PUTNAM'S DIAGONAL ARGUMENT AND THE IMPOSSIBILITY OF A UNIVERSAL LEARNING MACHINE

TOM F. STERKENBURG

Abstract. Putnam (1963) construed the aim of Carnap's program of inductive logic as the specification of an "optimum" or "universal" learning machine, and presented a diagonal proof against the very possibility of such a thing. Yet the ideas of Solomonoff (1964) and Levin (1970) lead to a mathematical foundation of precisely those aspects of Carnap's program that Putnam took issue with, and in particular, resurrect the notion of a universal learning machine.

This paper takes up the question whether the Solomonoff-Levin proposal is successful in this respect. I expose the general strategy to evade Putnam's argument, leading to a broader discussion of the outer limits of mechanized Bayesian induction. I argue that this strategy ultimately still succumbs to diagonalization, reinforcing Putnam's impossibility claim.

1. Introduction

Putnam (1963a) famously challenged the feasibility of Carnap's program of inductive logic on the grounds that a quantitative definition of "degree of confirmation" can never be adequate as a rational reconstruction of inductive reasoning. Specifically, he formulated two conditions of adequacy, and proceeded to give a *diagonal proof* to the effect that no Carnapian measure function can satisfy both. In (1963b), Putnam assumed the view that "the task of inductive logic is the construction of a 'universal learning machine'" (303), and accordingly presented his proof as showing the impossibility of this notion. What was shown, in these terms, is that there can be no *learning machine* that is also *universal*: no measure function that is effectively computable, that is also able to eventually detect any pattern that is effectively computable.

Independently of the work of Putnam, the suggestions of Solomonoff (1964) towards an "optimum induction system" gave rise to a definition that is very much in this spirit. The elements that Solomonoff took from Carnap's program, and those that he added to it — most importantly, the central role of effective computability — are the very elements that Putnam presumed in his challenge to it. Solomonoff's ideas found a secure mathematical footing in the work by Levin (1970), resulting in what qualifies, perhaps, as the definition of a universal learning machine. Namely, the *Solomonoff-Levin measure* does manage to unite versions of Putnam's two adequacy conditions — though, crucially, involving a weakened notion of effective computability.

In this paper I investigate whether the Solomonoff-Levin proposal indeed gives a definition of an "optimum," "cleverest possible," or *universal* learning machine. More broadly, this is an investigation into the possibility of a (Bayesian) definition

Date: January 5, 2017.

of a perfectly general and purely mechanic rule for extrapolating data — against the lesson that has generally been taken from Putnam that "[t]here is no universal algorithm" for induction (Dawid, 1985, 341; also see van Fraassen, 2000, 260). I will argue that there is promise in the general strategy that underlies the Solomonoff-Levin proposal, which is to try and identify a natural class of effective elements that is immune to diagonalization. This opens the prospect of attaining plausible versions of Putnam's two conditions that *are* compatible, and that enable a notion of a learning machine that is universal in a Reichenbachian sense: this *optimal* learning machine will learn successfully if any learning machine does. I will then show, however, that Putnam's lesson prevails: on a closer inspection of the proper interpretation of the relevant elements we see that this general strategy cannot escape diagonalization after all.

2. Overview

First, in section 3, I will introduce Putnam's original argument, which shows that no measure function can fulfill both of two conditions to qualify as a universal learning machine: the first on its convergence to any effectively computable hypothesis, the second on it being effectively computable itself. This is only one part of Putnam's charge; the other is that this is a defect peculiar to measure functions, because other methods, that respect the role of scientific theories (in particular, the hypothetico-deductive or HD method), *can* satisfy it. Next, in section 4, I explain how Solomonoff took his cue from Carnap's project, and went on to develop his ideas in a direction that (perhaps unlike Carnap's own approach) falls squarely within the general outlook and formal set-up that Putnam assumed for his argument. This raises the question how the resulting Solomonoff-Levin measure evades the diagonal argument.

The only way around Putnam's argument is to argue for a weakening of at least one of the two conditions that he showed are incompatible. Hence the question is what weakening the Solomonoff-Levin proposal introduces, and whether it can be given a proper motivation. To be in a position to answer this question, we need to go through a technical interlude, section 5, that traces the way to the exact definition of the Solomonoff-Levin measure. Here we will encounter the general strategy of identifying a class of effective measure functions that cannot be diagonalized; the Solomonoff-Levin measure is a universal element in this class. This definition does satisfy Putnam's first condition on convergence to any computable hypothesis, but it is effective in too weak a sense to still satisfy Putnam's second condition.

Turning to the question whether an accordingly weakened condition is defensible, we must first consider the second component of Putnam's charge. This is the claim that the conjunction of the two original conditions is not unreasonably strong, since the HD procedure does satisfy it. The conclusion that I reach in section 6 is that this claim does not stand up to scrutiny: drawing a distinction between specific *methods* and an underlying *architecture*, we see that the HD approach and the Bayesian approach of measure functions are in the exact same predicament. Given, then, that no specific method whatsoever can satisfy this pair of conditions, it stands to reason to explore the possibility of a notion of a universal learning machine that only satisfies a weaker pair. This we do in the final part of the paper, through an evaluation of the Solomonoff-Levin proposal.

I start in section 7 with the question whether the Solomonoff-Levin measure, in the spirit of the first condition, can detect all reasonable patterns. The naive interpretation of this question fails to be convincing, which prompts a different and much more natural interpretation. This Reichenbachian interpretation, pursued in section 8, takes the Solomonoff-Levin measure as *optimal* among all possible learning machines. If the original class of effective measure functions represents all possible learning machines, then the Solomonoff-Levin measure, as a universal element, is in a precise sense at least as good as any possible learning machine. In general, the identification of an undiagonalizable class of elements, if conjoined with a successful argument that it represents all possible learning machines, yields a notion of a universal learning machine.

Unfortunately, this strategy is obstructed by the fact that prediction methods should actually be identified with confirmation functions, i.e., *conditional* measure functions. This fact might sound innocuous, but it impacts the effectiveness properties of prediction methods. We will see that Putnam's original argument implies that this indeed blocks the central strategy of identifying a class of effective elements that cannot be diagonalized. Thus, as I conclude in section 9, our analysis provides further support to Putnam's case: there can be no such thing as a universal learning machine.

3. Putnam's argument

Consider a simple first-order language with a single monadic predicate G and an ordered infinity of individuals x_i, $i \in \mathbb{N}$. Let a *computable hypothesis* h be a computable set of sentences $h(x_i)$ for each individual x_i, where $h(x_i)$ equals one of Gx_i and $\neg Gx_i$. Now, if a given Carnapian measure function is supposed to be a rational reconstruction of our inductive practice, then, since our actual inductive methods would be sure to discern any computable pattern eventually, so should this given measure function. Hence a condition of adequacy on such a measure function P is that

(I) For any true computable hypothesis h, the *instance confirmation* $P(h(x_{n+1}) \mid h(x_0), ..., h(x_n))$ should pass and remain above threshold 0.5 after sufficiently many confirming individuals $x_0, ..., x_n$.

But for any measure function P that itself satisfies a weak condition of effective computability (so as to qualify, with the Church-Turing thesis, as an explicit method at all):

(II) For any true computable hypothesis h, for every n, it must be possible to compute an m such that if $h(x_{n+1}), ..., h(x_{n+m})$ hold, then $P(h(x_{n+m+1}) \mid h(x_0), ..., h(x_{n+m}))$ exceeds 0.5,

one can prove by diagonalization P's violation of (I). This is Putnam's diagonal argument: if the ideal inductive policy is to fulfill (I) and (II), then it is provably impossible to reconstruct it as a measure function.

We can treat condition (I) as an instance, for measure functions, of the general condition on an inductive method M that

(I*) M converges to any true computable hypothesis.

Moreover, in later expositions of the argument (Kelly, 2016, 701f), the slightly cumbersome condition (II) is often replaced by the (stronger) condition that P is

simply a computable function. The general condition on an inductive method M is that

(II*) M is computable.

The diagonal proof of the incompatiblity of (I*) and (II*) for measure functions is straightforward. Given candidate computable measure function P, we construct a computable hypothesis h such that P fails to converge on h, as follows. Starting with the first individual x_0, compute $P(Gx_0)$ and let $h(x_0)$ be $\neg Gx_0$ precisely if $P(Gx_0) > 0.5$. For each new individual x_{n+1}, proceed in the same fashion: compute $P(Gx_{n+1} \mid h(x_0), \ldots, h(x_n))$ and let $h(x_{n+1})$ be $\neg Gx_{n+1}$ precisely if this probability is greater than 0.5. The hypothesis h is clearly computable, but by construction the instance confirmation it is given by P never remains above 0.5: indeed, it never even goes above 0.5. Thus, again, if the ideal inductive policy is to be able to converge to any true computable hypothesis, *and* be computable itself, then it is impossible to reconstruct it as a measure function.

