Foundations of AI **Spring 2026**

Professor Serena Booth[1] **Knowledge Representation**

Probability Primer

Up until this point, we have worked with search problems where we have a known state space (X), start state (S), transition function (T), and goal state (G). These problems are well-defined, and we can solve them systematically. However, in the real world, things aren't always so clear-cut. Often, we don't know the exact state of the system (X) or how it will change over time (T). The physical world is filled with **uncertainty**, which makes reasoning and decision-making more challenging.

For the remainder of this class, we will deal with this uncertainty by introducing **probabilities**. Probabilities allow us to model situations where we don't have complete knowledge and make predictions based on partial information. In AI, handling uncertainty is fundamental, and probabilistic reasoning helps us make better decisions in complex environments.

1 Summarizing Uncertainty

Let's consider the example of diagnosing a dental patient with a toothache to explore uncertain reasoning. In any field—whether it's medicine, auto repair, or another—diagnosis often involves uncertainty. To understand where a logical approach might fall short, let's attempt to create rules for dental diagnosis using propositional logic. One might start with a basic rule like this:

$$\text{Toothache} \implies \text{Cavity}$$

However, this rule is flawed. Not every patient with a toothache has a cavity; some may suffer from gum disease, an abscess, or other issues. We could revise the rule as follows:

$$\text{Toothache} \implies \text{Cavity} \lor \text{GumProblem} \lor \text{Abscess} \ldots$$

Even this approach poses challenges, as it would require a long list of possible causes. Another option might be to reverse the rule:

$$\text{Cavity} \implies \text{Toothache}$$

But this, too, is inaccurate—not all cavities result in pain. To fully address this complexity, we would need to consider every condition that could cause a cavity to produce a toothache, making the rule excessively detailed. Moreover, medical knowledge does not provide complete theories for every condition, which adds to the difficulty.

Clearly, pure logic doesn't work well in such uncertain environments. Instead, we use **probability** to represent uncertainty, allowing us to handle this complexity more effectively.

Example: Consider the probability of flipping a coin and it landing heads, which is 0.5. What does this number represent? This probability, like all probabilities, is a value between 0 and 1. It reflects uncertainty by indicating that the outcome is not guaranteed to always be heads (true) or tails (false). Instead, it

[1] These notes were compiled by Professor Eric Ewing and Professor Amy Greenwald.

suggests that the coin will land heads half the time and tails the other half, capturing the idea that the result varies with each flip.

2 Conditional Probability

Let's revisit the question: Does a toothache always imply a cavity?

Not necessarily—a toothache could result from a cavity, gum issues, or even a chipped tooth. This is called the **qualification problem**, where the logical formulas needed to describe all possible outcomes would be incredibly long and complex.

Instead of using pure logic, we can use **conditional probability** to represent this uncertainty. Consider the following probability:

$$P(\text{toothache} \mid \text{cavity}) = \textit{some number}$$

This number represents the probability that a toothache will occur *given that you have a cavity.*

Definitions:

- **Prior Probability**: e.g., $P(\text{cavity}) = 0.2$: This is the probability of having a cavity **without considering additional information**. You can interpret this in two ways:
  - Out of 10 randomly sampled people, 2 would likely have a cavity (i.e., frequency interpretation).
  - If you have no prior knowledge about an individual, the probability that they have a cavity is 20% (i.e., probability represents *uncertainty*).
- **Posterior Probability**: e.g., $P(\text{cavity} \mid \text{toothache}) = 0.6$: This is the probability of having a cavity *given* that you have a toothache. This is a posterior probability because it **considers additional evidence** to update the likelihood. Note that it doesn't change the value of the prior.

For the rest of the semester, we will use this probabilistic foundation in machine learning and acting under uncertainty, where we predict probabilities (e.g., a certain class) based on available evidence (e.g., an image).

3 Joint Probability

Joint Probability refers to the probability of two events occurring together. For two random variables, say a and b, the joint probability is denoted as $P(a, b)$ or $P(a \wedge b)$, which represents the probability of both a and b happening simultaneously.

Example: Consider two independent events:

- a: The event that a die shows 3.
- b: The event that a coin lands heads.

