Foundations of Induction

Marcus Hutter
Australian National University
Canberra, ACT, 0200, Australia
http://www.hutter1.net/

Contents

Abstract

Humans and many other intelligent systems (have to) learn from experience, build models of the environment from the acquired knowledge, and use these models for prediction. In philosophy this is called inductive inference, in statistics it is called estimation and prediction, and in computer science it is addressed by machine learning.

I will first review unsuccessful attempts and unsuitable approaches towards a general theory of induction, including Popper's falsificationism and denial of confirmation, frequentist statistics and much of statistical learning theory, subjective Bayesianism, Carnap's confirmation theory, the data paradigm, eliminative induction, and deductive approaches. I will also debunk some other misguided views, such as the no-free-lunch myth and pluralism.

I will then turn to Solomonoff's formal, general, complete, and essentially unique theory of universal induction and prediction, rooted in algorithmic information theory and based on the philosophical and technical ideas of Ockham, Epicurus, Bayes, Turing, and Kolmogorov.

This theory provably addresses most issues that have plagued other inductive approaches, and essentially constitutes a conceptual solution to the induction problem. Some theoretical guarantees, extensions to (re)active learning, practical approximations, applications, and experimental results are mentioned in passing, but they are not the focus of this talk.

I will conclude with some general advice to philosophers and scientists interested in the foundations of induction.

Induction/Prediction Examples

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

Model selection: Are planetary orbits circles or ellipses?
How many wavelets do I need to describe my picture well?
Which genes can predict cancer?

Parameter estimation: Bias of my coin. Eccentricity of earth's orbit.

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

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

Question: Is there a general & formal & complete & consistent theory for induction & prediction?

Beyond induction: active/reward learning, fct. optimization, game theory.

The Need for a Unified Theory

Why do we need or should want a unified theory of induction?

- Finding new rules for every particular (new) problem is cumbersome.
- A plurality of theories is prone to disagreement or contradiction.
- Axiomatization boosted mathematics&logic&deduction and so (should) induction.
- Provides a convincing story and conceptual tools for outsiders.
- Automatize induction&science (that's what machine learning does)
- By relating it to existing narrow/heuristic/practical approaches we deepen our understanding of and can improve them.
- Necessary for resolving philosophical problems.
- Unified/universal theories are often beautiful gems.
- There is no convincing argument that the goal is unattainable.

Math $\Leftrightarrow$ Words

"There is nothing that can be said by mathematical symbols and relations which cannot also be said by words.

The converse, however, is false.
Much that can be and is said by words cannot be put into equations,

because it is nonsense."

(Clifford A. Truesdell, 1966)

Math ⇔ Words

"There is nothing that can be said by mathematical symbols and relations which cannot also be said by words.

The converse, however, is false.
Much that can be and is said by words
cannot be put into equations,

because it is non~~sense~~xxxxx-science."

Induction $\Leftrightarrow$ Deduction

Approximate correspondence between
the most important concepts in induction and deduction.

	Induction	$\Leftrightarrow$	Deduction
Type of inference:	generalization/prediction	$\Leftrightarrow$	specialization/derivation
Framework:	probability axioms	$\widehat{=}$	logical axioms
Assumptions:	prior	$\widehat{=}$	non-logical axioms
Inference rule:	Bayes rule	$\widehat{=}$	modus ponens
Results:	posterior	$\widehat{=}$	theorems
Universal scheme:	Solomonoff probability	$\widehat{=}$	Zermelo-Fraenkel set theory
Universal inference:	universal induction	$\widehat{=}$	universal theorem prover
Limitation:	incomputable	$\widehat{=}$	incomplete (Gödel)
In practice:	approximations	$\widehat{=}$	semi-formal proofs
Operation:	computation	$\widehat{=}$	proof

The foundations of induction are as solid as those for deduction.

Critique

Popper's Philosophy of Science is Flawed

- Popper was good at popularizing philosophy of science outside of philosophy.
- Popper's appeal: simple ideas, clearly expressed. Noble and heroic vision of science.
- This made him a pop star among many scientists.
- Unfortunately his ideas (falsificationism, corroboration) are seriously flawed.
- Further, there have been better philosophy/philosophers before, during, and after Popper (but also many worse ones!)
- Fazit: It's time to move on and change your idol.
- References: Godfrey-Smith (2003) Chp.4, Gardner (2001), Salmon (1981), Putnam (1974), Schilpp (1974).

