# Bayesian Networks III: BN Parameter Learning

**For Self-learning Purposes**

**Jiashu Chen**

March 10, 2024

# Contents

# Chapter 1

# Maximum Likelihood Estimation

## 1.1 The Likelihood Function

- The likelihood function is the joint probability distribution of observed data D given that it is generated by a specified model with parameter $\theta$: $P(D|\theta)$
- It determines how likely the data to be generated by the model with parameter $\theta$.
- To emphasize the fact that the likeihood is a function of the parameters $\theta$, it is often denoted by L as $L(\theta|D)$ or $L(\theta : D)$
- The parameters that maximize the likelihood are known as the **maximum likelihood estimate**:

$$\theta^* = argmax_\theta L(\theta|D)$$

## 1.2 Coin Toss Example

- Suppose in Coin Toss, $P(H) = \theta$, $P(T) = 1 - \theta$
- 

| # | X |
|---|---|
| 1 | T |
| 2 | H |
| 3 | H |
| 4 | T |
| 5 | H |
| 6 | T |
| 7 | H |
| 8 | T |
| 9 | H |
| 10 | H |

- Assumption 1: The coin tosses are independent, given the parameter.
  - P(Tail appeared at the first instance) = $P(X[1] = T|\theta) = 1 - \theta$
  - P(Head appeared at the second instance) = $P(X[2] = H|\theta) = \theta$
  - P(Head appeared at the second instance, given that the tail appeared at the first instance) = $P(X[2] = H|X[1] = T, \theta) = P(X[2] = H|\theta) = \theta$

- Assumption 2: The coin tosses are identically distributed.
  - $P(X[2] = H|\theta) = P(X[3] = H|\theta) = P(X[5] = H|\theta) = \theta$
- Given the data, we could calculate the likelihood function

$$\begin{aligned} L(\theta|D) &= P(D|\theta) \\ &= (1-\theta)\theta\theta(1-\theta)\theta(1-\theta)\theta\theta \\ &= \theta^6(1-\theta)^4 \end{aligned} \tag{1.1}$$

- To obtain the parameter that maximizes the likehood

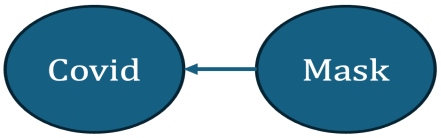
$$\begin{aligned} &\frac{\partial L}{\partial \theta} = 0 \\ &\rightarrow 6\theta^5(1-\theta)^4 - 4\theta^6(1-\theta)^3 \\ &= \theta^5(1-\theta)^3(6-10\theta) = 0 \\ &\rightarrow \theta^*_{MLE} = argmaxL(\theta : D) = 0.6 \\ &\Rightarrow \theta^*_{MLE} = \frac{N[H]}{N[H]+N[T]} = \frac{N[H]}{N} \end{aligned} \tag{1.2}$$

## 1.3 Parameterization of a Simple Bayesian Network

- How to learn the parameters in the COVID-MASK problem?
- First, parameterize the Bayesian network

$$P(M, C) = P(C|M)P(M)$$

| # | Covid | Mask |
|---|---|---|
| 1 | 1 | 0 |
| 2 | 0 | 1 |
| 3 | 1 | 0 |
| 4 | 0 | 1 |
| 5 | 0 | 1 |
| 6 | 0 | 1 |
| 7 | 1 | 0 |
| 8 | 0 | 1 |
| 9 | 0 | 1 |
| 10 | 0 | 0 |
| 11 | 0 | 1 |
| 12 | 1 | 1 |




| M\C | C0 | C1 |
|---|---|---|
| M0 | $\theta_{c0\|m0}$ | $\theta_{c1\|m0}$ |
| M1 | $\theta_{c0\|m1}$ | $\theta_{c1\|m1}$ |

| M | P(M) |
|---|---|
| M0 | $\theta_{m0}$ |
| M1 | $\theta_{m1}$ |

- In total, we have 6 parameters: $\theta_{m^0}, \theta_{m^1}, \theta_{c^0|m^0}, \theta_{c^1|m^0}, \theta_{c^0|m^1}, \theta_{c^1|m^1}$

$$\theta_{c^0|m^0} + \theta_{c^1|m^0} = 1 \tag{1.3}$$

$$\theta_{c^0|m^0} + \theta_{c^1|m^0} = 1 \tag{1.4}$$

$$\theta_{c^0|m^1} + \theta_{c^1|m^1} = 1 \tag{1.5}$$

- The data is D = {m(t), c(t) | t = 1, . . . , 12}
- We can use maximum likelihood estimation:

$$L(\theta|D) = P(D|\theta) = P(m(1), c(1), m(2), c(2), \ldots, m(12), c(12)|\theta)$$

- Independencies between instances imply

$$\begin{aligned}
L(\theta|D) &= \prod_{t=1}^{12} P(m(t), c(t)|\theta) \\
&= \prod_{t=1}^{12} P(m(t)|\theta)P(c(t)|m(t), \theta) \\
&= (\prod_{t=1}^{12} P(m(t)|\theta))(\prod_{t=1}^{12} P(c(t)|m(t), \theta)) \\
&= (\theta_{m^0}^{N[m^0]}\theta_{m^1}^{N[m^1]})(\theta_{c^0|m^0}^{N[c^0,m^0]}\theta_{c^1|m^0}^{N[c^0,m^0]}\theta_{c^0|m^1}^{N[c^0,m^1]}\theta_{c^1|m^1}^{N[c^1,m^1]}) \\
&= (\theta_{m^0}^{4}\theta_{m^1}^{8})(\theta_{c^0|m^0}^{1}\theta_{c^1|m^0}^{3}\theta_{c^0|m^1}^{7}\theta_{c^1|m^1}^{1}) \\
&= \theta_{m^0}^{4}(1-\theta_{m^0})^{8}\theta_{c^0|m^0}^{1}(1-\theta_{c^0|m^0})^{3}\theta_{c^0|m^1}^{7}(1-\theta_{c^0|m^1})^{1}
\end{aligned} \tag{1.6}$$

- Take derivative on each parameter

$$\begin{aligned}
&\frac{\partial L}{\partial \theta_{m^0}} = 0 \\
&\quad \rightarrow 4\theta_{m^0}^{3}(1-\theta_{m^0})^{8} - 8\theta_{m^0}^{4}(1-\theta_{m^0})^{7} \\
&\quad = \theta_{m^0}^{3}(1-\theta_{m^0})^{7}(4-12\theta_{m^0}) = 0 \\
&\quad \rightarrow \theta_{m^0}^{*} = \frac{4}{12} \\
&\quad \rightarrow \theta_{m^1}^{*} = \frac{8}{12}
\end{aligned} \tag{1.7}$$

$$\begin{aligned}
&\frac{\partial L}{\partial \theta_{c^0|m^0}} = 0 \\
&\quad \rightarrow (1-\theta_{c^0|m^0})^{3} - 3\theta_{c^0|m^0}(1-\theta_{c^0|m^0})^{2} \\
&\quad = (1-\theta_{c^0|m^0})^{2}(1-4\theta_{c^0|m^0}) = 0 \\
&\quad \rightarrow \theta_{c^0|m^0}^{*} = \frac{1}{4} \\
&\quad \rightarrow \theta_{c^1|m^0}^{*} = \frac{3}{4}
\end{aligned} \tag{1.8}$$

$$\begin{aligned}
&\frac{\partial L}{\partial \theta_{c^0|m^1}} = 0 \\
&\quad \rightarrow 7\theta_{c^0|m^1}^{6}(1-\theta_{c^0|m^1}) - \theta_{c^0|m^1}^{7} \\
&\quad = \theta_{c^0|m^1}^{6}(7-8\theta_{c^0|m^1}) = 0 \\
&\quad \rightarrow \theta_{c^0|m^1}^{*} = \frac{7}{8} \\
&\quad \rightarrow \theta_{c^1|m^1}^{*} = \frac{1}{8}
\end{aligned} \tag{1.9}$$

- More generally

$$\begin{aligned}
&\frac{\partial L}{\partial \theta_{m^0}} = 0 \\
&\quad \rightarrow N[m^0]\theta_{m^0}^{N[m^0]-1}(1-\theta_{m^0})^{N[m^1]} - N[m^1](1-\theta_{m^0})^{N[m^1]-1}\theta_{m^0}^{N[m^0]} = 0 \\
&\quad \rightarrow \theta_{m^0}^{N[m^0]-1}(1-\theta_{m^0})^{N[m^1]-1}(N[m^0] - (N[m^0]+N[m^1])\theta_{m^0}) = 0 \\
&\quad \rightarrow \theta^*_{m^0} = \frac{N[m^0]}{N[m^0]+N[m^1]} = \frac{N[m^0]}{N} \\
&\quad \rightarrow \theta^*_{m^1} = \frac{N[m^1]}{N[m^1]+N[m^1]} = \frac{N[m^1]}{N}
\end{aligned} \tag{1.10}$$

$$\begin{aligned}
&\frac{\partial L}{\partial \theta_{c^0|m^0}} = 0 \\
&\quad \rightarrow \theta^*_{c^0|m^0} = \frac{N[c^0,m^0]}{N[c^0,m^0]+N[c^1,m^0]} = \frac{N[c^0,m^0]}{N[m^0]} \\
&\quad \rightarrow \theta^*_{c^1|m^0} = \frac{N[c^1,m^0]}{N[c^0,m^0]+N[c^1,m^0]} = \frac{N[c^1,m^0]}{N[m^0]}
\end{aligned} \tag{1.11}$$

- Remark: The optimal parameters based on the MLE method are obtained by enumerating over the dataset.

