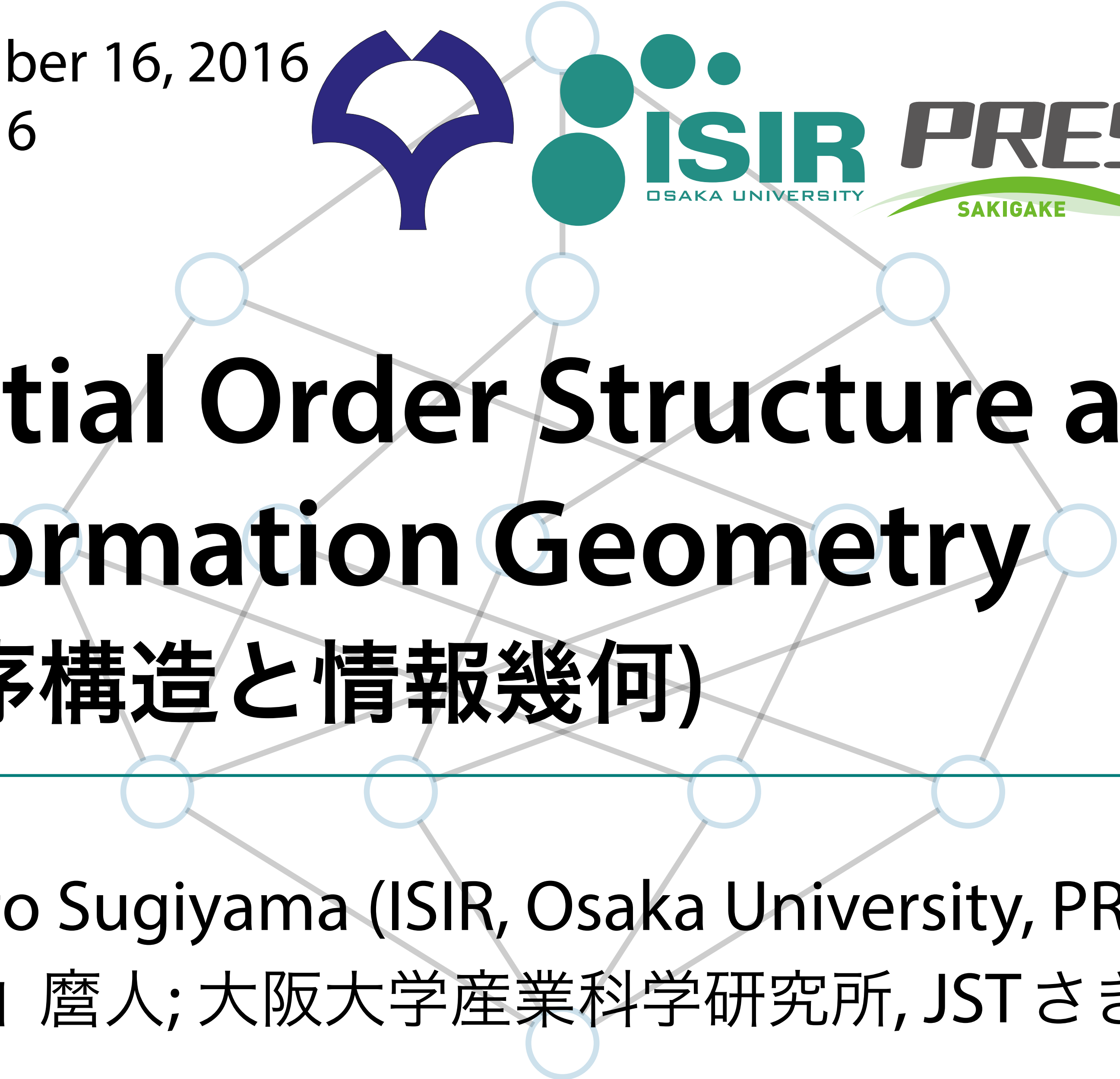
November 16, 2016
IBIS2016



Partial Order Structure and Information Geometry

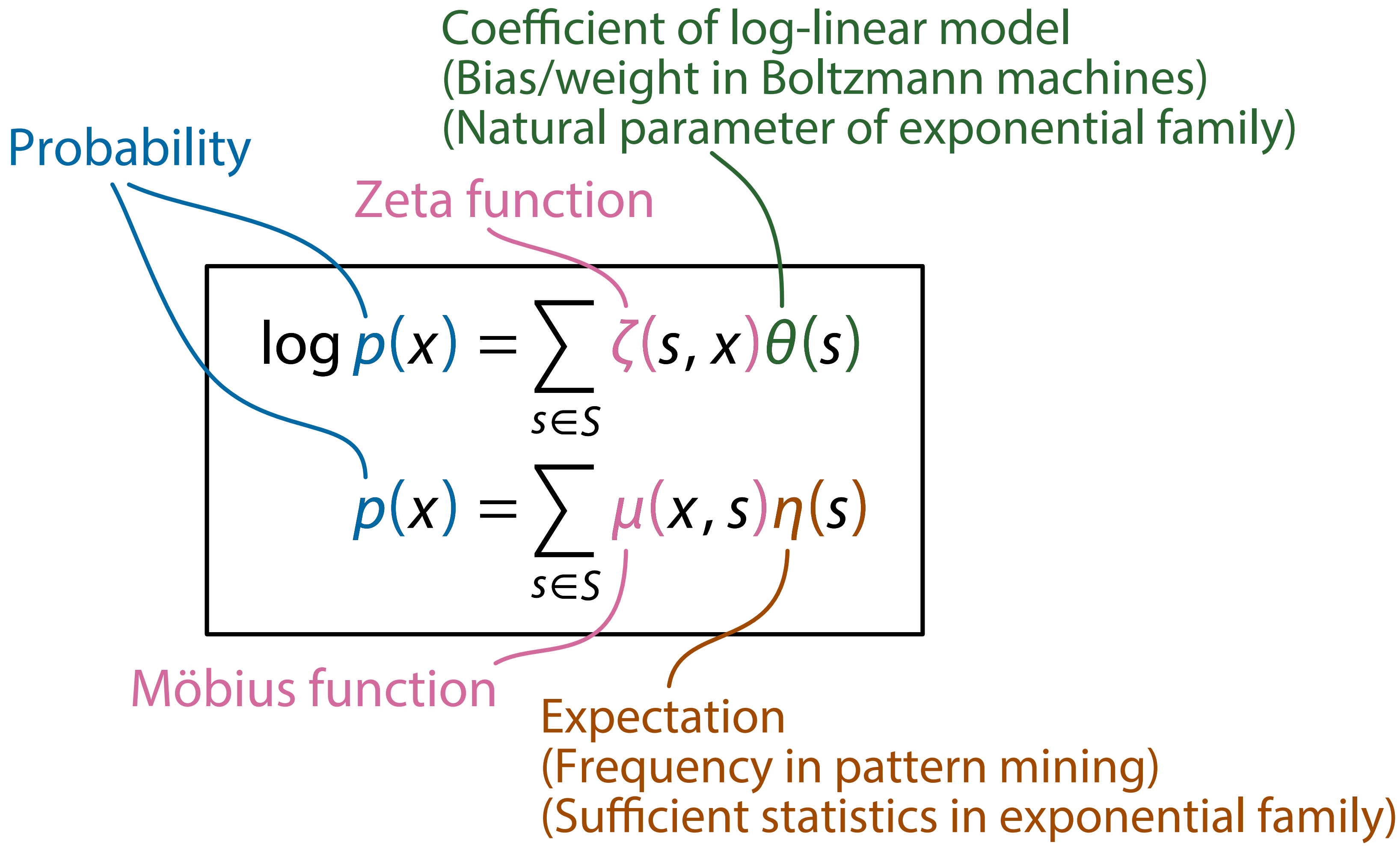
(順序構造と情報幾何)

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

Today’s Model on Poset $(S, \leq)$

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

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



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

Probability

Zeta function

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

Möbius function

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

Outcome

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

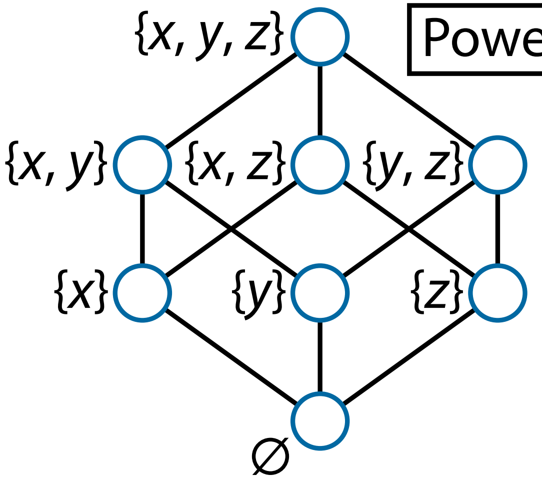
1. KL divergence decomposition:

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

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

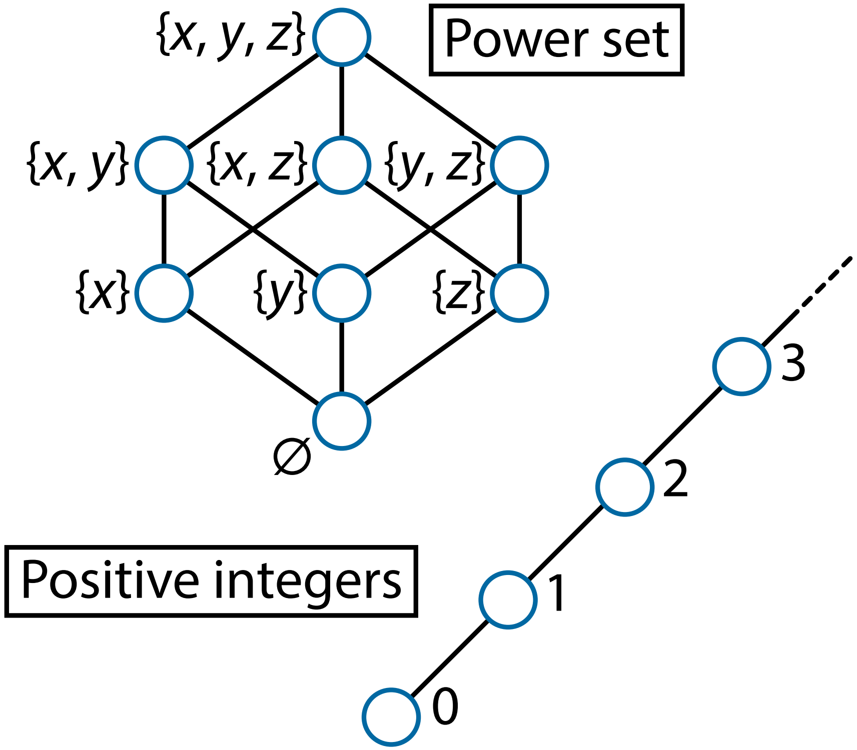
2. The set of probability distributions on $(S, \leq)$ is a dually flat manifold w.r.t. θ and η
   - p, θ, and η are coordinate systems
   - θ and η are orthogonal
   - θ introduces the structure of exponential family
   - η introduces the structure of mixture family