Popper's Falsificationism

- Demarcation problem: What is the difference between a scientific and a non-scientific theory?
- Popper's solution: Falsificationism: A hypothesis is scientific if and only if it can be refuted by some possible observation. Falsification is a matter of deductive logic.
- Problem 1: Stochastic models can never be falsified in Popper's strong deductive sense, since stochastic models can only become unlikely but never inconsistent with data.
- Problem 2: Falsificationism alone cannot prefer to use a well-tested theory (e.g. how to build bridges) over a brand-new untested one, since both have not been falsified.

Popper on Simplicity

- Why should we a-priori prefer to investigate “reasonable” theories over “obscure” theories.

- Popper prefers simple over complex theories because he believes that simple theories are easier to falsify.

- Popper equates simplicity with falsifiability, so is not advocating a simplicity bias proper.

- Problem: A complex theory with fixed parameters is as easy to falsify as a simple theory.

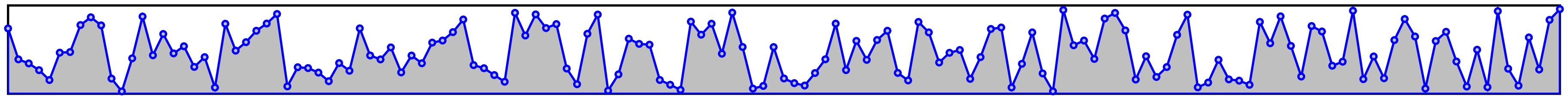
Popper's Corroboration / (Non)Confirmation

- Popper0 (fallibilism): We can never be completely certain about factual issues (✓)
- Popper1 (skepticism): Scientific confirmation is a myth.
- Popper2 (no confirmation): We cannot even increase our confidence in the truth of a theory when it passes observational tests.
- Popper3 (no reason to worry):
  Induction is a myth, but science does not need it anyway.
- Popper4 (corroboration): A theory that has survived many attempts to falsify it is "corroborated", and it is rational to choose more corroborated theories.
- Problem: Corroboration is just a new name for confirmation or meaningless.

The No Free Lunch (NFL) Theorem/Myth

- Consider algorithms for finding the maximum of a function, and compare their performance uniformly averaged over all functions over some fixed finite domain.

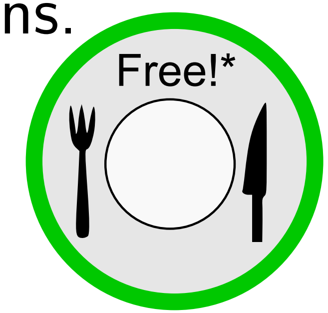
- Since sampling uniformly leads with (very) high probability to a totally random function (white noise), it is clear that on average no optimization algorithm can perform better than exhaustive search.



⇒ All reasonable optimization algorithms are equally good/bad on average.

- Conclusion correct, but obviously no practical implication, since nobody cares about the maximum of white noise functions.

- Uniform and universal sampling are both (non)assumptions, but only **universal sampling** makes sense and offers a free lunch.



*Subject to computation fees

Problems with Frequentism

- Definition: The probability of event E is the limiting relative frequency of its occurrence. $P(E) := \lim_{n\to\infty} \#_n(E)/n$.
- Circularity of definition: Limit exists only with probability 1. So we have explained "Probability of E" in terms of "Probability 1". What does probability 1 mean? [Cournot's principle can help]
- Limitation to i.i.d.: Requires independent and identically distributed (i.i.d) samples. But the real world is **not** i.i.d.
- Reference class problem: Example: Counting the frequency of some disease among "similar" patients. Considering all we know (symptoms, weight, age, ancestry, ...) there are no two similar patients. [Machine learning via feature selection can help]

Statistical Learning Theory

- Statistical Learning Theory predominantly considers i.i.d. data.
- E.g. Empirical Risk Minimization, PAC bounds, VC-dimension, Rademacher complexity, Cross-Validation is mostly developed for i.i.d. data.
- Applications: There are enough applications with data close to i.i.d. for Frequentists to thrive, and they are pushing their frontiers too.
- But the Real World is **not** (even close to) i.i.d.
- Real Life is a single long non-stationary non-ergodic trajectory of experience.

