1 Motivation

Last week we discussed reasoning patterns of the form,

\[
P_1 \\
\vdots \\
P_n \\
\hline C
\]

where \(C\) necessarily followed from the premises \(P_1, \ldots, P_n\), in the sense that, as long as the premises are true, the conclusion cannot fail to be true. Simple examples of this were the valid syllogisms, including arguments like:

Some A are B
No B are C
Some A are not C

But other seemingly good arguments we discussed are not valid in this sense. Examples of this were the so called \textit{inductive} inferences like:

- Dogs have sesamoid bones
- Nematodes have sesamoid bones
- Fish have sesamoid bones
- Crickets have sesamoid bones
- Salamanders have sesamoid bones

The conclusion here does not logically follow from the premises. Yet we might think the premises somehow make the conclusion \textit{more likely} or \textit{more probable}. This already shows that, even to understand the sorts of inference patterns we were discussing last week, we need more than plain validity.

In fact, probabilistic judgments are ubiquitous, and arguably critical to understanding how intelligent beings like us get around in the world. Most
of what we believe\textsuperscript{1} appears to be “merely” probabilistic. Consider crossing El Camino Real when the walk sign is on. Doing this would only make sense if I believe that the drivers all see a red light and that no one is going to run the red light and head right for me. Could I be certain that this will not happen? Does it follow by some valid argument from what I know that I will not be struck by a car? Apparently not. Instead, all of what I know about this situation makes it seem to me very probable that it will not happen. If I learn more about the situation—say, that there is a crazy person on the loose driving down El Camino—this may change my assessment of how probable it is that I will be hit. Once you start thinking about ordinary assumptions we make, and beliefs we have about the world, you can find indications of this kind of probabilistic judgment everywhere. This might lead us to suspect that probabilistic mental states will form an important component of what goes into “common sense” intelligence.

2 Probability as a Mathematical Framework

The basic notions of probability theory can be (re)introduced using the same tools we discussed last week. Imagine we have some set \( S \) of possible situations, \( s_1, s_2, s_3, \) and \( s_4 \):

\[
\begin{array}{c}
  s_1 \\
  s_2 \\
  s_3 \\
  s_4 
\end{array}
\]

As a concrete example, imagine they stand for the following states of affairs:

\( s_1 \): “The train is running, but it is late, and it is raining.”

\( s_2 \): “The train is not running, and it is raining.”

\( s_3 \): “The train is running, but it is not late, and it is not raining.”

\( s_4 \): “The train is not running, and it is not raining.”

Since in this example\textsuperscript{2} the set \( S \) of possible states is finite, to define a probability distribution on this state space we only need to assign each of

\textsuperscript{1}Note the use of folk psychological concepts here!

\textsuperscript{2}And in most every example we will discuss in this class.
$s_1$, $s_2$, $s_3$, and $s_4$ a real number value in the range $[0, 1]$, such that they all sum to one. For instance, here might be such an assignment (call it $P$, with the numbers assigned to each state in violet):

![Diagram]

We can then say, e.g., that $P(\{s_1\}) = 0.4$ and $P(\{s_2\}) = 0.3$, or more informally, but also more suggestively, that:

\[ P(\text{"The train is running, but it is late, and it is raining"}) = 0.4 \]

\[ P(\text{"The train is not running, and it is raining"}) = 0.3 \]

To determine the probability that some statement is true, we simply add up the probabilities of all the states where the statement is true. In both of these cases there is only one. But we can also read off the probability, e.g., of a statement like, “The train is running.” There are exactly two states where this is true—$s_1$ and $s_3$—so we simply add these up and conclude:

\[ P(\text{"The train is running"}) = 0.55 \]

This is the probability of being in a state where the sentence is true, i.e., $P(\{s_1, s_3\})$, the probability that we are in one of $s_1$ or $s_3$.

Likewise, we can notice that the probability of the train not running happens to be $1 - P(\text{"The train is running"})$:

\[ P(\text{"The train is not running"}) = 0.45 \]

Using the set notation from last week, this is just $P(\{s_2, s_4\}) = P(\neg\{s_1, s_3\})$. In fact, it is generally true, that in any probability distribution $P$, we have $P(\text{Not } A) = 1 - P(A)$, i.e., for any set $X$ of states, $P(\neg X) = 1 - P(X)$.

