An Introduction to Solomonoff Induction With an Application to Scientific Confirmation Theory

by Daniel Alexander Herrmann

Submitted in partial fulfillment of the requirements for the
Degree of
Bachelor of Arts and Sciences
Quest University Canada

and pertaining to the Question

How should we create artificial general intelligence?

June 4, 2017

Concentration Mentor	Daniel Herrmann

Acknowledgments

I would like to thank my mentor Darcy Otto for his frequent critical and constructive feedback, his attention to detail (especially in typographical matters), and his willingness to talk through philosophical problems with me. I would also like to thank him for the immense amount of support he has given to me these past two (really four) years. Not only has he given me large amounts of his time, energy, and thought, but he also treated me like a scholar in my own right. He has played a key rôle in fostering my intellectual development. For this, and more, I thank him.

I would like to thank my parents, Patrick and Olga Herrmann. They shared their love of learning and philosophy with me from an early age, and are always excited to talk with me about my studies. They have taught me integrity and perseverance, and it is with these traits that I approach my academic work.

A number of other people have also contributed to the success of this Keystone. I want to thank Mackenzie Marcotte for discussing the subject of my Keystone in detail, and for his willingness to dig deep into the mathematics. I want to thank Andrew Hamilton for meeting with me to discuss particle physics, statistics, and philosophy of science, and for reading a draft of my Keystone. His insights were invaluable. I want to thank Max Notarangelo and Michael Geuenich for reading over full early drafts and giving me feedback. I want to thank Neder Gatmon-Segal, Shiyu Huang, and Athena White for reading the Bayesian sequence prediction section and providing suggestions to make it more accessible. Finally, I want to thank my close friends not yet mentioned: Josie Bauman, Barbara Fernandes, Colin Wilt, Lars Laichter, and Stuart Lantz. They supported me when I needed it, and made me laugh the rest of the time.

1 Introduction

Experimental science, and broadly speaking inductive reasoning in general, can be viewed as aligning our beliefs about the world with the way the world actually is. Our intent is that the more science we do, the more accurate our beliefs will become. We generally think that the closer our theory (also called model or hypothesis) about the laws that govern our world is to the actual laws that govern our world, the better aligned our beliefs are. This approach is called model identification [Hut07], because the goal is to identify models closer and closer to the actual environment. However, there is another way to measure the success of inductive reasoning—quality of prediction. Instead of placing importance on the model, we can place importance on our prediction. The approach in which the quality of belief is measured by quality of prediction is called prequential or transductive [Hut07]. For example, instead of identifying the laws that govern the motion of the Sun with respect to Earth, we care about the accuracy of our prediction of where the Sun will be relative to the Earth.

One of the most well-studied frameworks for inductive reasoning is the Bayesian framework [Hut07]. Based on Bayes' theorem, it describes the best way for agents to update their beliefs (whether about models or the future) given new information. However, it does not describe what initial beliefs agents should hold (an agent's starting beliefs are called a "prior"). In the 1960s Ray Solomonoff proposed and argued for a universally optimal set of prior beliefs, which we now call Solomonoff's prior [Hut07]. This universal prior used in combination with the Bayesian sequence prediction framework is called Solomonoff induction, and is a strong candidate for the optimal way to learn about the world from past data (experience). It therefore has consequences for a wide range of fields including philosophy of science, artificial intelligence [Hut09], cognitive science [CV03], and information theory [CT12].

In [Hut09] Hutter gives a comprehensive survey of the open problems in universal induction, including the Zero Prior Problem.[1] Although he gives a brief sketch of a likely solution using Solomonoff induction in [Hut09] and [Hut07], he writes that the "discussion needs to be rolled out much further, say, at least one generally accessible article per one allegedly open problem" ([Hut09], p. 2).

Thus, there are three main goals of this paper: (1) to introduce readers to the Bayesian sequence prediction framework, (2) to introduce readers to Solomonoff's prior (and provide a brief summary of arguments for why it is universally optimal), and (3) to demonstrate how Solomonoff's prior solves the Zero Prior Problem in scientific confirmation theory. (3) will help strengthen the claim that Solomonoff's prior (a set of initial beliefs) is the universally optimal way to predict sequences (read "the future") by exploring a case in which it has a clear advantage over other prior beliefs. Specifically, §5.3 explores in detail how Solomonoff induction compares to a continuous hypothesis class. Furthermore,

[1] The Zero Prior Problem will be discussed in great length in §3. In general, it is the problem that if one starts out with a belief of 0 in any hypothesis (§2.3), then no matter how much evidence one observes in favour of the hypothesis, one will always be certain that the hypothesis is incorrect. In particular, this affects hypothesis classes (§2.4) with continuous parameters.

(3) makes this paper particularly interesting for scientists, as it explores the impact that different methodologies have on the claims science can make.

Although many of these ideas are not new, the method in which I will introduce these ideas will make them more accessible to a wider audience of generally educated people. In particular, I hope that it will be a good resource for scientists, for whom papers about methodology are infrequently written.

2 The Bayesian Framework

The Bayesian framework for sequence prediction allows one to predict the future (of a sequence) in provably optimal ways [Hut07]. In this section I will introduce the key parts of the Bayesian framework. The power of the framework comes from its property that if the true environment is contained in the hypothesis class then the framework's prediction will converge to the optimal prediction relatively quickly. For the sake of brevity, I omit these proofs. For the proofs see [Hut07] and [LV97].

