

# Notes on probability

Tim Bretl

Department of Aerospace Engineering  
Beckman Institute for Advanced Science and Technology  
University of Illinois at Urbana-Champaign

AE498MPA

September 18, 2007

# Elements of a probabilistic model

A probabilistic model is a formal way to describe an uncertain process:

- The uncertain process is the **experiment**.
- The experiment produces exactly one **outcome**  $s$ .
- A set of outcomes is an **event**  $A$ . If the experiment produces any outcome  $s \in A$ , then “the event  $A$  occurs.”
- The set of all possible outcomes is the **sample space**  $\Omega$ . Each outcome is an element  $s \in \Omega$ . Each event is a subset  $A \subset \Omega$ .
- The **probability** that an event  $A \subset \Omega$  occurs is a number  $P(A)$ .

# Set operations and probability

We define set operations (for events) like this:

- $A \cup B = \{s \in \Omega \mid s \in A \text{ or } s \in B\}$  is the **union** of two sets.
- $A \cap B = \{s \in \Omega \mid s \in A \text{ and } s \in B\}$  is the **intersection** of two sets.
- $A^c = \{s \in \Omega \mid s \notin A\}$  is the **complement** of a set.

We define the corresponding probabilities like this:

- $P(A \cup B)$  means the probability that **A or B** occurs.
- $P(A \cap B)$  means the probability that **A and B** occurs.
- $P(A^c)$  means the probability that **A does not** occur.

# Axioms of probability

For the probabilistic model to make sense, the probabilities  $P(A)$  must satisfy the following three axioms:

- $P(A) \geq 0$  for all  $A \subset \Omega$
- $P(\Omega) = 1$
- If  $A \cap B = \emptyset$ , then  $P(A \cup B) = P(A) + P(B)$

# Explaining the axioms in terms of relative frequency

To get some intuition, think of  $P(A)$  as the **relative frequency** of the event  $A$ . So if we conduct the same experiment  $n$  times, we expect  $A$  to occur  $P(A) \cdot n$  times on average.

- (First axiom.) No event can occur a negative number of times.
- (Second axiom.) Some outcome occurs every time, and the sample space  $\Omega$  is an event that contains all outcomes.
- (Third axiom.) If event  $A$  happens  $P(A) \cdot n$  times and event  $B$  happens  $P(B) \cdot n$  times, and if events  $A$  and  $B$  never happen at the same time, then the fraction of times that  $A$  or  $B$  occurs is

$$\frac{P(A) \cdot n + P(B) \cdot n}{n} = P(A) + P(B).$$

# Why is probability defined for events and not outcomes?

- If the sample space is finite, it doesn't make any difference. We can express any event as a finite set of outcomes  $A = \{s_1, \dots, s_n\}$ . So the third axiom says that the probability of event  $A$  is

$$P(A) = \sum_{i=1}^n P(\{s_i\}),$$

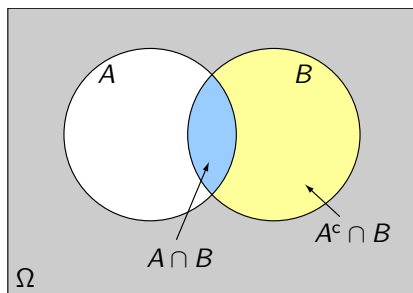
the sum of the probabilities of each outcome.

- If the sample space is infinite, we have to be careful. Consider an experiment with the sample space  $\Omega = [0, 1]$  where all outcomes are equally likely. Then the probability of each outcome *must* be zero. Otherwise the third axiom says that the probability of  $\Omega$  is infinite.
- We fix this problem by defining probability for events, not outcomes. As an example, we might define it by  $P(A) = \int_A ds$  for any event  $A \subset [0, 1]$ . So if  $A = [a, b]$ , then  $P(A) = b - a$ . This definition satisfies all three axioms.

# What if two events are not disjoint? (1)

## Lemma

For any  $A, B \subset \Omega$  we have  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ .



## What if two events are not disjoint? (2)

Proof.

- Express  $B$  and  $A \cup B$  as unions of disjoint events:

$$B = (A \cap B) \cup (A^c \cap B)$$
$$A \cup B = A \cup (A^c \cap B)$$

- Apply the third axiom (additivity):

$$P(B) = P(A \cap B) + P(A^c \cap B)$$
$$P(A \cup B) = P(A) + P(A^c \cap B)$$

- Solve for  $P(A \cup B)$ :

$$P(A \cup B) - P(B) = P(A) - P(A \cap B)$$
$$\Rightarrow P(A \cup B) = P(A) + P(B) - P(A \cap B)$$



## Other immediate consequences

The following are true:

- For any  $A, B \subset \Omega$  we have  $P(A \cup B) \leq P(A) + P(B)$
- $P(\emptyset) = 0$

# Conditional probability

Suppose we *know* that an event  $B$  occurs (so by necessity  $P(B) \neq 0$ ). This knowledge affects the probability that an event  $A$  also occurs:

The **conditional probability** of  $A$  given  $B$  is

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

So knowledge of  $B$  reduces the set of possible outcomes from  $A$  to  $A \cap B$  and scales (or *normalizes*) the probability of what's left by  $1/P(B)$ .

