

August 6–11, 2017  
ICML 2017



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



# Tensor Balancing on Statistical Manifold

---

Mahito Sugiyama<sup>1,2</sup> Hiroyuki Nakahara<sup>3</sup> Koji Tsuda<sup>4,5,6</sup>

<sup>1</sup>NII, <sup>2</sup>JST PRESTO, <sup>3</sup>RIKEN BSI, <sup>4</sup>UTokyo, <sup>5</sup>RIKEN AIP, <sup>6</sup>NIMS

# Results

---

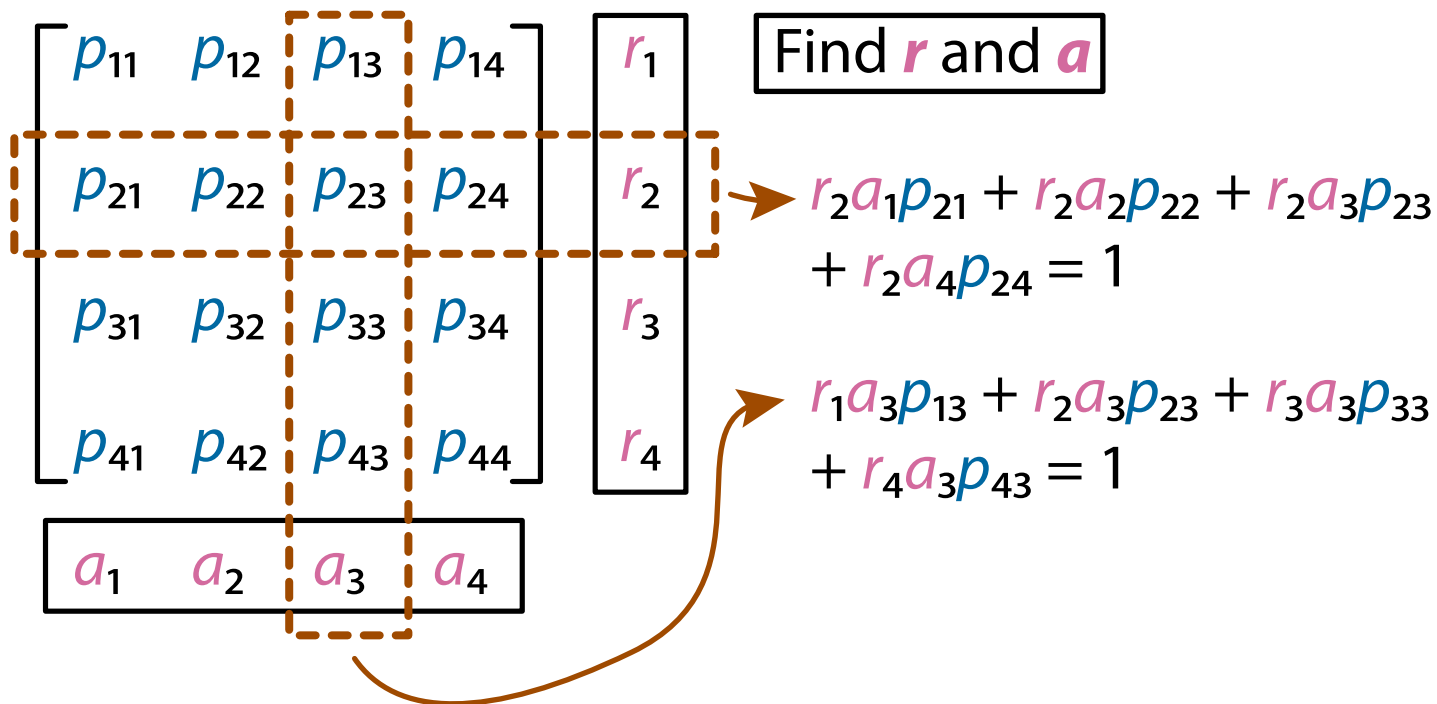
- **Balancing** of higher order (more than two) tensors is firstly (theoretically) achieved
  - We present a balancing algorithm and prove its global convergence
- A fast balancing algorithm with **quadratic convergence** using **Newton's method**
  - An existing algorithm is linear convergence
- **[Theory]** 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



# Matrix Balancing

---

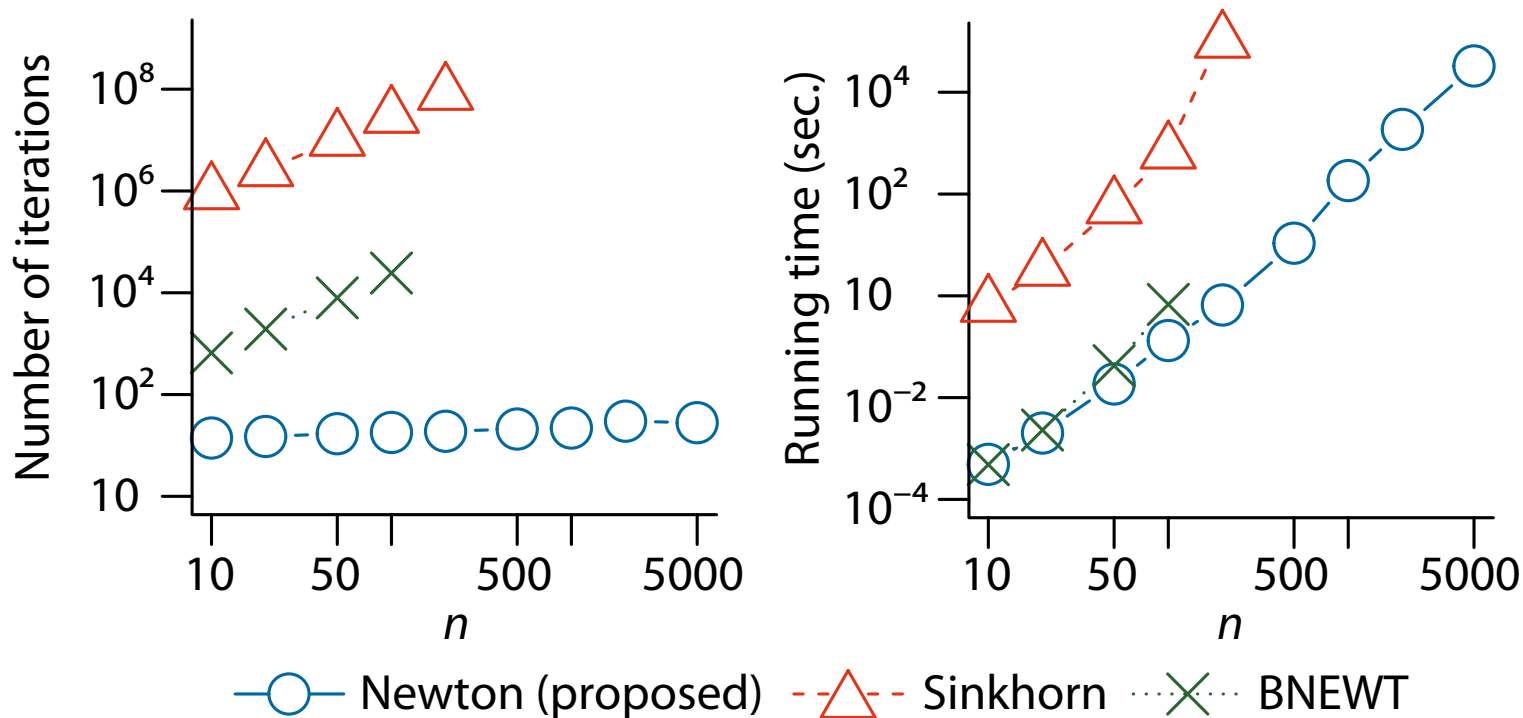
- Problem setting:

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

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

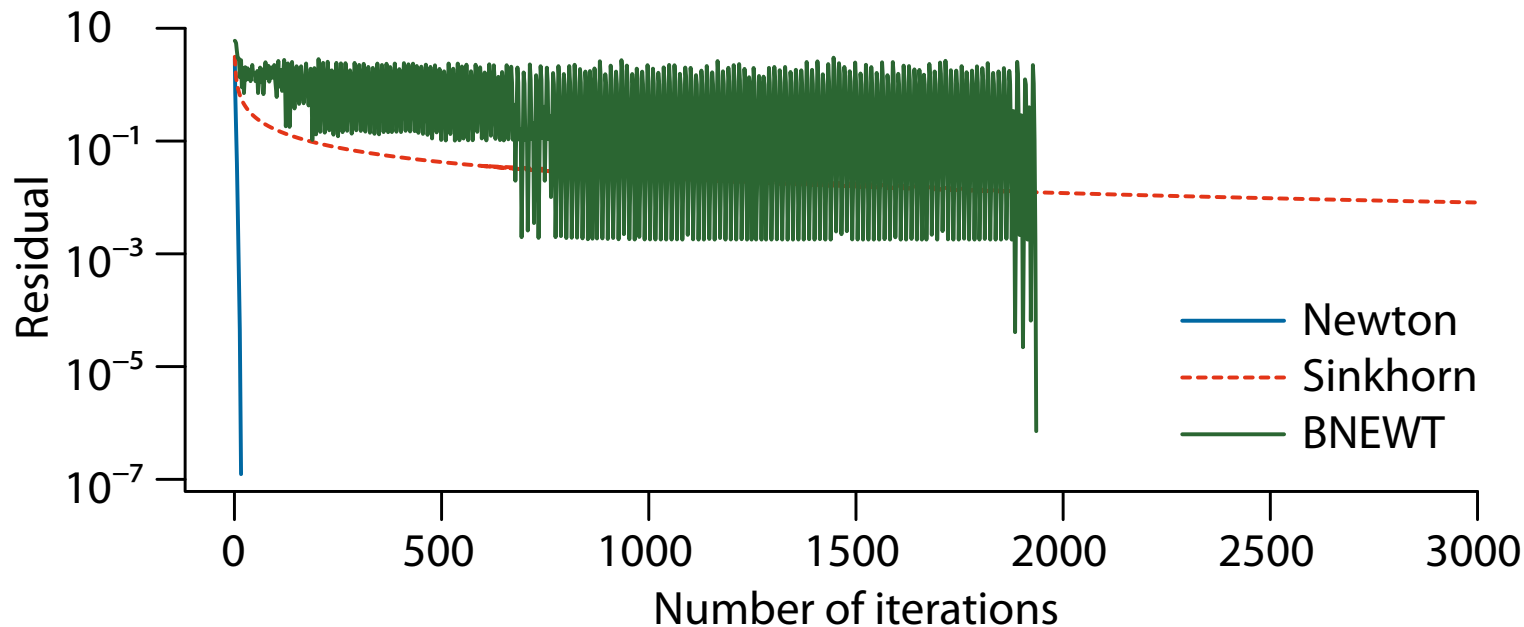
- $R = \text{diag}(r)$ ,  $S = \text{diag}(s)$
  - Each entry is given as  $p'_{ij} = p_{ij}r_i s_j$
- A fundamental process to analyze and compare matrices in a wide range of applications
  - Input-output analysis in economics, seat assignments in elections, Hi-C data analysis, Sudoku puzzle
  - Approximate Wasserstein distance

