August 4, 2017
ERATO 感謝祭 SeasonIV

Inter-University Research Institute Corporation / Research Organization of Information and Systems
National Institute of Informatics

PRESTO
SAKIGAKE

# Tensor Balancing on Statistical Manifold (ICML 2017)

Mahito Sugiyama[1,2] Hiroyuki Nakahara[3] Koji Tsuda[4,5,6]
[1]NII, [2]JST PRESTO, [3]RIKEN BSI, [4]UTokyo, [5]RIKEN AIP, [6]NIMS

# Results

- Balancing of higher order (more than two) tensors is firstly (theoretically) achieved
  - We present a balancing algorithm and prove its global convergence
- A fast balancing algorithm with quadratic convergence using Newton's method
  - An existing algorithm is linear convergence
- **[Theory]** We provide dually flat Riemannian manifold of probability distributions with the structured outcome space
  - Tensor balancing is an instance

# Matrix Balancing

$$\begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} & p_{34} \\ p_{41} & p_{42} & p_{43} & p_{44} \end{bmatrix}$$

# Matrix Balancing

$$
\begin{array}{cc}
\begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} & p_{34} \\ p_{41} & p_{42} & p_{43} & p_{44} \end{bmatrix} & \boxed{\begin{matrix} r_1 \\ r_2 \\ r_3 \\ r_4 \end{matrix}} \\
\boxed{\begin{matrix} a_1 & a_2 & a_3 & a_4 \end{matrix}} &
\end{array}
$$

$\boxed{\text{Find } \boldsymbol{r} \text{ and } \boldsymbol{a}}$

$$r_2 a_1 p_{21} + r_2 a_2 p_{22} + r_2 a_3 p_{23} + r_2 a_4 p_{24} = 1$$

$$r_1 a_3 p_{13} + r_2 a_3 p_{23} + r_3 a_3 p_{33} + r_4 a_3 p_{43} = 1$$

# Matrix Balancing

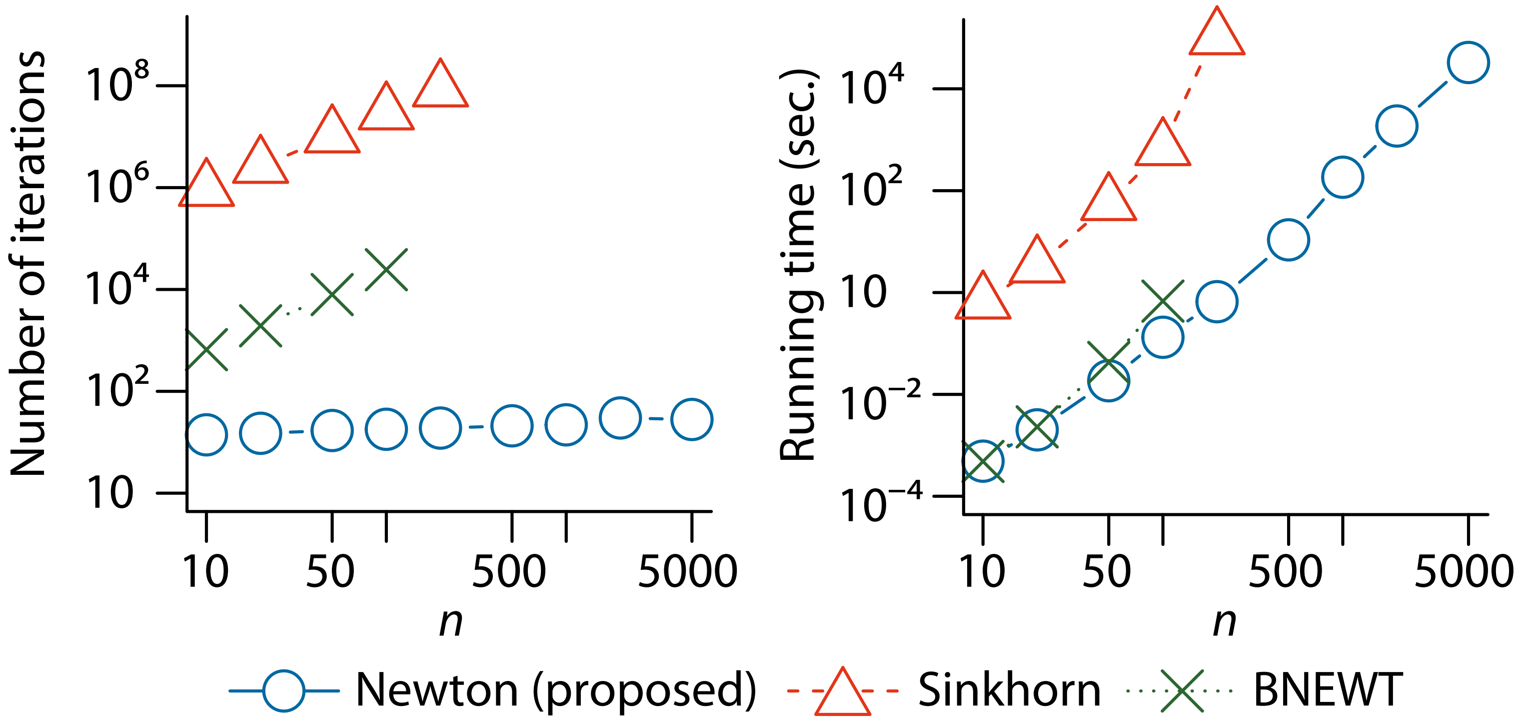
- Problem setting:
  Given a nonnegative matrix $P = (p_{ij}) \in \mathbb{R}_+^{n \times n}$, find $\boldsymbol{r}, \boldsymbol{s} \in \mathbb{R}^n$ s.t.

  $$(RPS)\mathbf{1} = \mathbf{1} \quad \text{and} \quad (RPS)^T\mathbf{1} = \mathbf{1}$$

