November 10, 2025

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

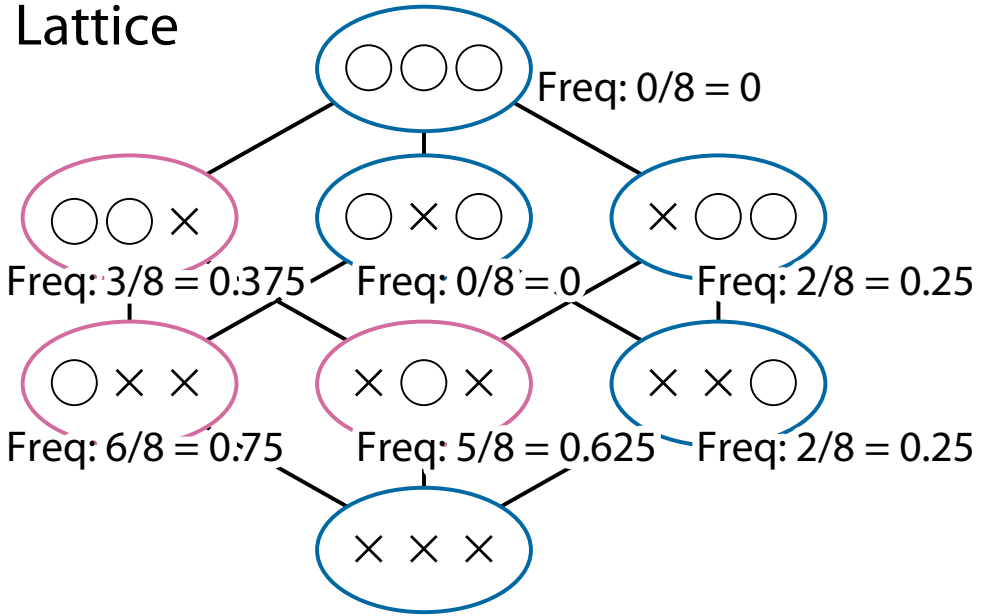
Mahito Sugiyama

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

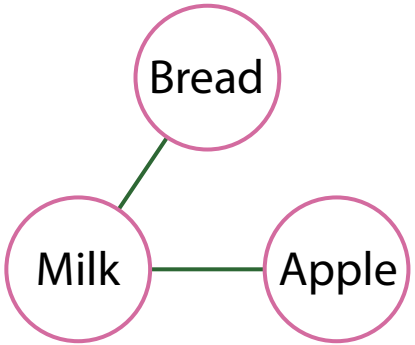


Learning Hierarchical Distribution (2/2)

MLE

log(prob.) = –10.41 + 9.43[Bread] + 8.52[Milk] – 9.84[Apple]
– 9.03[Bread&Milk] + 9.43[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)

$$\begin{bmatrix} r_1 & 0 \\ 0 & r_2 \end{bmatrix} \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix} \begin{bmatrix} s_1 & 0 \\ 0 & s_2 \end{bmatrix}$$

$$= \begin{bmatrix} r_1 s_1 p_{11} & r_1 s_2 p_{12} \\ r_2 s_1 p_{21} & r_2 s_2 p_{22} \end{bmatrix} \begin{matrix} \rightarrow \sum_j r_1 s_j p_{1j} = 1 \\ \rightarrow \sum_j r_2 s_j p_{2j} = 1 \end{matrix}$$

$$\downarrow \qquad \downarrow$$

$$\sum_i r_i s_1 p_{i1} = 1 \qquad \sum_i r_i s_2 p_{i2} = 1$$

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]

