Mar. 18, 2025
FDIG2025

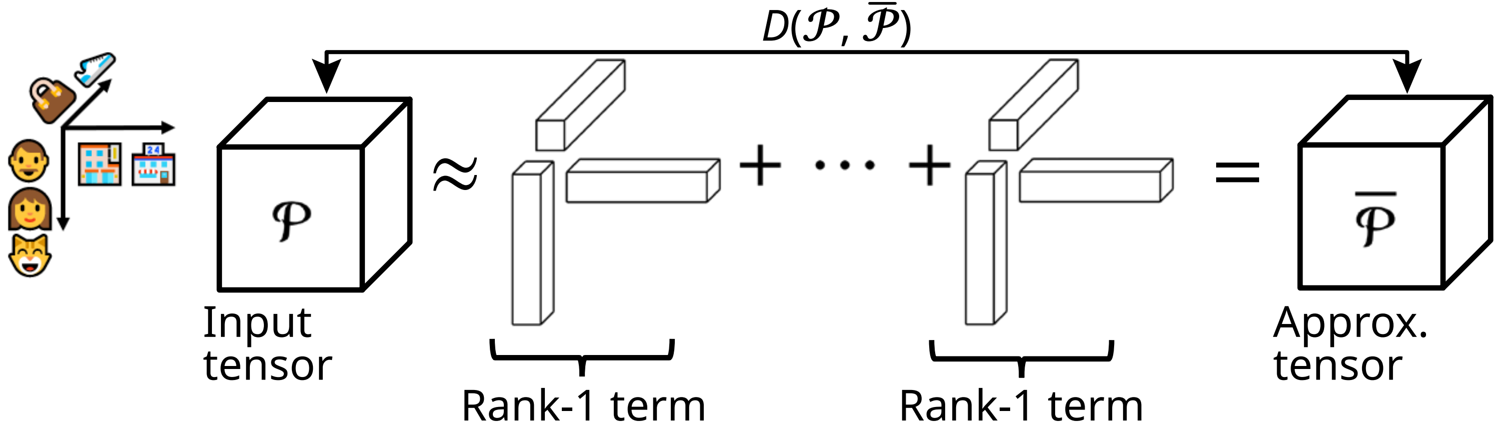
Inter-University Research Institute Corporation / Research Organization of Information and Systems
National Institute of Informatics

A Unified Information Geometric Perspective on Machine Learning in Structured Spaces

Mahito Sugiyama (NII) https://mahito.nii.ac.jp/

Nonnegative Tensor Decomposition

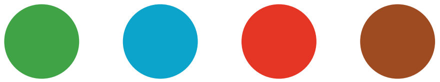
- Low-rank decomposition is commonly used in analysis of multi-dimensional arrays such as matrices and tensors
  - Approximate them by a linear combination of bases
  - Tune the number of bases (called rank) as a hyper-parameter



Many-Body Approximation for Tensors

One-body approximation

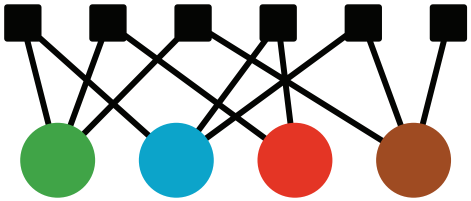
$\mathcal{P}_{ijkl} = p_i^{(1)} p_j^{(2)} p_k^{(3)} p_l^{(4)}$



= Rank-1 approximation
(mean-field approximation)

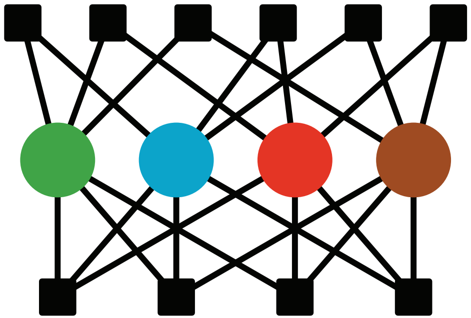
Two-body approximation

$\mathcal{P}_{ijkl} = X_{ij}^{(12)} X_{ik}^{(13)} \cdots X_{kl}^{(34)}$



Three-body approximation

$\mathcal{P}_{ijkl} = \chi_{ijk}^{(123)} \chi_{ijl}^{(124)} \chi_{ikl}^{(134)} \chi_{jkl}^{(234)}$



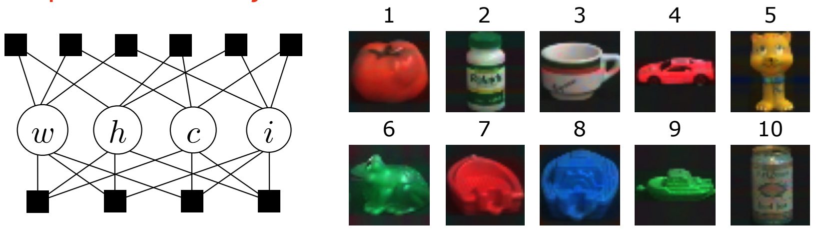
→ Larger capability

Ghalamkari, Sugiyama, & Kawahara, **Many-body Approximation for Non-negative Tensors**, NeurIPS2023 [Code (Julia)] [Code (Python)]

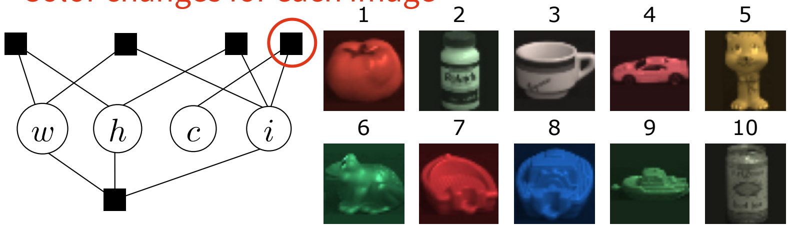
Application to Images

a. Up to four-body

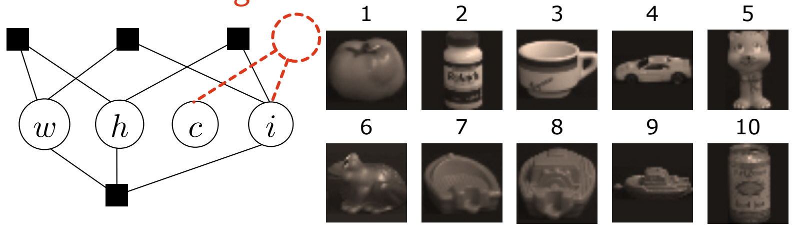
b. Up to three-body



c. Color changes for each image



d. When the interaction between colors and image indices is removed...



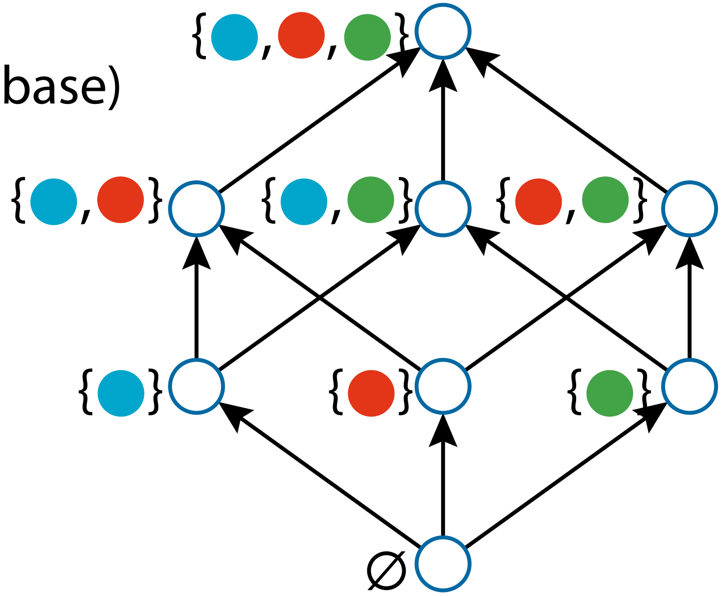
Properties and Related Work

- Optimization is always convex
- Generalization of Boltzmann machines
- Linear algebraic operations can be achieved by identifying a sample space with a tensor
  - Fast balancing [Sugiyama et al. ICML2017]
  - Legendre decomposition [Sugiyama et al. NeurIPS2018]
  - Fast and stable tensor low-rank approximation [Ghalamkari & Sugiyama, NeurIPS2021, Info.Geo. (2023)]
- An application to quantum chemistry calculations [Hagai et al. Digital Discovery (2023)]

