Probalistic Graphic Model[1][2] Notes

Yan JIN

November 30, 2016

Contents

1 Representation

1.1 Introduction and Overview

Model: The model is a **declarative representation** of our understanding of the world. It's **declarative** means that the representation stands on its own, which means that we can look into it and make sense of it **aside from any algorithm** that we might choose to apply on.

1. Representation
   - Directed and undirected
   - Temperal and plate models
2. Inference
   - Exact and approximate
   - Decision making
3. Learning
   - Parameters and structure
   - With and without complete data

1.1.1 Distributions (Chapters 2.1.1 to 2.1.3)

1.1.2 Factors (Chapter 4.2.1)

1.1.3 Quiz

1. Factor product.

Let X, Y and Z be binary variables.

If $\phi_1(X,Y)$ and $\phi_2(Y,Z)$ are the factors shown below, compute the selected entries (marked by a '?') in the factor $\psi(X,Y,Z) = \phi_1(X,Y) \cdot \phi_2(Y,Z)$, giving your answer according to the ordering of assignments to variables as shown below.

Separate each of the 3 entries of the factor with spaces, e.g., an answer of

0.1 0.2 0.3

means that $\psi(1,1,1) = 0.1$, $\psi(1,2,1) = 0.2$, and $\psi(2,2,2) = 0.3$. Give your answers as exact decimals without any trailing zeroes.

X	Y	$\phi_1(X,Y)$
1	1	0.8
1	2	0.5
2	1	0.5
2	2	0.6

×

Y	Z	$\phi_2(Y,Z)$
1	1	0.2
1	2	0.2
2	1	0.9
2	2	1.0

=

X	Y	Z	$\psi(X,Y,Z)$
1	1	1	?
1	1	2	
1	2	1	?
1	2	2	
2	1	1	
2	1	2	
2	2	1	
2	2	2	?

Fig. 1: Exercise 01-01-01

2. Factor reduction.

Let X, Z be binary variables, and let Y be a variable that takes on values 1, 2, or 3.

Now say we observe $Y = 3$. If $\phi(X, Y, Z)$ is the factor shown below, compute the missing entries of the reduced factor $\psi(X, Z)$ given that $Y = 3$, giving your answer according to the ordering of assignments to variables as shown below.

As before, you may separate the 4 entries of the factor by spaces.

X	Y	Z	$\phi(X,Y,Z)$
1	1	1	14
1	1	2	60
1	2	1	40
1	2	2	27
1	3	1	42
1	3	2	85
2	1	1	4
2	1	2	59
2	2	1	54
2	2	2	3
2	3	1	96
2	3	2	30

X	Z	$\psi(X,Z)$
1	1	?
1	2	?
2	1	?
2	2	?

Fig. 2: Exercise 01-01-02

3. **Properties of independent variables.**

Assume that A and B are independent random variables. Which of the following options are always true? You may select 1 or more options.

- ☐ $P(B|A) = P(B)$
- ☐ $P(A, B) = P(A) \times P(B)$
- ☐ $P(A) = P(B)$
- ☐ $P(A) \neq P(B)$

Fig. 3: Exercise 01-01-03

4. Factor marginalization.

Let X, Z be binary variables, and let Y be a variable that takes on values 1, 2, or 3.

If $\phi(X, Y, Z)$ is the factor shown below, compute the entries of the factor

$\psi(Y, Z) = \sum_X \phi(X, Y, Z)$,

giving your answer according to the ordering of assignments to variables as shown below.

Separate the 4 entries of the factor with spaces, and do not add any extra trailing or leading zeroes or decimal points.

X	Y	Z	$\phi(X,Y,Z)$
1	1	1	68
1	1	2	95
1	2	1	65
1	2	2	63
1	3	1	57
1	3	2	5
2	1	1	40
2	1	2	40
2	2	1	14
2	2	2	78
2	3	1	16
2	3	2	89

Y	Z	$\psi(Y,Z)$
1	1	?
1	2	?
2	1	?
2	2	?
3	1	
3	2	

Fig. 4: Exercise 01-01-04

1.1.4 Answers

01-01-01: 0.16 0.45 0.6;

01-01-02: 42 85 96 30;

01-01-03: 1st, 2nd;

01-01-04: 108 135 79 141;

1.2 Baysian Network (Directed Models)

1.2.1 Semantics and Factorization (Chapters 3.2.1, 3.2.2)

If you are unfamiliar with genetic inheritance, please watch this short Khan Academy video for some background.

CPD: Conditional Probability Distribution;

DAG: Directed Acyclic Graph;

P **factorizes** over Graph G if: $P(X_1, ..., X_n) = \prod_i \ P(X_i|Par_G(X_i))$

1.2.2 Reasoning Patterns (Chapter 3.2.1.2)

Causal Reasoning(Father to Son), Evidential Reasoning(Son to Father), Intercausal Reasoning;

1.2.3 Flow of Probabilistic Influence (Chapter 3.3.1)

When influence can flow from X to Y via Z, we can say that the trail $X \rightleftharpoons Z \rightleftharpoons Y$ is **active**;

v-structure;

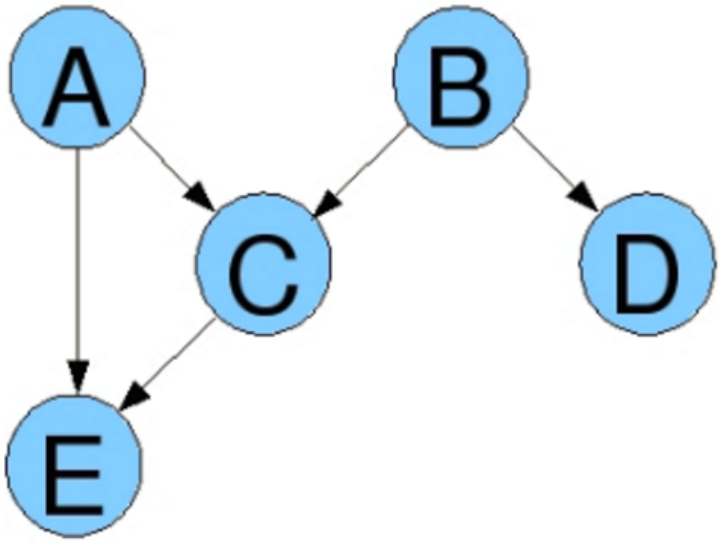
(Definition 3.6) Let $\mathcal{G}$ be a BN structure, and $X_1 \rightleftharpoons ... \rightleftharpoons X_n$ a trail in $\mathcal{G}$. Let **Z** be a subset of observed variables. The trail $X_1 \rightleftharpoons ... \rightleftharpoons X_n$ **is active given Z** if:

- Whenever we have a v-structure $X_{i-1} \rightarrow X_i \leftarrow X_{i+1}$, then X_i or one of its descendants are in **Z**;
- No other node along the trail is in **Z**;

1.2.4 Quiz

1. Factorization.

Given the same model as above, which of these is an appropriate decomposition of the joint distribution $P(A, B, C, E)$?

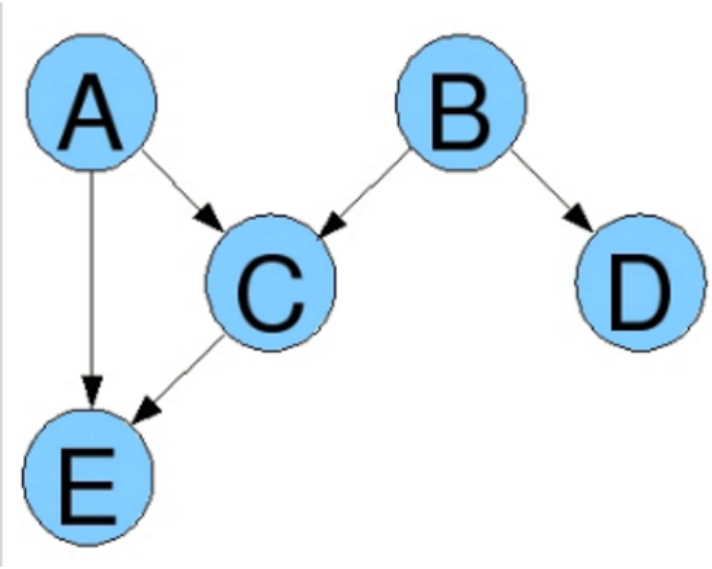


- $P(A, B, C, E) = P(A)P(B)P(A, B|C)P(A, C|E)$
- $P(A, B, C, E) = P(A)P(B)P(C|A, B)P(E|A, C)$
- $P(A, B, C, E) = P(A)P(B)P(C|A)P(C|B)P(E|A)P(E|C)$
- $P(A, B, C, E) = P(A)P(B)P(C)P(E)$

Fig. 5: Exercise 01-02-01

2. Independent parameters.

How many independent parameters are required to uniquely define the CPD of C (the conditional probability distribution associated with the variable C) in the same graphical model as above, if A, B, and D are binary, and C and E have three values each?



If you haven't come across the term before, here's a brief explanation: A multinomial distribution over m possibilities $x_1, \ldots, x_m$ has m parameters, but $m-1$ independent parameters, because we have the constraint that all parameters must sum to 1, so that if you specify $m-1$ of the parameters, the final one is fixed. In a CPD $P(X|Y)$, if X has m values and Y has k values, then we have k distinct multinomial distributions, one for each value of Y, and we have $m-1$ independent parameters in each of them, for a total of $k(m-1)$. More generally, in a CPD $P(X|Y_1, \ldots, Y_r)$, if each Y_i has k_i values, we have a total of $k_1 \times \ldots \times k_r \times (m-1)$ independent parameters.

Example: Let's say we have a graphical model that just had $X \rightarrow Y$, where both variables are binary. In this scenario, we need 1 parameter to define the CPD of X. The CPD of X contains two entries $P(X = 0)$ and $P(X = 1)$. Since the sum of these two entries has to be equal to 1, we only need one parameter to define the CPD.

Now we look at Y. The CPD for Y contains 4 entries which correspond to: $P(Y = 0|X = 0), P(Y = 1|X = 0), P(Y = 0|X = 1), P(Y = 1|X = 1)$. Note that $P(Y = 0|X = 0)$ and $P(Y = 1|X = 0)$ should sum to one, so we need 1 independent parameter to describe those two entries; likewise, $P(Y = 0|X = 1)$ and $P(Y = 1|X = 1)$ should also sum to 1, so we need 1 independent parameter for those two entries.

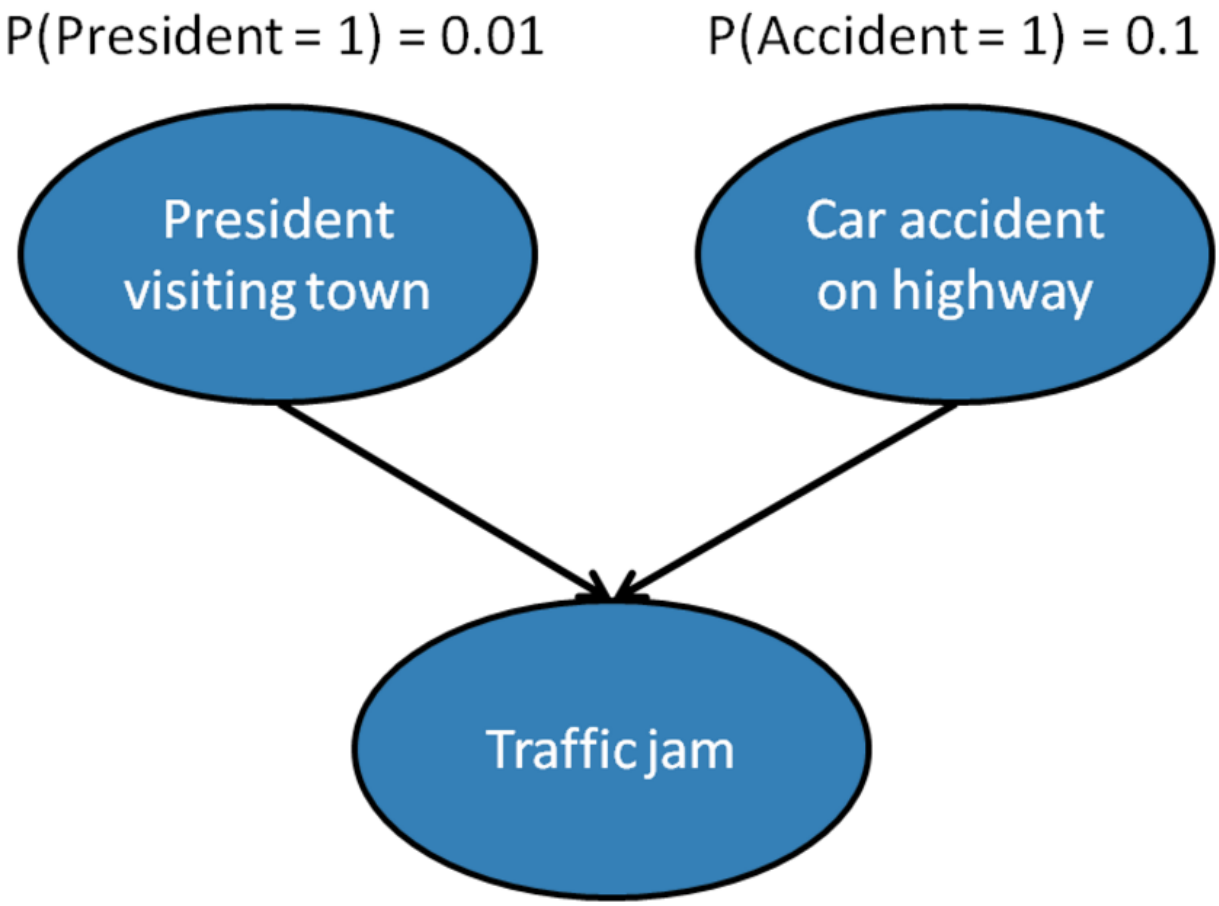
Therefore, we need 1 independent parameter to define the CPD of X and 2 independent parameters to define the CPD of Y.

