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

TOM F. STERKENBURG

Abstract. The diagonalization argument of Putnam (1963) denies the possibility of a universal learning machine. Yet the proposal of Solomonoff (1964) and Levin (1970) promises precisely such a thing. In this paper I discuss how their proposed measure function manages to evade Putnam's diagonalization in one respect, only to fatally fall prey to it in another.

1.

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. Consider a simple language with a monadic predicate G and an ordered infinity of individuals x_i, $i \in \mathbb{N}$, and 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 $\neg Gx_i$. Then a natural condition of adequacy on any proposed 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 $x_0, ..., x_n$.

(Since our actual inductive methods are sure to discern a computable pattern eventually, so should a reconstruction of them.) But for any measure function P that is itself at least effectively computable in a weak sense (so as to qualify, with the Church-Turing thesis, as an explicit method):

(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).

Thus, if the ideal inductive policy fulfills (I) and (II), then it is provably impossible to reconstruct it as a measure function. But maybe such a policy is so idealized as to thwart 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 in effect 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:

Date: May 14, 2016.

(III) For any true computable hypothesis h, if h is ever proposed, then method 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 seemed reluctant to commit himself to the idea of an "inductive machine" (see Carnap, 1950, 192-99). Nevertheless, in his Radio Free Europe address (1963b), Putnam 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" (1963b, 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 diagonalization 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).

2.

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). We seek 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 $x_1, x_2, x_3, \ldots$ 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-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 c-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 function $Q_{\mathbf{f}}$ 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 $Q_{\mathbf{f}}$ of *any true computable hypothesis* will converge to 1, thus fulfilling (I).

3.

How does Solomonoff evade Putnam's diagonalization?

If $Q_{\mathbf{f}}$ 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 $Q_{\mathbf{f}}$ 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 $Q_{\mathbf{f}}$ as a purported "optimum," or *universal* learning machine.

Specifically, we will work in this section towards the precise specification of $Q_{\mathbf{f}}$, 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 $\{0,1\}$. A computable measure on $\{0,1\}^\omega$ is generated in the standard way (according to "Method I" in Rogers, 1970, 9ff; also see Reimann, 2009, 249-256; Nies, 2009, 68-70) from a computable premeasure. A premeasure that generates a probability measure is a function $m : \{0,1\}^* \to [0,1]$ on the finite sequences that satisfies $m(\epsilon) = 1$ for the empty sequence ϵ and $m(x0) + m(x1) = m(x)$ for all $x \in \{0,1\}^*$; the resulting measure μ_m will then satisfy $\mu_m([\![x]\!]) = m(x)$ for every cylinder $[\![x]\!] = \{X \in \{0,1\}^\omega : x \preccurlyeq X\}$, i.e., class of infinite extensions of finite x. I will sometimes be sloppy and simply write "$\mu(x)$" for "$\mu([\![x]\!])$." A premeasure is computable if its values are uniformly computable reals: there is a computable $f : \{0,1\}^* \times \mathbb{N} \to \mathbb{Q}$ such that $|f(x,t) - m(x)| < 2^{-t}$ for all $x \in \{0,1\}^*, t \in \mathbb{N}$. A computable measure is also called a Δ^0_1 measure.

We will see below that the Solomonoff-Levin measure function $Q_{\mathbf{f}}$ has the following property.

(I') For any true Δ^0_1 measure μ, with probability 1, the values $Q_{\mathbf{f}}(x_{t+1} \mid x^t)$ for $x_{t+1} \in \{0,1\}$ converge to the values $\mu(x_{t+1} \mid x^t)$ as t goes to infinity.

This is a generalization of Putnam's condition (I) from "deterministic" computable hypotheses or single infinite computable sequences to computable probability measures on infinite sequences.

We proceed with our discussion of computable measures. The most basic (Δ^0_1) measure on Cantor space is the *uniform* measure λ. It is generated from the premeasure with $m(x) = 2^{-|x|}$ for all x, where $|x|$ denotes x's length. We can obtain other measures as *transformations* of λ by Borel functions $F : \{0,1\}^\omega \to \{0,1\}^\omega$. A transformation of λ by Borel function F, written λ_F, is defined by $\lambda_F(A) =$

$\lambda(F^{-1}(A))$. Every other Borel measure μ on Cantor space can be obtained as a transformation $\lambda_F = \mu$ by some Borel function F.

