Introduction to Machine Learning

- Prof. Balaraman Ravindran | IIT Madras

Problem Solving Session (Week-9)

Shreya Bansal

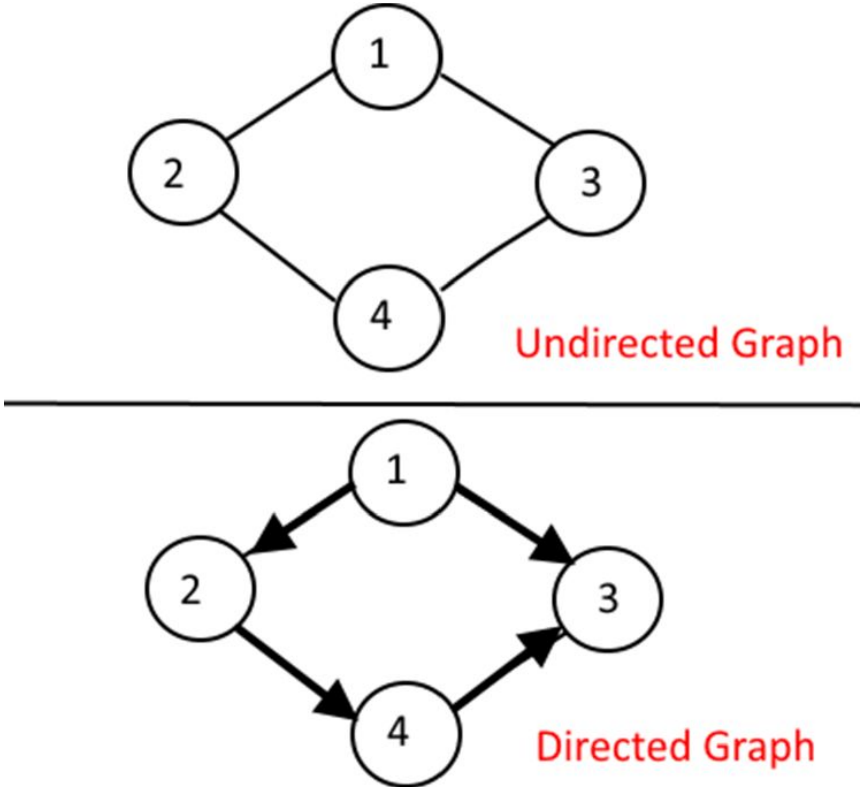
PMRF PhD Scholar
IIT Ropar

Week-9 Contents

---

1. Undirected Graph Models
2. Hidden Markov Model
3. Variable elimination
4. Tree width and Belief Propagation

Introduction to Undirected Graphical Models



Introduction to Undirected Graphical Models

- Undirected Graphical Models (UGMs) represent joint probability distributions using potential functions.
- Unlike Directed Graphical Models (Bayesian Networks), UGMs do not impose restrictions on the form of potential functions.
- UGMs are used to model relationships where directionality is not inherent (e.g., pixel neighborhoods in images).

Potential Functions

- Potential functions $\Psi c(Xc)$ are associated with cliques in the graph.

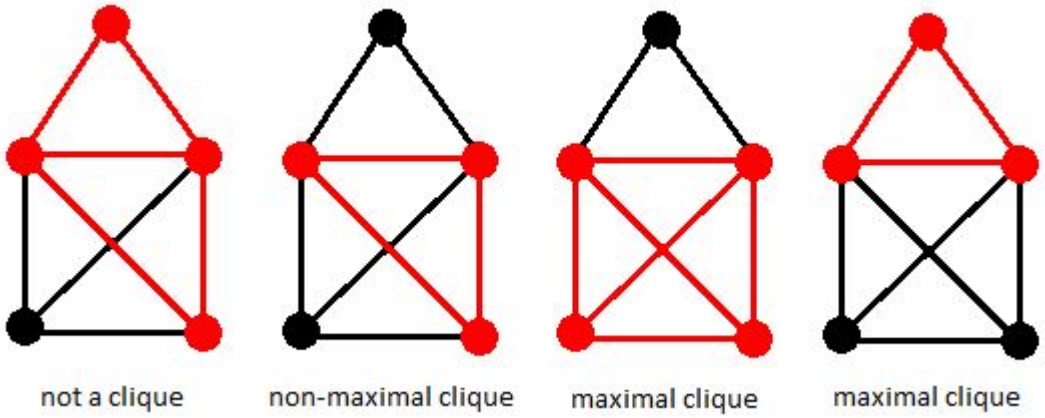
Cliques in Undirected Graphical Models

- A clique in an undirected graph is a subset of nodes where:
    - Every pair of nodes in the subset is connected by an edge. (Fully connected)
    - The subset is maximal (no additional node can be added to the subset while maintaining full connectivity).

Types of Cliques

Maximal Clique: A clique that cannot be extended by adding another node.

Non-Maximal Clique: A clique that is part of a larger clique.



Potential Functions

- Potential functions $\Psi c(Xc)$ are associated with **cliques** in the graph.
- They capture the **"compatibility" or "affinity"** between variables in a clique.

Compatibility and Affinity

- **Compatibility:**
- Refers to how well the values of variables in a clique "agree" with each other.
- High compatibility means the configuration is likely or desirable.

- **Affinity:**
- Refers to the strength of the relationship between variables in a clique.
- High affinity means the variables strongly influence each other.

Potential Functions

- Potential functions $\Psi c(Xc)$ are associated with **cliques** in the graph.
- They capture the **"compatibility" or "affinity"** between variables in a clique.
- **Properties:**
- $\Psi_c(X_c) \geq 0$(non-negative).
- Not restricted to being probabilities (unlike conditional probabilities in directed models).
- **Example:**
- For a clique c={X1,X2} , $\Psi_c(X_1, X_2)$ measures how likely X1 and X2 are to take specific values together.

Compatibility in a Clique

- **Clique**: $\{X_1, X_2\}$ (two variables).
- **Potential Function**:

$$\Psi_c(X_1, X_2) = \begin{cases} 5.0 & \text{if } X_1 = 0 \text{ and } X_2 = 0, \\ 2.0 & \text{if } X_1 = 0 \text{ and } X_2 = 1, \\ 1.0 & \text{if } X_1 = 1 \text{ and } X_2 = 0, \\ 3.0 & \text{if } X_1 = 1 \text{ and } X_2 = 1. \end{cases}$$

- **Interpretation**:
  - The configuration $X_1 = 0, X_2 = 0$ has the highest compatibility (value = 5.0).
  - The configuration $X_1 = 1, X_2 = 0$ has the lowest compatibility (value = 1.0).

Example: Pixel Labeling in Images

- Clique: Two neighboring pixels in an image.
- Potential Function:
- Assign high compatibility if neighboring pixels have the same label (e.g., both foreground or both background).
- Assign low compatibility if neighboring pixels have different labels.
- Result:
- Encourages smoothness in pixel labeling (e.g., large regions of foreground or background).

Probability Distribution in UGMs

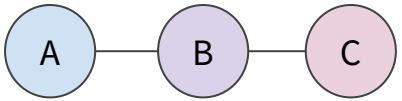
- **Formula:**
- $P(X)=(1/Z)\ \Pi_c \Psi_c(X_c)$
- $\Psi_c(X_c)$: **Potential function for clique c.**
- **Z: Partition function (normalizing constant).**
- **Ψc(Xc) must be non-negative.**
- **Z ensures P(X) is a valid probability distribution.**

Partition Function (Z)

- $Z = \sum_X \prod_c \Psi_c(X_c)$

- Z sums over all possible configurations of X.
- Computationally expensive for large graphs (e.g., 2^n for n binary variables).

Example: Product Over Cliques



- Cliques : {A,B} and {B,C}
- Variables: A, B, and C are binary variables (can take values 0 or 1).

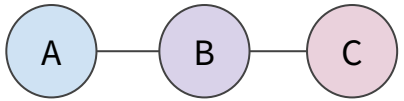
- **Potential Functions**:
  - For clique $\{A, B\}$:

$$\Psi_1(A,B) = \begin{cases} 2.0 & \text{if } A = 0 \text{ and } B = 0, \\ 1.5 & \text{if } A = 0 \text{ and } B = 1, \\ 1.0 & \text{if } A = 1 \text{ and } B = 0, \\ 3.0 & \text{if } A = 1 \text{ and } B = 1. \end{cases}$$

  - For clique $\{B, C\}$:

$$\Psi_2(B,C) = \begin{cases} 1.0 & \text{if } B = 0 \text{ and } C = 0, \\ 2.0 & \text{if } B = 0 \text{ and } C = 1, \\ 1.5 & \text{if } B = 1 \text{ and } C = 0, \\ 2.5 & \text{if } B = 1 \text{ and } C = 1. \end{cases}$$

Example: Product Over Cliques

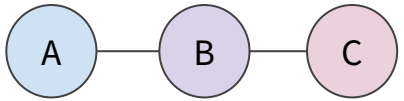


- **Calculate the Partition Function Z**
  - The partition function Z is the sum of the product of potential functions over all cliques for all possible configurations of A, B, and C:

$$Z = \sum_{A,B,C} \Psi_1(A, B) \cdot \Psi_2(B, C)$$

  - Compute Z:
    - There are $2^3 = 8$ possible configurations for A, B, and C.
    - Calculate $\Psi_1(A, B) \cdot \Psi_2(B, C)$ for each configuration and sum them up.

