Clipped Matrix Completion [6]
(Exact Recovery Guarantee)

Takeshi Teshima

December 24, 2018

The University of Tokyo

Ceiling Effect

Ceiling effect

Measurement limitation that observations are clipped at a threshold at the time of observation.

- Ex. Questionnaire

- Too many people answer with "5" (max value).
  - ⇒ questionnaire may not be measuring the domain correctly.
- There may exist some more variation within "5".

Ceiling Effect in ML Benchmark

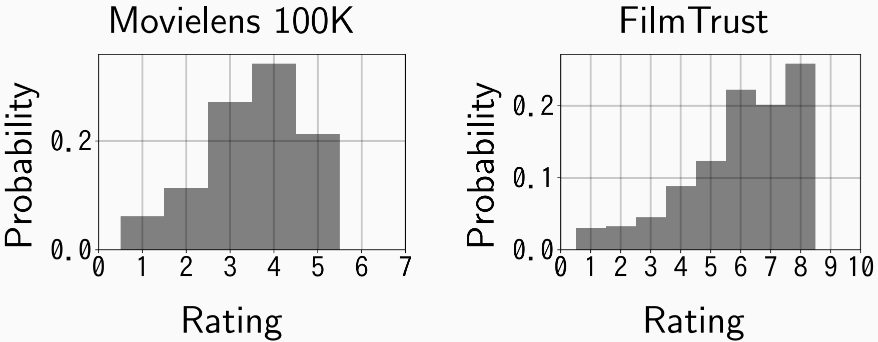


Figure 1: Histograms of benchmark recommendation systems data.

- Right-truncated histogram $\cdots$ typical for variable under ceiling effects.

Matrix Completion (MC)

- Recover matrix from missing, noise, etc.

	5		4
5		1	5
5	1	3	3

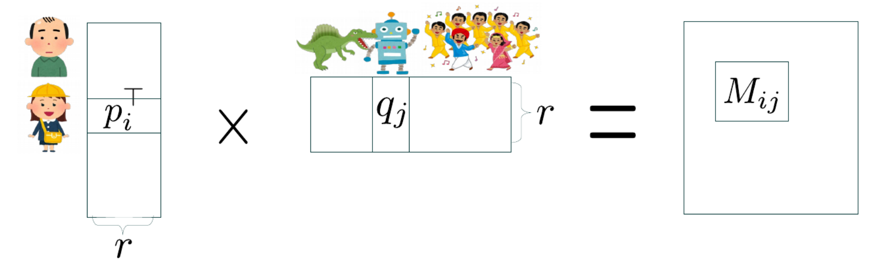
MC: example application

- Movie recommendation

	5		4
5		1	5
5	1	3	3

MC: Why would it ever work?

- Assume the matrix has a low rank.
- Principle of low-rank completion



- Low-rank = few latent factors dominate.
- Estimate the latent vectors, then one can impute values.

MC: The algorithm

We want to do

Rank minimization

$\min_{\mathbf{X} \in \mathbb{R}^{n_1 \times n_2}} \operatorname{rank}(\mathbf{X})$ s.t. ($\mathbf{X}$ complies with observation)

However, rank minimization is intractable. Instead:

Trace-norm minimization

$\min_{\mathbf{X} \in \mathbb{R}^{n_1 \times n_2}} \|\mathbf{X}\|_*$ s.t. ($\mathbf{X}$ complies with observation)

- Rank is count $\sum_k \mathbf{1}\{\sigma_k > 0\}$, Trace-norm is sum $\sum_k \sigma_k$.
- By [2], trace-norm minimization was given a guarantee that "$\widehat{\mathbf{M}}$ completely recovers $\mathbf{M}$."

Problem of Clipped Matrix Completion

- $\mathbf{M} \in \mathbb{R}^{n_1 \times n_2}$: the ground-truth matrix.
- $C \in \mathbb{R}$: the clipping threshold.
- $\mathrm{Clip}(\cdot) := \min\{C, \cdot\}$: the clipping operator (element-wise).
- $\mathbf{M}^{\mathrm{c}} := \mathrm{Clip}(\mathbf{M})$: full clipped matrix.
- Ω: the random set of observed indices (details later).

Problem (Clipped matrix completion (CMC))

Accurately recover $\mathbf{M}$ *from* $\mathbf{M}^{\mathrm{c}}_{\Omega} := \{M^{\mathrm{c}}_{ij}\}_{(i,j)\in\Omega}$ *and* C.

Illustration of CMC

4	7	4	7	4
0	3	6	15	12
4	6	2	2	0
2	6	7	16	12
8	13	6	9	4

(a) True matrix $\mathbf{M}$

	7	4	7	4
0	3	6	10	10
4	6	2	2	0
2	6	7	10	10
8	10	6	9	4

(b) Observed $\mathbf{M}^{\mathrm{c}}_{\Omega}$

4.0	7.0	4.0	7.0	4.0
-0.0	3.0	6.0	14.9	11.9
4.0	6.0	2.0	2.0	0.0
2.0	6.0	7.0	15.9	11.9
8.0	13.0	6.0	9.0	4.0

(c) Restored $\widehat{\mathbf{M}}$

Figure 2: The true low-rank matrix $\mathbf{M}$ has a distinct structure of large values. However, the observed data $\mathbf{M}^{\mathrm{c}}_{\Omega}$ is clipped at a predefined threshold $C = 10$. The goal of CMC is to restore $\mathbf{M}$ from the value of C and $\mathbf{M}^{\mathrm{c}}_{\Omega}$. The restored matrix $\widehat{\mathbf{M}}$ is an actual result of applying a proposed method (Fro-CMC).

Trace-norm minimization for CMC

$$\widehat{\mathbf{M}} \in \underset{\mathbf{X}\in\mathbb{R}^{n_1\times n_2}}{\arg\min} \|\mathbf{X}\|_{\mathrm{tr}} \text{ s.t. } \begin{cases} \mathcal{P}_{\Omega\setminus\mathcal{C}}(\mathbf{X}) = \mathcal{P}_{\Omega\setminus\mathcal{C}}(\mathbf{M}^{\mathrm{c}}), \\ \mathcal{P}_{\mathcal{C}}(\mathbf{M}^{\mathrm{c}}) \leq \mathcal{P}_{\mathcal{C}}(\mathbf{X}). \end{cases} \tag{1}$$

- Research question: can we prove $\widehat{\mathbf{M}} = \mathbf{M}$ (w.h.p.)?

Rough statement of the main theorem

Rough statement of the theorem

Assume

- $\mathbf{M}$ has nice properties (small information loss by clipping, incoherent, low-rank)
- observations are independent with probability p.
- p is large enough

Then, $\widehat{\mathbf{M}} = \mathbf{M}$ with high probability.

😄 CMC is feasible under a sufficient condition!

Coffee break

Quantities required for the statement

We need to define

- Coherence of $\mathbf{M}$
- Information loss of $\mathbf{M}$

Quantity 1: Coherence

Definition (Leverage scores [3])

Let $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2}$ have a skinny singular value decomposition $\mathbf{X} = \tilde{\mathbf{U}}\tilde{\mathbf{\Sigma}}\tilde{\mathbf{V}}^\top$. We define

$$\mu^{\mathrm{U}}(\mathbf{X}) := \max_{i \in [n_1]} \|\tilde{\mathbf{U}}_{i,\cdot}\|^2, \quad \mu^{\mathrm{V}}(\mathbf{X}) := \max_{j \in [n_2]} \|\tilde{\mathbf{V}}_{j,\cdot}\|^2,$$

where $\tilde{\mathbf{U}}_{i,\cdot}$ ($\tilde{\mathbf{V}}_{j,\cdot}$) is the i-th (resp. j-th) row of $\tilde{\mathbf{U}}$ (resp. $\tilde{\mathbf{V}}$).

- These are used to define the coherence of $\mathbf{M}$.

Quantity 1: Coherence (cont.)

Definition (Coherence and joint coherence [3])

Now the coherence of $\mathbf{M}$ is defined by

$$\mu_0 := \max\left\{\frac{n_1}{r}\mu^{\mathrm{U}}(\mathbf{M}), \frac{n_2}{r}\mu^{\mathrm{V}}(\mathbf{M})\right\}.$$

In addition, we define the following joint coherence:

$$\mu_1 := \sqrt{\frac{n_1 n_2}{r}}\|\mathbf{U}\mathbf{V}^\top\|_\infty.$$

What does coherence mean?

- Note

$$\begin{aligned}\|\mathbf{U}_{i,\cdot}\|^2 &= \sum_k \langle \mathbf{U}_{\cdot,k}, \boldsymbol{e}_i\rangle^2 \\ &= \|\sum_k \mathbf{U}_{\cdot,k}\langle \mathbf{U}_{\cdot,k}, \boldsymbol{e}_i\rangle\|^2 \\ &= \|\mathbf{U}\mathbf{U}^\top e_i\|^2 \\ &= \|\mathcal{P}_U(e_i)\|^2,\end{aligned}$$

where $U := \mathrm{Span}(u_1, \ldots, u_r)$.

- Therefore, a small coherence implies that there is no element in U that is "aligned" with e_i.
    - In other words, no element in U are too sparse.

What does coherence mean? (cont.)

- As a result, the components $\boldsymbol{u}_k \boldsymbol{v}_k^\top$ that $\mathbf{M}$ is composed of (as $\mathbf{M} = \mathbf{U\Sigma V}^\top$) cannot be sparse.

- The condition that coherence is small excludes the possibility that $\mathbf{M}$ is "spiky".
- The condition of $\mathbf{M}$ being low-rank is not enough to guarantee recovery.
  - e.g., a matrix with only the $1, 1$-entry being one and all others being zeros is also rank-one.
  - Incoherence condition (coherence being small) excludes such a possibility.

What does coherence mean? (cont.)