The goal of the framework is to predict the next symbol in a sequence. The sequence could be a sequence of numbers (1, 1, 2, 3, 5, ...), observations of the colour of ravens (black, black, black, ...), the money someone spends each day ($100.00, $96.73, $102.55, ...), *et cetera*. Informally, we measure the success of our predictions by the fraction of correct predictions we make (relative to the fraction of correct predictions if we knew the laws governing the sequence[2]).

2.1 The Sequence

Consider having observed a sequence of symbols (also called a string[3]) $x_{1:n} = x_1x_2, \ldots, x_n$ composed of only "1"s and "0"s (or any finite alphabet of symbols), of which we want to predict the x_{n+1th} symbol as well as possible. As an *in concreto* situation, imagine the binary sequence "101010" which represents the outcomes of the first six tosses of a coin, where heads is "1" and tails is "0". We want to predict, as well as possible, what the next symbol will be. If the coin is (somehow) deterministic, then after seeing enough of the sequence we would want to perfectly predict the next toss. If the coin is probabilistic (more likely with coins[4]), with a probability of 0.5 heads and 0.5 tails, then after seeing enough of the sequence the best we can reasonably expect is to get half of the predictions correct. However, if the coin had, for example, a 0.7 chance to come down heads and a 0.3 chance to come down tails, then we would hope to be able to correctly predict the result of the next toss about 70% of the time.

[2]This is μ prediction, which I describe in §2.5

[3]For the purposes of this paper, I assume that any sequence can be represented by a string.

[4]Strictly speaking, of course, this is not true. When I flip a coin, I apply certain forces in certain ways, which cause the coin to land on either heads or tails. If I were to flip the coin again in the exact same way the result would be the same. Thus, coins are deterministic. However, I find the example of the coin an intuitive way to understand probabilistic environments, because coins often seem probabalistic.

2.2 Environments

Informally speaking, an environment ν (also called model, measure, or hypothesis) is a possible world that can output a sequence. For example, consider a perfectly fair coin. This is an example of an environment. The coin is spun, and lands on either heads "1" or tails "0". Observing successive coin flips builds up a sequence, "101111001010...", of which we might want to predict the next element. In a more scientific example, environments are essentially the same as the models used in fields such as behavioral economics and statistical physics.

Environments can be either probabilistic or deterministic. An atom decaying is probabilistic, while the output of an idealized Turing machine is deterministic. The probability of observing an environment ν output a sequence x is $\nu(x)$. For deterministic environments, $\nu(x) = 1$ for exactly one sequence, and 0 for all others. For probabilistic environments $\nu(x)$ can be any real number between 0 and 1, and $\sum_x \nu(x) = 1$. For example, if there is a fair coin (call it η), then $\eta(1) = 0.5$, $\eta(10) = 0.25$, $\eta(100) = 0.125$, and so on.

2.3 Beliefs

When talking about beliefs in a Bayesian sense, one is talking about subjective probabilities. A subjective probability is a probability an agent gives to some event [Hut05]. This is contrasted with an objective probability of an event (for example, the probability that an atom will decay at some point) [Hut05]. For example, consider the sequence "12345". Without any additional information, what is the likely continuation of the sequence? My guess is that you have a fairly high belief that the next number is 6, though if I told you it was 9, you wouldn't think it impossible. A Bayesian agent would say that you had a higher belief that it would be 6, but you were not certain. This is the sense in which I use the term "belief" (which is equivalent to a subjective probability).

Above, it might have seemed that I used the term"belief" in an odd way. We usually think of belief as binary—you either believe something or you do not. However, belief in the Bayesian sense is better characterized as a level of confidence. When a Bayesian says that she has a higher belief in some proposition X than some other proposition Y, then what she is saying is that she has a higher confidence that X is true.

Thus, in the situation in which I am spinning what I know to be a perfectly fair coin, my belief that it will come up heads is 0.5 (on a scale of 0–1), and my belief that it will come up tails is also 0.5. If the coin is truly fair, then the best I can do is have my belief about the next digit (0.5 that the next digit is "1") match up with the objective probability of the environment (0.5 that the next digit is "1").

In both of the above examples, we examined beliefs in outcomes (the next symbol in a sequence). However, we can also have beliefs in environments. When a physicist talks about being fairly confident in general relativity, a Bayesian would say he is describing his level of belief in the model of general relativity. I denote the (Bayesian) belief in an environment ν with w_ν. Beliefs can be any

real number between 0 (believing that something is impossible) to 1 (thinking that something is necessary).

2.4 The Hypothesis Class M

We start with a set $M = \{\nu_1, \nu_2, \nu_3, \ldots\}$ of environments, which we believe holds the true environment (the real world), μ. M is called the hypothesis class. These environments can be either probabilistic or deterministic. For example, if we are trying to predict coin flips, our hypothesis class migh contain two environments, ν_1 and ν_2. ν_1 might be a fair coin, and ν_2 might be a coin weighted such that it comes down heads 80% of the time.

We also need to have a belief in each environment ν, by which I mean we have a belief that ν is the true environment. As in §2.3, let $w_\nu > 0$ be our prior belief that ν is the true environment (before we have seen any of the sequence we want to predict), and where all of our beliefs in each environment added together are less than or equal to 1 ($\sum_{\nu \in M} w_\nu \leq 1$). The set of environments M, combined with our belief in each environment ν in M, is the starting point of the framework, called the "prior."

