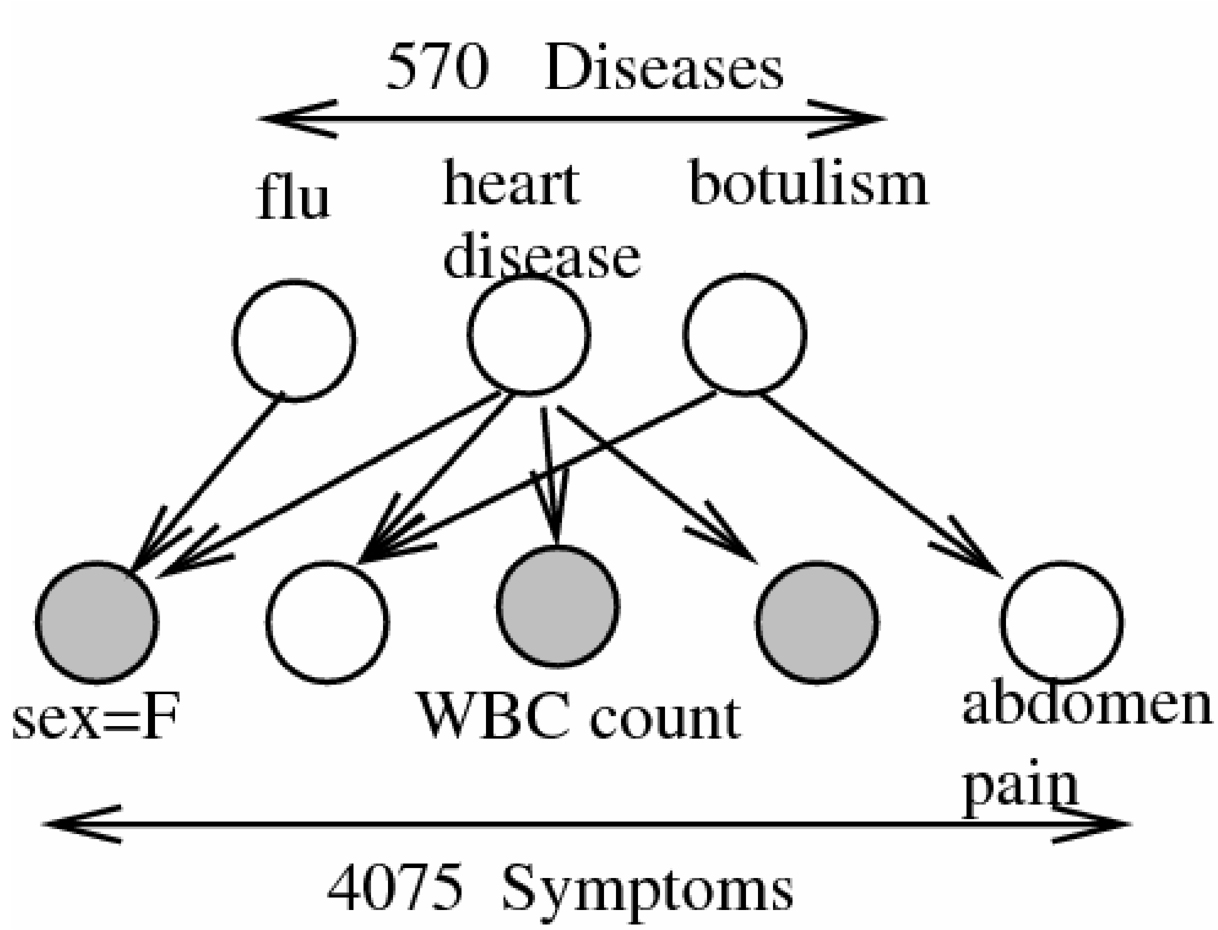
CS340 Machine learning
QMR

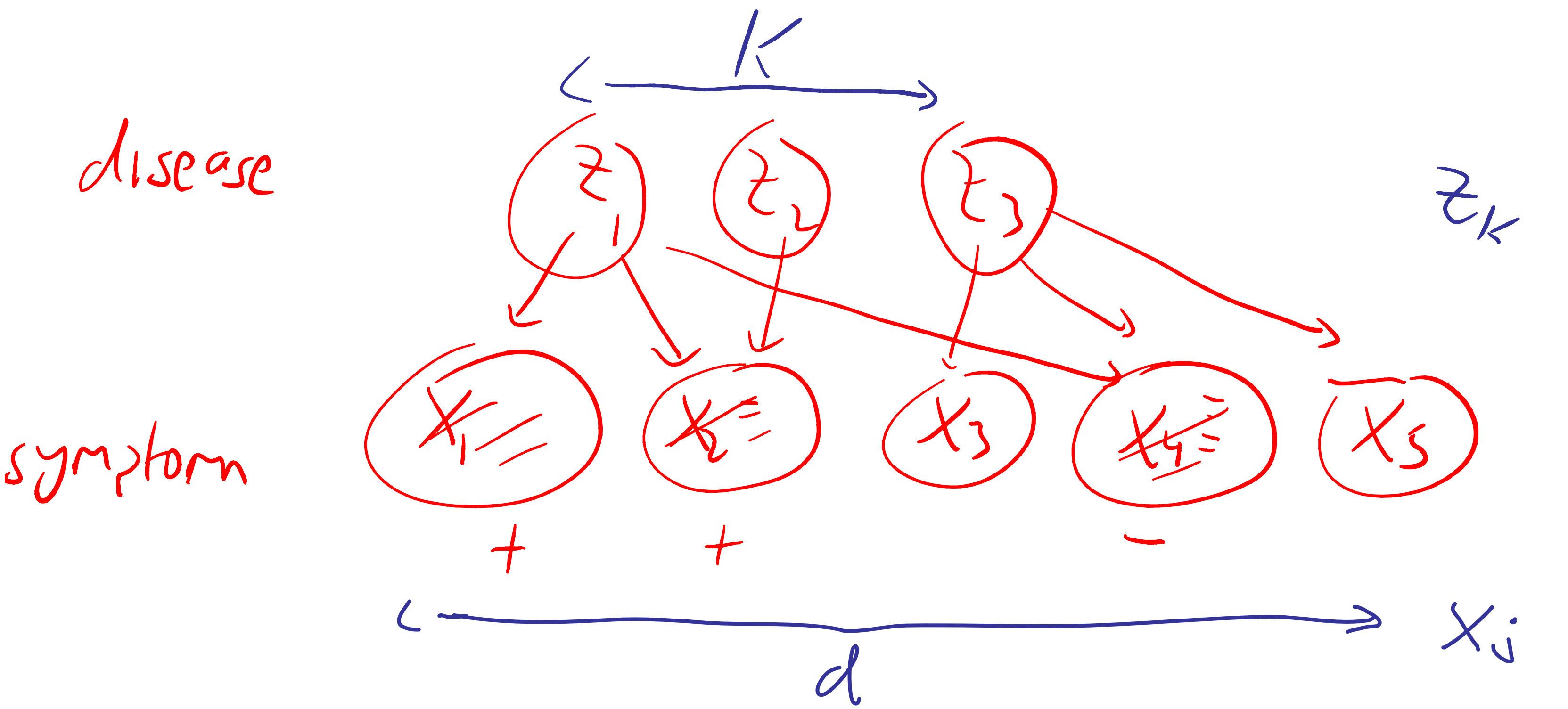
Quick Medical Reference

- Probabilistic expert system encoded as a DGM.
- Nodes are binary. Parameters hand-coded.



