Matrix Scaling, Entropic Optimal Transport, and Sinkhorn Algorithm: Theory, Applications, and New Perspectives

Hanbaek Lyu

Department of Mathematics

University of Wisconsin – Madison

UW-Madison AI Seminar

Apr. 20, 2026

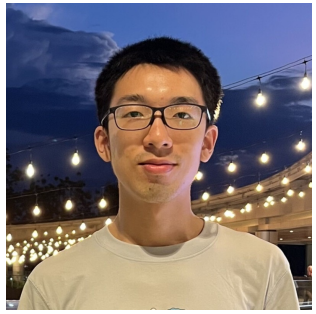
Collaborators and Students

Jakwang Kim
(U. HK, Shenzhen)

Sumit Mukherjee
(Columbia)

William Powell
(5th year Ph.D.)

Danny Duan
(5th year Ph.D.)



Shuqi Bi
(2nd year Ph.D.)

Rahul Choudhary
(5th year CS Ph.D.)

Introduction

Matrix Scaling
Sinkhorn Algorithm
Entropic OT
Wasserstein Distances

- Matrix Scaling Problem:

  Given a nonnegative matrix $\mathbf{A}$ and target row and column sums ($\mathbf{r}$, $\mathbf{c}$),

  find diagonal matrices $\mathbf{D_1}$ and $\mathbf{D_2}$ s.t. $\mathbf{D_1 A D_2}$ satisfies margin ($\mathbf{r}$, $\mathbf{c}$)

- Matrix Scaling Problem:

  Given a nonnegative matrix **A** and target row and column sums (**r**, **c**),

  find diagonal matrices $\mathbf{D_1}$ and $\mathbf{D_2}$ s.t. $\mathbf{D_1 A D_2}$ satisfies margin (**r**, **c**)

  - Sinkhorn's matrix scaling algorithm (1964)
    - Normalize the rows of **A** to match the target row sum **r;** Obtain **A'**
    - Normalize the columns of **A'** to match the target column sum **c;** Obtain **A"**
    - Repeat

- Matrix Scaling Problem:

  Given a nonnegative matrix **A** and target row and column sums (**r**, **c**), find diagonal matrices $\mathbf{D_1}$ and $\mathbf{D_2}$ s.t. $\mathbf{D_1 A D_2}$ satisfies margin (**r**, **c**)

  - Sinkhorn's matrix scaling algorithm (1964)
    - Normalize the rows of **A** to match the target row sum **r;** Obtain **A'**
    - Normalize the columns of **A'** to match the target column sum **c;** Obtain **A"**
    - Repeat
  - Sinkhorn's algorithm is known to solve the following relative entropy minimization problem:

$$\underset{\mathbf{Z}\in\mathcal{T}(\mathbf{r},\mathbf{c})}{\arg\min}\left[D_{KL}(\mathbf{Z}\,\|\,\mathbf{\Lambda})=\sum_{ij}\mathbf{Z}_{ij}\log\left(\mathbf{Z}_{ij}/\mathbf{\Lambda}_{ij}\right)\right]$$

Transportation polytope with margin (**r**, **c**)

- Matrix Scaling Problem:

  Given a nonnegative matrix **A** and target row and column sums (**r**, **c**),

  find diagonal matrices $\mathbf{D_1}$ and $\mathbf{D_2}$ s.t. $\mathbf{D_1 A D_2}$ satisfies margin (**r**, **c**)

  - Sinkhorn's matrix scaling algorithm (1964)
    - Normalize the rows of **A** to match the target row sum **r;** Obtain **A'**
    - Normalize the columns of **A'** to match the target column sum **c;** Obtain **A"**
    - Repeat
  - Sinkhorn's algorithm is known to solve the following relative entropy minimization problem:

$$\underset{\mathbf{Z}\in\mathcal{T}(\mathbf{r},\mathbf{c})}{\arg\min}\left[D_{KL}(\mathbf{Z}\,\|\,\mathbf{\Lambda})=\sum_{ij}\mathbf{Z}_{ij}\log\left(\mathbf{Z}_{ij}/\mathbf{\Lambda}_{ij}\right)\right]$$

Transportation polytope with margin (**r**, **c**)

  - Why? It is in fact the **alternating maximization** on the "dual"!

- Kantorovich Duality and Sinkhorn's algorithm

Primal

$$\min_{\mathbf{Z}\in\mathcal{T}(\mathbf{r},\mathbf{c})} \left[D_{KL}(\mathbf{Z}\,\|\mathbf{W}) \right]$$

Rescaled Matrix

- Kantorovich Duality and Sinkhorn's algorithm

Primal **Dual**

$$\min_{\mathbf{Z}\in\mathcal{T}(\mathbf{r},\mathbf{c})} \left[D_{KL}(\mathbf{Z}\,\|\mathbf{W})\right] = \max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \left[\langle \mathbf{r}, \boldsymbol{\alpha}\rangle + \langle \boldsymbol{\beta}, \mathbf{c}\rangle - \langle \exp(\boldsymbol{\alpha}\oplus\boldsymbol{\beta}), \mathbf{W}\rangle\right]$$

Rescaled Matrix

$\boldsymbol{\alpha}$ = Lagrange multipliers for the row-sum constraint

$\boldsymbol{\beta}$ = Lagrange multipliers for the column-sum constraint

(Schrödinger potentials)

- Kantorovich Duality and Sinkhorn's algorithm

Primal **Dual**

$$\min_{\mathbf{Z}\in\mathcal{T}(\mathbf{r},\mathbf{c})} \left[D_{KL}(\mathbf{Z}\,\|\mathbf{W})\right] = \max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \left[\langle \mathbf{r}, \boldsymbol{\alpha}\rangle + \langle \boldsymbol{\beta}, \mathbf{c}\rangle - \langle \exp(\boldsymbol{\alpha}\oplus\boldsymbol{\beta}), \mathbf{W}\rangle\right]$$

Rescaled Matrix

$\boldsymbol{\alpha}$ = Lagrange multipliers for the row-sum constraint

$\boldsymbol{\beta}$ = Lagrange multipliers for the column-sum constraint

(Schrödinger potentials)

$$\pi^* = \mathrm{Diag}(\exp(\boldsymbol{\alpha}^*))\,\mathbf{W}\,\mathrm{Diag}(\exp(\boldsymbol{\beta}^*))$$

- Kantorovich Duality and Sinkhorn's algorithm

Primal **Dual**

$$\min_{\mathbf{Z}\in\mathcal{T}(\mathbf{r},\mathbf{c})} \left[D_{KL}(\mathbf{Z}\,\|\mathbf{W})\right] = \max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \left[\langle \mathbf{r}, \boldsymbol{\alpha}\rangle + \langle \boldsymbol{\beta}, \mathbf{c}\rangle - \langle \exp(\boldsymbol{\alpha}\oplus\boldsymbol{\beta}), \mathbf{W}\rangle\right]$$

Rescaled Matrix

$\boldsymbol{\alpha}$ = Lagrange multipliers for the row-sum constraint

$\boldsymbol{\beta}$ = Lagrange multipliers for the column-sum constraint

(Schrödinger potentials)

$$\pi^* = \mathrm{Diag}(\exp(\boldsymbol{\alpha}^*))\,\mathbf{W}\,\mathrm{Diag}(\exp(\boldsymbol{\beta}^*))$$

Matrix scaling algorithm
(Iterative Proportional Fitting)

Fit the row sums;
Fit the column sums; etc.

- Kantorovich Duality and Sinkhorn's algorithm

Primal **Dual**

$$\min_{\mathbf{Z}\in\mathcal{T}(\mathbf{r},\mathbf{c})} \left[D_{KL}(\mathbf{Z}\,\|\mathbf{W})\right] = \max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \left[\langle \mathbf{r}, \boldsymbol{\alpha}\rangle + \langle \boldsymbol{\beta}, \mathbf{c}\rangle - \langle \exp(\boldsymbol{\alpha}\oplus\boldsymbol{\beta}), \mathbf{W}\rangle\right]$$

Rescaled Matrix

$\boldsymbol{\alpha}$ = Lagrange multipliers for the row-sum constraint

$\boldsymbol{\beta}$ = Lagrange multipliers for the column-sum constraint

(Schrödinger potentials)

$$\pi^* = \mathrm{Diag}(\exp(\boldsymbol{\alpha}^*))\,\mathbf{W}\,\mathrm{Diag}(\exp(\boldsymbol{\beta}^*))$$

Alternating maximization

Matrix scaling algorithm
(Iterative Proportional Fitting)

Fit the row sums;
Fit the column sums; etc.

Sinkhorn Algorithm

$$\begin{cases} \forall 1\le i\le n,\ \boldsymbol{\beta}_k(j) \leftarrow \log \dfrac{\mathbf{c}(j)}{\sum_{i=1}^m \mathbf{W}_{ij}\exp(\boldsymbol{\alpha}_{k-1}(i))}, \\ \forall 1\le i\le m,\ \boldsymbol{\alpha}_k(i) \leftarrow \log \dfrac{\mathbf{r}(i)}{\sum_{j=1}^n \mathbf{W}_{ij}\exp(\boldsymbol{\beta}_k(j))}. \end{cases}$$

- Quick math for the convergence of Sinkhorn's Algorithm (Dual interpretation of Franklin and Lorenz '89)

Sinkhorn Algorithm

$$\begin{cases} \forall 1 \le i \le n,\ \boldsymbol{\beta}_k(j) \leftarrow \log \frac{\mathbf{c}(j)}{\sum_{i=1}^{m} \mathbf{W}_{ij} \exp(\boldsymbol{\alpha}_{k-1}(i))}, \\ \forall 1 \le i \le m,\ \boldsymbol{\alpha}_k(i) \leftarrow \log \frac{\mathbf{r}(i)}{\sum_{j=1}^{n} \mathbf{W}_{ij} \exp(\boldsymbol{\beta}_k(j))}. \end{cases} \quad = \quad$$

Alternating maximization

$$\begin{cases} \boldsymbol{\beta}_k \leftarrow \arg\max_{\boldsymbol{\beta}} f(\boldsymbol{\alpha}_{k-1}, \boldsymbol{\beta}) \\ \boldsymbol{\alpha}_k \leftarrow \arg\max_{\boldsymbol{\alpha}} f(\boldsymbol{\alpha}, \boldsymbol{\beta}_k) \end{cases}$$

- Quick math for the convergence of Sinkhorn's Algorithm (Dual interpretation of Franklin and Lorenz '89)

Sinkhorn Algorithm

$$\begin{cases} \forall 1 \le i \le n,\ \boldsymbol{\beta}_k(j) \leftarrow \log \frac{\mathbf{c}(j)}{\sum_{i=1}^m \mathbf{W}_{ij} \exp(\boldsymbol{\alpha}_{k-1}(i))}, \\ \forall 1 \le i \le m,\ \boldsymbol{\alpha}_k(i) \leftarrow \log \frac{\mathbf{r}(i)}{\sum_{j=1}^n \mathbf{W}_{ij} \exp(\boldsymbol{\beta}_k(j))}. \end{cases} \quad = \quad \begin{cases} \boldsymbol{\beta}_k \leftarrow \arg\max_{\boldsymbol{\beta}} f(\boldsymbol{\alpha}_{k-1}, \boldsymbol{\beta}) \\ \boldsymbol{\alpha}_k \leftarrow \arg\max_{\boldsymbol{\alpha}} f(\boldsymbol{\alpha}, \boldsymbol{\beta}_k) \end{cases}$$

Alternating maximization

Step 1: ℓ_∞- non-expansion of the potentials

$$\|\boldsymbol{\alpha}_k - \boldsymbol{\alpha}^*\|_\infty = \|\xi(\boldsymbol{\beta}_{k-1}) - \xi(\boldsymbol{\beta}^*)\|_\infty$$

($\xi: \beta \mapsto \alpha$ update map)

- Quick math for the convergence of Sinkhorn's Algorithm (Dual interpretation of Franklin and Lorenz '89)

Sinkhorn Algorithm

$$\begin{cases} \forall 1 \le i \le n,\ \boldsymbol{\beta}_k(j) \leftarrow \log \frac{\mathbf{c}(j)}{\sum_{i=1}^m \mathbf{W}_{ij} \exp(\boldsymbol{\alpha}_{k-1}(i))}, \\ \forall 1 \le i \le m,\ \boldsymbol{\alpha}_k(i) \leftarrow \log \frac{\mathbf{r}(i)}{\sum_{j=1}^n \mathbf{W}_{ij} \exp(\boldsymbol{\beta}_k(j))}. \end{cases} \quad = \quad \begin{cases} \boldsymbol{\beta}_k \leftarrow \arg\max_{\boldsymbol{\beta}} f(\boldsymbol{\alpha}_{k-1}, \boldsymbol{\beta}) \\ \boldsymbol{\alpha}_k \leftarrow \arg\max_{\boldsymbol{\alpha}} f(\boldsymbol{\alpha}, \boldsymbol{\beta}_k) \end{cases}$$

Alternating maximization

Step 1: ℓ_∞- non-expansion of the potentials

$$\|\boldsymbol{\alpha}_k - \boldsymbol{\alpha}^*\|_\infty = \|\xi(\boldsymbol{\beta}_{k-1}) - \xi(\boldsymbol{\beta}^*)\|_\infty \qquad (\ \xi : \beta \mapsto \alpha \text{ update map}\ )$$

$$= \left\| \int_0^1 J_{\xi;\boldsymbol{\beta}}(\boldsymbol{\beta}_{k-1;s})\, ds (\boldsymbol{\beta}_{k-1} - \boldsymbol{\beta}^*) \right\|_\infty$$

- Quick math for the convergence of Sinkhorn's Algorithm (Dual interpretation of Franklin and Lorenz '89)

Sinkhorn Algorithm

$$\begin{cases} \forall 1 \le i \le n,\ \boldsymbol{\beta}_k(j) \leftarrow \log \frac{\mathbf{c}(j)}{\sum_{i=1}^m \mathbf{W}_{ij} \exp(\boldsymbol{\alpha}_{k-1}(i))}, \\ \forall 1 \le i \le m,\ \boldsymbol{\alpha}_k(i) \leftarrow \log \frac{\mathbf{r}(i)}{\sum_{j=1}^n \mathbf{W}_{ij} \exp(\boldsymbol{\beta}_k(j))}. \end{cases} \quad = \quad \begin{cases} \boldsymbol{\beta}_k \leftarrow \arg\max_{\boldsymbol{\beta}} f(\boldsymbol{\alpha}_{k-1}, \boldsymbol{\beta}) \\ \boldsymbol{\alpha}_k \leftarrow \arg\max_{\boldsymbol{\alpha}} f(\boldsymbol{\alpha}, \boldsymbol{\beta}_k) \end{cases}$$

Alternating maximization

Step 1: ℓ_∞- non-expansion of the potentials

$$\begin{aligned} \|\boldsymbol{\alpha}_k - \boldsymbol{\alpha}^*\|_\infty &= \|\xi(\boldsymbol{\beta}_{k-1}) - \xi(\boldsymbol{\beta}^*)\|_\infty \qquad (\ \xi: \beta \mapsto \alpha \text{ update map}\ ) \\ &= \left\| \int_0^1 J_{\xi;\boldsymbol{\beta}}(\boldsymbol{\beta}_{k-1;s})\, ds (\boldsymbol{\beta}_{k-1} - \boldsymbol{\beta}^*) \right\|_\infty \\ &\le \left(\int_0^1 \| \underbrace{J_{\xi;\boldsymbol{\beta}}(\boldsymbol{\beta}_{k-1;s})}_{=\text{row-stochastic}} \|_{\infty\to\infty} ds \right) \|\boldsymbol{\beta}_{k-1} - \boldsymbol{\beta}^*)\|_\infty \end{aligned}$$

- Quick math for the convergence of Sinkhorn's Algorithm (Dual interpretation of Franklin and Lorenz '89)