Itemset 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

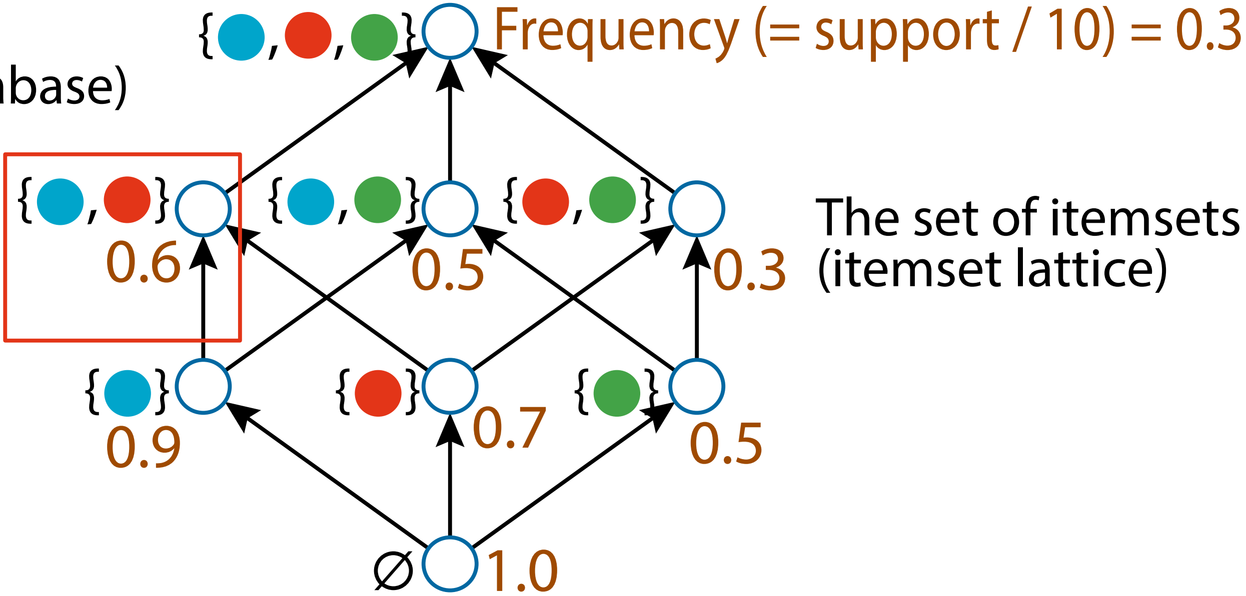


The set of itemsets
(itemset lattice)

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

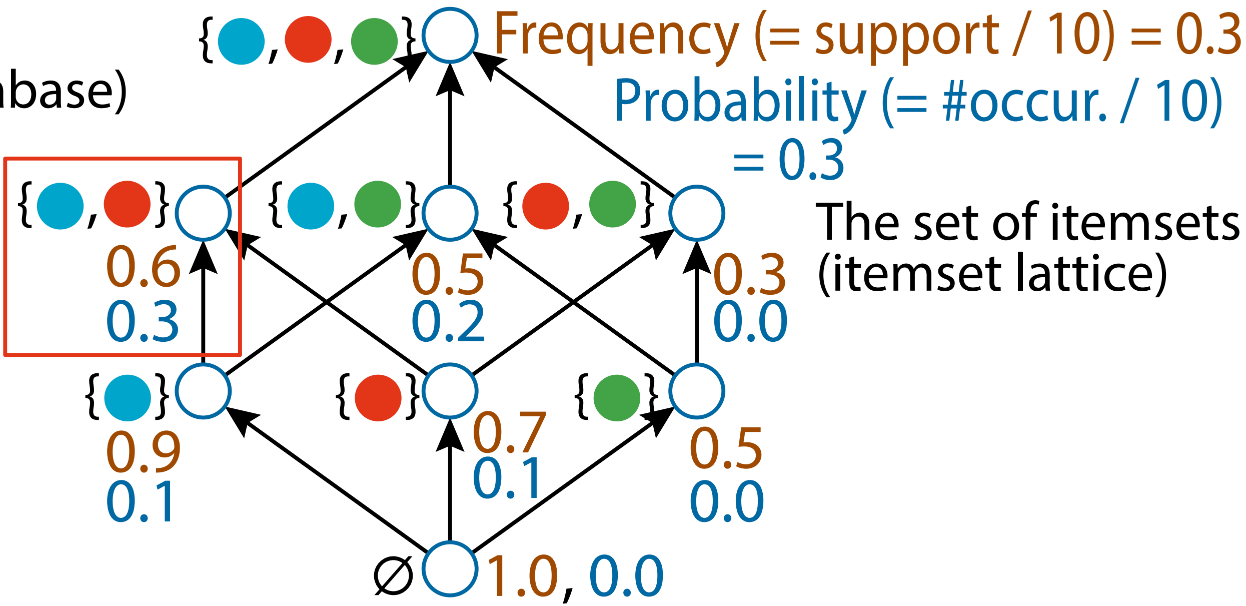


The set of itemsets
(itemset lattice)

Probability Distribution on Itemset Lattice

Binary vectors
(Transaction database)

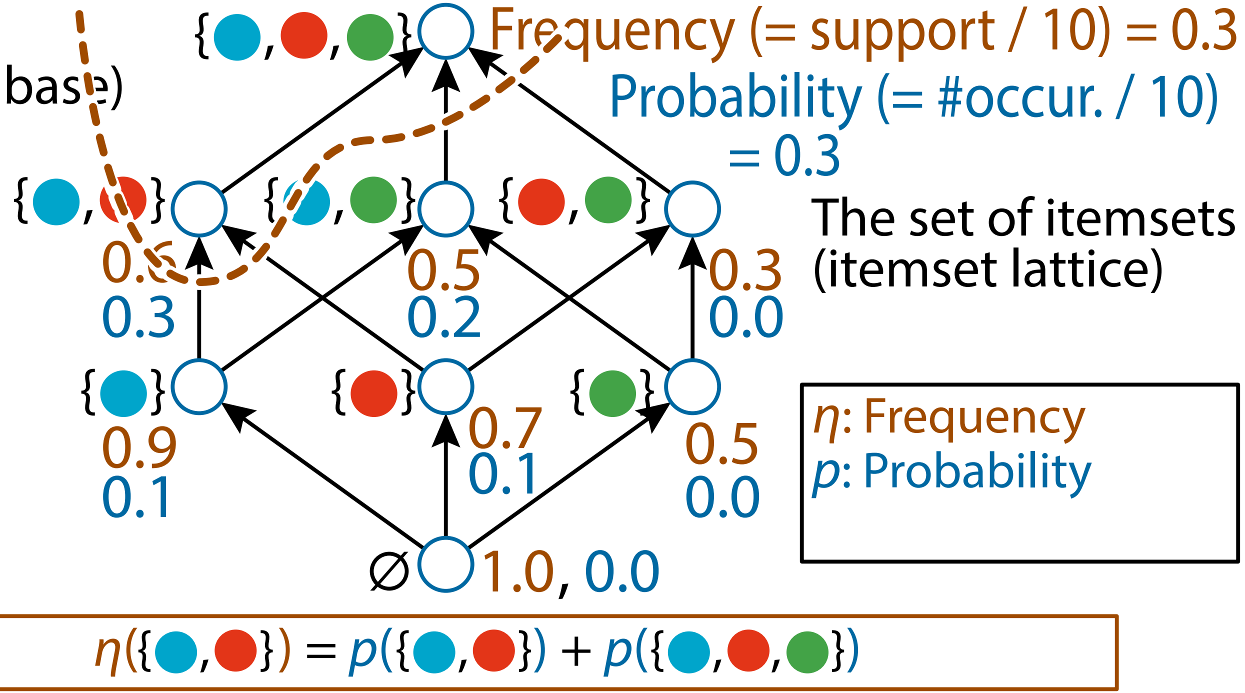
ID 1:	1	1	0
ID 2:	1	1	1
ID 3:	1	1	0
ID 4:	1	1	1
ID 5:	1	1	0
ID 6:	1	0	1
ID 7:	1	0	1
ID 8:	1	1	1
ID 9:	1	0	0
ID10:	0	1	0



