Probabilistic Reasoning and uncertainty

Sailee N. Panchbhai

Syllabus

Quantifying Uncertainty, Acting under Uncertainty, Basic Probability Notation, Inference Using Full Joint Distributions, Independence, Bayes' Rule, and Its Use, The Wumpus World Revisited, Probabilistic Reasoning, Representing Knowledge in an Uncertain Domain, The Semantics of Bayesian Networks, Efficient Representation of Conditional Distributions Exact Inference in Bayesian Networks, Approximate Inference in Bayesian Networks, Relational and First-Order Probability Models, and Other Approaches to Uncertain Reasoning.

Quantifying Uncertainty

- Talking about uncertainty means some weakness in the strength of belief of the expert in the conclusion.
- Two sides of a triangle are equal then the triangle is isosceles. (No uncertainty in this conclusion.)
- However, because of several reasons in a real-life scenario like medical reasoning, or automobile diagnosis we can not be very sure about either the evidence or if the evidences are known and its relationship with the conclusion there is some uncertainty lying somewhere.
- In this chapter we will see how such uncertainty can be handled in the case of AI

The Doorbell Problem

- The doorbell rang at 12’o clock at the midnight.
  - Was someone at the door?
  - Did Mohan wake up?

- Proposition 1: AtDoor(x) → Doorbell
- Proposition 2: Doorbell → Wake(Mohan)

Reasoning about Doorbell 1

- Given Doorbell, can we say AtDoor(X), because AtDoor(x) →Doorbell?

P → Q (if P is true then Q is necessarily true)
(but if P is false then Q may be true or false)
(Q is true the P may or may not be true)

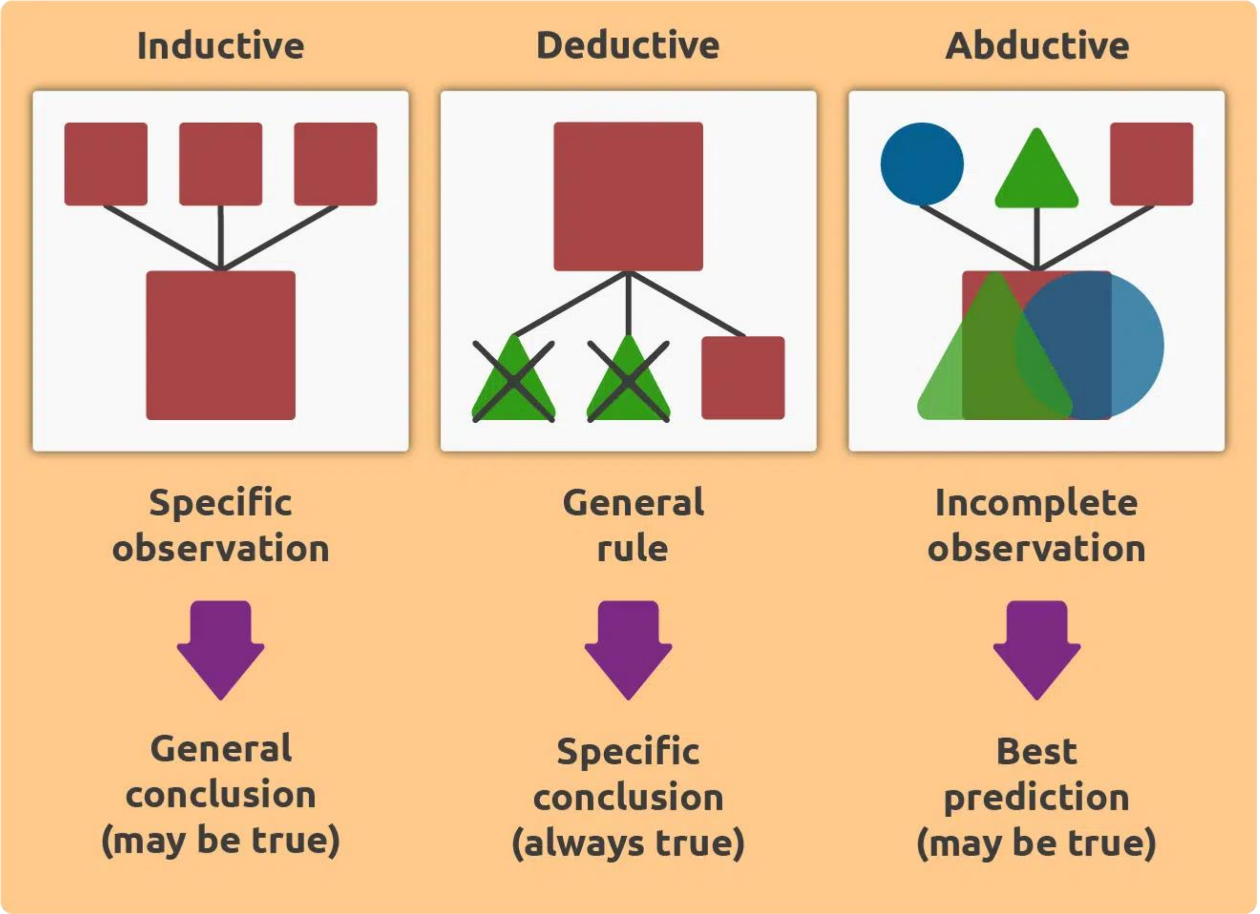
- According to deductive reasoning if there is a doorbell ringing we can not for sure say at door X.
- But using our common sense often we tend to conclude that someone is at the door if there is a doorbell. Although according to deductive reasoning may not be true.

- P → Q and we find Q is true and we infer P. This may not be correct according to a deduction but that is what we often do and this sort of reasoning is known as abductive reasoning
- The doorbell might start ringing due to some other reason e.g.

    - Short circuit

    - wind

    - Animal


Inductive
Deductive
Abductive
Specific observation
General conclusion (may be true)
General rule
Specific conclusion (always true)
Incomplete observation
Best prediction (may be true)

Reasoning about Doorbell 2

- Given Doorbell, can we say Wake(Mohan), because Doorbell → Wake(Mohan)?
- According to deductive reasoning yes.
- But again is this implication doorbell implies waking up Mohan always true?
- However, in general Mohan may not always wake up, even if the bell rings.
- Doorbell → Wake(Mohan) this is mostly correct but may not be always correct.

- Therefore we cannot answer either of the questions with certainty.
- Proposition 1 is incomplete. Modifying it as

AtDoor(x) V ShortCkt V Wind …..→ Doorbell

Does not help because the list of possible causes on the left is huge (infinite?)

- Proposition 2 is often true, but not a tautology.

Any way out?

- However, problems like that of the doorbell are very common in real life. In AI we often need to reason under such circumstances.

- We solve it by proper modeling of uncertainty and impreciseness and developing appropriate reasoning techniques.

Uncertainty:

- A situation in which something is not known, or something that is not known.
- So to represent uncertain knowledge, where we are not sure about the predicates, we need uncertain reasoning or probabilistic reasoning.

Sources of uncertainty

- Implication may be weak.

Doorbell(0.8) →Wake (Mohan)

Quantification of frequency

- Imprecise language like often, rarely, sometimes

  -Need to quantify these terms of frequencies

- Precise information may be too complex
  - Too many antecedents or consequents
  - AtDoor(x) V ShortCkt V Wind …..→ Doorbell

- Incomplete knowledge.
  - We may not know or guess all the possible antecedents or consequents
  - ‘The bell rang due to some spooky reason’

- Conflicting information

  - Experts often provide conflicting information

Quantification of measure of belief

Acting under uncertainty:

- Agents may need to handle uncertainty, whether due to partial observability, non-determinism, or a combination of the two.
- A agent may never know for certain what state it is in or where it will end up after a sequence of actions.
- **Probability theory** provides the **basis for our treatment of systems that reasons under uncertainty.**
- Four types of uncertainty in the decision-making process: data, prediction, judgment, and action uncertainty.

Causes of uncertainty

1. **Laziness:**

   - It is too much work to list the complete set of antecedents or consequents needed to ensure an exceptionless rule.

2. **Theoretical ignorance:**

   - Expertise in the area may not be sufficient to have a complete theory for the domain.

3. **Practical ignorance:**

   - Even if we know all the rules, we might be uncertain about a particular case because not all the necessary tests have been or can be run.

- The agent’s knowledge can at best provide only a **degree of belief** in the relevant sentence

Summarizing uncertainty:

- Dealing with a **degree of belief** is what **is done through the probability theory,** which assigns a numerical degree of belief between 0 and 1 to sentences.
- **Probability provides a way of summarizing the uncertainty** that comes from our laziness and ignorance.
- To make such choices, an agent must first have a **Preference** between the different possible **outcomes** of various plans.
- We use the **Utility theory** to represent and reason with preference.

- **Utility theory** says that every state has a degree of usefulness, to an agent and that agents will prefer states with higher utility.

Preference: Options, choices, what is more preferred.