But maybe such a policy is so idealized as to escape any formalization? To seal the fate of Carnap's program, Putnam proceeds to give an example of an inductive method that is *not* based on a measure function and that *does* satisfy the two requirements. This method M is the *hypothetico-deductive method*: supposing some enumeration of hypotheses that are proposed over time, at each point in time select and use for prediction (*accept*) the hypothesis first in line among those that have been consistent with past data. Then it satisfies (I*), or more precisely:

(I†) For any true computable hypothesis h, if h is ever proposed, then M will eventually come to (and forever remain to) accept it.

The distinctive feature of M is that it relies on the hypotheses that are actually proposed. To Putnam, this is as it should be. Not only does it conform to scientific practice: more fundamentally, it does justice to the "*indispensability of theories* as instruments of prediction" (ibid., 778). This appears to be the overarching reason why Putnam takes issue with Carnap's program: "certainly it appears implausible to say that there is a *rule* whereby one can go from the observational facts ... to the observational prediction without any 'detour' into the realm of theory. But this is a consequence of the supposition that degree of confirmation can be 'adequately defined'" (ibid., 780). Incredulously: "we get the further consequence that it is possible in principle to build an electronic computer such that, if it could somehow be given all the observational facts, it would always make the best prediction—i.e. the prediction that would be made by the best possible scientist if he had the best possible theories. *Science could in principle be done by a moron* (or an electronic computer)" (ibid., 781).

Here Putnam is still careful not to attribute to Carnap too strong a view: "Of course, I am not accusing Carnap of believing or stating that such a rule exists; the existence of such a rule is a *disguised* consequence of the assumption that [degree of confirmation] can be 'adequately defined'" (ibid., 780). Carnap indeed showed some reluctance in committing himself to the idea of an "inductive machine" (1950, 192-99); though in other places he does appear close to endorsing it (see especially his 1966, 33-34; suggestive, too, is his use of the image of an inductive robot in 1962, 309ff). In any case, in his *Radio Free Europe* address (1963b), Putnam simply declares that "we may think of a system of inductive logic as a design for a 'learning machine': that is to say, a design for a computing machine that can extrapolate certain kinds of empirical regularities from the data with which it is supplied"

(ibid., 297); and "if there is such a thing as a correct 'degree of confirmation' which can be fixed once and for all, then a machine which predicted in accordance with the degree of confirmation would be an *optimal*, that is to say, a cleverest possible learning machine" (ibid., 298). Again, the diagonal proof would show that there can be no such thing: it is "an argument against the existence – that is, against the possible existence – of a 'cleverest possible' learning machine" (ibid., 299).

4. Solomonoff's new start

Solomonoff (1964) aimed to describe precisely that: an "optimum" learning machine, a formal system of inductive inference that "is at least as good as any other that may be proposed" (ibid., 5). His ideas can indeed be seen as a particular offspring of Carnap's inductive logic; one that takes Putnam's picture of a learning machine seriously.

Solomonoff's mission statement is clear: "The problem dealt with will be the extrapolation of a long sequence of symbols" (ibid., 2). What is the probability that a given (long) sequence T is followed by a (one-symbol) sequence a? "In the language of Carnap (1950), we want $c(a, T)$, the degree of confirmation of the hypothesis that a will follow, given the evidence that T has just occurred" (ibid.). The underlying motivation is also very much in accord with things Carnap writes in his 1950 book. Solomonoff's suggestion that "all problems in inductive inference ... can be expressed in the form of the extrapolation of a long sequence of symbols" (ibid.) parallels Carnap's insistence on the primacy of the predictive inference — "the most important and fundamental inductive inference" (1950, 207). And Carnap's discussion under the header "Are Laws Needed for Making Predictions?" (ibid., 574-75) — conclusion: "the use of laws is not indispensable" — is easily read as informing Solomonoff's proclamation that his proposed methods are "meant to bypass the explicit formulation of scientific laws, and use the data of the past directly to make inductive inferences about specific future events" (1964, 16).

This already very much resembles the picture that Putnam painted in order to challenge it. What is more, the problem setting of sequence extrapolation is readily translatable into the formal set-up that Putnam presupposes in his paper. Let us suppose, as is customary in modern discussions of Solomonoff's theory, that we have an alphabet of only two symbols, '0' and '1.' Now Putnam assumes with Carnap a monadic predicate language L, but with an *ordered* domain $x_0, x_1, x_2, \ldots$ of individuals. Let L have a single monadic predicate G. Identifying the individuals with positions in a sequence as Putnam does (1963a, 766), we can have a '1' at the $i+1$-th position express the fact that individual x_i satisfies G, and a '0' that it does not. Thus we translate a symbol sequence of length n into the observation of the first n individuals.

Solomonoff's setting is then fully within the scope of Putnam's argument. This in contrast to that of Carnap, who could still resort to the defense that in his works he does *not* assume an ordered domain, and so "the difficulties which Putnam discusses do not apply to the inductive methods which I have presented in my publications" (1963a, 986). Nevertheless, Carnap does acknowledge at various places the need for taking into consideration the order of individuals in explicating degree of confirmation (e.g., 1950, 62-65; 1963b, 225-26); and he envisioned for this future project the same kind of "coordinate language" that Putnam assumes (also see Skyrms, 1991). For such a language, Carnap should have agreed with

Putnam's charge that an inductive system that is "not 'clever' enough to learn that position in the sequence is relevant" is too weak to be adequate. The difference in opinion then ultimately comes down to *what* regularities in the observed individuals should be extrapolated (i.e., *what* hypotheses or patterns should gain higher instance confirmation from supporting observations).

Carnap states in (1963a, 987; 1963b, 226) that he would only consider "laws of finite span." In terms of symbol sequence extrapolation, these are the hypotheses that make the probability of a certain symbol's occurrence at a certain position only depend on the immediately preceding subsequence of a fixed finite length (i.e., a Markov chain of certain order). In particular, hypotheses must not refer to *absolute* coordinates, which immediately rules out Putnam's example of the hypothesis that "the prime numbers are occupied by red" (1963a, 765). In Carnap's view, "no physicist would seriously consider a law like Putnam's prime number law" (1963a, 987), hence "it is hardly worthwhile to take account of such laws in adequacy conditions for [confirmation functions]" (1963b, 226). According to Putnam, however, "existing inductive methods are capable of establishing the correctness of such a hypothesis ... and so must any adequate 'reconstruction' of these methods" (1963a, 765). Indeed, the same goes for *any* effectively computable pattern; this is his adequacy condition (I*).

Others have charged Carnap's confirmation functions with an inability to meet various adequacy conditions on recognizing regularities (notably Achinstein, 1963; in fact the critique of Goodman, 1946, 1947 can be seen as an early instance of this line of attack). What is distinctive about Putnam's adequacy conditions is the emphasis on effective computability. Interestingly, this notion of effective computability is also the fundamental ingredient in Solomonoff's proposal. It is this aspect that genuinely sets Solomonoff's approach apart from Carnap's. The measure functions that Solomonoff proposed in (1964), and that evolved in the modern definition of a measure Q_U that we will see below, were explicitly defined in terms of the inputs to a universal Turing machine. Moreover, one can show that the instance confirmation via Q_U of *any true computable hypothesis* will converge to 1, thus fulfilling (I*).

5. The Solomonoff-Levin measure

How does Solomonoff evade Putnam's diagonalization? If Q_U is within the scope of Putnam's argument, and it still fulfills (I*), then it must give way with respect to (II*). To see how Q_U fulfills (I*) but not (II*), we will need to go into the details. This we do in the current section; in the next section we return to the main thread and ask ourselves what this means for Q_U as a purported "optimum," or *universal* learning machine.

Specifically, we will work in this section towards the precise specification of Q_U, and prove that it satisfies (I*). For a large part this amounts to retracing the formal setting that was developed in the landmark paper of Zvonkin and Levin (1970), based on Levin's doctoral thesis (translated as Levin, 2010).