Limitations of Other Approaches

- Subjective Bayes: No formal procedure/theory to get prior.
- Objective Bayes: Right in spirit, but limited to small classes unless community embraces information theory.
- MDL/MML: practical approximations of universal induction.
- Pluralism is globally inconsistent.
- Deductive Logic: not strong enough to allow for induction.
- Non-monotonic reasoning, inductive logic, default reasoning do not properly take uncertainty into account.
- Carnap's confirmation theory: Only for exchangeable data. Cannot confirm universal hypotheses.
- Data paradigm: data may be more important than algorithms for "simple" problems, but a "lookup-table" AGI will not work.
- Eliminative induction: ignores uncertainty and information theory.

Summary

The criticized approaches
cannot serve as a general foundation of induction.

Conciliation

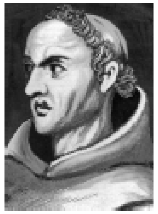
Of course most of the criticized approaches
do work in their limited domains, and
are trying to push their boundaries towards more generality.

And What Now?

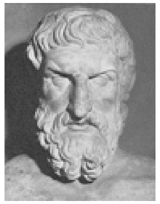
Criticizing others is easy and in itself a bit pointless.
The crucial question is whether there is something better out there.
And indeed there is, which I will turn to now.

Universal Induction

Foundations of Universal Induction



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

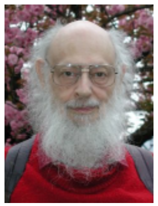


Epicurus' principle of multiple explanations
If more than one theory is consistent with the observations, keep all theories.

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

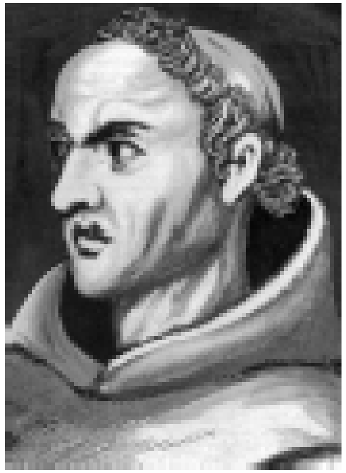
Turing's universal machine
Everything computable by a human using a fixed procedure can also be computed by a (universal) Turing machine.

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.
$\Rightarrow$ universal induction, formal Occam,AIT,MML,MDL,SRM,...

Science ≈ Induction ≈ Occam's Razor



- Grue Emerald Paradox:

  Hypothesis 1: All emeralds are green.

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

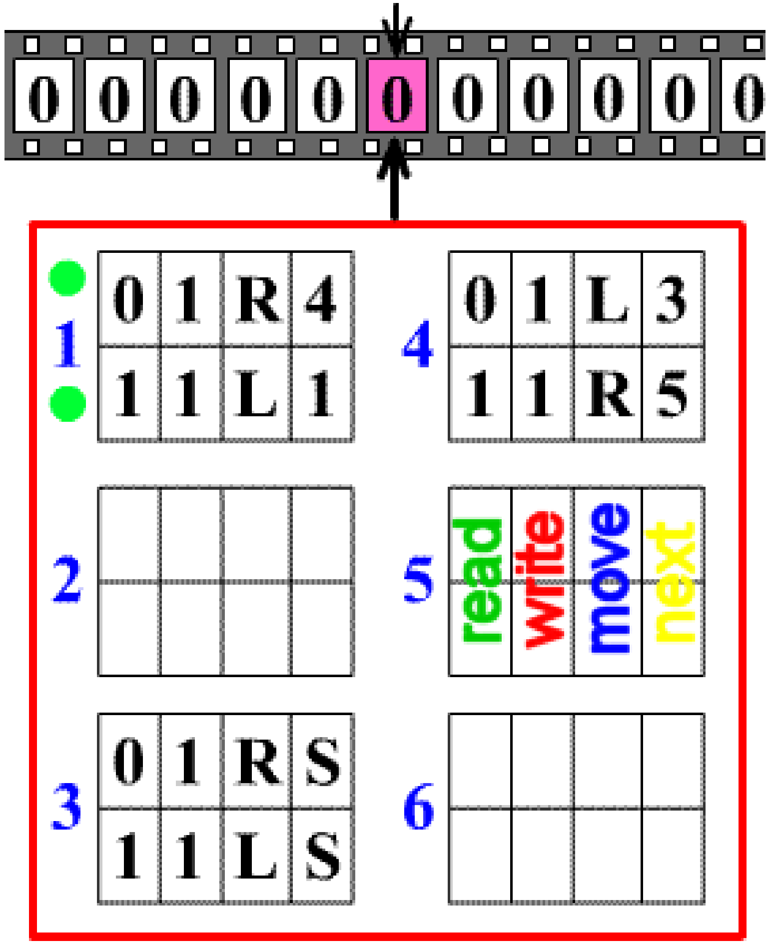
- Which hypothesis is more plausible? H1! Justification?