Outcome: completely specified state

Decision theory = Probability Theory + Utility Theory

Basic Probability Notation

- Notation for **describing the degree of belief.**
- **Probability** can be defined as a **chance that an uncertain event will occur.** It is the **numerical measure of the likelihood** that an event will occur. The value of probability always remains between 0 and 1 which represents ideal uncertainties.
- $0 \leq P(A) \leq 1$, where $P(A)$ is the probability of an event A.
- $P(A) = 0$, indicates total uncertainty in an event A.
- $P(A) = 1$, indicates total certainty in an event A.

- The **dependence on experience** is reflected in the syntactic distinction between :

  - **Prior probability statements** which apply before any evidence is obtained.
  - **Conditional probability** statement which includes the evidence explicitly.

- Sample space: **The set of all possible worlds i.e., all possible outcomes** is referred to as sample space.
- Notation:
- Ω : $Sample\ Space$
- ω: $An\ element\ in\ sample\ space$
- φ: $An\ event$

An event φ *is a subset of sample space* Ω: $\varphi \subseteq \Omega$

Example : Two Dice adding up to 11 is an event

φ = { (5,6) , (6,5) }

Unconditional and conditional probability:

- **Unconditional probability** is when you **don't consider any other information** except for the object in the question.

Example: Two Dice – red and blue; consider only one – red

- In **Conditional probability** we have evidence i.e., **extra information already revealed.**

Example: Rolling two dice – one is 6; the sum can not be 5!

Prior Probability:

- Use the notation **P(A) for the unconditional or prior probability** that proposition A true.

For example, if Fever denotes the proposition that a particular patient has a fever,

P(Fever) = 0.1

This means that in the absence of any other information, the agent will assign a probability of 0.1 (a 10% chance) to the event of the patient having a fever.

- P(A) can only be used when there is no other information. As soon as some new information B is known, we have to reason with the conditional probability of A given B instead of P(A)

Random Variable

- Variables in probability theory are called random variables.
- Every random variable has a domain - a set of possible values that it can take
- For example, let's say we have the random variable **Total** that calculates the sum of two dice: then the domain is the set {2,…,12}
- A Boolean random variable has the domain {True, False}
- For propositions involving random variables; for example, if we are **concerned about the random variable Weather,** we might have

P(Weather = Sunny) = 0.7

P(Weather = Cloudy) =0.08

- Can view **proposition symbols as random variables** as well, if we assume that they have a domain (true, false)
- For example: The expression P(Fever) can be viewed as shorthand P(Fever = True).
- Similarly, P(¬Fever) is shorthand for P(Fever = False).

Probability Distribution

- A **probability distribution** is when we want to talk about **all the possible values of a random variable.** Usually indicated by a bold **P.**
- A **Discrete Random Variable** is a random variable that takes a **finite number of distinct values.**
- For example,

An expression such as P(Weather), denotes a vector of values for the probabilities of each individual state of the weather.

For example, we could write

P(Weather) = (0.7, 0.2,0.08,0.02)

This statement defines the **probability distribution.**

Probability density function:

- A **continuous random variable** is a random variable that takes an **infinite number of distinct values.**

For example:

$P(Temp = X) = Uniform_{(18C,26C)}(X)$

Expresses that the temperature is distributed uniformly between 18 and 26 degrees.

This is called the **probability density function.**

Intersection:

- Two different experiments
  - Rolling a die
  - Flipping a coin
- Identify events from each experiment
  - A = Rolling an even number
  - B = Flipping heads
- **Intersection:** occurs whenever event A **and** event B both occur
  - Notation: Either "A and B" or "A ∩ B"
  - Rolling an even number **and** flipping heads

Joint Probability:

- Two different experiments
  - Rolling a die
  - Flipping a coin
- Identify events from each experiment
  - A = Rolling an even number
  - B = Flipping heads

- **Joint Probability**: Probability that the intersection of two events occurs
  - <u>Notation</u>: Either "P(A and B)" or "P(A $\cap$ B)"

Joint Probability:

- Experiment 1: Rolling a die
  - Event A1: Rolling an even number
  - Event A2: Rolling an odd number
- Experiment 2: Flipping a coin
  - Event B1: Flipping heads
  - Event B2: Flipping tails
- Four possible intersections

1. Rolling even and flipping heads
2. Rolling even and flipping tails
3. Rolling odd and flipping heads
4. Rolling odd and flipping tails

Table of Probabilities:

- Outcomes for one experiment are listed along the rows.
- Outcomes for other experiments are listed at the top of the column.
- Joint probabilities go inside the table.

		Coin Flip	
		Heads (B1)	**Tails (B2)**
Die Roll	**Even (A1)**	P(A1 ∩ B1)	P(A1 ∩ B2)
	Odd (A2)	P(A2 ∩ B1)	P(A2 ∩ B2)

Calculating joint probabilities:

- Table of probabilities:

Gender	Smoker? Yes	Smoker? No
Male	0.08	0.32
Female	0.12	0.48

- <u>Question</u>: what is the probability that a person is male and does not smoke?
- <u>Answer</u>: P(Male ∩ No) = 0.32

Marginal probabilities:

- **Marginal Probability**: Probability that an individual event from one experiment occurs, regardless of the outcomes from another experiment
  - Computed by adding the probabilities across the row (or down the column) of the desired event
  - Always involve only one experiment
  - Get their names from the fact that they are written in the margins of the table
  - <u>Notation</u>: P(A1)

Calculating marginal probabilities:

Classification of bank employees				
Gender	Rank			
	R1	R2	R3	Total
Male	30	80	90	200
Female	20	40	40	100
Total	50	120	130	300

- Question: After selecting a random employee in the bank, what is the probability that he/she is a rank 1 holder employee?
- Answer: P(R1) = Total no of rank 1 holders/total no of employees.

  = 50/300 = 0.1666

Conditional Probability:

- **Conditional probability:** the probability that a second event (B) will occur given that we know that the first event (A) has already occurred.
  - Note: A and B come from two different experiments
  - Notation: $P(B|A)$ → vertical bar "|" means "given"

Calculating conditional probabilities:

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

An event we want the probability for

An event that has already occurred

- To calculate the conditional8907-87-87-87-87-887-77887-87-8777-20203645190106 7890w probability :
  - Find the joint probability of A and B
  - Find the marginal probability of the event that has already occurred (Event A)
  - Divide the joint probability by the marginal probability

Calculating conditional probabilities:

- College students were asked if they have ever cheated on an exam. The result was broken down by gender.

Cheated on college exam?

Gender	Yes	No	Total
Male	0.32	0.22	0.54
Female	0.28	0.18	0.46
Total	0.60	0.40	1.00

- Question: Given that person has cheated, what is the probability he is male?
- Answer: $P(\text{Male}|\text{Cheater}) = P(\text{Male} \cap \text{Cheater})/P(\text{Cheater})$

$$= 0.32/0.60 = 0.533$$

Calculating conditional probabilities:

- Conditional probabilities:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

- This equation can also be written as follows, which called the product rule.

$$P(A \cap B) = P(A|B)\, P(B)$$

Axioms of Probability:

1. All probabilities are between 0 and 1

$$0 < P(A) < 1$$

2. Necessarily true (i.e., valid) propositions have probability 1, and necessarily false (i.e., unsatisfiable) propositions have probability 0.

P (True) = 1 ; P (False) = 0

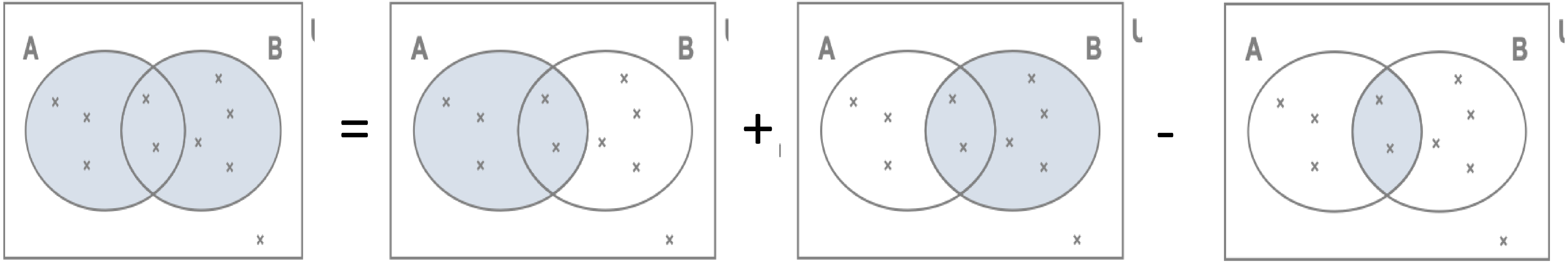
3. The probability of disjunction is given by

$$P(A \vee B) = P(A) + P(B) \ - P(A \wedge B)$$

Axioms of Probability:

3. The probability of disjunction is given by

$$P(A \vee B) = P(A) + P(B) - P(A \wedge B)$$



Axioms of Probability:

- From these three axioms, we can derive all other properties of probabilities.

