PROBABILISTIC MODELS – PART 3: BAYESIAN NETWORKS REPRESENTATION & SEMANTICS

Gerhard Widmer

Institute of Computational Perception
Johannes Kepler University
Linz, Austria

gerhard.widmer@jku.at
www.cp.jku.at/people/widmer

October 13, 2025

Presentation partly based on and inspired by [Koller & Friedman, 2009] and [Russell & Norvig, 2021], including the use of some figures from their books and/or lecture slides.

Many thanks to Daphne Koller, Nir Friedman, Stuart Russell, and Peter Norvig for making these available (`pgm.stanford.edu`; `aima.cs.berkeley.edu`).

Goals of this Lecture

- Start with a simple example of a Bayesian Network (BN)
- Explore what kinds of reasoning can be performed on it (intuitively)
- Give a formal definition of the semantics of a BN, in two parts:
  - what (conditional) independencies it encodes
  - how it models a full joint distribution
- Demonstrate reduction in complexity afforded by this representation
- Show how probabilities of events can be calculated from a BN
- Briefly mention some issues related to building a BN
- Discuss how to model continuous (real-valued) domains

Outline

A Simple Story

Consider the following story:[1]

A company wants to hire a college graduate. The company's goal is to hire intelligent persons, but there is no way to test intelligence directly. Instead, the company asks the applicants to supply the results of their SAT test[a], and their grade in some relevant course at the university.

[a] **S**cholastic **A**ssessment **T**est – a standardised test for university admission in the United States.

The company's goal:

- Use these items of evidence (SAT result; course grade) to estimate whether an applicant has a high level of intelligence.

Your task (as experts in Probabilistic Modelling and AI):

- Formalise this scenario in terms of a formal, quantitative model that supports algorithms that can compute these estimates.

[1] from Koller & Friedman, 2009

A Simple Story (Ed. GW)

Consider the following story:

A company wants to hire a college graduate. The company's goal is to hire truly interested and motivated persons, but there is no way to test motivation directly. Instead, the company asks the applicants to supply the results of their SAT test[a], and their grade in some relevant course at the university.

[a] **S**cholastic **A**ssessment **T**est – a standardised test for university admission in the United States.

The company's goal:

- Use these items of evidence (SAT result; course grade) to estimate whether an applicant has a high level of interest and motivation.

Your task (as experts in Probabilistic Modelling and AI):

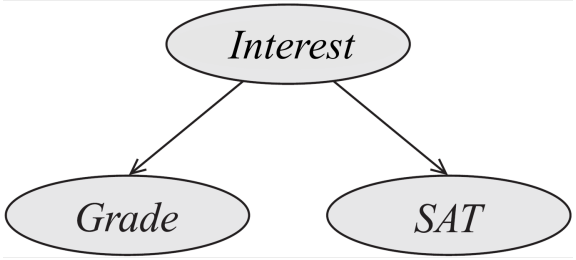
- Formalise this scenario in terms of a formal, quantitative model that supports algorithms that can compute these estimates.

First Steps Towards a Model

1. **Define the concepts (variables) of interest**
   - Three main concepts relevant for decision making:
     student's **motivation level** (which we want to determine/estimate);
     SAT score and **course grade** (which we may have access to).
   - Model these via 3 discrete random variables (with low resolution, for now):

$$\begin{aligned} Interest/Motivation\ (I) \quad &\text{with} \quad Val(I) = \{lo, hi\} = \{i^0, i^1\} \\ SATScore\ (S) \quad &\text{with} \quad Val(S) = \{lo, hi\} = \{s^0, s^1\} \\ Grade\ (G) \quad &\text{with} \quad Val(G) = \{A, B, C\} = \{g^1, g^2, g^3\} \end{aligned}$$

2. **Model the (causal) dependency structure between these variables**



Intuition: The SAT score a person achieves depends (among other things) on the person's motivation; likewise, the course grade is influenced by the interest and motivation level.

A Slightly More Complex Scenario

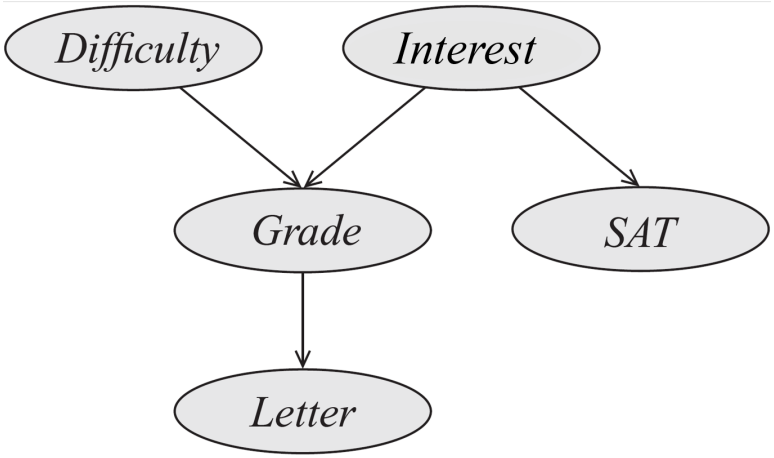
The Story (ctd.):

The grade depends not only on the student's motivation, but also on the **difficulty of the course**.
Student asks professor of the course for a **recommendation letter**; the professor writes that letter based on the course grade only (does not remember particular students because there are too many of these).

$\Rightarrow$ **Two additional variables:**

$$\begin{aligned} \mathit{Difficulty}\ (D) \quad &\text{with} \quad Val(D) = \{easy, hard\} = \{d^0, d^1\} \\ \mathit{Letter}\ (L) \quad &\text{with} \quad Val(L) = \{weak, strong\} = \{l^0, l^1\} \end{aligned}$$

