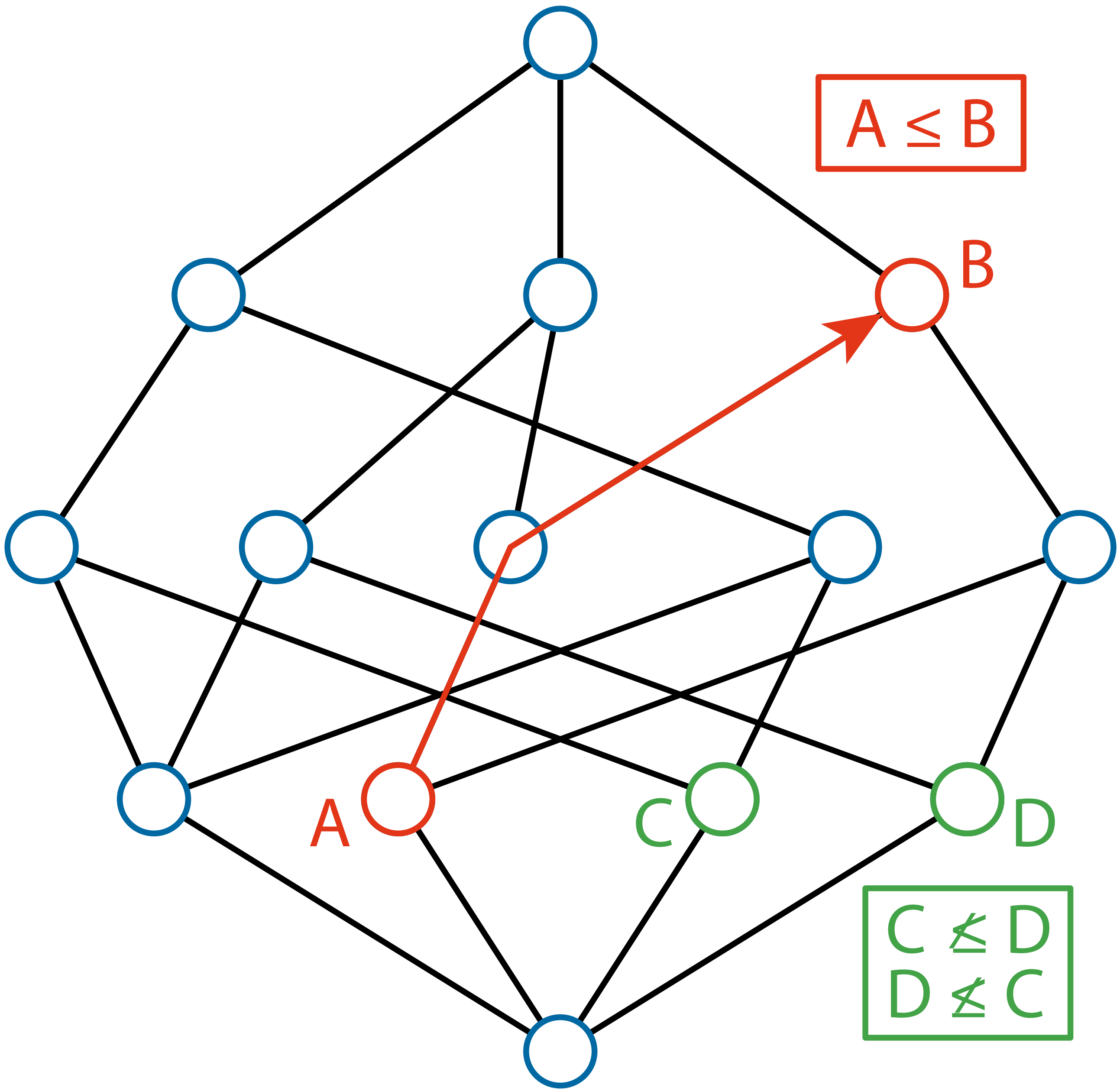
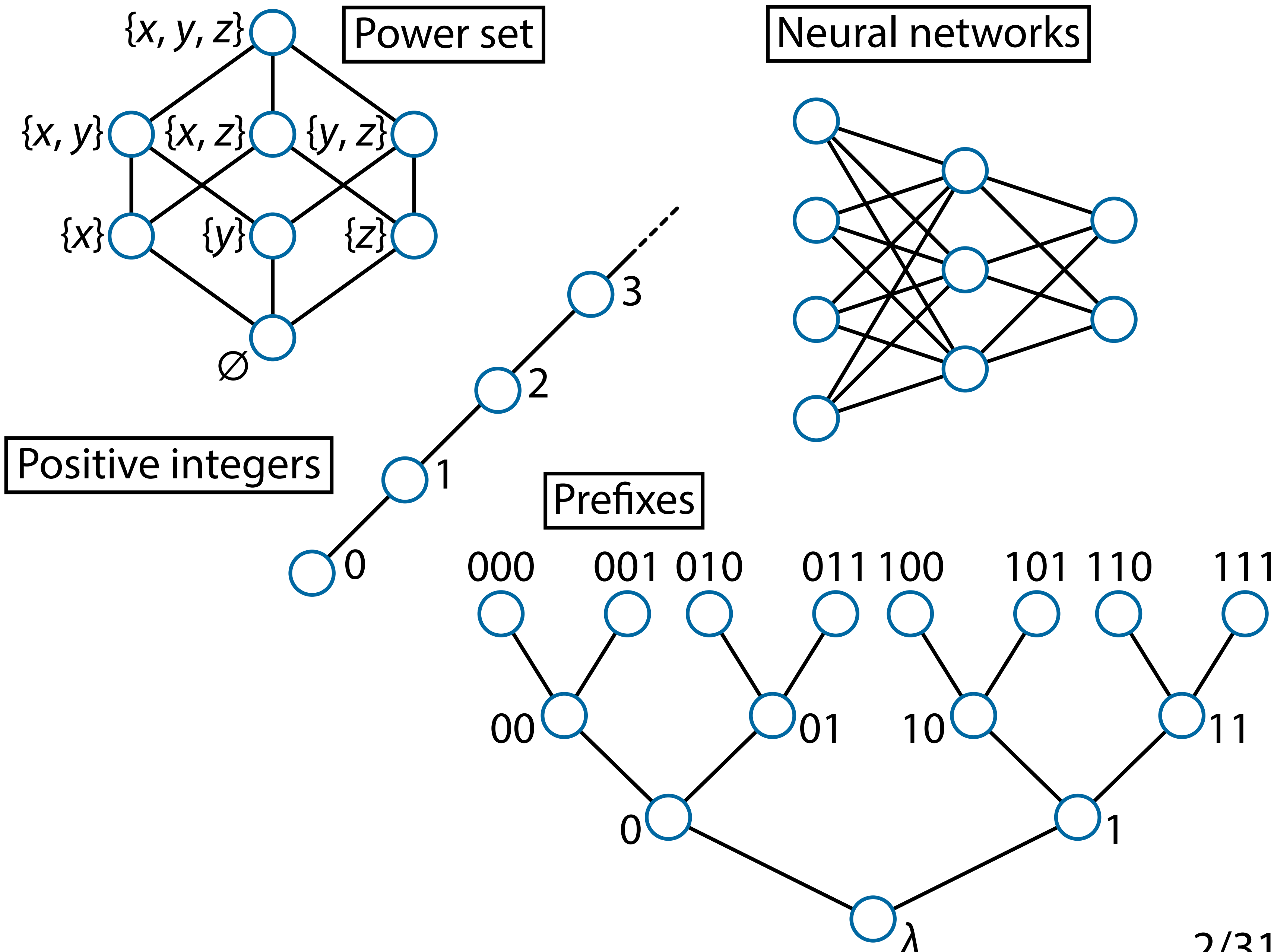
March 17, 2016
@RIKEN BSI

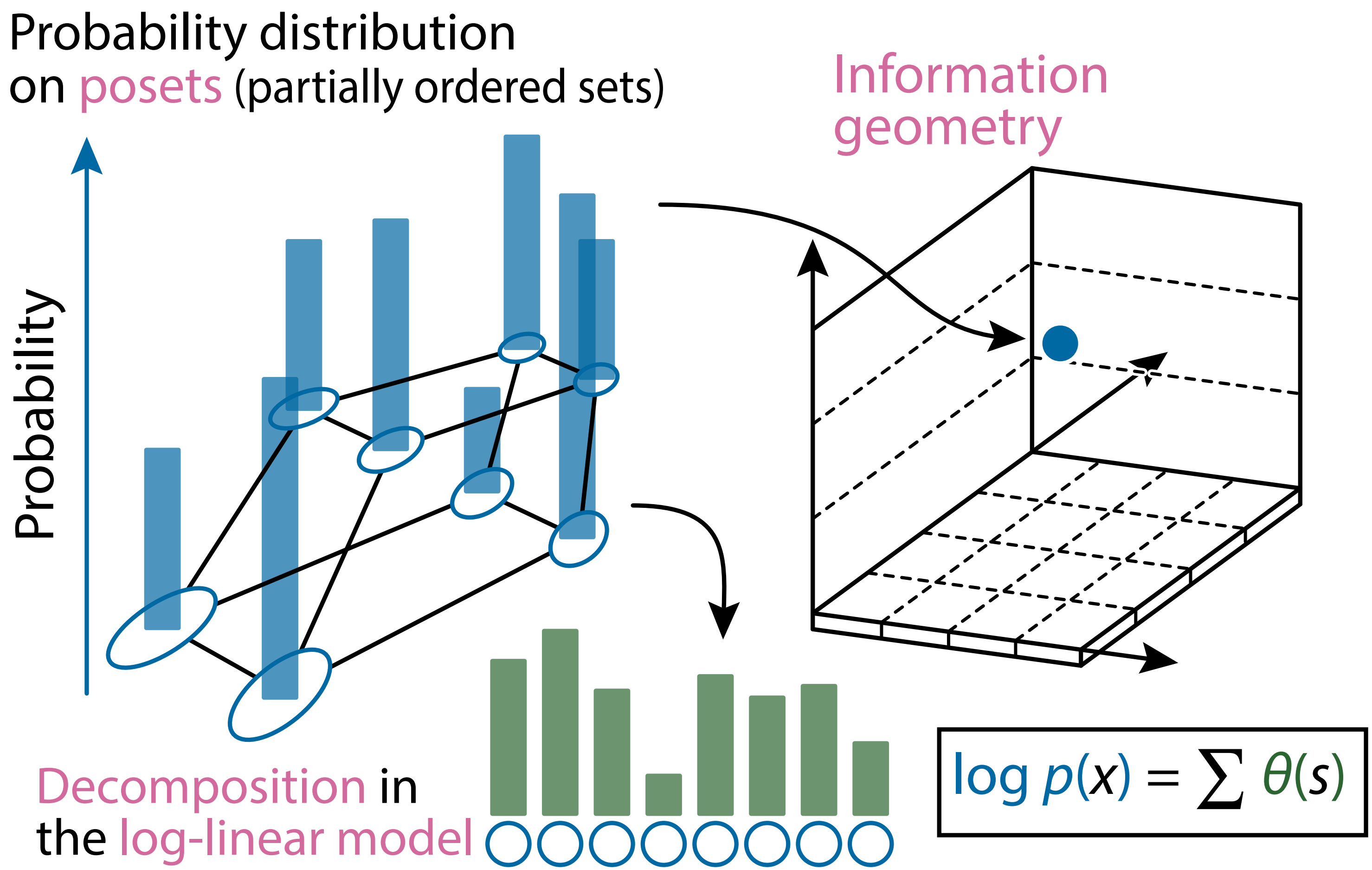
Statistical Analysis on Order Structures

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


$A \leq B$
B
A
C
D
$C \not\leq D$
$D \not\leq C$


Power set
{x, y, z}
{x, y}
{x, z}
{y, z}
{x}
{y}
{z}
∅
Neural networks
Positive integers
0
1
2
3
Prefixes
000
001
010
011
100
101
110
111
00
01
10
11
0
1
λ

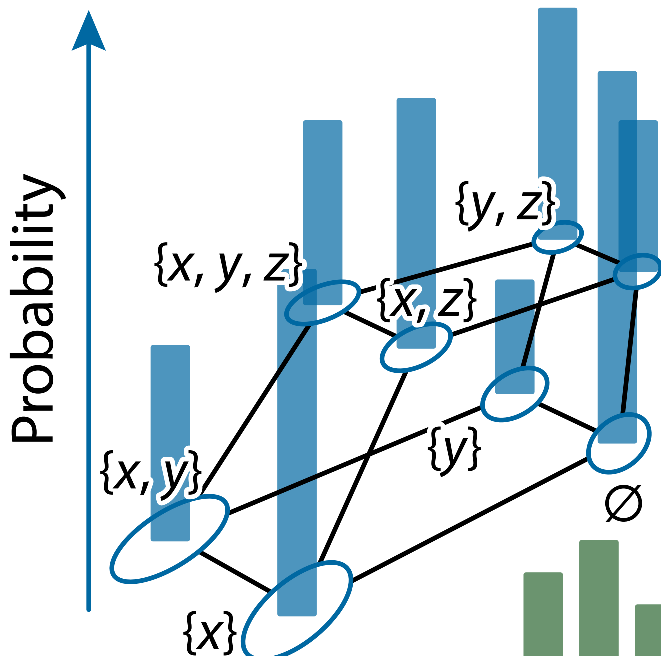
Summary



S. Amari, Information geometry on hierarchy of probability distributions, IEEE TIT 2001
M. Sugiyama, H. Nakahara, K. Tsuda, Information Decomposition on Structured Space, arXiv 2016

Summary

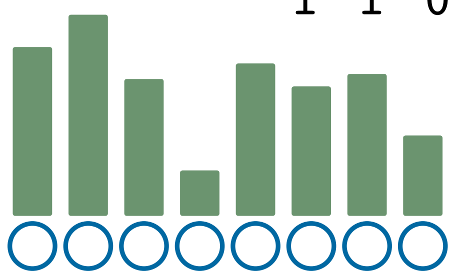
Probability distribution
on posets (partially ordered sets)