For example, the Probability of negation

$$P(A \vee \neg A) = P(A) + P(\neg A) \ - P(A \ \wedge \neg A)$$
$$P(True) = P(A) + P(\neg A) \ - P(False)$$
$$1 = P(A) + P(\neg A)$$
$$P(\neg A) \ = 1 - P(A)$$

Inference using full joint distribution:

- To study method for probabilistic inference

We begin with a simple example: a domain consisting of just the three Boolean variables *Toothache*, *Cavity*, and *Catch* (the dentist's nasty steel probe catches in my tooth). The full

	$toothache$		$\neg toothache$	
	$catch$	$\neg catch$	$catch$	$\neg catch$
$cavity$	0.108	0.012	0.072	0.008
$\neg cavity$	0.016	0.064	0.144	0.576

Figure 13.3 A full joint distribution for the *Toothache*, *Cavity*, *Catch* world.

- How to determine the probability of any proposition?

ity of any proposition, simple or complex: simply identify those possible worlds in which the proposition is true and add up their probabilities. For example, there are six possible worlds in which $cavity \lor toothache$ holds:

$$P(cavity \lor toothache) = 0.108 + 0.012 + 0.072 + 0.008 + 0.016 + 0.064 = 0.28 .$$

Full joint probability distribution:

- Distributions on multiple variables
- For example,
  - Weather = {sunny, rain, cloudy, snow}
  - Cavity = {cavity, ¬cavity}
  - Joint probability distribution of weather and cavity

$$
\begin{aligned}
P(W = sunny \wedge C = true) &= P(W = sunny | C = true)\, P(C = true) \\
P(W = rain \wedge C = true) &= P(W = rain | C = true)\, P(C = true) \\
P(W = cloudy \wedge C = true) &= P(W = cloudy | C = true)\, P(C = true) \\
P(W = snow \wedge C = true) &= P(W = snow | C = true)\, P(C = true) \\
P(W = sunny \wedge C = false) &= P(W = sunny | C = false)\, P(C = false) \\
P(W = rain \wedge C = false) &= P(W = rain | C = false)\, P(C = false) \\
P(W = cloudy \wedge C = false) &= P(W = cloudy | C = false)\, P(C = false) \\
P(W = snow \wedge C = false) &= P(W = snow | C = false)\, P(C = false)\, .
\end{aligned}
$$

  - Can be written as single equation:

$$\mathbf{P}(Weather, Cavity) = \mathbf{P}(Weather \mid Cavity)\mathbf{P}(Cavity)$$

Inference using full joint distribution:

- To study method for probabilistic inference

We begin with a simple example: a domain consisting of just the three Boolean variables *Toothache*, *Cavity*, and *Catch* (the dentist's nasty steel probe catches in my tooth). The full

	$toothache$		$\neg toothache$	
	$catch$	$\neg catch$	$catch$	$\neg catch$
$cavity$	0.108	0.012	0.072	0.008
$\neg cavity$	0.016	0.064	0.144	0.576

Figure 13.3 A full joint distribution for the *Toothache*, *Cavity*, *Catch* world.

- How to determine the probability of any proposition?

ity of any proposition, simple or complex: simply identify those possible worlds in which the proposition is true and add up their probabilities. For example, there are six possible worlds in which $cavity \vee toothache$ holds:

$$P(cavity \vee toothache) = 0.108 + 0.012 + 0.072 + 0.008 + 0.016 + 0.064 = 0.28 .$$

Inference using full joint distribution:

- To compute conditional probability:

Equation (13.3) to obtain an expression in terms of unconditional probabilities and then evaluating the expression from the full joint distribution. For example, we can compute the probability of a cavity, given evidence of a toothache, as follows:

$$\begin{aligned} P(cavity \mid toothache) &= \frac{P(cavity \wedge toothache)}{P(toothache)} \\ &= \frac{0.108 + 0.012}{0.108 + 0.012 + 0.016 + 0.064} = 0.6 \,. \end{aligned}$$

Just to check, we can also compute the probability that there is no cavity, given a toothache:

$$\begin{aligned} P(\neg cavity \mid toothache) &= \frac{P(\neg cavity \wedge toothache)}{P(toothache)} \\ &= \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064} = 0.4 \,. \end{aligned}$$

Inference using full joint distribution:

- Normalization

The two values sum to 1.0, as they should. Notice that in these two calculations the term $1/P(toothache)$ remains constant, no matter which value of $Cavity$ we calculate. In fact, it can be viewed as a **normalization** constant for the distribution $\mathbf{P}(Cavity \mid toothache)$, ensuring that it adds up to 1. Throughout the chapters dealing with probability, we use α to denote such constants. With this notation, we can write the two preceding equations in one:

$$\begin{aligned}\mathbf{P}(Cavity \mid toothache) &= \alpha\, \mathbf{P}(Cavity, toothache) \\ &= \alpha\, [\mathbf{P}(Cavity, toothache, catch) + \mathbf{P}(Cavity, toothache, \neg catch)] \\ &= \alpha\, [\langle 0.108, 0.016\rangle + \langle 0.012, 0.064\rangle] = \alpha\, \langle 0.12, 0.08\rangle = \langle 0.6, 0.4\rangle\ .\end{aligned}$$

In other words, we can calculate $\mathbf{P}(Cavity \mid toothache)$ even if we don't know the value of $P(toothache)$! We temporarily forget about the factor $1/P(toothache)$ and add up the values for $cavity$ and $\neg cavity$, getting 0.12 and 0.08. Those are the correct relative proportions, but they don't sum to 1, so we normalize them by dividing each one by $0.12 + 0.08$, getting the true probabilities of 0.6 and 0.4. Normalization turns out to be a useful shortcut in many probability calculations, both to make the computation easier and to allow us to proceed when some probability assessment (such as $P(toothache)$) is not available.

Independence:

- <u>Dependent event:</u>

An example of a dependent event is the probability of the clouds in the sky and the probability of rain on that day. The probability of clouds in the sky has an impact on the probability of rain that day. They are, therefore, dependent events.

- <u>Independent event:</u>

An example of an independent event is the probability of getting heads on two coin tosses. The probability of getting heads on the first coin toss does not have an impact on the probability of getting heads on the second coin toss.

Independence:

- Independence between propositions a and b can be written as,

$$P(a \mid b) = P(a) \quad \text{or} \quad P(b \mid a) = P(b) \quad \text{or} \quad P(a \wedge b) = P(a)P(b) \,.$$

Independence:

- Example, add Weather variable in our previous example,

$$\mathbf{P}(Toothache, Catch, Cavity, Weather)$$

- To find the P(toothache, catch, cavity, cloudy)

and $P(toothache, catch, cavity)$ related? We can use the product rule:

$$\begin{aligned} &P(toothache, catch, cavity, cloudy) \\ &\quad = P(cloudy \mid toothache, catch, cavity)P(toothache, catch, cavity) . \end{aligned}$$

- It seems safe to say that the weather does not influence the dental variables. Therefore the following assertion seems reasonable

$$P(cloudy \mid toothache, catch, cavity) = P(cloudy) . \quad (1$$

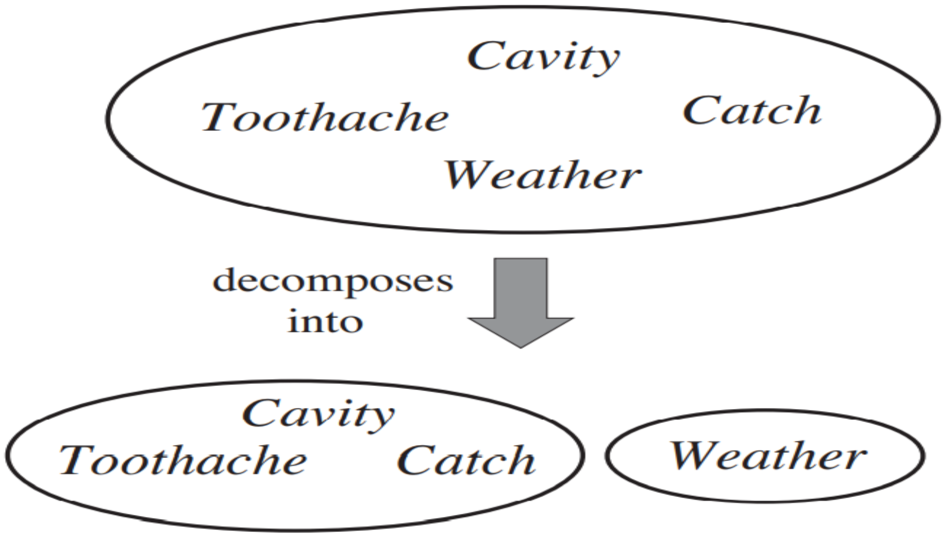
From this, we can deduce

$$P(toothache, catch, cavity, cloudy) = P(cloudy)P(toothache, catch, cavity) .$$

Independence:

$$\mathbf{P}(Toothache, Catch, Cavity, Weather) = \mathbf{P}(Toothache, Catch, Cavity)\mathbf{P}(Weather)\,.$$

Thus, the 32-element table for four variables can be constructed from one 8-element table and one 4-element table. This decomposition is illustrated schematically in Figure 13.4(a).



