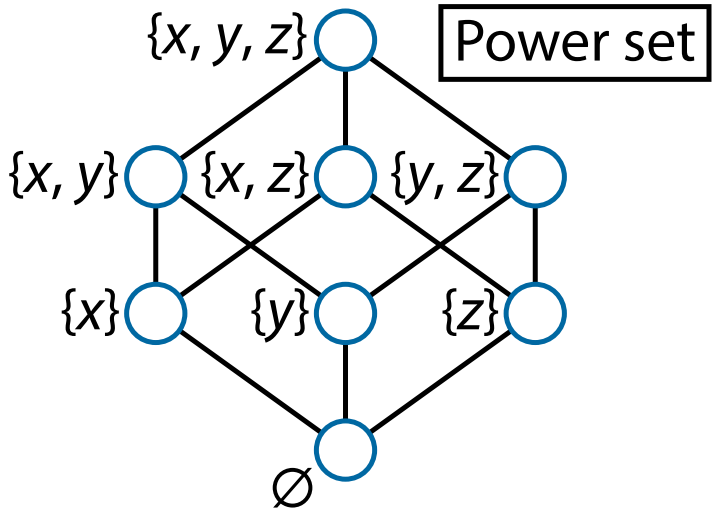
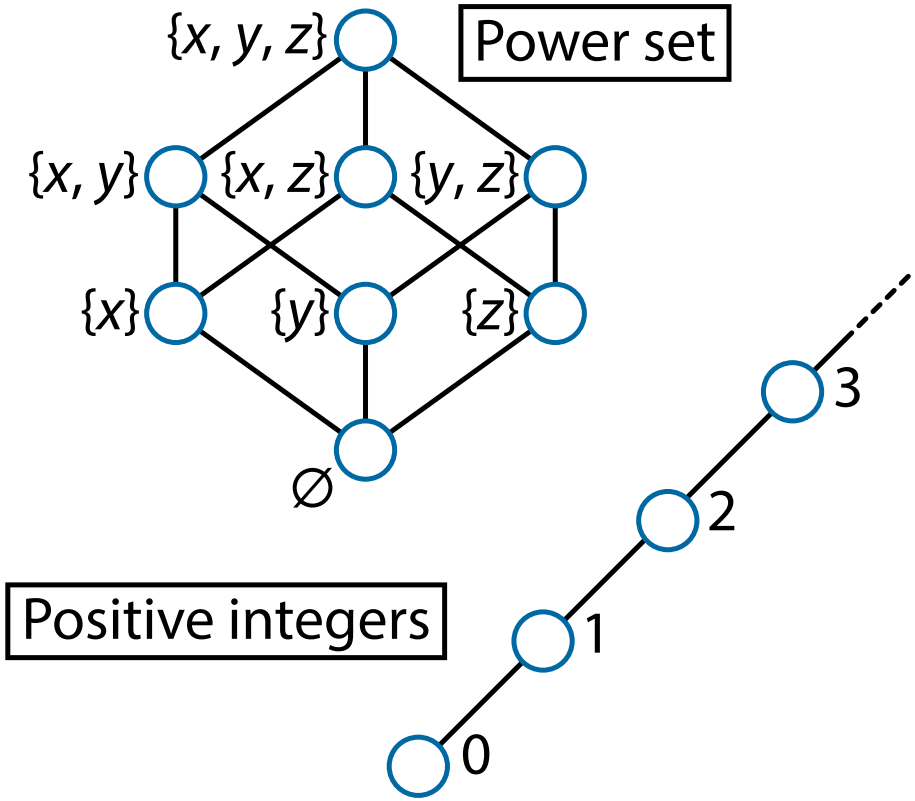
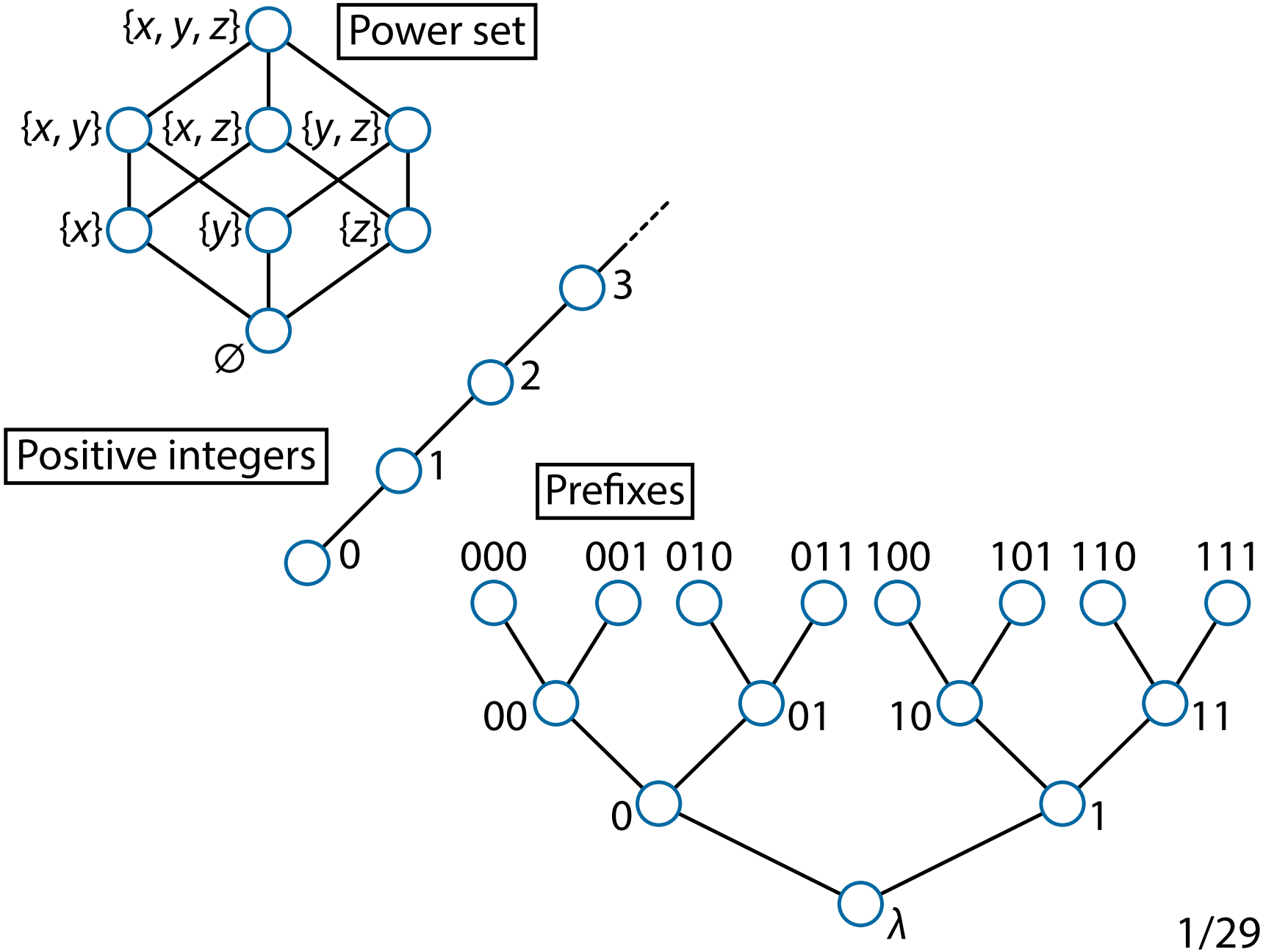
July 11, 2016
ISIT 2016

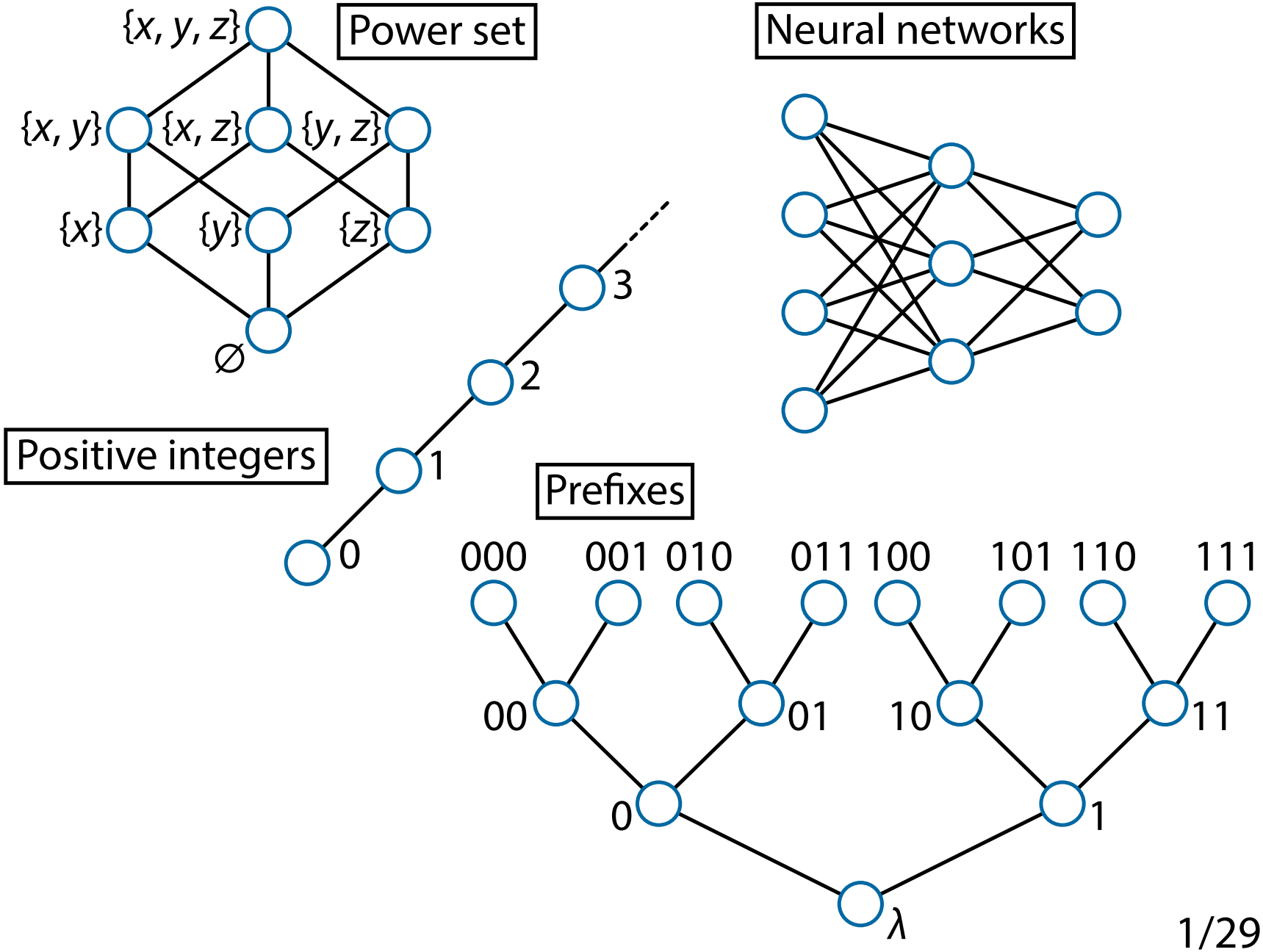
Information Decomposition on Structured Space

Mahito Sugiyama (Osaka Univ.)
Hiroyuki Nakahara (RIKEN), Koji Tsuda (UTokyo)


Power set
{x, y, z}
{x, y}
{x, z}
{y, z}
{x}
{y}
{z}
∅

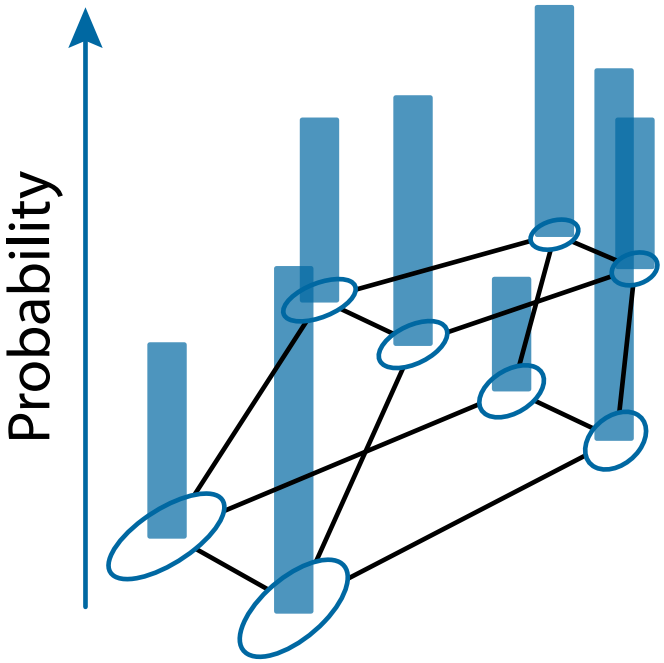

Power set
{x, y, z}
{x, y}
{x, z}
{y, z}
{x}
{y}
{z}
∅
Positive integers
0
1
2
3


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


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

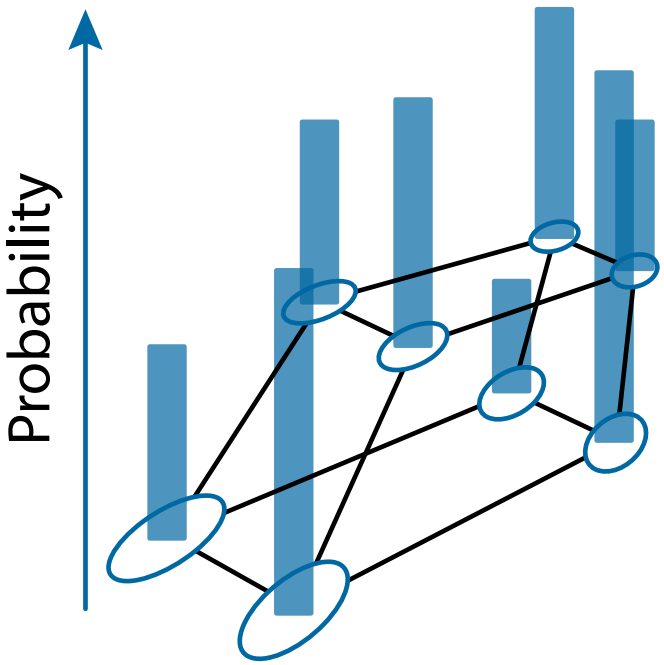
Summary

Probability distribution
on posets (partially ordered sets)

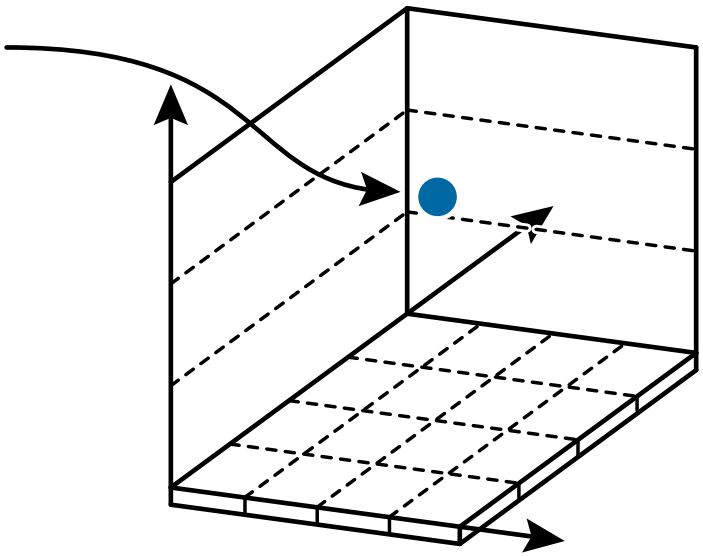


Summary

Probability distribution
on posets (partially ordered sets)

