# Epistemology and theory of machine learning

## Part 1: Impossibility results

Tom Sterkenburg

CWI
Centrum Wiskunde & Informatica

Emmy Noether-Programm
DFG Deutsche Forschungsgemeinschaft

LMU
LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
FAKULTÄT FÜR PHILOSOPHIE, WISSENSCHAFTSTHEORIE UND RELIGIONSWISSENSCHAFT
MUNICH CENTER FOR MATHEMATICAL PHILOSOPHY

HLRS Summer School Trust and ML
Stuttgart, July 2023

# Can we trust our machine learning algorithms?

- Do we have reasons for thinking that our machine learning algorithms *learn well*?
  - **Epistemic** trust—**reliability**.
  - The **mathematical theory** of machine learning promises formal learning guarantees.

# Philosophy of science

- Methodology of science.
- Is there something like **scientific method**?
- If so, can we **justify** this method, that is, provide reason for why it is good?

- Formal methodology of science.
- Can we give a formal account of scientific method?
- Can we formally justify this method?

# The problem of induction

- Scientific reasoning is **inductive** reasoning.
- Hume: we cannot have a justification for inductive inferences.
- ▷ No purely mathematical argument will cut it.
- ▷ No extra-mathematical (empirical) argument will cut it.

- But how does this leave room for any formal theory of scientific method and its justification?

# The problem of induction and ML

- Machine learning algorithms likewise take an inductive or **generalization** leap.
- So they are likewise susceptible to Hume's argument.

- How does this leave room for any formal theory of machine learning methods and their justification?

# A puzzle?

- The **no-free-lunch theorems** of supervised learning suggest a *skeptical* conclusion about machine learning algorithms.
  - "All learning algorithms are equally lacking in epistemic justification."
  - "A standard procedure like empirical risk minimization is just as good as empirical risk *maximization*."

- At the same time, the business of **learning theory** is to show that some possible algorithms *are* better than others.
  - "We can *prove* that empirical risk minimization is a good method (and we couldn't for empirical risk *maximization*)."

# The plan

1. The no-free-lunch theorems.
2. The road to skepticism.
3. Ways out? Universal prediction.

# The plan

**1.** The no-free-lunch theorems.

**2.** The road to skepticism.

**3.** Ways out? Universal prediction.

**II.** Statistical learning theory and beyond.

# The plan

1. **The no-free-lunch theorems.**
2. The road to skepticism.
3. Ways out? Universal prediction.

# The no-free-lunch (NFL) theorems

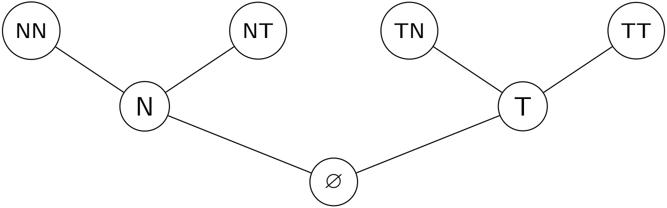
- Wolpert (1993,1996): "no free lunch theorems for supervised learning."
  - "All learning algorithms are a priori equivalent."
- Schaffer (1994): "conservation law of generalization performance."

# A simple version

- Every day we try to predict whether our breakfast will be tasty $(\mathrm{T})$, or not $(\mathrm{N})$.
- Our **learning algorithm** makes a guess whether breakfast will be tasty today, based on the days past.

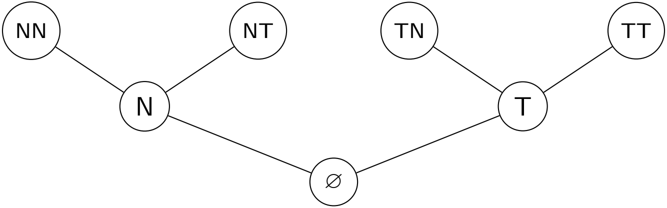
# A simple version

- Consider histories of two consecutive days.
- There are $2^2$ such histories or **learning situations**.



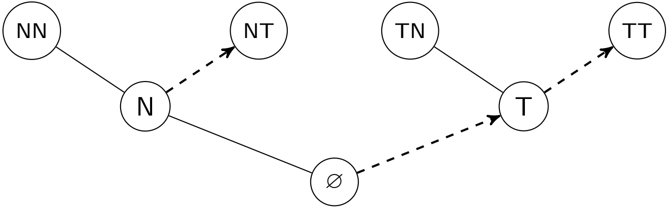
# A simple version

- Consider histories of two consecutive days.
- There are $2^2$ such histories or **learning situations**.
- There are $2^3$ different possible **learning algorithms** (functions from $\{\emptyset, \mathrm{T}, \mathrm{N}\}$ to $\{\mathrm{T}, \mathrm{N}\}$).



