Sequential Decisions based on Algorithmic Probability

Marcus Hutter
Canberra, ACT, 0200, Australia

Course COMP4670/COMP6467: Reinforcement Learning and Planning under Uncertainty
Overview

- **Setup:** Given (non)iid data $D = (x_1, ..., x_n)$, predict $x_{n+1}$
- **Ultimate goal** is to maximize profit or minimize loss
- **Consider Models/Hypothesis** $H_i \in \mathcal{M}$
- **Max.Likelihood:** $H_{best} = \arg \max_i p(D|H_i)$ (overfits if $\mathcal{M}$ large)
- **Bayes:** Posterior probability of $H_i$ is $p(H_i|D) \propto p(D|H_i)p(H_i)$
- **MDL:** $H_{best} = \arg \min_{H_i} \{\text{CodeLength}(D|H_i) + \text{CodeLength}(H_i)\}$
  (Complexity penalization)
- Bayes needs $\text{prior}(H_i)$, MDL needs $\text{CodeLength}(H_i)$
- **Occam+Epicurus:** High prior for simple models with short codes.
- **Kolmogorov/Solomonoff:** Quantification of simplicity/complexity
- **MDL & Bayes** work if $D$ is sampled from $H_{true} \in \mathcal{M}$
- **Universal AI** = Universal Induction + Sequential Decision Theory
Abstract

Machine learning is concerned with developing algorithms that learn from experience, build models of the environment from the acquired knowledge, and use these models for prediction. Machine learning is usually taught as a bunch of methods that can solve a bunch of problems (see my Introduction to SML last week). The following lecture takes a step back and asks about the foundations of machine learning, in particular the (philosophical) problem of inductive inference, (Bayesian) statistics, and artificial intelligence. The lecture concentrates on principled, unified, and exact methods.
Table of Contents

• Overview
• Philosophical Issues
• Bayesian Sequence Prediction
• Universal Inductive Inference
• Minimum Description Length Principle
• The Universal Similarity Metric
• Universal Artificial Intelligence
• Wrap Up
• Literature
Philosophical Issues: Contents

- Philosophical Problems
- On the Foundations of Machine Learning
- Example 1: Probability of Sunrise Tomorrow
- Example 2: Digits of a Computable Number
- Example 3: Number Sequences
- Occam’s Razor to the Rescue
- Grue Emerald and Confirmation Paradoxes
- What this Lecture is (Not) About
- Sequential/Online Prediction – Setup
Philosophical Issues: Abstract

I start by considering the philosophical problems concerning machine learning in general and induction in particular. I illustrate the problems and their intuitive solution on various (classical) induction examples. The common principle to their solution is Occam’s simplicity principle. Based on Occam’s and Epicurus’ principle, Bayesian probability theory, and Turing’s universal machine, Solomonoff developed a formal theory of induction. I describe the sequential/online setup considered in this lecture and place it into the wider machine learning context.
Philosophical Problems

- Does inductive inference work? Why? How?
- When is an individual sequence random?
- What does probability mean?
- How to choose the model class?
- How to choose the prior?
- How to make optimal decisions in unknown environments?
- What is intelligence?
On the Foundations of Machine Learning

• Example: *Algorithm/complexity theory*: The goal is to find fast algorithms solving problems and to show lower bounds on their computation time. Everything is rigorously defined: algorithm, Turing machine, problem classes, computation time, ...

• Most *disciplines* start with an informal way of attacking a subject. With time they get more and more formalized often to a point where they are completely rigorous. Examples: set theory, logical reasoning, proof theory, probability theory, infinitesimal calculus, energy, temperature, quantum field theory, ...

• *Machine learning*: Tries to build and understand systems that learn from past data, make good prediction, are able to generalize, act intelligently, ... Many terms are only vaguely defined or there are many alternate definitions.
**Example 1: Probability of Sunrise Tomorrow**

What is the probability \( p(1|1^d) \) that the sun will rise tomorrow? 
\( (d = \text{past \# days sun rose}, \ 1 = \text{sun rises. \ 0 = sun will not rise}) \)

- \( p \) is undefined, because there has never been an experiment that tested the existence of the sun *tomorrow* (ref. class problem).

- The \( p = 1 \), because the sun rose in all past experiments.

- \( p = 1 - \epsilon \), where \( \epsilon \) is the proportion of stars that explode per day.

- \( p = \frac{d+1}{d+2} \), which is Laplace rule derived from Bayes rule.

- Derive \( p \) from the type, age, size and temperature of the sun, even though we never observed another star with those exact properties.

**Conclusion:** We predict that the sun will rise tomorrow with high probability independent of the justification.
Example 2: Digits of a Computable Number

- Extend 14159265358979323846264338327950288419716939937?

- Looks random?!

- Frequency estimate: $n =$ length of sequence. $k_i =$ number of occurred $i \implies$ Probability of next digit being $i$ is $\frac{i}{n}$. Asymptotically $\frac{i}{n} \to \frac{1}{10}$ (seems to be) true.

- But we have the strong feeling that (i.e. with high probability) the next digit will be 5 because the previous digits were the expansion of $\pi$.

- Conclusion: We prefer answer 5, since we see more structure in the sequence than just random digits.
Example 3: Number Sequences

Sequence: \( x_1, x_2, x_3, x_4, x_5, \ldots \)

- \( x_5 = 5 \), since \( x_i = i \) for \( i = 1\ldots 4 \).
- \( x_5 = 29 \), since \( x_i = i^4 - 10i^3 + 35i^2 - 49i + 24 \).

Conclusion: We prefer 5, since linear relation involves less arbitrary parameters than 4th-order polynomial.

Sequence: \( 2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41, 43, 47, 53, 59, \ldots \)

- 61, since this is the next prime
- 60, since this is the order of the next simple group

Conclusion: We prefer answer 61, since primes are a more familiar concept than simple groups.

On-Line Encyclopedia of Integer Sequences:
http://www.research.att.com/~njas/sequences/
Occam’s Razor to the Rescue

• Is there a unique principle which allows us to formally arrive at a prediction which
  - coincides (always?) with our intuitive guess -or- even better,
  - which is (in some sense) most likely the best or correct answer?

• Yes! Occam’s razor: Use the simplest explanation consistent with past data (and use it for prediction).

• Works! For examples presented and for many more.

• Actually Occam’s razor can serve as a foundation of machine learning in general, and is even a fundamental principle (or maybe even the mere definition) of science.

• Problem: Not a formal/mathematical objective principle. What is simple for one may be complicated for another.
Grue Emerald Paradox

Hypothesis 1: All emeralds are green.

Hypothesis 2: All emeralds found till y2010 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.
**Confirmation Paradox**

(i) $R \rightarrow B$ is confirmed by an $R$-instance with property $B$

(ii) $\neg B \rightarrow \neg R$ is confirmed by a $\neg B$-instance with property $\neg R$.

(iii) Since $R \rightarrow B$ and $\neg B \rightarrow \neg R$ are logically equivalent, $R \rightarrow B$ is also confirmed by a $\neg B$-instance with property $\neg R$.

**Example:** Hypothesis (o): All ravens are black ($R=$Raven, $B=$Black).

(i) observing a Black Raven confirms Hypothesis (o).

(iii) observing a White Sock also confirms that all Ravens are Black, since a White Sock is a non-Raven which is non-Black.

This conclusion sounds absurd.
Problem Setup

- Induction problems can be phrased as sequence prediction tasks.
- Classification is a special case of sequence prediction.
  (With some tricks the other direction is also true)
- This lecture focusses on maximizing profit (minimizing loss).
  We’re not (primarily) interested in finding a (true/predictive/causal) model.
- Separating noise from data is not necessary in this setting!
## What This Lecture is (Not) About

Dichotomies in Artificial Intelligence & Machine Learning

<table>
<thead>
<tr>
<th>scope of my lecture</th>
<th>scope of other lectures</th>
</tr>
</thead>
<tbody>
<tr>
<td>(machine) learning</td>
<td>(GOFAI) knowledge-based</td>
</tr>
<tr>
<td>statistical</td>
<td>logic-based</td>
</tr>
<tr>
<td>decision ⇔ prediction</td>
<td>induction ⇔ action</td>
</tr>
<tr>
<td>classification</td>
<td>regression</td>
</tr>
<tr>
<td>sequential / non-iid</td>
<td>independent identically distributed</td>
</tr>
<tr>
<td>online learning</td>
<td>offline/batch learning</td>
</tr>
<tr>
<td>passive prediction</td>
<td>active learning</td>
</tr>
<tr>
<td>Bayes ⇔ MDL</td>
<td>Expert ⇔ Frequentist</td>
</tr>
<tr>
<td>uninformed / universal</td>
<td>informed / problem-specific</td>
</tr>
<tr>
<td>conceptual/mathematical issues</td>
<td>computational issues</td>
</tr>
<tr>
<td>exact/principled</td>
<td>heuristic</td>
</tr>
<tr>
<td>supervised learning</td>
<td>unsupervised ⇔ RL learning</td>
</tr>
<tr>
<td>exploitation</td>
<td>exploration</td>
</tr>
</tbody>
</table>
Sequential/Online Prediction – Setup

In sequential or online prediction, for times \( t = 1, 2, 3, \ldots \), our predictor \( p \) makes a prediction \( y^p_t \in \mathcal{Y} \) based on past observations \( x_1, \ldots, x_{t-1} \).

Thereafter \( x_t \in \mathcal{X} \) is observed and \( p \) suffers \( \text{Loss}(x_t, y^p_t) \).

The goal is to design predictors with small total loss or cumulative loss:

\[
\text{Loss}_{1:T}(p) := \sum_{t=1}^{T} \text{Loss}(x_t, y^p_t).
\]

Applications are abundant, e.g. weather or stock market forecasting.

