

November 16, 2016
IBIS2016



Partial Order Structure and Information Geometry (順序構造と情報幾何)

Mahito Sugiyama (ISIR, Osaka University, PRESTO)
(杉山 磨人; 大阪大学産業科学研究所, JST さきがけ)

Today's Model on Poset (S, \leq)

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

Today's Model on Poset (S, \leq)

Probability

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

Zeta function

Möbius function

Expectation
(Frequency in pattern mining)
(Sufficient statistics in exponential family)

$$\begin{aligned}\log p(x) &= \sum_{s \in S} \zeta(s, x) \theta(s) \\ p(x) &= \sum_{s \in S} \mu(x, s) \eta(s)\end{aligned}$$

Outcome

- Given a poset (S, \leq) and consider distributions on S
 - The least element $\perp \in S$

1. KL divergence decomposition:

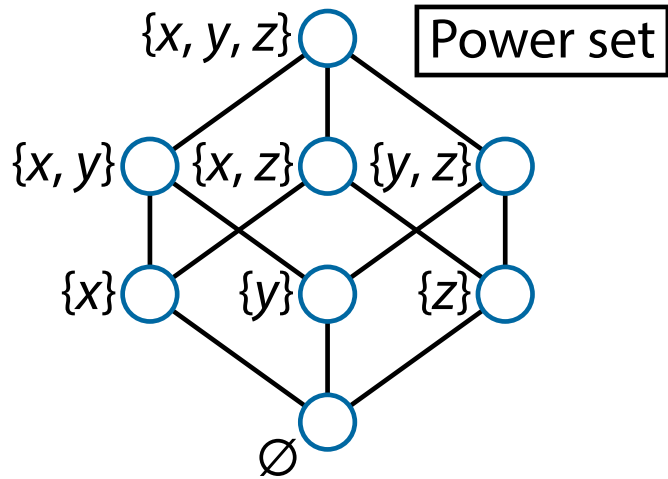
$$D_{\text{KL}}[P, R] = D_{\text{KL}}[P, Q] + D_{\text{KL}}[Q, R]$$

with Q s.t. $\theta_Q(x) = \theta_R(x)$ or $\eta_Q(x) = \eta_P(x)$ for all $x \in S \setminus \{\perp\}$

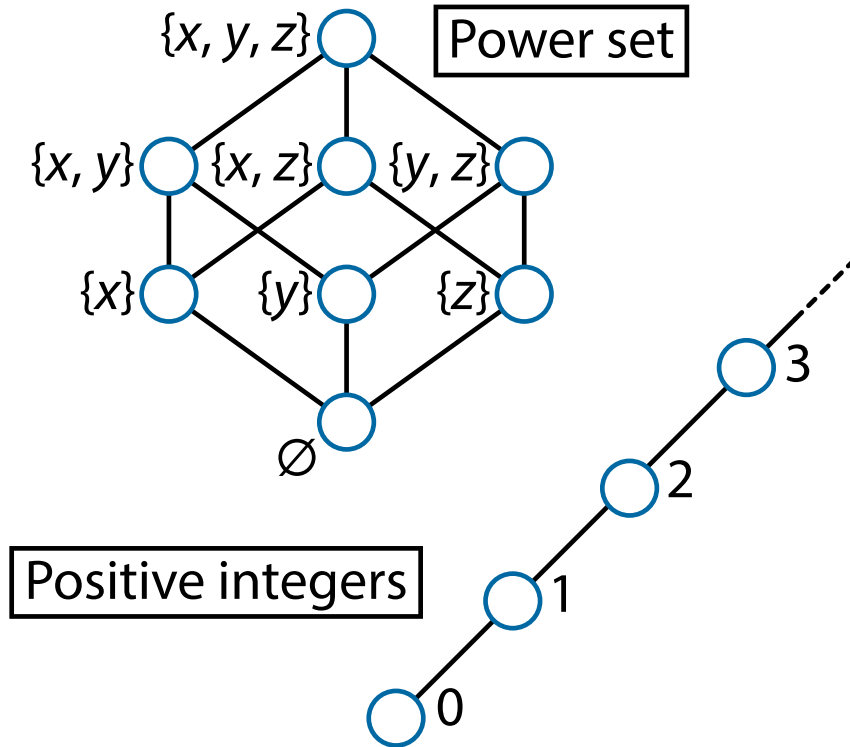
2. The set of probability distributions on (S, \leq) is a dually flat manifold w.r.t. θ and η

- p , θ , and η are coordinate systems
- θ and η are orthogonal
- θ introduces the structure of exponential family
- η introduces the structure of mixture family

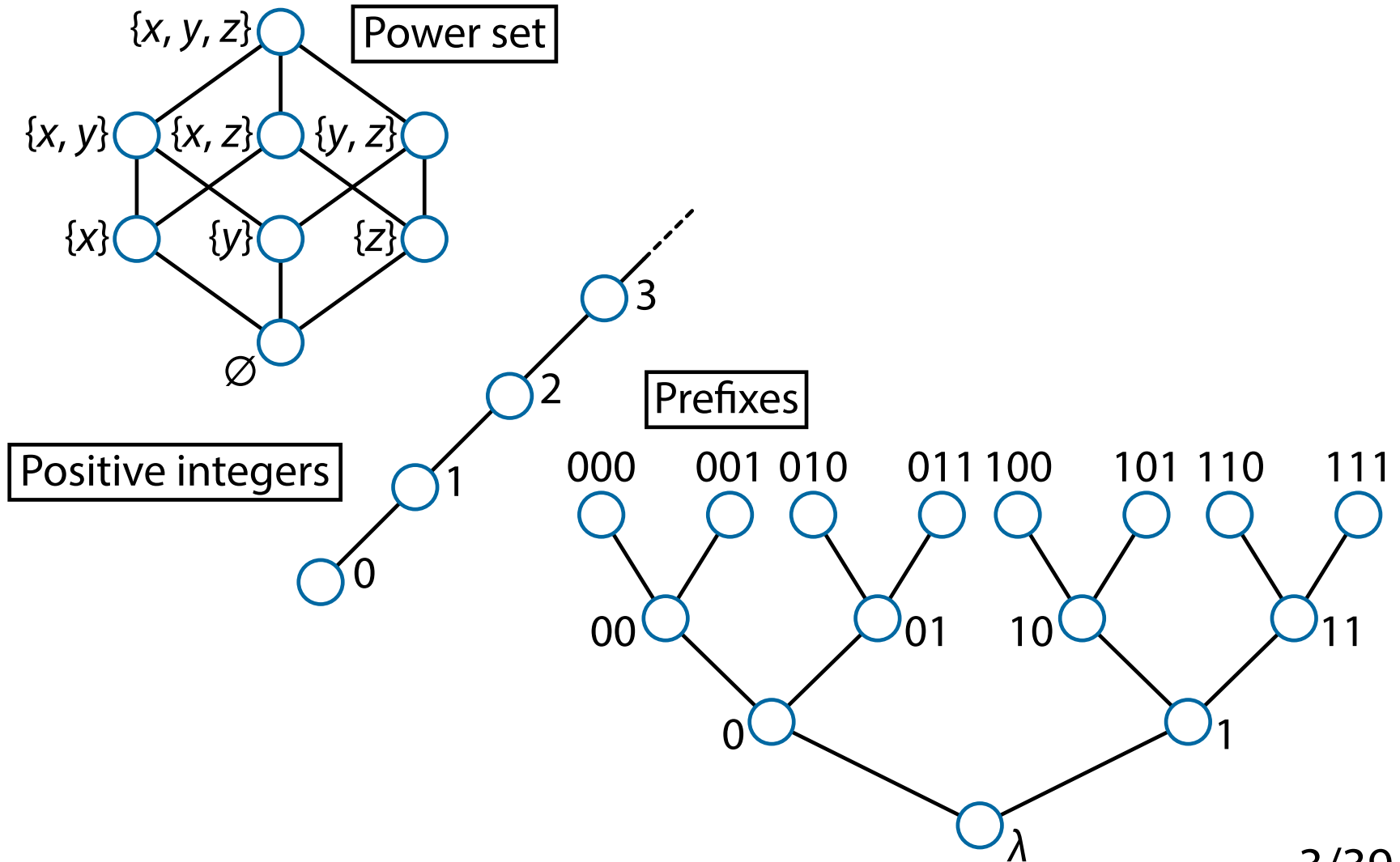
Partially Ordered Sets



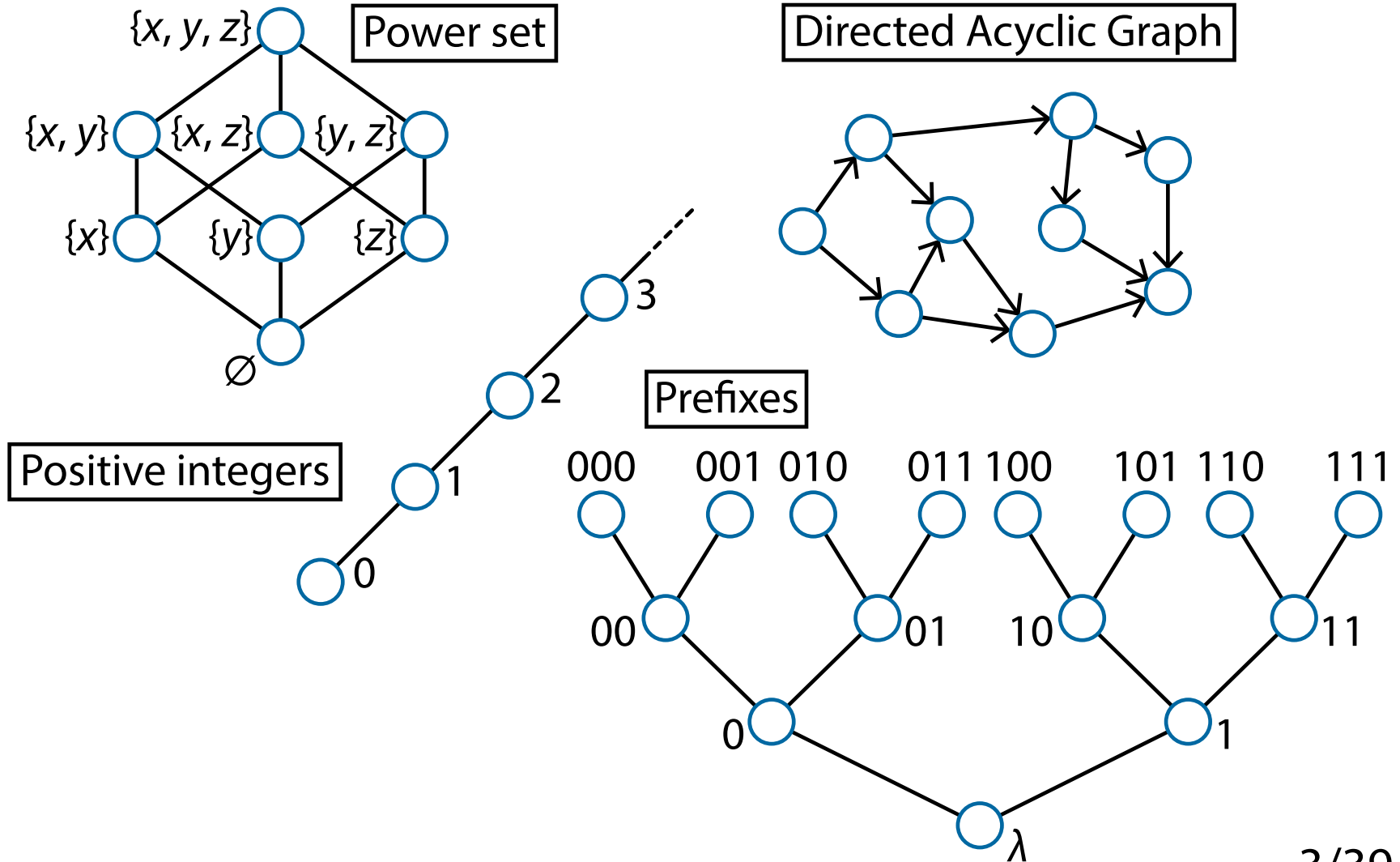
Partially Ordered Sets



Partially Ordered Sets

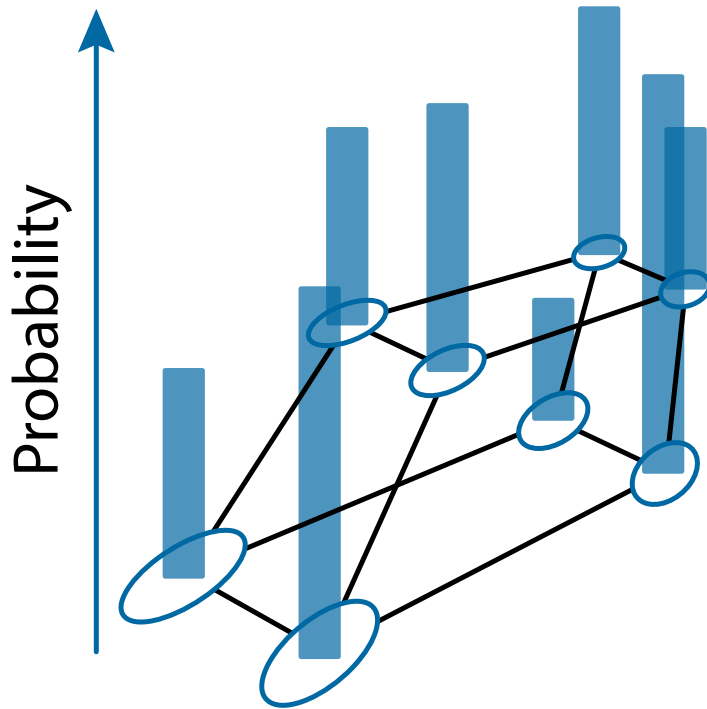


Partially Ordered Sets



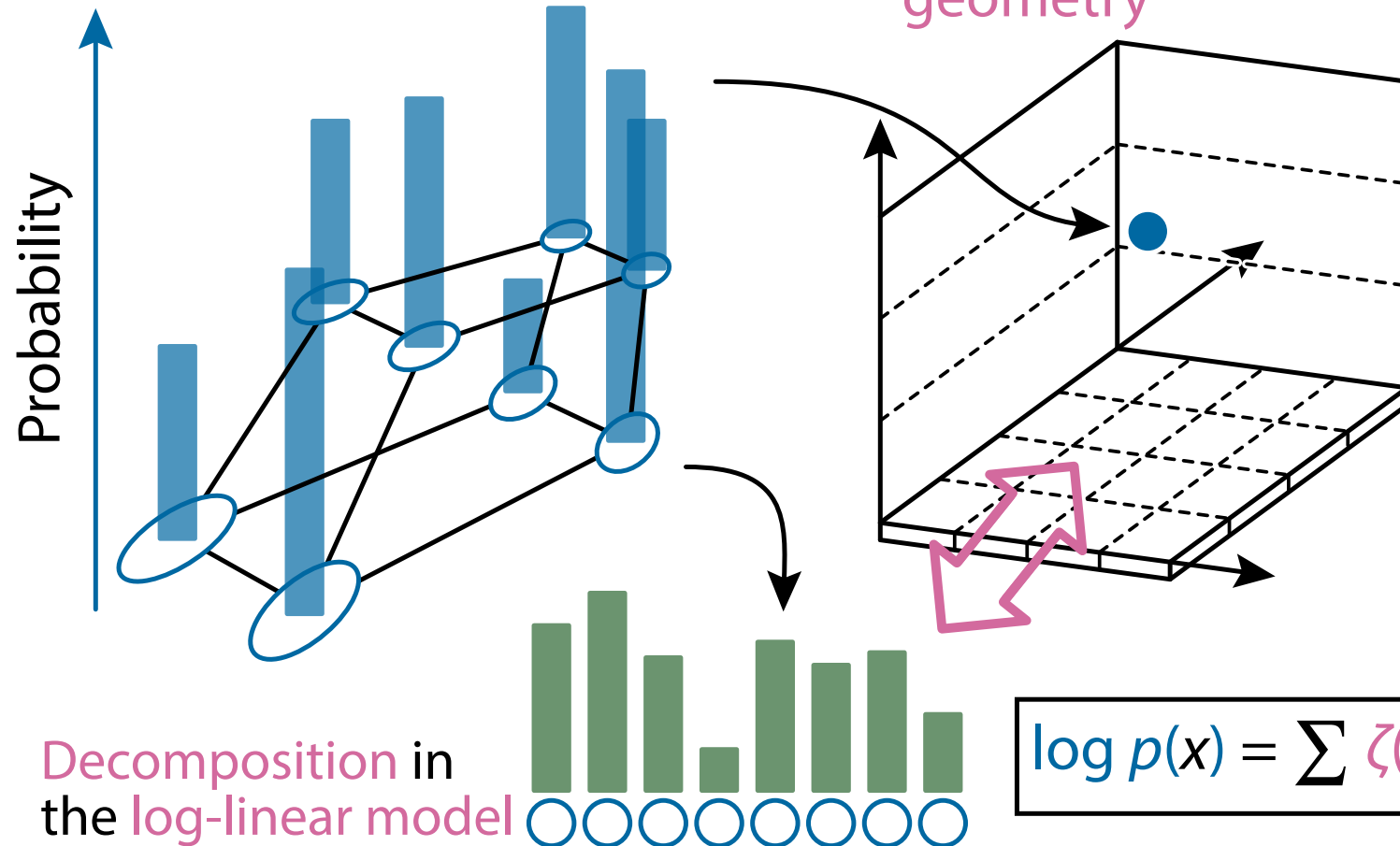
Posets with Probability Distribution

Probability distribution
on **posets** (partially ordered sets)



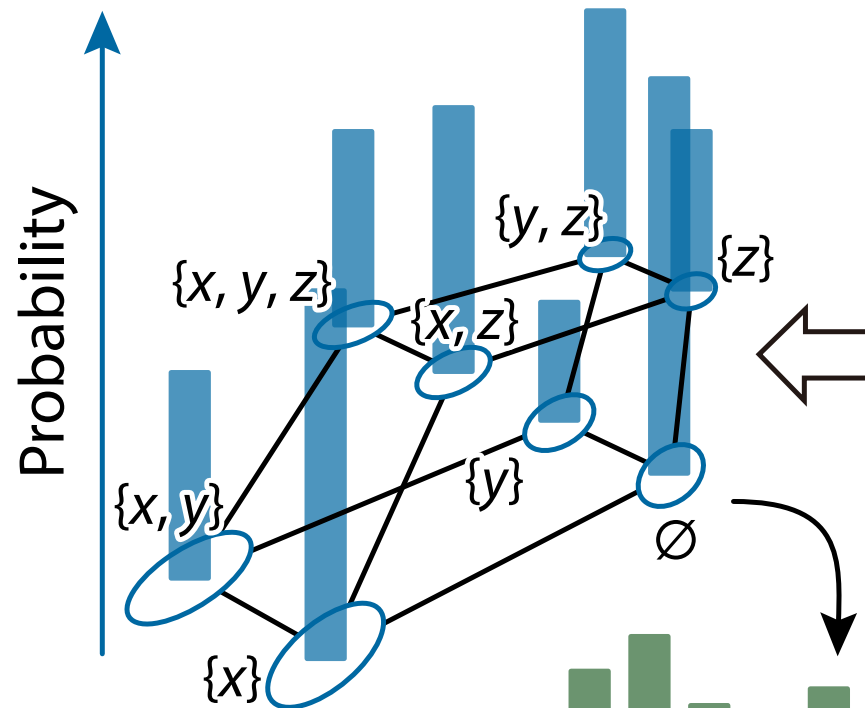
Posets with Probability Distribution

Probability distribution
on **posets** (partially ordered sets)



Posets with Probability Distribution

Probability distribution
on **posets** (partially ordered sets)

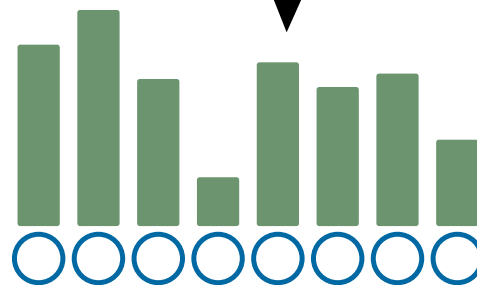


x y z (e.g. Neurons, SNPs, ...)

