TTIC 31170: Planning, Learning, and Estimation for Robotics and Artificial Intelligence

Spring 2021

Matthew Walter
TTI-Chicago

Lecture 2: Graphical Models (Continued)

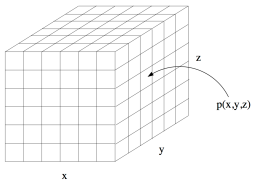
Key challenges

- **Represent** the world as a collection of random variables $X_1, \ldots, X_n$ with joint distribution $P(X_1, \ldots, X_n)$
  - How do you compactly represent the joint distribution?
  - Directed and undirected graphical models
- **Infer** distribution $P(X_i \mid X_1 = x_1, \ldots, X_m = x_m)$ over random variable(s) based upon evidence
  - How do you make inference tractable?
  - Exact vs. approximate
- **Learn** the distribution from available data

What is a good representation?

- Properties of a good representation:
  - Explicit
  - Modular (generalizable)
  - Permits efficient computation (i.e., learning & inference)

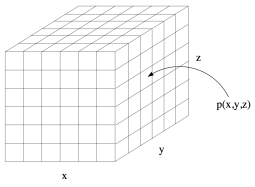
x_1	x_2	x_3	$p(x_1, x_2, x_3)$
0	0	0	.11
0	0	1	.02
	⋮		
1	1	1	.05



What is a good representation?

- Properties of a good representation:
  - Explicit
  - Modular (generalizable)
  - Permits efficient computation (i.e., learning & inference)

x_1	x_2	x_3	$p(x_1, x_2, x_3)$
0	0	0	.11
0	0	1	.02
	⋮		
1	1	1	.05

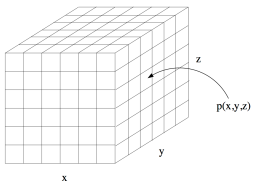


A naive representation requires a number of parameters that is exponential in the number of variables

What is a good representation?

- Properties of a good representation:
  - Explicit
  - Modular (generalizable)
  - Permits efficient computation (i.e., learning & inference)

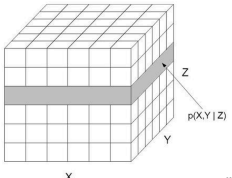
x_1	x_2	x_3	$p(x_1, x_2, x_3)$
0	0	0	.11
0	0	1	.02
	⋮		
1	1	1	.05



The amount of data required to learn these parameters is exponential in the number of variables

What is a good representation?

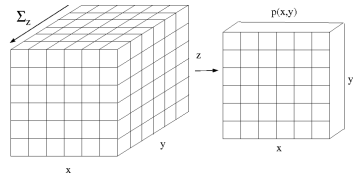
- Properties of a good representation:
  - Explicit
  - Modular (generalizable)
  - Permits efficient computation (i.e., learning & inference)



$$p(x, y \mid Z = z) = \frac{p(x, y, z)}{p(z)}$$

What is a good representation?

- Properties of a good representation:
  - Explicit
  - Modular (generalizable)
  - Permits efficient computation (i.e., learning & inference)

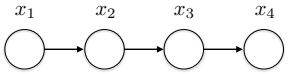


Cost of inference is exponential in the number of random variables

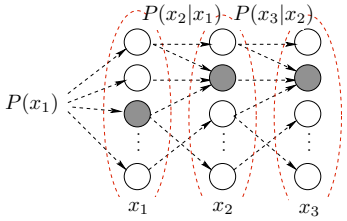
$$p(x,y) = \sum_{z \in \mathcal{Z}} p(x,y,z) \qquad p(x) = \sum_{y \in \mathcal{Y}} p(x,y)$$

Representation: Explicit

- Graphical model: Representation in terms of variables & dependencies:

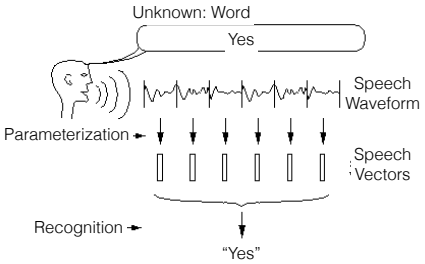


- Transition diagram: Representation in terms of state transitions



Graphical models: Examples

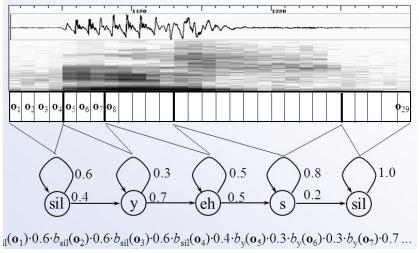
- Hidden Markov model as a Bayesian network
  - Example: Speech recognition



Graphical models: Examples

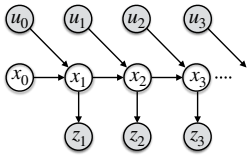
- Hidden Markov model as a Bayesian network
  - Example: Speech recognition

$$O = \{o_1, o_2, \ldots o_T\} \qquad P(w_i|O) = \frac{P(O,w_i)}{P(O)} = \frac{P(O|w_i)P(w_i)}{P(O)}$$



Graphical models: Examples

- Dynamic Bayesian networks
  - Example: Robot localization (and mapping)



x_t: robot's pose

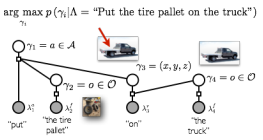
u_t: control input (action)

z_t: measurement

Graphical models: Examples

- Factor graph
  - Example: Natural language symbol grounding

Factor Graph

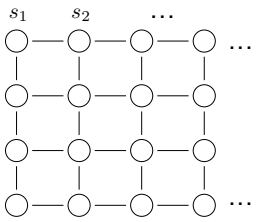


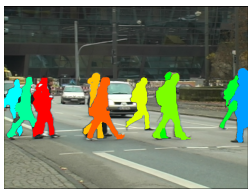
γ_i: Groundings (objects, actions)

