Foundations of Machine Learning

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

ANU RSISE NICTA

Machine Learning Summer School
MLSS-2008, 2 – 15 March, Kioloa

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 $p(H_i|D) \propto p(D|H_i)p(H_i)$
- Bayes needs prior(H_i)
- Occam+Epicurus: High prior for simple models.
- Kolmogorov/Solomonoff: Quantification of simplicity/complexity
- Bayes works 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 tutorial 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 tutorial concentrates on principled, unified, and exact methods.

Table of Contents

- Overview
- Philosophical Issues
- Bayesian Sequence Prediction
- Universal Inductive Inference
- 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 Tutorial 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 tutorial and place it into the wider machine learning context.

Philosophical Problems

- Does inductive inference work? Why? How?
- 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?
($d =$ past # days sun rose, $1 =$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 ϵ 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 occured $i \Longrightarrow$ Probability of next digit being i is $\frac{i}{n}$. Asymptotically $\frac{i}{n} \rightarrow \frac{1}{10}$ (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 π.

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

Example 3: Number Sequences

Sequence:

$x_1,$	$x_2,$	$x_3,$	$x_4,$	$x_5,$	...
1,	2,	3,	4,	?,	...

- $x_5 = 5$, since $x_i = i$ for $i = 1..4$.
- $x_5 = 29$, 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,?

- 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) $R \to B$ is confirmed by an R-instance with property B

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

(iii) Since $R \to B$ and $\neg B \to \neg R$ are logically equivalent,
$R \to B$ 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 tutorial 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 Tutorial is (Not) About

Dichotomies in Artificial Intelligence & Machine Learning

scope of my tutorial	⇔	**scope of other tutorials**
(machine) learning	⇔	(GOFAI) knowledge-based
statistical	⇔	logic-based
decision ⇔ prediction	⇔	induction ⇔ action
classification	⇔	regression
sequential / non-iid	⇔	independent identically distributed
online learning	⇔	offline/batch learning
passive prediction	⇔	active learning
Bayes ⇔ MDL	⇔	Expert ⇔ Frequentist
uninformed / universal	⇔	informed / problem-specific
conceptual/mathematical issues	⇔	computational issues
exact/principled	⇔	heuristic
supervised learning	⇔	unsupervised ⇔ RL learning
exploitation	⇔	exploration

Sequential/Online Prediction – Setup

In sequential or online prediction, for times $t = 1, 2, 3, ...$,

our predictor p makes a prediction $y_t^p \in \mathcal{Y}$

based on past observations $x_1, ..., x_{t-1}$.

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

The goal is to design predictors with small total loss or cumulative $\text{Loss}_{1:T}(p) := \sum_{t=1}^{T} \text{Loss}(x_t, y_t^p)$.

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

Example:

$\text{Loss}(x,y)$	$\mathcal{X} = \{\text{sunny}$,	$\text{rainy}\}$
$\mathcal{Y} = \{$ umbrella	0.1	0.3
sunglasses $\}$	0.0	1.0

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 Ω 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 repr. 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: $\boxed{\text{Beliefs follow the same rules as probabilities}}$

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 $H_i\cap H_j=\{\}\ \forall i\neq j$ 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.)

$$\boxed{\text{Solution:}\quad p(H_i|D)=\frac{p(D|H_i)p(H_i)}{\sum_{i\in I}p(D|H_i)p(H_i)}}$$

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 θ, i.e. H_θ :=Bernoulli(θ) with $\theta \in \Theta := [0,1]$.

Finite sequence: $x = x_1 x_2 ... x_n$ with n_1 ones and n_0 zeros.

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

Basic event: $\Gamma_x = \{\omega : \omega_1 = x_1, ..., \omega_n = x_n\}$ = 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: $p(\theta) := p(H_\theta) = 1$

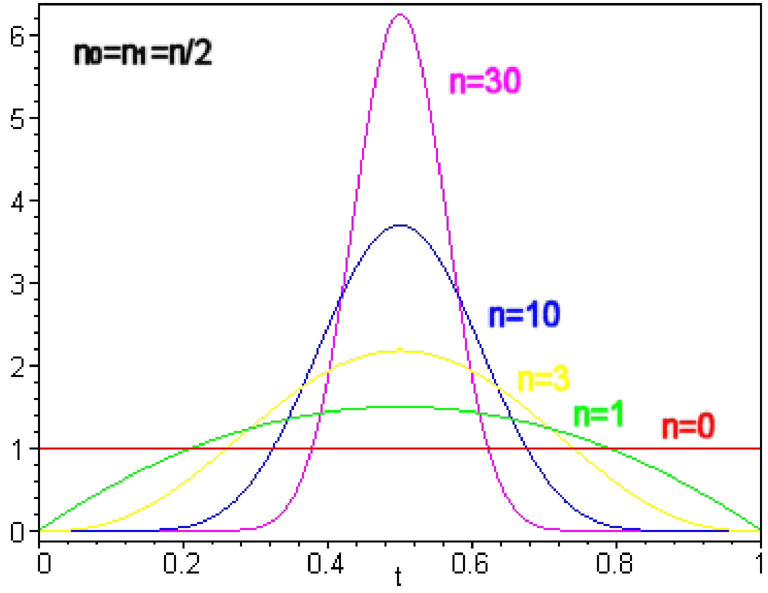
($\int_0^1 p(\theta)\,d\theta = 1$ instead $\sum_{i\in I} p(H_i) = 1$)

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 θ 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?

$$p(x_{n+1}=1|x_1...x_n) = \frac{p(x1)}{p(x)} = \frac{n_1+1}{n+2}$$

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.

The Bayes-Mixture Distribution ξ

- Assumption: The true (objective) environment μ is unknown.

- Bayesian approach: Replace true probability distribution μ by a Bayes-mixture ξ.

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

- The Bayes-mixture ξ is defined as

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

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

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

Convergence and Decisions

Goal: Given seq. $x_{1:t-1} \equiv x_{<t} \equiv x_1 x_2 ... x_{t-1}$, predict continuation x_t.

Expectation w.r.t. μ: $\mathbf{E}[f(\omega_{1:n})] := \sum_{x\in\mathcal{X}^n} \mu(x) f(x)$

KL-divergence: $D_n(\mu||\xi) := \mathbf{E}[\ln\frac{\mu(\omega_{1:n})}{\xi(\omega_{1:n})}] \leq \ln w_\mu^{-1}\ \forall n$

Hellinger distance: $h_t(\omega_{<t}) := \sum_{a\in\mathcal{X}}(\sqrt{\xi(a|\omega_{<t})} - \sqrt{\mu(a|\omega_{<t})})^2$

Rapid convergence: $\boxed{\sum_{t=1}^\infty \mathbf{E}[h_t(\omega_{<t})] \leq D_\infty \leq \ln w_\mu^{-1} < \infty}$ implies $\xi(x_t|\omega_{<t}) \to \mu(x_t|\omega_{<t})$, i.e. ξ is a good substitute for unknown μ.

Bayesian decisions: Bayes-optimal predictor Λ_ξ suffers instantaneous loss $l_t^{\Lambda_\xi} \in [0,1]$ at t only slightly larger than the μ-optimal predictor Λ_μ: $\sum_{t=1}^\infty \mathbf{E}[(\sqrt{l_t^{\Lambda_\xi}} - \sqrt{l_t^{\Lambda_\mu}})^2] \leq \sum_{t=1}^\infty 2\mathbf{E}[h_t] < \infty$ implies rapid $l_t^{\Lambda_\xi} \to l_t^{\Lambda_\mu}$.

