Extended Solomonoff Induction: Universal Induction With Infinitesimal Credences

Sam Korn

May 11, 2018

Abstract

Solomonoff induction, a theoretical model of universal induction, is defined solely in terms of classical, Bayesian probability. However, standard probabilistic formulations of partial belief often do not adequately address infinitesimal credences. I propose and formalize Extended Solomonoff Induction (ESI), an enrichment of Solmonoff induction that incorporates infinitesimal credences, and defend its advantages over other standard models of universal induction as well as other non-standard models built on the hyperreal numbers.

Keywords infinitesimal credence, universal induction, Solomonoff induction, hyperreals

Contents

1 Introduction

One of the central tasks to the field of modern artificial intelligence is the development of a system capable of learning. Machine learning algorithms have garnered immense popularity for their benefits to data analysis, with applications to marketing, natural language processing, and autonomous vehicle navigation. Almost every industry has been affected by machine learning, and new algorithms continue to be developed every day.

Although many machine learning algorithms focus specifically on extracting predictions from data, there are many in the field who've set their sights on a loftier goal — universal induction. *Inductive reasoning*, the process of drawing hypotheses from evidence, is a task crucial to human learning. The ability to generalize from instances is how infants learn that crying always brings a parent, and how children learn from being burned once by a hot pan that touching any hot objects is painful. Any time novel circumstances are encountered, humans rely on generalization and inference from examples to hypotheses to predict outcomes and make decisions. *Universal induction* refers to induction over all statements in a model, as opposed to a limited subset of information. It is this form of inductive reasoning that is considered so valuable for imitating human-scale learning processes, but it is a much more difficult task than limited inductive inference.

In order to formulate hypotheses from evidence, a system has to be able to represent both the hypothesis and the evidence internally. However, representing a proposition itself is not sufficient for induction; an inductive system must assert some relationship between statements and the world. One way to do this is to maintain a list of all possible statements, and assign each statement a truth value. For example, if statements are being represented as text strings, the list might contain the tuple `(``The sky is blue.'', TRUE)`. But, in reality, this is not an expressive enough representation. After looking at a weather report, one could reasonably believe that it might rain the following day and simultaneously believe that it might not. Representing the statement "It will rain tomorrow" as either true or false mischaracterizes one's dispositional attitude. For that reason, it is common to relate statements, not to a binary value of true or false, but to a probability that represents the degree of certainty one has that the statement is true. This is typically denoted $\mathbf{P}(A) = p$, where A is a propositional statement of the same sort as "The sky is blue" and "It will rain tomorrow," and p is a real number between 0 and 1 (including 0 and 1). This formulation of belief is known as *partial belief*.

There is a rich literature on philosophy of belief, which lends a precise vocabulary and several useful formalisms to the notion of partial belief. The degree of belief or degree of certainty of a proposition is commonly referred to as *credence*. For example, a proposition about which one is fairly certain would have a high credence, such as 0.9. A proposition about which one is maximally uncertain might have a credence of 0.5. And, a proposition one believes to be completely impossible might have a credence of 0.

Solomonoff induction, named after Ray Solomonoff, is a powerful theoretical

framework for universal induction that calculates real-valued credences for all hypotheses (represented as binary strings). Due to its adaptability and universality, Solomonoff induction has been heavily researched, and finite approximations of it have been used to play reinforcement learning games with huge success.

However, there are some situations in which real-valued credences do not seem to work. When there are an infinite number of possible outcomes, some models choose to assign infinitesimal probability value to certain outcomes. An *infinitesimal* is any quantity that is explicitly non-zero, but is less than or equal to than every real quantity. [1, 6]

In this paper I describe and advocate for Extended Solomonoff Induction (ESI), an enrichment of the Solomonoff induction framework that incorporates infinitesimal credences. **Section 2** presents a background to classical and infinitesimal probability, and lays out a formal groundwork for hyperreal credences. In **Section 3**, I introduce Solomonoff induction, and motivate the addition of infinitesimals to the framework. In **Section 4**, I propose and formalize ESI, and in **Section 5**, I defend the advantages of ESI by examining several common arguments against infinitesimal credences. I conclude in **Section 6** with a discussion of the triviality problem for finite models of ESI, which is a potential topic for further research.

2 Infinitesimal Credence

2.1 Classical Probability Theory

Axiomatic perspectives on probability come in two basic flavors: *epistemic* and *objective*. The epistemic approach views propositions as atomic object of probability. $\mathbf{P}(A)$ is construed as the credence or degree of belief that A is true, relative to some epistemic agent. Often, uppercase Greek letters, such as Ψ, are used instead of A to denote a proposition, and conjunction and disjunction are represented with logical connectives ($\wedge$ and $\vee$ respectively). The objective, or frequentist approach views events as the atomic object of probability, and $P(A)$ is taken to mean the likelihood that event A will happen. An event algebra is used to represent the probability space, and conjunction and disjunction are thereby represented with the set-theoretic intersection ($\cap$) and union ($\cup$) notations.

This paper will primarily present probability objectively, but the distinction is less important because Solomonoff induction acts on sequences of symbols, and thus abstracts away many of the formal details of a language.

2.1.1 Kolmogorov Probability Axioms

There are several axioms of probability which are commonly taken to characterize the rules of probability. These axioms are presented in many different ways, but most formulations of the axioms are equivalent. Presented below is one formulation of these axioms, attributed to the Andrey Kolmogorov [6]:

First Axiom The first axiom states that the probability of any statement must be greater than or equal to 0.

$$\mathbf{P}(A) \geq 0$$

Second Axiom The second axiom states that the probability of the sample space, Ω is 1.

$$\mathbf{P}(\Omega) = 1$$

Third Axiom The third axiom states that, for any disjoint propositions, A and B, the probability of one or the other or both occurring is the sum of their probabilities.

$$\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B)$$

Fourth Axiom Let

$$A = \bigcup_{n \in \mathbb{N}} A_n,$$

where $A_n \subseteq A_{n+1}$ are elements of the power set of Ω; then

$$\mathbf{P}(A) = \lim_{n \to \infty} \mathbf{P}(A_n)$$

The fourth axiom, together with the other axioms is equivalent to requiring σ-additivity. [1]

From these axioms, it is possible to derive many other important rules of probability, such as $\mathbf{P}(\emptyset) = 0$ (the probability of the empty set is zero). It is also simple to arrive at the classical rule of equiprobable outcomes, which states that if there are k possible outcomes, and each outcome is equally likely, then the probability of each outcome is $1/k$.