We start with the notion of a computable (probability) measure on the Cantor space $\{0,1\}^\omega$, the set of all infinite sequences of symbols in $\{0,1\}$. More accurately, a measure on Cantor space is defined on a tuple $(\{0,1\}^\omega, \mathfrak{F})$, with $\mathfrak{F}$ a σ-algebra on $\{0,1\}^\omega$. Then a probability measure on $(\{0,1\}^\omega, \mathfrak{F})$ is a countably additive function $\mu : \mathfrak{F} \to [0,1]$ with $\mu(\{0,1\}^\omega) = 1$. Let the *basic cylinder* $[\![\boldsymbol{x}]\!]$ be the

class of all infinite extensions in $\{0,1\}^\omega$ of the *finite* sequence $\boldsymbol{x} \in \{0,1\}^*$. It is convenient to view a measure (as well as the associated σ-algebra $\mathfrak{F}$) as being generated from an assignment of probability values to just the basic cylinders $[\![\boldsymbol{x}]\!]$ for all finite sequences $\boldsymbol{x}$. That is, we view a measure as being generated from a *pre-measure*, a function $m : \mathbb{B}^* \to [0,1]$ on the finite sequences that satisfies $m(\boldsymbol{\varnothing}) = 1$ for the *empty* sequence $\boldsymbol{\varnothing}$ and $m(\boldsymbol{x}0) + m(\boldsymbol{x}1) = m(\boldsymbol{x})$ for all $\boldsymbol{x} \in \{0,1\}^*$. The extension theorem due to Carathéodory (cf. Tao, 2011, 148ff) then gives a σ-algebra $\mathfrak{F}$ over $\{0,1\}^\omega$ (which includes all Borel classes) and unique measure μ_m on $\mathfrak{F}$ with $\mu_m([\![\boldsymbol{x}]\!]) = m(\boldsymbol{x})$. I will follow the custom of simply writing "$\mu(\boldsymbol{x})$" for "$\mu([\![\boldsymbol{x}]\!])$." (See Reimann, 2009, 249-256; Nies, 2009, 68-70 for more details.)

The most basic example of a measure on Cantor space is the *uniform* measure λ. It is generated from the premeasure with $m(\boldsymbol{x}) = 2^{-|\boldsymbol{x}|}$ for all $\boldsymbol{x}$, where $|\boldsymbol{x}|$ denotes $\boldsymbol{x}$'s length.

Now a measure is *computable* if it is generated from a computable pre-measure. A pre-measure is computable if its values can be uniformly computed up to any given precision. That is, there is a computable $f : \{0,1\}^* \times \mathbb{N} \to \mathbb{Q}$ such that $|f(\boldsymbol{x}, s) - m(\boldsymbol{x})| < 2^{-s}$ for all $\boldsymbol{x} \in \{0,1\}^*, s \in \mathbb{N}$ (cf. Downey and Hirschfeldt, 2010, 202-03). I will adopt the nomenclature of the *arithmetical hierarchy* of levels of effective computability (Kleene, 1943; Mostovski, 1947; see Soare, 1987, 60ff) and henceforth refer to the computable measures as the Δ_1 ("delta-one") measures.

We will see below that the Solomonoff-Levin measure Q_U has the property that for any true Δ_1 measure μ, with probability 1 ("μ-almost surely"), the values $Q_U(x_{n+1} \mid \boldsymbol{x}^n)$ for $x_{n+1} \in \{0,1\}, \boldsymbol{x}^n \in \{0,1\}^n$ converge to the values $\mu(x_{n+1} \mid \boldsymbol{x}^n)$ as n goes to infinity. That is, Q_U satisfies the following condition on a measure function P:

(I: Δ_1) P converges μ-almost surely to any true Δ_1 measure μ.

This is an instance of condition (I*) on a measure function, that at the same time generalizes from "deterministic" computable hypotheses or single infinite computable sequences to probability measures on infinite sequences. (Every computable infinite sequence $\boldsymbol{x}^\omega$ corresponds to a Δ_1 measure that assigns probability 1 to every initial segment $\boldsymbol{x}^n$ of $\boldsymbol{x}^\omega$.) Moreover, we can rephrase condition (II*) on a measure function as

(II: Δ_1) P is Δ_1.

This condition is *not* satisfied by Q_U. It is effectively computable in a weaker sense, that we turn to now.

Namely, we proceed with the notion of a *semi-computable* or Σ_1 ("sigma-one") measure on the extended space $\{0,1\}^\omega \cup \{0,1\}^*$ of infinite *and finite* sequences. This notion will strike those who see it for the first time as cumbersome, if not downright awkward; I will try to explain how it is both natural and important. First I will briefly describe how this class of measures comes about as precisely the *effective transformations* of the uniform measure on the Cantor space. Then I will discuss the crucial property of this class that *it cannot be diagonalized*, meaning that it contains *universal elements*. The Solomonoff-Levin measure is such a universal element.

Let a *transformation* λ_F of the uniform measure by Borel function $F : \{0,1\}^\omega \to \{0,1\}^\omega$ be defined by $\lambda_F(A) = \lambda(F^{-1}(A))$. It is a basic fact that every Borel measure μ on Cantor space can be obtained as a transformation of λ by some Borel function. What if we consider transformations by functions that are *computable*?

There are some details involved in the need to downscale these transformations to functions f on *finite* sequences, in order to impose the property of computability (see Reimann, 2009, 253f); in the end we are led to precisely those functions that can be represented by a particular type of Turing machine. Originally dubbed an *algorithmic process* (Zvonkin and Levin, 1970, 99), this type of machine is now better known as a *monotone* machine (see Shen et al. 2014, 139-142): it can be visualized as operating on a steady stream of input symbols, producing an (in)finite output sequence in the process. We then indeed have an effective analogue to the earlier statement: every Δ_1 measure can be obtained as a transformation λ_M of the uniform measure by some monotone machine M (Zvonkin and Levin, 1970, 100-01).

The monotone machines leading to the Δ_1 measures have the special property that they are "almost total," meaning that they produce an unending sequence on λ-almost all infinite input streams (ibid.). In general, however, a monotone machine M can fail to do so. This translates into the possibility that $\lambda_M(\boldsymbol{x})$ is strictly greater than $\lambda_M(\boldsymbol{x}0) + \lambda_M(\boldsymbol{x}1)$ for some $\boldsymbol{x}$. In that case we can say that λ_M assigns positive probability to the *finite* sequence $\boldsymbol{x}$. A function λ_M can thus be interpreted as a measure on the collection of infinite *and* finite sequences.

Levin calls the class of transformations λ_M by all monotone machines M the class of *semi-computable* measures on $\{0,1\}^\omega \cup \{0,1\}^*$. This is because the pre-measures corresponding to these transformations are precisely the functions $m : \{0,1\}^* \to [0,1]$ with $m(\boldsymbol{x}) \geq m(\boldsymbol{x}0) + m(\boldsymbol{x}1)$ for all $\boldsymbol{x}$ that satisfy a weaker requirement of computability, that we may paraphrase as *computable approximability from below* (Zvonkin and Levin, 1970, 102-03). In exact terms (cf. Downey and Hirschfeldt, 2010, 202-03), we call m (lower) semi-computable if there is a computable $f : \{0,1\}^* \times \mathbb{N} \to \mathbb{Q}$ such that for all $\boldsymbol{x} \in \{0,1\}^*$ we have $f(\boldsymbol{x}, s) \leq f(\boldsymbol{x}, s+1)$ for all $s \in \mathbb{N}$ and $\lim_{s\to\infty} f(\boldsymbol{x}, s) = m(\boldsymbol{x})$. Equivalently, the so-called *left-cut* $\{(q, \boldsymbol{x}) \in \mathbb{Q} \times \{0,1\}^* : q < m(\boldsymbol{x})\}$ is computably enumerable or Σ_1. For that reason I will refer to a semi-computable measure as a Σ_1 measure.

Let me reiterate the parallel between, on the one hand, the expansion from the Δ_1 to the Σ_1 measures, and, on the other, the expansion from the *total* computable (t.c.) to the *partial* computable (p.c.) functions. It is well-known since Turing (1936) that the class of t.c. functions is diagonalizable, and that this is overcome by enlarging the class to the p.c. functions (cf. Soare, 1987, 10ff). More precisely: under the assumption that there exists a *universal* t.c. function $\mathring{f}$ that can emulate every other t.c. function (meaning that $\mathring{f}(i, x) = f_i(x)$ for a listing $\{f_i\}_{i\in\mathbb{N}}$ of all t.c. functions), we can directly infer a *diagonal function* g (say $g(x) := \mathring{f}(x, x) + 1$) that is t.c. yet distinct from every single f_i (because $g(i) = f_i(i) + 1 \neq f_i(i)$ for all i), which is a contradiction. (Note the similarity to the argument in section 3.) To say that the class of t.c. functions is diagonalizable is therefore to say that there can be no such universal $\mathring{f}$, hence no effective listing of all elements: *the class is not effectively enumerable*. The introduction of partiality, however, defeats the construction of a diagonal function (consider: what if $f_i(i)$ is undefined?); and indeed the class of p.c. functions *is* effectively enumerable, *does* contain universal elements. Likewise, the class of Δ_1 measures is not effectively enumerable, does not contain universal elements; the larger class of Σ_1 measures is and does. We now turn to these universal Σ_1 measures (Zvonkin and Levin, 1970, 103-04).