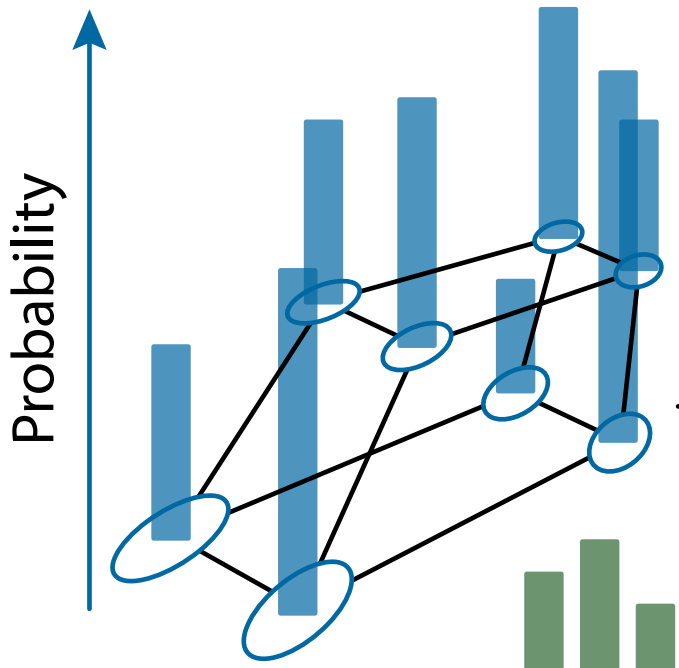


Information
geometry

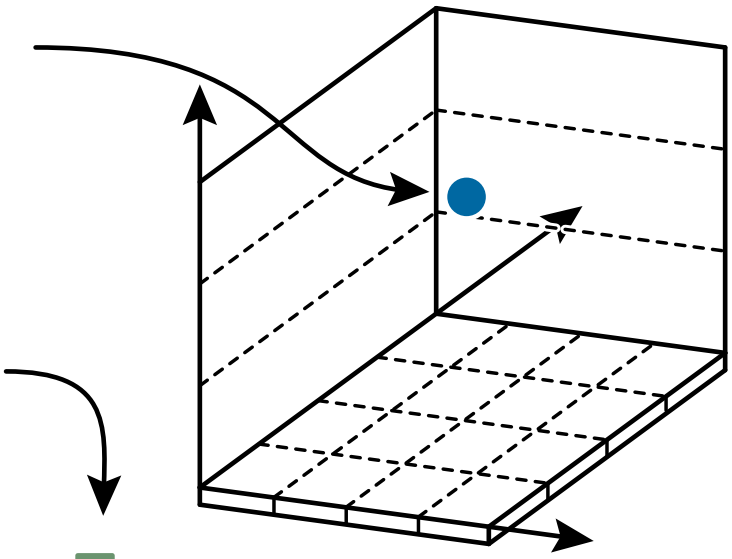


Summary

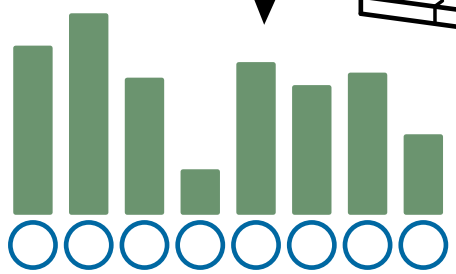
Probability distribution
on posets (partially ordered sets)



Information
geometry



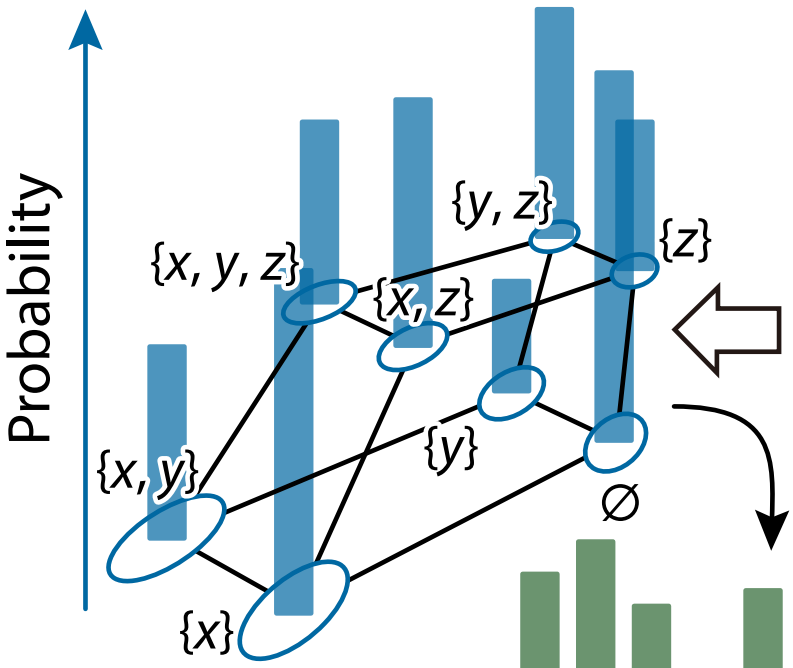
Decomposition in
the log-linear model



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

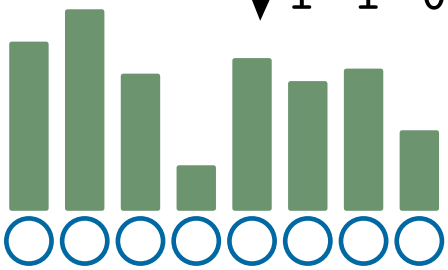
Summary

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

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

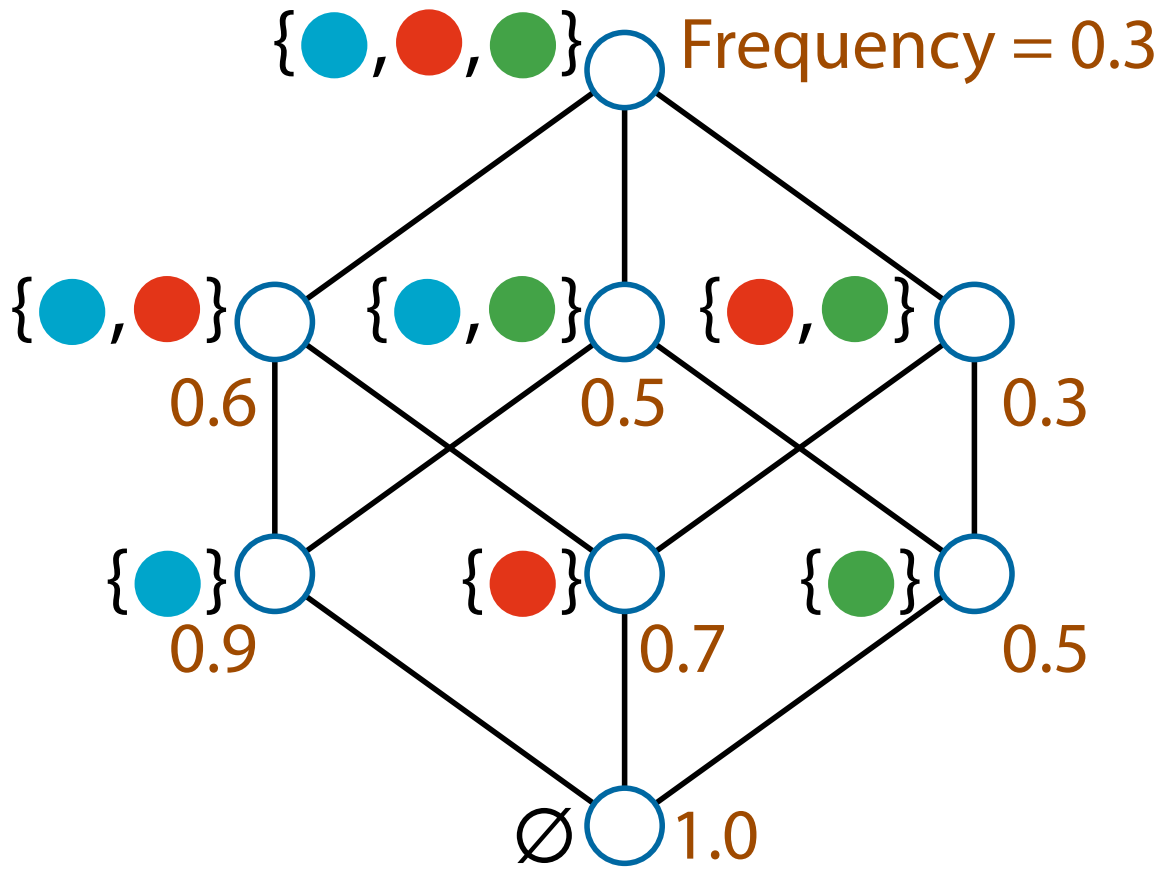
Transaction database

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

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

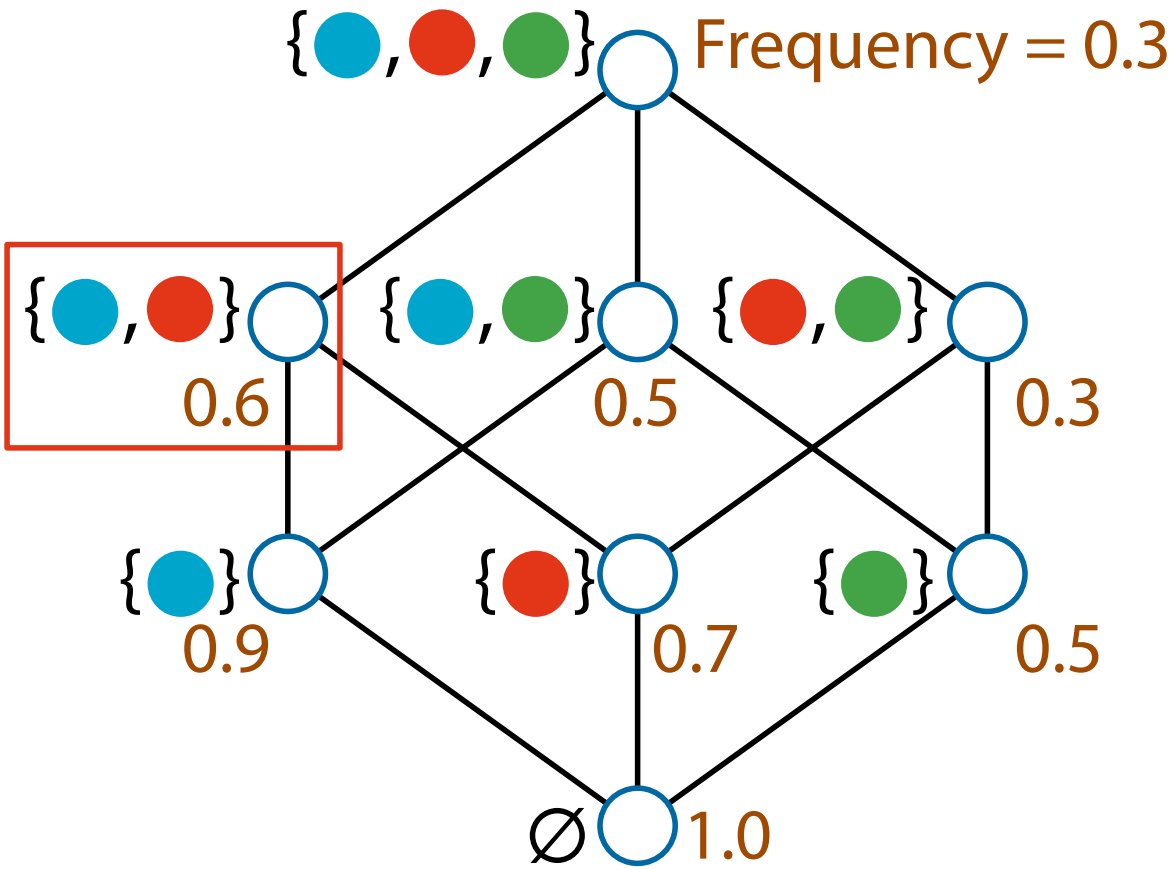
Itemset lattice



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

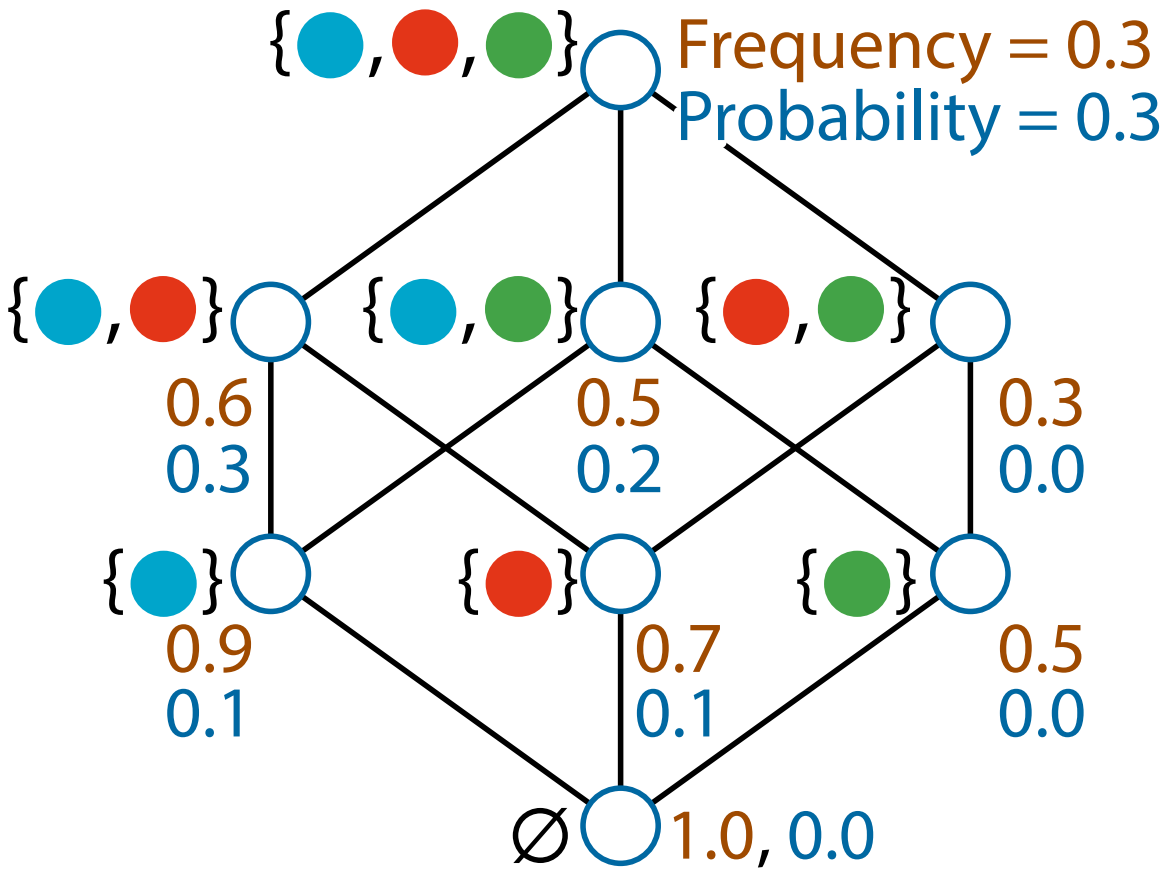
Itemset lattice



Transaction database

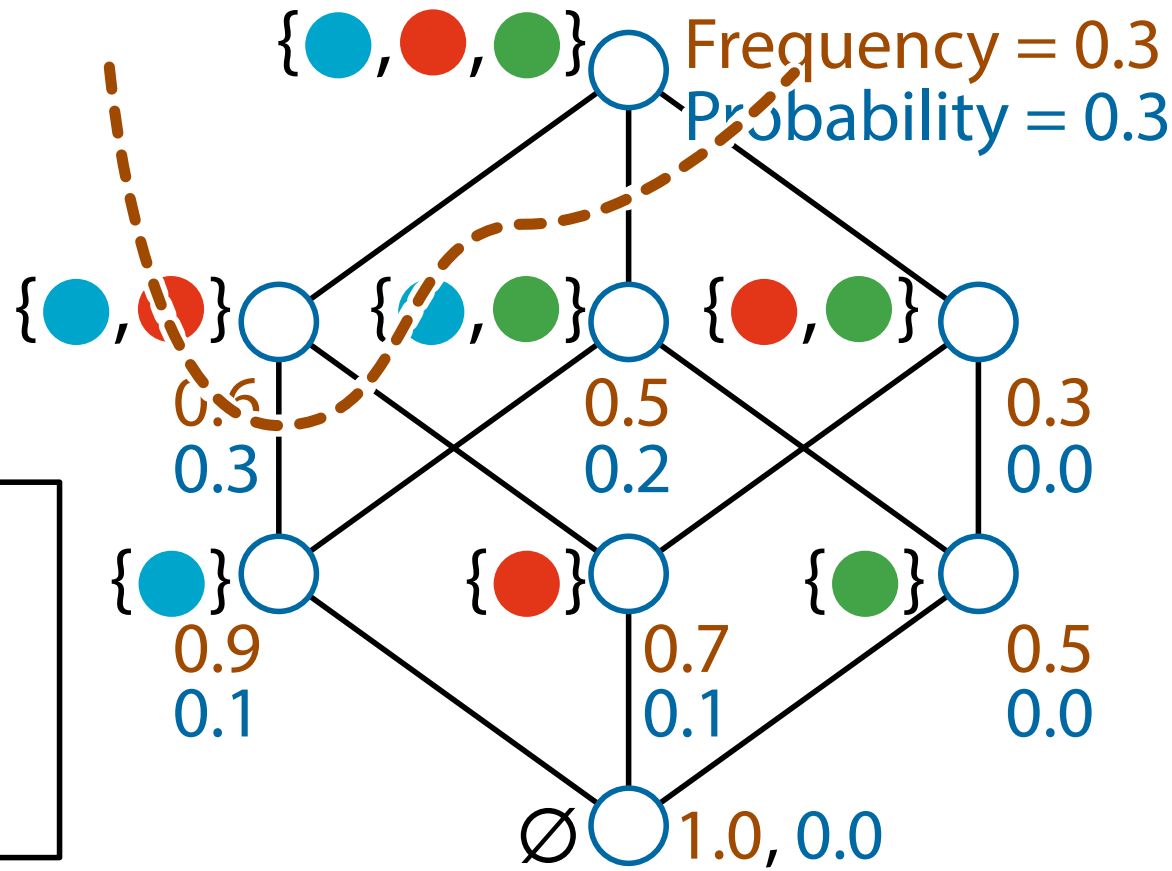
ID 1:	1	1	0
ID 2:	1	1	1
ID 3:	1	1	0
ID 4:	1	1	1
ID 5:	1	1	0
ID 6:	1	0	1
ID 7:	1	0	1
ID 8:	1	1	1
ID 9:	1	0	0
ID10:	0	1	0