Example:

\[
\begin{array}{c|cc}
\text{Loss}(x, y) & \mathcal{X} = \{\text{sunny, rainy}\} \\
\mathcal{Y} = \{\text{umbrella, sunglasses}\} & 0.1 & 0.3 \\
& 0.0 & 1.0
\end{array}
\]

Setup also includes: Classification and Regression problems.
Bayesian Sequence Prediction: Contents

- Uncertainty and Probability
- Frequency Interpretation: Counting
- Objective Interpretation: Uncertain Events
- Subjective Interpretation: Degrees of Belief
- Bayes’ and Laplace’s Rules
- Envelope Paradox
- The Bayes-mixture distribution
- Relative Entropy and Bound
- Predictive Convergence
- Sequential Decisions and Loss Bounds
- Generalization: Continuous Probability Classes
- Summary
Bayesian Sequence Prediction: Abstract

The aim of probability theory is to describe uncertainty. There are various sources and interpretations of uncertainty. I compare the frequency, objective, and subjective probabilities, and show that they all respect the same rules, and derive Bayes’ and Laplace’s famous and fundamental rules. Then I concentrate on general sequence prediction tasks. I define the Bayes mixture distribution and show that the posterior converges rapidly to the true posterior by exploiting some bounds on the relative entropy. Finally I show that the mixture predictor is also optimal in a decision-theoretic sense w.r.t. any bounded loss function.
Uncertainty and Probability

The aim of probability theory is to describe uncertainty.

Sources/interpretations for uncertainty:

- **Frequentist**: probabilities are relative frequencies.
  (e.g. the relative frequency of tossing head.)

- **Objectivist**: probabilities are real aspects of the world.
  (e.g. the probability that some atom decays in the next hour)

- **Subjectivist**: probabilities describe an agent’s degree of belief.
  (e.g. it is (im)plausible that extraterrestrians exist)
Frequency Interpretation: Counting

- The frequentist interprets probabilities as relative frequencies.
- If in a sequence of \( n \) independent identically distributed (i.i.d.) experiments (trials) an event occurs \( k(n) \) times, the relative frequency of the event is \( k(n)/n \).
- The limit \( \lim_{n \to \infty} k(n)/n \) is defined as the probability of the event.
- For instance, the probability of the event head in a sequence of repeatedly tossing a fair coin is \( \frac{1}{2} \).
- The frequentist position is the easiest to grasp, but it has several shortcomings:
  - Problems: definition circular, limited to i.i.d, reference class problem.
Objective Interpretation: Uncertain Events

- For the objectivist probabilities are real aspects of the world.

- The outcome of an observation or an experiment is not deterministic, but involves physical random processes.

- The set $\Omega$ of all possible outcomes is called the sample space.

- It is said that an event $E \subset \Omega$ occurred if the outcome is in $E$.

- In the case of i.i.d. experiments the probabilities $p$ assigned to events $E$ should be interpretable as limiting frequencies, but the application is not limited to this case.

- (Some) probability axioms:
  
  $p(\Omega) = 1$ and $p(\{\}) = 0$ and $0 \leq p(E) \leq 1$.
  
  $p(A \cup B) = p(A) + p(B) - p(A \cap B)$.
  
  $p(B|A) = \frac{p(A \cap B)}{p(A)}$ is the probability of $B$ given event $A$ occurred.
Subjective Interpretation: Degrees of Belief

- The subjectivist uses probabilities to characterize an agent’s degree of belief in something, rather than to characterize physical random processes.

- This is the most relevant interpretation of probabilities in AI.

- We define the plausibility of an event as the degree of belief in the event, or the subjective probability of the event.

- It is natural to assume that plausibilities/beliefs $\text{Bel}(\cdot|\cdot)$ can be represented by real numbers, that the rules qualitatively correspond to common sense, and that the rules are mathematically consistent. $\Rightarrow$

- Cox’s theorem: $\text{Bel}(\cdot|A)$ is isomorphic to a probability function $p(\cdot|\cdot)$ that satisfies the axioms of (objective) probabilities.

- Conclusion: Beliefs follow the same rules as probabilities
Bayes’ Famous Rule

Let \( D \) be some possible data (i.e. \( D \) is event with \( p(D) > 0 \)) and \{\( H_i \)\}_{i \in I} be a countable complete class of mutually exclusive hypotheses (i.e. \( H_i \) are events with \( H_i \cap H_j = \emptyset \) \( \forall i \neq j \) and \( \bigcup_{i \in I} H_i = \Omega \)).

Given: \( p(H_i) = \) a priori plausibility of hypotheses \( H_i \) (subj. prob.)

Given: \( p(D|H_i) = \) likelihood of data \( D \) under hypothesis \( H_i \) (obj. prob.)

Goal: \( p(H_i|D) = \) a posteriori plausibility of hypothesis \( H_i \) (subj. prob.)

\[
\text{Solution: } p(H_i|D) = \frac{p(D|H_i)p(H_i)}{\sum_{i \in I} p(D|H_i)p(H_i)}
\]

Proof: From the definition of conditional probability and

\[
\sum_{i \in I} p(H_i|...) = 1 \quad \Rightarrow \quad \sum_{i \in I} p(D|H_i)p(H_i) = \sum_{i \in I} p(H_i|D)p(D) = p(D)
\]
Example: Bayes’ and Laplace’s Rule

Assume data is generated by a biased coin with head probability $\theta$, i.e. 
$H_\theta := \text{Bernoulli}(\theta)$ with $\theta \in \Theta := [0, 1]$.

Finite sequence: $x = x_1x_2...x_n$ with $n_1$ ones and $n_0$ zeros.

Sample infinite sequence: $\omega \in \Omega = \{0, 1\}^\infty$

Basic event: $\Gamma_x = \{\omega : \omega_1 = x_1, ..., \omega_n = x_n\} = \text{set of all sequences starting with } x$.

Data likelihood: $p_\theta(x) := p(\Gamma_x | H_\theta) = \theta^{n_1}(1 - \theta)^{n_0}$.

Bayes (1763): Uniform prior plausibility: $p(\theta) := p(H_\theta) = 1$

Evidence: $p(x) = \int_0^1 p_\theta(x)p(\theta) d\theta = \int_0^1 \theta^{n_1}(1 - \theta)^{n_0} d\theta = \frac{n_1!n_0!}{(n_0+n_1+1)!}$
Example: Bayes’ and Laplace’s Rule

Bayes: Posterior plausibility of $\theta$ after seeing $x$ is:

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} = \frac{(n+1)!}{n_1!n_0!} \theta^{n_1}(1-\theta)^{n_0}$$

Laplace: What is the probability of seeing 1 after having observed $x$?

$$p(x_{n+1} = 1|x_1...x_n) = \frac{p(x_1)}{p(x)} = \frac{n_1+1}{n+2}$$

Laplace believed that the sun had risen for $5000$ years $= 1'826'213$ days, so he concluded that the probability of doomsday tomorrow is $\frac{1}{1826215}$. 
Exercise: Envelope Paradox

• I offer you two closed envelopes, one of them contains twice the amount of money than the other. You are allowed to pick one and open it. Now you have two options. Keep the money or decide for the other envelope (which could double or half your gain).

• Symmetry argument: It doesn't matter whether you switch, the expected gain is the same.

• Refutation: With probability $p = 1/2$, the other envelope contains twice/half the amount, i.e. if you switch your expected gain increases by a factor $1.25 = (1/2)*2+(1/2)*(1/2)$.

• Present a Bayesian solution.
Notation: Strings & Probabilities

Strings: \( x = x_1 x_2 \ldots x_n \) with \( x_t \in \mathcal{X} \) and \( x_{1:n} := x_1 x_2 \ldots x_{n-1} x_n \) and \( x_{<n} := x_1 \ldots x_{n-1} \).

Probabilities: \( \sigma(x_1 \ldots x_n) \) is the probability that an (infinite) sequence starts with \( x_1 \ldots x_n \).

Conditional probability:

\[
\sigma_n := \sigma(x_n|x_{<n}) = \sigma(x_{1:n})/\sigma(x_{<n}),
\]

\[
\sigma(x_1 \ldots x_n) = \sigma(x_1) \cdot \sigma(x_2|x_1) \ldots \cdot \sigma(x_n|x_1 \ldots x_{n-1}).
\]

True data generating distribution: \( \mu \)
The Bayes-Mixture Distribution $\xi$

- Assumption: The true (objective) environment $\mu$ is unknown.

- Bayesian approach: Replace true probability distribution $\mu$ by a Bayes-mixture $\xi$.

- Assumption: We know that the true environment $\mu$ is contained in some known countable (in)finite set $\mathcal{M}$ of environments.

- The Bayes-mixture $\xi$ is defined as
  \[ \xi(x_{1:m}) := \sum_{\nu \in \mathcal{M}} w_{\nu} \nu(x_{1:m}) \quad \text{with} \quad \sum_{\nu \in \mathcal{M}} w_{\nu} = 1, \quad w_{\nu} > 0 \quad \forall \nu \]

- The weights $w_{\nu}$ may be interpreted as the prior degree of belief that the true environment is $\nu$, or $k_{\nu} = \ln w_{\nu}^{-1}$ as a complexity penalty (prefix code length) of environment $\nu$.

- Then $\xi(x_{1:m})$ could be interpreted as the prior subjective belief probability in observing $x_{1:m}$.
Relative Entropy

Relative entropy:  \( D(p||q) := \sum_i p_i \ln \frac{p_i}{q_i} \)

Properties:  \( D(p||q) \geq 0 \) and  \( D(p||q) = 0 \iff p = q \)

Instantaneous relative entropy:  \( d_t(x_{<t}) := \sum_{x_t \in \mathcal{X}} \mu(x_t|x_{<t}) \ln \frac{\mu(x_t|x_{<t})}{\xi(x_t|x_{<t})} \)

Total relative entropy:  \( D_n := \sum_{t=1}^{n} \mathbb{E}[d_t] \leq \ln w_{\mu}^{-1} \)

\( \mathbb{E}[f] \) = Expectation of  \( f \) w.r.t. the true distribution  \( \mu \), e.g.

If  \( f : \mathcal{X}^n \rightarrow \mathbb{R} \), then  \( \mathbb{E}[f] := \sum_{x_{1:n}} \mu(x_{1:n}) f(x_{1:n}) \).