We are interested in transformations by functions that are effectively computable. To that end we introduce mappings $f : \{0,1\}^* \to \{0,1\}^*$ on *finite* sequences, mappings that have to satisfy a condition of *monotonicity*: if $x \preccurlyeq y$ then also $f(x) \preccurlyeq f(y)$. Consider the function $\Phi_f : X \mapsto \sup_{\preccurlyeq}\{f(x) : x \preccurlyeq X\}$ induced by f. If $\sup_{\preccurlyeq}\{f(x) : x \preccurlyeq X\}$ is an infinite sequence for all infinite X, then Φ_f gives a total function $F : \{0,1\}^\omega \to \{0,1\}^\omega$. If not, then we have to restrict the domain and Φ_f is a partial function on $\{0,1\}^\omega$. Alternatively, we can treat Φ_f as a total function $\{0,1\}^\omega \cup \{0,1\}^* \to \{0,1\}^\omega \cup \{0,1\}^*$ on the collection of infinite *and* finite sequences.

To specify a *computable* monotone mapping, we note that f can be represented by the set $M_f \subseteq \{0,1\}^* \times \{0,1\}^*$ of pairs of sequences (x, y) such that $f(x) \succcurlyeq y$. Then a computable monotone mapping is a monotone mapping f with c.e. M_f. One can visualize a computable monotone mapping as a particular type of Turing machine, one that operates on a steady stream of input symbols, producing an (in)finite output sequence in the process. Originally dubbed an *algorithmic process* (Zvonkin and Levin, 1970, 99), this type of machine is now better known as a *monotone* machine. (Also see Shen, Uspenky, and Vereshchagin 2014, 141-44.)

The transformation λ_f by monotone mapping f is given by the premeasure $m : y \mapsto \lambda(\llbracket\{x : f(x) \succcurlyeq y\}\rrbracket)$, mapping to each sequence y the uniform measure of the input sequences x that lead f to produce it (Zvonkin and Levin, 1970, 100). If computable monotone mapping f produces an infinite sequence with uniform probability 1 (i.e., the class of X with infinite $\Phi_f(X)$ has uniform measure 1), then the transformation λ_f is a premeasure that again generates a Δ^0_1 measure on $\{0,1\}^\omega$ (ibid.). Every other Δ^0_1 measure can be obtained as a λ-transformation $\lambda_f = \mu$ of some such "almost total" computable monotone mapping f.

However, if there is some finite y such that with positive uniform probability machine f stops producing more symbols after y (that is, the class of X with *finite* $\Phi_f(X)$ has positive uniform probability), then $\lambda_f(y)$ is strictly greater than $\lambda_f(y0)+\lambda_f(y1)$. A function λ_f can thus be interpreted as a premeasure generating a measure on the collection of infinite *and* finite sequences (ibid., 102). (Alternatively, one can interpret such a function as a "semimeasure" on $\{0,1\}^\omega$ (Levin and V'yugin, 1977, 360), a "defective" probability measure. See Li and Vitányi (2008, 264; 331-32).)

Levin calls the class of (measures generated from the) transformations λ_f by all monotone machines f the class of *semi-computable* measures on $\{0,1\}^\omega \cup \{0,1\}^*$. This is because these transformations are precisely the functions $m : \{0,1\}^* \to [0,1]$ with $m(x) \geq m(x0) + m(x1)$ for all 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 (also see 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 $x \in \{0,1\}^*$ we have $f(x,t) \leq f(x,t+1)$ for all $t \in \mathbb{N}$ and $\lim_{t\to\infty} f(x,t) = m(x)$. Equivalently, the left-cut $\{(q,x) \in \mathbb{Q} \times \{0,1\}^* : q < m(x)\}$ is c.e. A semi-computable measure is also called a Σ^0_1 measure.

