Information geometry of operator scaling

Takeru Matsuda (RIKEN Center for Brain Science)

Tasuku Soma (Massachusetts Institute of Technology)

Abstract

Matrix scaling

- classical problem with many applications
- solved by an iterative algorithm called the Sinkhorn algorithm
- Csiszár (1975): Sinkhorn algorithm = alternating e-projections

Operator scaling

- generalization of matrix scaling to positive maps
- Gurvits (2004): Sinkhorn algorithm for operator scaling

- We investigate the operator Sinkhorn algorithm from the viewpoint of quantum information geometry.

matrix scaling	KL divergence	Fisher metric
operator scaling	?	?

- paper: M. and Soma. *Linear Algebra and its Applications*, 2022.

Matrix scaling and Sinkhorn algorithm

Matrix scaling problem

- Input: $A \in \mathbb{R}_+^{n \times n}$, $r \in \mathbb{R}_+^n$, $c \in \mathbb{R}_+^n$
- Output: diagonal matrices $L \in \mathbb{R}_+^{n \times n}$ and $R \in \mathbb{R}_+^{n \times n}$ such that LAR has the specified row/column sums:

$$\sum_{j=1}^{n} (LAR)_{ij} = r_i \quad (i = 1, \dots, n)$$

$$\sum_{i=1}^{n} (LAR)_{ij} = c_j \quad (j = 1, \dots, n)$$

- application
  - Markov chain estimation (Sinkhorn, 1964)
  - contingency table analysis (Morioka and Tsuda, 2011)
  - optimal transport (Peyré and Cuturi, 2019)
  - data assimilation (Reich, 2019)
  - and more (Idel, 2016)

Matrix scaling problem: example

$$A = \begin{pmatrix} 0.5 & 0.4 & 0.1 \\ 0.4 & 0.2 & 0.7 \\ 0.1 & 0.4 & 0.2 \end{pmatrix} \quad r = c = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

$$L = \begin{pmatrix} 1.0136 & 0 & 0 \\ 0 & 0.7324 & 0 \\ 0 & 0 & 1.4826 \end{pmatrix}$$

$$R = \begin{pmatrix} 1.0548 & 0 & 0 \\ 0 & 0.8734 & 0 \\ 0 & 0 & 1.0982 \end{pmatrix}$$

$$LAR = \begin{pmatrix} 0.5346 & 0.3541 & 0.1113 \\ 0.3090 & 0.1279 & 0.5630 \\ 0.1564 & 0.5180 & 0.3256 \end{pmatrix}$$

Sinkhorn algorithm (a.k.a. RAS method)

- Initialize $A^{(0)} = A$, $L = I$ and $R = I$
- Iterate the following for $k = 0, 1, 2, \ldots$ until convergence

$$A^{(0)} \to A^{(1)} \to A^{(2)} \to \cdots \to A^*$$

- Row-scaling (= left multiplication by a diagonal matrix)

$$A_{ij}^{(2k+1)} = \frac{r_i A_{ij}^{(2k)}}{\sum_{j'} A_{ij'}^{(2k)}}, \quad L_{ii} \leftarrow L_{ii} \frac{r_i}{\sum_{j'} A_{ij'}^{(2k)}}$$

- Column-scaling (= right multiplication by a diagonal matrix)

$$A_{ij}^{(2k+2)} = \frac{c_j A_{ij}^{(2k+1)}}{\sum_{i'} A_{i'j}^{(2k+1)}}, \quad R_{jj} \leftarrow R_{jj} \frac{c_j}{\sum_{i'} A_{i'j}^{(2k+1)}}$$

Sinkhorn algorithm: example

$$A = \begin{pmatrix} 0.5 & 0.4 & 0.1 \\ 0.4 & 0.2 & 0.7 \\ 0.1 & 0.4 & 0.2 \end{pmatrix} \quad r = c = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

$$A^{(1)} = \begin{pmatrix} 0.5000 & 0.4000 & 0.1000 \\ 0.3077 & 0.1538 & 0.5385 \\ 0.1429 & 0.5714 & 0.2857 \end{pmatrix}$$

$$A^{(2)} = \begin{pmatrix} 0.5260 & 0.3555 & 0.1082 \\ 0.3237 & 0.1367 & 0.5826 \\ 0.1503 & 0.5078 & 0.3092 \end{pmatrix}$$

$$A^{*} = \begin{pmatrix} 0.5346 & 0.3541 & 0.1113 \\ 0.3090 & 0.1279 & 0.5630 \\ 0.1564 & 0.5180 & 0.3256 \end{pmatrix}$$

