

October 19, 2017
Kyoto University
Informatics Seminar



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



Information Geometric Analysis on Partial Order Structures

Mahito Sugiyama (NII, JST PRESTO)

Slide Is Available

- http://mahito.info/files/Sugiyama_KU_InfoSeminar_2017.pdf

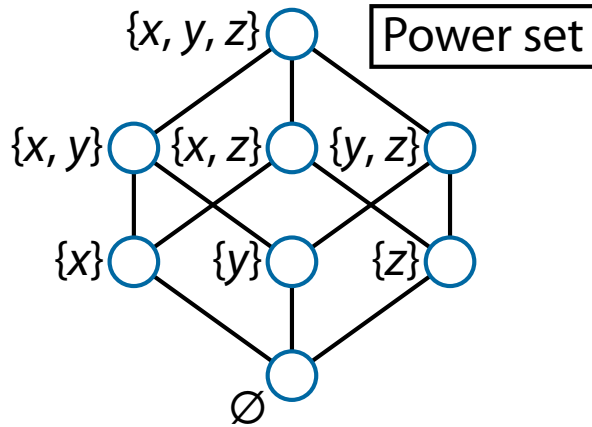
Outline

- Introduction to Pattern mining & Boltzmann machines
- Information geometric formulation
- Projection in statistical manifold
- Application to Matrix balancing
- Conclusion

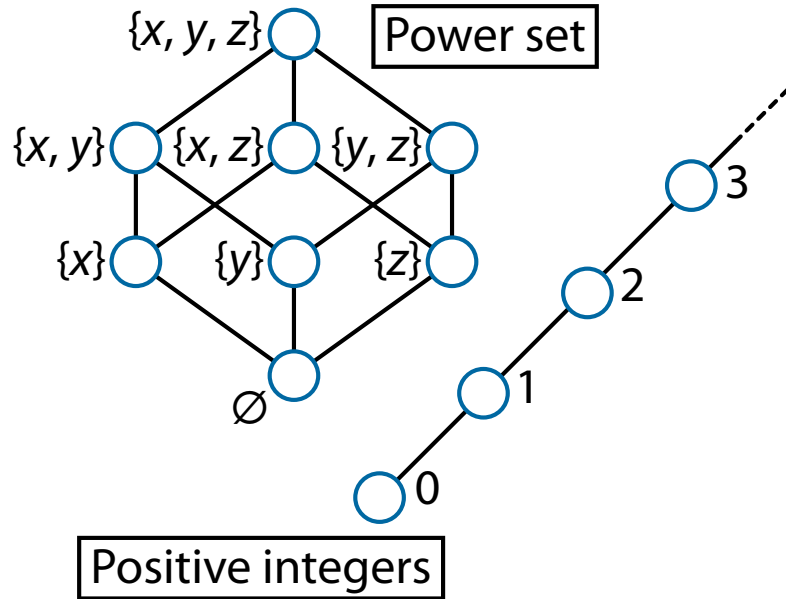
Outline

- Introduction to Pattern mining & Boltzmann machines
- Information geometric formulation
- Projection in statistical manifold
- Application to Matrix balancing
- Conclusion

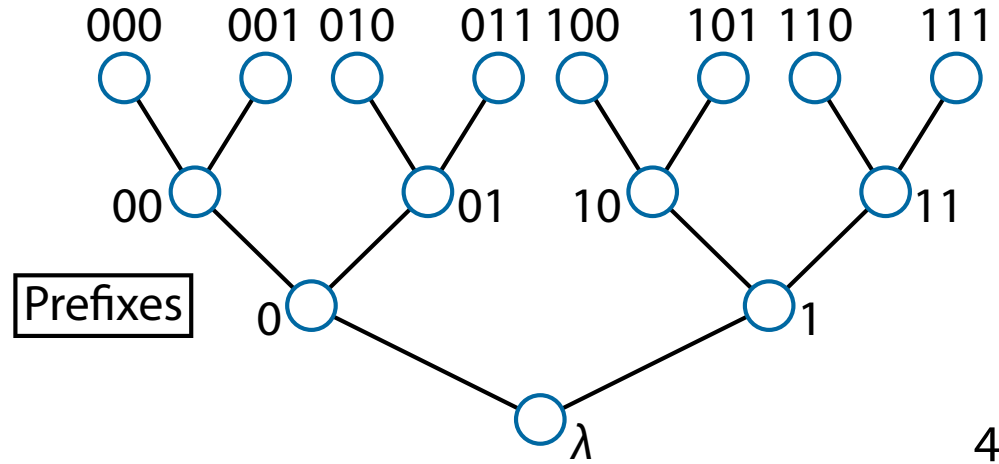
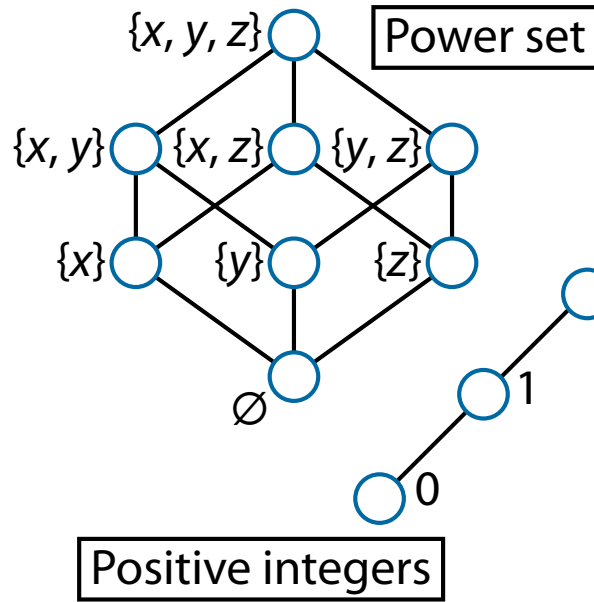
Various Posets (Partially Ordered Sets)



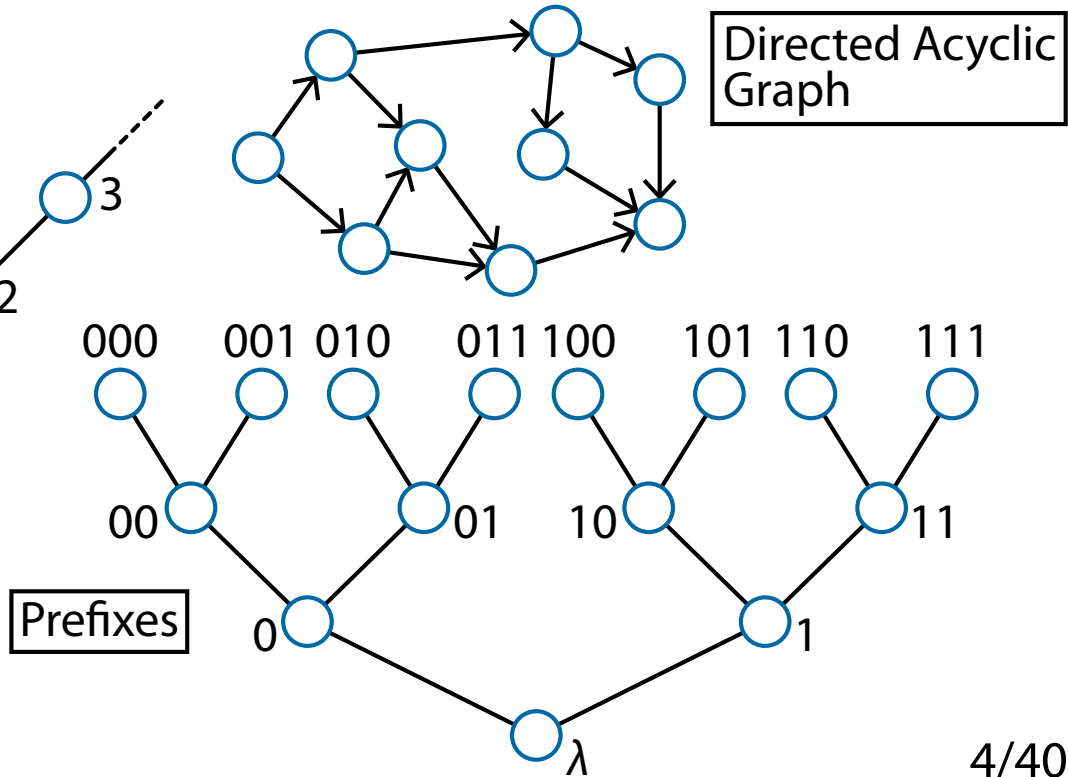
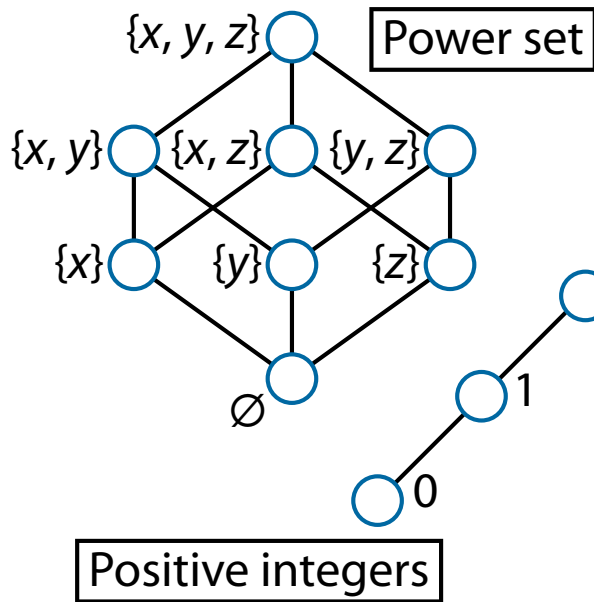
Various Posets (Partially Ordered Sets)