Partially Ordered Sets



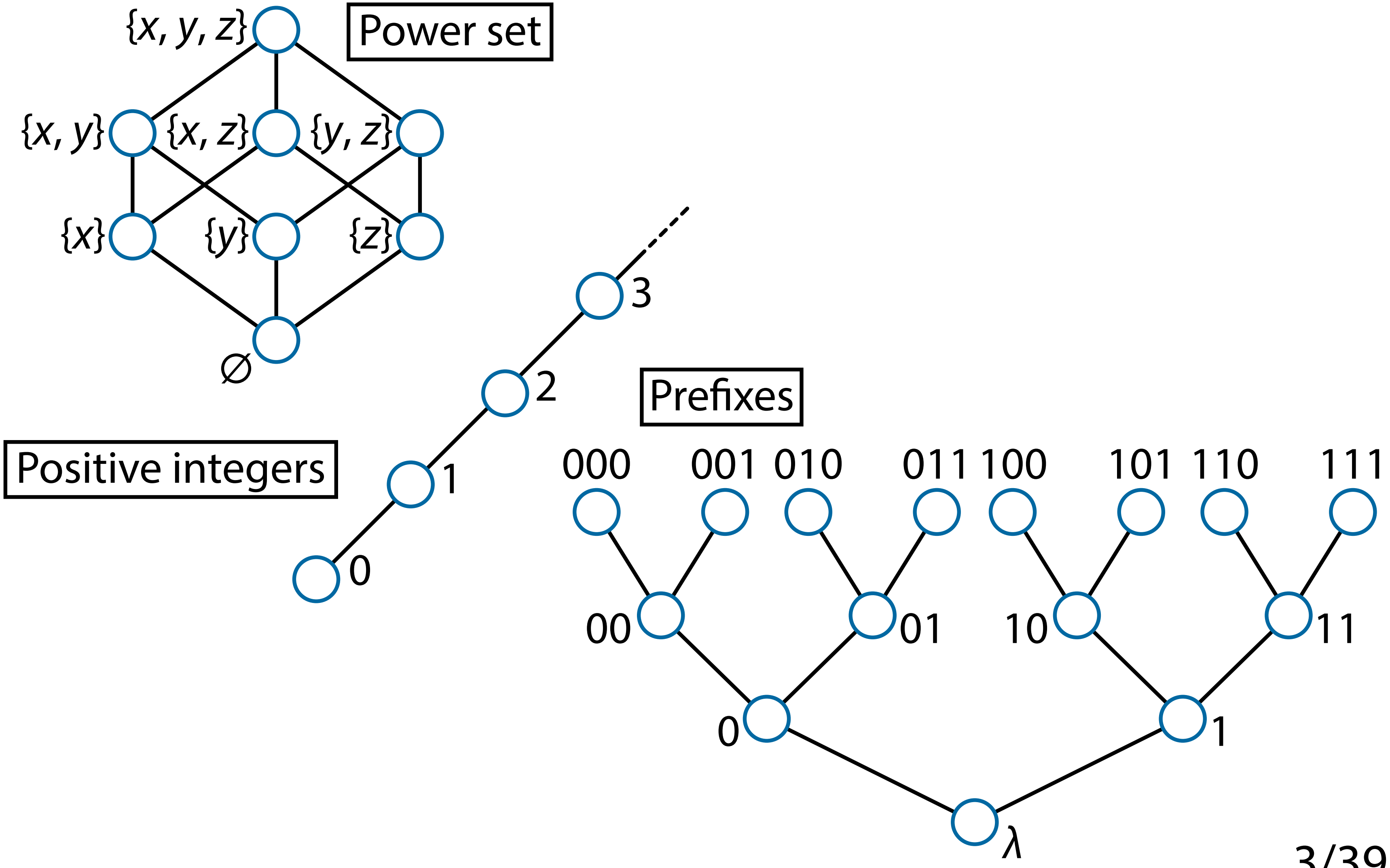
Power set

Partially Ordered Sets


Power set

Positive integers

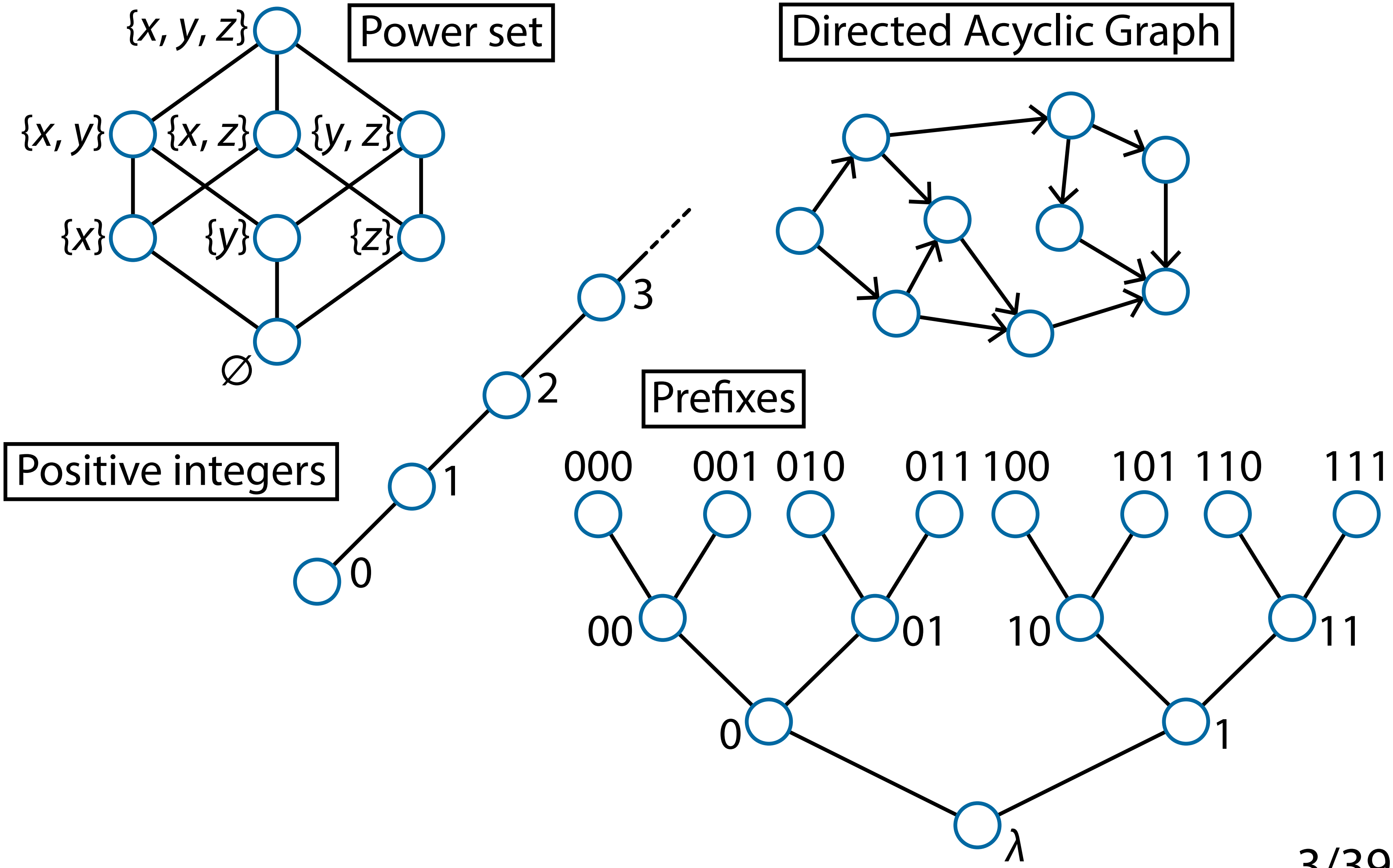
Partially Ordered Sets


Power set

Positive integers

Prefixes

Partially Ordered Sets



Power set

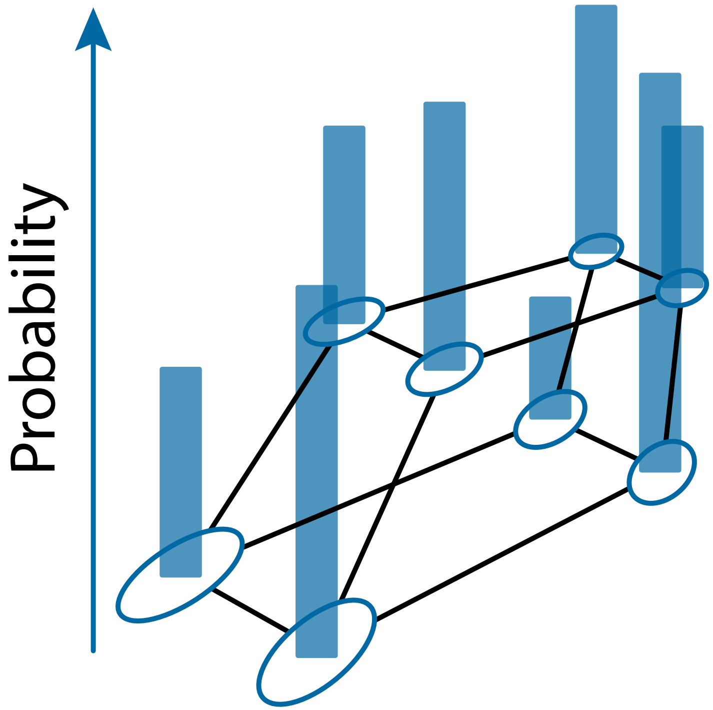
Directed Acyclic Graph

Positive integers

Prefixes

Posets with Probability Distribution

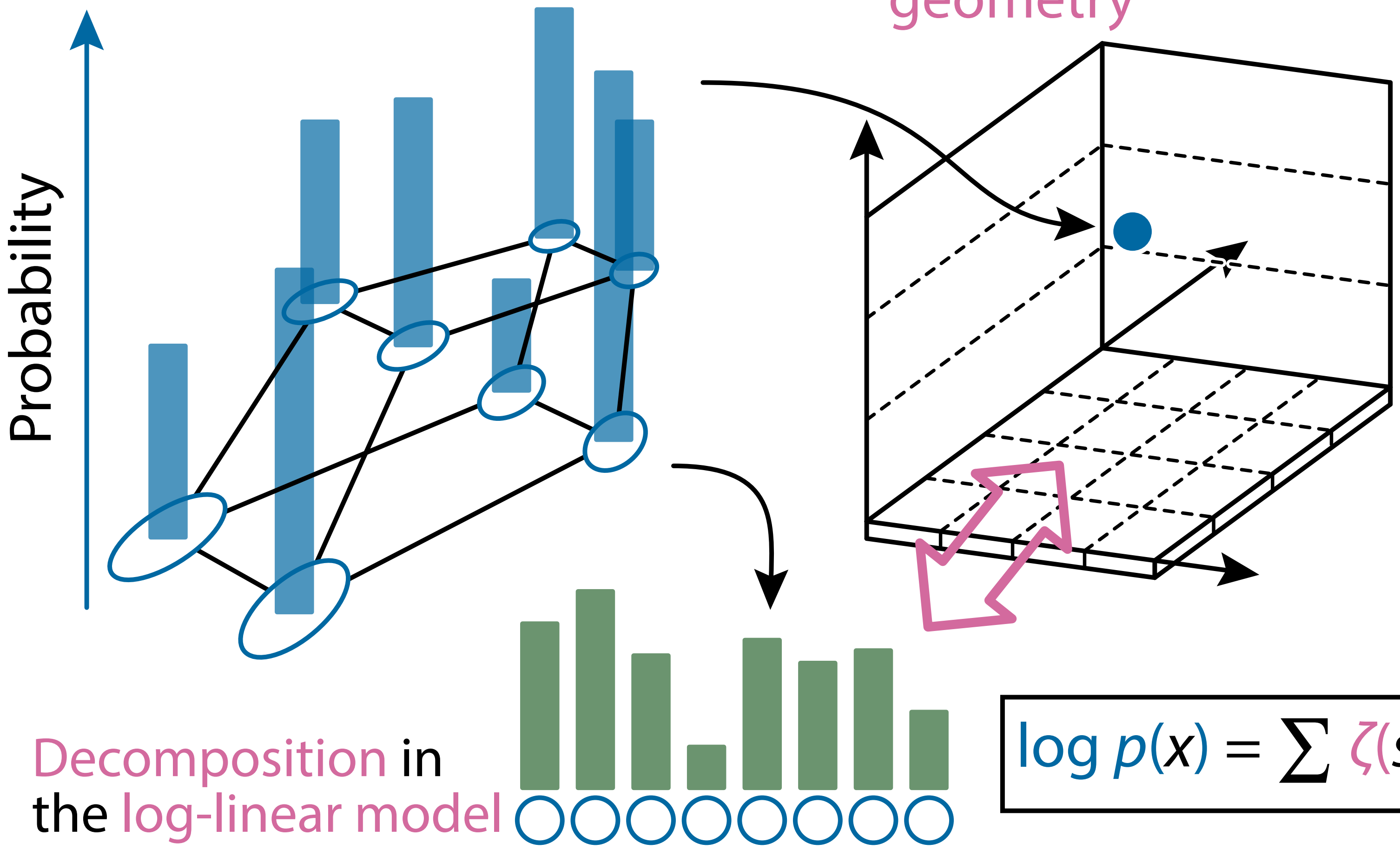
Probability distribution
on posets (partially ordered sets)



Posets with Probability Distribution

Probability distribution
on posets (partially ordered sets)

Information
geometry

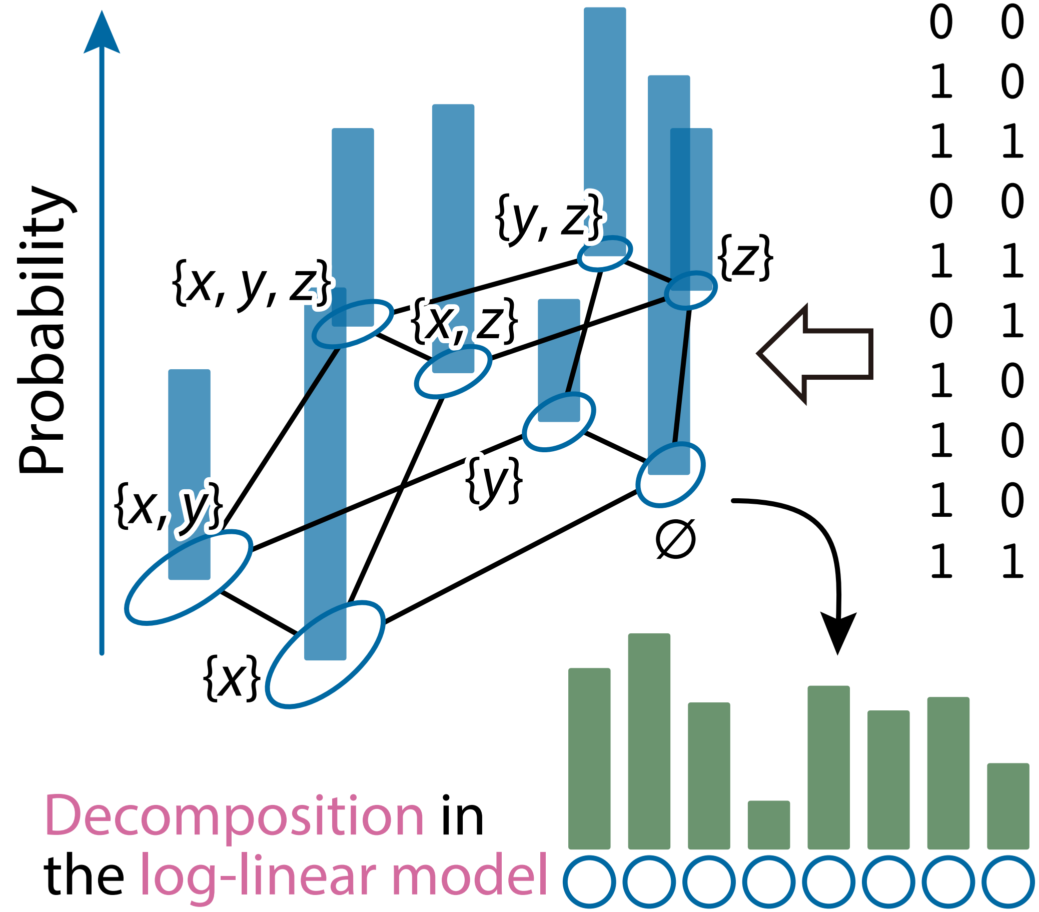


Decomposition in
the log-linear model

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

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 intractions

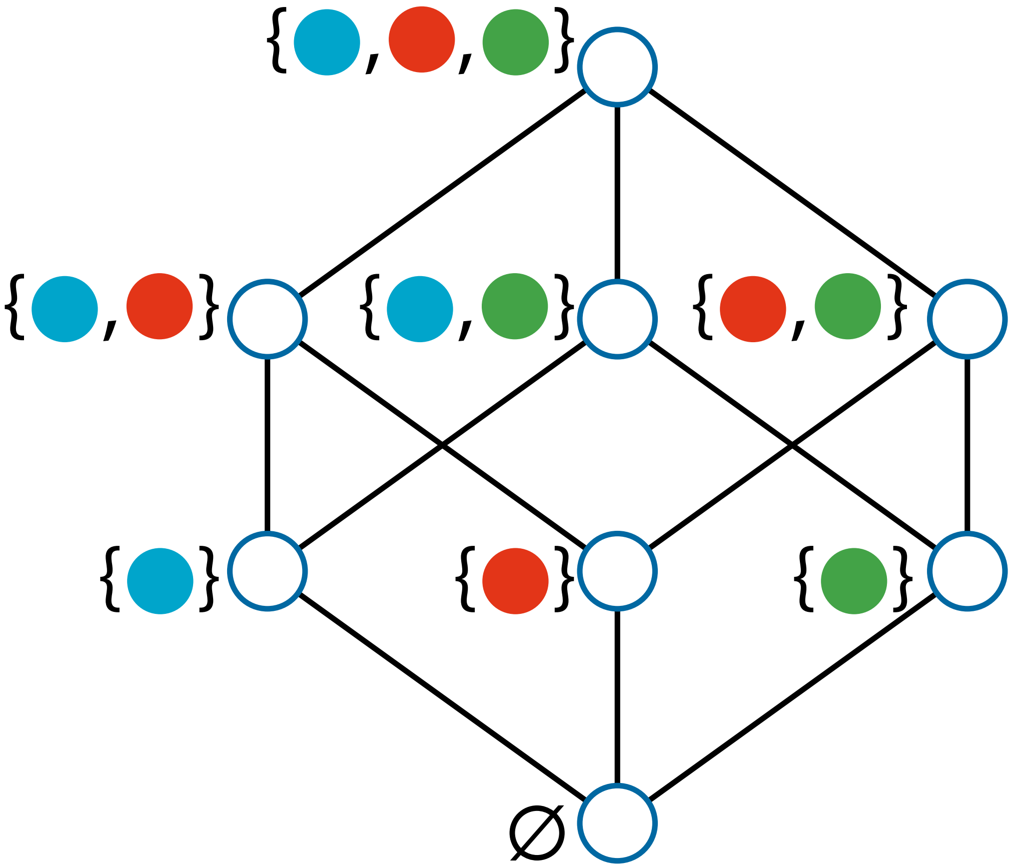
Decomposition in the log-linear model

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

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

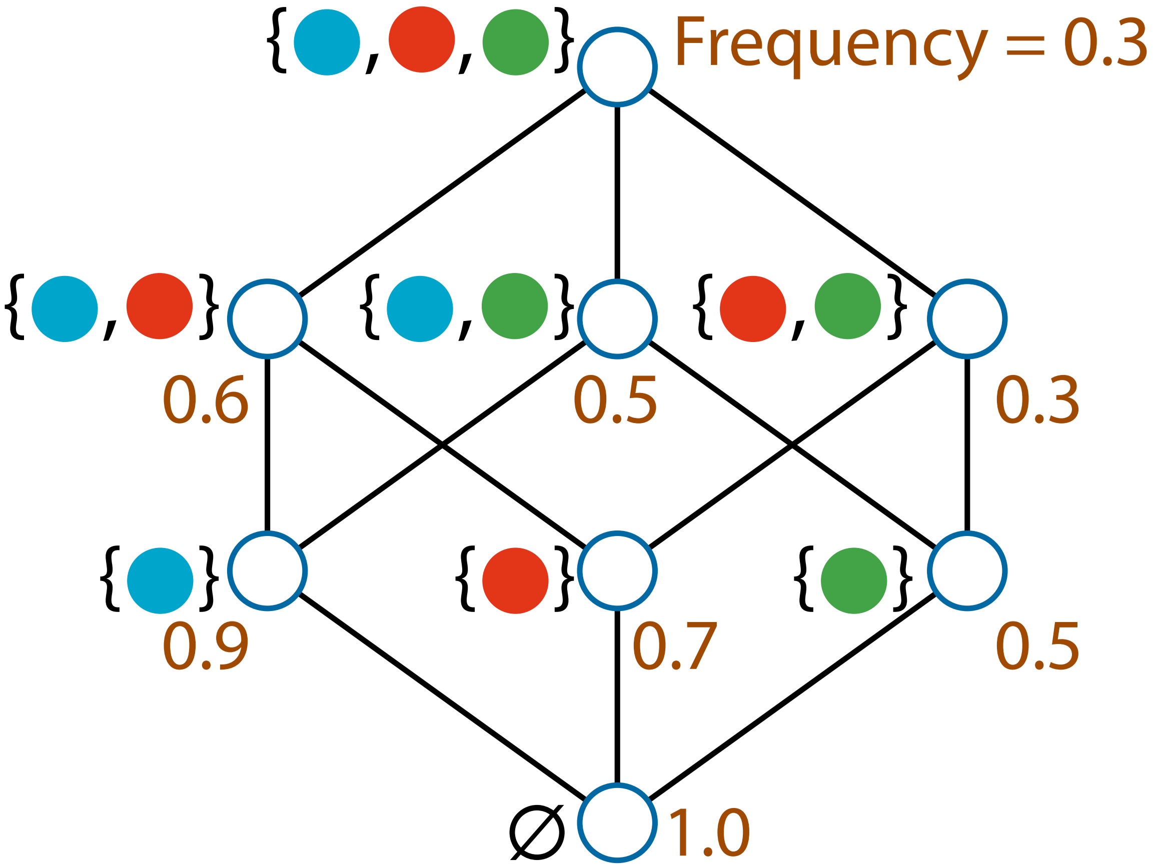
Poset (itemset lattice)



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

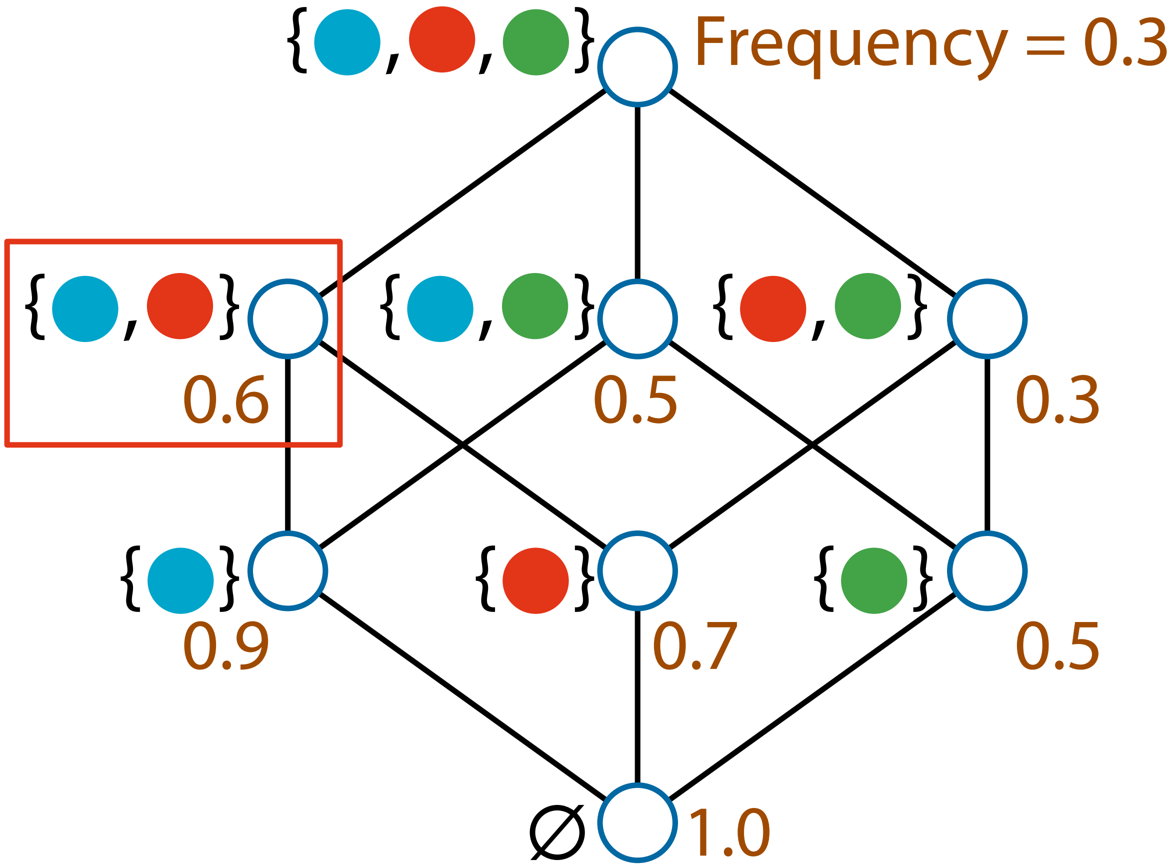
Poset (itemset lattice)



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

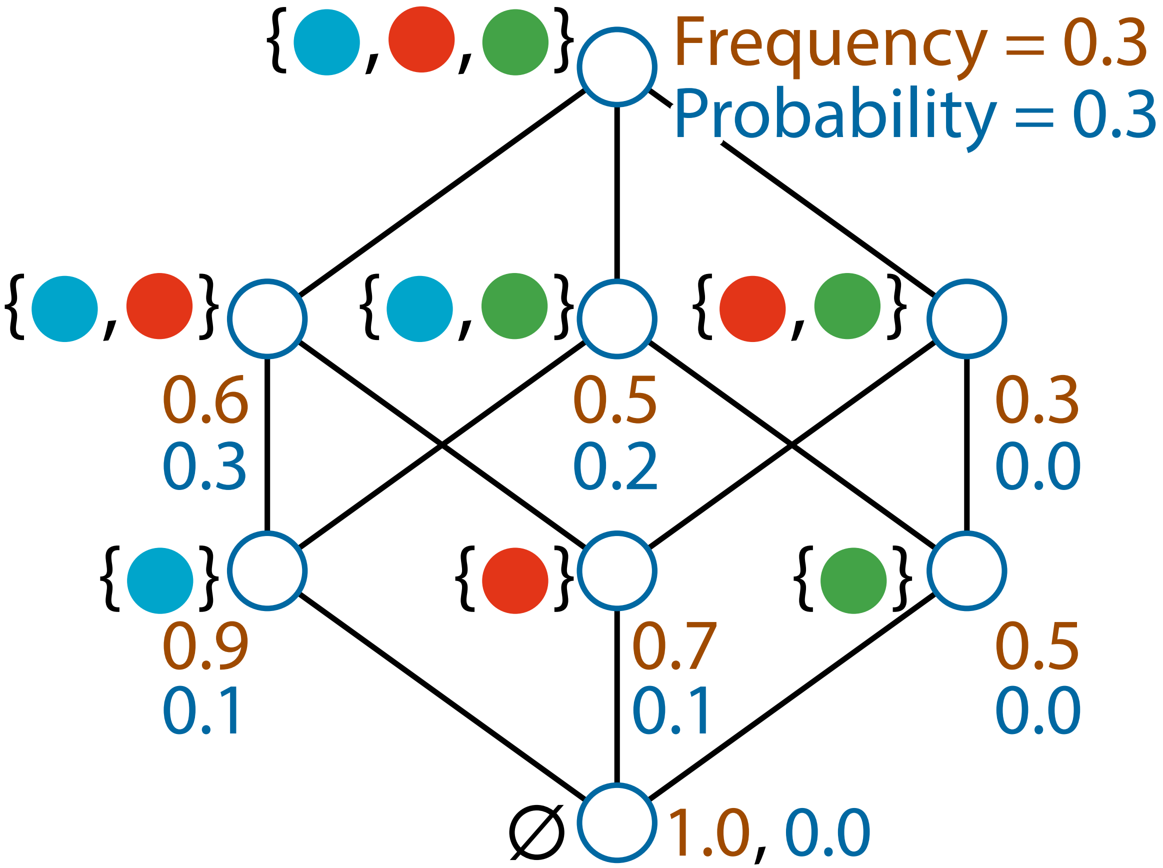
Poset (itemset lattice)