Pareto-optimality of Λ_ξ: Every predictor with loss smaller than Λ_ξ in some environment $\mu \in \mathcal{M}$ must be worse in another environment.

Generalization: Continuous Classes $\mathcal{M}$

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

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

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

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

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 μ.
- Bound on the relative entropy between ξ and μ.

$\Rightarrow$ posterior of ξ converges rapidly to the true posterior μ.

- ξ 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
- Fast Convergence
- How to Choose the Prior – Universal
- Kolmogorov Complexity
- How to Choose the Model Class – Universal
- Universal is Better than Continuous Class
- 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: Fast convergence, and in contrast to most classical continuous prior densities has no zero p(oste)rior 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?

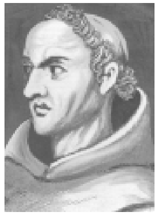
These are instances of the important problem of inductive inference or time-series forecasting or sequence prediction.

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.

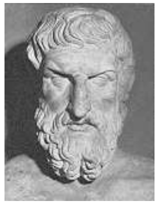
Beyond induction: active/reward learning, fct. optimization, game theory.

Foundations of Universal Induction



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.

⇒ 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 μ, i.e. $\mu(x) := \mathrm{P}[x|\mu]$ is the μ-probability that ω starts with $x \in \mathcal{X}^n$.
- Model class: We assume that μ is unknown but known to belong to a countable class of environments=models=measures $\mathcal{M} = \{\nu_1, \nu_2, ...\}$. [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: $w_\nu := \mathrm{P}[H_\nu]$ is our prior belief in H_ν

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

⇒ Posterior: $w_\nu(x) := \mathrm{P}[H_\nu|x] = \frac{\mathrm{P}[x|H_\nu]\mathrm{P}[H_\nu]}{\mathrm{P}[x]}$ is our posterior belief in ν (Bayes' rule).

How to Choose the Prior?

- Subjective: quantifying personal prior belief (not further discussed)
- Objective: based on rational principles (agreed on by everyone)
- Indifference or symmetry principle: Choose $w_\nu = \frac{1}{|\mathcal{M}|}$ for finite $\mathcal{M}$.
- Jeffreys or Bernardo's prior: Analogue for compact parametric spaces $\mathcal{M}$.
- Problem: The principles typically provide good objective priors for small discrete or compact spaces, but not for "large" model classes like countably infinite, non-compact, and non-parametric $\mathcal{M}$.
- Solution: Occam favors simplicity $\Rightarrow$ Assign high (low) prior to simple (complex) hypotheses.
- Problem: Quantitative and universal measure of simplicity/complexity.

Kolmogorov Complexity K(x)

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

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

For non-string objects o (like numbers and functions) we define $K(o):=K(\langle o\rangle)$, where $\langle o\rangle\in\mathcal{X}^*$ 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.

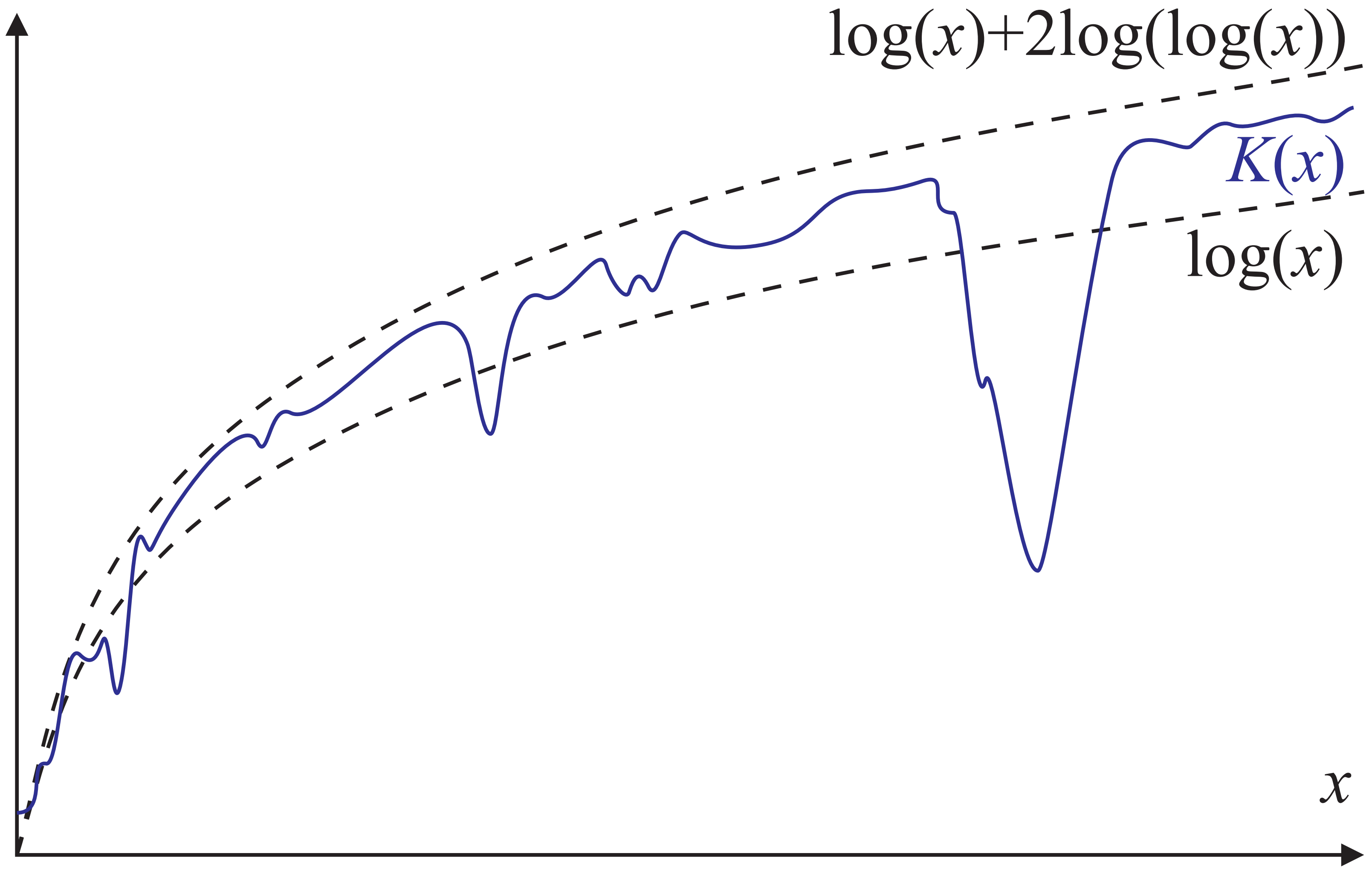
+ K shares many properties with Shannon entropy but is superior.

− $K(x)$ 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



The Universal Prior

- Quantify the complexity of an environment ν or hypothesis H_ν by its Kolmogorov complexity $K(\nu)$.
- Universal prior: $w_\nu = \boxed{w_\nu^U := 2^{-K(\nu)}}$ is a decreasing function in the model's complexity, and sums to (less than) one.

