July 18, 2019

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

National Institute of Informatics

PRESTO

SAKIGAKE

Machine Learning with Information Geometry

Mahito Sugiyama

Summary

- We study machine learning with its applications
  - We focus on the relationship between computational processes, discrete structures, and machine learning models
- Ongoing topics
  - Machine Learning with Information Geometry
  - Machine Learning with Discrete Structure
  - Significant Pattern Mining
  - Machine Learning Applications

Tensor (Matrix) Balancing

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

Tensor (Matrix) Balancing

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

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

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

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

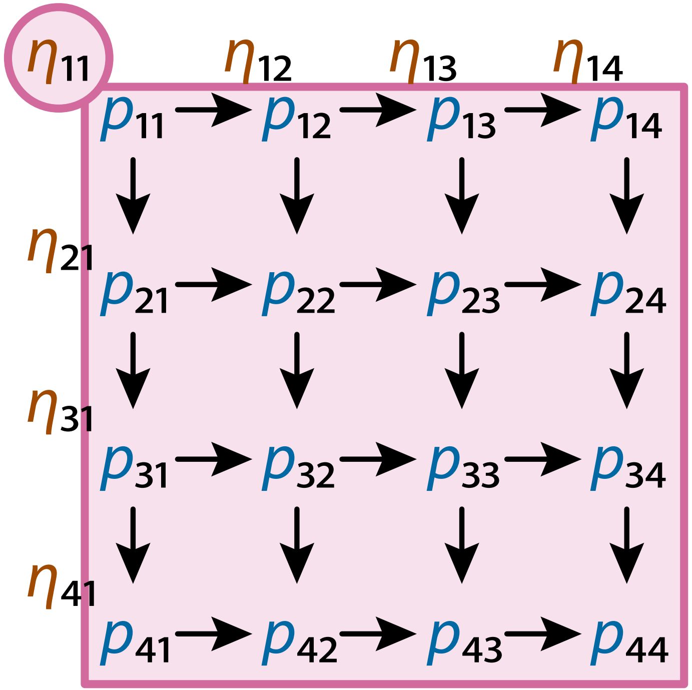
$$r_1 a_3 p_{13} + r_2 a_3 p_{23} + r_3 a_3 p_{33} + r_4 a_3 p_{43} = 1$$

