Matrix Scaling & Entropy-Regularized Optimal Transport

Rafael Oliveira
University of Waterloo

Fields Institute
July 6, 2025

Kantorovich OT

- Given probabilities, $\vec{r}, \vec{c} \in \mathbb{R}^n_{\geq 0}$, let

$$\Pi(\vec{r}, \vec{c}) := \{P \in (\mathbb{R}_{\geq 0})^{n \times n} \mid P\vec{1} = \vec{r},\ P^T\vec{1} = \vec{c}\}$$

  i.e. the set of transport plans from $\vec{r}$ to $\vec{c}$.

Kantorovich OT

- Given probabilities, $\vec{r}, \vec{c} \in \mathbb{R}^n_{\geq 0}$, let

$$\Pi(\vec{r}, \vec{c}) := \{P \in (\mathbb{R}_{\geq 0})^{n \times n} \mid P\vec{1} = \vec{r},\ P^T\vec{1} = \vec{c}\}$$

  i.e. the set of transport plans from $\vec{r}$ to $\vec{c}$.

- Kantorovich OT with cost matrix $C \in (\mathbb{R} \cup \{\infty\})^{n \times n}$

$$\begin{aligned} L_C(\vec{r}, \vec{c}) \quad &:= \quad \min \langle P, C \rangle \\ &\qquad \text{s.t. } P \in \Pi(\vec{r}, \vec{c}) \end{aligned}$$

  where $\langle P, C \rangle := \sum_{i,j} P_{ij} \cdot C_{ij}$

Entropic Regularization of OT

- Entropy of a transport plan:

$$H(P) := -\sum_{i,j} P_{ij} \cdot (\log P_{ij} - 1)$$

Entropic Regularization of OT

- Entropy of a transport plan:

$$H(P) := -\sum_{i,j} P_{ij} \cdot (\log P_{ij} - 1)$$

- Entropic regularization:

$$L_C^\lambda(\vec{r}, \vec{c}) := \min \langle P, C \rangle - \lambda \cdot H(P)$$
$$\text{s.t. } P \in \Pi(\vec{r}, \vec{c})$$

Entropic Regularization of OT

- Entropy of a transport plan:

$$H(P) := -\sum_{i,j} P_{ij} \cdot (\log P_{ij} - 1)$$

- Entropic regularization:

$$L_C^\lambda(\vec{r}, \vec{c}) := \min \langle P, C \rangle - \lambda \cdot H(P)$$
$$\text{s.t. } P \in \Pi(\vec{r}, \vec{c})$$

- Why regularize? (part 1)

  Entropy is 1-strongly concave, and thus $\langle P, C \rangle - \lambda \cdot H(P)$ becomes λ-strongly convex!

  - Unique minimizer
  - Minimizer converges to (an) optimal solution as $\lambda \to 0$

    Optimal solution with maximal entropy.
  - But do we gain anything computationally?

Entropic Regularization of OT

Optimal solution has the form: (Lagrange multipliers)

$$\hat{P} = \operatorname{diag}(\vec{x}) A \operatorname{diag}(\vec{y})$$

where $\vec{x}, \vec{y} \in \mathbb{R}^n_{>0}$ and $\hat{P} \cdot \vec{1} = \vec{r}$, $(\hat{P})^T \cdot \vec{1} = \vec{c}$.

$\hat{P}$ is a scaling of A with correct marginals.

Matrix Scaling

- Simplex:

$$\Sigma_r := \{\vec{a} \in \mathbb{R}^n_{\geq 0} \mid \|\vec{a}\|_1 = 1\}$$

- For any matrix B, let $r(B) := B \cdot \vec{1}$ and $c(B) := B^T \cdot \vec{1}$.

Matrix Scaling

- Simplex:

$$\Sigma_r := \{\vec{a} \in \mathbb{R}^n_{\geq 0} \mid \|\vec{a}\|_1 = 1\}$$

- For any matrix B, let $r(B) := B \cdot \vec{1}$ and $c(B) := B^T \cdot \vec{1}$.
- Distance to marginals $(\vec{r}, \vec{c}) \in \Sigma_n \times \Sigma_n$

$$\Delta_{\vec{r},\vec{c}}(A) := \sqrt{\sum_{i=1}^{n}(r_i(A) - \vec{r}_i)^2 + \sum_{j=1}^{n}(c_j(A) - \vec{c}_j)^2}$$

Matrix Scaling

- Simplex:

$$\Sigma_r := \{\vec{a} \in \mathbb{R}^n_{\geq 0} \mid \|\vec{a}\|_1 = 1\}$$

- Distance to marginals $(\vec{r}, \vec{c}) \in \Sigma_n \times \Sigma_n$

$$\Delta_{\vec{r},\vec{c}}(A) := \sqrt{\sum_{i=1}^{n}(r_i(A) - \vec{r}_i)^2 + \sum_{j=1}^{n}(c_j(A) - \vec{c}_j)^2}$$

