Nov. 20–23, 2019
IBIS 2019

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

PRESTO
SAKIGAKE

隣接代数と双対平坦構造を用いた学習

杉山 麿人（国立情報学研究所，JSTさきがけ研究者）

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)

Revisit Matrix Balancing [ICML2017]

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

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

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

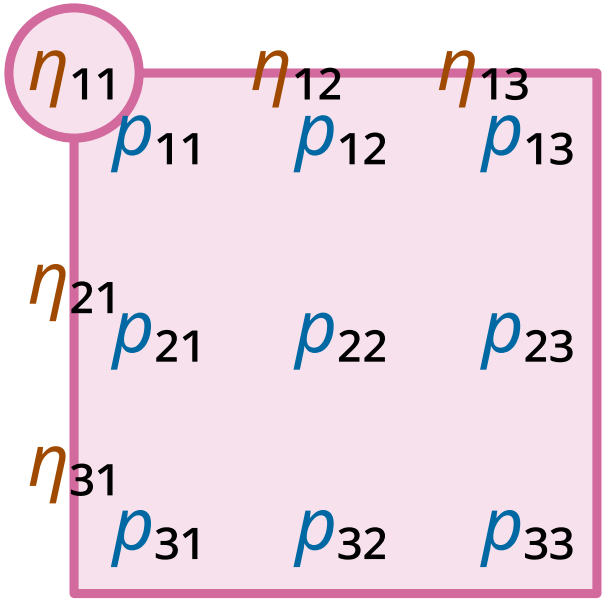
Introduce η

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

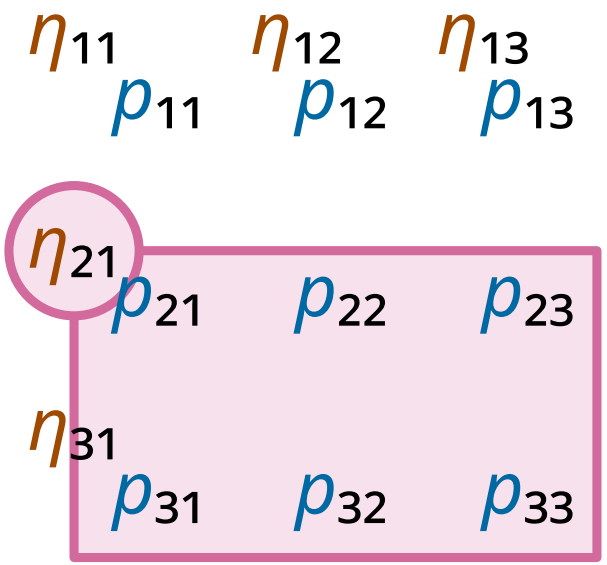
η_{21}
p_{21} p_{22} p_{23}

η_{31}
p_{31} p_{32} p_{33}

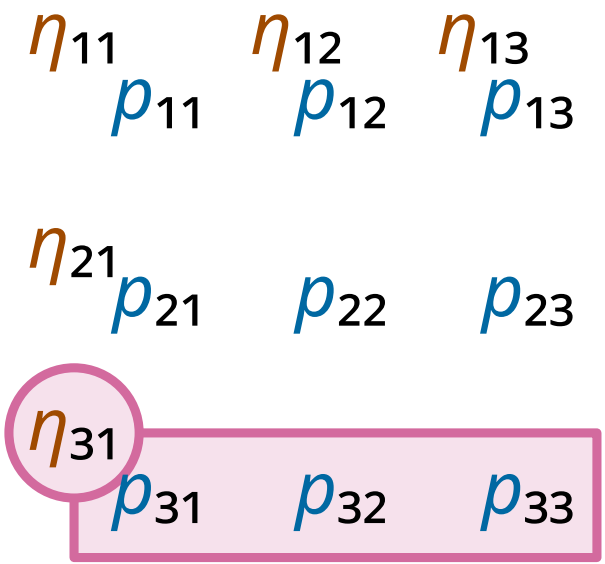
Introduce η



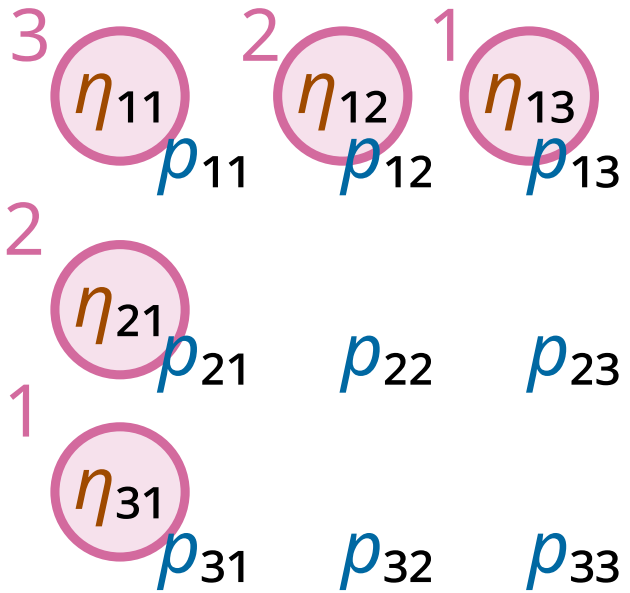
Introduce η



Introduce η



Introduce η

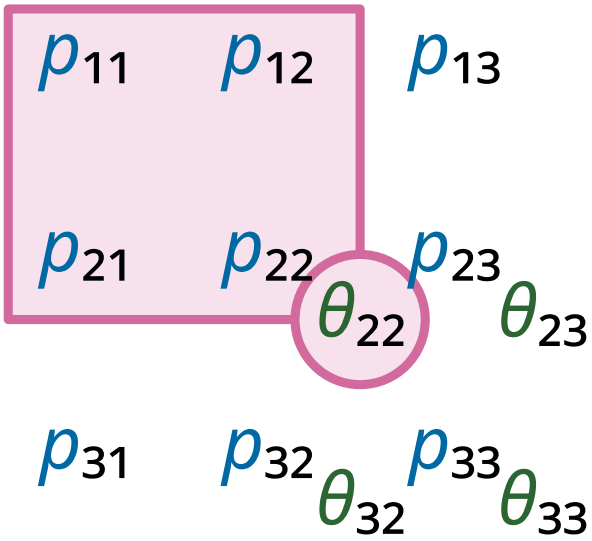


Introduce θ

$$\begin{array}{ccc} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ & \theta_{22} & \theta_{23} \\ p_{31} & p_{32} & p_{33} \\ & \theta_{32} & \theta_{33} \end{array}$$

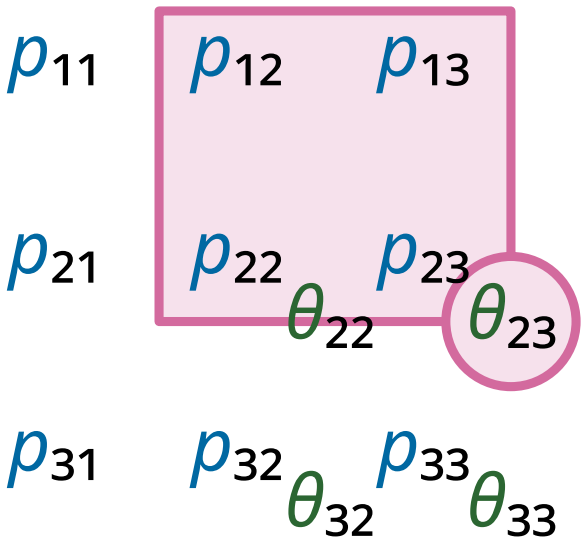
$$\theta_{ij} = \log p_{ij} - \log p_{i-1j} - \log p_{ij-1} + \log p_{i-1j-1}$$

Introduce θ



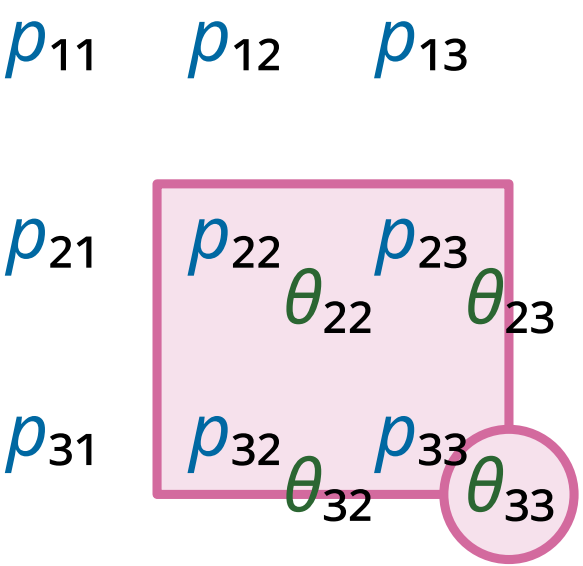
$$\theta_{ij} = \log p_{ij} - \log p_{i-1j} - \log p_{ij-1} + \log p_{i-1j-1}$$

Introduce θ



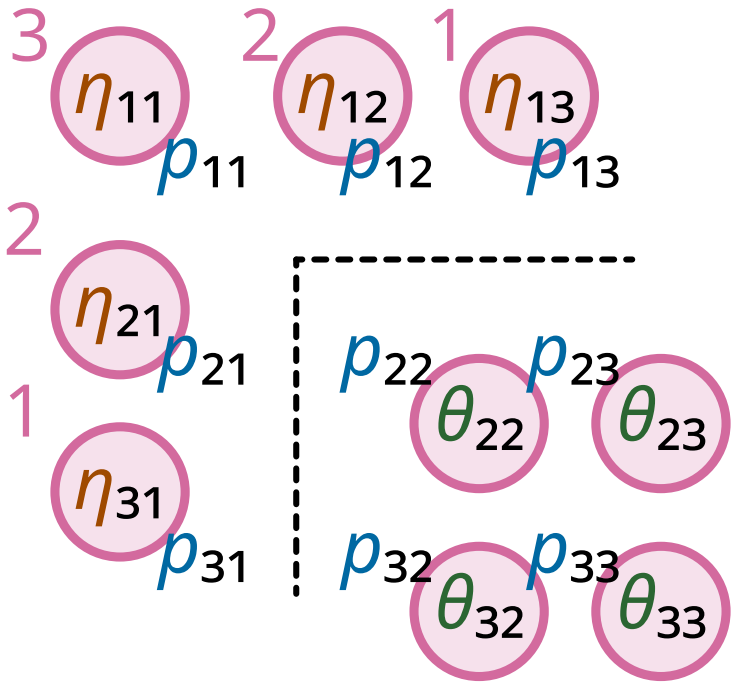
$$\theta_{ij} = \log p_{ij} - \log p_{i-1j} - \log p_{ij-1} + \log p_{i-1j-1}$$