Bayes' rule and its use:

- Two forms of the product rule

$$P(a \wedge b) = P(a \,|\, b)P(b) \qquad \text{and} \qquad P(a \wedge b) = P(b \,|\, a)P(a)\ .$$

Equating the two right-hand sides and dividing by $P(a)$, we get

$$P(b \,|\, a) = \frac{P(a \,|\, b)P(b)}{P(a)}\ . \tag{13.12}$$

This equation is known as **Bayes' rule** (also Bayes' law or Bayes' theorem). This simple equation underlies most modern AI systems for probabilistic inference.

Applying Bayes' Rule

- Example,

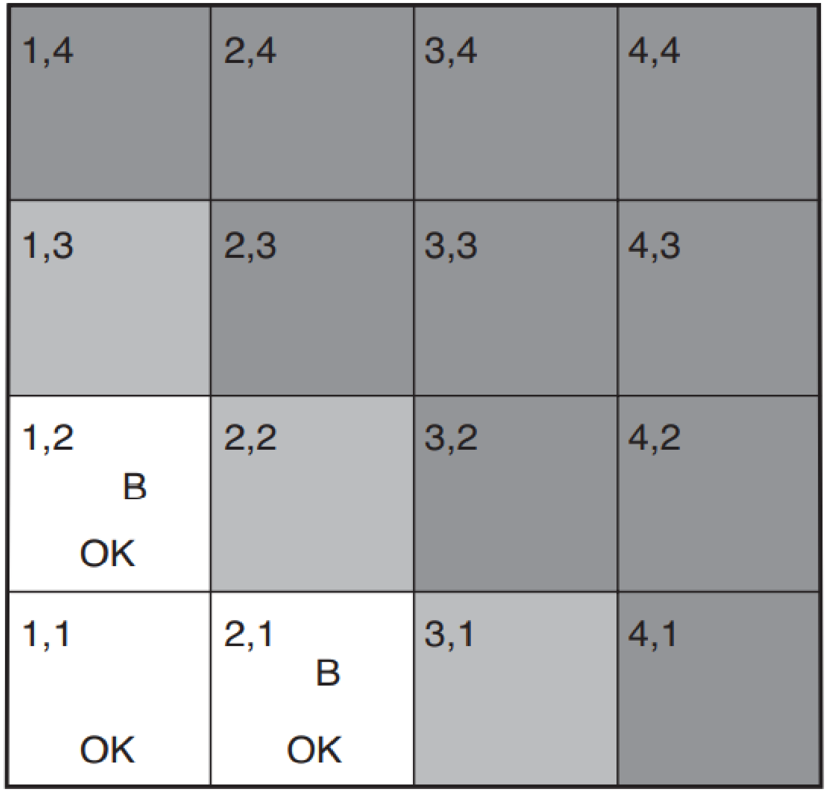
diagnosis, we often have conditional probabilities on causal relationships (that is, the doctor knows $P(symptoms \mid disease)$) and want to derive a diagnosis, $P(disease \mid symptoms)$. For example, a doctor knows that the disease meningitis causes the patient to have a stiff neck, say, 70% of the time. The doctor also knows some unconditional facts: the prior probability that a patient has meningitis is 1/50,000, and the prior probability that any patient has a stiff neck is 1%. Letting s be the proposition that the patient has a stiff neck and m be the proposition that the patient has meningitis, we have

$$
\begin{aligned}
P(s \mid m) &= 0.7 \\
P(m) &= 1/50000 \\
P(s) &= 0.01 \\
P(m \mid s) &= \frac{P(s \mid m)P(m)}{P(s)} = \frac{0.7 \times 1/50000}{0.01} = 0.0014 \,. \qquad (13.14)
\end{aligned}
$$

That is, we expect less than 1 in 700 patients with a stiff neck to have meningitis. Notice that

The Wumpus World Revisited:

- Aim – To calculate the probability that each of the three squares contains a pit.



(a)

Figure 13.5 (a): After finding a breeze in both [1,2] and [2,1], the agent is stuck-there is no safe place to explore

The Wumpus World Revisited:

- <u>Variables</u>
  - Boolean variable $P_{i,j}$ for each square, which is true if square [i,j] actually contains a pit.
  - Boolean variable Bi,j that are true if square [i,j] is breezy; we include this variable only for the observed squares – in this case, [1,1], [1,2], and [2,1].

- <u>Properties of the world</u>
  1) A pit causes breezes in all neighboring squares,
  2) Each square other than [1,1] contains a pit with a probability 0.2

The Wumpus World Revisited:

- Recapitulation

$$P(cavity \mid toothache) = \frac{P(cavity \wedge toothache)}{P(toothache)} = \frac{0.108 + 0.012}{0.108 + 0.012 + 0.016 + 0.064} = 0.6 .$$

Just to check, we can also compute the probability that there is no cavity, given a toothache:

$$P(\neg cavity \mid toothache) = \frac{P(\neg cavity \wedge toothache)}{P(toothache)} = \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064} = 0.4 .$$

$$\begin{aligned}\mathbf{P}(Cavity \mid toothache) &= \alpha\, \mathbf{P}(Cavity, toothache)\\ &= \alpha\, [\mathbf{P}(Cavity, toothache, catch) + \mathbf{P}(Cavity, toothache, \neg catch)]\\ &= \alpha\, [\langle 0.108, 0.016\rangle + \langle 0.012, 0.064\rangle] = \alpha\, \langle 0.12, 0.08\rangle = \langle 0.6, 0.4\rangle .\end{aligned}$$

	toothache		*¬toothache*	
	catch	*¬catch*	*catch*	*¬catch*
cavity	0.108	0.012	0.072	0.008
¬cavity	0.016	0.064	0.144	0.576

From the example, we can extract a general inference procedure. We begin with the case in which the query involves a single variable, X (*Cavity* in the example). Let $\mathbf{E}$ be the list of evidence variables (just *Toothache* in the example), let $\mathbf{e}$ be the list of observed values for them, and let $\mathbf{Y}$ be the remaining unobserved variables (just *Catch* in the example). The query is $\mathbf{P}(X \mid \mathbf{e})$ and can be evaluated as

$$\mathbf{P}(X \mid \mathbf{e}) = \alpha\, \mathbf{P}(X, \mathbf{e}) = \alpha \sum_{\mathbf{y}} \mathbf{P}(X, \mathbf{e}, \mathbf{y}) , \qquad (13.9)$$

The Wumpus World Revisited:

The next step is to specify the full joint distribution, $\mathbf{P}(P_{1,1}, \ldots, P_{4,4}, B_{1,1}, B_{1,2}, B_{2,1})$. Applying the product rule, we have

$$\mathbf{P}(P_{1,1}, \ldots, P_{4,4}, B_{1,1}, B_{1,2}, B_{2,1}) = \\ \mathbf{P}(B_{1,1}, B_{1,2}, B_{2,1} \mid P_{1,1}, \ldots, P_{4,4})\mathbf{P}(P_{1,1}, \ldots, P_{4,4})\ .$$

This decomposition makes it easy to see what the joint probability values should be. The first term is the conditional probability distribution of a breeze configuration, given a pit configuration; its values are 1 if the breezes are adjacent to the pits and 0 otherwise. The second term is the prior probability of a pit configuration. Each square contains a pit with probability 0.2, independently of the other squares; hence,

$$\mathbf{P}(P_{1,1}, \ldots, P_{4,4}) = \prod_{i,j=1,1}^{4,4} \mathbf{P}(P_{i,j})\ . \qquad (13.20)$$

For a particular configuration with exactly n pits, $P(P_{1,1}, \ldots, P_{4,4}) = 0.2^n \times 0.8^{16-n}$.

The Wumpus World Revisited:

In the situation in Figure 13.5(a), the evidence consists of the observed breeze (or its absence) in each square that is visited, combined with the fact that each such square contains no pit. We abbreviate these facts as $b = \neg b_{1,1} \wedge b_{1,2} \wedge b_{2,1}$ and $known = \neg p_{1,1} \wedge \neg p_{1,2} \wedge \neg p_{2,1}$. We are interested in answering queries such as $\mathbf{P}(P_{1,3} \mid known, b)$: how likely is it that [1,3] contains a pit, given the observations so far?

