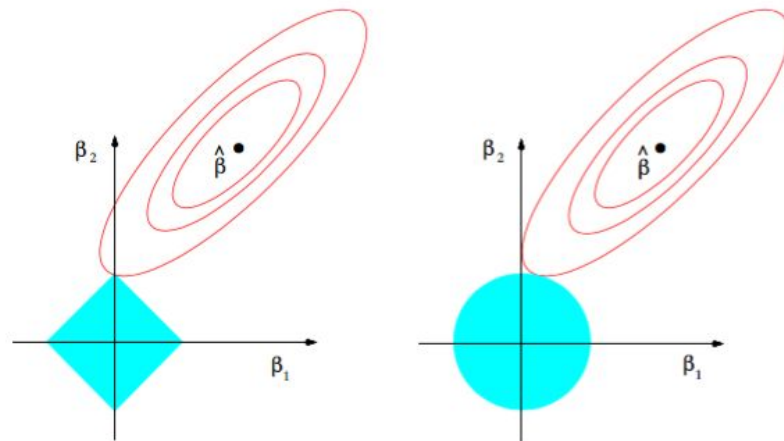
$$||X||_1 = |3| + |4| = 7$$



Using the same example, the L2 norm is calculated by

$$||X||_2 = \sqrt{(|3|^2 + |4|^2)} = \sqrt{9+16} = \sqrt{25} = 5$$

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



Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions $|\beta_1| + |\beta_2| \leq t$ and $\beta_1^2 + \beta_2^2 \leq t^2$, respectively, while the red ellipses are the contours of the least squares error function.

Norms as a unit circle:

For l_1 norm:

If $x_1 = 1$, then $\|(x_1, x_2)\| = 1$
iff. $x_2 = 0$

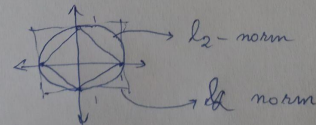
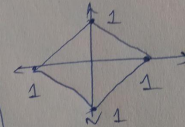
$$\Rightarrow |x_1| + |x_2| = 1$$

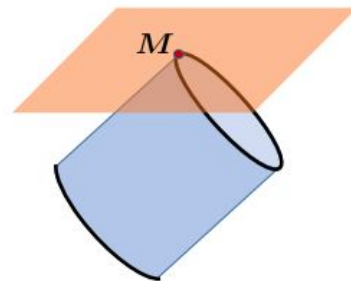
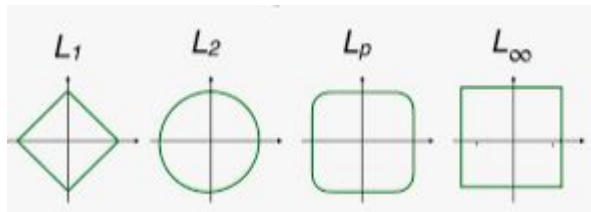
$$\text{for } Q_1, x_2 = 1 - x_1$$

$$Q_2: x_2 = 1 + x_1$$

$$Q_3: x_2 = -x_1 - 1$$

$$Q_4: x_2 = x_1 - 1$$





Geometric illustration of nuclear norm minimization:

The cylinder represents level sets of the nuclear norm;

The hyperplane represents the measurement constraint.

The two sets intersect at the thickened edges, which correspond to low-rank solutions.

Matrix completion via nonconvex optimization

- To reduce computations further whose complexities scale more favorably in n
-

Why is it nonconvex optimization?

A non-convex optimization problem is **any problem where the objective or any of the constraints are non-convex.**

Such a problem may have multiple feasible regions and multiple locally optimal points within each region.

- A function is concave if $-f$ is convex
- **A non-convex function "curves up and down" -- it is neither convex nor concave.**

Features

Based on gradient descent using a proper initialization.

