January 28, 2019

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

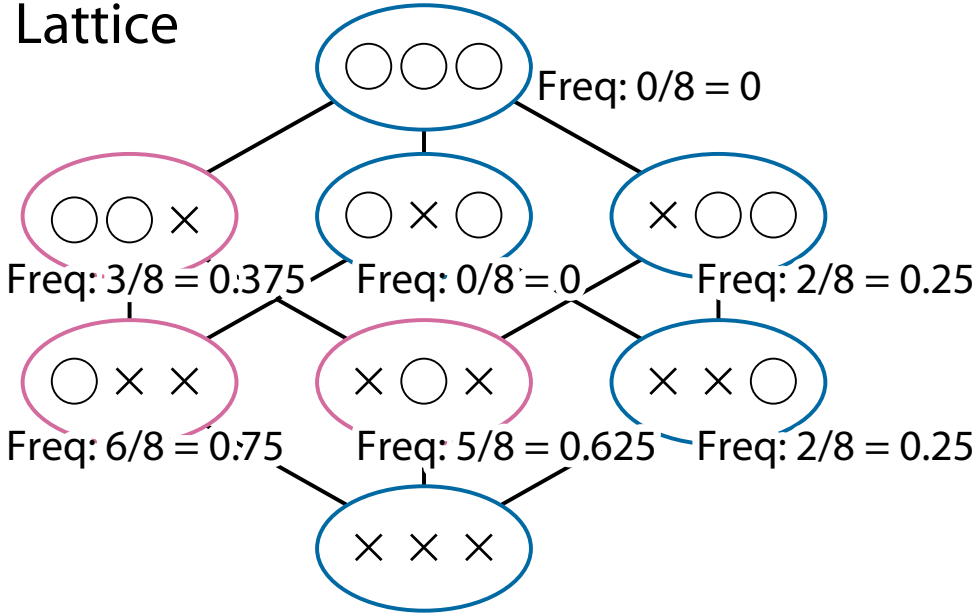
Mahito Sugiyama

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

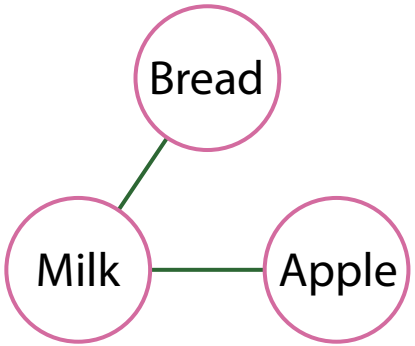


Learning Hierarchical Distribution (2/2)