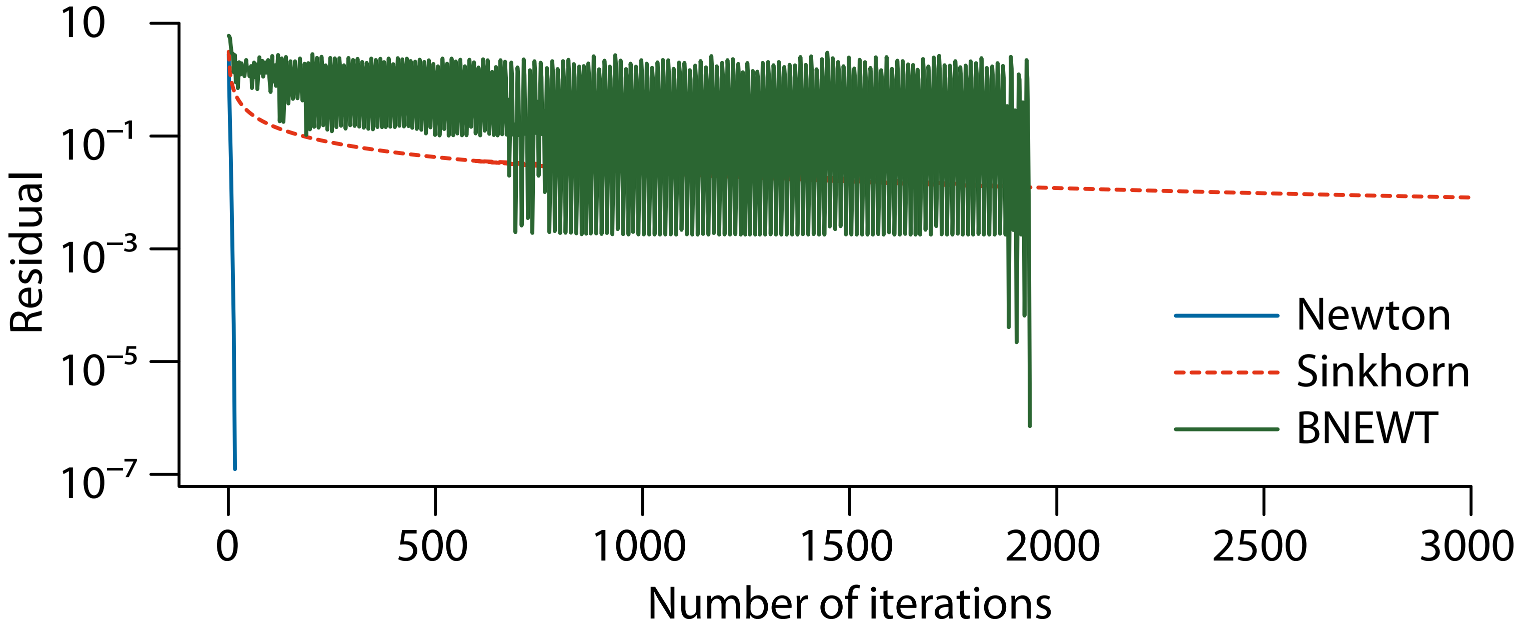
  - $R = \mathrm{diag}(\boldsymbol{r})$, $S = \mathrm{diag}(\boldsymbol{s})$
  - Each entry is given as $p'_{ij} = p_{ij} r_i s_j$

- A fundamental process to analyze and compare matrices in a wide range of applications
  - Input-output analysis in economics, seat assignments in elections, Hi-C data analysis, Sudoku puzzle
  - Approximate Wasserstein distance