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

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

[The Grue problem goes much deeper. This is only half of the story]

Turing Machines & Effective Enumeration



- Turing machine (TM) = (mathematical model for an) idealized computer.
- See e.g. textbook [HMU06]

  Show Turing Machine in Action: TuringBeispielAnimated.gif

- Instruction i: If symbol on tape under head is 0/1, write 0/1/- and move head left/right/not and goto instruction=state j.
- {partial recursive functions } $\equiv$ {functions computable with a TM}.
- A set of objects $S = \{o_1, o_2, o_3, ...\}$ can be (effectively) enumerated $:\Longleftrightarrow \exists$ TM machine mapping i to $\langle o_i \rangle$, where $\langle\rangle$ is some (often omitted) default coding of elements in S.

Information Theory & Kolmogorov Complexity

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

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

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

Kolmogorov-complexity$(x) \;=\; K(x) \;=\; K_U(x) \qquad \leq K_T(x) + c_T$

Bayesian Probability Theory

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

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

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

Solution:

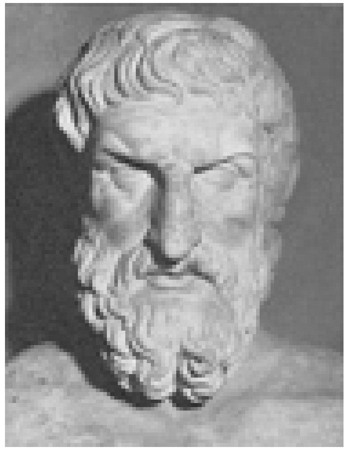
Bayes' rule:

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

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

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

Algorithmic Probability Theory

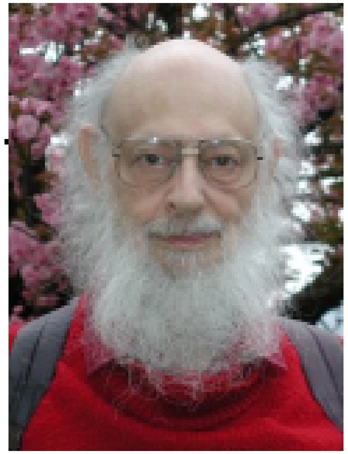


- Epicurus: If more than one theory is consistent with the observations, keep all theories.
- $\Rightarrow$ uniform prior over all H_i?
- Refinement with Occam's razor quantified in terms of Kolmogorov complexity:

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

- Fixing T we have a complete theory for prediction.
  Problem: How to choose T.
- Choosing U we have a universal theory for prediction.
  Observation: Particular choice of U does not matter much.
  Problem: Incomputable.

Inductive Inference & Universal Forecasting



- Solomonoff combined Occam, Epicurus, Bayes, and Turing into one formal theory of sequential prediction.
- $M(x)$ = probability that a universal Turing machine outputs x when provided with fair coin flips on the input tape.
- A posteriori probability of y given x is $M(y|x) = M(xy)/M(x)$.
- Given $x_1, .., x_{t-1}$, the probability of x_t is $M(x_t|x_1...x_{t-1})$.
- Immediate "applications":
  - Weather forecasting: $x_t \in \{\text{sun,rain}\}$.
  - Stock-market prediction: $x_t \in \{\text{bear,bull}\}$.
  - Continuing number sequences in an IQ test: $x_t \in I\!N$.
- Optimal universal inductive reasoning system!

Some Prediction Bounds for M

$M(x)$ = universal distribution. $h_n := \sum_{x_n}(M(x_n|x_{<n}) - \mu(x_n|x_{<n}))^2$

$\mu(x)$ = unknown true comp. distr. (no i.i.d. or any other assumptions)