λ_i: Phrases in instruction

Graphical models: Examples

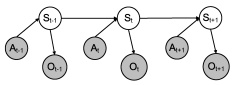
- Lattice models (e.g., Ising model) as a Markov random field
  - Example: Semantic image segmentation

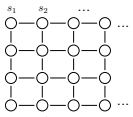


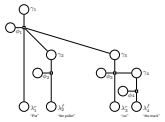


Graphical models

- Graph semantics:
  - graph ⟶ separation properties ⟶ independence
- Expression of probability distributions:
  - independence ⟶ family of distributions
- Inference and estimation:
  - graph structure ⟶ efficient computation







Topics

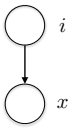
- Representations and graphical models
- Directed graphical models (Bayesian networks)
  - graphs and independence
- Undirected graphical models (Markov random fields)
  - graphs, independence
  - Bayesian networks as undirected models
- Inference

Bayesian networks

- Directed acyclic graphs (DAGs)
  - Nodes represent variables
  - Directed edges express dependencies

A mixture model as a Bayesian network

"i influences x"
"i causes x"
"x depends on i"



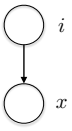
$$p(x, i) = p(x|i)p(i)$$

Bayesian networks

- Directed acyclic graphs (DAGs)
  - Nodes represent variables
  - Directed edges express dependencies

A mixture model as a Bayesian network

"i influences x"
"i causes x"
"x depends on i"



$$p(x, i) = p(x|i)p(i)$$

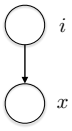
- Graph semantics:
  - graph ⟶ separation properties ⟶ independence
- Expression of probability distributions:
  - independence ⟶ family of distributions

Bayesian networks

- Directed acyclic graphs (DAGs)
  - Nodes represent variables
  - Directed edges express dependencies

A mixture model as a Bayesian network

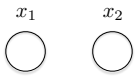
"i influences x"
"i causes x"
"x depends on i"



$$p(x, i) = p(x|i)p(i)$$

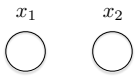
$$p(x_{\mathcal{V}}) = \prod_{v \in \mathcal{V}} p(x_v | x_{\pi_v}) \quad \text{where } \pi_v \text{ are the parents of } v$$

Example: Coin tosses



Example: Coin tosses

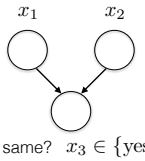
$P(x_1) = \begin{matrix} 0.5 \\ 0.5 \end{matrix}$



$P(x_2) = \begin{matrix} 0.5 \\ 0.5 \end{matrix}$

Example: Coin tosses

$P(x_1) = \begin{matrix} 0.5 \\ 0.5 \end{matrix}$



$P(x_2) = \begin{matrix} 0.5 \\ 0.5 \end{matrix}$

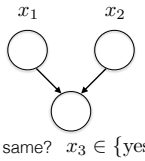
same? $x_3 \in \{\text{yes, no}\}$

$P(x_3|x_1, x_2) =$

	hh	ht	th	tt
y	1.0	0.0	0.0	1.0
n	0.0	1.0	1.0	0.0

Example: Coin tosses

$P(x_1) = \begin{matrix} 0.5 \\ 0.5 \end{matrix}$



$P(x_2) = \begin{matrix} 0.5 \\ 0.5 \end{matrix}$

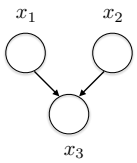
same? $x_3 \in \{\text{yes, no}\}$

$P(x_3|x_1, x_2) =$

	hh	ht	th	tt
y	1.0	0.0	0.0	1.0
n	0.0	1.0	1.0	0.0

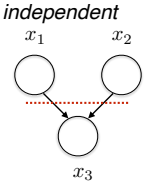
- Two levels of description
  1. graph structure (dependencies and independencies)
  2. associated probability distribution

What does the graph tell us?



What does the graph tell us?

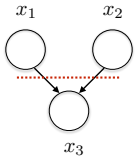
- x_1 and x_2 are ***marginally independent***



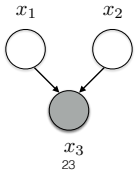
Knowing x_1 doesn't tell us anything about x_2

What does the graph tell us?

- x_1 and x_2 are *marginally independent*

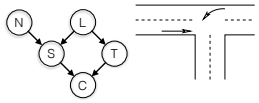


- x_1 and $x_2$become ***conditionally dependent*** if we know x_3



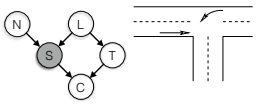
Driving example

N: X obeys rules of road
L: Traffic light
S: X decides to stop?
T: The other car turns left?
C: Crash?



Driving example

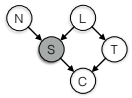
N: X obeys rules of road
L: Traffic light
S: X decides to stop?
T: The other car turns left?
C: Crash?



- If we know that X decided to stop (S), can knowing wether X obeys the law (N) tell us anything about the other car turning (T)?

Graph, independence, d-separation

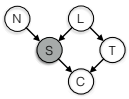
- Are N and T independent given S?



N: X obeys rules of road
L: Traffic light
S: X decides to stop?
T: The other car turns left?
C: Crash?

Graph, independence, d-separation

- Are N and T independent given S?

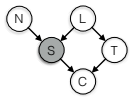


Definition: Variables N and T are ***d-separated*** given S if S separates them in the moralized ancestral graph

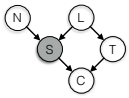
(d in d-separation stands for “dependence”)

Graph, independence, d-separation

- Are N and T independent given S?



