Optimality of Universal Bayesian Prediction for General Loss and Alphabet

Marcus Hutter

Istituto Dalle Molle di Studi sull’Intelligenza Artificiale (IDSIA), Galleria 2, CH-6928 Manno-Lugano, Switzerland
marcus@idsia.ch, http://www.idsia.ch/~marcus

2000 – 2002

Universal Induction = Ockham + Epicurus + Bayes

\[
\frac{\text{Loss(Universal Prediction Scheme)}}{\text{Loss(Any other Prediction Scheme)}} \leq 1 + o(1)
\]
Table of Contents

• The Philosophical Dilemma of Predicting the Future
• (Conditional) Probabilities and their Interpretation
• Probability that the Sun will rise Tomorrow
• Kolmogorov Complexity
• Universal Probability Distribution
• Universal Sequence Prediction
• Loss Bounds & Optimality
• Application to Games of Chance (Bound on Winning Time)
• Generalization: Continuous Probability Classes
• Generalization: The Universal AI Model
• Further Generalizations, Outlook, Conclusions
Problem Setup

• Every induction problem can be phrased as a sequence prediction.

• Classification is a special case of sequence prediction. (With some tricks the other direction is also true)

• I’m interested in maximizing profit (minimizing loss).
  I’m not (primarily) interested in finding a (true/predictive/causal) model.

• Separating noise from data is *not* necessary in this setting!

My Position to Occam

• Most of us belief in or at least use the axioms of logic, proof theory, set theory, natural numbers when doing science, without questioning them.

• We should/must add Occam’s razor in some quantified form, because it is the foundation of machine learning and science.

• There is (yet) no mathematical proof of Occam’s razor, and it is an independent axiom, but there is lots of evidence that this
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 class, ... 

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

• Machine learning: Tries to build and understand systems which learn from past data, to make good prediction, which are able to generalize. Many terms only vaguely defined or there are many alternatives.
Occam to the Rescue

- Is it possible to give machine learning a rigorous mathematical framework/definition?
- Yes! Use Occam’s razor, quantified in terms of Kolmogorov complexity, and combine it with Bayes, and possibly sequential decision theory.
- There is at the moment no alternative suggestion of how to define machine learning rigorously.

My view of (future) Machine learning

- Application = Solve learning tasks by approximating Kolmogorov complexity (MML, MDL, SRM, and much more specific ones, like SVM).
- Theory = Proof theorems, especially on convergence and approximations.
- Non-standard ML = Modify “Occam’s axiom” with the goal to find something better.
Induction = Predicting the Future

Extrapolate past observations to the future, but how can we know something about the future?

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

Ockhams' razor (simplicity) principle
Entities should not be multiplied beyond necessity.

Hume’s negation of Induction
The only form of induction possible is deduction as the conclusion is already logically contained in the premises.

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

Solomonoff’s universal prior
Solves the question of how to choose the prior if nothing is known.
Strings and Conditional Probabilities

Strings: \( x = x_1x_2...x_n \) with \( x_t \in \mathcal{X} \) and \( x_{1:m} := x_1x_2...x_{m-1}x_m \).

\( \rho(x_1...x_n) \) is the probability that an (infinite) sequence starts with \( x_1...x_n \).

Heavy use of Bayes’ rule in the following forms:

\[
\rho(x_n|x_{1:n}) = \frac{\rho(x_{1:n})}{\rho(x_{1:n})},
\]

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

If the true prior probability \( \mu(x_1...x_n) \) is known, then the optimal scheme is to minimize the \( \mu \)-expected loss.

Interpretation of Probabilities

Frequentist: Probabilities come from experiments.
Objectivists: Probabilities are real aspects of the world.
Subjectivists: Probabilities describe ones believe.
Probability of Sunrise Tomorrow

What is the probability that the sun will rise tomorrow? It is $\mu(1|d)$, where $d$ is the lifetime of the sun in days.

$1 = \text{sun raised. } 0 = \text{sun will not raise.}$

- The probability is undefined, because there has never been an experiment that tested the existence of the sun tomorrow (reference class problem).
- The probability is 1, because in all experiments that have been done (past days) the sun raised.
- The probability is $1 - \epsilon$, where $\epsilon$ is the proportion of stars in the universe that explode in a supernova per day.
- The probability is $(d + 1)/(d + 2)$ (Laplace estimate by assuming a Bernoulli process with uniformly distributed raising prior probability $p$).
- The probability can be derived from the type, age, size and temperature of the sun, even though we never have observed another star with those exact properties.

Solomonoff solved the problem of unknown prior $\mu$ by introducing a universal probability distribution $\xi$ related to Algorithmic Information Theory.
Kolmogorov Complexity

The Kolmogorov Complexity of a string $x$ is the length of the shortest program producing $x$.