For example, consider the hypothesis class with the two models from above. We might start off with a belief of 0.5 in ν_1 and 0.5 in ν_2. This would be our prior.

Unlike how we usually think of science, in which there are scientists working on new models to explain observations better than old models, the Bayesian sequence prediction framework has a full set of models at the beginning. At no point when using the framework do we add a model to M. Although this might seem very limiting and against common sense (how do we know that the true model μ is in M?), we will see in §4 how Solomonoff's prior provides a model class large enough for any sequence prediction tasks (thus making it universal).

2.5 μ Prediction

Given our hypothesis class M, our prior beliefs in each hypothesis $\nu \in M$, and a sequence x, we can begin predicting the sequence. In order to understand how we can use M and our prior beliefs to predict the sequence, it is useful to consider how we would optimally predict the sequence if we knew the true environment μ. Bayes' theorem gives us the tools we need. Bayes' theorem is

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

where A and B are different events, the (prior) probability that event A will happen is $P(A)$, and $P(A|B)$ is the probability of event A given event B.

If we know the true environment μ and we have observed the first n symbols of the sequence, then we can use Bayes' theorem to predict the $n + 1^{\text{th}}$ symbol. What we want to know is $P(x_{1:n+1}|x_{1:n}) = \mu(x_{1:n+1}|x_{1:n})$, or the so-called μ-probability of the the sequence we have observed with an additional $n + 1^{\text{th}}$ symbol after having observed the first n symbols. Replacing all of the Ps with

μs and the events A and B with the string the probability of which we want to know and the observed string respectively in Bayes' theorem we get

$$\mu(x_{1:n+1}|x_{1:n}) = \frac{\mu(x_{1:n}|x_{1:n+1})\mu(x_{1:n+1})}{\mu(x_{1:n})} = \frac{\mu(x_{1:n+1})}{\mu(x_{1:n})}$$

in which $\mu(x_{1:n}|x_{1:n+1})$ disappears, as it is clearly 1. Given perfect information in the form of the true environment μ, this is the best prediction possible.

For example, imagine that μ is a fair coin. We have observed the sequence "10010", and we want to know how likely the next digit is to be a "1". If we plug this into the above equation above we get:

$$\mu(100101|10010) = \frac{\mu(100101)}{\mu(10010)} = \frac{0.5^6}{0.5^5} = 0.5$$

This is exactly what we would expect. Since each coin flip is independent, the chance that the next digit will be a "1" is 0.5.

2.6 Sequence Prediction

We saw above how to find the probability of the next digit in a sequence if we knew the true environment μ. How should we predict if we do not know μ? The key lies in the fact that μ is known[5] to belong to our hypothesis class M (§2.4). In order to use M to give us a prior belief in a certain sequence x, we take the probability $\nu(x)$ for each environment ν in M, multiply that probability by our prior belief w_ν in ν, and add the result for each ν in M. The result

$$\xi(x) := \sum_{\nu \in M} w_\nu \nu(x)$$

represents our initial subjective probability (belief) of sequence x, as it is based on our beliefs w_ν in each environment ν in M.

For example, consider our coin model class from §2.4. We have two hypotheses, ν_1 and ν_2. ν_1 is a fair coin, whereas ν_2 lands heads 80% of the time. We have a belief of 0.5 in each hypothesis. Thus, in this case

$$\xi(1) = \sum_{\nu \in M} w_\nu \nu(1) = 0.5 * 0.5 + 0.5 * 0.8 = 0.65$$

We have a 0.65 belief that the coin will land heads.

A key property of ξ is that it dominates all environments ν in M (including μ), which means that:

$$\forall x \forall \nu \in M \; \xi(x) \geq w_\nu \nu(x)$$

This is because $\xi(x)$ is a sum of each $w_\nu \nu(x)$ and probabilities are never negative. Thus $\xi(x)$ will be at least $w_\mu \mu(x)$, and greater if there is any other nonzero

[5] This may seem to the reader like it is too large an assumption—how do we know μ is in M? We will see in §4.1 how Solomonoff's prior solves this issue.

$w_\nu \nu(x)$. This dominance of $\xi(x)$ over all its environments will be very important later.

Just as we were able to use Bayes' theorem to predict the likelihood of the next digit using μ, we can use it to predict using ξ:

$$\xi(x_{1:n+1}|x_{1:n}) = \frac{\xi(x_{1:n+1})}{\xi(x_{1:n})}$$

2.7 Belief Updating

The last piece of the Bayesian sequence prediction framework is belief update. Not only do we want to predict the likely continuation of the sequence given our prior beliefs (whatever they may be), but we also want to use the information contained in the sequence we have observed to update our beliefs in each hypothesis. Intuitively, if I have flipped a coin 100 times and observed it to come down heads each time, I should have a higher belief that the coin is unfair than when I started.