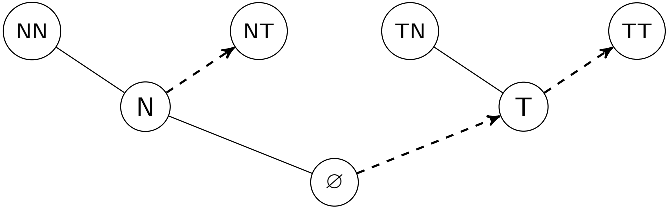
# A simple version

- Consider histories of two consecutive days.
- There are $2^2$ such histories or **learning situations**.
- There are $2^3$ different possible **learning algorithms** (functions from $\{\emptyset, \mathrm{T}, \mathrm{N}\}$ to $\{\mathrm{T}, \mathrm{N}\}$).



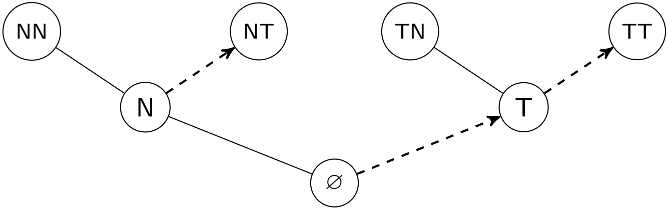
# A simple version

- A learning algorithm's **error** in a particular learning situation is its mean number of mistakes.
- Here, then, is an NFL statement: **every prediction algorithm attains the same error in *equally many* learning situations**.



# A simple version

- A learning algorithm's **error** in a particular learning situation is its mean number of mistakes.
- Here, then, is an NFL statement: **every prediction algorithm attains the same error in *equally many* learning situations**.
- Assume a *uniform* distribution on learning situations.
- Then we can say that **every learning method has the same expected error 1/2**.



# A reformulation

- The assumption of a uniform distribution on learning situations is not really well-motivated.
- In fact, this is, for the purpose of learning, really a *worst-case* assumption (cf. Boole, Peirce, Carnap, . . . )
  - "In a universe where learning is impossible, every learning algorithm is equivalent." Well, yes . . .

- But this assumption is actually not essential for a skeptical conclusion . . .

# A reformulation

- For every learning algorithm, there is a learning situation in which it is *not* successful, yet in which *another* learning algorithm *is* successful.
- There is no **universal** learning algorithm.
- ▷ Many modern formulations are of this form (e.g., Shalev-Shwartz & Ben-David, 2014).
- Every learning algorithm must come with some restrictive **inductive bias**.

# The plan

1. The no-free-lunch theorems.
2. **The road to skepticism.**
3. Ways out? Universal prediction.

# The road to skepticism

- We are concerned with a limited set of standard, generic, algorithms.
- What justification do we have for these standard learning algorithms?
  - NFL: these algorithms must have specific biases.
  - So, how do we justify these biases..?
- The world must have a structure that neatly matches these biases . . .
  - E.g., Giraud-Carrier and Provost's (2005) "weak assumption of machine learning" that "the process that presents us with learning problems [. . . ] induces a non-uniform probability distribution [over learning situations]."
  - OK, but how to justify such an assumption?

# The road to skepticism

- Hume's argument for inductive skepticism.
  - Inductive reasoning must proceed upon the supposition that the universe is *induction-friendly*.
  - What reason can we give for this supposition?
  - We certainly cannot give any *deductive*, a priori reason, because it's logically possible that the universe is *not* induction-friendly.
  - But we also cannot give a good *inductive* reason, because that would be circular!
  - Specifically, we cannot conclude from the success of inductive method so far (past evidence for induction-friendliness) that inductive method will remain successful (that the universe is, in fact, induction-friendly).
- So we're stuck.

# The road to skepticism

- We are concerned with a limited set of standard, generic, algorithms.
- What justification do we have for these standard learning algorithms?
  - NFL: these algorithms must have specific biases.
  - So, how do we justify these biases..?
- The world must have a structure that neatly matches these biases . . .
  - E.g., Giraud-Carrier and Provost's (2005) "weak assumption of machine learning" that "the process that presents us with learning problems [. . . ] induces a non-uniform probability distribution [over learning situations]."
  - OK, but how to justify such an assumption..?
- So we're stuck.

# The plan

1. The no-free-lunch theorems.
2. The road to skepticism.
3. **Ways out? Universal prediction.**

# Ways out . . . ?

- Maybe we have just been overly demanding?