Introduce η

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

Introduce η

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



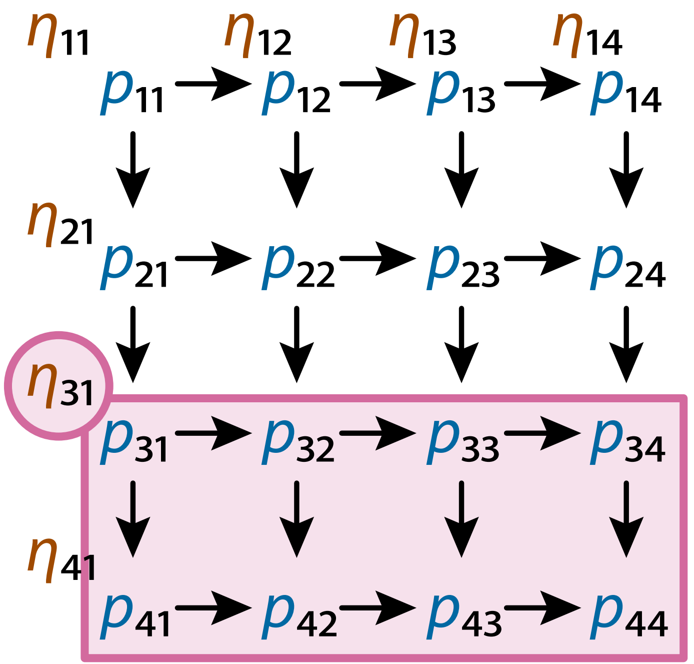
Introduce η

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

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

Introduce η

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



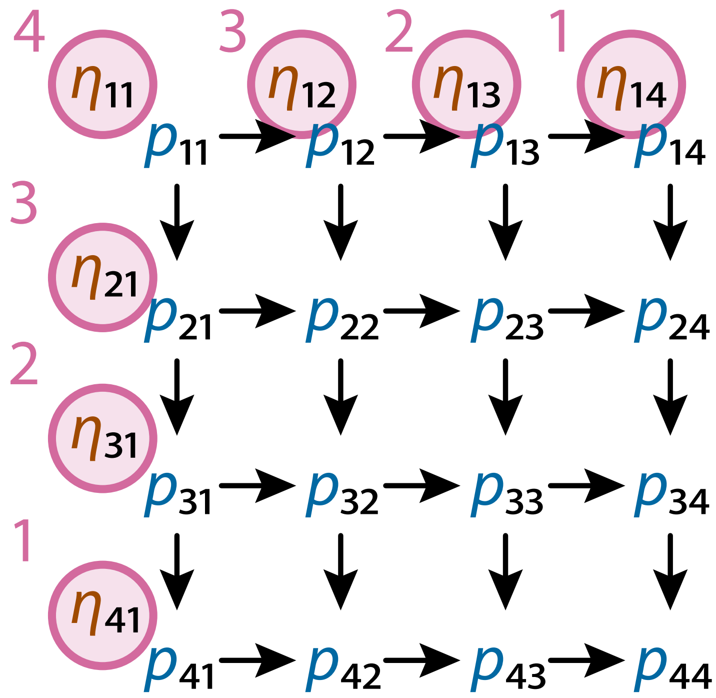
Introduce η

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

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

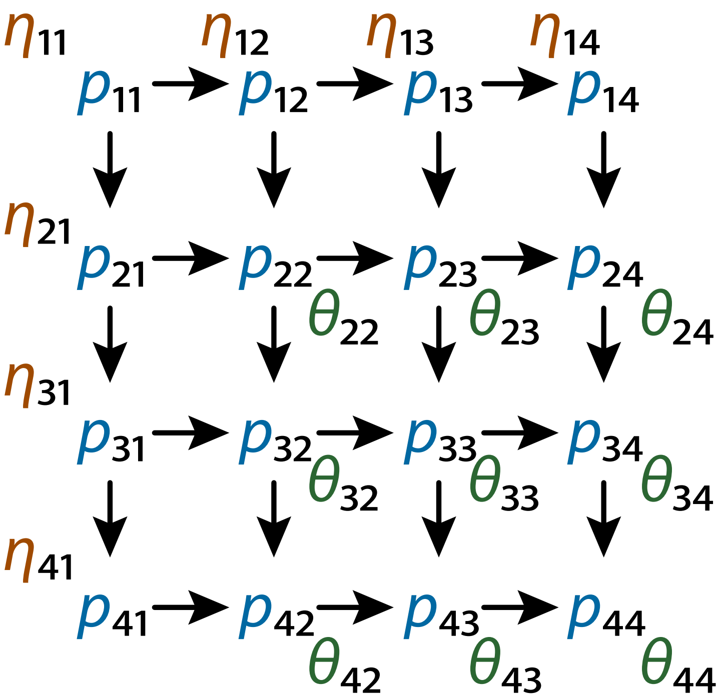
Constraints on η

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



Introduce θ

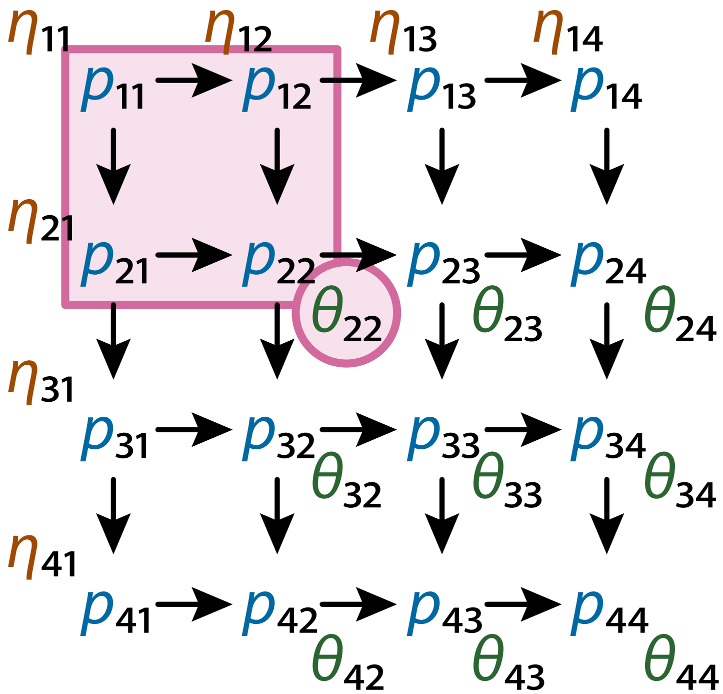
$$\begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} & p_{34} \\ p_{41} & p_{42} & p_{43} & p_{44} \end{bmatrix}$$



$$\boxed{\theta_{ij} = \log p_{ij} - \log p_{i-1j} - \log p_{ij-1} + \log p_{i-1j-1}}$$

Introduce θ

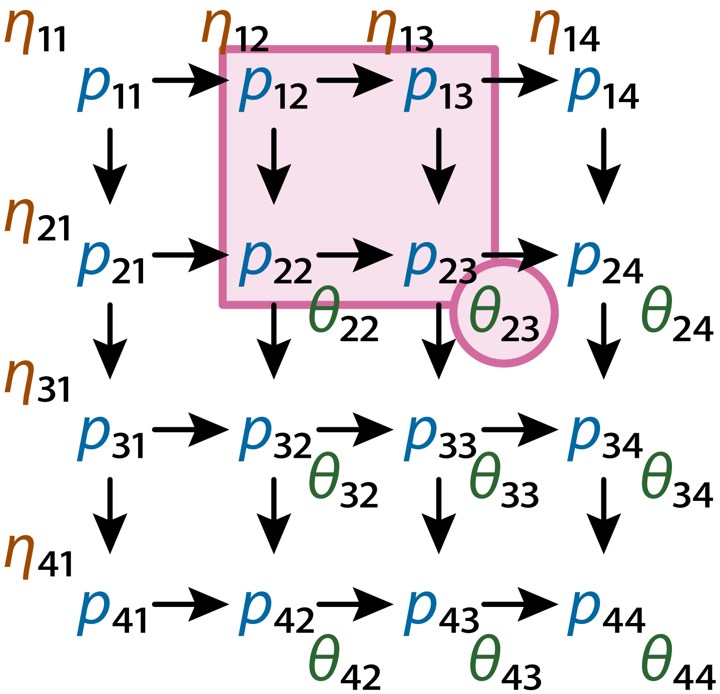
$$\begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} & p_{34} \\ p_{41} & p_{42} & p_{43} & p_{44} \end{bmatrix} \longrightarrow$$