Inference in QMR

- Infer probability of each disease given observations on subset of symptoms

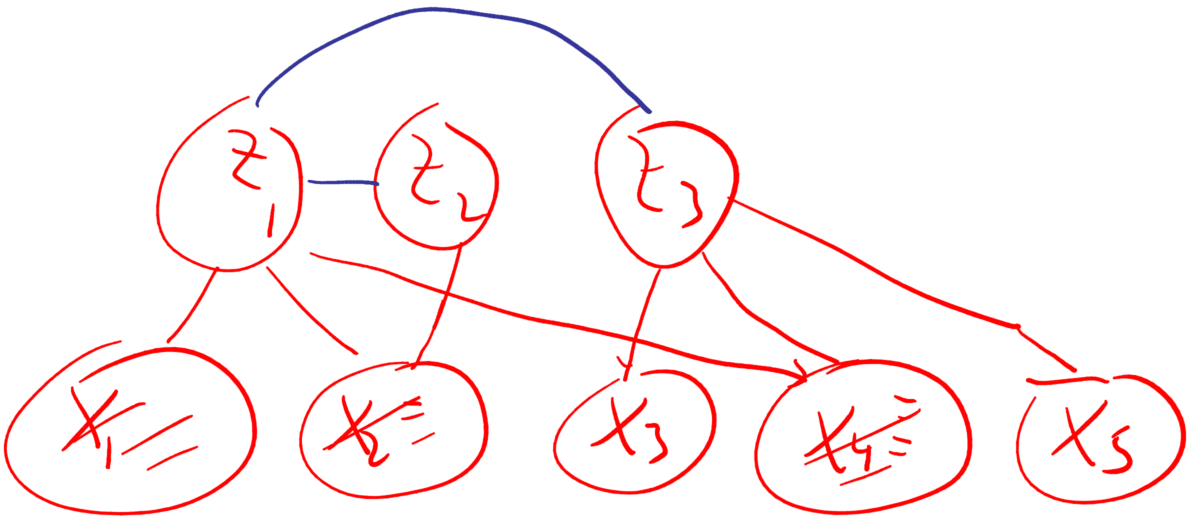


$$p(Z_k = + | X_1 = +, X_2 = +, X_4 = -)$$

Complexity of inference

- The disease nodes become dependent in the posterior due to explaining away. Thus exact inference takes $O(2^w)$ time, where w is (lower bounded by) the size of the largest clique of the moralized graph.

