Low Rank Matrix Completion using Alternating Minimization

Prateek Jain, Praneeth Netrapalli, Sujay Sanghavi

Presented by Chicheng Zhang
University of California San Diego

Nov, 2016

Outline

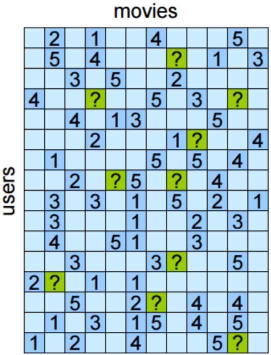
Recommender Systems

amazon.com

NETFLIX

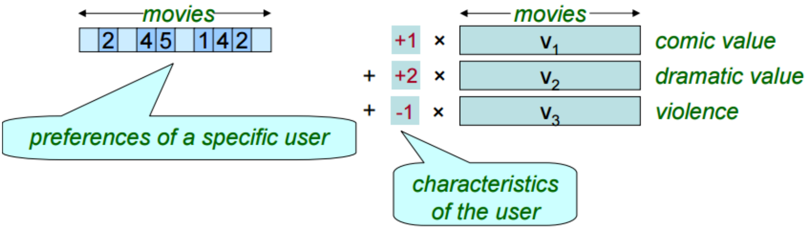
- Given m users and n items, order history
- Would like to recommend the users items they may like

Matrix Completion



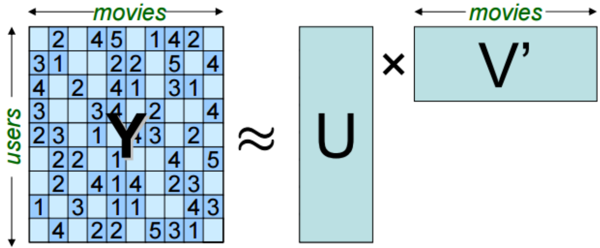
- Statistical / machine learning problem setup
- Given a matrix with entries observed at random
- Fill out the missing entries

Linear Factor Model



- Assumption: the users' rating to items is determined by a few "linear factors"
- The users and the items are both modeled as vectors in $\mathbb{R}^k$ $(k \ll m, n)$
- The rating of user $u_i \in \mathbb{R}^k$ to item $v_j \in \mathbb{R}^k$ is $\langle u_i, v_j \rangle$

Low Rank Matrix Completion: Formal Setup



- Matrix $Y \in \mathbb{R}^{m \times n}$ with rank k
- Observe entries Ω sampled from $[m] \times [n]$
- Goal: (Approximately) recover Y

Low Rank Matrix Completion: Performance Metrics

- Goal: recover Y
- **Sample Complexity:** how many entries needed
- **Computational Complexity:** how many arithmetic operations needed
- Trade off data efficiency and time efficiency

What makes matrix completion hard?

Observation 1: Sampling Probability

$$Y = \begin{bmatrix} Y_{11} & Y_{12} & - & - & \dots & - \\ Y_{21} & - & Y_{23} & - & \dots & - \\ - & - & - & - & \dots & - \\ & & \dots & & & \\ - & - & Y_{m3} & - & \dots & Y_{mn} \end{bmatrix}$$

- Completely miss column $j \Rightarrow$ large error on column j
- Need $\Omega(m+n)$ samples for small error

Observation 2: Coherence

A bad case:

$$Y = \begin{bmatrix} 1 & 0 & 0 & 0 & \dots & 1 \\ 0 & 1 & 1 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 \\ & & \dots & & & \\ 0 & 0 & 0 & 0 & \dots & 0 \end{bmatrix}$$

- Y has rank 2
- For column j, even observing a constant fraction of its entries does not help
- Need $\Omega(mn)$ samples for small error

