國立成功大學
National Cheng Kung University
1931

Introduction to

Artificial Intelligence

Chapter 12
Quantifying Uncertainty

Wei-Ta Chu (朱威達)

Acting Under Uncertainty

- Agents may need to handle *uncertainty*, whether due to partial observability, nondeterminism, or a combination of the two. An agent may never know for certain what state it's in or where it will end up after a sequence of actions.
- An example of uncertain reasoning: diagnosing a dental patient's toothache. Let us try to write rules for dental diagnosis using propositional logic, so that we can see how the logical approach breaks down.

Acting Under Uncertainty

- Consider the following simple rule:

$$Toothache \Rightarrow Cavity$$ (蛀牙)

- The problem is that this rule is wrong. Not all patients with toothaches have cavities; some of them have gum disease, an abscess, or one of several other problems:

$$Toothache \Rightarrow Cavity \lor GumProblem \lor Abscess \ldots$$ (潰瘍)

- Unfortunately, in order to make the rule true, we have to add an almost unlimited list of possible problems.

Acting Under Uncertainty

- We could try turning the rule into a causal rule:

$$Cavity \Rightarrow Toothache$$

- But this rule is not right either; not all cavities cause pain. The only way to fix the rule is to make it logically exhaustive: to augment the left-hand side with all the qualifications required for a cavity to cause a toothache.

Acting Under Uncertainty

- Trying to use logic to cope with a domain like medical diagnosis thus fails for three main reasons:
  - Laziness: It is too much work to list the complete set of antecedents (前項) or consequents (後項).
  - Theoretical ignorance: Medical science has no complete theory for the domain.
  - Practical ignorance: Even if we know all the rules, we might be uncertain about a particular patient because not all the necessary tests have been or can be run.

Acting Under Uncertainty

- Our main tool for dealing with degrees of belief is *probability theory*.
- Probability provides a way of summarizing the uncertainty that comes from our laziness and ignorance, thereby solving the qualification problem.
- This belief could be derived from statistical data—80% of the toothache patients seen so far have had cavities—or from some general dental knowledge, or from a combination of evidence.

Acting Under Uncertainty

- Probability statements are made with respect to a knowledge state, not with respect to the real world. We say “The probability that the patient has a cavity, given that she has a toothache, is 0.8.”
- If we later learn that the patient has a history of gum disease, we can make a different statement: “The probability that the patient has a cavity, given that she has a toothache and a history of gum disease, is 0.4.”
- If we gather further conclusive evidence against a cavity, we can say “The probability that the patient has a cavity, given all we now know, is almost 0.” Note that these statements do not contradict each other; each is a separate assertion about a different knowledge state.

Uncertainty and rational decisions

- An agent must first have **preferences** between the different possible **outcomes** of the various plans.
- We use **utility theory** to represent and reason with preferences. (The term **utility** is used here in the sense of "the quality of being useful".)
- Utility theory says that every state has a degree of usefulness, or utility, to an agent and that the agent will prefer states with higher utility.

Uncertainty and rational decisions

- Preferences, as expressed by utilities, are combined with probabilities in the general theory of rational decisions called **decision theory**:

$$\mathit{Decision\ theory} = \mathit{probability\ theory} + \mathit{utility\ theory}$$

- The fundamental idea of decision theory is that *an agent is rational if and only if it chooses the action that yields the highest expected utility, averaged over all the possible outcomes of the action*. This is called the principle of **maximum expected utility** (MEU).

What probabilities are about

- Probabilistic assertions (主張) talk about how probable the various worlds are.
- In probability theory, the set of all possible worlds is called the *sample space*. The possible worlds are mutually exclusive and exhaustive.
- For example, if we are about to roll two (distinguishable) dice, there are 36 possible worlds to consider: (1,1), (1,2), ..., (6,6). The Greek letter Ω (uppercase omega) is used to refer to the sample space, and *ω* (lowercase omega) refers to elements of the space, that is, particular possible worlds.

What probabilities are about

- A fully specified probability model associates a numerical probability $P(\omega)$ with each possible world. The basic axioms of probability theory:

$$0 \leq P(\omega) \leq 1 \text{ for every } \omega \text{ and } \sum_{\omega \in \Omega} P(\omega) = 1$$

- For example, if we assume that each die is fair and the rolls don't interfere with each other, then each of the possible worlds (1,1), (1,2), ..., (6,6) has probability 1/36.

What probabilities are about

- Probabilistic assertions and queries are not usually about particular possible worlds, but about sets of them. For example, we might be interested in the cases where the two dice add up to 11, the cases where doubles are rolled, and so on. In probability theory, these sets are called **events**. In AI, the sets are always described by **propositions** (命題) in a formal language.
- For each proposition, the corresponding set contains just those possible worlds in which the proposition holds. The probability associated with a proposition is defined to be the sum of the probabilities of the worlds in which it holds:

$$\text{For any proposition } \phi,\ P(\phi) = \sum_{\omega \in \phi} P(\omega) \qquad (13.2)$$

What probabilities are about

- For example, when rolling fair dice, we have
  $P(\text{Total} = 11) = P((5, 6)) + P((6, 5)) = 1/36 + 1/36 = 1/18.$
- Probabilities such as $P(\text{Total} = 11)$ and $P(\text{doubles})$ are called **unconditional** or **prior probabilities** (and sometimes just “priors” for short); they refer to degrees of belief in propositions *in the absence of any other information*.