0	0	1	...
1	0	0	...
1	1	1	...
0	0	0	...
1	1	0	...
0	1	1	...
1	0	1	...
1	0	1	...
1	0	1	...
1	1	0	...

Numerical score
(KL divergence)
and the p -value
for higher-order
interactions

Decomposition in
the log-linear model



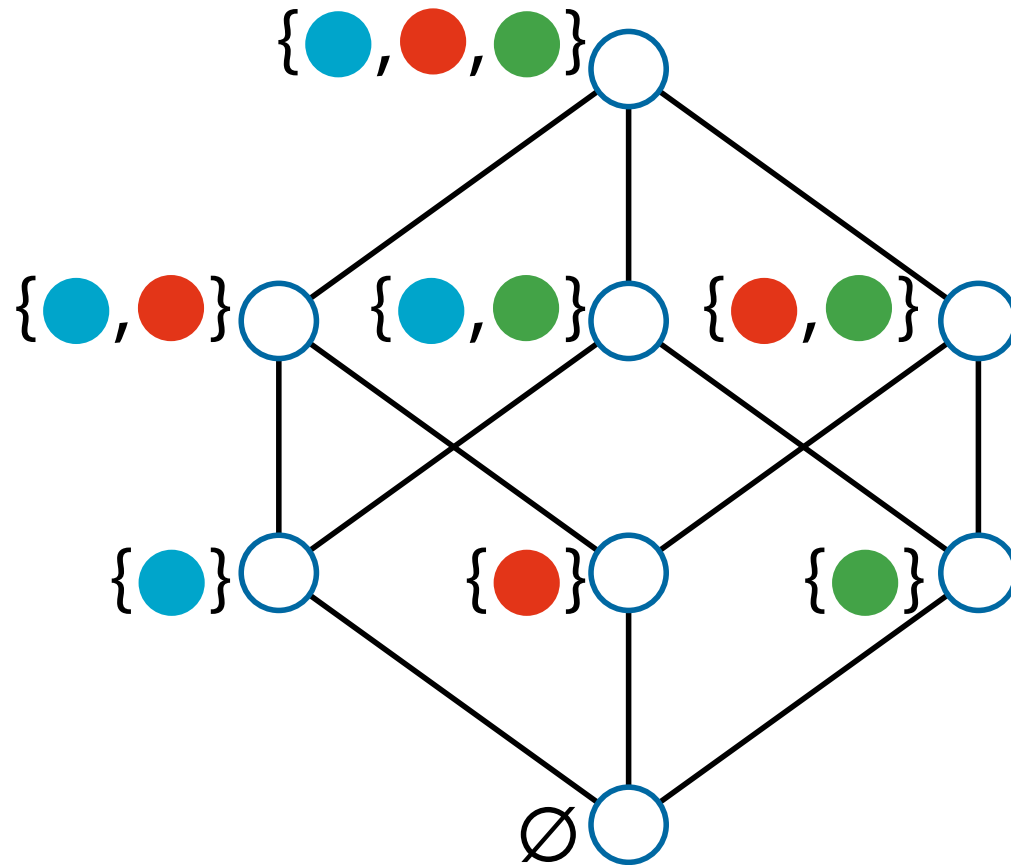
$$\log p(x) = \sum \zeta(s, x) \theta(s)$$

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

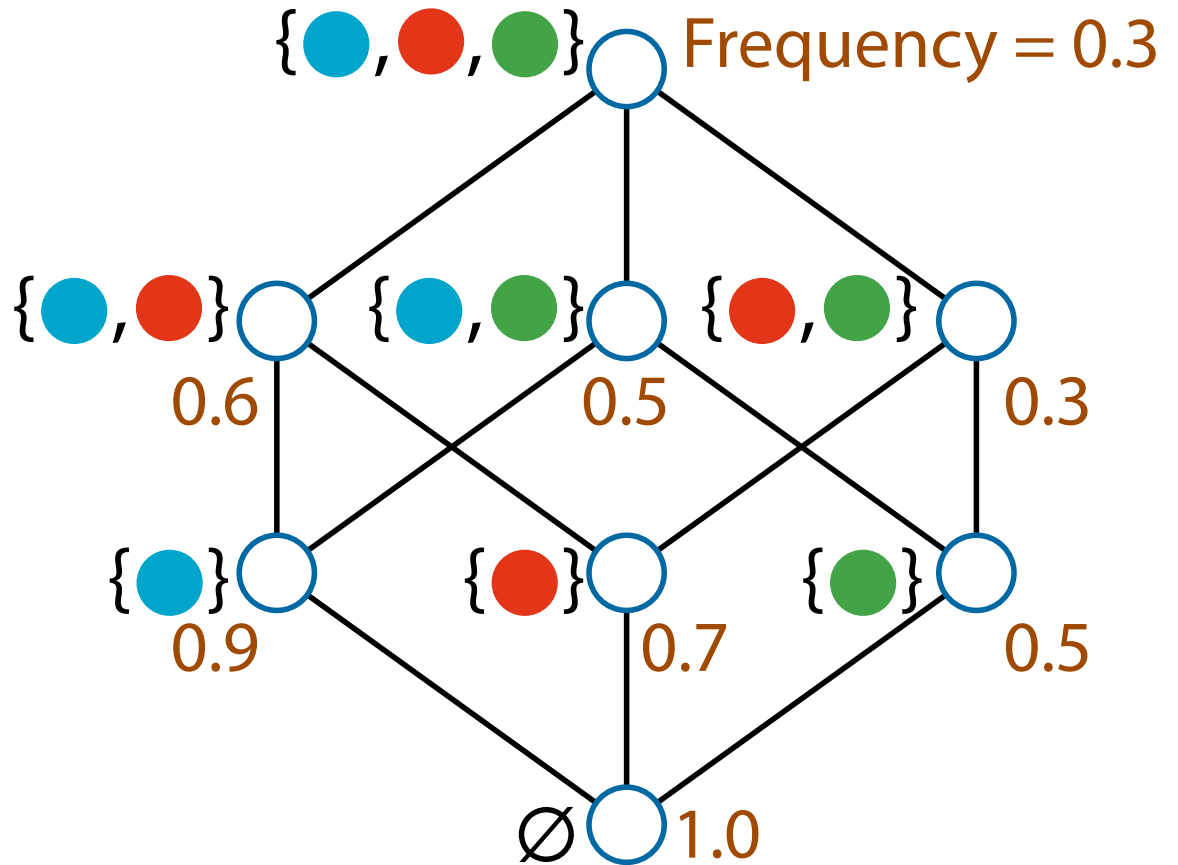
Poset (itemset lattice)






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

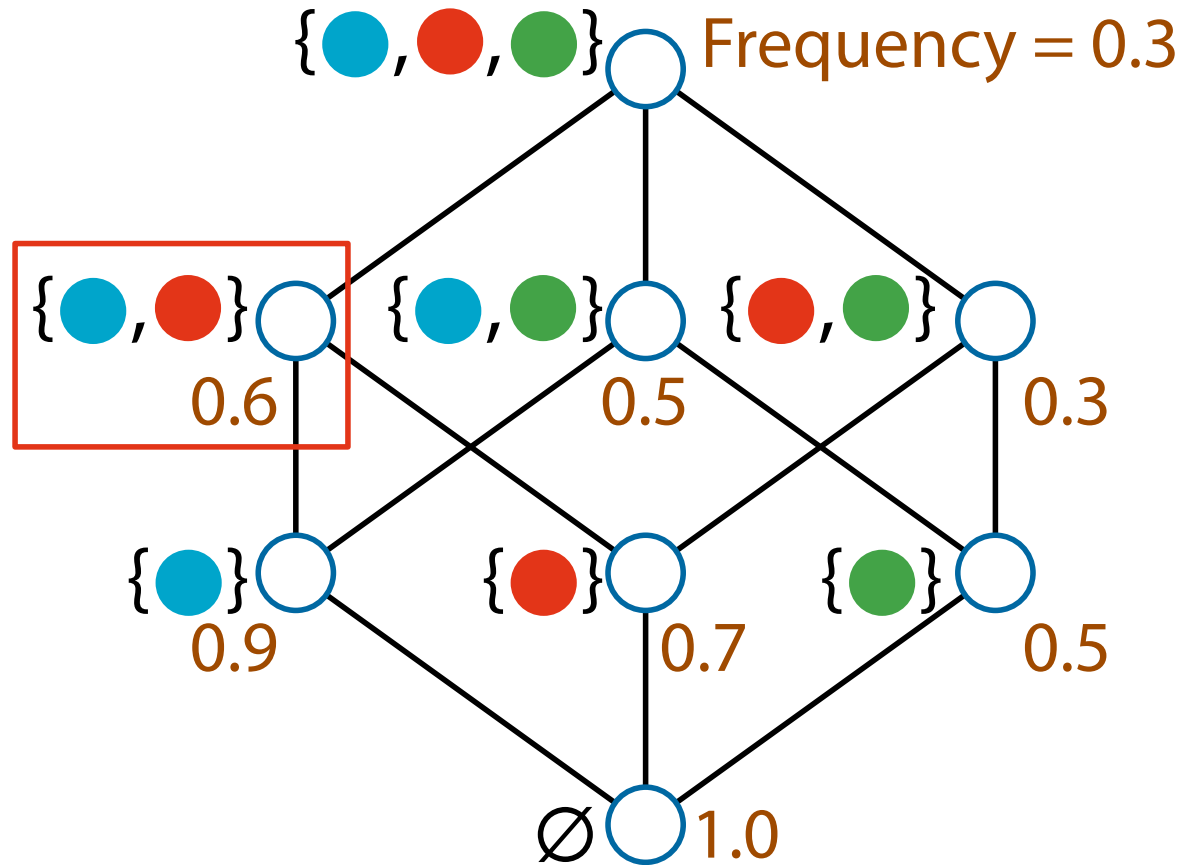
Poset (itemset lattice)



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)

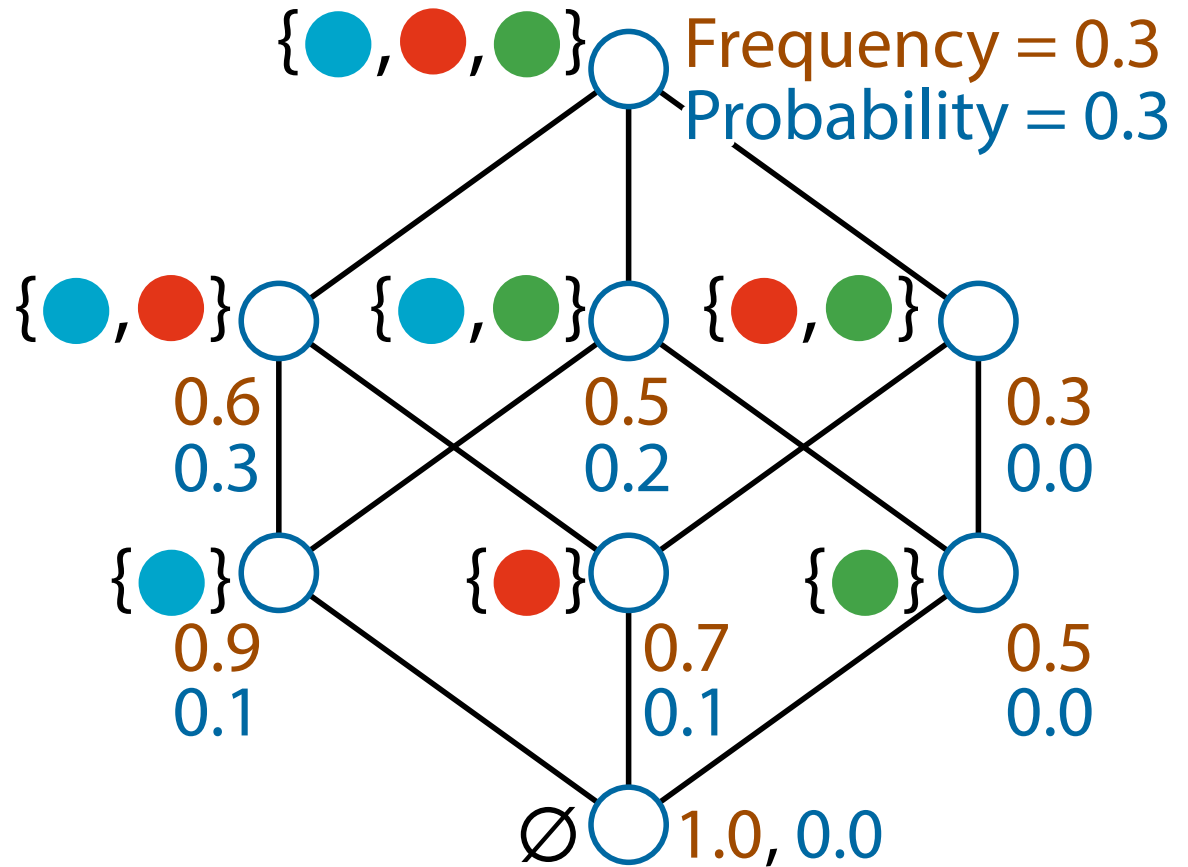


Binary vectors
(Transaction
database)



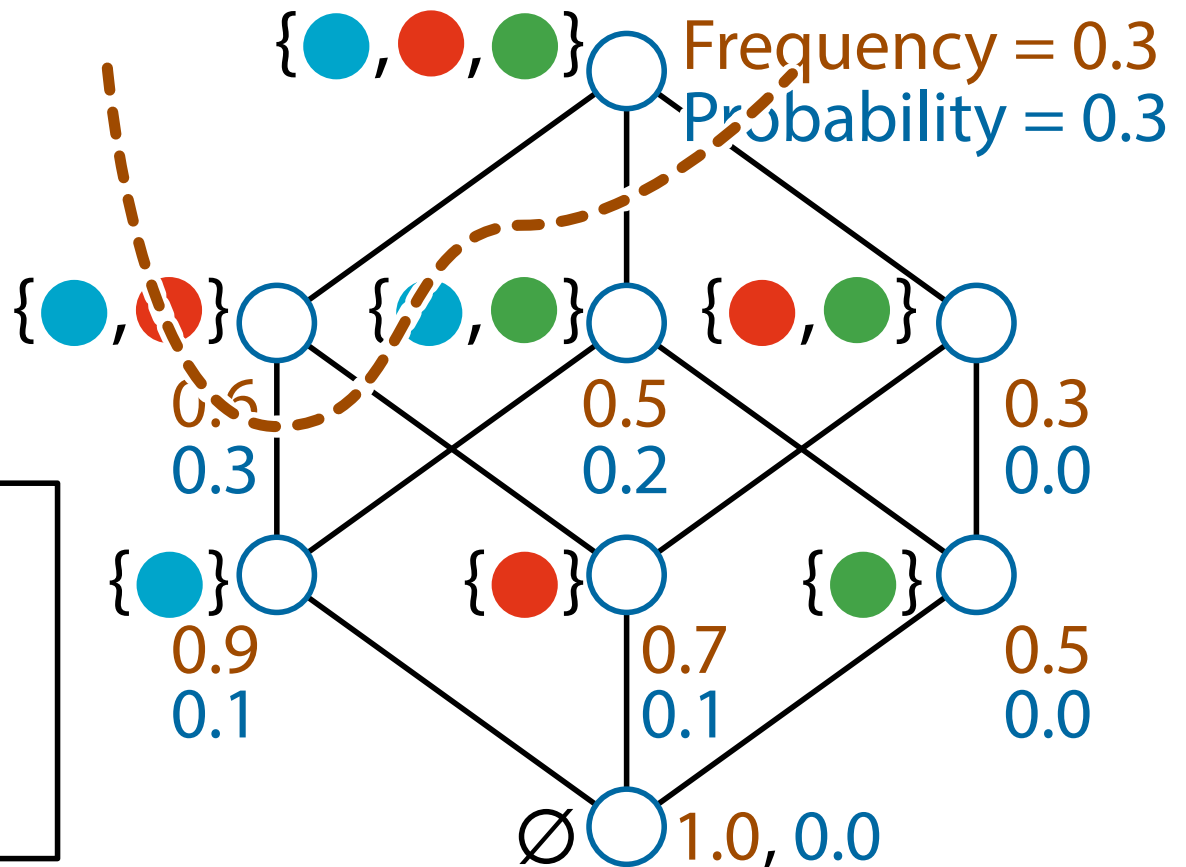
ID 1:	1	1	0
ID 2:	1	1	1
ID 3:	1	1	0
ID 4:	1	1	1
ID 5:	1	1	0
ID 6:	1	0	1
ID 7:	1	0	1
ID 8:	1	1	1
ID 9:	1	0	0
ID10:	0	1	0

Poset (itemset lattice)



Upward =
Pattern mining

Poset (itemset lattice)

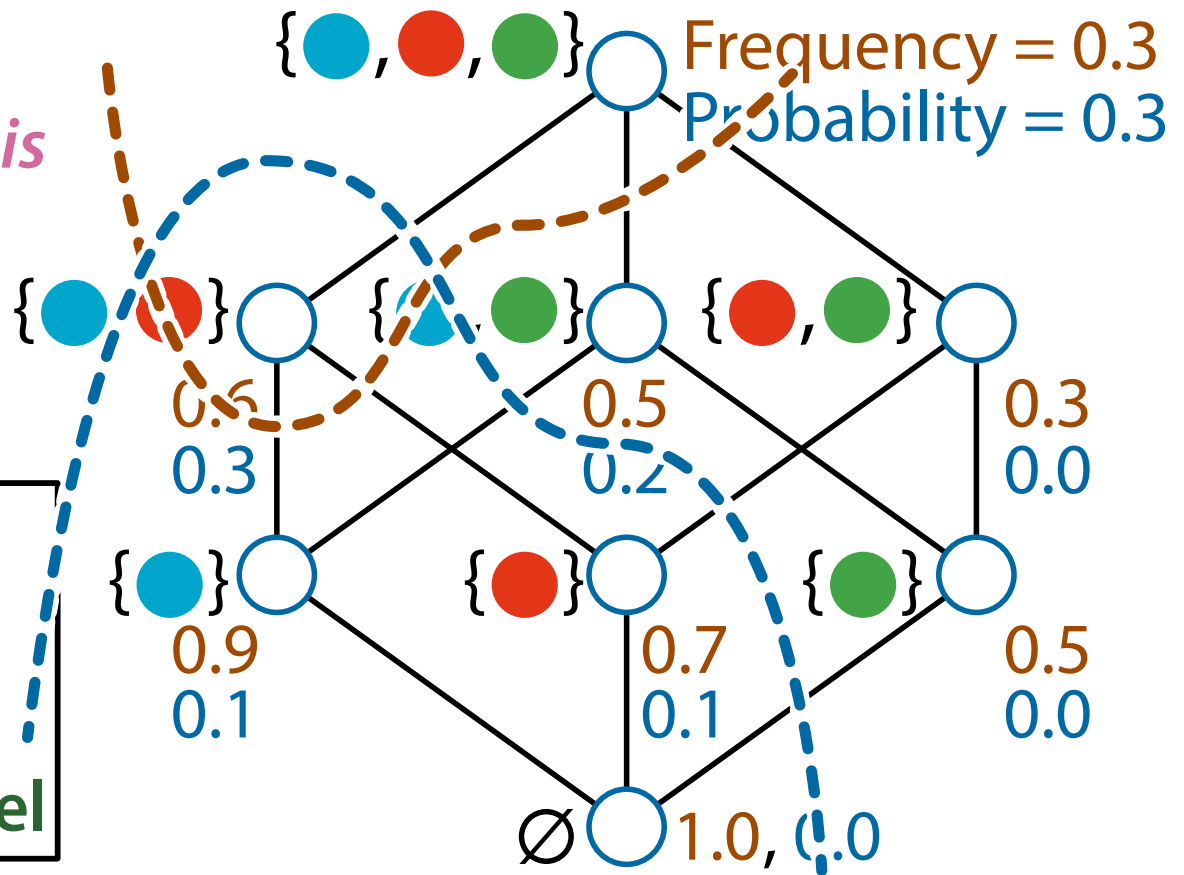


η : Frequency
 p : Probability

$$\eta(\{\text{blue}, \text{red}\}) = p(\{\text{blue}, \text{red}\}) + p(\{\text{blue}, \text{red}, \text{green}\})$$