- Matrix Scaling Problem
  - **Input:** $A \in (\mathbb{R}_{\geq 0})^{n \times n}$, $\vec{r} \in \Sigma_n$, $\vec{c} \in \Sigma_n$, $\varepsilon > 0$
  - **Output:**
    - YES, if there are $\mathbf{x}_\varepsilon \in \mathbb{R}^n_{>0}, \mathbf{y}_\varepsilon \in \mathbb{R}^n_{>0}$ such that

$$\Delta_{\vec{r},\vec{c}}(\mathrm{diag}(\mathbf{x}_\varepsilon) A \,\mathrm{diag}(\mathbf{y}_\varepsilon)) \leq \varepsilon$$

      in this case, also give me $\mathbf{x}_\varepsilon, \mathbf{y}_\varepsilon$.
    - NO, otherwise

Matrix Scaling

- Simplex:

$$\Sigma_r := \{\vec{a} \in \mathbb{R}^n_{\geq 0} \mid \|\vec{a}\|_1 = 1\}$$

- Matrix Scaling Problem
  - **Input:** $A \in (\mathbb{R}_{\geq 0})^{n \times n}$, $\vec{r} \in \Sigma_n$, $\vec{c} \in \Sigma_n$, $\varepsilon > 0$
  - **Output:**
    - YES, if there are $\mathbf{x}_\varepsilon \in \mathbb{R}^n_{>0}, \mathbf{y}_\varepsilon \in \mathbb{R}^n_{>0}$ such that

$$\Delta_{\vec{r},\vec{c}}(\operatorname{diag}(\mathbf{x}_\varepsilon) A \operatorname{diag}(\mathbf{y}_\varepsilon)) \leq \varepsilon$$

      in this case, also give me $\mathbf{x}_\varepsilon, \mathbf{y}_\varepsilon$.
    - NO, otherwise
- Can we solve this efficiently?

Matrix Scaling

- Simplex:

$$\Sigma_r := \{\vec{a} \in \mathbb{R}^n_{\geq 0} \mid \|\vec{a}\|_1 = 1\}$$

- Matrix Scaling Problem
  - **Input:** $A \in (\mathbb{R}_{\geq 0})^{n \times n}$, $\vec{r} \in \Sigma_n$, $\vec{c} \in \Sigma_n$, $\varepsilon > 0$
  - **Output:**
    - YES, if there are $\mathbf{x}_\varepsilon \in \mathbb{R}^n_{>0}, \mathbf{y}_\varepsilon \in \mathbb{R}^n_{>0}$ such that

$$\Delta_{\vec{r},\vec{c}}(\operatorname{diag}(\mathbf{x}_\varepsilon) A \operatorname{diag}(\mathbf{y}_\varepsilon)) \leq \varepsilon$$

      in this case, also give me $\mathbf{x}_\varepsilon, \mathbf{y}_\varepsilon$.
    - NO, otherwise
- Can we solve this efficiently?
- From now on: $\vec{r} = \vec{c} = \vec{1}/n$.
  In this case:

$$\Delta_{ds}(A) := \Delta_{\vec{1},\vec{1}}(A) := \sqrt{\sum_{i=1}^{n}(r_i(A) - 1/n)^2 + \sum_{j=1}^{n}(c_j(A) - 1/n)^2}$$

Matrix Scaling (in and) outside of OT

- Schrödinger's bridge problem (1931)
- Kruithof 1937: telephone engineering
- Stone 1962: economics & accounting
- Sinkhorn 1964: statistics
- preconditioning subroutine in numerical computations
- Cuturi 2013: computational optimal transport
  - machine learning
  - computer graphics
- Linial, Samorodnitski, Wigderson 2000:
  - deterministic approximation of permanent
  - bipartite matching
- Extremal combinatorics: fractional Sylvester-Gallai configurations
- more – **[Idel 2016]**, **[Garg O. 2018]**, **[Peyre Cuturi 2019]**

Matrix Scaling - (Simplified) Algorithm

Basic operations on a matrix $B \in (\mathbb{R}_{\geq 0})^{n \times n}$:

- $r(B) := (r_1(B), \ldots, r_n(B))$
- $c(B) := (c_1(B), \ldots, c_n(B))$

$$\text{row-normalize}(B) := \det(\text{diag}(r(B)))^{1/n} \cdot \text{diag}(r(B))^{-1} \cdot B$$

$$\text{col-normalize}(B) := \det(\text{diag}(c(B)))^{1/n} \cdot B \cdot \text{diag}(c(B))^{-1}$$

Only scaling by determinant 1 matrices.

Matrix Scaling - (Simplified) Algorithm

- On input $A \in (\mathbb{R}_{\geq 0})^{n \times n}, \vec{r} = \vec{c} = \vec{1}/n$ and $\varepsilon > 0$
- **Algorithm 1:**
  1. Let $A^{(0)} := \text{col-normalize}(A)$
  2. For $t = 0, \ldots, T$:
     - $A^{(2t+1)} = \text{row-normalize}(A^{(2t)})$
     - $A^{(2t+2)} = \text{col-normalize}(A^{(2t+1)})$
  3. If for any $\tau \in [2T + 2]$ we have $\Delta_{ds}(A^{(\tau)}) \leq \varepsilon$ return YES. Else, return NO.

