

October 24, 2022



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

National Institute of Informatics

Machine Learning and Information Geometry

Introduction to Intelligent Systems Science II

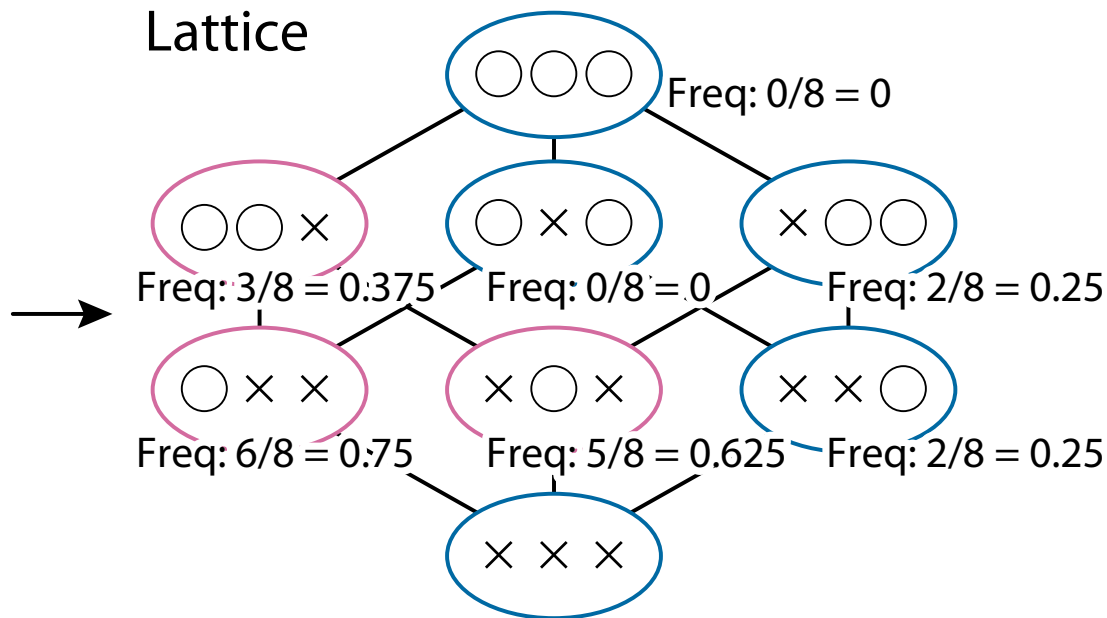
Mahito Sugiyama

Learning Hierarchical Distribution (1/2)

Dataset

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

Lattice

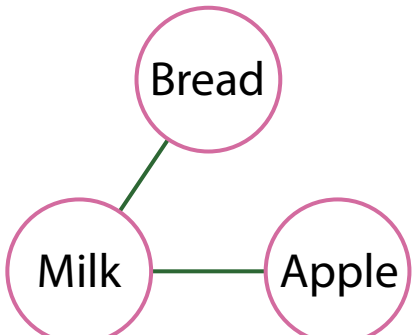


Learning Hierarchical Distribution (2/2)

MLE $\rightarrow$

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

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

Matrix Balancing

$$\begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix}$$

Matrix Balancing

Find r and s :
(Make doubly stochastic matrix)

$$\begin{bmatrix} r_1 & 0 \\ 0 & r_2 \end{bmatrix} \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix} \begin{bmatrix} s_1 & 0 \\ 0 & s_2 \end{bmatrix}$$
$$= \begin{bmatrix} r_1 s_1 p_{11} & r_1 s_2 p_{12} \\ r_2 s_1 p_{21} & r_2 s_2 p_{22} \end{bmatrix} \rightarrow \begin{aligned} \sum_j r_1 s_j p_{1j} &= 1 \\ \sum_j r_2 s_j p_{2j} &= 1 \end{aligned}$$
$$\downarrow \qquad \downarrow$$
$$\sum_i r_i s_1 p_{i1} = 1 \quad \sum_i r_i s_2 p_{i2} = 1$$

Sinkhorn-Knopp Algorithm

- Alternately rescale all rows and columns of a matrix P to sum to 1
- Commonly used to compute entropy-regularized Optimal transport (Wasserstein distance)
 - [Cuturi, 2013]

Revisit Matrix Balancing

p_{11} p_{12} p_{13}

p_{21} p_{22} p_{23}

p_{31} p_{32} p_{33}

[Sugiyama, Nakahara, Tsuda, ICML2017]

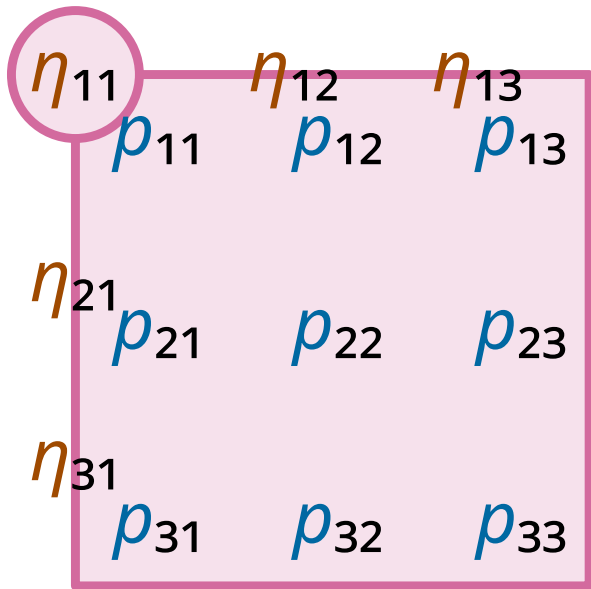
Introduce η

$$\begin{matrix} \eta_{11} & \eta_{12} & \eta_{13} \\ p_{11} & p_{12} & p_{13} \end{matrix}$$

$$\begin{matrix} \eta_{21} & & \\ p_{21} & p_{22} & p_{23} \end{matrix}$$

$$\begin{matrix} \eta_{31} & & \\ p_{31} & p_{32} & p_{33} \end{matrix}$$

Introduce η



Introduce η

$$\begin{array}{ccc} \eta_{11} & \eta_{12} & \eta_{13} \\ p_{11} & p_{12} & p_{13} \end{array}$$

$$\begin{array}{ccc} \eta_{21} & p_{22} & p_{23} \\ p_{21} & & \end{array}$$
$$\begin{array}{ccc} \eta_{31} & p_{32} & p_{33} \\ p_{31} & & \end{array}$$

