# Table of Contents

1. **A SHORT TOUR THROUGH THE COURSE**
2. **INFORMATION THEORY & KOLMOGOROV COMPLEXITY**
3. **BAYESIAN PROBABILITY THEORY**
4. **ALGORITHMIC PROBABILITY & UNIVERSAL INDUCTION**
5. **MINIMUM DESCRIPTION LENGTH**
6. **THE UNIVERSAL SIMILARITY METRIC**
7. **BAYESIAN SEQUENCE PREDICTION**
8. **UNIVERSAL RATIONAL AGENTS**
9. **THEORY OF RATIONAL AGENTS**
10. **APPROXIMATIONS & APPLICATIONS**
11. **DISCUSSION**
1 A SHORT TOUR THROUGH THE COURSE
Informal Definition of (Artificial) Intelligence

Intelligence measures an agent’s ability to achieve goals in a wide range of environments. [S. Legg and M. Hutter]

Emergent: Features such as the ability to learn and adapt, or to understand, are implicit in the above definition as these capacities enable an agent to succeed in a wide range of environments.

The science of Artificial Intelligence is concerned with the construction of intelligent systems/artifacts/agents and their analysis.

What next? Substantiate all terms above: agent, ability, utility, goal, success, learn, adapt, environment, ...

Never trust a theory if it is not supported by an experiment.
Induction $\rightarrow$ Prediction $\rightarrow$ Decision $\rightarrow$ Action

Having or acquiring or learning or inducing a model of the environment an agent interacts with allows the agent to make predictions and utilize them in its decision process of finding a good next action.

Induction infers general models from specific observations/facts/data, usually exhibiting regularities or properties or relations in the latter.

Example

Induction: Find a model of the world economy.

Prediction: Use the model for predicting the future stock market.

Decision: Decide whether to invest assets in stocks or bonds.

Action: Trading large quantities of stocks influences the market.
Science \approx Induction \approx Occam’s Razor

- Grue Emerald Paradox:

  *Hypothesis 1:* All emeralds are green.
  
  *Hypothesis 2:* All emeralds found till y2020 are green, thereafter all emeralds are blue.

- Which hypothesis is more plausible? **H1!** Justification?

- Occam's razor: take simplest hypothesis consistent with data. Is the most important principle in machine learning and science.

- Problem: How to quantify “simplicity”? Beauty? Elegance? Description Length!

[The Grue problem goes much deeper. This is only half of the story]
Information Theory & Kolmogorov Complexity

- Quantification/interpretation of Occam’s razor:
- Shortest description of object is best explanation.
- Shortest program for a string on a Turing machine $T$ leads to best extrapolation=prediction.

$$K_T(x) = \min_p \{ \ell(p) : T(p) = x \}$$

- Prediction is best for a universal Turing machine $U$.

$$\text{Kolmogorov-complexity}(x) = K(x) = K_U(x) \leq K_T(x) + c_T$$
Bayesian Probability Theory

Given (1): Models $P(D|H_i)$ for probability of observing data $D$, when $H_i$ is true.

Given (2): Prior probability over hypotheses $P(H_i)$.

Goal: Posterior probability $P(H_i|D)$ of $H_i$, after having seen data $D$.

Solution:

Bayes’ rule:

$$P(H_i|D) = \frac{P(D|H_i) \cdot P(H_i)}{\sum_i P(D|H_i) \cdot P(H_i)}$$

(1) Models $P(D|H_i)$ usually easy to describe (objective probabilities)

(2) But Bayesian prob. theory does not tell us how to choose the prior $P(H_i)$ (subjective probabilities)
Algorithmic Probability Theory

• **Epicurus:** If more than one theory is consistent with the observations, keep all theories.

• ⇒ uniform prior over all $H_i$?

• Refinement with **Occam’s razor** quantified in terms of **Kolmogorov complexity**:

$$P(H_i) := 2^{-K_{T/U}(H_i)}$$

• **Fixing** $T$ we have a complete theory for prediction.  
  **Problem:** How to choose $T$.

• **Choosing** $U$ we have a universal theory for prediction.  
  **Observation:** Particular choice of $U$ does not matter much.  
  **Problem:** Incomputable.
Inductive Inference & Universal Forecasting

- Solomonoff combined Occam, Epicurus, Bayes, and Turing into one formal theory of sequential prediction.

\[ M(x) = \text{probability that a universal Turing machine outputs } x \text{ when provided with fair coin flips on the input tape}. \]

- A posteriori probability of \( y \) given \( x \) is \( M(y|x) = M(xy)/M(x) \).

- Given \( \dot{x}_1, \ldots, \dot{x}_{t-1} \), the probability of \( x_t \) is \( M(x_t|\dot{x}_1\ldots\dot{x}_{t-1}) \).

- Immediate “applications”:
  - Weather forecasting: \( x_t \in \{\text{sun, rain}\} \).
  - Stock-market prediction: \( x_t \in \{\text{bear, bull}\} \).
  - Continuing number sequences in an IQ test: \( x_t \in \mathbb{N} \).