Various Posets (Partially Ordered Sets)



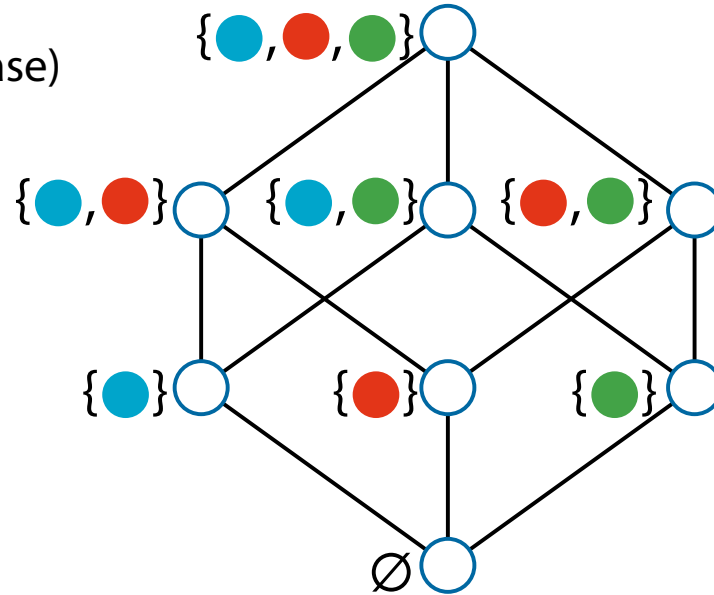
Various Posets (Partially Ordered Sets)



Pattern Mining

Binary vectors
(Transaction database)




	●	●	●
ID 1:	1	1	0
ID 2:	1	1	1
ID 3:	1	1	0
ID 4:	1	1	1
ID 5:	1	1	0
ID 6:	1	0	1
ID 7:	1	0	1
ID 8:	1	1	1
ID 9:	1	0	0
ID 10:	0	1	0

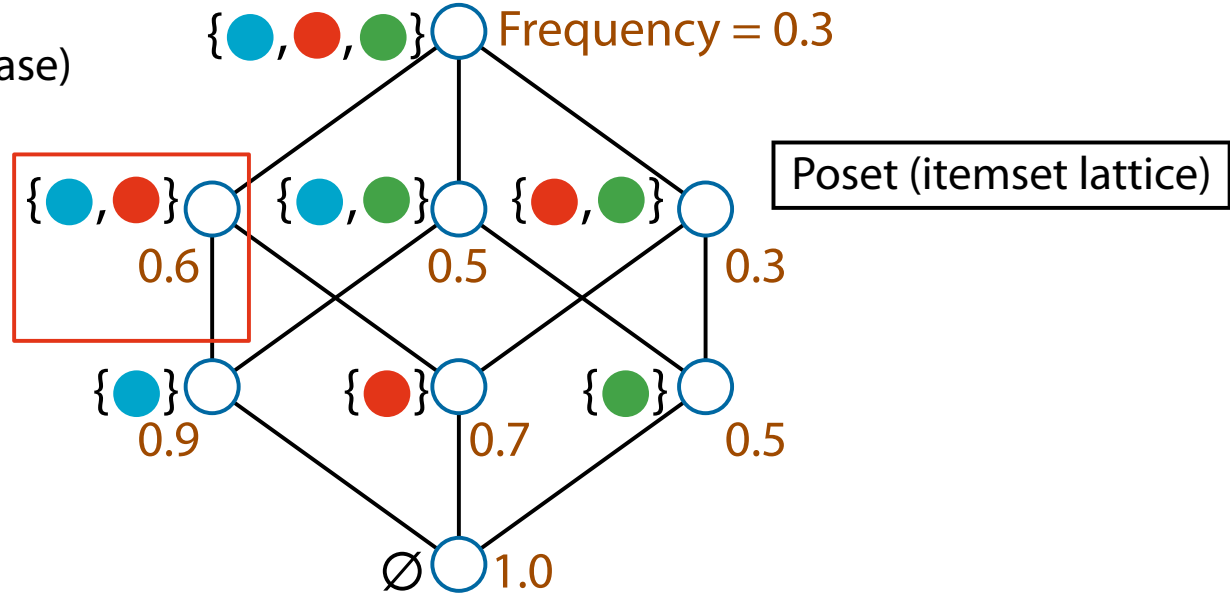


Poset (itemset lattice)

Frequency as Importance Measure




Binary vectors
(Transaction database)

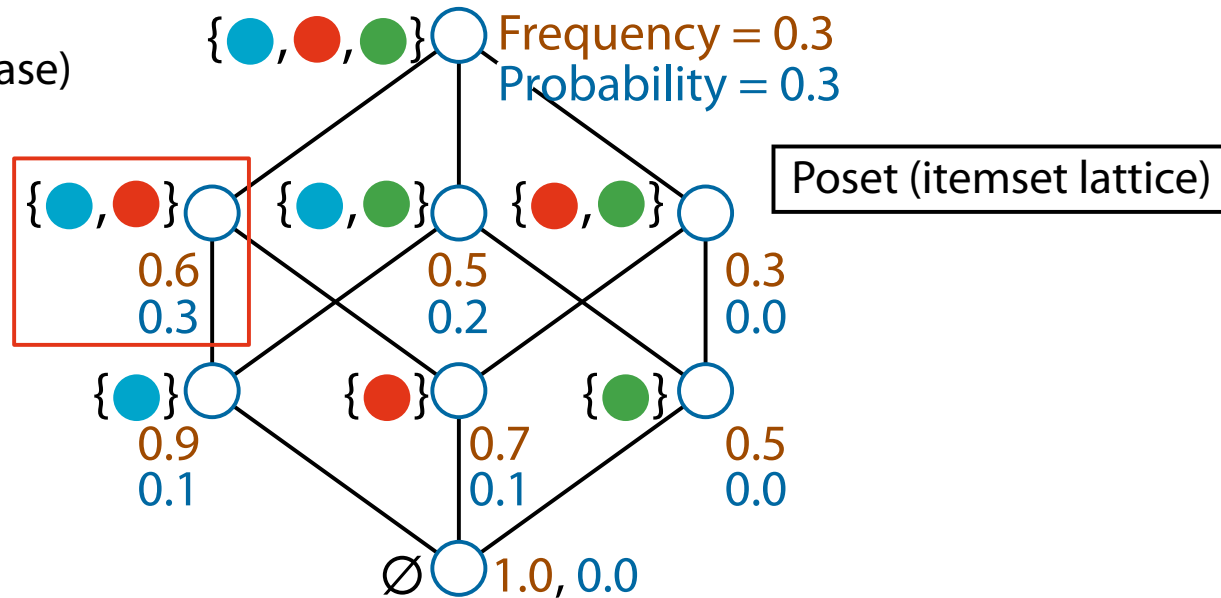
			
ID 1:	1	1	0
ID 2:	1	1	1
ID 3:	1	1	0
ID 4:	1	1	1
ID 5:	1	1	0
ID 6:	1	0	1
ID 7:	1	0	1
ID 8:	1	1	1
ID 9:	1	0	0
ID 10:	0	1	0



Probability on Poset

Binary vectors
(Transaction database)