## 1.4 Parameter Learning of Bayesian Networks

- Given a Bayesian network with structure G and joint distribution P over random variables X1, . . . , Xn.
- By parameter learning, we mean
  - Estimating the CPDs $P(X_i|Pax_i)$ known as the parameters and denoted by
    $$\theta_{x_i|pax_i} = P(x_i|pax_i)$$
  - By using dataset D = $\{\xi[1], \ldots, \xi[N]\}$, where $\xi[i]$ is instance i.
  - Then, using Bayesian networks, we could calculate joint probability distribution.
    $$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i|Pax_i)$$
    for arbitrary assignments $x_1, \ldots, x_n$.
- Parameterization of CPDs
  - Let $\boldsymbol{\theta}_{x_i|Pax_i}$ be the vector of all realizations $\theta_{x_i|pax_i}$:
    $$\boldsymbol{\theta}_{x_i|Pax_i} = (\theta_{x_i|Pax_i})_{x_i \in X_i}$$
    Examples of the right side: $\theta_{c^1|m^0}, \theta_{c^0|m^0}, etc$
  - And
    $$\boldsymbol{\theta} = (\boldsymbol{\theta}_{x_1|Pax_1}, \boldsymbol{\theta}_{x_2|Pax_2}, \ldots, \boldsymbol{\theta}_{x_n|Pax_n}$$

- The likelihood function then equals

$$\begin{aligned} L(\theta|D) &= P(D|\theta) \\ &= \prod_{t=1}^{N}\prod_{i=1}^{n} P(x_i[t]|pax_i[t],\boldsymbol{\theta}) \\ &= \prod_{i=1}^{n}[\prod_{t=1}^{N} P(x_i[t]|pax_i[t],\boldsymbol{\theta})] \\ &= \prod_{i=1}^{n}[\prod_{t=1}^{N} P(x_i[t]|pax_i[t],\boldsymbol{\theta}_{x_i|pax_i})] \end{aligned} \tag{1.12}$$

  N is the number of instance, n is the number of nodes

- Define the local likelihood function

$$\begin{aligned} L_i(\theta_{x_i|pax_i} : D) &= \prod_{t=1}^{N} P(x_i[t]|pax_i[t],\boldsymbol{\theta}_{x_i|pax_i}) \\ &= \prod_{pax_i}\prod_{x_i} \theta_{x_i|pax_i}^{N[x_i,pax]} \end{aligned} \tag{1.13}$$

- Then

$$L(\boldsymbol{\theta}|D) = \prod_{i=1}^{n} L_i(\boldsymbol{\theta}_{x_i|pax_i}|D) \tag{1.14}$$

- This **global decomposition** allows MLE to be obtained by maximizing the local likelihoods independently

$$\frac{\partial L}{\partial \boldsymbol{\theta}_{x_i|pax_i}} = 0 \rightarrow \frac{\partial L_i}{\partial \boldsymbol{\theta}_{x_i|pax_i}} = 0 \tag{1.15}$$

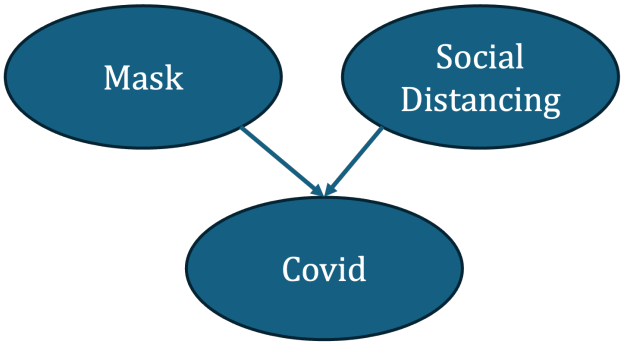
- Under the assumption $N[pax_i] \neq 0$, this results in:

  The MLE of CPD parameters $\theta_{x_i|pax_i}$ is obtained by $\theta^*_{x_i|pax_i} = \frac{N[x_i,pax_i]}{N[pax_i]}$

  Extremely simple, just count!

## 1.5 Practice: COVID Example

- For example, how to find the CPD corresponding to COVID, when the person wears a mask but does not keep social distancing, $\boldsymbol{\theta}_{C|m^1,d^0}$




| # | C | M | D |
|---|---|---|---|
| 1 | 1 | 1 | 0 |
| 2 | 0 | 1 | 0 |
| 3 | 1 | 1 | 1 |
| 4 | 0 | 1 | 1 |
| 5 | 0 | 1 | 0 |
| 6 | 0 | 1 | 1 |
| 7 | 1 | 1 | 1 |
| 8 | 0 | 1 | 0 |
| 9 | 0 | 1 | 1 |
| 10 | 0 | 1 | 0 |
| 11 | 0 | 1 | 1 |
| 12 | 1 | 1 | 0 |

- 
- Suppose the data includes a total of 12 instances with M = 1, 6 of which have D = 0, 2 of which have C = 1

$$\theta^*_{c^1|m^1,d^0} = \frac{N[m^1,d^0]}{N[c^1,m^1,d^0]} = \frac{2}{6} \tag{1.16}$$

$$\theta^*_{c^0|m^1,d^0} = 1 - \theta_{c^1|m^1,d^0} = \frac{4}{6} \tag{1.17}$$

- A special case: what if there was no individual reported in the dataset who wore a mask but did not keep social distancing? $N[c^1,m^1,d^1] = N[m^1,d^0] = 0$
- We may no longer use

$$\theta^*_{c^1|m^1,d^0} = \frac{N[m^1,d^0]}{N[c^1,m^1,d^0]} \tag{1.18}$$

- Here, contracting or not contracting COVID is equally probable when M = 1 and D = 0, because no information is given:

$$\theta^*_{c^1|m^1,d^0} = \theta^*_{c^0|m^1,d^0} = 0.5 \tag{1.19}$$

# Chapter 2

# Missing Values: The Effect on the Likelihood Function

## 2.1 How Does the Likelihood Function Change?

- In the simple COVID-Mask case, for complete data we had

$$L(\theta, \theta_{c^1|m^0}, \theta_{c^1|m^1}|D) = (1-\theta_{m^1})^4 \theta_{m^1}^8 (1-\theta_{c^1|m^0})^1 \theta_{c^1|m^0}^3 (1-\theta_{c^1|m^1})^7 \theta_{c^1|m^1}^1 \tag{2.1}$$

| # | C | M |
|---|---|---|
| 1 | 1 | ? |
| 2 | 0 | 1 |
| 3 | 1 | 1 |
| 4 | 0 | 1 |
| 5 | 0 | 1 |
| 6 | 0 | 1 |
| 7 | 1 | 1 |
| 8 | 0 | 1 |
| 9 | 0 | 1 |
| 10 | 0 | 1 |
| 11 | 0 | 1 |
| 12 | 1 | 1 |

- If instead of C[1] = 1, M[1] = 0, we had C[1] = 1, M[1] = ?, then
  - If M[1] = 0, the likelihood L does not change;
  - If M[1] = 1, the likelihood becomes:

$$L^{'} = (1-\theta_{m^1})^3 \theta_{m^1}^9 (1-\theta_{c^1|m^0})^1 \theta_{c^1|m^0}^2 (1-\theta_{c^1|m^1})^7 \theta_{c^1|m^1}^2 \tag{2.2}$$

- To obtain the likelihood under the missing value, we need to marginalize over M[1], that is, to sum two likelihood

$$\begin{aligned} L_{missing} &= L + L^{'} \\ &= (1-\theta_{m^1})^3 \theta_{m^1}^8 \\ &* (1-\theta_{c^1|m^0})^1 \theta_{c^1|m^0}^2 \\ &* (1-\theta_{c^1|m^1})^7 \theta_{c^1|m^1}^1 \\ &* ((1-\theta_{m^1})\theta_{c^1|m^0} + \theta_{m^1}\theta_{c^1|m^1}) \end{aligned} \tag{2.3}$$

  This is not decomposable!

## 2.2 Likelihood: Partially Observed Data

- Let O[t] and H[t] be the observed and missing variables at instance t
- Proposition:

If the data satisfies the iid assumption, the likelihood becomes

$$L(\boldsymbol{\theta}|D) = \prod_{t=1}^{N} P(\boldsymbol{o}[t]|\boldsymbol{\theta}) = \prod_{t=1}^{N} \sum_{h_t} P(\boldsymbol{o}[t], \boldsymbol{h}[t]|\theta) \tag{2.4}$$

- The sum makes the likelihood function multi-modal, **Unimodality** of the likelihood function is lost.
- The number of possible assignments h[t] in the sum is exponential in the number of missing values at instance t.
- Decomposibility of the likelihoods for different parameters is lost.
- All of these make it computationally challenging to find the maximum of the likelihood (MLE)
- We should use other approaches for parameter learning in the presence of partially observed data.

