Artificial Intelligence 2
Summer Semester 2024

– Lecture Notes –
Part V: Reasoning with Uncertain Knowledge

Prof. Dr. Michael Kohlhase

Professur für Wissensrepräsentation und -verarbeitung
Informatik, FAU Erlangen-Nürnberg
Michael.Kohlhase@FAU.de

2023-05-02

This document contains Part V of the course notes for the course "Artificial Intelligence 2" held at FAU Erlangen-Nürnberg in the Summer Semesters 2017 ff. and something here. This part of the course notes addresses inference and agent decision making in partially observable environments, i.e. where we only know probabilities instead of certainties whether propositions are true/false. We cover basic probability theory and – based on that – Bayesian Networks and simple decision making in such environments. Finally we extend this to probabilistic temporal models and their decision theory. Other parts of the lecture notes can be found at `http://kwarc.info/teaching/AI/notes-*.pdf`. **Syllabus and Schedule – Summer Semester 2023:**

#	date	until	slide	page
1	18. 4.	recap, Overview, some admin, ALeA		
2	19. 4.	Agents & Uncertainty	48	14
3	25. 4.	Utility Agents, Probabilities, Indepencence		
4	26. 4.	Chain rule, Marginalization, Normalization, conditional independence	87	33

Contents

Chapter 1

Quantifying Uncertainty

In this chapter we develop a machinery for dealing with uncertainty: Instead of thinking about what we know to be true, we must think about what is likely to be true.

1.1 Dealing with Uncertainty: Probabilities

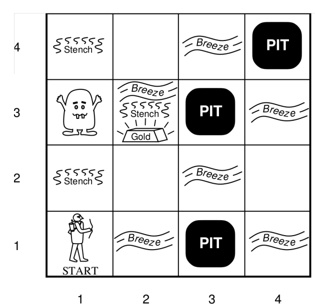
Before we go into the technical machinery in section 1.1, let us contemplate the sources of uncertainty our agents might have to deal with (subsection 1.1.1) and how the agent models need to be extended to cope with that (section 21.4 (Agent Architectures based on Belief States) in the AI lecture notes).

1.1.1 Sources of Uncertainty

A Video Nugget covering this subsection can be found at `https://fau.tv/clip/id/27582`.

Sources of Uncertainty in Decision-Making

Where's that d. . . Wumpus?
And where am I, anyway??



- ▷ **Non-deterministic actions:**
  - ▷ "When I try to go forward in this dark cave, I might actually go forward-left or forward-right."
- ▷ **Partial observability with unreliable sensors:**
  - ▷ "Did I feel a breeze right now?";
  - ▷ "I think I might smell a Wumpus here, but I got a cold and my nose is blocked."
  - ▷ "According to the heat scanner, the Wumpus is probably in cell [2,3]."

- **Uncertainty about the domain behavior:**
  - "Are you *sure* the Wumpus never moves?"

Unreliable Sensors

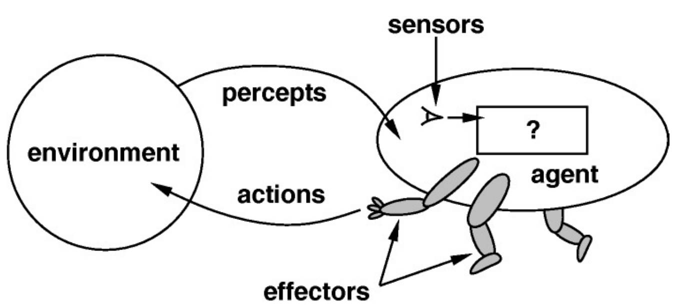
- **Robot Localization:** Suppose we want to support localization using landmarks to narrow down the area.
- **Example 1.1.1.** *If you see the Eiffel tower, then you're in Paris.*
- **Difficulty:** Sensors can be imprecise.
  - Even if a landmark is perceived, we cannot conclude with certainty that the robot is at that location.
  - *This is the half-scale Las Vegas copy, you dummy.*
  - Even if a landmark is *not* perceived, we cannot conclude with certainty that the robot is *not* at that location.
  - *Top of Eiffel tower hidden in the clouds.*
- Only the probability of being at a location increases or decreases.

1.1.2 Recap: Rational Agents as a Conceptual Framework

A Video Nugget covering this subsection can be found at `https://fau.tv/clip/id/27585`.

Agents and Environments

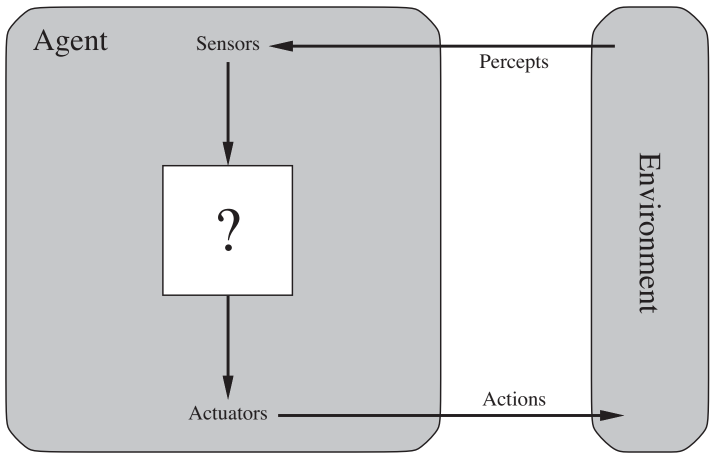
- **Definition 1.1.2.** An agent is anything that
  - perceives its environment via sensors (a means of sensing the environment)
  - acts on it with actuators (means of changing the environment).



- **Example 1.1.3.** Agents include humans, robots, softbots, thermostats, etc.

Agent Schema: Visualizing the Internal Agent Structure

▷ **Agent Schema:** We will use the following kind of schema to visualize the internal structure of an agent:



Different agents differ on the contents of the white box in the center.

Rationality

▷ **Idea:** Try to design agents that are successful! (aka. "do the right thing")

▷ **Definition 1.1.4.** A performance measure is a function that evaluates a sequence of environments.

▷ **Example 1.1.5.** A performance measure for the vacuum cleaner world could

- ▷ award one point per square cleaned up in time T?
- ▷ award one point per clean square per time step, minus one per move?
- ▷ penalize for $> k$ dirty squares?

▷ **Definition 1.1.6.** An agent is called rational, if it chooses whichever action maximizes the expected value of the performance measure given the percept sequence to date.

▷ **Question:** Why is rationality a good quality to aim for?

Consequences of Rationality: Exploration, Learning, Autonomy

▷ **Note:** a rational agent need not be perfect

- ▷ only needs to maximize expected value (rational $\neq$ omniscient)
  - ▷ need not predict e.g. very unlikely but catastrophic events in the future
- ▷ percepts may not supply all relevant information (rational $\neq$ clairvoyant)
  - ▷ if we cannot perceive things we do not need to react to them.
  - ▷ but we may need to try to find out about hidden dangers (exploration)
- ▷ action outcomes may not be as expected (rational $\neq$ successful)
  - ▷ but we may need to take action to ensure that they do (more often) (learning)

▷ **Note:** rational $\rightsquigarrow$ exploration, learning, autonomy

▷ **Definition 1.1.7.** An agent is called autonomous, if it does not rely on the prior knowledge about the environment of the designer.

▷ Autonomy avoids fixed behaviors that can become unsuccessful in a changing environment. (anything else would be irrational)

▷ The agent has to learning agentlearn all relevant traits, invariants, properties of the environment and actions.

PEAS: Describing the Task Environment

▷ **Observation:** To design a rational agent, we must specify the task environment in terms of performance measure, environment, actuators, and sensors, together called the PEAS components.

▷ **Example 1.1.8.** When designing an automated taxi:

- ▷ **Performance measure:** safety, destination, profits, legality, comfort, ...
- ▷ **Environment:** US streets/freeways, traffic, pedestrians, weather, ...
- ▷ **Actuators:** steering, accelerator, brake, horn, speaker/display, ...
- ▷ **Sensors:** video, accelerometers, gauges, engine sensors, keyboard, GPS, ...

▷ **Example 1.1.9 (Internet Shopping Agent).**
The task environment:

- ▷ Performance measure: price, quality, appropriateness, efficiency
- ▷ Environment: current and future WWW sites, vendors, shippers
- ▷ Actuators: display to user, follow URL, fill in form
- ▷ Sensors: HTML pages (text, graphics, scripts)

Environment types

▷ **Observation 1.1.10.** *Agent design is largely determined by the type of environment it is intended for.*

▷ **Problem:**

There is a vast number of possible kinds of environments in AI.

▷ **Solution:** Classify along a few "dimensions". (independent characteristics)

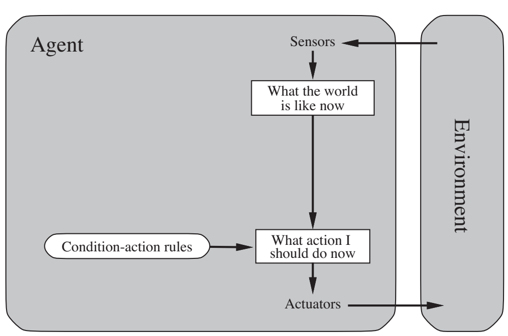
▷ **Definition 1.1.11.** For an agent a we classify the environment e of a by its type, which is one of the following. We call e

1. fully observable, iff the a's sensors give it access to the complete state of the environment at any point in time, else partially observable.
2. deterministic, iff the next state of the environment is completely determined by the current state and a's action, else stochastic.
3. episodic, iff a's experience is divided into atomic episodes, where it perceives and then performs a single action. Crucially the next episode does not depend on previous ones. Non-episodic environments are called sequential.
4. dynamic, iff the environment can change without an action performed by a, else static. If the environment does not change but a's performance measure does, we call e semidynamic.
5. discrete, iff the sets of e's state and a's actions are countable, else continuous.
6. single agent, iff only a acts on e; else multi agent (when must we count parts of e as agents?)

Simple reflex agents

▷ **Definition 1.1.12.** A simple reflex agent is an agent a that only bases its actions on the last percept: so the agent function simplifies to $f_a\colon \mathcal{P}\to\mathcal{A}$.

▷ **Agent Schema:**

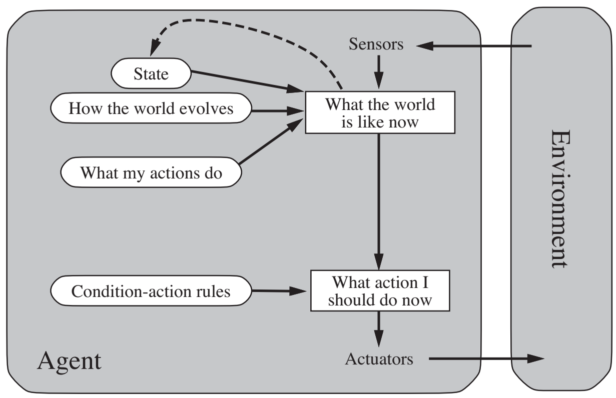


▷ **Example 1.1.13 (Agent Program).**

procedure Reflex—Vacuum—Agent [location,status] **returns** an action
  if status = Dirty **then** . . .

Model-based Reflex Agents: Idea

- **Idea:** Keep track of the state of the world we cannot see in an internal model.
- **Agent Schema:**



Model-based Reflex Agents: Definition

- **Definition 1.1.14.** A model based agent (also called reflex agent with state) is an agent whose function depends on
  - a world model: a set $\mathcal{S}$ of possible states.
  - a sensor model S that given a state s and percepts determines a new state s'.
  - (optionally) a transition model T, that predicts a new state s'' from a state s' and an action a .
  - An action function f that maps (new) states to actions.

  The agent function is iteratively computed via $e{\mapsto}f(S(s,e))$.
- **Note:** As different percept sequences lead to different states, so the agent function $f_a\colon\mathcal{P}^*{\rightarrow}\mathcal{A}$ no longer depends only on the last percept.
- **Example 1.1.15 (Tail Lights Again).** Model based agents can do the section 93 (Types of Agents) in the AI lecture notes if the states include a concept of tail light brightness.

1.1.3 Agent Architectures based on Belief States

A Video Nugget covering this subsection can be found at `https://fau.tv/clip/id/29041`.
We are now ready to proceed to environments which can only partially observed and where are our actions are non deterministic. Both sources of uncertainty conspire to allow us only partial

knowledge about the world, so that we can only optimize "expected utility" instead of "actual utility" of our actions.

World Models for Uncertainty

- **Problem:** We do not know with certainty what state the world is in!
- **Idea:** Just keep track of all the possible states it could be in.
- **Definition 1.1.16.** A model based agent has a world model consisting of
  - a belief state that has information about the possible states the world may be in, and
  - a sensor model that updates the belief state based on sensor information
  - a transition model that updates the belief state based on actions.
- **Idea:** The agent environment determines what the world model can be.
- In a fully observable, deterministic environment,
  - we can observe the initial state and subsequent states are given by the actions alone.
  - thus the belief state is a singleton set (we call its member the world state) and the transition model is a function from states and actions to states: a transition function.

That is exactly what we have been doing until now: we have been studying methods that build on descriptions of the "actual" world, and have been concentrating on the progression from atomic to factored and ultimately structured representations. Tellingly, we spoke of "world states" instead of "belief states"; we have now justified this practice in the brave new belief-based world models by the (re-) definition of "world states" above. To fortify our intuitions, let us recap from a belief-state-model perspective.

World Models by Agent Type in AI-1

- **Note:** All of these considerations only give requirements to the world model. What we can do with it depends on representation and inference.
- **Search-based Agents:** In a fully observable, deterministic environment
  - goal based agent with world state $\widehat{=}$ "current state"
  - no inference. (goal $\widehat{=}$ goal state from search problem)
- **CSP-based Agents:** In a fully observable, deterministic environment
  - goal based agent withworld state $\widehat{=}$ constraint network,
  - inference $\widehat{=}$ constraint propagation. (goal $\widehat{=}$ satisfying assignment)
- **Logic-based Agents:** In a fully observable, deterministic environment
  - model based agent with world state $\widehat{=}$ logical formula

- ▹ inference $\widehat{=}$ e.g. DPLL or resolution. (no decision theory covered in AI-1)

▷ **Planning Agents:** In a fully observable, deterministic, environment

- ▹ goal based agent with world state $\widehat{=}$ PL0, transition model $\widehat{=}$ STRIPS,
- ▹ inference $\widehat{=}$ state/plan space search. (goal: complete plan/execution)

Let us now see what happens when we lift the restrictions of total observability and determinism.

World Models for Complex Environments

▷ In a fully observable, but stochastic environment,

- ▹ the belief state must deal with a set of possible states.
- ▹ $\rightsquigarrow$ generalize the transition function to a transition relation.

▷ **Note:** This even applies to online problem solving, where we can just perceive the state. (e.g. when we want to optimize utility)

▷ In a deterministic, but partially observable environment,

- ▹ the belief state must deal with a set of possible states.
- ▹ we can use transition functions.
- ▹ We need a sensor model, which predicts the influence of percepts on the belief state – during update.

▷ In a stochastic, partially observable environment,

- ▹ mix the ideas from the last two. (sensor model + transition relation)

Preview: New World Models (Belief) $\rightsquigarrow$ new Agent Types

▷ **Probabilistic Agents:** In a partially observable environment

- ▹ belief state $\widehat{=}$ Bayesian networks,
- ▹ inference $\widehat{=}$ probabilistic inference.

▷ **Decision-Theoretic Agents:**
In a partially observable, stochastic environment

- ▹ belief state + transition model $\widehat{=}$ decision networks,
- ▹ inference $\widehat{=}$ maximizing expected utility.