Moral graph

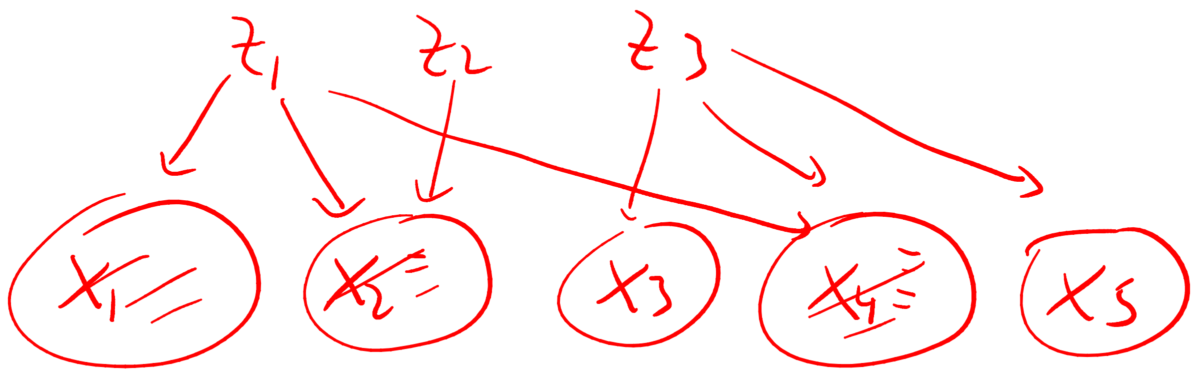


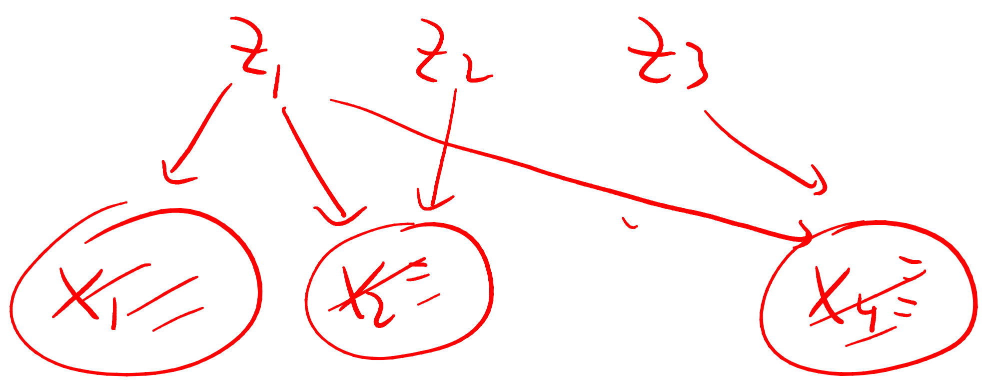
Complexity of inference

- The disease nodes become dependent in the posterior due to explaining away. Thus exact inference takes $O(2^w)$ time, where w is (lower bounded by) the size of the largest clique of the moralized graph.
- For QMR, w ~ 151, so exact inference is intractable.

Barren nodes

- We can remove leaves with no evidence, since their CPDs sum to one: $\sum_{x_3} p(x_3|z_1, z_2, z_3) = 1$