Sinkhorn's theorem

- Assume that A is positive ($A \in \mathbb{R}_{++}^{n\times n}$):

$$A_{ij} > 0 \quad (i = 1, \ldots, n;\ j = 1 \ldots, n)$$

Theorem (Sinkhorn, 1964)

For a positive matrix A, the solution (L, R) exists and it is unique up to constant ($L \to \lambda L, R \to \lambda^{-1} R$). The Sinkhorn algorithm converges to the solution.

- extension to nonnegative matix: Sinkhorn and Knopp (1967)
- There are many approaches to prove this theorem (Idel, 2016)
  - convex duality
  - nonlinear Perron-Frobenius
  - potential optimization
  - information geometry

Sinkhorn = alternating e-projections

Sinkhorn = alternating e-projections

- We consider the following problem for convenience

- Input: $A \in \mathbb{R}_{++}^{n \times n}$
- Output: diagonal matrices $L \in \mathbb{R}_{++}^{n \times n}$ and $R \in \mathbb{R}_{++}^{n \times n}$ such that

$$\sum_{i=1}^{n} (LAR)_{ij} = \frac{1}{n} \quad (j = 1, \ldots, n)$$

$$\sum_{j=1}^{n} (LAR)_{ij} = \frac{1}{n} \quad (i = 1, \ldots, n)$$

Notation

$$\Pi = \left\{ A \in \mathbb{R}_{++}^{n\times n} \,\middle|\, \sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij} = 1 \right\}$$

$$\Pi_1 = \left\{ A \in \mathbb{R}_{++}^{n\times n} \,\middle|\, \sum_{j=1}^{n} A_{ij} = \frac{1}{n} \quad (i = 1, \cdots, n) \right\} \subset \Pi$$

$$\Pi_2 = \left\{ A \in \mathbb{R}_{++}^{n\times n} \,\middle|\, \sum_{i=1}^{n} A_{ij} = \frac{1}{n} \quad (j = 1, \cdots, n) \right\} \subset \Pi$$

- Π is viewed as a multinomial model
  - The Fisher metric and e/m connections are naturally introduced (Amari and Nagaoka, 2000)
- Both Π_1 and Π_2 are m-flat subspaces

Sinkhorn = alternating e-projections

- From the viewpoint of information geometry, the Sinkhorn algorithm is interpreted as alternating e-projections !

Theorem (Csiszár, 1975)

Each iteration of the Sinkhorn algorithm is the e-projection: the e-geodesic from $A^{(2k)}$ to $A^{(2k+1)}$ (from $A^{(2k+1)}$ to $A^{(2k+2)}$) is orthogonal to Π_1 (Π_2) w.r.t. the Fisher metric.

- Note: since Π_1 and Π_2 are m-flat, the e-projection is unique
  - generalized Pythagorean theorem

Proof

- The e-geodesic from $A^{(2k)}$ to $A^{(2k+1)}$ is

$$\begin{aligned}A(t) &= C(t)^{-1} \exp((1-t)\log A^{(2k)} + t \log A^{(2k+1)}) \\ &= C(t)^{-1} (A_{ij}^{(2k)})^{1-t} (A_{ij}^{(2k+1)})^{t},\end{aligned}$$

  where $0 \leq t \leq 1$, $C(t) = \sum_{i,j} (A_{ij}^{(2k)})^{1-t} (A_{ij}^{(2k+1)})^{t}$, and each operation is element-wise.
- Thus,

$$\frac{\mathrm{d}}{\mathrm{d}t} \log A(t)_{ij} = -C'(t) + \log A_{ij}^{(2k+1)} - \log A_{ij}^{(2k)}.$$