2.1.2 Bayes Theorem and Variations

Bayess' Theorem, named for Thomas Bayes, and 18th century English statistician, describes the relationships between the conditional probabilities of two events.

Conditional probability is the probability of one event given that another occurred, is occurring, or will occur. For example, one might ask what the probability of passing a test is given that one studied for the test for several hours. This might be represented as $\mathbf{P}(pass|study)$. Conditional probability is critical to inductive inference, as the credence of a hypothesis is only relevant given a specific set of evidence.

Joint probability (the probability of two events cooccurring) is sometimes defined in terms of conditional probability:

$$\mathbf{P}(A \cap B) = \mathbf{P}(A|B)\mathbf{P}(B) = \mathbf{P}(B|A)\mathbf{P}(A)$$

This makes sense, as the probability of both A and B occurring should be the probability of A, times the probability that B will happen if A does (and vice versa).

Using the aforementioned axioms of probability and conditional probability, one can derive Bayes' theorem. The standard presentation of the theorem states:

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

Bayes' theorem can be directly proven from the above axiom of joint probability, and is also often represented using joint probability:

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

Using a technique known as *marginalization*, a singular probability can be rewritten in terms a set of exhaustive, mutually exclusive propositions, $\bar{B} = \{B_k : \forall i, j\ \mathbf{P}(B_i \cap B_j) = \emptyset \ \wedge \ \sum_i \mathbf{P}(B_i) = 1\}$:

$$\mathbf{P}(A) = \mathbf{P}(A|B_0)\mathbf{P}(B_0) + \mathbf{P}(A|B_1)\mathbf{P}(B_1) + \ldots = \sum_{B_k \in \bar{B}} \mathbf{P}(A \cap B_k)$$

Informally, this means that the probability of an event, A, is the sum of the probabilities of A cooccurring with all of the events $\{B_0, B_1, \ldots\}$. This is also often used to rewrite Bayes' theorem:

$$\mathbf{P}(A|B_0) = \frac{\mathbf{P}(B_0|A)P(A)}{\sum_{B_k \in \bar{B}} \mathbf{P}(B_k|A)\mathbf{P}(A)}$$

2.2 Infinitesimal Probability Theory

In most situations, the standard rules and notations of probability are able to completely describe the behavior of partial belief. However, any model of true universal induction must somehow address the issues of infinite hypothesis spaces and infinitesimal credence, which do not have neatly-defined behavior in standard models of probability.

Under the standard formulation of probability, propositions cannot have infinitesimal credence because credences must be real numbers, and there are no infinitesimal reals. To understand why this is, consider any contender for 'smallest positive real number.' Such a number, if it existed, would be infinitesimal, because it would be less than all other real numbers. Of course, such a number cannot exist, because simply dividing its value in half would result in a smaller number. As such, infinitesimals cannot be real numbers.

To see how infinitesimal probabilities might arise in partial belief, consider the following two examples:

> **Dart Board** A dart is thrown at a circular dart board. All points on the board have an equal probability of being hit by the dart. The probability of the dart landing on any given point, $\mathbf{P}(x)$, is therefore $1/|D|$, where D is set of all possible points the dart might hit, and $|D|$ is the total number of possible points on the board.

However, there are an infinite number of possible points that the dart might potentially land on, so for any candidate credence, $\mathbf{P}(x) = k \in [0,1]$, there is some $N \in \mathbb{R}$ such that $N < |D|$ and $1/N \leq k$, so $1/|D| < 1/N \leq k$ and by transitivity, $1/|D| < k$. Therefore, $k \neq 1/|D|$. Because $\mathbf{P}(x)$ is smaller than any real number, it must be infinitesimal.

Coin Flips A fair coin is flipped an infinite number of times, and the resulting sequence is recorded. Each coin flip is an independent event, meaning the probability of the coin landing on heads the second time has nothing to do with the probability of landing on heads the first time, and so on. The probability of two independent events is the product of the individual probabilities ($\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B)$), and the probability of landing on heads each step is $1/2$. Therefore, the probability of getting heads N times in a row is $\mathbf{P}(\overrightarrow{H_N}) = \prod_{i=1}^{N} \mathbf{P}(H_i) = (1/2)^N$. If N is infinite, $\mathbf{P}(\overrightarrow{H_N})$ must, again, be smaller than any real number, and thus infinitesimal.

In both of these examples, an infinitesimal probability arises when the set of possible outcomes is infinite. The above cases are a bit contrived, but the concept of infinite outcome spaces and infinitesimal credences arises with alarming frequency in many real world models of universal induction.

2.3 Hyperreals

The hyperreals are a strict superset of the real numbers, first formulated by Abraham Robinson in the 1960s. [5] Many philosophers consider the hyperreals to be the best candidate for infinitesimal credences, notably suggested by Skyrms and Lewis and further developed by Bernstein and Wattenberg. [3] For reasons that will soon be made clear, the word 'the' in 'the hyperreals' is a bit of a misnomer, since there are actually many structures which could rightly be called 'the hyperreals.'

2.3.1 Constructing the Hyperreals

The hyperreals are constructed from the set of all sequences of real numbers $\langle r_n \rangle = \langle r_1, r_2, \ldots \rangle \in \mathbb{R}^{\mathbb{N}}$. Intuitively, the idea is to let sequences for which $\lim_{n\to\infty} r_n = 0$ represent infinitely small numbers (infinitesimals) and sequences for which $\lim_{n\to\infty} r_n = \infty$ represent infinitely large numbers. [7]

In order to compare these sequences, there must be a way to determine which of the sequences has a larger set of indices at which the real number at that index is greater for one than the other. However, for infinite sequences, one cannot directly compare the size of infinite sets, and it may be that there are an infinite number of indices at which the element at the index is greater than, and at which the element is equal to, and the element is less than the element at the same index in the other sequence.