It is instructive to note the analogy between, on the one hand, the expansion from the Δ^0_1 to the Σ^0_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 that the class of t.c. functions is diagonalizable, and that this is overcome by enlarging the class to the p.c. functions. In other words, the class of all t.c. functions is not effectively enumerable; the class of p.c. functions is. Likewise, the class of Δ_1^0 measures is not effectively enumerable; the class of Σ_1^0 measures is. This analogy between Σ_1^0 measures and p.c. functions is indeed an equivalence in the sense that an effective enumeration of all Σ_1^0 measures is naturally obtained from an effective enumeration of all p.c. functions (cf. Li and Vitányi 2008, 261; 267).

The effective enumerability of the Σ_1^0 measures is crucial, because it allows for the construction of *universal* Σ_1^0 measures (Zvonkin and Levin, 1970, 103-04). Informally, such a 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_1^0 measure $\boldsymbol{\mu}$ is such that it *dominates* every other Σ_1^0 measure: for every $\mu_i \in \Sigma_1^0$ there is a constant $c_i \in \mathbb{N}$ such that for all $x \in \{0,1\}^*$ it holds that $\boldsymbol{\mu}(x) \geq \mu_i(x)/c_i$. "This fact is one of the reasons for introducing the concept of semi-computable measure" (ibid.) — we may take it as the main reason. In fact, the expansion to Σ_1^0 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_1^0 *randomness tests*: a sequence X is random if it passes a universal such test. Vovk (2001b), in defining his notion of *predictive complexity*, employed the class of Σ_1^0 *loss processes*: the predictive complexity of X 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_1^0 elements is then "a very natural solution to the problem of non-existence of a smallest computable loss process" (ibid.).

An example of a universal Σ_1^0 measure is easily given. Since the computable monotone mappings are also effectively enumerable, we can likewise specify *universal* such mappings. Let $\{z_i\}_{i\in\mathbb{N}}$ some computable prefix-free (i.e., $z_i \preccurlyeq z_j \Rightarrow i = j$) encoding of all computable monotone mappings f_i. The corresponding universal monotone mapping $\mathbf{f}$ is defined by $(z_i x, y) \in M_{\mathbf{f}} :\Leftrightarrow (x, y) \in M_{f_i}$. It is straightforward to show that the transformation $\lambda_{\mathbf{f}}$ of λ by universal $\mathbf{f}$ gives a universal Σ_1^0 measure.

We have finally arrived at the definition of the Solomonoff-Levin measure function. The measure $Q_{\mathbf{f}}$ is precisely the transformation of λ by universal monotone mapping $\mathbf{f}$.

Definition 1. $Q_{\mathbf{f}} := \lambda_{\mathbf{f}}$.

So there are in fact infinitely many such measures $Q_{\mathbf{f}}$, one for each choice of universal monotone mapping $\mathbf{f}$. Each is a universal Σ_1^0 measure. It is this property that is exploited in the adequacy result.

Proposition 2. $Q_{\mathbf{f}}$ fulfills (I').

Proof. Let $\mu \in \Delta_1^0$. The fact that $Q_{\mathbf{f}}$ dominates μ entails that μ is absolutely continuous with respect to $Q_{\mathbf{f}}$ (i.e., $\mu(A) > 0$ implies $Q_{\mathbf{f}}(A)$ for all A in the σ-algebra $\mathfrak{F}$), which by the classical result of Blackwell and Dubins (1962) entails that μ-a.s. the variational distance $\sup_{A\in\mathfrak{F}} |\mu(A \mid x^t) - Q_{\mathbf{f}}(A \mid x^t)| \to 0$ as $t \to \infty$ (see Huttegger, 2015, 617-18), so in particular (I'). □

4.

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_1^0 measures is to evade diagonalization — to obtain universal elements. The measure $Q_{\mathbf{f}}$ is a universal element; as such, it tracks every Δ_1^0 measure in the sense of (I'). The downside is that, as a universal Σ_1^0 element, $Q_{\mathbf{f}}$ is itself no longer Δ_1^0 (or the class of Δ_1^0 would already have universal elements). Worse, as we will see in more detail in the penultimate section below, $Q_{\mathbf{f}}$ fails to satisfy the even weaker effectiveness condition (II).