Example: Product Over Cliques

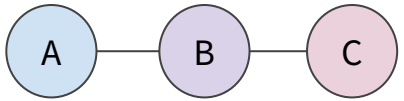


Configuration (A, B, C)	$\Psi_1(A, B)$	$\Psi_2(B, C)$	Product $\Psi_1 \cdot \Psi_2$
$(0, 0, 0)$	2.0	1.0	$2.0 \cdot 1.0 = 2.0$
$(0, 0, 1)$	2.0	2.0	$2.0 \cdot 2.0 = 4.0$
$(0, 1, 0)$	1.5	1.5	$1.5 \cdot 1.5 = 2.25$
$(0, 1, 1)$	1.5	2.5	$1.5 \cdot 2.5 = 3.75$
$(1, 0, 0)$	1.0	1.0	$1.0 \cdot 1.0 = 1.0$
$(1, 0, 1)$	1.0	2.0	$1.0 \cdot 2.0 = 2.0$
$(1, 1, 0)$	3.0	1.5	$3.0 \cdot 1.5 = 4.5$
$(1, 1, 1)$	3.0	2.5	$3.0 \cdot 2.5 = 7.5$

- Sum of products:

$$Z = 2.0 + 4.0 + 2.25 + 3.75 + 1.0 + 2.0 + 4.5 + 7.5 = 27.0$$

Example: Product Over Cliques



Calculate P(X) for Each Configuration

- The joint probability distribution $P(X)$ is given by:

$$P(X) = \frac{1}{Z}\Psi_1(A,B) \cdot \Psi_2(B,C)$$

- Compute $P(X)$ for each configuration:

Configuration (A, B, C)	Product $\Psi_1 \cdot \Psi_2$	Probability $P(X)$
$(0,0,0)$	2.0	$\frac{2.0}{27.0} = 0.0741$
$(0,0,1)$	4.0	$\frac{4.0}{27.0} = 0.1481$
$(0,1,0)$	2.25	$\frac{2.25}{27.0} = 0.0833$
$(0,1,1)$	3.75	$\frac{3.75}{27.0} = 0.1389$
$(1,0,0)$	1.0	$\frac{1.0}{27.0} = 0.0370$
$(1,0,1)$	2.0	$\frac{2.0}{27.0} = 0.0741$
$(1,1,0)$	4.5	$\frac{4.5}{27.0} = 0.1667$
$(1,1,1)$	7.5	$\frac{7.5}{27.0} = 0.2778$

Step 3: Verify Normalization

- Ensure that the probabilities sum to 1:

$$0.0741 + 0.1481 + 0.0833 + 0.1389 + 0.0370 + 0.0741 + 0.1667 + 0.2778 = 1.0$$

Challenges in UGMs

- Potential functions $\Psi_c(X_c)$ are unrestricted (unlike conditional probabilities in directed models).
- Normalization is required to ensure $P(X)$ is a valid probability distribution.
- Inference and computation of Z are computationally intensive.
- Example: For a graph with 20 binary variables, Z requires summing over 2^{20}=1,048,576 configurations.

Factorization in Undirected vs. Directed Graphs

- **Directed Graphs (Bayesian Networks)**: Factorize as a product of conditional probabilities:

$$P(X1,X2,...,Xn)=\prod_i P(X_i \mid \text{Parents}(X_i))$$

- Dependencies are encoded through parent-child relationships.
- Example: Disease diagnosis models.
- **Undirected Graphs (MRFs & CRFs) :** Factorize as a product of potential functions over cliques:

$$P(X)=(1/Z)\prod_c \Psi_c(X_c)$$

- No directionality; dependencies are symmetric.
- Example: Image segmentation using MRFs.
- **Key Differences**
- Directed Graphs: Encode causal relationships, easy sampling.
- Undirected Graphs: More flexible but harder for inference.

Why factorization use cliques

- Factorization in undirected graphs uses cliques because of the Hammersley–Clifford Theorem, which states that if a probability distribution $P(X)$ is positive and follows the Markov property with respect to an undirected graph, then it can be factorized into a product of potential functions over the maximal cliques of the graph.

$$P(X)=(1/Z)\,\Pi_c \Psi_c(X_c)$$

Why Cliques?

- **Complete Local Dependencies:**
  - A clique is a fully connected subset of nodes, meaning all variables within the clique directly interact.
  - Factorization over cliques ensures that no dependencies are ignored.
- **Markov Property Preservation:**
  - The clique potentials maintain local conditional independence, ensuring consistency with the graphical structure.
- **Simplified Computation:**
  - Using maximal cliques reduces redundancy and avoids unnecessary factorization over non-maximal subsets.

Hammersley-Clifford Theorem

- **Statement:**
- **Any probability distribution consistent with the factorization over a graph can be expressed using potential functions of the form:**

$$\Psi_c(X_c)=\exp\{-E(X_c)\}$$

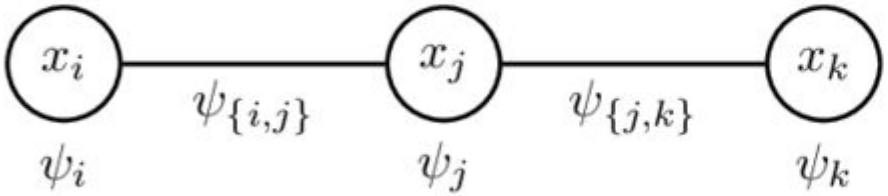
- **$E(X_c)$: Energy function (can be any real-valued function).**

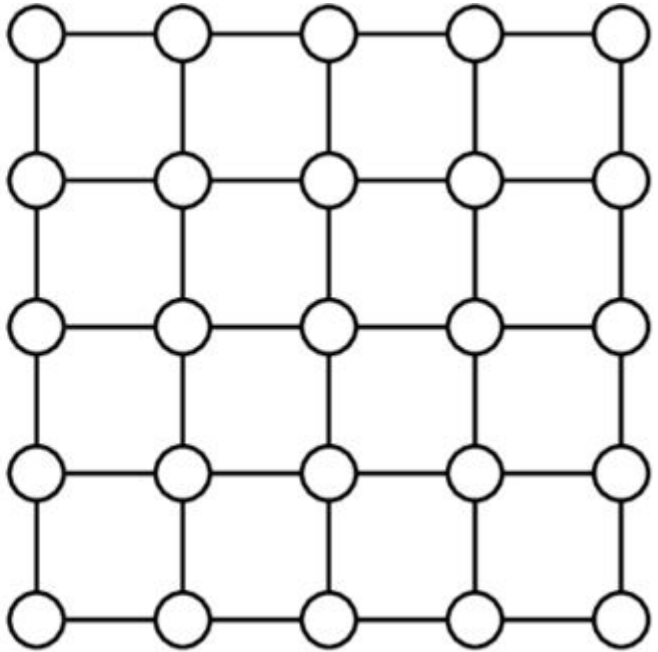
- **Key Points:**
- **Energy functions simplify the representation of potential functions.**
- **High energy corresponds to low probability, and vice versa.**

Energy Functions and Intuition

- Energy functions $E(X_c)$ are derived from data.
- High-count configurations in data are assigned low energy (high probability).
- Energy functions are not restricted to being normalized.

Markov Random Fields (MRFs)





Markov Random Fields (MRFs)

- **Definition:**
- **Undirected graphical models are also called Markov Random Fields.**
- **Variables are independent of non-neighbors given their immediate neighbors.**

- **Key Points:**
- **MRFs are commonly used in image processing (e.g., pixel labeling).**
- **Example: Lattice structure for modeling images.**

Pixel Labeling with MRFs

- Each pixel is a random variable (e.g., foreground or background).
- Potential functions are defined for edges (pairs of neighboring pixels).
- Inference involves finding the configuration of labels with the lowest energy.
- Example:
  - For a 3x3 image, there are 9 pixels, each with a label (foreground or background).
  - The goal is to assign labels such that the overall energy is minimized.

Pixel Labeling with MRFs

(a) Original frame

(b) Foreground pixel distribution

Training MRFs

- Node potentials: Derived from observed pixel values.
- Edge potentials: Learned from co-occurrence statistics in the data.
- Energy functions are often learned using maximum likelihood or logistic regression.

Example of MRF Training

- **Dataset Preparation:**
- Given an image dataset where each pixel needs to be labeled (e.g., foreground/background segmentation).
- **Defining Potentials:**
- Node potentials: Derived from observed pixel values using a classifier (e.g., logistic regression).
- Edge potentials: Learned from co-occurrence statistics in labeled training data.

Example of MRF Training

- **Energy Function:**
- **The energy function is formulated as:**

$$E(X)=\sum_i \psi i(Xi)+\sum_{(i,j)\in E} \psi ij(Xi,Xj)$$

- **Where $\psi i(Xi)$represents node potentials and $\psi ij(Xi,Xj)$ represents edge potentials.**
- **Parameter Learning:**
- **Use Maximum Likelihood Estimation (MLE) to estimate parameters.**
- **Gradient-based optimization methods (e.g., Stochastic Gradient Descent) help refine parameters.**
- **Inference for Prediction:**
- **Given a new image, inference (e.g., Graph Cuts, Belief Propagation) finds the most probable pixel labels.**

Example of MRF Training

- **Problem:**

Given a noisy grayscale image, classify each pixel as foreground (1) or background (0) using MRF.

- **Step 1: Define the Graph Structure**

Each pixel is a node in an undirected graph.

Edges connect neighboring pixels (e.g., 4-nearest neighbors).

Each node X_i takes values 0 (background) or 1 (foreground).

Example of MRF Training

- **Step 2: Define Node and Edge Potentials**
- **Node Potential $\psi i(Xi)$:** The probability of a pixel being foreground or background based on its observed intensity Yi .
- **Example: $\psi i(Xi)=P(Yi|Xi=1)$ if foreground or $P(Yi|Xi=0)$if background where probabilities can be estimated from histograms of training images.**
- **Edge Potential $\psi ij(Xi,Xj)$: Encourages smoothness (neighboring pixels prefer similar labels)**
- **$\psi ij(Xi,Xj)=e^{-\beta|Yi-Yj|}$ where β controls smoothness.**

Example of MRF Training

- Step 3: Compute the Energy Function
- The total energy is: $E(X) = \sum_i \psi i(Xi) + \sum_{(i,j)\in E} \psi ij(Xi, Xj)$
- For a 3x3 pixel patch:
- Observed intensities:

$$Y = \begin{bmatrix} 230 & 200 & 220 \\ 50 & 60 & 55 \\ 240 & 215 & 225 \end{bmatrix}$$

- Assumed foreground intensity range: 200–255
- Assumed background intensity range: 0–100

Challenges in Inference

- Exact inference is computationally hard, especially in graphs with loops.
- Approximate inference methods are often used.
- Trees allow for exact inference, but loops complicate the process.