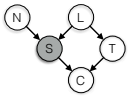
Definition: Variables N and T are ***d-separated*** given S if S separates them in the moralized ancestral graph



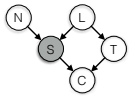
original

Graph, independence, d-separation

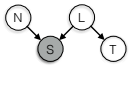
- Are N and T independent given S?



Definition: Variables N and T are ***d-separated*** given S if S separates them in the moralized ancestral graph



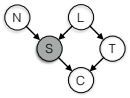
original



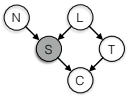
ancestral

Graph, independence, d-separation

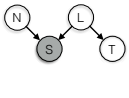
- Are N and T independent given S?



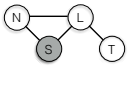
Definition: Variables N and T are ***d-separated*** given S if S separates them in the moralized ancestral graph



original



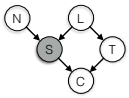
ancestral



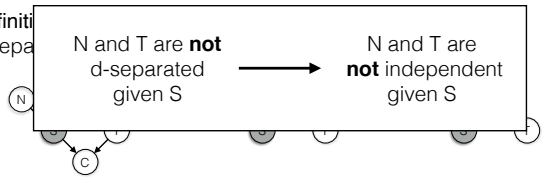
moralized ancestral

Graph, independence, d-separation

- Are N and T independent given S?



Definiti
S sepa



original

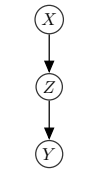
ancestral

moralized ancestral

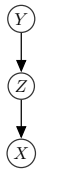
Graph, independence, d-separation

- Bayesian network structure implies conditional independencies

Cascade
(Markov chain; causal/evidential trail)

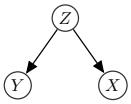


$(X \perp Y \mid Z)$



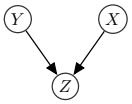
$(Y \perp X \mid Z)$

Common Cause



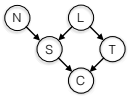
$(X \perp Y \mid Z)$

Common Effect
(v-structure)



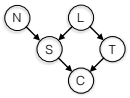
$(X \not\perp Y \mid Z)$

Graphs and distributions



- A graph is a compact representation of a large collection of independence properties

Graphs and distributions



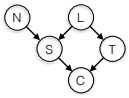
- A graph is a compact representation of a large collection of independence properties

Theorem: Any probability distribution that is consistent with a directed graph G has to factor according to "node given parents":

$$P(X|G) = \prod_{i=1}^{d} P(x_i|x_{\pi_i})$$

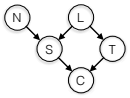
where π_i are the parents of x_i and d is the number of nodes in the graph

Graphs and distributions



$$P(X|G) = \prod_{i=1}^{d} P(x_i|x_{\pi_i})$$

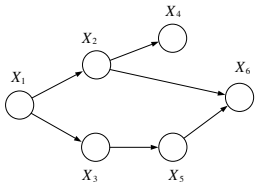
Graphs and distributions



$$P(X|G) = \prod_{i=1}^{d} P(x_i|x_{\pi_i})$$

$$P(CSTNL) = P(C|S,T)P(T|L)P(S|N,L)P(N)P(L)$$

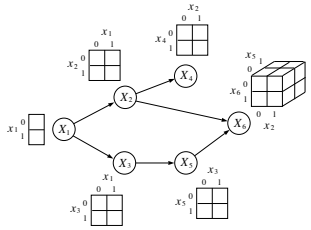
Graphs and distributions



Representational (storage, learning, & inference) complexity

- **Joint distribution**: Exponential in the number of variables
- **Bayesian Network**: Exponential in the number of parents of each node, linear in the number of nodes

Graphs and distributions

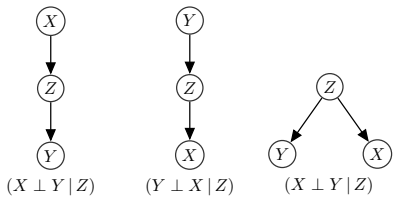


Representational (storage, learning, & inference) complexity

- **Joint distribution**: Exponential in the number of variables
- **Bayesian Network**: Exponential in the number of parents of each node, linear in the number of nodes

Graphs and distributions

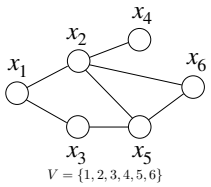
- Different Bayesian networks can encode the same conditional independencies (and the same distributions)



Topics

- Representations and graphical models
- Directed graphical models (Bayesian networks)
  - graphs and independence
- Undirected graphical models (Markov random fields)
  - graphs, independence
  - Bayesian networks as undirected models
- Inference

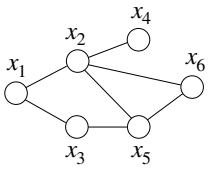
Undirected graphical models



$V = \{1, 2, 3, 4, 5, 6\}$

$G = (V, E)$

Undirected graphical models

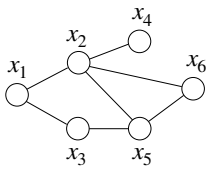


$G = (V, E)$

Definition: A ***clique*** $C \subseteq V$ is a subset of fully-connected nodes

$$C = \big\{\{1,2\}, \{1,3\}, \{2,4\}, \{3,5\}, \{2,5,6\}\big\}$$

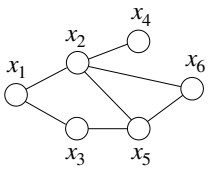
Undirected graphical models



$G = (V, E)$

Definition: A non-negative ***potential function*** $\psi_C(x_C)$ is associated with each clique C