To answer this query, we can follow the standard approach of Equation (13.9), namely, summing over entries from the full joint distribution. Let $Unknown$ be the set of $P_{i,j}$ vari-

The Wumpus World Revisited:

- Let **Unknown** be the set of $P_{i,j}$ variables for squares other than the **Known** squares and the **query** square [1,3].
- Let **Frontier** be the pit variables (other than query variable) that are adjacent to visited squares
- Also, let **Other** be the pit variables for the other unknown squares; in this case, there are 10 other squares, as shown in fig 13.5(b)

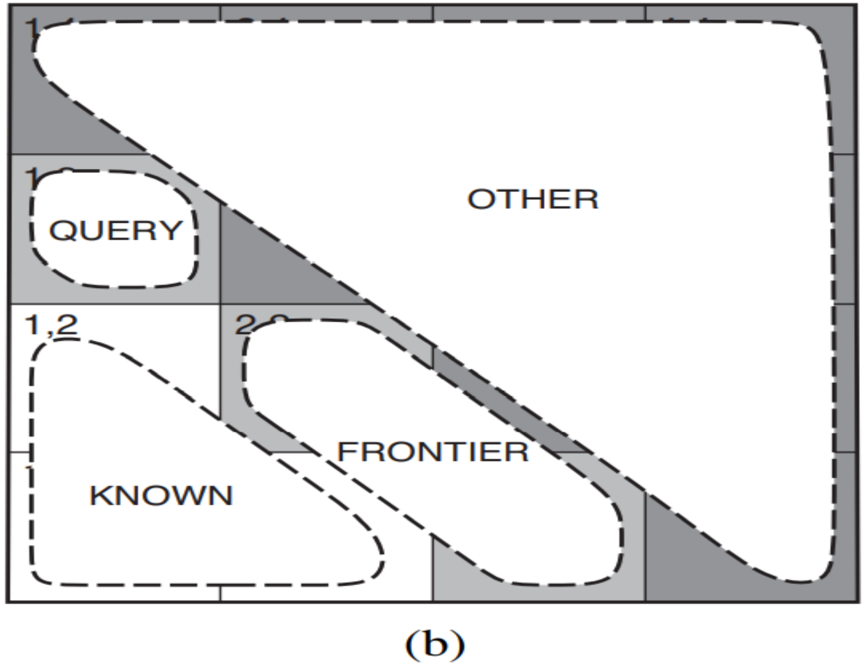


Figure 13.5(b) Division of the squares into Known, Frontier, and Other, for a query about [1,3].

The Wumpus World Revisited:

- Recapitulation

$$P(cavity \mid toothache) = \frac{P(cavity \wedge toothache)}{P(toothache)} = \frac{0.108 + 0.012}{0.108 + 0.012 + 0.016 + 0.064} = 0.6 .$$

Just to check, we can also compute the probability that there is no cavity, given a toothache:

$$P(\neg cavity \mid toothache) = \frac{P(\neg cavity \wedge toothache)}{P(toothache)} = \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064} = 0.4 .$$

$$\begin{aligned}\mathbf{P}(Cavity \mid toothache) &= \alpha\, \mathbf{P}(Cavity, toothache)\\ &= \alpha\, [\mathbf{P}(Cavity, toothache, catch) + \mathbf{P}(Cavity, toothache, \neg catch)]\\ &= \alpha\, [\langle 0.108, 0.016\rangle + \langle 0.012, 0.064\rangle] = \alpha\, \langle 0.12, 0.08\rangle = \langle 0.6, 0.4\rangle .\end{aligned}$$

	toothache		¬toothache	
	catch	¬catch	catch	¬catch
cavity	0.108	0.012	0.072	0.008
¬cavity	0.016	0.064	0.144	0.576

From the example, we can extract a general inference procedure. We begin with the case in which the query involves a single variable, X ($Cavity$ in the example). Let $\mathbf{E}$ be the list of evidence variables (just $Toothache$ in the example), let $\mathbf{e}$ be the list of observed values for them, and let $\mathbf{Y}$ be the remaining unobserved variables (just $Catch$ in the example). The query is $\mathbf{P}(X \mid \mathbf{e})$ and can be evaluated as

$$\mathbf{P}(X \mid \mathbf{e}) = \alpha\, \mathbf{P}(X, \mathbf{e}) = \alpha \sum_{\mathbf{y}} \mathbf{P}(X, \mathbf{e}, \mathbf{y}) , \qquad (13.9)$$

The Wumpus World Revisited:

ables for squares other than the *Known* squares and the query square [1,3]. Then, by Equation (13.9), we have

$$\mathbf{P}(P_{1,3} \mid known, b) = \alpha \sum_{unknown} \mathbf{P}(P_{1,3}, unknown, known, b)\ .$$

The full joint probabilities have already been specified, so we are done—that is, unless we care about computation. There are 12 unknown squares; hence the summation contains 10 other squares, as shown in Figure 13.5(b). The key insight is that the observed breezes are *conditionally independent* of the other variables, given the known, frontier, and query variables. To use the insight, we manipulate the query formula into a form in which the breezes are conditioned on all the other variables, and then we apply conditional independence:

$$\begin{aligned}
&\mathbf{P}(P_{1,3} \mid known, b) \\
&= \alpha \sum_{unknown} \mathbf{P}(P_{1,3}, known, b, unknown) \qquad \text{(by Equation (13.9))} \\
&= \alpha \sum_{unknown} \mathbf{P}(b \mid P_{1,3}, known, unknown)\mathbf{P}(P_{1,3}, known, unknown) \\
&\qquad\qquad \text{(by the product rule)}
\end{aligned}$$

$$= \alpha \sum_{frontier} \sum_{other} \mathbf{P}(b \mid known, P_{1,3}, frontier, other)\mathbf{P}(P_{1,3}, known, frontier, other)$$

$$= \alpha \sum_{frontier} \sum_{other} \mathbf{P}(b \mid known, P_{1,3}, frontier)\mathbf{P}(P_{1,3}, known, frontier, other)\ ,$$

where the final step uses conditional independence: b is independent of $other$ given $known$, $P_{1,3}$, and $frontier$. Now, the first term in this expression does not depend on the $Other$ variables, so we can move the summation inward:

$$\mathbf{P}(P_{1,3} \mid known, b)$$
$$= \alpha \sum_{frontier} \mathbf{P}(b \mid known, P_{1,3}, frontier) \sum_{other} \mathbf{P}(P_{1,3}, known, frontier, other)\ .$$

By independence, as in Equation (13.20), the prior term can be factored, and then the terms can be reordered:

$$\mathbf{P}(P_{1,3} \mid known, b)$$
$$= \alpha \sum_{frontier} \mathbf{P}(b \mid known, P_{1,3}, frontier) \sum_{other} \mathbf{P}(P_{1,3})P(known)P(frontier)P(other)$$
$$= \alpha\, P(known)\mathbf{P}(P_{1,3}) \sum_{frontier} \mathbf{P}(b \mid known, P_{1,3}, frontier)P(frontier) \sum_{other} P(other)$$
$$= \alpha'\, \mathbf{P}(P_{1,3}) \sum_{frontier} \mathbf{P}(b \mid known, P_{1,3}, frontier)P(frontier)\ ,$$

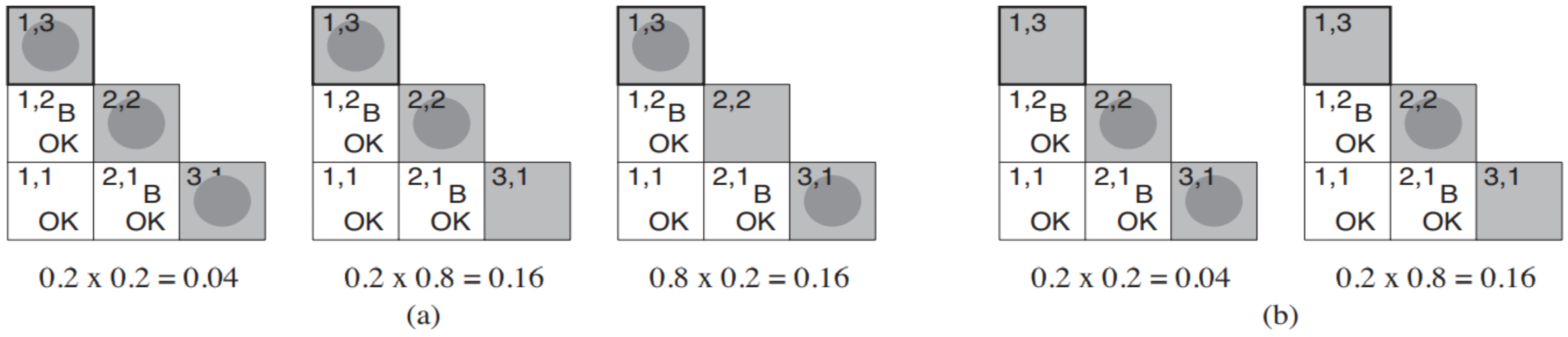


Figure 13.6 Consistent models for the frontier variables $P_{2,2}$ and $P_{3,1}$, showing $P(frontier)$ for each model: (a) three models with $P_{1,3} = true$ showing two or three pits, and (b) two models with $P_{1,3} = false$ showing one or two pits.

