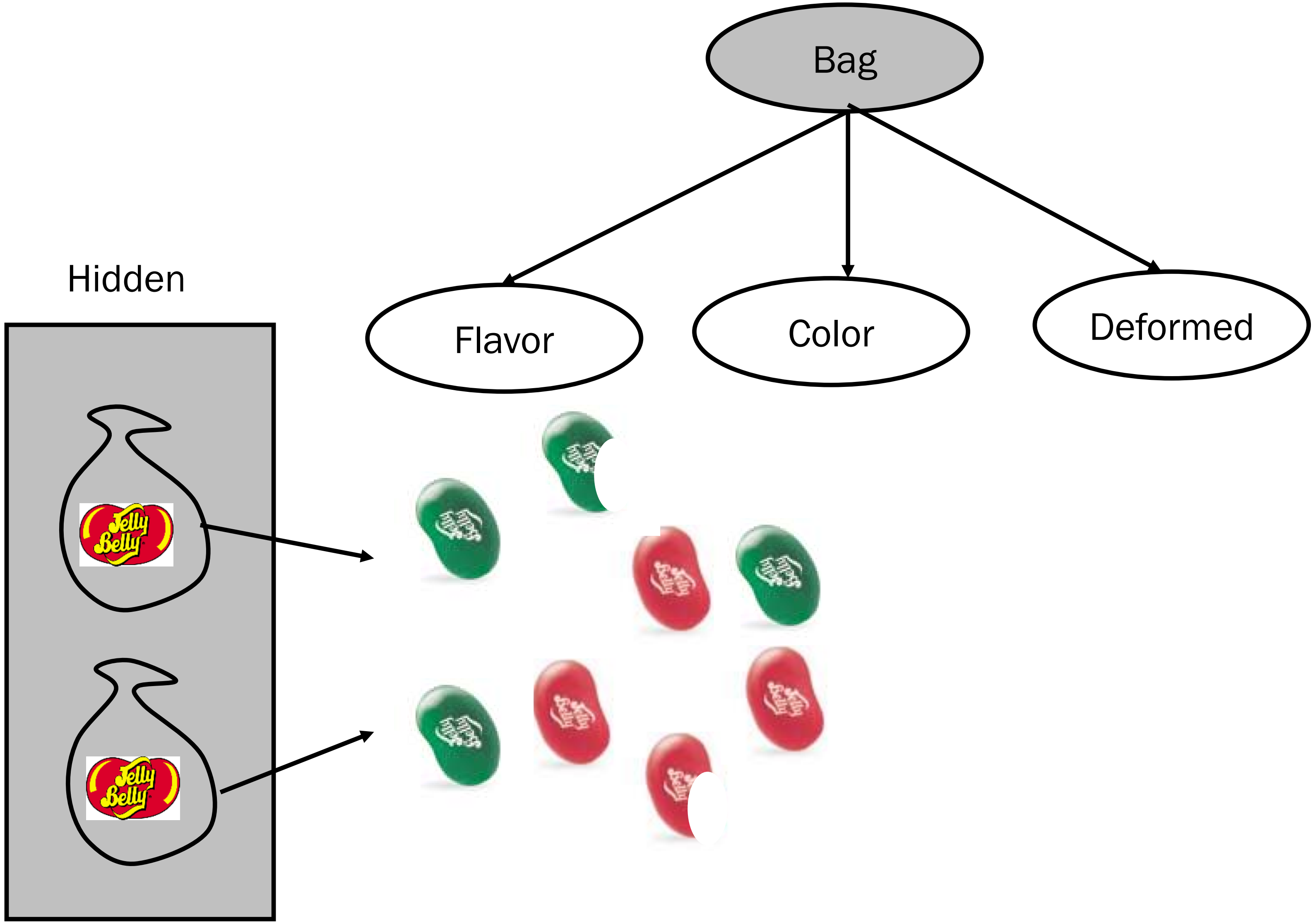
FUDAN SUMMER SCHOOL INTRODUCTION TO AI

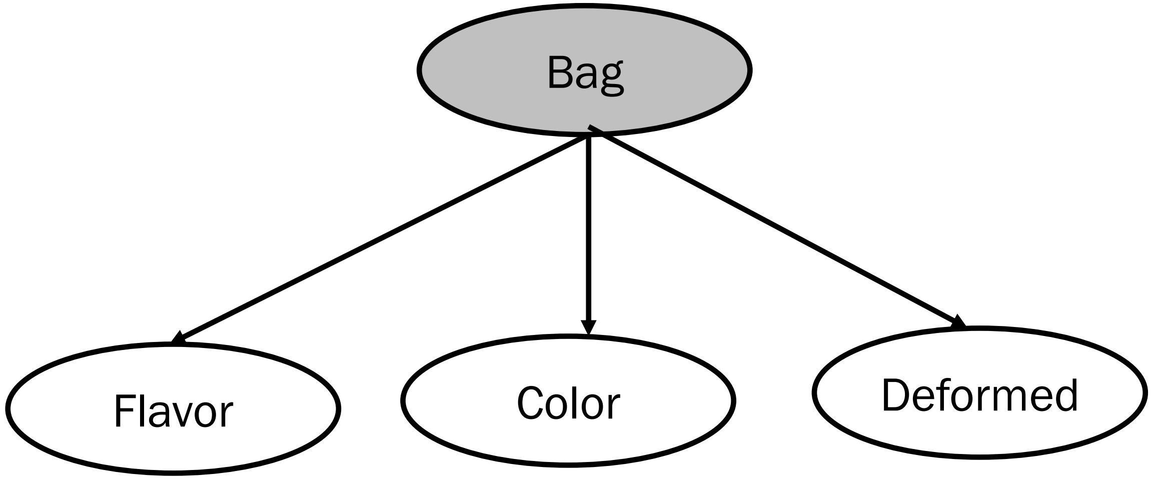
Probabilistic Reasoning and Decision Making

Day 9 –Expectation Maximization (EM) Algorithm

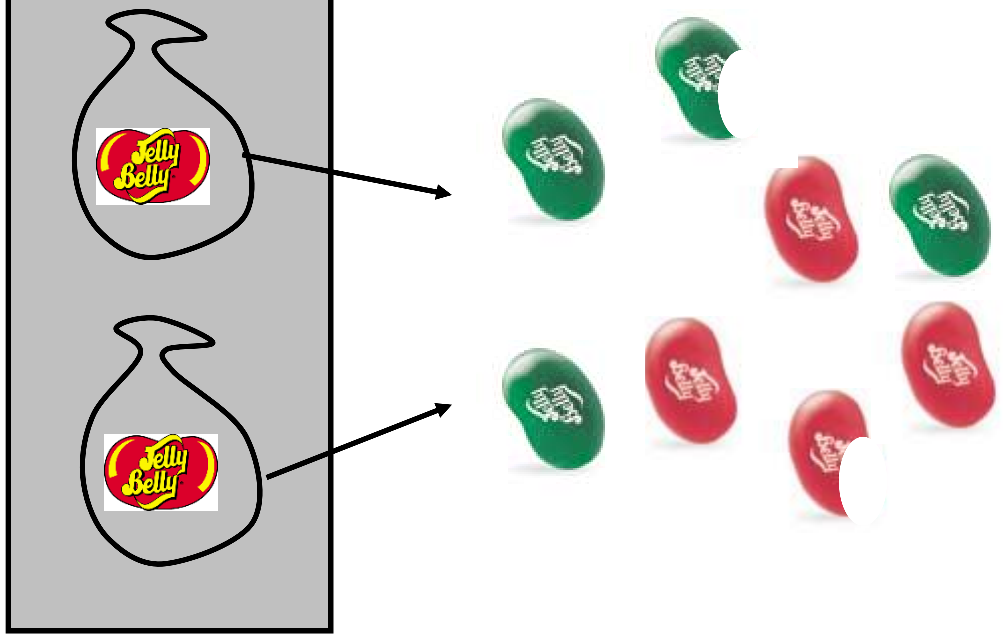
Learning from Partially Observed Data, Example