$$\boxed{\begin{aligned}\theta_{ij} = \log p_{ij} - \log p_{i-1j} - \log p_{ij-1} \\ + \log p_{i-1j-1}\end{aligned}}$$

Introduce θ

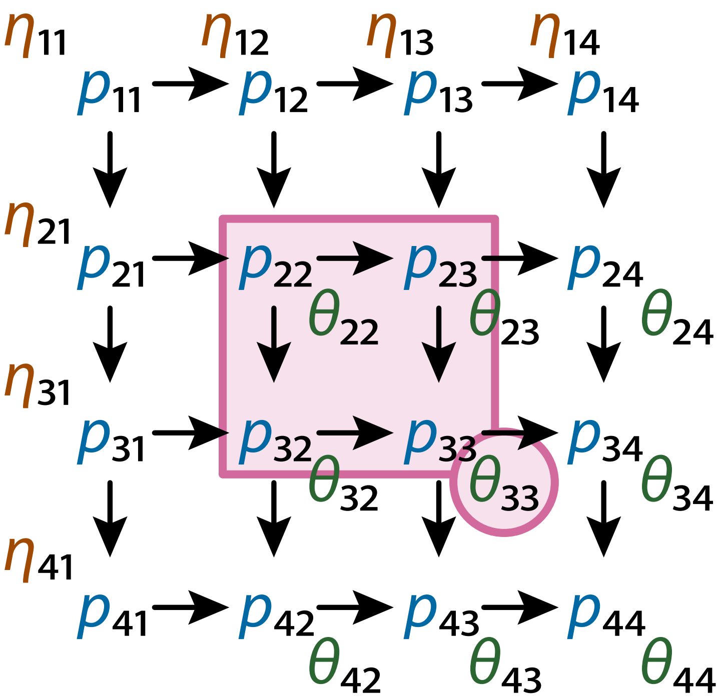
$$\begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} & p_{34} \\ p_{41} & p_{42} & p_{43} & p_{44} \end{bmatrix} \longrightarrow$$



$$\boxed{\begin{aligned}\theta_{ij} = \log p_{ij} - \log p_{i-1j} - \log p_{ij-1} \\ + \log p_{i-1j-1}\end{aligned}}$$

Introduce $\boldsymbol{\theta}$

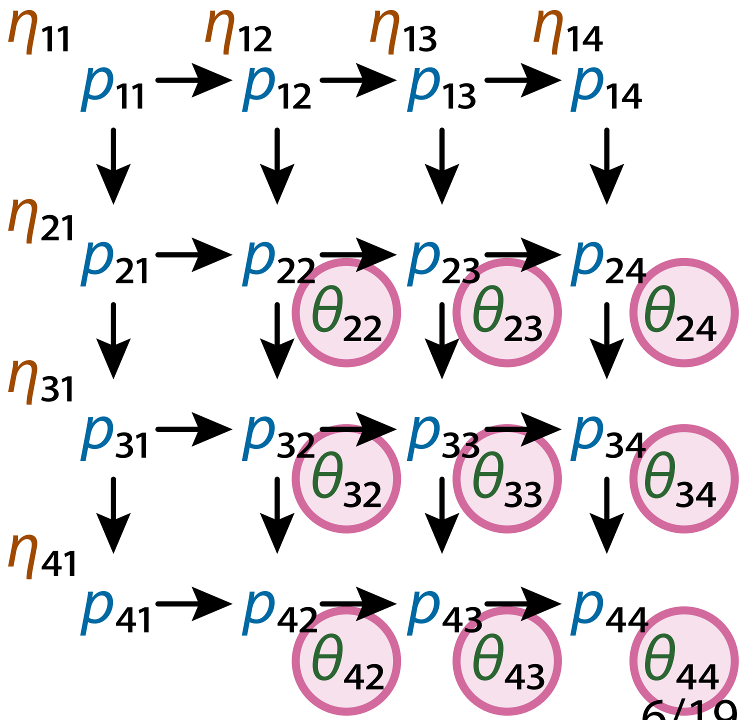
$$\begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} & p_{34} \\ p_{41} & p_{42} & p_{43} & p_{44} \end{bmatrix} \longrightarrow$$



$$\boxed{\begin{aligned}\theta_{ij} = \log p_{ij} - \log p_{i-1j} - \log p_{ij-1} \\ + \log p_{i-1j-1}\end{aligned}}$$