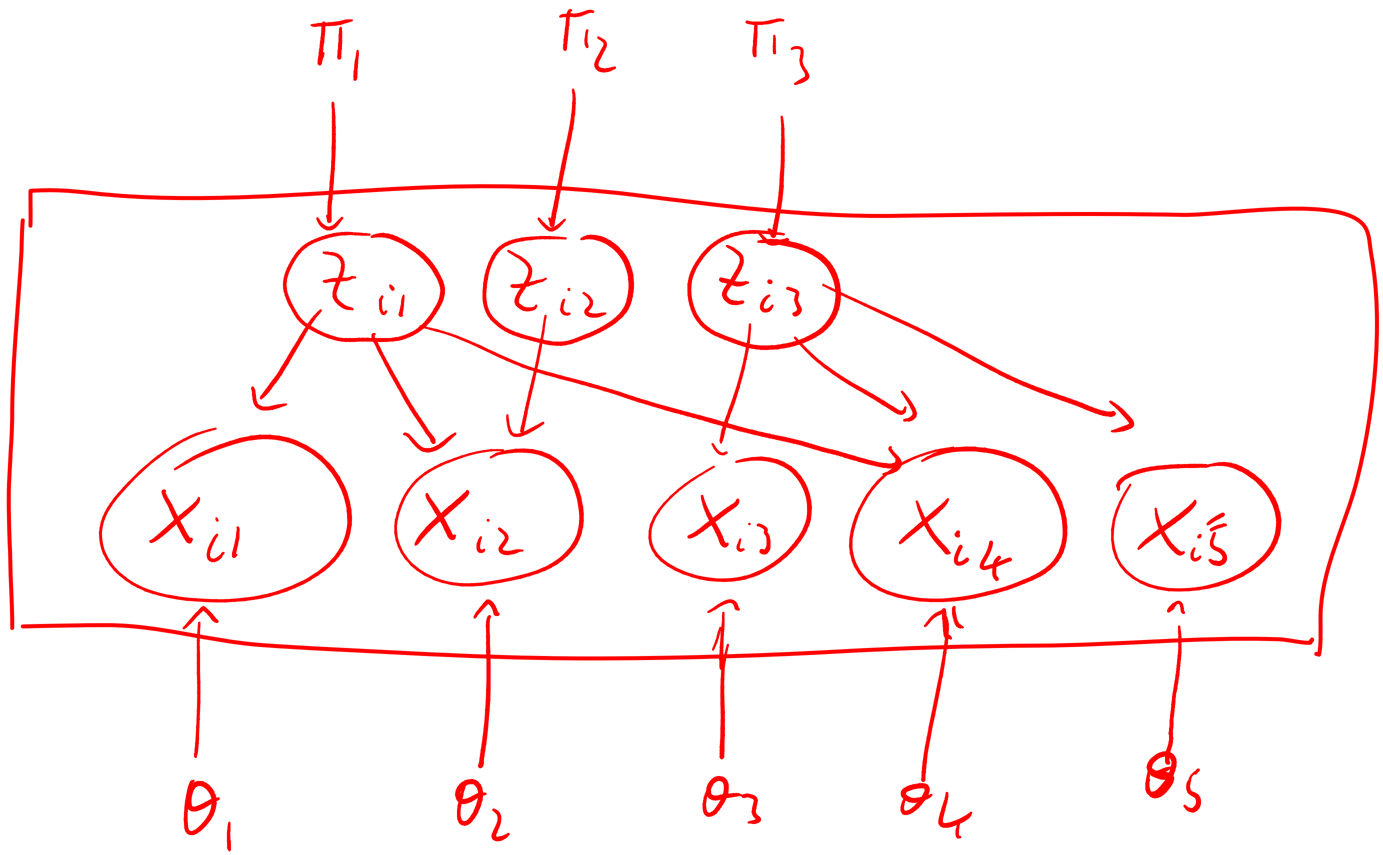


This can reduce the size of the cliques in the moral graph.

Quickscore algorithm

- The quickscore algorithm exploits the special structure (noisy-OR: see later) of the symptom CPDs, but still takes $O(2^p)$ time, where p = #positive findings. For QMR, $p > 20$.
- Many approximate methods have been developed for this model.
- In HW5, you will use exact inference on a small model.

Parameters of the QMR model



Parameter estimation

- Let $D = (X_{ij}, Z_{ik})_{i=1:n, j=1:d, k=1:K}$ be the training data.
- Let us assume no missing data.
- By global parameter independence, the posterior factorizes

$$p(\boldsymbol{\theta}, \boldsymbol{\pi} | D) \quad \propto \quad \prod_{k=1}^{K} p(\pi_k) p(D|\pi_k) \prod_{j=1}^{d} p(\boldsymbol{\theta}_j) p(D|\boldsymbol{\theta}_j)$$

Root CPDs in QMR

- CPDs = conditional probability distribution, P(node|parents)
- Root nodes have Bernoulli distribution, representing base rate of the disease.

$$p(Z_k = 1) = \pi_k$$

- Likelihood

$$p(D|\pi_k) = \prod_{i=1}^{n} \pi_k^{I(z_{ik}=1)} (1 - \pi_k)^{I(z_{ik}=0)}$$

- Prior

$$p(\pi_k) = Beta(\pi_k | a_k, b_k)$$

- Posterior

$$p(\boldsymbol{\pi}|D) = \prod_k Beta(\pi_k | a_k + N(Z_k = 1), b_k + N(Z_k = 0))$$

Leaf CPDs in QMR

- Let θ_j be the parameters of $p(X_j|pa(X_j))$.
- Representing $p(X_j|pa(X_j))$ as a table would need $2^{\#parents}$ parameters. Instead we use a noisy-OR parameterization, which has #parents parameters. (Could also use logistic regression.)

Noisy-ORs

- If parent Z_k is on, it will turn on its child X_j.
- But with probability q_{kj}, the “wire” from Z_k to X_j may fail, and the on parent will be inhibited.
- We assume such failures occur independently.
- Deterministic OR corresponds to all q_{kj}=0.

$$p(X_j = 0|Z_{\pi_j}) = \prod_{k \in \pi_j} q_{kj}^{I(Z_k=1)} = \prod_{k \in \pi_j : Z_k = 1} q_{kj}$$

$$\begin{tikzcd} z_1 \arrow[dr] & & z_2 \arrow[dl] \\ & X_j & \end{tikzcd}$$

Z_1	Z_2	$P(X_j = 0\|Z_1, Z_2)$	$P(X_j = 1\|Z_1, Z_2)$
0	0	1	0
1	0	q_{1j}	$1 - q_{1j}$
0	1	q_{2j}	$1 - q_{2j}$
1	1	$q_{1j}q_{2j}$	$1 - q_{1j}q_{2j}$