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



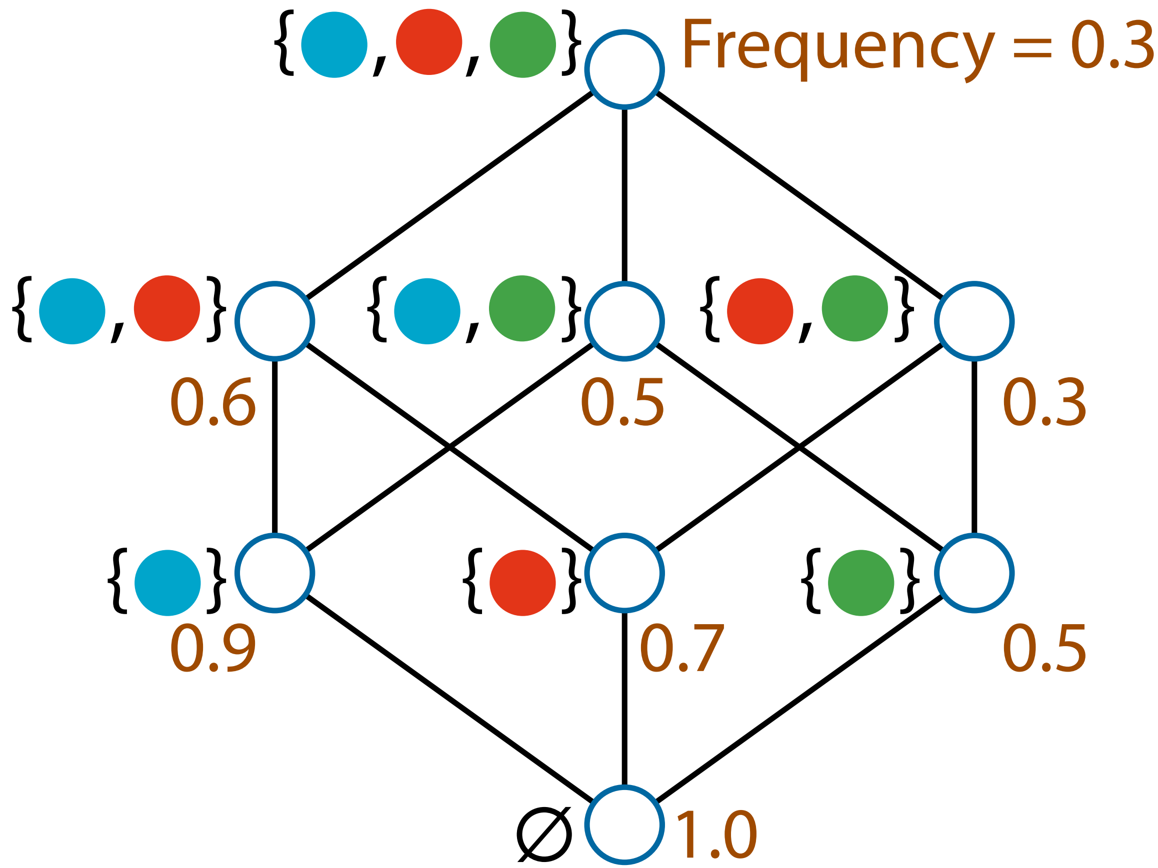
$$\log p(x) = \sum \theta(s)$$

S. Amari, Information geometry on hierarchy of probability distributions, IEEE TIT 2001
M. Sugiyama, H. Nakahara, K. Tsuda, Information Decomposition on Structured Space, arXiv 2016

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

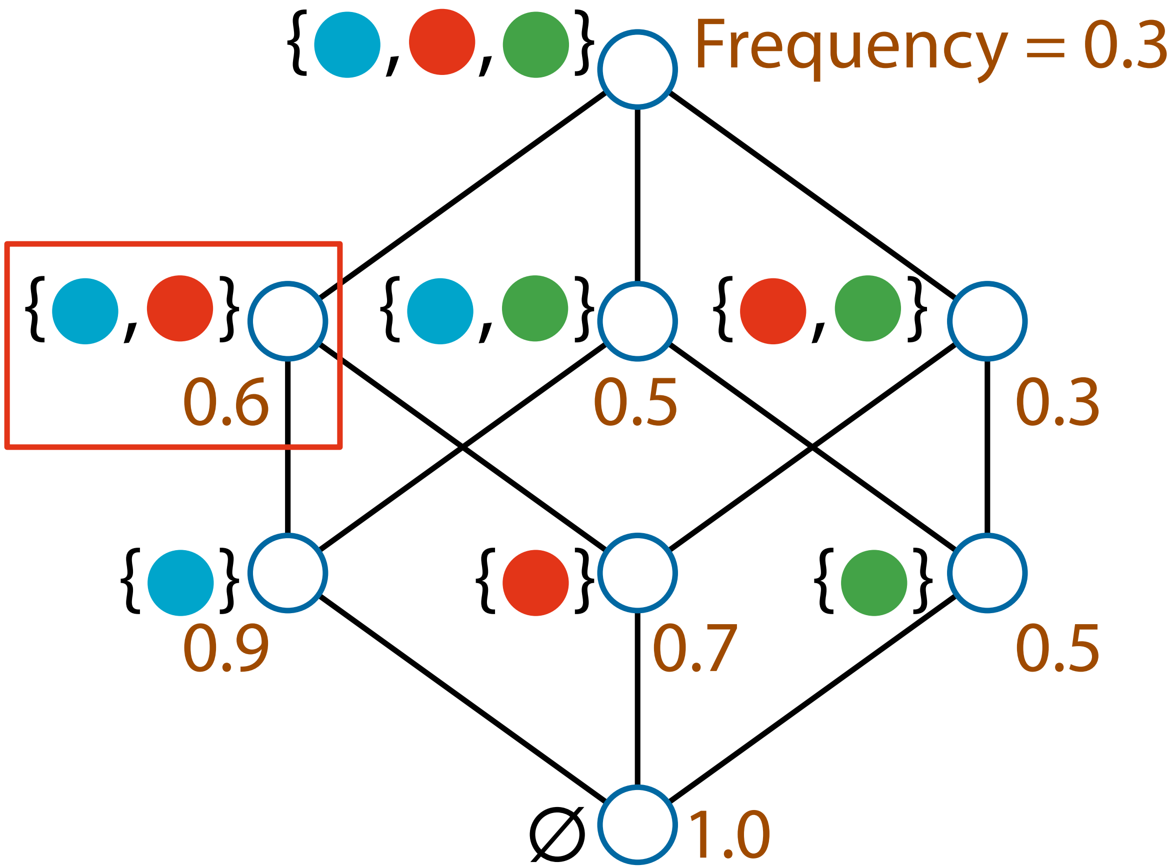
Itemset lattice



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 lattice



Transaction database

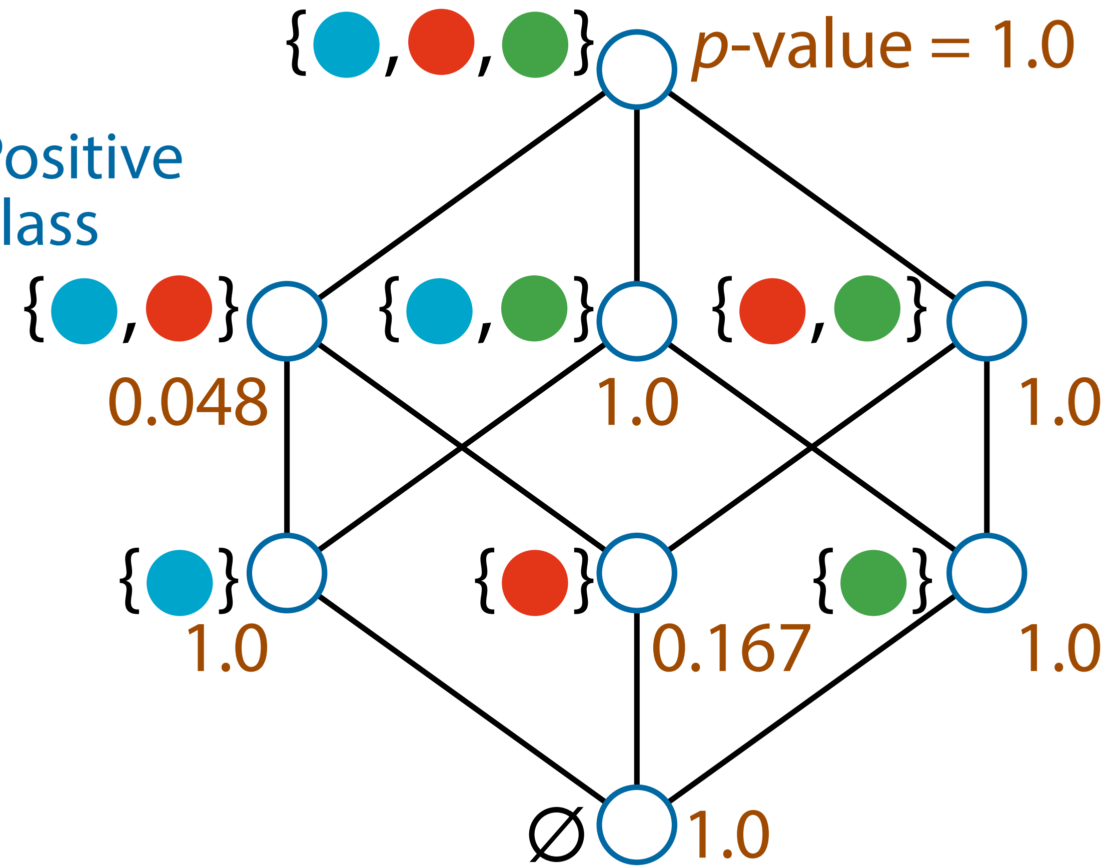
Positive class

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

Negative class

ID 6:	1	0	1
ID 7:	1	0	1
ID 8:	1	1	1
ID 9:	1	0	0
ID10:	0	1	0

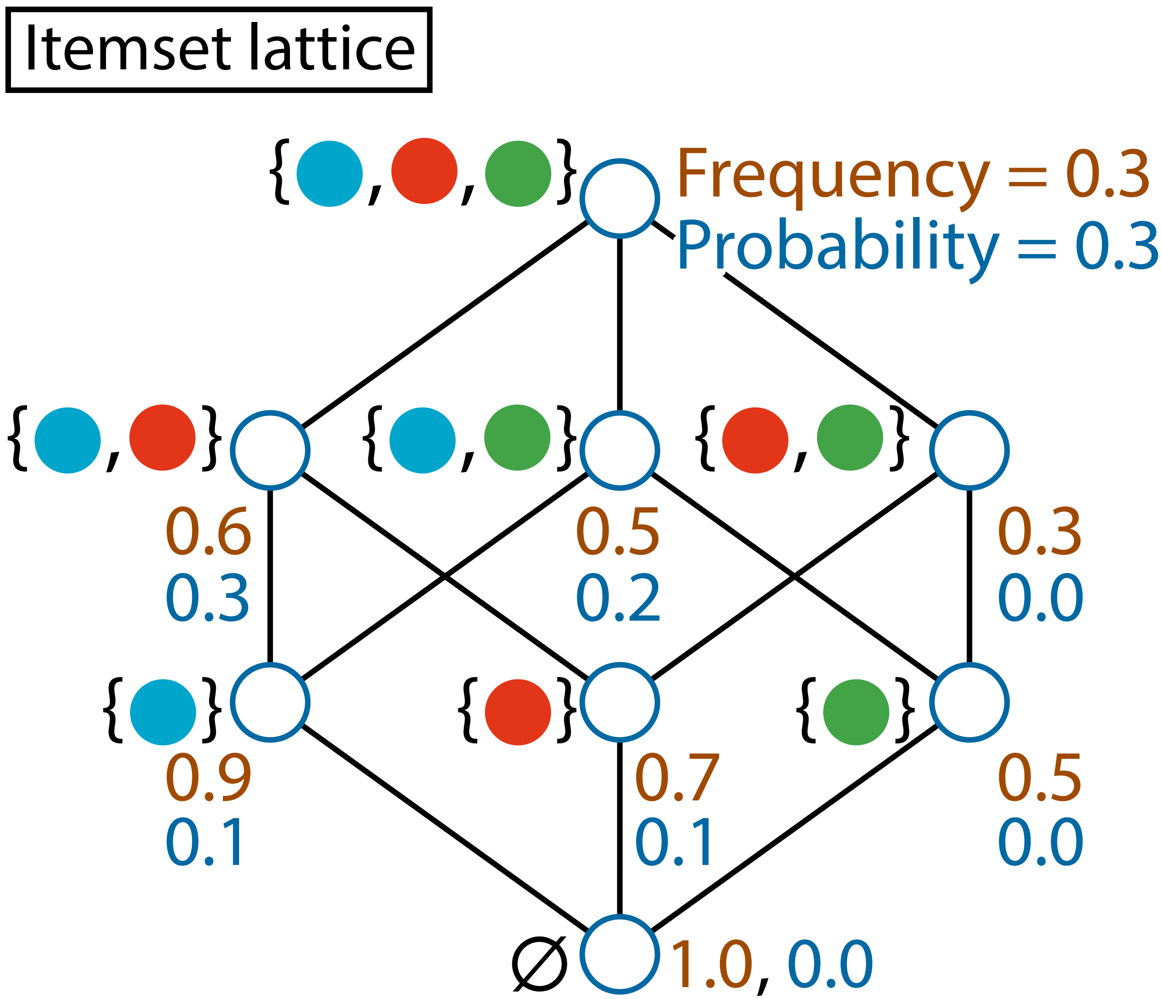
Itemset lattice



LAMP (Terada et al. PNAS 2013)
Westfall-Young light (Llinares-López et al. KDD 2015)