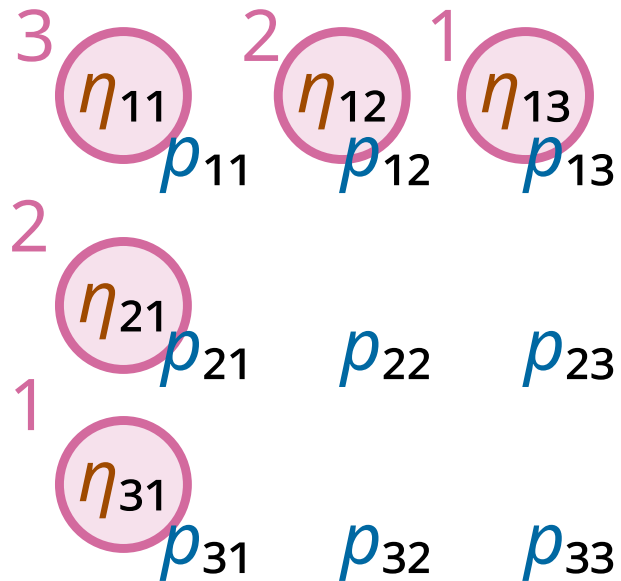
Introduce η

$$\begin{array}{ccc} \eta_{11} & \eta_{12} & \eta_{13} \\ p_{11} & p_{12} & p_{13} \end{array}$$

$$\begin{array}{ccc} \eta_{21} & & \\ p_{21} & p_{22} & p_{23} \end{array}$$

$$\begin{array}{ccc} \eta_{31} & & \\ p_{31} & p_{32} & p_{33} \end{array}$$

Introduce η



Introduce θ

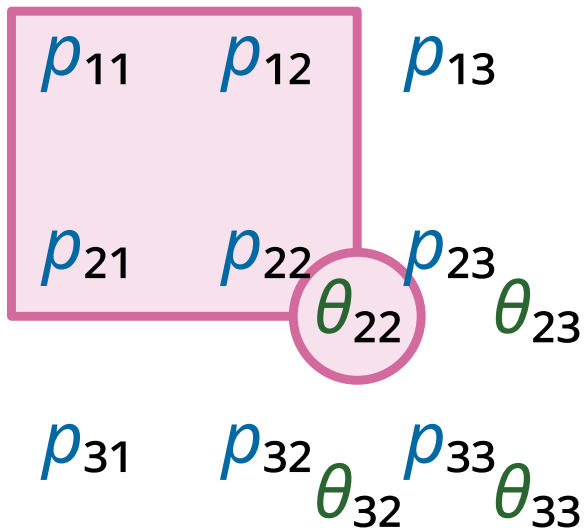
$$p_{11} \quad p_{12} \quad p_{13}$$

$$p_{21} \quad p_{22} \quad p_{23} \\ \theta_{22} \quad \theta_{23}$$

$$p_{31} \quad p_{32} \quad p_{33} \\ \theta_{32} \quad \theta_{33}$$

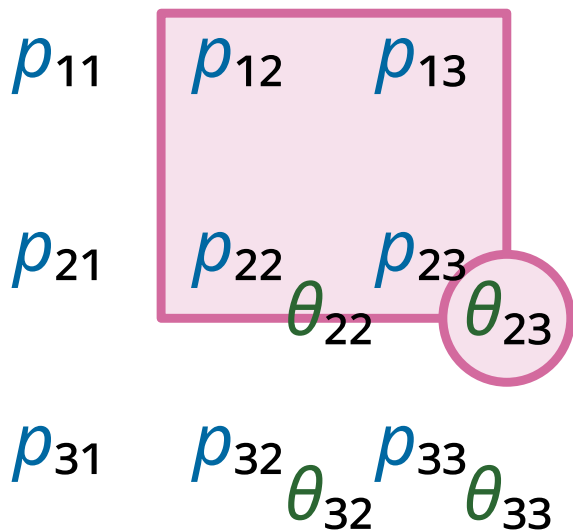
$$\begin{aligned} \theta_{ij} = & \log p_{ij} \\ & - \log p_{i-1j} - \log p_{ij-1} \\ & + \log p_{i-1j-1} \end{aligned}$$

Introduce θ



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

Introduce θ



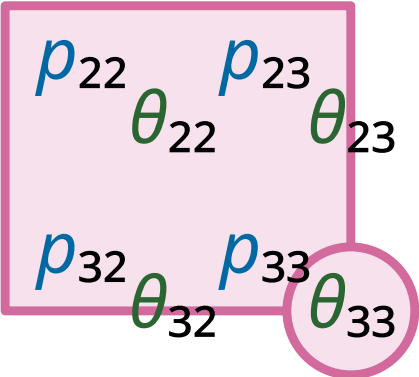
$$\begin{aligned}\theta_{ij} = & \log p_{ij} \\ & - \log p_{i-1j} - \log p_{ij-1} \\ & + \log p_{i-1j-1}\end{aligned}$$