Luckily, we can use Bayes' theorem to update our beliefs as well. Whereas before what we wanted was our belief in a sequence given an environment or set of environments (for example, μ-prediction in §2.5), what we want now is our belief in an environment given a sequence. We get

$$P(\nu|x_{1:n}) = \frac{P(x_{1:n}|\nu)P(\nu)}{P(x_{1:n})} = \frac{\nu(x_{1:n})w_\nu}{\xi(x_{1:n})} = w_\nu(x_{1:n})$$

where $w_\nu(x_{1:n})$ is our subjective probability in environment ν after having seen the first n symbols of x. In general, w_ν is the current belief in environment ν (after having seen any arbitrary amount of x). Every time we see a new digit in the sequence we update our beliefs in each environment, which we can then use to find the likely continuation of the sequence as in §2.6.

For example, consider observing a coin land heads in our coin example from §2.4 and §2.6. We want to update our belief in ν_1, the hypothesis that the coin is fair. We get

$$P(\nu|1) = \frac{\nu(1)w_{\nu_1}}{\xi(1)} = \frac{0.5 * 0.5}{0.65} \approx 0.385$$

Thus, after seeing the coin land heads, we can update our beliefs in each environment.

3 The Zero Prior Problem

3.1 Universal Hypotheses

Science aims at making universal hypotheses, and increasing confidence in these hypotheses by observing examples of what they predict. For example, one might have a hypothesis that all ravens are black or that all electrons have a rest mass of 9.109×10^{-31}kg. These are both examples of universal hypotheses because

they make claims about all members of a certain class of objects.[6] Furthermore, it often only takes a relatively small[7] number of examples with no counter-examples for scientists to be fairly confident in the hypothesis. For example, although scientists have only calculated the mass of a small fraction of the amount of electrons in the observable universe, they are fairly confident that all electrons have the same mass.

3.2 Universal Hypotheses in Sequence Prediction

Consider a scientist observing a sequence of ravens. This particular scientist is interested in the question, what is the fraction of ravens which are black? The scientist considers all the possible real numbers between 0 and 1 (inclusive) as decimal representations of the fraction of ravens that are black. If the scientist were using a model identification approach (as described in §1), she would care about how high her belief in the true environment (which she does not know) is. For example, if it were the case that all ravens are black, then what she is hoping is that the more she observes the more confident she becomes in the hypothesis "all ravens are black." On the contrary, if half of the ravens are black, then she would hope that the more ravens she sees the more confident she is in the hypothesis "half of all ravens are black."

However, she could also view take a prequential (§1) approach. Consider a sequence of numbers $x_1x_2x_3x_4\ldots$ where x_n is the n^{th} example of a raven she has seen, with each possible value for x_n corresponding to whether or not the raven was black. If she has observed n ravens, each digit in the sequence might have the value either "0" or "1", where "0" means "not-black" and "1" means "black." In the prequential approach, she wants to predict the probability that the $n+1^{\text{th}}$ digit is a "1", or, in words, that the next raven she observes will be black. In the prequential approach, the universal hypothesis "all ravens are black" is phrased only in terms of observable quantities. For example, it might be phrased as "every raven I will ever observe will be black." In terms of the sequence, we can think of this as the hypothesis that every digit in the sequence will be a "1."

3.3 Problem Setup

Broadly speaking, the Zero Prior Problem is the problem that, if the initial belief in X is zero, then no matter how much evidence is observed in favour of X the agent will still have a belief of zero in X. Informally, if X is thought to be impossible, then an agent's belief in X can never increase. In order to clarify this problem I consider a simple example.

Imagine the scientist in § 3.2 as she observes the ravens. She is curious about the fraction of ravens that are black. She is using the Bayesian sequence

[6] Another way to think of a universal hypothesis is a statement in formal logic that has the universal quantifier "∀".

[7] The number or examples is relatively small with respect to the size of the population. For a more precise notion of "relatively small" see §3.5.

prediction framework (as in §2) and thus needs a hypothesis class M. If she knows that the order in which she sees ravens is random,[8] then it seems natural for her hypothesis class to consider all environments ν_θ where $\nu_\theta(1) = \theta$, in which θ is real number between 0 and 1, inclusive. Thus,

$$M = \{\nu_\theta : \theta \in \mathbb{R} \ \& \ 0 \leq \theta \geq 1\}$$

is her model class. θ is the fraction of ravens that are black. One can imagine a similar case with a person drawing coloured balls out of a (infinitely) large bag. The person is curious about the fraction of balls that are black. Without any other prior information, the person considers all real numbers between 0 and 1, inclusive, as candidates for the fraction of balls that are black.

In order to deal with this continuous hypothesis space, we need to introduce a slight modification to the Bayesian sequence prediction framework as give in §2. The framework originally presented deals only with countable hypotheses, whereas the scientist wants to consider all real numbers between 0 and 1, which are uncountably infinite. Above, when we were working with countable hypotheses, we described our belief in a sequence x as

$$\xi(x) := \sum_{\nu \in M} w_\nu \nu(x)$$

which sums over our belief in each hypothesis multiplied by the probability of x under the hypothesis. However, with uncountable hypotheses, we cannot sum. Instead, we must take the integral. Thus, in a continuous environment class with a single parameter ranging from a to b, we get

$$\xi(x) := \int_a^b w(\theta)\nu_\theta(x)d\theta$$

In the case of our scientist, a is 0 and b is 1. w_ν is replaced with $w(\theta)$ because, in order to deal with continuous parameters, we must use a continuous weight density instead of a probability.[9]

Furthermore, it seems reasonable that before she has made any observations she should have an equal degree of belief in each environment. This choice of uniform prior belief is argued for by Laplace [Lap20], and seems quite natural. Without any other information, why should she be more confident in one hypothesis over another? Thus, as in §2.4, we want our beliefs in each environment to sum to 1 ($\sum_{\nu \in M} w_\nu \leq 1$). In order to deal with the continuity of the hypothesis space, we ingrate instead of sum to get

$$\int_a^b w(\theta)d\theta = 1$$

[8] This is generally in line with scientific practice—scientists try to remove bias by sampling randomly from a population.