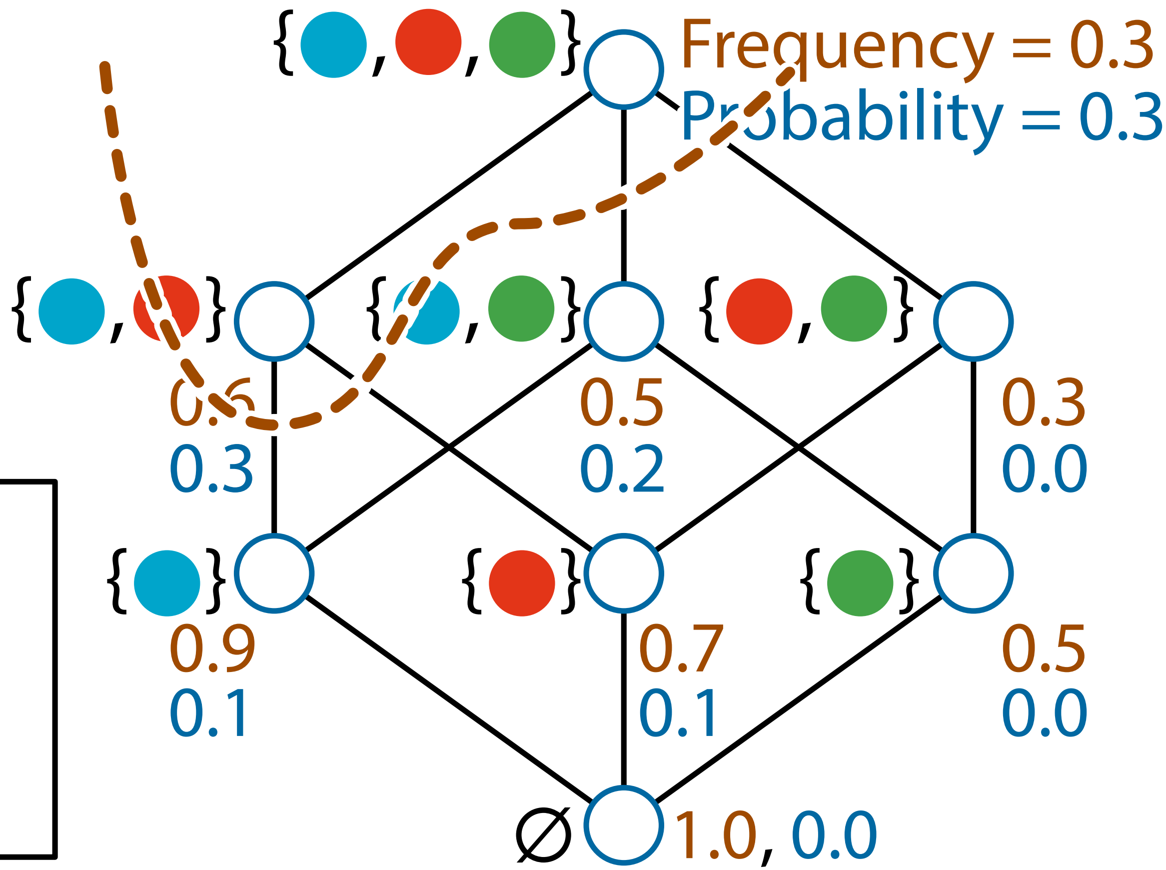
Transaction database

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



Upward = Pattern mining

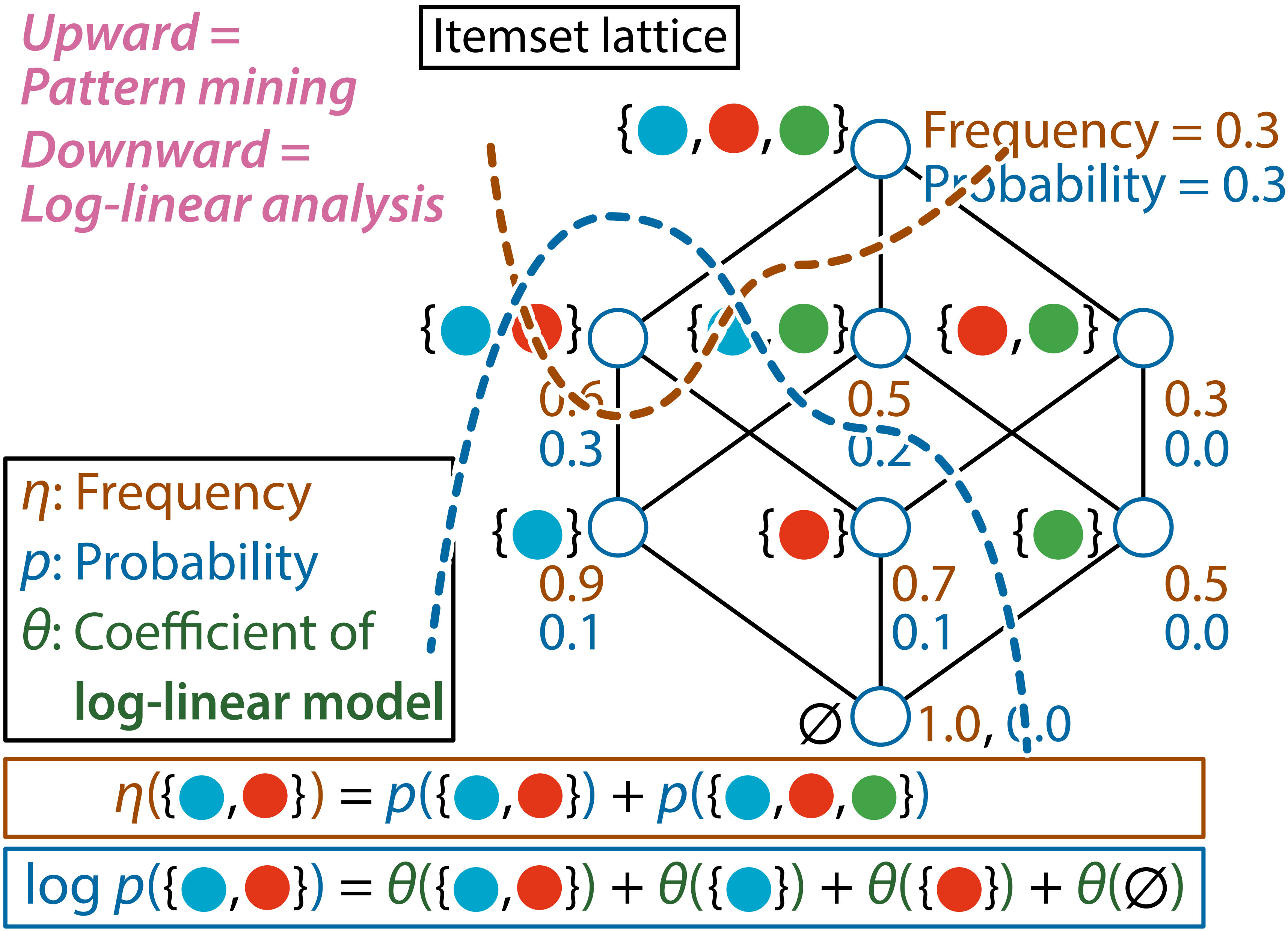
Itemset lattice

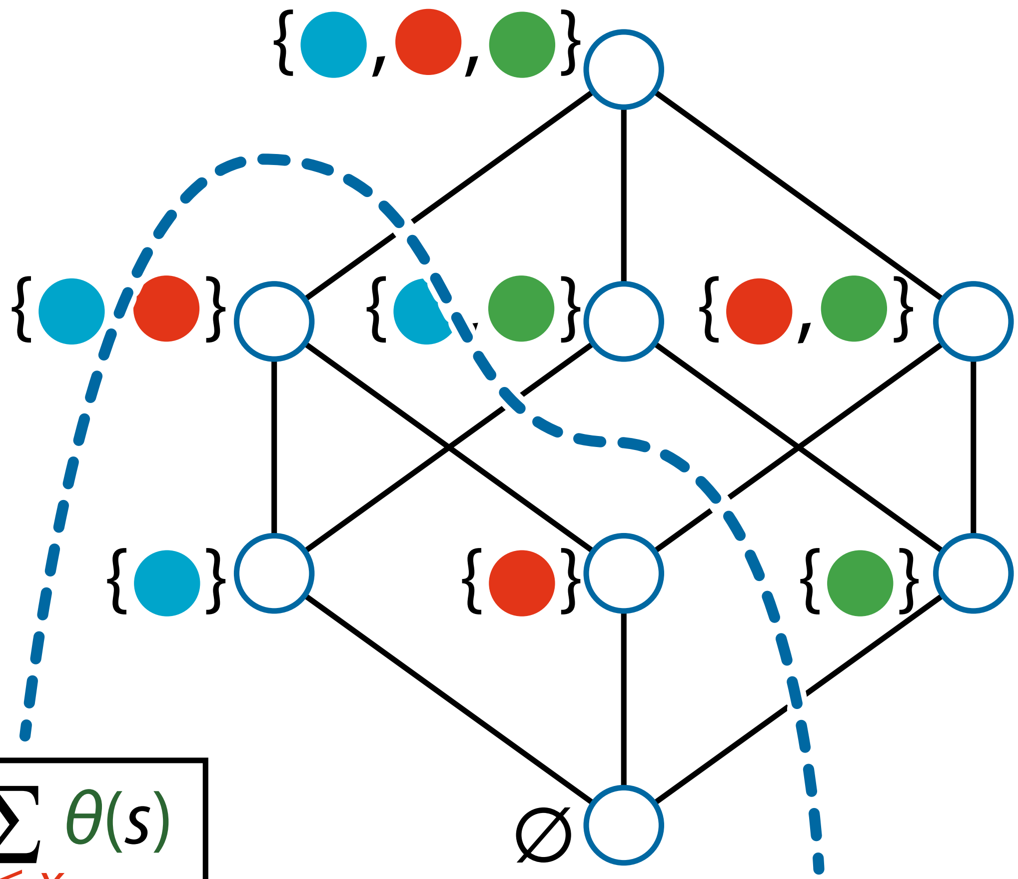


η: Frequency

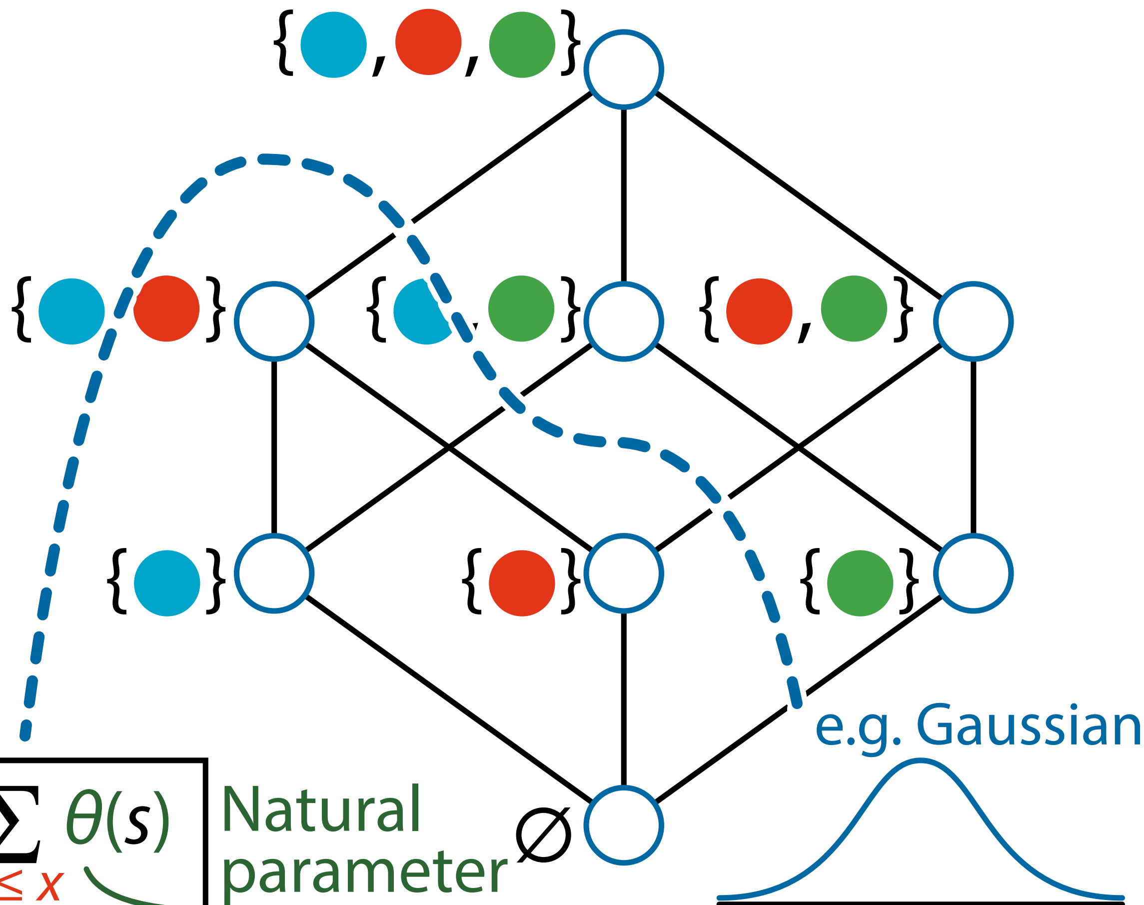
p: Probability

$\eta(\{●,●\}) = p(\{●,●\}) + p(\{●,●,●\})$


Upward = Pattern mining
Downward = Log-linear analysis
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
η: Frequency
p: Probability
θ: Coefficient of log-linear model



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



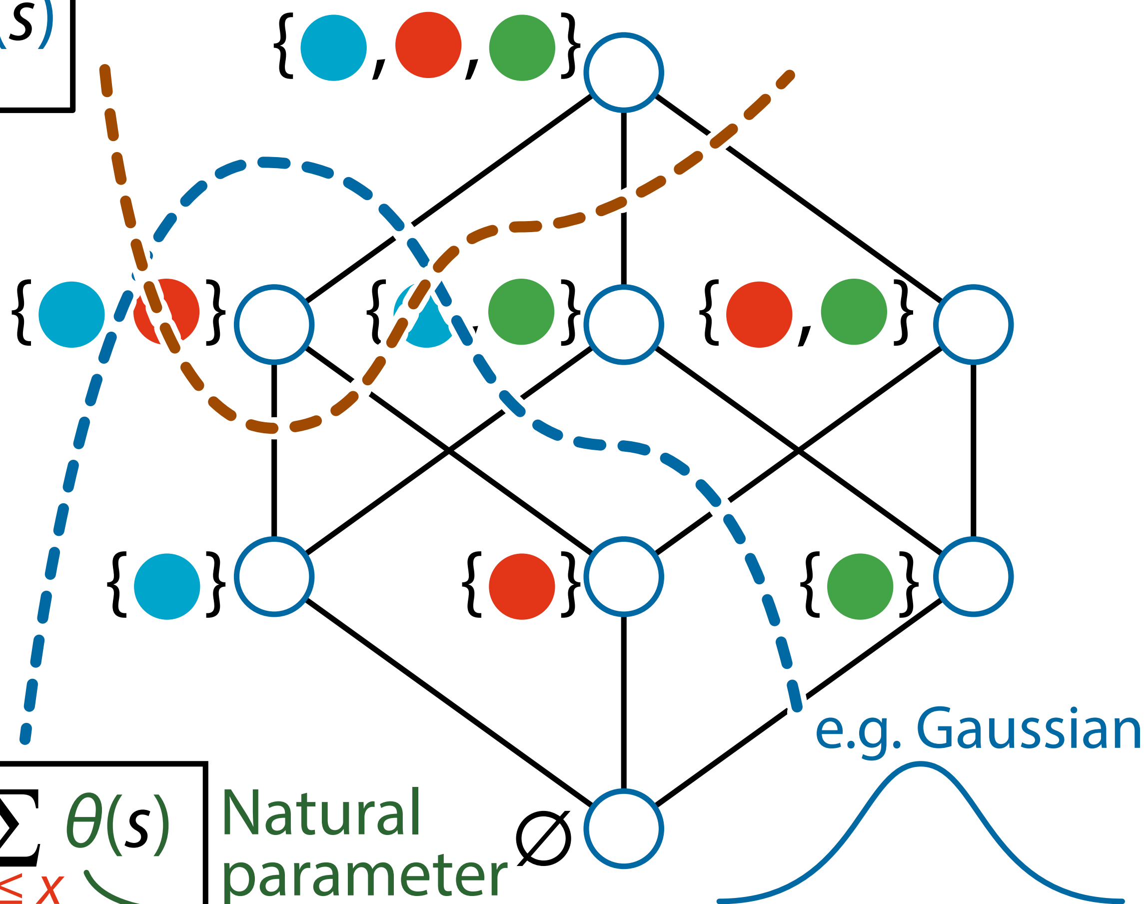
$$\log p(x) = \sum_{s \le x} \theta(s)$$