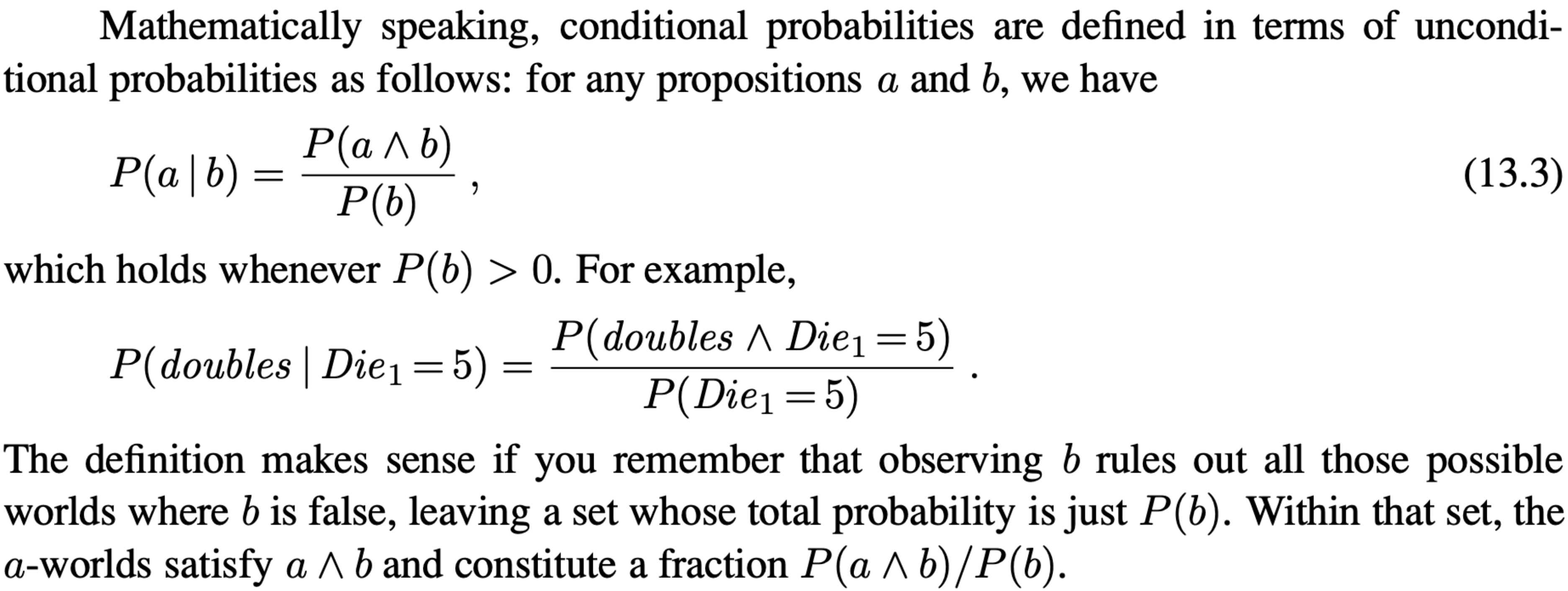
What probabilities are about

- Most of the time, however, we have some information, usually called **evidence**, that has already been revealed. For example, the first die may already be showing a 5.
- In that case, we are interested not in the unconditional probability of rolling doubles, but the **conditional** or **posterior** probability (or just "posterior" for short) of rolling doubles given that the first die is a 5. This probability is written $P(\text{doubles}|\text{Die}_1{=}5)$, where the " | " is pronounced "given."

What probabilities are about

- Similarly, if I am going to the dentist for a regular checkup, the probability $P(cavity) = 0.2$ might be of interest; but if I go to the dentist because I have a toothache, it's $P(cavity \mid toothache) = 0.6$ that matters.
- It is important to understand that $P(cavity)$=0.2 is still valid after toothache is observed. When making decisions, an agent needs to condition on all the evidence it has observed.

What probabilities are about



Mathematically speaking, conditional probabilities are defined in terms of unconditional probabilities as follows: for any propositions a and b, we have

$$P(a \mid b) = \frac{P(a \wedge b)}{P(b)}\,, \tag{13.3}$$

which holds whenever $P(b) > 0$. For example,

$$P(doubles \mid Die_1 = 5) = \frac{P(doubles \wedge Die_1 = 5)}{P(Die_1 = 5)}\,.$$

The definition makes sense if you remember that observing b rules out all those possible worlds where b is false, leaving a set whose total probability is just $P(b)$. Within that set, the a-worlds satisfy $a \wedge b$ and constitute a fraction $P(a \wedge b)/P(b)$.

What probabilities are about

- The definition of conditional probability, Equation (13.3), can be written in a different form called the **product rule**:

$$P(a \wedge b) = P(a \mid b)P(b)$$

- The product rule is perhaps easier to remember: it comes from the fact that, for *a* and *b* to be true, we need *b* to be true, and we also need *a* to be true given *b*.

Lang. of propositions in prob. assertions

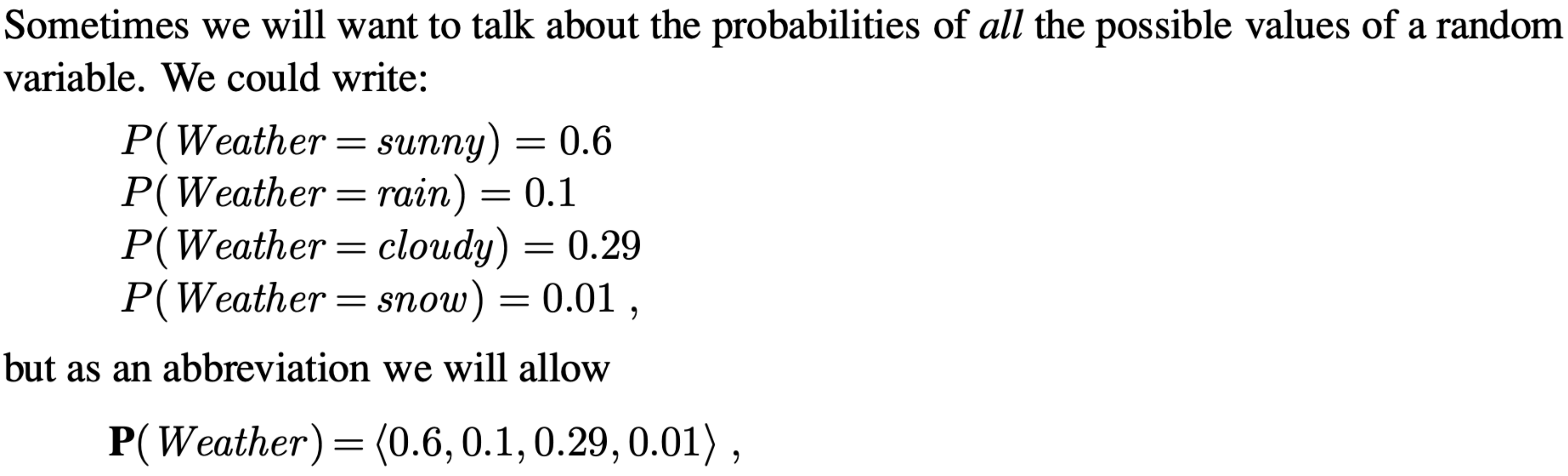
- Variables in probability theory are called **random variables** and their names begin with an uppercase letter. In the dice example, *Total* and Die_1 are random variables. Every random variable has a **domain**—the set of possible values it can take on. The domain of *Total* for two dice is the set {2,...,12} and the domain of Die_1 is {1,...,6}.
- A Boolean random variable has the domain {*true*, *false*}. For example, the proposition that doubles are rolled can be written as *Doubles* = *true*. By convention, propositions of the form *A*=*true* are abbreviated simply as *a*, while *A*=*false* is abbreviated as ¬*a*.

