Finding Hall blockers by matrix scaling

based on

--- convergence analysis on "divergent" Sinkhorn iteration ---

Hiroshi Hirai
Graduate School of Mathematics
Nagoya University
hirai.hiroshi@math.nagoya-u.ac.jp

Joint work with
Koyo Hayashi, Keiya Sakabe

arXiv:2204.07425 v2

Discrete Optimization and Machine Learning
GRIPS, August 8, 2023

Matrix Scaling（Sinkhorn 1964）

A: $n \times n$ nonnegative matrix

Can we scale A to doubly stochastic matrix RAC
by multiplying positive diagonal matrices R, C ?

Goal: Find R, C s.t. $RAC\,\mathbf{1} \approx \mathbf{1}, (RAC)^{\top}\mathbf{1} \approx \mathbf{1}$

Applications:

- Markov chain
- Preprocessing for solving linear equation
- Optimal transport, entropic regularization (Cuturi 2013)

and more

Sinkhorn Algorithm

- Row normalization: $A \leftarrow RA$ s.t. $(RA)\mathbf{1} = \mathbf{1}$; $R = \mathrm{diag}(\cdots (\sum_k A_{ik})^{-1} \cdots)$
- Col normalization: $A \leftarrow AC$ s.t. $(AC)^{\top}\mathbf{1} = \mathbf{1}$; $C = \mathrm{diag}(\cdots \left(\sum_k A_{kj}\right)^{-1} \cdots)$
- Repeat it

$$
\begin{array}{cc}
 & \begin{matrix} 4 & 10 & 1 & 1 \end{matrix} \\
\begin{matrix} 6 \\ 3 \\ 4 \\ 3 \end{matrix} & \begin{pmatrix} & 6 & & \\ 1 & & 1 & 1 \\ & 4 & & \\ 3 & & & \end{pmatrix}
\end{array}
\xrightarrow{\text{Row normalize}}
\begin{array}{cc}
 & \begin{matrix} 4/3 & 2 & 1/3 & 1/3 \end{matrix} \\
\begin{matrix} 1 \\ 1 \\ 1 \\ 1 \end{matrix} & \begin{pmatrix} & 1 & & \\ 1/3 & & 1/3 & 1/3 \\ & 1 & & \\ 1 & & & \end{pmatrix}
\end{array}
$$

$$
\xrightarrow{\text{Col normalize}}
\begin{array}{cc}
 & \begin{matrix} 1 & 1 & 1 & 1 \end{matrix} \\
\begin{matrix} 1/2 \\ 9/4 \\ 1/2 \\ 3/4 \end{matrix} & \begin{pmatrix} & 1/2 & & \\ 1/4 & & 1 & 1 \\ & 1/2 & & \\ 3/4 & & & \end{pmatrix}
\end{array}
\xrightarrow{\text{Row normalize}} \cdots
$$

Characterization of Scalability

Sinkhorn, Knopp 1967, Rothblum, Schneider 1989

Thm [SK1967, RS1989]: The following are equivalent:

- A is (approximately) doubly stochastic scalable:

$$\forall \epsilon > 0, \exists R, C \text{ s.t. } \|RAC\mathbf{1} - \mathbf{1}\| < \epsilon, \|(RAC)^\top \mathbf{1} - \mathbf{1}\| < \epsilon$$

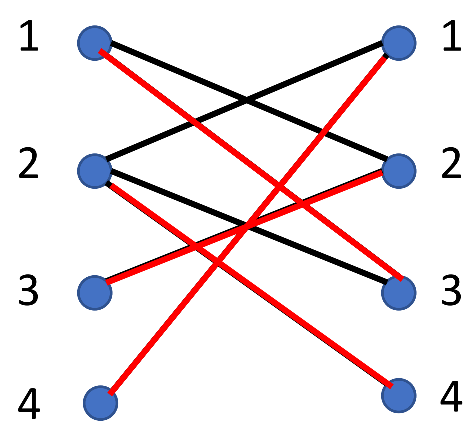
- Sinkhorn algorithm **converges**; $A \rightarrow$ doubly stochastic