Undirected graphical models



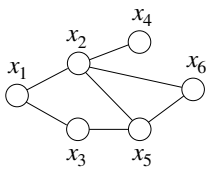
$G = (V, E)$

Theorem: The probability distribution for an undirected graph can be expressed as a product of clique potentials

$$p(x_V) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(x_C)$$

where Z is a normalization constant

Undirected graphical models

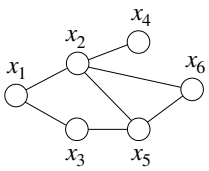


$G = (V, E)$

$$p(x_V) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(x_C)$$

$$p(x_V) = \frac{1}{Z} \psi(x_1, x_2) \psi(x_1, x_3) \psi(x_2, x_4) \psi(x_3, x_5) \psi(x_2, x_5, x_6)$$

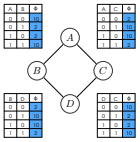
Undirected graphical models



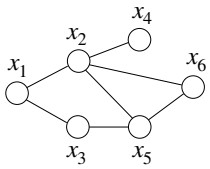
$G = (V, E)$

Definition: A non-negative *potential function* $\psi_C(x_C)$ is associated with each clique C

Example: Tabular



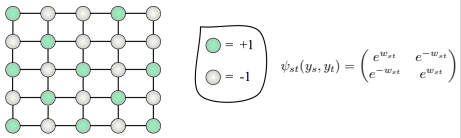
Undirected graphical models



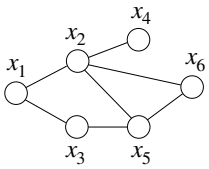
$G = (V, E)$

Definition: A non-negative *potential function* $\psi_C(x_C)$ is associated with each clique C

Example: Ising Model



Undirected graphical models



$G = (V, E)$

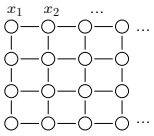
Definition: A non-negative *potential function* $\psi_C(x_C)$ is associated with each clique C

Example: Gaussian Markov random field

$$\prod_i \exp\left(-\frac{1}{2}(\Lambda_{ii}x_i^2 + 2\eta_i x_i)\right) \cdot \prod_{i,j:i\neq j} \exp\left(-\Lambda_{ij}x_i x_j\right)$$

Undirected graphical models

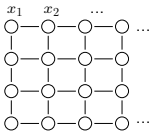
- For example, a lattice model with binary variables $x_i \in \{-1, 1\}$ (e.g., foreground vs. background) and pairwise edges



$$p(x_1, \ldots, x_n) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(x_C)$$
$$= \frac{1}{Z} \prod_{(i,j) \in E} \psi_{i,j}(x_i, x_j)$$

Undirected graphical models

- For example, a lattice model with binary variables $x_i \in \{-1, 1\}$ (e.g., foreground vs. background) and pairwise edges

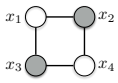


$$p(x_1, \ldots, x_n) = \frac{1}{Z} \prod_{(i,j) \in E} \psi_{i,j}(x_i, x_j)$$
$$= \frac{1}{Z} \prod_{(i,j) \in E} \exp\left(J_{ij} x_i, x_j\right)$$

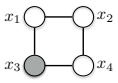
where J_{ij} specifies the "interaction strength" between x_i and x_j

Graph semantics

- Graph semantics for undirected graph come from separation



x_1 and x_4 are independent given x_2 and x_3

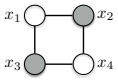


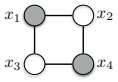
x_1 and x_4 are not independent given x_3

Graph semantics: Comparison

- Directed and undirected graphs are complementary

The following two independence properties can not be captured *simultaneously* with a Bayesian network



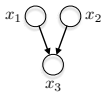


Graph semantics: Comparison

- Directed and undirected graphs are complementary

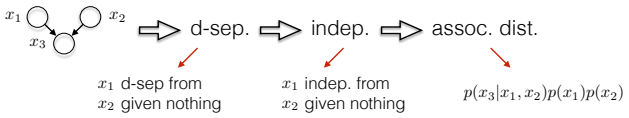
  The following two independence properties can not be captured *simultaneously* with a Bayesian network

  An undirected graph can not express marginal independence



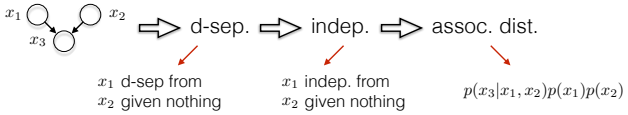
Summary: Graphs and Probabilities

- Directed graphical models (Bayesian networks)

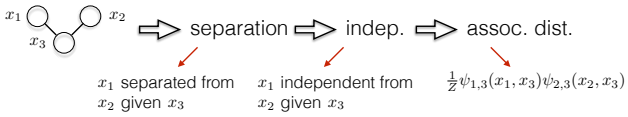


Summary: Graphs and Probabilities

- Directed graphical models (Bayesian networks)

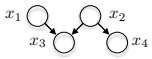


- Undirected graphical models (Markov random fields)



Graph transformations

- Directed graphical models can be converted into undirected graphical models via ***moralization*** (aka "marry the parents"):



$p(x_3|x_1,x_2)p(x_4|x_2)p(x_1)p(x_2)$

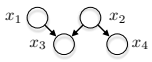
Graph transformations

- Directed graphical models can be converted into undirected graphical models via ***moralization*** (aka "marry the parents"):

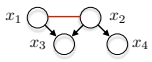
$p(x_3|x_1,x_2)p(x_4|x_2)p(x_1)p(x_2)$

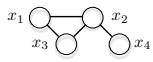
Graph transformations

- Directed graphical models can be converted into undirected graphical models via ***moralization*** (aka "marry the parents"):



$p(x_3|x_1,x_2)p(x_4|x_2)p(x_1)p(x_2)$



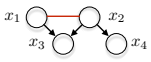


$\frac{1}{Z}\psi_{1,2,3}(x_1,x_2,x_3)\psi_{2,4}(x_2,x_4)$

Graph transformations

- Directed graphical models can be converted into undirected graphical models via ***moralization*** (aka "marry the parents"):