The force of Putnam's diagonalization proof is that no measure function can satisfy both (I) and (II), and $Q_{\mathbf{f}}$ is no exception. The Solomonoff-Levin measure function is powerful enough to avoid diagonalization and fulfill (I), but the price to pay is that $Q_{\mathbf{f}}$ might be said to be *too* powerful: it is no longer computable in the sense of (II). Does this invalidate $Q_{\mathbf{f}}$ as an inductive method — let alone a universal one?

One reply is that we cannot hold this against $Q_{\mathbf{f}}$ just like that, 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, consider the question of $Q_{\mathbf{f}}$'s adequacy afresh.

In order to assess Putnam's statement of the HD-method's adequacy, we have to consider the following two conditions, the generalizations of (I) and (II) to methods that are not necessarily measure functions (the original conditions are retrieved by inserting "instance confirmation greater than 0.5" for "accepted", cf. Putnam, 1963a, 771):

(I*) Any true computable hypothesis h is (and forever remains) accepted after sufficiently many confirming individuals $x_0, ..., x_n$.

(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 h is accepted after time $n + m$.

Again, it is important for Putnam's case against Carnap that the conditions (I*) and (II*) 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) and (II) are mutually exclusive — but other methods can: and the hypothetico-deductive (HD) method that Putnam describes is to be the case in point. On a closer look, however, this assertion will appear a bit murky.

Recall that Putnam's HD method depends on the hypotheses that are actually proposed in the course of time. The HD method fulfills (III), 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 (in the current context, computable infinite sequences) 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 (II*), 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; also see the simple derivation of this fact in 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? Moreover, "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., $\xi_w^{\mathcal{H}}(x) := \sum_{h\in\mathcal{H}} w(h)h(x)$.

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 come to accept (and forever stick to) the hypothesis if it is true (in the sense of (I)). Just like the HD architecture is guaranteed to accept and stick to every true deterministic hypothesis, *whenever* it is included in the hypothesis stream, so the classical Bayesian architecture is guaranteed to accept 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, whenever it is in $\mathcal{H}$, *with (true) probability 1*. Or, to put it more succinctly, a Bayes$^{\mathcal{H}}$ method will *almost surely converge on* the true hypothesis 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 $x \in \mathbb{B}^*$ that $\xi_w^{\mathcal{H}}(x) \geq w(h)h(x)$.

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.

5.

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 $Q_{\mathbf{f}}$ out of hand simply because it does not satisfy the special cases (I) and (II) — that it cannot do the impossible. Instead, let us conclude our survey with a fresh look at the question: could $Q_{\mathbf{f}}$ 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 $Q_{\mathbf{f}}$ be able to accept every reasonable (reasonably effective) hypothesis, if it is true? Second, in the spirit of (II), is $Q_{\mathbf{f}}$ itself still a reasonable (reasonably effective) method?

To start with the first. The best vantage point to address this question is to view $Q_{\mathbf{f}}$ as an instantiation of the classical Bayesian architecture that we saw in the previous section. It turns out that the measure functions $Q_{\mathbf{f}}$ are the classical Bayesian methods that employ the class of all Σ^0_1 hypotheses (also see Sterkenburg, 2016). To be exact, the measure functions $Q_{\mathbf{f}}$ are precisely the Bayes$^{\mathcal{H}_{\Sigma^0_1}}$ methods $\xi^{\Sigma^0_1}_w$ with semicomputable prior w over the hypothesis class $\mathcal{H}_{\Sigma^0_1}$ of all Σ^0_1 measures. (In particular, the choice of universal transformation $\mathbf{f}$ corresponds to the choice of semicomputable prior w over $\mathcal{H}_{\Sigma^0_1}$.) By Bayesian consistency, it follows that $Q_{\mathbf{f}}$ will almost surely converge on any true Σ^0_1 hypothesis. (This is again, in essence, Proposition 2 above, though I only stated it for Δ^0_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 first question: is the hypothesis class $\mathcal{H}_{\Sigma^0_1}$ sufficiently wide, sufficiently general?