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 $\Leftrightarrow$

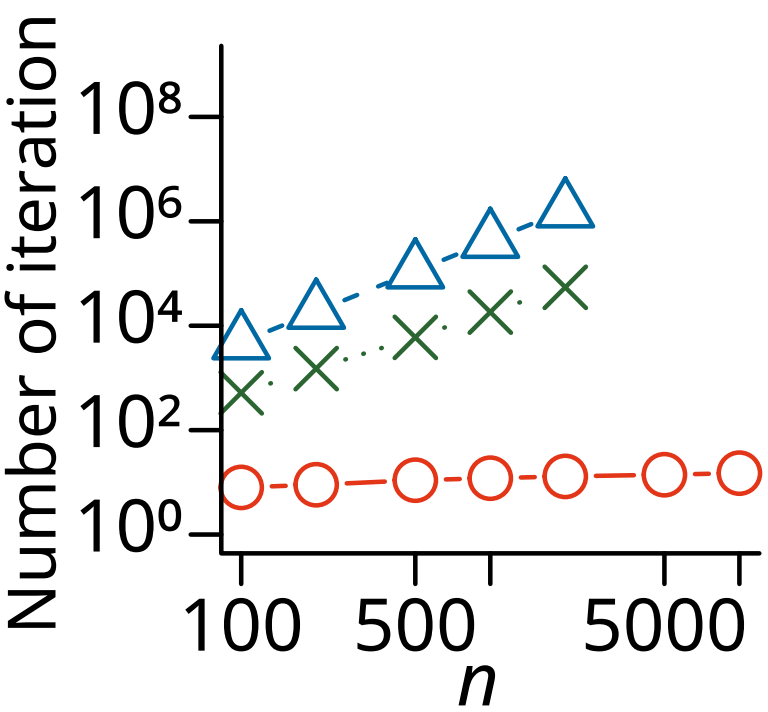
Satisfy $\eta_{i1} = \eta_{1i} = 3 - i + 1$

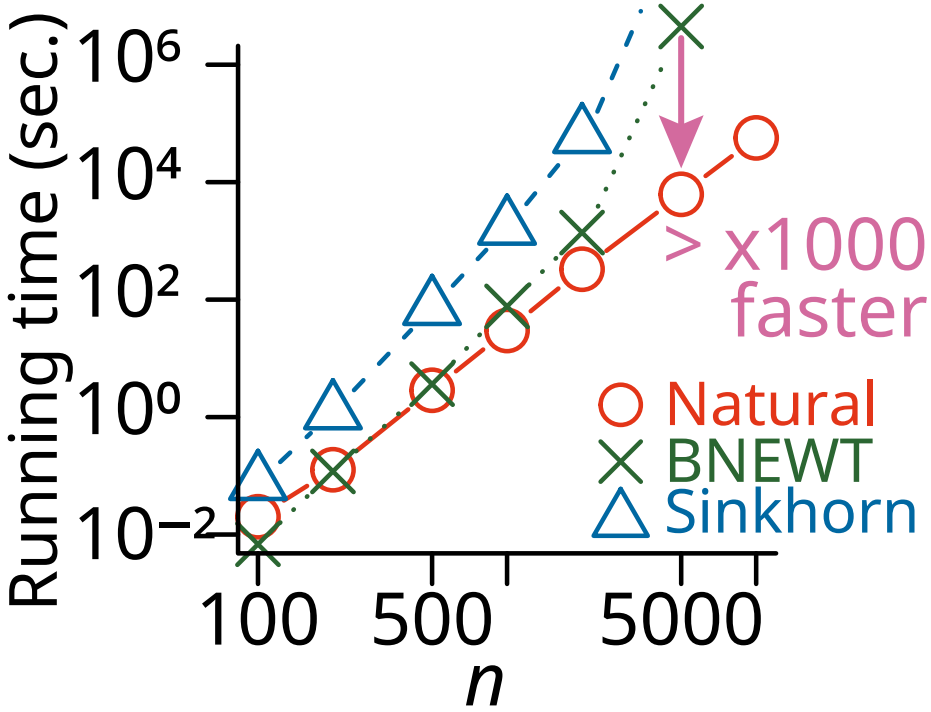
with keeping all θ_{ij}

Natural Gradient

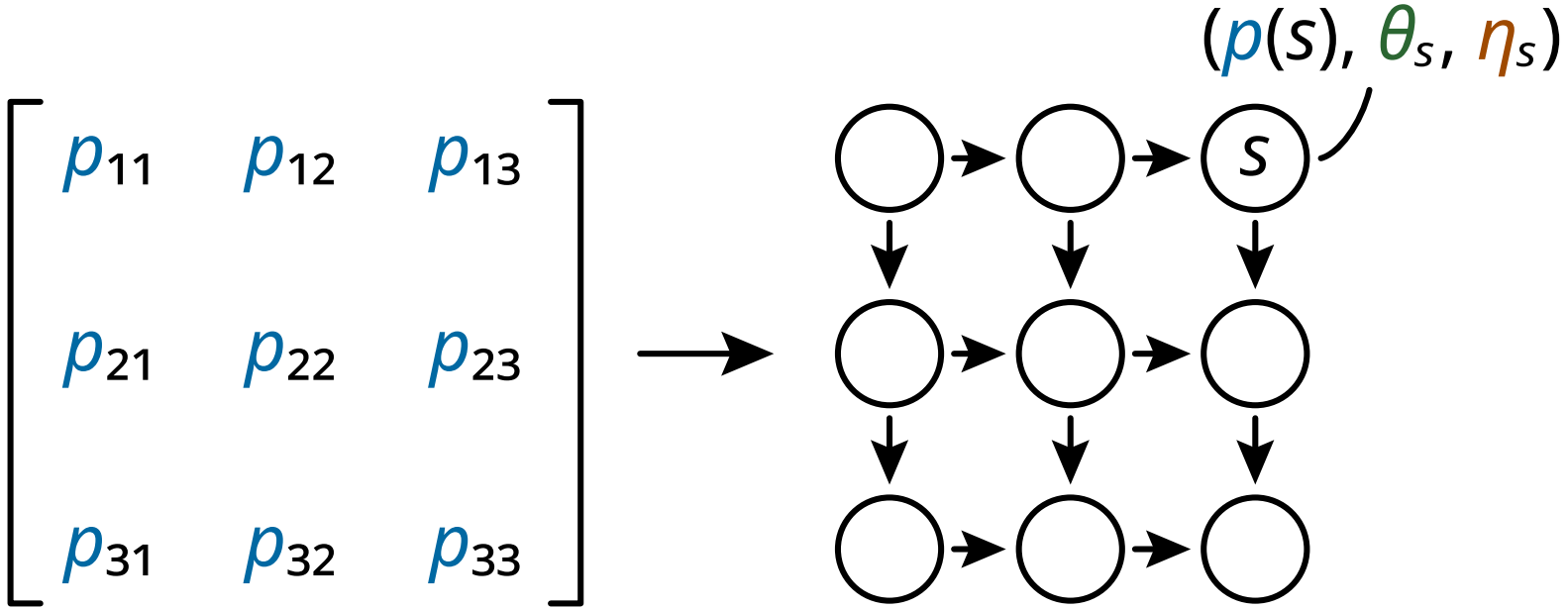
- Given $P \in \mathbb{R}^{n \times n}$, introduce (θ, η) as
  $\log p_{ij} = \sum_{k \le i} \sum_{l \le j} \theta_{kl}, \quad \eta_{ij} = \sum_{k \ge i} \sum_{l \ge j} p_{kl}$
- Let $I = \{11, 12, \dots, 1n, 21, \dots, n1\}$, $\theta = (\theta_i)_{i \in I}^T$, $\eta = (\eta_i)_{i \in I}^T$
- Using Fisher information matrix $G \in \mathbb{R}^{|I| \times |I|}$ s.t.
  $g_{(ij)(kl)} = \eta_{\max\{i,k\} \max\{j,l\}} - \eta_{ij} \eta_{kl}$, update formula is
  $\theta_{\text{next}} = \theta - G^{-1}(\eta - \eta_{\text{correct}})$

Results on Hessenberg Matrix

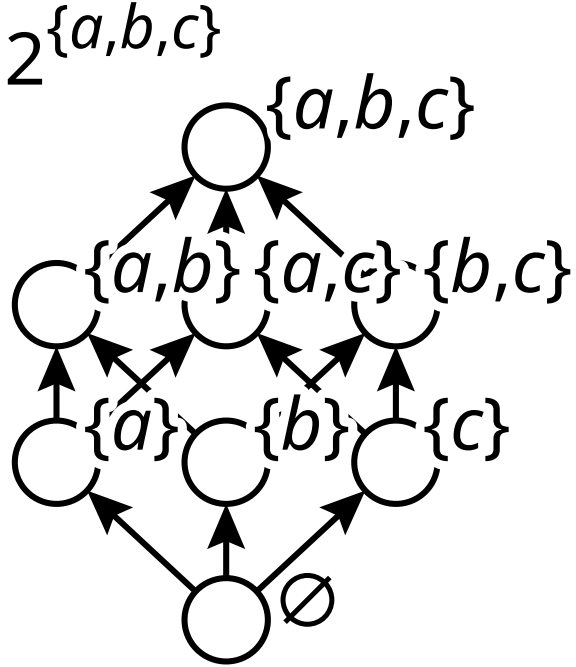


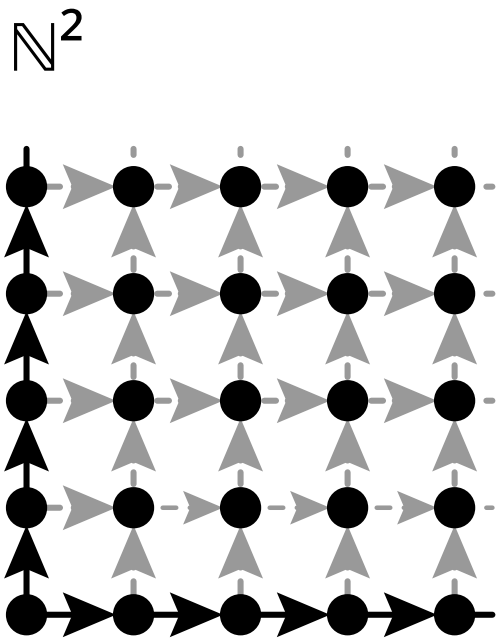


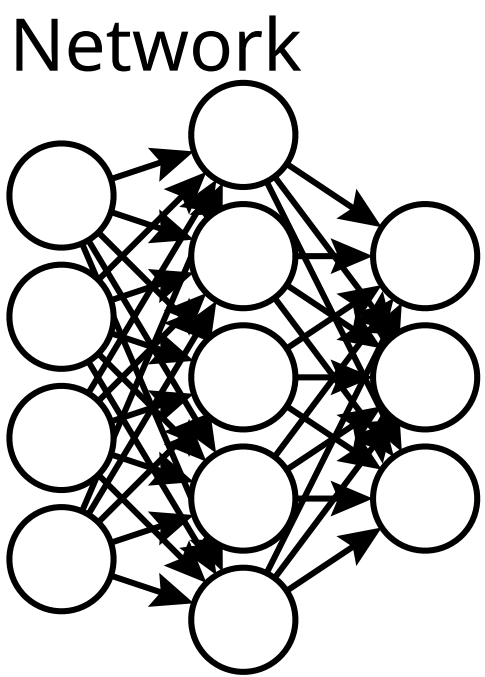
Introduce Partial Order Structure



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 $f(a, b)$ from intervals $[a, b]$ to a scalar with

$$(f + g)(a, b) = f(a, b) + g(a, b)$$

$$(fg)(a, b) = \sum_{a \leq x \leq b} f(a, x)g(x, b) \quad \text{(convolution)}$$

