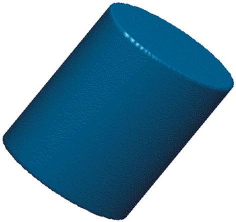
Computational Principles for High-dim Data Analysis

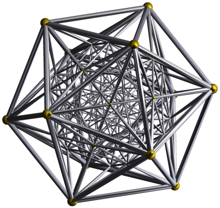
(Lecture Ten)

Yi Ma and Jiantao Jiao

EECS Department, UC Berkeley

September 30, 2021





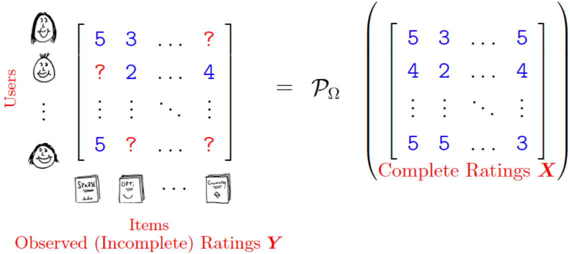
Convex Methods for Low-Rank Matrix Recovery
(Matrix Completion)

1. Motivating Example
2. Nuclear Norm Minimization
3. Algorithm: Augmented Lagrange Multiplier
4. Conditions for Success
5. Stable Matrix Completion

"Mathematics is the art of giving the same name to different things."
– Henri Poincaré

Example of Low-rank Matrix Completion

Recommendation Systems (how internet companies make money):



We observe:

$$\underset{\text{Observed ratings}}{\boldsymbol{Y}} = \mathcal{P}_\Omega\left[\underset{\text{Complete ratings}}{\boldsymbol{X}}\right],$$

where $\Omega \doteq \{(i,j) \mid \text{user } i \text{ has rated product } j\}$.

Nuclear Norm Minimization

Problem (Matrix Completion)

Let $\boldsymbol{X}_o \in \mathbb{R}^{n \times n}$ be a low-rank matrix. Suppose we are given $\boldsymbol{Y} = \mathcal{P}_\Omega[\boldsymbol{X}_o]$, where $\Omega \subseteq [n] \times [n]$. Fill in the missing entries of $\boldsymbol{X}_o$.

Notice: If $(i, j) \notin \Omega$, $\mathcal{P}_\Omega[\boldsymbol{E}_{ij}] = \mathbf{0}$. So $\mathcal{P}_\Omega$ has matrices of rank one in its null space! So, $\mathcal{P}_\Omega$ *cannot be rank-RIP* for any rank $r > 0$ with $\delta < 1$.

Question: can we still find $\boldsymbol{X}_o$ by solving the nuclear norm minimization:

$$\min \|\boldsymbol{X}\|_* \quad \text{subject to} \quad \mathcal{P}_\Omega[\boldsymbol{X}] = \boldsymbol{Y}? \tag{1}$$

Simulations lead the way of investigation – need an algorithm...

Algorithm via Augmented Lagrange Multiplier

Nuclear norm minimization for matrix completion:

$$\min \underbrace{\|\boldsymbol{X}\|_*}_{f(\boldsymbol{x})} \quad \text{subject to} \quad \underbrace{\mathcal{P}_\Omega[\boldsymbol{X}] = \boldsymbol{Y}}_{g(\boldsymbol{x})=0}. \tag{2}$$

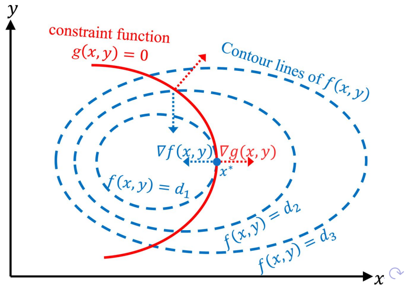
The Lagrangian method:

$$\mathcal{L}(\boldsymbol{X}, \boldsymbol{\Lambda}) = \|\boldsymbol{X}\|_* + \langle \boldsymbol{\Lambda}, \boldsymbol{Y} - \mathcal{P}_\Omega[\boldsymbol{X}] \rangle. \tag{3}$$

Optimality conditions:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{X}} = 0, \quad \frac{\partial \mathcal{L}}{\partial \boldsymbol{\Lambda}} = 0. \tag{4}$$

However, it only holds
at the point of the optimal solution $\boldsymbol{x}^\star$.