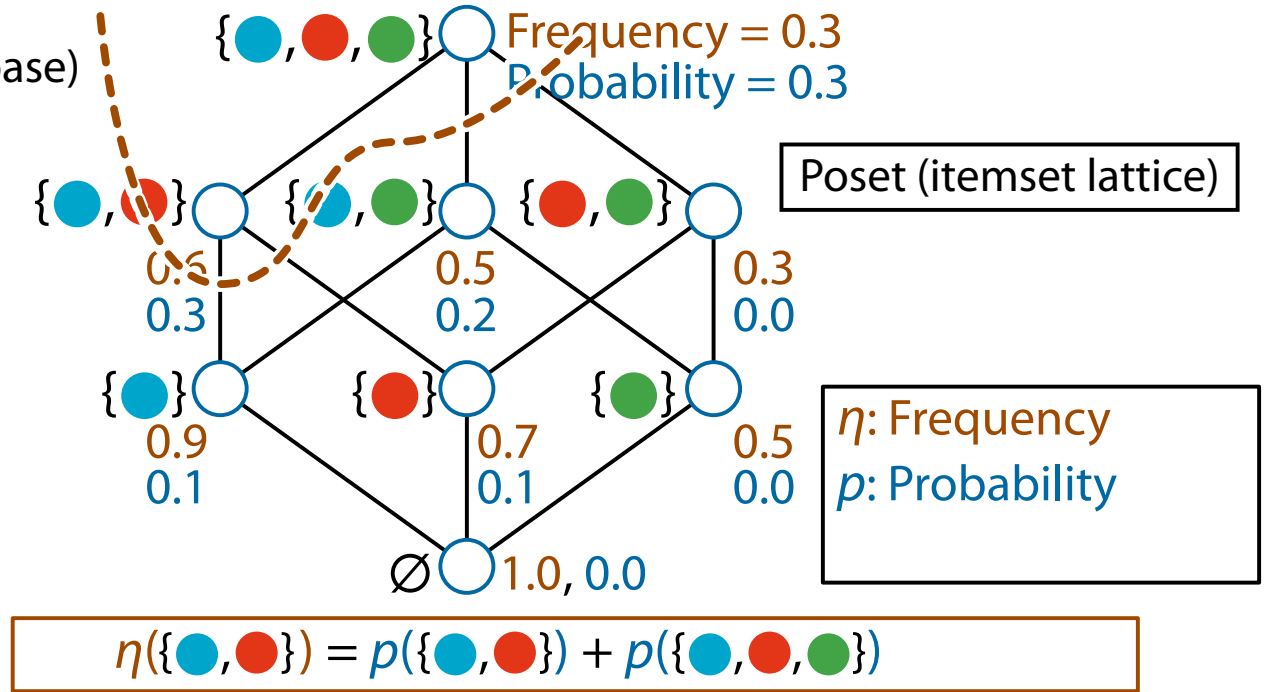
			
ID 1:	1	1	0
ID 2:	1	1	1
ID 3:	1	1	0
ID 4:	1	1	1
ID 5:	1	1	0
ID 6:	1	0	1
ID 7:	1	0	1
ID 8:	1	1	1
ID 9:	1	0	0
ID 10:	0	1	0



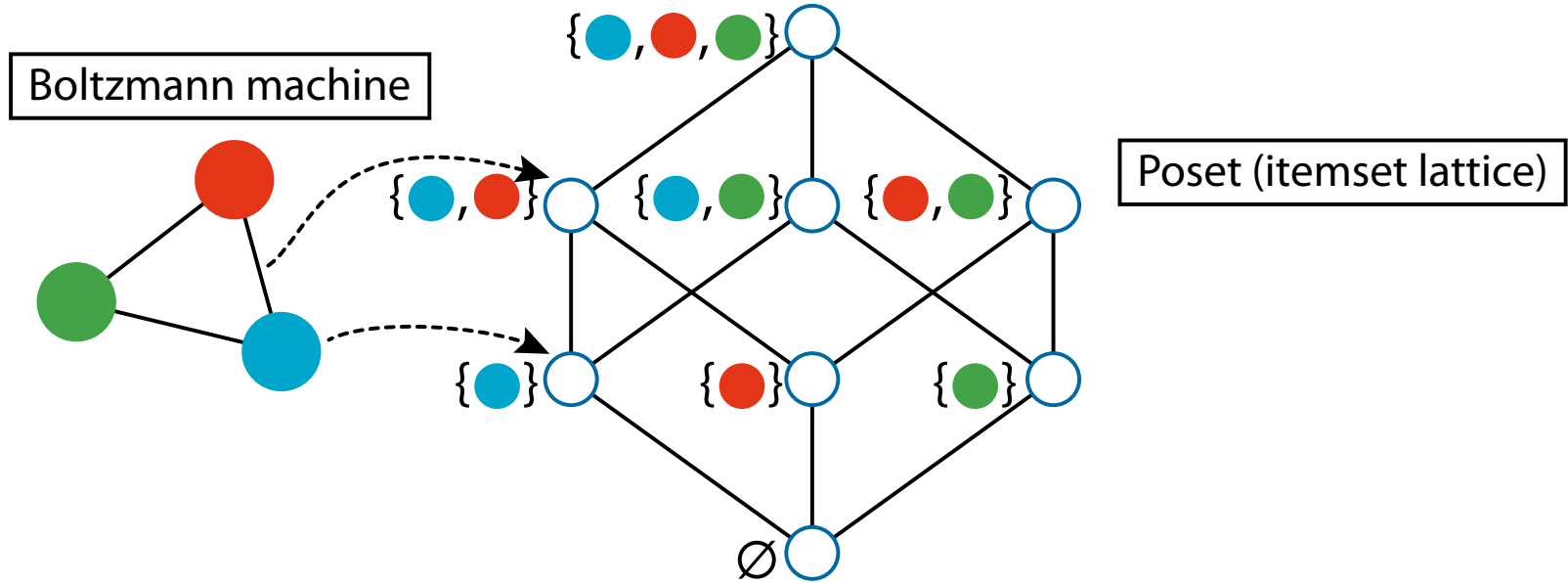
Pattern Mining → Upward Analysis

Binary vectors
(Transaction database)

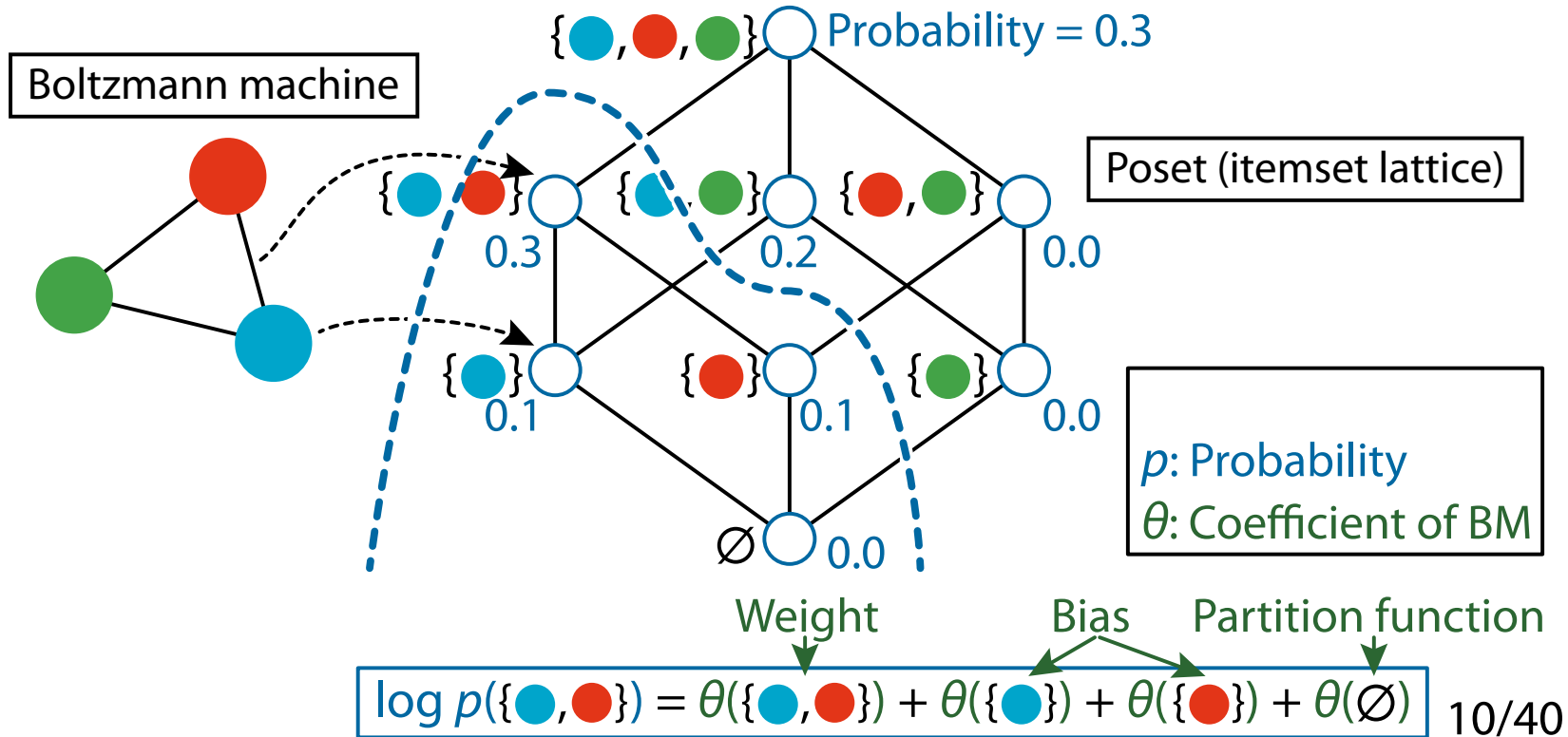
	●	●	●
ID 1:	1	1	0
ID 2:	1	1	1
ID 3:	1	1	0
ID 4:	1	1	1
ID 5:	1	1	0
ID 6:	1	0	1
ID 7:	1	0	1
ID 8:	1	1	1
ID 9:	1	0	0
ID10:	0	1	0



