An Introduction to Induction from an AIT View*

Xi Li

CLLC @ PKU
Beijing, 100871, China
xli@pku.edu.cn

1/4/2012

Contents

*This talk is based on Rathmanner and Hutter's paper:A Philosophical Treatise of Universal Induction 2011

Abstract

Understanding inductive reasoning is a problem that has engaged mankind for thousands of years. It has been tackled by many great minds ranging from Occam to Epicurus to Bacon to Hume to Mill to Bayes to Laplace to Keynes to Popper to Carnap to Solomonoff. In this article we analyse the history and philosophical foundations of Solomonoff Induction, which is a formal inductive framework which combines algorithmic information theory(AIT) with the Bayesian framework. After that we introduce Hutter's general reinforcement learning agents—AIXI based on Solomonoff's work. At last, we introduce some approximate approaches and applications.

Keywords

Universal Induction; Sequence Prediction; ML; ME; MDL; AIXI; Bayes rule; Occam's razor; Epicurus; Black raven paradox.

"Make everything as simple as possible, but not simpler"

— Einstein

1 Introduction

"As far as the laws of mathematics refer to reality, they are not certain, and as far as they are certain, they do not refer to reality."

— Einstein

Generally speaking, incomplete induction can only draw uncertain conclusions, while complete induction is both theoretically reducible to deduction and practically impossible. Just as the consistency of the deductive system can't be proved by itself but to resort to deduction reluctantly, the legality of deduction can't be justified circularly but to turn to deduction for help although seeming impossible.

Induction versus Deduction

	Induction	⇔	**Deduction**
Type of inference	generalization/prediction	⇔	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	$\widehat{=}$	ZFC
Universal inference	universal induction	$\widehat{=}$	universal theorem prover
Limitation	incomputable	$\widehat{=}$	imcomplete
In practice	approximions	$\widehat{=}$	semi-formal proofs
Operation	computation	$\widehat{=}$	proof

1.1 History

Occam insisted that "Entities should not be multiplied beyond necessity", while Epicurus thought that if more than one theory is consistent with the observations, we should keep them all.

Bacon thought that incomplete induction is not only possible, but can even reach certain conclusions, if pursued by means of an axiomatic analysis of the essence of the available data, through a classification by means of tables listing presence or absence of relevant features.

According to Hume, induction is just a mental habit, and necessity is something in the mind and not in the events, one can never demonstrate the necessity of a cause.

Leibniz raised the following curve-fitting paradox: since for any finite number of points there are always infinitely many curves going through them, any finite set of data is compatible with infinitely many inductive generalizations. The paradox raised the problem of which inductive generalization should be chosen, among the many available ones fitting the data. Wittgenstein had similar ideas in the following form: since any finite course of action is in accord with infinitely many rules, no universal rule can be learned by examples. Goodman made it more against our intuition: past observation on all emeralds

so far discovered supports conclusions giving rise to opposite predictions, such as ‘all emeralds are green’ or ‘all emeralds discovered till 2050 are green, and blue thereafter’.

Mill attempted to turn induction into a deduction, by adding principles about the world (such as ’the future resembles the past’, or ’space-time is homogeneous’) that would make inductive conclusions follow logically from their premises.

Keynes tried to assign to inductive generalizations, according to available evidence, probabilities that should converge to 1 as the generalizations are supported by more and more independent events.

Popper thought that single observational events may prove hypotheses wrong, but no finite sequence of events can verify them correct. Thus induction is theoretically unjustifiable and becomes in practice the choice of the simplest generalization that resists falsification, on the ground that the simpler a hypothesis is the easier it is to falsify it. But falsifiability is as subjective as simplicity. And sometimes a complex theory with fixed parameters is as easy to falsify as a simple theory.

Carnap weakened the verification theory into the following confirmation theory: observational events provide, if not proofs, at least positive confirmations of scientific hypotheses. We can evaluate the probability of a hypothesis, under certain assumptions. Thus induction becomes a choice of a generalization among the ones that confirm more evidence.

But how many evidences are strong enough to hold our belief?

The theory of confirmation is subject to peculiar paradoxes, which appear when it is mixed with the usual logic of deduction:

- $\because (\alpha \to \beta) \to (\beta \text{ confirms } \alpha)$ $\quad \{ \quad \alpha \text{ confirms } \bot \quad \}$
  $\because (\alpha \to \beta) \wedge (\gamma \text{ confirms } \alpha)) \to (\gamma \text{ confirms } \beta)$
  $\because (\alpha \wedge \beta) \to \alpha$
  $\therefore \alpha \text{ confirms } \alpha \wedge \beta$
  $\because (\alpha \wedge \beta) \to \beta$
  $\therefore \alpha \text{ confirms } \beta$

- $\because A(x) \wedge B(x) \text{ confirms } \forall x(A(x) \to B(x))$
  $\because \neg A(x) \wedge \neg B(x) \text{ confirms } \forall x(\neg B(x) \to \neg A(x))$
  $\because \forall x(A(x) \to B(x)) \leftrightarrow \forall x(\neg B(x) \to \neg A(x))$
  $\therefore \neg A(x) \wedge \neg B(x) \text{ confirms } \forall x(A(x) \to B(x))$

1.2 Logic versus Probability

On one hand, one can view Bayesian probability as a special case of classical logic by taking a frequentist interpretation. In this interpretation, one views the actual universe as just one of a large number of possible universes. Each universe is in one of the possible states; the probability assigned to each state is then the proportion of the possible universes in which that state is attained. Each new measurement eliminates some fraction of the universes in a given state, depending on how likely or unlikely that state was to actually produce that measurement; the surviving universes then have a new posterior probability

distribution, which is related to the prior distribution by Bayes' formula. For example, Carnap's approach: If the predicates $(Q_1, \cdots, Q_k)$ are defined so that they have different relative widths q_i, such that $q_1 + \cdots + q_k = 1$, then $c(Q_i(a_{n+1}|a_{1:n})) = \frac{n_i + \lambda q_i}{n+\lambda}$. But the problem is how to assign the initial prior distribution?

On the other hand, we can view Bayesian probability as a quantitative refinement of classical logic.

$\vdash A \to B$	$P(B\|A) = 1$
$\vdash A \to \neg B$	$P(B\|A) = 0$
$\nvdash (A \to B) \to (B \to A)$	$P(B\|A) = 1 \nRightarrow P(A\|B) = 1$
$\nvdash (A \to B) \to (B \to A)$	$P(B\|A) > P(B) \Rightarrow P(A\|B) > P(A)$
$\vdash (A \to B) \to (\neg B \to \neg A)$	$P(\neg A) \geq 1 - \frac{1 - P(\neg B)}{P(B\|A)}$
$\vdash (A \to B) \wedge A \to B$	$P(B) \geq P(B\|A)P(A)$
$\vdash (A \to B) \to (B \to C) \to (A \to C)$	$P(C\|A) \geq P(C\|B)P(B\|A)$
$\phi(0) \wedge \forall n(\phi(n) \to \phi(n+1)) \to \forall n \phi(n)$	$P(A_n) \geq \prod_{i=1}^{n-1} P(A_{i+1}\|A_i)$
$\phi(0) \wedge \forall n(\phi(n) \to \phi(n+1)) \to \forall n \phi(n)$	$\forall i(P(A_{i+1}\|A_i) < 1) \Rightarrow \lim_{n\to\infty} \prod_{i=1}^{n-1} P(A_{i+1}\|A_i) = 0$
$\phi(0) \wedge \forall n(\phi(n) \to \phi(n+1)) \to \forall n \phi(n)$	$\forall i(P(X = X_i) \geq c) \Rightarrow$ $P(X = X_1 = \cdots = X_n \| X_1 = \cdots = X_n) \geq \frac{c^n}{c^n + (1-c)^n}$

If we consider the predicate "confirm" as conditional probability like Carnap $c(h, e) = \frac{m(h\&e)}{m(e)}$ rather than as some kind of "implication" as in the raven paradox, then the paradox disappear. For example,
H:All ravens are black.
$\bar{H}$: All non-black objects are non-ravens.
H': Half the ravens are black.
D:A randomly selected raven is black.
A:A randomly selected non-black object is non-raven.

$$P(D|H) = P(A|H) = 1$$
$$P(D|H') = \frac{1}{2}$$
$$P(A|H') \approx 1 \text{ but } P(A|H') < 1$$
$$P(D) \neq P(A) \Rightarrow P(H|D) = \frac{P(D|H)P(H)}{P(D)} \neq \frac{P(A|\bar{H})P(\bar{H})}{P(A)} = P(\bar{H}|A)$$
$$\frac{P(H|A)}{P(H'|A)} = \frac{P(A|H)P(H)}{P(A|H')P(H')} \approx \frac{P(H)}{P(H')}$$
$$\frac{P(H|D)}{P(H'|D)} = \frac{P(D|H)P(H)}{P(D|H')P(H')} = 2\frac{P(H)}{P(H')}$$

It is the ratio $\frac{P(D|H)}{P(D|H')}$ determines the strength of the evidence: a strong piece of evidence needs to be plausible under hypothesis H, while simultaneously being implausible under rival hypotheses.
This corresponds with the *maximum likelihood principle*(ML) for hypothesis testing.

$$\widehat{H} = \arg\max_H P(D|H)$$

We use ML rather than

$$\widehat{H} = \arg_H[P(D|H) \approx 1]$$

1.3 Subjective Probability

The Dutch book argument shows that if an agent's beliefs are inconsistent (contradict the axioms) then a set of bets can be formulated which the agent finds favorable according to its beliefs but which guarantees that it will lose. Cox theorem gaves a formal rigorous justification that a rational belief system must obey the standard probability axioms.