Introduce θ

p_{11} p_{12} p_{13}

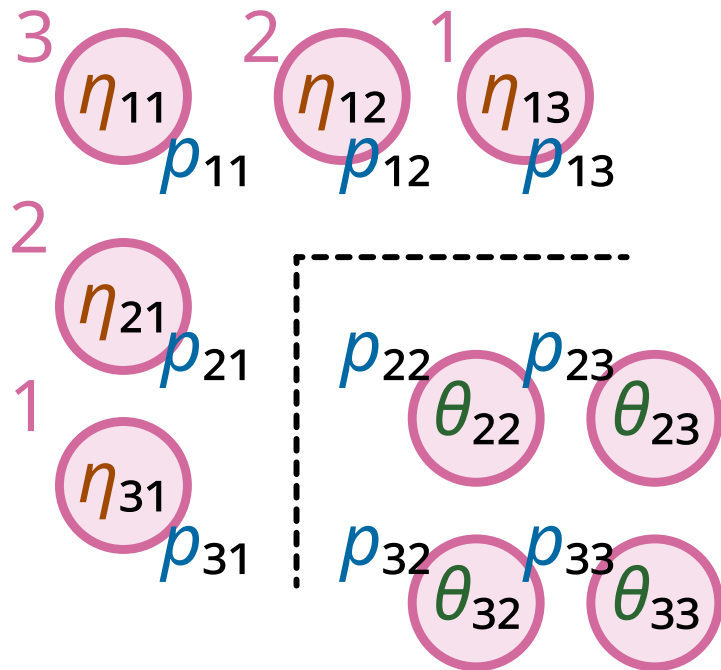
p_{21} p_{22} p_{23}
 θ_{22} θ_{23}

p_{31} p_{32} p_{33}
 θ_{32} θ_{33}



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

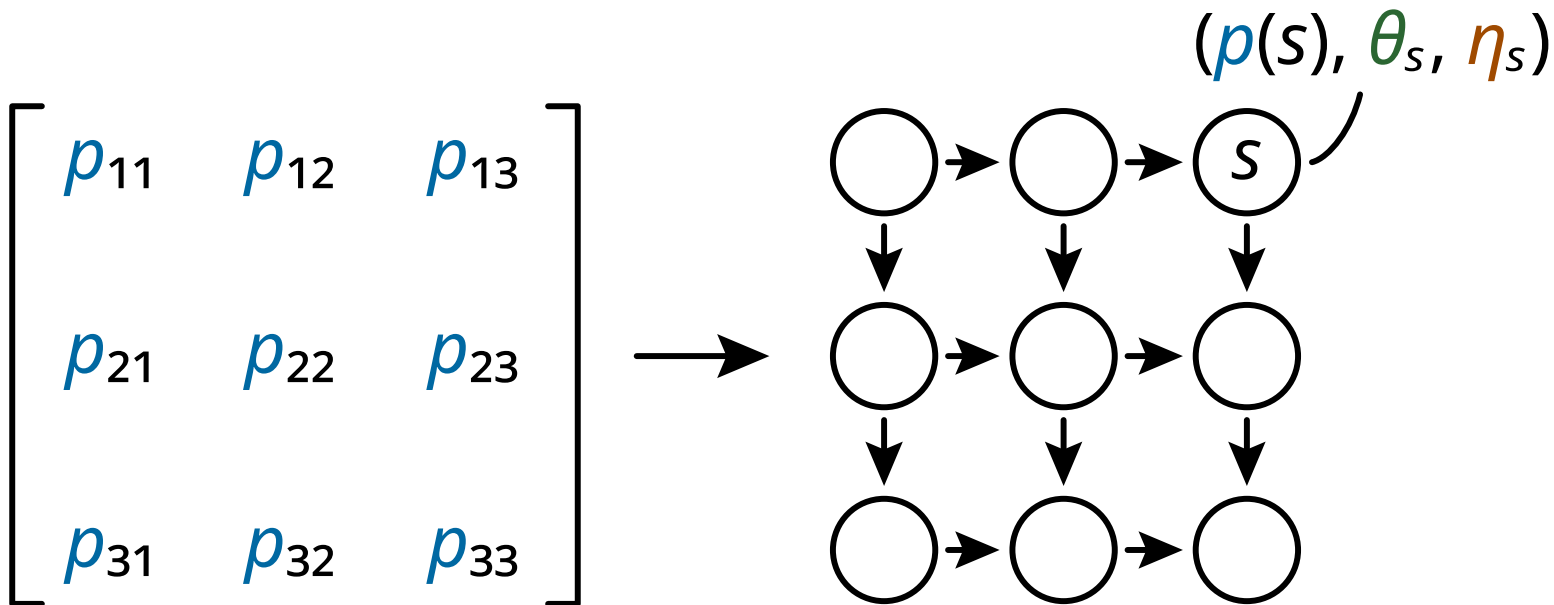
Balancing as Constraints on η and θ



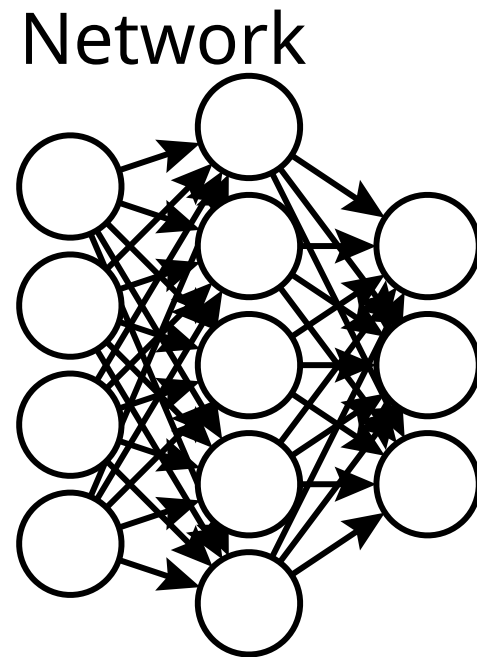
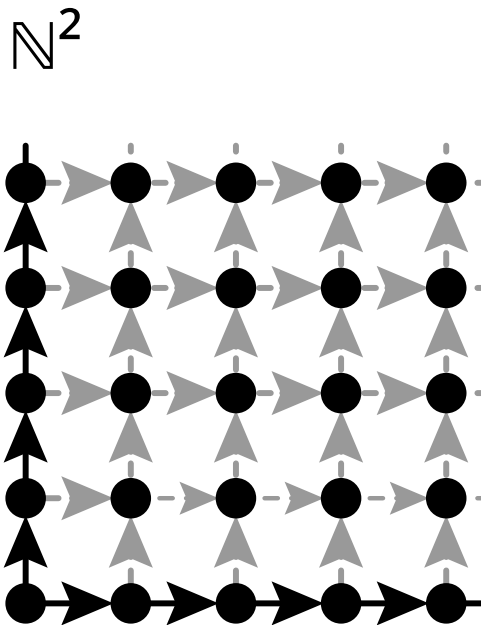
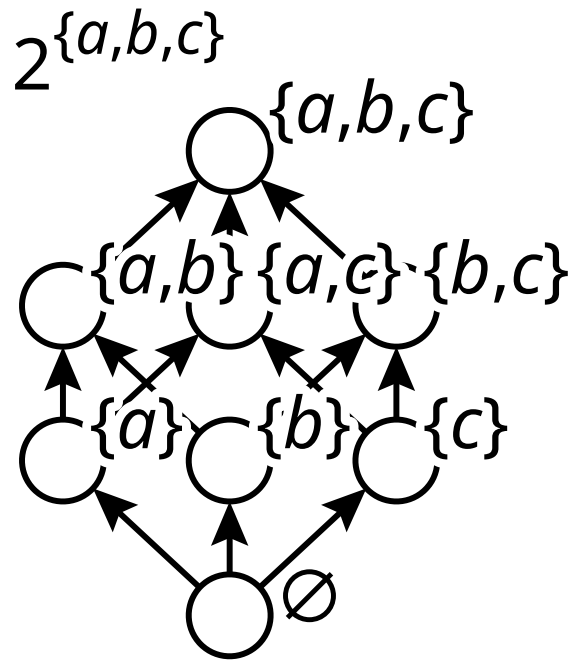
$$\theta_{ij} = \log p_{ij} - \log p_{i-1j} - \log p_{ij-1} + \log p_{i-1j-1}$$