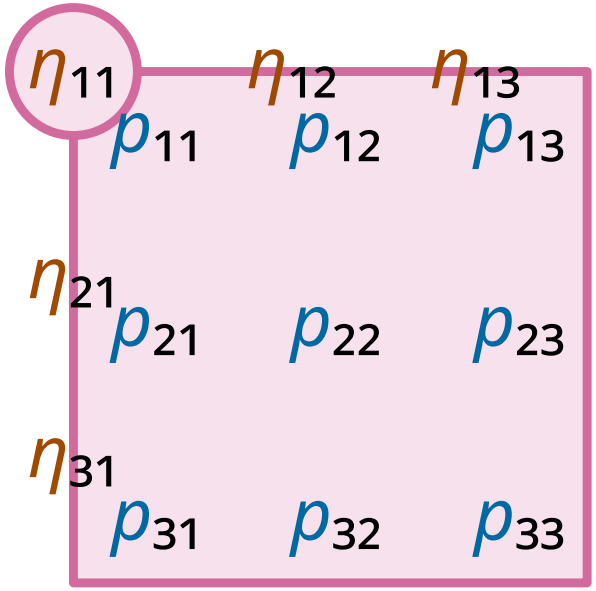
Introduce η

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

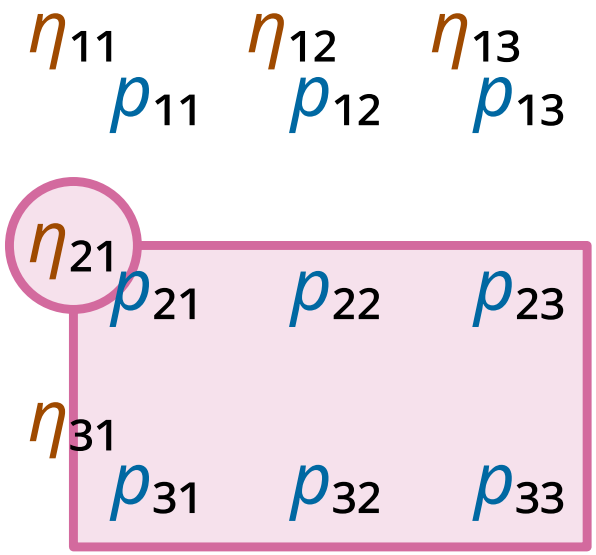
η_{21}
p_{21} p_{22} p_{23}

η_{31}
p_{31} p_{32} p_{33}

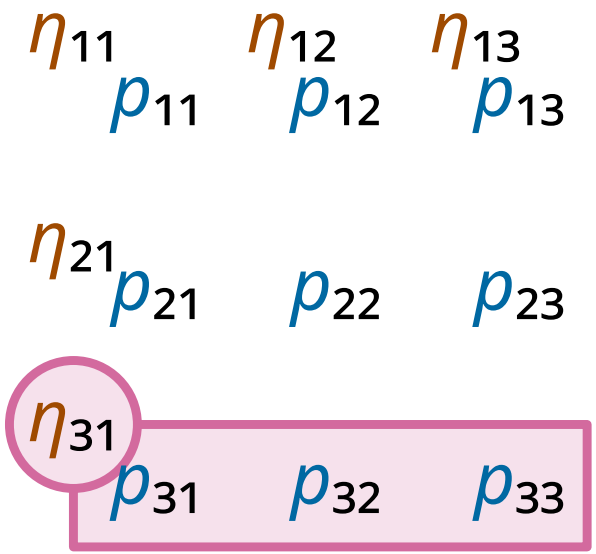
Introduce η



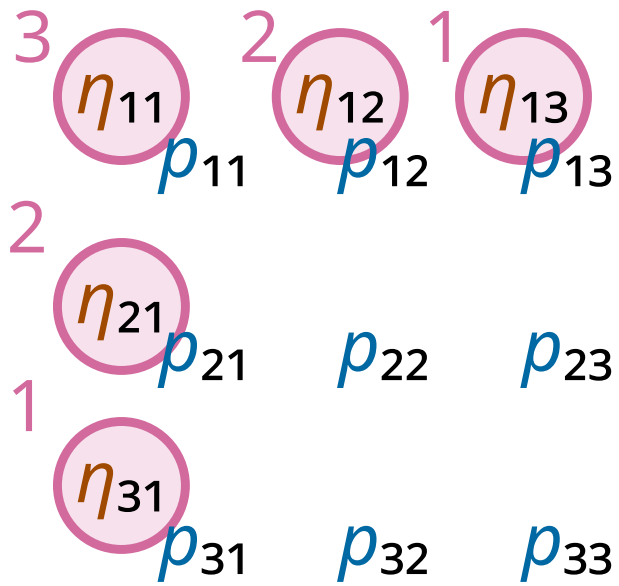
Introduce η



Introduce η



Introduce η



Introduce θ

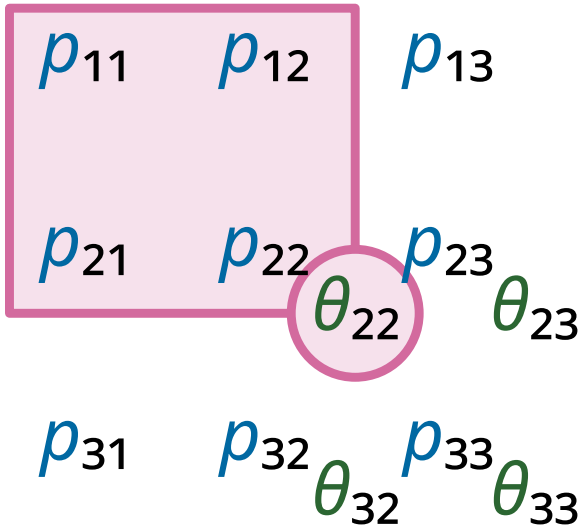
$$p_{11} \quad p_{12} \quad p_{13}$$

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

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

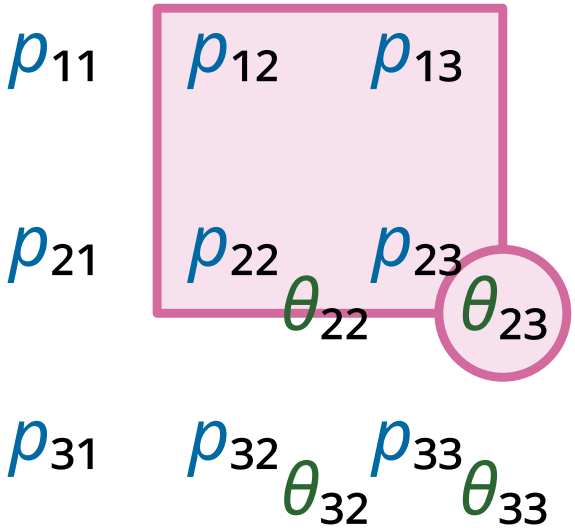
$$\theta_{ij} = \log p_{ij} - \log p_{i-1j} - \log p_{ij-1} + \log p_{i-1j-1}$$

Introduce θ



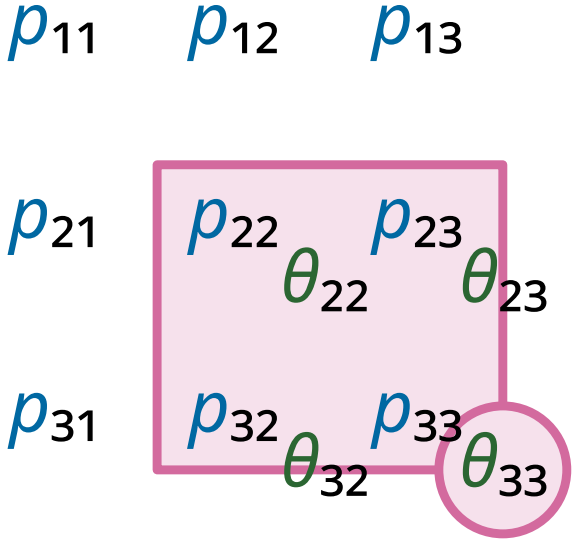
$$\theta_{ij} = \log p_{ij} - \log p_{i-1j} - \log p_{ij-1} + \log p_{i-1j-1}$$