Informally, a universal Σ_1 measure "is 'larger' than any other measure, and is concentrated on the widest subset of $[\{0,1\}^\omega \cup \{0,1\}^*]$" (ibid., 104). Formally, a universal Σ_1 measure $\mathring{\mu}$ is such that it *dominates* every other Σ_1 measure: for every $\mu_i \in \Sigma_1$ there is a constant $c_i \in [0,1]$ such that for all $\boldsymbol{x} \in \{0,1\}^*$ it holds that $\mathring{\mu}(\boldsymbol{x}) \geq c_i \mu_i(\boldsymbol{x})$. "This fact is one of the reasons for introducing the concept of semi-computable measure" (ibid.) — we may take it as the main reason. Indeed, the expansion to Σ_1 objects in order to obtain universal elements is a move that returns in many related contexts. Martin-Löf (1966), in defining his influential notion of *algorithmic randomness*, employed the class of all Σ_1 *randomness tests*: a sequence $\boldsymbol{x}^\omega$ is random if it passes a universal such test. Vovk (2001b), in defining his notion of *predictive complexity*, employed the class of Σ_1 *loss processes*: the predictive complexity of $\boldsymbol{x}^\omega$ is the loss incurred by a universal such process. Vovk and Watkins (1998, 17): "It would be ideal if the class of computable loss processes contains a smallest (say, to within an additive constant) element. Unfortunately ... such a smallest element does not exist." Levin's suggestion to widen the class to the Σ_1 elements is then "a very natural solution to the problem of non-existence of a smallest computable loss process" (ibid.).

The straightforward way of obtaining a universal Σ_1 measure is the following. Since the monotone machines are also effectively enumerable, we can likewise specify *universal* such machines. A transformation λ_U of λ by universal U then yields a universal Σ_1 measure.

We have finally arrived at the definition of the Solomonoff-Levin measure. The measure Q_U is precisely the transformation of λ by universal monotone machine U.

Definition 1. $Q_U := \lambda_U$.

So there are in fact infinitely many such measures, one for each choice of universal monotone machine U. Each is a universal Σ_1 measure. It is this property that is exploited in the adequacy result.

Proposition 2. Q_U fulfills (I*).

Proof. Let $\mu \in \Delta_1$. The fact that Q_U dominates μ entails that μ is absolutely continuous with respect to Q_U (i.e., $\mu(A) > 0$ implies $Q_U(A) > 0$ for all A in the σ-algebra $\mathcal{B}$), which by the classical result of Blackwell and Dubins (1962) entails that μ-almost surely the variational distance $\sup_{A \in \mathcal{B}} |\mu(A \mid \boldsymbol{x}^n) - Q_U(A \mid \boldsymbol{x}^n)| \to 0$ as $n \to \infty$ (see Huttegger, 2015, 617-18), so in particular (I: Δ_1). □

6. HD-methods and Bayesian methods

So how does the Solomonoff-Levin function evade Putnam's diagonalization? As we saw above, the very motivation for the expansion to the class of Σ_1 measures is to evade diagonalization — to obtain universal elements. The measure Q_U is a universal element; as such, it tracks every Δ_1 measure in the sense of (I: Δ_1). The downside is that, as a universal Σ_1 element, Q_U is itself no longer Δ_1 (or the class of Δ_1 measures would already have universal elements).

The force of Putnam's diagonal proof is that no measure function can satisfy both (I*) and (II*), and Q_U is no exception. The Solomonoff-Levin measure is powerful enough to avoid diagonalization and fulfill (I: Δ_1), but the price to pay is that Q_U might be said to be *too* powerful. It is no longer effective in the sense of (II: Δ_1), but only in the sense of

(II: Σ_1) P is Σ_1.

Does this invalidate Q_U as a learning machine — let alone a universal one?

One reply is that we cannot hold this against Q_U, since, after all, Putnam has shown that this incomputability is really a *necessary condition* for a policy to be optimal in the sense of (I*): "an optimal strategy, if such a strategy should exist, cannot be computable ... any optimal inductive strategy must exhibit recursive undecidability" (Hintikka, 1965, 283, fn. 22). However, this reply seems to miss the second component of Putnam's charge. This is the claim that, while no *measure function* can fulfill both adequacy conditions, *other methods* could — in particular, the HD-method.

In the current section we turn our attention to this claim. As discussed already in some detail by Kelly et al. (1994, 99-112), it actually turns out to be the weak spot in Putnam's argument. When we have this claim out of the way, we can, in the next section, follow up on the above reply and consider the question of Q_U's adequacy afresh.

Recall that we formulated (I*) and (II*) as conditions on inductive methods in general, not just measure functions. Again, Putnam (1963a, 770ff) takes it to be important for his case against Carnap that these conditions are not supposed to be mutually exclusive *a priori*; or it would be a rather moot charge that indeed no measure function can satisfy them in tandem. No measure function can satisfy both — conditions (I: Δ_1) and (II: Δ_1) are mutually exclusive — but other methods can: and the hypothetico-deductive (HD) method that Putnam describes is to be the case in point.

Crucially, however, Putnam's HD method depends on the hypotheses that are actually proposed in the course of time. The HD method fulfills (I†), which is so phrased as to accommodate this dependency: the method will come to accept (and forever stick to) any true computable hypothesis, *if* this hypothesis is ever proposed. Thus the HD method relies on some "hypothesis stream" (Kelly et al., 1994, 107) that is external to the method itself; and the method will come to embrace a true hypothesis whenever this hypothesis is part of the hypothesis stream.

In computability-theoretic terminology, the method uses the hypothesis stream as an *oracle*. The HD method is a simple set of rules, so obviously computable — *given* the oracle. But the oracle itself might be incomputable. Indeed, since the computable hypotheses are not effectively enumerable, any hypothesis stream that contains all computable hypotheses *is* incomputable. This is why Putnam must view the oracle as external to the HD method. The alternative is to view the generation of a particular hypotheses stream η as *part of the method itself*; but if any such HD-with-particular-hypothesis-stream-η method — let us simply say "HD$^\eta$ method" — is powerful enough to satisfy (I*), then the hypothesis stream and hence the method HD$^\eta$ as a whole must be incomputable. Putnam is well aware of this: "it is easily seen that any method that shares with Carnap the feature: what one will predict 'next' depends *only* on what has so far been observed, will also share the defect: either what one should predict will not in practice be *computable*, or some law will elude the method altogether" (Putnam, 1963a, 773). The diagonal proof described in section 3 readily generalizes to any method M: simply construct a computable sequence that goes against M's computable predictions at each point in time (cf. Kelly et al., 1994, 102-03).

In short, the HD$^\eta$ methods are in exactly the same predicament as Carnap's measure functions. Conditions (I*) and (II*) *are* mutually exclusive — unless we allow the method to be such that "the acceptance of a hypothesis also depends on *which* hypotheses are actually proposed" (Putnam, 1963a, 773), i.e., allow the method access to an external hypothesis stream.

But Putnam's assumption of an (incomputable) external oracle does, of course, raise questions of its own. The idea would be that we identify the oracle with the elusive process of the invention of hypotheses, the unanalyzable "context of discovery"; ultimately rooted, maybe, in "creative intuition" (Kelly et al., 1994, 108) or something of the sort. Is this process somehow incomputable? How would we know? More importantly, "if Putnam's favourite method is provided access to a powerful oracle, then why are Carnap's methods denied the same privilege?" (ibid., 107).

Kelly et al. offer Putnam the interpretation that the HD method provides an "architecture," a recipe for building particular methods (in our above terminology, HD$^\eta$ methods), that is "universal" in the sense that for every computable hypothesis, there is a particular computable instantiation of the architecture (a particular computable HD$^\eta$ method) that will come to accept (and forever stick to) the hypothesis if its true. "A scientist wedded to a universal architecture is shielded from Putnam's charges of inadequacy, since ... there is nothing one could have done by violating the strictures of the architecture that one could not have done by honoring them" (ibid., 110). Kelly et al. are not convinced, though, that their suggestion saves Putnam's argument, for the reason that it makes little sense for Putnam to endorse a universal architecture while calling every particular instance inadequate and therefore "*ridiculous*" (ibid., 110-11; here they quote Putnam, 1974, 238). There is, however, a more fundamental objection. Again, Putnam's argument against Carnap would only be completed if the above way out for the HD method were not open to measure functions. That is, it would only succeed if measure functions could not be likewise seen as instantiations of some universal architecture. But as a matter of fact, they can. They can be seen as instantiations of the *classical Bayesian* architecture. (Cf. Romeijn, 2004. I follow Diaconis and Freedman, 1986, 11 in adopting the designation "*classical* Bayesian." Also see Skyrms, 1996.)