- 4
- 11
- 8
- 6
- 12
- 7
- 3

Fig. 6: Exercise 01-02-02

3. ***Inter-causal reasoning.**

Consider the following model for traffic jams in a small town, which we assume can be caused by a car accident, or by a visit from the president (and the accompanying security motorcade).



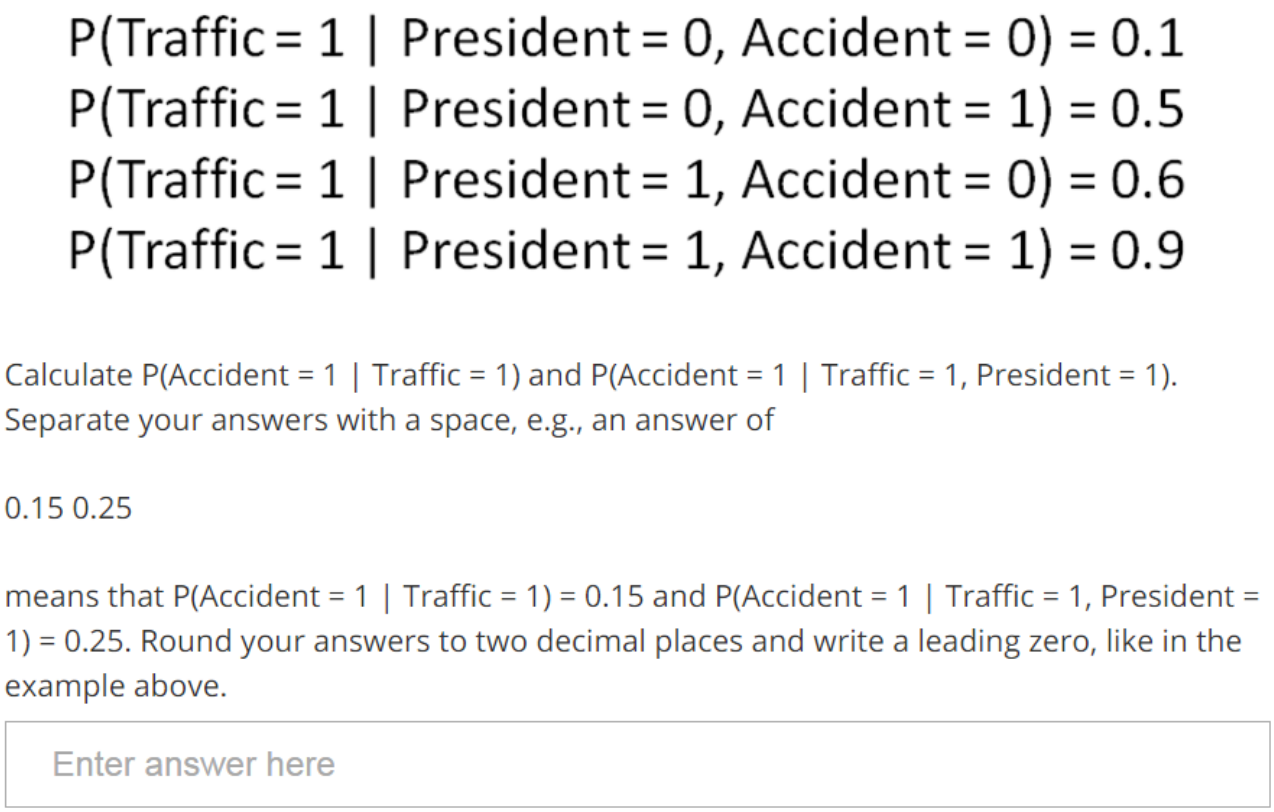


Fig. 7: Exercise 01-02-03

1.2.5 Answers

01-02-01: 2nd;

01-02-02: (3-1)*2*2 = 8;

01-02-03: 0.34 0.14;

$$
\begin{aligned}
P(A=1|T=1) &= \frac{P(A=1,T=1)}{P(T=1)} \\
&= \frac{P(A=1,T=1,P=0)+P(A=1,T=1,P=1)}{P(T=1)} \\
&= \frac{0.1*0.99*0.5+0.9*0.1*0.01}{0.1*0.99*0.5+0.9*0.1*0.01+0.1*0.99*0.9+0.6*0.01*0.9} = 0.34 \\
P(A=1|T=1,P=1) &= \frac{P(A=1,T=1,P=1)}{P(T=1,P=1)} \\
&= \frac{P(A=1,T=1,P=1)}{P(A=0,T=1,P=1)+P(A=1,T=1,P=1)} \\
&= \frac{P(P=1)P(A=1)P(T=1|A=1,P=1)}{P(P=1)P(A=0)P(T=1|A=0,P=1)+P(P=1)P(A=1)P(T=1|A=1,P=1)} \\
&= \frac{0.01*0.1*0.9}{0.01*0.9*0.6+0.01*0.1*0.9} = 0.14
\end{aligned}
$$

1.2.6 Conditional Independence (Chapters 2.1.4, 3.1)

1.2.7 Independencies in Bayesian Networks (Chapter 3.3.1)

(Definition 3.7) Let $\mathbf{X}$,$\mathbf{Y}$,$\mathbf{Z}$ be three set of nodes in $\mathcal{G}$. X and Y are **d-separated** give $\mathbf{Z}$, if there is no active trail between andy node $X \in \mathbf{X}$ and $Y \in \mathbf{Y}$ given $\mathbf{Z}$.

Theorem: If P factorizes over G, and d-$sep_G(\mathbf{X}, \mathbf{Y}|\mathbf{Z})$ then P satisfies $(\mathbf{X} \perp \mathbf{Y}|\mathbf{Z})$;

Any node is d-separated from its non-descendants given its parents; Then from the theorem above, we have: If P factorizes over G, then in P, any variable is independent of its non-descendants given its parents.

Definition: If P satisfies I(G), then G is an **I-map**(independency map) of P; where

$$I(G) = \{(\mathbf{X} \perp \mathbf{Y}|\mathbf{Z}) : d\text{-}sep_G(\mathbf{X}, \mathbf{Y}|\mathbf{Z})\}$$

Factorization $\Rightarrow$ Independence:

Theorem: If P factorizes over G, then G is an I-map for P;

Independence $\Rightarrow$ Factorization:

Theorem: If G is an I-map for P, then P factorizes over G;

Summary:

Two equivalent views of graph structure:

- Factorization: G allows P to be represented;
- I-map: Independencies encoded by G hold in P;

1.2.8 Naive Bayes (Chapter 3.1.3)

What independence assumption does the Naive Bayes model make?

- Given the other observed variables, each observed variable is independent of the class variable.
- Given all of the other observed variables, the class variable is independent of each observed variable.
- Given the class variable, each observed variable is independent of the other observed variables.

Correct
If we observe the class variable, influence cannot flow between any of the other variables, because the only path between two other variables is through the class variable. If we don't observe the class variable, however, all the variables are dependent.

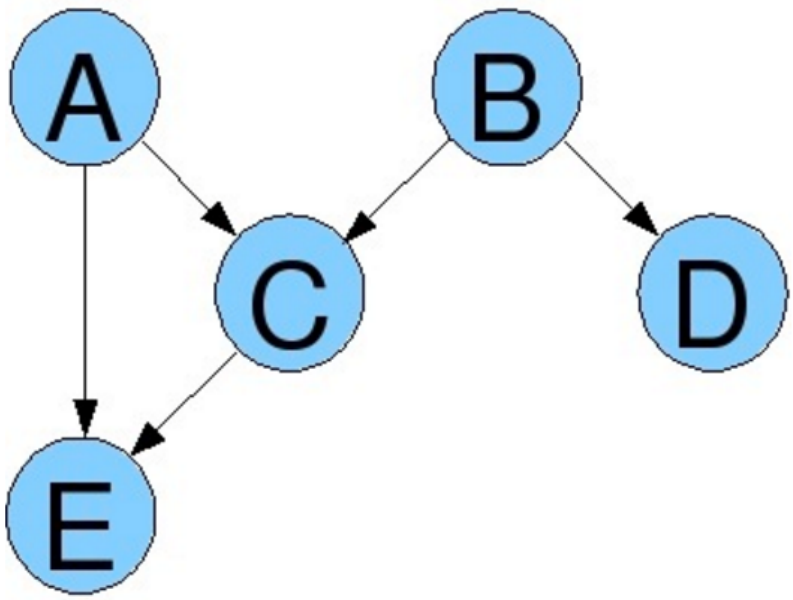
- The Naive Bayes model does not make any independence assumptions.

Fig. 8: Naive Bayes Assumption

1.2.9 Quiz

1. Independencies in a graph.

Which pairs of variables are independent in the graphical model below, given that none of them have been observed? You may select 1 or more options.

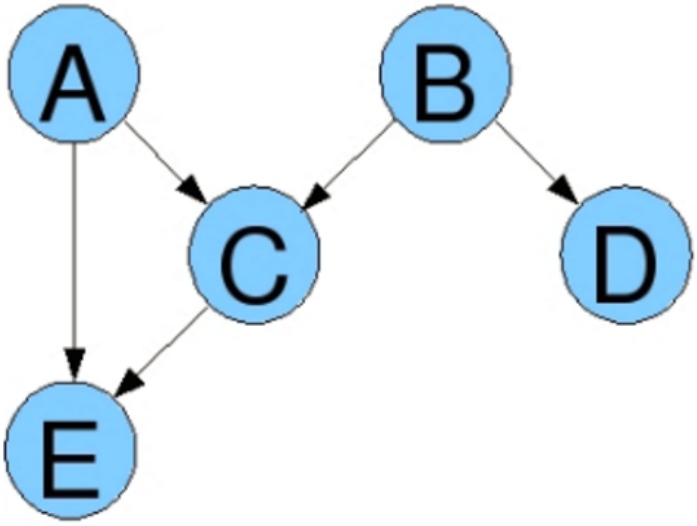


- [ ] A, B
- [ ] A, E
- [ ] D, E
- [ ] A, C
- [ ] None - there are no pairs of independent variables.

Fig. 9: Exercise 01-03-01

2. ***Independencies in a graph.** (An asterisk marks a question that is more challenging. Congratulations if you get it right!)

Now assume that the value of E is known. (E is observed. A, B, C, and D are not observed.) Which pairs of variables (not including E) are independent in the same graphical model, given E? You may select 1 or more options.



- [ ] None - given E, there are no pairs of variables that are independent.
- [ ] A, B
- [ ] A, C
- [ ] A, D
- [ ] B, D
- [ ] D, C
- [ ] B, C

Fig. 10: Exercise 01-03-02

3. **I-maps.** I-maps can also be defined directly on graphs as follows. Let $I(G)$ be the set of independencies encoded by a graph G. Then G_1 is an I-map for G_2 if $I(G_1) \subseteq I(G_2)$.

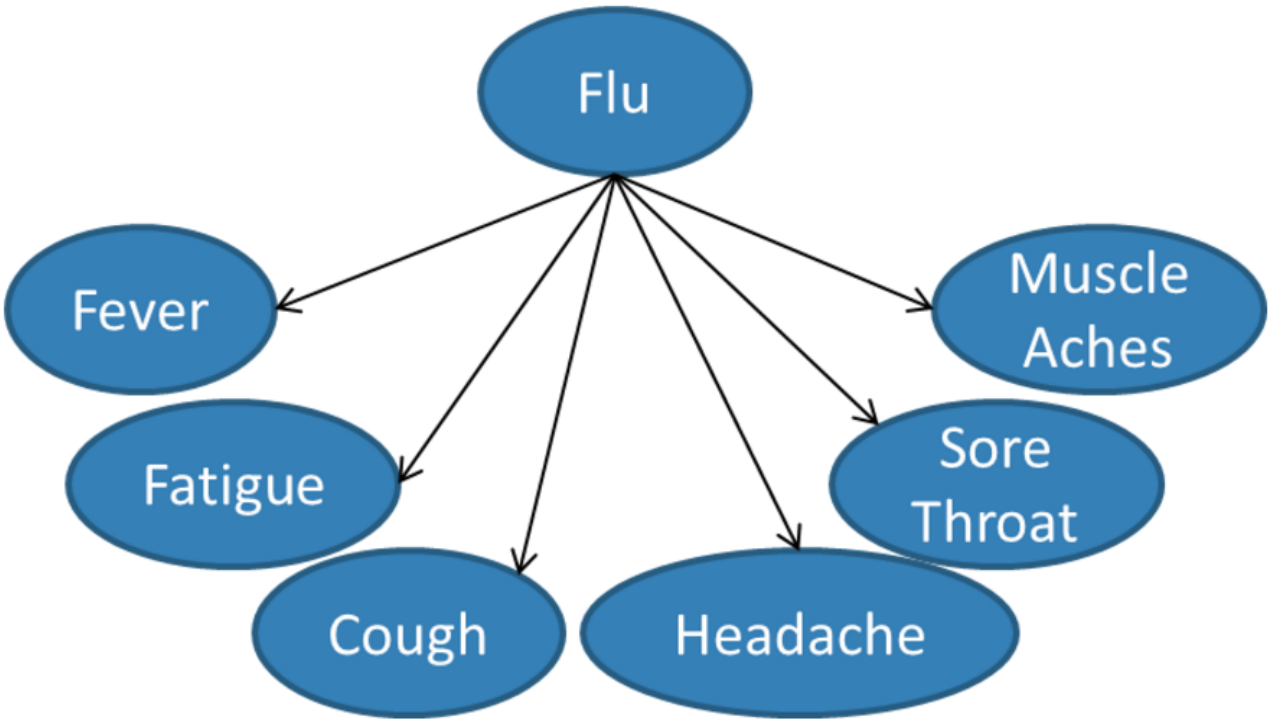
Which of the following statements about I-maps are true? You may select 1 or more options.