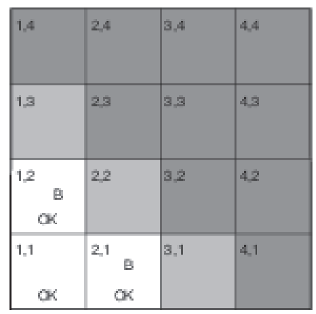
▷ We will study them in detailthis semester.

1.1.4 Modeling Uncertainty

A Video Nugget covering this subsection can be found at `https://fau.tv/clip/id/29043`. So we have extended the agent's world models to use sets of possible worlds instead of single (deterministic) world states. Let us evaluate whether this is enough for them to survive in the world.

Wumpus World Revisited

- **Recall:** We have updated agents with world/transition models with possible worlds.
- **Problem:** But pure sets of possible worlds are not enough
- **Example 1.1.17 (Beware of the Pit).**
  - We have a maze with pits that are detected in neighbouring squares via breeze (Wumpus and gold will not be assumed now).
  - Where does the agent should go, if there is breeze at (1,2) and (2,1)?
  - **Problem**: (1.3), (2,2), and (3.1) are all unsafe! (there are possible worlds with pits in any of them)



- **Idea:** We need world models that estimate the pit-likelyhood in cells!

Uncertainty and Logic

- **Example 1.1.18 (Diagnosis).** We want to build an expert dental diagnosis system, that deduces the cause (the disease) from the symptoms.
- Can we base this on logic?
- **Attempt 1:** Say we have a toothache. How's about:

$$\forall p.\mathsf{Symptom}(p,\mathsf{toothache}) \Rightarrow \mathsf{Disease}(p,\mathsf{cavity})$$

  - Is this rule correct?
  - No, toothaches may have different causes ("cavity" $\widehat{=}$ "Loch im Zahn").
- **Attempt 2:** So what about this:

$$\forall p.\mathsf{Symptom}(p,\mathsf{toothache}) \Rightarrow (\mathsf{Disease}(p,\mathsf{cavity}) \vee \mathsf{Disease}(p,\mathsf{gingivitis}) \vee \ldots)$$

  - We don't know all possible causes.
  - And we'd like to be able to deduce which causes are more plausible!

Uncertainty and Logic, ctd.

- **Attempt 3:** Perhaps a "causal" rule is better?

$$\forall p.\mathrm{Disease}(p, \mathrm{cavity}) \Rightarrow \mathrm{Symptom}(p, \mathrm{toothache})$$

- **Question:** Is this rule correct?
- **Answer:** No, not all cavities cause toothaches.
- **Question:** Does this rule allow to deduce a cause from a symptom?
- **Answer:** No, setting $\mathrm{Symptom}(p, \mathrm{toothache})$ to true here has no consequence on the truth of $\mathrm{Disease}(p, \mathrm{cavity})$.
- **Note:** If $\mathrm{Symptom}(p, \mathrm{toothache})$ is *false*, we would conclude $\neg\mathrm{Disease}(p, \mathrm{cavity})$ ... which would be incorrect, cf. previous question.
- Anyway, this still doesn't allow to compare the plausibility of different causes.
- **Summary:** Logic does not allow to weigh different alternatives, and it does not allow to express incomplete knowledge ("cavity does not always come with a toothache, nor vice versa").

Beliefs and Probabilities

- **Question:** What do we model with probabilities?
- **Answer:** Incomplete knowledge!
  - We are certain, but we *believe to a certain degree* that something is true.
  - Probability $\widehat{=}$ Our degree of belief, given our current knowledge.
- **Example 1.1.19 (Diagnosis).**
  - $\mathrm{Symptom}(p, \mathrm{toothache}) \Rightarrow \mathrm{Disease}(p, \mathrm{cavity})$ with 80% probability.
  - But, for any given p, in reality we do, or do not, have cavity: 1 or 0!
  - The "probability" depends on our knowledge!
  - The "80%" refers to the fraction of cavities within the set of all p' that are indistinguishable from p based on our knowledge.
  - If we receive new knowledge (e.g., $\mathrm{Disease}(p, \mathrm{gingivitis})$), the probability changes!
- Probabilities represent and measure the uncertainty that stems from lack of knowledge.

How to Obtain Probabilities?

- **Assessing probabilities through statistics:**
  - The agent is 90% convinced by its sensor information. (in 9 out of 10 cases, the information is correct)
  - $\text{Disease}(p, \text{cavity}) \Rightarrow \text{Symptom}(p, \text{toothache})$ with 80% probability $:\widehat{=}$ 8 out of 10 persons with a cavity have toothache.
- **Definition 1.1.20.** The process of estimating a probability P using statistics is called assessing P.
- **Observation:** Assessing even a single P can require huge effort!
- **Example 1.1.21.** The likelihood of making it to the university within 10 minutes.
- **What is probabilistic reasoning?** Deducing probabilities from knowledge about *other* probabilities.
- **Idea:** Probabilistic reasoning determines, based on probabilities that are (relatively) easy to assess, probabilities that are difficult to assess.

1.1.5 Acting under Uncertainty

A Video Nugget covering this subsection can be found at `https://fau.tv/clip/id/29044`.

Decision-Making Under Uncertainty

- **Example 1.1.22 (Giving a lecture).**
  - **Goal**: Be in HS002 at 10:15 to give a lecture.
  - **Possible plans**:
    - P_1: *Get up at 8:00, leave at 8:40, arrive at 9:00.*
    - P_2: *Get up at 9:50, leave at 10:05, arrive at 10:15.*
  - **Decision**: Both plans are correct, but P_2 succeeds only with probability 50%, and giving a lecture is important, so P_1 is the plan of choice.
- **Example 1.1.23 (Better Example).** Which train to take to Frankfurt airport?

Uncertainty and Rational Decisions

- **Here:** We're only concerned with deducing the likelihood of facts, not with action choice. In general, selecting actions is of course important.
- **Rational Agents:**
  - We have a choice of actions: go to FRA early or go to FRA just in time.
  - These can lead to different solutions with different probabilities.
  - The actions have different costs.

- ▷ The results have different utilities (safe timing/dislike airport food).

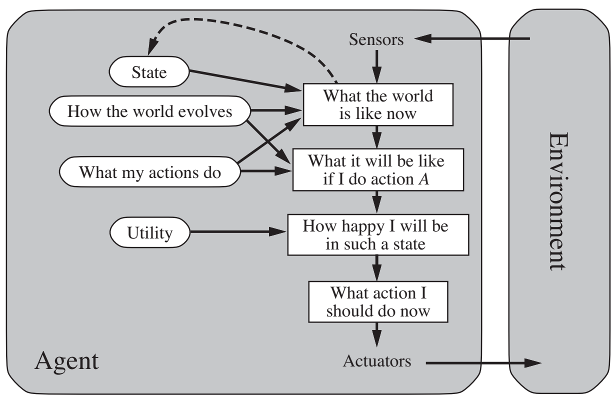
▷ A rational agent chooses the action with the maximum expected utility.

▷ Decision Theory $\widehat{=}$ Utility Theory + Probability Theory.

Utility-based agents

▷ **Definition 1.1.24.** A utility based agent uses a world model along with a utility function that models its preferences among the states of that world. It chooses the action that leads to the best expected utility.

▷ **Agent Schema:**



Decision-Theoretic Agent

▷ **Example 1.1.25 (A particular kind of utility-based agent).**

function DT-AGENT(*percept*) **returns** an *action*
  persistent: *belief_state*, probabilistic beliefs about the current state of the world
    action, the agent's action

  update *belief_state* based on *action* and *percept*
  calculate outcome probabilities for actions,
    given action descriptions and current *belief_state*
  select *action* with highest expected utility
    given probabilities of outcomes and utility information
  return *action*

1.1.6 Agenda for this Chapter: Basics of Probability Theory

A Video Nugget covering this subsection can be found at https://fau.tv/clip/id/29046.

Our Agenda for This Topic

- Our treatment of the topic "probabilistic reasoning" consists of this Chapter and the next.
  - This Chapter: All the basic machinery at use in Bayesian networks.
  - chapter 2: Bayesian networks: What they are, how to build them, how to use them.
- Bayesian networks are the most widespread and successful practical framework for probabilistic reasoning.

Our Agenda for This Chapter

- **Unconditional Probabilities and Conditional Probabilities:** Which concepts and properties of probabilities will be used?
  - Mostly a recap of things you're familiar with from school.
- **Independence and Basic Probabilistic Reasoning Methods:** What simple methods are there to avoid enumeration and to deduce probabilities from other probabilities?
  - A basic tool set we'll need. (Still familiar from school?)
- **Bayes' Rule:** What's that "Bayes"? How is it used and why is it important?
  - The basic insight about how to invert the "direction" of conditional probabilities.
- **Conditional Independence:** How to capture and exploit complex relations between random variables?
  - Explains the difficulties arising when using Bayes' rule on multiple evidences. conditional independence is used to ameliorate these difficulties.

1.2 Unconditional Probabilities

Video Nuggets covering this section can be found at https://fau.tv/clip/id/29047 and https://fau.tv/clip/id/29048.

Probabilistic Models

- **Definition 1.2.1.** A probability theory is an assertion language for talking about possible worlds and an inference method for quantifying the degree of belief in such assertions.

- **Remark:** Like logic, but for non binary belief degree.
- The possible worlds are
  - mutually exclusive: different possible worlds cannot both be the case and
  - exhaustive: one possible world must be the case.
- This determines the set of possible worlds.
- **Example 1.2.2.** If we roll two (distinguishable) dice with six sides, then we have 36 possible worlds: $(1{,}1)$, $(2{,}1)$, ..., $(6{,}6)$.
- We will restrict ourselves to a discrete, countable sample space. (others more complicated, less useful in AI)
- **Definition 1.2.3.** A probability model $\langle\Omega, P\rangle$ consists of a countable set Ω of possible worlds called the sample space and a probability function $P\colon \Omega\to\mathbb{R}$, such that $0\leq P(\omega)\leq 1$ for all $\omega\in\Omega$ and $\sum_{\omega\in\Omega} P(\omega) = 1$.

Unconditional Probabilities, Random Variables, and Events

- **Definition 1.2.4.** A random variable (also called random quantity, aleatory variable, or stochastic variable) is a variable quantity whose value depends on possible outcomes of unknown variables and processes we do not understand.
- **Definition 1.2.5.** If X is a random variable and x a possible value, we will refer to the fact $X = x$ as an outcome and a set of outcomes as an event. The set of possible outcomes of X is called the domain of X.
- The notation *uppercase "X"* for a random variable, and *lowercase "x"* for one of its values will be used frequently. (following Russel/Norvig)
- **Definition 1.2.6.** Given a random variable X, $P(X = x)$ denotes the prior probability, or unconditional probability, that X has value x in the absence of any other information.
- **Example 1.2.7.** $P(\text{Cavity} = \mathsf{T}) = 0.2$, where Cavity is a random variable whose value is true iff some given person has a cavity.

Types of Random Variables

- **Definition 1.2.8.** We say that a random variable X is finite domain, iff the domain D of X is finite and Boolean, iff $D = \{\mathsf{T}, \mathsf{F}\}$.
- **Note:** In general, random variables can have arbitrary domains. In AI-2, we restrict ourselves to finite domain and Boolean random variables.

▷ **Example 1.2.9.** Some prior probabilities

$$P(\text{Weather} = \text{sunny}) = 0.7$$
$$P(\text{Weather} = \text{rain}) = 0.2$$
$$P(\text{Weather} = \text{cloudy}) = 0.08$$
$$P(\text{Weather} = \text{snow}) = 0.02$$
$$P(\text{Headache} = \text{T}) = 0.1$$

Unlike us, Russel and Norvig live in California . . . :-(:-(

▷ **Convenience Notations:**

- ▷ By convention, we denote Boolean random variables with A, B, and more general finite domain random variables with X, Y.
- ▷ For a Boolean random variable Name, we write name for the outcome Name = T and ¬name for Name = F. (Follows Russel/Norvig as well)

Probability Distributions

▷ **Definition 1.2.10.** The probability distribution for a random variable X, written $\mathbf{P}(X)$, is the vector of probabilities for the (ordered) domain of X.

▷ **Example 1.2.11.** Probability distributions for finite domain and Boolean random variables

$$\mathbf{P}(\text{Headache}) = \langle 0.1, 0.9\rangle$$
$$\mathbf{P}(\text{Weather}) = \langle 0.7, 0.2, 0.08, 0.02\rangle$$

define the probability distribution for the random variables Headache and Weather.

▷ **Definition 1.2.12.**

Given a subset $\mathbf{Z} \subseteq \{X_1, \ldots, X_n\}$ of random variables, an event is an assignment of values to the variables in $\mathbf{Z}$. The joint probability distribution, written $\mathbf{P}(\mathbf{Z})$, lists the probabilities of all events.

▷ **Example 1.2.13.** $\mathbf{P}(\text{Headache}, \text{Weather})$ is

	Headache = T	$Headache$ = F
Weather = sunny	$P(W = \text{sunny} \wedge \text{headache})$	$P(W = \text{sunny} \wedge \neg\text{headache})$
Weather = rain		
Weather = cloudy		
Weather = snow		

The Full Joint Probability Distribution

▷ **Definition 1.2.14.**

Given random variables $\{X_1, \ldots, X_n\}$, an atomic event is an assignment of values to all variables.

▷ **Example 1.2.15.** If A and B are Boolean random variables, then we have four atomic events: $a \wedge b$, $a \wedge \neg b$, $\neg a \wedge b$, $\neg a \wedge \neg b$.

▷ **Definition 1.2.16.**

Given random variables $\{X_1, \ldots, X_n\}$, the full joint probability distribution, denoted $\mathbf{P}(X_1, \ldots, X_n)$, lists the probabilities of all atomic events.

▷ **Observation:**

Given random variables $X_1, \ldots, X_n$ with domains $D_1, \ldots, D_n$, the full joint probability distribution is an n-dimensional array of size $\langle D_1, \ldots, D_n \rangle$.

▷ **Example 1.2.17.** $\mathbf{P}(Cavity, Toothache)$

	toothache	¬toothache
cavity	0.12	0.08
¬cavity	0.08	0.72

▷ **Note:** All atomic events are disjoint (their pairwise conjunctions all are equivalent to F); the sum of all fields is 1 (the disjunction over all atomic events is T).

Probabilities of Propositional Formulae

▷ **Definition 1.2.18.**

Given random variables $\{X_1, \ldots, X_n\}$, a proposition is a PL^0 wff over the atoms $X_i = x_i$ where the x_i are values in the domains of X_i.

A function P that maps propositions into $[0,1]$ is a probability measure if

1. $P(\top) = 1$ and
2. for all propositions A, $P(A) = \sum_{e \models A} P(e)$ where e is an atomic event.

▷ Propositions represent sets of atomic events: the interpretations satisfying the formula.

▷ **Example 1.2.19.** $P(\text{cavity} \wedge \text{toothache}) = 0.12$ is the probability that some given person has both a cavity and a toothache. (Note the use of cavity for Cavity $=$ T and toothache for Toothache $=$ T.)

▷ **Notes:**

- ▷ Instead of $P(a \wedge b)$, we often write $P(a, b)$.
- ▷ Propositions can be viewed as Boolean random variables; we will denote them with A, B as well.

The role of clause 2 in Definition 1.2.18 is for P to "make sense": intuitively, the probability weight of a formula should be the sum of the weights of the interpretations satisfying it. Imagine this was not so; then, for example, we could have $P(A) = 0.2$ and $P(A \wedge B) = 0.8$. The role of 1 here is to "normalize" P so that the maximum probability is 1. (The minimum probability is 0 simply because of 1: the empty sum has weight 0).