# Results on Hessenberg Matrix

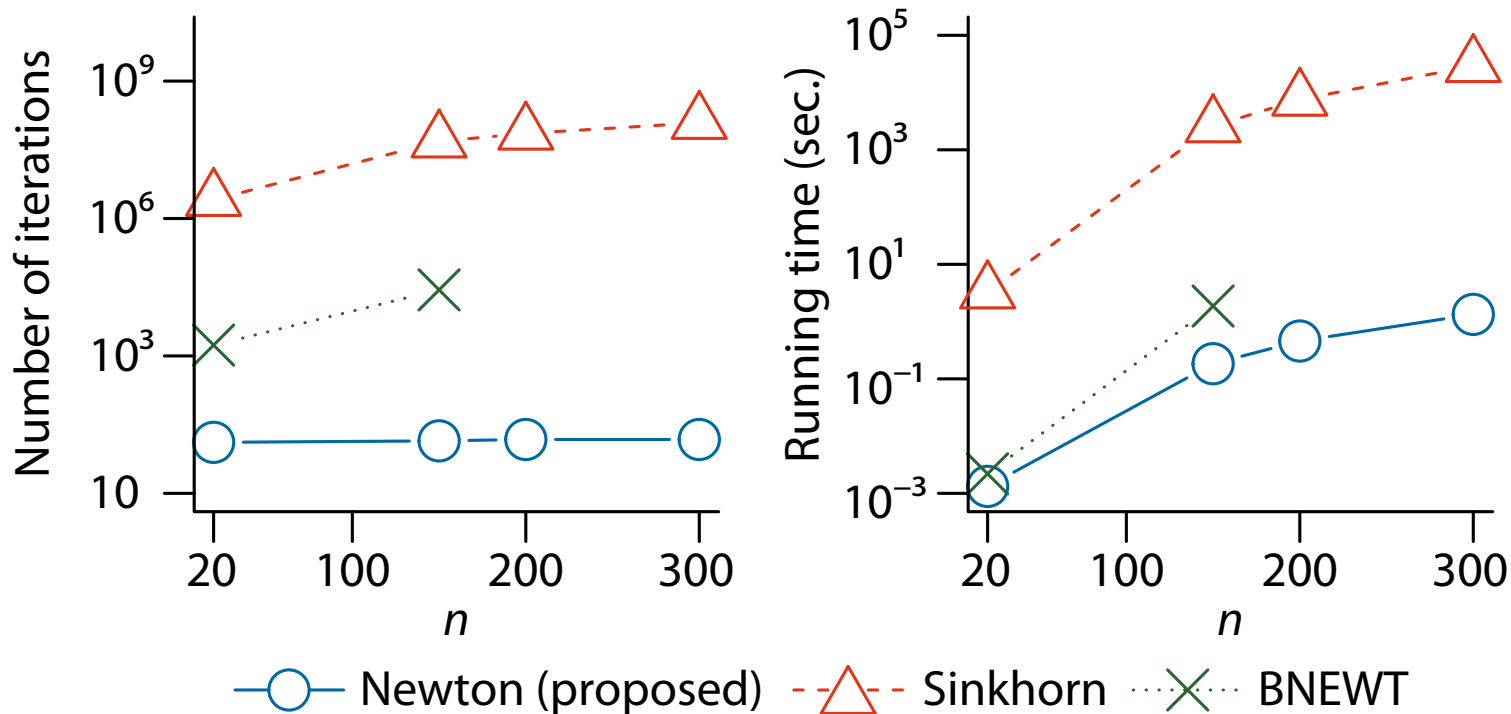


# Results on Hessenberg Matrix ( $n = 20$ )

---

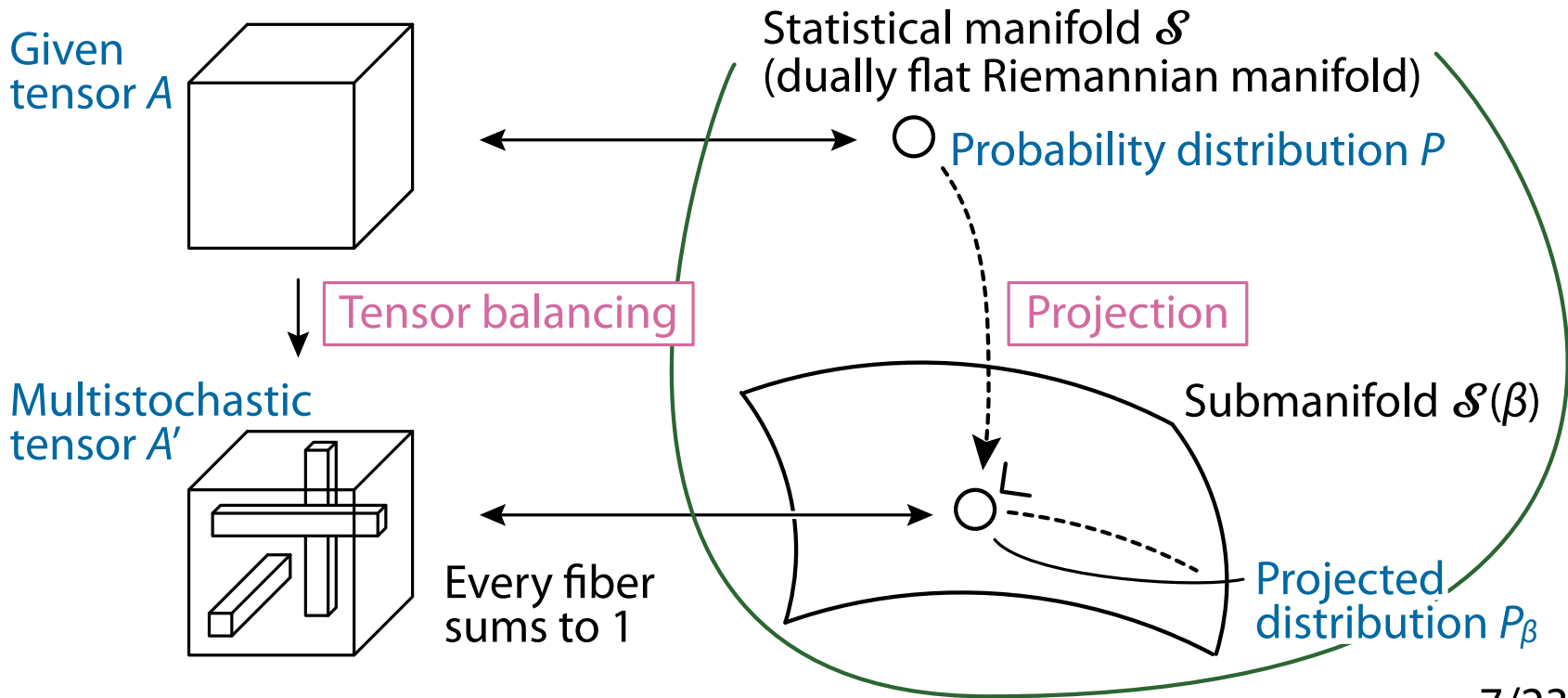


# Results on Trefethen Matrix



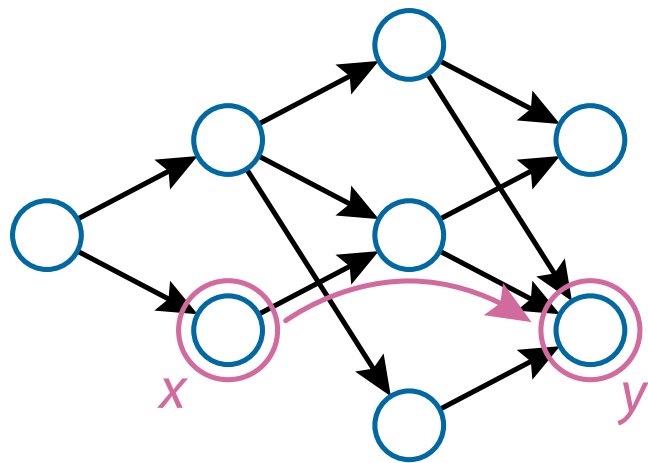


# Overview



# Partially Ordered Set

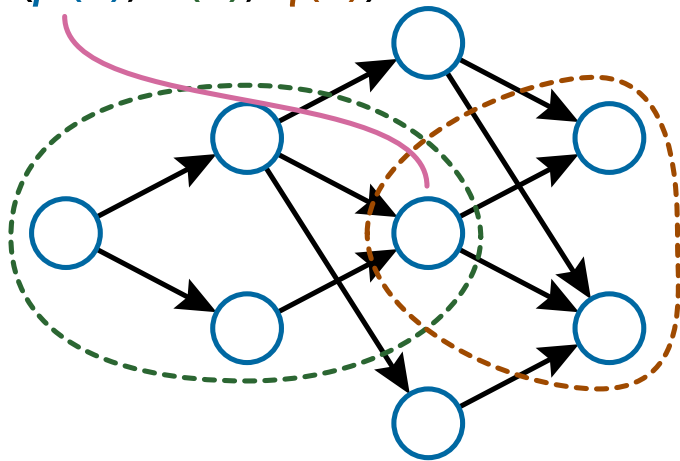
---



- 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  $x$

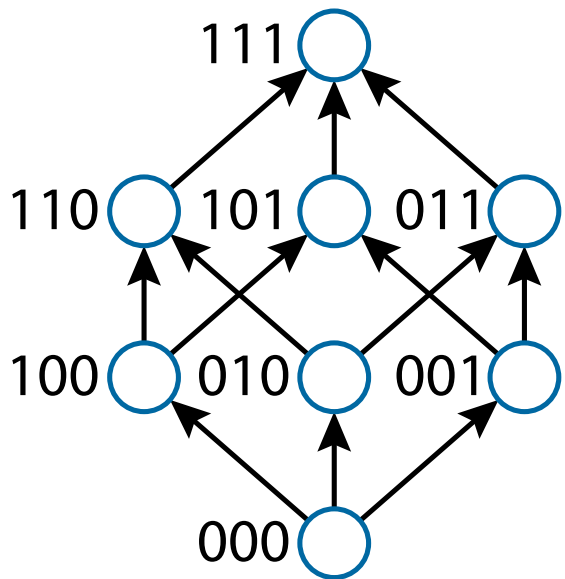
# Log-Linear Model on Poset

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



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

# Our Model Includes Binary Case



- Our model:

$$\log p(\mathbf{x}) = \sum_{s \leq \mathbf{x}} \theta(s), \quad \eta(\mathbf{x}) = \sum_{s \geq \mathbf{x}} p(s)$$

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

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

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

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

# Dually Flat Structure

---

- $\theta$  and  $\eta$  form a dual coordinate system:

$$\nabla\psi(\theta) = \eta, \quad \nabla\varphi(\eta) = \theta$$

- $\psi(\theta) = -\theta(\perp) = -\log p(\perp), \quad \varphi(\eta) = \sum_{x \in \mathcal{S}} p(x) \log p(x)$