# Chapter 3

# Missing Values: Missing Completely at Random

## 3.1 How does the Likelihood Function Change?

- Can we simply ignore the instance with the missing value and obtain the decomposable likelihood as before?
- In general, no. Because ignoring the whole instance means to also ignore info on COVID: C = 1
  - This makes C = 1 less likely in our model than in the dataset.
  - So we may obtain a lower estimate of the true value of $\theta_{c^1|M}$
- What if we did not want to estimate $\theta_{c^1|M}$, but just $\theta_{m^1}$?
-

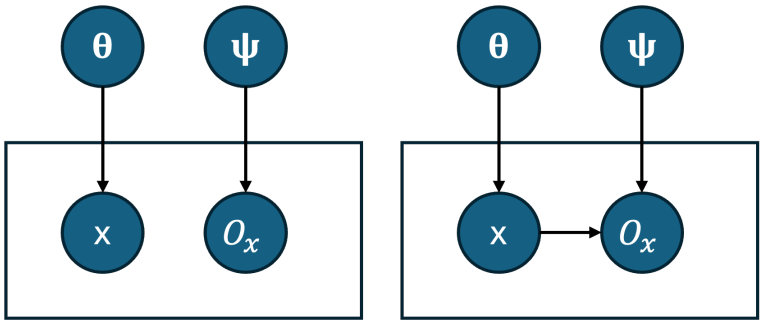
| # | C | M |
|---|---|---|
| 1 | 1 | ? |
| 2 | 0 | 1 |
| 3 | 1 | 1 |
| 4 | 0 | 1 |
| 5 | 0 | 1 |
| 6 | 0 | 1 |
| 7 | 1 | 1 |
| 8 | 0 | 1 |
| 9 | 0 | 1 |
| 10 | 0 | 1 |
| 11 | 0 | 1 |
| 12 | 1 | 1 |

## 3.2 Missing Value: Random or Deliberate

### Observability variable $O_x$

- $O_x = o^1$ means that the value of X is observed.
- $O_x = o^0$ means that the value of X is missing
- $O_x \perp X$: Random missing value

- $O_x \not\perp X$: Deliberate missing value
- $\psi = P(O_x = o^1)$ denotes the probability of observing x.




Random missing value  Deliberate missing value

## 3.3 Likelihood for Random Missing Values

- D = T, H, ?, T, H, T, H, ?, H, H
- ? becomes a legitimate value
- Let Y be a deterministic function of X and $O_x$ defined as

$$D_{it} = \begin{cases} X & O_x = o^1 \\ ? & O_x = o^0 \end{cases} \tag{3.1}$$

- For random missing values:

$$\begin{aligned} P(Y = H) = P(X = H, O_x = o^1) = P(X = H)P(O_x = o^1) = \theta\psi \\ P(Y = T) = P(X = T, O_x = o^0) = P(X = H)P(O_x = o^0) = (1-\theta)\psi \\ P(Y = ?) = P(O_x = o^0) = (1-\psi) \end{aligned} \tag{3.2}$$

- So the extended likelihood that includes the missing values equals

$$\begin{aligned} L_{missing}(\theta, \psi | D) &= (\theta\psi)^{N[H]}((1-\theta)\psi)^{N[T]}(1-\psi)^{N[?]} \\ &= \theta^{N[H]}(1-\theta)^{N[T]}\psi^{N[H]+N[T]}(1-\psi)^{N[?]} \\ &= \theta^5(1-\theta)^3\psi^8(1-\psi)^2 \end{aligned} \tag{3.3}$$

  Which is decomposable, the previous idea used for complete dataset is still valid.

- Decomposibility of the Likelihood
  - Consider

$$\begin{aligned} \text{D} &= \{\text{T, H, ?, T, H, T, H, ?, H, H}\} \\ \hat{\text{D}} &= \{o^1, o^1, o^0, o^1, o^1, o^1, o^1, o^0, o^1, o^1\} \end{aligned}$$

  - By omitting missing values from D, we have

$$\bar{D} = \{\text{T, H, T, H, T, H, H, H}\}$$

$$\begin{aligned} L_{missing}(\theta, \psi | D) &= L(\theta | \bar{D}) L(\psi | \hat{\text{D}}) \\ &= \theta^{N[H]}(1-\theta)^{N[T]}\psi^{N[H]+H[T]}(1-\psi)^{N[?]} \end{aligned} \tag{3.4}$$

where the likelihood of the observed data: $L(\theta|\hat{\mathrm{D}}) = \theta^{N[H]}(1-N[H])^{N[T]}$
likelihood of the observability of the data: $L(\psi|\bar{D}) = \psi^{N[H]+N[T]}(1-\psi)^{N[?]}$

- N[H] = 5, N[T] = 3, N[?] = 2

$$\begin{aligned} &\text{Probability of a head} \\ L(\theta|\bar{D}) = \Theta^*_{MLE} &= \frac{N[H]}{N[H]+N[T]} = \frac{5}{8} \\ &\text{Probability of observing an instance} \\ L(\psi|\bar{D}) = \psi^*_{MLE} &= \frac{N[H]+N[T]}{N[H]+N[T]+N[?]} = \frac{8}{10} \end{aligned} \tag{3.5}$$

One of the approach for learning with random missing values is **to ignore the missing values!**

## 3.4 The Observation Mechanism

- Why did it work? rather than marginalizing out the missing values (?), we accepted them as a valid value for X.
- Definition: Consider the set of random variables $X = X_1, \ldots, X_n$ and the set of their observability $O_X = O_{X_1}, \ldots, O_{X_n}$
  - The missing data model (observability model) is a joint distribution

$$P_{missing}(X, O_x) = P_{missing}(O_x|\boldsymbol{X})P(\boldsymbol{X}) \tag{3.6}$$

  - The actual observation is Y = $Y_1, \ldots, Y_n$, where

$$Y_i = \begin{cases} X_i & O_{x_i} = o^1 \\ ? & O_{x_i} = o^0 \end{cases} \tag{3.7}$$

  - Correspondingly, X can be partitioned into
    - observed variables: $\boldsymbol{X^y_{obs}} = X_i|y_i \neq ?$
    - hidden variables: $\boldsymbol{X^y_{hidden}} = X_i|y_i = ?$

## 3.5 The (extended) Likelihood Function under Missing Values

- Note that $P_{missing}(\boldsymbol{Y}) = P_{missing}(\boldsymbol{X_{obs}}, O_{\boldsymbol{X}})$
- Hence, under the iid assumption, the likelihood function equals

$$\begin{aligned} L_{missing}(\theta, \psi|D) &= \prod_{t=1}^{N} P_{missing}(\boldsymbol{y}_{[}\boldsymbol{t}]|\boldsymbol{\theta}, \boldsymbol{\psi}) \\ &= \prod_{t=1}^{N} P_{missing}(\boldsymbol{x_{obs}}[\boldsymbol{t}], \boldsymbol{o_x}[\boldsymbol{t}]|\boldsymbol{\theta}, \boldsymbol{\psi}) \end{aligned} \tag{3.8}$$

- Definition: **MCAR**: The missing data model is missing completely at random (MCAR) if

$$P_{missing} \models (O_{\boldsymbol{X}} \perp \boldsymbol{X}) \tag{3.9}$$

## 3.6 MCAR: Decoupling the Likelihood

### Decoupling the Likelihood

- Under MCAR, the likelihood of $\boldsymbol{X}$ and $O_X$ decomposes as

$$P_{missing}(\boldsymbol{X}, O_{\boldsymbol{X}}) = P_{missing}(O_{\boldsymbol{X}})P(X) \tag{3.10}$$

- Hence,

$$\begin{aligned} L_{missing}(\theta, \psi|D) &= \prod_{t=1}^{N} P_{missing}(\boldsymbol{y}_{[}\boldsymbol{t}]|\boldsymbol{\theta}, \boldsymbol{\psi}) \\ &= \prod_{t=1}^{N} P_{missing}(\boldsymbol{x_{obs}}[\boldsymbol{t}], \boldsymbol{o_x}[\boldsymbol{t}]|\boldsymbol{\theta}, \boldsymbol{\psi}) \\ &= \prod_{t=1}^{N} P_{missing}(\boldsymbol{x_{obs}}[\boldsymbol{t}]|\boldsymbol{\theta}, \boldsymbol{\psi}) P_{missing}(\boldsymbol{o_x}[\boldsymbol{t}]|\boldsymbol{\theta}, \boldsymbol{\psi}) \\ &= \prod_{t=1}^{N} P_{missing}(\boldsymbol{x_{obs}}[\boldsymbol{t}]|\boldsymbol{\theta}) P_{missing}(\boldsymbol{o_x}[\boldsymbol{t}]|\boldsymbol{\psi}) \\ &= \prod_{t=1}^{N} P_{missing}(\boldsymbol{x_{obs}}[\boldsymbol{t}]|\boldsymbol{\theta}) \prod_{t=1}^{N} P_{missing}(\boldsymbol{o_x}[\boldsymbol{t}]|\boldsymbol{\psi}) \end{aligned} \tag{3.11}$$

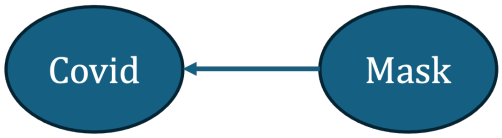
- Each part could be maximized separately.
- Therefore, we could ignore the mechanism that generates the missing value and maximize the first term alone. $\prod_{t=1}^{N} P_{missing}(\boldsymbol{x_{obs}}[\boldsymbol{t}]|\boldsymbol{\theta})$
- In the previous Covid example in section 3.1, we implicitly ignored the observation mechanism. We assume that missing value is independent with wearing masks. Otherwise, we cannot assume that the missing value being one (M = 1) happens with the probability P(M = 1) = $\theta_{m^1}$. So even to get the non-decomposible form, we need an assumption such as MCAR to ignore the observation mechanism.