Incoherence Assumption

Rank k matrix Y has SVD:

$$Y = U^*\Sigma^* {V^*}^T = \begin{bmatrix} u_1^* & u_2^* & \dots & u_k^* \end{bmatrix} \begin{bmatrix} \sigma_1^* & & & \\ & \sigma_2^* & & \\ & & \dots & \\ & & & \sigma_k^* \end{bmatrix} \begin{bmatrix} {v_1^*}^T \\ {v_2^*}^T \\ \dots \\ {v_k^*}^T \end{bmatrix}$$

Definition

A subspace spanned by orthonormal $U \in \mathbb{R}^{m \times k}$ is μ-coherent if

$$\max_{i \in [m]} \|e_i^T U\| \leq \mu \sqrt{\frac{k}{m}}$$

- Matrix Y is μ-coherent if both its row and column spaces (U^* and V^*) are μ-coherent.
- Enforces "denseness" of Y

Incoherence Assumption

$$\max_{i \in [m]} \| e_i^T U^* \| \leq \mu \sqrt{\frac{k}{m}}$$

Ideally ($\mu = O(1)$):

$$U^* = \begin{bmatrix} O(\sqrt{\frac{1}{m}}) & \dots & O(\sqrt{\frac{1}{m}}) \\ O(\sqrt{\frac{1}{m}}) & \dots & O(\sqrt{\frac{1}{m}}) \\ & \dots & \\ O(\sqrt{\frac{1}{m}}) & \dots & O(\sqrt{\frac{1}{m}}) \end{bmatrix}$$

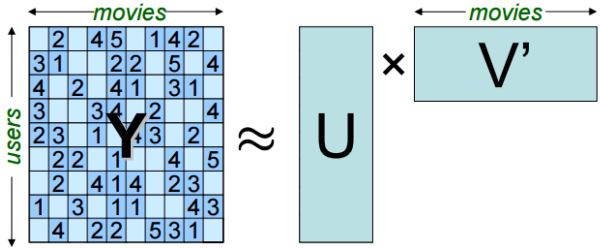
Bad Example ($\mu = \sqrt{\frac{m}{2}}$):

$$U^* = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ \dots & \\ 0 & 0 \end{bmatrix}$$

- $1 \leq \mu \leq \sqrt{\frac{\max(m,n)}{k}}$; expect "easy case" if μ is constant
- Coherence is invariant under rotation, therefore a property of subspace

Outline

Objective Function



Idea: formulate the matrix completion problem as a "factorization" problem

$$\min_{U \in \mathbb{R}^{m \times k}, V \in \mathbb{R}^{n \times k}} F_\Omega(U, V)$$

where

$$F_\Omega(U, V) := \sum_{(i,j) \in \Omega} (Y_{i,j} - (UV^T)_{i,j})^2$$

Algorithm: Part I

Algorithm **AltMin**(Ω, U_0, T)

- For $t = 1, 2, \ldots, T$:

$$V_t \leftarrow \arg\min_{V \in \mathbb{R}^{n \times k}} F_\Omega(U_{t-1}, V),$$

$$U_t \leftarrow \arg\min_{U \in \mathbb{R}^{m \times k}} F_\Omega(U, V_t).$$

- Return (U_T, V_T)

Algorithm: Part II

Algorithm **AltMinComplete**

- Initialize: $(U_0, \Sigma_0, V_0) \leftarrow \mathrm{svd}_k(P_\Omega(Y))$,
  where $P_\Omega(Y)_{i,j} := \begin{cases} Y_{i,j} & (i,j) \in \Omega \\ 0 & (i,j) \notin \Omega \end{cases}$
- $(U_T, V_T) \leftarrow$ **AltMin**(Ω, U, T)
- Return $X = U_T V_T^T$.

Performance Guarantees

Theorem

Suppose matrix Y has rank k, μ-coherent, and Ω is a random subset of $[m] \times [n]$ of size $\tilde{O}(\kappa(Y)^4 k^{4.5} \mu^2 n)$. Then, AltMinComplete outputs X such that $\|X - Y\|_F \leq \epsilon$ in $T = \tilde{O}(\log \frac{1}{\epsilon})$ iterations.