The Cox's axioms for beliefs are as follows:

- The degree of belief in an event B, given that event A has occurred can be characterized by a real-valued function $Bel(B|A)$.
- $Bel(\Omega \setminus B|A)$ is a twice differentiable function of $Bel(B|A)$ for $A \neq \emptyset$.
- $Bel(B \cap C|A)$ is a twice differentiable function of $Bel(C|B \cap A)$ and $Bel(B|A)$ for $B \cap A \neq \emptyset$.

$\rho : \mathcal{X}^* \to [0,1]$ is a semimeasure if $\rho(x) \geq \sum_{|a|=1} \rho(xa) \, \forall x \in \mathcal{X}^*$, and a (probability) measure if equality holds and $\rho(\epsilon) = 1$. We assume measures, hypotheses, models and environments express the same concept. The class of all possible considered environments is denoted $\mathcal{M}$. Assume that $w_\nu := P[H_\nu]$ is the given prior belief in H_ν. Assume that sequence $\omega = \omega_{1:\infty} \in \mathcal{X}^\infty$ is sampled from the "true" probability measure $\mu \in \mathcal{M}$, i.e. $\mu(x_{1:n}) := H_\mu(x_{1:n}) := P[\omega_{1:n} = x_{1:n}|H_\mu]$ is the μ-probability that ω starts with $x_{1:n}$. I denote expectations w.r.t. μ by $\mathbb{E}$, for $f : \mathcal{X}^n \to \mathbb{R}$, $\mathbb{E}[f] := \sum_{x_{1:n} \in \mathcal{X}^n} \mu(x_{1:n}) f(x_{1:n})$.

Bayes Theorem Let H_ν be a hypothesis from class $\mathcal{M}$ which is required to be mutually exclusive and complete. Bayes mixture $\xi(x) := P[x] = \sum_{\nu \in \mathcal{M}} P[x|H_\nu]P[H_\nu]$ must be our (prior) belief in observational data x, and Bayes formula below can be understood as our posterior belief in ν given data x.

$$w_\nu(x) := P(H_\nu|x) \;=\; \frac{P(x|H_\nu)P(H_\nu)}{P(x)} \;=\; \frac{P(x|H_\nu)P(H_\nu)}{\sum_{H_i \in \mathcal{M}} P(x|H_i)P(H_i)}$$

A Bayesian is a subjectivist, believing that our beliefs and hence probabilities are a result of our personal history. To be able to update beliefs consistently a Bayesian must first decide on the set of all explanations that may be possible. For universal induction, we are interested in finding the true governing process behind our entire reality and to do this we consider all possible worlds in a certain sense. No matter what the problem is we

can always consider it to consist of an agent in some unknown environment. The agent must have some prior belief in these explanations before the updating process begins, in other words, before any observations have been made.

This understanding of probability can be troubling as it suggests that we can never be certain of any truth about reality, however this corresponds exactly with the philosophy of science. In science it is not possible to ever prove a hypothesis, it is only possible to disprove it. What are often stated as physical laws are actually only strongly believed and heavily tested hypotheses.

Surprise/Uncertainty/Ignorance/Entropy Surprise springs from Ignorance. If we believe that it is probable that something will happen, we will not very surprise when it happens.

If we assume the surprise function $S(P)$ satisfy the following axioms, then $S(P) = -C \log_2 P$.

- $S(1) = 0$
- $P < Q \Rightarrow S(P) > S(Q)$
- S is a continuous function of P
- $S(PQ) = S(P) + S(Q) \quad 0 < P \leq 1, 0 < Q \leq 1$

Let X be a random variable taking values in $\mathcal{X}^*$ with distribution P. We assume $C = 1$, the expected surprise of a random variable X is $H(X) = -\sum_{x \in \mathcal{X}^*} P(x) \log_2 P(x)$, i.e. the Shannon entropy of X.

1.4 Induction versus Prediction

Inductive inference [Odi99] can be described as a particular step from chaos to order, or from effects to causes. In such a process we can isolate three distinct aspects:

- data

  They provide bits of information about the phenomena to be inferred, at given instants of time. According to whether the order in which they occur is considered to be relevant or not, the flow of data can be taken to be a function (or a set). For example, the evolution of a physical system in time can be identified with a function, while a language can be identified with a set of strings.

- goals of inference

  A theoretical goal is to be able to explain the phenomena to which the data refer, while a practical goal is to be able to predict the flow of data from a certain point on. In the first case, one is interested in understanding causes, in the second in reproducing effects.

- time needed for inference

  An inference must be completed in a finite time, which could be specified in advance, or at least one should know that the process will be completed even if sometimes he can't figure out in advance when.

As a approximation to an inference process of the kind described above, we picture time as consisting of discrete intervals, and events as being codifiable by natural numbers. Thus a phenomenon to be inferred may be thought of as a function on the natural numbers. Any such function f is given by a sequence of values

$$f(0), \cdots, f(n), \cdots$$

The function can be inferred if this is not just a sequence of accidents, but rather it has an intrinsic necessity. We can specify this internal structure of the sequence of values in at least two ways, corresponding to the two goals described above:

- On the one hand, we can ask for a finite description that would compress the finite amount of information contained in the sequence of values.

  Find a total recursive function g such that,

  $$\phi_{\lim_{n\to\infty} g(\langle f(0),\cdots,f(n)\rangle)} \simeq f$$

  or for almost every n,

  $$\phi_{g(\langle f(0),\cdots,f(n)\rangle)} \simeq f$$

- On the other hand, we can ask for a method that would allow us to predict the next value $f(n+1)$, once the values $f(0), \cdots, f(n)$ have been exhibited, for an arbitrary n.

  Find a total recursive function g such that, for almost every n,

  $$f(n+1) = g(\langle f(0), \cdots, f(n)\rangle)$$

Induction can be understood to include the process of drawing conclusions about some given data, or as the process of predicting the future. Can general induction problems be rephrased as prediction problems?

Regression is the problem of finding the function that is responsible for generating some given data points, often accounting for some noise or imprecision. The data is a set of (feature,value) tuples $\{(x_1, f(x_1)), (x_2, f(x_2)),....(x_n, f(x_n))\}$. In machine learning this problem is often tackled by constructing a function that is the 'best' estimate of the true function according to the data seen so far. Alternatively, it can be formalized directly in terms of sequential prediction by writing the input data as a sequence and appending it with a new point x_{n+1} for which we want to find the functional value. In other words the problem becomes: "What is the next value in the sequence $x_1, f(x_1), x_2, f(x_2), ...x_n, f(x_n), x_{n+1}, ?$". Although this approach does not produce the function explicitly, it is essentially equivalent, since $f(x)$ for any x can be obtained by choosing $x_{n+1} = x$.

1.5 Kolmogorov Complexity

Notation I write $\mathcal{X}^*$ for the set of finite strings over some alphabet $\mathcal{X}$, and $\mathcal{X}^\infty$ for the set of infinite sequences. For a string $x \in \mathcal{X}^*$ of length $|x| = n$ I write $x_1x_2...x_n$ with $x_t \in \mathcal{X}$, and further abbreviate $x_{t:n} := x_t x_{t+1}...x_{n-1}x_n$ and $x_{<n} := x_1...x_{n-1}$. The notation generalises for blocks of symbols: e.g. $ax_{1:n}$ denotes $a_1x_1a_2x_2\dots a_nx_n$ and $ax_{<j}$ denotes $a_1x_1a_2x_2\dots a_{j-1}x_{j-1}$. The empty string is denoted by ϵ. The concatenation of two strings s and r is denoted by sr.

A function $f : \mathcal{X}^* \to \mathbb{R}\cup\{\pm\infty\}$ is said to be lower semi-computable (or enumerable) if the set $\{(x, y) \in \mathcal{X}^* \times \mathbb{Q} : y < f(x)\}$ is recursively enumerable. f is upper semi-computable (or co-enumerable) if $-f$ is enumerable. f is computable (or recursive) if f and $-f$ are enumerable. The set of (co)enumerable functions is recursively enumerable.

I write $f(x) \stackrel{+}{\leq} g(x)$ for $f(x) \leq g(x) + O(1)$ and $f(x) \stackrel{\times}{\leq} g(x)$ for $f(x) \leq O(g(x))$.

We say that a property $A(n)$ holds for *most* n, if $\lim_{n\to\infty} \frac{|\{t\leq n : A(t)\}|}{n} = 1$.

Definition 1 (Kolmogorov complexity)**.**

$$K(x) \;:=\; \min\{|p| : U(p) = x\} \tag{1}$$

Where U is a universal prefix Turing Machine.

It has the following properties:

K is not computable, but only co-enumerable; (K1)

$K(n) \stackrel{+}{\leq} \log_2 n + 2\log_2\log_2 n$; (K2)

$$\sum_x 2^{-K(x)} \leq 1; \tag{K3}$$

$K(f(x)) \stackrel{+}{\leq} K(x) + K(f)$ for recursive $f : \mathcal{X}^* \to \mathcal{X}^*$; (K4)

$K(x) \stackrel{+}{\leq} -\log_2 P(x) + K(P)$ if $P : \mathcal{X}^* \to [0, 1]$ is enumerable and $\sum_x P(x) \leq 1$; (K5)

$$\sum_{x:f(x)=y} 2^{-K(x)} \stackrel{\times}{=} 2^{-K(y)} \quad \text{if } f \text{ is recursive.} \tag{K6}$$

Shannon entropy equals the expected value of Kolmogorov complexity, up to a constant term that only depends on the distribution P.