We can also deduce what the probabilities of other logical combinations ought to be. For instance, suppose we wanted to know the probability that either the train is late or it is raining. As before, we count up the probabilities of all the states in which this statement is true:

\[ P(\text{"The train is late or it is raining"}) = 0.7 \]
This is just $P(\{s_1, s_2\})$, which notice is the union of the states where “The train is late” is true, and those where “It is raining” is true. Thinking of this in terms of the disjunction, we have

$$P(\text{“The train is late or it is raining”}) = P(\text{“The train is late”}) + P(\text{“It is raining”}) - P(\text{“The train is late and it is raining”})$$

which makes sense, given that

$$P(\{s_1\} \cup \{s_1, s_2\}) = P(\{s_1\}) + P(\{s_1, s_2\}) - P(\{s_1\})$$

Abstracting away from this specific example, we can say what a finite probability distribution is in general. Suppose we have some universe $U$, and a family of subsets of $U$: $X, Y, Z, \ldots$, which might look as follows:

A probability $P$ will map each such subset to some real number in the range $[0, 1]$, and we always require that:

1. $P(U) = 1$ — the probability of the whole space is 1.
2. $P(\neg X) = 1 - P(X)$ — the probability of not $X$ is 1 minus the probability of $X$.
3. $P(X \cup Y) = P(X) + P(Y)$, provided $X \cap Y = \emptyset$ — the probability of $X$ or $Y$, when they are disjoint (meaning that they cannot both be true), is the sum of their probabilities.

Given the correspondence discussed last week between set operations and logical connectives, we could also rewrite these using logical notation. For instance, 3 could be written:

$$3’. P(X \text{ or } Y) = P(X) + P(Y), \text{ provided } X \text{ and } Y \text{ could not both hold.}$$
3 Where Do These Probabilities Come From?

What exactly are these probabilities supposed to mean, and where do they come from? In familiar “textbook” cases, the answer seems almost obvious. If we take a fair two-sided coin, we standardly assume there are just two possible outcomes—heads or tails—and that the probability is split between the two, so that either has probability 0.5. Similarly with a six-sided dice: we assume the dice has equal probability of landing on each of the six sides, so that they all have probability 1/6. This then allows us to ask the usual questions in probability classes, like, “What is the probability of rolling an even number?” or “What is the probability in seven rolls of getting between two and four odd numbers?” and so on.

It is an interesting philosophical and foundational question why we assume these particular probabilities for the textbook cases of die, coins, etc., but we want to focus attention on even murkier—but also quite useful—questions. Suppose, for instance, we wanted to estimate the probability of any given word in English: what is the probability of hearing or reading a given word, say, the word ‘banal’? This is indeed something that researchers have need for in natural language processing, psycholinguistics, and related areas. If we had a large enough corpus of English, we might try to assign probabilities to words by counting up all the word occurrences in the entire corpus, then counting the number of occurrences of our word, and dividing the latter by the former. So, for instance, we would estimate:

\[
P('banal') = \frac{\text{number of occurrences of 'banal'}}{\text{number of word occurrences overall}}
\]

And this will typically be some very small number.

One thing you might notice about this way of assigning probabilities is that it will depend heavily on what the data you are using look like. You would expect a much higher probability assignment for the word ‘lawsuit’ if your corpus were based on legal documents than if it were based on Broadway musical reviews. You would also not expect it to be a very good estimate of the “real” probability of a word if the corpus is too small. A major problem with this simple method is when the word does not occur at all, for example. Researchers in these disciplines have ways of dealing
with these issues, and in fact the issues are quite general and shared among
disciplines. The problem of making a probability estimate based on some
set of data is one of the central problems of statistics. No matter what your
major, you will likely encounter (or have already encountered) a number of
methods for addressing this kind of problem.

There are many problems where relevant data are hard to come by, yet
we might still want to make a probability judgment. Consider the Big Bang
hypothesis. Many scientists consider it very probable that the universe
expanded from a very dense, high temperature state. But what is this judg-
ment based on? Definitely not the “count and divide” method: observing a
bunch of beginnings of the universe and seeing how many were of this sort!
Rather less direct methods go into these judgments. They also seem to be
relatively subjective, in that different scientists with the same evidence seem
to be able to make different estimates of the probabilities.