Introduce θ

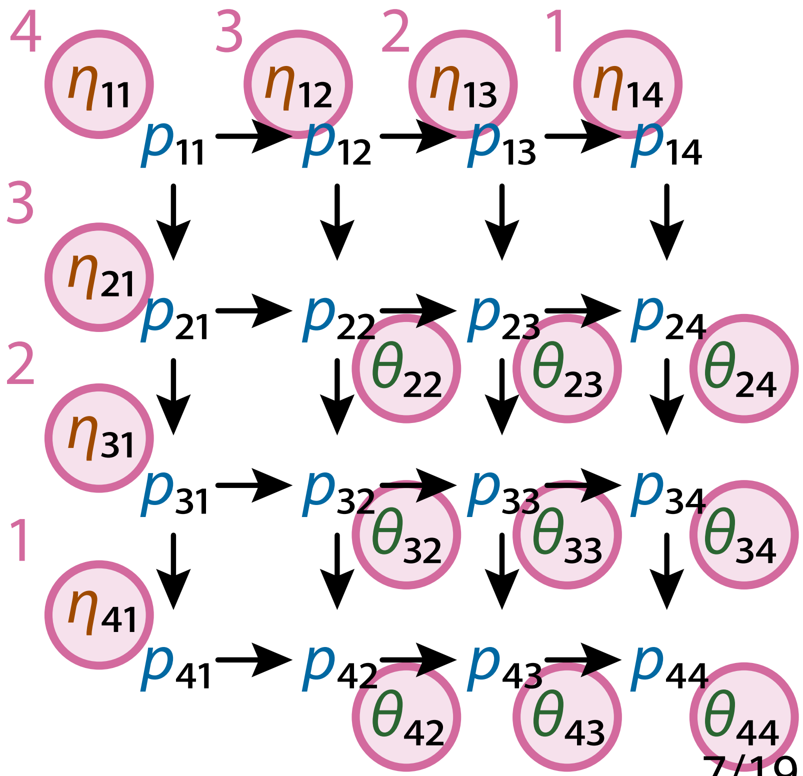
$$\begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} & p_{34} \\ p_{41} & p_{42} & p_{43} & p_{44} \end{bmatrix} \longrightarrow$$



$$\boxed{\begin{aligned}\theta_{ij} = \log p_{ij} - \log p_{i-1j} - \log p_{ij-1} \\ + \log p_{i-1j-1}\end{aligned}}$$