Algorithm via Augmented Lagrange Multiplier

The *augmented* Lagrangian is to regularize the landscape around the optimal solution $\boldsymbol{x}^\star$:

$$\mathcal{L}_\mu(\boldsymbol{X}, \boldsymbol{\Lambda}) = \|\boldsymbol{X}\|_* + \langle \boldsymbol{\Lambda}, \boldsymbol{Y} - \mathcal{P}_\Omega[\boldsymbol{X}] \rangle + \frac{\mu}{2} \|\boldsymbol{Y} - \mathcal{P}_\Omega[\boldsymbol{X}]\|_F^2. \quad (5)$$

Amenable for alternating optimization to converge to the optimal solution $\boldsymbol{x}^\star$ more easily and efficiently:

$$\text{Primal:} \quad \boldsymbol{X}_{k+1} \in \arg\min_{\boldsymbol{X}} \mathcal{L}_\mu(\boldsymbol{X}, \boldsymbol{\Lambda}_k), \quad (6)$$

$$\text{Dual:} \quad \boldsymbol{\Lambda}_{k+1} = \boldsymbol{\Lambda}_k + \mu \mathcal{P}_\Omega\big[\boldsymbol{Y} - \boldsymbol{X}_{k+1}\big]. \quad (7)$$

Algorithm: Proximal Gradient Descent

How to minimize the augmented Lagrangian $\mathcal{L}_\mu$:

$$\min_{\boldsymbol{X}} F(\boldsymbol{X}) \doteq \underbrace{\|\boldsymbol{X}\|_*}_{g(\boldsymbol{X})\text{ convex}} + \underbrace{\langle \boldsymbol{\Lambda}, \boldsymbol{Y} - \mathcal{P}_\Omega[\boldsymbol{X}]\rangle + \frac{\mu}{2}\|\boldsymbol{Y} - \mathcal{P}_\Omega[\boldsymbol{X}]\|_F^2}_{f(\boldsymbol{X})\text{ smooth, convex, }\mu\text{-Lipschitz}}. \quad (8)$$

At each iterate $\boldsymbol{X}_k$, construct a local (quadratic) upper bound for F:

$$\hat{F}(\boldsymbol{X}, \boldsymbol{X}_k) = g(\boldsymbol{X}) + f(\boldsymbol{X}_k) + \langle \nabla f(\boldsymbol{X}_k), \boldsymbol{X} - \boldsymbol{X}_k\rangle + \frac{\mu}{2}\|\boldsymbol{X} - \boldsymbol{X}_k\|_2^2. \quad (9)$$

Proximal gradient descent: the next iterate $\boldsymbol{X}_{k+1}$ is computed as

$$\boldsymbol{X}_{k+1} = \arg\min_{\boldsymbol{X}} \left\{ g(\boldsymbol{X}) + \frac{\mu}{2}\left\| \boldsymbol{X} - \underbrace{\left(\boldsymbol{X}_k - \tfrac{1}{\mu}\nabla f(\boldsymbol{X}_k)\right)}_{\boldsymbol{M}} \right\|_F^2 \right\} \quad (10)$$

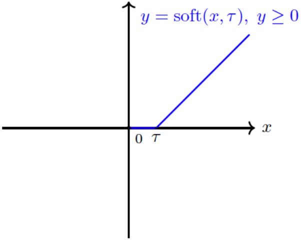
$$= \operatorname{prox}_{g/\mu}(\boldsymbol{M}) \quad \text{(see details in Chapter 8).} \quad (11)$$

Algorithm: Proximal Operator for Nuclear Norm

For a matrix $\boldsymbol{M}$ with SVD $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$, its singular value thresholding operator is:

$$\mathcal{D}_\tau[\boldsymbol{M}] = \boldsymbol{U}\mathcal{S}_\tau\left[\boldsymbol{\Sigma}\right]\boldsymbol{V}^*,$$

where $\mathcal{S}_\tau[\boldsymbol{X}] = \mathrm{sign}(\boldsymbol{X}) \circ (|\boldsymbol{X}| - \tau)_+$ is the entry-wise soft thresholding operator.