Applications of MRFs

- **Image segmentation and labeling.**
- **Conditional Random Fields (CRFs): Extension of MRFs for structured prediction tasks.**
- **Widely used in computer vision and natural language processing.**

Introduction to HMMs

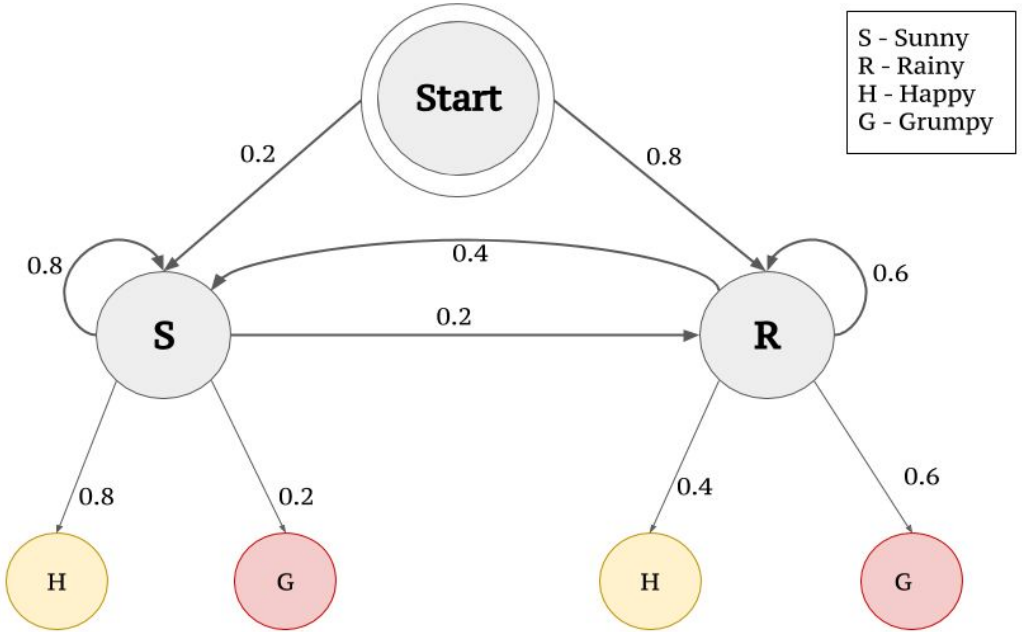
- **What is an HMM?**
- **A graphical model representing probabilistic dependencies.**
- **Hidden states generate observed sequences.**
- **Based on the Markov assumption: The present state depends only on the previous state.**

Components of an HMM

1. States (X): Hidden variables.
2. Observations (Y): The observable sequence.
3. Transition Probabilities (A): Probabilities of moving from one state to another.
4. Emission Probabilities (B): Probability of an observation given a state.
5. Initial Probabilities (π): Probability distribution of the starting state.

Graphical Representation

- A directed chain structure where each state depends only on the previous state.
- Observations depend only on the corresponding hidden state.



Example - Part-of-Speech Tagging

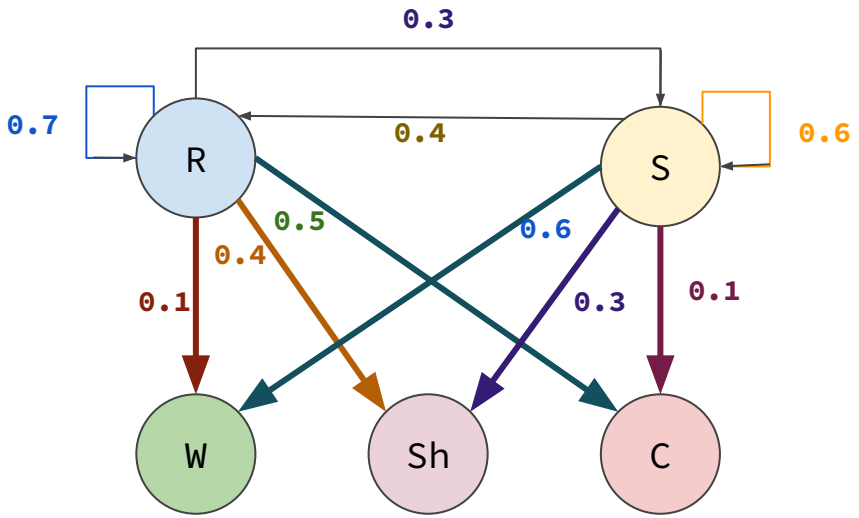
- States (X): Noun (N), Verb (V), Adjective (A)
- Observations (Y): Words in a sentence ("dog", "barks", "loudly")
- Transition Probabilities (A):

  $$P(N \rightarrow V) = 0.5, P(N \rightarrow A) = 0.3, \text{etc.}$$

- Emission Probabilities (B):

  $$P(\text{"dog"} \mid N) = 0.6, P(\text{"barks"} \mid V) = 0.7, \text{etc.}$$

Calculated Example



Given:

States: Rainy (R), Sunny (S)

Observations: Walk (W), Shop (Sh), Clean (C)

Transition Probabilities (A):

P(R|R) = 0.7, P(S|R) = 0.3

P(R|S) = 0.4, P(S|S) = 0.6

Emission Probabilities (B):

P(W|R) = 0.1, P(Sh|R) = 0.4, P(C|R) = 0.5

P(W|S) = 0.6, P(Sh|S) = 0.3, P(C|S) = 0.1

Initial Probabilities (π):P(R) = 0.6, P(S) = 0.4

Calculated Example

- **Question**

  What is the probability of observing the sequence (Walk, Shop) given the model?

Calculated Example

Step 1: Compute Forward Probabilities

1. **At time t=1:**
   - $\alpha_1(R) = \pi(R) * P(W \mid R) = 0.6 * 0.1 = 0.06$
   - $\alpha_1(S) = \pi(S) * P(W \mid S) = 0.4 * 0.6 = 0.24$
2. **At time t=2:**
   - $\alpha_2(R) = [\alpha_1(R) * P(R \mid R) + \alpha_1(S) * P(R \mid S)] * P(Sh \mid R) = [(0.06 * 0.7) + (0.24 * 0.4)] * 0.4 = (0.042 + 0.096) * 0.4 = 0.0552$
   - $\alpha_2(S) = [\alpha_1(R) * P(S \mid R) + \alpha_1(S) * P(S \mid S)] * P(Sh \mid S) = [(0.06 * 0.3) + (0.24 * 0.6)] * 0.3 = (0.018 + 0.144) * 0.3 = 0.0486$

Final Probability:

$P(Walk, Shop) = \alpha_2(R) + \alpha_2(S) = 0.0552 + 0.0486 = \mathbf{0.1038}$

Inference in Graphical Models

- **Two Core Problems:**
- Inference: Given a model, answer queries (e.g., marginals, conditionals).
    - Example: P(Job | Grade)
- Learning:
    - Parameter Learning: Estimate potentials/CPDs given the graph.
    - Structure Learning: Discover the graph from data (harder).
- **Challenge:**
- Marginalizing over large joint distributions is computationally expensive.

Example Model (Student Scenario)

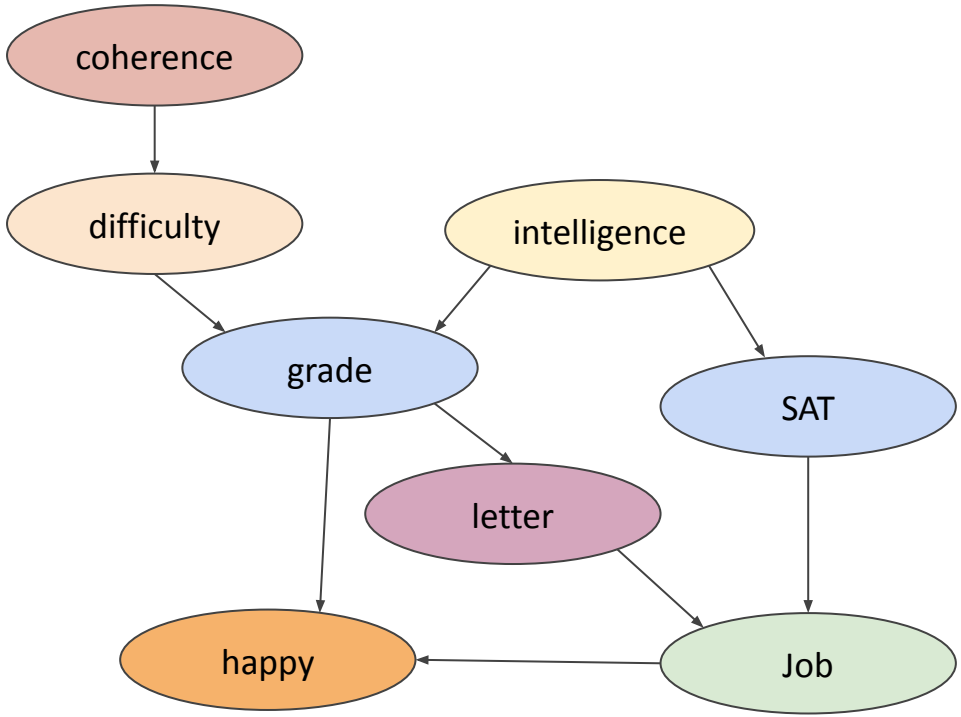
$P(C,D,I,G,S,L,J,H)=$

$P(C)\ P(D|C)\ P(I)\ P(G|I,D)\ P(S|I)$
$P(L|G)\ P(J|L,S)\ P(H|G,J)$

Inference Query:

$P(J)=\sum_{C,D,I,G,S,L,H} P(C,D,I,G,S,L,J,H)$

Naive summation: $O(2^8)$ for binary variables



Variable Elimination (Intuition)

- **Goal: Push sums inward to minimize computation.**
- **Key Idea:**
- **Factorize: Express joint as product of local potentials (CPDs/factors).**
- **Eliminate Variables Sequentially: Marginalize one variable at a time, updating remaining factors.**
- **Example Elimination Order:**
- **C→D→I→S→L→G→H**

Step-by-Step Elimination-1

- Eliminate C
- $P(C)\ P(D|C)\ P(I)\ P(G|I,D)\ P(S|I)\ P(L|G)\ P(J|L,S)\ P(H|G,J)$

- Compute $\tau_1(D)=\sum_C P(C)P(D|C)$
- New factor over D

Step-by-Step Elimination-2

- Eliminate D
- $P(C)\ P(D|C)\ P(I)\ P(G|I,D)\ P(S|I)\ P(L|G)\ P(J|L,S)\ P(H|G,J)$
- $\tau_1(D)\ P(I)\ P(G|I,D)\ P(S|I)\ P(L|G)\ P(J|L,S)\ P(H|G,J)$

- Compute $\tau_2(G,I)=\sum_D P(G|I,D)\tau 1(D)$
- New factor over G,I.