The probability of rolling a 3 on a six-sided die is:

$$P(a) = \frac{1}{6}$$

The probability of flipping a heads on a fair coin is:

$$P(b) = \frac{1}{2}$$

Since the die roll and the coin flip are independent events, the joint probability is simply the product of the two probabilities:

$$P(a \wedge b) = P(a) \times P(b) = \frac{1}{6} \times \frac{1}{2} = \frac{1}{12}$$

Thus, the joint probability of the die showing 3 and the coin landing heads is $P(a \wedge b) = \frac{1}{12}$.

Example: Consider two dependent events:

- a: The event that the die shows 3.
- b: The event that the die shows an odd number.

Let's list out all the possible outcomes and count up the number of times both events are true. The possible outcomes for a six-sided die are: 1, 2, 3, 4, 5, and 6. The odd numbers are 1, 3, and 5. We want to calculate the joint probability $P(a \wedge b)$, which represents the probability that both a (the die shows 3) and b (the die shows an odd number) are true.

#rolled	a (die = 3)	b (die = odd)
1	0	1
2	0	0
3	1	1
4	0	0
5	0	1
6	0	0

Table 1: Possible outcomes for a die roll showing 3 and an odd number.

Each row in the table corresponds to one of the six possible outcomes of the die. The probability for each row is $\frac{1}{6}$, since each outcome is equally likely.

In this case, $P(a \wedge b)$ is only true in row 3, where both a (die = 3) and b (die is odd) are 1. Therefore, the joint probability $P(a \wedge b) = \frac{1}{6}$.

Let's also consider the conditional probability $P(a \mid b)$, which is the probability that the die shows 3 given that it is odd. We eliminate the possibility of even numbers (2, 4, 6) and are left with the odd numbers (1, 3, 5). Given that the die is odd, there are 3 equally likely outcomes, and only one of them is a 3:

$$P(a \mid b) = \frac{1}{3}$$

Notice, however, that listing all possible outcomes can quickly become overwhelming as the number of variables increases. In many cases, we may not even be aware of all the possibilities that exist in the real world, making this approach impractical.

4 Bayes' Rule

We will derive Bayes' Rule to help us avoid this generally impractical method of listing all possible outcomes. Specifically, it helps us update probabilities based on new evidence without needing to explicitly compute every joint probability.

In general, the conditional probability of a given b is defined as:

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

This formula represents the probability of a and b occurring together (joint probability $P(a \land b)$), normalized by the probability of b occurring.

We can manipulate the conditional probability formula to derive **Bayes' Rule**. Starting with:

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

We know that $P(a \land b)$ is symmetric, meaning:

$$P(a \land b) = P(b \mid a) \times P(a)$$

Substituting this back into the equation gives:

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

This is **Bayes' Rule**, which allows us to reverse conditional probabilities when the opposite condition is known. It is widely used in machine learning and probabilistic reasoning, especially when new evidence (like b) changes our belief about a.

$$\text{Bayes Rule} : \boxed{P(a \mid b) = \frac{P(b \mid a) \times P(a)}{P(b)}}$$

Where:

- $P(a \mid b)$: The posterior probability of a, given b.
- $P(b \mid a)$: The likelihood of observing b, given that a is true.
- $P(a)$: The prior probability of a (before considering b).
- $P(b)$: The evidence or normalization factor, which ensures the probabilities sum to 1.

5 Independence

Independence is a core assumption in probability (although it doesn't always hold true in the real world).

5.1 Formal Definition

In probability, two events a and b are independent if the occurrence of one does not affect the probability of the other. The formal definition of independence is:

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

This means the probability of both events occurring together (their joint probability) is simply the product of their individual probabilities.

Example: Consider rolling a die and flipping a coin. The outcome of the die roll (e.g., landing on a 3) is independent of the outcome of the coin flip (e.g., heads or tails). Therefore:

$$P(\text{die} = 3 \wedge \text{coin} = \text{heads}) = P(\text{die} = 3) \times P(\text{coin} = \text{heads})$$

Since the probability of the die showing 3 is $\frac{1}{6}$ and the probability of heads is $\frac{1}{2}$, the joint probability is:

$$P(\text{die} = 3 \wedge \text{coin} = \text{heads}) = \frac{1}{6} \times \frac{1}{2} = \frac{1}{12}$$

Thus, the events are independent because one outcome does not affect the other.

5.2 Conditional Definition

If two events are independent, then knowing that b occurred does not change the probability of a. Thus, the conditional probability of a given b is just the same as the probability of a by itself:

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

This means that adding evidence from b does not provide any new information about a.

