Philosophy and AI
Lecture 3: Epistemology - Probability

Marco Degano

11 February 2025

Readings

Required:

- Glymour 2015: chapter 8, pp. 185 – 209:
  - Sections: The Theory of Probability, Bernoulli Trials; The Binominal Distribution; Bayes, Price, and Hume; The Modern Revival; Bounded Rationality and Bayesian Problems.

Optional:

- Lin, Hanti, "Bayesian Epistemology", The Stanford Encyclopedia of Philosophy, `https://plato.stanford.edu/archives/sum2024/entries/epistemology-bayesian/`

Outline

1. Probability Theory

2. Induction and Probability

3. Bayesian Solutions

Outline

1. Probability Theory
2. Induction and Probability
3. Bayesian Solutions

Intro

- Skepticism claims that there is no reliable procedure to obtain knowledge about the world.
- In this and the next lecture, we look at solutions: spelling out the idea that we don't need reliability in all worlds, but only in those we care about.
- In this lecture, we see how probability theory and Bayesianism can provide such a solution to (Hume's) inductive skepticism.
- Reliability is expressed by a **probability distribution**, so this method will consist of deriving a reliable probability distribution based on observations.

Probability Theory

- Probability theory is the mathematical theory of assigning probabilities to events and reasoning with them.
- A probability measure assigns probabilities (i.e., real numbers $0 \leq p \leq 1$) to events.

Probability Theory: Example

- We have an urn with marbles of the same weight and size.
- There are b-many black marbles, and w-many white marbles (and there are no other marbles).
- An 'experiment' is: shuffle, take a marble (blindly), record it, replace it, shuffle again.
- The possible outcomes (or samples) of an experiment are drawing a white marble (W) and drawing a black marble (B).
- Since there are no other outcomes, the outcome space is $X = \{W, B\}$.

Probability Theory: Events

- An event is the occurrence of an outcome: we write $\{B\}$ for the event that the outcome B occurs.
- The probability of this event in our experiment is $p(\{B\}) = \frac{b}{b+w}$. And similarly: $p(\{W\}) = \frac{w}{b+w}$.
- Logical combinations of events are also events:
  - The event of drawing either a black or a white marble is their union $\{B\} \cup \{W\} = \{B, W\}$ $(= X)$.
  - For 'and' it is the intersection: $\{B\} \cap \{W\}$ $(= \emptyset)$.
  - The event of not drawing a black marble is the complement $X \setminus \{B\} = \{W\}$.
- The set of events is thus the set of all subsets of the outcome space X: written as $\mathscr{P}(X)$ (the powerset of X).

Probability Theory: Probability Space

- The trivial event that an outcome occurs is X and has a probability of 1 (is certain).
- If two events A and B are incompatible ($A \cap B = \emptyset$), then $p(A \cup B) = p(A) + p(B)$.
- More generally:

Definition

A probability measure over a finite set X is a function $p : \mathscr{P}(X) \to [0,1]$ such that

- $p(X) = 1$, and
- $p(A \cup B) = p(A) + p(B)$ if $A \cap B = \emptyset$.

We call $(X, \mathscr{P}(X), p)$ a probability space.

Probability Theory

- For the mathematicians:
- The general definition of a probability space $(X, \mathscr{F}, p)$ with a potentially infinite outcome space X (e.g., $X = \mathbb{R}$) is more complex:
  - One may allow only certain subsets of $\mathscr{P}(X)$ (called σ-algebras) as the set of events $\mathscr{F}$, and
  - p must be countably additive (instead of merely finitely additive).
- If we need this more general case, we will address it only 'intuitively' i.e., not formally.
- Terminology: In the finite case, one may also speak of a probability distribution rather than a probability measure. If $x \in X$, we write $p(x)$ for $p(\{x\})$.

Outline

The Problem of Induction, Probabilistically I

The connection to induction is the following analogy:

empirical law	—	probability measure (the urn of nature)
instance/observation	—	event (drawing from the urn of nature)

- For example: An ‘experiment’ could be observing, on a given day, whether the sun rises (outcome 1) or not (outcome 0).
- The observation/instance that the sun rises corresponds to the event $\{1\}$; that it does not rise corresponds to $\{0\}$.
- The empirical law that the sun always rises then corresponds to the probability measure p such that the experiment is designed so $p(\{1\}) = 1$ (the sun rising is certain).

The Problem of Induction, Probabilistically II

- Determining whether the empirical law holds (through empirical research) corresponds to determining the probability measure of the urn of nature (i.e., the true probability).
- In other words, according to this analogy, empirical research is akin to determining the true probability measure of an urn (the ratio of white to black marbles) by conducting repeated experiments.
- Repeated experiments with an urn are known as Bernoulli trials.