Introduce θ



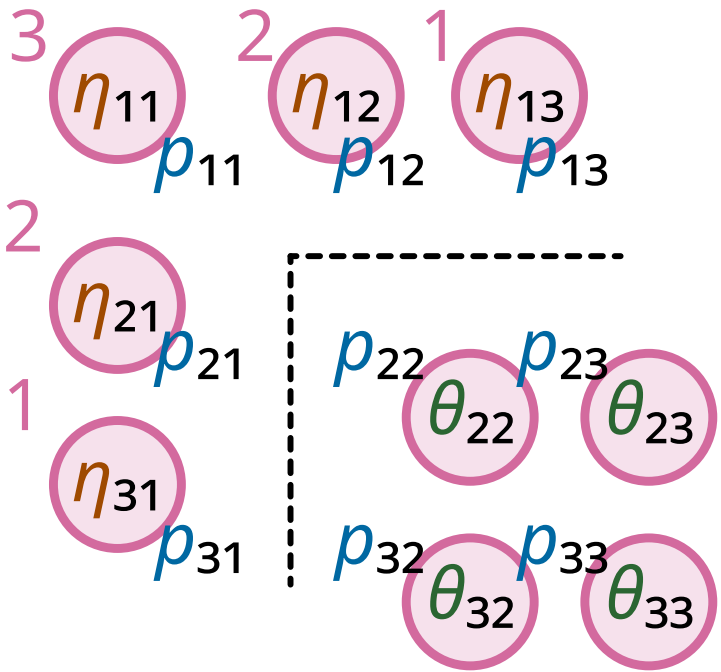
$$\theta_{ij} = \log p_{ij} - \log p_{i-1j} - \log p_{ij-1} + \log p_{i-1j-1}$$

Introduce θ



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

Balancing as Constraints on η and θ



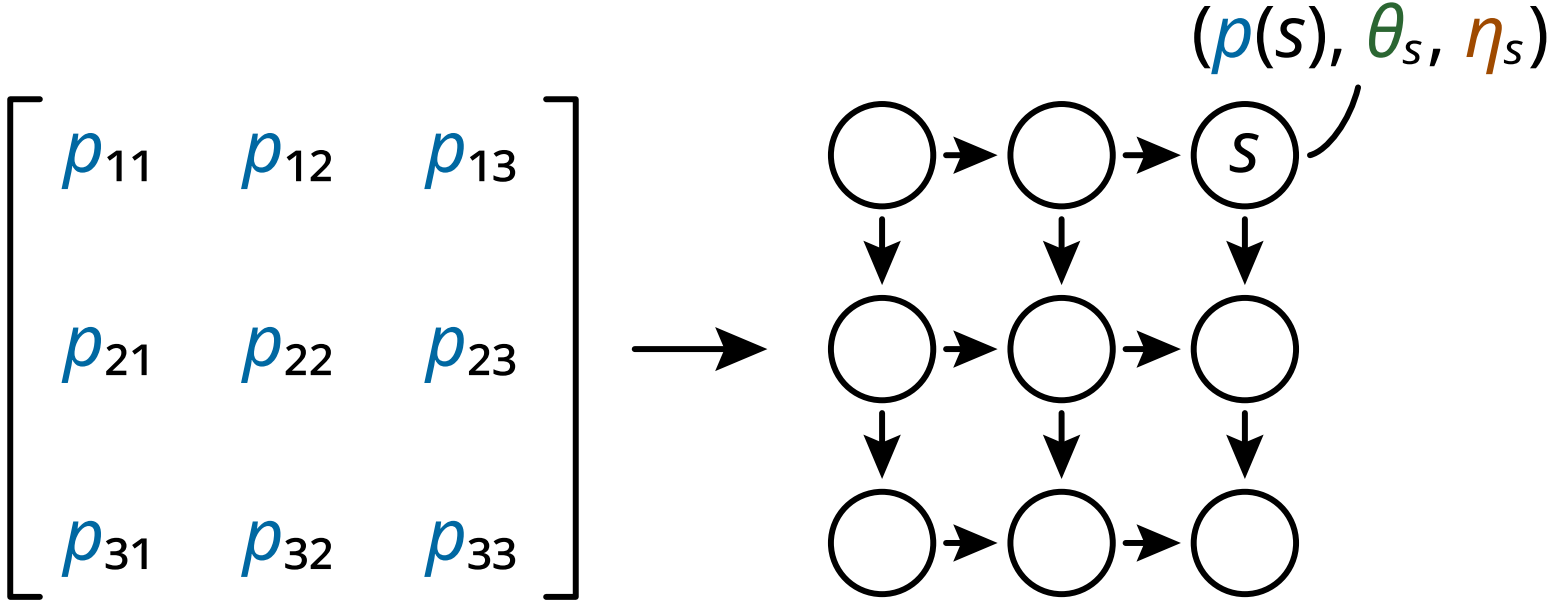
$$\theta_{ij} = \log p_{ij} - \log p_{i-1j} - \log p_{ij-1} + \log p_{i-1j-1}$$