Notice that the expression $\mathbf{P}(b \mid known, P_{1,3}, frontier)$ is 1 when the frontier is consistent with the breeze observations, and 0 otherwise. Thus, for each value of $P_{1,3}$, we sum over the *logical models* for the frontier variables that are consistent with the known facts. (Compare with the enumeration over models in Figure 7.5 on page 241.) The models and their associated prior probabilities—$P(frontier)$—are shown in Figure 13.6. We have

$$\mathbf{P}(P_{1,3} \mid known, b) = \alpha' \langle 0.2(0.04 + 0.16 + 0.16),\ 0.8(0.04 + 0.16) \rangle \approx \langle 0.31, 0.69 \rangle\ .$$

That is, [1,3] (and [3,1] by symmetry) contains a pit with roughly 31% probability. A similar calculation, which the reader might wish to perform, shows that [2,2] contains a pit with roughly 86% probability. The wumpus agent should definitely avoid [2,2]! Note that our

Representing knowledge in an uncertain domain:

Bayesian network:

- The full joint probability distribution can answer any question about the domain, but can become intractably large as the number of variables grows.
- This section introduces the Bayesian network to represent the dependencies among variables.
- Bayesian network is also known as belief network, probabilistic network, causal network, and knowledge map.

Bayesian network:

- A Bayesian network is a directed graph in which each node is annotated with quantitative probability information. The full specification is as follows:

1. Each node corresponds to a random variable, which may be discrete or continuous.
2. A set of directed links or arrows connect pairs of nodes. If there is an arrow from node X to node Y, X is said to be a parent of Y. The graph has no directed cycles, hence it is a directed acyclic graph or DAG.
3. Each node X_i has a conditional probability distribution $\mathbf{P(X_i|parents(X_i))}$ that quantifies the effect of the parents on the node.

Bayesian network:

- The intuitive meaning of an arrow is typically that X has a direct influence on Y.
- It is usually easy for a domain expert to decide what direct influences exist in the domain.
- Once the topology of the Bayesian network is laid out, we need only specify a conditional probability distribution for each variable, given its parents.

Bayesian Belief Network –Example.

- You have a new burglar alarm installed at home.
- It is fairly reliable at detecting burglary, but also sometimes responds to minor earthquakes.
- You have two neighbors, John and Merry, who promised to call you at work when they hear the alarm.
- Jhon always calls when he hears the alarm, but sometimes confuses telephone ringing with alarm and calls too.
- Merry likes loud music and sometimes misses the alarm.
- Given the evidence of who has or has not called, we would like to estimate the probability of a burglary.

Bayesian Belief Network –Example

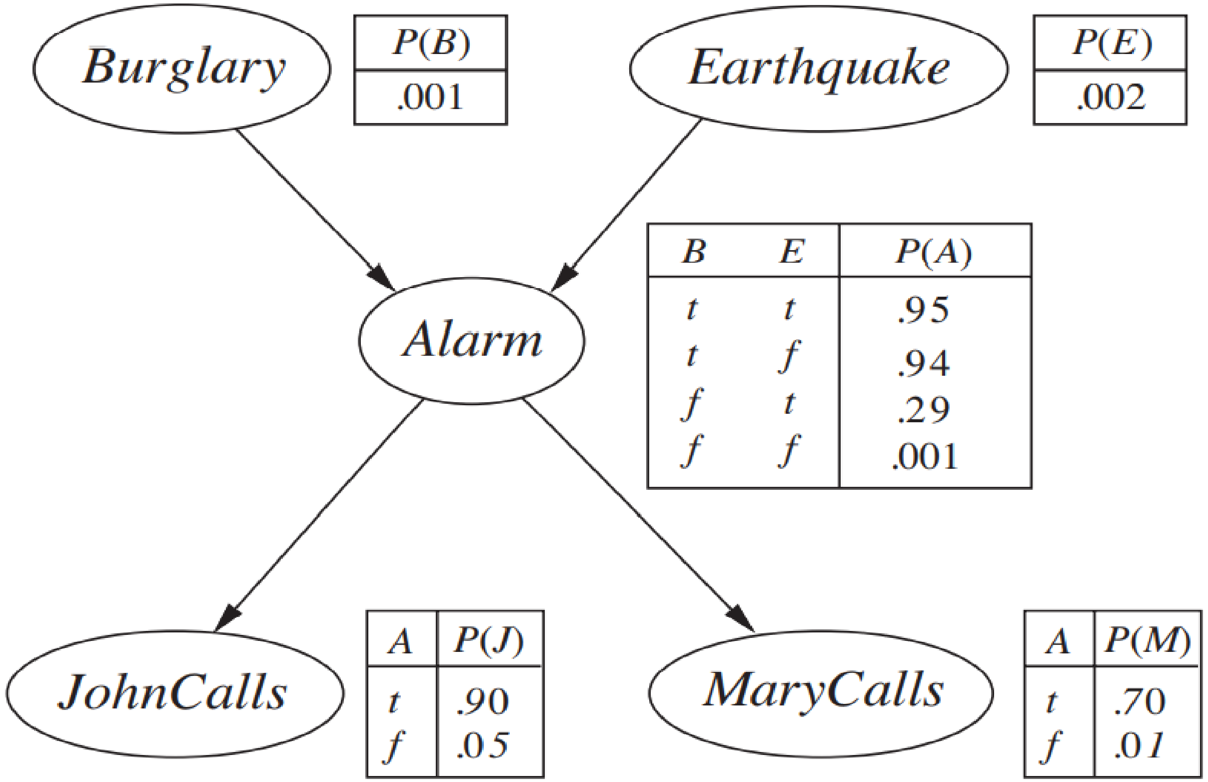


Figure 14.2 A typical Bayesian network, showing both the topology and the conditional probability tables (CPTs). In the CPTs, the letters B, E, A, J, and M stand for *Burglary*, *Earthquake*, *Alarm*, *JohnCalls*, and *MaryCalls*, respectively.

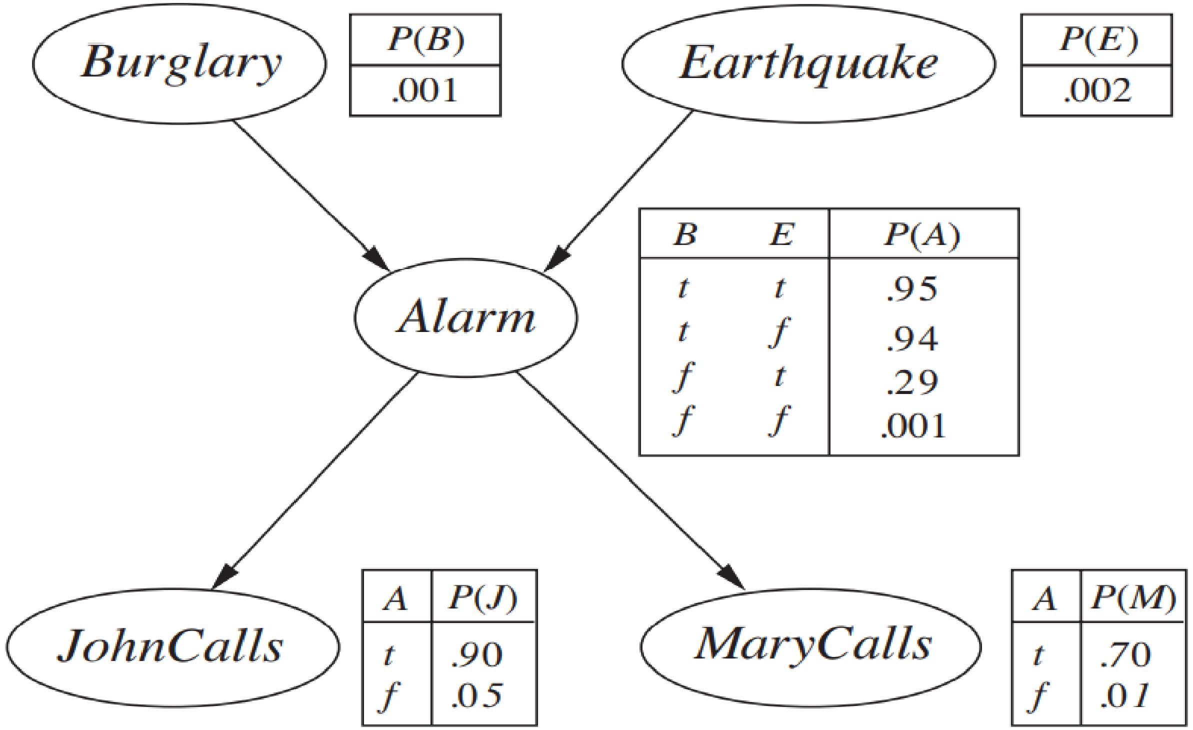
The semantics of Bayesian Network:

- A Bayesian network is a directed acyclic graph with some numeric parameters attached to each node.
- The network semantics – is to define the way in which it represents a specific joint distribution over all the variables.
- We know that **$P(X_i|parents(X_i))$.**
- Joint distribution is the probability of a conjunction of particular assignments to each variable , such as

$$P(X_1 = x_1 \wedge \cdots \wedge Xn = xn)$$

$$P(x_{1,} \ldots, xn) = \prod_{i=1}^{n} P(xi|parents(xi))$$

The semantics of Bayesian Network:



P(B)
.001

P(E)
.002

B	E	P(A)
t	t	.95
t	f	.94
f	t	.29
f	f	.001

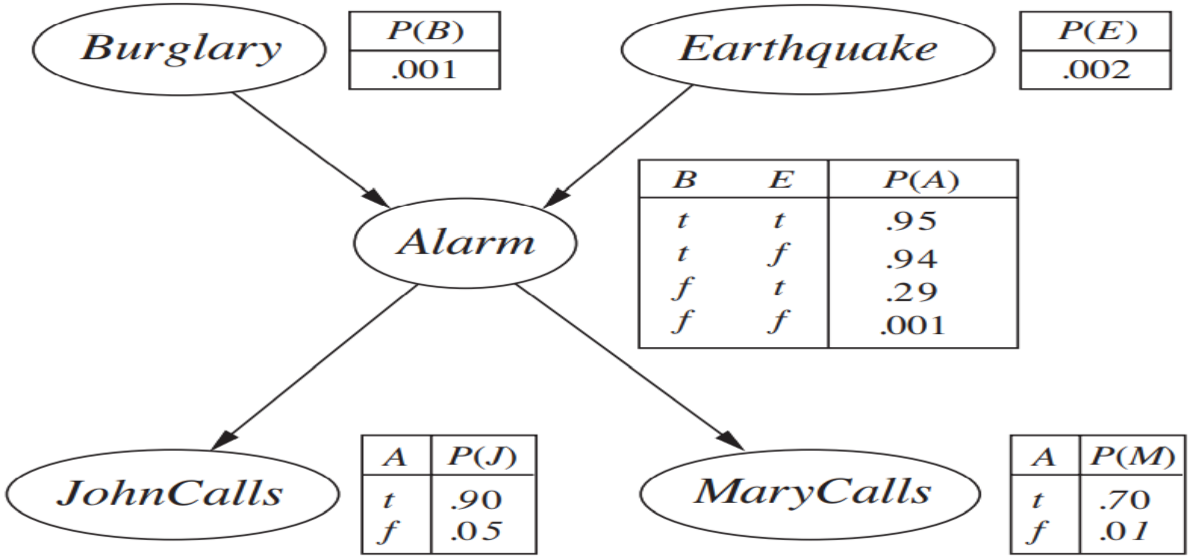
A	P(J)
t	.90
f	.05

A	P(M)
t	.70
f	.01

Figure 14.2 A typical Bayesian network, showing both the topology and the conditional probability tables (CPTs). In the CPTs, the letters B, E, A, J, and M stand for *Burglary*, *Earthquake*, *Alarm*, *JohnCalls*, and *MaryCalls*, respectively.

The semantics of Bayesian Network:

1. What is the probability that the alarm has sounded but neither a burglary nor an earthquake has occurred, and both John and Merry call?



Solution:

$$P(j \wedge m \wedge a \wedge \neg b \wedge \neg e) = P(j|a)P(m|a)P(a|\neg b, \neg e)P(\neg b)P(\neg e)$$

$$= 0.90 \times 0.70 \times 0.001 \times 0.999 \times 0.998$$

$$= 0.00062$$

Exact Inference in Bayesian Networks:

- The graphical independence representation

    - yields an efficient inference scheme

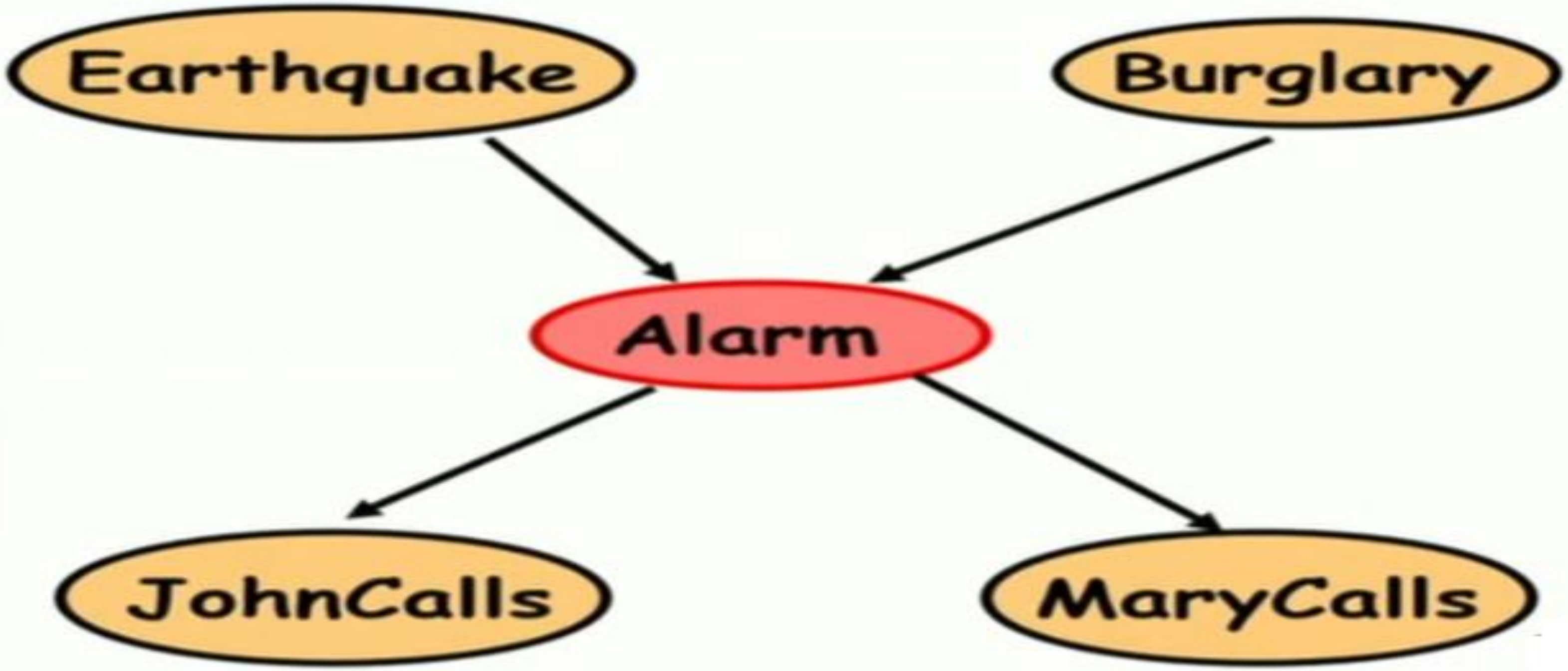
- We generally want to compute

    - Marginal probability *Pr(Z),*

    - *Pr(Z|**E**)* where **E** is (conjunctive) evidence.

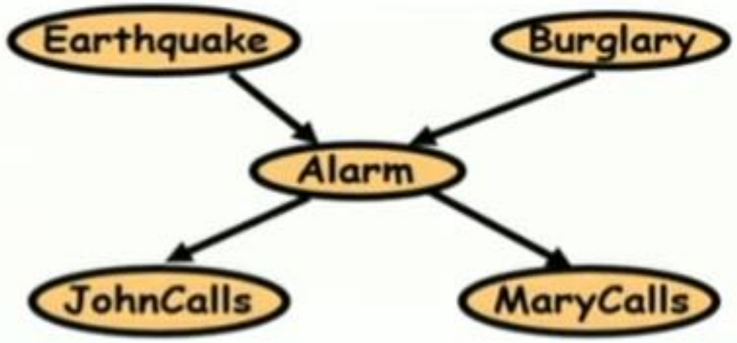
    - Z : query variable(s),
    - E : evidence variable(s),
    - everything else: hidden variable