$\Rightarrow$ $D_n \leq K(\mu)\ln 2$, i.e. the number of ε-deviations of ξ from μ or l^{Λ_ξ} from l^{Λ_μ} is proportional to the complexity of the environment.

- No other semi-computable prior leads to better prediction (bounds).
- For continuous $\mathcal{M}$, we can assign a (proper) universal prior (not density) $w_\theta^U = 2^{-K(\theta)} > 0$ for computable θ, and 0 for uncomp. θ.
- This effectively reduces $\mathcal{M}$ to a discrete class $\{\nu_\theta \in \mathcal{M} : w_\theta^U > 0\}$ which is typically dense in $\mathcal{M}$.
- This prior has many advantages over the classical prior (densities).

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, ξ_U is semi-computable [ZL70].

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

- Formally: $\boxed{M(x) := \sum_{p\,:\,U(p)=x*} 2^{-\ell(p)}}$ 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 ν_p. ($\nu_p(x) = 1$ if $U(p) = x*$ and 0 else)

- $M(x)$ coincides with $\xi_U(x)$ within an irrelevant multiplicative constant.

Universal is better than Continuous Class&Prior

- Problem of zero prior / confirmation of universal hypotheses:

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

- 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 $\mathcal{M}$.

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

Convergence and Loss Bounds

- Total (loss) bounds: $\sum_{n=1}^{\infty}\mathbf{E}[h_n] \overset{\times}{<} K(\mu)\ln 2$, 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:
  $\mathbf{E}[h_n] \overset{\times}{<} \frac{1}{n}\ln w(\mu)^{-1}$ and $\mathbf{E}[h_n] \overset{\times}{<} \frac{1}{n}\ln w_\mu^{-1} = \frac{1}{n}K(\mu)\ln 2$.
- 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 $\bar{x}_n \neq x_n$: $2^{-K(n)} \overset{\times}{<} M(\bar{x}_n|x_{<n}) \overset{\times}{<} 2^{2K(x_{1:n}*)-K(n)}$
- Magic instance numbers: e.g. $M(0|1^n) \overset{\times}{=} 2^{-K(n)} \to 0$, 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: $\sum_{t=n+1}^{\infty}\mathbf{E}[h_t|\omega_{1:n}] \overset{+}{<} [K(\mu|\omega_{1:n})+K(n)]\ln 2$

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,
+ the D_n bound for continuous classes,
+ the problem of zero p(oste)rior & 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 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 ⇔ x can be easily (re)constructed from y ⇔ Kolmogorov complexity $K(x|y) := \min\{\ell(p) : U(p,y) = x\}$ is small

Examples:

1) x is very similar to itself ($K(x|x) \overset{+}{=} 0$)

2) A processed x is similar to x ($K(f(x)|x) \overset{+}{=} 0$ 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 ($K(\text{random}|y)$ =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 $K(x)\equiv K_U(x)\approx K_T(x)$ $\Longrightarrow$ 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 $x_1, ..., x_n$ 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)

Genomics & Phylogeny: Mammals

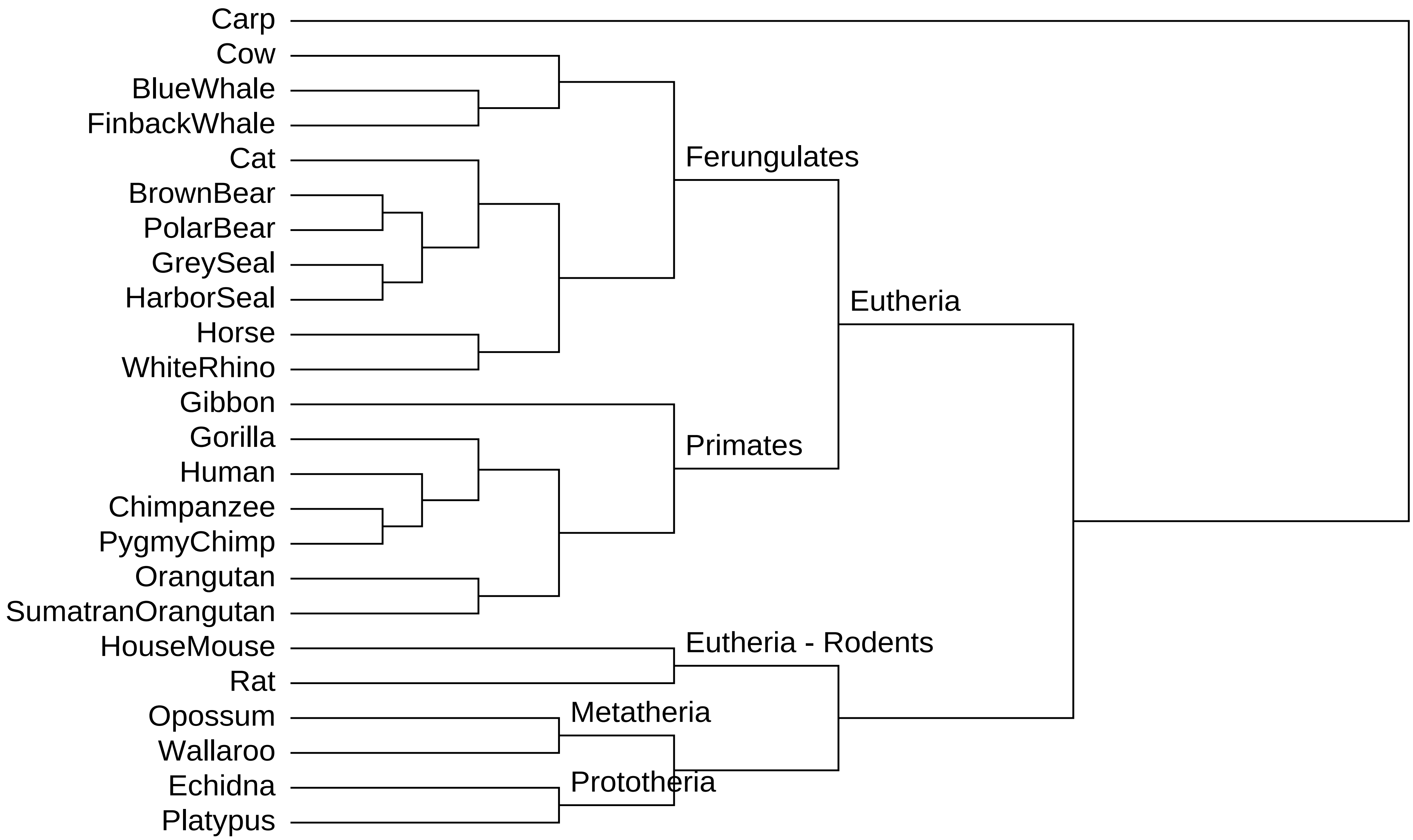
Let $x_1, ..., x_n$ be mitochondrial genome sequences of different mammals:

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

	BrownBear	Carp	Cat	Chimpanzee	Cow	Echidna	FinWhale	Gibbon	Gorilla	HouseMouse	Human	...
BrownBear	0.002	0.943	0.887	0.935	0.906	0.944	0.915	0.939	0.940	0.934	0.930	...
Carp	0.943	0.006	0.946	0.954	0.947	0.955	0.952	0.951	0.957	0.956	0.946	...
Cat	0.887	0.946	0.003	0.926	0.897	0.942	0.905	0.928	0.931	0.919	0.922	...
Chimpanzee	0.935	0.954	0.926	0.006	0.926	0.948	0.926	0.849	0.731	0.943	0.667	...
Cow	0.906	0.947	0.897	0.926	0.006	0.936	0.885	0.931	0.927	0.925	0.920	...
Echidna	0.944	0.955	0.942	0.948	0.936	0.005	0.936	0.947	0.947	0.941	0.939	...
FinbackWhale	0.915	0.952	0.905	0.926	0.885	0.936	0.005	0.930	0.931	0.933	0.922	...
Gibbon	0.939	0.951	0.928	0.849	0.931	0.947	0.930	0.005	0.859	0.948	0.844	...
Gorilla	0.940	0.957	0.931	0.731	0.927	0.947	0.931	0.859	0.006	0.944	0.737	...
HouseMouse	0.934	0.956	0.919	0.943	0.925	0.941	0.933	0.948	0.944	0.006	0.932	...
Human	0.930	0.946	0.922	0.667	0.920	0.939	0.922	0.844	0.737	0.932	0.005	...
...	...	...	...	...	...	...	...	...	...	...	...	...

Genomics & Phylogeny: Mammals

Evolutionary tree built from complete mammalian mtDNA of 24 species:

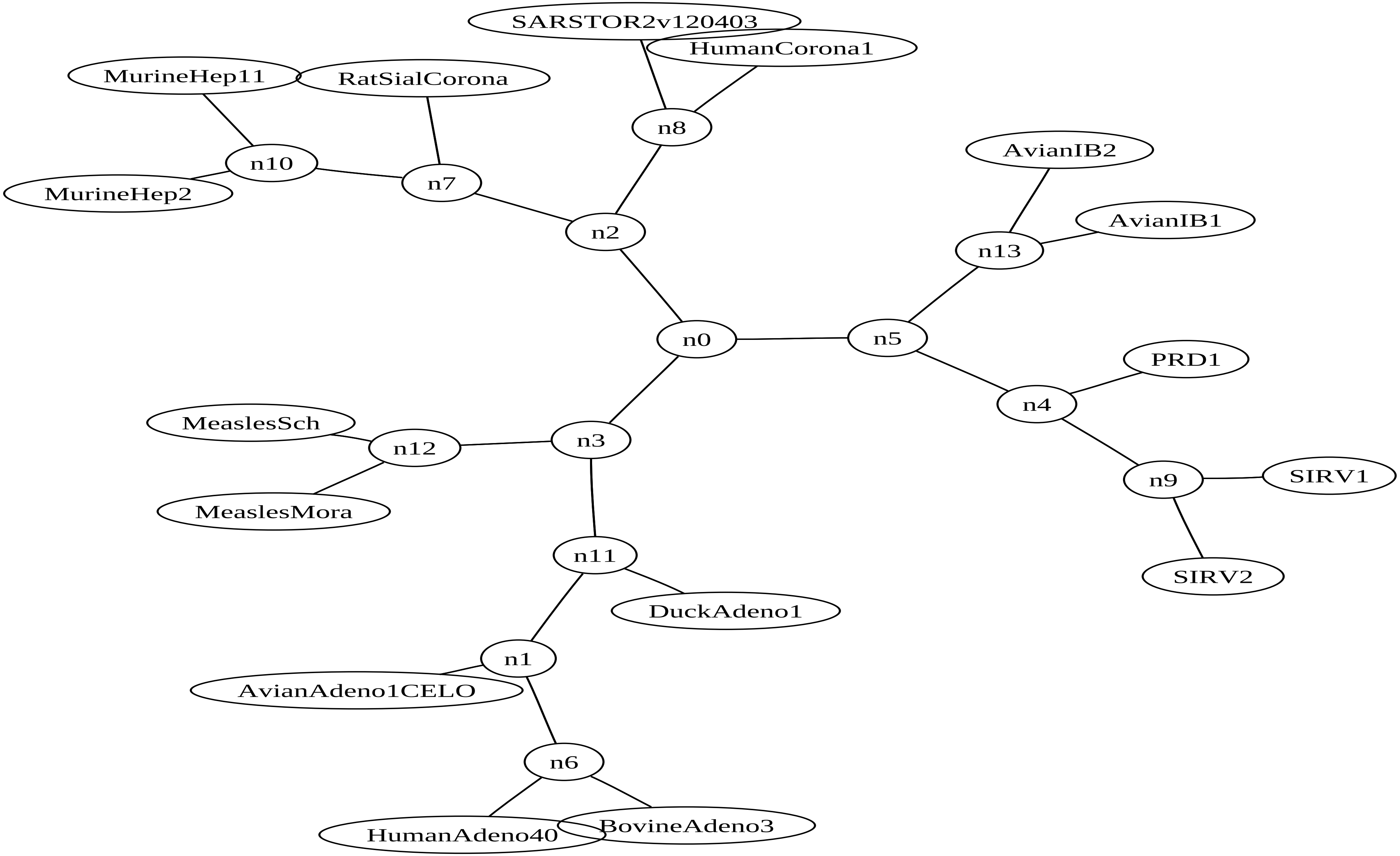


Genomics & Phylogeny: SARS Virus and Others

- 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



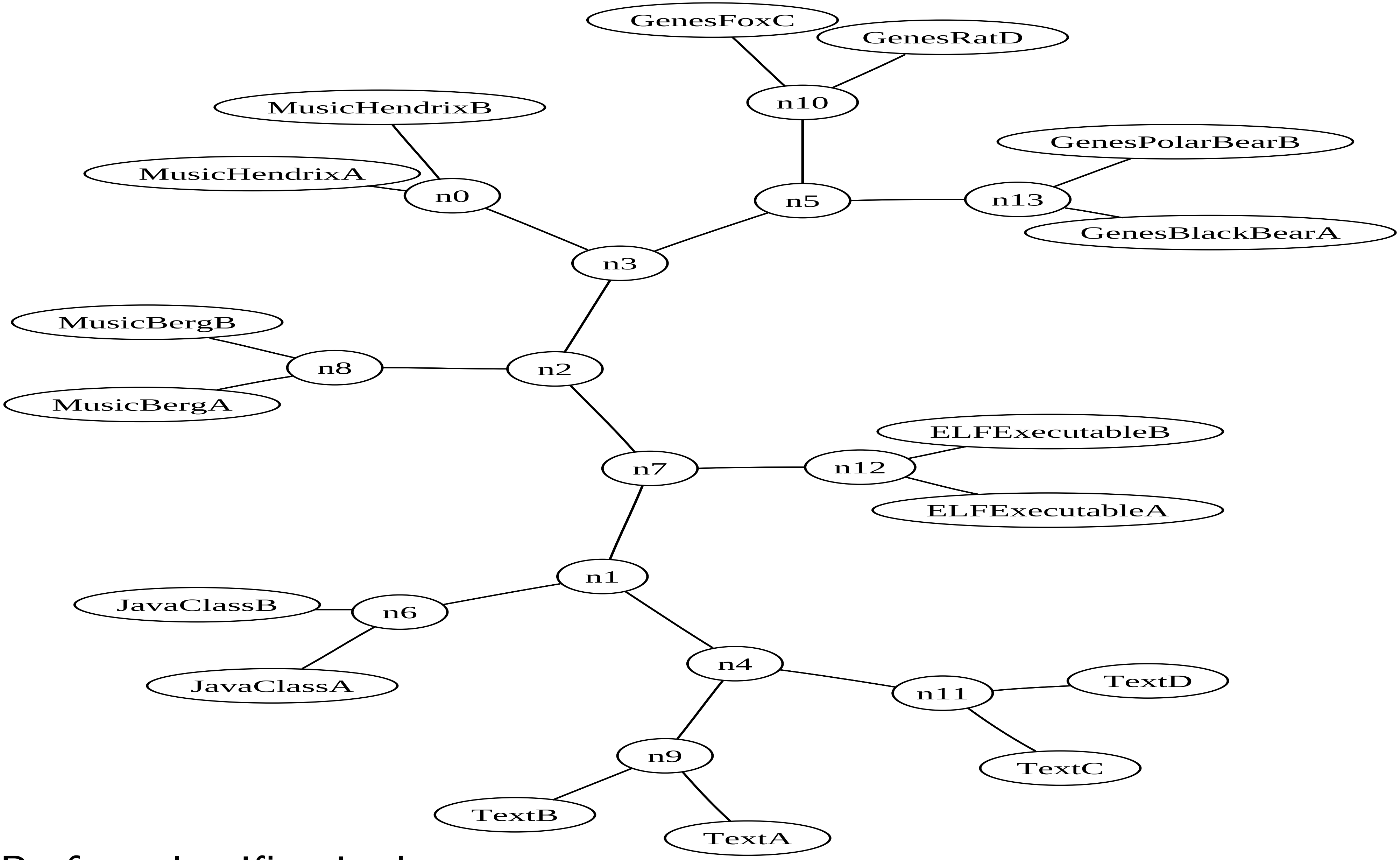
Classification of Different File Types

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



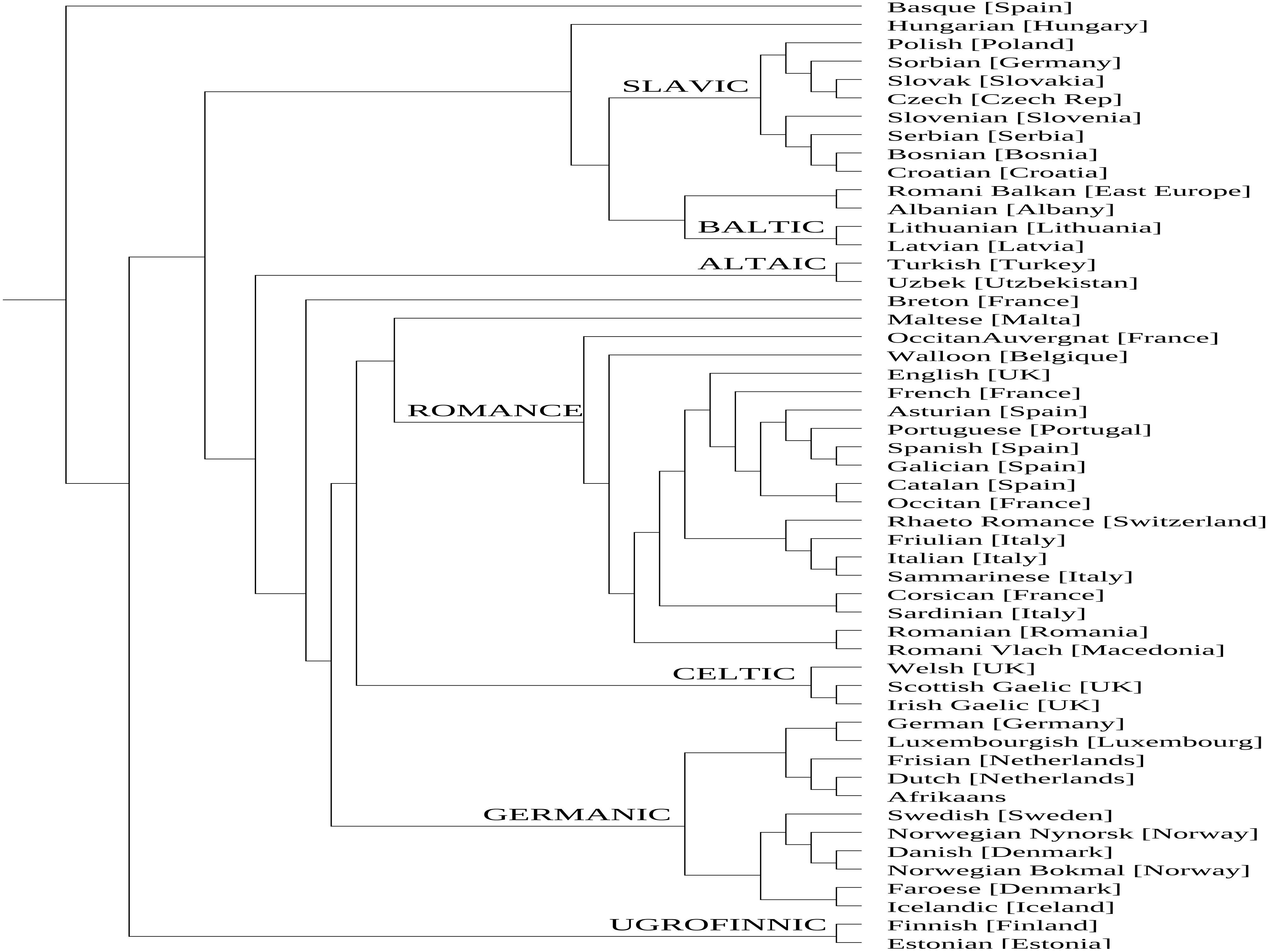
Perfect classification!

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)


Basque [Spain]
Hungarian [Hungary]
SLAVIC
Polish [Poland]
Sorbian [Germany]
Slovak [Slovakia]
Czech [Czech Rep]
Slovenian [Slovenia]
Serbian [Serbia]
Bosnian [Bosnia]
Croatian [Croatia]
Romani Balkan [East Europe]
Albanian [Albany]
BALTIC
Lithuanian [Lithuania]
Latvian [Latvia]
ALTAIC
Turkish [Turkey]
Uzbek [Utzbekistan]
Breton [France]
Maltese [Malta]
OccitanAuvergnat [France]
Walloon [Belgique]
ROMANCE
English [UK]
French [France]
Asturian [Spain]
Portuguese [Portugal]
Spanish [Spain]
Galician [Spain]
Catalan [Spain]
Occitan [France]
Rhaeto Romance [Switzerland]
Friulian [Italy]
Italian [Italy]
Sammarinese [Italy]
Corsican [France]
Sardinian [Italy]
Romanian [Romania]
Romani Vlach [Macedonia]
CELTIC
Welsh [UK]
Scottish Gaelic [UK]
Irish Gaelic [UK]
German [Germany]
Luxembourgish [Luxembourg]
Frisian [Netherlands]
Dutch [Netherlands]
Afrikaans
GERMANIC
Swedish [Sweden]
Norwegian Nynorsk [Norway]
Danish [Denmark]
Norwegian Bokmal [Norway]
Faroese [Denmark]
Icelandic [Iceland]
UGROFINNIC
Finnish [Finland]
Estonian [Estonia]