Analogy to Matrix Multiplication

- For $[a, b]$, define $\boldsymbol{f}, \boldsymbol{g} \in \mathbb{R}^{|[a,b]|}$ as

$$\boldsymbol{f} = \begin{bmatrix} f(a,a) \\ \vdots \\ f(a,b) \end{bmatrix}, \quad \boldsymbol{g} = \begin{bmatrix} g(a,a) \\ \vdots \\ g(b,a) \end{bmatrix}$$

- For f and g,

$$(f g)(a,b) = \boldsymbol{f}^T \boldsymbol{g}$$

Special Elements

- Delta function δ:

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

- Zeta function ζ: (integral)

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

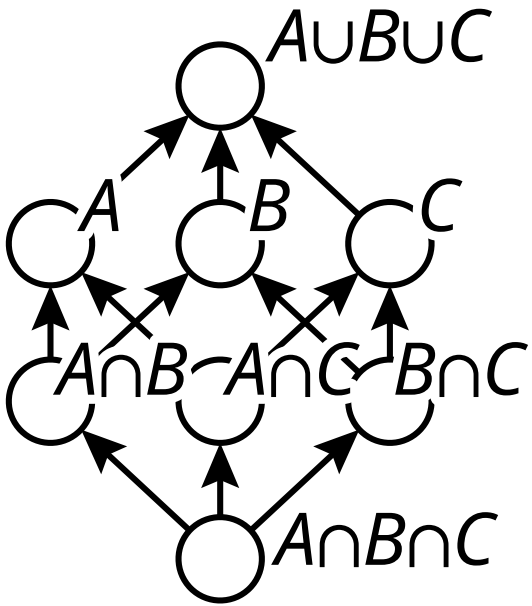
- Möbius function $\mu = \zeta^{-1}$: $\zeta\mu = \delta$ (differential)

Möbius Inversion Formula

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

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

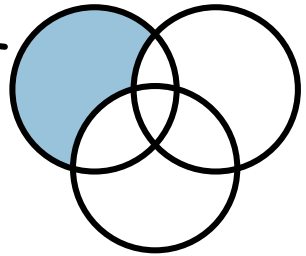
E.g.1: Inclusion-Exclusion Principle



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

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

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



E.g.2: Divisibility

- Divisibility poset: $a \leq b \iff b|a$ (a divides b)
- The Möbius function: $n = b/a$ and

$$\mu(n) = \begin{cases} (-1)^k & \text{if } n = p_1 p_2 \ldots p_k \text{ for } k \text{ distinct primes} \\ 0 & \text{otherwise} \end{cases}$$

  - $\mu(a, b) = \mu(b|a)$, the Möbius function in number theory
- The Riemann zeta function ζ is given by

$$1/\zeta(s) = \sum_{n=1}^{\infty} \mu(n)/n^s$$

Log-Linear Model on Poset [ICML2017]

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

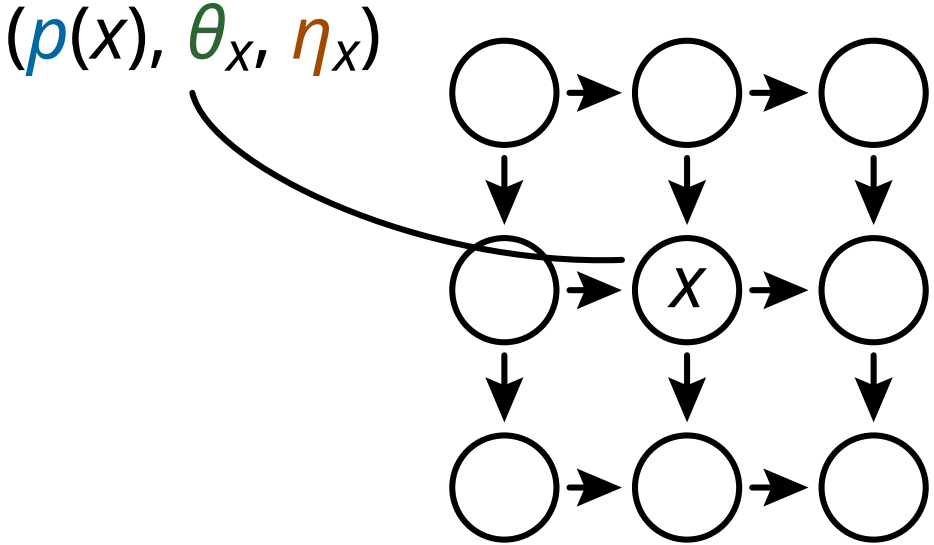
  $$\theta_x = \sum_{s \in S} \mu(s, x) \log p(s),$$

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

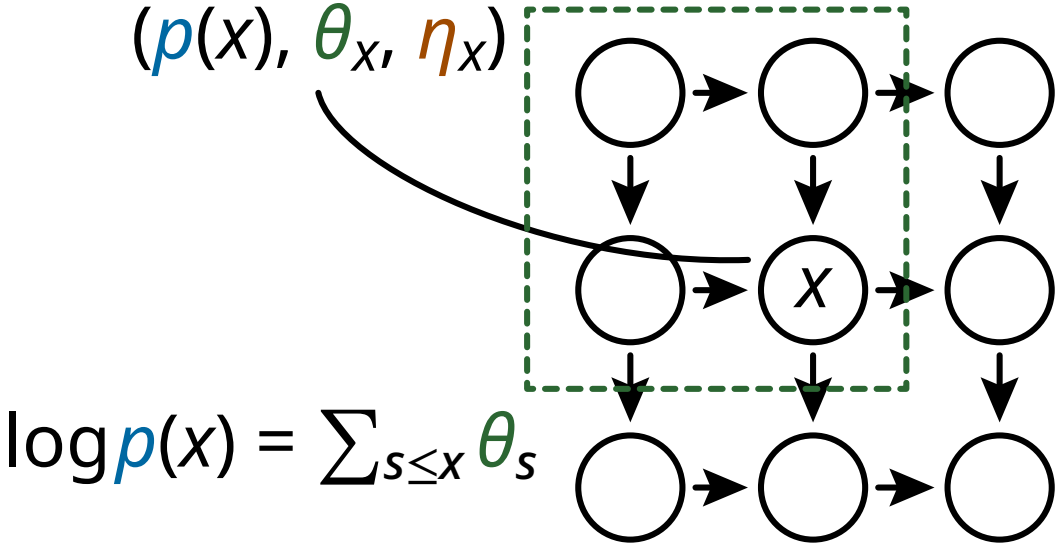
- From the Möbius inversion formula, log-linear model is:

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

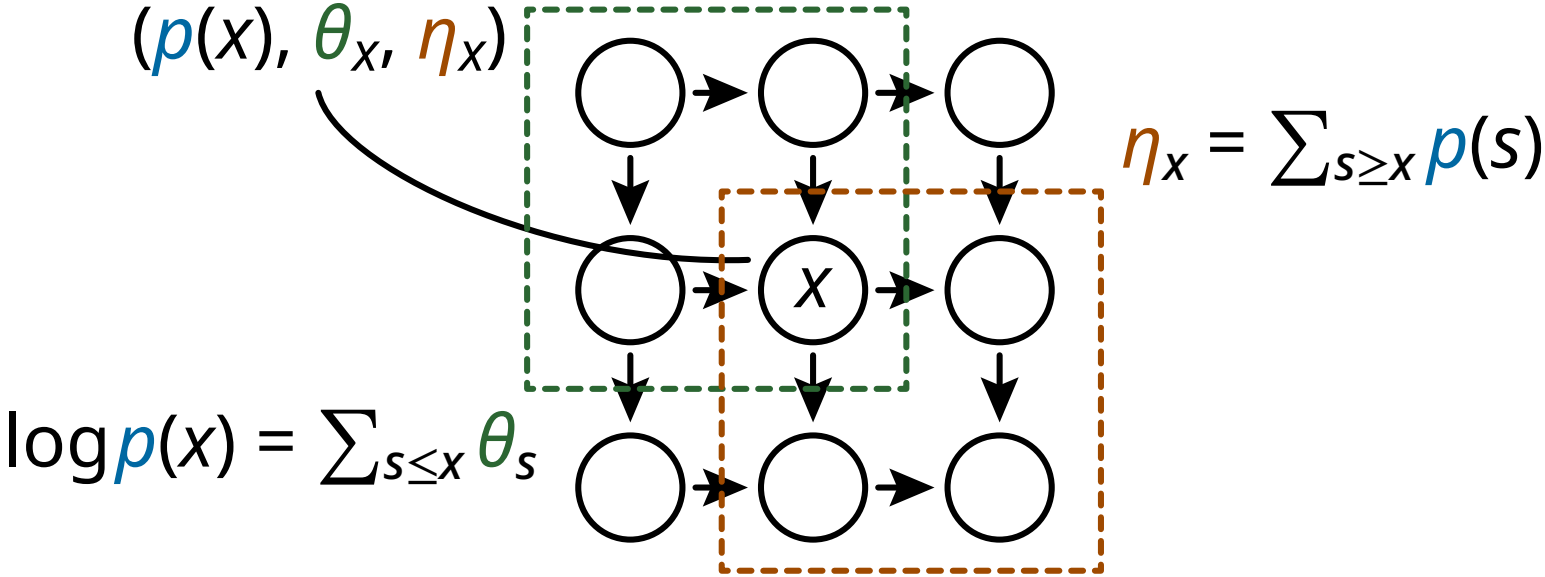
Log-Linear Model on Poset



Log-Linear Model on Poset



Log-Linear Model on Poset

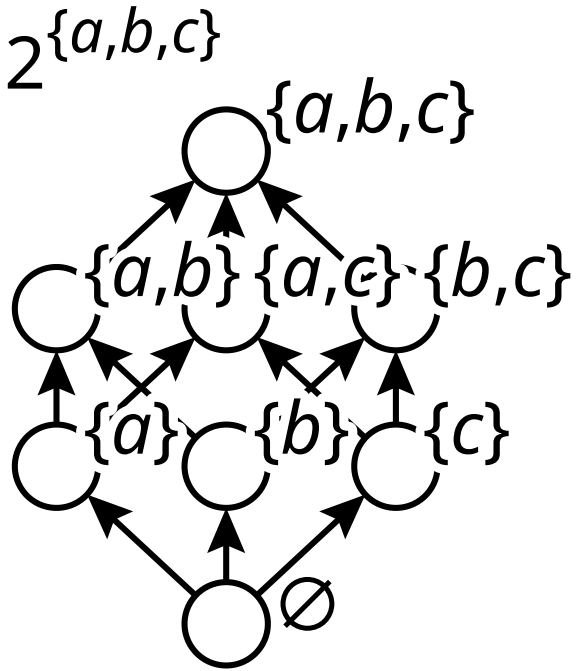


Exponential Family

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

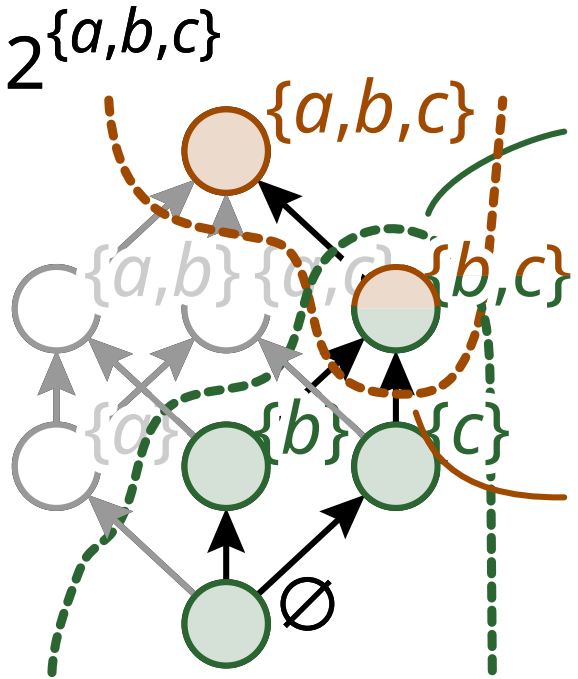
- θ : Natural parameter
- η : Expectation parameter

Binary Log-Linear Model [AAAI2019]



(= Boltzmann machine)

Binary Log-Linear Model [AAAI2019]

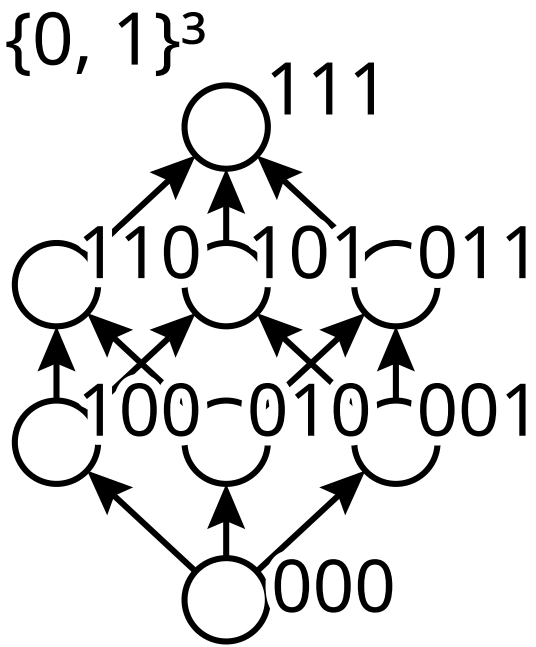


(= Boltzmann machine)

$$\log p(x) = \sum_{s \le x} \theta_s$$

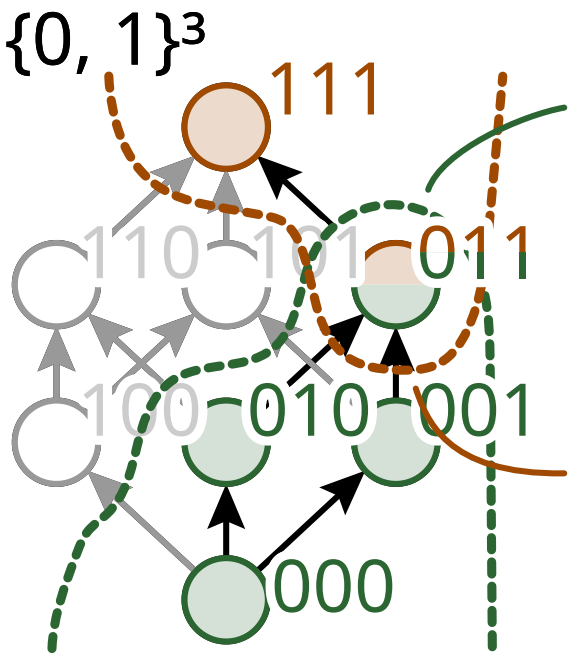
$$\eta_x = \sum_{s \ge x} p(s)$$

Binary Log-Linear Model [AAAI2019]



(= Boltzmann machine)

Binary Log-Linear Model [AAAI2019]



(= Boltzmann machine)

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

For $\boldsymbol{x}$ with $x_{i1} = \cdots = x_{ik} = 1$,

$$\eta_x = \sum_{s \ge x} p(s)$$
$$= \mathbb{E}[x_{i1} \cdots x_{ik}] = \Pr(x_{i1} = \cdots = x_{ik} = 1)$$

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)

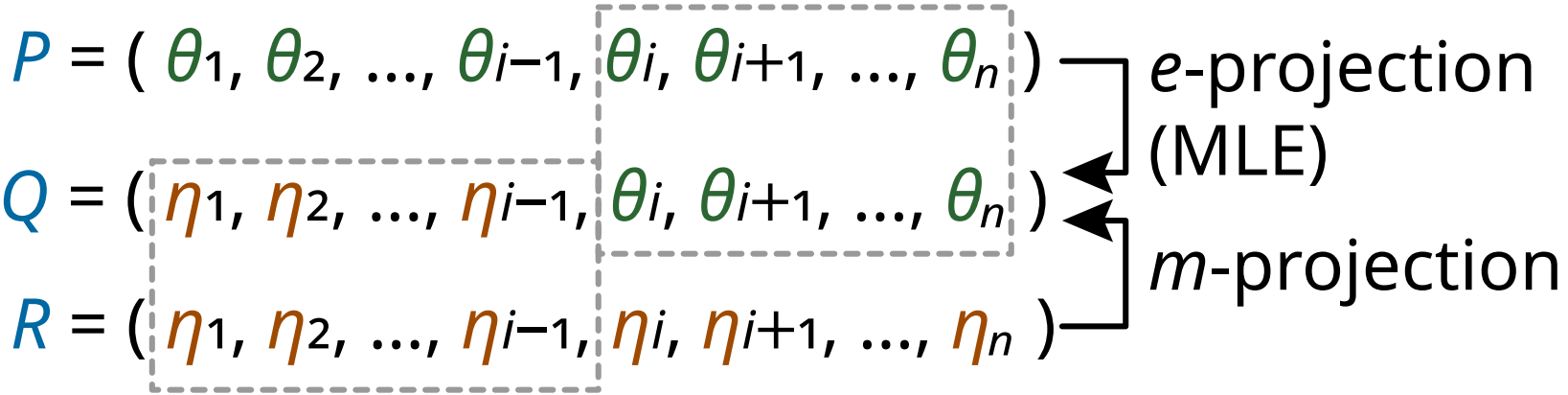
$$\mathbb{E}_s\left[\frac{\partial}{\partial\theta_x}\log p(s)\frac{\partial}{\partial\theta_y}\log p(s)\right] = \sum_{s\in S}\zeta(x,s)\zeta(y,s)p(s) - \eta_x\eta_y$$

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

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

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.

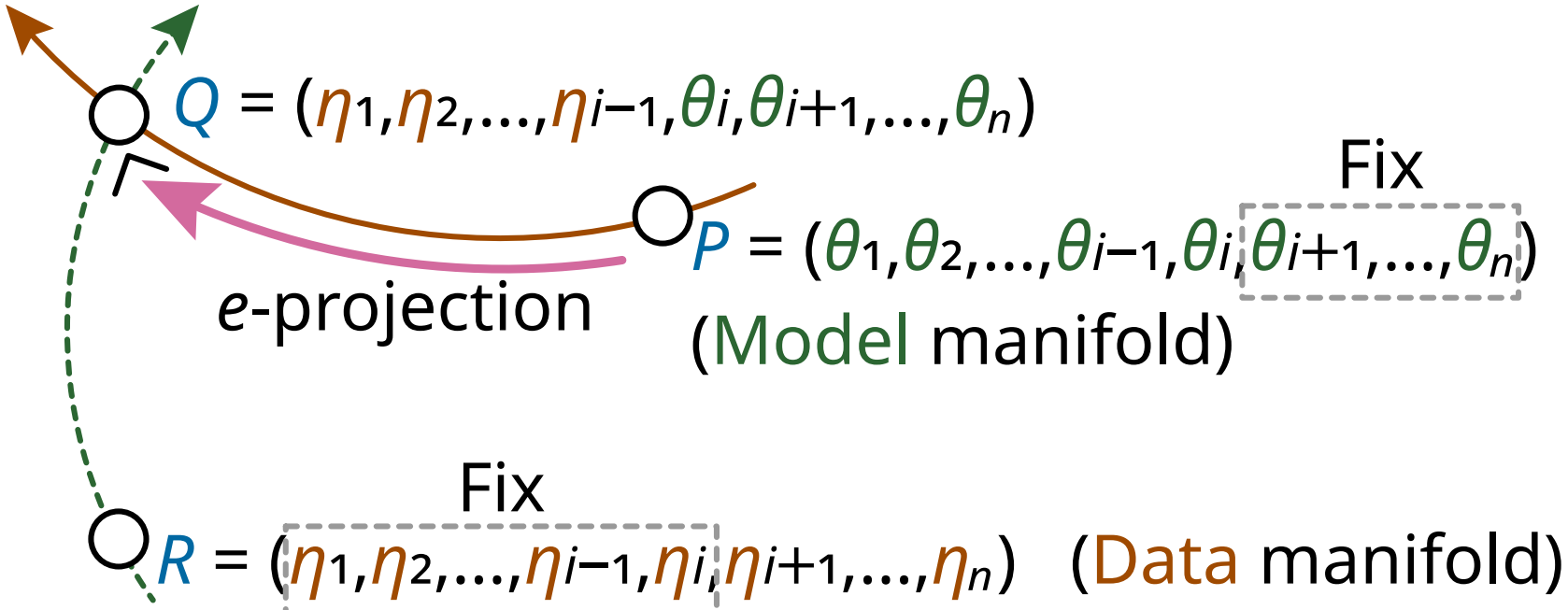
$Q = (\hat{\eta}_1, \hat{\eta}_2, \ldots, \hat{\eta}_{i-1}, 0, 0, \ldots, 0)$