[9] We use a density across multiple (all) hypothesis instead of a probability assigned to a single hypothesis because we are operating in a continuous space. For a more detailed assessment of how this interacts with other parts of the Bayesian framework, see [Hut07].

[Hut03] shows that this framework for continuous hypothesis classes also converges quickly to the true environment. Other than these differences, the continuous framework functions the same as the framework in §2.

3.4 The Problem

As stated in § 3.3, in general, the Zero Prior Problem is the problem that if you start with a prior belief of zero in something, then no matter how much favourable evidence you observe, your posterior belief (the belief after obtaining new information) will still be zero.

We see this in the scientist's setup above. Although the prior belief that the environment lies in the range p to q (where p and q are both real numbers between 0 and 1) is non-zero, the prior belief that any particular number θ is the fraction of ravens which are black is zero. Formally,

$$\int_{\theta}^{\theta} w(\theta) d\theta = 0$$

for all θ. Thus, even though the scientist can be confident that the true fraction of black ravens falls in a range of values, she always has a confidence of zero that it falls on any particular value. Thus, the hypotheses "all ravens are black" or even (by extension) "all electrons have a mass of 9.109×10^{-31}kg" can never have any non-zero degree of confidence assigned to them.

Above, the problem was formulated in a model identification framework, because we were concerned that we could never have any degree of belief in a particular model. We might wonder if using a prequential approach avoids the problem, by referring only to observations. As stated above, we might express the hypothesis "all ravens are black" as "every raven I will ever observe will be black." Referring to the sequence, this can be expressed as "every digit of the infinite sequence x will be '1'."

As [Bay63] showed, in the setup above

$$\xi(x) = \int_0^1 w(\theta)\nu_\theta(x) d\theta = \frac{n_1! n_0!}{(n+1)!}$$

where n_1 is the number of 1s in the sequence x, n_0 is the number of "0"s in x, and $n = n_0 + n_1$. Recall that $\xi(x)$ is the prior probability that we will observe sequence x. Thus, the probability of a sequence of n "1"s (denoted at 1^n) is

$$\xi(1^n) = \frac{n_1! n_0!}{(n+1)!} = \frac{n_1!}{(n+1)!} = \frac{1}{(n+1)}$$

Thus, if we use the sequence prediction setup as in § 2.6, the probability of observing another k "1"s after we have observed n "1"s is

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

which, if k is fixed as a certain number, will converge to 1 as the number n of observations grows.

However, the scientist is not interested in a fixed k, but in an infinite k—she doesn't only want to be confident that the next 100 (fixed $k = 100$) ravens will be black, but she wants to be confident that *all* the ravens she could ever see will be black. In this case, she would have to let k approach infinity. Clearly,

$$lim_{k\to\infty}\xi(1^k|1^n) = lim_{k\to\infty}\frac{n+1}{n+k+1} = 0$$

for any number of observations n. Thus even in a prequential framework there is a Zero Prior Problem—the scientist always has zero confidence that all ravens are black.

3.5 Finite Population Size

As [Hut07] pointed out, one might object to the problem and say that the issue is the assumed infinite population size. However, as he further points out, even if we restrict ourselves to thinking of finite population sizes, there are still issues. If N is the total population of ravens, and our scientist has observed n ravens, all of which were black, then

$$\xi(1^{N-n}|1^n) = \frac{n+1}{N+1}$$

As [Hut] notes, the confidence of seeing a further $N - n$ black ravens is only close to 1 (a high degree of belief) if a large fraction of the population has been observed. This defies scientific practice, as scientists are often fairly confident in universal claims of which they have seen a relatively low[10] number of examples.

3.6 Solution Criteria

The Zero Prior Problem poses a serious challenge to the confirmation of universal hypotheses. A solution to this problem must meet certain criteria:

1. It must allow scientists to have a non-zero posterior belief in universal hypotheses.

2. It must allow scientists to be reasonably confident in universal hypotheses without having observed most of the objects of inquiry.

3. In general, it must perform approximately as well or better than the initial attempt as described in §3.3.

[10]The number of examples, n, is small relative to the population size, N. The equation is this section quantifies this: when n is not very close to N, we are not very confident. Scientists, however, are often confident in their hypotheses when n is a small fraction of N. For example, although scientists have only calculated the mass of a small fraction of the amount of electrons in the observable universe, they are fairly confident that all electrons have the same mass.

4 Solomonoff Induction

The Bayesian sequence prediction framework as described in §2 tells one how to update one's beliefs in hypotheses based on new information. One starts with a set of environments, M, and a belief w_ν in each environment ν. However, there is still an open question—which environments should be in M, and how much should we believe each environment *a priori*?

We saw a partial attempted answer in section §3.3: in that specific case where environments are thought of as fractions of a population with a certain property, the scientist considered all real numbers between 0 and 1, and had equal *a priori* belief in each environment. However, as we saw in §3.4, this approach is not without problems. Thus, we must ask ourselves, can we do better?