Matrix balancing $\Leftrightarrow$
 Satisfy $\eta_{i1} = \eta_{1i} = 3 - i + 1$
 with keeping all θ_{ij}

Introduce Partial Order Structure



Partially Ordered Sets (Posets)



Incidence Algebra

- Incidence algebra is defined over a poset $(S, \leq)$
 - (Closed) Interval $[a, b] = \{s \in S \mid a \leq s \leq b\}$
- Members of the incidence algebra are functions $\alpha(a, b)$ from intervals $[a, b]$ to a scalar with

$$(\alpha + \beta)(a, b) = \alpha(a, b) + \beta(a, b)$$

$$(\alpha\beta)(a, b) = \sum_{a \leq x \leq b} \alpha(a, x)\beta(x, b) \quad (\text{convolution})$$

Special Elements

- Delta function δ :

$$\delta(a, b) = \begin{cases} 1 & \text{if } a = b \\ 0 & \text{otherwise} \end{cases}$$

- Zeta function ζ : (integral)

$$\zeta(a, b) = \begin{cases} 1 & \text{if } a \leq b \\ 0 & \text{otherwise} \end{cases}$$

- Möbius function $\mu = \zeta^{-1}$: $\zeta\mu = \delta$ (differential)

(ζ, μ) Leads to Non-Singularity

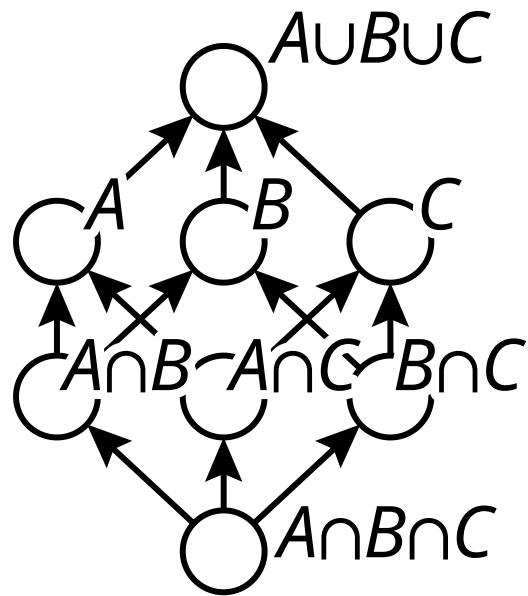
- For a poset $(S, \leq)$, let $S = \{s_1, s_2, \dots, s_n\}$
- Let us define the **zeta matrix** $\mathbf{Z} \in \{0, 1\}^{n \times n}$ as
$$z_{ij} = \zeta(s_i, s_j)$$
- Relationship between ζ and μ guarantees that $\mathbf{Z}$ is always **regular**:
$$\mathbf{Z}^{-1} = \mathbf{M} \quad \text{such that} \quad m_{ij} = \mu(s_i, s_j)$$
 - $\mathbf{M} \in \mathbb{Z}^{n \times n}$

Möbius Inversion Formula

- Given a poset S , for any functions $f, g : S \rightarrow \mathbb{R}$, the Möbius inversion formula is given as

$$\left\{ \begin{array}{l} g(x) = \sum_{s \in S} \zeta(s, x) f(s) = \sum_{s \leq x} \zeta(s, x) f(s) = \sum_{s \leq x} f(s) \\ f(x) = \sum_{s \in S} \mu(s, x) g(s) = \sum_{s \leq x} \mu(s, x) g(s) \end{array} \right.$$

E.g.1: Inclusion-Exclusion Principle

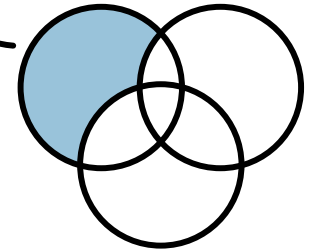


$$|A \cup B \cup C| = |A| + |B| + |C| - |A \cap B| - |A \cap C| - |B \cap C| + |A \cap B \cap C|$$

$$f(X) = |X| \quad g(X) = |X \setminus \bigcup_{Y \subset X} Y|$$

$$f(X) = \sum_{Y \leq X} g(Y)$$

$$g(X) = \sum_{Y \leq X} \mu(Y, X) f(Y)$$



Log-Linear Model on Poset [ICML2017]

- For probability $p:S \rightarrow (0, 1)$ with $\sum_{x \in S} p(x) = 1$, introduce θ and η as

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

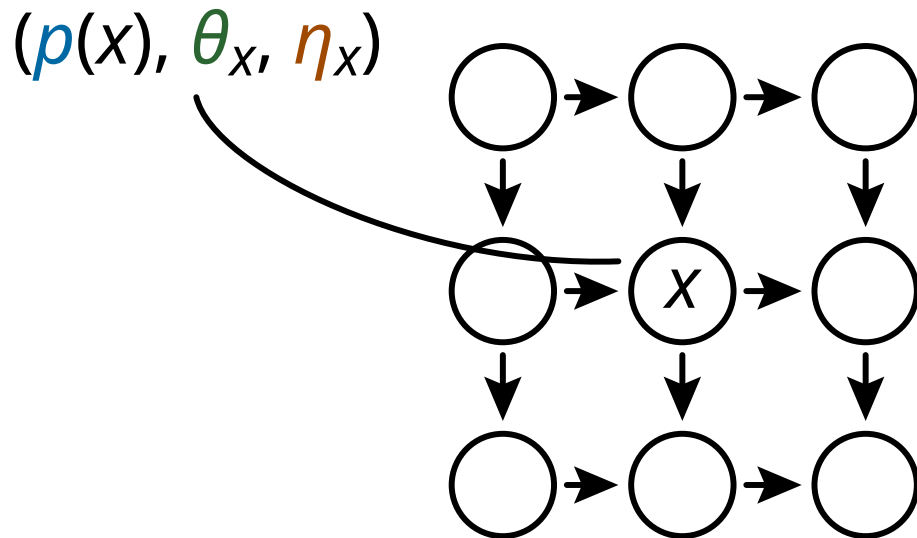
$$\eta_x = \sum_{s \in S} \zeta(x, s) p(s) = \sum_{s \geq x} p(s)$$