MLE

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

Boltzmann machine

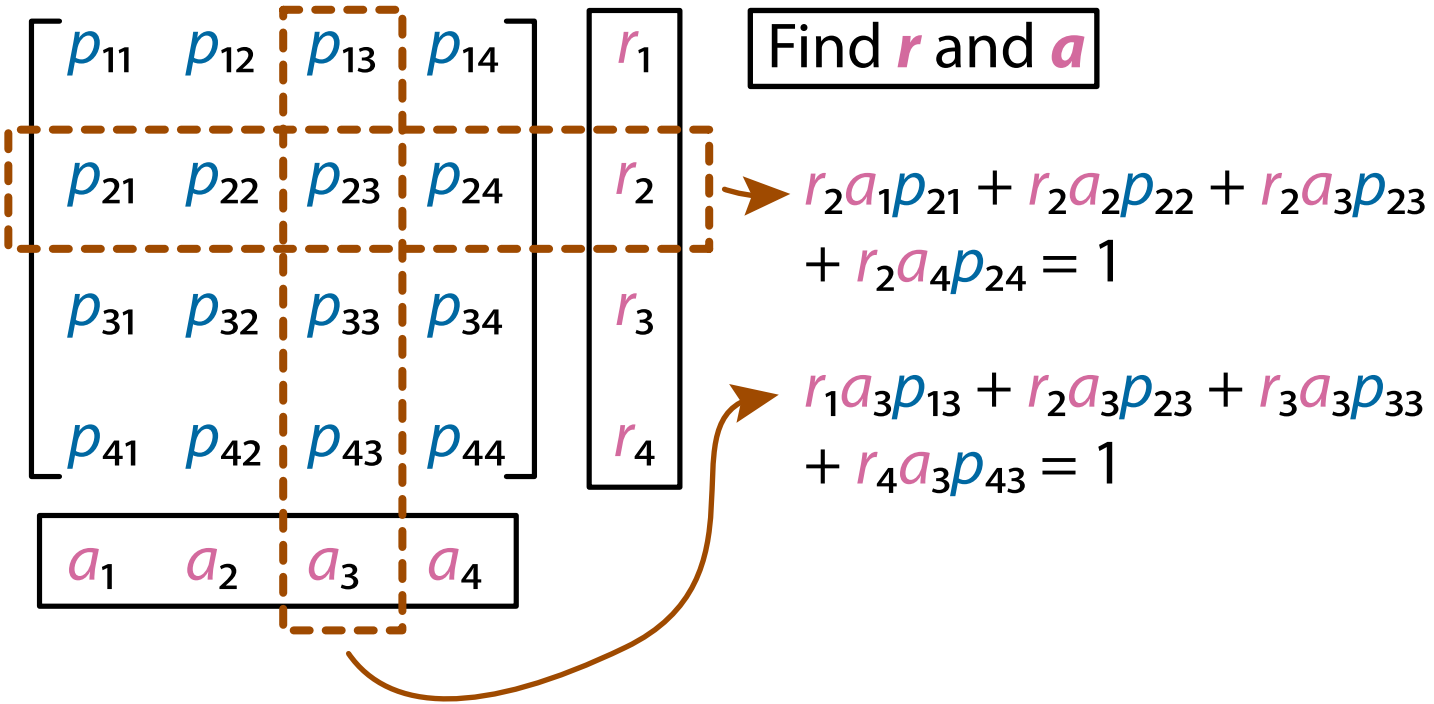


Bread	Milk	Apple	Prob. from data	Learned prob.
×	×	×	?	0.0000300109
○	×	×	0.375	0.3749599867
×	○	×	?	0.1499903954
×	×	○	?	0.0000000016
○	○	×	0.375	0.2250096042
○	×	○	?	0.0000200043
×	○	○	0.25	0.0999895960
○	○	○	?	0.1500004008

Tensor (Matrix) Balancing

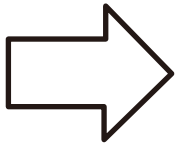
$$\begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} & p_{34} \\ p_{41} & p_{42} & p_{43} & p_{44} \end{bmatrix}$$

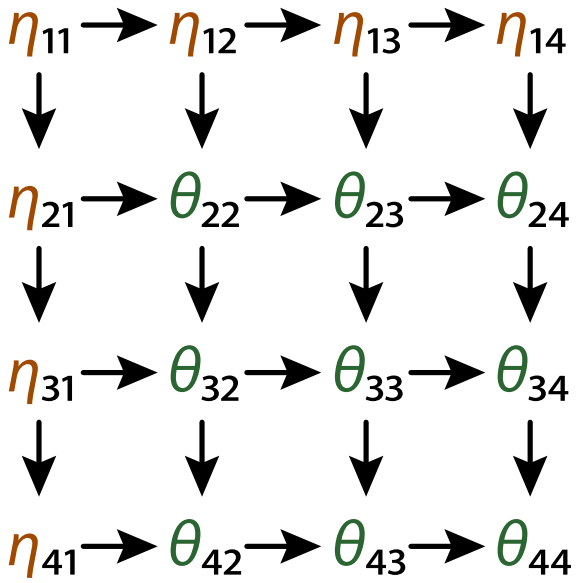
Tensor (Matrix) Balancing



Introduce η and θ

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





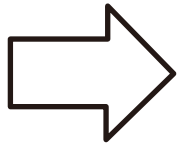
Matrix balancing is achieved if:

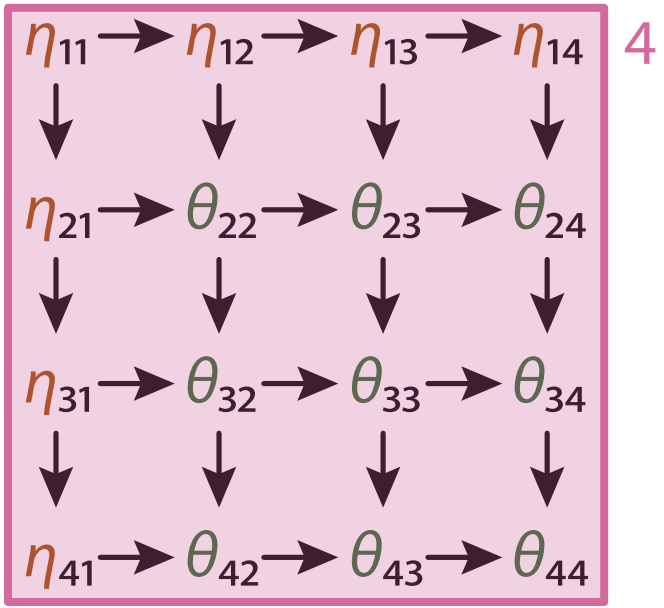
$\eta_{11} = 4, \eta_{21} = 3, \eta_{31} = 2, \eta_{41} = 1$

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

Definition of η

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





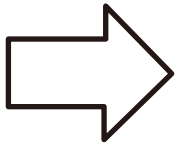
Matrix balancing is achieved if:

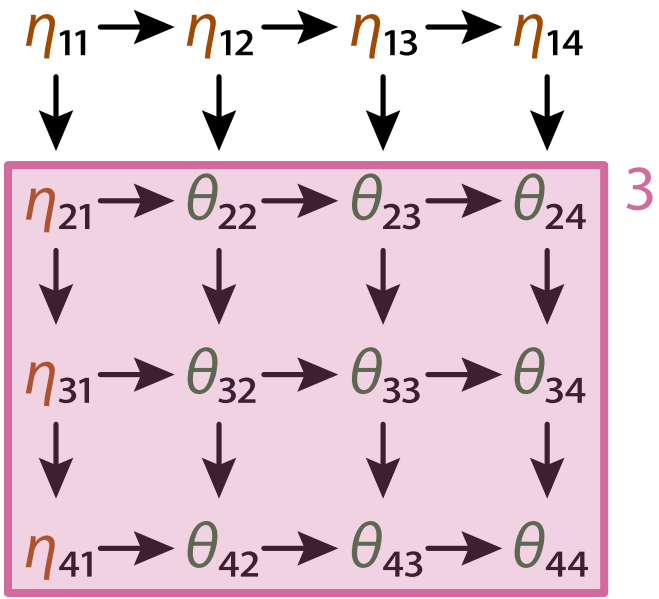
$\eta_{11} = 4, \eta_{21} = 3, \eta_{31} = 2, \eta_{41} = 1$

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

Definition of η

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





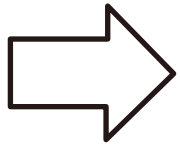
Matrix balancing is achieved if:

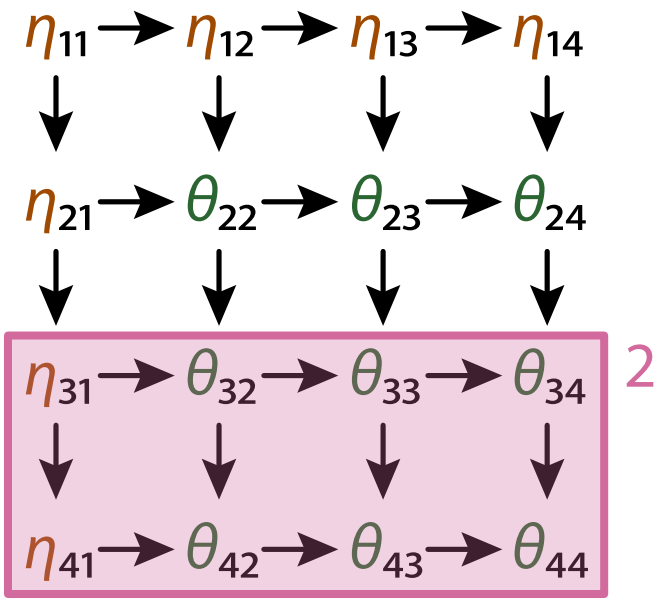
$\eta_{11} = 4, \eta_{21} = 3, \eta_{31} = 2, \eta_{41} = 1$

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

Definition of η

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





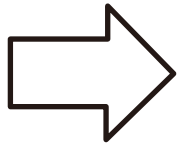
Matrix balancing is achieved if:

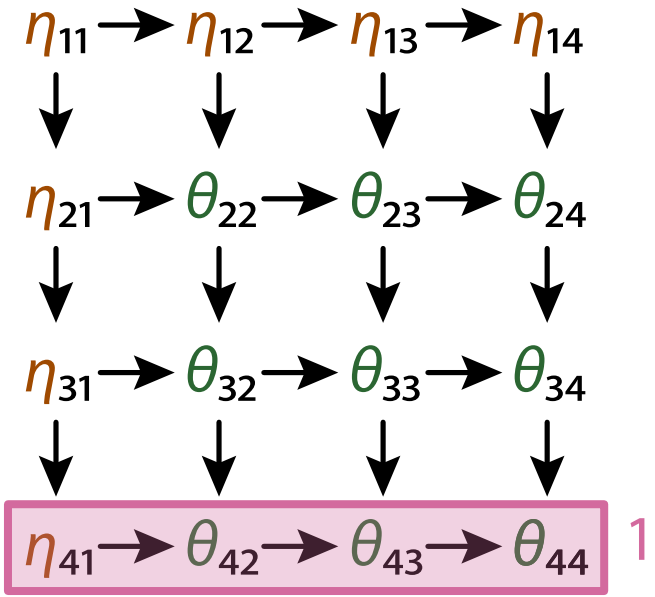
$\eta_{11} = 4, \eta_{21} = 3, \eta_{31} = 2, \eta_{41} = 1$

