February 4, 2019

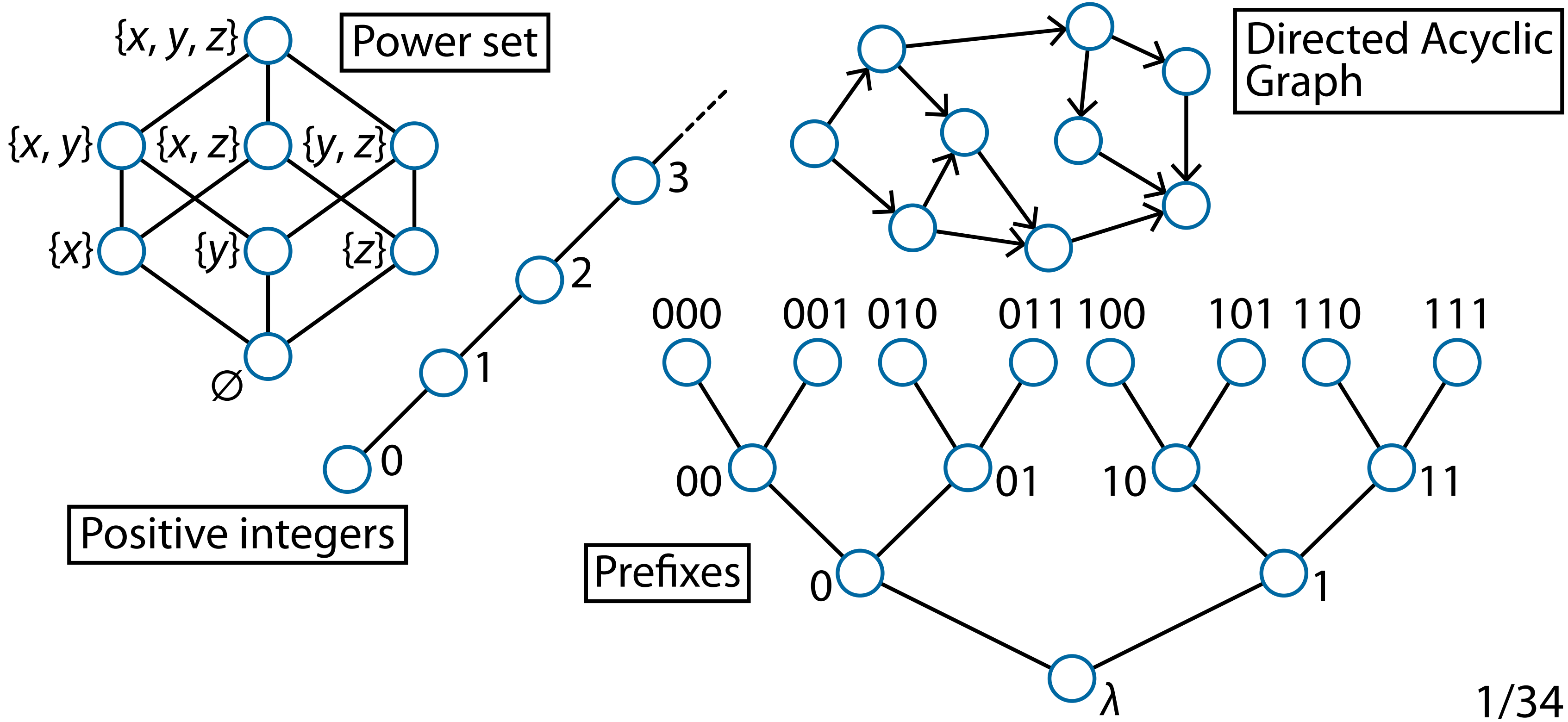
Inter-University Research Institute Corporation / Research Organization of Information and Systems
National Institute of Informatics

Machine Learning and Information Geometry II

Introduction to Intelligent Systems Science II

Mahito Sugiyama

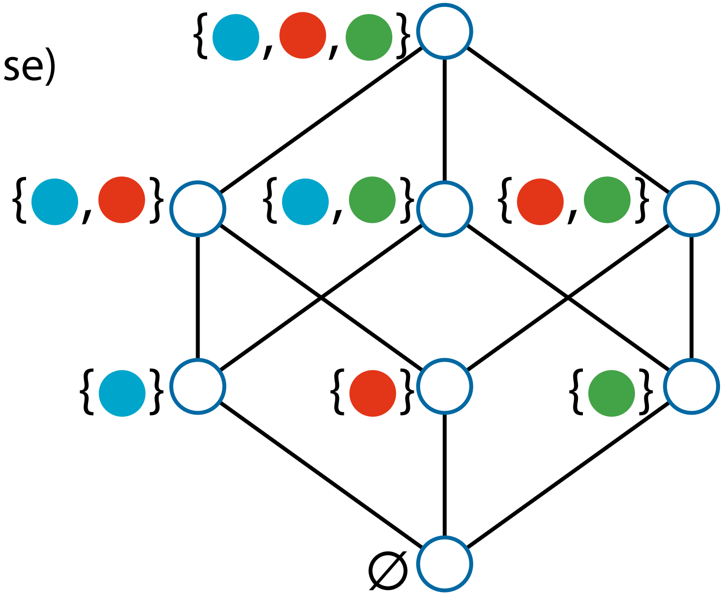
Various Hierarchical Models as Posets



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
ID10:	0	1	0

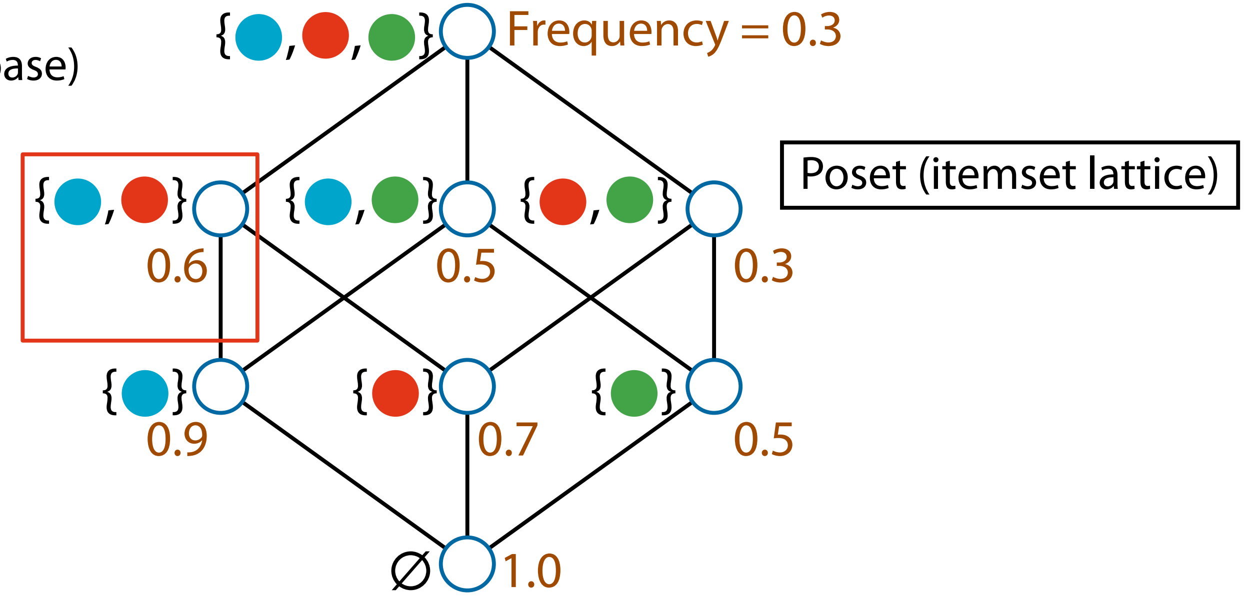


Poset (itemset lattice)