In other, more mundane cases, we may simply lack the relevant data for
other reasons, or maybe we could not hold all the data in our heads. For
instance, in my estimation of whether a car will run a red light as I cross
the street, I do find it sufficiently probable that I will not be hit. But is this
because I literally divide the number of times I have been hit by the number
of times I have crossed the street? Surely not. I may never have been hit
at all, yet still consider it possible (if unlikely) that I could be hit. And also
in this case, it seems that two people—even with all the same data about
the world—could reasonably make different estimates about how likely it
is. The probability estimates in this case, as in the last case, have a feel of
being somehow less objective than the earlier examples of die, coins, etc.

These kinds of probability estimates that are not necessarily (but may
be in part) derived from some objective external source (such as empirical
data) are called subjective probabilities. The basic idea is that many intelli-
gent agents—human agents being a paradigm case—form probabilistic judg-
ments about matters they are uncertain about, and that these judgments
help guide them in deciding what to do.\footnote{This will be the theme of next week, when we talk about decision and action.} This includes not just paradigm
cases from probability textbooks like coin tosses and roulette wheels, but
also important everyday matters like whether it is safe to cross the street or
whether the train is running on time, all the way to the most esoteric scient-
ific hypotheses and theories like the Big Band, or the Riemann hypothesis
in mathematics.\footnote{The details of what this hypothesis says are not important. The important point is
that many mathematicians would assign a very high probability to the statement being
true. But again, such judgments are not based on data in the simple straightforward


6
much as think about it, we can also assess the probability of it. Because many of these judgments are likely to be different for different agents the probabilities are called ‘subjective’.

The assumption that a reasonable agent architecture will include subjective probabilistic judgments is common across many SymSys disciplines. In AI many agents are endowed with probability judgments so that they can make smart decisions.\(^6\) In cognitive psychology and in linguistics (especially at Stanford), it is often assumed that much of what we believe about the world can be understood in terms of subjective probabilities, and that this helps explain how we interact with one another, interpret one another’s speech and actions, and so on. Many contemporary epistemologists and philosophers of language, psychology, science, and mathematics also assume that we all harbor subjective probabilities.

4 And Whence the Axioms?

When probabilities are derived in a straightforward, “objective” way—like when we just assume we know the possible atomic states (heads or tails) and their probabilities (0.5, 0.5), or when we use the “count and divide” method—the probability assignment automatically conform to the rules 1, 2, 3 from §2 above. For instance, in the count-and-divide method, if we assess the probability of ‘banal’ to be

\[
P(\text{‘banal’}) = \frac{\# \text{ occurrences of ‘banal}}{\# \text{ overall word occurrences}}
\]

then we cannot help but have that

\[
P(\text{‘banal’ or ‘slime’}) = \frac{\# \text{ occurrences of ‘banal}}{\# \text{ overall word occurrences}} + \frac{\# \text{ occurrences of ‘slime}}{\# \text{ overall word occurrences}}
\]

\[
= \frac{\# \text{ occurrences of ‘banal’ or ‘slime’}}{\# \text{ overall word occurrences}}
\]

way that probabilities based on the “count and divide” method—or even any of its more sophisticated cousins, like “Laplace’s method”, etc.—are.

\(^6\)We already mentioned one instance of this in NLP. For another instance, in robotics an agent might maintain a probability distribution over its location on a map of some environment from which it receives only noisy, intermittent data.
or for example that

\[
P(\text{a word other than \textquote{banal}}) = \frac{\# \text{ occurrences of \textquote{slime}}}{\# \text{ overall word occurrences}} + \frac{\# \text{ occurrences of \textquote{bachelor}}}{\# \text{ overall word occurrences}} + \ldots
\]

\[
= \frac{\# \text{ occurrences of words other than \textquote{banal}}}{\# \text{ overall word occurrences}}
\]

\[
= 1 - P(\textquote{banal})
\]

Rule 1, 2, and 3 simply hold by the way we are defining the probabilities.