Itemset Mining

Binary vectors
(Transaction database)

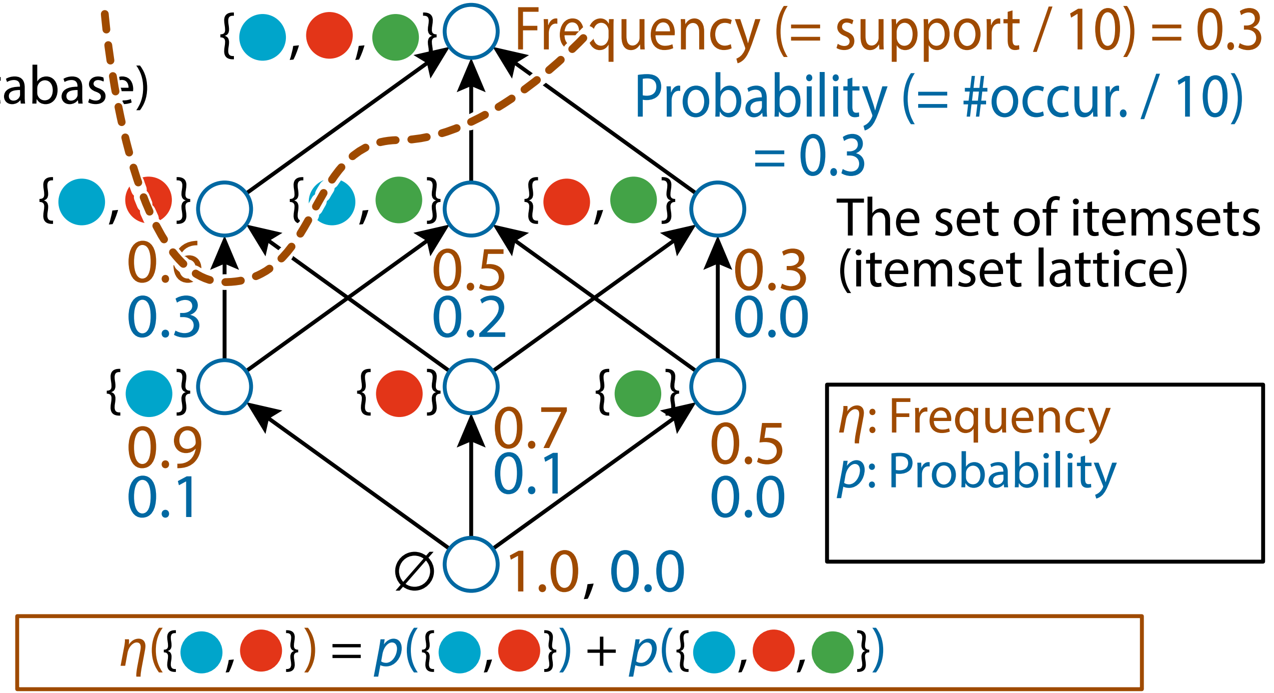
	●	●	●
ID 1:	1	1	0
ID 2:	1	1	1
ID 3:	1	1	0
ID 4:	1	1	1
ID 5:	1	1	0
ID 6:	1	0	1
ID 7:	1	0	1
ID 8:	1	1	1
ID 9:	1	0	0
ID10:	0	1	0



Itemset Mining → Upward Analysis

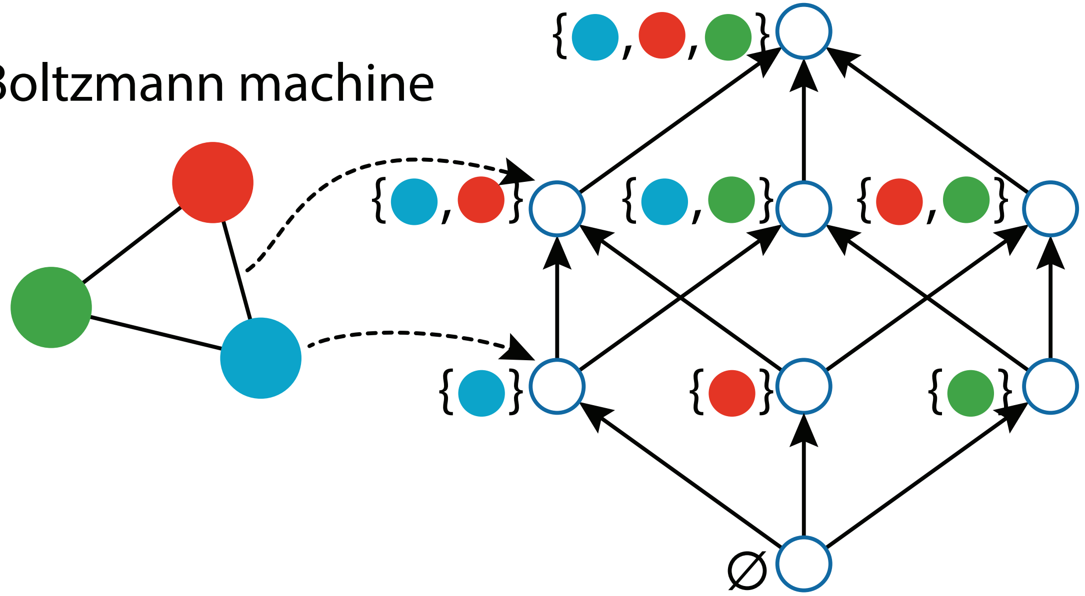
Binary vectors
(Transaction database)