Frequency as Importance Measure

Binary vectors
(Transaction database)

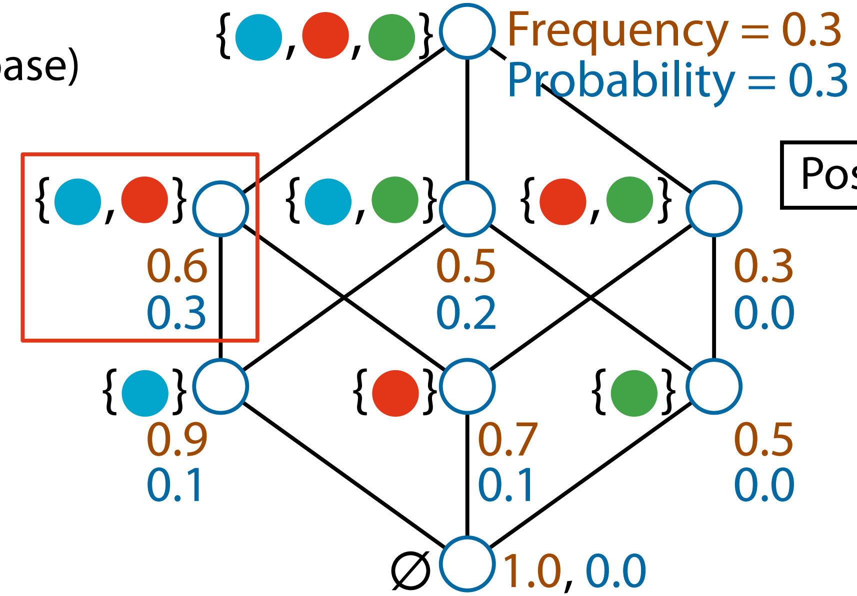
ID 1:	1	1	0
ID 2:	1	1	1
ID 3:	1	1	0
ID 4:	1	1	1
ID 5:	1	1	0
ID 6:	1	0	1
ID 7:	1	0	1
ID 8:	1	1	1
ID 9:	1	0	0
ID10:	0	1	0



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
ID10:	0	1	0

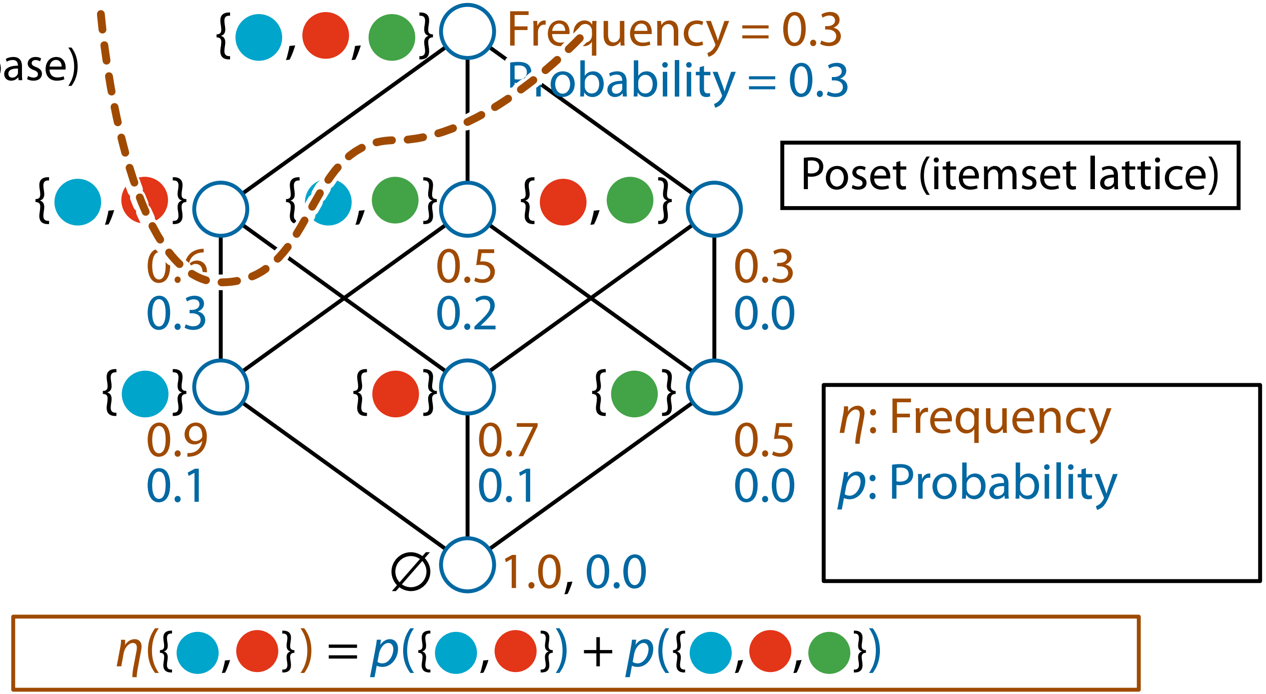


Poset (itemset lattice)