### Partial Likelihood Decomposition

- To maximize the first term, $\prod_{t=1}^{N} P_{missing}(\boldsymbol{x_{obs}}[\boldsymbol{t}]|\boldsymbol{\theta})$, we cannot simply ignore all missing values, because the first term does not decompose according to the BN structure.(need to include all values, not just observed values)

$$\begin{aligned} \prod_{t=1}^{N} P_{missing}(\boldsymbol{x_{obs}}[\boldsymbol{t}]|\boldsymbol{\theta}) &= \prod_{t=1}^{N} \sum_{x_{hidden}} P_{missing}(\boldsymbol{x_{obs}}[\boldsymbol{t}], \boldsymbol{x_{hidden}}|\boldsymbol{\theta}) \\ &= \prod_{t=1}^{N} \sum_{x_{hidden}} \prod_{i=1}^{n} P_{missing}(x_i[t]|pa_{x_i}[t], \theta) \end{aligned} \tag{3.12}$$

- Covid Example: if the missing value is Mask



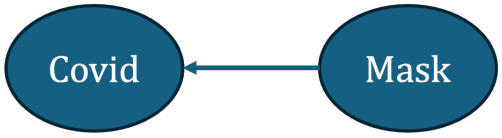

| # | C | M |
|---|---|---|
| 1 | 1 | ? |
| 2 | 0 | 1 |
| 3 | 1 | 1 |
| 4 | 0 | 1 |
| 5 | 0 | 1 |
| 6 | 0 | 1 |
| 7 | 1 | 1 |
| 8 | 0 | 1 |
| 9 | 0 | 1 |
| 10 | 0 | 1 |
| 11 | 0 | 1 |
| 12 | 1 | 1 |

- 
  - We cannot simply ignore the instance with missing value and decompose likelihood.
  - For $\theta_{c^1|M}$, ignoring the whole instance means to ignore one instance C = 1. So, we might underestimate $\theta_{c^1|M}$.
  - For $\theta_{m^1}$, we still cannot ignore the instance. The data on covid could partly determine the missing value.

$$\begin{aligned} L_{missing} = L + L' = (1-\theta_{m^1})^3\theta^8_{m^1} \\ * (1-\theta_{c^1|m^0})^1\theta^2_{c^1|m^0} \\ * (1-\theta_{c^1|m^1})^7\theta^1_{c^1|m^1} \\ * ((1-\theta_{m^1})\theta_{c^1|m^0} + \theta_{m^1}\theta c^1|m^1) \end{aligned} \tag{3.13}$$

  In the equation, $\theta_{m^1}$ cannot be factorized from the likelihood. This is because mask is the parent of Covid.

- Covid Example: if the missing value is Covid




| # | C | M |
|---|---|---|
| 1 | ? | 1 |
| 2 | 0 | 1 |
| 3 | 1 | 1 |
| 4 | 0 | 1 |
| 5 | 0 | 1 |
| 6 | 0 | 1 |
| 7 | 1 | 1 |
| 8 | 0 | 1 |
| 9 | 0 | 1 |
| 10 | 0 | 1 |
| 11 | 0 | 1 |
| 12 | 1 | 1 |

- 

  - For $P(M = m^1)$, which is $\theta_{m^1}$.

$$P(?|m^1)P(m^1) = P(c^1|m^1)P(m^1) + P(c^0|m^1)P(m^1) = P(m^1) = \theta_{m^1} \tag{3.14}$$

  - For $\theta_{C|M}$, we could ignore the whole instance.

$$\begin{aligned} L_{missing} = L + L' = (1 - \theta_{m^1})^3 \theta_{m^1}^8 \\ * (1 - \theta_{c^1|m^0})^1 \theta_{c^1|m^0}^2 \\ * (1 - \theta_{c^1|m^1})^7 \theta_{c^1|m^1}^1 \\ * \theta_{m^1} \end{aligned} \tag{3.15}$$

    In the equation, $\theta_{m^1}$ is decomposable

- Therefore,

$$\begin{aligned} \prod_{t=1}^{N} P_{missing}(\boldsymbol{x_{obs}[t]}|\boldsymbol{\theta}) &= \prod_{t=1}^{N} \sum_{x_{hidden}} P_{missing}(\boldsymbol{x_{obs}[t]}, \boldsymbol{x_{hidden}}|\boldsymbol{\theta}) \\ &= \prod_{t=1}^{N} \sum_{x_{hidden}} \prod_{i=1}^{n} P_{missing}(x_i[t]|pa_{x_i}[t], \theta) \\ &= \prod_{t=1}^{N} \sum_{x_{hidden}} \theta_{x_i[t]|pa_{x_i}[t]} \end{aligned} \tag{3.16}$$

  - For parameter $\theta_{x_i[t]|pa_{x_i}[t]}$, if neither $X_i$ nor any of its parents have a missing value in the dataset, then it can be taken out of the summation and optimized separately as before.
  - If the value of only $X_i$ (not its parents) is missing at some instances, and $X_i$ has no child, then we can simply ignore those instances and optimize $\theta_{x_i[t]|pa_{x_i}[t]}$ as before.
  - Otherwise, local likelihood is not decomposable.

## Proposition: MLE under MCAR

Given the variable $X_i$, assume that $N[pa_{x_i}] \neq 0$. Under MCAR, if at least one of the following condition hold:

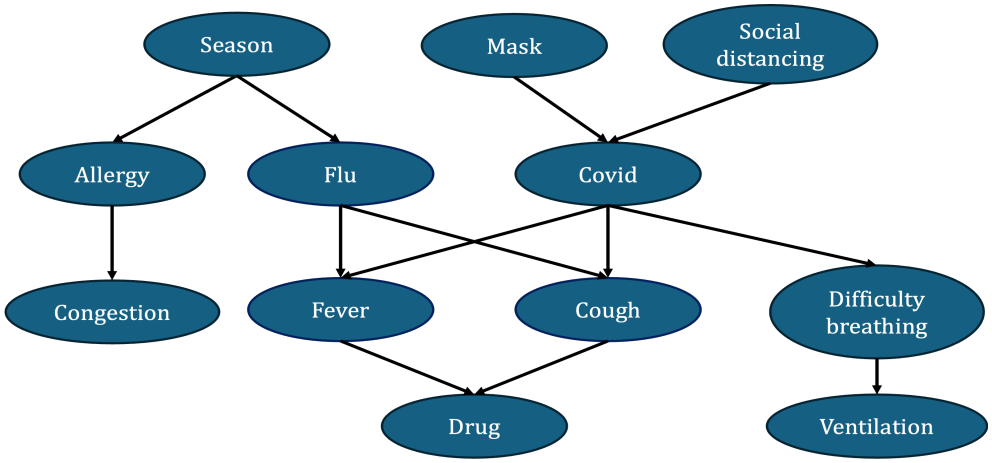
- Both $X_i$ and its parents are always observed in the whole dataset, or
- All parents of $X_i$ are observed in the whole dataset and $X_i$ does not have a child.

then the MLE of the CPD parameters $\theta_{X_i|pa_{x_i}}$ is obtained by

$$\theta^*_{x_i|pa_{x_i}} = \frac{N[x_i, pa_{x_i}]}{N[pa_{x_i}]} \tag{3.17}$$

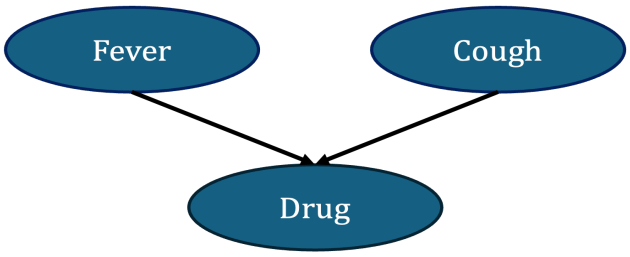
- Under the above condition, we can ignore the missing values of the variables other than $X_i$ and its parents.
- Unlike many other models, no imputation is needed to fill in the missing values in the case!