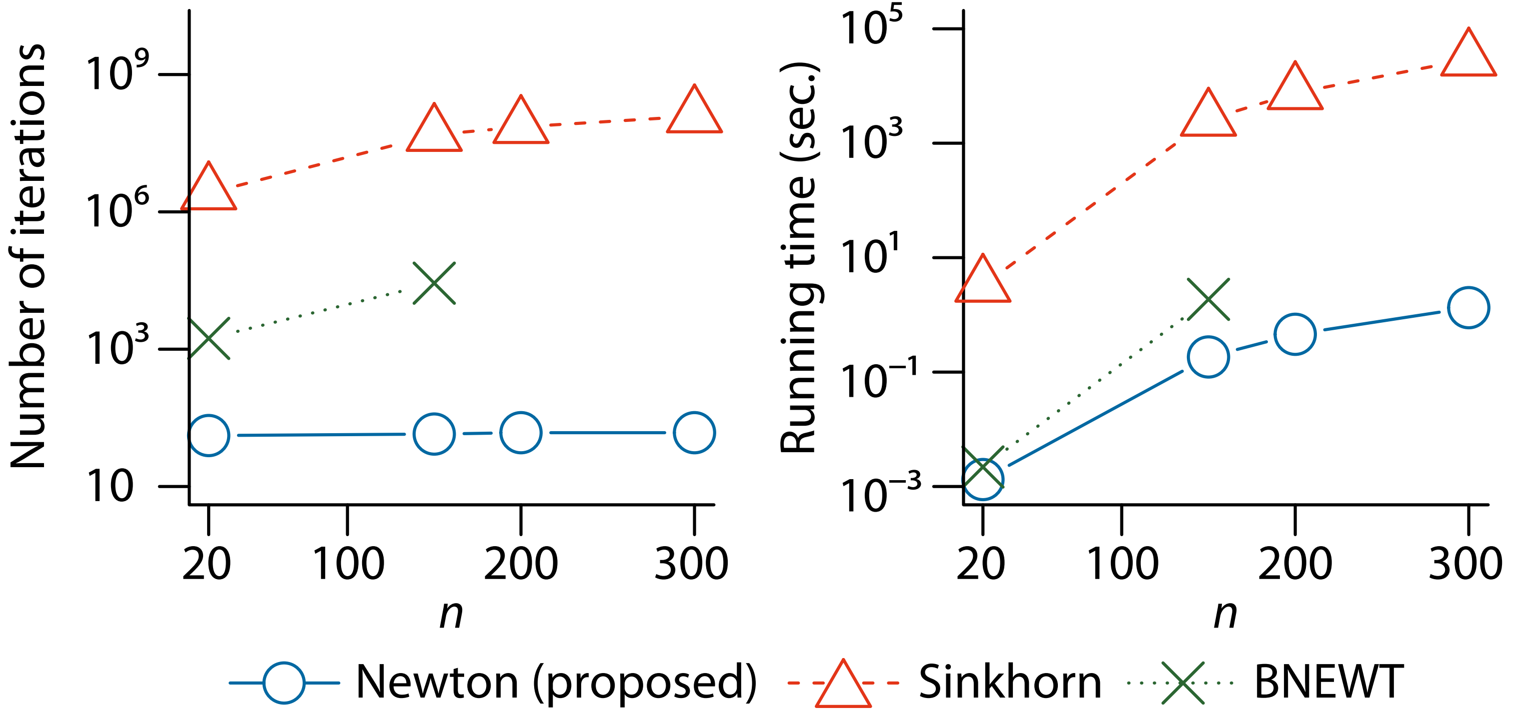
# Results on Hessenberg Matrix



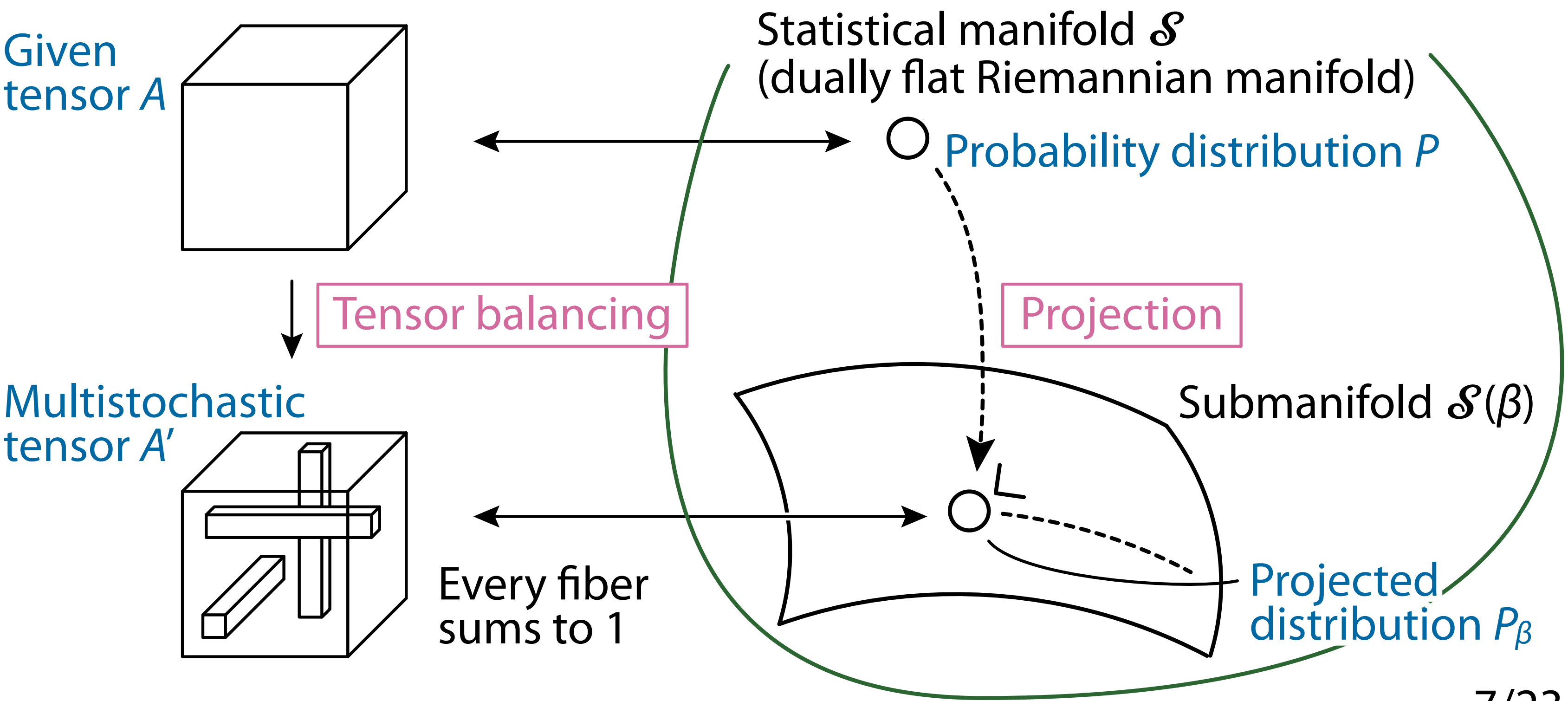
# Results on Hessenberg Matrix ($n = 20$)



# Results on Trefethen Matrix

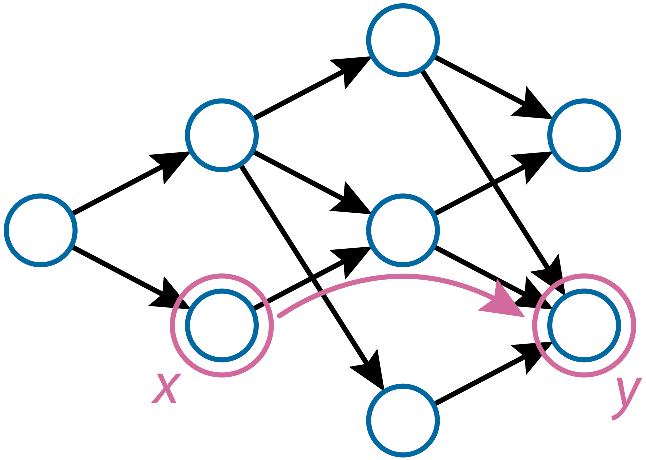


# Overview



Given
tensor $A$
Tensor balancing
Multistochastic
tensor $A'$
Every fiber
sums to 1
Statistical manifold $\mathcal{S}$
(dually flat Riemannian manifold)
Probability distribution $P$
Projection
Submanifold $\mathcal{S}(\beta)$
Projected
distribution $P_\beta$

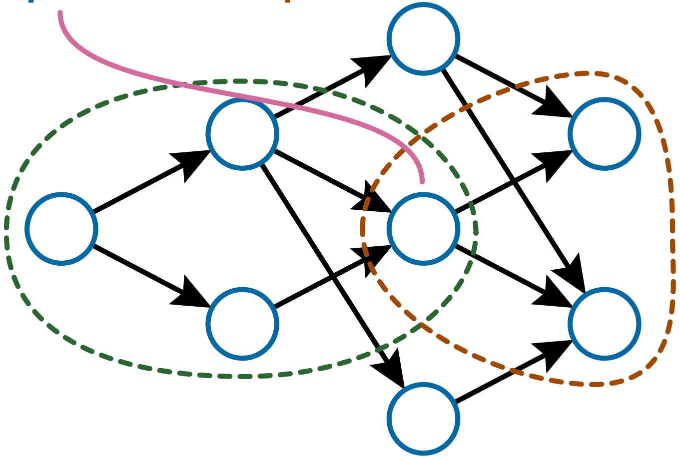
# Partially Ordered Set




- Partially ordered set (poset) $(S, \leq)$
  - (i) $x \leq x$ (reflexivity)
  - (ii) $x \leq y, y \leq x \Rightarrow x = y$ (antisymmetry)
  - (iii) $x \leq y, y \leq z \Rightarrow x \leq z$ (transitivity)
  - – We assume that $S$ is finite and includes the least element (bottom) $\perp \in S$
- Equivalent to a DAG
  - – Each $x \in S$ is a node
  - – $x \leq y \iff y$ is reachable from $x$