Lang. of propositions in prob. assertions

- Variables can have infinite domains—either discrete (like the integers) or continuous (like the reals).
- Finally, we can combine these sorts of elementary propositions by using the connectives of propositional logic. For example, we can express “The probability that the patient has a cavity, given that she is a teenager with no toothache, is 0.1” as follows:

$$P(cavity \mid \neg toothache \wedge teen) = 0.1$$

Lang. of propositions in prob. assertions



Sometimes we will want to talk about the probabilities of *all* the possible values of a random variable. We could write:

$$
\begin{aligned}
&P(Weather = sunny) = 0.6 \\
&P(Weather = rain) = 0.1 \\
&P(Weather = cloudy) = 0.29 \\
&P(Weather = snow) = 0.01 \ ,
\end{aligned}
$$

but as an abbreviation we will allow

$$\mathbf{P}(Weather) = \langle 0.6, 0.1, 0.29, 0.01 \rangle \ ,$$

- where the bold **P** indicates that the result is a vector of numbers. We say that the **P** statement defines a **probability distribution** for the random variable *Weather*. The **P** notation is also used for conditional distributions: $\mathbf{P}(X \mid Y)$ gives the values of $\mathbf{P}(X = x_i \mid Y = y_j)$ for each possible i, j pair.

Lang. of propositions in prob. assertions

- For continuous variables, it is not possible to write out the entire distribution as a vector, because there are infinitely many values. Instead, we can define the probability that a random variable takes on some value *x* as a parameterized function of *x*. For example, the sentence

$$P(NoonTemp = x) = Uniform_{[18C,26C]}(x)$$

expresses the belief that the temperature at noon is distributed uniformly between 18 and 26 degrees Celsius. We call this a **probability density function**.

Lang. of propositions in prob. assertions

- Probability density functions (sometimes called **pdfs**) differ in meaning from discrete distributions. Saying that the probability density is uniform from 18C to 26C means that there is a 100% chance that the temperature will fall somewhere in that 8C-wide region and a 50% chance that it will fall in any 4C-wide region, and so on.
- We write the probability density for a continuous random variable *X* at value *x* as *P*(*X* = *x*) or just *P*(*x*); the intuitive definition of *P*(*x*) is the probability that *X* falls within an arbitrarily small region beginning at *x*, divided by the width of the region: $P(x) = \lim_{dx \to 0} P(x \leq X \leq x + dx)/dx$

Lang. of propositions in prob. assertions

- In addition to distributions on single variables, we need notation for distributions on multiple variables. Commas are used for this. For example, **P**(*Weather*, *Cavity*) denotes the probabilities of all combinations of the values of *Weather* and *Cavity*. This is a 4×2 table of probabilities called the **joint probability distribution** of *Weather* and *Cavity*.
- We can also mix variables with and without values; **P**(*sunny*, *Cavity*) would be a two-element vector giving the probabilities of a sunny day with a cavity and a sunny day with no cavity.

Lang. of propositions in prob. assertions

- The product rules for all possible values of Weather and Cavity can be written as a single equation:

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

  instead of as these 4 × 2 = 8 equations (using abbreviations W and C):

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

Lang. of propositions in prob. assertions

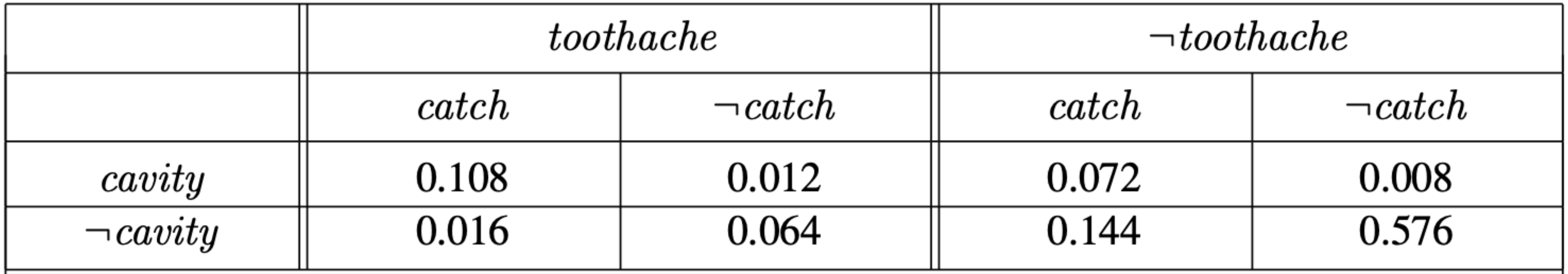
- As a degenerate case, **P**(*sunny*, *cavity*) has no variables and thus is a one-element vector that is the probability of a sunny day with a cavity, which could also be written as *P*(*sunny*, *cavity*) or *P*(*sunny* ∧ *cavity*).

Lang. of propositions in prob. assertions

- A possible world is defined to be an assignment of values to all of the random variables under consideration. For example, if the random variables are *Cavity*, *Toothache*, and *Weather*, then there are 2 × 2 × 4 = 16 possible worlds.
- A probability model is completely determined by the joint distribution for all of the random variables. For example, if the variables are *Cavity*, *Toothache*, and *Weather*, then the full joint distribution is given by **P**(*Cavity*, *Toothache*, *Weather*). This joint distribution can be represented as a 2 × 2 × 4 table with 16 entries.

Inference Using Full Joint Distributions

- We describe a simple method for **probabilistic inference**—that is, the computation of posterior probabilities for query propositions given observed evidence.
- We begin with a simple example: a domain consisting of just the three Boolean variables *Toothache*, *Cavity*, and *Catch*. The full joint distribution is a 2 × 2 × 2 table as shown in Figure 13.3.