Learning from Partially Observed Data, Example

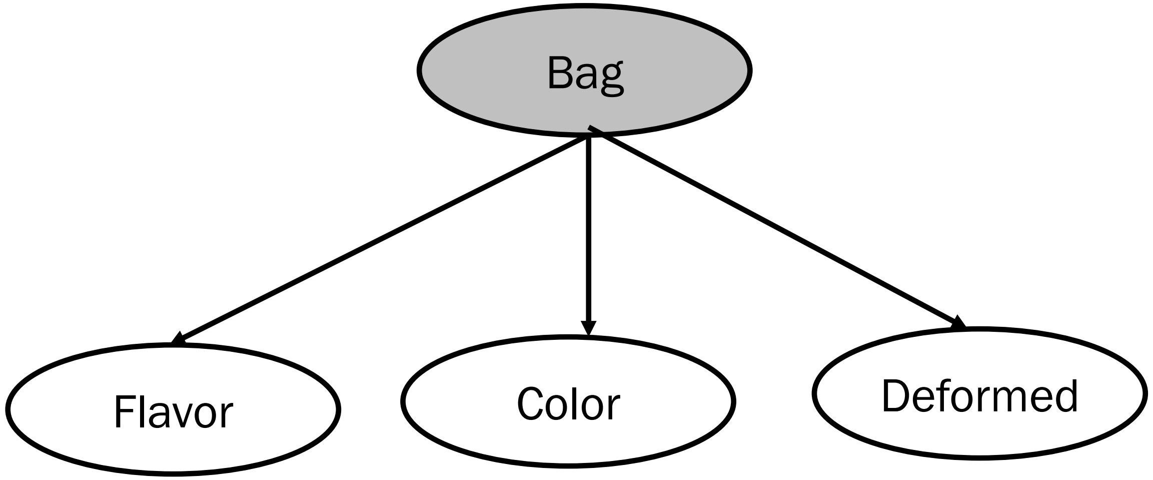


Hidden



	C=red		C=green	
	D=1	D=0	D=1	D=0
F=cherry	273	93	104	90
F=jalapeño	79	100	94	167

Learning from Partially Observed Data, Example

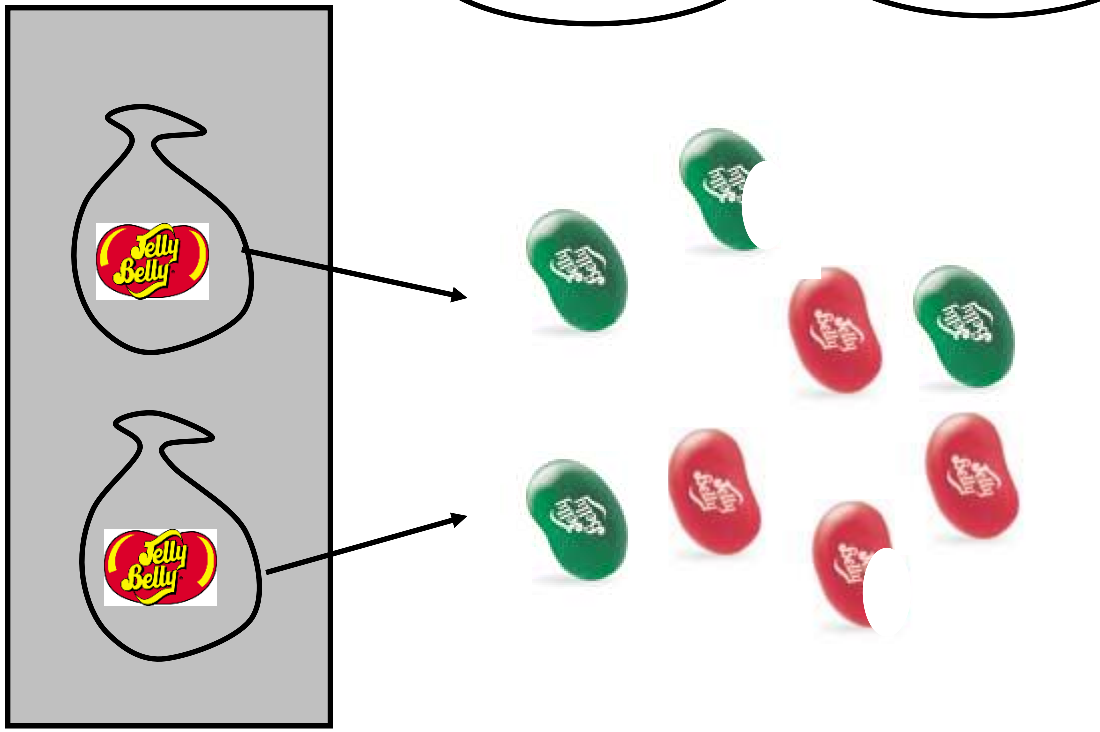

Bag

Flavor

Color

Deformed

Hidden



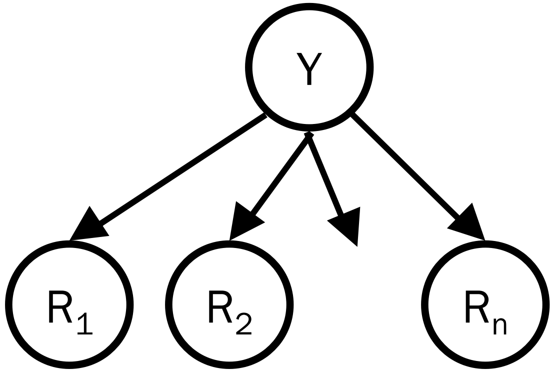
If we knew which beans came from which bag, we could trivially determine the distribution of the bag

If we knew the distribution of each bag, we could probabilistically assign beans to bags

We don't know either!

Learning with partial data

Another example: Movie Recommender System



Learning with partial data, generally

Let $\{X_1, X_2, \ldots, X_n\}$ denote all the nodes in a Bayes Net.
Let H denote the subset of hidden (unobserved) nodes.
Let V denote the subset of visible (observed) nodes.

$$V \cup H = \{X_1, X_2, \ldots, X_n\}$$

Goal: Estimate the CPTs in the BN to maximize the probability of the partially observed data

Learning with partial data, generally

Let $\{X_1, X_2, \dots, X_n\}$ denote all the nodes in a Bayes Net.
Let H denote the subset of hidden (unobserved) nodes.
Let V denote the subset of visible (observed) nodes.

$$V \cup H = \{X_1, X_2, \dots, X_n\}$$

Goal: Estimate the CPTs in the BN to maximize the probability of the partially observed data

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

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

What should we do next?