- In particular, $C'(1)$ coincides with the Kullback–Leibler divergence:

$$C'(1) = D(A^{(2k+1)} || A^{(2k)}) = \sum_{i,j} A_{ij}^{(2k+1)} \log \frac{A_{ij}^{(2k+1)}}{A_{ij}^{(2k)}}.$$

Proof

- Therefore, the e-representation of the tangent vector of the e-geodesic at $A^{(2k+1)}$ is

$$
\begin{aligned}
X^{(e)} &= \frac{\mathrm{d}}{\mathrm{d}t} \log A(t)_{ij} \bigg|_{t=1} \\
&= -D(A^{(2k+1)} || A^{(2k)}) + \log A_{ij}^{(2k+1)} - \log A_{ij}^{(2k)}.
\end{aligned}
$$

- From the definition of the Sinkhorn algorithm,

$$
\log A_{ij}^{(2k+1)} - \log A_{ij}^{(2k)} = -\log \left(\sum_{j'} A_{ij'}^{(2k)} \right),
$$

which depends only on i.

- Hence, each row vector of $X^{(e)}$ is parallel to the all-one vector $(1, \cdots, 1)$.

Proof

- On the other hand, consider the m-representation $Y^{(m)}$ of any tangent vector of Π_1 at $A^{(2k+1)}$.
- Then, from definition of Π_1, each row vector of $Y^{(m)}$ is orthogonal to the all-one vector.
- Therefore, X and Y are orthogonal with respect to the Fisher metric:
$$\langle X, Y \rangle = \sum_{i,j} X_{ij}^{(e)} Y_{ij}^{(m)} = 0.$$
- Hence, the e-geodesic from $A^{(2k)}$ to $A^{(2k+1)}$ is orthogonal to Π_1.

Sinkhorn minimizes KL divergence

- Kullback–Leibler divergence

$$D(B||A) = \sum_{i,j} \left(B_{ij} \log \frac{B_{ij}}{A_{ij}} - B_{ij} + A_{ij} \right)$$

Corollary (Csiszár, 1975)

Each iteration of the Sinkhorn algorithm minimizes KL divergence:

$$D(A^{(2k+1)}||A^{(2k)}) = \min_{B \in \Pi_1} D(B||A^{(2k)})$$

$$D(A^{(2k+2)}||A^{(2k+1)}) = \min_{B \in \Pi_2} D(B||A^{(2k+1)})$$

Convergence of Sinkhorn algorithm

Theorem (Csiszár, 1975)

The Sinkhorn algorithm converges to the e-projection A^* of A onto $\Pi_1 \cap \Pi_2$:

$$D(A^*||A) = \min_{B \in \Pi_1 \cap \Pi_2} D(B||A)$$

- Proof is based on the generalized Pythagorean theorem:

$$D(A^*||A) = D(A^*||A^{(K)}) + \sum_{k=1}^{K} D(A^{(k)}||A^{(k-1)}), \quad K \to \infty$$

Generalized Pythagorean theorem

If the e-geodesic from A_1 to A_2 and the m-geodesic from A_2 to A_3 are orthogonal at A_2 w.r.t. the Fisher metric, then

$$D(A_3||A_1) = D(A_3||A_2) + D(A_2||A_1)$$

Operator scaling

From matrix to operator

- Recently, generalization of matrix scaling to positive maps (operator scaling) is becoming more and more important.
  - theoretical computer science (Gurvits, 2004; Garg et al., 2019+)
  - mathematical physics (Georgiou and Pavon, 2015)

- Gurvits (2004) extended the Sinkhorn algorithm to operator scaling and several authors have investigated its properties (Idel, 2016; Garg et al., 2019+).