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

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

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

Sufficient statistics of exponential family

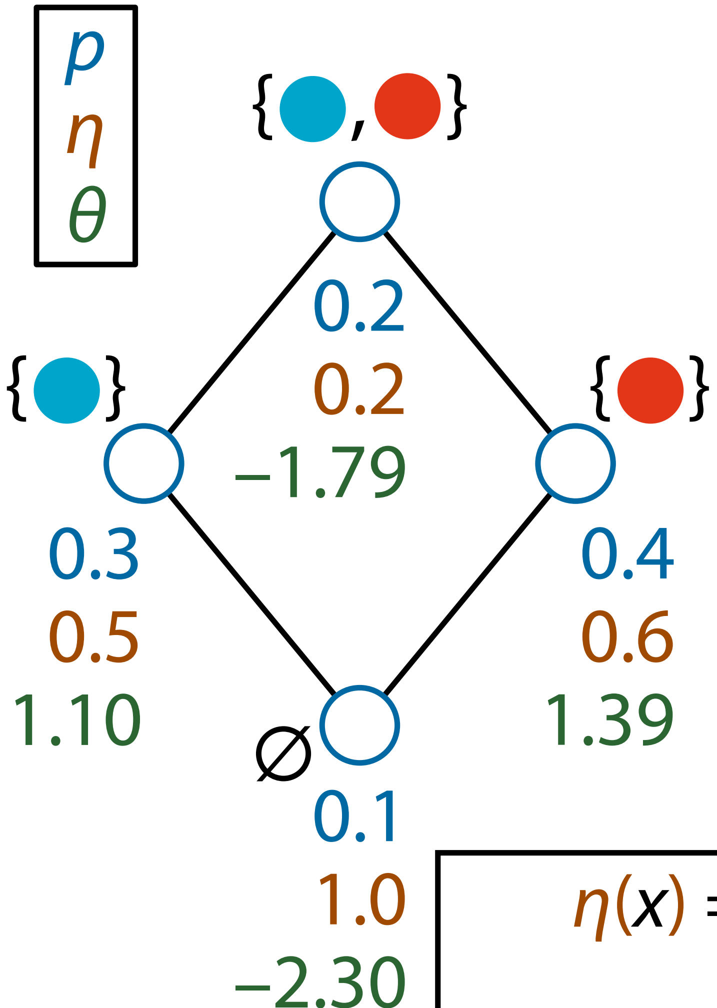


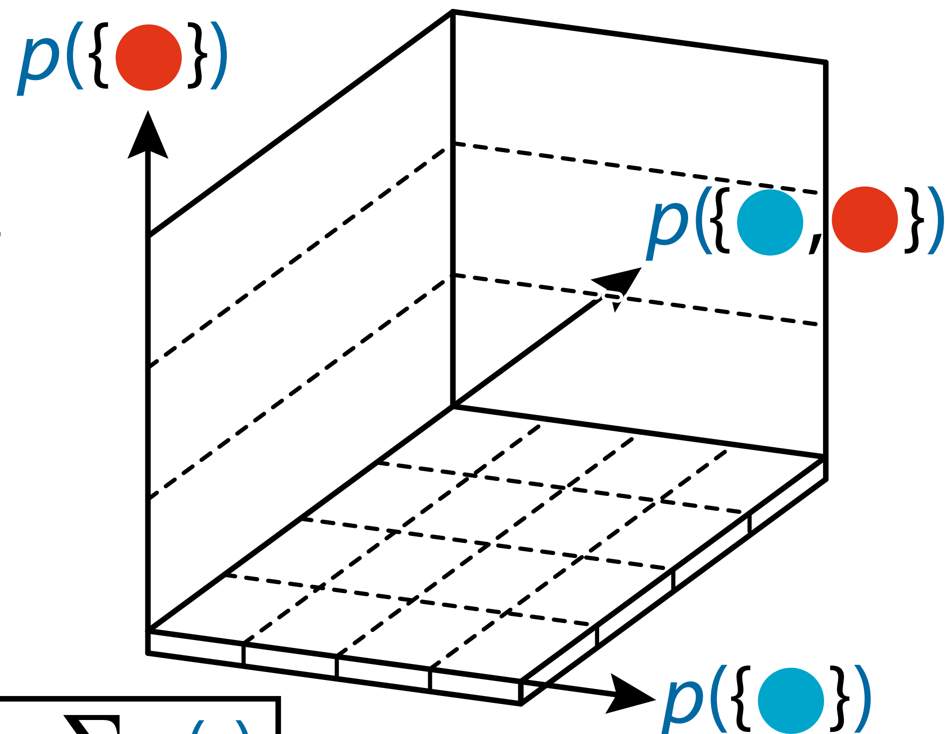
$$\log p(x) = \sum_{s \leq x} \theta(s)$$

Natural parameter

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

Triple for each node

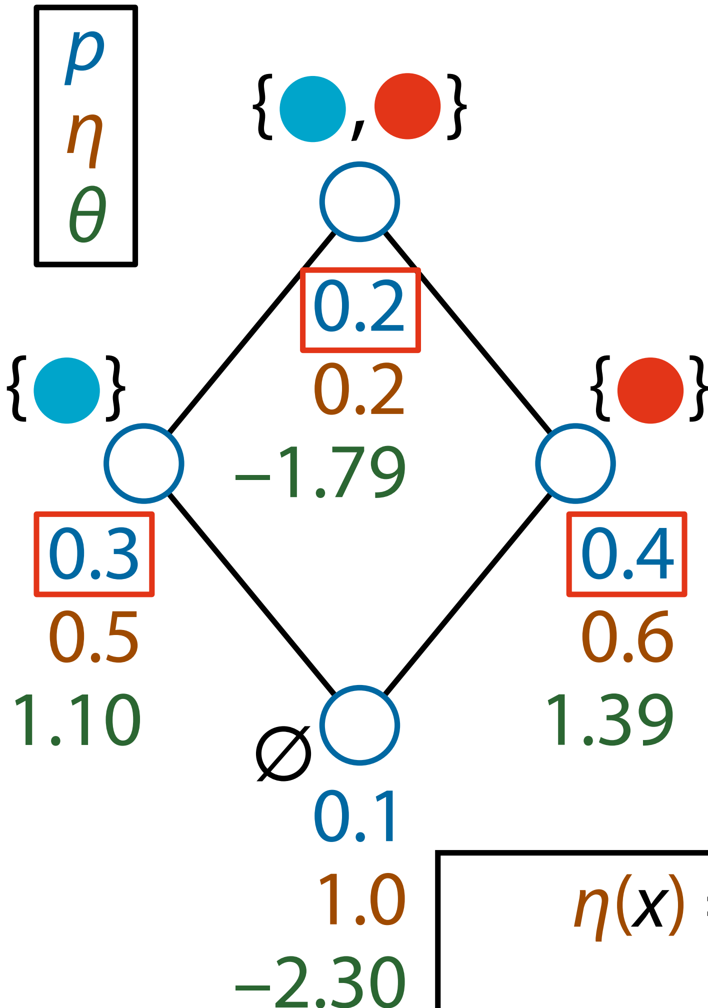




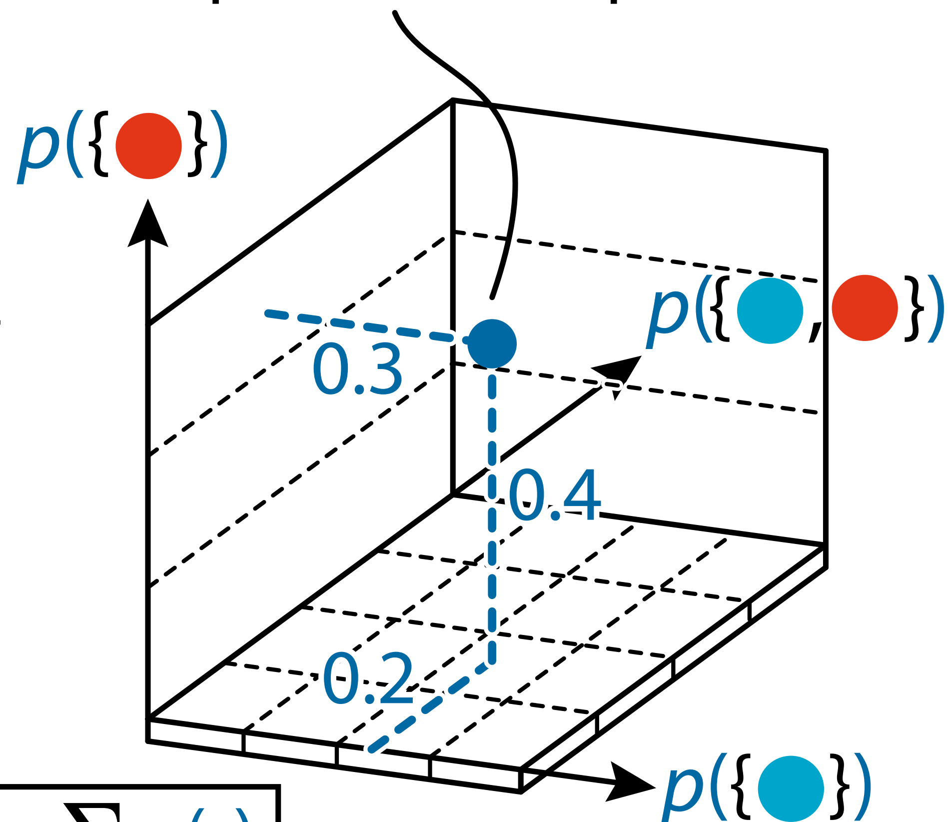
$$\eta(x) = \sum_{s \geq x} p(s)$$

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

Triple for each node



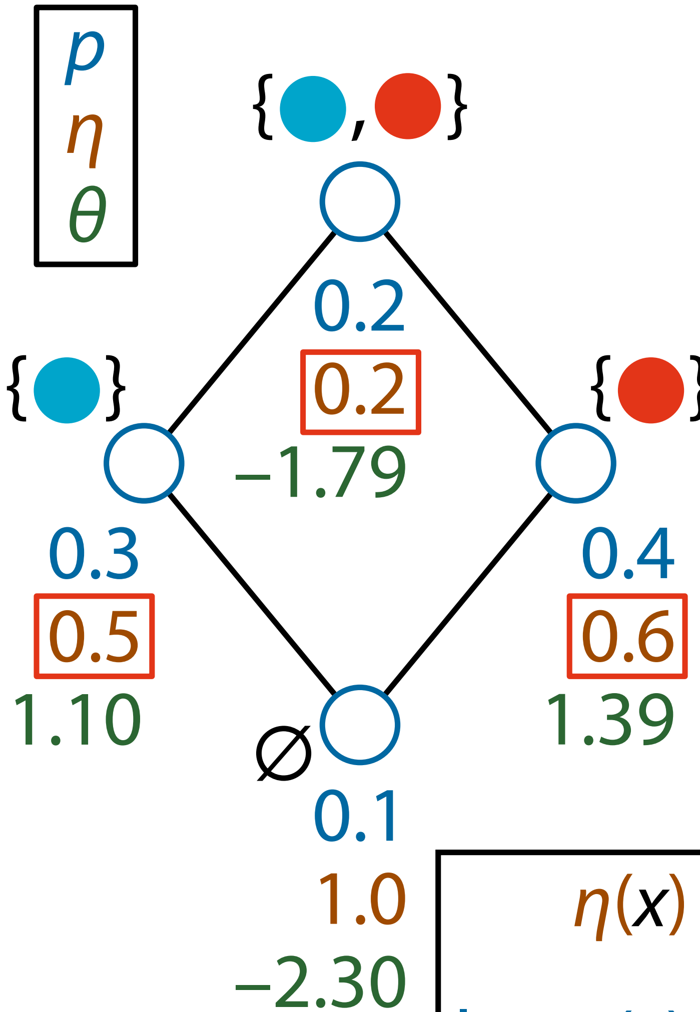
Probability distribution is a "point" in 3D space



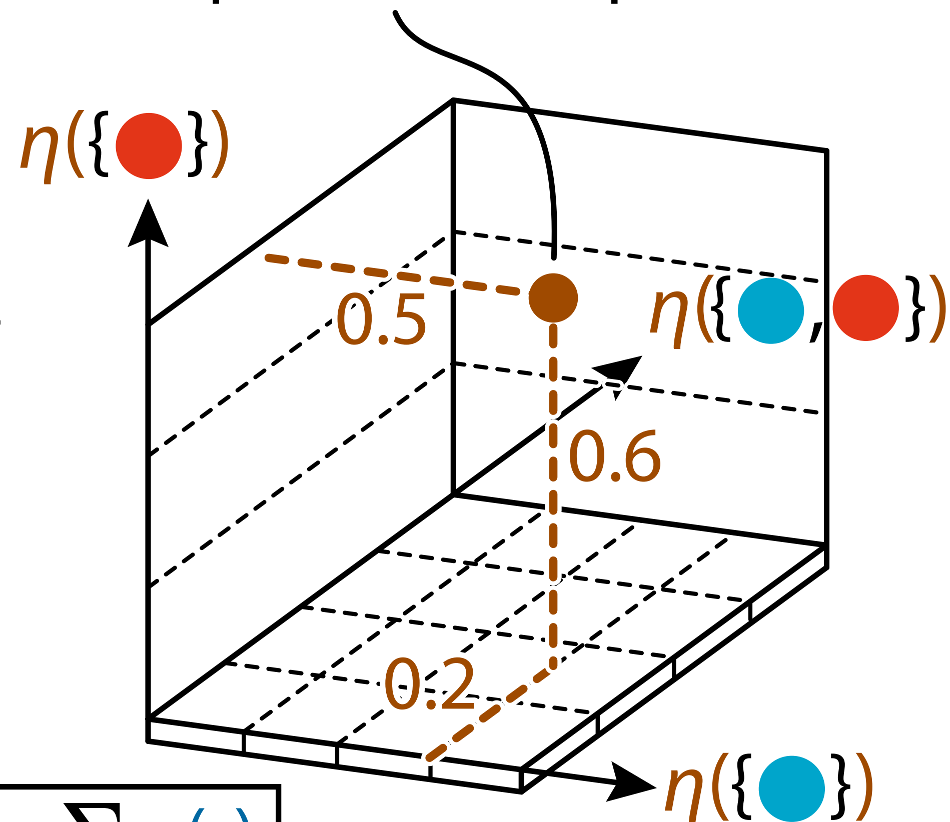
$$\eta(x) = \sum_{s \geq x} p(s)$$

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

Triple for each node



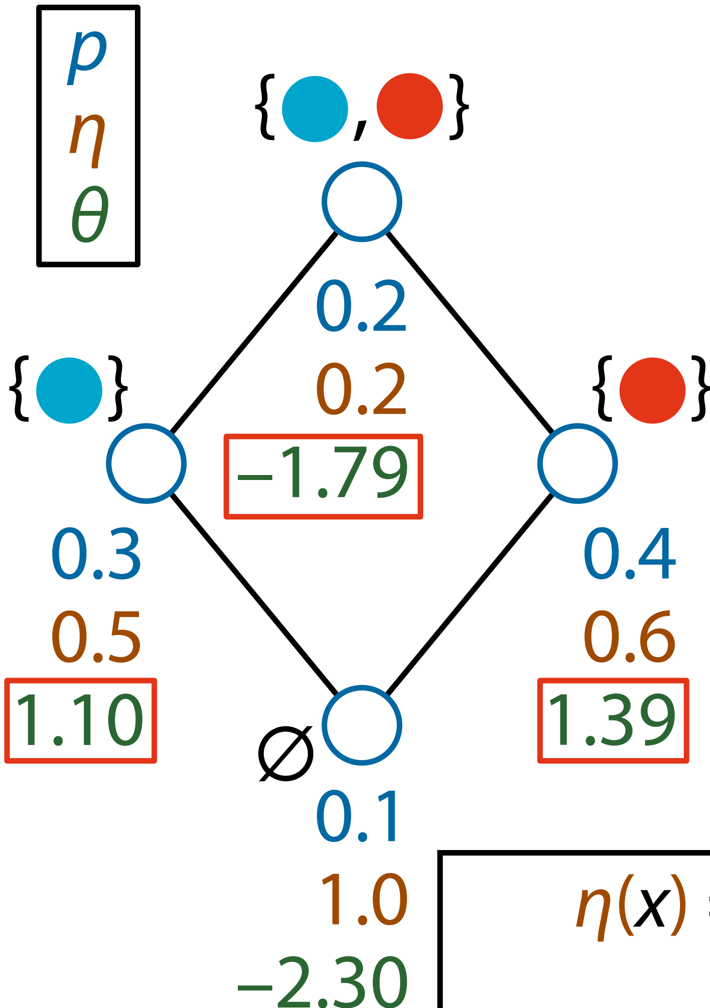
Probability distribution is a "point" in 3D space