Matrix Scaling - (Simplified) Algorithm

- On input $A \in (\mathbb{R}_{\geq 0})^{n \times n}, \vec{r} = \vec{c} = \vec{1}/n$ and $\varepsilon > 0$
- **Algorithm 1:**
  1. Let $A^{(0)} := \text{col-normalize}(A)$
  2. For $t = 0, \ldots, T$:
     - $A^{(2t+1)} = \text{row-normalize}(A^{(2t)})$
     - $A^{(2t+2)} = \text{col-normalize}(A^{(2t+1)})$
  3. If for any $\tau \in [2T + 2]$ we have $\Delta_{ds}(A^{(\tau)}) \leq \varepsilon$ return YES. Else, return NO.
- Extremely useful in practice (independently discovered multiple times)

Matrix Scaling - (Simplified) Algorithm

- On input $A \in (\mathbb{R}_{\geq 0})^{n \times n}, \vec{r} = \vec{c} = \vec{1}/n$ and $\varepsilon > 0$
- **Algorithm 1:**
  1. Let $A^{(0)} := \text{col-normalize}(A)$
  2. For $t = 0, \ldots, T$:
     - $A^{(2t+1)} = \text{row-normalize}(A^{(2t)})$
     - $A^{(2t+2)} = \text{col-normalize}(A^{(2t+1)})$
  3. If for any $\tau \in [2T + 2]$ we have $\Delta_{ds}(A^{(\tau)}) \leq \varepsilon$ return YES. Else, return NO.
- Extremely useful in practice (independently discovered multiple times)
- **Questions:** if we want to decide whether A is approximately scalable...
  - does this algorithm always work?
  - how should we choose T and ε?

Examples

- Example 1:

Examples

- Example 2:

Examples

- Are these the only cases?

Matrix Scaling - Version 2

- Matrix Scaling Problem (norm minimization)
  - **Input:** $A \in (\mathbb{R}_{\geq 0})^{n \times n}$, $\varepsilon > 0$
  - **Output:**
    - YES, if there are $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n_{>0}$ such that

$$\frac{\|\mathrm{diag}(\mathbf{x}) A \,\mathrm{diag}(\mathbf{y})\|}{\det(\mathrm{diag}(\mathbf{x}))^{1/n} \cdot \det(\mathrm{diag}(\mathbf{y}))^{1/n}} \leq \varepsilon$$

      in this case, also give me $\mathbf{x}, \mathbf{y}$.
    - NO, otherwise

Matrix Scaling - Version 2

- Matrix Scaling Problem (norm minimization)
  - **Input:** $A \in (\mathbb{R}_{\geq 0})^{n \times n}$, $\varepsilon > 0$
  - **Output:**
    - YES, if there are $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n_{>0}$ such that

$$\frac{\|\mathrm{diag}(\mathbf{x}) A \,\mathrm{diag}(\mathbf{y})\|}{\det(\mathrm{diag}(\mathbf{x}))^{1/n} \cdot \det(\mathrm{diag}(\mathbf{y}))^{1/n}} \leq \varepsilon$$

      in this case, also give me $\mathbf{x}, \mathbf{y}$.
    - NO, otherwise

Matrix Scaling - Version 2

- Matrix Scaling Problem (norm minimization)
  - **Input:** $A \in (\mathbb{R}_{\geq 0})^{n \times n}$, $\varepsilon > 0$
  - **Output:**
    - YES, if there are $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n_{>0}$ such that

$$\frac{\|\mathrm{diag}(\mathbf{x}) A \,\mathrm{diag}(\mathbf{y})\|}{\det(\mathrm{diag}(\mathbf{x}))^{1/n} \cdot \det(\mathrm{diag}(\mathbf{y}))^{1/n}} \leq \varepsilon$$

      in this case, also give me $\mathbf{x}, \mathbf{y}$.
    - NO, otherwise
- Function of interest:

$$F_A(\mathbf{x}, \mathbf{y}) := \log\left(\frac{\|\mathrm{diag}(\mathbf{x}) A \,\mathrm{diag}(\mathbf{y})\|}{\det(\mathrm{diag}(\mathbf{x}))^{1/n} \cdot \det(\mathrm{diag}(\mathbf{y}))^{1/n}}\right)$$

$$F_A(\mathbf{x}, \mathbf{y}) = \log\left(\|\mathrm{diag}(\mathbf{x}) A \,\mathrm{diag}(\mathbf{y})\|\right) - \frac{1}{n} \cdot \log\left(\det(\mathrm{diag}(\mathbf{x}) \cdot \mathrm{diag}(\mathbf{y}))\right)$$