- $\psi(\theta)$  and  $\varphi(\eta)$  are connected via the Legendre transformation:

$$\varphi(\eta) = \max_{\theta'} \left( \theta' \eta - \psi(\theta') \right), \quad \theta' \eta = \sum_{x \in \mathcal{S} \setminus \{\perp\}} \theta'(x) \eta(x)$$

- $\psi(\theta)$  and  $\varphi(\eta)$  should be convex

# Gradient and Riemannian Manifold

---

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

$$\left\{ \begin{array}{l} g_{xy}(\theta) = \frac{\partial \eta(x)}{\partial \theta(y)} = \sum_{s \in \mathcal{S}} \zeta(x, s) \zeta(y, s) p(s) - \eta(x) \eta(y) \\ g_{xy}(\eta) = \frac{\partial \theta(x)}{\partial \eta(y)} = \sum_{s \in \mathcal{S}} \mu(s, x) \mu(s, y) p(s)^{-1} \end{array} \right.$$

- $\zeta$  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,

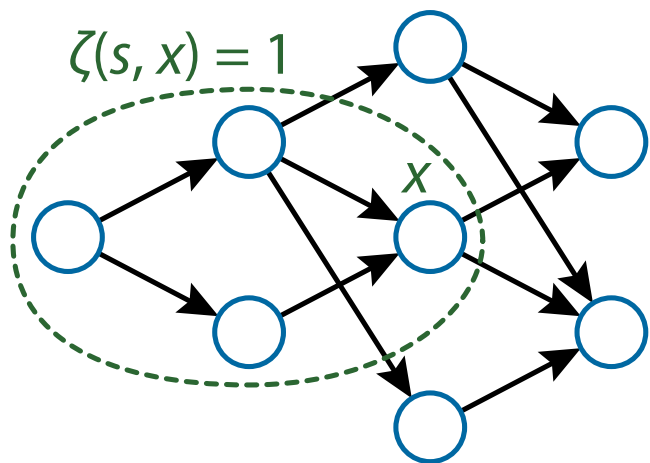
$$\mathbf{E} \left[ \frac{\partial}{\partial \theta(x)} \log p(s) \frac{\partial}{\partial \theta(y)} \log p(s) \right] = \sum_{s \in S} \zeta(x, s) \zeta(y, s) p(s) - \eta(x) \eta(y),$$

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

- $\theta$  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**  $\zeta: S \times S \rightarrow \{0, 1\}$