Itemset lattice



Upward = Pattern mining

Itemset lattice



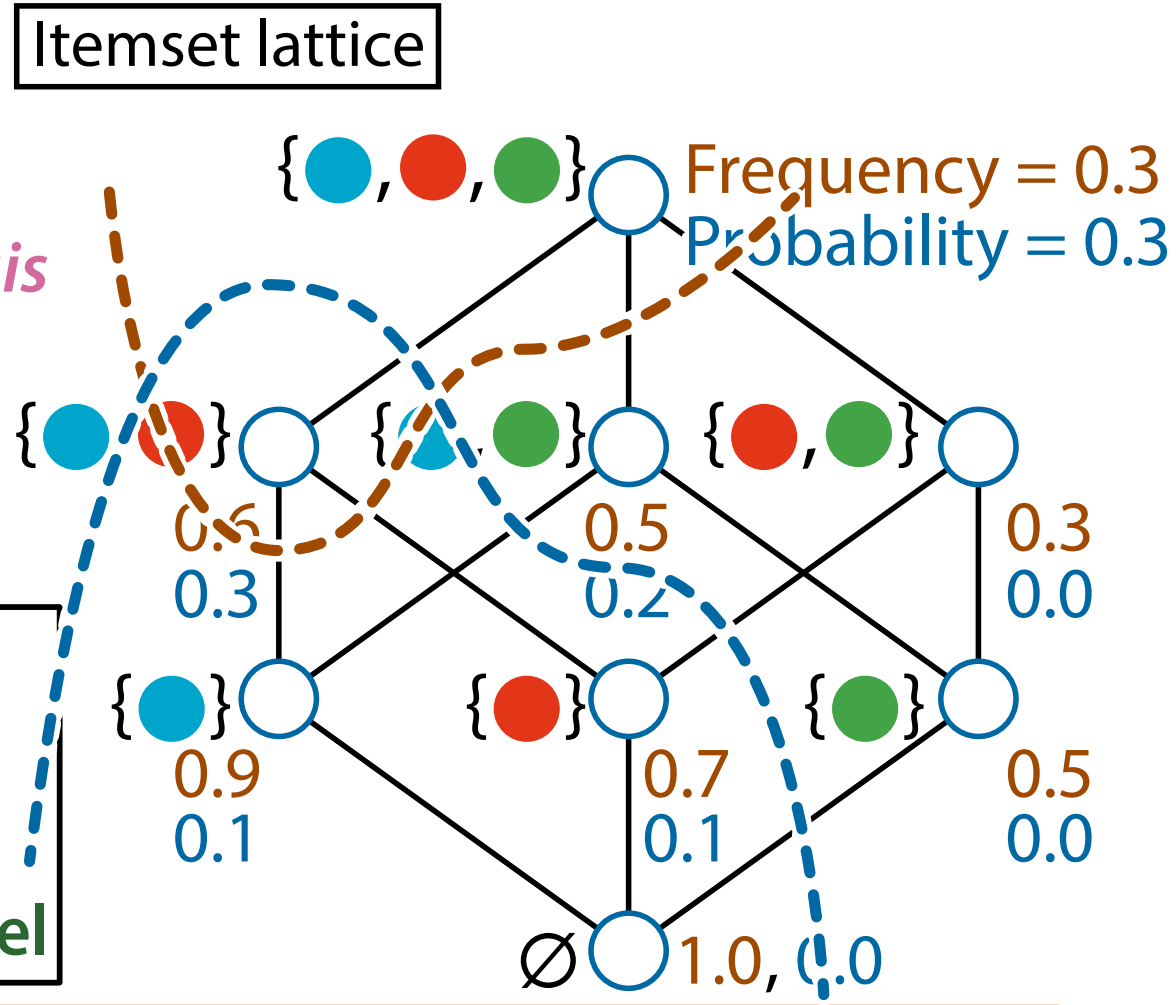
η: Frequency

p: Probability

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

Upward = Pattern mining

Downward = Log-linear analysis



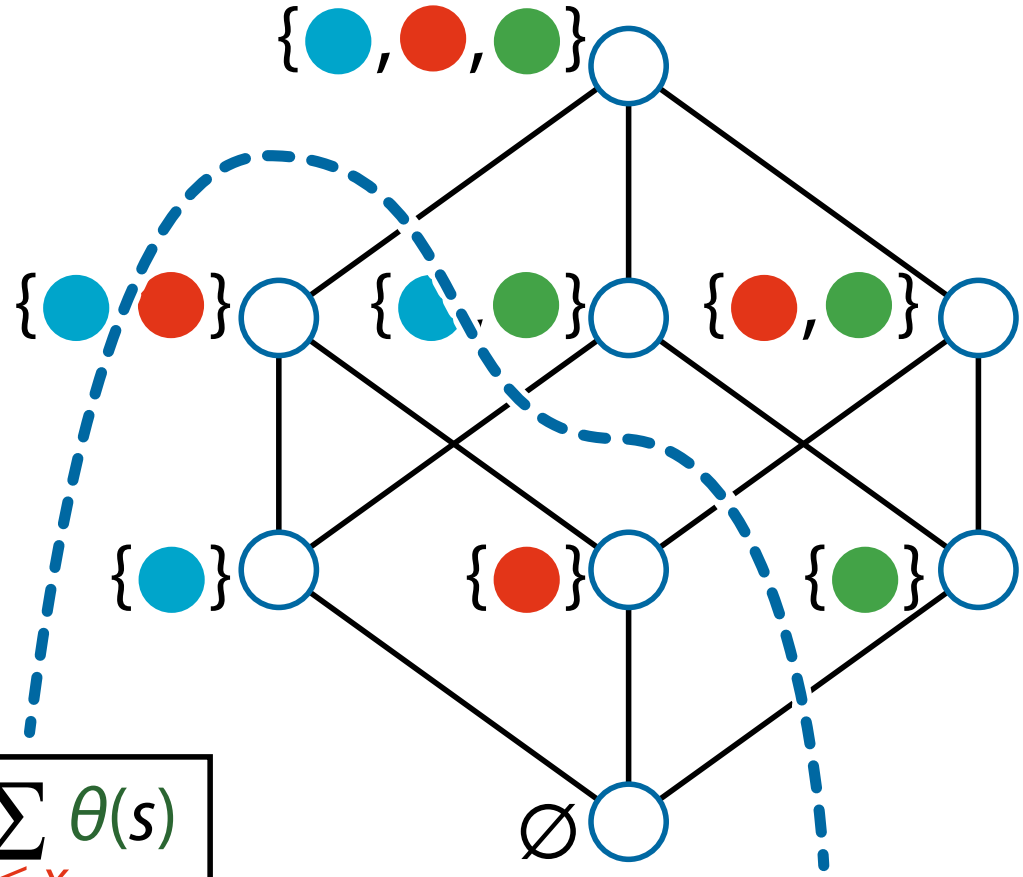
η: Frequency

p: Probability

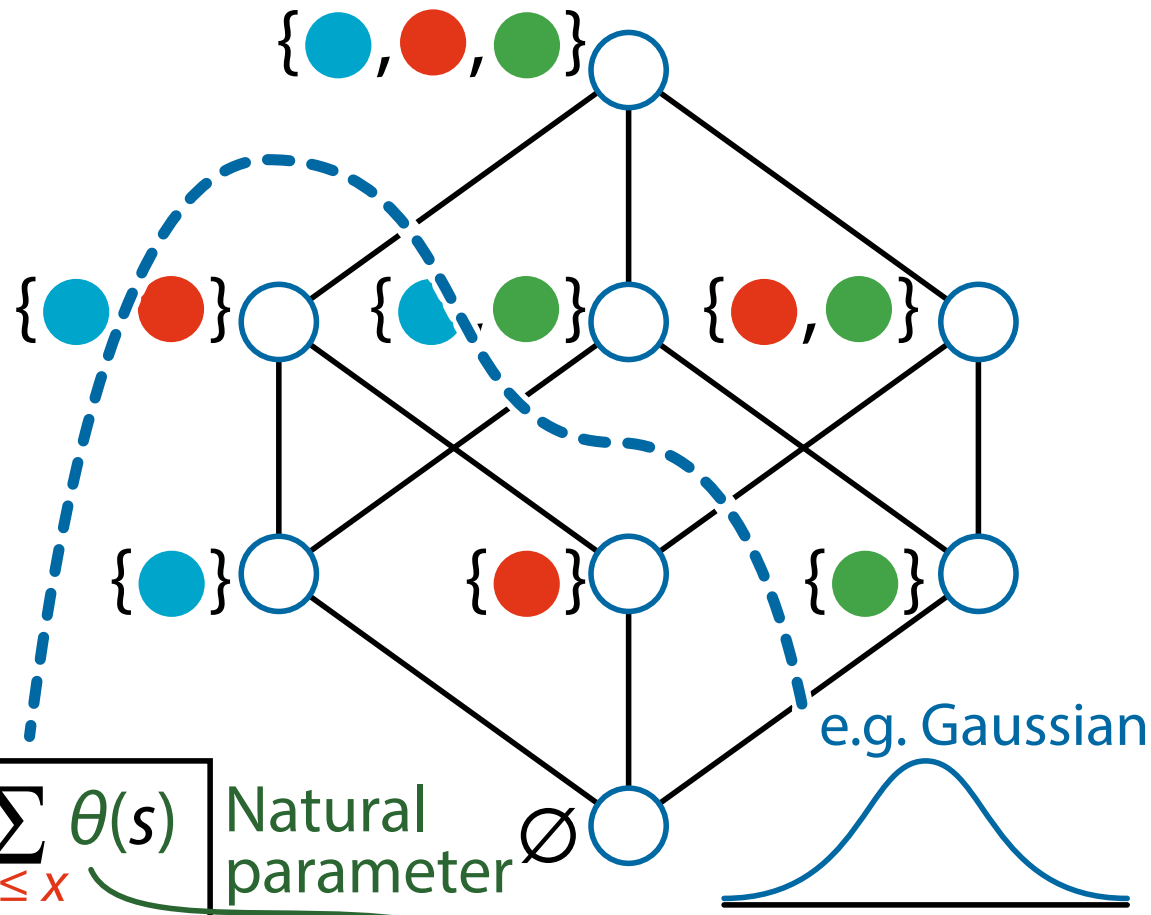
θ: Coefficient of **log-linear model**

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

$$\log p(\{●,●\}) = \theta(\{●,●\}) + \theta(\{●\}) + \theta(\{●\}) + \theta(\varnothing)$$