Kolmogorov and Negation

▷ **Theorem 1.2.20 (Kolmogorow).** *A function P that maps propositions into $[0,1]$ is a probability measure if and only if*

i $P(\top) = 1$ *and*

ii' *for all propositions A, B:* $P(a \vee b) = P(a) + P(b) - P(a \wedge b)$.

▷ **Observation:** We can equivalently replace

ii for all propositions A, $P(A) = \sum_{I \models A} P(I)$ with Kolmogorow's (*ii'*).

▷ **Question:** Assume we have

iii $P(\bot) = 0$.

How to derive from (*i*), (*ii'*), and (*iii*) that, for all propositions A, $P(\neg a) = 1 - P(a)$?

▷ **Answer:** reserved for the plenary sessions $\leadsto$ be there!

Believing in Kolmogorov?

▷ **Reminder 1:** (*i*) $P(\top) = 1$; (*ii'*) $P(a \vee b) = P(a) + P(b) - P(a \wedge b)$.

▷ **Reminder 2:** "Probabilities model our belief."

- ▹ If P represents an objectively observable probability, the axioms clearly make sense.
- ▹ But why should an agent respect these axioms, when modeling its subjective own belief?

▷ **Question:** Do you believe in Kolmogorow's axioms?

▷ **Answer:** reserved for the plenary sessions $\leadsto$ be there!

1.3 Conditional Probabilities

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/29049`.

Conditional Probabilities: Intuition

▷ Do probabilities change as we gather new knowledge?

▷ Yes! Probabilities model our *belief*, thus they depend on our knowledge.

▷ **Example 1.3.1.** Your "probability of missing the connection train" increases when you are informed that your current train has 30 minutes delay.

▷ **Example 1.3.2.** The "probability of cavity" increases when the doctor is informed that the patient has a toothache.

▷ In the presence of additional information, we can no longer use the unconditional (*prior!*) probabilities.

▷ Given propositions A and B, $P(a|b)$ denotes the conditional probability of a (i.e., $A = \mathsf{T}$) given that all we know is b (i.e., $B = \mathsf{T}$).

▷ **Example 1.3.3.** $P(\text{cavity}) = 0.2$ vs. $P(\text{cavity}|\text{toothache}) = 0.6$.

▷ **Example 1.3.4.** $P(\text{cavity}|\text{toothache} \wedge \neg\text{cavity}) = 0$

Conditional Probabilities: Definition

▷ **Definition 1.3.5.** Given propositions A and B where $P(b) \neq 0$, the conditional probability, or posterior probability, of a given b, written $P(a|b)$, is defined as:

$$P(a|b) := \frac{P(a \wedge b)}{P(b)}$$

▷ **Intuition:** The likelihood of having a and b, within the set of outcomes where we have b.

▷ **Example 1.3.6.** $P(\text{cavity} \wedge \text{toothache}) = 0.12$ and $P(\text{toothache}) = 0.2$ yield $P(\text{cavity}|\text{toothache}) = 0.6$.

Conditional Probability Distributions

▷ **Definition 1.3.7.** Given random variables X and Y, the conditional probability distribution of X given Y, written $\mathbf{P}(X|Y)$, i.e. with a boldface P, is the table of all conditional probabilities of values of X given values of Y.

▷ For sets of variables: $\mathbf{P}(X_1, \ldots, X_n | Y_1, \ldots, Y_m)$.

▷ **Example 1.3.8.** $\mathbf{P}(\text{Weather}|\text{Headache}) =$

	Headache $= \mathsf{T}$	$Headache = \mathsf{F}$
Weather = sunny	$P(W = \text{sunny}|\text{headache})$	$P(W = \text{sunny}|\neg\text{headache})$
Weather = rain		
Weather = cloudy		
Weather = snow		

What is *The probability of sunshine given that I have a headache?*

▷ If you're susceptible to headaches depending on weather conditions, this makes sense. Otherwise, the two variables are independent. (see next section)

1.4 Independence

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/29050`.

Working with the Full Joint Probability Distribution

▷ **Example 1.4.1.** Consider the following full joint probability distribution:

	toothache	¬toothache
cavity	0.12	0.08
¬cavity	0.08	0.72

- ▷ How to compute $P(\text{cavity})$?
- ▷ Sum across the row:

$$P(\text{cavity} \wedge \text{toothache}) + P(\text{cavity} \wedge \neg\text{toothache}) = 0.2$$

- ▷ How to compute $P(\text{cavity} \vee \text{toothache})$?
- ▷ Sum across atomic events:

$$P(\text{cavity}\wedge\text{toothache})+P(\neg\text{cavity}\wedge\text{toothache})+P(\text{cavity}\wedge\neg\text{toothache}) = 0.28$$

- ▷ How to compute $P(\text{cavity}|\text{toothache})$?
- ▷ $\frac{P(\text{cavity}\wedge\text{toothache})}{P(\text{toothache})}$
- ▷ All relevant probabilities can be computed using the full joint probability distribution, by expressing propositions as disjunctions of atomic events.

Working with the Full Joint Probability Distribution??

▷ **Question:** Is it a good idea to use the full joint probability distribution?

▷ **Answer:** No:

- ▷ Given n random variables with k values each, the full joint probability distribution contains k^n probabilities.
- ▷ Computational cost of dealing with this size.
- ▷ Practically impossible to assess all these probabilities.

▷ **Question:** So, is there a compact way to represent the full joint probability distribution? Is there an efficient method to work with that representation?

▷ **Answer:** Not in general, but it works in many cases. We can work directly with conditional probabilities, and exploit conditional independence.

▷ **Eventually:** Bayesian networks. (First, we do the simple case)

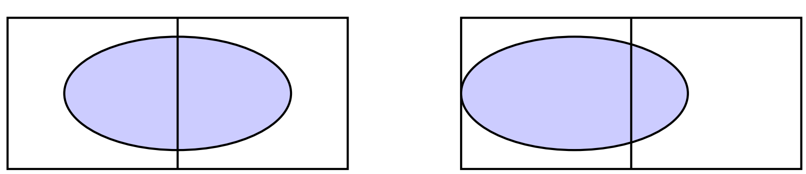
Independence of Events and Random Variables

- **Definition 1.4.2.** Events a and b are independent if $P(a \wedge b) = P(a) \cdot P(b)$.
- Given independent events a and b where $P(b) \neq 0$, we have $P(a|b) = P(a)$.
- *Proof:*
  1. By definition, $P(a|b) = \frac{P(a \wedge b)}{P(b)}$,
  2. which by independence is equal to $\frac{P(a) \cdot P(b)}{P(b)} = P(a)$.
- Similarly, if $P(a) \neq 0$, we have $P(b|a) = P(b)$.
- **Definition 1.4.3.** Random variables X and Y are independent if $\mathbf{P}(X, Y) = \mathbf{P}(X) \otimes \mathbf{P}(Y)$. (System of equations given by outer product!)

Independence (Examples)

- **Example 1.4.4.**
  - $P(\text{Die1} = 6 \wedge \text{Die2} = 6) = 1/36$.
  - $P(W = \text{sunny}|\text{headache}) = P(W = \text{sunny})$ (unless you're weather-sensitive; cf. slide 70)
  - But toothache and cavity are NOT independent.
  - The fraction of "cavity" is higher within "toothache" than within "$\neg$toothache". $P(\text{toothache}) = 0.2$ and $P(\text{cavity}) = 0.2$, but $P(\text{toothache} \wedge \text{cavity}) = 0.12 > 0.04$.
- **Intuition:**

Independent Dependent



Oval independent of rectangle, iff split equally

Illustration: Exploiting Independence

- **Example 1.4.5.** Consider (again) the following full joint probability distribution:

	toothache	¬toothache
cavity	0.12	0.08
¬cavity	0.08	0.72

Adding variable Weather with values sunny, rain, cloudy, snow, the full joint probability distribution contains 16 probabilities.

But your teeth do not influence the weather, nor vice versa!

- ▷ Weather is independent of each of Cavity and Toothache: For all value combinations (c,t) of Cavity and Toothache, and for all values w of Weather, we have $P(c \wedge t \wedge w) = P(c \wedge t) \cdot P(w)$.
- ▷ $\mathbf{P}(\text{Cavity}, \text{Toothache}, \text{Weather})$ can be reconstructed from the separate tables $\mathbf{P}(\text{Cavity}, \text{Toothache})$ and $\mathbf{P}(\text{Weather})$. (8 probabilities)

▷ Independence can be exploited to represent the full joint probability distribution more compactly.

▷ Sometimes, variables are independent only under particular conditions: conditional independence. (see later)

1.5 Basic Probabilistic Reasoning Methods

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/29051`.

The Product Rule

▷ **Definition 1.5.1.** The following identity is called the product rule:
Given propositions a and b, $P(a \wedge b) = P(a|b) \cdot P(b)$.

▷ **Note:** The product rule is a direct consequence of the definition of conditional probability

▷ **Example 1.5.2.** $P(\text{cavity} \wedge \text{toothache}) = P(\text{toothache}|\text{cavity}) \cdot P(\text{cavity})$.

▷ If we know the values of $P(a|b)$ and $P(b)$, then we can compute $P(a \wedge b)$.

▷ Similarly, $P(a \wedge b) = P(b|a) \cdot P(a)$.

▷ **Definition 1.5.3.** We use the component wise array product (bold dot)
$\mathbf{P}(X,Y) = \mathbf{P}(X|Y) \boldsymbol{\cdot} \mathbf{P}(Y)$ as a summary notation for the equation system $\mathbf{P}(x_i, y_j) = \mathbf{P}(x_i|y_j) \cdot \mathbf{P}(y_j)$ where i, j range over domain sizes of X and Y.

▷ **Example 1.5.4.** $\mathbf{P}(\text{Weather}, \text{Ache}) = \mathbf{P}(\text{Weather}|\text{Ache}) \boldsymbol{\cdot} \mathbf{P}(\text{Ache})$ is

$$\begin{aligned}
P(W = \text{sunny} \wedge \text{ache}) &= P(W = \text{sunny}|\text{ache}) \cdot P(\text{ache}) \\
P(W = \text{rain} \wedge \text{ache}) &= P(W = \text{rain}|\text{ache}) \cdot P(\text{ache}) \\
\cdots &= \cdots \\
P(W = \text{snow} \wedge \neg\text{ache}) &= P(W = \text{snow}|\neg\text{ache}) \cdot P(\neg\text{ache})
\end{aligned}$$

▷ **Note:** The outer product in $\mathbf{P}(X,Y) = \mathbf{P}(X) \boldsymbol{\cdot} \mathbf{P}(Y)$ is just by conincidence, we will use $\mathbf{P}(X,Y) = \mathbf{P}(X) \boldsymbol{\cdot} \mathbf{P}(Y)$ instead.

The component wise array product from Definition 1.5.3 is something that Russell/Norvig (and the literature in general) glosses over and sweeps under the rug. The problem is that it is not a real mathematical operator, that can be defined notation independently, because it depends on the indices in the representation. But the notation is just too convenient to bypass.
It is just a coincidence that we can use the outer product in probability distributions $\mathbf{P}(X, Y) = \mathbf{P}(X) \cdot \mathbf{P}(Y)$. Here, the outer product and component wise array product co-incide.

The Chain Rule

- ▷ **Lemma 1.5.5 (Chain Rule).** *Given random variables $X_1, \ldots, X_n$, we have*

  $$\mathbf{P}(X_1, \ldots, X_n) = \mathbf{P}(X_n | X_{n-1}, \ldots, X_1) \cdot \mathbf{P}(X_{n-1} | X_{n-2}, \ldots, X_1) \cdot \ldots \cdot \mathbf{P}(X_2 | X_1) \cdot \mathbf{P}(X_1)$$

  This identity is called the chain rule.

- ▷ **Example 1.5.6.**

$$\begin{aligned} & P(\neg\text{brush} \wedge \text{cavity} \wedge \text{toothache}) \\ = \; & P(\text{toothache}|\text{cavity}, \neg\text{brush}) \cdot P(\text{cavity}, \neg\text{brush}) \\ = \; & P(\text{toothache}|\text{cavity}, \neg\text{brush}) \cdot P(\text{cavity}|\neg\text{brush}) \cdot P(\neg\text{brush}) \end{aligned}$$

- ▷ *Proof:* Iterated application of the product rule
  1. $\mathbf{P}(X_1, \ldots, X_n) = \mathbf{P}(X_n | X_{n-1}, \ldots, X_1) \cdot \mathbf{P}(X_{n-1}, \ldots, X_1)$ by the product rule.
  2. In turn, $\mathbf{P}(X_{n-1}, \ldots, X_1) = \mathbf{P}(X_{n-1} | X_{n-2}, \ldots, X_1) \cdot \mathbf{P}(X_{n-2}, \ldots, X_1)$, etc.

- ▷ **Note:** This works *for any ordering* of the variables.
  - ▷ We can recover the probability of atomic events from sequenced conditional probabilities for any ordering of the variables.
  - ▷ First of the four basic techniques in Bayesian networks.

Marginalization

- ▷ Extracting a sub-distribution from a larger joint distribution:
- ▷ Given sets $\mathbf{X}$ and $\mathbf{Y}$ of random variables, we have:

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

  where $\sum_{\mathbf{y} \in \mathbf{Y}}$ sums over all possible value combinations of $\mathbf{Y}$.

▷ **Example 1.5.7.** (Note: Equation system!)

$$\mathbf{P}(\mathsf{Cavity}) = \sum_{y \in \mathsf{Toothache}} \mathbf{P}(\mathsf{Cavity}, y)$$
$$P(\mathsf{cavity}) = P(\mathsf{cavity}, \mathsf{toothache}) + P(\mathsf{cavity}, \neg\mathsf{toothache})$$
$$P(\neg\mathsf{cavity}) = P(\neg\mathsf{cavity}, \mathsf{toothache}) + P(\neg\mathsf{cavity}, \neg\mathsf{toothache})$$

Questionnaire: Rules of Probabilistic Reasoning

▷ Say $P(\mathsf{dog}) = 0.4$, $(\neg\mathsf{dog}) \Leftrightarrow \mathsf{cat}$, and $P(\mathsf{likeslasagna}|\mathsf{cat}) = 0.5$.

▷ **Question:** Is $P(\mathsf{likeslasagna} \wedge \mathsf{cat})$ is A: 0.2, B: 0.5, C: 0.475, D: 0.3

▷ **Answer:** reserved for the plenary sessions ⇝ be there!

▷ **Question:** Can we compute the value of $P(\mathsf{likeslasagna})$, given the above informations?

▷ **Answer:** reserved for the plenary sessions ⇝ be there!

We now come to a very important technique of computing unknown probabilities, which looks almost like magic. Before we formally define it on the next slide, we will get an intuition by considering it in the context of our dentistry example.

Normalization: Idea

▷ **Problem:** We know $P(\mathsf{cavity} \wedge \mathsf{toothache})$ but don't know $P(\mathsf{toothache})$.