# Log-Linear Model on Poset

Each $x \in S$ has a triple:
$(p(x), \theta(x), \eta(x))$

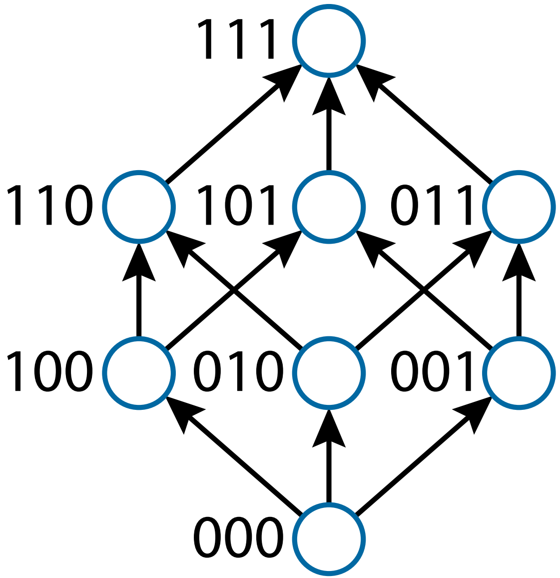


- A probability vector $p : S \to (0, 1)$ s.t. $\sum_{x \in S} p(x) = 1$
  - (Normalized) weight for each node
- We introduce $\theta : S \to \mathbb{R}$ and $\eta : S \to \mathbb{R}$ as

$$\log p(x) = \sum_{s \le x} \theta(s),$$

$$\eta(x) = \sum_{s \ge x} p(s)$$

# Our Model Includes Binary Case




- Our model:

$$\log p(x) = \sum_{s \le x} \theta(s), \qquad \eta(x) = \sum_{s \ge x} p(s)$$

is generalization of the log-linear model on binary vectors with $\boldsymbol{x} \in \{0, 1\}^n = S$:

$$\log p(\boldsymbol{x}) = \sum_i \theta^i x^i + \sum_{i<j} \theta^{ij} x^i x^j + \dots$$
$$+ \theta^{1\dots n} x^1 x^2 \dots x^n - \psi,$$

$$\eta^i = \mathbf{E}[x^i] = \Pr(x^i = 1),$$
$$\eta^{ij} = \mathbf{E}[x^i x^j] = \Pr(x^i = x^j = 1), \dots$$

# Dually Flat Structure

- $\theta$ and $\eta$ form a dual coordinate system:
  $\nabla\psi(\theta) = \eta,\ \nabla\varphi(\eta) = \theta$
  - $\psi(\theta) = -\theta(\perp) = -\log p(\perp),\ \varphi(\eta) = \sum_{x \in S} p(x) \log p(x)$
  - $\psi(\theta)$ and $\varphi(\eta)$ are connected via the Legendre transformation:
    $\varphi(\eta) = \max_{\theta'} \left(\theta'\eta - \psi(\theta')\right), \quad \theta'\eta = \sum_{x \in S \setminus \{\perp\}} \theta'(x)\eta(x)$
    - $\psi(\theta)$ and $\varphi(\eta)$ should be convex

# Gradient and Riemannian Manifold

- The gradients: $g(\theta) = \nabla\nabla\psi(\theta) = \nabla\eta$, $g(\eta) = \nabla\nabla\varphi(\eta) = \nabla\theta$

$$
\begin{cases}
g_{xy}(\theta) = \dfrac{\partial\eta(x)}{\partial\theta(y)} = \displaystyle\sum_{s\in S} \zeta(x,s)\zeta(y,s)p(s) - \eta(x)\eta(y) \\
g_{xy}(\eta) = \dfrac{\partial\theta(x)}{\partial\eta(y)} = \displaystyle\sum_{s\in S} \mu(s,x)\mu(s,y)p(s)^{-1}
\end{cases}
$$

  - $\zeta$ and $\mu$ are the zeta function and the Möbius function determined by the partial order (DAG) structure
  - The manifold $(\mathcal{S}, g(\xi))$ is a Riemannian manifold with the set $\mathcal{S}$ of probability vectors and the Riemannian metric $g(\xi)$

# Fisher Information Matrix and Orthogonality

- Since $g(\xi)$ coincides with the Fisher information matrix,

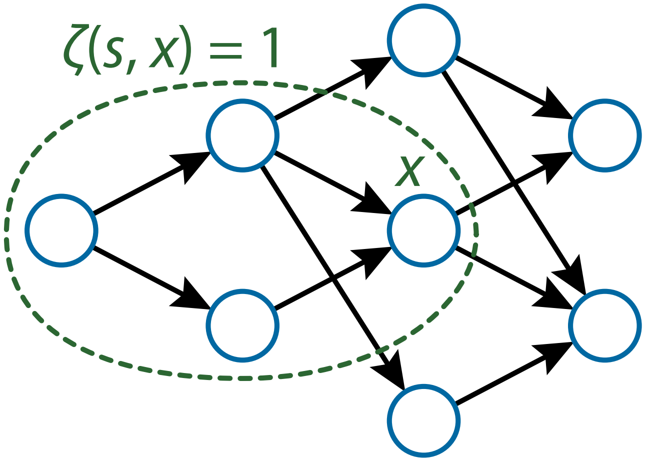
$$\mathbf{E}\left[\frac{\partial}{\partial\theta(x)}\log p(s)\frac{\partial}{\partial\theta(y)}\log p(s)\right]=\sum_{s\in S}\zeta(x,s)\zeta(y,s)p(s)-\eta(x)\eta(y),$$

$$\mathbf{E}\left[\frac{\partial}{\partial\eta(x)}\log p(s)\frac{\partial}{\partial\eta(y)}\log p(s)\right]=\sum_{s\in S}\mu(s,x)\mu(s,y)p(s)^{-1}$$

- $\theta$ and $\eta$ are orthogonal, i.e.,

$$\mathbf{E}\left[\frac{\partial}{\partial\theta(x)}\log p(s)\frac{\partial}{\partial\eta(y)}\log p(s)\right]=\sum_{s\in S}\zeta(x,s)\mu(s,y)=\delta_{xy}$$

# Möbius Function on Poset




- Zeta function $\zeta : S \times S \to \{0, 1\}$

$$\zeta(s, x) = \begin{cases} 1 & \text{if } s \le x, \\ 0 & \text{otherwise.} \end{cases}$$