A Model of the Causal Dependency Structure



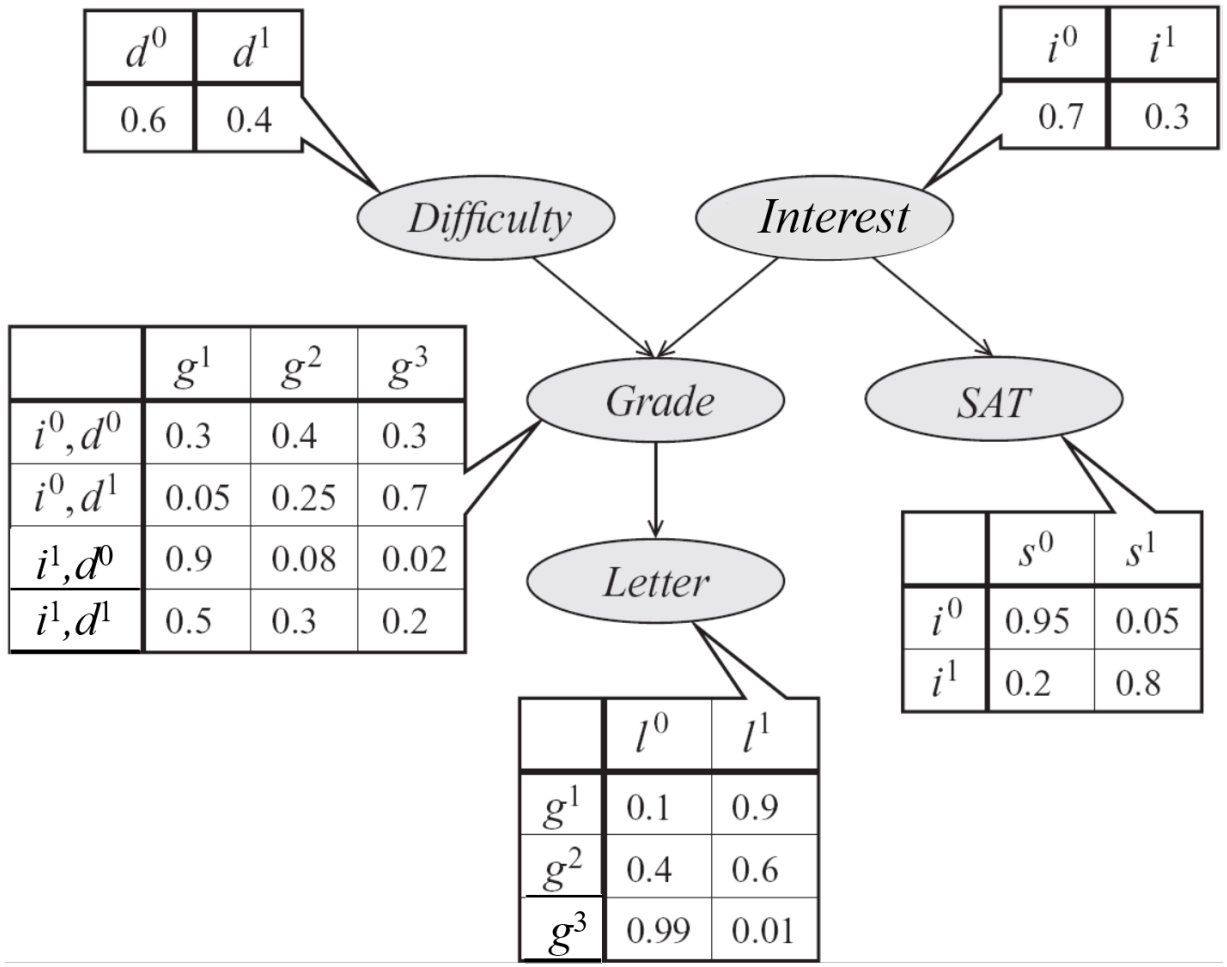
Obvious: These are not *deterministic* relationships:

- Motivation has an influence on SAT score, but is not the only determining factor (and one SAT test may be easier than another)
- Motivation and course difficulty do not uniquely determine the grade a student will achieve
- Professor may be absent-minded and give high recommendation despite low course grade ...

☞ Will try to model these uncertainties via **probabilities**.

A Full Model with Probabilities

Student Network with (Conditional) Probability Distributions (CPDs):



How to Read CPDs

Unconditional probabilities associated with variables without parents:

Example: $P(Interest)$:

i^0	i^1
0.7	**0.3**

- $P(i^1) =$
  the probability that a random student is highly motivated is $= 0.3$ etc. ...

Conditional probabilities for variables that depend on parent variables:

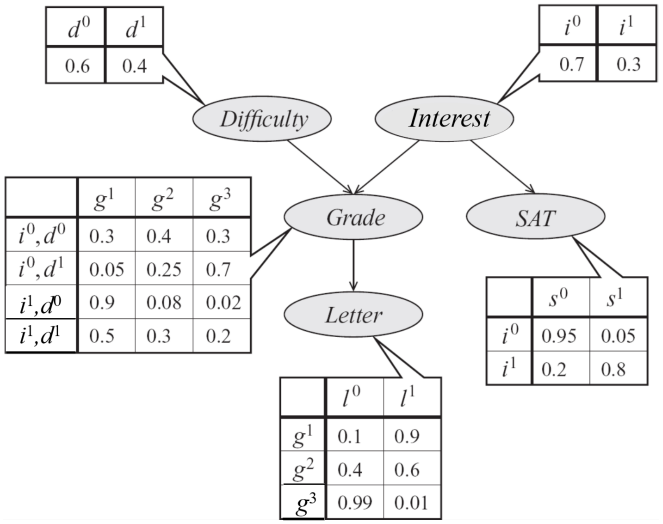
Example: $P(SAT \mid Interest)$:

	s^0	s^1
i^0	0.95	0.05
i^1	**0.2**	0.8

- $P(s^0|i^1) =$
  the probability that a motivated student gets a low SAT score $= 0.2$ etc.

☞ Each row in each of the tables is a **(conditional) probability distribution**.

Preview



In the rest of this lecture, we will show that

- The structure of the model supports different **patterns of reasoning**
- The model reflects all the **independencies** that hold in this story
- The model is a **compact representation of the full joint distribution** over all variables
- It is these independencies that makes the representation so **compact**
- As a representation of the full joint distribution, the model permits us to compute the **answer to any probabilistic query**

The Probabilistic Query

Remember:
We will be interested in answering questions of the following form

Definition

A **Probabilistic Query** involves computing the **Conditional Probability Distribution**

$$P(\boldsymbol{X} \mid \boldsymbol{E} = \boldsymbol{e})$$
$$\text{or} \quad P(\boldsymbol{X} \mid \boldsymbol{e}) \qquad \text{for short}$$

for some sets of variables $\boldsymbol{X}, \boldsymbol{E} \subseteq \mathcal{X}$ and a specific value assignment $\boldsymbol{E} = \boldsymbol{e}$.

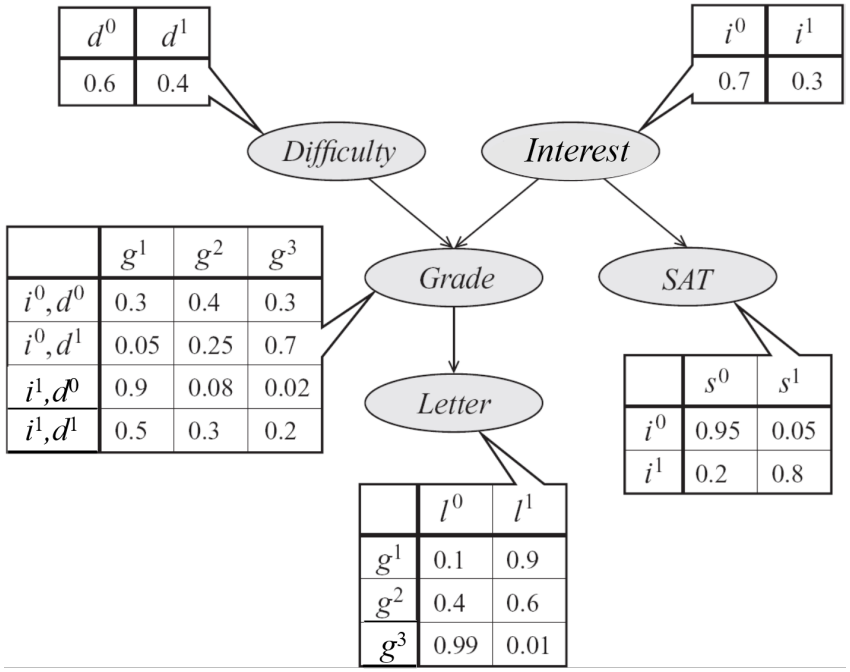
IN WORDS:

Given that we have observed the specific values $\boldsymbol{e}$ of the variables $\boldsymbol{E}$ (the '**evidence variables**'), what are the probabilities for the different value combinations of the $\boldsymbol{X}$ (the '**query variables**')?