CP map and Kraus representation

- A linear map $T : \mathbb{C}^{n\times n} \to \mathbb{C}^{n\times n}$ is called completely positive (CP) if it has the Kraus representation:

$$T(X) = \sum_k V_k X V_k^\dagger$$

- Then, the dual map $T^* : \mathbb{C}^{n\times n} \to \mathbb{C}^{n\times n}$ is also CP with the Kraus representation

$$T^*(X) = \sum_k V_k^\dagger X V_k$$

- In quantum information theory, quantum operations are described by trace-preserving CP (TPCP) maps.

$$X \succeq O,\ \mathrm{tr}\, X = 1 \ \Rightarrow\ T(X) \succeq O,\ \mathrm{tr}\, T(X) = 1$$

Choi–Jamiolkowski representation

- linear map $T : \mathbb{C}^{n\times n} \to \mathbb{C}^{n\times n}$
- Choi–Jamiolkowski representation of T ($n^2 \times n^2$ matrix)
  - E_{ij} : $n \times n$ matrix with $(E_{ij})_{i'j'} = \delta_{ii'}\delta_{jj'}$

$$\mathrm{CH}(T) = \begin{pmatrix} T(E_{11}) & T(E_{12}) & \cdots & T(E_{1n}) \\ T(E_{21}) & T(E_{22}) & \cdots & T(E_{2n}) \\ \vdots & \vdots & \ddots & \vdots \\ T(E_{n1}) & T(E_{n2}) & \cdots & T(E_{nn}) \end{pmatrix}$$

Theorem (Choi, 1975)

$$T : \text{completely positive} \;\Leftrightarrow\; \mathrm{CH}(T) \succeq O$$

- We identify each CP map T with its Choi–Jamiolkowski representation $\mathrm{CH}(T)$ in the following.

Kronecker product and partial trace

- Kronecker product $\otimes : \mathbb{C}^{n\times n} \times \mathbb{C}^{n\times n} \to \mathbb{C}^{n^2\times n^2}$

$$A \otimes B = \begin{pmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{n1}B & \cdots & a_{nn}B \end{pmatrix}$$

- partial trace $\mathrm{tr}_k : \mathbb{C}^{n^2\times n^2} \to \mathbb{C}^{n\times n}$ (linear)

$$\mathrm{tr}_1(A \otimes B) = (\mathrm{tr} A)B, \quad \mathrm{tr}_2(A \otimes B) = (\mathrm{tr} B)A$$

- When $n = 2$ and $C \in \mathbb{C}^{4\times 4}$,

$$\mathrm{tr}_1(C) = \begin{pmatrix} C_{11} + C_{33} & C_{12} + C_{34} \\ C_{21} + C_{43} & C_{22} + C_{44} \end{pmatrix}$$

$$\mathrm{tr}_2(C) = \begin{pmatrix} C_{11} + C_{22} & C_{13} + C_{24} \\ C_{31} + C_{42} & C_{33} + C_{44} \end{pmatrix}$$

Operator scaling problem

- Input: CP map T ($\mathbb{C}^{n\times n} \to \mathbb{C}^{n\times n}$)
- Output: $L \in \mathbb{C}^{n\times n}$ and $R \in \mathbb{C}^{n\times n}$ such that

$$T_{L,R}(I) = T^*_{L,R}(I) = I,$$

where

$$T_{L,R}(X) := LT(RXR^\dagger)L^\dagger = \sum_k \tilde{V}_k X \tilde{V}_k^\dagger, \quad \tilde{V}_k = LV_kR$$

- In the Choi–Jamiolkowski representation,

$$\mathrm{CH}(T_{L,R}) = (R \otimes L)\mathrm{CH}(T)(R \otimes L)$$

$$T_{L,R}(I) = I \;\Leftrightarrow\; \mathrm{tr}_1(\mathrm{CH}(T_{L,R})) = I$$

$$T^*_{L,R}(I) = I \;\Leftrightarrow\; \mathrm{tr}_2(\mathrm{CH}(T_{L,R})) = I$$

Remark: relation to Edmonds problem

Edmonds problem (Edmonds, 1964)

For $n \times n$ matrices $A_1, \ldots, A_k$, decide if

$$P_A(x_1, \ldots, x_k) = \det\left(\sum_{i=1}^{k} x_i A_i\right) \equiv 0.$$

- This problem has a randomized polynomial time algorithm.
  - random substitution of $x_1, \ldots, x_k$
- However, it is not known whether a deterministic polynomial time algorithm exists or not.

- Gurvits (2004) gave a deterministic polynomial time algorithm for certain classes of inputs (Edmonds–Rado class) by extending the Sinkhorn algorithm to operator scaling.
  - Garg et al. (2019+) presented a detailed complexity analysis.