The classical Bayesian architecture employs a countable *hypothesis class* (where hypotheses are again measures over Cantor space), as well as a *prior distribution* that gives positive probability to every element of this hypothesis class. Given a hypothesis class $\mathcal{H}$ and prior w, the corresponding Bayes-with-particular-hypothesis-class-$\mathcal{H}$ method $\xi_w^{\mathcal{H}}$ — let us say "Bayes$^{\mathcal{H}}$ method" $\xi_w^{\mathcal{H}}$ — is the measure function that is simply the w-weighted mean over the hypotheses in $\mathcal{H}$, i.e., $\xi_w^{\mathcal{H}}(\boldsymbol{x}) := \sum_{h\in\mathcal{H}} w(h)h(\boldsymbol{x})$.

The classical Bayesian architecture is a universal architecture because for every (computable) deterministic hypothesis, there is a particular (computable) instantiation of the architecture (a Bayes$^{\mathcal{H}}$ method where $\mathcal{H}$ contains the hypothesis) that will converge on it when it is true. Just like the HD architecture is guaranteed to converge on (i.e, accept and stick to) every true deterministic hypothesis, *whenever* it is included in the hypothesis stream, so the classical Bayesian architecture is guaranteed to converge on every true deterministic hypothesis, *whenever* it is included in the hypothesis class. More generally, to also cover the case where the true hypothesis is in fact probabilistic, a Bayes$^{\mathcal{H}}$ method will come to accept

and forever stick to any true hypothesis μ, with probability 1 ("μ-almost surely"), whenever it is in $\mathcal{H}$. This property is also known as Bayesian *consistency*. It follows from the exact same argument as the proof of Theorem 2, given the fact that $\xi_w^{\mathcal{H}}$ *dominates* every element in $\mathcal{H}$: for every $h \in \mathcal{H}$ we clearly have for all $\boldsymbol{x} \in \mathbb{B}^*$ that $\xi_w^{\mathcal{H}}(\boldsymbol{x}) \geq w(h)h(\boldsymbol{x})$.

Every measure function over Cantor space corresponds to a Bayes$^{\mathcal{H}}$ method for some $\mathcal{H}$ and w. We can thus interpret any measure function as relying on a class of hypotheses — meeting Putnam's insistence on the indispensability of theory. Moreover, this point of view naturally accommodates a *simplicity ordering* of hypotheses that Putnam (inspired by Kemeny, 1953) envisages a refined HD method to employ (1963a, 775-77), and that in (1963b, 301-02) he proposes as a line of further investigation for inductive logic: "given a simplicity ordering of some hypotheses, to construct a c-function which will be in agreement with that simplicity ordering, that is, which will permit one to extrapolate any one of those hypotheses, and which will give the preference always to the earliest hypothesis in the ordering which is compatible with the data" (ibid., 302). The solution to this problem is the measure function Bayes$^{\mathcal{H}}$ with a prior w that expresses the desired simplicity ordering on the hypotheses in $\mathcal{H}$, assigning lower probability to hypotheses further away in the ordering.

In conclusion of this discussion, there is a perfect analogy between the situation for the HD method and for the classical Bayesian method. No *particular* measure function — Bayes$^{\mathcal{H}}$ method — can satisfy both (I*) and (II*). But, similarly, no *particular* HD$^{\eta}$ method can satisfy both (I*) and (II*). Nevertheless, the HD *architecture* is universal. But, similarly, the classical Bayesian *architecture* is universal. From this perspective, Putnam's argument, purporting to show that measure functions have fundamental shortcomings that other methods do not, fails.

7. A universal learning machine

We have observed that (I*) and (II*) are mutually exclusive: no particular method can satisfy both. Let us then follow up on the earlier suggestion to not dismiss the Solomonoff-Levin function Q_U out of hand simply because it does not satisfy the special cases (I: Δ_1) and (II: Δ_1) — that it cannot do the impossible. Instead, let us conclude our investigation with a fresh look at the question: could Q_U be an adequate characterization of a "cleverest possible," a *universal* learning machine?

We can still, with Putnam, divide this question into two parts. First, in the spirit of (I*), will Q_U be able to accept every reasonable (reasonably effective) hypothesis, if it is true? Second, in the spirit of (II*), is Q_U itself still a reasonable (reasonably effective) method?

To start with the first. Could Q_U be called universal in the sense that it is able to track *any* reasonable hypothesis? The best vantage point to address this question is to view Q_U as an instantiation of the classical Bayesian architecture that we saw in the previous section. It turns out that the measure functions Q_U are the classical Bayesian methods that employ the class of all Σ_1 hypotheses (see Sterkenburg, 2016). To be exact, the measure functions Q_U are precisely the Bayes$^{\mathcal{H}_{\Sigma_1}}$ methods $\xi_w^{\Sigma_1}$ with semi-computable prior w over the hypothesis class $\mathcal{H}_{\Sigma_1}$ of all Σ_1 measures. (In particular, the choice of universal machine U corresponds to the choice of semi-computable prior w over $\mathcal{H}_{\Sigma_1}$.) By Bayesian consistency, it follows that Q_U will

almost surely converge on any true Σ_1 hypothesis. (This is again, in essence, Theorem 2 above, though I only stated it for Δ_1 measures. See the Appendix for details.)

The hypothesis class embodies the regularities that can be extrapolated, the patterns that should gain higher instance confirmation from supporting instances. Thus we may rephrase our question: is the hypothesis class $\mathcal{H}_{\Sigma_1}$ sufficiently wide, sufficiently general?

Before we turn to an answer, we connect this question to an important alternative perspective on Q_U. This is the interpretation of Q_U as an "a priori" distribution over the symbol sequences. Measure Q_U "corresponds to what we intuitively understand by the words 'a priori probability,'" Zvonkin and Levin (1970, 104) write, because "if nothing is known in advance about the properties of [a] sequence, then the only (weakest) assertion we can make regarding it is that it can be obtained randomly with respect to $[Q_U]$". This is an illustration of how the question of the generality of $\mathcal{H}_{\Sigma_1}$ — the class of candidate measures that may be assumed to generate the data — is related to the question of the adequacy of Q_U as an a priori probability assignment on the data sequences. Ultimately, the latter perspective is associated with the idea that inductive reasoning attains justification from some objective or rational starting point. It is in this spirit that Carnap (1962) writes that against our credences that are derived from a rational initial credence function (i.e., measure function), "Hume's objection does not hold, because [we] can give rational reasons for it" (ibid., 317): the rationality requirements that are codified as axioms constraining the measure function. It also seems in this spirit that Li and Vitányi (2008), presenting Q_U as a "universal prior distribution," make reference to Hume and claim that the "perfect theory of induction" invented by Solomonoff "may give a rigorous and satisfactory solution to this old problem in philosophy" (ibid., 347).

The problem with this idea is, to begin, that there is still subjectivity involved in pinning down the exact starting point. The choice of initial credence function (measure function) is "guided (*though not uniquely determined*) by the axioms of inductive logic" (Carnap, 1971, 30, emphasis mine). Likewise, the definition of Q_U still leaves open the choice of U — from the classical Bayesian perspective, the choice of semicomputable prior w over $\mathcal{H}_{\Sigma_1}$ (cf. Sterkenburg, 2016, 471-74). One might reply, with Jeffrey (1973, 302), that what we have here is only a "latituditarianism" where different possible choices "are sufficiently similar so that their differences are swamped out by experience," a result of having priors over the same hypothesis class. Indeed, bracketing this issue here, we still face the more fundamental problem: the problem of justifying the stipulated constraints on the measure functions, i.e., from the classical Bayesian perspective, the problem of justifying the choice of hypothesis class. And that brings us back to the question of the generality of the hypothesis class.