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

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

Inference Using Full Joint Distributions

- Notice that the probabilities in the joint distribution sum to 1, as required by the axioms of probability. Equation (13.2) gives us a direct way to calculate the probability: 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* ∨ *toothache* holds:

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

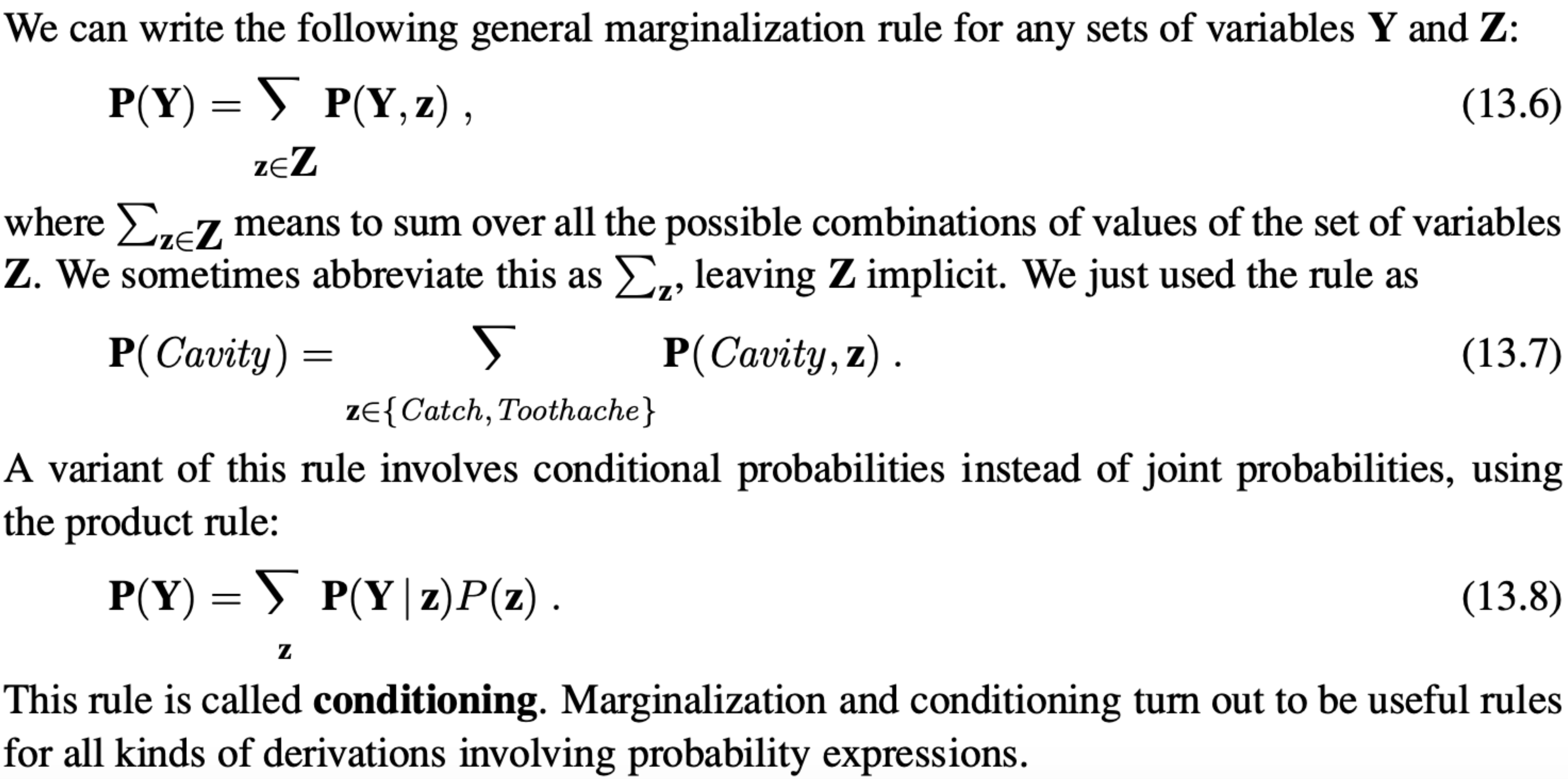
Inference Using Full Joint Distributions

- One particularly common task is to extract the distribution over some subset of variables or a single variable. For example, adding the entries in the first row gives the unconditional or **marginal probability** of cavity:

$$P(cavity) = 0.108 + 0.012 + 0.072 + 0.008 = 0.2$$

- This process is called **marginalization**, or **summing out**—because we sum up the probabilities for each possible value of the other variables, thereby taking them out of the equation.

Inference Using Full Joint Distributions



We can write the following general marginalization rule for any sets of variables $\mathbf{Y}$ and $\mathbf{Z}$:

$$\mathbf{P}(\mathbf{Y}) = \sum_{\mathbf{z}\in\mathbf{Z}} \mathbf{P}(\mathbf{Y},\mathbf{z}) \, , \tag{13.6}$$

where $\sum_{\mathbf{z}\in\mathbf{Z}}$ means to sum over all the possible combinations of values of the set of variables $\mathbf{Z}$. We sometimes abbreviate this as $\sum_{\mathbf{z}}$, leaving $\mathbf{Z}$ implicit. We just used the rule as

$$\mathbf{P}(Cavity) = \sum_{\mathbf{z}\in\{Catch, Toothache\}} \mathbf{P}(Cavity, \mathbf{z}) \, . \tag{13.7}$$

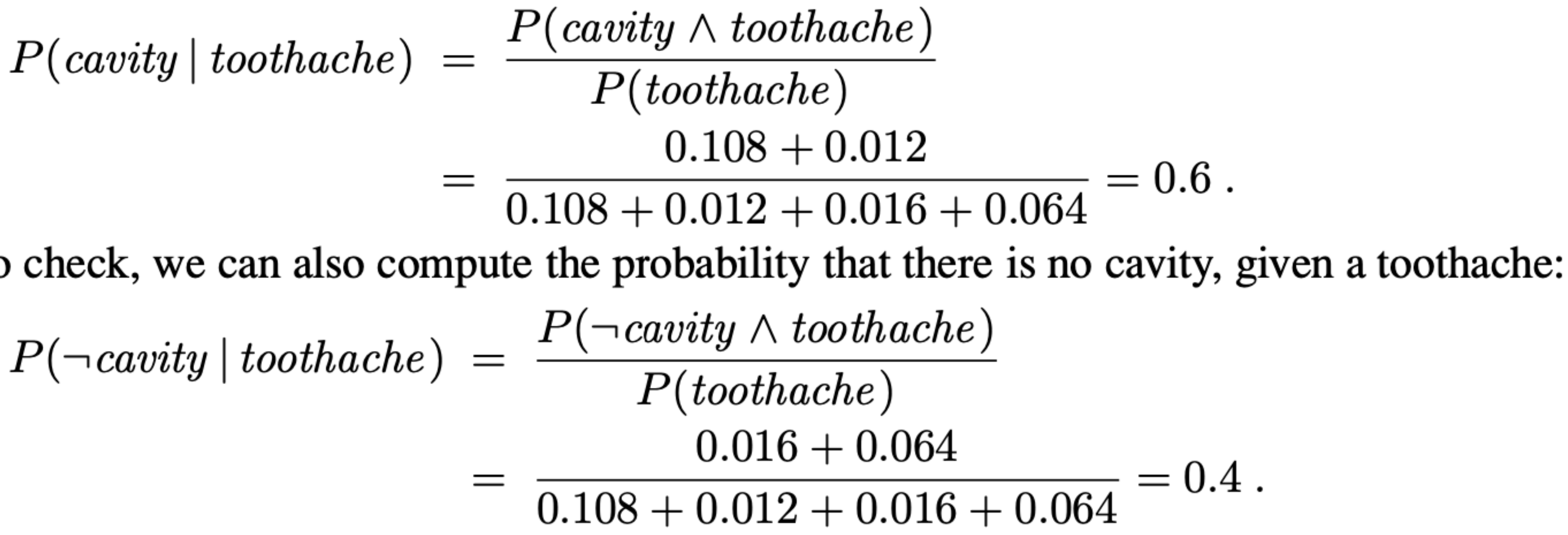
A variant of this rule involves conditional probabilities instead of joint probabilities, using the product rule:

$$\mathbf{P}(\mathbf{Y}) = \sum_{\mathbf{z}} \mathbf{P}(\mathbf{Y} \mid \mathbf{z}) P(\mathbf{z}) \, . \tag{13.8}$$

This rule is called **conditioning**. Marginalization and conditioning turn out to be useful rules for all kinds of derivations involving probability expressions.

Inference Using Full Joint Distributions

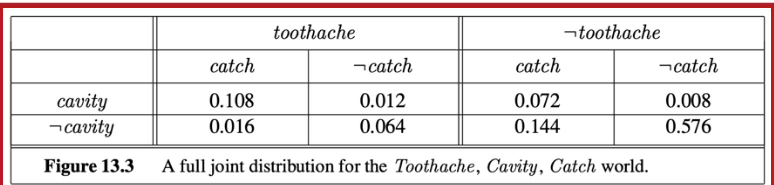
- In most cases, we are interested in computing conditional probabilities of some variables, given evidence about others. For example, we can compute the probability of a cavity, given evidence of a toothache, as follows:



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

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

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



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

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

Inference Using Full Joint Distributions

- 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. It can be viewed as a normalization constant for the distribution **P**(*Cavity* | *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:

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

Inference Using Full Joint Distributions

- In other words, we can calculate **P**(*Cavity* | *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 ¬*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.

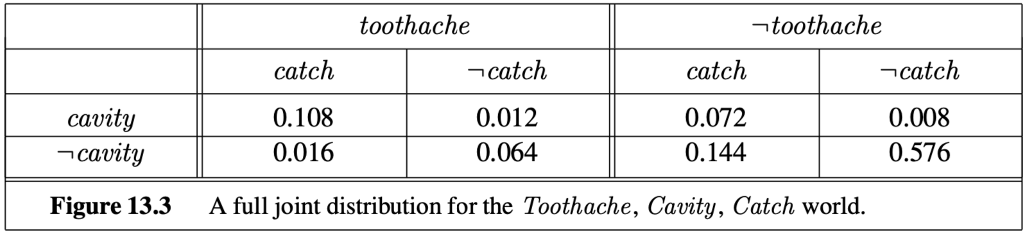
Inference Using Full Joint Distributions

- If the query involves a single variable, X (*Cavity* in the example). Let **E** be the list of evidence variables (just *Toothache* in the example), let **e** be the list of observed values for them, and let **Y** be the remaining unobserved variables (just *Catch* in the example). The query is **P**(X | **e**) and can be evaluated as

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

where the summation is over all possible **y**'s (i.e., all possible combinations of values of the unobserved variables **Y**). Notice that together the variables X, **E**, and **Y** constitute the complete set of variables for the domain, so **P**(X, **e**, **y**) is simply a subset of probabilities from the full joint distribution.

Independence



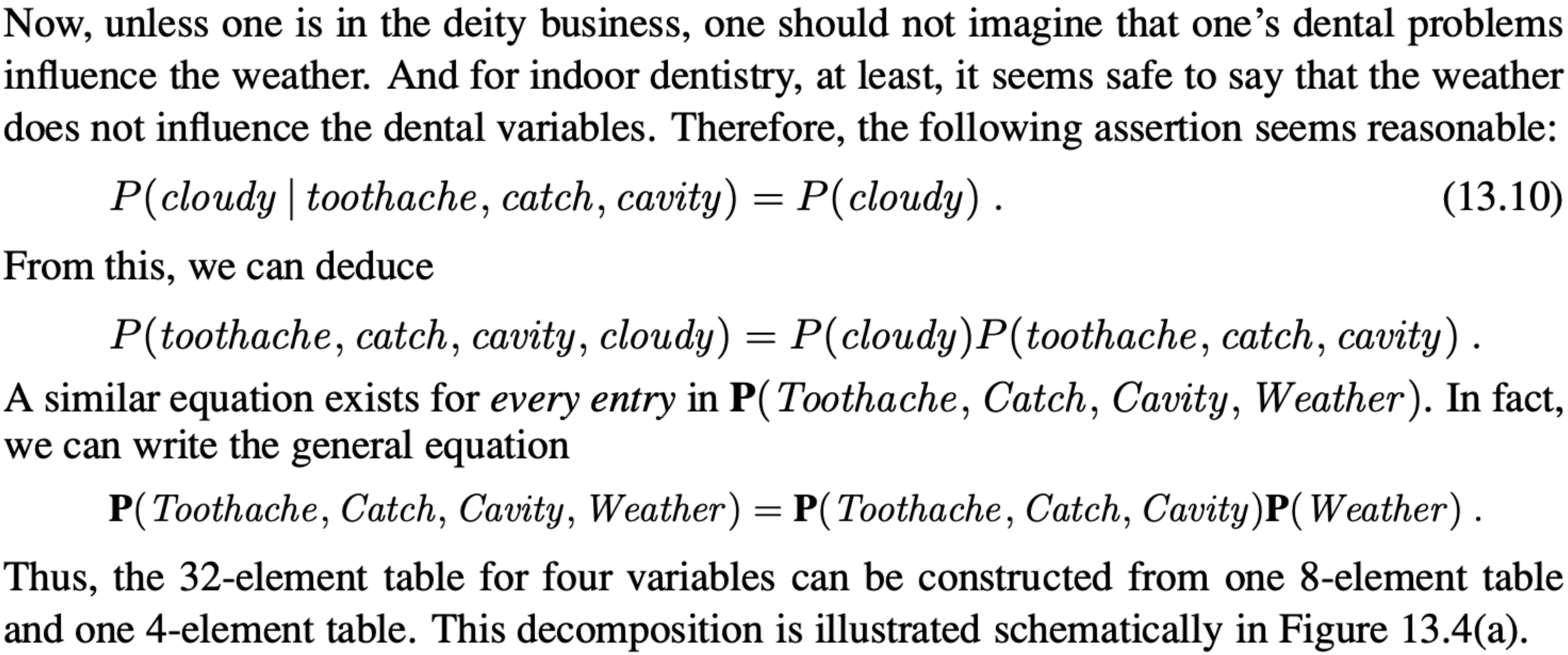
	toothache		*¬toothache*	
	catch	*¬catch*	*catch*	*¬catch*
cavity	0.108	0.012	0.072	0.008
¬cavity	0.016	0.064	0.144	0.576