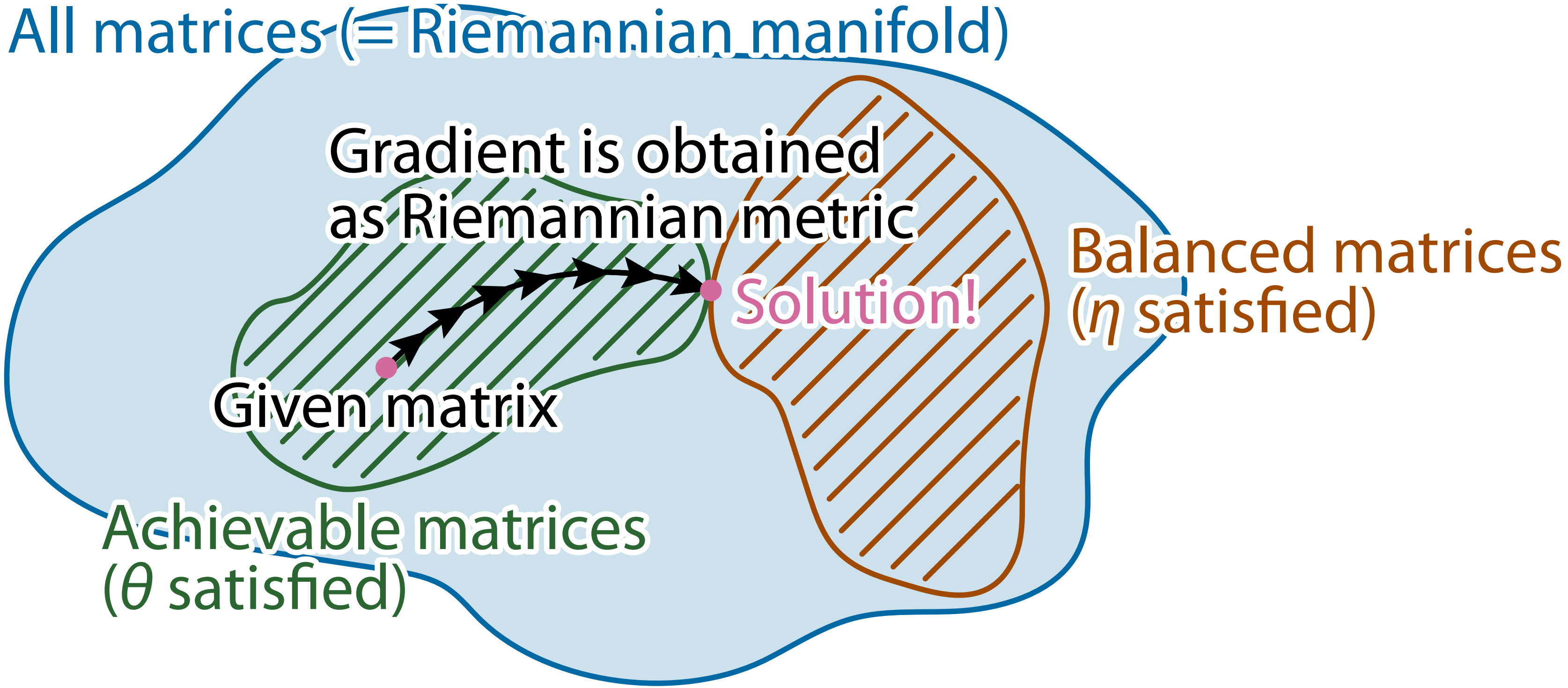
Balancing as Constraints on η and θ

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

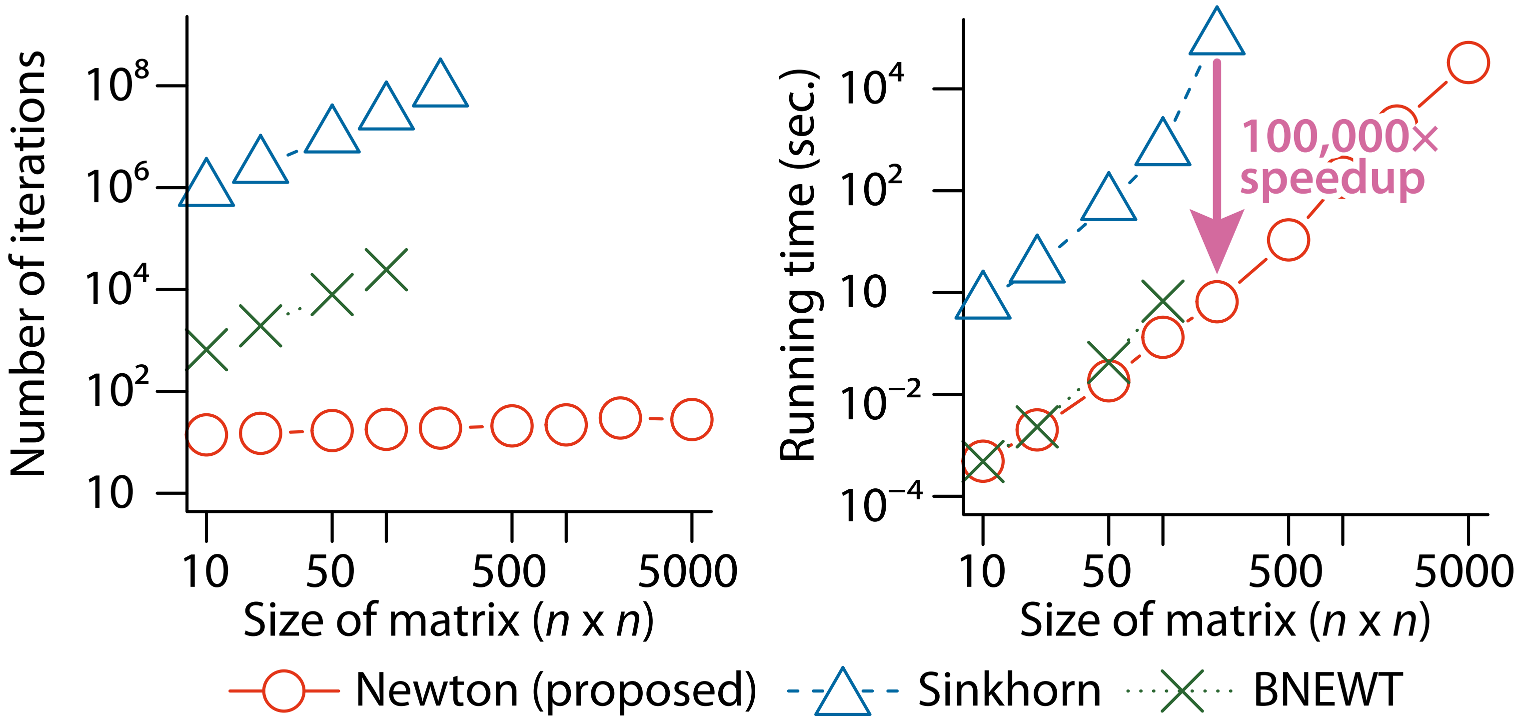


Matrix balancing is achieved iff:
$\eta_{11} = 4$, $\eta_{21} = 3$, ..., $\eta_{41} = 1$
without changing any θ_{ij}