In the 1960s Ray Solomonoff proposed and argued for a universally optimal hypotheses class M and an optimal way to assign an *a priori* degree of belief to each hypothesis ν in M [LV97]. In short, Solomonoff considers all possible computable hypotheses and assigns lower prior beliefs the more complex the hypothesis is. In this section I introduce Solomonoff's prior and the intuition behind it. Furthermore, I show how it can be used within the Bayesian sequence prediction framework with success in any environment.

4.1 Choosing a Universal M

As we saw in §2, the success of the framework depends on the true environment μ being one of the hypotheses in M. This must have left the reader with an uneasy feeling—how can we be certain that μ is in M, without any prior information about the system we are observing?

Solomonoff's approach to this question was to make M very large. The intuition behind this is that the more hypotheses you have in M, the less restrictive are your assumptions. For example, if one only considers environments that only ever print a finite number of "1"s, then one is not able to consider an environment that outputs an infinite number of "1"s.

To achieve this, Solomonoff used the class of every computable environment [Hut07] [LV97].[11] Thus Solomonoff's universal M^S includes all computable hypotheses.

4.2 Universal Prior Belief

Now that we have our universal set of hypotheses M^S we are in a position to assign a prior belief in each hypothesis $\nu \in M^S$. When deciding between hypotheses equally consistent with observed data scientists often favour the simpler of the two hypotheses. When they do this they are appealing to Occam's Razor, which states that the simplest hypothesis is the more likely one [Lau97]. This intuition lead Solomonoff to have lower prior beliefs the greater the complexity of a hypothesis.

[11] Although this was Solomonoff's original approach, M has since been extended by Levin to include other realted hypotheses. For a detailed account see [ZL70].

In order to formalize this intuition Solomonoff used Kolmogorov complexity to measure the complexity of a hypothesis. Kolmogorov complexity quantifies the complexity of a string. Consider, for example, the sequence

$$111111111111111111111111111111$$

This string does not seem that complex. On the other hand, the string

$$11001011100101000111101010 1001$$

seems more complex. We can think of the complexity of a string as the shortest description of that string. Imagine, for example, trying to send descriptions of both of the strings above to your friends such that they could reconstruct the string. To describe the first string you could write something like "thirty '1's in a row." However, to describe the second string, it seems like you would have to send the entire string itself (or something else close to that length)!

We can formalize the notion of this kind of complexity using a universal Turing machine. Let U be some universal Turing machine with a binary input tape. We can think of the input p as a program ran by U. The Kolmogorov complexity of a string[12] x is defined as

$$K(x) := \min_p \{l(p) : U(p) = x\}$$

where $l(p)$ is the number of bits in p [Hut, p. 9]. This lines up with the idea above of sending a description to a friend; our friend is analogous to U and our description of the string is analogous to p. The first string above can be generated by a fairly short program, whereas the second string would need a longer program.

One might point out that the length of the shortest program p that outputs x is dependent on the choice of U. This is correct. However, Kolmogorov complexity has the important property that, if we change our choice of universal Turing machine from U to a different universal Turing machine U', the Kolmogorov complexity of a string x increases at most by a constant independent of x.

This seems intuitive when considering an interpreter for a programming language L written in another language P. For all programs written in L that are significantly shorter than the obvious equivalent program written in P, one can simply write an L interpreter in P, and then input the L program. Thus, the shortest program written in P to output a given string will be longer than the shortest program written in L that outputs the same string by at most the length of the shortest L interpreter written in P.

Solomonoff used Kolmogorov complexity to generate a prior belief biased towards simple models [LV97]. He used the Kolmogorov complexity of an environment ν to formalize the notions of complexity and simplicity. He wanted the

[12]Technically, in order to receive many of the benefits of using Kolmogorov complexity, Solomonoff induction uses a modified version called *prefix Kolmogorov complexity*, based on a prefix universal Turing machine. However, the subtleties of this distinction is beyond the scope and need of this paper. For an excellent in-depth account of prefix Kolmogorov complexity see [LV97].

belief in a hypothesis to decrease the greater the complexity of the hypothesis. Thus, the prior belief w_ν in hypothesis ν is $2^{-K(\nu)}$. Furthermore, Kolmogorov complexity has the property that $\sum_x 2^{-K(x)} \leq 1$, which satisfies the requirement for our initial set of beliefs as described in §2.4.

4.3 Solomonoff Sequence Prediction

Other than the specification of M^S and the initial beliefs, Solomonoff induction works the same as Bayesian sequence prediction as described in § 2. Thus, in Solomonoff induction,

$$\xi(x) := \sum_{\nu \in M^S} w_\nu \nu(x) = \sum_{\nu \in M} 2^{-K(\nu)} \nu(x)$$

Given that Bayesian sequence prediction works when the true environment μ is in M, and the large size of Solomonoff's prior, Solomonoff Induction will work in any case in which the sequence is drawn from any computable probability distribution [Hut07]. This is because it dominates all computable environments (see §2.6). This seems reasonable/sufficient for (mostly) all induction problems. It is in this sense that Solomonoff induction is universal. From herein, I will denote the $\xi(x)$ defined in this section as $\xi^S(x)$ to avoid confusion between $\xi^S(x)$ and other $\xi(x)$.

