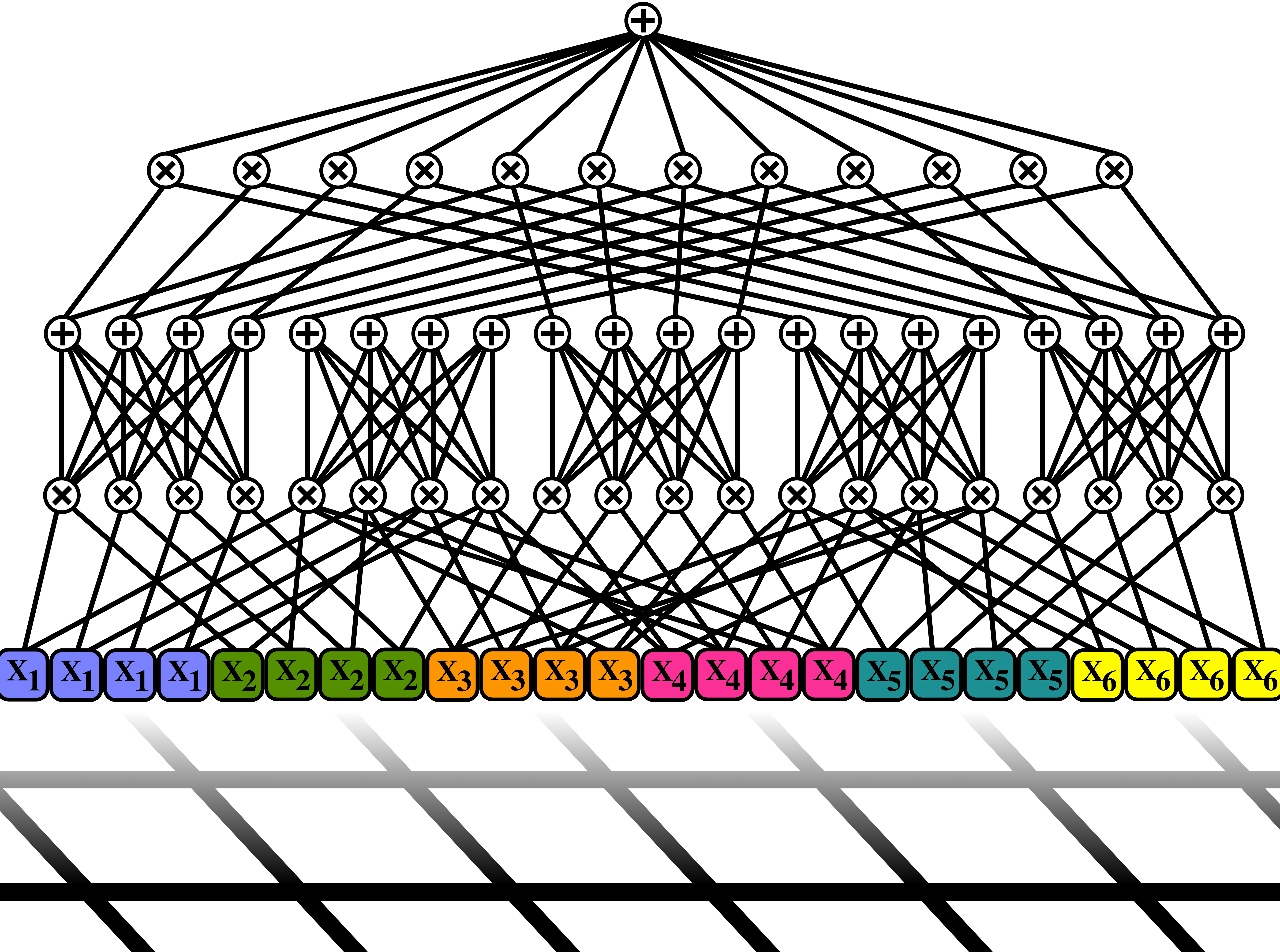
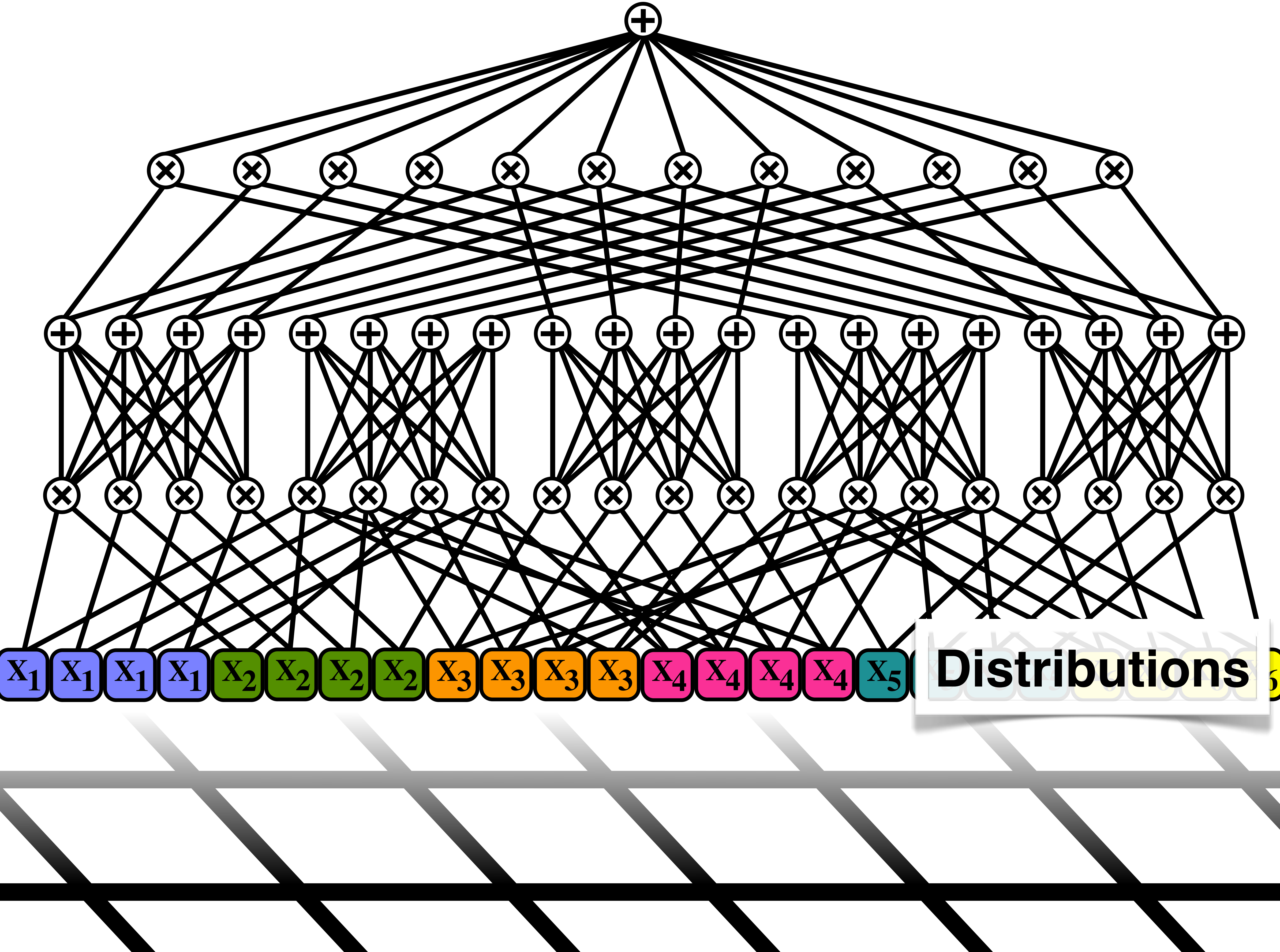
Discriminative Learning of Sum-Product Networks

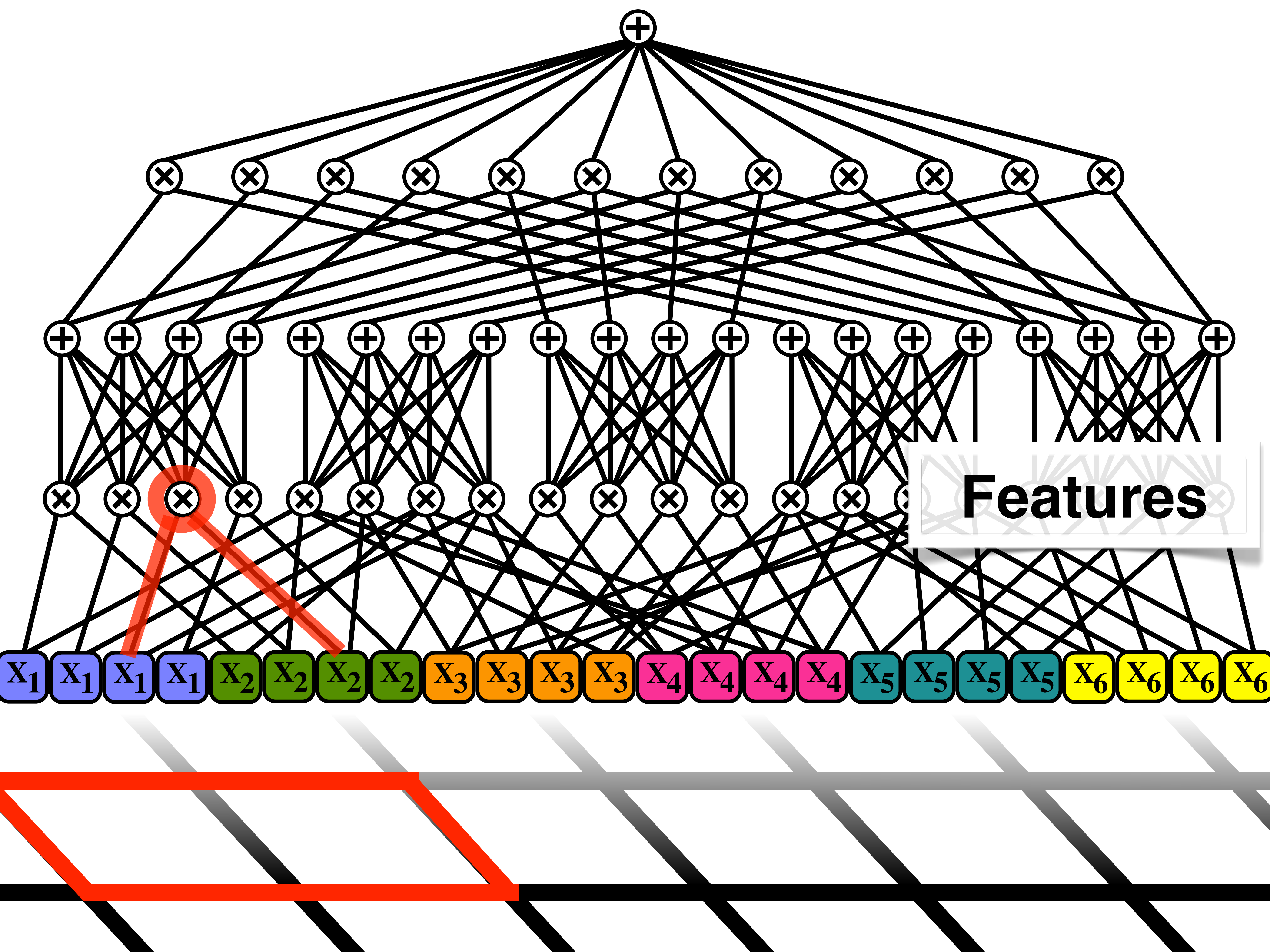
Robert Gens
Pedro Domingos

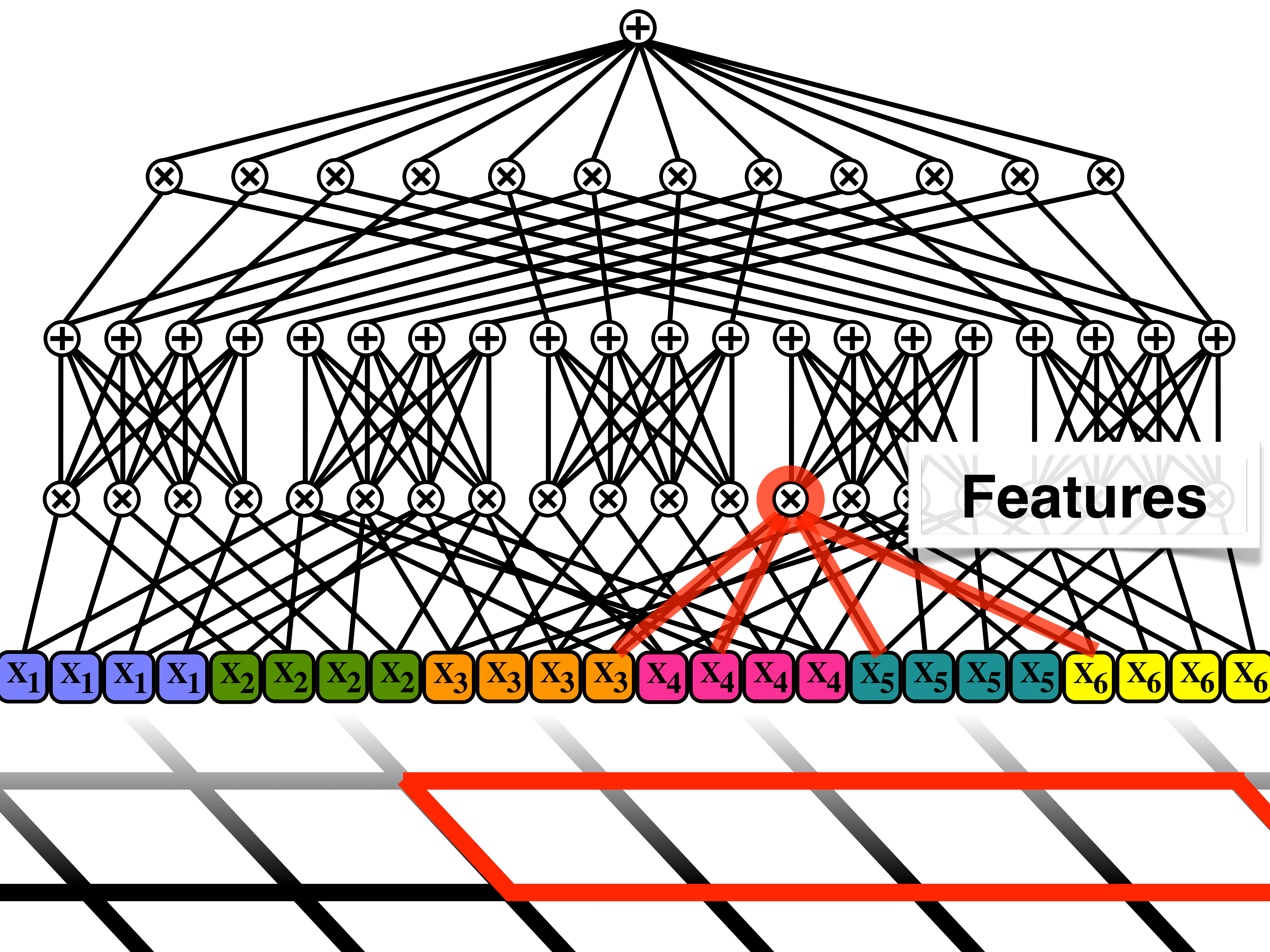
UNIVERSITY of WASHINGTON
University of Washington Computer Science & Engineering



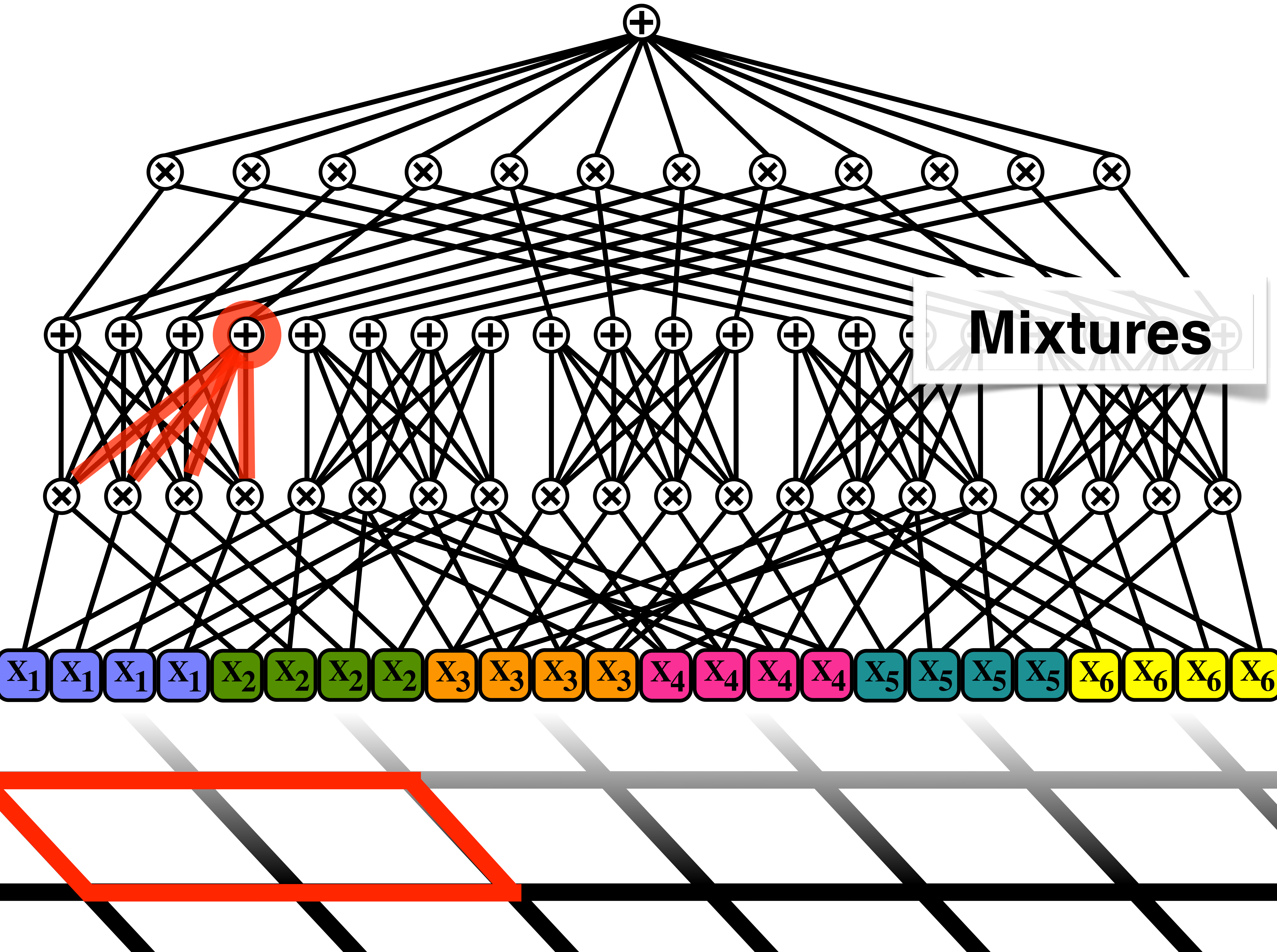
X1
X1
X1
X1
X2
X2
X2
X2
X3
X3
X3
X3
X4
X4
X4
X4
X5
X5
X5
X5
X6
X6
X6
X6

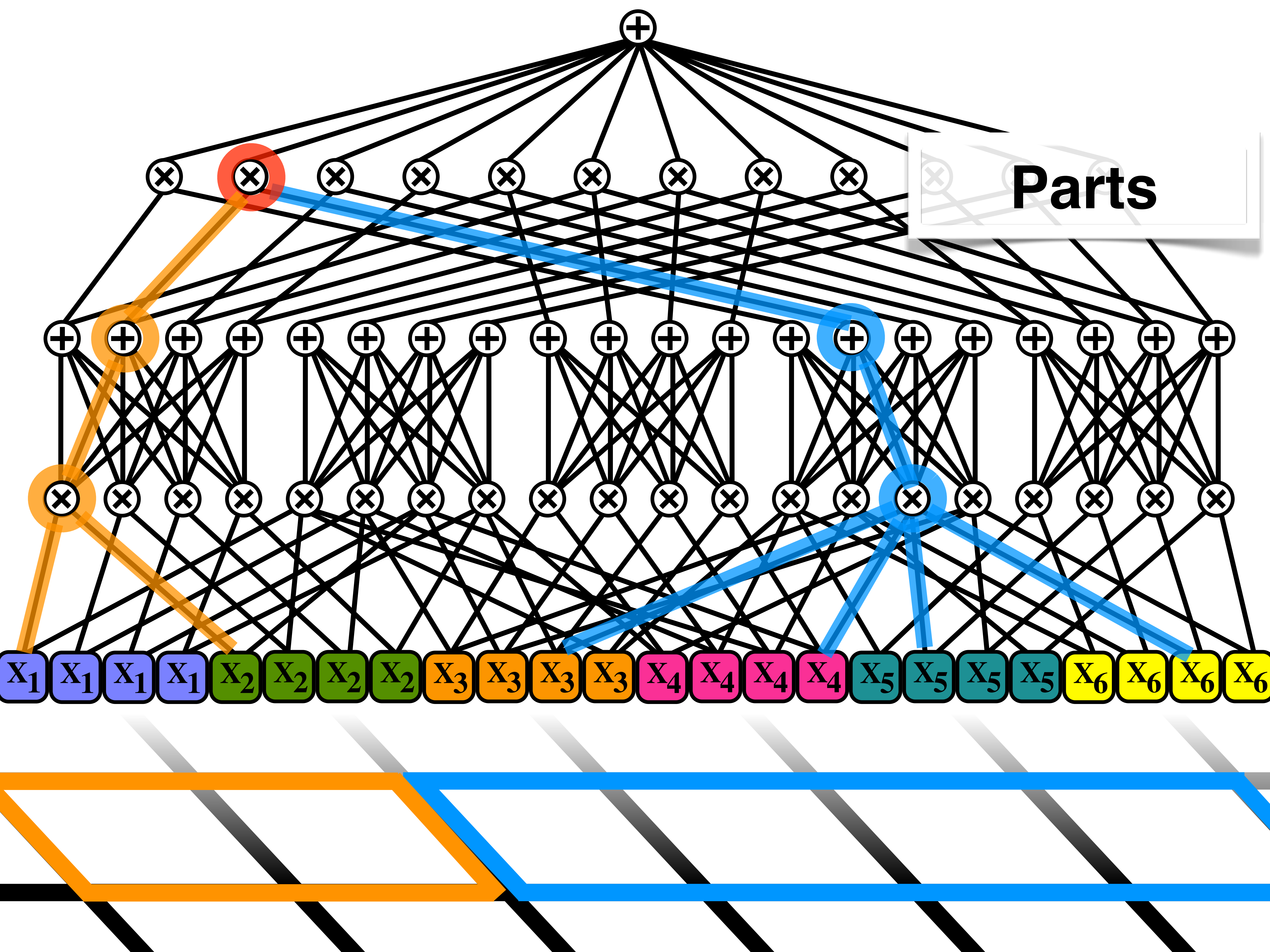

X1 X1 X1 X1 X2 X2 X2 X2 X3 X3 X3 X3 X4 X4 X4 X4 X5
Distributions


Features
X1 X1 X1 X1 X2 X2 X2 X2 X3 X3 X3 X3 X4 X4 X4 X4 X5 X5 X5 X5 X6 X6 X6 X6

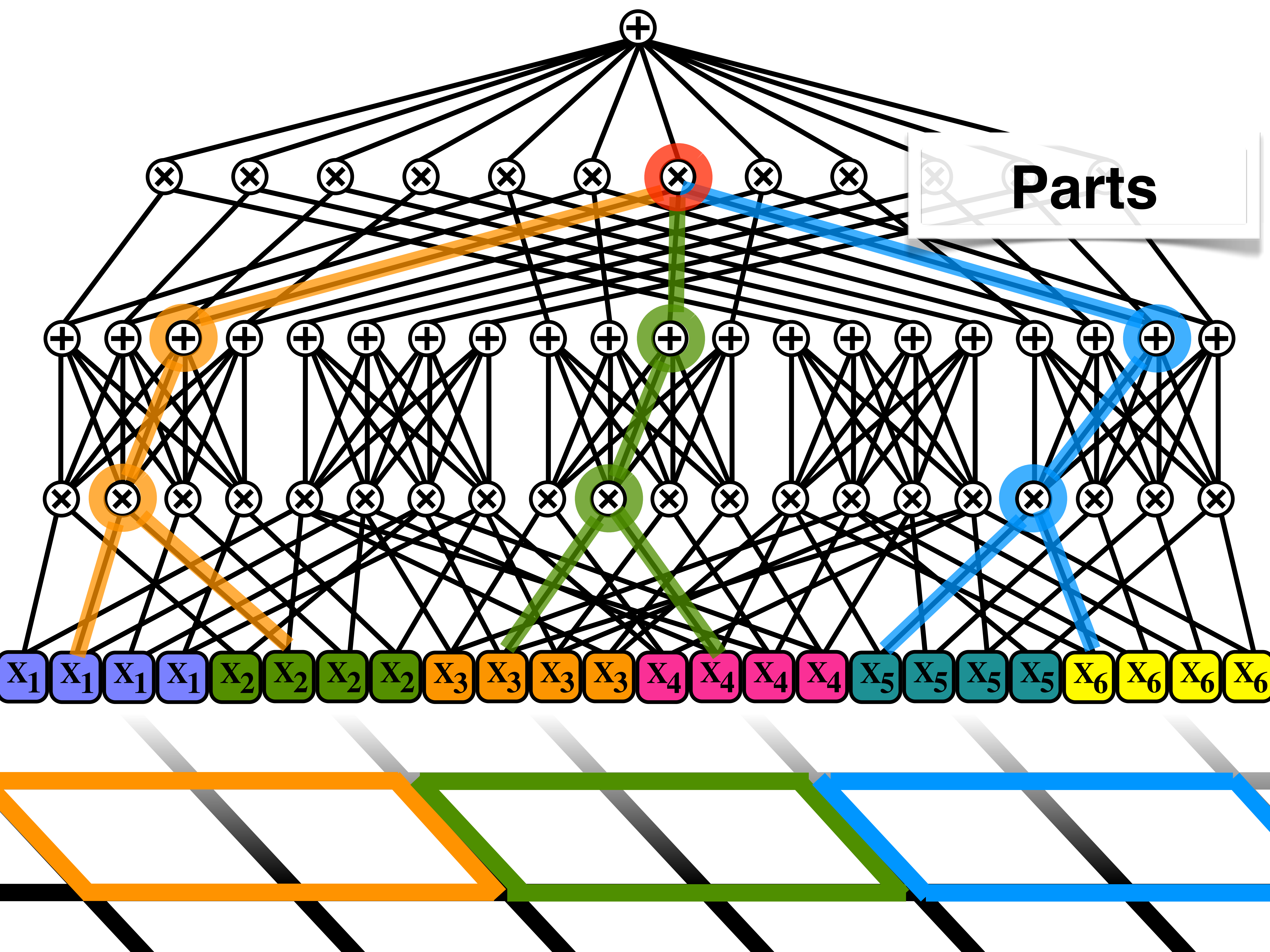


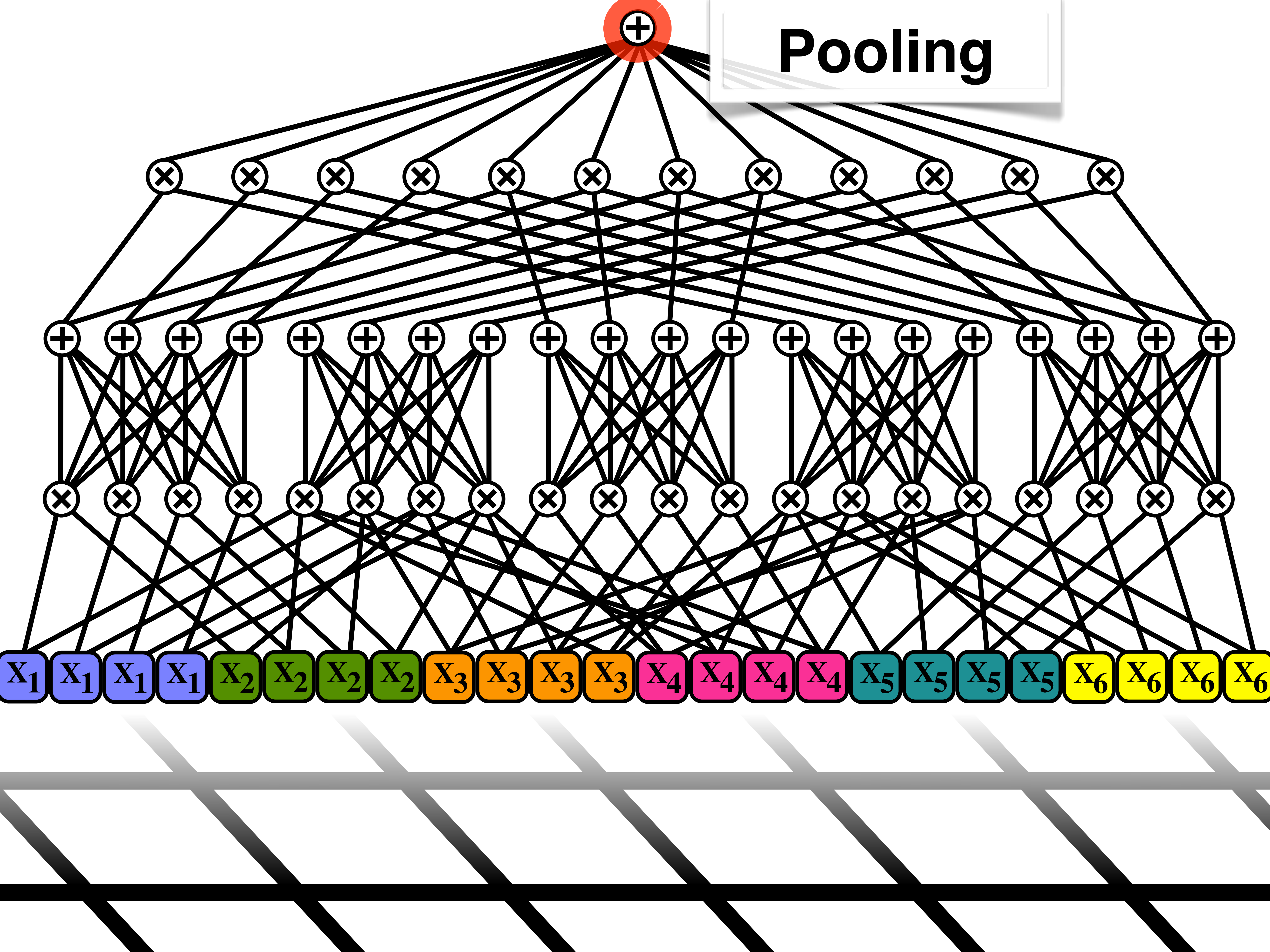
Features
X1 X1 X1 X1 X2 X2 X2 X2 X3 X3 X3 X3 X4 X4 X4 X4 X5 X5 X5 X5 X6 X6 X6 X6