Upward =
Pattern mining
Downward =
Log-linear analysis

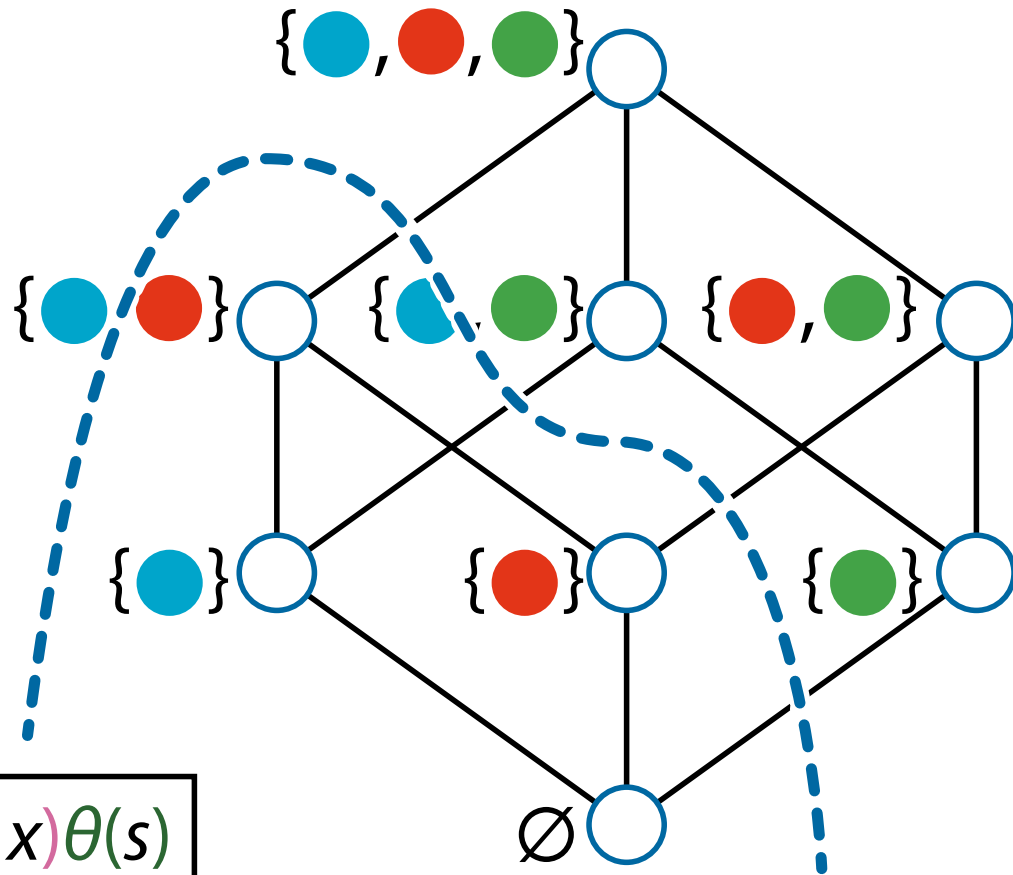
Poset (itemset lattice)



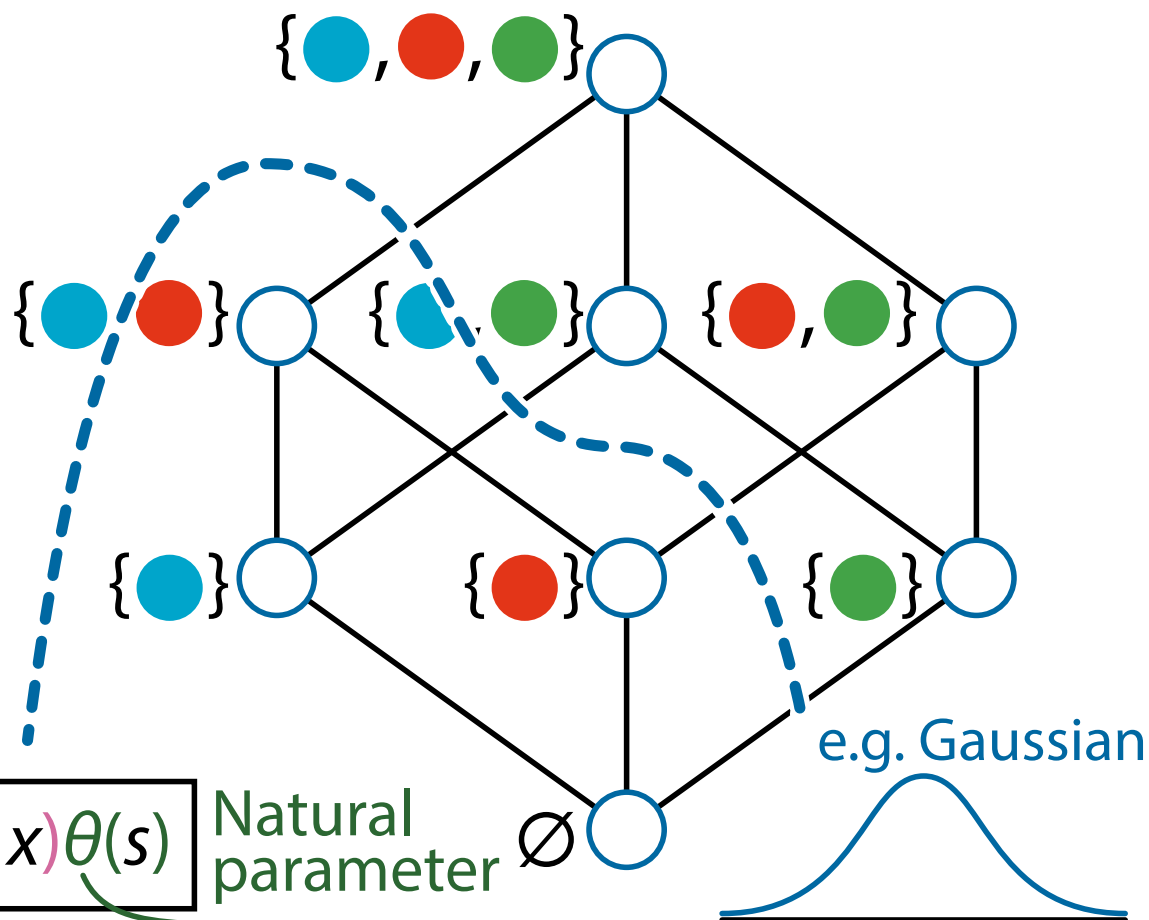
η : Frequency
 p : Probability
 θ : Coefficient of
log-linear model

$$\eta(\{\text{blue}, \text{red}\}) = p(\{\text{blue}, \text{red}\}) + p(\{\text{blue}, \text{red}, \text{green}\})$$

$$\log p(\{\text{blue}, \text{red}\}) = \theta(\{\text{blue}, \text{red}\}) + \theta(\{\text{blue}\}) + \theta(\{\text{red}\}) + \theta(\emptyset)$$



$$\log p(x) = \sum \zeta(s, x) \theta(s)$$



$$\log p(x) = \sum \zeta(s, x) \theta(s)$$

Natural parameter $\theta(s)$

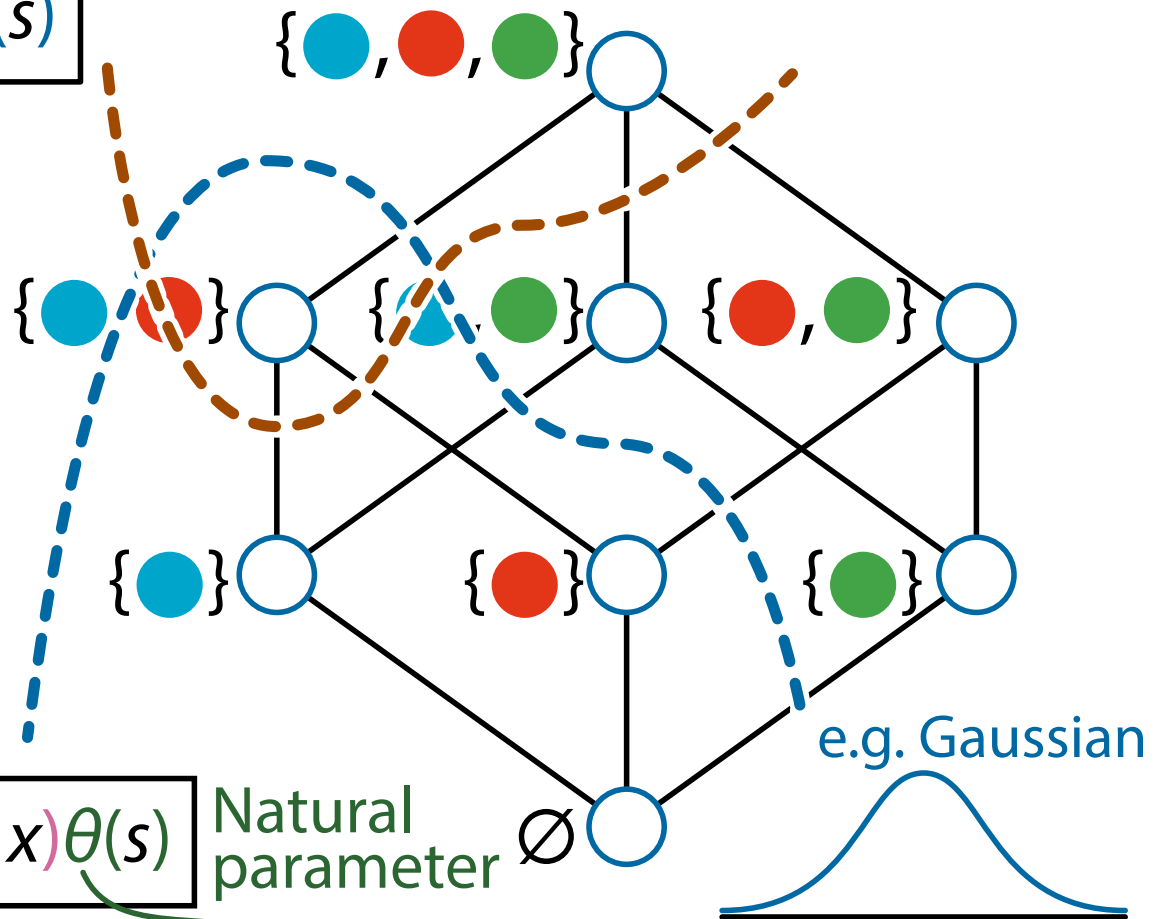
Exponential family:

$$p(x) = \exp\left(\sum \theta(s) F_s(x) - \psi(\theta)\right)$$

$$\eta(x) = \sum \zeta(x, s)p(s)$$

$$\eta(x) = \mathbb{E}[F_x(s)]$$

Sufficient
statistics of
exponential
family



$$\log p(x) = \sum \zeta(s, x)\theta(s)$$

Natural parameter

Exponential family:

$$p(x) = \exp\left(\sum \theta(s)F_s(x) - \psi(\theta)\right)$$

Möbius Inversion on Posets

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

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

- **Möbius function** $\mu: S \times S \rightarrow \mathbb{Z}$, defined as $\mu = \zeta^{-1}$:

$$\mu(x, y) = \begin{cases} 1, & \text{if } x = y, \\ -\sum_{x \leq s < y} \mu(x, s) & \text{if } x < y, \\ 0 & \text{otherwise} \end{cases}$$

- The **Möbius inversion formula** [Rota (1964)]:

$$g(x) = \sum_{s \in S} \zeta(s, x) f(s) \iff f(x) = \sum_{s \in S} \mu(s, x) g(s)$$

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, \dots, A_n ,

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

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

Mathematical Formulation

- Log-linear model and its sufficient statistics:

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

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

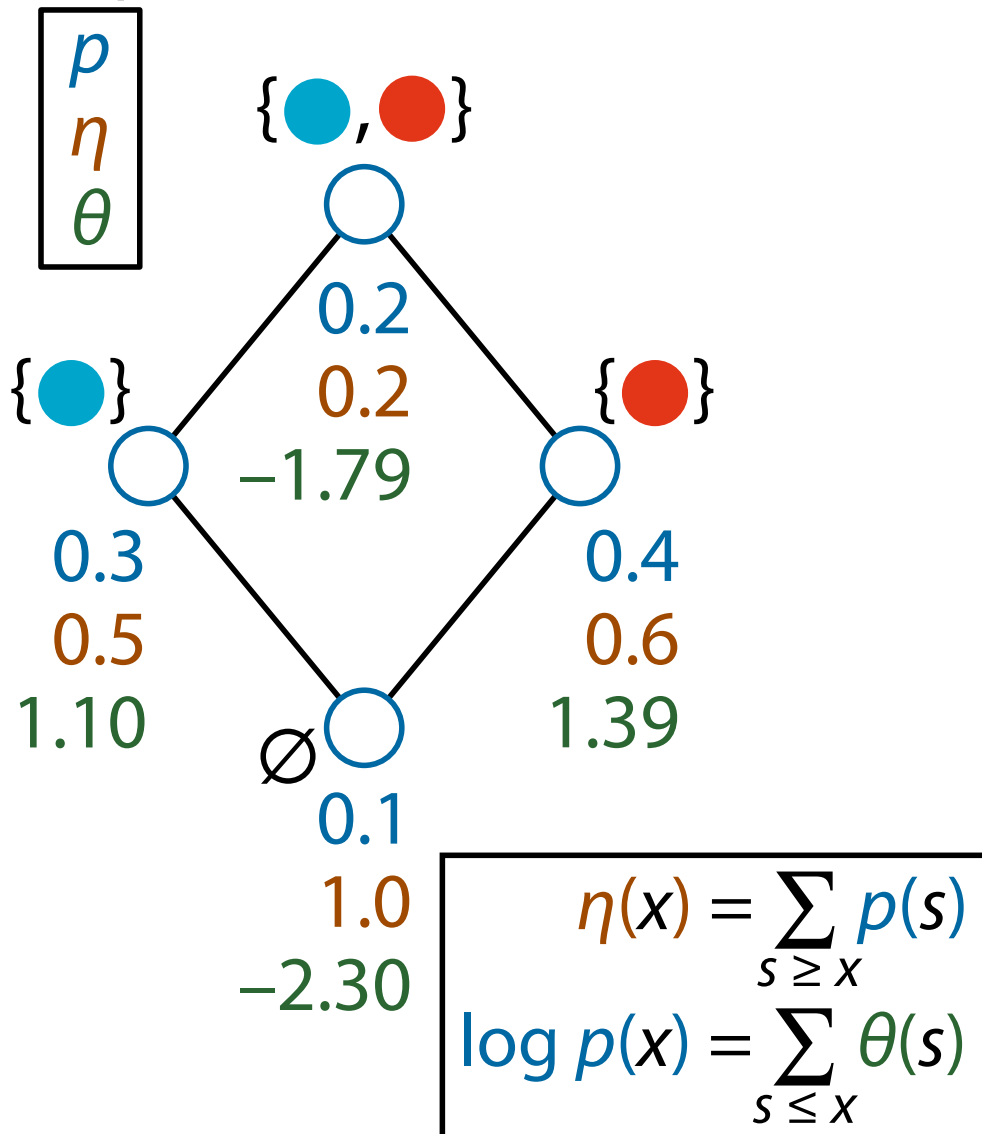
- Generalization of the log-linear model on binary vectors:

$$\log p(\mathbf{x}) = \sum_i \theta^i x^i + \sum_{i < j} \theta^{ij} x^i x^j + \dots + \theta^{1\dots n} x^1 x^2 \dots x^n,$$