Theorem

The unique solution $\boldsymbol{X}_\star$ to the program:

$$\min_{\boldsymbol{X}}\{\|\boldsymbol{X}\|_* + \tfrac{\mu}{2}\|\boldsymbol{X} - \boldsymbol{M}\|_F^2\}, \tag{12}$$

is given by

$$\boldsymbol{X}_\star = \mathcal{D}_{\mu^{-1}}[\boldsymbol{M}]. \tag{13}$$

Algorithm via Augmented Lagrange Multiplier

Outer Loop: Matrix Completion by ALM

1: **initialize:** $\boldsymbol{X}_0 = \boldsymbol{\Lambda}_0 = 0, \mu > 0$.
2: **while** not converged **do**
3: compute $\boldsymbol{X}_{k+1} \in \arg\min_{\boldsymbol{X}} \mathcal{L}_\mu\big(\boldsymbol{X}, \boldsymbol{\Lambda}_k\big)$ (say by PG);
4: compute $\boldsymbol{\Lambda}_{k+1} = \boldsymbol{\Lambda}_k + \mu\big(\boldsymbol{Y} - \mathcal{P}_\Omega\big[\boldsymbol{X}_{k+1}\big]\big)$.
5: **end while**

Inner Loop: Proximal Gradient

1: **initialize:** $\boldsymbol{X}_0$ starts with the $\boldsymbol{X}_k$ from the outer loop.
2: **while** not converged **do**
3: compute

$$\begin{aligned}\boldsymbol{X}_{\ell+1} &= \operatorname{prox}_{g/\mu}\big(\boldsymbol{X}_\ell - \mu^{-1}\nabla f(\boldsymbol{X}_\ell)\big) \\ &= \mathcal{D}_{\mu^{-1}}\Big[\underbrace{\mathcal{P}_{\Omega^c}[\boldsymbol{X}_\ell] + \boldsymbol{Y} + \mu^{-1}\mathcal{P}_\Omega[\boldsymbol{\Lambda}_k]}_{\text{exercise}}\Big].\end{aligned}$$

4: **end while**

Similar Phenomena of Success

Comparison: low-rank matrix recovery from random linear measurements versus matrix completion from random sampled entries.

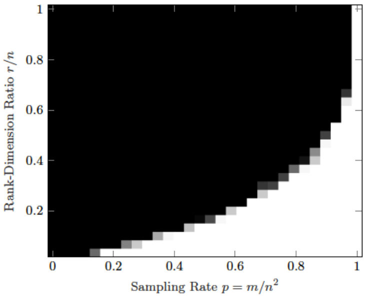


Figure: Left: phase transition for matrix recovery; Right: phase transition for matrix completion.

When Nuclear Norm Minimization Succeeds?

When it fails?

1. if $\boldsymbol{X}_o$ is itself *sparse* (as in the example of $\boldsymbol{E}_{ij}$)
2. if Ω is chosen adversarially (e.g., an entire row or column of $\boldsymbol{X}_o$).

Notice for any rank-r orthogonal matrix $\boldsymbol{U}$:

$$\sum_i \|\boldsymbol{e}_i^* \boldsymbol{U}\|_2^2 = \|\boldsymbol{U}\|_F^2 = r \quad \Longrightarrow \quad \max_i \|\boldsymbol{e}_i^* \boldsymbol{U}\|_2^2 \geq r/n.$$

Definition

We say that $\boldsymbol{X}_o = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$ is *ν-incoherent* if the following hold:

$$\forall\, i \in [n], \quad \|\boldsymbol{e}_i^* \boldsymbol{U}\|_2^2 \leq \nu r/n, \tag{14}$$

$$\forall\, j \in [n], \quad \|\boldsymbol{e}_j^* \boldsymbol{V}\|_2^2 \leq \nu r/n. \tag{15}$$

When Nuclear Norm Minimization Succeeds?

