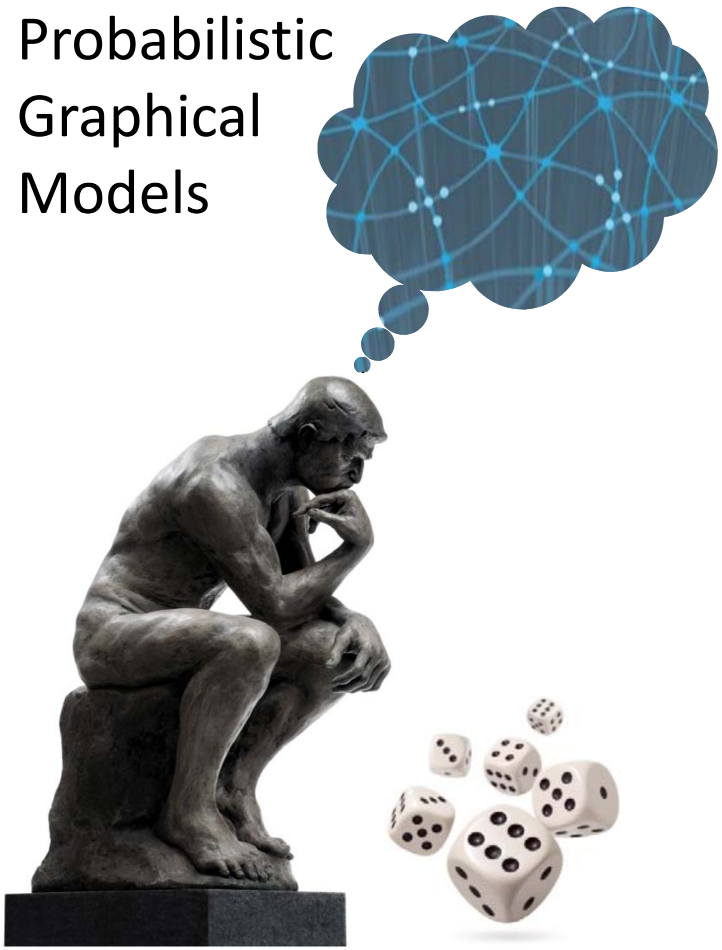
Probabilistic
Graphical
Models



Learning

Parameter Estimation

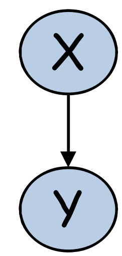
Max-Likelihood for BNs

MLE for Bayesian Networks

- Parameters: $\theta_{x^0}, \theta_{x^1}$
  $\theta_{y^0|x^0}, \theta_{y^1|x^0}, \theta_{y^0|x^1}, \theta_{y^1|x^1}$
- Data instances: <x[m],y[m]>

X	
x^0	x^1
0.7	0.3

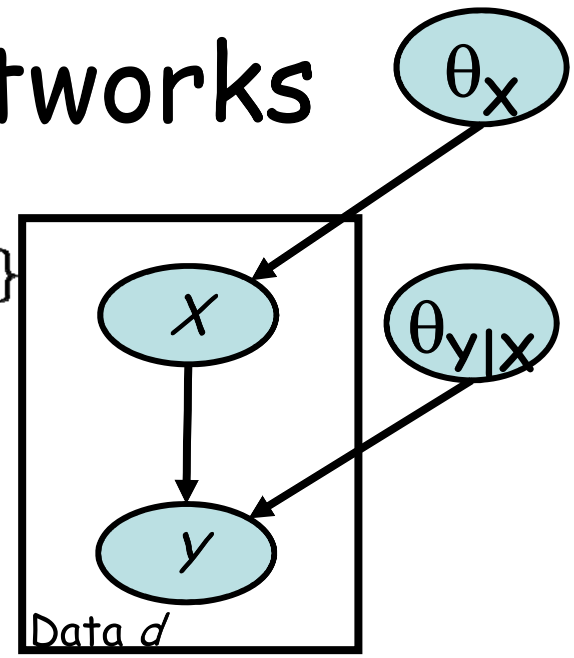
P(X)