Classify Music w.r.t. Composer

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

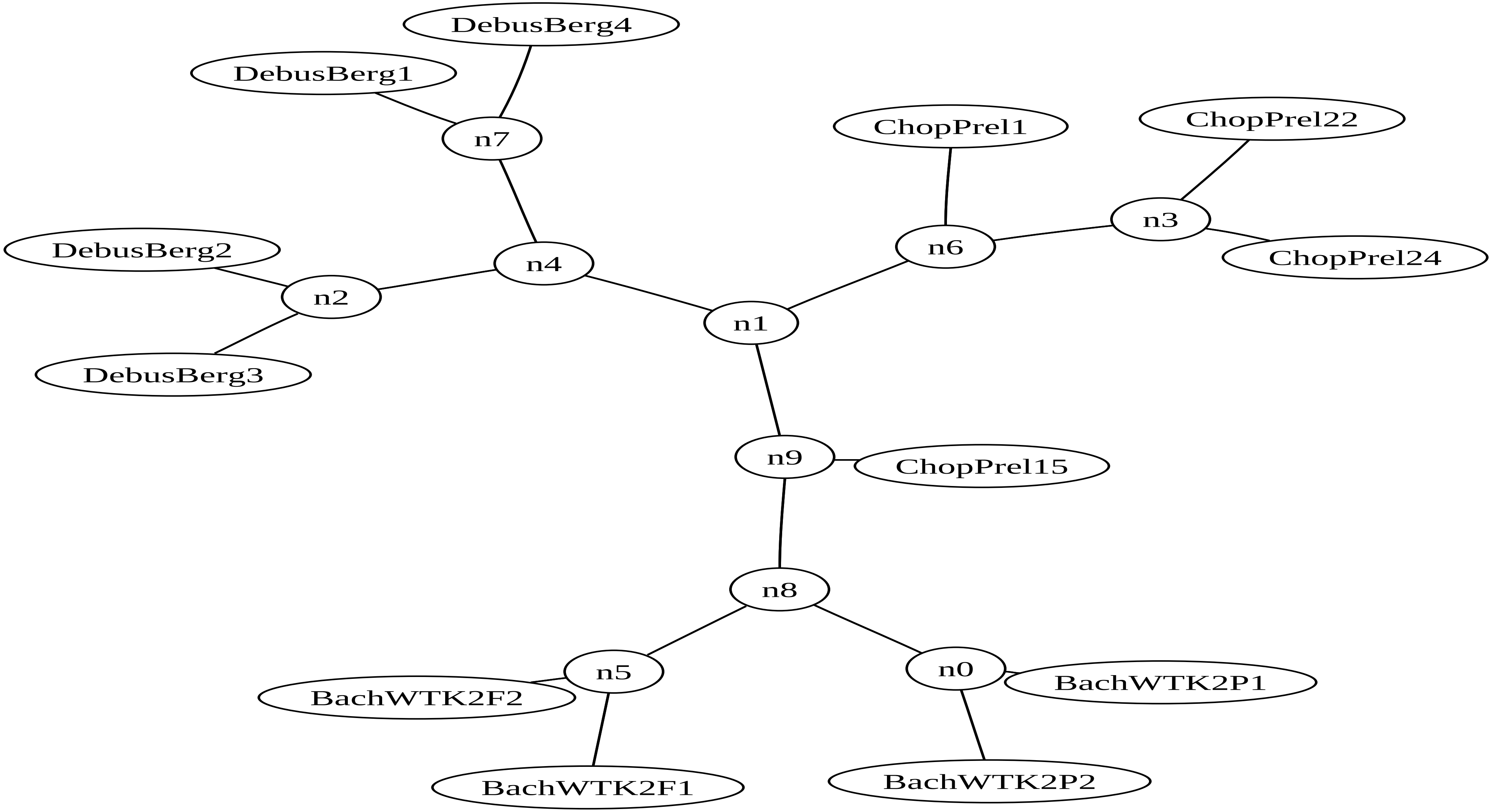
Preprocessing the MIDI files:

- 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 $x_1, ..., x_n$ still sounded like the original.

Classify Music w.r.t. Composer

12 pieces of music: 4×Bach + 4×Chopin + 4×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

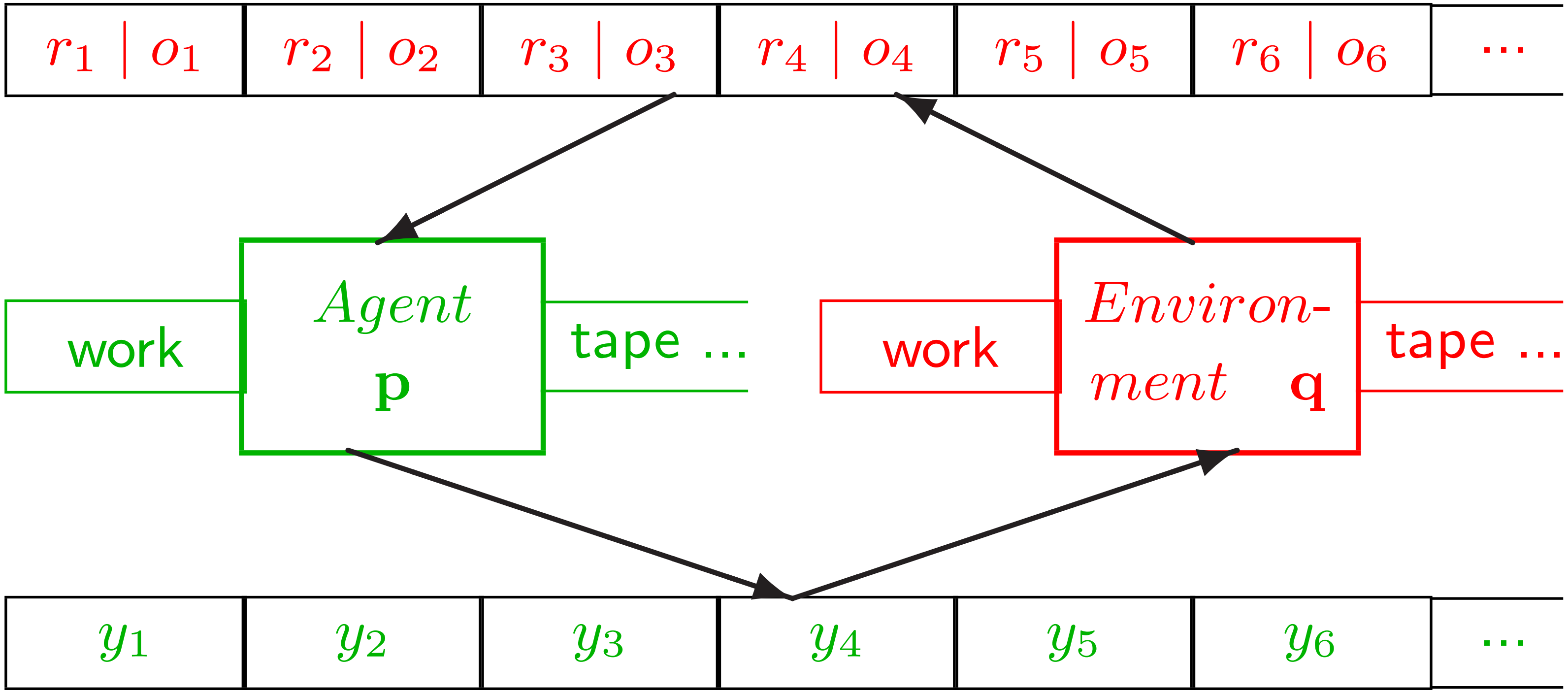
Universal Rational Agents: Abstract

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 Agent Model



Most if not all AI problems can be formulated within the agent framework

Rational Agents in Deterministic Environments

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

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

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

- Value $V_{km}^{pq} := r_k + ... + r_m$,
  optimal policy $p^{best} := \arg\max_p V_{1m}^{pq}$,
  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)

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

The average value of policy p with horizon m in environment σ is defined as

$$V_\sigma^p \;:=\; \frac{1}{m}\sum_{x_{1:m}}(r_1+\,...\,+r_m)\sigma(x_{1:m}|y_{1:m})_{|y_{1:m}=p(x_{<m})}$$

The goal of the agent should be to maximize the value.