- $\exists$ **perfect matching** in bipartite graph G_A

$$V(G_A) \coloneqq [n] \sqcup [n],\ E(G_A) \coloneqq \{\, ij \mid A_{ij} \neq 0 \,\}$$

$$A = \begin{array}{c@{}c} & \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \end{array} & \begin{pmatrix} & 6 & 3 & \\ 1 & & 1 & 1 \\ & 4 & & \\ 3 & & & \end{pmatrix} \end{array}$$

$G_A =$



Testing Perfect Matching by Sinkhorn

Linial, Samorodnitsky, Wigderson 2000

$G = (U \sqcup V, E)$: bipartite graph, $|U| = |V| = n$

$$A(G)_{ij} := \begin{cases} 1 & \text{if } ij \in E \\ 0 & \text{otherwise} \end{cases}$$

Initialization: $A \leftarrow A(G)$

1. Row-normalize: $A \leftarrow \mathrm{diag}(1/\sum_k A_{ik})A$
2. Col-normalize: $A \leftarrow A\,\mathrm{diag}\big(1/\sum_k A_{kj}\big)$
3. If $\|A\mathbf{1} - \mathbf{1}\|_2 < 1/\sqrt{n}$, then stop.
4. Go to 1.

$$\begin{array}{c|c|} & 1\;1\;1\;\cdots\;1 \\ \hline p_1 & \\ p_2 & \\ \vdots & A \\ p_n & \\ \hline \end{array}$$

Thm [LSW2000]

$\exists$ perfect matching in G → terminates within $O(n^2 \log n)$ iterations

$\nexists$ perfect matching in G → not terminate

- LSW algorithm is slower than augmenting path
  But interesting in its conceptual difference & extraordinary simplicity.

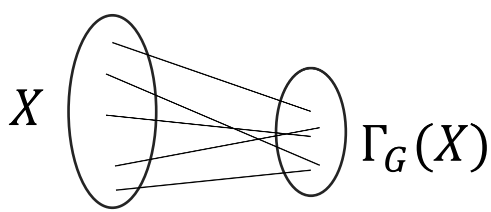
- links to recent development on *operator scaling* & *noncommutative PIT*

  Garg, Gurvits, Oliveira, Wigderson FOCS2016

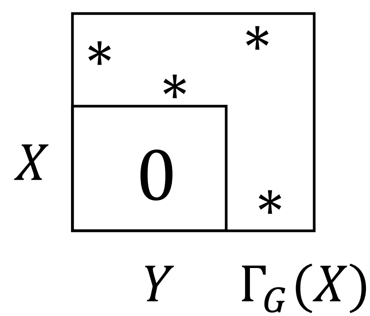
- LSW algorithm outputs **neither** perfect matching **nor** *certificate of nonexistence*

Hall's marriage theorem：∃ perfect matching ⇔ ∄ *Hall blocker*

Hall blocker



$|X| > |\Gamma_G(X)|$



$|X| + |Y| > n$

Can we identify Hall blockers from (divergent !) Sinkhorn iteration ?

Significant in operator scaling generalization: Franks, Soma, Goemans SODA2023
But not well-understood even in matrix scaling setting

Result 1

$O(n^2 \log n)$ Sinkhorn iterations identify a Hall blocker
*from **scaling matrices R, C***

0. $A \leftarrow A(G),\ R = C = I$

1. Row-normalize: $R \leftarrow \mathrm{diag}\left(\left(\sum_j A_{1j}C_{jj}\right)^{-1}, \left(\sum_j A_{2j}C_{jj}\right)^{-1}, \dots, \left(\sum_j A_{nj}C_{jj}\right)^{-1}\right)$

2. Col-normalize: $C \leftarrow \mathrm{diag}\left(\left(\sum_i R_{ii}A_{i1}\right)^{-1}, \left(\sum_i R_{ii}A_{i2}\right)^{-1}, \dots, \left(\sum_i R_{ii}A_{in}\right)^{-1}\right)$