Theorem 1. *For a computable probability distribution P*

$$0 \leq \sum_x P(x)K(x) + \sum_x P(x)\log_2 P(x) \stackrel{+}{\leq} K(P)$$

Proof. The first " $\leq$ " follows directly from Shannon's Noiseless Coding Theorem.

$$\begin{aligned} &\sum_x P(x)K(x) + \sum_x P(x)\log_2 P(x) \\ &= \sum_x P(x)\left(K(x) + \log_2 P(x)\right) \qquad \text{[Equation (K5)]} \end{aligned}$$

$$\stackrel{+}{\leq} \sum_x P(x)K(P)$$
$$\leq K(P)$$

□

2 Bayesianism for Prediction

"The most incomprehensible thing about the world is that it is comprehensible."

— Einstein

2.1 Convergence Results

Deterministic For a deterministic environment it is sufficient to know the unique observation sequence α that must be generated, since $\mu(\alpha_{1:n}) = 1$ for all n, and $\mu(x) = 0$ for any x that is not a prefix of α. In this case we identify μ with α and the following holds:

Theorem 2.

$$\sum_{t=1}^{\infty} |1-\xi(\alpha_t|\alpha_{<t})| \;\leq\; \ln w_\mu^{-1} \;\; \textit{and} \;\; \xi(\alpha_{t:n}|\alpha_t) \to 1 \;\; \textit{for} \;\; n \geq t \to \infty \tag{2}$$

Proof.

$$\because \xi(\alpha_{1:n}) = \Sigma_{\nu\in\mathcal{M}} w_\nu \nu(\alpha_{1:n}) \geq w_\mu \mu(\alpha_{1:n}) = w_\mu > 0$$
$$\because \sum_{t=1}^{n} |1 - \xi(x_t|x_{<t})| \leq -\sum_{t=1}^{n} \ln \xi(x_t|x_{<t}) = -\ln \xi(x_{1:n})$$
$$\therefore \sum_{t=1}^{\infty} |1-\xi(\alpha_t|\alpha_{<t})| \;\leq\; \ln w_\mu^{-1}$$

□

Non-deterministic

Lemma 1. *For any discrete probability measure μ and discrete semimeasure ξ over $\mathcal{X}^*$*

$$\sum_{|a|=1} (\sqrt{\mu(a)} - \sqrt{\xi(a)})^2 \leq \sum_{|a|=1} \mu(a) \ln \frac{\mu(a)}{\xi(a)}$$

Proof.

$$f(x,y) := x \ln \frac{x}{y} - (\sqrt{x} - \sqrt{y})^2 + y - x$$

$$\therefore \frac{f(x,y)}{2x} = -\ln\sqrt{\frac{y}{x}} + \sqrt{\frac{y}{x}} - 1$$
$$\because \ln x \le x - 1 \quad \text{for} \quad 0 < x \le 1$$
$$\therefore \frac{f(x,y)}{2x} \ge 0$$
$$\therefore f(x,y) \ge 0$$

□

Theorem 3. *In non-deterministic environments the following result holds:*

$$\sum_{t=1}^{n} \mathbb{E}\left[\left(\sqrt{\tfrac{\xi(x_t|x_{<t})}{\mu(x_t|x_{<t})}} - 1\right)^2\right] \;\le\; \sum_{t=1}^{n} \mathbb{E}[h_t] \;\le\; D_n(\mu\|\xi) := \mathbb{E}[\ln \tfrac{\mu(x_{1:n})}{\xi(x_{1:n})}] \;\le\; \ln w_\mu^{-1} \;<\; \infty \qquad (3)$$

where

$$h_t(x_{<t}) \;:=\; \sum_{|a|=1} \Big(\sqrt{\xi(a|x_{<t})} - \sqrt{\mu(a|x_{<t})}\Big)^2 \qquad (4)$$

Proof. for the first " ≤ ",

$$(\sqrt{\frac{\xi}{\mu}} - 1)^2 = \mu^{-1}(\sqrt{\xi} - \sqrt{\mu})^2$$

for the second " ≤ ", use Lemma 1.

for the third " ≤ ",

$$\xi(x) \ge w_\mu \mu(x)$$

□

These bounds (with $n = \infty$) imply $h_t \to 0$ and hence

$$\xi(x_t|x_{<t}) - \mu(x_t|x_{<t}) \;\longrightarrow\; 0$$

and

$$\frac{\xi(x_t|x_{<t})}{\mu(x_t|x_{<t})} \;\longrightarrow\; 1$$

both rapidly with probability 1 for $t \to \infty$.

2.2 Bayesian Decisions

Let $\text{Loss}(x_t, y_t) \in [0, 1]$ be the received loss when y_t has been predicted and x_t was the correct observation.

Given this loss function the optimal predictor Λ_ρ for environment ρ after seeing observations $x_{<t}$ is defined as the prediction or decision or action y_t that minimizes the ρ-expected loss. This is the action that we expect to be least bad according to environment ρ.

$$y_t^{\Lambda_\rho}(x_{<t}) \;:=\; \arg\min_{y_t} \sum_{x_t} \rho(x_t|x_{<t})\text{Loss}(x_t, y_t)$$

Given this optimal predictor, the expected instantaneous loss at time t and the total expected loss from the first n predictions are defined as follows.

$$\begin{aligned} loss_t^{\Lambda_\rho} \;&:=\; \mathbb{E}[\text{Loss}(x_t, y_t^{\Lambda_\rho})] \\ Loss_n^{\Lambda_\rho} \;&:=\; \sum_{t=1}^{n} \mathbb{E}[\text{Loss}(x_t, y_t^{\Lambda_\rho})] \end{aligned}$$

Obviously the best predictor possible is the optimal predictor for the true environment Λ_μ, however as μ is generally unknown, the best available option is the optimal predictor Λ_ξ for Bayes mixture ξ for which the following result holds:

Theorem 4.

$$\left(\sqrt{Loss_n^{\Lambda_\xi}} - \sqrt{Loss_n^{\Lambda_\mu}}\right)^2 \;\leq\; \sum_{t=1}^{n} \mathbb{E}\left[\left(\sqrt{loss_t^{\Lambda_\xi}} - \sqrt{loss_t^{\Lambda_\mu}}\right)^2\right] \;\leq\; 2\ln(w_\mu^{-1}) \;<\; \infty$$

One can also show that Λ_ξ is Pareto-optimal in the sense that every other predictor with smaller loss than Λ_ξ in some environment $\nu \in \mathcal{M}$ must be worse in another environment.

2.3 Continuous Environment Classes

Although the results above were proved assuming that the model class is countable, analogous results hold for the case that the model class $\mathcal{M}$ is uncountable such as continuous parameter classes.

$$\xi(x) \;=\; \int_{\nu\in\mathcal{M}} \nu(x)w(\nu)d\nu$$

where $w(\nu)$ is (now) a prior probability *density* over $\nu \in \mathcal{M}$. One problem with this is that the dominance $\xi(x) \geq w(\mu)\mu(x)$ is no longer valid since the prior probability (not the density) is zero for any single point. To avoid this problem the Bayesian mixture is instead shown to dominate the integral over a small vicinity around the true environment μ. By making some weak assumptions about the smoothness of the parametric model class $\mathcal{M}$, a weaker type of dominance makes it possible to prove the following:

Theorem 5.

$$\sum_{t=1}^{n} \mathbb{E}[h_t] \;\leq\; \ln(w(\mu)^{-1}) + O(\log(n))$$

2.4 Choosing the Model Class and Priors

The above results demonstrate that the Bayesian framework is highly effective and essentially optimal given the available information. Unfortunately the operation and performance of this framework is sensitive to the initial choice of hypothesis class and prior. As long as they are non zero, the chosen priors will not affect the asymptotic performance of the Bayesian mixture as the observations eventually wash out this initial belief value. However in short-term applications they can have a significant impact.

The only restriction on the hypothesis class is that it must contain the true environment. But adding unnecessarily small priors leads to a high error bound which may affect short-term performance.

For these reasons, the general guideline is to choose the smallest model class that will contain the true environment and priors that best reflect a rational a-priori belief in each of these environments. Occam's razor in conjunction with Epicurus' principle of multiple explanations, quantified by Kolmogorov complexity will lead us to the universal prior.

3 How to Choose the Prior

"God always takes the simplest way."

— Einstein

3.1 Indifference Principle

Quantifying Epicurus's principle of multiple explanations leads to the *indifference principle*(IP) which assumes that if there is no evidence favoring any particular hypothesis then we should weight them all as equally likely. When told that an urn contains either all black balls or all white balls and no other information, it seems natural to assign a probability of 0.5 to each hypothesis before any balls have been observed. This can be extended to any finite hypothesis class by assigning probability $\frac{1}{|\mathcal{M}|}$ to each hypothesis where $|\mathcal{M}|$ is the number of hypotheses in $\mathcal{M}$.

For a continuous hypothesis class the analogous approach is to assign a uniform prior density which must integrate to 1 to be a proper probability density.

