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

Matrix completion

• Motivation

• Theoretical aspects:
  – nuclear norm
  – low-rank matrix sensing
  – low-rank matrix completion

• efficient algorithm
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 $\mathcal{M} \in \mathbb{R}^{n_1 \times n_2}$ to represent the Netflix data set, we may model it through factorization:

  $\begin{bmatrix}
  \times & \times \\
  \times & \times \\
  \times & \times 
  \end{bmatrix} \mathcal{M} \approx \begin{bmatrix}
  \end{bmatrix}$

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

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

- Observe partial information about distances:

\[ M_{i,j} = \|x_i - x_j\|_2^2 \]

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

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

then

\[
M_{i,j} = \mathbf{x}_i^T \mathbf{x}_i + \mathbf{x}_j^T \mathbf{x}_j - 2\mathbf{x}_i^T \mathbf{x}_j
\]

• Matrix completion problem: Let \( Y = XX^T \). The distance matrix \( M \in \mathbb{R}^{n \times n} \) between points can be written as

\[
M = \text{diag}(Y)e^T + ed\text{diag}(Y)^T - 2Y
\]

• The rank \( r \) of \( M \) 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.
Tomasi and Kanade's factorization: Given $n$ points $\mathbf{x}_{i,j}^T \in \mathbb{R}^2$ corresponding to the location of the $i$th point in the $j$th frame, define the matrix

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

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

$$\mathbf{M} = \begin{bmatrix} \mathbf{s}_1^T \\ \vdots \\ \mathbf{s}_n^T \end{bmatrix} \begin{bmatrix} \mathbf{w}_1 & \mathbf{w}_2 & \cdots & \mathbf{w}_{2m} \end{bmatrix}$$

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

Due to occlusions, there are many missing entries in the matrix $\mathbf{M}$. Can we complete the missing entries?
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
Let $M \in \mathbb{R}^{n \times n}$ (square case for simplicity) be a matrix of rank $r \ll n$.

The Singular Value Decomposition (SVD) of $M$ is given as

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

where $\{\sigma_i\}_{i=1}^{r}$ are the singular values; and $\{u_i\}_{i=1}^{r}, \{v_i\}_{i=1}^{r}$ are the singular vectors.

$M$ has $(2n - r)r$ degrees of freedom.
We make linear measurements of $M$:

$$y_i = \langle A_i, M \rangle = \text{Tr}(A_i^T M), \quad i = 1, \ldots m,$$

which can be written more concisely in an operator form:

$$y = A(M)$$

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

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

The problem of rank minimization:

$$\hat{M} = \arg\min_X \text{rank}(X) \quad \text{s.t.} \quad y = A(X).$$
Nuclear norm

Just as $\ell_1$ norm provides a convex relaxation to cardinality minimization, we use the nuclear norm which gives a convex relaxation to rank minimization.

**Definition 1.** The nuclear norm of $X$ is defined as

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

where $\sigma_i(X)$ is the $i$th largest singular value of $X$.

- Since the rank is $\sum_{i=1}^{n} 1(\sigma_i(X) \neq 0)$, the nuclear norm can be thought as an $\ell_1$ norm relaxation of the vector of singular values.

- This is a norm. Relationships between different norms:

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

- Tightest convex relaxation: $\{X : \|X\|_* \leq 1\}$ is the convex hull of rank-1 matrices obeying $\|xy^T\| \leq 1$. 
Additivity of the nuclear norm

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

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

This is similar to the $\ell_1$ norm when $x$ and $y$ have disjoint support:

$$\|x + y\|_1 = \|x\|_1 + \|y\|_1$$

which is essentially all we need to get the proof of $\ell_1$ minimization with RIP...
Lemma 2.