3. Sort as $R_{11} \geq R_{22} \geq \cdots \geq R_{nn}, C_{11} \leq C_{22} \leq \cdots \leq C_{nn}$,

   Choose k s.t. $(R_{kk}C_{kk})^n \geq R_{11}R_{22}\cdots R_{nn}C_{11}C_{22}\cdots C_{nn}$,

   Output $X := \{1,2,\dots,k\}$

4. Go to 1.

Rounding produce by
Franks,Soma Goemans SODA2023

Thm [This work] Suppose that G has no perfect matching

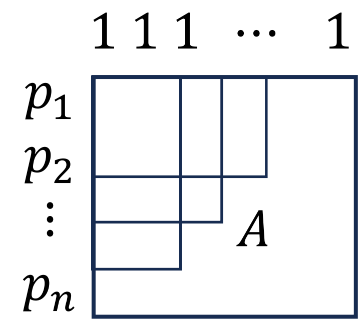
X is a Hall blocker within $O(n^2 \log n)$ iterations

Result 2

*$O(n^6 \log n)$ Sinkhorn iterations identify all parametric Hall blockers from **row-marginals** $\boldsymbol{p} \coloneqq \boldsymbol{A1}$*

0. $A \leftarrow A(G)$
1. Row-normalize: $A \leftarrow \mathrm{diag}(1/\sum_k A_{ik})A$
2. Col-normalize: $A \leftarrow A\,\mathrm{diag}\left(1/\sum_k A_{kj}\right)$
3. Sort $p \coloneqq A\mathbf{1}$ as $p_1 \geq p_2 \geq \cdots \geq p_n$

   Output $\mathcal{X} \coloneqq \{\{1,2,\ldots,k\} \mid k = 1,2,\ldots,n\}$.
4. Go to 1.



Thm [This work] Suppose that G has no perfect matching.

$\mathcal{X}$ contains all *parametric Hall blockers* within $O(n^6 \log n)$ iterations

maximizing $|X| - \alpha|\Gamma_G(X)|$ $(\alpha \in \mathbb{R}_+)$

Proof idea

Sinkhorn iteration $=$ Alternating minimization

Result 1 : $\boldsymbol{O(n^2 \log n)}$ iterations identify "a" Hall blocker

$\sim$ geometric programming interpretation

$$\text{inf.} \quad \log \frac{(x^\top A y)^n}{\prod_i x_i \prod_j y_j} \quad \text{s.t.} \quad x > 0, y > 0.$$

Fix y optimize x. Fix x optimize y $\cdots$

Result 2 : $\boldsymbol{O(n^6 \log n)}$ iterations identify "all parametric" Hall blockers

$\sim$ KL-divergence minimization interpretation

$$\text{inf.} \sum_{ij} M_{ij} \log \frac{M_{ij}}{N_{ij}} \quad \text{s.t.} \; M\mathbf{1} = \mathbf{1}, N^\top \mathbf{1} = \mathbf{1},$$

$$M, N \geq 0, \;\; \operatorname{supp} M, \operatorname{supp} N \subseteq \operatorname{supp} A$$

Fix N optimize M. Fix M optimize N $\cdots$

Matrix Scaling as Geometric Programming

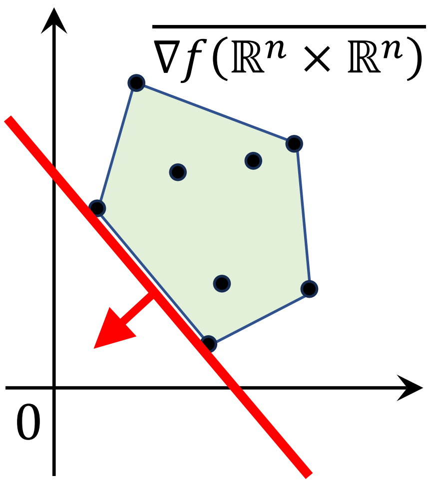
Thm (SK1967, RS1989) The following are equivalent:

1. A is approximately scalable.
2. Sinkhorn converges.
3. Geometric programming is ***bounded below:***

$$\inf_{s.t\in\mathbb{R}^n} \quad f(s,t) \coloneqq \log \sum_{i,j} A_{ij} e^{s_i+t_j} - \frac{1}{n}\mathbf{1}^\top s - \frac{1}{n}\mathbf{1}^\top t \quad > -\infty$$

Scaling matrix $R, C \quad \Leftrightarrow \quad \mathrm{diag}(e^{s_i}),\ \mathrm{diag}(e^{t_j})$

Sinkhorn algorithm $\Leftrightarrow$ alternating minimization



$$\inf_{s.t\in\mathbb{R}^n} f(s,t) > -\infty \Leftrightarrow 0 \in \overline{\nabla f(\mathbb{R}^n \times \mathbb{R}^n)}$$

$$= -\infty \Leftrightarrow 0 \notin \overline{\nabla f(\mathbb{R}^n \times \mathbb{R}^n)}$$

$$||$$

Hall blocker ≈ separating facet of $\mathrm{conv}\left\{ (e_i, e_j) - \frac{1}{n}(\mathbf{1},\mathbf{1}) \mid i,j : A_{ij} > 0 \right\}$

Unbounded Certificate in Geometric Programming

Lem/Obs [This work ?]

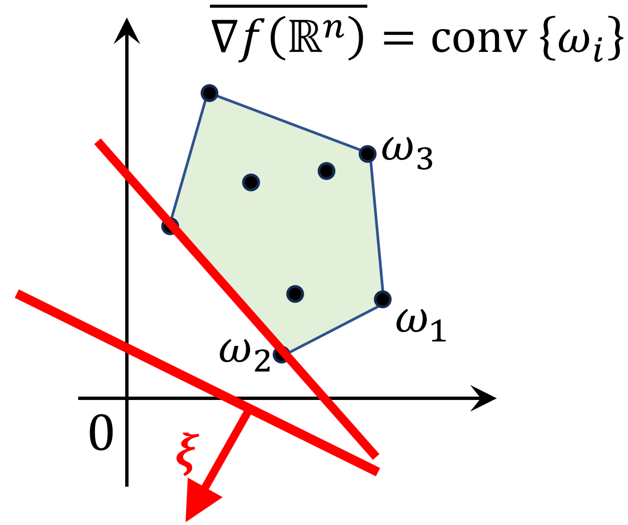
$$g(\xi) := \log\left(\sum_{i=1}^{m} q_i e^{\omega_i^\top \xi}\right), q_i > 0$$

If $g(\xi) \leq \log\min\{q_1 \dots, q_m\}$,

then $\{0\}$ and $\overline{\nabla f(\mathbb{R}^n)}$ are separated

by the hyperplane of normal vector ξ

→ $\inf g = -\infty$



$\because\ \omega_k^\top \xi \geq 0 (\exists k) \rightarrow\ g(\xi) = \log\left(\sum_{i=1}^{m} q_i e^{\omega_i^\top \xi}\right) > \log q_k$

We analyze

- If unbounded, *"many" iterations* make $\xi = (s, t)$ a separating hyperplane
- *Round* the hyperplane to a separating facet (≈ Hall blocker)

Rounding procedure: Franks, Soma, Goemans SODA2023

Analysis (can be skipped)

Decrement of one Sinkhorn iteration: $s, t \to s', t \to s', t'$; $A \underset{\text{Row}}{\to} A' \underset{\text{Col}}{\to} A''$

Lem [folkrore ?; Altschuler, Weed, Rigolle NIPS2017]

$$f(s,t) - f(s',t') = D_{KL}(\mathbf{1}|A\mathbf{1})/n + D_{KL}\left(\mathbf{1}|A'^{\top}\mathbf{1}\right)/n$$

Pinsker's ineq $\geq \frac{1}{2n^2}\{\|A\mathbf{1} - \mathbf{1}\|_1^2 + \left\|A'^{\top}\mathbf{1} - \mathbf{1}\right\|_1^2\}$

Lem [Gurvits, Leake STOC2021]:

A: nonnegative, supp $A = E(G)$

$$\|A\mathbf{1} - \mathbf{1}\|_1 + \|A^{\top}\mathbf{1} - \mathbf{1}\|_1 \geq 2\max_X |X| - |\Gamma_G(\mathrm{X})|$$

≥ 2 If no perfect matching

- $A \coloneqq A_G$: 0-1 matrix for G without perfect matching
- $f(0,0) = O(\log n)$, decrement of one iteration $= \Omega(1/n^2)$
- $f(s,t) \leq \log 1 = 0$ after $O(n^2 \log n)$ iterations
- s, t: separating hyperplane $\to$ Hall blocker

Proof idea

Sinkhorn iteration $=$ Alternating minimization

Result 1 : $\boldsymbol{O(n^2 \log n)}$ iterations identify "a" Hall blocker

$\sim$ geometric programming interpretation

$$\text{inf.} \quad \log \frac{x^\top A y}{\prod_i x_i \prod_j y_j} \quad \text{s.t.} \quad x > 0, y > 0.$$

Fix y optimize x. Fix x optimize y $\cdots$

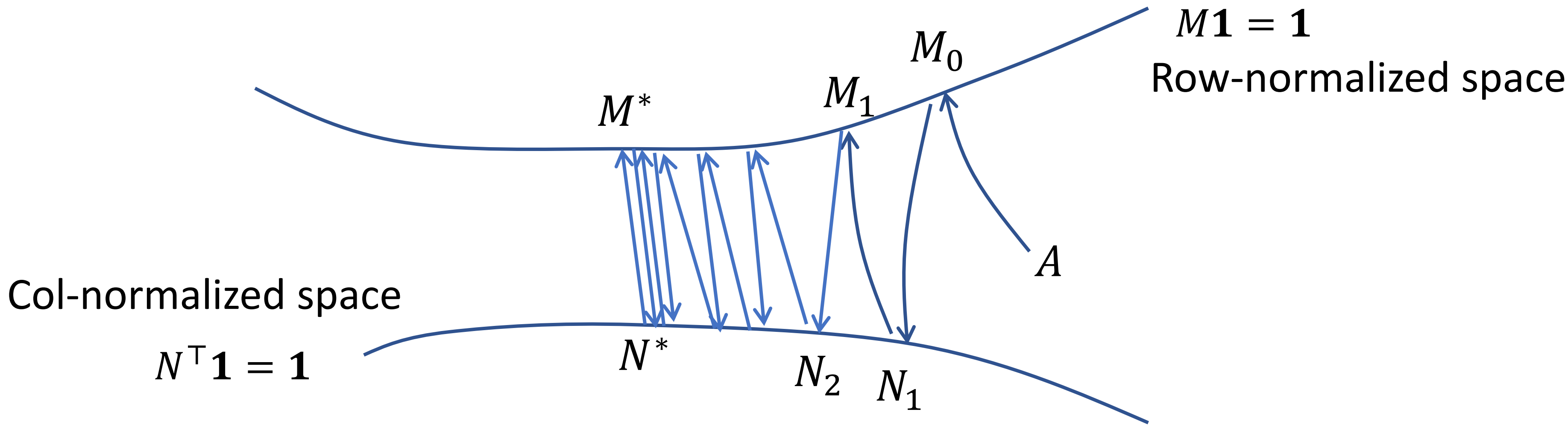
Result 2 : $\boldsymbol{O(n^6 \log n)}$ iterations identify "all parametric" Hall blockers

$\sim$ KL-divergence minimization interpretation

$$\text{inf.} \sum_{ij} M_{ij} \log \frac{M_{ij}}{N_{ij}} \quad \text{s.t.} \; M\mathbf{1} = \mathbf{1}, N^\top \mathbf{1} = \mathbf{1},$$
$$M, N \geq 0, \;\; \operatorname{supp} M, \operatorname{supp} N \subseteq \operatorname{supp} A$$