- Total bound: $\sum_{n=1}^{\infty}\mathbf{E}[h_n] \leq K(\mu)\ln 2$, which implies
  Convergence: $M(x_n|x_{<n}) \to \mu(x_n|x_{<n})$ w.μ.p.1.
- Instantaneous i.i.d. bounds: For i.i.d. $\mathcal{M}$ with continuous, discrete, and universal prior, respectively:
  $\mathbf{E}[h_n] \stackrel{\times}{\leq} \frac{1}{n}\ln w(\mu)^{-1}$ and $\mathbf{E}[h_n] \stackrel{\times}{\leq} \frac{1}{n}\ln w_\mu^{-1} = \frac{1}{n}K(\mu)\ln 2$.
- Bounds for computable environments: Rapidly $M(x_t|x_{<t}) \to 1$ on every computable sequence $x_{1:\infty}$ (whichsoever, e.g. 1^∞ or the digits of π or e), i.e. M quickly recognizes the structure of the sequence.
- Weak instantaneous bounds: valid for all n and $x_{1:n}$ and $\bar{x}_n \neq x_n$:
  $2^{-K(n)} \stackrel{\times}{\leq} M(\bar{x}_n|x_{<n}) \stackrel{\times}{\leq} 2^{2K(x_{1:n}*)-K(n)}$
- Magic instance numbers: e.g. $M(0|1^n) \stackrel{\times}{=} 2^{-K(n)} \to 0$, 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 μ, we make few errors in future:
  $\sum_{t=n+1}^{\infty}\mathbf{E}[h_t|\omega_{1:n}] \stackrel{+}{\leq} [K(\mu|\omega_{1:n})+K(n)]\ln 2$

Some other Properties of M

- Problem of zero prior / confirmation of universal hypotheses:

  $$\text{P[All ravens black}|n\text{ black ravens]} \begin{cases} \equiv 0 \text{ in Bayes-Laplace model} \\ \stackrel{fast}{\longrightarrow} 1 \text{ for universal prior } w_\theta^U \end{cases}$$

- Reparametrization and regrouping invariance: $w_\theta^U = 2^{-K(\theta)}$ always exists and is invariant w.r.t. all computable reparametrizations f. (Jeffrey prior only w.r.t. bijections, and does not always exist)

- The Problem of Old Evidence: No risk of biasing the prior towards past data, since w_θ^U is fixed and independent of model class $\mathcal{M}$.

- The Problem of New Theories: Updating of $\mathcal{M}$ is not necessary, since universal class $\mathcal{M}_U$ includes already all.

- M predicts better than all other mixture predictors based on **any** (continuous or discrete) model class and prior, even in non-computable environments.

More Stuff / Critique / Problems

- Prior knowledge y can be incorporated by using "subjective" prior $w^U_{\nu|y} = 2^{-K(\nu|y)}$ 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 Artificial Intelligence

Induction→Prediction→Decision→Action

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

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

Example

Induction: Find a model of the world economy.

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

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

Action: Trading large quantities of stocks influences the market.

Sequential Decision Theory

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

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

(1) predict/make decision y_t,

(2) observe x_t,

(3) suffer loss Loss(x_t, y_t),

(4) $t \to t + 1$, goto (1)

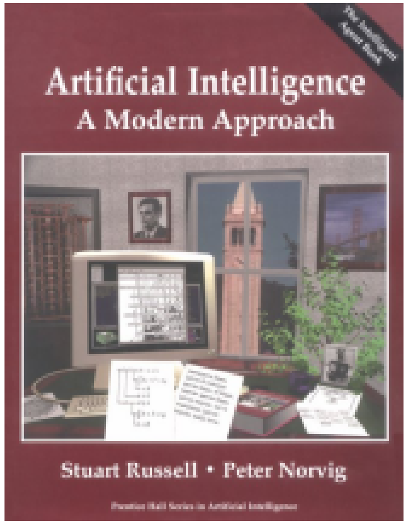
Goal: Minimize expected Loss.

Greedy minimization of expected loss is optimal if:

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

Problem: Expectation w.r.t. what?