- Optimal universal inductive reasoning system!
The Minimum Description Length Principle

• **Approximation** of Solomonoff, since $M$ is incomputable:
  
  $M(x) \approx 2^{-K_U(x)}$ (quite good)

• $K_U(x) \approx K_T(x)$ (very crude)

• Predict $y$ of highest $M(y|x)$ is approximately same as

• **MDL:** Predict $y$ of smallest $K_T(xy)$. 
Application: Universal Clustering

- **Question:** When is object $x$ similar to object $y$?

- **Universal solution:** $x$ similar to $y$
  $\iff$ $x$ can be easily (re)constructed from $y$
  $\iff$ $K(x|y) := \min\{\ell(p) : U(p, y) = x\}$ is small.

- **Universal Similarity:** Symmetrize&normalize $K(x|y)$.

- **Normalized compression distance:** Approximate $K \equiv K_U$ by $K_T$.

- **Practice:** For $T$ choose (de)compressor like lzw or gzip or bzip(2).

- **Multiple objects $\Rightarrow$ similarity matrix $\Rightarrow$ similarity tree.

- **Applications:** Completely automatic reconstruction (a) of the evolutionary tree of 24 mammals based on complete mtDNA, and (b) of the classification tree of 52 languages based on the declaration of human rights and (c) many others. [Cilibrasi&Vitanyi'05]
Sequential Decision Theory

Setup: For $t = 1, 2, 3, 4, ...$

Given sequence $x_1, x_2, ..., x_{t-1}$

1. predict/make decision $y_t$,
2. observe $x_t$,
3. suffer loss $\text{Loss}(x_t, y_t)$,
4. $t \rightarrow t + 1$, goto (1)

Goal: Minimize expected Loss.

Greedy minimization of expected loss is optimal if:

Important: Decision $y_t$ does not influence env. (future observations).
Loss function is known.

Problem: Expectation w.r.t. what?

Solution: W.r.t. universal distribution $M$ if true distr. is unknown.
Example: Weather Forecasting

Observation \( x_t \in \mathcal{X} = \{\text{sunny, rainy}\} \)

Decision \( y_t \in \mathcal{Y} = \{\text{umbrella, sunglasses}\} \)

<table>
<thead>
<tr>
<th>Loss</th>
<th>sunny</th>
<th>rainy</th>
</tr>
</thead>
<tbody>
<tr>
<td>umbrella</td>
<td>0.1</td>
<td>0.3</td>
</tr>
<tr>
<td>sunglasses</td>
<td>0.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Taking umbrella/sunglasses does not influence future weather
(ignoring butterfly effect)
Agent Model with Reward

if actions/decisions \( a \) influence the environment \( q \)
Rational Agents in Known Environment

- **Setup:** Known deterministic or probabilistic environment
- **Fields:** AI planning & sequential decision theory & control theory
- **Greedy maximization of reward** $r (=-\text{Loss})$ no longer optimal.
  - **Example:** Chess
- **Exploration versus exploitation problem.**
  - $\Rightarrow$ Agent has to be farsighted.
- **Optimal solution:** Maximize future (expected) reward sum, called value.
- **Problem:** Things drastically change if environment is unknown
Rational Agents in Unknown Environment

Additional problem: (probabilistic) environment unknown.

Fields: reinforcement learning and adaptive control theory

Bayesian approach: Mixture distribution $\xi$.


2. Computationally very hard problem.

3. Choice of horizon? Immortal agents are lazy.

Universal Solomonoff mixture $\Rightarrow$ universal agent AIXI.

Represents a formal (math., non-comp.) solution to the AI problem? Most (all AI?) problems are easily phrased within AIXI.
Computational Issues: Universal Search

- **Levin search**: Fastest algorithm for inversion and optimization problems.

- **Theoretical application**: Assume somebody found a non-constructive proof of P=NP, then Levin-search is a polynomial time algorithm for every NP (complete) problem.

- **Practical (OOPS) applications** (J. Schmidhuber)
  Mazes, towers of hanoi, robotics, ...

- **FastPrg**: The asymptotically fastest and shortest algorithm for all well-defined problems.

- **Computable Approximations of AIXI**: AIXItl and AIξ and MC-AIXI-CTW and ΦMDP.

- **Human Knowledge Compression Prize**: (50’000€)
without providing any domain knowledge, the same agent is able to self-adapt to a diverse range of interactive environments.

www.youtube.com/watch?v=yfsMHtmGDKE
Discussion at End of Course

- What has been achieved?
- Made assumptions.
- General and personal remarks.
- Open problems.
- Philosophical issues.
Exercises

1. [C10] What is the probability $p$ that the sun will rise tomorrow,

2. [C15] Justify Laplace’ rule ($p = \frac{n+1}{n+2}$, where $n =$ #days sun rose in past)

3. [C05] Predict sequences:
   - 2,3,5,7,11,13,17,19,23,29,31,37,41,43,47,53,59,?
   - 3,1,4,1,5,9,2,6,5,3,?,
   - 1,2,3,4,?

4. [C10] Argue in (1) and (3) for different continuations.
Introductory Literature


See [http://www.hutter1.net/ai/introref.htm](http://www.hutter1.net/ai/introref.htm) for more.