# COMS W4701 HW5

Dawei He (dh3027)

May 3, 2022

## Problem 1

### (a)

$$Pr(D_0, D_1, ..., D_n) = \prod_{i=0}^{n} P(D_i|D_0, ..., D_{i-1})$$

The size of full joint distribution is $2^{n+1}$.
There are 2 entries are nonzero.

### (b)

| $D_0$ | $Pr(+d_n\|D_0)$ |
|---|---|
| $+d_0$ | 1 |
| $-d_0$ | 0 |

$$Pr(+d_n|D_0) = \frac{\sum_{D_1, D_2, ..., D_{n-1}} Pr(D_0, D_1, ..., D_n)}{Pr(+d_n| + d_0) + Pr(+d_n| - d_0)}$$

### (c)

$$\begin{aligned} Pr(A, B, D_0, +d_n) &= Pr(A)Pr(B)Pr(D_0|A, B)Pr(+d_n|A, B, D_0, ..., D_{n-1}) \\ &= Pr(A)Pr(B)Pr(D_0|A, B)Pr(+d_n|D_{n-1}) \end{aligned}$$

We can treat CPT of $Pr(D_i|D_{i-1})$ as a transition matrix. In this case, the matrix will be a diagonal matrix with all elements on diagonal are equal to 1. So we can rewrite above equation:

$$\begin{aligned} Pr(A, B, D_0, +d_n) &= Pr(A)Pr(B)Pr(D_0|A, B)Pr(+d_n|D_{n-1}) \\ &= Pr(A)Pr(B)Pr(D_0|A, B)Pr(+d_n|D_{n-1})Pr(D_{n-1}|D_{n-2}) \\ &= Pr(A)Pr(B)Pr(D_0|A, B)Pr(+d_n|D_{n-2}) \\ &= Pr(A)Pr(B)Pr(D_0|A, B)Pr(+d_n|D_{n-2})Pr(D_{n-2}|D_{n-3}) \\ &... \\ &= Pr(A)Pr(B)Pr(D_0|A, B)Pr(+d_n|D_0) \end{aligned}$$

### (d)

We can get the CPT of $Pr(A, B, D_0, +d_n)$ from (c).

| A | B | $D_0$ | $Pr(A, B, D_0, +d_n)$ |
|---|---|---|---|
| +a | +b | $+d_0$ | 0.5*0.5*1*1 |
| +a | +b | $-d_0$ | 0.5*0.5*0*0 |
| +a | -b | $+d_0$ | 0.5*0.5*0.5*1 |
| +a | -b | $-d_0$ | 0.5*0.5*0.5*0 |
| -a | +b | $+d_0$ | 0.5*0.5*0.5*1 |
| -a | +b | $-d_0$ | 0.5*0.5*0.5*0 |
| -a | -b | $+d_0$ | 0.5*0.5*0*1 |
| -a | -b | $-d_0$ | 0.5*0.5*1*0 |

Sum over $D_0$ we can get:

| A | B | $Pr(A, B, +d_n)$ |
|---|---|---|
| +a | +b | 0.5*0.5*1*1 |
| +a | -b | 0.5*0.5*0.5*1 |
| -a | +b | 0.5*0.5*0.5*1 |
| -a | -b | 0.5*0.5*0*1 |

We can normalize it to get $Pr(A, B| + d_n)$.

| A | B | $Pr(A, B| + d_n)$ |
|---|---|---|
| +a | +b | $\frac{1}{2}$ |
| +a | -b | $\frac{1}{4}$ |
| -a | +b | $\frac{1}{4}$ |
| -a | -b | 0 |

Sum over B we can get:

| A | $Pr(A| + d_n)$ |
|---|---|
| +a | $\frac{3}{4}$ |
| -a | $\frac{1}{4}$ |

Sum over A we can get:

| B | $Pr(B| + d_n)$ |
|---|---|
| +b | $\frac{3}{4}$ |
| -b | $\frac{1}{4}$ |

If we $Pr(A| + d_n) \times Pr(B| + d_n)$.

| A | B | $Pr(A, B| + d_n)$ |
|---|---|---|
| +a | +b | $\frac{9}{16}$ |
| +a | -b | $\frac{3}{16}$ |
| -a | +b | $\frac{3}{16}$ |
| -a | -b | $\frac{1}{16}$ |

Clearly $Pr(A| + d_n) \times Pr(B| + d_n) \neq Pr(A, B| + d_n)$, so A and B are not independent conditioned on $D_n$.