Bernoulli $\mathrm{Ber}(p)$ **sampling model**: each entry (i, j) belongs to the observed set Ω independently with probability $p \in [0, 1]$. Hence, the expected number of observed entries is:

$$m = \mathbb{E}\big[\, |\Omega| \,\big] = pn^2. \tag{16}$$

Theorem (Matrix Completion via Nuclear Norm Minimization)

Let $\boldsymbol{X}_o \in \mathbb{R}^{n \times n}$ be a rank-r matrix with incoherence parameter ν. Suppose that we observe $\boldsymbol{Y} = \mathcal{P}_\Omega[\boldsymbol{X}_o]$, with Ω sampled according to the Bernoulli model with probability

$$p \geq C_1 \frac{\nu r \log^2(n)}{n}. \tag{17}$$

Then with probability at least $1 - C_2 n^{-c_3}$, $\boldsymbol{X}_o$ is the unique optimal solution to

$$\text{minimize} \quad \|\boldsymbol{X}\|_* \quad \text{subject to} \quad \mathcal{P}_\Omega[\boldsymbol{X}] = \boldsymbol{Y}. \tag{18}$$

When Nuclear Norm Minimization Succeeds?

Lemma (Subdifferential of nuclear norm)

Let $\boldsymbol{X} \in \mathbb{R}^{n \times n}$ have compact singular value decomposition $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$. The subdifferential of the nuclear norm at $\boldsymbol{X}$ is given by

$$\partial \|\cdot\|_* (\boldsymbol{X}) = \left\{ \boldsymbol{Z} \mid \mathcal{P}_{\mathsf{T}}[\boldsymbol{Z}] = \boldsymbol{U}\boldsymbol{V}^*, \ \|\mathcal{P}_{\mathsf{T}^\perp}[\boldsymbol{Z}]\| \leq 1 \right\}. \tag{19}$$

When Nuclear Norm Minimization Succeeds?

Key ideas for the Theorem:

For the program:

$$\min \|\boldsymbol{X}\|_* \quad \text{subject to} \quad \mathcal{P}_\Omega[\boldsymbol{X}] = \mathcal{P}_\Omega[\boldsymbol{X}_o]. \tag{20}$$

Similar to the ℓ^1 case, find a dual certificate $\boldsymbol{\Lambda}$ that satisfies (the KKT condition):

- (i) $\boldsymbol{\Lambda}$ is supported on Ω: $\mathcal{P}_\Omega[\boldsymbol{\Lambda}] = \boldsymbol{\Lambda}$ and
- (ii) $\boldsymbol{\Lambda} \in \partial \|\cdot\|_* (\boldsymbol{X}_o)$ – i.e., $\mathcal{P}_{\mathsf{T}}[\boldsymbol{\Lambda}] = \boldsymbol{U}\boldsymbol{V}^*$ and $\|\mathcal{P}_{\mathsf{T}^\perp}[\boldsymbol{\Lambda}]\| \le 1$,

Strategy: look for a matrix $\boldsymbol{\Lambda}$ of smallest 2-norm that satisfies the equality constraints

$$\mathcal{P}_{\Omega^c}[\boldsymbol{\Lambda}] = \mathbf{0}, \quad \mathcal{P}_{\mathsf{T}}[\boldsymbol{\Lambda}] = \boldsymbol{U}\boldsymbol{V}^*, \tag{21}$$

and then hope to check that it satisfies the inequality constraints

$$\|\mathcal{P}_{\mathsf{T}^\perp}[\boldsymbol{\Lambda}]\| \le 1.$$

When Nuclear Norm Minimization Succeeds?