A. Use the product rule

B. Express $P\left(V = v^{(t)}\right)$ using the conditional independence in the BN

C. Use marginalization

D. Use Bayes Rule

Learning with partial data, generally

Let $\{X_1, X_2, \dots, X_n\}$ denote all the nodes in a Bayes Net.
Let H denote the subset of hidden (unobserved) nodes.
Let V denote the subset of visible (observed) nodes.

$$V \cup H = \{X_1, X_2, \dots, X_n\}$$

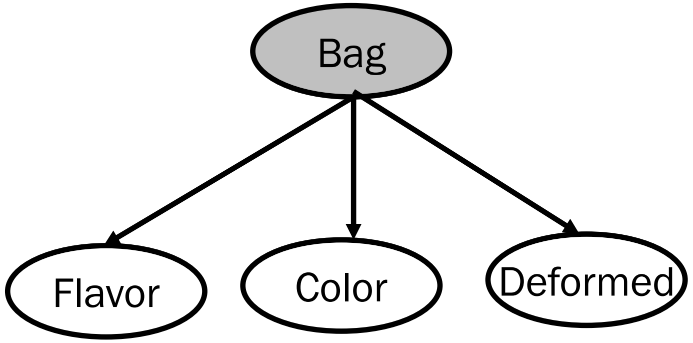
Goal: Estimate the CPTs in the BN to maximize the probability of the partially observed data

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

$$= \sum_{t=1}^{T} \log P\left(V = v^{(t)}\right) = \sum_{t=1}^{T} \log \sum_{h} P\left(V = v^{(t)}, H = h\right)$$

Expectation Maximization: Central Idea

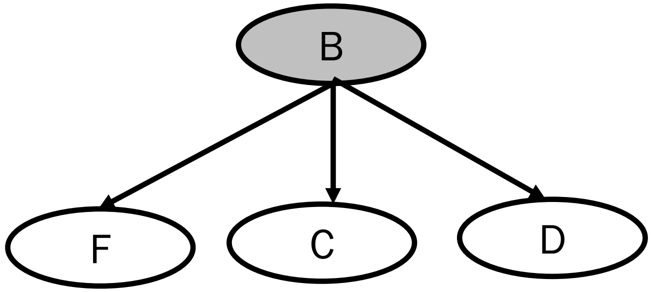
- Expectation Step (E-step): Compute *expected* values for missing data based on the observed data
- Maximization Step (M-step): Re-compute parameters using ML formula for fully observed data, using expected values as if they were observations



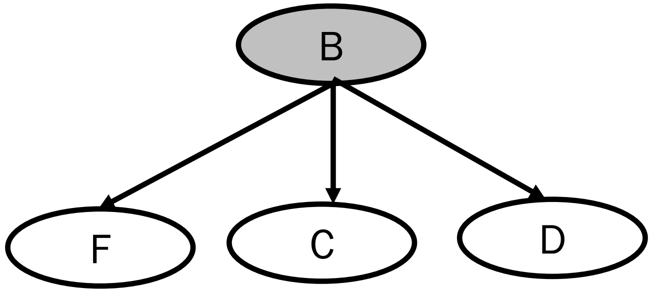
1. Guess the parameters of the model (really, just GUESS!)
2. Use those parameters to calculate expected counts of beans from each bag
3. Use those expected counts to update parameters
4. Lather, rinse, repeat!

Specifying the problem

What are the parameters of this model that we want to learn?



EM: Example

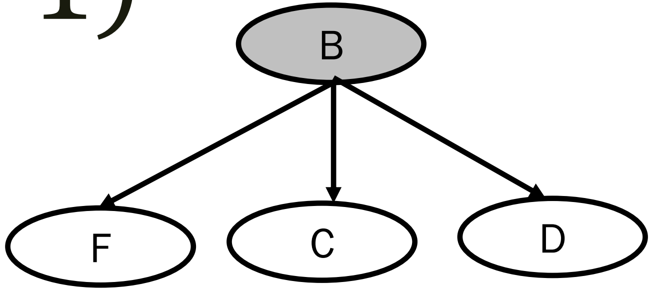


Guess (initialize) the parameters of the model
Must be valid probabilities, don't have to be right
For simplicity we might choose:

Initialization

$$P(B=1)=0.6$$
$$P(F=cherry|B=1)=P(C=red|B=1)=P(D=1|B=1)=0.6$$
$$P(F=cherry|B=2)=P(C=red|B=2)=P(D=1|B=2)=0.3$$

EM: Example: Updating $P(Bag = 1)$

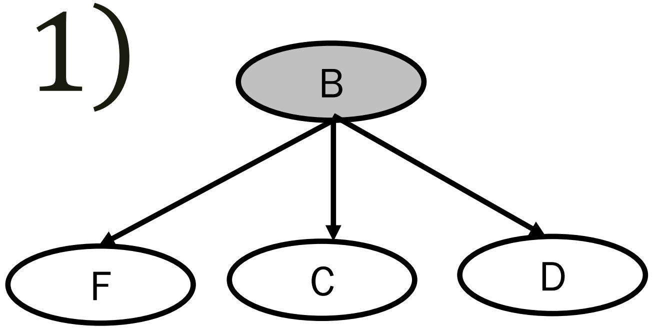


What would be the ML estimate if we knew which beans came from which bags?

But we don't know this, so we need an estimate for this value. Use the expectation!

E-step

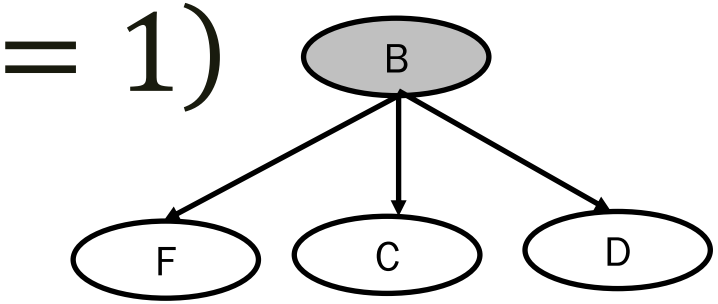
EM: Example: Updating $P(Bag = 1)$



$$\widehat{\text{count}}(B = 1) = \sum_{t=1}^{T} P\left(B = 1 \middle| F = f^{(t)}, C = c^{(t)}, D = d^{(t)}\right)$$

E-step

EM: Example: Updating $P(Bag = 1)$



E-step

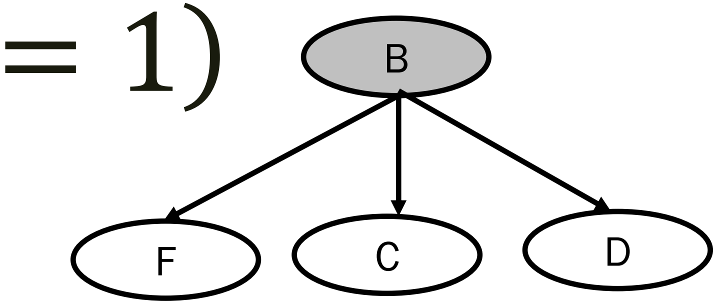
$$\widehat{\text{count}}(B = 1) = \sum_{t=1}^{T} P\left(B = 1 \middle| F = f^{(t)}, C = c^{(t)}, D = d^{(t)}\right)$$

$$= \sum_{t=1}^{T} \frac{P(B = 1)P\left(F = f^{(t)} \middle| B = 1\right)P\left(C = c^{(t)} \middle| B = 1\right)P\left(D = d^{(t)} \middle| B = 1\right)}{\sum_b P(B = b)P\left(F = f^{(t)} \middle| B = b\right)P\left(C = c^{(t)} \middle| B = b\right)P\left(D = d^{(t)} \middle| B = b\right)}$$

Where do we get these values?