Step-by-Step Elimination-3

- Eliminate I
- $P(C)\ P(D|C)\ P(I)\ P(G|I,D)\ P(S|I)\ P(L|G)\ P(J|L,S)\ P(H|G,J)$
- $\tau_1(D)\ P(I)\ P(G|I,D)\ P(S|I)\ P(L|G)\ P(J|L,S)\ P(H|G,J)$
- $P(I)\ \tau 2(G,I)\ P(S|I)\ P(L|G)\ P(J|L,S)\ P(H|G,J)$

- Compute $\tau_3(G,S)=\sum_I \tau 2(G,I)P(I)P(S|I)$

Step-by-Step Elimination-4

- Eliminate S
- $P(C)\,P(D|C)\,P(I)\,P(G|I,D)\,P(S|I)\,P(L|G)\,P(J|L,S)\,P(H|G,J)$
- $\tau_1(D)\,P(I)\,P(G|I,D)\,P(S|I)\,P(L|G)\,P(J|L,S)\,P(H|G,J)$
- $P(I)\,\tau 2(G,I)\,P(S|I)\,P(L|G)\,P(J|L,S)\,P(H|G,J)$
- $\tau_3(G,S)P(L|G)\,P(J|L,S)\,P(H|G,J)$

- Compute $\tau_4(J,L,G)=\sum_S \tau 3(G,S)P(J|L,S)$

Step-by-Step Elimination-5

- Eliminate L
- $P(C)\ P(D|C)\ P(I)\ P(G|I,D)\ P(S|I)\ P(L|G)\ P(J|L,S)\ P(H|G,J)$
- $\tau_1(D)\ P(I)\ P(G|I,D)\ P(S|I)\ P(L|G)\ P(J|L,S)\ P(H|G,J)$
- $P(I)\ \tau 2(G,I)\ P(S|I)\ P(L|G)\ P(J|L,S)\ P(H|G,J)$
- $\tau_3(G,S) P(L|G)\ P(J|L,S)\ P(H|G,J)$
- $\tau 4(J,L,G) P(L|G)\ P(H|G,J)$

- Compute $\tau_5(J,G)=\sum_L \tau 4(J,L,G) P(L|G)$

Step-by-Step Elimination-6

- Eliminate G
- $P(C)\,P(D|C)\,P(I)\,P(G|I,D)\,P(S|I)\,P(L|G)\,P(J|L,S)\,P(H|G,J)$
- $\tau_1(D)\,P(I)\,P(G|I,D)\,P(S|I)\,P(L|G)\,P(J|L,S)\,P(H|G,J)$
- $P(I)\,\tau_2(G,I)\,P(S|I)\,P(L|G)\,P(J|L,S)\,P(H|G,J)$
- $\tau_3(G,S)P(L|G)\,P(J|L,S)\,P(H|G,J)$
- $\tau_4(J,L,G)P(L|G)\,P(H|G,J)$
- **$\tau_5(J,G)P(H|G,J)$**

- Compute $\tau_6(J,H)=\sum_G \tau_5(J,G)P(H|G,J)$

Step-by-Step Elimination-6

- Eliminate H
- $P(C)\ P(D|C)\ P(I)\ P(G|I,D)\ P(S|I)\ P(L|G)\ P(J|L,S)\ P(H|G,J)$
- $\tau_1(D)\ P(I)\ P(G|I,D)\ P(S|I)\ P(L|G)\ P(J|L,S)\ P(H|G,J)$
- $P(I)\ \tau_2(G,I)\ P(S|I)\ P(L|G)\ P(J|L,S)\ P(H|G,J)$
- $\tau_3(G,S) P(L|G)\ P(J|L,S)\ P(H|G,J)$
- $\tau_4(J,L,G) P(L|G)\ P(H|G,J)$
- $\tau_5(J,G) P(H|G,J)$
- $\tau_6(J,H)$

- Compute $P(J)=\Sigma_H \tau_6(J,H)$

Computational Complexity

- Induced Width: Size of the largest clique formed during elimination.
- Determines complexity: O(exp(induced width))
- Good Order:Always eliminate variables that disconnect the graph first (like C).
- Rule of Thumb: Start with variables that have the fewest connections!
- Fill-in Edges: Edges added during elimination to form cliques.

Treewidth and Induced Width

- **Definition:**
- Treewidth: Minimal induced width across all possible elimination orderings.
- Induced Width: Size of the largest clique formed during variable elimination.

- **Key Points:**
- Measures graph complexity for inference.
- For trees: Treewidth = 1 (if defined as max clique size - 1).
- Elimination order impacts induced width (leaf-to-root is optimal for trees).

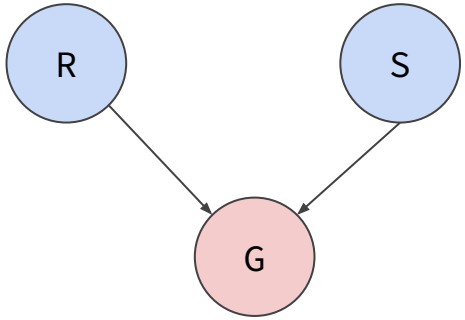
Limitations of Variable Elimination

- **Problem:**
- **Recomputing intermediates for new queries (e.g., P(H)after P(J)).**
- **No reuse of calculations.**

Belief Propagation (The "Time-Saver")

- Cache intermediate results ("messages") between nodes.
- Like writing down sub-totals while cleaning:
    - Compute "message" from C to D once.
    - Reuse it for all future queries involving D
- Works perfectly for trees (no loops).
- Approximate for complex graphs (but still faster).

Example of Belief Propagation



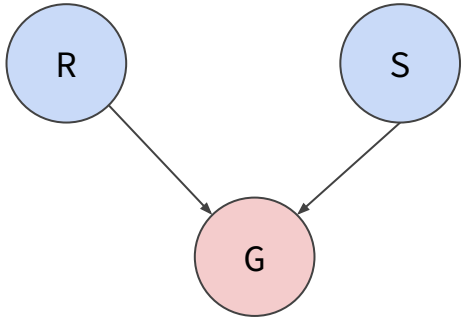
Consider this Bayesian Network (Tree):
Rain (R) → Wet Grass (G)
Sprinkler (S) → Wet Grass (G)

We want to compute:

P(R=1|G=1) (Probability it rained given the grass is wet).

P(S=1|G=1) (Probability the sprinkler was on given the grass is wet).

Step 1: Define Probabilities



- Assume binary variables (0/1)
- Priors: P(R=1)=0.2 P(S=1)=0.1
- Conditional Probabilities:
- P(G=1|R=1,S=0)=0.9 (Grass gets wet if it rains)
- P(G=1|R=0,S=1)=0.8 (Grass gets wet if sprinkler is on)
- P(G=1|R=1,S=1)=0.99 (Both rain and sprinkler)
- P(G=1|R=0,S=0)=0.01 (Neither)

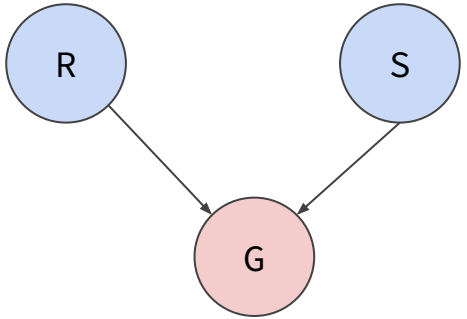
Step 1: Factor Graph Representation

- The joint distribution factors as: $P(R,S,G)=P(R)\cdot P(S)\cdot P(G|R,S)$
- Priors:

  $(R)=[0.8,0.2]$ (for R=0,R=1)

  $(S)=[0.9,0.1]$ (for S=0,S=1)

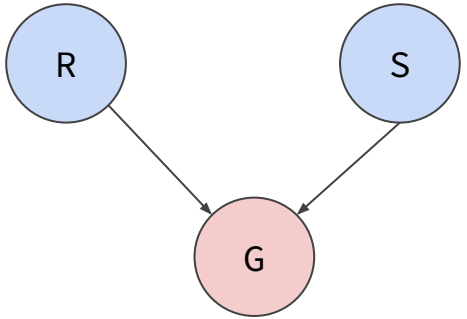
- Conditional Probability Table (CPT) for G (R,S,G): Given by the problem (see table below).



