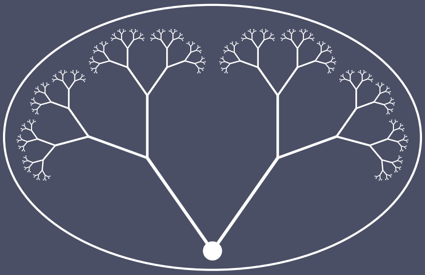
UNIVERSAL PREDICTION

TOM F. STERKENBURG

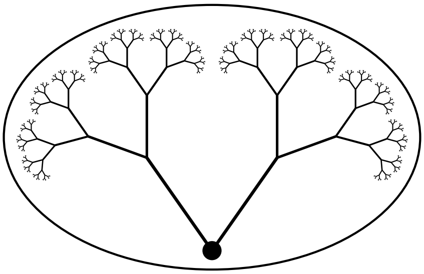


UNIVERSAL PREDICTION

A PHILOSOPHICAL INVESTIGATION

UNIVERSELE VOORSPELLING

EEN WIJSGERIGE ONDERZOEKING



TOM F. STERKENBURG

ISBN 978-94-034-0319-9 (printed version)
ISBN 978-94-034-0318-2 (electronic version)

rijksuniversiteit
groningen

Universal Prediction

A Philosophical Investigation

Proefschrift

ter verkrijging van de graad van doctor aan de
Rijksuniversiteit Groningen
op gezag van de
rector magnificus prof. dr. E. Sterken
en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op

donderdag 18 januari 2018 om 14.30 uur

door

Tom Florian Sterkenburg

geboren op 18 april 1986
te Purmerend

Promotores
Prof. dr. J.-W. Romeijn
Prof. dr. P.D. Grünwald

Beoordelingscommissie
Prof. dr. dr. H. Leitgeb
Prof. dr. A.J.M. Peijnenburg
Prof. dr. S.L. Zabell

Contents

Dankbetuiging / Acknowledgements

I have delegated specific acknowledgements to the *endnotes of thanks*—see if your name is there on page 208!

Ik wil mijn begeleiders Peter Grünwald en Jan-Willem Romeijn bedanken voor de vrijheid die ze me gaven om mijn eigen weg te vinden, voor de hulp die ze me gaven wanneer ik daarnaar verlangde, en voor het gevoel dat ze me gaven dat dit een waardevolle onderneming was. Ik durf niet te zeggen wat het belangrijkst is geweest.

I also want to thank the members of the assessment committee: Hannes Leitgeb, Jeanne Peijnenburg, and Sandy Zabell.

I was lucky to have two academic homes during this project, and the benefits of two circles of colleagues. Many thanks to the mathematicians/computer scientists in Amsterdam[1] and the philosophers in Groningen.[2] Part of the final writing I did while I was visiting the Center for Formal Epistemology at CMU, Pittsburgh.[3]

I made much use of the amazing library of the CWI. It convinced me of the importance, the more so in a hostile digital age,[4] of a physical library, a collective memory one can actually walk around in.

*

Basic notions and notation

This is an overview of notions and notations that will be used throughout the thesis. It serves as a reference: all notions will be properly introduced and explained in the main text.

Binary sequences. In this thesis I only consider sequences that are built from an alphabet of just two symbols, '0' and '1.' I use the variables 'x,' 'y', 'z' to refer to individual symbols; the variables '$\boldsymbol{x}$,' '$\boldsymbol{y}$,' '$\boldsymbol{z}$' denote sequences of symbols. The empty sequence is $\varnothing$. For two sequences $\boldsymbol{x}$ and $\boldsymbol{y}$, their *concatenation* is simply written '$\boldsymbol{xy}$.' I write '$\boldsymbol{x} \preccurlyeq \boldsymbol{y}$' if $\boldsymbol{x}$ is an *initial segment* or *prefix* of $\boldsymbol{y}$ (so there is an $\boldsymbol{z}$ such that $\boldsymbol{xz} = \boldsymbol{y}$; if $\boldsymbol{z} \neq \varnothing$ then I write '$\boldsymbol{x} \prec \boldsymbol{y}$'). I often write '$\boldsymbol{x}^t$' to indicate that the sequence has length t; sometimes it conveys the more specific fact that $\boldsymbol{x}^t$ is the prefix of length t of the (longer) sequence $\boldsymbol{x}$, also written '$\boldsymbol{x} \upharpoonright_t$.' Occasionally I refer to the length of $\boldsymbol{x}$ by '$|\boldsymbol{x}|$.' The sequence $\boldsymbol{x}^-$ is the initial segment of $\boldsymbol{x}$ of length $|\boldsymbol{x}| - 1$. An infinite sequence is denoted by adding the superscript 'ω' to a variable name, like so: '$\boldsymbol{x}^\omega$,' '$\boldsymbol{y}^\omega$,' '$\boldsymbol{z}^\omega$.' The i-th symbol of $\boldsymbol{x}$ is $\boldsymbol{x}(i)$. Sequences $\boldsymbol{x}$ and $\boldsymbol{y}$ are *comparable*, written '$\boldsymbol{x} \sim \boldsymbol{y}$,' if $\boldsymbol{x} \preccurlyeq \boldsymbol{y}$ or $\boldsymbol{y} \prec \boldsymbol{x}$; if $\boldsymbol{x}$ and $\boldsymbol{y}$ are not comparable this is written '$\boldsymbol{x} \mid \boldsymbol{y}$.' The *lexicographical ordering* arranges all finite sequences in the natural increasing-length ordering $\varnothing, 0, 1, 00, 01, 10, 11, 000, \ldots$; I write '$\boldsymbol{x} <_L \boldsymbol{y}$' if $\boldsymbol{x}$ precedes $\boldsymbol{y}$ in this ordering. The number of occurrences of symbol x in sequence $\boldsymbol{x}$ is denoted '$\#_x \boldsymbol{x}$'.

Let $\mathbb{B} := \{0, 1\}$ denote the set of symbols. Then $\mathbb{B}^t$ is the set of all symbol sequences of length t (and likewise we have $\mathbb{B}^{\leq t}$ and $\mathbb{B}^{<t}$). $\mathbb{B}^* = \cup_{t \in \mathbb{N}} \mathbb{B}^t$ is the set of all finite sequences; $\mathbb{B}^\omega$ the class of all infinite sequences. A subset $A \subseteq \mathbb{B}^*$ of finite sequences is *prefix-free* if $\boldsymbol{x} \mid \boldsymbol{y}$ for every two different $\boldsymbol{x}, \boldsymbol{y} \in A$. For set A of finite sequences, its *bottom* $\lfloor A \rfloor := \{\boldsymbol{x} \in A : \forall \boldsymbol{y} \in A.\ \boldsymbol{y} \preccurlyeq \boldsymbol{x} \Rightarrow \boldsymbol{y} = \boldsymbol{x}\}$ is the prefix-free subset of *minimal* sequences in A that have no strict prefixes in A.

For given finite sequence $\boldsymbol{x}$, the class $[\![\boldsymbol{x}]\!] := \{\boldsymbol{x}^\omega \in \mathbb{B}^\omega : \boldsymbol{x}^\omega \succcurlyeq \boldsymbol{x}\}$ is the class of infinite extensions of $\boldsymbol{x}$. Likewise, for set $A \subseteq \mathbb{B}^*$ of finite sequences, let $[\![A]\!] := \{\boldsymbol{x}^\omega \in \mathbb{B}^\omega : \boldsymbol{x} \in A \ \&\ \boldsymbol{x}^\omega \succcurlyeq \boldsymbol{x}\}$.

Prediction methods. A prediction method (alternatively, *prediction strategy/rule/system*, or simply *predictor*) is a function

$$\mathsf{p} : \mathbb{B}^* \to \mathcal{P}$$

from the finite data sequences to *predictions*, distributions over $\mathbb{B}$. I often specify a prediction $p \in \mathcal{P}$ by $p = (a_0, a_1)$, meaning $p(0) = a_0$ and $p(1) = a_1$. I also use the shorthand

$$\mathsf{p}(x, \boldsymbol{x}) := \mathsf{p}(\boldsymbol{x})(x).$$

Probability measures. Strictly formally, I consider measures μ on $\mathbb{B}^\omega$, also known as the *Cantor space*. However, to keep things simple where I can, I usually treat a measure as a function $\mu : \mathbb{B}^* \to [0,1]$ that assigns probability values to the *finite sequences*, and that satisfies

$$\mu(\varnothing) = 1;$$
$$\mu(\boldsymbol{x}0) + \mu(\boldsymbol{x}1) = \mu(\boldsymbol{x}) \text{ for all } \boldsymbol{x} \in \mathbb{B}^*.$$

(Such a function is called a “probabilistic source” in Grünwald, 2007, 53. Strictly formally, again, it is the pre-measure m that generates a measure; this is described in 2.1.1.)

I sometimes denote by ‘$\mu^{\boldsymbol{x}}$’ the measure μ *conditional* on $\boldsymbol{x}$, i.e., the measure $\mu(\cdot \mid \boldsymbol{x})$. See 2.1.1.3 for details on the definition of conditional measures in sequential prediction: notably, there is the convention of writing ‘$\mu(\boldsymbol{y} \mid \boldsymbol{x})$’ for $\mu(\boldsymbol{xy} \mid \boldsymbol{x})$. I denote by ‘$\mu^t$’ the *distribution* over $\mathbb{B}^t$ that is given by $\mu^t(\boldsymbol{x}^t) = \mu(\boldsymbol{x}^t)$, and likewise I denote by $\mu^1(\cdot \mid \boldsymbol{x})$ the one-step conditional measure that is a distribution over $\mathbb{B}$.

Order notation. I regularly use the standard ‘big-O’ notation for functions $f, g : \mathbb{N} \to \mathbb{R}$, where $f(t) = O(g(t))$, ‘f is big O of g,’ means that there is a constant $c > 0$ such that $|f(n)| \leq c\,|g(n)|$ for all $n \in \mathbb{N}$. In particular, I often use $f(n) = O(1)$ to signify that there is single constant c such that $f(n) < c$ for all n.

Somewhat less standard is the notation ‘$f \leq^+ g$’ to express that f *additively minorizes* g, meaning that $f(n) = g(n) + O(1)$, i.e., there is a constant c such that for all n, $f(n) \leq g(n) + c$. (Equivalently, g additively *majorizes* or *dominates* f.) Likewise, ‘$f \leq^\times g$’ expresses that f *multiplicatively* minorizes g: there is a constant c such that $f(n) \leq c \cdot g(n)$. Moreover, ‘$f =^+ g$’ and ‘$f =^\times g$’ express that f and g additively and multiplicatively minorize (equivalently, majorize) each other, respectively.

Computability. I sketch the model of a *Turing machine* in I.4; for a more detailed specification see for instance Soare (2016, 7ff).

Computable functions. A Turing machine specifies a (possibly partial) function $T : \mathbb{B}^* \to \mathbb{B}^*$. In fact, since we can effectively map $\mathbb{B}^*$ onto any desired class of finite objects (e.g., the bits $\mathbb{B}$, the natural numbers $\mathbb{N}$, the rational numbers $\mathbb{Q}$, the finite sets of finite sequences $\mathcal{P}^{\text{fin}}(\mathbb{B}^*)$, ...), we can say more generally that for any given classes $\mathbb{A}, \mathbb{A}'$ of finite objects, a Turing machine defines a (possibly partial) function $T : \mathbb{A} \to \mathbb{A}'$. (As we can have $\mathbb{A}$ be the Cartesian product of a finite number n of finite sets of objects, we can also have a Turing machine define an n-place function.) Now a function $\varphi : \mathbb{A} \to \mathbb{A}'$ is *computable* if there is a Turing machine that specifies it. The Turing machines thus specify the *partial computable* (*p.c.*) functions, that I denote by the letters 'φ', 'ψ,' .... If φ is not defined on input a (the Turing machine that specifies it does not halt on input a), we say that φ *diverges* on a and write '$\varphi(a) \uparrow$.' Likewise, if φ *converges* on a with output a' we can write '$\varphi(a) =\downarrow a'$.' I write '$\varphi(a) \simeq \psi(a)$' to mean that $\varphi = \psi$, i.e., for all $a \in \mathbb{A}$, either $\varphi(n) \downarrow= \psi(a)$ or $\varphi(a) \uparrow$ and $\psi(a) \uparrow$. The *total computable* (*t.c.*) functions, denoted by the letters 'f,' 'g,' ..., are computable functions that are everywhere defined.

Acceptable enumerations. One can indeed define an effective list all Turing machines $\{T_e\}_{e\in\mathbb{N}}$ by coding them onto the integers: this induces an *acceptable enumeration* $\{\varphi_e\}_{e\in\mathbb{N}}$ of all p.c. functions. We can now also define a *universal Turing machine* that takes (the code for) an index of a machine and another input, and then reconstructs and runs this machine on this input. Thus the p.c. functions are *uniformly* computable: there is a single p.c. function $\mathring{\varphi}$ (a *universal* p.c. function, specified by a universal Turing machine) such that $\mathring{\varphi}(e, n) \simeq \varphi_e(n)$.

Sets and sequences. A *set* $A \subseteq \mathbb{A}$ of finite objects is *computable* if there exists a computable *characteristic function* $\chi_A : \mathbb{A} \to \mathbb{B}$ such that $\chi_A(a) = 1$ iff $a \in A$. Similarly, an *infinite sequence* $\boldsymbol{x}^\omega \in \mathbb{B}^\omega$ is *computable* if there exists a computable characteristic function $\chi_{\boldsymbol{x}^\omega} : \mathbb{N} \to \mathbb{B}$ that returns the correct symbol for each given position, $\chi_{\boldsymbol{x}^\omega}(n) = \boldsymbol{x}^\omega(n)$. Set $A \subseteq \mathbb{A}$ is *computably enumerable* (*c.e.*) if there exists a computable procedure to enumerate its elements, or equivalently: if it is the domain of a p.c. function.

Real-valued functions. A *real number* r is *computable* if we can computably approximate it to any desired accuracy: there is a computable function $f : \mathbb{N} \to \mathbb{Q}$ such that $|r - f(s)| < 2^{-s}$. Equivalently, the set $\{q : q < r\}$ of *left-cuts* is computable. A *function* $f : \mathbb{A} \to \mathbb{R}$ on the reals is *computable* if its values are uniformly computable: there is a two-place computable function $g : \mathbb{N} \times \mathbb{N} \to \mathbb{Q}$ such that $|f(a) - g(a, s)| < 2^{-s}$. Equivalently, the set $\{(q, a) : q < f(a)\}$ is computable. (See Downey and Hirschfeldt, 2010, 197ff.)

Predictors and measures. The notion of a *computable prediction method* is now defined: it is a prediction method p such that the set $\{(q, x, \boldsymbol{x}) : q < \mathsf{p}(x, \boldsymbol{x})\}$ is computable. The notion of a *computable pre-measure* is likewise defined. A *computable measure* is a measure μ_m that is induced from a computable pre-measure m.

*

Introduction

Mathematical philosophy. Philosophy can deal with contentious topics. To some, the discipline of philosophy itself is a contentious topic. So it can happen that the author of a textbook on the otherwise rather dry subject of measure theoretic probability spices up his work with stabs at practitioners of philosophy of probability, including the lament that

> Since philosophers are pompous where we are precise, they are thought to think deeply . . . (Williams, 1991, 25)

Whatever its further merits, this declaration did inspire me towards an informal characterization of the field this thesis is in. This is the field of *mathematical philosophy*: the treatment of philosophical—pompous?—questions with precise, mathematical, means.

The question. Here is a pompous question: can there be such a thing as a *universal prediction method*?

* * *

Universal prediction. This thesis is concerned with the possibility of universal methods of prediction. From the outset I restrict attention to the simple abstract setting of *sequential prediction* with a binary alphabet. In this setting one makes predictions on a stream of data consisting of instances of just two possible symbols, say 0 and 1. More precisely, in each successive trial one of two possible symbols is revealed; and a prediction method must give at each trial—and only based on the sequence of symbols revealed so far—a (probabilistic) prediction which symbol will appear next.

A *universal* prediction method is, to a first approximation, a method that performs well in *all* cases. This can be construed as the requirement: whatever the actual or *true* data-generating mechanism, the method performs not much worse than one could if one knew the mechanism at work. I call this universal *reliability*. Alternatively, this can be construed as universal *optimality*: whatever the data stream, the method performs not much worse than any other possible prediction method. This is all still quite informal: a perfectly precise characterization of the notion of universality of a prediction method, including what it should mean for a predictor to perform well, is one of the things that I will develop later in this thesis.

The focus in this thesis lies on a proposed definition of a universal prediction method that goes back to Solomonoff (1964). One component that stands out in Solomonoff's proposal is the relation that is forged between universality and effective *computability*. Another main component is the relation that is suggested with a preference for *simplicity*. While, however, the philosophical import of Solomonoff's proposal has repeatedly been emphasized by authors in theoretical computer science, attention in the philosophical literature has so far been largely restricted to the occasional mention in overview works. The main aim of this thesis is to position Solomonoff's proposal in a broader philosophical context, and thereby to address the main question on the possibility of a universal prediction method.