A. We can't, we're stuck
B. From the last iteration
C. Directly from the data

EM: Example: Updating $P(Bag = 1)$



$$\widehat{\text{count}}(B = 1) = \sum_{t=1}^{T} P\left(B = 1 \middle| F = f^{(t)}, C = c^{(t)}, D = d^{(t)}\right)$$

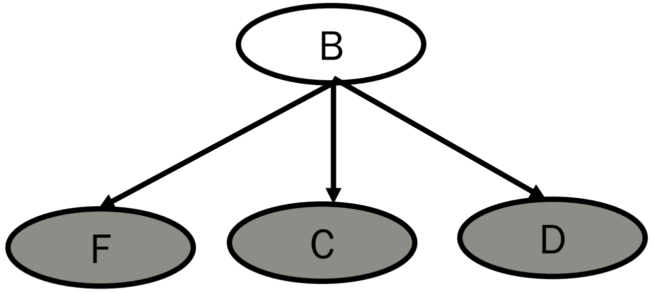
E-step

$$= \sum_{t=1}^{T} \frac{P(B = 1)P\left(F = f^{(t)} \middle| B = 1\right)P\left(C = c^{(t)} \middle| B = 1\right)P\left(D = d^{(t)} \middle| B = 1\right)}{\sum_b P(B = b)P\left(F = f^{(t)} \middle| B = b\right)P\left(C = c^{(t)} \middle| B = b\right)P\left(D = d^{(t)} \middle| B = b\right)}$$

$$P(Bag = 1) \leftarrow \frac{\widehat{\text{count}}(B = 1)}{T}$$

M-step

EM: Initialization

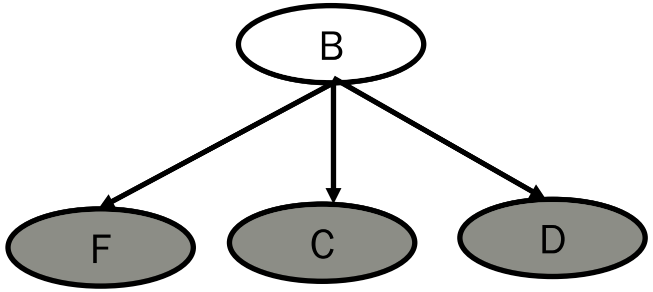


Param.	Iter 0	Iter 1	Iter 2	…
$P(B = 1)$				
$P(F = cherry \mid B = 1)$				
$P(F = cherry \mid B = 2)$				
$P(C = red \mid B = 1)$				
$P(C = red \mid B = 2)$				
$P(D = 1 \mid B = 1)$				
$P(D = 1 \mid B = 2)$				

EM algorithm (alternate expression):

```
Initialize the parameters of the model
While parameter values have not converged:
    For each parameter, p:
        • E-step: Use parameter values from last iteration to
          calculate expected counts necessary to update p
        • M-step: Use those expected counts to update p
```

EM: Calculate $P(B = 1)$ for iter 1



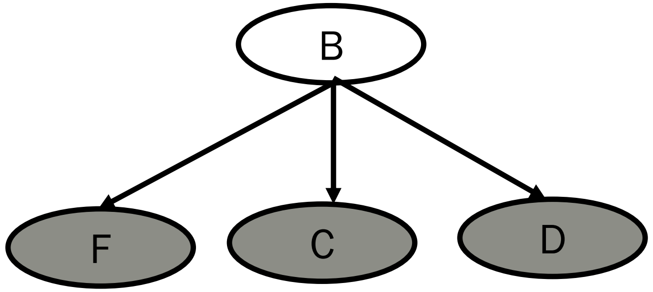
Param.	Iter 0	Iter 1	Iter 2	…
$P(B = 1)$	0.6			
$P(F = cherry\|B = 1)$	0.6			
$P(F = cherry\|B = 2)$	0.3			
$P(C = red\|B = 1)$	0.6			
$P(C = red\|B = 2)$	0.3			
$P(D = 1\|B = 1)$	0.6			
$P(D = 1\|B = 2)$	0.3			

```
EM algorithm (alternate expression):
Initialize the parameters of the model
While parameter values have not converged:
    For each parameter, p:
        • E-step: Use parameter values from last iteration to
          calculate expected counts necessary to update p
        • M-step: Use those expected counts to update p
```

$$\widehat{\text{count}}(B = 1) = \sum_{t=1}^{T} P\left(B = 1 \middle| F = f^{(t)}, C = c^{(t)}, D = d^{(t)}\right)$$

$$= \sum_{t=1}^{T} \frac{P(B = 1)P\left(F = f^{(t)} \middle| B = 1\right)P\left(C = c^{(t)} \middle| B = 1\right)P\left(D = d^{(t)} \middle| B = 1\right)}{\sum_b P(B = b)P\left(F = f^{(t)} \middle| B = b\right)P\left(C = c^{(t)} \middle| B = b\right)P\left(D = d^{(t)} \middle| B = b\right)}$$

EM: Calculate $P(B=1)$ for iter 1



Param.	Iter 0	Iter 1	Iter 2	…
$P(B=1)$	0.6			
$P(F=cherry\|B=1)$	0.6			
$P(F=cherry\|B=2)$	0.3			
$P(C=red\|B=1)$	0.6			
$P(C=red\|B=2)$	0.3			
$P(D=1\|B=1)$	0.6			
$P(D=1\|B=2)$	0.3			

<u>EM algorithm (alternate expression):</u>

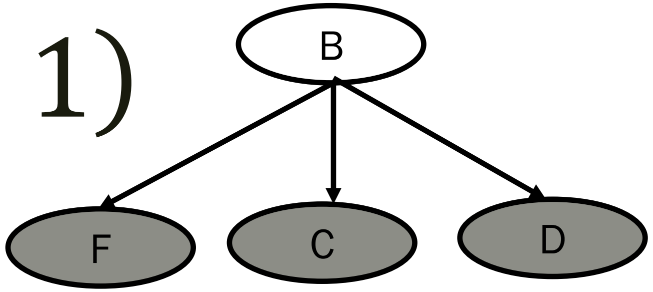
```
Initialize the parameters of the model
While parameter values have not converged:
    For each parameter, p:
        • E-step: Use parameter values from last iteration to
          calculate expected counts necessary to update p
        • M-step: Use those expected counts to update p
```

$$\widehat{\text{count}}(B=1) = \sum_{t=1}^{T} P\left(B=1 \middle| F=f^{(t)}, C=c^{(t)}, D=d^{(t)}\right)$$

$$= \sum_{t=1}^{T} \frac{P(B=1)P\left(F=f^{(t)}\middle|B=1\right)P\left(C=c^{(t)}\middle|B=1\right)P\left(D=d^{(t)}\middle|B=1\right)}{\sum_b P(B=b)P\left(F=f^{(t)}\middle|B=b\right)P\left(C=c^{(t)}\middle|B=b\right)P\left(D=d^{(t)}\middle|B=b\right)}$$

$$P(Bag=1)_{iter1} \leftarrow \frac{\widehat{\text{count}}(B=1)}{T}$$

EM: Example: Updating $P(Bag = 1)$



E-step + M-step

$$P(Bag = 1) \leftarrow \frac{1}{T}\sum_{t=1}^{T} \frac{P(B=1)P\left(F=f^{(t)}\middle|B=1\right)P\left(C=c^{(t)}\middle|B=1\right)P\left(D=d^{(t)}\middle|B=1\right)}{\sum_b P(B=b)P\left(F=f^{(t)}\middle|B=b\right)P\left(C=c^{(t)}\middle|B=b\right)P\left(D=d^{(t)}\middle|B=b\right)}$$

Current parameter values:
$P(B = 1) = 0.6$
$P(F = cherry|B = 1) = P(C = red|B = 1) = P(D = 1|B = 1) = 0.6$
$P(F = cherry|B = 2) = P(C = red|B = 2) = P(D = 1|B = 2) = 0.3$

What is the contribution from the red, deformed, cherry candies to this total?