Mixtures
X1 X1 X1 X1 X2 X2 X2 X2 X3 X3 X3 X3 X4 X4 X4 X4 X5 X5 X5 X5 X6 X6 X6 X6

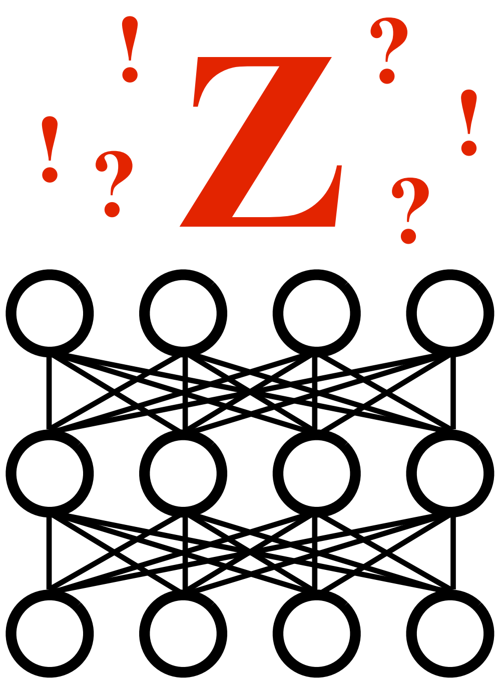


Parts
X1 X1 X1 X1 X2 X2 X2 X2 X3 X3 X3 X3 X4 X4 X4 X4 X5 X5 X5 X5 X6 X6 X6 X6

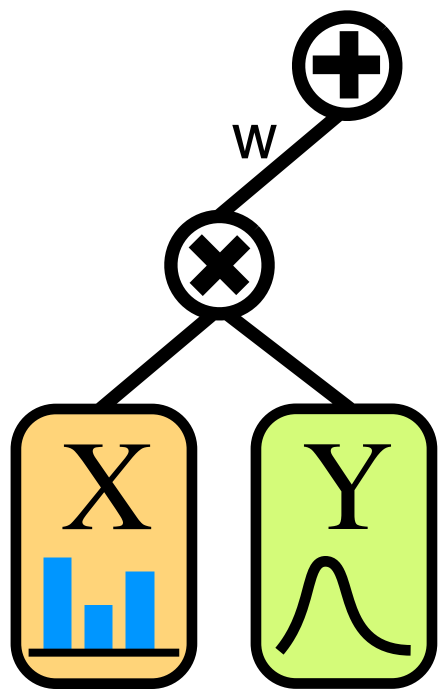

Parts
X1 X1 X1 X1 X2 X2 X2 X2 X3 X3 X3 X3 X4 X4 X4 X4 X5 X5 X5 X5 X6 X6 X6 X6



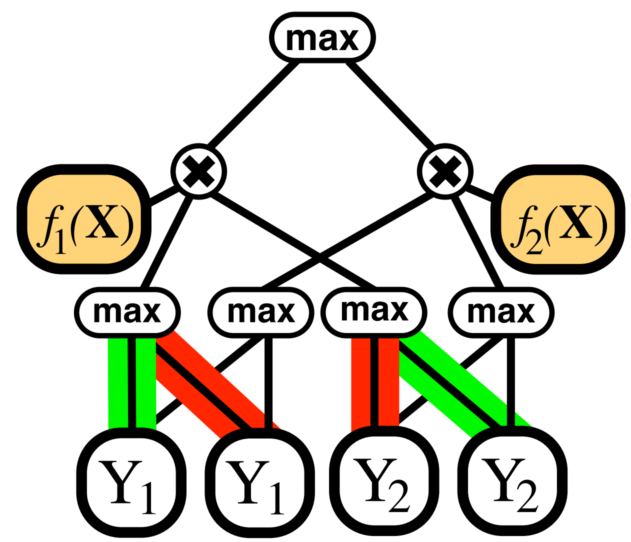
Pooling
X1 X1 X1 X1 X2 X2 X2 X2 X3 X3 X3 X3 X4 X4 X4 X4 X5 X5 X5 X5 X6 X6 X6 X6



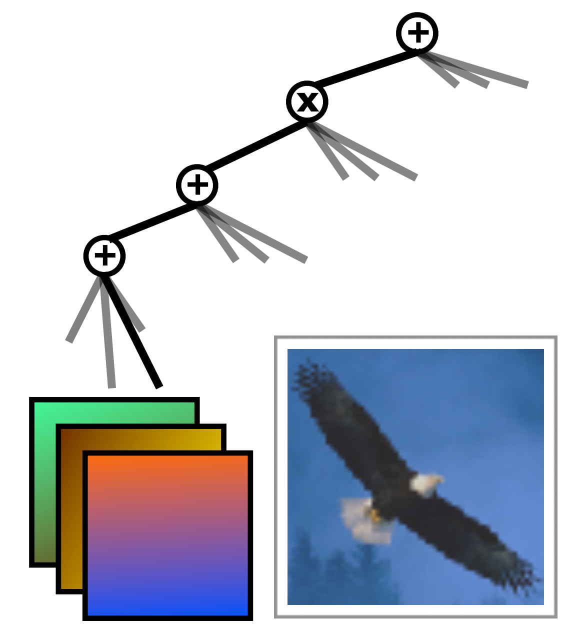
Motivation



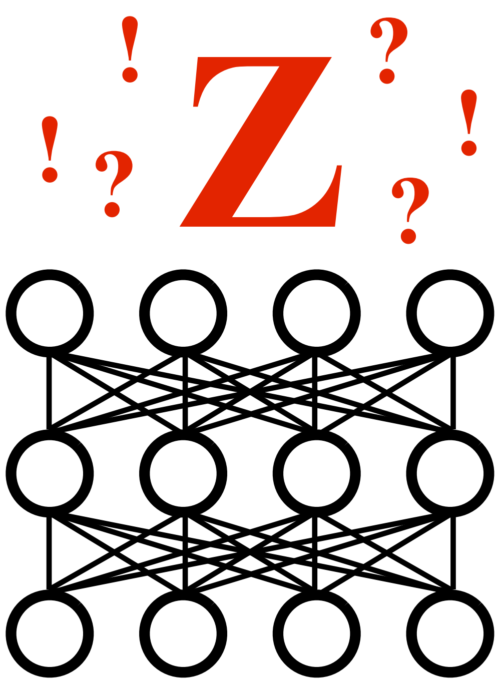
SPN Review



Discriminative Training



Experiments



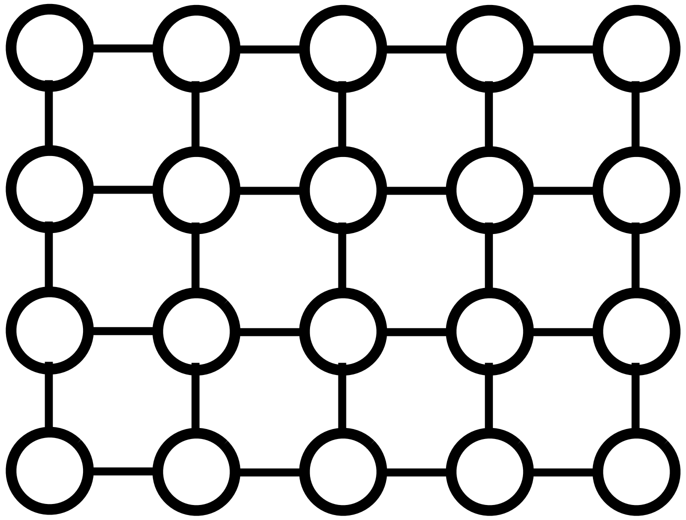
Motivation

SPN Review

Discriminative Training

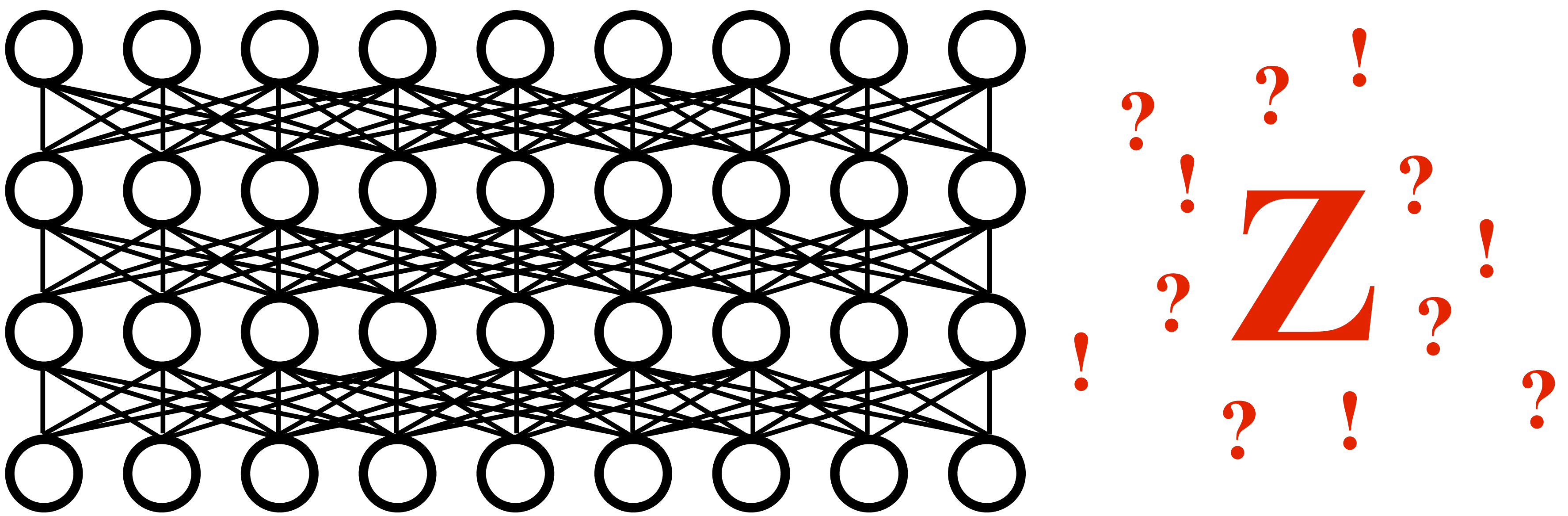
Experiments

Graphical Models



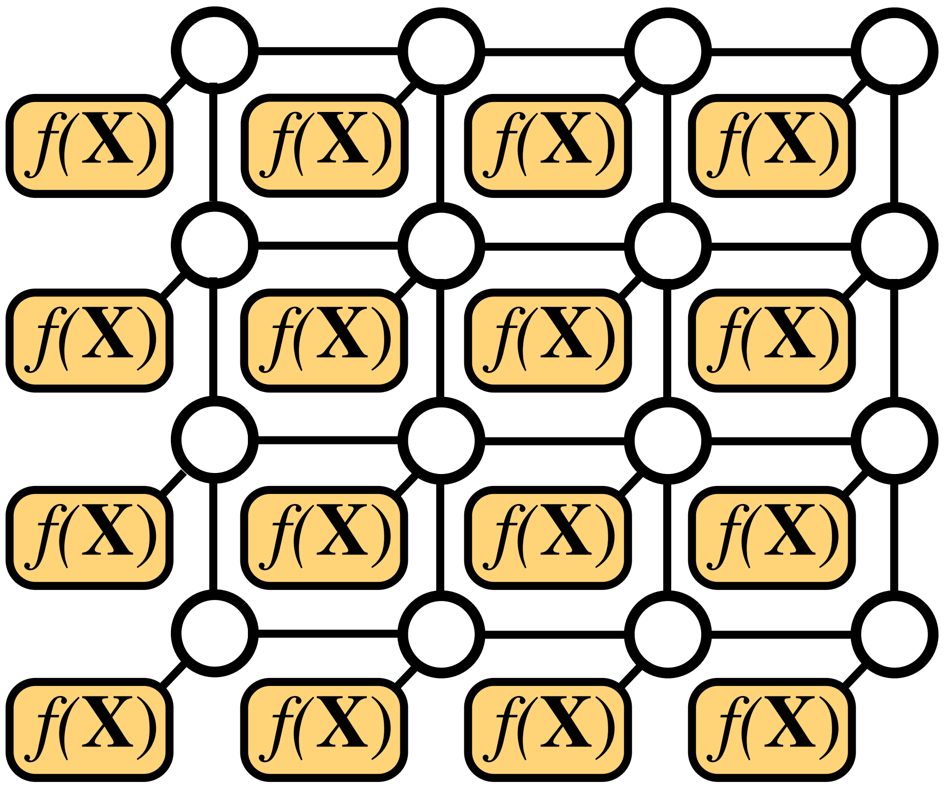
➡SPNs perform fast, exact inference on high treewidth models

Deep Architectures



➡SPNs have full probabilistic semantics and tractable inference over many layers

Discriminative Learning

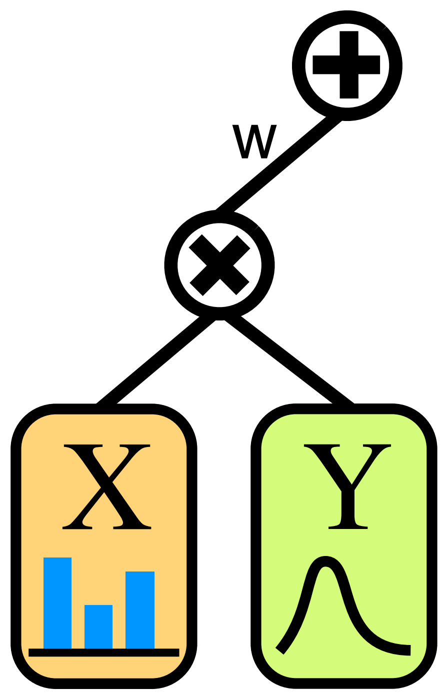


$Z(X)$?! ?! ?

NIPS'12

➡SPNs combine features with fast, exact inference over high treewidth models

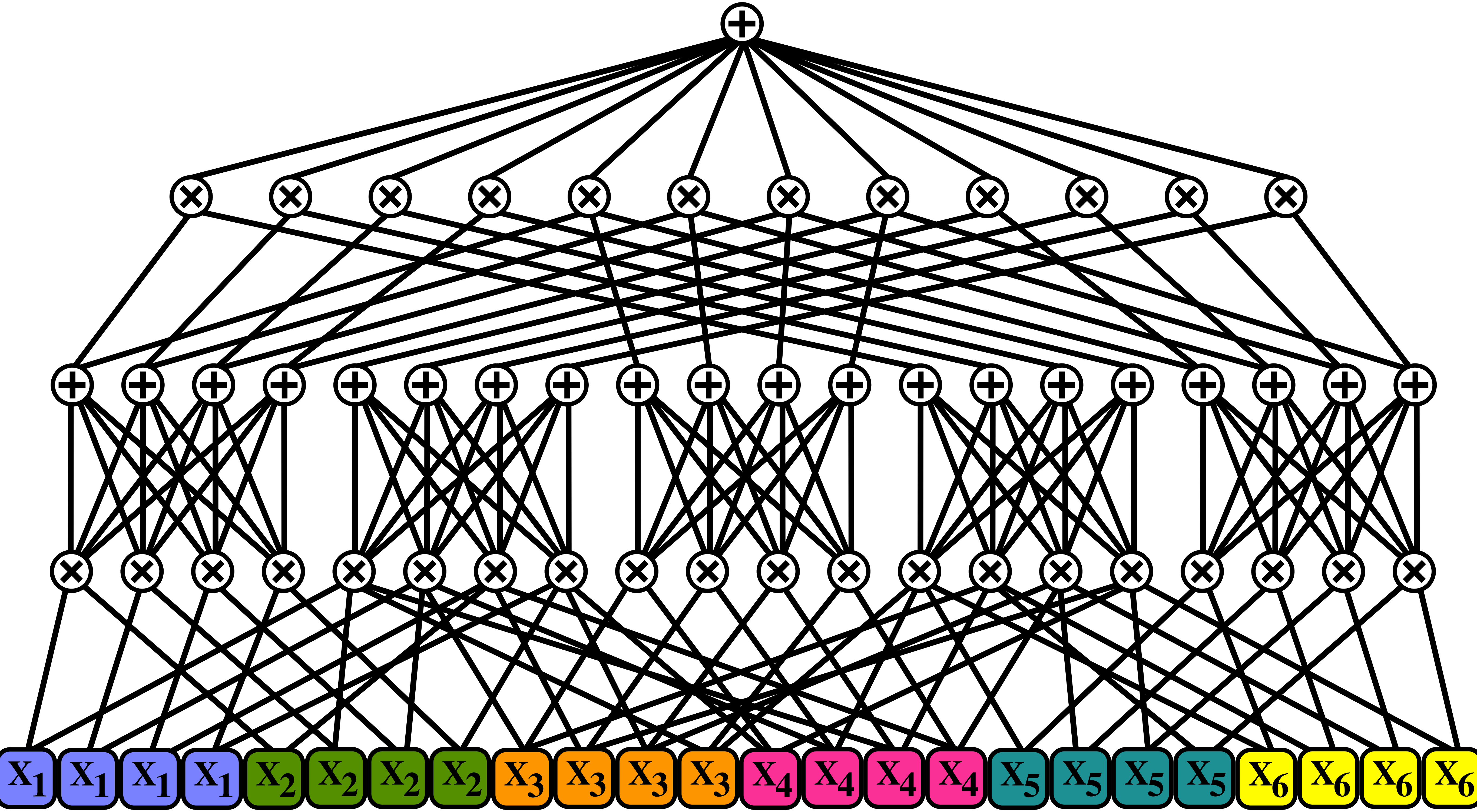
Motivation



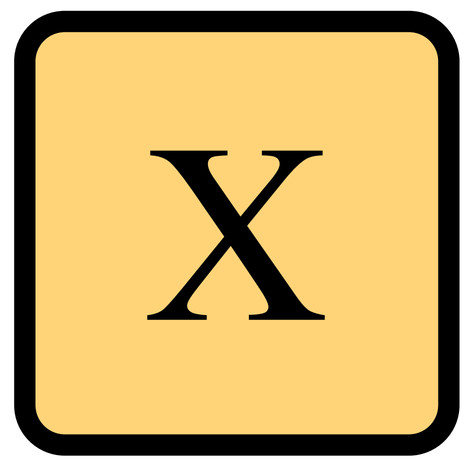
SPN Review

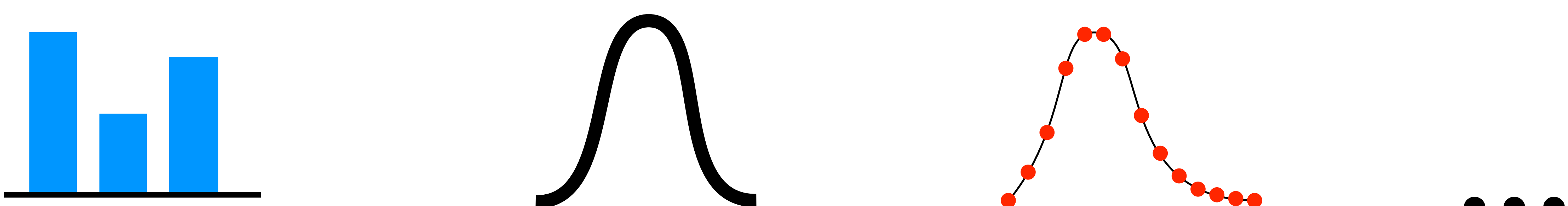
Discriminative Training

Experiments


X1 X1 X1 X1 X2 X2 X2 X2 X3 X3 X3 X3 X4 X4 X4 X4 X5 X5 X5 X5 X6 X6 X6 X6

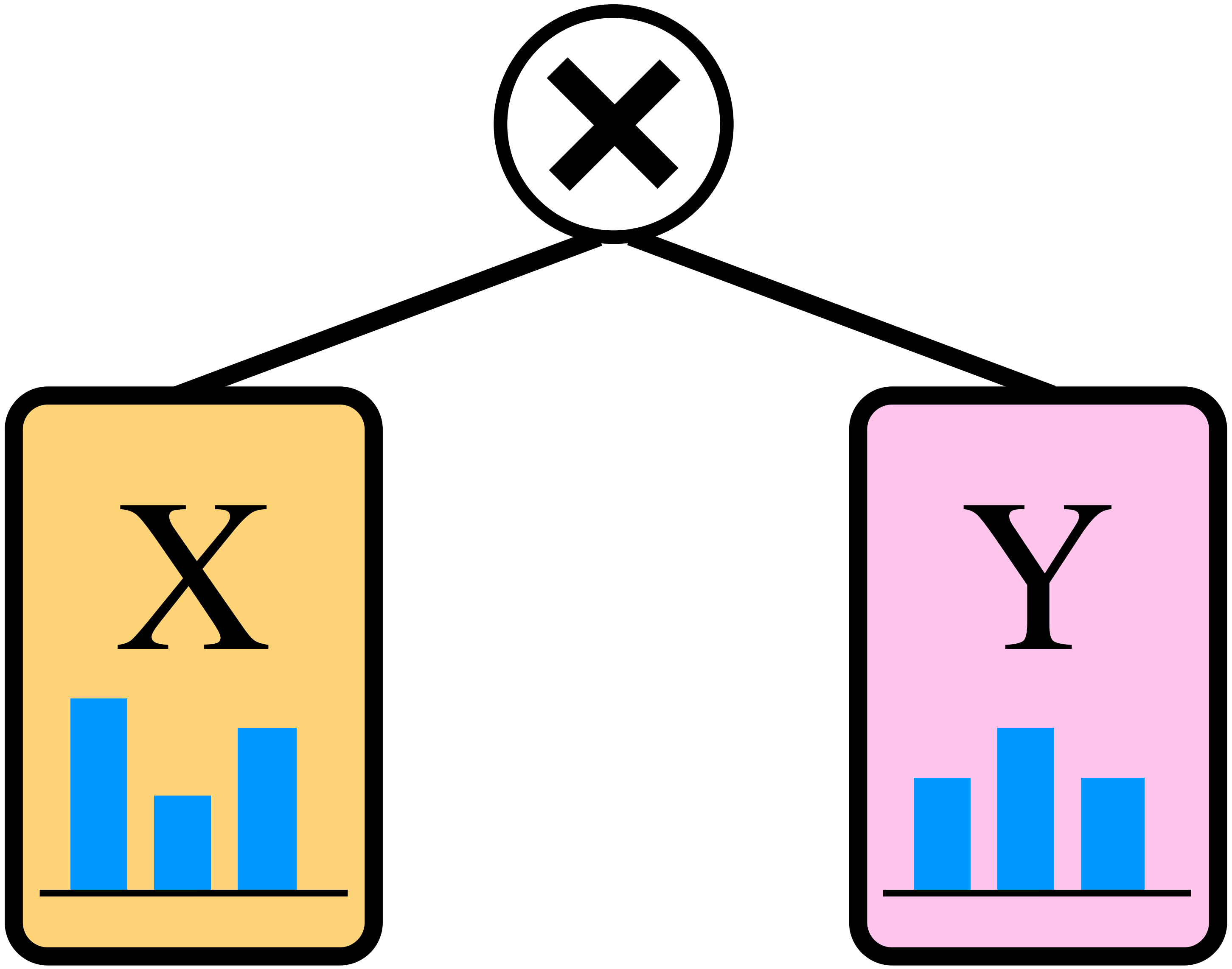
A Univariate Distribution Is an SPN.



