Probabilistic Graphical Models and Causal Inference

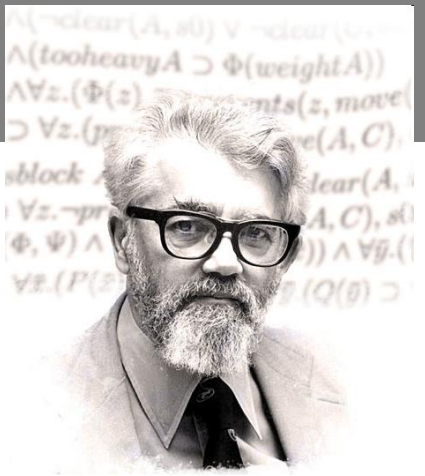
Episode 1: Probabilistic Graphical Models

Marco Piastra

This presentation can be downloaded here:
https://vision.unipv.it/AI/AIRG.html

Prologue:
Causal Inference ?

"Artificial Intelligence" (first appearance of the term)



[Image from Wikipedia]

"We propose that a two-month, ten-man study
of **artificial intelligence** carried out during the summer of 1956 [...]

The study is to proceed on the basis of the conjecture
that every aspect of learning or any other feature of **intelligence**
can in principle be ***so precisely described***
that a machine can be made to ***simulate*** it. "

[John McCarthy et al., 1955, emphasis added]

Reasoning as a Process

A line of reasoning, in formal logic, is translated into a *process* made of *steps*

Each *step* is of *inference*: from some elements, derive some others

- What is the purpose of logic?

  To distinguish correct reasoning from incorrect reasoning

  "No false conclusions from true premises"

Causal Model

Namely, the objective of what we are talking about

A *causal model* is a conceptual tool that we can align with actual *observations*,

that allows us to perform virtual *interventions* and estimate their effects

and to evaluate possible *counterfactual* worlds
("*What if one element was different from actuality?*")

All of this in a precise and formal framework,
in which each inference step *can be performed,*
under specific prerequisites

Using probability theory *as the basic formalism*

Apropos dependence and *independence*

Random variables: notation

(*) In a finitary setting

- Random variables, events and event spaces

A random variable X assumes values in a set $\mathcal{X}$
and generates an event space $X = x,\ x \in \mathcal{X}$

Probability:

This is the probability measure over the event space *generated by the* random variable X

$$P(X)$$

This the probability (i.e., a value in $[0,1]$ *) associated to the* event $X = x$

$$P(X = x)$$

This is the probability measure over the event space *generated by the* random variables X and Y *as joint occurrences of* X *and the event* $Y = y$

$$P(X, Y = y)$$

Marginalization

(*) In a finitary setting

Removing a random variable from a joint distribution

Given a joint probability distribution

$$P(X, Y)$$

The *marginal probability* $P(X)$ is obtained via a summation:

$$P(X) := \sum_{y} P(X, Y = y)$$

A marginal probability can be a joint probability as well …

$$P(X, Y) := \sum_{z} P(X, Y, Z = z)$$

Marginal probability, shorthand notation with values of Y omitted:

$$P(X) = \sum_{Y} P(X, Y)$$

Conditionalization

(*) In a finitary setting

Exploring 'what if' something becomes known

Given a joint probability distribution

$$P(X, Y)$$

The *conditional probability* $P(X \mid Y = y)$ is defined as:

$$P(X \mid Y = y) := \frac{P(X, Y = y)}{P(Y = y)}$$

A conditional probability can be a joint probability as well …

$$P(X, Y \mid Z = z) := \frac{P(X, Y, Z = z)}{P(Z = z)}$$

Conditional probability, more general notation:

$$P(X \mid Y) := \frac{P(X, Y)}{P(Y)}$$

Conditionalization

(*) In a finitary setting

Exploring 'what if' something becomes known

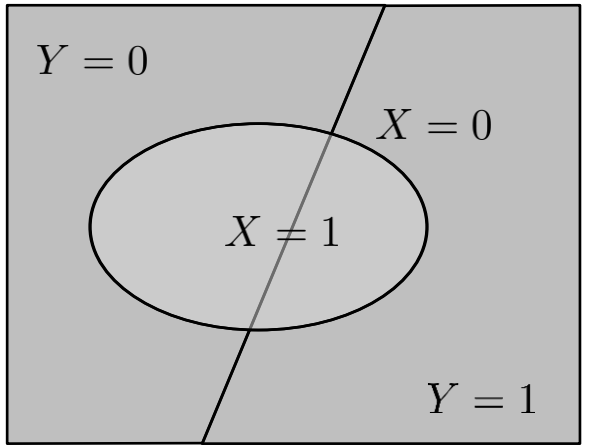
Conditional probability, more general notation:

$$P(X \mid Y) \;:=\; \frac{P(X,Y)}{P(Y)}$$

Assume both variables are binary $X, Y \in \{0,1\}$

$$P(X \mid Y=0) \;:=\; \frac{P(X,Y)}{P(Y=0)}$$

$$P(X \mid Y=1) \;:=\; \frac{P(X,Y)}{P(Y=1)}$$



Each value of the conditioning variable defines a distinct event (sub)space

Remarkable Rules

(*) In a finitary setting

Following from the definition of conditionalization:

$$P(X \mid Y) := \frac{P(X,Y)}{P(Y)}$$

- **Chain Rule**

$$P(X,Y) = P(X \mid Y)P(Y)$$
$$P(X,Y) = P(Y \mid X)P(X)$$

- **Bayes' Theorem**

$$P(X \mid Y)P(Y) = P(Y \mid X)P(X)$$

$$P(X \mid Y) = \frac{P(Y \mid X)P(X)}{P(Y)}$$

$$P(X \mid Y) = \frac{P(Y \mid X)P(X)}{\sum_X P(Y \mid X)P(X)}$$

Probabilistic Inference (no *learning*)

- **General Structure**

  Starting from a fully-specified joint probability distribution

  $$P(X_1, X_2, \ldots, X_n)$$

  In an *inference* problem, random variables $\{X_1, X_2, \ldots, X_n\}$ are divided into three categories:

  1) *Observed variables* $\{X_o\}$, having a definite value
  2) *Irrelevant variables* $\{X_i\}$, which are not directly part of the answer
  3) *Relevant variables* $\{X_r\}$, which are part of the answer we seek

The general solution is:

$$P(\{X_r\}|\{X_o\}) = \sum_{\{X_i\}} P(\{X_r\}, \{X_i\}|\{X_o\})$$

- Computability is always guaranteed (*in a finitary setting)
- Computational efficiency can be a problem

  The number of combinations grows exponentially with the number of random variables

Independence, conditional independence

- **Independence** (also *marginal independence*)