	C=red		C=green	
	D=1	D=0	D=1	D=0
F=cherry	273	93	104	90
F=jalapeño	79	100	94	167

	C=red		C=green	
	D=1	D=0	D=1	D=0
F=cherry	273	93	104	90
F=jalapeño	79	100	94	167

$$P(Bag = 1)_{iter1} \leftarrow \frac{\widehat{\text{count}}(B = 1)}{T}$$

Param.	Iter 0
$P(B = 1)$	0.6
$P(F = cherry \mid B = 1)$	0.6
$P(F = cherry \mid B = 2)$	0.3
$P(C = red \mid B = 1)$	0.6
$P(C = red \mid B = 2)$	0.3
$P(D = 1 \mid B = 1)$	0.6
$P(D = 1 \mid B = 2)$	0.3

$$\widehat{\text{count}}(B = 1)$$
$$= \sum_{t=1}^{T} \frac{P(B = 1)P\left(F = f^{(t)} \middle| B = 1\right)P\left(C = c^{(t)} \middle| B = 1\right)P\left(D = d^{(t)} \middle| B = 1\right)}{\sum_b P(B = b)P\left(F = f^{(t)} \middle| B = b\right)P\left(C = c^{(t)} \middle| B = b\right)P\left(D = d^{(t)} \middle| B = b\right)}$$

F	C	D	contribution
cherry	red	1	
cherry	red	0	
cherry	green	1	
cherry	green	0	
jalapeno	red	1	
jalapeno	red	0	
jalapeno	green	1	
jalapeno	green	0	

	C=red		C=green	
	D=1	D=0	D=1	D=0
F=cherry	273	93	104	90
F=jalapeño	79	100	94	167

$$P(Bag=1)_{iter1} \leftarrow \frac{\widehat{\text{count}}(B=1)}{T}$$

Param.	Iter 0
$P(B=1)$	0.6
$P(F=cherry\|B=1)$	0.6
$P(F=cherry\|B=2)$	0.3
$P(C=red\|B=1)$	0.6
$P(C=red\|B=2)$	0.3
$P(D=1\|B=1)$	0.6
$P(D=1\|B=2)$	0.3

$$\widehat{\text{count}}(B=1)$$

$$= \sum_{t=1}^{T} \frac{P(B=1)P\left(F=f^{(t)} \middle| B=1\right)P\left(C=c^{(t)} \middle| B=1\right)P\left(D=d^{(t)} \middle| B=1\right)}{\sum_b P(B=b)P\left(F=f^{(t)} \middle| B=b\right)P\left(C=c^{(t)} \middle| B=b\right)P\left(D=d^{(t)} \middle| B=b\right)}$$

F	C	D	contribution
cherry	red	1	
cherry	red	0	
cherry	green	1	
cherry	green	0	
jalapeno	red	1	
jalapeno	red	0	
jalapeno	green	1	
jalapeno	green	0	

	C=red		C=green	
	D=1	D=0	D=1	D=0
F=cherry	273	93	104	90
F=jalapeño	79	100	94	167

$$P(Bag = 1)_{iter1} \leftarrow \frac{\widehat{\text{count}}(B = 1)}{T}$$

Param.	Iter 0
$P(B = 1)$	0.6
$P(F = cherry \| B = 1)$	0.6
$P(F = cherry \| B = 2)$	0.3
$P(C = red \| B = 1)$	0.6
$P(C = red \| B = 2)$	0.3
$P(D = 1 \| B = 1)$	0.6
$P(D = 1 \| B = 2)$	0.3

$$\widehat{\text{count}}(B = 1) = \sum_{t=1}^{T} \frac{P(B = 1)P\left(F = f^{(t)} \middle| B = 1\right)P\left(C = c^{(t)} \middle| B = 1\right)P\left(D = d^{(t)} \middle| B = 1\right)}{\sum_b P(B = b)P\left(F = f^{(t)} \middle| B = b\right)P\left(C = c^{(t)} \middle| B = b\right)P\left(D = d^{(t)} \middle| B = b\right)}$$

F	C	D	contribution
cherry	red	1	
cherry	red	0	
cherry	green	1	
cherry	green	0	
jalapeno	red	1	
jalapeno	red	0	
jalapeno	green	1	
jalapeno	green	0	

	C=red		C=green	
	D=1	D=0	D=1	D=0
F=cherry	273	93	104	90
F=jalapeño	79	100	94	167

$$P(Bag = 1)_{iter1} \leftarrow \frac{\widehat{\text{count}}(B = 1)}{T}$$

Param.	Iter 0
$P(B = 1)$	0.6
$P(F = cherry \mid B = 1)$	0.6
$P(F = cherry \mid B = 2)$	0.3
$P(C = red \mid B = 1)$	0.6
$P(C = red \mid B = 2)$	0.3
$P(D = 1 \mid B = 1)$	0.6
$P(D = 1 \mid B = 2)$	0.3

$$\widehat{\text{count}}(B = 1)$$

$$= \sum_{t=1}^{T} \frac{P(B = 1)P\left(F = f^{(t)} \middle| B = 1\right)P\left(C = c^{(t)} \middle| B = 1\right)P\left(D = d^{(t)} \middle| B = 1\right)}{\sum_b P(B = b)P\left(F = f^{(t)} \middle| B = b\right)P\left(C = c^{(t)} \middle| B = b\right)P\left(D = d^{(t)} \middle| B = b\right)}$$

F	C	D	contribution
cherry	red	1	
cherry	red	0	
cherry	green	1	
cherry	green	0	
jalapeno	red	1	
jalapeno	red	0	
jalapeno	green	1	
jalapeno	green	0	

	C=red		C=green	
	D=1	D=0	D=1	D=0
F=cherry	273	93	104	90
F=jalapeño	79	100	94	167

$$P(Bag = 1)_{iter1} \leftarrow \frac{\widehat{\text{count}}(B = 1)}{T}$$

Param.	Iter 0
$P(B = 1)$	0.6
$P(F = cherry \mid B = 1)$	0.6
$P(F = cherry \mid B = 2)$	0.3
$P(C = red \mid B = 1)$	0.6
$P(C = red \mid B = 2)$	0.3
$P(D = 1 \mid B = 1)$	0.6
$P(D = 1 \mid B = 2)$	0.3

$$\widehat{\text{count}}(B = 1)$$

$$= \sum_{t=1}^{T} \frac{P(B = 1)P\left(F = f^{(t)} \middle| B = 1\right)P\left(C = c^{(t)} \middle| B = 1\right)P\left(D = d^{(t)} \middle| B = 1\right)}{\sum_b P(B = b)P\left(F = f^{(t)} \middle| B = b\right)P\left(C = c^{(t)} \middle| B = b\right)P\left(D = d^{(t)} \middle| B = b\right)}$$

F	C	D	contribution
cherry	red	1	
cherry	red	0	
cherry	green	1	
cherry	green	0	
jalapeno	red	1	
jalapeno	red	0	
jalapeno	green	1	
jalapeno	green	0	

	C=red		C=green	
	D=1	D=0	D=1	D=0
F=cherry	273	93	104	90
F=jalapeño	79	100	94	167

$$P(Bag = 1)_{iter1} \leftarrow \frac{\widehat{\text{count}}(B = 1)}{T}$$

Param.	Iter 0
$P(B = 1)$	0.6
$P(F = cherry \vert B = 1)$	0.6
$P(F = cherry \vert B = 2)$	0.3
$P(C = red \vert B = 1)$	0.6
$P(C = red \vert B = 2)$	0.3
$P(D = 1 \vert B = 1)$	0.6
$P(D = 1 \vert B = 2)$	0.3

$$\widehat{\text{count}}(B = 1) = \sum_{t=1}^{T} \frac{P(B = 1)P\left(F = f^{(t)} \middle| B = 1\right)P\left(C = c^{(t)} \middle| B = 1\right)P\left(D = d^{(t)} \middle| B = 1\right)}{\sum_b P(B = b)P\left(F = f^{(t)} \middle| B = b\right)P\left(C = c^{(t)} \middle| B = b\right)P\left(D = d^{(t)} \middle| B = b\right)}$$

F	C	D	contribution
cherry	red	1	
cherry	red	0	
cherry	green	1	
cherry	green	0	
jalapeno	red	1	
jalapeno	red	0	
jalapeno	green	1	
jalapeno	green	0	