General Reasoning Patterns

Some Examples of Typical Reasoning Patterns in the Student Network

- Demonstrate by example different lines of reasoning
- Will only give an intuitive account here
- The next chapter will show how all this (and more) can be automatically calculated from the network model.



Starting Point: An Unconditional Query

Meet a random student named Joe:

Starting Point:
How likely is it that Joe's recommendation letter is strong (l^1)?
Assume we currently know nothing at all about Joe.

☞ **Query:**

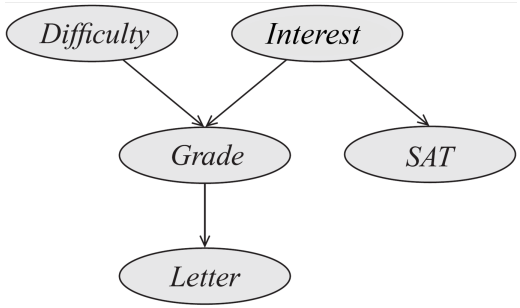
$$P(l^1) = ? \quad \text{(a marginal probability)}$$

☞ **Answer:**

$$P(l^1) \approx 0.502$$

(Will show in next chapter how this is computed from our network model.)

Pattern 1: Causal Reasoning



Assume we

- ... now find out that Joe is not a very interested student (i^0).
  Consequence: Probability of strong recommendation goes down:

$$P(l^1 \mid i^0) \approx 0.389$$

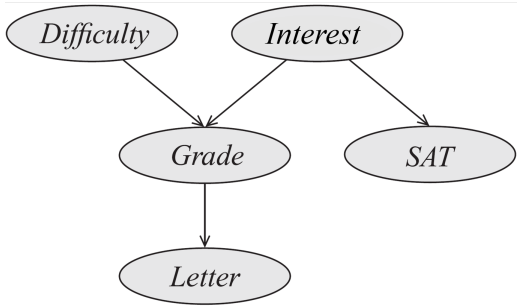
- ... find out in addition that the course "AI" was an easy class (d^0) that year.
  Consequence: Probability of strong recommendation goes up again:

$$P(l^1 \mid i^0, d^0) \approx 0.513$$

(Why? Because an easy class raises the probability that he got a good grade (g^1) despite his i^0, which in turn increases the probability of l^1)

☞ **This kind of reasoning from causes to effects ('downwards' in the network) is called Causal Reasoning or Prediction**

Pattern 2: Evidential Reasoning



Starting point: What is the probability that Joe is highly motivated (i^1)?

$$P(i^1) = 0.3 \quad \text{(directly given by the model)}$$

Now assume we

- ... find out that Joe got a C in class "AI" (g^3).

  Consequence: Probability that Joe is motivated goes down significantly:

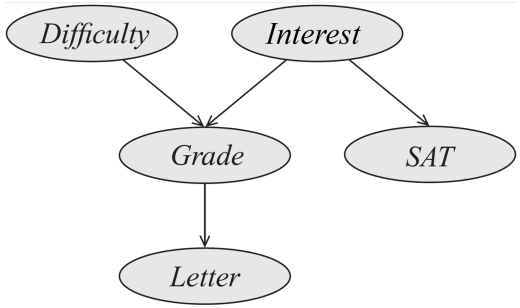
  $$P(i^1 \mid g^3) \approx 0.079$$

  Also: Probability that AI is difficult (d^1) also goes up – from

  $$P(d^1) = 0.4 \text{ to } P(d^1 \mid g^3) \approx 0.63$$

☞ **Reasoning from effects (observations) to possible causes ('upwards' in the network) is called Evidential Reasoning or Explanation**

Pattern 3: Intercausal Reasoning



Assume that in addition to his poor AI grade (g^3), Joe also submits his SAT score, which (surprisingly) is high (s^1) ...

Consequence: Probability that Joe is motivated goes up, from 0.079 to

$$P(i^1 \mid g^3, s^1) \approx 0.578$$

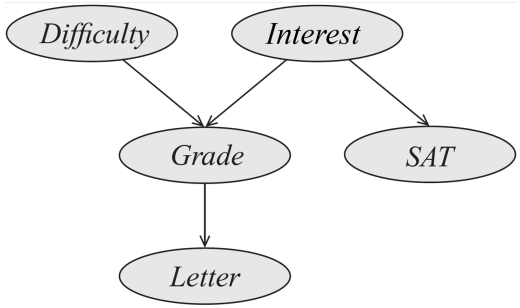
Intuitive Explanation (check CPD tables in model):

- High SAT score outweighs poor grade – students with low motivation are *extremely* unlikely to get a good SAT score: $P(s^1|i^0) = 0.05$
- while students with high motivation can still get C grades

But also:

- Motivated students are much more likely to get C's in hard classes ...
  $\Rightarrow P(d^1 \mid g^3, s^1)$ **goes up to** 0.76 **(compared to** $P(d^1 \mid g^3) = 0.63$**)**

Pattern 3: Intercausal Reasoning



Let's look at this last pattern again:

$$P(d^1 \mid g^3, s^1) > P(d^1 \mid g^3)$$

- SAT gives us information about $Interest/Motivation$
- $Interest$, together with known $Grade$, changes our belief in $Difficulty$

In words: One causal factor for $Grade$ — $Interest$ — gives us information about another causal factor for $Grade$: $Difficulty$!

(Intuitively: A poor grade is a possible indicator of a difficult class, but could also be due to low motivation. But when we learn that the student is (probably) highly motivated, then probably the course must really have been difficult ...)

☞ **This reasoning pattern is called "Explaining Away".**
'Explaining Away' is a special case of Intercausal Reasoning.