For example, how might one compare the sequences $A = \langle 1, 0, 1, 0, 1, \ldots \rangle$ and $B = \langle 0, 1, 0, 1, 0, \ldots \rangle$. The answer is, one must construct an *ultrafilter*. An ultrafilter, $\mathfrak{F}$, is a subset of the power set of the set of natural numbers, to which which following apply:

- If $X \in \mathfrak{F}$ and $X \subseteq Y \subseteq \mathbb{N}$, then $Y \in \mathfrak{F}$
- If $X \in \mathfrak{F}$ and $Y \in \mathfrak{F}$, then $X \cap Y \in \mathfrak{F}$
- $\mathbb{N} \in \mathfrak{F}$
- $\emptyset \notin \mathfrak{F}$
- For any $A \subset \mathbb{N}$, $\mathfrak{F}$ contains exactly one of A and $\mathbb{N} \backslash A$

Additionally, an ultrafilter is *free* if it contains no finite subsets of $\mathbb{N}$.

For a free ultrafilter, $\mathfrak{F}$, let the equivalence relation $\equiv$ be defined on sequences of real numbers as:

$$\langle a_n \rangle \equiv \langle b_n \rangle \iff \{n \in \mathbb{N} \mid a_n = b_n\} \in \mathfrak{F}$$

The set of hyperreal numbers, ${}^*\mathbb{R}$, is thus defined as the modulo of the set of real sequences by $\equiv$:

$${}^*\mathbb{R} = \{[r] \mid r \in \mathbb{R}^{\mathbb{N}}\} = \mathbb{R}^{\mathbb{N}} / \equiv$$

From there, addition, multiplication, and an ordering operation are defined accordingly. [7]

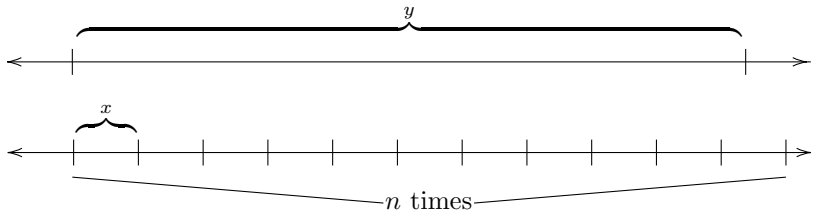
2.4 Measure Theory Objection

2.4.1 Archimedean Property

As support for infinitesimal-valued credences has grown within the academic community, so has opposition to their use. This is, in part, because any probability measure that allows for infinitesimals violates the *Archimedean property*. The Archimedean property is a property of certain algebraic structures. For an ordered field, $\mathcal{F}$, the principle states that for any positive elements, x and y in $\mathcal{F}$, there is some integer, n, such that $nx > y$. [1] Formally:

$$\forall x, y \in \mathcal{F} \ \exists n \in \mathbb{Z} \ \ nx > y$$

Plainly speaking, the Archimedean property requires that any element can be measured in terms of a smaller element, by stacking up the smaller element a finite number of times until it is the same size as or greater than the larger element.



Another way of thinking about the property is that it simply rejects any fields that include infinite or infinitesimal spaces, because any infinitesimal multiplied by an integer is still infinitesimal.

The Archimedean property is required to give meaning to the concept of 'measurement,' because non-Archimedean fields do not have *comparability* — not every pair of elements can be compared in a meaningful way. Infinitesimal elements can be compared with other infinitesimal elements, but a comparison of relative size between an infinitesimal element and a standard-valued element will never return a finite measurement.

2.4.2 Infinite Sums of Measure Zero

Many opponents of infinitesimal probability argue against it on the grounds that it simply isn't necessary. They offer *measure theory* as an alternative framework for explaining the probability of minuscule propositions without assigning infinitesimal probabilities.

Measure theory provides an analytical method for assigning a notion of 'size' to subsets of a set. Some subsets of infinite sets may be assigned *measure zero*, which means that, relative to the size of the set, the subset is incomparably small. The union of an infinite number of subsets of measure zero does not necessarily result in a measure zero set. A common example of this is in integration. The area under a single point in a function has a value of zero, and the subset consisting of a single point has measure zero relative to the set of all points that comprise the function. And yet, the sum of the area under all points of a function has a finite, real value (namely the integral).

Proponents of measure theory reject the necessity of infinitesimals by pointing out a supposed flaw in the argument for infinitesimals. In order to argue that infinitesimals are necessary, one must argue that the sum of probabilities of minuscule propositions must not be minuscule, but will be unless minuscule propositions have infinitesimal credences. Measure theorists suggest that minuscule propositions may have zero probability but can still can sum to finite values.

The measure theory argument seems intuitive because there are many examples of sets of measure zero that are defined for functions which obtain value only at a limit. For example, area under a curve isn't just the sum of the areas under every point, it's the limit of a set of finite areas as the number of areas approaches infinity and the size of each area approaches zero. Properties such as area are *limit-valued*.

Unlike area, however, the probability of a set of propositions is exactly the sum of the probabilities of each proposition. Probability is not limit-valued, and therefore if minuscule propositions all have credence 0, then even an infinite sum of the probabilities of minuscule propositions will still always equal 0.

3 Solomonoff Induction

3.1 History and Background

In 1964, Ray Solomonoff published "A Formal Theory of Inductive Inference," in which he proposed a new formal framework for universal inductive inference. The framework merged the Bayesian probability framework with *algorithmic probability*, which refers to the probability of binary sequences being produced by a specific type of machine known as a *Universal Turing Machine*. [8, 9]

A *Turing machine* is a conceptual machine which has a binary tape, read-write head, and a finite set of rules that determines how the machine manipulates the symbols on the tape. A Universal Turing Machine generalizes the concept of a Turing machine. Given a specific input program, a Universal Turing Machine can emulate any other Turing Machine. As such, Universal Turing Machines are the conceptual prototypes for all computers, and given an adequate amount of time, can perform any computation that a modern computer can.

Given a black box 'environment,' a so-called *Solomonoff machine* feeds a binary input into the environment, and observes the binary output. It uses these observations of input and output to construct a *reference machine*, to probabilistically model the observed relationships between input and output, and thereby report a probability of a given output being produced, and predict the next bit of output given an input string. The input strings correspond to the notion of hypotheses, which, when given to the real world environment, produce evidence as output.