Matrix balancing $\Leftrightarrow$

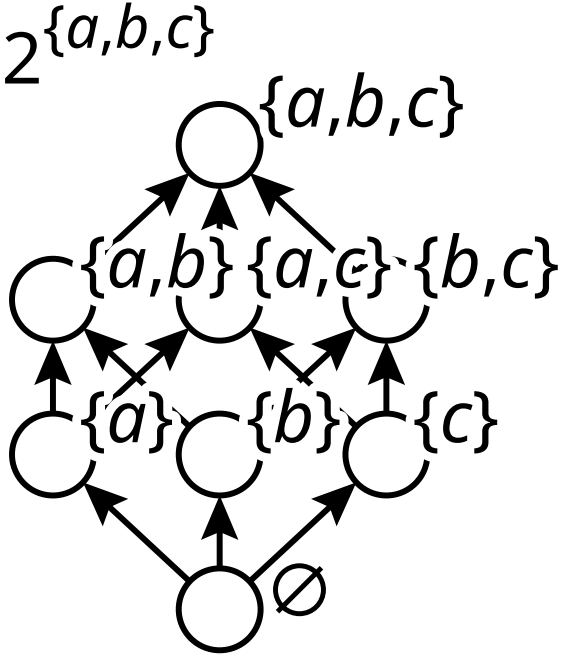
Satisfy $\eta_{i1} = \eta_{1i} = 3 - i + 1$

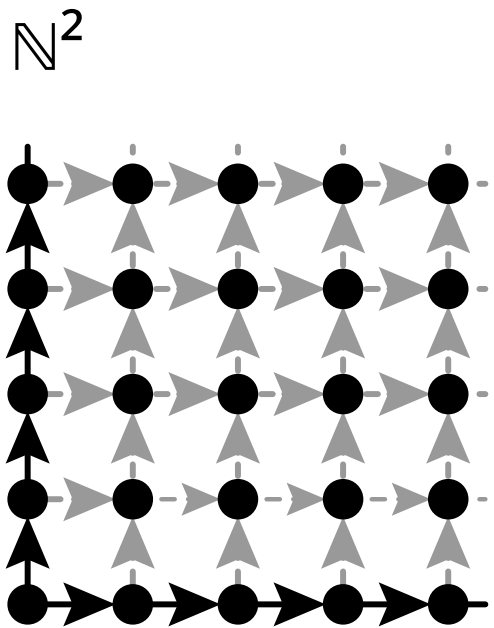
with keeping all θ_{ij}

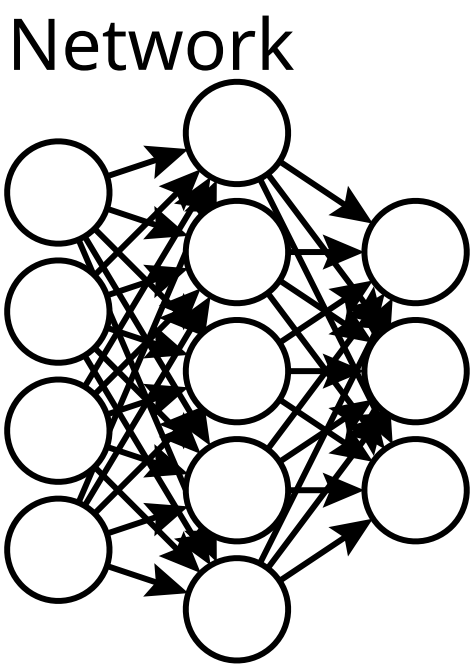
Introduce Partial Order Structure



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 $S = \{s_1, s_2, \ldots, s_n\}$
- 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:

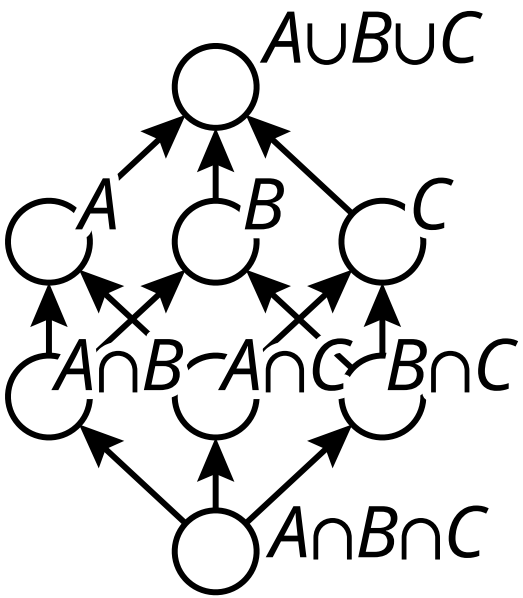
  $\mathbf{Z}^{-1} = \mathbf{M}$ such that $m_{ij} = \mu(s_i, s_j)$
  - $\mathbf{M} \in \mathbb{Z}^{n \times n}$

Möbius Inversion Formula

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

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

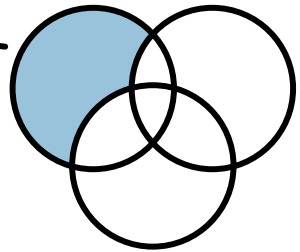
E.g.1: Inclusion-Exclusion Principle



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

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

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



Log-Linear Model on Poset [ICML2017]

- For probability $p: S \to (0, 1)$ with $\sum_{x \in S} p(x) = 1$, introduce $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ as

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

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

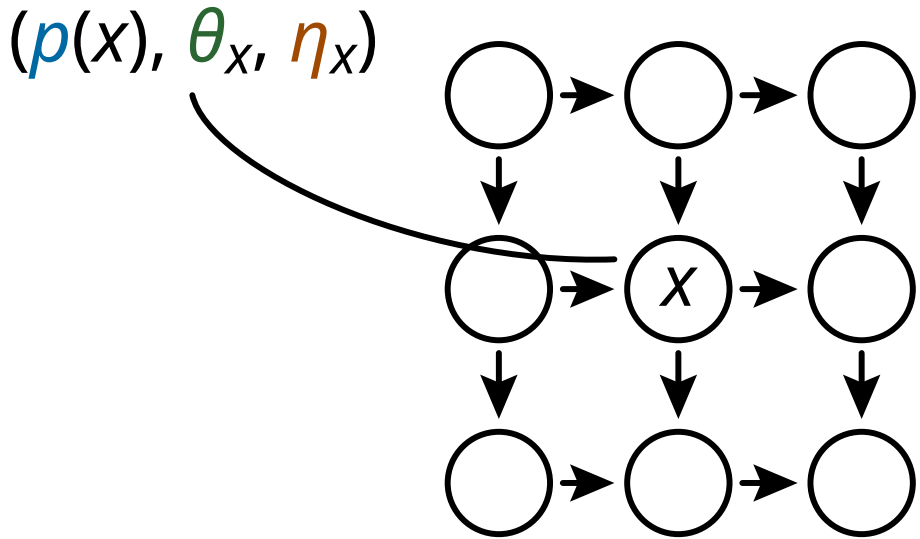
- From the Möbius inversion formula, log-linear model is:

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

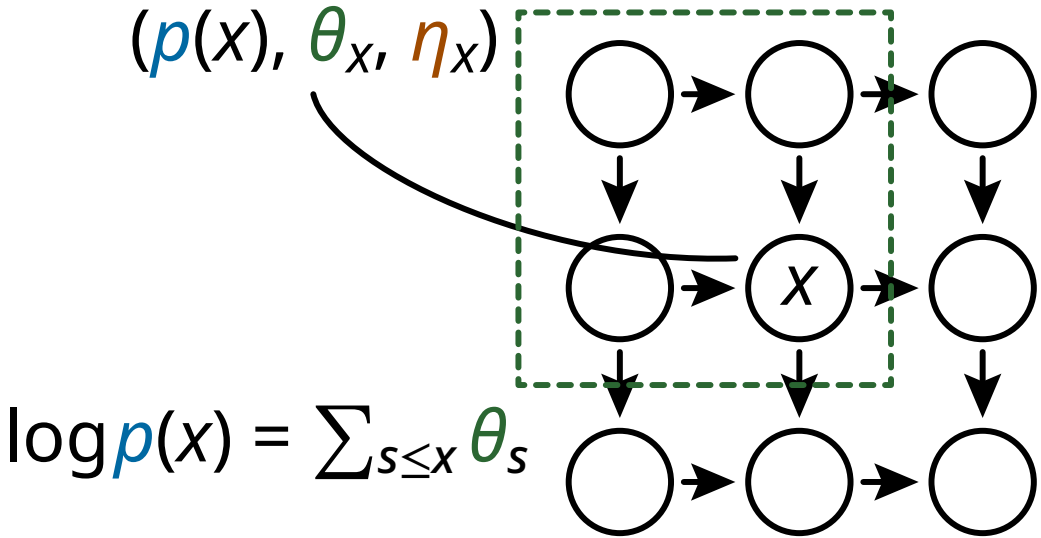
- In the matrix form, it can be written as