## 3.7 Covid Example




| # | Covid | Mask | Social Distancing | FlU | COugh | Fever | Ventilation | Season | ConGestion | Difficulty Breathing | DRug | Allergy |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | True | False | True | False | True | True | True | Spring | True | True | False | False |
| 2 | False | ? | ? | False | False | True | False | Summer | ? | False | True | False |
| 3 | True | False | False | True | True | False | False | Fall | False | True | ? | False |
| 4 | False | True | True | False | False | True | False | Winter | True | True | False | True |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 1000 | ? | True | ? | False | True | False | True | Spring | ? | False | True | True |

- Suppose the data is **missing completely at random**.
- Case 1: how to find the CPD corresponding to drug, when the person coughs but does not have fever? i.e. $\theta_{R|o^1,f^0}$?




| # | R | O | F |
|---|---|---|---|
| 1 | 1 | 1 | 0 |
| 3 | 1 | 1 | 1 |
| 5 | 0 | 1 | 0 |
| 6 | 0 | 1 | 1 |
| 8 | 0 | 1 | 0 |
| 9 | 0 | 1 | 1 |
| 11 | 0 | 1 | 1 |
| 12 | 1 | 1 | 0 |

- 
  - The missing values are completely at random
  - There is no missing value in the parents of R, and R has no child in the BN.
  - Suppose the data includes a total of 12 instances with O = 1, four of which do not have a value for R.

- So $\theta_{r^1|o^1,f^0}$ can be calculated as before

$$\begin{aligned} \theta^*_{r^1|o^1,f^0} &= \frac{N[r^1, o^1, f^0]}{N[o^1, f^0]} = \frac{2}{4} = 0.5 \\ \rightarrow \theta^*_{r^0|o^1,f^0} &= 1 - \theta^*_{r^1|o^1,f^0} = 0.5 \end{aligned} \tag{3.18}$$

- Case 2: If we want to find the CPD corresponding to Covid, when the person wears mask, but does not keep social distancing. The parent mask and social distancing have missing values, so we cannot ignore the instances with missing values and may not use the previous method.

# Chapter 4

# Missing Values: Missing at Random

Following the previous Covid example, what if the dataset with missing values are not completely at random? For instance, some of those who reported coughing did not report drug. Some of those who reported to wear a mask did not take the covid test.

## 4.1 Missing Not Completely at Random

- The MCAR condition does not hold in many cases, often $O_x$ depends on X.
- Example: Registration System
  - Consider a product registration system and the corresponding dataset that includes the hour of registration and the weight of the product. In the second hour, the weight of some products were not saved due to a fault in the system (no longer MCAR!)
  - So the probability of missing values in the product weight for each hour is independent of the weight:
$$P_{missing} \models (O_W \perp W|H) \tag{4.1}$$
  - Implying that $P(O_w|w,h) = P(O_w|h)$
  - The hour is always observed
$$P_{missing}(O_H = O^1) = 1 \tag{4.2}$$
  - And weight and hour are independent.
  - So we only need the following parameters
$$\begin{aligned} \theta_H &= P(H=1) \\ \theta_W &= P(W = Heavy) \\ \psi_{O_W|h^1} &= P_{missing}(O_w = O^1|H=1) \\ \psi_{O_W|h^2} &= P_{missing}(O_w = O^1|H=2) \end{aligned} \tag{4.3}$$

| # | Hour | Weight |
|---|---|---|
| 1 | 1 | Heavy |
| 2 | 1 | Light |
| 3 | 0 | Light |
| 4 | 0 | ? |
| 5 | 0 | Heavy |
| 6 | 0 | ? |
| 7 | 0 | Light |

–

$$
\begin{aligned}
L &= \theta_H \theta_W \psi_{O_{W|h^1}} \\
&\quad * (1-\theta_H)(1-\theta_W)\psi_{\theta_{w|h^2}} \\
&\quad * \theta_H(1-\theta_W)\psi_{\theta_{w|h^1}} \\
&\quad * (1-\theta_H)(1-\psi_{\theta_{w|h^2}}) \\
&\quad * (1-\theta_H)\theta_w \psi_{\theta_{w|h^2}} \\
&\quad * (1-\theta_H)(1-\psi_{\theta_{w|h^2}}) \\
&\quad * \theta_H(1-\theta_W)\psi_{\theta_{w|h^1}} \\
&= \theta_H^3 * (1-\theta_H)^4 * \theta_W^2 * (1-\theta_w)^3 * \psi^3_{\theta_w|h^1} * (1-\psi_{O_{w|h^1}})^0 * \psi^2_{\theta_w|h^2} * (1-\psi_{O_{w|h^2}})^2
\end{aligned} \tag{4.4}
$$

– From the data, we have:
N[H=1, W=heavy] + N[H=1, W=light] = 3
N[H=2, W=heavy] + N[H=2, W=light] = 2
N[H=1, W=?] = 0
N[H=2, W=?] = 0

– The likelihood is decomposed into a term of exclusively $\theta$ and a term of exclusively $\psi$, which can be maximized separately.

## 4.2 Relaxing the MCAR Assumption

- In general,

$$
\begin{aligned}
P_{missing}(y) &= P_{missing}(x) \\
&= \sum_{x^y_{hidden}} P_{missing}(x^y_{obs}, o_x, x^y_{hidden}) \\
&= \sum_{x^y_{hidden}} P_{missing}(x^y_{obs}, x^y_{hidden}) P_{missing}(o_x | x^y_{obs}, x^y_{hidden})
\end{aligned} \tag{4.5}
$$

- Note that $x^y_{obs}$, the superscript y means not all variables of the dataset are included, but only observed variables at particular instance y.

- Now under the assumption $P_{missing} \models (o_x \perp x^y_{hidden} | x^y_{obs})$

$$\begin{aligned} P_{missing}(y) &= \sum_{x^y_{hidden}} P_{missing}(x^y_{obs}, x^y_{hidden}) P_{missing}(o_x | x^y_{obs}) \\ &= P_{missing}(o_x | x^y_{obs}) \sum_{x^y_{hidden}} P_{missing}(x^y_{obs}, x^y_{hidden}) \\ &= P_{missing}(o_x | x^y_{obs}) P(x^y_{obs}) \end{aligned} \tag{4.6}$$

- $P_{missing}(o_x | x^y_{obs})$ depends only on $\psi$
- $P(x^y_{obs})$ dependes only on $\theta$

## Missing at Random (MAR)

- Definition: MAR

The missing data model $P_{missing}$ is **Missing at random (MAR)** if for every non-trival observation y, i.e. $P_{missing}(y) > 0$

$$P_{missing} \models (o_X \perp x^y_{hidden} | x^y_{obs}), \forall x^y_{hidden} \in Val(X^y_{hidden}) \tag{4.7}$$

Where $o_x$ are the specific values of the observability vairiables given y.

- The condition implies that given the observed values, the variables' observability do not depend on the hidden values.
- Remark: MAR is written at the **event-level conditional independence** since every instance may have a different pattern of observed variables.

## MAR: Decomposibility

- Theorem: If $P_{missing}$ satisfies MAR, the likelihood function is decomposed as

$$L_{missing}(\theta, \psi | D) = L(\theta | D) L(\psi | D) \tag{4.8}$$

- Again, the observation mechanism may be ignored.

## Proposition: MLE under MAR

Given the variable $X_i$, assume that $N[pa_{x_i}] \neq 0$. Under MCAR, if at least one of the following condition hold:

- Both $X_i$ and its parents are always observed in the whole dataset, or
- All parents of $X_i$ are observed in the whole dataset and $X_i$ does not have a child.

then the MLE of the CPD parameters $\theta_{X_i | pa_{x_i}}$ is obtained by

$$\theta^*_{x_i | pa_{x_i}} = \frac{N[x_i, pa_{x_i}]}{N[pa_{x_i}]} \tag{4.9}$$

## 4.3 Covid Example

- some of those who reported coughing did not report drug, so the observability of drug depends on $O = o^1$ that is an observed value of cough ($O_R \perp x_{hidden} | x_{obs}$)
- Some of those who reported to wear a mask did not take the covid test. The observability of covid depends on whether they wore a mask ($M = m^1$), which is an observed variable.
- Other missing values were MCAR

# Chapter 5

# Missing at Random: Gradient Ascent

## 5.1 Recall

### Proposition: MLE under MAR

Given the variable $X_i$, assume that $N[pa_{x_i}] \neq 0$. Under MCAR, if at least one of the following condition hold:

- Both $X_i$ and its parents are always observed in the whole dataset, or
- All parents of $X_i$ are observed in the whole dataset and $X_i$ does not have a child.

then the MLE of the CPD parameters $\theta_{X_i|pa_{x_i}}$ is obtained by

$$\theta^*_{x_i|pa_{x_i}} = \frac{N[x_i, pa_{x_i}]}{N[pa_{x_i}]} \tag{5.1}$$

- But what if the parents have missing values?
- We can use (1) gradient ascent; (2) expectation maximization;

## 5.2 Parameter Estimation: Gradient Ascent

- The goal is to estimate the $\theta$ that maximizes the likeihood (that is, an approximation for the MLE)
- The gradient ascent is a standard approach for optimization and can be summarized in the following steps:
  - Initialize the estimate $\hat{\theta} = \theta^0$ randomly.
  - Compute the gradient of log-likelihood l at the current solution $\hat{\theta}$

$$\frac{\partial l(\theta|D)}{\partial \theta}|\theta = \hat{\theta} \tag{5.2}$$

  - Update the solution using some positive scaler $\eta$ as

$$\hat{\theta} \leftarrow \hat{\theta} + \eta \frac{\partial l(\theta|D)}{\partial \theta}|\theta = \hat{\theta} \tag{5.3}$$

  - Repeat steps 2 & 3 until convergence (changes in $\hat{\theta}$ falls short than a speicified threshold)

### Constraints on Parameters

- For discrete CPDs, the optimal $\hat{\theta}$ represents a probability distribution and hence, must satisfy two constraints:
  - All entries of $\hat{\theta}$ must belong to [0, 1]
  - The entries of each CPD must add up to one.