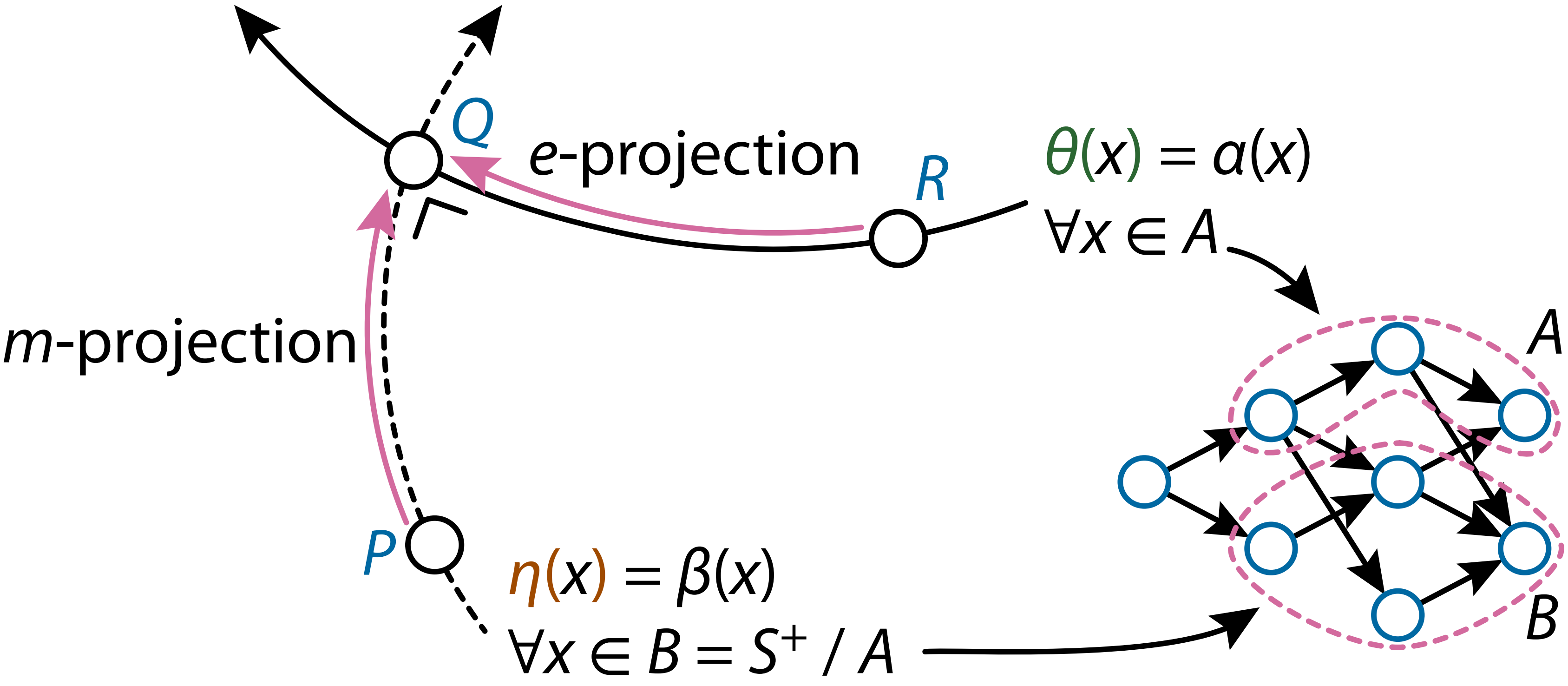
- Möbius function $\mu : S \times S \to \mathbb{Z}$

$$\mu(x, y) = \begin{cases} 1 & \text{if } x = y, \\ -\sum_{x \le s < y} \mu(x, s) & \text{if } x < y, \\ 0 & \text{otherwise} \end{cases}$$

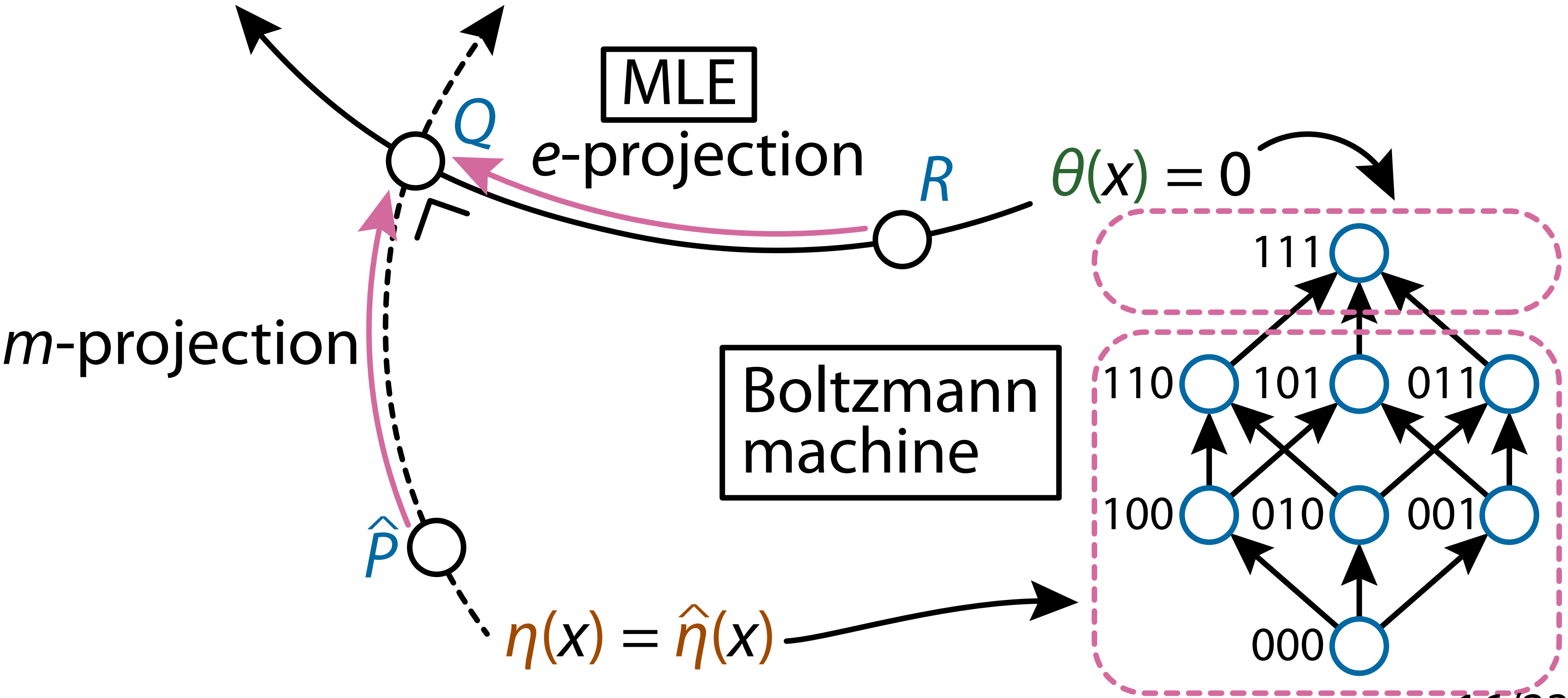
  – We have $\zeta\mu = I$, i.e.,

  $$\sum_{s \in S} \zeta(s, y)\mu(x, s) = \sum_{x \le s \le y} \mu(x, s) = \delta_{xy}$$

# *e*-Projection and *m*-Projection



# *e*-Projection and *m*-Projection



# Compute $e$-Projection by Newton's Method

- Each step of Newton's method:

$$\begin{bmatrix} \vdots \\ \eta^{(t)}_{P_\beta}(x) - \beta(x) \\ \vdots \\ \vdots \end{bmatrix} + J \begin{bmatrix} \vdots \\ \vdots \\ \theta^{(t+1)}_{P_\beta}(y) - \theta^{(t)}_{P_\beta}(y) \\ \vdots \end{bmatrix} = \mathbf{0},$$