R	S	$G=0$	$G=1$
0	0	0.99	0.01
0	1	0.2	0.8
1	0	0.1	0.9
1	1	0.01	0.99

Step 1: Factor Graph Representation

- **Parent vs. Child Nodes:**
    - **Parents (R,S): Directly influence the child (G).**
    - **Child (G): Depends on its parents (via P(G|R,S)).**
- **Message Types:**
    - **Child-to-parent message: Sent from G to R (or S).**
    - **"Given my observed value, here's how likely each of your states is."**
    - **Parent-to-child message: Sent from R or S to G.**
    - **"Here's my current belief about my state."**
- **Goal: Compute how observing G=1 updates beliefs about parents R and S.**



R	S	$G=0$	$G=1$
0	0	0.99	0.01
0	1	0.2	0.8
1	0	0.1	0.9
1	1	0.01	0.99

Message G to Parent R

Step-by-Step: Message from Child G to Parent R

1. Observe $G = 1$

- We fix $G = 1$ and want to know how this affects R.

2. What Does G Say to R?

The message $\mu_{G \to R}(R)$ answers:
"For each possible state of R, how well does it explain my observed state $G = 1$, averaged over all possible states of my other parent S?"

3. Formula

$$\mu_{G \to R}(R) = \sum_S P(G = 1 \mid R, S) \cdot \mu_{S \to G}(S)$$

- $\mu_{S \to G}(S)$: Current belief about S (initially its prior: $[0.9, 0.1]$).

Message G to Parent R

4. Calculation

- **For $R = 0$:**

$$\mu_{G \to R}(0) = P(G = 1 \mid R = 0, S = 0) \cdot P(S = 0) + P(G = 1 \mid R = 0, S = 1) \cdot P(S = 1)$$
$$= 0.01 \times 0.9 + 0.8 \times 0.1 = 0.089$$

  - *Interpretation*: If $R = 0$, the combined probability of $G = 1$ (via sprinkler or neither) is **0.089**.

- **For $R = 1$:**

$$\mu_{G \to R}(1) = P(G = 1 \mid R = 1, S = 0) \cdot P(S = 0) + P(G = 1 \mid R = 1, S = 1) \cdot P(S = 1)$$
$$= 0.9 \times 0.9 + 0.99 \times 0.1 = 0.909$$

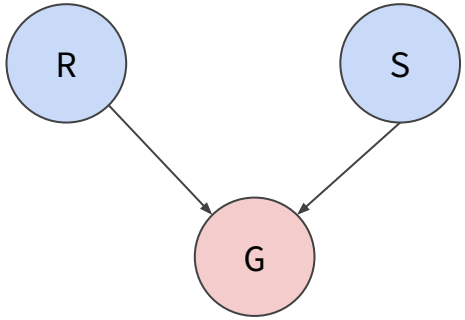
  - *Interpretation*: If $R = 1$, the probability of $G = 1$ is **0.909** (much higher!).

5. Resulting Message

$$\mu_{G \to R}(R) = [0.089, 0.909] \quad (\text{for } R = 0 \text{ and } R = 1)$$

- This tells R: "*Your state $R = 1$ explains $G = 1$ much better than $R = 0$.*"

How R Updates Its Belief



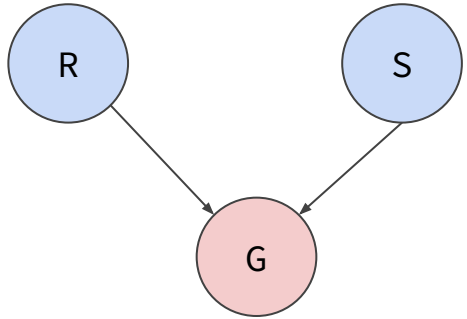
1. **Prior belief about** R: $[0.8, 0.2]$ (from f_R).
2. **Multiply by message**:

$$P(R \mid G = 1) \propto [0.8 \times 0.089, 0.2 \times 0.909] = [0.0712, 0.1818]$$

3. **Normalize**:

$$P(R = 1 \mid G = 1) = \frac{0.1818}{0.0712 + 0.1818} = 71.86\%.$$

Message from Child G to Parent S



The same logic applies for S:

$$\mu_{G \to S}(S) = \sum_{R} P(G = 1 \mid R, S) \cdot \mu_{R \to G}(R)$$

- **For** $S = 0$:

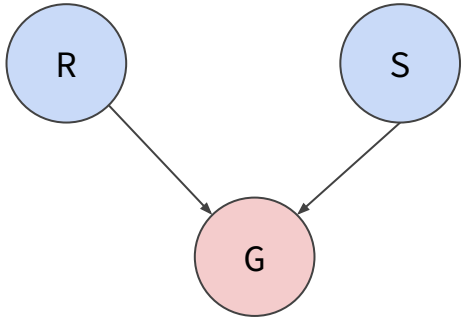
$$0.01 \times 0.8 + 0.9 \times 0.2 = 0.188$$

- **For** $S = 1$:

$$0.8 \times 0.8 + 0.99 \times 0.2 = 0.838$$

- Result: $\mu_{G \to S}(S) = [0.188, 0.838]$.

Key Intuition



- **Child-to-parent message: Summarizes how well each parent state explains the observed child, accounting for uncertainty in the other parents.**
- **Parent-to-child message: Shares the parent's current belief about itself.**

Key Formula

Key Formulas:

1. **Message from parent X to child Y:**

$$\mu_{X \to Y}(X) = P(X)$$

2. **Message from child Y to parent X** (if observed):

$$\mu_{Y \to X}(X) = \sum_{S \in \text{Siblings}} P(Y \mid X, S) \cdot \mu_{S \to Y}(S)$$

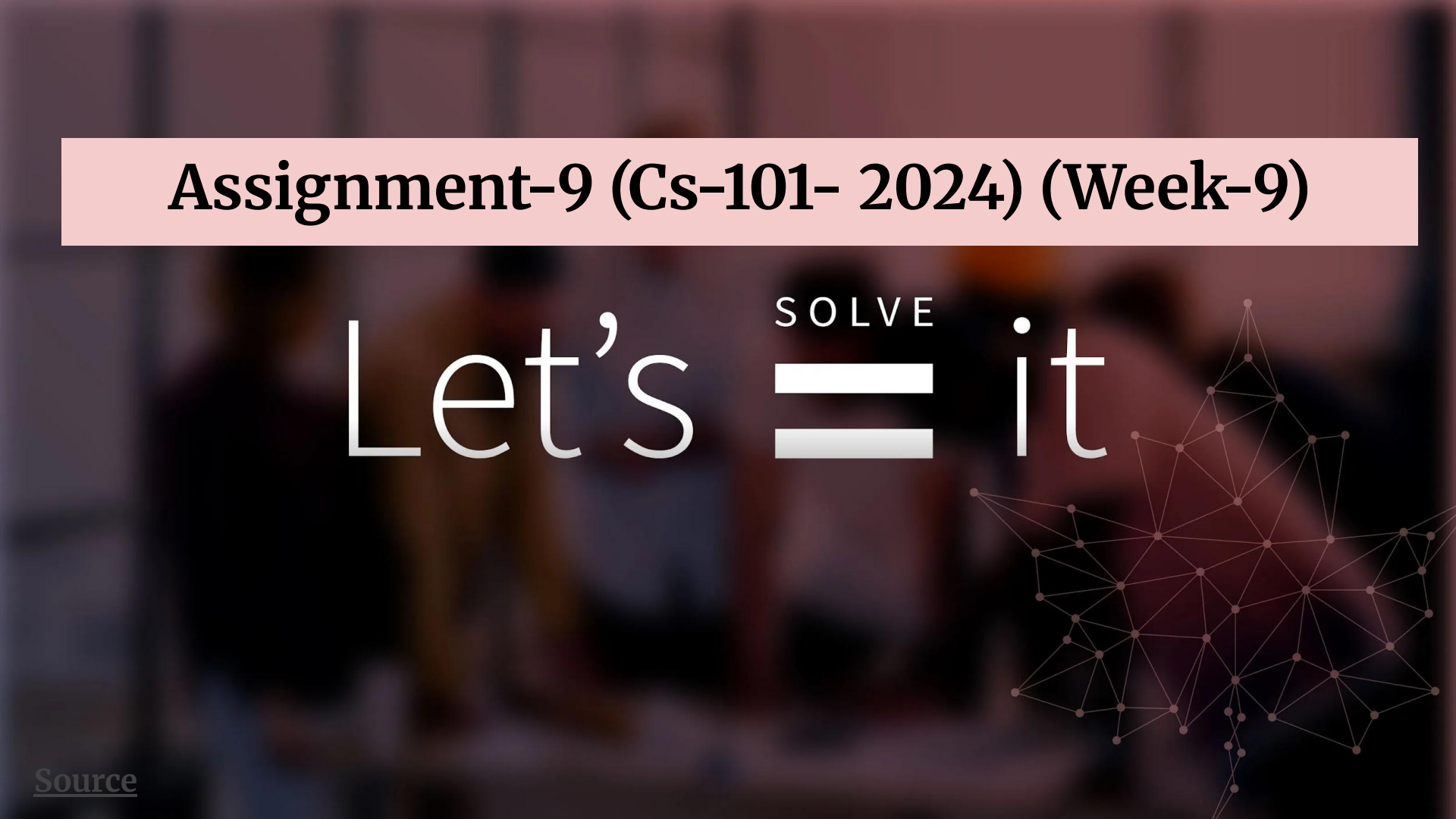
3. **Marginal**:

$$P(X \mid \text{evidence}) \propto \mu_{X \to Y}(X) \cdot \mu_{Y \to X}(X)$$

MAP Inference (Finding the "Best" Configuration)

- **Goal: Find the most probable assignment (e.g., best job/happiness combo).**
- **Trick: Replace sums with max in variable elimination:**
    - **Compute $\max_C P(C)P(D|C)$ instead of $\sum C$.**
    - **Track which C value gave the max (e.g., C=1).**
    - **Repeat for all variables.**
- **Result: Probability of the best configuration + the configuration itself.**

Assignment-9 (Cs-101- 2024) (Week-9)

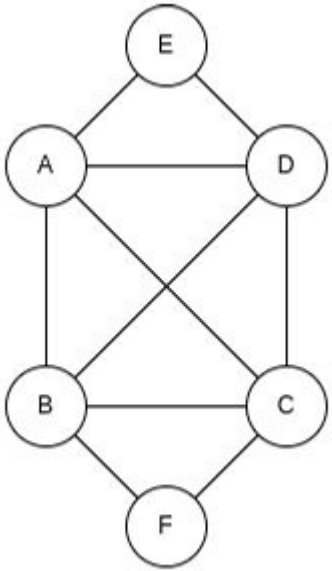


Source

Question-1

01:00

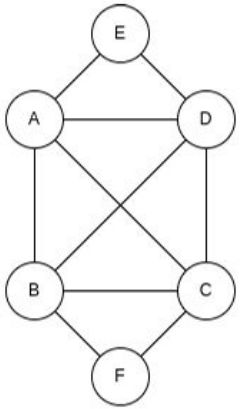
In the undirected graph given below, how many terms will be there in its potential function factorization?



a) 7
b) 3
c) 5
d) 9
e) None

Question-1- Correct answer

In the undirected graph given below, how many terms will be there in its potential function factorization?