- Spiky matrix is possible when there is a sparse component $u_k v_k^\top$.
- Sparsity of $u_k v_k^\top$ means that there is a sparse u_k or v_k.
    - Let's say u_k is sparse.
    - Then, considering the normalization property of U (column vectors are normalized to norm-one), there must be a gathered mass in some dimension i of u_k.

Quantity 2: The information subspace

- We will define the information subspace T of $\mathbf{M}$.
- T is important because...
  1. $\mathbf{M} \in T$.
  2. T is used for explicit expression of $\partial \|\mathbf{M}\|_{\mathrm{tr}}$.
- Let $\mathbf{M} = \mathbf{U}\Sigma\mathbf{V}^\top$. Then $\mathbf{U}\mathbf{V}^\top \in T$ and $\partial \|\mathbf{M}\|_{\mathrm{tr}} = \{\mathbf{W} + \mathbf{U}\mathbf{V}^\top : \mathbf{W} \in T^\perp, \|\mathbf{W}\|_{\mathrm{op}} \leq 1\}$.
- $\mathbf{U}\mathbf{V}^\top \in T$.
- The feasibility of recovery depends upon the amount of information we have about T.

Quantity 2: The information subspace (cont.)

Definition (The information subspace of $\mathbf{M}$ [2])

- $\mathbf{M} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$: skinny singular value decomposition (SVD) ($\mathbf{U} \in \mathbb{R}^{n_1 \times r}, \boldsymbol{\Sigma} \in \mathbb{R}^{r \times r}$ and $\mathbf{V} \in \mathbb{R}^{n_2 \times r}$).
- Define the information subspace of $\mathbf{M}$ by
  $T := \mathrm{span}\big(\{\boldsymbol{u}_k \boldsymbol{y}^\top : k \in [r], \boldsymbol{y} \in \mathbb{R}^{n_2}\} \cup \{\boldsymbol{x}\boldsymbol{v}_k^\top : k \in [r], \boldsymbol{x} \in \mathbb{R}^{n_1}\}\big)$
- where $\boldsymbol{u}_k, \boldsymbol{v}_k$ are the k-th column of $\mathbf{U}$ and $\mathbf{V}$ (resp.).

- $\mathcal{P}_T, \mathcal{P}_{T^\perp}$: the projections onto T and $T^\perp$, resp.

(Key) quantity 3: Information loss

- Using T, we capture the information loss.
- The loss are measured in three different norms: $\|\cdot\|_{\mathrm{F}}$, $\|\cdot\|_{\mathrm{op}}$, and $\|\cdot\|_{\mathrm{tr}}$.
- To express the factor of clipping, we define an element-wise transformation $\mathcal{P}^*$.
- $\mathcal{P}^*$ describes the amount of information left after clipping

(Key) quantity 3: Information loss (cont.)

- In the theorem of exact recovery guarantee, we will assume: information loss is small and enough information is left by $\mathcal{P}^*$.

(Key) quantity 3: Information loss (cont.)

Definition (The information loss)

$$\rho_{\mathrm{F}} := \sup_{\mathbf{Z}\in T\setminus\{\mathbf{O}\}:\|\mathbf{Z}\|_{\mathrm{F}}\le\|\mathbf{U}\mathbf{V}^\top\|_{\mathrm{F}}} \frac{\|\mathcal{P}_T\mathcal{P}^*(\mathbf{Z})-\mathbf{Z}\|_{\mathrm{F}}}{\|\mathbf{Z}\|_{\mathrm{F}}},$$

$$\rho_{\infty} := \sup_{\mathbf{Z}\in T\setminus\{\mathbf{O}\}:\|\mathbf{Z}\|_{\infty}\le\|\mathbf{U}\mathbf{V}^\top\|_{\infty}} \frac{\|\mathcal{P}_T\mathcal{P}^*(\mathbf{Z})-\mathbf{Z}\|_{\infty}}{\|\mathbf{Z}\|_{\infty}},$$

$$\rho_{\mathrm{op}} := \sqrt{r}\mu_1 \left(\sup_{\substack{\mathbf{Z}\in T\setminus\{\mathbf{O}\}:\\ \|\mathbf{Z}\|_{\mathrm{op}}\le\sqrt{n_1 n_2}\|\mathbf{U}\mathbf{V}^\top\|_{\mathrm{op}}}} \frac{\|\mathcal{P}^*(\mathbf{Z})-\mathbf{Z}\|_{\mathrm{op}}}{\|\mathbf{Z}\|_{\mathrm{op}}} \right),$$

$$(\mathcal{P}^*(\mathbf{Z}))_{ij} = \begin{cases} Z_{ij} & \text{if } M_{ij} < C, \\ \max\{Z_{ij}, 0\} & \text{if } M_{ij} = C, \\ 0 & \text{otherwise.} \end{cases}$$

(Key) quantity 4: The importance of $\mathcal{B}$

- Another quantity $\nu_{\mathcal{B}}$ to measure the information loss is required.
- If this quantity is small, enough information of T may be left in non-clipped entries.

Definition (The importance of clipped entries)

Define

$$\nu_{\mathcal{B}} := \|\mathcal{P}_T \mathcal{P}_{\mathcal{B}} \mathcal{P}_T - \mathcal{P}_T\|_{\mathrm{op}},$$

where $\mathcal{B} := \{(i, j) : M_{ij} < C\}$.

Assumption on the observation scheme

Assumption (Assumption on the observation scheme)

- $p \in [0, 1]$, $k_0 := \lceil \log_2(2\sqrt{2}\sqrt{n_1 n_2 r}) \rceil$, *and* $q := 1 - (1-p)^{1/k_0}$.
- *For each* $k = 1, \ldots, k_0$,
  - $\Omega_k \subset [n_1] \times [n_2]$*: a random set of matrix indices such that*
  - *sampled according to* $\mathbb{P}((i,j) \in \Omega_k) = q$
  - $\{(i,j) \in \Omega_k\}$ *are all independent.*
- *Then,* Ω *was generated by* $\Omega = \bigcup_{k=1}^{k_0} \Omega_k$.

The need for Assumption 1 is technical [3].

The theorem

Theorem (Exact recovery guarantee for CMC)

Assume $\rho_{\mathrm{F}} < \frac{1}{2}, \rho_{\mathrm{op}} < \frac{1}{4}, \rho_{\infty} < \frac{1}{2}, \nu_{\mathcal{B}} < \frac{1}{2}$, *and Assumption 1 for some* $p \in [0, 1]$. *For simplicity of the statement, assume* $n_1, n_2 \geq 2$ *and* $p \geq \frac{1}{n_1 n_2}$. *If, additionally,*

$$p \geq \min\left\{1, c_\rho \max({\mu_1}^2, \mu_0) r f(n_1, n_2)\right\}$$

is satisfied, then...

The theorem (cont.)

Theorem (Exact recovery guarantee for CMC)

. . . the solution of Eq. (1) *is unique and equal to* M *with probability at least* $1-\delta$, *where*

$$
\begin{aligned}
c_\rho \quad &= \max\left\{\frac{24}{(1/2-\rho_{\mathrm{F}})^2}, \frac{8}{(1/4-\rho_{\mathrm{op}})^2}, \right.\\
&\qquad \left. \frac{8}{(1/2-\rho_\infty)^2}, \frac{8}{(1/2-\nu_{\mathcal{B}})^2}\right\},\\
f(n_1,n_2) &= \mathcal{O}\left(\frac{(n_1+n_2)(\log(n_1 n_2))^2}{n_1 n_2}\right),\\
\delta \quad &= \mathcal{O}\left(\frac{\log(n_1,n_2)}{n_1+n_2}\right)(n_1+n_2)^{-1}.
\end{aligned}
$$

Coffee break

Linear algebra and functional analysis

- Matrix inner product: $\langle \mathbf{X}, \mathbf{Y} \rangle = \sum_{ij} X_{ij} Y_{ij}$.
- Matrix norms:
  - $\|\mathbf{X}\|_{\mathrm{F}} := \sqrt{\langle \mathbf{X}, \mathbf{X} \rangle}$
  - $\|\mathbf{X}\|_{\mathrm{tr}} := \sum_k \sigma_k$ (σ_k: singular values)
  - $\|\mathbf{X}\|_{\mathrm{op}} := \sup_{\boldsymbol{v}:\|\boldsymbol{v}\|=1} \|\mathbf{X}\boldsymbol{v}\|$

- $\|\cdot\|_{\mathrm{tr}}$ and $\|\cdot\|_{\mathrm{op}}$ are dual.
    - $|\langle \mathbf{X}, \mathbf{Y}\rangle| \leq \|\mathbf{X}\|_{\mathrm{op}}\|\mathbf{Y}\|_{\mathrm{tr}}$
- Let $S \subset \mathbb{R}^{n_1 \times n_2}$: subspace. For each $\mathbf{Y} \in S$, there exists $\mathbf{X} \in S$ such that
    - $\|\mathbf{X}\|_{\mathrm{op}} = 1$
    - $\langle \mathbf{X}, \mathbf{Y}\rangle = \|\mathbf{X}\|_{\mathrm{op}}\|\mathbf{Y}\|_{\mathrm{tr}}$
- $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$: (skinny) SVD. Then,
    - $\mathbf{U}\mathbf{V}^\top \in T$.
    - $\partial\|\mathbf{X}\|_{\mathrm{tr}} = \{\mathbf{W} + \mathbf{U}\mathbf{V}^\top : \mathbf{W} \in T^\perp, \|\mathbf{W}\|_{\mathrm{op}} \leq 1\}$ (subgradients are (1) $\mathbf{U}\mathbf{V}^\top$ on T (2) small norm on $T^\perp$).

