Review

- Learning in BNs
- ML estimation from complete data

examples $\{(x_1^{(t)}, x_2^{(t)}, \ldots, x_n^{(t)})\}_{t=1}^{T}$

$$P_{ML}(X_i = x \mid pa_i = \pi) = \frac{count(X_i = x,\ pa_i = \pi)}{count(pa_i = \pi)}$$

$$= \frac{\sum_{t=1}^{T} I(x_i^{(t)}, x)\, I(pa_i^{(t)}, \pi)}{\sum_{t=1}^{T} I(pa_i^{(t)}, \pi)}$$

I: indicator function tests for equality

- ML estimation from incomplete data

examples $t = 1, 2, \ldots, T$
Hidden nodes $H^{(t)}$
Visible nodes $V^{(t)}$

t	X_1	X_2	...	X_n
1	1	2	5	3
2	2	?	0	3
3	?	1	1	8
⋮	?	?	5	5
T	2	6	3	9

? : hidden values

Choose CPTs to maximize log-likelihood

$$\mathcal{L} = \sum_{t=1}^{T} \log P(V = v^{(t)})$$

$$= \sum_{t} \log \sum_{h} P(V = v^{(t)}, H^{(t)} = h)$$

$$= \sum_{t} \log \sum_{h} \prod_{i=1}^{n} P(X_i = x \mid pa_i = \pi)\Big|_{V = v^{(t)},\ H = h}$$

How to maximize?

- **Expectation-Maximization (EM) algorithm**

  Iterative procedure to maximize $\sum_t \log P(V = v^{(t)})$ for incomplete data in terms of CPTs of BN.

- **Algorithm:**

  (1) Initialize all CPTs with (possibly) random values

  (2) Do until log-likelihood stops increasing:

  E-step: compute posterior probabilities (inference)

  $$P(X_i = x, pa_i = \pi \mid V = v^{(t)}) \quad \text{for } t = 1, 2, \ldots, T$$

  M-step: update CPTs

  $$P(X_i = x \mid pa_i = \pi) \leftarrow \frac{\sum_t P(X_i = x, pa_i = \pi \mid v^{(t)})}{\sum_t P(pa_i = \pi \mid v^{(t)})}$$

Intuition

- expected statistics under $P(H \mid v^{(t)})$ are filling in "missing values" of incomplete data
- expected counts are substituting for observed counts in complete data case

Iterate E + M steps until convergence. Why iterate?

RHS depends on current CPTs.

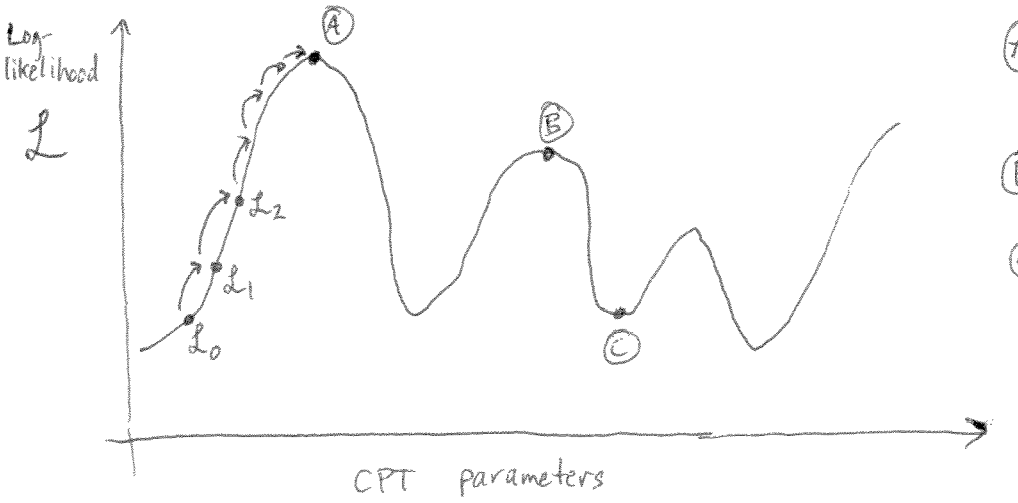
Key properties

- monotonic convergence

  Each iteration of EM improves log-likelihood $\ell = \sum_t \log P(v^{(t)})$

  If ℓ_k is log-likelihood at k^{th} iteration, then $\ell_k \geq \ell_{k-1}$.

- converges to stationary point of log-likelihood where gradient vanishes



(A) global max: most desirable, not guaranteed

(B) local max: usual outcome

(C) local min: possible in theory, but never in practice

- no tuning parameters: no step sizes, learning rates, backtracking

Example  A, C are observed (visible), B is hidden

Incomplete data set $\{(a_t, c_t)\}_{t=1}^{T}$

t	A	B	C
1	a_1	?	c_1
2	a_2	?	c_2
⋮	⋮	⋮	⋮
T	a_T	?	c_T

Log-likelihood:

$$\ell = \sum_t \log P(A=a_t, C=c_t)$$

$$= \sum_t \log \sum_b P(A=a_t, B=b, C=c_t) \quad \text{marginalization}$$

$$= \sum_t \log \left\{ \sum_b P(a_t) P(b|a_t) P(c_t|b) \right\} \quad \text{product rule, CI, shorthand}$$

General EM algorithm:

M-step $P(X_i=x \mid pa_i=\pi) \leftarrow$ E-step $\overline{P(X_i=x, pa_i=\pi \mid v^{(t)})}$ / $\sum_t P(pa_i=\pi \mid v^{(t)})$

Now apply to this example...