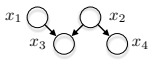
$p(x_3|x_1,x_2)p(x_4|x_2)p(x_1)p(x_2)$



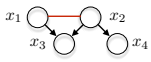
$\frac{1}{Z}\psi_{1,2,3}(x_1,x_2,x_3)\psi_{2,4}(x_2,x_4)$

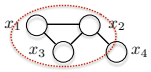
Graph transformations

- Directed graphical models can be converted into undirected graphical models via ***moralization*** (aka "marry the parents"):



$p(x_3|x_1,x_2)p(x_4|x_2)p(x_1)p(x_2)$





$\frac{1}{Z}\psi_{1,2,3}(x_1,x_2,x_3)\psi_{2,4}(x_2,x_4)$

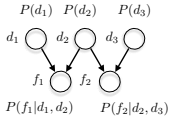
- The distribution doesn't change, only the graph
- The undirected graph is consistent with the distribution associated with the original graph

Topics

- Representations and graphical models
- Directed graphical models (Bayesian networks)
  - graphs and independence
- Undirected graphical models (Markov Random Fields)
  - graphs, independence
  - Bayesian networks as undirected models
- Inference

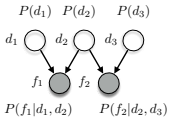
Inference

- Event prediction example: Binary event variables d and possible observations f



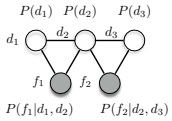
Inference

- Event prediction example: Binary event variables d and possible observations f



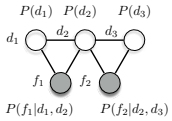
Inference

- Moralize the directed graphical model



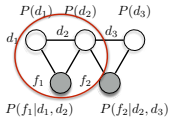
Inference

- Moralize the directed graphical model



Inference

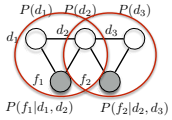
- Moralize the directed graphical model



$$\psi_{1,2}(d_1,d_2) = P(d_1)P(d_2)P(f_1^*|d_1,d_2)$$

Inference

- Moralize the directed graphical model

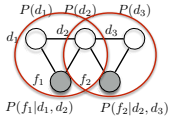


$$\psi_{1,2}(d_1,d_2) = P(d_1)P(d_2)P(f_1^*|d_1,d_2)$$

$$\psi_{2,3}(d_2,d_3) = P(d_3)P(f_2^*|d_2,d_3)$$

Inference

- Moralize the directed graphical model



$$\psi_{1,2}(d_1,d_2) = P(d_1)P(d_2)P(f_1^*|d_1,d_2)$$

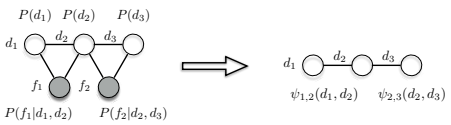
$$\psi_{2,3}(d_2,d_3) = P(d_3)P(f_2^*|d_2,d_3)$$

- Join distribution as a product of "interaction potentials"

$$P(d_1,d_2,d_3,\text{data}) = \psi_{1,2}(d_1,d_2)\cdot\psi_{2,3}(d_2,d_3)$$

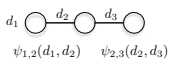
Inference

- Moralize the directed graphical model



$$P(d_1,d_2,d_3,\text{data}) = \psi_{1,2}(d_1,d_2)\cdot\psi_{2,3}(d_2,d_3)$$

Marginalization and messages



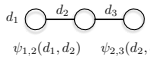
- It suffices to calculate the following marginals:

$$P(d_1,\text{data}) = \sum_{d_2,d_3} P(d_1,d_2,d_3,\text{data})$$
$$P(d_2,\text{data}) = \sum_{d_1,d_3} P(d_1,d_2,d_3,\text{data})$$
$$P(d_3,\text{data}) = \sum_{d_1,d_2} P(d_1,d_2,d_3,\text{data})$$

We can use these marginals to calculate desired probabilities:

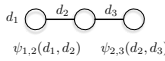
$$P(d_1|\text{data}) = \frac{P(d_1,\text{data})}{P(\text{data})} = \frac{P(d_1,\text{data})}{\sum_{d_1'} P(d_1',\text{data})}$$

Marginalization and messages



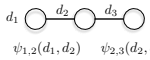
$$P(d_2, d_3, \text{data}) = \sum_{d_1} P(d_1, d_2, d_3, \text{data})$$

Marginalization and messages



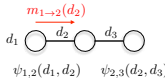
$$\begin{aligned} P(d_2, d_3, \text{data}) &= \sum_{d_1} P(d_1, d_2, d_3, \text{data}) \\ &= \sum_{d_1} \psi_{12}(d_1, d_2) \cdot \psi_{23}(d_2, d_3) \end{aligned}$$

Marginalization and messages



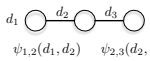
$$\begin{aligned} P(d_2, d_3, \text{data}) &= \sum_{d_1} P(d_1, d_2, d_3, \text{data}) \\ &= \sum_{d_1} \psi_{12}(d_1, d_2) \cdot \psi_{23}(d_2, d_3) \\ &= \left[\sum_{d_1} \psi_{12}(d_1, d_2) \right] \cdot \psi_{23}(d_2, d_3) \end{aligned}$$

Marginalization and messages



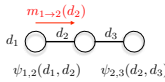
$$\begin{aligned} P(d_2, d_3, \text{data}) &= \sum_{d_1} P(d_1, d_2, d_3, \text{data}) \\ &= \sum_{d_1} \psi_{12}(d_1, d_2) \cdot \psi_{23}(d_2, d_3) \\ &= \left[\sum_{d_1} \psi_{12}(d_1, d_2) \right] \cdot \psi_{23}(d_2, d_3) \\ &= m_{1\to 2}(d_2) \cdot \psi_{23}(d_2, d_3) \end{aligned}$$