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

Definition of η

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





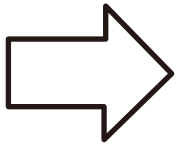
Matrix balancing is achieved if:

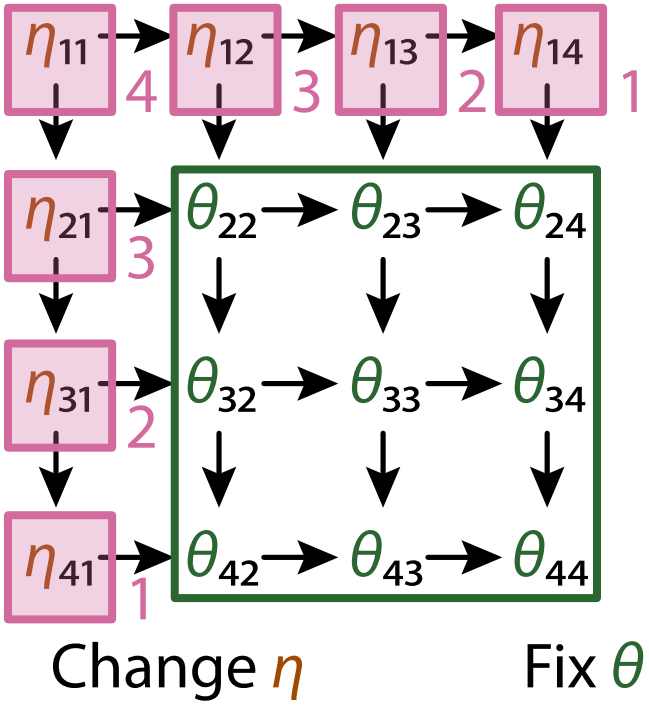
$\eta_{11} = 4, \eta_{21} = 3, \eta_{31} = 2, \eta_{41} = 1$

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

Balancing as Constraints on η

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



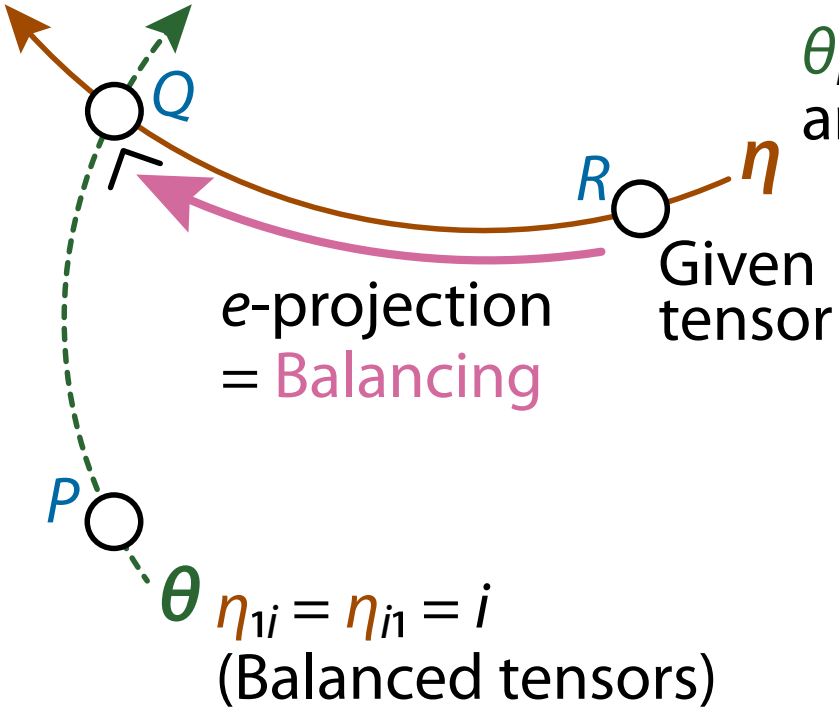


Matrix balancing is achieved if:

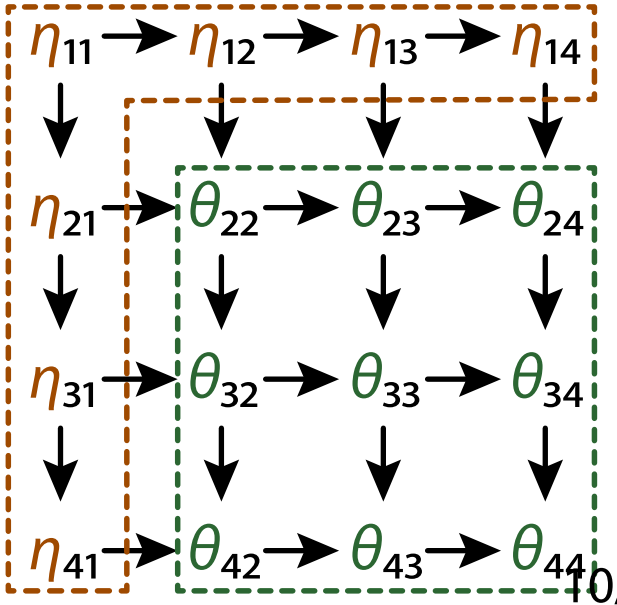
$\eta_{11} = 4, \eta_{21} = 3, \eta_{31} = 2, \eta_{41} = 1$