▷ **Step 1:** Case distinction over values of Cavity: ($P(\mathsf{toothache})$ as an unknown)

$$\begin{aligned} P(\mathsf{cavity}|\mathsf{toothache}) &= \frac{P(\mathsf{cavity} \wedge \mathsf{toothache})}{P(\mathsf{toothache})} = \frac{0.12}{P(\mathsf{toothache})} \\ P(\neg\mathsf{cavity}|\mathsf{toothache}) &= \frac{P(\neg\mathsf{cavity} \wedge \mathsf{toothache})}{P(\mathsf{toothache})} = \frac{0.08}{P(\mathsf{toothache})} \end{aligned}$$

▷ **Step 2:** Assuming placeholder $\alpha := 1/P(\mathsf{toothache})$:

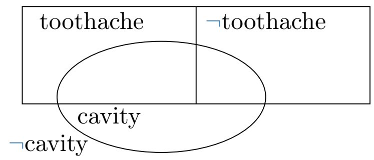
$$\begin{aligned} P(\mathsf{cavity}|\mathsf{toothache}) &= \alpha P(\mathsf{cavity} \wedge \mathsf{toothache}) = \alpha 0.12 \\ P(\neg\mathsf{cavity}|\mathsf{toothache}) &= \alpha P(\neg\mathsf{cavity} \wedge \mathsf{toothache}) = \alpha 0.08 \end{aligned}$$

▷ **Step 3:** Fixing toothache to be true, view $P(\mathsf{cavity} \wedge \mathsf{toothache})$ vs. $P(\neg\mathsf{cavity} \wedge \mathsf{toothache})$ as the "relative weights of $P(\mathsf{cavity})$ vs. $P(\neg\mathsf{cavity})$ within toothache".

Then normalize their summed-up weight to 1:
$1 = \alpha(0.12 + 0.08) \rightsquigarrow \alpha = \frac{1}{0.12+0.08} = \frac{1}{0.2} = 5$

▷ α is a normalization constant scaling the sum of relative weights to 1.

To understand what is going on, consider the situation in the following diagram:



Now consider the areas of $A_1 = \text{toothache} \wedge \text{cavity}$ and $A_2 = \text{toothache} \wedge \neg\text{cavity}$ then $A_1 \cup A_2 = \text{toothache}$; this is exactly what we will exploit (see next slide), but we notate it slightly differently in what will be a convenient manner in step 1.

In step 2 we only introduce a convenient placeholder α that makes subsequent argumentation easier.

In step 3, we view A_1 and A_2 as "relative weights"; say that we perceive the left half as "1" (because we already know toothache and don't need to worry about ¬toothache), and we re-normalize to get the desired sum $\alpha A_1 + \alpha A_2 = 1$.

Normalization

▷ **Question:** Say we know $P(\text{likeschappi} \wedge \text{dog}) = 0.32$ and $P(\neg\text{likeschappi} \wedge \text{dog}) = 0.08$. Can we compute $P(\text{likeschappi}|\text{dog})$? (Chappi $\widehat{=}$ popular dog food)

▷ **Answer:** reserved for the plenary sessions ⇝ be there!

▷ **Question:** So what is $P(\text{likeschappi}|\text{dog})$?

▷ **Answer:** reserved for the plenary sessions ⇝ be there!

Normalization: Formal

▷ **Definition 1.5.8.**

Given a vector $\langle w_1, \ldots, w_k \rangle$ of numbers in $[0,1]$ where $\sum_{i=1}^{k} w_i \leq 1$, the normalization constant α is $\alpha\langle w_1, \ldots, w_k \rangle := \frac{1}{\sum_{i=1}^{k} w_i}$.

▷ **Note:**

The condition $\sum_{i=1}^{k} w_i \leq 1$ is needed because these will be relative weights, i.e. case distinction over a subset of all worlds (the one fixed by the knowledge in our conditional probability).

▷ **Example 1.5.9.** $\alpha\langle 0.12, 0.08 \rangle = 5\langle 0.12, 0.08 \rangle = \langle 0.6, 0.4 \rangle$.

▷ Given a random variable X and an event $\mathbf{e}$, we have $\mathbf{P}(X|\mathbf{e}) = \alpha\mathbf{P}(X, \mathbf{e})$.

Proof:

1. For each value x of X, $P(X = x|\mathbf{e}) = P(X = x \wedge \mathbf{e})/P(\mathbf{e})$.
2. So all we need to prove is that $\alpha = 1/P(\mathbf{e})$.

3. By definition, $\alpha = 1/\sum_x P(X = x \wedge \mathbf{e})$, so we need to prove

$$P(\mathbf{e}) = \sum_x P(X = x \wedge \mathbf{e})$$

which holds by marginalization.

Normalization: Formal

▷ **Example 1.5.10.** $\alpha\langle P(\text{cavity}\wedge\text{toothache}), P(\neg\text{cavity}\wedge\text{toothache})\rangle = \alpha\langle 0.12, 0.08\rangle$, so $P(\text{cavity}|\text{toothache}) = 0.6$, and $P(\neg\text{cavity}|\text{toothache}) = 0.4$.

▷ Another way of saying this is: "We use α as a placeholder for $1/P(\mathbf{e})$, which we compute using the sum of relative weights by Marginalization."

▷ **Computation Rule: Normalization+Marginalization**
Given "query variable" X, "observed event" $\mathbf{e}$, and "hidden variables" set $\mathbf{Y}$:

$$\mathbf{P}(X|\mathbf{e}) = \alpha \cdot \mathbf{P}(X, \mathbf{e}) = \alpha \cdot (\sum_{\mathbf{y}\in\mathbf{Y}} \mathbf{P}(X, \mathbf{e}, \mathbf{y}))$$

▷ Second of the four basic techniques in Bayesian networks.

1.6 Bayes' Rule

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/29053`.

Bayes' Rule

▷ **Definition 1.6.1 (Bayes' Rule).** Given propositions A and B where $P(a) \neq 0$ and $P(b) \neq 0$, we have:

$$P(a|b) = \frac{P(b|a) \cdot P(a)}{P(b)}$$

This equation is called Bayes' rule.

▷ *Proof:*

1. By definition, $P(a|b) = \frac{P(a\wedge b)}{P(b)}$
2. by the product rule $P(a \wedge b) = P(b|a) \cdot P(a)$ is equal to the claim.

▷ **Notation:** This is a system of equations!

$$\mathbf{P}(X|Y) = \frac{\mathbf{P}(Y|X) \cdot \mathbf{P}(X)}{\mathbf{P}(Y)}$$

Applying Bayes' Rule

- **Example 1.6.2.** Say we know that $P(\text{toothache}|\text{cavity}) = 0.6$, $P(\text{cavity}) = 0.2$, and $P(\text{toothache}) = 0.2$.

  We can we compute $P(\text{cavity}|\text{toothache})$: By Bayes' rule, $P(\text{cavity}|\text{toothache}) = \frac{P(\text{toothache}|\text{cavity}) \cdot P(\text{cavity})}{P(\text{toothache})} = \frac{0.6 \cdot 0.2}{0.2} = 0.6$.
- **Ok, but:** Why don't we simply assess $P(\text{cavity}|\text{toothache})$ directly?
- **Definition 1.6.3.** We have to take cause and effect into account (cavities cause toothache)
  - $P(\text{toothache}|\text{cavity})$ is causal,
  - $P(\text{cavity}|\text{toothache})$ is diagnostic.
- **Intuition:** Causal dependencies are robust over frequency of the causes.
- **Example 1.6.4.** If there is a cavity epidemic then $P(\text{cavity}|\text{toothache})$ increases, but $P(\text{toothache}|\text{cavity})$ remains the same. (only depends on how cavities "work")
- Also, causal dependencies are often easier to assess.
- **Intuition:** "reason about causes in order to draw conclusions about symptoms".
- Bayes' rule allows to perform diagnosis (observing a symptom, what is the cause?) based on prior probabilities and causal dependencies.

Extended Example: Bayes' Rule and Meningitis

- **Facts known to doctors:**
  - The prior probabilities of meningitis (m) and stiff neck (s) are $P(m) = 0.00002$ and $P(s) = 0.01$.
  - Meningitis causes a stiff neck 70% of the time: $P(s|m) = 0.7$.
- **Doctor d uses Bayes' Rule:**
  $P(m|s) = \frac{P(s|m) \cdot P(m)}{P(s)} = \frac{0.7 \cdot 0.00002}{0.01} = 0.0014 \sim \frac{1}{700}$.
  - Even though stiff neck is strongly indicated by meningitis ($P(s|m) = 0.7$)
  - the probability of meningitis in the patient remains small.
  - The prior probability of stiff necks is much higher than that of meningitis.
- Doctor d' knows $P(m|s)$ from observation; she does not need Bayes' rule!
- Indeed, but what if a meningitis epidemic erupts
- Then d knows that $P(m|s)$ grows proportionally with $P(m)$ (d' clueless)

Bayes Rule for Dogs

- Say $P(\text{dog}) = 0.4$, $P(\text{likeschappi}|\text{dog}) = 0.8$, and $P(\text{likeschappi}) = 0.5$.
- **Question:** What is $P(\text{dog}|\text{likeschappi})$?
  A: 0.8 B: 0.64 C: 0.9 D: 0.32?
- **Answer:** reserved for the plenary sessions $\rightsquigarrow$ be there!
- **Question:** Is $P(\text{dog}|\text{likeschappi})$ causal or diagnostic?
- **Answer:** reserved for the plenary sessions $\rightsquigarrow$ be there!
- **Question:** Is $P(\text{likeschappi}|\text{dog})$ causal or diagnostic?
- **Answer:** reserved for the plenary sessions $\rightsquigarrow$ be there!

1.7 Conditional Independence

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/29054`.

Bayes' Rule with Multiple Evidence

- **Example 1.7.1.** Say we know from medicinical studies that $P(\text{cavity}) = 0.2$, $P(\text{toothache}|\text{cavity}) = 0.6$, $P(\text{toothache}|\neg\text{cavity}) = 0.1$, $P(\text{catch}|\text{cavity}) = 0.9$, and $P(\text{catch}|\neg\text{cavity}) = 0.2$.

  Now, in case we did observe the symptoms toothache and catch (the dentist's probe catches in the aching tooth), what would be the likelihood of having a cavity? What is $P(\text{cavity}|\text{toothache} \wedge \text{catch})$?

  - **Trial 1:** Bayes' rule

    $$P(\text{cavity}|\text{toothache} \wedge \text{catch}) = \frac{P(\text{toothache} \wedge \text{catch}|\text{cavity}) \cdot P(\text{cavity})}{P(\text{toothache} \wedge \text{catch})}$$

  - **Trial 2:** Normalization $\mathbf{P}(X|\mathbf{e}) = \alpha\mathbf{P}(X,\mathbf{e})$ then Product Rule $\mathbf{P}(X,\mathbf{e}) = \mathbf{P}(\mathbf{e}|X) \cdot \mathbf{P}(X)$, with $X = \text{Cavity}$, $\mathbf{e} = \text{toothache} \wedge \text{catch}$:

    $$\begin{aligned}\mathbf{P}(\text{Cavity}|\text{catch} \wedge \text{toothache}) &= \alpha \cdot \mathbf{P}(\text{toothache} \wedge \text{catch}|\text{Cavity}) \cdot \mathbf{P}(\text{Cavity})\\ P(\text{cavity}|\text{catch} \wedge \text{toothache}) &= \alpha \cdot P(\text{toothache} \wedge \text{catch}|\text{cavity}) \cdot P(\text{cavity})\\ P(\neg\text{cavity}|\text{catch} \wedge \text{toothache}) &= \alpha P(\text{toothache} \wedge \text{catch}|\neg\text{cavity}) P(\neg\text{cavity})\end{aligned}$$

Bayes' Rule with Multiple Evidence, ctd.

- $\mathbf{P}(\text{Cavity}|\text{toothache} \wedge \text{catch}) = \alpha\mathbf{P}(\text{toothache} \wedge \text{catch}|\text{Cavity}) \cdot \mathbf{P}(\text{Cavity})$
- **Question:** So, is everything fine?

▷ **Answer:** No! We need $\mathbf{P}(\text{toothache} \wedge \text{catch}|\text{Cavity})$, i.e. causal dependencies for all combinations of symptoms! ($\gg 2$, in general)

▷ **Question:** Are Toothache and Catch independent?

▷ **Answer:** No. If a probe catches, we probably have a cavity which probably causes toothache.

▷ **But:** They are conditionally independent given the presence or absence of a cavity!

Conditional Independence

▷ **Definition 1.7.2.** Given sets of random variables $\mathbb{Z}_1$, $\mathbb{Z}_2$, and $\mathbf{Z}$, we say that $\mathbb{Z}_1$ and $\mathbb{Z}_2$ are conditionally independent given $\mathbf{Z}$ if:

$$\mathbf{P}(\mathbb{Z}_1, \mathbb{Z}_2|\mathbf{Z}) = \mathbf{P}(\mathbb{Z}_1|\mathbf{Z}) \cdot \mathbf{P}(\mathbb{Z}_2|\mathbf{Z})$$

We alternatively say that $\mathbb{Z}_1$ is conditionally independent of $\mathbb{Z}_2$ given $\mathbf{Z}$.

▷ **Example 1.7.3.** Catch and Toothache are conditionally independent given Cavity.

- ▷ For cavity: this may cause both, but they don't influence each other.
- ▷ For $\neg$cavity: something else causes catch and/or toothache.

So we have:

$$\mathbf{P}(\text{Toothache}, \text{Catch}|\text{cavity}) = \mathbf{P}(\text{Toothache}|\text{cavity}) \cdot \mathbf{P}(\text{Catch}|\text{cavity})$$
$$\mathbf{P}(\text{Toothache}, \text{Catch}|\neg\text{cavity}) = \mathbf{P}(\text{Toothache}|\neg\text{cavity}) \cdot \mathbf{P}(\text{Catch}|\neg\text{cavity})$$

▷ **Note:** The definition is symmetric regarding the roles of $\mathbb{Z}_1$ and $\mathbb{Z}_2$: Toothache is conditionally independent of Cavity.

▷ But there may be dependencies *within* $\mathbb{Z}_1$ or $\mathbb{Z}_2$, e.g. $\mathbb{Z}_2 = \{\text{Toothache}, \text{Sleeplessness}\}$.

Conditional Independence, ctd.

▷ If $\mathbb{Z}_1$ and $\mathbb{Z}_2$ are conditionally independent given $\mathbf{Z}$, then $\mathbf{P}(\mathbb{Z}_1|\mathbb{Z}_2, \mathbf{Z}) = \mathbf{P}(\mathbb{Z}_1|\mathbf{Z})$.

▷ *Proof:*

1. By definition, $\mathbf{P}(\mathbb{Z}_1|\mathbb{Z}_2, \mathbf{Z}) = \frac{\mathbf{P}(\mathbb{Z}_1, \mathbb{Z}_2, \mathbf{Z})}{\mathbf{P}(\mathbb{Z}_2, \mathbf{Z})}$
2. which by product rule is equal to $\frac{\mathbf{P}(\mathbb{Z}_1, \mathbb{Z}_2|\mathbf{Z}) \cdot \mathbf{P}(\mathbf{Z})}{\mathbf{P}(\mathbb{Z}_2, \mathbf{Z})}$
3. which by conditional independence is equal to $\frac{\mathbf{P}(\mathbb{Z}_1|\mathbf{Z}) \cdot \mathbf{P}(\mathbb{Z}_2|\mathbf{Z}) \cdot \mathbf{P}(\mathbf{Z})}{\mathbf{P}(\mathbb{Z}_2, \mathbf{Z})}$.
4. Since $\frac{\mathbf{P}(\mathbb{Z}_2|\mathbf{Z}) \cdot \mathbf{P}(\mathbf{Z})}{\mathbf{P}(\mathbb{Z}_2, \mathbf{Z})} = 1$ this proves the claim.

▷ **Example 1.7.4.** Using $\{\text{Toothache}\}$ as $\mathbf{Z_1}$, $\{\text{Catch}\}$ as $\mathbf{Z_2}$, and $\{\text{Cavity}\}$ as $\mathbf{Z}$: $\mathbf{P}(\text{Toothache}|\text{Catch}, \text{Cavity}) = \mathbf{P}(\text{Toothache}|\text{Cavity})$.