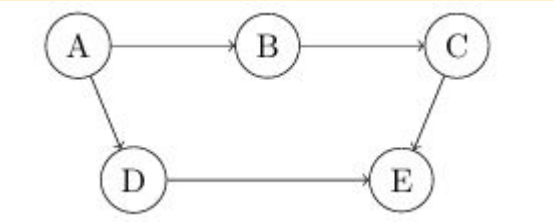
a) 7
b) 3
c) 5
d) 9
e) none

Correct options: (b)-three cliques {A,D,E}, {A,B,C,D}, {B,C,F}

Question-2

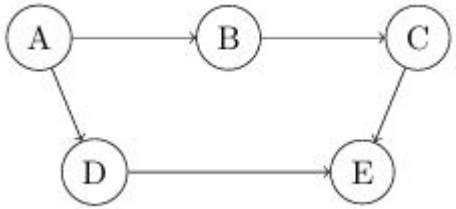
01:00

Consider the following directed graph:
Based on the d-separation rules, which of the following statements is true?



a) A and C are conditionally independent given B
b) A and E are conditionally independent given D
c) B and E are conditionally dependent given C
d) A and C are conditionally dependent given D and E

Question-2-Explanation



a) A and C are conditionally independent given B: A→B→C

Chain Structure: if observed→ block (independent)

b) A and E are conditionally independent given D

Chain Structure: if observed→ block (independent)

c) B and E are conditionally dependent given C

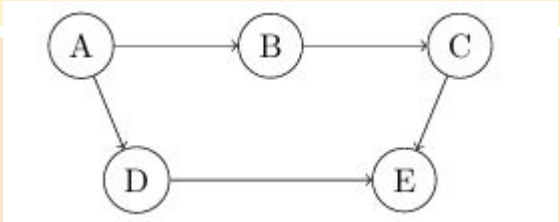
Chain Structure: if observed→ block (independent)

d) A and C are conditionally dependent given D and E

A→D→E← C: if E,D observes , collider path, observe means dependent

Question-2- Correct answer

Consider the following directed graph: Based on the d-separation rules, which of the following statements is true?



a) A and C are conditionally independent given B

b) A and E are conditionally independent given D

c) B and E are conditionally dependent given C

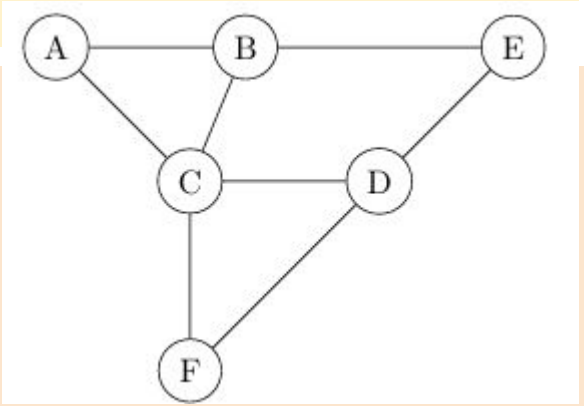
d) A and C are conditionally dependent given D and E

Correct options: (a) (b) (d)

Question-3

01:00

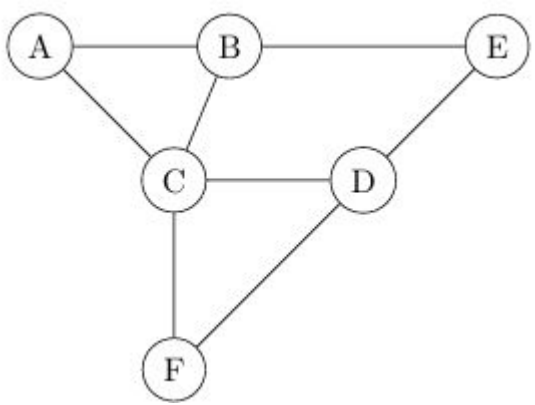
Consider the following undirected graph: In the undirected graph given above, which nodes are conditionally independent of each other given C? Select all that apply.



a) A, E
b) B, F
c) A, D
d) B, D
e) None of the above

Question-3 - Correct answer

Consider the following undirected graph: In the undirected graph given above, which nodes are conditionally independent of each other given C? Select all that apply.



a) A, E
b) B, F
c) A, D
d) B, D
e) None of the above

Correct options: (e)-All pairs have an alternate route to each other that does not pass through C

Question-4

03:00

Consider the following statements about Hidden Markov Models (HMMs):
I. The ”Hidden” in HMM refers to the fact that the state transition probabilities are unknown.
II. The ”Markov” property means that the current state depends only on the previous state.
III. The ”Hidden” aspect relates to the underlying state sequence that is not directly observable.
IV. The ”Markov” in HMM indicates that the model uses matrix operations for calculations.
Which of the statements correctly describe the ”Hidden” and ”Markov” aspects of Hidden Markov Models?

a) I and II
b) I and IV
c) II and III
d) III and IV

Question-4

03:00

I. The ”Hidden” in HMM refers to the fact that the state transition probabilities are unknown. — it is known

II. The ”Markov” property means that the current state depends only on the previous state.

III. The ”Hidden” aspect relates to the underlying state sequence that is not directly observable.

IV. The ”Markov” in HMM indicates that the model uses matrix operations for calculations. – Markov means future state dependence on current state only

Question-4 - Correct answer

Consider the following statements about Hidden Markov Models (HMMs):
I. The ”Hidden” in HMM refers to the fact that the state transition probabilities are unknown.
II. The ”Markov” property means that the current state depends only on the previous state.
III. The ”Hidden” aspect relates to the underlying state sequence that is not directly observable.
IV. The ”Markov” in HMM indicates that the model uses matrix operations for calculations.
Which of the statements correctly describe the ”Hidden” and ”Markov” aspects of Hidden Markov Models?

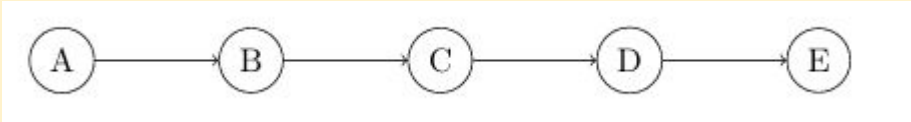
a) I and II
b) I and IV
c) II and III
d) III and IV

Correct options: (c)

Question-5

01:00

For the given graphical model, what is the optimal variable elimination order when trying to calculate P(E=e)



a) A, B, C, D

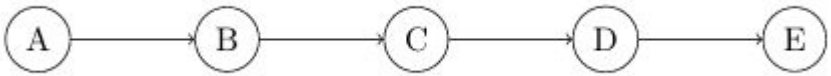
b) D, C, B, A

c) A, D, B, C

d) D, A, C, A

Question-5 - Correct answer

For the given graphical model, what is the optimal variable elimination order when trying to calculate P(E=e)



a) A, B, C, D

b) D, C, B, A

c) A, D, B, C

d) D, A, C, A

Correct options: (a)

Question-6

01:00

Consider the following statements regarding belief propagation:
I. Belief propagation is used to compute marginal probabilities in graphical models.
II. Belief propagation can be applied to both directed and undirected graphical models.
III. Belief propagation guarantees an exact solution when applied to loopy graphs.
IV. Belief propagation works by passing messages between nodes in a graph.
Which of the statements correctly describe the use of belief propagation?

a) I and II
b) II and III
c) I, II, and IV
d) I, III, and IV
e) II, III, and IV

Question-6

01:00

I. Belief propagation is used to compute marginal probabilities in graphical models.

II. Belief propagation can be applied to both directed and undirected graphical models.

III. Belief propagation guarantees an exact solution when applied to loopy graphs.-- exact solution for tree not loops

IV. Belief propagation works by passing messages between nodes in a graph.

Question-6 - Correct answer

Consider the following statements regarding belief propagation:
I. Belief propagation is used to compute marginal probabilities in graphical models.
II. Belief propagation can be applied to both directed and undirected graphical models.
III. Belief propagation guarantees an exact solution when applied to loopy graphs.
IV. Belief propagation works by passing messages between nodes in a graph.
Which of the statements correctly describe the use of belief propagation?

a) I and II
b) II and III
c) I, II, and IV
d) I, III, and IV
e) II, III, and IV

Correct options: (c)

Question-7

03:00

HMMs are used for finding these. Select all that apply

a) Probability of a given observation sequence
b) All possible hidden state sequences given an observation sequence
c) Most probable observation sequence given the hidden states
d) Most probable hidden states given the observation sequence