The starting point in this thesis is the connection of Solomonoff's proposal to Carnap's program of inductive logic. More specifically, the thesis sets off from an influential argument of Putnam against Carnap's program, a mathematical proof that is generally understood to demonstrate the impossibility of a universal prediction method.

I will say some more about this starting point below, after which I outline the main themes, the contributions, and the structure of this thesis. But first, to start off on a truly general footing, I give a little sketch of the history of universal prediction.

* * *

Some broad strokes of history. Leibniz famously imagined how all scientific disputes would be solved in a purely mechanical way. If only we had a universal calculus, conjoined with a "universal characteristic" to represent any scientific proposition, then we could establish the truth of any such proposition by simply saying: *calculemus!* Leibniz's proposal of this universal symbolic language and idealized *calculus ratiocinator*—as well as more down-to-earth calculating devices, including his actual construction of a 'stepped reckoner'—are an early articulation of ideas that were finally gaining momentum in the 19th century, and that evolved into modern symbolic logic and computer science in the 20th. Babbage gave a design for a general-purpose computer in 1837; Boole (1847, 1854) and others developed a purely syntactic or symbolic logic in the image of algebra. Frege (1879) significantly extended the latter work to what is in effect the language of *first-order logic*, and initiated the *logicist* ideal of reducing all of mathematics to pure logic. Others set out to formalize branches of mathematics in axiomatic theories, some motivated by logicism and some by Hilbert's *finitist* ideal to ground all of mathematics on a small number of axioms and proving its consistency by constructive means. A central challenge was what Hilbert (1928) would call the *Entscheidungsproblem*: can there exist a mechanical procedure, an *algorithm*, that decides the truth of any given mathematical statement, any given expression in first-order logic? This required a precise definition of the notion of algorithm or *effective computability*, which was provided in a convincing way by Turing (1936). His *universal Turing machine* gave a mathematical model of a device that can implement any conceivable mechanical calculation. It is a mathematical model that came to be instantiated in the digital computer, that now manages all our calculations and has indeed started to turn mechanized mathematical reasoning—*automated theorem proving*—into reality.

Thus runs (in *very* broad strokes) the story of *formalizing* and *mechanizing* deductive or mathematical reasoning. But this story misses something important about Leibniz's original ideal (Hacking, 2006, 135):

> Most readers of Leibniz have taken this to be the cry of some alien rationalism which assumes that every issue can be settled by deductive proof. Quite the contrary. Leibniz was not in general speaking of proving propositions but only of finding out which are most probable *ex datis*.

On Hacking's reading, Leibniz envisioned a logic of *induction*, specifically, a universal calculus for *probabilistic* reasoning. This is a logic of *partial entailment* where we can derive that one statement entails or confirms another to a certain numerical extent, viz., a probability. (From this perspective deductive logic is the special case that only figures probabilities 1.) In fact, Boole in *The Laws of Thought* likewise extended his symbolic logic to a calculus of probabilistic reasoning. He was one of the early proponents of the *logical interpretation of probability*, where the probability of a proposition stands for the degree of belief that a rational agent, on purely logical grounds, *should* attach to it. The subsequent development of mathematical logic completely disregarded probability and inductive reasoning, but its great success in formalizing deductive reasoning still inspired a number of philosophers to try and place probability on the same firmly logical footing: notably Keynes (1921) and Johnson (1932), Wittgenstein and Waismann (1930), and Carnap and co-workers (1945; 1950; 1952; . . .). Keynes's 1921 book attempts an axiomatization of logical probability in the spirit of Russell and Whiteheads *Principia*, and was in general very influential; but by far the most formidable pursuit of the logical approach to probability was the work done within Carnap's program of inductive logic, that lasted several decades and that still has outgrowths today.

Carnap (with Hempel, Reichenbach, Feigl, and others) belonged to the *logical empiricists*, a group of philosophers that for some time in the mid-20th century represented the "received view" in the philosophy of science (Suppe, 1977). They were broadly concerned with exposing 'the logic of scientific inference,' employing the apparatus of formal logic as well as—and increasingly so—mathematical probability and statistics. They were thus concerned with *formalizing* scientific method, or at least that part that belonged to the objective "context of justification" rather than the messy psychological "context of discovery" (Reichenbach, 1938). Perhaps the most important object of their study was the notion of *confirmation* of a scientific assertion by a body of data.

Carnap with his inductive logic indeed sought to give a *quantitative* explication of *degree of confirmation*: this was his logical probability. If successful, this would actually yield a formalization of the most bare form—to Carnap, the most fundamental form—of scientific inference: the extrapolation from current data to a more general conclusion; in particular, to a probabilistic prediction about yet unknown data, the *predictive inference*. It gives a rational and objective *induction rule* for directly going from data to predictions. In the words of van Fraassen (1989, 132),

> Here is the ideal of induction: of a rule of calculation, that extrapolates from particular data to general (or at least ampliative) conclusions. Parts of the ideal are (a) that it is a *rule*, (b) that it is *rationally compelling*, and (c) that it is *objective* in the sense of being independent of the historical or psychological context in which the data appear, and finally (d) that it is *ampliative*. If this ideal is correct, then support of general conclusions by the data is able to guide our opinion, without recourse to anything outside the data—such as laws, necessities, universals, or what have you.

Van Fraassen continues: "Critique of this ideal is made no easier by the fact that this rule of induction does not exist ... Sketches of rules of this sort have been presented, with a good deal of hand-waving, but none has ever been seriously advocated for long." Carnap was ultimately unsuccessful in this regard, too, but his and his coworkers' continued struggle with induction and their engagement with work from mathematical probability and statistics did have an enduring impact on the philosophical debate. According to Zabell (2011, 305, emphasis mine), "Carnap's most lasting influence was more subtle but also more important: he largely shaped the way current philosophy views the nature and the role of probability, in particular its widespread acceptance of the *Bayesian* paradigm." The Bayesian framework is nowadays the most popular unified account of all aspects of scientific reasoning.

The modeling of scientific reasoning in a Bayesian manner evokes the picture of scientists following "Bayesian algorithms" (though the—subjective—input to these algorithms would still be part of something like the context of discovery, cf. Salmon, 1990 in response to Kuhn, 1977). In general, with the rise of the digital computer, any project of formalizing scientific reasoning soon evokes the Leibnizian ideal of *mechanizing* or *automating* scientific reasoning. The image one finds in the early literature is that of the 'learning machine' or 'inductive robot'—often rather in an attempt to bring out the absurdity of the idea of mechanizing all of science. Carnap granted the absurdity of automating the process of coming up with a scientific theory, but stuck to his belief that inductive reasoning can be formalized and indeed be automated into a rule of calculation (1966, 33f):

> One cannot simply follow a mechanical procedure based on fixed rules to devise a new system of theoretical concepts, and with

> its help a theory. Creative ingenuity is required. This point is sometimes expressed by saying that there cannot be an inductive machine—a computer into which we can put all the relevant observational sentences and get, as an output, a neat system of laws that will explain the observed phenomenon.
>
> I agree that there cannot be an inductive machine if the purpose of the machine is to invent new theories. I believe however, that there can be an inductive machine with a much more modest aim. Given certain observations e and a hypothesis h (in the form, say, of a prediction ...), then I believe it is in many cases possible to determine, by mechanical procedures, the logical probability, or degree of confirmation, of h on the basis of e.

Meanwhile, the first humble practical steps towards the ideal of mechanical learning were taken in the nascent field of *artificial intelligence.* Oddly, though, work in artificial intelligence went back to observing a strict separation between a logical and a probabilistic approach, where the rule- and knowledge-based approach dominated up to the point that by the 1980's the probabilistic approach had been all but purged from the field. However, the latter approach reorganized and reemerged under the header of *machine learning*, and its tremendous advance in recent years has resurrected and indeed brought to popular awareness the ideal of automated learning, or automated inductive reasoning.

Without the digital computer that came to instantiate Turing's mathematical model modern science would be unimaginable; the next step, some now speculate, is that the computer can do it all. *Big data* promises a purely data-driven science; and even if we need theory, the context of discovery might yield to automation as well (Gillies, 2001a; Schmidt and Lipson, 2009). But if it is true that the whole of science can be automated, it must in the end take the form of a particular algorithm that extrapolates data to predictions, a modern formulation of the age-old ideal of the induction rule. The "master algorithm," to take a term from a recent popular book on machine learning (Domingos, 2015, 25):

> *All knowledge—past, present, and future—can be derived from data by a single, universal learning algorithm.*

* * *

Carnap, Putnam, and Solomonoff. Putnam in (1963a; 1963b) construed the aim of Carnap's program of inductive logic as the specification of a universal inductive method, and presented a formal proof against the very possibility of such a notion.

Specifically, Putnam (1963a) formulated two conditions of adequacy on any reconstruction of "the judgements an ideal inductive judge would make" (ibid., 778), and proceeded to give a *diagonal proof* to the effect that no Carnapian definition can satisfy both. In (1963b), Putnam explicitly assumed the view that “the task of inductive logic is the construction of a ‘universal learning machine’” (ibid., 303), and accordingly presented his proof as showing the impossibility of this notion. What he had shown, in these terms, is that there can be no *learning machine* that is also *universal*: no inductive method that is effectively computable, that is also able to eventually detect any pattern that is effectively computable.

In 1956, around the same time that Putnam first wrote down his argument, McCarthy organized the Dartmouth workshop that marks the birth of the field of artificial intelligence. The select list of participants included such influential figures as Minsky, Shannon, Newell, Simon—and Solomonoff. Solomonoff, taking inspiration from interactions at this workshop, as well as earlier interactions with Carnap (who was in Chicago when Solomonoff was a student there), spent a number of years thinking about mechanized inductive reasoning and published his findings in (1964). The ideas in this paper, that later found a more secure mathematical footing in the work by Kolmogorov’s student Levin (1970), are important for a number of reasons.

First, they include the earliest formulation of the founding notion of the field of *algorithmic information theory* (also: *Kolmogorov complexity*) in theoretical computer science. Second, they include ideas on universal prediction that have a direct line to developments in modern theoretical machine learning. Third, and this is the focal point of this thesis, they lead to a formal foundation of precisely those aspects of Carnap’s program that Putnam took issue with, and in particular, resurrect the notion of a universal mechanic rule for induction. The resulting *Solomonoff-Levin predictor* qualifies, perhaps, as the definition of a universal inductive machine.

* * *

The Solomonoff-Levin definition. There are two main distinct yet equivalent modern formulations of the Solomonoff-Levin definition. I designate the mathematical result establishing their equivalence, theorem 2.16, as a *representation theorem*, and make much use of it in this thesis.

I give here a rough description of both formulations. In chapter 2 I explain both definitions in detail.

The Solomonoff-Levin definition (1). First, the Solomonoff-Levin definition is a Bayesian mixture—a weighted mean—over a very general class of probability measures over data sequences. Namely, it is a mixture, with a semi-computable prior or weight function, over the class of all semi-computable measures over (finite and infinite) data sequences. Here semi-computability is a weakening of full-blown computability that can be understood as ‘computable approximability from below.’

The Solomonoff-Levin definition (2). Second, the Solomonoff-Levin definition is a transformation of the uniform measure by a universal monotone

Turing machine. More concretely, it assigns to each sequence the probability that it is generated by a universal monotone Turing machine, when this machine is given uniformly random input. Phrased somewhat differently, the probability it assigns to each sequence is given by the input sequences to a universal machine that lead the machine to generate the sequence (the sequence's *descriptions*), where shorter descriptions contribute more probability.

* * *

This thesis (1). In this thesis I investigate whether and how the Solomonoff-Levin proposal can avoid Putnam's diagonal argument to yield a definition of an "optimum," "cleverest possible," or *universal* inductive machine. More broadly, this is a philosophical and historical investigation into the possibility of a perfectly general and purely mechanic rule for extrapolating data: a *universal prediction method.*

This thesis (2). Furthermore, I investigate the common association of the Solomonoff-Levin proposal, and algorithmic information theory in general, with a notion of *simplicity* in terms of *datacompression.* I investigate a suggested justification of the principle of *Occam's razor*, as well as the more recent notion of the *predictive complexity* of data sequences.

* * *

Contributions of this thesis. The main contribution of the current work is a clarification of the philosophical and formal aspects of the Solomonoff-Levin proposal for universal sequential prediction. This includes an explication of the following aspects.

- The historical and conceptual connection of the Solomonoff-Levin proposal to Carnap's program of inductive logic and Putnam's reconstruction and critique of the latter.
- The different possible interpretations of prediction methods and the Solomonoff-Levin method in particular, most importantly as a Bayesian mixture predictor operating under a particular inductive assumption and as an aggregating predictor over a pool of competing prediction methods.
- The notion of universality in sequential prediction, and the distinction between universal reliability and optimality. The interpretation of effective computability as leading to a universal inductive assumption or as leading to a universal pool of prediction methods.
- The weaker notion of semi-computability that is central to the Solomonoff-Levin proposal, and that appears to provide an opening to evade Putnam's diagonal argument.

- The role of simplicity in the Solomonoff-Levin proposal, and the relation to universality. The interpretation of data-compression and the role of the logarithmic loss function in sequential prediction.
- The place of Solomonoff's theory on prediction in the wider area of algorithmic information theory.

This thesis also presents a number of new mathematical results about the Solomonoff-Levin definition, that function to support the philosophical observations. The most important result is theorem 2.13, that gives a generalized characterization of the Solomonoff-Levin measure as a universal transformation.

It sometimes seems like progress in philosophy is mainly of a negative nature: option X cannot work, and option Y is problematic, too. I do not think this is necessarily the case: I think the above main contribution is a positive one and represents genuine progress. Nonetheless, the main conlusions of this thesis are negative:

- The Solomonoff-Levin proposal ultimately fails to escape Putnam's argument, and this failure generalizes: there cannot be a universal prediction method.
- The suggested justification of Occam's razor via the Solomonoff-Levin definition does not succeed. The supposed formalization of Occam's razor in the Solomonoff-Levin definition does not actually go beyond the property of universality.
- The formal notion of predictive complexity falls short of its aim.

* * *

Organization of this thesis. I have divided the thesis into four parts. The two core parts are devoted to the two main themes of universal prediction and of simplicity, respectively. These core parts are preceded by a more informal prelude part, and succeeded by a more formal appendix part that contains auxiliary material and proofs.

Throughout the thesis I have prefixed some section headers with a '*': this is to indicate sections that disrupt the flow of the main text by making a peripheral or technical point, and that can be safely skipped by the reader.

Part I. Prelude. Aims to explain and motivate in an easy-going fashion the central concepts of this thesis, thus setting the stage for the further parts.

This part consists of the following seven sections: on the game of sequential prediction (I.1), the assumption of a deterministic hypothesis (I.2), the assumption of a probabilistic hypothesis (I.3), the constraint of computability (I.4), universal optimality (I.5), the Solomonoff-Levin proposal for universal prediction (I.6), and the Solomonoff-Levin proposal and simplicity (I.7).

Part II. Universality. On the theme of the Solomonoff-Levin definition as a proposed universal prediction method, vis-à-vis Putnam's diagonal argument against the possibility of such a definition.

This part consists of the following four chapters.

Chapter 1. Introduces Putnam's diagonal argument (1.1), explains Carnap's program of inductive logic (1.2), and introduces and positions Solomonoff's approach (1.3).

Chapter 2. A technical chapter. Sets out the definition of the Solomonoff-Levin measure (2.1), and discusses the equivalent mixture definition and presents new results that generalize both (2.2).

Chapter 3. The most important chapter. Charts different interpretations of prediction methods: as stemming from a priori measures (3.1), as mixtures over hypotheses (3.2), and as mixtures over predictors (3.3).

Chapter 4. Wraps up the Carnap-Putnam-Solomonoff storyline. Revisits and critically discusses Putnam's argument (4.1), dismisses the universal *reliability* of the Solomonoff-Levin predictor (4.2), and discusses and finally dismisses the universal *optimality* of the Solomonoff-Levin predictor (4.3).

Part III. Complexity. On the theme of the Solomonoff-Levin definition as providing a formalization of simplicity and a justification of Occam's razor.

This part consists of the following two chapters.

Chapter 5. On the association of the Solomonoff-Levin definition with Occam's razor. Reconstructs and refutes the suggested justification of Occam's razor (5.1), and challenges the simplicity interpretation itself (5.2).

Chapter 6. On more recent work related to the Solomonoff-Levin definition, in particular the notion of predictive complexity. Discusses the theory of prediction with expert advice (6.1) that generalizes universal prediction to different loss functions, introduces and criticizes the resulting notion of predictive complexity (6.2), and points out some further directions of research that arise from the work in this thesis (6.3).