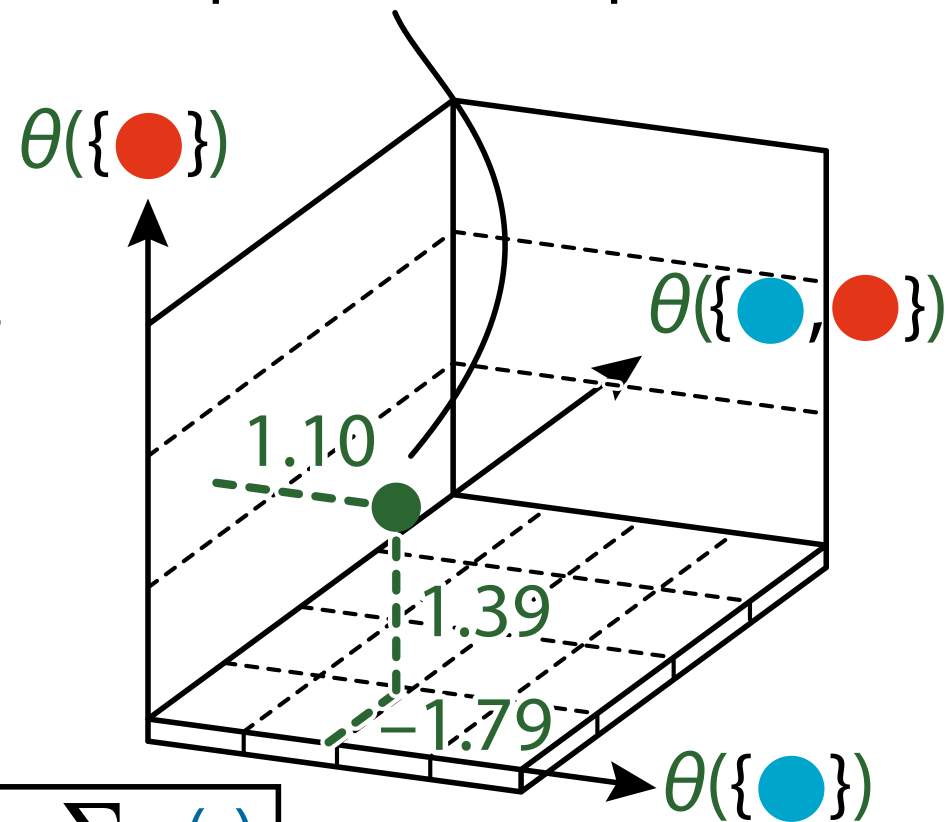
$$\eta(x) = \sum_{s \geq x} p(s)$$

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

Triple for each node



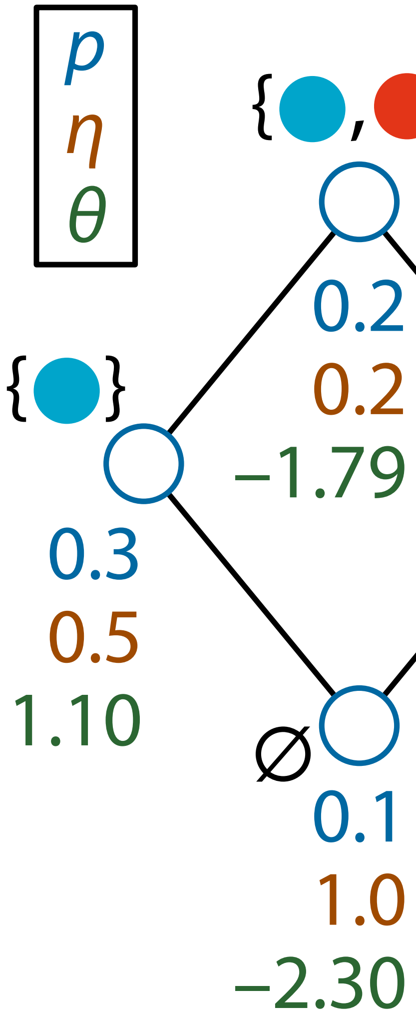
Probability distribution is a "point" in 3D space



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

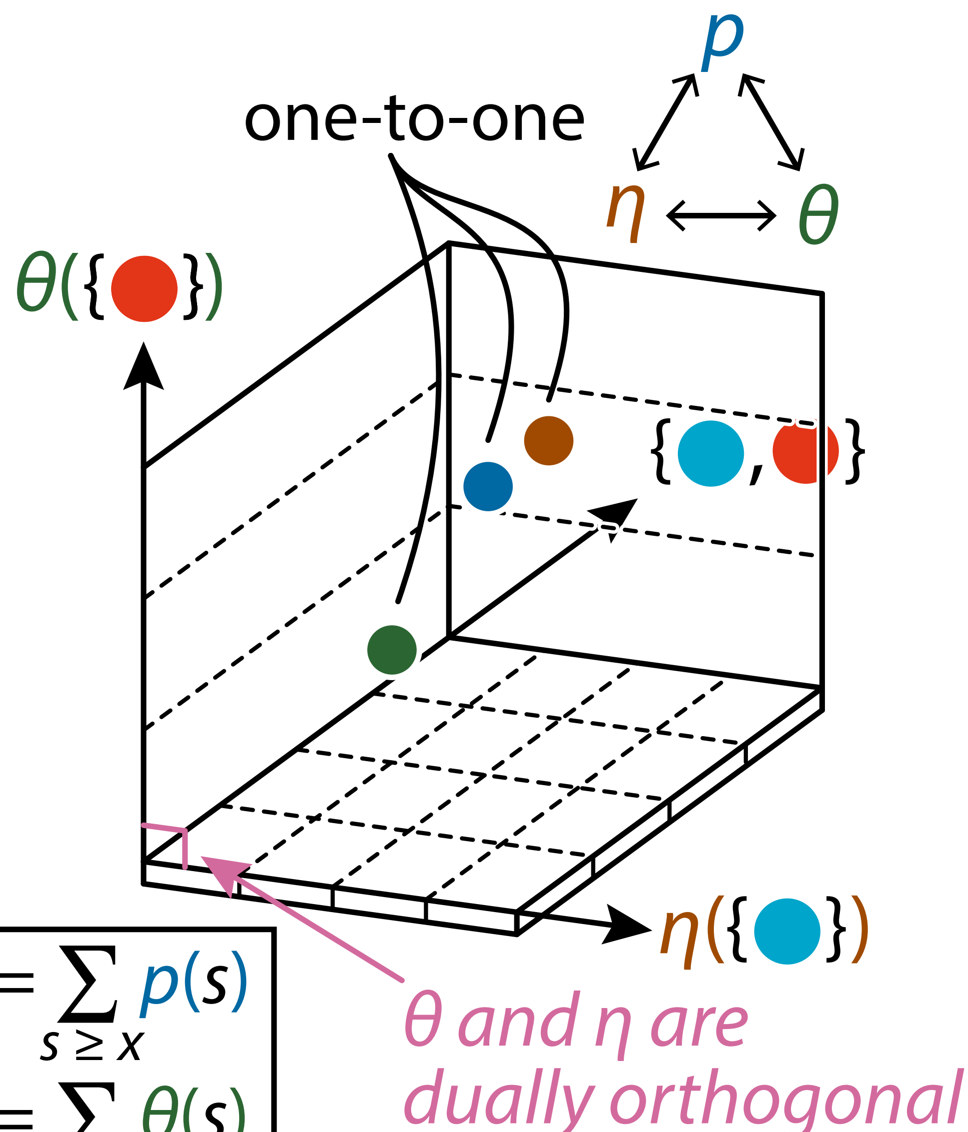
$$\log p(x) = \sum_{s \leq x} \theta(s)$$