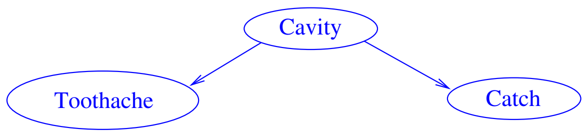
▷ In the presence of conditional independence, we can drop variables from the right-hand side of conditional probabilities.

▷ Third of the four basic techniques in Bayesian networks.

▷ **Last missing technique:** "Capture variable dependencies in a graph"; illustration see next slide, details see chapter 2

Exploiting Conditional Independence: Overview

▷ **1. Graph captures variable dependencies:** (Variables $X_1, \ldots, X_n$)



- ▷ Given evidence $\mathbf{e}$, want to know $\mathbf{P}(X|e)$.
- ▷ Remaining vars: $\mathbf{Y}$.

▷ **2. Normalization+Marginalization:**

$$\mathbf{P}(X|\mathbf{e}) = \alpha \cdot \mathbf{P}(X, \mathbf{e}); \text{ if } \mathbf{Y} \neq \emptyset \text{ then } \mathbf{P}(X|\mathbf{e}) = \alpha \cdot (\sum_{\mathbf{y} \in \mathbf{Y}} \mathbf{P}(X, \mathbf{e}, \mathbf{y}))$$

- ▷ A sum over atomic events!

▷ **3. Chain rule:** Order $X_1, \ldots, X_n$ consistently with dependency graph.

$$\mathbf{P}(X_1, \ldots, X_n) = \mathbf{P}(X_n|X_{n-1}, \ldots, X_1) \cdot \mathbf{P}(X_{n-1}|X_{n-2}, \ldots, X_1) \cdot \ldots \cdot \mathbf{P}(X_1)$$

▷ **4. Exploit Conditional Independence:** Instead of $\mathbf{P}(X_i|X_{i-1}, \ldots, X_1)$, with previous slide we can use $\mathbf{P}(X_i|\text{Parents}(X_i))$.

- ▷ Bayesian networks!

Exploiting Conditional Independence: Example

▷ **1. Graph captures variable dependencies:** (See previous slide.)

- ▷ Given toothache, catch, want $\mathbf{P}(\text{Cavity}|\text{toothache}, \text{catch})$. Remaining vars: $\emptyset$.

▷ **2. Normalization+Marginalization:**

$$\mathbf{P}(\text{Cavity}|\text{toothache}, \text{catch}) = \alpha \cdot \mathbf{P}(\text{Cavity}, \text{toothache}, \text{catch})$$

▷ **3. Chain rule:**
Order $X_1 =$ Cavity, $X_2 =$ Toothache, $X_3 =$ Catch.

$$\mathbf{P}(\text{Cavity}, \text{toothache}, \text{catch}) = \\ \mathbf{P}(\text{catch}|\text{toothache}, \text{Cavity}) \cdot \mathbf{P}(\text{toothache}|\text{Cavity}) \cdot \mathbf{P}(\text{Cavity})$$

▷ **4. Exploit Conditional independence:**

Instead of $\mathbf{P}(\text{catch}|\text{toothache}, \text{Cavity})$ use $\mathbf{P}(\text{catch}|\text{Cavity})$.

▷ **Thus:**

$$\begin{aligned} &\mathbf{P}(\text{Cavity}|\text{toothache}, \text{catch}) \\ &= \alpha \cdot \mathbf{P}(\text{catch}|\text{Cavity}) \cdot \mathbf{P}(\text{toothache}|\text{Cavity}) \cdot \mathbf{P}(\text{Cavity}) \\ &= \alpha \cdot \langle 0.9 \cdot 0.6 \cdot 0.2, 0.2 \cdot 0.1 \cdot 0.8 \rangle \\ &= \alpha \cdot \langle 0.108, 0.016 \rangle \end{aligned}$$

▷ **So:** $\alpha \approx 8.06$ and $\mathbf{P}(\text{cavity}|\text{toothache} \wedge \text{catch}) \approx 0.87$.

Naive Bayes Models

▷ **Definition 1.7.5.** A Bayesian network in which a single cause directly influences a number of effects, all of which are conditionally independent, given the cause is called a naive Bayes model or Bayesian classifier.

▷ **Observation 1.7.6.** *In a naive Bayes model, the full joint probability distribution can be written as*

$$\mathbf{P}(cause|effect_1, \ldots, effect_n) = \alpha\langle effect_1, \ldots, effect_n\rangle \cdot \mathbf{P}(cause) \cdot \prod_i \mathbf{P}(effect_i|cause)$$

▷ **Note:** This kind of model is called "naive" since it is often used as a simplifying model if the effects are not conditionally independent after all.

▷ It is also called idiot Bayes model by Bayesian fundamentalists.

▷ In practice, naive Bayes models can work surprisingly well, even when the conditional independence assumption is not true.

▷ **Example 1.7.7.** The dentistry example is a (true) naive Bayes model.

Questionnaire

▷ Consider the random variables $X_1 =$ Animal, $X_2 =$ LikesChappi, and $X_3 =$ LoudNoise, and X_1 has values $\{\text{dog}, \text{cat}, \text{other}\}$, X_2 and X_3 are Boolean.

▷ **Question:** Which statements are correct?

(A) Animal is independent of LikesChappi.

(B) LoudNoise is independent of LikesChappi.

(C) Animal is conditionally independent of LikesChappi given LoudNoise.

(D) LikesChappi is conditionally independent of LoudNoise given Animal.

Think about this intuitively: Given both values for variable X, are the chances of Y being true higher for one of these (fixing value of the third variable where specified)?

▷ **Answer:** reserved for the plenary sessions $\leadsto$ be there!

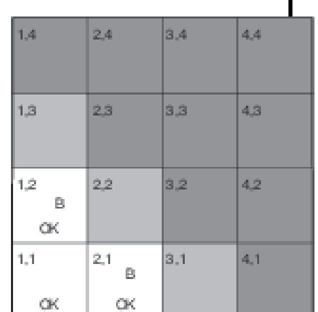
1.8 The Wumpus World Revisited

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/29055`. We will fortify our intuition about naive Bayes models with a variant of the Wumpus world we looked at Example 1.1.17 to understand whether logic was up to the job of guiding an agent in the Wumpus cave.

Wumpus World Revisited

▷ **Example 1.8.1 (The Wumpus is Back).**

- ▷ We have a maze where
  - ▷ pits cause a breeze in neighboring cells
  - ▷ Every cell except $[1,1]$ has a 20% pit probability. (unfair otherwise)
  - ▷ we forget the wumpus and the gold for now (simpler)
- ▷ Where does the agent should go, if there is breeze at $[1,2]$ and $[2,1]$?
- ▷ Pure logical inference can conclude nothing about which square is most likely to be safe!

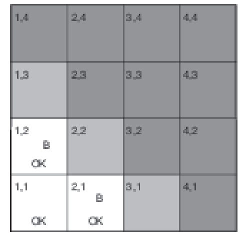


▷ **Idea:** Let's evaluate our probabilistic reasoning machinery, if that can help!

Wumpus: Probabilistic Model

▷ Boolean random variables (only for the observed squares)

- $P_{i,j}$: pit at square $[i, j]$
- $B_{i,j}$: breeze at square $[i, j]$



▷ Full joint probability distribution

1. $\mathbf{P}(P_{1,1}, \ldots, P_{4,4}, B_{1,1}, B_{1,2}, B_{2,1}) = \mathbf{P}(B_{1,1}, B_{1,2}, B_{2,1} | P_{1,1}, \ldots, P_{4,4}) \cdot \mathbf{P}(P_{1,1}, \ldots, P_{4,4})$ (Product Rule)
2. $\mathbf{P}(P_{1,1}, \ldots P_{4,4}) = \prod_{i,j=1,1}^{4,4} \mathbf{P}(P_{i,j})$ (pits are spread independently)
3. For a particular configuration $p_{1,1}, \ldots, p_{4,4}$ with $p_{i,j} \in \{\mathsf{T}, \mathsf{F}\}$, n pits, and $P(p_{i,j}) = 0.2$ we have $P(p_{1,1}, \ldots, p_{4,4}) = 0.2^n \cdot 0.8^{16-n}$

Wumpus: Query and Simple Reasoning

▷ We have evidence in our example:

- $b = \neg b_{1,1} \wedge b_{1,2} \wedge b_{2,1}$ and
- $\kappa = \neg p_{1,1} \wedge \neg p_{1,2} \wedge \neg p_{2,1}$

We are interested in answering queries such as $P(P_{1,3} | \kappa, b)$. (pit in $(1, 3)$ given evidence)

▷ **Observation:** The answer can be computed by enumeration of the full joint probability distribution.

▷ **Standard Approach:** Let U be the variables $P_{i,j}$ except $P_{1,3}$ and κ, then

$$P(P_{1,3} | \kappa, b) = \sum_{u \in U} \mathbf{P}(P_{1,3}, u, \kappa, b)$$

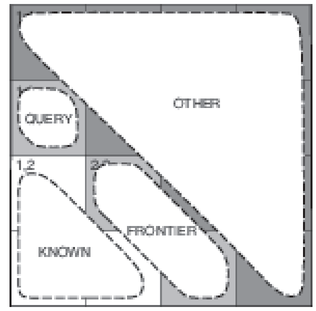
▷ **Problem:** Need to explore all possible values of variables in U ($2^{12} = 4096$ terms!)

▷ Can we do better? (faster; with less computation)

Wumpus: Conditional Independence

▷ **Observation 1.8.2.**

The observed breezes are conditionally independent of the other variables given the known, frontier, and query variables.



- ▷ We split the set of hidden variables into fringe and other variables: $U = F \cup O$ where F is the fringe and O the rest.
- ▷ **Corollary 1.8.3.** $P(b|P_{1,3},\kappa,U) = P(b|P_{1,3},\kappa,F)$ (by conditional independence)
- ▷ **Now:** let us exploit this formula.

Wumpus: Reasoning

- ▷ We calculate:

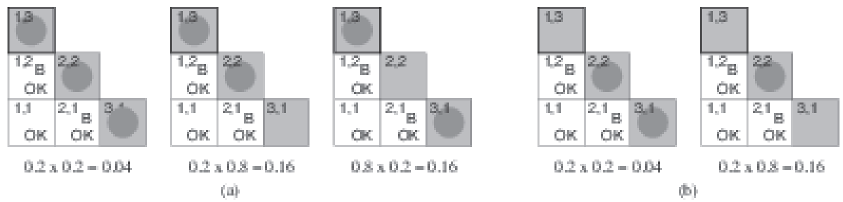
$$
\begin{aligned}
P(P_{1,3}|\kappa,b) &= \alpha(\sum_{u\in U}\mathbf{P}(P_{1,3},u,\kappa,b))\\
&= \alpha(\sum_{u\in U}\mathbf{P}(b|P_{1,3},\kappa,u)\cdot\mathbf{P}(P_{1,3},\kappa,u))\\
&= \alpha(\sum_{f\in F}\sum_{o\in O}\mathbf{P}(b|P_{1,3},\kappa,f,o)\cdot\mathbf{P}(P_{1,3},\kappa,f,o))\\
&= \alpha(\sum_{f\in F}\mathbf{P}(b|P_{1,3},\kappa,f)\cdot(\sum_{o\in O}\mathbf{P}(P_{1,3},\kappa,f,o)))\\
&= \alpha(\sum_{f\in F}\mathbf{P}(b|P_{1,3},\kappa,f)\cdot(\sum_{o\in O}\mathbf{P}(P_{1,3})\cdot P(\kappa)\cdot P(f)\cdot P(o)))\\
&= \alpha\mathbf{P}(P_{1,3})P(\kappa)(\sum_{f\in F}\mathbf{P}(b|P_{1,3},\kappa,f)\cdot P(f)\cdot(\sum_{o\in O}P(o)))\\
&= \alpha' P(P_{1,3})(\sum_{f\in F}\mathbf{P}(b|P_{1,3},\kappa,f)\cdot P(f))
\end{aligned}
$$

for $\alpha' {:=} \alpha P(\kappa)$ as $\sum_{o\in O}P(o)=1$.

Wumpus: Solution

- ▷ We calculate using the product rule and conditional independence (see above) $P(P_{1,3}|\kappa,b) = \alpha'\cdot P(P_{1,3})\cdot(\sum_{f\in F}\mathbf{P}(b|P_{1,3},\kappa,f)\cdot P(f))$
- ▷ Let us explore possible models (values) of Fringe that are F compatible with ob-

servation b.



▷ $\mathbf{P}(P_{1,3}|\kappa, b) = \alpha' \cdot \langle 0.2 \cdot (0.04 + 0.16 + 0.16), 0.8 \cdot (0.04 + 0.16)\rangle = \langle 0.31, 0.69\rangle$

▷ $\mathbf{P}(P_{3,1}|\kappa, b) = \langle 0.31, 0.69\rangle$ by symmetry

▷ $\mathbf{P}(P_{2,2}|\kappa, b) = \langle 0.86, 0.14\rangle$ (definitely avoid)

1.9 Conclusion

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/29056`.

Summary

▷ Uncertainty is unavoidable in many environments, namely whenever agents do not have perfect knowledge.

▷ Probabilities express the degree of belief of an agent, given its knowledge, into an event.

▷ Conditional probabilities express the likelihood of an event given observed evidence.

▷ Assessing a probability $\widehat{=}$ use statistics to approximate the likelihood of an event.

▷ Bayes' rule allows us to derive, from probabilities that are easy to assess, probabilities that aren't easy to assess.

▷ Given multiple evidence, we can exploit conditional independence.

▷ Bayesian networks (up next) do this, in a comprehensive, computational manner.

Reading: *Chapter 13: Quantifying Uncertainty* [RN03].

Content: Sections 13.1 and 13.2 roughly correspond to my "Introduction" and "Probability Theory Concepts". Section 13.3 and 13.4 roughly correspond to my "Basic Probabilistic Inference". Section 13.5 roughly corresponds to my "Bayes' Rule" and "Multiple Evidence".

In Section 13.6, RN go back to the Wumpus world and discuss some inferences in a probabilistic version thereof.

Overall, the content is quite similar. I have added some examples, have tried to make a few subtle points more explicit, and I indicate already how these techniques will be used in Bayesian networks. RN gives many complementary explanations, nice as additional background reading.

Chapter 2

Probabilistic Reasoning: Bayesian Networks

2.1 Introduction

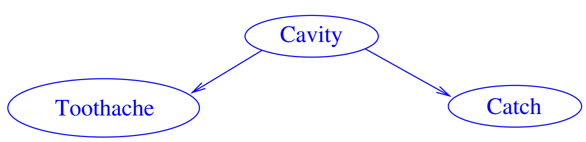
A Video Nugget covering this section can be found at `https://fau.tv/clip/id/29218`.

Reminder: Our Agenda for This Topic

- Our treatment of the topic "probabilistic reasoning" consists of this and last section.
  - chapter 1: All the basic machinery at use in Bayesian networks.
  - **This section**: Bayesian networks: What they are, how to build them, how to use them.
  - The most wide-spread and successful practical framework for probabilistic reasoning.

Reminder: Our Machinery

1. **Graph captures variable dependencies:** (Variables $X_1, \ldots, X_n$)



   - Given evidence $\mathbf{e}$, want to know $\mathbf{P}(X|e)$. Remaining vars: $\mathbf{Y}$.

2. **Normalization+Marginalization:**

$$\mathbf{P}(X|\mathbf{e}) = \alpha\mathbf{P}(X,\mathbf{e}) = \alpha \sum_{\mathbf{y}\in\mathbf{Y}} \mathbf{P}(X,\mathbf{e},\mathbf{y})$$

   - A sum over atomic events!