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)



{🔵,🔴,🟢} Frequency = 0.3
Probability = 0.3

{🔵,🔴} 0.6 0.3

{🔵,🟢} 0.5 0.2

{🔴,🟢} 0.3 0.0

{🔵} 0.9 0.1

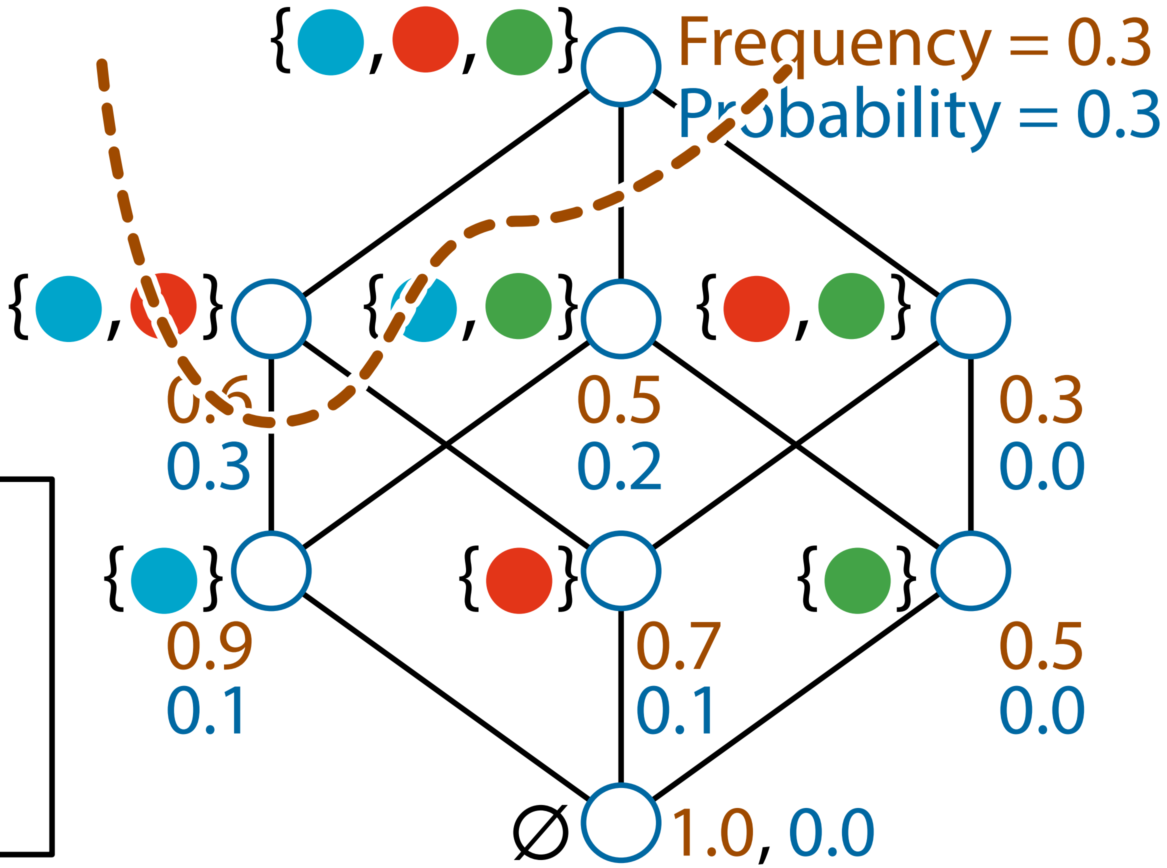
{🔴} 0.7 0.1

{🟢} 0.5 0.0

∅ 1.0, 0.0

Upward = *Pattern mining*

Poset (itemset lattice)



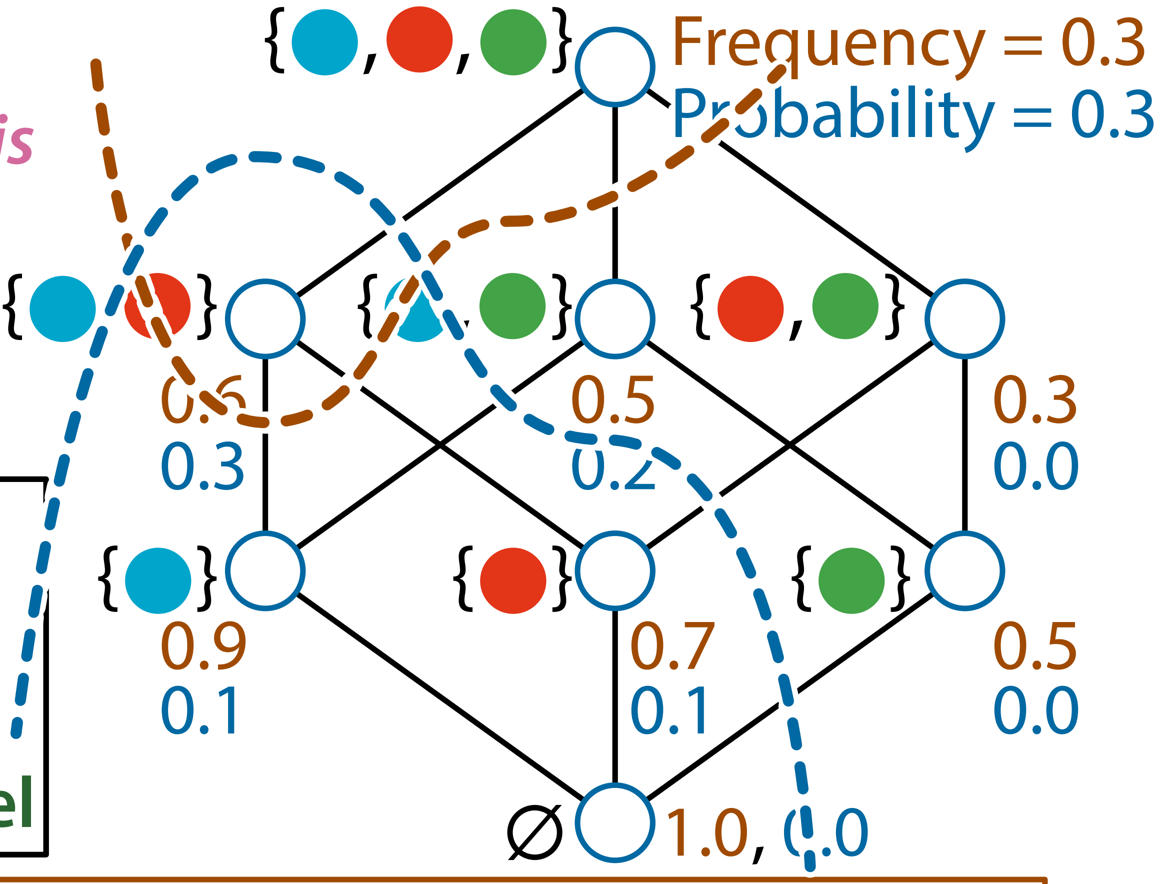
η: Frequency

p: Probability

η({🔵,🔴}) = *p*({🔵,🔴}) + *p*({🔵,🔴,🟢})

Upward =
Pattern mining
Downward =
Log-linear analysis

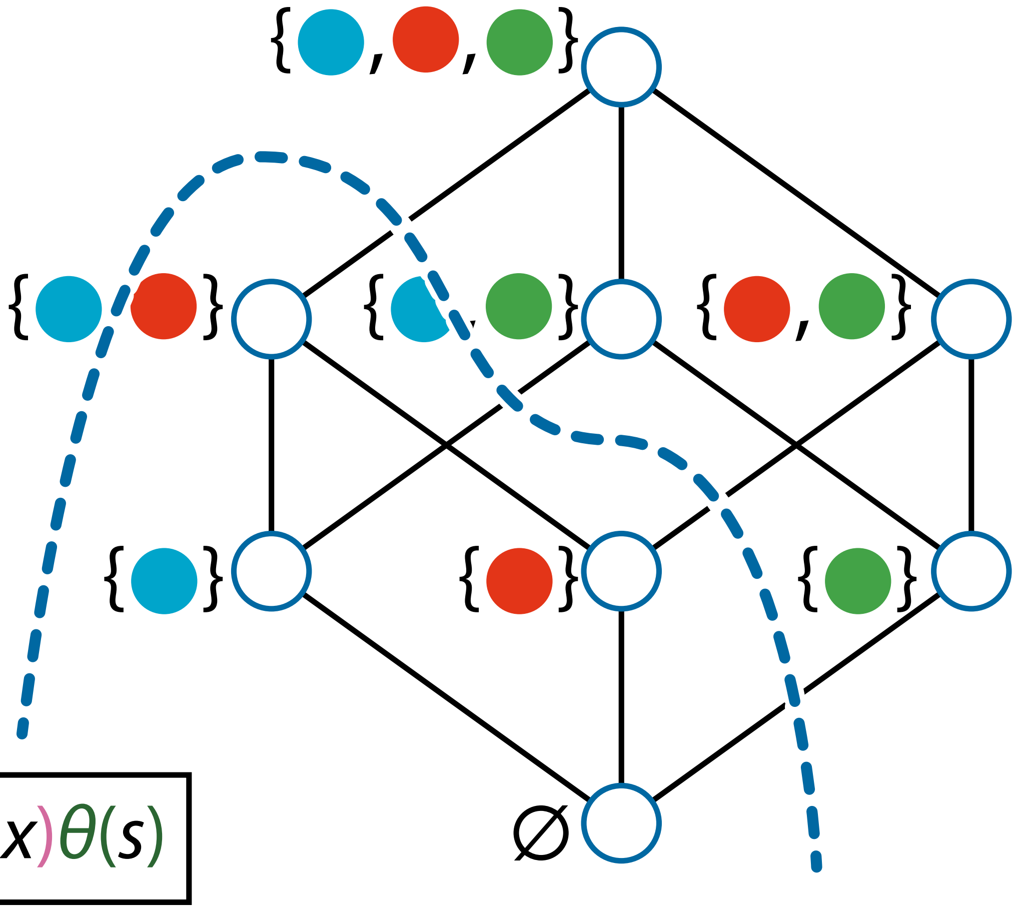
Poset (itemset lattice)



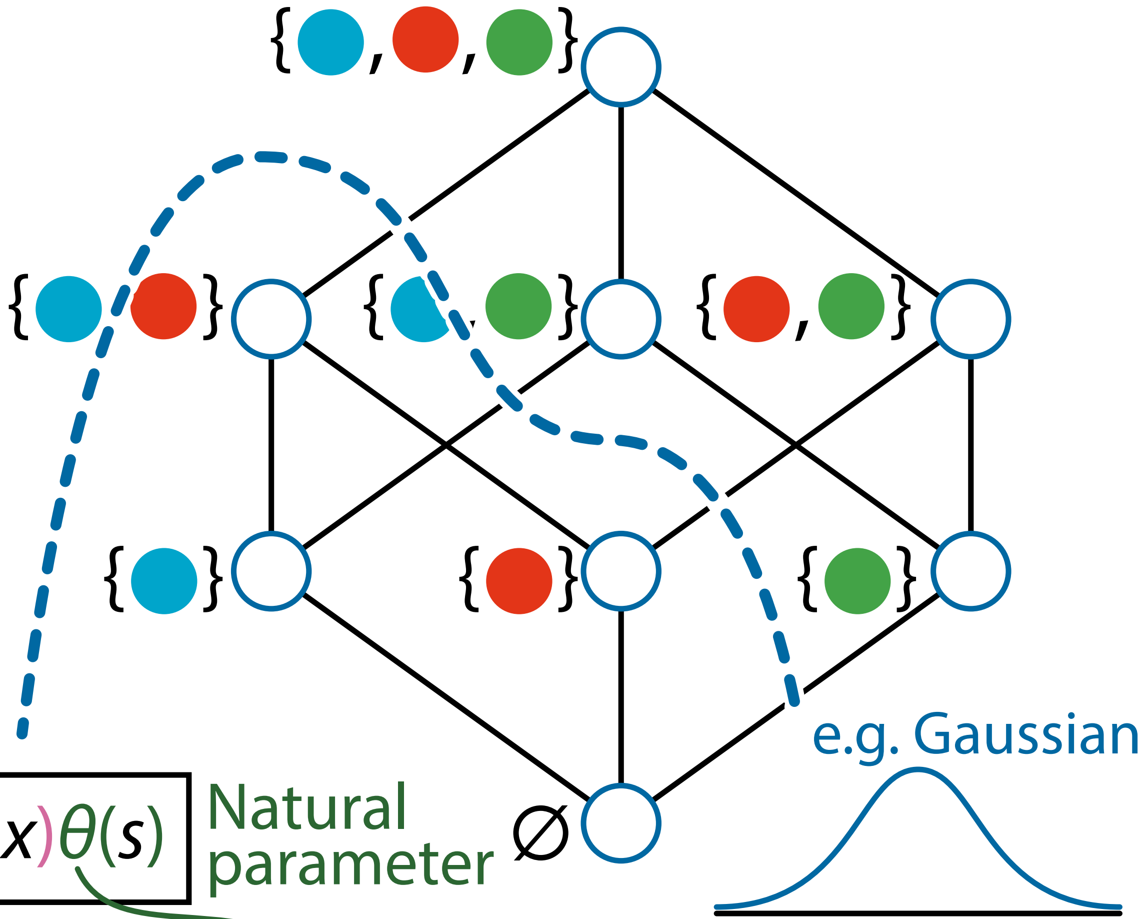
η: Frequency
p: Probability
θ: Coefficient of
log-linear model

η({🔵,🔴}) = p({🔵,🔴}) + p({🔵,🔴,🟢})

log p({🔵,🔴}) = θ({🔵,🔴}) + θ({🔵}) + θ({🔴}) + $\theta(\varnothing)$



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



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

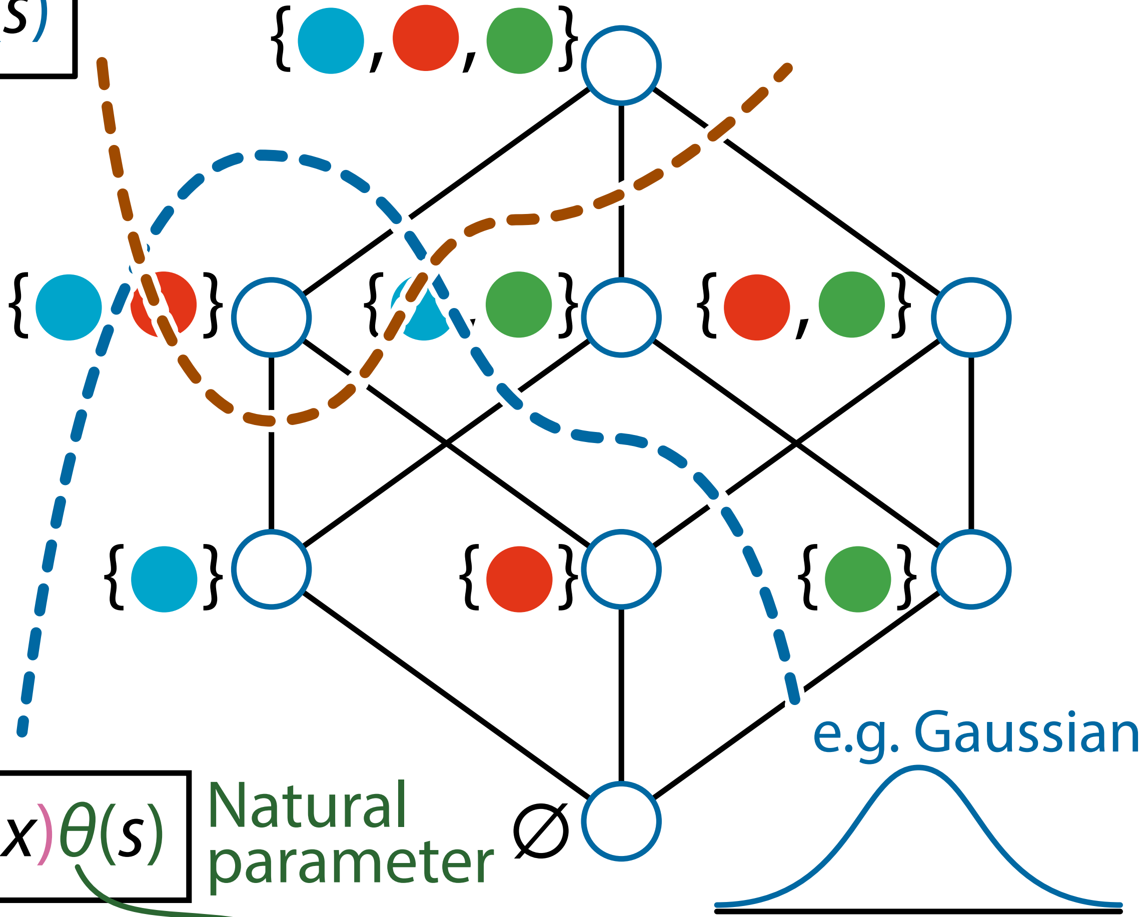
Natural parameter

Exponential family: $$p(x) = \exp\Big(\sum \theta(s) F_s(x) - \psi(\theta)\Big)$$

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

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

Sufficient statistics of exponential family



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

Natural parameter

$$\boxed{\text{Exponential family:}\quad p(x) = \exp\Big(\sum \theta(s)F_s(x) - \psi(\theta)\Big)}$$

Möbius Inversion on Posets

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

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

- Möbius function $\mu : S \times S \to \mathbb{Z}$, 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 $A_1, A_2, \ldots, A_n$,

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

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

Mathematical Formulation

- Log-linear model and its sufficient statistics:

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

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

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

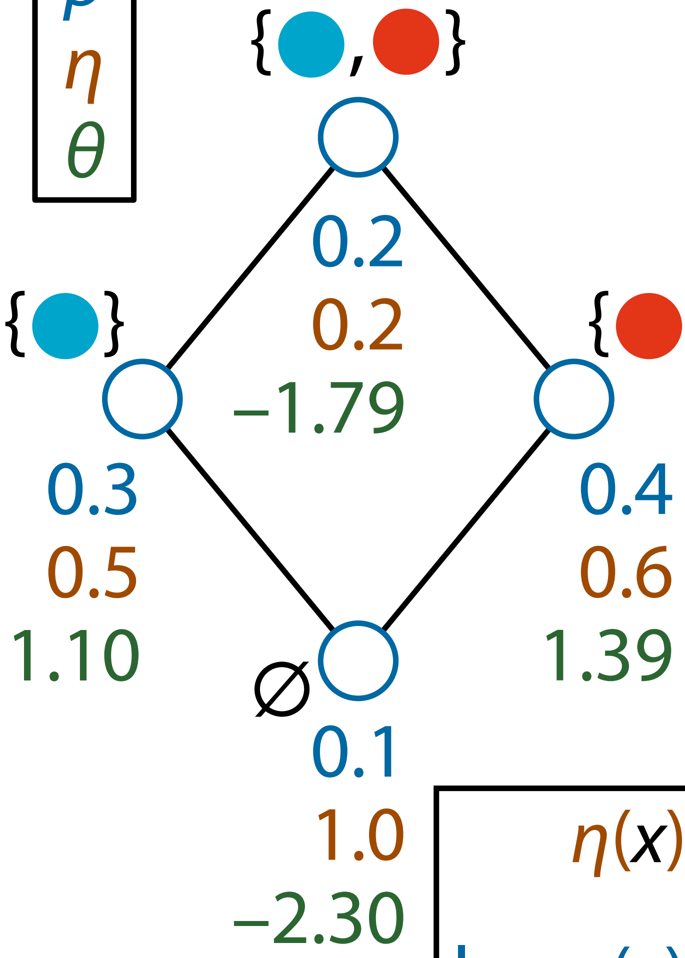
$$\log p(\boldsymbol{x}) = \sum_{i} \theta^{i} x^{i} + \sum_{i<j} \theta^{ij} x^{i} x^{j} + \cdots + \theta^{1\ldots n} x^{1} x^{2} \ldots x^{n},$$