As Howson (2000) argues at length, the choice of prior distribution — what hypotheses are assigned nonzero prior probability — constitutes our inevitable "Humean inductive assumptions." "According to Hume's circularity thesis, every inductive argument has a concealed or explicit circularity. In the case of probabilistic arguments ... this would manifest itself on analysis in some sort of prior loading in favour of the sorts of 'resemblance' between past and future we thought desirable. Well, of course, we have seen exactly that: *the prior loading is supplied by*

the prior probabilities" (ibid., 88). (Also see Romeijn, 2004, 357ff.) It is important for the observation that Bayesian methods cannot escape Hume's argument that inductive assumptions must be *restrictive*: that it is impossible to have a prior over *everything* that could be true. That is, from the classical Bayesian perspective, it must be the case that no hypothesis class $\mathcal{H}$ can contain every possible hypothesis, that no $\mathcal{H}$ is fully general.

Could $\mathcal{H}_{\Sigma_1}$, then, escape Hume's argument — is $\mathcal{H}_{\Sigma_1}$ fully general? Naturally, it is not. As a restriction on what hypotheses could ever be *true*, really a *metaphysical* assumption on the world, not only would the restriction to any specific level of effective computability (Δ_1, Σ_1, ...) look arbitrary: the assumption of effective computability itself is a stipulation that wants motivation.

8. An optimal learning machine

There is, however, an alternative interpretation still. This interpretation is to take the elements of the class $\mathcal{H}_{\Sigma_1}$, not as hypotheses about the origin of the data, but as *competing learning machines*.

This interpretation is actually more in line with Putnam's demand that the cleverest possible learning machine should be able to eventually pick up any pattern *that our actual inductive methods would*. It is also more in line with Solomonoff's original aim that given "a very large body of data, the model is *at least as good as any other that may be proposed*" (1964, 5, emphasis mine). (Noteworthy, moreover, is that Solomonoff's basic idea of sequential prediction by a mixture over the elements of a general class $\mathcal{H}$ is the starting point of a currently very active branch of machine learning; here the stated goal is indeed to predict at least as well as any member of a pool $\mathcal{H}$ of competing "experts" without assumptions on the origin of the data. See, for instance, Vovk, 2001a; Cesa-Bianchi and Lugosi, 2006.)

Let us see what we get when we thus reinterpret the Σ_1 measures as *all possible learning machines*. As a start, Theorem 2 could be reinterpreted as a fully general *merging-of-opinions* result (see Huttegger, 2015): every learning machine anticipates with certainty that Q_U's confirmation values converge to its own. Moreover, it is easy to derive the following more "absolute" fact. For any learning machine ν, there is a constant bound on the surplus *logarithmic loss* (expressing the divergence between the given confirmation values and the symbols that actually obtain) incurred by Q_U relative to this learning machine ν, on *any* symbol sequence (see the Appendix for details). Thus, if we take the Σ_1 measures as all possible learning machines, then Q_U is a universal learning machine in the following powerful sense: *it is a learning machine that compared to* any *other learning machine will always come to perform at least as well*.

We may brand this the *optimality* interpretation: rather than *reliable* (guaranteed with certainty to converge on the true hypothesis), Q_U is *optimal* in the sense that it is guaranteed to converge on the true hypothesis *if any learning machine does*. The learning machine Q_U is *vindicated* in the sense of Reichenbach (see Salmon, 1991).

If we accept this, then Q_U *is* a universal learning machine — defying Putnam's lesson that there can be no such thing (see, in particular, the discussion of van Fraassen, 2000, 257ff against a Reichenbachian conception of a universal inductive rule). As we have seen, the crucial move to unlock this possibility after all, hence the crucial precondition to our optimality interpretation, is the expansion to the

nondiagonalizable class of Σ_1 elements. The moment has come to answer the question whether this move is reasonable at all. Specifically, we need to answer the question that is the analogue in this interpretation to the first question we started the previous section with: is it reasonable to identify all possible learning machines with the Σ_1 measures?

Most importantly, is the class of Σ_1 measures not *too* wide — does a Σ_1 measure that fails to be Δ_1 still constitute a proper learning machine? As a special case, we have returned to the second question we started the previous section with: does Q_U itself constitute a reasonable (reasonably effective) method?

Now an incomputable measure function is certainly "impractical" (Cover et al., 1989, 863), or indeed "of no use to anybody" (Putnam, 1963a, 768) in any practical way — but that already goes for any measure function that *is* computable but not in some sense *efficiently* so. The minimal requirement that Putnam was after is computability *in principle*, i.e., given an unlimited amount of space and time. Indeed, under the Church-Turing thesis, computability is just what it *means* to be (in principle) implementable as an explicit method — computability is the minimal requirement to be a method at all. On this view, a Δ_1 measure is a measure that corresponds to a method that (given unlimited resources) for any finite sequence returns the probability that the measure assigns to it. But, likewise, a Σ_1 measure still corresponds to a method that (given unlimited resources) for any finite sequence returns *increasingly accurate approximations* of its probability. So, albeit in a weaker sense, a Σ_1 measure is still connected to some explicit method. (Cf. Martin-Löf, 1969, 268 on his choice of Σ_1 randomness tests: "on the basis of Church's thesis it seems safe to say that this is the most general definition we can imagine as long as we confine ourselves to tests which can actually be carried out and are not pure set theoretic abstractions.")

This seems good — but we passed over a crucial detail. This is the fact that for the purpose of inductive reasoning, we are actually interested in the *conditional* probabilities issued by the measure functions: those are the confirmation values. For that reason inductive methods or learning machines should actually be identified with two-place *confirmation functions* rather than the underlying one-place measure functions. But this has repercussions for the level of effectiveness.

This aspect is easy to oversee, because for the Δ_1 measures it makes no difference. If a measure μ is Δ_1, so μ as a function on finite sequences is computable, then (and only then) the two-place function $\mu(\cdot \mid \cdot)$, given by $\mu(x_{n+1} \mid \boldsymbol{x}^n) = \mu(\boldsymbol{x}^{n+1})/\mu(\boldsymbol{x}^n)$, is computable as well. Thus the Δ_1 measures correspond precisely to the Δ_1 conditional measures, or confirmation functions. However, for the Σ_1 measures this *does* make a difference. In particular, the *conditional* Solomonoff-Levin function $Q_U(\cdot \mid \cdot)$ is no longer Σ_1.

As a matter of fact, this follows from Putnam's original diagonalization argument, that shows the incompatibility of the conditions (I) and (II) that we started with in section 3. Namely, if $Q_U(\cdot \mid \cdot)$ were Σ_1, then Q_U would satisfy (II): if, on some infinite sequence $\boldsymbol{x}^\omega$, the value $Q_U(x_{n+1} \mid \boldsymbol{x}^n)$ will for large enough n always be above 0.5, then we can computably locate an m with $Q_U(x_{m+1} \mid \boldsymbol{x}^m) > 0.5$. For completeness, the following proof recounts the details of the diagonalization (cf. Putnam, 1963a, 768f, Putnam, 1963b, 299). (A different proof has been given by Leike and Hutter, 2015, 370-71.)

Proposition 3. $Q_U(\cdot \mid \cdot) \notin \Sigma_1$.

Proof. Suppose towards a contradiction that $Q_U(\cdot \mid \cdot)$ is Σ_1, so that (II) holds for Q_U. We can now construct a computable infinite sequence $\boldsymbol{x}^\omega$ as follows. Start calculating $Q_U(0 \mid 0^n)$ from below in dovetailing fashion for increasing $n \in \mathbb{N}$, until an n_0 such that $Q_U(0 \mid 0^{n_0}) > 0.5$ is found (since Q_U satisfies (I) such n_0 must exist). Next, calculate $Q(0 \mid 0^{n_0}10^n)$ for increasing n until an n_1 with $Q(0 \mid 0^{n_0}10^{n_1}) > 0.5$ is found. Continuing like this, we obtain a list $n_0, n_1, n_2, ...$ of positions; let $\boldsymbol{x}^\omega := 0^{n_0}10^{n_1}10^{n_2}1\ldots$. Sequence $\boldsymbol{x}^\omega$ is computable, but by construction the instance confirmation of $\boldsymbol{x}^\omega$ will never remain above 0.5, contradicting (I). □

Now we could argue that $Q_U(\cdot \mid \cdot)$ is still Δ_2 or *limit computable*, meaning that it still corresponds to a method that converges to any given finite sequence's probability in the limit (cf. ibid., 365). But the problem runs deeper. The problem is that we cannot recover the optimality interpretation for conditional measures.

Namely, if we accept that a Δ_2 confirmation function (i.e., a Δ_2 conditional measure) still counts as a possible learning machine, then we should identify the possible learning machines with the class of Δ_2 confirmation functions (rather than the original class of confirmation functions with underlying Σ_1 measure functions). That means that the sought-for optimality would have to be relative to *this* class. But $Q_U(\cdot \mid \cdot)$ is not optimal among the Δ_2 confirmation functions — *no* Δ_2 confirmation function is. This is because the class of Δ_2 *measure* functions, that precisely induces the class of Δ_2 *confirmation* functions, *is* diagonalizable: just like in the Δ_1 case, one can, for any given Δ_2 measure function, construct a Δ_2 sequence that it will never converge on.