- $\kappa(Y) := \frac{\sigma_1^*}{\sigma_k^*}$ is the condition number of Y
- Implication: If coherence μ is constant, then only need $\tilde{O}(1)$ samples per row for recovery.

Comparison with Previous work:

Algorithms	Time	#Samples
Trace Norm Minimization	$O(\|\Omega\| n \sqrt{\frac{1}{\epsilon}})$	$\tilde{O}(k \mu^2 n)$
AltMinComplete	$O(\|\Omega\| k^2 \log \frac{1}{\epsilon})$	$\tilde{O}(\kappa(Y)^4 k^{4.5} \mu^2 n)$

Outline

Orthonormalization

$$F_\Omega(U,V) := \sum_{(i,j)\in\Omega} (Y_{i,j} - (UV^T)_{i,j})^2$$

Claim

During the execution of AltMinComplete, any invertible column linear transformations of the iterates $U_t (V_t)$ do not change the final output X.

Proof.

If U_t had been transformed to $U_t R$ then V_{t+1} would become $V_{t+1} R^{-T}$ in the next iteration to compensate. Vice versa. □

Implications:

- Without loss of generality we can perform orthonormalization after each iteration (Gram-Schmidt/QR/..)
- We are really learning subspaces of U^* (V^*)

What is AltMinComplete Doing?

- Given orthonormal $U \in \mathbb{R}^{m \times k}$, Ω sampled from $[m] \times [n]$,

$$V \leftarrow \arg \min_{V \in \mathbb{R}^{n \times k}} \sum_{i,j \in \Omega} (Y_{i,j} - (UV^T)_{i,j})^2$$

- If $\Omega = [m] \times [n]$, then the update is:

$$V \leftarrow Y^T U \qquad \text{Power Iteration}$$

- If $\Omega \subset [m] \times [n]$, expect the update to be

$$V \leftarrow Y^T U + G, \|G\| \approx 0 \qquad \text{Approximate Power Iteration}$$

Detour: Classical Power Iteration and Analysis

Classical Power Iteration

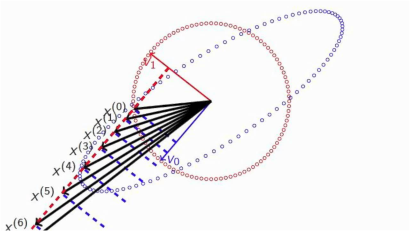
- Problem: Given matrix $A \in \mathbb{R}^{m \times n}$ $(m \leq n)$ with SVD

$$A = U^* \Sigma^* {V^*}^T = \begin{bmatrix} u_1^* & u_2^* & \dots & u_m^* \end{bmatrix} \begin{bmatrix} \sigma_1^* & & & \\ & \sigma_2^* & & \\ & & \dots & \\ & & & \sigma_m^* \end{bmatrix} \begin{bmatrix} {v_1^*}^T \\ {v_2^*}^T \\ \dots \\ {v_m^*}^T \end{bmatrix}$$

$$= U_k^* \Sigma_k^* {V_k^*}^T + U_\perp^* \Sigma_\perp^* {V_\perp^*}^T$$

- Goal: approximately compute the subspace spanned by its top k singular vectors $U_k^* = \begin{bmatrix} u_1^* & \dots & u_k^* \end{bmatrix}$
- Has a wide range of applications, e.g. PCA

Classical Power Iteration



- Algorithm:
  - Randomly initialize $U_0 \in \mathbb{R}^{m \times k}$.
  - For $t = 1, 2, \ldots, T$:
    $$V_t \leftarrow \text{orth}(A^T U_{t-1}),$$
    $$U_t \leftarrow \text{orth}(AV_t)$$
  - Return U_T, V_T.
- How do we analyze this algorithm?
- What is the notion of closeness between subspaces?