Preview: Reasoning with Bayesian Networks

The important (and nice) thing about Bayesian Networks is that they naturally support all these forms of reasoning — via one general, uniform inference algorithm

- All of the above (including “explaining away”) follow naturally from the model
- No specialised algorithms needed for different kinds of reasoning
- A Bayes Net reasoning algorithm can derive *every* consequence that follows from *any combination* of given information (observations)
- No notion of cause and effect needed: dependencies in a network can be modelled in arbitrary directions (as long as you get the numbers right ...)
- Extremely general knowledge representation & inference model

☞ **See next chapter (“Inference”)**

The Semantics of Bayesian Networks (1): Independencies

Remember:

- We will show that BNs are a model of the full joint distribution.
- Last lecture: learned that independencies permit compact representation of joint distribution:

Definition

Let $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z} \subset \mathcal{X}$ be sets of random variables.
$\boldsymbol{X}$ and $\boldsymbol{Y}$ are **conditionally independent** given $\boldsymbol{Z}$, denoted as $(\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z})$, if

- for all values $\boldsymbol{x} \in Val(\boldsymbol{X}), \boldsymbol{y} \in Val(\boldsymbol{Y}), \boldsymbol{z} \in Val(\boldsymbol{Z})$:

$$P(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{z}) = P(\boldsymbol{x} \mid \boldsymbol{z})$$

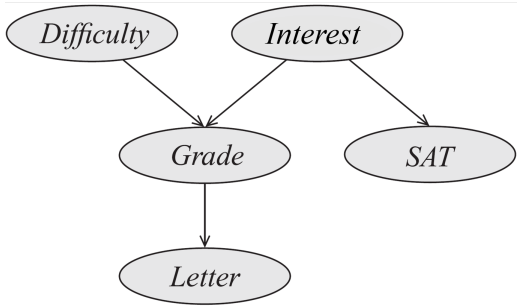
- This is equivalent to saying that

$$P(\boldsymbol{X}, \boldsymbol{Y} \mid \boldsymbol{Z}) = P(\boldsymbol{X}|\boldsymbol{Z})P(\boldsymbol{Y}|\boldsymbol{Z})$$

Consequence:

- Joint distribution over $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$ can be factorised into product of lower-dimensional (conditional) distributions.

Independencies in the Student Model – Intuitive Analysis



Let's start with variable $Letter$:

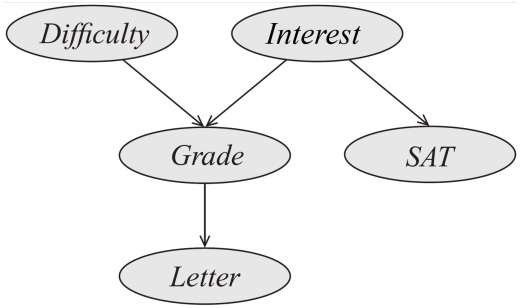
- Assume that probability of student getting a high recommendation depends only on her grade in the class (that's what the model says)
- So: **if we know** the student's grade and the professor's tendencies (the probabilities with which s/he writes a strong letter depending on the grade = the CPDs associated with variable L), then we can predict the probability of a good or bad recommendation

... and learning any of the other aspects (e.g., whether student is motivated, or had a good SAT score) **will not change our belief** regarding the letter.

☞ $Letter$ **is conditionally independent of the other variables, given** $Grade$**:**

$$(L \perp D, I, S \mid G)$$

Independencies in the Student Model



Is $Letter$ *completely* independent of the other variables?

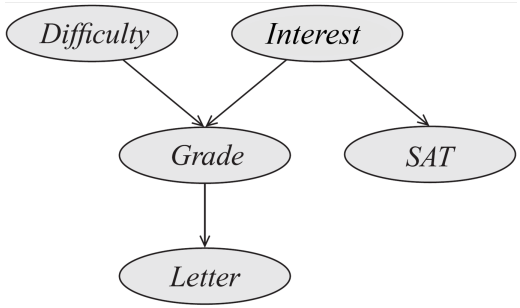
$$(L \perp D, I, S) \ ?$$

- If we *don't* know the grade, then the student's motivation, the difficulty of the course etc. do tell us something about the probability of getting a good recommendation (via $Grade$ and causal reasoning)
- But if $Grade$ has a known fixed value, the other variables don't tell us anything in addition!

☞ **L is conditionally independent of D, I, S, given G, but not marginally independent:**

$$(L \perp D, I, S \mid G) \quad \text{but} \quad (L \not\perp D, I, S)$$

Independencies in the Student Model



Variable SAT:

- In our story, motivation is the only directly determining factor for the obtainable SAT score
- In other words: **if we know** the student's motivation level, we can predict her SAT score (probabilistically)

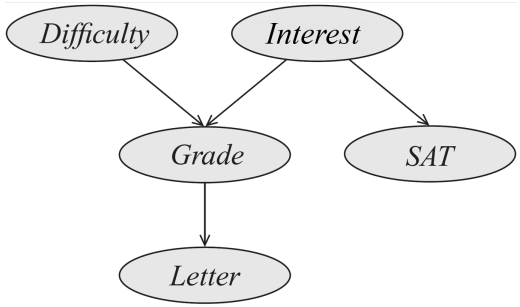
... and learning about any of the other aspects (e.g., the student's grade in the AI course) **will not change our belief** regarding the SAT score.

☞ SAT **is conditionally independent of** D, G, L**, given** $Interest$:

$$(S \perp D, G, L \mid I), \quad \text{but} \quad (S \not\perp G, L)$$

(For a discussion of $(S \perp D)$, see the slide on $Difficulty$ below)

Independencies in the Student Model



Variable $Grade$:

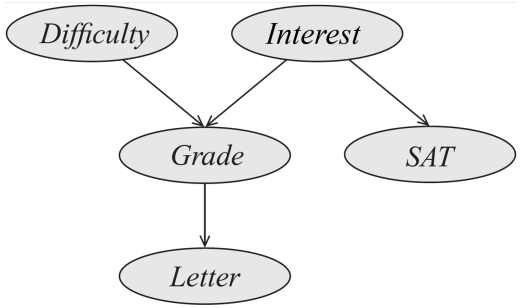
- In our story, the grade depends only on the student's interest / motivation and on course difficulty.

$$\boxed{\Rightarrow (G \perp L, S \mid I, D) \quad ?}$$

NO!

This is false both intuitively (see next slide), and for the specific distribution represented by the model (check with our inference algorithms in the next chapter).

Independencies in the Student Model



G is not independent of L:

- Consider a smart student in a difficult class: i^1, d^1
- Consequence: Probability of an A grade $P(g^1 \mid i^1, d^1)$ is medium (because of class difficulty)
- But if we learn that l^1 (good recommendation), our belief in a good grade goes up:

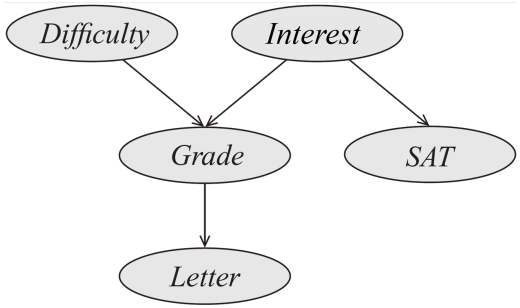
$$\boxed{P(g^1 \mid i^1, d^1, l^1) > P(g^1 \mid i^1, d^1)}$$

(in our model: $P(g^1 \mid i^1, d^1) = 0.5; \quad P(g^1 \mid i^1, d^1, l^1) = 0.712$)

- Knowledge of L helps us to better guess G, **even if we know** I and D

$$(G \not\perp L \mid I, D)$$

Independencies in the Student Model



Is G independent of S ?