- From the Möbius inversion formula, log-linear model is:

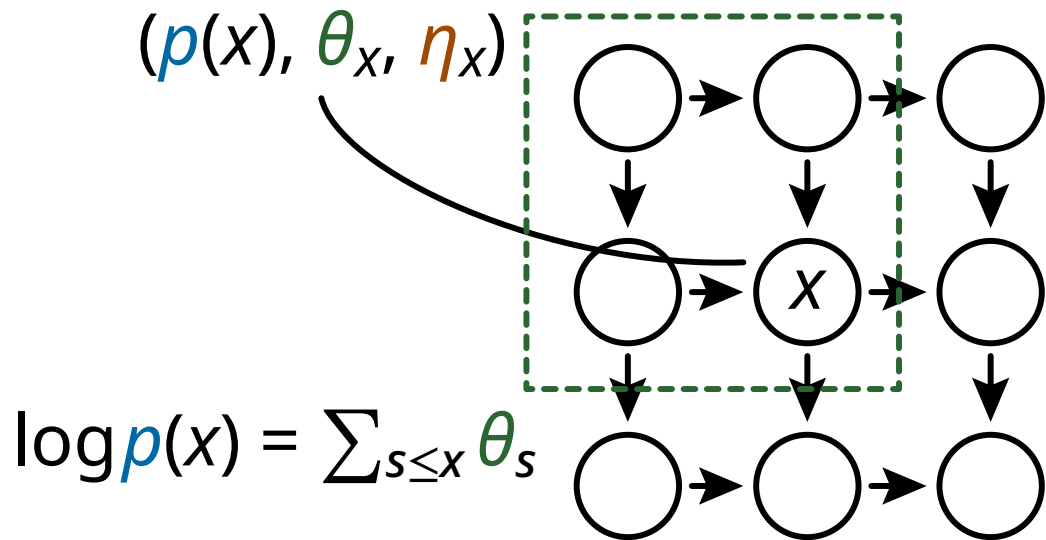
$$\log p(x) = \sum_{s \in S} \zeta(s, x) \theta_s = \sum_{s \leq x} \theta_s$$

- In the matrix form: $\log p = Z\theta$

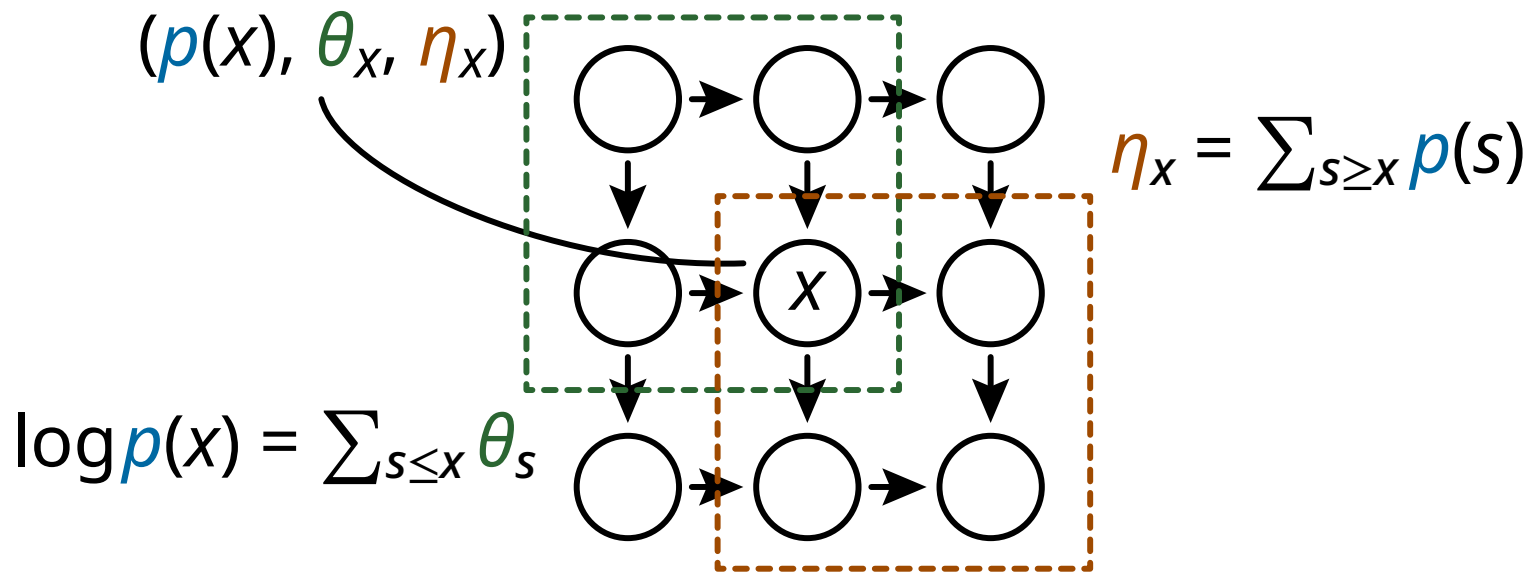
Log-Linear Model on Poset



Log-Linear Model on Poset



Log-Linear Model on Poset

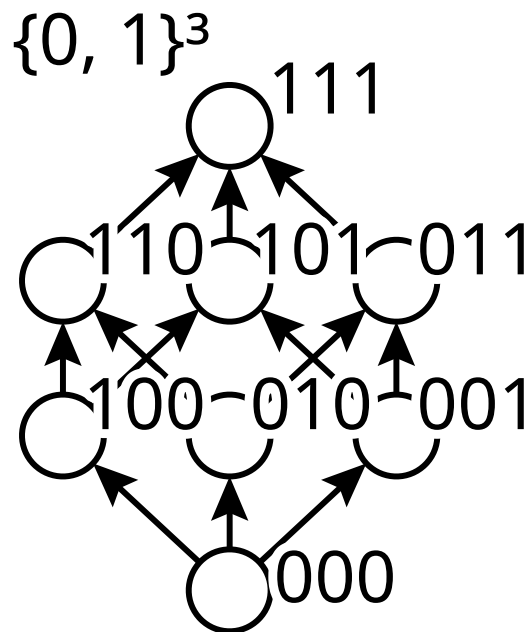


Exponential Family

- The log-linear model on posets belongs to the exponential family
- θ : Natural parameter
- η : Expectation parameter