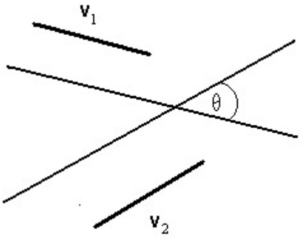
Angles between Linear Subspaces

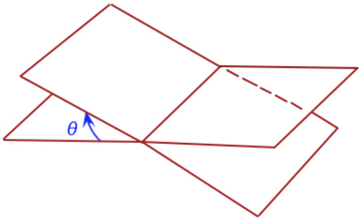
- Given two linear subspaces spanned by orthonormal bases $E, F \in \mathbb{R}^{n\times k}$
- The angle between E and F is defined as:

$$\theta(E,F) = \max_{x\in \text{span}(E)} \min_{y\in\text{span}(F)} \theta(x,y),$$

where $\theta(x,y) = \arccos \frac{|\langle x,y\rangle|}{\|x\|\|y\|}$.

Examples for $k = 1, 2$:





Properties of Subspace Angles

$$
\begin{aligned}
\cos\theta(E,F) &= \min_{\|x\|=1}\max_{\|y\|=1} |y^T F^T E x| \\
&= \sigma_k(F^T E)
\end{aligned}
$$

- Invariant under rotation
  $\Rightarrow \theta(E,F)$ is a property on subspaces
- Symmetry: $\theta(E,F) = \theta(F,E)$
- $\tan\theta(E,F) = \|(F_\perp^T E)(F^T E)^{-1}\|$
  where E does not have to be orthonormal in the last equation

Classical Analysis of Power Iteration

Suppose $A = U^* \Sigma^* {V^*}^T = U_k^* \Sigma_k^* {V_k^*}^T + U_\perp^* \Sigma_\perp^* {V_\perp^*}^T$

Lemma

If $V = A^T U$, then

$$\tan\theta(V, V_k^*) \leq \frac{\sigma_{k+1}}{\sigma_k} \tan\theta(U, U_k^*)$$

Proof.

$$\tan\theta(V, V_k^*) \quad = \quad \|({V_\perp^*}^T V)({V_k^*}^T V)^{-1}\|$$

Classical Analysis of Power Iteration

Suppose $A = U^*\Sigma^* {V^*}^T = U_k^*\Sigma_k^* {V_k^*}^T + U_\perp^*\Sigma_\perp^* {V_\perp^*}^T$

Lemma

If $V = A^T U$, then

$$\tan\theta(V, V_k^*) \leq \frac{\sigma_{k+1}}{\sigma_k} \tan\theta(U, U_k^*)$$

Proof.

$$\begin{aligned}
\tan\theta(V, V_k^*) &= \|({V_\perp^*}^T V)({V_k^*}^T V)^{-1}\| \\
&= \|({V_\perp^*}^T A^T U)({V_k^*}^T A^T U)^{-1}\|
\end{aligned}$$

Classical Analysis of Power Iteration

Suppose $A = U^* \Sigma^* {V^*}^T = U_k^* \Sigma_k^* {V_k^*}^T + U_\perp^* \Sigma_\perp^* {V_\perp^*}^T$

Lemma

If $V = A^T U$, then

$$\tan\theta(V, V_k^*) \le \frac{\sigma_{k+1}}{\sigma_k} \tan\theta(U, U_k^*)$$

Proof.

$$\begin{aligned}
\tan\theta(V, V_k^*) &= \|({V_\perp^*}^T V)({V_k^*}^T V)^{-1}\| \\
&= \|({V_\perp^*}^T A^T U)({V_k^*}^T A^T U)^{-1}\| \\
&= \|(\Sigma_\perp {U_\perp}^T U)(\Sigma_k^* {U_k^*}^T U)^{-1}\|
\end{aligned}$$

Classical Analysis of Power Iteration

Suppose $A = U^* \Sigma^* {V^*}^T = U_k^* \Sigma_k^* {V_k^*}^T + U_\perp^* \Sigma_\perp^* {V_\perp^*}^T$