Proof based on dominance or universality:  \( \xi(x) \geq w_{\mu} \mu(x) \).
Proof of the Entropy Bound

\[ D_n \equiv \sum_{t=1}^{n} \sum_{x_{<t}} \mu(x_{t}) \cdot d_t(x_{t}) \quad (a) \quad \sum_{t=1}^{n} \sum_{x_{1:t}} \mu(x_{1:t}) \ln \frac{\mu(x_t|x_{<t})}{\xi(x_t|x_{<t})} = \]

\[ \sum_{x_{1:n}} \mu(x_{1:n}) \ln \prod_{t=1}^{n} \frac{\mu(x_t|x_{<t})}{\xi(x_t|x_{<t})} \quad (c) \quad \sum_{x_{1:n}} \mu(x_{1:n}) \ln \frac{\mu(x_{1:n})}{\xi(x_{1:n})} \quad (d) \leq \ln w_{\mu}^{-1} \]

(a) Insert def. of \( d_t \) and used chain rule \( \mu(x_{<t}) \cdot \mu(x_t|x_{<t}) = \mu(x_{1:t}) \).

(b) \( \sum_{x_{1:t}} \mu(x_{1:t}) = \sum_{x_{1:n}} \mu(x_{1:n}) \) and argument of log is independent of \( x_{t+1:n} \). The \( t \) sum can now be exchanged with the \( x_{1:n} \) sum and transforms to a product inside the logarithm.

(c) Use chain rule again for \( \mu \) and \( \xi \).

(d) Use dominance \( \xi(x) \geq w_{\mu} \mu(x) \).
**Predictive Convergence**

Theorem: \( \xi(x_t|x_{<t}) \rightarrow \mu(x_t|x_{<t}) \) rapid w.p.1 for \( t \rightarrow \infty \)

Proof: \( D_\infty \equiv \sum_{t=1}^{\infty} E[d_t] \leq \ln w^{-1}_\mu \) and \( d_t \geq 0 \)

\[ \implies d_t \xrightarrow{t \rightarrow \infty} 0 \iff \xi_t \rightarrow \mu_t. \]

Fazit: \( \xi \) is excellent universal predictor if unknown \( \mu \) belongs to \( \mathcal{M} \).

How to choose \( \mathcal{M} \) and \( w_\mu \)? Both as large as possible?! More later.
**Sequential Decisions**

A prediction is very often the basis for some decision. The decision results in an action, which itself leads to some reward or loss.

Let \( \text{Loss}(x_t, y_t) \in [0, 1] \) be the received loss when taking action \( y_t \in \mathcal{Y} \) and \( x_t \in \mathcal{X} \) is the \( t^{th} \) symbol of the sequence.

For instance, decision \( \mathcal{Y} = \{ \text{umbrella, sunglasses} \} \) based on weather forecasts \( \mathcal{X} = \{ \text{sunny, rainy} \} \).

<table>
<thead>
<tr>
<th></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>

The goal is to minimize the \( \mu \)-expected loss. More generally we define the \( \Lambda_{\sigma} \) prediction scheme, which minimizes the \( \sigma \)-expected loss:

\[
y_t^{\Lambda_{\sigma}} := \arg \min_{y_t \in \mathcal{Y}} \sum_{x_t} \sigma(x_t | x_{<t}) \text{Loss}(x_t, y_t)
\]
Loss Bounds

- **Definition:** \( \mu \)-expected loss when \( \Lambda_\sigma \) predicts the \( t^{th} \) symbol:
  \[
  \text{Loss}_t(\Lambda_\sigma)(x_{<t}) := \sum_{x_t} \mu(x_t|x_{<t}) \text{Loss}(x_t, y_t^{\Lambda_\sigma})
  \]

- \( \text{Loss}_t(\Lambda_\mu/\xi) \) made by the informed/universal scheme \( \Lambda_\mu/\xi \):
  \[ \text{Loss}_t(\Lambda_\mu) \leq \text{Loss}_t(\Lambda) \quad \forall t, \Lambda. \]

- **Theorem:**
  \[
  0 \leq \text{Loss}_t(\Lambda_\xi) - \text{Loss}_t(\Lambda_\mu) \leq \sum_{x_t} |\xi_t - \mu_t| \leq \sqrt{2d_t} \xrightarrow{w.p.1} 0
  \]

- **Total Loss**
  \[
  \text{Loss}_{1:n}(\Lambda_\sigma) := \sum_{t=1}^{n} \mathbb{E}[\text{Loss}_t(\Lambda_\sigma)].
  \]

- **Theorem:**
  \[
  \text{Loss}_{1:n}(\Lambda_\xi) - \text{Loss}_{1:n}(\Lambda_\mu) \leq 2D_n + 2\sqrt{\text{Loss}_{1:n}(\Lambda_\mu)D_n}
  \]

- **Corollary:**
  If \( \text{Loss}_{1:\infty}(\Lambda_\mu) \) is finite, then \( \text{Loss}_{1:\infty}(\Lambda_\xi) \) is finite, and
  \[
  \frac{\text{Loss}_{1:n}(\Lambda_\xi)}{\text{Loss}_{1:\infty}(\Lambda_\mu)} \to 1 \quad \text{if} \quad \text{Loss}_{1:\infty}(\Lambda_\mu) \to \infty.
  \]

- **Remark:**
  Holds for any loss function \( \in [0, 1] \) with no assumptions (like i.i.d., Markovian, stationary, ergodic, ...) on \( \mu \in \mathcal{M} \).
Proof of Instantaneous Loss Bounds

Abbreviations: \( X = \{1, \ldots, N\}, \quad N = |X|, \quad i = x_t, \quad y_i = \mu(x_t|x_{<t}), \quad z_i = \xi(x_t|x_{<t}), \quad m = y_t^\Lambda\mu, \quad s = y_t^\Lambda\xi, \quad \ell_{xy} = \text{Loss}(x, y). \)

This and definition of \( y_t^\Lambda\mu \) and \( y_t^\Lambda\xi \) and \( \sum_i z_i \ell_{is} \leq \sum_i z_i \ell_{ij} \forall j \) implies

\[
\text{Loss}_t(\Lambda_\xi) - \text{Loss}_t(\Lambda_\mu) = \sum_i y_i \ell_{is} - \sum_i y_i \ell_{im} \leq \sum_i (y_i - z_i)(\ell_{is} - \ell_{im}) \\
\leq \sum_i |y_i - z_i| \cdot |\ell_{is} - \ell_{im}| \leq \sum_i |y_i - z_i| \leq \sqrt{\sum_i y_i \ln \frac{y_i}{z_i}} \equiv \sqrt{2d_t(x_{<t})}
\]

(a) We added \( \sum_i z_i(\ell_{im} - \ell_{is}) \geq 0. \)

(b) \( |\ell_{is} - \ell_{im}| \leq 1 \) since \( \ell \in [0, 1]. \)

(c) Pinsker’s inequality (elementary, but not trivial)
**Generalization: Continuous Classes $\mathcal{M}$**

In statistical parameter estimation one often has a continuous hypothesis class (e.g. a Bernoulli$(\theta)$ process with unknown $\theta \in [0, 1]$).

$$\mathcal{M} := \{\nu_\theta : \theta \in \mathbb{R}^d\}, \quad \xi(x) := \int_{\mathbb{R}^d} d\theta \, w(\theta) \, \nu_\theta(x), \quad \int_{\mathbb{R}^d} d\theta \, w(\theta) = 1$$

Under weak regularity conditions [CB90,H'03]:

Theorem: $D_n(\mu||\xi) \leq \ln w(\mu)^{-1} + \frac{d}{2} \ln \frac{n}{2\pi} + O(1)$

where $O(1)$ depends on the local curvature (parametric complexity) of $\ln \nu_\theta$, and is independent $n$ for many reasonable classes, including all stationary ($k^{th}$-order) finite-state Markov processes ($k = 0$ is i.i.d.).

$D_n \propto \log(n) = o(n)$ still implies excellent prediction and decision for most $n.$ [RH'07]
Bayesian Sequence Prediction: Summary

- The aim of probability theory is to describe uncertainty.
- Various sources and interpretations of uncertainty: frequency, objective, and subjective probabilities.
- They all respect the same rules.
- General sequence prediction: Use known (subj.) Bayes mixture \( \xi = \sum_{\nu \in \mathcal{M}} w_{\nu} \nu \) in place of unknown (obj.) true distribution \( \mu \).
- Bound on the relative entropy between \( \xi \) and \( \mu \).
  \[ \Rightarrow \text{posterior of } \xi \text{ converges rapidly to the true posterior } \mu. \]
- \( \xi \) is also optimal in a decision-theoretic sense w.r.t. any bounded loss function.
- No structural assumptions on \( \mathcal{M} \) and \( \nu \in \mathcal{M} \).
Universal Inductive Inferences: Contents

- Foundations of Universal Induction
- Bayesian Sequence Prediction and Confirmation
- Convergence and Decisions
- How to Choose the Prior – Universal
- Kolmogorov Complexity
- How to Choose the Model Class – Universal
- The Problem of Zero Prior
- Reparametrization and Regrouping Invariance
- The Problem of Old Evidence / New Theories
- Universal is Better than Continuous Class
- More Bounds / Stuff / Critique / Problems
- Summary / Outlook / Literature
Universal Inductive Inferences: Abstract

Solomonoff completed the Bayesian framework by providing a rigorous, unique, formal, and universal choice for the model class and the prior. I will discuss in breadth how and in which sense universal (non-i.i.d.) sequence prediction solves various (philosophical) problems of traditional Bayesian sequence prediction. I show that Solomonoff’s model possesses many desirable properties: Strong total and weak instantaneous bounds, and in contrast to most classical continuous prior densities has no zero posterior problem, i.e. can confirm universal hypotheses, is reparametrization and regrouping invariant, and avoids the old-evidence and updating problem. It even performs well (actually better) in non-computable environments.
**Induction Examples**

**Sequence prediction:** Predict weather/stock-quote/... tomorrow, based on past sequence. Continue IQ test sequence like 1,4,9,16,?

**Classification:** Predict whether email is spam. Classification can be reduced to sequence prediction.

**Hypothesis testing/identification:** Does treatment X cure cancer? Do observations of white swans confirm that all ravens are black?

These are instances of the important problem of inductive inference or time-series forecasting or sequence prediction.

**Problem:** Finding prediction rules for every particular (new) problem is possible but cumbersome and prone to disagreement or contradiction.

**Goal:** A single, formal, general, complete theory for prediction.

**Beyond induction:** active/reward learning, fct. optimization, game theory.
Foundations of Universal Induction

Ockham’s razor (simplicity) principle
Entities should not be multiplied beyond necessity.