Solomonoff induction proceeds by defining a *universal prior* — a probability distribution over the set of all binary strings. This probability is assigned in such a way to reflect *Occam's Razor*, an epistemic principle that holds that, *a priori* (without prior knowledge), 'simpler' hypotheses are more probable. Solomonoff formalized this principle using *Kolmogorov complexity*. Kolmogorov complexity, $\kappa(x)$, is a measure of the algorithmic complexity of a string, x, and is defined as the length of the shortest program which outputs the given string. For example, Kolmogorov complexity is able to encode the difference in complexity between the following strings:

1. `111111111111111111111111`
2. `100101111010110101011100110`

One might try to encode each of these strings with an algorithm in a C-like language:

1. `for(i=0;i<25;i++) print('1');`

2. `print('10010111101011011011100110');`

For the first string, it is possible to construct an algorithm that encodes the string without having to write the entire string, in this case using 29 characters. For the second string, no such algorithm is easily available, meaning the the complexity must be at least the length of the string itself, in this case 35 characters. Of course, the 'length' of a program depends on the language the string is expressed in. But there are a number of proven bounds to Kolmogorov complexity, which make it an invaluable tool in measuring the relative simplicity of information.

Solomonoff induction assigns priors to each binary string, inversely weighted by the Kolmogorov complexity. The probability of a hypothesis, H_0, is given as

$$\mathbf{P}(H_0) = 2^{-\kappa(H_0)}$$

where $\kappa(H_0)$ is the Kolmogorov complexity of H_0. If the hypothesis is considered as a binary input string encoding an algorithm, the probability of H_0 is based simply on the length:

$$\mathbf{P}(H_0) = 2^{-length(H_0)}$$

This corresponds with the probability that the string would occur randomly. The shorter the input string, the more likely it is to occur randomly, and vice versa.

The probability of an output string, E_0 (which can be thought of as evidence or data produced by the hypothesis), is the sum of the probabilities of all of the input strings which, when fed to the Universal Turing Machine, U, produce a string that starts with E_0 (denoted E_0^*):

$$\mathbf{P}(E_0) = \sum_{\{H|U(H)=E_0^*\}} 2^{-length(H)}$$

This can be thought of as equivalent to the marginalization technique described in **Section 2.1.2**.

By requiring that the input codes be *prefix-free* (meaning no code is the prefix for any other), shorter codes are padded out with random binary strings. If the input string `010` matches a given output string, then `0100` and `0101` both will as well, and so on. Thus, a matching hypothesis that is one bit shorter than another contributes twice as much to the sum.

Typically, Solomonoff induction is considered to be deterministic — given a specific input and output string, the machine either will or will not produce the output given the input. In other words $\mathbf{P}(E|H)$ is always a 1 or a 0. In this scenario, the conditional probability can be calculated from Bayes formula:

$$\mathbf{P}(H_0|E_0) = \frac{\mathbf{P}(E_0|H_0)\mathbf{P}(H_0)}{\mathbf{P}(E_0)} = \frac{2^{-length(H)}}{\sum_{\{H|U(H)=E_0^*\}} 2^{-length(H)}}$$

When Solomonoff induction is non-deterministic, there are well-researched bounds on algorithmic complexity.[9] For the most part, this paper will assume a non-deterministic model, as that most closely matches the general form of universal induction seen as desirable for modeling learning.

3.2 Motivating ESI: Infinite Hypothesis Spaces

Most limited models of induction follow a *supervised learning* paradigm, meaning the system is given a clear representation of the data and the expected output, which are generated by a human programmer. The set of possible outputs is usually already specified, and requires no creativity or generation of new possible outputs on the part of the inductive system. Following a supervised learning paradigm, an inductive system would typically be provided with a data set and a finite set of possible hypotheses, and would output the probabilistic confidence in each hypothesis (or simply output the highest-confidence hypothesis).

A model of true universal induction might instead consider as the hypothesis space the entire set of representable propositions, given a language for representing propositions. Given a finite amount of time and a finite amount of state memory, this hypothesis space will, of course, be finite. However, it may also be possible for a system to comprehend that there are certain propositions that are possible (and should be considered when computing credences) that are not currently representable or algorithmically reachable states given the machine's current computational limitations.

To give a concrete example of this, let us presume a model based on a Bayesian framework of conditional probability. The system is asked to determine the probability that a coin is fair (denoted $\mathbf{P}(F)$), given that it has been tossed four times resulting in the sequence $\overrightarrow{H_4} = \langle HHHH \rangle$:

$$\mathbf{P}(F|\overrightarrow{H_4}) = \frac{\mathbf{P}(\overrightarrow{H_4}|F)\mathbf{P}(F)}{\mathbf{P}(\overrightarrow{H_4})}$$

It is easy to determine the probability of the sequence given that the coin is fair:

$$\mathbf{P}(\overrightarrow{H_4}|F) = \mathbf{P}(H_1|F) \times \mathbf{P}(H_2|F) \times \mathbf{P}(H_3|F) \times \mathbf{P}(H_4|F) = (1/2)^4 = 1/16$$

However, it is much more difficult to determine the *prior probabilities* $\mathbf{P}(F)$ and $\mathbf{P}(\overrightarrow{H_4})$. For the sake of simplicity, let us assume that the system already has a prior probability for $\mathbf{P}(F)$, even if its just 0.5, or some other value based on previous data and assumptions. The only sensible way to compute $\mathbf{P}(\overrightarrow{H_4})$ is to use the marginalization technique discussed in **Section 2.1.2**. But what set of propositions should be marginalized over? The set must include the hypothesis F. Another hypothesis that might be desirable would be $R_{1.0}$, which corresponds to the statement "The coin is biased to land on heads 100% of the time." Similarly, one might want to include a weakened form, $R_{0.75}$, or "The coin is biased to land on heads 75% of the time." Following this pattern, it is reasonable to assume that the set should contain all propositions of the form R_k for all $k \in [0, 1]$.

Interestingly, with just this set of propositions $\bar{R} = \{R_k : k \in [0, 1]\}$, it is possible to solve for $P(R_k)$ using a class of machine learning algorithms known as *Expectation Maximization (EM)* algorithms. These algorithms rely on the

assumption that the probability values are generated according to a specific family of distributions (such as the binomial distribution, which would describe the coin flips). However, there are other propositions that must be considered as well, such as "The fair coin is flipped by a magician who forces the coin to land on heads each time," or "The first coin flip in a sequence is fair and the outcome of each subsequent toss is identical to the previous toss." A system must consider all of these hypotheses, and many more, in order to faithfully model universal induction. In fact, any proposition that the model's language can express could potentially have a causal relationship with the sequence of coin tosses, and so the prior must be marginalized over every expressible proposition.