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



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

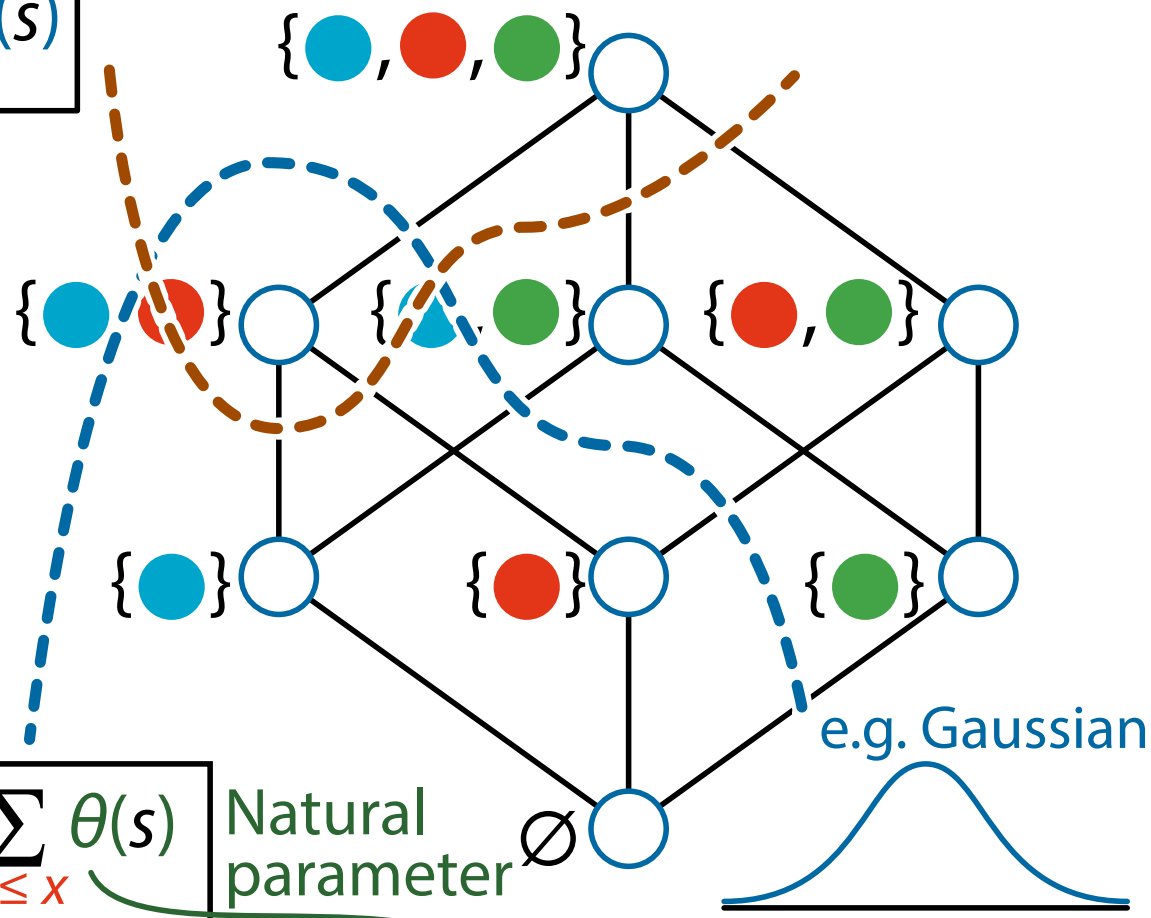
Natural parameter

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

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

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

Sufficient statistics of exponential family

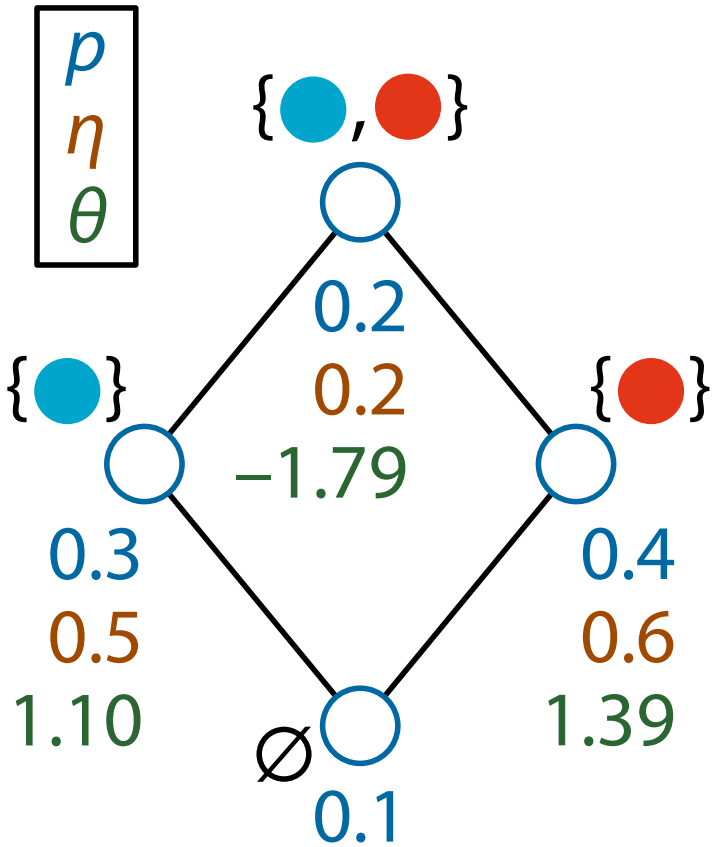


$$\log p(x) = \sum_{s \leq x} \theta(s)$$

Natural parameter

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

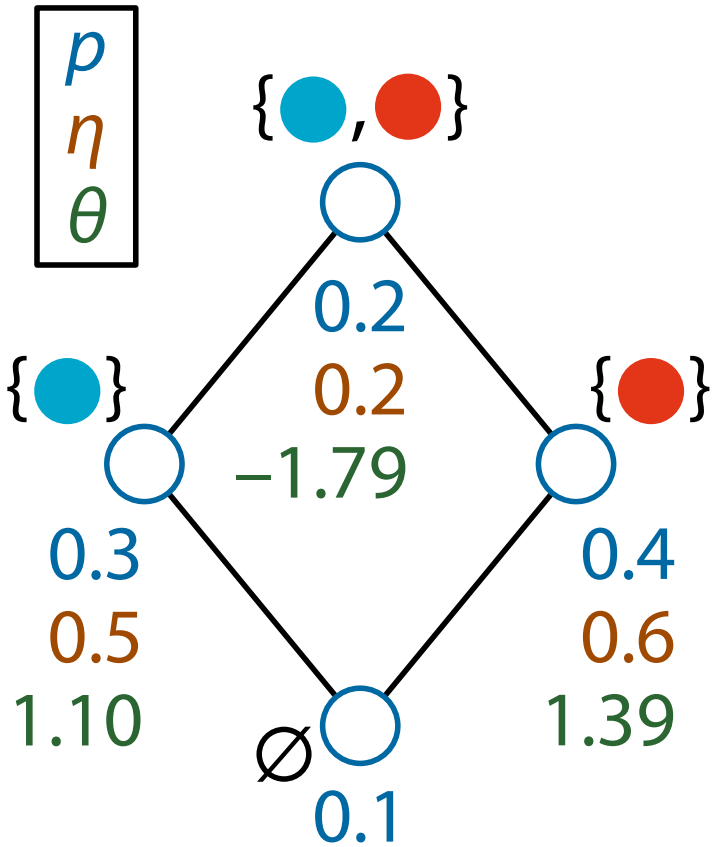
Triple for each node

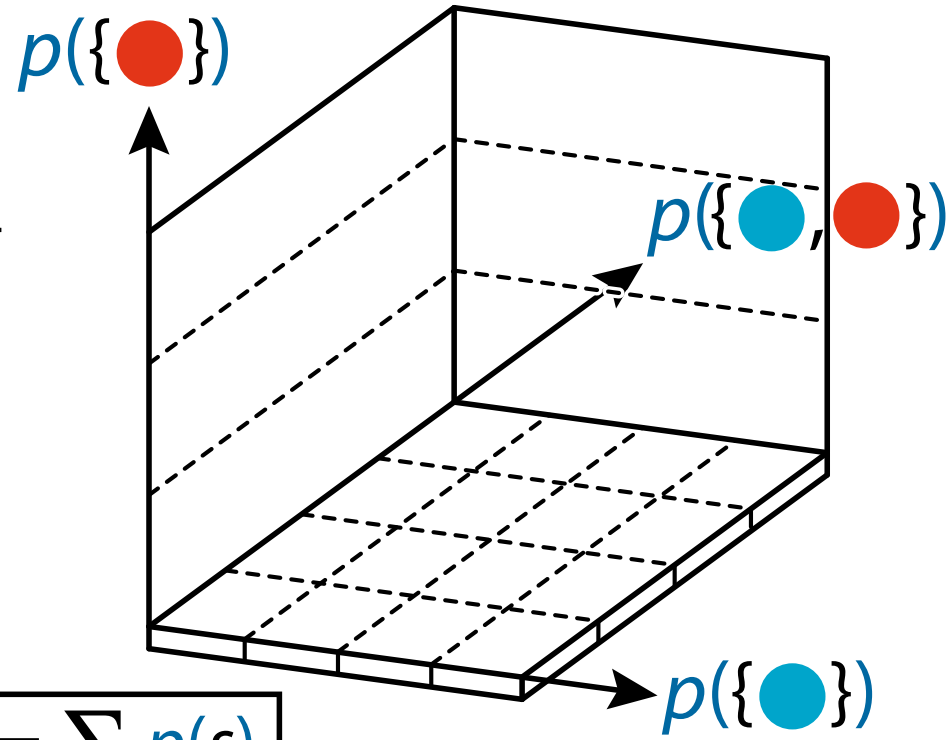


$$\eta(x) = \sum_{s \geq x} p(s)$$

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

Triple for each node

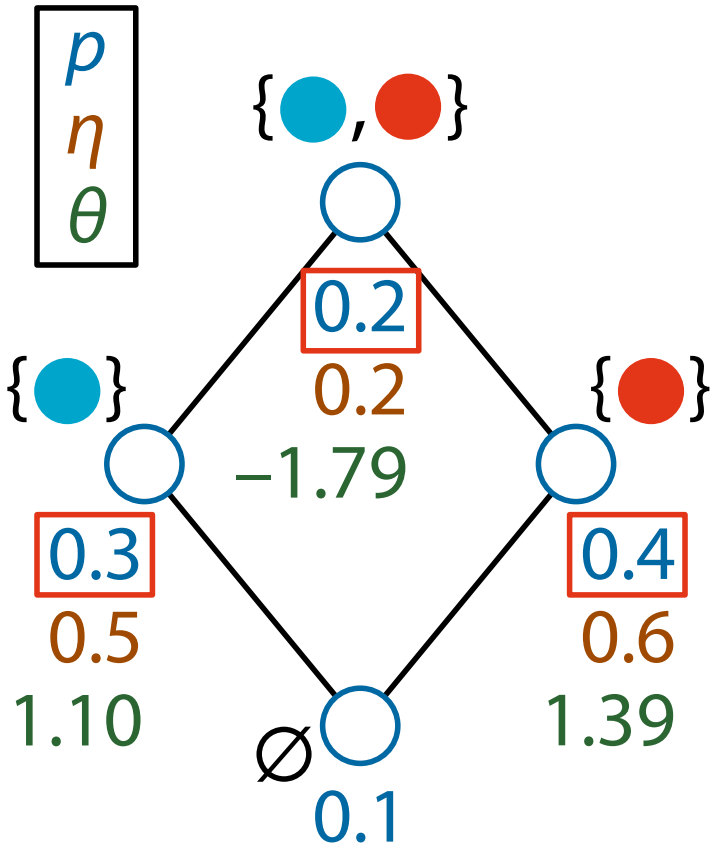




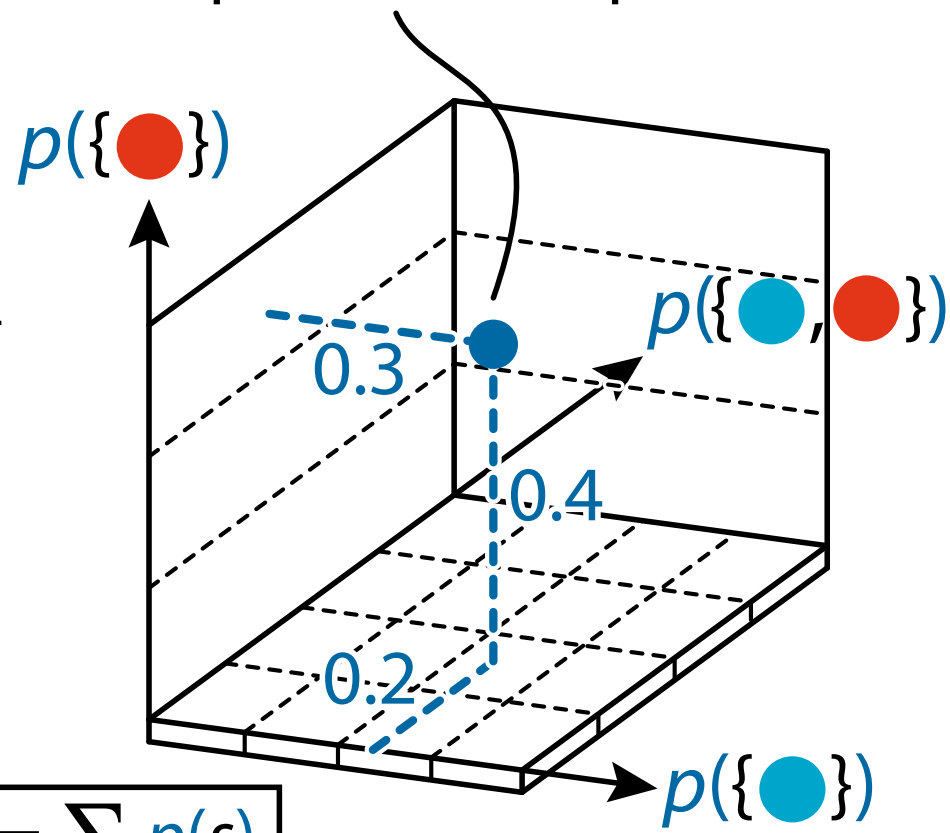
$$\eta(x) = \sum_{s \geq x} p(s)$$

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

Triple for each node



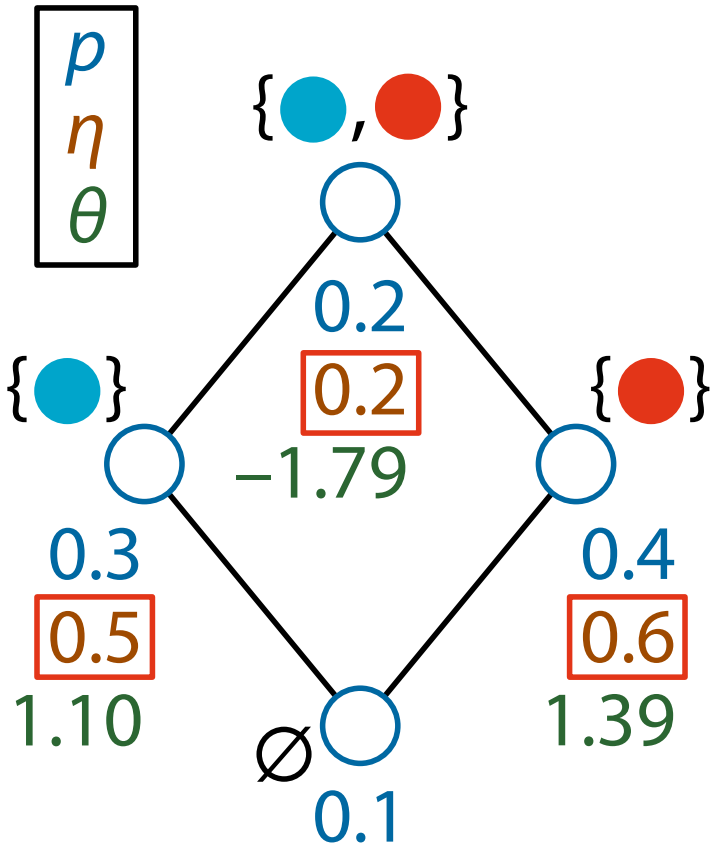
Probability distribution is a "point" in 3D space



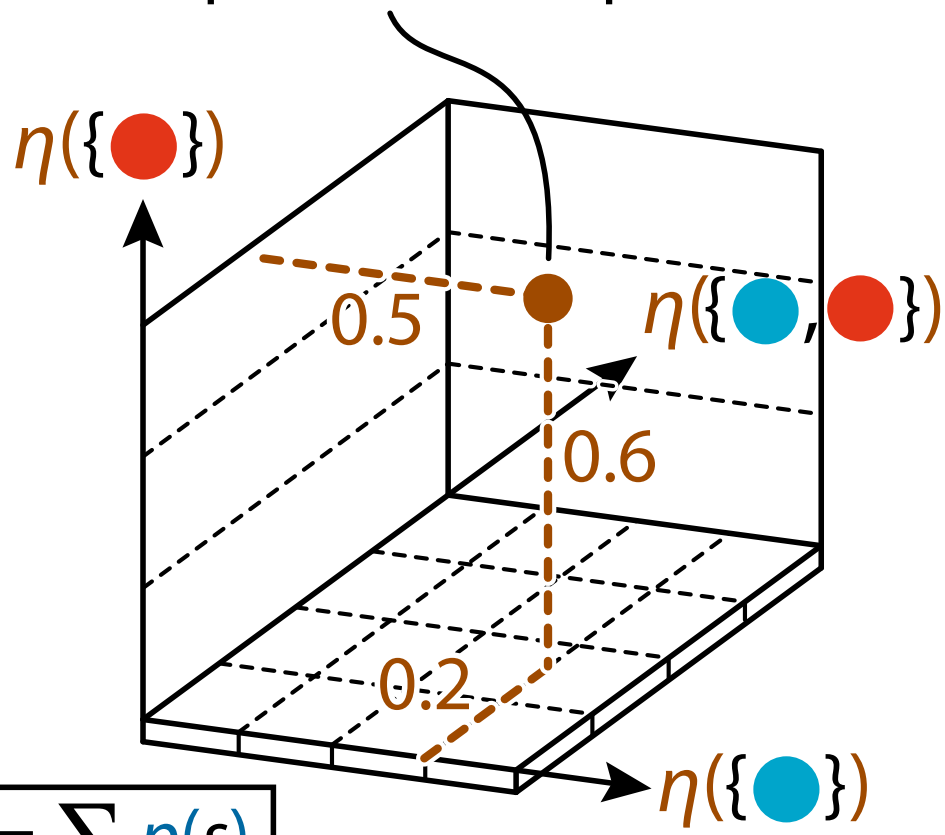
$$\eta(x) = \sum_{s \geq x} p(s)$$

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

Triple for each node



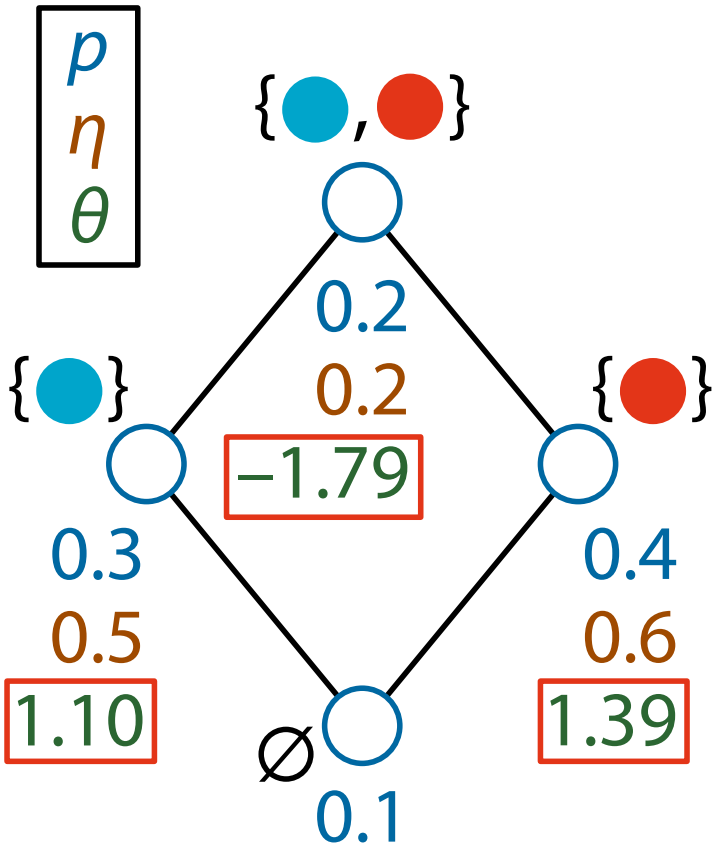
Probability distribution is a "point" in 3D space



