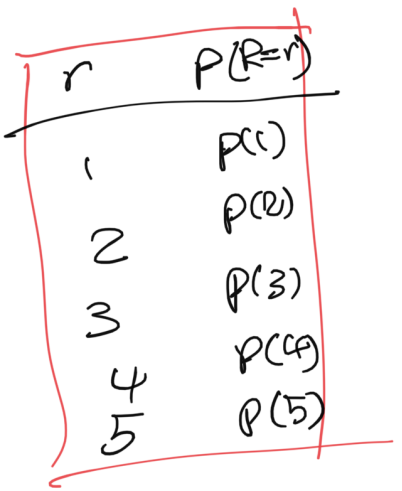
Example: one variable

Setup:

- One variable R representing the rating of a movie $\{1, 2, 3, 4, 5\}$

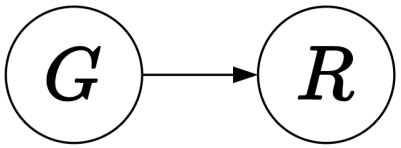
$$\mathbb{P}(R = r) = p(r)$$



Parameters:

$$\theta = (p(1), p(2), p(3), p(4), p(5))$$

Example: two variables



$$\mathbb{P}(G = g, R = r) = p_G(g)p_R(r \mid g)$$

$$\mathcal{D}_{\text{train}} = \{(\text{d}, 4), (\text{d}, 4), (\text{d}, 5), (\text{c}, 1), (\text{c}, 5)\}$$

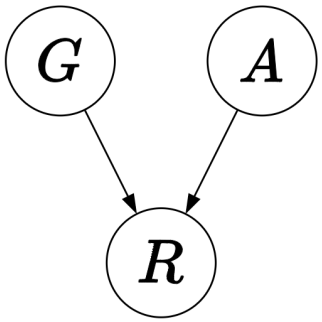
Intuitive strategy: Estimate each local conditional distribution (p_G and p_R) separately

θ:

g	$p_G(g)$
d	3/5
c	2/5

g	r	$\text{count}_R(g, r)$
d	4	2
d	5	1
c	1	1
c	5	1

Example: v-structure



$$\mathcal{D}_{\text{train}} = \{(\text{d}, 0, 3), (\text{d}, 1, 5), (\text{d}, 0, 1), (\text{c}, 0, 5), (\text{c}, 1, 4)\}$$

Parameters: $\theta = (p_G, p_A, p_R)$

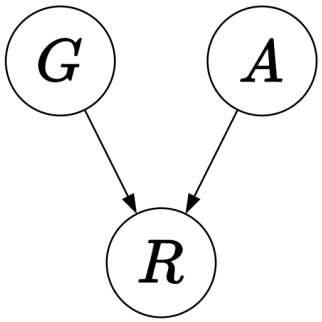
θ:

g	$p_G(g)$
d	3/5
c	2/5

a	$p_A(a)$
0	3/5
1	2/5

g	a	r	$\text{count}_R(g, a, r)$
d	0	1	1
d	0	3	1
d	1	5	1
c	0	5	1
c	1	4	1

Example: v-structure



$$\mathcal{D}_{\text{train}} = \{(\text{d}, 0, 3), (\text{d}, 1, 5), (\text{d}, 0, 1), (\text{c}, 0, 5), (\text{c}, 1, 4)\}$$

Parameters: $\theta = (p_G, p_A, p_R)$

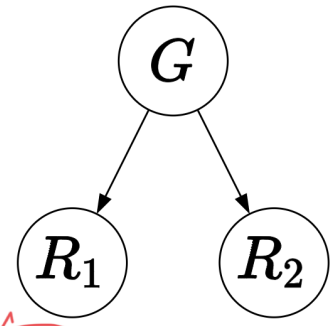
θ:

g	$p_G(g)$
d	3/5
c	2/5

a	$p_A(a)$
0	3/5
1	2/5

g	a	r	$\text{count}_R(g, a, r)$
d	0	1	1
d	0	3	1
d	1	5	1
c	0	5	1
c	1	4	1

Example: inverted-v structure



$P(R_1|G) = P(R_2|G)$

$$\mathcal{D}_{\text{train}} = \{(\text{d}, 4, 5), (\text{d}, 4, 4), (\text{d}, 5, 3), (\text{c}, 1, 2), (\text{c}, 5, 4)\}$$