X	*Y*	
	y^0	y^1
x^0	0.95	0.05
x^1	0.2	0.8

P(Y|X)

MLE for Bayesian Networks



- Parameters: $\{\theta_x : x \in Val(X)\}$
  $\{\theta_{y|x} : x \in Val(X), y \in Val(Y)\}$

$$L(\Theta : D) = \prod_{m=1}^{M} P(x[m], y[m] : \theta)$$

chain rule for BNs

$$= \prod_{m=1}^{M} P(x[m] : \theta) P(y[m] \mid x[m] : \theta)$$

$$= \left(\prod_{m=1}^{M} P(x[m] : \theta)\right)\left(\prod_{m=1}^{M} P(y[m] \mid x[m] : \theta)\right)$$

product of two local likelihood

$$= \left(\prod_{m=1}^{M} P(x[m] : \theta_X)\right)\left(\prod_{m=1}^{M} P(y[m] \mid x[m] : \theta_{Y|X})\right)$$

Daphne Koller

MLE for Bayesian Networks

- Likelihood for Bayesian network

parents of X_i

chain rule

$$
\begin{aligned}
L(\Theta : D) &= \prod_m P(x[m] : \Theta) \\
&= \prod_m \prod_i P(x_i[m] \mid \mathrm{U}_i[m] : \Theta_i) \\
&= \prod_i \underbrace{\prod_m P(x_i[m] \mid \mathrm{U}_i[m] : \Theta_i)}_{L_i(\theta_{X_i|U_i} : D)} \\
&= \prod_i L_i(\Theta_i : D)
\end{aligned}
$$

local likelihood

⇨ if $\theta_{X_i|U_i}$ are disjoint, then MLE can be computed by maximizing each local likelihood separately

Daphne Koller

MLE for Table CPDs

$$\prod_{m=1}^{M} P(x[m] \mid \mathbf{u}[m] : \theta) = \prod_{m=1}^{M} P(x[m] \mid \mathbf{u}[m] : \underline{\theta_{X|\mathbf{U}}})$$

$$= \prod_{x,\mathbf{u}} \left(\prod_{m : x[m]=x,\, \mathbf{u}[m]=\mathbf{u}} P(x[m] \mid \mathbf{u}[m] : \theta_{X|\mathbf{U}}) \right)$$

$P(X[m]=x \mid u[m]=u : \theta_{x|u}) = \theta_{x|u}$

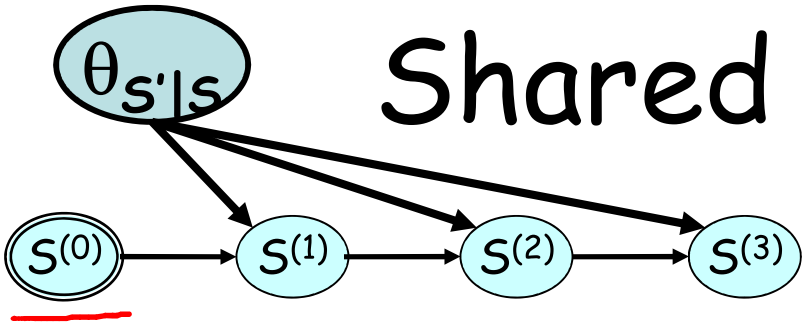
$$= \prod_{x,\mathbf{u}} \left(\prod_{m : x[m]=x,\, \mathbf{u}[m]=\mathbf{u}} \theta_{x|\mathbf{u}} \right)$$

$$= \prod_{x,\mathbf{u}} \underbrace{\theta_{x|\mathbf{u}}}_{P(x|u)}{}^{M[x,\mathbf{u}]}$$

fraction of $X=x$ among cases where $\bar{U}=\bar{u}$

$$\theta_{x|\mathbf{u}} = \frac{M[x,\mathbf{u}]}{\sum_{x'} M[x',\mathbf{u}]} = \frac{M[x,\mathbf{u}]}{M[\mathbf{u}]}$$