Sinkhorn Algorithm

$$\begin{cases} \forall 1 \le i \le n,\ \boldsymbol{\beta}_k(j) \leftarrow \log \frac{\mathbf{c}(j)}{\sum_{i=1}^m \mathbf{W}_{ij} \exp(\boldsymbol{\alpha}_{k-1}(i))}, \\ \forall 1 \le i \le m,\ \boldsymbol{\alpha}_k(i) \leftarrow \log \frac{\mathbf{r}(i)}{\sum_{j=1}^n \mathbf{W}_{ij} \exp(\boldsymbol{\beta}_k(j))}. \end{cases} \quad = \quad \begin{cases} \boldsymbol{\beta}_k \leftarrow \arg\max_{\boldsymbol{\beta}} f(\boldsymbol{\alpha}_{k-1}, \boldsymbol{\beta}) \\ \boldsymbol{\alpha}_k \leftarrow \arg\max_{\boldsymbol{\alpha}} f(\boldsymbol{\alpha}, \boldsymbol{\beta}_k) \end{cases}$$

Alternating maximization

Step 1: ℓ_∞- non-expansion of the potentials

$$\begin{aligned} \|\boldsymbol{\alpha}_k - \boldsymbol{\alpha}^*\|_\infty &= \|\xi(\boldsymbol{\beta}_{k-1}) - \xi(\boldsymbol{\beta}^*)\|_\infty \qquad (\ \xi: \beta \mapsto \alpha \text{ update map}\ ) \\ &= \left\| \int_0^1 J_{\xi;\boldsymbol{\beta}}(\boldsymbol{\beta}_{k-1;s})\, ds (\boldsymbol{\beta}_{k-1} - \boldsymbol{\beta}^*) \right\|_\infty \\ &\le \left(\int_0^1 \| \underbrace{J_{\xi;\boldsymbol{\beta}}(\boldsymbol{\beta}_{k-1;s})}_{=\text{row-stochastic}} \|_{\infty\to\infty} ds \right) \|\boldsymbol{\beta}_{k-1} - \boldsymbol{\beta}^*)\|_\infty \\ &\le \|\boldsymbol{\beta}_{k-1} - \boldsymbol{\beta}^*)\|_\infty \end{aligned}$$

- Quick math for the convergence of Sinkhorn's Algorithm (Dual interpretation of Franklin and Lorenz '89)

Sinkhorn Algorithm

$$\begin{cases} \forall 1 \le i \le n,\ \boldsymbol{\beta}_k(j) \leftarrow \log \frac{\mathbf{c}(j)}{\sum_{i=1}^m \mathbf{W}_{ij} \exp(\boldsymbol{\alpha}_{k-1}(i))}, \\ \forall 1 \le i \le m,\ \boldsymbol{\alpha}_k(i) \leftarrow \log \frac{\mathbf{r}(i)}{\sum_{j=1}^n \mathbf{W}_{ij} \exp(\boldsymbol{\beta}_k(j))}. \end{cases} \quad = \quad \begin{cases} \boldsymbol{\beta}_k \leftarrow \arg\max_{\boldsymbol{\beta}} f(\boldsymbol{\alpha}_{k-1}, \boldsymbol{\beta}) \\ \boldsymbol{\alpha}_k \leftarrow \arg\max_{\boldsymbol{\alpha}} f(\boldsymbol{\alpha}, \boldsymbol{\beta}_k) \end{cases}$$

Alternating maximization

Step 1: ℓ_∞- non-expansion of the potentials

$$\begin{aligned} \|\boldsymbol{\alpha}_k - \boldsymbol{\alpha}^*\|_\infty &= \|\xi(\boldsymbol{\beta}_{k-1}) - \xi(\boldsymbol{\beta}^*)\|_\infty \qquad (\ \xi: \beta \mapsto \alpha \text{ update map}\ ) \\ &= \left\| \int_0^1 J_{\xi;\boldsymbol{\beta}}(\boldsymbol{\beta}_{k-1;s})\, ds (\boldsymbol{\beta}_{k-1} - \boldsymbol{\beta}^*) \right\|_\infty \\ &\le \left(\int_0^1 \| \underbrace{J_{\xi;\boldsymbol{\beta}}(\boldsymbol{\beta}_{k-1;s})}_{=\text{row-stochastic}} \|_{\infty\to\infty} ds \right) \|\boldsymbol{\beta}_{k-1} - \boldsymbol{\beta}^*)\|_\infty \\ &\le \|\boldsymbol{\beta}_{k-1} - \boldsymbol{\beta}^*)\|_\infty \\ &\le \cdots \le \|\boldsymbol{\alpha}_0 - \boldsymbol{\alpha}^*\|_\infty \end{aligned}$$

- Quick math for the convergence of Sinkhorn's Algorithm

Step 2: span-contraction of the potentials

$$\|\boldsymbol{\alpha}_k - \boldsymbol{\alpha}^*\|_{sp} = \|\xi(\boldsymbol{\beta}_{k-1}) - \xi(\boldsymbol{\beta}^*)\|_{sp}$$

$$(\ sp(x) \coloneqq \max x\ - \min x\ )$$

- Quick math for the convergence of Sinkhorn's Algorithm

Step 2: span-contraction of the potentials

$$\|\boldsymbol{\alpha}_k - \boldsymbol{\alpha}^*\|_{sp} = \|\xi(\boldsymbol{\beta}_{k-1}) - \xi(\boldsymbol{\beta}^*)\|_{sp} \qquad (\, sp(x) := \max x \, - \min x \,)$$

$$= \left\| \int_0^1 J_{\xi;\boldsymbol{\beta}}(\boldsymbol{\beta}_{k-1;s})\, ds (\boldsymbol{\beta}_{k-1} - \boldsymbol{\beta}^*) \right\|_{sp}$$

$$\leq \Big(\int_0^1 \Big\| \underbrace{J_{\xi;\boldsymbol{\beta}}(\boldsymbol{\beta}_{k-1;s})}_{\substack{=\text{row-stochastic, nonzero} \\ \leq \tanh(\Delta(J_{\xi;\boldsymbol{\beta}}(\boldsymbol{\beta}_{k-1;s}))/2) < 1}} \Big\|_{sp \to sp} ds \Big) \|\boldsymbol{\beta}_{k-1} - \boldsymbol{\beta}^*)\|_{sp}$$

(Birkoff's contraction Thm.)

$$\Delta(P) := \max_{i,i',j,j'} \log \frac{P_{ij}\, P_{i'j'}}{P_{ij'}\, P_{i'j}}$$

- Quick math for the convergence of Sinkhorn's Algorithm

Step 2: span-contraction of the potentials

$$\|\boldsymbol{\alpha}_k - \boldsymbol{\alpha}^*\|_{sp} = \|\xi(\boldsymbol{\beta}_{k-1}) - \xi(\boldsymbol{\beta}^*)\|_{sp} \qquad (\, sp(x) := \max x - \min x \,)$$

$$= \left\| \int_0^1 J_{\xi;\boldsymbol{\beta}}(\boldsymbol{\beta}_{k-1;s})\, ds (\boldsymbol{\beta}_{k-1} - \boldsymbol{\beta}^*) \right\|_{sp}$$

$$\le \Big(\int_0^1 \Big\| \underbrace{J_{\xi;\boldsymbol{\beta}}(\boldsymbol{\beta}_{k-1;s})}_{\substack{=\text{row-stochastic, nonzero} \\ \le \tanh(\Delta(J_{\xi;\boldsymbol{\beta}}(\boldsymbol{\beta}_{k-1;s}))/2) < 1}} \Big\|_{sp \to sp} ds \Big) \|\boldsymbol{\beta}_{k-1} - \boldsymbol{\beta}^*)\|_{sp}$$

(Birkoff's contraction Thm.)

$$\le \rho(\text{range of potentials})\, \|\boldsymbol{\beta}_{k-1} - \boldsymbol{\beta}^*)\|_{sp} \qquad \Delta(P) := \max_{i,i',j,j'} \log \frac{P_{ij}\, P_{i'j'}}{P_{ij'}\, P_{i'j}}$$

- Quick math for the convergence of Sinkhorn's Algorithm

Step 2: span-contraction of the potentials

$$\|\boldsymbol{\alpha}_k - \boldsymbol{\alpha}^*\|_{sp} = \|\xi(\boldsymbol{\beta}_{k-1}) - \xi(\boldsymbol{\beta}^*)\|_{sp} \qquad (\, sp(x) := \max x - \min x \,)$$

$$= \left\| \int_0^1 J_{\xi;\boldsymbol{\beta}}(\boldsymbol{\beta}_{k-1;s})\, ds (\boldsymbol{\beta}_{k-1} - \boldsymbol{\beta}^*) \right\|_{sp}$$

$$\le \Big(\int_0^1 \Big\| \underbrace{J_{\xi;\boldsymbol{\beta}}(\boldsymbol{\beta}_{k-1;s})}_{\substack{=\text{row-stochastic, nonzero} \\ \le \tanh(\Delta(J_{\xi;\boldsymbol{\beta}}(\boldsymbol{\beta}_{k-1;s}))/2) < 1}} \Big\|_{sp \to sp} ds \Big) \|\boldsymbol{\beta}_{k-1} - \boldsymbol{\beta}^*)\|_{sp} \qquad \text{(Birkoff's contraction Thm.)}$$

$$\le \rho(\text{range of potentials})\, \|\boldsymbol{\beta}_{k-1} - \boldsymbol{\beta}^*)\|_{sp} \qquad \Delta(P) := \max_{i,i',j,j'} \log \frac{P_{ij} P_{i'j'}}{P_{ij'} P_{i'j}}$$

$$\le \cdots \le \rho(\text{range of potentials})^k \, \|\boldsymbol{\alpha}_0 - \boldsymbol{\alpha}^*\|_{sp}$$

- Monge-Kantorovich Optimal Transport

  Given a margin ($\mathbf{r}$, $\mathbf{c}$) and cost matrix $\mathbf{C} = (c_{ij})$, c_{ij}=cost of moving a unit mass from location i to j, find the most efficient coupling $\boldsymbol{\pi}^*$:

$$\pi^* = \underset{\pi \in \Pi(\mathbf{r},\mathbf{c})}{\arg\min} \langle \pi, \mathbf{C} \rangle$$

- Monge-Kantorovich Optimal Transport

  Given a margin $(\mathbf{r}, \mathbf{c})$ and cost matrix $\mathbf{C} = (c_{ij})$, c_{ij}=cost of moving a unit mass from location i to j, find the most efficient coupling $\boldsymbol{\pi}^*$:

$$\pi^* = \underset{\pi \in \Pi(\mathbf{r},\mathbf{c})}{\arg\min} \langle \pi, \mathbf{C} \rangle$$

- Entropic Optimal Transport

  - Entropic Regularization (Cuturi, NeurIPS '13)

$$\pi^* = \underset{\pi \in \Pi(\mathbf{r},\mathbf{c})}{\arg\min} \langle \pi, \mathbf{C} \rangle + \varepsilon \, \mathrm{KL}(\pi \| \mathbf{r} \otimes \mathbf{c})$$

- Monge-Kantorovich Optimal Transport

  Given a margin $(\mathbf{r}, \mathbf{c})$ and cost matrix $\mathbf{C} = (c_{ij})$, c_{ij}=cost of moving a unit mass from location i to j, find the most efficient coupling $\boldsymbol{\pi}^*$:

$$\pi^* = \underset{\pi \in \Pi(\mathbf{r},\mathbf{c})}{\arg\min} \langle \pi, \mathbf{C} \rangle$$

- Entropic Optimal Transport

  • Entropic Regularization (Cuturi, NeurIPS '13)

$$\pi^* = \underset{\pi \in \Pi(\mathbf{r},\mathbf{c})}{\arg\min} \langle \pi, \mathbf{C} \rangle + \varepsilon \, \mathrm{KL}(\pi \| \mathbf{r} \otimes \mathbf{c})$$

$$= \underset{\pi \in \Pi(\mathbf{r},\mathbf{c})}{\arg\min} \left[D_{KL}(\pi \,\|\, \mathbf{W}) \text{ with } \mathbf{W} = \exp(-\mathbf{C}/\varepsilon) \odot (\mathbf{r} \otimes \mathbf{c}) \right]$$

(Matrix Scaling Problem, so solve it by Sinkhorn!)

- Wasserstein distances

  μ, ν = Probability measures on spaces X and Y

  κ: $X \times Y \to [0, \infty]$ cost function

$$W_\kappa(\mu, \nu) := \min_{\pi \in \Pi(\mu,\nu)} \left[\langle \pi, \kappa \rangle = \int \kappa(x, y)\, d\pi(x, y) \right]$$

- Wasserstein distances

  μ, ν = Probability measures on spaces X and Y

  κ: $X \times Y \to [0, \infty]$ cost function

$$W_\kappa(\mu, \nu) := \min_{\pi \in \Pi(\mu,\nu)} \left[\langle \pi, \kappa \rangle = \int \kappa(x, y)\, d\pi(x, y) \right]$$

Cons: Nonsmooth and high sample complexity ($n^{-1/\dim}$), LP computation

→ Not suitable for ML uses

- Wasserstein distances

  μ, ν = Probability measures on spaces X and Y

  κ: $X \times Y \to [0, \infty]$ cost function

$$W_\kappa(\mu,\nu) := \min_{\pi \in \Pi(\mu,\nu)} \left[\langle \pi, \kappa \rangle = \int \kappa(x,y)\, d\pi(x,y) \right]$$

Cons: Nonsmooth and high sample complexity ($n^{-1/\dim}$), LP computation

→ Not suitable for ML uses

- Sinkhorn Divergence (Cuturi' 13)

$$S_\varepsilon(\mu,\nu) = \mathrm{OT}_\varepsilon(\mu,\nu) - \tfrac{1}{2}\mathrm{OT}_\varepsilon(\mu,\mu) - \tfrac{1}{2}\mathrm{OT}_\varepsilon(\nu,\nu).$$

$$\mathrm{OT}_\kappa(\mu,\nu) := \min_{\pi \in \Pi(\mu,\nu)} \left[\langle \pi, \kappa \rangle + \varepsilon D_{KL}(\pi, \mu \otimes \nu) \right]$$ (Computed by Empirical Sinkhorn)

- Wasserstein distances

  μ, ν = Probability measures on spaces X and Y

  $\kappa: X \times Y \to [0, \infty]$ cost function

$$W_\kappa(\mu,\nu) := \min_{\pi \in \Pi(\mu,\nu)} \left[\langle \pi, \kappa \rangle = \int \kappa(x,y)\, d\pi(x,y) \right]$$

Cons: Nonsmooth and high sample complexity ($n^{-1/\dim}$), LP computation

→ Not suitable for ML uses

- Sinkhorn Divergence (Cuturi' 13)

$$S_\varepsilon(\mu,\nu) = \mathrm{OT}_\varepsilon(\mu,\nu) - \tfrac{1}{2}\mathrm{OT}_\varepsilon(\mu,\mu) - \tfrac{1}{2}\mathrm{OT}_\varepsilon(\nu,\nu).$$

$$\mathrm{OT}_\kappa(\mu,\nu) := \min_{\pi \in \Pi(\mu,\nu)} \left[\langle \pi, \kappa \rangle + \varepsilon D_{KL}(\pi, \mu \otimes \nu) \right]$$ (Computed by Empirical Sinkhorn)