  Two variables are independent
  iff their joint probability can be factorized into the product of *marginals*

$$< X \perp Y > \quad \Leftrightarrow \quad P(X,Y) = P(X)P(Y) \quad \Leftrightarrow \quad < Y \perp X >$$

$$\Rightarrow \quad P(X|Y) = \frac{P(X,Y)}{P(Y)} = \frac{P(X)P(Y)}{P(Y)} = P(X)$$

- **Conditional independence**

  Two variables are conditionally independent given a third variable,
  iff their joint conditional probability can be factorized into the product of *conditional marginals*

$$< X \perp Y \mid Z > \quad \Leftrightarrow \quad P(X,Y|Z) = P(X|Z)P(Y|Z) \quad \Leftrightarrow \quad < Y \perp X \mid Z >$$

$$\Rightarrow \quad P(X|Y,Z) = \frac{P(X,Y|Z)}{P(Y|Z)} = \frac{P(X|Z)P(Y|Z)}{P(Y|Z)} = P(X|Z)$$

CAUTION: *the two forms of independence are distinct!*

$$< X \perp Y > \not\Rightarrow < X \perp Y \mid Z > \qquad < X \perp Y \mid Z > \not\Rightarrow < X \perp Y >$$

Say it with graphs

Chain Factorization

- **Univariate factorization of a Joint Probability Distribution**

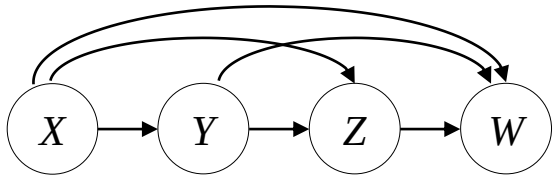
From the definition of conditional probability (*through chain rule*)

$$P(X, Y, Z, W) = P(X)P(Y|X)P(Z|X, Y)P(W|Y, X, Z)$$

Any joint probability distribution can be factorized in a way such that
each factor is *univariate* (i.e., one random variable as independent) conditional distribution.

- Each factorization depends on an arbitrary *sequence* of the *random variables*
- Hence factorizations are not *unique*: any sequence produces a legitimate factorization of the same kind

Graphical equivalent of a *univariate factorization*



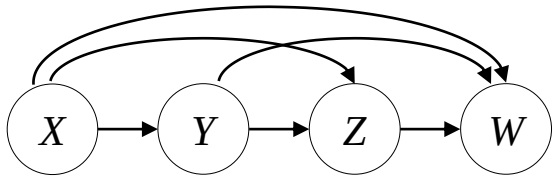
In this oriented graph:

- each node represents a random variable (and the corresponding *univariate* factor)
- each arc represents a conditioning of a random variable over another one (i.e. *dependence*)

Chain Factorization

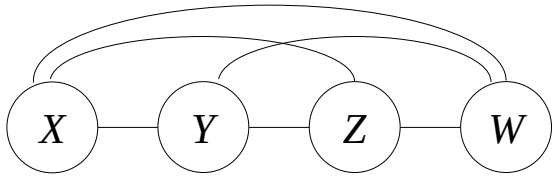
- **Graphical model**

$$P(X,Y,Z,W) = P(X)P(Y|X)P(Z|X,Y)P(W|Y,X,Z)$$



This graph:

- is *acyclic*: if you follow the arrows, you will never return to the same node
- is *completely connected*: if you ignore arc orientations, every node is connected to any other node



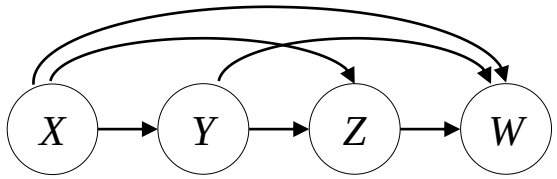
Any *univariate factorization* can be represented by a *graphical model*

Every *completely connected*, *acyclic* and *oriented graph* represents a *univariate factorization*

Chain Factorization and Independence Assumptions

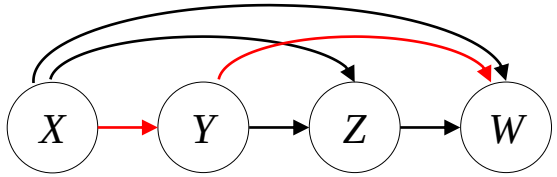
- **Graphical model**

$$P(X, Y, Z, W) = P(X)P(Y|X)P(Z|X, Y)P(W|Y, X, Z)$$



- **Independence**

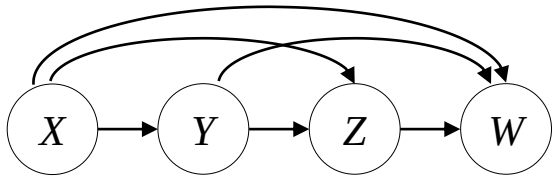
Let's remove a few arcs from the graph and rewrite the factorization accordingly



Chain Factorization and Independence Assumptions

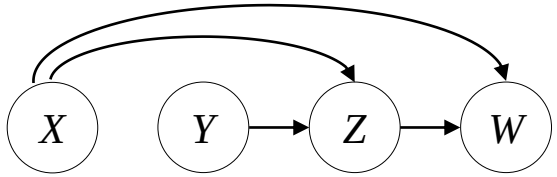
- **Graphical model**

$$P(X,Y,Z,W) = P(X)P(Y|X)P(Z|X,Y)P(W|Y,X,Z)$$



- **Independence**

Let's remove a few arcs from the graph and rewrite the factorization accordingly



$$P(X,Y,Z,W) = P(X)P(Y)P(Z|X,Y)P(W|X,Z)$$

which is true only if

$$P(Y|X) = P(Y) \qquad \langle X \perp Y \rangle$$

$$P(W|X,Y,Z) = P(W|X,Z) \qquad \langle Y \perp W \,|X,Z \rangle$$

Independence

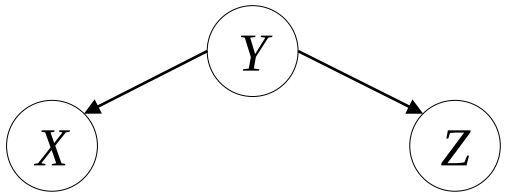
Conditional Independence

Graphical models and independence assumptions

- **Structural Equivalence**

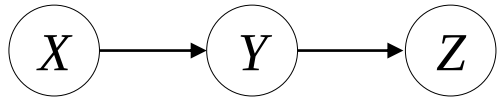
  Different *structures*, different factorizations, same *independence* assumptions:

'fork'



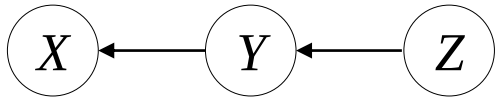
$$P(Y)P(X|Y)P(Z|Y) \Rightarrow \langle X \perp Z|Y \rangle$$