Notation

- $\omega_{ij} := \mathbf{1}\{(i,j) \in \Omega\}$, $\omega_{ij}^{(k)} := \mathbf{1}\{(i,j) \in \Omega_k\}$
- $\mathcal{R}_\Omega := \frac{1}{p}\mathcal{P}_\Omega, \mathcal{R}_\Omega^{\frac{1}{2}} := \frac{1}{\sqrt{p}}\mathcal{P}_\Omega, \mathcal{R}_\mathcal{C} := \frac{1}{p}\mathcal{P}_\mathcal{C}$, and $\mathcal{R}_{\Omega_k} := \frac{1}{q}\mathcal{P}_{\Omega_k}$
  - Note: $\mathcal{P}_{\Omega\backslash\mathcal{C}}, \mathcal{P}_\mathcal{C}, \mathcal{P}_\Omega, \mathcal{R}_\Omega, \mathcal{R}_\Omega^{\frac{1}{2}}$ are all self-adjoint.
- $\{\boldsymbol{e}_i\}_{i=1}^{n_1}, \{\boldsymbol{f}_j\}_{j=1}^{n_2}$: The standard bases of $\mathbb{R}^{n_1}$ and $\mathbb{R}^{n_2}$ (resp.).

Detailed Form of the Theorem

Theorem 8 is a simplified version of the following.

Theorem

Assume $\rho_{\mathrm{F}} < \frac{1}{2}, \rho_{\mathrm{op}} < \frac{1}{4}, \rho_{\infty} < \frac{1}{2}$*, and* $\nu_{\mathcal{B}} < \frac{1}{2}$*, and assume the independent and uniform sampling scheme as in Assumption 1. If for some* $\beta > \max\{1, 1/(4\log(n_1 n_2)), 1 + (\log 2/\log(n_1 n_2))\}$,

$$p \geq \min\left\{1, \max\left\{\frac{1}{n_1 n_2}, p_{\min}^{\mathrm{F}}, p_{\min}^{\mathrm{op},1}, p_{\min}^{\mathrm{op},2}, p_{\min}^{\infty}, p_{\min}^{\mathrm{main}}\right\}\right\} \quad (2)$$

where. . .

Detailed Form of the Theorem (cont.)

Theorem

$$
\begin{aligned}
p_{\min}^{\mathrm{F}} &= \frac{8k_0\mu_0\beta r}{(1/2-\rho_{\mathrm{F}})^2}\,\frac{(n_1+n_2)\log(n_1 n_2)}{n_1 n_2},\\
p_{\min}^{\mathrm{op},1} &= \frac{8k_0\beta}{3(1/4-\rho_{\mathrm{op}})^2}\,\frac{\log(n_1+n_2)}{\max(n_1,n_2)},\\
p_{\min}^{\mathrm{op},2} &= \frac{8k_0\beta r{\mu_1}^2}{3(1/4-\rho_{\mathrm{op}})^2}\,\frac{\max(n_1,n_2)\log(n_1+n_2)}{n_1 n_2},\\
p_{\min}^{\infty} &= \frac{8k_0\mu_0 r\beta}{3(1/2-\rho_{\infty})^2}\,\frac{(n_1+n_2)\log(n_1 n_2)}{n_1 n_2},\\
p_{\min}^{\mathrm{main}} &= \frac{8\beta r\mu_0}{3(1/2-\nu_{\mathcal{B}})^2}\,\frac{(n_1+n_2)\log(n_1 n_2)}{n_1 n_2},
\end{aligned}
$$

is satisfied, then...

Detailed Form of the Theorem (cont.)

Theorem

. . . the minimizer of Eq. (1) *is unique and equal to* $\mathbf{M}$ *with probability at least* $1 - k_0(e^{\frac{1}{4}}(n_1 n_2)^{-\beta} + 2(n_1 n_2)^{1-\beta} + (n_1 + n_2)^{1-\beta}) - 2(n_1 n_2)^{1-\beta}$.

Road map

1. We want to prove $\forall \widehat{\mathbf{M}} \neq \mathbf{M} : \|\widehat{\mathbf{M}}\|_{\mathrm{tr}} > \|\mathbf{M}\|_{\mathrm{tr}}$ w.h.p.
2. To do so, we use $\partial \|\mathbf{M}\|_{\mathrm{tr}}$.
   - Let $\mathbf{Z} \in \partial \|\mathbf{M}\|_{\mathrm{tr}}$, then we can do $\|\widehat{\mathbf{M}}\|_{\mathrm{tr}} \geq \langle \mathbf{Z}, \widehat{\mathbf{M}} - \mathbf{M} \rangle + \|\mathbf{M}\|_{\mathrm{tr}}$.
   - $\partial \|\mathbf{M}\|_{\mathrm{tr}}$ has a known expression using $\mathbf{U}\mathbf{V}^\top$ and $T^\perp$.
3. Then our objective becomes $\langle \mathbf{Z}, \widehat{\mathbf{M}} - \mathbf{M} \rangle > 0$.

- We actually use "approximate" subgradient $\mathbf{Y}$ for 3.
  - This $\mathbf{Y}$ is called the dual certificate.

Road map

Main lemma (informal)

If $\mathbf{Y}$: dual certificate exists, then $\|\widehat{\mathbf{M}}\|_{\mathrm{tr}} > \|\mathbf{M}\|_{\mathrm{tr}}$ unless $\widehat{\mathbf{M}} = \mathbf{M}$.

Existence of dual certificate w.h.p. (informal)

1. Construct a candidate $\mathbf{Y}$ by golfing scheme.
2. Prove that $\mathbf{Y}$ is actually a dual certificate.
    - based on concentration inequalities (Bernstein-type).

Definition of dual certificate

Definition (Dual certificate)

We say that $\mathbf{Y} \in \mathbb{R}^{n_1 \times n_2}$ is a dual certificate if it satisfies

1. $\mathbf{Y} \in \mathrm{range} \mathcal{P}_\Omega^*$
2. $\|\mathbf{U}\mathbf{V}^\top - \mathcal{P}_T \mathbf{Y}\|_{\mathrm{F}} \leq \frac{\sqrt{p}}{2\sqrt{2}}$
3. $\|\mathcal{P}_{T^\perp} \mathbf{Y}\|_{\mathrm{op}} < \frac{1}{2}$

By definition of $\mathcal{P}^*$, we have $\langle \mathcal{P}_\Omega(\mathbf{M}^{\mathrm{c}} - \mathbf{M}), \mathbf{Y} \rangle \geq 0$.

Main lemma

Given a dual certificate $\mathbf{Y}$ (and a little more condition), we can have the following result.

Lemma (Main lemma)

Assume that

1. *a dual certificate* $\mathbf{Y}$ *exists*
2. $\|\mathcal{P}_T \mathcal{P}_\Omega \mathcal{P}_\mathcal{B} \mathcal{P}_T - \mathcal{P}_T \mathcal{P}_\mathcal{B} \mathcal{P}_T\|_{\mathrm{op}} \leq \frac{1}{2} - \nu_\mathcal{B}$.

Then, the minimizer of trace-norm minimization (Eq. (1)*) is unique and is equal to* $\mathbf{M}$.

Proof (Lemma 13)

(Proof)

- Note that $\mathbf{M}$ is in the feasibility set of Eq. (1).
- Let $\widehat{\mathbf{M}} \in \mathbb{R}^{n_1 \times n_2}$ be another matrix (different from $\mathbf{M}$) in the feasibility set
- denote $\mathbf{H} := \widehat{\mathbf{M}} - \mathbf{M}$.
- Since the trace-norm is dual to the operator norm [5, Proposition 2.1],

Proof (Lemma 13) (cont.)

- there exists $\mathbf{W} \in T^{\perp}$ which satisfies $\|\mathbf{W}\|_{\mathrm{op}} = 1$ and $\langle \mathbf{W}, \mathcal{P}_{T^{\perp}} \mathbf{H} \rangle = \|\mathcal{P}_{T^{\perp}} \mathbf{H}\|_{\mathrm{tr}}$.
- It is also known that by using this $\mathbf{W}$, $\mathbf{U}\mathbf{V}^{\top} + \mathbf{W}$ is a subgradient of $\|\cdot\|_{\mathrm{tr}}$ at $\mathbf{M}$ [2].
- Therefore, we can calculate

Proof (Lemma 13) (cont.)

$$
\begin{aligned}
\|\widehat{\mathbf{M}}\|_{\mathrm{tr}} &= \|\mathbf{M}+\mathbf{H}\|_{\mathrm{tr}} \\
&\geq \|\mathbf{M}\|_{\mathrm{tr}} + \langle \mathbf{H}, \mathbf{U}\mathbf{V}^\top + \mathbf{W} \rangle \\
&= \|\mathbf{M}\|_{\mathrm{tr}} + \langle \mathbf{H}, \mathbf{U}\mathbf{V}^\top - \mathcal{P}_T \mathbf{Y} \rangle + \langle \mathbf{H}, \mathbf{W} - \mathcal{P}_{T^\perp} \mathbf{Y} \rangle + \langle \mathbf{H}, \mathbf{Y} \rangle \\
&\geq \|\mathbf{M}\|_{\mathrm{tr}} + \langle \mathcal{P}_T \mathbf{H}, \mathbf{U}\mathbf{V}^\top - \mathcal{P}_T \mathbf{Y} \rangle + \langle \mathcal{P}_{T^\perp} \mathbf{H}, \mathbf{W} - \mathcal{P}_{T^\perp} \mathbf{Y} \rangle + \langle \mathbf{H}, \mathbf{Y} \rangle,
\end{aligned}
\tag{3}
$$

where we used the self-adjointness of the projection operators, as well as $\mathbf{U}\mathbf{V}^\top \in T$.

From here, we will bound each term in the rightmost equation of Eq. (3).

[Lower-bounding $\langle \mathbf{H}, \mathbf{Y} \rangle$ with 0]

Proof (Lemma 13) (cont.)