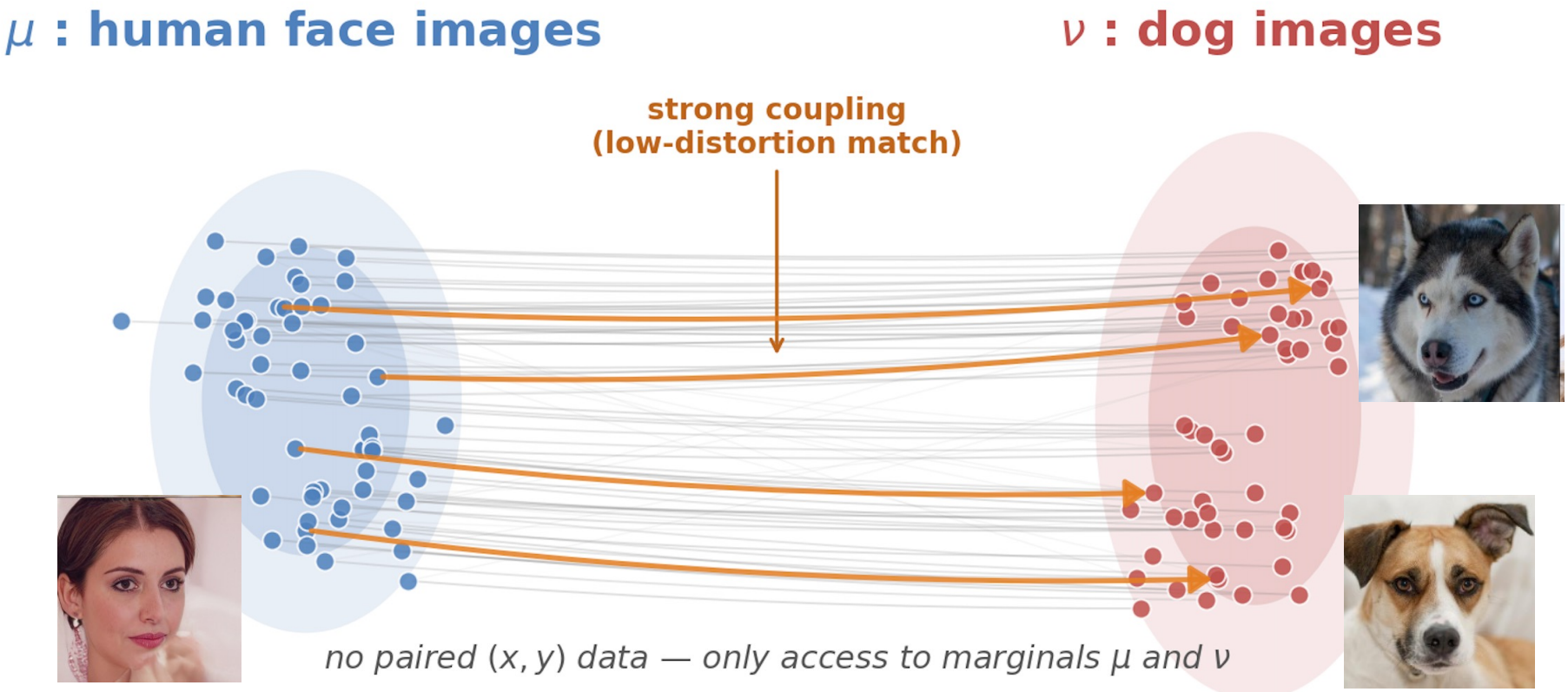
Pros: smooth, low sample complexity ($n^{-1/2}$), easily computed, $W_\kappa(\mu,\nu) = \lim_{\varepsilon \to 0} \mathrm{S}_\kappa(\mu,\nu)$

- Cuturi, *Sinkhorn Distances: Lightspeed Computation of OT*, NeurIPS 2013.
- Genevay, Peyré, Cuturi, *Learning Generative Models with Sinkhorn Divergences*, AISTATS 2018.
- Feydy, Séjourné, Vialard, Amari, Trouvé, Peyré, *Interpolating between OT and MMD using Sinkhorn divergences*, AISTATS 2019.
- Genevay, Chizat, Bach, Cuturi, Peyré, *Sample Complexity of Sinkhorn divergences*, AISTATS 2019.
- Peyré & Cuturi, *Computational Optimal Transport*, FnT in ML, 2019 (the standard reference book).

ML Applications on Generative modeling

Image translation
Knowledge distillation

- Generative modeling: Unpaired Image Translation



- Generative modeling: Unpaired Image Translation

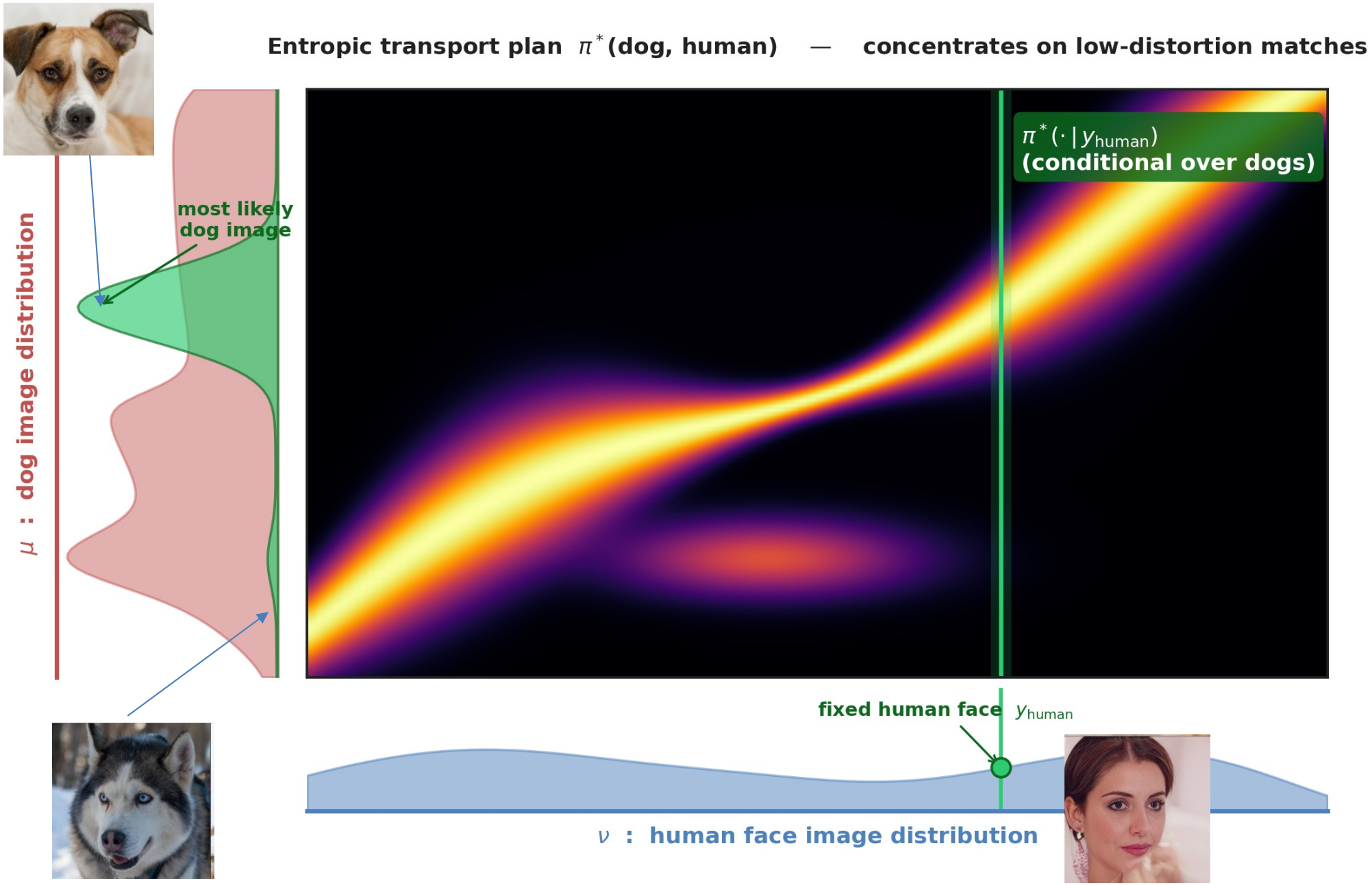
Results from REU 2025 @ UW-Madison on generative modeling

Dog ↔ Human image translation

- Generative modeling: Unpaired Image Translation

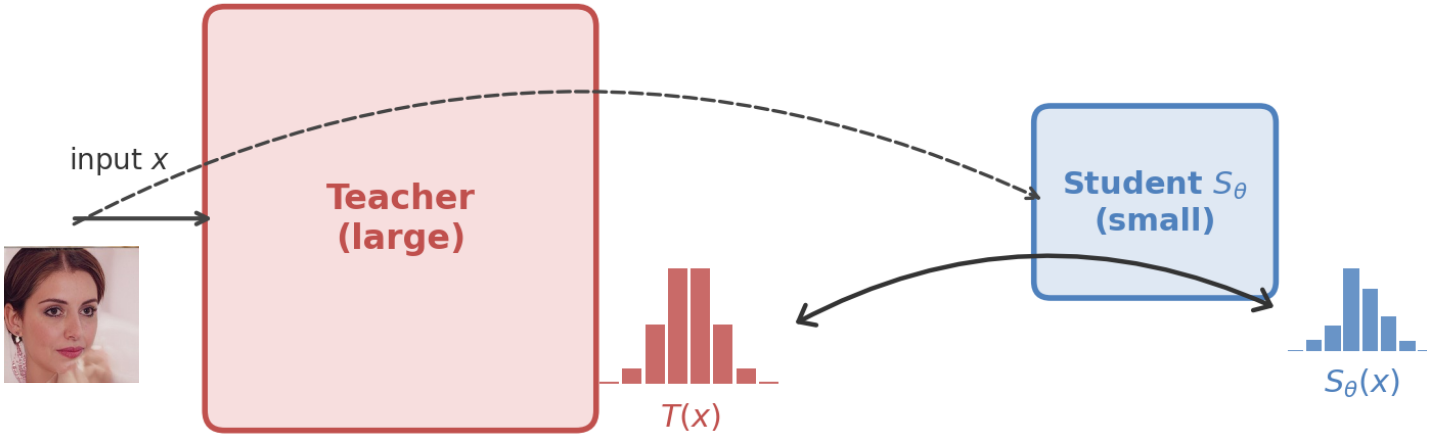
Results from REU 2025 @ UW-Madison on generative modeling

Adult ↔ Children image translation


Entropic transport plan π^*(dog, human) — concentrates on low-distortion matches
μ : dog image distribution
most likely dog image
$\pi^*(\cdot \mid y_{\text{human}})$ (conditional over dogs)
fixed human face y_{human}
ν : human face image distribution

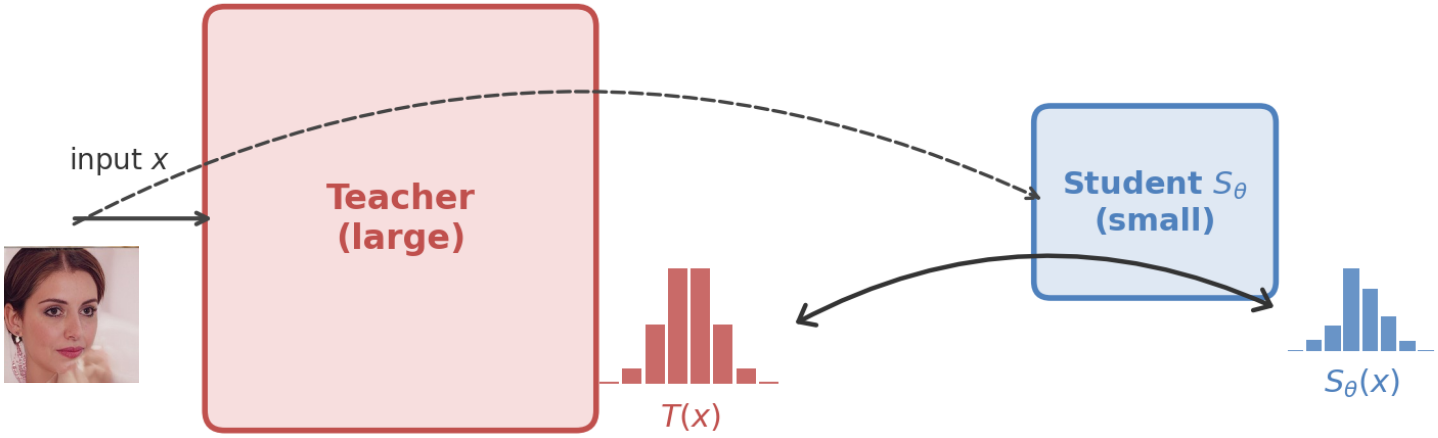
Generative modeling: Knowledge Distillation

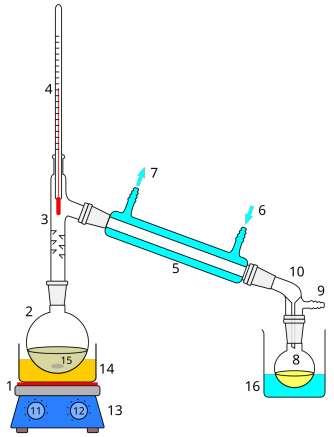
- Compress a large teacher T into a small student S_θ while preserving behavior



Generative modeling: Knowledge Distillation

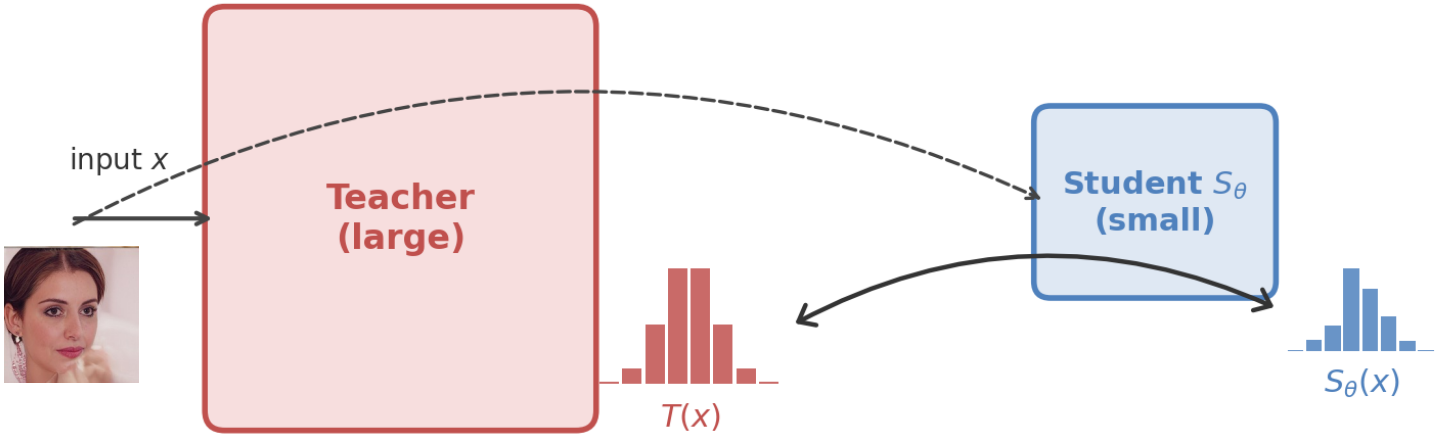
- Compress a large teacher T into a small student S_θ while preserving behavior

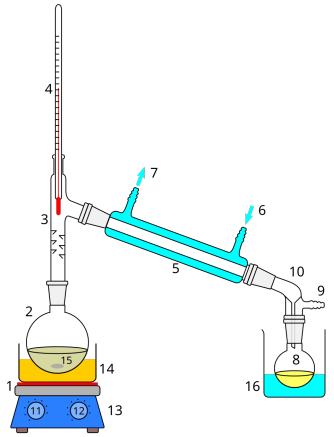




Generative modeling: Knowledge Distillation

- Compress a large teacher T into a small student S_θ while preserving behavior





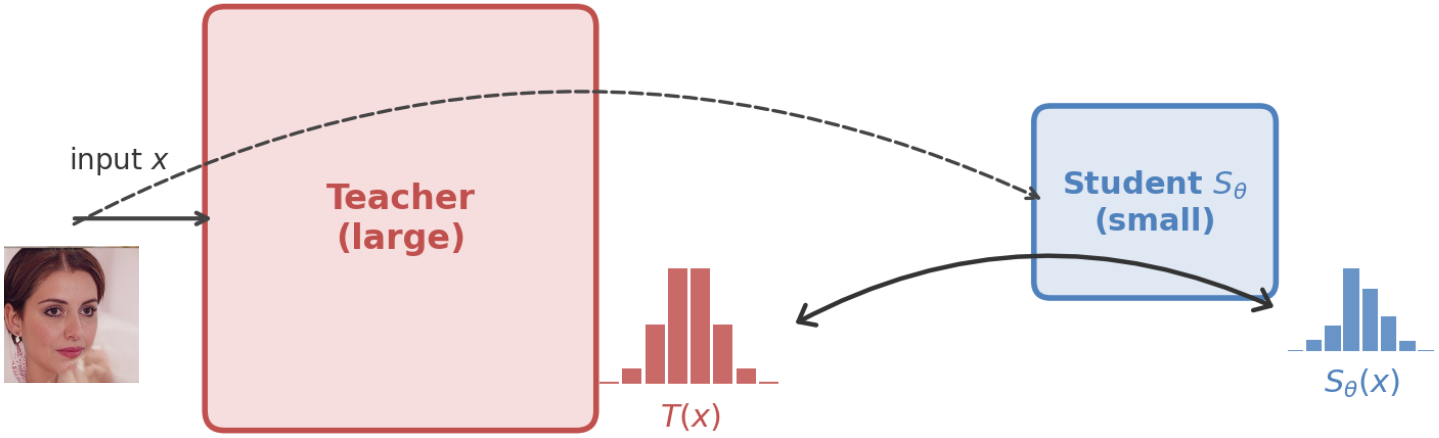
Distillation is being used to make capable small LLM from large models

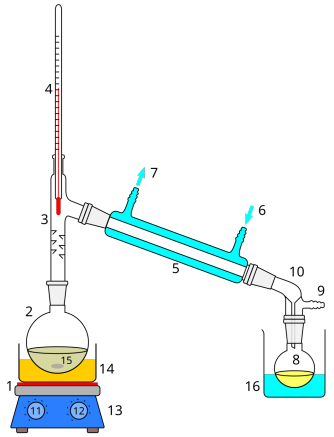
Refs: Cui–Wu–Chen '22 (Wasserstein KD); Park–Kim–Lu '19 (Relational KD); Bhardwaj et al. '22 (OT for LLM distillation)

- **Llama 3.1 → Llama 3.2** — the 1B and 3B were distilled from the 8B/70B.