- [ ] A graph K is an I-map for a graph G if and only if all of the independencies encoded by K are also encoded by G.
- [ ] A graph K is an I-map for a graph G if and only if K encodes all of the independences that G has and more.
- [ ] An I-map is a function f that maps a graph G to itself, i.e., $f(G) = G$.
- [ ] The graph K that is the same as the graph G, except that all of the edges are oriented in the opposite direction as the corresponding edges in G, is always an I-map for G, regardless of the structure of G.
- [ ] I-maps are Apple's answer to Google Maps.

Fig. 11: Exercise 01-03-03

4. ***Naive Bayes.**

Consider the following Naive Bayes model for flu diagnosis:



Assume a population size of 10,000. Which of the following statements are true in this model? You may select 1 or more options.

- [ ] Say we observe that 1000 people have the flu, out of which 500 people have a headache (and possibly other symptoms) and 500 have a fever (and possibly other symptoms).

  We would expect that approximately 250 people with the flu also have both a headache and fever.

- [ ] Say we observe that 1000 people have a headache (and possibly other symptoms), out of which 500 people have the flu (and possibly other symptoms), and 500 people have a fever (and possibly other symptoms).

  We would expect that approximately 250 people with a headache also have both the flu and a fever.

- [ ] Say we observe that 1000 people have the flu, out of which 500 people have a headache (and possibly other symptoms) and 500 have a fever (and possibly other symptoms).

  We can conclude that exactly 250 people with the flu also have both a headache and fever.

- [ ] Say we observe that 1000 people have the flu, out of which 500 people have a headache (and possibly other symptoms) and 500 people have a fever (and possibly other symptoms).

  Without more information, we cannot estimate how many people with the flu also have both a headache and fever.

Fig. 12: Exercise 01-03-04

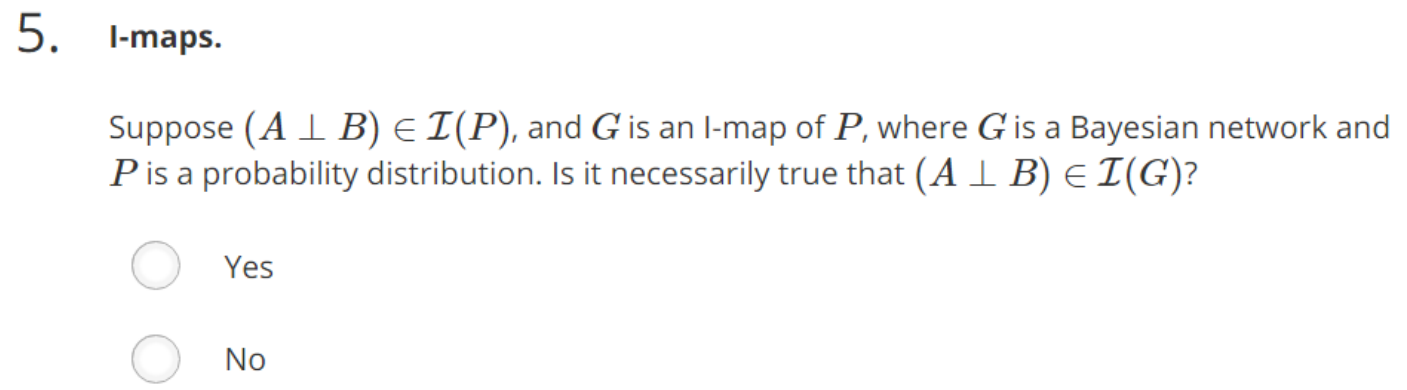


Fig. 13: Exercise 01-03-05

1.2.10 Answers

01-03-01: 1st;

D,E: There is an active trail connecting D and E that goes through B and C;

01-03-02: 1st;

A,D: Observing E activates the V-structures around C and E. Hence, influence can flow from A to B through C, and therefore from A to D through C and B.

D,C: Influence can flow along the active trail $D \leftarrow B \rightarrow C$.

01-03-03: 1st;

For the 4th item: This is not always true; consider the V-structure $A \rightarrow B \leftarrow C$.

01-03-04: 1st;

$$P(Headache = 1, Fever = 1|Flu = 1) = P(Headache = 1|Flu = 1) \times P(Fever = 1|Flu = 1)$$

01-03-05: No;

Since G is an I-map of P, all independencies in G are also in P. However, this doesn't mean that all independencies in P are also in G. An easy way to remember this is that **the complete graph**, which has no independencies, **is an I-map of all distributions**.

1.2.11 Bayesian Networks: Knowledge Engineering

1.2.11.1 Application - Medical Diagnosis Chapter 3.2: Box 3.D (p. 67)

1.3 Template Models for Bayesian Networks

1.3.1 Overview (Chapter 6.1)

Template Variables is instantiated(duplicated) multiple times;

Template Models is a Language that tell us how template variables can be the dependency models from template.

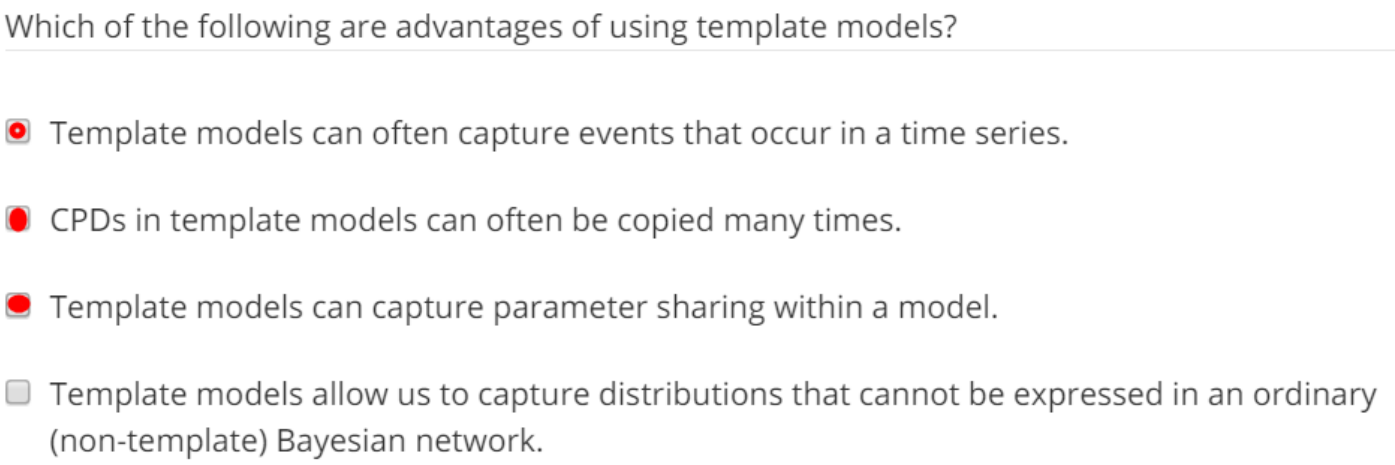


Fig. 14: Template Models

Template models are a convenient way of representing Bayesian networks that have a high amount of parameter sharing and structure. However, they are merely compact representations of a fully unrolled Bayesian network, thus have no additional representative powers.

- Temperal Models: for dealing with temporal processes, for example, where we have replication over time.
  - Dynamic Bayesian networks (DBN); Hidden Markov Models (HMM);
- Object relation models:
  - Directed Models: Plate models;
  - Undirected Models;

1.3.2 Temporal Models - DBNs (Chapters 6.2, 6.3)

Markov Assumption;

Time Invariance, so we can replicated the **Template probability model** for ALL t;

2-time-slice Bayesian Network (2TBN);

Dynamic Bayesian Network (DBN);

1.3.3 Temporal Models - HMMs (Chapters 6.2, 6.3)

Applications:

- Robot localization
- Speech recognition
- Biological sequence analysis
- Text annotation

1.3.4 Plate Models (Chapters 6.4.1)

1.3.5 Quiz

1. Markov Assumption.

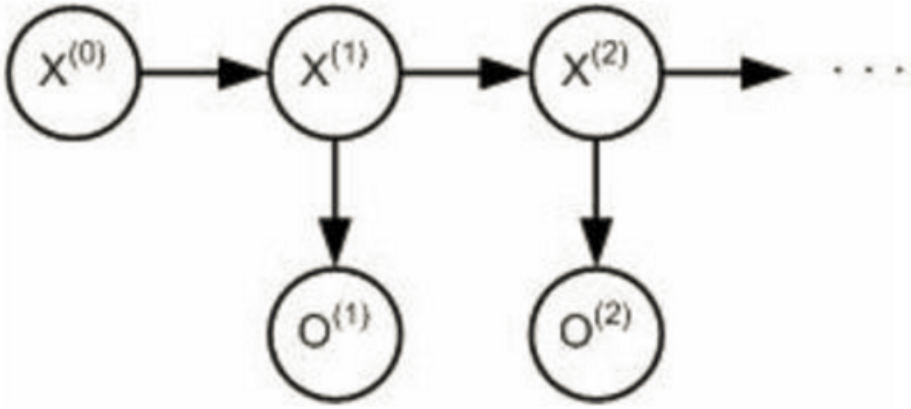
If a dynamic system X satisfies the Markov assumption for all time $t \geq 0$, which of the following statements must be true? You may select 1 or more options.

- ☐ $(X^{(t+1)} \perp X^{(0:(t-1))})$
- ☐ $(X^{(t+1)} \perp X^{(t)})$
- ☐ $(X^{(t+1)} \perp X^{(0:(t-1))} | X^{(t)})$

Fig. 15: Exercise 01-04-01

2. Independencies in DBNs.

In the following DBN, which of the following independence assumptions are true? You may select 1 or more options.



- ☐ $(O^{(t)} \perp O^{(t-1)} \mid X^{(t)})$
- ☐ $(O^{(t)} \perp X^{(t+1)} \mid X^{(t)})$
- ☐ $(O^{(t)} \perp X^{(t-1)} \mid X^{(t)})$
- ☐ $(O^{(t)} \perp O^{(t-1)})$

Fig. 16: Exercise 01-04-02

3. **Applications of DBNs.**

For which of the following applications might one use a DBN (i.e. the Markov assumption is satisfied)? You may select 1 or more options.

- ☐ Predicting the probability that today will be a snow day (school will be closed because of the snow), when this probability depends only on whether yesterday was a snow day.
- ☐ Modeling time-series data, where the events at each time-point are influenced by only the events at the one time-point directly before it
- ☐ Modeling the behavior of people, where a person's behavior is influenced by only the behavior of people in the same generation and the people in his/her parents' generation.
- ☐ Modeling data taken at different locations along a road, where the data at each location is influenced by only the data at the same location and at the location directly to the East

Fig. 17: Exercise 01-04-03

4. **Plate Semantics.**

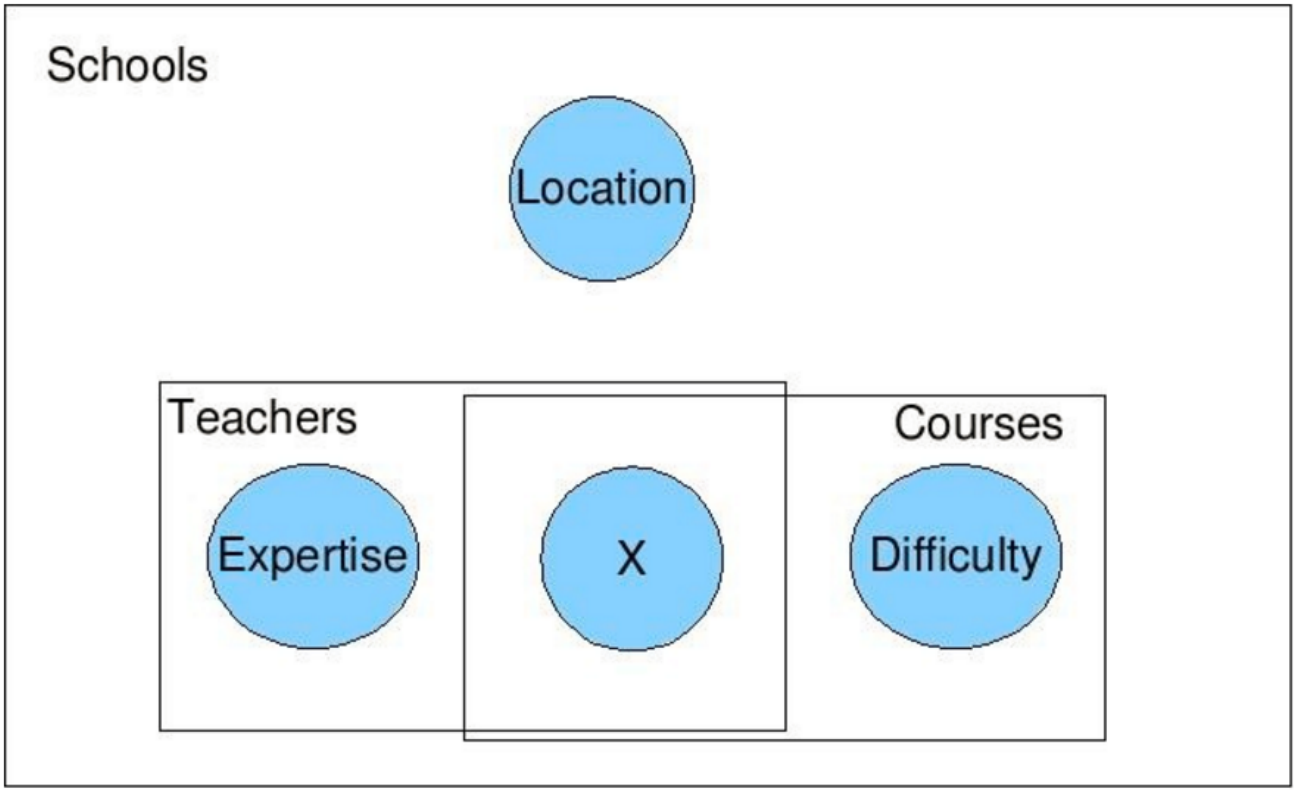
"Let A and B be random variables inside a common plate indexed by i. Which of the following statements must be true? You may select 1 or more options.