Even the set $\bar{R}$, which might be the hypothesis space in a very minimal language, is infinite. In fact, $\bar{R}$ is *uncountably infinite*. *Countably infinite* sets are those whose elements correspond 1-to-1 with the set of natural numbers $\mathbb{N} = \{0, 1, 2, 3, \ldots\}$, and thus can be 'counted' one at a time. Uncountably infinite sets have a greater number of elements, and thus their elements cannot be counted. Listing the numbers in order, by any ordering scheme, will never contain every number in the set. The set of real numbers, $\mathbb{R}$, is the quintessential uncountable set, and because $\bar{R}$ corresponds exactly with $\mathbb{R}$, it too is uncountably infinite.

It is worth noting that, in computational models of induction, the hypothesis space can be considered countable, because every unique machine state has a finite binary representation, and these binary representations can be ordered by length and concatenated.

4 Extended Solomonoff Induction (ESI)

In this section, I propose an extension to Solmonoff Induction, ESI, which specifies infinitesimal-valued credences as the universal prior for infinitely complex hypotheses.

The probability function, $\mathbf{P}$, maps the set of all binary strings, Ω, to an extension of the real field. This field is formed, not by all sequences of real numbers, but by the union of real numbers with the unlimited number,

$$\Upsilon = \langle 2^n \rangle = \langle 2^1, 2^2, \ldots, 2^k, \ldots \rangle,$$

together with the addition and multiplication operations (defined element-wise, as with the hyperreals). The multiplicative inverse of Υ is ε:

$$\varepsilon = \langle 2^{-n} \rangle = \langle (1/2)^1, (1/2)^2, \ldots, (1/2)^k, \ldots \rangle.$$

As defined, Υ encodes the "number" of binary strings of countably infinite length, and ε encodes the Solomonoff universal prior probability of an infinitely long binary string.

Every real number, $r \in \mathbb{R}$, is also represented as an infinite sequence $r \equiv \langle r, r, r, \ldots \rangle$. The 0-element of the field is therefore $0 \equiv \langle 0, 0, 0, \ldots \rangle$ and the 1-element is $1 \equiv \langle 1, 1, 1, \ldots \rangle$. Let the set of all elements in the closure the set $\mathbb{R} \cup \Upsilon$ under countable addition and finite multiplication define the new set Δ.

Δ, defined as such, along with addition and multiplication, satisfies all of the field axioms.

Each element of the extended field can be represented by an ordered pair, (s, n), where s is the standard (real) part and n is the non-standard part. A lexicographic ordering can now be imposed to form an ordered field. The ordering operation, $<$, first considers the standard part, and if the standard parts are equivalent, considers the non-standard part. Equivalently, any two elements of the field can be compared element-wise. Unlike the hyperreals, however, only a finite number of indices in a representative sequence can be non-monotonic. Therefore, no ultrafilter is needed to compare sequences, because only one set of indices (those that are larger, those that are equal and those that are smaller) can be infinite.

The set Δ, along with the ordering operation, $<$, and the addition and multiplication operations define an ordered field, denoted ${}^{\delta}\mathbb{R}$. This extended ordered field will be used as the codomain of the ESI probability measure: $\mathbf{P} : \Omega \to [0, 1]_{{}^{\delta}\mathbb{R}}$.

The probability axioms can be constructed for ESI in the exact same manner as with Classical Probability Theory, except that the codomain of $\mathbf{P}$ is $[0, 1]_{{}^{\delta}\mathbb{R}}$ instead of $[0, 1]_{\mathbb{R}}$.

4.1 Example: Infinite Coin Toss Sequence

Using ESI, it is possible to precisely calculate the probabilities of infinite sequences. Let $\vec{X}$ be an random, infinite sequence of coin tosses of a fair coin (i.e. $\langle HTHHTHTHTTTTHT\ldots\rangle$). Let $S_{\vec{X}}$ be the hypothesis that simply produces the $\vec{X}$ by reprinting it character-by-character. Let $C_{\vec{X}}$ be a hypothesis corresponding to a compression of $\vec{X}$ with a 50% compression ratio. And let $P_{\vec{X}}$ be a finite hypothesis that outputs $\vec{X}$.

The Kolmogorov complexity of each of hypothesis is related to the length of the program used to encode it. $\kappa(S_{\vec{X}}) = \Upsilon + c_0$, where c_0 is a positive integer constant, corresponding to the length of the `print()` method of the algorithm. $\kappa(C_{\vec{X}}) = \frac{\Upsilon}{2} + c_1$, where c_1 is a positive integer constant corresponding to the length of the decoding algorithm. And $\kappa(P_{\vec{X}}) = k \in \mathbb{N}$.

The prior probabilities for each hypothesis can be directly calculated:

$$\mathbf{P}(S_{\vec{X}}) = 2^{-\kappa(S_{\vec{X}})} = \xi \qquad \mathbf{P}(C_{\vec{X}}) = 2^{-\kappa(C_{\vec{X}})} = \sqrt{\xi} \qquad \mathbf{P}(P_{\vec{X}}) = 2^{-\kappa(P_{\vec{X}})} = q$$

where ξ is an infinitesimal, and $q \in \mathbb{Q}$ is a finite rational number.

Now, a few different scenarios can be considered. In the first scenario, the only hypothesis which outputs $\vec{X}$ is $S_{\vec{X}}$. In this case, $\mathbf{P}(X) = \xi$, and

$$\mathbf{P}(S_{\vec{X}}|\vec{X}) = \frac{\xi}{\xi} = 1$$

The same can be done for $C_{\vec{X}}$ and $P_{\vec{X}}$. This is the simplest scenario, and it makes sense — if there is only one matching hypothesis for the observed evidence

in the entire hypothesis space, conditionalizing the hypothesis on the evidence should result in complete confidence in the hypothesis.