- 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

p
η
θ

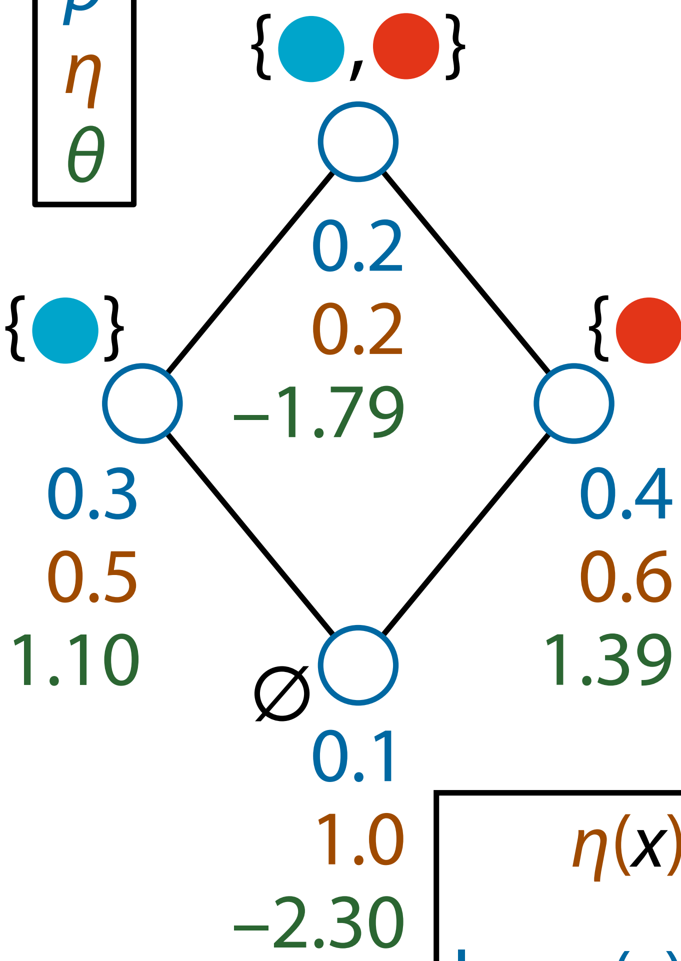


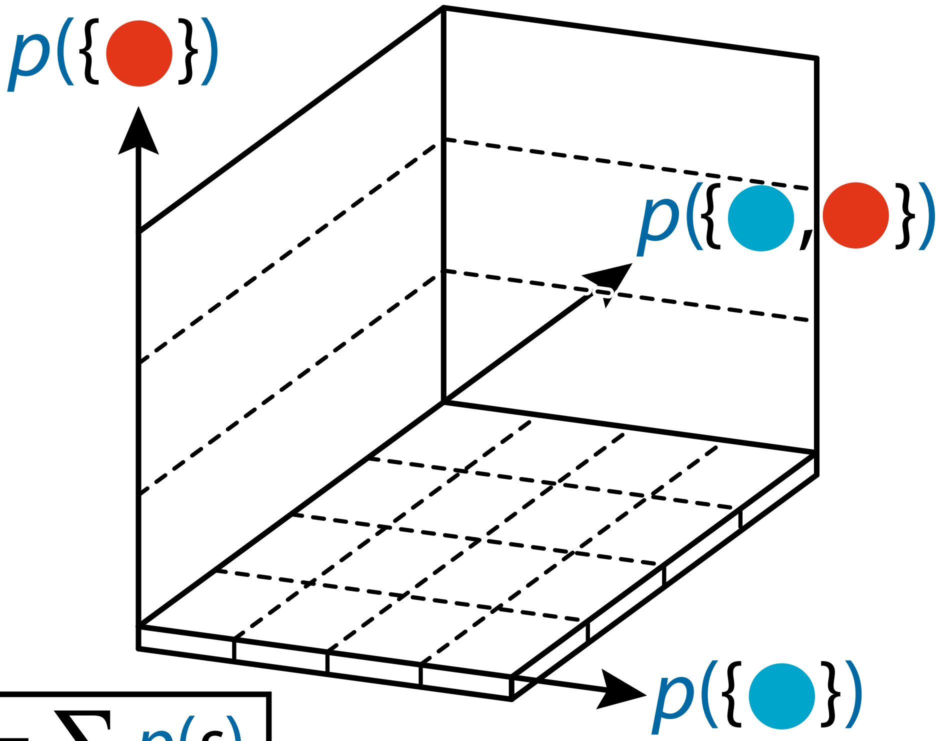
$$\eta(x) = \sum_{s \geq x} p(s)$$

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

Triple for each node

p
η
θ

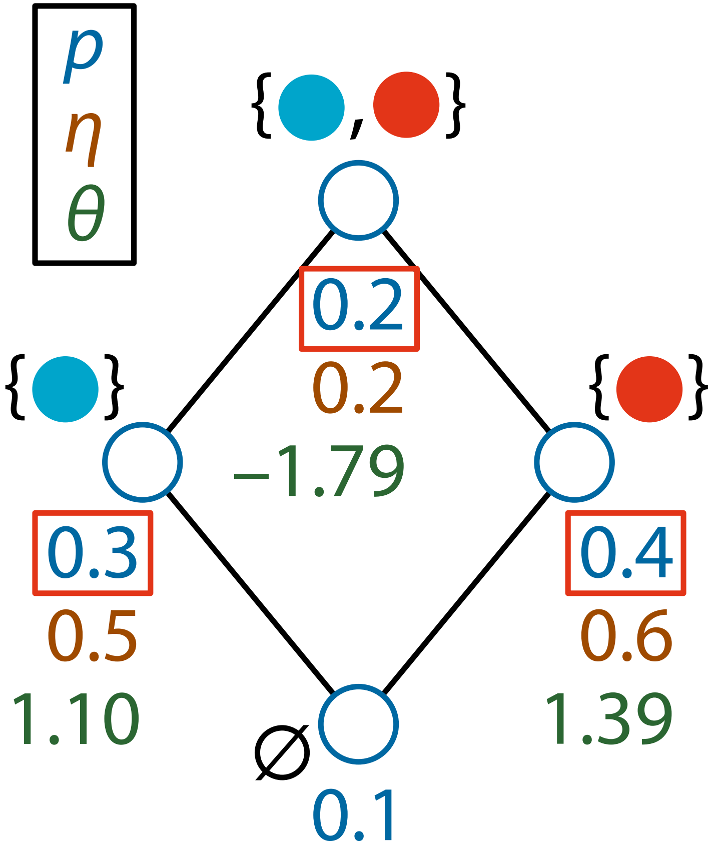




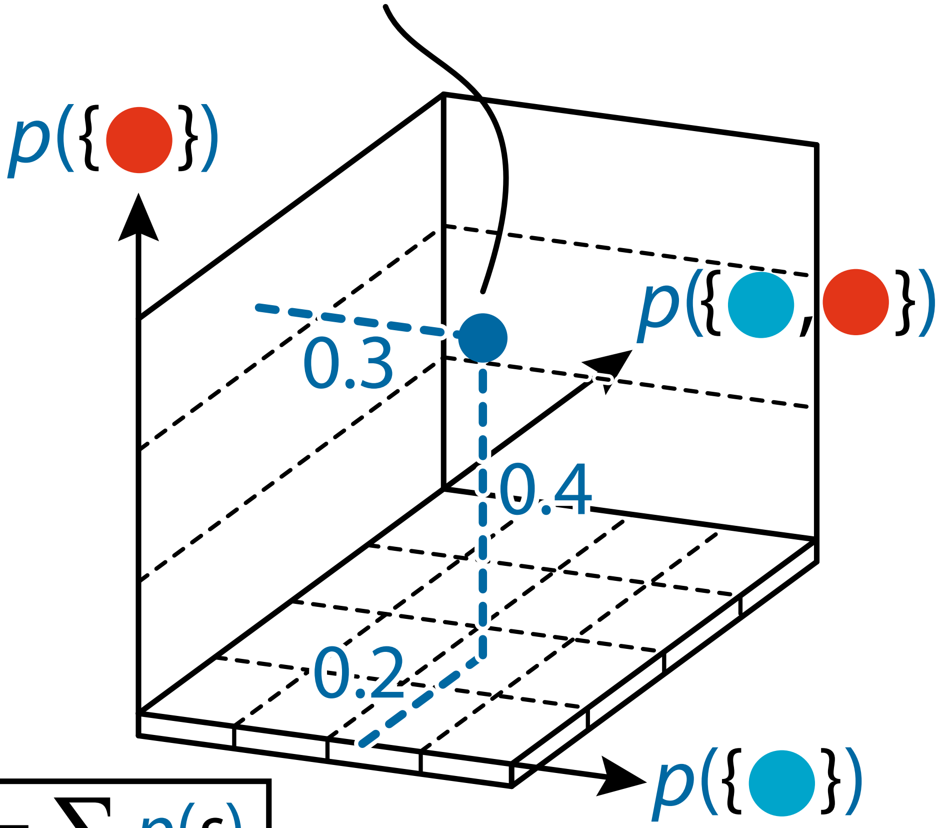
$$\eta(x) = \sum_{s \geq x} p(s)$$

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

Triple for each node



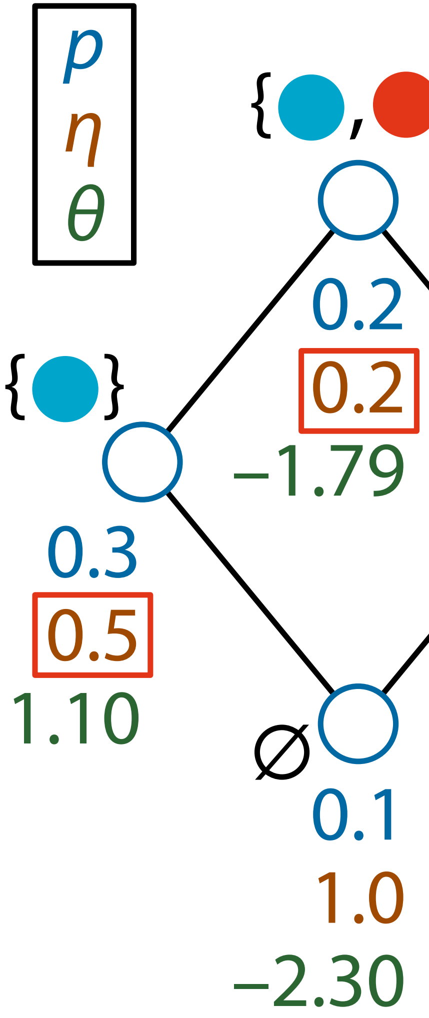
Probability distribution is a “point” in 3D space



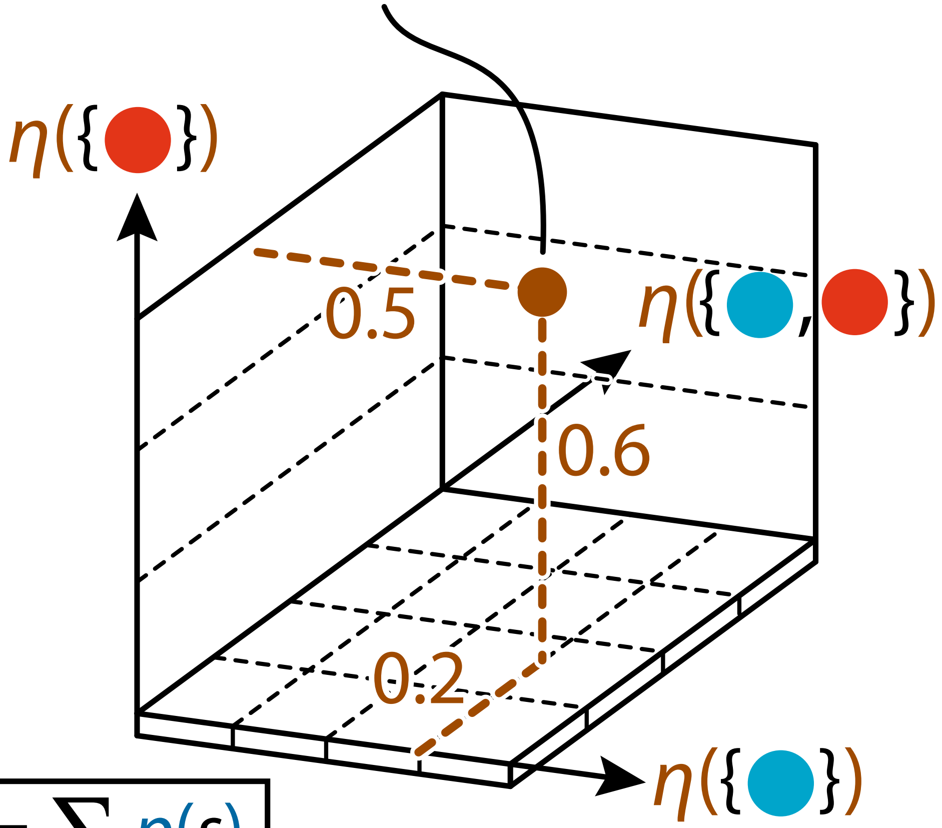
$$\eta(x) = \sum_{s \geq x} p(s)$$

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

Triple for each node



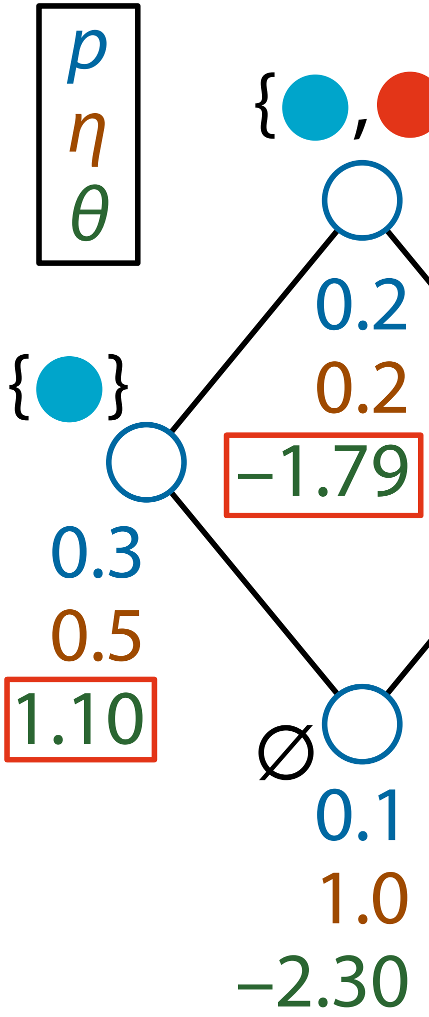
Probability distribution is a “point” in 3D space



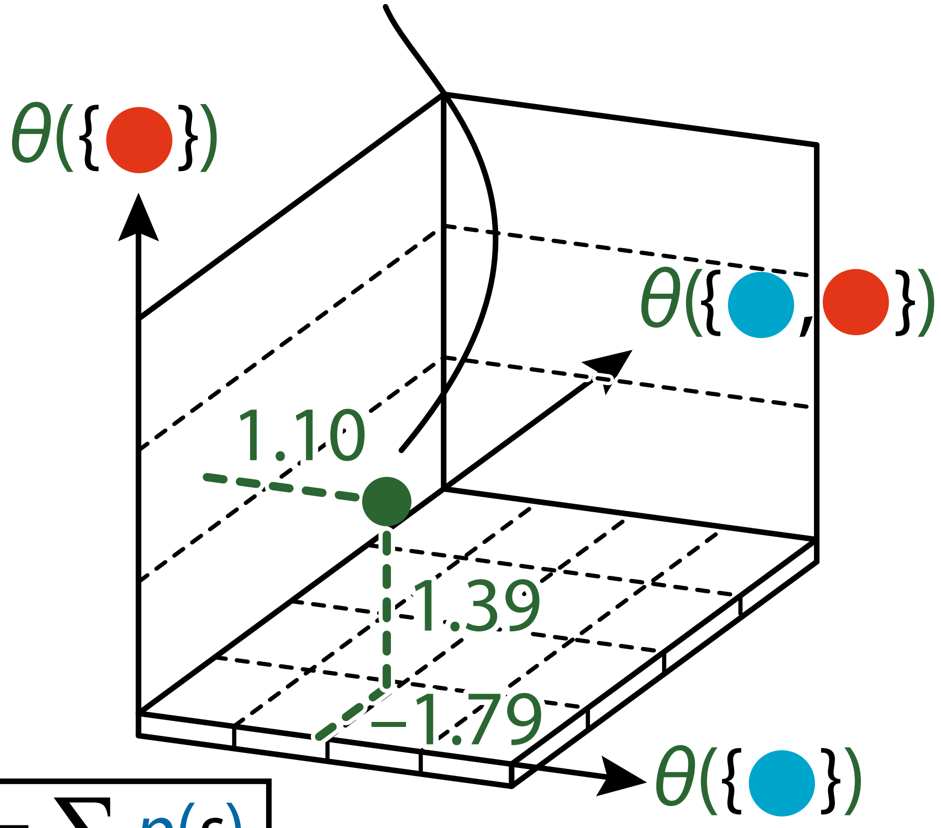
$$\eta(x) = \sum_{s \geq x} p(s)$$

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

Triple for each node



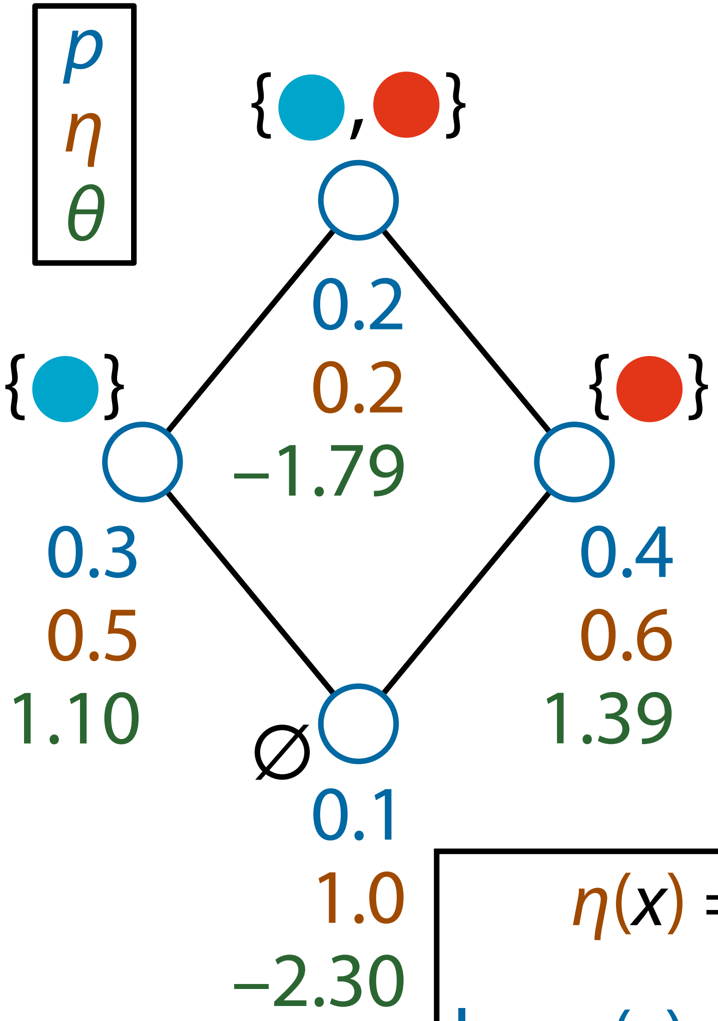
Probability distribution is a “point” in 3D space