- Suppose we lower our aims to **estimating the limiting relative frequency** of tasty breakfasts, as the days go by?
  - That is, we are interested in estimating $\lim_{\text{days}\to\infty} \frac{\text{tasty days}}{\text{days}}$.
  - For instance, if breakfasts are only tasty in the weekends, the limiting relative frequency is 2/7.
  - Let's adopt the **straight rule** method that always outputs the estimate $\frac{\text{tasty days}}{\text{days}}$.
- The estimates of the straight rule will get closer and closer to the limiting relative frequency (if it exists).

# The pragmatic justification of induction

- Reichenbach: the inductive method is successful, whenever success is attainable at all.
  - The straight rule converges on the limiting relative frequency, whenever there exists a limiting relative frequency at all.

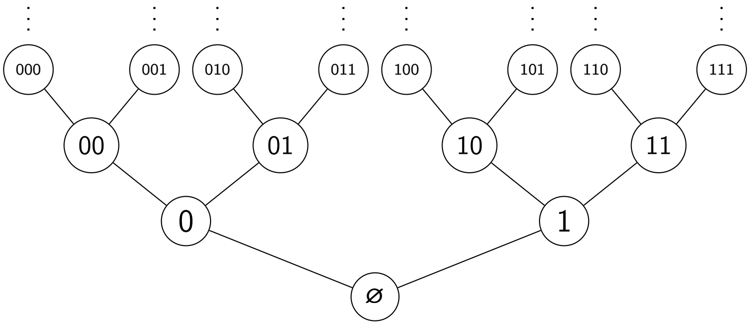
- The straight rule method is a **universal method** for Reichenbach's estimation problem.
  - So the no-free-lunch theorem does not hold for all learning problems.
- But Reichenbach's learning problem is slightly trivial . . .

# Ways out . . . ?

- Let's try to be a little more demanding again . . .

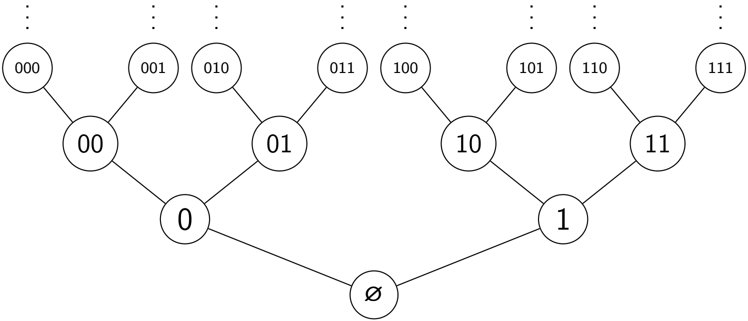
- Suppose we still want to successfully predict next outcomes, but we weaken our success criterion to predicting successfully in the limit.
  - Let's also suppose our predictions can be probabilistic (we always issue a probability of the next breakfast being tasty).
  - Then we want the predicted probabilities to converge to 1 for tasty days and 0 for nontasty days.

# Sequential prediction



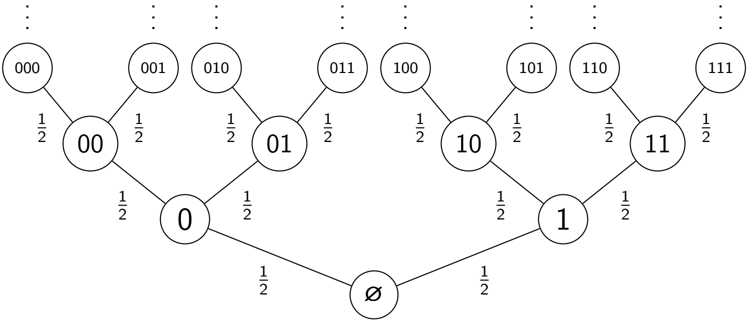
# Sequential prediction

- A **prediction method** is a function $\mathrm{pred}$ from $\{0,1\}^*$ to $[0,1]$



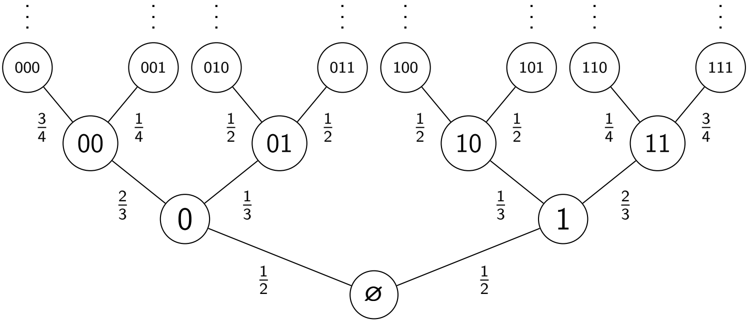
# Sequential prediction

- A **prediction method** is a function $\mathrm{pred}$ from $\{0,1\}^*$ to $[0,1]$
- Example: always fifty-fifty, $\mathsf{p}(\boldsymbol{x}) = \frac{1}{2}$ for all $\boldsymbol{x} \in \{0,1\}^*$.