	C=red		C=green	
	D=1	D=0	D=1	D=0
F=cherry	273	93	104	90
F=jalapeño	79	100	94	167

$$P(Bag = 1)_{iter1} \leftarrow \frac{\widehat{\text{count}}(B = 1)}{T}$$

Param.	Iter 0
$P(B = 1)$	0.6
$P(F = cherry \mid B = 1)$	0.6
$P(F = cherry \mid B = 2)$	0.3
$P(C = red \mid B = 1)$	0.6
$P(C = red \mid B = 2)$	0.3
$P(D = 1 \mid B = 1)$	0.6
$P(D = 1 \mid B = 2)$	0.3

$$\widehat{\text{count}}(B = 1)$$
$$= \sum_{t=1}^{T} \frac{P(B = 1)P\left(F = f^{(t)} \middle| B = 1\right)P\left(C = c^{(t)} \middle| B = 1\right)P\left(D = d^{(t)} \middle| B = 1\right)}{\sum_b P(B = b)P\left(F = f^{(t)} \middle| B = b\right)P\left(C = c^{(t)} \middle| B = b\right)P\left(D = d^{(t)} \middle| B = b\right)}$$

F	C	D	contribution
cherry	red	1	
cherry	red	0	
cherry	green	1	
cherry	green	0	
jalapeno	red	1	
jalapeno	red	0	
jalapeno	green	1	
jalapeno	green	0	

	C=red		C=green	
	D=1	D=0	D=1	D=0
F=cherry	273	93	104	90
F=jalapeño	79	100	94	167

$$P(Bag = 1)_{iter1} \leftarrow \frac{\widehat{\text{count}}(B = 1)}{T}$$

Param.	Iter 0
$P(B = 1)$	0.6
$P(F = cherry \vert B = 1)$	0.6
$P(F = cherry \vert B = 2)$	0.3
$P(C = red \vert B = 1)$	0.6
$P(C = red \vert B = 2)$	0.3
$P(D = 1 \vert B = 1)$	0.6
$P(D = 1 \vert B = 2)$	0.3

$$\widehat{\text{count}}(B = 1) = \sum_{t=1}^{T} \frac{P(B = 1)P\left(F = f^{(t)} \middle| B = 1\right)P\left(C = c^{(t)} \middle| B = 1\right)P\left(D = d^{(t)} \middle| B = 1\right)}{\sum_b P(B = b)P\left(F = f^{(t)} \middle| B = b\right)P\left(C = c^{(t)} \middle| B = b\right)P\left(D = d^{(t)} \middle| B = b\right)}$$

F	C	D	contribution
cherry	red	1	
cherry	red	0	
cherry	green	1	
cherry	green	0	
jalapeno	red	1	
jalapeno	red	0	
jalapeno	green	1	
jalapeno	green	0	

What about the conditionals?

How do we calculate $P(F = cherry|B = 1)$

What is the ML formula for this parameter, if we had fully observed data?

What is the estimate for count(*F=cherry, B=1*)?

A. $\sum_{t=1}^{T} P\left(B = 1, F = cherry|F = f^{(t)}, C = c^{(t)}, D = d^{(t)}\right)$
B. $\sum_{t=1}^{T} P\left(B = 1|F = f^{(t)}, C = c^{(t)}, D = d^{(t)}\right)$
C. $\sum_{t=1}^{T} P\left(F = f^{(t)}|B = 1\right)$
D. $\sum_{t=1}^{T} P(F = cherry|B = 1)$
E. I have no idea

What about the conditionals?

How do we calculate $P(F = cherry|B = 1)$

$$\widehat{\text{count}}(B = 1, F = cherry) = \sum_{t=1}^{T} P\left(B = 1, F = cherry|F = f^{(t)}, C = c^{(t)}, D = d^{(t)}\right)$$

EM: General update formula for BN params

Nodes with parents:

$$P(X_i = x | Pa_i = \pi)_{iter\_i+1} = \frac{\widehat{\text{count}}(X_i = x, Pa_i = \pi)}{\widehat{\text{count}}(Pa_i = \pi)}$$

Root nodes:

$$P(X_i = x)_{iter\_i+1} = \frac{\widehat{\text{count}}(X_i = x)}{T}$$

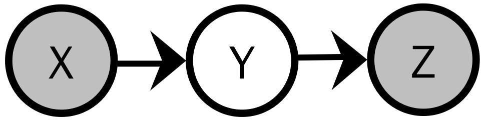
EM algorithm (alternate expression):

```
Initialize the parameters of the model
While parameter values have not converged:
    For each parameter, p:
        • E-step: Use parameter values from last iteration to
          calculate expected counts necessary to update p
        • M-step: Use those expected counts to update p
```

Properties of EM

- Monotonic convergence: Each iteration of EM increases (or does not change) log-likelihood of observed data.
- No tuning parameters, no learning rates, no backtracking
- Converges to a local or global maximum. Often depends on initialization values.

EM Practice on Second Simple Example



Which parameters of this network can you estimate directly from the data (in one step—no iteration required)?

A. $P(X)$

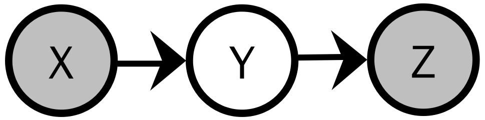
B. $P(Y|X)$

C. $P(Z|Y)$

D. Both A and C

E. None of them

EM Practice on Second Simple Example



$V = \{X, Z\}$ $H = \{Y\}$

Your turn! Express $P(Z = z|Y = y)$ in terms of $I\left(x, x^{(t)}\right)$, $I\left(z, z^{(t)}\right)$ and $P\left(Y = y\middle|X = x^{(t)}, Z = z^{(t)}\right)$

Consider a model with the following structure:

$$A \leftarrow B \rightarrow C$$

Suppose all of the variables are binary, and suppose only A and C are observable in the dataset. If we have T observations, then the dataset is of the form $\{(a_t, c_t)\}_{t=1}^{T}$.

(a) If we want to apply EM to this dataset, list all of the CPT entries that would need to be estimated. (Hint: there are 5 CPT entries in this model.)

Consider a model with the following structure:

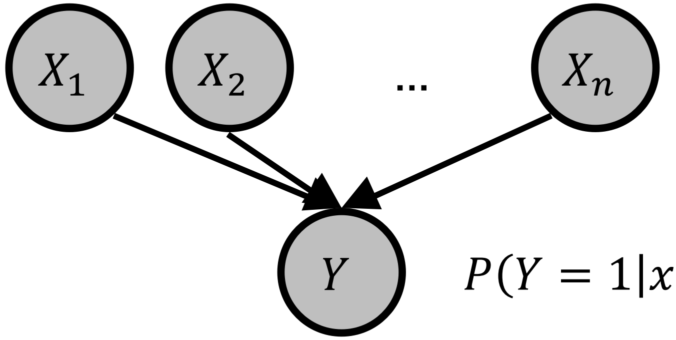
$$A \leftarrow B \rightarrow C$$

Suppose all of the variables are binary, and suppose only A and C are observable in the dataset. If we have T observations, then the dataset is of the form $\{(a_t, c_t)\}_{t=1}^{T}$.

(b) Find formulas for the EM update rules for the CPT entries listed in part (a). (Hint: to simplify the formulas, you can first express the EM updates in terms of equality-testing functions and the probability $P(B|A, C)$. Then, you can find a formula for $P(B|A, C)$ in terms of the CPT entries.)