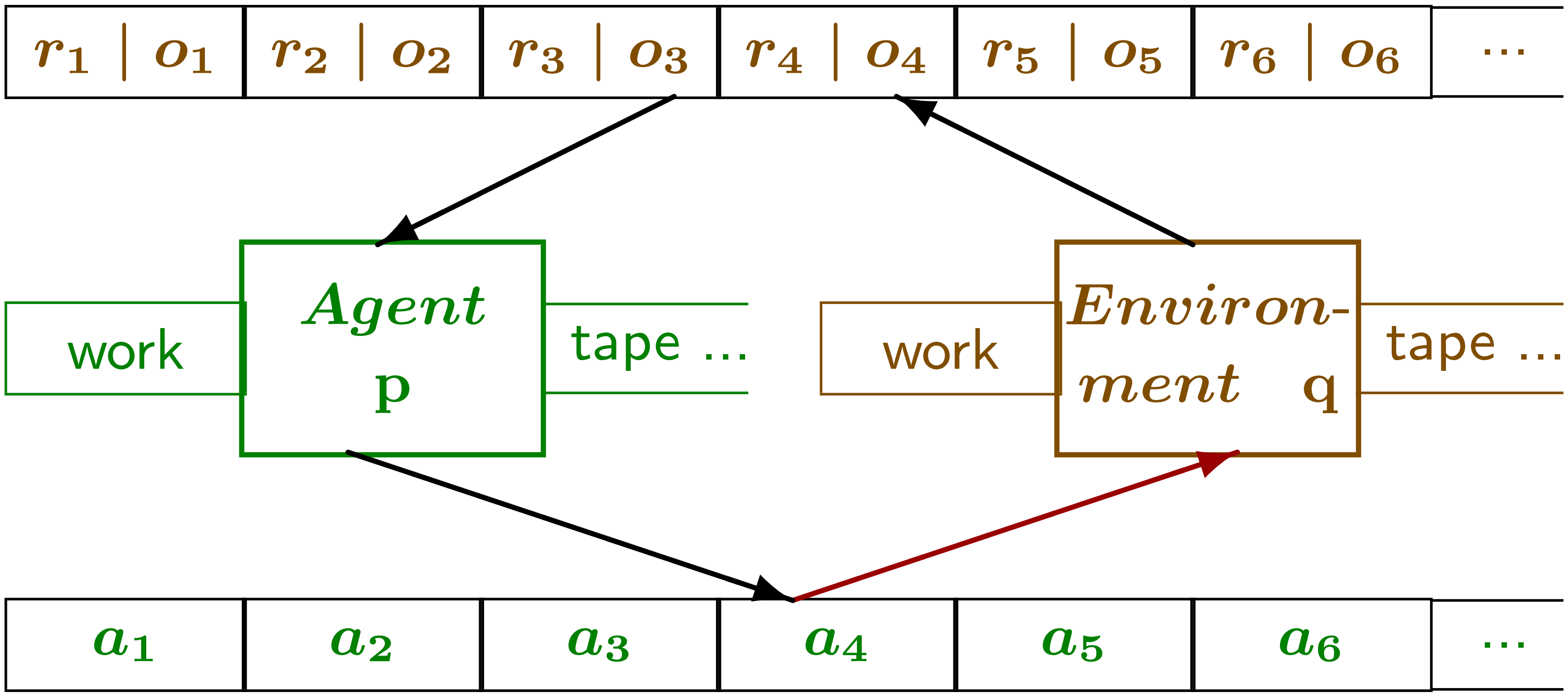
Solution: W.r.t. universal distribution M if true distr. is unknown.



Agent Model with Reward

if actions/decisions a

influence the environment q



Universal Artificial Intelligence

Key idea: Optimal action/plan/policy based on the simplest world model consistent with history. Formally ...

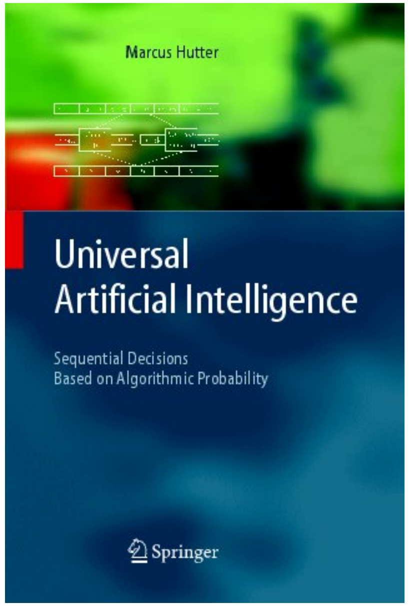
AIXI: $$a_k := \arg\max_{a_k} \sum_{o_k r_k} \ldots \max_{a_m} \sum_{o_m r_m} [r_k + \ldots + r_m] \sum_{p\,:\,U(p,a_1..a_m)=o_1 r_1..o_m r_m} 2^{-length(p)}$$

k=now, $a$ction, $o$bservation, $r$eward, Universal TM, $p$rogram, m=lifespan

AIXI is an elegant, complete, essentially unique, and limit-computable mathematical theory of AI.

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

Proof: For formalizations, quantifications, proofs see $\Rightarrow$



Problem: Computationally intractable.