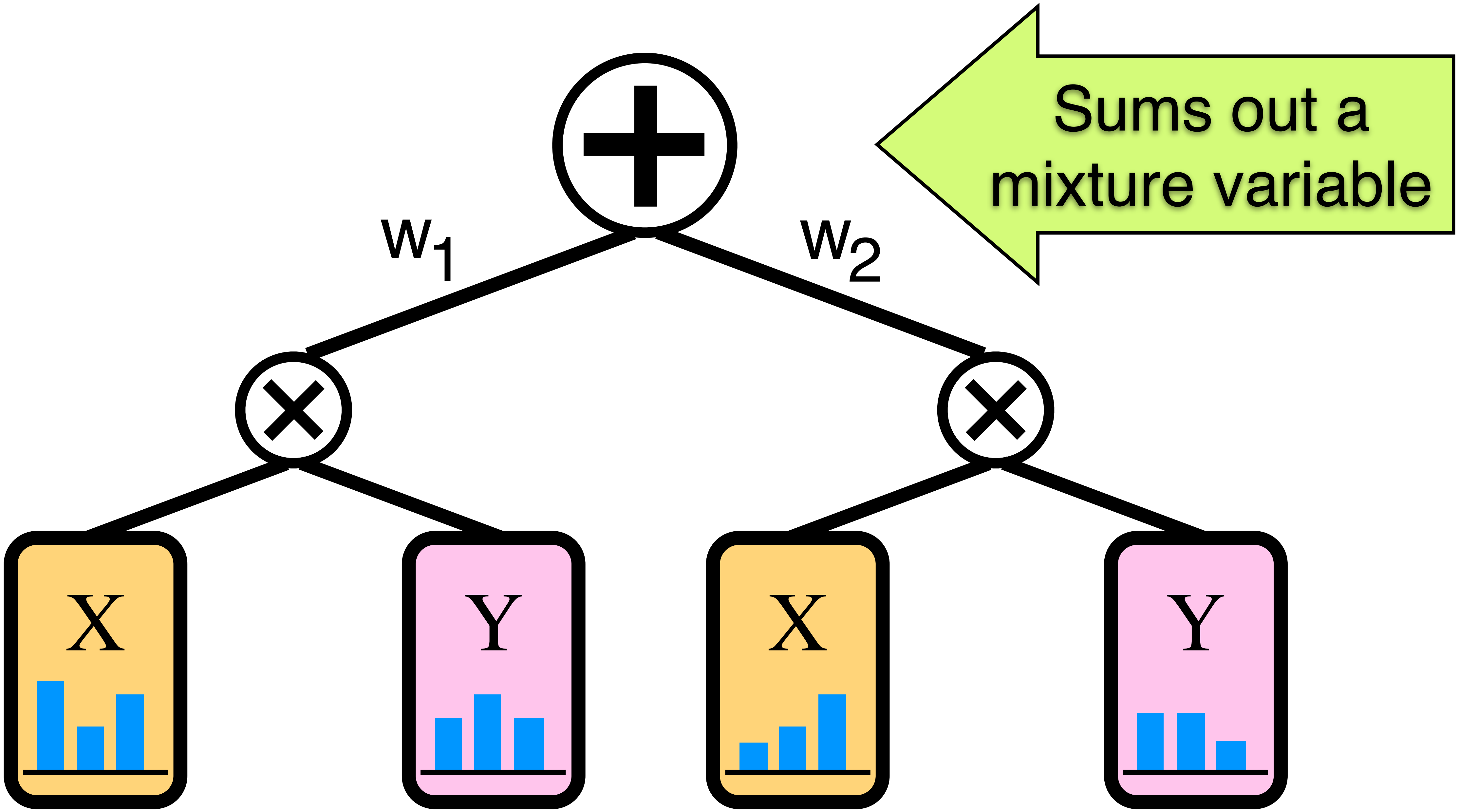


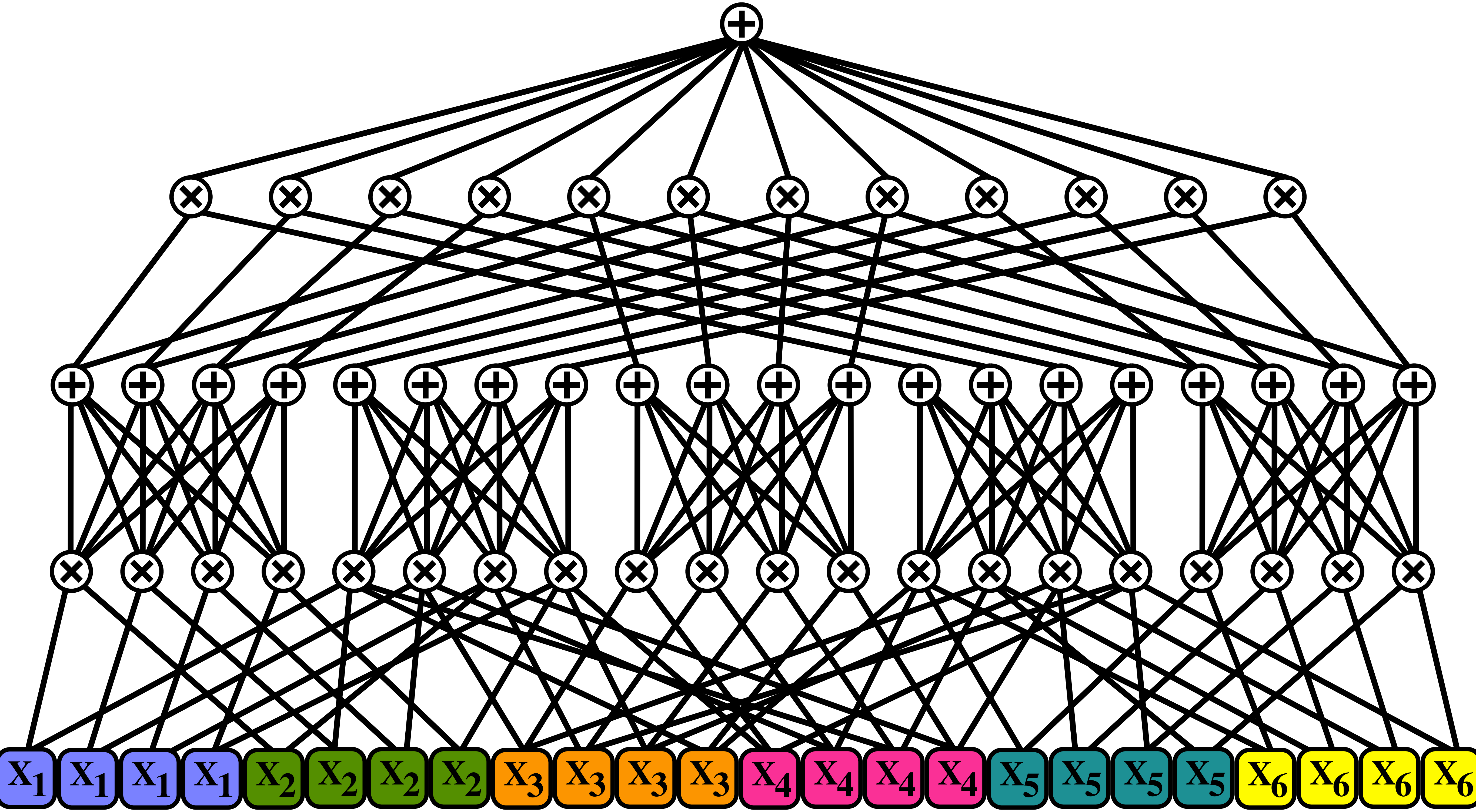
Multinomial Gaussian Poisson ...

A Product of SPNs over Disjoint Variables Is an SPN.

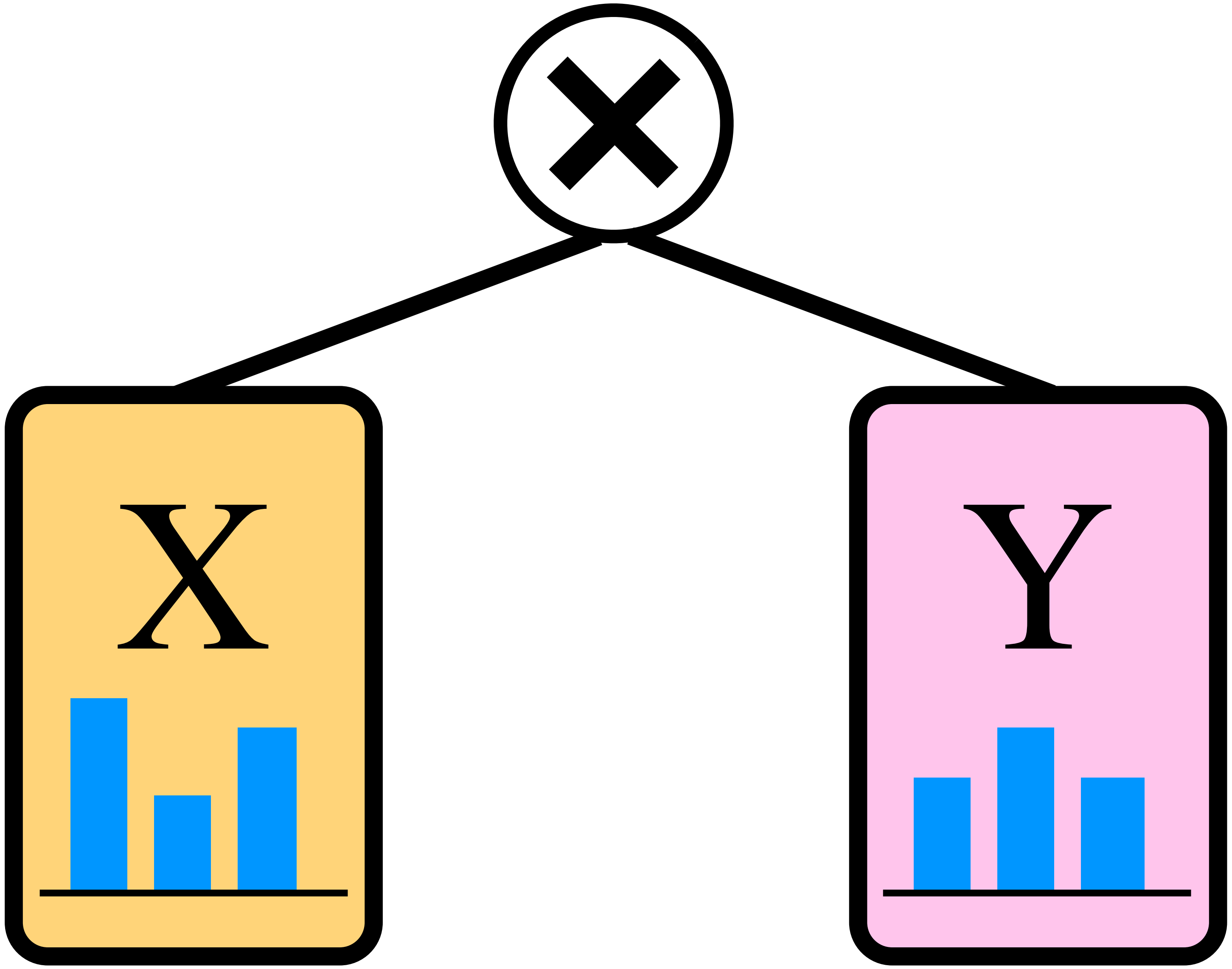


A Weighted Sum of SPNs over the Same Variables Is an SPN.

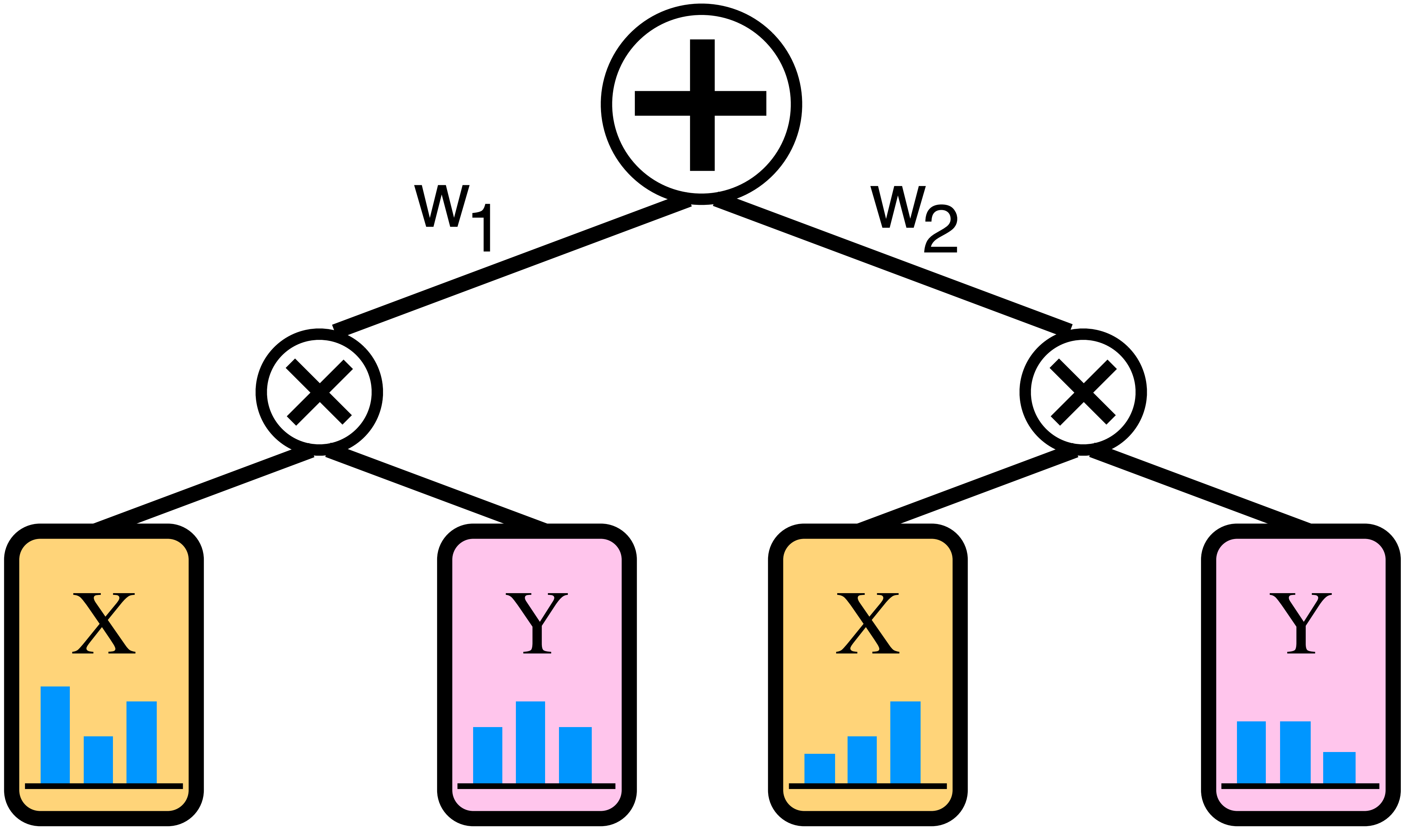



X1 X1 X1 X1 X2 X2 X2 X2 X3 X3 X3 X3 X4 X4 X4 X4 X5 X5 X5 X5 X6 X6 X6 X6

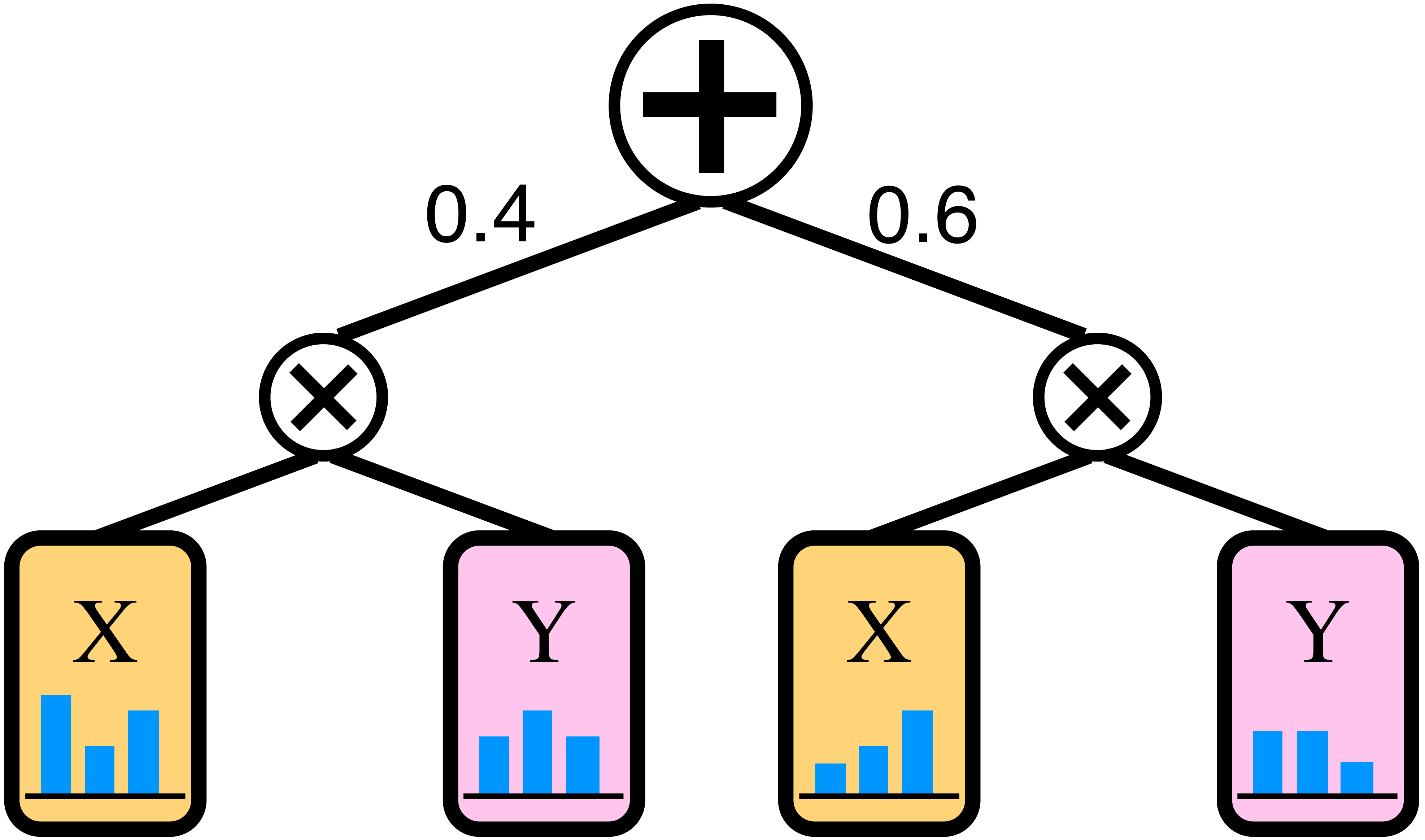
All Marginals Are Computable in Linear Time



All Marginals Are Computable in Linear Time

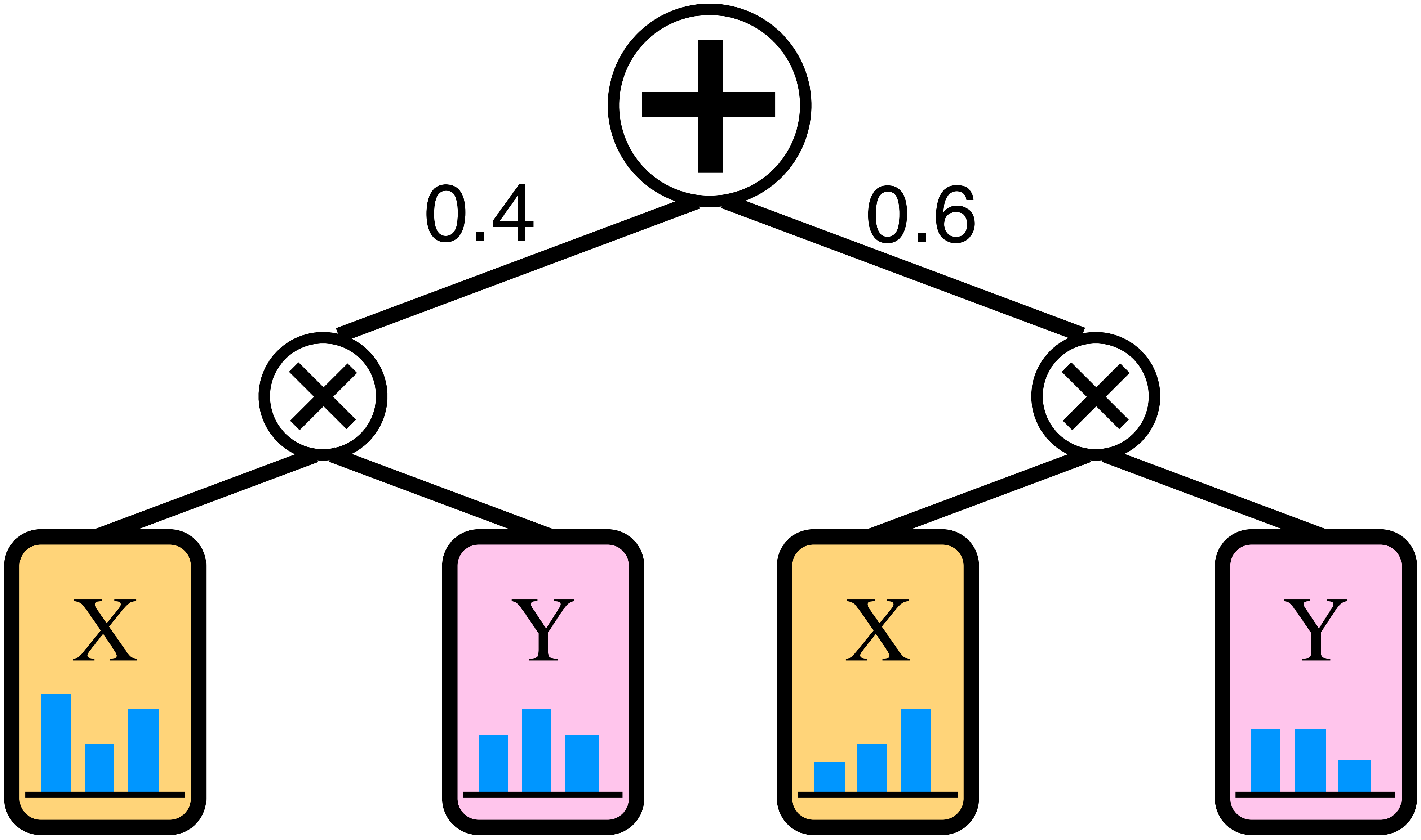


All Marginals Are Computable in Linear Time



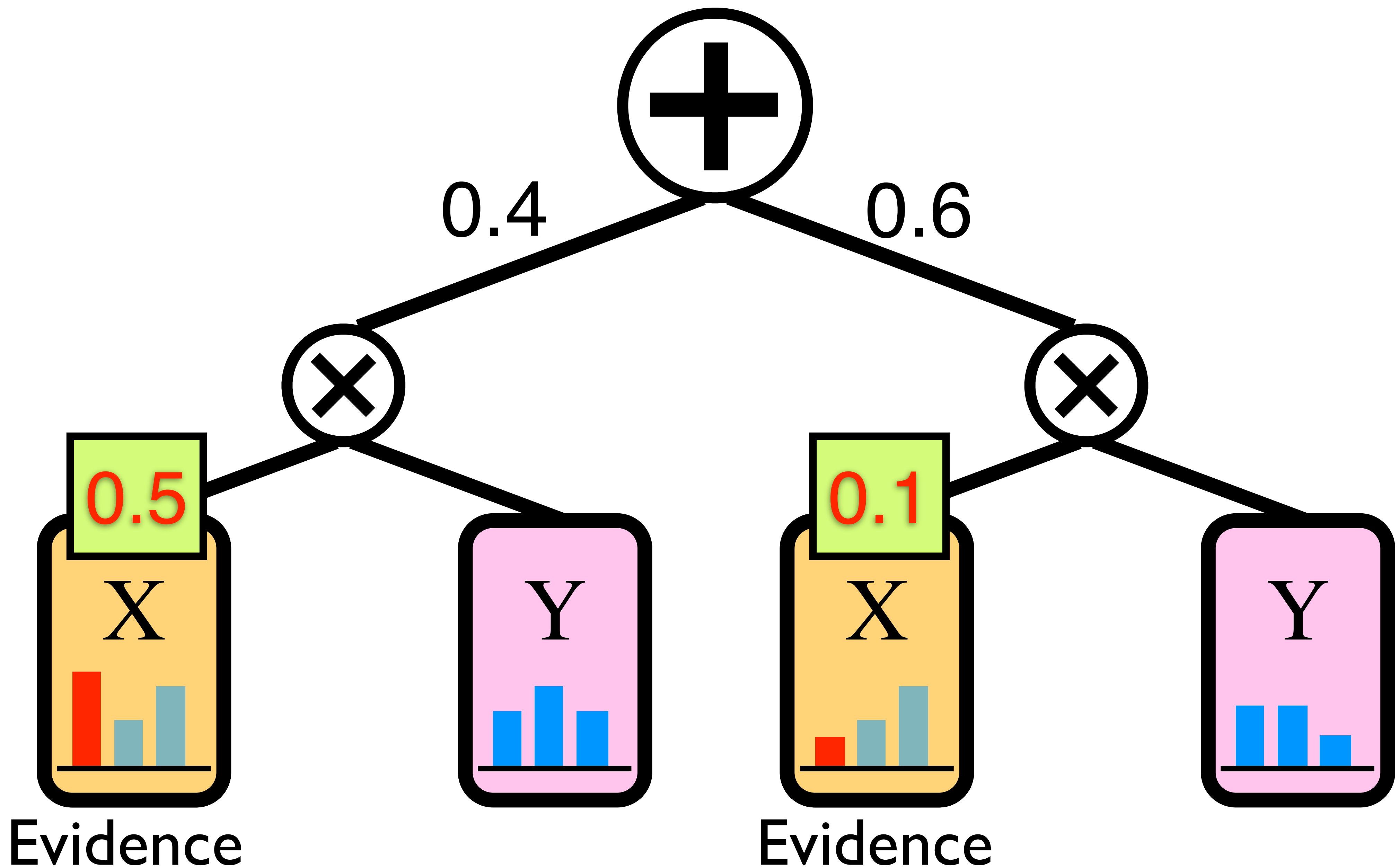
All Marginals Are Computable in Linear Time

$P(X=0)$?



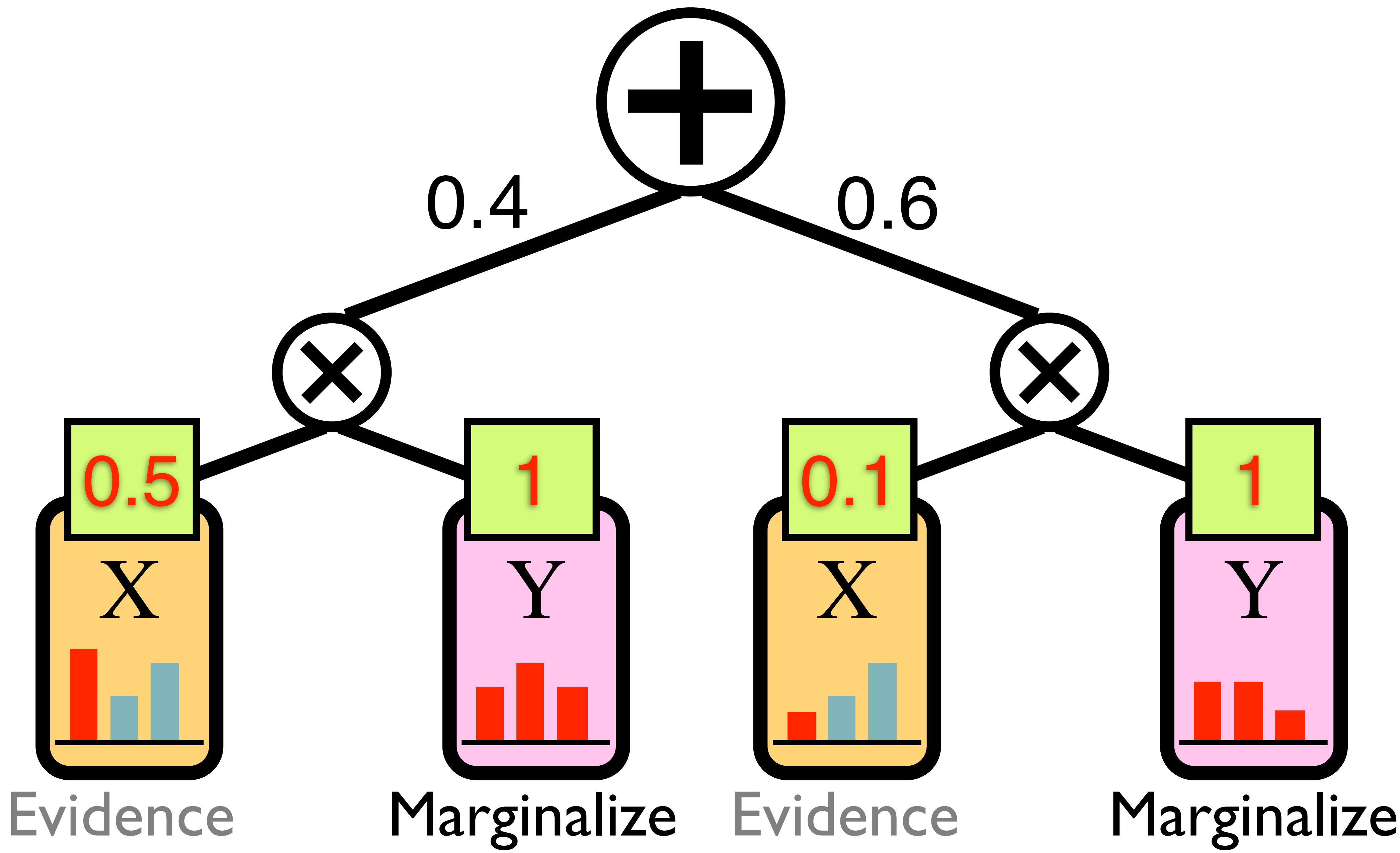
All Marginals Are Computable in Linear Time

$P(X=0)$?