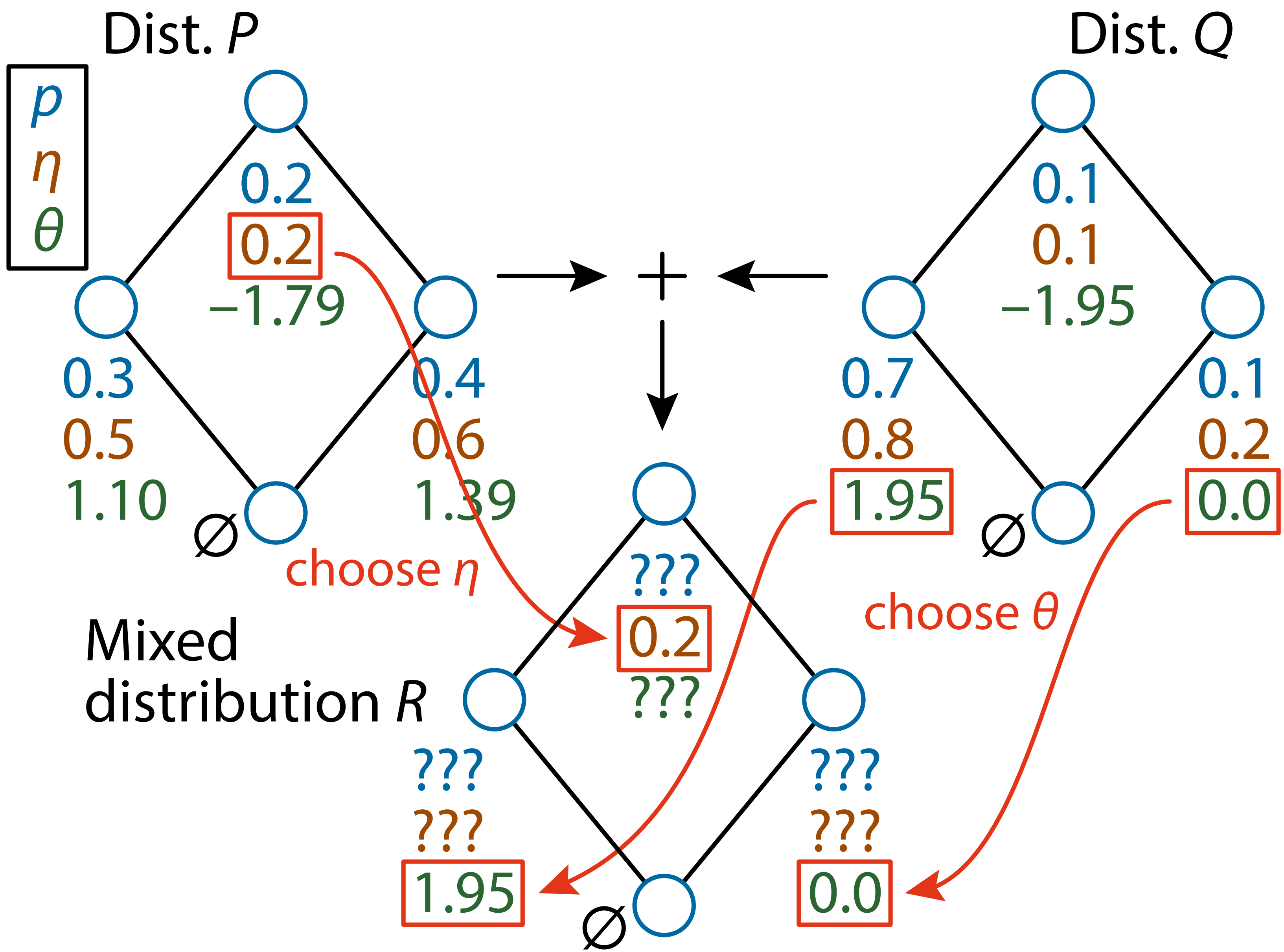
Triple for each node



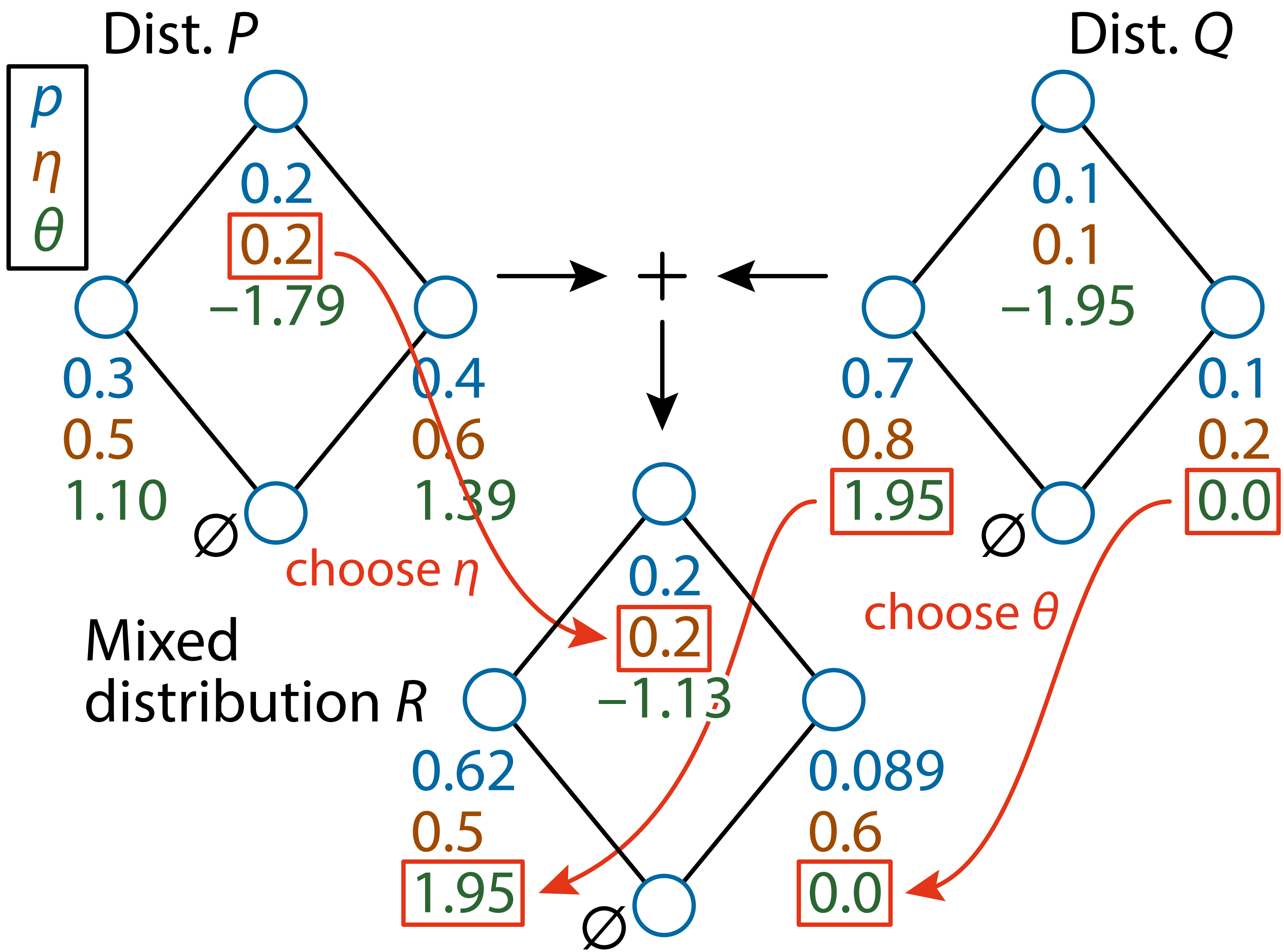
$$\eta(x) = \sum_{s \geq x} p(s)$$

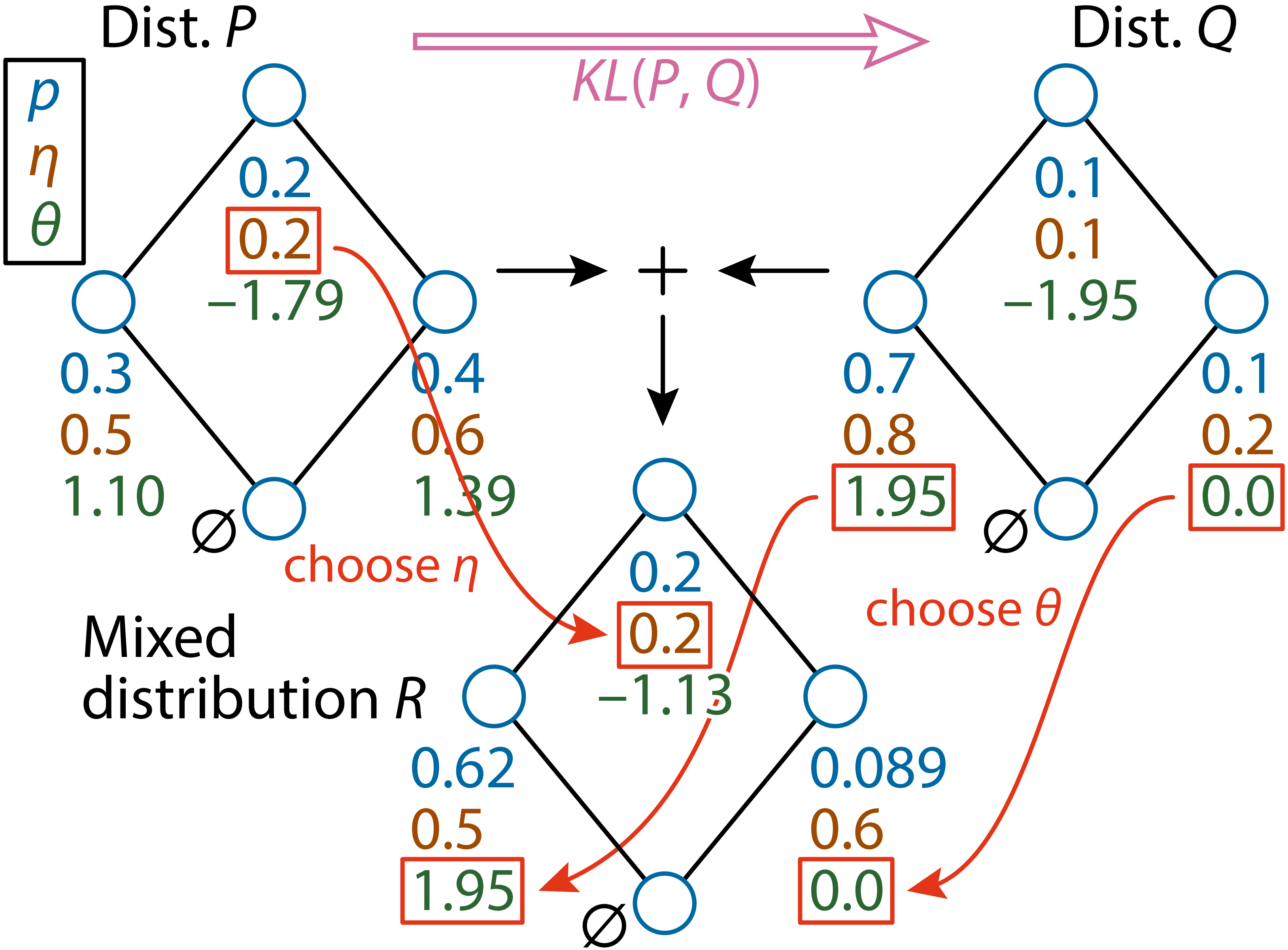
$$\log p(x) = \sum_{s \leq x} \theta(s)$$



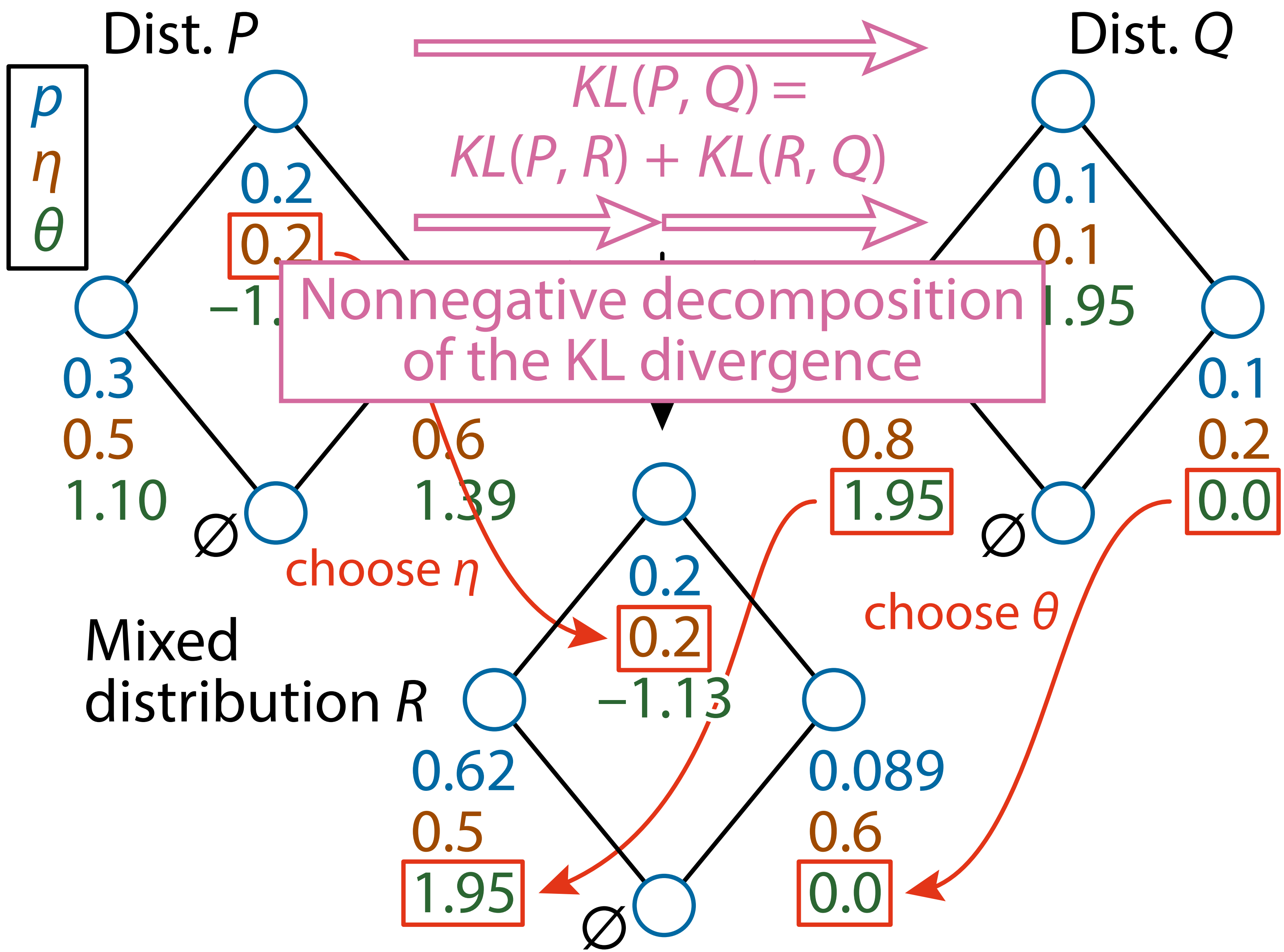


Dist. P
p
η
θ
0.2
0.2
−1.79
0.3
0.5
1.10
0.4
0.6
1.39
∅
choose η
+
Dist. Q
0.1
0.1
−1.95
0.7
0.8
1.95
0.1
0.2
0.0
∅
choose θ
Mixed distribution R
???
0.2
???
???
???
1.95
???
???
0.0
∅

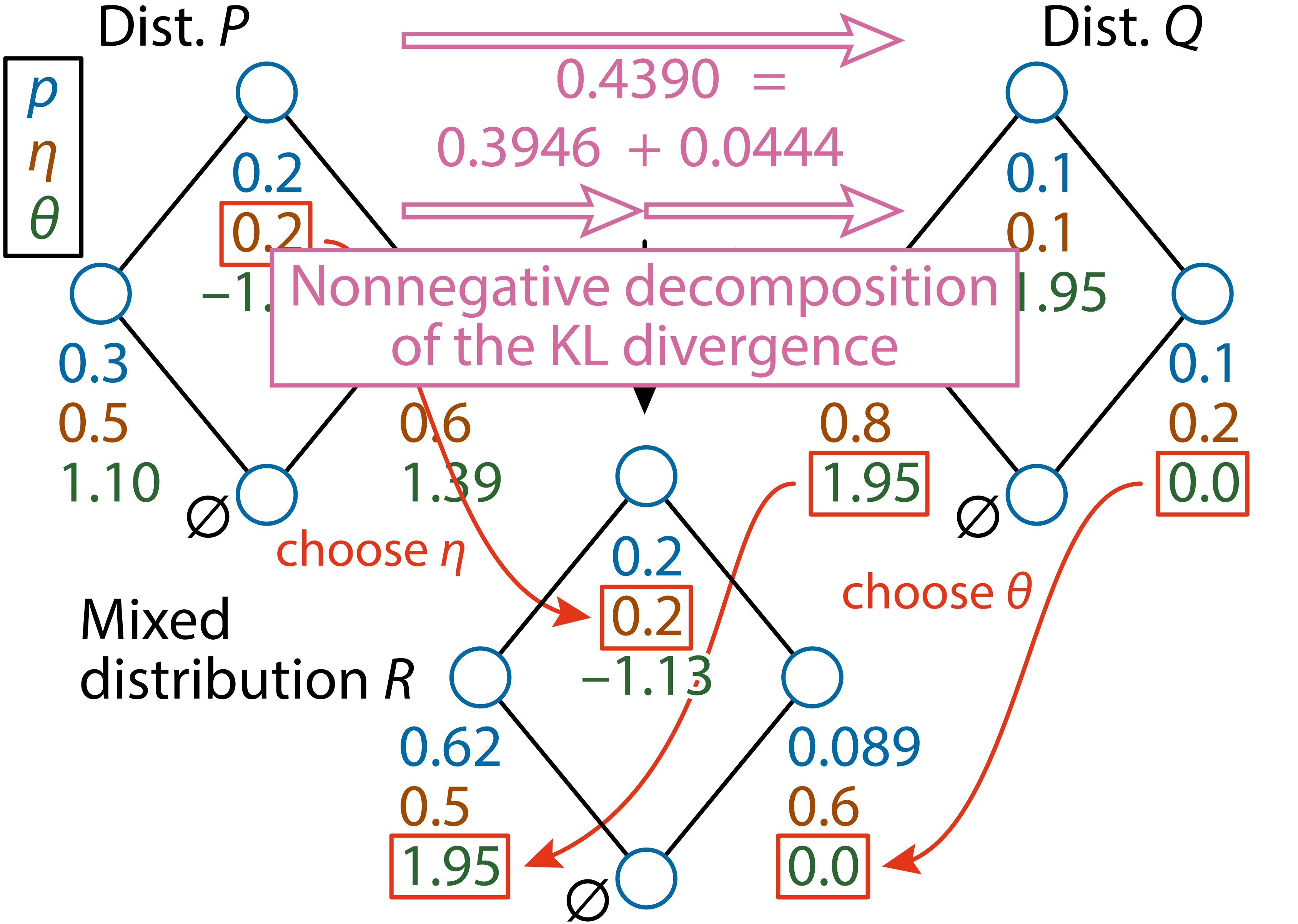

Dist. P
Dist. Q
p
η
θ
0.2
0.2
−1.79
0.3
0.5
1.10
0.4
0.6
1.39
∅
0.1
0.1
−1.95
0.7
0.8
1.95
0.1
0.2
0.0
∅
+
choose η
choose θ
Mixed distribution R
0.2
0.2
−1.13
0.62
0.5
1.95
0.089
0.6
0.0
∅