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

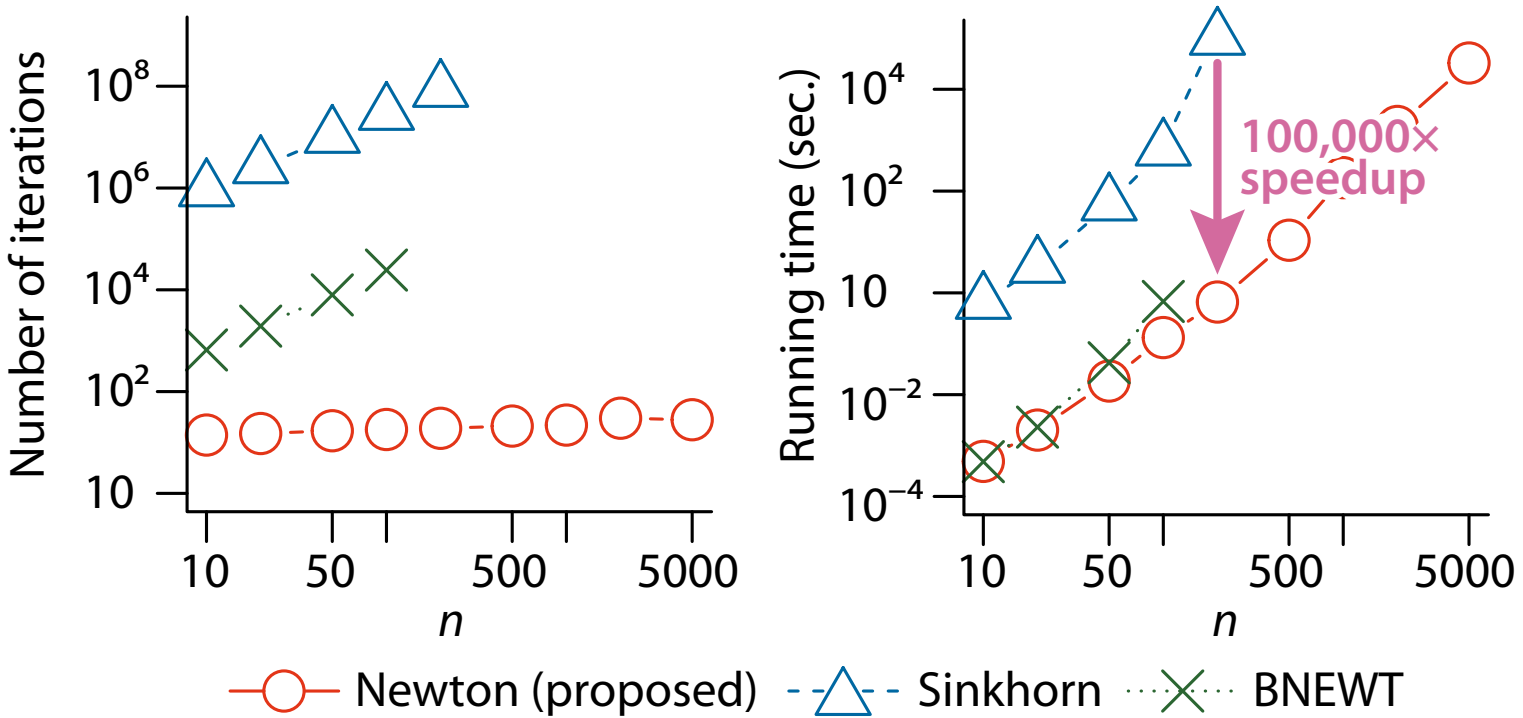
e-Projection = Balancing



θ_{ij} $(i \neq 1, i \neq 1)$ are fixed

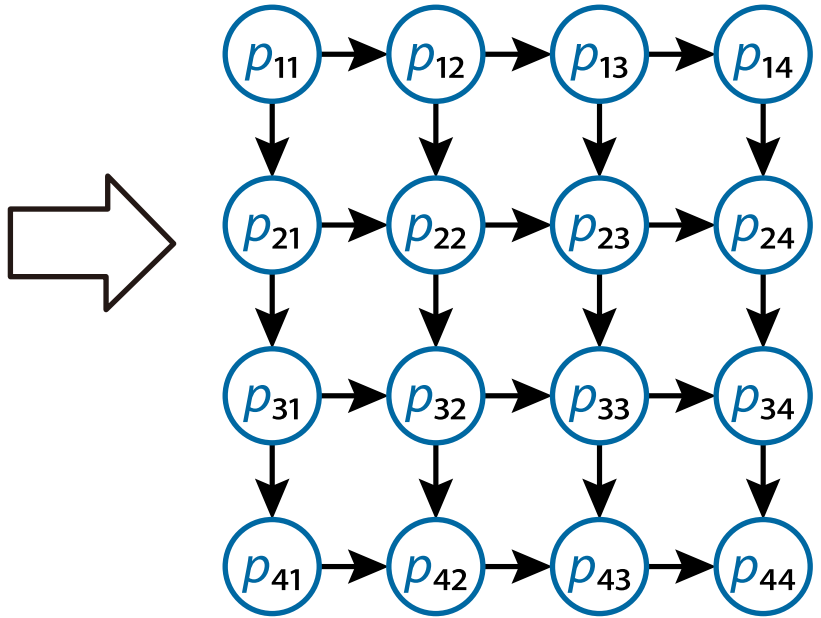


Speedup on Hessenberg Matrix

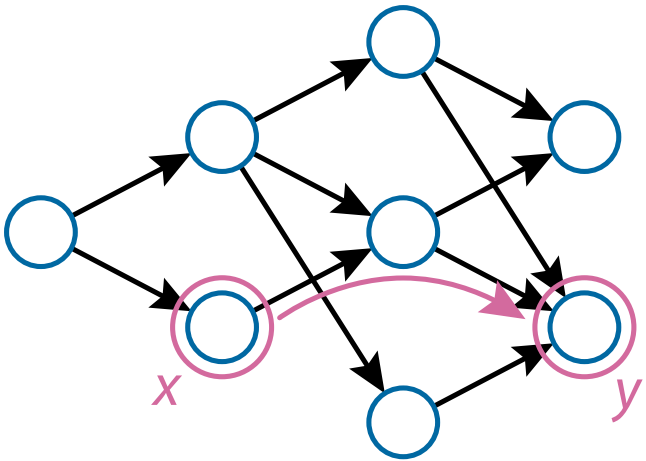


From Matrix to Poset (DAG)

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