Epicurus’ principle of multiple explanations
If more than one theory is consistent with the observations, keep all theories.

Bayes’ rule for conditional probabilities
Given the prior belief/probability one can predict all future probabilities.

Turing’s universal machine
Everything computable by a human using a fixed procedure can also be computed by a (universal) Turing machine.

Kolmogorov’s complexity
The complexity or information content of an object is the length of its shortest description on a universal Turing machine.

Solomonoff’s universal prior = Ockham + Epicurus + Bayes + Turing
Solves the question of how to choose the prior if nothing is known. ⇒ universal induction, formal Occam, AIT, MML, MDL, SRM, …
Bayesian Sequence Prediction and Confirmation

• **Assumption:** Sequence $\omega \in \mathcal{X}^\infty$ is sampled from the “true” probability measure $\mu$, i.e. $\mu(x) := P[x|\mu]$ is the $\mu$-probability that $\omega$ starts with $x \in \mathcal{X}^n$.

• **Model class:** We assume that $\mu$ is unknown but known to belong to a countable class of environments=models=measures $\mathcal{M} = \{\nu_1, \nu_2, \ldots\}$. [no i.i.d./ergodic/stationary assumption]

• **Hypothesis class:** $\{H\nu : \nu \in \mathcal{M}\}$ forms a mutually exclusive and complete class of hypotheses.

• **Prior:** $w_\nu := P[H\nu]$ is our prior belief in $H\nu$

$\Rightarrow$ **Evidence:** $\xi(x) := P[x] = \sum_{\nu \in \mathcal{M}} P[x|H\nu]P[H\nu] = \sum_{\nu} w_\nu \nu(x)$ must be our (prior) belief in $x$.

$\Rightarrow$ **Posterior:** $w_\nu(x) := P[H\nu|x] = \frac{P[x|H\nu]P[H\nu]}{P[x]}$ is our posterior belief in $\nu$ (Bayes’ rule).
When is a Sequence Random?

a) Is 0110010100101101001111011 generated by a fair coin flip?
b) Is 1111111111111111111111 generated by a fair coin flip?
c) Is 110010010001111110110101010 generated by a fair coin flip?
d) Is 0101010101010101010101010101 generated by a fair coin flip?

- Intuitively: (a) and (c) look random, but (b) and (d) look unlikely.
- Problem: Formally (a-d) have equal probability $\left(\frac{1}{2}\right)^{\text{length}}$.
- Classical solution: Consider hypothesis class $H := \{\text{Bernoulli}(p) : p \in \Theta \subseteq [0, 1]\}$ and determine $p$ for which sequence has maximum likelihood $\implies$ (a,c,d) are fair Bernoulli($\frac{1}{2}$) coins, (b) not.
- Problem: (d) is non-random, also (c) is binary expansion of $\pi$.
- Solution: Choose $H$ larger, but how large? Overfitting? MDL?
- AIT Solution: A sequence is random iff it is incompressible.
What does Probability Mean?

Naive frequency interpretation is circular:

- Probability of event $E$ is $p := \lim_{n \to \infty} \frac{k_n(E)}{n}$,
  $n = \# \text{i.i.d. trials}, k_n(E) = \# \text{occurrences of event } E \text{ in } n \text{ trials}$.

- Problem: Limit may be anything (or nothing):
  e.g. a fair coin can give: Head, Head, Head, Head, Head, ... $\Rightarrow p = 1$.

- Of course, for a fair coin this sequence is “unlikely”.
  For fair coin, $p = 1/2$ with “high probability”.

- But to make this statement rigorous we need to formally know what “high probability” means. Circularity!

Also: In complex domains typical for AI, sample size is often 1.
(e.g. a single non-iid historic weather data sequences is given).
We want to know whether certain properties hold for this particular seq.
How to Choose the Prior?

The probability axioms allow relating probabilities and plausibilities of different events, but they do not uniquely fix a numerical value for each event, except for the sure event $\Omega$ and the empty event $\emptyset$.

We need new principles for determining values for at least some basis events from which others can then be computed.

There seem to be only 3 general principles:

- The principle of indifference — the symmetry principle
- The maximum entropy principle
- Occam’s razor — the simplicity principle

Concrete: How shall we choose the hypothesis space $\{H_i\}$ and their prior $p(H_i)$ — or — $\mathcal{M} = \{\nu\}$ and their weight $w_\nu$. 
Indifference or Symmetry Principle

Assign same probability to all hypotheses:
\[ p(H_i) = \frac{1}{|I|} \] for finite \( I \)
\[ p(H_\theta) = [\text{Vol}(\Theta)]^{-1} \] for compact and measurable \( \Theta \).
\[ \Rightarrow p(H_i|D) \propto p(D|H_i) \overset{\wedge}{=} \text{classical Hypothesis testing (Max.Likelihood)}.
\]

Prev. Example: \( H_\theta = \text{Bernoulli}(\theta) \) with \( p(\theta) = 1 \) for \( \theta \in \Theta := [0, 1] \).

Problems: Does not work for “large” hypothesis spaces:
(a) Uniform distr. on infinite \( I = \mathbb{N} \) or noncompact \( \Theta \) not possible!
(b) Reparametrization: \( \theta \sim f(\theta) \). Uniform in \( \theta \) is not uniform in \( f(\theta) \).

Example: “Uniform” distr. on space of all (binary) sequences \( \{0, 1\}^\infty \):
\[ p(x_1...x_n) = \left(\frac{1}{2}\right)^n \forall n \forall x_1...x_n \Rightarrow p(x_{n+1} = 1|x_1...x_n) = \frac{1}{2} \] always!
Inference so not possible (No-Free-Lunch myth).

Predictive setting: All we need is \( p(x) \).
Occam’s Razor — The Simplicity Principle

- Only Occam’s razor (in combination with Epicurus’ principle) is general enough to assign prior probabilities in every situation.

- The idea is to assign high (subjective) probability to simple events, and low probability to complex events.

- Simple events (strings) are more plausible a priori than complex ones.

- This gives (approximately) justice to both Occam’s razor and Epicurus’ principle.
Prefix Sets/Codes

String $x$ is (proper) prefix of $y$ :⇔ $\exists z (\neq \epsilon)$ such that $xz = y$.

Set $\mathcal{P}$ is prefix-free or a prefix code :⇔ no element is a proper prefix of another.

Example: A self-delimiting code is prefix-free.

Kraft Inequality

For a prefix code $\mathcal{P}$ we have $\sum_{x \in \mathcal{P}} 2^{-\ell(x)} \leq 1$.

Conversely, let $l_1, l_2, \ldots$ be a countable sequence of natural numbers such that Kraft's inequality $\sum_k 2^{-l_k} \leq 1$ is satisfied. Then there exists a prefix code $\mathcal{P}$ with these lengths of its binary code.
Proof of the Kraft-Inequality

Proof $\Rightarrow$: Assign to each $x \in \mathcal{P}$ the interval $\Gamma_x := [0.x, 0.x + 2^{-\ell(x)})$. Length of interval $\Gamma_x$ is $2^{-\ell(x)}$. Intervals are disjoint, since $\mathcal{P}$ is prefix free, hence

$$\sum_{x \in \mathcal{P}} 2^{-\ell(x)} = \sum_{x \in \mathcal{P}} \text{Length}(\Gamma_x) \leq \text{Length}([0, 1]) = 1$$

$\Leftarrow$: Idea: Choose $l_1, l_2, \ldots$ in increasing order. Successively chop off intervals of lengths $2^{-l_1}, 2^{-l_2}, \ldots$ from left to right from $[0, 1)$ and define left interval boundary as code.
Priors from Prefix Codes

- Let $\text{Code}(H_\nu)$ be a prefix code of hypothesis $H_\nu$.

- Define complexity $Kw(\nu) := \text{Length}(\text{Code}(H_\nu))$.

- Choose prior $w_\nu = p(H_\nu) = 2^{-Kw(\nu)}$.
  \[ \Rightarrow \sum_{\nu \in \mathcal{M}} w_\nu \leq 1 \text{ is semi-probability (by Kraft).} \]

- How to choose a Code and hypothesis space $\mathcal{M}$?

- Praxis: Choose a code which is reasonable for your problem and $\mathcal{M}$ large enough to contain the true model.

- Theory: Choose a universal code and consider “all” hypotheses ...
Kolmogorov Complexity $K(x)$

K. of string $x$ is the length of the shortest (prefix) program producing $x$:

$$K(x) := \min_p \{l(p) : U(p) = x\}, \quad U = \text{universal TM}$$

For non-string objects $o$ (like numbers and functions) we define

$$K(o) := K(\langle o \rangle), \text{ where } \langle o \rangle \in \mathcal{X}^* \text{ is some standard code for } o.$$

+ Simple strings like 000...0 have small $K$,
  irregular (e.g. random) strings have large $K$.

• The definition is nearly independent of the choice of $U$.

+ $K$ satisfies most properties an information measure should satisfy.

+ $K$ shares many properties with Shannon entropy but is superior.

− $K(x)$ is not computable, but only semi-computable from above.

Fazit: \( K \) is an excellent universal complexity measure, suitable for quantifying Occam’s razor.
Schematic Graph of Kolmogorov Complexity

Although $K(x)$ is incomputable, we can draw a schematic graph

$$\log(x) + 2\log(\log(x))$$

$K(x)$

$\log(x)$

$x$
The Universal Prior

- Quantify the complexity of an environment $\nu$ or hypothesis $H_\nu$ by its Kolmogorov complexity $K(\nu)$.

- **Universal prior:** $w_\nu = \underbrace{w^U_\nu := 2^{-K(\nu)}}$ is a decreasing function in the model’s complexity, and sums to (less than) one.

  $$D_n \leq K(\mu) \ln 2,$$
  i.e. the number of $\varepsilon$-deviations of $\xi$ from $\mu$ or $l^{A\xi}$ from $l^{A\mu}$ is proportional to the complexity of the environment.

- No other semi-computable prior leads to better prediction (bounds).

- For **continuous** $M$, we can assign a (proper) universal prior (not density) $w^U_\theta = 2^{-K(\theta)} > 0$ for computable $\theta$, and $0$ for uncomp. $\theta$.

- This effectively reduces $M$ to a discrete class $\left\{ \nu_\theta \in M : w^U_\theta > 0 \right\}$ which is typically dense in $M$.