Leak nodes

- Sometimes a child is on even if all its parents are off, since there may be some other "hidden" cause.
- To explain this, we assume every child has an extra "leak" or background parent that is always on. This will turn the child on unless it is inhibited w.p. q_{0j}.

$$p(X_j = 0|Z_{\pi_j}) = q_{0j} \prod_{k \in \pi_j} q_{kj}^{I(Z_k=1)}$$

B	Z_1	Z_2	$P(X_j = 0 \mid Z_1, Z_2)$	$P(X_j = 1 \mid Z_1, Z_2)$
1	0	0	q_{0j}	$1 - q_{0j}$
1	1	0	$q_{0j}q_{1j}$	$1 - q_{0j}q_{1j}$
1	0	1	$q_{0j}q_{2j}$	$1 - q_{0j}q_{2j}$
1	1	1	$q_{0j}q_{1j}q_{2j}$	$1 - q_{0j}q_{1j}q_{2j}$

Alternative parameterization

- q_{kj} = prob k fails to cause j.
- Let $w_{kj} = 1 - q_{kj}$ = prob k causes j. Then

$$
\begin{aligned}
p(X_j = 1 | Z_{\pi_j}) &= 1 - \prod_{k \in \pi_j} q_{kj}^{I(Z_k=1)} \\
&= 1 - \prod_{k \in \pi_j} (1 - w_{kj})^{I(Z_k=1)}
\end{aligned}
$$

Parameter estimation for noisy-ORs

- Consider the case of a single cause and a single effect.

B	C	$P(E=0 \mid C, \mathbf{w})$	$P(E=1 \mid C, \mathbf{w})$
1	0	$1 - w_0$	$1 - (1 - w_0)$
1	1	$(1 - w_0)(1 - w_1)$	$1 - (1 - w_0)(1 - w_1)$

- We want to estimate w from a contingency table of counts.

	Effect absent $E = 0$	Effect present $E = 1$
Cause absent $C = 0$	$N(E = 0, C = 0)$	$N(C = 0, E = 1)$
Cause present $C = 1$	$N(E = 0, C = 1)$	$N(C = 1, E = 1)$

Maximum likelihood estimation

- Let p(e|c) be the empirical probabilities (derived from the counts N(e,c)).
- Let p(e|c,w) be the model-predicted probabilities.
- The MLE is gotten by finding the w that minimizes the KL divergence

$$\mathbf{w} = \arg\min_{\mathbf{w}} KL(p(e|c)||p(e|c,\mathbf{w}))$$

- Hence we require

$$\begin{aligned} p(e=1|c=1,\mathbf{w}) &= p(e=1|c=1) \\ p(e=1|c=0,\mathbf{w}) &= p(e=1|c=0) \end{aligned}$$

“Structure and strength in causal induction”, Griffiths and Tenenbaum, Cognitive Psychology, 51:334-384, 2005

MLE for w0

- Recall

$$p(e=1|c,\mathbf{w}) \quad = \quad 1-(1-w_0)(1-w_1)^{I(c=1)}$$

- Set

$$w_0 \quad = \quad p(e=1|c=0) = \frac{N(e=1,c=0)}{N(e=1,c=0)+N(e=0,c=0)}$$

- Then

$$\begin{aligned} p(e=1|c=0,\mathbf{w}) \quad &= \quad 1-(1-w_0) \\ &= \quad 1-(1-p(e=1|c=0)) = p(e=1|c=0) \end{aligned}$$

MLE for w1

- Recall

$$p(e=1|c,\mathbf{w}) \quad = \quad 1-(1-w_0)(1-w_1)^{I(c=1)}$$

- Set

$$w_1 \quad = \quad \frac{p(e=1|c=1)-p(e=1|c=0)}{1-p(e=1|c=0)}$$

"causal power"