Since given data D, the posterior belief of the hypothesis H is

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

if indifference principle holds, it justifies the *maximum likelihood principle*.

$$P(H|D) \propto P(D|H)$$
$$\widehat{H} = \arg\max_H P(H|D) = \arg\max_H P(D|H)$$

However, we know that maximum likelihood principle is confronted with the overfitting problem when the model class $\mathcal{M}$ is too large. We will show that indifference principle

is not reparametrization invariant. Jeffreys' solution is to find a symmetry group of the problem (like permutations for finite $\mathcal{M}$) and require the prior to be *invariant under group transformations*. For instance, if $\theta \in \mathbb{R}$ is a location parameter (e.g. the mean) it is natural to require a translation-invariant prior. Problems are that there may be no obvious symmetry, the resulting prior may be improper (like for the translation group), and the result can depend on which parameters are treated as nuisance parameters.

The *maximum entropy principle*(ME) extends the symmetry principle by allowing certain types of constraints on the parameters. The ME principle selects the estimated values $\widehat{\theta} = (\widehat{p_1}, \cdots, \widehat{p_k})$ that maximize the entropy function

$$H(p_1, \cdots, p_k) = -\sum_{i=1}^{k} p_i \log p_i$$

suject to

$$\sum_{i=1}^{k} p_i = 1$$

and some other constraints provided by empirical data or considerations of symmetry, probabilistic laws, and so on.

When we are totally ignorant, the indifference principle follows from the maximum entropy principle.

$$\widehat{\theta} = (\widehat{p_1}, \cdots, \widehat{p_k}) = \arg \max_{(p_1, \cdots, p_k)} \sum_{i=1}^{k} p_i \log \frac{1}{p_i} = (\frac{1}{k}, \cdots, \frac{1}{k})$$

Both IP and ME can be considered as special cases of the *minimum description length principle*(MDL).

$$\widehat{H} = \arg \min_{H} K(D|H) + K(H) \approx \arg \max_{H} P(D|H)P(H) = \arg \max_{H} P(H|D)$$

Conjugate priors are classes of priors such that the posteriors are themselves again in the class. While this can lead to interesting classes, the principle itself is not selective, since e.g. the class of all priors forms a conjugate class.

3.2 Laplace and the Rule of Succession

Supose the model class is $\mathcal{M} = \{\theta | \theta \in [0, 1]\}$. We estimate the true probability using Bayes mixture which represents our subjective probability. This involves integrating over our prior belief density $w(\theta) = P(\theta)$ to give

$$\xi(x) \;=\; P(x) \;=\; \int_0^1 P(x|\theta) w(\theta) d\theta$$

We assume the prior distribution to be uniform and proper:

$$\int_0^1 w(\theta)d\theta \;=\; 1 \quad \text{and} \quad w(\theta) \;=\; w(\theta') \text{ for all } \theta \text{ and } \theta' \in \mathcal{M}$$

This results in the density $\forall\theta \in [0,1](w(\theta) = 1)$.

$$\therefore P(x) \;=\; \int_0^1 P(x|\theta)d\theta \;=\; \int_0^1 \theta^s(1-\theta)^f d\theta \;=\; \frac{s!f!}{(s+f+1)!}$$

$$\therefore P(x_{n+1} = 1|x_{1:n}) \;=\; \frac{P(x_{1:n}1)}{P(x_{1:n})} \;=\; \frac{\frac{(s+1)!f!}{(s+1+f+1)!}}{\frac{s!f!}{(s+f+1)!}} \;=\; \frac{s+1}{s+f+2} \;=\; \frac{s+1}{n+2}$$

3.3 Confirmation Problem

For some hypothesis H and evidence E Bayes rule states $P(H|E) = P(E|H)P(H)/P(E)$. Therefore it is clear that if $P(H) = 0$ then regardless of the evidence E our posterior evidence $P(H|E)$ must remain identically zero. Imagine we are observing the color of ravens and θ is the percentage of ravens that are black. The hypothesis "All ravens are black" therefore might be associated with $\theta = 1$, but even after observing one million black ravens and no non-black ravens $P(\theta = 1|x) = 0$.

Instead of $\theta = 1$ it is also possible to formulate the hypothesis "all ravens are black" as the observation sequence of an infinite number of black ravens, i.e. $H' = x = 1^\infty$ where a 1 is a black raven. If $x_{1:n} = 1^n$ is a sequence of n black ravens, then $P(x_{1:n}) = \frac{n!}{(n+1)!} = \frac{1}{n+1}$. Therefore

$$P(1^k|1^n) \;=\; \frac{P(1^{n+k})}{P(1^n)} \;=\; \frac{n}{n+k}$$

However for the above hypothesis of "all ravens are black" k is infinite and $P(1^{k=\infty}|1^n) = 0$ for any number n of observed ravens.

If we instead consider the composite or partial hypothesis $\theta_p = \{\theta|\theta \in (1-\varepsilon, 1]\}$ for any arbitrarily small ε, then $P(\theta|x)$ converges to 1 as the number of observed black ravens increases. This is called a soft hypothesis and intuitively it is the hypothesis that the percentage of black ravens is 1 or very close to 1. The reason our belief in this hypothesis can converge to 1 is that the probability is now the integral over a small interval which has a-priori non-zero mass $P(\theta_p) = \varepsilon > 0$ and a-posteriori asymptotically all mass $P(\theta_p|1^n) \to 1$.

But this implies that we are certain that everything has exceptions, which is unreasonable. We can not be certain about their truth or falsity.

Since the indifference principle gives rise to the zero prior problem and hence the confirmation problem, there is another solution that assigns a non-zero weight to the point $\theta = 1$. Consider, for instance, the improper density $w(\theta) = \frac{1}{2}(1 + \delta(1-\theta))$, where δ is the Dirac-$\delta(\int f(\theta)\delta(\theta - a)\,d\theta = f(a))$, or equivalently $P(\theta \geq a) = 1 - \frac{1}{2}a$ with $a \in [0,1]$, which gives $P(\theta = 1) = \frac{1}{2}$. Using this approach results in the following Bayesian mixture

distribution:

$$\xi(x_{1:n}) \;=\; \frac{1}{2}\left(\frac{s!f!}{(n+1)!}+\delta_{s,n}\right) \quad \text{where} \quad \delta_{s,n}=\begin{cases}1 & \text{if } s=n\\ 0 & \text{otherwise}\end{cases}$$

Therefore, if all ravens observed are black, the Bayesian mixture gives $\xi(1^n) = \frac{1}{2}(\frac{n!0!}{(n+1)!} + 1) = \frac{1}{2}\cdot\frac{n+2}{n+1}$, which is much larger than the $\xi(1^n) = \frac{1}{n+1}$ given by the uniform prior.

$$P(1^k|1^n) \;=\; \xi(1^k|1^n) \;=\; \frac{\xi(1^{n+k})}{\xi(1^n)} \;=\; \frac{n+k+2}{n+k+1}\cdot\frac{n+1}{n+2}$$
$$P(H'|1^n) \;=\; P(1^\infty|1^n) \;=\; \lim_{k\to\infty} P(1^k|1^n) \;=\; \frac{n+1}{n+2}$$

It is clear that the chosen "improper density" solution is biased towards universal generalizations, in this case to the hypothesis "all ravens are black". The question is then why not design the density to also be able to confirm "no ravens are black", or "exactly half the ravens are black"? It seems that we are intuitively biased towards hypotheses corresponding to simpler values. Then why not assign non-zero prior to all computable θ?

3.4 Reparametrization Invariance

By applying some general principle to a parameter θ of hypothesis class $\mathcal{M}$ we arrive at prior $w(\theta)$. If we consider some new parametrization θ' which is related to θ via some bijection $f : \theta' = f(\theta)$, then there would be two ways to arrive at a prior which focuses on this new parameter θ'. Firstly we can directly apply the same principle to this new parametrization to get prior $w'(\theta')$. The second way is to transform the original prior using this same bijection. If both of these ways lead to the same prior we say the principle we used satisfy the reparametrization invariance principle(RIP).

It is clear that the indifference principle does not satisfy RIP in the case of densities, although it does satisfy RIP for finite model classes $\mathcal{M}$.

3.5 Regrouping Invariance

Regrouping invariance can be thought of as a generalization of the concept of reparametrization invariance, with the function f that is not necessarily bijective and hence can lead to a many to one or one to many correspondence. Because the function f is not bijective anymore, the transformation of the prior $w(\theta)$ to some new parametrization θ' now involves an integral or sum of the priors over all values of θ for which $f(\theta) = \theta'$. Formally, for discrete class $\mathcal{M}$ we have $\tilde{w}_{\theta'} = \sum_{\theta:f(\theta)=\theta'} w_\theta$, and similarly for continuous parametric classes we have $\tilde{w}(\theta') = \int \delta(f(\theta)-\theta')w(\theta)d\theta$. As with reparametrization invariance before, for a principle to be regrouping invariant, we require that $\tilde{w}(\theta') = w'(\theta')$ where $w'(\theta')$ is obtained by applying the same principle to the new parametrization.

For example, for an i.i.d. with d possible observations the parameter space is $\triangle^{d-1} := \{\vec{\theta} \equiv (\theta_1, ..., \theta_d) \in [0,1]^d : \sum_{i=1}^d \theta_i = 1\}$. The probability of $x_{1:n}$, with n_i occurrences of observation i, is given by $P(x_{1:n}|\vec{\theta}) = \prod_{i=1}^d \theta_i^{n_i}$.