- This prior has many advantages over the classical prior (densities).
The Problem of Zero Prior
= the problem of confirmation of universal hypotheses

Problem: If the prior is zero, then the posterior is necessarily also zero.

Example: Consider the hypothesis $H = H_1$ that all balls in some urn or all ravens are black (=1) or that the sun rises every day.

Starting with a prior density as $w(\theta) = 1$ implies that prior $P[H_\theta] = 0$ for all $\theta$, hence posterior $P[H_\theta|1..1] = 0$, hence $H$ never gets confirmed.

3 non-solutions: define $H = \{\omega = 1^\infty\}$ | use finite population | abandon strict/logical/all-quantified/universal hypotheses in favor of soft hyp.

Solution: Assign non-zero prior to $\theta = 1 \Rightarrow P[H|1^n] \rightarrow 1$.

Generalization: Assign non-zero prior to all “special” $\theta$, like $\frac{1}{2}$ and $\frac{1}{6}$, which may naturally appear in a hypothesis, like “is the coin or die fair”.

Universal solution: Assign non-zero prior to all comp. $\theta$, e.g. $w^U_\theta = 2^{-K(\theta)}$
Reparametrization Invariance

• New parametrization e.g. \(\psi = \sqrt{\theta}\), then the \(\psi\)-density \(w'(\psi) = 2\sqrt{\theta} w(\theta)\) is no longer uniform if \(w(\theta) = 1\) is uniform \(\Rightarrow\) indifference principle is not reparametrization invariant (RIP).

• Jeffrey’s and Bernardo’s principle satisfy RIP w.r.t. differentiable bijective transformations \(\psi = f^{-1}(\theta)\).

• The universal prior \(w_\theta^U = 2^{-K(\theta)}\) also satisfies RIP w.r.t. simple computable \(f\). (within a multiplicative constant)
Regrouping Invariance

- **Non-bijective transformations:**
  E.g. grouping ball colors into categories black/non-black.

- **No classical principle is regrouping invariant.**

- Regrouping invariance is regarded as a very important and desirable property. [Walley’s (1996) solution: sets of priors]

- The universal prior $w^U_\theta = 2^{-K(\theta)}$ is invariant under regrouping, and more generally under all simple [computable with complexity $O(1)$] even non-bijective transformations. (within a multiplicative constant)

- **Note:** Reparametrization and regrouping invariance hold for arbitrary classes and are not limited to the i.i.d. case.
Universal Choice of Class $\mathcal{M}$

- The larger $\mathcal{M}$ the less restrictive is the assumption $\mu \in \mathcal{M}$.

- The class $\mathcal{M}_U$ of all (semi)computable (semi)measures, although only countable, is pretty large, since it includes all valid physics theories. Further, $\xi_U$ is semi-computable [ZL70].

- Solomonoff’s universal prior $M(x) := \text{probability that the output of a universal TM } U \text{ with random input starts with } x$.

- Formally: $M(x) := \sum_{p: U(p) = x} 2^{-\ell(p)}$ where the sum is over all (minimal) programs $p$ for which $U$ outputs a string starting with $x$.

- $M$ may be regarded as a $2^{-\ell(p)}$-weighted mixture over all deterministic environments $\nu_p$. ($\nu_p(x) = 1$ if $U(p) = x^*$ and 0 else)

- $M(x)$ coincides with $\xi_U(x)$ within an irrelevant multiplicative constant.
The Problem of Old Evidence / New Theories

• What if some evidence $E \equiv x$ (e.g. Mercury's perihelion advance) is known well-before the correct hypothesis/theory/model $H \equiv \mu$ (Einstein's general relativity theory) is found?

• How shall $H$ be added to the Bayesian machinery a posteriori?

• What should the “prior” of $H$ be?

• Should it be the belief in $H$ in a hypothetical counterfactual world in which $E$ is not known?

• Can old evidence $E$ confirm $H$?

• After all, $H$ could simply be constructed/biased/fitted towards “explaining” $E$. 
Solution of the Old-Evidence Problem

• The universal class $\mathcal{M}_U$ and universal prior $w^U$ formally solves this problem.

• The universal prior of $H$ is $2^{-K(H)}$ independent of $\mathcal{M}$ and of whether $E$ is known or not.

• Updating $\mathcal{M}$ is unproblematic, and even not necessary when starting with $\mathcal{M}_U$, since it includes all hypothesis (including yet unknown or unnamed ones) a priori.
Universal is Better than Continuous $\mathcal{M}$

- Although $\nu_\theta()$ and $w_\theta$ are incomp. for cont. classes $\mathcal{M}$ for most $\theta$, $\xi()$ is typically computable. (exactly as for Laplace or numerically)

$$\Rightarrow D_n(\mu||\mathcal{M}) \overset{+}{<} D_n(\mu||\xi) + K(\xi) \ln 2 \text{ for all } \mu$$

- That is, $\mathcal{M}$ is superior to all computable mixture predictors $\xi$ based on any (continuous or discrete) model class $\mathcal{M}$ and weight $w(\theta)$, save an additive constant $K(\xi) \ln 2 = O(1)$, even if environment $\mu$ is not computable.

- While $D_n(\mu||\xi) \sim \frac{d}{2} \ln n$ for all $\mu \in \mathcal{M}$, $D_n(\mu||\mathcal{M}) \leq K(\mu) \ln 2$ is even finite for computable $\mu$.

Fazit: Solomonoff prediction works also in non-computable environments
Convergence and Loss Bounds

- **Total (loss) bounds:** $\sum_{n=1}^{\infty} \mathbb{E}[h_n] \leq K(\mu) \ln 2$, where $h_t(\omega_{<t}) := \sum_{a \in \mathcal{X}} (\sqrt{\xi(a|\omega_{<t})} - \sqrt{\mu(a|\omega_{<t})})^2$.

- **Instantaneous i.i.d. bounds:** For i.i.d. $\mathcal{M}$ with continuous, discrete, and universal prior, respectively:
  
  $\mathbb{E}[h_n] \leq \frac{1}{n} \ln w(\mu)^{-1}$ and $\mathbb{E}[h_n] \leq \frac{1}{n} \ln w_{\mu}^{-1} = \frac{1}{n} K(\mu) \ln 2$.

- **Bounds for computable environments:** Rapidly $M(x_t|x_{<t}) \to 1$ on every computable sequence $x_{1:}\infty$ (whichsoever, e.g. $1^\infty$ or the digits of $\pi$ or $e$), i.e. $M$ quickly recognizes the structure of the sequence.

- **Weak instantaneous bounds:** valid for all $n$ and $x_{1:n}$ and $\bar{x}_n \neq x_n$:
  
  $2^{-K(n)} \leq M(\bar{x}_n|x_{<n}) \leq 2^{2K(x_{1:n}*)-K(n)}$.

- **Magic instance numbers:** e.g. $M(0|1^n) \leq 2^{-K(n)} \to 0$, but spikes up for simple $n$. $M$ is cautious at magic instance numbers $n$.

- **Future bounds / errors to come:** If our past observations $\omega_{1:n}$ contain a lot of information about $\mu$, we make few errors in future:
  
  $\sum_{t=n+1}^{\infty} \mathbb{E}[h_t|\omega_{1:n}] \leq [K(\mu|\omega_{1:n}) + K(n)] \ln 2$.
More Stuff / Critique / Problems

- Prior knowledge $y$ can be incorporated by using “subjective” prior $w_{\nu|y}^U = 2^{-K(\nu|y)}$ or by prefixing observation $x$ by $y$.

- Additive/multiplicative constant fudges and $U$-dependence is often (but not always) harmless.

- Incomputability: $K$ and $M$ can serve as “gold standards” which practitioners should aim at, but have to be (crudely) approximated in practice (MDL [Ris89], MML [Wal05], LZW [LZ76], CTW [WSTT95], NCD [CV05]).
Universal Inductive Inference: Summary

Universal Solomonoff prediction solves/avoids/meliorates many problems of (Bayesian) induction. We discussed:

+ general total bounds for generic class, prior, and loss,
+ i.i.d./universal-specific instantaneous and future bounds,
+ the $D_n$ bound for continuous classes,
+ indifference/symmetry principles,
+ the problem of zero prior & confirm. of universal hypotheses,
+ reparametrization and regrouping invariance,
+ the problem of old evidence and updating,
+ that $M$ works even in non-computable environments,
+ how to incorporate prior knowledge,
  - the prediction of short sequences,
  - the constant fudges in all results and the $U$-dependence,
  - $M$’s incomputability and crude practical approximations.
Outlook

• Relation to Prediction with Expert Advice
• Relation to the Minimal Description Length (MDL) Principle
• Generalization to Active/Reinforcement learning (AIXI)

Open Problems

• Prediction of selected bits
• Convergence of $M$ on Martin-Loef random sequences
• Better instantaneous bounds for $M$
• Future bounds for Bayes (general $\xi$)
Minimum Description Length: Contents

- The Minimum Description Length Principle
- Application: Sequence Prediction
- Application: Regression / Polynomial Fitting
- Summary
Minimum Description Length: Abstract

The Minimum Description/Message Length principle is one of the most important concepts in Machine Learning, and serves as a scientific guide, in general. The motivation is as follows: To make predictions involves finding regularities in past data, regularities in data allows for compression, hence short descriptions of data should help in making predictions. In this lecture we approach MDL from a Bayesian perspective and relate it to a MAP (maximum a posteriori) model choice. The Bayesian prior is chosen in accordance with Occam and Epicurus and the posterior is approximated by the MAP solution. We reconsider (un)fair coin flips and compare the M(D)L to Bayes-Laplace’s solution, and similarly for general sequence prediction tasks. Finally I present an application to regression / polynomial fitting.
MDL as Approximation of Solomonoff’s $M$

- Approximation of Solomonoff, since $M$ incomputable:
  
  - $M(x) \approx 2^{-K(x)}$ (excellent approximation)

- $K(x) \equiv K_U(x) \approx K_T(x)$
  (approximation quality depends on $T$ and $x$)

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

- MDL: Predict $y$ of smallest complexity $K_T(xy)$.

- Examples for $x$: Daily weather or stock market data.