Marginalization and messages



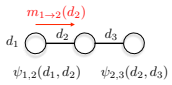
$$P(d_3, \text{data}) = \sum_{d_2} P(d_2, d_3, \text{data})$$

Marginalization and messages



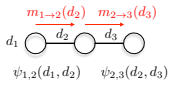
$$\begin{aligned} P(d_3, \text{data}) &= \sum_{d_2} P(d_2, d_3, \text{data}) \\ &= \sum_{d_2} m_{1\to 2}(d_2) \cdot \psi_{23}(d_2, d_3) \end{aligned}$$

Marginalization and messages



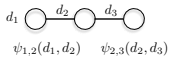
$$
\begin{aligned}
P(d_3, \text{data}) &= \sum_{d_2} P(d_2, d_3, \text{data}) \\
&= \sum_{d_2} m_{1\to 2}(d_2) \cdot \psi_{23}(d_2, d_3) \\
&= \left[\sum_{d_2} m_{1\to 2}(d_2) \cdot \psi_{23}(d_2, d_3) \right]
\end{aligned}
$$

Marginalization and messages



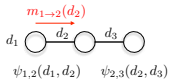
$$
\begin{aligned}
P(d_3, \text{data}) &= \sum_{d_2} P(d_2, d_3, \text{data}) \\
&= \sum_{d_2} m_{1\to 2}(d_2) \cdot \psi_{23}(d_2, d_3) \\
&= \left[\sum_{d_2} m_{1\to 2}(d_2) \cdot \psi_{23}(d_2, d_3) \right] \\
&= m_{2\to 3}(d_3)
\end{aligned}
$$

Marginalization and messages



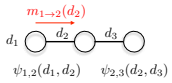
$$
\begin{aligned}
P(d_2, \text{data}) &= \sum_{d_3} P(d_2, d_3, \text{data})
\end{aligned}
$$

Marginalization and messages



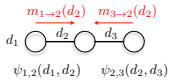
$$
\begin{aligned}
P(d_2, \text{data}) &= \sum_{d_3} P(d_2, d_3, \text{data}) \\
&= \sum_{d_3} m_{1\to 2}(d_2) \cdot \psi_{23}(d_2, d_3)
\end{aligned}
$$

Marginalization and messages



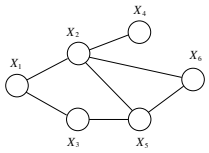
$$
\begin{aligned}
P(d_2, \text{data}) &= \sum_{d_3} P(d_2, d_3, \text{data}) \\
&= \sum_{d_3} m_{1\to 2}(d_2) \cdot \psi_{23}(d_2, d_3) \\
&= m_{1\to 2}(d_2) \cdot \left[\sum_{d_3} \psi_{23}(d_2, d_3) \right]
\end{aligned}
$$

Marginalization and messages

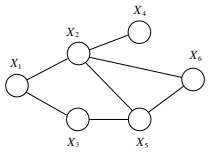


$$
\begin{aligned}
P(d_2, \text{data}) &= \sum_{d_3} P(d_2, d_3, \text{data}) \\
&= \sum_{d_3} m_{1\to 2}(d_2) \cdot \psi_{23}(d_2, d_3) \\
&= m_{1\to 2}(d_2) \cdot \left[\sum_{d_3} \psi_{23}(d_2, d_3) \right] \\
&= m_{1\to 2}(d_2) \cdot m_{3\to 2}(d_2)
\end{aligned}
$$

Marginalization

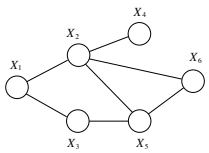


Marginalization



$$p(x_1) = \sum_{x_2}\sum_{x_3}\sum_{x_5}\sum_{x_5}\sum_{x_6} p(x_1, x_2, x_3, x_4, x_5, x_6)$$
$$\propto \sum_{x_2}\sum_{x_3}\sum_{x_5}\sum_{x_5}\sum_{x_6} \psi(x_1, x_2)\psi(x_1, x_3)\psi(x_2, x_4)\psi(x_3, x_5)\psi(x_2, x_5, x_6)$$

Marginalization



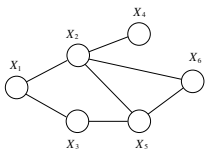
$$p(x_1) = \sum_{x_2}\sum_{x_3}\sum_{x_5}\sum_{x_5}\sum_{x_6} p(x_1, x_2, x_3, x_4, x_5, x_6)$$
$$\propto \sum_{x_2}\sum_{x_3}\sum_{x_5}\sum_{x_5}\sum_{x_6} \psi(x_1, x_2)\psi(x_1, x_3)\psi(x_2, x_4)\psi(x_3, x_5)\psi(x_2, x_5, x_6)$$

Nominally, each sum is applied to 6 variables $\Rightarrow$ $\mathcal{O}(r^6)$

Marginalization

- We can reduce the complexity by exploiting the distribution's structure to be smart about the marginalization order

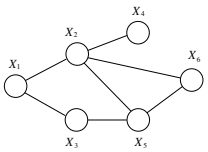
$$p(x_1) = \frac{1}{Z}\sum_{x_2}\psi(x_1, x_2)\sum_{x_3}\psi(x_1, x_3)\sum_{x_4}\psi(x_2, x_4)\sum_{x_5}\psi(x_3, x_5)\sum_{x_6}\psi(x_2, x_5, x_6)$$



Marginalization

- We can reduce the complexity by exploiting the distribution's structure to be smart about the marginalization order