Operator Sinkhorn algorithm (Gurvits, 2004)

- Initialize $T_0 = T$ and $A = B = I$
- Iterate the following for $k = 0, 1, 2, \dots$ until convergence

$$T_0 \to T_1 \to T_2 \to \cdots \to T_*$$

- left normalization ($\to T_{2k+1}(I) = I$)

$$T_{2k+1}(X) = T_{2k}(I)^{-1/2} T_{2k}(X) T_{2k}(I)^{-1/2}$$

$$L \leftarrow T_{2k}(I)^{-1/2} L$$

- right normalization ($\to T^*_{2k+2}(I) = I$)

$$T_{2k+2}(X) = T_{2k+1}(T^*_{2k+1}(I)^{-1/2} X T^*_{2k+1}(I)^{-1/2})$$

$$R \leftarrow T^*_{2k+1}(I)^{-1/2} R$$

Operator Sinkhorn algorithm

- We use the Choi–Jamiolkowski representation for convenience.

$$
\begin{aligned}
\Pi &= \{\mathrm{CH}(T) \mid T : \text{completely positive},\ \mathrm{tr}\, T(I) = \mathrm{tr}\, T^*(I) = n\} \\
&= \{\rho \succeq O \mid \mathrm{tr}\, \rho = n\} \\
\Pi_1 &= \{\mathrm{CH}(T) \mid T : \text{completely positive},\ T(I) = I\} \\
&= \{\rho \succeq O \mid \mathrm{tr}_1(\rho) = I\} \subset \Pi \\
\Pi_2 &= \{\mathrm{CH}(T) \mid T : \text{completely positive},\ T^*(I) = I\} \\
&= \{\rho \succeq O \mid \mathrm{tr}_2(\rho) = I\} \subset \Pi
\end{aligned}
$$

- Putting $\rho_k := \mathrm{CH}(T_k)$, each iteration of operator Sinkhorn is written as

$$\rho_{2k+1} = (I \otimes T_{2k}(I)^{-1/2})\rho_{2k}(I \otimes T_{2k}(I)^{-1/2}) \in \Pi_1$$

$$\rho_{2k+2} = (T^*_{2k+1}(I)^{-1/2} \otimes I)\rho_{2k+1}(T^*_{2k+1}(I)^{-1/2} \otimes I) \in \Pi_2$$

Information geometry of operator Sinkhorn algorithm

operator Sinkhorn = alternating projection?

- It is not clear whether the operator Sinkhorn algorithm can be viewed as alternating projection w.r.t. some divergence.
  - open problem (Georgiou and Pavon, 2015; Gurvits, 2004; Idel, 2016)

matrix scaling	KL divergence
operator scaling	?

- We investigate the operator Sinkhorn algorithm from the viewpoint of quantum information geometry (Amari and Nagaoka, 2000; Fujiwara, 2015).

classical	quantum
$p \geq 0, \sum_k p_k = 1$	$\rho \succeq O, \mathrm{tr}\, \rho = 1$

Density matrix

- In quantum information theory, a quantum state is described by a density matrix ρ satisfying

$$\rho \succeq O, \quad \mathrm{tr}\ \rho = 1.$$

- The operator Sinkhorn algorithm is viewed as updating density matrices:

$$\begin{aligned} \Pi &= \{\rho \succeq O \mid \mathrm{tr}\ \rho = n\} \\ \Pi_1 &= \{\rho \succeq O \mid \mathrm{tr}_1(\rho) = I\} \subset \Pi \\ \Pi_2 &= \{\rho \succeq O \mid \mathrm{tr}_2(\rho) = I\} \subset \Pi \end{aligned}$$

$$\rho_{2k+1} = \mathrm{CH}(T_{2k+1}) \in \Pi_1, \quad \rho_{2k+2} = \mathrm{CH}(T_{2k+2}) \in \Pi_2$$

Riemannian metric for quantum states

- In classical information geometry, the Fisher metric is the only monotone metric (Cencov's theorem).

- However, in quantum information geometry, monotone metrics are not unique.
  - Each monotone metric is characterized by an operator monotone function (Petz, 1996).
  - Each monotone metric induces its own e-connection.