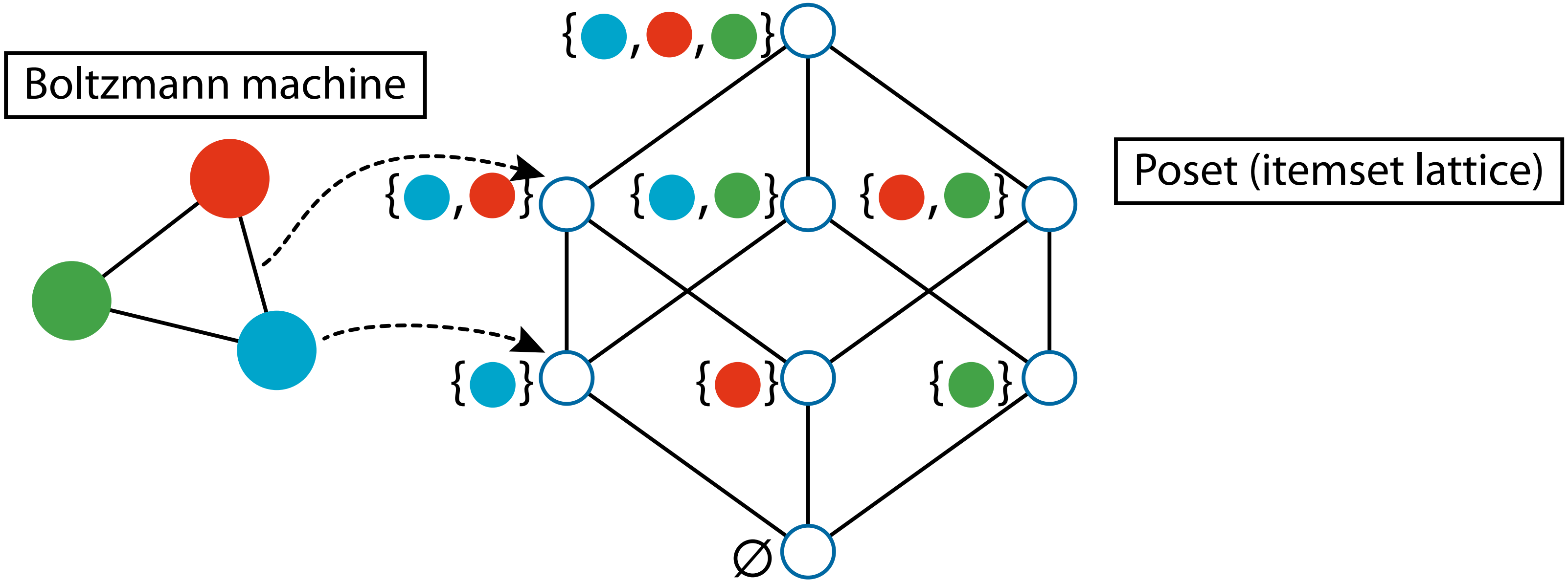
Pattern Mining → Upward Analysis

Binary vectors
(Transaction database)

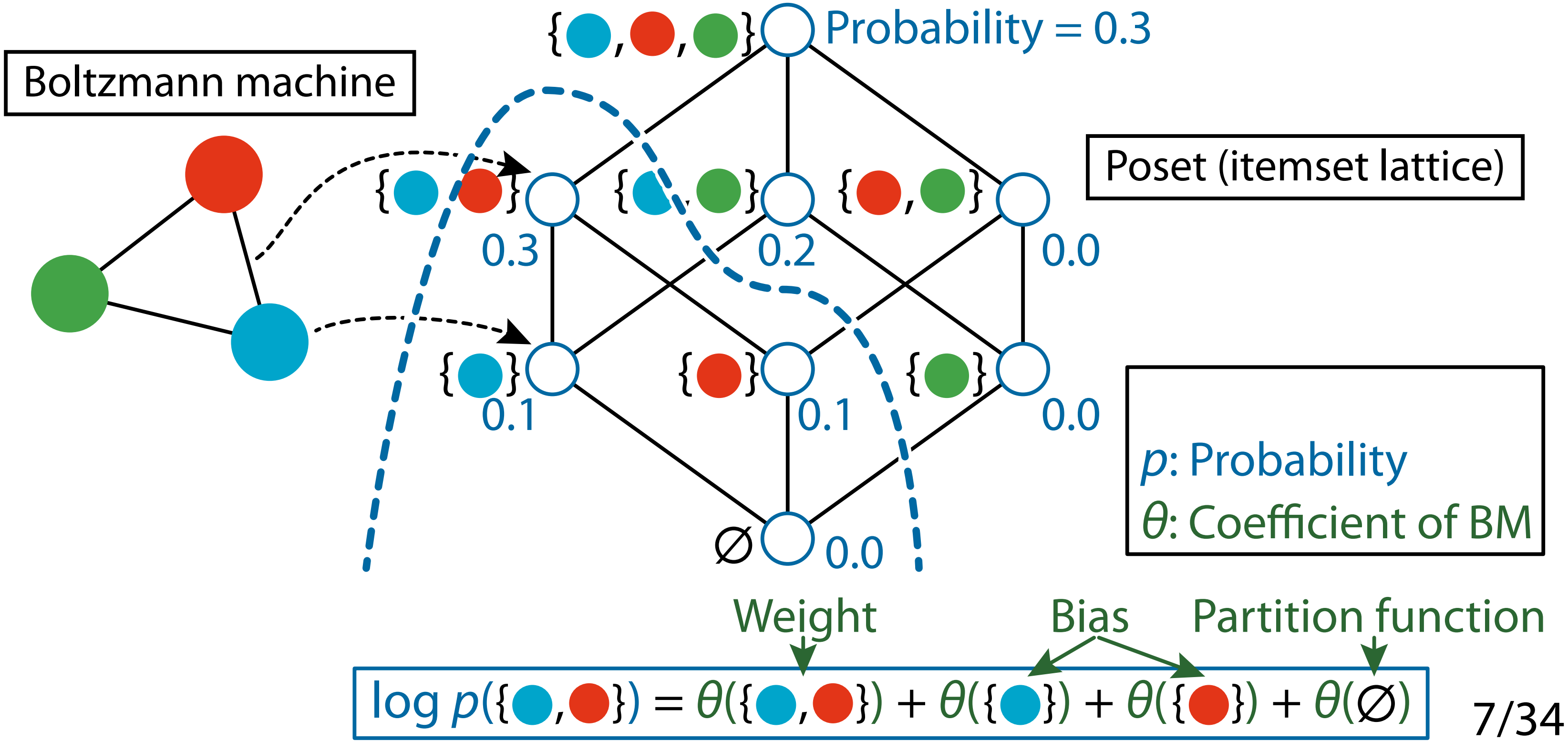
ID 1:	1	1	0
ID 2:	1	1	1
ID 3:	1	1	0
ID 4:	1	1	1
ID 5:	1	1	0
ID 6:	1	0	1
ID 7:	1	0	1
ID 8:	1	1	1
ID 9:	1	0	0
ID10:	0	1	0