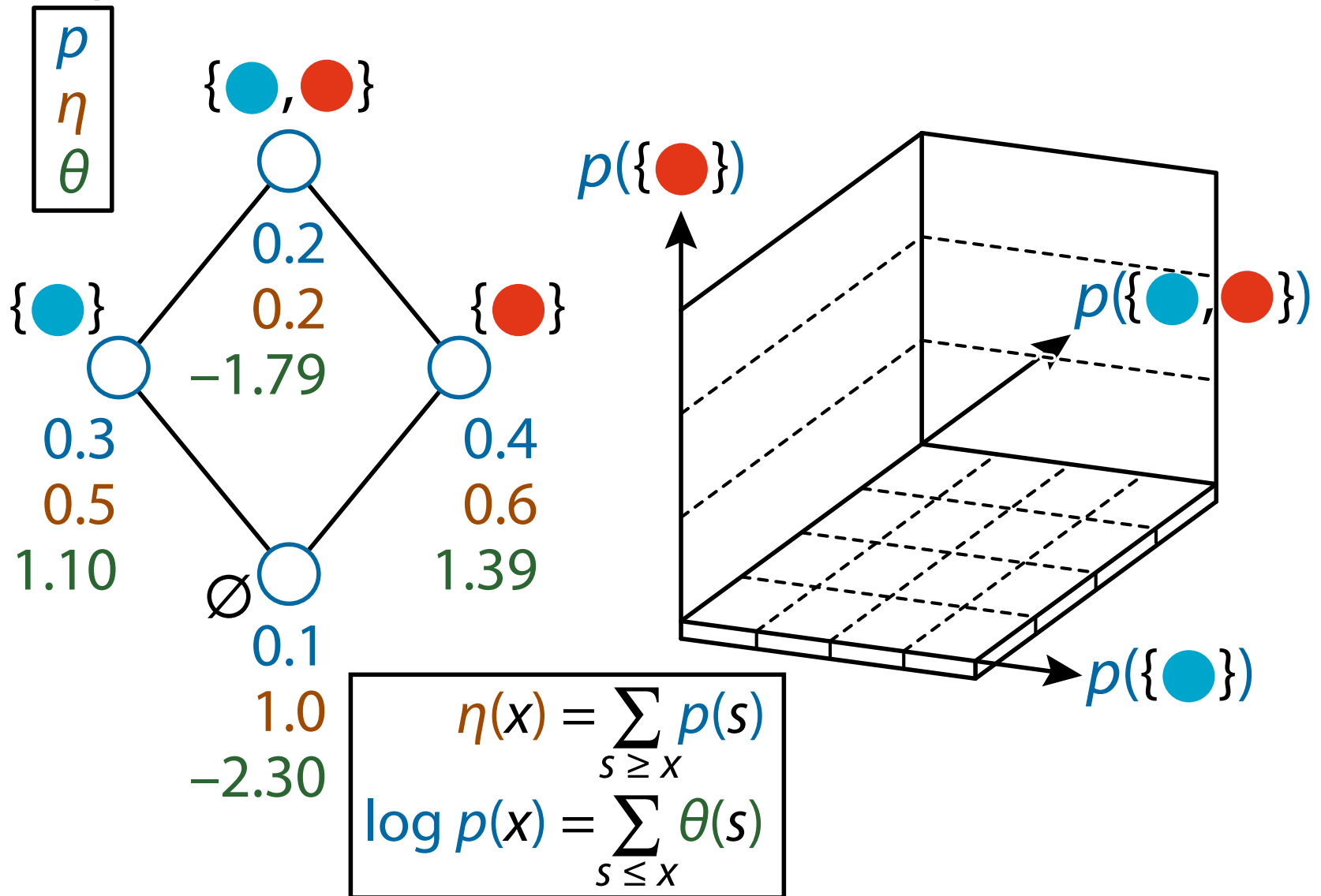
- From the Möbius inversion formula,

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

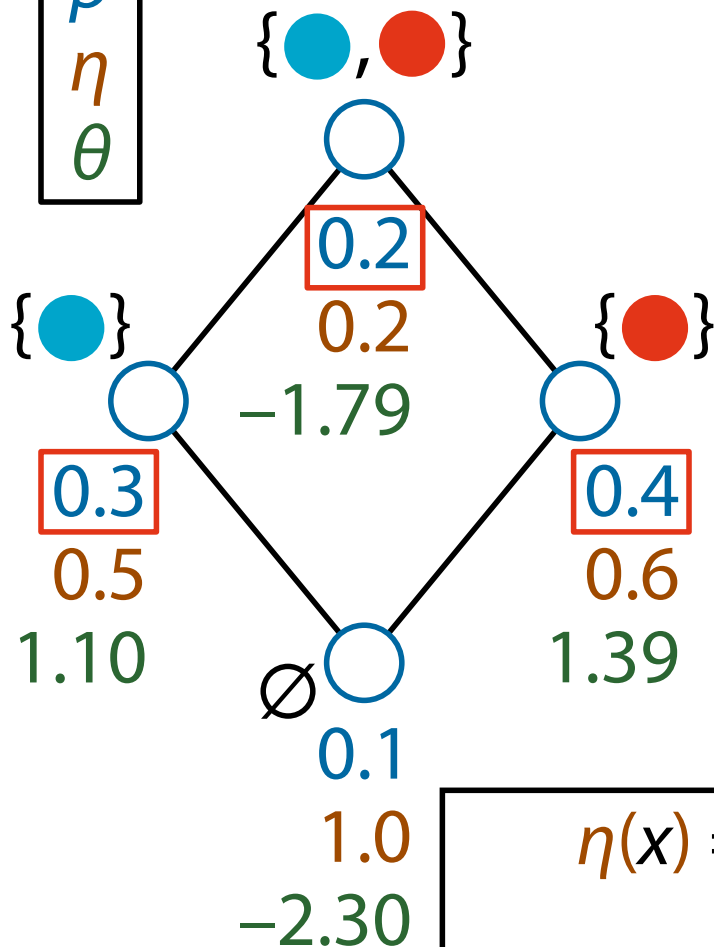
Triple for each node



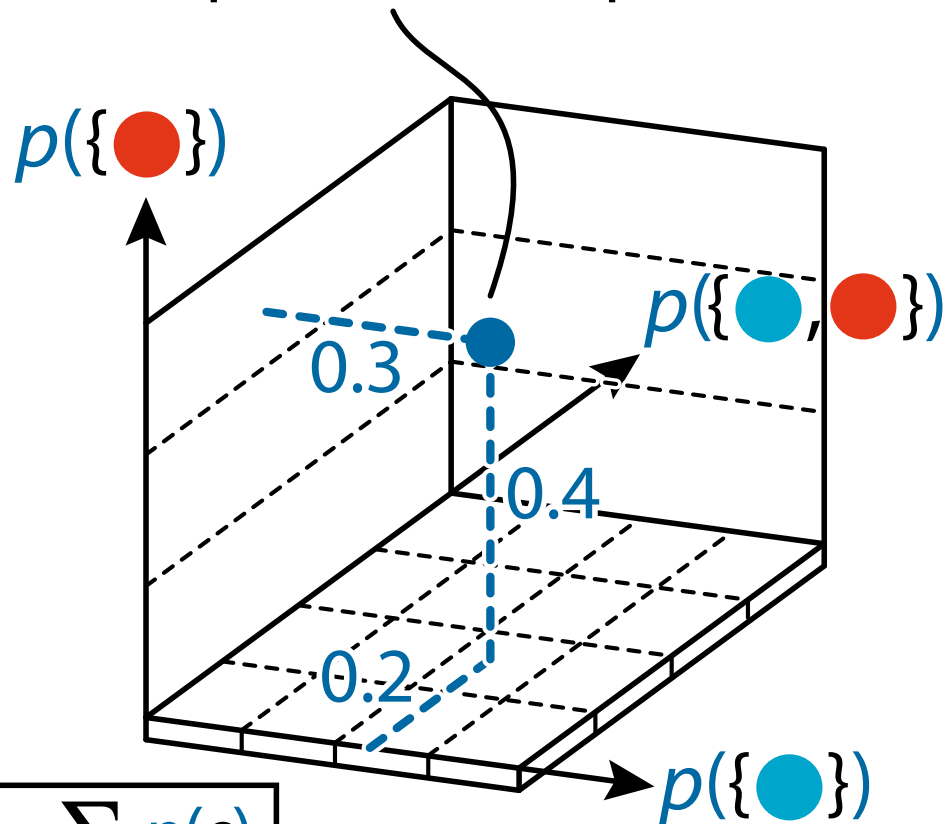
Triple for each node



Triple for each node



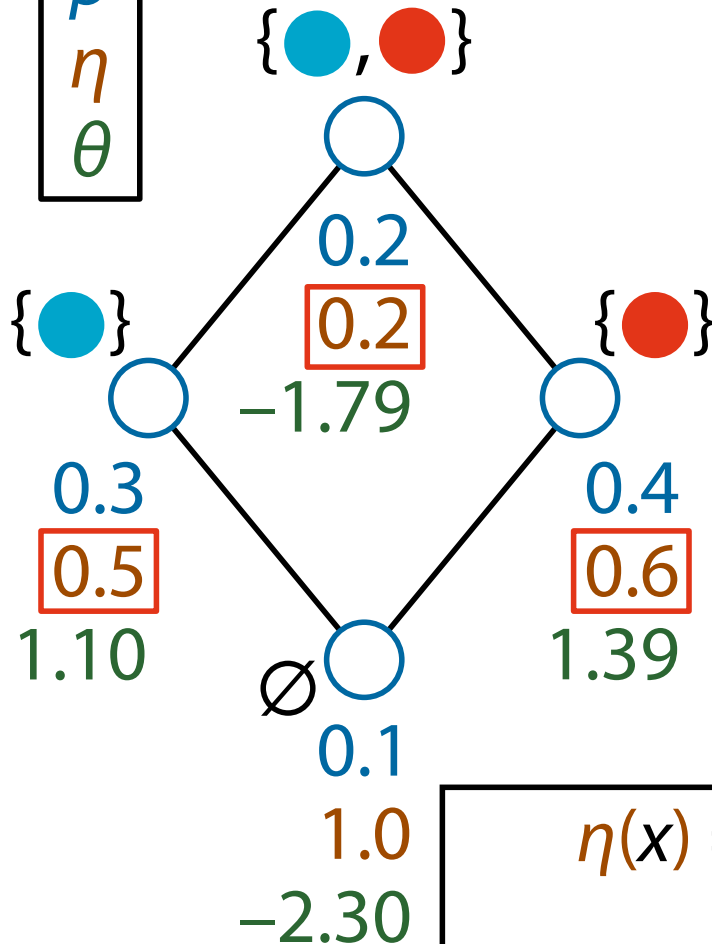
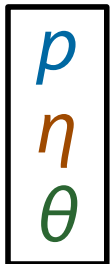
Probability distribution is a "point" in 3D space



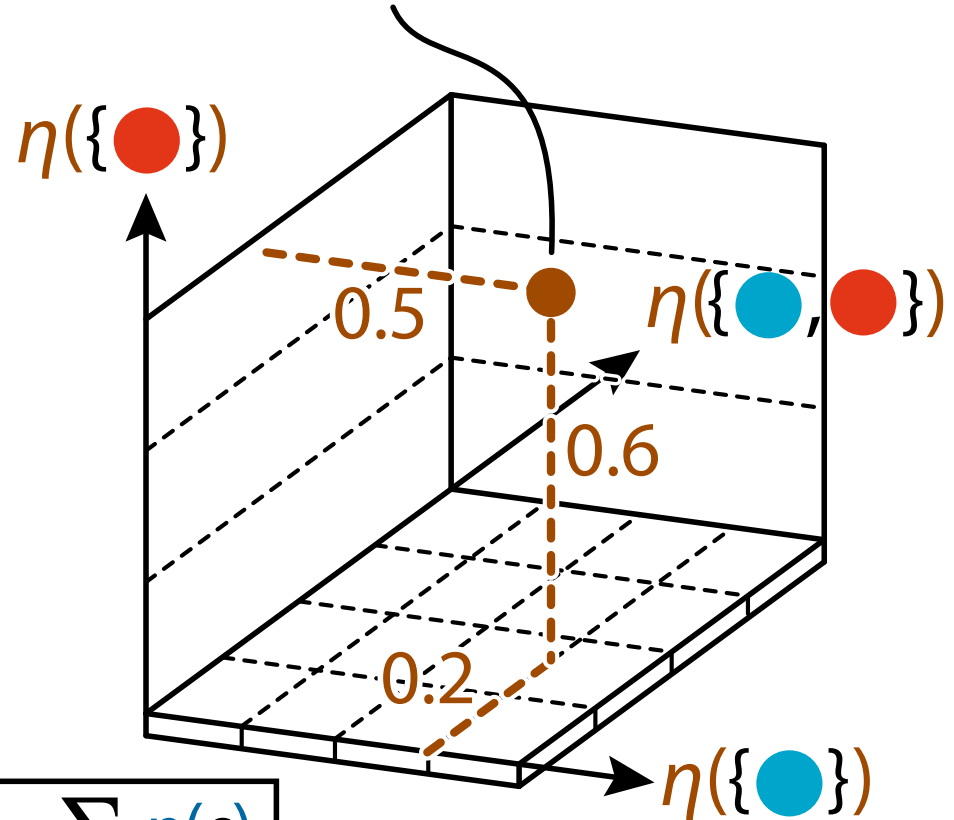
$$\eta(x) = \sum_{s \geq x} p(s)$$

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

Triple for each node



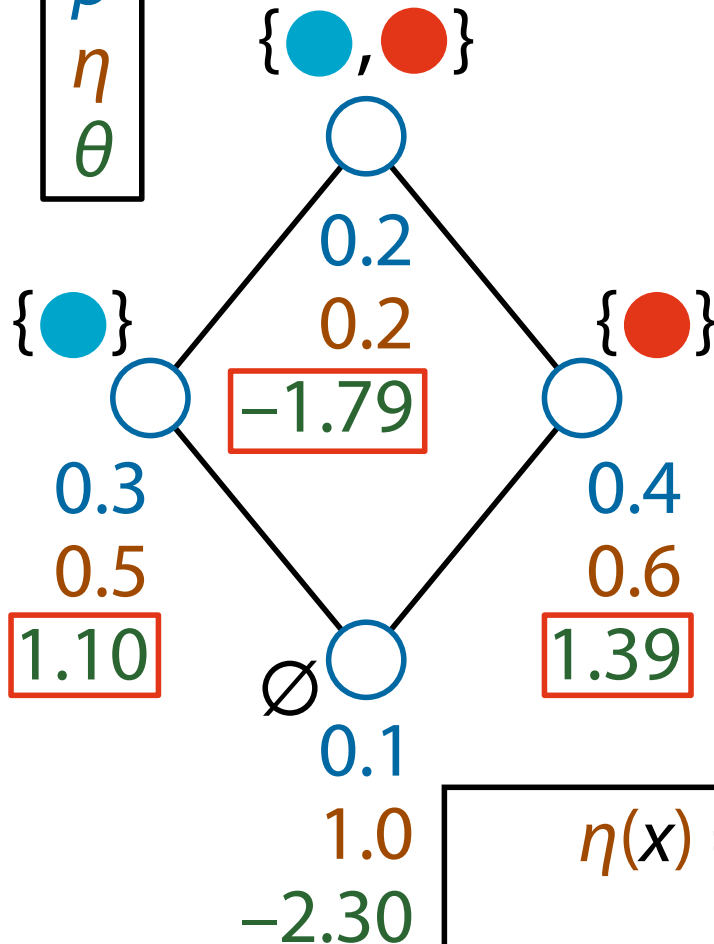
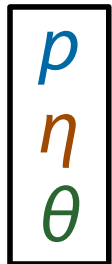
Probability distribution is a "point" in 3D space



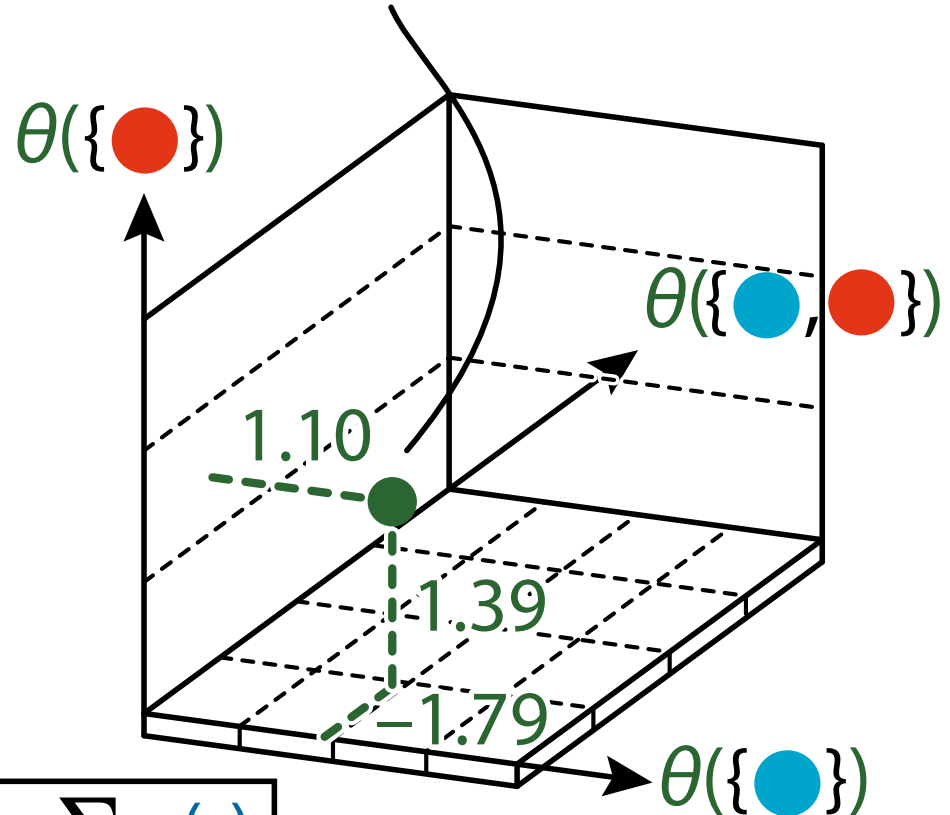
$$\eta(x) = \sum_{s \geq x} p(s)$$

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

Triple for each node



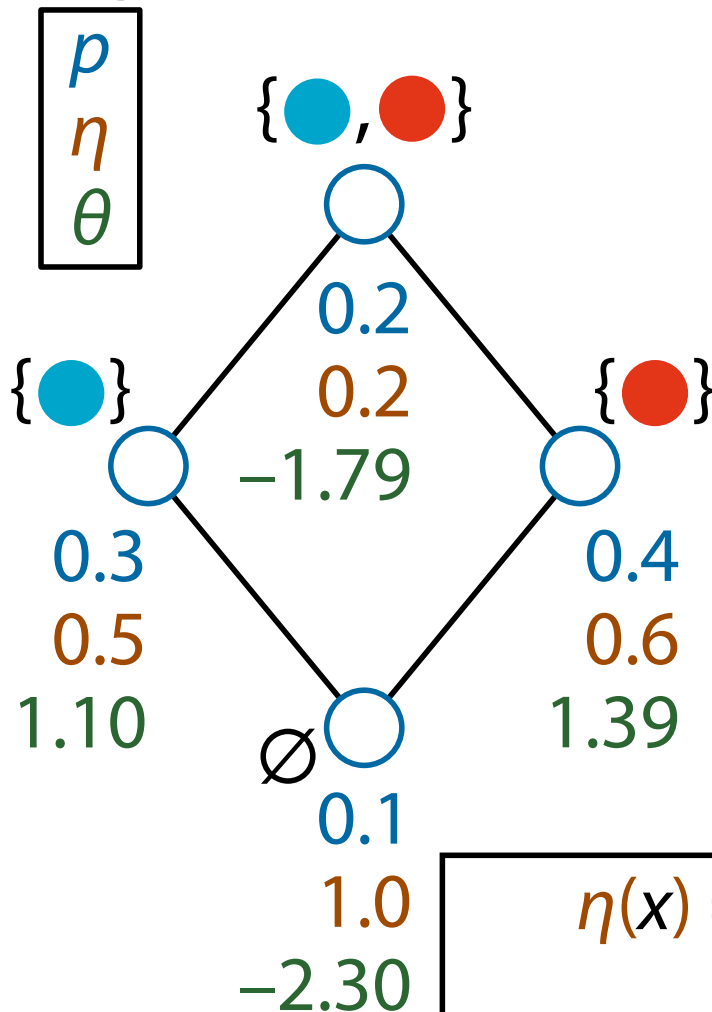
Probability distribution is a "point" in 3D space



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

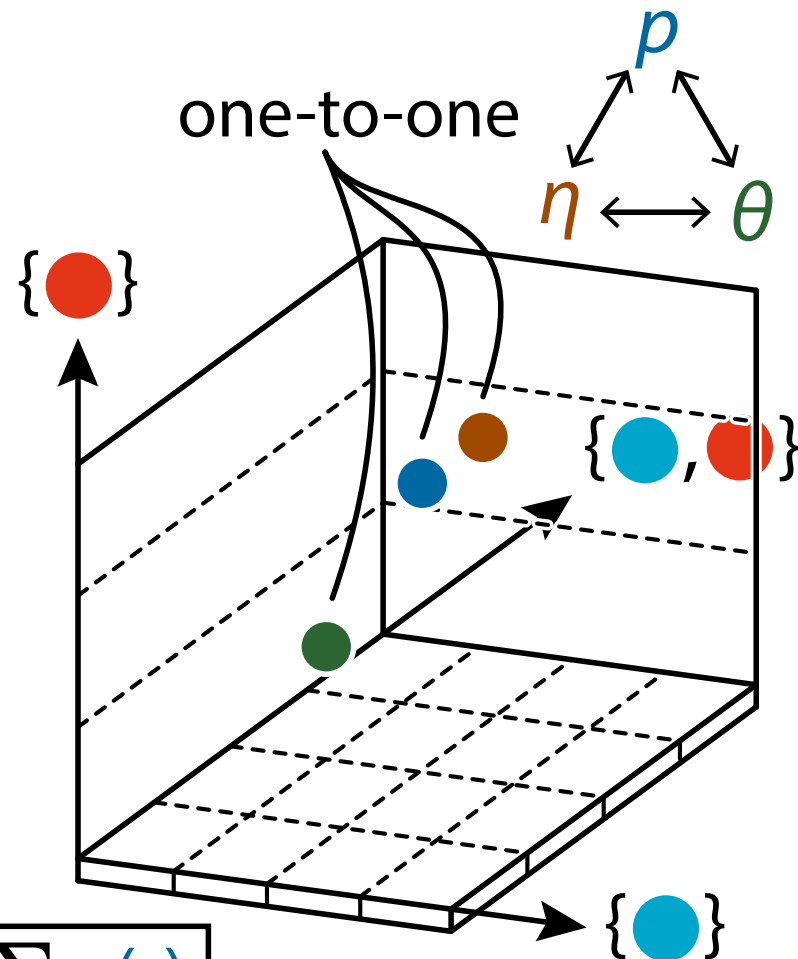
$$\log p(x) = \sum_{s \leq x} \theta(s)$$