Parameters: $\theta = (p_G, p_R)$

θ:

g	$p_G(g)$
d	3/5
c	2/5

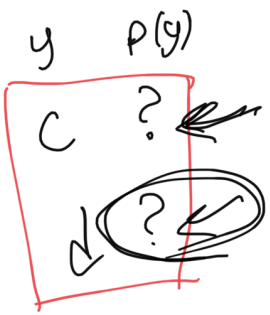
g	r	$\text{count}_R(g, r)$
d	3	1
d	4	3
d	5	2
c	1	1
c	2	1
c	4	1
c	5	1

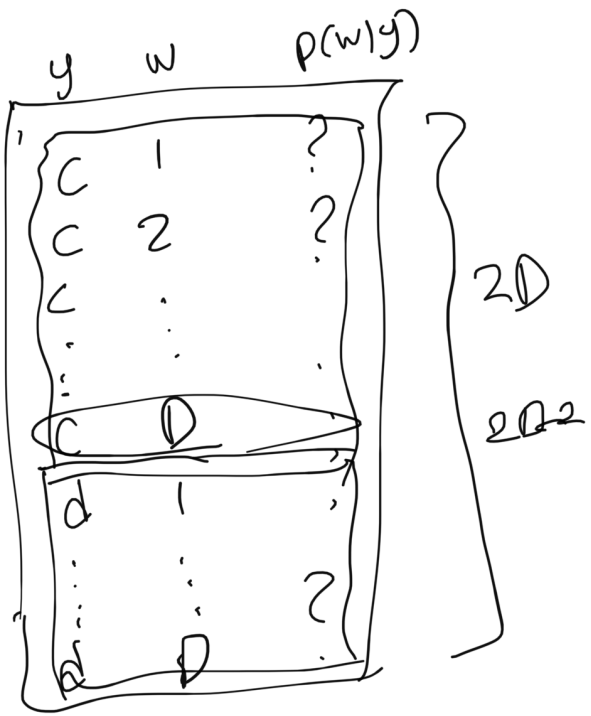
answer in chat

Question

If Y can take on 2 values and each W_j can take on D values, how many parameters are there?

activate deactivate reset report





General case

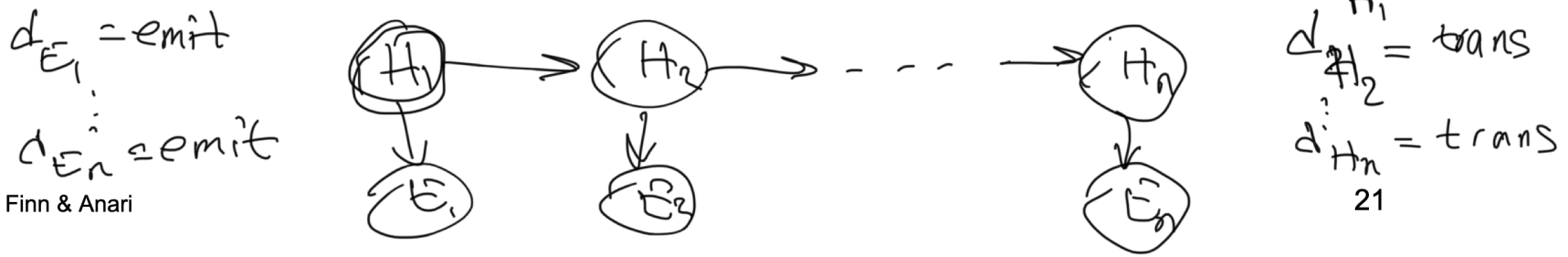
Bayesian network: variables $X_1, \ldots, X_n$

Parameters: collection of distributions $\theta = \{p_d : d \in D\}$
(e.g., $D = \{\text{start}, \text{trans}, \text{emit}\}$)

Each variable X_i is generated from distribution p_{d_i}:

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} p_{d_i}\left(x_i \mid x_{\text{Parents}(i)}\right)$$