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

- Let us expand the full joint distribution in Figure 13.3 by adding a fourth variable, *Weather*. The full joint distribution then becomes **P**(*Toothache*, *Catch*, *Cavity*, *Weather*), which has 2 × 2 × 2 × 4 = 32 entries. It contains four "editions" of the table shown in Figure 13.3, one for each kind of weather.
- How are *P*(*toothache*, *catch*, *cavity*, *cloudy*) and *P*(*toothache*, *catch*, *cavity*) related? We can use the product rule:

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

Independence



Now, unless one is in the deity business, one should not imagine that one's dental problems influence the weather. And for indoor dentistry, at least, it seems safe to say that the weather does not influence the dental variables. Therefore, the following assertion seems reasonable:

$$P(cloudy \mid toothache, catch, cavity) = P(cloudy) . \tag{13.10}$$

From this, we can deduce

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

A similar equation exists for *every entry* in $\mathbf{P}(Toothache, Catch, Cavity, Weather)$. In fact, we can write the general equation

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

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).

Independence

- The property we used in Equation (13.10) is called **independence** (also **marginal independence** and **absolute independence**). In particular, the weather is independent of one's dental problems. Independence between propositions *a* and *b* can be written as

$$P(a \mid b) = P(a) \quad \text{or} \quad P(b \mid a) = P(b) \quad \text{or} \quad P(a \land b) = P(a)P(b) \qquad (13.11)$$

- All these forms are equivalent. Independence between variables *X* and *Y* can be written as follows:

$$\mathbf{P}(X \mid Y) = \mathbf{P}(X) \quad \text{or} \quad \mathbf{P}(Y \mid X) = \mathbf{P}(Y) \quad \text{or} \quad \mathbf{P}(X, Y) = \mathbf{P}(X)\mathbf{P}(Y)$$

Independence

- Independence assertions are usually based on knowledge of the domain. As the toothache– weather example illustrates, they can dramatically reduce the amount of information necessary to specify the full joint distribution. If the complete set of variables can be divided into independent subsets, then the full joint distribution can be factored into separate joint distributions on those subsets.
- For example, the full joint distribution on the outcome of n independent coin flips, $P(C_1, \ldots, C_n)$, has 2^n entries, but it can be represented as the product of n single-variable distributions $P(C_i)$.

Independence

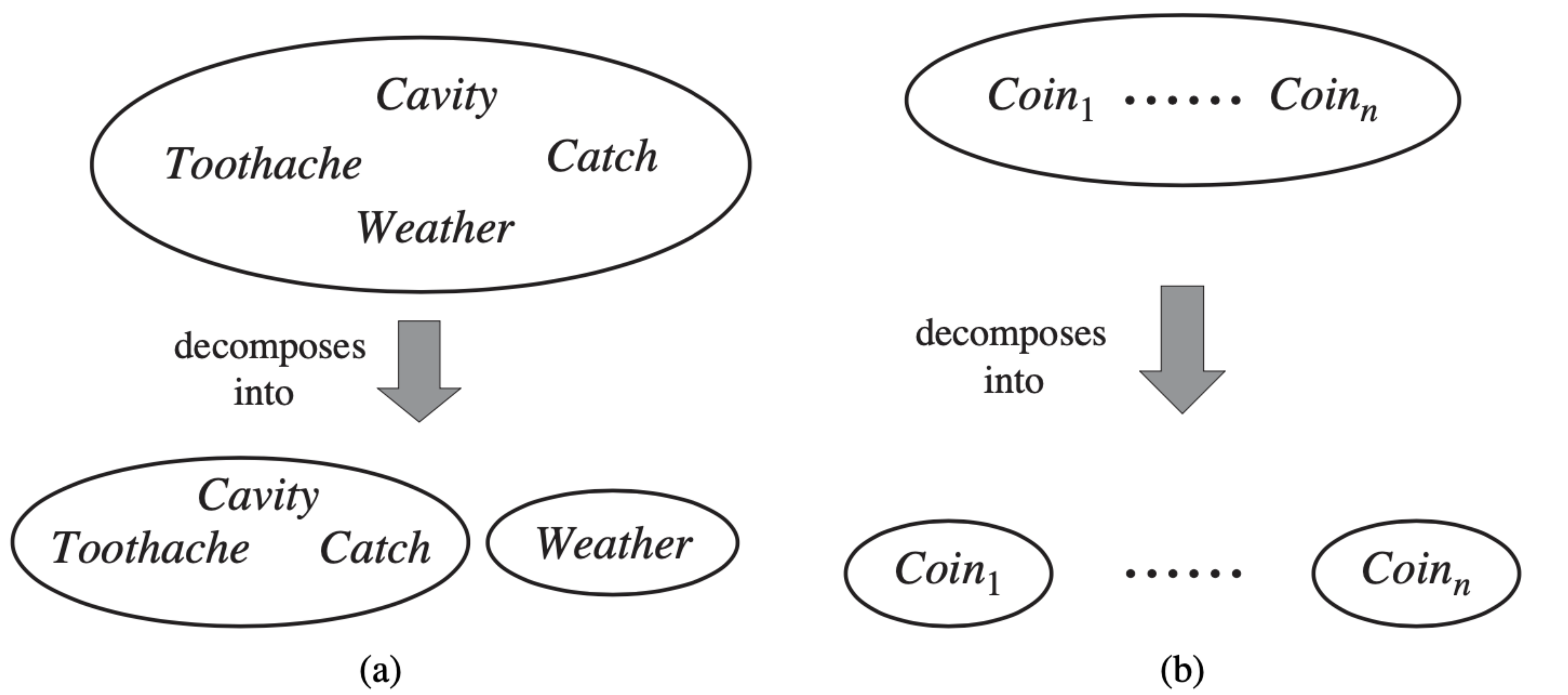


Figure 13.4 Two examples of factoring a large joint distribution into smaller distributions, using absolute independence. (a) Weather and dental problems are independent. (b) Coin flips are independent.