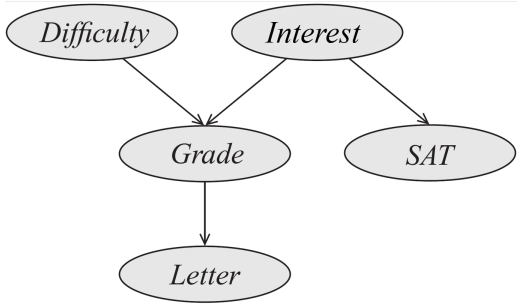
- Intuitively: yes, conditionally.
- If we know a student's motivation and the course difficulty, we have certain expectations regarding the student's grade.
- Learning about his SAT score doesn't change our belief regarding the intelligence level (because we *know* it)...
- ... and thus doesn't change our belief regarding the grade.

☞ **G is conditionally independent of S, given I and D:**

$$(G \perp S \mid I, D)$$

(Again, this will be confirmed when we compute the exact probabilities from the model)

Independencies in the Student Model



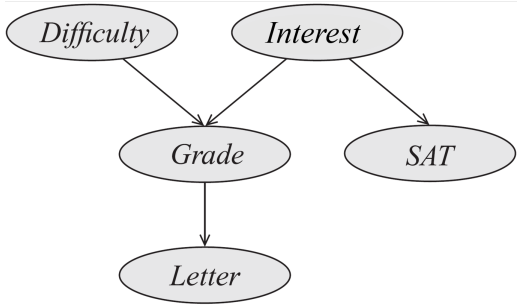
Variables $Difficulty, Interest$:

- Have no parents in the model.
- Each of them can be affected by information about some of their descendants in the model (via *evidential* or *intercausal reasoning*).
- But without information about any of the other variables, I tells us nothing about D, and vice versa.

☞ **D and I (and by the same kind of argument, D and SAT) are independent:**

$$(D \perp I, S)$$

Bayes Nets as a Model of Independencies



Common pattern in the above analysis:

- X's parents "shield" X from causal influences further up in the network (e.g., from its grandparents).
- Given the values of its parents, a variable X is independent from all other variables in the network that are not its children and, more generally, its descendants.
- However, information about X's descendants can change our belief about X (via an *evidential reasoning* process).

Bayesian Network Semantics I: Structure & Independencies

Definition Part I: The Graph Structure

A **Bayesian Network Structure** $\mathcal{G}$ is a **directed acyclic graph** whose nodes represent a set of random variables $\mathcal{X} = \{X_1, \ldots, X_n\}$.
Given structure $\mathcal{G}$, the following **conditional independencies** hold:[a]

$$(X_i \perp \text{NonDesc}(X_i) \mid \text{Pa}(X_i)) \quad \text{for all } X_i \in \mathcal{X}$$

where

- $\text{Pa}(X_i)$ denotes the parents of X_i in $\mathcal{G}$
- $\text{NonDesc}(X_i)$ are the variables that are not descendants of X_i in the graph.

[a] ... and possibly additional ones

IN WORDS:

In a Bayesian Network, **every variable is conditionally independent of all its non-descendants, given its parents**

Bayesian Network Semantics II: BNs as a Factorised Distribution

Question:

- How do the conditional probability distributions (CPDs) in a BN relate to the overall probability distribution P over the variables $\mathcal{X}$?

Answer (sketch of proof see next slide):

- If the network graph correctly represents the conditional independencies in the joint distribution P over $\mathcal{X}$, then this full joint distribution P can be reconstructed from products of the local CPDs.

BNs as a Model of a Joint Distribution

Sketch of Proof

- Consider full joint distribution $P(X_1, \ldots, X_n)$ over $\mathcal{X}$
- Assume that $X_1, \ldots, X_n$ is a **topological ordering** of $\mathcal{X}$ relative to $\mathcal{G}$[2]
- Apply chain rule to decompose P into a product of conditional factors:

$$\begin{aligned} P(X_1, \ldots, X_n) &= P(X_1)P(X_2|X_1)P(X_3|X_1, X_2)\cdots \\ &= \prod_{i=1}^{n} P(X_i \mid X_1, \ldots, X_{i-1}) \end{aligned}$$

- Because of the topological ordering, the set $\{X_1, \ldots, X_{i-1}\}$ contains all parents of X_i, possibly some of X_i's other non-descendants, but none of its descendants
- Thus, by the central property of BNs $(X_i \perp \mathrm{NonDesc}(X_i) \mid \mathrm{Pa}(X_i))$, each factor simplifies to

$$P(X_i \mid X_1, \ldots, X_{i-1}) = P(X_i \mid \mathrm{Pa}(X_i))$$

[2] In words: Assume that $\mathcal{X}$ is sorted 'from top to bottom': for each variable X_i, all of its parents $\mathrm{Pa}(X_i)$ appear before X_i in the list. It is always possible to find such an ordering in an acyclic graph.

BNs as a Model of a Joint Distribution

Consequence:

THE CHAIN RULE FOR BAYESIAN NETWORKS

$$
\begin{aligned}
P(X_1, \ldots, X_n) &= \prod_{i=1}^{n} P(X_i \mid X_1, \ldots, X_{i-1}) \\
&= \prod_{i=1}^{n} P(X_i \mid \mathrm{Pa}(X_i))
\end{aligned}
$$

IN WORDS:

- The CPDs $P(X_i \mid \mathrm{Pa}(X_i))$ jointly define a full probability distribution over the variable space $\mathcal{X}$
- This full joint distribution can be represented as a product of the CPDs
- Any individual entry of the full joint distribution can be calculated as a product over entries from the local CPDs.

Bayesian Network Semantics II: CPDs & the Distribution

Definition Part II: The Factorised Probability Distribution

A **Bayesian Network** $\mathcal{B}$ is a BN structure graph $\mathcal{G}$ where each node (variable) X_i is associated with a set of **conditional probability distributions (CPDs)** $P(X_i \mid \mathrm{Pa}(X_i))$.

The network represents a **full joint distribution** P over $\mathcal{X}$ as follows:

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid \mathrm{Pa}(X_i))$$

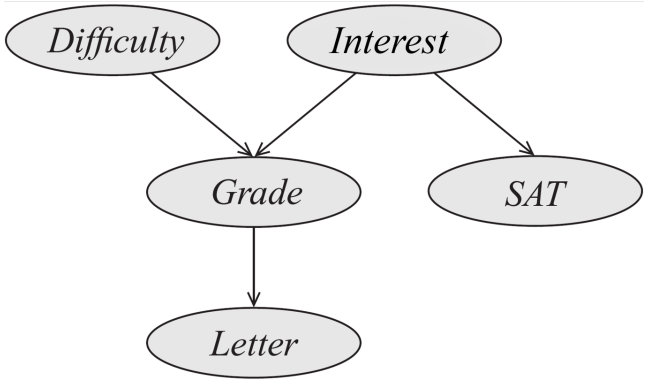
NOTES:

- We say that the joint distribution P **factorises over network structure** $\mathcal{G}$
- A BN is a **factorised representation** of the joint distribution over $\mathcal{X}$
- Each CPD table $P(X_i \mid \mathrm{Pa}(X_i))$ is called a **factor**.