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

The definition is "nearly" independent of the choice of $U$

$$|K_U(x) - K_{U'}(x)| < c_{UU'}, \quad K_U(x) \equiv K_{U'}$$

$\equiv$ indicates equality up to a constant $c_{UU'}$ independent of $x$.

$K$ satisfies most properties an information measure should satisfy:

$$K(xy) \leq K(x) + K(y).$$

$K(x)$ is not computable, but only co-enumerable (semi-computable).
Universal Probability Distribution

The universal semimeasure is the probability that output of $U$ starts with $x$ when the input is provided with fair coin flips

$$\xi(x) = \sum_{\mu_i \in \mathcal{M}} w_{\mu_i} \cdot \mu_i(x) \overset{\times}{=} \sum_{p : U(p) = x^*} 2^{-l(p)}, \text{ e.g. } w_{\mu_i} = 2^{-K(\mu_i)}$$

[Solomonoff 64]

Universality property of $\xi$: $\xi$ dominates every computable probability

$$\xi(x) \geq w_{\mu_i} \cdot \mu_i(x) \quad \forall \mu_i \in \mathcal{M}$$

Furthermore, the $\mu$ expected squared distance sum between $\xi$ and computable $\mu$

$$\sum_{t=1}^{\infty} \sum_{x_{1:t}} \mu(x_{<t}) (\xi(x_t|x_{<t}) - \mu(x_t|x_{<t}))^2 \leq \ln 2$$

[Solomonoff 78] (for binary alphabet)

$$\Rightarrow \xi(x_n|x_{<n}) \overset{n \to \infty}{\rightarrow} \mu(x_n|x_{<n}) \text{ with } \mu \text{ probability } 1 \quad \Rightarrow \quad \xi$$

is good guess for $\mu$. 


**Convergence Theorem**

The universal conditional probability $\xi(x_t|x_{<t})$ of the next symbol is related to the true conditional probability $\mu(x_t|x_{<t})$ in the following ways:

1. $\sum_{t=1}^{n} \mathbb{E} \left[ \sum_{x_t} \left( \mu(x_t|x_{<t}) - \xi(x_t|x_{<t}) \right)^2 \right] \equiv S_n \leq D_n \leq \ln w^{-1} \mu < \infty$

2. $\sum_{x_t} \left( \mu(x_t|x_{<t}) - \xi(x_t|x_{<t}) \right)^2 \equiv s_t(x_{<t}) \leq d_t(x_{<t}) \rightarrow 0$ for $t \rightarrow \infty$

3. $\xi(x'_t|x_{<t}) \rightarrow \mu(x'_t|x_{<t})$ for $t \rightarrow \infty$ with $\mu$ probability 1

4. $\sum_{t=1}^{n} \mathbb{E} \left[ \left( \frac{\sqrt{\xi(x_t|x_{<t})}}{\mu(x_t|x_{<t})} - 1 \right)^2 \right] \leq D_n \leq \ln w^{-1} \mu < \infty$

5. $\frac{\xi(x_t|x_{<t})}{\mu(x_t|x_{<t})} \rightarrow 1$ for $t \rightarrow \infty$ with $\mu$ probability 1

where $d_t = \sum_{x_n} \mu(x_n|x_{<n}) \ln \frac{\mu(x_n|x_{<n})}{\xi(x_n|x_{<n})}$ and $D_n = \sum_{x_{1:n}} \mu(x_{1:n})$ are relative entropies, and $w_\mu$ is the weight of $\mu$ in $\xi$. 

Universal Sequence Prediction

A prediction is very often the basis for some decision. The decision itself leads to some reward or loss. Let $\ell_{x_t y_t} \in [0, 1]$ be the 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>Loss</th>
<th>sunny</th>
<th>rainy</th>
</tr>
</thead>
<tbody>
<tr>
<td>umbrella</td>
<td>0.3</td>
<td>0.1</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 prediction scheme

$$y_t^{\Lambda_{\rho}} := \arg\min_{y_t \in \mathcal{Y}} \sum_{x_t} \rho(x_t|x_{<t}) \ell_{x_t y_t}$$

which minimizes the $\rho$-expected loss. The actual $\mu$-expected loss when $\Lambda_{\rho}$ predicts the $t^{th}$ symbol and the total $\mu$-expected loss in the first $n$ predictions are

$$l_t^{\Lambda_{\rho}}(x_{<t}) := \sum_{x_t} \mu(x_t|x_{<t}) \ell_{x_t y_t^{\Lambda_{\rho}}}, \quad L_n^{\Lambda_{\rho}} := \sum_{t=1}^{n} \sum_{x_{<t}} \mu(x_{<t}) \ell_{x_t y_t^{\Lambda_{\rho}}}.$$
Loss Bounds (Main Theorem)

\[ L_n^{\Lambda_{\mu}} \] made by the informed scheme \( \Lambda_{\mu} \),
\[ L_n^{\Lambda_{\xi}} \] made by the universal scheme \( \Lambda_{\xi} \),
\[ L_n^{\Lambda} \] made by any (causal) prediction scheme \( \Lambda \).