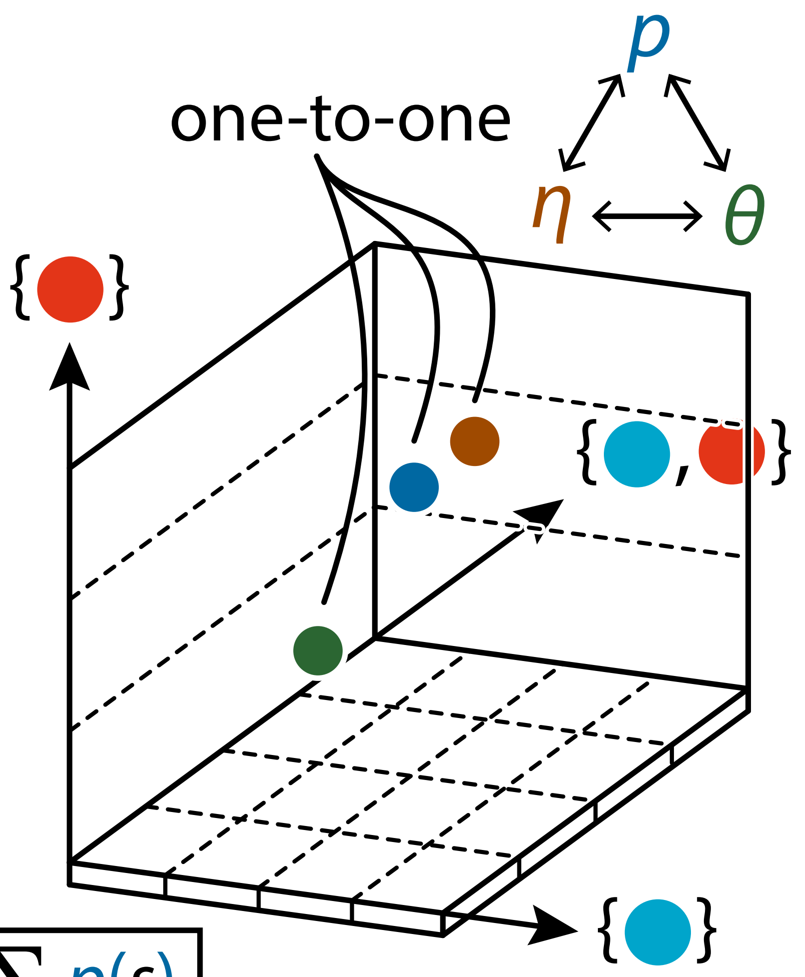


$$\eta(x) = \sum_{s \geq x} p(s)$$

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

Triple for each node

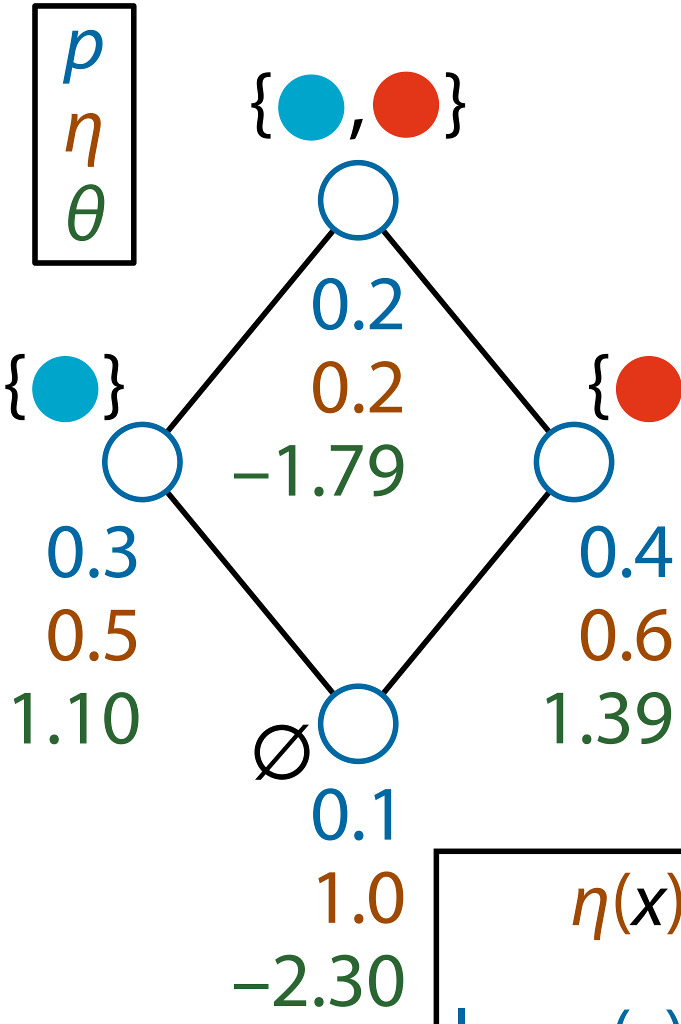


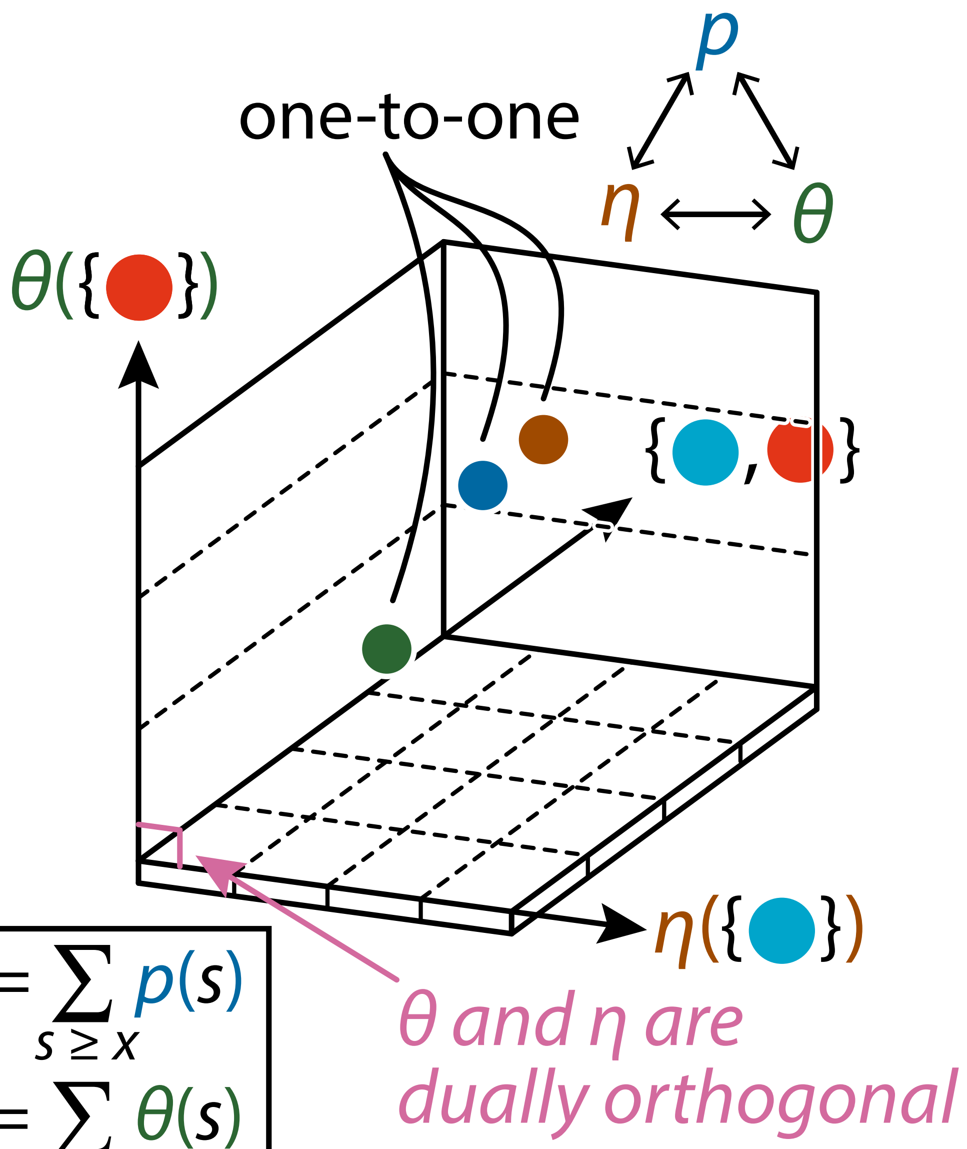


$$\eta(x) = \sum_{s \geq x} p(s)$$

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

Triple for each node





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

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

Orthogonality of θ and η

- From Möbius inversion,

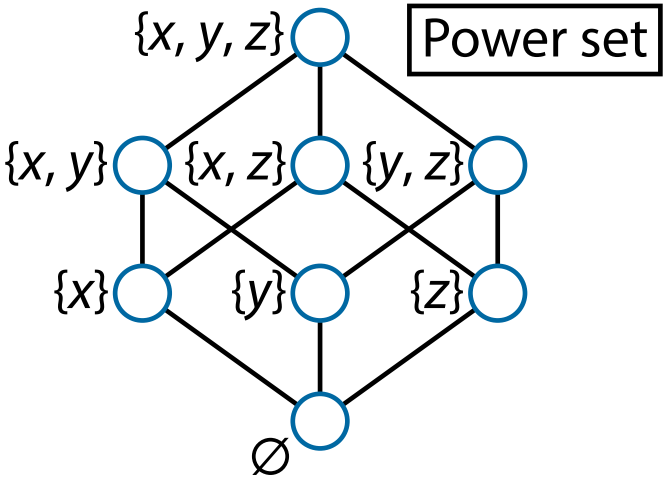
$$\sum_{s \in S} \zeta(x, s)\mu(s, y) = \delta_{x,y}, \quad \delta_{x,y} = \begin{cases} 1 & \text{if } x = y, \\ 0 & \text{otherwise} \end{cases}$$

- θ and η are dually orthogonal:

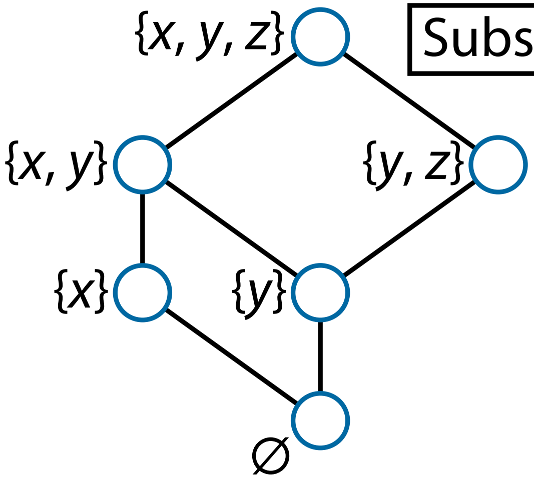
$$\mathrm{E}\left[\frac{\partial}{\partial\theta(x)} \log p(s) \frac{\partial}{\partial\eta(y)} \log p(s)\right] = \sum_{s \in S} \zeta(x, s)\mu(s, y) = \delta_{x,y}$$

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

Existing Approach Limited To Power Set

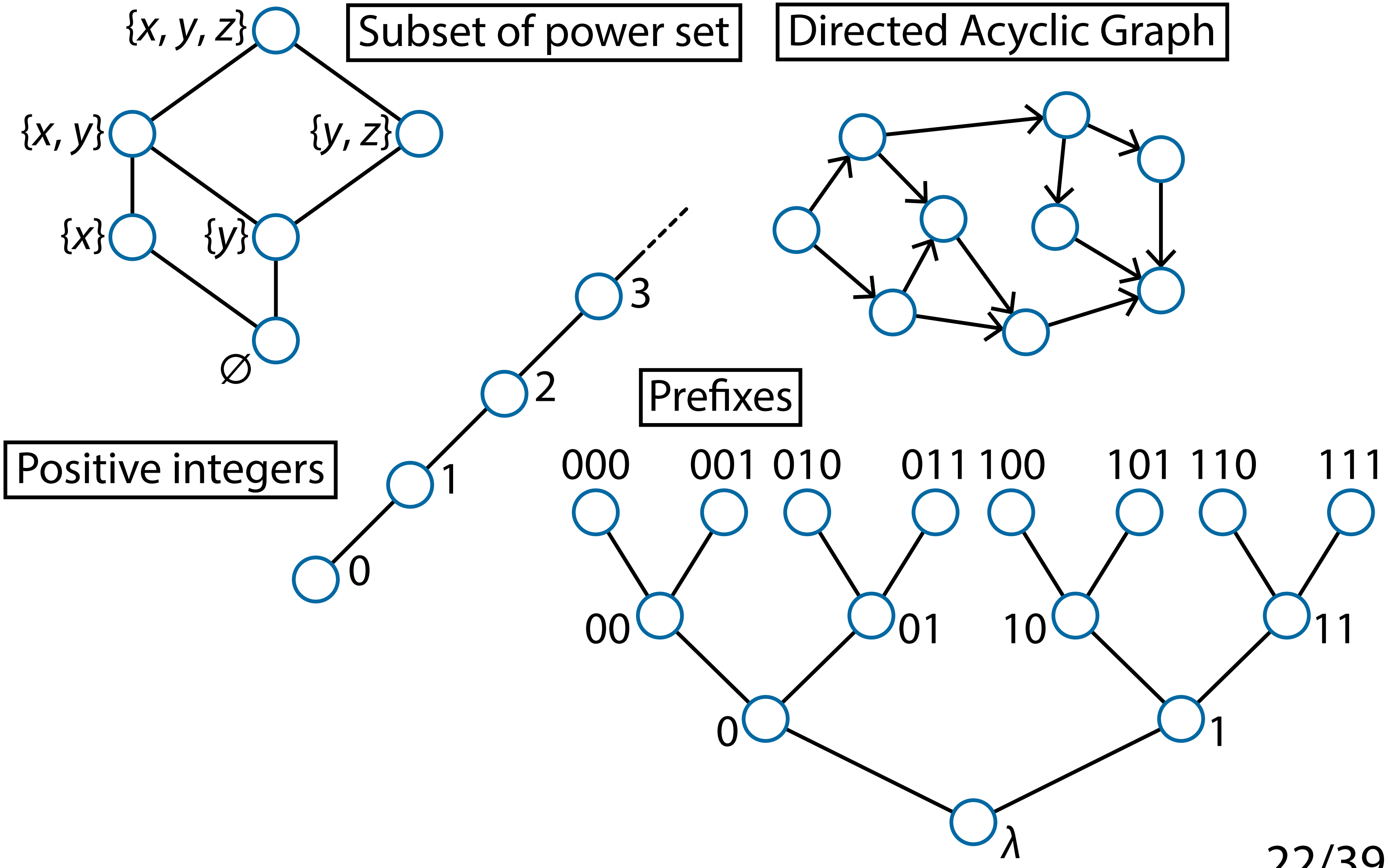


Our Approach Applies To Any Posets



Subset of power set

Our Approach Applies To Any Posets



Subset of power set

Directed Acyclic Graph

Positive integers

Prefixes

KL Divergence Decomposition

- KL divergence decomposition:

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

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

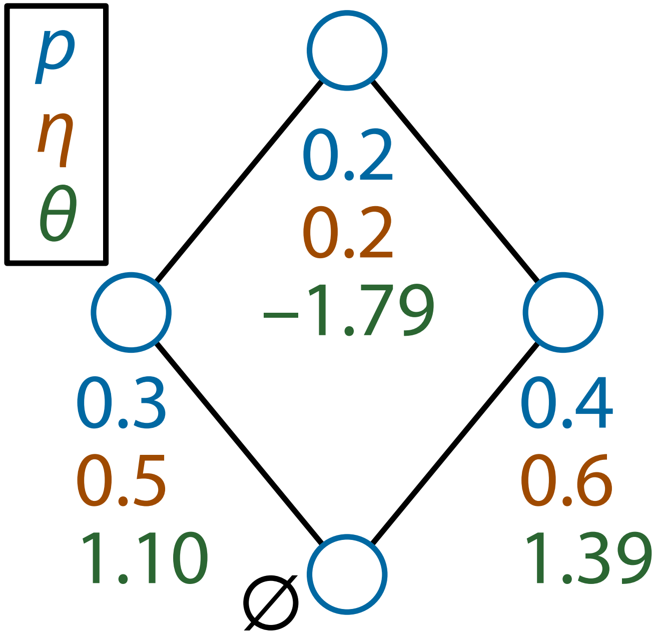
  - Q is called the mixed distribution of (P, R)
  - It is known as the (generalized) Pythagoras theorem in Information Geometry

- We can derive from Möbius inversion:

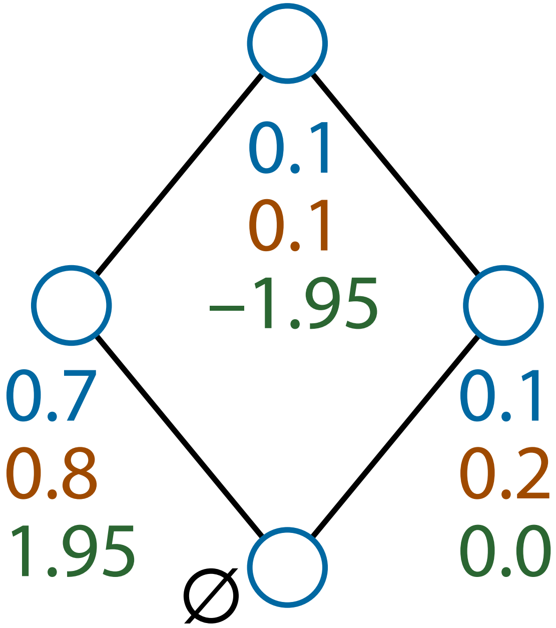
$$D_{\mathrm{KL}}[P, Q] + D_{\mathrm{KL}}[Q, R] - D_{\mathrm{KL}}[P, R]$$