Nor can we take a step back and settle for the class of Σ_1 confirmation functions. Once again it follows from Putnam's argument above that there cannot exist universal elements in the class of measure functions that induce the Σ_1 confirmation functions: for any given Σ_1 conditional measure one can construct a Σ_1 conditional measure (indeed even a computable sequence) it will never converge on.

All of this easily generalizes to higher levels ("relativizes" in computability-theoretic jargon): the strategy for optimality cannot work on any level in the arithmetical hierarchy.

9. Conclusion

Thus we conclude our story on an unhappy note. We have discussed how Putnam's diagonal argument shows that no method whatsoever — not just the methods expressed as measure functions — can satisfy at the same time two conditions to qualify as a universal learning machine: the one on the ability to detect every true effectively computable pattern, the other on the effective computability of the method itself. Faced with this impossibility result, we allowed ourselves to consider as candidate universal learning machines measure functions that only satisfy a weaker pair of conditions; specifically, we considered the Solomonoff-Levin measure. The overarching strategy we identified to bring versions of the two conditions together is to locate a natural class of effective measure functions that cannot be diagonalized, i.e., that contains universal elements. If one could reasonably identify this class of measure functions with all possible learning machines, then the universal elements would be vindicated as universally optimal learning machines: they constitute learning machines that are in a strong sense at least as good as

any other learning machine. In particular, we saw that the Solomonoff-Levin measures were constructed as universal elements among the Σ_1 measures — and so, our hope ran, they could qualify as such optimal learning machines. Unfortunately, we found a fatal flaw in this strategy: learning machines should be identified with two-place confirmation (conditional measure) functions rather than the underlying one-place measure functions. This affects their effectiveness properties, which ultimately means that no level in the arithmetical hierarchy yields an undiagonalizable class of learning machines. Putnam's argument stands.

Appendix

Theorem 2 is in the literature (Li and Vitányi, 2008, 352-56; Hutter, 2003, 2062; Poland and Hutter, 2005, 3781) usually presented as a consequence of (variations of) the following stronger result, first shown by Solomonoff (1978, 426-27). Let us introduce as a measure of the divergence between two distributions P_1 and P_2 over $\{0,1\}$ the squared *Hellinger distance*

$$H(P_1, P_2) := \sum_{x\in\{0,1\}} \left(\sqrt{P_1(x)} - \sqrt{P_2(x)}\right)^2. \tag{1}$$

Then, for every $\mu \in \Delta_1$, the expected infinite sum of divergences between Q_U and μ

$$\mathbf{E}_{X^\omega\sim\mu}\left[\sum_{n=0}^{\infty} H\left(\mu(\cdot \mid X^n), Q_U(\cdot \mid X^n)\right)\right] \tag{2}$$

is bounded by a constant.

To see how (I: Δ_1) follows from this constant bound, suppose that Q_U does not satisfy (I: Δ_1): there is a $\mu \in \Delta_1$ such that with probability $\epsilon > 0$ there is a $\delta > 0$ such that $|\mu(x_{n+1} \mid \boldsymbol{x}^n) - Q_U(x_{n+1} \mid \boldsymbol{x}^n)| > \delta$ infinitely often. But that means that with positive probability the infinite sum of squared Hellinger distances is infinite, and the expectation (2) cannot be bounded by a constant.

The proof of the constant bound on (2) starts with the observation that the distance $H(P_1, P_2)$ is bounded by the *Kullback-Leibler divergence*

$$D(P_1 \parallel P_2) := \mathbf{E}_{X\sim P_1}\left[\ln \frac{P_1(X)}{P_2(X)}\right]. \tag{3}$$

The term $-\ln P(\boldsymbol{x})$ expresses the *logarithmic loss* of P on sequence $\boldsymbol{x}$, a standard measure of prediction error; the difference $-\ln P_2(\boldsymbol{x}) - (-\ln P_1(\boldsymbol{x})) = \ln \frac{P_1(\boldsymbol{x})}{P_2(\boldsymbol{x})}$ expresses the surplus prediction error or *regret* of P_2 relative to P_1 on sequence $\boldsymbol{x}$. Thus the Kullback-Leibler divergence (3) expresses the expected regret of P_2 relative to P_1.

Using $H(P_1, P_2) \leq D(P_1 \parallel P_2)$ one can work out that (2) is bounded by

$$\mathbf{E}_{X^\omega\sim\mu}\left[\sum_{n=0}^{\infty} \ln \frac{\mu(X_{n+1} \mid X^n)}{Q_U(X_{n+1} \mid X^n)}\right]. \tag{4}$$

Now by the universality of Q_U in the class of Σ_1 measures we know that Q_U dominates μ: for every finite $\boldsymbol{x}$ there is a constant c such that $Q_U(\boldsymbol{x}) \geq \mu(\boldsymbol{x})/c$. Indeed we can identify c with $1/w(\mu)$, where w is the prior over hypothesis class $\mathcal{H}_{\Sigma_1}$ in the classical Bayesian representation $\xi_w^{\Sigma_1}$ of Q_U. This fact allows us to derive that *for every sequence*

$\boldsymbol{x}^m$ *of any length* m

$$\begin{aligned}\sum_{n=0}^{m-1} \ln \frac{\mu(x_{n+1} \mid \boldsymbol{x}^n)}{Q_U(x_{n+1} \mid \boldsymbol{x}^n)} &= \ln \prod_{n=0}^{m-1} \frac{\mu(x_{n+1} \mid \boldsymbol{x}^n)}{Q_U(x_{n+1} \mid \boldsymbol{x}^n)} \\ &= \ln \frac{\mu(\boldsymbol{x}^m)}{Q_U(\boldsymbol{x}^m)} \\ &\leq -\ln w(\mu). \qquad (5)\end{aligned}$$

This concludes the proof that (2) is bounded by a constant: since the bound (5) holds for any individual sequence of any length, it also holds for (4) and thus for (2).

Theorem 2 was in the main text only stated for measures μ in Δ_1: measures over $\{0,1\}^\omega$. To retrieve the merging-of-opinions variant of this result mentioned in the main text, we need to make it go through for Σ_1 measures, measures over $\{0,1\}^\omega \cup \{0,1\}^*$ — indeed we need to make precise what "almost surely" should mean for such "semi-measures." We can do this as follows. Let a $\nu \in \Sigma_1$ be represented by a measure ν' over $\{0,1,\mathsf{s}\}^\omega$, with '$\mathsf{s}$' a "stopping symbol": we have $\nu'(\boldsymbol{x}0) + \nu'(\boldsymbol{x}1) + \nu'(\boldsymbol{x}\mathsf{s}) = \nu'(\boldsymbol{x})$ and we stipulate $\nu'(\boldsymbol{x}) = \nu(\boldsymbol{x})$ and $\nu'(\boldsymbol{x}\mathsf{ss}) = \nu'(\boldsymbol{x}\mathsf{s})$ for all $\boldsymbol{x} \in \{0,1\}^*$. Then for all $\nu \in \Sigma_1$ we have that Q'_U dominates ν', hence $\nu' \ll Q'_U$ and the Blackwell-Dubins theorem applies as before.

The absolute optimality property mentioned in the main text is just the individual sequence bound (5) above. To reformulate, for any $\nu \in \Sigma_1$, the sum of surplus prediction errors (regrets) of Q_U relative to ν will *always* (for any sequence $\boldsymbol{x}^m$ of any length m) be bounded by a constant:

$$\sum_{n=0}^{m-1} \left(-\ln Q_U(x_{n+1} \mid \boldsymbol{x}^n) - \left(-\ln \nu(x_{n+1} \mid \boldsymbol{x}^n)\right)\right) \leq -\ln w(\nu).$$

References

P. Achinstein. Confirmation theory, order, and periodicity. *Philosophy of Science*, 30:17–35, 1963.

D. Blackwell and L. Dubins. Merging of opinion with increasing information. *The Annals of Mathematical Statistics*, 33:882–886, 1962.

R. Carnap. *Logical Foundations of Probability*. The University of Chicago Press, Chicago, IL, 1950.

R. Carnap. The aim of inductive logic. In E. Nagel, P. Suppes, and A. Tarski, editors, *Logic, Methodology and Philosophy of Science. Proceedings of the 1960 International Congress*, pages 303–318. Stanford University Press, Stanford, CA, 1962.