Part IV. Appendices. Consisting of the following.

Appendix A. Contains brief expositions of concepts and results in the periphery of the Solomonoff-Levin theory that come up at various places in the thesis: relating to the Σ_1 semi-distributions (A.1), description systems (A.2), Kolmogorov complexity (A.3), and Martin-Löf randomness (A.4).

Appendix B. Contains all proofs of the results in this thesis, divided into those on the framework of Σ_1 measures and semi-distributions (B.1) and those on sequential prediction (B.2).

*

Part I

Prelude

This part builds up and motivates the main concepts and themes of this thesis. As such, it precedes—forms a prelude to—the more detailed work done in the subsequent parts.

In I.1, I introduce the setting of sequential prediction and point out the basic problems with induction. In I.2, I further illustrate these problems by means of a diagonal argument, and introduce the idea of constraining the problem by assuming a class of possible deterministic hypotheses. In I.3, I consider the more general case of probabilistic hypotheses, and introduce the Bayesian approach to sequential prediction.

In I.4, I introduce effective computability as an inductive assumption about Nature. In I.5, I explain that effective computability is more convincing as a constraint on prediction methods, which leads to the new goal of universal optimality. Unfortunately, the most straightforward way of defining a universally optimal prediction method is blocked by Putnam's diagonal argument.

In I.6, I introduce the Solomonoff-Levin proposal as an attempt to avoid diagonalization and thereby obtain a definition of a universal prediction method. This provides the basis for part II of the thesis. In I.7, I introduce the association of the Solomonoff-Levin proposal with Occam's razor. This provides the basis for part III of the thesis.

I.1. Sequential prediction

A game of prediction. Imagine prediction as a game we play against Nature. The latter repeatedly issues a symbol, either 0 or 1: it is helpful to visualize this as the tracing of an upward path through a binary tree, figure 1. Our task, at each such successive trial, is to predict the symbol Nature will play. To spell this out (for an overview of notation, see page v): at each trial $t+1$,

- we issue a *prediction* p, based on the sequence $\boldsymbol{x}^t$ of symbols that Nature has generated so far;
- Nature reveals the next outcome x_{t+1};
- we suffer a *loss* $\ell(p, x_{t+1})$ that quantifies how much our prediction was off.

Our predictions are *probabilistic*: p is a probability distribution over $\mathbb{B} = \{0,1\}$, the possible outcomes. A *prediction strategy* specifies a prediction for each possible state we might find ourselves in: for each node in the binary tree it assigns a probability to both outgoing branches. Thus a prediction strategy p is a function from $\mathbb{B}^*$, the finite sequences, to $\mathcal{P}$, the distributions over $\mathbb{B}$.

The simplest of examples of a prediction rule is the *indifferent rule*, that always says fifty-fifty: $\mathsf{p}(\boldsymbol{x}) = (\frac{1}{2}, \frac{1}{2})$ for each $\boldsymbol{x} \in \mathbb{B}^*$. A slightly more sophisticated example is a *rule of succession*, a prediction rule that takes into account the relative fequency of symbols in the data so far. For instance, *Laplace's* rule

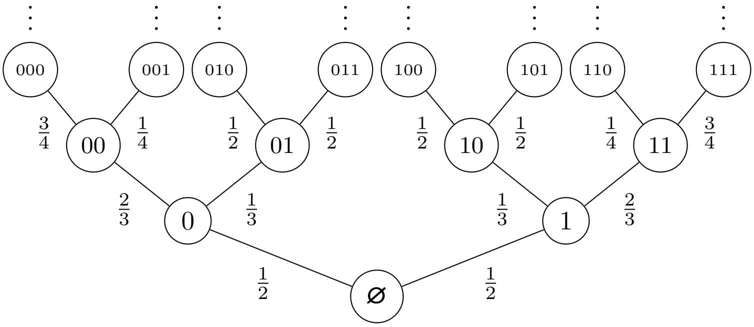


FIGURE 1. The binary tree of the prediction game. Nature sets out on a path from the root upwards; we try to predict every next symbol using a prediction strategy that assigns probabilities to each node's two upward branches. The values depicted here are those given by Laplace's rule of succession.

of succession is defined by

$$\mathsf{p}(\boldsymbol{x}^t) = \left(\frac{\#_0\boldsymbol{x}^t + 1}{t+2}, \frac{\#_1\boldsymbol{x}^t + 1}{t+2} \right).$$

Figure 1 shows the values it gives for sequences up to length 2.

A standard loss function is the *logarithmic loss function*, defined by

$$\ell(p, x_{t+1}) := -\log p(x_{t+1}).$$

So if we assigned probability 1 to what was to become the actual outcome x_{t+1}, we incur loss 0; as we were more careful and assigned less probability to x_{t+1}, our loss increases; and if we made an extreme prediction the other way and assigned probability 0 to x_{t+1}, we incur loss *infinity.* I will for now, by way of illustration, assume the logarithmic loss function; but later in the thesis (specifically, chapter 6) I will also discuss other loss functions.

This is the basic framework of the prediction game that I assume throughout this thesis.

The generality of this setting. The starting assumption of this thesis is that a maximally general and abstract setting is useful for the study of foundational questions—ultimately, in our case, the fundamental question of epistemology: *what can we know?* Being granted this, however, we face the problem of producing a framework that attains that generality. Does the above framework of sequential prediction suit our goal, if this goal is to examine the limits of scientific or statistical inference?

One can, to begin with, object that scientific inquiry consists not so much in producing forecasts as in inferring general conclusions: not so much in prediction as in the identification and the confirmation of hypotheses and theories. (That *is* studied within a similar abstract framework in *formal learning theory*, see Kelly, 1996.) One can thus object to the generality of the problem setting of prediction; one can further object to the way predictive inference is rendered in our framework. As Dawid (1985b, 279) writes, when he introduces this framework under the label of "prequential forecasting,"

> This formalism may appear to be an uncomfortable straightjacket into which to squeeze statistical theory. The data may arrive *en bloc*, rather than in a natural order; if they come from a time-series, it may be impossible, or not obviously desirable, to analyse them at every point of time, or to formulate one-step ahead forecasts; and the restriction whereby all uncertainty about the next observation is to be encoded in a probability distribution, while acceptable to Bayesians, may not appeal to others.

In addition, we stipulate a binary alphabet to express the data, rather than allowing for any countable or even continuous alphabet (though this is not an essential limitation, cf. Hutter, 2003b), or indeed an *unknown* alphabet (the *sampling of species* problem, see Zabell, 1992). Finally, this setting of passive prediction leaves out the component of active data-gathering, which is taken into account in *reinforcement learning* (see Sutton and Barto, 1998).

While there certainly is a case to be made that prediction is at most a subsidiary part of science, there is also, as I highlighted in my historical sketch in the introduction (page 2), an important opposed tradition, fashionable in parts of machine learning today, that takes it that scientific inference ultimately comes down to inductive inference from particulars to particulars, or predictive inference. (In machine learning, the term *transduction* is sometimes used to distinguish the inference to particulars from *induction*, which then takes the specific meaning of inference to general conclusions, see Vapnik, 1998.) This is an important motivation for investigating the limits of inference within the general setting of prediction—for investigating the possibility of universal prediction. Moreover, while our framework of sequential prediction certainly cannot accommodate everything there is to say about prediction, I think, and I assume in this thesis, that it possesses a level of generality that lends significance to the conclusions we draw from it. For the rest, I will side with Dawid, who concludes his above enumeration of concerns (ibid.):

> All these are valid *prima facie* objections; but I would respond by suggesting that, if you will tentatively join me in following through the implications of the prequential approach, you may find that it offers new insights enough to offset such disquiet.

The goal: a universal prediction strategy. So if prediction is a game, how do we *win*? In other words, what is our goal in the prediction game?

Informally, our goal is to predict well; and this means, in the current framework, to keep our losses to a minimum. But how and to what extent can we achieve that?

Let us set off from a basic intuition about what is required for good prediction. What makes the indifferent rule a silly prediction method? It is the fact that in a clear sense, it never *learns* anything. No matter the moves Nature makes, no matter the regularity the resulting data sequence exhibits, the indifferent method remains unmoved and sticks to the exact same forecasts—and every single trial it incurs the same positive amount of loss. A rule of succession is more sophisticated because it does allow itself to be informed by the sequence: it adjusts its predictions to the observed relative frequencies. It extrapolates a regularity from the past in its predictions about the future. In that sense, it can learn from the data. However, in the same sense, it is extremely limited in the things it can learn. Its predictive probabilities are still completely uninformed by any *order* effects in the data. Even in the favourable case that the sequence that Nature constructs exhibits a stable relative frequency, and the method's predictions eventually converge on this frequency, there is (unless the relative frequency is an extreme value 0 or 1) at least a specific positive amount of loss it keeps on incurring every single trial.

The next step would be a method that can learn enough from the data about the sequence that is being constructed to actually make its losses eventually *go down*. That is, a method that makes the loss it incurs every trial converge to 0.

Let us tentatively formulate this as our goal in the game (the 'winning condition'): to make the losses go to 0. Then a *universal* prediction strategy would be a prediction strategy that *always* manages to attain this goal, that is, that manages to make the losses go to 0, *no matter what* Nature does. Intuitively, such a universal method should always be able to learn from the data; it should always be able to eventually discover the regularity in the past and predict well by extrapolating it.

But if we think about this just a little more, we soon realize how wildly overambitious this goal is.

The problems with induction. The problem with the simple prescription to extrapolate the pattern of the past is that at any given time, there is any number of regularities we can recognize in the data so far (Goodman, 1954, 82):

> To say that valid predictions are those based on past regularities, without being able to say *which* regularities, is thus quite pointless. Regularities are where you can find them, and you can find them anywhere.

The sequence $\boldsymbol{x}^9 = 010011000$ follows the pattern 'repeat for $n = 0, 1, 2, \ldots$: $n-1$ times 0 and $n-1$ times 1' (so the next symbol would be 1); but it also follows the pattern 'repeat for $n = 0, 1, 2, \ldots$: 2^n times 0 and 2^n times 1' (so the next symbol would be 0). The simple fact that any finite evidence can be generalized in an infinite number of ways has been the subject of discussion by earlier authors, for instance Jeffreys (1939, 3) and Poincaré (1902, 173); but Goodman (1946) first explicitly formulated it as a problem for Carnap's confirmation theory, and related it to the problem of induction (1954, 59ff). Goodman's *new riddle of induction* (ibid.) then reads: *granted* that it is a good idea to predict by extrapolating from the past, then we still do not know *which* of these many patterns to extrapolate. The original problem of induction, going back to Hume, is that what is granted in this statement of the new riddle: do we actually have any good reason—any *justification*—for trying to extrapolate patterns of the past?

Hume's skeptical argument starts from the observation that inductive reasoning must "proceed upon the supposition, that the future will be conformable to the past" (1748, 62). But what rational reason can we give for adopting this 'principle of the uniformity of Nature's strategy'? We cannot justify it on the grounds that it has held in the past, because this "must be evidently going in a circle, and taking that for granted, which is the very point in question" (ibid., 63). Or, if this is not directly circular, we need a principle of uniformity on a higher level ('if extrapolating patterns of the past has been successful, then it will remain so'), which for its justification requires yet a higher principle: and we are led into an infinite regress. Nor, of course, can we justify induction deductively: "it implies no contradiction, that the course of nature may change" (ibid., 61). To be sure, if the only constraint on Nature is the bare framework of the prediction game, then Nature can basically do whatever, whenever. Nature can indeed be *adversarial*: it can explictly sabotage our predictions. Namely, to take the most extreme case, it can play symbol x whenever we give it predictive probability no more than 0.5; thus making sure that every trial we incur at least a same high amount of loss. In other words, it is possible that our basic starting point is false: there is just nothing that can be learned from the data.

The problem of induction. Hume's problem of induction as I sketched it within our framework of sequential prediction might very well leave the impression of a purely logical observation: induction cannot be justified because as a matter of logic anything can happen. Note again, though, that this is only the second part of the argument, saying that we cannot give deductive reasons for induction; the other part is that we cannot give *inductive* reasons for induction. This is what makes it an extremely powerful argument; an argument that so far has withstood any attempt at a solution. Nevertheless, it is hard to shake off a first impression of the problem of induction as something of a frivolous puzzle, and I should say some more on why it is a genuine *philosophical* problem, indeed the central problem in the philosophy of science.

It is a genuine problem because inductive reasoning, the procedure of extrapolating observational data to more general conclusions, is to many philosophers (although not to all: Popper, for one, famously disagreed) the very hallmark of science. On this view, scientific reasoning *is* inductive reasoning. But then it is a *profound* problem that science, supposedly the most rational of human enterprises, is at heart a procedure that cannot be rationally justified, cannot be supported by good reasons. It suggests that science is ultimately also only a leap of faith, no better than reading tea leaves or any other irrational practice. This is clearly unsatisfactory: and some answer to the problem of induction would therefore "not only be of fundamental *epistemological* importance; it would also be of fundamental *cultural* importance as part of the enterprise of enhancing scientific rationality" (Schurz, 2008, 280); see especially the lucid explanation of Salmon (1967, 54ff) of the significance of the problem of induction.

This signifance stands in contrast, again, to the deceptively simple form of Hume's argument, and the fact that no scientist will (or should) feel compelled to suspend his activities for it. Recognizing this, Howson (2000, 10) sets apart the original problem of induction from what he calls *Hume's problem*, "the problem of reconciling the continuing failure to rebut Hume's argument with the undoubted fact that induction not only seemed to work but to work surpassingly well."

Thus, to the extent that our framework of sequential prediction captures the essence of inductive inference, it is important to try and direct our search for a universal prediction method to a possible justification of induction (see I.5 below; and 4.2-4.3); and if this fails, to try and understand *why* it fails. That is a main aim of this thesis.

* * *

I.2. Deterministic hypotheses

We return to the problems with induction in our framework of sequential prediction. It seems that these problems leave us no option but to impose constraints on the prediction game—and worry about the justification for those later.

A first encounter with diagonalization. As a start, to counter the above-mentioned possibility of Nature explicitly sabotaging us, how about we deny Nature access to our predictions? This is not enough, it turns out: Nature does not even have to be reactive to our particular prediction strategy to be adversarial.

We will now assume that there are only countably many possible prediction strategies, an assumption that will later be motivated in some detail. Under that assumption, Nature can generate data sequences that will make us fail to converge *no matter* what prediction method we choose to follow. This can be shown by a diagonal argument, the type of argument (as mentioned in the introduction, page 5) that Putnam used against Carnap.

Since we assumed there exists a countable number of possible predictors, we can assume Nature keeps a list $\{\mathsf{p}_i\}_{i\in\mathbb{N}}$ of all of them. Now the most straightforward diagonal history has predictor p_i fail at trial $i+1$, which is to say that Nature selects the next outcome x_{i+1} such that $\mathsf{p}_i(x_{i+1}, \boldsymbol{x}^i) \leq 0.5$. This guarantees that every predictor fails at some point—though it still leaves open the possibility that predictors will converge to correct predictions after this single failure. A more refined diagonalization, depicted in figure 2, makes sure that every predictor keeps on failing, and so never makes its loss go to 0 (also see Sudbury, 1973). Rather than continuing to p_2 after making p_1 fail for the first time, Nature backtracks and first makes p_0 and p_1 fail a second time. Then, after making p_2 fail for the first time, before turning to p_3, Nature backtracks again and makes the first three predictors fail another time. So it continues, each time extending to one more predictor before backtracking and making each of the previous predictors fail one more time. The result is that each predictor will fail infinitely often, and no predictor makes its loss go to 0.

I used here a dynamical language that still paints Nature as pursuing a strategy that reacts to predictions, in this case the predictions of all strategies. But the argument establishes an existence claim that is independent of what we, the player making the forecasts, do. The earlier adversariality, with Nature reacting to our particular predictions, gives the statement: for every prediction strategy, there exists a history that makes it fail infinitely often. (And this history depends on the prediction strategy.) The argument of this section gives the statement: there exists a history that makes every prediction method fail infinitely often.

The procedure can be extended by having Nature intersperse the diagonalization moves with playing the successive symbols of some given infinite sequence $\boldsymbol{x}^\omega$ (Schervish, 1985b). So every odd trial it makes the next move in the original diagonalization; every even trial it plays the next symbol of $\boldsymbol{x}^\omega$. Since there are uncountably many infinite sequences $\boldsymbol{x}^\omega$, there are uncountably many ways of generating such a history, each of which makes each predictor fail infinitely often. So here we have another expression of the impossibility of universal prediction in the naive sense: there are uncountably many histories that make each prediction method fail infinitely often, that are unlearnable.

Making assumptions: deterministic hypotheses. If enforcing constraints on Nature is the way we choose to go, these constraints need to go beyond just denying Nature access to our predictions. We would actually have to stipulate that Nature can only choose from a limited number of ways of generating the data.