However, this is not so automatic when considering subjective probabilities in general. Suppose I make the following judgments about how likely each of the following teams is to win a given national championship:

\[
P(\text{Stanford}) = 0.2 \quad \quad P(\text{Oregon}) = 0.1
\]
\[
P(\text{Berkeley}) = 0.01 \quad \quad P(\text{Oregon State}) = 0.02
\]
\[
P(\text{UCLA}) = 0.015 \quad \quad P(\text{USC}) = 0.008
\]
\[
P(\text{Arizona}) = 0.01 \quad \quad P(\text{Utah}) = 0.009
\]
\[
P(\text{Arizona State}) = 0.02 \quad \quad P(\text{Washington}) = 0.02
\]
\[
P(\text{Colorado}) = 0.009 \quad \quad P(\text{Washington State}) = 0.01
\]

Again, it goes without saying that any two people will differ at least somewhat in these subjective estimates. But the question we are now asking is whether these judgments must satisfy the probability axioms. What would that mean in this context? Here it would mean, e.g., that the probability of Stanford or Berkeley winning must be 0.21. It also means, for example, that the probability of some PAC-12 team winning must be 0.431, the sum of all these numbers. To stress the point, if this is to be a proper probability assignment, the agent who makes the 12 judgments above cannot assign any number except 0.431 to the statement that a PAC-12 team will win. Since these judgments are now allowed to be purely subjective, what would compel us to require these axioms hold? Put otherwise, what would be irrational about a person making the 12 judgments above and nonetheless judging that there is, say, a 0.44 chance that a PAC-12 team will win? Someone with slightly different judgments about the 12 individual teams may indeed have judged that \(P(\text{PAC-12}) = 0.44\). Why (if at all?) would there be something

\footnote{It is important for this example that the PAC-12 is comprised of these 12 teams.}
“wrong” if one and the same person made these 12 judgments but also the judgment that $P(\text{PAC-12}) = 0.44$?

Or to take a scientific case, suppose that a climate scientist judges that there is a 0.3 probability that the global average temperature will increase by at least 1 degree Celsius within 10 years. Why might it be unreasonable or irrational for her to simultaneously judge that there is a 0.8 probability (or anything other than $1 - 0.3 = 0.7$ probability) that this will not happen? Surely some climate scientists could reasonably judge that there is a 0.2 probability of 1 degree increase, and therefore a 0.8 probability of this not happening. So there is nothing inherently unreasonable about that judgment. But the probability axioms—in the case of subjective probability—say that someone who judges the increase to have 0.3 probability must judge the negation of this to have 0.7 probability. Why?

Answers to these questions have been proposed by mathematicians and philosophers over the years. If we have time, we may discuss some of these arguments in class. Whatever the justification, most cognitive scientists and AI researchers using probabilities to model and design agents do assume that these numbers obey the probability axioms, even when they are subjective probabilities. It is an open, foundational question why.

5 “Updating” a Probability Distribution

Provided an agent’s probability assignment does satisfy the axioms, there is a widely assumed canonical way of “updating” the agent’s distribution with new information. Suppose our agent assigns probability $P$ to a bunch of events, $X, Y, Z,$ and so on. And then suppose she learns that $Y$ is definitely true. What should she then say about the probabilities of $X, Z,$ and all the rest? Many researchers think that probabilistic conditioning gives the most reasonable answer to this question. Define the following notation:

$$P(X | Y) = \frac{P(X \text{ and } Y)}{P(Y)} = \frac{P(X \cap Y)}{P(Y)}$$

We read $P(X | Y)$ as giving the probability of $X,$ conditional on $Y,$ or having learned that $Y$ is true. In a picture (on the next page), we are interested in the ratio of the purple (blue+red) region’s probability to the red blue region’s probability. Having learned that $Y$ is true means that we can restrict attention to the blue region: we know the “real world” is one in $Y$. We then look at the probability of still being in an $X$-world, given that we are restricting attention to the $Y$ worlds.
For a concrete example, let us return to the example from §2 about trains and rain. Suppose I learn that the train is running, i.e., that one of the worlds $s_1$ or $s_3$ is the real world (either the train is late and it is raining, or the train is not late and it is not raining). Then we can altogether disregard $s_2$ and $s_4$, and consider just a subset of the points:

The problem with the picture now is that this is not a proper probability distribution, since the numbers do not sum to 1. A natural way of fixing this problem is simply to add all the numbers up, and divide the old numbers by this sum, so that they will definitely sum to 1. (The technical term for this move is renormalization.)