### Computing the Gradient of the Log-likehood

- Consider a Bayesian network over the variables $X_1, X_2, \ldots, X_n$, and let $\mathbf{o}$ be a tuple of observations. If $P(x_i|pa_{x_i}) > 0$,

$$\frac{\partial P(\mathbf{o})}{\partial P(x_i|pa_{x_i})} = \frac{P(x_i, pa_{x_i}, \mathbf{o})}{P(x_i|pa_{x_i})} \tag{5.4}$$

- Consider a Bayesian network over the variables $X_1, X_2, \ldots, X_n$ and a partially observable dataset D = {**o**[1], . . . , **o**[N]}. Then,

$$\frac{\partial l(\boldsymbol{\theta}|D)}{\partial P(x_i|pa_{x_i})} = \frac{1}{P(x_i|pa_{x_i})} \sum_{t=1}^{N} P(x_i, pa_{x_i}|\mathbf{o}[t], \boldsymbol{\theta}) \tag{5.5}$$

- Chain rule of derivatives

$$\frac{\partial l(\boldsymbol{\theta}|D)}{\partial \theta} = \sum_{x_i, pa_{x_i}} \frac{\partial l(\boldsymbol{\theta}|D)}{\partial P(x_i|pa_{x_i})} \frac{\partial P(x_i|pa_{x_i})}{\partial \theta} \tag{5.6}$$

## 5.3 Gradient Ascent: Drawbacks

- MAR (or MCAR) is still assumed. Otherwise, the observation mechanism should be modelled first.
- Gradient ascent is guaranteed to achieve a local maximum
- Global maximum is not guaranteed.
- Repeating the algorithm for multiple random starting points can help.

# Chapter 6

# Missing at Random: Expectation Maximization

## 6.1 Expectation Maximization

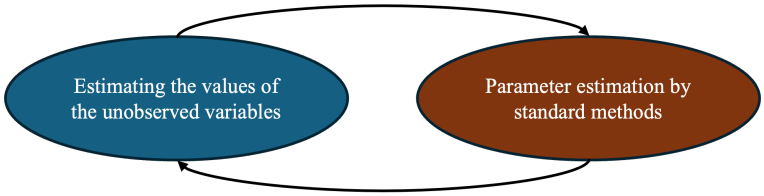
- One way to lear the parameters is to first fill in (impute) the missing values with appropriate values.
- Then we can use standard parameter-learning methods such as MLE.

| # | C | M |
|---|---|---|
| 1 | 1 | ? |
| 2 | 0 | 1 |
| 3 | 1 | 1 |
| 4 | 0 | 1 |
| 5 | ? | 1 |
| 6 | 0 | 1 |
| 7 | 1 | 1 |
| 8 | 0 | ? |
| 9 | 0 | 1 |
| 10 | ? | 1 |
| 11 | 0 | 1 |
| 12 | 1 | 1 |

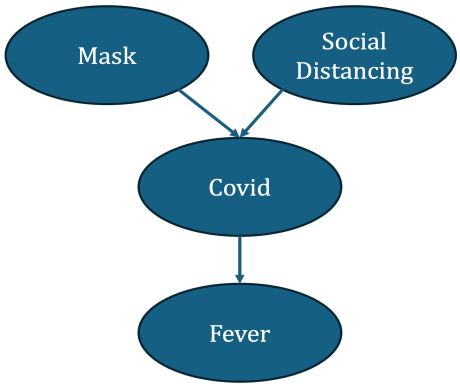
Impute

| # | C | M |
|---|---|---|
| 1 | 1 | 0 |
| 2 | 0 | 1 |
| 3 | 1 | 0 |
| 4 | 0 | 1 |
| 5 | 0 | 1 |
| 6 | 0 | 1 |
| 7 | 1 | 0 |
| 8 | 0 | 1 |
| 9 | 0 | 1 |
| 10 | 0 | 0 |
| 11 | 0 | 1 |
| 12 | 1 | 1 |

- Expectation Maximization (EM) is an iterative method for doing so. At each iteration, it performs two tasks:
  - Estimating the values of the unobserved variables
  - Learning the parameters from the resulting complete data.



## 6.2 EM Example




- Suppose we want to estimate $\theta_{f^1|c^0} = P(F = f^1|C = c^0)$
- If we have fully observed data

$$\theta^*_{f^1|c^0} = \frac{N[f^1, c^0]}{N[c^0]} \tag{6.1}$$

- If we have partially observed data with two instances:

| # | Mask | Social Distancing | Covid | Fever |
|---|---|---|---|---|
| 1 | 1 | ? | ? | 0 |
| 2 | ? | 1 | ? | 1 |

$$\begin{aligned} \xi[1] &=< m^1, ?, ?, f^0 > \\ \xi[2] &=<?, d^1, ?, f^1 > \end{aligned} \tag{6.2}$$

- Assume that our initial guesses for $\theta^0$ is

| $\theta_{m^1}$ | $\theta_{d^1}$ | $\theta_{c^1\|m^0,d^0}$ | $\theta_{c^1\|m^0,d^1}$ | $\theta_{c^1\|m^1,d^0}$ | $\theta_{c^1\|m^1,d^1}$ | $\theta_{f^1\|c^0}$ | $\theta_{f^1\|c^1}$ |
|---|---|---|---|---|---|---|---|
| 0.7 | 0.7 | 0.5 | 0.17 | 0.125 | 0.02 | 0.3 | 0.8 |

-

### 6.2.1 Expectation Phase: Filling in the Missing Values

- For $\xi[1] =< m^1, ?, ?, f^0 >$, we have four cases for <D, C>

$$< d^0, c^0 >, < d^0, c^1 >, < d^1, c^0 >, < d^1, c^1 > \tag{6.3}$$

- We could determine the chances of each case, given the observed values and the current parameters:

$$\begin{aligned} P(D, C|\mathbf{o}[1], \boldsymbol{\theta}) &= P(D, C|m^1, f^0, \boldsymbol{\theta}) \\ &= \frac{P(m^1)P(D)P(C|m^1, D)P(f^0|C)}{\sum_{D,C} P(m^1)P(D)P(C|m^1, D)P(f^0|C)} \end{aligned} \tag{6.4}$$

$$
\begin{aligned}
P(d^0, c^0|\mathbf{o}[1], \boldsymbol{\theta}) &= \frac{0.7 * 0.3 * 0.875 * 0.7}{0.28} = 0.46 \\
P(d^0, c^1|\mathbf{o}[1], \boldsymbol{\theta}) &= \frac{0.7 * 0.3 * 0.125 * 0.7}{0.28} = 0.02 \\
P(d^1, c^0|\mathbf{o}[1], \boldsymbol{\theta}) &= \frac{0.7 * 0.7 * 0.98 * 0.3}{0.28} = 0.51 \\
P(d^1, c^1|\mathbf{o}[1], \boldsymbol{\theta}) &= \frac{0.7 * 0.7 * 0.02 * 0.2}{0.28} = 0.01
\end{aligned} \tag{6.5}
$$

- We can extend and fill in the table accordingly

| # | Weight | Mask | Social Distancing | Covid | Fever |
|---|---|---|---|---|---|
| 1 | 0.46 | 1 | 0 | 0 | 0 |
| 1 | 0.02 | 1 | 0 | 1 | 0 |
| 1 | 0.51 | 1 | 1 | 0 | 0 |
| 1 | 0.01 | 1 | 1 | 1 | 0 |
| 2 | 1 | ? | 1 | ? | 1 |

- For $\xi[2] = <?, d^1, ?, f^1>$, we have four cases for <M, C>

$$
< m^0, c^0 >, < m^0, c^1 >, < m^1, c^0 >, < m^1, c^1 > \tag{6.6}
$$

- Similar to previous case, we could determine the chances of each case, given the observed values and the current parameters:

$$
P(M, C|\mathbf{o}[2], \boldsymbol{\theta}) = P(M, C|d^1, f^1, \boldsymbol{\theta}) \tag{6.7}
$$

$$
\begin{aligned}
P(m^0, c^0|d^1, f^1, \boldsymbol{\theta}) &= 0.23 \\
P(m^0, c^1|d^1, f^1, \boldsymbol{\theta}) &= 0.12 \\
P(m^1, c^0|d^1, f^1, \boldsymbol{\theta}) &= 0.62 \\
P(m^1, c^1|d^1, f^1, \boldsymbol{\theta}) &= 0.03
\end{aligned} \tag{6.8}
$$