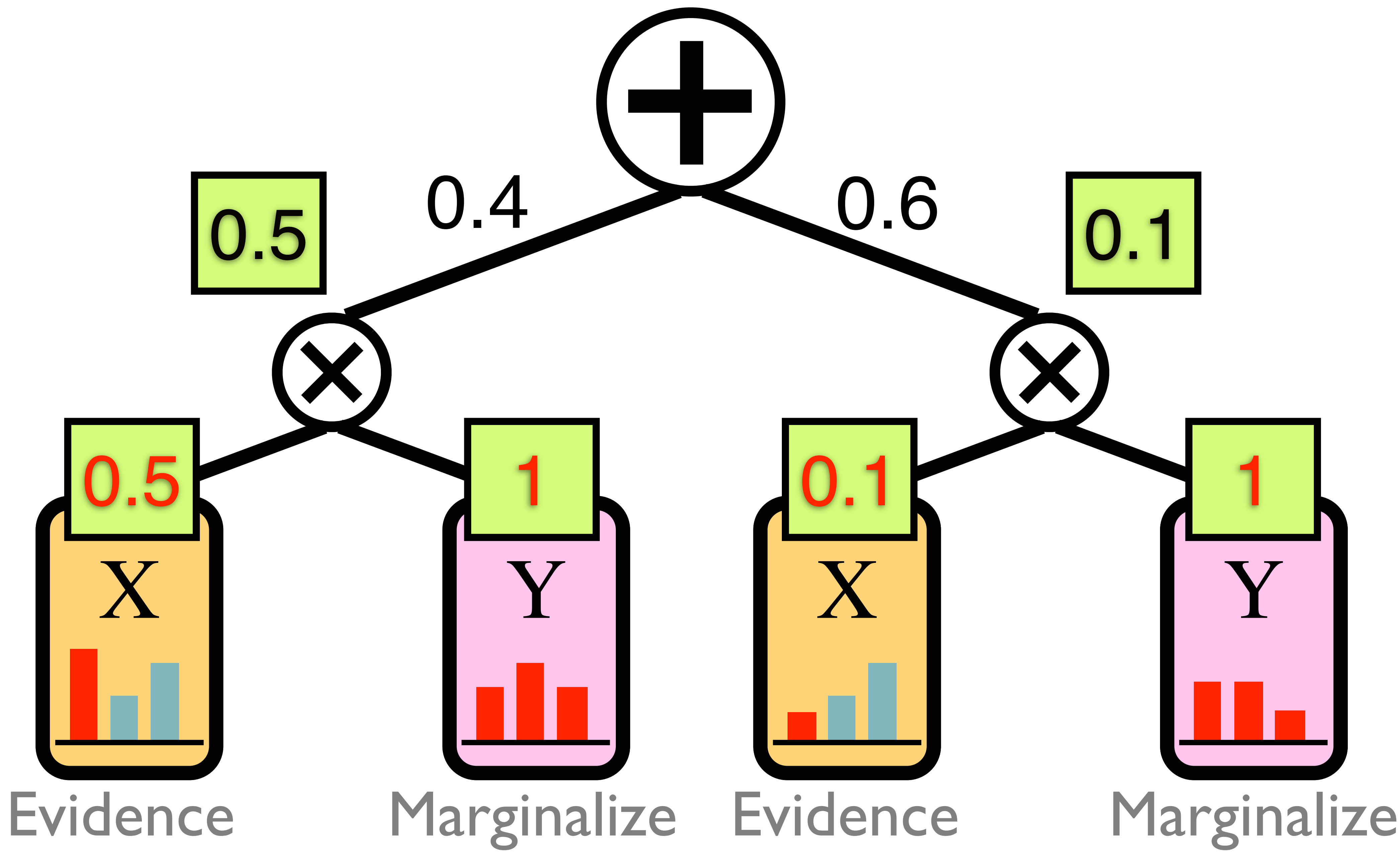
All Marginals Are Computable in Linear Time

$P(X=0)$?



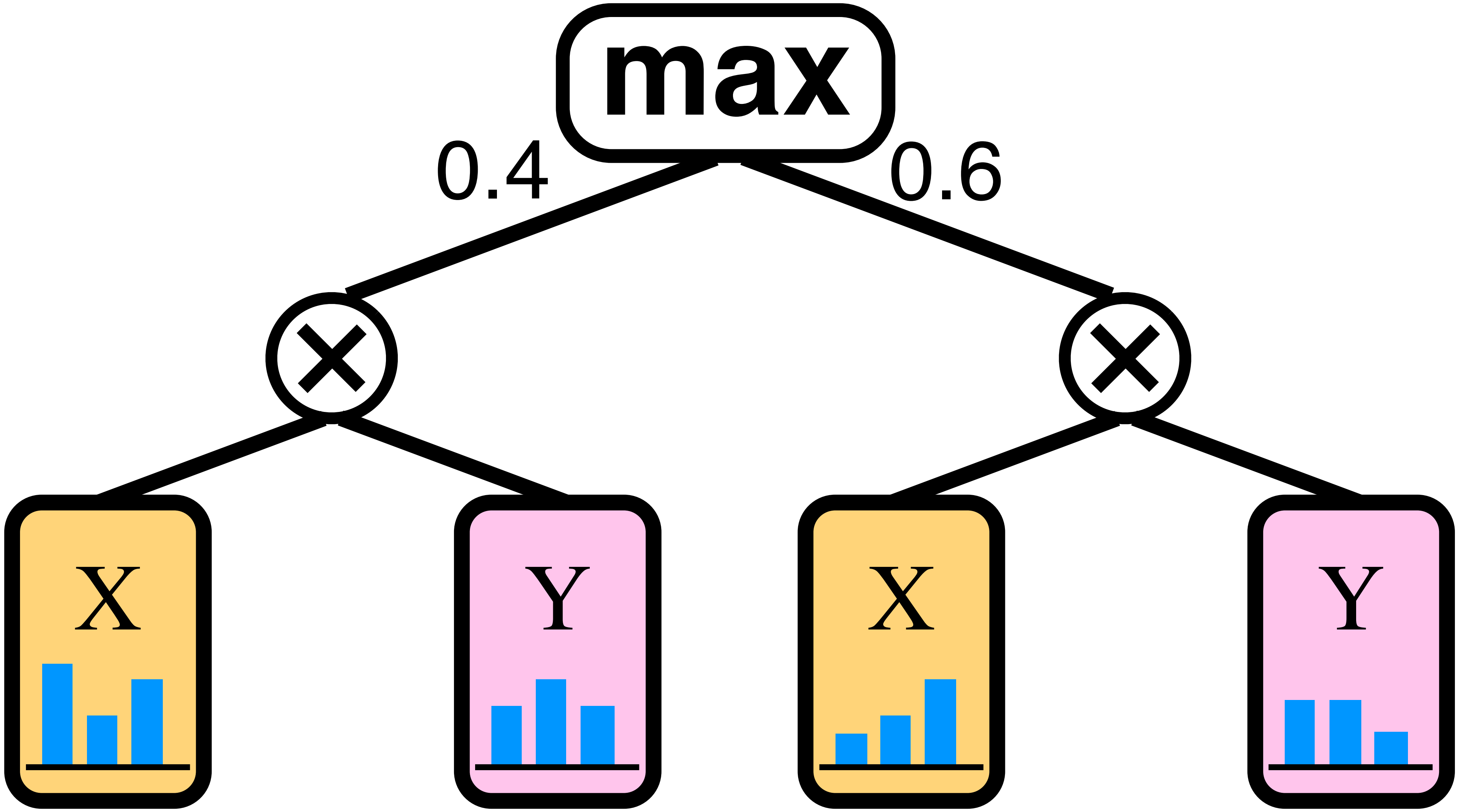
All Marginals Are Computable in Linear Time

$P(X=0) =$ 0.26



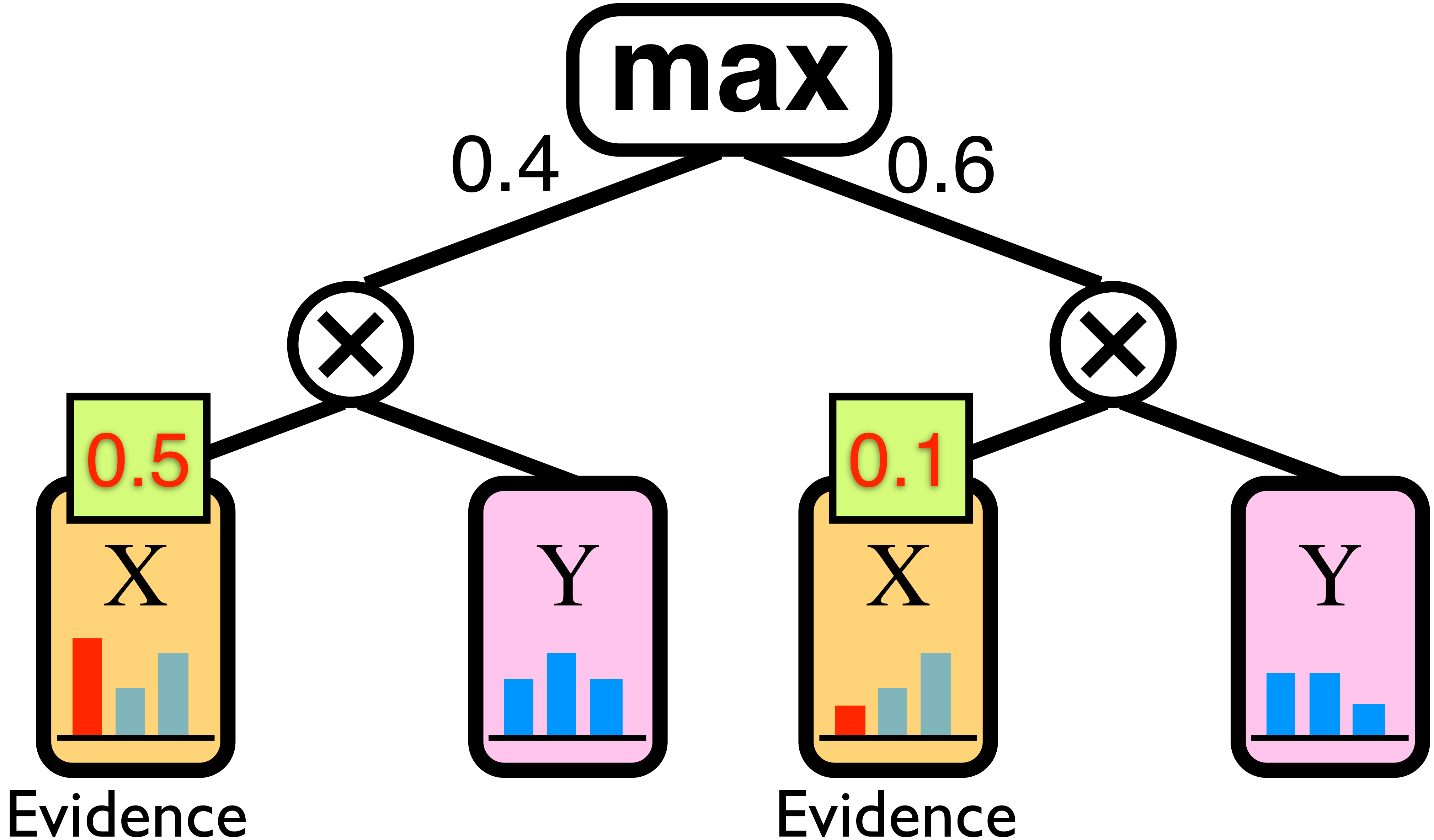
All MAP States Are Computable in Linear Time

$$\max_y P(X=0, Y=y) \;?$$



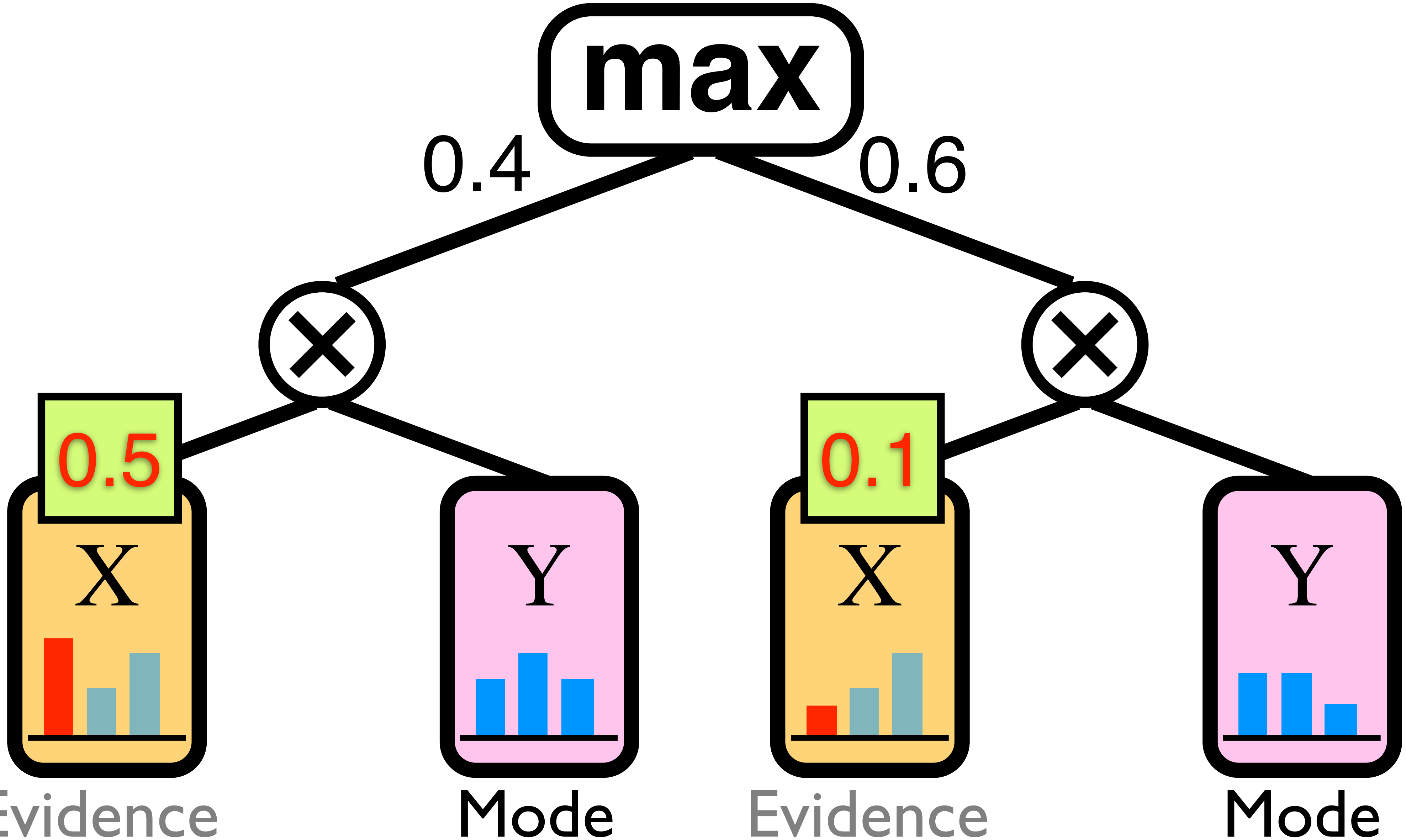
All MAP States Are Computable in Linear Time

$$\max_y P(X=0, Y=y)\ ?$$



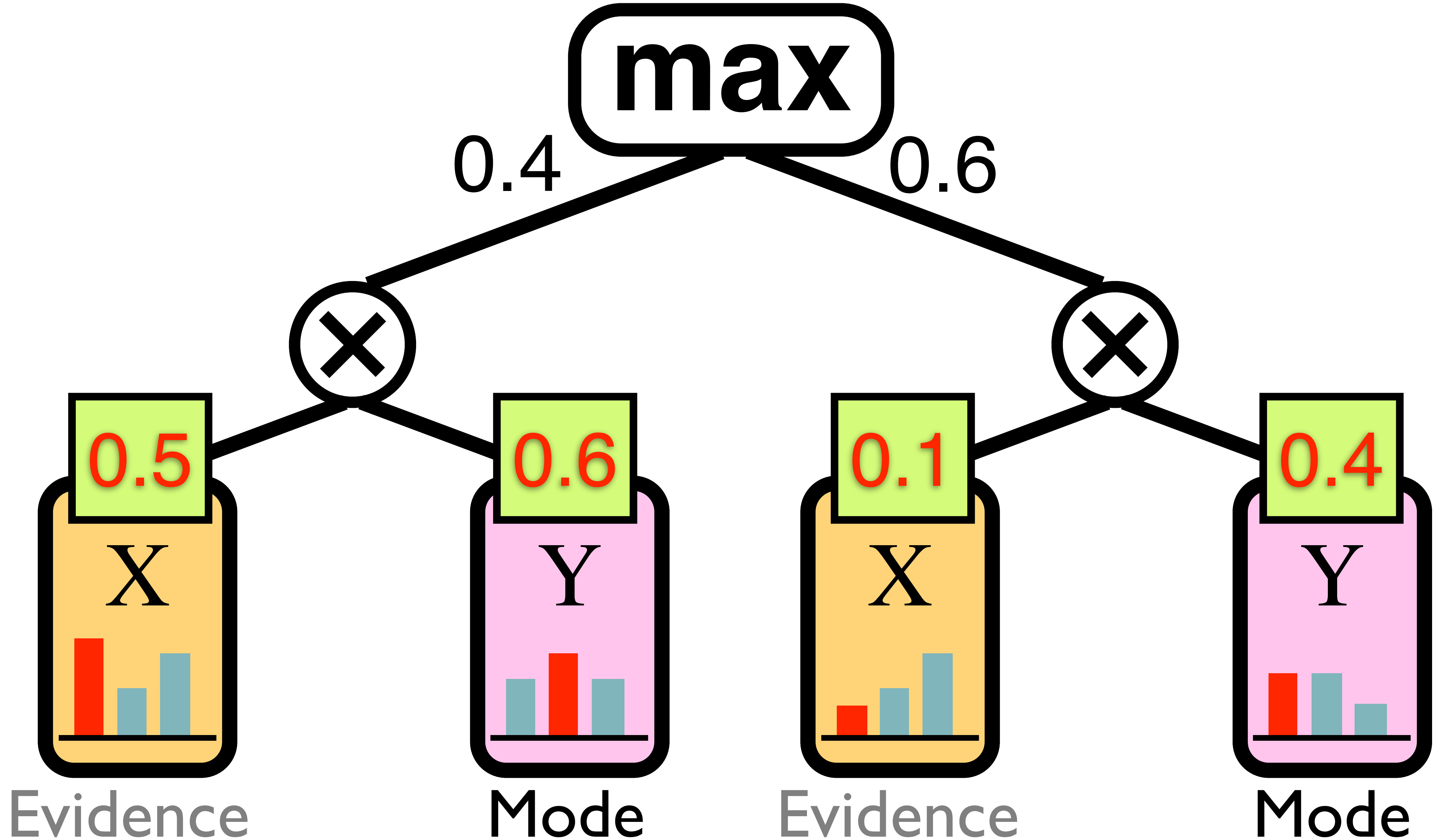
All MAP States Are Computable in Linear Time

$$\max_y P(X{=}0, Y{=}y)\ ?$$



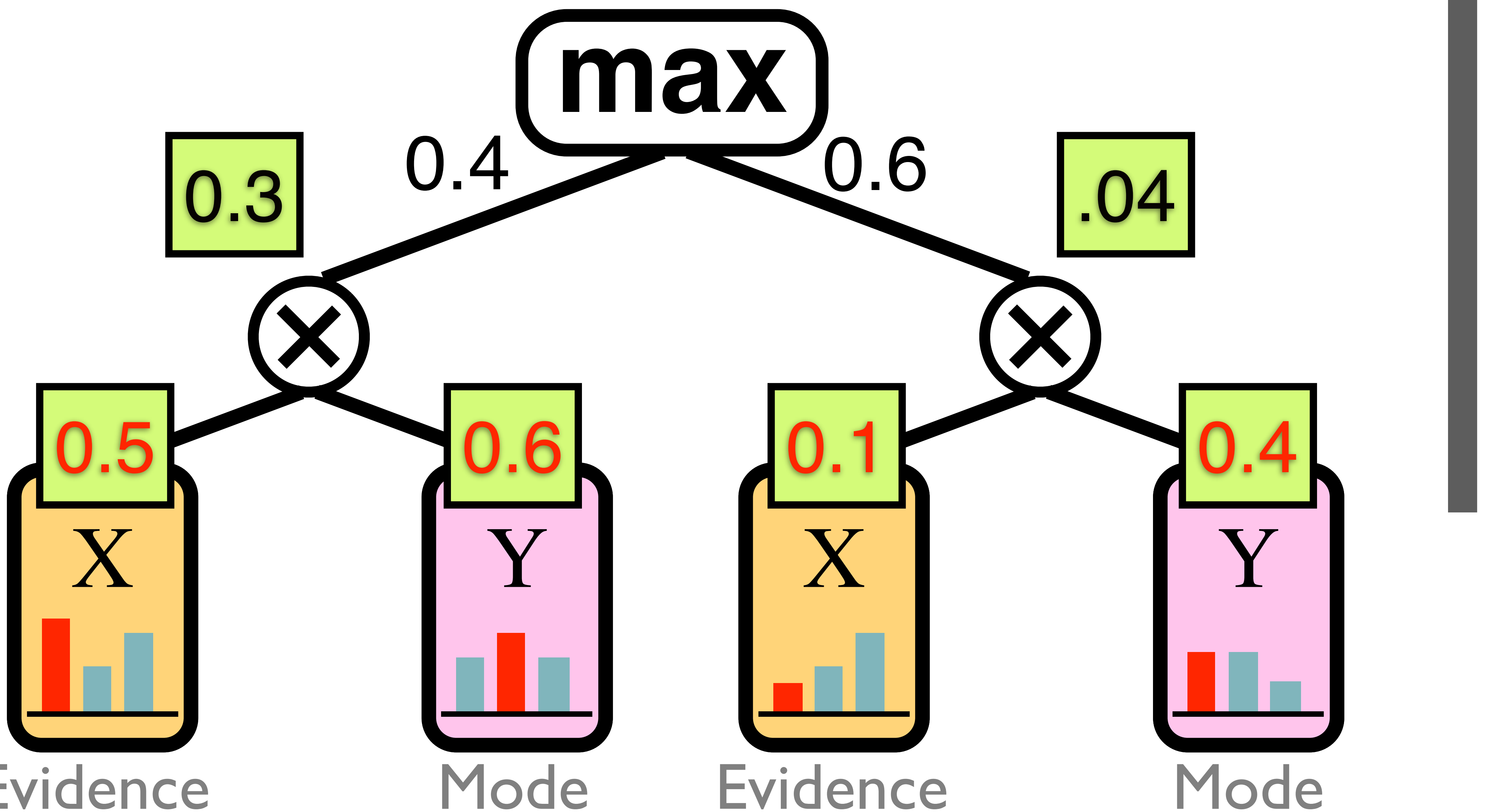
All MAP States Are Computable in Linear Time

$$\max_y P(X{=}0, Y{=}y)\ ?$$



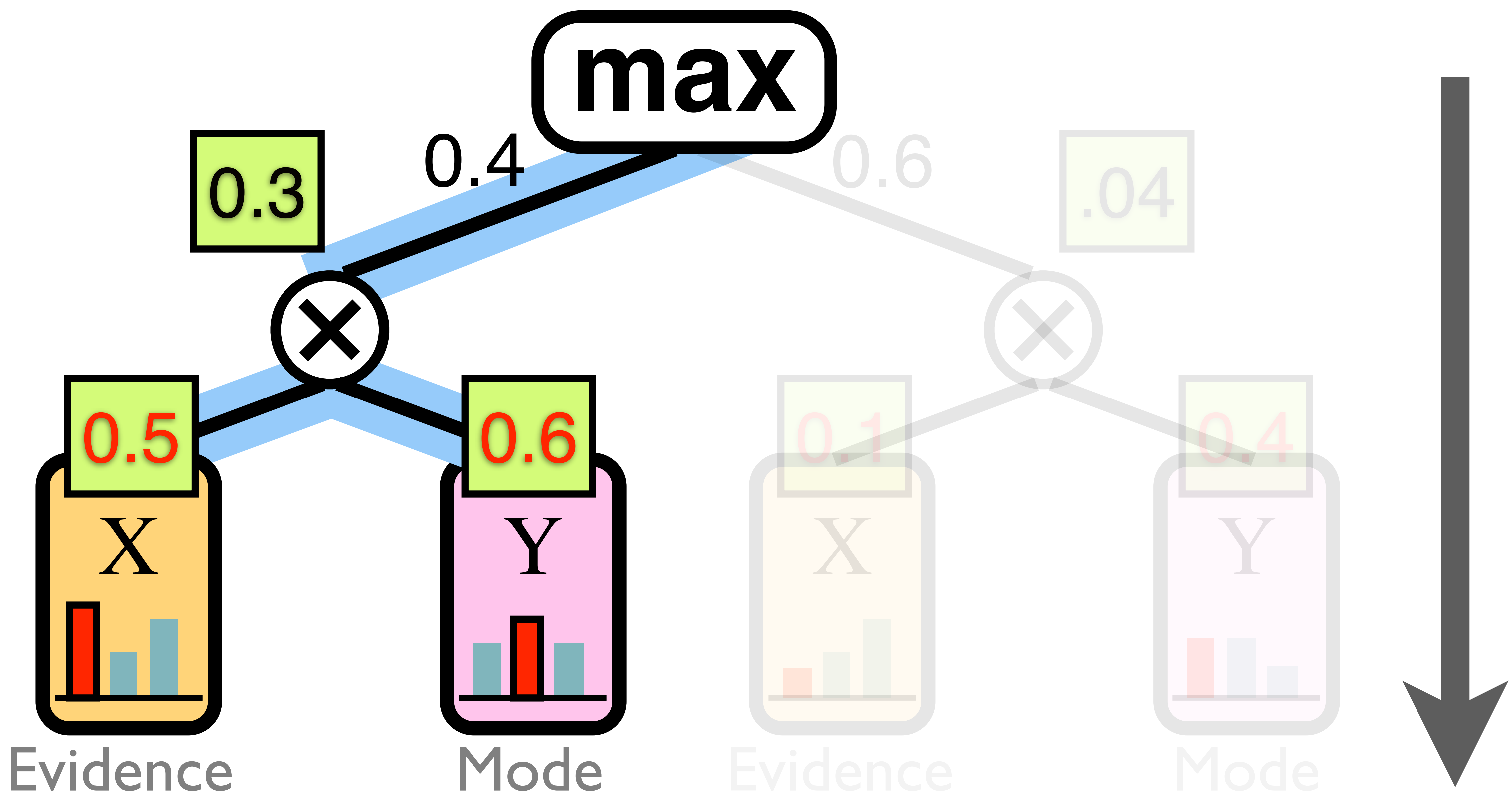
All MAP States Are Computable in Linear Time

$$\max_y P(X=0, Y=y) = 0.12$$



All MAP States Are Computable in Linear Time

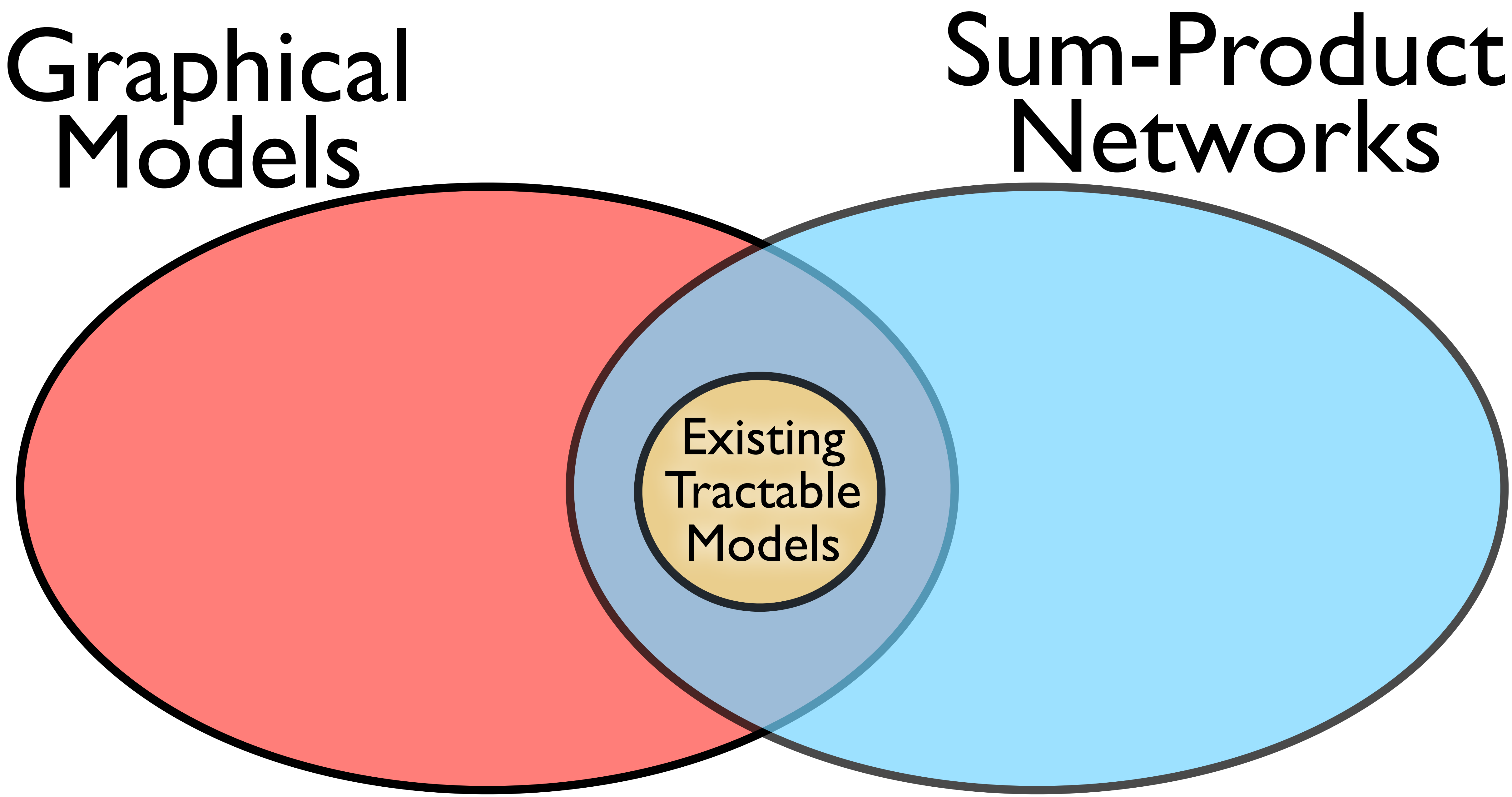
$$\max_y P(X{=}0, Y{=}y) = 0.12$$



Special Cases of SPNs

- Junction trees
- Hierarchical mixture models
- Non-recursive probabilistic context-free grammars
- Models with context-specific independence
- Models with determinism
- Other high-treewidth models