3. **Chain rule:** $X_1, \ldots, X_n$ consistently with dependency graph.

$$\mathbf{P}(X_1,\ldots,X_n) = \mathbf{P}(X_n|X_{n-1},\ldots,X_1) \cdot \mathbf{P}(X_{n-1}|X_{n-2},\ldots,X_1) \cdot \ldots \cdot \mathbf{P}(X_1)$$

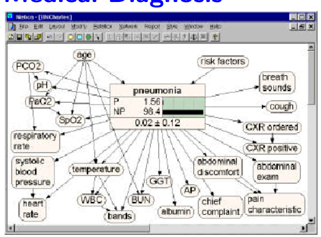
4. **Exploit conditional independence:** Instead of $\mathbf{P}(X_i|X_{i-1},\ldots,X_1)$, we can use $\mathbf{P}(X_i|\text{Parents}(X_i))$.

   ▷ Bayesian networks!

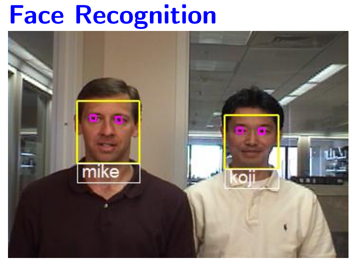
Some Applications

▷ A ubiquitous problem: Observe "symptoms", need to infer "causes".

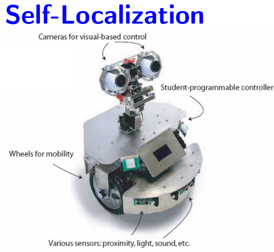
Medical Diagnosis



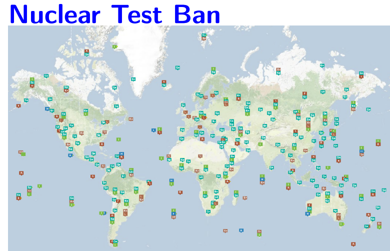
Face Recognition



Self-Localization



Nuclear Test Ban



Our Agenda for This Chapter

▷ **What is a Bayesian Network?:** i.e. What is the syntax?

  ▷ Tells you what Bayesian networks look like.

▷ **What is the Meaning of a Bayesian Network?:** What is the semantics?

  ▷ Makes the intuitive meaning precise.

▷ **Constructing Bayesian Networks:** How do we design these networks? What effect do our choices have on their size?

  ▷ Before you can start doing inference, you need to model your domain.

▷ **Inference in Bayesian Networks:** How do we use these networks? What is the associated complexity?

▷ Inference is our primary purpose. It is important to understand its complexities and how it can be improved.

2.2 What is a Bayesian Network?

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/29221`.

What is a Bayesian Network? (Short: BN)

▷ What do the others say?

- ▷ *"A Bayesian network is a methodology for representing the full joint probability distribution. In some cases, that representation is compact."*
- ▷ *"A Bayesian network is a graph whose nodes are random variables X_i and whose edges $\langle X_j, X_i\rangle$ denote a direct influence of X_j on X_i. Each node X_i is associated with a conditional probability table (CPT), specifying $\mathbf{P}(X_i|Parents(X_i))$."*
- ▷ *"A Bayesian network is a graphical way to depict conditional independence relations within a set of random variables."*

▷ A Bayesian network (BN) represents the structure of a given domain. Probabilistic inference exploits that structure for improved efficiency.

▷ BN inference: Determine the distribution of a query variable X given observed evidence $\mathbf{e}$: $\mathbf{P}(X|\mathbf{e})$.

John, Mary, and My Brand-New Alarm

▷ **Example 2.2.1 (From Russell/Norvig).**

- ▷ I got very valuable stuff at home. So I bought an alarm. Unfortunately, the alarm just rings at home, doesn't call me on my mobile.
- ▷ I've got two neighbors, Mary and John, who'll call me if they hear the alarm.
- ▷ The problem is that, sometimes, the alarm is caused by an earthquake.
- ▷ Also, John might confuse the alarm with his telephone, and Mary might miss the alarm altogether because she typically listens to loud music.

▷ **Question:** Given that both John and Mary call me, what is the probability of a burglary?

John, Mary, and My Alarm: Designing the Network

▷ **Cooking Recipe:**

(1) Design the random variables $X_1, \ldots, X_n$;

(2) Identify their dependencies;

(3) Insert the conditional probability tables $\mathbf{P}(X_i|\text{Parents}(X_i))$.

▷ **Example 2.2.2 (Let's cook!).** Using this recipe on Example 2.2.1, ...

(1) Random variables: Burglary, Earthquake, Alarm, JohnCalls, MaryCalls.

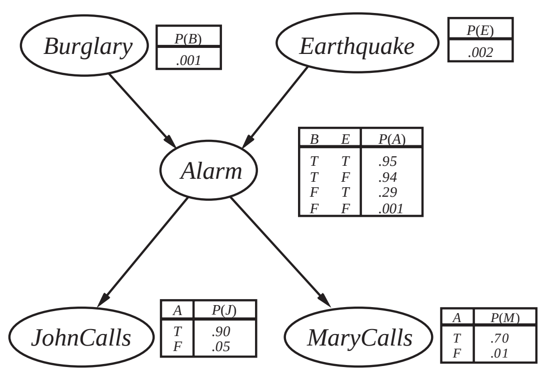
(2) **Dependencies**: Burglaries and earthquakes are independent. (this is actually debatable ⇝ design decision!)

The alarm might be activated by either. John and Mary call if and only if they hear the alarm. (they don't care about earthquakes)

(3) **Conditional probability tables**: Assess the probabilities, see next slide.

John, Mary, and My Alarm: The Bayesian network

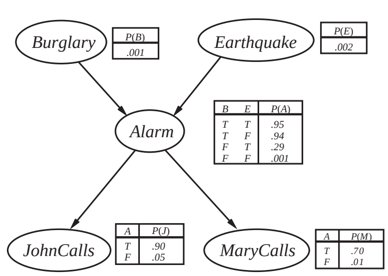
▷ **Example 2.2.3.** Continuing Example 2.2.2 we obtain



▷ **Note:** In each $\mathbf{P}(X_i|\text{Parents}(X_i))$, we show only $\mathbf{P}(X_i = \mathsf{T}|\text{Parents}(X_i))$. We don't show $\mathbf{P}(X_i = \mathsf{F}|\text{Parents}(X_i))$ which is $1 - \mathbf{P}(X_i = \mathsf{T}|\text{Parents}(X_i))$.

The Syntax of Bayesian Networks

▷ To fix the exact definition of Bayesian networks recall the **??**:



▷ **Definition 2.2.4 (Bayesian Network).** Given random variables $X_1, \ldots, X_n$ with finite domains $D_1, \ldots, D_n$, a Bayesian network (also belief network or probabilistic network) is a node labeled DAG $\mathcal{B} := \langle \{X_1, \ldots, X_n\}, E, \mathsf{CPT} \rangle$.

Each X_i is labeled with a function

$$\mathsf{CPT}(X_i) \colon D_i \times \prod_{X_j \in \mathsf{Parents}(X_i)} D_j \to [0,1]$$

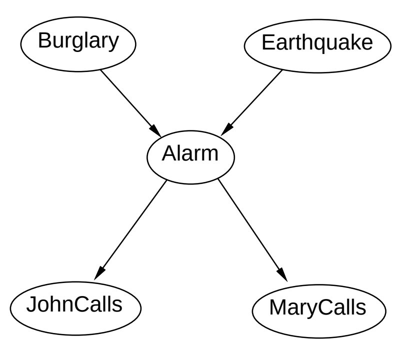
where $\mathsf{Parents}(X_i) := \{X_j | (X_j, X_i) \in E\}$ it is called the conditional probability table at X_i.

▷ **Definition 2.2.5.** Bayesian networks and related formalisms summed up under the term graphical models.

2.3 What is the Meaning of a Bayesian Network?

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/29223`.

The Semantics of Bayesian Networks: Illustration



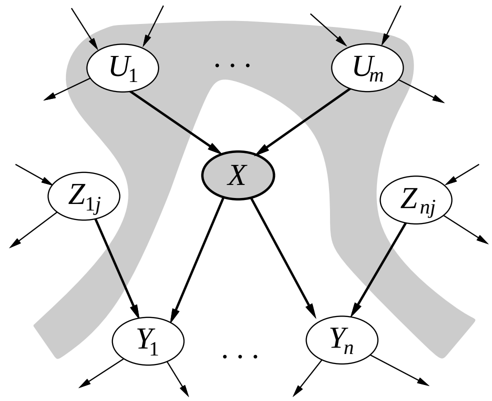
▷ Alarm depends on Burglary and Earthquake.

▷ MaryCalls only depends on Alarm. $\mathbf{P}(\mathsf{MaryCalls}|\mathsf{Alarm}, \mathsf{Burglary}) = \mathbf{P}(\mathsf{MaryCalls}|\mathsf{Alarm})$

▷ Bayesian networks represent sets of conditional independence assumptions.

The Semantics of Bayesian Networks: Illustration, ctd.

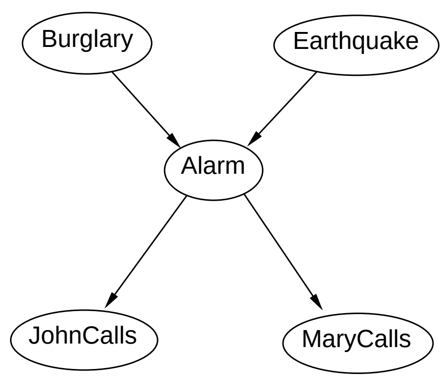
▷ **Observation 2.3.1.** *Each node X in a BN is conditionally independent of its non-descendants given its parents $\mathsf{Parents}(X)$.*



▷ **Question:** Why non-descendants of X?

▷ **Intuition:** Given that BNs are acyclic, these are exactly those nodes that *could* have an edge into X.

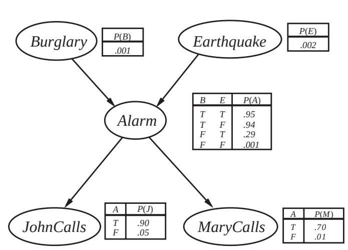
The Semantics of BNs



▷ **Question:** Given the value of Alarm, MaryCalls is independent of?

▷ **Answer:** reserved for the plenary sessions ⇝ be there!

The Semantics of Bayesian Networks: Formal



▷ **Definition 2.3.2.** Let $\langle\mathcal{X},E\rangle$ be a Bayesian network, $X\in\mathcal{X}$, and E^* the transitive reflexive closure of E, then $\mathrm{NonDesc}(X){:=}\{Y|(X,Y)\notin E^*\}\backslash\mathrm{Parents}(X)$ is the set of non-descendents of X.

▷ **Definition 2.3.3.** Given a Bayesian network $\mathcal{B}{:=}\langle\mathcal{X},E\rangle$, we identify $\mathcal{B}$ with the following two assumptions:

(A) $X\in\mathcal{X}$ is conditionally independent of $\mathrm{NonDesc}(X)$ given $\mathrm{Parents}(X)$.

(B) For all values x of $X\in\mathcal{X}$, and all value combinations of $\mathrm{Parents}(X)$, we have $P(x|\mathrm{Parents}(X)) = \mathrm{CPT}(x,\mathrm{Parents}(X))$.

Recovering the Full Joint Probability Distribution

▷ **Intuition:** A Bayesian network is a methodology for representing the full joint probability distribution.

▷ **Problem:** How to recover the full joint probability distribution $\mathbf{P}(X_1,\ldots,X_n)$ from $\mathcal{B}{:=}\langle\{X_1,\ldots,X_n\},E\rangle$?

▷ **Chain Rule:** For *any* ordering $X_1,\ldots,X_n$, we have:

$$\mathbf{P}(X_1,\ldots,X_n) = \mathbf{P}(X_n|X_{n-1},\ldots,X_1)\cdot\mathbf{P}(X_{n-1}|X_{n-2},\ldots,X_1)\cdot\ldots\cdot\mathbf{P}(X_1)$$

Choose $X_1,\ldots,X_n$ consistent with $\mathcal{B}$: $X_j\in\mathrm{Parents}(X_i) \leadsto j<i$.

▷ **Observation 2.3.4 (Exploiting Conditional Independence).** *With Definition 2.3.3 (A), we can use* $\mathbf{P}(X_i|\mathit{Parents}(X_i))$ *instead of* $\mathbf{P}(X_i|X_{i-1},\ldots,X_1)$*:*

$$\mathbf{P}(X_1,\ldots,X_n) = \prod_{i=1}^{n}\mathbf{P}(X_i|\mathit{Parents}(X_i))$$

The distributions $\mathbf{P}(X_i|\mathit{Parents}(X_i))$ *are given by Definition 2.3.3 (B).*

▷ Same for atomic events $P(X_1,\ldots,X_n)$.

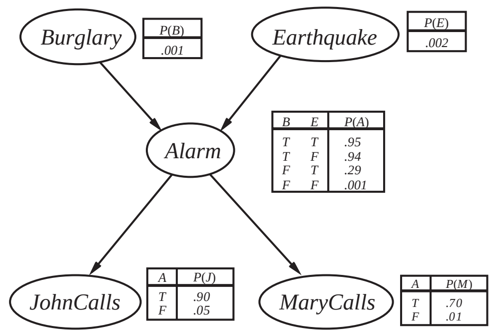
▷ **Observation 2.3.5 (Why "acyclic"?).** *For cyclic $\mathcal{B}$, this does NOT hold, indeed cyclic BNs may be self contradictory. (need a consistent ordering)*

Note: If there is a cycle, then any ordering $X_1,\ldots,X_n$ will not be consistent with the BN; so in the chain rule on $X_1,\ldots,X_n$ there comes a point where we have $\mathbf{P}(X_i|X_{i-1},\ldots,X_1)$ in the chain but $\mathbf{P}(X_i|\mathrm{Parents}(X_i))$ in the definition of distribution, and $\mathrm{Parents}(X_i)\not\subseteq\{X_{i-1},\ldots,X_1\}$ but then the products are different. So the chain rule can no longer be used to prove that we can reconstruct the full joint probability distribution. In fact, cyclic Bayesian network contain ambiguities (several interpretations possible) and may be self-contradictory (no probability distribution matches the Bayesian network).

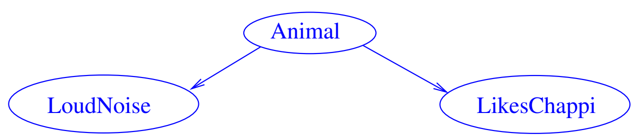
Recovering a Probability for John, Mary, and the Alarm

▷ **Example 2.3.6.** John and Mary called because there was an alarm, but no earthquake or burglary

$$\begin{aligned} P(j,m,a,\neg b,\neg e) &= P(j|a)\cdot P(m|a)\cdot P(a|\neg b,\neg e)\cdot P(\neg b)\cdot P(\neg e) \\ &= 0.9*0.7*0.001*0.999*0.998 \\ &= 0.00062 \end{aligned}$$



Meaning of Bayesian Networks

▷ 

Say $\mathcal{B}$ is the Bayesian network above. Which statements are correct?

(A) Animal is independent of LikesChappi.

(B) LoudNoise is independent of LikesChappi.

(C) Animal is conditionally independent of LikesChappi given LoudNoise.

(D) LikesChappi is conditionally independent of LoudNoise given Animal.

Think about this intuitively: Given both values for variable X, is the chances of Y being true higher for one of these (fixing value of third var where specified)?