The regrouping problem arises when we want to make inferences about the hypothesis "all ravens are black" when the setup is now to record the extra information of whether a raven is colored or white. When we make an inference that only looks at the 'blackness' of a raven, the observations are collapsed into blackness or non-blackness as before by mapping black to success and either white or colored to failure. Now $P(x_{1:n}|\theta) = \theta^s(1-\theta)^f$. However, since we assumed indifference over the parameter vectors in $\triangle^2$, by regrouping the prior belief is skewed towards higher proportions of non-black ravens. Therefore $\tilde{w}(\theta') = 2(1-\theta') \neq 1 = w'(\theta')$ for $\theta' \in [0,1]$.

In fact, it was shown by Wallace that there is no acceptable prior density that solves this problem universally.

3.6 Universal Prior

The universal prior is designed to do justice to both Occam and Epicurus as well as be applicable to any computable environment. To do justice to Epicurus' principle of multiple explanations we must regard all environments as possible, which means the prior for each environment must be non zero. To do justice to Occam we must regard simpler hypotheses as more plausible than complex ones. To be a valid prior it must also sum to (less than or equal to) one. Since the prefix Kolmogorov complexity satisfies Kraft's inequality, the following is a valid prior.

$$w_\nu^U \;:=\; 2^{-K(\nu)}$$

This prior is monotonically decreasing in the complexity of ν and is non-zero for all computable ν.

When the bounds for Bayesian prediction are re-examined in the context of the Universal Prior we see that the upper bounds on the deviation of the Bayesian mixture from the true environment are $\ln(w_\mu^{-1}) = \ln(2^{K(\mu)}) = K(\mu)\ln(2)$.

The universality of Kolmogorov complexity bestows the universal prior w_ν^U with remarkable properties. First, any other reasonable prior w_ν gives approximately the same or weaker bounds. Second, the universal prior approximately satisfies both, reparametrization and regrouping invariance. This is possible, since it is not a density.

Theorem 6. *w_ν^U is reparametrization invariant.*

Proof.

$$\tilde{w}_{\theta'}^U = w_{f^{-1}(\theta')}^U = 2^{-K(f^{-1}(\theta'))} \stackrel{\times}{=} 2^{-K(\theta')} = w'^U_{\theta'}$$

□

Theorem 7. *w_ν^U is regrouping invariant.*

Proof.

$$\tilde{w}^U_{\theta'} \;=\; \sum_{\theta:f(\theta)=\theta'} 2^{-K(\theta)} \;\stackrel{\times}{=}\; 2^{-K(\theta')} \;=\; w'^U_{\theta'} \qquad [Equation(K6)]$$

□

4 Solomonoff's Universal Probability

"God does not play dice."
"Reality is merely an illusion, albeit a very persistent one."

— Einstein

Solomonoff's [Sol78] induction scheme completes the general Bayesian framework by choosing the model class $\mathcal{M}$ to be the class of all computable measures and taking the universal prior over this class.

4.1 How to Choose the Model Class

The class of all computable distributions, although only countable, is pretty large from a practical point of view. Finding a non-computable physical system would overturn the Church-Turing thesis. It is the largest class, relevant from a computational point of view. However, this class is not recursively enumerable, since the class of total computable functions $f : \mathcal{X}^* \to \mathbb{R}$ is not recursively enumerable because of halting problem. Levin "slightly" extends the class to include also lower semi-computable semimeasures. One can show that this class $\mathcal{M}_U = \{\nu_1, \nu_2, ...\}$ is recursively enumerable, hence the universal Bayesian mixture

$$\xi_U(x) \;=\; \sum_{\nu\in\mathcal{M}_U} w^U_\nu \nu(x) \tag{5}$$

is itself lower semi-computable.

Theorem 8. $\xi_U \in \mathcal{M}_U$

Proof.

$\because w^U_\nu = 2^{-K(\nu)}$ and ν are lower semi-computable.
$\therefore \exists n(\lim_{n\to\infty} w^n_\nu = w^U_\nu)$ and $\exists n(\lim_{n\to\infty} \nu^n(x) = \nu(x))$
$\because \xi^n_U(x) = \sum_{\nu\in\mathcal{M}_U} w^n_\nu \nu^n(x)$
$\therefore \lim_{n\to\infty} \xi^n_U(x) = \xi_U(x)$
$\because \xi^n_U$ is increasing in n.
$\therefore \xi_U$ is lower semi-computable.

$$\because \xi_U(x) = \sum_{\nu\in\mathcal{M}_U} w_\nu^U \nu(x) \geq \sum_{\nu\in\mathcal{M}_U} w_\nu^U (\sum_{|a|=1} \nu(xa)) = \sum_{|a|=1} \sum_{\nu\in\mathcal{M}_U} w_\nu^U \nu(xa) = \sum_{|a|=1} \xi_U(xa)$$

$\therefore \xi_U$ is semi-measure.

□

Obviously,

$$\xi_U(x) \geq w_\nu^U \nu(x) \tag{6}$$

One of the problems with the Bayesian framework is dealing with new hypotheses H that were not in the original class $\mathcal{M}$. In science it is natural to come up with a new explanation of some data which cannot be satisfactorily explained by any of the current models. Unfortunately the Bayesian framework describes how to update our belief in a hypothesis according to evidence but not how to assign a belief if the hypothesis was created to fit the data. By choosing the universal class $\mathcal{M}_U$ this problem is formally solved. Theoretically it can no longer occur since this class is complete in the sense that it already contains any reasonable hypothesis.

4.2 Deterministic Representation

The above definition is a mixture over all semi-computable stochastic environments using the universal prior as weights. It is however possible to think about ξ_U in a completely different way. To do this we assume that the world is governed by some deterministic computable process. In other words, suppose the world is created by a God who flips a coin instead of playing dice, equipped with a Universal Monotone Turing Machine which reads a 1 for heads and a 0 for tails. In this case the probability of x should be:

$$M(x) \;:=\; \sum_{p:U(p)=x*} 2^{-|p|}$$

where $U(p) = x*$ means p is a minimal program printing a string starting with x.

Obviously, It can be regarded as a $2^{-|p|}$-weighted mixture over all computable deterministic environments ν_p ($\nu_p(x) = 1$ if $U(p) = x*$ and 0 otherwise).

It can also be seen as follow:$M(x) = \lim_{n\to\infty} \frac{|\{p\in\mathbb{B}^n : U(p)=x*\}|}{2^n}$.

It turns out that $\xi_U \stackrel{\times}{=} M$. One can also get an explicit enumeration of all lower semi-computable semimeasures $\mathcal{M}_U = \{\nu_1, \nu_2, ...\}$ by means of $\nu_i(x) := \sum_{p:T_i(p)=x*} 2^{-|p|}$, where $T_i(p) \equiv U(\langle i\rangle p)$, $i = 1, 2, ...$ is an enumeration of all monotone Turing machines.

Lemma 2. *For $\nu \in \mathcal{M}_U$,*

$$M(x) \stackrel{\times}{\geq} 2^{-K(\nu)} \nu(x) \tag{7}$$

Proof.

$$M(x) = \sum_{p:U(p)=x*} 2^{-|p|}$$

$$
\begin{aligned}
&\geq \sum_{q:U(\langle T\rangle q)=x*} 2^{-|q|} \\
&= 2^{-|T|} \sum_{q:T(q)=x*} 2^{-|q|} \\
&\overset{\times}{=} 2^{-K(\nu)}\nu(x)
\end{aligned}
$$

□

4.3 Total Bounds

Since Solomonoff's approach is simply the Bayesian framework with the universal model class and prior, the bounds for Bayes mixture remain valid for ξ_U, therefore also for M.

In the case that the true distribution μ is deterministic the following bound holds.

Theorem 9.

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

where the monotone complexity $Km(x) := \min\{|p| : U(p) = x*\}$

Proof. Follows from theorem 2 and $w_\mu^U = 2^{-K(\mu)} = 2^{-K(x_{1:\infty})}$ and $K(x_{1:\infty}) \leq Km(x_{1:\infty})$ □

If $x_{1:\infty}$ is a computable sequence, then $Km(x_{1:\infty})$ is finite, which implies $M(x_t|x_{<t}) \to 1$. In particular, observing an increasing number of black ravens, $M(1|1^n) \to 1$ ($Km(1^\infty) = O(1)$), and we become rapidly confident that future ravens are black.

For the non-deterministic case, similar results hold.