Lemma

If $V = A^T U$, then

$$\tan \theta(V, V_k^*) \leq \frac{\sigma_{k+1}}{\sigma_k} \tan \theta(U, U_k^*)$$

Proof.

$$\begin{aligned}
\tan \theta(V, V_k^*) &= \|({V_\perp^*}^T V)({V_k^*}^T V)^{-1}\| \\
&= \|({V_\perp^*}^T A^T U)({V_k^*}^T A^T U)^{-1}\| \\
&= \|(\Sigma_\perp {U_\perp}^T U)(\Sigma_k^* {U_k^*}^T U)^{-1}\| \\
&= \|\Sigma_\perp \cdot ({U_\perp}^T U)({U_k^*}^T U)^{-1} \cdot {\Sigma_k^*}^{-1}\|
\end{aligned}$$

Classical Analysis of Power Iteration

Suppose $A = U^* \Sigma^* {V^*}^T = U_k^* \Sigma_k^* {V_k^*}^T + U_\perp^* \Sigma_\perp^* {V_\perp^*}^T$

Lemma

If $V = A^T U$, then

$$\tan \theta(V, V_k^*) \leq \frac{\sigma_{k+1}}{\sigma_k} \tan \theta(U, U_k^*)$$

Proof.

$$\begin{aligned}
\tan \theta(V, V_k^*) &= \|({V_\perp^*}^T V)({V_k^*}^T V)^{-1}\| \\
&= \|({V_\perp^*}^T A^T U)({V_k^*}^T A^T U)^{-1}\| \\
&= \|(\Sigma_\perp {U_\perp}^T U)(\Sigma_k^* {U_k^*}^T U)^{-1}\| \\
&= \|\Sigma_\perp \cdot ({U_\perp}^T U)({U_k^*}^T U)^{-1} \cdot {\Sigma_k^*}^{-1}\| \\
&\leq \frac{\sigma_{k+1}}{\sigma_k} \tan \theta(U, U_k^*) \quad \square
\end{aligned}$$

Classical Analysis of Power Iteration

- The lemma above implies linear convergence of $\tan\theta(V_t, V_k^*)$ and $\tan\theta(U_t, U_k^*)$ in power iteration.
- After $T = O(\frac{\sigma_k}{\sigma_k - \sigma_{k+1}} \ln \frac{1}{\epsilon})$ iterations:

$$\|(I - U_T U_T^T) U_k^*\| = \sin\theta(U_T, U_k^*) \leq \tan\theta(U_T, U_k^*) \leq \epsilon$$

- Subspace angle is a handy tool for analyzing power iteration-type updates

..Back to Matrix Completion

Convergence of AltMinComplete: High Level Idea

- **Base Case:** Initialization U_0 falls into "basin of attraction" Z.
- **Inductive Case:**
  1. If iterate U_{t-1} is in Z, then V_t improves over U_{t-1} and is still in Z;
  2. Similarly if V_t is in Z, then U_t improves over V_t and is still in Z.

$$V_t \leftarrow \arg\min_{V \in \mathbb{R}^{n \times k}} F_\Omega(U_{t-1}, V),$$

$$U_t \leftarrow \arg\min_{U \in \mathbb{R}^{m \times k}} F_\Omega(U, V_t).$$

Local Convergence of AltMinComplete

$Y = U^* \Sigma^* V^*$, $U^* \in \mathbb{R}^{m \times k}$, $V^* \in \mathbb{R}^{n \times k}$ is μ-coherent
Let $\mu_1 = 4\mu\sqrt{k}\kappa(Y)$.