- Example for $T$: Lempel-Ziv decompressor.
The Minimum Description Length Principle

Identification of probabilistic model “best” describing data:

Probabilistic model (≡ hypothesis) $H_\nu$ with $\nu \in \mathcal{M}$ and data $D$.

Most probable model is $\nu^{\text{MDL}} = \arg\max_{\nu \in \mathcal{M}} p(H_\nu | D)$.

Bayes’ rule: $p(H_\nu | D) = p(D | H_\nu) \cdot p(H_\nu) / p(D)$.

Occam’s razor: $p(H_\nu) = 2^{-Kw(\nu)}$.

By definition: $p(D | H_\nu) = \nu(x), \quad D = x = \text{data-seq.}, \quad p(D) = \text{const.}$

Take logarithm $\implies \nu^{\text{MDL}} = \arg\min_{\nu \in \mathcal{M}} \{K\nu(x) + Kw(\nu)\}$

$K\nu(x) := -\log_2 \nu(x) = \text{length of Shannon-Fano code of } x \text{ given } H_\nu$.

$Kw(\nu) = \text{length of model } H_\nu$.

Names: Two-part MDL or MAP or MML (∃ “slight” differences)
Predict with Best Model

- Use **best model** from class of models $\mathcal{M}$ for prediction:

- Predict $y$ with probability $\nu^{\text{MDL}}(y|x) = \frac{\nu^{\text{MDL}}(xy)}{\nu^{\text{MDL}}(x)}$ (3 variants)

- $y^{\text{MDL}} = \arg\max_y \nu^{\text{MDL}}(y|x)$ is most likely continuation of $x$

- **Special case:** $Kw(\nu) = \text{const.}$
  $\implies \text{MDL} \sim \text{ML} := \text{Maximum likelihood principle.}$

- **Example:** $H_\theta = \text{Bernoulli}(\theta)$ with $\theta \in [0, 1]$ and $Kw(\theta) := \text{const.}$ and $\nu(x_1:n) = \theta^{n_1}(1 - \theta)^{n_0}$ with $n_1 = x_1 + \ldots + x_n = n - n_0$.
  $\Rightarrow \theta^{\text{MDL}} = \arg\min_\theta \{-\log_2 \theta^{n_1}(1 - \theta)^{n_0} + Kw(\theta)\} = \text{ML}$
  (overconfident, e.g. $n_1 = 0$)

- Compare with Laplace’ rule based on Bayes’ rule: $\theta^{\text{Laplace}} = \frac{n_1 + 1}{n + 2}$. 
**Application: Sequence Prediction**

- Instead of Bayes mixture $\xi(x) = \sum_{\nu} w_{\nu} \nu(x)$, consider MAP/MDL

- $\nu^{\text{MDL}}(x) = \max \{ w_{\nu} \nu(x) : \nu \in \mathcal{M} \} = \arg \min_{\nu \in \mathcal{M}} \{ K \nu(x) + K w(\nu) \}$.

- \[
\sum_{t=1}^{\infty} \mathbb{E} \left[ \sum_{x_t} \left( \mu(x_t|x_{<t}) - \nu^{\text{MDL}}(x_t|x_{<t}) \right)^2 \right] \leq 8w_{\mu}^{-1} \quad \Leftarrow \quad \text{no log as for } \xi!
\]

$\Rightarrow$ MDL converges, but speed can be exponentially worse than Bayes

$\Rightarrow$ be careful (bound is tight).

- For continuous smooth model class $\mathcal{M}$ and prior $w_\nu$, MDL is as good as Bayes.
Application: Regression / Polynomial Fitting

- **Data** $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$

- **Fit polynomial** $f_d(x) := a_0 + a_1 x + a_2 x^2 + \ldots + a_d x^d$ of degree $d$ through points $D$

- **Measure of error:** $SQ(a_0 \ldots a_d) = \sum_{i=1}^{n} (y_i - f_d(x_i))^2$

- **Given** $d$, minimize $SQ(a_0 : d)$ w.r.t. parameters $a_0 \ldots a_d$.

- **This classical approach does not tell us how to choose** $d$? ($d \geq n - 1$ gives perfect fit)
MDL Solution to Polynomial Fitting

Assume $y$ is Gaussian with variance $\sigma^2$ and mean $f_d(x)$, i.e.

$$P((x, y) | f_d) := P(y | x, f_d) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(y - f_d(x))^2}{2\sigma^2}\right)$$

$$\implies P(D | f_d) = \prod_{i=1}^{d} P((x_i, y_i) | f_d) = \frac{e^{-SQ(a_{0:d})/2\sigma^2}}{(2\pi\sigma^2)^{n/2}}$$

The larger the error $SQ$, the less likely the data.

**Occam:** $P(f_d) = 2^{-Kw(f_d)}$. Simple coding: $Kw(f_d) \approx (d + 1) \cdot C$, where $C$ is the description length=accuracy of each coefficient $a_k$ in bits $\implies$

$$f^{MDL} = \arg\min_{f} \left\{ -\log_2 P(D | f) + Kw(f) \right\} = \arg\min_{d, a_{0:d}} \left\{ \frac{SQ(a_{0:d})}{2\sigma^2 \ln 2} + (d+1)C \right\}$$

Fixed $d$ $\implies a^{ML}_{0:d} = \arg\min_{a_{0:d}} SQ(a_{0:d}) = \text{classical solution}$

(by linear invariance of $\arg\min$)
MDL Polynomial Fitting: Determine Degree $d$

Determine $d$ ($\min_f = \min_d \min_{f,d}$):

$$d = \arg \min_d \left\{ \frac{1}{2\sigma^2 \ln 2} \, SQ(a^{ML}_{0:d}) + \frac{n}{2} \log_2(2\pi\sigma^2) + (d+1)C \right\}$$

Interpretation: Tradeoff between SQuare error and complexity penalty

Minimization w.r.t. $\sigma$ leads to $\sigma^2 = SQ(d) := SQ(a^{ML}_{0:d})$, hence

$$d = \arg \min_d \left\{ \frac{n}{2} \ln SQ(d) + (d+1)C \right\}.$$ 

With subtle arguments one can derive $C \doteq \frac{1}{2} \ln n$.

Numerically find minimum of r.h.s.
Minimum Description Length: Summary

- Probability axioms give no guidance of how to choose the prior.
- Occam's razor is the only general (always applicable) principle for determining priors, especially in complex domains typical for AI.
- \[ \text{Prior} = 2^{-\text{descr.length}} \quad \text{— Universal prior} = 2^{-\text{Kolmogorov complexity}}. \]
- Prediction \( \hat{=} \) finding regularities \( \hat{=} \) compression \( \hat{=} \) MDL.
- MDL principle: from a model class, a model is chosen that: minimizes the joint description length of the model and the data observed so far given the model.
- Similar to (Bayesian) Maximum a Posteriori (MAP) principle.
- MDL often as good as Bayes but not always.
The Universal Similarity Metric: Contents

- Kolmogorov Complexity
- The Universal Similarity Metric
- Tree-Based Clustering
- Genomics & Phylogeny: Mammals, SARS Virus & Others
- Classification of Different File Types
- Language Tree (Re)construction
- Classify Music w.r.t. Composer
- Further Applications
- Summary
The Universal Similarity Metric: Abstract

The MDL method has been studied from very concrete and highly tuned practical applications to general theoretical assertions. Sequence prediction is just one application of MDL. The MDL idea has also been used to define the so called information distance or universal similarity metric, measuring the similarity between two individual objects. I will present some very impressive recent clustering applications based on standard Lempel-Ziv or bzip2 compression, including a 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) others.

Based on [Cilibrasi&Vitanyi’05]
Conditional Kolmogorov Complexity

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

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

**Examples:**

1) $x$ is very similar to itself ($K(x|x) \pm 0$)

2) A processed $x$ is similar to $x$ ($K(f(x)|x) \pm 0$ if $K(f) = O(1)$).  
   e.g. doubling, reverting, inverting, encrypting, partially deleting $x$.

3) A random string is with high probability not similar to any other string ($K(random|y) = \text{length}(random)$).

The problem with $K(x|y)$ as similarity=distance measure is that it is neither symmetric nor normalized nor computable.
The Universal Similarity Metric

- Symmetrization and normalization leads to a/the universal metric \( d \):

\[
0 \leq d(x, y) := \frac{\max\{K(x|y), K(y|x)\}}{\max\{K(x), K(y)\}} \leq 1
\]

- Every effective similarity between \( x \) and \( y \) is detected by \( d \)

- Use \( K(x|y) \approx K(xy) - K(y) \) (coding \( T \)) and \( K(x) \equiv K_U(x) \approx K_T(x) \)

\[\implies\] computable approximation: Normalized compression distance:

\[
d(x, y) \approx \frac{K_T(xy) - \min\{K_T(x), K_T(y)\}}{\max\{K_T(x), K_T(y)\}} \lesssim 1
\]

- For \( T \) choose Lempel-Ziv or gzip or bzip(2) (de)compressor in the applications below.

- Theory: Lempel-Ziv compresses asymptotically better than any probabilistic finite state automaton predictor/compressor.
Tree-Based Clustering

- If many objects $x_1, \ldots, x_n$ need to be compared, determine the similarity matrix $M_{ij} = d(x_i, x_j)$ for $1 \leq i, j \leq n$

- Now cluster similar objects.

- There are various clustering techniques.