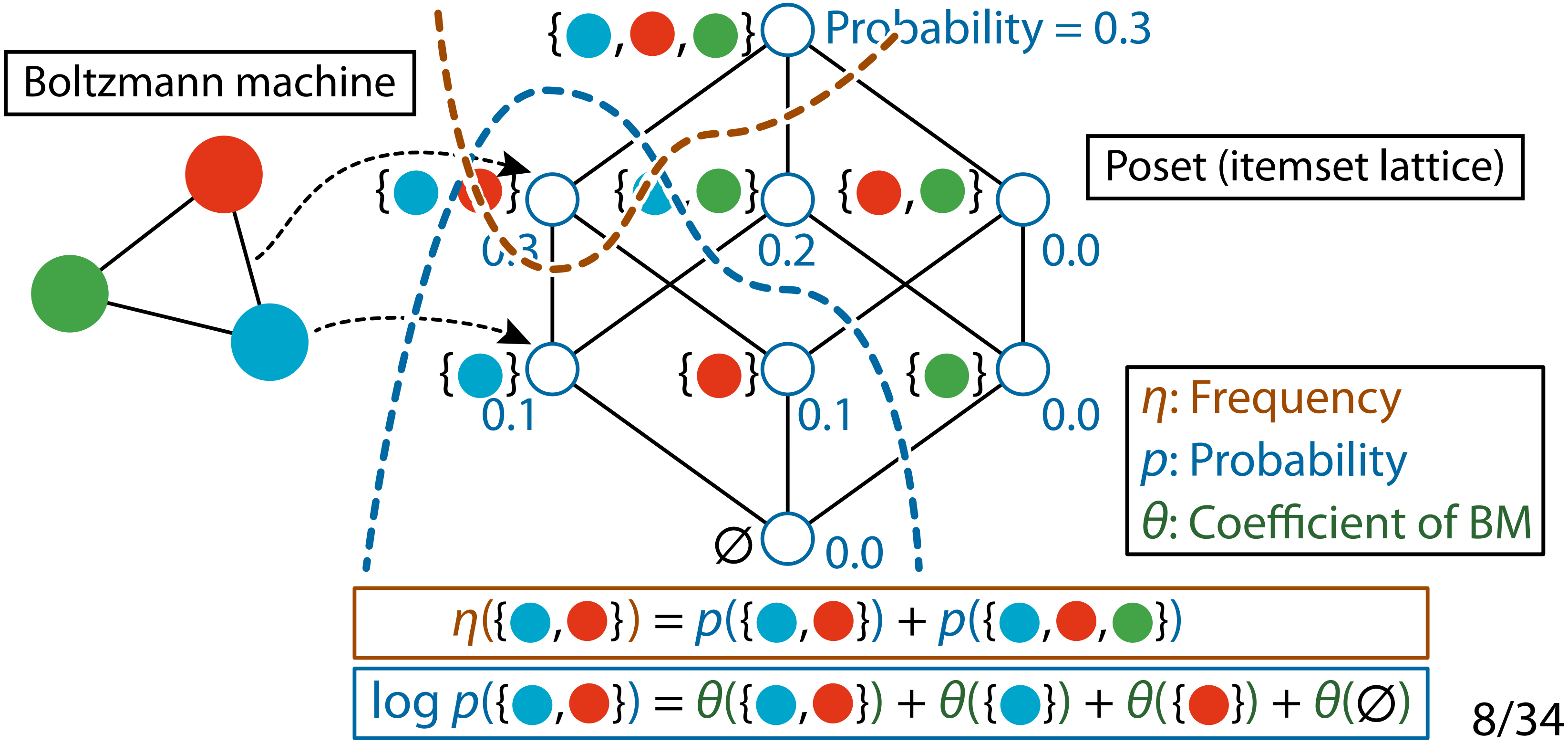
Boltzmann Machines



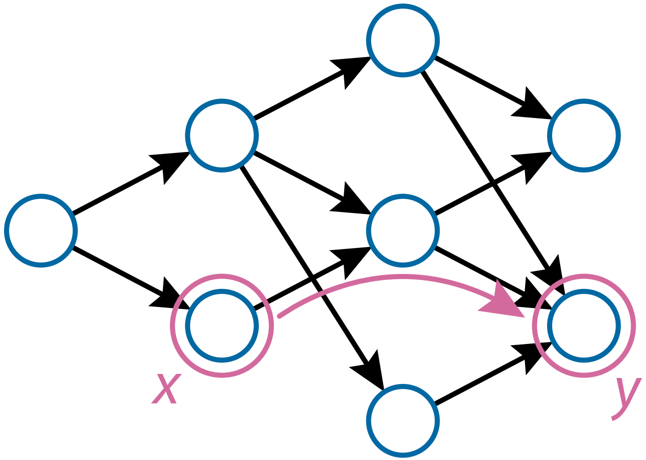
Boltzmann Machines → Downward Analysis



Pattern Mining & Boltzmann Machines



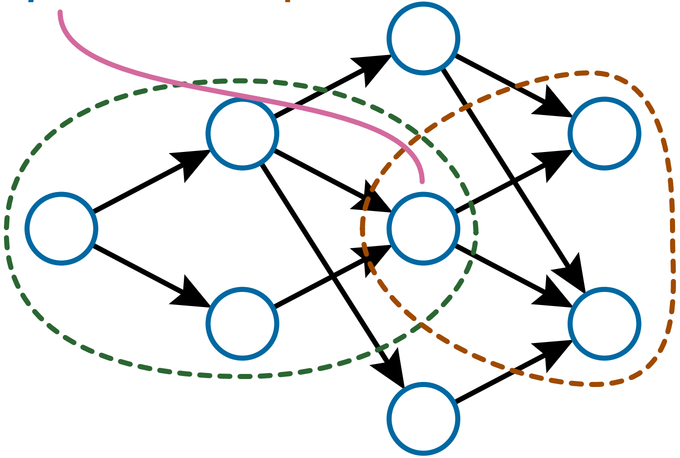
Partially Ordered Set



- 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

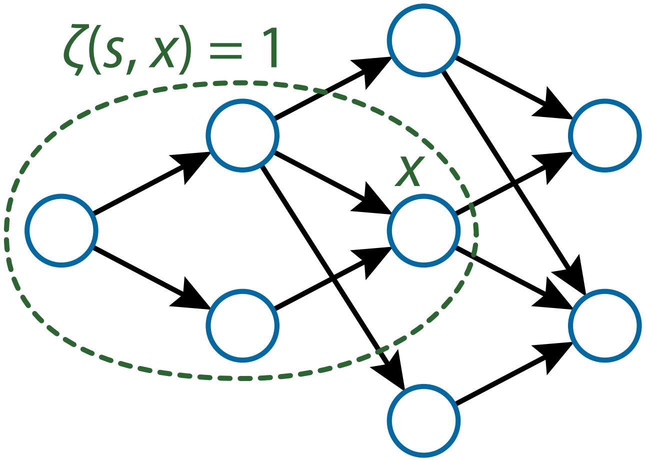
Log-Linear Model on Poset

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



- A probability vector $p: S \to (0, 1)$ 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
  $$\theta(x) = \sum_{s \in S} \mu(s, x) \log p(s), \;\; \eta(x) = \sum_{s \geq x} p(s)$$
- From the Möbius inversion formula:
  $$\log p(x) = \sum_{s \leq x} \theta(s), \;\; p(x) = \sum_{s \in S} \mu(x, s) \eta(s)$$

Möbius Function on Poset



- 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 $\zeta\mu = I$ (convolutional inverse):

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

Dually Flat Structure

- θ and η form a dual coordinate system:

  $\nabla\psi(\theta) = \eta,\ \nabla\varphi(\eta) = \theta$

  - $\psi(\theta) = -\theta(\perp) = -\log p(\perp),\ \varphi(\eta) = \sum_{x\in S} p(x)\log p(x)$
  - $\psi(\theta)$ and $\varphi(\eta)$ are connected via the Legendre transformation:

    $\varphi(\eta) = \max_{\theta'}\left(\theta'\eta - \psi(\theta')\right), \quad \theta'\eta = \sum_{x\in S\setminus\{\perp\}} \theta'(x)\eta(x)$

    - $\psi(\theta)$ and $\varphi(\eta)$ should be convex

Gradient and Riemannian Manifold

- The gradients: $g(\theta) = \nabla\nabla\psi(\theta) = \nabla\eta$, $g(\eta) = \nabla\nabla\varphi(\eta) = \nabla\theta$

$$
\begin{cases}
g_{xy}(\theta) = \dfrac{\partial\eta(x)}{\partial\theta(y)} = \displaystyle\sum_{s\in S} \zeta(x,s)\zeta(y,s)p(s) - \eta(x)\eta(y) \\
g_{xy}(\eta) = \dfrac{\partial\theta(x)}{\partial\eta(y)} = \displaystyle\sum_{s\in S} \mu(s,x)\mu(s,y)p(s)^{-1}
\end{cases}
$$