To use a better term, that stays clear of suggesting that we can actively enforce metaphysical constraints on Nature: we would have to *assume* that

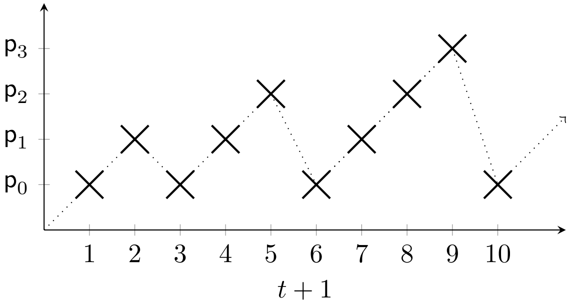


FIGURE 2. The construction of a diagonal history. The horizontal axis marks the trials; the vertical axis an enumeration of all prediction methods. A cross at $(\mathsf{p}_i, t+1)$ signifies that Nature makes p_i fail at trial $t+1$, by issuing symbol x_{t+1} when $\mathsf{p}_i(x_{t+1}, \boldsymbol{x}^t) \leq 0.5$.

Nature chooses from a limited number of ways of generating the data. These possible data-generating strategies we call our *hypotheses*.

For instance, we can assume that Nature chooses only one of countably many infinite sequences: these are our *deterministic* hypotheses. This does the trick—the following prediction strategy will, under that assumption, be sure to make our losses converge to 0. Keep an ordered list of all the hypotheses, i.e., infinite sequences; at each trial throw out the sequences that are refuted by the previous symbol, and assign a predictive probability $1-2^{-i}$ to the next symbol of the sequence ranking first in the updated list, where i is the number of trials the sequence has been at the top of the list already. Then at some point all the incorrect sequences that we originally listed ahead of the sequence Nature chose are refuted, and we will give increasing probability—indeed, converging to 1—to the symbols Nature actually selects. Hence the loss we incur at each trial converges to 0. (Why not simply assign probability 1 to the symbol given by the first sequence in the list? Because we need to guard against incurring logarithmic loss infinity if this sequence is refuted. Of course, this is not an issue if we instead use a bounded loss function.)

Who gets to go first? Thus assuming that Nature is limited to countably many (deterministic) strategies is enough to enable us to specify a prediction strategy that, under that assumption, will always succeed. Note that this mirrors the earlier diagonalization result: that limiting *us* to countably many prediction strategies is enough to enable Nature to specify a history that will always make us fail. These two results pull in different directions, potentially leading to funny consequences if we are not careful about what gets constrained first and why.

For instance, what if we include among our hypotheses one of the unlearnable diagonal sequences of the previous section? Then the simple method we just saw will make our losses converge to 0 if Nature plays this sequence—it

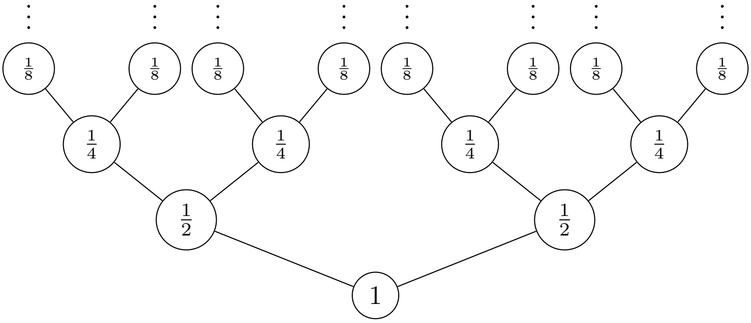


FIGURE 3. The Lebesgue or uniform measure on the binary tree. The nodes that as before represent the possible finite sequences are now labeled with their probabilities according to the uniform measure.

will learn the unlearnable sequence! What has happened here is that this new prediction method must fall outside of the earlier fixed class of *all* methods. But *why* cannot this procedure count as a proper prediction method? Clearly, we need to be more precise about what we admit as proper prediction methods, and this will in fact be a crucial step further below.

But first we need to consider the case of *probabilistic* hypotheses, which will bring us to the *Bayesian* approach to sequential prediction.

* * *

I.3. Probabilistic hypotheses

Rather than revealing the successive symbols of a fixed sequence, Nature might itself proceed probabilistically, at some steps (or each of them) tossing the proverbial coin to decide on the next symbol.

Making assumptions: probabilistic hypotheses. In full generality, such a probabilistic data-generating strategy is given by a *probability measure*, an assignment of probabilility to each node in the binary tree, where the total probability at each level is normalized to 1 (the probabilities assigned to all same-level nodes sum to 1). (Figure 3 depicts as an example the 'fully random'—*uniform*—measure where each same-length sequence has the exact same probability.) Formally, this is a function $\mu : \mathbb{B}^* \to [0,1]$ such that

$$\mu(\varnothing) = 1;$$
$$\mu(\boldsymbol{x}0) + \mu(\boldsymbol{x}1) = \mu(\boldsymbol{x}) \text{ for all } \boldsymbol{x} \in \mathbb{B}^*.$$

(*Really* formally, it corresponds to a probability measure on the Cantor space, the class of infinite sequences. See 2.1.1.)

A deterministic strategy—an infinite sequence—is a special case of a probabilistic strategy, where the infinite sequence's initial segments are all assigned probability 1.

Hypotheses and strategies. I defined a probabilistic hypothesis here as a measure on the full binary tree, but we can also see it as a function that to each node assigns the distribution governing which symbol is generated next. That is, *formally*, a probabilistic hypothesis or data-generating strategy is equivalent to a prediction strategy, that is also a function from the finite sequences to distributions on $\mathbb{B}$. I return to this formal equivalence between hypotheses and prediction strategies in I.5 below.

Losses and regrets. In the face of probabilistic data-generating strategies, the goal of reducing the losses to 0 becomes utterly unfeasible. Consider the fully random data-generating strategy (figure 3). Even if we knew that Nature plays this strategy, and we always issue the *correct* predictive probabilities, those that coincide with the actual probabilities (in this example, this is actually our naive *indifferent* strategy, $\mathsf{p}(\boldsymbol{x}) = \left(\frac{1}{2}, \frac{1}{2}\right)$), we would still incur the same positive loss each single trial. One could say: *Nature itself* would be unable to make its losses go to 0.

(Of course, for each finite sequence that is randomly generated, we *could* have kept our loss arbitrarily low by somehow having assigned arbitrarily high predictive probabilities to the actual symbols of this sequence. From this *ex post facto* perspective, the strategy that issues the actual probabilities was not the best possible strategy. But there is a clear intuition that the actual probabilities, those aligned with the actual data-generating strategy, are the best possible predictions: and they are in a precise way, namely *in expectation*. Consequently, we normally assume a loss function to be such that in expectation, the loss is minimized by issuing the actual probabilities. Such loss functions—among them the logarithmic loss function—are called *proper*. See 6.1.2.)

A more reasonable measure is therefore the *surplus* loss relative to the best possible strategy, the strategy p_μ such that $\mathsf{p}_\mu(\boldsymbol{x}) = \mu(\cdot \mid \boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{B}^*$. This surplus loss $\ell_\mathsf{p} - \ell_{\mathsf{p}_\mu}$ we call the *regret* (*relative to* μ) of p. (Note that a deterministic hypothesis μ is the special case where p's regret is its loss.)

A universal prediction method we then take to be a method that always—no matter the μ that Nature plays—makes its regrets relative to μ go to 0.

Bayesian prediction. An important instance of the general approach of making assumption about Nature, with the added benefit of naturally accommodating probabilistic data-generating strategies, is the *Bayesian* approach to sequential prediction (see Dawid, 1984, 280).

We start out again with a limited number of hypotheses, that are now probability measures. For ease of presentation and with an eye to what follows below, I will again take this to be a countable number; so we have some indexed *hypothesis class* $\mathcal{H} = \{\mu_i\}_{i \in I}$ for countable index set I. We then put a *prior probability distribution* over this class: a function w over the indices in I that is everywhere positive and that sums to 1.

As we observe the sequence Nature presents, we update the prior distribution to a *posterior* distribution over the hypotheses. We follow *Bayes's rule* in equating this posterior with the conditional prior $w(\cdot \mid \cdot)$, which by Bayes's theorem is given by

$$w(i \mid \boldsymbol{x}) = \frac{\mu_i(\boldsymbol{x})w(i)}{\sum_i \mu_i(\boldsymbol{x})w(i)}.$$

The Bayesian mixture predictor issues at each trial the posterior-weighted average of the probabilities given by the hypotheses,

$$\mathsf{p}_{\mathsf{bayes}}(\boldsymbol{x}) = \sum_i w(i \mid \boldsymbol{x})\mu_i(\cdot \mid \boldsymbol{x}). \tag{1}$$

Now, importantly, one can prove (again, for the countable case) that if Nature chooses a strategy that is a hypothesis μ in $\mathcal{H}$, then the Bayesian mixture method indeed makes its regrets relative to μ converge to 0.

(All of this is treated in more detail in 3.2.2.)

Bayesianisms. The term 'Bayesian' is slightly treacherous, because it can refer to any of at least 46,656 different things. Not even the update rule that bears Bayes's name is undisputed among all self-professed Bayesians. But if there is a common core to all varieties of Bayesianism *in philosophy*, it is the allowance for a particular interpretation of probability: the *epistemic* interpretation as an agent's *degrees of belief*.

This can be seen to subsume the logical interpretation pursued by Carnap, that I mentioned in the historical sketch in the introduction, page 2. On the logical interpretation—in its strongest form—probabilities are the logical-*objective* degrees of belief of the *uniquely* rational agent. Various Bayesian interpretations can indeed be seen as taking various positions on a scale of objective-subjective, where at the objective end lies the logical interpretation and at the purely subjectivist end the only rationality constraints left are those of *coherence*, or adherence to the Kolmogorov axioms of probability. As a matter of historical fact, Carnap would drop more and more rationality constraints and so moved in the direction of—and helped popularize—the subjective Bayesian philosophy (see Zabell, 2011 and also 3.2.5).

Our postulation of a class of probabilistic hypotheses about the actual *objective* state or strategy of Nature—what Diaconis and Freedman (1986, 11) call the *classical* Bayesian interpretation, because the assignment of a prior probability assignment to the unknown parameters of a statistical model goes back to Bayes and Laplace—is actually anathema to truly subjective Bayesians like de Finetti, who believe that the very concept of an unknown objective probability is meaningless. On the other hand, modern-day Bayesian approaches in statistics take a much more pragmatic perspective, where not even the interpretation of degrees of belief is necessarily retained (e.g., Gelman and Shalizi, 2013).

This interpretation of probability as degree of belief is something I also do not necessarily want to assume when referring to the prediction method given by (1). It may be a natural interpretation of a prior over a hypothesis class (these are the things we—to various degrees—*believe* Nature might do), but it does not seem necessary (perhaps there are things we believe possible but we prefer not to think about?). On a minimal interpretation, these hypotheses are simply the *possibilities that we take into consideration*, with different *weights*. For that reason, I will mostly prefer to refer to predictor (1) by the more neutral denotation '*mixture predictor* p_{mix},' and to refer to the prior as the 'weight function.' Chapter 3 gives a much more detailed account of possible interpretations of (mixture) prediction methods.

Hume, Bayes, and Goodman. The new riddle of induction asks *what* patterns we should extrapolate when we do induction. With a (Bayesian) mixture prediction method we answer Goodman's riddle by stipulation: those (probabilistic) patterns that are given by the hypotheses in our class, i.e., those that we assign positive prior probability or weight. (Somewhat more precisely: at each trial, those patterns that are given by the hypotheses that have retained positive posterior probability, and weighted by *how much* posterior probability.)

In the terminology of Howson (2000), the choice of prior distribution constitutes our inevitable "Humean inductive assumptions." (Also see Romeijn, 2004, 357ff.) Howson (ibid., 88):

> 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.*

Thus the great merit of the Bayesian formal approach is that it locates our inductive assumptions very precisely: in the prior. (See 3.2.2.)

* * *

I.4. Computability

Hacking (2001, 184f) writes,

> Here is an odd fact, a coincidental (?) relation between the early days of the Bayesian philosophy and the early days of computer science.

> During the Second World War, Savage was the chief 'statistical' assistant to John von Neumann, the great mathematician who built the first electronic computer, and introduced the modern age of computers and information.
>
> There was one other great advocate of Bayesian ideas directly after World War II, the English probability theorist I.J. Good. I.J. Good was an assistant to A.M. Turing. Turing defined the idea of the ideal computer and proved the fundamental theorem about ideal computation. That is why today we speak about Turing machines.
>
> It is as if the modern Bayesian idea is a byproduct of the age of computers.

Here I will consider the following inductive assumption, to be formalized in a Bayesian mixture strategy: Nature's possible data-generating strategies are *effectively computable*.

Since computability plays a central role in this thesis, I start with a discussion of the fundamentals: the notion of Turing machine and the Church-Turing thesis.

Turing-computability. The formal notion of computability is studied in mathematical logic under the header of *recursion theory* or (more recently, see Soare, 1996, 1999) *computability theory* (classic textbooks are Rogers, 1967; Soare, 1987; Odifreddi, 1989; a new textbook is Soare, 2016). The founding notion is Turing's (1936) model of a computing machine.

A *Turing machine* models an idealized 'computor' (that is: a person) who, only aided by pen and paper, numbly and tirelessly follows a set of basic instructions. Specifically (see ibid., 249ff), the computor works on a potentially infinite paper tape divided into squares, and what she does at each step is determined by the symbol she reads at the current square and her current "state of mind" (there are finitely many possible symbols and states). More specific still, at each step she consults a finite list of instructions that are of the form: if in this state and reading this symbol, then do this (either: move to left or right, or write new symbol; and/or change state). (If there is no matching instruction, the computor gets stuck. It can also happen that the computor goes into an infinite loop, which occurs if she ends up in a *configuration*—a combination of state and all symbols on the tape—she was in before.) The machine is thus specified by the instructions; the input to the computation is given by the symbols on the tape at the start, and the output is given by the symbols on the tape when (if at all!) the computor arrives in the unique halting state. In the standard modern presentation of the Turing machine, the human computor is replaced by an abstract movable read/write head with a set of internal states (see Odifreddi, 1989, 46ff, Soare, 2016, 7ff).

Of particular significance is Turing's specification of a *universal* computing machine (1936, 241), a *universal Turing machine* that can emulate every other Turing machine.

Turing's theoretical model predates the spectacular rise of the digital computer, but the obvious analogy to our modern computing devices is striking. A Turing machine corresponds to a single algorithm, or computer program, and a universal Turing machine to a familiar all-purpose computer, that can execute any algorithm given the right program.

Let us call a function (relation, set, problem, ...) *Turing-computable* if it is computable by some Turing machine, or equivalently, by any universal Turing machine. Turing showed that some well-defined problems are *not* Turing-computable: in particular, the *Halting problem*—given a (code for a) Turing machine and an input, does this machine eventually halt on this input?— is not solvable by any Turing machine (1936, 247f). But Turing's main target was the Entscheidungsproblem (see page 2). He showed (ibid., 259ff) that if the Entscheidungsproblem—given a (code for a) formula in first-order logic, is this formula provable?—were solvable by a Turing machine, then so would the Halting problem: hence the Entscheidungsproblem is not solvable by any Turing machine.

The Church-Turing thesis. But does that mean that the Entscheidungsproblem is not effectively solvable, full stop? What I am asking here is: can we equate the formal notion Turing-computability with the informal notion of effective (mechanical, algorithmic, ...) calculability? The *Church-Turing thesis* is the statement that we can.

There are a number of reasons in support of the identification of effective computability with Turing-computability (*Turing's thesis* in Kleene, 1952, 376, and first dubbed the *Church-Turing thesis* in Kleene, 1967, 232). One reason is the remarkable *confluence* of several different proposed models (including Church's λ-definability, Gödel's general recursiveness, and Turing-computability) that all turned out to be extensionally equivalent (Gandy, 1988). (Although one could also say that the very fact that they are mathematically equivalent shows that they were not so different to begin with, see Sieg, 2008, 563.) Another is the fact that we have to date not discovered any algorithm that would not be implementable on a Turing-machine. (Although why would this give reason to think that we will never find such an algorithm: the problem of induction, ibid.) A reason that is also sometimes mentioned (Piccinini, 2011, 738), is the fact that the class of Turing-computable functions *cannot be diagonalized*—a fact that we will return to below. But perhaps the most compelling reason and certainly the distinctive appeal of Turing's model was suggested by Turing himself (1936, 249ff; also see Church, 1937, 43): it just seems evident that we could reproduce or program every step of any effective procedure directly into an instruction for a Turing machine.