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



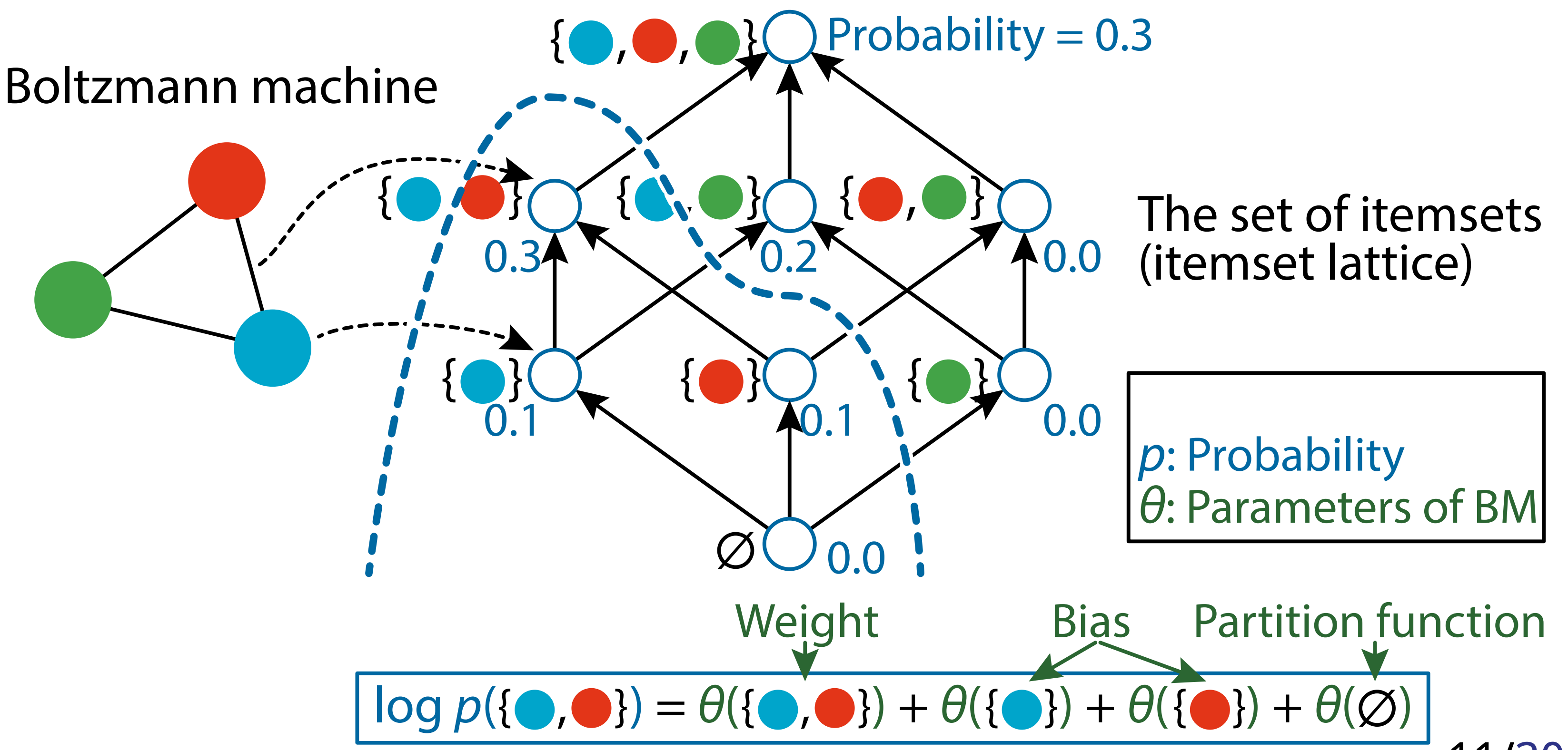
Boltzmann Machines

Boltzmann machine

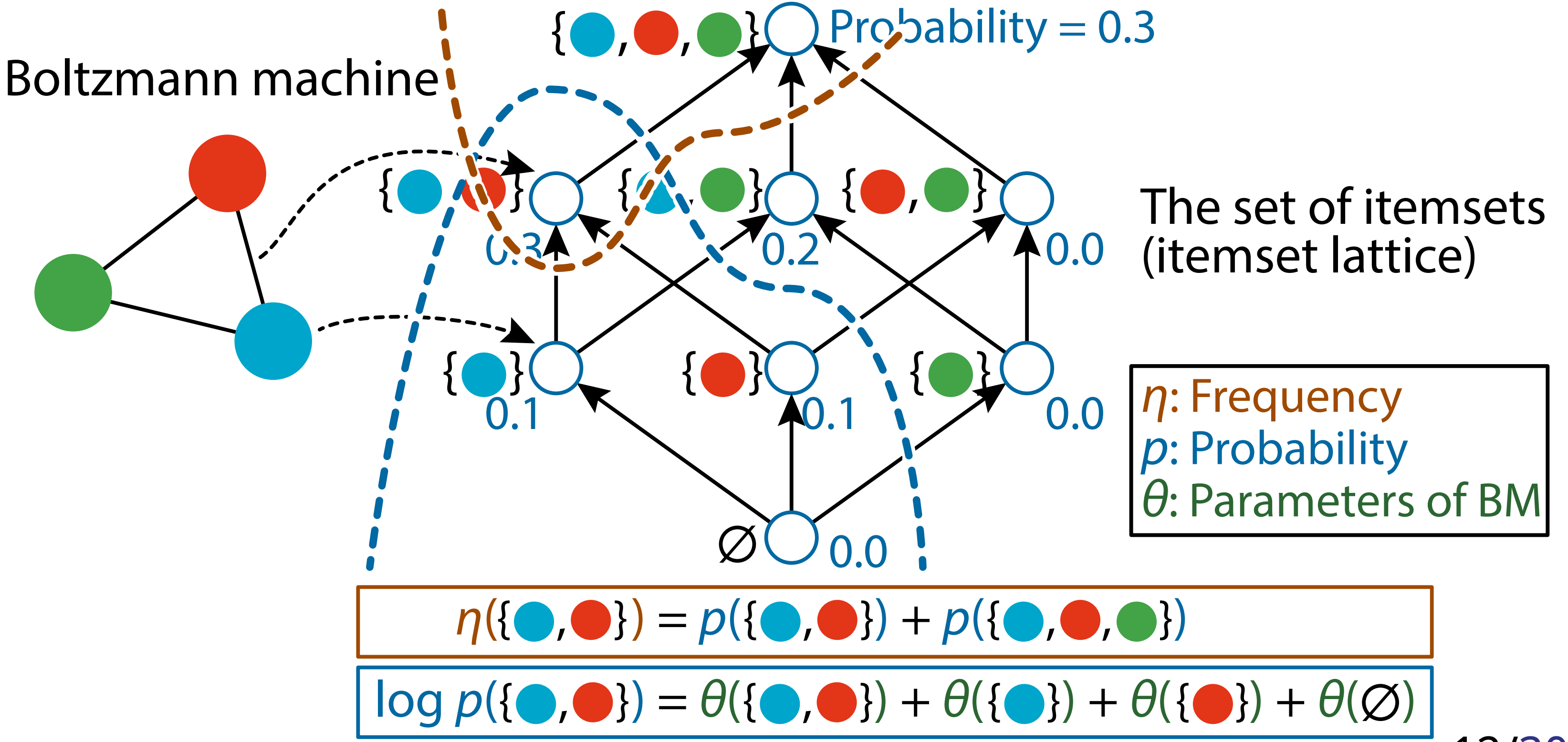


The set of itemsets
(itemset lattice)

Boltzmann Machines → Downward Analysis

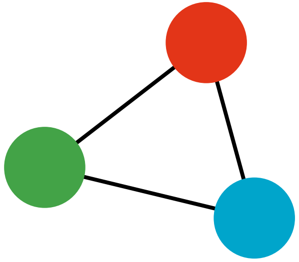


Itemset Mining & Boltzmann Machines



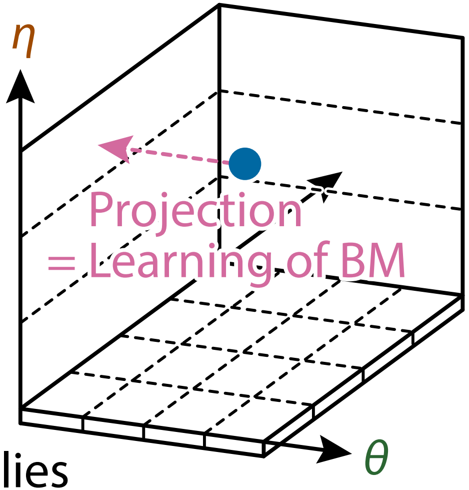
Information Geometric Understanding

Boltzmann machine



Information geometric understanding

Statistical manifold
(each point is a distribution)



Representation by graphical model

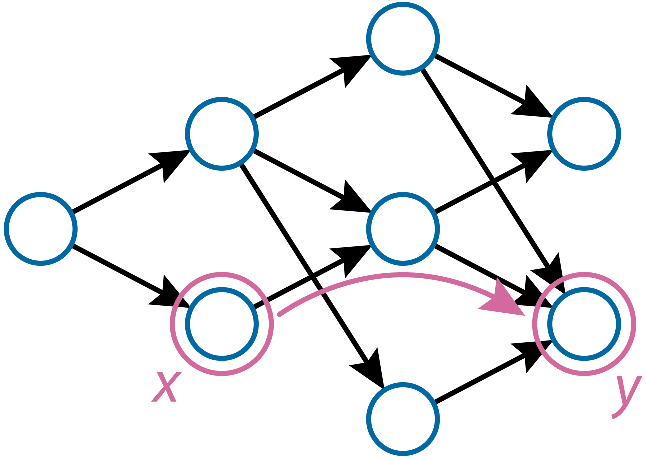
(θ, η) coordinate system for exponential families

Gibbs distribution:

$$\log p(\boldsymbol{x}) = \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_{12} x_1 x_2 + \theta_{13} x_1 x_3 + \theta_{23} x_2 x_3 + \psi$$

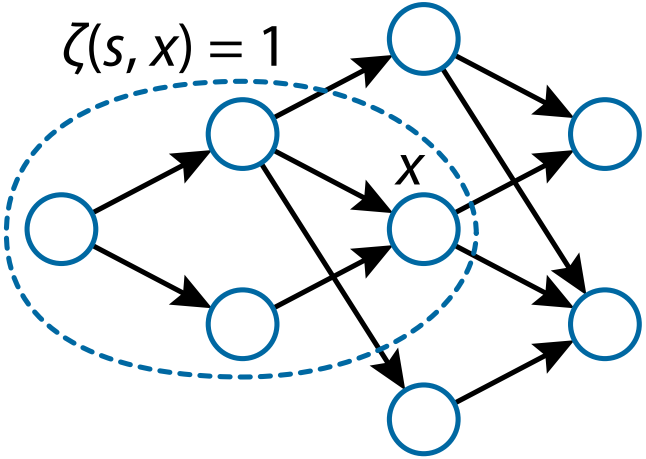
for each $\boldsymbol{x} = (x_1, x_2, x_3) \in \{0, 1\}^3$