- Tree-based clustering: Create a tree connecting similar objects,

- e.g. quartet method (for clustering)
Let $x_1, \ldots, x_n$ be mitochondrial genome sequences of different mammals:

**Partial distance matrix $M_{ij}$ using bzip2(?)**

<table>
<thead>
<tr>
<th></th>
<th>Cat</th>
<th>Echidna</th>
<th>Gorilla</th>
<th>Human</th>
</tr>
</thead>
<tbody>
<tr>
<td>BrownBear</td>
<td>0.002</td>
<td>0.943</td>
<td>0.887</td>
<td>0.935</td>
</tr>
<tr>
<td>Carp</td>
<td>0.943</td>
<td>0.006</td>
<td>0.946</td>
<td>0.954</td>
</tr>
<tr>
<td>Cat</td>
<td>0.887</td>
<td>0.946</td>
<td>0.003</td>
<td>0.926</td>
</tr>
<tr>
<td>Chimpanzee</td>
<td>0.935</td>
<td>0.954</td>
<td>0.926</td>
<td>0.006</td>
</tr>
<tr>
<td>Cow</td>
<td>0.906</td>
<td>0.947</td>
<td>0.897</td>
<td>0.926</td>
</tr>
<tr>
<td>Echidna</td>
<td>0.944</td>
<td>0.955</td>
<td>0.942</td>
<td>0.948</td>
</tr>
<tr>
<td>FinbackWhale</td>
<td>0.915</td>
<td>0.952</td>
<td>0.905</td>
<td>0.926</td>
</tr>
<tr>
<td>Gibbon</td>
<td>0.939</td>
<td>0.951</td>
<td>0.928</td>
<td>0.849</td>
</tr>
<tr>
<td>Gorilla</td>
<td>0.940</td>
<td>0.957</td>
<td>0.931</td>
<td>0.731</td>
</tr>
<tr>
<td>HouseMouse</td>
<td>0.934</td>
<td>0.956</td>
<td>0.919</td>
<td>0.943</td>
</tr>
<tr>
<td>Human</td>
<td>0.930</td>
<td>0.946</td>
<td>0.922</td>
<td>0.667</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Genomics & Phylogeny: Mammals

Evolutionary tree built from complete mammalian mtDNA of 24 species:
Genomics & Phylogeny: SARS Virus and Others

- Clustering of SARS virus in relation to potential similar virii based on complete sequenced genome(s) using bzip2:

- The relations are very similar to the definitive tree based on medical-macrobio-genomics analysis from biologists.
Genomics & Phylogeny: SARS Virus and Others
Classification of Different File Types

Classification of files based on markedly different file types using bzip2

- Four mitochondrial gene sequences
- Four excerpts from the novel “The Zeppelin’s Passenger”
- Four MIDI files without further processing
- Two Linux x86 ELF executables (the `cp` and `rm` commands)
- Two compiled Java class files

No features of any specific domain of application are used!
Classification of Different File Types

Perfect classification!
Language Tree (Re)construction

• Let $x_1, ..., x_n$ be the “The Universal Declaration of Human Rights” in various languages $1, ..., n$.

• Distance matrix $M_{ij}$ based on gzip. Language tree constructed from $M_{ij}$ by the Fitch-Margoliash method [Li&al’03]

• All main linguistic groups can be recognized (next slide)
Classify Music w.r.t. Composer

Let $m_1, \ldots, m_n$ be pieces of music in MIDI format.

Preprocessing the MIDI files:

- Delete identifying information (composer, title, ...), instrument indicators, MIDI control signals, tempo variations, ...
- Keep only note-on and note-off information.
- A note, $k \in \mathbb{Z}$ half-tones above the average note is coded as a signed byte with value $k$.
- The whole piece is quantized in $0.05$ second intervals.
- Tracks are sorted according to decreasing average volume, and then output in succession.

Processed files $x_1, \ldots, x_n$ still sounded like the original.
Classify Music w.r.t. Composer

12 pieces of music: $4 \times \text{Bach} + 4 \times \text{Chopin} + 4 \times \text{Debussy}$. Class. by bzip2

Perfect grouping of processed MIDI files w.r.t. composers.
Further Applications

• Classification of Fungi
• Optical character recognition
• Classification of Galaxies
• Clustering of novels w.r.t. authors
• Larger data sets

See [Cilibrasi & Vitanyi’03]
The Clustering Method: Summary

- based on the universal similarity metric,
- based on Kolmogorov complexity,
- approximated by bzip2,
- with the similarity matrix represented by tree,
- approximated by the quartet method

- leads to excellent classification in many domains.
Universal Rational Agents: Contents

- Rational agents
- Sequential decision theory
- Reinforcement learning
- Value function
- Universal Bayes mixture and AIXI model
- Self-optimizing policies
- Pareto-optimality
- Environmental Classes
- The horizon problem
- Computational Issues
Universal Rational Agents: Abstract

Sequential decision theory formally solves the problem of rational agents in uncertain worlds if the true environmental prior probability distribution is known. Solomonoff’s theory of universal induction formally solves the problem of sequence prediction for unknown prior distribution.

Here we combine both ideas and develop an elegant parameter-free theory of an optimal reinforcement learning agent embedded in an arbitrary unknown environment that possesses essentially all aspects of rational intelligence. The theory reduces all conceptual AI problems to pure computational ones.

We give strong arguments that the resulting AIXI model is the most intelligent unbiased agent possible. Other discussed topics are relations between problem classes, the horizon problem, and computational issues.
Most if not all AI problems can be formulated within the agent framework.
Rational Agents in Deterministic Environments

- \( p : \mathcal{X}^* \rightarrow \mathcal{Y}^* \) is deterministic policy of the agent,
  \( p(x_{<k}) = y_{1:k} \) with \( x_{<k} \equiv x_1 \ldots x_{k-1} \).

- \( q : \mathcal{Y}^* \rightarrow \mathcal{X}^* \) is deterministic environment,
  \( q(y_{1:k}) = x_{1:k} \) with \( y_{1:k} \equiv y_1 \ldots y_k \).

- Input \( x_k \equiv r_k o_k \) consists of a regular informative part \( o_k \) and reward \( r_k \in [0..r_{\text{max}}] \).

- Value \( V_{km}^{pq} := r_k + \ldots + r_m \),
  optimal policy \( p^{\text{best}} := \arg \max_p V_{1m}^{pq} \),
  Lifespan or initial horizon \( m \).
Agents in Probabilistic Environments

Given history $y_{1:k} x_{<k}$, the probability that the environment leads to perception $x_k$ in cycle $k$ is (by definition) $\sigma(x_k | y_{1:k} x_{<k})$.

Abbreviation (chain rule)

$$\sigma(x_{1:m} | y_{1:m}) = \sigma(x_1 | y_1) \cdot \sigma(x_2 | y_{1:2} x_1) \cdot \ldots \cdot \sigma(x_m | y_{1:m} x_{<m})$$

The average value of policy $p$ with horizon $m$ in environment $\sigma$ is defined as

$$V_p^\sigma := \frac{1}{m} \sum_{x_{1:m}} (r_1 + \ldots + r_m) \sigma(x_{1:m} | y_{1:m}) | y_{1:m} = p(x_{<m})$$

The goal of the agent should be to maximize the value.
The $\sigma$-optimal policy $p^\sigma := \arg\max_p V^p_\sigma$ maximizes $V^p_\sigma \leq V^*_\sigma := V^{p^\sigma}_\sigma$.

Explicit expressions for the action $y_k$ in cycle $k$ of the $\sigma$-optimal policy $p^\sigma$ and their value $V^*_\sigma$ are

$$y_k = \arg\max_{y_k} \sum_{x_k} \max_{y_{k+1}} \sum_{x_{k+1}} \ldots \max_{y_m} \sum_{x_m} \left( r_k + \ldots + r_m \right) \sigma(x_k:m|y_1:m \mathcal{X} < k),$$

and

$$V^*_\sigma = \frac{1}{m} \max_{y_1} \sum_{x_1} \max_{y_2} \sum_{x_2} \ldots \max_{y_m} \sum_{x_m} \left( r_1 + \ldots + r_m \right) \sigma(x_1:m|y_1:m).$$

Keyword: Expectimax tree/algorithm.
**Expectimax Tree/Algorithm**

\[ V^*_\sigma(yx < k) = \max_{y_k} V^*_\sigma(yx < k y_k) \]

\[ V^*_\sigma(yx < k y_k) = \sum_{x_k} [r_k + V^*_\sigma(yx_1:k)] \sigma(x_k | yx < k y_k) \]

\( \sigma \) expected reward \( r_k \) and observation \( o_k \).

\[ V^*_\sigma(yx_1:k) = \max_{y_{k+1}} V^*_\sigma(yx_1:k y_{k+1}) \]
Known environment $\mu$

- Assumption: $\mu$ is the true environment in which the agent operates

- Then, policy $p^{\mu}$ is optimal in the sense that no other policy for an agent leads to higher $\mu^{AI}$-expected reward.

- Special choices of $\mu$: deterministic or adversarial environments, Markov decision processes (MDPs), adversarial environments.

- There is no principle problem in computing the optimal action $y_k$ as long as $\mu^{AI}$ is known and computable and $\mathcal{X}$, $\mathcal{Y}$ and $m$ are finite.

- Things drastically change if $\mu^{AI}$ is unknown ...
The Bayes-mixture distribution $\xi$

Assumption: The true environment $\mu$ is unknown.

Bayesian approach: The true probability distribution $\mu^{AI}$ is not learned directly, but is replaced by a Bayes-mixture $\xi^{AI}$.

Assumption: We know that the true environment $\mu$ is contained in some known (finite or countable) set $\mathcal{M}$ of environments.

The Bayes-mixture $\xi$ is defined as

$$\xi(x_{1:m}|y_{1:m}) := \sum_{\nu \in \mathcal{M}} w_{\nu} \nu(x_{1:m}|y_{1:m}) \quad \text{with} \quad \sum_{\nu \in \mathcal{M}} w_{\nu} = 1, \quad w_{\nu} > 0 \ \forall \nu$$

The weights $w_{\nu}$ may be interpreted as the prior degree of belief that the true environment is $\nu$.

Then $\xi(x_{1:m}|y_{1:m})$ could be interpreted as the prior subjective belief probability in observing $x_{1:m}$, given actions $y_{1:m}$. 
Questions of Interest

• It is natural to follow the policy \( p^\xi \) which maximizes \( V_\xi^p \).

• If \( \mu \) is the true environment the expected reward when following policy \( p^\xi \) will be \( V_\mu^{p^\xi} \).

• The optimal (but infeasible) policy \( p^\mu \) yields reward \( V_\mu^{p^\mu} \equiv V_\mu^* \).

• Are there policies with uniformly larger value than \( V_\mu^{p^\xi} \)?

• How close is \( V_\mu^{p^\xi} \) to \( V_\mu^* \)?