Then we can see from the picture, e.g., that $P$(Rain | The train is running) = $8/11$. This could also be derived directly with the formula above:
\[ P(\text{Rain} \mid \text{The train is running}) = \frac{P(\text{It is raining and the train is running})}{P(\text{The train is running})} \]
\[ = \frac{0.4}{0.55} \]
\[ = \frac{8}{11} \]

Note that \( P(X \mid Y) \) is just a defined notion. It is a further substantive claim that the right way to update probabilities having learned that \( Y \) is true is to use this definition. The vast majority of researchers who use probability do assume this, but there are also people who think it is not the right way to think of probabilistic learning.

On the other hand, if we are convinced that this is the right picture of learning—that \( P(X \mid Y) \) in general captures what an agent’s probability for \( X \) should be after having learned \( Y \)—then we can avail ourselves of a very powerful tool, namely Bayes Theorem. This is a very elementary result about probability (hardly a theorem!), which says:

\[ P(X \mid Y) = \frac{P(Y \mid X)P(X)}{P(Y)} \quad (1) \]

Why does this hold? It is actually quite simple. Because \( P(X \mid Y) = P(X \cap Y)/P(Y) \), and thus \( P(X \cap Y) = P(X \mid Y) P(Y) \), and because also \( P(Y \mid X) = P(Y \cap X)/P(X) \), and hence \( P(X \cap Y) = P(Y \mid X) P(X) \), we therefore have \( P(X \mid Y) P(Y) = P(Y \mid X) P(X) \), which easily gives (1).

And what does it mean? In some cases we may know—or have opinions about—only some information. Bayes Theorem allows us to infer new information from what we already know or believe.

Here is a simple example from social reasoning, to foreshadow some of what is to come in the course. Suppose my friend says to me, “You seem awfully chipper this morning!” (call this sentence \( S \)), and I want to assess how likely it is that my friend is trying to insult me. That is, I am trying to assess:

\[ P(\text{My friend is insulting me} \mid \text{My friend said } S) \]

Then Bayes Theorem tells me I can determine this just from knowing the following information:

\[ P(\text{My friend would say } S) \]
\[ P(\text{My friend would try to insult me}) \]
\[ P(\text{My friend would say } S \text{ if this friend were trying to insult me}) \]
Because by Bayes Theorem, we have:

\[
P(\text{Trying to insult me} \mid \text{Said } S) = \frac{P(\text{Said } S \mid \text{Trying to insult me}) P(\text{Trying to insult me})}{P(\text{Said } S)}
\]

We will see many more examples of this kind of reasoning.

Another paradigm example of Bayesian reasoning is in scientific theory choice. Suppose, for example, that theory \( T_1 \) makes the observed data \( D \) very probable, i.e., \( P(D \mid T_1) \) is high, while on the face of it we find \( T_1 \) rather implausible, i.e., \( P(T_1) \) is low. Suppose the only competitor theory is \( T_2 \), which does not make the observed data quite as probable, i.e., \( P(D \mid T_2) \) is rather low, even though it is initially a very plausible theory, i.e., \( P(T_2) \) is high. Given all the data, how should we decide which theory is more probably the correct one? Bayes Theorem actually tells us:

\[
P(T_1 \mid D) = \frac{P(D \mid T_1) P(T_1)}{P(D)} \quad P(T_2 \mid D) = \frac{P(D \mid T_2) P(T_2)}{P(D)}
\]

And in fact, if we are only interesting in comparing \( T_1 \) and \( T_2 \), then we can ignore the term \( P(D) \), and just compare the two numbers:

\[
P(D \mid T_1) P(T_1) > ? <? P(D \mid T_2) P(T_2)
\]

If the left is larger, this means we should prefer \( T_1 \), whereas if the right is larger we should prefer \( T_2 \). The intuition here is that how good a theory is should depend both on how initially plausible it is, and also how well it accounts for the data (how probable it makes the data). This method is known as Bayesian theory comparison.

Starting already on Thursday, Bayesian reasoning will play an important role in understanding some concrete cognitive phenomena, in areas as seemingly diverse as language understanding, social reasoning, and some aspects of perception. The aim of these notes is to ensure that you understand what is behind the assumption that people maintain probabilistic beliefs to begin with, and what this means. In particular, for Bayesian reasoning to work, people must have mental states that can be understood as representing probability distributions, satisfying the axioms above. It is a further empirical hypothesis—which we will also look at in more detail—that the human mind somehow updates these probabilistic beliefs by something like conditioning, often invoking Bayes Theorem.