Reconstructing the Joint Distribution: An Example

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid \mathrm{Pa}(X_i))$$

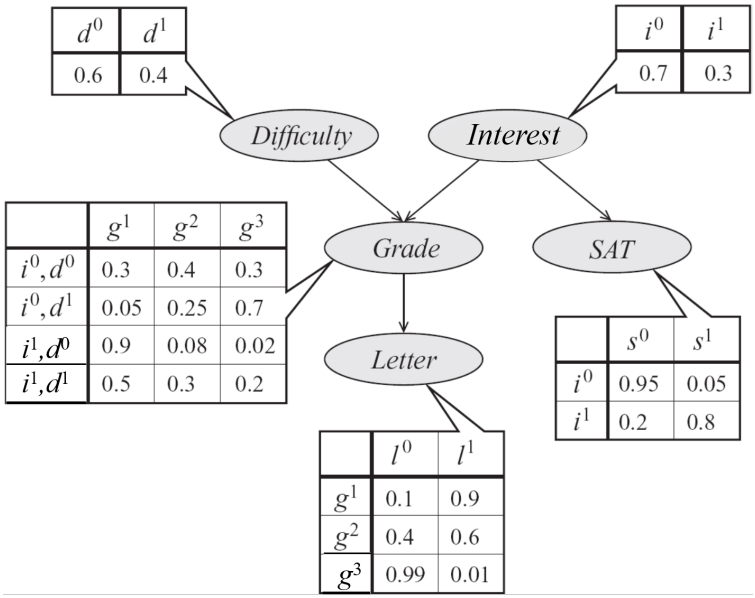
Thus, for our student network:



$$P(D, I, G, S, L) = P(D) \cdot P(I) \cdot P(G|D, I) \cdot P(S|I) \cdot P(L|G)$$

☞ **Each individual entry of the full joint distribution can be computed as a product over entries from the local CPDs.**

Reconstructing the Joint Distribution: An Example



$$\begin{aligned} \underline{P(d^0, i^0, g^2, s^0, l^1)} &= P(d^0) \cdot P(i^0) \cdot P(g^2|d^0, i^0) \cdot P(s^0|i^0) \cdot P(l^1|g^2) \\ &= 0.6 \times 0.7 \times 0.4 \times 0.95 \times 0.6 \\ &= \underline{0.09576} \end{aligned}$$

In words: *The probability of a random student being less motivated, the AI course being easy, the student achieving a B grade in this course and a low SAT score, and still receiving a good recommendation letter from the professor is* 9.58%.

Compactness of the Representation

Remember:

Complexity of the Full Joint Distribution

Consider a set $\mathcal{X} = \{X_1, \ldots, X_N\}$ of N boolean random variables.
The **Full Joint Distribution** $P(X_1, \ldots, X_N)$ over $\mathcal{X}$

- has 2^N **entries** (for the 2^N different atomic events over $\mathcal{X}$)
- Specifying these requires $2^N - 1$ **independent parameters**[a]

[a] The last parameter is redundant because the distribution must sum to 1.0

☞ **Representing the Full Joint explicitly (e.g., in the form of a table) is infeasible for reasonably large worlds $\mathcal{X}$**

Compactness of the Representation

Now consider a Bayesian Network with 'sparse connectivity':

Complexity of a Sparsely Connected Bayesian Network

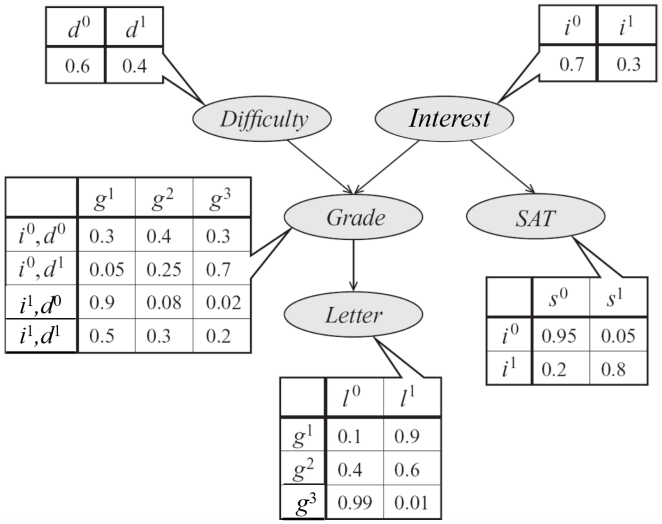
Consider a Bayesian Network $\mathcal{B}$ over N boolean variables $\mathcal{X} = \{X_1, \ldots, X_N\}$. Assume that the network is **sparsely connected** in that each variable has **at most** $k \ll N$ **parents**, where k is a **fixed constant**.

- Each row in the CPD table of a variable X_i requires 1 number (the probability p_i^0 for $X_i = x_i^0$; the probability for x_i^1 is simply $1 - p_i^0$)
- If X_i has k parents, the CPD of X_i has 2^k rows (one for each possible value combination of the parents) $\Rightarrow$ the CPD requires 2^k numbers
- If each of the N variables in the model has at most k parents, all the CPDs together require $\leq N \times 2^k$ **numbers**.

Consequence:

- For a fixed k, $2^k N = O(N)$ is **linear** in N!
- **Exponential reduction in complexity** from $O(2^N)$ to $O(N)$ (if $k \ll N$)!

Example: The Student Network



- 5 variables, $2 \times 2 \times 3 \times 2 \times 2 = 48$ atomic events
- Explicit representation of the full joint distribution requires $48 - 1 = 47$ independent parameters
- Representation of the joint distribution via the BN requires only $1 + 1 + 8 + 3 + 2 = 15$ independent parameters

An Extreme Example: Independent Variables

Consider tossing 10 independent coins:

- 2^{10} possible atomic events (combinations of heads and tails)
- Explicit representation of the full joint distribution requires $2^{10} - 1 = 1023$ independent parameters
- Modelling the joint distribution as a Bayes Net: 10 independent variables $\Rightarrow$ only $10 \times 1 = 10$ independent parameters required!

A Realistic Example

Consider a model for the diagnosis of a complex technical system:

- Might have 50 variables
  (modelling the state (ok/defective) of system components)

  (see ICU model below)
- of which each (directly) depends on ≤ 5 others.
- Explicit representation of the full joint distribution requires

$$2^{50} - 1 \approx 10^{15}$$

  independent parameters.
- The factorised BN representation requires only

$$\leq 50 \times 2^5 = 1,600$$

  independent parameters ...

Constructing BNs: Some General Considerations

To construct a Bayesian Network, we need to do 3 things:

1. Decide on the **random variables** to be used, and their **values (domains)**
2. Specify the **structure** of the network
3. Specify the **conditional probability distributions (CPDs)** for each node.

Step 1 must always be done by the system designer.