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

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

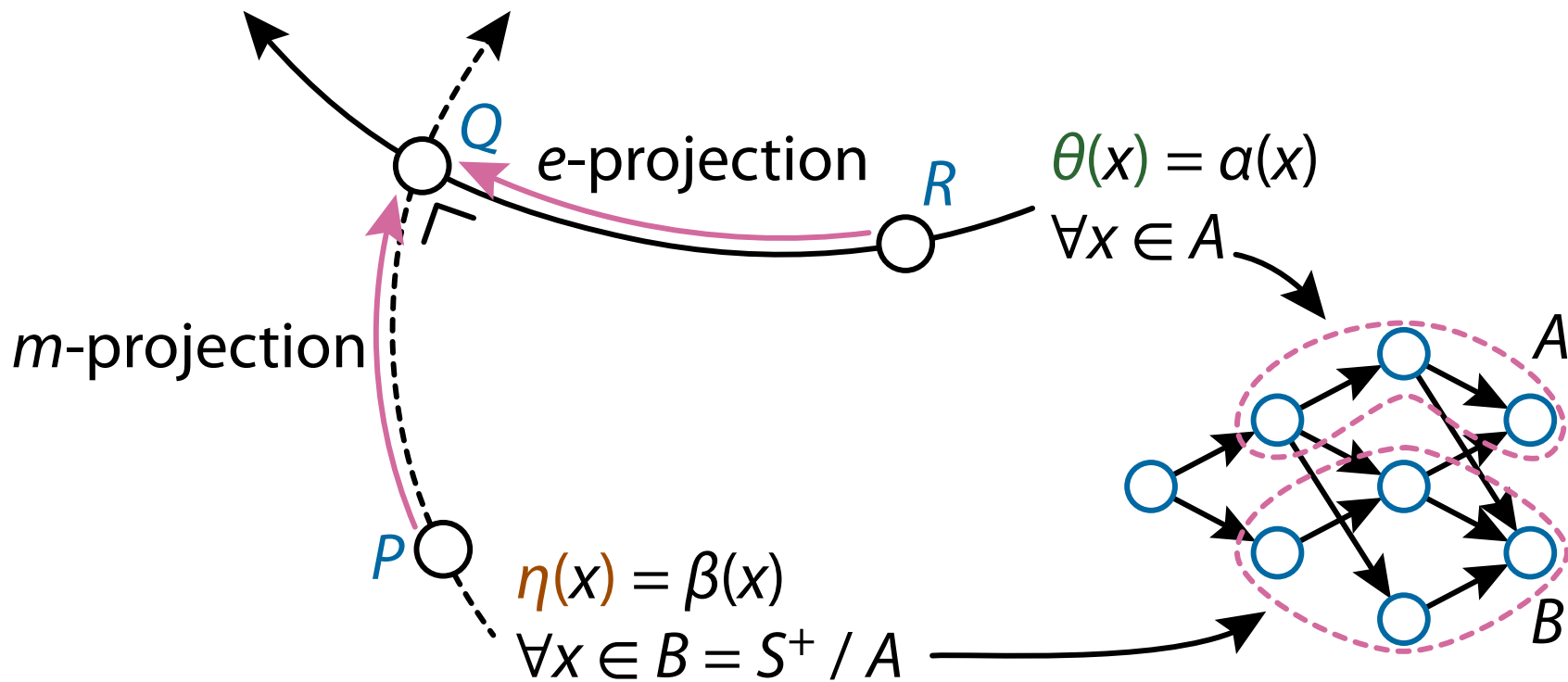
$$\mu(x, y) = \begin{cases} 1 & \text{if } x = y, \\ -\sum_{x \leq s < y} \mu(x, s) & \text{if } x < y, \\ 0 & \text{otherwise} \end{cases}$$

- We have  $\zeta\mu = I$ , i.e.,

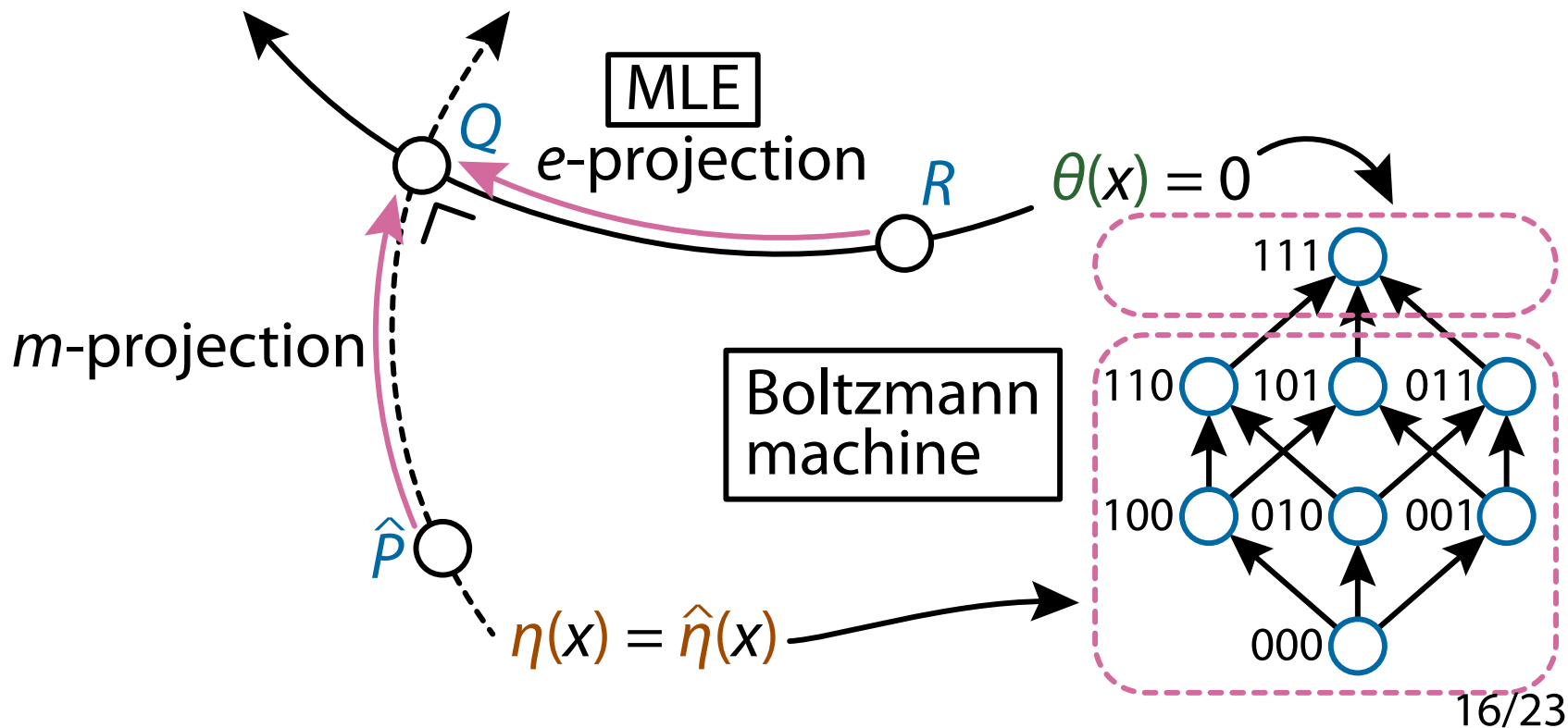
$$\sum_{s \in S} \zeta(s, y) \mu(x, s) = \sum_{x \leq s \leq y} \mu(x, s) = \delta_{xy}$$



# e-Projection and *m*-Projection



# e-Projection and $m$ -Projection



# Compute e-Projection by Newton's Method

---

- Each step of Newton's method:

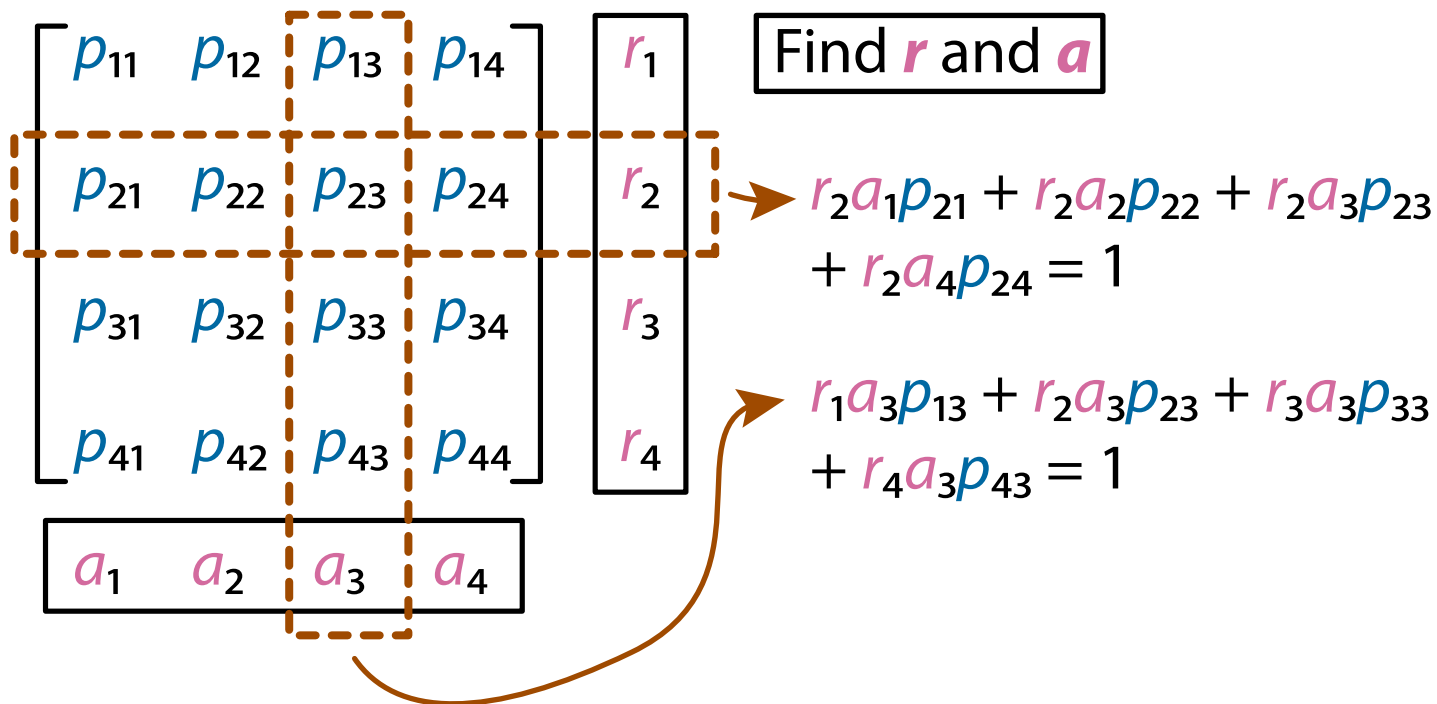
$$\begin{bmatrix} \vdots \\ \eta_{P_\beta}^{(t)}(x) - \beta(x) \\ \vdots \\ \vdots \end{bmatrix} + J \begin{bmatrix} \vdots \\ \theta_{P_\beta}^{(t+1)}(y) - \theta_{P_\beta}^{(t)}(y) \\ \vdots \\ \vdots \end{bmatrix} = \mathbf{0},$$

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

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

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

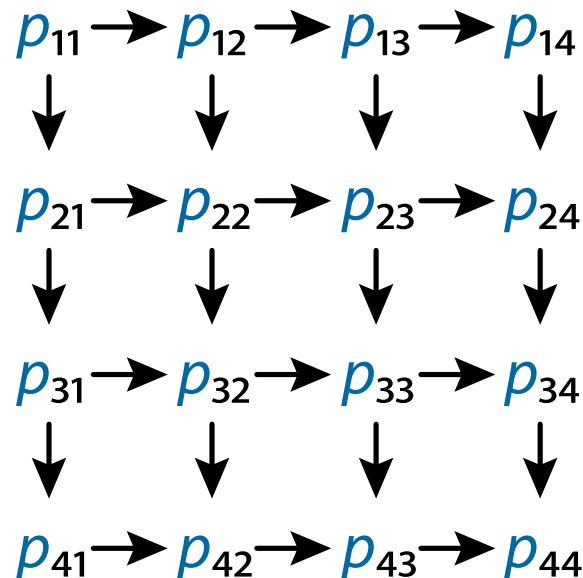
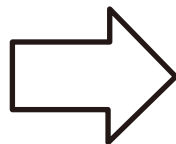
# Problem Setting



# View Matrix as Poset

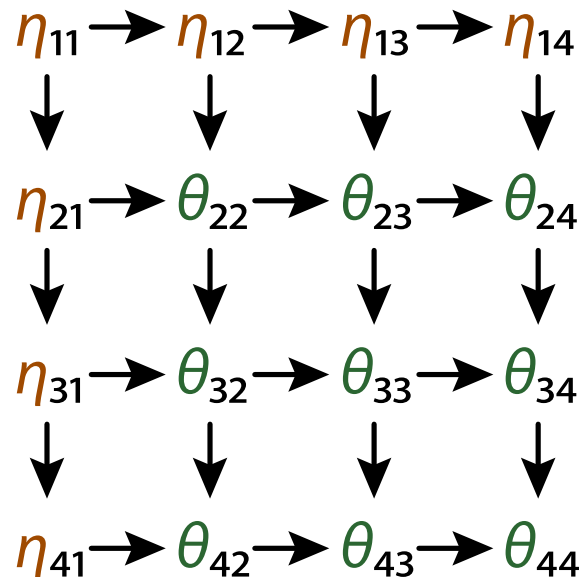
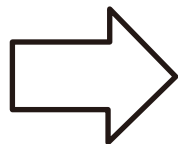
---