# Problem 2

## (a)

Given no observations in the Bayes net, pair $(tampering, fire), (tampering, smoke)$ are independent.

Now suppose Alarm is observed.
pair $(tampering, fire), (tampering, smoke)$ are not independent, but they are independent given no observations.
pair $(tampering, leaving), (tampering, report)$ are independent, but they are not independent given no observations.
pair $(smoke, leaving), (smoke, report)$ are independent, but they are not independent given no observations.
pair $(fire, leaving), (fire, report)$ are independent, but they are not independent given no observations.

## (b)

$$P(smoke|report) \propto \sum_{fire,tampering,alarm,leaving} P(fire)P(smoke|fire)P(tampering)$$
$$P(alarm|tampering,fire)P(leaving|alarm)P(report|leaving)$$

The size of the table should be 4.

## (c)

Greatest number of operation:

$$\sum_{leaving} P(report|leaving) \sum_{tampering} P(tampering) \sum_{alarm} P(leaving|alarm) \sum_{fire} P(fire)P(smoke|fire)P(alarm|tampering,fire)$$

The max table size is 16.

Fewest number of operationg:

$$\sum_{fire} P(fire)P(smoke|fire) \sum_{leaving} P(report|leaving) \sum_{alarm} P(leaving|alarm) \sum_{tampering} P(tampering)P(alarm|tampering,fire)$$

The max table size is 8.

## (d)

| tampering | fire | alarm | $P(alarm, tampering|fire)$ |
|---|---|---|---|
| + | + | + | 0.02*0.01*0.5=0.0001 |
| + | + | - | 0.02*0.01*0.5=0.0001 |
| + | - | + | 0.02*0.99*0.85=0.01683 |
| + | - | - | 0.02*0.99*0.15=0.00297 |
| - | + | + | 0.98*0.01*0.99=0.009702 |
| - | + | - | 0.98*0.01*0.01=0.000098 |
| - | - | + | 0.98*0.99*0=0 |
| - | - | - | 0.98*0.99*1=0.9702 |

$$\sum_{tampering} \downarrow$$

| fire | alarm | $P(alarm|fire)$ |
|---|---|---|
| + | + | 0.009802 |
| + | - | 0.000198 |
| - | + | 0.01683 |
| - | - | 0.97317 |

$P(leaving|alarm) \downarrow$

| leaving | fire | alarm | $P(leaving, alarm\|fire)$ |
|---|---|---|---|
| + | + | + | 0.009802*0.88 |
| + | + | - | 0.000198*0 |
| + | - | + | 0.01683*0.88 |
| + | - | - | 0.97317*0 |
| - | + | + | 0.009802*0.12 |
| - | + | - | 0.000198*1 |
| - | - | + | 0.01683*0.12 |
| - | - | - | 0.97317*1 |

$\sum_{alarm} \downarrow$

| leaving | fire | $P(leaving\|fire)$ |
|---|---|---|
| + | + | 0.00862576 |
| + | - | 0.0148104 |
| - | + | 0.00137424 |
| - | - | 0.9751896 |

$P(report|leaving) \downarrow$

| report | leaving | fire | $P(report, leaving\|fire)$ |
|---|---|---|---|
| + | + | + | 0.00862576*0.75 |
| + | + | - | 0.0148104*0.75 |
| + | - | + | 0.00137424*0.01 |
| + | - | - | 0.9751896*0.01 |
| - | + | + | 0.00862576*0.25 |
| - | + | - | 0.0148104*0.25 |
| - | - | + | 0.00137424*0.99 |
| - | - | - | 0.9751896*0.99 |

$\sum_{leaving} \downarrow$

| report | fire | $P(report\|fire)$ |
|---|---|---|
| + | + | 0.0064830624 |
| + | - | 0.020859696 |
| - | + | 0.0035169376 |
| - | - | 0.969140304 |

$P(fire)P(smoke|fire) \downarrow$

| smoke | report | fire | $P(smoke, report, fire)$ |
|---|---|---|---|
| + | + | + | 0.0064830624*0.01*0.9 |
| + | + | - | 0.020859696*0.99*0.01 |
| + | - | + | 0.0035169376*0.01*0.9 |
| + | - | - | 0.969140304*0.99*0.01 |
| - | + | + | 0.0064830624*0.01*0.1 |
| - | + | - | 0.020859696*0.99*0.99 |
| - | - | + | 0.0035169376*0.01*0.1 |
| - | - | - | 0.969140304*0.99*0.99 |