▷ **Answers:** reserved for the plenary sessions ⇝ be there!

2.4 Constructing Bayesian Networks

Video Nuggets covering this section can be found at `https://fau.tv/clip/id/29224` and `https://fau.tv/clip/id/29226`.

Constructing Bayesian Networks

▷ **BN construction algorithm:**

1. Initialize $BN{:=}\langle\{X_1,\ldots,X_n\},E\rangle$ where $E=\emptyset$.

2. Fix any order of the variables, $X_1, \ldots, X_n$.
3. **for** $i := 1, \ldots, n$ **do**
   a. Choose a minimal set Parents$(X_i) \subseteq \{X_1, \ldots, X_{i-1}\}$ so that

   $$\mathbf{P}(X_i|X_{i-1}, \ldots, X_1) = \mathbf{P}(X_i|\text{Parents}(X_i))$$

   b. For each $X_j \in \text{Parents}(X_i)$, insert (X_j, X_i) into E.
   c. Associate X_i with CPT(X_i) corresponding to $\mathbf{P}(X_i|\text{Parents}(X_i))$.

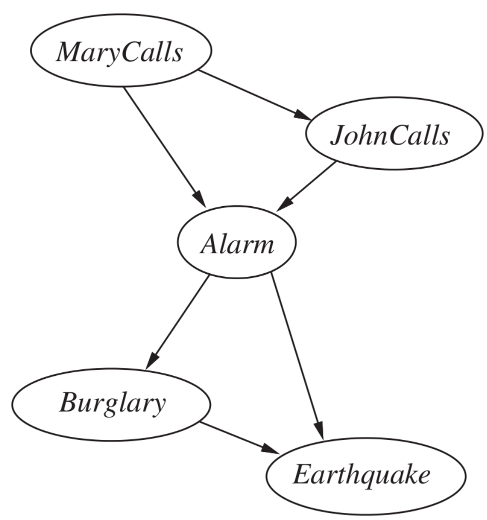
▷ **Attention:** *Which* variables we need to include into Parents(X_i) depends on what "$\{X_1, \ldots, X_{i-1}\}$" is . . . !

▷ The size of the resulting BN depends on the chosen order $X_1, \ldots, X_n$.

▷ The size of a Bayesian network is *not* a fixed property of the domain. It depends on the skill of the designer.

John and Mary Depend on the Variable Order!

▷ **Example 2.4.1.** MaryCalls, JohnCalls, Alarm, Burglary, Earthquake.



Note: For **??** we try to determine whether – given different value assignments to potential parents – the probability of X_i being true differs? If yes, we include these parents. In the particular case:

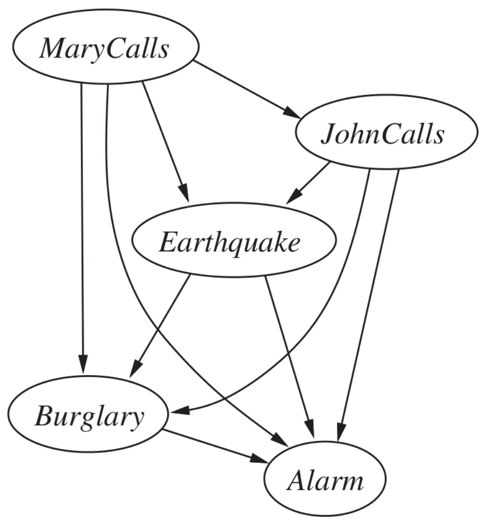
1. M to J yes because the common cause may be the alarm.
2. M, J to A yes because they may have heard alarm.
3. A to B yes because if A then higher chance of B.
4. However, M/J to B no because M/J only react to the alarm so if we have the value of A then values of M/J don't provide more information about B.

5. A to E yes because if A then higher chance of E.

6. B to E yes because, if A and not B then chances of E are higher than if A and B.

John and Mary Depend on the Variable Order! Ctd.

▷ **Example 2.4.2.** MaryCalls, JohnCalls, Earthquake, Burglary, Alarm.



Again: Given different value assignments to potential parents, does the probability of X_i being true differ? If yes, include these parents.

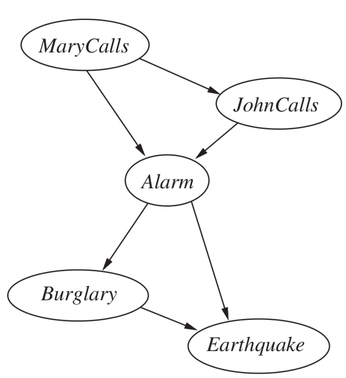
1. M to J as before.

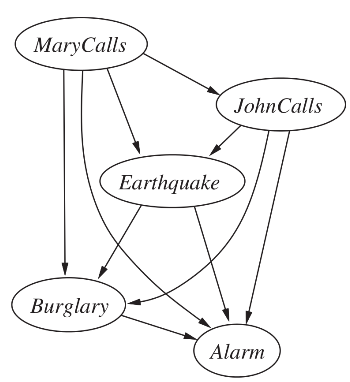
2. M, J to E as probability of E is higher if M/J is true.

3. Same for B; E to B because, given M and J are true, if E is true as well then prob of B is lower than if E is false.

4. $M/J/B/E$ to A because if $M/J/B/E$ is true (even when changing the value of just one of these) then probability of A is higher.

John and Mary, What Went Wrong?





- **Intuition:** These BNs link from symptoms to causes! ($\mathbf{P}(\text{Cavity}|\text{Toothache})$) Even though M and J are conditionally independent given A, they are *not* independent without any additional evidence; thus we don't "see" their conditional independence unless we ordered A before M and J! $\rightsquigarrow$ We organized the domain in the wrong way here.

  We fail to identify many conditional independence relations (e.g., get dependencies between conditionally independent symptoms).

- **Also recall:** Conditional probabilities $\mathbf{P}(\text{Symptom}|\text{Cause})$ are more robust and often easier to assess than $\mathbf{P}(\text{Cause}|\text{Symptom})$.

- **Rule of Thumb:** We should order causes before symptoms.

Compactness of Bayesian Networks

- **Definition 2.4.3.** Given random variables $X_1, \ldots, X_n$ with finite domains $D_1, \ldots, D_n$ the size of $\mathcal{B}{:=}\langle\{X_1, \ldots, X_n\}, E\rangle$ is defined as

$$\text{size}(\mathcal{B}){:=}\sum_{i=1}^{n} \#(D_i) \cdot \prod_{X_j \in \text{Parents}(X_i)} \#(D_j)$$

- **Note:** $\text{size}(\mathcal{B}) \mathrel{\hat{=}}$ The total number of entries in the CPTs.

- **Note:** Smaller BN $\rightsquigarrow$ need to assess less probabilities, more efficient inference.

- **Observation 2.4.4.** *Explicit full joint probability distribution has size* $\prod_{i=1}^{n} \#(D_i)$.

- **Observation 2.4.5.** *If* $\#(\text{Parents}(X_i)) \leq k$ *for every* X_i*, and* $D_{\max}$ *is the largest random variable domain, then* $\text{size}(\mathcal{B}) \leq n\#(D_{\max})^{k+1}$.

- **Example 2.4.6.** For $\#(D_{\max}) = 2$, $n = 20$, $k = 4$ we have $2^{20} = 1048576$ probabilities, but a Bayesian network of size $\leq 20 \cdot 2^5 = 640 \ldots$!

- In the *worst case*, $\text{size}(\mathcal{B}) = n \cdot \prod_{i=1}^{n} \#(D_i)$, namely if every variable depends on all its predecessors in the chosen order.

- **Intuition:** BNs are compact if each variable is directly influenced only by few of its predecessor variables.

Constructing Bayesian Networks

- **Question:** What is the Bayesian network we get by constructing according to the ordering

1. $X_1 = \text{LoudNoise}, X_2 = \text{Animal}, X_3 = \text{LikesChappi}$?
2. $X_1 = \text{LoudNoise}, X_2 = \text{LikesChappi}, X_3 = \text{Animal}$?

▷ **Answer:** reserved for the plenary sessions ⇝ be there!

2.5 Constructing Bayesian Networks

Video Nuggets covering this section can be found at https://fau.tv/clip/id/29224 and https://fau.tv/clip/id/29226.

Constructing Bayesian Networks

▷ **BN construction algorithm:**

1. Initialize $BN{:=}\langle\{X_1,\ldots,X_n\},E\rangle$ where $E=\emptyset$.
2. Fix any order of the variables, $X_1,\ldots,X_n$.
3. **for** $i:=1,\ldots,n$ **do**
   a. Choose a minimal set $\text{Parents}(X_i)\subseteq\{X_1,\ldots,X_{i-1}\}$ so that
   $$\mathbf{P}(X_i|X_{i-1},\ldots,X_1)=\mathbf{P}(X_i|\text{Parents}(X_i))$$
   b. For each $X_j{\in}\text{Parents}(X_i)$, insert (X_j,X_i) into E.
   c. Associate X_i with $\text{CPT}(X_i)$ corresponding to $\mathbf{P}(X_i|\text{Parents}(X_i))$.

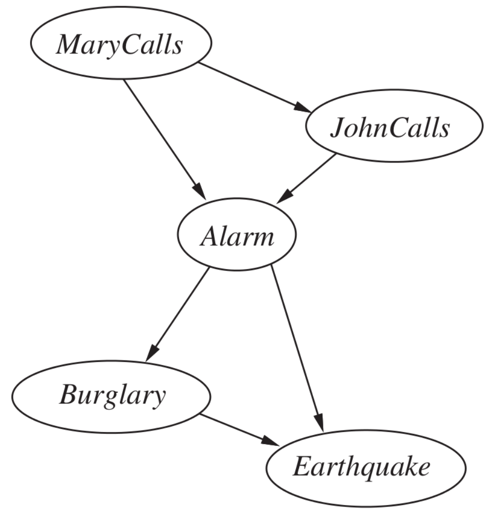
▷ **Attention:** *Which* variables we need to include into $\text{Parents}(X_i)$ depends on what "$\{X_1,\ldots,X_{i-1}\}$" is . . . !

▷ The size of the resulting BN depends on the chosen order $X_1,\ldots,X_n$.

▷ The size of a Bayesian network is *not* a fixed property of the domain. It depends on the skill of the designer.

John and Mary Depend on the Variable Order!

▷ **Example 2.5.1.** MaryCalls, JohnCalls, Alarm, Burglary, Earthquake.

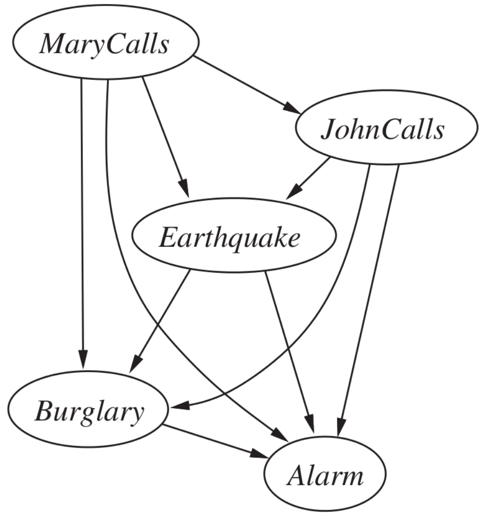


Note: For ?? we try to determine whether – given different value assignments to potential parents – the probability of X_i being true differs? If yes, we include these parents. In the particular case:

1. M to J yes because the common cause may be the alarm.
2. M, J to A yes because they may have heard alarm.
3. A to B yes because if A then higher chance of B.
4. However, M/J to B no because M/J only react to the alarm so if we have the value of A then values of M/J don't provide more information about B.
5. A to E yes because if A then higher chance of E.
6. B to E yes because, if A and not B then chances of E are higher than if A and B.

John and Mary Depend on the Variable Order! Ctd.

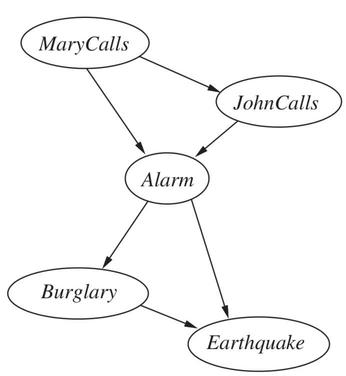
▷ **Example 2.5.2.** MaryCalls, JohnCalls, Earthquake, Burglary, Alarm.

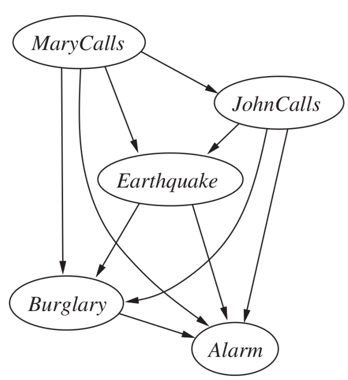


Again: Given different value assignments to potential parents, does the probability of X_i being true differ? If yes, include these parents.

1. M to J as before.
2. M, J to E as probability of E is higher if M/J is true.
3. Same for B; E to B because, given M and J are true, if E is true as well then prob of B is lower than if E is false.
4. $M/J/B/E$ to A because if $M/J/B/E$ is true (even when changing the value of just one of these) then probability of A is higher.

John and Mary, What Went Wrong?





▷ **Intuition:** These BNs link from symptoms to causes! ($\mathbf{P}(\text{Cavity}|\text{Toothache})$) Even though M and J are conditionally independent given A, they are *not* independent without any additional evidence; thus we don't "see" their conditional independence unless we ordered A before M and J! $\leadsto$ We organized the domain in the wrong way here.

We fail to identify many conditional independence relations (e.g., get dependencies between conditionally independent symptoms).

▷ **Also recall:** Conditional probabilities $\mathbf{P}(\text{Symptom}|\text{Cause})$ are more robust and often easier to assess than $\mathbf{P}(\text{Cause}|\text{Symptom})$.

▷ **Rule of Thumb:** We should order causes before symptoms.

Compactness of Bayesian Networks

▷ **Definition 2.5.3.** Given random variables $X_1, \ldots, X_n$ with finite domains $D_1, \ldots, D_n$ the size of $\mathcal{B} := \langle \{X_1, \ldots, X_n\}, E \rangle$ is defined as

$$\text{size}(\mathcal{B}) := \sum_{i=1}^{n} \#(D_i) \cdot \prod_{X_j \in \text{Parents}(X_i)} \#(D_j)$$

▷ **Note:** $\text{size}(\mathcal{B}) \;\widehat{=}\;$ The total number of entries in the CPTs.

▷ **Note:** Smaller BN $\leadsto$ need to assess less probabilities, more efficient inference.

▷ **Observation 2.5.4.** *Explicit full joint probability distribution has size* $\prod_{i=1}^{n} \#(D_i)$.

▷ **Observation 2.5.5.** *If* $\#(\text{Parents}(X_i)) \leq k$ *for every* X_i*, and* $D_{\max}$ *is the largest random variable domain, then* $\text{size}(\mathcal{B}) \leq n\#(D_{\max})^{k+1}$.

▷ **Example 2.5.6.** For $\#(D_{\max}) = 2$, $n = 20$, $k = 4$ we have $2^{20} = 1048576$ probabilities, but a Bayesian network of size $\leq 20 \cdot 2^5 = 640$ …!

▷ In the *worst case*, $\text{size}(\mathcal{B}) = n \cdot \prod_{i=1}^{n} \#(D_i)$, namely if every variable depends on all its predecessors in the chosen order.

▷ **Intuition:** BNs are compact if each variable is directly influenced only by few of its predecessor variables.

