# Error Bounds for Universal Solomonoff Sequence Prediction

Marcus Hutter

IDSIA, Galleria 2, CH-6928 Manno-Lugano, Switzerland
marcus@hutter1.de http://www.hutter1.de

12th July, 2000

**Abstract**

Universal Induction = Ockham + Epicur + Bayes

$$\frac{\#\text{ Errors(Universal Prediction Scheme)}}{\#\text{ Errors(Any other Prediction Scheme)}} \leq 1+o(1)$$

# Table of Contents

# Induction = Predicting the Future

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

Philosophical Dilemma:

- 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 start configuration.

- Bayes' rule for conditional probabilities

Given sequence $x_1...x_{k-1}$ what is the next letter $x_k$?

# Strings and Conditional Probabilities

Binary strings: $x = x_1 x_2 ... x_n$ with $x_k \in \{0,1\}$.
$x_{1:m} := x_1 x_2 ... x_{m-1} x_m \quad x_{<n} := x_1 ... x_{n-1}$.

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

An underlined argument $\underline{x_k}$ is a probability variable. Non-underlined arguments $x_k$ represent conditions. With this convention, Bayes' rule has the form:

$$\rho(x_{<n}\underline{x_n}) = \rho(\underline{x}_{1:n})/\rho(\underline{x}_{<n}),$$
$$\rho(\underline{x_1 ... x_n}) = \rho(\underline{x}_1)\cdot\rho(x_1\underline{x}_2)\cdot ... \cdot\rho(x_1 ... x_{n-1}\underline{x}_n).$$

If the true prior probability $\mu(\underline{x}_1 ... \underline{x}_n)$ is known, then the optimal scheme is to predict the $x_k$ with highest conditional $\mu$ probability if $x_{<k}$ is known, i.e. $\text{maxarg}_{x_k}\, \mu(x_{<k}\underline{x}_k)$.

# Interpretation of Probabilities

Frequentist: Probabilities come from experiments.
Objectivist: Probabilities are real aspects of the world.
Subjectivist: Probabilities describe ones believe.

# Probability of Sunrise Tomorrow

What is the probability that the sun will rise tomorrow? It is $\mu(1^d\underline{0})$. $d =$ actual lifetime of the sun in days. $1 =$ sun raised. $0 =$ sun will not raise.

- The probability is undefined, because there has never been an experiment that tested the existence of the sum $tomorrow$ (reference class problem).
- The probability is 1, because in all experiments that have been done (on 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(p) 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 $\xi$ based on Algorithmic Information Theory.

# Kolmogorov Complexity

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

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

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

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

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

$K$ satisfies most properties an information measure should satisfy, e.g. $K(xy) \stackrel{+}{\leq} K(x) + K(y)$.

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

# 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

$$\xi(\underline{x}) \;:=\; \sum_{p\,:\,U(p)=x*} 2^{-l(p)} \;\stackrel{\times}{=}\; \sum_{\rho} 2^{-K(\rho)}\rho(x)$$

[Solomonoff 64]

Universality property of $\xi$: $\xi$ maximizes every computable probability distribution

$$\xi(\underline{x}) \;\stackrel{\times}{\geq}\; 2^{-K(\rho)}\cdot\rho(\underline{x}) \quad \forall\rho$$

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

$$\sum_{k=1}^{\infty}\sum_{x_{1:k}}\mu(\underline{x}_{1:k})(\xi(x_{<k}\underline{x}_k)-\mu(x_{<k}\underline{x}_k))^2 \;\stackrel{+}{<}\; \tfrac{1}{2}\ln 2\cdot K(\mu)$$

[Solomonoff 78] (for binary alphabet)

# Universal Sequence Prediction

$\Rightarrow \xi(x_{<n}\underline{x}_n) \stackrel{n\to\infty}{\longrightarrow} \mu(x_{<n}\underline{x}_n)$ with $\mu$ probability 1.

$\Rightarrow$ Replacing $\mu$ by $\xi$ might not introduce many additional prediction errors.

General scheme: Predict $x_k$ with prob. $\rho(x_{<k}\underline{x}_k)$.