$\sum_{fire} \downarrow$

| smoke | report | $P(smoke, report)$ |
|---|---|---|
| + | + | 0.000264858552 |
| + | - | 0.009626141448 |
| - | + | 0.02045107111 |
| - | - | 0.9498579289 |

$Normalize \downarrow$

| smoke | report | $P(smoke\|report)$ |
|---|---|---|
| + | + | 0.00027021 |
| + | - | 0.00982059 |
| - | + | 0.02086418 |
| - | - | 0.96904502 |

# problem 3

## (a)

1. Given influenza, $S_1$={sore throat}, $s_2$={fever}, $S_3$={bronchitis, smokes, coughing, wheezing}
2. Given bronchitis, $S_1$={coughing}, $S_2$={wheezing}, $S_3$={sore throat, fever, influenza, smokes}
3. Given both influenza and bronchitis, $S_1$={sore throat}, $S_2$={fever}, $S_3$={smokes}, $S_4$={coughing, wheezing}

## (b)

If influenza and smokes are observed, the distribution of the resulting samples will be very close to the target distribution, all weights will be identical.

If coughing and wheezing are observed, the evidence will not guide the generation of samples, so the distribution of the resulting samples will not reflect target distribution. The weights will be different, and most corrections done by the weights.

## (c)

$$P(influenza|mb(influenza)) \propto P(influenza)P(+st|influenza)P(-f|influenza)P(-b|influenza, +s)$$

The max size of the table should be 2.

## (d)

| I | $P(I\|sample)$ |
|---|---|
| + | 0.000015 |
| - | 0.00027075 |

=

| I | $P(I)$ |
|---|---|
| + | 0.05 |
| - | 0.95 |

×

| I | $P(+st\|I)$ |
|---|---|
| + | 0.3 |
| - | 0.001 |

×

| I | $P(-f\|I)$ |
|---|---|
| + | 0.1 |
| - | 0.95 |

×

| I | $P(-b\|I, +s)$ |
|---|---|
| + | 0.01 |
| - | 0.3 |

# Problem 4

## 4.1

### (a)

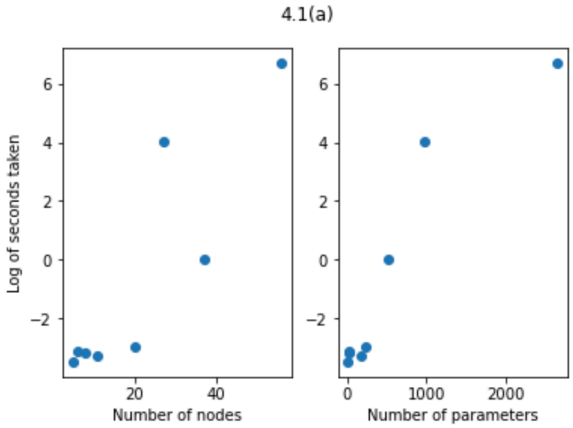



Figure 1: i) log of seconds taken vs number of nodes, and ii) log of seconds taken vs number of Bayes net parameters.

We can see that the log of seconds taken has quadratic relation with number of nodes and log of seconds taken has linear relation with number of parameters. So exact inference is generally exponential in Bayes net size.

### (b)

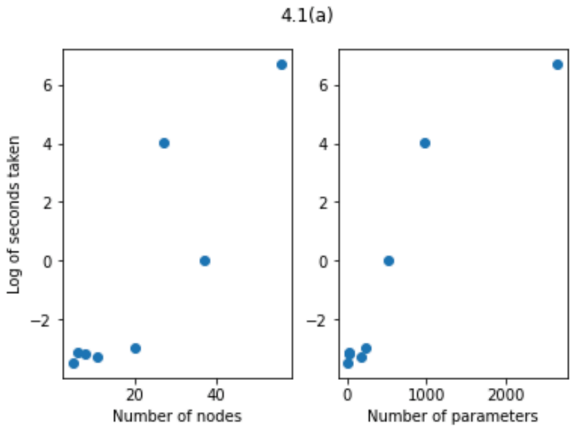



Figure 2: i) log of seconds taken vs number of nodes, and ii) log of seconds taken vs number of Bayes net parameters with the option elimination order=MinNeighbors

We can see that in plot (b), each point has less log of seconds taken. So elimination order is important to the efficiency of inference. If we can find a good elimination order, we can have smaller size of tables during the calculation, which can save us a lot of time and memory.