Compactly Representable Probability Distributions



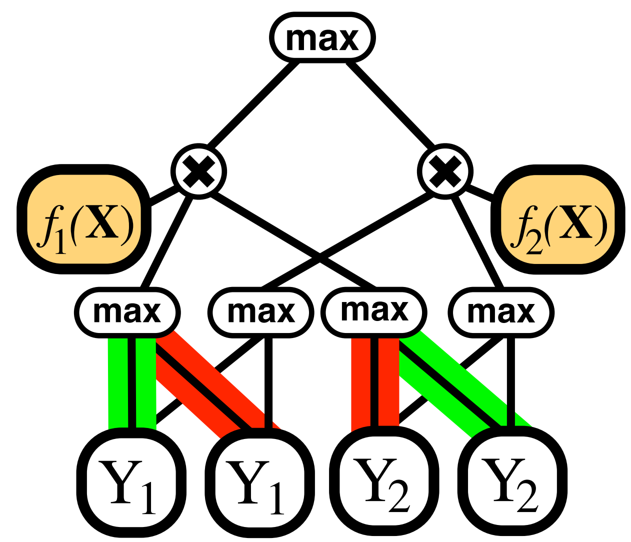
Learning SPNs

Update	**Soft Inference** (Marginals)	**Hard Inference** (MAP States)
Gen. EM	✓	✓
Gen. Gradient	✓	
Disc. Gradient		

Poon & Domingos, UAI 2011

Motivation

SPN Review



Discriminative Training

Experiments

Discriminative SPNs

$$P(\mathbf{Y}|\mathbf{X})$$

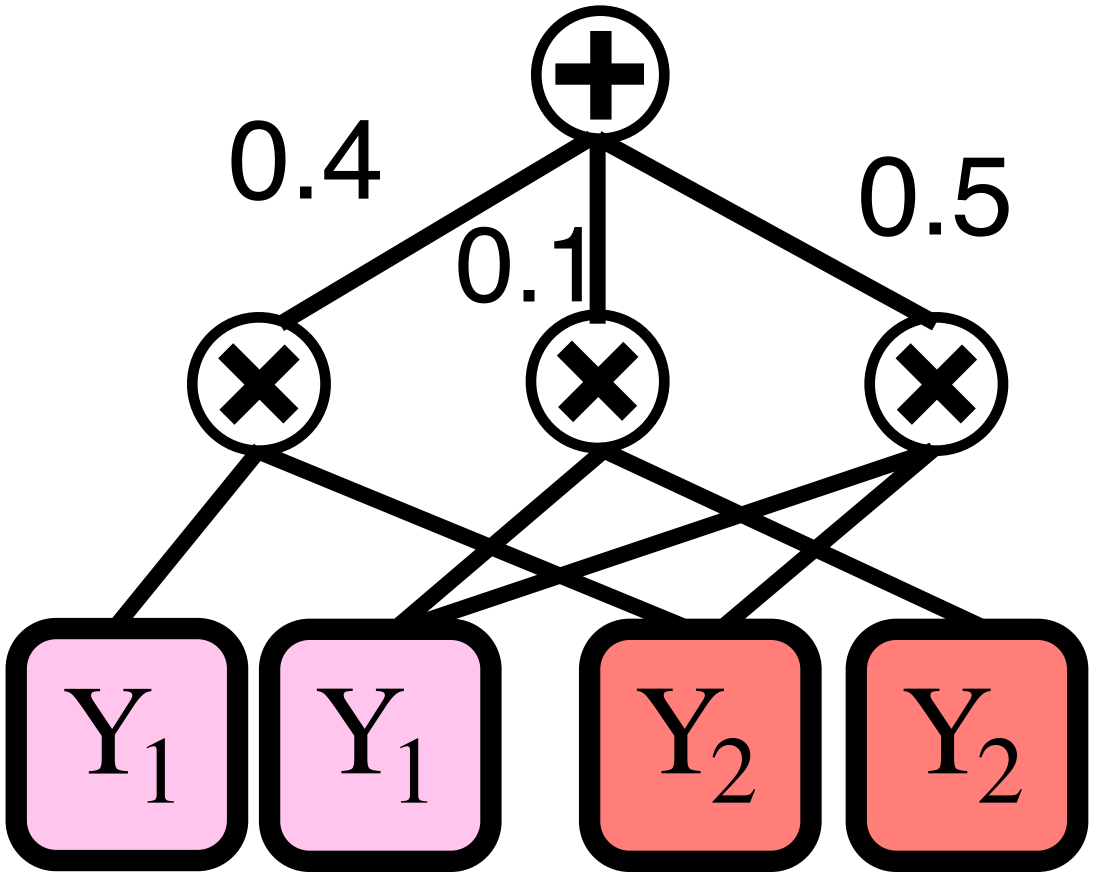
$\mathbf{Y}$ Query

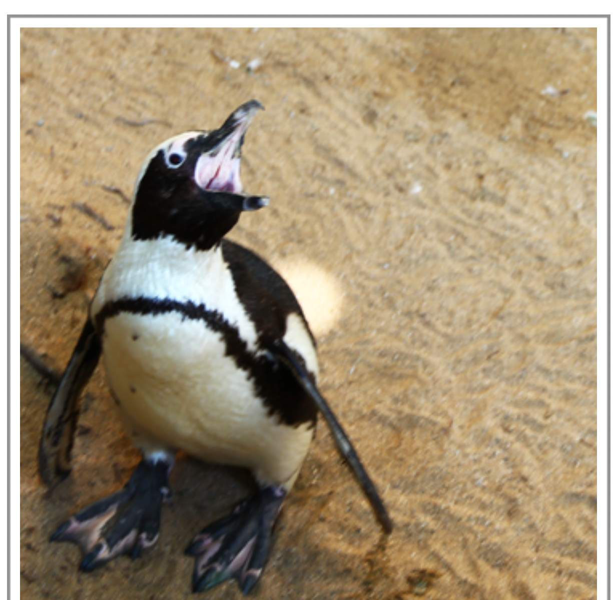
$\mathbf{H}$ Hidden

$\mathbf{X}$ Evidence

Treat as constants

Discriminative SPNs



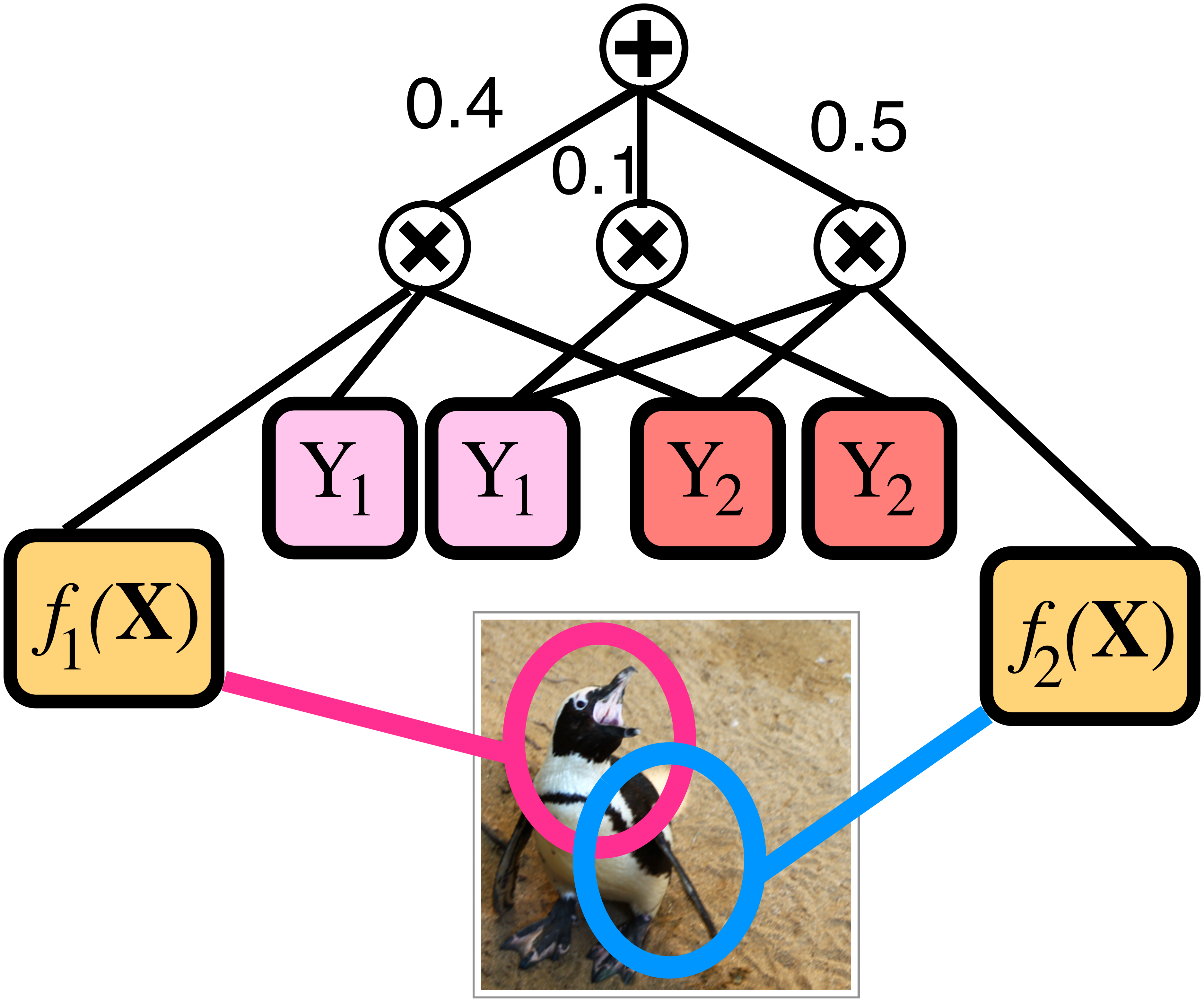


H Hidden

Y Query

X Evidence

Discriminative SPNs



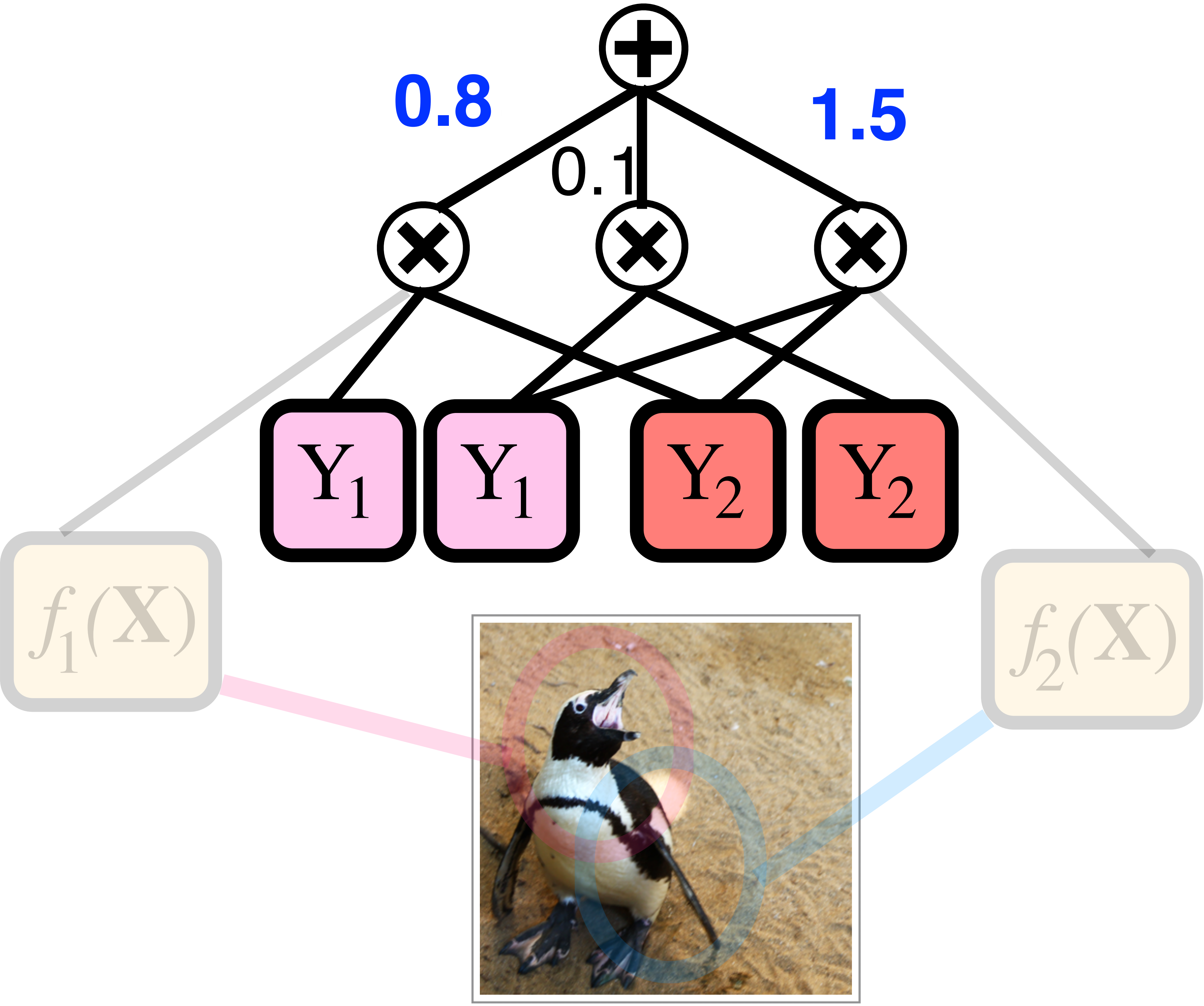
$\mathbf{H}$ Hidden

$\mathbf{Y}$ Query

$f(\mathbf{X})$ Features (non-negative)

$\mathbf{X}$ Evidence

Discriminative SPNs



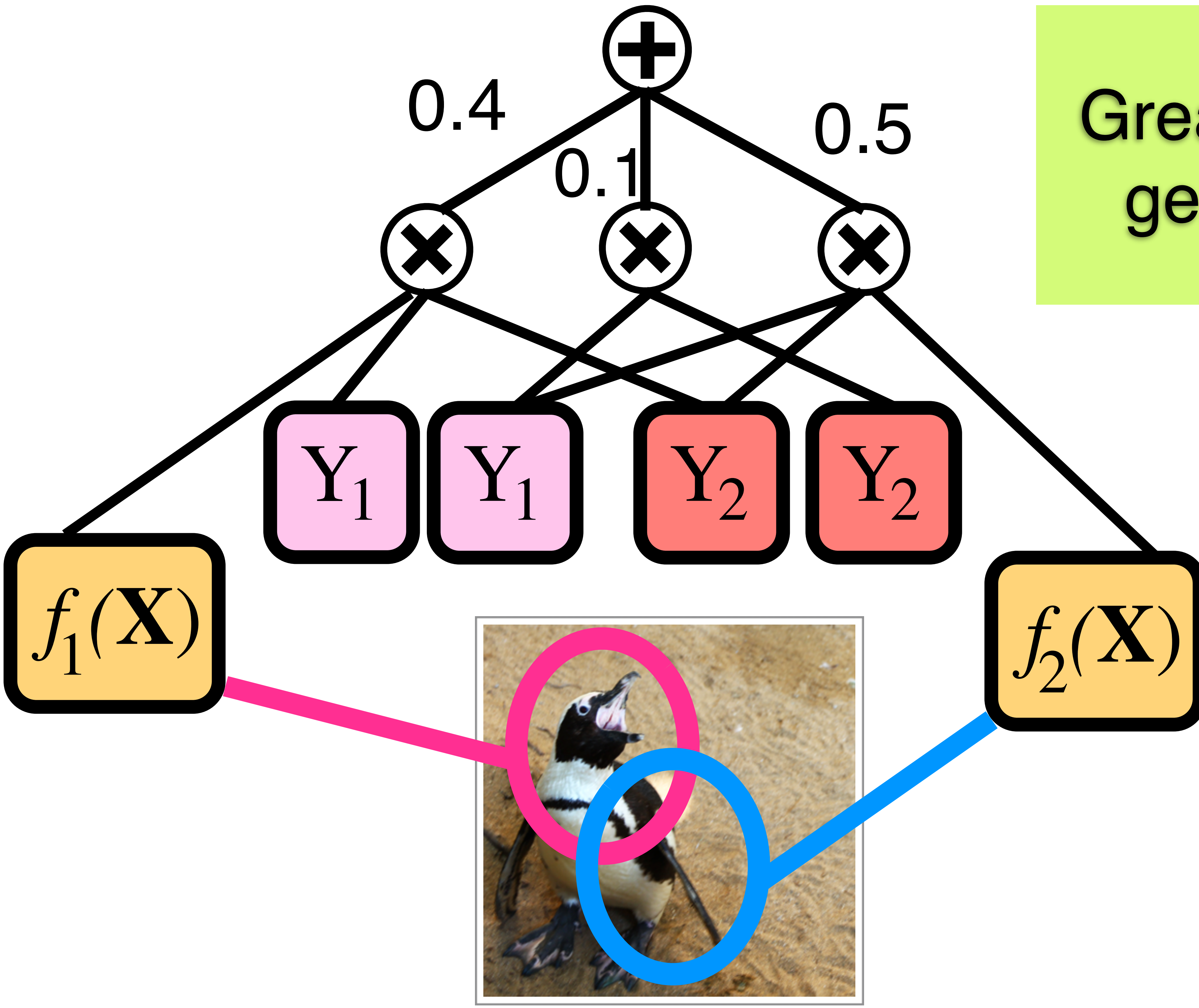
H Hidden

Y Query

$f(\mathbf{X})$ Features (non-negative)

X Evidence

Discriminative SPNs



Greater variety than generative SPNs

Y Query

$f(\mathbf{X})$ Features (non-negative)

X Evidence

Discriminative Training

$$\nabla \log P(\mathbf{y}|\mathbf{x}) = \nabla \log \frac{P(\mathbf{y},\mathbf{x})}{P(\mathbf{x})} =$$

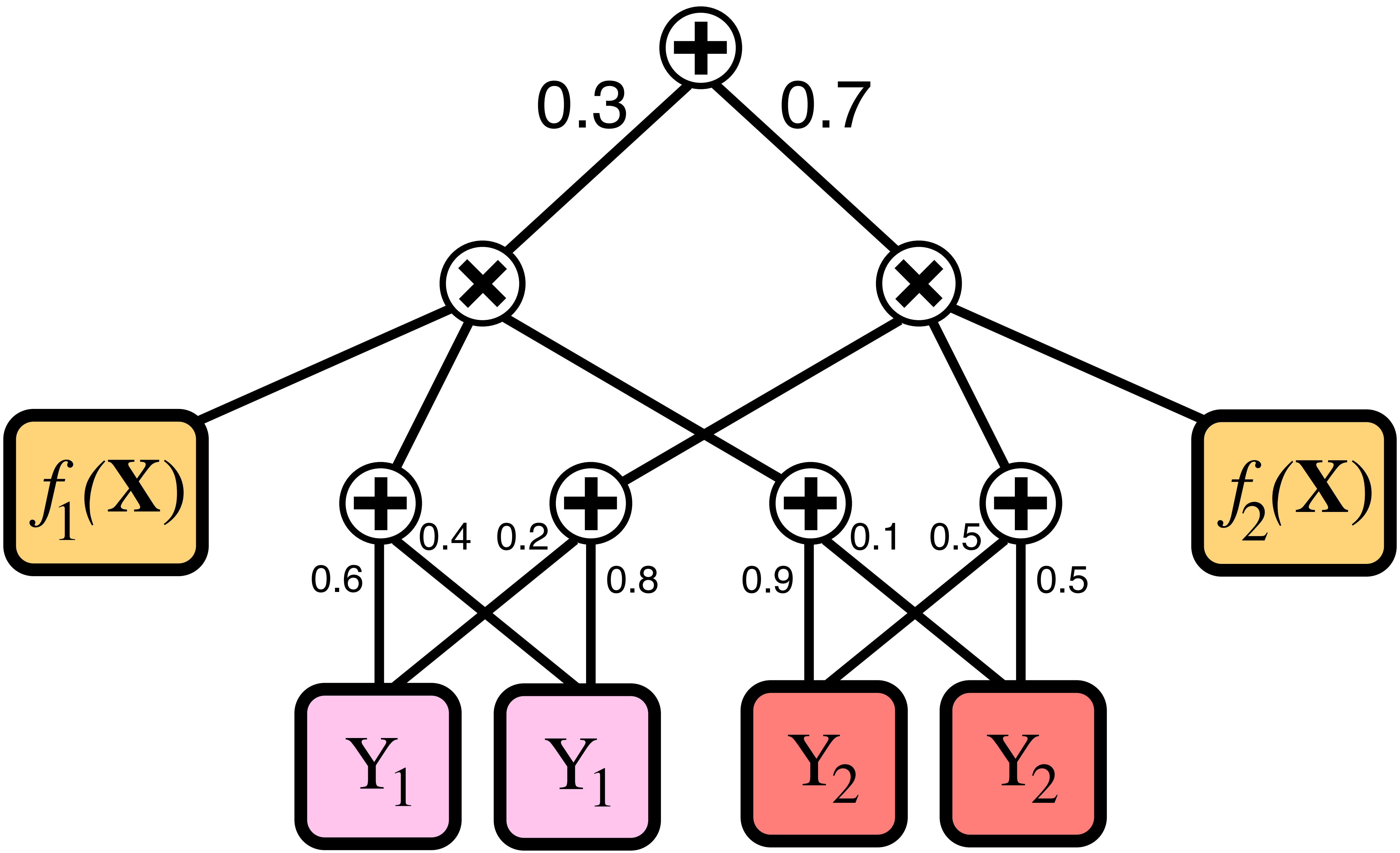
$$\nabla \log \sum_{\mathbf{h}} P(\mathbf{y},\mathbf{h},\mathbf{x}) - \nabla \log \sum_{\mathbf{y}',\mathbf{h}} P(\mathbf{y}',\mathbf{h},\mathbf{x})$$

Tractable!