Matrix Scaling - Version 2

- Matrix Scaling Problem (norm minimization)
  - **Input:** $A \in (\mathbb{R}_{\geq 0})^{n \times n}$, $\varepsilon > 0$
  - **Output:**
    - YES, if there are $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n_{>0}$ such that

$$\frac{\|\mathrm{diag}(\mathbf{x}) A \,\mathrm{diag}(\mathbf{y})\|}{\det(\mathrm{diag}(\mathbf{x}))^{1/n} \cdot \det(\mathrm{diag}(\mathbf{y}))^{1/n}} \leq \varepsilon$$

      in this case, also give me $\mathbf{x}, \mathbf{y}$.
    - NO, otherwise
- Function of interest:

$$F_A(\mathbf{x}, \mathbf{y}) := \log \left(\frac{\|\mathrm{diag}(\mathbf{x}) A \,\mathrm{diag}(\mathbf{y})\|}{\det(\mathrm{diag}(\mathbf{x}))^{1/n} \cdot \det(\mathrm{diag}(\mathbf{y}))^{1/n}} \right)$$

if B scaling of A by determinant 1 diagonal matrices $\mathrm{diag}(\mathbf{x}), \mathrm{diag}(\mathbf{y})$, we have

$$F_A(\mathbf{x}, \mathbf{y}) = F_B(\vec{1}, \vec{1}) = \log \|B\|$$

Matrix Scaling - Version 2

- Matrix Scaling Problem (norm minimization)
  - **Input:** $A \in (\mathbb{R}_{\geq 0})^{n \times n}$, $\varepsilon > 0$
  - **Output:**
    - YES, if there are $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n_{>0}$ such that

$$\frac{\|\mathrm{diag}(\mathbf{x}) A \,\mathrm{diag}(\mathbf{y})\|}{\det(\mathrm{diag}(\mathbf{x}))^{1/n} \cdot \det(\mathrm{diag}(\mathbf{y}))^{1/n}} \leq \varepsilon$$

      in this case, also give me $\mathbf{x}, \mathbf{y}$.
    - NO, otherwise
- Function of interest:

$$F_A(\mathbf{x}, \mathbf{y}) := \log \left(\frac{\|\mathrm{diag}(\mathbf{x}) A \,\mathrm{diag}(\mathbf{y})\|}{\det(\mathrm{diag}(\mathbf{x}))^{1/n} \cdot \det(\mathrm{diag}(\mathbf{y}))^{1/n}} \right)$$

- Want to compute

$$\mathrm{logcap}(A) := \inf_{\mathbf{x}, \mathbf{y} \in \mathbb{R}^n_{>0}} F_A(\mathbf{x}, \mathbf{y})$$

Matrix Scaling - Optimization

- **Question:** Why are these two versions equivalent?

Matrix Scaling - Optimization

- **Question:** Why are these two versions equivalent?
- **Answer:** Convexity of F_A!

Matrix Scaling - Optimization

- **Question:** Why are these two versions equivalent?
- **Answer:** Convexity of F_A!
- Can think of $F_A : \mathbb{R}^{2n} \to \mathbb{R}$ via composition with the exponential map

$$F_A(e^{\mathbf{u}}, e^{\mathbf{v}}) = \log\left(\frac{\|\mathrm{diag}(e^{\mathbf{u}})A\,\mathrm{diag}(e^{\mathbf{v}})\|}{\det(\mathrm{diag}(e^{\mathbf{u}}))^{1/n} \cdot \det(\mathrm{diag}(e^{\mathbf{v}}))^{1/n}}\right)$$

where now $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$.

Matrix Scaling - Optimization

- **Question:** Why are these two versions equivalent?
- **Answer:** Convexity of F_A!
- Gradient of F_A (at $(\vec{1}, \vec{1})$):

$$\langle \nabla F_A, (\mathbf{u}, \mathbf{v}) \rangle = \partial_{t=0} F_A(e^{t\mathbf{u}}, e^{t\mathbf{v}})$$

$$= \frac{1}{\|A\|} \cdot \sum_{i,j=1}^{n} A_{ij}(u_i + v_j) - \frac{1}{n} \cdot \left(\sum_i u_i + \sum_j v_j \right)$$

Matrix Scaling - Optimization

- **Question:** Why are these two versions equivalent?
- **Answer:** Convexity of F_A!
- Gradient of F_A (at $(\vec{1}, \vec{1})$):

$$\begin{aligned}\langle \nabla F_A, (\mathbf{u}, \mathbf{v}) \rangle &= \partial_{t=0} F_A(e^{t\mathbf{u}}, e^{t\mathbf{v}}) \\ &= \frac{1}{\|A\|} \cdot \sum_{i,j=1}^{n} A_{ij}(u_i + v_j) - \frac{1}{n} \cdot \left(\sum_i u_i + \sum_j v_j \right)\end{aligned}$$

- Hence

$$\nabla F_A = \left(\mathbf{r}(A) - \frac{1}{n} \cdot \vec{1},\ \mathbf{c}(A) - \frac{1}{n} \cdot \vec{1} \right)$$

Gradient $\leftrightarrow$ normalized row and column sums!