• What is the most general class \( \mathcal{M} \) and weights \( w_\nu \).
A universal choice of $\xi$ and $\mathcal{M}$

- We have to assume the existence of some structure on the environment to avoid the No-Free-Lunch Theorems [Wolpert 96].
- We can only unravel effective structures which are describable by (semi)computable probability distributions.
- So we may include all (semi)computable (semi)distributions in $\mathcal{M}$.
- Occam’s razor and Epicurus’ principle of multiple explanations tell us to assign high prior belief to simple environments.
- Using Kolmogorov’s universal complexity measure $K(\nu)$ for environments $\nu$ one should set $w_\nu \sim 2^{-K(\nu)}$, where $K(\nu)$ is the length of the shortest program on a universal TM computing $\nu$.
- The resulting AIXI model [Hutter:00] is a unification of (Bellman’s) sequential decision and Solomonoff’s universal induction theory.
The AIXI Model in one Line

The most intelligent unbiased learning agent

\[
y_k = \arg \max_y \sum_{x_k} \ldots \max_{y_m} \sum_{x_m} [r(x_k) + \ldots + r(x_m)] \sum_{q: U(q,y_1:m) = x_1:m} 2^{-\ell(q)}
\]

is an elegant mathematical theory of AI

Claim: AIXI is the most intelligent environmental independent, i.e. universally optimal, agent possible.

Proof: For formalizations, quantifications, and proofs, see [Hut05].

Applications: Strategic Games, Function Minimization, Supervised Learning from Examples, Sequence Prediction, Classification.

In the following we consider generic \( \mathcal{M} \) and \( w_\nu \).
Pareto-Optimality of $p^\xi$

Policy $p^\xi$ is Pareto-optimal in the sense that there is no other policy $p$ with $V_{\nu}^p \geq V_{\nu}^{p^\xi}$ for all $\nu \in \mathcal{M}$ and strict inequality for at least one $\nu$.

Self-optimizing Policies

Under which circumstances does the value of the universal policy $p^\xi$ converge to optimum?

$$V_{\nu}^{p^\xi} \to V_{\nu}^* \quad \text{for horizon} \quad m \to \infty \quad \text{for all} \quad \nu \in \mathcal{M}. \quad (1)$$

The least we must demand from $\mathcal{M}$ to have a chance that (1) is true is that there exists some policy $\tilde{p}$ at all with this property, i.e.

$$\exists \tilde{p} : V_{\nu}^{\tilde{p}} \to V_{\nu}^* \quad \text{for horizon} \quad m \to \infty \quad \text{for all} \quad \nu \in \mathcal{M}. \quad (2)$$

Main result: $(2) \Rightarrow (1)$: The necessary condition of the existence of a self-optimizing policy $\tilde{p}$ is also sufficient for $p^\xi$ to be self-optimizing.
Environments w. (Non)Self-Optimizing Policies

Non Self-Optimizing

- nth order MDP
- Strongly Ergodic nth order MDP
- Repeated Strategic Game
- Classification
- Bandit
- Markov Process
- Function Minimisation
- 1.i.d. Process

Self-Optimizing

- POMDP
- MDP
- Ergodic MDP
- Totally Passive Environment
- Sequence Prediction

- Chronological Environment
- Passive Environment
Particularly Interesting Environments

- **Sequence Prediction**, e.g. weather or stock-market prediction.  
  Strong result: \( V_\mu^* - V_\mu^p_{\xi} = O\left(\sqrt{\frac{K(\mu)}{m}}\right), \quad m = \text{horizon} \).

- **Strategic Games**: Learn to play well (minimax) strategic zero-sum games (like chess) or even exploit limited capabilities of opponent.

- **Optimization**: Find (approximate) minimum of function with as few function calls as possible. Difficult exploration versus exploitation problem.

- **Supervised learning**: Learn functions by presenting \((z, f(z))\) pairs and ask for function values of \(z'\) by presenting \((z', ?)\) pairs.  
  Supervised learning is much faster than reinforcement learning.

\(\xi\) quickly learns to **predict**, **play games**, **optimize**, and **learn supervised**.
Future Value and the Right Discounting

- Eliminate the arbitrary horizon parameter $m$ by discounting the rewards $r_k \sim \gamma_k r_k$ with $\Gamma_k := \sum_{i=k}^{\infty} \gamma_i < \infty$ and letting $m \to \infty$:

$$V_{k\gamma}^{\pi\sigma} := \frac{1}{\Gamma_k} \lim_{m \to \infty} \sum_{x_{k:m}} (\gamma_k r_k + \ldots + \gamma_m r_m) \sigma(x_{k:m}|y_{1:m}x < k)|y_{1:m} = p(x < m)$$

- If there exists a self-optimizing policy for $\mathcal{M}$, then $p^\xi$ is self-optimizing: If $\exists \pi_k \forall \nu : V_{k\gamma}^{\pi_k \nu} \xrightarrow{k \to \infty} V_{k\gamma}^* \nu \Rightarrow V_{k\gamma}^{p^\xi \mu} \xrightarrow{k \to \infty} V_{k\gamma}^* \mu$.

- Standard geometric discounting: $\gamma_k = \gamma^k$ with $0 < \gamma < 1$.
  Problem: Most environments do not possess self-optimizing policies under this discounting, since effective horizon $h_{k}^{\text{eff}}$ is finite.

- Power discounting: $\gamma_k = k^{-2} \Rightarrow h_{k}^{\text{eff}} \sim k = \text{agent’s age}$.
  Universal discounting: $\gamma_k = 2^{-K(k)} \Rightarrow h_{k}^{\text{eff}} \sim \text{Ackermann}(k)$

- Result: Policy $p^\xi$ is self-optimizing for ergodic MDPs if $\frac{\gamma_{k+1}}{\gamma_k} \to 1$. 
The Timebounded AI$\xi$ Model

An algorithm $p^{\text{best}}$ has been constructed for which the following holds:

- Let $p$ be any (extended chronological) policy
- with length $\ell(p) \leq \tilde{l}$ and computation time per cycle $t(p) \leq \tilde{t}$
- for which there exists a proof of length $\leq l_P$ that $p$ is a valid approximation.
- Then an algorithm $p^{\text{best}}$ can be constructed, depending on $\tilde{l}, \tilde{t}$ and $l_P$ but not on knowing $p$
- which is effectively more or equally intelligent according to $\succeq_c$ than any such $p$.
- The size of $p^{\text{best}}$ is $\ell(p^{\text{best}}) = O(\ln(\tilde{l} \cdot \tilde{t} \cdot l_P))$,
- the setup-time is $t_{\text{setup}}(p^{\text{best}}) = O(l_P^2 \cdot 2^{l_P})$,
- the computation time per cycle is $t_{\text{cycle}}(p^{\text{best}}) = O(2^{\tilde{l}} \cdot \tilde{t})$. 
Universal Rational Agents: Summary

- **Setup:** Agents acting in general probabilistic environments with reinforcement feedback.

- **Assumptions:** True environment $\mu$ belongs to a known class of environments $\mathcal{M}$, but is otherwise unknown.

- **Results:** The Bayes-optimal policy $p^\xi$ based on the Bayes-mixture $\xi = \sum_{\nu \in \mathcal{M}} w_\nu \nu$ is Pareto-optimal and self-optimizing if $\mathcal{M}$ admits self-optimizing policies.

- **Application:** The class of ergodic MDPs admits self-optimizing policies.

- **New:** Policy $p^\xi$ with unbounded effective horizon is the first purely Bayesian self-optimizing consistent policy for ergodic MDPs.

- **Learn:** The combined conditions $\Gamma_k < \infty$ and $\frac{\gamma_k+1}{\gamma_k} \to 1$ allow a consistent self-optimizing Bayes-optimal policy based on mixtures.
Universal Rational Agents: Remarks

- We have developed a parameterless AI model based on sequential decisions and algorithmic probability.
- We have reduced the AI problem to pure computational questions.
- AI $\xi$ seems not to lack any important known methodology of AI, apart from computational aspects.
- There is no need for implementing extra knowledge, as this can be learned by presenting it in $o_k$ in any form.
- The learning process itself is an important aspect of AI.
- Noise or irrelevant information in the inputs do not disturb the AI $\xi$ system.
- Philosophical questions: relevance of non-computational physics (Penrose), number of wisdom $\Omega$ (Chaitin), consciousness, social consequences.
Universal Rational Agents: Outlook

- Continuous classes $\mathcal{M}$.
- Restricted policy classes.
- Non-asymptotic bounds.
- Tighter bounds by exploiting extra properties of the environments, like the mixing rate of MDPs.
- Search for other performance criteria [Hutter:00].
- Instead of convergence of the expected reward sum, study convergence with high probability of the actually realized reward sum.
- Other environmental classes (separability concepts, downscaling).
Wrap Up

- **Setup:** Given (non)iid data $D = (x_1, ..., x_n)$, predict $x_{n+1}$
- **Ultimate goal** is to maximize profit or minimize loss
- **Consider** Models/Hypothesis $H_i \in \mathcal{M}$
- **Max.Likelihood:** $H_{best} = \arg \max_i p(D|H_i)$ (overfits if $\mathcal{M}$ large)
- **Bayes:** Posterior probability of $H_i$ is $p(H_i|D) \propto p(D|H_i)p(H_i)$
- **MDL:** $H_{best} = \arg \min_{H_i} \left\{ \text{CodeLength}(D|H_i) + \text{CodeLength}(H_i) \right\}$ (Complexity penalization)
- Bayes needs prior($H_i$), MDL needs CodeLength($H_i$)
- **Occam+Epicurus:** High prior for simple models with short codes.
- **Kolmogorov/Solomonoff:** Quantification of simplicity/complexity
- **MDL & Bayes** work if $D$ is sampled from $H_{true} \in \mathcal{M}$
- **Universal AI** = Universal Induction + Sequential Decision Theory
Literature

http://arXiv.org/abs/cs/0312044


http://www.hutter1.net/ai/uaibook.htm.

http://arxiv.org/abs/0709.1516

http://dx.doi.org/10.1007/s11023-007-9079-x
Thanks! Questions? Details:

**Jobs:** PostDoc and PhD positions at RSISE and NICTA, Australia

**Projects** at http://www.hutter1.net/

**A Unified View of Artificial Intelligence**

\[
\text{Decision Theory} = \text{Probability} + \text{Utility Theory} + \text{Universal Induction} = \text{Ockham} + \text{Bayes} + \text{Turing}
\]

**Open research problems** at www.hutter1.net/ai/uaibook.htm

**Compression competition** with 50’000 Euro prize at prize.hutter1.net