5 Solution to the Zero Prior Problem

As stated in § 3.6 there are three properties we want a solution to the Zero Prior Problem to possess. In this section I will show how Solomonoff's prior adapted for the formulation of the Zero Prior Problem in §3.4 fulfills these three requirements.

5.1 Non-Zero Beliefs in Universal Hypotheses

The first requirement was that scientists must be able to have non-zero posterior beliefs in universal hypotheses. The problem of a zero prior occurs in §3.4 because of the continuous hypothesis space. Solomonoff induction, by considering only computable (and thus countable) hypotheses, clearly does not have this issue. If we take the class of hypotheses given in §3.3 and remove all hypotheses based on a non-computable θ, then we are left with a countable set of hypotheses. Since Solomonoff's prior belief in each hypothesis ν is $2^{-K(\nu)}$, there is never a hypothesis under consideration with a Zero Prior Problem, including the hypothesis that all "ravens are black" ($\theta = 1$) since the number 1 is computable.

Thus, by assigning a non-zero prior belief to universal hypotheses, Solomonoff induction clearly fulfills the first requirement for a solution to the Zero Prior Problem.

5.2 High Confidence With Few Examples

The second requirement we would like a solution to have was that it must allow scientists to be reasonably confident in universal hypotheses without having observed most of the objects of inquiry. For example, although scientists have only calculated the mass of a small fraction of the amount of electrons in the observable universe, they are fairly confident that all electrons have the same mass.

Solomonoff induction is biased towards simpler hypotheses (using the Kolomogorov complexity formalization of simplicity and complexity). It assigns a higher prior belief the simpler a hypothesis is. Hypotheses such as "all ravens are black" and "all electrons have the same mass" are quite simple. The larger the prior belief in something (whether an environment or a sequence), the fewer the observations that must be made to move towards a high degree of belief in it. Thus, although the mathematical analysis is beyond the scope of this paper, we can reasonably expect that because Solomonoff induction is biased towards simple hypotheses, it allows scientists to be fairly confident in hypotheses after having seen only a small fraction of the population. This is consistent with scientific practice.

5.3 Solomonoff Induction Compared to Continuous ξ

The third requirement is that the solution must perform at least as well as the continuous framework given in §3.3. In other words, can Solomonoff induction, a framework with a countable model class, perform at least as well as a model with an uncountable hypothesis class in a continuous hypothesis space? What if the true environment μ is uncomputable?

Even if the model class contains uncomputable environments (for example, if $\mu(x)$ is uncomputable and in the model class), $\xi(x)$ itself is often computable because the integral across the continuous hypothesis class can be approximated to arbitrary precision by a computable function ([Hut07], p. 17). Thus, even if μ itself is uncomputable, the function actually used for prediction, ξ, is computable. Since, as shown in §4.3, Solomonoff's prior dominates all computable functions, it also dominates the ξ based on the continuous hypothesis class (which includes μ). Thus,

$$\xi^S(x) = \sum_{\nu \in M^S} 2^{-K(\nu)} \nu(x) \geq 2^{-K(\xi)} \xi(x)$$

because ξ is in M^S.

Thus, even if we are using a countable hypothesis class of only computable environments and μ is not computable, we can still predict roughly as well as a continuous class that contains μ. We can make this more explicit by introducing a formal notion of distance between probability functions [Hut03]. That is, we quantify the distance between the subjective probability $\xi(x_n|x_{1:n-1})$ and the objective probability $\mu(x_n|x_{1:n-1})$. We use the relative entropy between two environments ξ and μ at the tth element of the sequence defined as ([Hut05] p.

73):

$$d(x_{1:t-1}) := \sum_{x_t} \mu(x_t|x_{1:t-1}) ln \frac{\mu(x_t|x_{1:t-1})}{\xi(x_t|x_{1:t-1})}$$

This gives a distance between two environments at a specific time. In order to get the total distance between two environments we use an expectation function as defined in ([Hut07], p. 3):

$$E[f(x_{1:n})] = \sum_{x_{1:n}} \mu(x_{1:n}) f(x_{1:n})$$

where f is some function we are trying to minimize. Since we want to minimize the distance between ξ and μ at all times in the sequence, we define a total distance measure D_n as

$$D_n(\mu||\xi) := \sum_{t=1}^{n} E[d_t(x_{1:t-1})]$$

which can be reduced to

$$D_n(\mu||\xi) = E[ln \frac{\mu(x_{1:n})}{\xi(x_{1:n})}]$$

(as shown in [Hut07] p.6; [Hut05] p. 73).

Now that we have a distance function between two different environments, we can find the distance between the true, possibly uncomputable environment μ, and Solomonoff's $\xi^S(x)$, $D_n(\mu||\xi^S)$. We want ξ^S to do at least as well as the ξ based on the continuous model class. Thus, if $D_n(\mu||\xi^S)$ is not significantly different from $D_n(\mu||\xi)$, then we have satisfied the final requirement of the solution.

$$D_n(\mu||\xi^S) = E[ln \frac{\mu(x_{1:n})}{\xi^S(x_{1:n})}]$$

can be expanded easily to

$$E[ln \frac{\mu(x_{1:n})}{\xi(x_{1:n})} \frac{\xi(x_{1:n})}{\xi^S(x_{1:n})}] = E[ln \frac{\mu(x_{1:n})}{\xi(x_{1:n})} + ln \frac{\xi(x_{1:n})}{\xi^S(x_{1:n})}]$$