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



# Introduce $\theta$ and $\eta$

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



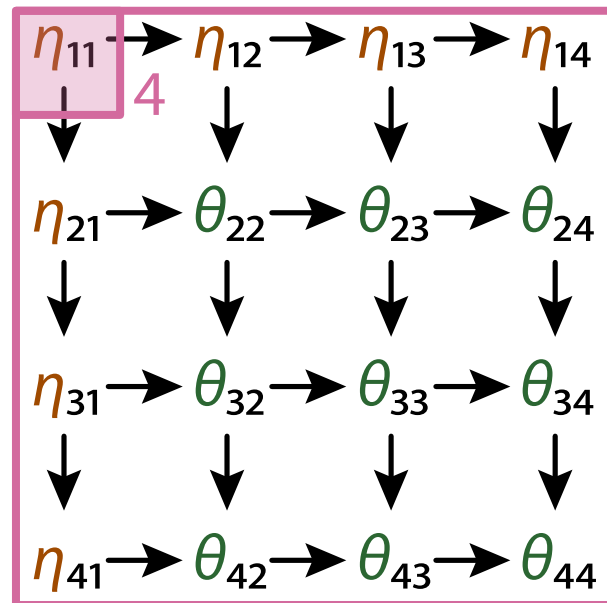
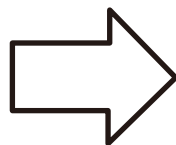
Matrix balancing is achieved if:

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

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

# Introduce $\theta$ and $\eta$

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



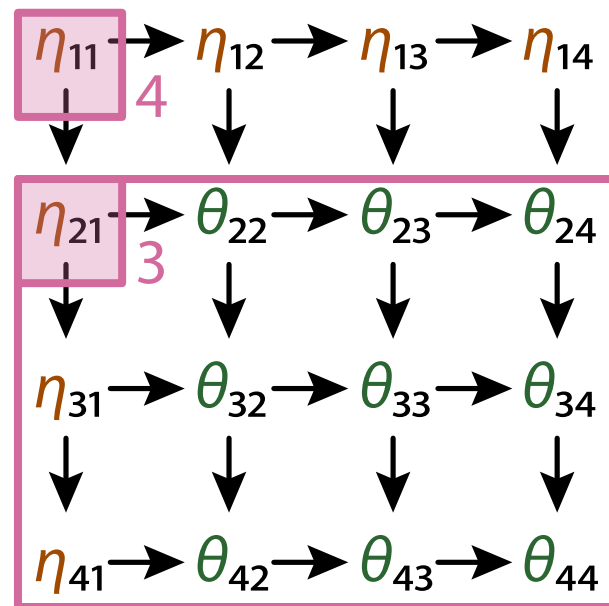
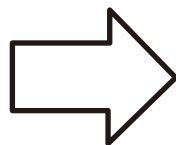
Matrix balancing is achieved if:

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

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

# Introduce $\theta$ 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:

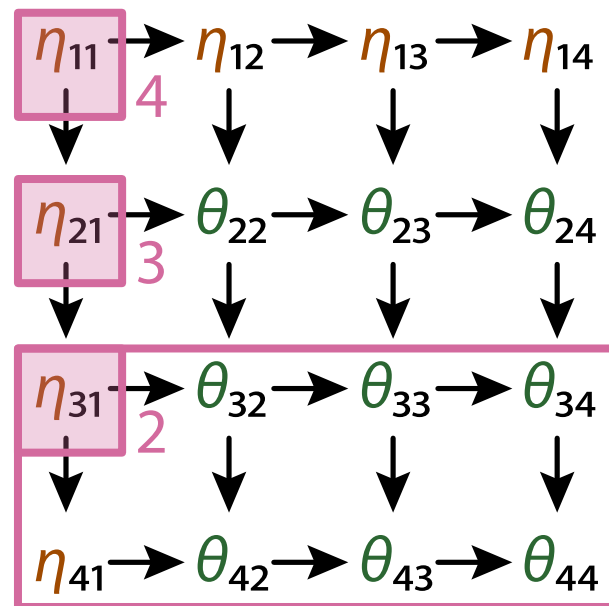
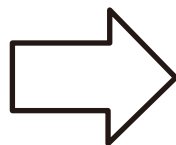
$$\eta_{11} = 4, \eta_{21} = 3, \eta_{31} = 2, \eta_{41} = 1$$

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



# Introduce $\theta$ and $\eta$

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



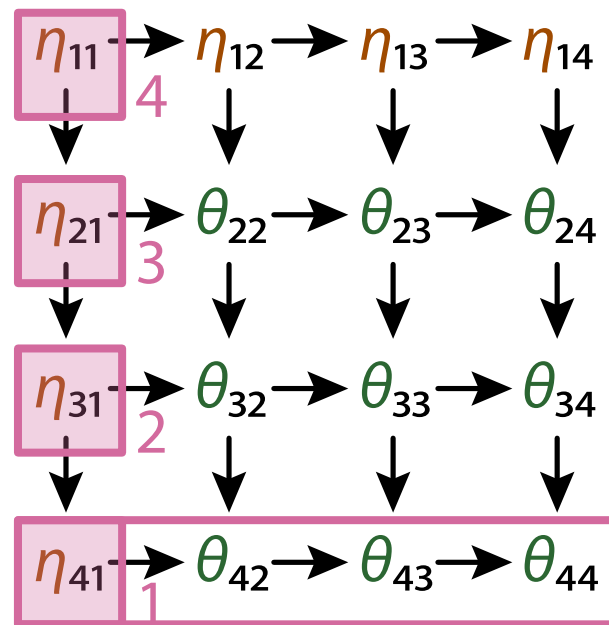
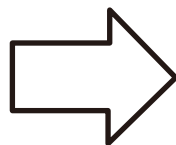
Matrix balancing is achieved if:

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

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

# Introduce $\theta$ 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:

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

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

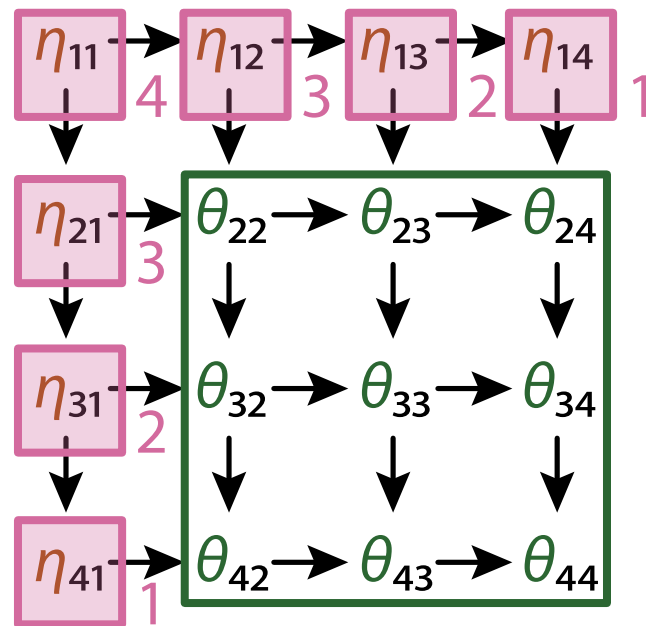
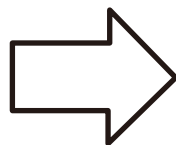
# e-Projection = Balancing