Triple for each node

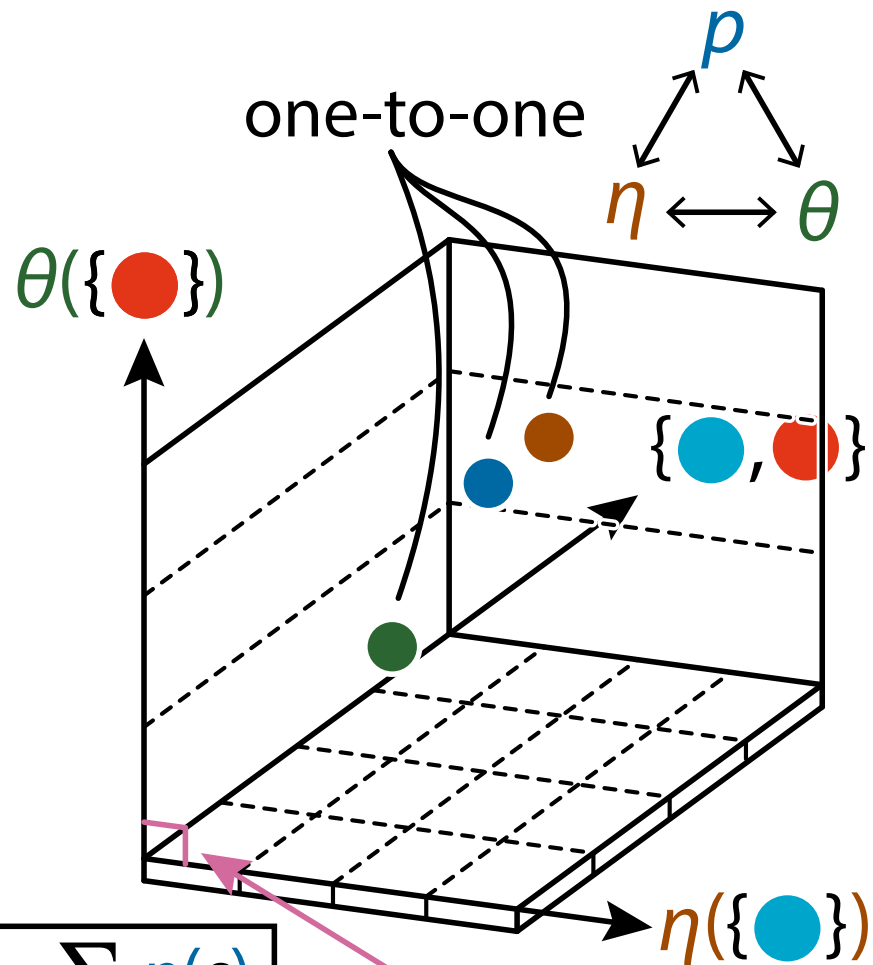
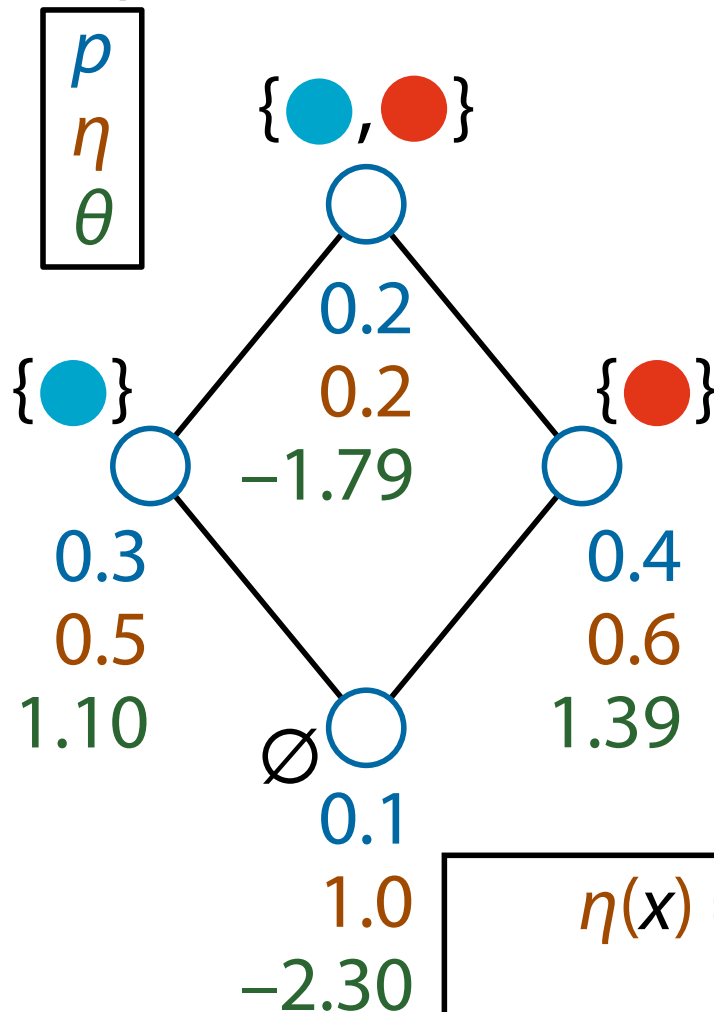


$$\eta(x) = \sum_{s \geq x} p(s)$$

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



Triple for each node



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

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

θ and η are dually orthogonal

Orthogonality of θ and η

- From Möbius inversion,

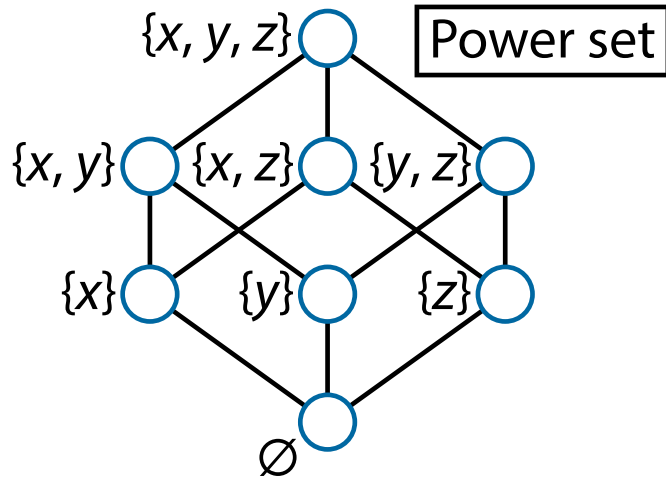
$$\sum_{s \in S} \zeta(x, s) \mu(s, y) = \delta_{x, y}, \quad \delta_{x, y} = \begin{cases} 1 & \text{if } x = y, \\ 0 & \text{otherwise} \end{cases}$$

- θ and η are dually orthogonal:

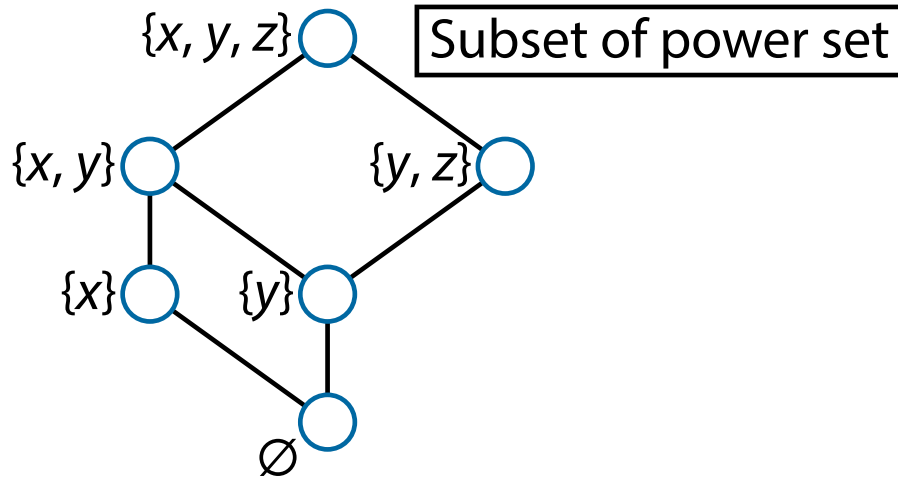
$$\mathbb{E} \left[\frac{\partial}{\partial \theta(x)} \log p(s) \frac{\partial}{\partial \eta(y)} \log p(s) \right] = \sum_{s \in S} \zeta(x, s) \mu(s, y) = \delta_{x, y}$$

- Partial order structure leads to the same dually flat structure with the exponential family

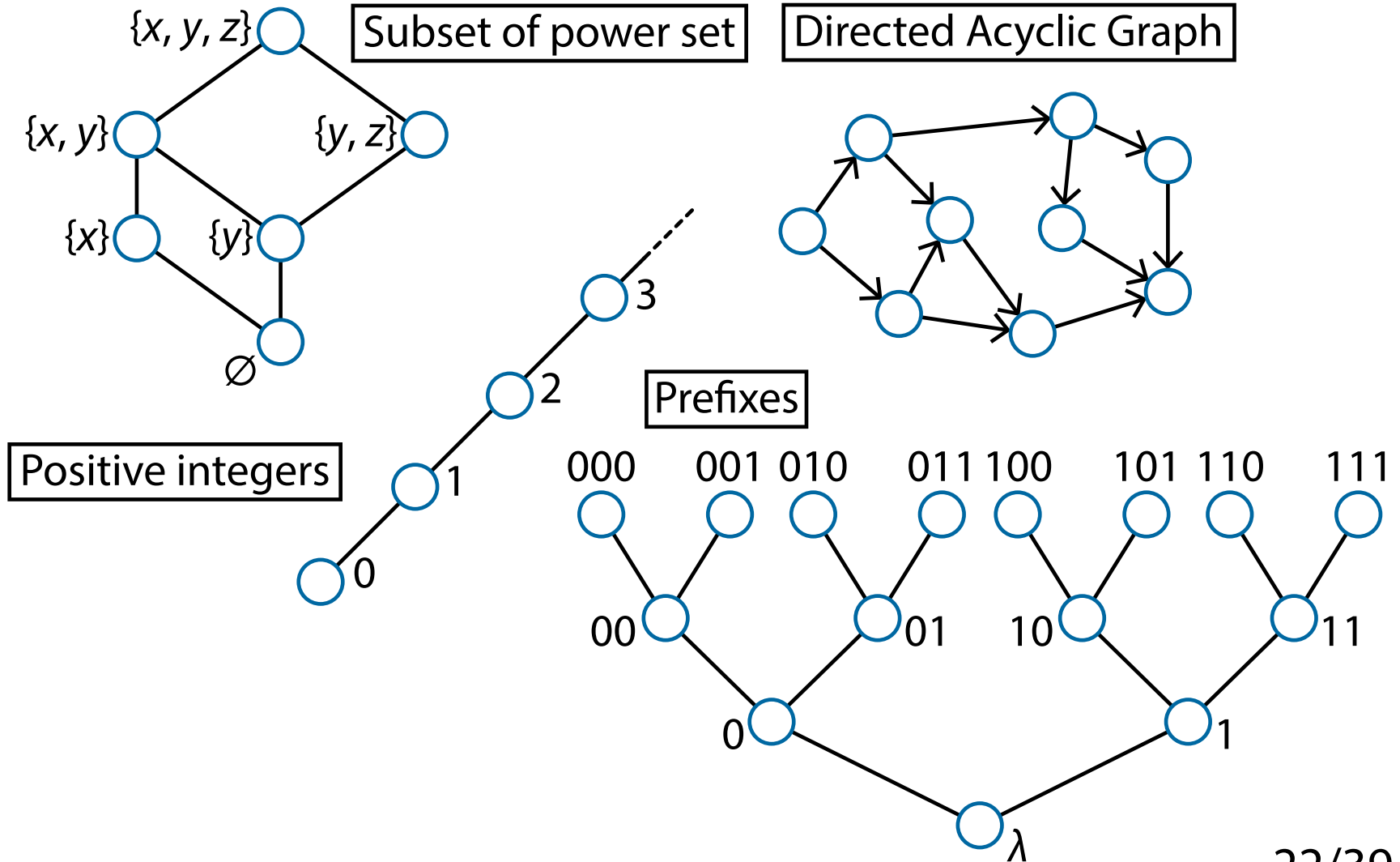
Existing Approach Limited To Power Set



Our Approach Applies To Any Posets



Our Approach Applies To Any Posets



KL Divergence Decomposition

- KL divergence decomposition:

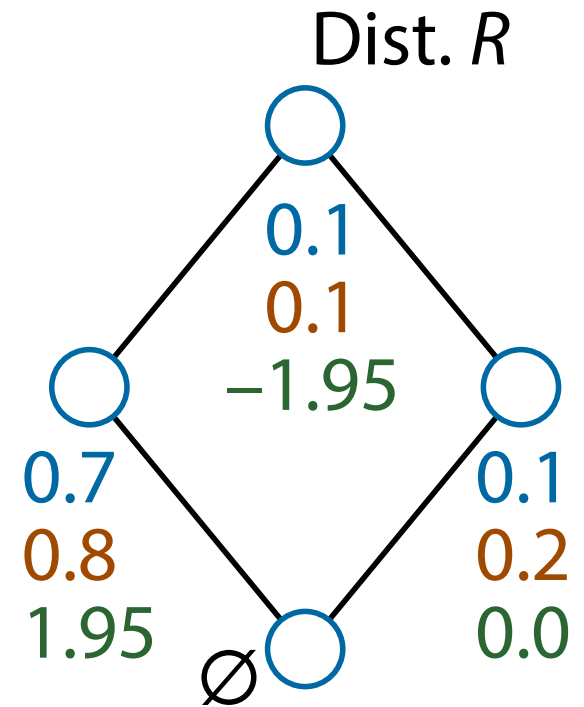
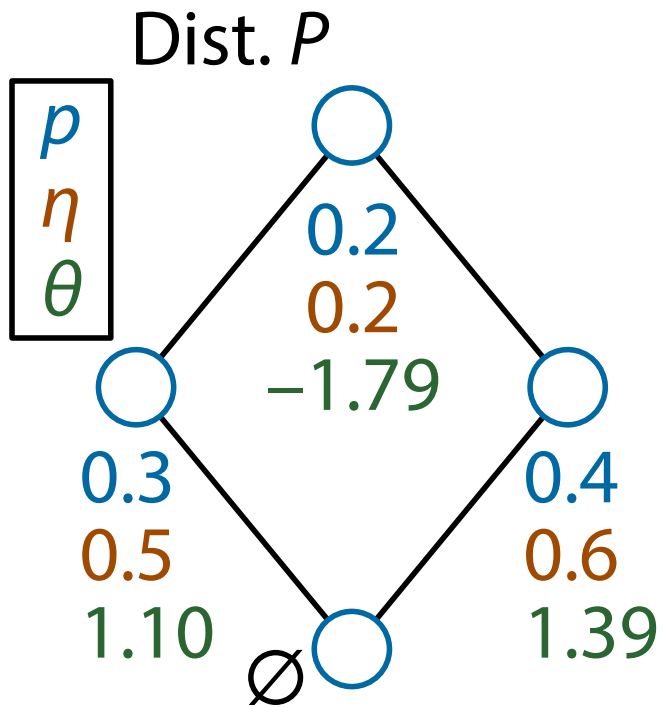
$$D_{\text{KL}}[P, R] = D_{\text{KL}}[P, Q] + D_{\text{KL}}[Q, R]$$

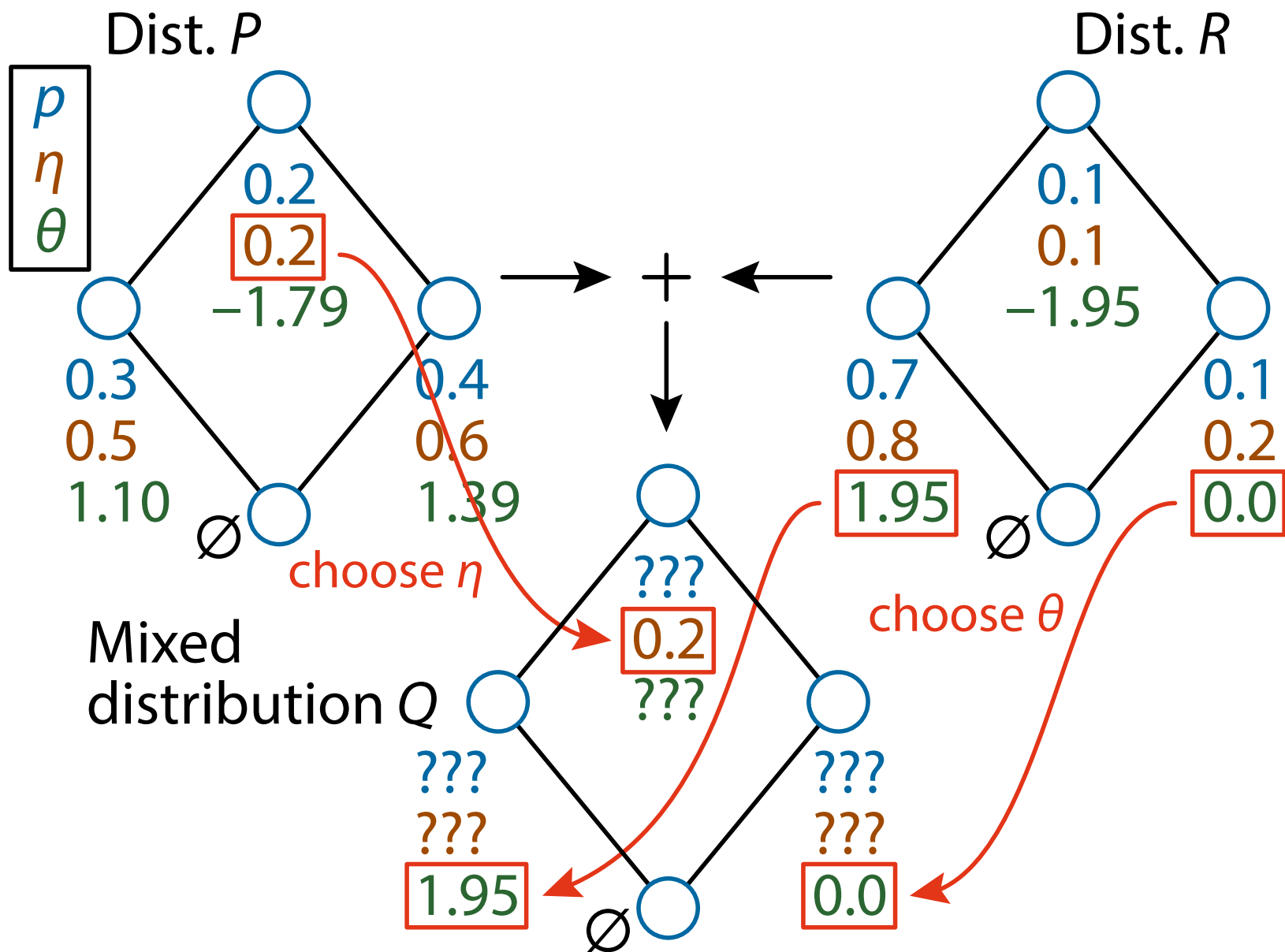
with Q s.t. $\theta_Q(x) = \theta_R(x)$ or $\eta_Q(x) = \eta_P(x)$ for all $x \in S \setminus \{\perp\}$