Lemma

Suppose

1. $\tan \theta(U^*, U) \leq 1/2$,
2. *U is μ_1-coherent,*
3. *Ω is a random subset of $[m] \times [n]$ of size $\tilde{O}(\kappa(Y)^4 k^{4.5} \mu^2 n)$.*

Then, update rule $V \leftarrow \arg\min F_\Omega(U, V)$ has the guarantee that:

1. $\tan \theta(V^*, V) \leq \frac{\tan \theta(U^*, U)}{4}$,
2. *V is μ_1-coherent.*

Local Convergence of AltMinComplete

$Y = U^* \Sigma^* V^*$, $U^* \in \mathbb{R}^{m \times k}$, $V^* \in \mathbb{R}^{n \times k}$ is μ-coherent
Let $\mu_1 = 4\mu\sqrt{k}\kappa(Y)$.

Lemma

Suppose

1. $\tan \theta(U^*, U) \leq 1/2$,
2. *U is μ_1-coherent,*

} *U is in Z* (items 1–2)

3. *Ω is a random subset of $[m] \times [n]$ of size $\tilde{O}(\kappa(Y)^4 k^{4.5} \mu^2 n)$.*

Then, update rule $V \leftarrow \arg\min F_\Omega(U, V)$ has the guarantee that:

1. $\tan \theta(V^*, V) \leq \frac{\tan \theta(U^*, U)}{4}$,
2. *V is μ_1-coherent.*

Local Convergence of AltMinComplete

$Y = U^*\Sigma^*V^*$, $U^* \in \mathbb{R}^{m\times k}$, $V^* \in \mathbb{R}^{n\times k}$ is μ-coherent
Let $\mu_1 = 4\mu\sqrt{k}\kappa(Y)$.

Lemma

Suppose

1. $\tan\theta(U^*, U) \le 1/2$,
2. *U is μ_1-coherent,*

} *U is in Z*

3. *Ω is a random subset of $[m]\times[n]$ of size $\tilde{O}(\kappa(Y)^4 k^{4.5}\mu^2 n)$.*

Then, update rule $V \leftarrow \arg\min F_\Omega(U, V)$ has the guarantee that:

1. $\tan\theta(V^*, V) \le \frac{\tan\theta(U^*,U)}{4}$,
2. *V is μ_1-coherent.*

} *V is in Z and V improves over U*

Proof Idea 1: Incoherence $\Rightarrow$ Low Noise with Low Sample Complexity

Claim

If U is μ_1-coherent, and Ω is a random subset of $[m] \times [n]$ of size $\tilde{O}(\kappa(Y)^4 k^{4.5} \mu^2 n)$, then least squares update $V = \arg\min F_{\Omega}(U, V)$ can be written as:

$$V = Y^T U + G, \quad \|G\| \leq \sigma_k^* \tan\theta(U, U^*)$$

Proof.

By standard matrix concentration. □

- The update is Approximate Power Iteration
- If μ is large ($O(\sqrt{\frac{n}{k}})$) then the sample complexity can be $O(n^2)$

Proof Idea 2: Angle Contraction

- Following the analysis in classical power iteration,

$$\tan\theta(V^*, V) = \|({V_\perp^*}^T V)({V^*}^T V)^{-1}\|$$

Proof Idea 2: Angle Contraction

- Following the analysis in classical power iteration,

$$\begin{aligned}\tan\theta(V^*, V) &= \|({V_\perp^*}^T V)({V^*}^T V)^{-1}\| \\ &\le \frac{\sigma_1({V_\perp^*}^T V)}{\sigma_k({V^*}^T V)}\end{aligned}$$

Proof Idea 2: Angle Contraction

- Following the analysis in classical power iteration,

$$\begin{aligned}\tan\theta(V^*, V) &= \|({V_\perp^*}^T V)({V^*}^T V)^{-1}\| \\ &\le \frac{\sigma_1({V_\perp^*}^T V)}{\sigma_k({V^*}^T V)} \\ &= \frac{\sigma_1({V_\perp^*}^T G)}{\sigma_k(\Sigma^* {V^*}^T V + {V_\perp^*}^T G)}\end{aligned}$$