  – $J$ is the $|\mathrm{dom}(\beta)| \times |\mathrm{dom}(\beta)|$ Jacobian matrix given as

$$J_{xy} = \frac{\partial \eta^{(t)}_{P_\beta}(x)}{\partial \theta^{(t)}_{P_\beta}(y)} = \sum_{s \in S} \zeta(x,s)\zeta(y,s){p_\beta}^{(t)}(s) - \eta^{(t)}_{P_\beta}(x)\eta^{(t)}_{P_\beta}(y)$$

  for each $x, y \in \mathrm{dom}(\beta)$

# Problem Setting

$$\begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} & p_{34} \\ p_{41} & p_{42} & p_{43} & p_{44} \end{bmatrix} \quad \boxed{\begin{matrix} r_1 \\ r_2 \\ r_3 \\ r_4 \end{matrix}}$$

$$\boxed{\begin{matrix} a_1 & a_2 & a_3 & a_4 \end{matrix}}$$

Find ***r*** and ***a***

$$r_2 a_1 p_{21} + r_2 a_2 p_{22} + r_2 a_3 p_{23} + r_2 a_4 p_{24} = 1$$

$$r_1 a_3 p_{13} + r_2 a_3 p_{23} + r_3 a_3 p_{33} + r_4 a_3 p_{43} = 1$$

# View Matrix as Poset

$$
\begin{bmatrix}
p_{11} & p_{12} & p_{13} & p_{14} \\
p_{21} & p_{22} & p_{23} & p_{24} \\
p_{31} & p_{32} & p_{33} & p_{34} \\
p_{41} & p_{42} & p_{43} & p_{44}
\end{bmatrix}
\quad \Rightarrow \quad
\begin{tikzcd}
p_{11} \arrow[r] \arrow[d] & p_{12} \arrow[r] \arrow[d] & p_{13} \arrow[r] \arrow[d] & p_{14} \arrow[d] \\
p_{21} \arrow[r] \arrow[d] & p_{22} \arrow[r] \arrow[d] & p_{23} \arrow[r] \arrow[d] & p_{24} \arrow[d] \\
p_{31} \arrow[r] \arrow[d] & p_{32} \arrow[r] \arrow[d] & p_{33} \arrow[r] \arrow[d] & p_{34} \arrow[d] \\
p_{41} \arrow[r] & p_{42} \arrow[r] & p_{43} \arrow[r] & p_{44}
\end{tikzcd}
$$

# Introduce $\theta$ and $\eta$

$$\begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} & p_{34} \\ p_{41} & p_{42} & p_{43} & p_{44} \end{bmatrix} \Rightarrow$$

$$\begin{tikzcd}
\eta_{11} \arrow[r] \arrow[d] & \eta_{12} \arrow[r] \arrow[d] & \eta_{13} \arrow[r] \arrow[d] & \eta_{14} \arrow[d] \\
\eta_{21} \arrow[r] \arrow[d] & \theta_{22} \arrow[r] \arrow[d] & \theta_{23} \arrow[r] \arrow[d] & \theta_{24} \arrow[d] \\
\eta_{31} \arrow[r] \arrow[d] & \theta_{32} \arrow[r] \arrow[d] & \theta_{33} \arrow[r] \arrow[d] & \theta_{34} \arrow[d] \\
\eta_{41} \arrow[r] & \theta_{42} \arrow[r] & \theta_{43} \arrow[r] & \theta_{44}
\end{tikzcd}$$

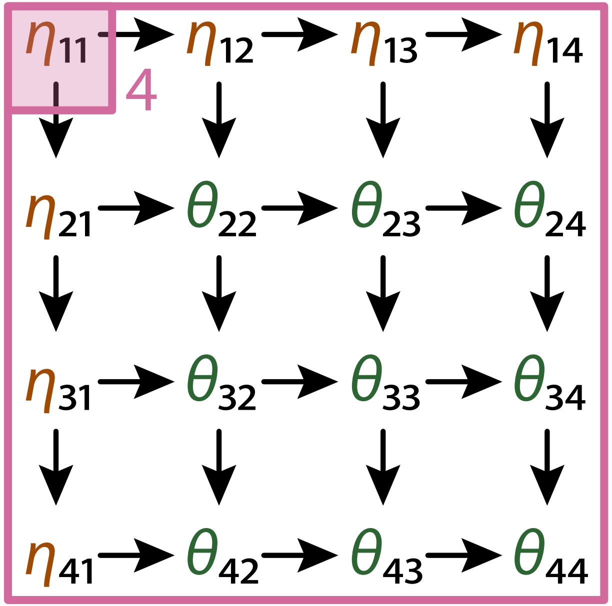
Matrix balancing is achieved if:

$\eta_{11} = 4, \eta_{21} = 3, \eta_{31} = 2, \eta_{41} = 1$

$\eta_{11} = 4, \eta_{12} = 3, \eta_{13} = 2, \eta_{41} = 1$

# Introduce $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$

$$\begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} & p_{34} \\ p_{41} & p_{42} & p_{43} & p_{44} \end{bmatrix}$$



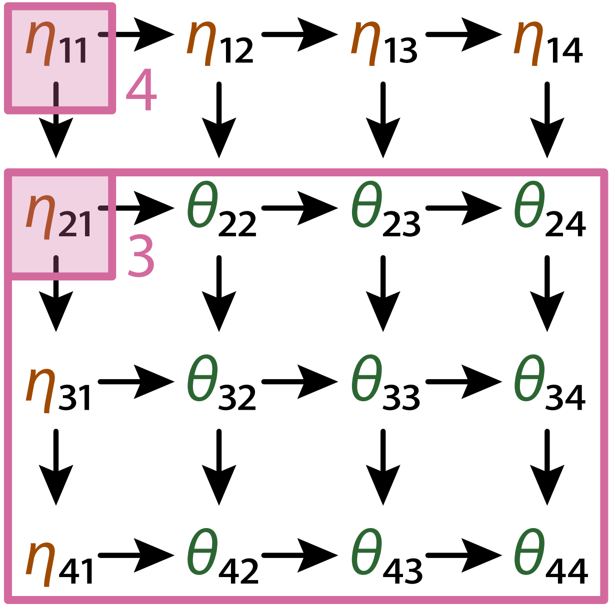

Matrix balancing is achieved if:

$\eta_{11} = 4, \eta_{21} = 3, \eta_{31} = 2, \eta_{41} = 1$

$\eta_{11} = 4, \eta_{12} = 3, \eta_{13} = 2, \eta_{41} = 1$

# Introduce $\theta$ and $\eta$

$$\begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} & p_{34} \\ p_{41} & p_{42} & p_{43} & p_{44} \end{bmatrix}$$