  $$\log \boldsymbol{p} = \mathbf{Z} \boldsymbol{\theta}$$

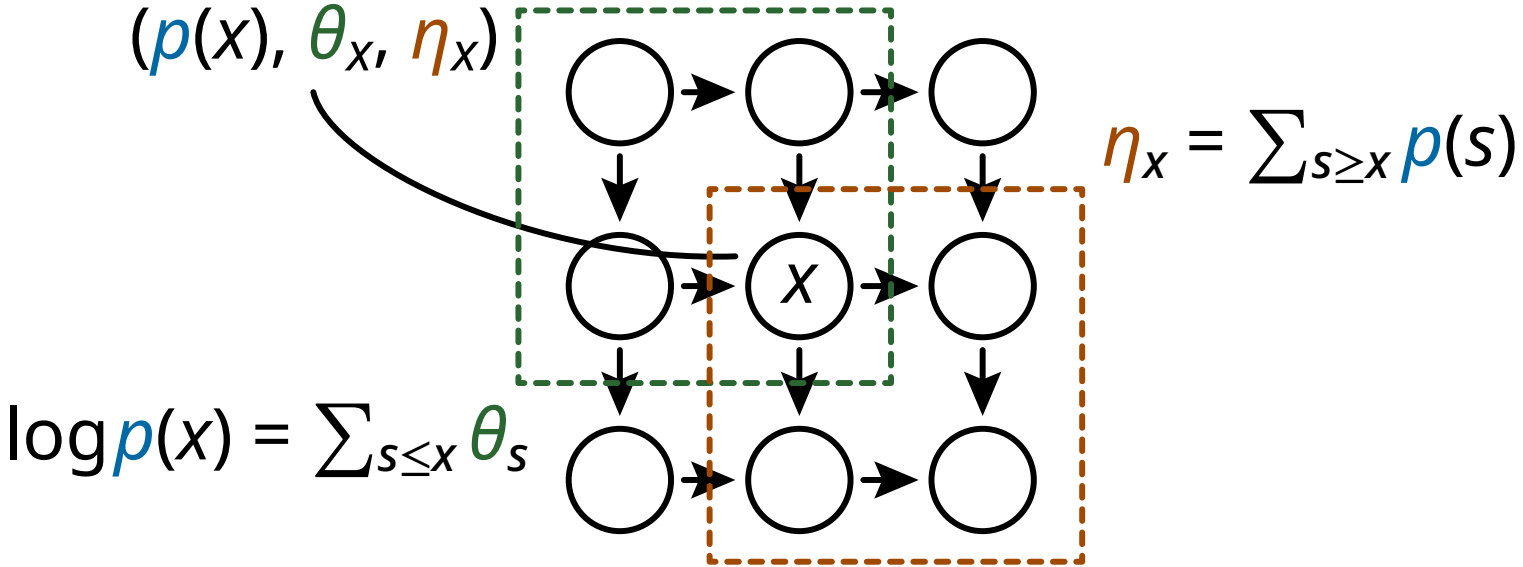
Log-Linear Model on Poset



Log-Linear Model on Poset



Log-Linear Model on Poset

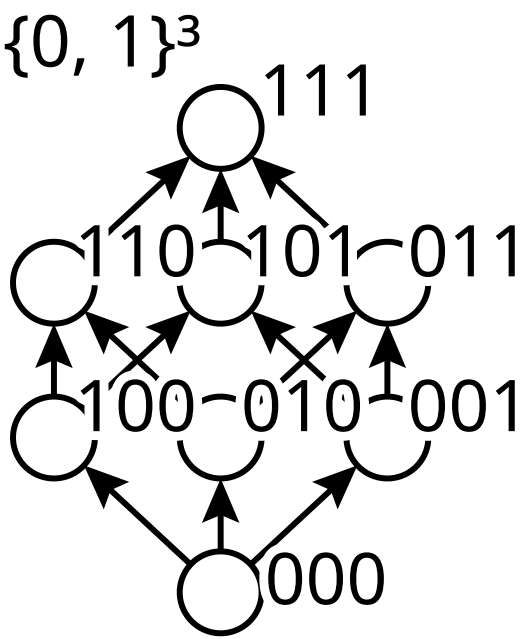


Exponential Family

- The log-linear model on posets belongs to the exponential family

- θ : Natural parameter
- η : Expectation parameter

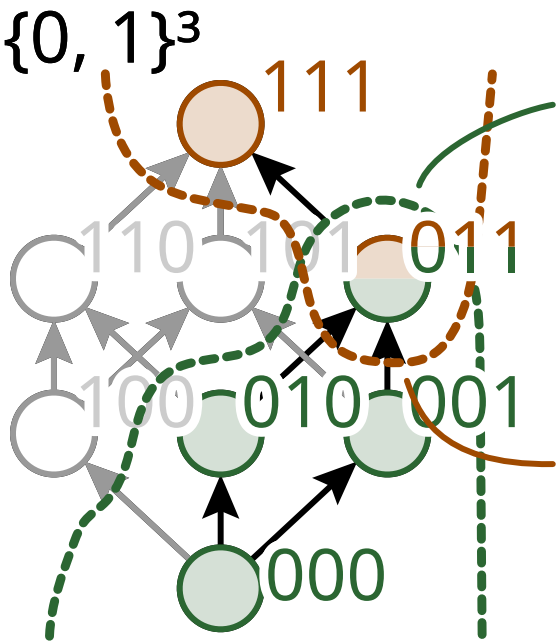
Binary Log-Linear Model



(= Boltzmann machine)

[Luo & Sugiyama, AAAI2019]

Binary Log-Linear Model



(= Boltzmann machine)

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

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

$$\eta_{\boldsymbol{x}} = \mathbb{E}[x_{i_1} \cdots x_{i_k}] = \Pr(x_{i_1} = \cdots = x_{i_k} = 1)$$
$$= \sum_{s \ge x} p(s)$$

[Luo & Sugiyama, AAAI2019]

Dually Flat Structure

- Let $\psi(\theta) = -\theta(\bot)$ (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)

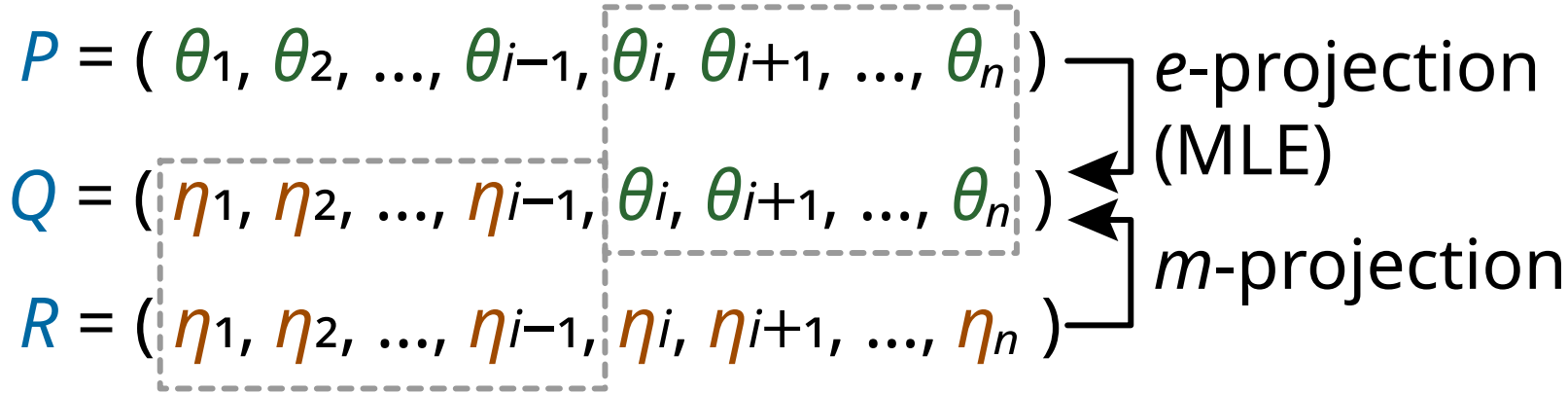
$$\mathbb{E}_s\left[\frac{\partial}{\partial\theta_x}\log p(s)\frac{\partial}{\partial\theta_y}\log p(s)\right] = \sum_{s\in S}\zeta(x,s)\zeta(y,s)p(s) - \eta_x\eta_y$$

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

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

Mixed Coordinate System

- Many problems are formulated as coordinate mixing



Pythagorean theorem: (Q is always unique)

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

Mixed Coordinate System (Example)

- Many problems are formulated as coordinate mixing

$P = (0, 0, \ldots, 0, 0, 0, \ldots, 0) \rightarrow$ Uniform dist.