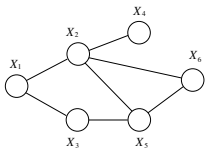
$$p(x_1) = \frac{1}{Z}\sum_{x_2}\psi(x_1, x_2)\sum_{x_3}\psi(x_1, x_3)\sum_{x_4}\psi(x_2, x_4)\sum_{x_5}\psi(x_3, x_5)\sum_{x_6}\psi(x_2, x_5, x_6)$$
$$= \frac{1}{Z}\sum_{x_2}\psi(x_1, x_2)\sum_{x_3}\psi(x_1, x_3)\sum_{x_4}\psi(x_2, x_4)\sum_{x_5}\psi(x_3, x_5)m_6(x_2, x_5)$$



Marginalization

- We can reduce the complexity by exploiting the distribution's structure to be smart about the marginalization order

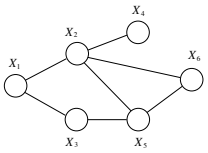
$$p(x_1) = \frac{1}{Z}\sum_{x_2}\psi(x_1, x_2)\sum_{x_3}\psi(x_1, x_3)\sum_{x_4}\psi(x_2, x_4)\sum_{x_5}\psi(x_3, x_5)\sum_{x_6}\psi(x_2, x_5, x_6)$$
$$= \frac{1}{Z}\sum_{x_2}\psi(x_1, x_2)\sum_{x_3}\psi(x_1, x_3)\sum_{x_4}\psi(x_2, x_4)\sum_{x_5}\psi(x_3, x_5)m_6(x_2, x_5)$$
$$= \frac{1}{Z}\sum_{x_2}\psi(x_1, x_2)\sum_{x_3}\psi(x_1, x_3)m_5(x_2, x_3)\sum_{x_4}\psi(x_2, x_4)$$



Marginalization

- We can reduce the complexity by exploiting the distribution's structure to be smart about the marginalization order

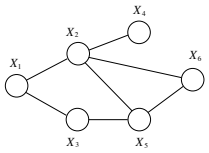
$$
\begin{aligned}
p(x_1) &= \frac{1}{Z}\sum_{x_2}\psi(x_1,x_2)\sum_{x_3}\psi(x_1,x_3)\sum_{x_4}\psi(x_2,x_4)\sum_{x_5}\psi(x_3,x_5)\sum_{x_6}\psi(x_2,x_5,x_6) \\
&= \frac{1}{Z}\sum_{x_2}\psi(x_1,x_2)\sum_{x_3}\psi(x_1,x_3)\sum_{x_4}\psi(x_2,x_4)\sum_{x_5}\psi(x_3,x_5)m_6(x_2,x_5) \\
&= \frac{1}{Z}\sum_{x_2}\psi(x_1,x_2)\sum_{x_3}\psi(x_1,x_3)m_5(x_2,x_3)\sum_{x_4}\psi(x_2,x_4) \\
&= \frac{1}{Z}\sum_{x_2}\psi(x_1,x_2)m_4(x_2)\sum_{x_3}\psi(x_1,x_3)m_5(x_2,x_3)
\end{aligned}
$$



Marginalization

- We can reduce the complexity by exploiting the distribution's structure to be smart about the marginalization order

$$
\begin{aligned}
p(x_1) &= \frac{1}{Z}\sum_{x_2}\psi(x_1,x_2)\sum_{x_3}\psi(x_1,x_3)\sum_{x_4}\psi(x_2,x_4)\sum_{x_5}\psi(x_3,x_5)\sum_{x_6}\psi(x_2,x_5,x_6) \\
&= \frac{1}{Z}\sum_{x_2}\psi(x_1,x_2)\sum_{x_3}\psi(x_1,x_3)\sum_{x_4}\psi(x_2,x_4)\sum_{x_5}\psi(x_3,x_5)m_6(x_2,x_5) \\
&= \frac{1}{Z}\sum_{x_2}\psi(x_1,x_2)\sum_{x_3}\psi(x_1,x_3)m_5(x_2,x_3)\sum_{x_4}\psi(x_2,x_4) \\
&= \frac{1}{Z}\sum_{x_2}\psi(x_1,x_2)m_4(x_2)\sum_{x_3}\psi(x_1,x_3)m_5(x_2,x_3) \\
&= \frac{1}{Z}\sum_{x_2}\psi(x_1,x_2)m_4(x_2)m_3(x_1,x_2)
\end{aligned}
$$

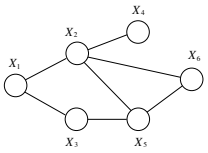


Marginalization

- We can reduce the complexity by exploiting the distribution's structure to be smart about the marginalization order

$$
\begin{aligned}
p(x_1) &= \frac{1}{Z}\sum_{x_2}\psi(x_1,x_2)\sum_{x_3}\psi(x_1,x_3)\sum_{x_4}\psi(x_2,x_4)\sum_{x_5}\psi(x_3,x_5)\boxed{\sum_{x_6}\psi(x_2,x_5,x_6)} \\
&= \frac{1}{Z}\sum_{x_2}\psi(x_1,x_2)\sum_{x_3}\psi(x_1,x_3)\sum_{x_4}\psi(x_2,x_4)\sum_{x_5}\psi(x_3,x_5)m_6(x_2,x_5) \\
&= \frac{1}{Z}\sum_{x_2}\psi(x_1,x_2)\sum_{x_3}\psi(x_1,x_3)m_5(x_2,x_3)\sum_{x_4}\psi(x_2,x_4) \\
&= \frac{1}{Z}\sum_{x_2}\psi(x_1,x_2)m_4(x_2)\sum_{x_3}\psi(x_1,x_3)m_5(x_2,x_3) \\
&= \frac{1}{Z}\sum_{x_2}\psi(x_1,x_2)m_4(x_2)m_3(x_1,x_2) \\
&= \frac{1}{Z}m_2(x_1),
\end{aligned}
$$