We have $\langle \mathbf{H}, \mathbf{Y} \rangle \geq \langle \mathbf{M}^{\mathrm{c}} - \mathbf{M}, \mathbf{Y} \rangle \geq 0$, since

$$\langle \mathbf{H}, \mathbf{Y} \rangle - \langle \mathbf{M}^{\mathrm{c}} - \mathbf{M}, \mathbf{Y} \rangle = \langle \widehat{\mathbf{M}} - \mathbf{M}^{\mathrm{c}}, \mathbf{Y} \rangle = \langle \mathcal{P}_{\Omega}(\widehat{\mathbf{M}} - \mathbf{M}^{\mathrm{c}}), \mathbf{Y} \rangle \geq 0$$

can be seen by considering the signs element-wise.

[Lower-bounding $\langle \mathcal{P}_{T^\perp} \mathbf{H}, \mathcal{P}_{T^\perp}(\mathbf{W} - \mathbf{Y}) \rangle$ with $\|\mathcal{P}_{T^\perp} \mathbf{H}\|_{\mathrm{F}}$]

We have

$$\begin{aligned}
\langle \mathcal{P}_{T^\perp} \mathbf{H}, \mathcal{P}_{T^\perp}(\mathbf{W} - \mathbf{Y}) \rangle &= \|\mathcal{P}_{T^\perp} \mathbf{H}\|_{\mathrm{tr}} - \langle \mathcal{P}_{T^\perp} \mathbf{H}, \mathcal{P}_{T^\perp} \mathbf{Y} \rangle \\
&\geq (1 - \|\mathcal{P}_{T^\perp} \mathbf{Y}\|_{\mathrm{op}}) \|\mathcal{P}_{T^\perp} \mathbf{H}\|_{\mathrm{tr}} \\
&\geq (1 - \|\mathcal{P}_{T^\perp} \mathbf{Y}\|_{\mathrm{op}}) \|\mathcal{P}_{T^\perp} \mathbf{H}\|_{\mathrm{F}}.
\end{aligned}$$

Proof (Lemma 13) (cont.)

[Lower-bounding $\langle \mathcal{P}_T \mathbf{H}, \mathbf{U}\mathbf{V}^\top - \mathcal{P}_T \mathbf{Y} \rangle$ with $\|\mathcal{P}_{T^\perp} \mathbf{H}\|_{\mathrm{F}}$]
Now note

$$\langle \mathcal{P}_T \mathbf{H}, \mathbf{U}\mathbf{V}^\top - \mathcal{P}_T \mathbf{Y} \rangle \geq -\|\mathcal{P}_T \mathbf{H}\|_{\mathrm{F}} \|\mathbf{U}\mathbf{V}^\top - \mathcal{P}_T \mathbf{Y}\|_{\mathrm{F}},$$

- We go on to upper-bound $\|\mathcal{P}_T \mathbf{H}\|_{\mathrm{F}}$ by $\|\mathcal{P}_{T^\perp} \mathbf{H}\|_{\mathrm{F}}$.
- Note $0 = \|\mathcal{R}_\Omega^{\frac{1}{2}} \mathcal{P}_{\mathcal{B}} \mathbf{H}\|_{\mathrm{F}} \geq \|\mathcal{R}_\Omega^{\frac{1}{2}} \mathcal{P}_{\mathcal{B}} \mathcal{P}_T \mathbf{H}\|_{\mathrm{F}} - \|\mathcal{R}_\Omega^{\frac{1}{2}} \mathcal{P}_{\mathcal{B}} \mathcal{P}_{T^\perp} \mathbf{H}\|_{\mathrm{F}}$.

Proof (Lemma 13) (cont.)

- Therefore, $\|\mathcal{R}_\Omega^{\frac{1}{2}}\mathcal{P}_\mathcal{B}\mathcal{P}_T\mathbf{H}\|_\mathrm{F} \geq \|\mathcal{R}_\Omega^{\frac{1}{2}}\mathcal{P}_\mathcal{B}\mathcal{P}_{T^\perp}\mathbf{H}\|_\mathrm{F}$.

Now

$$\begin{aligned}
&\|\mathcal{R}_\Omega^{\frac{1}{2}}\mathcal{P}_\mathcal{B}\mathcal{P}_T\mathbf{H}\|_\mathrm{F}^2 \\
&= \langle \mathcal{R}_\Omega\mathcal{P}_\mathcal{B}\mathcal{P}_T\mathbf{H}, \mathcal{P}_\mathcal{B}\mathcal{P}_T\mathbf{H}\rangle \\
&= \langle \mathcal{R}_\Omega\mathcal{P}_\mathcal{B}\mathcal{P}_T\mathbf{H}, \mathcal{P}_T\mathbf{H}\rangle \\
&= \|\mathcal{P}_T\mathbf{H}\|_\mathrm{F}^2 + \langle \mathcal{P}_T(\mathcal{R}_\Omega\mathcal{P}_\mathcal{B}\mathcal{P}_T - \mathcal{P}_T)\mathcal{P}_T\mathbf{H}, \mathcal{P}_T\mathbf{H}\rangle \\
&\geq \|\mathcal{P}_T\mathbf{H}\|_\mathrm{F}^2 - \|\mathcal{P}_T\mathcal{R}_\Omega\mathcal{P}_\mathcal{B}\mathcal{P}_T - \mathcal{P}_T\|_\mathrm{op}\|\mathcal{P}_T\mathbf{H}\|_\mathrm{F}^2 \\
&\geq \|\mathcal{P}_T\mathbf{H}\|_\mathrm{F}^2 \\
&\quad - (\|\mathcal{P}_T\mathcal{R}_\Omega\mathcal{P}_\mathcal{B}\mathcal{P}_T - \mathcal{P}_T\mathcal{P}_\mathcal{B}\mathcal{P}_T\|_\mathrm{op} + \|\mathcal{P}_T\mathcal{P}_\mathcal{B}\mathcal{P}_T - \mathcal{P}_T\|_\mathrm{op})\|\mathcal{P}_T\mathbf{H}\|_\mathrm{F}^2 \\
&\geq \left(1 - \left(\frac{1}{2} - \nu_\mathcal{B}\right) - \nu_\mathcal{B}\right)\|\mathcal{P}_T\mathbf{H}\|_\mathrm{F}^2 \\
&= \frac{1}{2}\|\mathcal{P}_T\mathbf{H}\|_\mathrm{F}^2.
\end{aligned}$$

Proof (Lemma 13) (cont.)

On the other hand,

$$\|\mathcal{R}_{\Omega}^{\frac{1}{2}}\mathcal{P}_{\mathcal{B}}\mathcal{P}_{T^{\perp}}\mathbf{H}\|_{\mathrm{F}} \leq \frac{1}{\sqrt{p}}\|\mathcal{P}_{T^{\perp}}\mathbf{H}\|_{\mathrm{F}}.$$

Therefore, we have

$$-\|\mathcal{P}_{T}\mathbf{H}\|_{\mathrm{F}} \geq -\sqrt{\frac{2}{p}}\|\mathcal{P}_{T^{\perp}}\mathbf{H}\|_{\mathrm{F}}.$$

[Finishing the proof]

Proof (Lemma 13) (cont.)

Now we are ready to continue the calculation of Eq. (3) as

$$\begin{aligned}
\|\widehat{\mathbf{M}}\|_{\mathrm{tr}} &\geq \|\mathbf{M}\|_{\mathrm{tr}} - \|\mathbf{U}\mathbf{V}^\top - \mathcal{P}_T\mathbf{Y}\|_{\mathrm{F}}\|\mathcal{P}_T\mathbf{H}\|_{\mathrm{F}} + (1 - \|\mathcal{P}_{T^\perp}\mathbf{Y}\|_{\mathrm{op}})\|\mathcal{P}_{T^\perp}\mathbf{H}\|_{\mathrm{F}} + 0 \\
&\geq \|\mathbf{M}\|_{\mathrm{tr}} - \|\mathbf{U}\mathbf{V}^\top - \mathcal{P}_T\mathbf{Y}\|_{\mathrm{F}}\sqrt{\frac{2}{p}}\|\mathcal{P}_{T^\perp}\mathbf{H}\|_{\mathrm{F}} + (1 - \|\mathcal{P}_{T^\perp}\mathbf{Y}\|_{\mathrm{op}})\|\mathcal{P}_{T^\perp}\mathbf{H}\|_{\mathrm{F}} \\
&\geq \|\mathbf{M}\|_{\mathrm{tr}} + \left(1 - \|\mathcal{P}_{T^\perp}\mathbf{Y}\|_{\mathrm{op}} - \|\mathbf{U}\mathbf{V}^\top - \mathcal{P}_T\mathbf{Y}\|_{\mathrm{F}}\sqrt{\frac{2}{p}}\right)\|\mathcal{P}_{T^\perp}\mathbf{H}\|_{\mathrm{F}} \\
&> \|\mathbf{M}\|_{\mathrm{tr}} + \left(1 - \frac{1}{2} - \frac{1}{2}\right)\|\mathcal{P}_{T^\perp}\mathbf{H}\|_{\mathrm{F}} \\
&= \|\mathbf{M}\|_{\mathrm{tr}}.
\end{aligned}$$

Therefore, $\mathbf{M}$ is the unique minimizer of Eq. (1). □

What's next?

From here, we will

- construct a candidate of dual certificate $\mathbf{Y}$ by golfing scheme.
- and prove that $\mathbf{Y}$ is actually a dual certificate.

Coffee break

Strategy

- Find a candidate of dual certificate $\mathbf{Y}$ by golfing scheme.
  - Golfing scheme is like a theoretical SGD.
- We then prove that $\mathbf{Y}$ is actually a dual certificate.
  - The proof uses concentration inequalities and information loss.