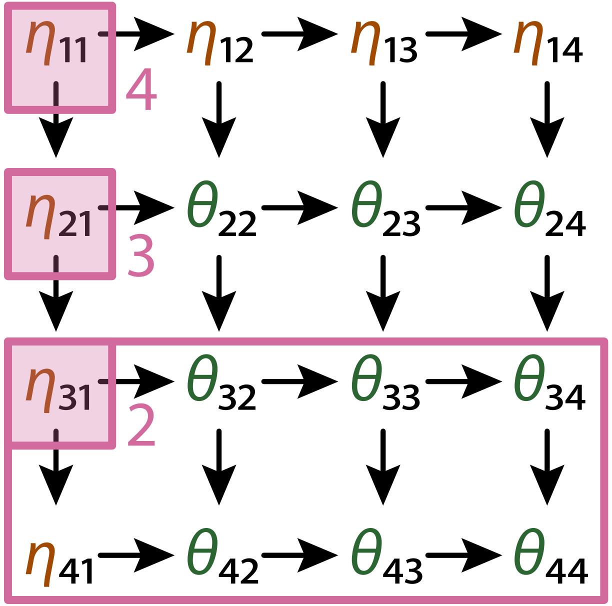

Matrix balancing is achieved if:

$\eta_{11} = 4, \eta_{21} = 3, \eta_{31} = 2, \eta_{41} = 1$

$\eta_{11} = 4, \eta_{12} = 3, \eta_{13} = 2, \eta_{41} = 1$

# Introduce $\theta$ and $\eta$

$$\begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} & p_{34} \\ p_{41} & p_{42} & p_{43} & p_{44} \end{bmatrix}$$



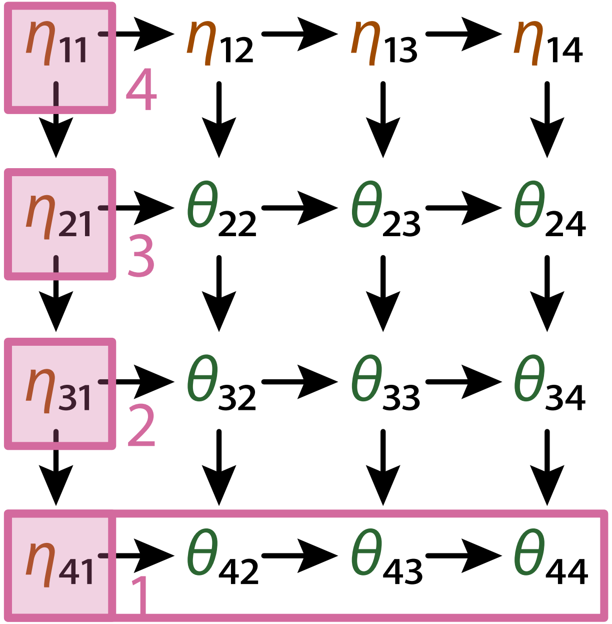

Matrix balancing is achieved if:

$\eta_{11} = 4, \eta_{21} = 3, \eta_{31} = 2, \eta_{41} = 1$

$\eta_{11} = 4, \eta_{12} = 3, \eta_{13} = 2, \eta_{41} = 1$

# Introduce $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$

$$\begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} & p_{34} \\ p_{41} & p_{42} & p_{43} & p_{44} \end{bmatrix}$$



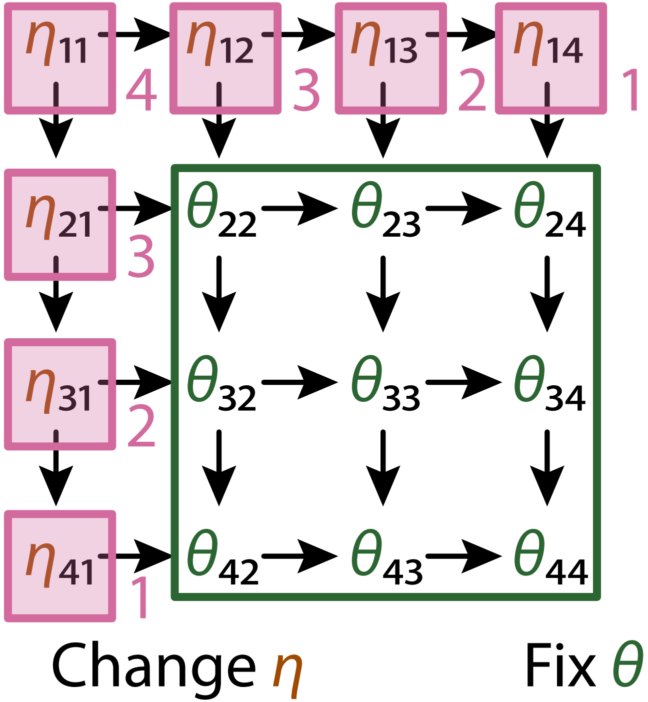

Matrix balancing is achieved if:

$\eta_{11} = 4, \eta_{21} = 3, \eta_{31} = 2, \eta_{41} = 1$

$\eta_{11} = 4, \eta_{12} = 3, \eta_{13} = 2, \eta_{41} = 1$

# $e$-Projection = Balancing

$$\begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} & p_{34} \\ p_{41} & p_{42} & p_{43} & p_{44} \end{bmatrix}$$



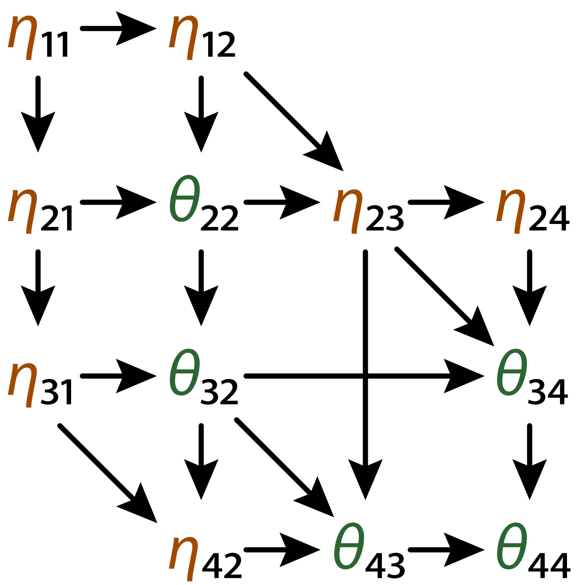

Matrix balancing is achieved if:

$\eta_{11} = 4, \eta_{21} = 3, \eta_{31} = 2, \eta_{41} = 1$

$\eta_{11} = 4, \eta_{12} = 3, \eta_{13} = 2, \eta_{41} = 1$

# Remove Zeros If Exists

$$\begin{bmatrix} p_{11} & p_{12} & 0 & 0 \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & 0 & p_{34} \\ 0 & p_{42} & p_{43} & p_{44} \end{bmatrix} \Rightarrow$$




Matrix balancing is achieved if:

$\eta_{11} = 4, \eta_{21} = 3, \eta_{31} = 2, \eta_{41} = 1$

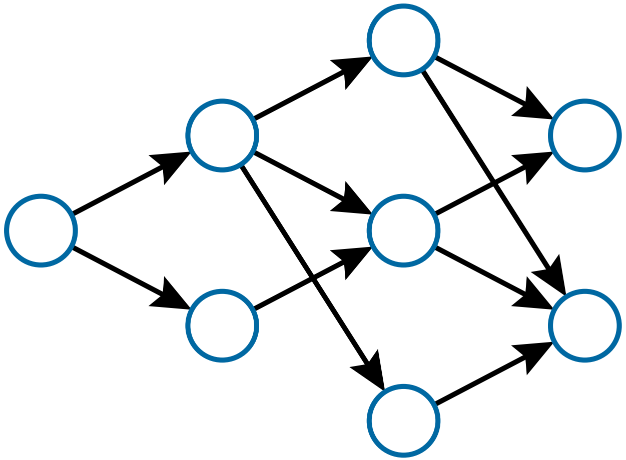
$\eta_{11} = 4, \eta_{12} = 3, \eta_{13} = 2, \eta_{41} = 1$

# Conclusion

- We have achieved efficient tensor balancing with Newton's method
- We have introduced the dually flat structure into distribution of partially ordered outcome space
- **Discrete structure + Information Geometry**
  **= original and significant data analysis methods!**

# Appendix

# Möbius Inversion



- The Möbius inversion formula [Rota (1964)]:

$$g(x) = \sum_{s \in S} \zeta(s,x) f(s) = \sum_{s \leq x} f(s)$$

$$\Leftrightarrow f(x) = \sum_{s \in S} \mu(s,x) g(s),$$

# Möbius Function Is Generalization of Inclusion-Exclusion Principle

- For sets $A, B, C$,

  $$|A \cup B \cup C| = |A| + |B| + |C| - |A \cap B| - |B \cap C| - |A \cap C| + |A \cap B \cap C|$$

- In general, for $A_1, A_2, \ldots, A_n$,

  $$\left|\bigcup_i A_i\right| = \sum_{J \subseteq \{1,\ldots,n\}, J \neq \varnothing} (-1)^{|J|-1} \left|\bigcap_{j \in J} A_j\right|$$

- The Möbius function $\mu$ is the generalization of "$(-1)^{|J|-1}$"

# *m*-Projection

- Submanifold by $\beta$: $\boldsymbol{\mathcal{S}}(\beta) = \{P \in \boldsymbol{\mathcal{S}} \mid \theta_P(x) = \beta(x),\ \forall x \in \mathrm{dom}(\beta)\}$

- *m*-projection of $P \in \boldsymbol{\mathcal{S}}$ onto $\boldsymbol{\mathcal{S}}(\beta)$ is $P_\beta \in \boldsymbol{\mathcal{S}}(\beta)$ s.t.

$$\begin{cases} \theta_{P_\beta}(x) = \beta(x) & \text{if } x \in \mathrm{dom}(\beta), \\ \eta_{P_\beta}(x) = \eta_P(x) & \text{if } x \in (S \setminus \{\perp\}) \setminus \mathrm{dom}(\beta) \end{cases}$$

  - This is the minimizer of the KL divergence from $P$ to $\boldsymbol{\mathcal{S}}(\beta)$:

    $P_\beta = \mathrm{argmin}_{Q \in \boldsymbol{\mathcal{S}}(\beta)} D_{\mathrm{KL}}[P, Q]$
  - The projected distribution $P_\beta$ always uniquely exists

- Pythagorean theorem: $D_{\mathrm{KL}}[P, Q] = D_{\mathrm{KL}}[P, P_\beta] + D_{\mathrm{KL}}[P_\beta, Q]$
  for all $Q \in \boldsymbol{\mathcal{S}}(\beta)$

A-3/A-6

# *e*-Projection

- Submanifold by $\beta$: $\mathcal{S}(\beta) = \{P \in \mathcal{S} \mid \eta_P(x) = \beta(x),\ \forall x \in \mathrm{dom}(\beta)\}$

- *e*-projection of $P \in \mathcal{S}$ onto $\mathcal{S}(\beta)$ is $P_\beta \in \mathcal{S}(\beta)$ s.t.

$$\begin{cases} \theta_{P_\beta}(x) = \theta_P(x) & \text{if } x \in (S \setminus \{\bot\}) \setminus \mathrm{dom}(\beta), \\ \eta_{P_\beta}(x) = \beta(x) & \text{if } x \in \mathrm{dom}(\beta) \end{cases}$$

  - This is the minimizer of the KL divergence from $P$ to $\mathcal{S}(\beta)$:

    $$P_\beta = \mathrm{argmin}_{Q \in \mathcal{S}(\beta)} D_{\mathrm{KL}}[P, Q]$$

  - The projected distribution $P_\beta$ always uniquely exists

- Pythagorean theorem: $D_{\mathrm{KL}}[P, Q] = D_{\mathrm{KL}}[P, P_\beta] + D_{\mathrm{KL}}[P_\beta, Q]$ for all $Q \in \mathcal{S}(\beta)$

A-4/A-6

# Computation of *e*-Projection

- Given $P$ and $\beta$, we compute $P_\beta$ such that

$$\begin{cases} \theta_{P_\beta}(x) = \theta_P(x) & \text{if } x \in (S \setminus \{\perp\}) \setminus \mathrm{dom}(\beta), \\ \eta_{P_\beta}(x) = \beta(x) & \text{if } x \in \mathrm{dom}(\beta) \end{cases}$$

- Initialize with $P_\beta^{(0)} = P$ and, at each step $t$, update $\eta_{P_\beta}^{(t)}(x)$ for $x \in \mathrm{dom}(\beta)$
  - Since $\theta$ and $\eta$ are orthogonal, we can change $\eta_{P_\beta}^{(t)}(x)$ while fixing $\theta_{P_\beta}^{(t)}(y)$ for $y \notin \mathrm{dom}(\beta)$

# Matrix And Tensor Balancing

- Given a nonnegative matrix $P = (p_{ij}) \in \mathbb{R}_{+}^{n \times n}$, find $\boldsymbol{r}, \boldsymbol{s} \in \mathbb{R}^{n}$ s.t.
  $(RPS)\mathbf{1} = \mathbf{1}$ and $(RPS)^{T}\mathbf{1} = \mathbf{1}$, where $R = \mathrm{diag}(\boldsymbol{r})$, $S = \mathrm{diag}(\boldsymbol{s})$
- Given a tensor $P \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_N}$ with $n_1 = \cdots = n_N = n$, find $(N-1)$ order tensors $R^1, R^2, \ldots, R^N$ s.t. $\forall m \in [N]$
  $P' \times_m \mathbf{1} = \mathbf{1} \; (\in \mathbb{R}^{n_1 \times \cdots \times n_{m-1} \times n_{m+1} \times \cdots \times n_N})$
  - Each entry $p'_{i_1 i_2 \ldots i_N}$ of the balanced tensor $P'$ is given as
    $$p'_{i_1 i_2 \ldots i_N} = p_{i_1 i_2 \ldots i_N} \prod_{m \in [N]} R^{m}_{i_1 \ldots i_{m-1} i_{m+1} \ldots i_N}$$
  - The balanced tensor $P'$ is called multistochastic