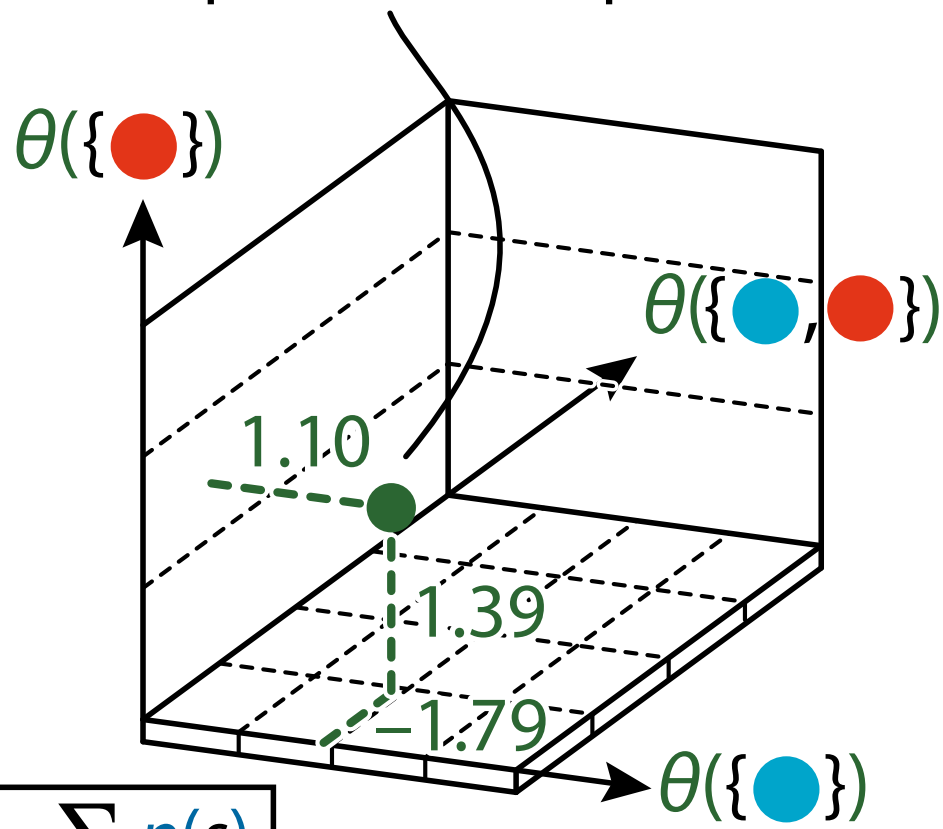
$$\eta(x) = \sum_{s \geq x} p(s)$$

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

Triple for each node



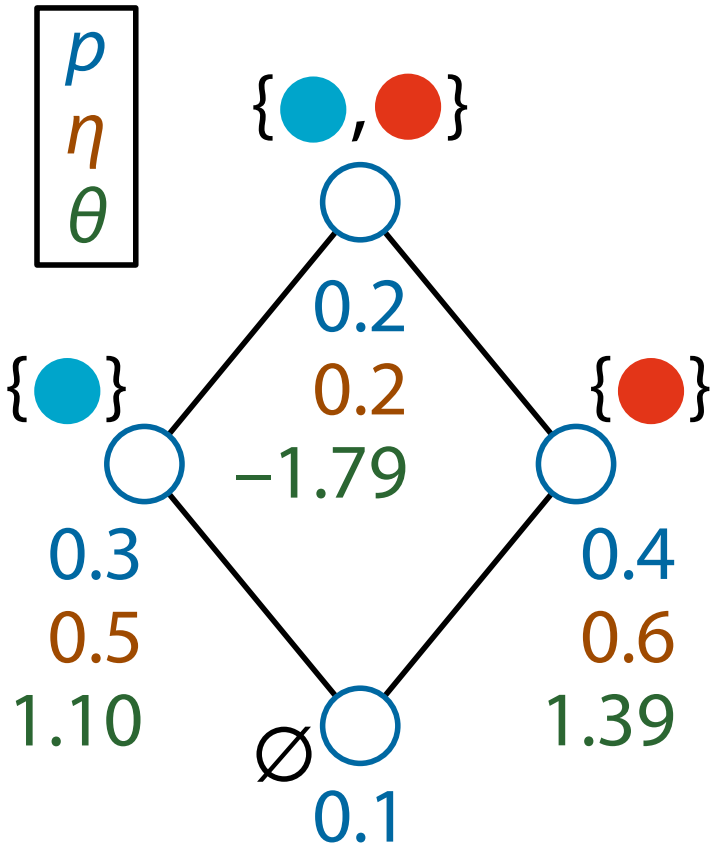
Probability distribution is a "point" in 3D space

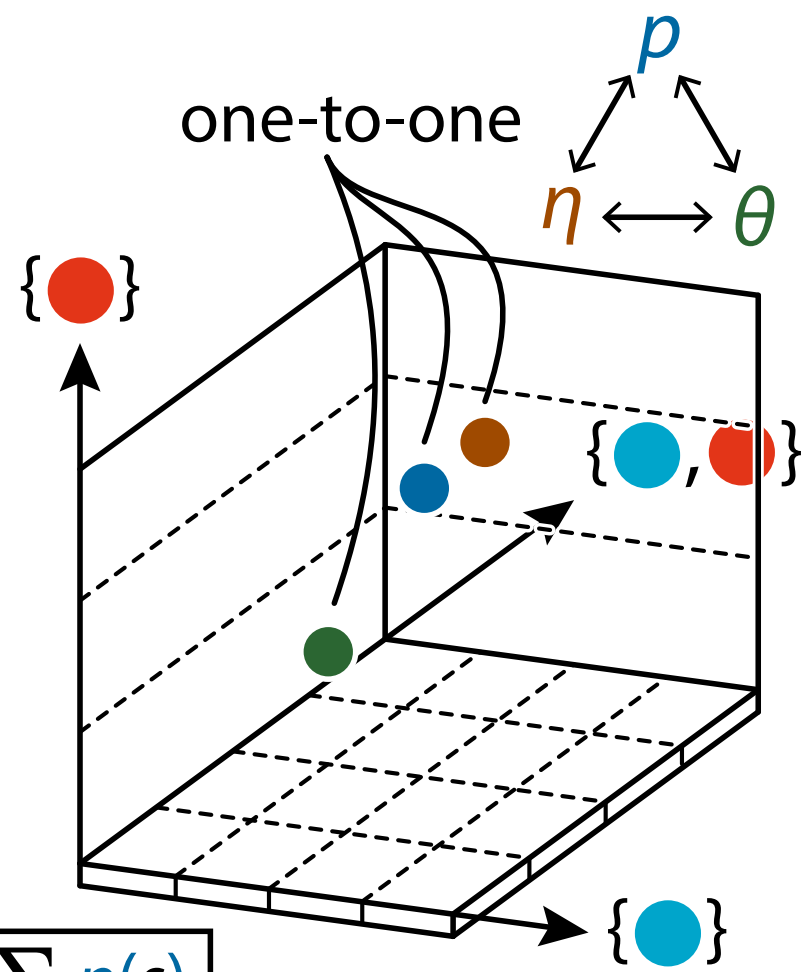


$$\eta(x) = \sum_{s \geq x} p(s)$$

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

Triple for each node

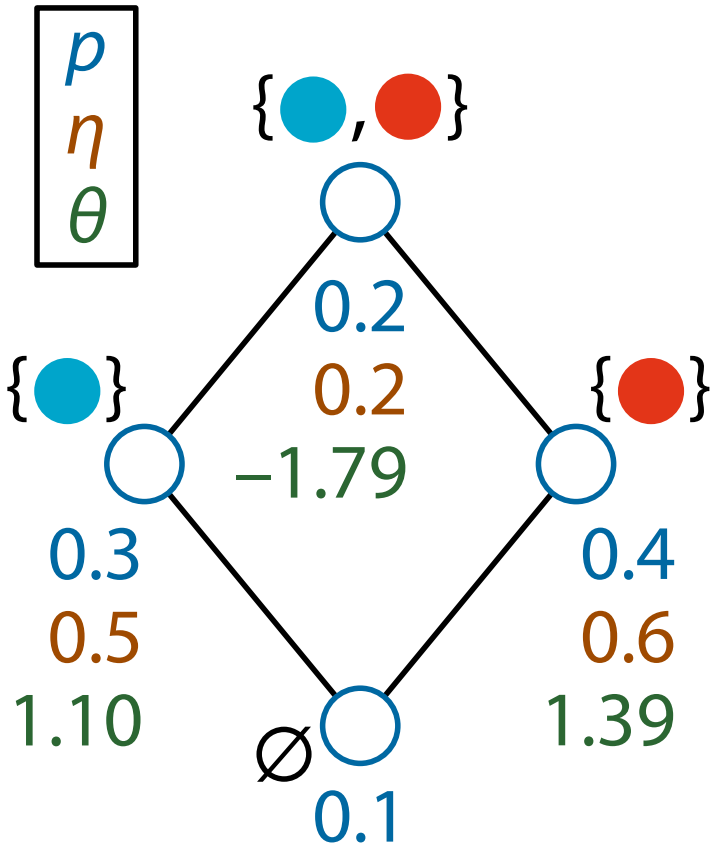




$$\eta(x) = \sum_{s \geq x} p(s)$$

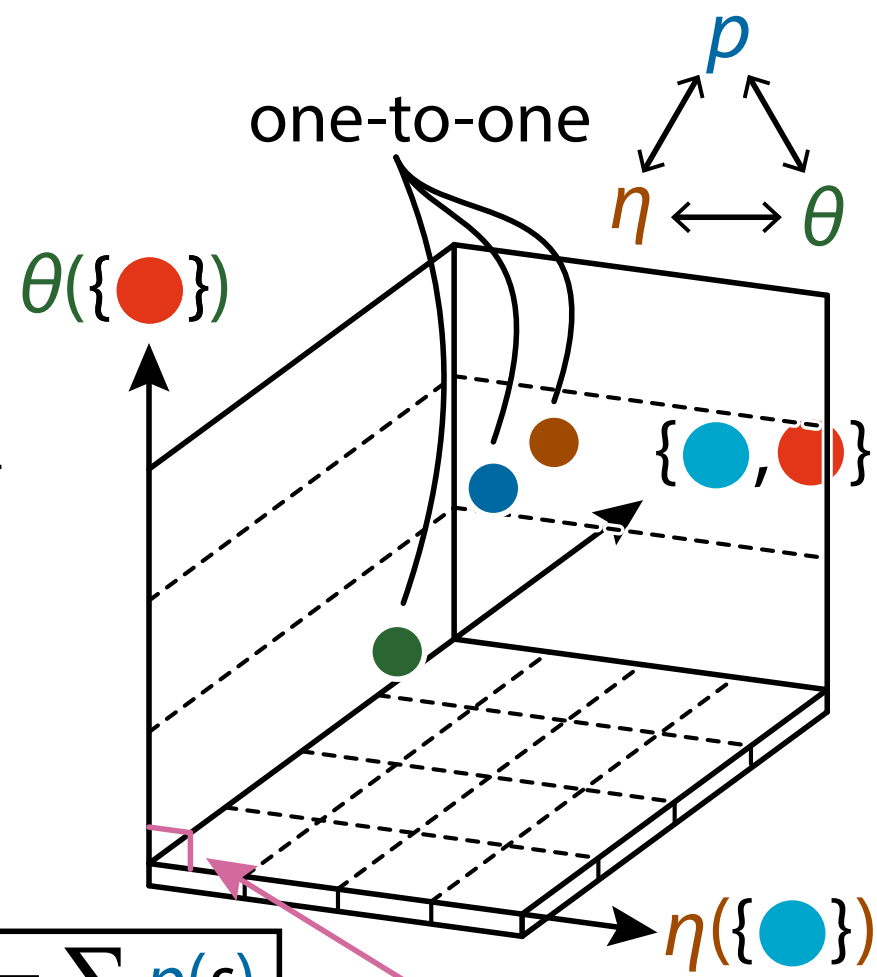
$$\log p(x) = \sum_{s \leq x} \theta(s)$$

Triple for each node

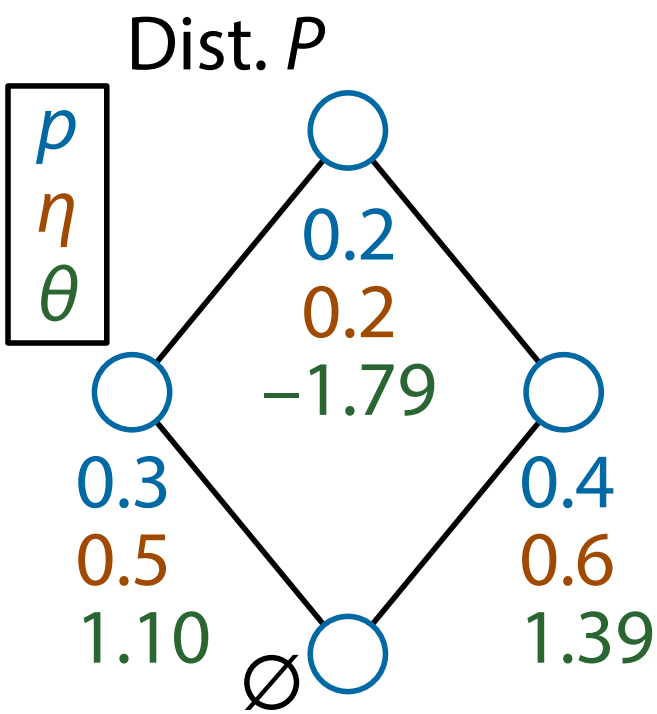


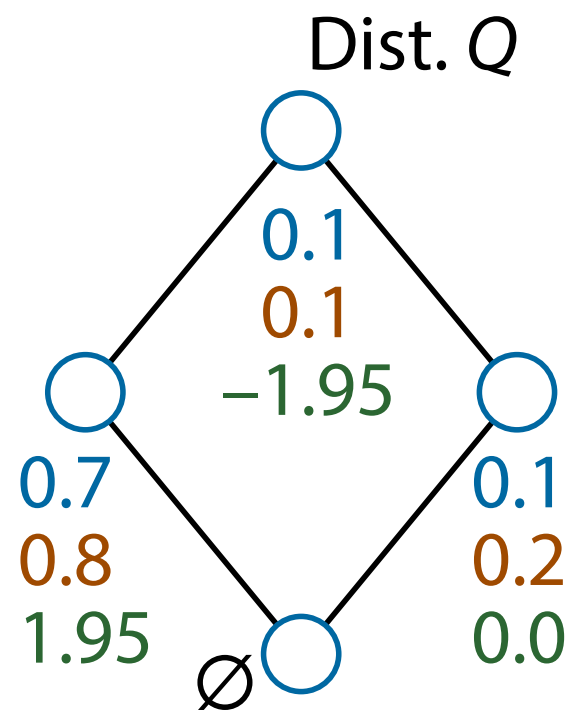
$$\eta(x) = \sum_{s \geq x} p(s)$$

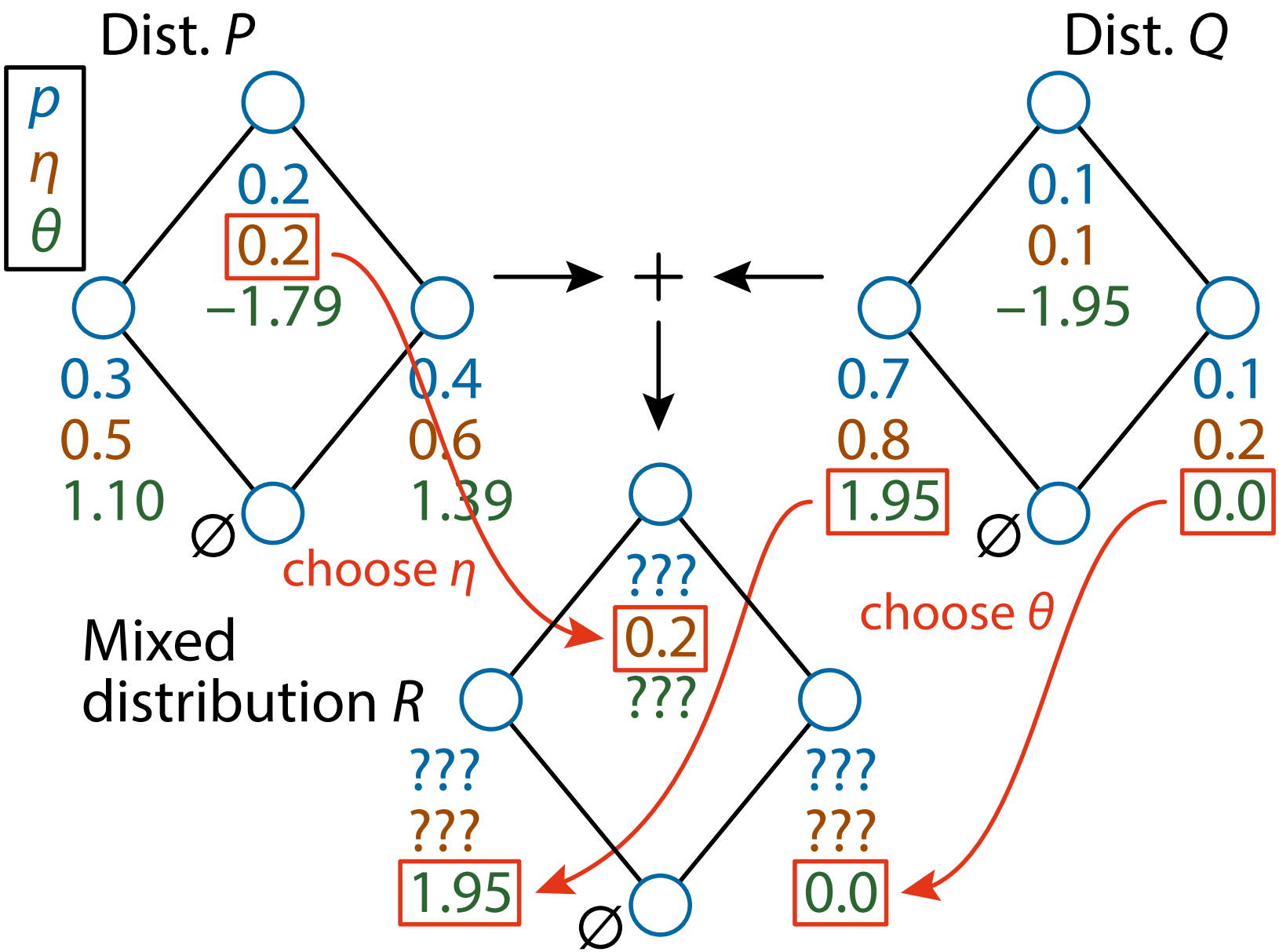
$$\log p(x) = \sum_{s \leq x} \theta(s)$$