## 4.2

### (a)

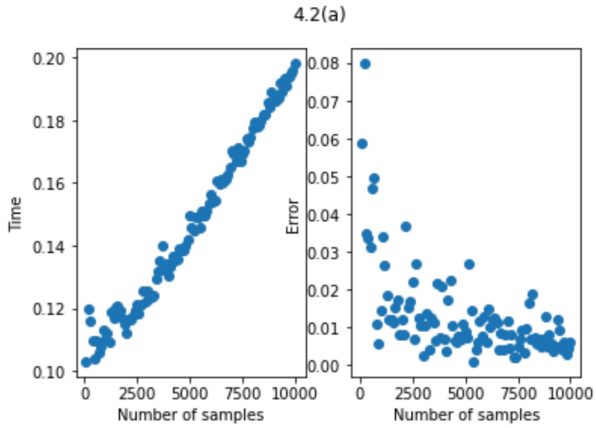



Figure 3: Time taken per number of samples and error per number of samples

### (b)

Computational complexity has positive linear relation with number of samples. Using sampling has a faster speed than computing exact inference but it has relative large error if the sample size is small. After 7500 number of samples the error stop improving. The average error to the exact inference is around 0.005 after 7500 number of samples.

### (c)

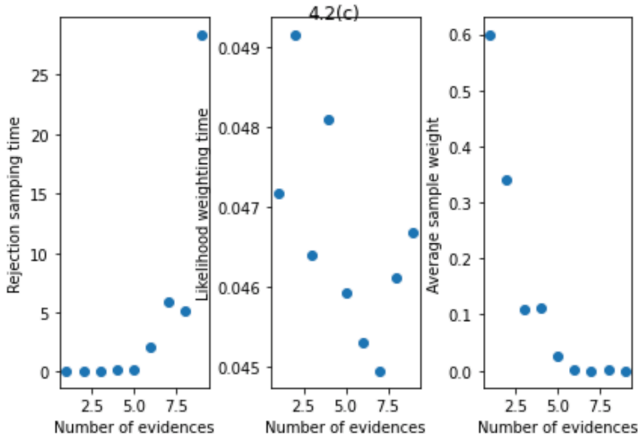
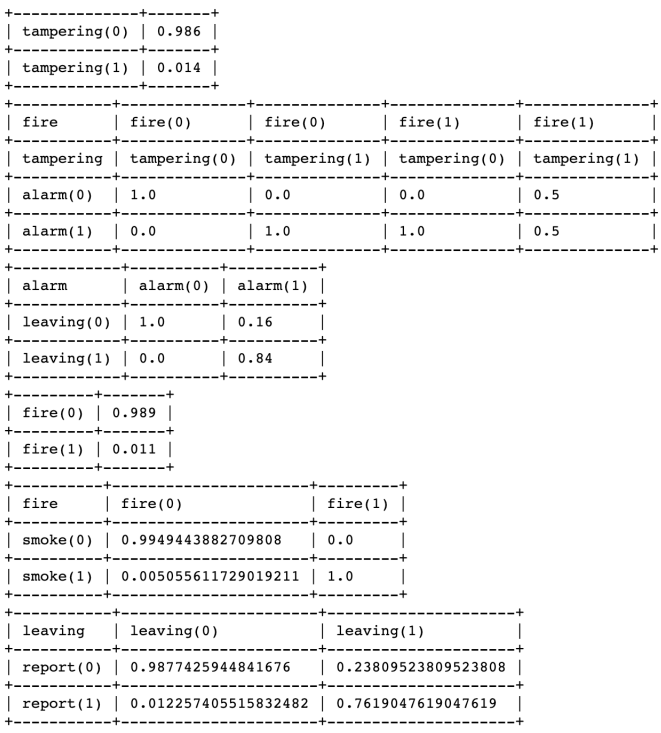



Figure 4: Time taken vs number of evidence variables for rejection sampling. Time taken vs number of evidence variables for likelihood weighting. Average sample weight vs number of evidence variables (for likelihood weighting)

### (d)

For rejection sampling, the time needed to perform sampling increase exponentially with the number of evidence. Because if the number of evidences increase, we have to reject a lot of sample to meet the requirement of observed evidence, so the time will increase a lot. For likelihood weighting, the time needed to perform sampling are almost the same regardless of the number of evidence. Because every sample generated meet the requirements of observed evidence and we will accept them all, so we spent almost the time for each evidence set. The average sample weight decrease with the number of evidence. Because more evidence means we have to times more p(e—parent(e)) in weight calculation, and they are all less than 1. So weights will approach 0 if more evidences are given.