$r \times r \times r$ matrix



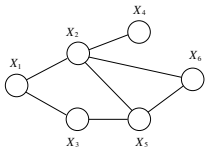
Marginalization

- We can reduce the complexity by exploiting the distribution's structure to be smart about the marginalization order

$$
\begin{aligned}
p(x_1) &= \frac{1}{Z}\sum_{x_2}\psi(x_1,x_2)\sum_{x_3}\psi(x_1,x_3)\sum_{x_4}\psi(x_2,x_4)\sum_{x_5}\psi(x_3,x_5)\boxed{\sum_{x_6}\psi(x_2,x_5,x_6)} \\
&= \frac{1}{Z}\sum_{x_2}\psi(x_1,x_2)\sum_{x_3}\psi(x_1,x_3)\sum_{x_4}\psi(x_2,x_4)\sum_{x_5}\psi(x_3,x_5)m_6(x_2,x_5) \\
&= \frac{1}{Z}\sum_{x_2}\psi(x_1,x_2)\sum_{x_3}\psi(x_1,x_3)m_5(x_2,x_3)\sum_{x_4}\psi(x_2,x_4) \\
&= \frac{1}{Z}\sum_{x_2}\psi(x_1,x_2)m_4(x_2)\sum_{x_3}\psi(x_1,x_3)m_5(x_2,x_3) \\
&= \frac{1}{Z}\sum_{x_2}\psi(x_1,x_2)m_4(x_2)m_3(x_1,x_2) \\
&= \frac{1}{Z}m_2(x_1),
\end{aligned}
$$

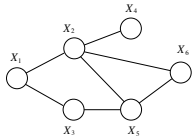
$r \times r \times r$ matrix

$\mathcal{O}(r^3)$



Elimination order

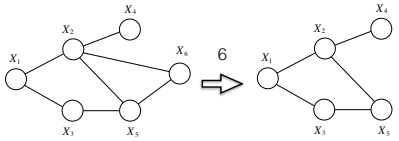
- Summations (eliminations) introduce intermediate edges



$$
p(x_1) = \frac{1}{Z}\sum_{x_2}\psi(x_1,x_2)\sum_{x_3}\psi(x_1,x_3)\sum_{x_4}\psi(x_2,x_4)\sum_{x_5}\psi(x_3,x_5)\sum_{x_6}\psi(x_2,x_5,x_6)
$$

Elimination order

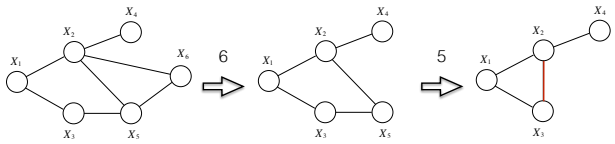
- Summations (eliminations) introduce intermediate edges



$$
\begin{aligned}
p(x_1) &= \frac{1}{Z}\sum_{x_2}\psi(x_1,x_2)\sum_{x_3}\psi(x_1,x_3)\sum_{x_4}\psi(x_2,x_4)\sum_{x_5}\psi(x_3,x_5)\sum_{x_6}\psi(x_2,x_5,x_6) \\
&= \frac{1}{Z}\sum_{x_2}\psi(x_1,x_2)\sum_{x_3}\psi(x_1,x_3)\sum_{x_4}\psi(x_2,x_4)\sum_{x_5}\psi(x_3,x_5)m_6(x_2,x_5)
\end{aligned}
$$

Elimination order

- Summations (eliminations) introduce intermediate edges



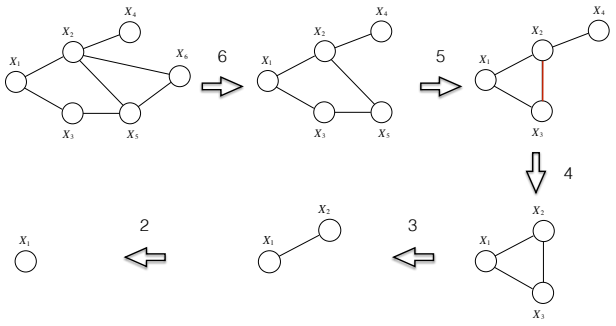
$$p(x_1) = \frac{1}{Z}\sum_{x_2}\psi(x_1,x_2)\sum_{x_3}\psi(x_1,x_3)\sum_{x_4}\psi(x_2,x_4)\sum_{x_5}\psi(x_3,x_5)\sum_{x_6}\psi(x_2,x_5,x_6)$$

$$= \frac{1}{Z}\sum_{x_2}\psi(x_1,x_2)\sum_{x_3}\psi(x_1,x_3)\sum_{x_4}\psi(x_2,x_4)\sum_{x_5}\psi(x_3,x_5)m_6(x_2,x_5)$$

$$= \frac{1}{Z}\sum_{x_2}\psi(x_1,x_2)\sum_{x_3}\psi(x_1,x_3)m_5(x_2,x_3)\sum_{x_4}\psi(x_2,x_4)$$

Elimination order

- Summations (eliminations) introduce intermediate edges



Elimination order

- Worst-case summation complexity is a function of the largest intermediate clique (e.g., 3 element clique)
- Goal: Choose summation order (*elimination order*) with smallest intermediate clique
  - e.g., we went from $\mathcal{O}(r^6)$ to $\mathcal{O}(r^3)$

Definition: The *treewidth* is the minimum over elimination order of the maximal clique size

- Note parallels with solving set of linear equations

Elimination order

- Worst-case summation complexity is a function of the largest intermediate clique (e.g., 3 element clique)
- Goa… st inte…
  - e…

Finding elimination order that achieves treewidth is NP-hard

Defi… of the maximal clique size

- Note parallels with solving set of linear equations