  - 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,

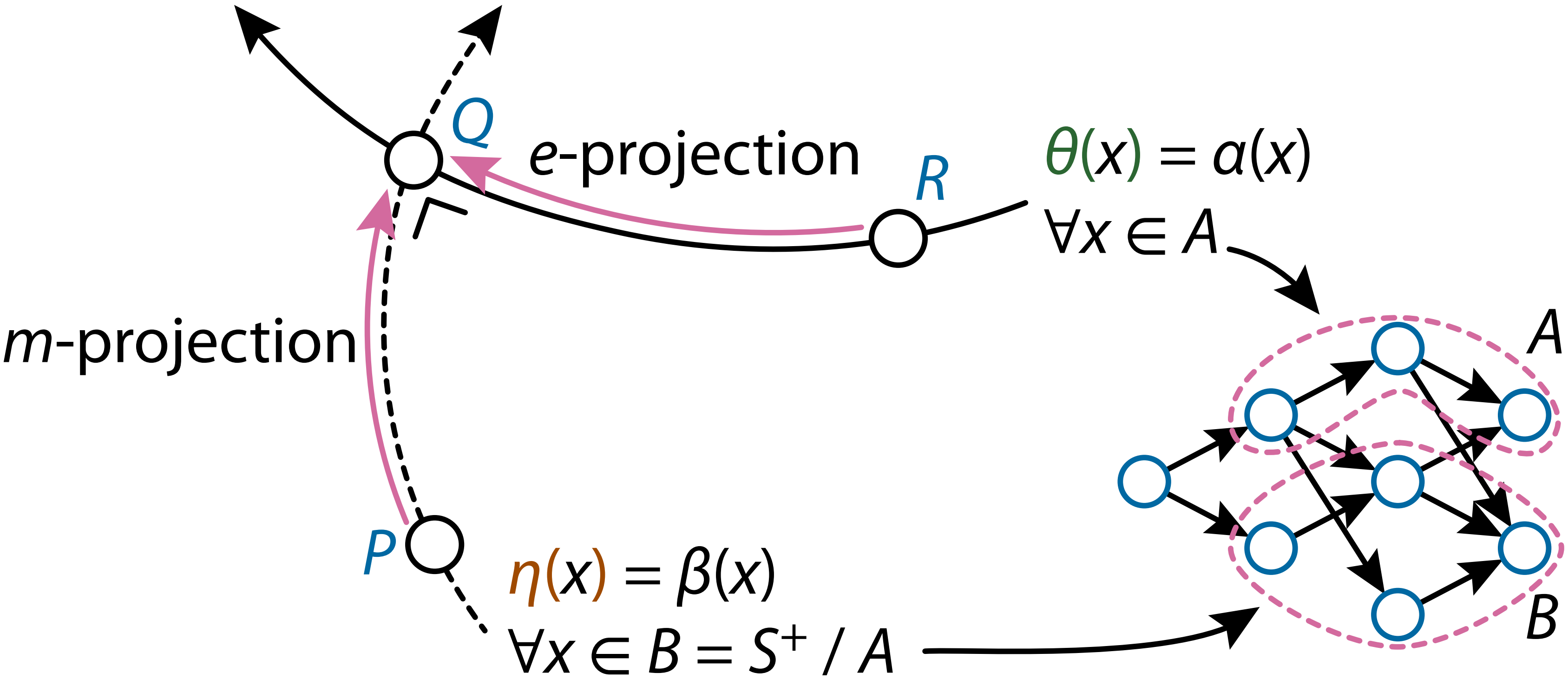
$$\mathbf{E}\left[\frac{\partial}{\partial\theta(x)}\log p(s)\frac{\partial}{\partial\theta(y)}\log p(s)\right] = g_{xy}(\theta),$$

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

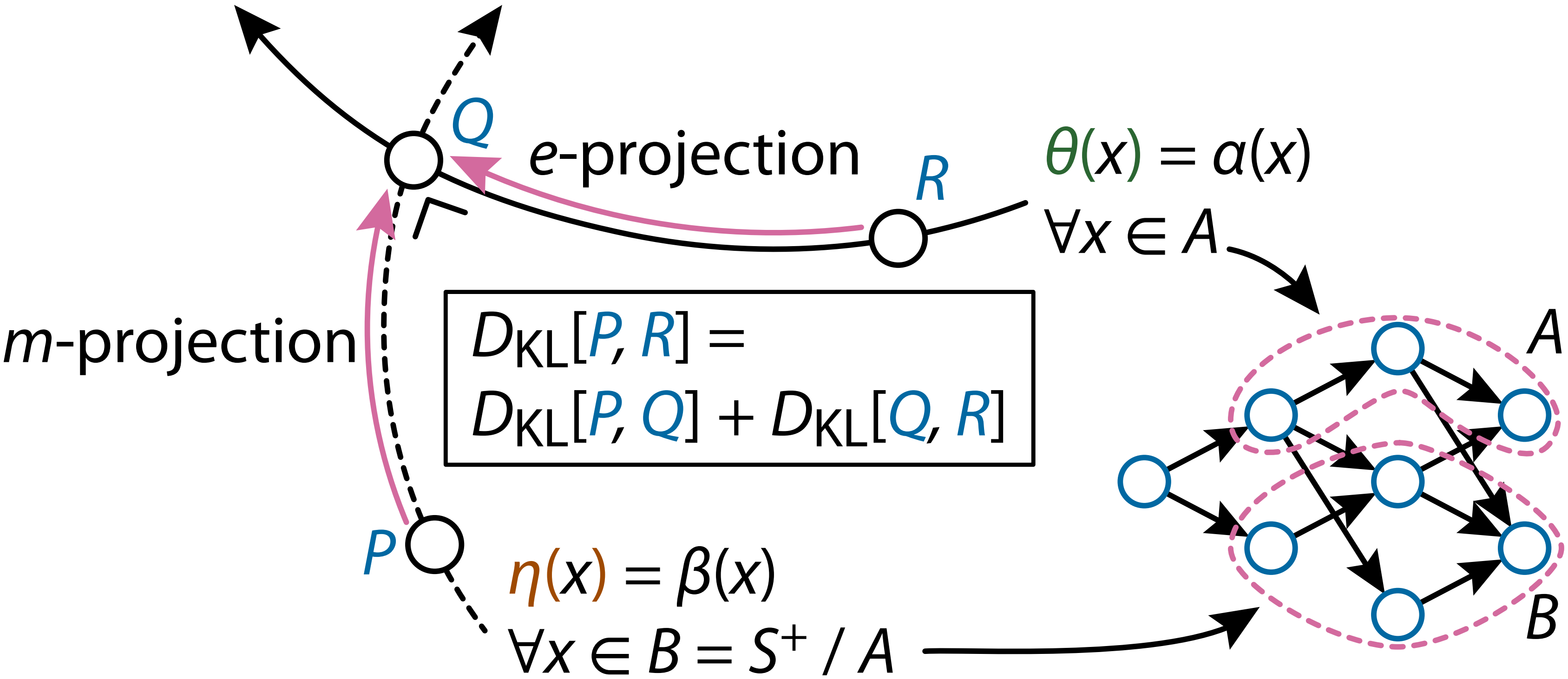
- θ and η are orthogonal, i.e.,

$$\mathbf{E}\left[\frac{\partial}{\partial\theta(x)}\log p(s)\frac{\partial}{\partial\eta(y)}\log p(s)\right] = \sum_{s\in S}\zeta(x,s)\mu(s,y) = \delta_{xy}$$