Shared Parameters



$$L(\theta : S^{(0:T)}) = \prod_{t=1}^{T} P(S^{(t)} \mid S^{(t-1)} : \theta)$$

$$= \prod_{i,j} \prod_{t:\underbrace{S^{(t)}=s^i, S^{(t+1)}=s^j}_{s^i \to s^j}} P(S^{(t+1)} \mid S^{(t)} : \theta_{S'|S})$$

$$= \prod_{i,j} \prod_{t:S^{(t)}=s^i, S^{(t+1)}=s^j} \theta_{s^i \to s^j}$$

$$= \prod_{i,j} \theta_{s^i \to s^j}^{M[s^i \to s^j]}$$

$$\hat{\theta}_{s^i \to s^j} = \frac{M[s^i \to s^j]}{M[s^i]}$$

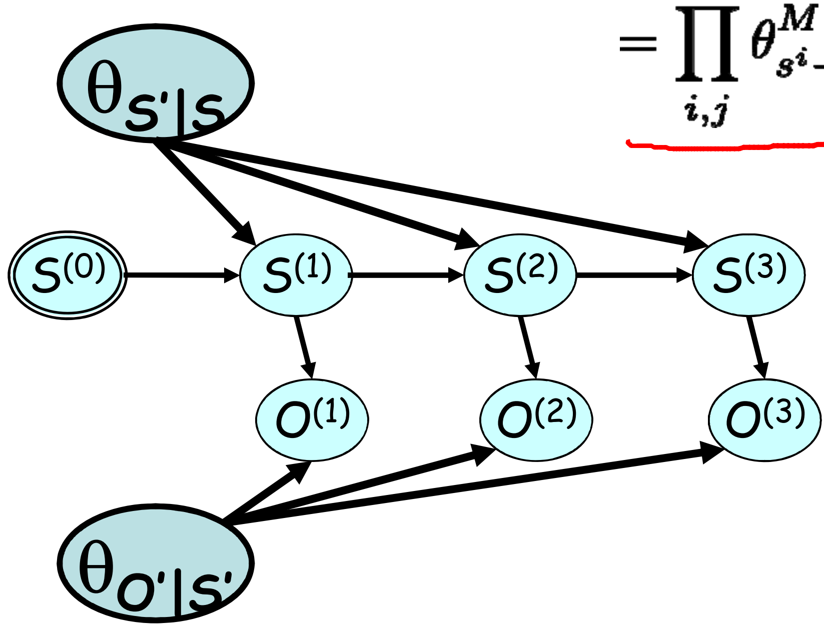
$$M[s^i \to s^j] = |\{t \ : \ S^{(t)} = s^i, S^{(t+1)} = s^j\}|$$

Daphne Koller

Shared Parameters

$$L(\Theta : S^{(0:T)}, O^{(0:T)}) = \underbrace{\prod_{t=1}^{T} P(S^{(t)} \mid S^{(t-1)} : \theta_{S'|S})}\ \underbrace{\prod_{t=1}^{T} P(O^{(t)} \mid S^{(t)} : \theta_{O'|S'})}$$

$$= \underbrace{\prod_{i,j} \theta_{s^i \to s^j}^{M[s^i \to s^j]}}\ \prod_{i,k} \theta_{o^k|s^i}^{M[o^k, s^i]}$$



$$M[s^i \to s^j] = |\{t \ : \ S^{(t)} = s^i, S^{(t+1)} = s^j\}|$$

$$\underline{M[o^k, s^i]} = |\{t \ : \ S^{(t)} = \underline{s^i}, O^{(t)} = \underline{o^k}\}|$$

Daphne Koller

Summary

- For BN with disjoint sets of parameters in CPDs, likelihood decomposes as product of local likelihood functions, one per variable
- For table CPDs, local likelihood further decomposes as product of likelihood for multinomials, one for each parent combination
- For networks with shared CPDs, sufficient statistics accumulate over all uses of CPD

Fragmentation & Overfitting

$$\theta_{x|\mathbf{u}} = \frac{M[x,\mathbf{u}]}{\sum_{x'} M[x',\mathbf{u}]} = \frac{M[x,\mathbf{u}]}{M[\mathbf{u}]}$$

- # of "buckets" increases exponentially with |U|
- For large |U|, most "buckets" will have very few instances u
  - ⇨ **very poor parameter estimates**
- **With limited data, we often get better generalization with simpler structures**

even when wrong