Fix N optimize M. Fix M optimize N $\cdots$

Information geometry of Sinkhorn iteration



$$\inf.\ D_{KL}(N|M) \ \text{ s.t. } M\mathbf{1} = \mathbf{1}, N^\top\mathbf{1} = \mathbf{1},$$
$$M, N \geq 0, \ \text{supp}\, M, \text{supp}\, N \subseteq \text{supp}\, A$$

Thm (Csiszar, Tusnady 1984, Gietl, Reffel 2013)

$\{(M_k, N_k)\}_{k=1,2,\dots}$ converges to a minimum KL-divergence pair (M^*, N^*)

- $M^* = N^*$ if Sinkhorn converges
- $M^* \leftrightarrow N^*$ if Sinkhorn does not converge → DEMO
  oscillating

Rate of Convergence

We point out that proof argument of CT1984 implies:

Lem [CT1984; This work]

$$D_{KL}(N_k|M_k) - D_{KL}(N^*|M^*) \le \frac{D(N^*|M_0)}{k}$$

$$\parallel$$

$p^k := N_k\mathbf{1}$

$p^* := N^*\mathbf{1}$

$= \mathrm{argmin}_{N^\top \mathbf{1}=\mathbf{1}} D_{KL}(N\mathbf{1}|\mathbf{1})$

$$D_{KL}(p^k|\mathbf{1}) - D_{KL}(p^*|\mathbf{1}) \ \ge\ D_{KL}(p^k|p^*) \ \ge\ \frac{1}{2n}\left\|p^k - p^*\right\|_1^2$$

Pinsker

Pythagorean thm

Lem [This work]

$$\left\|p^k - p^*\right\|_1 \ \le \sqrt{\frac{2nD_{KL}(N^*|M_0)}{k}}$$

We give an *explicit* formula of the marginal limit $p^* = N^*\mathbf{1}$ in terms of **DM-decomposition & parametric Hall blockers**

The Sinkhorn Limit $N^* \leftrightarrow M^*$

Aas 2014: The limits $N^* \leftrightarrow M^*$ are **block diagonalized** so that each block is oscillated as

$$\begin{array}{c|c|} & 1\ 1\ 1\ \cdots\ 1 \\ \hline \alpha & \\ \alpha & \\ \vdots & \\ \alpha & \\ \hline \end{array} \quad \Leftrightarrow \quad \begin{array}{c|c|} & \frac{1}{\alpha}\ \frac{1}{\alpha}\ \frac{1}{\alpha}\ \cdots\ \frac{1}{\alpha} \\ \hline 1 & \\ 1 & \\ \vdots & \\ 1 & \\ \hline \end{array}$$

α :=col number / row number

This work:

This block diagonalization = extended **Dulmage-Mendelsohn decomposition**

Canonical form of matrices under row/column permutation

→DEMO

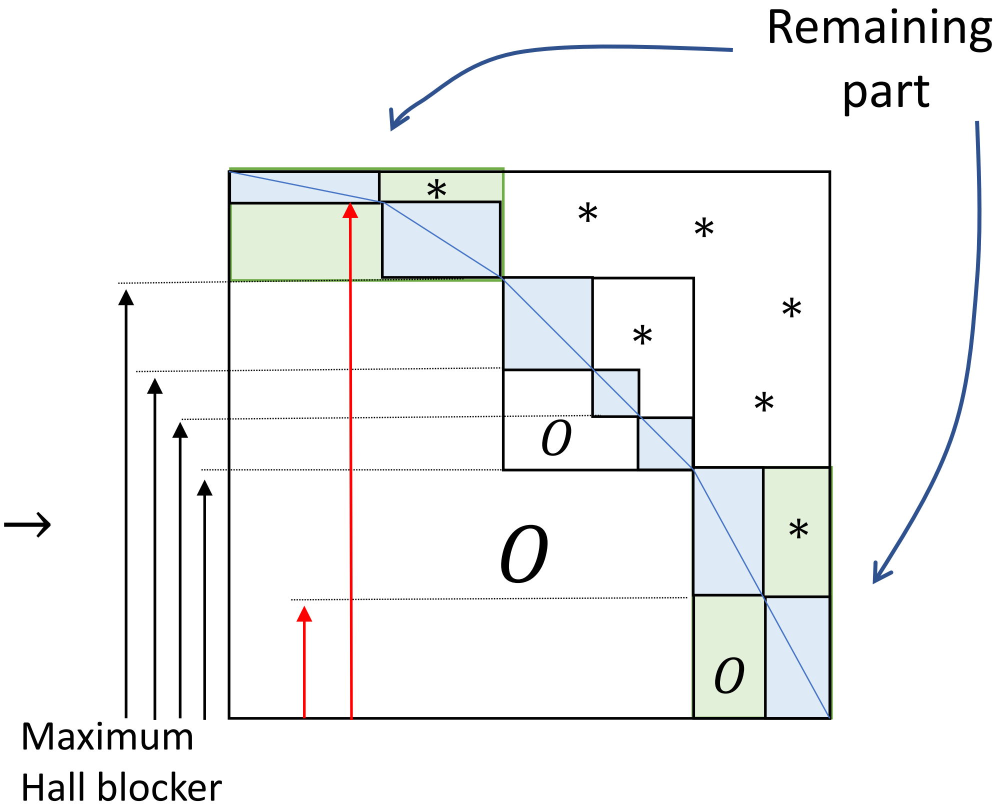
The DM-decomposition

Dulmage-Mendelson 1958

DM-decomposition =

Permute rows & columns along maximum-size Hall blockers

Size of zero block
$:=$ # row+ # col
$\approx\ |X| - |\Gamma(X)| + \text{const}$

$A \rightarrow$ 

Tomizawa 1977 (unpublished)

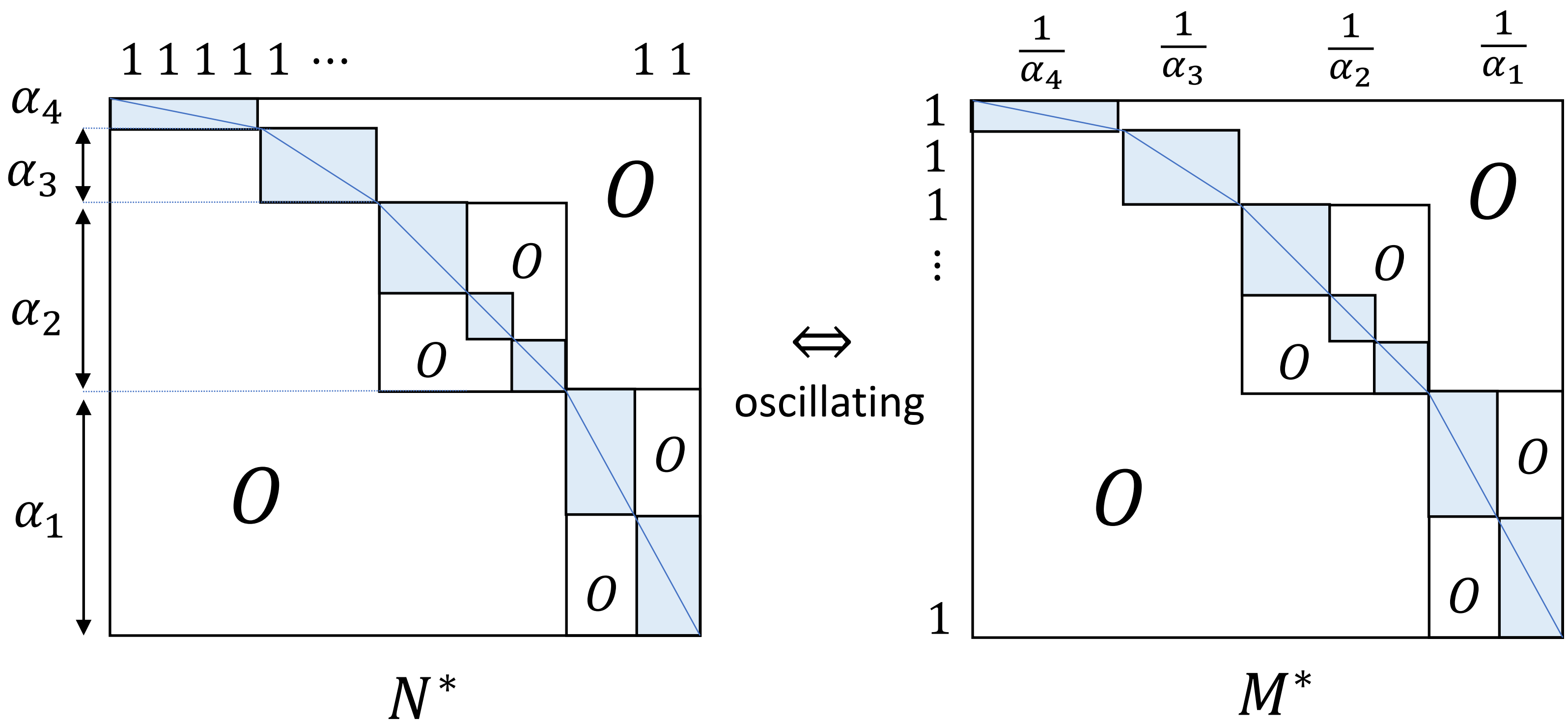
Extended DM decomposition = decompose "remaining parts"