Though here it must be stressed again that Turing analyzed what a human computor could possibly calculate; this is maybe not the same as what discrete mechanical devices or machines could possibly calculate. Copeland

(2000; 2006) argues that Turing was strictly concerned with the first identification; Hodges (2006) objects that it is not clear at all that Turing and others at the time were attached to this distinction. Turing's student Gandy (1980) did draw the distinction between a *Thesis T* on human computability and a *thesis M* on machine-computability, yet he also abated the distinction by formulating a number of requirements on discrete machines and showing that Turing-computability in fact follows from those (these conditions of locality and boundedness as axioms for computability are further streamlined by Sieg 2002a; 2002b; 2008, 586ff). Both theses are, in any case, concerned with purely mechanical computability, which sets them apart from the much stronger thesis about *physical* computability, that would be refuted by the existence of (analogue) devices that can harness incomputable (results of) physical processes. The Entscheidungsproblem motivated an *epistemological* thesis about what a mathematician can possibly prove, and the Church-Turing thesis as the union of Theses T and M is still an epistemological thesis about what can be possibly calculated (by man or machine) in a purely mechanical manner.

It is safe to say that the Church-Turing thesis in this form is generally accepted, and I will assume it here, too. Apart from the *unavoidable* use of the nontrival direction of the Church-Turing thesis (where we infer Turing-computability from computability, or, as in the the case of the Entscheidungsproblem, *in*computability from Turing-*in*computability), there is also the common *lazy* use in proofs where we infer the existence of a particular Turing machine from a description or merely a sketch of an algorithm (terminology by Boolos et al., 2007, 83). I will henceforth mostly leave the use of the Church-Turing thesis implicit and treat 'effective computability' and 'Turing-computability' as synonymous.

Computable hypotheses. A computable deterministic hypothesis is a computable infinite sequence. A computable probabilistic hypothesis is such that its probability values are given by a computable function. (See page vii and also 2.1.1.)

The class of computable hypotheses is nice because it is at the same time very small and very large. It is very small because it is countable (there are only countably many Turing machines), a speck in the uncountable vastness of all logically possible measures. Yet it is very large because it contains every hypothesis that we could ever specify in sufficient detail to actually calculate its values.

If we thus put the constraint—the inductive assumption—of computability on Nature, we can specify a universal prediction strategy. In particular, the Bayesian prediction method with the class of all computable hypotheses will make its regrets relative to the actual measure go to 0—under the inductive assumption of computability.

'Relative universality.' The Bayesian prediction method over the class of computable hypotheses is thus a universal prediction method—for the class of computable hypotheses. In general, what we have uncovered so far is the possibility of universal prediction methods for a given limited (countable) class of hypotheses. This is, sure enough, straining the word 'universality' (also see Grünwald, 2007, 175)—as conveyed by the oxymoron 'relative universality,' relative to a given hypothesis class. 'True universality' is only universality relative to the universal class of *all* hypotheses.

The problem of induction indicates that it is impossible to specify a mixture over the class of *all* hypotheses: any inductive assumption must be restrictive, universality must be relative. An a priori restriction or assumption on Nature is inevitable—but any such restriction must lack justification (see Howson, 2000 and 4.2).

The inductive assumption of computability. What about the restriction of computability? Rathmanner and Hutter (2011, 1118) write:

> It should be appreciated that according to the Church-Turing thesis, the class of all computable measures includes essentially any conceivable natural environment.

Howson (2000, 77), when discussing the claim that only the computable hypotheses represent "genuine discussable hypotheses," demurs:

> it is just not true that we can consider only denumerably many hypotheses. We have seen that in the language of ordinary analysis hypothesis spaces of uncountably many elements are dealt with as a matter of course. The fact is that these are all possibilities and they cannot be ignored at the behest of an arbitrary restriction on language.

Certainly a statistician will feel uncomfortable with being restricted to only countably many hypotheses: already the lowly model of Bernoulli distributions with parameters in the interval $[0, 1]$—uncountably many real values, hence most of them incomputable!—would be prohibited. But the relevant issue here is not so much whether or not *we* can conceive of these possibilities or genuinely discuss them (though this will be important in I.5 below!): the point is rather that *these are all possibilities*. There are uncountably many possible things Nature can do, and our restriction to the *computable* possibilities is just that: a restriction of possibilities.

This restriction is definitely not equivalent to the Church-Turing thesis. We would need some kind of physical variant of the Church-Turing thesis, and a "bold" one at that, that not only says that what Nature can *compute* must be Turing-computable but indeed that what Nature can *do* must be Turing-computable (Piccinini, 2011). This is a fertile topic for speculation, but at the end of the day there simply seems little justification for promoting the eminently *epistemological* notion of computability to a *metaphysical* constraint on the world.

* * *

I.5. Optimality

We have so far aimed for universal *reliability*: to predict successfully no matter what Nature does. But truly universal reliability is impossible because Nature might proceed in such a way that there is simply nothing to be learned from the data. This is a possibility, there is no way around it—yet it is a possibility that in a way is not so interesting. If there is nothing we can do, there is nothing we can do. The interesting case is when Nature is such that it *is* possible to learn. Accordingly, we can aim for a universal method to be successful in the more interesting case: *whenever it is possible at all to be successful*, whenever *some method* is successful. Rather than aiming for a universally reliable method, we aim for a universally *optimal* method.

Reichenbach's vindication of induction. This is the basic idea behind Reichenbach's attempted *pragmatic justification* or *vindication* of induction: whenever it is possible at all to be successful, the inductive method will be successful (Reichenbach, 1933, 421f; 1935, 410ff 1938, 348ff; Feigl, 1950; see Salmon, 1967, 52ff, 85ff; Salmon, 1974; 1991). "Die Induktionsregel ist die günstigste Setzung, weil sie die einzige Setzung ist, von der wir wissen: wenn es überhaupt möglich ist, Zukunftsaussagen zu machen, so werden wir sie durch diese Setzung finden" (1935, 418). In more poetic language (ibid., 420):

> Ein Blinder, der sich im Gebirge verirrt hat, tastet mit seinem Stock einen Pfad. Er weiß nicht, wohin ihn der Pfad führt, auch nicht, ob der Pfad ihn nicht so nah an den Abgrund führt, daß er hinunterstürzen wird. Und doch wird er, indem er sich mit seinem Stock von Schritt zu Schritt weitertastet, dem Pfade folgen und weitergehen. Denn wenn es für ihn überhaupt eine Möglichkeit gibt, aus der Felswildnis heraus zu kommen, dann ist es das Tasten entlang diesem Pfad. Als Blinde stehen wir vor der Zukunft; aber wir tasten einen Pfad, und wir wissen: wenn wir überhaupt einen Weg durch die Zukunft finden können, dann geschieht es durch Tasten entlang diesem Pfad.

Vindication by optimality. Reichenbach's idea is evocative, but as it stands also "impossibly vague" (Salmon, 1967, 53). To start with, what is "the inductive method" supposed to be? Reichenbach did advance a particular rule of induction, that was motivated by his thoroughly probabilistic epistemology and his *frequentist* interpretation of probability, identifying probabilities with limiting relative frequencies. This prediction rule infers actual probabilities through *induction by enumeration*: it estimates the limiting relative frequencies by the *current* relative frequencies. This gives the most straightforward of rules of succession—indeed called the *straight rule* by Carnap—that issues predictive probabilities that are the observed relative frequencies,

$$\mathsf{p}(\boldsymbol{x}^t) = \left(\frac{\#_0\boldsymbol{x}^t}{t}, \frac{\#_1\boldsymbol{x}^t}{t} \right). \tag{2}$$

Now Reichenbach's argument is as follows. Either Nature's strategy is *uniform*—a limiting relative frequency exists—or it is not. If it is, then the inductive method (the straight rule) will be successful: it will converge on the correct limiting frequency. If it is not, if no limiting relative frequency exists, then obviously *no* rule will be successful in this sense. Hence the inductive rule is successful if *any* rule is.

Clearly, there are many weak spots in this argument. There is the sweeping reduction of scientific inference to estimating limiting relative frequencies (Sellars, 1964, 212ff; Skyrms, 1965, 254ff); there is the fact that there are infinitely many other rules that are likewise guaranteed to be successful in converging to an existing limiting relative frequency (Reichenbach, 1938, 113ff). There is the fact, too, that it is just not true that no method can be successful if there is no limiting relative frequency (Herz, 1936): sequences that never converge on a particular relative frequency of symbols might still be successfully—even perfectly—predicted by one or another method. To this last objection Reichenbach (1938, 358f) replied that a successful such alternative method p *has a high relative frequency of successful predictions*, which implies that his inductive method posits that p's predictions will continue to be accurate (by enumerative induction the *limiting* relative frequency of p's predictions being successful is inferred to be high). Reichenbach did not proceed to make this idea more precise, though: and indeed it does not seem feasible to reconcile the different levels at which his straight rule is now working—both the *object*-level of the data and the *meta*-level of other methods—in such a way that his method can be vindicated as desired (Skyrms, 1965, 260f; Skyrms, 2000, 44ff; Schurz, 2008, 281). Nevertheless, the idea of following those methods that have been successful so far is a very powerful idea, and as we will see this is how a method that is optimal relative to a pool of other methods operates. It is how a *universally* optimal method—that is vindicated as guaranteed to be successful whenever *any* method is—must operate.

Universal optimality. We define a method that is universally optimal as a method that will come to predict at least as successfully as *any* other prediction method, *no matter* what Nature does. More precisely, a universally optimal method is such that, *for any other prediction method*, the regrets *relative to this method* converge to 0 or—the universal method could do strictly better than this method!—less, always.

Consider again the mixture predictor over all computable hypotheses, introduced in I.4 above. We will now attempt to reinterpret this prediction method as a universally *optimal* method. This takes a couple of steps, that follow next: but the basic idea is that we reinterpret the class of computable hypotheses as the pool of computable prediction methods.

Reinterpretation: hypotheses and prediction methods. As I mentioned before in I.3 above, there is a formal correspondence between hypotheses (i.e., measures over the sequences) and prediction methods. (Roughly, the conditional probabilities of a measure define a prediction method and vice versa. See 3.1.) That means we can reinterpret an hypothesis as a prediction method, and a pool of hypotheses as a pool of competing prediction methods. Consequently, a constraint on Nature (an inductive assumption) can be reinterpreted as a constraint on possible prediction methods.

Reinterpretation: mixtures over prediction methods. In particular, we can reinterpet a mixture predictor over a pool of hypotheses as a mixture predictor that aggregates over a pool of predictors. On this reinterpretation, a prior or weight function gives weights to competing predictors rather than possible courses of Nature. (See 3.3.)

Relative optimality. Analogous to the case of universal reliability, we can show that this mixture predictor is optimal: for any predictor in the pool, its regrets relative to this predictor will converge to 0 or less, always. So, analogous to the case of reliability, we have uncovered the possibility of methods that exhibit 'relative universality' in the sense of optimality. (See again 3.3.)

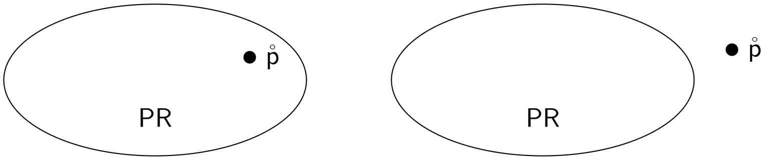
In the case of optimality, however, we appear within reach of *true* universality.

Computable prediction methods. Namely, while the conclusion of I.4 above was that the epistemic constraint of computability is an unwarranted metaphysical constraint on what Nature can do, it *does* appear a plausible constraint on our possible prediction methods. "It is reasonable to claim that any possible statistical analysis, formal or informal, must be computable" (Dawid, 1985b, 340; 1985a, 1260). Surely any prediction strategy we can ever devise must be implementable on a computer, hence captured by some algorithm, hence (Church-Turing thesis!) computable.

(What about the earlier point that in statistics uncountable hypothesis classes are the rule rather than the exception? That is true: but actual methods for manipulating these classes and generating samples or predictions from them will still be computable. The uniform mixture over the class of Bernoulli distributions, to give a simple example, is a computable object. Also see Freer and Roy, 2012.)

If we thus stipulate that all prediction methods must be computable, then the countable pool of computable prediction methods is the pool of *all* prediction methods.

Paul Oktopus. In this thesis I will take for granted that effective computability is a plausible constraint on possible prediction methods. It must be noted, though, that this claim still does not reduce to the original Church-Turing thesis—not without (at least) the further stipulation that a prediction method must be expressible as an explicit set of instructions to calculate its predictions, or not be a method at all. On the one hand, this seems plausible enough (again: surely any prediction strategy must be implementable on a



(a) Universal elements are included in the pool.

(b) Universal elements lie outside of the pool.

FIGURE 4. The pool PR of all prediction methods: two possibilities.

computer), on the other, it excludes the possibility of prediction methods that employ or are given by processes *in the world*. If the claim is to extend to such methods, it must again involve some variant of the physical Church-Turing thesis.

During the 2010 World Cup, many were following the predictions of a common octopus, Paul, who would identify the winning team of an upcoming soccer match by moving to one of two marked food containers in its tank. Is Paul a computable prediction method?

A universal element in the pool of all predictors. If we thus identify the pool of computable prediction methods with the pool of *all* predictors, then a universally optimal method relative to this pool (for instance, the mixture predictor over this pool) is a truly universally optimal method—at least, if it is still an actual method itself! Specifically, if we identify the pool of all predictors with the computable ones, we need the optimal method to be computable, too. Thus we need the pool of all predictors to actually *contain* its universal elements, as depicted in figure 4a.

Unfortunately, Putnam's diagonal argument shows that the situation is rather like figure 4b: any element that is universal for the pool of computable predictors, can no longer be a computable element itself.

Putnam's diagonal argument. Putnam's diagonal argument (1963a)—actually already a much simplified version of this argument (see Kelly, 2004 and 1.1) —shows the incompatability of two conditions on a prediction method:

(1) it is universal for all computable prediction methods;
(2) it is computable.

Suppose, for a contradiction, that we *do* have a computable prediction method, p_{U}, that is universal for all computable prediction methods. That means, in particular, that for every computable sequence that Nature plays (which corresponds to a computable *predictor* that predicts with certainty the symbols of this sequence), universal p_{U}'s losses (regret relative to this perfect predictor) should converge to 0. It now suffices to show that Nature can be adversarial against p_{U} in the extreme sense of making it fail at each trial, *and* that it can do this is in a computable way. But this is easy if p_{U} is computable:

at each trial $t+1$ simply compute and reveal the x_{t+1} with $\mathsf{p}_{\mathrm{U}}(x_{t+1}, \boldsymbol{x}^t) \leq 0.5$. This generates a computable sequence that p_{U} will never converge on.

There exists no universal algorithm. In his discussion of Reichenbach's attempted vindication of induction, van Fraassen (2000) points out a natural way of going beyond the simplistic straight rule that bears a resemblance to our story so far. His suggestion is that pursuing Reichenbach's idea must lead us to hope for a "'universal' forecast system" (ibid., 260) that is *computably calibrated* whenever any computable prediction method is. While this is a different notion from what we have discussed so far, the general idea and the demonstration of its impossibility follow the same lines, and I will now rehearse these as a summary of the foregoing. (Let me, nevertheless, briefly sketch this new notion. Calibration of method p on an infinite data sequence $\boldsymbol{x}^\omega$ means, roughly, that for each possible prediction p, the relative frequency of symbols in the infinite subsequence of $\boldsymbol{x}^\omega$ formed by taking those trials for which p issued a prediction close to p, is indeed close to the distribution p. *Computable* calibration, very roughly, demands a match, for all infinite subsequences $\boldsymbol{y}^\omega$ of $\boldsymbol{x}^\omega$ that are extracted by a computable *selection rule*, between the mean predictive probabilities given by p the relative frequency of symbols.)

The theory of (computable) calibration is discussed in detail by Dawid (1982; 1985a). He concludes the first paper with the conjecture—"doubtful, although not impossible"—of a single prediction method that is always calibrated; Oakes (1985) shows that this *is* impossible, using a simple adversarial sequence of the kind we saw earlier. Dawid, in a comment aptly titled "The Impossiblity of Inductive Inference" (1985b), notes the resemblance to Putnam's argument, and then remarks in quick succession that

- "there will always be data sequences that we are simply unable to track" (the impossibility of universal reliability); that
- "if we believe only that *some* computable distribution has given rise to the data, we cannot guarantee to obtain valid probability forecasts from *any* computable analysis" (the impossibility of a computable method that is universal relative to the computable hypotheses); and finally that
- there can be no computable method that is computably calibrated for any sequence *some* method is computably calibrated for (in effect, the impossibility of a computable method that is universally optimal relative to the computable methods).

Van Fraassen (2000, 260):

> The hope was formulated as a certainty by Reichenbach: with numerical induction properly tuned we have a method that will lead us to the truth *if any rule will* . . . Is this true as we have come to construe it? No, it is not.

Dawid again (1985b, 341):

> It would essentially solve the problem of (probabilistic) induction if we could construct a computable "universal algorithm" that, operating sequentially on any sequence of data for which such objective forecasts exist, would always output, asymptotically, these objectively valid forecasts. Unfortunately, the preceding arguments demonstrate that this ideal is unattainable: There exists no universal algorithm.