R. Carnap. Replies and systematic expositions. In P. A. Schilpp, editor, *The Philosophy of Rudolf Carnap*, volume XI of *The Library of Living Philosophers*, pages 859–1013. Open Court, LaSalle, IL, 1963a.

R. Carnap. Variety, analogy, and periodicity in inductive logic. *Philosophy of Science*, 30(3): 222–227, 1963b.

R. Carnap. *Philosophical Foundations of Physics: An Introduction to the Philosophy of Science*. Basic Books, New York, 1966.

R. Carnap. Inductive logic and rational decisions. In R. Carnap and R. C. Jeffrey, editors, *Studies in Inductive Logic and Probability*, volume 1, pages 5–31. University of California Press, 1971.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning and Games*. Cambridge University Press, Cambridge, 2006.

T. M. Cover, P. Gács, and R. M. Gray. Kolmogorov's contributions to information theory and algorithmic complexity. *The Annals of Probability*, 17(3):840–865, 1989.

A. P. Dawid. The impossibility of inductive inference. Comment on Oakes (1985). *Journal of the American Statistical Association*, 80(390):340–341, 1985.

P. Diaconis and D. Freedman. On the consistency of Bayes estimates. *The Annals of Statistics*, 14(1):1–26, 1986.

R. G. Downey and D. R. Hirschfeldt. *Algorithmic Randomness and Complexity*, volume 1 of *Theory and Applications of Computability*. Springer, New York, 2010.

N. Goodman. A query on confirmation. *The Journal of Philosophy*, 43(14):383–385, 1946.

N. Goodman. On infirmities of confirmation-theory. *Philosophy and Phenomenological Research*, 8(1):149–151, 1947.

J. Hintikka. Towards a theory of inductive generalization. In Y. Bar-Hillel, editor, *Logic, Methodology and Philosophy of Science. Proceedings of the 1964 International Congress*, Studies in Logic and the Foundations of Mathematics, pages 274–288. North-Holland, Amsterdam, 1965.

C. Howson. *Hume's Problem: Induction and the Justification of Belief*. Oxford University Press, New York, 2000.

S. M. Huttegger. Merging of opinions and probability kinematics. *The Review of Symbolic Logic*, 8(4):611–648, 2015.

M. Hutter. Convergence and loss bounds for Bayesian sequence prediction. *IEEE Transactions on Information Theory*, 49(8):2061–2067, 2003.

R. C. Jeffrey. Carnap's inductive logic. *Synthese*, 25:299–306, 1973.

K. T. Kelly. Learning theory and epistemology. In H. Arló-Costa, V. F. Hendricks, and J. F. A. K. van Benthem, editors, *Readings in Formal Epistemology*, volume 1 of *Graduate Texts in Philosophy*, pages 695–716. Springer, 2016.

K. T. Kelly, C. F. Juhl, and C. Glymour. Reliability, realism, and relativism. In P. Clark and B. Hale, editors, *Reading Putnam*, pages 98–160. Blackwell, Oxford, 1994.

J. Kemeny. The use of simplicity in induction. *Philosophical Review*, 62(3):391–408, 1953.

S. C. Kleene. Recursive predicates and quantifiers. *Transactions of the American Mathematical Society*, 53:41–73, 1943.

J. Leike and M. Hutter. On the computability of Solomonoff induction and knowledge-seeking. In K. Chaudhuri, C. Gentile, and S. Zilles, editors, *Algorithmic Learning Theory. Proceedings of the 26th International Conference, ALT 2015*, volume 9355 of *Lecture Notes in Artificial Intelligence*, pages 364–378. Springer, 2015.

L. A. Levin. Some theorems on the algorithmic approach to probability theory and information theory. *Annals of Pure and Applied Logic*, 162:224–235, 2010. Translation of PhD dissertation, Moscow State University, Russia, 1971.

M. Li and P. M. B. Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications*. Texts in Computer Science. Springer, New York, third edition, 2008.

P. Martin-Löf. The definition of random sequences. *Information and Control*, 9:602–619, 1966.

P. Martin-Löf. Algorithms and randomness. *Review of the International Statistical Institute*, 37 (3):265–272, 1969.

A. Mostovski. On definable sets of positive integers. *Fundamenta Mathematicae*, 34:81–112, 1947.

A. Nies. *Computability and Randomness*, volume 51 of *Oxford Logic Guides*. Oxford University Press, 2009.

J. Poland and M. Hutter. Asymptotics of discrete MDL for online prediction. *IEEE Transactions on Information Theory*, 51(11):3780–3795, 2005.

H. Putnam. 'Degree of confirmation' and inductive logic. In P. A. Schilpp, editor, *The Philosophy of Rudolf Carnap*, volume XI of *The Library of Living Philosophers*, pages 761–783. Open Court, LaSalle, IL, 1963a.

H. Putnam. Probability and confirmation. In *The Voice of America Forum Lectures, Philosophy of Science Series 10*. U.S. Information Agency, Washington, D.C., 1963b. Page numbers refer to reprint in *Mathematics, Matter, and Method*, volume 1, Cambridge University Press, Cambridge, 1975, pages 293-304.

H. Putnam. The "corroboration" of theories. In P. A. Schilpp, editor, *The Philosophy of Karl Popper, Book I*, volume XIV of *The Library of Living Philosophers*, pages 221–240. Open Court, La Salle, IL, 1974.

J. Reimann. Randomness—beyond Lebesgue measure. In S. B. Cooper, H. Geuvers, A. Pillay, and J. Väänänen, editors, *Logic Colloquium 2006*, volume 32 of *Lecture Notes in Logic*, pages 247–279. Association for Symbolic Logic, Chicago, IL, 2009.

J.-W. Romeijn. Hypotheses and inductive predictions. *Synthese*, 141(3):333–364, 2004.

W. C. Salmon. Hans Reichenbach's vindication of induction. *Erkenntnis*, 35:99–122, 1991.

A. K. Shen, V. A. Uspensky, and N. K. Vereshchagin. Kolmogorov complexity and algorithmic randomness. Unpublished translation (available at http://www.lirmm.fr/~ashen/) of Russian edition, MCCME Publishing House, Moscow, Russia, 2014.

B. Skyrms. Carnapian inductive logic for Markov chains. *Erkenntnis*, 35:439–460, 1991.

B. Skyrms. Carnapian inductive logic and Bayesian statistics. In T. Ferguson, L. Shapley, and J. MacQueen, editors, *Statistics, Probability and Game Theory: Papers in Honor of David Blackwell*, volume 30 of *Lecture Notes - Monograph Series*, pages 321–336. Institute of Mathematical Statistics, 1996.

R. I. Soare. *Recursively Enumerable Sets and Degrees: A Study of Computable Functions and Computably Generated Sets*. Perspectives in Mathematical Logic. Springer, 1987.

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

R. J. Solomonoff. Complexity-based induction systems: Comparisons and convergence theorems. *IEEE Transactions on Information Theory*, IT-24(4):422–432, 1978.

T. F. Sterkenburg. Solomonoff prediction and Occam's razor. *Philosophy of Science*, 83(4):459–479, 2016.

T. Tao. *An Introduction to Measure Theory*, volume 126 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2011.

A. M. Turing. On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, 2(42):230–265, 1936.

B. C. van Fraassen. The false hopes of traditional epistemology. *Philosophy and Phenomenological Research*, 60(2):253–280, 2000.

V. Vovk. Competitive on-line statistics. *International Statistical Review*, 69(2):213–248, 2001a.

V. Vovk. Probability theory for the Brier game. *Theoretical Computer Science*, 261:57–79, 2001b. A preliminary version appeared in M. Li and A. Maruoka, eds., *Algorithmic Learning Theory. Proceedings of the 8th International Conference, ALT 1997*, volume 1316 of *Lecture Notes in Computer Science*, 323-338. Springer, 1997.

V. Vovk and C. Watkins. Universal portfolio selection. In P. Bartlett and Y. Mansour, editors, *Proceedings of the Eleventh Annual Conference on Computational Learning Theory, COLT 1998*, pages 12–23. ACM, 1998.

A. K. Zvonkin and L. A. Levin. The complexity of finite objects and the development of the concepts of information and randomness by means of the theory of algorithms. *Russian Mathematical Surveys*, 26(6):83–124, 1970. Translation of the Russian original in *Uspekhi Matematicheskikh Nauk*, 25(6):85-127, 1970.

Machine Learning Group, Centrum Wiskunde & Informatica, Amsterdam; Faculty of Philosophy, University of Groningen

E-mail address: tom@cwi.nl