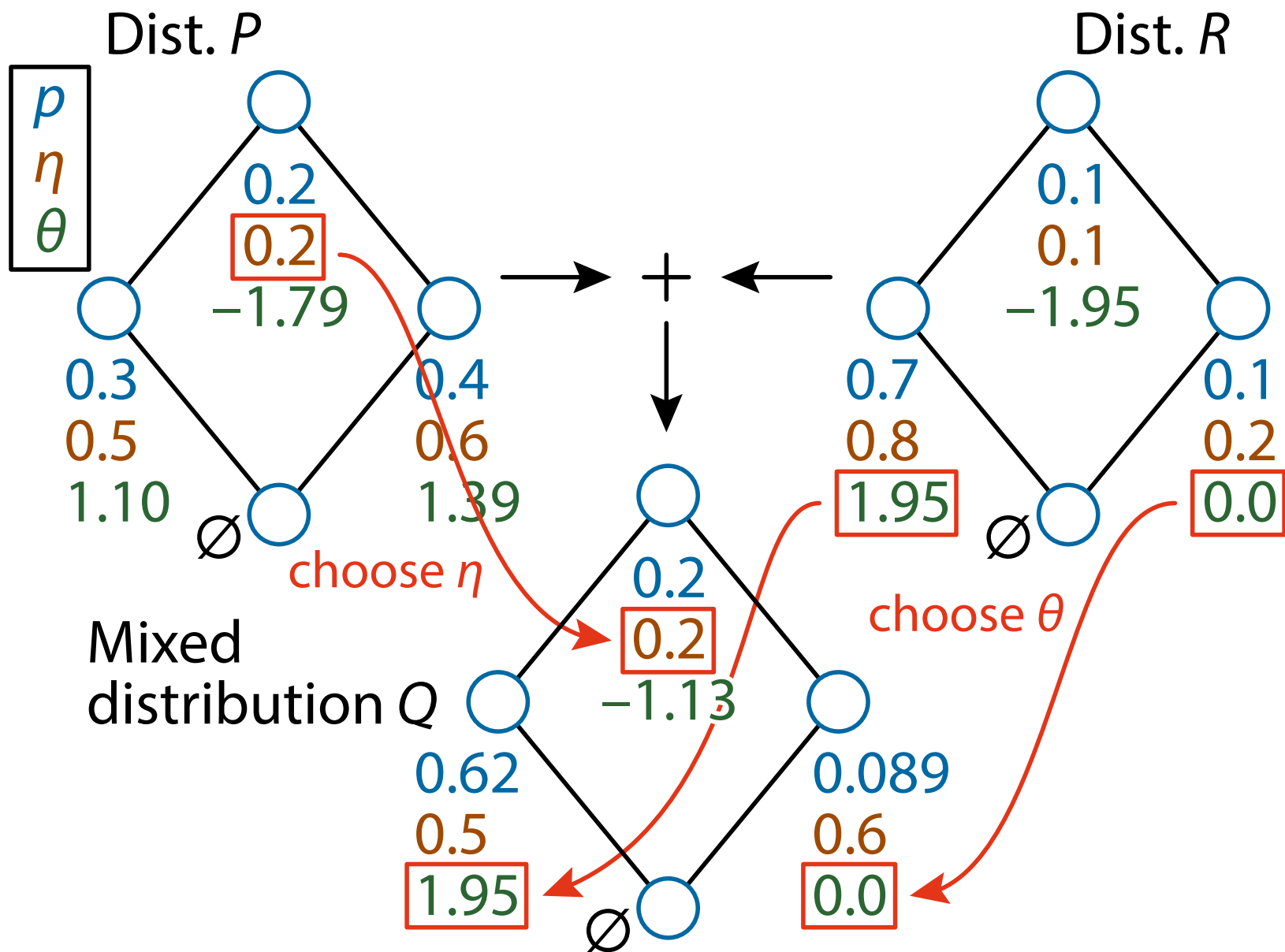
- Q is called the **mixed distribution** of (P, R)
- It is known as the (generalized) **Pythagoras theorem** in Information Geometry

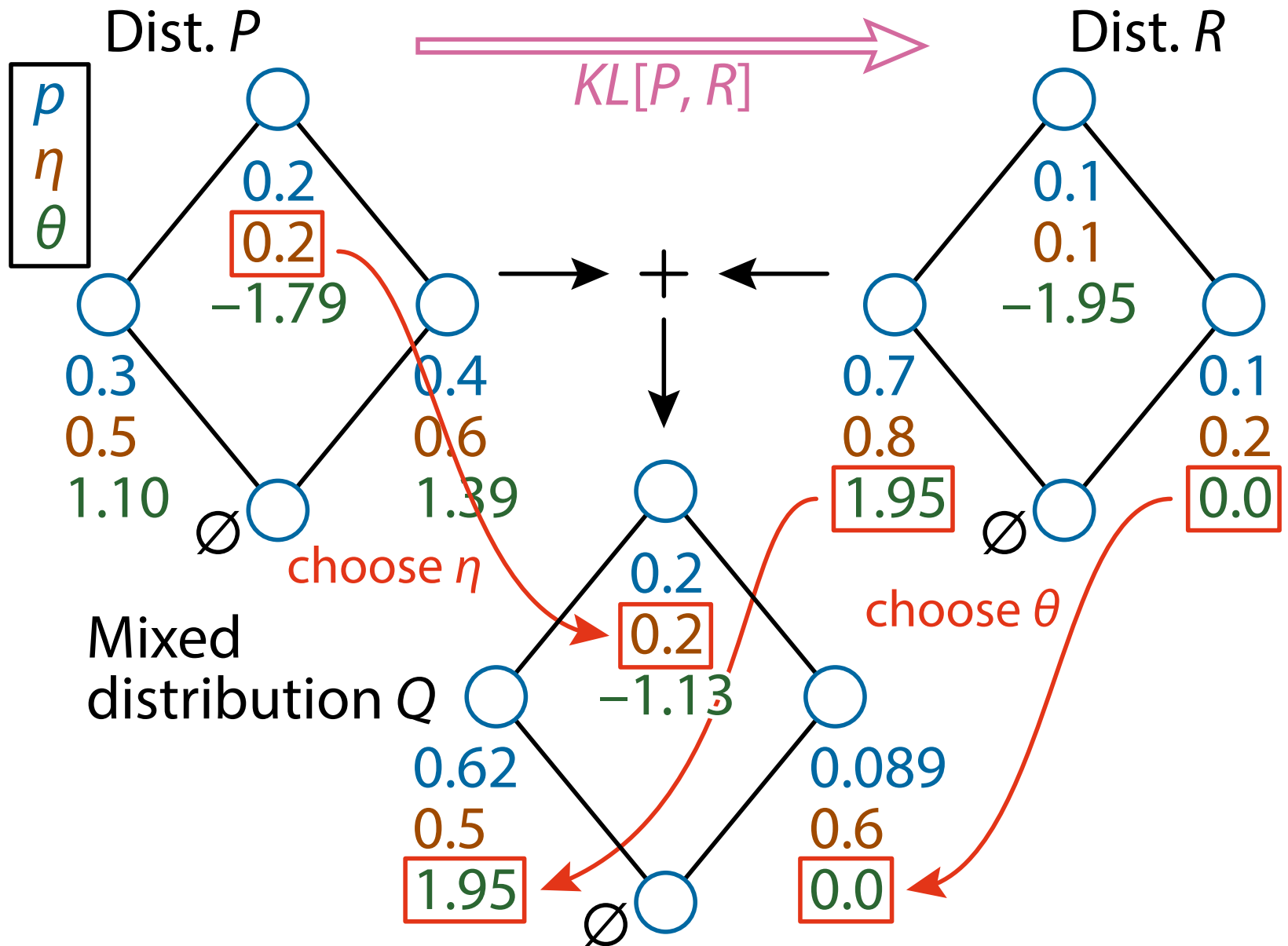
- We can derive from Möbius inversion:

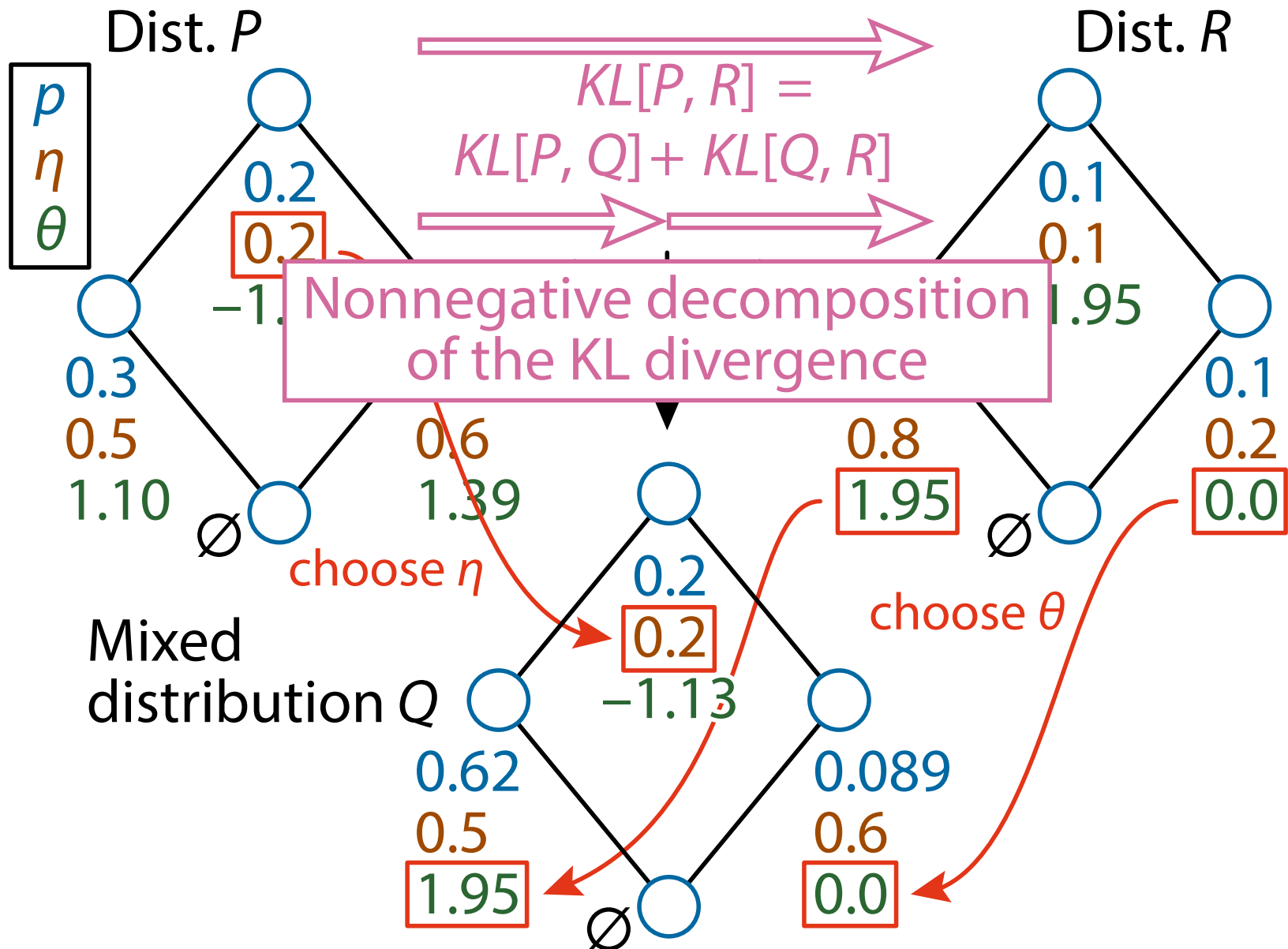
$$\begin{aligned} & D_{\text{KL}}[P, Q] + D_{\text{KL}}[Q, R] - D_{\text{KL}}[P, R] \\ &= \sum_{s \in S} (\eta_Q(s) - \eta_P(s)) (\theta_Q(s) - \theta_R(s)) \end{aligned}$$

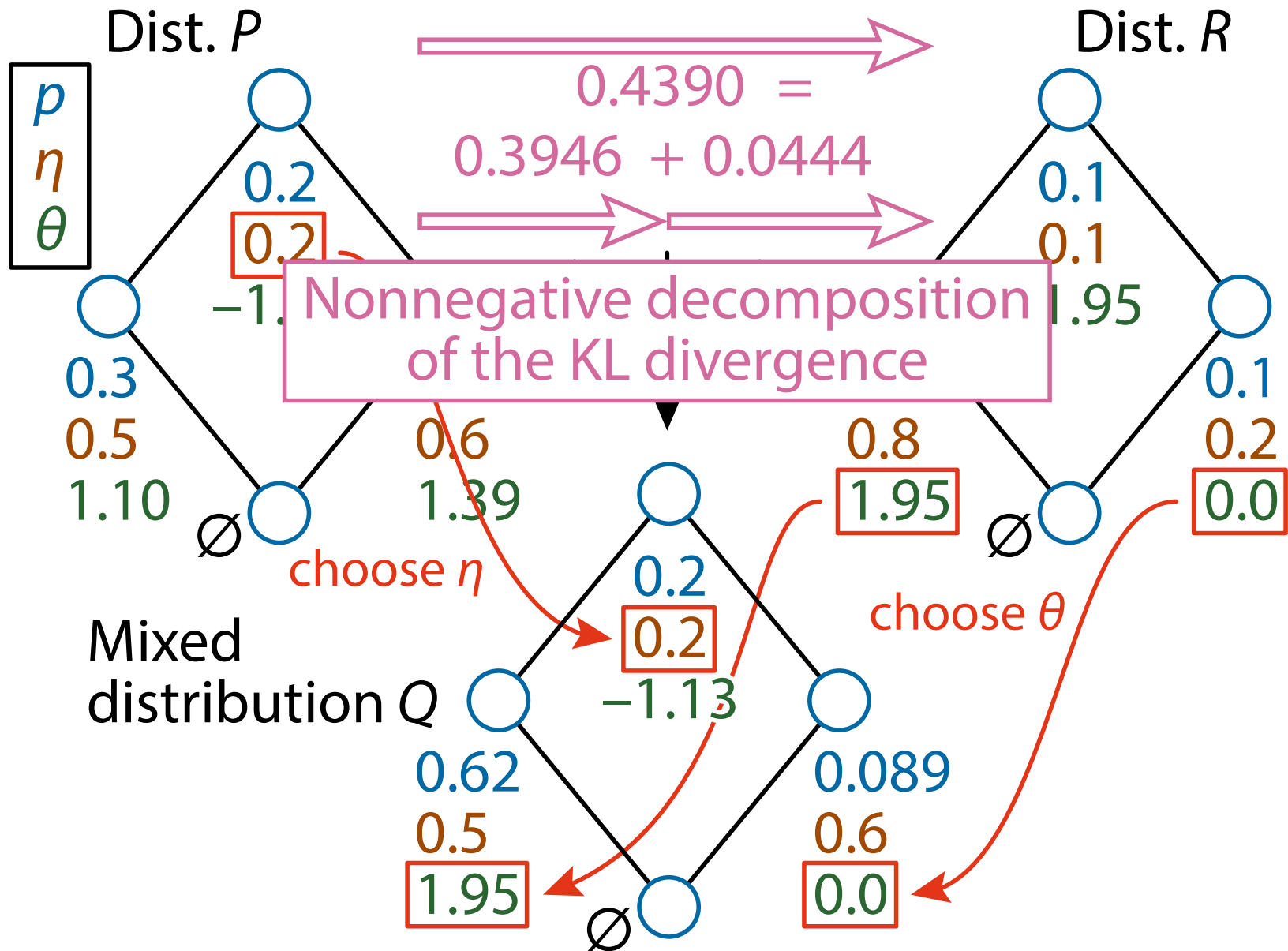


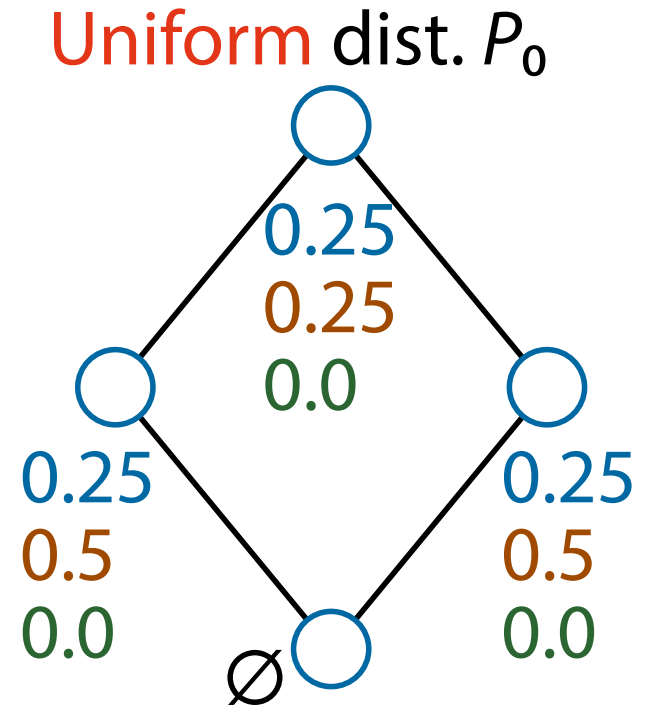
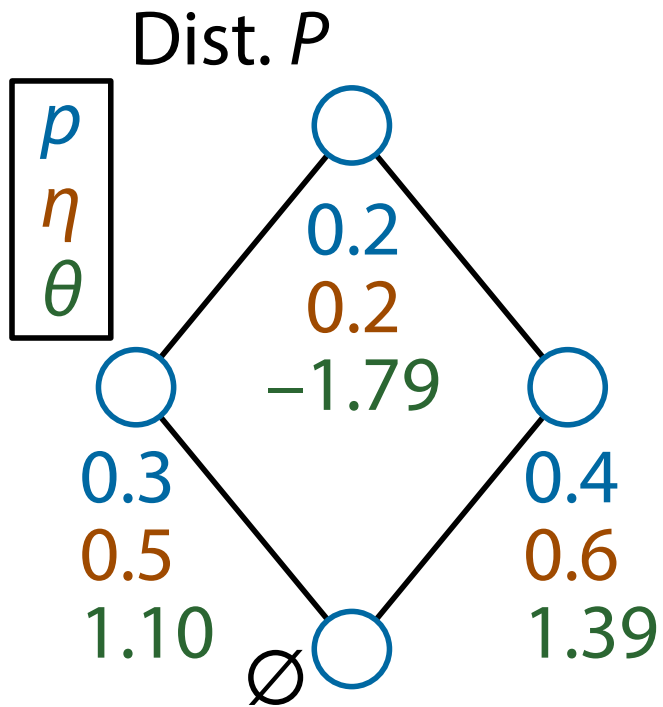