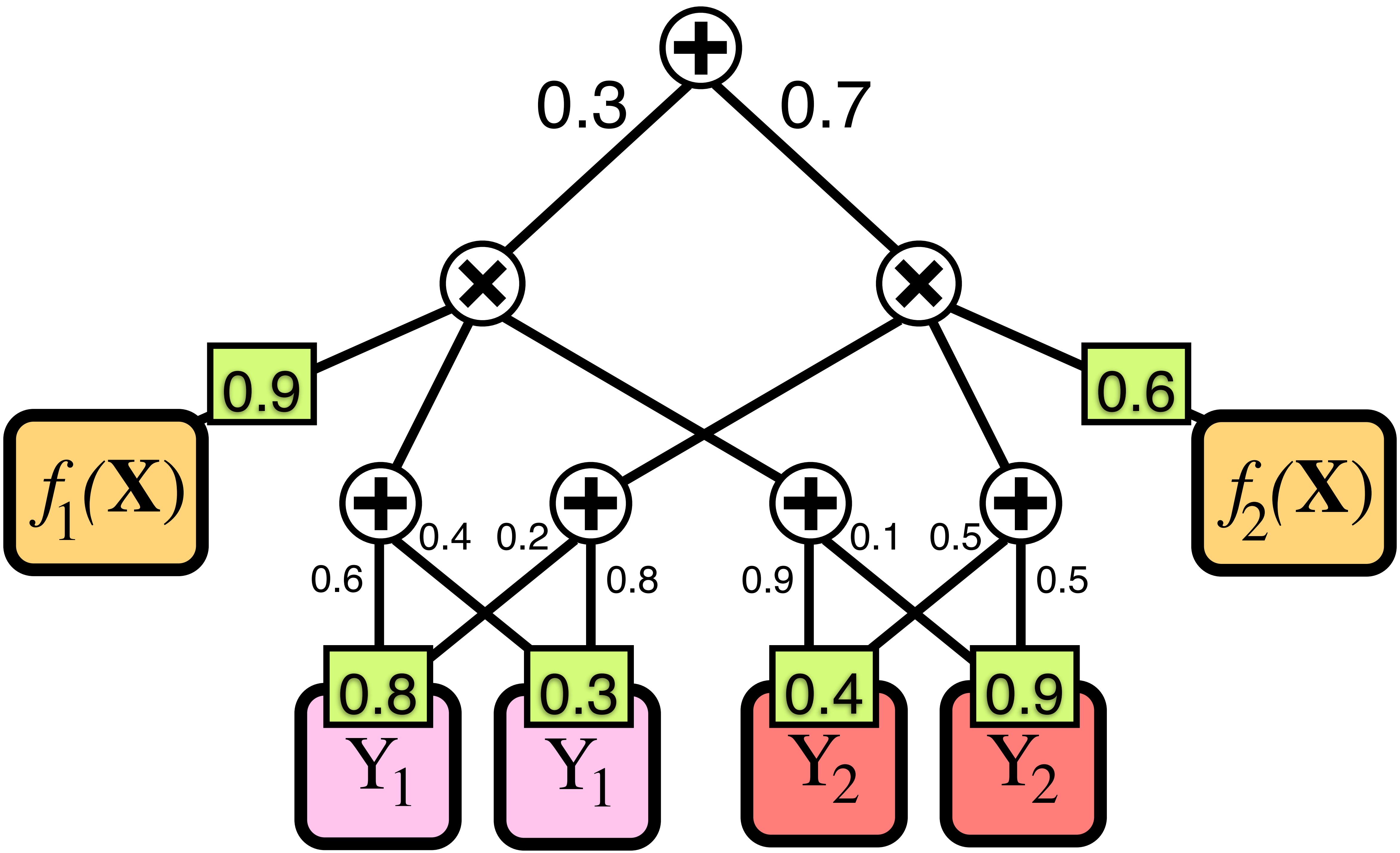
Correct label

Best guess

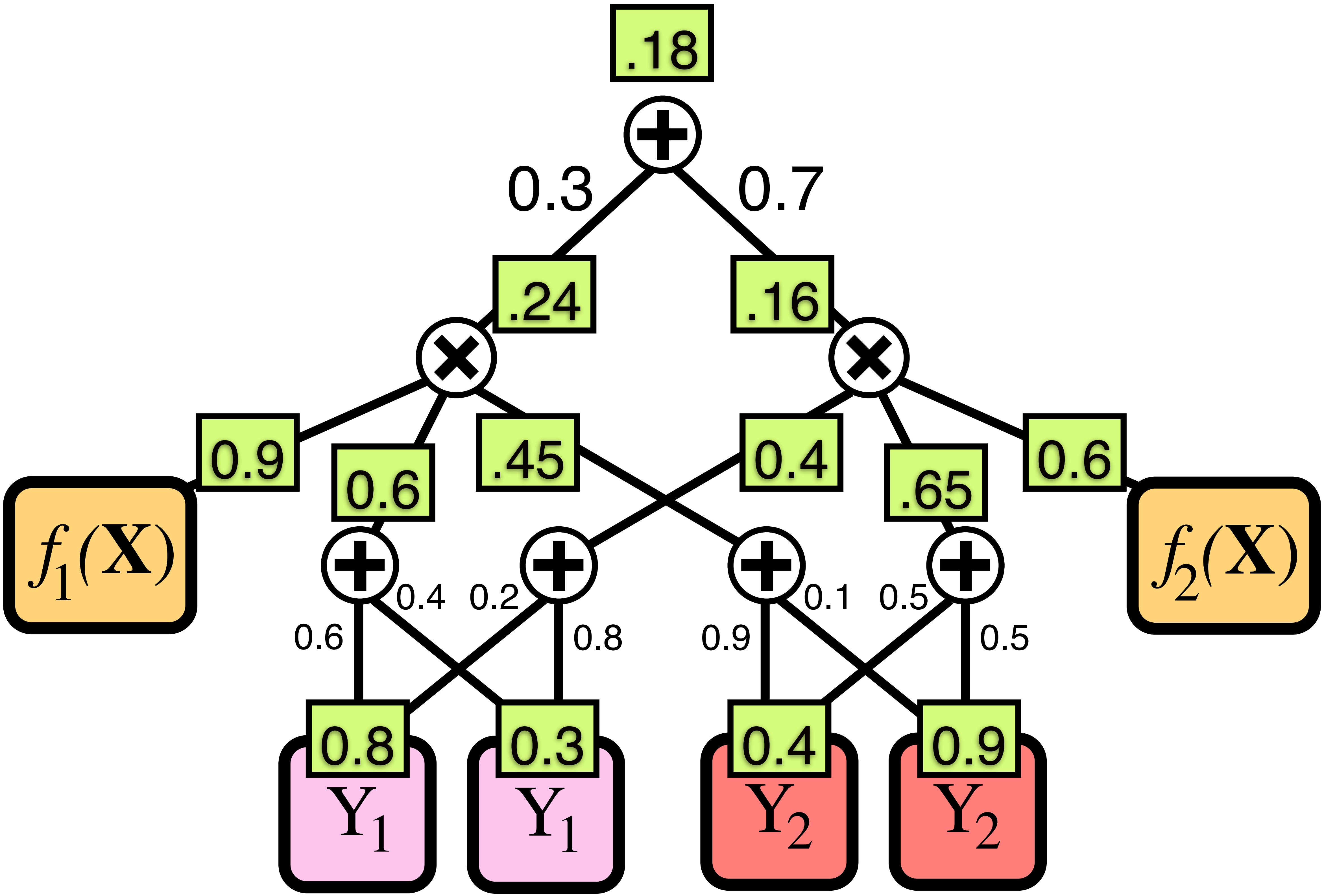
SPN Backpropagation



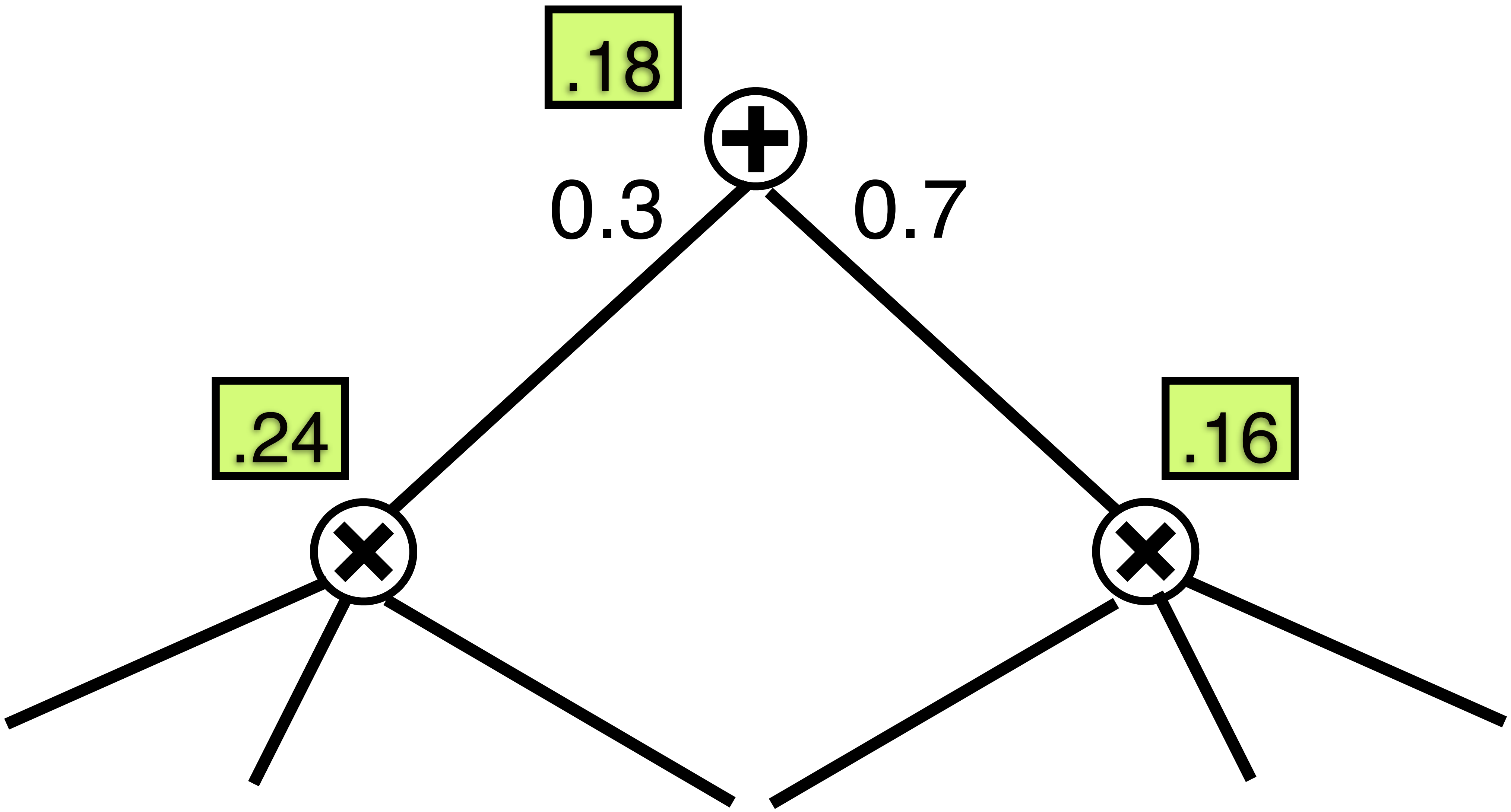
SPN Backpropagation



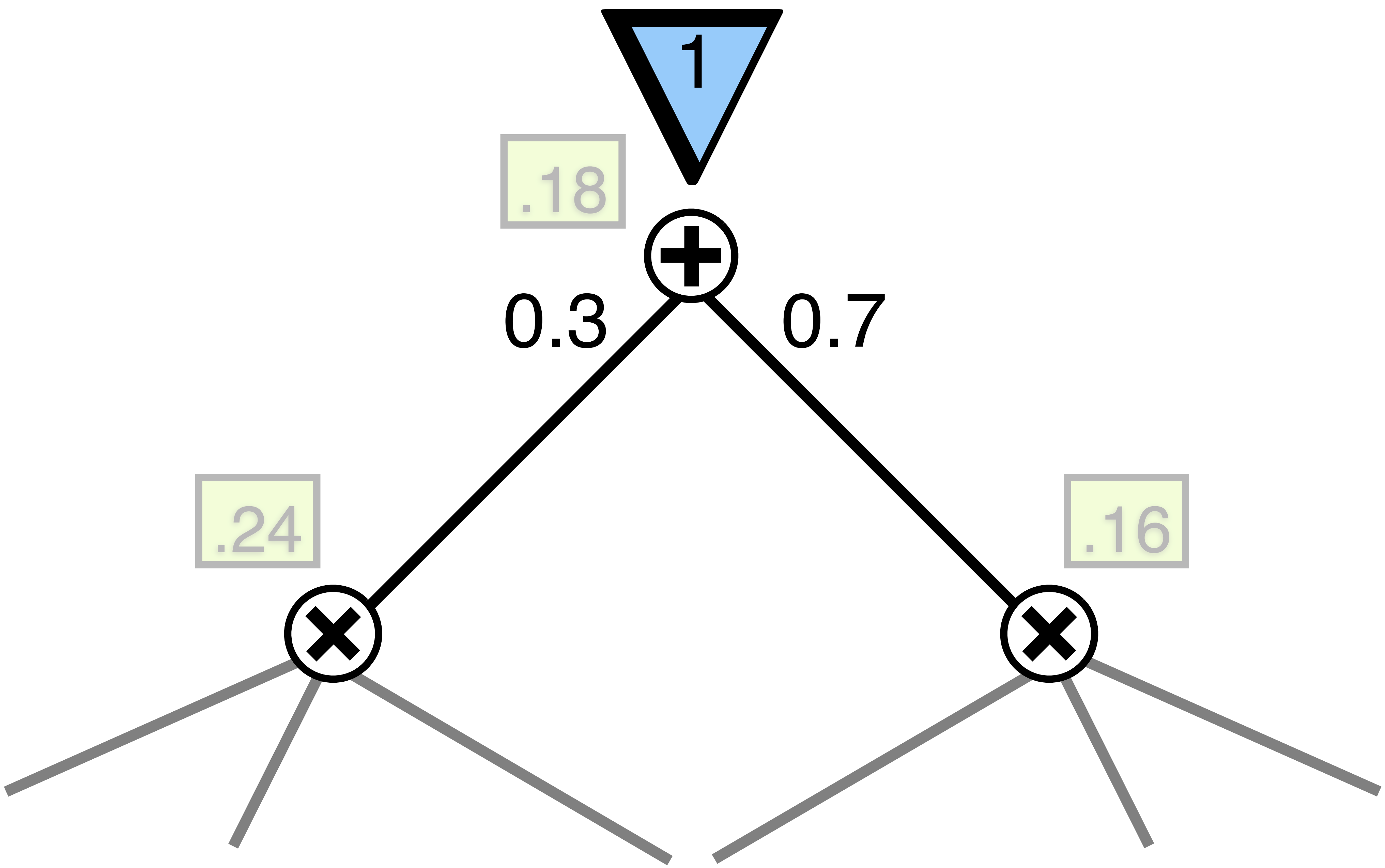
SPN Backpropagation



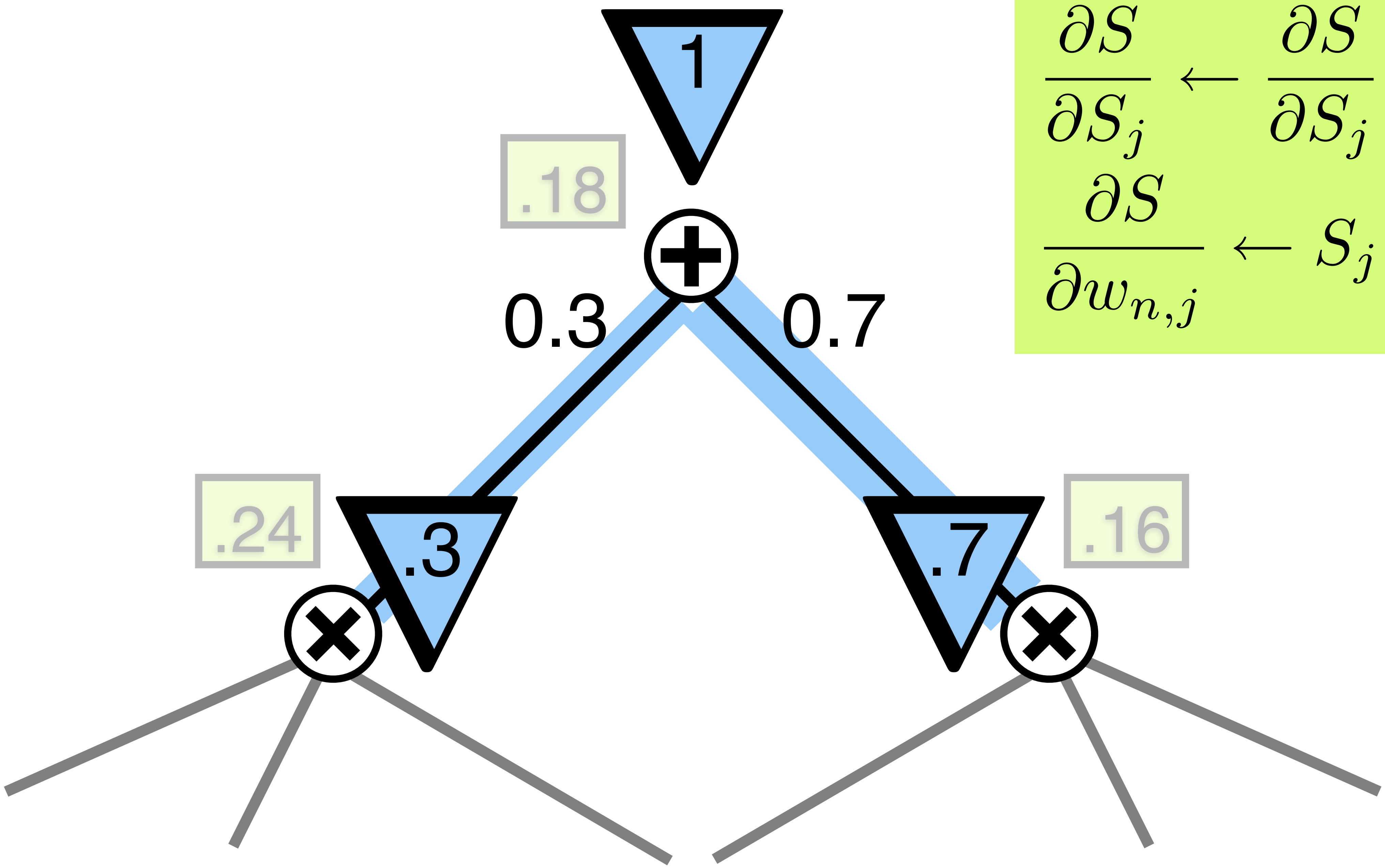
SPN Backpropagation



SPN Backpropagation



SPN Backpropagation

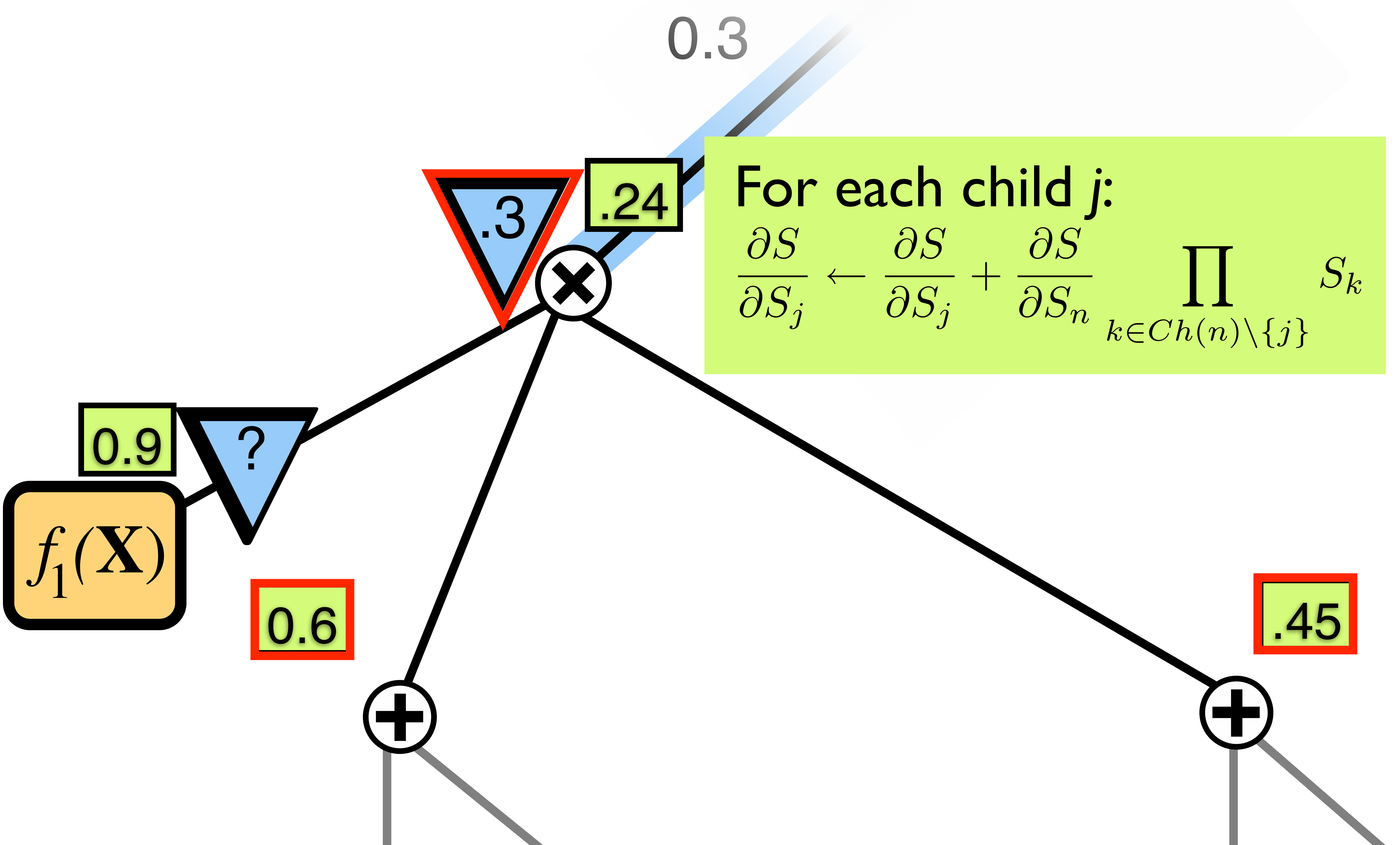


For each child *j*:

$$\frac{\partial S}{\partial S_j} \leftarrow \frac{\partial S}{\partial S_j} + w_{n,j} \frac{\partial S}{\partial S_n}$$

$$\frac{\partial S}{\partial w_{n,j}} \leftarrow S_j \frac{\partial S}{\partial S_n}$$

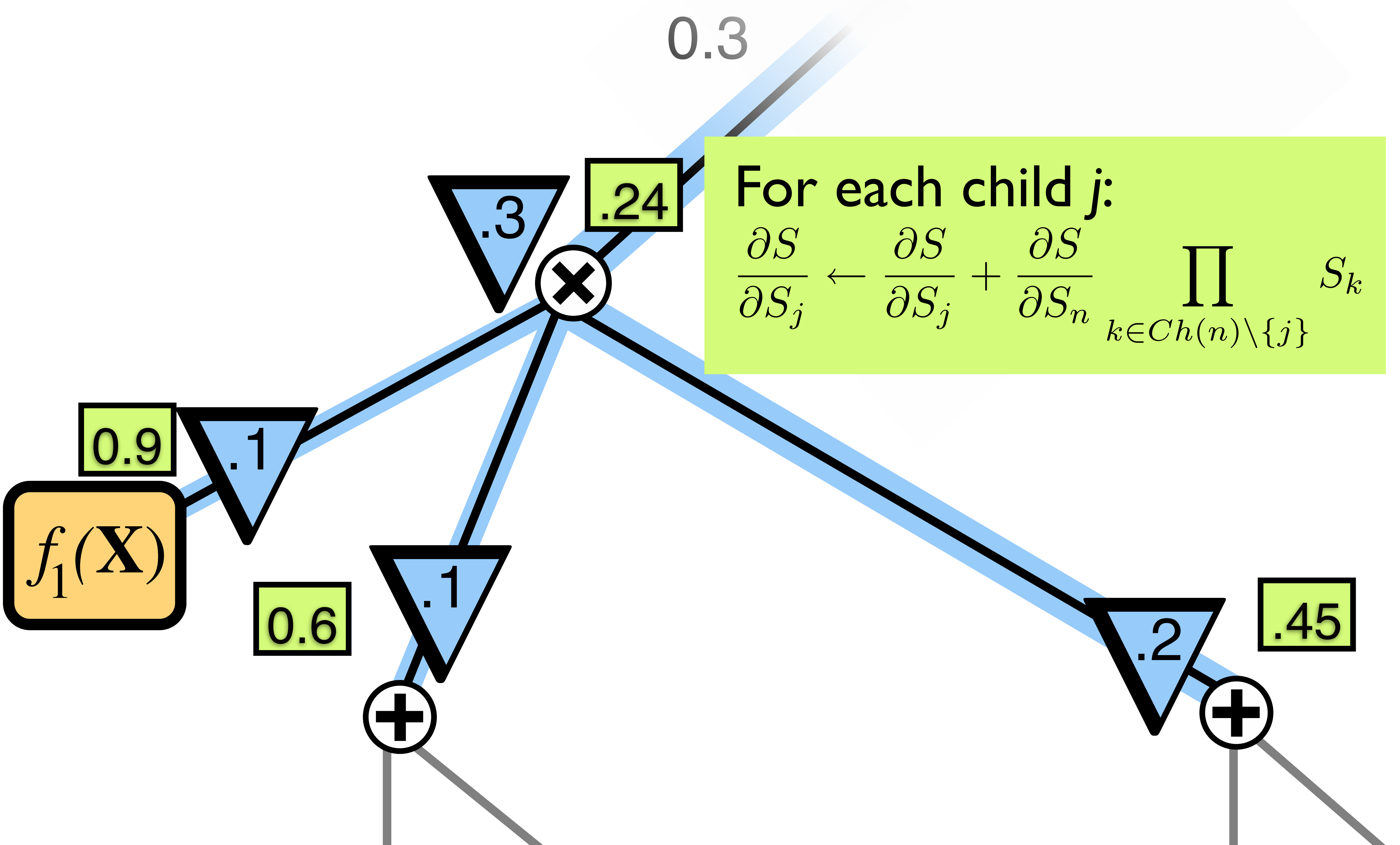
SPN Backpropagation



For each child *j*:

$$\frac{\partial S}{\partial S_j} \leftarrow \frac{\partial S}{\partial S_j} + \frac{\partial S}{\partial S_n} \prod_{k \in Ch(n) \setminus \{j\}} S_k$$

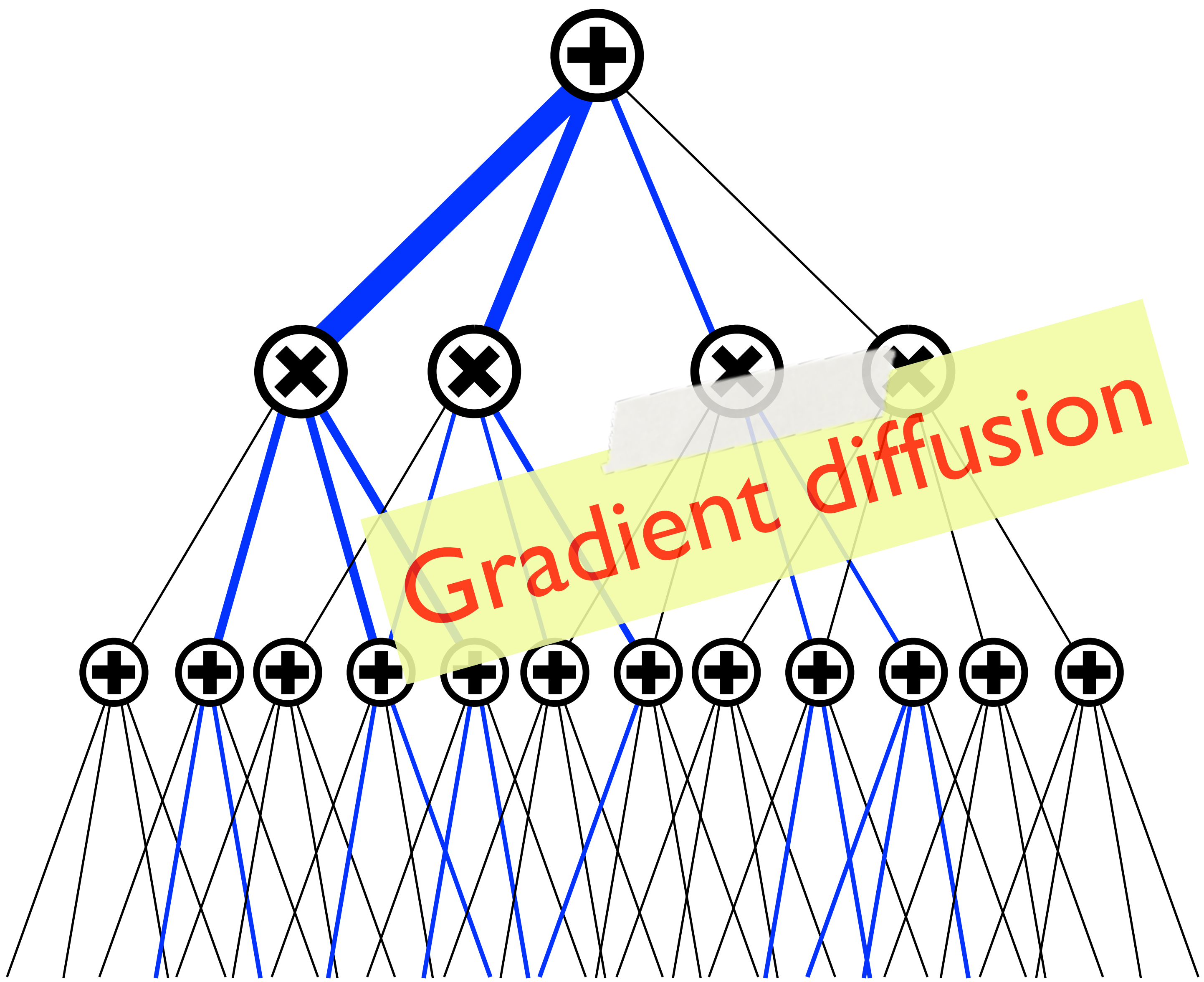
SPN Backpropagation



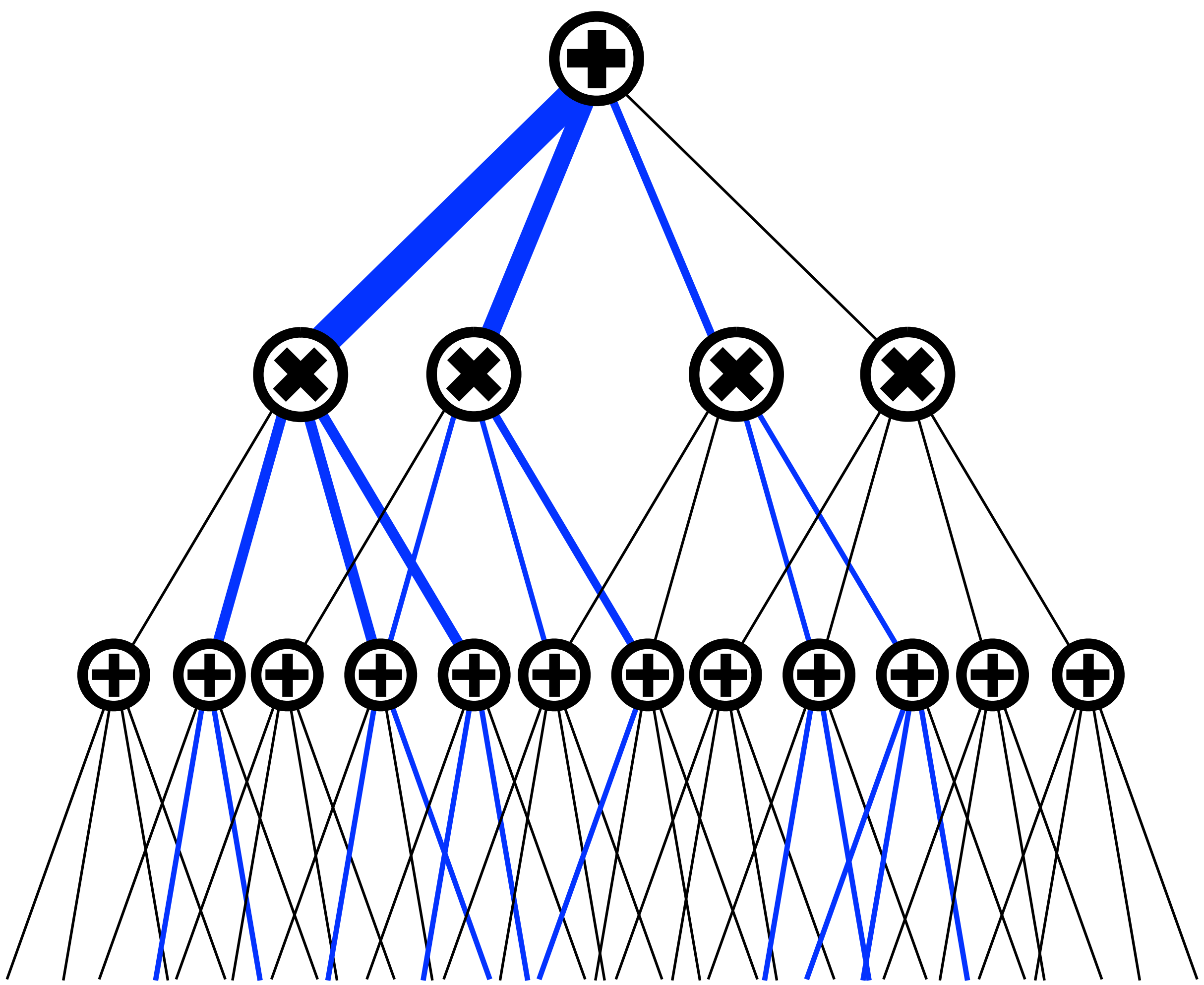
For each child *j*:

$$\frac{\partial S}{\partial S_j} \leftarrow \frac{\partial S}{\partial S_j} + \frac{\partial S}{\partial S_n} \prod_{k \in Ch(n) \setminus \{j\}} S_k$$