i) \[ L_n^{\Lambda_{\mu}} \leq L_n^{\Lambda} \] for any (causal) prediction scheme \( \Lambda \).

ii) \[ 0 \leq L_n^{\Lambda_{\xi}} - L_n^{\Lambda_{\mu}} \leq 2D_n + 2\sqrt{L_n^{\Lambda_{\mu}} D_n} \]

iii) if \( L_{\infty,\Lambda_{\mu}} \) is finite, then \( L_{\infty,\Lambda_{\xi}} \) is finite.

iv) \[ \frac{L_n^{\Lambda_{\xi}}}{L_n^{\Lambda_{\mu}}} = 1 + O((L_n^{\Lambda_{\mu}})^{-1/2}) \quad L_n^{\Lambda_{\mu}} \to \infty \]

v) \[ \sum_{t=1}^{n} \mathbb{E}[(l_{t\Lambda_{\xi}}(x_{<t}) - l_{t\Lambda_{\mu}}(x_{<t}))^2] \leq 2D_n \leq 2 \ln w_{\mu} \]

vi) \[ 0 \leq l_{t\Lambda_{\xi}}(x_{<t}) - l_{t\Lambda_{\mu}}(x_{<t}) \leq \left\{ \frac{\sqrt{2d_t(x_{<t})}}{2d_t(x_{<t}) + 2\sqrt{l_{t\Lambda_{\mu}}(x_{<t})}} \right\} \]

where \( D_n := \sum_{x_1:n} \mu(x_1:n) \ln \frac{\mu(x_1:n)}{\xi(x_1:n)} \leq \ln \frac{1}{w_{\mu}} = \ln 2 \cdot K(\mu) \).

Remark: The bound is valid for any loss function \( \in [0, 1] \) with no assumptions (like i.i.d., Markovian, stationary, ergodic, ...) on the structure of the \mu_i \in M.
Example Application

A dealer has two dice, one with 2 white and 4 black faces, the other with 4 white and 2 black faces. He chooses a die according to some deterministic rule. In every round, we bet $s = $3 on white or black and receive $r = $5 for every correct prediction.

If we know $\mu$, i.e. the die the dealer chooses, we should predict the color which is on 4 sides and win money. Expected Profit ($\equiv -$Loss): $P_{n\Lambda_\mu}/n = \frac{1}{3}$

If we don’t know $\mu$ we can use Solomonoff prediction scheme $\Lambda_\xi$ with asymptotically the same profit:

$$\frac{P_{n\Lambda_\xi}}{P_{n\Lambda_\mu}} = 1 - O(n^{-1/2})$$

Bound on Winning Time

Estimate of the number of rounds before reaching the winning zone:

$$P_{n\Lambda_\xi} > 0 \quad \text{if} \quad L_n^{\Lambda_\xi} < 0 \quad \text{if} \quad n > 330 \ln 2 \cdot K(\mu) + O(1)$$

$\Lambda_\xi$ is asymptotically optimal with rapid convergence.
General Bound for Winning Time

For every (passive) game of chance for which there exists a winning strategy, you can make money by using $\Lambda_\xi$ even if you don’t know the underlying process/algorithm.

$\Lambda_\xi$ finds and exploits every regularity.

The time $n$ needed to reach the winning zone is

$$n \leq \left( \frac{2p\Delta}{\bar{p}_n\Lambda_\mu} \right)^2 \cdot \ln \frac{1}{w_\mu}, \quad \bar{p}_n\Lambda_\mu := \frac{1}{n} \sum_{t=1}^{n} p_t\Lambda_\mu, \quad p\Delta = \ldots$$
Generalization: Continuous Probability Classes

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

$$M := \{\mu_\theta : \theta \in \mathbb{R}^d\}, \quad \xi(x_{1:n}) := \int_{\mathbb{R}^d} d\theta w(\theta) \cdot \mu_\theta(x_{1:n}),$$

The only property of $\xi$ needed was $\xi(x_{1:n}) \geq w_{\mu_i} \cdot \mu_i(x_{1:n})$ which was obtained by dropping the sum over $\mu_i$. Here, restrict the integral over $\mathbb{R}^d$ to a small vicinity of $\theta$. For sufficiently smooth $\mu_\theta$ and $w(\theta)$ we expect

$$\xi(x_{1:n}) \gtrsim |N_\delta_n| \cdot w(\theta) \cdot \mu_\theta(x_{1:n}) \implies D_n \lesssim \ln \frac{1}{w_\mu} + \ln \frac{1}{|N_\delta_n|}.$$