Generative modeling: Knowledge Distillation

- Compress a large teacher T into a small student S_θ while preserving behavior





Distillation is being used to make capable small LLM from large models

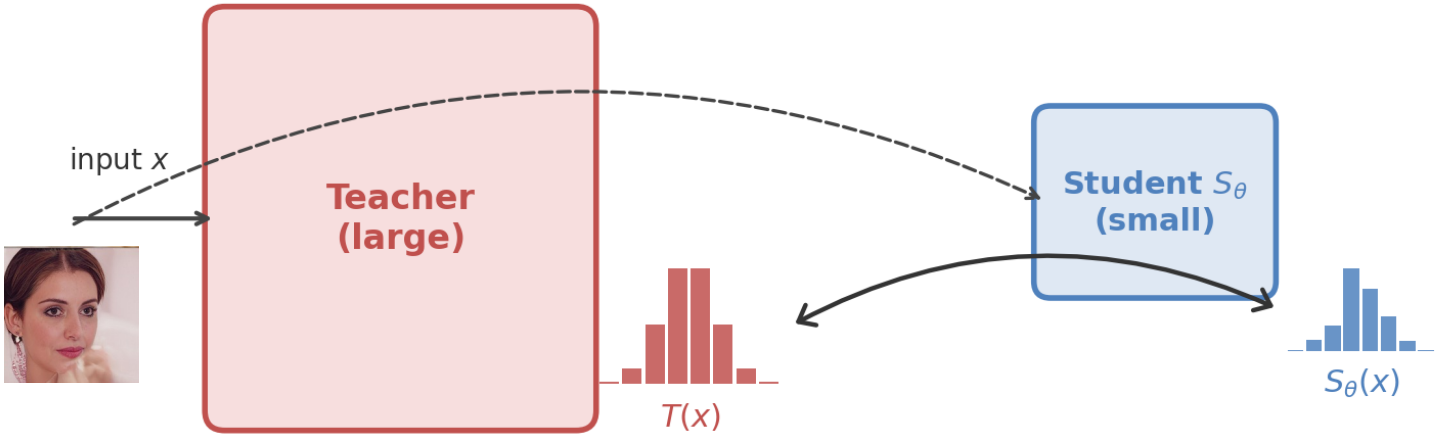
Refs: Cui–Wu–Chen '22 (Wasserstein KD); Park–Kim–Lu '19 (Relational KD); Bhardwaj et al. '22 (OT for LLM distillation)

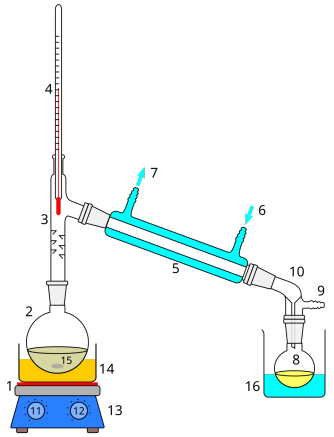
•**Llama 3.1 → Llama 3.2** — the 1B and 3B were distilled from the 8B/70B.

- Classical (Hinton '15): Find θ by minimizing $\mathrm{D}_{KL}(\, S_\theta(x) \,\|\, T(x)\,)$

Generative modeling: Knowledge Distillation

- Compress a large teacher T into a small student S_θ while preserving behavior





Distillation is being used to make capable small LLM from large models

Refs: Cui–Wu–Chen '22 (Wasserstein KD); Park–Kim–Lu '19 (Relational KD); Bhardwaj et al. '22 (OT for LLM distillation)

•**Llama 3.1 → Llama 3.2** — the 1B and 3B were distilled from the 8B/70B.

- Classical (Hinton '15): Find θ by minimizing $\mathrm{D}_{KL}(\, S_\theta(x) \,\|\, T(x)\,)$

EOT-based KD: Sinkhorn divergence handles disjoint supports; KL blows up

$$\mathcal{L}_{\mathrm{KD}}(\theta) \;=\; \mathrm{Sinkhorn}_\varepsilon\big(S_\theta(x),\, T(x);\, C\big)$$

New Perspective I: Random Coordinate Sinkhorn

- Motivating Question: EOT on large graph G=(V,E)

  $G = (V, E)$ a large graph, μ, ν two probability distributions on the node set V.

  Find the most efficient coupling $\boldsymbol{\pi}^*$ in terms of the shortest-path distance on G

$$\pi^* = \underset{\pi \in \Pi(\mu,\nu)}{\arg\min} \langle \pi, d_G \rangle + \varepsilon D_{\mathrm{KL}}(\pi \| \mu \otimes \nu)$$

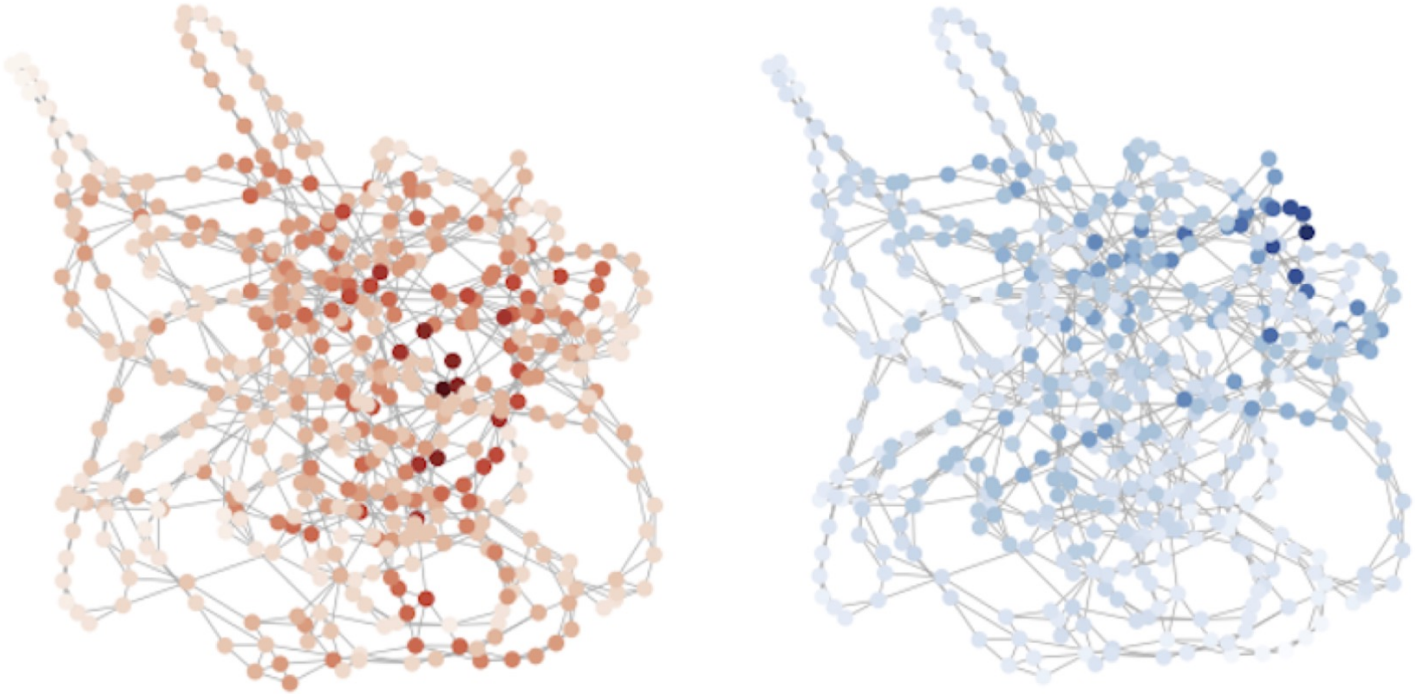


FIGURE 1. A 500-node graph G and two probability distributions μ and ν on the node set shown. Color shade indicates mass at each node.

- Motivating Question: EOT on large graph G=(V,E)

$G = (V, E)$ a large graph, μ, ν two probability distributions on the node set V.

Find the most efficient coupling $\boldsymbol{\pi}^*$ in terms of the shortest-path distance on G

$$\pi^* = \underset{\pi \in \Pi(\mu,\nu)}{\arg\min} \langle \pi, d_G \rangle + \varepsilon D_{\mathrm{KL}}(\pi \| \mu \otimes \nu)$$

Shortest-path distance matrix

For large $G = (V, E)$, even evaluating the objective is challenging

Computational Cost: The cost matrix d_G requires an All-Pairs Shortest Path (APSP) computation = $O(n^3)$ or $O(n^2 \log n + n|E|)$

Memory Cost: The matrix d_G is dense, requiring $O(n^2)$ storage, which is prohibitive for graphs with millions of nodes.

- Open Question: EOT on large graph G=(V,E)

$$(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) \in \underset{\boldsymbol{\alpha}, \boldsymbol{\beta}: V \to \mathbb{R}}{\operatorname{argmax}} f_G(\boldsymbol{\alpha}, \boldsymbol{\beta}) := \langle \boldsymbol{\alpha}, \mu \rangle + \langle \boldsymbol{\beta}, \nu \rangle - \langle \exp(\boldsymbol{\alpha} \oplus \boldsymbol{\beta}), K_G \odot (\mu \otimes \nu) \rangle$$

$$K_G(x, y) := \exp(-d_G(x, y)/\varepsilon)$$

- Open Question: EOT on large graph G=(V,E)

$$(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) \in \underset{\boldsymbol{\alpha},\boldsymbol{\beta}:V\to\mathbb{R}}{\arg\max}\ f_G(\boldsymbol{\alpha},\boldsymbol{\beta}) := \langle \boldsymbol{\alpha}, \mu\rangle + \langle \boldsymbol{\beta}, \nu\rangle - \langle \exp(\boldsymbol{\alpha}\oplus\boldsymbol{\beta}), K_G \odot (\mu\otimes\nu)\rangle$$

$$K_G(x,y) := \exp(-d_G(x,y)/\varepsilon)$$

Main Idea: Solve an easier stochastic surrogate

$$(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) \in \underset{\boldsymbol{\alpha},\boldsymbol{\beta}:V\to\mathbb{R}}{\arg\max} \left(\hat{f}_G(\boldsymbol{\alpha},\boldsymbol{\beta}) := \mathbb{E}_{H\sim\pi}\left[\hat{f}_H(\boldsymbol{\alpha},\boldsymbol{\beta})\right]\right)$$

$$= \langle \boldsymbol{\alpha}, \mu\rangle + \langle \boldsymbol{\beta}, \nu\rangle - \langle \exp(\boldsymbol{\alpha}\oplus\boldsymbol{\beta}), \hat{K} \odot (\mu\otimes\nu)\rangle$$

where $\hat{f}_H$ = local dual objective on a random subgraph H:

- Open Question: EOT on large graph G=(V,E)

$$(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) \in \operatorname*{argmax}_{\boldsymbol{\alpha},\boldsymbol{\beta}:V\to\mathbb{R}} f_G(\boldsymbol{\alpha},\boldsymbol{\beta}) := \langle \boldsymbol{\alpha}, \mu\rangle + \langle \boldsymbol{\beta}, \nu\rangle - \langle \exp(\boldsymbol{\alpha}\oplus\boldsymbol{\beta}), K_G \odot (\mu\otimes\nu)\rangle$$

$$K_G(x,y) := \exp(-d_G(x,y)/\varepsilon)$$

Main Idea: Solve an easier stochastic surrogate

$$(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) \in \operatorname*{argmax}_{\boldsymbol{\alpha},\boldsymbol{\beta}:V\to\mathbb{R}} \left(\hat{f}_G(\boldsymbol{\alpha},\boldsymbol{\beta}) := \mathbb{E}_{H\sim\pi}\left[\hat{f}_H(\boldsymbol{\alpha},\boldsymbol{\beta})\right]\right)$$

$$= \langle \boldsymbol{\alpha}, \mu\rangle + \langle \boldsymbol{\beta}, \nu\rangle - \langle \exp(\boldsymbol{\alpha}\oplus\boldsymbol{\beta}), \hat{K} \odot (\mu\otimes\nu)\rangle$$

where $\hat{f}_H$ = local dual objective on a random subgraph H:

$$\hat{f}_H(\boldsymbol{\alpha},\boldsymbol{\beta}) = \sum_{x\in V(H)} \boldsymbol{\alpha}(x)\frac{\mu(x)}{p_x} + \sum_{y\in V(H)} \boldsymbol{\beta}(y)\frac{\nu(y)}{p_y} - \sum_{x,y\in V(H)} e^{\boldsymbol{\alpha}(x)+\boldsymbol{\beta}(y)} K_H(x,y)\frac{\mu(x)\nu(y)}{p_{xy}}$$

$\hat{K}$ = Expected surrogate Gibbs kernel

$$\hat{K}(x,y) = \mathbb{E}_H\left[K_H(x,y) \,|\, x,y\in V(H)\right]$$

$$K_H(x,y) = \exp(-d_H(x,y)/\varepsilon) \qquad \text{for } x,y\in V(H)$$

Local distance matrix on H

- Open Question: EOT on large graph G=(V,E)

$$(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) \in \underset{\boldsymbol{\alpha},\boldsymbol{\beta}:V\to\mathbb{R}}{\arg\max}\ f_G(\boldsymbol{\alpha},\boldsymbol{\beta}) := \langle \boldsymbol{\alpha}, \mu\rangle + \langle \boldsymbol{\beta}, \nu\rangle - \langle \exp(\boldsymbol{\alpha}\oplus\boldsymbol{\beta}), K_G \odot (\mu\otimes\nu)\rangle$$

$$K_G(x,y) := \exp(-d_G(x,y)/\varepsilon)$$

Main Idea: Solve an easier stochastic surrogate

$$(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) \in \underset{\boldsymbol{\alpha},\boldsymbol{\beta}:V\to\mathbb{R}}{\arg\max}\left(\hat{f}_G(\boldsymbol{\alpha},\boldsymbol{\beta}) := \mathbb{E}_{H\sim\pi}\left[\hat{f}_H(\boldsymbol{\alpha},\boldsymbol{\beta})\right]\right)$$

$$= \langle \boldsymbol{\alpha}, \mu\rangle + \langle \boldsymbol{\beta}, \nu\rangle - \langle \exp(\boldsymbol{\alpha}\oplus\boldsymbol{\beta}), \hat{K} \odot (\mu\otimes\nu)\rangle$$

Q1: How do we control the stochastic relaxation bias $f_G - \hat{f}_G$?