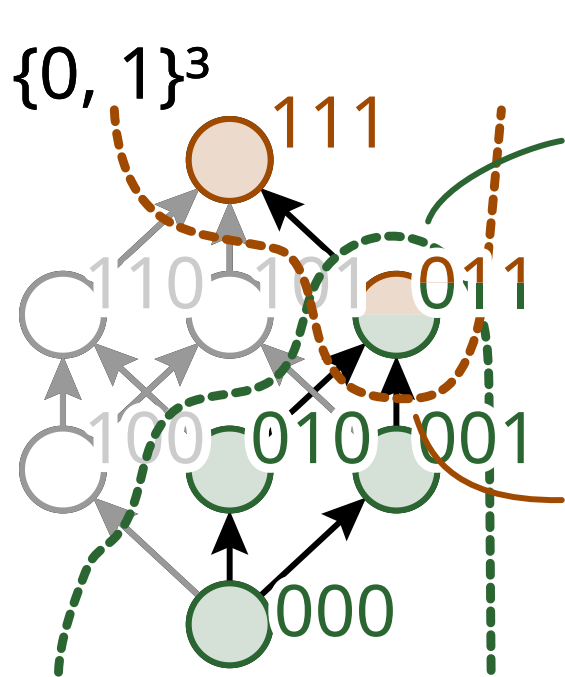
Binary Log-Linear Model

(= Boltzmann machine)



[Luo & Sugiyama, AAAI2019]

Binary Log-Linear Model



(= Boltzmann machine)

$$\log p(x) = -\psi + \sum_i \theta_i x_i + \sum_{i,j} \theta_{ij} x_i x_j + \dots$$
$$= \sum_{s \leq x} \theta_s$$

For x with $x_{i_1} = \dots = x_{i_k} = 1$,

$$\eta_x = \mathbb{E}[x_{i_1} \dots x_{i_k}] = \Pr(x_{i_1} = \dots = x_{i_k} = 1)$$
$$= \sum_{s \geq x} p(s)$$

[Luo & Sugiyama, AAAI2019]

Dually Flat Structure

- Let $\psi(\theta) = -\theta(\perp)$ (convex, partition function)

$$\psi(\theta) \xrightarrow{\text{Legendre transformation}} \phi(\eta) = \sum_{x \in S} p(x) \log p(x)$$

- $(\psi(\theta), \phi(\eta))$ leads to dually flat coordinate system (θ, η) :

$$\nabla \psi(\theta) = \eta, \quad \frac{\partial}{\partial \theta_x} \psi(\theta) = \eta_x$$

$$\nabla \phi(\eta) = \theta, \quad \frac{\partial}{\partial \eta_x} \phi(\eta) = \theta_x$$

Riemannian Metric (Fisher Information)

$$\frac{\partial}{\partial \theta_x} \frac{\partial}{\partial \theta_y} \psi(\theta) = \frac{\partial}{\partial \theta_x} \eta_y = \sum_{s \in S} \zeta(x, s) \zeta(y, s) p(s) - \eta_x \eta_y$$

$$\frac{\partial}{\partial \eta_x} \frac{\partial}{\partial \eta_y} \phi(\eta) = \frac{\partial}{\partial \eta_x} \theta_y = \sum_{s \in S} \mu(s, x) \mu(s, y) p(s)^{-1}$$

$$\mathbb{E}_s \left[\frac{\partial}{\partial \theta_x} \log p(s) \frac{\partial}{\partial \eta_y} \log p(s) \right] = \delta(x, y)$$

Riemannian Metric (Fisher Information)

$$\mathbb{E}_s \left[\frac{\partial}{\partial \theta_x} \log p(s) \frac{\partial}{\partial \theta_y} \log p(s) \right] = \sum_{s \in S} \zeta(x, s) \zeta(y, s) p(s) - \eta_x \eta_y$$

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

$$\mathbb{E}_s \left[\frac{\partial}{\partial \theta_x} \log p(s) \frac{\partial}{\partial \eta_y} \log p(s) \right] = \delta(x, y)$$

Mixed Coordinate System

- Many problems are formulated as coordinate mixing

$$\begin{aligned} P &= (\theta_1, \theta_2, \dots, \theta_{i-1}, \theta_i, \theta_{i+1}, \dots, \theta_n) \\ Q &= (\eta_1, \eta_2, \dots, \eta_{i-1}, \theta_i, \theta_{i+1}, \dots, \theta_n) \\ R &= (\eta_1, \eta_2, \dots, \eta_{i-1}, \eta_i, \eta_{i+1}, \dots, \eta_n) \end{aligned}$$

$\left. \begin{array}{l} \text{e-projection} \\ \text{(MLE)} \end{array} \right\}$

$\left. \begin{array}{l} \text{m-projection} \end{array} \right\}$

Pythagorean theorem: (Q is always unique)

$$\text{KL}(P, R) = \text{KL}(P, Q) + \text{KL}(Q, R)$$

Mixed Coordinate System (Example)

- Many problems are formulated as coordinate mixing

$$P = (\underset{\text{green}}{0} , \underset{\text{green}}{0} , \dots, \underset{\text{green}}{0} , \underset{\text{green}}{0}, \underset{\text{green}}{0} , \dots, \underset{\text{green}}{0}) \rightarrow \text{Uniform dist.}$$