Matrix Scaling - Optimization

- **Question:** Why are these two versions equivalent?
- **Answer:** Convexity of F_A!
- Gradient of F_A (at $(\vec{1}, \vec{1})$):

$$\begin{aligned}\langle \nabla F_A, (\mathbf{u}, \mathbf{v}) \rangle &= \partial_{t=0} F_A(e^{t\mathbf{u}}, e^{t\mathbf{v}}) \\ &= \frac{1}{\|A\|} \cdot \sum_{i,j=1}^{n} A_{ij}(u_i + v_j) - \frac{1}{n} \cdot \left(\sum_i u_i + \sum_j v_j \right)\end{aligned}$$

- Hence

$$\nabla F_A = \left(\mathbf{r}(A) - \frac{1}{n} \cdot \vec{1},\ \mathbf{c}(A) - \frac{1}{n} \cdot \vec{1} \right)$$

Gradient $\leftrightarrow$ normalized row and column sums!

- Moreover:

$$\Delta_{ds}(A) = \|\nabla F_A\|_2$$

Matrix Scaling

- Can prove that F_A is convex! (exercise)
- Thus we get:
  - Minimizing the norm of gradient of $F_A \leftrightarrow$ minimizing F_A
  - Two formulations are "equivalent"!

Matrix Scaling

- Can prove that F_A is convex! (exercise)
- Thus we get:
  - Minimizing the norm of gradient of $F_A \leftrightarrow$ minimizing F_A
  - Two formulations are "equivalent"!
  - Algorithm 1 is block-coordinate gradient descent!
    - optimize over "row variables," then over "column variables"

Matrix Scaling

- Can prove that F_A is convex! (exercise)
- Thus we get:
  - Minimizing the norm of gradient of $F_A \leftrightarrow$ minimizing F_A
  - Two formulations are "equivalent"!
  - Algorithm 1 is block-coordinate gradient descent!
    - optimize over "row variables," then over "column variables"
  - **Questions:**
    - Can we get a tight quantitative connection?
    - Can we prove convergence of Algorithm 1?

Matrix Scaling - Observations

1. $\operatorname{logcap}(A) = -\infty \Leftrightarrow$ can scale A arbitrarily close to to 0

Matrix Scaling - Observations

1. $\operatorname{logcap}(A) = -\infty \Leftrightarrow$ can scale A arbitrarily close to to 0
2. if $\operatorname{logcap}(A) > -\infty$ gradient cannot be large all the time

Matrix Scaling - Observations

1. $\operatorname{logcap}(A) = -\infty \Leftrightarrow$ can scale A arbitrarily close to to 0
2. if $\operatorname{logcap}(A) > -\infty$ gradient cannot be large all the time
3. Invariant property of scaling action:

$$B := \frac{\operatorname{diag}(\mathbf{x}) A \operatorname{diag}(\mathbf{y})}{\det(\operatorname{diag}(\mathbf{x}))^{1/n} \cdot \det(\operatorname{diag}(\mathbf{y}))^{1/n}}$$

Then for any permutation $\sigma \in S_n$

$$\prod_{i=1}^{n} B_{i\sigma(i)} = \prod_{i=1}^{n} A_{i\sigma(i)}$$

Matrix Scaling - Observations

1. $\operatorname{logcap}(A) = -\infty \Leftrightarrow$ can scale A arbitrarily close to to 0
2. if $\operatorname{logcap}(A) > -\infty$ gradient cannot be large all the time
3. Invariant property of scaling action:

$$B := \frac{\operatorname{diag}(\mathbf{x}) A \operatorname{diag}(\mathbf{y})}{\det(\operatorname{diag}(\mathbf{x}))^{1/n} \cdot \det(\operatorname{diag}(\mathbf{y}))^{1/n}}$$

Then for any permutation $\sigma \in S_n$

$$\prod_{i=1}^{n} B_{i\sigma(i)} = \prod_{i=1}^{n} A_{i\sigma(i)}$$

4. If support of A contains a permutation, then cannot get "too close" to 0 via scaling!

Matrix Scaling - Observations

1. $\operatorname{logcap}(A) = -\infty \Leftrightarrow$ can scale A arbitrarily close to to 0
2. if $\operatorname{logcap}(A) > -\infty$ gradient cannot be large all the time
3. Invariant property of scaling action:

$$B := \frac{\operatorname{diag}(\mathbf{x}) A \operatorname{diag}(\mathbf{y})}{\det(\operatorname{diag}(\mathbf{x}))^{1/n} \cdot \det(\operatorname{diag}(\mathbf{y}))^{1/n}}$$

Then for any permutation $\sigma \in S_n$

$$\prod_{i=1}^{n} B_{i\sigma(i)} = \prod_{i=1}^{n} A_{i\sigma(i)}$$

4. If support of A contains a permutation, then cannot get "too close" to 0 via scaling!
5. Hall's theorem: bipartite graph has perfect matching $\Leftrightarrow$ adjacency matrix has no "big block of zeroes"!