* * *

I.6. A Formal Theory of Inductive Inference

This is the title of Solomonoff's pioneering 1964 paper, that reports ideas he developed over the course of several years. (And that he already distributed occasionally in the form of technical reports; see, for instance, the discussion by Minsky, 1961, 27f.) In his paper Solomonoff proposes a number of different "models of induction," prediction methods that are intended to be "optimum with respect to all other conceivable models" (1964, 17) and that are constrained by computability.

Here I will present the most important of his proposed models as exploiting a basic strategy to avoid diagonalization. To that end, I will first turn to the diagonal argument due to Turing.

Turing's diagonal argument . . . The simplified version of Putnam's argument above really exposes a basic computability-theoretic fact, that goes back all the way to the founding paper of Turing (1936). This is the impossibility of a computable enumeration of the class of all computable functions. In our case, it is the impossibility of a computable enumeration of all computable measures (prediction methods). In other words, the pool of all computable hypotheses (prediction methods) is itself incomputable, which in particular implies that the mixture over this pool is incomputable.

Turing's original argument, in terms of computable real numbers (infinite sequences), is as follows (ibid., 246). If it *were* possible to computably enumerate all computable infinite sequences, then there would be a computable function $\mathring{g}$ such that $\mathring{g}(i,n) = \boldsymbol{x}_i^\omega(n)$ for all $i, n \in \mathbb{N}$, with $\{\boldsymbol{x}_i^\omega\}_{i\in\mathbb{N}}$ containing all and only computable sequences. But then we could define a computable infinite sequence $\boldsymbol{y}^\omega$ by

$$\boldsymbol{y}^\omega(n) := \begin{cases} 1 & \text{if } \mathring{g}(n,n) = 0; \\ 0 & \text{otherwise ,} \end{cases} \tag{3}$$

a computable diagonal sequence that differs from each computable sequence in the assumed exhaustive list because for each $i \in \mathbb{N}$,

$$\boldsymbol{x}_i^\omega(i) = \mathring{g}(i,i) \neq \boldsymbol{y}^\omega(i).$$

It is a computable version of Cantor's ur-diagonal argument (1891), that showed that there must be more than enumerably many real numbers. But the similar conclusion—that there must be uncountably many computable numbers—is not an option for us here, because we know there are only countably many Turing machines: hence we must reject the assumption that the computable sequences can be enumerated *computably*.

. . . and the way out. Computable infinite sequences correspond to computable functions $f : \mathbb{N} \rightarrow \mathbb{B}$ that are *total*, or defined on the whole domain $\mathbb{N}$. Turing's argument thus shows that there can be no computable enumeration of total computable (t.c.) functions: the existence of a *universal* t.c. function $\mathring{g} : i, n \mapsto g_i(n)$ that can emulate any other t.c. function g_i would allow us to define a t.c. diagonal function as in (3).

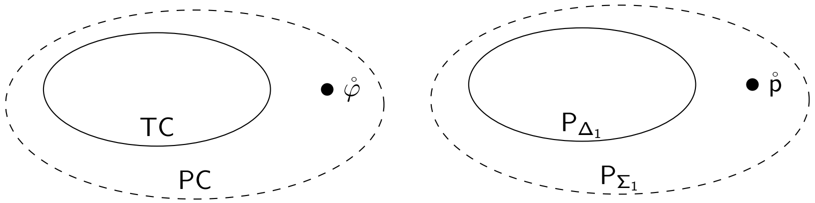
However, "[*w*]*e can avoid the diagonalization difficulty by allowing sets of instructions for nontotal partial functions as well as for total functions*" (Rogers, 1967, 11; also see Soare, 2016, 4f; Odifreddi, 1989, 145ff). Indeed, the functions given by Turing machines need not be total. A Turing machine might on some input values get stuck in an infinite loop and never halt, meaning that the corresponding function is only *partially* defined. "It is more natural to consider partial computable functions anyway, because . . . certain algorithms may be naturally defined only on *some* but not all arguments" (Soare, 2016, 4). (The partial computable functions correspond to a genuinely richer class of algorithms: they do not all equal some t.c. function with restricted domain, Kleene, 1938, 151.) And the partial computable (p.c.) functions *can* be effectively enumerated: the class of p.c. functions is immune to diagonalization.

To see why this is so, let us try to diagonalize an effective enumeration $\{\varphi_i\}_{i \in \mathbb{N}}$ of the p.c. functions. We define again the diagonal function

$$\hat{\varphi}(n) := \begin{cases} 1 & \text{if } \varphi_n(n) \downarrow = 0; \\ 0 & \text{if } \varphi_n(n) \downarrow = 1, \end{cases}$$

where the notation '$\varphi_n(n) \downarrow = 0$' means '$\varphi_n(n)$ converges with output 0.' For those n on which $\varphi_n(n)$ is undefined, our function $\hat{\varphi}$ is thus undefined, too: it is itself a p.c. function. But this property blocks the crucial step from any given index i to the observation that $\hat{\varphi}$ cannot equal ϕ_i because they differ on input i: they might actually both diverge on input i! (Nor can we avoid this by first checking whether $\varphi_n(n)$ is in fact defined: the *Halting problem* is undecidable.) "Thus *the notion of partial recursive function seems to have a built-in defense against diagonalization*" (Odifreddi, 1989, 152).

There exists a universal algorithm. What happened is that we escaped diagonalization by going to a larger class that is not diagonalizable. The class of t.c. functions is diagonalizable, so it cannot contain universal elements; the larger class of p.c. functions is not diagonalizable and it does contain universal elements (figure 5a). Such a universal element, a p.c. function $\mathring{\varphi} : i, n \mapsto \phi_i(n)$ that emulates every other p.c. function, is the function that corresponds to a



(a) From the class of t.c. functions to the class of p.c. functions.

(b) From the class of computable measures to the pool of semi-computable measures.

FIGURE 5. From a diagonalizable class of total computable objects to a nondiagonalizable class of partially computable objects.

universal Turing machine. Turing's fundamental observation is that there do exist universal algorithms: those given by universal Turing machines, computing universal p.c. functions.

What about algorithms for prediction?

Enter algorithmic information theory. A function $\mathsf{p} : \mathbb{B}^* \to [0,1]$ that is only partially defined, though, does not seem a very viable prediction method. How do we make sense of a method that at some nodes might not return predictions at all? (Also see Kelly et al., 1994, 104.)

However, when we pursue the analogous extension of the computable *measures* (figure 5b; throughout the thesis I also refer to these as the 'Δ_1 measures,' see 2.1.1), the objects we obtain are measures ('Σ_1 measures,' see 2.1.2) that are actually still defined on every node of the tree. These measures then give (via the equivalence between measures and prediction methods mentioned in I.5 above) prediction methods that *are* defined on every node. What we lose is the guarantee that we can compute these measures' values to any accuracy: we are only able to compute increasingly accurate lower bounds. This 'lower semi-computability' superficially looks less detrimental to the notion of a prediction method than sheer undefinedness. And, crucially, the mixtures over all elements in this class are still within the class: universal elements that we then interpret as giving universal prediction methods.

This is, anyway, how we must interpret the universal prediction methods that arise from the work of Solomonoff (1964), that initiated the field of *algorithmic information theory* (the standard textbook is Li and Vitányi, 2008).

Solomonoff. As mentioned above, Solomonoff (1964) described a number of different "models" for prediction. One of those models actually comes down again to the mixture over computable measures ("probability evaluation methods," ibid., 19ff), which of course "is not 'effectively computable' (in the sense of Turing (193[6)]) and so it does not include itself in the summation" (ibid., 21).

The familiar general idea here—defining universal methods for sequential prediction by a mixture over hypotheses or prediction methods—is an idea that extends to modern work in machine learning, where Solomonoff still receives credit for it (e.g., Vovk, 1998, 167; Ryabko, 2010, 583). Still, it is an idea that from various angles one arrives at relatively easily, and that perhaps would not be associated with Solomonoff were it not for the impact of his *other* proposed models. (As a matter of historical interest, I note here the paper by Howard (1975), that independently develops what amounts to the idea of Bayesian prediction with computable measures, but that stops at the observation of the impossibility of a universal computable mixture.)

Solomonoff's other proposed models are given by measures that are directly defined in terms of universal Turing machines. (This corresponds to the second definition I described in the introduction, page 6.) Specifically (ibid., 3),

> A priori probabilities are assigned to strings of symbols by examining the manner in which these strings might be produced by a universal Turing machine. Strings with short and/or numerous "descriptions" (a "description" of a string being an input to the machine that yields that string as output) are assigned high a priori probabilities. Strings with long, and/or few descriptions are assigned small a priori probabilities.

The intuition here—an intuition inspired by Shannon's *information theory*—is that sequences that are easier to describe (easier to *code*) should be more likely. Solomonoff's fundamental step is the idea to identify the descriptional complexity of a sequence with the minimal required length of input to a universal Turing machine to produce the sequence, *plus* the realization (ibid., 11ff) that this notion is to a certain extent independent of the particular choice of universal Turing machine. This is the founding idea of the field of algorithmic information theory (see Li and Vitányi, 2008, 95ff, 192).

However, Solomonoff's presentation suffers from a certain lack of rigor (also see Bienvenu et al., 2009, 17), which left the task to others to make the latter idea—and its interplay with the idea of universal mixtures—perfectly precise.

Kolmogorov. The founding idea of algorithmic information theory—in fact, the field itself—is best known under the header of *Kolmogorov complexity*, because it was independently arrived at by Kolmogorov (1965). His motivation was an entirely different one: the formalization of *randomness*.

The problem of randomness goes back to von Mises's attempt to base probability theory on a frequentist interpretation. He *defined* the probability of an outcome as the limiting relative frequency of the same type of outcome in an unending series of trials. This he modeled by a *Kollektiv*, an infinite *random sequence* (say an element $\boldsymbol{x}^\omega \in \mathbb{B}^\omega$, in case of two types of possible outcomes), that satisfies two properties corresponding to the empirical laws of chance processes: (1) it has a limiting relative frequency of 0's and 1's and (2) the *Prinzip vom ausgeschlossenen Spielsystem*: there is no betting strategy on the successive outcomes that is guaranteed to make unbounded gains. The latter was formalized in the requirement that no *admissable selection rule* (corresponding to a betting strategy that picks out certain trials to put money on) would give an infinite subsequence with a *different* limiting relative frequency. But when is a selection rule 'admissable'?—von Mises was happy to leave this an 'intensional' notion to be further specified in the particular empirical situation at hand (see Van Lambalgen, 1987a, 29ff), but many others saw here a serious problem. Church (1940) proposed to identify admissable with *computable*; but by that time Ville (1936; 1939) had already shown that the definition of selection rule is too weak to capture all feasible betting strategies, and von Mises's program had largely been deserted, anyway, with Kolmogorov's (1933) introduction of the measure-theoretic axiomatization of probability (see Van Lambalgen, 1987a,b; Bienvenu et al., 2009). Slightly ironically so, because Kolmogorov himself adhered to a frequency interpretation (1963, 369):

> I have already expressed the view ... that the basis for the applicability of the results of the mathematical theory of probability to real 'random phenomena' must depend on some form of the frequency concept of probability, the unavoidable nature of which has been established by von Mises in a spirited manner.

He long believed, however, that the frequency interpretation cannot escape the dilemma that a concept based on limiting relative frequency "does not contribute anything to substantiate the applicability of the results of probability theory to real practical problems where we have always to deal with a finite number of trials," whereas a *finite* frequentism "does not admit a rigorous formal exposition within the framework of pure mathematics" (ibid.). In the 1960's he changed his mind on the latter point: "I have come to realize that the concept of random distribution of a property in a large finite population can have a strict formal mathematical exposition." Roughly, a random finite sequence is random if *suffciently simple* selection rules or *algorithms* (of which as a combinatorial fact there cannot be many) leave the relative frequencies almost intact. "Such a conception in its full development requires the introduction of a measure of the complexity of the algorithm" (ibid.). In (1965, 7), Kolmogorov proposes to identify this with the complexity notion he introduced there: the minimal required length of input to a universal p.c. function, the same idea as Solomonoff's.

What proved to be the most influential strand of work building on Kolmogorov's idea is (again somewhat ironically) the theory of random *infinite* sequences. Martin-Löf (1966) gave a first characterization of infinite random sequences in terms of effective *statistical tests*. Schnorr (1971a; 1971b) gave alternative characterizations in terms of effective betting strategies or Ville's notion of *martingales*. Schnorr (1973), and independently Kolmogorov's student Levin (1973), also first provided the appropriate *complexity* concept to characterize random infinite sequences. The computability-theoretic properties and interactions of different formal notions of randomness—notions that can often be equivalently defined via any of these three 'paradigms'—are the subject of the modern theory of *algorithmic randomness* (see the textbooks Nies, 2009; Downey and Hirschfeldt, 2010).

Levin. Kolmogorov's idea also spawned a tradition of work by Russian mathematicians on descriptional complexity. (A new textbook on algorithmic information theory from authors in that tradition is forthcoming in translation, Shen et al., 20xx.) An important overview of early results of the Russian school is the joint paper by Zvonkin and Levin (1970), that is based for a significant part on results from Levin's thesis (translated as Levin, 2010).

Among these is the introduction of the notion of Σ_1 measure, which makes precise Solomonoff's above idea of associating probabilities with the lengths of a machine's inputs. Thus Levin makes precise the class depicted in figure 5b, including the universal elements that he constructs as mixtures over all elements in the class. (This corresponds to the first definition I described in the introduction, page 6.) I will therefore call these universal elements (following Li and Vitányi, 1989, 172; 1992b, 356) the *Solomonoff-Levin measures.*

This thesis (1). The first main strand in this thesis, the topic of part II, is the interpretation of Solomonoff's proposal as a theory of universal prediction: in particular, the Reichenbachian interpretation of the Solomonoff-Levin predictors as optimal among all possible predictors.

I presented the Solomonoff-Levin measures as arising from an explicit attempt to identify a class of effective elements that is immune to diagonalization, the procedure of Putnam's simplified argument. The question is whether this nondiagonalizable class and its universal elements, depicted in figure 5b, is indeed susceptible to the desired interpretation.

The conclusion is negative: this interpretation does not work, and the main reason is a mismatch between the level of effectiveness of the Solomonoff-Levin *measures* and the Solomonoff-Levin *predictors.* The latter are *not* susceptible to the desired interpretation, a fact that turns out to be exposed by Putnam's original and more complex diagonal argument.

* * *

I.7. The Use of Simplicity in Induction

This is the title of the famous paper by Kemeny (1953), in which he discusses the preference in inductive reasoning for *simple* hypotheses. In particular, in the problem setting of identifying within a class of hypotheses the correct one, he formulates "what any scientist would do" (ibid., 396) as the rule:

> Select a hypothesis which is as well in agreement with the observed values as possible; if there is any choice left, choose the simplest possible hypothesis.

Choosing the simplest of options is certainly the standard way of breaking the stalemate of Goodman's riddle that there are always infinitely many ways of generalizing from finite observation data: both Poincaré (1902) and Jeffreys (1939), for example, took this route (also see Watkins, 1984, 105ff). But Kemeny mentions two immediate problems for this idea, problems that are still very much unresolved today: how exactly must we *define* the notion of simplicity, and how can we actually *justify* the use of simplicity in induction?

Part III of this thesis is concerned with the relevance of the Solomonoff-Levin proposal to these problems. In this section I provide some more context to the problems of giving a formalization and a justification for the principle of preferring simplicity, the principle of Occam's razor, and then outline how Solomonoff's ideas have been cast as providing these.

'Occam's razor.' The principle of scientific method to prefer simplicity is commonly referred to as *Occam's razor*. The connection to the scholastic philosopher is supposed to be that Ockham (latinized 'Occam') first used a principle of parsimony—"don't multiply entities beyond necessity"—to defend a nominalist position in the medieval debate about universalia, stating that universals (general properties, like being human, red, . . .) are just names that do not have a real existence. The actual context is, unsurprisingly, more subtle (see, e.g., Spade, 1999); for one thing, Occam and his contemporaries understood the principle of parsimony he invoked to be well-established and going back at least to Aristotle. But more importantly, observe that already in this basic guise of advocating a minimal ontology, it is not so clear what the razor principle actually achieves. No one would insist on positing unneccesary entities—that is what makes them unnecessary (equally sensible is an *anti-razor* that tells us not to posit *less* entities than necessary). The pertinent issue is, of course, *what* is unnecessary. Left open, moreover, is how to proceed with those unneccesary entities: does the principle tell us to simply refrain from claiming they exist (the "agnostic" interpretation), or should we assert that they do *not* exist (the "atheistic" interpretation, terminology Sober, 1981)? The first instruction is again quite trivial; the second almost sounds like Leibniz's principle of sufficient reason ('nothing can exist without a reason')—a rather heavy metaphysical commitment.