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

Matrix balancing is achieved if:

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

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

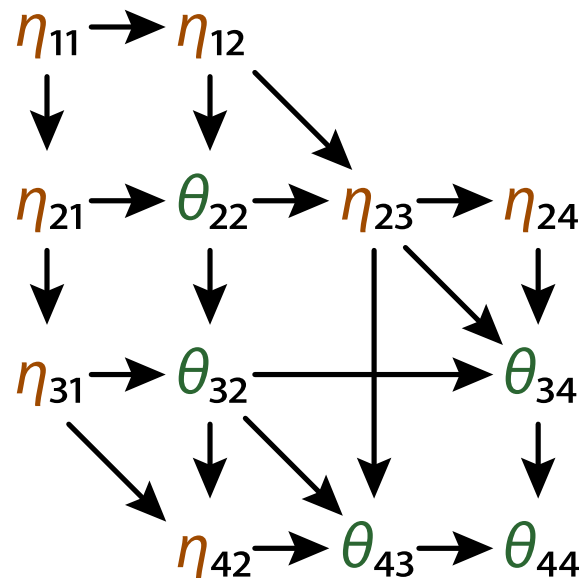
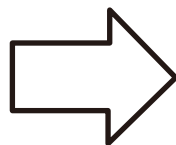


Change  $\eta$

Fix  $\theta$

# Remove Zeros If Exists

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



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

---

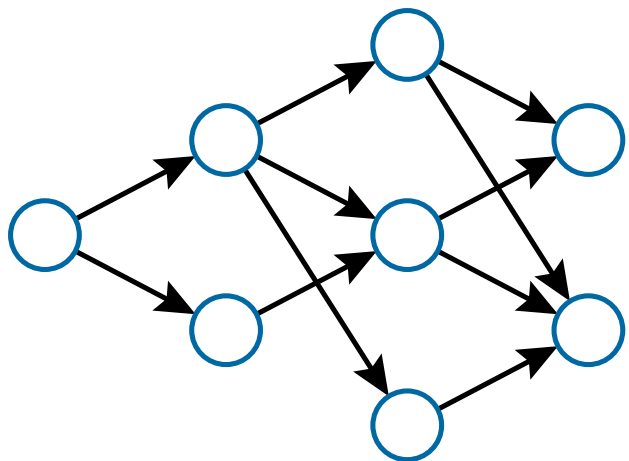
- We have achieved **efficient tensor balancing** with **Newton's method**
- We have introduced the **dually flat structure** into distribution of **partially ordered outcome space**
- **Discrete structure + Information Geometry**  
= **original and significant data analysis methods!**

# 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  $A_1, A_2, \dots, A_n$ ,

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

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



# $m$ -Projection

---

- Submanifold by  $\beta$ :  $\mathcal{S}(\beta) = \{P \in \mathcal{S} \mid \theta_P(x) = \beta(x), \forall x \in \text{dom}(\beta)\}$
- $m$ -projection of  $P \in \mathcal{S}$  onto  $\mathcal{S}(\beta)$  is  $P_\beta \in \mathcal{S}(\beta)$  s.t.  
$$\begin{cases} \theta_{P_\beta}(x) = \beta(x) & \text{if } x \in \text{dom}(\beta), \\ \eta_{P_\beta}(x) = \eta_P(x) & \text{if } x \in (S \setminus \{\perp\}) \setminus \text{dom}(\beta) \end{cases}$$
  - This is the **minimizer** of the KL divergence from  $P$  to  $\mathcal{S}(\beta)$ :  
$$P_\beta = \operatorname{argmin}_{Q \in \mathcal{S}(\beta)} D_{\text{KL}}[P, Q]$$
  - The projected distribution  $P_\beta$  **always uniquely exists**
- **Pythagorean theorem**:  $D_{\text{KL}}[P, Q] = D_{\text{KL}}[P, P_\beta] + D_{\text{KL}}[P_\beta, Q]$   
for all  $Q \in \mathcal{S}(\beta)$

# e-Projection

---

- Submanifold by  $\beta$ :  $\mathcal{S}(\beta) = \{P \in \mathcal{S} \mid \eta_P(x) = \beta(x), \forall x \in \text{dom}(\beta)\}$
- **e-projection** of  $P \in \mathcal{S}$  onto  $\mathcal{S}(\beta)$  is  $P_\beta \in \mathcal{S}(\beta)$  s.t.  
$$\begin{cases} \theta_{P_\beta}(x) = \theta_P(x) & \text{if } x \in (S \setminus \{\perp\}) \setminus \text{dom}(\beta), \\ \eta_{P_\beta}(x) = \beta(x) & \text{if } x \in \text{dom}(\beta) \end{cases}$$
  - This is the **minimizer** of the KL divergence from  $P$  to  $\mathcal{S}(\beta)$ :  
$$P_\beta = \operatorname{argmin}_{Q \in \mathcal{S}(\beta)} D_{\text{KL}}[P, Q]$$
  - The projected distribution  $P_\beta$  **always uniquely exists**
- **Pythagorean theorem**:  $D_{\text{KL}}[P, Q] = D_{\text{KL}}[P, P_\beta] + D_{\text{KL}}[P_\beta, Q]$   
for all  $Q \in \mathcal{S}(\beta)$

# Computation of e-Projection

---

- Given  $P$  and  $\beta$ , we compute  $P_\beta$  such that
$$\begin{cases} \theta_{P_\beta}(x) = \theta_P(x) & \text{if } x \in (S \setminus \{\perp\}) \setminus \text{dom}(\beta), \\ \eta_{P_\beta}(x) = \beta(x) & \text{if } x \in \text{dom}(\beta) \end{cases}$$
- Initialize with  $P_\beta^{(0)} = P$  and, at each step  $t$ ,  
update  $\eta_{P_\beta}^{(t)}(x)$  for  $x \in \text{dom}(\beta)$ 
  - Since  $\theta$  and  $\eta$  are **orthogonal**, we can change  $\eta_{P_\beta}^{(t)}(x)$   
while fixing  $\theta_{P_\beta}^{(t)}(y)$  for  $y \notin \text{dom}(\beta)$

# Matrix And Tensor Balancing

---

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