Lemma 3 (Entropy Inequality). *Let μ and ρ be two probability distributions over $\mathcal{X}^*$. Then we have*

$$
\sum_{|a|=1} (\mu(a) - \rho(a))^2 \leq \sum_{|a|=1} \mu(a) \ln \frac{\mu(a)}{\rho(a)} \tag{8}
$$

Proof. for $\mathcal{X} = \mathbb{B} = \{0, 1\}$

$$
\begin{aligned}
&f(p, q) := p \ln \frac{p}{q} + (1 - p) \ln \frac{1 - p}{1 - q} - 2(p - q)^2 \\
&\because \frac{\partial f}{\partial q} = (q - p) \frac{4(q - \frac{1}{2})^2}{q(1 - q)} \\
&\therefore f(p, q) \geq 0
\end{aligned}
$$

□

Theorem 10 (Completeness Theorem).

$$
\sum_{t=1}^{\infty} \sum_{x_{1:t} \in \mathbb{B}^t} \mu(x_{<t}) \Big(M(x_t|x_{<t}) - \mu(x_t|x_{<t}) \Big)^2 \overset{+}{\leq} K(\mu) \ln 2 < \infty \tag{9}
$$

Proof.

$$M'(\epsilon) := 1$$
$$M'(xa) := M'(x)\frac{M(xa)}{\sum_{|a|=1} M(xa)}$$

$$\begin{aligned}
&\sum_{t=1}^{n}\sum_{x_{1:t}\in\mathbb{B}^t}\mu(x_{<t})\Big(M(x_t|x_{<t})-\mu(x_t|x_{<t})\Big)^2\\
&=\sum_{t=1}^{n}\sum_{x_{<t}}\mu(x_{<t})\sum_{x_t}\Big(M(x_t|x_{<t})-\mu(x_t|x_{<t})\Big)^2\\
&\le\sum_{t=1}^{n}\sum_{x_{<t}}\mu(x_{<t})\sum_{x_t}\mu(x_t|x_{<t})\ln\frac{\mu(x_t|x_{<t})}{M'(x_t|x_{<t})} \qquad \text{[Lemma 3]}\\
&\le\sum_{t=1}^{n}\sum_{x_{<t}}\mu(x_{<t})\sum_{x_t}\mu(x_t|x_{<t})\ln\frac{\mu(x_t|x_{<t})}{M(x_t|x_{<t})}\\
&=\sum_{t=1}^{n}\sum_{x_{1:t}}\mu(x_{1:t})\ln\frac{\mu(x_t|x_{<t})}{M(x_t|x_{<t})}\\
&=\sum_{t=1}^{n}\sum_{x_{1:t}}\Big(\sum_{x_{t+1:n}}\mu(x_{1:n})\Big)\ln\frac{\mu(x_t|x_{<t})}{M(x_t|x_{<t})}\\
&=\sum_{t=1}^{n}\sum_{x_{1:n}}\mu(x_{1:n})\ln\frac{\mu(x_t|x_{<t})}{M(x_t|x_{<t})}\\
&=\sum_{x_{1:n}}\mu(x_{1:n})\sum_{t=1}^{n}\ln\frac{\mu(x_t|x_{<t})}{M(x_t|x_{<t})}\\
&=\sum_{x_{1:n}}\mu(x_{1:n})\ln\frac{\mu(x_{1:n})}{M(x_{1:n})}\\
&\overset{+}{\le} K(\mu)\ln 2 \qquad \text{[Lemma 2]}
\end{aligned}$$

□

Both results consider one-step lookahead prediction but are easily extendible to multi-step lookahead prediction. Exploiting absolute continuity of μ w.r.t. M, asymptotic convergence can be shown even for infinite lookahead and any computable μ:

Theorem 11.

$$\sup_{A\subseteq\mathcal{X}^\infty}\Big|M(A|x_{<t})-\mu(A|x_{<t})\Big| \;\longrightarrow\; 0 \quad \textit{with } \mu\textit{-probability } 1$$

4.4 Instantaneous Bounds

The previous bounds give excellent guarantees over some initial n predictions but say nothing about the nth prediction itself. The following instantaneous bound for computable x also holds:

Theorem 12.

$$2^{-K(n)} \;\leq\; (1 - M(x_n|x_{<n})) \;\leq\; 2^{2Km(x_{1:n})-K(n)}$$

In particular for $x_{1:\infty} = 1^\infty$ we get

$$M(0|1^n) \;\stackrel{\times}{=}\; 2^{-K(n)}$$

which means that M quickly disbelieves in non-black ravens.

4.5 Future Bounds

When looking at an agent's performance it is often important to consider not only the total and instantaneous performance but also the total future performance bounds. In other words it can be important to estimate how many errors it is going to make from now on.

Theorem 13.

$$\sum_{t=n+1}^{\infty} \mathbb{E}[h_t|x_{1:n}] \;\stackrel{+}{\leq}\; (K(\mu|x_{1:n}) + K(n)) \ln 2$$

4.6 Universal is Better than Continuous

It can be proved that M can do as good as any Bayesian mixture ξ over any model class $\mathcal{M}$, continuous or discrete, and prior function $w()$ over this class. The reason for this is that although a specific environment ν in $\mathcal{M}$ may be incomputable and its prior w_ν may be zero, the prior function $w()$ and the overall mixture ξ generally remain computable. This computability of ξ implies the following general result for any, possibly incomputable, environment μ:

Theorem 14.

$$D_n(\mu||M) := \mathbb{E}[\ln \frac{\mu}{M}] = \mathbb{E}[\ln \frac{\mu}{\xi}] + \mathbb{E}[\ln \frac{\xi}{M}] \;\stackrel{+}{\leq}\; D_n(\mu||\xi) + K(\xi) \ln 2$$

So any bound valid for $D_n(\mu||\xi)$ is directly valid for $D_n(\mu||M)$. If there exists any computable predictor that converges to μ, then so does M, whether μ is computable or not.

5 Hutter's AIXI

"God is subtle but he is not malicious."

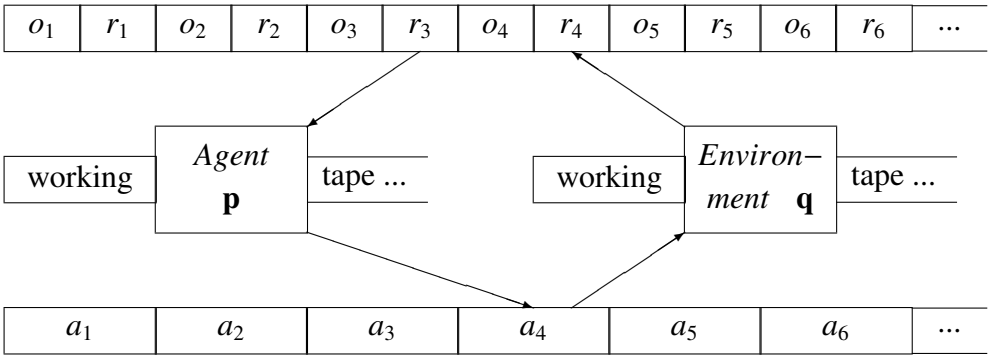
— Einstein

The General Reinforcement Learning Problem Consider an agent that exists within some unknown environment. The agent interacts with the environment in cycles. In each cycle, the agent executes an action and in turn receives an observation and a reward. The only information available to the agent is its history of previous interactions. The *general reinforcement learning problem* is to construct an agent that, over time, collects as much reward as possible from the (unknown) environment.

The AIXI Agent To achieve generality, the environment is assumed to be an unknown but computable function; i.e. the observations and rewards received by the agent, given its past actions, can be computed by some program running on a Turing machine.

More formally, let $U(q, a_1 a_2 \dots a_n)$ denote the output of a universal Turing machine U supplied with program q and input $a_1 a_2 \dots a_n$, $m \in \mathbb{N}$ a finite lookahead horizon, and $|q|$ the length in bits of program q. The action picked by AIXI [VNH11] at time t, having executed actions $a_1 a_2 \dots a_{t-1}$ and having received the sequence of observation-reward pairs $o_1 r_1 o_2 r_2 \dots o_{t-1} r_{t-1}$ from the environment, is given by:

$$a_t^* = \arg\max_{a_t} \sum_{o_t r_t} \dots \max_{a_{t+m}} \sum_{o_{t+m} r_{t+m}} [r_t + \dots + r_{t+m}] \sum_{q : U(q, a_1 \dots a_{t+m}) = o_1 r_1 \dots o_{t+m} r_{t+m}} 2^{-|q|}. \tag{10}$$



5.1 The Agent Setting

Agent Setting The (finite) action, observation, and reward spaces are denoted by $\mathcal{A}, \mathcal{O}$, and $\mathcal{R}$ respectively. Also, $\mathcal{X}$ denotes the joint perception space $\mathcal{O} \times \mathcal{R}$.

Definition 2. *A history h is an element of* $(\mathcal{A} \times \mathcal{X})^* \cup (\mathcal{A} \times \mathcal{X})^* \times \mathcal{A}$.

Definition 3. *An environment ρ is a sequence of conditional probability functions $\{\rho_0, \rho_1, \rho_2, \dots\}$, where $\rho_n \colon \mathcal{A}^n \to$ Density $(\mathcal{X}^n)$, that satisfies*

$$\forall a_{1:n} \forall x_{<n} : \rho_{n-1}(x_{<n} | a_{<n}) = \sum_{x_n \in \mathcal{X}} \rho_n(x_{1:n} | a_{1:n}). \tag{11}$$

In the base case, we have $\rho_0(\epsilon|\epsilon) = 1$.

Equation (11), called the chronological condition in [Hut05], captures the natural constraint that action a_n has no effect on earlier perceptions $x_{<n}$. For convenience, we drop the index n in ρ_n from here onwards.

Given an environment ρ, we define the predictive probability

$$\forall a_{1:n} \forall x_{1:n} : \rho(x_{<n}|a_{<n}) > 0 \Longrightarrow \rho(x_n|ax_{<n}a_n) := \frac{\rho(x_{1:n}|a_{1:n})}{\rho(x_{<n}|a_{<n})} \tag{12}$$

$$\rho(x_{1:n}|a_{1:n}) = \rho(x_1|a_1)\rho(x_2|ax_1a_2)\cdots\rho(x_n|ax_{<n}a_n) \tag{13}$$

Reward, Policy and Value Functions The agent's goal is to accumulate as much reward as it can during its lifetime. More precisely, the agent seeks a *policy* that will allow it to maximise its expected future reward up to a fixed, finite, but arbitrarily large horizon $m \in \mathbb{N}$. The instantaneous reward values are assumed to be bounded.