- First: posterior probability $P(H|V)$

Shorthand: $P(B=b \mid A=a_t, C=c_t) \rightarrow P(b \mid a_t, c_t)$

$$P(b \mid a_t, c_t) = \frac{P(c_t \mid a_t, b) P(b \mid a_t)}{P(c_t \mid a_t)}$$ Bayes rule

$$= \frac{P(c_t \mid b) P(b \mid a_t)}{\sum_{b'} P(b', c_t \mid a_t)}$$ CI + marginalization

$$= \frac{P(c_t \mid b) P(b \mid a_t)}{\sum_{b'} P(c_t \mid b') P(b' \mid a_t)}$$ product rule + CI

- Update CPTs $P(A)$, $P(B|A)$, $P(C|B)$

$$P(A=a) = \frac{\sum_t \text{count}(A=a)}{T}$$

M-step:

$$P(B=b \mid A=a) \leftarrow \frac{\sum_t P(A=a, B=b \mid A=a_t, C=c_t)}{\sum_t P(A=a \mid A=a_t, C=c_t)}$$

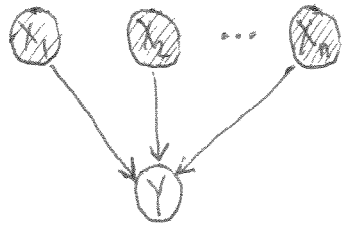
Simplify: $$P(b|a) \leftarrow \frac{\sum_t I(a, a_t) P(b \mid a_t, c_t)}{\sum_t I(a, a_t)}$$ ← posterior computed above, in terms of CPTs

$$P(C=c \mid B=b) \leftarrow \frac{\sum_{t=1}^{I} P(B=b, C=c \mid A=a_t, C=c_t)}{\sum_{t=1}^{I} P(B=b \mid A=a_t, C=c_t)}$$

Simplify: $$P(c \mid b) \leftarrow \frac{\sum_t I(c, c_t) P(b \mid a_t, c_t)}{\sum_t P(b \mid a_t, c_t)}$$ → posterior computed above

---

Noisy-OR model



e.g. disease $X_i \in \{0,1\}$
symptom $Y \in \{0,1\}$

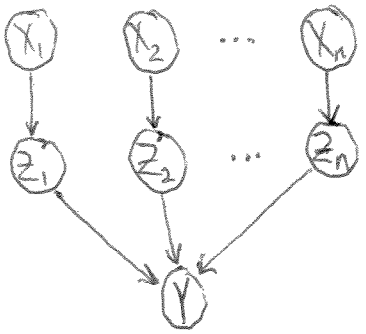
$$P(Y=1 \mid X_1, \ldots, X_n) = 1 - \prod_{i=1}^{n} (1-p_i)^{X_i} \quad \text{where } p_i \in [0,1]$$

- From complete data $\{(\vec{x}_t, y_t)\}_{t=1}^{T}$, how do we estimate $p_i \in [0,1]$?

  Noisy-OR is a "parametric" model of CPT (n values parameterize 2^n CPT entries)

  No simple, closed-form ML estimate for $p_i \in [0,1]$.

- Alternative formulation (motivation: use EM to estimate p_i)



$$P(Y \mid Z_1, \ldots, Z_n) = OR(Z_1, \ldots, Z_n)$$

logical-OR (deterministic)

$$P(Z_i=1 \mid X_i=1) = p_i$$
$$P(Z_i=1 \mid X_i=0) = 0$$

Equivalently: $$P(Z_i=0 \mid X_i) = (1-p_i)^{X_i} = \begin{cases} 1-p_i, & X_i=1 \\ 1, & X_i=0 \end{cases}$$

What is $P(Y=1|\vec{X})$ in this new model?

$$P(Y=1|\vec{X}) = \sum_{\vec{z}\in\{0,1\}^n} P(Y=1, \vec{z}|\vec{X}) \qquad \text{marginalization}$$

$$= \sum_{\vec{z}} P(Y=1|\vec{z},\vec{X}) P(\vec{z}|\vec{X}) \qquad \text{product rule}$$

$$= \sum_{\vec{z}} P(Y=1|\vec{z}) P(\vec{z}|\vec{X}) \qquad \text{conditional independence}$$

$$= \sum_{\vec{z}\neq\vec{0}} P(\vec{z}|\vec{X}) \qquad \text{because } Y = OR(\vec{z})$$

$$= 1 - P(\vec{z}=\vec{0}|\vec{X}) \qquad \text{from normalization}$$

$$= 1 - \prod_{i=1}^{n} P(z_i=0|X_i) \qquad \text{conditional independence}$$

$$= 1 - \prod_{i=1}^{n} (1-p_i)^{X_i} \qquad \underline{\text{Same as original Noisy-OR BN!}}$$

- Posterior probability

$$P(z_i=1|\vec{X},Y) = \frac{P(Y|\vec{X},z_i=1)\,P(z_i=1|\vec{X})}{P(Y|\vec{X})}$$

($P(Y|\vec{X},z_i=1)$: 1 if $y=1$, 0 if $y=0$; $P(z_i=1|\vec{X})$: p_i if $X_i=1$, 0 if $X_i=0$)

$P(Y|\vec{X})$: $1 - \prod_{i=1}^{n} (1-p_i)^{X_i}$ if $y=1$

$$= \frac{Y p_i X_i}{1 - \prod_{i=1}^{n} (1-p_i)^{X_i}}$$