Parameter sharing: d_i could be same for multiple i



Maximum likelihood

$$\mathcal{D}_{\text{train}} = \{(\text{d}, 4), (\text{d}, 5), (\text{c}, 5)\}$$

$$\max_{p_G(\cdot), p_R(\cdot|\text{c}), p_R(\cdot|\text{d})} \left(p_G(\text{d}) p_R(4 \mid \text{d}) p_G(\text{d}) p_R(5 \mid \text{d}) p_G(\text{c}) p_R(5 \mid \text{c})\right)$$

subject to

$$\sum_g p_G(g) = 1$$

$$\sum_r p_R(r \mid G = c) = 1$$

$$\sum_r p_R(r \mid G = d) = 1$$

Maximum likelihood

$$\mathcal{D}_{\text{train}} = \{(\text{d}, 4), (\text{d}, 5), (\text{c}, 5)\}$$

$$\max_{p_G(\cdot)} (p_G(\text{d}) p_G(\text{d}) p_G(\text{c})) \max_{p_R(\cdot|\text{c})} p_R(5 \mid \text{c}) \max_{p_R(\cdot|\text{d})} (p_R(4 \mid \text{d}) p_R(5 \mid \text{d}))$$

$\sum_g p_G(g) = 1$ $\sum_r p_R(r \mid G = \text{c}) = 1$ $\sum_r p_R(r \mid G = \text{d}) = 1$

Regularization: Laplace smoothing

Key idea: Laplace smoothing

For each distribution d and partial assignment $(x_{\text{Parents}(i)}, x_i)$, add λ to $\text{count}_d(x_{\text{Parents}(i)}, x_i)$.
Then normalize to get probability estimates.

Interpretation: hallucinate λ occurrences of each local assignment

Larger $\lambda \Rightarrow$ more smoothing $\Rightarrow$ probabilities closer to uniform.

for each $(x_{parents}, y_i)$ we add λ imaginary data points where $(x_{parents}, x_i)$ was "observed".

Maximum likelihood

$$\mathcal{D}_{\text{train}} = \{(\text{d}, 4), (\text{d}, 5), (\text{c}, 5)\}$$

$$\max_{p_G(\cdot)} (p_G(\text{d})p_G(\text{d})p_G(\text{c})) \max_{p_R(\cdot|\text{c})} p_R(5 \mid \text{c}) \max_{p_R(\cdot|\text{d})} (p_R(4 \mid \text{d})p_R(5 \mid \text{d}))$$

$P_G(d)^2 P_G(c)^1$

$P_R(1|c)^0 P_R(2|c)^0 \cdots P_R(5|c)^1$

$P_G(d) + P_G(c) = 1$

$P_R(1|c) + P_R(2|c) + \cdots + P_R(5|c) = 1$

Solution:

$$p_G(\text{d}) = \frac{2}{3}, p_G(\text{c}) = \frac{1}{3}, p_R(5 \mid \text{c}) = 1, p_R(4 \mid \text{d}) = \frac{1}{2}, p_R(5 \mid \text{d}) = \frac{1}{2}$$

- Key: decomposes into subproblems, one for each distribution d and assignment x_{Parents}
- For each subproblem, solve in closed form

$$\max x^5 y^2 z^3 \text{ subject to } x + y + z = 1$$

Maximum likelihood

$$\mathcal{D}_{\text{train}} = \{(\text{d}, 4), (\text{d}, 5), (\text{c}, 5)\}$$

$$\max_{p_G(\cdot)} (p_G(\text{d}) p_G(\text{d}) p_G(\text{c})) \max_{p_R(\cdot|\text{c})} p_R(5 \mid \text{c}) \max_{p_R(\cdot|\text{d})} (p_R(4 \mid \text{d}) p_R(5 \mid \text{d}))$$

Solution:

$$p_G(\text{d}) = \frac{2}{3}, p_G(\text{c}) = \frac{1}{3}, p_R(5 \mid \text{c}) = 1, p_R(4 \mid \text{d}) = \frac{1}{2}, p_R(5 \mid \text{d}) = \frac{1}{2}$$