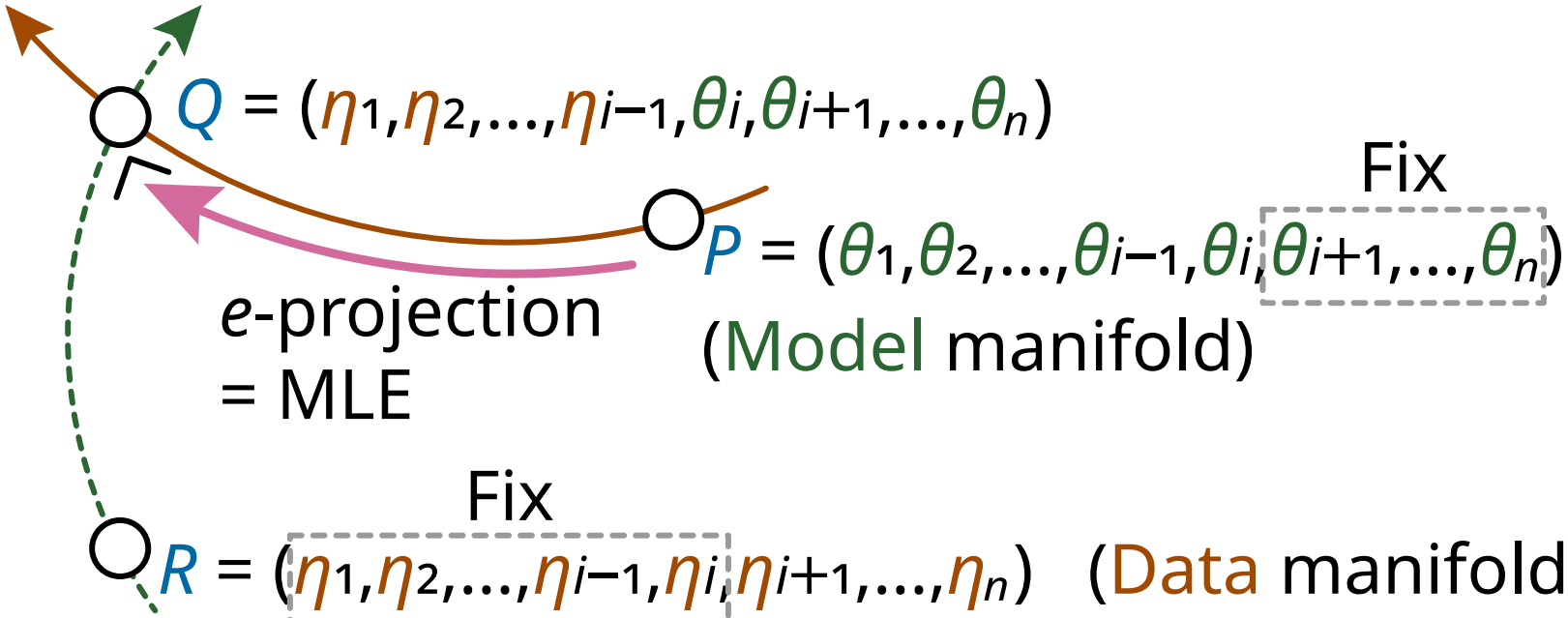
$Q = (\hat{\eta}_1, \hat{\eta}_2, \ldots, \hat{\eta}_{i-1}, 0, 0, \ldots, 0)$

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

Pythagorean theorem: (Q is always unique)

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

Two Submanifolds



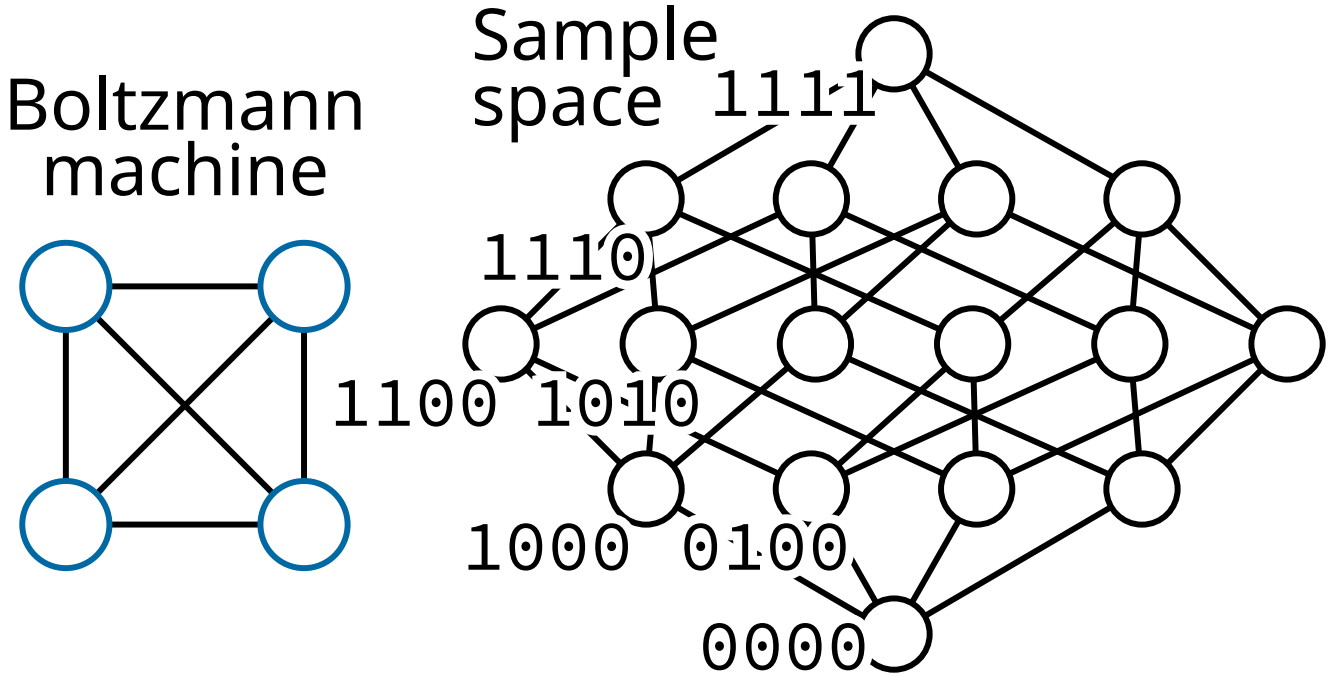
Gradient methods for *e*-projection

- *e*-projection is convex optimization
- Gradient descent (first-order):