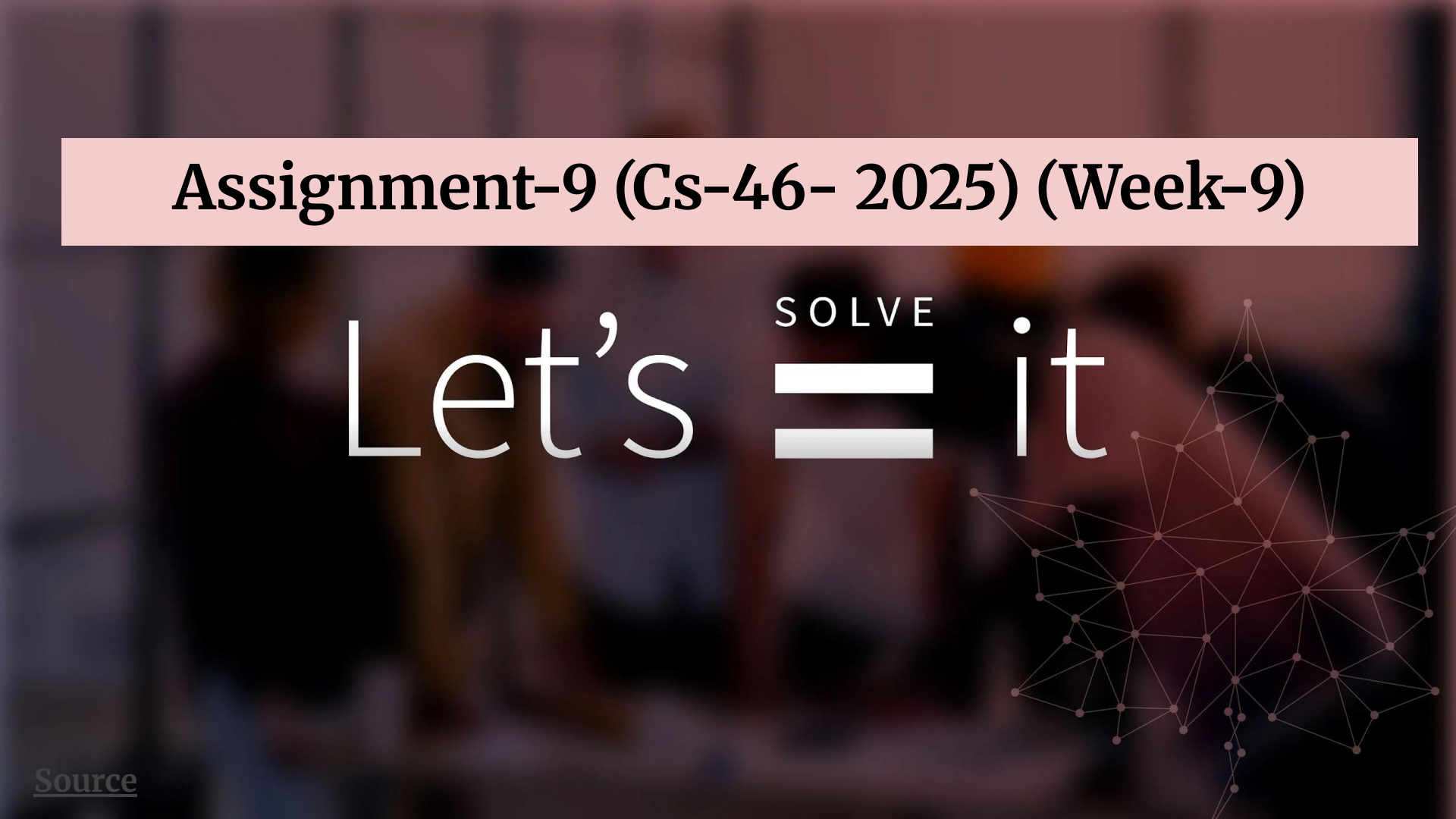
Question-7 - Correct answer

HMMs are used for finding these. Select all that apply

a) Probability of a given observation sequence
b) All possible hidden state sequences given an observation sequence
c) Most probable observation sequence given the hidden states
d) Most probable hidden states given the observation sequence

Correct options: (a)(d)

Assignment-9 (Cs-46- 2025) (Week-9)



Source

Question-1

01:00

Consider the Markov Random Field given below. We need to delete one edge (without deleting any nodes) so that in the resulting graph, B and F are independent given A. Which of these edges could be deleted to achieve this independence?

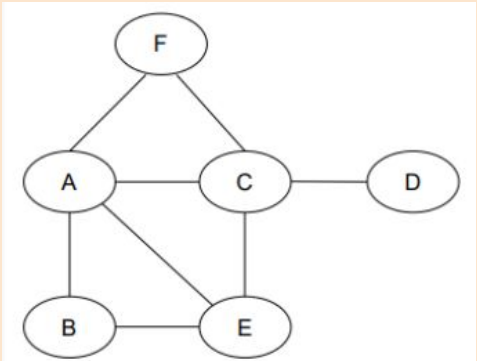
Note: In each option, we only delete one edge from the original graph.

a) AC

b) BE

c) CE

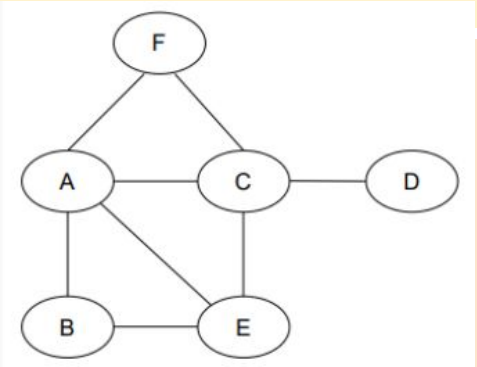
d) AE



Question-1- Correct answer

Consider the Markov Random Field given below. We need to delete one edge (without deleting any nodes) so that in the resulting graph, B and F are independent given A. Which of these edges could be deleted to achieve this independence?

Note: In each option, we only delete one edge from the original graph.



a) AC

b) BE

c) CE

d) AE

Correct options: (b,c)- break B-E-C-F path

Question-2

01:00

Consider the Markov Random Field from question 1. We need to delete one node (and also delete the edges incident with that node) so that in the resulting graph, B and C are independent given A. Which of these nodes could be deleted to achieve this independence?

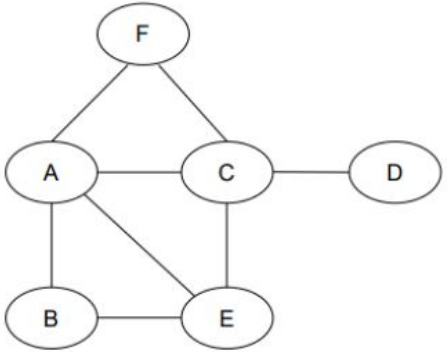
Note: In each option, we only delete one node and its incident edges from the original graph.

a) D

b) E

c) F

d) None of the above



Question-2- Correct answer

Consider the Markov Random Field from question 1. We need to delete one node (and also delete the edges incident with that node) so that in the resulting graph, B and C are independent given A. Which of these nodes could be deleted to achieve this independence?

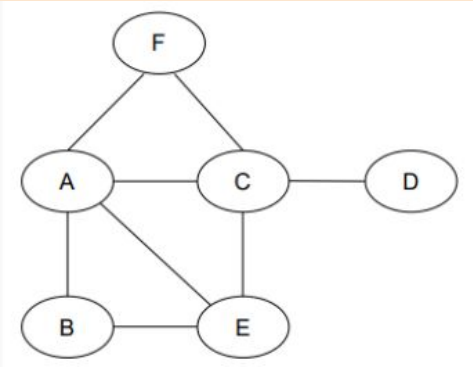
Note: In each option, we only delete one node and its incident edges from the original graph.

a) D

b) E

c) F

d) None of the above



Correct options: (b)

Question-3

01:00

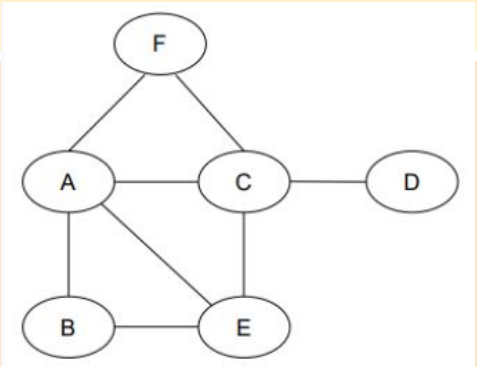
Consider the Markov Random Field from question 1. Which of the nodes has / have the largest Markov blanket (i.e. the Markov blanket with the most number of nodes)?

a) A
b) B
c) C
d) D
e) E
f) F

Question-3 - Correct answer

Consider the Markov Random Field from question 1. Which of the nodes has / have the largest Markov blanket (i.e. the Markov blanket with the most number of nodes)?

a) A
b) B
c) C
d) D
e) E
f) F

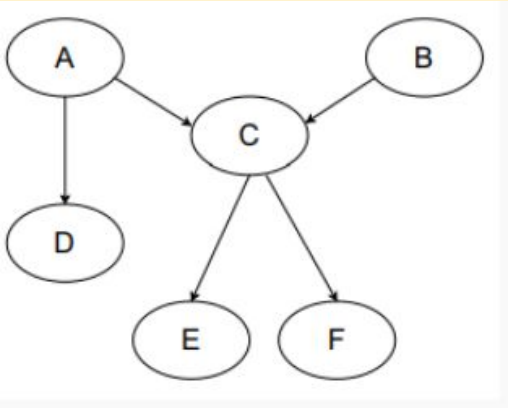


Correct options: (a,c)

Question-4

03:00

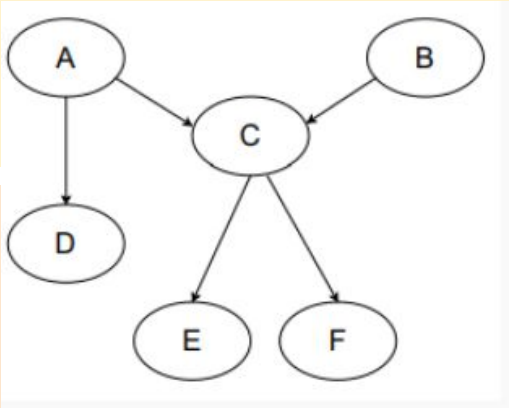
Consider the Bayesian Network given below. Which of the following independence relations hold?