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 $\perp \in S$
- Equivalent to a DAG
  - Each $x \in S$ is a node
  - $x \leq y \iff y$ is reachable from x
- Itemset lattice is a poset, where "$\leq$" is "$\subseteq$"

Zeta and Möbius Functions



- Zeta function $\zeta : S \times S \to \{0, 1\}$

  $$\zeta(s, x) = \begin{cases} 1 & \text{if } s \leq x, \\ 0 & \text{otherwise.} \end{cases}$$

  - (integral)
- Möbius function $\mu = \zeta^{-1}$
  - $\zeta\mu = \delta$, where $\delta_{xy} = 1$ if $x = y$ and $\delta_{xy} = 0$ otherwise
  - (differential)
- Incidence algebra is induced

The Log-Linear Model on Posets

Probability

Zeta function
(encoding partial order structure)

Coefficient of log-linear model
(Bias/weight in Boltzmann machines)
(Natural parameter of exponential family)

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

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

Möbius function

η θ Dual coordinates in IG

Expectation
(Frequency in itemset mining)
(Expectation parameter in exponential family)

Sugiyama, Nakahara & Tsuda, **Tensor Balancing on Statistical Manifold**, ICML2017

Mixed Coordinate System

- Many problems are formulated as coordinate mixing

$$P = (\theta_1, \theta_2, \ldots, \theta_{i-1}, \theta_i, \theta_{i+1}, \ldots, \theta_n)$$

$$Q = (\eta_1, \eta_2, \ldots, \eta_{i-1}, \theta_i, \theta_{i+1}, \ldots, \theta_n)$$

$$R = (\eta_1, \eta_2, \ldots, \eta_{i-1}, \eta_i, \eta_{i+1}, \ldots, \eta_n)$$

e-projection (MLE)

m-projection

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

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

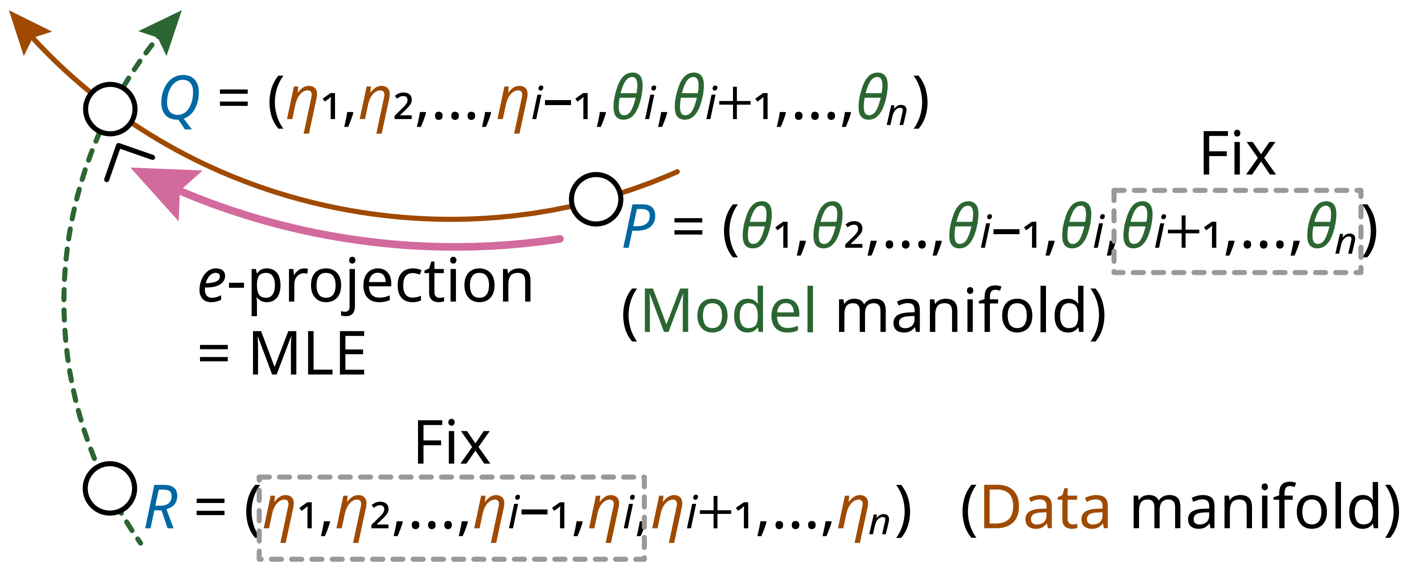
$$Q = (\hat{\eta}_1, \hat{\eta}_2, \ldots, \hat{\eta}_{i-1}, 0, 0, \ldots, 0)$$

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

Pythagorean theorem: (Q is always unique)