Information Geometric View

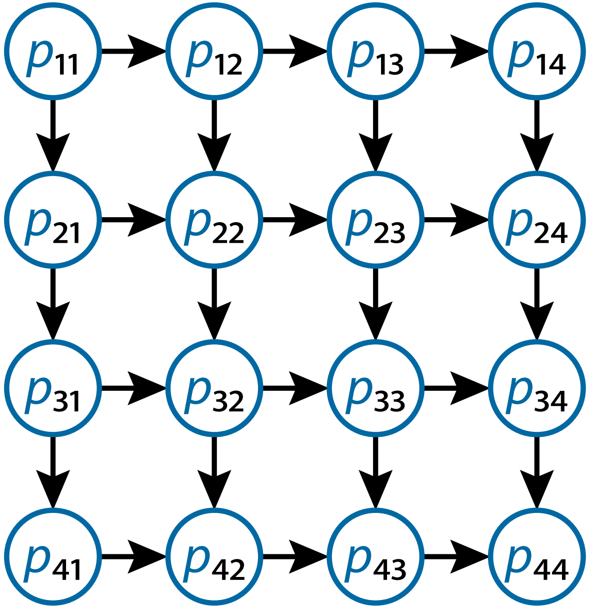


Empirical Performance

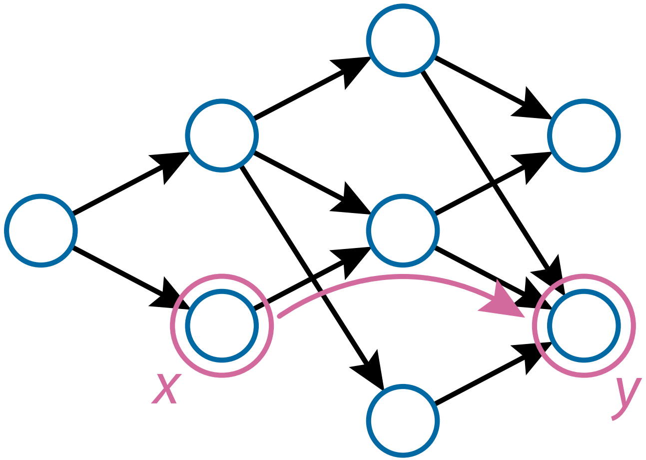


From Matrix to Poset (DAG)

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



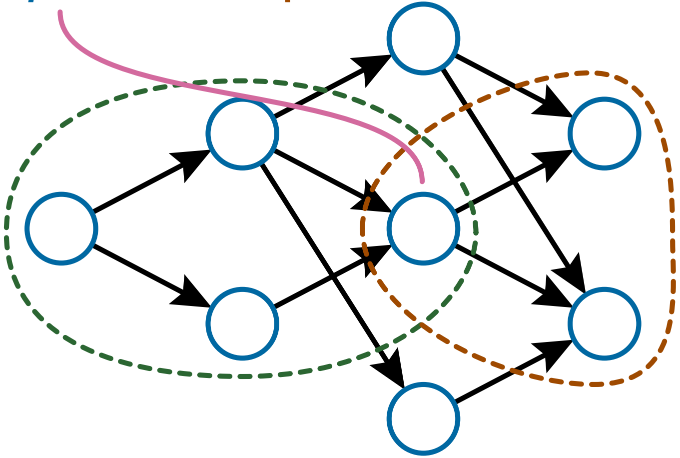
Partially Ordered Set



- 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) $\bot \in S$
- 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:
$(p(x), \theta(x), \eta(x))$



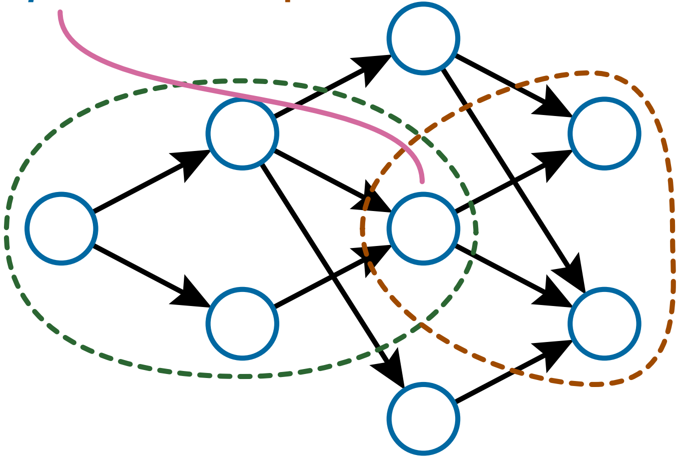
- A probability vector $p : S \to (0, 1)$ s.t. $\sum_{x \in S} p(x) = 1$
  - (Normalized) weight for each node
- We introduce $\theta : S \to \mathbb{R}$ and $\eta : S \to \mathbb{R}$ as

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

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

Log-Linear Model on Poset

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



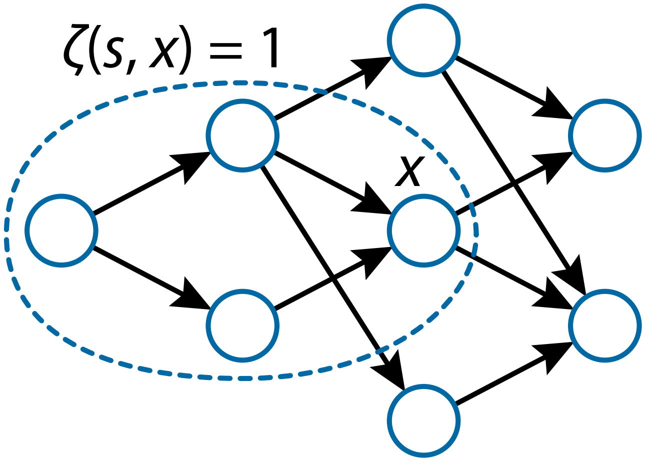
- A probability vector $p: S \to (0, 1)$ s.t. $\sum_{x \in S} p(x) = 1$
  - (Normalized) weight for each node