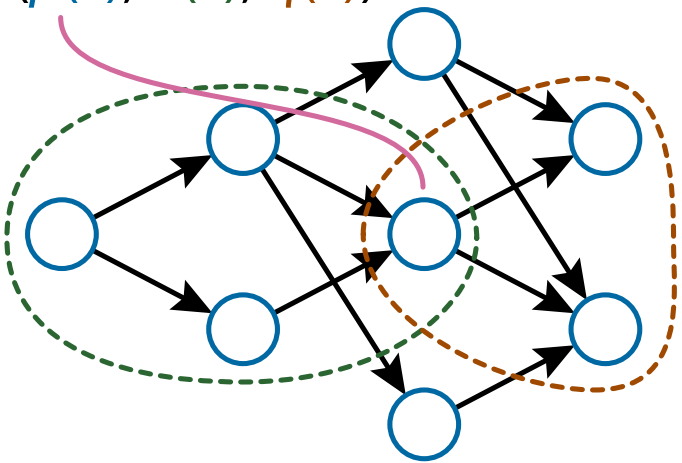
Partially Ordered Set



- 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:
$(p(x), \theta(x), \eta(x))$



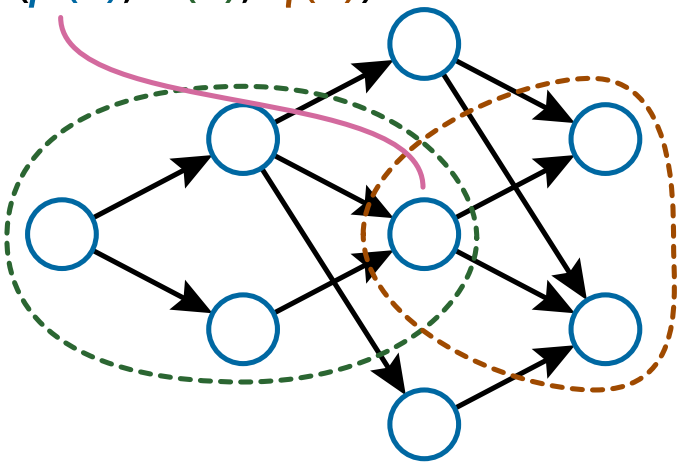
- A probability vector $p{:}S \to (0, 1)$ s.t. $\sum_{x \in S} p(x) = 1$
  - (Normalized) weight for each node
- We introduce $\theta{:}S \to \mathbb{R}$ and $\eta{:}S \to \mathbb{R}$ as