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

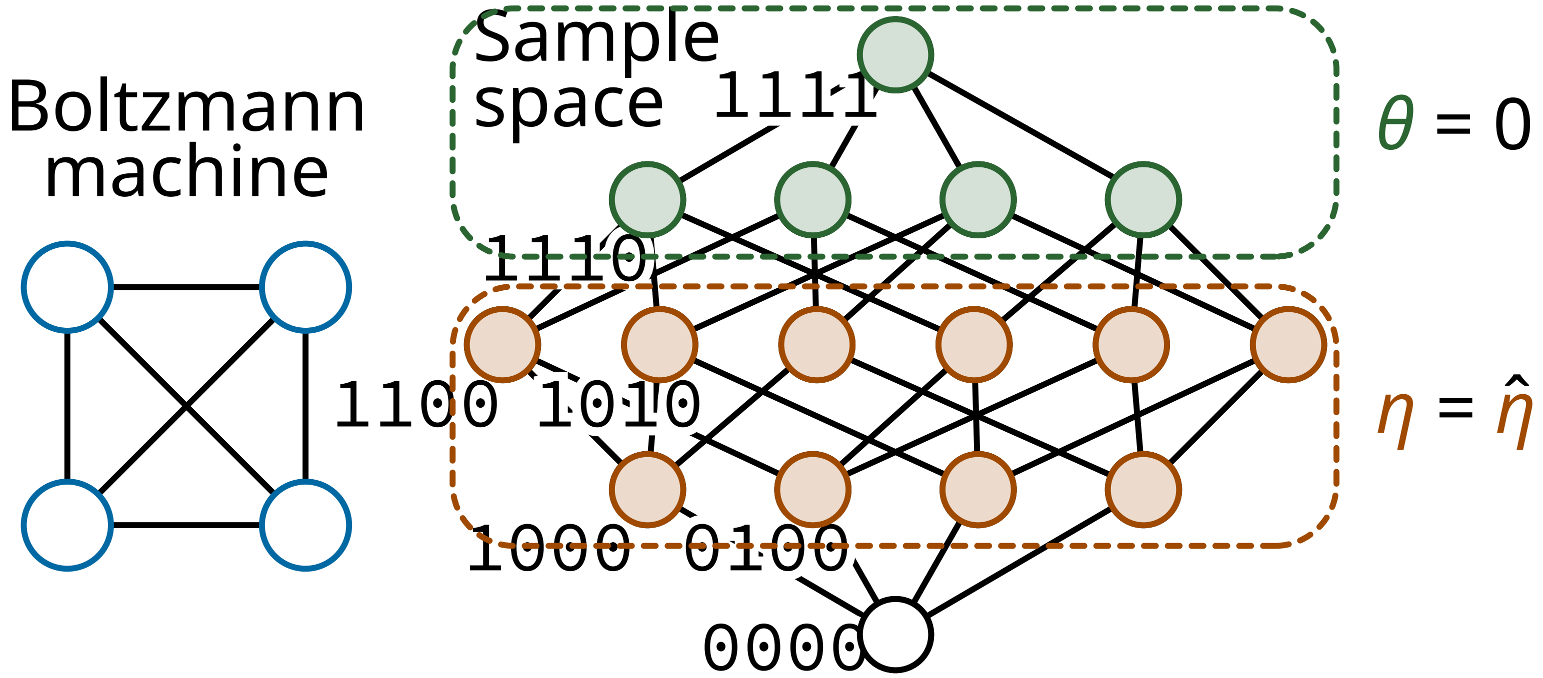
Two Submanifolds



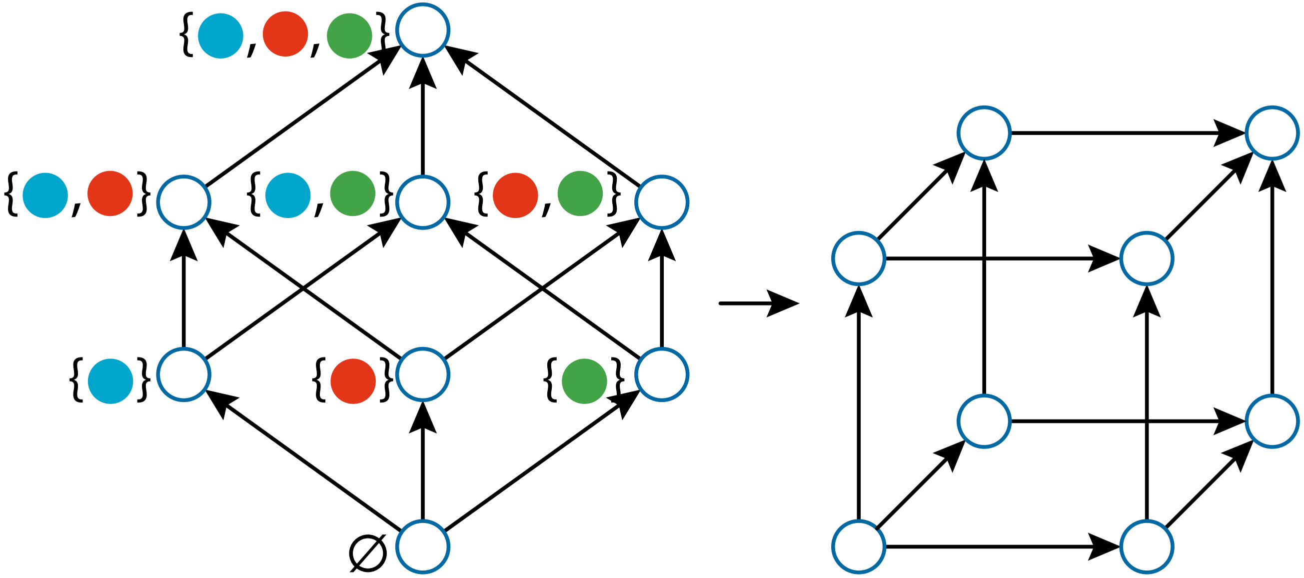
$\theta_{\text{next}} \leftarrow \theta - \varepsilon(\eta - \hat{\eta}_{\text{target}})$ or

$\theta_{\text{next}} \leftarrow \theta - G^{-1}(\eta - \hat{\eta}_{\text{target}})$ (natural grad., G: FIM)

Boltzmann Machine Training

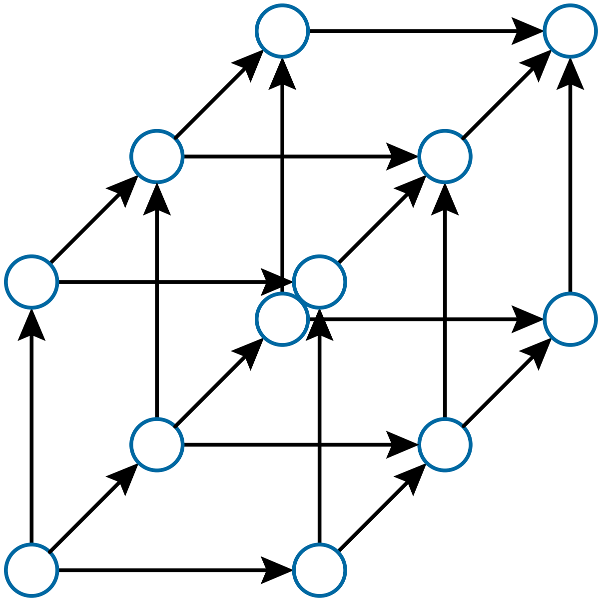


From Itemset Lattices to Tensors

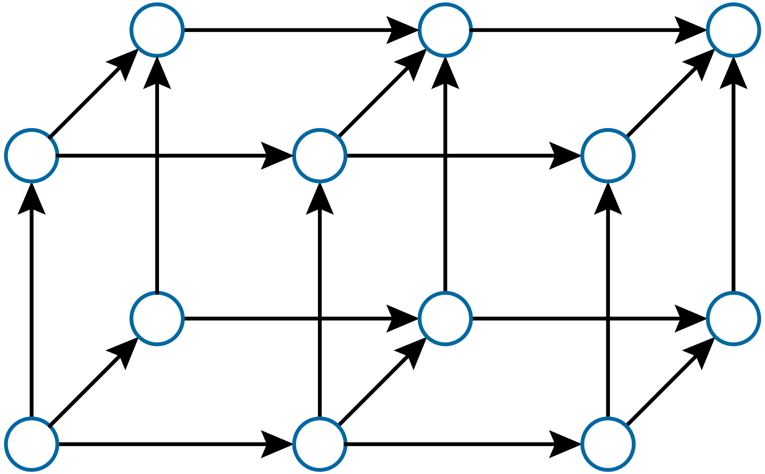


The number of features = The number of modes

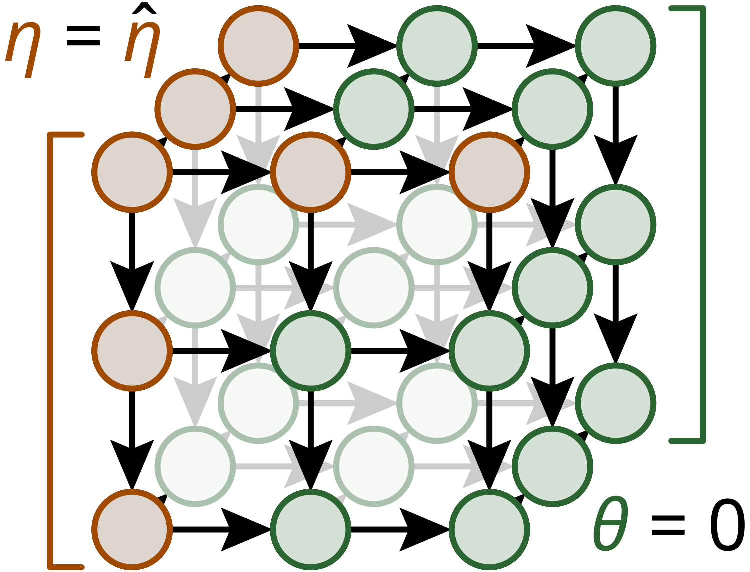
Any Tensor Can Be Treated



Any Tensor Can Be Treated



Rank-1 (or One-Body) Approximation



Set all θ to be 0 except for corners
⇒ • Rank-1 best approximation
  • Mean-field approximation

A closed form solution is available
(gradient method is not needed)

Apply rank-1 approximation
to sub-tensors
⇒ Any low-rank approximation is
achieved without a gradient method

Formulation of Many-body Approximation

- We explicitly model associations between features/modes (no latent variables, hence convex)