- Key: decomposes into subproblems, one for each distribution d and assignment x_{Parents}
- For each subproblem, solve in closed form

$$\max x^5 y^2 z^3 \text{ subject to } x + y + z = 1$$

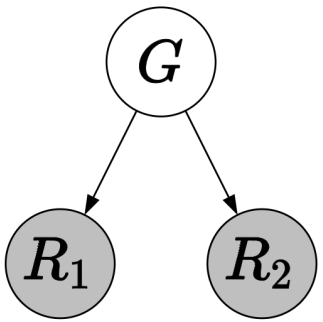
$$x = \frac{5}{5+2+3} \quad y = \frac{2}{5+2+3} \quad z = \frac{3}{5+2+3}$$

is achieved when $x \propto 5, y \propto 2, z \propto 3$

Maximum marginal likelihood

Variables: H is hidden, $E = e$ is observed

Example:



$$H = G \quad E = (R_1, R_2) \quad e = (1, 2)$$
$$\theta = (p_G, p_R)$$

$P(E=e) =$

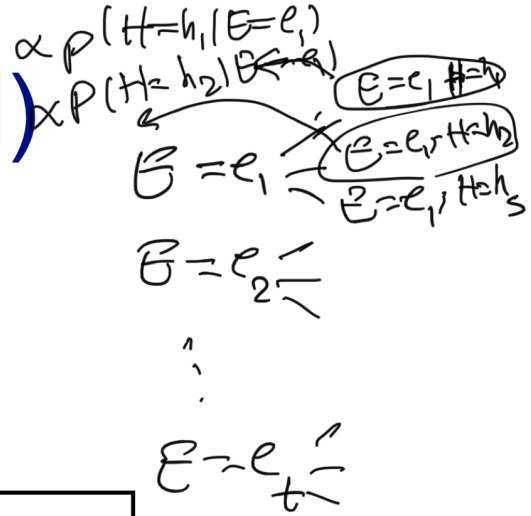
$\sum_h P(E=e \wedge H=h)$

Maximum marginal likelihood objective:

$$\max_\theta \prod_{e \in \mathcal{D}_{\text{train}}} \mathbb{P}(E = e; \theta)$$

$$= \max_\theta \prod_{e \in \mathcal{D}_{\text{train}}} \sum_h \mathbb{P}(H = h, E = e; \theta)$$

Expectation Maximization (EM)



Intuition: generalization of the K-means algorithm

Variables: H is hidden, $E = e$ is observed

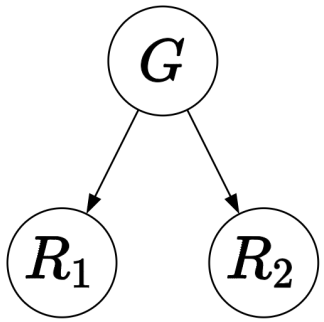
Algorithm: Expectation Maximization (EM)

Initialize θ

E-step:

- Compute $q(h) = \mathbb{P}(H = h \mid E = e; \theta)$ for each h (use any probabilistic inference algorithm)
- Create weighted points: (h, e) with weight $q(h)$

Example: one iteration of EM



$$\mathcal{D}_{\text{train}} = \{(?, 2, 2), (?, 1, 2)\}$$

$\propto P[G = g, R_1 = r_1, R_2 = r_2]$

θ:

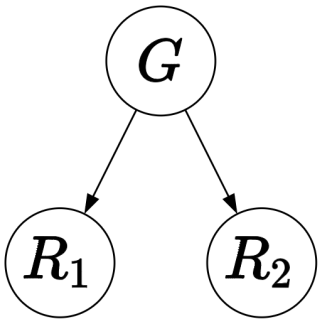
g	$p_G(g)$
c	0.5
d	0.5

g	r	$p_R(r \mid g)$
c	1	0.4
c	2	0.6
d	1	0.6
d	2	0.4

E-step

(r_1, r_2)	g	$\mathbb{P}(G = g, R_1 = r_1, R_2 = r_2)$	$q(g)$
(2, 2)	c	$0.5 \cdot 0.6 \cdot 0.6 = 0.18$	$\frac{0.18}{0.18+0.08} = 0.69$
(2, 2)	d	$0.5 \cdot 0.4 \cdot 0.4 = 0.08$	$\frac{0.08}{0.18+0.08} = 0.31$
(1, 2)	c	$0.5 \cdot 0.4 \cdot 0.6 = 0.12$	$\frac{0.12}{0.12+0.12} = 0.5$
(1, 2)	d	$0.5 \cdot 0.6 \cdot 0.4 = 0.12$	$\frac{0.12}{0.12+0.12} = 0.5$