- ☐ There is an instance of A and an instance of B for every i.
- ☐ If there is an instance of A for some i, then there is no instance of B for that i.
- ☐ For each i, A(i) and B(i) have the same CPDs.
- ☐ For each i, A(i) and B(i) have edges connecting them to the same variables outside of the plate.

Fig. 18: Exercise 01-04-04

5. ***Plate Interpretation.**

Consider the plate model below (with edges removed). Which of the following might a given instance of X possibly represent in the grounded model? (You may select 1 or more options. Keep in mind that this question addresses the variable's semantics, not its CPD.)



- [ ] Whether a teacher with expertise E taught a course of difficulty D
- [ ] Whether a specific teacher T taught a specific course C at school S
- [ ] Whether someone with expertise E taught something of difficulty D at school S
- [ ] Whether a specific teacher T is a tough grader
- [ ] None of these options can represent X in the grounded model

Fig. 19: Exercise 01-04-05

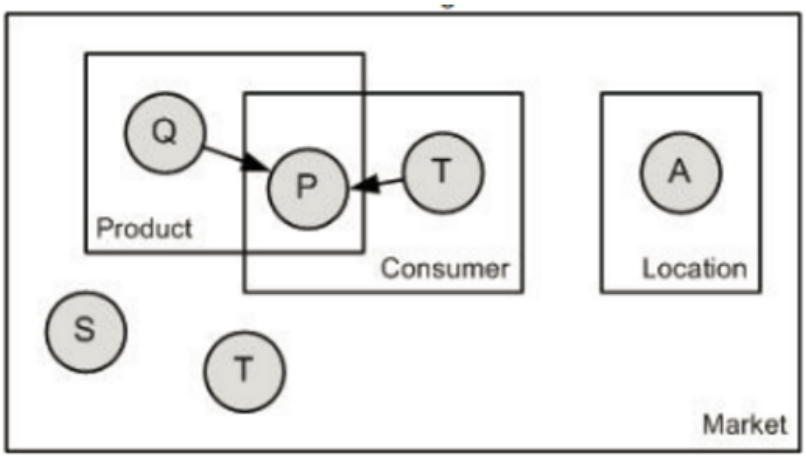
6. **Grounded Plates.**

Using the same plate model, now assume that there are s schools, t teachers in each school, and c courses taught by each teacher. How many instances of the Location variable are there?

- s
- s^2
- stc
- t

Fig. 20: Exercise 01-04-06

7. **Template Models.** Consider the plate model shown below. Assume we are given K Markets, L Products, M Consumers and N Locations. What is the total number of instances of the variable P in the grounded BN?



- $K \cdot L \cdot M$
- $K \cdot L \cdot M \cdot N$
- $(L \cdot M)^K$
- $K \cdot (N + (L \cdot M))$

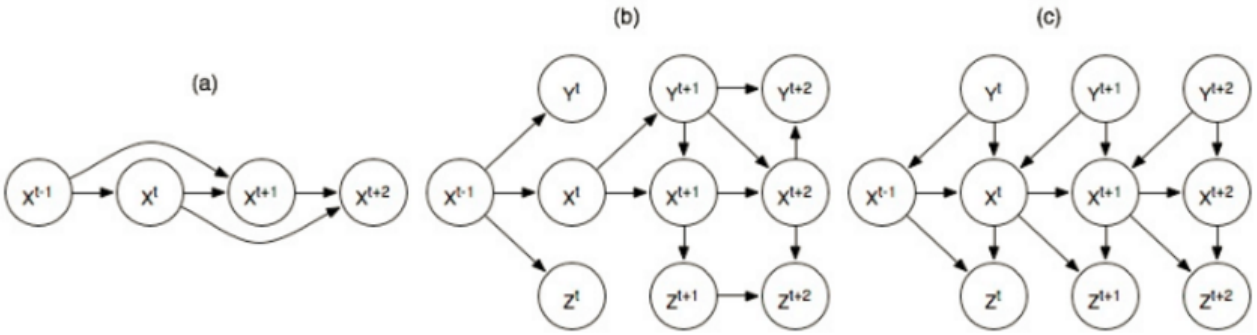
Fig. 21: Exercise 01-04-07

8. **Template Models.** Consider the plate model from the previous question. What might P represent?

- Whether a specific product PROD was consumed by consumer C in market M
- Whether a specific product PROD was consumed by consumer C in market M that is supervised by supervisor S (assuming that there is exactly 1 unique supervisor per market) and has target audience T (assuming that there is exactly 1 unique target audience per market)
- Whether a specific product PROD was consumed by consumer C in all markets
- Whether a specific product of brand q was consumed by a consumer with age t in a market of type m

Fig. 22: Exercise 01-04-08

9. **Time-Series Graphs.** Which of the time-series graphs satisfies the Markov assumption? You may select 1 or more options.

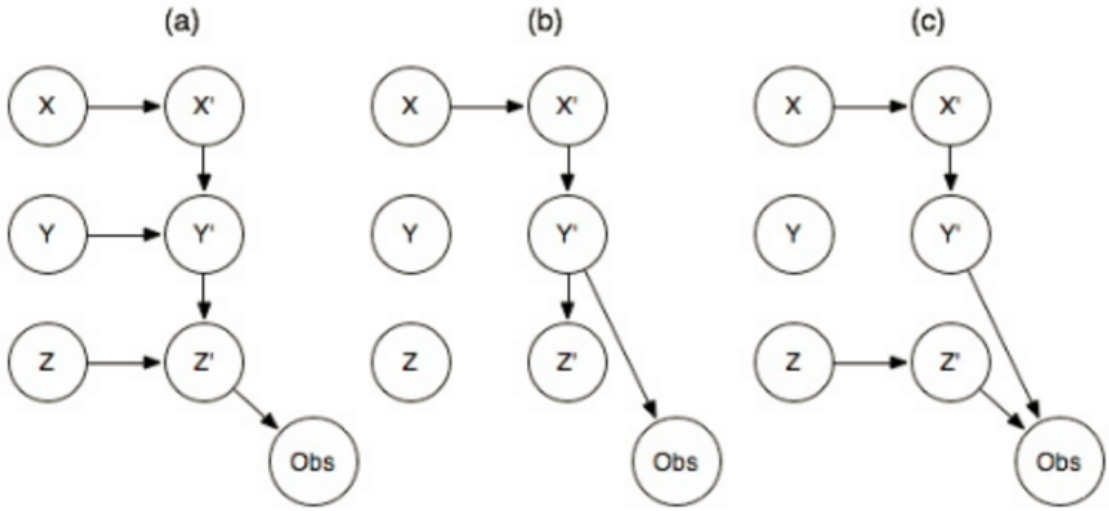


- [ ] (a)
- [ ] (b)
- [ ] (c)

Fig. 23: Exercise 01-04-09

10. ***Unrolling DBNs.** Below are 2-TBNs that could be unrolled into DBNs. Consider these unrolled DBNs (note that there are no edges within the first time-point). In which of them will $(X^{(t)} \perp Z^{(t)} \mid Y^{(t)})$ hold for all t, assuming $Obs^{(t)}$ is observed for all t and $X^{(t)}$ and $Z^{(t)}$ are never observed? You may select 1 or more options.

Hint: Unroll these 2-TBNs into DBNs that are at least 3 time steps long (i.e., involving variables from $t-1, t, t+1$).



- [ ] (a)
- [ ] (b)
- [ ] (c)

Fig. 24: Exercise 01-04-10

1.3.6 Answers

01-04-01: 3th;

01-04-02: 1st,2nd,3th;

$$(O^t \perp O^{t-1}) is wrong because there is an active trail from O^t to O^{t-1} through X^t and X^{t-1}.$$

01-04-03: 1st,2nd,3th,4th;

01-04-04: 1st;

01-04-05: 2nd;

For 3: In the grounded model, there will be an instance of X for each combination of Teacher and Class, and there is a combination like this for each School. Thus, we are looking at a random variable that will say

something about a specific teacher, class, and school combination, not a particular expertise, difficulty, and school combination.

01-04-06: s;

01-04-07: KLM;

For 4: P is not in the location plate.

01-04-08: 4th;

For 2: S and T do not have arrows to P, and P is not in a supervisor or target audience plate.

01-04-09: (b);

01-04-10: (b);