Definition 4. *Given history $ax_{1:t}$, the m-horizon expected future reward of an agent acting under policy $\pi\colon (\mathcal{A}\times\mathcal{X})^* \to \mathcal{A}$ with respect to an environment ρ is:*

$$v^m_\rho(\pi, ax_{1:t}) := \mathbb{E}_\rho\left[\sum_{i=t+1}^{t+m} R_i(ax_{\leq t+m}) \;\middle|\; x_{1:t}\right], \tag{14}$$

where for $t < k \leq t+m$, $a_k := \pi(ax_{<k})$ and $R_i(aor_{\leq n}) := r_k$ for $1 \leq i \leq n$. The quantity $v^m_\rho(\pi, ax_{1:t}a_{t+1})$ is defined similarly, except that a_{t+1} is now no longer defined by π.

The optimal policy π^* is the policy that maximises the expected future reward. The maximal achievable expected future reward of an agent with history h in environment ρ looking m steps ahead is $V^m_\rho(h) := v^m_\rho(\pi^*, h)$. It is easy to see that if $h \in (\mathcal{A}\times\mathcal{X})^t$, then

$$V^m_\rho(h) = \max_{a_{t+1}} \sum_{x_{t+1}} \rho(x_{t+1}|ha_{t+1}) \cdots \max_{a_{t+m}} \sum_{x_{t+m}} \rho(x_{t+m}|hax_{t+1:t+m-1}a_{t+m}) \left[\sum_{i=t+1}^{t+m} r_i\right] \tag{15}$$

For convenience, we will often refer to Equation (15) as the *expectimax operation*. Furthermore, the m-horizon optimal action a^*_{t+1} at time $t+1$ is related to the expectimax operation by

$$a^*_{t+1} = \arg\max_{a_{t+1}} V^m_\rho(ax_{1:t}a_{t+1}). \tag{16}$$

5.2 Bayesian Agents

Since we are assuming that the agent does not initially know the true environment, we desire subjective models whose predictive performance improves as the agent gains experience.

Definition 5. *Given a countable model class $\mathcal{M} := \{\rho_1, \rho_2, \dots\}$ and a prior weight $w^\rho_0 > 0$ for each $\rho \in \mathcal{M}$ such that $\sum_{\rho\in\mathcal{M}} w^\rho_0 = 1$, the mixture environment model is $\xi(x_{1:n}|a_{1:n}) := \sum_{\rho\in\mathcal{M}} w^\rho_0 \rho(x_{1:n}|a_{1:n})$.*

Proposition 1. *A mixture environment model is an environment model.*

Proof. $\forall a_{1:n} \in \mathcal{A}^n$ and $\forall x_{<n} \in \mathcal{X}^{n-1}$ we have that

$$\sum_{x_n \in \mathcal{X}} \xi(x_{1:n}|a_{1:n}) = \sum_{x_n \in \mathcal{X}} \sum_{\rho \in \mathcal{M}} w_0^\rho \rho(x_{1:n}|a_{1:n}) = \sum_{\rho \in \mathcal{M}} w_0^\rho \sum_{x_n \in \mathcal{X}} \rho(x_{1:n}|a_{1:n}) = \xi(x_{<n}|a_{<n})$$

where the final step follows from application of Equation (11) and Definition 5. □

Prediction with a Mixture Environment Model As a mixture environment model is an environment model, we can simply use:

$$\xi(x_n|ax_{<n}a_n) = \frac{\xi(x_{1:n}|a_{1:n})}{\xi(x_{<n}|a_{<n})} \tag{17}$$

to predict the next observation reward pair. Equation (17) can also be expressed in terms of a convex combination of model predictions, with each model weighted by its posterior, from

$$\xi(x_n|ax_{<n}a_n) = \frac{\sum_{\rho \in \mathcal{M}} w_0^\rho \rho(x_{1:n}|a_{1:n})}{\sum_{\rho \in \mathcal{M}} w_0^\rho \rho(x_{<n}|a_{<n})} = \sum_{\rho \in \mathcal{M}} w_{n-1}^\rho \rho(x_n|ax_{<n}a_n), \tag{18}$$

where the posterior weight w_{n-1}^ρ for environment model ρ is given by

$$w_{n-1}^\rho := \frac{w_0^\rho \rho(x_{<n}|a_{<n})}{\sum_{\nu \in \mathcal{M}} w_0^\nu \nu(x_{<n}|a_{<n})} = \Pr(\rho|ax_{<n}) \tag{19}$$

If $|\mathcal{M}|$ is finite, Equations (17) and (18) can be maintained online in $O(|\mathcal{M}|)$ time by using the fact that

$$\rho(x_{1:n}|a_{1:n}) = \rho(x_{<n}|a_{<n})\rho(x_n|ax_{<n}a),$$

which follows from Equation (13), to incrementally maintain the likelihood term for each model.

Theoretical Properties We now show that if there is a good model of the (unknown) environment in $\mathcal{M}$, an agent using the mixture environment model

$$\xi(x_{1:n}|a_{1:n}) := \sum_{\rho \in \mathcal{M}} w_0^\rho \rho(x_{1:n}|a_{1:n}) \tag{20}$$

will predict well.

Theorem 15. *Let μ be the true environment. The μ-expected squared difference of μ and ξ is bounded as follows. For all $n \in \mathbb{N}$, for all $a_{1:n}$,*

$$\sum_{k=1}^{n} \sum_{x_{1:k}} \mu(x_{<k}|a_{<k}) \Big(\mu(x_k|ax_{<k}a_k) - \xi(x_k|ax_{<k}a_k)\Big)^2 \leq \min_{\rho \in \mathcal{M}} \Big\{ -\ln w_0^\rho + D_{1:n}(\mu \,\|\, \rho) \Big\} \tag{21}$$

where $D_{1:n}(\mu \,\|\, \rho) := \sum_{x_{1:n}} \mu(x_{1:n}|a_{1:n}) \ln \frac{\mu(x_{1:n}|a_{1:n})}{\rho(x_{1:n}|a_{1:n})}$ is the KL divergence of $\mu(\cdot|a_{1:n})$ and $\rho(\cdot|a_{1:n})$.

Proof. Combining [Hut05, §3.2.8 and §5.1.3] we get

$$
\begin{aligned}
&\sum_{k=1}^{n}\sum_{x_{1:k}}\mu(x_{<k}|a_{<k})\Big(\mu(x_k|ax_{<k}a_k)-\xi(x_k|ax_{<k}a_k)\Big)^2\\
&=\sum_{k=1}^{n}\sum_{x_{<k}}\mu(x_{<k}|a_{<k})\sum_{x_k}\Big(\mu(x_k|ax_{<k}a_k)-\xi(x_k|ax_{<k}a_k)\Big)^2\\
&\le\sum_{k=1}^{n}\sum_{x_{<k}}\mu(x_{<k}|a_{<k})\sum_{x_k}\mu(x_k|ax_{<k}a_k)\ln\frac{\mu(x_k|ax_{<k}a_k)}{\xi(x_k|ax_{<k}a_k)} && \text{[Lemma 3]}\\
&=\sum_{k=1}^{n}\sum_{x_{1:k}}\mu(x_{1:k}|a_{1:k})\ln\frac{\mu(x_k|ax_{<k}a_k)}{\xi(x_k|ax_{<k}a_k)} && \text{[Equation (12)]}\\
&=\sum_{k=1}^{n}\sum_{x_{1:k}}\Big(\sum_{x_{k+1:n}}\mu(x_{1:n}|a_{1:n})\Big)\ln\frac{\mu(x_k|ax_{<k}a_k)}{\xi(x_k|ax_{<k}a_k)} && \text{[Equation (11)]}\\
&=\sum_{k=1}^{n}\sum_{x_{1:n}}\mu(x_{1:n}|a_{1:n})\ln\frac{\mu(x_k|ax_{<k}a_k)}{\xi(x_k|ax_{<k}a_k)}\\
&=\sum_{x_{1:n}}\mu(x_{1:n}|a_{1:n})\sum_{k=1}^{n}\ln\frac{\mu(x_k|ax_{<k}a_k)}{\xi(x_k|ax_{<k}a_k)}\\
&=\sum_{x_{1:n}}\mu(x_{1:n}|a_{1:n})\ln\frac{\mu(x_{1:n}|a_{1:n})}{\xi(x_{1:n}|a_{1:n})} && \text{[Equation (13)]}\\
&=\sum_{x_{1:n}}\mu(x_{1:n}|a_{1:n})\ln\left[\frac{\mu(x_{1:n}|a_{1:n})}{\rho(x_{1:n}|a_{1:n})}\frac{\rho(x_{1:n}|a_{1:n})}{\xi(x_{1:n}|a_{1:n})}\right] && \text{[arbitrary } \rho\in\mathcal{M}\text{]}\\
&=\sum_{x_{1:n}}\mu(x_{1:n}|a_{1:n})\ln\frac{\mu(x_{1:n}|a_{1:n})}{\rho(x_{1:n}|a_{1:n})}+\sum_{x_{1:n}}\mu(x_{1:n}|a_{1:n})\ln\frac{\rho(x_{1:n}|a_{1:n})}{\xi(x_{1:n}|a_{1:n})}\\
&\le D_{1:n}(\mu\,\|\,\rho)+\sum_{x_{1:n}}\mu(x_{1:n}|a_{1:n})\ln\frac{\rho(x_{1:n}|a_{1:n})}{w_0^\rho\rho(x_{1:n}|a_{1:n})} && \text{[Definition 5]}\\
&=D_{1:n}(\mu\,\|\,\rho)-\ln w_0^\rho.
\end{aligned}
$$

Since the inequality holds for arbitrary $\rho\in\mathcal{M}$, it holds for the minimising ρ. □

In Theorem 15, take the supremum over n in the r.h.s and then the limit $n\to\infty$ on the l.h.s. If $\sup_n D_{1:n}(\mu\,\|\,\rho)<\infty$ for the minimising ρ, the infinite sum on the l.h.s can only be finite if $\xi(x_k|ax_{<k}a_k)$ converges sufficiently fast to $\mu(x_k|ax_{<k}a_k)$ for $k\to\infty$ with probability 1, hence ξ predicts μ with rapid convergence.