- symmetric logarithmic derivative (SLD) metric

$$g_\rho^{\mathrm{S}}(X,Y) = \frac{1}{2}\mathrm{tr}\,(L_X^{\mathrm{S}}\rho + \rho L_X^{\mathrm{S}})L_Y^{\mathrm{S}}, \quad X\rho = \frac{1}{2}(L_X^{\mathrm{S}}\rho + \rho L_X^{\mathrm{S}})$$

- right logarithmic derivative (RLD) metric
- Bogoliubov metric

$$g_\rho^{\mathrm{B}}(X,Y) = \mathrm{tr}\,(X\rho)(Y\log\rho)$$

operator Sinkhorn = alternating e-projections (SLD) !

- SLD metric

$$g^{\mathrm{S}}_{\rho}(X,Y) = \frac{1}{2}\mathrm{tr}\,(L^{\mathrm{S}}_X\rho + \rho L^{\mathrm{S}}_X)L^{\mathrm{S}}_Y, \quad X\rho = \frac{1}{2}(L^{\mathrm{S}}_X\rho + \rho L^{\mathrm{S}}_X)$$

- e-geodesic from ρ to σ under the SLD metric ($0 \leq t \leq 1$)

$$\gamma(t) = K^t \rho K^t, \quad K = \rho^{-1} \# \sigma$$

Theorem (M. and Soma, 2022)

Each iteration of the operator Sinkhorn algorithm is the unique e-projection w.r.t. SLD metric: the e-geodesic from ρ_{2k} to ρ_{2k+1} (from ρ_{2k+1} to ρ_{2k+2}) is orthogonal to Π_1 (Π_2) w.r.t. the SLD metric.

- Does it minimize some divergence ??

proof

- The e-geodesic from ρ_{2k} to ρ_{2k+1} is

$$\gamma(t) = K^t \rho_{2k} K^t, \quad K = \rho_{2k}^{-1} \# \rho_{2k+1} = I \otimes T_{2k}(I)^{-1/2}$$

- The e-representation L_X^{S} of the tangent vector X of the e-geodesic at ρ_{2k+1} is the solution of the Lyapunov equation:

$$\frac{1}{2}(L_X^{\mathrm{S}} \rho_{2k+1} + \rho_{2k+1} L_X^{\mathrm{S}}) = \gamma'(1) = (\log K)\rho_{2k+1} + \rho_{2k+1}(\log K)$$

- Since the solution of the Lyapunov equation is unique,

$$L_X^{\mathrm{S}} = 2 \log K = -I \otimes \log T_{2k}(I)$$

- Therefore, X is orthogonal to Π_1 w.r.t. SLD metric.
- Uniqueness is shown similarly.
  - not from generalized Pythagorean theorem

Quantum relative entropy (failed)

- quantum relative entropy
  - a quantum analogue of KL divergence

$$D(\rho||\sigma) = \mathrm{tr}\ \rho(\log\rho - \log\sigma)$$

- It induces a dually flat structure with the Bogoliubov metric.
  - It is the only case where the e-connection becomes torsion-free (Nagaoka)
- However, this e-projection does not coincide with the operator Sinkhorn iteration..
- 「皮肉なことに、この幾何構造は、これまでの量子統計学の進展の中で何らの重要性も見いだされていないのである。つまり量子統計学的に意味を持つ情報幾何構造の探求を目指すならば、それは必然的に双対平坦多様体という楽園からの訣別を伴うことになる。」（藤原, 2015）

Another divergence

- capacity

$$\mathrm{cap}(T) = \inf_{X \succ O} \frac{\det T(X)}{\det X} \geq 0$$

- From an analogy to matrix scaling, we expect

$$-\log \mathrm{cap}(T) = \min_{\rho \in \Pi_1 \cap \Pi_2} D(\rho || \mathrm{CH}(T))$$