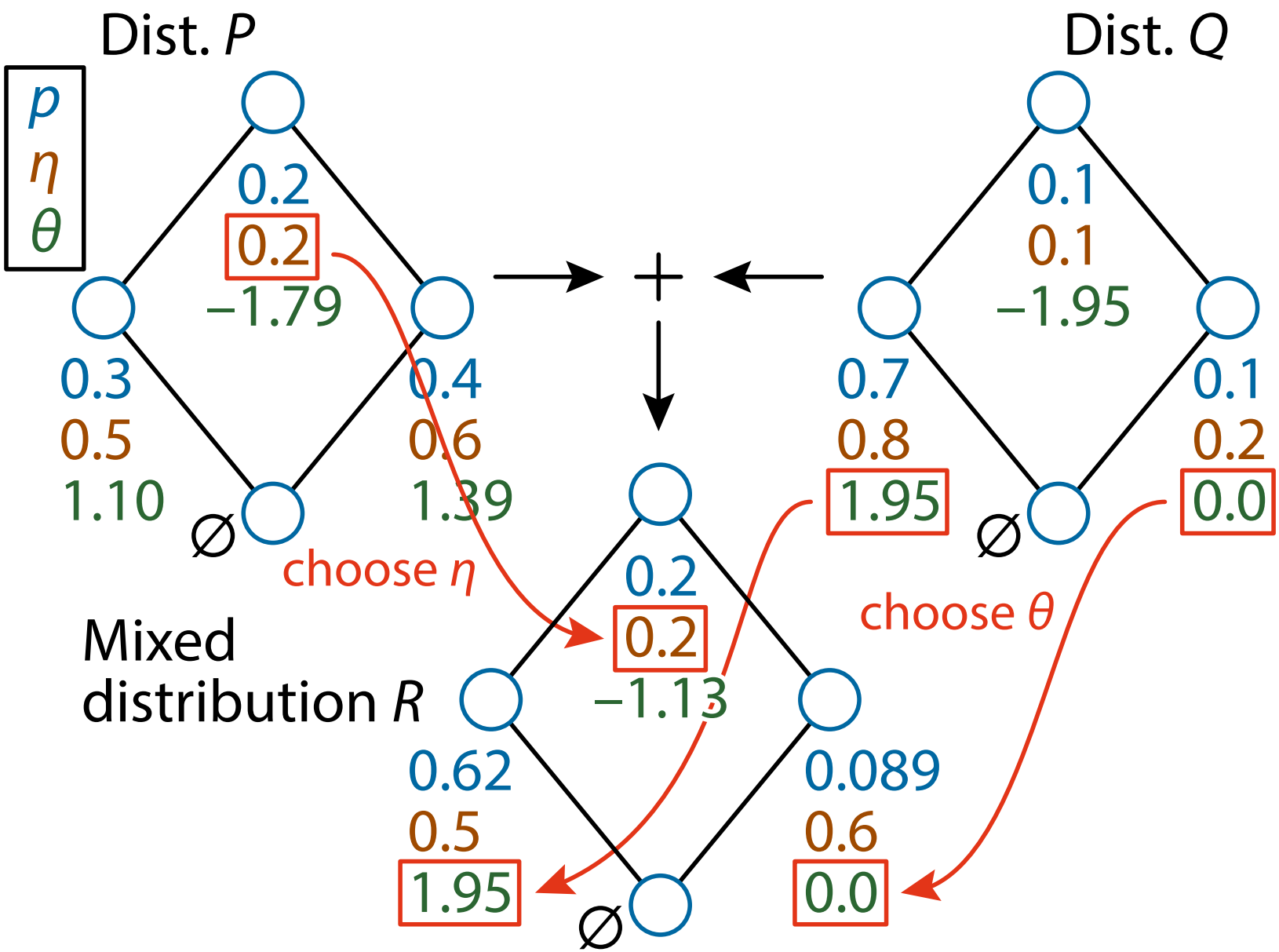


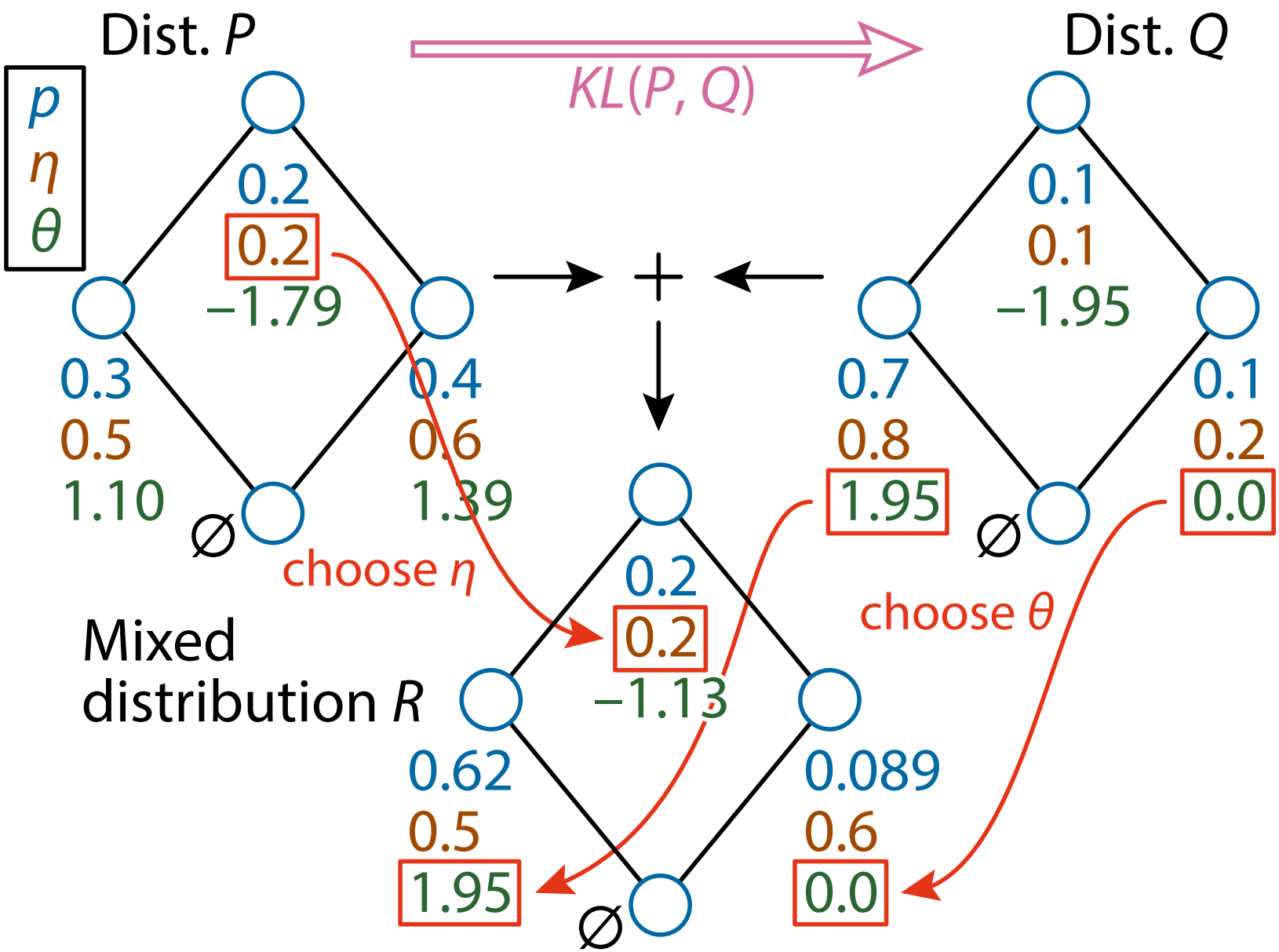
θ and η are dually orthogonal

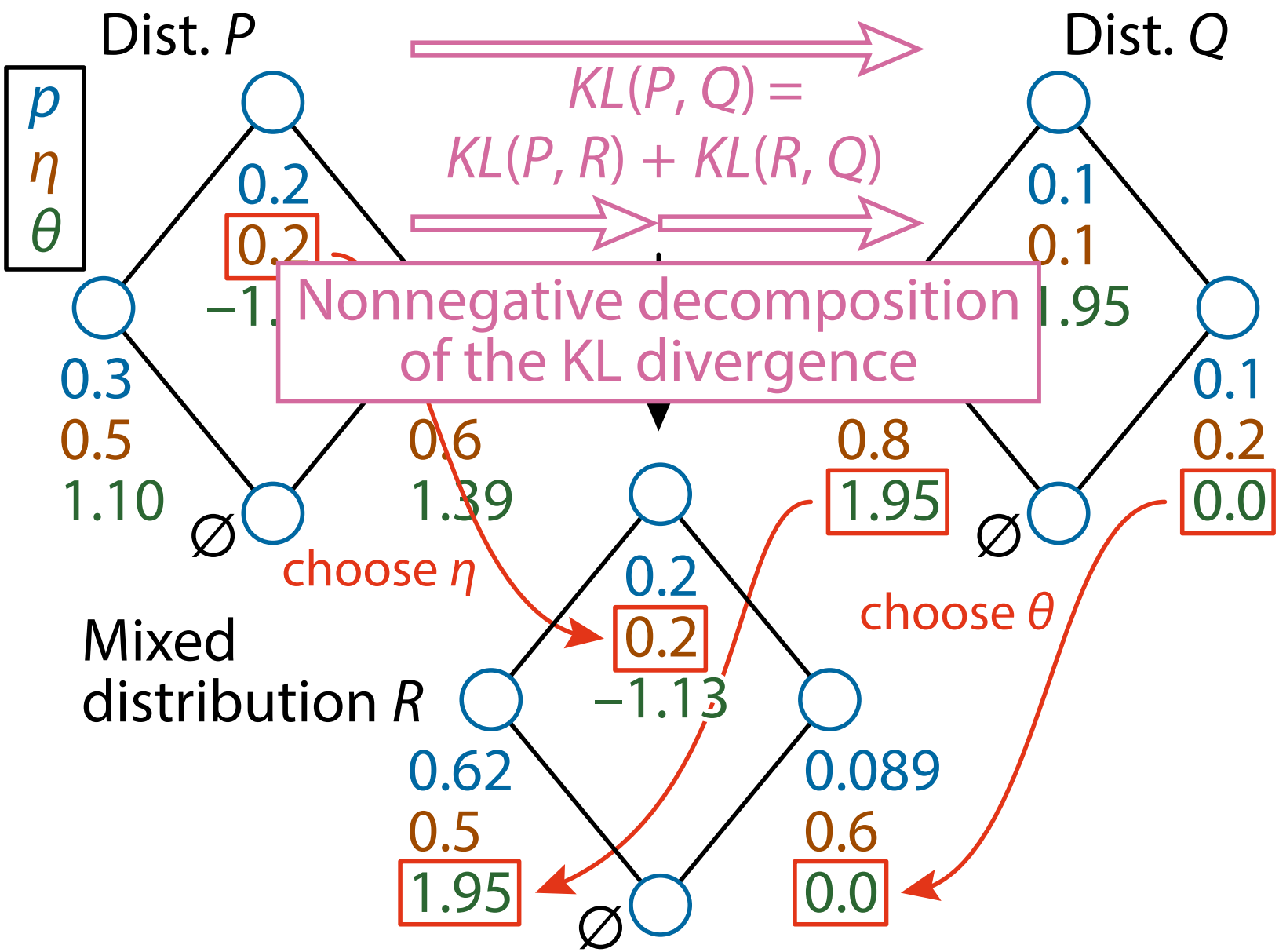

Dist. P
p
η
θ
0.2
0.2
−1.79
0.3
0.5
1.10
0.4
0.6
1.39
∅

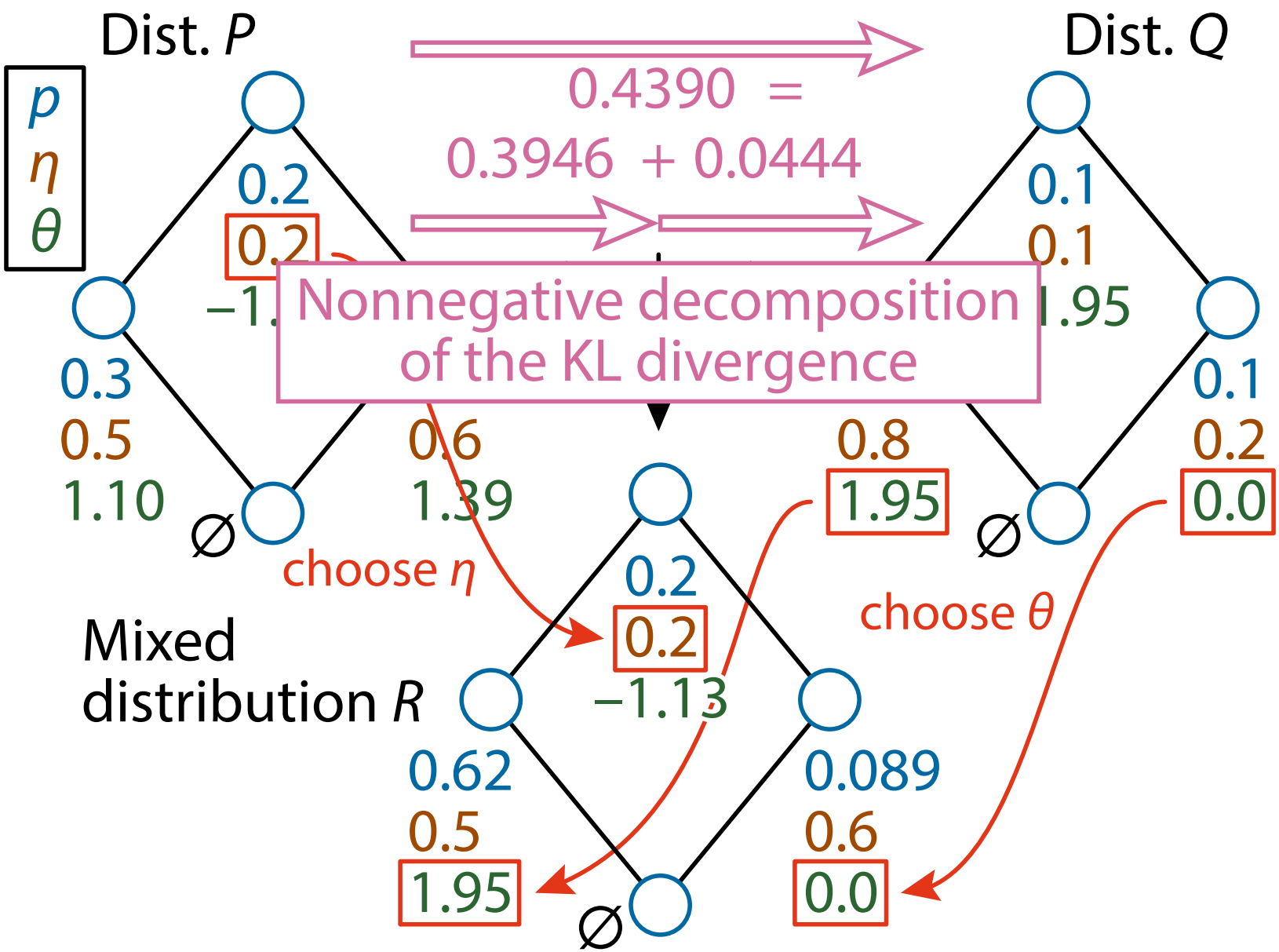

Dist. Q
0.1
0.1
−1.95
0.7
0.8
1.95
0.1
0.2
0.0
∅

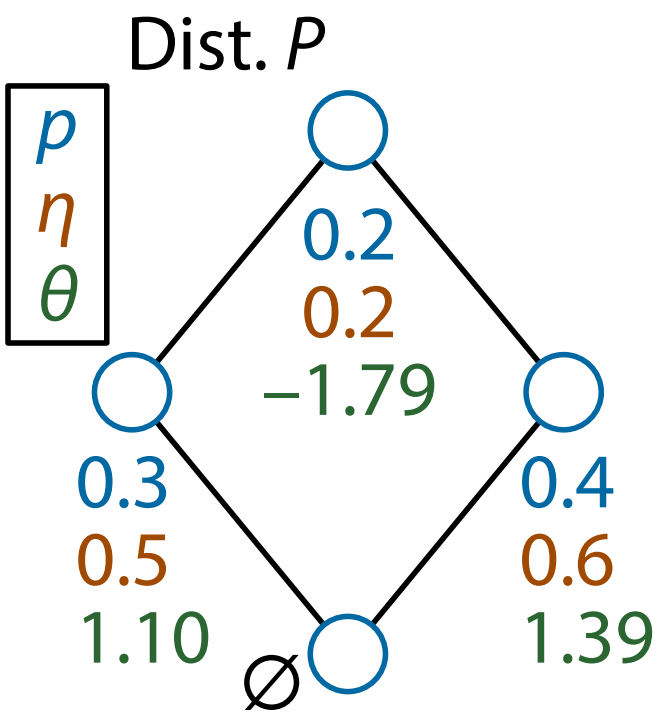

Dist. P
Dist. Q
p
η
θ
0.2
0.2
−1.79
0.3
0.5
1.10
0.4
0.6
1.39
∅
0.1
0.1
−1.95
0.7
0.8
1.95
0.1
0.2
0.0
∅
choose η
choose θ
Mixed distribution R
???
0.2
???
???
???
1.95
???
???
0.0
∅