1.4 Structured CPDs for Bayesian Networks

1.4.1 Overview (Chapter 5.1, 5.2)

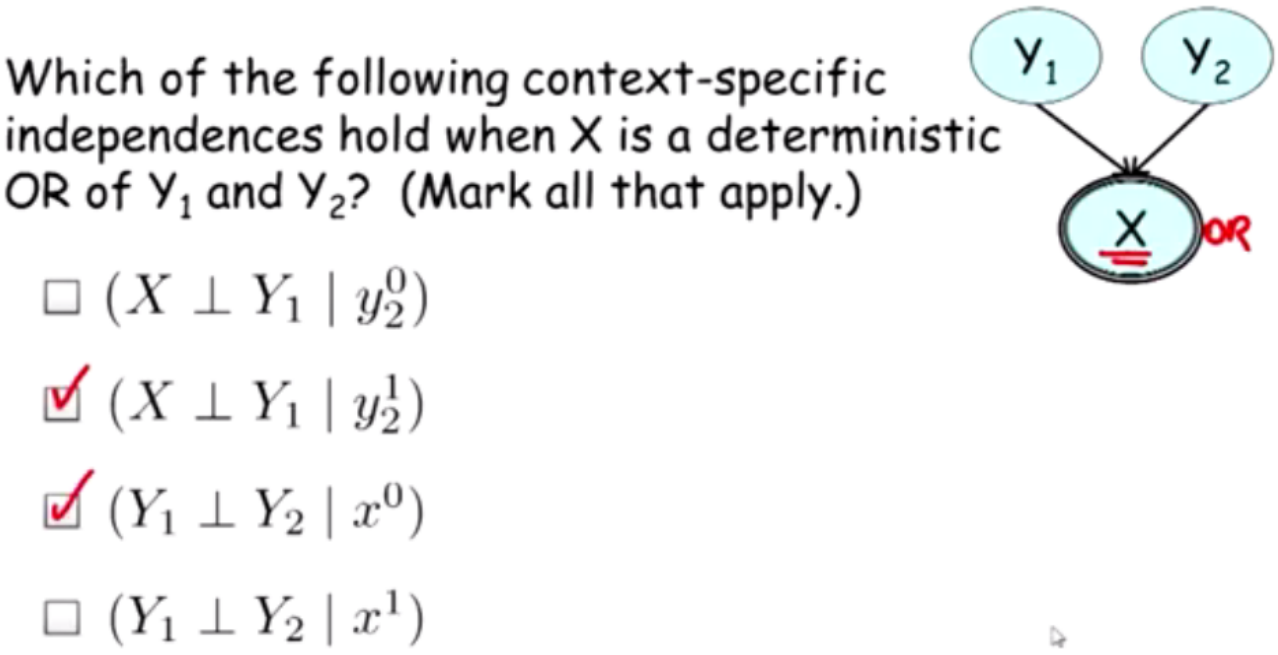


Fig. 25: An example of Context-Specific Independence

1.4.2 Tree-Structured CPDs (Chapter 5.3)

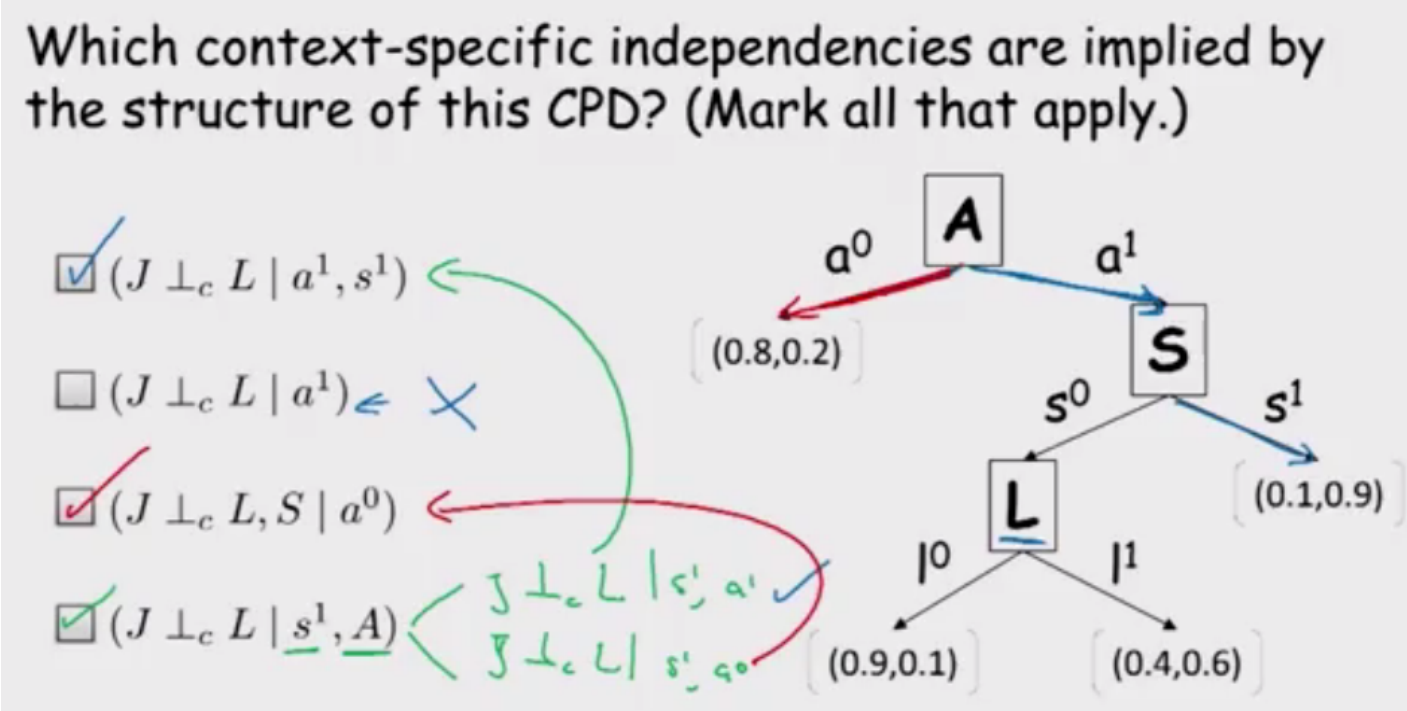


Fig. 26: An example of Context-Specific Independence for Tree Structured CPDs

1.4.3 Independence of Causal Influence (Chapter 5.4)

1.4.4 Continuous Variables (Chapter 5.5)

Linear Gaussian;

Conditional Linear Gaussian;

Nonlinear Gaussian - Robot Localization, Robot Motion Model;

1.4.5 Quiz

1. **Causal Influence.** Consider the CPD below. What is the probability that $E = e_0$ in the following graph, given an observation $A = a_0, B = b_0, C = c_1, D = d_0$? Note that, for the pairs of probabilities that make up the leaves, the probability on the left is the probability of e_0, and the probability on the right is the probability of e_1.

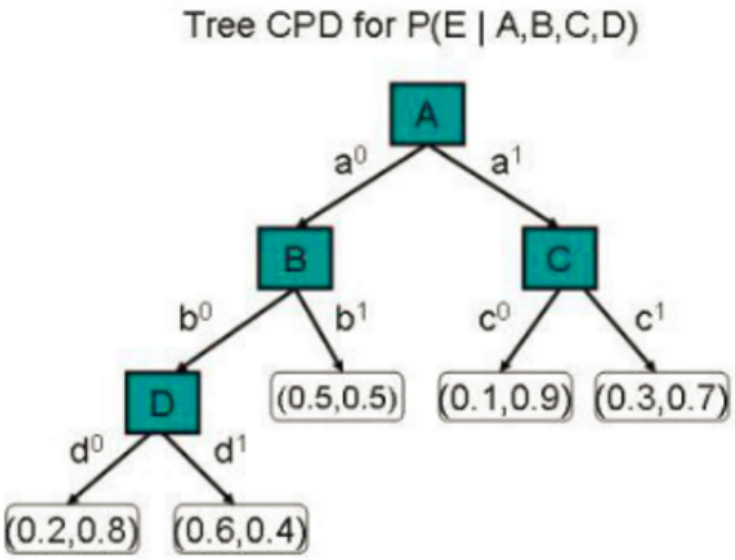
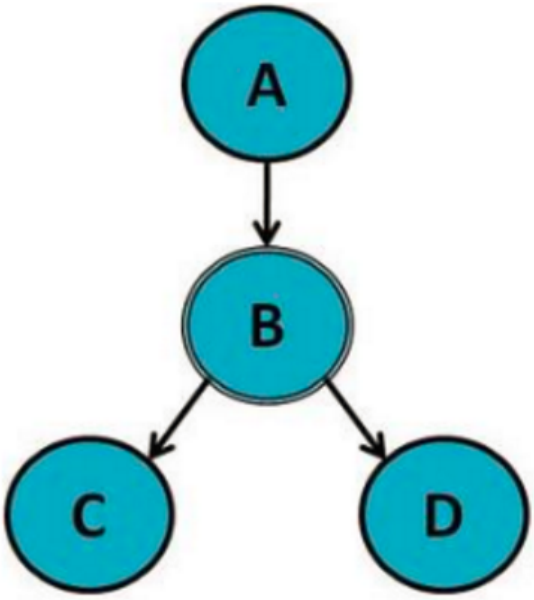


Fig. 27: Exercise 01-05-01

2. **Independencies with Deterministic Functions.** In the following Bayesian network, the node B is a deterministic function of its parent A. Which of the following is an independence statement that holds in the network? You may select 1 or more options.



- ☐ $(A \perp D \mid B)$
- ☐ $(C \perp D \mid A)$
- ☐ $(C \perp D \mid B)$
- ☐ $(A \perp D \mid C)$

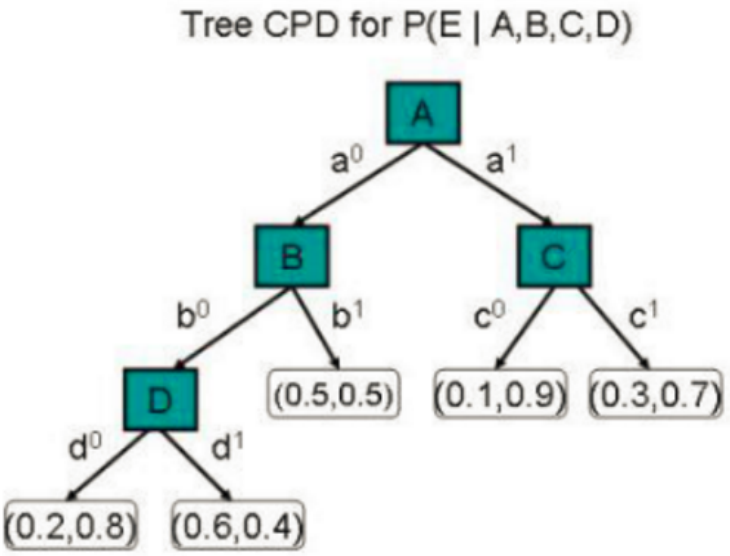
Fig. 28: Exercise 01-05-02

3. **Independencies in Bayesian Networks.** For the network in the previous question, let B no longer be a deterministic function of its parent A. Which of the following is an independence statement that holds in the modified Bayesian network? You may select 1 or more options.

- ☐ $(A \perp D \mid C)$
- ☐ $(A \perp D \mid B)$
- ☐ $(A \perp B \mid C, D)$
- ☐ $(C \perp D \mid A)$

Fig. 29: Exercise 01-05-03

4. **Context-Specific Independencies in Bayesian Networks.** Which of the following are context-specific independences that **do** exist in the tree CPD below? (Note: Only consider independencies in this CPD, ignoring other possible paths in the network that are not shown here. You may select 1 or more options.)



- ☐ $(E \perp_c D \mid b^1)$
- ☐ $(A \perp_c D \mid B)$
- ☐ $(E \perp_c C \mid a^0, b^0)$
- ☐ $(E \perp_c C \mid b^0, d^0)$

Fig. 30: Exercise 01-05-04

1.4.6 Answers

01-05-01: $P(E = e_0|A = a_0, B = b_0, C = c_1, D = d_0) = 0.2$

01-05-02: 1st,2nd,3rd;

01-05-03: 2nd;

01-05-04: 1st,3rd;

For 2: This option is wrong because the tree CPD represents P(E—A,B,C,D), so it does not give any information about whether A and D are independent.

A variable X is independent of E given conditioning assignments z if all paths consistent with z traversed in the tree CPD reach a leaf without querying X.

1.5 Markov Network (Undirected Models)

1.5.1 Pairwise Markov Networks (Chapter 4.1)

1.5.2 General Gibbs Distribution (Chapter 4.2.2)

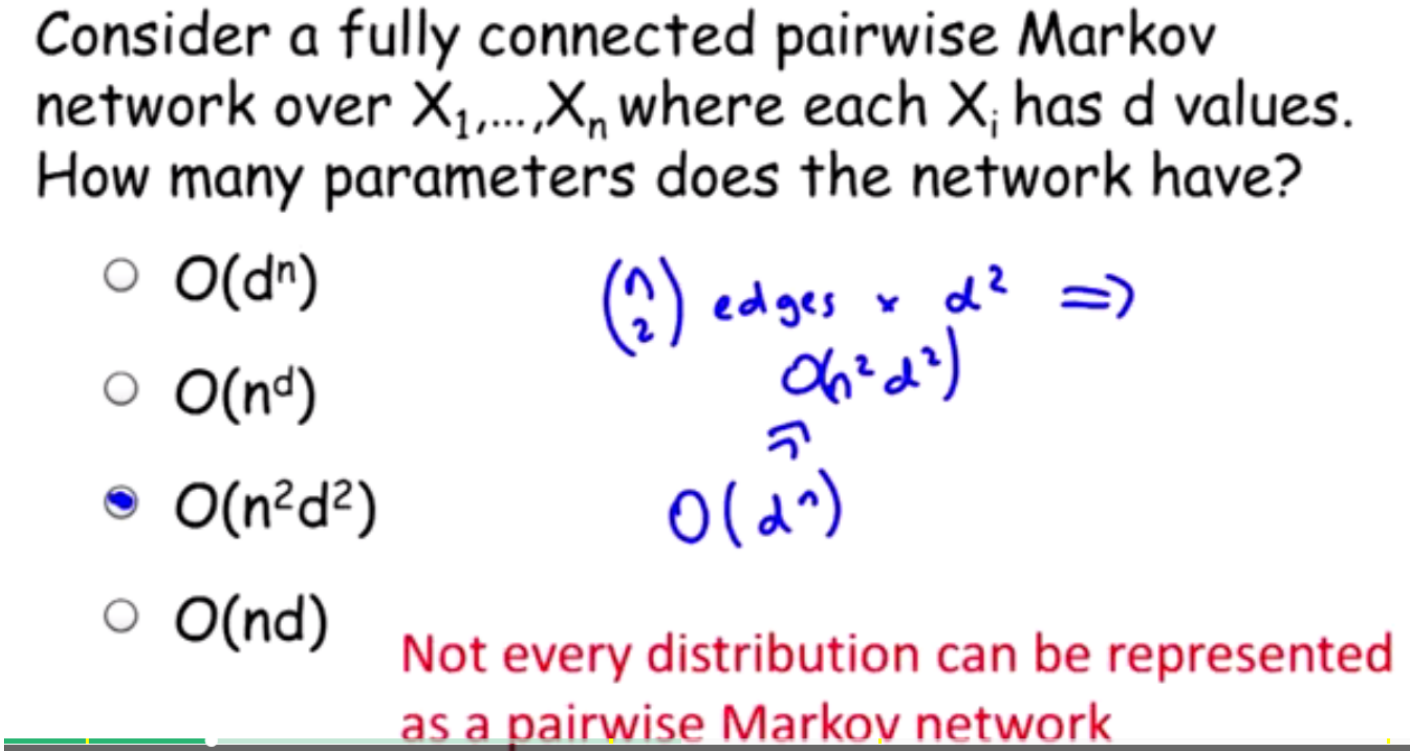
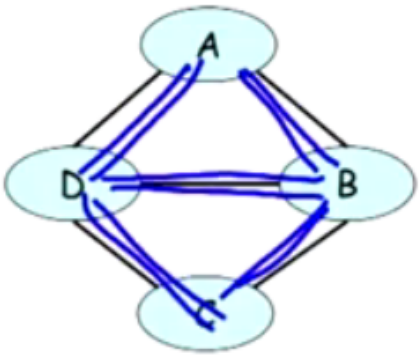


Fig. 31: Why General Gibbs Distribution

Which Gibbs distribution would induce the graph H?