'sequence'



$$P(X)P(Y|X)P(Z|Y) \Rightarrow \langle X \perp Z|Y \rangle$$

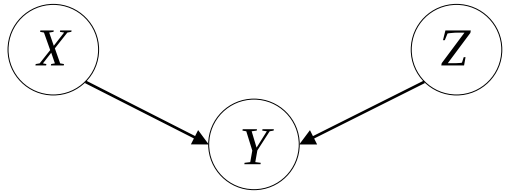
'sequence'



$$P(Z)P(Y|Z)P(X|Y) \Rightarrow \langle X \perp Z|Y \rangle$$

Yet, this other *structure* implies a different independence assumption:

'join' or 'collider'



$$P(X)P(Z)P(Y|X,Z) \Rightarrow \langle X \perp Z \rangle$$

Graphical models and independence assumptions

Equivalence criterion, in general

- **Markov Equivalence Class**

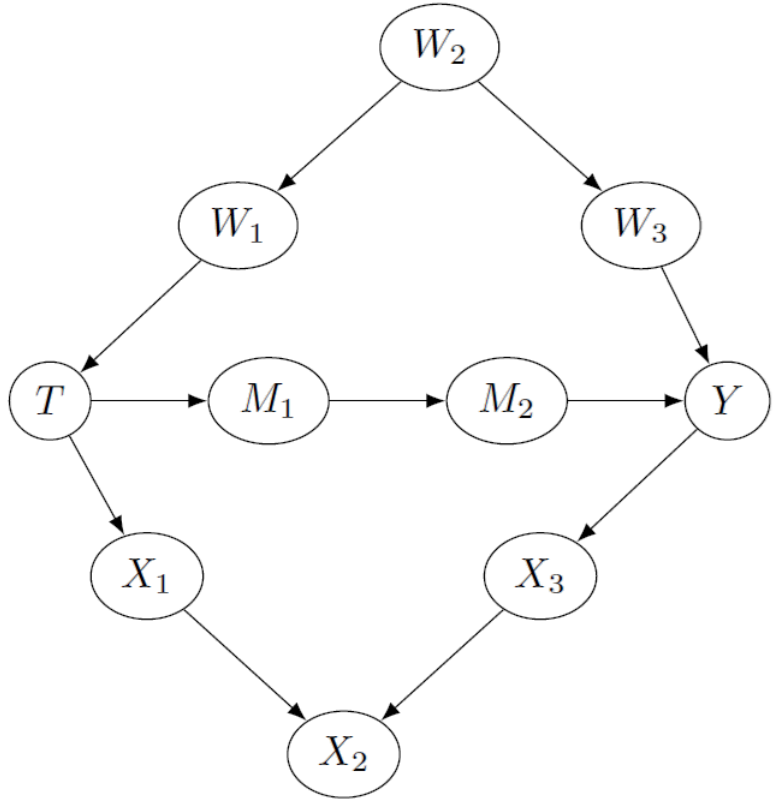
  Two graphical models share the same independence assumptions when:

  1) they share the same *undirected* structure (i.e., *skeleton*)
  2) they share the same *joins* (a.k.a. *colliders*)

  (*) This holds true when some independence is expressed (i.e., if some links are missing).
  Any DAG built out of a clique will be equivalent, regardless of joins
  (i.e., no independence assumptions represented anyway)

Graphical Models as Univariate Factorization

Any graphical model represents a univariate factorization



- **Factorization for a graphical model**

  The general formula is:

$$P(X_1, X_2, \ldots, X_n) = \prod_i P(X_i \mid parents(X_i))$$

  where *parents* are all nodes having a direct, incoming dependence arc

For the example in figure:

$$P(W_1, W_2, W_3, T, M_1, M_2, Y, X_1, X_2, X_3) =$$
$$P(W_2)P(W_1|W_2)P(W_3|W_2)P(T|W1)P(M_1|T)P(M_2|M_1)P(Y|W_3, M_2)P(X_1|T)P(X_3|Y)P(X_3|X_1, X_2)$$

Paths in Graphical Models

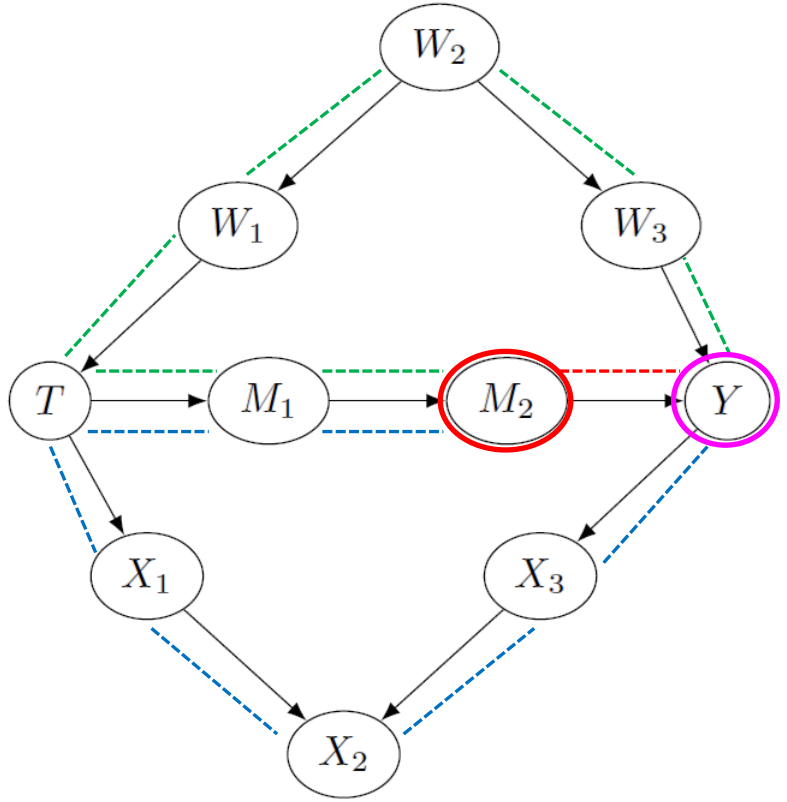
In a graphical model

Consider any two nodes X and Y

A *path* between X and Y
is a path in the graph ignoring orientations (i.e., arrows)

Example:

In the graph on the right,
consider all paths between M_2 and Y



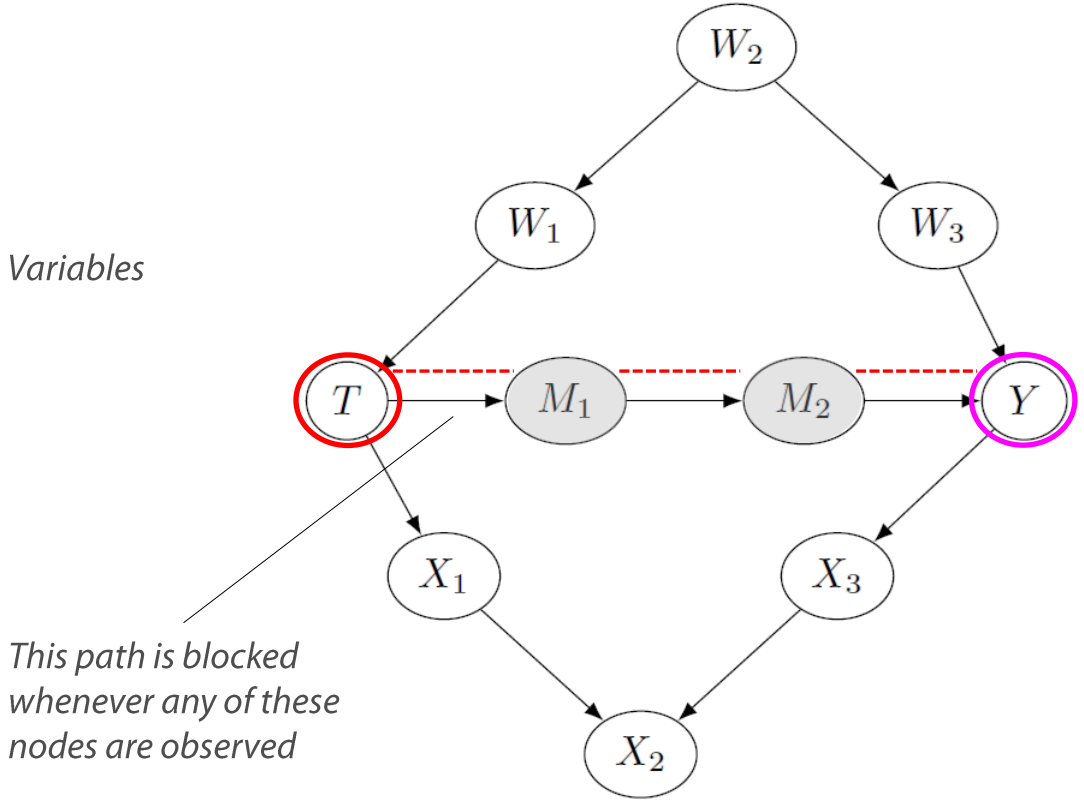
Blocked Paths in Graphical Models

In a graphical model

A *path* between any two nodes X and Y is ***blocked*** whenever the observations $\{Z_o\}$ are such that the path contains either:

Observed Variables

1) a *sequence* or a *fork* for which $\{Z_o\}$ contains the node in between
2) a *collider* for which $\{Z_o\}$ does <u>not</u> contain the observation of the join node nor of any of its descendants



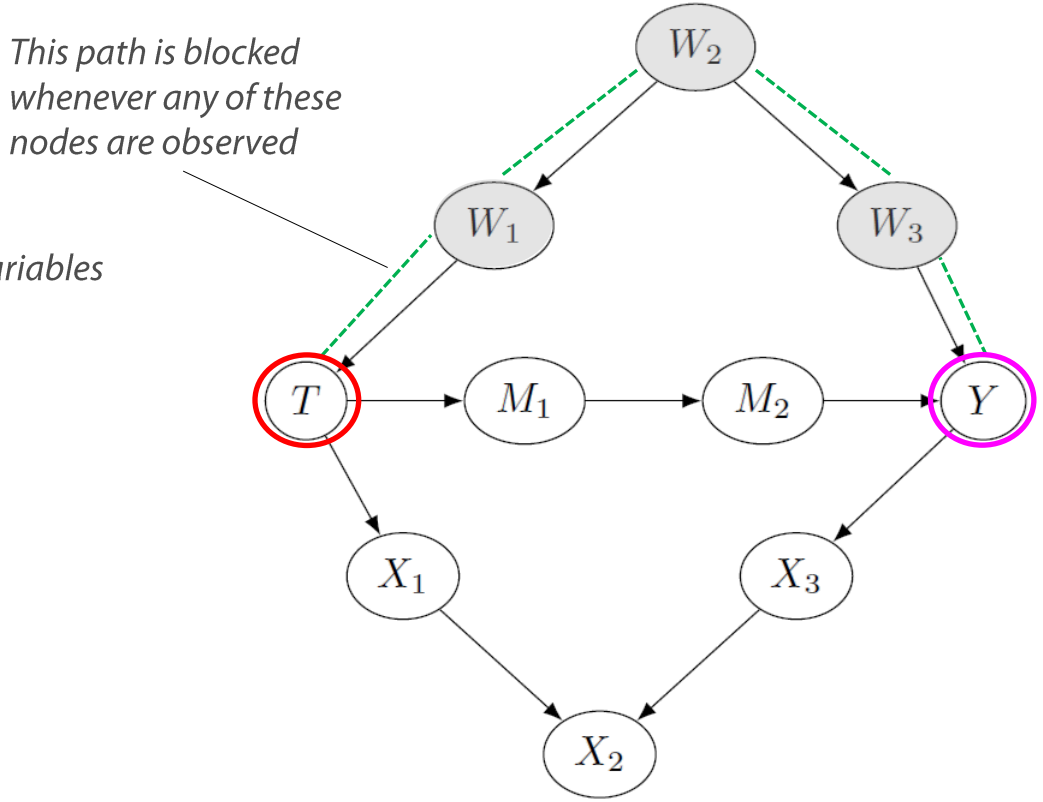
Blocked Paths in Graphical Models

In a graphical model

A *path* between any two nodes X and Y is ***blocked*** whenever the observations $\{Z_o\}$ are such that the path contains either:

Observed Variables

1) a *sequence* or a *fork* for which $\{Z_o\}$ contains the node in between
2) a *collider* for which $\{Z_o\}$ does <u>not</u> contain the observation of the join node nor of any of its descendants