Before we turn to an answer, we connect this question to an important alternative perspective on $Q_{\mathbf{f}}$. This is the interpretation of $Q_{\mathbf{f}}$ as an "a priori" distribution over the symbol sequences. Measure $Q_{\mathbf{f}}$ "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 $[Q_{\mathbf{f}}]$". This is an illustration of how the question of the generality of $\mathcal{H}_{\Sigma^0_1}$ — the class of candidate measures that may be assumed to generate the data — is related to the question of the adequacy of $Q_{\mathbf{f}}$ 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 $Q_{\mathbf{f}}$ 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 $Q_{\mathbf{f}}$ still leaves open the choice of $\mathbf{f}$ — from the classical Bayesian perspective, the choice of semicomputable prior w over $\mathcal{H}_{\Sigma_1^0}$. (See Sterkenburg (2016) for more on this.) Bracketing this issue here, we still face a more fundamental problem: the problem of justifying the stipulated constraints on the measure functions. From the classical Bayesian perspective, this is 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 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 $\mathcal{H}_{\Sigma_1^0}$, then, escape Hume's argument — is $\mathcal{H}_{\Sigma_1^0}$ fully general? Naturally, it is not. As a restriction on what hypotheses could be *true*, a *metaphysical* assumption on the world, not only would the restriction to any specific level of effective computability (Δ_1^0, Σ_1^0, ...) look arbitrary: the assumption of effective computability itself is a stipulation that wants motivation.

6.

There is, however, an alternative interpretation still. This interpretation is to take the elements of the class $\mathcal{H}_{\Sigma_1^0}$, not as hypotheses about the origin of the data, but as *competing inductive methods* (cf. Sterkenburg, 2016).

This interpretation is actually more in line with Putnam's demand that the ideal inductive policy or the universal 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 being developed in great depth as a vibrant 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 Vovk, 2001a; Cesa-Bianchi and Lugosi, 2006).)

Let us see what we get when we thus reinterpret the Σ_1^0 measures as *all possible inductive methods*. As a start, Proposition 2 could be reinterpreted as a fully general *merging-of-opinions* result (see Huttegger, 2015): every inductive method anticipates with certainty that $Q_{\mathbf{f}}$'s confirmation values converge to its own. Moreover, it is easy to derive the following more "absolute" fact. For any inductive

method 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 $Q_{\mathbf{f}}$ relative to this method ν, on *any* symbol sequence (see the Appendix for details). Thus, if we take the Σ^0_1 measures as all possible inductive methods, then $Q_{\mathbf{f}}$ is a universal inductive method in the following powerful sense: *it is an inductive method that compared to any other inductive method 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), $Q_{\mathbf{f}}$ is *optimal* in the sense that it is guaranteed to converge on the true hypothesis *if any method does*. The inductive method $Q_{\mathbf{f}}$ is *vindicated* in the sense of Reichenbach (see Salmon, 1991).

In this interpretation, $Q_{\mathbf{f}}$ *is* a universal learning machine — defying the lesson that has generally been taken from Putnam's proof that there can be no such thing (cf. Dawid, 1985, 341). 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^0_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 inductive methods with the Σ^0_1 measures?

Most importantly, is the class of Σ^0_1 measures not *too* wide — does a Σ^0_1 measure that fails to be Δ^0_1 still constitute a proper method? As a special case, we have returned to the second question we started the previous section with: does $Q_{\mathbf{f}}$ 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^0_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^0_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^0_1 measure is still connected to some explicit method. (Cf. Martin-Löf on his choice of Σ^0_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" (1969, 268).)

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 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. Namely, in particular, the *conditional* Solomonoff-Levin function $Q_{\mathbf{f}}(\cdot \mid \cdot)$, given by $Q_{\mathbf{f}}(\tau \mid \sigma) = Q_{\mathbf{f}}(\sigma\tau)/Q_{\mathbf{f}}(\sigma)$, is no longer Σ_1^0.