- We can extend and fill in the table accordingly

| # | Weight | Mask | Social Distancing | Covid | Fever |
|---|---|---|---|---|---|
| 1 | 0.46 | 1 | 0 | 0 | 0 |
| 1 | 0.02 | 1 | 0 | 1 | 0 |
| 1 | 0.51 | 1 | 1 | 0 | 0 |
| 1 | 0.01 | 1 | 1 | 1 | 0 |
| 2 | 0.23 | 0 | 1 | 0 | 1 |
| 2 | 0.12 | 0 | 1 | 1 | 1 |
| 2 | 0.62 | 1 | 1 | 0 | 1 |
| 2 | 0.03 | 1 | 1 | 1 | 1 |

- Now, to obtain $\theta_{f^1|c^0}$, we count as we did when using MLE, with the difference that now each row counts as much as its weight. For example, $N_{\theta^0}[f^1, c^0] = 0.23 + 0.62 = 0.85$, $N_{\theta^0}[c^0] = 0.46 + 0.51 + 0.23 + 0.62 = 1.82$. Thus,

$$\theta^1_{f^1|c^0} = \frac{N_{\theta^0}[f^1, c^0]}{N_{\theta^0}[c^0]} = 0.46 \tag{6.9}$$

- The quantities $N_{\theta^0}[f^1, c^0]$ and $N_{\theta^0}[c^0]$ are referred to as **expected sufficient statistics**.

## Expected Sufficient Statistics

- Consider the set of random variables **X** and partially observed dataset D = $\{\xi[1], \ldots, \xi[N]\}$.
- Let **H**[t] and **O**[t] denote the missing and observed variables at instance t.
- Let **Y** be a susbet of the variables **X**.
- The **expected sufficient statistics** of $y \in Val(Y)$ given the data D is the expected number of times that y appears in the data instances, given the observed values and parameters.

$$\begin{aligned} N_\theta[y] = \sum_{t=1}^{N} \sum_{\mathbf{h}[t]} P(\mathbf{h}[t]|\mathbf{o}[t], \boldsymbol{\theta}) * 1\{\xi_h[t] < \mathbf{Y} >= \mathbf{y}\} \\ P(\mathbf{h}[t]|\mathbf{o}[t], \boldsymbol{\theta}) \text{is instance weight.} \\ 1\{\xi_h[t] < \mathbf{Y} >= \mathbf{y}\} \text{is instance eligibility.} \end{aligned} \tag{6.10}$$

- $\xi_h[t]$ is the augmented instance t that additionally includes the hidden variables h[t].
- $\xi_h[t] < Y >$ returns the values of Y at augmented instance t.
- $1\{.\}$ returns 1 if the argument is true and 0 otherwise.
- In the example, the quantities $N_{\theta^0}[f^1, c^0]$ and $N_{\theta^0}[c^0]$ are referred to as expected sufficient statistics.

$$\begin{aligned} N_{\theta^0}[f^1, c^0] &= \sum_{s,c} P(s, c|m^1, f^0, \boldsymbol{\theta^0}) * 1\{\xi[1] < F, C >= (f^1, c^0)\} + \\ &\sum_{m,c} P(m, c|d^1, f^1, \boldsymbol{\theta^0}) * 1\{\xi[2] < F, C >= (f^1, c^0)\} \\ &= P(m^0, c^0|d^1, f^1, \boldsymbol{\theta^0}) + P(m^1, c^0|d^1, f^1, \boldsymbol{\theta^0}) \\ &= 0.23 + 0.62 \\ &= 0.85 \end{aligned} \tag{6.11}$$

## 6.3 The EM Algorith for Bayesian Networks

Starting from some initial value $\theta^0$, perform the following steps for k = 0,1,2, . . .

### 6.3.1 Expectation Phase (E-step)

Using current parameter $\theta^k$, compute expected sufficient statistics for each parameter $\theta_{x|pax_x}$. i.e. $N_{\theta^k}[x, pa_x]$ and $N_{\theta^k}[pa_x]$, by either of the following methods:

- Extending the dataset
  - first, fill in the missing values by extending the dataset and finding the probabilities of each possible assignment to the hidden variables given the observed variables.
  - second, sum the weights of the instance that match x, $pa_x$
- Using the formulas

$$\begin{aligned} N_\theta[x, pa_x] &= \sum_t \sum_{h[t]} P(h[t]|\mathbf{o}[t], \boldsymbol{\theta}) * 1\{\xi_h[t] < X, P_{a_x} >= (x, pa_x)\} \\ &= \sum_t P(x, pa_x|\mathbf{o}[t], \boldsymbol{\theta^k}) \\ N_\theta[pa_x] &= \sum_x N_{\theta^k}[x, pa_x] \end{aligned} \tag{6.12}$$

### 6.3.2 Maximization Phase (M-step)

Find the next parameter $\theta^(k+1)$ by maximizing likelihood

$$\theta^{t+1}_{x|pa_x} = \frac{N_{\theta^k}[x, pa_x]}{N_{\theta^k[pa_x]}} \tag{6.13}$$

## 6.4 Conclusion

- EM is guaranteed to increase the likelihood at each iteration.

$$l(\boldsymbol{\theta^k}|D) <= l(\theta^(k+1)|D) \tag{6.14}$$

- EM also cannot guarantee global maximum.
- The maximization step uses standard estimation methods for fully observed data. M-step is straightforward.
- The expectation step requires some inferences. The number of inferences grows exponentially with the number of variables and their parents. E-step is more difficult.

# Chapter 7

# Bayesian Parameter Estimation

## 7.1 Prior Knowledge About Parameters

There are some drawbacks with using MLE

- It assigns an absolute zero to parameters whose corresponding instances did not appear in the dataset. $N[x, pa_x = 0]$ in
$$\theta^*_X = \frac{N[X, pa_X]}{N[pa_X]} \tag{7.1}$$
- This means that the corresponding event never happens in reality, which is often wrong and a consequence of a limited dataset.
- This is also problematic when there is no instance in the dataset on the parents of a variable, makeing the denominator equal to zero.
- If we have prior knowlege about the parameter, we cannot incorporate it into our estimation. For instance, even if we know that a coin is unfair. As long as the result of two tosses is H and T, MLE suggests a fair coin: $\theta = 0.5$.

## 7.2 Prior and Posterior Distributions

- If we treat $\theta$ as a random variable, then by using the Bayes rule,
$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)} \tag{7.2}$$
  - $P(\theta|D)$: posterior parameter distribution
  - $P(D|\theta)$: likelihood function
  - $P(\theta)$: prior parameter distribution
  - $P(D)$: probability of data; act as normalizing factor and can be calculated as
$$P(D) = \int P(D|\theta)P(\theta)d\theta \tag{7.3}$$
- The posterior is a distribution, not a single value.
- Similar to MLE, one may use the mode (maximum), called maximum a posterior (MAP), as the optimal parameter.
$$\theta_{MAP} = argmax_\theta P(\theta|D) \tag{7.4}$$

- The mean or other statistics can be used as well.
- The main difference between the posterior $P(\theta|D)$ and the likelihood $P(D|\theta)$ is that the posterior additionally includes the prior $P(\theta)$

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)} \tag{7.5}$$

## The Coin Toss Example

- Assume we observe N coin tosses: D = {x[1], . . . , x[N]}.
- Let $\theta$ be the probability of heads (H)
- We want to use the mean of the posterior as the optimal value for the parameter $\theta$

$$\overline{\theta} = \int \theta P(\theta|D) d\theta \tag{7.6}$$

- For a uniform prior distribution

$$\begin{aligned} \overline{\theta} &= \int \theta \frac{P(D|\theta)}{P(D)} d\theta \\ &= \frac{1}{P(D)} \int \theta(\theta^{N[H]}(1-\theta)^{N[T]}) d\theta \end{aligned} \tag{7.7}$$

- Using integration by part, we have:

$$\begin{aligned} \int \theta^{N[H]+1}(1-\theta)^{N[T]} d\theta = \int_0^1 (N[H]+1)\theta^{N[H]} \frac{(1-\theta)^{N[T]+1}}{N[T]+1} d\theta \\ -\theta^{N[H]+1} \frac{(1-\theta)^{N[T]+1}}{N[T]+1} \end{aligned} \tag{7.8}$$

$$\begin{aligned} \overline{\theta} &= \frac{1}{P(D)} \frac{N[H]+1}{N[T]+1} \int_0^1 \theta^{N[H]}(1-\theta)^{N[T]+1} d\theta \\ &= \frac{1}{P(D)} \frac{N[H]+1}{N[T]+1} \int_0^1 \theta^{N[H]}(1-\theta)^{N[T]}(1-\theta) d\theta \\ &= \frac{1}{P(D)} \frac{N[H]+1}{N[T]+1} [\int_0^1 \theta^{N[H]}(1-\theta)^{N[T]} d\theta - \int_0^1 \theta^{N[H]+1}(1-\theta)^{N[T]} d\theta] \\ &= \frac{1}{P(D)} \frac{N[H]+1}{N[T]+1} [\int_0^1 P(D|\theta) d\theta - \int_0^1 \theta^{N[H]+1}(1-\theta)^{N[T]} d\theta] \\ &= \frac{1}{P(D)} \frac{N[H]+1}{N[T]+1} [P(D) - P(D)\overline{\theta}] \\ &= \frac{N[H]+1}{N[T]+1} - \frac{N[H]+1}{N[T]+1} \overline{\theta} \end{aligned} \tag{7.9}$$