Definition of the generalized golfing scheme

Definition (Generalized golfing scheme)

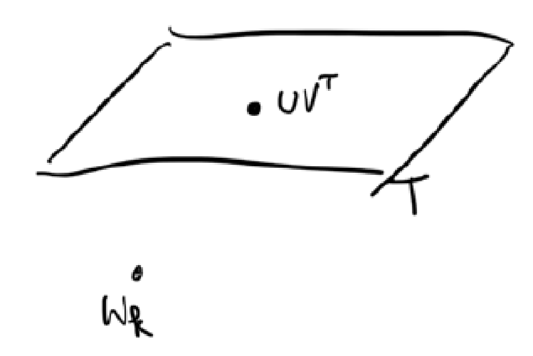
We recursively define $\{\mathbf{W}_k\}_{k=0}^{k_0}$ by

$$\begin{cases}\mathbf{W}_0 : & = \mathbf{O} \\ \Delta_k : & = \mathbf{U}\mathbf{V}^\top - \mathbf{W}_k \\ \mathbf{W}_k : & = \mathbf{W}_{k-1} + \mathcal{R}^*_{\Omega_k}\mathcal{P}_T\Delta_{k-1} = \mathbf{U}\mathbf{V}^\top - (\mathcal{I} - \mathcal{R}^*_{\Omega_k}\mathcal{P}_T)\Delta_{k-1}\end{cases}$$

where $\mathcal{R}^*_{\Omega_k}(\cdot) := \mathcal{R}_{\Omega_k}(\mathcal{P}^*(\cdot))$, and define $\mathbf{Y} := \mathbf{W}_{k_0}$.

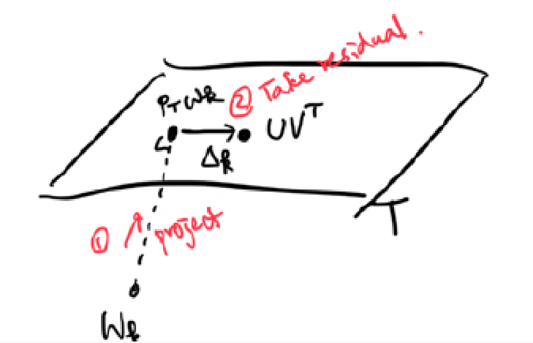
- The idea: next slide

The idea of golfing scheme



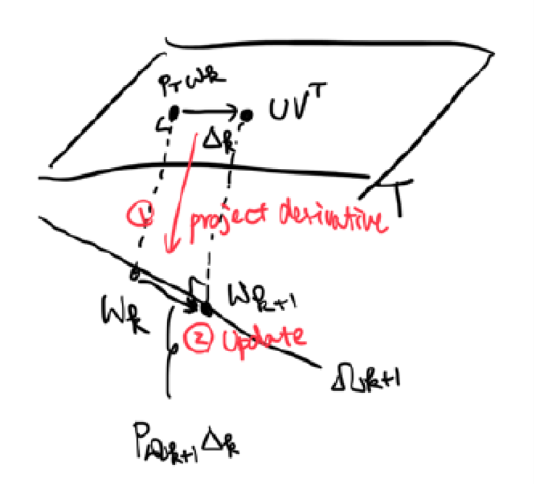
- $\mathbf{W}_k := \mathbf{W}_{k-1} + \mathcal{R}^*_{\Omega_k}\mathcal{P}_T\Delta_{k-1} \quad (\Delta_k := \mathbf{U}\mathbf{V}^\top - \mathbf{W}_k)$
- Goal: approximate $\mathbf{U}\mathbf{V}^\top$ on T while keeping small $\|\mathcal{P}_{T^\perp} \cdot \|_{\mathrm{op}}$.

The idea of golfing scheme



- $\mathbf{W}_k := \mathbf{W}_{k-1} + \mathcal{R}^*_{\Omega_k}\mathcal{P}_T\Delta_{k-1} \quad (\Delta_k := \mathbf{U}\mathbf{V}^\top - \mathbf{W}_k)$

The idea of golfing scheme



- $\mathbf{W}_k := \mathbf{W}_{k-1} + \mathcal{R}^*_{\Omega_k}\mathcal{P}_T\Delta_{k-1} \quad (\Delta_k := \mathbf{U}\mathbf{V}^\top - \mathbf{W}_k)$

$\mathbf{Y}$ is actually a dual certificate.

Lemma ($\mathbf{Y}$ is a dual certificate)

If for some
$\beta > \max\{1, 1/(4\log(n_1 n_2)), 1 + (\log 2/\log(n_1 n_2))\}$,

$$p \geq \min\left\{1, \max\left\{\frac{1}{n_1 n_2}, p^{\mathrm{F}}_{\min}, p^{\mathrm{op},1}_{\min}, p^{\mathrm{op},2}_{\min}, p^{\infty}_{\min}\right\}\right\} \tag{4}$$

is satisfied, then the matrix $\mathbf{Y} \in \mathbb{R}^{n_1 \times n_2}$ *defined by Def. 14 is a dual certificate (Def. 12) with probability at least*
$1 - k_0(e^{\frac{1}{4}}(n_1 n_2)^{-\beta} + 2(n_1 n_2)^{1-\beta} + (n_1 + n_2)^{1-\beta})$.

Proof ($\mathbf{Y}$ is a dual certificate)

(Proof)

- By construction, we have $\mathbf{Y} \in \mathrm{range}\mathcal{P}_{\Omega}^{*}$.
- From here, we show the other two conditions of the dual certificate.
- In the proof, we will use Prop. 1 below.

Proof ($\mathbf{Y}$ is a dual certificate) (cont.)

Prop.

$$\rho_{\mathrm{op}} \geq \|\mathbf{U}\mathbf{V}^\top\|_\infty \left(\sup_{\mathbf{Z} \in T \setminus \{\mathbf{O}\} : \|\mathbf{Z}\|_\infty \leq \|\mathbf{U}\mathbf{V}^\top\|_\infty} \frac{\|\mathcal{P}^*\mathbf{Z} - \mathbf{Z}\|_{\mathrm{op}}}{\|\mathbf{Z}\|_\infty} \right)$$

Also, by concentration inequalities, we can prove

Proof ($\mathbf{Y}$ is a dual certificate) (cont.)

Concentration inequalities

1. $\|\mathcal{P}_T\mathcal{P}^*\mathcal{P}_T\Delta_{k-1} - \mathcal{P}_T\mathcal{R}^*_{\Omega_k}\mathcal{P}_T\Delta_{k-1}\|_{\mathrm{F}} \le \left(\frac{1}{2} - \rho_{\mathrm{F}}\right)\|\mathcal{P}_T\Delta_{k-1}\|_{\mathrm{F}}$
2. $\|(\mathcal{R}^*_{\Omega_k} - \mathcal{P}^*)(\mathcal{P}_T\Delta_{k-1})\|_{\mathrm{op}} \le \left(\frac{1}{4} - \rho_{\mathrm{op}}\right)\frac{1}{\|\mathbf{U}\mathbf{V}^\top\|_\infty}\|\mathcal{P}_T\Delta_{k-1}\|_\infty$
3. $\|(\mathcal{P}_T\mathcal{R}^*_{\Omega_k}\mathcal{P}_T - \mathcal{P}_T\mathcal{P}^*\mathcal{P}_T)(\mathcal{P}_T\Delta_{k-1})\|_\infty \le \left(\frac{1}{2} - \rho_\infty\right)\|\mathcal{P}_T\Delta_{k-1}\|_\infty$

hold with the probability specified in the statement of the theorem.

We trust these inequalities here.

[Upper bounding $\|\mathbf{U}\mathbf{V}^\top - \mathcal{P}_T\mathbf{Y}\|_{\mathrm{F}}$]

Proof (Y is a dual certificate) (cont.)

We confirm by recursion that if Eq. (8) holds for all $k \in [k_0]$, then we have $\|\mathcal{P}_T\Delta_k\|_{\mathrm{F}} \leq \|\mathbf{U}\mathbf{V}^\top\|_{\mathrm{F}}$. First, we have $\|\mathcal{P}_T\Delta_0\|_{\mathrm{F}} = \|\mathbf{U}\mathbf{V}^\top\|_{\mathrm{F}}$. Second, if $\|\mathcal{P}_T\Delta_{k-1}\|_{\mathrm{F}} \leq \|\mathbf{U}\mathbf{V}^\top\|_{\mathrm{F}}$, then

$$
\begin{aligned}
\|\mathcal{P}_T\Delta_k\|_{\mathrm{F}} &= \|\mathcal{P}_T(\mathbf{U}\mathbf{V}^\top - \mathbf{W}_k)\|_{\mathrm{F}} \\
&= \|\mathbf{U}\mathbf{V}^\top - \mathcal{P}_T\mathbf{W}_{k-1} - \mathcal{P}_T\mathcal{R}^*_{\Omega_k}\mathcal{P}_T\Delta_{k-1}\|_{\mathrm{F}} \\
&\leq \|\mathcal{P}_T\Delta_{k-1} - \mathcal{P}_T\mathcal{P}^*\mathcal{P}_T\Delta_{k-1}\|_{\mathrm{F}} + \|\mathcal{P}_T\mathcal{P}^*\mathcal{P}_T\Delta_{k-1} - \mathcal{P}_T\mathcal{R}^*_{\Omega_k}\mathcal{P}_T\Delta_{k-1}\|_{\mathrm{F}} \\
&\leq \rho_{\mathrm{F}}\|\mathcal{P}_T\Delta_{k-1}\|_{\mathrm{F}} + \left(\frac{1}{2} - \rho_{\mathrm{F}}\right)\|\mathcal{P}_T\Delta_{k-1}\|_{\mathrm{F}} \\
&= \frac{1}{2}\|\mathcal{P}_T\Delta_{k-1}\|_{\mathrm{F}} \leq \|\mathbf{U}\mathbf{V}^\top\|_{\mathrm{F}}
\end{aligned}
$$

Proof (Y is a dual certificate) (cont.)