The average Fisher information $\bar{J}_n$ measures the curvature (parametric complexity) of $\ln \mu_\theta$. Under some weak regularity conditions on $\bar{J}_n$ one can show

$$D_n := \sum_{x_{1:n}} \mu(x_{1:n}) \ln \frac{\mu(x_{1:n})}{\xi(x_{1:n})} \leq \ln \frac{1}{w_\mu} + \frac{d}{2} \ln \frac{n}{2\pi} + \frac{1}{2} \ln \det \bar{J}_n + o(1)$$

i.e. $D_n$ grows only logarithmically with $n$. 

Optimality of the Universal Predictor

• There are $M$ and $\mu \in M$ and weights $w_\mu$ for which the loss bounds are tight.

• The universal prior $\xi$ is pareto-optimal, in the sense that there is no $\rho$ with $F(\mu_a, \rho) \leq F(\mu_a, \xi)$ for all $\mu_a \in M$ and strict inequality for at least one $\mu_a$, where $F$ is the instantaneous or total squared distance $s_t$, $S_n$, $d_t$, $D_n$, or error $e_t$, $E_n$, or loss $l_t$, $L_n$.

• $\xi$ is elastic pareto-optimal in the sense that by accepting a slight performance decrease in some environments one can only achieve a slight performance increase in other environments.

• Within the set of enumerable weight functions with short program, the universal weights $w_\nu = 2^{-K(\nu)}$ lead to the smallest performance bounds (to $\ln w^{-1}_\mu$) constant in all enumerable environments.

Does all this justify Occam’s razor?
Larger & Smaller Environmental Classes

- all finitely computable probability measures
  \((\xi \not\in M \text{ in no sense computable})\)

- all enumerable (approximable from below) semi-measures [Solomonoff 64, 78]
  \((\xi \in M \text{ enumerable})\)

- all cumulatively enumerable semi-measures [Schmidhuber 01]
  (distribution enumerable and \(\in M\))

- all approximable (asymptotically computable) measures [Schmidhuber 01]
  \((\xi \not\in M \text{ in no sense computable})\)

- Speed prior related to Levin complexity and Levin search [Schmidhuber 01]
  (which distributions are dominated?)

- finite-state automata instead of general Turing machines [Feder et al. 92]
  to Lempel-Ziv data compression \((\xi \not\in M)\)
Generalization: The Universal AI\(^\xi\)

Universal AI = Universal Induction + Sequential Decision Theory

Replace \(\mu^{AI}\) in decision theory model AI\(\mu\) by an appropriate generalization of \(\xi\).

\[
\xi(y_{x1:t}) := \sum_{q:q(y_{1:t})=x_{1:t}} 2^{-l(q)}
\]

\[
y_t = \arg\max_y \sum_{x_t} \max_{y_{t+1}} \sum_{x_{t+1}} \ldots \max_{y_m} \sum_{x_m} (r(x_t) + \ldots + r(x_m)) \cdot \xi(x_{t:m}|y_{t:m})
\]

Claim: AI\(\xi\) is the most intelligent environmental independent, i.e. universally optimal, agent possible.

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

[Proceedings of ECML-2001] and [http://www.hutter1.de/ai]
Further Generalizations:

- Time and history dependent loss function in general interval $[\ell_{\min}, \ell_{\max}]$.
- Infinite (countable and uncountable) action/decision space.
- Partial Sequence Prediction.
- Independent Experiments & Classification.

Outlook:

- Infinite (prediction) alphabet $\mathcal{X}$.
- Delayed and Probabilistic Sequence Prediction.
- Unification with (Lossbounds for) aggregating strategies.
- Determine suitable performance measures for universal AI $\xi$.
- Study learning aspect of $\Lambda_\xi$ and AI $\xi$.
- Information theoretic interpretation of winning time.
- Implementation and application of $\Lambda_\xi$ for specific finite $\mathcal{M}$.
- Downscale theory and results to MDL approach.
Conclusions

- Solomonoff’s prediction scheme, which is related to Kolomogorov complexity, formally solves the general problem of induction.

- We proved convergence and loss-bounds for Solomonoff prediction showing that it is well suited, even for difficult prediction problems.

- We proved several optimality properties for Solomonoff prediction.

- We made no structural assumptions on the probability distributions $\mu_i \in M$.

- The bounds are valid for any bounded loss function.

- We proved a bound on the time to win in games of chances.

- Discrete and continuous probability classes have been considered.

- Generalizations to active agents with reinforcement feedback have been suggested.

- At least all this is a lot of evidence that Occam’s razor is a useful principle.