Unfortunately, this straightforward strategy does not work out directly as solution to the equalities is not so easy to analyze...

An alternative strategy: an set of (relaxed) conditions for optimality:

Proposition (KKT Conditions – Approximate Version)

The matrix $\boldsymbol{X}_o$ is the unique optimal solution to the nuclear minimization problem (18) *if the following set of conditions hold*

1. *The operator norm of the operator $p^{-1}\mathcal{P}_{\mathsf{T}}\mathcal{P}_{\Omega}\mathcal{P}_{\mathsf{T}} - \mathcal{P}_{\mathsf{T}}$ is small:*

$$\left\| p^{-1}\mathcal{P}_{\mathsf{T}}\mathcal{P}_{\Omega}\mathcal{P}_{\mathsf{T}} - \mathcal{P}_{\mathsf{T}} \right\| \leq \tfrac{1}{2}.$$

2. *There exists a dual certificate $\boldsymbol{\Lambda}$ that satisfies $\mathcal{P}_{\Omega}[\boldsymbol{\Lambda}] = \boldsymbol{\Lambda}$ and*
   - *(a)* $\|\mathcal{P}_{\mathsf{T}^{\perp}}[\boldsymbol{\Lambda}]\| \leq \frac{1}{2}$;
   - *(b)* $\|\mathcal{P}_{\mathsf{T}}[\boldsymbol{\Lambda}] - \boldsymbol{U}\boldsymbol{V}^*\|_F \leq \frac{1}{4n}$.

Matrix Completion with Noise

Problem: the observed entries are often corrupted with some noise:

$$Y_{ij} = [\boldsymbol{X}_o]_{ij} + Z_{ij},\ (i,j) \in \Omega; \quad \text{or} \quad \mathcal{P}_\Omega[\boldsymbol{Y}] = \mathcal{P}_\Omega[\boldsymbol{X}_o] + \mathcal{P}_\Omega[\boldsymbol{Z}], \tag{22}$$

where Z_{ij} can be some small noise, say $\|\mathcal{P}_\Omega[\boldsymbol{Z}]\|_F < \epsilon$.

$$\min \|\boldsymbol{X}\|_* \quad \text{subject to} \quad \|\mathcal{P}_\Omega[\boldsymbol{X}] - \mathcal{P}_\Omega[\boldsymbol{Y}]\|_F < \epsilon. \tag{23}$$

Theorem (Stable Matrix Completion)

Let $\boldsymbol{X}_o \in \mathbb{R}^{n\times n}$ be a rank-r, ν-incoherent matrix. Suppose that we observe $\mathcal{P}_\Omega[\boldsymbol{Y}] = \mathcal{P}_\Omega[\boldsymbol{X}_o] + \mathcal{P}_\Omega[\boldsymbol{Z}]$, where Ω is uniformly sampled from subsets of size

$$m \geq C_1 \nu n r \log^2(n), \tag{24}$$

then with high probability, the optimal solution $\hat{\boldsymbol{X}}$ to the convex program (23) *satisfies*

$$\|\hat{\boldsymbol{X}} - \boldsymbol{X}_o\|_F \leq c\frac{n\sqrt{n}\log(n)}{\sqrt{m}}\epsilon \leq c'\frac{n}{\sqrt{r}}\epsilon, \quad \text{for some } c > 0. \tag{25}$$

Summary

Nuclear norm minimization can recover w.h.p. a low-rank matrix $\boldsymbol{X}_o$ from

1. $m = O(nr)$ random linear measurements: $\boldsymbol{y} = \mathcal{A}[\boldsymbol{X}]$;
2. $m = O(nr\log^2 n)$ randomly sampled entries: $\boldsymbol{Y} = \mathcal{P}_\Omega[\boldsymbol{X}]$;
3. the estimate $\hat{\boldsymbol{X}}$ is stable to small noise.

Assignments

- Reading: Section 4.4-4.6 of Chapter 4.
- Programming Homework # 2.