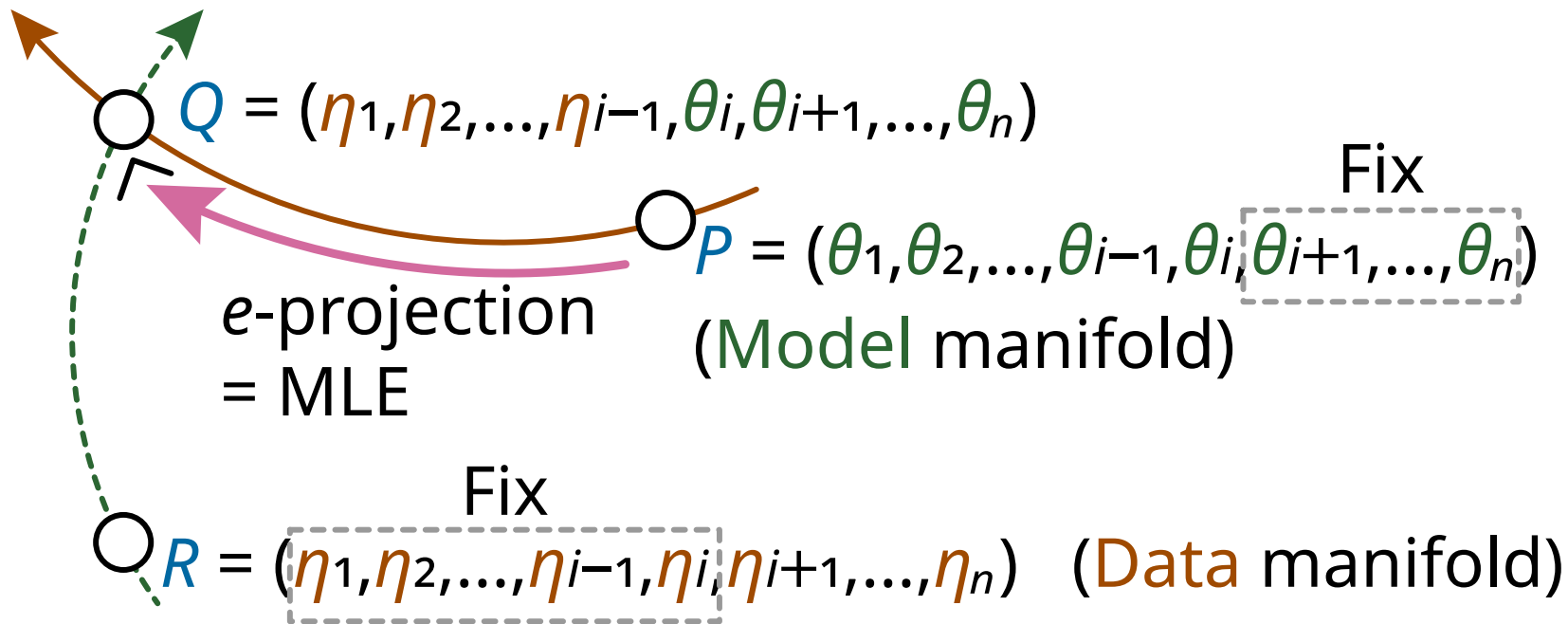
$$Q = (\underset{\text{orange}}{\hat{\eta}_1}, \underset{\text{orange}}{\hat{\eta}_2}, \dots, \underset{\text{orange}}{\hat{\eta}_{i-1}}, \underset{\text{green}}{0}, \underset{\text{green}}{0} , \dots, \underset{\text{green}}{0})$$

$$R = (\underset{\text{orange}}{\hat{\eta}_1}, \underset{\text{orange}}{\hat{\eta}_2}, \dots, \underset{\text{orange}}{\hat{\eta}_{i-1}}, \underset{\text{orange}}{\hat{\eta}_i}, \underset{\text{orange}}{\hat{\eta}_{i+1}}, \dots, \underset{\text{orange}}{\hat{\eta}_n}) \rightarrow \text{Empirical dist.}$$

Pythagorean theorem: $(Q \text{ is always unique})$

$$\text{KL}(\underset{\text{blue}}{P}, \underset{\text{blue}}{R}) = \text{KL}(\underset{\text{blue}}{P}, \underset{\text{blue}}{Q}) + \text{KL}(\underset{\text{blue}}{Q}, \underset{\text{blue}}{R})$$

Two Submanifolds



Gradient methods for e-projection

- e-projection is convex optimization

- Gradient descent (first-order):

$$\theta_{\text{next}} \leftarrow \theta - \varepsilon(\eta - \hat{\eta}_{\text{target}})$$

- Natural gradient (second-order)

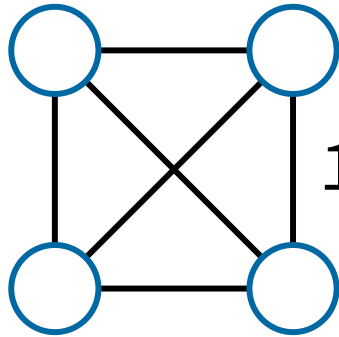
$$\theta_{\text{next}} \leftarrow \theta - G^{-1}(\eta - \hat{\eta}_{\text{target}})$$

– G is Fisher information matrix w.r.t. θ

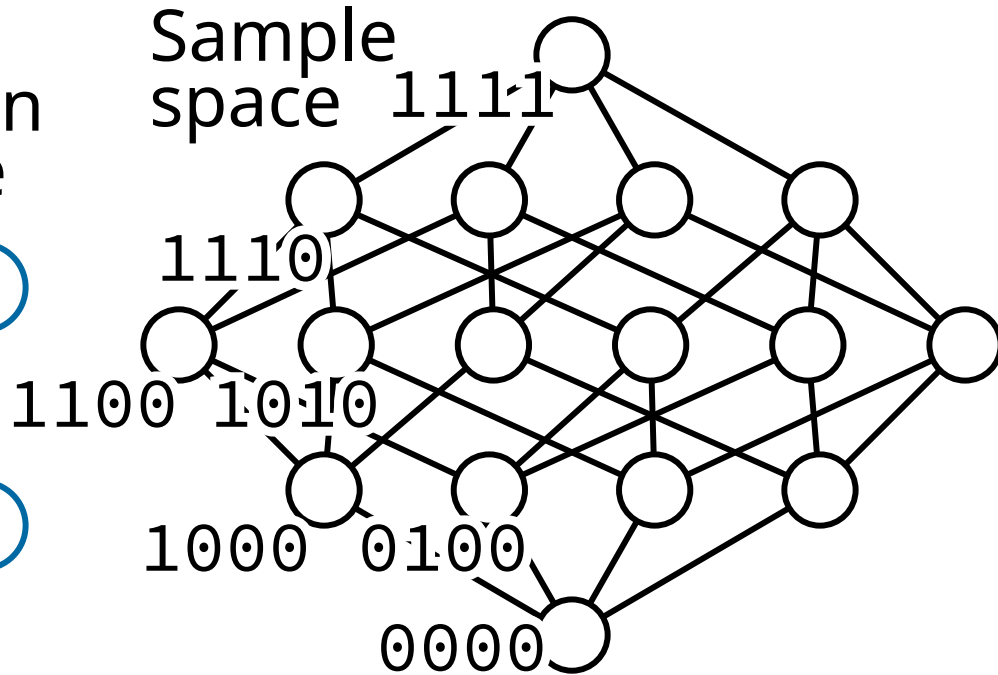
- Coordinate descent [Hayashi, Sugiyama, Matsushima, DSAA2020]