- ✓ $\phi_1(A, B, D), \phi_2(B, C, D)$
- ✓ $\phi_1(A, B), \phi_2(B, C), \phi_3(C, D), \phi_4(A, D), \phi_5(B, D)$
- ✓ $\phi_1(A, B, D), \phi_2(B, C), \phi_3(C, D)$
- ○ All of the above

Cannot read the factorization from the graph

Fig. 32: Gibbs distribution factoration

1.5.3 Conditional Random Fields (Chapter 4.6.1)

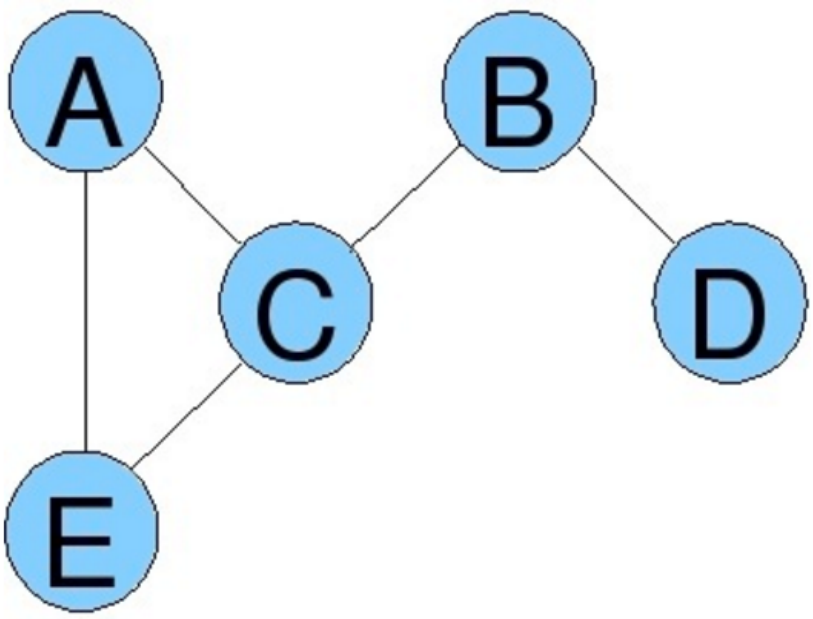
1.5.4 Quiz

1. **Factor Scope.** Let $\phi(c, e)$ be a factor in a graphical model, where c is a value of C and e is a value of E. What is the scope of ϕ?

   - ○ {A, B, C, E}
   - ○ {A, C, E}
   - ○ {C, E}
   - ○ {C}

Fig. 33: Exercise 01-06-01

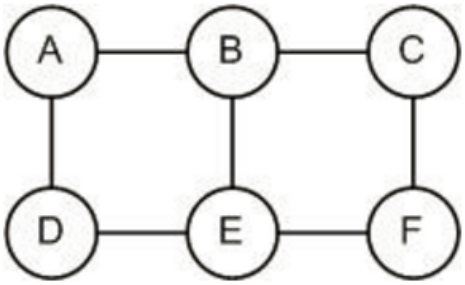
2. **Independence in Markov Networks.** Consider this graphical model from week 1's quizzes. This time, all of the edges are undirected (see modified graph below). Which pairs of variables are independent in this network? You may select 1 or more options.



- [ ] No pair of variables are independent on each other.
- [ ] D, E
- [ ] A, D

Fig. 34: Exercise 01-06-02

3. **Factorization.** Which of the following is a valid Gibbs distribution over this graph?



- $\frac{\phi(A,B,D)\times\phi(C,E,F)}{Z}$, where Z is the partition function
- $\phi(A,B,D)\times\phi(C,E,F)$
- $\phi(A)\times\phi(B)\times\phi(D)\times\phi(E)\times\phi(F)$
- $\frac{\phi(A)\times\phi(B)\times\phi(C)\times\phi(D)\times\phi(E)\times\phi(F)}{Z}$, where Z is the partition function

Fig. 35: Exercise 01-06-03

4. **Factors in Markov Network.** Let $\phi(A,B,C)$ be a factor in a probability distribution that factorizes over a Markov network. Which of the following must be true? You may select 1 or more options.

- [ ] A, B, and C do not form a clique in the network.
- [ ] A, B, and C form a clique in the network.
- [ ] $\phi(a,b,c)\geq 0$, where a is a value of A, b is a value of B, and c is a value of C.
- [ ] There is no path from A to B, no path from B to C, and no path from A to C in the network.
- [ ] There is a path from A to B, a path from B to C, and a path from A to C in the network.

Fig. 36: Exercise 01-06-04

1.5.5 Answers

01-06-01: {C,E};

01-06-02: No pairs of variables are independent in a fully connected Markov network;

01-06-03: 4th;

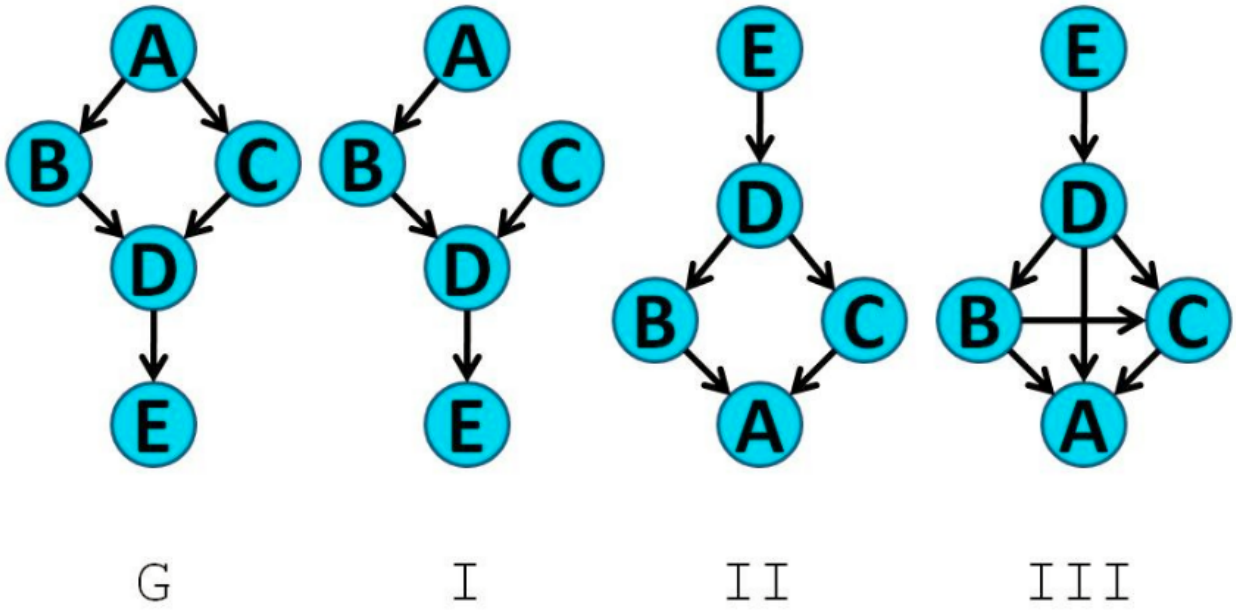
01-06-04: 2nd,3rd,5th;

1.5.6 Independencies in Markov Networks (Chapter 4.3.1)

1.5.7 I-Maps and Perfect Maps (Chapter 3.3.4)

1.5.8 Quiz

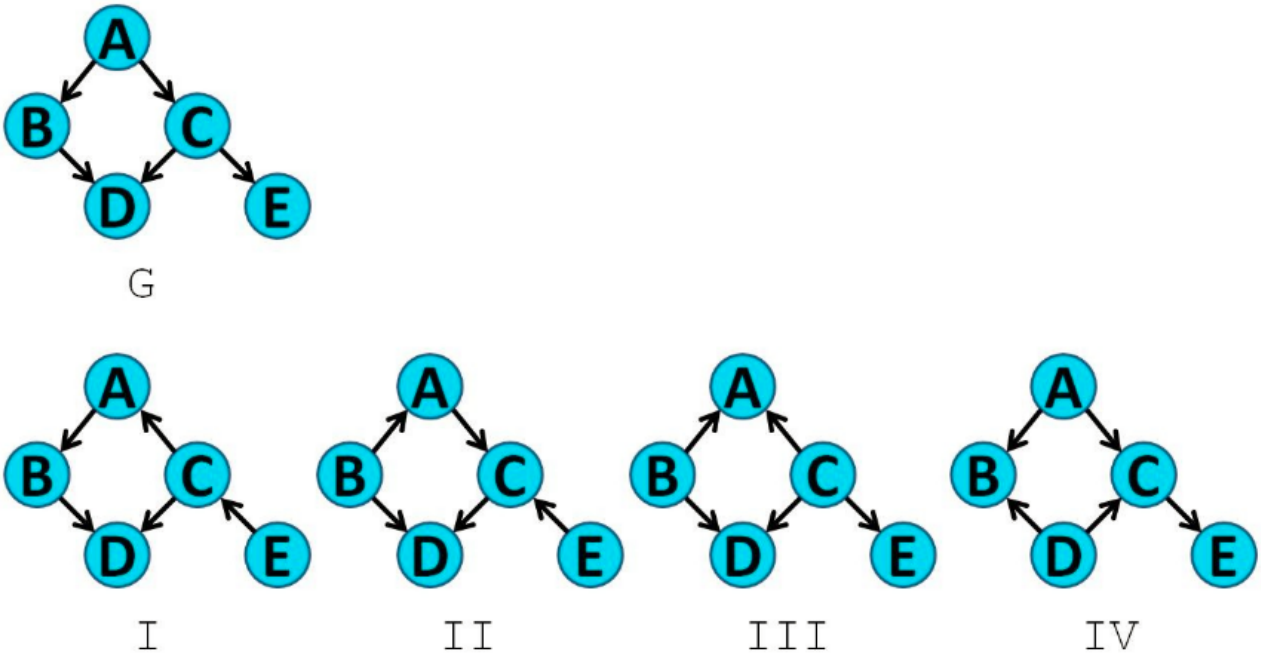
1. **I-Maps.** Graph G is a perfect I-map for distribution P, i.e. $\mathcal{I}(G) = \mathcal{I}(P)$. Which of the other graphs is a **perfect** I-map for P?



- None of the above
- II
- I
- I and III

Fig. 37: Exercise 01-07-01

2. **I-Equivalence.** In the figure below, graph G is I-equivalent to which other graph(s)?



- I
- I and III
- None of the above
- I and IV

Fig. 38: Exercise 01-07-02

3. ***I-Equivalence.** Let Bayesian network G be a simple directed chain $X_1 \to X_2 \to \ldots \to X_n$ for some number n. How many Bayesian networks are I-equivalent to G including G itself?

- n
- $2n-1$
- 0
- 1

Fig. 39: Exercise 01-07-03

1.5.9 Answers

01-07-01: None of the above;

I isn't because it has the extra independence $(A \perp C)$.

II has the extra independence relation $(B \perp C|D)$(among others).

III has no extra independencies but does not preserve an independence relationship in G.

01-07-02: I;

IV isn't because it has the extra independence relation $(B \perp C|D)$.

01-07-03: n;

The chain $X_1 \leftarrow ... \leftarrow X_i \rightarrow ... \rightarrow X_n$ is I-equivalent, where i can be 2 through n (when i=n, all arrows point left). Thus there are n-1 I-equivalent networks like this. Including the original network makes n.

1.5.10 Log-Linear Models (Chapter 4.4, p. 125)

1.5.11 Shared Features in Log-Linear Models (Chapter 4: Box 4.B (p. 112), Box 4.C (p. 126), Box 4.D (p. 127))

1.6 Decision Making

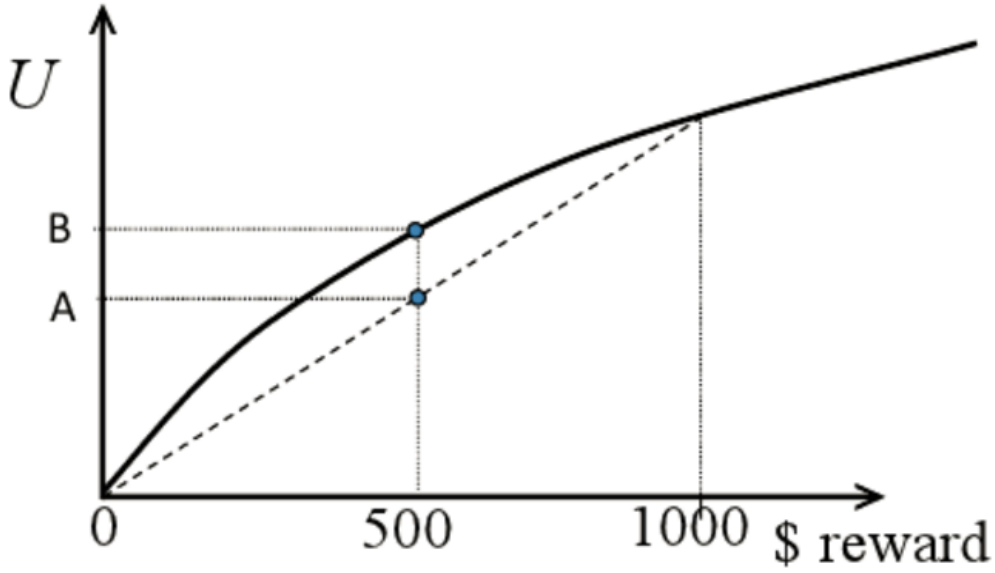
1.6.1 Maximum Expected Utility (Chapter 22.1.1, 23.2.104, 23.4.1-2, 23.5.1)

1.6.2 Utility Functions (Chapter 22.2.1-3, 22.3.2, 22.4.2)

1.6.3 Value of Perfect Information (Chapter 23.7.1-2)

1.6.4 Quiz

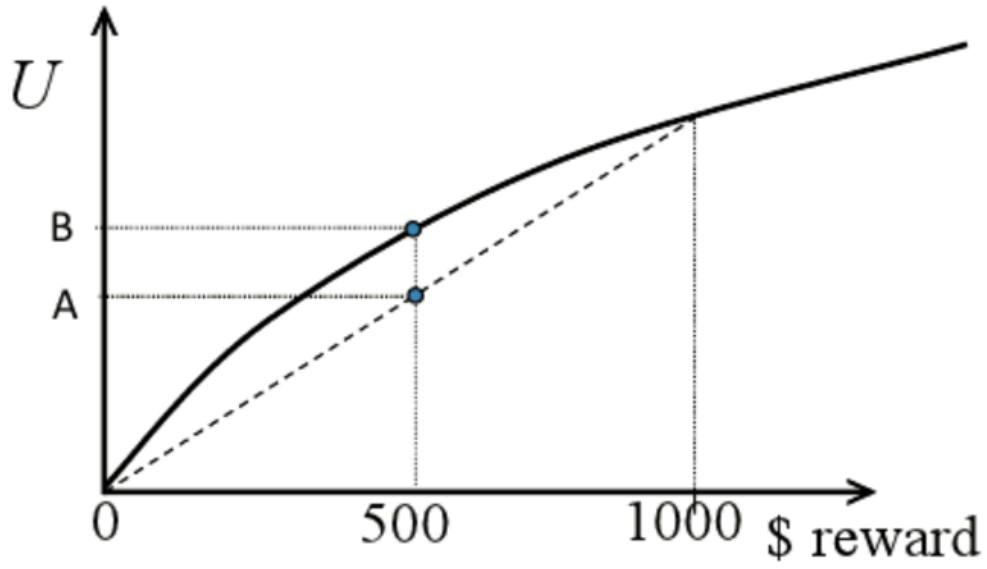
1. **Utility Curves.** What does the point marked A on the Y axis correspond to? (Mark all that apply.)