AIXI: The Universal Bayesian Agent Theorem 15 motivates the construction of Bayesian agents that use rich model classes. The AIXI agent can be seen as the limiting case of this viewpoint, by using the largest model class expressible on a Turing machine.

Note that AIXI can handle stochastic environments since Equation (10) can be shown to be formally equivalent to

$$a_t^* = \arg\max_{a_t} \sum_{o_t r_t} \ldots \max_{a_{t+m}} \sum_{o_{t+m} r_{t+m}} [r_t + \cdots + r_{t+m}] \sum_{\rho \in \mathcal{M}_U} 2^{-K(\rho)} \rho(x_{1:t+m} | a_{1:t+m}), \tag{22}$$

where $\rho(x_{1:t+m}|a_1 \ldots a_{t+m})$ is the probability of observing $x_1 x_2 \ldots x_{t+m}$ given actions $a_1 a_2 \ldots a_{t+m}$, class $\mathcal{M}_U$ consists of all enumerable chronological semimeasures, which includes all computable ρ, and $K(\rho)$ denotes the Kolmogorov complexity of ρ with respect to U. In the case where the environment is a computable function and

$$\xi_U(x_{1:t}|a_{1:t}) := \sum_{\rho \in \mathcal{M}_U} 2^{-K(\rho)} \rho(x_{1:t}|a_{1:t}), \tag{23}$$

Theorem 15 shows for all $n \in \mathbb{N}$ and for all $a_{1:n}$,

$$\sum_{k=1}^{n} \sum_{x_{1:k}} \mu(x_{<k}|a_{<k}) \Big(\mu(x_k|a x_{<k} a_k) - \xi_U(x_k|a x_{<k} a_k) \Big)^2 \leq K(\mu) \ln 2. \tag{24}$$

6 Approximations and Applications

Solomonoff's universal probability M as well as Kolmogorov complexity K are not computable, hence need to be approximated in practice. Levin complexity Kt is a down-scaled computable variant of K with nice theoretical properties, and the minimum description length principle(MDL) is a effective model selection principle based on Ockham's razor quantifying complexity using practical compressors. K and M have also been used to well-define the clustering and the AI problem.

6.1 Minimum Description Length Principle

$$\widehat{\nu} = \arg\min_{\nu} \{K(x|H_\nu) + K(H_\nu) : \nu \in \mathcal{M}\} = \max\{w_\nu \nu(x) : \nu \in \mathcal{M}\} \tag{25}$$

MDL converges, but speed can be exponentially worse than Bayes.

Theorem 16.

$$\sum_{t=1}^{\infty} \mathbb{E} \left[\sum_{x_t \in \mathcal{X}} \left(\widehat{\nu}(x_t|x_{<t}) - \mu(x_t|x_{<t}) \right)^2 \right] \overset{+}{\leq} 8 w_\mu^{-1} \tag{26}$$

Practical MDL achieves computational feasibility by restricting the hypotheses, which are the methods of data compression, to probabilistic Shannon-Fano based encoding schemes.

6.2 Resource Bounded Complexity and Prior

Levin Complexity Levin Complexity is a direct variant of algorithmic complexity which is computable because it bounds the resources available to the execution.

$$Kt(x) \;=\; \min_p\{|p| + \log t(p,x) \;:\; U(p) = x\}$$

Within a (typically large) factor, Levin search is the fastest algorithm for inverting a function $g : Y \to X$, if g can be evaluated quickly. Given x, an inversion algorithm p tries to find a $y \in Y$, called g-witness for x, with $g(y) = x$.

Levin Search LEVIN: Run all $\{p : |p| \leq i\}$ for $2^{i-|p|}$ steps in phase $i = 1, 2, 3, \ldots$ until it has inverted g on x.

Theorem 17. *All strings $\{x : Kt(x) \stackrel{+}{\leq} k\}$ can be generated and tested in 2^{k+1} steps.*

Proof. $\sum_{i=1}^{k} \sum_{|p|\leq i} 2^{i-|p|} = \sum_{U(p)\downarrow} 2^{-|p|} \sum_{i=1}^{k} 2^i \leq 2^{k+1}$ □

The time needed is $t^L(x) \leq 2^{K(k)+O(1)} t^+_{p_k}(x)$, where $t^+_{p_k}(x)$ is the runtime of $p_k(x)$ plus the time to verify the correctness of the result ($g(p(x)) = x$) by a *known* implementation for g.

There is another algorithm that can do nearly as good as Levin search.

SIMPLE: p_1 is run every second step, p_2 every second step in the remaining unused steps, p_3 every second step in the remaining unused steps, and so forth, i.e. according to the sequence of indices $121312141213121512\ldots$.

Levin search can be modified to handle time-limited optimization problems as well.

Definition 6. *$p \to x$ if p computes output starting with x, while no prefix of p outputs x. $p \to_i x$ if $p \to x$ in phase i of* LEVIN.

Definition 7 (Speed Prior).

$$S^n(x) := \sum_{i=1}^{n} 2^{-i} S_i(x)$$
$$S(x) := \lim_{n\to\infty} S^n(x)$$

where

$$S_i(\epsilon) = 1; \quad S_i(x) = \sum_{p\to_i x} 2^{-|p|}$$

The speed prior $S(x)$ defined in [Sch02] is also a semimeasure, and approximates $M(x)$.

6.3 Universal Similarity Measure

For objects x and y, the similarity metric is defined by symmetrizing and normalizing $K(x|y)$ as follows:

$$d(x,y) \;:=\; \frac{\max\{K(x|y),\, K(y|x)\}}{\max\{K(x),\, K(y)\}} \approx \frac{K(xy) - min\{K(x), K(y)\}}{max\{K(x), K(y)\}}$$

Although the Kolmogorov complexity is incomputable, it is possible to achieve an effective approximation of this metric by approximating K with a good compressor such as Lempel-Ziv, gzip or bzip, or it can be approximated by just counting the frequency p in the world wide web with Google and then use $-\log p$. The similarity metric defined above can lead to excellent classification in many domains.

For example, if there are n objects $x_1, x_2, \cdots, x_n$, we can figure out the similarity matrix $M_{ij} = \Big(d(x_i, x_j)\Big)_{ij}$ and then cluster similar objects.

7 Discussion

Our predictions of the future are dependent on a lifetime of observations. One of the problems with Solomonoff induction is that relevant background knowledge is not explicitly accounted for. There are two ways to modify Solomonoff induction to account for prior background knowledge y.

The first method is to resort to the conditional Kolmogorov complexity $K(v|y)$. The second method is to prefix the observation sequence x, which we wish to predict, with the prior knowledge y. We are therefore predicting the continuation of the sequence yx.

8 Conclusion

Advantages and Problems The following are the advantages (+) and problems (−) of Solomonoff's approach:

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

AIXI as a Principle As the AIXI agent is only asymptotically computable, it is by no means an algorithmic solution to the general reinforcement learning problem. Rather it is best understood as a Bayesian *optimality notion* for decision making in general unknown environments.

References

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

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

[Jay03] E. T. Jaynes. Probability Theory: The Logic of Science. *Cambridge University Press, Cambridge, MA*, 2003.

[LV08] M. Li and P. M. B. Vitányi. An Introduction to Kolmogorov Complexity and its Applications. *Springer, Berlin, 3rd edition*, 2008.

[Odi99] P.G. Odifreddi. Classical Recursion Theory. Volume2. *Elseviser*, 1999.

[Sch02] J. Schmidhuber. The speed prior: A new simplicity measure yielding near-optimal computable predictions. In *Proc. 15th Conf. on Computational Learning Theory (COLT'02)*, volume 2375 of *LNAI*, pages 216–228, Sydney, 2002. Springer, Berlin.

[Sol64] R. J. Solomonoff. A formal theory of inductive inference: Parts 1 and 2. *Information and Control*, 7:1–22 and 224–254, 1964.

[Sol78] R. J. Solomonoff. Complexity-based induction systems: Comparisons and convergence theorems. *IEEE Transactions on Information Theory*, IT-24:422–432, 1978.

[Sol03] R. J. Solomonoff. Progress In Incremental Machine Learning. TR IDSIA-16-03. 2003

[VNH11] J. Veness, K. S. Ng, M. Hutter, W. Uther, and D. Silver. A Monte Carlo AIXI approximation. *Journal of Artificial Intelligence Research*, 40:95–142, 2011.