- Then

$$\begin{aligned} p(e=1|c=1,\mathbf{w}) \quad &= \quad 1-(1-w_0)(1-w_1) \\ &= \quad p(e=1|c=1) \end{aligned}$$

Derivation left as homework exercise

Bayesian parameter estimation

- Since $0 \leq w_j \leq 1$, a suitable prior is

$$p(\mathbf{w}) = Beta(w_0|a_0,b_0)Beta(w_1|a_1,b_1)$$

- Likelihood

C	$P(E=0\|C,\mathbf{w})$	$P(E=1\|C,\mathbf{w})$
0	$\theta_{00} = 1 - w_0$	$\theta_{01} = 1 - (1 - w_0)$
1	$\theta_{10} = (1 - w_0)(1 - w_1)$	$\theta_{11} = 1 - (1 - w_0)(1 - w_1)$

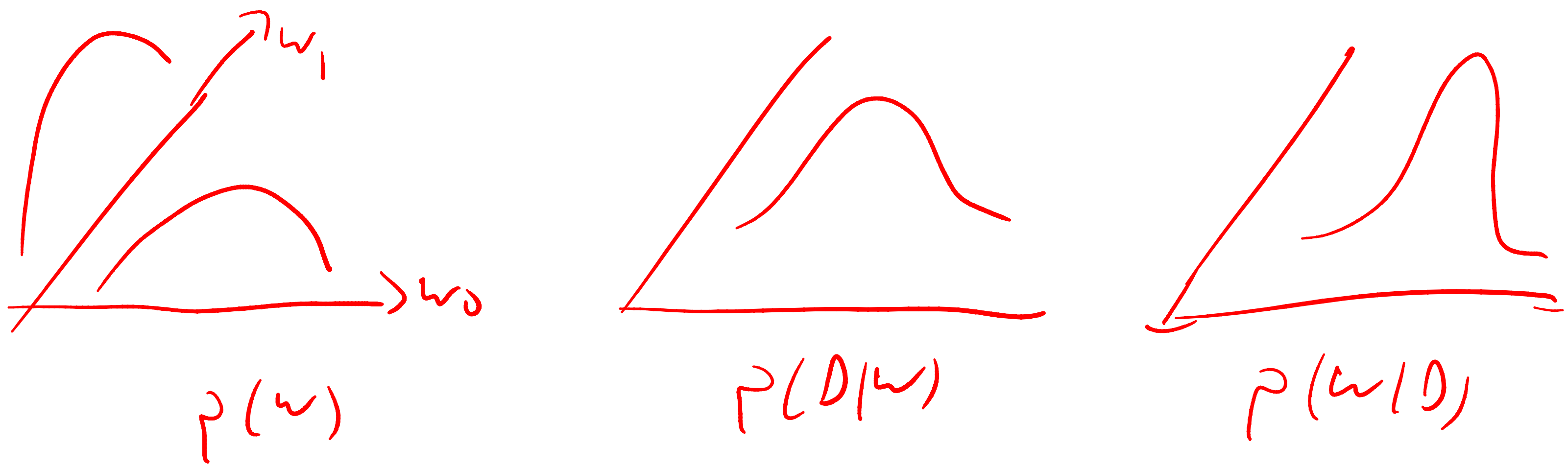
$$p(D|\mathbf{w}) = \prod_{i=1}^{n}\prod_{e=0}^{1}\prod_{c=0}^{1}\theta_{ec}^{I(c_i=c))I(e_i=e)}$$

$$= \prod_{e=0}^{1}\prod_{c=0}^{1}\theta_{ec}^{N(e,c)}$$

- Not conjugate ☹

Gridding

- We can compute $p(w_0,w_1|D)$ by gridding up the space.



- This is only tractable for 2 parameters.
- In general, need to use Monte Carlo or variational methods.