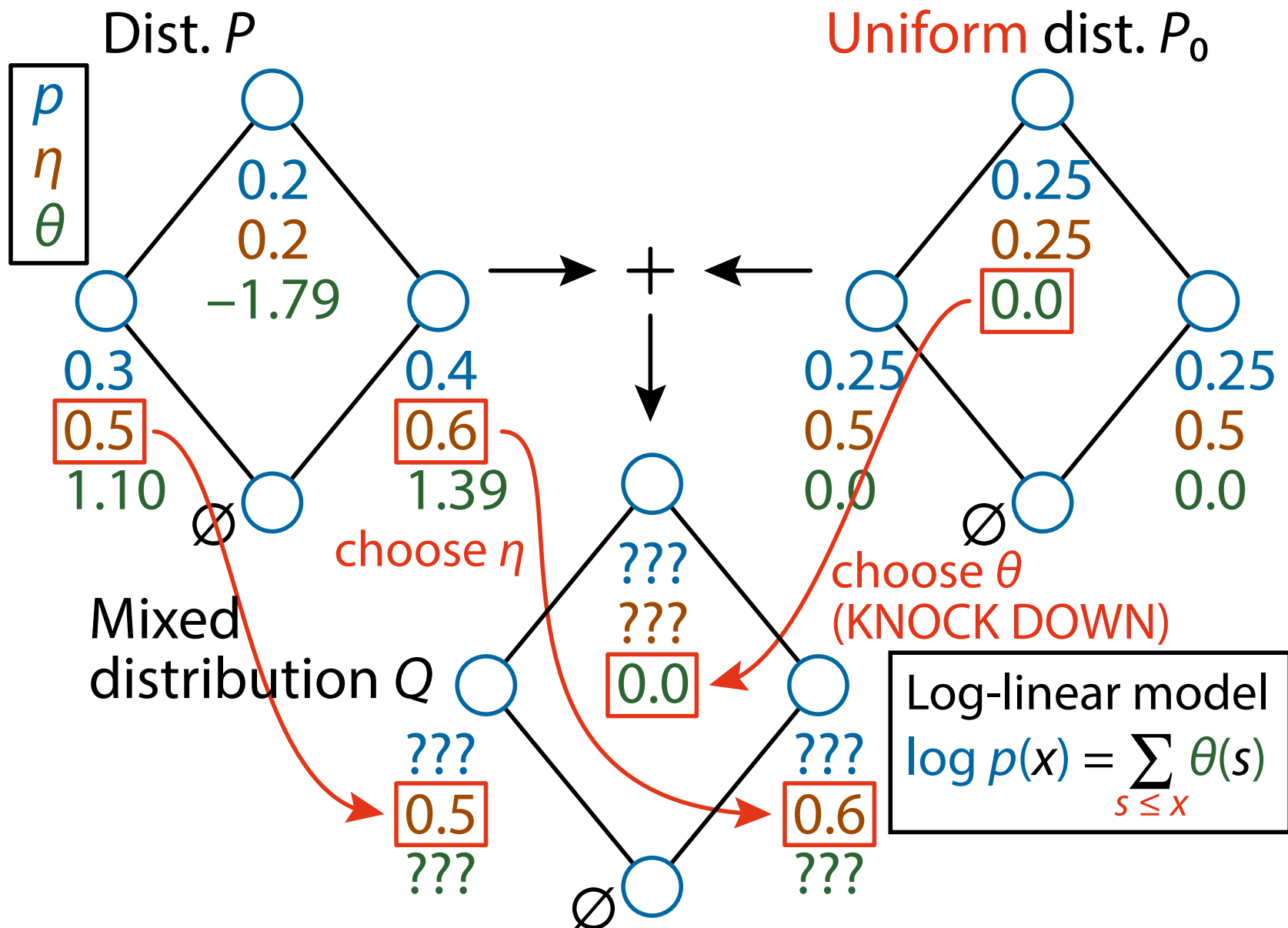


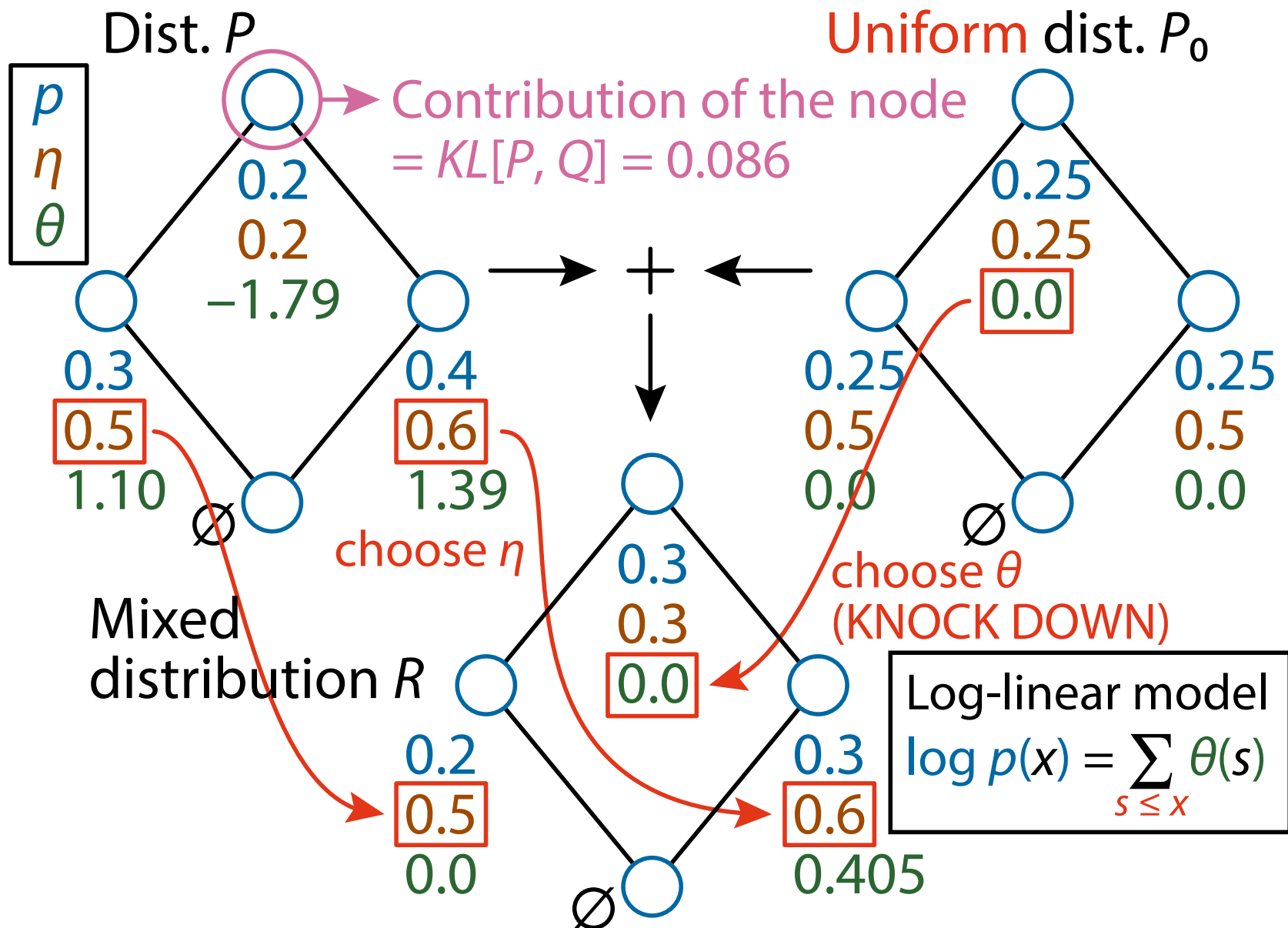


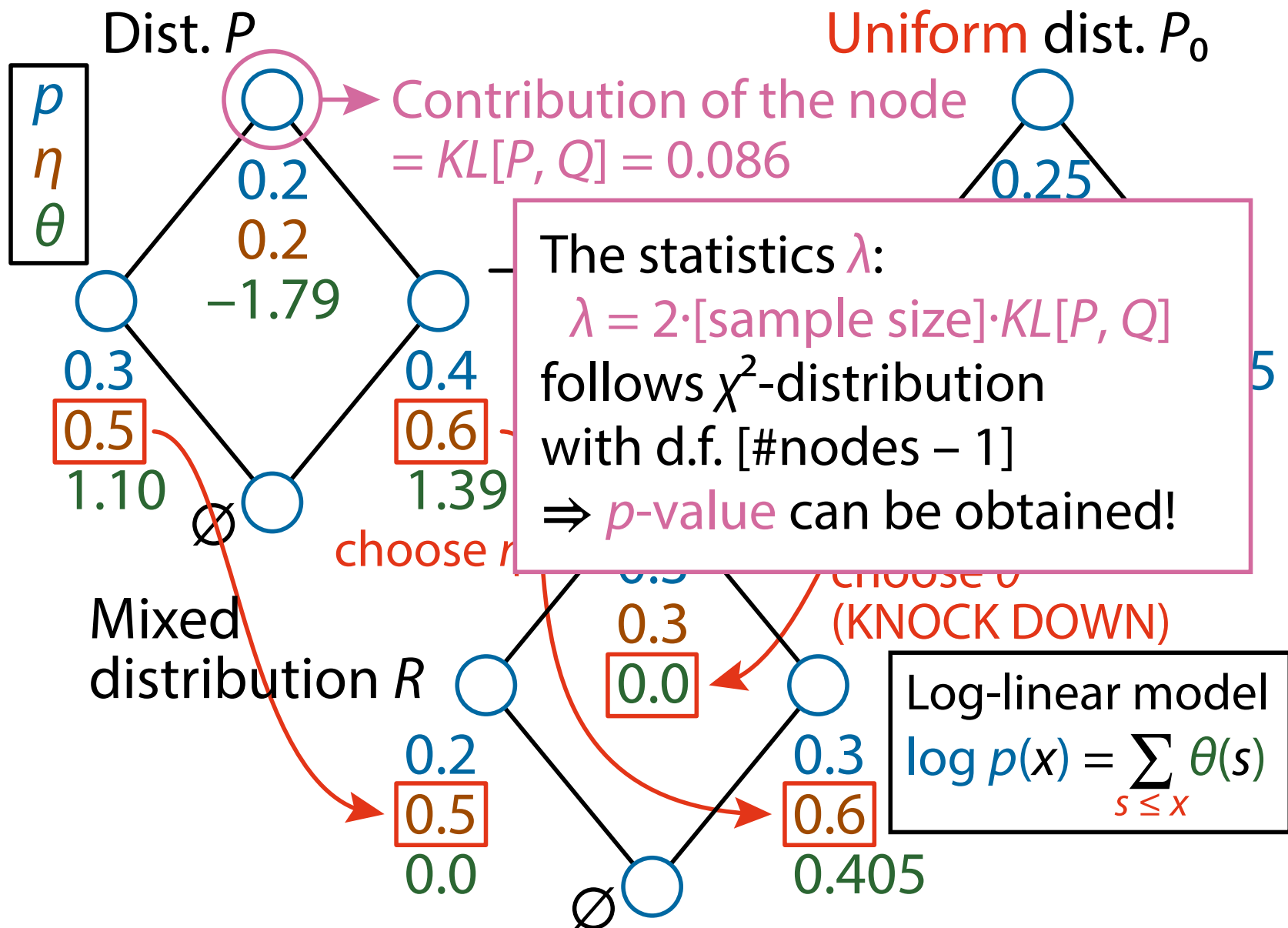




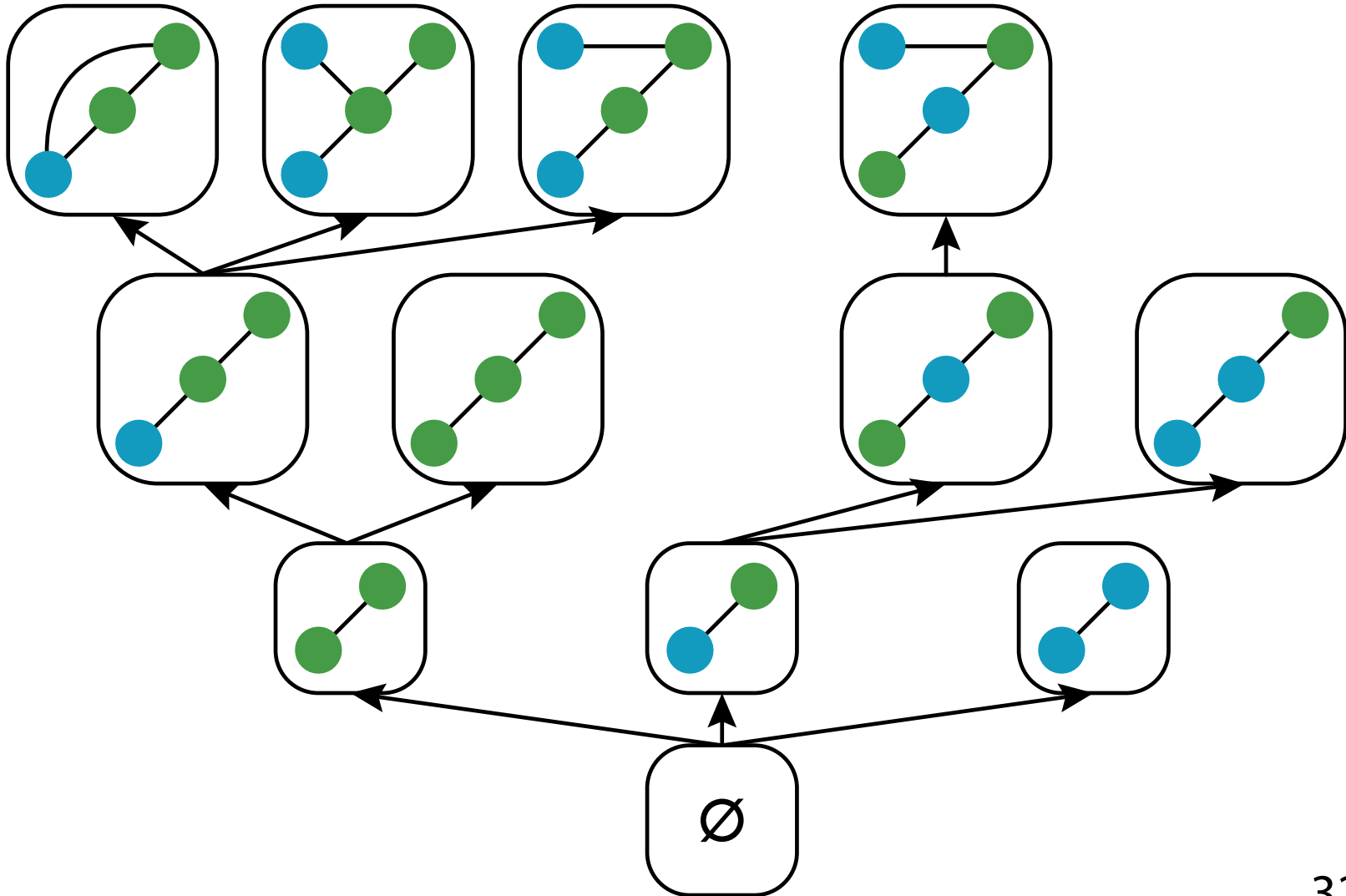




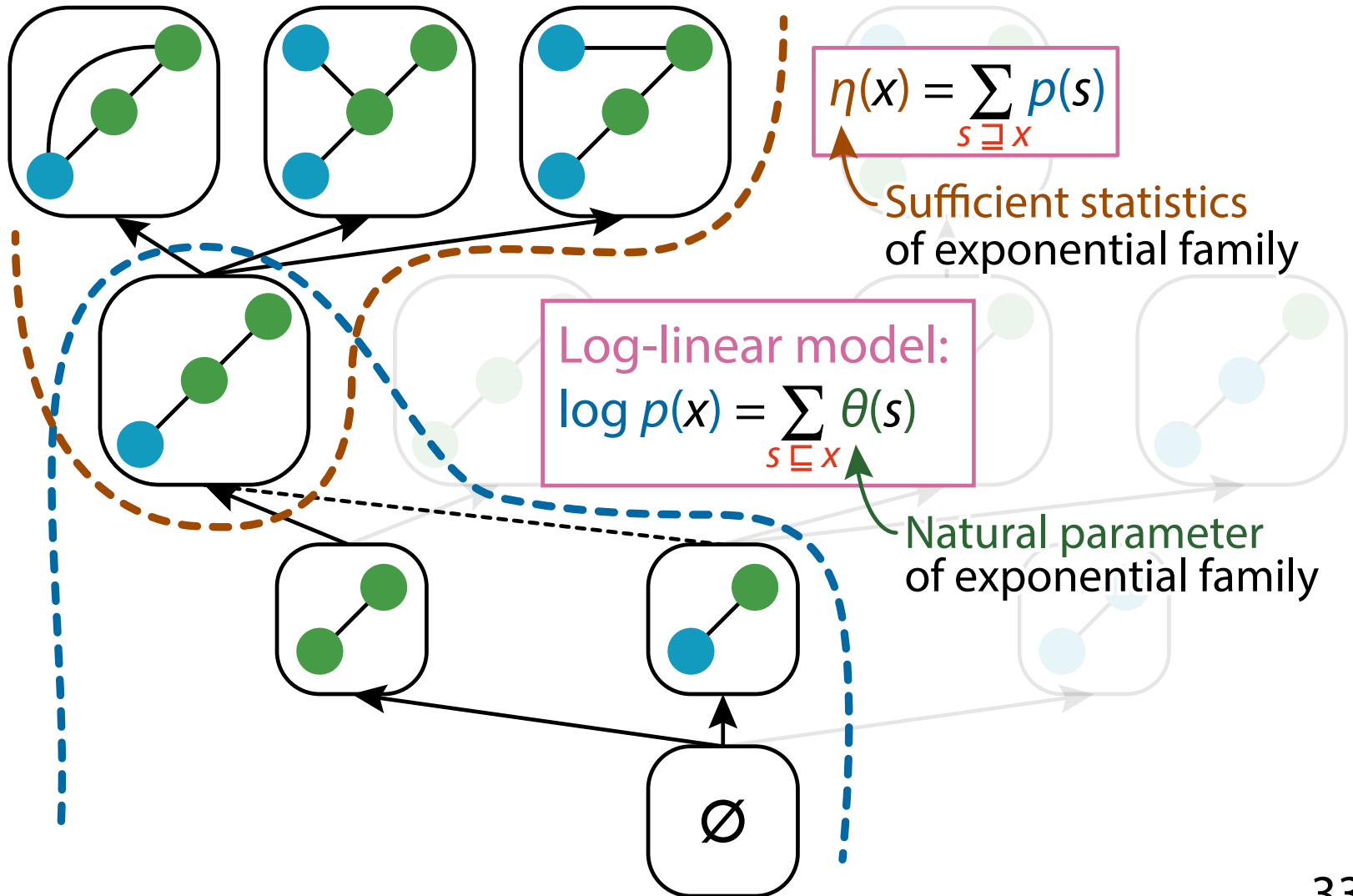




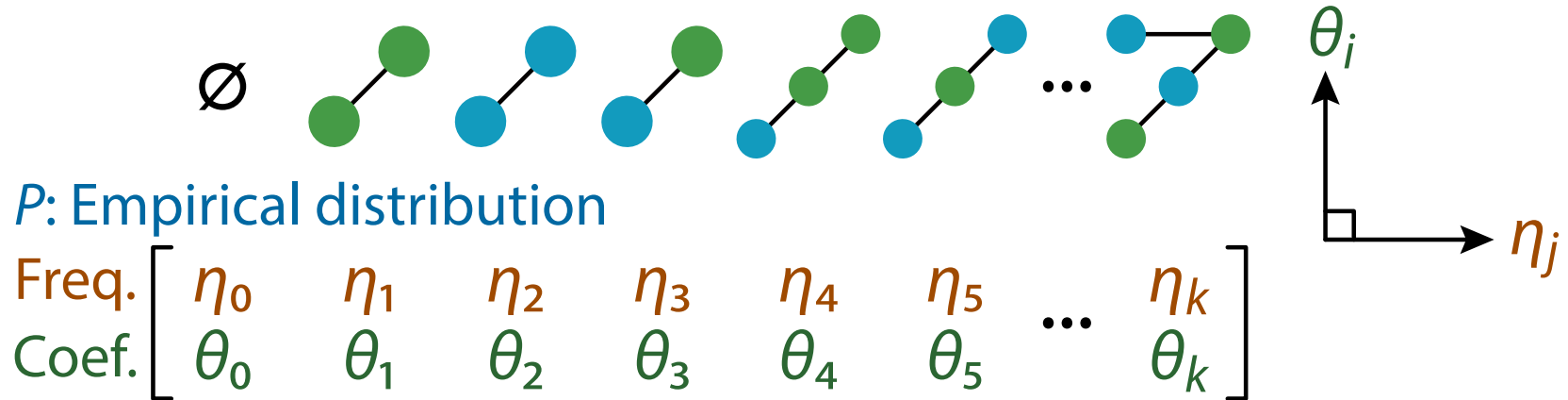
Poset of Subgraphs



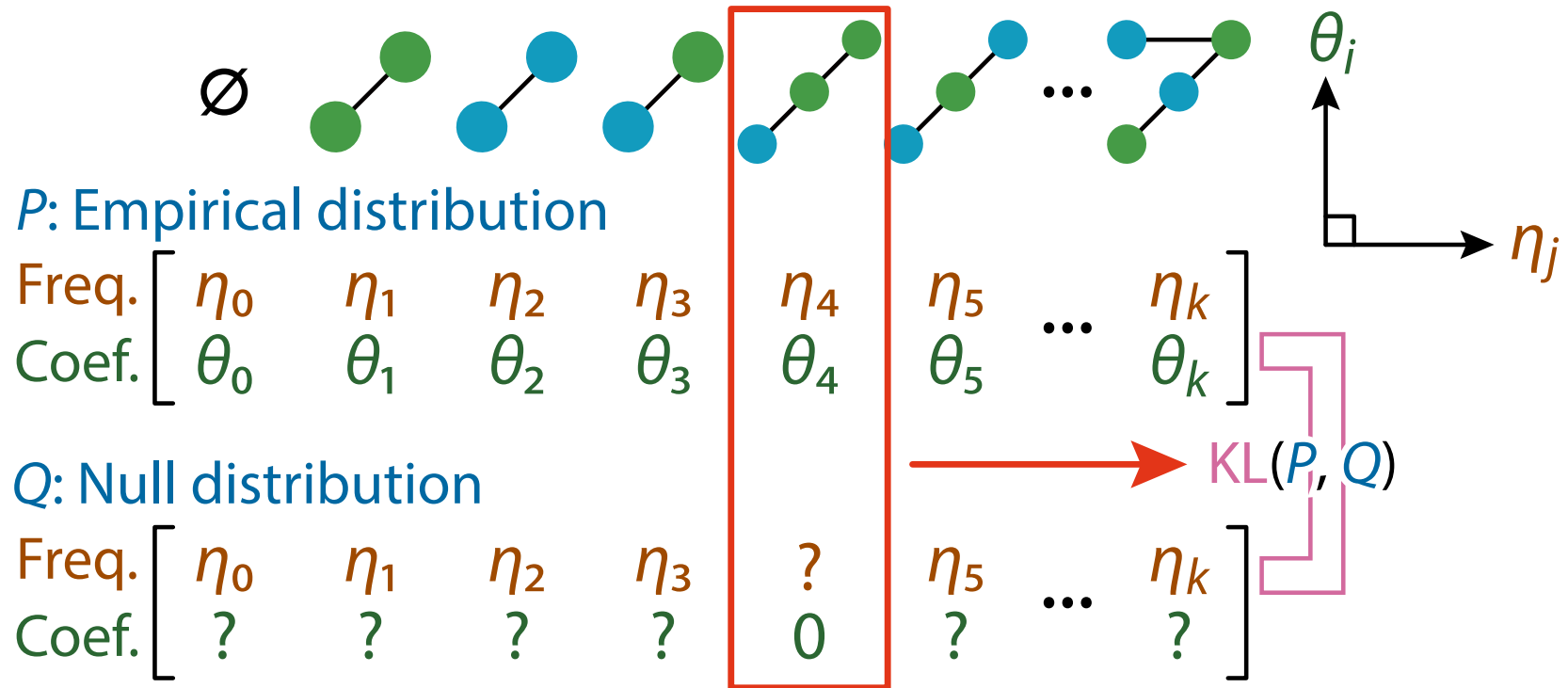
Log-Linear Model on Subgraphs



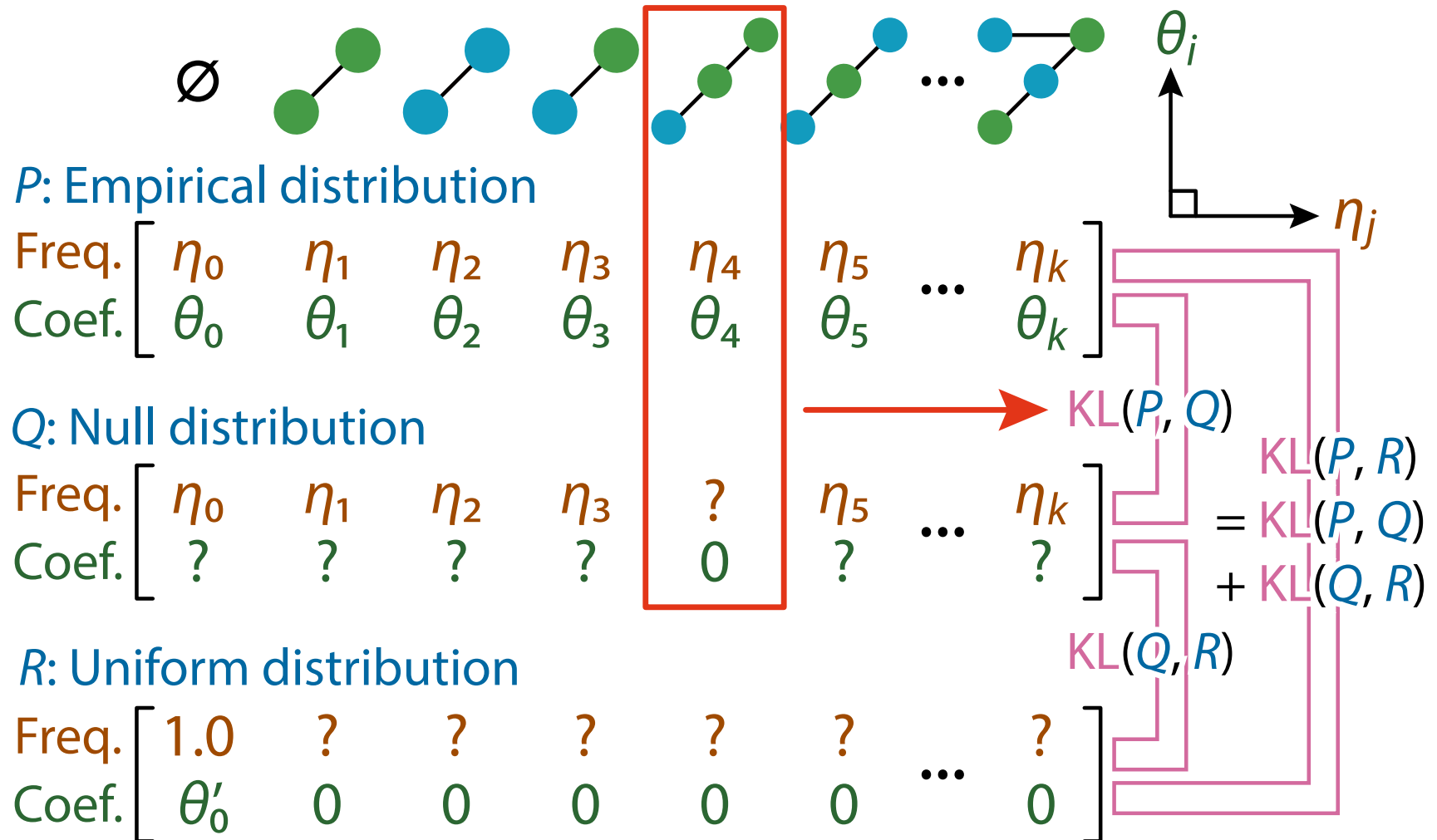
Information of Each Subgraph



Information of Each Subgraph






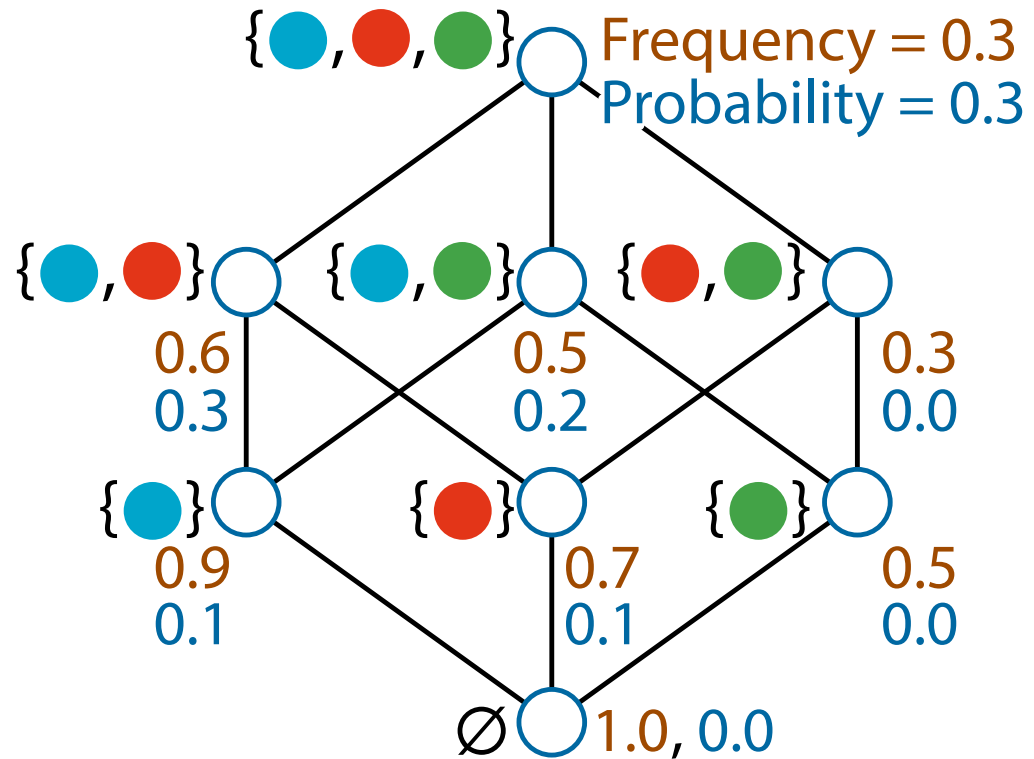
Information of Each Subgraph



Make a Poset from Data

Dataset




			
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

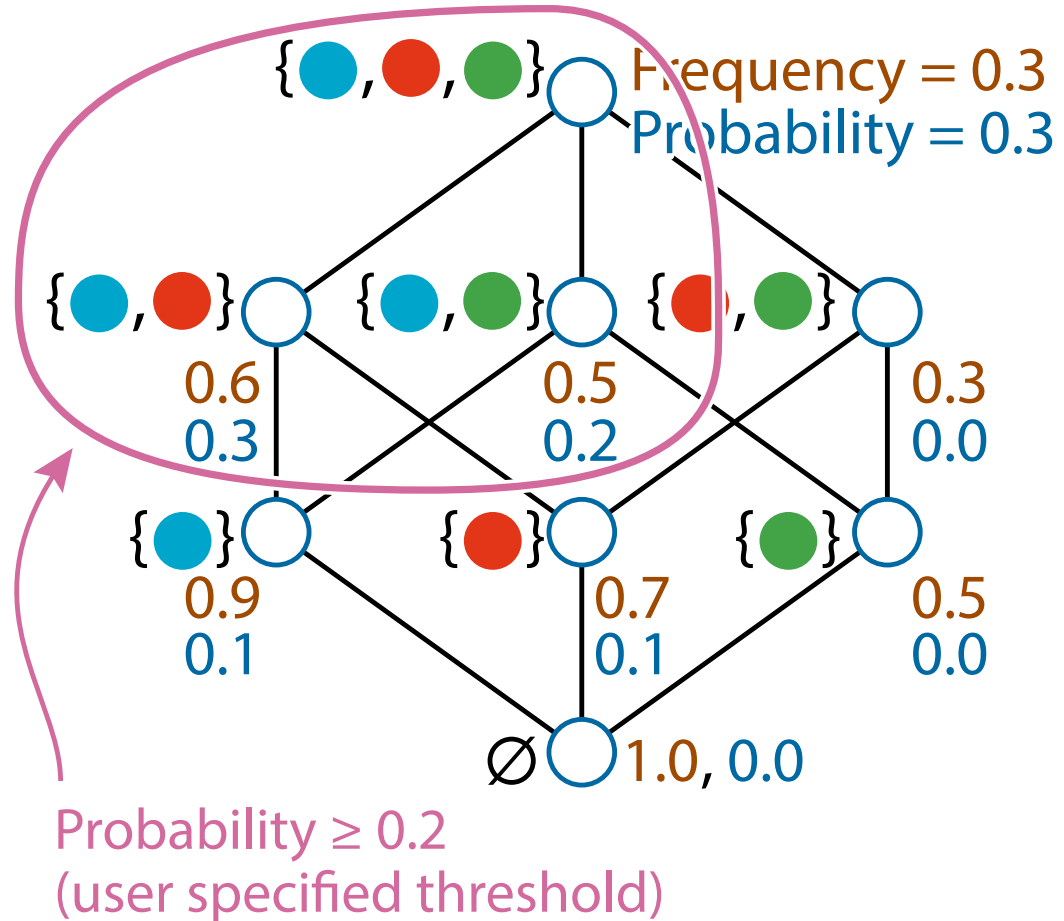


Number of nodes = $2^{\text{\#features}}$
 \Rightarrow combinatorial explosion!

Make a Poset from Data




Dataset

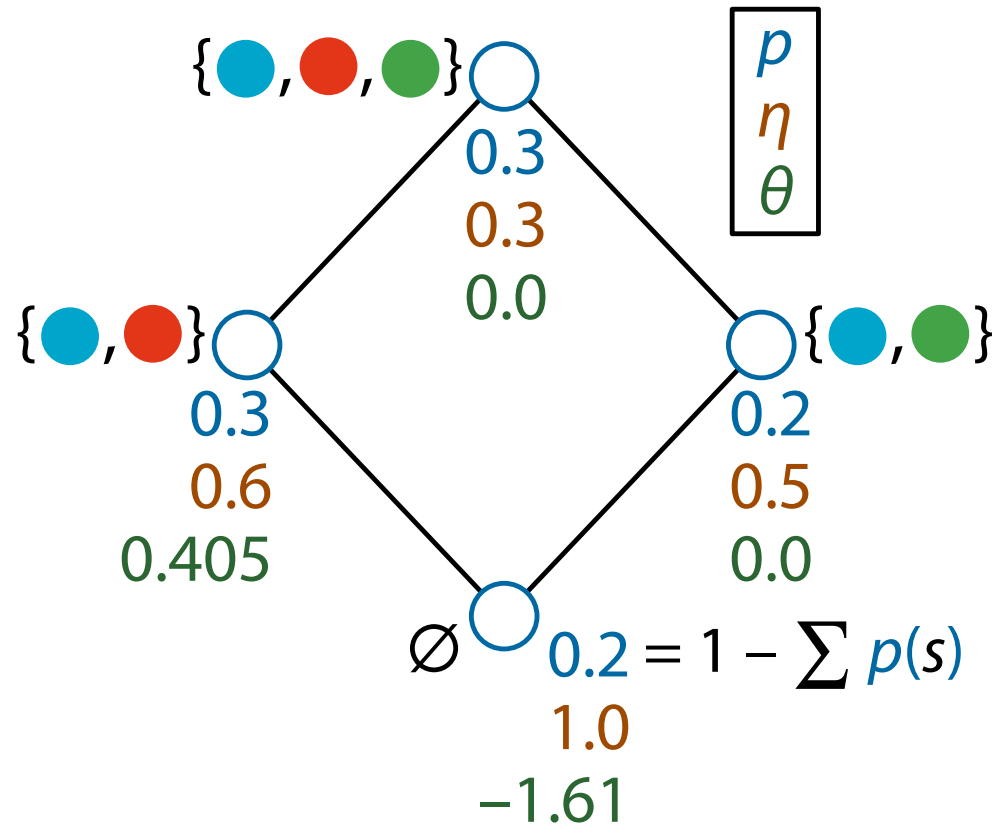
			
ID 1:	1	1	0
ID 2:	1	1	1
ID 3:	1	1	0
ID 4:	1	1	1
ID 5:	1	1	0
ID 6:	1	0	1
ID 7:	1	0	1
ID 8:	1	1	1
ID 9:	1	0	0
ID10:	0	1	0



Remove Nodes with Probability 0

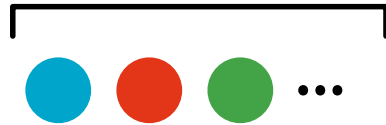
Dataset

			
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



Example on Real Data (kosarak)

features: 41,270

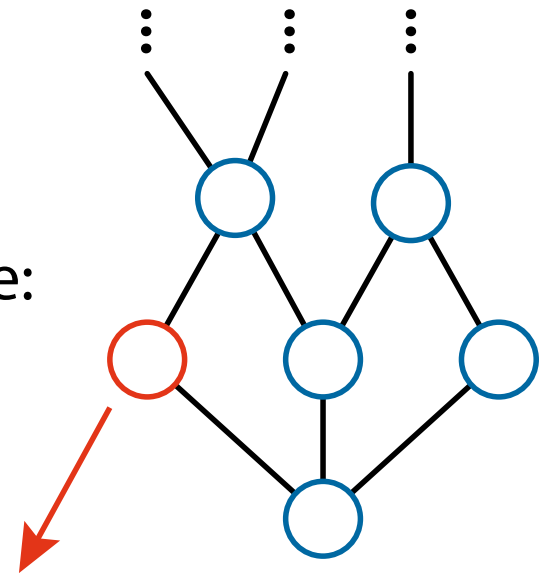