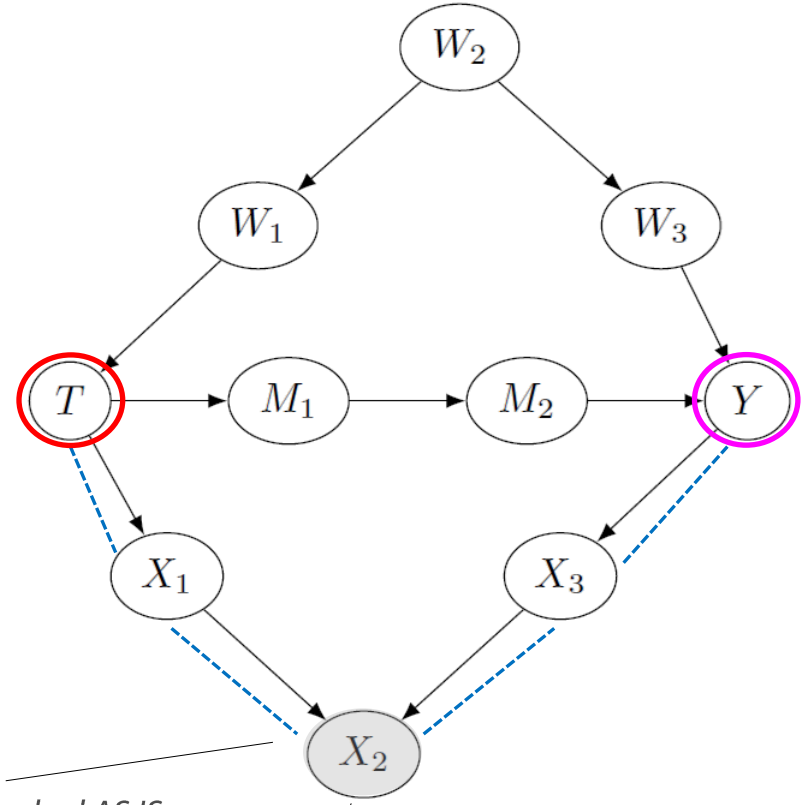


Blocked Paths in Graphical Models

In a graphical model

A *path* between any two nodes X and Y is ***blocked*** whenever the observations $\{Z_o\}$ are such that the path contains either: *Observed Variables*

1) a *sequence* or a *fork* for which $\{Z_o\}$ contains the node in between
2) a *collider* for which $\{Z_o\}$ does <u>not</u> contain the observation of the join node nor of any of its descendants



This path is <u>blocked</u> AS IS: the collider blocks it

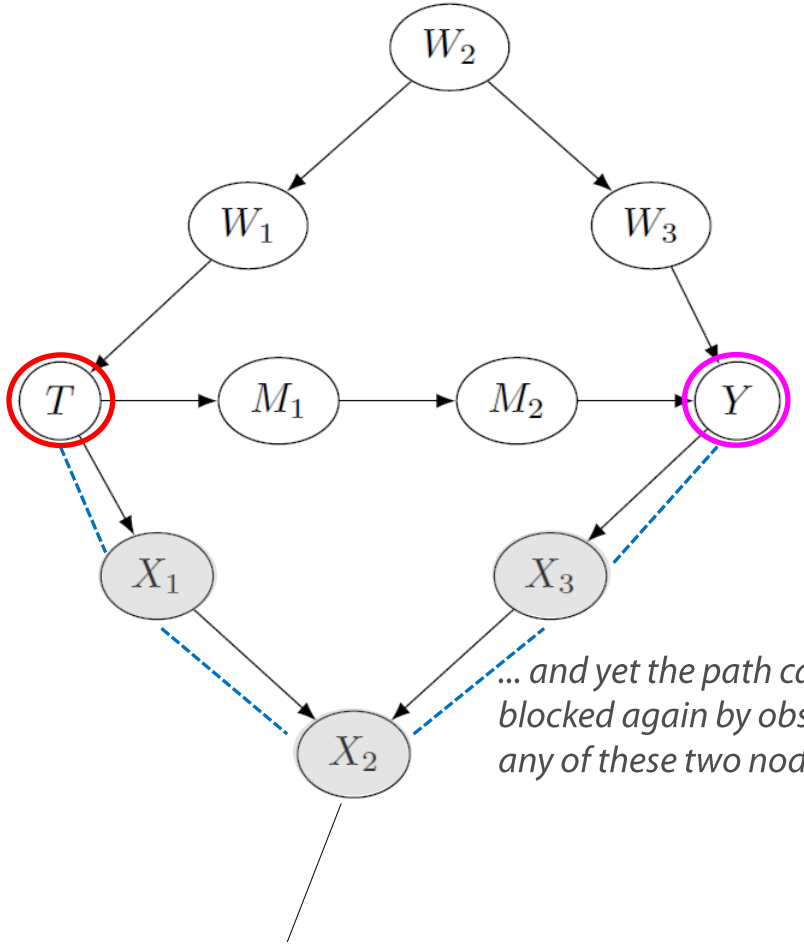
It becomes <u>unblocked</u> when this node is observed...

Blocked Paths in Graphical Models

In a graphical model

A *path* between any two nodes A and B is ***blocked*** whenever the observations $\{X_o\}$ are such that the path contains either: *Observed Variables*

1) a *sequence* or a *fork* for which $\{Z_o\}$ contains the node in between
2) a *collider* for which $\{X_o\}$ does <u>not</u> contain the observation of the join node nor of any of its descendants



D-Separation in Graphical Models

- **Dependency Separation (d-separation)**

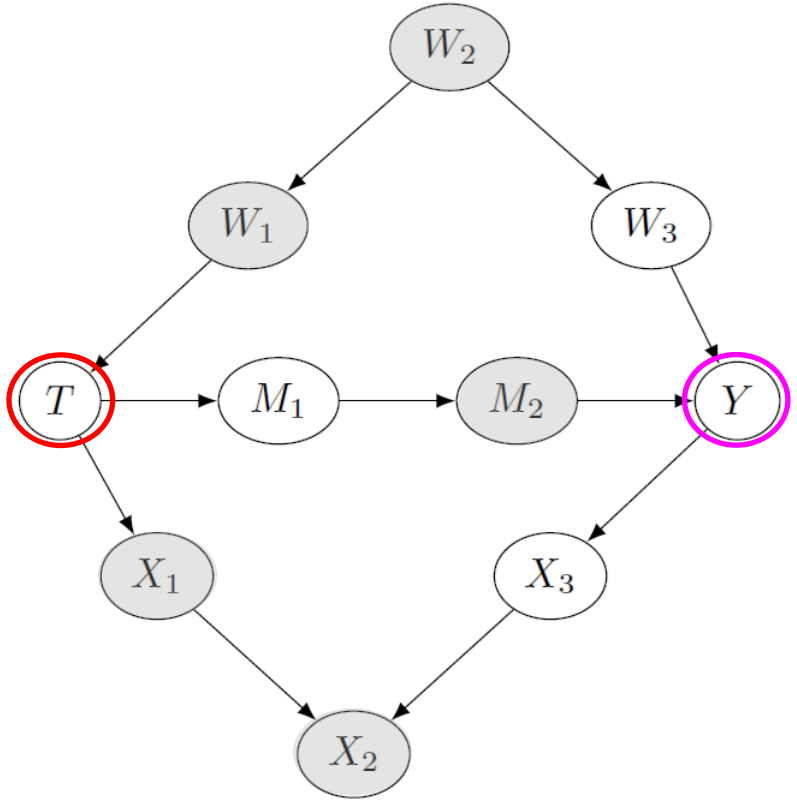
  Any two nodes X and Y in a graphical model are ***d-separated*** whenever the observations $\{Z_o\}$ are such that <u>all paths</u> between X and Y are blocked