- ☐ $U(\$500)$
- ☐ $0.5U(\$0) + 0.5U(\$1000)$
- ☐ $500
- ☐ $U(\ell)$ where ℓ is a lottery that pays $0 with probability 0.5 and $1000 with probability 0.5.

Fig. 40: Exercise 01-08-01

2. **Utility Curves.** What does the point marked B on the Y axis correspond to? (Mark all that apply.)

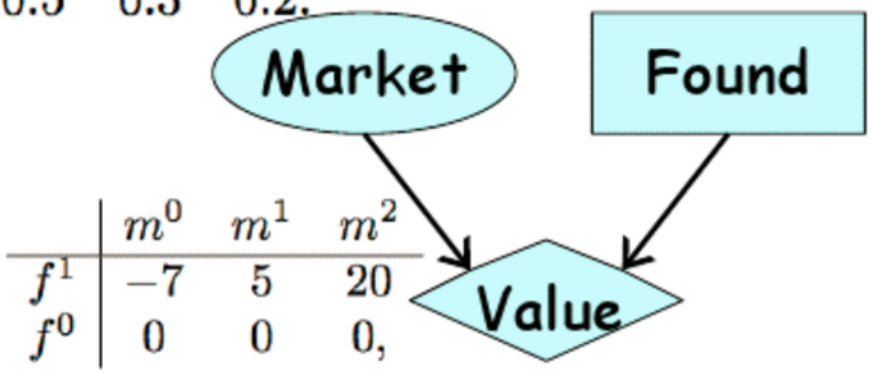


- [ ] $U(\ell)$ where ℓ is a lottery that pays \$0 with probability 0.5 and \$1000 with probability 0.5.
- [ ] $0.5U(\$0) + 0.5U(\$1000)$
- [ ] $U(\$500)$
- [ ] $\$500$

Fig. 41: Exercise 01-08-02

3. **Expected Utility.** In the simple influence diagram on the right, with the CPD for M and the utility function V, what is the expected utility of the action f^1?

m^0	m^1	m^2
0.5	0.3	0.2,

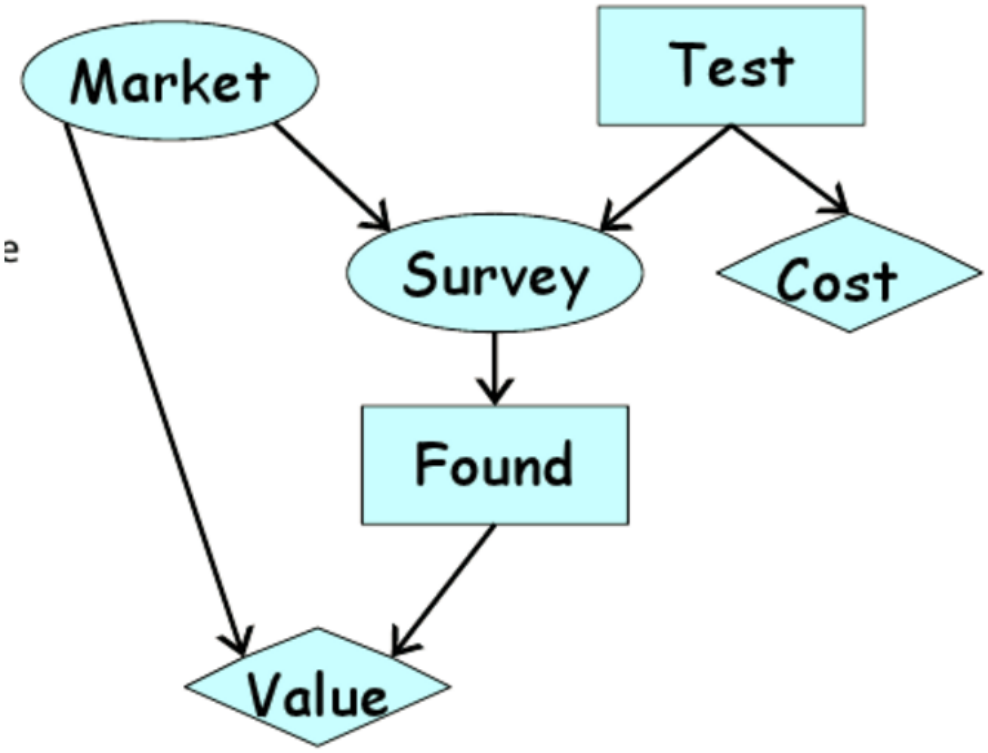


	m^0	m^1	m^2
f^1	-7	5	20
f^0	0	0	0,

- 20
- 2
- 0
- 5

Fig. 42: Exercise 01-08-03

4. ***Uninformative Variables.** In the influence diagram on the right, what is an appropriate way to have the model account for the fact that if the Test wasn't performed (t^0), then the survey is uninformative?



- Set $P(S|M,t^0)$ to be uniform.
- Set $P(S|M,t^0)$ so that S takes the value s^0 with probability 1.
- Set $P(S|M,t^0)$ so that S takes some new value "not performed" with probability 1.
- Set $P(S|M,t^0) = P(S|M,t^1)$.

Fig. 43: Exercise 01-08-04

1.6.5 Answers

01-08-01: 2nd,4th;

01-08-02: 3rd;

01-08-03: 2;

01-08-04: 3rd;

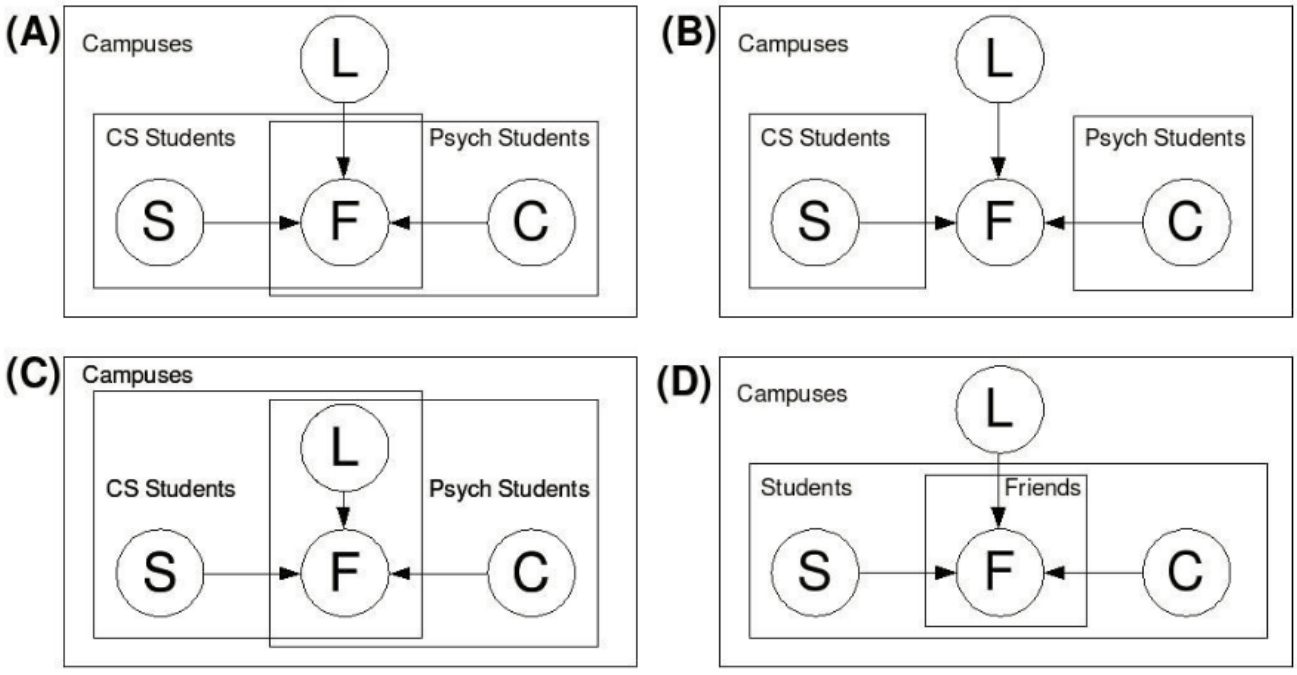
This is the appropriate action. Assigning S to any other value would not be desirable, as these other values may represent survey results, but we have not actually conducted the survey.

1.7 Knowledge Engineering & Summary

1.7.1 Quiz

1. **Template Model Representation.** Consider the following scenario:

On each campus there are several Computer Science students and several Psychology students (each student belongs to one xor the other group). We have a binary variable L for whether the campus is large, a binary variable S for whether the CS student is shy, a binary variable C for whether the Psychology student likes computers, and a binary variable F for whether the Computer Science student is friends with the Psychology student. Which of the following plate models can represent this scenario?



- (B)
- (A)
- (C)
- None of these plate models can represent this scenario

Fig. 44: Exercise 01-09-01

2. **Partition Function.** Which of the following is a use of the partition function?

- One can divide factor products by the partition function in order to convert them into probabilities.
- One can subtract the partition function from factor products in order to convert them into probabilities.
- The partition function is useless and should be ignored
- The partition function is used only in the context of Bayesian networks, not Markov networks.

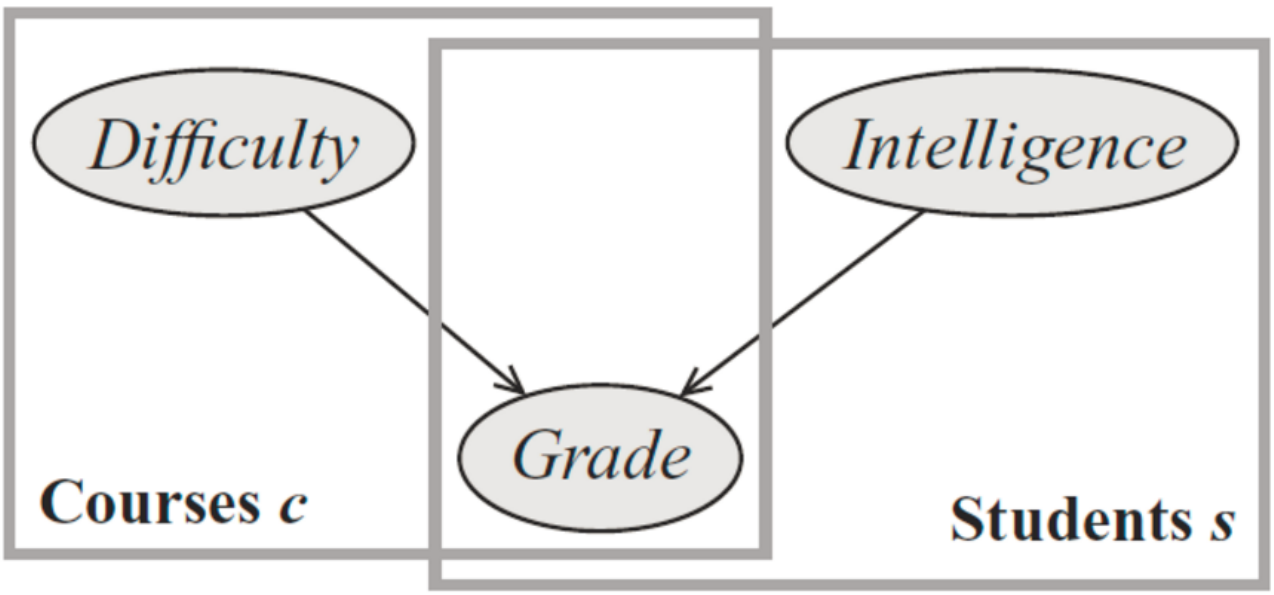
Fig. 45: Exercise 01-09-02

3. ***I-Equivalence.** Let T be any directed tree (not a polytree) over n nodes, where $n \geq 1$. A directed tree is a traditional tree, where each node has at most one parent and there is only one root, i.e., all but one node has exactly one parent. (In a polytree, nodes may have multiple parents.) How many networks (including itself) are I-equivalent to T?