ID 1:	1	1	0
ID 2:	1	1	1
ID 3:	1	1	0 ...
ID 4:	1	1	1
ID 5:	1	1	0
⋮	⋮		

Total runtime:
4.95 seconds

Sample size:
990,002

nodes: 3,253
(Threshold: 10^{-5})



significant interactions: **583**



Single feature: 537

Pairwise interactions: 41

Triple interactions: 5

Example on Real Data (accidents)

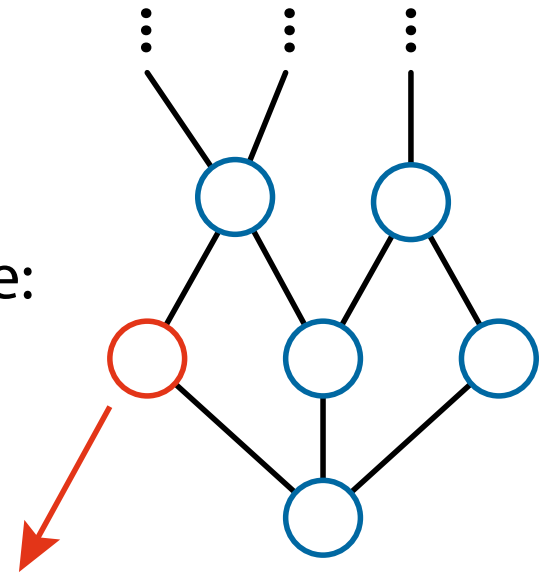
features: 468

				...
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	
⋮	⋮			

Total runtime:
4.95 seconds

Sample size:
340,183

nodes: 281
(Threshold: 5×10^{-6})



significant interactions: 280
features in each interaction
is between 26 to 41

Conclusion

- A close connection between the **partial order structure** and **information geometry**
 - **Möbius inversion** leads to the **dually flat manifolds**
 - M. Sugiyama, H. Nakahara, K. Tsuda, *Information Decomposition on Structured Space*, IEEE ISIT (2016)
 - S. Amari, *Information geometry on hierarchy of probability distributions*, IEEE Trans. Info. Theory (2001)
 - H. Nakahara, S. Amari, *Information-geometric measure for neural spikes*, Neural Computation (2002)
- We can decompose the KL divergence and asses the significance on any posets