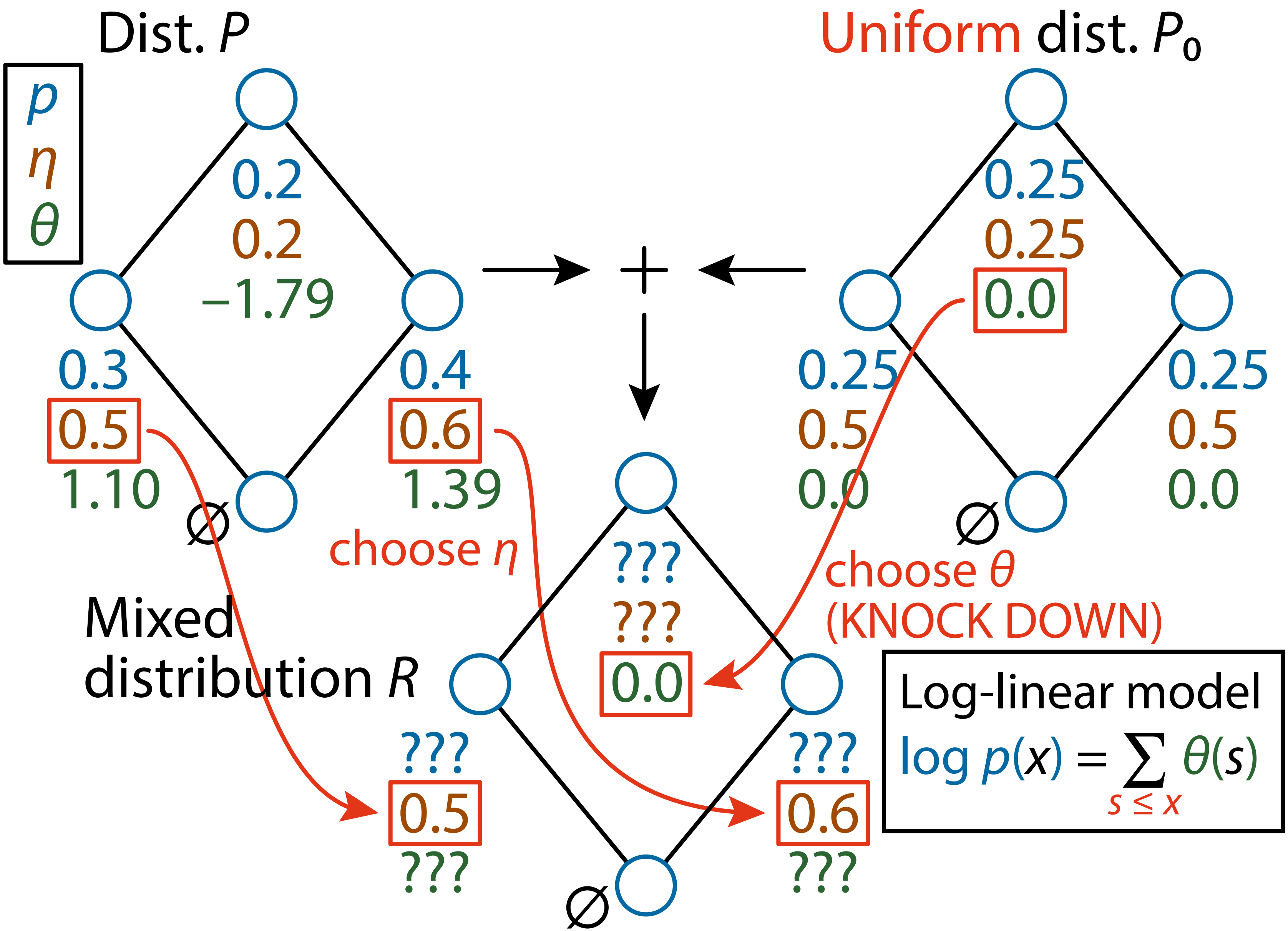
Dist. P
p
η
θ
$KL(P, Q)$
Dist. Q
0.2
0.2
−1.79
0.3
0.5
1.10
0.4
0.6
1.39
∅
0.1
0.1
−1.95
0.7
0.8
1.95
0.1
0.2
0.0
∅
choose η
choose θ
Mixed distribution R
0.2
0.2
−1.13
0.62
0.5
1.95
0.089
0.6
0.0
∅



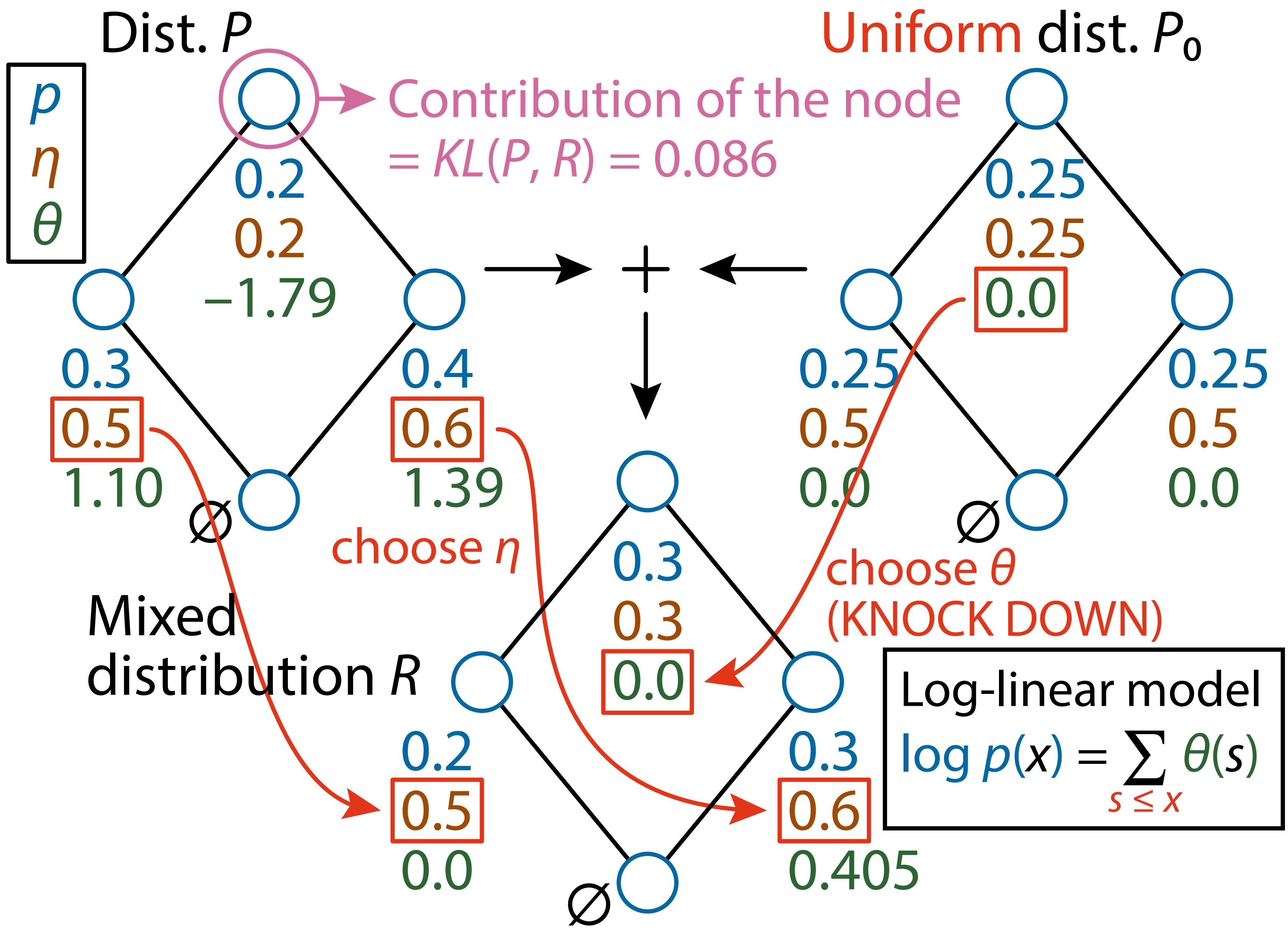
Dist. P
Dist. Q
p
η
θ
KL(P, Q) =
KL(P, R) + KL(R, Q)
Nonnegative decomposition
of the KL divergence
0.2
0.2
0.3
0.5
1.10
0.6
1.39
0.1
0.1
1.95
0.1
0.2
0.0
0.8
1.95
choose η
choose θ
Mixed
distribution R
0.2
0.2
−1.13
0.62
0.5
1.95
0.089
0.6
0.0



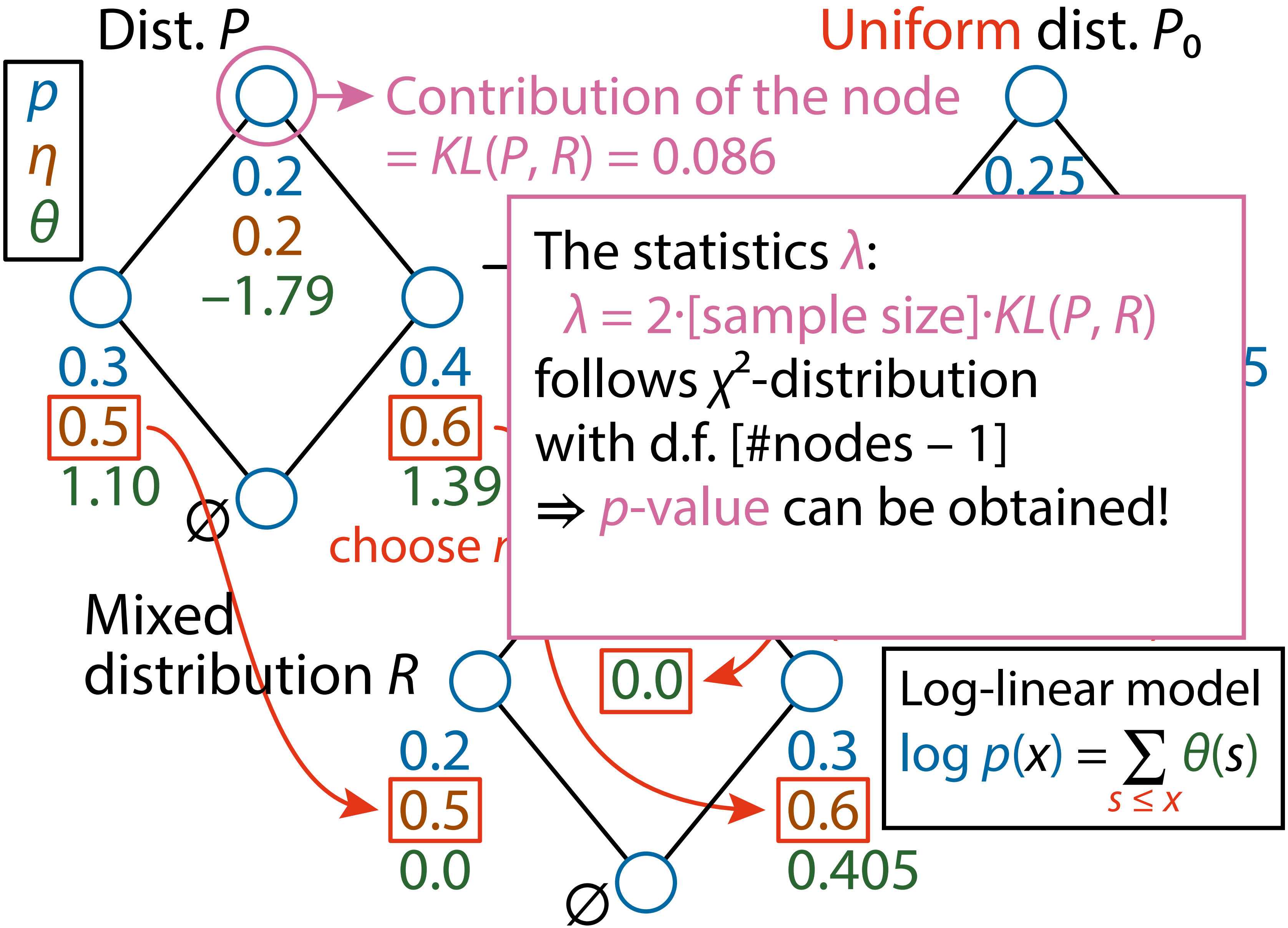
Dist. P
Dist. Q
p
η
θ
0.4390 = 0.3946 + 0.0444
Nonnegative decomposition of the KL divergence
0.2
0.2
0.3
0.5
1.10
0.6
1.39
0.1
0.1
1.95
0.1
0.2
0.0
0.8
1.95
choose η
choose θ
Mixed distribution R
0.2
0.2
−1.13
0.62
0.5
1.95
0.089
0.6
0.0



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
Uniform dist. P_0
0.25
0.25
0.0
0.25
0.5
0.0
0.25
0.5
0.0
∅
choose θ
(KNOCK DOWN)
Mixed distribution R
???
???
0.0
???
0.5
???
???
0.6
???
∅
Log-linear model
$\log p(x) = \sum_{s \le x} \theta(s)$

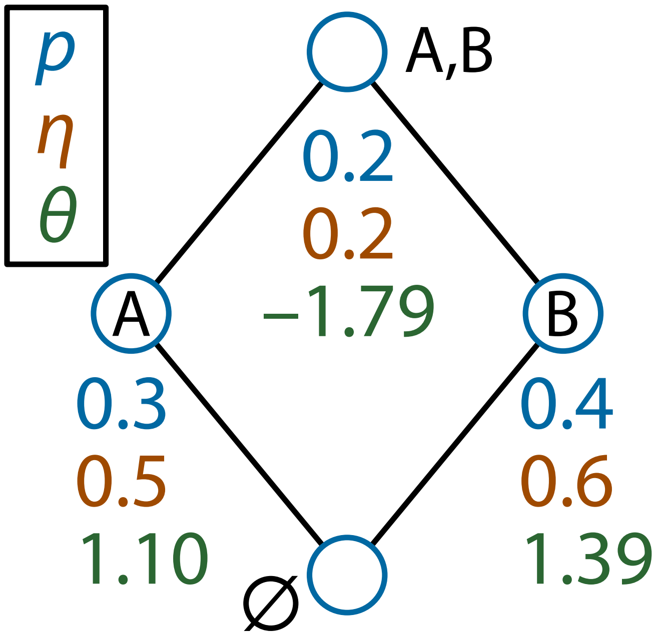


Dist. P
p
η
θ
Contribution of the node
$= KL(P, R) = 0.086$
0.2
0.2
−1.79
0.3
0.5
1.10
0.4
0.6
1.39
∅
Uniform dist. P_0
0.25
0.25
0.0
0.25
0.5
0.0
0.25
0.5
0.0
∅
choose η
choose θ
(KNOCK DOWN)
Mixed
distribution R
0.3
0.3
0.0
0.2
0.5
0.0
0.3
0.6
0.405
∅
Log-linear model
$\log p(x) = \sum_{s \le x} \theta(s)$