6 Sample Problems

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

Table 2: Joint probability distribution for toothache, cavity, and catch.

We can use table 2 to find the following probabilities:

1. $P(\text{cavity})$
2. $P(\text{cavity} \mid \text{toothache})$
3. $P(\neg\text{cavity} \mid \text{toothache})$
   - Note: $P(\text{cavity} \mid \text{toothache}) + P(\neg\text{cavity} \mid \text{toothache}) = 1$.
4. $P(\text{cavity} \mid \text{toothache} \wedge \text{instrument catches})$

Solutions

1. $P(\text{cavity})$

   To calculate $P(\text{cavity})$, sum all the probabilities in the cavity row:

   $$P(\text{cavity}) = 0.108 + 0.012 + 0.072 + 0.008 = 0.200$$

   So, the probability that a patient has a cavity is $P(\text{cavity}) = 0.200$.

2. $P(\text{cavity} \mid \text{toothache})$

   The formula for conditional probability is:

   $$P(\text{cavity} \mid \text{toothache}) = \frac{P(\text{cavity and toothache})}{P(\text{toothache})}$$

   From the table:

   $$P(\text{cavity and toothache}) = 0.108 + 0.012 = 0.120$$

   The total probability of a toothache is the sum of all values in the toothache columns:

   $$P(\text{toothache}) = (0.108 + 0.012) + (0.016 + 0.064) = 0.200$$

   Now, substitute into the conditional probability formula:

   $$P(\text{cavity} \mid \text{toothache}) = \frac{0.120}{0.200} = 0.600$$

   So, $P(\text{cavity} \mid \text{toothache}) = 0.600$.

3. $P(\neg\text{cavity} \mid \text{toothache})$

   From the conditional probability formula:

   $$P(\neg\text{cavity} \mid \text{toothache}) = \frac{P(\neg\text{cavity and toothache})}{P(\text{toothache})}$$

   From the table:

   $$P(\neg\text{cavity and toothache}) = 0.016 + 0.064 = 0.080$$

   We already know $P(\text{toothache}) = 0.200$. So:

   $$P(\neg\text{cavity} \mid \text{toothache}) = \frac{0.080}{0.200} = 0.400$$

   Therefore, $P(\neg\text{cavity} \mid \text{toothache}) = 0.400$.

   Check: Since $P(\text{cavity} \mid \text{toothache}) + P(\neg\text{cavity} \mid \text{toothache}) = 1$:

   $$0.600 + 0.400 = 1$$

   This confirms our calculations.

4. $P(\text{cavity} \mid \text{toothache} \wedge \text{instrument catches})$

   The formula for this conditional probability is:

   $$P(\text{cavity} \mid \text{toothache} \wedge \text{instrument catches}) = \frac{P(\text{cavity and toothache and catch})}{P(\text{toothache and catch})}$$

   From the table:

   $$P(\text{cavity and toothache and catch}) = 0.108$$

The total probability of toothache and instrument catching is:

$$P(\text{toothache and catch}) = 0.108 + 0.016 = 0.124$$

Now substitute into the conditional probability formula:

$$P(\text{cavity} \mid \text{toothache} \wedge \text{instrument catches}) = \frac{0.108}{0.124} \approx 0.871$$

Therefore, $P(\text{cavity} \mid \text{toothache} \wedge \text{instrument catches}) \approx 0.871$.

7 Conditional Probability Tables

A **Conditional Probability Table (CPT)** is a systematic way to represent the conditional probabilities of a random variable given its *parent* variables. CPTs are fundamental building blocks in many probabilistic models.

7.1 Structure of a CPT

A CPT for a variable X with parents $Y_1, Y_2, \ldots, Y_n$ contains an entry for every possible combination of parent values, specifying $P(X \mid Y_1, Y_2, \ldots, Y_n)$.

Example: Consider a simple medical diagnosis scenario with three variables:

- **Flu** (F): whether a patient has the flu (true/false)
- **Fever** (V): whether the patient has a fever (true/false)
- **Headache** (H): whether the patient has a headache (true/false)

Suppose we model this as: Flu → Fever and Flu → Headache (flu causes both symptoms).