The Problem of Induction, Probabilistically III

Thus, we can extend the analogy to what we can call the Bernoulli model of empirical research:

empirical law	—	probability measure (the urn of nature)
instance/observation	—	event (drawing from the urn of nature)
empirical research	—	repeated drawing (Bernoulli trials)

Two questions arise:

- How do we formally describe Bernoulli trials?
- Can this model be used to approximate the true probability distribution/measure?

Bernoulli Trials I

- Assume that outcome 1 (black marble, sunrise, etc.) has probability p; and outcome 0 (white marble, no sunrise, etc.) has probability $q = 1 - p$.
- We call outcome 1 a success and outcome 0 a failure.
- If we conduct the experiment twice, there are four possible outcomes: $11, 10, 01, 00$. The probabilities are:

$$p(\{11\}) = pp \quad p(\{10\}) = pq \quad p(\{01\}) = qp \quad p(\{00\}) = qq.$$

- For example, the probability of the event that there is (exactly) one success is $p(\{10, 01\})$, and that is

$$p(\{10\}) + p(\{01\}) = pq + qp = 2pq.$$

Bernoulli Trials II

- In general: If we conduct n experiments, the outcomes are binary strings of length n,
- and the probability of k successes in the entire trial is

$$\binom{n}{k} p^k q^{n-k}$$

where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ is the binomial coefficient ($n! = 1 \cdot \ldots \cdot n$ with $0! = 1$).

- $\binom{n}{k}$ describes how many ways there are to choose k objects out of n different objects.
- In the previous example (1 success in 2 trials: $n = 2$, $k = 1$), $\binom{2}{1} = 2$.

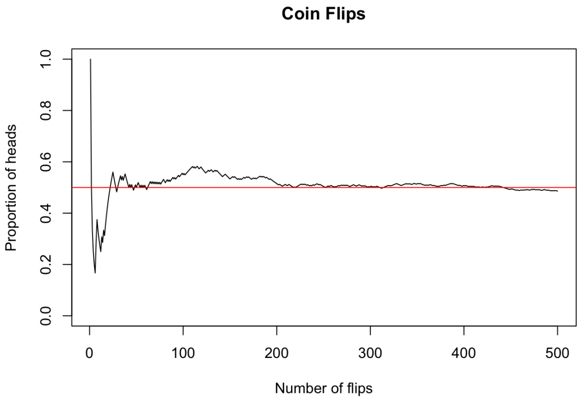
Bernoulli Trials III

- We described precisely what Bernoulli trials are.
- Can they approximate the true probability p?
- Yes! That is what the Bernoulli theorem is about, also known as

Theorem (The Weak Law of Large Numbers)

For any p (the true probability of the urn), for any real number $\varepsilon > 0$ (a small margin of error), the probability that the observed proportion of successes in n trials differs from the true proportion of successes p by more than ε converges to 0 as the number of trials n increases.

Illustration



- **Initial Variability:** At the beginning (few flips), the proportion of heads fluctuates greatly because the sample size is small, and random variation dominates.
- **Convergence Over Trials:** As the number of flips increases, the proportion of heads stabilizes and gets closer to the theoretical value (0.5).

Bernoulli Trials IV

- Bernoulli's response to inductive skepticism argues that while empirical research cannot provide absolute certainty, it can offer increasing confidence in probabilistic conclusions.
- Since empirical research involves a finite number of observations, we cannot determine the exact true probability of an event. However, repeated trials allow us to approximate it with increasing accuracy.
- As the number of trials grows, the observed proportion of successes becomes closer to the true probability, making our empirical estimates more reliable.
- Instead of achieving absolute knowledge, we attain progressively justified belief, which is practically sufficient for scientific inquiry and decision-making.

Problems with the Bernoulli Model

- The Bernoulli theorem establishes that, as the number of trials increases, the observed probability is the true probability within an arbitrarily small margin of error.
- However, this result is primarily qualitative: it guarantees that certainty approaches 1 but does not specify how this happens for a given number of trials.
- Bernoulli himself recognized this limitation-he lacked a precise method to quantify the accuracy of an observed proportion relative to the number of trials.
- This limitation is significant for addressing inductive skepticism. We not only need assurance that our confidence increases over time but also a way to measure how much certainty we have at any given stage.

Outline

1. Probability Theory
2. Induction and Probability
3. Bayesian Solutions

The Bayesian Turn

- The problem of induction, in probabilistic terms, is as follows:
  - Given observations, we estimate the true probability distribution p.
  - We seek to quantify how confident we can be that our estimate is correct.
- Bayes' key insight was to approach this using probabilities of probabilities, assigning a probability distribution to p itself.