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

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

**Proof:** on the blackboard.
Definition 2. For a given norm $\| \cdot \|_A$ in an inner product space $\langle \cdot, \cdot \rangle$, the dual norm is defined as

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

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

$$\langle X, Y \rangle \leq \|X\|_A \|Y\|_A^*.$$ 

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 $\| \cdot \|$;
rank minimization vs cardinality minimization:

<table>
<thead>
<tr>
<th>parsimony concept</th>
<th>cardinality</th>
<th>rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hilbert Space norm</td>
<td>Euclidean</td>
<td>Frobenius</td>
</tr>
<tr>
<td>sparsity inducing norm</td>
<td>$\ell_1$</td>
<td>nuclear</td>
</tr>
<tr>
<td>dual norm</td>
<td>$\ell_\infty$</td>
<td>operator</td>
</tr>
<tr>
<td>norm additivity</td>
<td>disjoint support</td>
<td>orthogonal row and column spaces</td>
</tr>
<tr>
<td>convex optimization</td>
<td>linear programming</td>
<td>semidefinite programming</td>
</tr>
</tbody>
</table>

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

\[
\hat{M} = \arg\min_X \text{rank}(X) \quad \text{s.t.} \quad y = A(X).
\]

We pose the following nuclear norm minimization algorithm:

\[
\hat{M} = \arg\min_X \|X\|_* \quad \text{s.t.} \quad y = A(X),
\]

which can be solved efficiently via SDP:

\[
\hat{M} = \arg\min_{X,W_1,W_2} \frac{1}{2} \text{Tr}(W_1) + \frac{1}{2} \text{Tr}(W_2) \quad \text{s.t.} \quad y = A(X), \quad \begin{bmatrix} W_1 & X \\ X^T & W_2 \end{bmatrix} \succeq 0.
\]
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*

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

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

- If $\{\mathcal{A}_i\}_{i=1}^m$ 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 $A$ 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 $M$ exactly.

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

- For the noisy case,

$$y = A(M) + w$$

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

$$\hat{M} = \arg\min_X \frac{1}{2} \| y - A(X) \|_2^2 + \lambda \| X \|_*,$$

where $\lambda$ is a regularization parameter.
• If $\|A^*(w)\| \leq \lambda/2$ and $\delta_{4r} < (3\sqrt{2} - 1)/17$, then

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

for some constant $C$. For the Gaussian case,

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

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

• If $M$ is an approximately low-rank matrix, we further have

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

with probability at least $1 - 2e^{-cn}$, in the Gaussian sampling case.
In the matrix completion setting, we are given partial observations of the entries of $\mathbf{M}$, 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

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

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

- Consider the following algorithm:

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

- The observation operator doesn’t satisfy RIP!
Consider a rank-one matrix $M = xy^T$ with the following sampling pattern:

\[
\begin{bmatrix}
\times & \times & \times & \times & \times & \times \\
\times & \times & \times & \times & \times & \times \\
\times & \times & \times & \times & \times & \times \\
\times & \times & \times & \times & \times & \times \\
\times & \times & \times & \times & \times & \times \\
\times & \times & \times & \times & \times & \times \\
\end{bmatrix}
\]

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.

$\implies$ 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:

\[
\begin{bmatrix}
1 & 0 & 0 & \cdots & 0 \\
0 & 0 & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \cdots & 0
\end{bmatrix}
\quad \text{vs} \quad
\begin{bmatrix}
1 & 1 & 1 & \cdots & 1 \\
1 & 1 & 1 & \cdots & 1 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & 1 & 1 & \cdots & 1
\end{bmatrix}
\quad \text{vs} \quad
\begin{bmatrix}
1 & 1 & 1 & \cdots & 1 \\
0 & 0 & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \cdots & 0
\end{bmatrix}
\]

The middle one would be “easier” to complete.

Column and row spaces cannot be aligned with basis vectors.
Coherence

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

Definition 4. [Coherence] Smallest scalar $\mu$ obeying

\[
\max_{1 \leq i \leq n} \|U^T e_i\|_2^2 \leq \frac{\mu}{n}, \quad \max_{1 \leq i \leq n} \|V^T e_i\|_2^2 \leq \frac{\mu}{n},
\]

where $e_i$ is the $i$th standard basis vector.

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

We would like $\mu = O(1)$. 
Information-theoretic lower bound

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

\[ m \leq \mu \times nr \times \log n \approx \text{dof} \times \mu \log n \]

Remarks:

• When \( \mu = O(1) \), we need \( m \leq nr \log n \).