Now, by the same recursion formula, we can show $\|\mathcal{P}_T\Delta_{k_0}\|_{\mathrm{F}} \leq \left(\frac{1}{2}\right)^{k_0}\|\mathcal{P}_T\Delta_0\|_{\mathrm{F}}$. Therefore, under the condition Eq. (4), by the union bound, we have Eq. (8) for all $k \in [k_0]$ with probability at least $1 - k_0 e^{\frac{1}{4}}(n_1 n_2)^{-\beta}$ and

$$\begin{aligned}
\|\mathbf{U}\mathbf{V}^\top - \mathcal{P}_T\mathbf{Y}\|_{\mathrm{F}} = \|\mathcal{P}_T\Delta_{k_0}\|_{\mathrm{F}} &\leq \left(\frac{1}{2}\right)^{k_0}\|\mathcal{P}_T\Delta_0\|_{\mathrm{F}} \\
&\leq \sqrt{\frac{1}{n_1 n_2 r}}\frac{1}{2\sqrt{2}}\|\mathbf{U}\mathbf{V}^\top\|_{\mathrm{F}} \\
&\leq \sqrt{\frac{p}{r}}\frac{1}{2\sqrt{2}}\|\mathbf{U}\mathbf{V}^\top\|_{\mathrm{F}} \\
&= \sqrt{\frac{p}{r}}\frac{1}{2\sqrt{2}}\sqrt{r}
\end{aligned}$$

Proof ($\mathbf{Y}$ is a dual certificate) (cont.)

because $k_0 = \lceil \log_2(2\sqrt{2}\sqrt{n_1 n_2 r}) \rceil$, where we used $\frac{1}{n_1 n_2} \leq p$.

[Upper bounding $\|\mathcal{P}_{T^\perp}\mathbf{Y}\|_{\mathrm{op}}$]

By a similar argument of recursion as above with Eq. (**??**) in Lemma **??**, we can prove that for all $k \in [k_0]$, $\|\mathcal{P}_T\Delta_k\|_\infty \leq \|\mathbf{U}\mathbf{V}^\top\|_\infty$ and $\|\mathcal{P}_T\Delta_k\|_\infty \leq \frac{1}{2}\|\mathcal{P}_T\Delta_{k-1}\|_\infty$, with probability at least $1 - k_0 2(n_1 n_2)^{1-\beta}$ under the condition Eq. (4). Similarly, with Eq. **??** in Lemma **??** and using Prop. 1, we obtain for all $k \in [k_0]$, $\|(\mathcal{R}^*_{\Omega_k} - \mathcal{I})(\mathcal{P}_T\Delta_{k-1})\|_{\mathrm{op}} \leq \frac{1}{4\|\mathbf{U}\mathbf{V}^\top\|_\infty}\|\mathcal{P}_T\Delta_{k-1}\|_\infty$, with

Proof (Y is a dual certificate) (cont.)

probability at least $1 - k_0(n_1 + n_2)^{1-\beta}$ under the condition Eq. (4). Therefore, under the condition Eq. (4), with probability at least $1 - k_0(2(n_1 n_2)^{1-\beta} + (n_1 + n_2)^{1-\beta})$, we have

$$
\begin{aligned}
\|\mathcal{P}_{T^\perp} Y\|_{\mathrm{op}} &= \left\| \mathcal{P}_{T^\perp} \sum_{k=1}^{k_0} \mathcal{R}^*_{\Omega_k} \mathcal{P}_T(\Delta_{k-1}) \right\|_{\mathrm{op}} \\
&\leq \sum_{k=1}^{k_0} \|\mathcal{P}_{T^\perp} \mathcal{R}^*_{\Omega_k} \mathcal{P}_T(\Delta_{k-1})\|_{\mathrm{op}} \\
&= \sum_{k=1}^{k_0} \|\mathcal{P}_{T^\perp} \mathcal{R}^*_{\Omega_k} \mathcal{P}_T(\Delta_{k-1}) - \mathcal{P}_{T^\perp} \mathcal{P}_T \Delta_{k-1}\|_{\mathrm{op}}
\end{aligned}
$$

Proof ($\mathbf{Y}$ is a dual certificate) (cont.)

$$
\begin{aligned}
&\le \sum_{k=1}^{k_0} \|(\mathcal{R}_{\Omega_k}^* - \mathcal{I})(\mathcal{P}_T \Delta_{k-1})\|_{\text{op}} \\
&\le \sum_{k=1}^{k_0} \frac{1}{4\|\mathbf{U}\mathbf{V}^\top\|_\infty} \|\mathcal{P}_T \Delta_{k-1}\|_\infty \\
&\le \sum_{k=1}^{k_0} 2^{-k+1} \frac{1}{4\|\mathbf{U}\mathbf{V}^\top\|_\infty} \|\mathcal{P}_T \Delta_0\|_\infty \\
&< \frac{1}{2}.
\end{aligned}
$$

By taking the union bound, we have the lemma.

□

Lemma Used in the Proof ($\mathbf{Y}$ is a dual certificate)

In the recursion formula, we have used the following property yielding from the definition of ρ_{op} (Def. 5).

Prop.

$$\rho_{\mathrm{op}} \geq \|\mathbf{U}\mathbf{V}^\top\|_\infty \left(\sup_{\mathbf{Z} \in T \setminus \{\mathbf{O}\} : \|\mathbf{Z}\|_\infty \leq \|\mathbf{U}\mathbf{V}^\top\|_\infty} \frac{\|\mathcal{P}^* \mathbf{Z} - \mathbf{Z}\|_{\mathrm{op}}}{\|\mathbf{Z}\|_\infty} \right)$$

Proof (Lemma Used in the Proof)

(Proof)

- We have $\{\mathbf{Z} \in T : \|\mathbf{Z}\|_\infty \leq \|\mathbf{U}\mathbf{V}^\top\|_\infty\} \subset \{\mathbf{Z} \in T : \|\mathbf{Z}\|_{\mathrm{op}} \leq \sqrt{n_1 n_2}\|\mathbf{U}\mathbf{V}^\top\|_{\mathrm{op}}\}$,
- because if $\|\mathbf{Z}\|_\infty \leq \|\mathbf{U}\mathbf{V}^\top\|_\infty$, then we can obtain $\|\mathbf{Z}\|_{\mathrm{op}} \leq \sqrt{n_1 n_2}\|\mathbf{Z}\|_\infty \leq \sqrt{n_1 n_2}\|\mathbf{U}\mathbf{V}^\top\|_\infty \leq \sqrt{n_1 n_2}\|\mathbf{U}\mathbf{V}^\top\|_{\mathrm{op}}$.
- (Here, we used $\|\mathbf{Z}\|_{\mathrm{op}} \leq \sqrt{n_1 n_2}\|\mathbf{Z}\|_\infty$ and $\|\mathbf{Z}\|_\infty \leq \|\mathbf{Z}\|_{\mathrm{op}}$).

Proof (Lemma Used in the Proof) (cont.)

- Therefore,

$$\begin{aligned}
\rho_{\mathrm{op}} &= \sqrt{r}\mu_1 \left(\sup_{\mathbf{Z}\in T\setminus\{\mathbf{O}\}:\|\mathbf{Z}\|_{\mathrm{op}}\le\sqrt{n_1 n_2}\|\mathbf{U}\mathbf{V}^\top\|_{\mathrm{op}}} \frac{\|\mathcal{P}^*\mathbf{Z}-\mathbf{Z}\|_{\mathrm{op}}}{\|\mathbf{Z}\|_{\mathrm{op}}} \right) \\
&= \sqrt{n_1 n_2}\|\mathbf{U}\mathbf{V}^\top\|_\infty \left(\sup_{\mathbf{Z}\in T\setminus\{\mathbf{O}\}:\|\mathbf{Z}\|_{\mathrm{op}}\le\sqrt{n_1 n_2}\|\mathbf{U}\mathbf{V}^\top\|_{\mathrm{op}}} \frac{\|\mathcal{P}^*\mathbf{Z}-\mathbf{Z}\|_{\mathrm{op}}}{\|\mathbf{Z}\|_{\mathrm{op}}} \right) \\
&\ge \|\mathbf{U}\mathbf{V}^\top\|_\infty \left(\sup_{\mathbf{Z}\in T\setminus\{\mathbf{O}\}:\|\mathbf{Z}\|_\infty\le\|\mathbf{U}\mathbf{V}^\top\|_\infty} \frac{\|\mathcal{P}^*\mathbf{Z}-\mathbf{Z}\|_{\mathrm{op}}}{\frac{1}{\sqrt{n_1 n_2}}\|\mathbf{Z}\|_{\mathrm{op}}} \right) \\
&\ge \|\mathbf{U}\mathbf{V}^\top\|_\infty \left(\sup_{\mathbf{Z}\in T\setminus\{\mathbf{O}\}:\|\mathbf{Z}\|_\infty\le\|\mathbf{U}\mathbf{V}^\top\|_\infty} \frac{\|\mathcal{P}^*\mathbf{Z}-\mathbf{Z}\|_{\mathrm{op}}}{\|\mathbf{Z}\|_\infty} \right).
\end{aligned}$$

□

Coffee break

This section

- Many concentration inequalities are shown for reference.
- In this talk, only the vector Bernstein inequality will be used.