The next scenario to consider is one in which $S_{\vec{X}}$ and $P_{\vec{X}}$ are the only matching hypotheses for $\vec{X}$:

$$\mathbf{P}(S_{\vec{X}}|\vec{X}) = \frac{\xi}{\xi + q} = \omega_0$$
$$\mathbf{P}(P_{\vec{X}}|\vec{X}) = \frac{q}{\xi + q} = 1 - \omega_0$$

Information about ω_0 can be obtained by taking the limit of $\mathbf{P}(S_{\vec{X}})$ as $\xi \to 0$, which is 0. This means that the sequence in $\mathbb{R}^{\mathbb{N}}$ corresponding to ω_0 is Cauchy, and thus ω_0 is an infinitesimal. If $\mathbf{P}(S_{\vec{X}})$, that means that $\mathbf{P}(P_{\vec{X}})$ is infinitely close to, but infinitesimally less than 1.

Similar probability distributions result if the only matching hypothesis are $S_{\vec{X}}$ and $C_{\vec{X}}$:

$$\mathbf{P}(S_{\vec{X}}|\vec{X}) = \frac{\xi}{\xi + \sqrt{\xi}} = \omega_1$$
$$\mathbf{P}(C_{\vec{X}}|\vec{X}) = \frac{\sqrt{\xi}}{\xi + \sqrt{\xi}} = 1 - \omega_1$$

and if all three identified hypotheses match:

$$\mathbf{P}(S_{\vec{X}}|\vec{X}) = \frac{\xi}{\xi + \sqrt{\xi} + q} = \omega_2$$
$$\mathbf{P}(C_{\vec{X}}|\vec{X}) = \frac{\sqrt{\xi}}{\xi + \sqrt{\xi} + q} = \frac{\omega_2}{\sqrt{\xi}}$$
$$\mathbf{P}(P_{\vec{X}}|\vec{X}) = \frac{q}{\xi + \sqrt{\xi} + q} = 1 - \omega_2 - \frac{\omega_2}{\sqrt{\xi}}.$$

What these probability distributions suggest is that, in line with Occam's Razor, a finitely complex hypothesis will always be more supported by evidence than an infinitely complex one, and that for two infinitely complex hypotheses, the relatively less complex of the two hypotheses will always be more supported by evidence.

The significance of these results is two-fold. First, many of the conditional probabilities above simply cannot be calculated in any standard Bayesian probability model, let alone in Solomonoff induction. In particular, in a standard model, $\mathbf{P}(S_{\vec{X}})$ and $\mathbf{P}(C_{\vec{X}})$ would have to take on a credence of 0. This already eliminates valuable information about the ordering properties of the hypotheses. This could lead an inductive system to come to the irrational conclusion that a hypothesis for which the corresponding algorithm begins by iterating through 100-million-bit-long list and doing nothing each time and then directly printing the output is as probable as a hypothesis for which the corresponding algorithm immediately prints the output. Especially because infinite sequences play a role in active, ongoing data generation, it is empirically possible for a standard

Solomonoff model to consider both such infinite hypotheses to have the same exact credence, and implement the 100-million-bit-iterator. Additionally, if the only hypotheses found are infinitely complex, conditional probability cannot be calculated at all because it would require division by zero.

Second, it is significant for finite approximations of ESI that a finitely complex hypothesis will dominate an infinitely complex hypothesis, because it means if any finitely complex hypotheses are found, the model no longer has to check for infinitely complex hypotheses. Even if there is an infinite set of matching infinitely complex hypotheses, no single infinitely complex hypothesis in the set can have a higher probability than the finite hypothesis. In that scenario, the model can compute the hypothesis with the highest credence, even if its impossible to compute the precise probability for each hypothesis.

5 Advantages of ESI

5.1 Unfiltered Codomain

One of the biggest advantages of ESI is that, by nature of its construction, the codomain of the probability distribution does not have to be constructed via an ultrafilter, unlike the hyperreals. This means several things - first, infinitesimal probability can be computed directly from information about the proposition. In order to construct, the model must specify an ultrafilter, which takes up an infinite amount of space. Moreover, different constructions of the hyperreals use different ultrafilters and it is not necessarily the case that all models constructed by different ultrafilters are isomorphic (this turns out to be equivalent to the axiom of choice). On top of that, in order to check for equality of two hyperreals, a model must confirm that a set indices is in the ultrafilter, which could take an infinite amount of time.

By contrast, ESI provides a simple, standardizable, and intuitive procedure for generating the prior probability for every proposition in the domain. Additionally, checking for equality of ESI infinitesimals only requires confirming that the set of indices at which elements do not match is finite.

5.2 Williamson's Coin Toss Argument

Regularity — the principle that if a proposition has credence 0, it must be epistemically impossible — is often used as a justification for infinitesimal credences. Although regularity holds in ESI, it is not necessary for justifying the use of infinitesimals. Because of that, ESI is advantageous in its ability to defend regularity without circular reasoning.

In his 2007 paper "How Probable Is an Infinite Sequence of Heads?," Williamson argues against regularity, by attempting to exhibit a scenario in which, even in a non-standard model, regularity fails.

His argument is as follows: consider a coin which is flipped an infinite number of times, at one second intervals. Let $\mathbf{P}(\overrightarrow{H_1})$ be the event that every toss comes

up heads. $\overrightarrow{H}$ is unlikely, but still epistemically possible, so by regularity it must have a non-zero credence. Additionally, let $\mathbf{P}(H_1)$ be the probability that the coin land on heads on the first toss, and $\mathbf{P}(\overrightarrow{H_2})$ be the probability that the coin land on heads every time after the first toss.

1. $\mathbf{P}(\overrightarrow{H_1}) = \mathbf{P}(H_1 \cap \overrightarrow{H_2})$
2. H_1 and $\overrightarrow{H_2}$ are independent, so $\mathbf{P}(H_1 \cap \overrightarrow{H_2}) = \mathbf{P}(H_1)\mathbf{P}(\overrightarrow{H_2})$
3. $\mathbf{P}(H_1) = \frac{1}{2}$
4. By (1)-(3), $\mathbf{P}(\overrightarrow{H_1}) = \frac{1}{2}\mathbf{P}(\overrightarrow{H_2})$
5. $\overrightarrow{H_1}$ and $\overrightarrow{H_2}$ are isomorphic events, so $\mathbf{P}(\overrightarrow{H_1}) = \mathbf{P}(\overrightarrow{H_2})$
6. By substitution on (4) and (5), $\mathbf{P}(\overrightarrow{H_1}) = \frac{1}{2}\mathbf{P}(\overrightarrow{H_1})$
7. $\therefore \mathbf{P}(\overrightarrow{H_1}) = 0$

The flawed premise in Williamson's argument is (5). In a standard model, it is true that $\overrightarrow{H_1}$ and $\overrightarrow{H_2}$ are isomorphic, because there exists a bijection between the sets of tosses. The function $I : \overrightarrow{H_1} \to \overrightarrow{H_2}$ can be defined as $I(H_k) = H_{k+1}$.