Achievement: Well-defines AI. Gold standard to aim at. Inspired practical algorithms. Cf. infeasible exact minimax. [H'00-05]

Applications & Approximations

Approximations

- Exact computation of probabilities often intractable.
- Approximate the solution, not the problem!
- Cheap surrogates will fail in critical situations.
- Any heuristic method should be gauged against Solomonoff/AIXI gold-standard.
- Established/good approximations: MDL, MML, CTW, Universal search, Monte Carlo, ...
- AGI needs anytime algorithms powerful enough to approach Solomonoff/AIXI in the limit of infinite comp.time.

The Minimum Description Length Principle

- Approximation of Solomonoff, since M is incomputable:
- $M(x) \approx 2^{-K_U(x)}$ (quite good)
- $K_U(x) \approx K_T(x)$ (very crude)
- Predict y of highest $M(y|x)$ is approximately same as
- MDL: Predict y of smallest $K_T(xy)$.

Universal Clustering

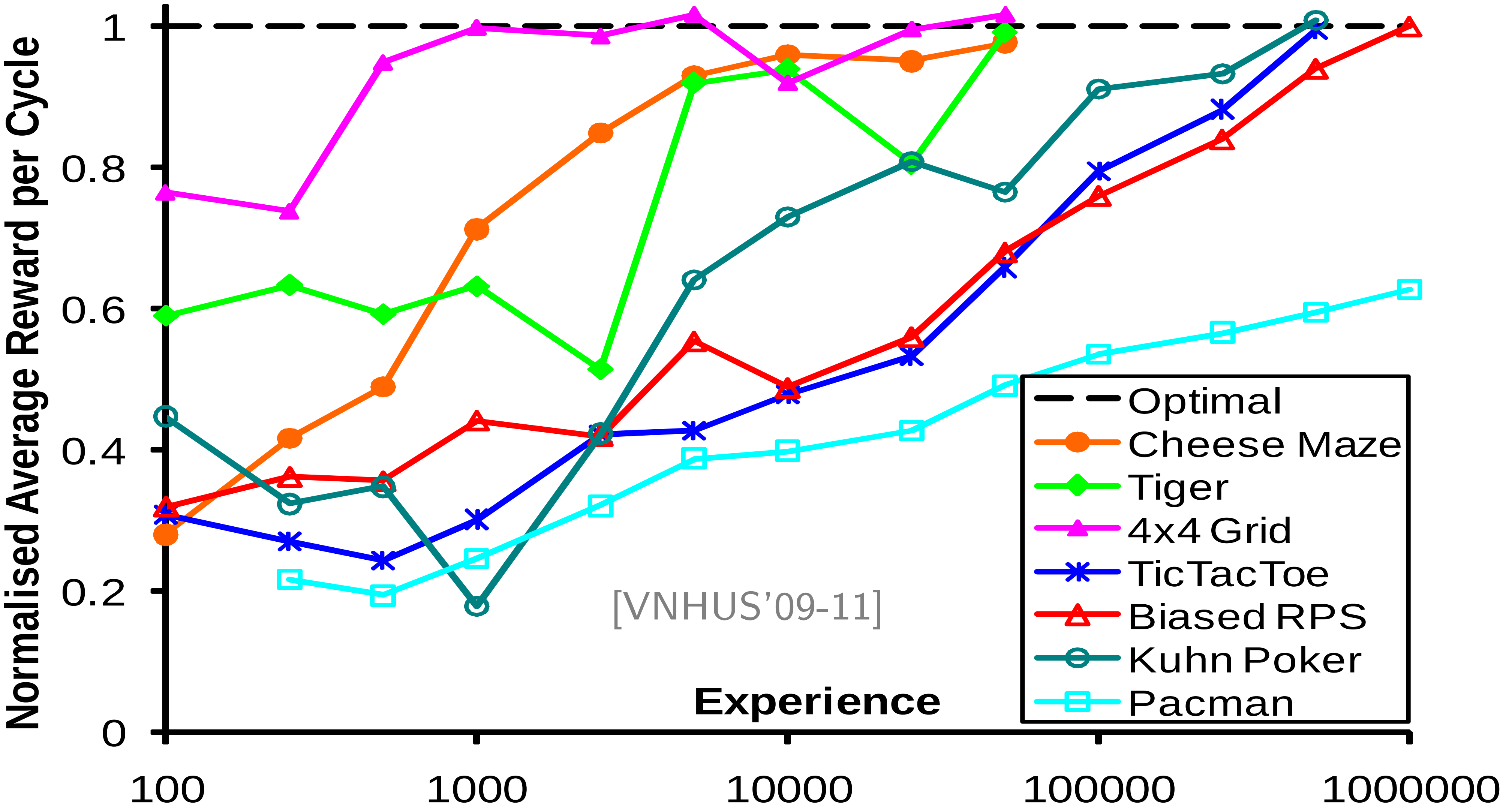
- Question: When is object x similar to object y?
- Universal solution: x similar to y
  $\Leftrightarrow$ x can be easily (re)constructed from y
  $\Leftrightarrow$ $K(x|y) := \min\{\ell(p) : U(p,y) = x\}$ is small.
- Universal Similarity: Symmetrize&normalize $K(x|y)$.
- Normalized compression distance: Approximate $K \equiv K_U$ by K_T.
- Practice: For T choose (de)compressor like lzw or gzip or bzip(2).
- Multiple objects $\Rightarrow$ similarity matrix $\Rightarrow$ similarity tree.
- Applications: Completely automatic reconstruction (a) of the evolutionary tree of 24 mammals based on complete mtDNA, and (b) of the classification tree of 52 languages based on the declaration of human rights and (c) many others. [Cilibrasi&Vitanyi'05]