Optimal Policy and Value

The σ-optimal policy $p^\sigma := \arg\max_p V_\sigma^p$ maximizes $V_\sigma^p \leq V_\sigma^* := V_\sigma^{p^\sigma}$.

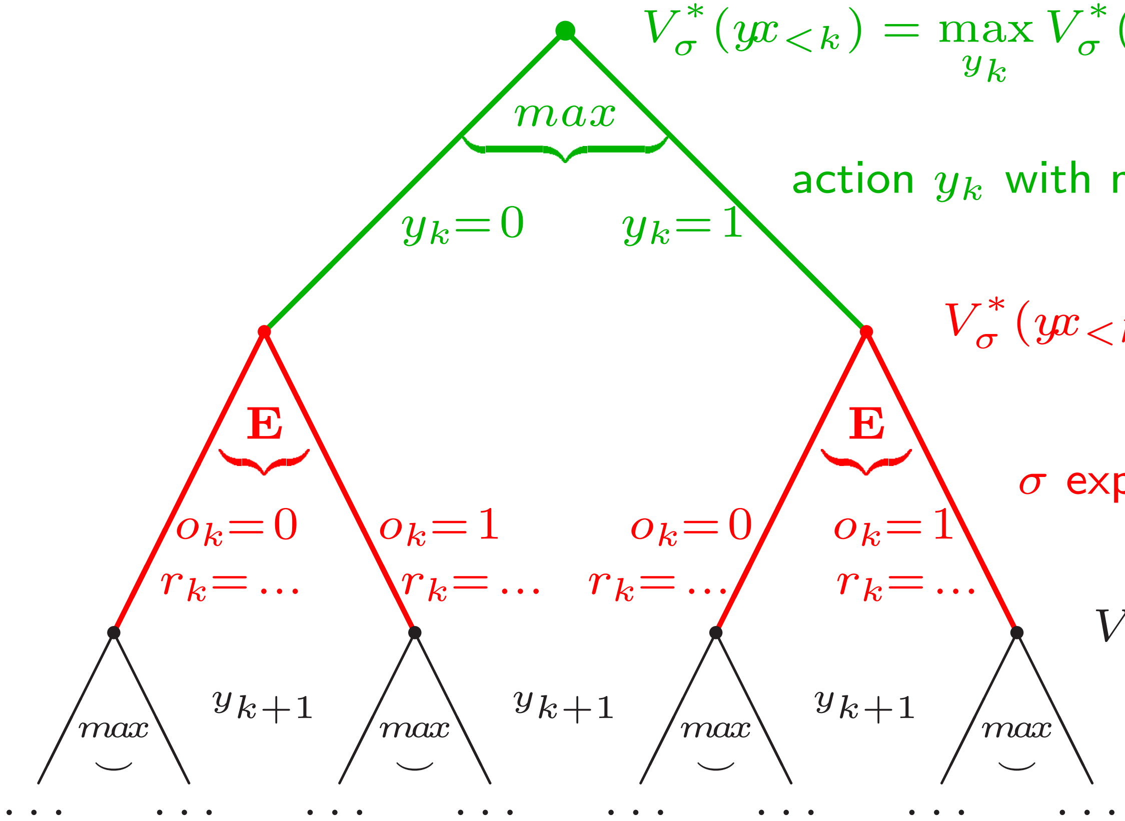
Explicit expressions for the action y_k in cycle k of the σ-optimal policy p^σ and their value V_σ^* are

$$y_k \;=\; \arg\max_{y_k}\sum_{x_k}\max_{y_{k+1}}\sum_{x_{k+1}}\;...\;\max_{y_m}\sum_{x_m}(r_k+...+r_m)\cdot\sigma(x_{k:m}|y_{1:m}x_{<k}),$$

$$V_\sigma^* \;=\; \frac{1}{m}\max_{y_1}\sum_{x_1}\max_{y_2}\sum_{x_2}\;...\;\max_{y_m}\sum_{x_m}(r_1+...+r_m)\cdot\sigma(x_{1:m}|y_{1:m}).$$

Keyword: Expectimax tree/algorithm.

Expectimax Tree/Algorithm



$$V^*_\sigma(yx_{<k}) = \max_{y_k} V^*_\sigma(yx_{<k}y_k)$$

action y_k with max value.

$$V^*_\sigma(yx_{<k}y_k) = \sum_{x_k}[r_k + V^*_\sigma(yx_{1:k})]\sigma(x_k|yx_{<k}y_k)$$

σ 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 μ

- Assumption: μ is the true environment in which the agent operates

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

- Special choices of μ: 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 μ^{AI} is known and computable and $\mathcal{X}$, $\mathcal{Y}$ and m are finite.

- Things drastically change if μ^{AI} is unknown ...

The Bayes-mixture distribution ξ

Assumption: The true environment μ is unknown.

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

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

The Bayes-mixture ξ 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_ν may be interpreted as the prior degree of belief that the true environment is ν.

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^ξ which maximizes V_ξ^p.
- If μ is the true environment the expected reward when following policy p^ξ will be $V_\mu^{p^\xi}$.
- The optimal (but infeasible) policy p^μ 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_μ^*?
- What is the most general class $\mathcal{M}$ and weights w_ν.

A universal choice of ξ 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 ν one should set $w_\nu \sim 2^{-K(\nu)}$, where $K(\nu)$ is the length of the shortest program on a universal TM computing ν.
- 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

$$\boxed{y_k \;=\; \arg\max_{y_k}\sum_{x_k}\ldots\max_{y_m}\sum_{x_m}[r(x_k)+\ldots+r(x_m)]\sum_{q\,:\,U(q,y_{1:m})=x_{1:m}} 2^{-\ell(q)}}$$

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_ν.

Pareto-Optimality of p^ξ

Policy p^ξ 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 ν.

Self-optimizing Policies

Under which circumstances does the value of the universal policy p^ξ converge to optimum?