$P(B \mid J=\text{true}, M=\text{true})$



$$P(b|j,m) = \alpha \sum_{e,a} P(b,j,m,e,a)$$

$$P(B|j,m) = \frac{P(B,j,m)}{P(j,m)}$$



$$= \alpha P(B,j,m)$$

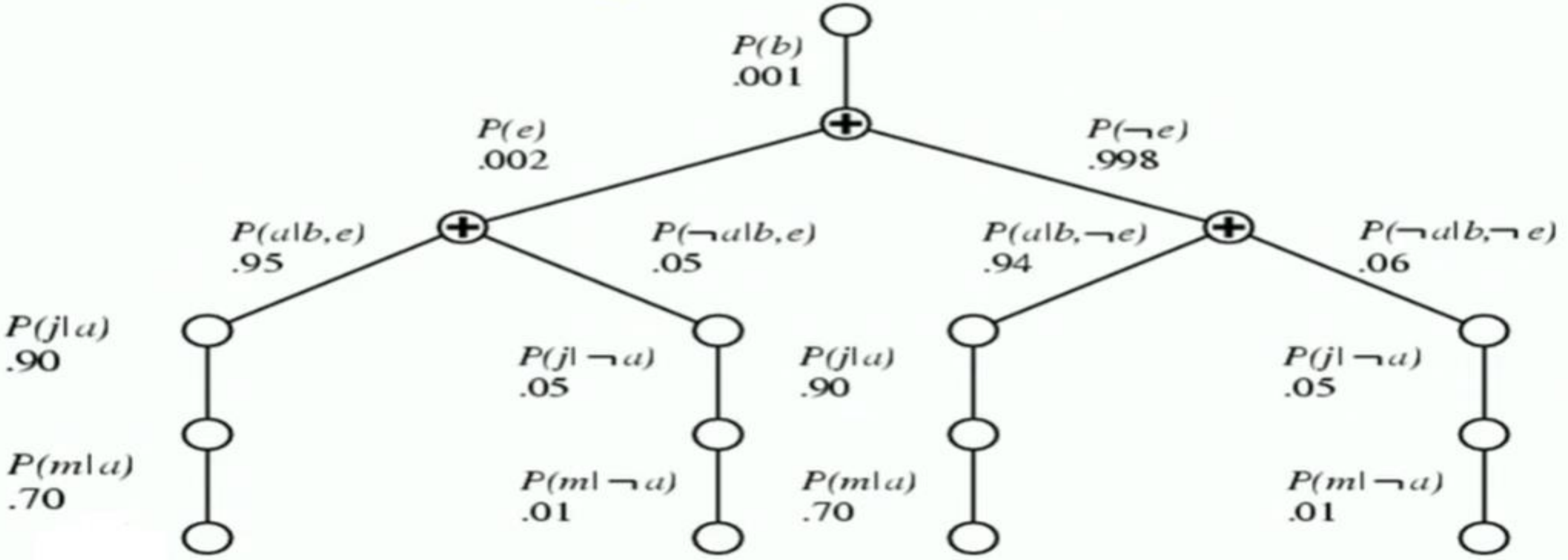
$$= \alpha \textstyle\sum_{E,A} P(B,E,A,j,m)$$

$$= \alpha \sum_{E,A} P(B)P(E)P(A|E,B)P(j|A)P(m|A)$$

$$= \alpha P(B) \sum_{E} P(E) \sum_{A} P(A|E,B)P(j|A)P(m|A)$$

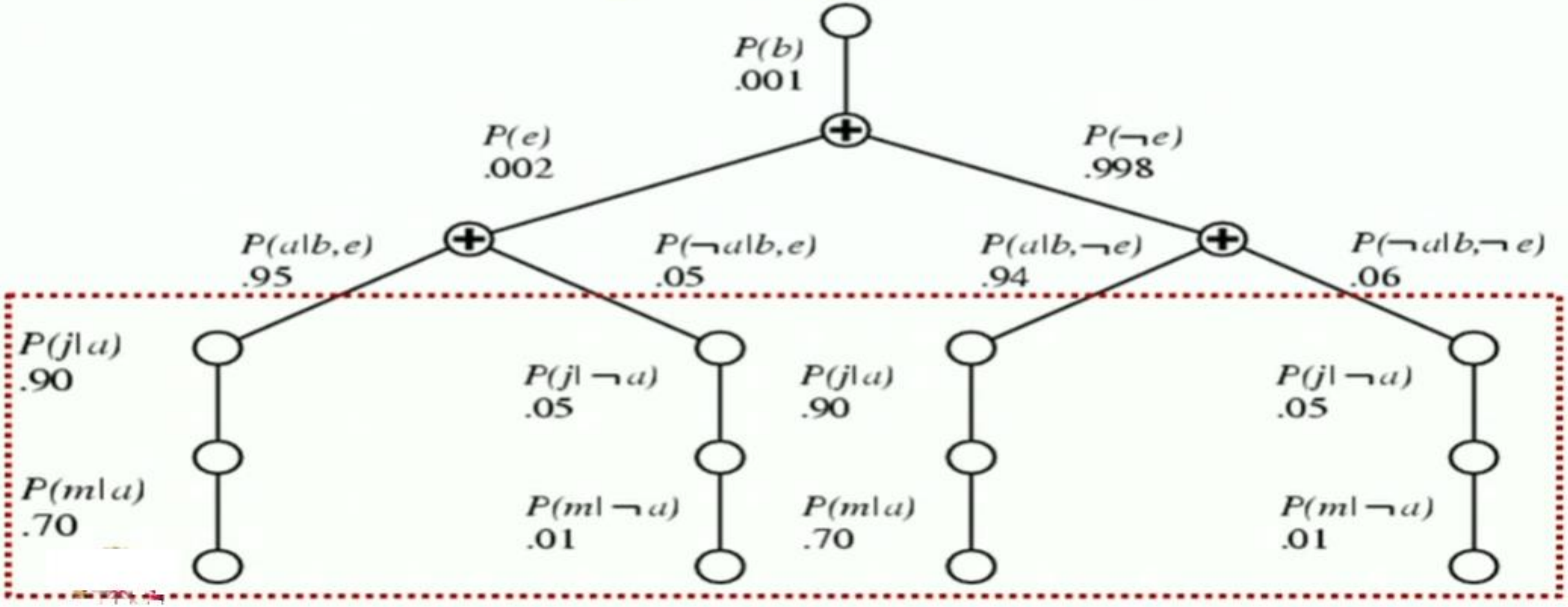
Variable elimination:

$$P(b|j,m) = \alpha P(b) \sum_E P(E) \sum_A P(A|B,E)P(j|A)P(m|A)$$



Variable Elimination

$$P(b|j,m) = \alpha P(b) \sum_E P(E) \sum_A P(A|B,E)P(j|A)P(m|A)$$



Variable Elimination

- A *factor* is a function from some set of variables into a specific value: e.g., *f(E,A,N1)*
  - CPTs are factors, e.g., *P(A|E,B)* function of *A,E,B*
- VE works by *eliminating* all variables in turn until there is a factor with only query variable
- To eliminate a variable:
  - *join* all factors containing that variable (like DB)
  - *sum out* the influence of the variable on new factor
  - exploits product form of joint distribution

Example of VE: P(JC)

$P(J)$

$= \Sigma_{M,A,B,E}\ P(J,M,A,B,E)$

$= \Sigma_{M,A,B,E}\ P(J|A)P(M|A)\ P(B)P(A|B,E)P(E)$

$= \Sigma_A P(J|A)\ \Sigma_M P(M|A)\ \Sigma_B P(B)\ \Sigma_E P(A|B,E)P(E)$

$= \Sigma_A P(J|A)\ \Sigma_M P(M|A)\ \Sigma_B P(B)\ f1(A,B)$

$= \Sigma_A P(J|A)\ \Sigma_M P(M|A)\ f2(A)$

$= \Sigma_A P(J|A)\ f3(A)$

$= f4(J)$

$$\alpha P(B) \sum_{E} P(E) \sum_{A} P(A|E,B) P(j|A) P(m|A)$$

	Pr(J|A)
a	0.9 (0.1)
$\bar{a}$	0.05 (0.95)

	Pr(M|A)
a	0.7 (0.3)
$\bar{a}$	0.01 (0.99)

	Pr(j|A)P(m|A)
a	0.9x0.7
$\bar{a}$	0.05x0.01

$$\alpha P(B) \sum_E P(E) \sum_A P(A|E,B) f1(A)$$

	Pr(J|A)
a	0.9 (0.1)
$\bar{a}$	0.05 (0.95)

	Pr(M|A)
a	0.7 (0.3)
$\bar{a}$	0.01 (0.99)

	f1(A)
a	0.63
$\bar{a}$	0.0005

$$\alpha P(B) \sum_{E} P(E) \sum_{A} P(A|E,B) f1(A)$$

	f1(A)
a	0.63
$\bar{a}$	0.0005

	Pr(A|E,B)
e,b	0.95 (0.05)
e,$\bar{b}$	0.29 (0.71)
$\bar{e}$,b	0.94 (0.06)
$\bar{e}$,$\bar{b}$	0.001 (0.999)

e,b	0.95x0.63+0.05x0.0005
e,$\bar{b}$	0.29x0.63+0.71x0.0005
$\bar{e}$,b	0.94x0.63+0.06x0.0005
$\bar{e}$,$\bar{b}$	0.001x0.63+0.999x0.0005

$$\alpha P(B) \sum_{E} P(E) f2(E, B)$$

	f1(A)
a	0.63
$\bar{a}$	0.0005

	Pr(A|E,B)
e,b	0.95 (0.05)
e,$\bar{b}$	0.29 (0.71)
$\bar{e}$,b	0.94 (0.06)
$\bar{e}$,$\bar{b}$	0.001 (0.999)

	f2(E,B)
e,b	0.60
e,$\bar{b}$	0.18
$\bar{e}$,b	0.59
$\bar{e}$,$\bar{b}$	0.001

αf3(B)

Pr(E=t)	Pr(E=f)
0.002	0.998

Pr(B=t)	Pr(B=f)
0.001	0.999

	f2(E,B)
e,b	0.60
e,$\bar{b}$	0.18
$\bar{e}$,b	0.59
$\bar{e}$,$\bar{b}$	0.001

	f3(B)
b	0.0006
$\bar{b}$	0.0013

$$\alpha f3(B) \rightarrow P(B|j, m)$$

	f3(B)
b	0.0006
$\bar{b}$	0.0013

$$N = 0.0006 + 0.0013$$
$$= 0.0019$$

	P(B|j,m)
b	0.32
$\bar{b}$	0.68