Universal Search

- Levin search: Fastest algorithm for inversion and optimization problems.
- Theoretical application:
  Assume somebody found a non-constructive proof of P=NP, then Levin-search is a polynomial time algorithm for every NP (complete) problem.
- Practical (OOPS) applications (J. Schmidhuber)
  Maze, towers of hanoi, robotics, ...
- FastPrg: The asymptotically fastest and shortest algorithm for all well-defined problems.
- Computable Approximations of AIXI:
  AIXItl and AIξ and MC-AIXI-CTW and ΦMDP.
- Human Knowledge Compression Prize: (50'000€)

A Monte-Carlo AIXI Approximation

based on Upper Confidence Tree (UCT) search for planning
and Context Tree Weighting (CTW) compression for learning



Conclusion

Summary

- Conceptually and mathematically the problem of induction is solved.
- Computational problems and some philosophical questions remain.
- Ingredients for induction and prediction:
  Ockham, Epicurus, Turing, Bayes, Kolmogorov, Solomonoff
- For decisions and actions: Include Bellman.
- Mathematical results: consistency, bounds, optimality, many others.
- Most Philosophical riddles around induction solved.
- Experimental results via practical compressors.

Induction $\approx$ Science $\approx$ Machine Learning $\approx$
Ockham's razor $\approx$ Compression $\approx$ Intelligence.

Advice

to philosophers and scientists interested in the foundations of induction

- Accept Universal Induction (UI) as the best conceptual solution of the induction problem so far.
- Stand on the shoulders of giants like Bayes, Shannon, Turing, Kolmogorov, Solomonoff, Wallace, Rissanen, Bellman.
- Work out defects / what is missing, and try to improve UI, or
- Work on alternatives but then benchmark your approach against state of the art UI.
- Cranks who have not understood the giants and try to reinvent the wheel from scratch can safely be ignored.

Never trust a ~~theory~~ experiment if it is not supported by an ~~experiment~~ theory

When it's OK to ignore UI

- if your pursued approaches already works sufficiently well
- if your problem is simple enough (e.g. i.i.d.)
- if you do not care about a principled/sound solution
- if you're happy to succeed by trial-and-error (with restrictions)

Information Theory

- Information Theory plays an even more significant role for induction than this presentation might suggest.
- Algorithmic Information Theory is superior to Shannon Information.
- There are AIT versions that even capture Meaningful Information.

Outlook

- Use compression size as general performance measure (like perplexity is used in speech)

- Via code-length view, many approaches become comparable, and may be regarded as approximations to UI.

- This should lead to better compression algorithms which in turn should lead to better learning algorithms.

- Address open problems in induction within the UI framework.

Thanks! Questions? Details:

[RH11] S. Rathmanner and M. Hutter. *A philosophical treatise of universal induction*. Entropy, 13(6):1076–1136, 2011.

[Hut07] M. Hutter. *On universal prediction and Bayesian confirmation*. Theoretical Computer Science, 384(1):33–48, 2007.

[LH07] S. Legg and M. Hutter. *Universal intelligence: A definition of machine intelligence.* Minds & Machines, 17(4):391–444, 2007.

[Hut12] M. Hutter. *One Decade of Universal Artificial Intelligence*. In Theoretical Foundations of Artificial General Intelligence, 4:67–88, 2012.

[Hut05] M. Hutter. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*. Springer, Berlin, 2005.

[GS03] P. Godfrey-Smith. *Theory and Reality: An Introduction to the Philosophy of Science*. University of Chicago, 2003.