EM for Learning Noisy-OR model

All variables are binary



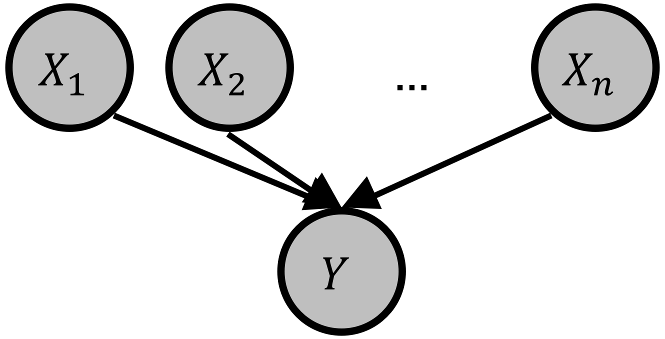
Complete the formula below:

$P(Y = 1 | x_1, x_2, \dots, x_n) =$

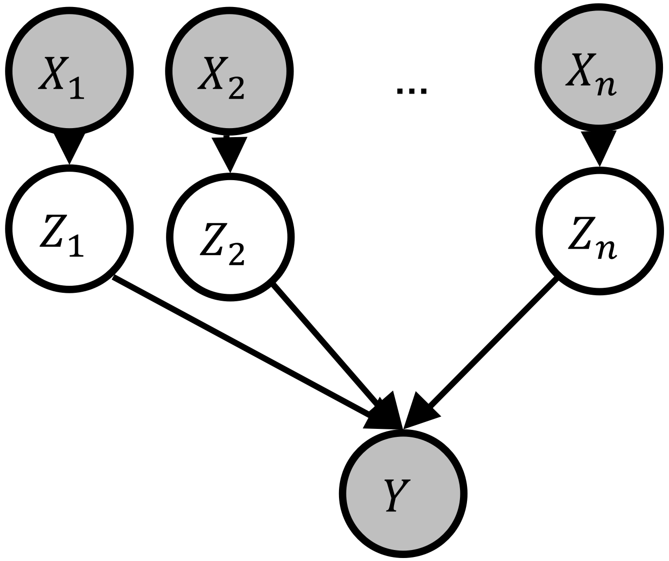
Problem: From *complete* data $\left\{\left({x_1}^{(t)}, {x_n}^{(t)}, \dots, {x_n}^{(t)}, y^{(t)}\right)\right\}_{t=1}^{T}$, estimate $p_i \in [0,1]$.

EM for Learning Noisy-OR: Alternate model

Problem: From (complete) data $\left\{\left({x_1}^{(t)}, {x_n}^{(t)}, \dots, {x_n}^{(t)}, y^{(t)}\right)\right\}_{t=1}^{T}$, estimate $p_i \in [0,1]$.



$$P(Y = 1|x_1, x_2, \dots, x_n) = 1 - \prod_{i=1}^{n}(1 - p_i)^{x_i}$$

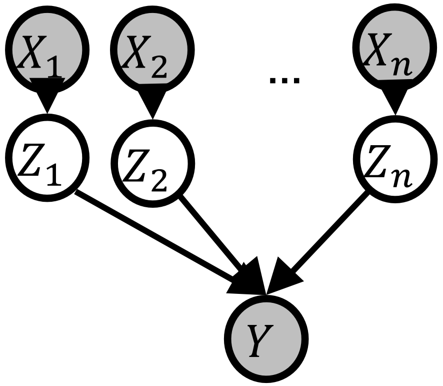


EM for Learning Noisy-OR: Alternate model

Problem: From (complete) data $\left\{\left({x_1}^{(t)}, {x_n}^{(t)}, \dots, {x_n}^{(t)}, y^{(t)}\right)\right\}_{t=1}^{T}$, estimate $p_i \in [0,1]$.

$P(Z_i = 0|X_i = 0) = 1$

$P(Z_i = 1|X_i = 1) = p_i$



$P(Y|\vec{z}) = OR(\vec{z})$

What is $P(Y = 1|x_1, x_2, \dots, x_n)$ in this alternate model?

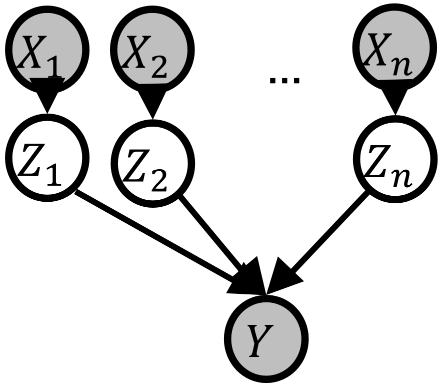
First show that $P(Y = 1|\vec{x}) = \sum_{\vec{z}} P(Y = 1|\vec{z})P(\vec{z}|\vec{x})$

EM for Learning Noisy-OR: Alternate model

Problem: From (complete) data $\left\{\left(x_1{}^{(t)}, x_n{}^{(t)}, \dots, x_n{}^{(t)}, y^{(t)}\right)\right\}_{t=1}^{T}$, estimate $p_i \in [0,1]$.

$P(Z_i = 0|X_i = 0) = 1$

$P(Z_i = 1|X_i = 1) = p_i$



$P(Y|\vec{z}) = OR(\vec{z})$

What is $P(Y = 1|x_1, x_2, \dots, x_n)$ in this alternate model?

$$P(Y = 1|\vec{x}) = \sum_{\vec{z}} \boxed{P(Y = 1|\vec{z})P(\vec{z}|\vec{x})}$$

When is the term in the red box 0?

A. When at least one z_i is 0

B. When all z_is are 0

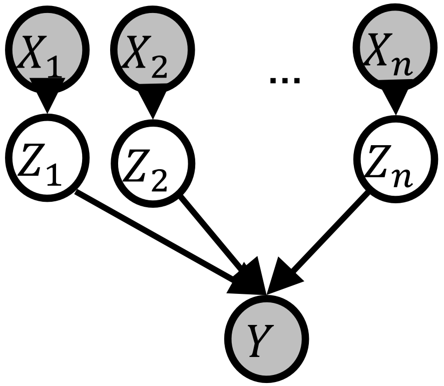
C. You can't tell from the information given

EM for Learning Noisy-OR: Alternate model

Problem: From (complete) data $\left\{\left(x_1{}^{(t)}, x_n{}^{(t)}, \ldots, x_n{}^{(t)}, y^{(t)}\right)\right\}_{t=1}^{T}$, estimate $p_i \in [0,1]$.

$P(Z_i = 0|X_i = 0) = 1$

$P(Z_i = 1|X_i = 1) = p_i$



$P(Y|\vec{z}) = OR(\vec{z})$

What is $P(Y = 1|x_1, x_2, \ldots, x_n)$ in this alternate model?

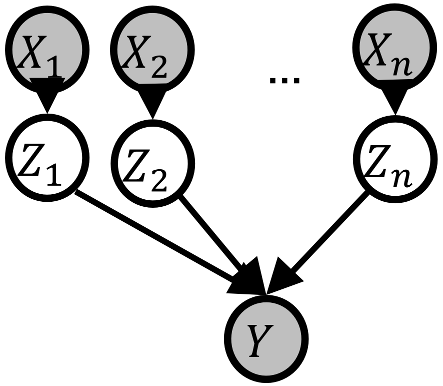
$$P(Y = 1|\vec{x}) = \sum_{\vec{z}} P(Y = 1|\vec{z}) P(\vec{z}|\vec{x})$$

EM for Learning Noisy-OR: Alternate model

Problem: From (complete) data $\left\{\left(x_1{}^{(t)}, x_n{}^{(t)}, \dots, x_n{}^{(t)}, y^{(t)}\right)\right\}_{t=1}^{T}$, estimate $p_i \in [0,1]$.