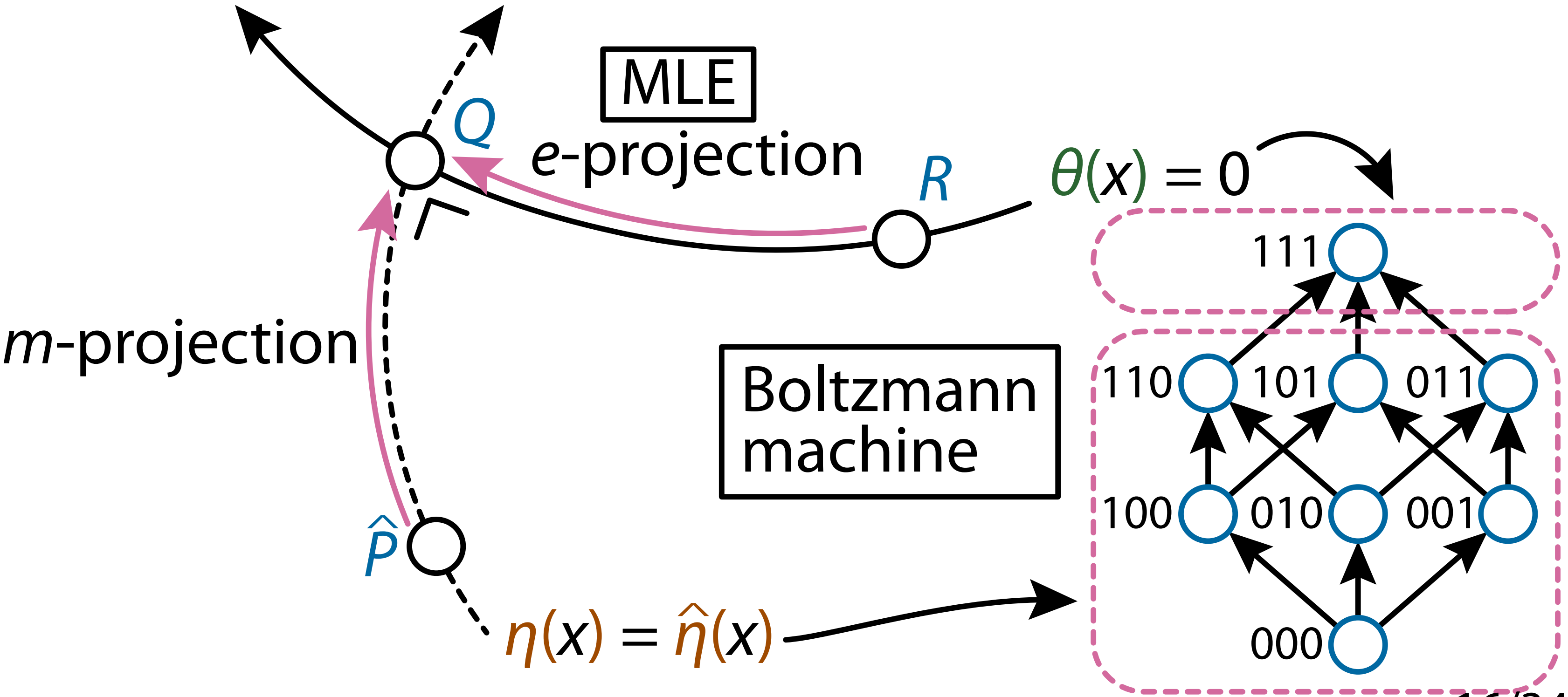
e-Projection and *m*-Projection



e-Projection and *m*-Projection



e-Projection and *m*-Projection



Computation of *e*-Projection

- Given P and β, we compute P_β such that

$$\begin{cases} \theta_{P_\beta}(x) = \theta_P(x) & \text{if } x \in (S \setminus \{\perp\}) \setminus \operatorname{dom}(\beta), \\ \eta_{P_\beta}(x) = \beta(x) & \text{if } x \in \operatorname{dom}(\beta) \end{cases}$$

- Initialize with $P_\beta^{(0)} = P$ and, at each step t, update $\eta_{P_\beta}^{(t)}(x)$ for $x \in \operatorname{dom}(\beta)$
  - Since θ and η are orthogonal, we can change $\eta_{P_\beta}^{(t)}(x)$ while fixing $\theta_{P_\beta}^{(t)}(y)$ for $y \notin \operatorname{dom}(\beta)$

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: $g(\theta) = \nabla\nabla\psi(\theta) = \nabla\eta$ and $g(\eta) = \nabla\nabla\varphi(\eta) = \nabla\theta$

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

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

Newton's Method (1/2)

- Each step of Newton's method:

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

  - J is the $|\mathrm{dom}(\beta)| \times |\mathrm{dom}(\beta)|$ 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 $x, y \in \mathrm{dom}(\beta)$

Newton's Method (2/2)

- Each update is

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

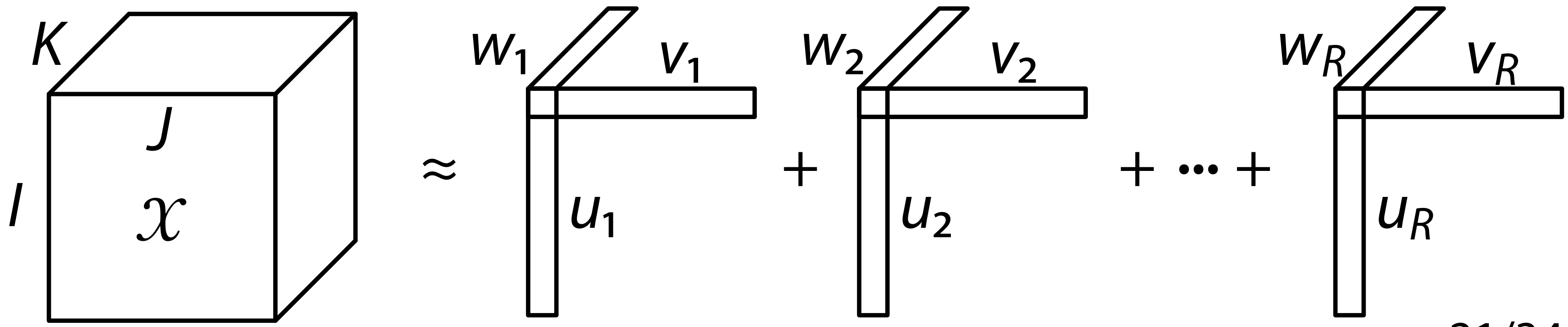
  - J^{-1} is the inverse of J
  - J is the $|\mathrm{dom}(\beta)| \times |\mathrm{dom}(\beta)|$ Jacobian matrix given as