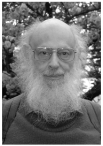
# Sequential prediction

- A **prediction method** is a function $\mathrm{pred}$ from $\{0,1\}^*$ to $[0,1]$.
- Example: always fifty-fifty, $\mathrm{pred}(\boldsymbol{x}) = \frac{1}{2}$ for all $\boldsymbol{x} \in \{0,1\}^*$.
- Example: Laplace's rule of succession, $\mathsf{p}(\boldsymbol{x}) = \frac{\#_1\boldsymbol{x}+1}{\#\boldsymbol{x}+2}$.



# Universal prediction

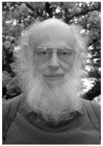
- The promise of **algorithmic information theory**, in particular the predictive theory developed by Solomonoff (1964) and Levin (1970).
  - A **universal prediction** method that predicts by **data compression**.
  - Formalized in terms of **Kolmogorov complexity**.



- For every **computable data-generating distribution**, this prediction method will (with probability 1) converge on the right probabilities.

# Universal prediction

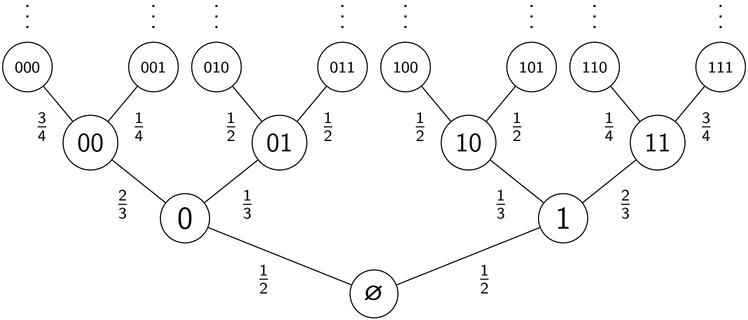
- The promise of **algorithmic information theory**, in particular the predictive theory developed by Solomonoff (1964) and Levin (1970).
  - A **universal prediction** method that predicts by **data compression**.
  - Formalized in terms of **Kolmogorov complexity**.



- For every **computable data-generating distribution**, this prediction method will (with probability 1) converge on the right probabilities.

# Carnap's inductive logic

# Carnap's inductive logic

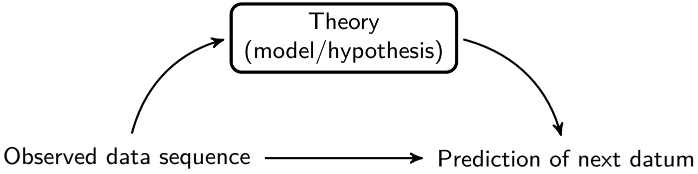


# Carnap's inductive logic

# Putnam v. Carnap's inductive logic

# Putnam:

- “Certainly it appears implausible to say that there is a *rule* whereby one can go from the observational facts (if only one had them all written out) to the observational prediction without any ‘detour’ into the realm of theory.”



Theory (model/hypothesis)

Observed data sequence

Prediction of next datum

- “. . . we get the further consequence that it is possible in principle to build an electronic computer such that it would always make the best prediction–i.e. the prediction that would be made by the best possible scientist if he had the best possible theories. *Science could in principle be done by a moron* (or an electronic computer).”

# Putnam's diagonal argument

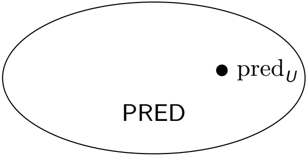
- The following two conditions on a universal prediction method are incompatible.
  - It should converge on **all computable patterns**.
  - It should be **computable**.

# Putnam's diagonal argument

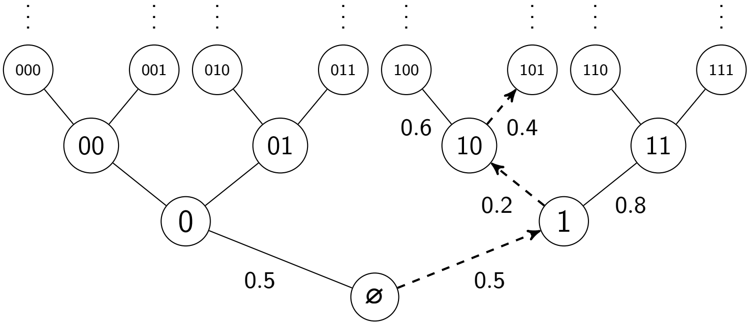
- The following two conditions on a universal prediction method are incompatible.
  - It should do well whenever *some* (**computable**!) method does well.
  - It should be **computable**.