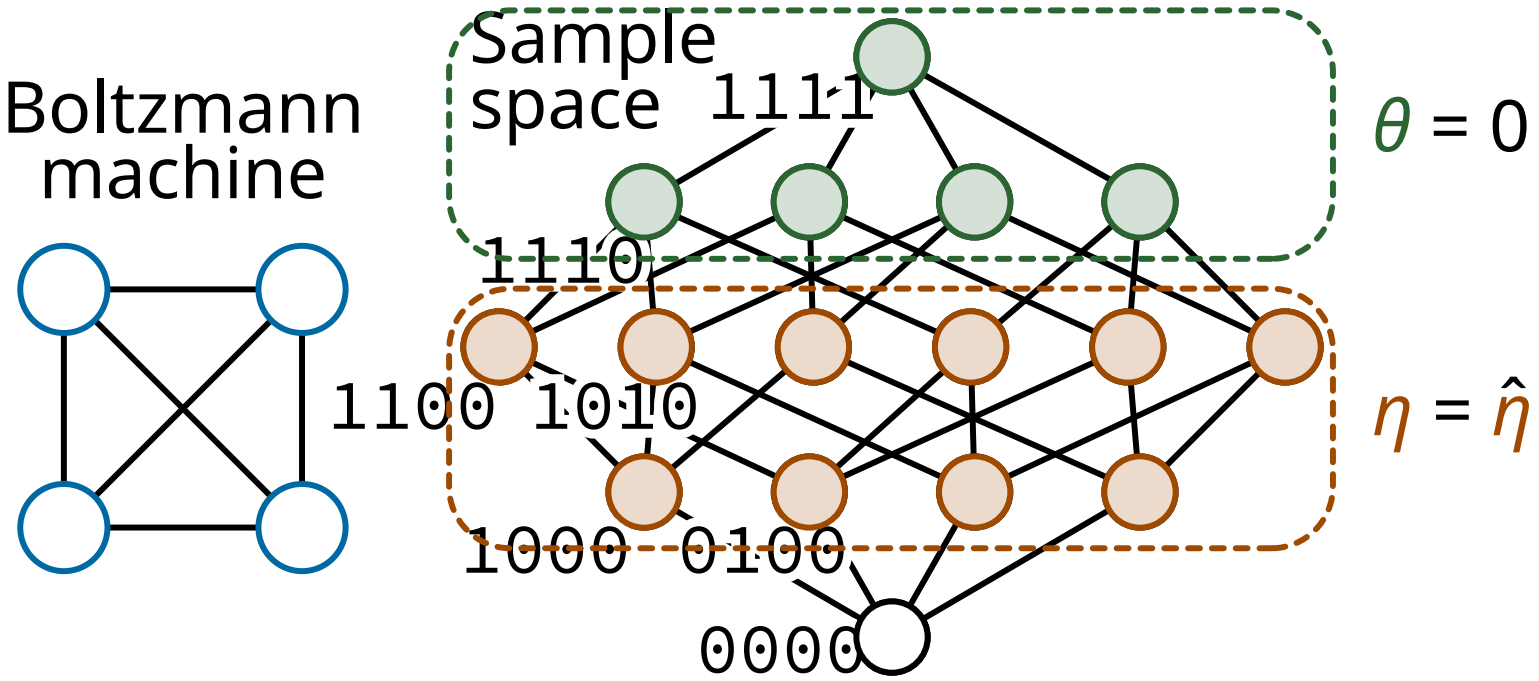
  $\boldsymbol{\theta}_{\text{next}} \leftarrow \boldsymbol{\theta} - \varepsilon(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}_{\text{target}})$
- Natural gradient (second-order)

  $\boldsymbol{\theta}_{\text{next}} \leftarrow \boldsymbol{\theta} - G^{-1}(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}_{\text{target}})$
  - G is Fisher information matrix w.r.t. θ
- Coordinate descent [Hayashi, Sugiyama, Matsushima, DSAA2020]

Boltzmann Machine Training

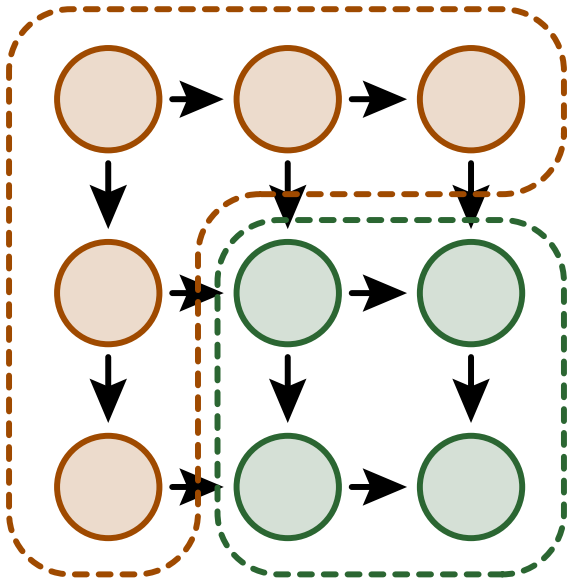


Boltzmann Machine Training



Matrix Balancing

3x3 matrix
as poset:



$\eta = 3 - i + 1$

$\theta = \hat{\theta}$

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