## Conditioning induces a new probabilistic model

Assume we are given a valid probability law  $P(A)$  defined on all  $A \subset \Omega$ . Then conditioning on a known event  $B \subset \Omega$  generates another valid probability law, on the “new” sample space  $B$ :

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \geq 0 \quad \text{non-negativity}$$

$$P(B|B) = 1 \quad \text{normalization}$$

$$\begin{aligned} P(A_1 \cup A_2|B) &= \frac{P((A_1 \cup A_2) \cap B)}{P(B)} \\ &= \frac{P((A_1 \cap B) \cup (A_2 \cap B))}{P(B)} \\ &= \frac{P(A_1 \cap B) + P(A_2 \cap B)}{P(B)} \\ &= P(A_1|B) + P(A_2|B) \quad \text{additivity} \end{aligned}$$

(This fact will allow us to add conditioners to things like Bayes' Rule.)

# Bayes' Rule

There is another way to compute  $P(A|B)$  that does not require direct knowledge either of  $P(A \cap B)$  or of  $P(B)$ . Notice that both of the following are true, from the definition of conditional probability:

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

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

Eliminating  $P(A \cap B)$  from these expressions, we arrive at...

Bayes' Rule

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

## Some comments on Bayes' Rule

- It relates  $P(A|B)$ , which we don't know, to its "inverse"  $P(B|A)$ , which we often do.
- You can think about  $A$  as a cause and  $B$  as an effect. Given an observed effect, Bayes' Rule allows us to infer the cause. (Many different events  $C_1, \dots, C_n$  may have caused an observed effect  $B$ . Bayes' Rule allows us to compute the probability that event  $A$  is the one that caused  $B$ .)
- We generally call  $P(A)$  the **prior** and  $P(A|B)$  the **posterior** (or the a-posteriori probability).

# The distinction between “facts” and “measurements”

It is important to understand that when we write  $P(A|B)$ , we are treating the event  $B$  as a **fact**. It has happened, period.

- Imagine our sample space includes every city in America (and one outcome that means “not a city.”) Say  $A$  is the event that you are in Champaign-Urbana and  $B$  is the event that you are in Illinois.
- When we write  $P(A|B)$ , we mean the probability that we are in Champaign-Urbana given that we are in Illinois.
- The fact that we are in Illinois may have been derived from an observation, for example by seeing a sign that says “Welcome to Illinois.” But  $B$  itself is a known event, not an observation. We are accepting it as given.
- So it does not make sense to ask about how certain we are that  $B$  is given (in other words, about “how good our sensor is”). It is a fact.

Later, when we talk about measurements, we will revisit these questions.

## But what is $P(B)$ ?

- We can write  $\Omega$  as a disjoint union:

$$\Omega = A \cup A^c$$

- As a consequence, we can also write  $B$  as a disjoint union:

$$\begin{aligned} B &= B \cap \Omega \\ &= B \cap (A \cup A^c) \\ &= (B \cap A) \cup (B \cap A^c) \end{aligned}$$

- So the third axiom of probability says the following:

$$P(B) = P(B \cap A) + P(B \cap A^c)$$

- And finally, from the definition of conditional probability...

$$P(B) = P(B|A)P(A) + P(B|A^c)P(A^c)$$

## Another form of Bayes' Rule

### Bayes' Rule

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}$$

- All we've done is plug in our expression for  $P(B)$ .
- But notice that now, *everything* on the right-hand-side is expressed in terms of **conditional probabilities** (how likely is the effect given the cause) and **priors** (how likely is the cause, all things being equal).
- Also,  $P(B|A)P(A)$  appears in *both* the numerator and denominator.
- And most importantly, the denominator would be *exactly the same* for  $P(A^c|B)$ .

# Partitions

- $\Omega = A \cup A^c$  is an example of a **partition** of the sample space.
- A partition of a set is a disjoint union that covers the entire set.
- There are generally many other choices for a partition.
- For example, imagine  $\Omega = \{1, 2, 3, 4\}$ . The sets  $\{1, 2\}$ ,  $\{2, 3\}$ , and  $\{4\}$  do not form a partition because they are not disjoint. The sets  $\{1\}$ ,  $\{2\}$ , and  $\{3\}$  do not form a partition because their union does not equal  $\Omega$ . The sets  $\{1, 2\}$ ,  $\{3\}$ , and  $\{4\}$  *do* form a valid partition.