This is essentially implied by the original diagonalization argument: one can verify that $Q_{\mathbf{f}}(\cdot \mid \cdot) \in \Sigma_1^0$ would mean that $Q_{\mathbf{f}}$ satisfies (II). For completeness, the following proof recounts the details of the diagonalization. (A different proof has been given by Leike and Hutter, 2015, 370-71.)

Proposition 3. $Q_{\mathbf{f}}(\cdot \mid \cdot) \notin \Sigma_1^0$.

Proof. Suppose towards a contradiction that $Q_{\mathbf{f}}(\cdot \mid \cdot)$ is Σ_1^0, so that (II) holds for $Q_{\mathbf{f}}$. We can now construct a computable infinite sequence x^ω as follows. Start calculating $Q_{\mathbf{f}}(0 \mid 0^n)$ from below in dovetailing fashion for increasing $n \in \mathbb{N}$, until an n_0 such that $Q_{\mathbf{f}}(0 \mid 0^{n_0}) > 0.5$ is found (since $Q_{\mathbf{f}}$ 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 $x^\omega := 0^{n_0}10^{n_1}10^{n_2}1\ldots$. Sequence x^ω is computable, but by construction the instance confirmation of x^ω will never remain above 0.5, contradicting (I). □

Now we could argue that $Q_{\mathbf{f}}(\cdot \mid \cdot)$ is still Δ_2^0 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_2^0 confirmation function (i.e., a Δ_2^0 conditional measure) still counts as a possible method, then we should identify the possible inductive methods with the class of Δ_2^0 confirmation functions (rather than the original class of confirmation functions with underlying Σ_1^0 measure functions). That means that the sought-for optimality would have to be relative to *this* class. But $Q_{\mathbf{f}}(\cdot \mid \cdot)$ is not optimal among the Δ_2^0 confirmation functions — *no* Δ_2^0 confirmation function is. This is because the class of Δ_2^0 *measure* functions, that in this case precisely induces the class of Δ_2^0 *confirmation* functions, *is* diagonalizable — just like the class of Δ_1^0 measure functions is. Nor can we take a step back and settle for the class of Σ_1^0 confirmation functions: once again one can show by the argument of Proposition 3 that there cannot exist universal elements in the class of measure functions that induce the Σ_1^0 confirmation functions. This easily relativizes: our strategy for optimality cannot work on any level in the arithmetical hierarchy.

7.

Thus we conclude our story on an unhappy note. We have discussed how Putnam's diagonalization argument shows that no method whatsoever — not just 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. On the principle that one should not aim for the impossible, we allowed ourselves to consider as candidate universal learning machines measure functions that satisfy the first condition but that fall short of the second; specifically, we considered the Solomonoff-Levin measure function. The overarching strategy we identified to evade Putnam's argument is to locate a sufficiently large class of measure functions that is immune to diagonalization, hence contains universal elements. If one could reasonably identify this class of measure functions with all possible inductive methods,

then the universal elements would be vindicated as optimal inductive methods, as universal learning machines: they constitute methods that are in a strong sense at least as good as any other method. In particular, we saw that the Solomonoff-Levin measure functions were constructed as universal elements among the Σ^0_1 measures — and so, our hope ran, they could qualify as such optimal methods. Unfortunately, we found a fatal flaw in this strategy: inductive methods 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 inductive methods. Putnam's diagonalization argument is not so easily disposed of.

Appendix

Proposition 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*

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

Then, for every $\mu \in \Delta^0_1$, the expected infinite sum of divergences between $Q_{\mathbf{f}}$ and μ

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

is bounded by a constant.

To see how (I') follows from this constant bound, suppose that $Q_{\mathbf{f}}$ does not satisfy (I'): there is a $\mu \in \Delta^0_1$ such that with probability $\epsilon > 0$ there is a $\delta > 0$ such that $\left| \mu(x_{t+1} \mid x^t) - Q_{\mathbf{f}}(x_{t+1} \mid x^t) \right| > \delta$ infinitely often. But that means that with positive probability the infinite sum of squared Hellinger distances is infinite, and the expectation (1) cannot be bounded by a constant.