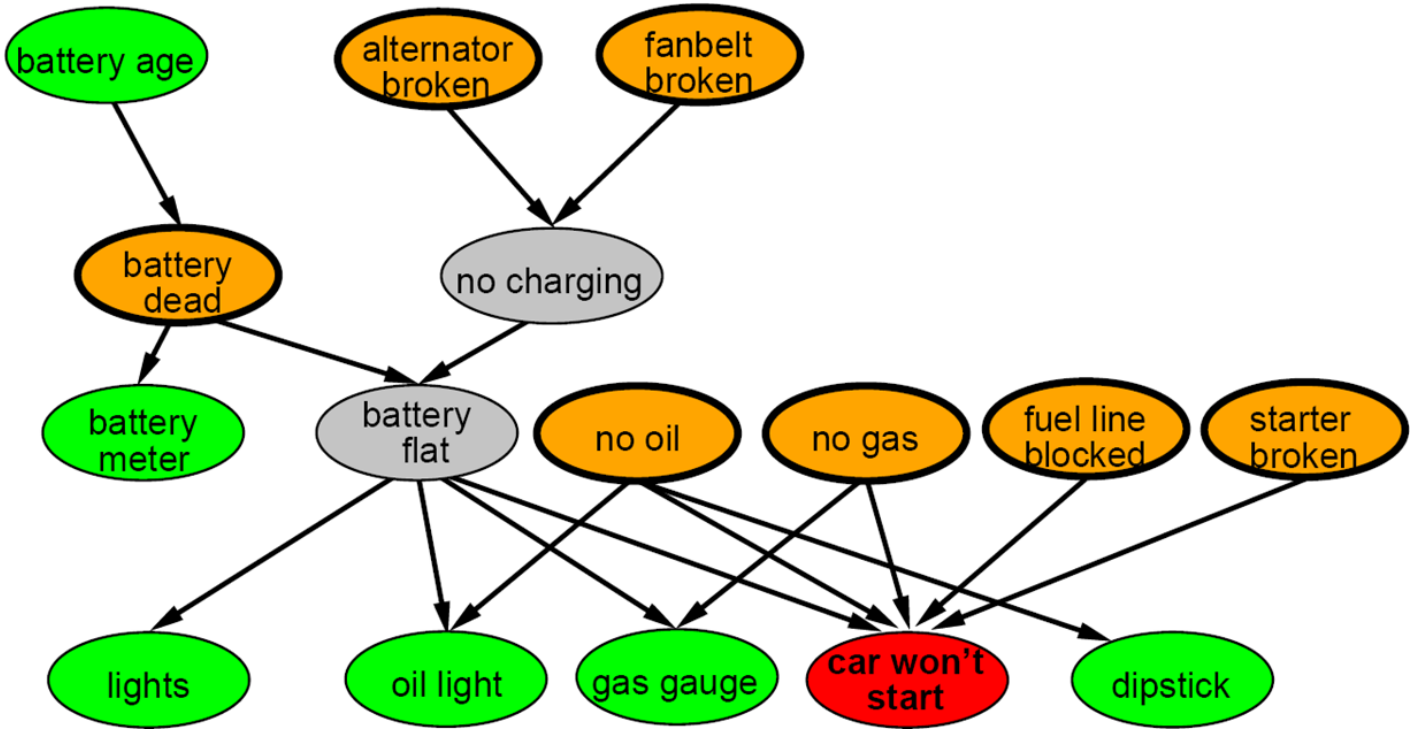
Step 2 can be done manually, in cooperation between system designer and a domain expert (☞ *'Knowledge Engineering'*), or automatically by the system, via learning from example situations (☞ *'Machine Learning'*)

Step 3 is usually done via machine learning (too complex to specify manually)

Important: Try to model in direction of causality!
(generally leads to simpler, sparser, more intuitive networks)[3]

[3] **Exercise:** Try to model our student world in an 'anti-causal' direction – e.g., such that L, G, S, D, I would be a topological ordering of the resulting network – and verify that you will need to introduce additional edges ...

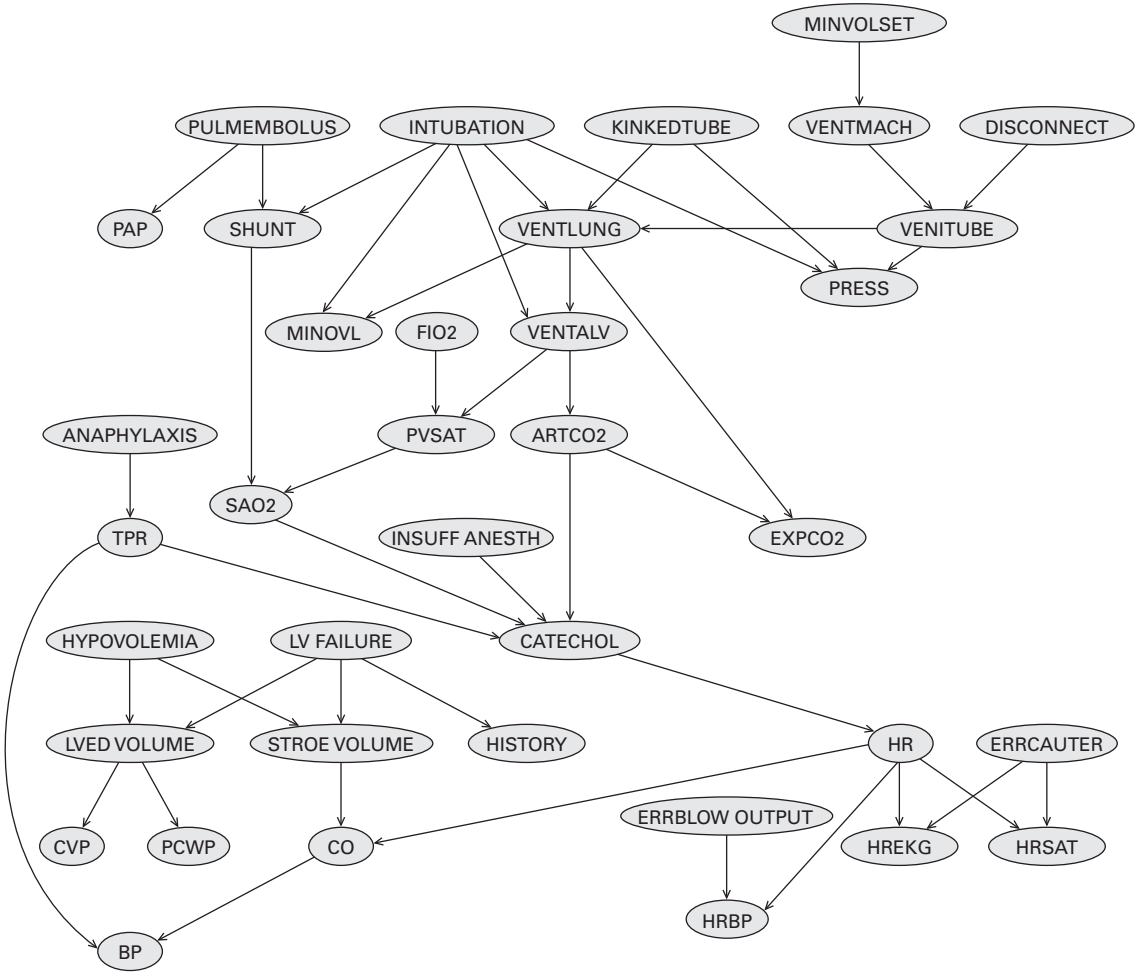
An Example: A Simplified Model for Car Diagnosis[4]



red:	problem to be explained (observable variable)
green:	'tests' – information that can be obtained (potentially observable)
orange:	'diagnoses' – possible explanations of the problem (not directly observable)
grey:	auxiliary concepts to make for a more structured, compact model