Proof Idea 2: Angle Contraction

- Following the analysis in classical power iteration,

$$
\begin{aligned}
\tan\theta(V^*, V) &= \|({V_\perp^*}^T V)({V^*}^T V)^{-1}\| \\
&\leq \frac{\sigma_1({V_\perp^*}^T V)}{\sigma_k({V^*}^T V)} \\
&= \frac{\sigma_1({V_\perp^*}^T G)}{\sigma_k(\Sigma^* {V^*}^T V + {V_\perp^*}^T G)} \\
&\leq \frac{\|G\|}{\sigma_k^* - \|G\|}
\end{aligned}
$$

Proof Idea 2: Angle Contraction

- Following the analysis in classical power iteration,

$$
\begin{aligned}
\tan\theta(V^*, V) &= \|({V_\perp^*}^T V)({V^*}^T V)^{-1}\| \\
&\le \frac{\sigma_1({V_\perp^*}^T V)}{\sigma_k({V^*}^T V)} \\
&= \frac{\sigma_1({V_\perp^*}^T G)}{\sigma_k(\Sigma^* {V^*}^T V + {V_\perp^*}^T G)} \\
&\le \frac{\|G\|}{\sigma_k^* - \|G\|} \\
&\le \frac{\tan\theta(U^*, U)}{4}
\end{aligned}
$$

Proof Idea 3: Bounding the Coherence

Claim

The subspace spanned by V is μ_1-coherent.

Proof Idea.

With sufficiently many samples, $V \approx Y^T U$, thus

$$\text{span}(V) \approx \text{span}(Y^T) = \text{span}(V^*)$$

therefore μ_1-coherent. □

Technical Details Omitted

- Initialization: taking SVD on $P_\Omega(Y)$ ensures that U_0 falls into Z (basin of attraction)
  1. $\tan\theta(U_0, U^*) \leq \frac{1}{2}$,
  2. U_0 is μ_1-coherent.
- Shown by standard matrix concentration and serves as the inductive basis
- Recovery: Closeness of Subspace $\Rightarrow$ Closeness of Completed Matrix

Outline

Summary

- This paper rigorously analyzes alterating minimization for matrix completion, a well-known heuristic
- Key idea: the optimzation algorithm can be seen as Approximate Power Iteration (See also [Hardt'14])
- Key tool: subspace angles measuring the closeness between subspaces

Thank you!

Explicit Form of Update

Taking derivatives yields the following normal equation:

$$\underbrace{\begin{bmatrix} (\langle u_s, u_t \rangle_{\Omega_1}) & & \\ & \cdots & \\ & & (\langle u_s, u_t \rangle_{\Omega_n}) \end{bmatrix}}_{B_\Omega} \underbrace{\begin{bmatrix} v^1 \\ \cdots \\ v^n \end{bmatrix}}_{\text{vec}(V)} = \underbrace{\begin{bmatrix} (\langle u_s, u_t^* \rangle_{\Omega_1}) & & \\ & \cdots & \\ & & (\langle u_s, u_t^* \rangle_{\Omega_n}) \end{bmatrix}}_{C_\Omega} \underbrace{\begin{bmatrix} \Sigma^* {v^*}^1 \\ \cdots \\ \Sigma^* {v^*}^n \end{bmatrix}}_{vec(\Sigma V^*)}$$

where $\Omega_j = \{i \in [m] : (i,j) \in \Omega\}$, and $\langle x, y \rangle_S = \sum_{i \in S} x_i y_i$.

- Sanity Check: if $\Omega = [m] \times [n]$, then:
  - $B_\Omega = I$
  - $C_\Omega = \text{diag}(U^T U^*, \ldots, U^T U^*)$
  - $V \leftarrow V^* \Sigma^* {U^*}^T U = Y^T U$.
- Row-wise Form: for all $j \in [n]$, $(U^T P_j U) v^j = (U^T P_j U^*) \Sigma^* {v^*}^j$.