However, in ESI, $\mathbf{P}(\overrightarrow{H_1})$ and $\mathbf{P}(\overrightarrow{H_2})$ can be computed directly from the propositions $\overrightarrow{H_1}$ and $\overrightarrow{H_2}$.

$$\mathbf{P}(\overrightarrow{H_1}) = 2^{-length(\overrightarrow{H_1})} = \varepsilon = \langle 2^{-1}, 2^{-2}, 2^{-3}, \ldots \rangle$$

In order to computer $\overrightarrow{H_2}$, observe that, for any sequence of fair coin flips, prepending a 100% biased coin flip does not change the probability. For example, let $\mathbf{P}(H^*)$ represent the probability of landing on heads with a coin that has heads on both sides. $\mathbf{P}(H_1^*)$ of course equals 1. Naturally, $\mathbf{P}(\langle HHH \rangle) = (1/2)(1/2)(1/2) = (1)(1/2)(1/2)(1/2) = \mathbf{P}(\langle H^*HHH \rangle)$. From this, we get $\mathbf{P}(\overrightarrow{H_2}) = \mathbf{P}(\langle H^* \rangle \overrightarrow{H_2})$. This sequence, $\langle H^* \rangle \overrightarrow{H_2}$, has the same 'length' relative to $\overrightarrow{H_1}$. Because $\mathbf{P}(H_1^*) = 2 * \mathbf{P}(H_1)$, $\mathbf{P}(\langle H^* \rangle \overrightarrow{H_2}) = 2 * \mathbf{P}(\overrightarrow{H_1})$, which means

$$\mathbf{P}(\overrightarrow{H_2}) = 2 * 2^{-length(\overrightarrow{H_1})} = 2\varepsilon = \langle 2^0, 2^{-1}, 2^{-2}, \ldots \rangle$$

.

Because $\varepsilon, 2\varepsilon \in {}^{\delta}\mathbb{R}$, they can be compared element-wise. Examining the sequences, it's clear $\mathbf{P}(\overrightarrow{H_2}) > \mathbf{P}(\overrightarrow{H_1})$. Because it is not true that $0 > 0$, one can conclude that $\mathbf{P}(\overrightarrow{H_2}) \neq 0 \neq \mathbf{P}(\overrightarrow{H_1})$.

5.3 Minuscule Conditionalization (Easwaran)

In "Regularity and Hyperreal Credences," Easwaran also argues against the use of infinitesimals by trying to condemn Regularity. He contends that in any model with minuscule credences and Bayesian conditionalization it can be proven that "only minuscule propositions can provide enough evidence for one to believe other minuscule propositions." He gives the following argument:

1. $\mathbf{P}(A \cap B)$ is minuscule when $\mathbf{P}(A)$ is minuscule.
2. When an agent learns B, she replaces her credence $\mathbf{P}(A)$ with $\mathbf{P}(A|B)$ for every proposition A.
3. $\mathbf{P}(A|B)$ is minuscule when $\mathbf{P}(A \cap B)$ is minuscule and $\mathbf{P}(B)$ is not minuscule.
4. Therefore, if A is minuscule, then the agent will never have high credence in A unless she learns some B that is also minuscule.

Easwaran concludes that the only way around this problematic result is to alter the rules for Bayesian conditionalization in such a way that invalidates a significant argument in favor of Regularity, which he describes as the argument from 'stubbornness.' [3]

Easwaran's argument may be valid, but the supposition that the conclusion — "Only minuscule propositions can provide enough evidence for one to believe other minuscule propositions" — is problematic seems to be based on an erroneous intuition. On the surface, it seems that there are lots of appreciable propositions which should be able to provide enough evidence for one to believe in an infinitesimal proposition.

Because of ESI's capability to describe infinitesimals and its relationship to Occam's Razor, it is clear why, within ESI, only minuscule propositions can conditionalize other minuscule propositions such that they become appreciable. The first example scenario given in **Section 4.1** demonstrates this perfectly.

Let us return, for another example, to the dartboard thought experiment. Let A be the proposition that a dart thrown at a circular dartboard will hit some exact point, (r, θ), on the board. Because there are, in theory, an uncountably infinite number of possible coordinates the dart might hit, A must be a minuscule proposition. Suppose that B is a newly learned proposition, which states that there is a magnet which draws any darts thrown at the board to (r, θ) 100% of the time. Ideally, we would want $\mathbf{P}(A|B)$ to reflect this updated likelihood, and assign a high future credence to A.

In this case, it is not immediately obvious that B must also be a minuscule proposition. But, consider all of the possible alternatives to B. For example, B could have been the proposition that the magnet pulls the dart to the radial coordinate $(r, \theta) = (0.000001, 0.4)$. There are just as many propositions similar to B as there are potential points for the dart to hit in A. Therefore, just as A must be minuscule, so must B, unless there is some additional evidence for B which increases its credence.

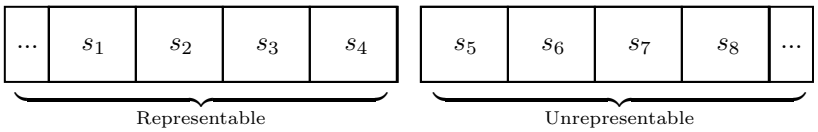
A more realistic alternative to B, call it B^*, is that the magnet will cause the dart to land somewhere inside a small radius around the exact center of the board. $\mathbf{P}(B^*)$ should certainly be a plausibly high credence. But $\mathbf{P}(A|B^*)$, in this case, should not be a high credence, because within that small radius, there are still an uncountable number of possible positions. It does seem desirable to claim that $\mathbf{P}(A|B^*) > \mathbf{P}(A)$, but if A does not have an infinitesimal credence, then this cannot be true unless there is some external structure imposed on probabilities of minuscule credences or non-Bayesian update rule.