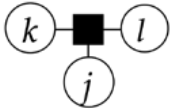
$$\mathcal{P}_{ijkl} = \exp\left[\sum_{i'=1}^{i}\sum_{j'=1}^{j}\sum_{k'=1}^{k}\sum_{l'=1}^{l}\theta_{i'j'k'l'}\right]$$

$$\sum_{k'=2}^{k}\sum_{l'=2}^{l}\theta_{11k'l'} \qquad \sum_{j'=2}^{j}\sum_{k'=2}^{k}\sum_{l'=2}^{l}\theta_{1j'k'l'}$$

$$= \exp\left[H_0 + H_i^{(1)} + \dots + H_l^{(4)} + H_{ij}^{(12)} + \dots + H_{kl}^{(34)} + H_{ijk}^{(123)} + \dots + H_{jkl}^{(234)} + H_{ijkl}^{(1234)}\right]$$

$$\sum_{l'=2}^{l}\theta_{111l'}$$

$\equiv H^{(34)}$



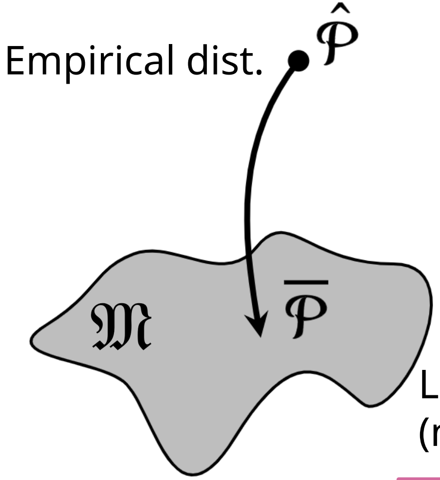
$\equiv H^{(234)}$

- This belongs to the energy-based model, or the exponential family

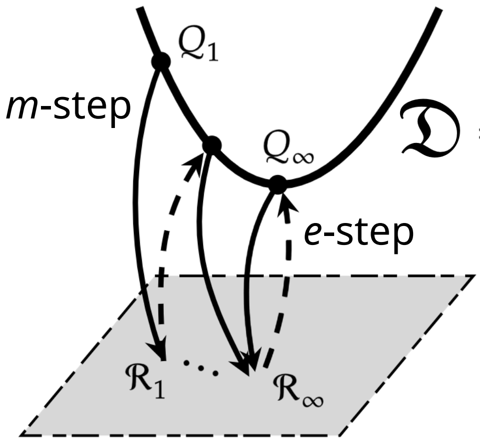
"Addition" with Latent Variables
[Ghalamkari et al. AAAI2025WS]

D-dim tensor space

(*D*+1)-dim tensor space



Equivalent



$\mathfrak{D} = \left\{ Q \mid \sum_r Q_{ijkr} = \hat{\mathcal{P}}_{ijk} \right\}$

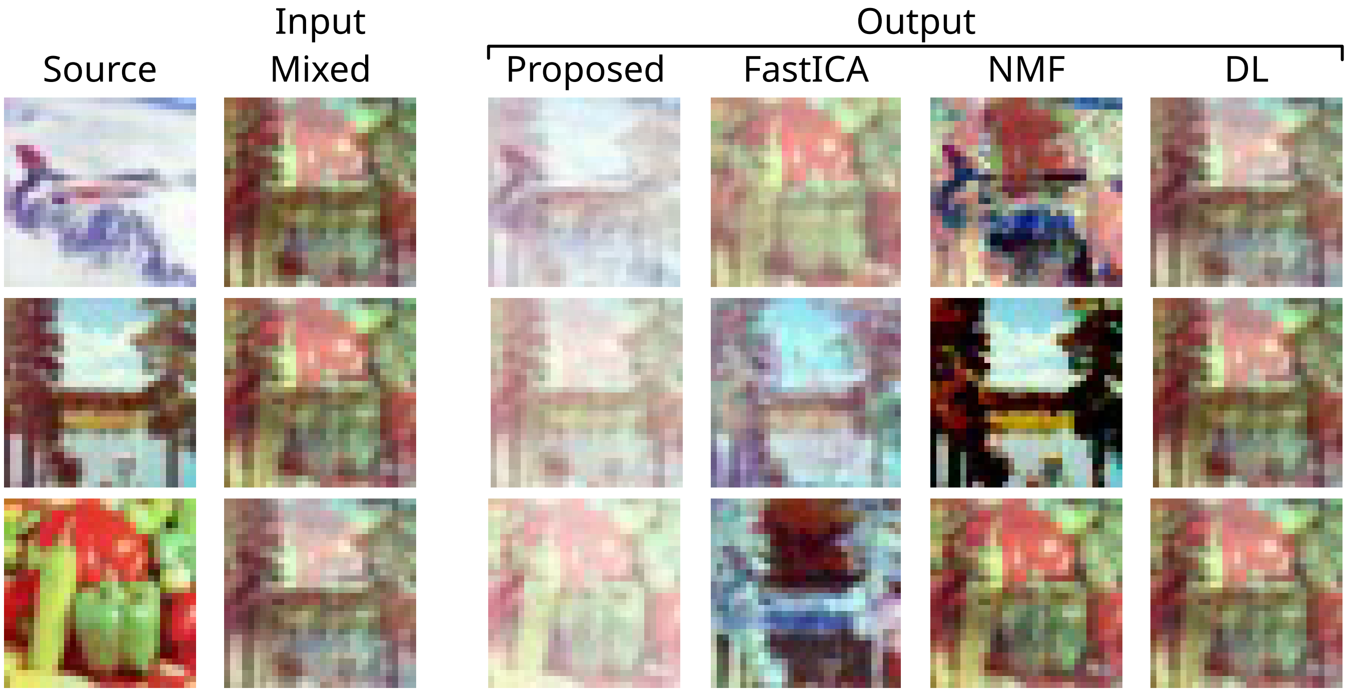
Low-rank tensors
(non-flat manifold)

Generalized
rank-1 tensors
(e-flat manifold)

$$\mathfrak{M} = \left\{ \mathcal{P} \mid \mathcal{P}_{ijk} = \sum_r A_{ir} B_{jr} C_{kr} \right\}$$

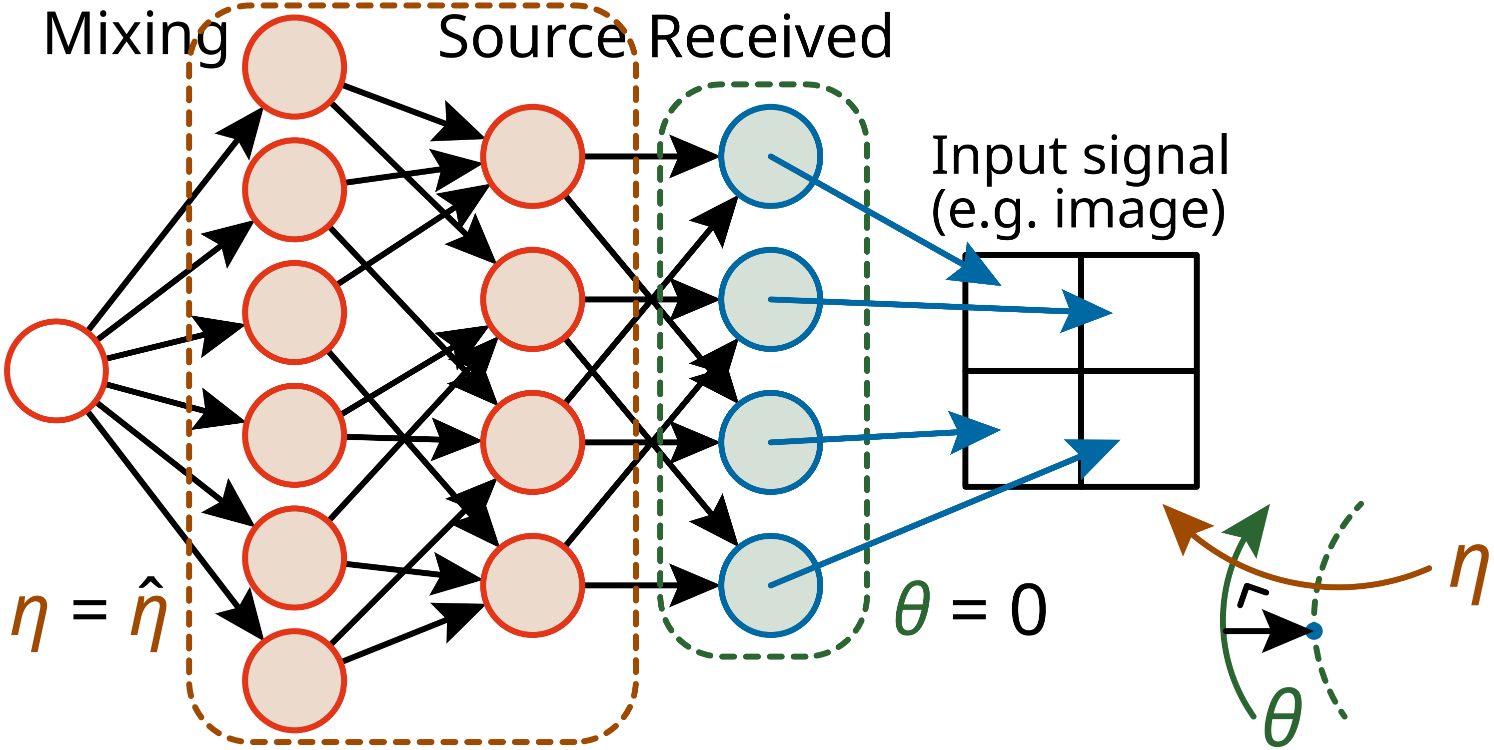
$$\mathfrak{B}^{\mathrm{CP}} = \{ \mathcal{R} \mid \mathcal{R}_{ijkr} = A_{ir} B_{jr} C_{kr} \}$$