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

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

Log-Linear Model on Poset

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

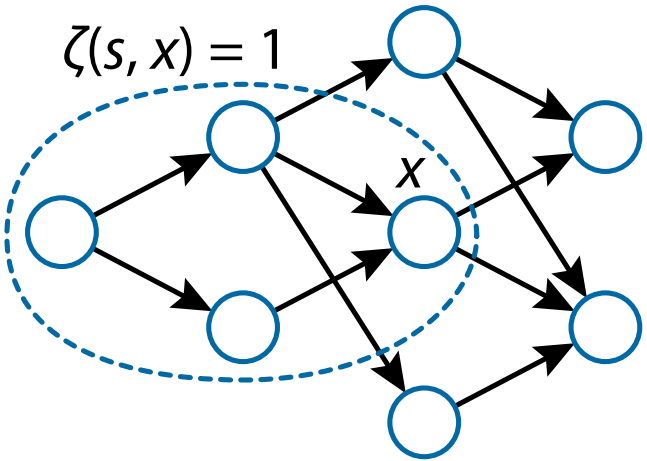


- A probability vector $p{:}S \to (0, 1)$ s.t. $\sum_{x \in S} p(x) = 1$
  - (Normalized) weight for each node
- We introduce $\theta{:}S \to \mathbb{R}$ and $\eta{:}S \to \mathbb{R}$ as

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

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

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,

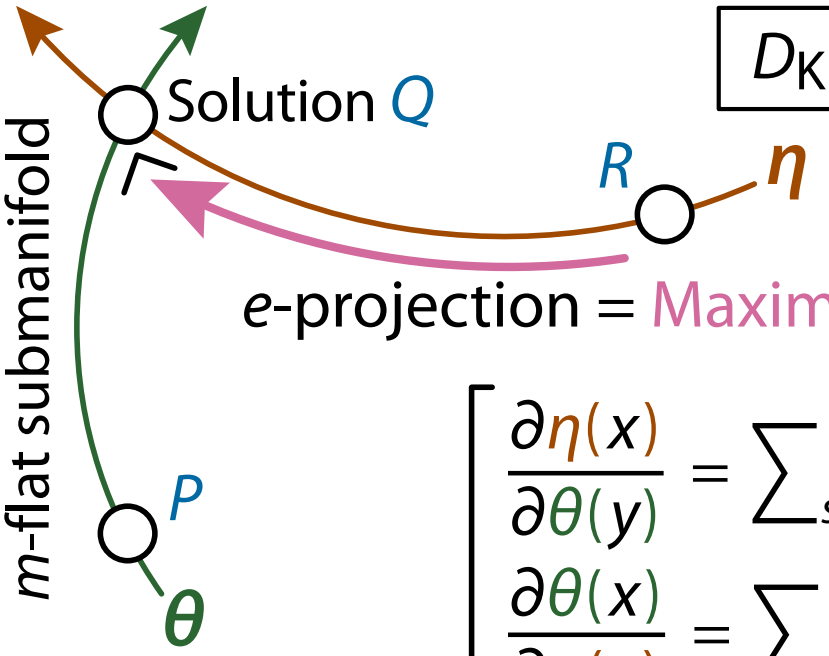
  $$|A \cup B \cup C| = |A| + |B| + |C| - |A \cap B| - |B \cap C| - |A \cap C| + |A \cap B \cap C|$$

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

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

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

Riemannian Manifold with Info. Geometry

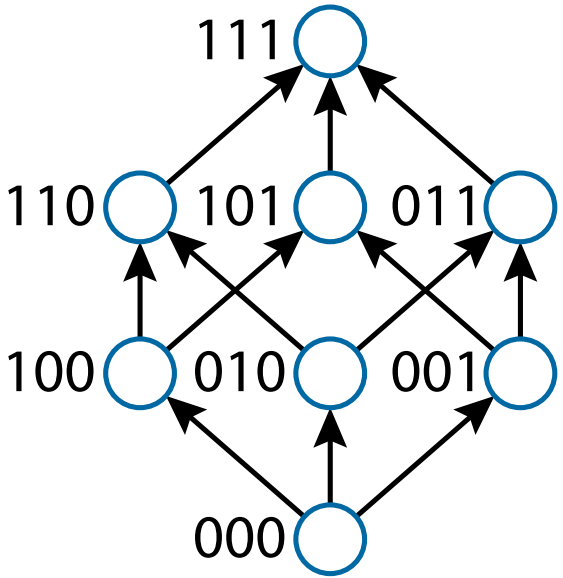


$$D_{\mathrm{KL}}[P, R] = D_{\mathrm{KL}}[P, Q] + D_{\mathrm{KL}}[Q, R]$$

e-projection = Maximum Likelihood Estimation

$$\begin{cases} \dfrac{\partial \eta(x)}{\partial \theta(y)} = \displaystyle\sum_{s \in S} \zeta(x, s)\zeta(y, s)p(s) - \eta(x)\eta(y) \\ \dfrac{\partial \theta(x)}{\partial \eta(y)} = \displaystyle\sum_{s \in S} \underbrace{\mu(s, x)\mu(s, y)}_{\text{Capture partial order structure}} p(s)^{-1} \end{cases}$$