Matrix Bernstein inequality

Theorem (Matrix Bernstein inequality [7])

Let $\{\mathbf{Z}_k\}_{k=1}^{L}$ *be independent random matrices with dimensions* $d_1 \times d_2$*. If* $\mathbb{E}(\mathbf{Z}_k) = \mathbf{O}$ *and* $\|\mathbf{Z}_k\|_{\mathrm{op}} \leq R$ *(a.s.), then define* $\sigma^2 :=$
$\max\left\{\left\|\sum_{k=1}^{L} \mathbb{E}(\mathbf{Z}_k^\top \mathbf{Z}_k)\right\|_{\mathrm{op}}, \left\|\sum_{k=1}^{L} \mathbb{E}(\mathbf{Z}_k \mathbf{Z}_k^\top)\right\|_{\mathrm{op}}\right\}$.
Then for all $t \in \left[0, \frac{\sigma^2}{R}\right]$,

$$\mathcal{P}\left\{\left\|\sum_{k=1}^{L} \mathbf{Z}_k\right\|_{\mathrm{op}} \geq t\right\} \leq (d_1 + d_2)\exp\left(\frac{-\frac{3}{8}t^2}{\sigma^2}\right)$$

holds.

Matrix Bernstein inequality (cont.)

Theorem (Matrix Bernstein inequality [7])

Therefore, if

$$\sqrt{\frac{8}{3}\left(\log\frac{d_1+d_2}{\delta}\right)\sigma^2} \leq \frac{\sigma^2}{R}, \tag{5}$$

then with probability at least $1-\delta$,

$$\left\|\sum_{k=1}^{L}\mathbf{Z}_k\right\|_{\mathrm{op}} \leq \sqrt{\frac{8}{3}\left(\log\frac{d_1+d_2}{\delta}\right)\sigma^2}$$

holds.

Vector Bernstein inequality

Theorem (Vector Bernstein inequality [4])

Let $\{\boldsymbol{v}_k\}_{k=1}^L$ *be independent random vectors in* $\mathbb{R}^d$*. Suppose that* $\mathbb{E}\boldsymbol{v}_k = \boldsymbol{o}$ *and* $\|\boldsymbol{v}_k\| \leq R$ *(a.s.) and put* $\sum_{k=1}^L \mathbb{E}\|\boldsymbol{v}_k\|^2 \leq \sigma^2$*. Then for all* $t \in \left[0, \frac{\sigma^2}{R}\right]$*,*

$$\mathbb{P}\left(\left\|\sum_{k=1}^L \boldsymbol{v}_k\right\| \geq t\right) \leq \exp\left(-\frac{(t/\sigma - 1)^2}{4}\right) \leq \exp\left(-\frac{t^2}{8\sigma^2} + \frac{1}{4}\right)$$

holds.

Vector Bernstein inequality (cont.)

Theorem (Vector Bernstein inequality [4])

Therefore, given

$$\sigma\sqrt{2+8\log\frac{1}{\delta}} \le \frac{\sigma^2}{R} \tag{6}$$

with probability at least $1-\delta$,

$$\left\|\sum_{k=1}^{L} \boldsymbol{v}_k\right\| \le \sigma\sqrt{2+8\log\frac{1}{\delta}}$$

holds.

Scalar Bernstein inequality

Theorem (Bernstein's inequality for scalars [1, Corollary 2.11])

Let $X_1, \ldots, X_n$ be independent real-valued random variables that satisfy $|X_i| \leq R$ (a.s.), $\mathbb{E}[X_i] = 0$, and $\sum_{i=1}^n \mathbb{E}[X_i^2] \leq \sigma^2$. Then for all $t \in \left[0, \frac{\sigma^2}{R}\right]$,

$$\mathbb{P}\left\{\left|\sum_{i=1}^n X_i\right| \geq t\right\} \leq 2\exp\left(-\frac{\frac{3}{8}t^2}{\sigma^2}\right).$$

holds.

Theorem (Bernstein's inequality for scalars [1, Corollary 2.11])

Therefore, given

$$\sqrt{\frac{8}{3}\sigma^2 \log\frac{2}{\delta}} \leq \frac{\sigma^2}{R}, \tag{7}$$

with probability at least $1-\delta$,

$$\left|\sum_{i=1}^{n} X_i\right| \leq \sqrt{\frac{8}{3}\sigma^2 \log\frac{2}{\delta}}.$$

holds.

Concentration properties

From here, we denote $P^*_{ij}(\cdot) := (\mathcal{P}^*(\cdot))_{ij}$.

We need to prove the following concentrations.

1. $\|\mathcal{P}_T\mathcal{P}^*\mathcal{P}_T\Delta_{k-1} - \mathcal{P}_T\mathcal{R}^*_{\Omega_k}\mathcal{P}_T\Delta_{k-1}\|_{\mathrm{F}} \le \left(\frac{1}{2} - \rho_{\mathrm{F}}\right)\|\mathcal{P}_T\Delta_{k-1}\|_{\mathrm{F}}$
2. $\|(\mathcal{R}^*_{\Omega_k} - \mathcal{P}^*)(\mathcal{P}_T\Delta_{k-1})\|_{\mathrm{op}} \le \left(\frac{1}{4} - \rho_{\mathrm{op}}\right)\frac{1}{\|\mathbf{U}\mathbf{V}^\top\|_\infty}\|\mathcal{P}_T\Delta_{k-1}\|_\infty$
3. $\|(\mathcal{P}_T\mathcal{R}^*_{\Omega_k}\mathcal{P}_T - \mathcal{P}_T\mathcal{P}^*\mathcal{P}_T)(\mathcal{P}_T\Delta_{k-1})\|_\infty \le \left(\frac{1}{2} - \rho_\infty\right)\|\mathcal{P}_T\Delta_{k-1}\|_\infty$
4. $\|\mathcal{P}_T\mathcal{R}_\Omega\mathcal{P}_\mathcal{B}\mathcal{P}_T - \mathcal{P}_T\mathcal{P}_\mathcal{B}\mathcal{P}_T\|_{\mathrm{op}} \le \frac{1}{2} - \nu_\mathcal{B}$

We will prove only 1. here.

Frobenius norm concentration

Frobenius norm concentration (cont.)

Lemma (Frobenius norm concentration)

Assume that $\rho_{\mathrm{F}} < \frac{1}{2}$*, and that for some* $\beta > 1/(4\log(n_1 n_2))$,

$$p \geq \min\left\{1, p_{\min}^{\mathrm{F}}\right\},$$

is satisfied. Let $k \in \{1, \ldots, k_0\}$*. Then, given* $\mathcal{P}_T \Delta_{k-1}$ *that is independent of* Ω_k*, we have, w.p.* $\geq 1 - e^{\frac{1}{4}}(n_1 n_2)^{-\beta}$,

$$\|\mathcal{P}_T \mathcal{P}^* \mathcal{P}_T \Delta_{k-1} - \mathcal{P}_T \mathcal{R}_{\Omega_k}^* \mathcal{P}_T \Delta_{k-1}\|_{\mathrm{F}} \leq \left(\frac{1}{2} - \rho_{\mathrm{F}}\right) \|\mathcal{P}_T \Delta_{k-1}\|_{\mathrm{F}} \tag{8}$$

Proof (Frobenius norm concentration)

Before moving on to the proof, let us note the following property of coherence to be used in the proof.

Prop.

$$\|\mathcal{P}_T(\boldsymbol{e}_i \boldsymbol{f}_j^\top)\|_{\mathrm{F}}^2 \le \frac{n_1 + n_2}{n_1 n_2} \mu_0 r$$

Proof (Frobenius norm concentration) (cont.)

Proof.

$$
\begin{aligned}
\|\mathcal{P}_T(\boldsymbol{e}_i\boldsymbol{f}_j^\top)\|_{\mathrm{F}}^2 &= \|\mathcal{P}_{\mathrm{U}}(\boldsymbol{e}_i\boldsymbol{f}_j^\top)\|_{\mathrm{F}}^2 + \|\mathcal{P}_{\mathrm{V}}(\boldsymbol{e}_i\boldsymbol{f}_j^\top)\|_{\mathrm{F}}^2 - \|\mathcal{P}_{\mathrm{U}}(\boldsymbol{e}_i\boldsymbol{f}_j^\top)\|_{\mathrm{F}}^2\|\mathcal{P}_{\mathrm{V}}(\boldsymbol{e}_i\boldsymbol{f}_j^\top)\|_{\mathrm{F}}^2 \\
&\le \|\mathcal{P}_{\mathrm{U}}(\boldsymbol{e}_i\boldsymbol{f}_j^\top)\|_{\mathrm{F}}^2 + \|\mathcal{P}_{\mathrm{V}}(\boldsymbol{e}_i\boldsymbol{f}_j^\top)\|_{\mathrm{F}}^2 \\
&\le \frac{n_1+n_2}{n_1 n_2}\mu_0 r.
\end{aligned}
$$

□

Proof (Frobenius norm concentration) (cont.)

Note that since $(1-p)^{1/k_0} \leq 1-(1/k_0)p$, it follows that $q \geq (1/k_0)p$. We will repeatedly use this relation in proving concentration properties.

Proof (Frobenius norm concentration)

(Proof)

- If $p = 1$, then we have $q = 1$, therefore Eq. (8) holds.

Thus, from here, we assume $1 \geq p \geq p_{\min}^{\mathrm{F}}$.