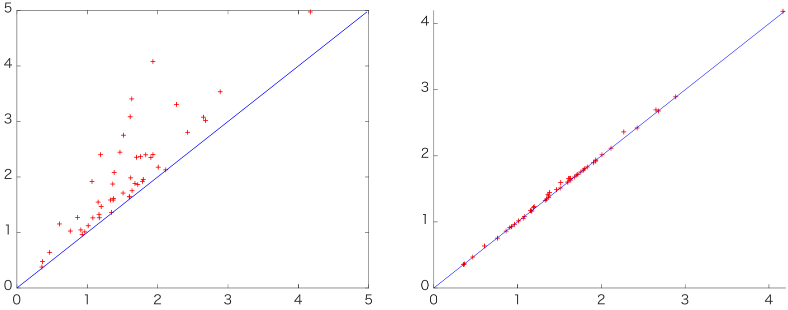
- By considering the convex duality, we can guess

$$D(\rho || \sigma) = 2 \,\mathrm{tr}\, \rho \log(\rho \# \sigma^{-1}),$$

  where $A \# B = A^{1/2}(A^{-1/2} B A^{-1/2})^{1/2} A^{1/2}$ is the matrix geometric mean (Bhatia, 1997).
- In fact, Nagaoka (1994) already discussed this divergence with relation to the SLD metric.
  - To avoid the torsion of e-connection, he considered one-dimensional manifolds.

Numerical check

- $-\log \mathrm{cap}(T)$ (x-axis) v.s. divergence (y-axis)
- left: quantum relative entropy
- right: geometric mean divergence



- The geometric mean divergence has better fit (but not exact?)

Summary

- Operator scaling is a generalization of matrix scaling with many applications.
- We investigated the operator Sinkhorn algorithm from the viewpoint of quantum information geometry.

matrix scaling	KL divergence	Fisher metric
operator scaling	???	SLD metric

- Future work: divergence in operator Sinkhorn
  - quantum analogue of KL divergence is not unique
  - generalized Pythagorean theorem for statistical manifolds admitting torsion (Henmi and Matsuzoe, 2019) ?

References

- S. Amari and H. Nagaoka. *Methods of Information Geometry*. Oxford University Press, 2000.
- R. Bhatia. *Matrix Analysis*. Springer, 1997.
- M. D. Choi. Completely positive linear maps on complex matrices. *Linear Algebra and its Applications*, 10, 285–290, 1975.
- I. Csiszár. I-divergence geometry of probability distributions. *The Annals of Probability*, 3, 146–158, 1975.
- A. Fujiwara. *Foundations of Information Geometry*. Makino-shoten, 2015. (in Japanese)
- A. Garg, L. Gurvits, R. Oliveira and A. Wigderson. Operator Scaling: Theory and Applications. *Foundations of Computational Mathematics*, to appear.

References

- T. T. Georgiou and M. Pavon. Positive contraction mappings for classical and quantum Schrödinger systems. *Journal of Mathematical Physics*, 56, 033301, 2015.
- L. Gurvits. Classical complexity and quantum entanglement. *Journal of Computer and System Sciences*, 69, 448–484, 2004.
- M. Idel. A review of matrix scaling and Sinkhorn's normal form for matrices and positive maps. arXiv:1609.06349.
- T. Matsuda and T. Soma. Information geometry of operator scaling. *Linear Algebra and Its Applications*, 649, 240–267, 2022.
- R. Morioka and K. Tsuda. Information geometry of input-output table. Technical Report IEICE, 110, 161–168, 2011. (in Japanese)
- H. Nagaoka. Differential geometrical aspects of quantum state estimation and relative entropy. In *Quantum Communication, Computing, and Measurement*, Plenum Press, 1994.

References

- D. Petz. Monotone metrics on matrix spaces. *Linear Algebra and its Applications*, 244, 81–96, 1996.
- G. Peyré and M. Cuturi. Computational Optimal Transport. *Foundations and Trends in Machine Learning*, 11, 355–607, 2019.
- S. Reich. Data assimilation: The Schrödinger perspective. *Acta Numerica*, 28, 635–711, 2019.
- R. Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices, *The Annals of Mathematical Statistics*, 35, 876–879, 1964.
- R. Sinkhorn and P. Knopp. Concerning nonnegative matrices and doubly stochastic matrices, *Pacific Journal of Mathematics*, 21, 343–348, 1967.