via parametric stable sets in bipartite graph $G(A)$

Weighted size of zero block $:=$ # row $+\ \alpha \times$ # col

$\approx |X| - \alpha|\Gamma(X)| + \text{const}$

The Sinkhorn Limit $N^* \longleftrightarrow M^*$



- From $\alpha_1 > \alpha_2 > \alpha_3 > \cdots$, the row-sum vector $p^* = N^*\mathbf{1}$ identifies the structure of DM-decomposition → parametric Hall blockers
- So does $p = A\mathbf{1}$ if it is close to $p^* = N^*\mathbf{1}$ after $O(n^6 \log n)$ iterations

← convergence rate $O(1/\sqrt{k})$

Summary

- Parametric Hall blockers from divergent Sinkhorn iteration
- Information geometry & DM-decomposition
- Can we improve $O(n^6 \log n)$?

Original motivation: Operator scaling (Gurvits 2004, Garg et al. 2020)

Completely positive operator $X \mapsto \sum_k A_k X A_k^\dagger$ $(A_1, A_2, \ldots, A_m \in \mathbb{C}^{n \times n})$

doubly-stochastic $\underset{\text{def}}{\Leftrightarrow}$ $\displaystyle\sum_k A_k A_k^\dagger = I \, , \sum_k A_k^\dagger A_k = I$

Find *scaling* $g, h \in GL_n(\mathbb{C})$ s.t. $X \mapsto \sum_k g A_k h \, X \, h^\dagger A_k^\dagger g^\dagger$ is doubly-stochastic

- Operator Sinkhorn algorithm (Gurvits algorithm)
- Shrunk subspace = a certificate of nonscalability (analogue of Hall blocker)

Operator Sinkhorn algorithm (Gurvits algorithm, flip-flop)

1. Row-normalize: $A_k \leftarrow gA_k \quad : g(\sum_k A_k A_k^\dagger)g^\dagger = I$
2. Col-normalize: $A_k \leftarrow A_k h^\dagger : h(\sum_k A_k^\dagger A_k)h^\dagger = I$
3. Go to 1.

Thm (Gurvitz 2004): The following condition are equivalent:

- $X \mapsto \sum_k A_k X A_k^\dagger$ is approximately scalable
- Operator Sinkhorn converges
- No vector subspace (*shrunk subspace*) $V \subseteq \mathbb{C}^n$: $\dim V > \dim \sum_k A_k V$

matrix scaling case
$|X| > |\Gamma_G(X)|$

Can we find shrunk subspaces by operator Sinkhorn iterations ?

Franks, Soma, Goemans SODA2023: **YES** by *modified* Sinkhorn iterations.

Still complicated & huge (polynomial) iterations

Thank you for your attention !