## 4.3



| tampering(0) | 0.986 |
|---|---|
| tampering(1) | 0.014 |

| fire | fire(0) | fire(0) | fire(1) | fire(1) |
|---|---|---|---|---|
| tampering | tampering(0) | tampering(1) | tampering(0) | tampering(1) |
| alarm(0) | 1.0 | 0.0 | 0.0 | 0.5 |
| alarm(1) | 0.0 | 1.0 | 1.0 | 0.5 |

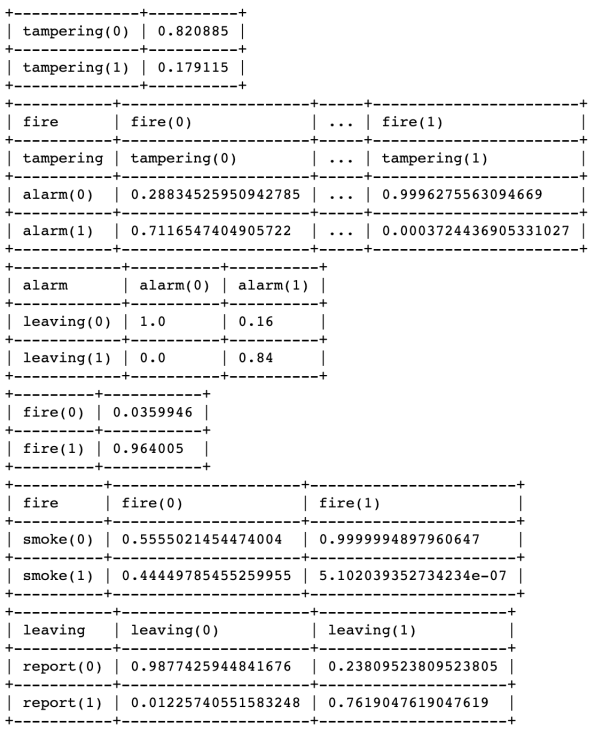
| alarm | alarm(0) | alarm(1) |
|---|---|---|
| leaving(0) | 1.0 | 0.16 |
| leaving(1) | 0.0 | 0.84 |

| fire(0) | 0.989 |
|---|---|
| fire(1) | 0.011 |

| fire | fire(0) | fire(1) |
|---|---|---|
| smoke(0) | 0.9949443882709808 | 0.0 |
| smoke(1) | 0.005055611729019211 | 1.0 |

| leaving | leaving(0) | leaving(1) |
|---|---|---|
| report(0) | 0.9877425944841676 | 0.23809523809523808 |
| report(1) | 0.012257405515832482 | 0.7619047619047619 |

(a) MLE



| tampering(0) | 0.820885 |
|---|---|
| tampering(1) | 0.179115 |

| fire | fire(0) | ... | fire(1) |
|---|---|---|---|
| tampering | tampering(0) | ... | tampering(1) |
| alarm(0) | 0.28834525950942785 | ... | 0.9996275563094669 |
| alarm(1) | 0.7116547404905722 | ... | 0.0003724436905331027 |

| alarm | alarm(0) | alarm(1) |
|---|---|---|
| leaving(0) | 1.0 | 0.16 |
| leaving(1) | 0.0 | 0.84 |

| fire(0) | 0.0359946 |
|---|---|
| fire(1) | 0.964005 |

| fire | fire(0) | fire(1) |
|---|---|---|
| smoke(0) | 0.5555021454474004 | 0.9999994897960647 |
| smoke(1) | 0.44449785455259955 | 5.102039352734234e-07 |

| leaving | leaving(0) | leaving(1) |
|---|---|---|
| report(0) | 0.9877425944841676 | 0.23809523809523805 |
| report(1) | 0.01225740551583248 | 0.7619047619047619 |

(b) EM

Because ME method have to first estimate the missing variables in the dataset and then maximize the parameters of the model using the data. However, the data we generate for the missing data could have a totally different distribution than their true distribution. So when we try to maximize the parameter in the model to fit the generated data, the parameters could be very different from the parameters in MLE model. Because MLE learn from original data directly and do not have to generate the data for latent feature.