• 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..] \[ \text{There exists universal constant } c_0, c_1, c_2 > 0 \text{ such that if} \]

\[ m \geq c_0 \mu n r \log^2 n, \]

\[ \text{then } \mathcal{M} \text{ is the unique optimal solution to the nuclear norm minimization problem with probability at least } 1 - c_1 n^{-c_2}. \]

Remark:

- This result is optimal up to a logarithmic factor in \( n \). See [Chen, Incoherence-Optimal Matrix Completion].
\[ \left\| \begin{bmatrix} x & y \\ y & z \end{bmatrix} \right\| \leq 1 \]
A few notations

- $P_U$ is the orthogonal projection to the column space of $M$;
- $P_V$ is the orthogonal projection to the row space of $M$;
- Let $T$ be the span of matrices of the form:
  $$T = \{UX^T + YV^T : X, Y \in \mathbb{R}^{n \times r}\}$$
- Let $\mathcal{P}_T$ be the orthogonal projection onto $T$:
  $$\mathcal{P}_T(X) = P_U X + X P_V - P_U X P_V$$
- The complement projection $\mathcal{P}_{T\perp} = \mathcal{I} - \mathcal{P}_T$:
  $$\mathcal{P}_{T\perp}(X) = (I - P_U) X (I - P_V)$$
The subgradient of $\| \cdot \|_*$ at $M$ can be written as

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

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

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

The subgradient doesn’t depend on the singular values of $M$. 
Basic consequence of incoherence

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

\[
\left\| P_T(e_i e_j^T) \right\|_F^2 \leq \frac{2\mu r}{n}.
\]

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

Define the operator

$$R_\Omega(X) = \sum_{k=1}^{m} \langle X, e_{i_k} e_{j_k}^T \rangle e_{i_k} e_{j_k}^T = \sum_{k=1}^{m} X_{i_k,j_k} e_{i_k} e_{j_k}^T$$

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

$$\|R_\Omega\| \leq \frac{8}{3} \beta \log(n)$$

with probability at least $1 - n^{2-2\beta}$. 
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 \( Y \) supported on \( \Omega \) such that \( Y \in \partial \| M \|_* \), i.e.

\[
\mathcal{P}_T(Y) = UV^T, \quad \| \mathcal{P}_{T\perp}(Y) \| \leq 1.
\]

The first equality constraint is not easy to satisfy, see [Candès 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

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

and there exists \( Y \) supported on \( \Omega \) such that

\[
\left\| \mathcal{P}_T(Y) - UV^T \right\|_F \leq \sqrt{\frac{r}{2n}}, \quad \left\| \mathcal{P}_T^\perp(Y) \right\| < \frac{1}{2},
\]

then \( M \) 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 $\beta > 1$,

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

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

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

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

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

$$\frac{n^2}{m} \left\| P_T \mathcal{R}_\Omega P_T - \frac{m}{n^2} P_T \right\| \leq \frac{1}{2}.$$
We introduce the clever golfing scheme proposed by David Gross.

\[ X_0 = \text{sgn}\rho \]
\[ -\mathcal{P}_T\mathcal{R}_1 X_0 \rightarrow X_1 \]
\[ X_2 - \mathcal{P}_T\mathcal{R}_2 X_1 \rightarrow 0 \]

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

\[ \hat{M} = \arg\min_X \frac{1}{2} \| y - A(X) \|_2^2 + \lambda \| X \|_* \]

- **Initialization:** \( x_0 = x_{-1} \in \mathbb{R}^n, \theta_0 = 1, \)

- **For** \( k = 1, 2, \ldots, \)

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

  \[ Y_k = X_{k-1} + \left( \frac{\theta_{k-1} - 1}{\theta_k} \right) (X_{k-1} - X_{k-2}) \]

  \[ X_k = \text{prox}_{t_k \lambda \| \cdot \|_*} (Y_k - t_k A^*(A(Y_k) - y)) \]

- **What is the proximal operator for \( \| \cdot \|_* ? \)**
Proposition 6.

\[
\text{prox}_{t_k \lambda \| \cdot \|_*}(X) = \arg \min_{Z} \left\{ \frac{1}{2}\| Z - X \|_2^2 + t_k \lambda \| Z \|_* \right\} = T_{t_k \lambda}(X)
\]

where

\[
T_{\tau}(X) = U T_{\tau}(\Sigma) V^T,
\]

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

\[
T_{\tau}(\Sigma) = \text{diag}(\{(\sigma_i - \tau)_+\}).
\]