For Example: Binary Case



- Our model:

$$\log p(x) = \sum_{s \le x} \theta(s), \qquad \eta(x) = \sum_{s \ge x} p(s)$$

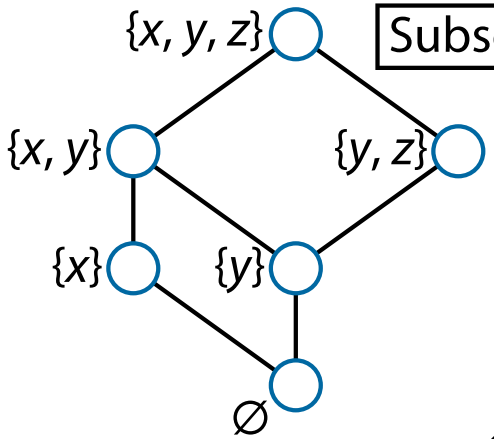
is generalization of the log-linear model on binary vectors with $\boldsymbol{x} \in \{0, 1\}^n = S$:

$$\log p(\boldsymbol{x}) = \sum\nolimits_i \theta^i x^i + \sum\nolimits_{i<j} \theta^{ij} x^i x^j + \ldots + \theta^{1\ldots n} x^1 x^2 \ldots x^n - \psi,$$

$$\eta^i = \mathbf{E}[x^i] = \Pr(x^i = 1),$$

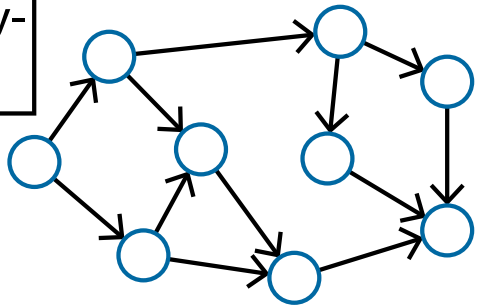
$$\eta^{ij} = \mathbf{E}[x^i x^j] = \Pr(x^i = x^j = 1), \ldots$$

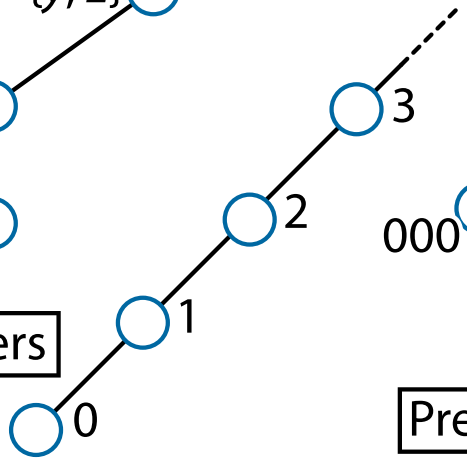
Various Posets



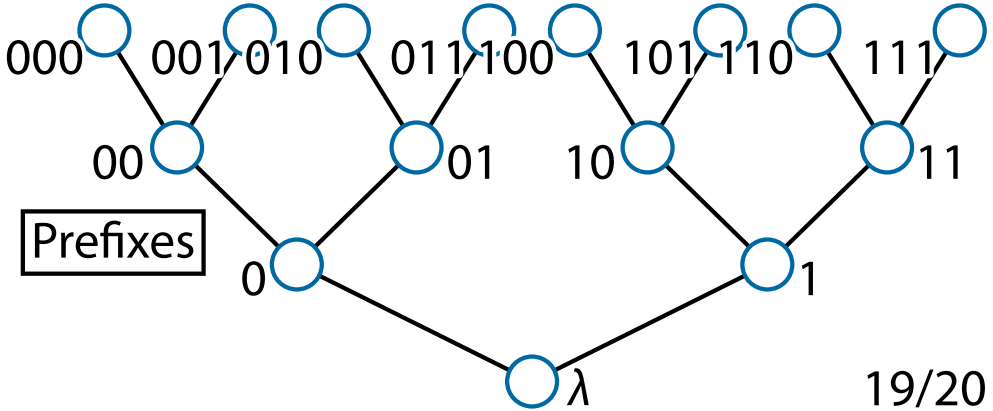
Subset of power set

Directed Acyclic Graph





Positive integers



Prefixes

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