Constructing Bayesian Networks

▷ **Question:** What is the Bayesian network we get by constructing according to the ordering

1. $X_1 = \text{LoudNoise}, X_2 = \text{Animal}, X_3 = \text{LikesChappi}$?
2. $X_1 = \text{LoudNoise}, X_2 = \text{LikesChappi}, X_3 = \text{Animal}$?

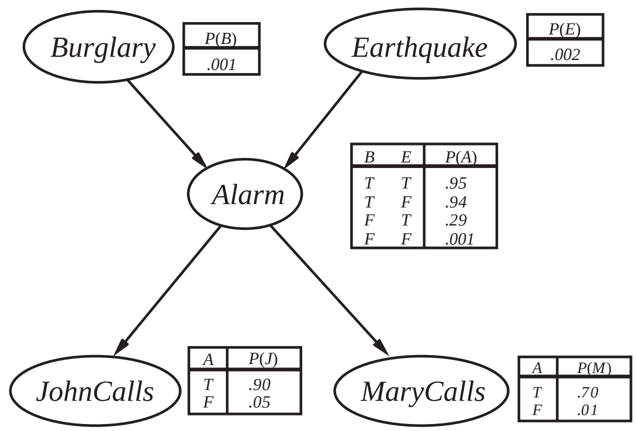
▷ **Answer:** reserved for the plenary sessions $\leadsto$ be there!

2.6 Inference in Bayesian Networks

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/29227`.

Inference for Mary and John

- **Intuition:** Observe evidence variables and draw conclusions on query variables.
- **Example 2.6.1.**



  - What is $\mathbf{P}(\text{Burglary}|\text{johncalls})$?
  - What is $\mathbf{P}(\text{Burglary}|\text{johncalls}, \text{marycalls})$?

Probabilistic Inference Tasks in Bayesian Networks

- **Definition 2.6.2 (Probabilistic Inference Task).** Given random variables $X_1, \ldots, X_n$, a probabilistic inference task consists of a set $\mathbf{X} \subseteq \{X_1, \ldots, X_n\}$ of query variables, a set $\mathbf{E} \subseteq \{X_1, \ldots, X_n\}$ of evidence variables, and an event $\mathbf{e}$ that assigns values to $\mathbf{E}$. We wish to compute the conditional probability distribution $\mathbf{P}(\mathbf{X}|\mathbf{e})$.

  $\mathbf{Y}:=\{X_1, \ldots, X_n\} \backslash \mathbf{X} \cup \mathbf{E}$ are the hidden variables.
- **Notes:**
  - We assume that a Bayesian network $\mathcal{B}$ for $X_1, \ldots, X_n$ is given.
  - In the remainder, for simplicity, $\mathbf{X} = \{X\}$ is a singleton.
- **Example 2.6.3.** In $\mathbf{P}(\text{Burglary}|\text{johncalls}, \text{marycalls})$, $X = \text{Burglary}$, $\mathbf{e} = \text{johncalls}, \text{marycalls}$, and $\mathbf{Y} = \{\text{Alarm}, \text{EarthQuake}\}$.

Inference by Enumeration: The Principle (A Reminder!)

- **Problem:** Given evidence $\mathbf{e}$, want to know $\mathbf{P}(X|\mathbf{e})$.
  Hidden variables: $\mathbf{Y}$.
- **1. Bayesian network:** Construct a Bayesian network $\mathcal{B}$ that captures variable

dependencies.

▷ **2. Normalization+Marginalization:**

$$\mathbf{P}(X|\mathbf{e}) = \alpha\mathbf{P}(X,\mathbf{e});\text{ if }\mathbf{Y}\neq\emptyset\text{ then }\mathbf{P}(X|\mathbf{e}) = \alpha(\sum_{\mathbf{y}\in\mathbf{Y}}\mathbf{P}(X,\mathbf{e},\mathbf{y}))$$

▷ Recover the summed-up probabilities $\mathbf{P}(X,\mathbf{e},\mathbf{y})$ from $\mathcal{B}$!

▷ **3. Chain Rule:** Order $X_1,\ldots,X_n$ consistent with $\mathcal{B}$.

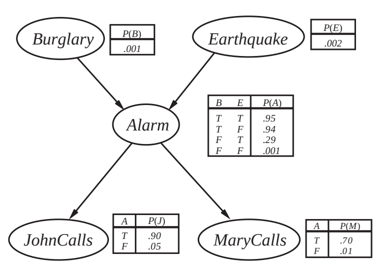
$$\mathbf{P}(X_1,\ldots,X_n) = \mathbf{P}(X_n|X_{n-1},\ldots,X_1)\cdot\mathbf{P}(X_{n-1}|X_{n-2},\ldots,X_1)\cdot\ldots\cdot\mathbf{P}(X_1)$$

▷ **4. Exploit conditional independence:** Instead of $\mathbf{P}(X_i|X_{i-1},\ldots,X_1)$, use $\mathbf{P}(X_i|\mathrm{Parents}(X_i))$.

▷ Given a Bayesian network $\mathcal{B}$, probabilistic inference tasks can be solved as sums of products of conditional probabilities from $\mathcal{B}$.

▷ Sum over all value combinations of hidden variables.

Inference by Enumeration: John and Mary



▷ **Want:** $\mathbf{P}(\text{Burglary}|\text{johncalls},\text{marycalls})$.
Hidden variables: $\mathbf{Y} = \{\text{Earthquake},\text{Alarm}\}$.

▷ **Normalization+Marginalization:**

$$\mathbf{P}(B|j,m) = \alpha\mathbf{P}(B,j,m) = \alpha(\sum_{v_E}\sum_{v_A}\mathbf{P}(B,j,m,v_E,v_A))$$

▷ **Order:** $X_1 = B$, $X_2 = E$, $X_3 = A$, $X_4 = J$, $X_5 = M$.

▷ **Chain rule and conditional independence:**

$$\mathbf{P}(B|j,m) = \alpha(\sum_{v_E}\sum_{v_A}\mathbf{P}(B)\cdot P(v_E)\cdot\mathbf{P}(v_A|B,v_E)\cdot P(j|v_A)\cdot P(m|v_A))$$

Inference by Enumeration: John and Mary, ctd.

- ▷ **Move variables outwards:** (until we hit the first parent):

$$\mathbf{P}(B|j,m) = \alpha \cdot \mathbf{P}(B) \cdot (\sum_{v_E} P(v_E) \cdot (\sum_{v_A} \mathbf{P}(v_A|B,v_E) \cdot P(j|v_A) \cdot P(m|v_A)))$$

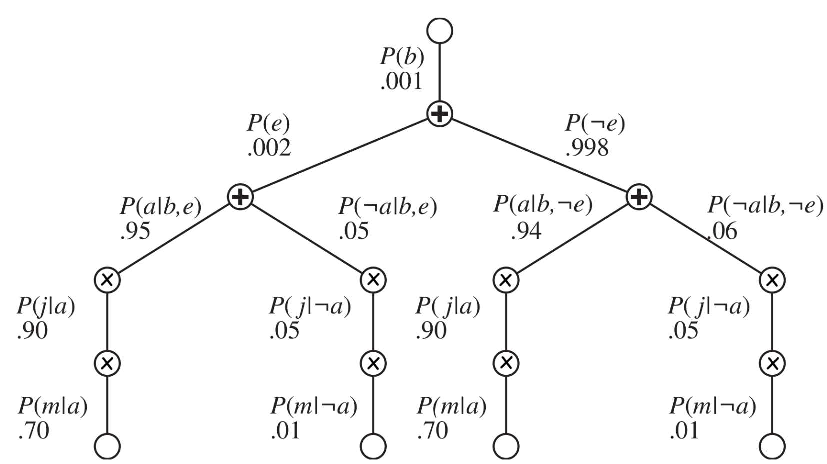
  Note: This step *is* actually done by the pseudo-code, implicitly in the sense that in the recursive calls to enumerate-all we multiply our own prob with all the rest. That is valid because, the variable ordering being consistent, all our parents are already here which is just another way of saying "my own prob does not depend on the variables in the rest of the order".

- ▷ The probabilities of the outside-variables multiply the entire "rest of the sum"
- ▷ **Chain rule and conditional independence, ctd.:**

$$\begin{aligned}
& \mathbf{P}(B|j,m) \\
= \; & \alpha \mathbf{P}(B)(\sum_{v_E} P(v_E)(\sum_{v_A} \mathbf{P}(v_A|B,v_E)P(j|v_A)P(m|v_A))) \\
= \; & \alpha \cdot P(b) \cdot \left(\begin{array}{c} P(e) \cdot \left(\begin{array}{l} \overbrace{P(a|b,e)P(j|a)P(m|a)}^{a} \\ + \; \underbrace{P(\neg a|b,e)P(j|\neg a)P(m|\neg a)}_{\neg a} \end{array} \right) \Bigg\} e \\ + \; P(\neg e) \cdot \left(\begin{array}{l} \overbrace{P(a|b,\neg e)P(j|a)P(m|a)}^{a} \\ + \; \underbrace{P(\neg a|b,\neg e)P(j|\neg a)P(m|\neg a)}_{\neg a} \end{array} \right) \Bigg\} \neg e \end{array} \right) \\
= \; & \alpha \langle 0.00059224, 0.0014919 \rangle \approx \langle 0.284, 0.716 \rangle
\end{aligned}$$

This computation can be viewed as a "search tree"! (see next slide)

The Evaluation of $P(b|j,m)$, as a "Search Tree"



- ▷ Inference by enumeration = a tree with "sum nodes" branching over values of hidden variables, and with non-branching "multiplication nodes".

Inference by Enumeration: Variable Elimination

- **Inference by Enumeration:**
  - Evaluates the tree in a depth-first manner.
  - space complexity: linear in the number of variables.
  - time complexity: exponential in the number of hidden variables, e.g. $\mathcal{O}(2^{\#(\mathbf{Y})})$ in case these variables are Boolean.
- Can we do better than this?
- **Definition 2.6.4.** Variable elimination is a BNI algorithm that avoids
  - repeated computation, and (see below)
  - irrelevant computation. (see below)
- In some special cases, variable elimination runs in polynomial time.

Variable Elimination: Sketch of Ideas

- **Avoiding repeated computation:** Evaluate expressions from right to left, storing all intermediate results.
- For query $P(B|j,m)$:
  1. CPTs of BN yield *factors* (probability tables):
  $$\mathbf{P}(B|j,m) = \alpha \cdot \underbrace{\mathbf{P}(B)}_{\mathbf{f}_1(B)} \cdot (\sum_{v_E} \underbrace{P(v_E)}_{\mathbf{f}_2(E)} \sum_{v_A} \underbrace{\mathbf{P}(v_A|B,v_E)}_{\mathbf{f}_3(A,B,E)} \cdot \underbrace{P(j|v_A)}_{\mathbf{f}_4(A)} \cdot \underbrace{P(m|v_A)}_{\mathbf{f}_5(A)})$$
  2. Then the computation is performed in terms of *factor product* and *summing out variables* from factors:
  $$\mathbf{P}(B|j,m) = \alpha \cdot \mathbf{f}_1(B) \cdot (\sum_{v_E} \mathbf{f}_2(E) \cdot (\sum_{v_A} \mathbf{f}_3(A,B,E) \cdot \mathbf{f}_4(A) \cdot \mathbf{f}_5(A)))$$
- **Avoiding irrelevant computation:** Repeatedly remove hidden variables that are leaf nodes.
- For query $P(\text{JohnCalls}|\text{burglary})$:
  $$\mathbf{P}(J|b) = \alpha \cdot P(b) \cdot (\sum_{v_E} P(v_E) \cdot (\sum_{v_A} P(v_A|b,v_E) \cdot \mathbf{P}(J|v_A) \cdot (\sum_{v_M} P(v_M|v_A))))$$
  - The rightmost sum equals 1 and can be dropped.

The Complexity of Exact Inference

- **Definition 2.6.5.** A graph G is called singly connected, or a polytree (otherwise multiply connected), if there is at most one undirected path between any two nodes in G.
- **Theorem 2.6.6 (Good News).** *On singly connected Bayesian networks, variable elimination runs in polynomial time.*
- Is our BN for Mary & John a polytree? (Yes.)
- **Theorem 2.6.7 (Bad News).** *For multiply connected Bayesian networks, probabilistic inference is* ***#P****-hard.* (***#P*** *is harder than* ***NP****, i.e.* ***NP*** $\subseteq$ ***#P***)
- **So?:** Life goes on ... In the hard cases, if need be we can throw exactitude to the winds and approximate.
- **Example 2.6.8.** Sampling techniques as in MCTS.

2.7 Conclusion

A Video Nugget covering this section can be found at `https://fau.tv/clip/id/29228`.

Summary

- Bayesian networks (BN) are a wide-spread tool to model uncertainty, and to reason about it. A BN represents conditional independence relations between random variables. It consists of a graph encoding the variable dependencies, and of conditional probability tables (CPTs).
- Given a variable order, the BN is small if every variable depends on only a few of its predecessors.
- Probabilistic inference requires to compute the probability distribution of a set of query variables, given a set of evidence variables whose values we know. The remaining variables are hidden.
- Inference by enumeration takes a BN as input, then applies Normalization+Marginalization, the chain rule, and exploits conditional independence. This can be viewed as a tree search that branches over all values of the hidden variables.
- Variable elimination avoids unnecessary computation. It runs in polynomial time for poly-tree BNs. In general, exact probabilistic inference is **#P**-hard. Approximate probabilistic inference methods exist.

Topics We Didn't Cover Here

▷ **Inference by sampling**: A whole zoo of methods for doing this exists.

▷ **Clustering**: Pre-combining subsets of variables to reduce the running time of inference.

▷ **Compilation to SAT**: More precisely, to "weighted model counting" in CNF formulas. Model counting extends DPLL with the ability to determine the number of satisfying interpretations. Weighted model counting allows to define a mass for each such interpretation (= the probability of an atomic event).

▷ **Dynamic BN**: BN with one slice of variables at each "time step", encoding probabilistic behavior over time.

▷ **Relational BN**: BN with predicates and object variables.

▷ **First-order BN**: Relational BN with quantification, i.e. probabilistic logic. E.g., the BLOG language developed by Stuart Russel and co-workers.

Reading:

- *Chapter 14: Probabilistic Reasoning* of [RN03].
  - Section 14.1 roughly corresponds to my "What is a Bayesian Network?".
  - Section 14.2 roughly corresponds to my "What is the Meaning of a Bayesian Network?" and "Constructing Bayesian Networks".The main change I made here is to *define* the semantics of the BN in terms of the conditional independence relations, which I find clearer than RN's definition that uses the reconstructed full joint probability distribution instead.
  - Section 14.4 roughly corresponds to my "Inference in Bayesian Networks". RN give full details on variable elimination, which makes for nice ongoing reading.
  - Section 14.3 discusses how CPTs are specified in practice.
  - Section 14.5 covers approximate sampling-based inference.
  - Section 14.6 briefly discusses relational and first-order BNs.
  - Section 14.7 briefly discusses other approaches to reasoning about uncertainty.

All of this is nice as additional background reading.

Bibliography

[DF31] B. De Finetti. “Sul significato soggettivo della probabilita”. In: *Fundamenta Mathematicae* 17 (1931), pp. 298–329.

[Pra+94] Malcolm Pradhan et al. “Knowledge Engineering for Large Belief Networks”. In: *Proceedings of the Tenth International Conference on Uncertainty in Artificial Intelligence*. UAI’94. Seattle, WA: Morgan Kaufmann Publishers Inc., 1994, pp. 484–490. ISBN: 1-55860-332-8. URL: `http://dl.acm.org/citation.cfm?id=2074394.2074456`.

[RN03] Stuart J. Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. 2nd ed. Pearso n Education, 2003. ISBN: 0137903952.