# The law of total probability

## Lemma (Law of Total Probability)

Let  $A_1, \dots, A_n$  be a partition of  $\Omega$  and let  $B \subset \Omega$ . Then

$$P(B) = \sum_{i=1}^n P(B|A_i)P(A_i).$$

## Proof.

Just as before, we can write  $B$  as the disjoint union

$$B = (A_1 \cap B) \cup \dots \cup (A_n \cap B).$$

So from the third axiom of probability,

$$P(B) = P(A_1 \cap B) + \dots + P(A_n \cap B).$$

And the result follows from the definition of conditional probability, namely  $P(A_i \cap B) = P(B|A_i)P(A_i)$ . □

## Yet another form of Bayes' Rule

### Bayes' Rule

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^n P(B|A_j)P(A_j)}$$

- All we've done is generalize the rule for an arbitrary partition  $A_1, \dots, A_n$  of the sample space  $\Omega$ .
- As before, notice that everything is expressed in terms of conditional probabilities and priors, and that  $P(B|A_i)P(A_i)$  appears in the numerator and denominator for every  $i$ .
- And again, the denominator is *exactly the same* for each  $A_i$ .

## Bayes' Rule in practice

- We will often be interested in computing the entire posterior, or in other words in computing  $P(A_i|B)$  for every  $A_i$ .
- In practice, we do not have to compute the denominator

$$\sum_{j=1}^n P(B|A_j)P(A_j)$$

every single time, since it is the same for each  $A_i$ .

- Instead, we simply compute  $P(B|A_i)P(A_i)$  for every  $i$ , and divide by their sum to get  $P(A_i|B)$ .
- We can think of the sum as a normalizing parameter, call it  $\eta$ . So we often write Bayes' Rule as follows:

Bayes' Rule

$$P(A_i|B) = \eta P(B|A_i)P(A_i)$$

## Probability given multiple facts (batch method)

- Rather than just one event  $B$ , let's say we know (for a fact) that two events  $B_1$  and  $B_2$  have occurred.
- We write the conditional probability of  $A_i$  given  $B_1$  and  $B_2$  as  $P(A_i|B_1, B_2)$ , where this is shorthand for  $P(A_i|B_1 \cap B_2)$ .
- We could apply the definition of conditional probability:

$$P(A_i|B_1, B_2) = \frac{P(A_i \cap (B_1 \cap B_2))}{P(B_1 \cap B_2)}$$

- We could also apply Bayes' Rule:

$$P(A_i|B_1, B_2) = \frac{P(B_1 \cap B_2|A_i)P(A_i)}{\sum_{j=1}^n P(B_1 \cap B_2|A_j)P(A_j)}$$

- Both of these are **batch** methods. We keep a list of facts (a list of given events  $B_1, \dots, B_m$ ). Every time we add a new fact (a new event  $B_{m+1}$ ), we recompute the conditional probability of each  $A_i$  using the entire list of facts.

## Probability given multiple facts (recursive method)

- But there is another way to find  $P(A_i|B_1, B_2)$  with Bayes' Rule. Imagine we are given  $B_1$  first. Then we know

$$P(A_i|B_1) = \frac{P(B_1|A_i)P(A_i)}{\sum_{j=1}^n P(B_1|A_j)P(A_j)}.$$

- Now imagine we are given  $B_2$ . Again we apply Bayes' Rule, but this time conditioned on  $B_1$ :

$$P(A_i|B_2, B_1) = \frac{P(B_2|A_i, B_1)P(A_i|B_1)}{\sum_{j=1}^n P(B_2|A_j, B_1)P(A_j|B_1)}.$$

- This is a **recursive** method. We begin with a prior  $P(A_i)$ . Given  $B_1$ , we compute the posterior  $P(A_i|B_1)$ . This posterior becomes our new prior, and given  $B_2$  we compute  $P(A_i|B_2, B_1)$ .
- The only remaining problem is the term  $P(B_2|A_i, B_1)$ . This is not recursive yet, since it depends on how  $B_1$  and  $B_2$  relate. We will see that for many practical problems  $P(B_2|A_i, B_1) = P(B_2|A_i)$ .

## For more information...

There are many good resources, both offline...

- Introduction to Probability (Bertsekas and Tsitsiklis, Athena Scientific, 2002)
- Probability and Random Processes, Third Edition (Grimmett and Stirzaker, Oxford, 2001)

and online...

- An Exploration of Random Processes for Engineers (Bruce Hajek's course notes from ECE534 here at Illinois)
- Introduction to Probability (Grinstead and Snell, AMS, 1997)
- Sanjay Lall's E207B Course Notes (actually a course on modern control, but taught from a probabilistic viewpoint)