- So, the mean of the posterior is

$$\overline{\theta} = \frac{N[H]+1}{N[H]+N[T]+2} \tag{7.10}$$

- The maximum likelihood estimation (MLE) was

$$\theta^*_{MLE} = \frac{N[H]}{N[H]+N[T]} \tag{7.11}$$

- They match at infinitely large data size

$$\lim_{N->\infty} \overline{\theta} = \lim_{N->\infty} \theta^*_{MLE} \tag{7.12}$$

## 7.3 Beta Distribution

- The beta distribution with two hyperparameters $\alpha_0$, $\alpha_1 > 0$ and a normalizing constant $\gamma$

$$\begin{aligned} \theta &\sim Beta(\alpha_1, \alpha_0) \\ PDF : P(\theta) &= \gamma \theta^{\alpha_1 - 1}(1-\theta)^{\alpha_0 - 1} \end{aligned} \tag{7.13}$$

- In the coin toss example, the posterior mean with the Beta prior is

$$\overline{\theta} = \frac{N[H] + \alpha_1}{N[H] + \alpha_1 + N[T] + \alpha_0} \tag{7.14}$$

- For example, for N[H] = 6 and N[T] = 4,

$$\begin{aligned} Beta(1,1) : \overline{\theta} &= \frac{N[H]+1}{N[H]+1+N[T]+1} = \frac{7}{12} = 0.58 \\ Beta(2,2) : \overline{\theta} &= \frac{N[H]+2}{N[H]+2+N[T]+2} = \frac{8}{14} = 0.57 \\ Beta(10,10) : \overline{\theta} &= \frac{N[H]+10}{N[H]+10+N[T]+10} = \frac{16}{30} = 0.53 \end{aligned} \tag{7.15}$$

- The MLE equals

$$\theta^*_{MLE} = \frac{N[H]}{N[H]+N[T]} = \frac{6}{10} = 0.6 \tag{7.16}$$

- For large enough data instances, the posterior mean with the beta prior approximate MLE.

## 7.4 Dirichlet Distribution

Categorical variables with more than two values. We could use drichlet distribution, which is a natural extension to beta distribution.

- Dirichlet distribution with hyperparameters $\alpha_1, \ldots, \alpha_k$:

$$\begin{aligned} \theta = \{\theta_1, \ldots, \theta_k\} &\sim Dirichlet(\alpha_1, \ldots, \alpha_k) \\ PDF : p(\theta) &\propto \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \\ E[\theta_k] = \overline{\theta_k} &= \frac{\alpha_k}{\sum \alpha_k} \\ model[\theta_k] &= \frac{\alpha_k - 1}{\sum \alpha a_k - K} \end{aligned} \tag{7.17}$$

- Proposition: Consider the random variable $X \in \{x^1, \ldots, x^K\}$ and parameters $\theta = (\theta_1, \ldots, \theta_k)$, where $P(X = x^k|\theta) = \theta_k$. If the prior is

$$P(\theta) = Dirichlet(\alpha_1, \ldots, \alpha_k) \tag{7.18}$$

  then the posterior equals

$$P(\theta|D) = Dirichlet(\alpha_1 + N[x^1], \ldots, \alpha_k + N[x^K]) \tag{7.19}$$

  where $N[x^K]$ is the number of occurrences of $x_K$ in the dataset D.

### 7.4.1 Posterior Predictive Distribution

- Using the posterior, we can predict new instances
- Consider the random variable X whose probability distribution is parameterized by $\theta$. Given the observed data D = x[1], . . . , x[N], the **posterior predictive distribution** or **bayesian estimator** is

$$\begin{aligned} P(x[N+1]|D) &= \int P(x[N+1], \theta|D) d\theta \\ &= \int P(x[N+1]|\theta, D) P(\theta|D) d\theta \\ &= \int P(x[N+1]|\theta) P(\theta|D) d\theta \end{aligned} \tag{7.20}$$

- In the case of a multinomial distribution, with $X \in \{x^1, \ldots, x^k\}$ and $P(X = x^i|\theta) = \theta_i$.

$$P(X[N+1] = x^i|D) = \int \theta_i P(\theta_i|D) d\theta_i = E[\theta_i|D] \tag{7.21}$$

  which is the mean of the posterior of $\theta_i$

### 7.4.2 Posterior Mean for Dirichlet Prior

- Consider the multinomial variable $X \in \{x^1, \ldots, x^K\}$ parameterized by $\theta = (\theta_1, \ldots, \theta_k)$, where $P(X_k = x^k|\theta) = \theta_k$. Consider the dataset D = {x[1], . . . , x[N]}. For a dirichlet prior $Dirichlet(\alpha_1, \ldots, \alpha_k)$, the posterior mean and the bayesian estimator equals

$$\overline{\theta_k} = P(X[N+1] = x^k|D) = \frac{N[x^k] + \alpha_k}{N + \alpha} \tag{7.22}$$

  where $\alpha = \sum_{k=1}^{K} \alpha_k$, and posterior mode equals

$$\theta_k^{MAP} = \frac{N[x^k] + \alpha_k - 1}{N + \alpha - K} \tag{7.23}$$

# Chapter 8

# Bayesian Parameter Estimation in Bayesian Networks

## 8.1 CPDs with Dirichlet Priors

- If the parameter prior distribution of a bayesian network satisfies

$$P(\theta_{X|pa_x}) = Dirichlet(\alpha_{x^1|pa_x}, \ldots, \alpha_{x^k|pa_x}) \tag{8.1}$$

then given the dataset $D = \{\xi[1], \ldots, \xi[N]\}$, the posterior satisfies

$$P(\theta_{X|pax_x}|D) = Dirichlet(a_{x^1|pa_x} + N[x^1, pa_x], \ldots, a_{x^K|pa_x} + N[x^K, pa_x]) \tag{8.2}$$

resulting in the following posterior mean and mode:

$$\begin{aligned} \overline{\theta}_{x^k|pa_x} &= \frac{N[x^k, pa_x] + \alpha_{x^k|pa_x}}{N[pa_x + \sum_k \alpha_{x^k|pa_x}]} \\ \theta^{MAP}_{x^k|pa_x} &= \frac{N[x^k, pa_x] + \alpha_{x^k|pa_x} - 1}{N[pa_x] + \sum_k \alpha_{x^k|pa_x} - K} \end{aligned} \tag{8.3}$$

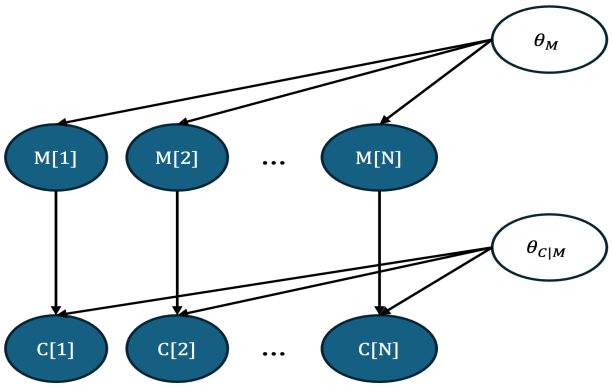
- We first need to see if the global and local decomposition of the posterior hold so that we can break $P(\theta|D)$ into the multiplications of $P(\theta_{x|pa_x}|D)$

## 8.2 Global Parameter Independence

- Consider the covid-mask problem: P(M,C) = P(C|M)P(M)
- With unknown vector parameters,

$$\theta_M = (\theta_{m^0}, \theta_{m^1})\theta_{C|M} = (\theta_{c^0|m^0}, \theta_{c^1|m^0}, \theta_{c^0|m^1}, \theta_{c^1|m^1}) \tag{8.4}$$

- We can form a bayesian network for iid samples from the network M -> C with $\theta_M \perp \theta_{C|M}$ for N samples:




- Satisfying sample independence

$$(M[t], C[t] \perp M[t\prime], C[t\prime]|\theta_M, \theta_{C|M})\forall t, t\prime \tag{8.5}$$

- Global Parameter Independence: Consider a bayesian network with parameters $\theta = (\theta_{X_1|Pa_{x_1}}, \ldots, \theta_{X_n|Pa_{x_n}})$. A prior $P(\theta)$ is said to satisfy **global parameter independence** if

$$P(\theta) = \prod_{i=1}^{n} P(\theta_{x_i|pa_{x_i}}) \tag{8.6}$$