- n
- $n+1$
- $2n$
- 2

Fig. 46: Exercise 01-09-03

4. ***Markov Network Construction.** Consider the unrolled network for the plate model shown below, where we have n students and m courses. Assume that we have observed the grade of all students in all courses. In general, what does a pairwise Markov network that is a minimal I-map for the conditional distribution look like? (Hint: the factors in the network are the CPDs reduced by the observed grades. We are interested in modeling the conditional distribution, so we do not need to explicitly include the Grade variables in this new network. Instead, we model their effect by appropriately choosing the factor values in the new network.)

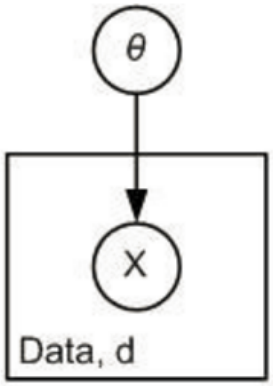


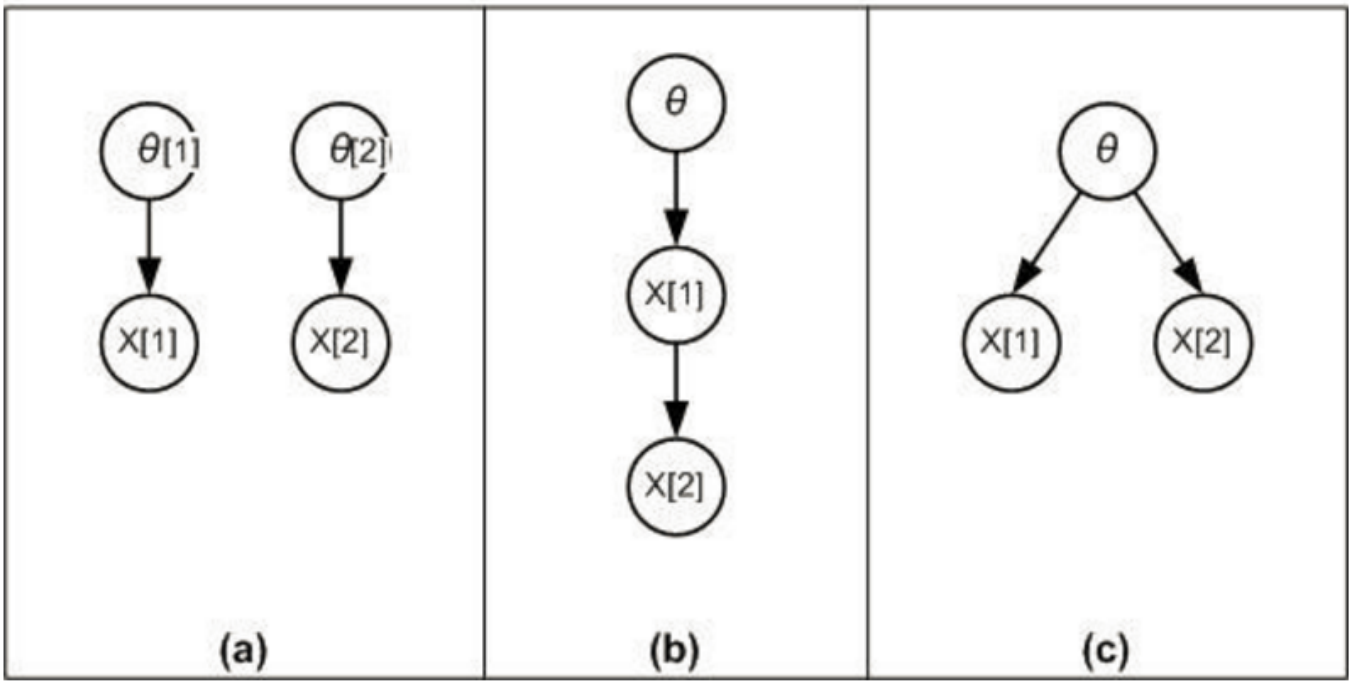
- A fully connected graph with instantiations of the Difficulty and Intelligence variables.
- Impossible to tell without more information on the exact grades observed.
- A fully connected bipartite graph where instantiations of the Difficulty variables are on one side and instantiations of the Intelligence variables are on the other side.
- A graph over instantiations of the Difficulty variables and instantiations of the Intelligence variables, not necessarily bipartite; there could be edges between different Difficulty variables, and there could also be edges between different Intelligence variables.
- A bipartite graph where instantiations of the Difficulty variables are on one side and instantiations of the Intelligence variables are on the other side. In general, this graph will not be fully connected.

Fig. 47: Exercise 01-09-04

5. Grounded Plates.

Which of the following is a valid grounded model for the plate shown? You may select 1 or more options.





- [ ] (b) -- watch out, options are not in order
- [ ] (a) -- watch out, options are not in order
- [ ] (c) -- watch out, options are not in order

Fig. 48: Exercise 01-09-05

6. Independencies in Markov Networks.

Consider the following set of factors:
$\Phi = \{\Phi_1(A, B), \Phi_2(B, C, D), \Phi_3(D), \Phi_4(C, E, F)\}$. Now, consider a Markov Network G such that P_Φ factorizes over G. Which of the following is an independence statement that holds in the network? You may select 1 or more options.

- [ ] $(C \perp D \mid A)$
- [ ] $(A \perp F \mid C)$
- [ ] $(B \perp E \mid A)$
- [ ] $(C \perp E \mid B)$
- [ ] $(B \perp E \mid C)$
- [ ] $(A \perp E \mid B)$

Fig. 49: Exercise 01-09-06

7. Factorization of Probability Distributions.

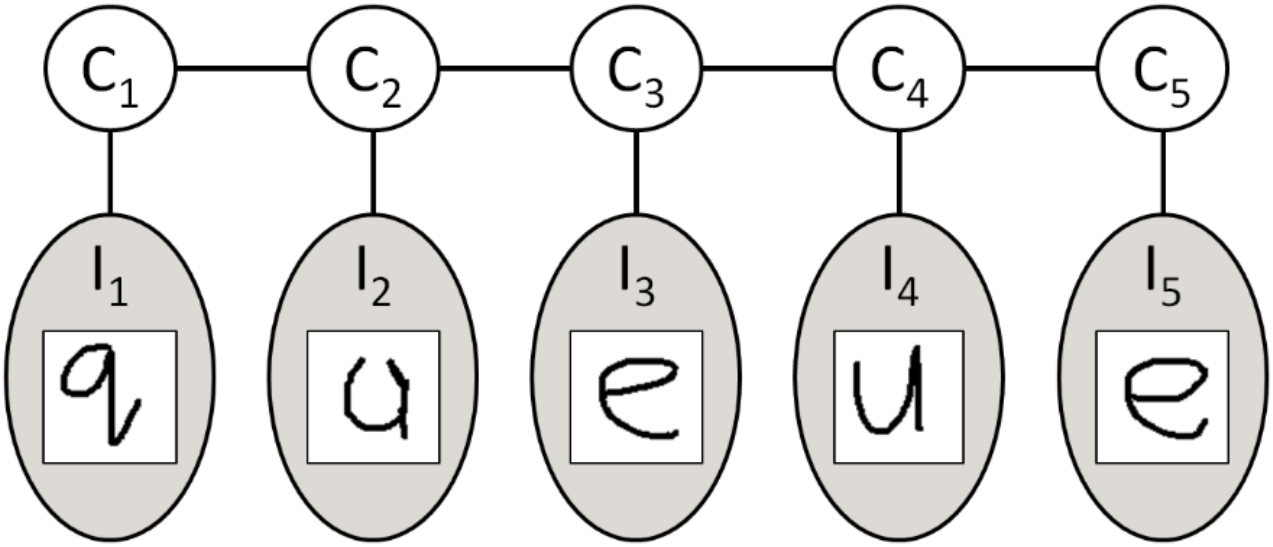
Consider a directed graph G. We construct a new graph G' by removing one edge from G. Which of the following is always true? You may select 1 or more options.

- [ ] If G and G' were undirected graphs, the answers to the other options would not change.
- [ ] Any probability distribution P that factorizes over G also factorizes over G'.
- [ ] Any probability distribution P that factorizes over G' also factorizes over G.
- [ ] No probability distribution P that factorizes over G also factorizes over G'.

Fig. 50: Exercise 01-09-07

8. Template Model in CRF.

The CRF model for OCR with only singleton and pairwise potentials that you played around with in PA3 and PA7 is an instance of a template model, with variables $C_1, \ldots, C_n$ over the characters and observed images $I_1, \ldots, I_n$. The model we used is a template model in that the singleton potentials are replicated across different C_i variables, and the pairwise potentials are replicated across character pairs. The structure of the model is shown below:



Now consider the advantages of this particular template model for the OCR task, as compared to a non-template model that has the same structure, but where there are distinct singleton potentials for each C_i variable, and distinct potentials for each pair of characters. Which of the following about the advantage of using a template model is true? You may select 1 or more options.

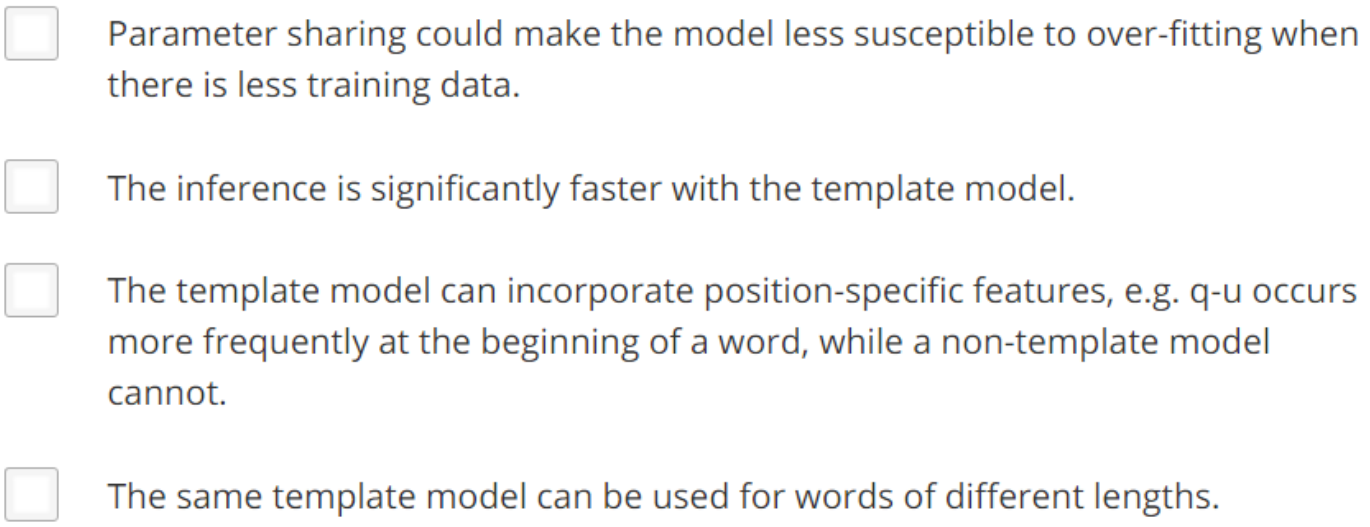


Fig. 51: Exercise 01-09-08

1.7.2 Answers

01-09-01: (A);
01-09-02: 1st;
01-09-03: n;
01-09-04: 3rd;
01-09-05: (c);
01-09-06: 2nd,5th,6th;
01-09-07: ???;1st and 4th not right! 4th not right!(Who know the answer, tell me! rockking.jy@gmail.com)
01-09-08: 1st, 4th;

2 Inference

2.1 Variable Elimination

2.1.1 Conditional Probability Queries (Chapter 9.3)

2.1.2 MAP Queries (Chapter 13.2.1)

Variable Elimination Algorithm. Chapter 9.2.
Variable Elimination Complexity. Chapter 9.4 through 9.4.2.3.
VE - Graph Based Perspective. Chapter 9.4.
Finding Elimination Orderings. Chapter 9.4.3.
Message Passing in Cluster Graphs
Belief Propagation. Chapter 11.3.2
Properties of Cluster Graphs. Chapter 11.3.2
Properties of Belief Propagation. Chapter 11.3.3

Clique Trees

Clique Tree Algorithm - Correctness. Chapter 10.2.1

Clique Tree Algorithm - Computation. Chapters 10.2.2, 10.3.3.1

Clique Trees and Independence. Chapter 10.1.2

Clique Trees and VE. Chapter 10.4.1

Optional: Loopy Belief Propagation

BP in Practice. Box 11.C

Loopy BP and Message Decoding. Box 11.A

MAP Message Passing (combined slides)

MAP Exact Inference. Chapter 13.2.1

Finding a MAP Assignment. Chapter 13.2.2

Optional: Other MAP Algorithms (combined slides)

Tractable MAP Problems. Chapter 13.6.

Dual Decomposition - Intuition. Dual Decomposition is not in the textbook, but for further information you may refer to the original paper: MRF Energy Minimization and Beyond via Dual Decomposition N. Komodakis, N.Paragios and G. Tziritas

Dual Decomposition - Algorithm.

Sampling Methods (combined slides)

Simple Sampling. Chapter 12.1.

Markov Chain Monte Carlo . Chapter 12.3 up to 12.3.2.2.

Using a Markov Chain. Chapter 12.3.5.

Gibbs Sampling. Review of Chapter 12.3.2 as applied to Gibbs Sampling.

Metropolis Hastings Algorithm. Chapter 12.3.4.2.

Inference In Temporal Models

Inference in Temporal Models. Chapter 15.

Reference

[1] Daphne Koller. Probabilistic graphical models.

[2] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.