The proof of the constant bound on (1) 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 $-\ln P(x^t)$ expresses the *logarithmic loss* of P on sequence x^t, a standard measure of prediction error; the difference $-\ln P_2(x^t) - (-\ln P_1(x^t)) = \ln \frac{P_1(x^t)}{P_2(x^t)}$ expresses the surplus prediction error or *regret* of P_2 relative to P_1 on sequence x^t. 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 (1) is bounded by

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

Now by the universality of $Q_{\mathbf{f}}$ in the class of Σ^0_1 measures we know that $Q_{\mathbf{f}}$ dominates μ: for every finite x there is a constant c such that $Q_{\mathbf{f}}(x) \geq \mu(x)/c$. Indeed we can identify c with $1/w(\mu)$, where w is the prior over hypothesis class $\mathcal{H}_{\Sigma^0_1}$ in the classical Bayesian representation $\xi^{\Sigma^0_1}_w$ of $Q_{\mathbf{f}}$. This fact allows us to derive that *for every sequence x^s of any*

length s

$$\begin{aligned}\sum_{t=0}^{s-1} \ln \frac{\mu(x_{t+1} \mid x^t)}{Q_{\mathbf{f}}(x_{t+1} \mid x^t)} &= \ln \prod_{t=0}^{s-1} \frac{\mu(x_{t+1} \mid x^t)}{Q_{\mathbf{f}}(x_{t+1} \mid x^t)} \\ &= \ln \frac{\mu(x^s)}{Q_{\mathbf{f}}(x^s)} \\ &\leq -\ln w(\mu). \qquad (5)\end{aligned}$$

This concludes the proof that (1) 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 (1).

Proposition 2 was in the main text only stated for measures μ in Δ_1^0: 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_1^0 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 $\nu \in \Sigma_1^0$ be represented by a measure ν' over $\{0,1,\varnothing\}^\omega$, with $\varnothing$ a "stopping symbol": we have $\nu'(\sigma 0) + \nu'(\sigma 1) + \nu'(\sigma\varnothing) = \nu'(\sigma)$ and we stipulate $\nu'(\sigma) = \nu(\sigma)$ and $\nu'(\sigma\varnothing\varnothing) = \nu'(\sigma\varnothing)$ for all $\sigma \in \{0,1\}^*$. Then for all $\nu \in \Sigma_1^0$ we have that $Q'_{\mathbf{f}}$ dominates ν', hence $\nu' \ll Q'_{\mathbf{f}}$ 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 $\nu \in \Sigma_1^0$, the sum of surplus prediction errors (regrets) of $Q_{\mathbf{f}}$ relative to ν will *always* (for any sequence x^s of any length s) be bounded by a constant:

$$\sum_{t=0}^{s-1} \big(-\ln Q_{\mathbf{f}}(x_{t+1} \mid x^t) - \big(-\ln \nu(x_{t+1} \mid x^t)\big)\big) \leq -\ln w(\nu).$$

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, Illinois, 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, California, 1962.

R. Carnap. Replies and basic expositions: Hilary Putnam on 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 983–989. Open Court, La Salle, Illinois, 1963a.

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

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. *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.

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.

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.

L. A. Levin and V. V. V'yugin. Invariant properties of information bulks. In G. Goos and J. Hartmanis, editors, *Proceedings of the Sixth Symposium on the Mathematical Foundations of Computer Science*, volume 53 of *Lecture Notes in Computer Science*, pages 359–364, Berlin, 1977. Springer.

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. 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, La Salle, 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 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.

C. A. Rogers. *Hausdorff measures*. Cambridge University Press, Cambridge, 1970.

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. Uspenky, and N. K. Vereshchagin. *Kolmogorov Complexity and Algorithmic Randomness*. 2014. Translation 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. 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. Forthcoming in *Philosophy of Science*, 83(3), 2016.

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.

Algorithms & Complexity Group, Centrum Wiskunde & Informatica, Amsterdam; Faculty of Philosophy, University of Groningen

E-mail address: tom@cwi.nl