  Observed Variables

  In that case we have

  $$\langle X \perp Y \,|\{Z_o\}\rangle$$

  REMEMBER: <u>all</u> paths need to be blocked



These observations make the two nodes <u>d-separated</u>

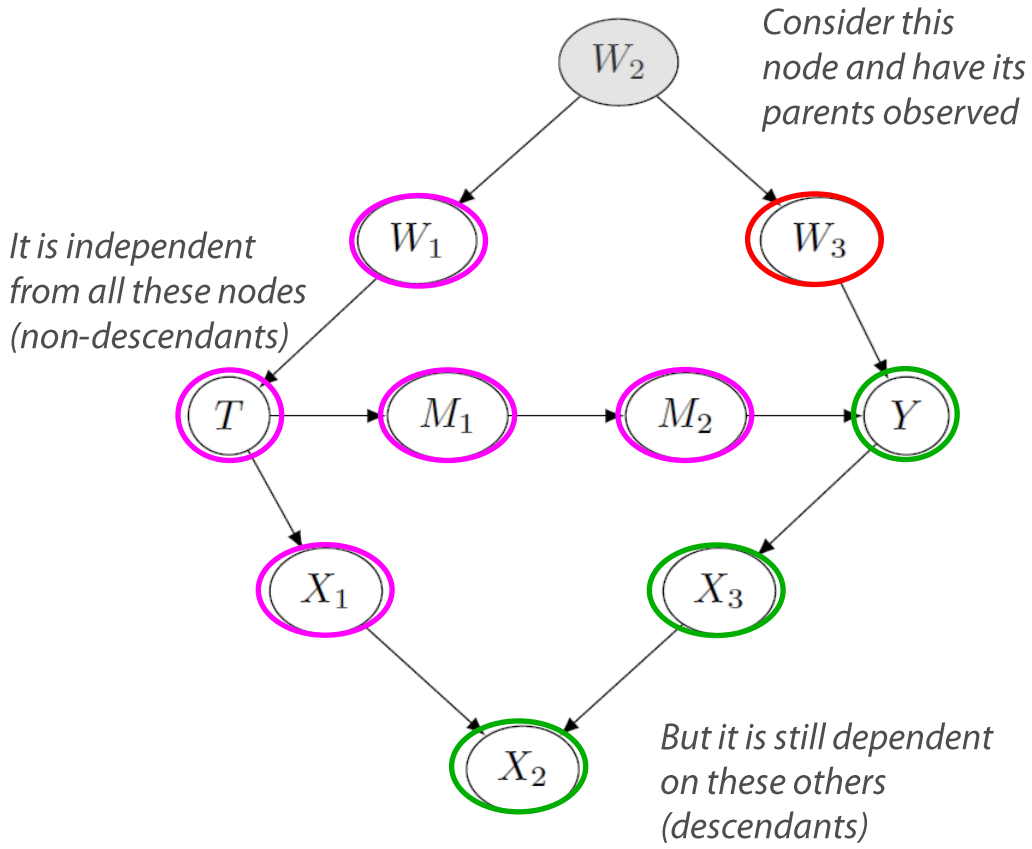
Graphical models: fundamental assumptions

- **Minimality**

  Adjacent nodes in the graph are dependent.

- **Local Markov Assumption**

  Given its parents in the graph, a node X is independent of all its non-descendants



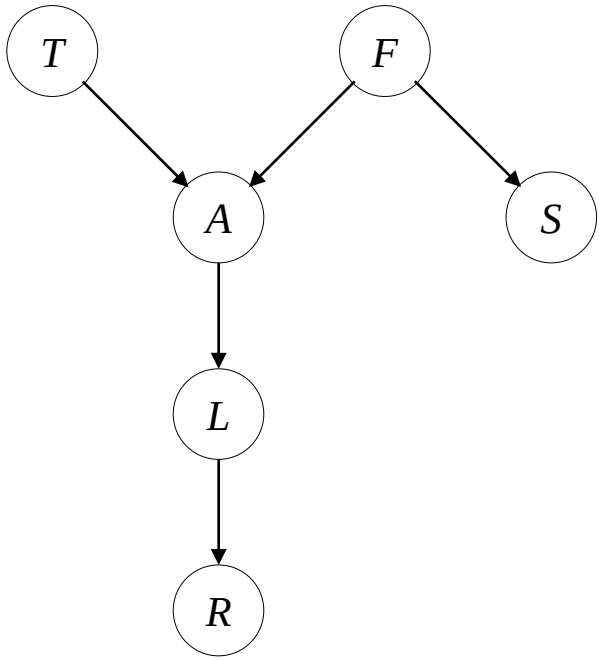
Learning Probabilities from Data

Graphical Models: Learning Parameters

Conditional Probability Factors

Given a graphical model

Example (assume that all random variables are binary):



Joint distribution (general formula)

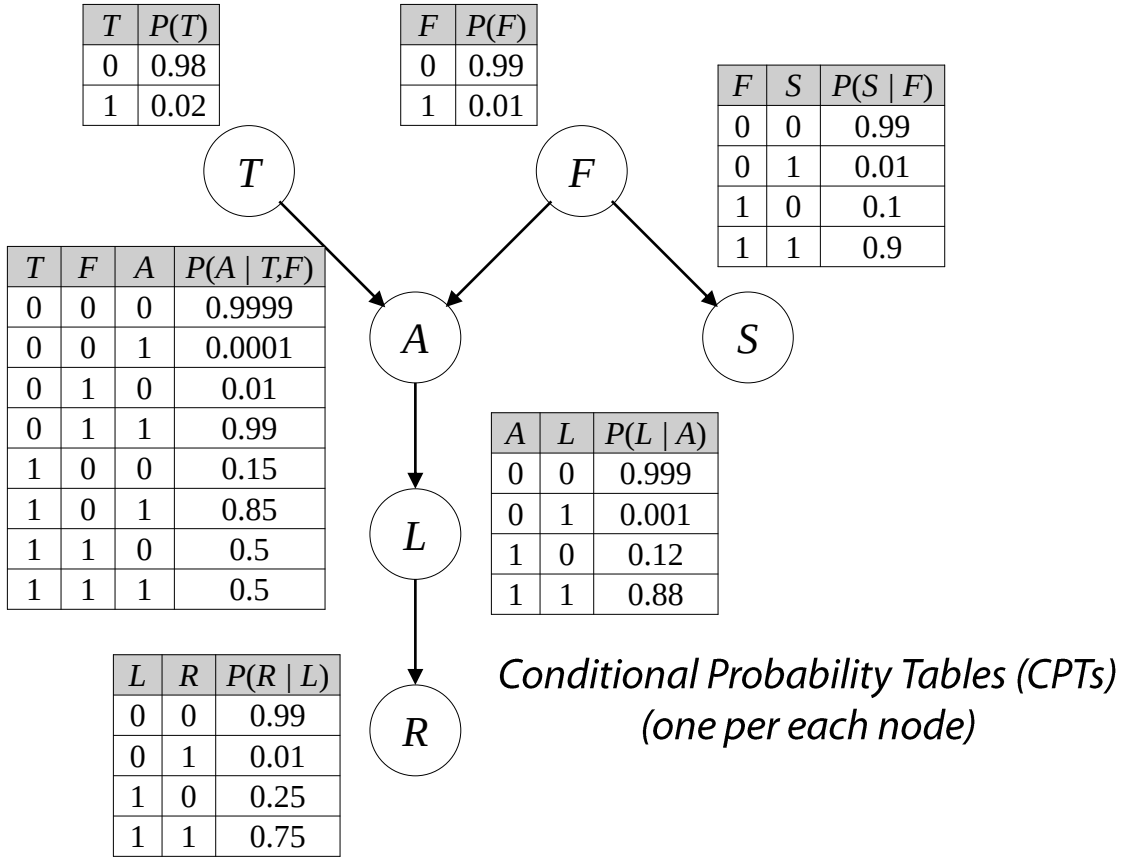
$$P(X_1, X_2, \ldots, X_n) = \prod_i P(X_i \mid parents(X_i))$$