The Bayesian Turn

- To formalize this, we distinguish between:
  - At the base level, p describes the probability in an experiment (e.g., drawing from an urn).
  - At the higher level, a probability distribution represents our uncertainty about the correct value of p.
- Without prior observations, we assume all values of p are equally likely (a uniform prior).
- As we collect more data, we refine our estimate of p, and we get a better idea of the correct value of p.

The Bayesian Turn

- (Don't worry, the mathematical details aren't very important. In the exams, you do not need to calculate any integral. But focus on what they intuitively mean and imply)
- Mathematically, we seek the probability that p lies within an interval $[a, b]$ given k successes in n trials:

$$P\big(a \leq p \leq b \mid k \text{ successes in } n \text{ trials}\big).$$

- Conditional probability $P(A|B)$ is defined as $P(A \cap B)/P(B)$, so we need to determine $P(A \cap B)$ and $P(B)$.

The Bayesian Turn

- What is $P(B) = P(k \text{ successes in } n \text{ trials})$?
- Idea: Calculate the probability of k successes in n trials given the probability of success p, and then average over all possible values of p.
- Some mathematical details:
  - Since p takes real values ($p \in [0,1]$), we cannot compute the average using a sum but need an integral.
  - Since we haven't observed anything yet, all values of p are equally likely. Thus:

$$P(B) = \int_0^1 \binom{n}{k} p^k (1-p)^{n-k} \; dp.$$

The Bayesian Turn

- What is $P(a \leq p \leq b \textit{ and } k$ successes in n trials), i.e., $P(A \cap B)$?
- The same idea, except now p is in $[a, b]$ instead of $[0, 1]$. Thus:

$$P(A \cap B) = \int_a^b \binom{n}{k} p^k (1-p)^{n-k} \, dp.$$

- We now have an expression for the desired $P(A|B)$. However, solving these integrals is difficult. Bayes solved the first: $P(B) = 1/(n+1)$.
- As an example of the second, consider our observation of the sunrise: i.e., where every observation is a success.

The Bayesian Turn

- Then $n = k$, so $\binom{n}{n} = 1$ and $n - k = 0$, so:

$$P(A \cap B) = \int_a^b \binom{n}{k} p^k (1-p)^{n-k}\ dp = \int_a^b p^n dp$$
$$= \left[\frac{1}{n+1} p^{n+1}\right]_a^b = \frac{1}{n+1}\left(b^{n+1} - a^{n+1}\right)$$

- Since $P(B) = 1/(n+1)$, and $P(A|B)$ is $P(A \cap B)/P(B)$, we have:

$$P\big(a \leq p \leq b \mid\ n \text{ successes in } n \text{ trials}\big) = b^{n+1} - a^{n+1}.$$

The Bayesian Turn

- Thus, after observing $n = 10$ sunrises, our confidence that the true probability (nature's urn) lies between 0.8 and 1 is $1^{11} - 0.8^{11} \approx 0.91$.
- After observing $n = 500$ sunrises, our confidence that the true probability lies between 0.99 and 1 is ≈ 0.993.
- This is the Bayesian answer to the problem of induction:
  - Not only do the observed probabilities converge to the true probability (as Bernoulli already showed).
  - We can also quantify how confident we are in our estimate given the number of observations.

Modern Bayesianism

(1) **A New Interpretation of Probability**

- The traditional interpretation, frequentism, defines probability as relative frequency-the limit of an event's occurrences over many trials.
- Bayesianism, in contrast, defines probability as rational degrees of belief-the confidence a rational agent assigns to an event.
- Degrees of belief take values in $[0,1]$ and must follow probability axioms.
- Rationality requires updating beliefs when encountering new evidence via conditional probability.
- Frequentism is objective (probability depends on observed frequencies).
- Bayesianism is subjective (probability applies even to unobserved events).

Modern Bayesianism

(2) **Beyond Bernoulli Trials: A More General Bayesian Framework**

- Bayesianism is not limited to simple yes/no (Bernoulli) trials.
- A prior probability distribution represents initial uncertainty about parameters.
- As data (evidence) is observed, a posterior distribution by conditionalizing on the data (or evidence) that we've observed.
- Under broad conditions, Bayesian updating converges to the true probability distribution.
- (Later, we will discuss cases where convergence may fail.)

Modern Bayesianism

(3) **A More General Model of Science**

- The Bernoulli model assumes empirical research involves binary outcomes (yes/no).
- In reality, probabilities may change over time, meaning trials are not always independent.
- Bayesian methods also work with general probability distributions.
- Additionally, Bayesian inference is not limited to simple yes/no questions-we can allow any possible hypotheses formulated in an appropriate formal language.