a) A and B are independent if C is given
b) A and B are independent if no other variables are given
c) C and D are not independent if A is given
d) A and F are independent if C is given

Question-4 - Correct answer

Consider the Bayesian Network given below. Which of the following independence relations hold?



a) A and B are independent if C is given– collider case
b) A and B are independent if no other variables are given
c) C and D are not independent if A is given
d) A and F are independent if C is given

Correct options: (b,d)

Question-5

01:00

In the Bayesian Network from question 4, assume that every variable is binary. What is the number of independent parameters required to represent all the probability tables for the distribution?

a) 8
b) 12
c) 16
d) 24
e) 36

Question-5 - Correct answer

In the Bayesian Network from question 4, assume that every variable is binary. What is the number of independent parameters required to represent all the probability tables for the distribution?

a) 8
b) 12
c) 16
d) 24
e) 36

Correct options: (b)

Question-6

01:00

In the Bayesian Network from question 4, suppose variables A, C, E can take four possible values, while variables B, D, F are binary. What is the number of independent parameters required to represent all the probability tables for the distribution?

a) 24
b) 36
c) 48
d) 64
e) 84

Question-6 - Correct answer

In the Bayesian Network from question 4, suppose variables A, C, E can take four possible values, while variables B, D, F are binary. What is the number of independent parameters required to represent all the probability tables for the distribution?

a) 24
b) 36
c) 48
d) 64
e) 84

Correct options: (c)

Question-7

03:00

In the Bayesian Network from question 4, suppose all variables can take 4 values. What is the number of independent parameters required to represent all the probability tables for the distribution?

a) 72
b) 90
c) 108
d) 128
e) 144

Question-7 - Correct answer

In the Bayesian Network from question 4, suppose all variables can take 4 values. What is the number of independent parameters required to represent all the probability tables for the distribution?

a) 72
b) 90
c) 108
d) 128
e) 144

Correct options: (b)

Question-8

01:00

Consider the Bayesian Network from question 4. which of the given options are valid factorizations to calculate the marginal P (E = e) using variable elimination (need not be the optimal order)?

- [ ] $\sum_B P(B) \sum_A P(A) \sum_D P(D|A) \sum_C P(C|A,B) \sum_F P(E=e|C) P(F|C)$
- [ ] $\sum_A P(A) \sum_D P(D|A) \sum_B P(B) \sum_C P(C|A,B) \sum_F P(E=e|C) P(F|C)$
- [ ] $\sum_B P(B) \sum_A P(D|A) \sum_D P(A) \sum_F P(C|A,B) \sum_C P(E=e|C) P(F|C)$
- [ ] $\sum_A P(B) \sum_B P(D|A) \sum_D P(A) \sum_F P(C|A,B) \sum_C P(E=e|C) P(F|C)$
- [ ] $\sum_A P(A) \sum_B P(B) \sum_C P(C|A,B) \sum_D P(D|A) \sum_F P(E=e|C) P(F|C)$

Question-8 - Correct answer

Consider the Bayesian Network from question 4. which of the given options are valid factorizations to calculate the marginal P (E = e) using variable elimination (need not be the optimal order)?

Accepted Answers:

$$\sum_B P(B) \sum_A P(A) \sum_D P(D|A) \sum_C P(C|A,B) \sum_F P(E=e|C) P(F|C)$$

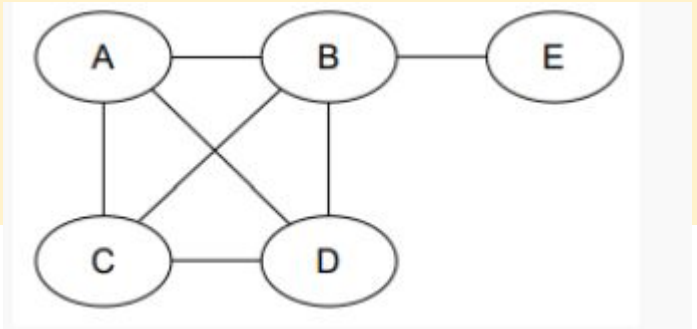
$$\sum_A P(A) \sum_D P(D|A) \sum_B P(B) \sum_C P(C|A,B) \sum_F P(E=e|C) P(F|C)$$

$$\sum_A P(A) \sum_B P(B) \sum_C P(C|A,B) \sum_D P(D|A) \sum_F P(E=e|C) P(F|C)$$

Question-9

01:00

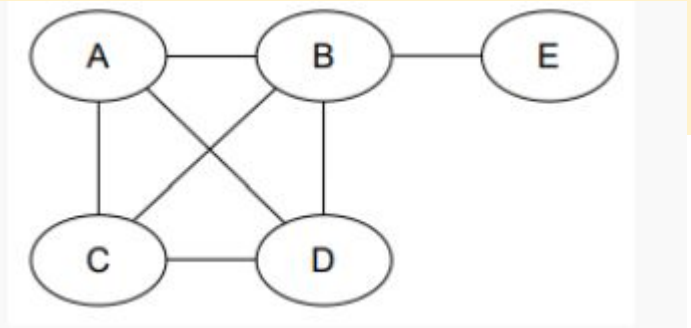
Consider the MRF given below. Which of the following factorization(s) of P (a, b, c, d, e) satisfies/satisfy the independence assumptions represented by this MRF?



- $P(a,b,c,d,e) = \frac{1}{Z}\psi_1(a,b,c,d)\psi_2(b,e)$
- $P(a,b,c,d,e) = \frac{1}{Z}\psi_1(b)\psi_2(a,c,d)\psi_3(a,b,e)$
- $P(a,b,c,d,e) = \frac{1}{Z}\psi_1(a,b)\psi_2(c,d)\psi_3(b,e)$
- $P(a,b,c,d,e) = \frac{1}{Z}\psi_1(a,b)\psi_2(c,d)\psi_3(b,d,e)$
- $P(a,b,c,d,e) = \frac{1}{Z}\psi_1(a,c)\psi_2(b,d)\psi_3(b,e)$
- $P(a,b,c,d,e) = \frac{1}{Z}\psi_1(c)\psi_2(b,e)\psi_3(b,a,d)$

Question-9 - Correct answer

Consider the MRF given below. Which of the following factorization(s) of P (a, b, c, d, e) satisfies/satisfy the independence assumptions represented by this MRF?



Accepted Answers:

$$P(a,b,c,d,e) = \frac{1}{Z}\psi_1(a,b,c,d)\psi_2(b,e)$$
$$P(a,b,c,d,e) = \frac{1}{Z}\psi_1(a,b)\psi_2(c,d)\psi_3(b,e)$$
$$P(a,b,c,d,e) = \frac{1}{Z}\psi_1(a,c)\psi_2(b,d)\psi_3(b,e)$$
$$P(a,b,c,d,e) = \frac{1}{Z}\psi_1(c)\psi_2(b,e)\psi_3(b,a,d)$$

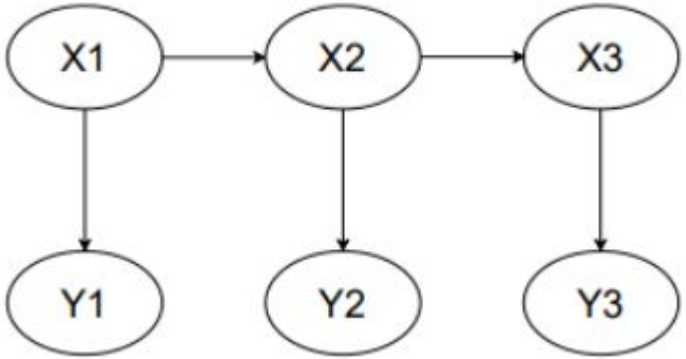
Question-10

03:00

The following figure shows an HMM for three time steps i

= 1, 2, 3. Suppose that it is used to perform part-of-speech tagging for a sentence. Which of the following statements is/are true?

a) The Xi variables represent parts-of-speech and the Yi variables represent the words in the sentence.
b) The Yi variables represent parts-of-speech and the Xi variables represent the words in the sentence.
c) The Xi variables are observed and the Yi variables need to be predicted.
d) The Yi variables are observed and the Xi variables need to be predicted.

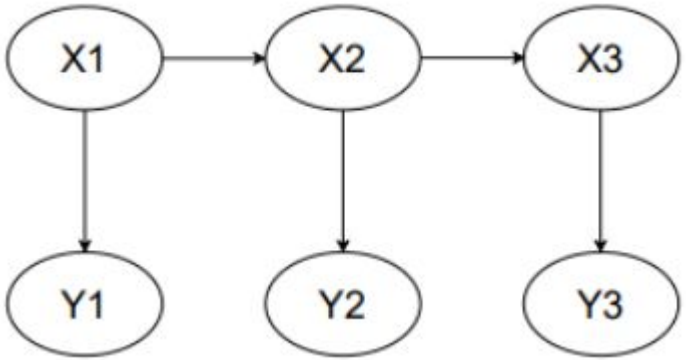


Question-10- Correct answer

The following figure shows an HMM for three time steps i

= 1, 2, 3. Suppose that it is used to perform part-of-speech tagging for a sentence. Which of the following statements is/are true?

a) The Xi variables represent parts-of-speech and the Yi variables represent the words in the sentence.
b) The Yi variables represent parts-of-speech and the Xi variables represent the words in the sentence.
c) The Xi variables are observed and the Yi variables need to be predicted.
d) The Yi variables are observed and the Xi variables need to be predicted



Correct options: (a,d)


THANK YOU

Suggestions and Feedback

Next Session:

Sunday:
28-Sept-2025
3:00 - 5:00 PM