Graphical Models: Learning Parameters

Conditional Probability Factors

Given a graphical model

Example (assume that all random variables are binary):



T	P(T)
0	0.98
1	0.02

F	P(F)
0	0.99
1	0.01

F	S	P(S | F)
0	0	0.99
0	1	0.01
1	0	0.1
1	1	0.9

T	F	A	P(A | T,F)
0	0	0	0.9999
0	0	1	0.0001
0	1	0	0.01
0	1	1	0.99
1	0	0	0.15
1	0	1	0.85
1	1	0	0.5
1	1	1	0.5

A	L	P(L | A)
0	0	0.999
0	1	0.001
1	0	0.12
1	1	0.88

L	R	P(R | L)
0	0	0.99
0	1	0.01
1	0	0.25
1	1	0.75

Conditional Probability Tables (CPTs)
(one per each node)

Joint distribution (general formula)

$$P(X_1, X_2, \ldots, X_n) = \prod_i P(X_i \mid parents(X_i))$$

Each CPT corresponds to a conditional factor

$$P(X_i \mid parents(X_i))$$

How can we learn all these factors *from* data*?*

Observations (Dataset)

For a given set of random variables $\{X_1, X_2, \ldots, X_n\}$

Consider a *dataset*

$$D := \{(x_1^{(i)}, x_2^{(i)}, \ldots, x_n^{(i)})\}_{i=1}^{N}$$

namely, a set of *data items*

$$d^{(i)} := (x_1^{(i)}, x_2^{(i)}, \ldots, x_n^{(i)})$$

- **Independent and Identically Distributed (IID) Dataset**

  1) All data items are independent from each other

  $$\langle d^{(i)} \perp d^{(j)} \rangle, \quad \forall i, j \ : \ i \neq j$$

  2) The generating distribution is the same for all data items

  $$d^{(i)} \sim P(X_1, X_2, \ldots, X_n), \quad \forall i$$

CAUTION: *Being* IID *is not an obvious property of observations*

for example, different measurements on different patients may be IID, but different measurements over time on the same patient are not IID

Complete Observations

For a given set of random variables $\{X_1, X_2, \ldots, X_n\}$

Consider a *dataset*

$$D := \{(x_1^{(i)}, x_2^{(i)}, \ldots, x_n^{(i)})\}_{i=1}^{N}$$

namely, a set of *data items*

$$d^{(i)} := (x_1^{(i)}, x_2^{(i)}, \ldots, x_n^{(i)})$$

- **Complete Observations**

  Are all observations *complete*?

  In other words, are all values $(x_1^{(i)}, x_2^{(i)}, \ldots, x_n^{(i)})$ completely specified or some of them are *missing*?

CAUTION: *Learning probabilities from a dataset with missing values is still possible but more difficult*

Maximum Likelihood Estimator (MLE)

Given dataset:

$$D := \{(x_1^{(i)}, x_2^{(i)}, \ldots, x_n^{(i)})\}_{i=1}^{N}$$

and a probability measure with parameters:

$$P(X_1, X_2, \ldots, X_n;\ \vartheta)$$

- **Likelihood of the dataset**

$$\mathcal{L}(\vartheta \mid D) := P(D \mid \vartheta)$$

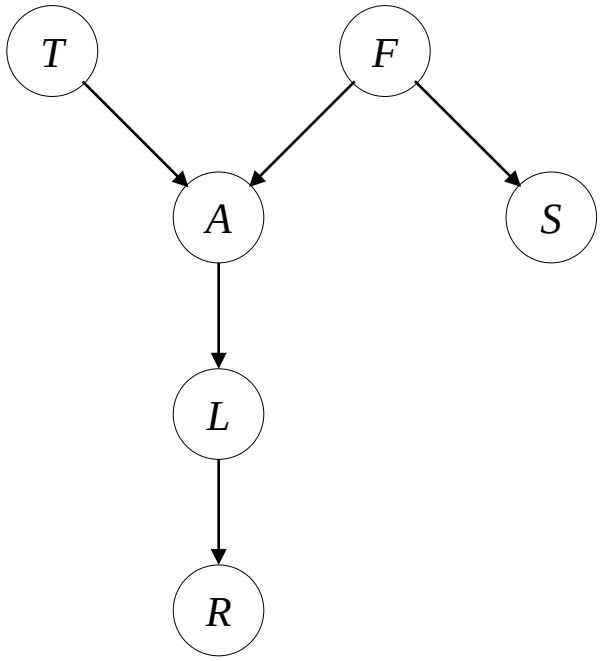
Notice the inversion: likelihood = probability of generating the dataset with specific parameter values

Learning as an optimization problem (*Maximum Likelihood Estimator - MLE*)

$$\vartheta^*_{\mathrm{ML}} := \operatorname*{argmax}_{\vartheta} \mathcal{L}(\vartheta \mid D)$$

Maximum Likelihood Estimator (MLE)