Since $E[f(x_{1:n})] = \sum_{x_{1:n}} \mu(x_{1:n}) f(x_{1:n})$ we can divide the one expectation measure into a sum of two:

$$E[ln \frac{\mu(x_{1:n})}{\xi(x_{1:n})} + ln \frac{\xi(x_{1:n})}{\xi^S(x_{1:n})}] = E[ln \frac{\mu(x_{1:n})}{\xi(x_{1:n})}] + E[ln \frac{\xi(x_{1:n})}{\xi^S(x_{1:n})}]$$

which means that

$$D_n(\mu||\xi^S) = E[ln \frac{\mu(x_{1:n})}{\xi(x_{1:n})}] + E[ln \frac{\xi(x_{1:n})}{\xi^S(x_{1:n})}]$$

This means that the difference between $D_n(\mu||\xi^S)$ and $D_n(\mu||\xi)$ is $E[ln \frac{\xi(x_{1:n})}{\xi^S(x_{1:n})}]$. If this difference is a constant, then we know that ξ^S is as good as ξ, except for

a constant. I show that this difference is a constant as follows. As shown earlier in this section, we know that

$$\xi^S(x) \geq 2^{-K(\xi)}\xi(x)$$

which we can rewrite as:

$$\xi^S(x)2^{K(\xi)} \geq \xi(x)$$

We can now substitute $\xi^S(x)2^{K(\xi)}$ for $\xi(x)$ in $E[ln\frac{\xi(x_{1:n})}{\xi^S(x_{1:n})}]$ to get:

$$E[ln\frac{\xi^S(x_{1:n})2^{K(x)}}{\xi^S(x_{1:n})}] = E[ln(2^{K(x)})] = E[K(\xi)ln2]$$

which is a constant. Remembering to account for the inequality we get

$$D_n(\mu||\xi^S) - D_n(\mu||\xi) \leq E[K(\xi)ln2] = K(\xi)ln2$$

showing that Solomonoff Induction performs as well when μ is uncomputable as the ξ based on the continuous model class, plus an additive constant. Furthermore, since $K(\xi)ln2$ is fixed and the n in D_n can grow arbitrarily large, ξ^S will perform poorer than ξ at only a finite number of points in the sequence.

Thus, Solomonoff induction fulfills the third and final requirement for it to be a solution to the Zero Prior Problem.

6 Conclusion

In this paper I have introduced readers to the Bayesian sequence prediction framework and Solomonoff induction. I have tried to show the power of the Bayesian sequence prediction framework—if the hypothesis class one uses contains the true environment, then one's subjective probability ξ will converge to the objective probability μ. Furthermore, I showed how Solomonoff induction is a strong contender for the optimal way to learn from past data. I demonstrated how it solves a problem concerning scientific confirmation theory within the Bayesian sequence prediction framework. I hope that by drawing out how Solomonoff induction solves the Zero Prior Problem this paper serves to strengthen the claim that Solomonoff induction is the universally optimal way to learn from past data. By giving a starting set of beliefs it completes the Bayesian framework, which describes the best way to update one's beliefs based on new information.

Although this paper focuses on how Solomonoff induction is relevant to science, it is also very relevant to the artificial intelligence (AI) project. Marcus Hutter has put forward a mathematically rigorous model of intelligence called "AIXI" [Hut05]. Hutter combines sequential decision theory and Solomonoff induction into one agent, which he argues makes optimal decisions based on optimal (learned) beliefs. Although AIXI is uncomputable, it is intended to serve as a goal for AI researchers. In this way, AIXI does for the AI field what Solomonoff induction does for science: it provides the ideal for which to strive.

References

[Bay63] Thomas Bayes. "Essay towards solving a problem in the doctrine of chances". In: *Philosophical Transactions of the Royal Society* 53 (1763). [Reprinted in *Biometrika*, 45, 296-315, 1958]., pp. 376–393.

[CV03] Nick Chater and Paul Vitányi. "Simplicity: a unifying principle in cognitive science?" In: *Trends in cognitive sciences* 7.1 (2003), pp. 19–22.

[CT12] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[Hut03] Marcus Hutter. "Optimality of universal Bayesian sequence prediction for general loss and alphabet". In: *Journal of Machine Learning Research* 4.Nov (2003), pp. 971–1000.

[Hut05] Marcus Hutter. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*. Berlin: Springer, 2005. ISBN: 3-540-22139-5. DOI: 10.1007/b138233.

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

[Hut09] Marcus Hutter. "Open problems in universal induction intelligence". In: *Algorithms* (2009), p. 906.

[Lap20] Pierre Simon de Laplace. *Théorie analytique des probabilités*. Vol. 7. Courcier, 1820.

[Lau97] Bernhard Lauth. "New Blades for Occam's Razor". In: *Erkenntnis (1975-)* 46.2 (1997), pp. 241–267. ISSN: 01650106, 15728420. URL: http://www.jstor.org/stable/20012762.

[LV97] Ming Li and Paul Vitanyi. *An introduction to Kolmogorov Complexity and its Applications*. Springer, 1997.

[ZL70] Alexander K Zvonkin and Leonid A Levin. "The complexity of finite objects and the development of the concepts of information and randomness by means of the theory of algorithms". In: *Russian Mathematical Surveys* 25.6 (1970), p. 83.