The predictor is deterministic, if $\rho(\underline{x}_{1:n}) \in \{0,1\}$.

Probability of making a wrong prediction:
$e_{n\rho}(x_{<n}) := \sum_{x_n\in\{0,1\}} \mu(x_{<n}\underline{x}_n)[1-\rho(x_{<n}\underline{x}_n)]$

Total $\mu$-expected errors in the first $n$ steps:
$E_{n\rho} := \sum_{k=1}^{n} \sum_{x_1\ldots x_{k-1}} \mu(\underline{x}_{<k})\cdot e_{k\rho}(x_{<k})$

Kullback Leibler distance between $\mu$ and $\xi$:
$h_n(x_{<n}) = \sum_{x_n} \mu(x_{<n}\underline{x}_n) \ln \frac{\mu(x_{<n}\underline{x}_n)}{\xi(x_{<n}\underline{x}_n)}$

$H_n$ is then defined as the sum-expectation:
$H_n := \sum_{k=1}^{n} \sum_{x_{<k}} \mu(\underline{x}_{<k})\cdot h_k(x_{<k}) \stackrel{+}{<} \ln 2\cdot K(\mu)$
[Solomonoff 78]

# Error Bounds

Comparison of the expected number of errors

$E_{n\mu}$ made by the informed scheme $\mu$,
$E_{n\xi}$ made by the universal scheme $\xi$,
$E_{n\rho}$ made by an arbitrary scheme $\rho$.

$$
\begin{aligned}
i) \quad & |E_{n\xi} - E_{n\mu}| \quad < \quad H_n + \sqrt{2E_{n\mu}H_n} \\
ii) \quad & E_{n\mu} \quad \leq \quad 2E_{n\rho} \quad , \quad e_{n\mu} \leq 2e_{n\rho} \\
iii) \quad & E_{n\xi} \quad < \quad 2E_{n\rho} + H_n + \sqrt{4E_{n\rho}H_n}
\end{aligned}
$$

[Hutter 99]

For computable $\mu$, i.e. for $K(\mu) < \infty$, the following statements immediately follow:

$$
\begin{aligned}
vii) \quad & \text{if } E_{\infty\mu} \text{ is finite, then } E_{\infty\xi} \text{ is finite} \\
viii) \quad & E_{n\xi}/E_{n\mu} \quad = \; 1 + O(E_{n\mu}^{-1/2}) \stackrel{E_{n\mu}\to\infty}{\longrightarrow} 1 \\
ix) \quad & E_{n\xi} - E_{n\mu} = \qquad O(\sqrt{E_{n\mu}}) \\
x) \quad & E_{n\xi}/E_{n\rho} \quad \leq \; 2 + O(E_{n\rho}^{-1/2})
\end{aligned}
$$

# Deterministic Sequence Prediction

$\Theta_\rho$ is defined to predict the $x_n$ with higher $\rho$ probability

$$\Theta_\rho(x_{<n}\underline{x}_n) := \begin{cases} 0 & \text{for} \quad \rho(x_{<n}\underline{x}_n) < \frac{1}{2} \\ 1 & \text{for} \quad \rho(x_{<n}\underline{x}_n) > \frac{1}{2} \end{cases}$$

Comparison of the expected number of errors

$E_{n\Theta_\mu}$ made by the informed scheme $\Theta_\mu$,
$E_{n\Theta_\xi}$ made by the universal scheme $\Theta_\xi$,
$E_{n\rho}$ made by an arbitrary scheme $\rho$.

$$\begin{aligned} &i) \quad 0 \leq E_{n\Theta_\xi} - E_{n\Theta_\mu} < H_n + \sqrt{4E_{n\Theta_\mu}H_n + H_n^2} \\ &ii) \quad E_{n\Theta_\mu} \leq E_{n\rho} \quad , \quad e_{n\Theta_\mu} \leq e_{n\rho} \\ &iii) \quad E_{n\Theta_\xi} < E_{n\rho} + H_n + \sqrt{4E_{n\rho}H_n + H_n^2} \end{aligned}$$