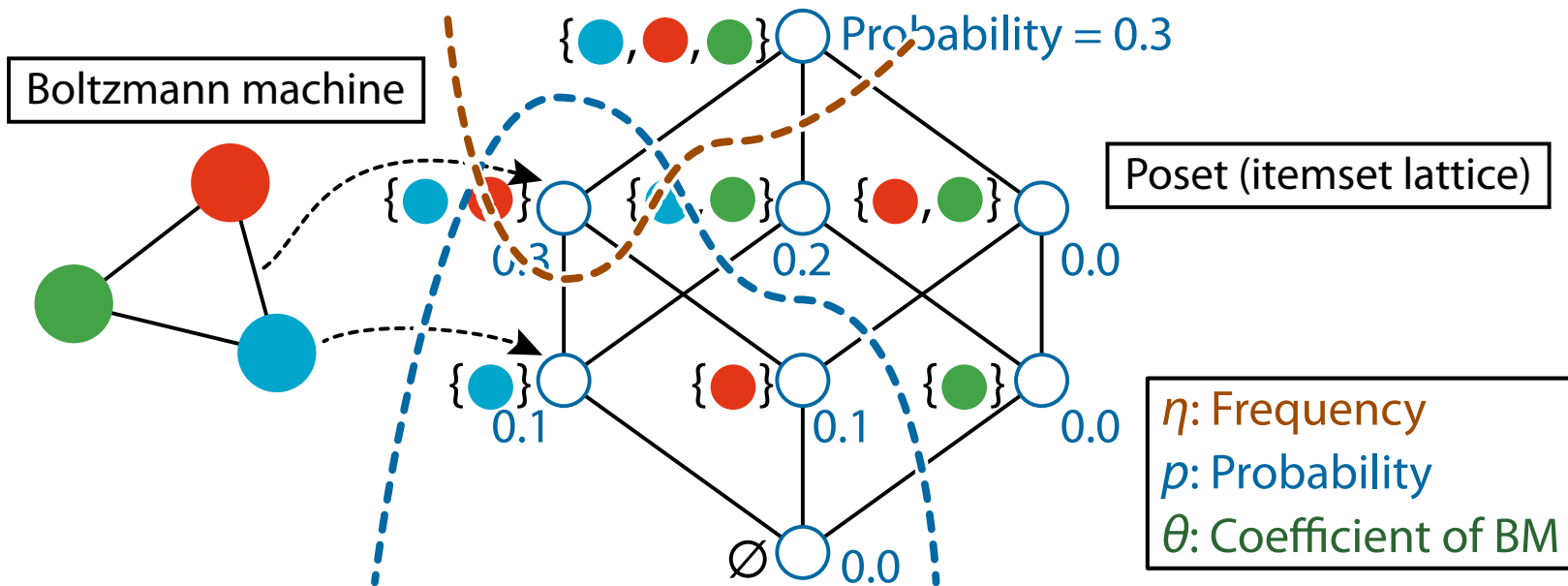
Boltzmann Machines



Boltzmann Machines → Downward Analysis



Pattern Mining & Boltzmann Machines



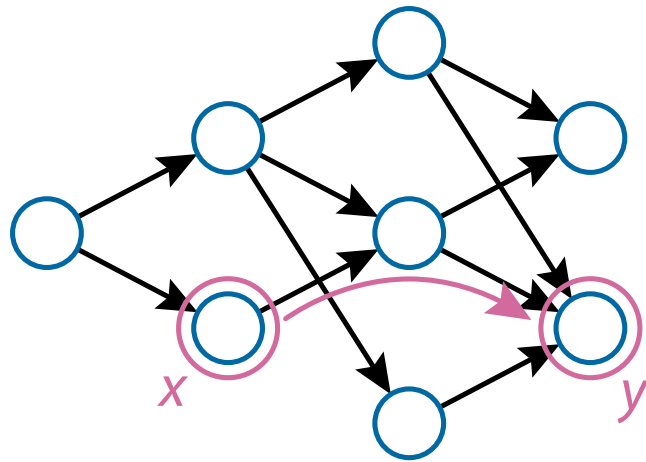
$$\eta(\{\bullet, \bullet\}) = p(\{\bullet, \bullet\}) + p(\{\bullet, \bullet, \bullet\})$$

$$\log p(\{\bullet, \bullet\}) = \theta(\{\bullet, \bullet\}) + \theta(\{\bullet\}) + \theta(\{\bullet\}) + \theta(\emptyset)$$

Outline

- Introduction to Pattern mining & Boltzmann machines
- Information geometric formulation
- Projection in statistical manifold
- Application to Matrix balancing
- Conclusion

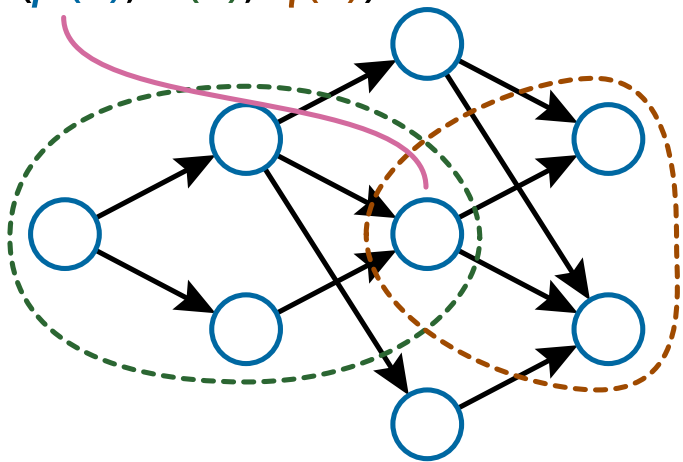
Partially Ordered Set



- 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

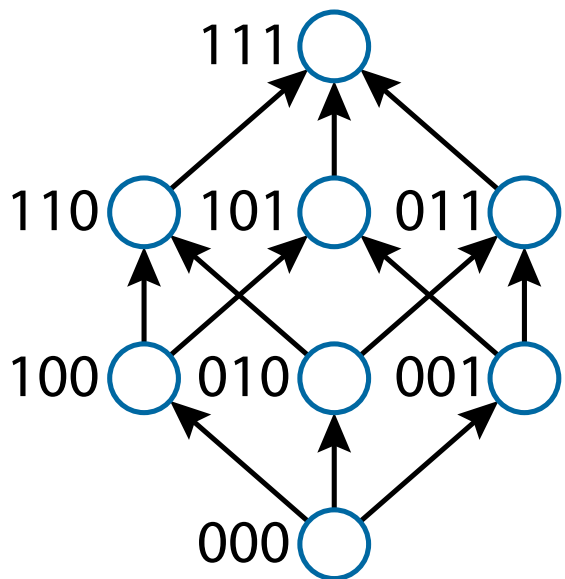
Log-Linear Model on Poset

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

- θ and η 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: $g(\theta) = \nabla\nabla\psi(\theta) = \nabla\eta$, $g(\eta) = \nabla\nabla\varphi(\eta) = \nabla\theta$

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

- ζ and μ 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,

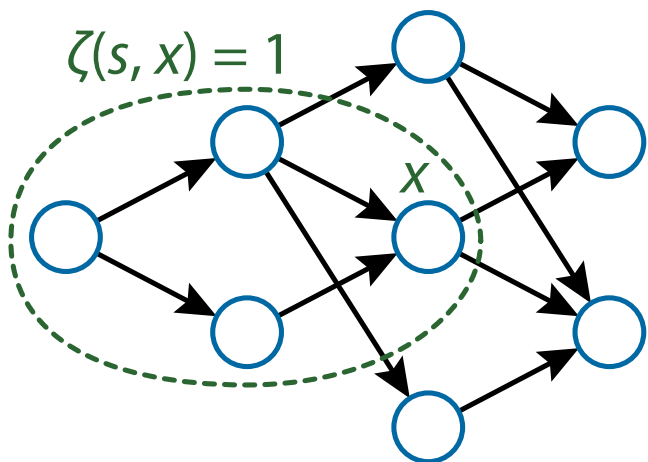
$$\mathbf{E} \left[\frac{\partial}{\partial \theta(x)} \log p(s) \frac{\partial}{\partial \theta(y)} \log p(s) \right] = g_{xy}(\theta),$$

$$\mathbf{E} \left[\frac{\partial}{\partial \eta(x)} \log p(s) \frac{\partial}{\partial \eta(y)} \log p(s) \right] = g_{xy}(\eta)$$

- θ and η 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 $\zeta\mu = I$ (**convolutional inverse**):