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

Pythagorean theorem: (Q is always unique)

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

Two Submanifolds



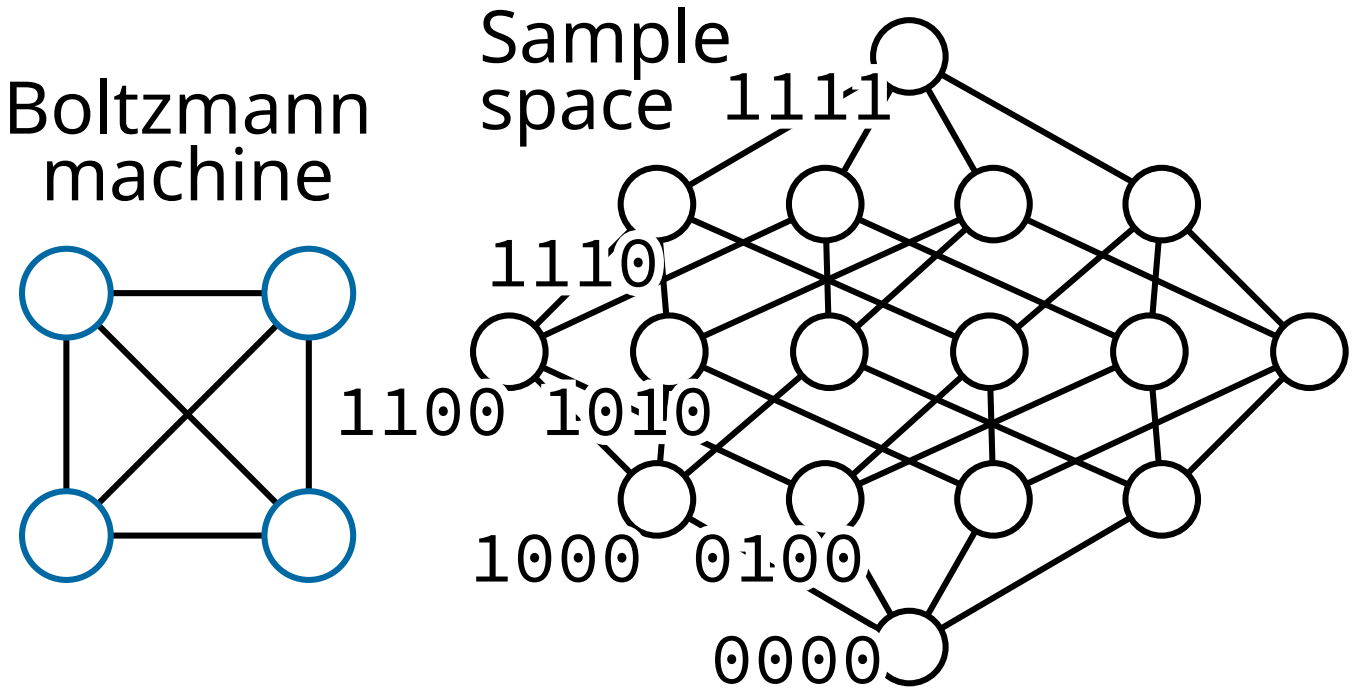
Gradient methods for *e*-projection

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

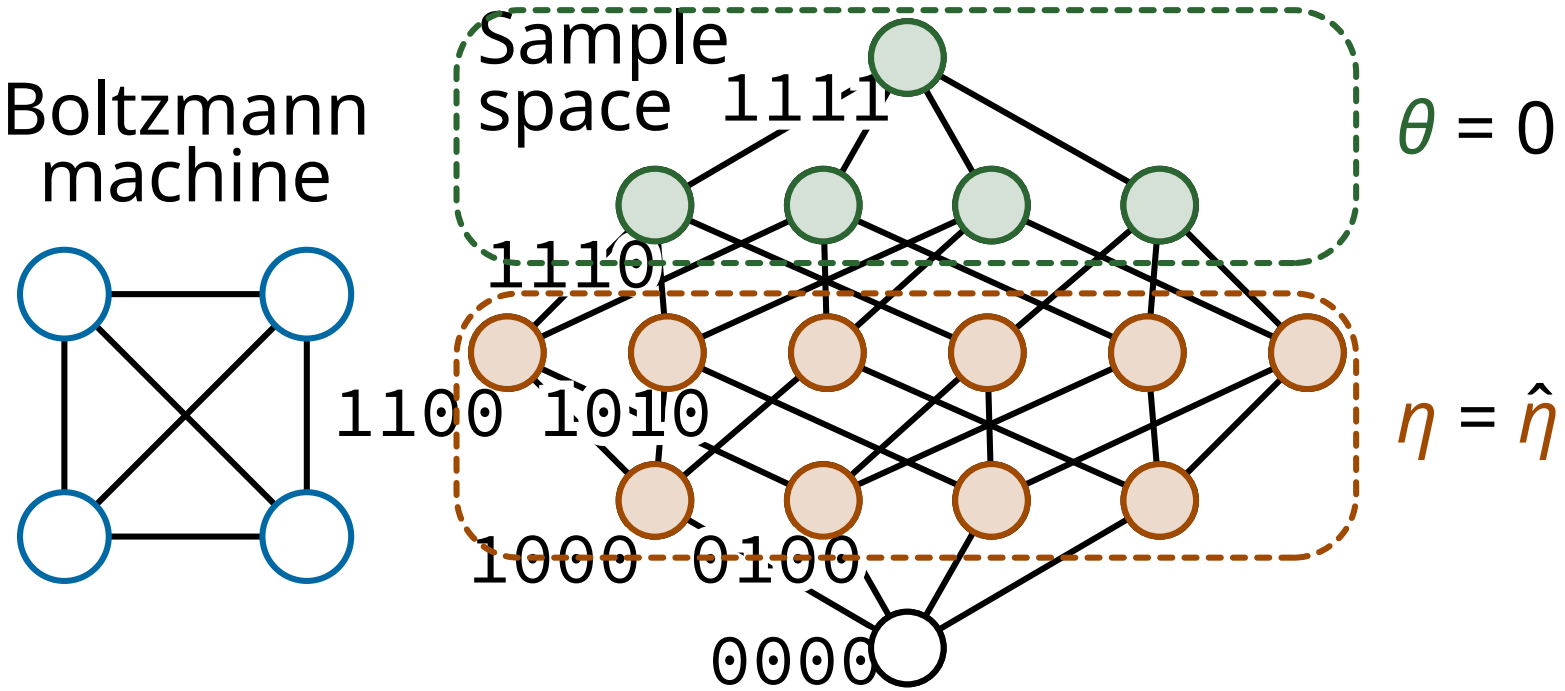
  $\theta_{\text{next},\, i} \leftarrow \theta_i - \epsilon(\eta_i - \hat{\eta}_{\text{target},\, i})$
- Natural gradient (second-order)

  $\theta_{\text{next}} \leftarrow \theta - G^{-1}(\eta - \eta_{\text{target}})$
  - G is Fisher information matrix w.r.t. θ
- Coordinate descent [IBIS2019]

Boltzmann Machine Training

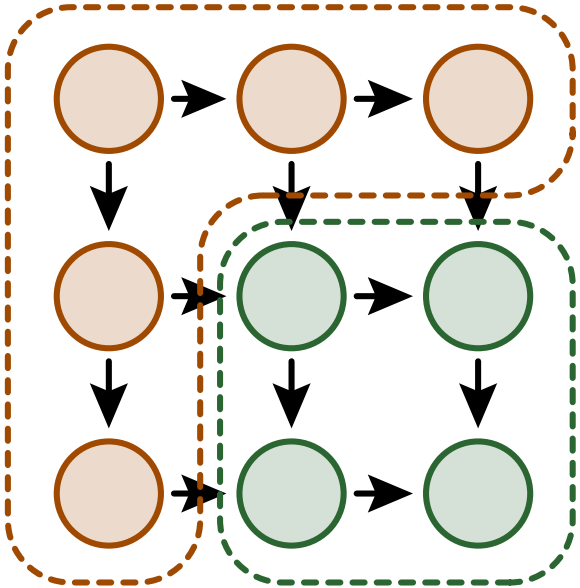


Boltzmann Machine Training



Matrix Balancing

3x3 matrix
as poset:

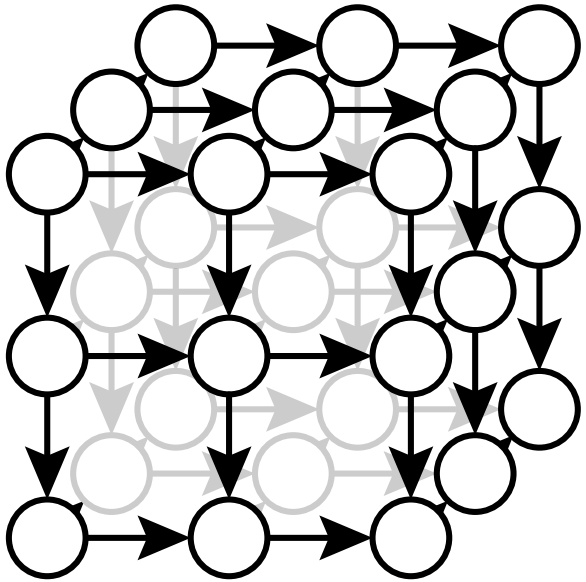


$\eta = 3 - i + 1$

$\theta = \hat{\theta}$

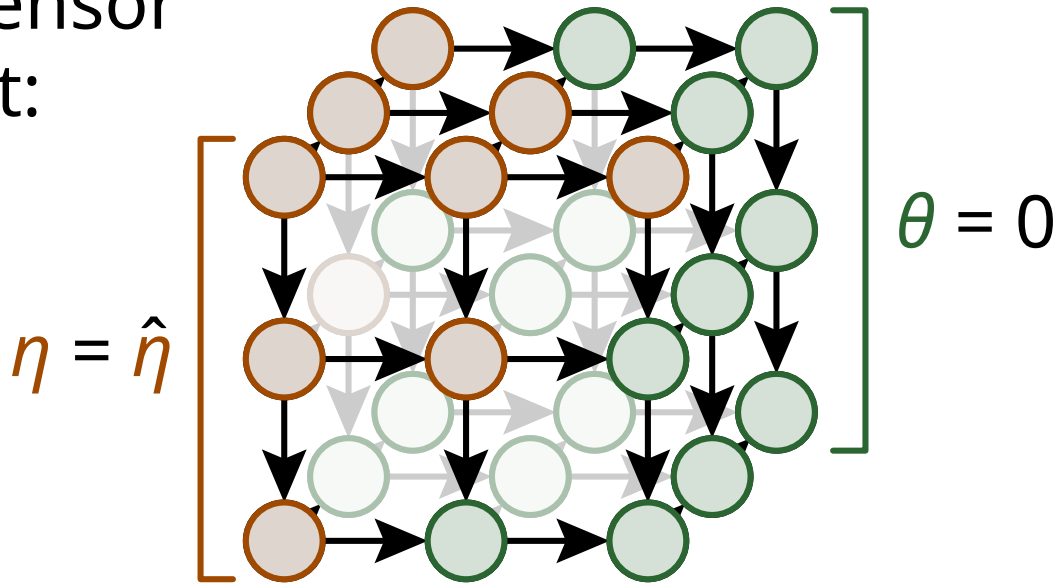
Legendre Decomposition [NeurIPS 2018]

3x3x3 tensor
as poset:



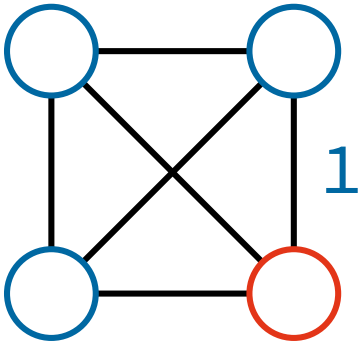
Legendre Decomposition [NeurIPS 2018]

3x3x3 tensor
as poset:

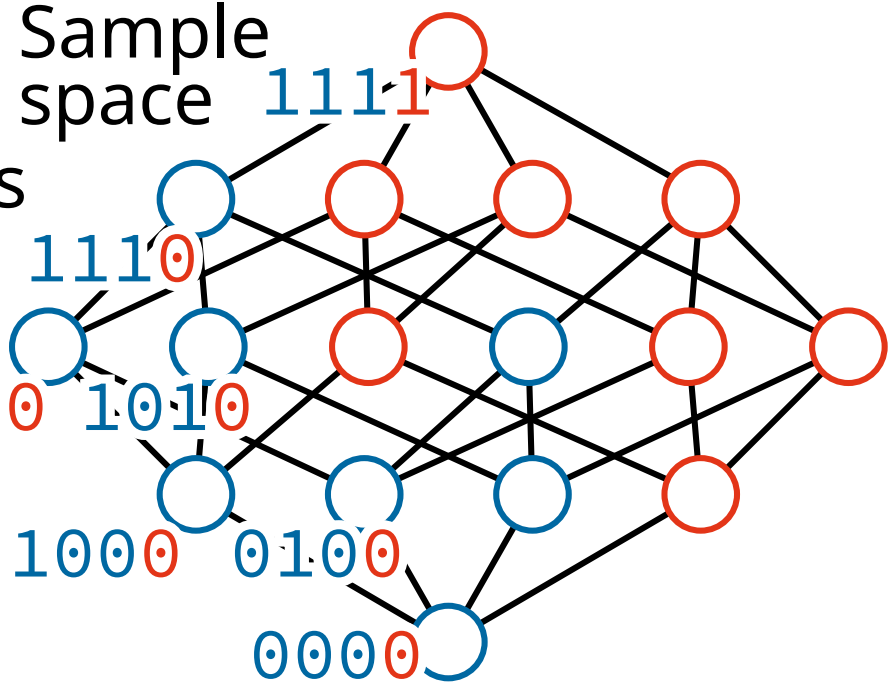


Introducing Hidden Variables in BM

BM with
hidden variables

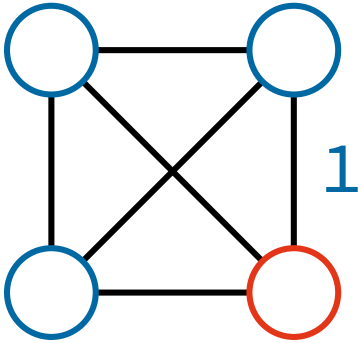


Sample
space

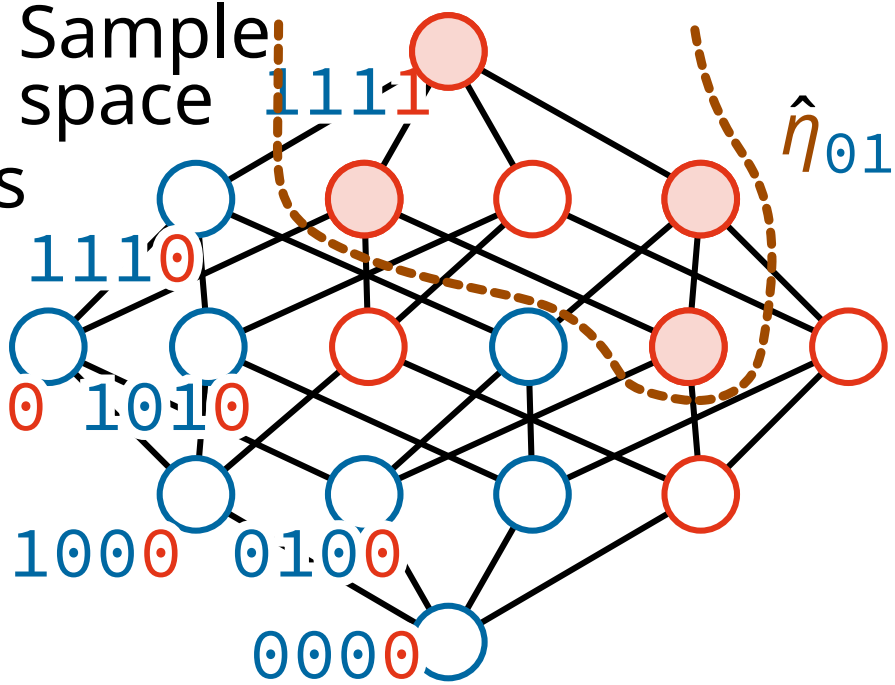


Introducing Hidden Variables in BM

BM with hidden variables

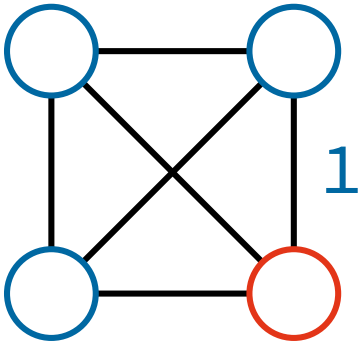


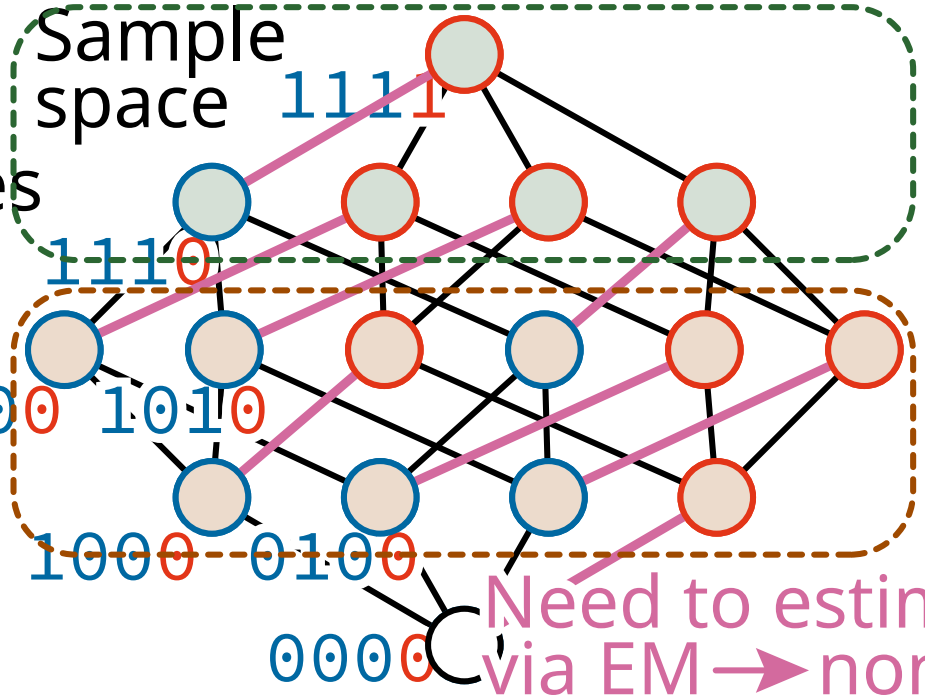
Sample space



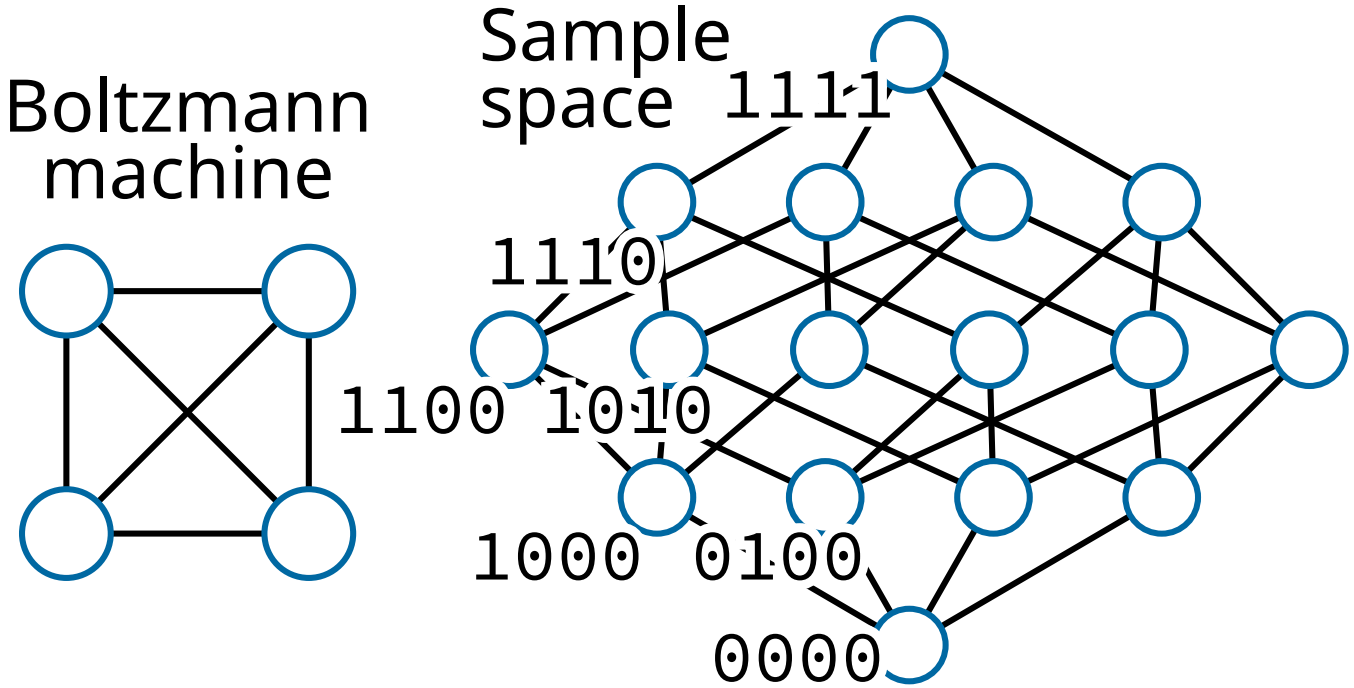
Introducing Hidden Variables in BM

BM with hidden variables

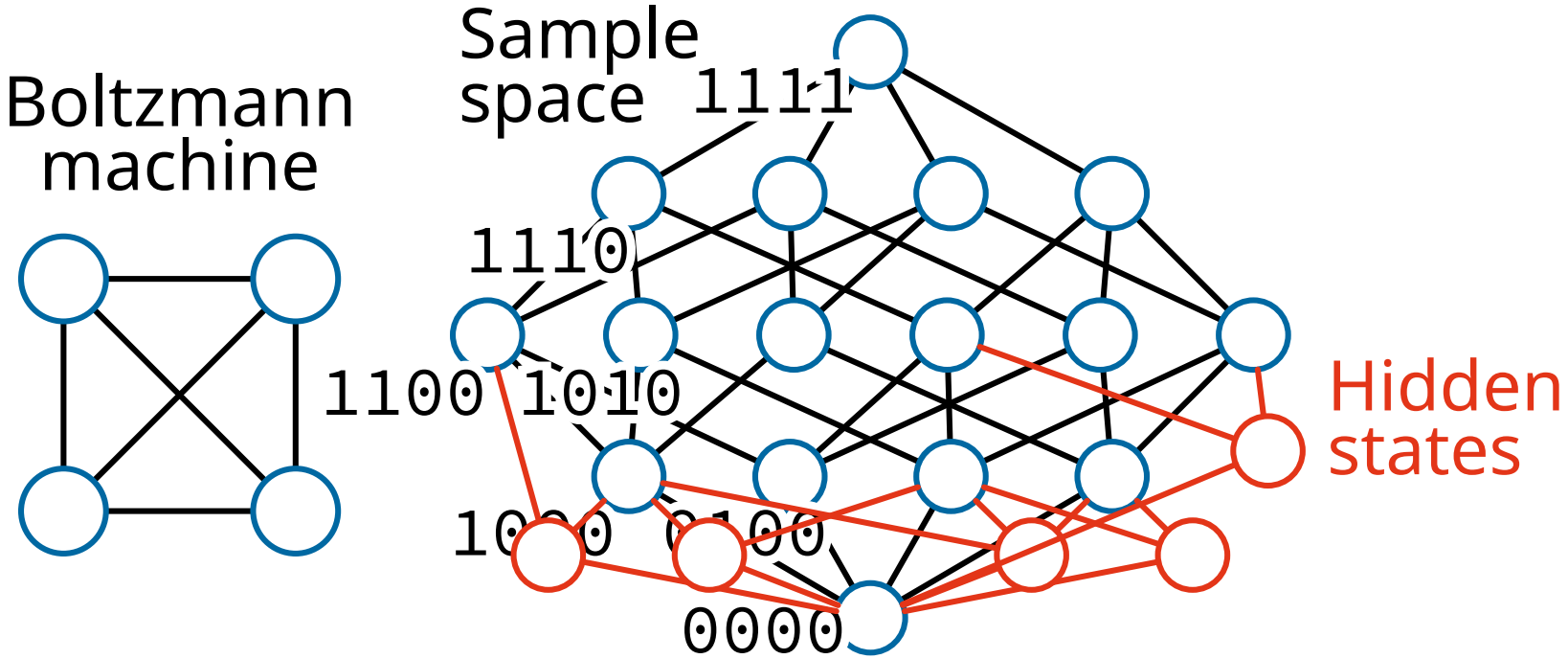




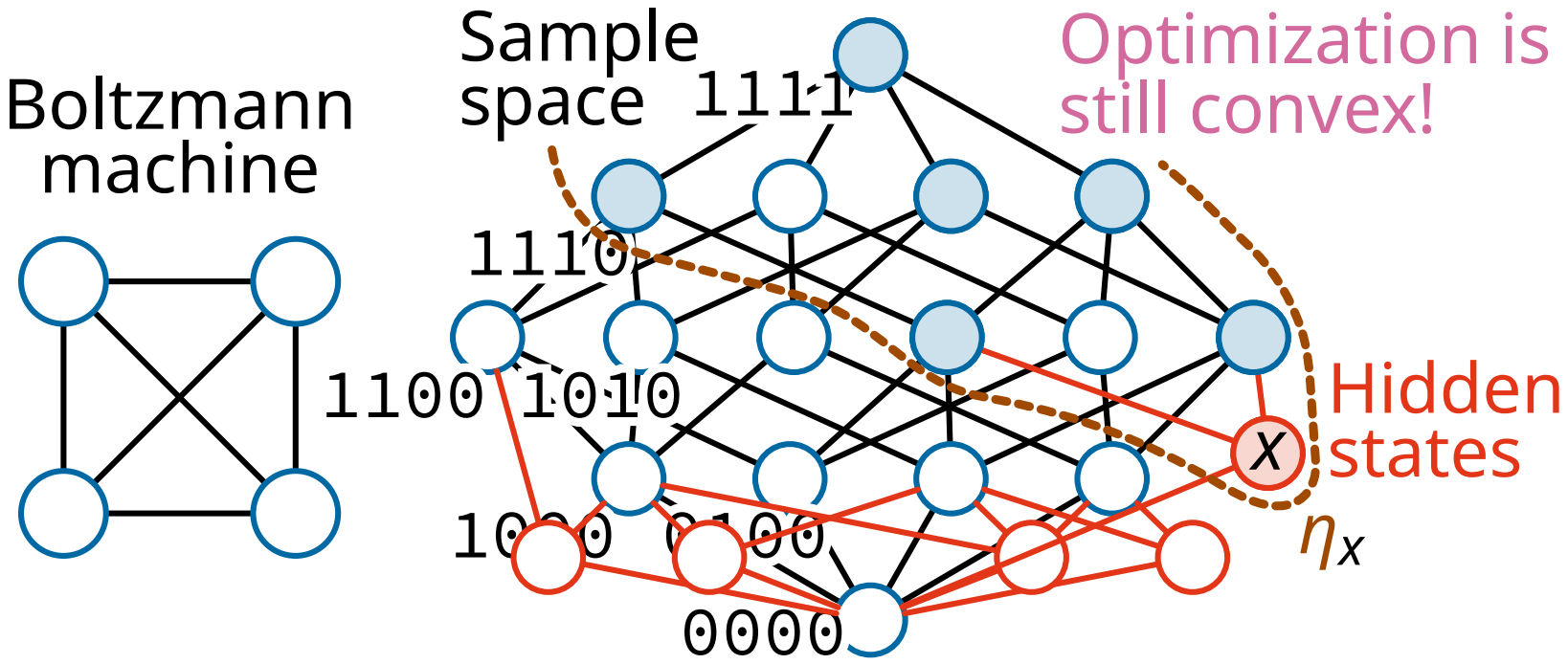
Introducing Hidden States



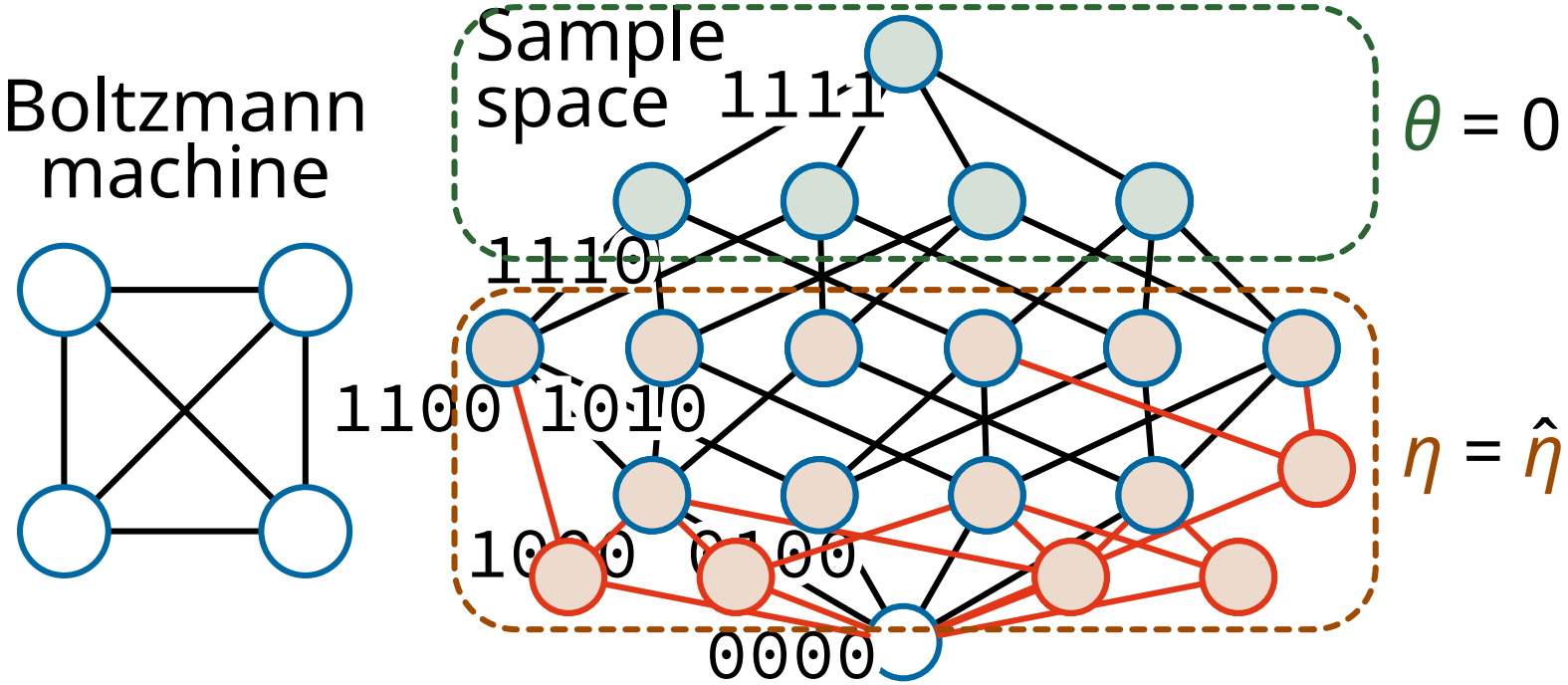
Introducing Hidden States



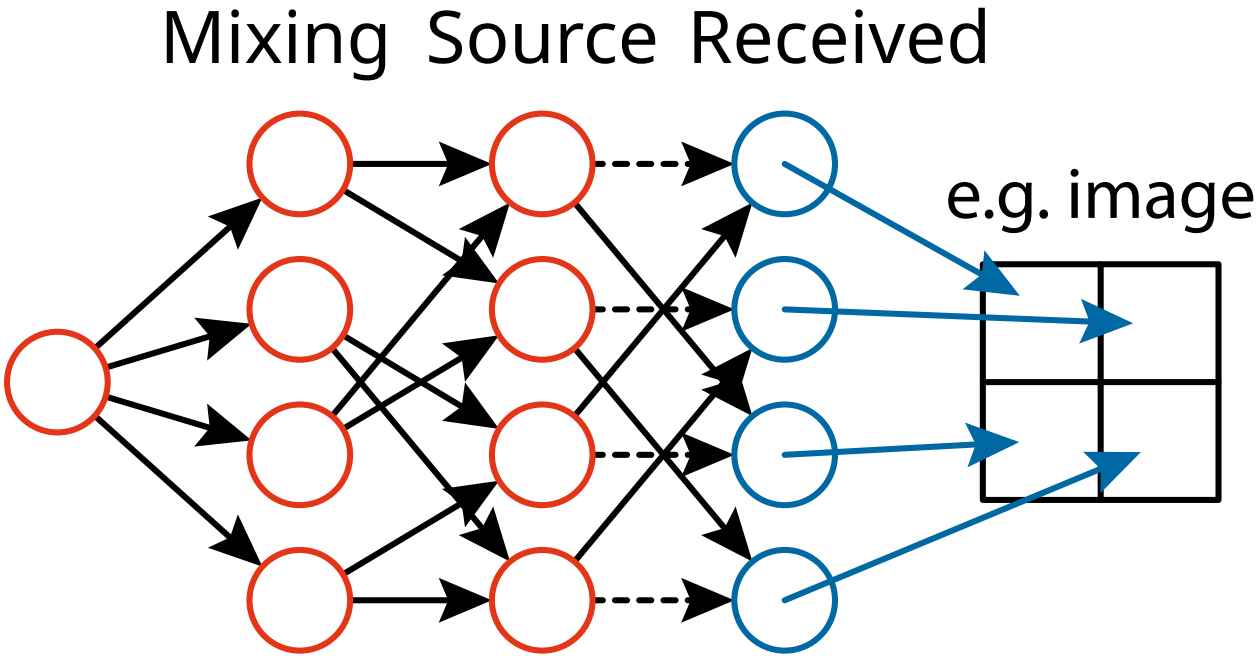
Introducing Hidden States



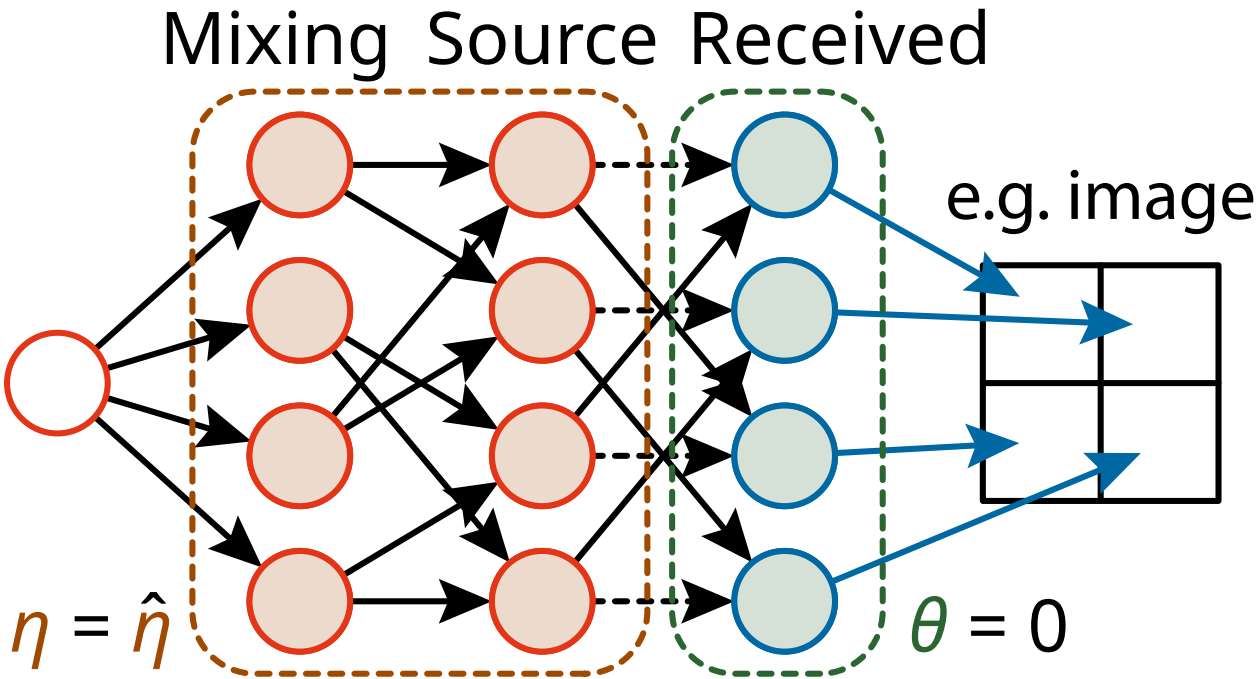
Introducing Hidden States



Blind Source Separation [arXiv]



Blind Source Separation [arXiv]

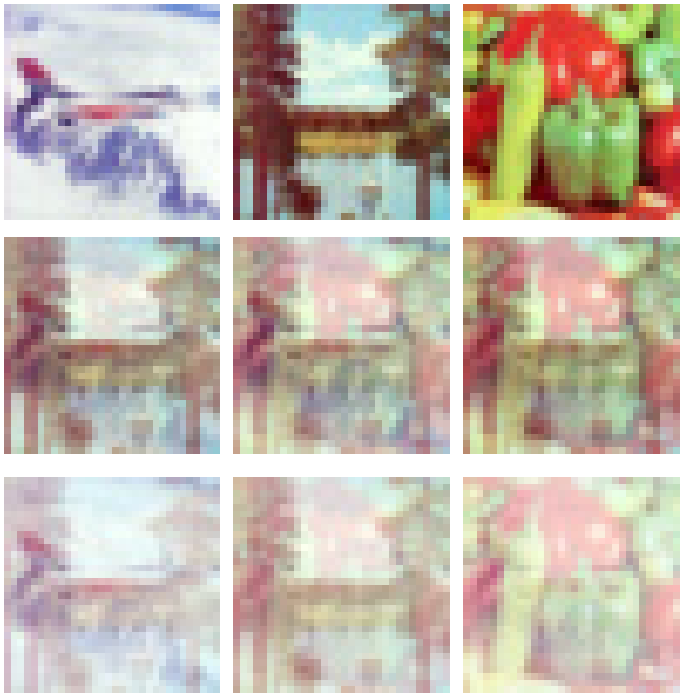


Results of BSS

Source

Received

Reconst.



Method	RMSE
IGBSS	**0.27032**
FastICA	0.43630
NMF	0.62195
DicLearn	0.37167

Relationship to Homology

- Möbius function is Euler characteristics
- Consider order complex $\Delta(S)$ of a poset S with $\bot, \top \in S$
  - (i) Vertices of $\Delta(S)$ are elements of S
  - (ii) Faces of $\Delta(S)$ are chains of S
- For the Euler characteristic $\chi(\Delta(S))$,

  $\mu(\bot, \top) = \chi(\Delta(S)) + 1$

  - Two spaces are homotopy equivalent
    $\Rightarrow$ Euler characteristics are the same

Summary: Recipe for Poset-LogLinear

1. Treat the target as a poset
2. Introduce the log-linear model on poset
3. Formulate the objective as coordinate mixing
4. Solve it by a gradient method

(This slide is at `https://mahito.nii.ac.jp/`)

Acknowledgment

- Tsuda, K. (UTokyo), Nakahara, H. (RIKEN CBS)
- Yamada, R. (KyotoU), Mimura, K. (HiroshimaCU)
- Luo, S., Azizi, L. (USydney)
- Hayashi, S., Matsushima, S. (UTokyo)
- Borgwardt, K. and his lab members (ETH Zürich)
- Lab members at NII