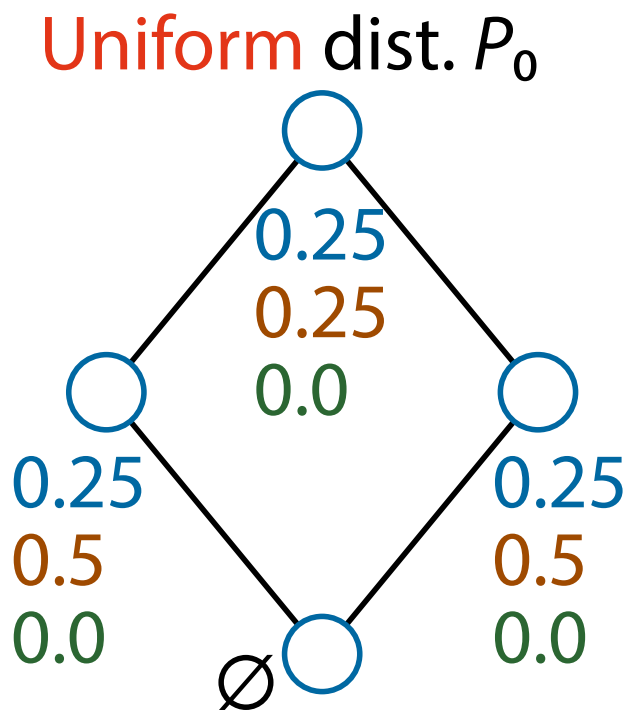

Dist. P
Dist. Q
p
η
θ
0.2
0.2
−1.79
0.3
0.5
1.10
0.4
0.6
1.39
∅
0.1
0.1
−1.95
0.7
0.8
1.95
0.1
0.2
0.0
∅
choose η
choose θ
Mixed distribution R
0.2
0.2
−1.13
0.62
0.5
1.95
0.089
0.6
0.0
∅

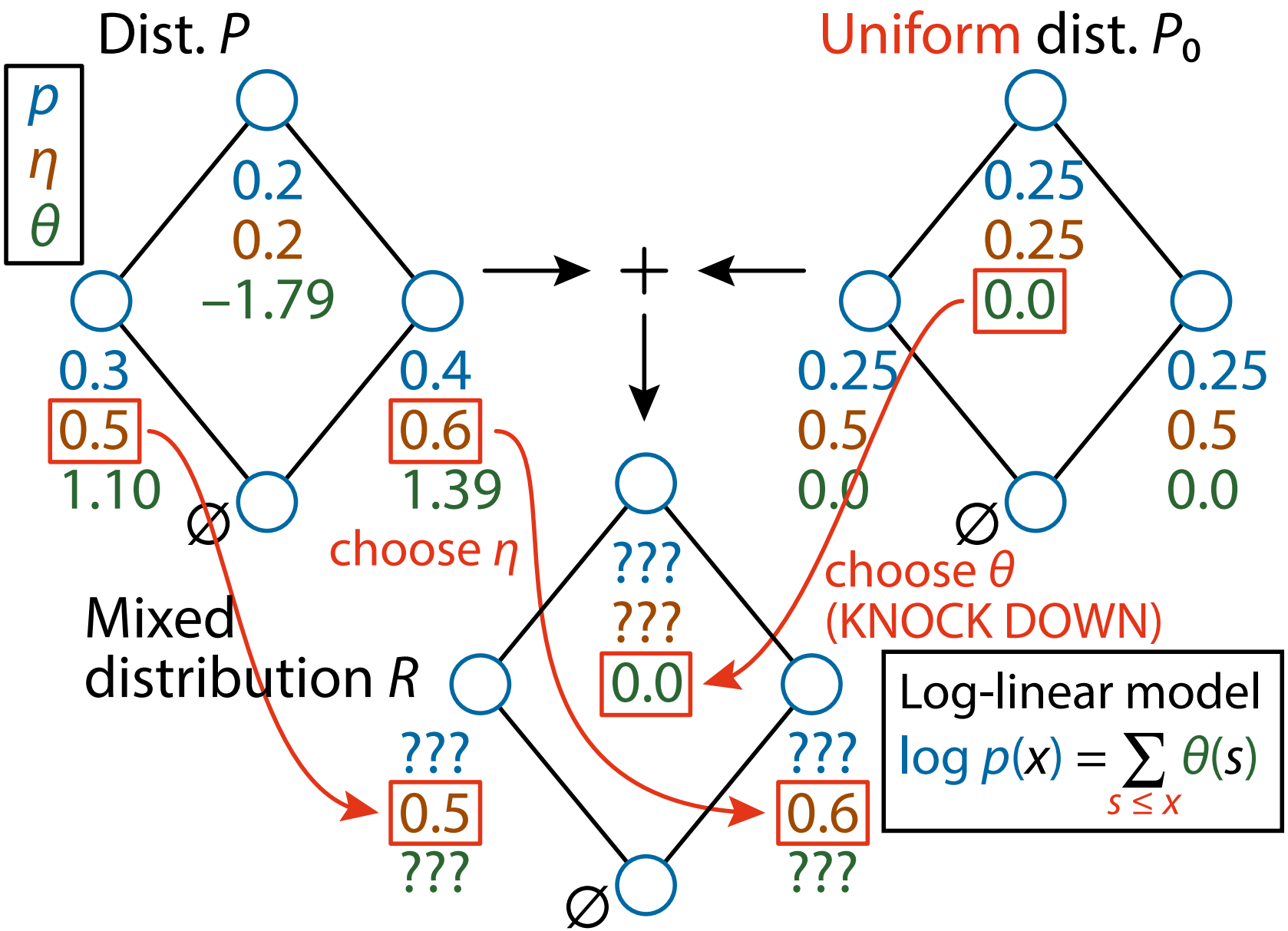

Dist. P
KL(P, Q)
Dist. Q
p
η
θ
0.2
0.2
−1.79
0.3
0.5
1.10
0.4
0.6
1.39
∅
0.1
0.1
−1.95
0.7
0.8
1.95
0.1
0.2
0.0
∅
choose η
choose θ
Mixed
distribution R
0.2
0.2
−1.13
0.62
0.5
1.95
0.089
0.6
0.0
∅

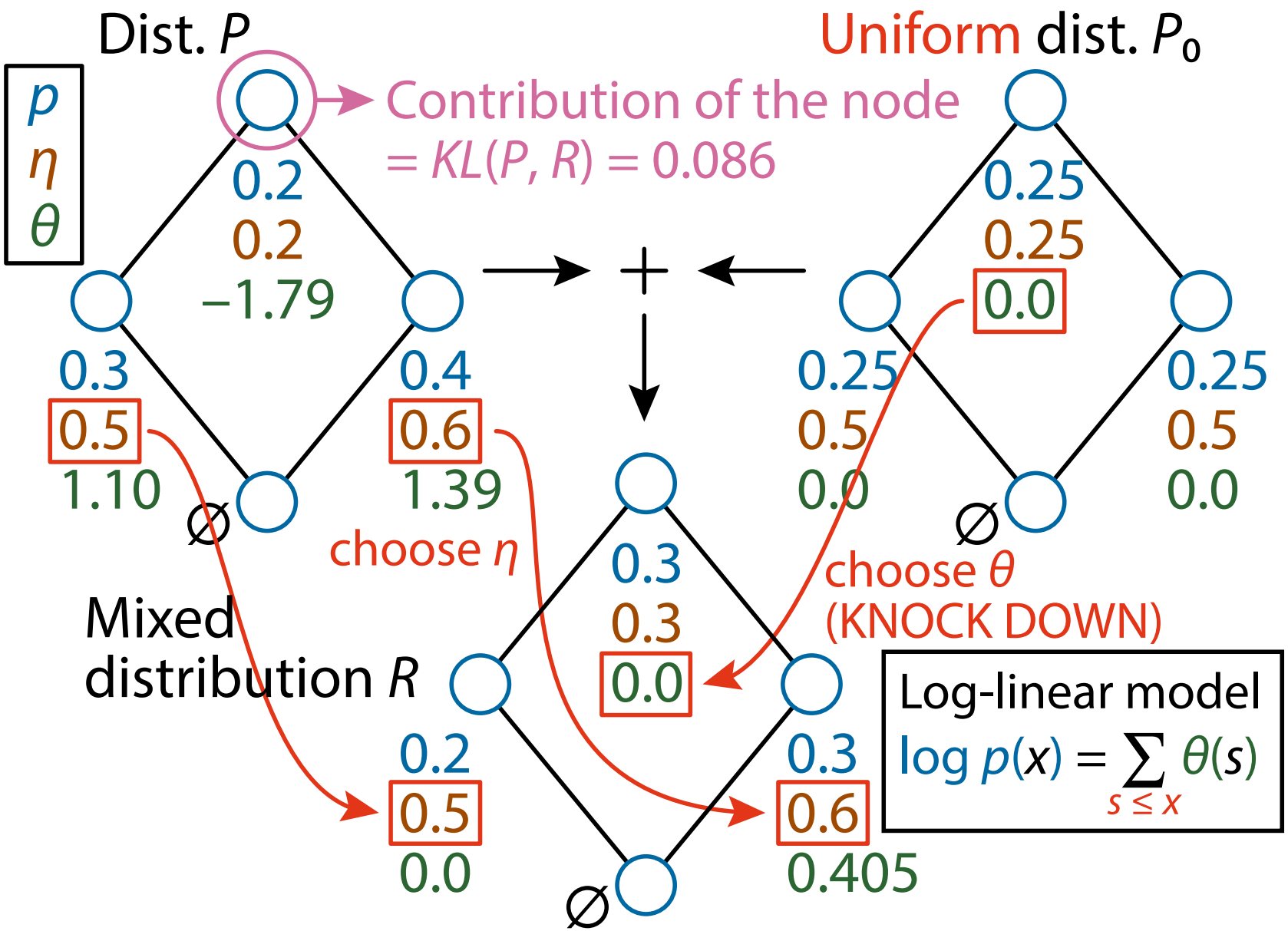

Dist. P
Dist. Q
p
η
θ
KL(P, Q) =
KL(P, R) + KL(R, Q)
Nonnegative decomposition
of the KL divergence
0.2
0.2
–1.
0.3
0.5
1.10
∅
0.6
1.39
0.1
0.1
1.95
0.1
0.2
0.0
0.8
1.95
∅
choose η
choose θ
Mixed
distribution R
0.2
0.2
–1.13
0.62
0.5
1.95
∅
0.089
0.6
0.0

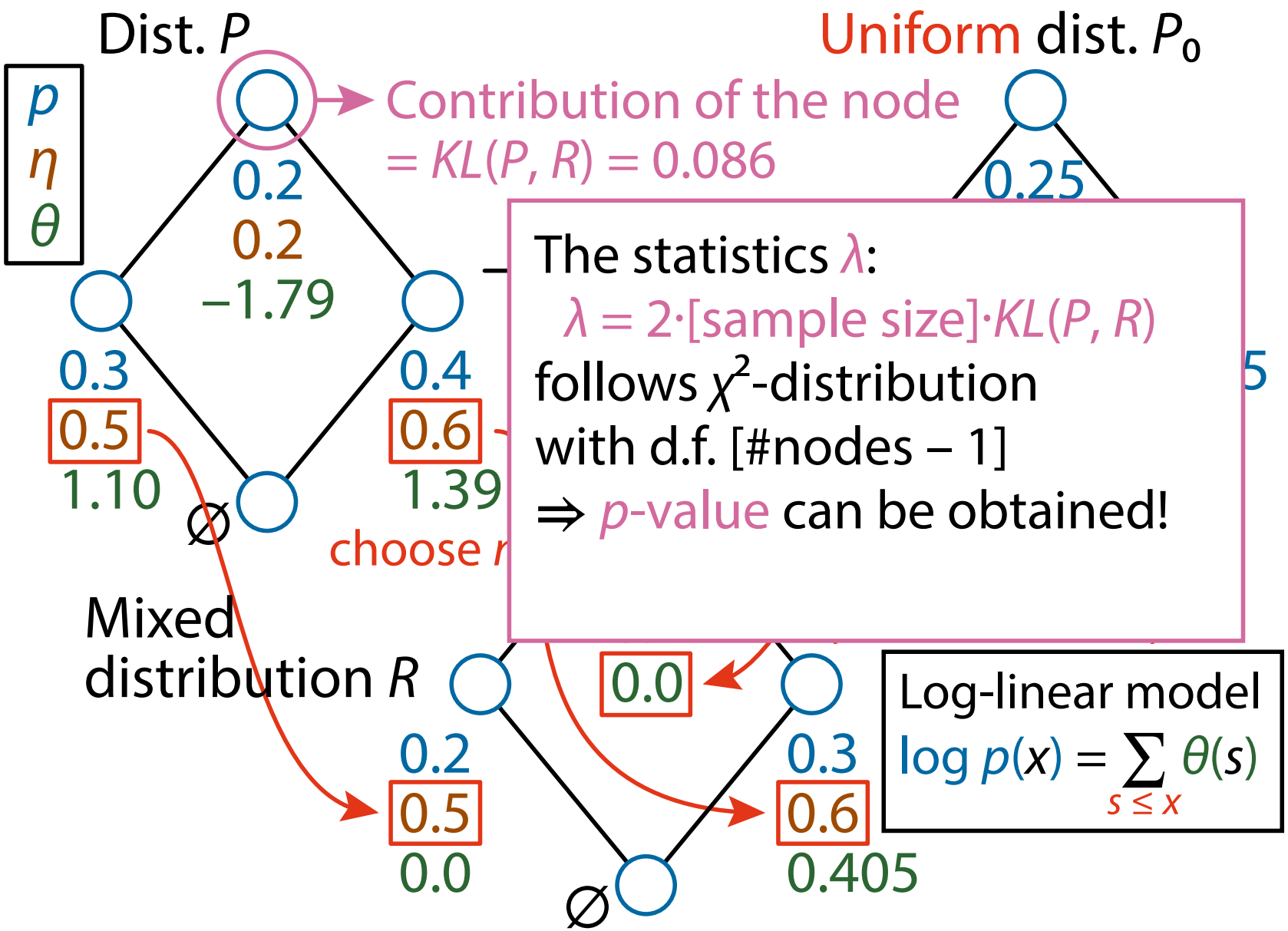

Dist. P
Dist. Q
p
η
θ
0.4390 =
0.3946 + 0.0444
Nonnegative decomposition
of the KL divergence
0.2
0.2
–1.
0.3
0.5
1.10
∅
0.6
1.39
0.1
0.1
1.95
0.1
0.2
0.0
∅
0.8
1.95
choose η
choose θ
Mixed
distribution R
0.2
0.2
–1.13
0.62
0.5
1.95
∅
0.089
0.6
0.0