Blind Source Separation [Luo et al. UAI2021]

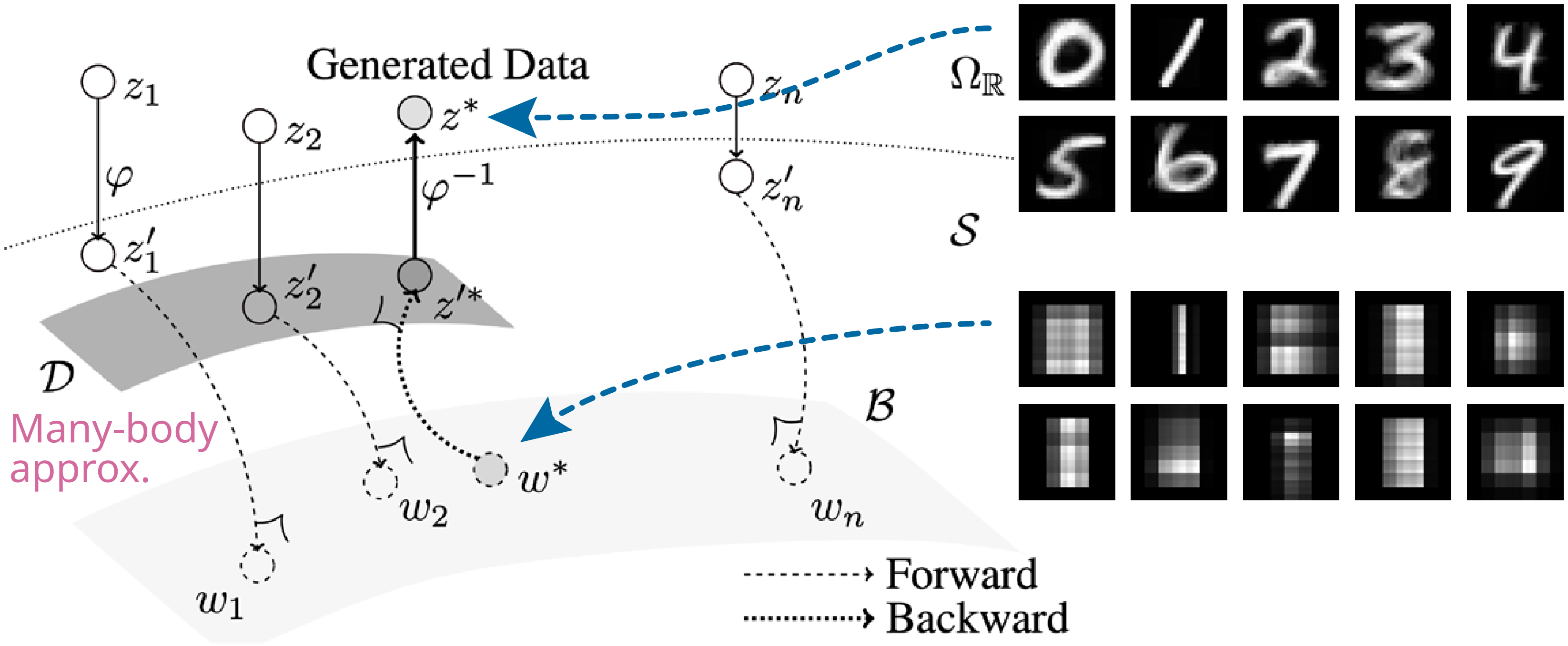


RMSE: **0.252±0.000** 0.260±0.111 0.362±0.030 0.612±0.000

Mixed Coordinates for BBS

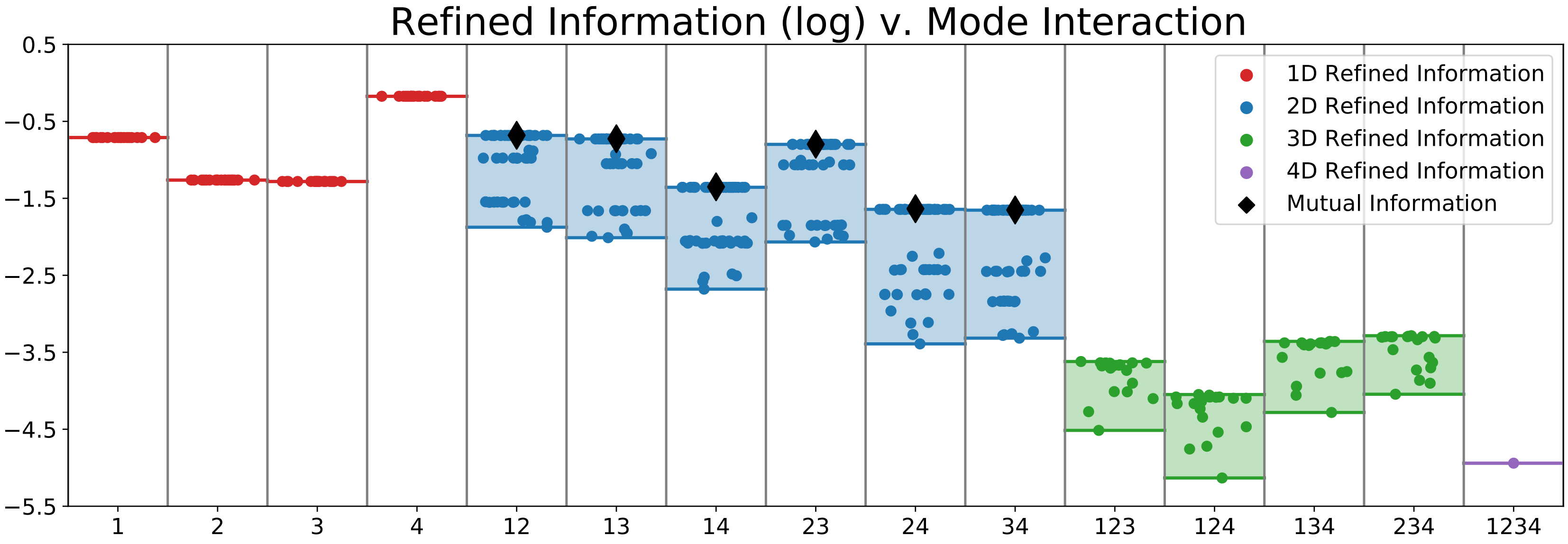


Data Argumentation [Hu & Sugiyama, arXiv:2410.00718]



Refined Information [Enouen&Sugiyama,arXiv:2410.11964]

- Refined information of mode (feature) interactions for tabular data



$$\mathrm{RI}_{\mathcal{J}\to\mathcal{I}}(p) = \mathrm{KL}(p; p_{\mathcal{J}}) - \mathrm{KL}(p; p_{\mathcal{I}})$$

Summary

- Probabilistic modeling with discrete structure
  - The log-linear model on posets
  - Fundamental concept from various fields: incidence algebra from order theory, frequency in pattern mining, parameters in Boltzmann machines, (θ, η) coordinates in information geometry, ...
- In tensor decomposition, we can treat each feature explicitly
  - e.g. decompose (A, B, C) into the product of (A, B), (B, C), and (C, A)
  - This has not been achieved by (low) rank-based approaches
  - Some relationship to tensor-and-circuites?
    - AAAI2025 WS: `https://april-tools.github.io/colorai/`
- **Slide:** `https://mahito.nii.ac.jp/pdf/FDIG2025.pdf`