- First decompose $\|\mathcal{P}_T \mathcal{P}^* \mathcal{P}_T \Delta_{k-1} - \mathcal{P}_T \mathcal{R}_{\Omega_k}^* \mathcal{P}_T \Delta_{k-1}\|_{\mathrm{F}}$ as

$$
\begin{aligned}
&\|\mathcal{P}_T \mathcal{P}^* \mathcal{P}_T \Delta_{k-1} - \mathcal{P}_T \mathcal{R}_{\Omega_k}^* \mathcal{P}_T \Delta_{k-1}\|_{\mathrm{F}} \\
&= \left\| \mathcal{P}_T \sum_{(i,j)} \left(1 - \frac{\omega_{ij}^{(k)}}{q}\right) P_{ij}^* (\langle \boldsymbol{e}_i \boldsymbol{f}_j^\top, \mathcal{P}_T \Delta_{k-1} \rangle) \boldsymbol{e}_i \boldsymbol{f}_j^\top \right\|_{\mathrm{F}} \\
&= \left\| \sum_{(i,j)} \left(1 - \frac{\omega_{ij}^{(k)}}{q}\right) P_{ij}^* (\langle \boldsymbol{e}_i \boldsymbol{f}_j^\top, \mathcal{P}_T \Delta_{k-1} \rangle) \mathcal{P}_T (\boldsymbol{e}_i \boldsymbol{f}_j^\top) \right\|_{\mathrm{F}} \\
&=: \left\| \sum \mathbf{S}_{ij} \right\| \quad .
\end{aligned}
$$

Proof (Frobenius norm concentration) (cont.)

- From here, we check the conditions for the vector Bernstein inequality (Theorem 19).

Now it is easy to verify that $\mathbb{E}[\mathbf{S}_{ij}] = \mathbf{O}$. We also have

$$\begin{aligned}
\|\mathbf{S}_{ij}\|_{\mathrm{F}} &= \left(1 - \frac{\omega_{ij}^{(k)}}{q}\right) |P_{ij}^{*}(\langle \boldsymbol{e}_i \boldsymbol{f}_j^{\top}, \mathcal{P}_T \Delta_{k-1}\rangle)| \|\mathcal{P}_T(\boldsymbol{e}_i \boldsymbol{f}_j^{\top})\|_{\mathrm{F}} \\
&\leq \frac{1}{q} |\langle \boldsymbol{e}_i \boldsymbol{f}_j^{\top}, \mathcal{P}_T \Delta_{k-1}\rangle| \|\mathcal{P}_T(\boldsymbol{e}_i \boldsymbol{f}_j^{\top})\|_{\mathrm{F}} \\
&\leq \frac{1}{q} \|\mathcal{P}_T(\boldsymbol{e}_i \boldsymbol{f}_j^{\top})\|_{\mathrm{F}}^2 \|\mathcal{P}_T \Delta_{k-1}\|_{\mathrm{F}} \\
&\leq \frac{1}{q} \frac{n_1 + n_2}{n_1 n_2} \mu_0 r \|\mathcal{P}_T \Delta_{k-1}\|_{\mathrm{F}}.
\end{aligned}$$

Proof (Frobenius norm concentration) (cont.)

On the other hand,

$$
\begin{aligned}
\sum_{(i,j)} \mathbb{E}\|\mathbf{S}_{ij}\|_{\mathrm{F}}^2 &= \sum_{(i,j)} \mathbb{E}\left[\left(1 - \frac{\omega_{ij}^{(k)}}{q}\right)^2\right] P_{ij}^*(\langle \boldsymbol{e}_i \boldsymbol{f}_j^\top, \mathcal{P}_T \Delta_{k-1}\rangle)^2 \|\mathcal{P}_T(\boldsymbol{e}_i \boldsymbol{f}_j^\top)\|_{\mathrm{F}}^2 \\
&= \frac{1-q}{q} \sum_{(i,j)} P_{ij}^*(\langle \boldsymbol{e}_i \boldsymbol{f}_j^\top, \mathcal{P}_T \Delta_{k-1}\rangle)^2 \|\mathcal{P}_T(\boldsymbol{e}_i \boldsymbol{f}_j^\top)\|_{\mathrm{F}}^2 \\
&\le \frac{1-q}{q} \sum_{(i,j)} \langle \boldsymbol{e}_i \boldsymbol{f}_j^\top, \mathcal{P}_T \Delta_{k-1}\rangle^2 \|\mathcal{P}_T(\boldsymbol{e}_i \boldsymbol{f}_j^\top)\|_{\mathrm{F}}^2 \\
&\le \frac{1-q}{q} \max_{(i,j)} \left\{\|\mathcal{P}_T(\boldsymbol{e}_i \boldsymbol{f}_j^\top)\|_{\mathrm{F}}^2\right\} \sum_{(i,j)} \langle \boldsymbol{e}_i \boldsymbol{f}_j^\top, \mathcal{P}_T \Delta_{k-1}\rangle^2 \\
&= \frac{1-q}{q} \max_{(i,j)} \|\mathcal{P}_T(\boldsymbol{e}_i \boldsymbol{f}_j^\top)\|_{\mathrm{F}}^2 \|\mathcal{P}_T \Delta_{k-1}\|_{\mathrm{F}}^2
\end{aligned}
$$

Proof (Frobenius norm concentration) (cont.)

$$\leq \frac{1}{q} \max_{(i,j)} \|\mathcal{P}_T(\boldsymbol{e}_i \boldsymbol{f}_j^\top)\|_{\mathrm{F}}^2 \|\mathcal{P}_T \Delta_{k-1}\|_{\mathrm{F}}^2$$
$$= \frac{1}{q} \left(\frac{n_1 + n_2}{n_1 n_2} \mu_0 r \right) \|\mathcal{P}_T \Delta_{k-1}\|_{\mathrm{F}}^2 .$$

Let
$R := \frac{n_1+n_2}{q n_1 n_2} \mu_0 r \|\mathcal{P}_T \Delta_{k-1}\|_{\mathrm{F}}, \sigma^2 := \frac{n_1+n_2}{q n_1 n_2} \mu_0 r \|\mathcal{P}_T \Delta_{k-1}\|_{\mathrm{F}}^2$,
and $\delta = e^{\frac{1}{4}} (n_1 n_2)^{-\beta}$. Under the condition

$$q \geq \frac{p}{k_0} \geq \frac{p_{\min}^{\mathrm{F}}}{k_0} = \frac{8 \mu_0 r}{(1/2 - \rho_{\mathrm{F}})^2} \beta \log(n_1 n_2) \frac{n_1 + n_2}{n_1 n_2},$$

Proof (Frobenius norm concentration) (cont.)

the condition (6) of Theorem 19 is satisfied, because

$$\sqrt{\left(2+8\log\frac{1}{\delta}\right)\sigma^2} = \sqrt{8\beta\log(n_1n_2)\frac{n_1+n_2}{qn_1n_2}\mu_0 r}\|\mathcal{P}_T\Delta_{k-1}\|_{\mathrm{F}}$$
$$\leq \left(\frac{1}{2}-\rho_{\mathrm{F}}\right)\|\mathcal{P}_T\Delta_{k-1}\|_{\mathrm{F}} \leq \|\mathcal{P}_T\Delta_{k-1}\|_{\mathrm{F}} = \frac{\sigma^2}{R}.$$

Therefore, applying Theorem 19 with $d = n_1n_2$, we obtain

$$\left\|\sum_{(i,j)}\mathbf{S}_{ij}\right\|_{\mathrm{F}} \leq \sqrt{\left(2+8\log\frac{1}{\delta}\right)\sigma^2} \leq \left(\frac{1}{2}-\rho_{\mathrm{F}}\right)\|\mathcal{P}_T\Delta_{k-1}\|_{\mathrm{F}}$$

Proof (Frobenius norm concentration) (cont.)

with probability at least $1 - e^{\frac{1}{4}}(n_1 n_2)^{-\beta}$.

□

Other concentration inequalities

For the concentration of

- $\|(\mathcal{R}^*_{\Omega_k} - \mathcal{P}^*)(\mathcal{P}_T\Delta_{k-1})\|_{\mathrm{op}} \leq \left(\frac{1}{4} - \rho_{\mathrm{op}}\right) \frac{1}{\|\mathbf{U}\mathbf{V}^\top\|_\infty} \|\mathcal{P}_T\Delta_{k-1}\|_\infty$
- $\|(\mathcal{P}_T\mathcal{R}^*_{\Omega_k}\mathcal{P}_T - \mathcal{P}_T\mathcal{P}^*\mathcal{P}_T)(\mathcal{P}_T\Delta_{k-1})\|_\infty \leq \left(\frac{1}{2} - \rho_\infty\right) \|\mathcal{P}_T\Delta_{k-1}\|_\infty$
- $\|\mathcal{P}_T\mathcal{R}_\Omega\mathcal{P}_\mathcal{B}\mathcal{P}_T - \mathcal{P}_T\mathcal{P}_\mathcal{B}\mathcal{P}_T\|_{\mathrm{op}} \leq \frac{1}{2} - \nu_\mathcal{B}$

please refer to the paper (they are similar but require different calculations).

Proof of Theorem thm:exact-recovery-guarantee

Proof.

The theorem immediately follows from the combination of Lemma 13, Lemma **??**, Lemma 15, and the union bound. □

References

[1] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart.
Concentration Inequalities: A Nonasymptotic Theory of Independence.
Oxford University Press, Oxford, 1st edition, 2013.

[2] Emmanuel J. Candès and Benjamin Recht.
Exact matrix completion via convex optimization.
Foundations of Computational mathematics, 9(6):717, 2009.

[3] Yudong Chen, Srinadh Bhojanapalli, Sujay Sanghavi, and Rachel Ward.
Completing any low-rank matrix, provably.
Journal of Machine Learning Research, 16:2999–3034, 2015.

References (cont.)

[4] David Gross.
Recovering low-rank matrices from few coefficients in any basis.
IEEE Transactions on Information Theory, 57(3):1548–1566, March 2011.

[5] B. Recht, M. Fazel, and P. Parrilo.
Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization.
SIAM Review, 52(3):471–501, January 2010.

References (cont.)

[6] Takeshi Teshima, Miao Xu, Issei Sato, and Masashi Sugiyama.
Clipped Matrix Completion: A Remedy for Ceiling Effects.
arXiv:1809.04997 [cs, stat], September 2018.

[7] Joel A. Tropp.
User-friendly tail bounds for sums of random matrices.
Foundations of Computational Mathematics, 12(4):389–434, August 2012.