- Open Question: EOT on large graph G=(V,E)

$$(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) \in \underset{\boldsymbol{\alpha},\boldsymbol{\beta}:V\to\mathbb{R}}{\operatorname{argmax}} f_G(\boldsymbol{\alpha},\boldsymbol{\beta}) := \langle \boldsymbol{\alpha}, \mu\rangle + \langle \boldsymbol{\beta}, \nu\rangle - \langle \exp(\boldsymbol{\alpha}\oplus\boldsymbol{\beta}), K_G \odot (\mu\otimes\nu)\rangle$$

$$K_G(x,y) := \exp(-d_G(x,y)/\varepsilon)$$

Main Idea: Solve an easier stochastic surrogate

$$(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) \in \underset{\boldsymbol{\alpha},\boldsymbol{\beta}:V\to\mathbb{R}}{\operatorname{argmax}} \left(\hat{f}_G(\boldsymbol{\alpha},\boldsymbol{\beta}) := \mathbb{E}_{H\sim\pi}\left[\hat{f}_H(\boldsymbol{\alpha},\boldsymbol{\beta})\right]\right)$$

$$= \langle \boldsymbol{\alpha}, \mu\rangle + \langle \boldsymbol{\beta}, \nu\rangle - \langle \exp(\boldsymbol{\alpha}\oplus\boldsymbol{\beta}), \hat{K} \odot (\mu\otimes\nu)\rangle$$

Q1: How do we control the stochastic relaxation bias $f_G - \hat{f}_G$?

Q2: How do we compute the stochastic surrogate potentials $(\hat{\alpha}, \hat{\beta})$?

- **Q1: How do we control the stochastic relaxation bias $f_G - \hat{f}_G$?**

$$(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) \in \operatorname*{argmax}_{\boldsymbol{\alpha},\boldsymbol{\beta}:V\to\mathbb{R}} f_G(\boldsymbol{\alpha},\boldsymbol{\beta}) := \langle \boldsymbol{\alpha}, \mu\rangle + \langle \boldsymbol{\beta}, \nu\rangle - \langle \exp(\boldsymbol{\alpha}\oplus\boldsymbol{\beta}), K_G \odot (\mu\otimes\nu)\rangle$$

$$(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) \in \operatorname*{argmax}_{\boldsymbol{\alpha},\boldsymbol{\beta}:V\to\mathbb{R}} \left(\hat{f}_G(\boldsymbol{\alpha},\boldsymbol{\beta}) = \langle \boldsymbol{\alpha}, \mu\rangle + \langle \boldsymbol{\beta}, \nu\rangle - \langle \exp(\boldsymbol{\alpha}\oplus\boldsymbol{\beta}), \hat{K} \odot (\mu\otimes\nu)\rangle\right.$$

- **Q1: How do we control the stochastic relaxation bias $f_G - \hat{f}_G$?**

$$(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) \in \operatorname*{argmax}_{\boldsymbol{\alpha},\boldsymbol{\beta}:V\to\mathbb{R}} f_G(\boldsymbol{\alpha},\boldsymbol{\beta}) := \langle \boldsymbol{\alpha}, \mu\rangle + \langle \boldsymbol{\beta}, \nu\rangle - \langle \exp(\boldsymbol{\alpha}\oplus\boldsymbol{\beta}), K_G \odot (\mu\otimes\nu)\rangle$$

$$(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) \in \operatorname*{argmax}_{\boldsymbol{\alpha},\boldsymbol{\beta}:V\to\mathbb{R}} \left(\hat{f}_G(\boldsymbol{\alpha},\boldsymbol{\beta}) = \langle \boldsymbol{\alpha}, \mu\rangle + \langle \boldsymbol{\beta}, \nu\rangle - \langle \exp(\boldsymbol{\alpha}\oplus\boldsymbol{\beta}), \hat{K} \odot (\mu\otimes\nu)\rangle\right.$$

Thm. (Bi, Kim, L. 26+) Let $\pi^* = e^{\alpha\oplus\beta} \odot (\mu\otimes\nu)$, $\hat{\pi}^* = e^{\hat{\alpha}\oplus\hat{\beta}} \odot (\mu\otimes\nu)$. Then

$$d_{\mathrm{Hell}}(\pi^*, \hat{\pi}^*) \le 2e^{\kappa_{max}} d_{\mathrm{Hell};\mu\otimes\nu}(K_G, \hat{K})$$

where $\kappa_{\max} = \max_{x,y} K_G(x,y) \vee \hat{K}(x,y)$, $\qquad d_{Hell;\mu\otimes\nu}(K,\tilde{K}) = \left(\sum_{ij}(\sqrt{K_{ij}} - \sqrt{\tilde{K}_{ij}})^2 \mu(i)\nu(j)\right)^{1/2}$

- **Q1: How do we control the stochastic relaxation bias $f_G - \hat{f}_G$?**

$$(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) \in \operatorname*{argmax}_{\boldsymbol{\alpha},\boldsymbol{\beta}:V\to\mathbb{R}} f_G(\boldsymbol{\alpha},\boldsymbol{\beta}) := \langle \boldsymbol{\alpha}, \mu\rangle + \langle \boldsymbol{\beta}, \nu\rangle - \langle \exp(\boldsymbol{\alpha}\oplus\boldsymbol{\beta}), K_G \odot (\mu\otimes\nu)\rangle$$

$$(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) \in \operatorname*{argmax}_{\boldsymbol{\alpha},\boldsymbol{\beta}:V\to\mathbb{R}} \left(\hat{f}_G(\boldsymbol{\alpha},\boldsymbol{\beta}) = \langle \boldsymbol{\alpha}, \mu\rangle + \langle \boldsymbol{\beta}, \nu\rangle - \langle \exp(\boldsymbol{\alpha}\oplus\boldsymbol{\beta}), \hat{K} \odot (\mu\otimes\nu)\rangle\right.$$

Thm. (Bi, Kim, L. 26+) Let $\pi^* = e^{\alpha\oplus\beta} \odot (\mu\otimes\nu)$, $\hat{\pi}^* = e^{\hat{\alpha}\oplus\hat{\beta}} \odot (\mu\otimes\nu)$. Then

$$d_{\mathrm{Hell}}(\pi^*, \hat{\pi}^*) \le 2e^{\kappa_{max}} d_{\mathrm{Hell};\mu\otimes\nu}(K_G, \hat{K})$$

where $\kappa_{\max} = \max_{x,y} K_G(x,y) \vee \hat{K}(x,y)$, $\quad d_{Hell;\mu\otimes\nu}(K,\tilde{K}) = \left(\sum_{ij}(\sqrt{K_{ij}} - \sqrt{\tilde{K}_{ij}})^2 \mu(i)\nu(j)\right)^{1/2}$

Follows directly from a stability bound on matrix scaling w.r.t. input matrix due to William Powell and Danny Duan (26+)

$d_{\mathrm{Hell};\mu\otimes\nu}(K_G, \hat{K})$ depends on G and the random subgraph H

e.g. H=k-ball neighborhood of a random node

- **Q2: How do we compute the stochastic surrogate potentials $(\hat{\alpha}, \hat{\beta})$?**

$$(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) \in \underset{\boldsymbol{\alpha}, \boldsymbol{\beta}: V \to \mathbb{R}}{\arg\max} \left(\hat{f}_G(\boldsymbol{\alpha}, \boldsymbol{\beta}) := \mathbb{E}_{H \sim \pi} \left[\hat{f}_H(\boldsymbol{\alpha}, \boldsymbol{\beta}) \right] \right)$$

$$= \langle \boldsymbol{\alpha}, \mu \rangle + \langle \boldsymbol{\beta}, \nu \rangle - \langle \exp(\boldsymbol{\alpha} \oplus \boldsymbol{\beta}), \hat{K} \odot (\mu \otimes \nu) \rangle$$

Alg. (Full Sinkhorn)

$$\begin{cases} \boldsymbol{\alpha}_k \leftarrow \arg\max_{\boldsymbol{\alpha}} f(\boldsymbol{\alpha}, \boldsymbol{\beta}_{k-1}) \\ \boldsymbol{\beta}_k \leftarrow \arg\max_{\boldsymbol{\beta}} f(\boldsymbol{\alpha}_k, \boldsymbol{\beta}) \end{cases}$$

- **Q2: How do we compute the stochastic surrogate potentials $(\hat{\alpha}, \hat{\beta})$?**

$$(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) \in \underset{\boldsymbol{\alpha}, \boldsymbol{\beta}: V \to \mathbb{R}}{\arg\max} \left(\hat{f}_G(\boldsymbol{\alpha}, \boldsymbol{\beta}) := \mathbb{E}_{H \sim \pi} \left[\hat{f}_H(\boldsymbol{\alpha}, \boldsymbol{\beta}) \right] \right)$$

$$= \langle \boldsymbol{\alpha}, \mu \rangle + \langle \boldsymbol{\beta}, \nu \rangle - \langle \exp(\boldsymbol{\alpha} \oplus \boldsymbol{\beta}), \hat{K} \odot (\mu \otimes \nu) \rangle$$

Alg. (Random Coordinate Sinkhorn)

$$\begin{cases} \boldsymbol{\alpha}_k \leftarrow \boldsymbol{\alpha}_{k;N_k} \approx \arg\max_{\boldsymbol{\alpha}} \hat{f}_G(\boldsymbol{\alpha}, \boldsymbol{\beta}_{k-1}) \\ \quad \text{for } i = 1, \dots, N_k: \\ \quad\quad \text{Sample } H_i \sim \pi \\ \quad\quad \boldsymbol{\alpha}_{k;i+\frac{1}{2}} \leftarrow \arg\max_{\boldsymbol{\alpha}} \hat{f}_{H_i}(\boldsymbol{\alpha}, \boldsymbol{\beta}_{k-1}) \\ \quad\quad \boldsymbol{\alpha}_{k;i} \leftarrow (1-\eta)\boldsymbol{\alpha}_{k;i-1} + \eta \boldsymbol{\alpha}_{k;i+\frac{1}{2}} \\ \boldsymbol{\beta}_k \leftarrow \boldsymbol{\beta}_{k;N_k} \approx \arg\max_{\boldsymbol{\beta}} \hat{f}_G(\boldsymbol{\alpha}_k, \boldsymbol{\beta}) \\ \quad \text{Similarly} \end{cases}$$

Alg. (Full Sinkhorn)

$$\begin{cases} \boldsymbol{\alpha}_k \leftarrow \arg\max_{\boldsymbol{\alpha}} f(\boldsymbol{\alpha}, \boldsymbol{\beta}_{k-1}) \\ \boldsymbol{\beta}_k \leftarrow \arg\max_{\boldsymbol{\beta}} f(\boldsymbol{\alpha}_k, \boldsymbol{\beta}) \end{cases}$$

- **Q2: How do we compute the stochastic surrogate potentials $(\hat{\alpha}, \hat{\beta})$?**

$$(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) \in \underset{\boldsymbol{\alpha}, \boldsymbol{\beta}: V \to \mathbb{R}}{\arg\max} \left(\hat{f}_G(\boldsymbol{\alpha}, \boldsymbol{\beta}) := \mathbb{E}_{H \sim \pi} \left[\hat{f}_H(\boldsymbol{\alpha}, \boldsymbol{\beta}) \right] \right)$$

$$= \langle \boldsymbol{\alpha}, \mu \rangle + \langle \boldsymbol{\beta}, \nu \rangle - \langle \exp(\boldsymbol{\alpha} \oplus \boldsymbol{\beta}), \hat{K} \odot (\mu \otimes \nu) \rangle$$

Alg. (Random Coordinate Sinkhorn)

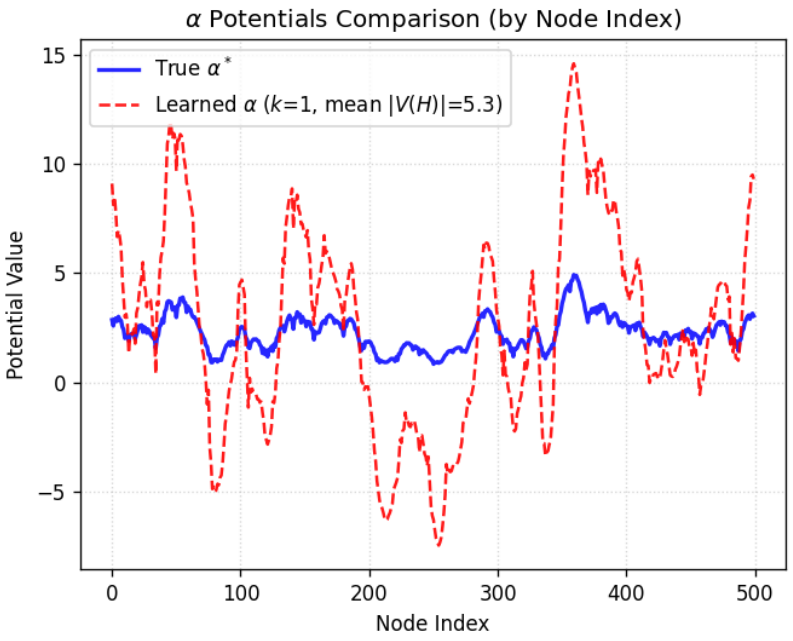
$$\begin{cases} \boldsymbol{\alpha}_k \leftarrow \boldsymbol{\alpha}_{k;N_k} \approx \arg\max_{\boldsymbol{\alpha}} \hat{f}_G(\boldsymbol{\alpha}, \boldsymbol{\beta}_{k-1}) \\ \quad \text{for } i = 1, \dots, N_k: \\ \quad\quad \text{Sample } H_i \sim \pi \\ \quad\quad \boldsymbol{\alpha}_{k;i+\frac{1}{2}} \leftarrow \arg\max_{\boldsymbol{\alpha}} \hat{f}_{H_i}(\boldsymbol{\alpha}, \boldsymbol{\beta}_{k-1}) \\ \quad\quad \boldsymbol{\alpha}_{k;i} \leftarrow (1-\eta)\boldsymbol{\alpha}_{k;i-1} + \eta \boldsymbol{\alpha}_{k;i+\frac{1}{2}} \\ \boldsymbol{\beta}_k \leftarrow \boldsymbol{\beta}_{k;N_k} \approx \arg\max_{\boldsymbol{\beta}} \hat{f}_G(\boldsymbol{\alpha}_k, \boldsymbol{\beta}) \\ \quad \text{Similarly} \end{cases}$$

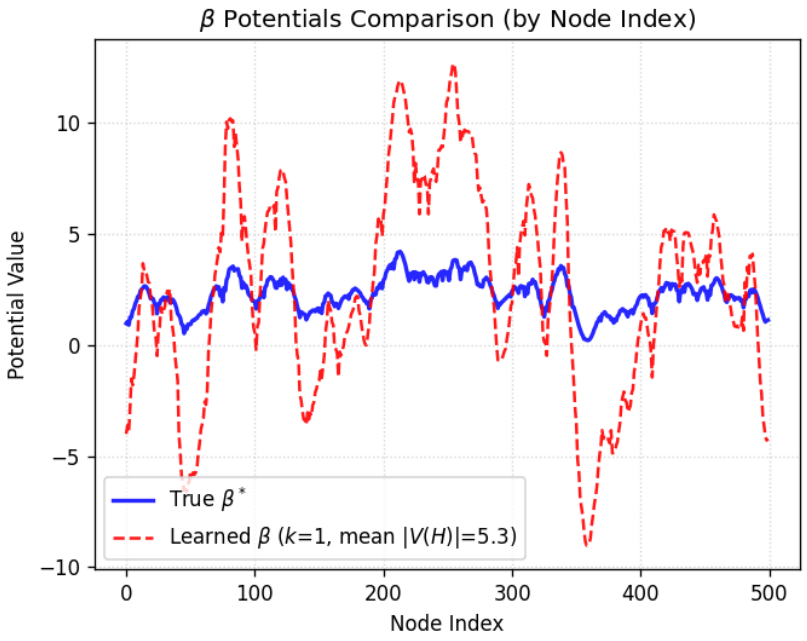
Alg. (Full Sinkhorn)