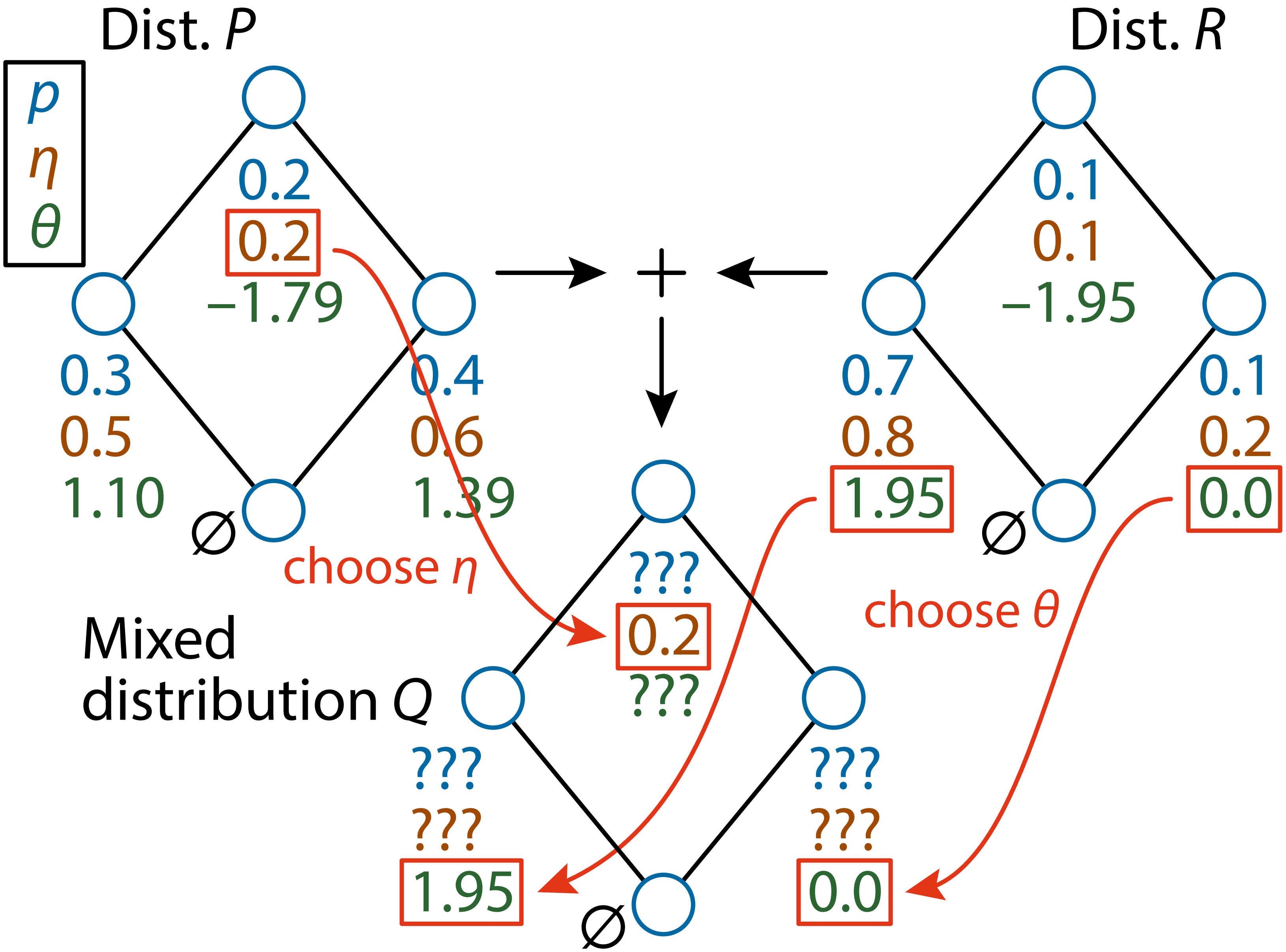
$$= \sum_{s \in S} (\eta_Q(s) - \eta_P(s)) (\theta_Q(s) - \theta_R(s))$$

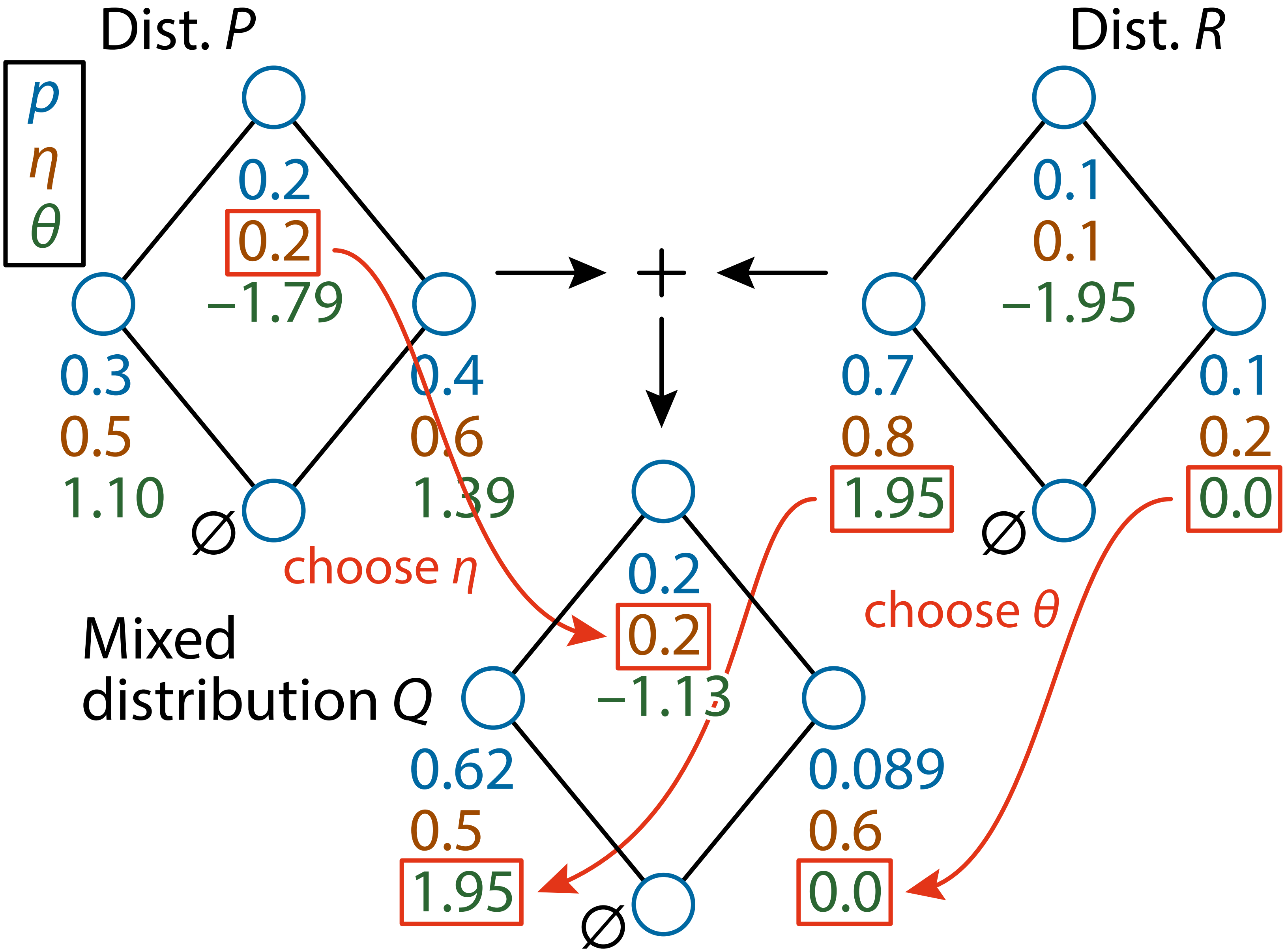
Dist. *P*

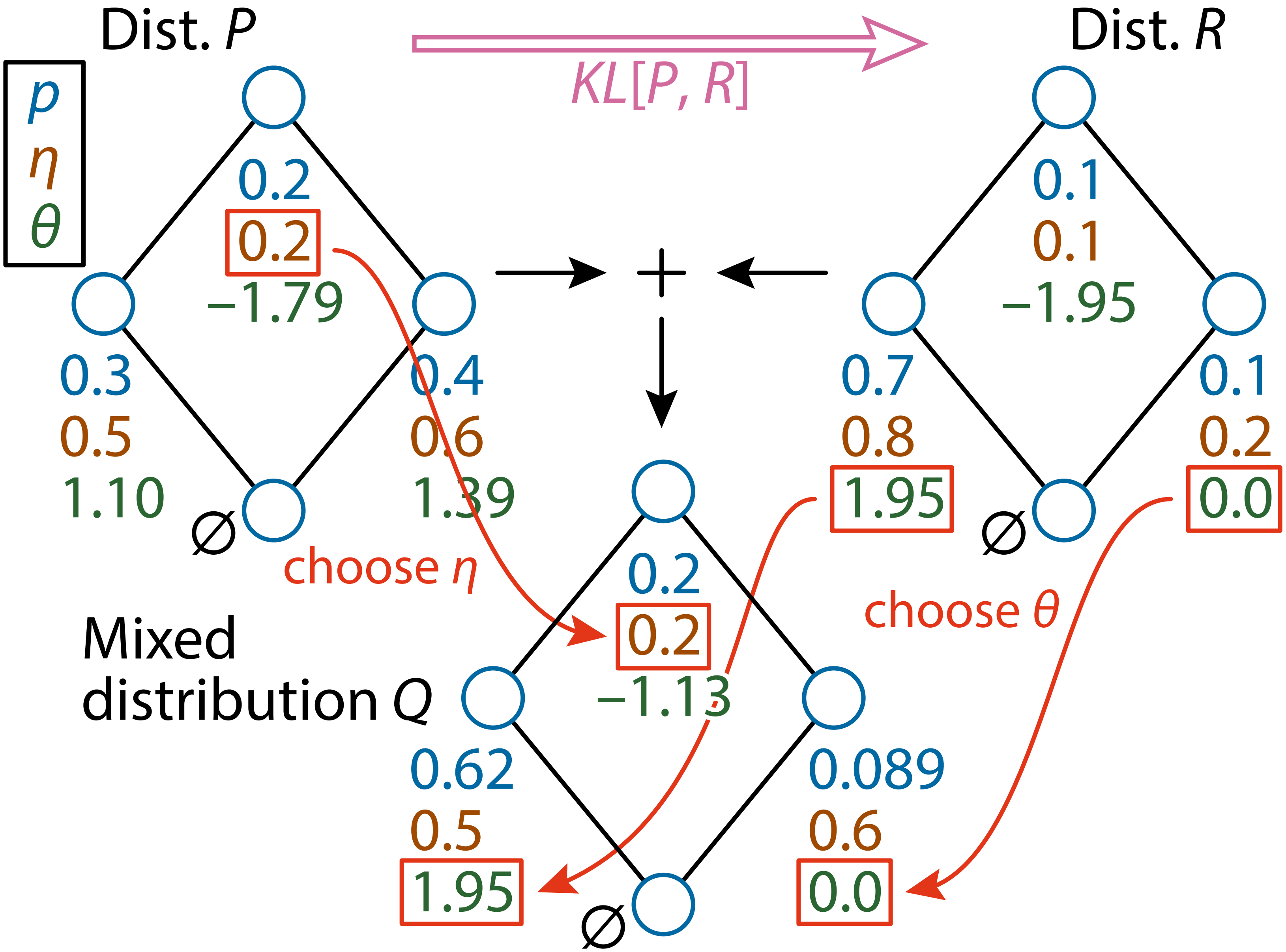

p
η
θ
0.2
0.2
−1.79
0.3
0.5
1.10
0.4
0.6
1.39
∅

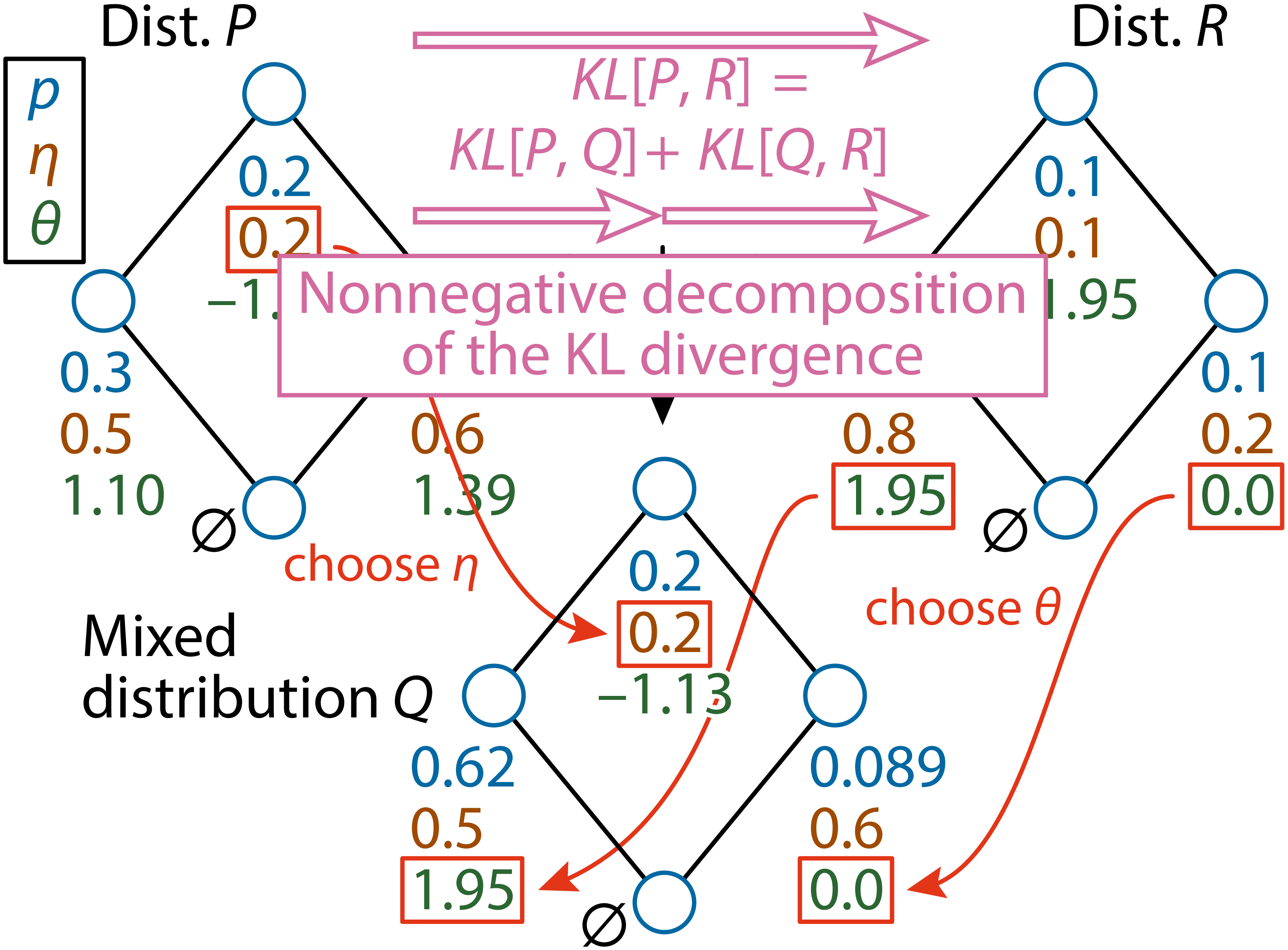
Dist. *R*


0.1
0.1
−1.95
0.7
0.8
1.95
0.1
0.2
0.0
∅


Dist. P
p
η
θ
0.2
0.2
−1.79
0.3
0.5
1.10
0.4
0.6
1.39
∅
choose η
Dist. R
0.1
0.1
−1.95
0.7
0.8
1.95
0.1
0.2
0.0
∅
choose θ
+
Mixed distribution Q
???
0.2
???
???
???
1.95
???
???
0.0
∅

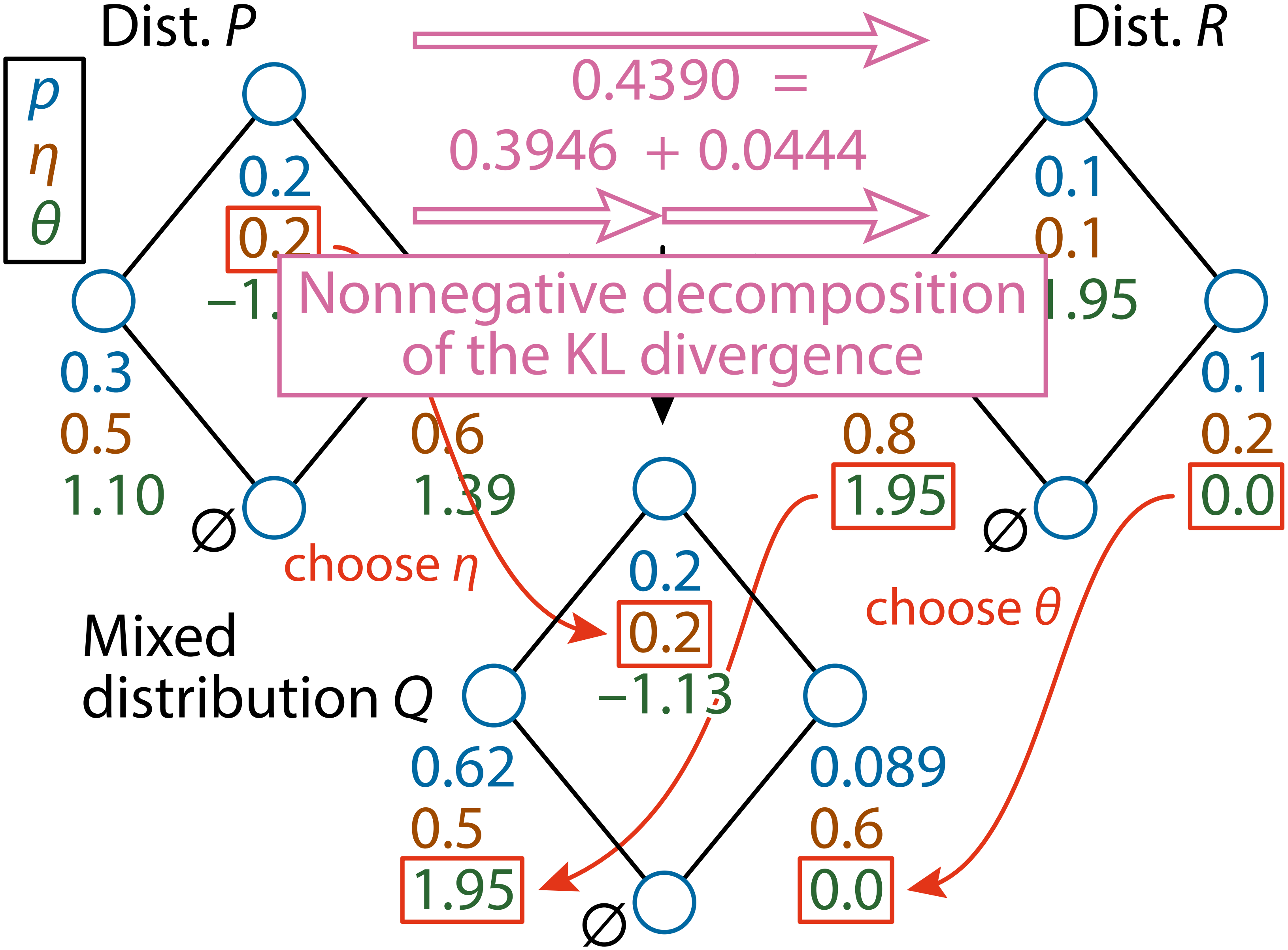

Dist. *P*
p
η
θ
0.2
0.2
−1.79
0.3
0.5
1.10
0.4
0.6
1.39
∅
choose *η*
Dist. *R*
0.1
0.1
−1.95
0.7
0.8
1.95
0.1
0.2
0.0
∅
choose *θ*
+
Mixed distribution *Q*
0.2
0.2
−1.13
0.62
0.5
1.95
0.089
0.6
0.0
∅


Dist. P
$KL[P, R]$
Dist. R
p
η
θ
0.2
0.2
−1.79
0.3
0.5
1.10
0.4
0.6
1.39
∅
choose η
0.1
0.1
−1.95
0.7
0.8
1.95
0.1
0.2
0.0
∅
choose θ
Mixed distribution Q
0.2
0.2
−1.13
0.62
0.5
1.95
0.089
0.6
0.0
∅