As a modern principle of scientific methodology, Occam's razor has acquired a broader meaning than purely relating to the parsimony of entities (objects, quantities, parameters, . . .) in our theories. It is taken to express a general preference for simplicity of theory and explanation, what is often labeled as a preference for *elegance*. This includes, for instance, the intuition advocated by Einstein that the physicist should strive for the simplest possible mathematical description of the world. But in this guise, too, the problem with Occam's razor is that as soon as we try to make it more precise, it starts looking either quite trivial, or way too strong.

Occam's razor as a philosophical problem. As a *pragmatic* principle, presenting simplicity as "only a matter of convenience, a laborsaving device" (Kemeny, 1953, 391), Occam's razor does not say much: of course we would prefer to work with simpler theories, other things being equal. A stronger expression of this pragmatic view is that the scientific enterprise is *per definition* the search for maximally economical descriptions of nature. Still, while this conception of science (that one can also find expressed by Poincaré, 1902, 156) could certainly be *explained* from our "biological craving for simple answers" (Hoffmann et al., 1996, 123), it is still lacking in epistemic *justification.* The interesting problem is to find justication for Occam's razor as an *epistemic* principle, where it comes with the promise of a relation between a simplicity preference and attaining the truth. The threat here is that, in this guise, the principle soon commits us to an unwarranted assumption that the world must in some sense be simple. (Einstein indeed took the position that "nature is the realisation of the simplest conceivable mathematical ideas," see Norton, 2000.) The challenge, then, is to provide a *justification* for the strong, epistemic version of Occam's razor: to show that we have epistemic grounds for a simplicity preference, *while avoiding* any simplicity assumption on the world.

But before that we would still have to address the problem of what simplicity actually *is.* Can we measure simplicity? Is it possible at all to compare theories on their simplicity in an objective fashion? Or is simplicity in the end an inherently vague or even subjective notion?

Occam's razor in statistics. Many philosophers have looked for a response to these challenges in probability theory and statistics (see Sober, 2015; Gauch, 2003). A central theme in statistics as well as machine learning is the trade-off between simplicity and goodness-of-fit of an hypothesis or a *model* (a class of hypotheses), and discussions of Occam's razor have centered on the lessons that may be drawn from various modern approaches in model selection. Different such approaches give different formalizations of the trade-off, and in particular of a model's simplicity; but they tend to share two features. First, there is some plausible notion of a model's simplicity as its size or rather its *richness* (two main examples of such complexity measures are the *Vapnik-Chervonenkis* or *VC dimension* in statistical learning theory, 1971, also see Vapnik, 1998; Harman and Kulkarni, 2007; and the *stochastic complexity* in minimum description length inference, Rissanen, 1986, 1987, also see Grünwald, 2007). Second, there is the support from formal and empirical results that are to show their good performance. The combination of these two ingredients yields the prospect of an honest justification of a simplicity preference: a demonstration that preferring simplicity leads to good results, without an assumption that the world must be simple.

I will not discuss here whether instantiations of this strategy are indeed successful; my purpose here is to set this strategy apart from the approach to Occam's razor in algorithmic information theory. While a justification of Occam's razor must have the same general form, there is an important respect in which the approach in algorithmic information theory is fundamentally different from the above.

Occam's razor in algorithmic information theory. Namely, rather than models, classes of objects, we are promised a general and objective quantification of simplicity of *individual objects*. The idea is that a data object, like the specification of a hypothesis, is simpler as it is more *compressible*, meaning that we can capture it in a shorter description. This idea is made formally precise in the definition of a data object's Kolmogorov complexity as the length of its shortest description (Li and Vitányi, 2008, 260):

> This gives an objective and absolute definition of 'simplicity' as 'low Kolmogorov complexity.' Consequently, one obtains an objective and absolute version of the classic maxim of William of Ockham.

As I noted earlier, the first published variant of Kolmogorov complexity to appear in the literature is Solomonoff's (1964) description of the Solomonoff-Levin measure. From the start this measure and the associated universal prediction method has been associated with a simplicity preference (ibid., 3):

> That [this definition] might be valid is suggested by "Occam's razor," one interpretation of which is that the more "simple" or "economical" of several hypotheses is the more likely. Turing machines are then used to explicate the concepts of "simplicity" or "economy"—the most "simple" hypothesis being that with the shortest "description."

Thus we appear to have a precise definition of simplicity, as well as a prediction method that implements a simplicity bias—implements Occam's razor—using this definition. What is more, the provable *universality* of the Solomonoff-Levin prediction method translates into a strong property of reliability or truth-convergence. Together, this seems to lead to a precise link between a preference for simplicity and finding the truth. It seems to meet the above goal of providing epistemic grounds for a simplicity preference *without* metaphysical simplicity assumptions (also see Grünwald and Vitányi, 2008, 314). In other words, it suggests a justification for the epistemic version of Occam's razor (Vitányi and Li, 2002, 154):

> This validates by mathematical proof a rigorous formal version of Occam's razor – the ancient simplicity-based method to infer the true cause of the data.

Predictive complexity. Solomonoff's ideas also have a direct link to the modern branch of theoretical machine learning that goes by the name of *prediction with expert advice* (founding papers are Littlestone and Warmuth, 1994; Cesa-Bianchi et al., 1997; Vovk, 1990, 1998; the standard textbook on the subject is Cesa-Bianchi and Lugosi, 2006). The object of study here is the design of prediction strategies that are optimal relative to a given pool of prediction methods, for a given loss function.

Of major importance is Vovk's specification of the *aggregating algorithm*, that generalizes the Bayesian mixture strategy and its optimality to loss functions other than the logarithmic loss. The aggregating algorithm for the log-loss function applied to the pool of Σ_1 measures in fact corresponds again to the Solomonoff-Levin predictor.

A further idea due to Vovk (1998; 2001b) is the following. Another and more direct way of associating a notion of simplicity with a Solomonoff-Levin predictor is to interpret the logarithmic loss it suffers on a given sequence $\boldsymbol{x}$ as a measure of the intrinsic difficulty of predicting $\boldsymbol{x}$. This measure is Vovk's *predictive complexity* in the particular case of the log-loss function; and this notion again generalizes via the aggregating algorithm to other loss functions. Vovk presents this as a measure of the intrinsic complexity of sequences (relative to a given loss function).

This thesis (2). The second main strand in this thesis, the topic of part III, is the association of the Solomonoff-Levin proposal with the elusive concept of simplicity: in particular, the idea that the Solomonoff-Levin predictors do not only consitute a *formalization* of a simplicity preference, the principle of Occam's razor, but also provide a *justification* for it.

I already briefly indicated the relevant notion of simplicity as compressibility, and the shape of the suggested epistemic justification of a simplicity preference that must avoid simplicity assumptions on the world. I will spell this out in much more detail, and then address the question whether this justification and indeed this formalization is convincing. Subsequently, I will investigate Vovk's notion of predictive complexity as an intrinsic notion of complexity of sequences.

My conclusions are again negative. The suggested justification does not work, precisely because the relevant simplicity preference constitutes a particular inductive assumption. In addition, I argue that the relevant definition of simplicity as compressibility does not convincingly lead to an objective formalization of a simplicity preference in prediction. While, moreover, the notion of predictive complexity has a more direct and therefore ostensibly less problematic interpretation, in the end it does not deliver on its promise.

*

Part II

Universality

CHAPTER 1

Confirmation and computation

This chapter stages Solomonoff's theory of universal prediction as an offspring of Carnap's program of inductive logic. Solomonoff's outlook bears a strong resemblance, especially, to the way Carnap's inductive logic was presented by Putnam, in order to subject it to his diagonal argument. The discussion in this chapter thus raises the question that drives the subsequent chapters: can Solomonoff's proposed universal prediction method avoid Putnam's argument, and how could it?

In 1.1, I introduce Putnam's diagonal argument and its motivation. In 1.2, I discuss in some detail Carnap's program of inductive logic. In 1.3, I introduce Solomonoff's ideas and relate these to the points of contention between Putnam and Carnap.

Innovations. While Putnam's argument has found a place in philosophers' collective memory, there actually seem to exist few in-depth evaluations of the reasons why Putnam believed his proof should spell the end of Carnap's program and why Carnap disagreed. Section 1.1 is mainly a summary of Putnam's original papers, but sections 1.2 and 1.3 include a more critical exposition of what I think are the main aspects of Carnap's program that conflict with the way Putnam sought to present it. The technical account in 1.2.2 and 1.2.3 of the Johnson-Carnap functions in the setting of binary sequential prediction draws directly from Carnap's original writings and secondary sources (particularly, Zabell, 2011 and Suppes, 2002). A main contribution of this thesis is a detailed positioning of Solomonoff's proposal relative to Carnap's program and Putnam's view, as initiated in section 1.3. (This chapter is based on part of Sterkenburg, 201x.)

1.1. Putnam's diagonal argument

Putnam, in his contribution to the volume of *The Library of Living Philosophers* devoted to Carnap (Schilpp, 1963), declares that Carnap had better give up his program of inductive logic (1963a, 761, 778). Putnam's reasoning goes beyond "intuitive considerations and plausible argument" (ibid., 761): he offers a *mathematical proof* that Carnap's objective is a formal impossibility.

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 $\neg Gx_i$. A Carnapian *confirmation function* $\mathfrak{c}$ gives the degree of confirmation—the logical probability—that one statement confers upon another. In particular,

$$\mathfrak{c}(h(x_{t+1}), h(x_0) \ \& \ \ldots \ \& \ h(x_t))$$

is the degree to which the statement that the next individual x_{t+1} satisfies h is confirmed by the fact that all of x_0 up to x_t do so already. (Carnap also calls this the *instance confirmation* of h.) Now, if a given Carnapian confirmation 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 confirmation function. Hence a condition of adequacy on such a confirmation function $\mathfrak{c}$ is that

(I) For any computable hypothesis h, the value for the instance confirmation $\mathfrak{c}(h(x_{t+1}), h(x_0) \ \& \ \ldots \ \& \ h(x_t))$ should converge to 1 as we observe a longer and longer succession of confirming individuals $x_0, \ldots, x_t$.

But for any confirmation function $\mathfrak{c}$ that itself satisfies a weak condition of effective computability (to not be "of no use to anybody," ibid., 768):

(II) For every t, it must be possible to compute an s such that if G holds for the next s individuals $x_{t+1}, \ldots, x_{t+s}$, then the instance confirmation $\mathfrak{c}\,(Gx_{t+s+1}, Gx_{t+1} \ \& \ \ldots \ \& \ G(x_{t+s}))$ exceeds 0.5,

one can prove by diagonalization $\mathfrak{c}$'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 Carnapian confirmation function.

Let me simplify things a little. (I return to the details of the original argument in chapter 4.) We can treat condition (I) as an instance of the condition on an 'inductive method' M, a condition left somewhat informal in its generality, that

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

Moreover, in later expositions of the argument (e.g., Earman, 1992, 207ff; Kelly, 2004, 701f), the slightly cumbersome condition (II) is often replaced by the (stronger) condition that $\mathfrak{c}$ 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 confirmation functions is straightforward (also see I.5 above). Given candidate computable confirmation function $\mathfrak{c}$, we construct a computable hypothesis h such that $\mathfrak{c}$ fails to converge on h, as follows. Starting with the first individual x_0, compute $\mathfrak{c}(Gx_0)$ and let $h(x_0)$ be $\neg Gx_0$ precisely if $\mathfrak{c}(Gx_0) > 0.5$. For each new individual x_{t+1}, proceed in the same fashion: compute $\mathfrak{c}(Gx_{t+1} \mid h(x_0), \ldots, h(x_t))$ and let $h(x_{t+1})$ be $\neg Gx_{t+1}$ precisely if this probability is greater than 0.5. The hypothesis h is clearly computable, but by construction the instance confirmation given by $\mathfrak{c}$ does not converge to 1: 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* is to be computable itself, then it is impossible to reconstruct it as a confirmation 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 confirmation function and that *does* satisfy the two requirements. This method HD 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 HD will eventually come to (and forever remain to) accept it.

The distinctive feature of HD 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 (ibid., 780):

> Certainly it appears implausible to say that there is a *rule* whereby one can go from the observational facts (if one only had them all written out) 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"; i.e. defined in such a way as to agree with the actual inductive judgements of good and careful scientists.

Incredulously (ibid., 781):

> 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).

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, 192ff), though his reservations mainly concern the possibility of mechanized formulation of hypotheses based on observation data (ibid., 193):

> I am completely in agreement that an inductive machine of *this* kind is not possible. However, I think we must be careful not to draw too far-reaching negative consequences from this fact. I do not believe that this fact excludes the possibility of a system of

> inductive logic with exact rules or the possibility of an inductive machine with a different, more limited, aim.

Such a more limited aim is what Putnam is after: that the values given by a confirmation function can be computed. But this Carnap actually also rejects: "$\mathfrak{c}$ is, in general, not a computable function" (ibid., 196), basically because inductive logic contains deductive logic and the latter is already undecidable for a sufficiently rich language. He then does, however, express confidence that $\mathfrak{c}$ is computable for "restricted classes of cases" (ibid.); a sentiment that returns in the passage from (1966) that I cited in the introduction, page 4: "I believe it is in many cases possible to determine, by mechanical procedures, the logical probability, or degree of confirmation, of h on the basis of e."

Carnap's reservations notwithstanding, Putnam, in his *Radio Free Europe* address (1963b, 297), 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.

Moreover (ibid., 298),

> 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.

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).

* * *

1.2. Carnap's inductive logic

This section discusses Carnap's program of inductive logic. The focus is on the formal framework of Carnap's early inductive logic, as it stood around the time of publication of Putnam's argument and Solomonoff's paper. This is in essence the framework Carnap described in the volume *Logical Foundations of Probability* (1950) and the booklet *The Continuum of Inductive Methods* (1952); and which he updated in the 1950's (see the 1955 lecture notes published as Carnap, 1973) to the system that is reported in the volume *Induktive Logik und Wahrscheinlichkeit* (Carnap and Stegmüller, 1959) and summarized in the Schilpp volume that contains Putnam's argument (Carnap, 1963a, 966ff).

(Putnam's paper, while only published in 1963, was actually already written sometime in the mid-1950's (see Carnap, 1958[5]; 1963a, 988); and apparently with knowledge of the updated system (see ibid., 974). Solomonoff in (1964) only cites the *Logical Foundations*.)

In 1.2.1, I introduce Carnap's program and its philosophical motivation. In 1.2.2, I describe the formal framework of his early inductive logic, and make the translation to our setting of binary sequential prediction. In 1.2.3, I specify the *Johnson-Carnap* confirmation functions for sequential prediction.

1.2.1. The program of inductive logic.

1.2.1.1. *The logical interpretation of probability.* The evidence e that all ravens we have seen so far have two wings certainly does not logically *entail* the hypothesis h that the next raven we will spot has two wings, too. Still—or so starts the logical approach to probability—e does *support* or *confirm* the hypothesis h to some extent: the evidence *partially* entails the hypothesis. This is logical probability: degree of partial entailment. In this sense, probability theory is an extension of logic. Probabilities pertain to nonempirical, *a priori* relations between statements; and because they pertain to relations—the extent to which h is confirmed *by* e—all probabilities are *conditional* (though see 1.2.2.3 below).

1.2.1.2. *Rational degrees of belief.* Carnap, like Keynes before him, further identifies logical probability with *rational degree of belief.* Probabilities are an agent's degrees of belief, but in a strictly normative sense: the *objective* degrees of belief an ideal agent *ought* to have, to count as perfectly rational. In the words of Keynes (1921, 4, quoted with assent by Carnap in 1950, 43),

> ... in the sense important to logic, probability is not subjective. It is not, that is to say, subject to human caprice. A proposition is not probable because we think so. When once the facts are given that determine our knowledge, what is probable or improbable in the circumstances has been fixed objectively, and is independent of our opinion. The Theory of Probability is logical, therefore, because it is concerned with the degree of belief which it is *rational* to entertain in given conditions and not merely with the actual beliefs of particular individuals, which may or may not be rational.

(The identification of logical probability and rational degree of belief is not indisputable: one can argue, for instance, that no finite evidence ('all ravens so far ... ') even partially entails an infinite hypothesis ('*all* ravens ... '), whereas it is perfectly rational to believe certain infinite hypotheses to some extent; also see Gillies, 2000, 30f.) Carnap later (1962b; 1963d, 67f; 1963a, 967ff) explicitly endorsed the explication of logical probability as degree of belief, including the operationalization in terms of fair betting quotients. This had the additional advantage of making clear why logical probability should be *probability* (the *Dutch book* argument that *incoherent* degrees of belief, i.e., that are in violation of the probability axioms, are sufficient for irrationally accepting bets that are guaranteed to lose you money); and of naturally tying in with a decision theory (see Carnap, 1971a), in accordance with the view that "probability is a guide in life" (1947a). But how do we determine the exact values of these logical probabilities?