Boltzmann Machine Training

Boltzmann
machine

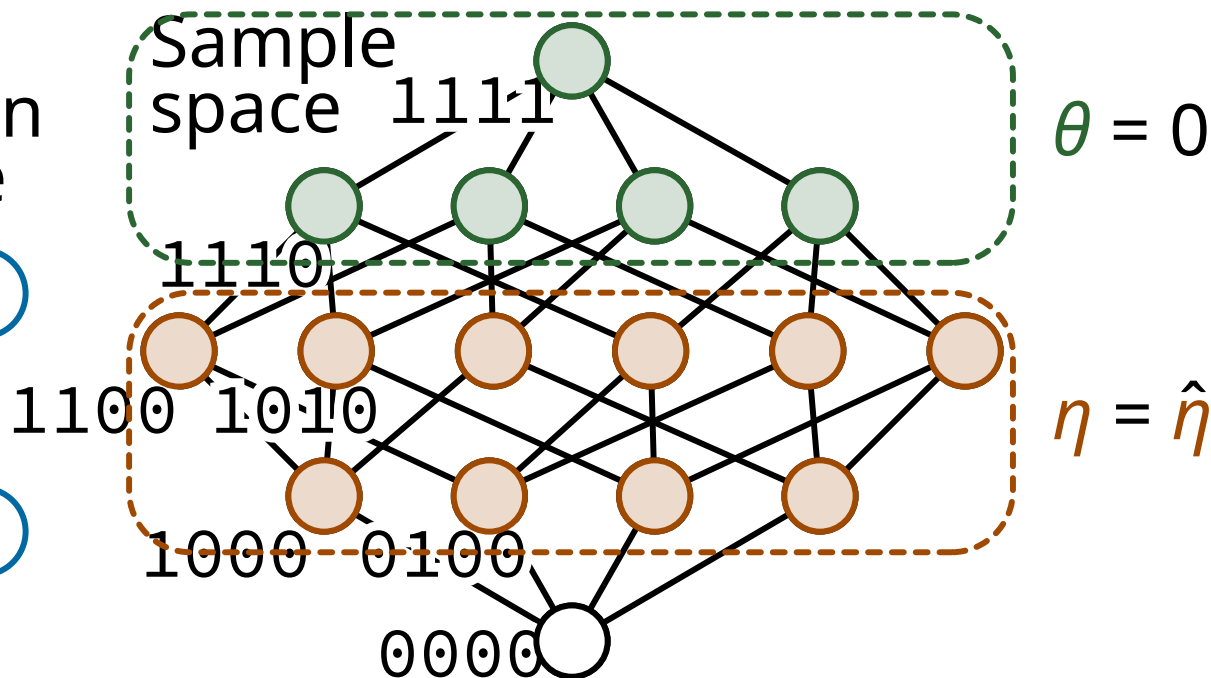
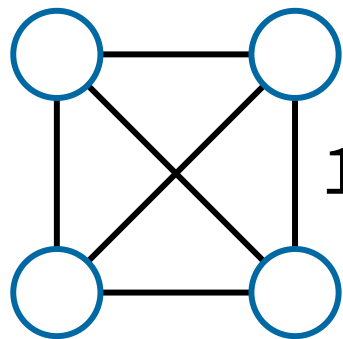


Sample
space



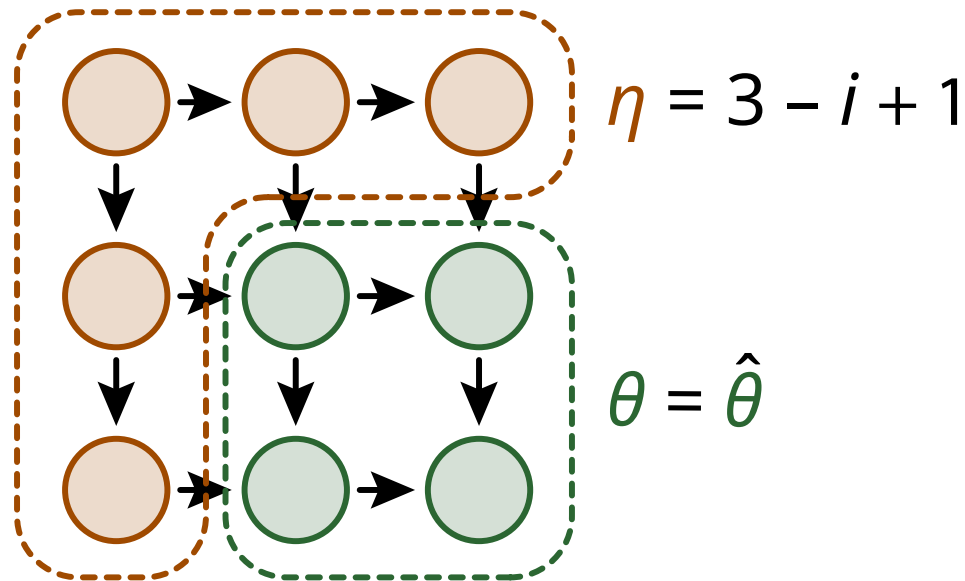
Boltzmann Machine Training

Boltzmann machine



Matrix Balancing

3x3 matrix
as poset:



Summary

- Information geometric formulation for partial order structures
 - Learning process can be achieved as a projection in the parameter space (dually flat manifold)
- Several applications
 - Boltzmann machines
 - Matrix (Tensor) balancing