$KL[P, R] = KL[P, Q] + KL[Q, R]$

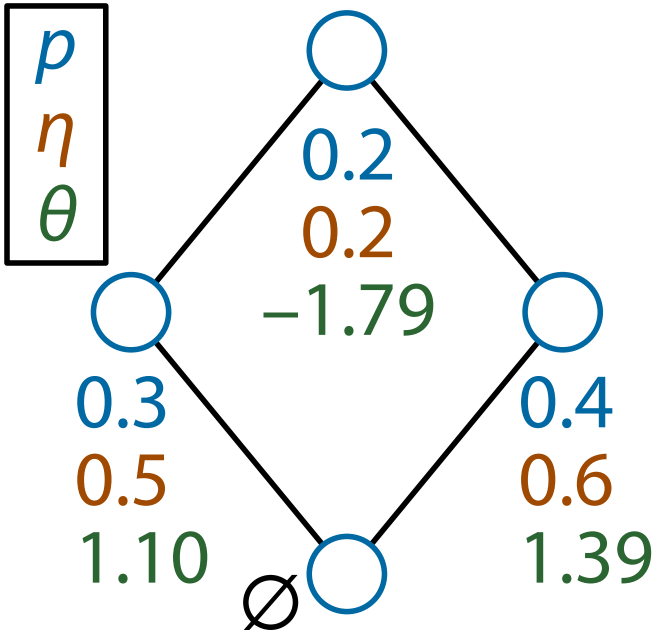
Nonnegative decomposition of the KL divergence



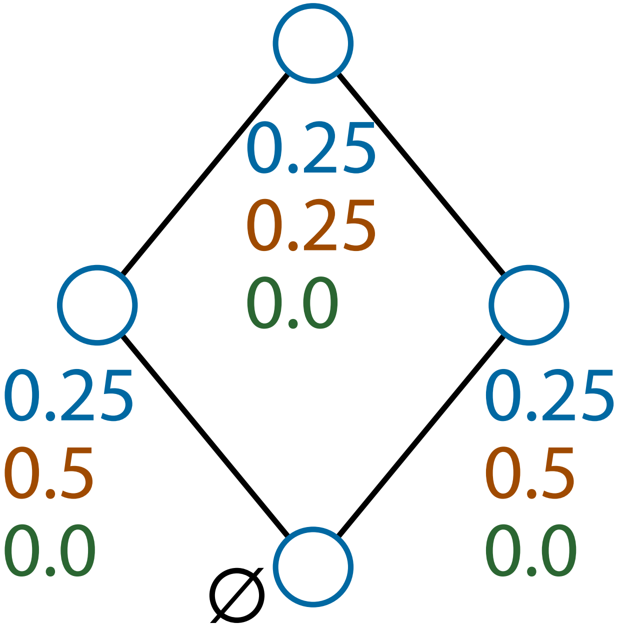
0.4390 = 0.3946 + 0.0444

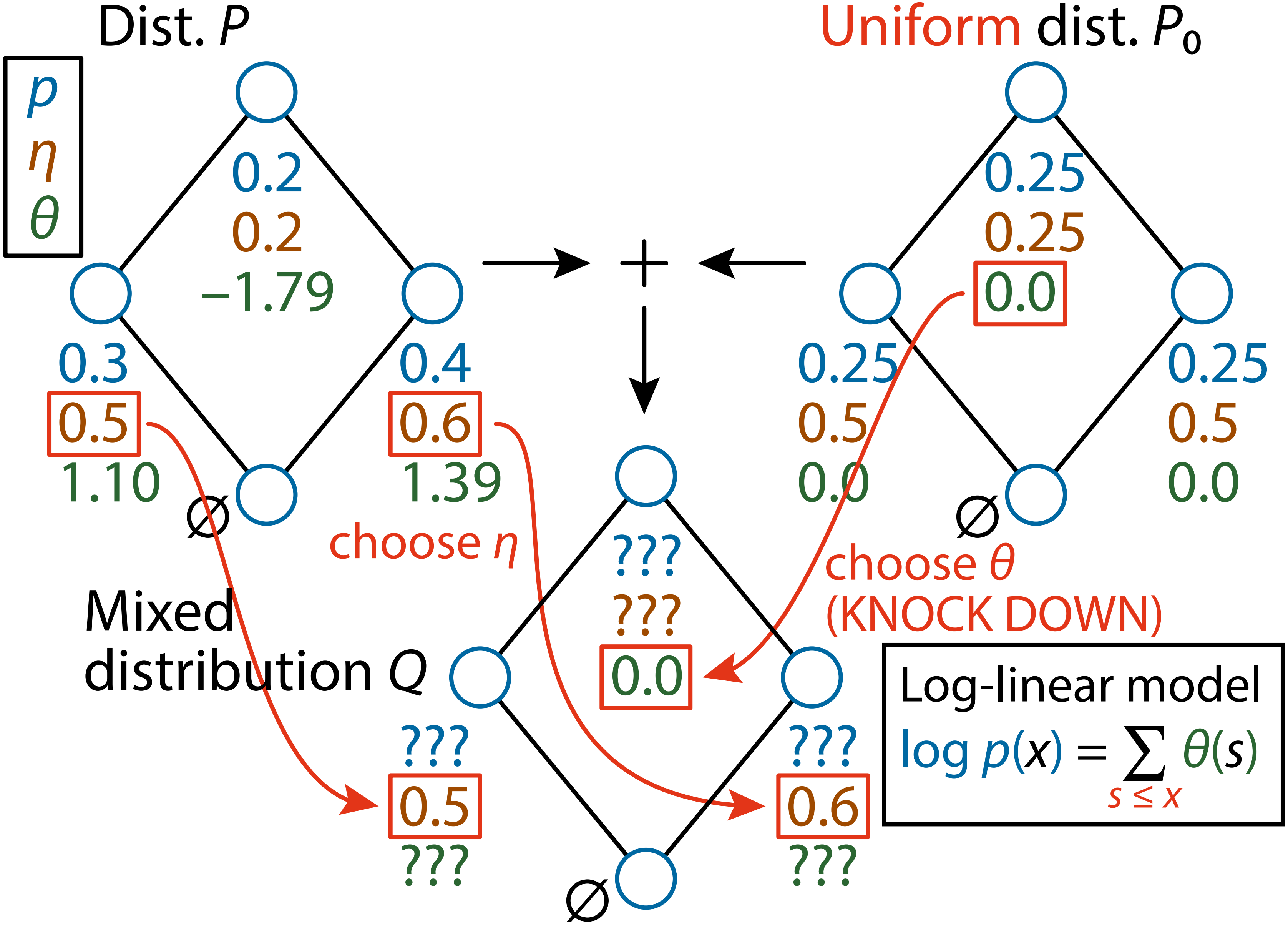
Nonnegative decomposition of the KL divergence

Dist. P



Uniform dist. P_0





Dist. P

Uniform dist. P_0

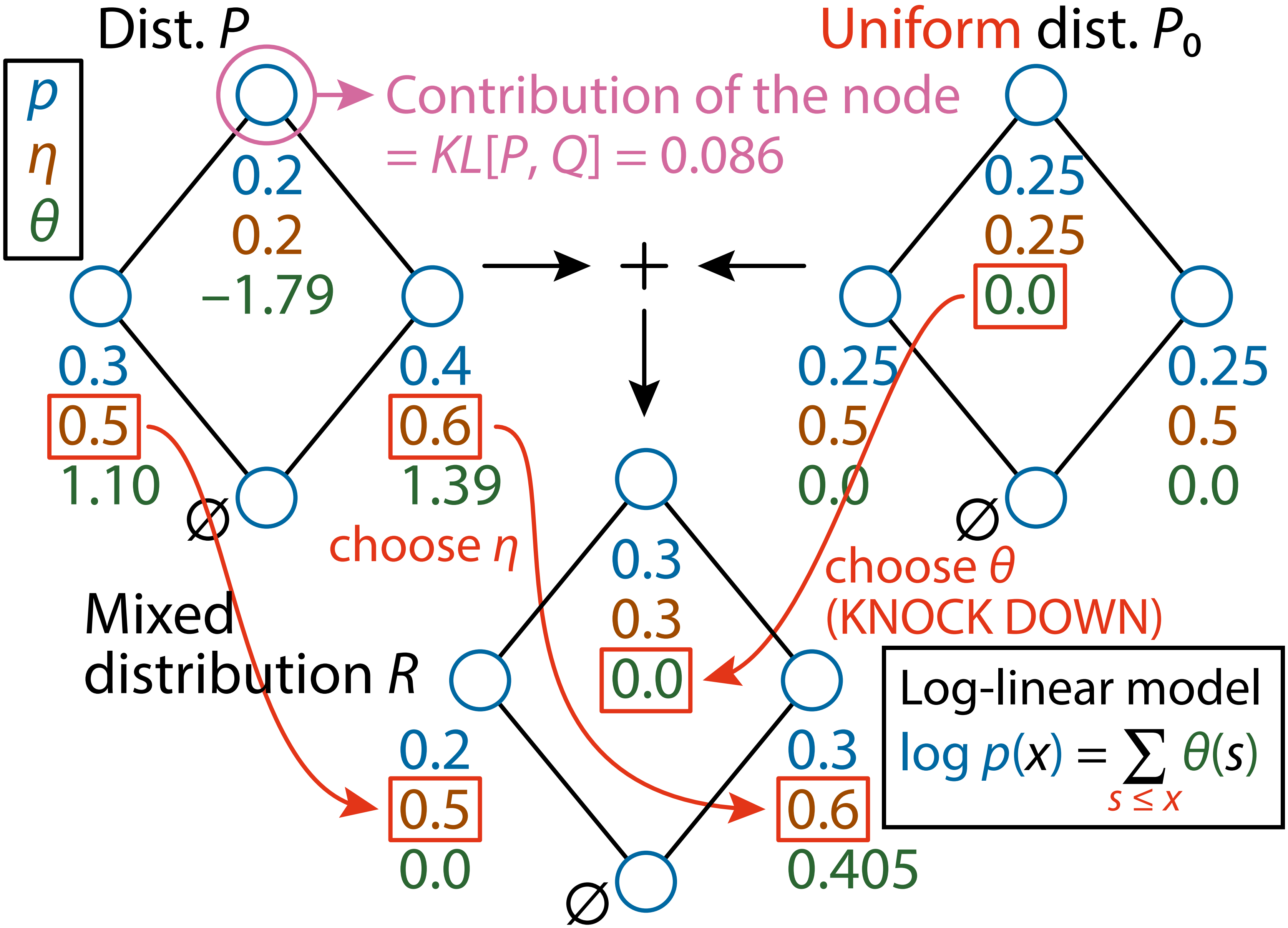
Mixed distribution Q

choose η

choose θ (KNOCK DOWN)

Log-linear model

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



Dist. P

Uniform dist. P_0

Contribution of the node $= KL[P, Q] = 0.086$

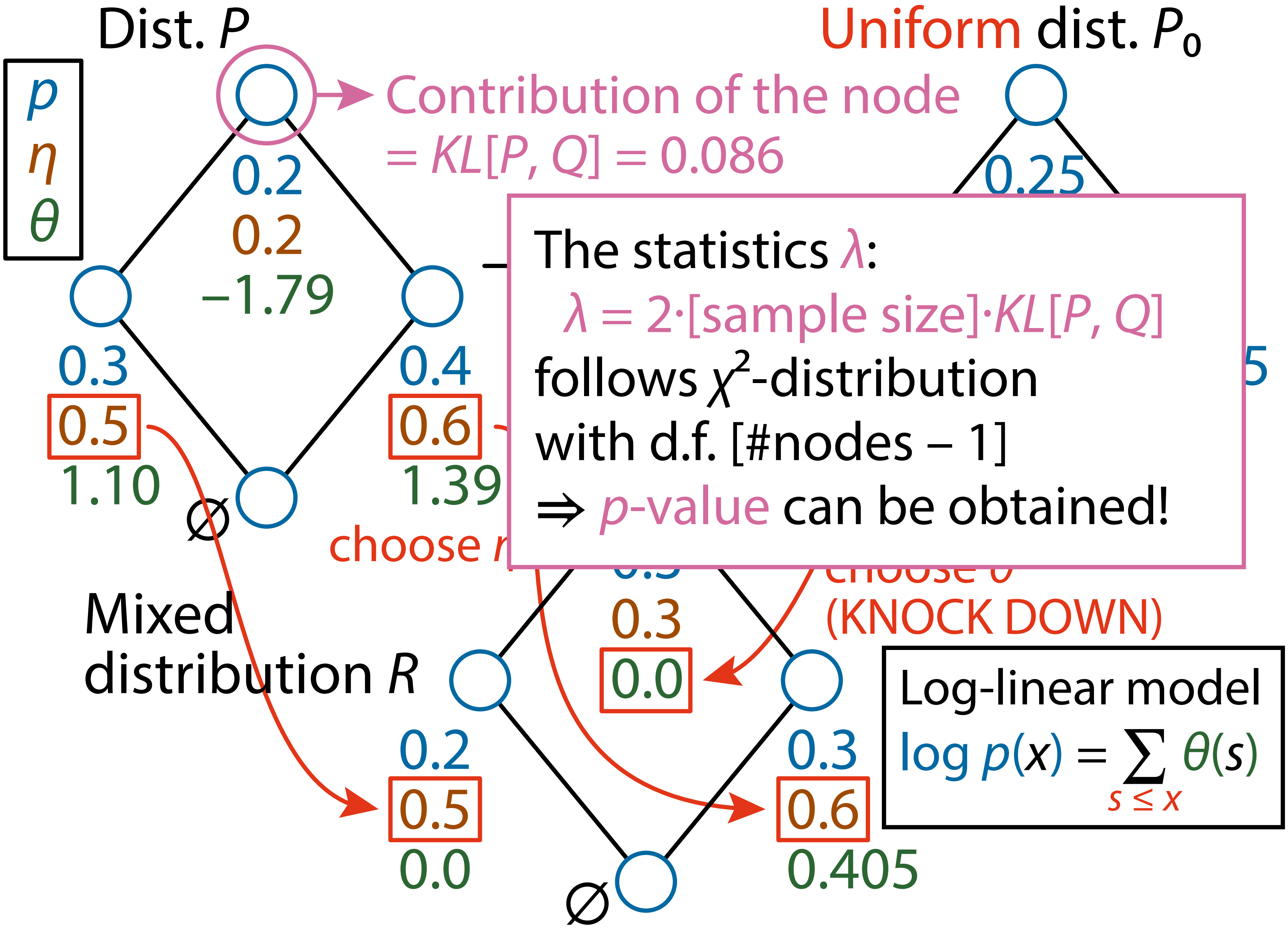
choose η

choose θ (KNOCK DOWN)

Mixed distribution R

Log-linear model

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



Contribution of the node $= KL[P, Q] = 0.086$

The statistics λ:

$\lambda = 2\cdot[\text{sample size}]\cdot KL[P, Q]$

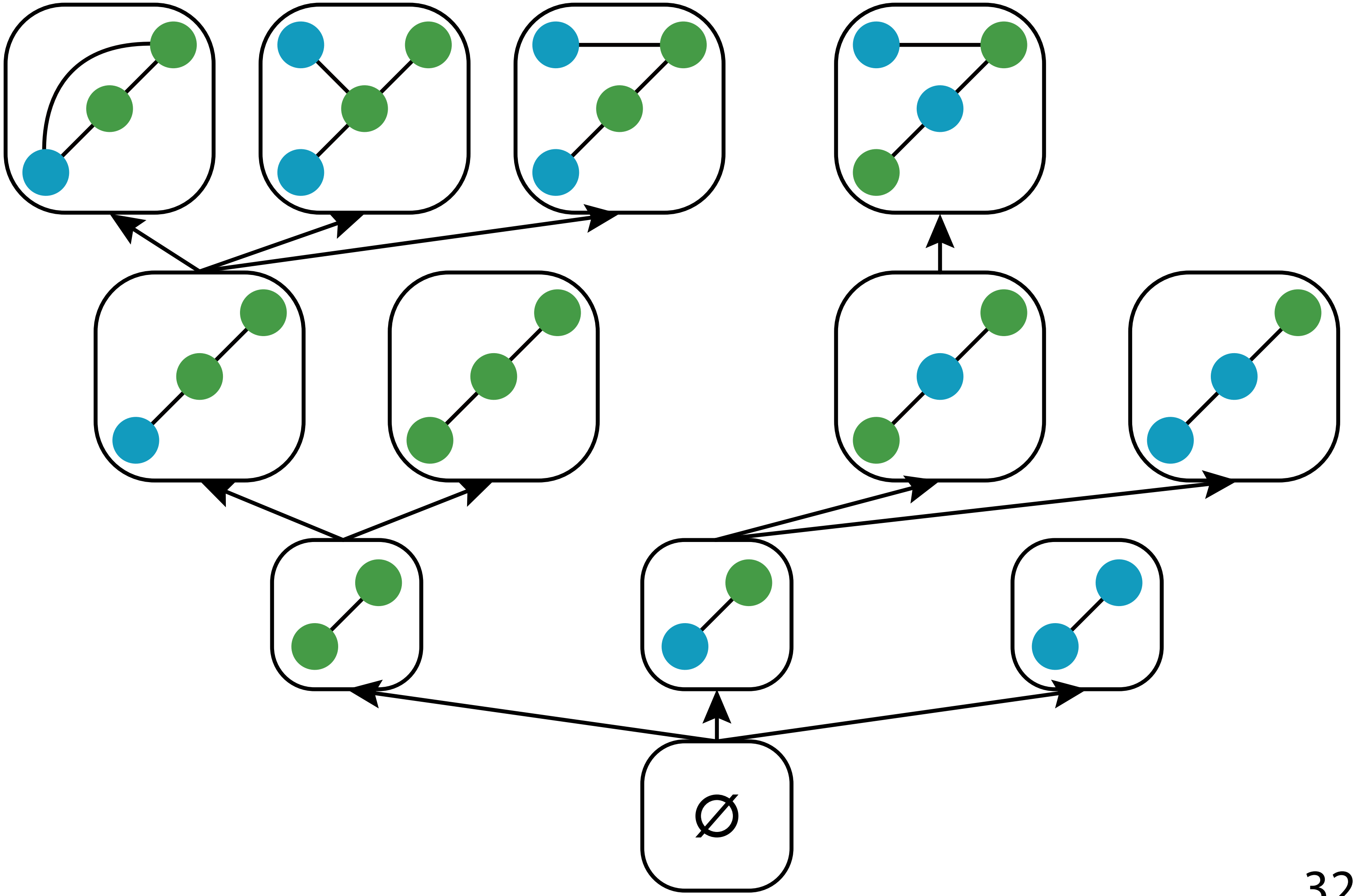
follows χ^2-distribution with d.f. [#nodes – 1]

⇒ p-value can be obtained!