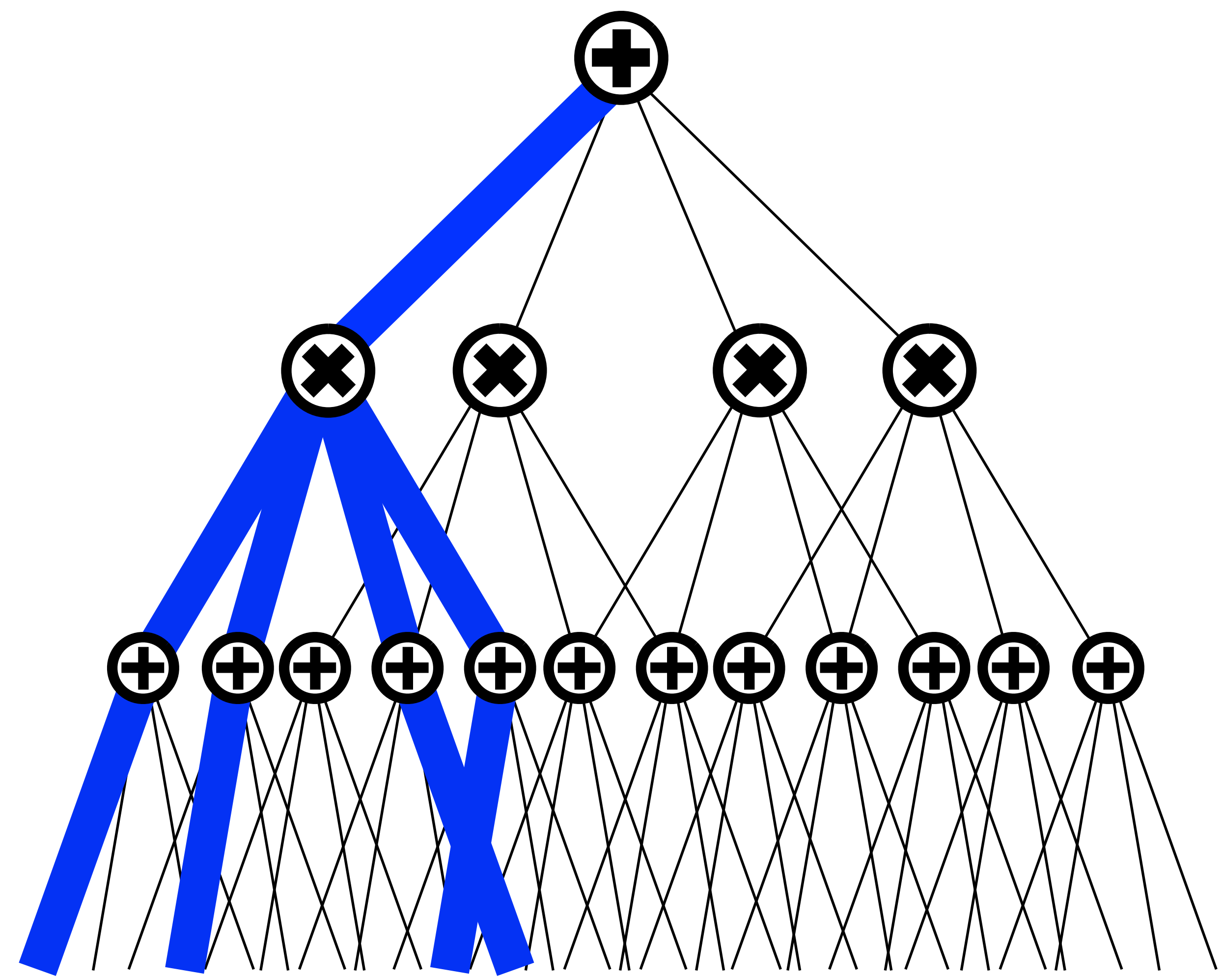
Problem with Backpropagation



Hard Inference Overcomes Gradient Diffusion



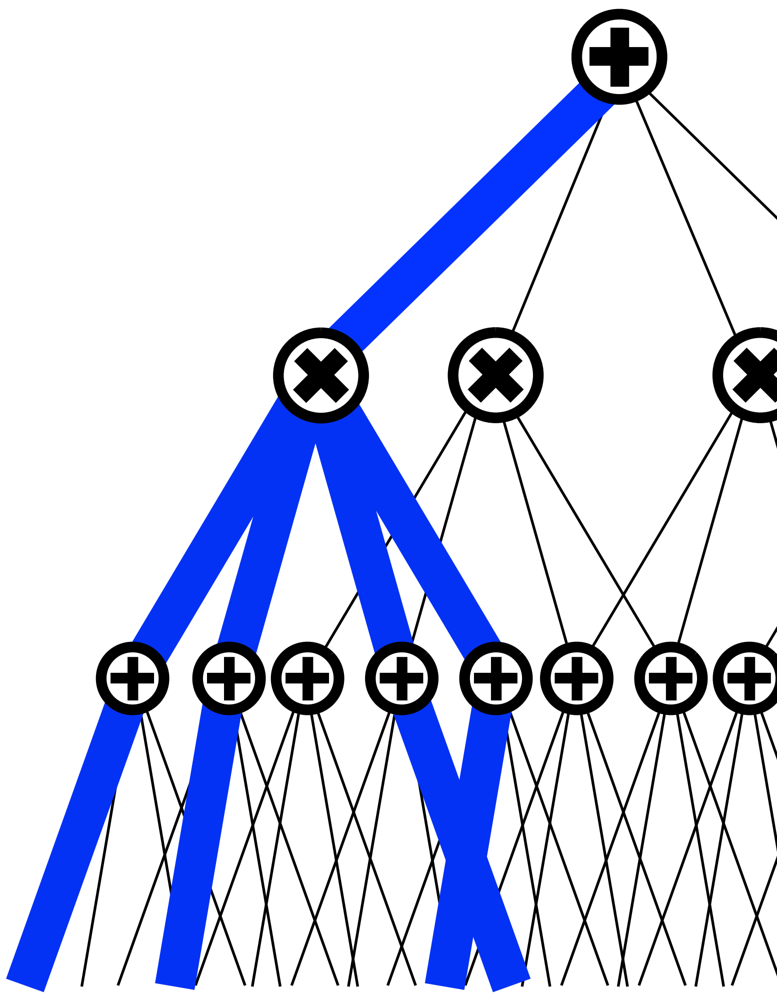
Soft Inference
(Marginals)



Hard Inference
(MAP States)

Reasons to Use Hard Inference

- To overcome gradient diffusion
- When goal is to predict most probable structure
- For speed or tractability



Hard Gradient

$$\nabla \log \tilde{P}(\mathbf{y}|\mathbf{x}) = \nabla \log \frac{\tilde{P}(\mathbf{y},\mathbf{x})}{\tilde{P}(\mathbf{x})} =$$

$$\nabla \log \max_{\mathbf{h}} P(\mathbf{y},\mathbf{h},\mathbf{x}) - \nabla \log \max_{\mathbf{y}',\mathbf{h}} P(\mathbf{y}',\mathbf{h},\mathbf{x})$$

Correct label

Best guess

Hard Gradient

$$\nabla \log \widetilde{P}(\mathbf{y}|\mathbf{x}) = \nabla \log \frac{\widetilde{P}(\mathbf{y}, \mathbf{x})}{\widetilde{P}(\mathbf{x})} =$$

$$\nabla \log \max_{\mathbf{h}} P(\mathbf{y}, \mathbf{h}, \mathbf{x}) \text{ - } \nabla \log \max_{\mathbf{y}', \mathbf{h}} P(\mathbf{y}', \mathbf{h}, \mathbf{x})$$

Hard Gradient

$$\nabla \log \tilde{P}(\mathbf{y}|\mathbf{x}) = \nabla \log \frac{\tilde{P}(\mathbf{y},\mathbf{x})}{\tilde{P}(\mathbf{x})} =$$

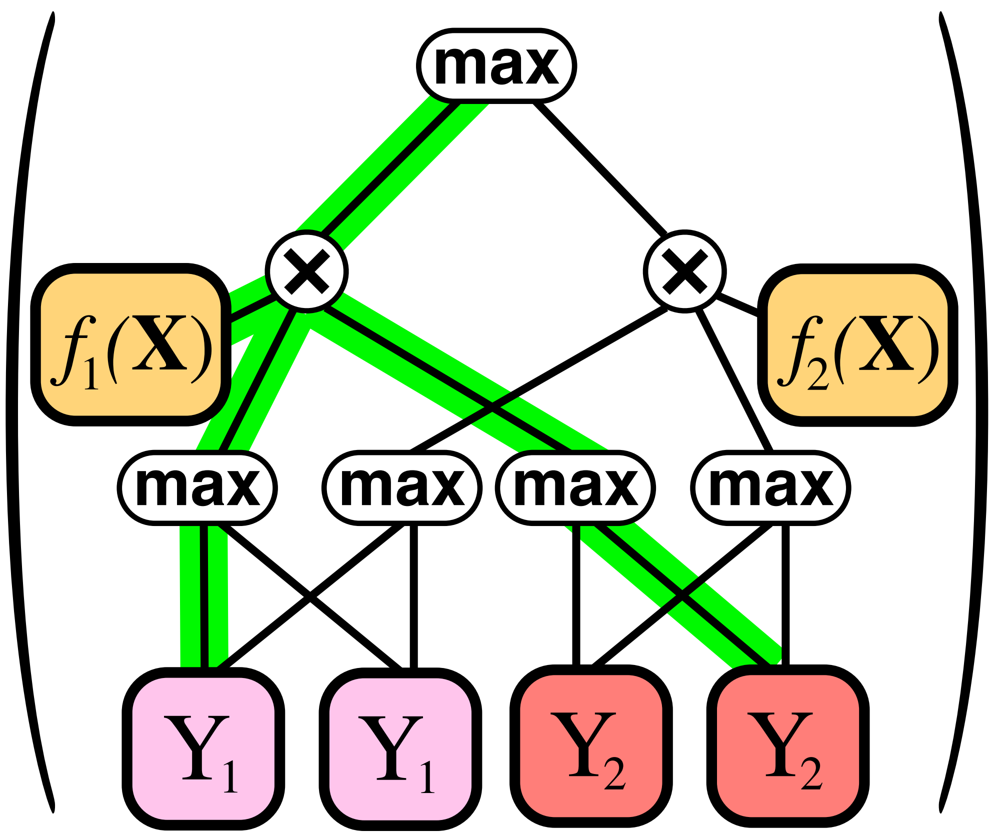
$$\nabla \log \left(\qquad \right) - \nabla \log \left(\qquad \right)$$

$$\max_{\mathbf{h}} P(\mathbf{y},\mathbf{h},\mathbf{x}) \qquad \max_{\mathbf{y}',\mathbf{h}} P(\mathbf{y}',\mathbf{h},\mathbf{x})$$

Hard Gradient

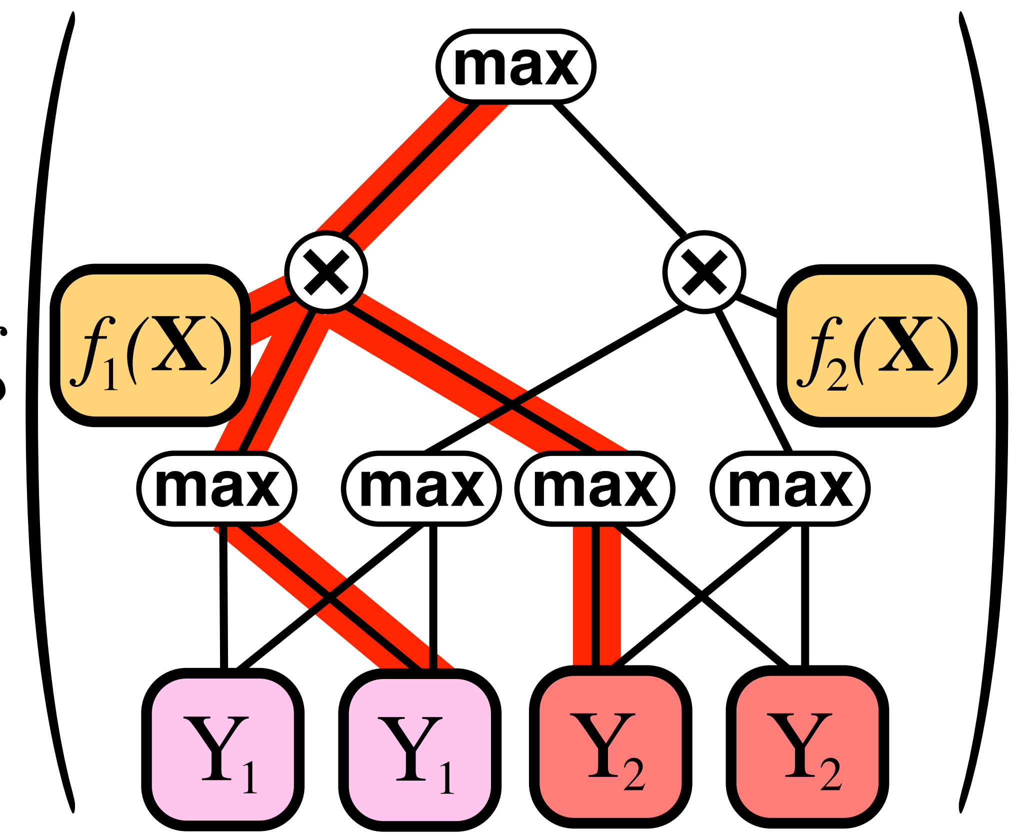
$$\nabla \log \tilde{P}(\mathbf{y}|\mathbf{x}) = \nabla \log \frac{\tilde{P}(\mathbf{y},\mathbf{x})}{\tilde{P}(\mathbf{x})} =$$

$\nabla \log$



$$\max_{\mathbf{h}} P(\mathbf{y},\mathbf{h},\mathbf{x})$$

\- $\nabla \log$

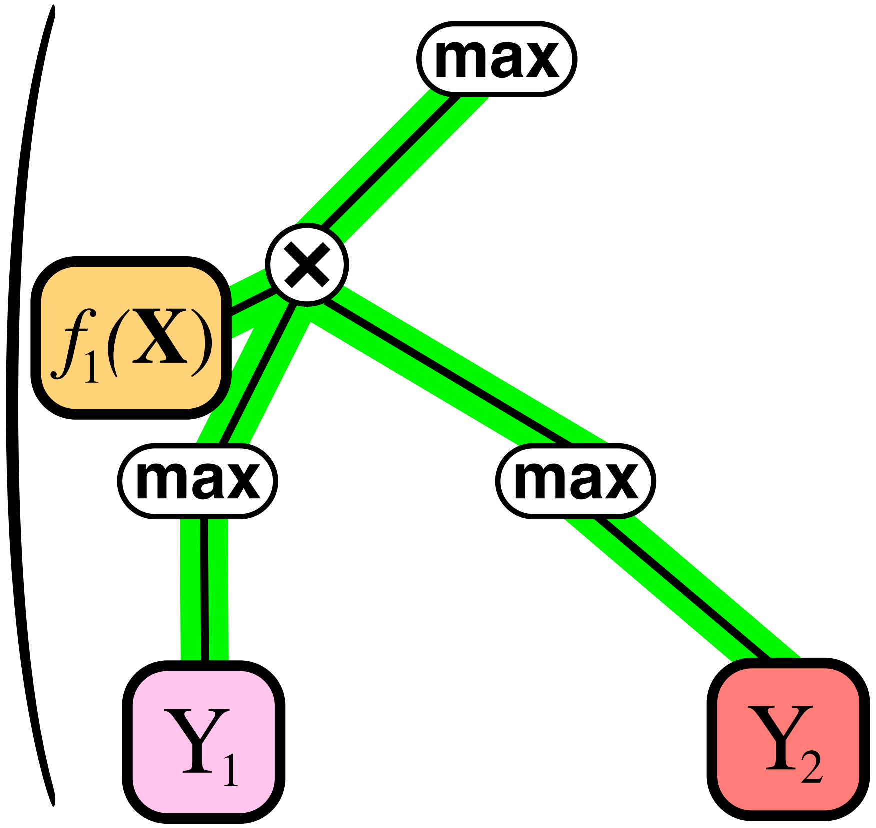


$$\max_{\mathbf{y}',\mathbf{h}} P(\mathbf{y}',\mathbf{h},\mathbf{x})$$

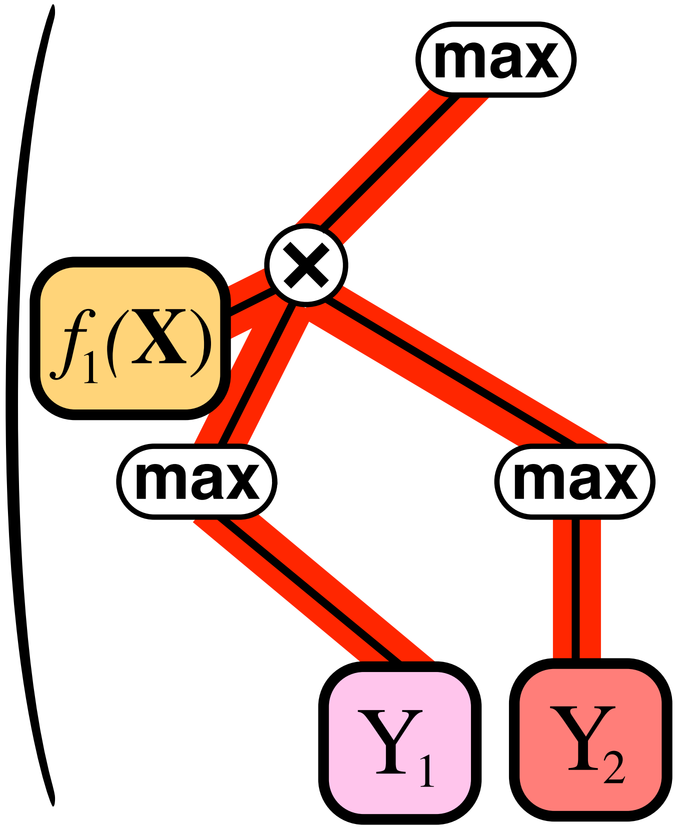
Hard Gradient

$$\nabla \log \tilde{P}(\mathbf{y}|\mathbf{x}) = \nabla \log \frac{\tilde{P}(\mathbf{y},\mathbf{x})}{\tilde{P}(\mathbf{x})} =$$

$\nabla \log$ (



) - $\nabla \log$ (



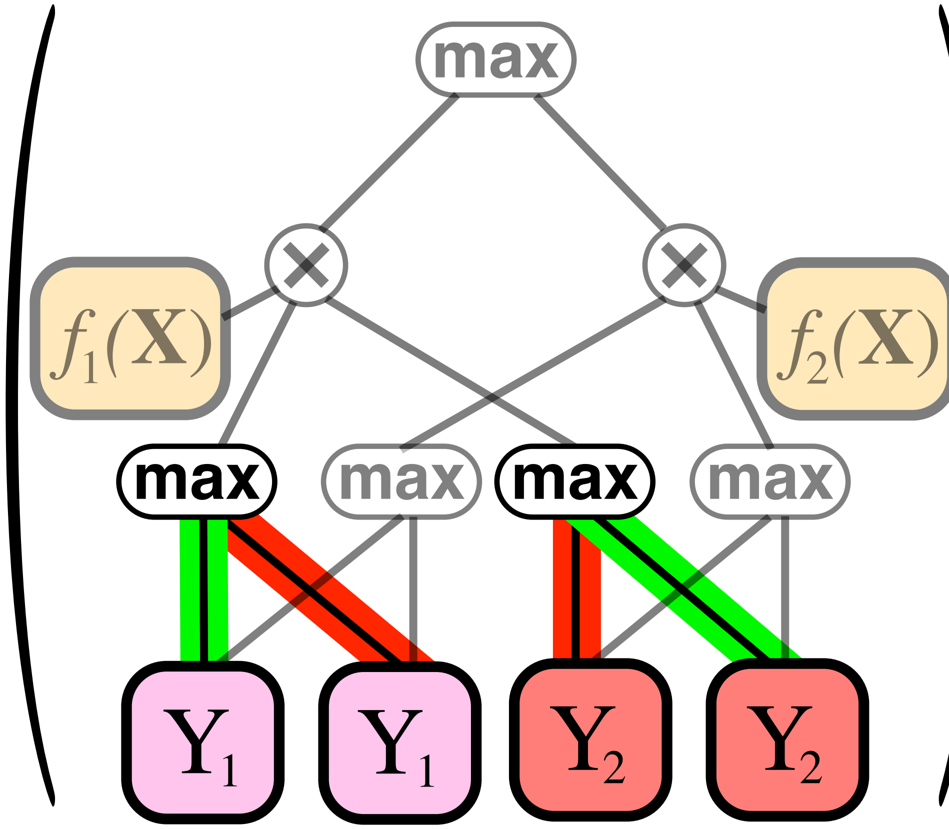
)

$$\max_{\mathbf{h}} P(\mathbf{y},\mathbf{h},\mathbf{x}) \qquad \max_{\mathbf{y}',\mathbf{h}} P(\mathbf{y}',\mathbf{h},\mathbf{x})$$

Hard Gradient

$$\nabla \log \tilde{P}(\mathbf{y}|\mathbf{x}) = \nabla \log \frac{\tilde{P}(\mathbf{y},\mathbf{x})}{\tilde{P}(\mathbf{x})} =$$

$\nabla \log$ (

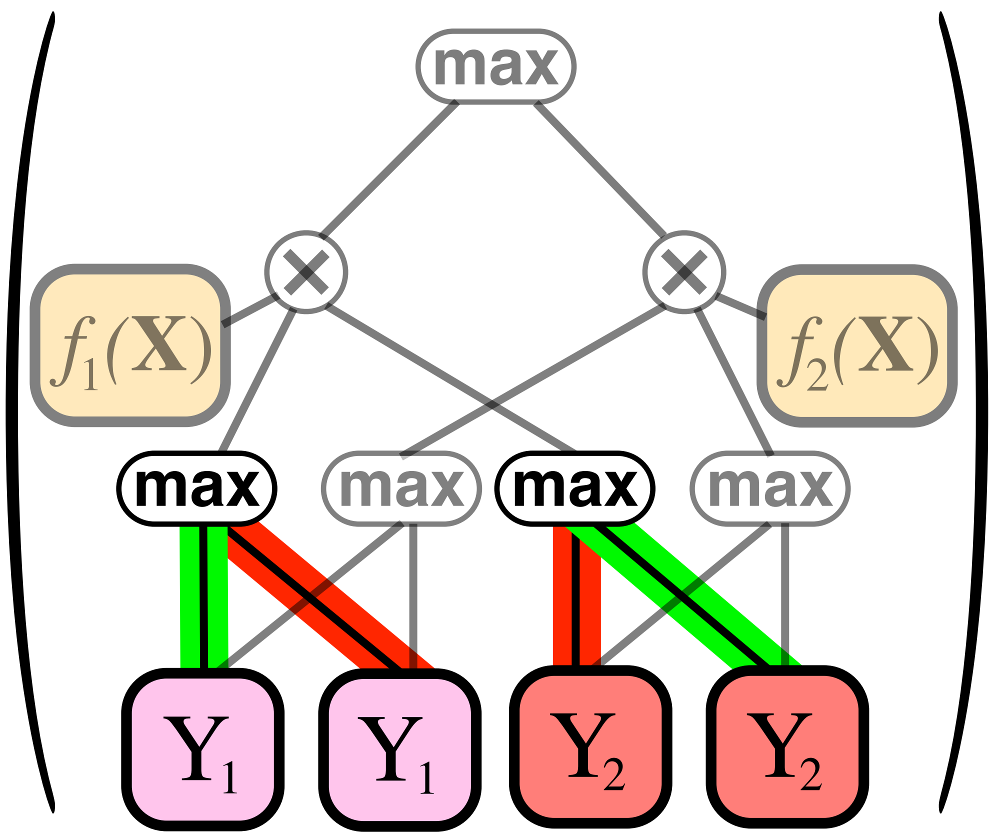


)

Hard Gradient

$$\nabla \log \tilde{P}(\mathbf{y}|\mathbf{x}) = \nabla \log \frac{\tilde{P}(\mathbf{y},\mathbf{x})}{\tilde{P}(\mathbf{x})} =$$

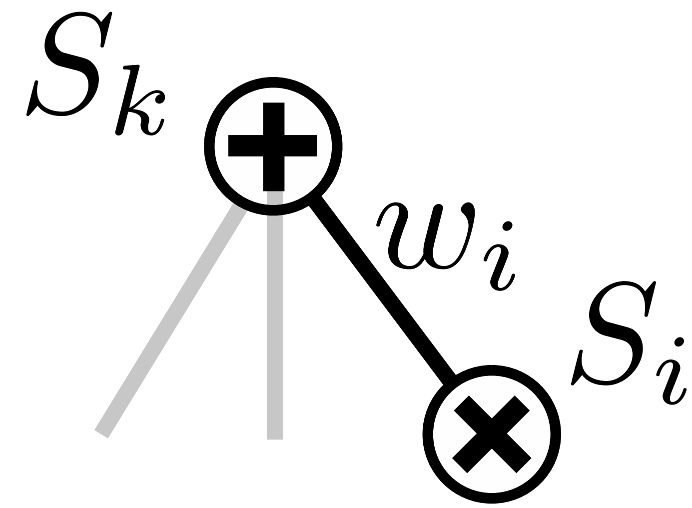
$\nabla \log$



w/ correct label _ # w/ model guess

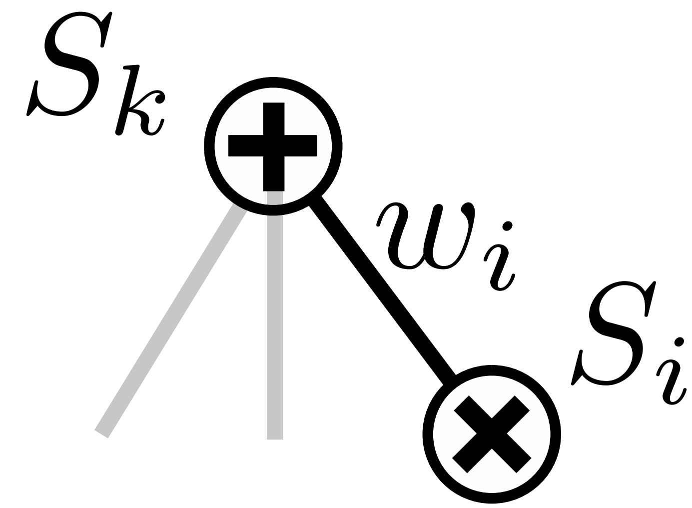
$$\frac{\partial}{\partial w_i} \log \tilde{P}(\mathbf{y}|\mathbf{x}) = \frac{\Delta c_i}{w_i}$$

Learning SPNs: Summary



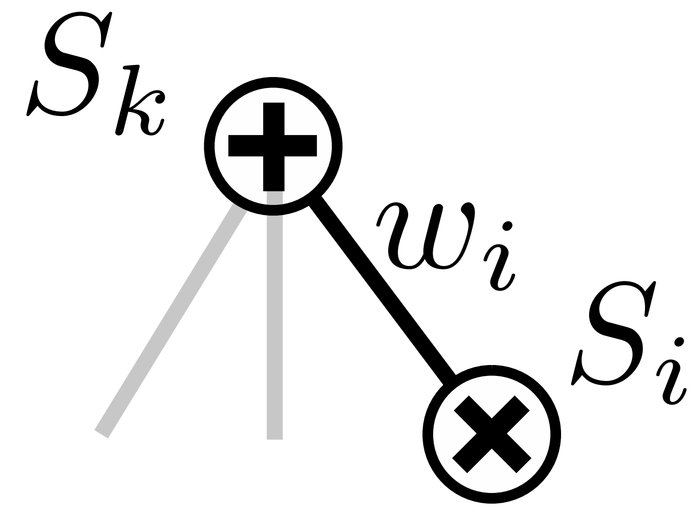
Update	**Soft Inference** (Marginals)	**Hard Inference** (MAP States)
Gen. EM	$\Delta w_i \propto w_i \frac{\partial S}{\partial S_k}$	$\Delta w_i = c_i$
Gen. Gradient	$\Delta w_i = \eta \frac{\partial S}{\partial S_k} S_i$	$\Delta w_i = \eta \frac{c_i}{w_i}$
Disc. Gradient	$\Delta w_i = \eta(\overbrace{\frac{S_i}{S}\frac{\partial S}{\partial S_k}}^{\text{true label}} - \overbrace{\frac{S_i}{S}\frac{\partial S}{\partial S_k}}^{\text{exp. label}})$	$\Delta w_i = \frac{\eta}{w_i}(\overbrace{c_i}^{\text{true}} - \overbrace{c_i}^{\text{test}})$