CPT for Flu (no parents)

Since Flu has no parents, its CPT is simply the prior probability:

Flu	$P(\text{Flu})$
true	0.1
false	0.9

Table 3: CPT for Flu

CPT for Fever (parent: Flu)

This table tells us:

- If someone has the flu, there's an 80% chance they'll have a fever
- If someone doesn't have the flu, there's only a 20% chance they'll have a fever

Flu	$P(\text{Fever} = \text{true} \mid \text{Flu})$	$P(\text{Fever} = \text{false} \mid \text{Flu})$
true	0.8	0.2
false	0.2	0.8

Table 4: CPT for Fever given Flu

CPT for Headache (parent: Flu)

Flu	$P(\text{Headache} = \text{true} \mid \text{Flu})$	$P(\text{Headache} = \text{false} \mid \text{Flu})$
true	0.7	0.3
false	0.1	0.9

Table 5: CPT for Headache given Flu

7.2 Using CPTs for Inference

CPTs allow us to compute any joint or conditional probability in the model. For example, to find $P(\text{Flu} = \text{true} \mid \text{Fever} = \text{true}, \text{Headache} = \text{true})$, we can use Bayes' rule combined with the CPT values.

Using the chain rule and our CPTs:

$$P(\text{Flu}, \text{Fever}, \text{Headache}) = P(\text{Flu}) \cdot P(\text{Fever} \mid \text{Flu}) \cdot P(\text{Headache} \mid \text{Flu})$$

For the specific case where all variables are true:

$$P(F = T, V = T, H = T) = 0.1 \times 0.8 \times 0.7 = 0.056$$

7.3 CPTs with Multiple Parents

When a variable has multiple parents, the CPT must specify probabilities for every combination of parent values.

Example: Suppose **Headache** depends on both **Flu** and **Stress**:

Flu	**Stress**	$P(\text{Headache} = \text{true} \mid \text{Flu}, \text{Stress})$	$P(\text{Headache} = \text{false} \mid \text{Flu}, \text{Stress})$
true	true	0.9	0.1
true	false	0.7	0.3
false	true	0.6	0.4
false	false	0.1	0.9

Table 6: CPT for Headache given Flu and Stress

This table shows that having both flu and stress leads to a 90% chance of headache, while having neither leads to only a 10% chance.

8 Extended Chain Rule

Recall the definition of conditional probability:

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

Equivalently,

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

More generally,

$$\begin{aligned} P(a \wedge b \wedge c) &= P(a \mid b \wedge c)P(b \wedge c) \\ &= P(a \mid b \wedge c)P(b \mid c)P(c) \end{aligned}$$

and

$$\begin{aligned} P(a \wedge b \wedge c \wedge d) &= P(a \mid b \wedge c \wedge d)P(b \wedge c \wedge d) \\ &= P(a \mid b \wedge c \wedge d)P(b \mid c \wedge d)P(c \wedge d) \\ &= P(a \mid b \wedge c \wedge d)P(b \mid c \wedge d)P(c \mid d)P(d) \end{aligned}$$

In other words, given a sequence of random variables $X_1, \ldots, X_n$,

$$\begin{aligned} P(X_1, \ldots, X_n) &= P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_2, X_1) \ldots P(X_n \mid X_{n-1}, \ldots, X_1) \\ &= P(X_1)\prod_{i=2}^{n} P(X_i \mid X_{i-1} \ldots X_1) \end{aligned}$$

9 Marginalization

A full joint distribution explains the probabilistic relationships among all the random variables in our model. But sometimes we are interested in relationships among specific, but not all, variables.

As a very simple example, consider a joint probability distribution over the weather W (sun or rain) and the temperature T (hot or cold). An example joint distribution over W and T is shown below.

W	T	$P(W,T)$
sun	hot	0.4
sun	cold	0.3
rain	hot	0.2
rain	cold	0.1

We can "marginalize over" T to compute $P(W)$; similarly, we can marginalize over W to compute $P(T)$:

W	$P(W)$
sun	0.7
rain	0.3

T	$P(T)$
hot	0.6
cold	0.4

In general, to marginalize over a random variable in a joint distribution is to tallly the joint probabilities over all its possible values: e.g., $P(W) = \sum_{t \in T} P(W, T = t)$ and $P(T) = \sum_{w \in W} P(T, W = w)$.

A very common use case for marginalization is a setting in which we are trying to infer a probability distribution over a query, given some evidence, but where some relevant information is unavailable. For example, a doctor may be trying to diagnose a disease from symptoms, and may have administered one of three possible tests. In this case, the query is the disease, the evidence includes the symptoms and the one test result, and the variables to marginalize over when inferring disease probabilities are the other two tests.