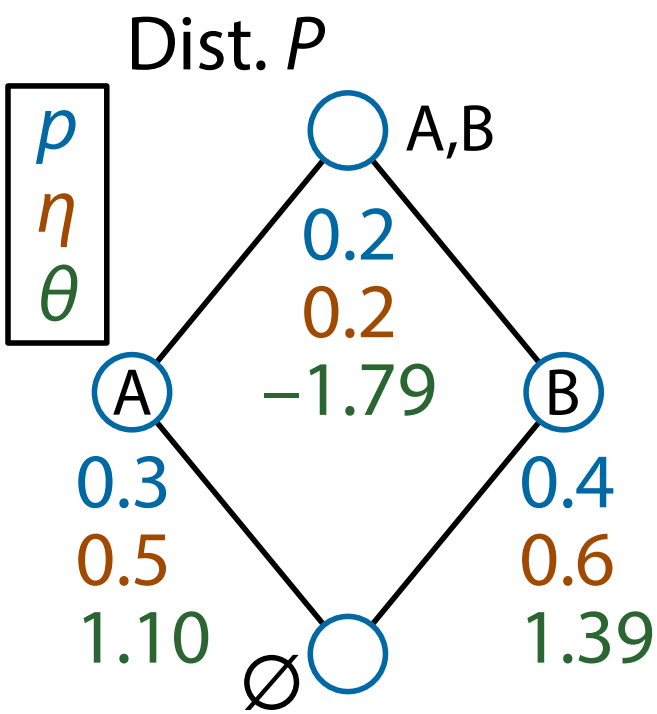

Dist. P
p
η
θ
0.2
0.2
−1.79
0.3
0.5
1.10
0.4
0.6
1.39
∅


Uniform dist. P_0
0.25
0.25
0.0
0.25
0.5
0.0
0.25
0.5
0.0
∅

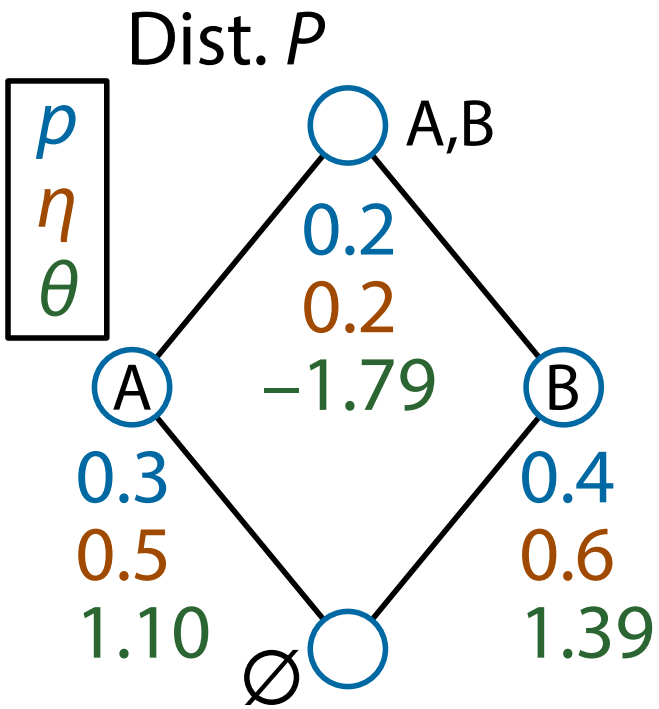

Dist. P
p
η
θ
0.2
0.2
−1.79
0.3
0.5
1.10
0.4
0.6
1.39
∅
Uniform dist. P_0
0.25
0.25
0.0
0.25
0.5
0.0
0.25
0.5
0.0
∅
choose η
choose θ
(KNOCK DOWN)
Mixed
distribution R
???
???
0.0
???
0.5
???
???
0.6
???
∅
Log-linear model
$\log p(x) = \sum_{s \leq x} \theta(s)$


Dist. P
Uniform dist. P_0
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
∅
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
$\log p(x) = \sum_{s \leq x} \theta(s)$


Dist. P
Uniform dist. P_0
p
η
θ
Contribution of the node
$= KL(P, R) = 0.086$
0.2
0.2
−1.79
0.25
0.3
0.5
1.10
0.4
0.6
1.39
∅
choose
The statistics λ:
$\lambda = 2 \cdot [\text{sample size}] \cdot KL(P, R)$
follows χ^2-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
$\log p(x) = \sum_{s \le x} \theta(s)$



$$\begin{array}{ccc} \text{A} & \text{B} & \text{A, B} \\ \end{array}$$
$$\begin{bmatrix} 0.3 & 0.4 & 0.2 \\ 0.5 & 0.6 & 0.2 \\ 1.10 & 1.39 & -1.79 \end{bmatrix}$$



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

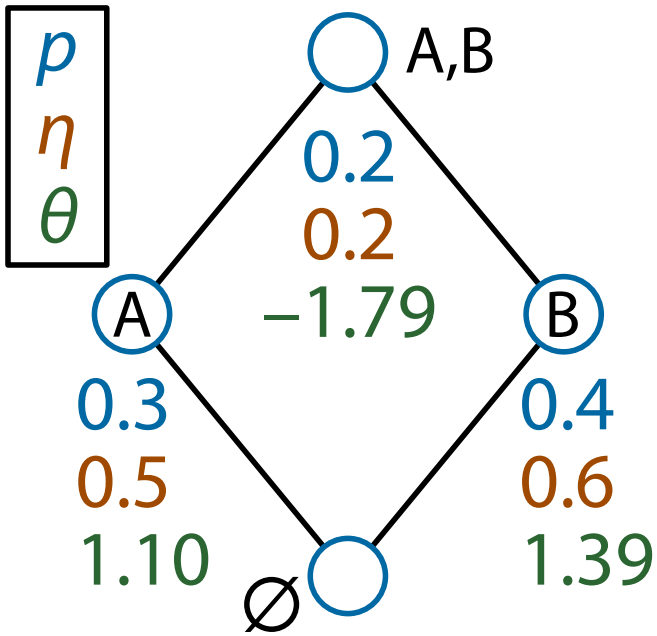
$$
\begin{bmatrix}
0.3 & 0.4 & 0.2 \\
0.5 & 0.6 & 0.2 \\
1.10 & 1.39 & -1.79
\end{bmatrix}
$$

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

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

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

Dist. P



	A	B	A, B
p	0.3	0.4	0.2
η	0.5	0.6	0.2
θ	1.10	1.39	–1.79

KL = Score of A

	A	B	A, B
p	???	???	???
η	???	0.6	0.2
θ	0.0	???	???

KL = Score of B

	A	B	A, B
p	???	???	???
η	0.5	???	0.2
θ	???	0.0	???

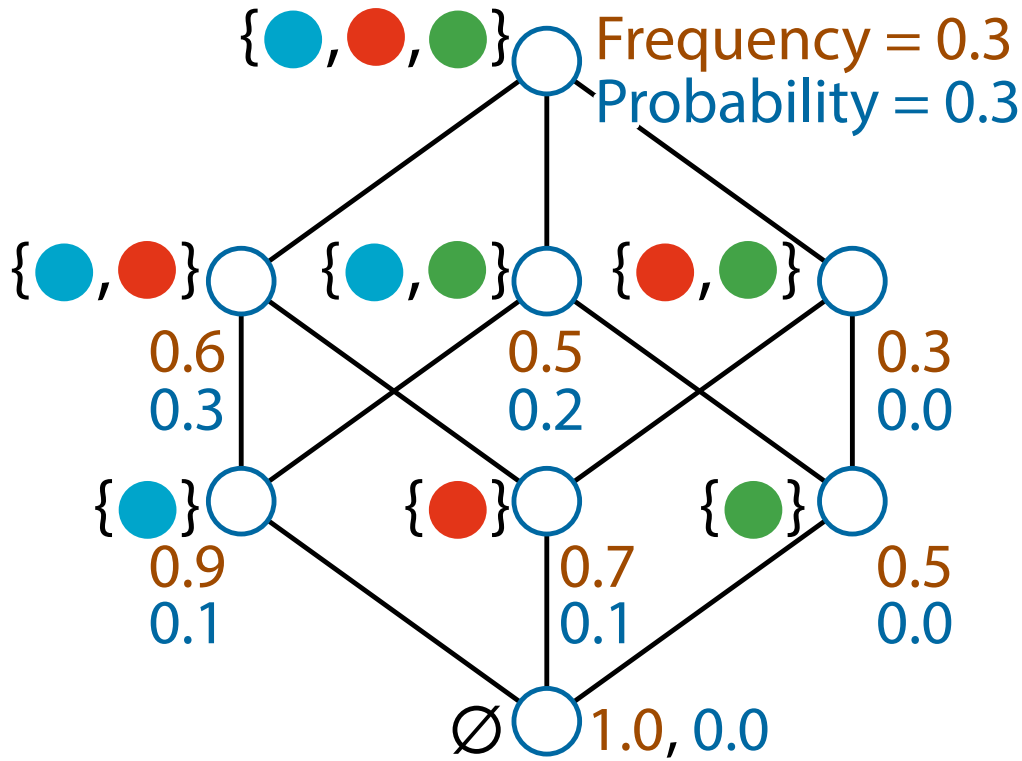
KL = Score of A, B

	A	B	A, B
p	???	???	???
η	0.5	0.6	???
θ	???	???	0.0

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

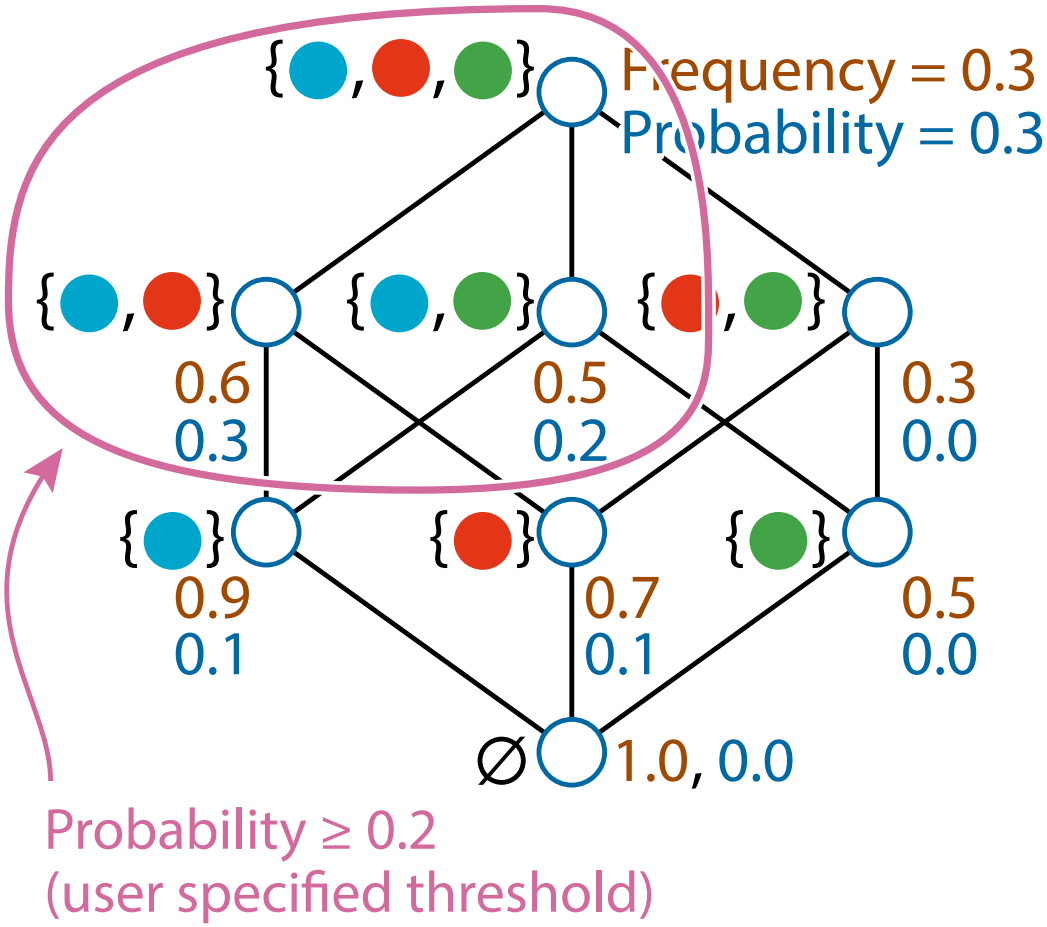


Number of nodes = $2^{\#\text{features}}$
⇒combinatorial explosion!

Make a Poset from Data

Dataset

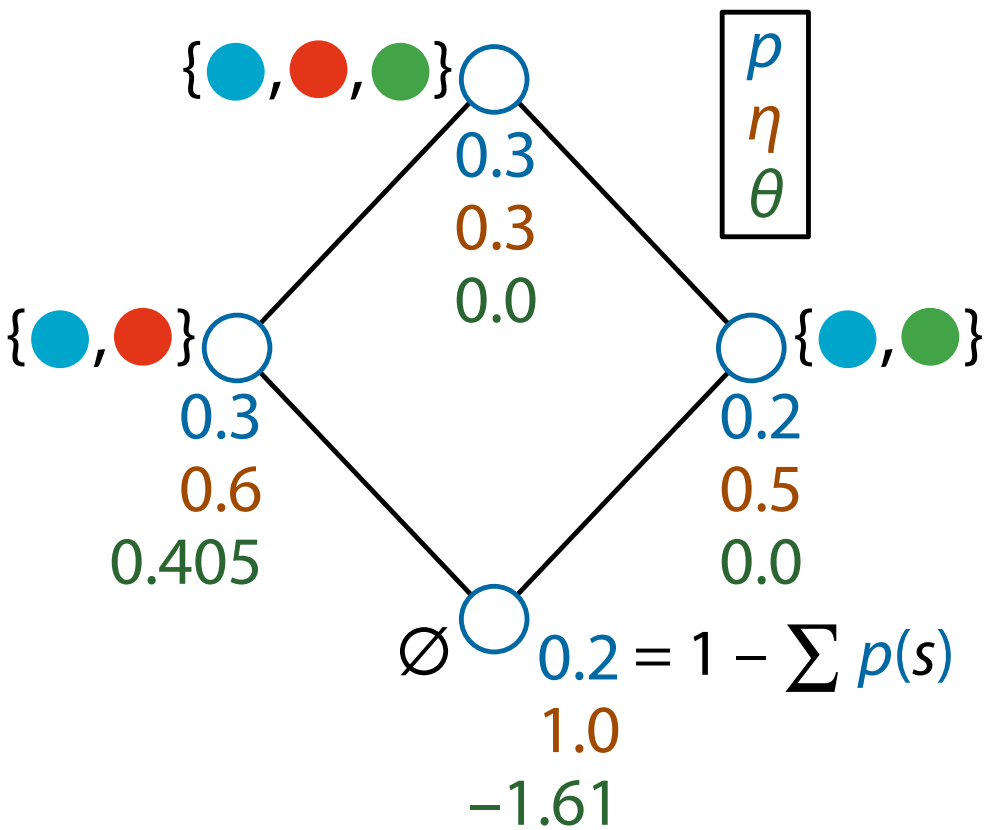
ID 1:	1	1	0
ID 2:	1	1	1
ID 3:	1	1	0
ID 4:	1	1	1
ID 5:	1	1	0
ID 6:	1	0	1
ID 7:	1	0	1
ID 8:	1	1	1
ID 9:	1	0	0
ID10:	0	1	0



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)

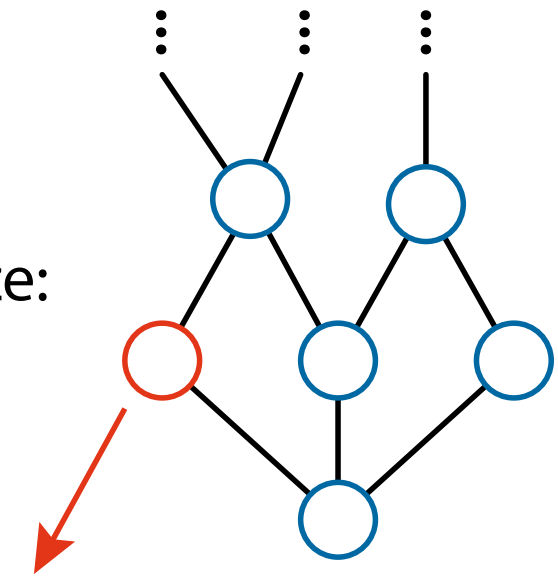
features: 41,270

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)

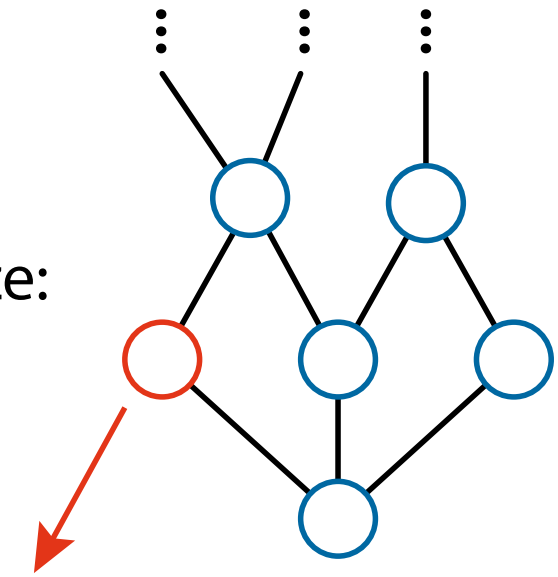
features: 468

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

- We build information geometry for posets (partially ordered sets)
  - Natural connection between the information geometric dual coordinates and the partial order structure
- We can decompose a probability distribution and asses the significance of any-order interactions beyond pairwise interactions