Bayes' Rule

- Product rule can actually be written in two forms:

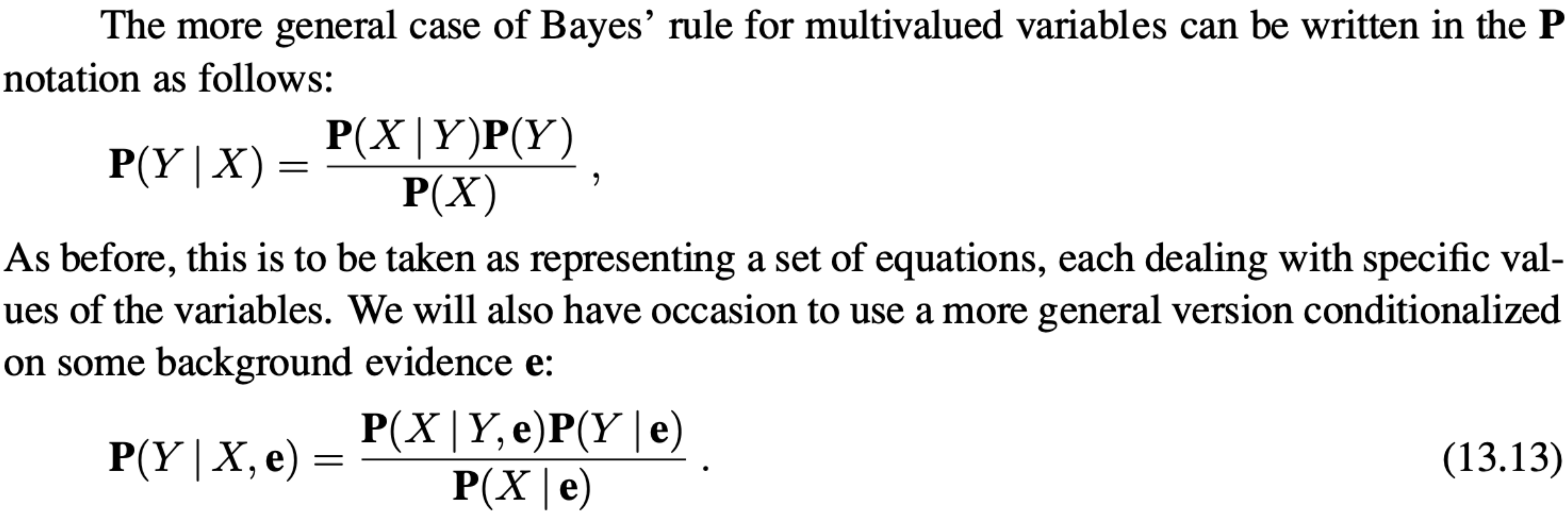
$$P(a \wedge b) = P(a \mid b)P(b) \qquad \text{and} \qquad P(a \wedge b) = P(b \mid a)P(a)$$

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

$$P(b \mid a) = \frac{P(a \mid b)P(b)}{P(a)}$$

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

Bayes' Rule



The more general case of Bayes' rule for multivalued variables can be written in the **P** notation as follows:

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

As before, this is to be taken as representing a set of equations, each dealing with specific values of the variables. We will also have occasion to use a more general version conditionalized on some background evidence **e**:

$$\mathbf{P}(Y \mid X, \mathbf{e}) = \frac{\mathbf{P}(X \mid Y, \mathbf{e})\mathbf{P}(Y \mid \mathbf{e})}{\mathbf{P}(X \mid \mathbf{e})} . \qquad (13.13)$$

Applying Bayes' rule: The simple case

- On the surface, Bayes' rule does not seem very useful. It allows us to compute the single term $P(b \mid a)$ in terms of three terms: $P(a \mid b)$, $P(b)$, and $P(a)$. That seems like two steps backwards, but Bayes' rule is useful in practice because there are many cases where we do have good probability estimates for these three numbers and need to compute the fourth.
- Often, we perceive as evidence the *effect* of some unknown *cause* and we would like to determine that cause. In that case, Bayes' rule becomes

$$P(cause \mid effect) = \frac{P(effect \mid cause)P(cause)}{P(effect)}$$

Applying Bayes' rule: The simple case

- The conditional probability *P*(*effect* | *cause*) quantifies the relationship in the **causal** direction, whereas *P*(*cause* | *effect*) describes the **diagnostic** direction.
- In a task such as medical diagnosis, we often have conditional probabilities on causal relationships (that is, the doctor knows *P*(*symptoms* | *disease*)) and want to derive a diagnosis, *P*(*disease* | *symptoms*).
- For example, a doctor knows that the disease meningitis (腦膜炎) causes the patient to have a stiff (僵硬) 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%.

Applying Bayes' rule: The simple case

- Letting *s* be the proposition that the patient has a stiff neck and *m* be the proposition that the patient has meningitis, we have

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

- Notice that even though a stiff neck is quite strongly indicated by meningitis (with probability 0.7), the probability of meningitis in the patient remains small. This is because the prior probability of stiff necks is much higher than that of meningitis.

Applying Bayes' rule: The simple case

- One can avoid assessing the prior probability of the evidence (here, $P(s)$) by instead computing a posterior probability for each value of the query variable (here, m and $\neg m$) and then normalizing the results. The same process can be applied when using Bayes' rule. We have

$$\mathbf{P}(M \mid s) = \alpha \, \langle P(s \mid m)P(m), P(s \mid \neg m)P(\neg m)\rangle$$

- Thus, to use this approach we need to estimate $P(s \mid \neg m)$ instead of $P(s)$. There is no free lunch—sometimes this is easier, sometimes it is harder.