Algorithm 1 - Recap & Remarks

1. Let $A^{(0)} := \text{col-normalize}(A)$
2. For $t = 0, \ldots, T$:
   - $A^{(2t+1)} = \text{row-normalize}(A^{(2t)})$
   - $A^{(2t+2)} = \text{col-normalize}(A^{(2t+1)})$
3. If for any $\tau \in [2T + 2]$ we have $\Delta_{ds}(A^{(\tau)}) \leq \varepsilon$ return YES. Else, return NO.

Algorithm 1 - Recap & Remarks

1. Let $A^{(0)} := \text{col-normalize}(A)$
2. For $t = 0, \ldots, T$:
   - $A^{(2t+1)} = \text{row-normalize}(A^{(2t)})$
   - $A^{(2t+2)} = \text{col-normalize}(A^{(2t+1)})$
3. If for any $\tau \in [2T+2]$ we have $\Delta_{ds}(A^{(\tau)}) \leq \varepsilon$ return YES. Else, return NO.

Remarks:

- A has "big block of zeros" $\Rightarrow \Delta_{ds}(A^{(\tau)}) \geq \frac{1}{2n^2}$ for all $\tau \in [T]$

So let's set $\varepsilon = 1/2n^2$

Algorithm 1 - Recap & Remarks

1. Let $A^{(0)} := \text{col-normalize}(A)$
2. For $t = 0, \ldots, T$:
   - $A^{(2t+1)} = \text{row-normalize}(A^{(2t)})$
   - $A^{(2t+2)} = \text{col-normalize}(A^{(2t+1)})$
3. If for any $\tau \in [2T+2]$ we have $\Delta_{ds}(A^{(\tau)}) \leq \varepsilon$ return YES. Else, return NO.

Remarks:

- A has "big block of zeros" $\Rightarrow \Delta_{ds}(A^{(\tau)}) \geq \frac{1}{2n^2}$ for all $\tau \in [T]$
- $A^{(\tau)}$ is a scaling of A by diagonal determinant 1 matrices!

Algorithm 1 - Recap & Remarks

1. Let $A^{(0)} := \text{col-normalize}(A)$
2. For $t = 0, \ldots, T$:
   - $A^{(2t+1)} = \text{row-normalize}(A^{(2t)})$
   - $A^{(2t+2)} = \text{col-normalize}(A^{(2t+1)})$
3. If for any $\tau \in [2T+2]$ we have $\Delta_{ds}(A^{(\tau)}) \leq \varepsilon$ return YES. Else, return NO.

Remarks:

- A has "big block of zeros" $\Rightarrow \Delta_{ds}(A^{(\tau)}) \geq \frac{1}{2n^2}$ for all $\tau \in [T]$
- $A^{(\tau)}$ is a scaling of A by diagonal determinant 1 matrices!
- If $A^{(\tau)} = \text{diag}(\mathbf{x}^{(\tau)}) A \, \text{diag}(\mathbf{y}^{(\tau)})$, let

$$f(\tau) := F_A(\mathbf{x}^{(\tau)}, \mathbf{y}^{(\tau)}) = F_{A^{(\tau)}}(\vec{1}, \vec{1}) = \log \left\| A^{(\tau)} \right\|$$

Algorithm 1 - Recap & Remarks

1. Let $A^{(0)} := \text{col-normalize}(A)$
2. For $t = 0, \ldots, T$:
   - $A^{(2t+1)} = \text{row-normalize}(A^{(2t)})$
   - $A^{(2t+2)} = \text{col-normalize}(A^{(2t+1)})$
3. If for any $\tau \in [2T+2]$ we have $\Delta_{ds}(A^{(\tau)}) \leq \varepsilon$ return YES. Else, return NO.

Remarks:

- A has "big block of zeros" $\Rightarrow \Delta_{ds}(A^{(\tau)}) \geq \frac{1}{2n^2}$ for all $\tau \in [T]$
- $A^{(\tau)}$ is a scaling of A by diagonal determinant 1 matrices!
- If $A^{(\tau)} = \text{diag}(\mathbf{x}^{(\tau)}) A \, \text{diag}(\mathbf{y}^{(\tau)})$, let

$$f(\tau) := F_A(\mathbf{x}^{(\tau)}, \mathbf{y}^{(\tau)}) = F_{A^{(\tau)}}(\vec{1}, \vec{1}) = \log \left\| A^{(\tau)} \right\|$$

- Hence,

$$f(\tau) - f(\tau - 1) = \frac{1}{n} \cdot \log \det \left(\text{diag}(\mathbf{r}(A^{(\tau-1)})) \cdot \text{diag}(\mathbf{c}(A)^{(\tau-1)}) \right)$$

Analysis of Algorithm 1

- Quantitative AM-GM: if $\alpha_1, \ldots, \alpha_n \in \mathbb{R}^n_{>0}$ are such that
  - $\sum_{i=1}^n \alpha_i = n$
  - $\sum_{i=1}^n (\alpha_i - 1)^2 = \alpha$

  then

$$\prod_{i=1}^n \alpha_i \leq \begin{cases} \exp(-\alpha/6), & \text{if } \alpha \leq 1, \\ \exp(-1/6), & \text{otherwise} \end{cases}$$