- We introduce $\theta: S \to \mathbb{R}$ and $\eta: S \to \mathbb{R}$ as

$$\log p(x) = \sum_{s \le x} \theta(s), \;\; \theta(x) = \sum_{s \in S} \mu(s, x) \log p(s)$$

$$\eta(x) = \sum_{s \ge x} p(s), \;\; p(x) = \sum_{s \in S} \mu(x, s) \eta(s)$$

Möbius Function



- 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 $\zeta\mu = I$, that is;
  $\sum_{s \in S} \zeta(s, y)\mu(x, s) = \sum_{x \le s \le y} \mu(x, s) = \delta_{xy}$

Möbius Function Is Generalization of Inclusion-Exclusion Principle

- For sets A, B, C,

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

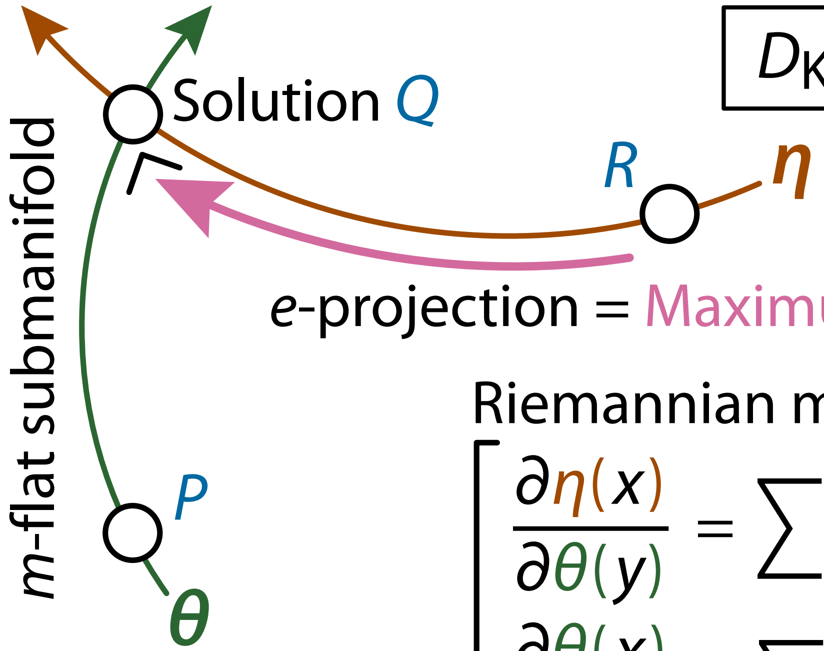
- In general, for $A_1, A_2, \ldots, A_n$,

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

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

Riemannian Manifold with Info. Geometry

$$\boxed{D_{\mathrm{KL}}[P, R] = D_{\mathrm{KL}}[P, Q] + D_{\mathrm{KL}}[Q, R]}$$



e-projection = Maximum Likelihood Estimation

Riemannian metric:

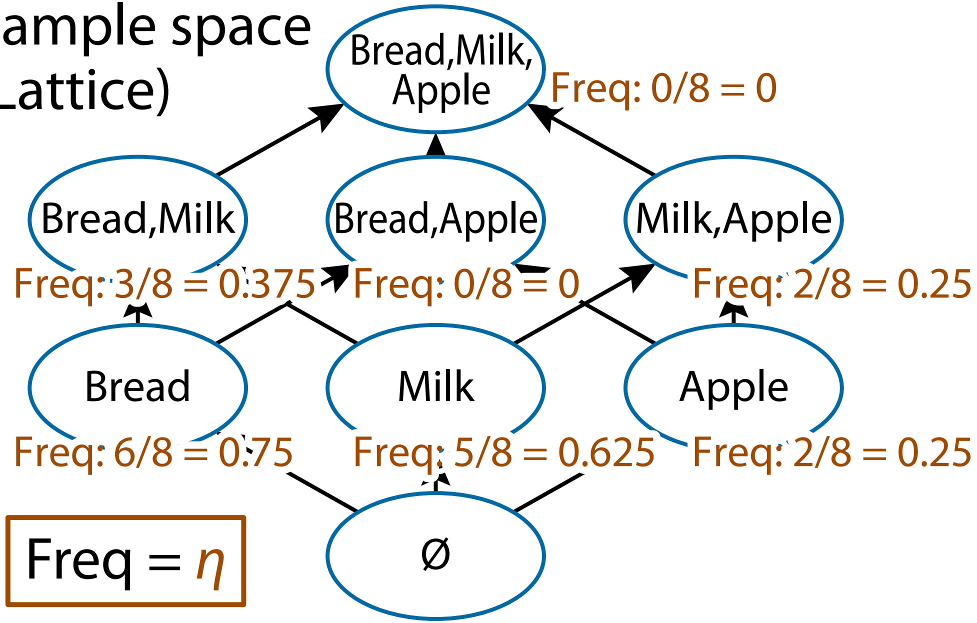
$$\left[\begin{aligned} \frac{\partial \eta(x)}{\partial \theta(y)} &= \sum\nolimits_{s \in S} \zeta(x, s)\zeta(y, s)p(s) - \eta(x)\eta(y) \\ \frac{\partial \theta(x)}{\partial \eta(y)} &= \sum\nolimits_{s \in S} \mu(s, x)\mu(s, y)p(s)^{-1} \end{aligned}\right.$$