# Putnam's diagonal argument

- The following two conditions on a universal prediction method are incompatible.
- It should be **universal** among all possible (**computable**!) methods.
- It should be **computable**.

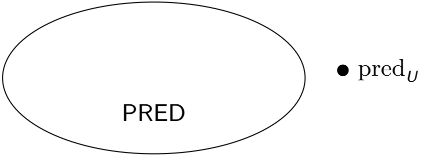


# Putnam's diagonal argument



# Putnam's diagonal argument

- The following two conditions on a universal prediction method are incompatible.
  - It should be **universal** among all possible (**computable**!) methods.
  - It should be **computable**.

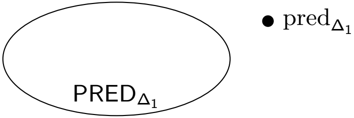


PRED

$\bullet \ \mathrm{pred}_U$

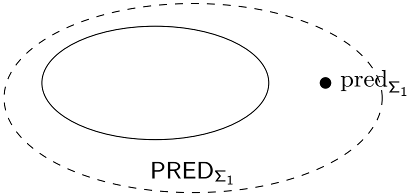
# The Solomonoff-Levin method

- Try to escape diagonalization by expanding to a class of computably approximable methods, that *does* contain ‘universal elements.’



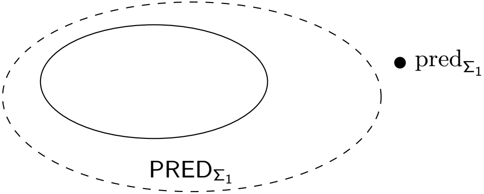
# The Solomonoff-Levin method

- Try to escape diagonalization by expanding to a class of computably approximable methods, that *does* contain ‘universal elements.’



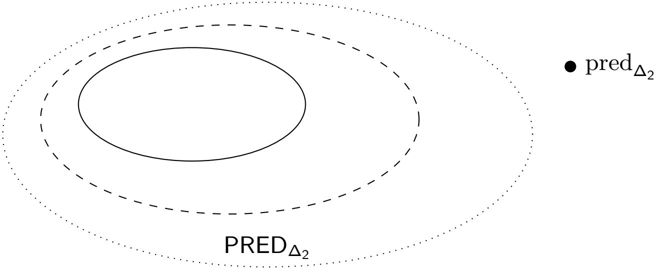
# The Solomonoff-Levin method

- Try to escape diagonalization by expanding to a class of computably approximable methods, that *does* contain 'universal elements.'



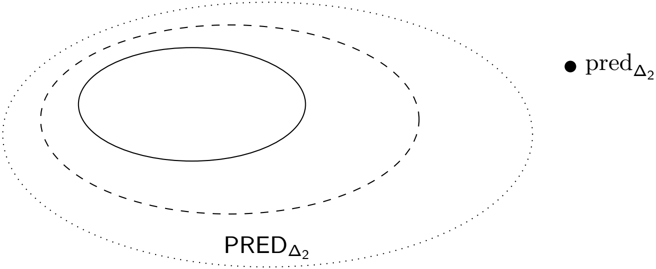
# The Solomonoff-Levin method

- Try to escape diagonalization by expanding to a class of computably approximable methods, that *does* contain 'universal elements.'



# The Solomonoff-Levin method

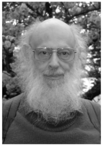
- Try to escape diagonalization by expanding to a class of computably approximable methods, that *does* contain ‘universal elements.’


$\bullet\ \mathrm{pred}_{\Delta_2}$
$\mathrm{PRED}_{\Delta_2}$

- In sum: it can’t be done.

# No universal prediction method

- So this attempt to escape the no-free-lunch theorems didn't work—algorithmic information theory can't really help us.



- In general, no-free-lunch results for interesting learning problems seem inescapable.
- But then how can machine learning theory still have a constructive story to tell—how can it give some kind of epistemic justification for our standard learning algorithms?

# To conclude: looking back and forwards

**In this lecture, we discussed the no-free-lunch theorems, and how they seem to obstruct the possibility of a justification for our machine learning algorithms.**

**In the next, we will see how these negative results do still leave room for a positive story.**

tom.sterkenburg@lmu.de