6 Further Research: Triviality in Finite Models

Infinite hypothesis spaces pose a very specific challenge to computational models of induction — finite models can never represent all of the possible values. In any model with an infinite hypothesis space, there will be some propositions that are unrepresentable. This means that the model cannot store the proposition in memory, and cannot calculate its value in a finite amount of time. It might also mean that there are propositions that the model's language is simply not expressive enough to represent in its current state. As a result, computational models of induction typically just ignore any propositions that are not representable in the model's current state, and treat the representable propositions as the full hypothesis space. This naive approach leads to a serious concern about the triviality of which propositions are assigned credence.

Just because a proposition is not currently representable, does not mean it is not a possible value, or that should not be factored into a system's calculations when determining credences. Returning to the coin flipping example from the previous section, imagine that a young child is playing a game where she watches a coin being flipped multiple times, and is asked to bet on the outcome of the each coin flip. Suppose the child has never even heard of weighted coins, and so the only hypothesis in her hypothesis space is F, "The coin is fair." If the coin is biased to always land on heads, but the child assumes that the only possible hypotheses are the representable hypotheses, she will continue to make random bets on the outcome of the next flip, because the sum of credences of all hypotheses in the sample space is 1, and so the child can only have 100% confidence in the hypothesis. Of course, if the child is taught about biased coins, her hypothesis space will expand. Similarly, if a computer is given extra memory, in the form of an additional memory card or drive, or the available representation language is expanded, a model's hypothesis space can expand.

If the model does not limit its hypothesis space to the representable states, the question then becomes, what credences should the unrepresentable states have? Take, as an example, this simplified depiction of a model with 4 representable states, labeled s_1, s_2, s_3, and s_4, stored in the last for memory blocks of the machine, and an infinite sequence of states, $s_5, s_6, \ldots$ that are not representable:



Assume that each state, s_i, stores a tuple containing a proposition, A_i, and a probability value, $P(A_i)$, and that all propositions are mutually exclusive ($\bigcap_{i=1}^{\infty} A_i = \emptyset$).

Although the probabilities of the unrepresentable states are not explicitly assigned, information about their probabilities can be deduced from the sum of the probability values of the representable states:

$$\sum_{i=1}^{4} P(A_i) = 1$$

The unrepresentable propositions still correspond to actual possibilities in the universal hypothesis space, Ω, the credences of which sum to 1 by the **First Axiom** of probability:

$$\sum_{i=1}^{\infty} P(A_i) = \sum_{A_i \in \Omega} P(A_i) = 1$$

The probabilities of the unrepresentable propositions sum to 0:

$$\sum_{i=5}^{\infty} P(A_i) = \sum_{i=1}^{\infty} P(A_i) - \sum_{i=1}^{4} P(A_i) = 0$$

Therefore, the probability of each unrepresentable proposition is 0:

$$\forall i \in \mathbb{N}, i \geq 5 \quad P(A_i) = 0$$

Although nothing about the naive approach violates the laws of probability, the triviality of the probability assignments it produces should raise concerns about the integrity of the approach. Why is it that A_4 has a non-zero credence, and A_5 does not? A_5 is a possible proposition and s_5 is a plausible state, and the only reason $P(A_5) = 0$ is that the machine does not have enough memory space to represent it. It's possible that, should the memory be expanded to include s_5 and the probability distribution reevaluated from scratch, A_5 might even have the highest credence of all of the propositions. I contend that the triviality of this distinction undermines the credibility of all other probability assignments made by a model following the naive approach.

7 Acknowledgements

This work would not have been possible without the mentorship of my thesis advisor, Dr. Zoltan Domotor, Director of Undergraduate Studies in Philosophy, who directed my research and encouraged me to write an interdisciplinary thesis. I would also like to thank Dr. Henry Towsner, Assistant Professor of Mathematics, for his guidance on non-standard analysis, and inspirational brainstorm sessions.

Additionally, I would like to thank Owen Gray, who introduced me to Solomonoff induction, and Dr. Daniel Singer, Assistant Professor of Philosophy, who recommended to me Easwaran's "Regularity and Hyperreal Credences" — the paper that started it all.

Perhaps most importantly, I am eternally grateful for the love and support of my girlfriend, Annie, who endured my endless rambling and complaining, and of my parents, who taught me to pursue my passions and follow them through to the end.

8 References

[1] Benci, V., Horsten, L., & Wenmackers, S. (2016). Infinitesimal probabilities. *The British Journal for the Philosophy of Science*, 1-44. https://doi.org/10.1093/bjps/axw013

[2] De Falco, I. A., Cioppa, A. D., Maisto, D., & Tarantino, E. (2006). A genetic programming approach to Solomonoff?s probabilistic induction. In *Lecture Notes in Computer Science* (pp. 24-35). Retrieved from ResearchGate database.

[3] Easwaran, K. (2014). Regularity and hyperreal credences. *Philosophical Review*, *123*(1), 1-41. https://doi.org/10.1215/00318108-2366479

[4] Easwaran, K., & Towsner, H. (2018, January 30). *Realism in mathematics: The case of the hyperreals*. Unpublished working paper.

[5] Goldblatt, R. (1998). *Graduate Texts in Mathematics: Vol. 188. Lectures on the hyperreals: An introduction to nonstandard analysis*. New York, NY: Springer-Verlag.

[6] Katz, M. G., & Sherry, D. (2013). Leibniz's infinitesimals: Their fictionality, their modern implementations, and their foes from Berkeley to Russell and beyond. *Erkenntnis*, 78(3), 571-625. Retrieved from JSTOR database. (Accession No. 42001448)

[7] Ødegaard, A. T. M. (n.d.). *Hyperreal calculus*. University of Oslo.

[8] Rathmanner, S., & Hutter, M. (2011). A philosophical treatise of universal induction. *Entropy*, 13(6), 1076-1136. https://doi.org/10.3390/e13061076

[9] Vallinder, A. (2012). *Solomonoff induction: A solution to the problem of the priors?* (Master's thesis, Lund University).

[10] Weintraub, R. (2008). How probable is an infinite sequence of heads? A reply to Williamson. *Analysis*, *68*(3), 247-250. Retrieved from JSTOR database.

[11] Wenmackers, S. (2011). *Philosophy of probability: Foundations, epistemology, and computation* (Doctoral dissertation). University of Groningen, Netherlands.

[12] Williamson, T. (2007). How probable is an infinite sequence of heads? *Analysis*, *67*(3), 173-180. Retrieved from JSTOR database.