Example of Learning Prob. Dist. (1/2)

Dataset

	Bread	Milk	Apple
ID 1	○	×	×
ID 2	○	○	×
ID 3	○	×	×
ID 4	×	○	○
ID 5	×	○	○
ID 6	○	○	×
ID 7	○	×	×
ID 8	○	○	×

Sample space (Lattice)



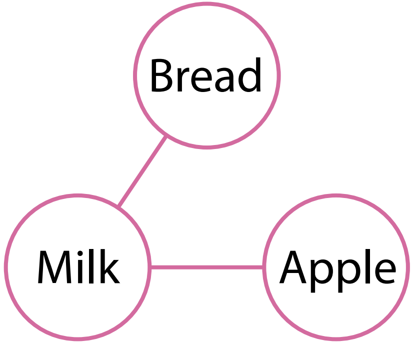
Example of Learning Prob. Dist. (2/2)

$$\log(\text{prob.}) = -10.41 + 9.43[\text{Bread}] + 8.52[\text{Milk}] - 9.84[\text{Apple}] - 9.03[\text{Bread,Milk}] + 9.43[\text{Milk,Apple}]$$

MLE

Parameter = θ

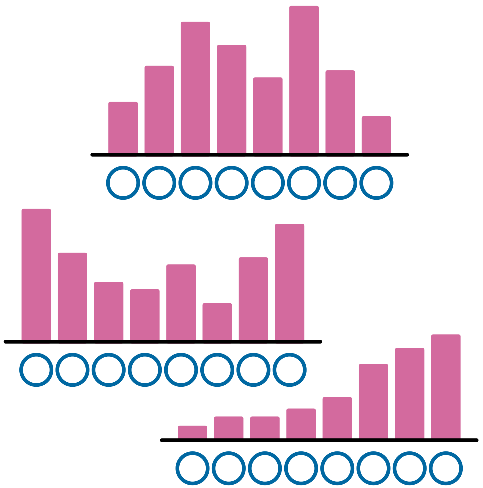
Boltzmann machine



Bread	Milk	Apple	Prob. from data	Learned prob.
×	×	×	?	0.0000300109
○	×	×	0.375	0.3749599867
×	○	×	?	0.1499903954
×	×	○	?	0.0000000016
○	○	×	0.375	0.2250096042
○	×	○	?	0.0000200043
×	○	○	0.25	0.0999895960
○	○	○	?	0.1500004008

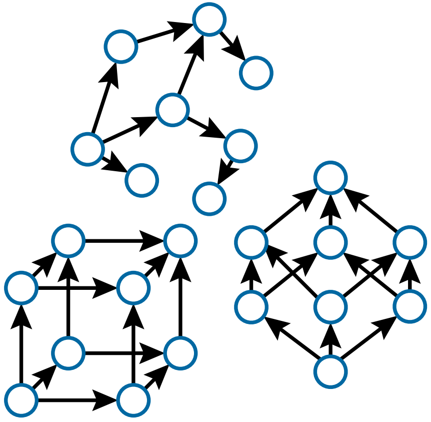
Summary of Our Approach

Distribution



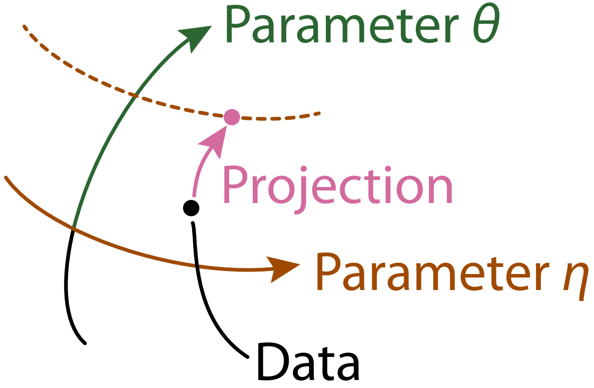
$+$

Partial order structure (DAG)



$=$

Manifold in Information Geometry



Projection: MLE, balancing

Conclusion

- We have established information geometric formulation for partial order structures
  - Learning process can be achieved as a projection in the parameter space (dually flat manifold)

- We have studied several applications
  - Sugiyama, M., Nakahara, H., Tsuda, K., **Tensor Balancing on Statistical Manifold**, ICML2017
  - Sugiyama, M., Nakahara, H., Tsuda, K., **Legendre Decomposition for Tensors**, NeurIPS2018
  - **Truncated Boltzmann machines** (submitted)