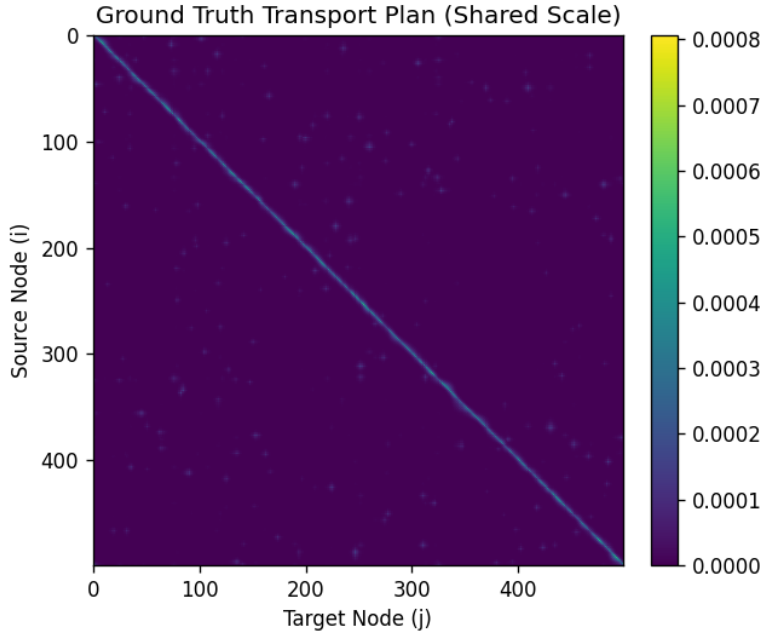
$$\begin{cases} \boldsymbol{\alpha}_k \leftarrow \arg\max_{\boldsymbol{\alpha}} f(\boldsymbol{\alpha}, \boldsymbol{\beta}_{k-1}) \\ \boldsymbol{\beta}_k \leftarrow \arg\max_{\boldsymbol{\beta}} f(\boldsymbol{\alpha}_k, \boldsymbol{\beta}) \end{cases}$$

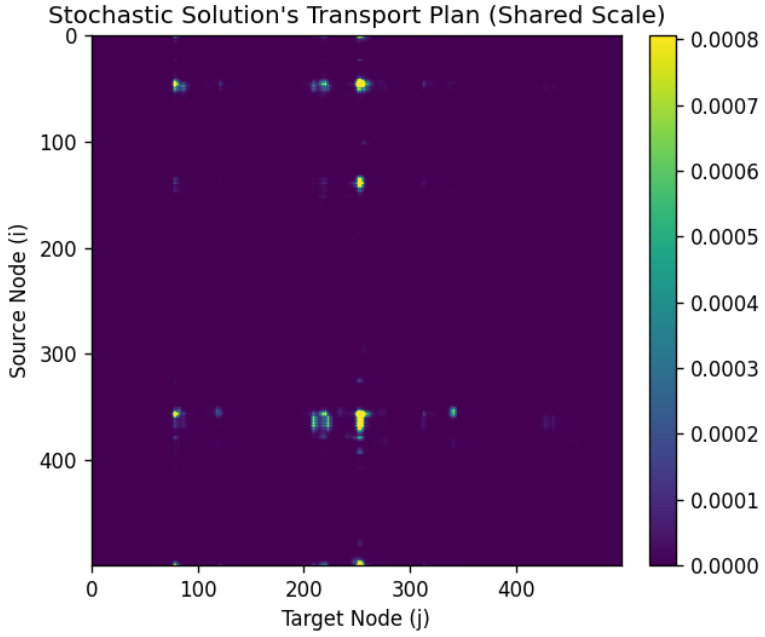
A practical choice of random subgraph H
= k-ball at a uniformly random center

RC-Sinkhorn (k-ball, k=1) vs. Ground Truth N=500, ε=1.0, K_outer=3000, N_inner=10, η=0.25 [K3000_N10_eta0p25]

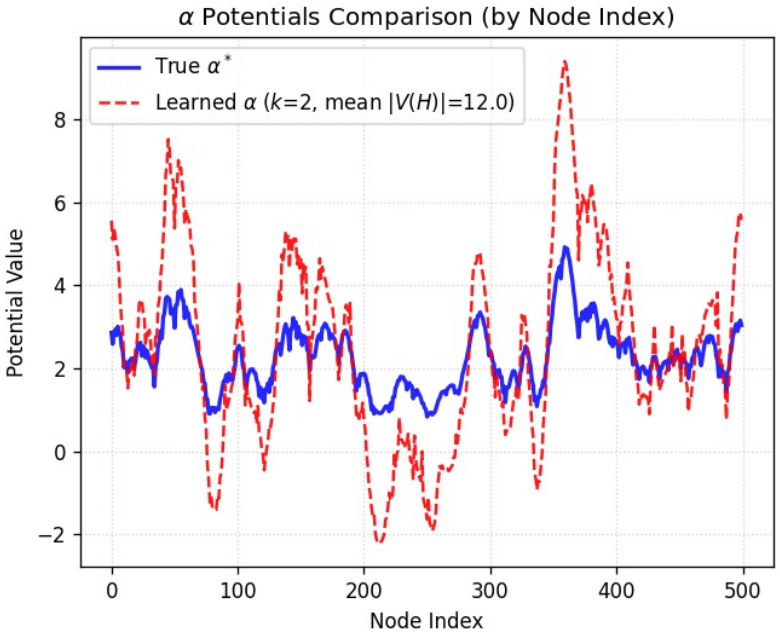


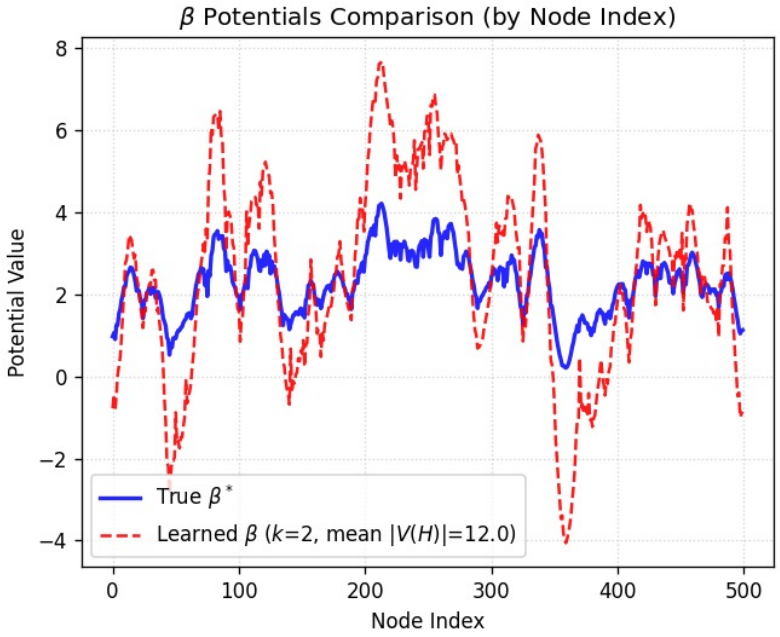


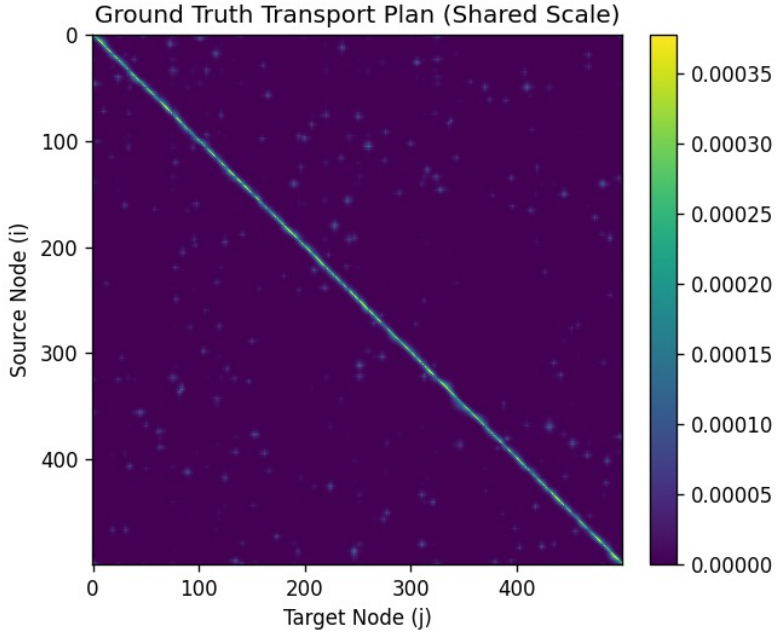


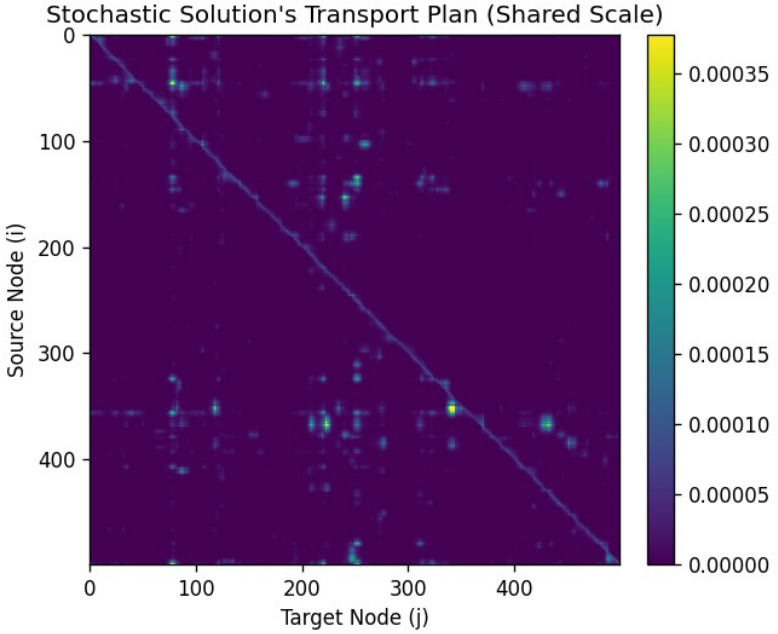


RC-Sinkhorn (k-ball, k=2) vs. Ground Truth N=500, ε=1.0, K_outer=3000, N_inner=10, η=0.25 [K3000_N10_eta0p25]

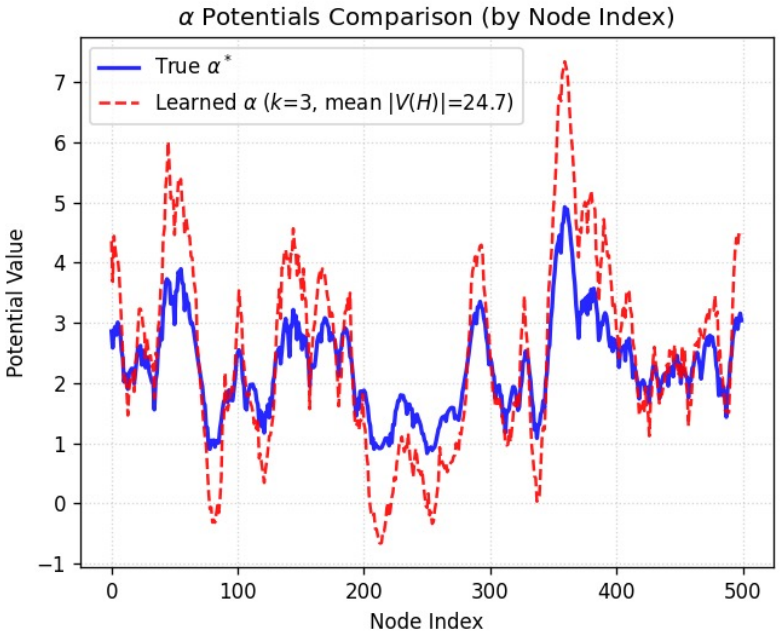


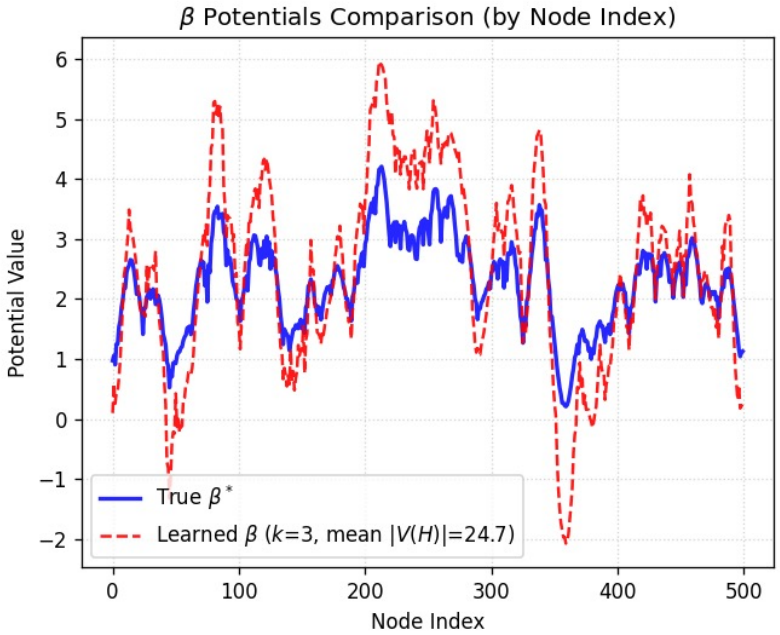


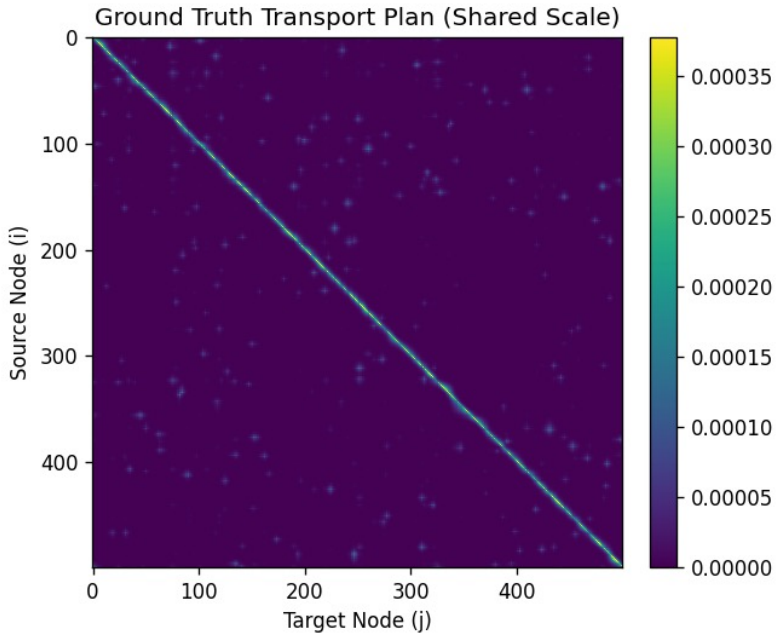


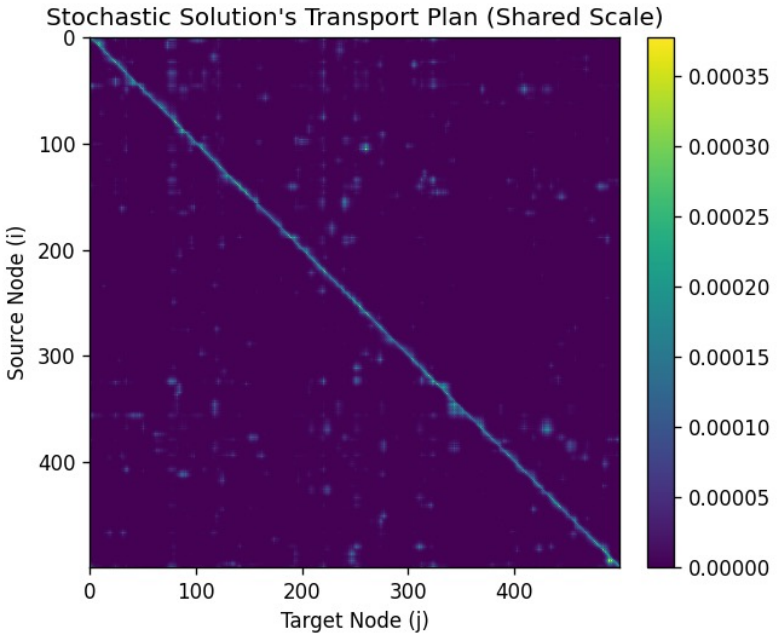


RC-Sinkhorn (k-ball, k=3) vs. Ground Truth N=500, ε=1.0, K_outer=3000, N_inner=10, η=0.25 [K3000_N10_eta0p25]