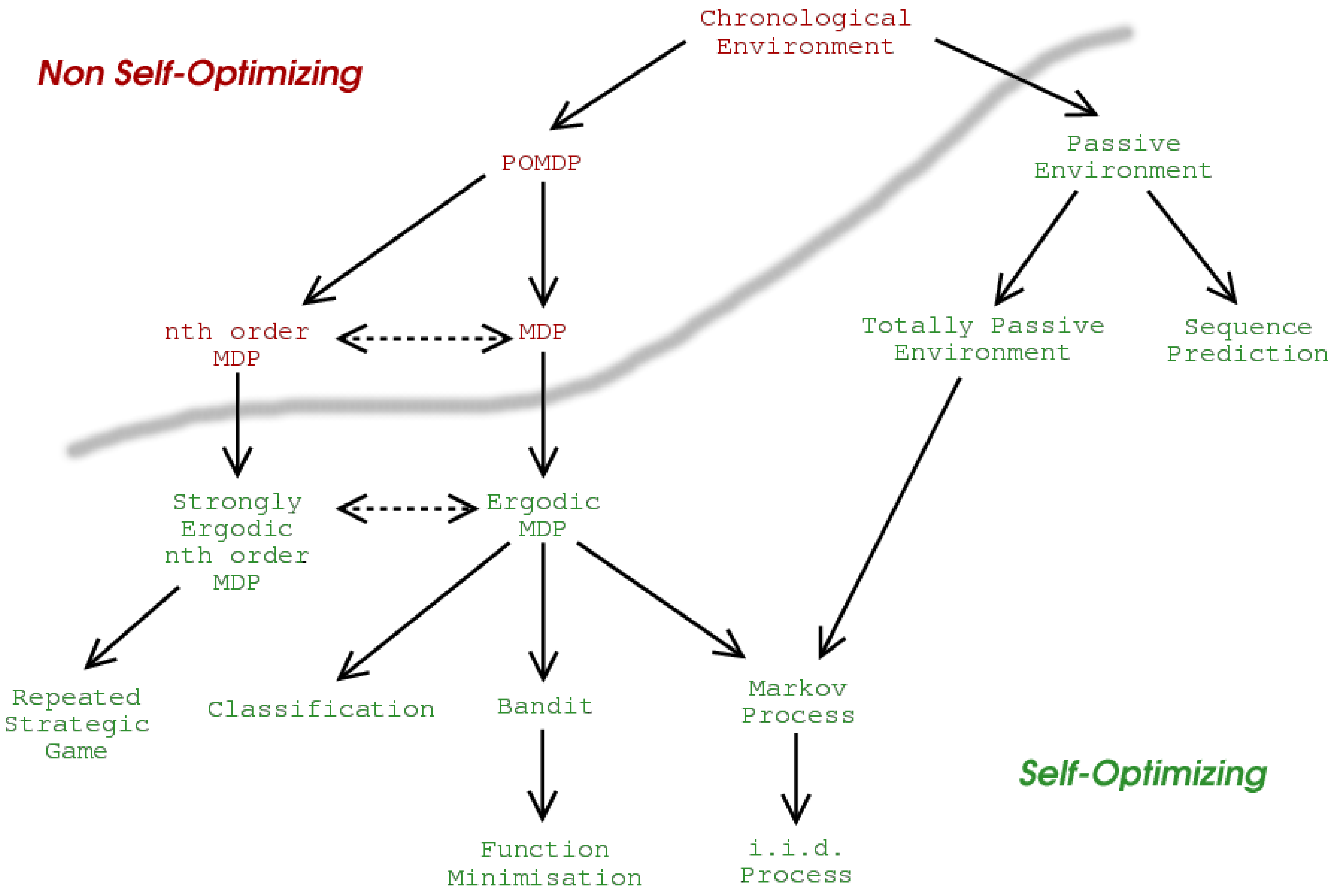
$$V_\nu^{p^\xi} \to V_\nu^* \quad \text{for horizon} \quad m \to \infty \quad \text{for all} \quad \nu \in \mathcal{M}. \qquad (1)$$

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.

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

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

Environments w. (Non)Self-Optimizing Policies



Particularly Interesting Environments

- Sequence Prediction, e.g. weather or stock-market prediction. Strong result: $V_\mu^* - V_\mu^{p^\xi} = O(\sqrt{\frac{K(\mu)}{m}})$, $m =$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.

AIξ quickly learns to predict, play games, optimize, and learn supervised.

Universal Rational Agents: Summary

- Setup: Agents acting in general probabilistic environments with reinforcement feedback.
- Assumptions: Unknown true environment μ belongs to a known class of environments $\mathcal{M}$.
- Results: The Bayes-optimal policy p^ξ 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.
- We have reduced the AI problem to pure computational questions (which are addressed in the time-bounded AIXItl).
- AIξ incorporates all aspects of intelligence (apart comp.-time).
- How to choose horizon: use future value and universal discounting.
- ToDo: prove (optimality) properties, scale down, implement.

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 $p(H_i|D) \propto p(D|H_i)p(H_i)$
- Bayes needs prior(H_i)
- Occam+Epicurus: High prior for simple models.
- Kolmogorov/Solomonoff: Quantification of simplicity/complexity
- Bayes works if D is sampled from $H_{true} \in \mathcal{M}$
- Universal AI $=$ Universal Induction $+$ Sequential Decision Theory

Literature

[CV05] R. Cilibrasi and P. M. B. Vitányi. Clustering by compression. *IEEE Trans. Information Theory*, 51(4):1523–1545, 2005.
http://arXiv.org/abs/cs/0312044

[Hut05] M. Hutter. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*. Springer, Berlin, 2005.
http://www.hutter1.net/ai/uaibook.htm.

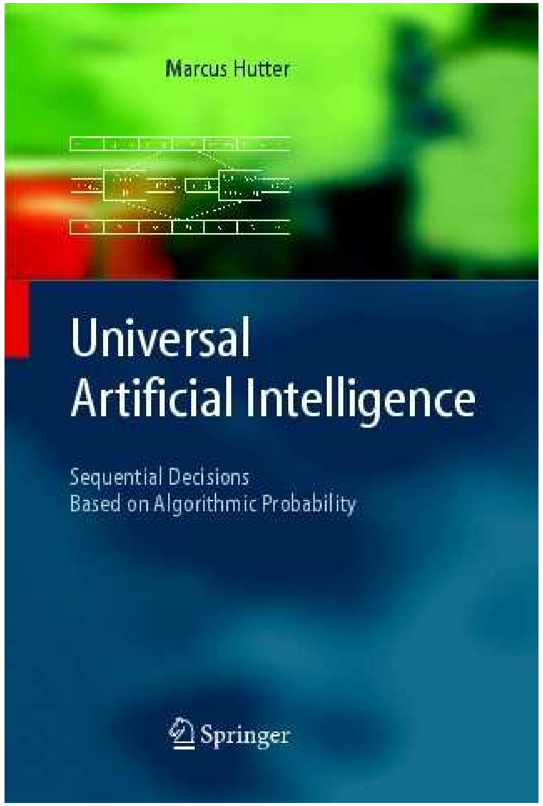
[Hut07] M. Hutter. On universal prediction and Bayesian confirmation. *Theoretical Computer Science*, 384(1):33–48, 2007.
http://arxiv.org/abs/0709.1516

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


Marcus Hutter
Universal Artificial Intelligence
Sequential Decisions Based on Algorithmic Probability
Springer

A Unified View of Artificial Intelligence

$$\begin{array}{ccc} \| & & \| \\ \text{Decision Theory} & = & \text{Probability} + \text{Utility Theory} \\ + & & + \\ \text{Universal Induction} & = & \text{Ockham} + \text{Bayes} + \text{Turing} \end{array}$$

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

Compression competition with 50'000 Euro prize at prize.hutter1.net