Analysis of Algorithm 1

- Quantitative AM-GM: if $\alpha_1, \ldots, \alpha_n \in \mathbb{R}^n_{>0}$ are such that
  - $\sum_{i=1}^n \alpha_i = n$
  - $\sum_{i=1}^n (\alpha_i - 1)^2 = \alpha$

  then

$$\prod_{i=1}^n \alpha_i \leq \begin{cases} \exp(-\alpha/6), & \text{if } \alpha \leq 1, \\ \exp(-1/6), & \text{otherwise} \end{cases}$$

- If $\Delta_{ds}(A^{(\tau-1)}) \geq 1/2n^2$ then

$$\det\left(\operatorname{diag}(\mathbf{r}(A^{(\tau-1)})) \cdot \operatorname{diag}(\mathbf{c}(A)^{(\tau-1)}))\right) \leq \exp(-1/12n^2)$$

Analysis of Algorithm 1

- Quantitative AM-GM: if $\alpha_1, \ldots, \alpha_n \in \mathbb{R}^n_{>0}$ are such that
  - $\sum_{i=1}^n \alpha_i = n$
  - $\sum_{i=1}^n (\alpha_i - 1)^2 = \alpha$

  then

$$\prod_{i=1}^n \alpha_i \leq \begin{cases} \exp(-\alpha/6), & \text{if } \alpha \leq 1, \\ \exp(-1/6), & \text{otherwise} \end{cases}$$

- If $\Delta_{ds}(A^{(\tau-1)}) \geq 1/2n^2$ then

$$\det\left(\operatorname{diag}(\mathbf{r}(A^{(\tau-1)})) \cdot \operatorname{diag}(\mathbf{c}(A)^{(\tau-1)}))\right) \leq \exp(-1/12n^2)$$

- Hence

$$f(\tau) - f(\tau-1) = \frac{1}{n} \cdot \log\det\left(\operatorname{diag}(\mathbf{r}(A^{(\tau-1)})) \cdot \operatorname{diag}(\mathbf{c}(A^{(\tau-1)}\right.$$

$$\leq -\frac{1}{12n^3}$$

Analysis of Algorithm 1

- Quantitative AM-GM: if $\alpha_1, \ldots, \alpha_n \in \mathbb{R}^n_{>0}$ are such that
  - $\sum_{i=1}^n \alpha_i = n$
  - $\sum_{i=1}^n (\alpha_i - 1)^2 = \alpha$

  then
$$\prod_{i=1}^n \alpha_i \leq \begin{cases} \exp(-\alpha/6), & \text{if } \alpha \leq 1, \\ \exp(-1/6), & \text{otherwise} \end{cases}$$
- If $\Delta_{ds}(A^{(\tau-1)}) \geq 1/2n^2$ then
$$\det\left(\operatorname{diag}(\mathbf{r}(A^{(\tau-1)})) \cdot \operatorname{diag}(\mathbf{c}(A)^{(\tau-1)})\right) \leq \exp(-1/12n^2)$$
- Hence
$$f(\tau) - f(\tau-1) = \frac{1}{n} \cdot \log\det\left(\operatorname{diag}(\mathbf{r}(A^{(\tau-1)})) \cdot \operatorname{diag}(\mathbf{c}(A^{(\tau-1)}\right.$$
$$\leq -\frac{1}{12n^3}$$
- $\Delta_{ds}(A^{(\tau)}) \geq 1/2n^2$ for all $\tau \in [T] \Rightarrow f(T) \leq f(0) - \frac{T}{12n^3}$

Analysis of Algorithm 1

- if $\Delta_{ds}(A^{(\tau)}) \geq 1/2n^2$ for all $\tau \in [T]$

$$f(T) \leq f(0) - \frac{T}{12n^3} = \log \|A\| - \frac{T}{12n^3}$$

Analysis of Algorithm 1

- if $\Delta_{ds}(A^{(\tau)}) \geq 1/2n^2$ for all $\tau \in [T]$

$$f(T) \leq f(0) - \frac{T}{12n^3} = \log \|A\| - \frac{T}{12n^3}$$

- If support of A has a permutation σ

$$\left\|A^{(\tau)}\right\|^n \geq \prod_{i=1}^{n} A^{(\tau)}_{i\sigma(i)} = \prod_{i=1}^{n} A_{i\sigma(i)} \geq \nu(A)^n$$

where

$$\nu(A) := \min_{(i,j)\in \mathrm{supp}(A)} A_{i,j}$$

Analysis of Algorithm 1

- if $\Delta_{ds}(A^{(\tau)}) \geq 1/2n^2$ for all $\tau \in [T]$

$$f(T) \leq f(0) - \frac{T}{12n^3} = \log \|A\| - \frac{T}{12n^3}$$

- If support of A has a permutation σ

$$\left\|A^{(\tau)}\right\|^n \geq \prod_{i=1}^{n} A^{(\tau)}_{i\sigma(i)} = \prod_{i=1}^{n} A_{i\sigma(i)} \geq \nu(A)^n$$

where

$$\nu(A) := \min_{(i,j)\in \mathrm{supp}(A)} A_{i,j}$$

- Thus

$$f(\tau) = \log \left\|A^{(\tau)}\right\| \geq \log \nu(A), \quad \forall \tau \geq 0$$