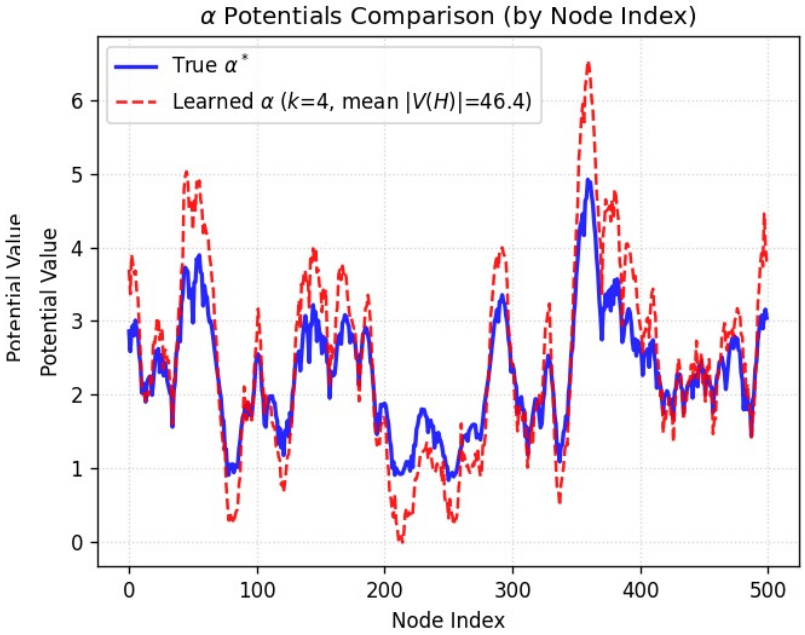


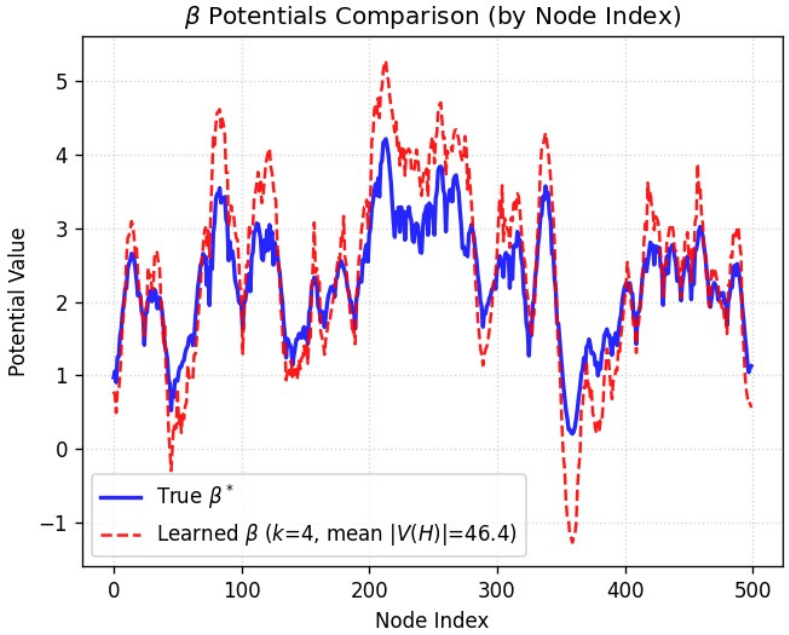


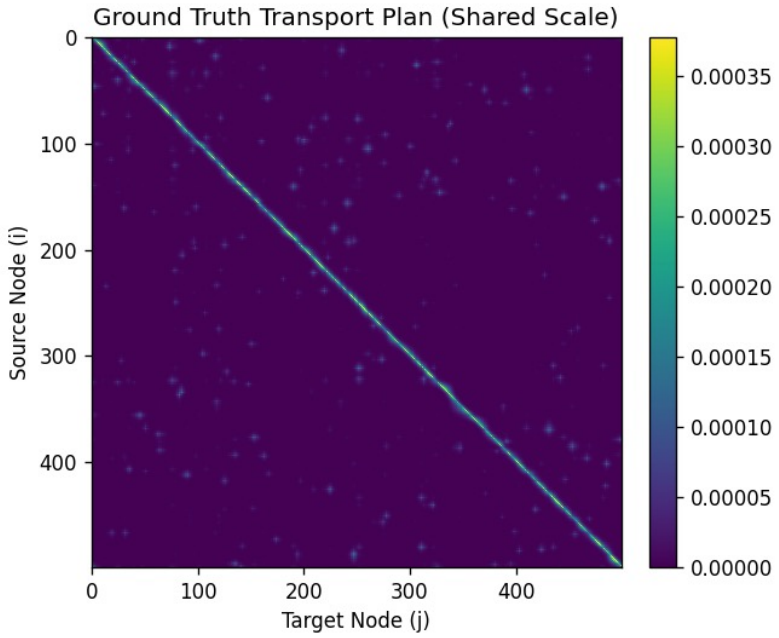


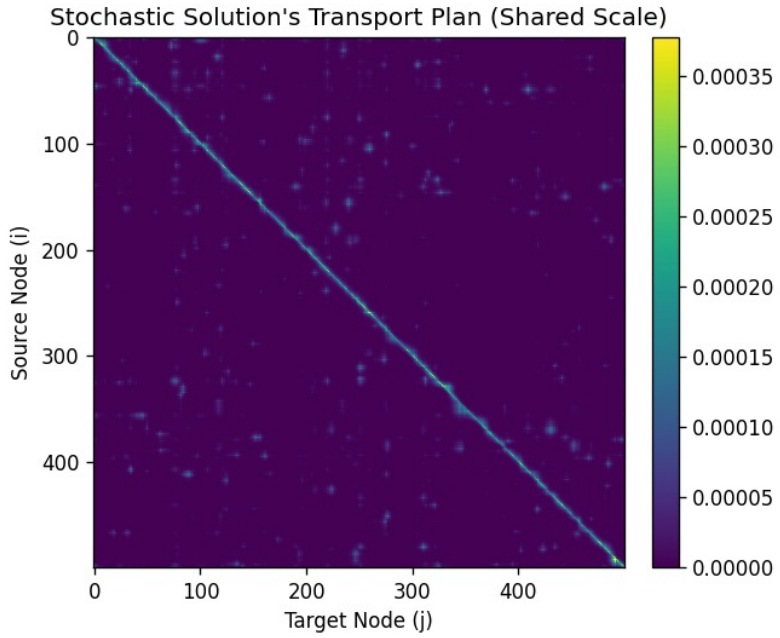


RC-Sinkhorn (k-ball, k=4) vs. Ground Truth N=500, ε=1.0, K_outer=3000, N_inner=10, η=0.25 [K3000_N10_eta0p25]









Convergence analysis

Alg. (Random Coordinate Sinkhorn)

$$
\begin{cases}
\boldsymbol{\alpha}_k \leftarrow \boldsymbol{\alpha}_{k;N_k} \approx \operatorname{argmax}_{\boldsymbol{\alpha}} \hat{f}_G(\boldsymbol{\alpha}, \boldsymbol{\beta}_{k-1}) \\
\quad \text{for } i = 1, \dots, N_k: \\
\quad\quad \text{Sample } H_i \sim \pi \\
\quad\quad \boldsymbol{\alpha}_{k;i+\frac{1}{2}} \leftarrow \operatorname{argmax}_{\boldsymbol{\alpha}} \hat{f}_{H_i}(\boldsymbol{\alpha}, \boldsymbol{\beta}_{k-1}) \\
\quad\quad \boldsymbol{\alpha}_{k;i} \leftarrow (1-\eta)\boldsymbol{\alpha}_{k;i-1} + \eta \boldsymbol{\alpha}_{k;i+\frac{1}{2}} \\
\boldsymbol{\beta}_k \leftarrow \boldsymbol{\beta}_{k;N_k} \approx \operatorname{argmax}_{\boldsymbol{\beta}} \hat{f}_G(\boldsymbol{\alpha}_k, \boldsymbol{\beta}) \\
\quad \text{Similarly}
\end{cases}
\subseteq
$$

Alg. (Inexact Sinkhorn)

$$
\begin{cases}
\boldsymbol{\alpha}_k \approx \operatorname{argmax}_{\boldsymbol{\alpha}} f(\boldsymbol{\alpha}, \boldsymbol{\beta}_{k-1}) \\
\boldsymbol{\beta}_k \approx \operatorname{argmax}_{\boldsymbol{\beta}} f(\boldsymbol{\alpha}_k, \boldsymbol{\beta})
\end{cases}
$$

▶ Convergence analysis

Alg. (Random Coordinate Sinkhorn)

$$
\begin{cases}
\boldsymbol{\alpha}_k \leftarrow \boldsymbol{\alpha}_{k;N_k} \approx \operatorname{argmax}_{\boldsymbol{\alpha}} \hat{f}_G(\boldsymbol{\alpha}, \boldsymbol{\beta}_{k-1}) \\
\quad \text{for } i = 1, \dots, N_k: \\
\quad\quad \text{Sample } H_i \sim \pi \\
\quad\quad \boldsymbol{\alpha}_{k;i+\frac{1}{2}} \leftarrow \operatorname{argmax}_{\boldsymbol{\alpha}} \hat{f}_{H_i}(\boldsymbol{\alpha}, \boldsymbol{\beta}_{k-1}) \\
\quad\quad \boldsymbol{\alpha}_{k;i} \leftarrow (1-\eta)\boldsymbol{\alpha}_{k;i-1} + \eta \boldsymbol{\alpha}_{k;i+\frac{1}{2}} \\
\boldsymbol{\beta}_k \leftarrow \boldsymbol{\beta}_{k;N_k} \approx \operatorname{argmax}_{\boldsymbol{\beta}} \hat{f}_G(\boldsymbol{\alpha}_k, \boldsymbol{\beta}) \\
\quad \text{Similarly}
\end{cases}
\quad \subseteq
$$

Alg. (Inexact Sinkhorn)

$$
\begin{cases}
\boldsymbol{\alpha}_k \approx \operatorname{argmax}_{\boldsymbol{\alpha}} f(\boldsymbol{\alpha}, \boldsymbol{\beta}_{k-1}) \\
\boldsymbol{\beta}_k \approx \operatorname{argmax}_{\boldsymbol{\beta}} f(\boldsymbol{\alpha}_k, \boldsymbol{\beta})
\end{cases}
$$

A key issue in full Sinkhorn analysis:

The dual objective $f(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \langle \mathbf{r}, \boldsymbol{\alpha} \rangle + \langle \boldsymbol{\beta}, \mathbf{c} \rangle - \langle \exp(\boldsymbol{\alpha} \oplus \boldsymbol{\beta}), \mathbf{W} \rangle$

is concave but has unbounded level sets

Unbounded smoothness and strong concavity parameters

▶ Convergence analysis

Alg. (Random Coordinate Sinkhorn)

$$
\begin{cases}
\boldsymbol{\alpha}_k \leftarrow \boldsymbol{\alpha}_{k;N_k} \approx \operatorname{argmax}_{\boldsymbol{\alpha}} \hat{f}_G(\boldsymbol{\alpha}, \boldsymbol{\beta}_{k-1}) \\
\quad \text{for } i = 1, \dots, N_k: \\
\quad\quad \text{Sample } H_i \sim \pi \\
\quad\quad \boldsymbol{\alpha}_{k;i+\frac{1}{2}} \leftarrow \operatorname{argmax}_{\boldsymbol{\alpha}} \hat{f}_{H_i}(\boldsymbol{\alpha}, \boldsymbol{\beta}_{k-1}) \\
\quad\quad \boldsymbol{\alpha}_{k;i} \leftarrow (1-\eta)\boldsymbol{\alpha}_{k;i-1} + \eta \boldsymbol{\alpha}_{k;i+\frac{1}{2}} \\
\boldsymbol{\beta}_k \leftarrow \boldsymbol{\beta}_{k;N_k} \approx \operatorname{argmax}_{\boldsymbol{\beta}} \hat{f}_G(\boldsymbol{\alpha}_k, \boldsymbol{\beta}) \\
\quad \text{Similarly}
\end{cases}
\quad \subseteq
$$

Alg. (Inexact Sinkhorn)

$$
\begin{cases}
\boldsymbol{\alpha}_k \approx \operatorname{argmax}_{\boldsymbol{\alpha}} f(\boldsymbol{\alpha}, \boldsymbol{\beta}_{k-1}) \\
\boldsymbol{\beta}_k \approx \operatorname{argmax}_{\boldsymbol{\beta}} f(\boldsymbol{\alpha}_k, \boldsymbol{\beta})
\end{cases}
$$

A key issue in full Sinkhorn analysis:

The dual objective $f(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \langle \mathbf{r}, \boldsymbol{\alpha} \rangle + \langle \boldsymbol{\beta}, \mathbf{c} \rangle - \langle \exp(\boldsymbol{\alpha} \oplus \boldsymbol{\beta}), \mathbf{W} \rangle$

is concave but has unbounded level sets

Unbounded smoothness and strong concavity parameters

A priori confinement analysis: $\| \boldsymbol{\alpha}_k - \boldsymbol{\alpha}^* \|_\infty, \| \boldsymbol{\beta}_k - \boldsymbol{\beta}^* \|_\infty \leq$ **non-increasing in** k

(The Jacobian of the Sinkhorn map is row-stochastic)

- **Convergence analysis**

Alg. (Inexact Sinkhorn)

$$\begin{cases} \boldsymbol{\alpha}_k \approx \arg\max_{\boldsymbol{\alpha}} f(\boldsymbol{\alpha}, \boldsymbol{\beta}_{k-1}) \\ \boldsymbol{\beta}_k \approx \arg\max_{\boldsymbol{\beta}} f(\boldsymbol{\alpha}_k, \boldsymbol{\beta}) \end{cases}$$

Thm. (Bi, Kim, L. 26+)

$\Delta_k = \sup f - f(\alpha_k, \beta_k) = O(k^{-2})$

provided asymptotically exact updates

- **Convergence analysis**

Alg. (Inexact Sinkhorn)

$$\begin{cases} \boldsymbol{\alpha}_k \approx \arg\max_{\boldsymbol{\alpha}} f(\boldsymbol{\alpha}, \boldsymbol{\beta}_{k-1}) \\ \boldsymbol{\beta}_k \approx \arg\max_{\boldsymbol{\beta}} f(\boldsymbol{\alpha}_k, \boldsymbol{\beta}) \end{cases}$$

Thm. (Bi, Kim, L. 26+)

$\Delta_k = \sup f - f(\alpha_k, \beta_k) = O(k^{-2})$

provided asymptotically exact updates

Can we show boundedness of iterates? $\|\boldsymbol{\alpha_k} - \boldsymbol{\alpha}^*\|_\infty, \|\boldsymbol{\beta_k} - \boldsymbol{\beta}^*\|_\infty \leq \boldsymbol{M}$?