$$J_{xy} = \frac{\partial \eta_{P_\beta}^{(t)}(x)}{\partial \theta_{P_\beta}^{(t)}(y)} = \sum_{s \in S} \zeta(x, s)\zeta(y, s) p_\beta^{(t)}(s) - \eta_{P_\beta}^{(t)}(x)\eta_{P_\beta}^{(t)}(y)$$

for each $x, y \in \mathrm{dom}(\beta)$

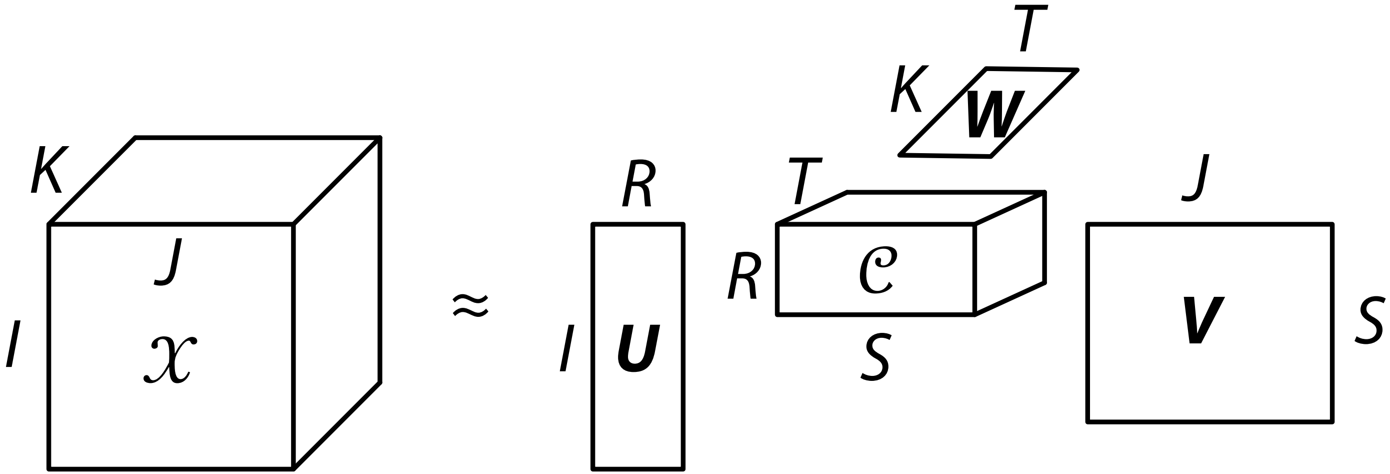
CP Decomposition

- Approximate a tensor $\mathcal{X}$ by R rank-1 tensors:
  $x_{ijk} \approx \sum_{r=1}^{R} u_{ir} v_{jr} w_{kr}$
  - Number of parameters $IJK \to R(I + J + K)$



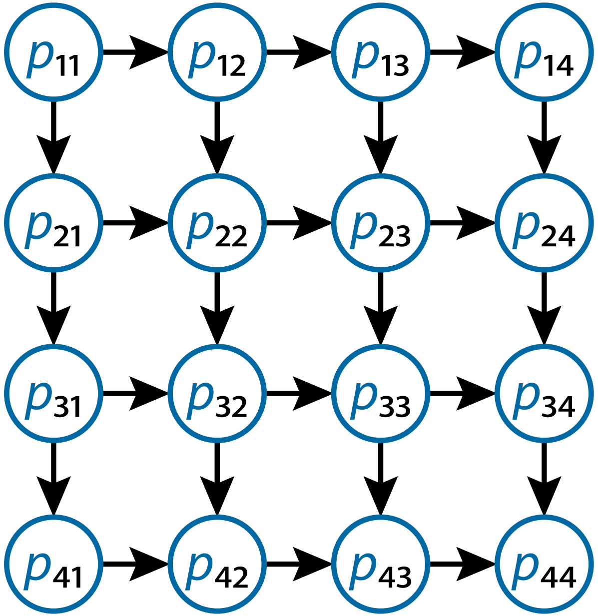
Tucker Decomposition

- Approximate a tensor $\mathcal{X}$ by three matrices and a core tensor:
  $x_{ijk} \approx \sum_{r=1}^{R} \sum_{s=1}^{S} \sum_{t=1}^{T} c_{rst} u_{ir} v_{js} w_{kt}$
  - Number of parameters $IJK \to RST + IR + JS + KT$



From Matrix to Poset (DAG)

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



Legendre Decomposition

Input matrix:

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

Legendre Decomposition

Input matrix:

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

Reconstructed matrix:

Legendre Decomposition

Input matrix:

Reconstructed matrix:

$$\begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} & p_{34} \\ p_{41} & p_{42} & p_{43} & p_{44} \end{bmatrix} \xrightarrow{\text{Legendre decomposition}} \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} & \theta_{14} \\ \theta_{21} & \theta_{22} & 0 & 0 \\ \theta_{31} & \theta_{32} & 0 & 0 \\ \theta_{41} & \theta_{42} & 0 & 0 \end{bmatrix}$$

θ: parameter

Decomposition basis (arbitrary subset of indices)

Legendre Decomposition

Input matrix: Reconstructed matrix:

$$\begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} & p_{34} \\ p_{41} & p_{42} & p_{43} & p_{44} \end{bmatrix} \xrightarrow{\text{Legendre decomposition}} \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} & \\ \theta_{21} & \theta_{22} & 0 & \\ \theta_{31} & \theta_{32} & \boxed{0} & \\ & & & \end{bmatrix}$$

$$q_{ij} = \prod_{k \leq i,\, l \leq j} \exp(\theta_{kl})$$

Legendre Decomposition

Input matrix:

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

$\xrightarrow{\text{Legendre decomposition}}$

Reconstructed matrix:

$$\begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} & \theta_{14} \\ \theta_{21} & \theta_{22} & 0 & \boxed{0} \\ & & & \\ & & & \end{bmatrix}$$

$$q_{ij} = \prod_{k \le i,\, l \le j} \exp(\theta_{kl})$$

Legendre Decomposition

Input matrix:

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

Legendre decomposition $\longrightarrow$

Reconstructed matrix:

$$\begin{bmatrix} \theta_{11} & \theta_{12} & \boxed{\theta_{13}} & \\ & & & \\ & & & \\ & & & \end{bmatrix}$$

$$q_{ij} = \prod_{k \le i,\, l \le j} \exp(\theta_{kl})$$

Legendre Decomposition

Input matrix:

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

Reconstructed matrix:

$$\begin{bmatrix} \eta_{11} & \eta_{12} & \eta_{13} & \eta_{14} \\ \eta_{21} & \eta_{22} & & \\ \eta_{31} & \eta_{32} & & \\ \eta_{41} & \eta_{42} & & \end{bmatrix}$$

η: constraint

Decomposition basis (arbitrary subset of indices)

Legendre Decomposition

Input matrix:

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

Legendre decomposition $\longrightarrow$

Reconstructed matrix:

$$\begin{bmatrix} & & & \\ & \boxed{q_{22}} & q_{23} & q_{24} \\ & q_{32} & q_{33} & q_{34} \\ & q_{42} & q_{43} & q_{44} \end{bmatrix}$$

$$\eta_{ij} = \sum_{k \geq i,\, l \geq j} q_{kl}$$

Legendre Decomposition

Input matrix:

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

Reconstructed matrix:

$$\begin{bmatrix} & & \boxed{q_{13}} & q_{14} \\ & & q_{23} & q_{24} \\ & & q_{33} & q_{34} \\ & & q_{43} & q_{44} \end{bmatrix}$$

$$\eta_{ij} = \sum_{k \geq i,\, l \geq j} q_{kl}$$

Legendre Decomposition

Input matrix:

$$\begin{bmatrix} \hat{\eta}_{11} & \hat{\eta}_{12} & \hat{\eta}_{13} & \hat{\eta}_{14} \\ \hat{\eta}_{21} & \hat{\eta}_{22} & & \\ \hat{\eta}_{31} & \hat{\eta}_{32} & & \\ \hat{\eta}_{41} & \hat{\eta}_{42} & & \end{bmatrix}$$

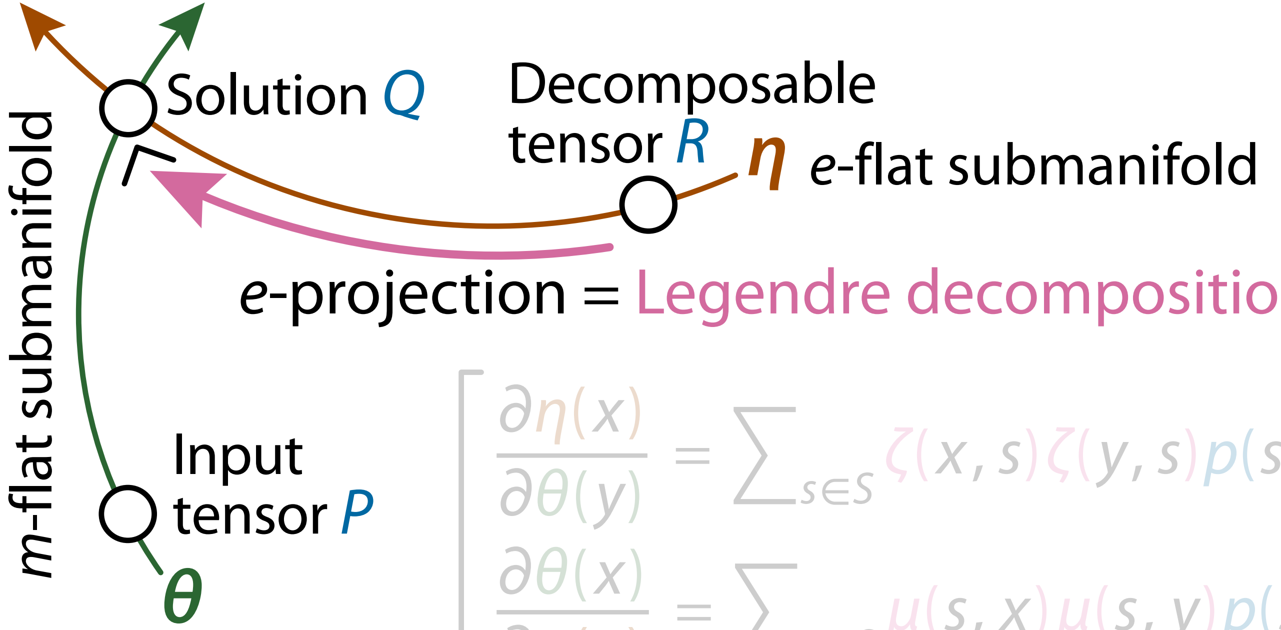
Legendre decomposition $\longrightarrow$

Find $\boldsymbol{\theta}$ s.t. $\hat{\boldsymbol{\eta}} = \boldsymbol{\eta}$

Reconstructed matrix:

$$\begin{bmatrix} \eta_{11} & \eta_{12} & \eta_{13} & \eta_{14} \\ \eta_{21} & \eta_{22} & & \\ \eta_{31} & \eta_{32} & & \\ \eta_{41} & \eta_{42} & & \end{bmatrix}$$

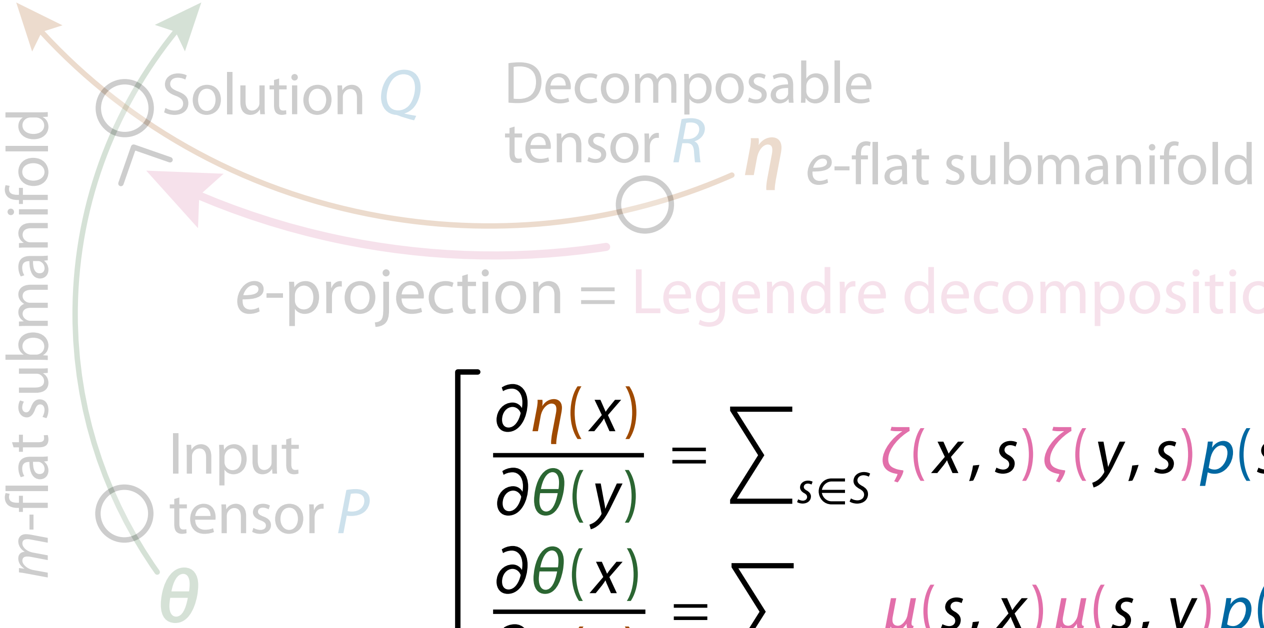
Information Geometry



e-projection = Legendre decomposition (= MLE)

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

Information Geometry



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

Summary of Lectrue

- Information geometric formulation connects pattern mining and Boltzmann machines
  - Applications including matrix balancing
  - Sugiyama, M., Nakahara, H., Tsuda, K.: **Tensor Balancing on Statistical Manifold**, *ICML 2017*

- **Discrete structure (posets)** + **Information Geometry** = **Strong formulation for data analysis**!

- Further application to tensor decomposition:
  - Sugiyama, M., Nakahara, H., Tsuda, K.: **Legendre Decomposition for Tensors**, *NeurIPS 2018*