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

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, \dots, A_n ,

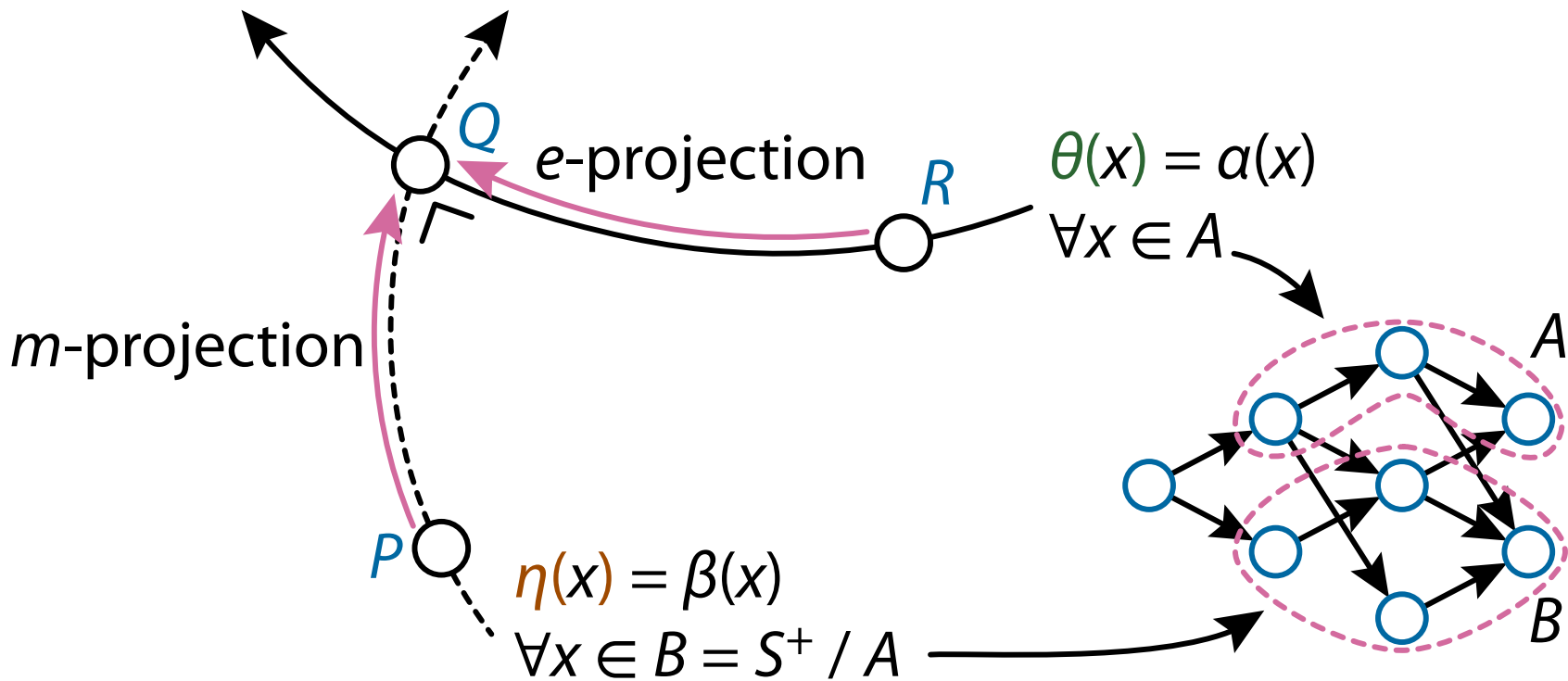
$$\left| \bigcup_i A_i \right| = \sum_{J \subseteq \{1, \dots, n\}, J \neq \emptyset} (-1)^{|J|-1} \left| \bigcap_{j \in J} A_j \right|$$

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

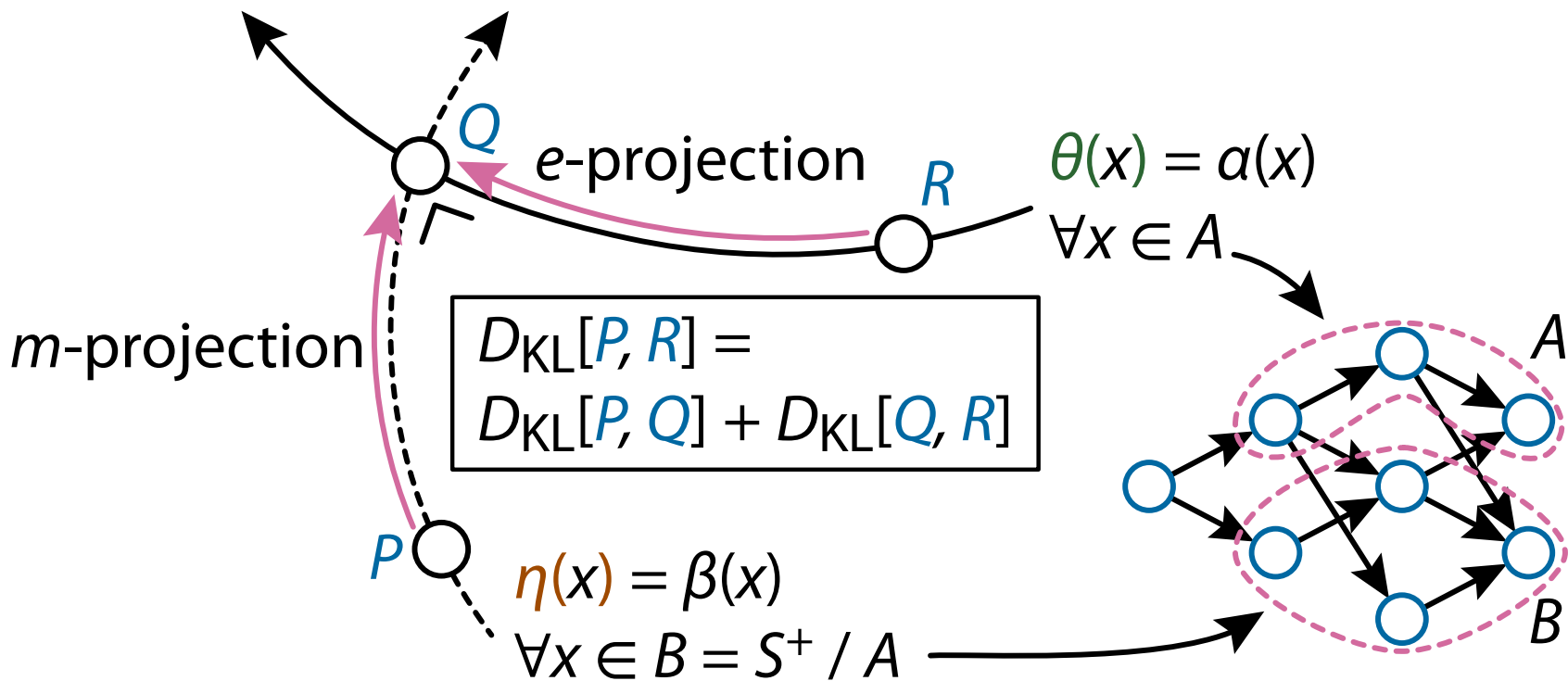
Outline

- Introduction to Pattern mining & Boltzmann machines
- Information geometric formulation
- Projection in statistical manifold
- Application to Matrix balancing
- Conclusion

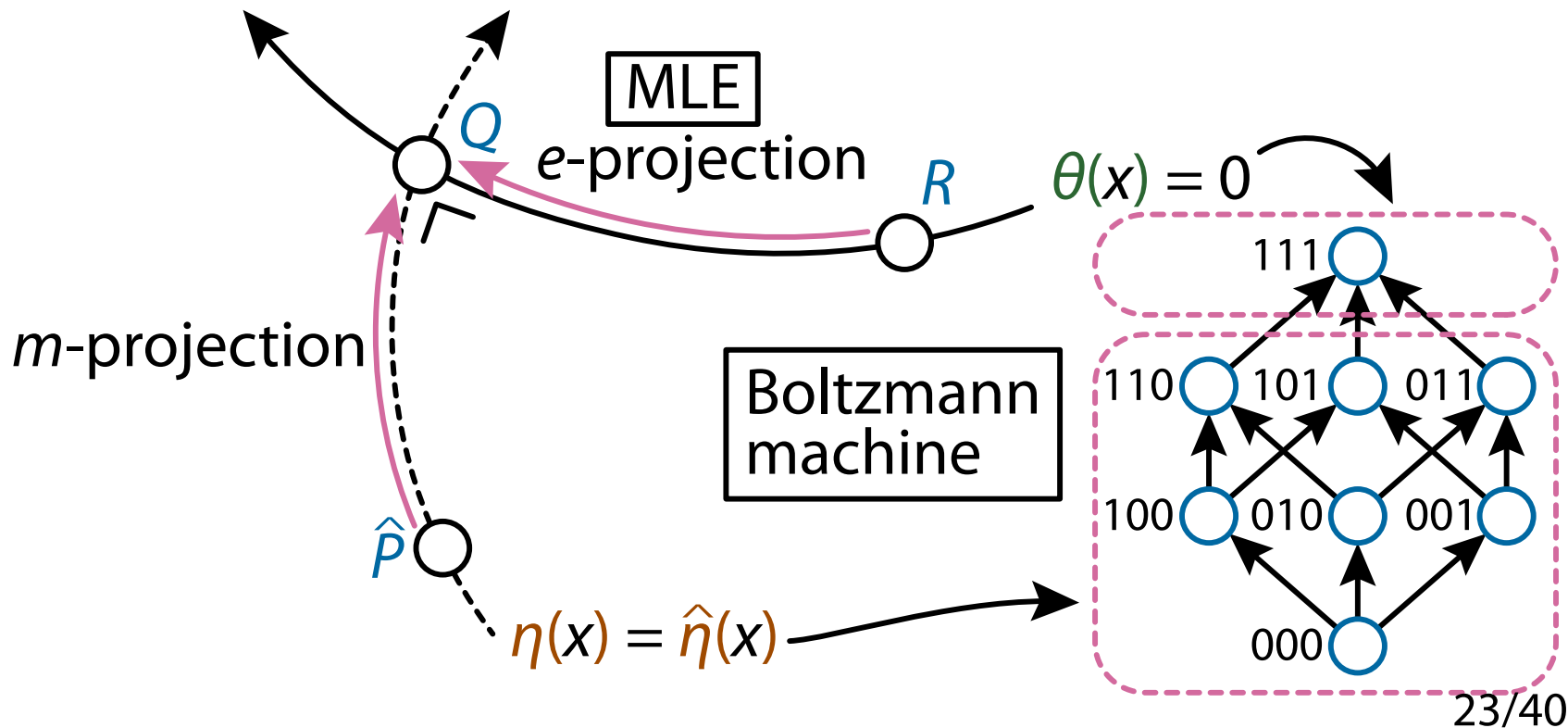
e-Projection and m-Projection



e-Projection and m-Projection



e-Projection and *m*-Projection



Computation of e-Projection

- Given P and β , we compute P_β 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 θ and η are **orthogonal**, we can change $\eta_{P_\beta}^{(t)}(x)$
while fixing $\theta_{P_\beta}^{(t)}(y)$ for $y \notin \text{dom}(\beta)$