WTS: $D_k = \max\left\{\|\alpha_k - \alpha^*\|_{sp}, \|\beta_k - \beta^*\|_{sp}\right\}$ is uniformly bdd

Convergence analysis

Alg. (Inexact Sinkhorn)

$$\begin{cases} \boldsymbol{\alpha}_k \approx \arg\max_{\boldsymbol{\alpha}} f(\boldsymbol{\alpha}, \boldsymbol{\beta}_{k-1}) \\ \boldsymbol{\beta}_k \approx \arg\max_{\boldsymbol{\beta}} f(\boldsymbol{\alpha}_k, \boldsymbol{\beta}) \end{cases}$$

Thm. (Bi, Kim, L. 26+)

$\Delta_k = \sup f - f(\alpha_k, \beta_k) = O(k^{-2})$

provided asymptotically exact updates

Can we show boundedness of iterates? $\|\boldsymbol{\alpha}_k - \boldsymbol{\alpha}^*\|_\infty, \|\boldsymbol{\beta}_k - \boldsymbol{\beta}^*\|_\infty \le \boldsymbol{M}$?

WTS: $D_k = \max\left\{\|\alpha_k - \alpha^*\|_{sp}, \|\beta_k - \beta^*\|_{sp}\right\}$ is uniformly bdd

We show the following recursion:

$$D_{k+1} \le \tau(D_k) D_k + c_k$$

where $c_k = \|\alpha_k - \alpha_k^{\text{exact}}(\hat{\beta}_{k-1})\|_{sp} + \|\beta_k - \beta_k^{\text{exact}}(\hat{\alpha}_{k-1})\|_{sp}$

$\tau(D) =$ Explicit sp contraction ratio $(< 1 \; for \; D < \infty)$

If c_k is summable (can be controlled), then the recursion gives $\sup_k D_k \le M$

New Perspective II: Random Matrix Scaling

- Recall matrix scaling..

Primal **Dual**

$$\min_{\mathbf{Z}\in\mathcal{T}(\mathbf{r},\mathbf{c})} \left[D_{KL}(\mathbf{Z}\,\|\mathbf{W})\right] = \max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \left[\langle \mathbf{r}, \boldsymbol{\alpha}\rangle + \langle \boldsymbol{\beta}, \mathbf{c}\rangle - \langle \exp(\boldsymbol{\alpha}\oplus\boldsymbol{\beta}), \mathbf{W}\rangle\right]$$

Rescaled Matrix

$\boldsymbol{\alpha}$ = Lagrange multipliers for the row-sum constraint

$\boldsymbol{\beta}$ = Lagrange multipliers for the column-sum constraint

(Schrödinger potentials)

$$\pi^* = \mathrm{Diag}(\exp(\boldsymbol{\alpha}^*))\,\mathbf{W}\,\mathrm{Diag}(\exp(\boldsymbol{\beta}^*))$$

Alternating maximization

Matrix scaling algorithm
(Iterative Proportional Fitting)

Fit the row sums;
Fit the column sums; etc.

Sinkhorn Algorithm

$$\begin{cases} \forall 1 \le i \le n,\ \boldsymbol{\beta}_k(j) \leftarrow \log \dfrac{\mathbf{c}(j)}{\sum_{i=1}^m \mathbf{W}_{ij}\exp(\boldsymbol{\alpha}_{k-1}(i))}, \\ \forall 1 \le i \le m,\ \boldsymbol{\alpha}_k(i) \leftarrow \log \dfrac{\mathbf{r}(i)}{\sum_{j=1}^n \mathbf{W}_{ij}\exp(\boldsymbol{\beta}_k(j))}. \end{cases}$$

- Sinkhorn-rescale a Random matrix X **??**

Primal **Dual**

$$\min_{\mathbf{Z}\in\mathcal{T}(\mathbf{r},\mathbf{c})} \left[D_{KL}(\mathbf{Z}\,\|\,\boldsymbol{X})\right] = \max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \left[\langle \mathbf{r}, \boldsymbol{\alpha}\rangle + \langle \boldsymbol{\beta}, \mathbf{c}\rangle - \langle \exp(\boldsymbol{\alpha}\oplus\boldsymbol{\beta}), \boldsymbol{X}\rangle\right]$$

Rescaled Matrix

$\boldsymbol{\alpha}$ = Lagrange multipliers for the row-sum constraint

$\boldsymbol{\beta}$ = Lagrange multipliers for the column-sum constraint

(Schrödinger potentials)

$$\boldsymbol{X}^{r,c} := \pi^* = \mathrm{Diag}(\exp(\boldsymbol{\alpha}^*))\,\boldsymbol{X}\,\mathrm{Diag}(\exp(\boldsymbol{\beta}^*))$$

Alternating maximization

Matrix scaling algorithm
(Iterative Proportional Fitting)

Fit the row sums;
Fit the column sums; etc.

Sinkhorn Algorithm

$$\begin{cases} \forall 1 \le i \le n,\ \boldsymbol{\beta}_k(j) \leftarrow \log \dfrac{\mathbf{c}(j)}{\sum_{i=1}^{m} \boldsymbol{X}_{ij} \exp(\boldsymbol{\alpha}_{k-1}(i))}, \\ \forall 1 \le i \le m,\ \boldsymbol{\alpha}_k(i) \leftarrow \log \dfrac{\mathbf{r}(i)}{\sum_{j=1}^{n} \boldsymbol{X}_{ij} \exp(\boldsymbol{\beta}_k(j))}. \end{cases}$$

- Sinkhorn-rescale a Random matrix X **??**

Primal **Dual**

$$\min_{\mathbf{Z}\in\mathcal{T}(\mathbf{r},\mathbf{c})} \left[D_{KL}(\mathbf{Z}\|\boldsymbol{X})\right] = \max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \left[\langle \mathbf{r},\boldsymbol{\alpha}\rangle + \langle \boldsymbol{\beta},\mathbf{c}\rangle - \langle \exp(\boldsymbol{\alpha}\oplus\boldsymbol{\beta}),\boldsymbol{X}\rangle\right]$$

Rescaled Matrix

$\boldsymbol{\alpha}$ = Lagrange multipliers for the row-sum constraint

$\boldsymbol{\beta}$ = Lagrange multipliers for the column-sum constraint

(Schrödinger potentials)

$$\boldsymbol{X}^{r,c} = \mathrm{Diag}(\exp(\boldsymbol{\alpha}^*))\,\boldsymbol{X}\,\mathrm{Diag}(\exp(\boldsymbol{\beta}^*))$$

Depends on X!

Alternating maximization

Matrix scaling algorithm
(Iterative Proportional Fitting)

Fit the row sums;
Fit the column sums; etc.

Sinkhorn Algorithm

$$\begin{cases} \forall 1\le i\le n,\ \boldsymbol{\beta}_k(j) \leftarrow \log \dfrac{\mathbf{c}(j)}{\sum_{i=1}^m \boldsymbol{X}_{ij}\exp(\boldsymbol{\alpha}_{k-1}(i))}, \\ \forall 1\le i\le m,\ \boldsymbol{\alpha}_k(i) \leftarrow \log \dfrac{\mathbf{r}(i)}{\sum_{j=1}^n \boldsymbol{X}_{ij}\exp(\boldsymbol{\beta}_k(j))}. \end{cases}$$

- Sinkhorn-rescale a Random matrix $\boldsymbol{X}$ **??**

$$\boldsymbol{X}^{\boldsymbol{r},\boldsymbol{c}} = \mathrm{Diag}(\exp(\boldsymbol{\alpha}^*))\,\boldsymbol{X}\,\mathrm{Diag}(\exp(\boldsymbol{\beta}^*))$$

Depends on X!

Thm. (Duan, Powell, L. 26+)

$\boldsymbol{X}^{\boldsymbol{r},\boldsymbol{c}} \approx (\mathbb{E}\boldsymbol{X})^{\boldsymbol{r},\boldsymbol{c}}$ (High-probability concentration in L1, L2 norm)

$\boldsymbol{X}^{\boldsymbol{r},\boldsymbol{c}} \to \pi^* = \underset{\pi\in\Pi(\mathbf{r}^\infty,\mathbf{c}^\infty)}{\mathrm{argmin}}\ D_{KL}(\pi \,\|\, \mathbb{E}\mathbf{X}^\infty)$ (weak convergence w/ explicit rates)

Empirical Singular Value Dist. $(\frac{mn}{N\sqrt{n}}(\boldsymbol{X}^{\boldsymbol{r},\boldsymbol{c}} - (\mathbb{E}\boldsymbol{X})^{\boldsymbol{r},\boldsymbol{c}})) \to$ limiting distribution

CLT for potentials $(\boldsymbol{\alpha},\boldsymbol{\beta})$ for mean$(\mathbf{X}_1,\dots,\boldsymbol{X}_M)$ and margin $(\boldsymbol{r},\boldsymbol{c})$

- Sinkhorn-rescale a Random matrix $\boldsymbol{X}$ **??**

$$\boldsymbol{X}^{\boldsymbol{r},\boldsymbol{c}} = \mathrm{Diag}(\exp(\boldsymbol{\alpha}^*))\,\boldsymbol{X}\,\mathrm{Diag}(\exp(\boldsymbol{\beta}^*)) \quad \approx \quad \hat{\mathbf{X}}^{\mathbf{r},\mathbf{c}} = \mathrm{Diag}(\exp(\overline{\boldsymbol{\alpha}}))\,\mathbf{X}\,\mathrm{Diag}(\exp(\overline{\boldsymbol{\beta}}))$$

Depends on X!

Depends on $\mathbb{E}X$!

Thm. (Duan, Powell, L. 26+)

$\boldsymbol{X}^{\boldsymbol{r},\boldsymbol{c}} \approx (\mathbb{E}\boldsymbol{X})^{\boldsymbol{r},\boldsymbol{c}}$ (High-probability concentration in L1, L2 norm)

$\boldsymbol{X}^{\boldsymbol{r},\boldsymbol{c}} \to \pi^* = \underset{\pi\in\Pi(\mathbf{r}^\infty,\mathbf{c}^\infty)}{\mathrm{argmin}}\ D_{KL}(\pi \,\|\, \mathbb{E}\mathbf{X}^\infty)$ (weak convergence w/ explicit rates)

Empirical Singular Value Dist. $(\frac{mn}{N\sqrt{n}}(\boldsymbol{X}^{\boldsymbol{r},\boldsymbol{c}} - (\mathbb{E}\boldsymbol{X})^{\boldsymbol{r},\boldsymbol{c}})) \to$ limiting distribution

CLT for potentials $(\boldsymbol{\alpha},\boldsymbol{\beta})$ for mean$(\mathbf{X}_1,\ldots,\boldsymbol{X}_M)$ and margin $(\boldsymbol{r},\boldsymbol{c})$

- Key ingredients: Stability Theory for Matrix Scaling

$$\mathbf{\Lambda}^{\mathbf{r},\mathbf{c}} = \underset{\mathbf{Z}\in\mathcal{T}(\mathbf{r},\mathbf{c})}{\arg\min}\, D_{KL}(\mathbf{Z}\,\|\,\mathbf{\Lambda}) = \mathrm{Diag}(\exp(\boldsymbol{\alpha}))\,\mathbf{\Lambda}\,\mathrm{Diag}(\exp(\boldsymbol{\beta}))$$

How much does the rescaled matrix / potentials change when we perturb the input $\mathbf{\Lambda}$ or margin $(\boldsymbol{r}, \boldsymbol{c})$?

- Key ingredients: Stability Theory for Matrix Scaling

$$\mathbf{\Lambda}^{\mathbf{r},\mathbf{c}} = \operatorname*{arg\,min}_{\mathbf{Z}\in\mathcal{T}(\mathbf{r},\mathbf{c})} D_{KL}(\mathbf{Z}\,\|\,\mathbf{\Lambda}) = \mathrm{Diag}(\exp(\boldsymbol{\alpha}))\,\mathbf{\Lambda}\,\mathrm{Diag}(\exp(\boldsymbol{\beta}))$$

How much does the rescaled matrix / potentials change when we perturb the input $\mathbf{\Lambda}$ or margin $(\boldsymbol{r},\boldsymbol{c})$?

Thm. (Duan, Powell, L. 26+)

$$d_{\text{Hellinger}}(\Lambda^{r,c},\Sigma^{r,c}) \le C\; d_{\text{Hellinger}}(\Lambda,\Sigma)$$

$$d_{\text{Hellinger}}\left(\Lambda^{\boldsymbol{r},\boldsymbol{c}},\Lambda^{\boldsymbol{r}',\boldsymbol{c}'}\right) \le C'\|(\boldsymbol{r},\boldsymbol{c})-(\boldsymbol{r}',\boldsymbol{c}')\,\|_1$$

$$\|\left(\boldsymbol{\alpha}_{\Lambda}^{\boldsymbol{r},\boldsymbol{c}},\boldsymbol{\beta}_{\Lambda}^{\boldsymbol{r},\boldsymbol{c}}\right)-\left(\boldsymbol{\alpha}_{\Lambda}^{\boldsymbol{r}',\boldsymbol{c}'},\boldsymbol{\beta}_{\Lambda}^{\boldsymbol{r}',\boldsymbol{c}'}\right)\|_\infty \le C''(\text{relative margin error})$$

New Perspective III: Poisson Contingency Tables

- $\boldsymbol{X}_{|\boldsymbol{r},\boldsymbol{c}} = \boldsymbol{X} \sim \text{Poisson}(\boldsymbol{\Lambda})$ conditioned on having margin ($\mathbf{r}$,$\mathbf{c}$) ??

  If $\boldsymbol{\Lambda}$ =rank 1, then $X_{|r,c} \sim$ Multivariate Hypergeometric Dist.

  For general $\boldsymbol{\Lambda}$, no explicit form, no approximation scheme, etc.

- $\boldsymbol{X}_{|\boldsymbol{r},\boldsymbol{c}} = \boldsymbol{X} \sim \text{Poisson}(\boldsymbol{\Lambda})$ conditioned on having margin $(\mathbf{r},\mathbf{c})$??

  If $\boldsymbol{\Lambda}$ =rank 1, then $X_{|r,c} \sim$ Multivariate Hypergeometric Dist.

  For general $\boldsymbol{\Lambda}$, no explicit form, no approximation scheme, etc.

Thm. (Duan, Powell, L., Mukherjee 26+)

$\boldsymbol{X}_{|\boldsymbol{r},\boldsymbol{c}} =_{\boldsymbol{d}}$ certain Gibbs random permutation of length N=total sum

$$\boldsymbol{X}_{|\boldsymbol{r},\boldsymbol{c}} \to \pi^* = \underset{\pi \in \Pi(\mathbf{r}^\infty, \mathbf{c}^\infty)}{\operatorname{argmin}} D_{KL}(\pi \,\|\, \mathbb{E}\mathbf{X}^\infty)$$

(weak convergence w/ explicit rates)

Empirical Singular Value Dist. $\left(\frac{mn}{N\sqrt{n}}(\boldsymbol{X}_{|\boldsymbol{r},\boldsymbol{c}} - \boldsymbol{\Lambda}^{\boldsymbol{r},\boldsymbol{c}})\right) \to$ limiting distribution

- $X_{|r,c} = X \sim \text{Poisson}(\Lambda)$ conditioned on having margin $(\mathbf{r},\mathbf{c})$??

  If Λ =rank 1, then $X_{|r,c} \sim$ Multivariate Hypergeometric Dist.

  For general Λ, no explicit form, no approximation scheme, etc.

Thm. (Duan, Powell, L., Mukherjee 26+)

$X_{|r,c} =_d$ certain Gibbs random permutation of length N=total sum

$$X_{|r,c} \to \pi^* = \operatorname*{argmin}_{\pi \in \Pi(\mathbf{r}^\infty, \mathbf{c}^\infty)} D_{KL}(\pi \,\|\, \mathbb{E}\mathbf{X}^\infty)$$

(weak convergence w/ explicit rates)

Empirical Singular Value Dist. $\left(\frac{mn}{N\sqrt{n}}(X_{|r,c} - \Lambda^{r,c})\right) \to$ limiting distribution

Proof uses: Transference principle, Varadhan's lemma, Barvinok's permanent identity

Thank you very much!