Example: one iteration of EM



$$\mathcal{D}_{\text{train}} = \{(?, 2, 2), (?, 1, 2)\}$$

θ:

g	$p_G(g)$
c	0.5
d	0.5

g	r	$p_R(r \mid g)$
c	1	0.4
c	2	0.6
d	1	0.6
d	2	0.4

E-step →

(r_1, r_2)	g	$\mathbb{P}(G = g, R_1 = r_1, R_2 = r_2)$	$q(g)$
(2, 2)	c	$0.5 \cdot 0.6 \cdot 0.6 = 0.18$	$\frac{0.18}{0.18+0.08} = 0.69$
(2, 2)	d	$0.5 \cdot 0.4 \cdot 0.4 = 0.08$	$\frac{0.08}{0.18+0.08} = 0.31$
(1, 2)	c	$0.5 \cdot 0.4 \cdot 0.6 = 0.12$	$\frac{0.12}{0.12+0.12} = 0.5$
(1, 2)	d	$0.5 \cdot 0.6 \cdot 0.4 = 0.12$	$\frac{0.12}{0.12+0.12} = 0.5$

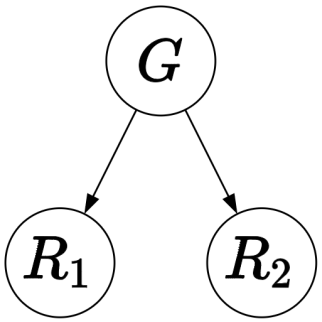
M-step →

θ:

g	count	$p_G(g)$
c	0.69 + 0.5	0.59
d	0.31 + 0.5	0.41

g	r	count	$p_R(r \mid g)$
c	1	0.5	0.21
c	2	0.5 + 0.69 + 0.69	0.79
d	1	0.5	0.31
d	2	0.5 + 0.31 + 0.31	0.69

Example: one iteration of EM



$$\mathcal{D}_{\text{train}} = \{(?, 2, 2), (?, 1, 2)\}$$

θ:

g	$p_G(g)$
c	0.5
d	0.5

g	r	$p_R(r \mid g)$
c	1	0.4
c	2	0.6
d	1	0.6
d	2	0.4

E-step →

(r_1, r_2)	g	$\mathbb{P}(G = g, R_1 = r_1, R_2 = r_2)$	$q(g)$
(2, 2)	c	$0.5 \cdot 0.6 \cdot 0.6 = 0.18$	$\frac{0.18}{0.18+0.08} = 0.69$
(2, 2)	d	$0.5 \cdot 0.4 \cdot 0.4 = 0.08$	$\frac{0.08}{0.18+0.08} = 0.31$
(1, 2)	c	$0.5 \cdot 0.4 \cdot 0.6 = 0.12$	$\frac{0.12}{0.12+0.12} = 0.5$
(1, 2)	d	$0.5 \cdot 0.6 \cdot 0.4 = 0.12$	$\frac{0.12}{0.12+0.12} = 0.5$

M-step →

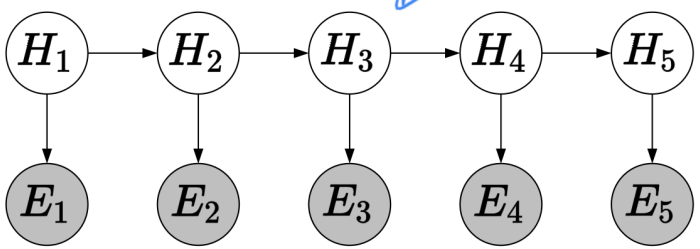
θ:

g	count	$p_G(g)$
c	0.69 + 0.5	0.59
d	0.31 + 0.5	0.41

g	r	count	$p_R(r \mid g)$
c	1	0.5	0.21
c	2	0.5 + 0.69 + 0.69	0.79
d	1	0.5	0.31
d	2	0.5 + 0.31 + 0.31	0.69

Increase marginal likelihood

Application: decipherment as an HMM



E-step: forward-backward algorithm computes

$$q_i(h) \stackrel{\text{def}}{=} \mathbb{P}(H_i = h \mid E_1 = e_1, \ldots E_n = e_n)$$

M-step: count (fractional) and normalize

$$\text{count}_{\text{emit}}(h, e) = \sum_{i=1}^{n} q_i(h) \cdot [e_i = e]$$

$$p_{\text{emit}}(e \mid h) \propto \text{count}_{\text{emit}}(h, e)$$