For a graphical model



Joint distribution (probability measure)

$$P(X_1, X_2, \ldots, X_n) = \prod_i P(X_i \mid parents(X_i))$$

Conditional Probability Factors (IID and complete observations)

$$\vartheta := \{P(X_i \mid parents(X_i))\}_{i=1}^{n}$$

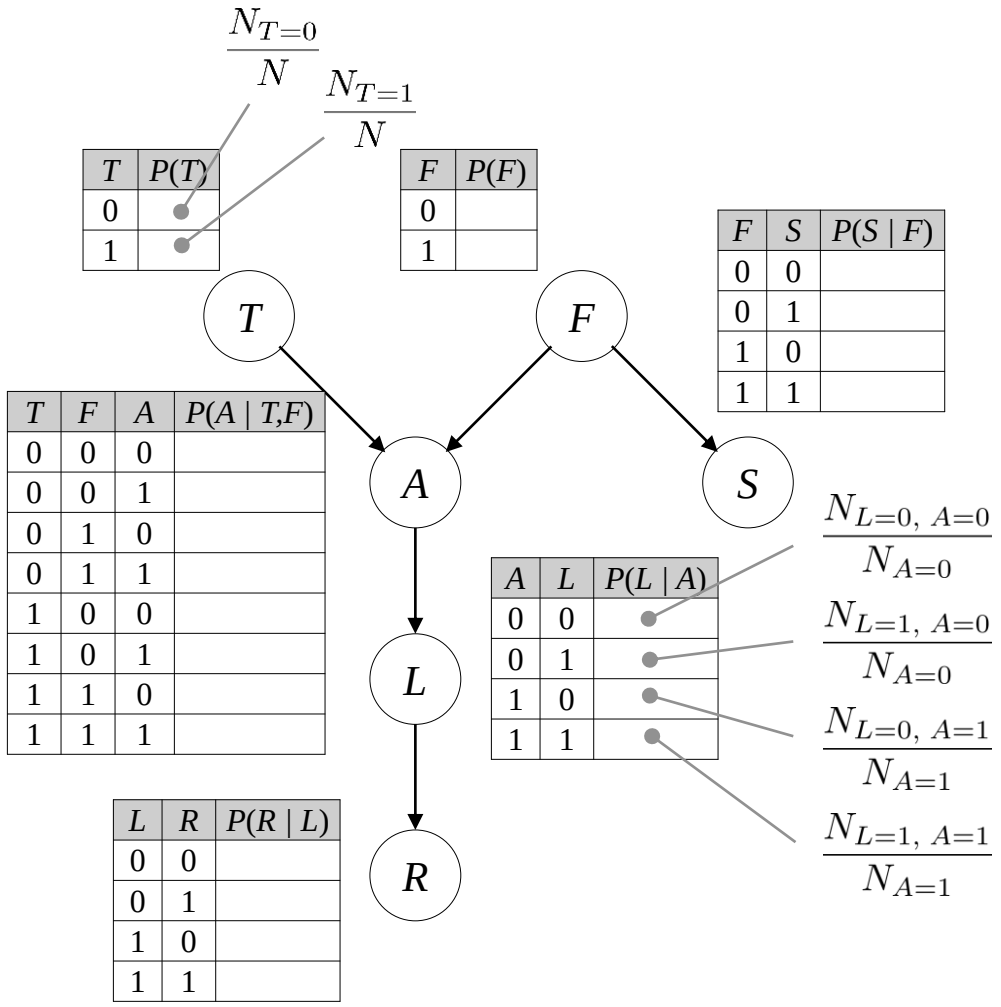
MLE – General Solution (IID and complete observations)

$$P(X_i = x_i \mid parents(X_i) = x) = \frac{N_{X_i = x_i,\ parents(X_i) = x}}{N_{parents(X_i) = x}}$$

In other terms, each probability factor can be computed by just counting a relative frequency of occurrence in the dataset

Maximum Likelihood Estimator (MLE)

For a graphical model



T	P(T)
0	
1	

F	P(F)
0	
1	

F	S	P(S | F)
0	0	
0	1	
1	0	
1	1	

T	F	A	P(A | T,F)
0	0	0	
0	0	1	
0	1	0	
0	1	1	
1	0	0	
1	0	1	
1	1	0	
1	1	1	

A	L	P(L | A)
0	0	
0	1	
1	0	
1	1	

L	R	P(R | L)
0	0	
0	1	
1	0	
1	1	

Joint distribution (probability measure)

$$P(X_1, X_2, \ldots, X_n) = \prod_i P(X_i \mid parents(X_i))$$

Conditional Probability Factors

$$\vartheta := \{P(X_i \mid parents(X_i))\}_{i=1}^{n}$$

MLE – General Solution (IID and complete observations)

$$P(X_i = x_i \mid parents(X_i) = x) = \frac{N_{X_i=x_i,\ parents(X_i)=x}}{N_{parents(X_i)=x}}$$

In other terms, each probability factor can be computed by just counting a relative frequency of occurrence in the dataset

Missing Data

For a given set of random variables $\{X_1, X_2, \ldots, X_n\}$

Consider a *dataset*

$$D := \{(x_1^{(i)}, x_2^{(i)}, \ldots, x_n^{(i)})\}_{i=1}^{N}$$

in which some data items may contain *missing values*

$$d^{(i)} = (x_1^{(i)}, ?, \ldots, x_n^{(i)})$$

Possible Completions

A possible completion of a data item is another data item in which there are no more missing values

Example (with binary values):

$$d^{(i)} = (x_1^{(i)} = 0, x_2^{(i)} = ?, x_3^{(i)} = 0)$$

Possible completions are:

$$\tilde{d}_1^{(i)} = (x_1^{(i)} = 0, x_2^{(i)} = 0, x_3^{(i)} = 0)$$
$$\tilde{d}_2^{(i)} = (x_1^{(i)} = 0, x_2^{(i)} = 1, x_3^{(i)} = 0)$$

For a data item with no missing data
the only possible completion is the data item itself

Expectation-Maximization (EM) Algorithm

Fundamental idea: *using probabilities of possible completions*

In the completely observed case: probability factor are estimated as frequencies of occurrence

$$\frac{N_{X_i=x_i,\ parents(X_i)=x}}{N_{parents(X_i)=x}}$$

In the Expectation-Maximization algorithm, *estimated* frequencies are used:

$$\frac{\tilde{N}_{X_i=x_i,\ parents(X_i)=x}}{\tilde{N}_{parents(X_i)=x}}$$

where:

$$\tilde{N}_D := \sum_{i=1}^{N} \sum_{\tilde{d}^{(i)}} P(\tilde{d}^{(i)} \mid d^{(i)};\ \vartheta)$$

Observed values

All possible completions

In words, any incomplete observations 'splits up' and contributes with the probabilities of possible completions
Note that, when all observations are complete:

$$\tilde{N}_D = N_D$$

Expectation-Maximization (EM) Algorithm

Fundamental idea: *using probabilities of possible completions*

Algorithm:

1) Assign parameters $\vartheta^{(0)}$ at random
2) Compute probabilities for all possible completions $\{P(\tilde{d}^{(i)} \mid d^{(i)} \; ; \; \vartheta^{(t)})\}$
3) Update all probability factors using *estimated frequencies*:

$$\vartheta^{(t+1)} = \left\{ P(X_i = x_i \mid parents(X_i) = x) = \frac{\tilde{N}_{X_i = x_i,\, parents(X_i) = x}}{\tilde{N}_{parents(X_i) = x}} \right\}$$

3) Go back to step 2) until some convergence criterion is met