For computable $\mu$, i.e. for $K(\mu) < \infty$, the following statements immediately follow:

$$\begin{aligned} &vii) \quad \text{if } E_{\infty\Theta_\mu} \text{ is finite, then } E_{\infty\Theta_\xi} \text{ is finite} \\ &viii) \quad E_{n\Theta_\xi}/E_{n\Theta_\mu} = 1 + O(E_{n\Theta_\mu}^{-1/2}) \overset{E_{n\Theta_\mu}\to\infty}{\longrightarrow} 1 \\ &ix) \quad E_{n\Theta_\xi} - E_{n\Theta_\mu} = O(\sqrt{E_{n\Theta_\mu}}) \\ &x) \quad E_{n\Theta_\xi}/E_{n\rho} \leq 1 + O(E_{n\rho}^{-1/2}) \end{aligned}$$

# 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.

Expected profit when using scheme $\rho$:
$P_{n\rho} := (n - E_{n\rho})r - ns = (2n - 5E_{n\rho})\$$

If we know $\mu$, i.e. the die the dealer chooses, we should predict the color which is on 4 sides and win money:
$E_{n\Theta_\mu}/n = \frac{1}{3}, \quad P_{n\Theta_\mu}/n = \frac{1}{3}\$ > 0$

With the probabilistic scheme we loose money:
$E_{n\mu}/n = \frac{1}{3}\cdot\frac{2}{3} + \frac{2}{3}\cdot\frac{1}{3}, \quad P_{n\mu}/n = -\frac{2}{9}\$ < 0$

If we don't know $\mu$ we can use Solomonoff prediction $\xi$ or $\Theta_\xi$ with asymptotically the same profit:
$P_{n\xi}/P_{n\mu} = 1 - O(n^{-1/2}) = P_{n\Theta_\xi}/P_{n\Theta_\mu},$

# Example Application (Profit Bound)

Estimate of the number of rounds before reaching the winning zone with the $\Theta_\xi$ system.

$P_{n\Theta_\xi} > 0$ if
$E_{n\Theta_\xi} < (1-s/r)n$ if
$E_{n\Theta_\mu} + H_n + \sqrt{4E_{n\Theta_\mu}H_n + H_n^2} \; < \; (1-s/r)\cdot n$ if
$n \; > \; 330 \ln 2 \cdot K(\mu) + O(1)$.

$\Theta_\xi$ is asymptotically optimal with rapid convergence.

# Generalization

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

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

# Definition of the Universal AI$\xi$ Model

Universal AI = Universal Induction + Decision Theory

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

$$\xi(\underline{yx}_{1:k}) := \sum_{q:q(y_{1:k})=x_{1:k}} 2^{-l(q)}$$

$$\dot{y}_k = \operatorname*{maxarg}_{y_k} \sum_{x_k} \max_{y_{k+1}} \sum_{x_{k+1}} ... \max_{y_{m_k}} \sum_{x_{m_k}}$$
$$(c(x_k)+ ... +c(x_{m_k}))\cdot\xi(\dot{y}\dot{x}_{<k}\underline{yx}_{k:m_k})$$

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.

# Outlook

- Generalize error bounds to non-binary sequences (done).
- Generalize error bounds to more general credit functions, i.e. credit $C_{ij}$ if outcome is $i$ and prediction was $j$.
- Error bounds for computable approximations to $\xi$.
- Determine suitable performance measures for the universal AI$\xi$ model and prove bounds.

# Conclusions

- We have proved several new error bounds for Solomonoff prediction $\xi$ in terms of informed prediction $\mu$ and in terms of general prediction schemes $\rho$.

- Theorem 1 and Corollary 1 summarize the results in the probabilistic case and Theorem 2 and Corollary 2 for the deterministic case.

- We have shown that in the probabilistic case $E_{n\xi}$ is asymptotically bounded by twice the number of errors of any other prediction scheme.

- In the deterministic variant of Solomonoff prediction this factor 2 is absent. It is well suited, even for difficult prediction problems, as the error probability $E_{\Theta_\xi}/n$ converges rapidly to that of the minimal possible error probability $E_{\Theta_\mu}/n$.