Incorporate rank of the matrix M

Consider a rank-constrained least-squares problem:

$$\min_{\Phi \in \mathbb{R}^{n_1 \times n_2}} \|\mathcal{P}_\Omega(\Phi - M)\|_F^2, \quad \text{s.t.} \quad \text{rank}(\Phi) \leq r,$$

Frobenius norm of a matrix

$$\min_{\Phi \in \mathbb{R}^{n_1 \times n_2}} \|\mathcal{P}_\Omega(\Phi - M)\|_F^2, \quad \text{s.t.} \quad \text{rank}(\Phi) \leq r,$$

$\|\cdot\|_F$:square root of the sum of the squares of the elements of the matrix

$$\min_{\Phi \in \mathbb{R}^{n_1 \times n_2}} \|\mathcal{P}_\Omega(\Phi - M)\|_F^2, \quad \text{s.t.} \quad \text{rank}(\Phi) \leq r,$$

Consider low-rank factorization,

$$\Phi = XY^\top, \text{ where } X \in \mathbb{R}^{n_1 \times r} \quad \text{and } Y \in \mathbb{R}^{n_2 \times r}$$

$$\min_{X, Y} f(X, Y) := \left\| \mathcal{P}_\Omega(XY^\top - M) \right\|_F^2.$$

Advantage:

memory complexities of X and Y are linear in n instead of quadratic in n when dealing with Φ .

How do we optimize the nonconvex loss $F(X, Y)$?

Initialization:

Consider the partially-observed matrix $\frac{1}{p} \mathcal{P}_{\Omega}(M)$

*The observation probability p , if not known, can be estimated by the sample proportion $|\Omega|/(n_1 n_2)$.

How do we optimize the nonconvex loss $F(X, Y)$?

Let $U_0 \Sigma_0 V_0^\top$ be the best rank- r approximation of $\frac{1}{p} \mathcal{P}_\Omega(M)$, where $U_0 \in \mathbb{R}^{n_1 \times r}$, $V_0 \in \mathbb{R}^{n_2 \times r}$ contain orthonormal columns and Σ_0 is an $r \times r$ diagonal matrix. The spectral initialization sets $X_0 = U_0 \Sigma_0^{1/2}$ and $Y_0 = V_0 \Sigma_0^{1/2}$.

Refine:

$$\begin{bmatrix} \mathbf{X}_{t+1} \\ \mathbf{Y}_{t+1} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_t \\ \mathbf{Y}_t \end{bmatrix} - \eta_t \begin{bmatrix} \nabla_{\mathbf{X}} F(\mathbf{X}_t, \mathbf{Y}_t) \\ \nabla_{\mathbf{Y}} F(\mathbf{X}_t, \mathbf{Y}_t) \end{bmatrix}, \quad (8)$$

where η_t is the step size, and $\nabla_{\mathbf{X}} F(\mathbf{X}, \mathbf{Y})$, $\nabla_{\mathbf{Y}} F(\mathbf{X}, \mathbf{Y})$ are the partial derivatives with respect to \mathbf{X} and \mathbf{Y} that can be derived easily.

① How knowledge of Rank can be exploited in matrix completion?

②
$$\begin{bmatrix} 3 & 0 & 2 & 1 \\ 2 & 0 & 0 & 0 \\ 1 & 0 & 0 & 4 \\ 5 & 7 & 9 & 2 \end{bmatrix}$$

- Is this matrix suitable for low rank completion? Justify

Hint: Check in terms of sparsity and DoF. etc.

3. Can $y = e^{2k-2}$ be used as the constraint/min function of convex optimization?

Hint: Check if the function is convex or not.

4. In the view of coherence, comment which matrix is better for LMC?

$$M_1 = \begin{bmatrix} 5 & 1 & 0 \\ 2 & 5 & 4 \\ 2 & 2 & 1 \end{bmatrix} \quad M_2 = \begin{bmatrix} 5 & 4 & 1 \\ 0 & 3 & 2 \\ 0 & 0 & 5 \end{bmatrix}$$

5. check the matrix if.
a) low rank b) semidefinite

$$\begin{bmatrix} 1 & 0 & 1 \\ -2 & -3 & 1 \\ 3 & 3 & 0 \end{bmatrix}$$

Hint: Check from the basic definition.