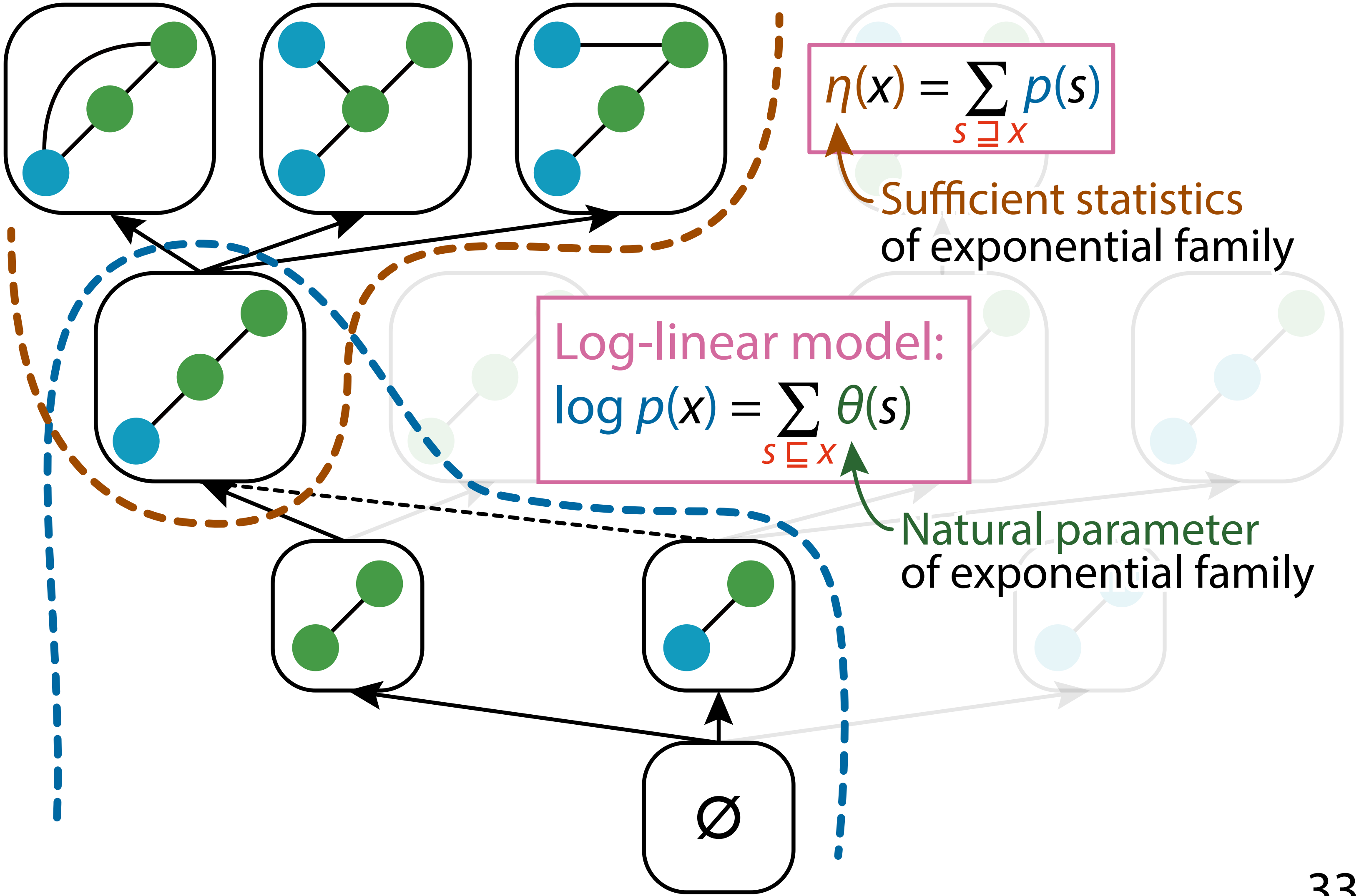
Log-linear model

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

Poset of Subgraphs



Log-Linear Model on Subgraphs



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

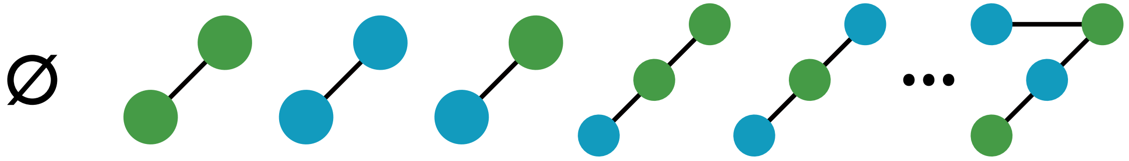
Sufficient statistics of exponential family

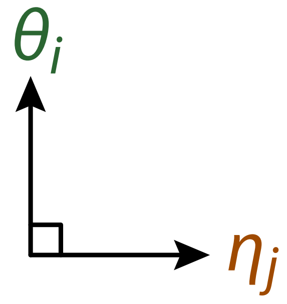
Log-linear model:

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

Natural parameter of exponential family

Information of Each Subgraph

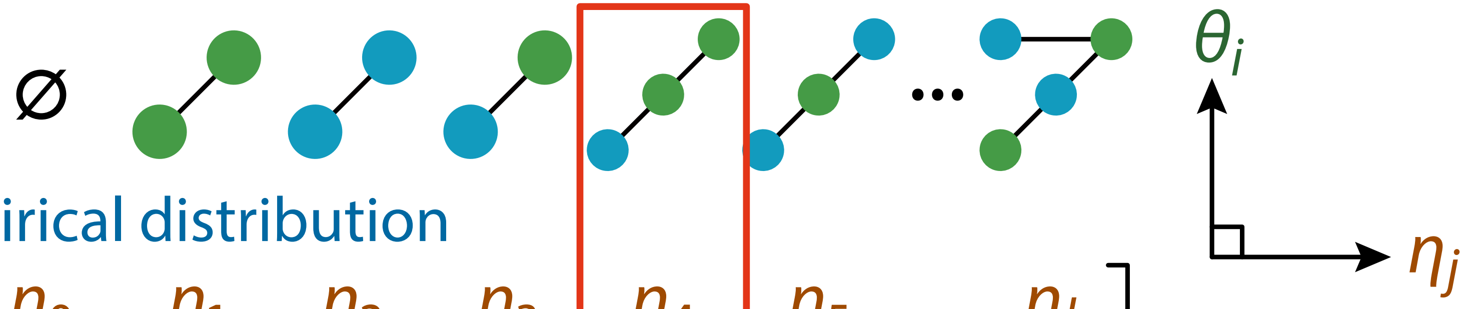




P: Empirical distribution

$$\begin{array}{c} \text{Freq.} \\ \text{Coef.} \end{array} \begin{bmatrix} \eta_0 & \eta_1 & \eta_2 & \eta_3 & \eta_4 & \eta_5 & \cdots & \eta_k \\ \theta_0 & \theta_1 & \theta_2 & \theta_3 & \theta_4 & \theta_5 & & \theta_k \end{bmatrix}$$

Information of Each Subgraph



P: Empirical distribution

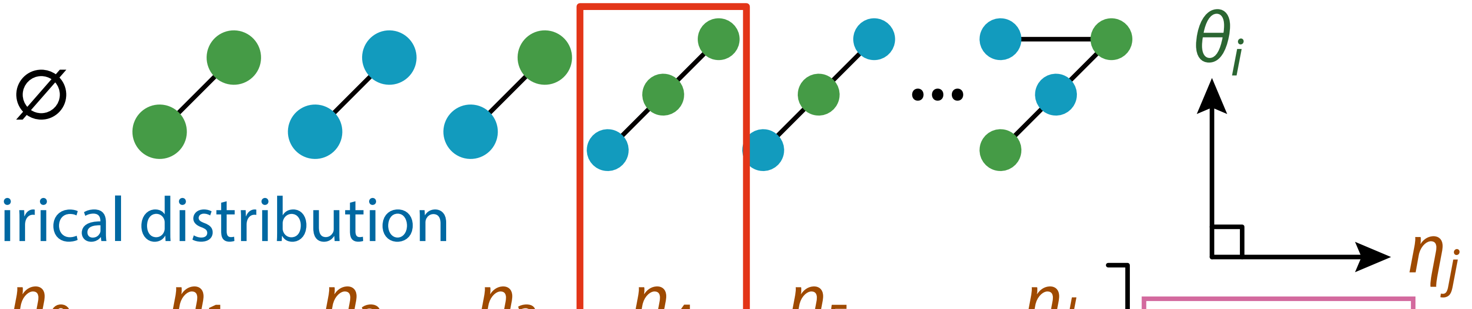
$$\begin{array}{l} \text{Freq.} \\ \text{Coef.} \end{array} \begin{bmatrix} \eta_0 & \eta_1 & \eta_2 & \eta_3 & \eta_4 & \eta_5 & \cdots & \eta_k \\ \theta_0 & \theta_1 & \theta_2 & \theta_3 & \theta_4 & \theta_5 & & \theta_k \end{bmatrix}$$

$$\longrightarrow \mathrm{KL}(P, Q)$$

Q: Null distribution

$$\begin{array}{l} \text{Freq.} \\ \text{Coef.} \end{array} \begin{bmatrix} \eta_0 & \eta_1 & \eta_2 & \eta_3 & ? & \eta_5 & \cdots & \eta_k \\ ? & ? & ? & ? & 0 & ? & & ? \end{bmatrix}$$

Information of Each Subgraph



P: Empirical distribution

$$\begin{array}{l} \text{Freq.} \\ \text{Coef.} \end{array} \left[\begin{array}{cccccccc} \eta_0 & \eta_1 & \eta_2 & \eta_3 & \eta_4 & \eta_5 & \cdots & \eta_k \\ \theta_0 & \theta_1 & \theta_2 & \theta_3 & \theta_4 & \theta_5 & \cdots & \theta_k \end{array} \right]$$

Q: Null distribution

$$\begin{array}{l} \text{Freq.} \\ \text{Coef.} \end{array} \left[\begin{array}{cccccccc} \eta_0 & \eta_1 & \eta_2 & \eta_3 & ? & \eta_5 & \cdots & \eta_k \\ ? & ? & ? & ? & 0 & ? & \cdots & ? \end{array} \right]$$

R: Uniform distribution

$$\begin{array}{l} \text{Freq.} \\ \text{Coef.} \end{array} \left[\begin{array}{cccccccc} 1.0 & ? & ? & ? & ? & ? & \cdots & ? \\ \theta'_0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 \end{array} \right]$$

$\longrightarrow \mathrm{KL}(P, Q)$

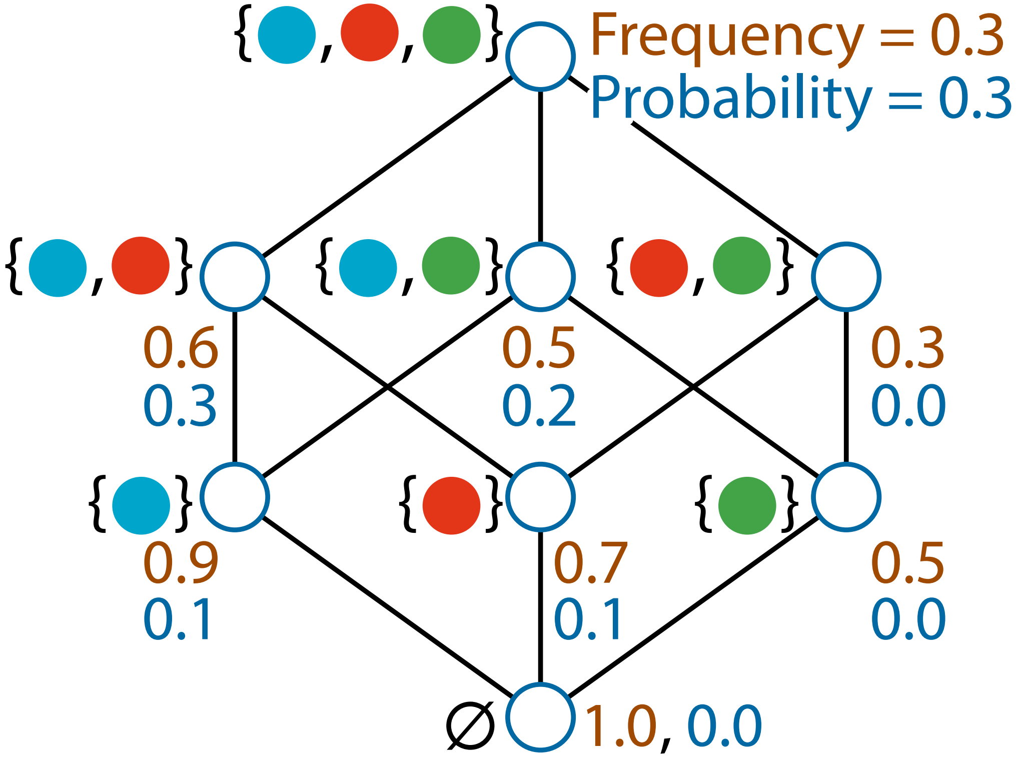
$\mathrm{KL}(Q, R)$

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

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

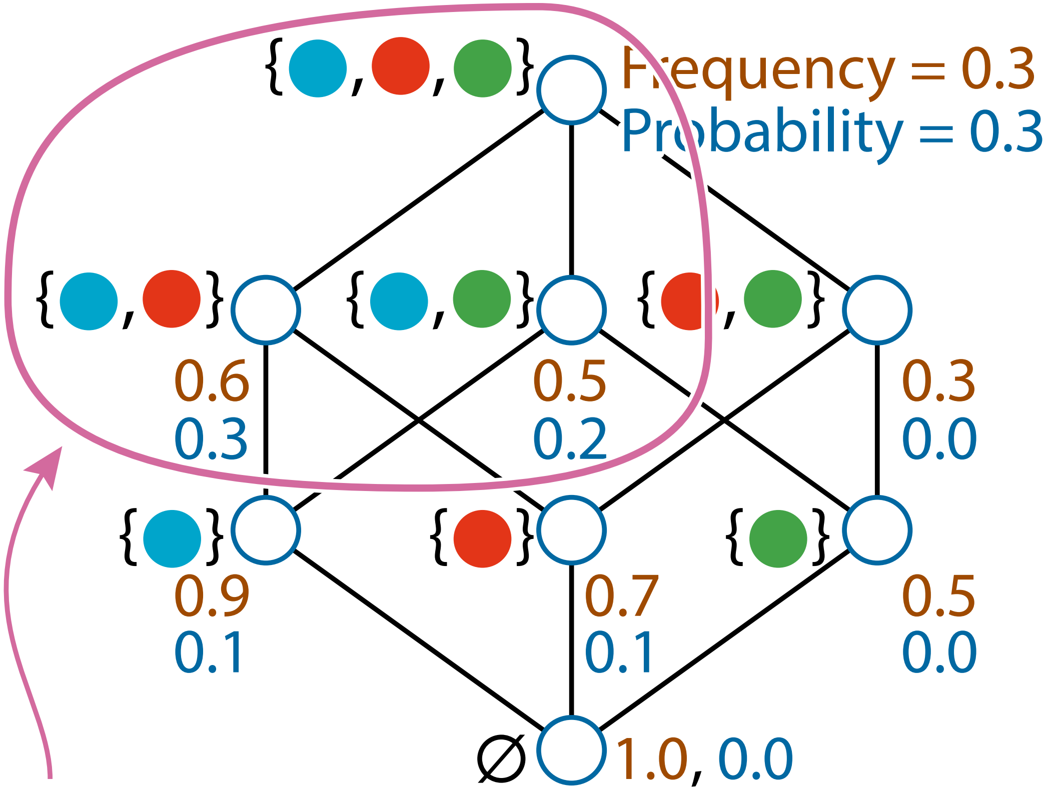


Number of nodes = $2^{\#\text{features}}$
⇒combinatorial explosion!

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

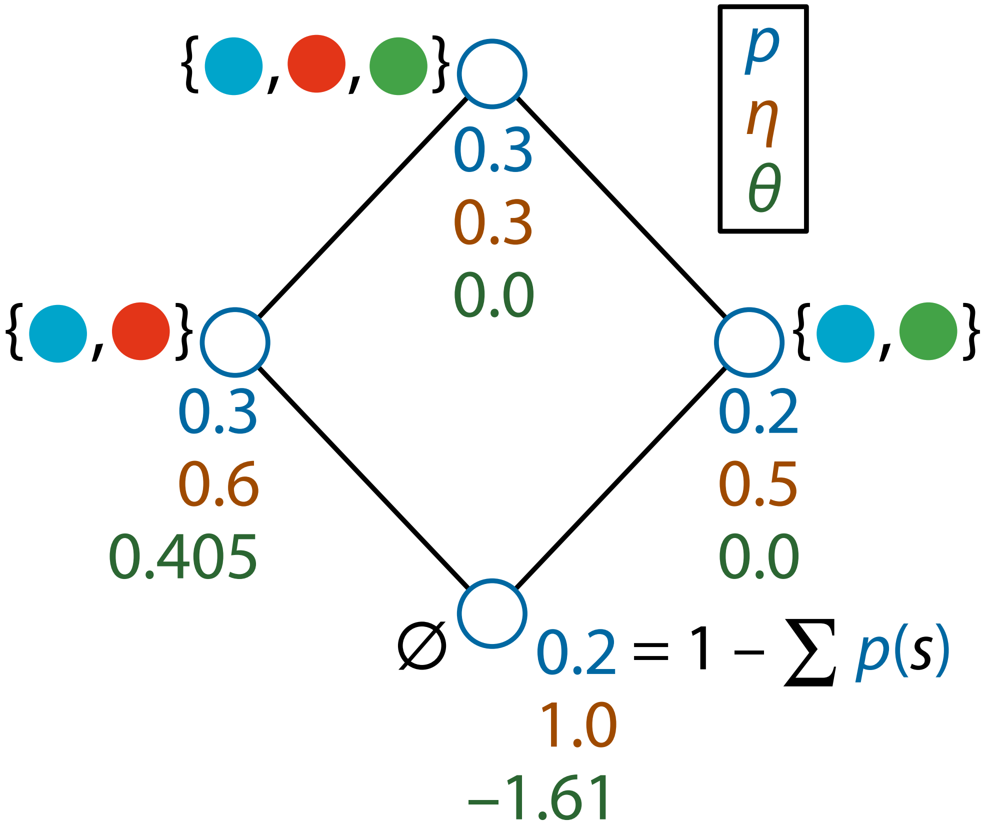


Probability ≥ 0.2
(user specified threshold)

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)

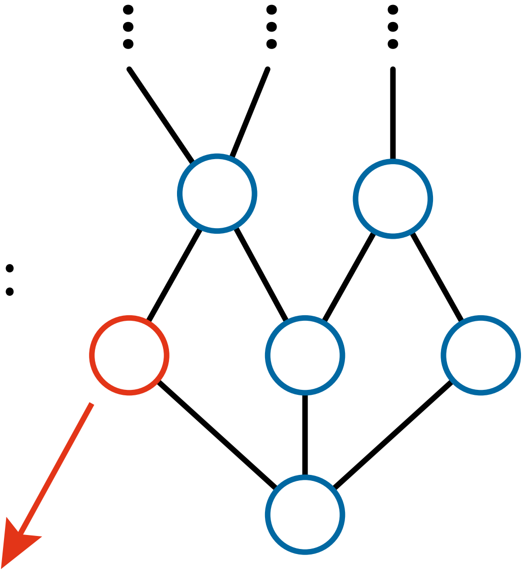
# features: 41,270

	●	●	●	
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	
⋮		⋮		

Sample size: 990,002

Total runtime: 4.95 seconds

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



significant interactions: 583

Single feature: 537

Pairwise interactions: 41

Triple interactions: 5

Example on Real Data (accidents)

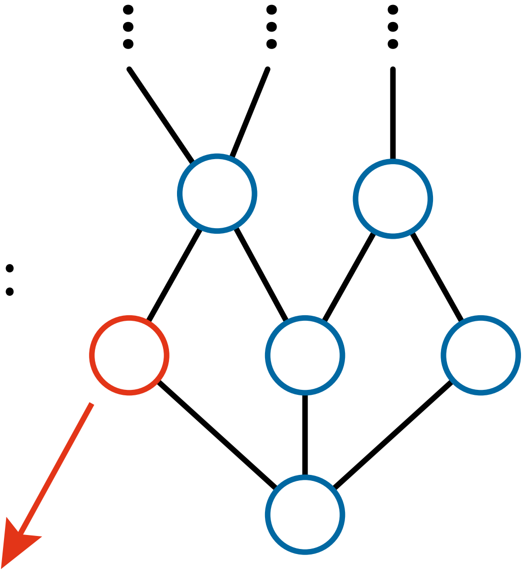
# features: 468

	●	●	●	…
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	
⋮		⋮		

Sample size: 340,183

Total runtime: 4.95 seconds

# 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