$P(Z_i = 0 | X_i = 0) = 1$

$P(Z_i = 1 | X_i = 1) = p_i$



$P(Y|\vec{z}) = OR(\vec{z})$

What is $P(Y = 1 | x_1, x_2, \dots, x_n)$ in this alternate model?

$$P(Y = 1|\vec{x}) = \sum_{\vec{z}} P(Y = 1|\vec{z}) P(\vec{z}|\vec{x}) = \sum_{\vec{z} \neq 0} P(\vec{z}|\vec{x}) =$$

A. $P(\vec{z} = 0|\vec{x})$

B. $1 - P(\vec{z} = 0|\vec{x})$

C. $P(\vec{z}|\vec{x} = 0)$

D. $1 - P(\vec{z}|\vec{x} = 0)$

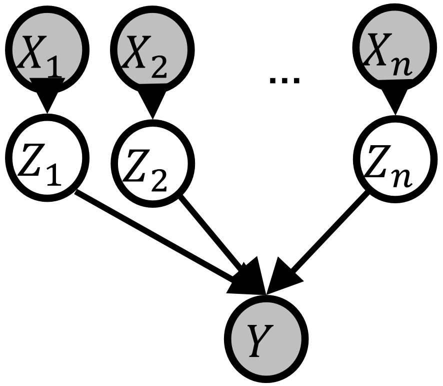
E. None of these

EM for Learning Noisy-OR: Alternate model

Problem: From (complete) data $\left\{\left({x_1}^{(t)}, {x_n}^{(t)}, \ldots, {x_n}^{(t)}, y^{(t)}\right)\right\}_{t=1}^{T}$, estimate $p_i \in [0,1]$.

$P(Z_i = 0|X_i = 0) = 1$

$P(Z_i = 1|X_i = 1) = p_i$



$P(Y|\vec{z}) = OR(\vec{z})$

What is $P(Y = 1|x_1, x_2, \ldots, x_n)$ in this alternate model?

$$P(Y = 1|\vec{x}) = \sum_{\vec{z}} P(Y = 1|\vec{z})P(\vec{z}|\vec{x}) = \sum_{\vec{z}\neq 0} P(\vec{z}|\vec{x}) = 1 - P(\vec{z} = 0|\vec{x})$$

Learning $P(Z_i = 1|X_i = 1) = p_i$ using EM

- If we had fully observed data:

Learning $P(Z_i = 1|X_i = 1) = p_i$ using EM

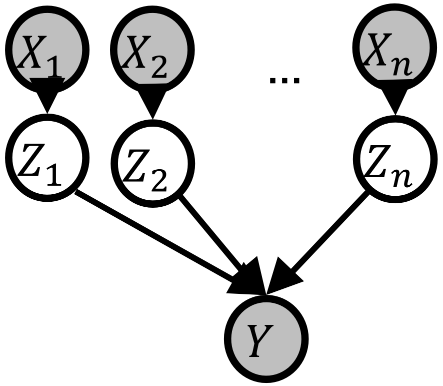
$$\widehat{count}(Z_i = 1, X_i = 1) = \sum_{t=1}^{T} P\left(Z_i = 1, X_i = 1 \middle| \vec{x}^{(t)}, y^{(t)}\right)$$

Learning $P(Z_i = 1|X_i = 1) = p_i$ using EM: Calculate $P\left(Z_i = 1\middle|\vec{x}^{(t)}, y^{(t)}\right)$ with current params

$P(Z_i = 0|X_i = 0) = 1$

$P(Z_i = 1|X_i = 1) = p_i$

$$\widehat{count}(Z_i = 1, X_i = 1) = \sum_{t=1}^{T} P\left(Z_i = 1, X_i = 1\middle|\vec{x}^{(t)}, y^{(t)}\right) = \sum_{t=1}^{T} x_i^{(t)} P\left(Z_i = 1\ \middle|\vec{x}^{(t)}, y^{(t)}\right)$$



$P(Y|\vec{z}) = OR(\vec{z})$

$$P\left(Z_i = 1\ \middle|\vec{x}^{(t)}, y^{(t)}\right) = \frac{P\left(Y = y^{(t)}\middle|Z_i = 1\right)P\left(Z_i = 1\middle|\vec{x}^{(t)}\right)}{P\left(Y = y^{(t)}\middle|\vec{x}^{(t)}\right)}$$

Which of the following is NOT a true statement?

A. $P\left(Y = y^{(t)}\middle|Z_i = 1\right) = 0$ when $y^{(t)} = 0$

B. $P\left(Y = y^{(t)}\middle|Z_i = 1\right) = 1$ when $y^{(t)} = 1$

C. $P\left(Z_i = 1\middle|\vec{x}^{(t)}\right) = 1$ when $x_i^{(t)} = 1$

D. $P\left(Z_i = 1\middle|\vec{x}^{(t)}\right) = 0$ when $x_i^{(t)} = 0$

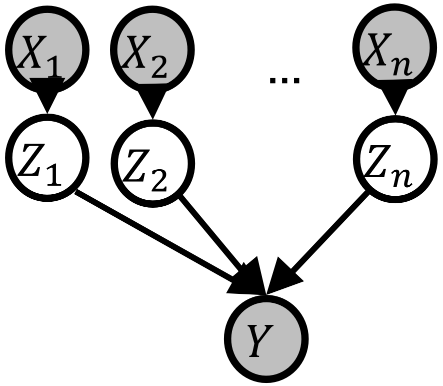
E. More than one of the above is NOT true

Learning $P(Z_i = 1|X_i = 1) = p_i$ using EM: Calculate $P\left(Z_i = 1\middle|\vec{x}^{(t)}, y^{(t)}\right)$ with current params

$P(Z_i = 0|X_i = 0) = 1$

$P(Z_i = 1|X_i = 1) = p_i$

$$\widehat{count}(Z_i = 1, X_i = 1) = \sum_{t=1}^{T} P\left(Z_i = 1, X_i = 1\middle|\vec{x}^{(t)}, y^{(t)}\right) = \sum_{t=1}^{T} x_i{}^{(t)} P\left(Z_i = 1\ \middle|\vec{x}^{(t)}, y^{(t)}\right)$$



$P(Y|\vec{z}) = OR(\vec{z})$

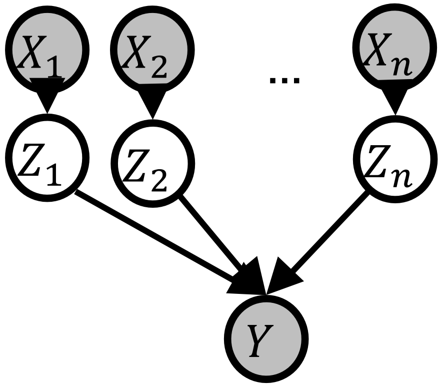
$$P\left(Z_i = 1\ \middle|\vec{x}^{(t)}, y^{(t)}\right) = \frac{P\left(Y = y^{(t)}\middle|Z_i = 1\right)P\left(Z_i = 1\middle|\vec{x}^{(t)}\right)}{P\left(Y = y^{(t)}\middle|\vec{x}^{(t)}\right)}$$

Learning $P(Z_i = 1|X_i = 1) = p_i$ using EM: Update rule

$P(Z_i = 0|X_i = 0) = 1$

$P(Z_i = 1|X_i = 1) = p_i$

$$\widehat{count}(Z_i = 1, X_i = 1) = \sum_{t=1}^{T}\left[\frac{p_i y^{(t)} x_i{}^{(t)}}{1 - \prod_{i=1}^{n}(1 - p_i)^{x_i{}^{(t)}}}\right]$$



$P(Y|\vec{z}) = OR(\vec{z})$

$$p_i \leftarrow \frac{\widehat{count}(Z_i = 1, X_i = 1)}{\widehat{count}(X_i = 1)} = \frac{p_i}{T_i}\sum_{t=1}^{T}\left[\frac{y^{(t)} x_i{}^{(t)}}{1 - \prod_{i=1}^{n}(1 - p_i)^{x_i{}^{(t)}}}\right]$$