Analysis of Algorithm 1

- if $\Delta_{ds}(A^{(\tau)}) \geq 1/2n^2$ for all $\tau \in [T]$

$$f(T) \leq f(0) - \frac{T}{12n^3} = \log \|A\| - \frac{T}{12n^3}$$

- If support of A has a permutation σ

$$\left\|A^{(\tau)}\right\|^n \geq \prod_{i=1}^{n} A^{(\tau)}_{i\sigma(i)} = \prod_{i=1}^{n} A_{i\sigma(i)} \geq \nu(A)^n$$

where

$$\nu(A) := \min_{(i,j)\in \mathrm{supp}(A)} A_{i,j}$$

- Thus

$$f(\tau) = \log \left\|A^{(\tau)}\right\| \geq \log \nu(A), \quad \forall \tau \geq 0$$

- Setting $T = 12n^3 \cdot \log(\|A\| / \nu(A))$ we are done.

Recap

- Proved that a non-negative matrix can be approximately scalable to doubly-stochastic iff its support contains a permutation (i.e. perfect bipartite matching)

Recap

- Proved that a non-negative matrix can be approximately scalable to doubly-stochastic iff its support contains a permutation (i.e. perfect bipartite matching)
- Used a polynomial invariant (permutation polynomial) to lower bound minimum (log) norm

Recap

- Proved that a non-negative matrix can be approximately scalable to doubly-stochastic iff its support contains a permutation (i.e. perfect bipartite matching)
- Used a polynomial invariant (permutation polynomial) to lower bound minimum (log) norm
- The log-norm function is convex (under proper parametrization)

Recap

- Proved that a non-negative matrix can be approximately scalable to doubly-stochastic iff its support contains a permutation (i.e. perfect bipartite matching)
- Used a polynomial invariant (permutation polynomial) to lower bound minimum (log) norm
- The log-norm function is convex (under proper parametrization)
- gradient of log-norm is distance to doubly stochastic

Recap

- Proved that a non-negative matrix can be approximately scalable to doubly-stochastic iff its support contains a permutation (i.e. perfect bipartite matching)
- Used a polynomial invariant (permutation polynomial) to lower bound minimum (log) norm
- The log-norm function is convex (under proper parametrization)
- gradient of log-norm is distance to doubly stochastic
- alternating minimization is block-coordinate descent

Recap

- Proved that a non-negative matrix can be approximately scalable to doubly-stochastic iff its support contains a permutation (i.e. perfect bipartite matching)
- Used a polynomial invariant (permutation polynomial) to lower bound minimum (log) norm
- The log-norm function is convex (under proper parametrization)
- gradient of log-norm is distance to doubly stochastic
- alternating minimization is block-coordinate descent

- Friday's lecture:

  Generalization to quantum operators (and maybe more)!

References I

Ayers, P. and Coudreuse, F. and Gerolin, A. and Oliveira, R. and Ramachandran, A. (2025)

Operator Scaling & Quantum Optimal Transport

Manuscript

Berta, F. and Fawzi, O. and Tomamichel, M. (2017)

On Variational Expressions for Quantum Relative Entropies

Letters in Mathematical Physics, 107, pp.2239-2265.

Gurvits, L. (2004)

Classical complexity and quantum entanglement.

Journal of Computer and System Sciences, 69(3), pp.448-484.

Idel, M. (2016)

A review of matrix scaling and Sinkhorn's normal form for matrices and positive maps

arXiv preprint arXiv:1609.06349

References II

Garg, A. and Oliveira, Rafael (2018)

Recent progress on scaling algorithms and applications

Bulletin of EATCS 2.125

Peyré, G. and Cuturi, M. (2019)

Computational optimal transport: With applications to data science.

Foundations and Trends in Machine Learning, 11(5-6), pp.355-607.

Sinkhorn, R. (1964)

A relationship between arbitrary positive matrices and doubly stochastic matrices.

The annals of mathematical statistics, 35(2), pp.876-879.

Silva Louzeiro, M., Bergmann, R. and Herzog, R. (2022)

Fenchel duality and a separation theorem on Hadamard manifolds.

SIAM Journal on Optimization, 32(2), pp.854-873.

References III

Keyl, M., 2006.

Quantum state estimation and large deviations.

Reviews in Mathematical Physics, 18(01), pp.19-60.

Keyl, M. and Werner, R.F., 2001.

Estimating the spectrum of a density operator.

Physical Review A, 64(5), p.052311.