Applying Bayes' rule: The simple case

- One obvious question to ask about Bayes' rule is why one might have available the conditional probability in one direction, but not the other. In the meningitis domain, perhaps the doctor knows that a stiff neck implies meningitis in 1 out of 5000 cases; that is, the doctor has quantitative information in the diagnostic direction from symptoms to causes. Such a doctor has no need to use Bayes' rule.
- Unfortunately, diagnostic knowledge is often more fragile (脆弱) than causal knowledge. If there is a sudden epidemic of meningitis, the unconditional probability of meningitis, $P(m)$, will go up.

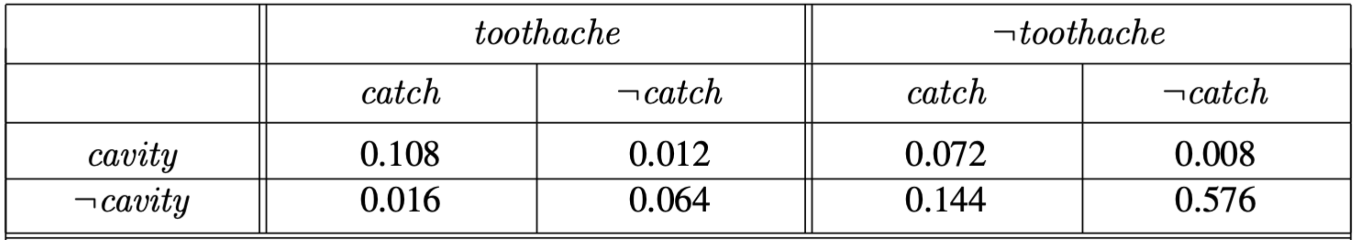
Applying Bayes’ rule: The simple case

- The doctor who derived the diagnostic probability $P(m \mid s)$ directly from statistical observation of patients before the epidemic will have no idea how to update the value, but the doctor who computes $P(m \mid s)$ from the other three values will see that $P(m \mid s)$ should go up proportionately with $P(m)$. Most important, the causal information $P(s \mid m)$ is unaffected by the epidemic, because it simply reflects the way meningitis works.

Using Bayes' rule: Combining evidence

- What happens when we have two or more pieces of evidence? For example, what can a dentist conclude if her nasty steel probe catches in the aching tooth of a patient? If we know the full joint distribution (Figure 13.3), we can read off the answer:

$$\mathbf{P}(Cavity \mid toothache \land catch) = \alpha \, \langle 0.108, 0.016 \rangle \approx \langle 0.871, 0.129 \rangle$$



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

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

Using Bayes' rule: Combining evidence

- We know, however, that such an approach does not scale up to larger numbers of variables. We can try using Bayes' rule to reformulate the problem:

$$\begin{aligned}&\mathbf{P}(Cavity \mid toothache \land catch)\\&\quad = \alpha\, \mathbf{P}(toothache \land catch \mid Cavity)\, \mathbf{P}(Cavity)\end{aligned}$$

- We need to know the conditional probabilities of the conjunction *toothache* ∧ *catch* for each value of *Cavity*.
- If there are *n* possible evidence variables (X rays, diet, oral hygiene, etc.), then there are 2^n possible combinations of observed values for which we would need to know conditional probabilities.

Using Bayes' rule: Combining evidence

- We need to find some additional assertions about the domain that will enable us to simplify the expressions. The notion of **independence** in Section 13.4 provides a clue. It would be nice if *Toothache* and *Catch* were independent, but they are not: if the probe catches in the tooth, then it is likely that the tooth has a cavity and that the cavity causes a toothache.
- *These variables are independent, however, given the presence or the absence of a cavity.*

Using Bayes' rule: Combining evidence

- Each is directly caused by the cavity, but neither has a direct effect on the other: toothache depends on the state of the nerves in the tooth, whereas the probe's accuracy depends on the dentist's skill, to which the toothache is irrelevant. This property is written as

$$\mathbf{P}(toothache \wedge catch \mid Cavity) = \mathbf{P}(toothache \mid Cavity)\mathbf{P}(catch \mid Cavity) \qquad (13.17)$$

- This equation expresses the **conditional independence** of *toothache* and *catch* given *Cavity*. We can plug it into Equation (13.16) to obtain the probability of a cavity:

$$\begin{aligned}&\mathbf{P}(Cavity \mid toothache \wedge catch) \qquad (13.18)\\ &\quad = \alpha\,\mathbf{P}(toothache \mid Cavity)\,\mathbf{P}(catch \mid Cavity)\,\mathbf{P}(Cavity)\end{aligned}$$

$$\boxed{\begin{aligned}&\mathbf{P}(Cavity \mid toothache \wedge catch)\\ &\quad = \alpha\,\mathbf{P}(toothache \wedge catch \mid Cavity)\,\mathbf{P}(Cavity)\end{aligned}}$$

Using Bayes' rule: Combining evidence

- The general definition of **conditional independence** of two variables *X* and *Y*, given a third variable *Z*, is

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

- In the dentist domain, it seems reasonable to assert conditional independence of the variables *Toothache* and *Catch*, given *Cavity*:

$$\mathbf{P}(Toothache, Catch \mid Cavity) = \mathbf{P}(Toothache \mid Cavity)\mathbf{P}(Catch \mid Cavity) \qquad (13.19)$$

Using Bayes' rule: Combining evidence

- Given the assertion in Equation (13.19), we can derive a decomposition:

$$\begin{aligned}&\mathbf{P}(Toothache, Catch, Cavity)\\ &\quad = \mathbf{P}(Toothache, Catch \mid Cavity)\mathbf{P}(Cavity) \quad \text{(product rule)}\\ &\quad = \mathbf{P}(Toothache \mid Cavity)\mathbf{P}(Catch \mid Cavity)\mathbf{P}(Cavity) \quad \text{(using 13.19).}\end{aligned}$$

- Conditional independence assertions can allow probabilistic systems to *scale up*; moreover, they are much more commonly available than absolute independence assertions.

Naïve Bayes Models

- The dentistry example illustrates a commonly occurring pattern in which a single cause directly influences a number of effects, all of which are conditionally independent, given the cause. The full joint distribution can be written as

$$\mathbf{P}(Cause, Effect_1, \ldots, Effect_n) = \mathbf{P}(Cause) \prod_i \mathbf{P}(Effect_i \mid Cause)$$

Naïve Bayes Models

- Such a probability distribution is called a **naive Bayes** model—“naive” because it is often used in cases where the “effect” variables are not strictly independent given the cause variable. (The naive Bayes model is sometimes called a **Bayesian classifier**.)
- In practice, naive Bayes systems often work very well, even when the conditional independence assumption is not strictly true.