Problems for Bayesianism

1. Convergence Issues: When Bayesian Updating Fails

(1) **Problem of Zero Prior Probability**

- If a Bayesian agent assigns prior probability 0 to the true value of p, no amount of evidence will update their belief toward it.
- Example: If we assume a prior that excludes $p = 1$ for the probability of sunrise, we will never converge to this truth, no matter how many sunrises we observe.
- This highlights the risk of overly restrictive or dogmatic priors.

Problems for Bayesianism

(2) **Underdetermination by Data**

- If two competing hypotheses predict identical probabilities for observed data, Bayesian updating does not favor one over the other.
- This leaves Bayesianism unable to refute metaphysical skepticism (e.g., brain-in-a-vat scenarios).
- Since all observations are equally likely under both hypotheses, Bayesianism permits strong belief even in brain-in-a-vat scenarios.

Problems for Bayesianism

2. Unrealistic Rationality Assumptions

(1) Assumption of Perfect Data Reliability

- Bayesian reasoning assumes that all observations are true.
- Problem: Perceptual errors exist (e.g., mistaking a color under different lighting conditions).
- A possible solution is to model sensory data separately, treating interpretation as another Bayesian process.

(2) Unbounded Computational Requirements

- Bayesian agents must track all hypotheses, compute probability updates, and reason without logical errors.
- This is computationally infeasible, especially as hypotheses become complex.
- Worse: Some Bayesian updates require solving non-computable problems, meaning even theoretical computers cannot perform them.

Problems for Bayesianism

Possible responses:

- It's not a problem per se to (first) focus on the 'idealized' case of rational agents (before looking at actual agents).
- One can understand Bayesianism as a normative theory: how agents should reason as opposed to how they actually do. (Many take logic to be normative in this sense: describing correct reasoning rather than the psychology of reasoning.)

Problems for Bayesianism

3. The Problem of Inductive Assumptions

- Bayesian inference **assumes** the uniformity of nature-that the future resembles the past.
- What if the true probability changes unpredictably every day? In such a world, no learning occurs, even with infinite data.
- Bayesian inference requires that all observations come from a stable probability distribution.
  - This is typically ensured by assuming independent and identically distributed (i.i.d.) samples.
  - While independence can sometimes be relaxed (e.g., Markov models), identical distribution remains essential.

Problems for Bayesianism

Possible Responses: Defending Bayesianism

- One defense is that assuming stability is optimal:
  - If the world is stable, Bayesian inference works best.
  - If the world is unstable, no method works better, making Bayesianism no worse than alternatives.
- Reichenbach argued that induction is not just useful but dominant:
  - In some worlds, it is superior to all other methods.
  - In no logically possible world is it worse than any other method.

Summary and Current State

- The Bayesian approach provides a structured solution to the problem of induction:
- While we can never achieve absolute certainty about an empirical law, we can rigorously quantify how our confidence increases with accumulating evidence.
- Through Bayesian updating, the probability assigned to a correct hypothesis grows as more observations are collected.
- Under broad conditions, this learning process converges to the truth, making Bayesian inference a powerful tool for scientific reasoning.

Summary and Current State

- This offers a nuanced solution to the problem of induction:
  - Instead of absolute certainty, we work with degrees of belief that continuously improve with evidence.
  - Traditional notions of knowledge, justification, and reliability are reframed in terms of probabilistic confidence.
  - Bayesian induction is not universally reliable in all logically possible worlds but is effective in the worlds that matter to us.
- In the next lecture, we will explore non-probabilistic approaches to induction and their philosophical implications.

Exercises

1. Do think the Bernoulli model of empirical inquiry is a plausible description of doing science? (E.g., what if the relevant part of nature is more complicated than just a binary yes/no question, i.e., an urn with two outcomes?) Are your objections solved by the more general Bayesian model? Do you have objections to it as well?
2. Discuss the Bernoulli reply to inductive skepticism: is increasingly justified belief really almost as good as knowledge?
3. Do some calculations in the slides 'The Bayesian twist' above: e.g., more examples of a and b as on slide to see how fast the convergence is.

Exercises

4. Compare the frequentist and Bayesian interpretation of probability. Do you think one is a better analysis of our concept of probability or do you think they are incomparable and capture two different meanings of ‘probability’?
5. Consider the Bayesian formalization of the concept of rationality. What do you think does it get right, what does it miss?
6. Discuss the non-computability objection to Bayesianism: e.g., how bad is theoretical non-computability for a philosophical theory if in most practical applications it can be avoided? Can you find further responses?