Learning SPNs: Summary



Update	**Soft Inference** (Marginals)	**Hard Inference** (MAP States)
Gen. EM	$\Delta w_i \propto w_i \frac{\partial S}{\partial S_k}$	$\Delta w_i = c_i$
Gen. Gradient	$\Delta w_i = \eta \frac{\partial S}{\partial S_k} S_i$	$\Delta w_i = \eta \frac{c_i}{w_i}$
Disc. Gradient	$\Delta w_i = \eta (\overbrace{\frac{S_i}{S} \frac{\partial S}{\partial S_k}}^{\text{true label}} - \overbrace{\frac{S_i}{S} \frac{\partial S}{\partial S_k}}^{\text{exp. label}})$	$\Delta w_i = \frac{\eta}{w_i} (\overbrace{c_i}^{\text{true}} - \overbrace{c_i}^{\text{test}})$

Learning SPNs: Summary

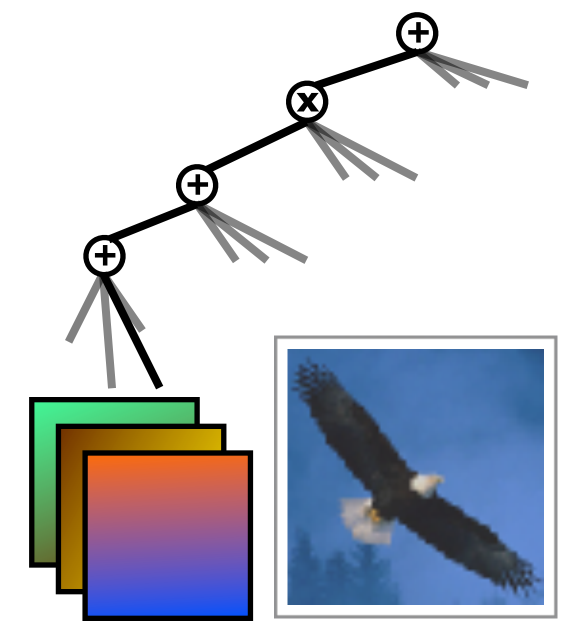


Update	**Soft Inference** (Marginals)	**Hard Inference** (MAP States)
Gen. EM	$\Delta w_i \propto w_i \frac{\partial S}{\partial S_k}$	$\Delta w_i = c_i$
Gen. Gradient	$\Delta w_i = \eta \frac{\partial S}{\partial S_k} S_i$	$\Delta w_i = \eta \frac{c_i}{w_i}$
Disc. Gradient	$\Delta w_i = \eta(\overbrace{\frac{S_i}{S}\frac{\partial S}{\partial S_k}}^{\text{true label}} - \overbrace{\frac{S_i}{S}\frac{\partial S}{\partial S_k}}^{\text{exp. label}})$	$\Delta w_i = \frac{\eta}{w_i}(\overbrace{c_i}^{\text{true}} - \overbrace{c_i}^{\text{test}})$

Motivation

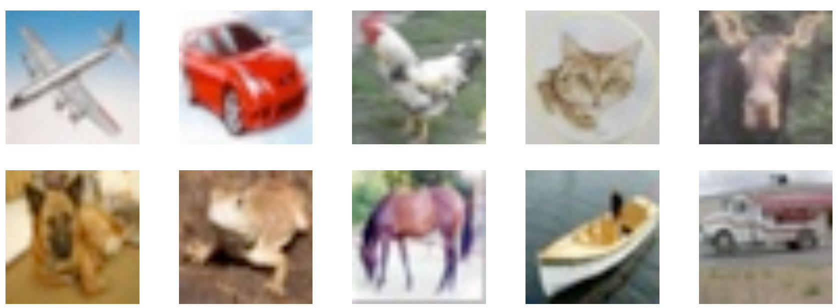
SPN Review

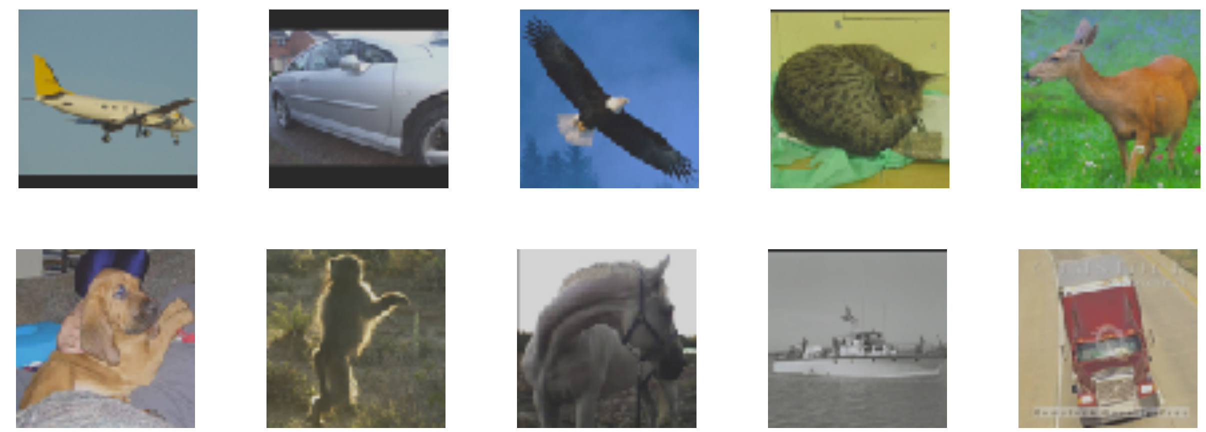
Discriminative Training



Experiments

Image Classification





CIFAR-10

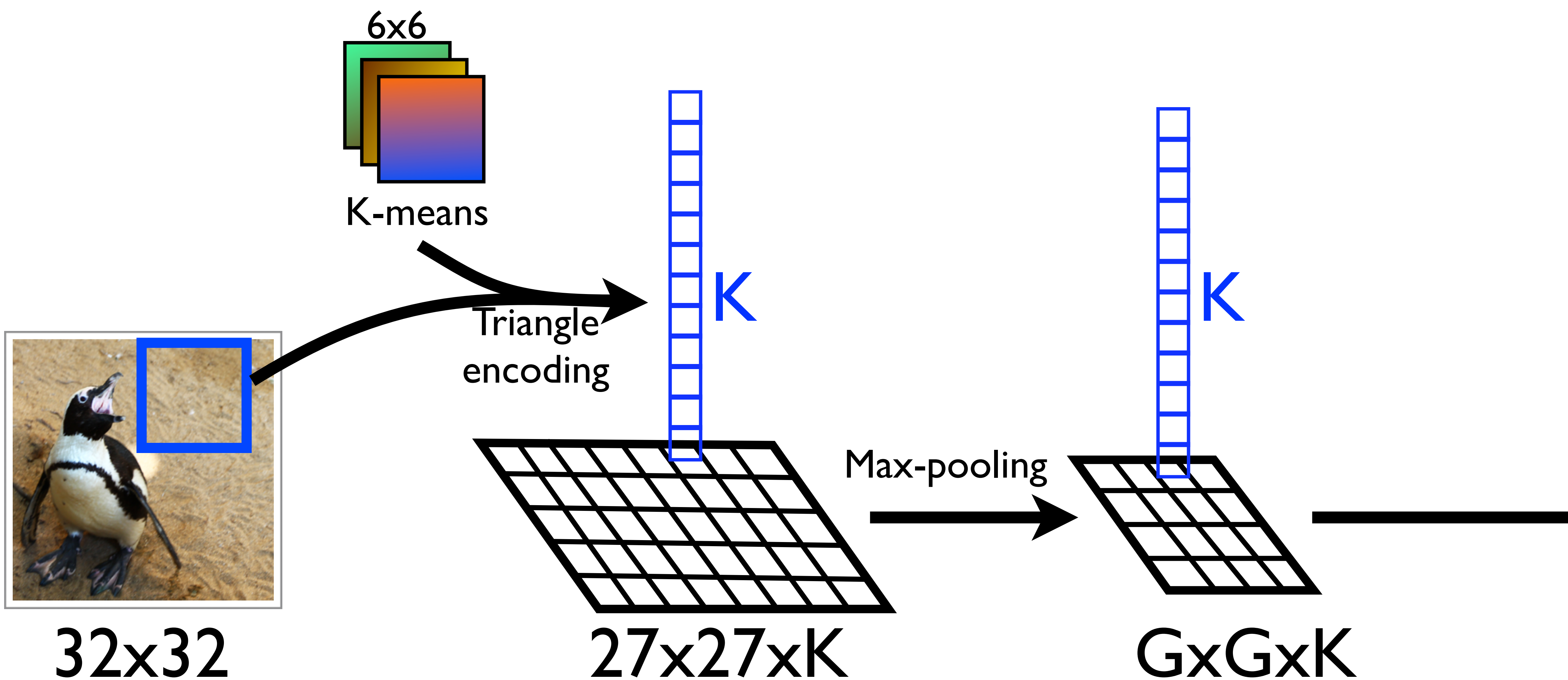
32x32px
50k train
10k test

STL-10

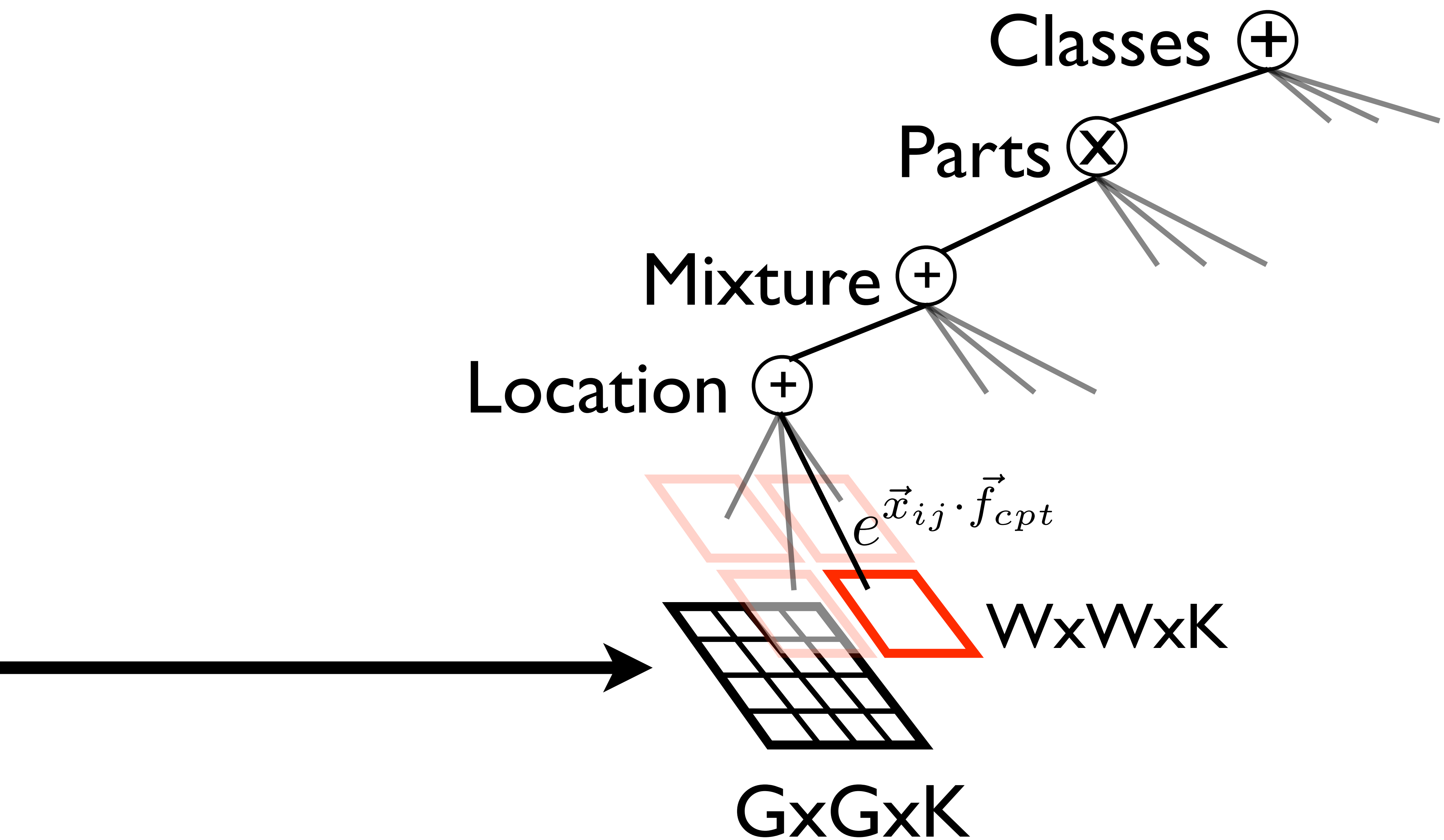
96x96px
5k train } 10 folds
8k test }
~~100k unlabeled~~

Feature Extraction

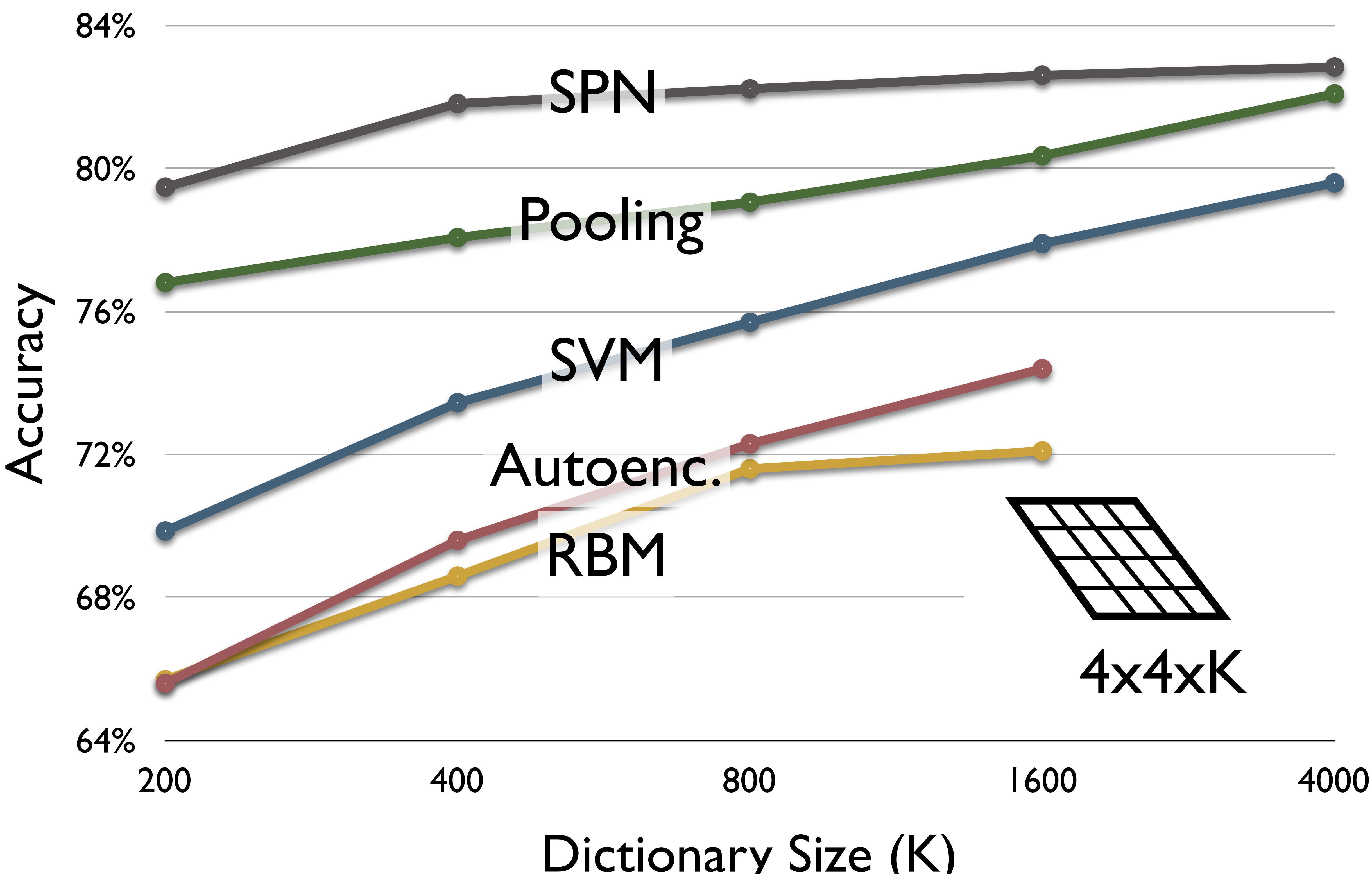
Coates et al., AISTATS 2011



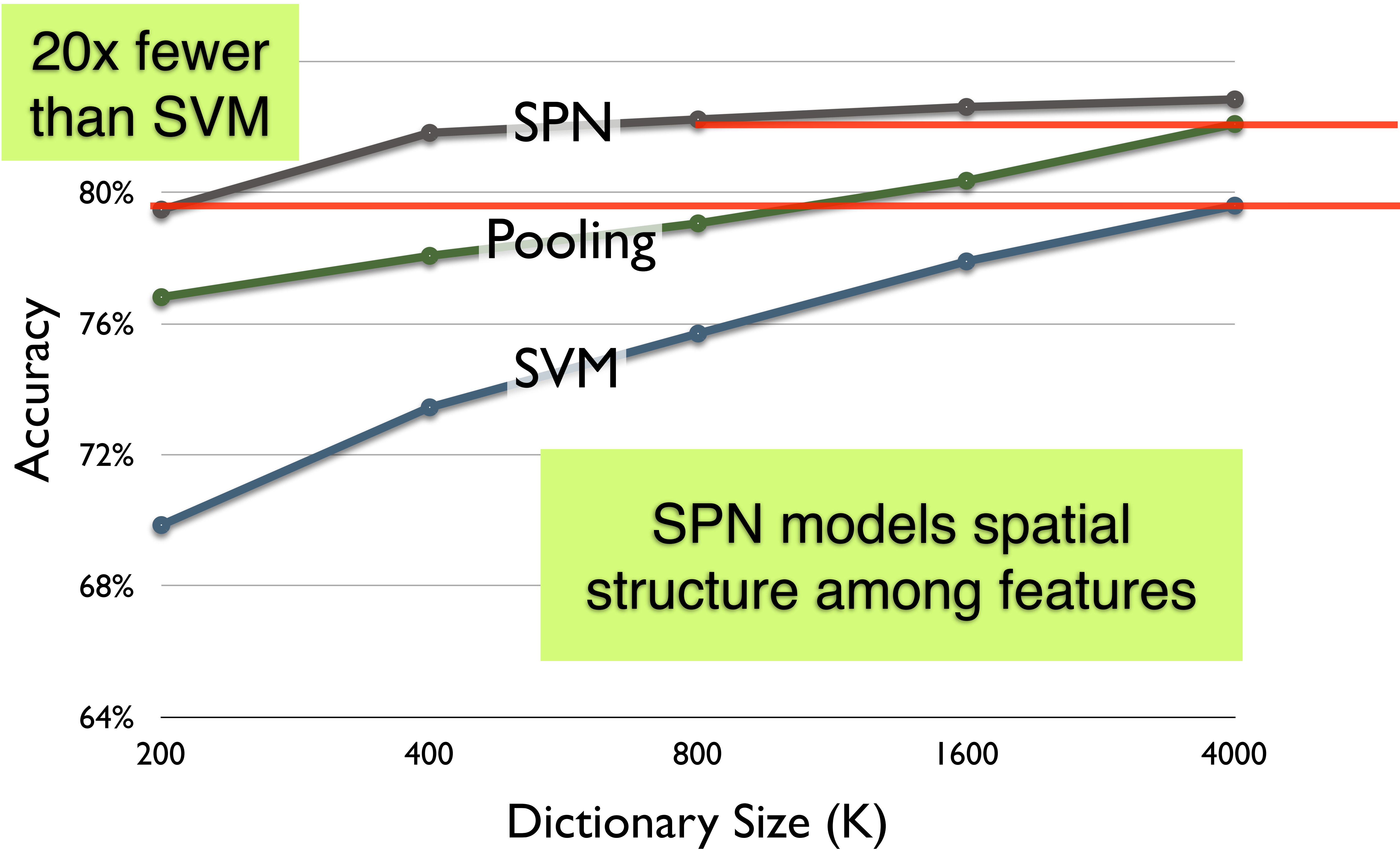
SPN Architecture



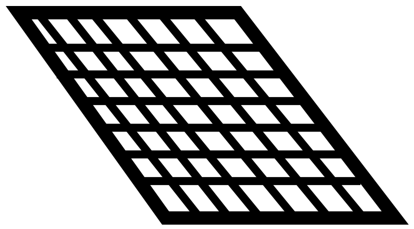
CIFAR-10 Results



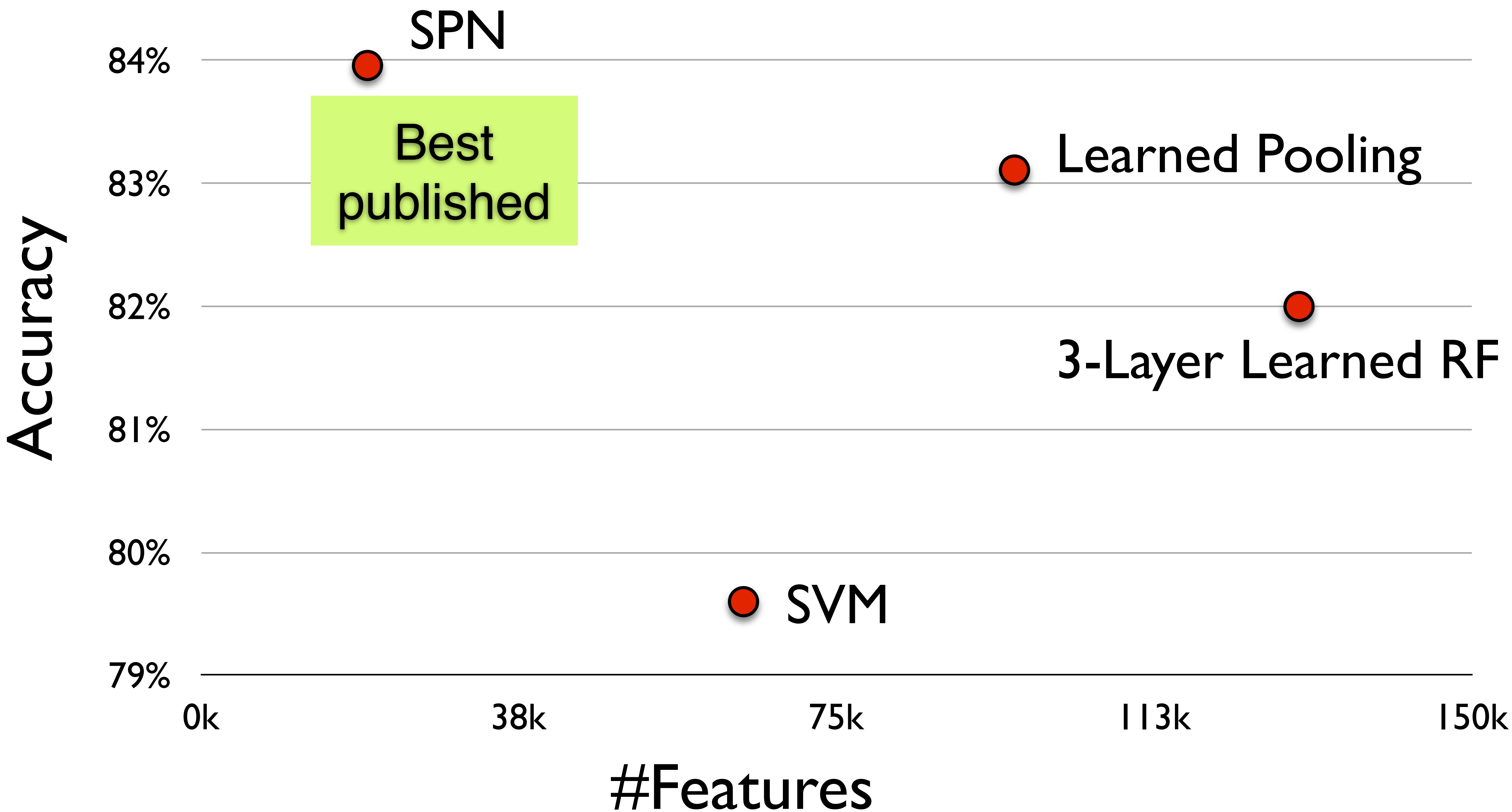
CIFAR-10 Results



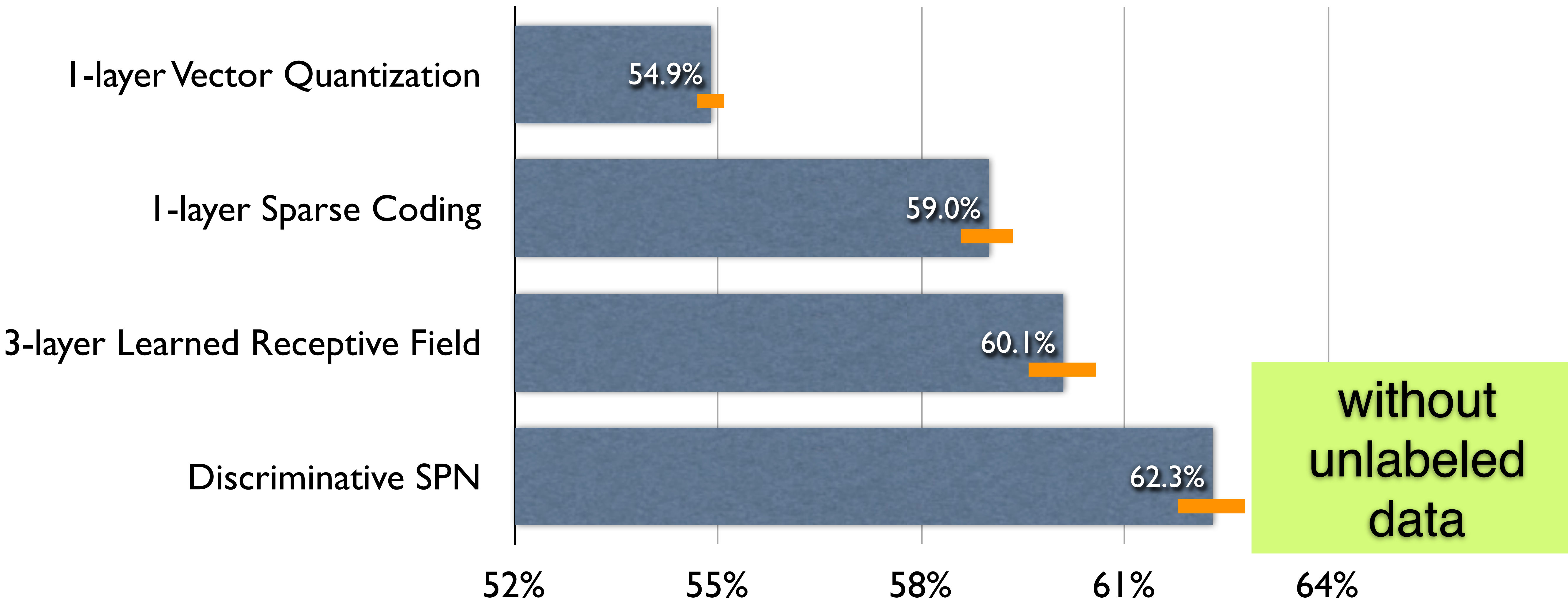
CIFAR-10 Results



7x7x400



STL-10 results



Future Work

- Max-margin SPNs
- Learning SPN structure
- Applying discriminative SPNs to structured prediction
- Approximate inference using SPNs

Poster W15

Summary

- Discriminative SPNs combine the advantages of
  - Tractable inference
  - Deep architectures
  - Discriminative learning
- Hard gradient combats diffusion in deep models
- Discriminative SPNs outperform SVMs and deep models on image classification benchmarks