Dist. P
p
η
θ
Contribution of the node
= KL(P, R) = 0.086
0.2
0.2
−1.79
0.3
0.5
1.10
0.4
0.6
1.39
∅
choose
Uniform dist. P0
0.25
5
The statistics λ:
λ = 2·[sample size]·KL(P, R)
follows χ2-distribution
with d.f. [#nodes – 1]
⇒ p-value can be obtained!
Mixed
distribution R
0.0
0.2
0.5
0.0
0.3
0.6
0.405
∅
Log-linear model
log p(x) = ∑ θ(s)
s ≤ x

Dist. P



$$
\begin{array}{ccc}
\text{A} & \text{B} & \text{A, B} \\
\end{array}
$$

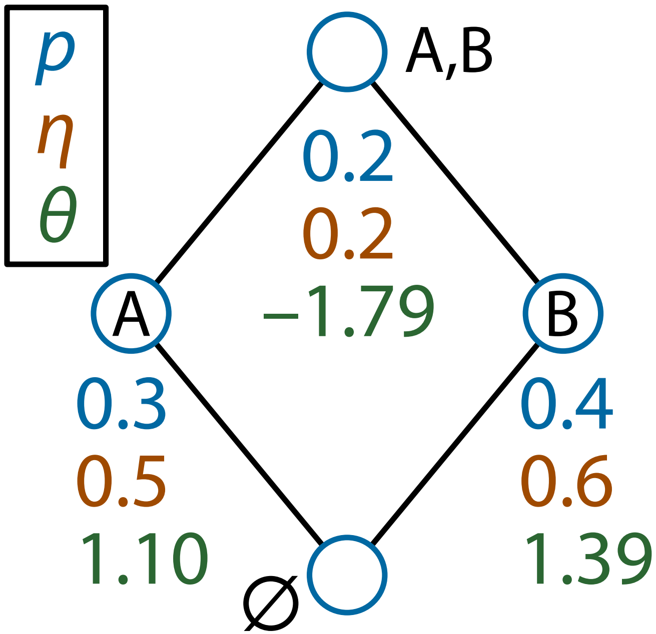
$$
\begin{bmatrix}
0.3 & 0.4 & 0.2 \\
0.5 & 0.6 & 0.2 \\
1.10 & 1.39 & -1.79
\end{bmatrix}
$$

$$
\begin{bmatrix}
??? & ??? & ??? \\
??? & 0.6 & 0.2 \\
0.0 & ??? & ???
\end{bmatrix}
$$

$$
\begin{bmatrix}
??? & ??? & ??? \\
0.5 & ??? & 0.2 \\
??? & 0.0 & ???
\end{bmatrix}
$$

$$
\begin{bmatrix}
??? & ??? & ??? \\
0.5 & 0.6 & ??? \\
??? & ??? & 0.0
\end{bmatrix}
$$

Dist. P



$$\begin{array}{c} \begin{array}{ccc} \text{A} & \text{B} & \text{A, B} \end{array} \\ \left[\begin{array}{ccc} 0.3 & 0.4 & 0.2 \\ 0.5 & 0.6 & 0.2 \\ 1.10 & 1.39 & -1.79 \end{array}\right] \end{array}$$

$$\left[\begin{array}{ccc} ??? & ??? & ??? \\ ??? & 0.6 & 0.2 \\ 0.0 & ??? & ??? \end{array}\right]$$

KL = Score of A

$$\left[\begin{array}{ccc} ??? & ??? & ??? \\ 0.5 & ??? & 0.2 \\ ??? & 0.0 & ??? \end{array}\right]$$

KL = Score of B

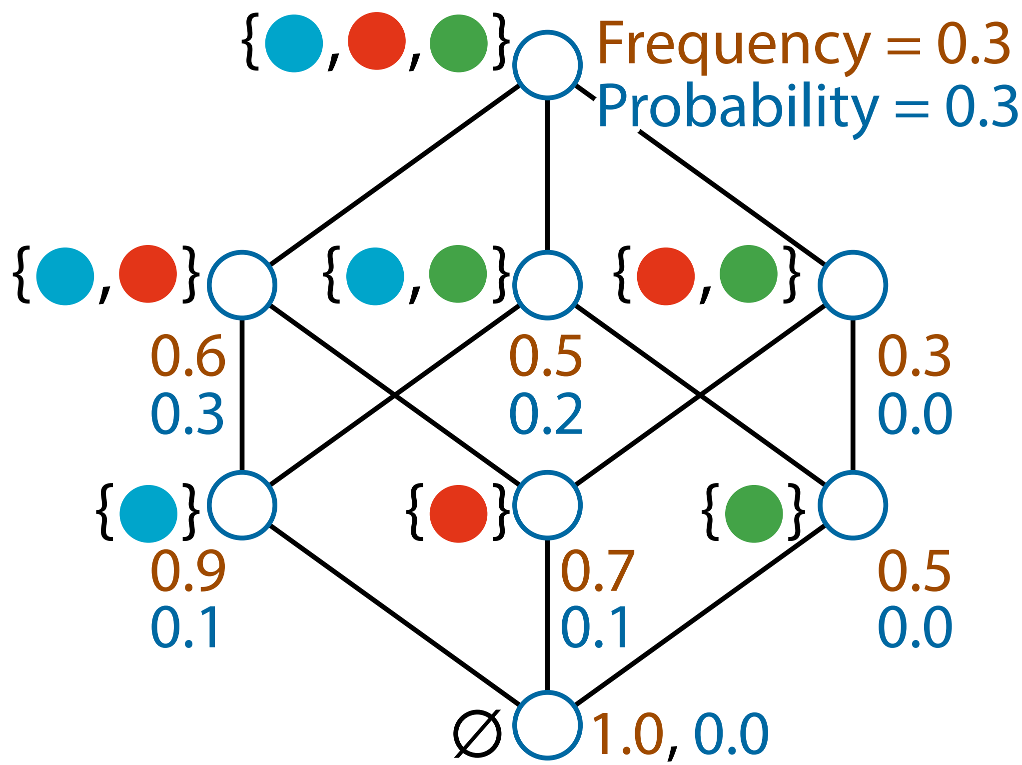
$$\left[\begin{array}{ccc} ??? & ??? & ??? \\ 0.5 & 0.6 & ??? \\ ??? & ??? & 0.0 \end{array}\right]$$

KL = Score of A, B

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

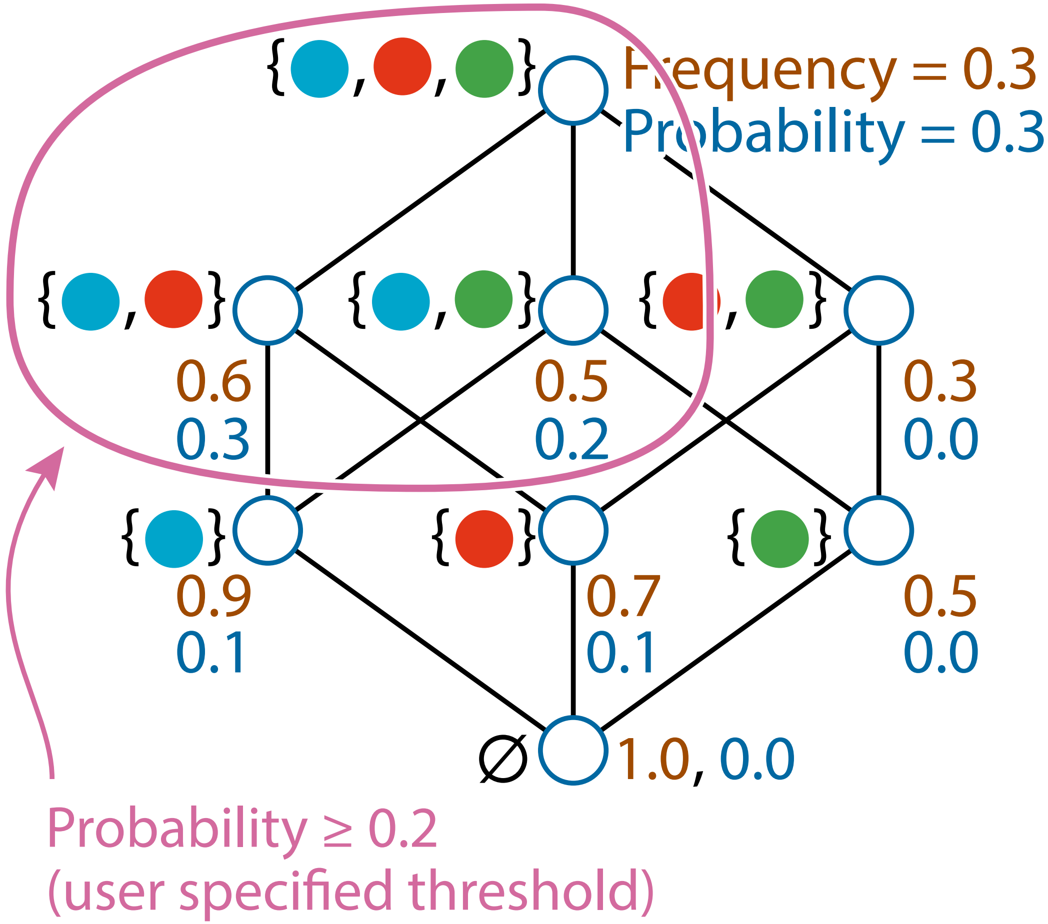


Number of nodes = $2^{\#\text{features}}$
⇒combinatorial explosion!

Make a Poset from Data

Dataset

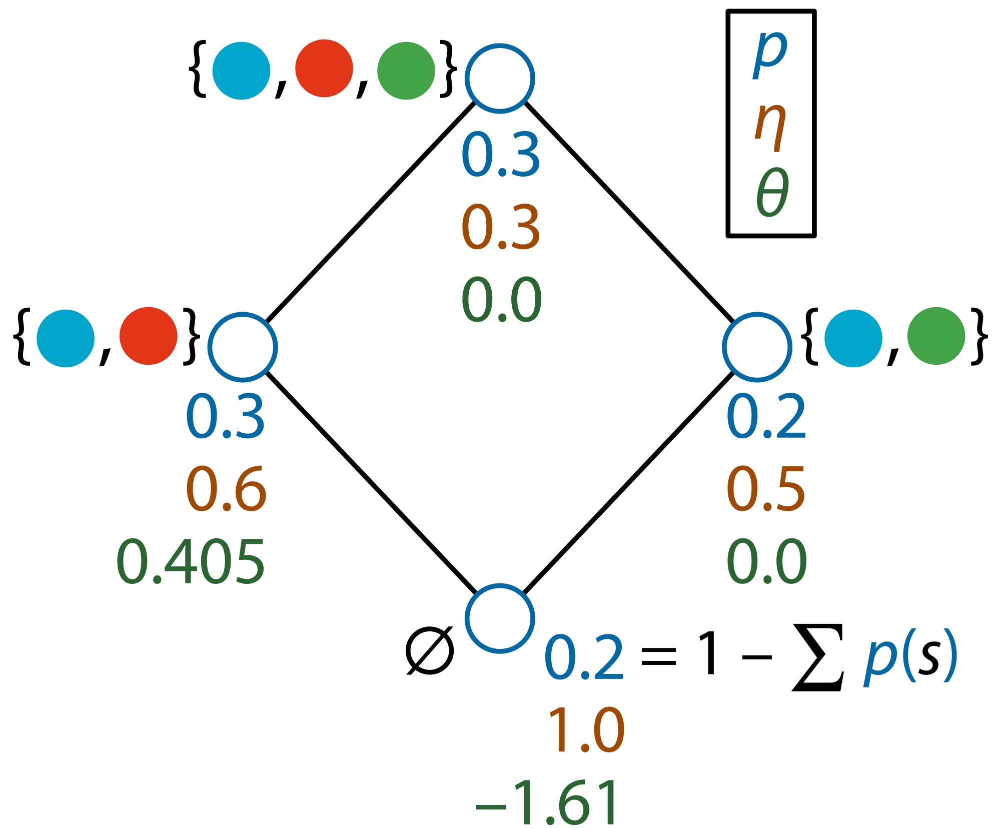
ID 1:	1	1	0
ID 2:	1	1	1
ID 3:	1	1	0
ID 4:	1	1	1
ID 5:	1	1	0
ID 6:	1	0	1
ID 7:	1	0	1
ID 8:	1	1	1
ID 9:	1	0	0
ID10:	0	1	0



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

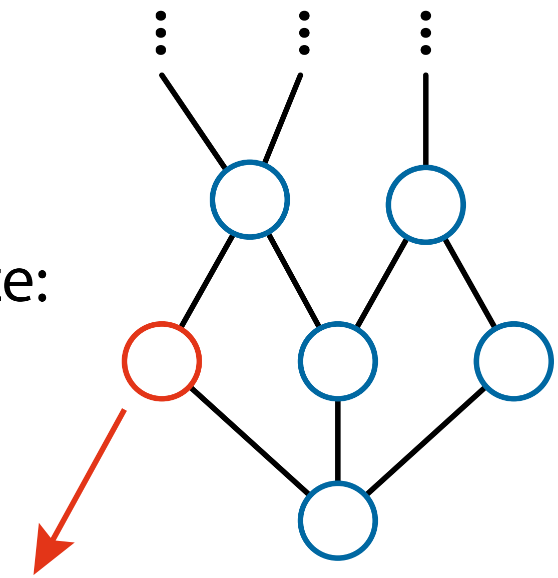
⋮ ⋮

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

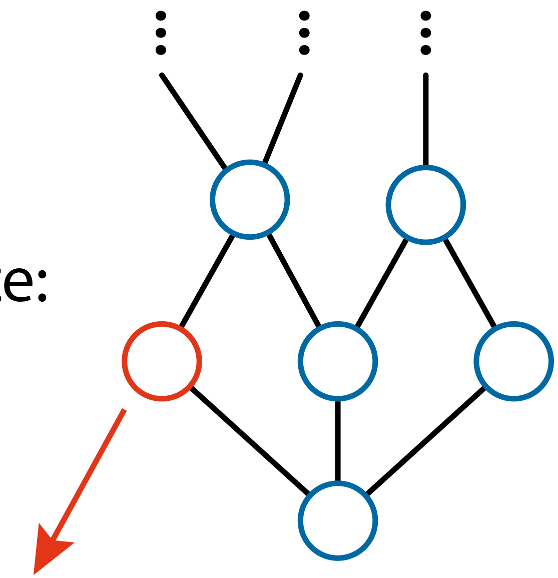
Example on Real Data (accidents)

features: 468

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

- We build information geometry for posets (partially ordered sets)
  - Natural connection between the information geometric dual coordinates and the partial order structure
    - M. Sugiyama, H. Nakahara, K. Tsuda, *Information Decomposition on Structured Space*, arXiv:1601.05533 (2016)
    - S. Amari, *Information geometry on hierarchy of probability distributions*, IEEE Trans. Info. Theory (2001)
- We can decompose a probability distribution and asses the significance of any-order interactions beyond pairwise interactions