[4] from lecture slides by Stuart J. Russell

A Model for Monitoring an Intensive Care Unit (ICU) [5]



[5] originally introduced by (Beinlich et al., 1989)

Preview:[6] Continuous Systems and Variables

So far:

- Assumed that all the variables in a model are *discrete* with a finite domain $Val(X)$
- Permitted us to represent conditional probability distributions $P(X \mid \mathrm{Pa}(X))$ in the form of a table.

Problem:

- Many systems and processes involve *continuous* aspects and measurements that are best modelled by *real-valued* variables
- Example: Position and motion of an airplane or a robot (and our measurements of their position and motion).

[6] More on this later, when we look at Kalman Filters

Continuous Systems and Variables

SOLUTION 1: Discretisation

- Split the domain of a variable into a fixed number of intervals and represent each continuous value by the label of its interval.

Advantage:

- All of our representation and reasoning machinery is directly applicable

Problems:

- Problem 1: **Loss of information** caused by this approximation
- Problem 2: **Complexity**.
- Example: For reasonably accurate robot tracking, might need to discretise position variables into $1,000$ intervals.
  $\Rightarrow 1,000,000$ different (x, y) positions.
  $\Rightarrow$ Imagine the resulting CPDs ...

Continuous Systems and Variables

SOLUTION 2: Work with continuous variables and CPDs

- Permit continuous random variables X with $Val(X) \subseteq \mathbb{R}$
- Model CPDs as *continuous density functions* $p(X \mid \mathrm{Pa}(X))$.

Main Questions:

- How to model the distribution of a continuous variable X?
- How to define *conditional* distributions $p(X \mid \mathrm{pa}(X))$, when the parents of X are also continuous variables?

Example: $X \longrightarrow Y$, with both X and Y continuous

- CPD $P(Y|X)$ contains a set of conditional distributions $p(Y|x)$ over Y, *one for every possible value* $x \in Val(X)$
- Must define an **infinite number of conditional distributions** $p(Y|x)$!

Linear Gaussian Networks

One Possible Solution:
Parameterisable density function, functional dependence of parameters

- Use a parameterisable family of density functions to model distributions
- Assume that the parameters of a conditional distribution $p(X|Y)$ are a **function** of the specific value y of the parent Y!
- Consequence: Need not explicitly represent (infinite number of) conditional distributions $p(X|y)$

Special Case: Linear Gaussian Networks

- Model all (prior and conditional) distributions over single continuous variables X_i as Gaussians $\mathcal{N}(\mu_i; \sigma_i^2)$
- Assume that the mean of Y in $p(Y|X)$ is a **linear function of** X, and that the variance σ^2 of Y **is independent of** X (i.e., fixed)

Example:

$$p(Y \mid x) = \mathcal{N}(3.5x - 0.9; 1.0)$$

Linear Gaussian Networks

Definition

A **Linear Gaussian Bayesian Network** is a BN all of whose variables are continuous, and where all of the variables have **Linear Gaussian Models**.

Definition

Let Y be a continuous variable with continuous parents $X_1, \ldots, X_k$.
We say that Y has a **Linear Gaussian Model** if there are parameters $\beta_0, \beta_1, \ldots, \beta_k$ and σ^2 such that

$$p(Y \mid x_1, \ldots, x_k) = \mathcal{N}(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \; ; \; \sigma^2)$$

In vector notation:

$$p(Y \mid \boldsymbol{x}) = \mathcal{N}(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x} \; ; \; \sigma^2)$$

An Example: The Simplified Bike Scenario

- A bicyclist rides along a road (for simplicity: road is a straight line)
- Position is measured once every second
- Model bike's position (in m) at second t by a continuous variable $X^{(t)}$
- Model the velocity of the bike (in m/sec) by a continuous variable $V^{(t)}$

Ideally:

- Would expect that $X^{(t+1)} = X^{(t)} + V^{(t)}$.
  (If bicycle is at meter 510.4 at time t, and its velocity is $5m/sec$, then we expect its position $X^{(t+1)}$ the next second to be at meter 515.4 ...)

But:

- There is inevitably some random imprecision in real-world motion
- More realistic to say: Bicycle's position $X^{(t+1)}$ is described by a **normal distribution** whose mean is 515.4 and whose variance is 0.8 meters.

Some Notes on Linear Gaussian Models

Alternative Interpretation of a Linear Model $P(Y \mid X_1, ..., X_k)$:

- Y is a linear function of the variables $X_1, \ldots, X_k$, with the addition of Gaussian noise (randomly distributed "error") ϵ:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon$$

where ϵ is a normally distributed random variable with mean 0 and variance σ^2.

Practical Relevance:

- The Gaussian, linear dependency, and fixed variance assumptions are rather severe restrictions
- May be satisfied in some processes and applications, and not in others
- ... but they are reasonably good models of many physical phenomena
- ... and they support simple and elegant reasoning algorithms.

Reasoning with Linear Gaussian Models

For a restricted family of models and a concrete reasoning algorithm, see our later chapter on Kalman Filters.

Advanced Topics (not treated here)

- **Full analysis of independence in BN graphs**
- **Algorithms for constructing graphs for given sets of independencies**
- **Other Types of Local Probability Models (CPDs):**
  Context-specific CPDs, Tree and Rule CPDs, Hybrid Discrete-Continuous Models, Logistic Models, ...

What you should remember of this section

- The probabilistic query
- Typical reasoning patterns
- Definition of Bayesian Networks
- Independencies encoded by Bayesian Networks
- The Chain Rule for Bayesian Networks
- How to compute event probabilities from BNs
- The notion of independent / non-redundant parameters
- How BNs support a compact representation of the Full Joint Distribution
- The basic idea of Linear Gaussian Models[7]

[7] You'll learn this later, when we talk about the Kalman Filter.

Literature

Koller, D. and Friedman, N. (2009).
Probabilistic Graphical Models: Principles and Techniques. Cambridge, MA: MIT Press.

Beinlich, L., Suermondt, H., Chavez, R., and Cooper, G. (1989).
The ALARM Monitoring System: A Case Study with Two Probabilistic Inference Techniches for Belief Networks. In *Proceedings of the 2nd European Conference on Arificial Intelligence in Medicine*, pp.247-256. Berlin: Springer Verlag.