Gradient

- We can use **Newton's method** as we can compute the derivatives $\partial\theta^{(t)}(x)/\partial\eta^{(t)}(y)$ and $\partial\eta^{(t)}(x)/\partial\theta^{(t)}(y)$, thanks to the **Möbius inversion**

- Gradient of θ and η is obtained as the Riemannian metric:

$$g(\theta) = \nabla\nabla\psi(\theta) = \nabla\eta \text{ and } g(\eta) = \nabla\nabla\varphi(\eta) = \nabla\theta$$

$$\frac{\partial\eta(x)}{\partial\theta(y)} = \sum_{s \in \mathcal{S}} \zeta(x, s)\zeta(y, s)p(s) - \eta(x)\eta(y),$$

$$\frac{\partial\theta(x)}{\partial\eta(y)} = \sum_{s \in \mathcal{S}} \mu(s, x)\mu(s, y)p(s)^{-1}$$

Newton's Method (1/2)

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

Newton's Method (2/2)

- Each update is

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

- J^{-1} is the inverse of J
- 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 $x, y \in \text{dom}(\beta)$

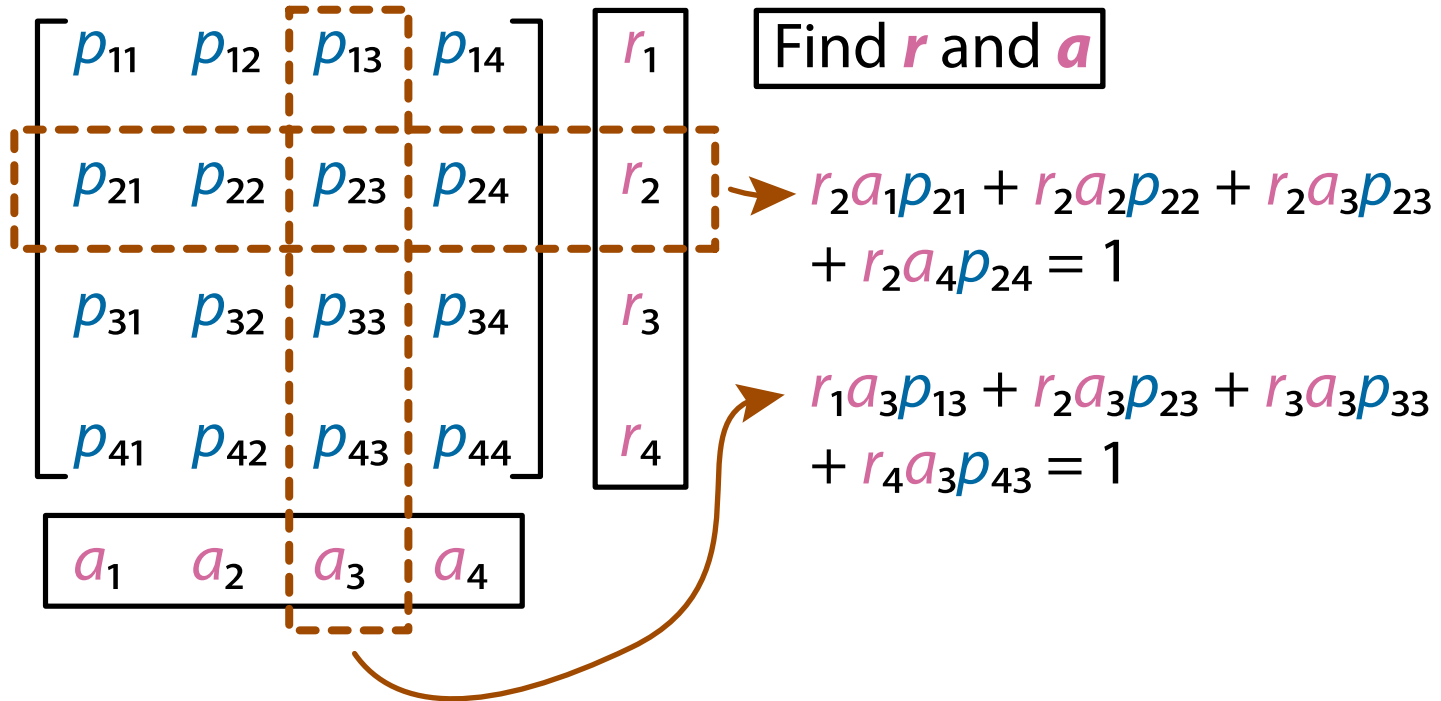
Outline

- Introduction to Pattern mining & Boltzmann machines
- Information geometric formulation
- Projection in statistical manifold
- Application to Matrix balancing
- Conclusion

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

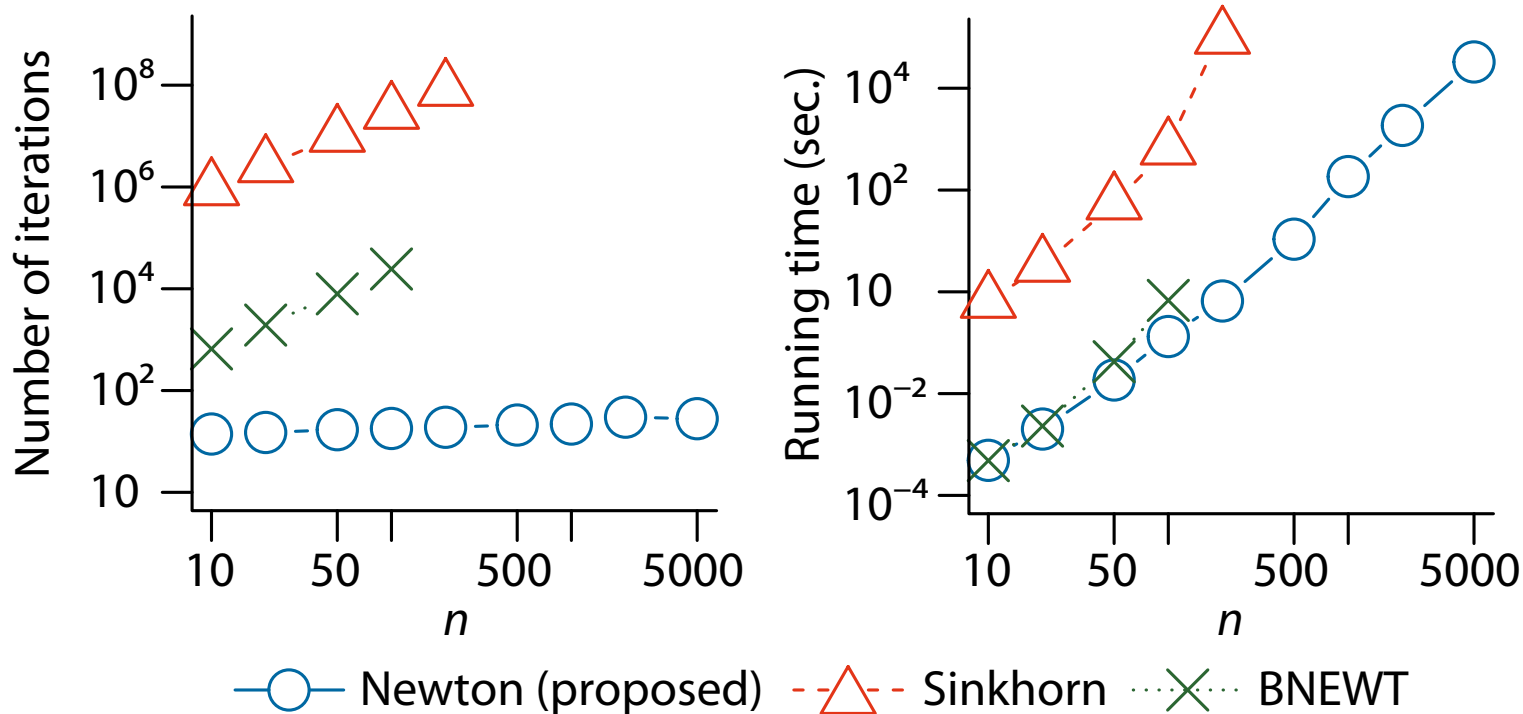
- Problem setting:

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

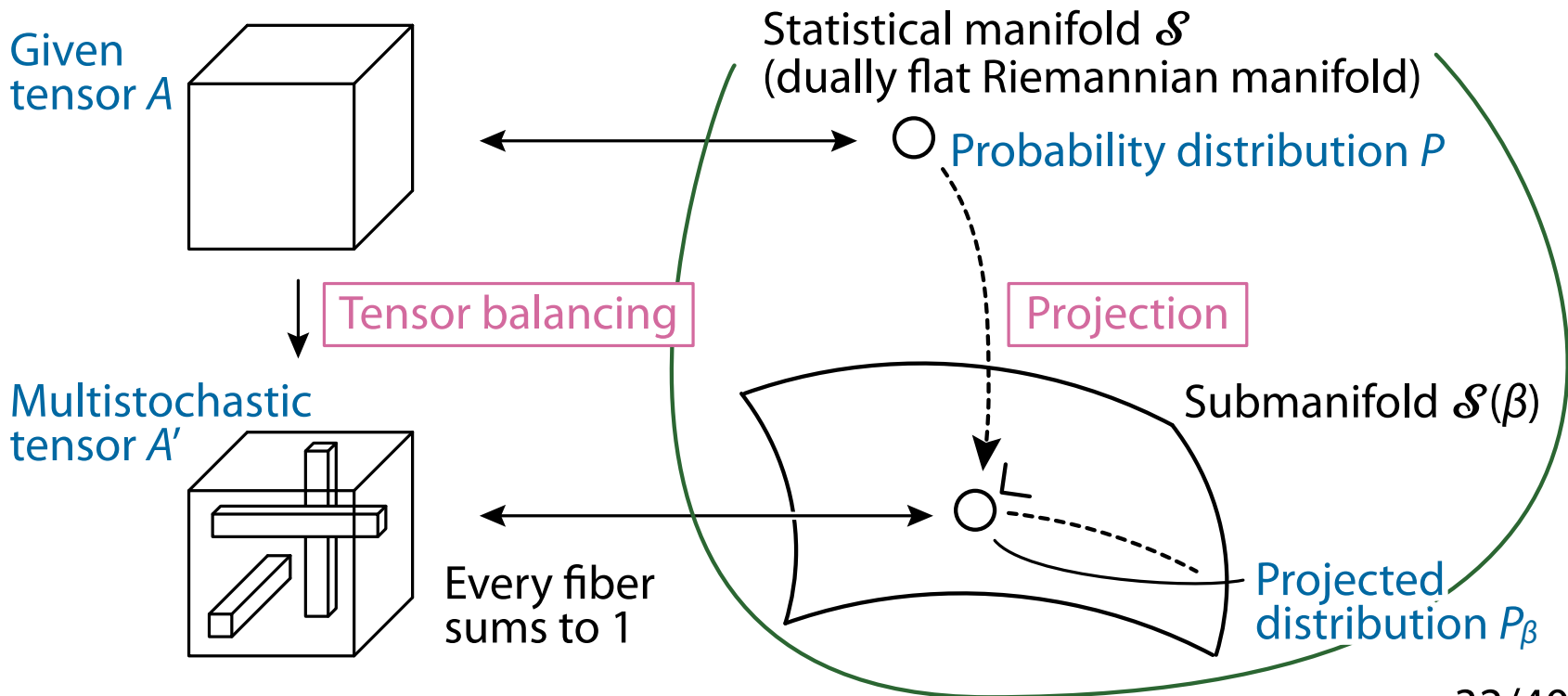
$$(RPS)\mathbf{1} = \mathbf{1} \quad \text{and} \quad (RPS)^T \mathbf{1} = \mathbf{1}$$

- $R = \text{diag}(\mathbf{r}), S = \text{diag}(\mathbf{s})$
 - 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

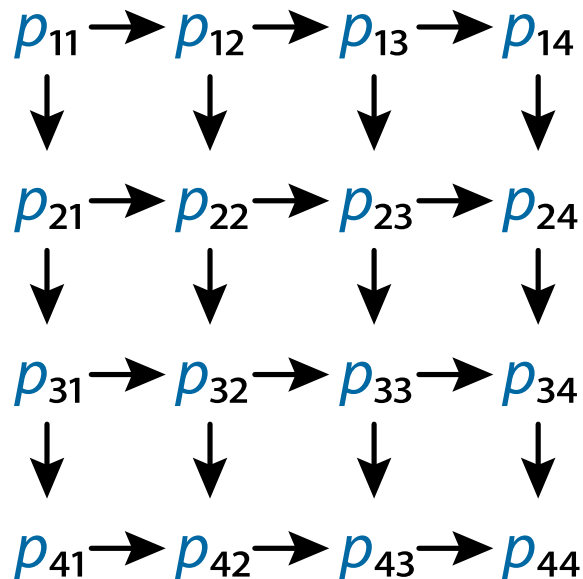
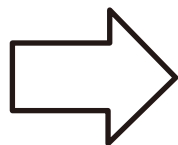


Overview of Our Approach



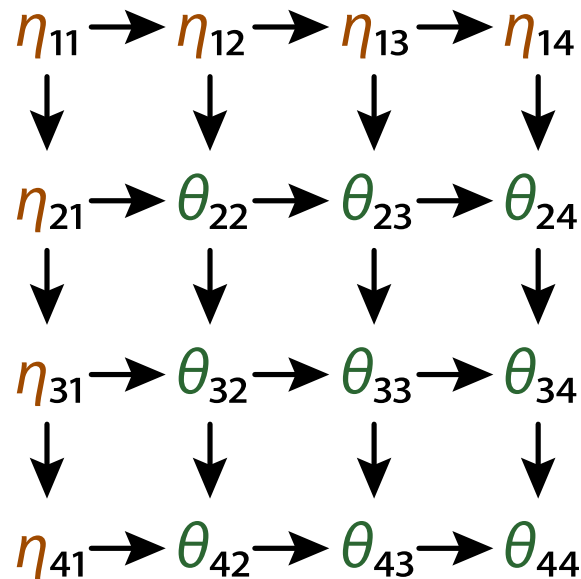
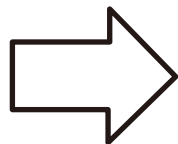
View Matrix as Poset

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



Introduce θ and η

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



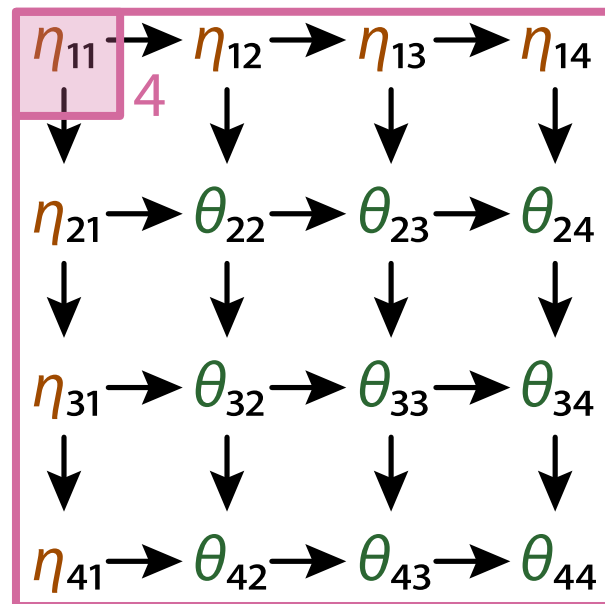
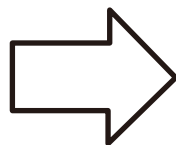
Matrix balancing is achieved if:

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

$$\eta_{11} = 4, \eta_{12} = 3, \eta_{13} = 2, \eta_{14} = 1$$

Introduce θ and η

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



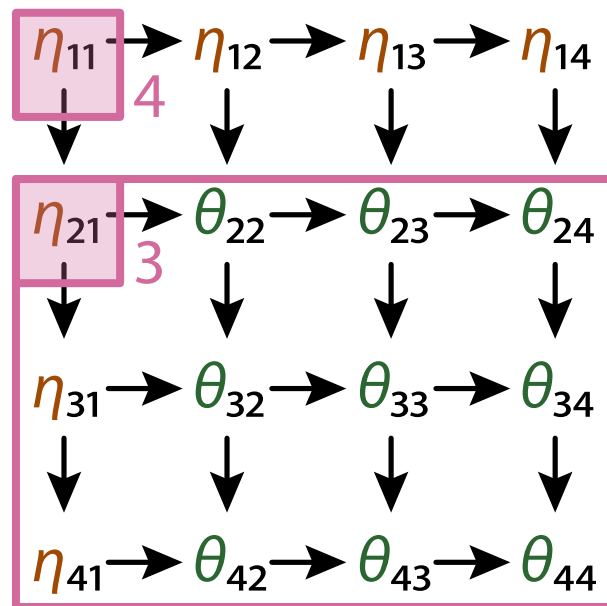
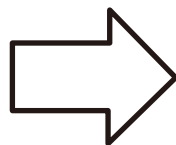
Matrix balancing is achieved if:

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

$$\eta_{11} = 4, \eta_{12} = 3, \eta_{13} = 2, \eta_{14} = 1$$

Introduce θ and η

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



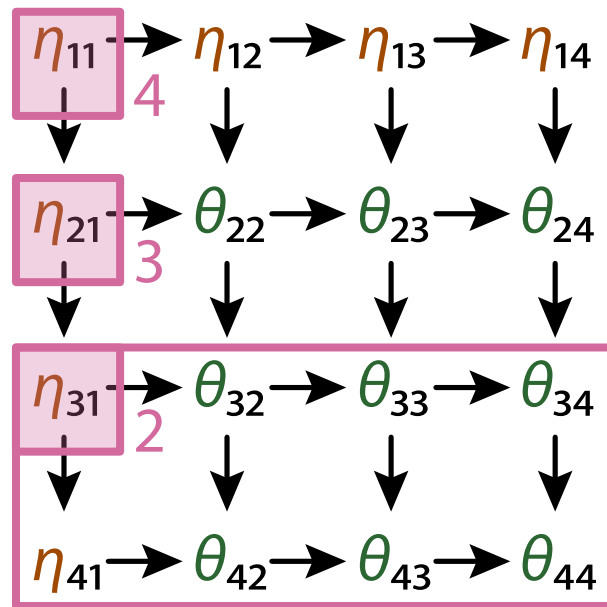
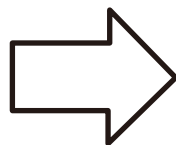
Matrix balancing is achieved if:

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

$$\eta_{11} = 4, \eta_{12} = 3, \eta_{13} = 2, \eta_{14} = 1$$

Introduce θ and η

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



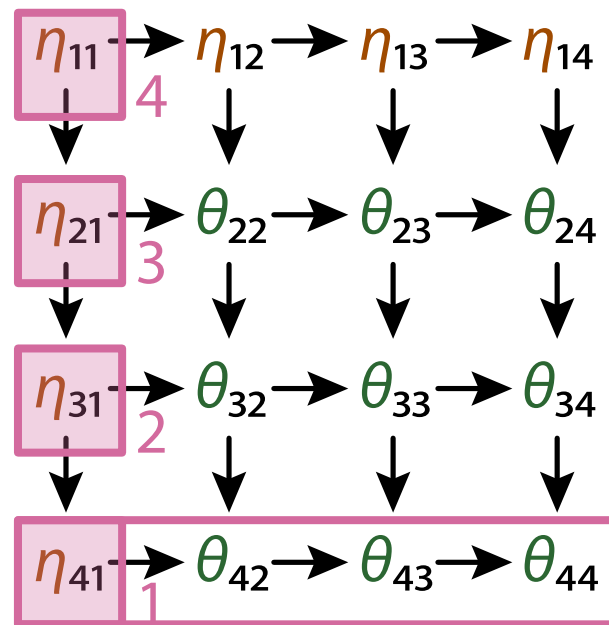
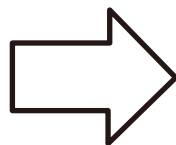
Matrix balancing is achieved if:

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

$$\eta_{11} = 4, \eta_{12} = 3, \eta_{13} = 2, \eta_{14} = 1$$

Introduce θ and η

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



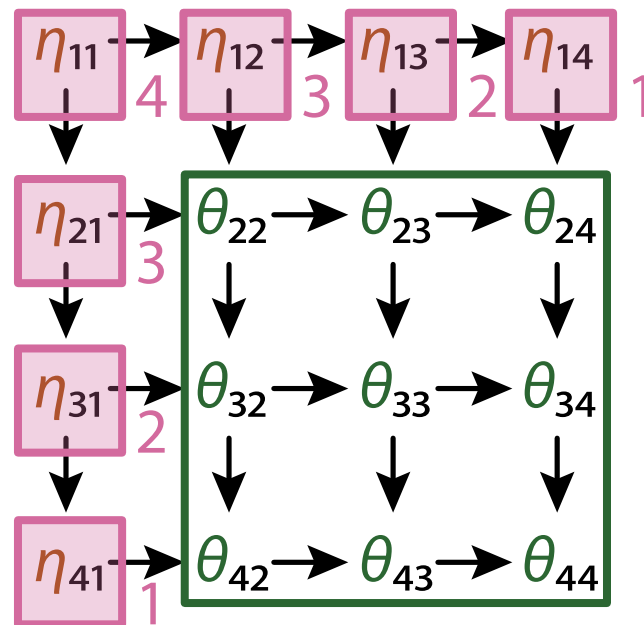
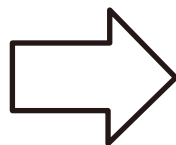
Matrix balancing is achieved if:

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

$$\eta_{11} = 4, \eta_{12} = 3, \eta_{13} = 2, \eta_{14} = 1$$

e-Projection = Balancing

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

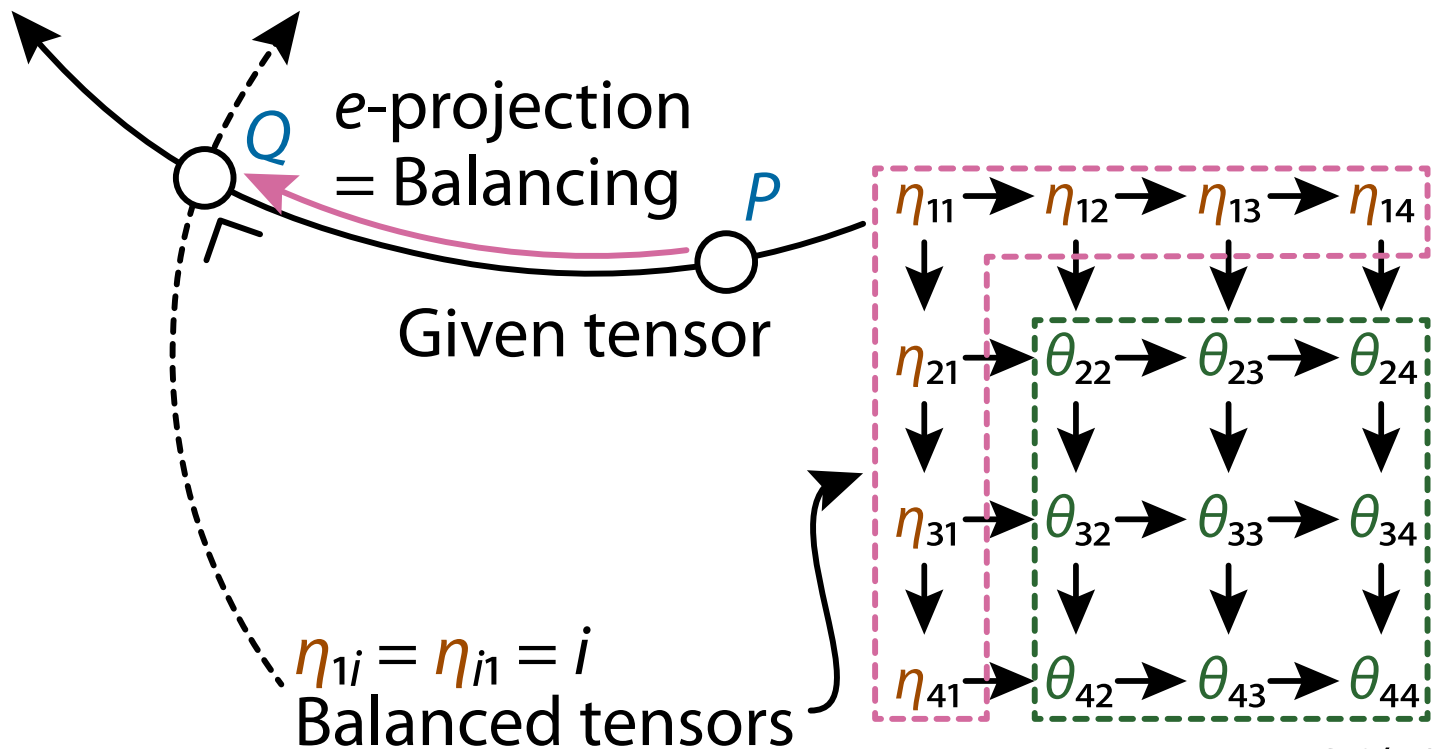


Matrix balancing is achieved if:

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

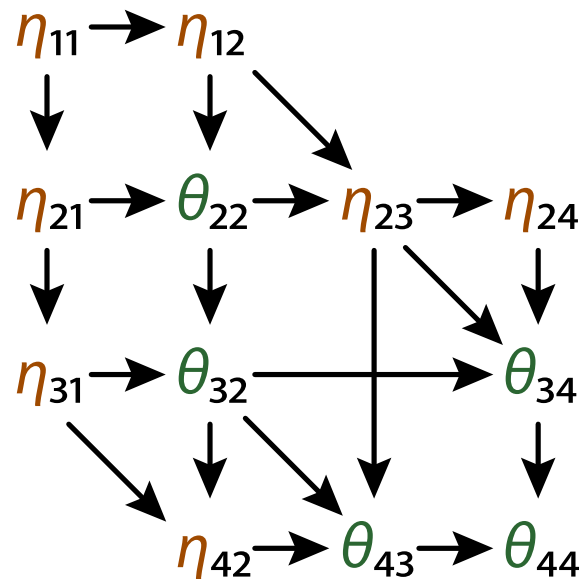
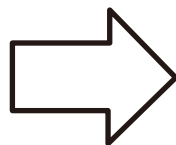
$$\eta_{11} = 4, \eta_{12} = 3, \eta_{13} = 2, \eta_{14} = 1$$

e-Projection = Balancing



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:

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

$$\eta_{11} = 4, \eta_{12} = 3, \eta_{13} = 2, \eta_{14} = 1$$

Outline

- Introduction to Pattern mining & Boltzmann machines
- Information geometric formulation
- Projection in statistical manifold
- Application to Matrix balancing
- Conclusion

Conclusion

- Information geometric formulation connects pattern mining and Boltzmann machines
- Applications including matrix balancing
- **Discrete structure (posets) + Information Geometry = Strong formulation for data analysis!**

Source

- This slide:
http://mahito.info/files/Sugiyama_KU_InfoSeminar_2017.pdf
- Sugiyama, M., Nakahara, H., Tsuda, K.:
Tensor Balancing on Statistical Manifold, ICML 2017
 - arXiv: <https://arxiv.org/abs/1702.08142>
 - GitHub: <https://github.com/mahito-sugiyama/newton-balancing>
- Sugiyama, M., Nakahara, H., Tsuda, K.:
Information Decomposition on Structured Space, IEEE ISIT 2016
 - arXiv: <http://arxiv.org/abs/1601.05533>