1.2.1.3. *The principle of indifference.* The logical approach to probability is in some important respects a successor of the *classical* interpretation of probability, that is mainly associated with Laplace and that is roughly given by the following two tenets. First, in the light of Laplace's doctrine of universal determinism, probability can only be an epistemic notion. Second, the way probabilities are calculated—indeed, probability is *defined*—is by the *principle of non-sufficient reason*: absent reason to think different outcomes are not equally possible, they have equal probability. This general definition is problematic for multiple reasons. If it is not circular (what does equally possible mean?) then it still magically infers knowledge from ignorance; it is inapplicable if we *do* have reason to think one outcome is more likely than the next; and it is *inconsistent* in the case of infinitely many outcomes, as revealed by a whole family of paradoxes (see Salmon, 1967, 66ff; Suppes, 2002, 163ff). Keynes, in fact, gave an authorative overview of these paradoxes (1921, 42ff); yet he still adopted much the same principle for his logical approach. He recast equipossibility from ignorance as the informed decision to treat cases as symmetrical, and rebranded it the principle of *indifference* (see Galavotti, 2005, 148f). Carnap, as we will see in detail in 1.2.3 below, likewise employed the principle of indifference to infer equal probabilities, again not from ignorance, but from the logical symmetry of cases (Carnap, 1953, 193f). Thus the principle of indifference is a cornerstone of the logical interpretation of probability, but because of its controversial status at once a main liability (see for instance van Fraassen, 1989, 293: "the story [of the principle of indifference and its problems] is especially important for philosophy, because it shows the impossibility of the ideal of logical probability.").

1.2.1.4. *Intuitions and axioms.* For Keynes, logical probabilities have an objective existence, in a Platonic sense: they exist even if we cannot know them (see Gillies, 2000, 31ff). Insofar as we can come to know these logical relations at all, we do so by "immediate logical intuition." However, some relations are more immediately perceivable than others, which would allow us to enshrine a number of such obvious relations as axioms from which the less obvious ones can be derived. This is to parallel the procedure in mathematical logic; but in that case already mathematicians' intuitions differ, and in the case of logical probability our intuitions—how exactly do we apply indifference, and indeed: *why*?—are much less clear still. Ramsey (1931, 65f) put it devastatingly simple when he criticized Keynes's reliance on this immediate intuitive grasp by noting that "there really do not seem to be any such things as the probability relations he describes":

> He supposes that, at any rate in certain cases, they can be perceived; but speaking for myself I feel confident this is not true. I do not perceive them, and if I am to be persuaded that they exist it must be by argument; morever, I shrewdly suspect that others do not perceive them either, because they are able to come

> to so little agreement as to which of them relates any two given propositions.

1.2.1.5. *Carnap's program.* For Carnap, establishing the logical probablity relations, in particular, establishing the axioms from which they follow, is part of an ongoing research program. Starting from the simplest of cases, expressed in the simplest of formal languages, we put down axioms and investigate their consequences, always checking whether they are sufficiently in line with—but at the same time having them inform and sharpen!—our inductive intuitions and statistical practice (cf. Jeffrey, 1973). This is Carnap's program of inductive logic.

1.2.1.6. *The formal language.* I have said little about the 'language' of binary symbols in sequential prediction. Methods for prediction as defined in this thesis operate on sequences of symbols in a purely *syntactical* way, in the sense that predictions do not depend on what the symbols actually refer to. But of course this ignores the fact that, unless the data in a particular situation already come to us in well-differentiated discrete chunks, there is the choice of how we put the sensations coming to us into such form (how we 'carve up the Humean mosaic'): and our predictions inevitably depend on how we did *that.* Carnap's inductive logic also very much *appears* purely syntactical in the sense that the confirmation values given by the axioms depend only on the formal properties of the given language (and since the Continuum, an additional choice of parameters). But things are much more subtle here, due to the fact that for Carnap, the formulation of the proper language in each situation is a crucial consideration. His reply to Goodman's new riddle (1946), for instance, is essentially that the choice of the right "qualitative" primitive predicates will determine (indeed, will *reveal*) what properties are projectible (1947b, 146ff; 1948). In later works one can even find suggestions that the adoption of certain axioms must depend on the meaning of the predicates (see Jeffrey, 1966, who also sets aside Hempel's "purely syntactical theory" of (qualitative) confirmation (1943) from Carnap's "more semantical one," 282). This is something to keep in mind when I discuss the formal framework and the syntactical axioms that define the Johnson-Carnap methods for sequential prediction in 1.2.2 and 1.2.3 below.

1.2.1.7. *The provisionary nature of the program.* Of relevance here, too, is Carnap's repeated insistence on the *provisionary* nature of the results in his program: the system as its stands is always only preliminary and bound to be superseded by more refined axioms for more complex languages. As Jeffrey (1972, 633) puts it:

> Carnap's goal was the definition of an adequate *c*-function for a very rich language in which one can discuss matters so diverse as theoretical physics and tomorrow's dinner. He seems to have believed that until the system either nears such maturity or enters *rigor mortis*, arguments pro and con are likely to be pointless.

1.2.1.8. *The indispensability of theory.* Carnap's answer to Putnam's central point of the indispensability of theory is characteristic of both the role of the language and the provisionary nature of the program (1963a, 987f):

> On this point I entirely agree with him; his belief that my conception here is "diametrically opposed" to his ... is not correct. In my publications I have discussed the problem of inductive logic not for a quantitative theoretical language, but only for simple language forms which may be regarded as constituting part of the qualitative observation language.
>
> ... it does not follow at all, as Putnam believes, that an adequate method of [degree of confirmation] is impossible, but rather that ... we must construct a new inductive logic which refers to the theoretical language instead of the observation language.

Thus it is not an entirely fair representation of Carnap's views to interpret his inductive logic as aiming for a purely syntactic 'universal prediction method'—though one can be forgiven for adopting this interpretation, based on much of what Carnap says in other places. Putnam certainly used this interpretation; when I introduce Solomonoff's views in 1.3 below, I relate these rather to *Putnam's interpretation of* Carnap's inductive logic.

1.2.1.9. **The subsequent development of Carnap's program.* My restriction in this thesis to Carnap's 'early' inductive logic leaves out some important later developments. In particular, the early phase precedes Carnap's transition from a formal logical language to the measure-theoretic framework of mathematical probability (see Jeffrey, 1971), which he employed in the formulation of his final "Basic System" (Carnap, 1971b, 1980; see Hilpinen, 1973). In the meantime Carnap's program had been subjected to some scathing criticism: I focus on Putnam (1963a, 761: "this particular project should be abandoned"), but other examples are Nagel (1963, 825: "he has not resolved the outstanding issues in the philosophy of induction, and his approach to the problems is not a promising one"), Lakatos (1968, 373: "the historian of thought may have to record a 'degenerating problem shift'"), and Hacking (1975, 142: "no foundation at all"). Despite such opposition, the program of inductive logic kept driving investigation and further refinement by an "invisible college" of followers (Zabell, 2011, 305). Even recent years have seen advances (e.g., Huttegger, 201x), accompanied, on the one side, by philosophical reappraisals of Carnap's inductive logic (e.g., Groves, 2015; Sznajder, 2016), on the other, by the mostly technical development of Carnapian inductive logic as a proper branch of mathematical logic (Paris and Vencovská, 2015).

1.2.2. Confirmation and prediction. Here I go through some details of Carnap's formal framework, as given in the *Logical Foundations* and the *Continuum* and for the relevant part preserved in the 1950's. I zoom in on the "singular predictive inference," and make the translation to our setting of sequential prediction. (Years cited without author refer to Carnap's works.)

1.2.2.1. *The formal language.* A language system $\mathfrak{L}_k^t$ consists of a finite number k of one-place *atomic predicates* $P_1, P_2, \dots, P_k$ and a finite number t of *individual constants* $a_1, a_2, \dots, a_t$. A particular combination $P_i a_j$ of a predicate P_i and an individual a_j is called an *atomic sentence*; all other sentences are generated from the atomic sentences under the customary logical connectives. Then a *state-description* $\mathfrak{Z}$, which gives a complete description of a particular state of affairs, is a conjunction of atomic sentences for each individual.

1.2.2.2. *Regular measure functions.* A *regular measure function* $\mathfrak{m}$ assigns to every state-description a positive real number, in such a way that the sum $\sum_{\mathfrak{Z}} \mathfrak{m}(\mathfrak{Z})$ over all state-descriptions equals 1. An assignment of $\mathfrak{m}$-values to the state-descriptions induces an assignment of $\mathfrak{m}$-values to all sentences (1950, 294ff; Kemeny, 1963, 721).

1.2.2.3. *Regular confirmation functions.* A *regular confirmation function* $\mathfrak{c}$ is defined from a regular measure function $\mathfrak{m}$ by

$$\mathfrak{c}(h, e) := \frac{\mathfrak{m}(e \mathbin{\&} h)}{\mathfrak{m}(e)}, \tag{4}$$

for all sentences e and h with $\mathfrak{m}(e) \neq 0$ (1950, 295; 1963a, 975). So even though it is a central component of the logical interpretation that probabilities are always *conditional* (1.2.1.1 above), Carnap still employs an unconditional measure function in their definition—which he interprets as the *null confirmation* $\mathfrak{c}_0$ (1950, 289, 307ff) or the confirmation with respect to the tautological evidence (also see 3.1.2 below).

1.2.2.4. *An infinite list of individuals.* In order to obtain a regular confirmation function ${}_\infty\mathfrak{c}$ for a language system $\mathfrak{L}_k^\infty$ with an unbounded number of individuals, we define in language systems $\mathfrak{L}_k^1, \mathfrak{L}_k^2, \dots$ with $\mathfrak{L}_k^t$ containing individuals $a_1, \dots, a_t$ a *fitting* sequence ${}_1\mathfrak{c}, {}_2\mathfrak{c}, \dots$ of regular confirmation functions such that every two functions agree on all sentences they both contain. Then ${}_\infty\mathfrak{c}(e, h)$ is simply defined as the limit $\lim_{t\to\infty} {}_t\mathfrak{c}(h, e)$. (See Carnap, 1950, 302ff, Carnap, 1963a, 975.) Note that a state-description in this language system is an infinite conjunction; it will be convenient to talk rather about the *t-state-descriptions* that are state-descriptions within $\mathfrak{L}_K^t$, i.e., complete descriptions of the first t individuals.

1.2.2.5. *The predictive inference.* We will now assume that the predicates are mutually exclusive and exhaustive (they form a *division*, 1950, 107f, or a *family*, 1963a, 973), meaning that each individual satisfies one and only one predicate. Let e^t denote a conjunction of atomic sentences for each of t individual constants $b_1, \dots, b_t$; and let h^s denote a conjunction of atomic sentences for another s individuals $c_1, \dots, c_s$ that are different from the previous ones (see Carnap, 1952, 12). The determination of the value of $\mathfrak{c}(h^s, e^t)$ is the *predictive inference*, "the most important and fundamental kind of inductive inference" (1950, 207; 568).

1.2.2.6. *The singular predictive inference.* Let e_j^t be formed from e^t by replacing every atomic predicate P_i with $i \neq j$ by the negation of P_j; and let h_j denote either the atomic sentence $P_j c$ or its negation for an individual constant c different from the $b_1, \ldots, b_t$ (see 1952, 12f; 1963a, 975). That is, e_j^t expresses for each of $b_1, \ldots, b_t$ whether or not it satisfies P_j; and h_j either states or denies that c satifies P_j. Then determining the value of $\mathfrak{c}(h_j, e_j^t)$ is the *singular predictive inference*, “the most important special case of the predictive inference” (1950, 568).

1.2.2.7. *Sequential prediction.* We now assume we have a family of just two predicates P_1 and P_2. Let us also assume that e_1^t as above concerns the first t individuals $a_1, \ldots, a_t$ of our language system $\mathfrak{L}_2^\infty$. Note that this means that the sentences e_1^t are precisely the t-state-descriptions. We can now encode e_1^t as a sequence $\boldsymbol{x}^t \in \mathbb{B}^t$, as follows. For each individual a_i, $i \leq t$, let $\boldsymbol{x}^t(i-1) = 1$ if $P_1 a_i$ is in e_1^t, and $\boldsymbol{x}^t(i-1) = 0$ otherwise (i.e., if $P_2 a_i$). Likewise, we assume that h_1 concerns the individual a_{t+1}, and we encode h_1 as ‘1’ if it is $P_1 a_{t+1}$ and as ‘0’ otherwise. In this reformulation the sequences $\boldsymbol{x}^t \in \mathbb{B}^t$ are precisely the t-state-descriptions, and the singular predictive inference is the problem of determining the value $\mathfrak{c}(x, \boldsymbol{x}^t)$ for given $x \in \mathbb{B}, \boldsymbol{x}^t \in \mathbb{B}^*$, that is, the problem of sequential prediction.

1.2.3. The Johnson-Carnap predictors.

Here I go through Carnap’s actual explications of logical probability for the singular predictive inference: or the specification of methods for sequential prediction.

Much of the same route that Carnap took was actually traced earlier, and unbeknownst to Carnap, by Johnson (1924; 1932; see Zabell, 2005, 2011.) For that reason I will refer to these prediction methods as the *Johnson-Carnap predictors.*

1.2.3.1. *State-symmetry.* Carnap’s first main stipulation on regular confirmation functions is that “an adequate concept of degree of confirmation should treat all individuals on a par” (Carnap, 1950, 483). That is to say, all *isomorphic* state-descriptions, that only differ from each other by an exchange of individual constants, should have the same $\mathfrak{m}$-value. Carnap calls these regular measure functions *symmetrical*; a more precise term used by Suppes (2002, 192) is *state-symmetry.* (Johnson 1924, 183 called this stipulation the *permutation postulate.*) In our setting, a state-symmetric measure function $\mathfrak{m}$ has the property that a sequence’s $\mathfrak{m}$-value only depends on the *number* of 0’s and 1’s it consists of, not on their order. To be precise,

$$\mathfrak{m}(\boldsymbol{x}^t) = \mathfrak{m}(\boldsymbol{y}^t) \text{ for every } \boldsymbol{x}^t, \boldsymbol{y}^t \text{ with } \#_0 \boldsymbol{x}^t = \#_0 \boldsymbol{y}^t. \tag{5}$$

For the induced confirmation function $\mathfrak{c}$ we then also have

$$\mathfrak{c}(\boldsymbol{x}^t) = \mathfrak{c}(\boldsymbol{y}^t) \text{ for } \boldsymbol{x}^t, \boldsymbol{y}^t \text{ with } \#_0 \boldsymbol{x}^t = \#_0 \boldsymbol{y}^t. \tag{6}$$

Carnap’s state-symmetry is nowadays better known as the condition of *exchangeability*. In modern statistical parlance, if we predict in accordance with (6), we would say that the frequency counts of 0’s and 1’s are a *sufficient statistic*.

1.2.3.2. *Structure-symmetry: the function $\mathfrak{c}^*$.* A *structure-description* is a maximal set of isomorphic state-descriptions; in our setting it is a set of all sequences with the same number of 0's and 1's. Note that the property of state-symmetry says that all sentences in each structure-description have equal probability. In (1950, 562ff), Carnap proposes the unique state-symmetric regular measure function $\mathfrak{m}^*$ that is also *structure*-symmetric: it assigns equal probability to each of the structure-descriptions. (Johnson called it the *combination postulate.*) In our setting, the induced confirmation function has the form

$$\mathfrak{c}^*(\boldsymbol{x}^t) := \left(\frac{\#_0\boldsymbol{x}^t + 1}{t+2}, \frac{\#_1\boldsymbol{x}^t + 1}{t+2} \right). \tag{7}$$

This is Laplace's rule of succession, that I introduced in I.1 above. It is a prime instance of an important appeal of Carnap's logical approach: that simply by capitalizing on neutral symmetries in the language, we end up with the positive result of a natural measure of confirmation, with a prediction method that can learn.

1.2.3.3. **The function $\mathfrak{c}\dagger$.* However, we actually lose the latter with the more stringent symmetry stipulation of assigning, for each t, equal probability to *all t-state-descriptions.* The resulting measure function $\mathfrak{m}\dagger$ is indeed discussed by Carnap (1950, 298f, 564f): in our setting, it reduces to the regular measure function

$$\mathfrak{m}\dagger(\boldsymbol{x}^t) := 2^{-t} \text{ for all } \boldsymbol{x}^t \in \mathbb{B}^*, \tag{8}$$

that is, the uniform measure λ. But $\mathfrak{c}\dagger(x \mid \boldsymbol{x}) = \lambda(x \mid \boldsymbol{x}) = 1/2$ for every $x \in \mathbb{B}, \boldsymbol{x} \in \mathbb{B}^*$, so $\mathfrak{c}^\dagger$ can never learn from experience, which "would obviously be in striking contradiction to the basic principle of all inductive reasoning" (ibid., 565). Also see 3.1.2.1 below.

1.2.3.4. *Towards the continuum.* In fact, Carnap did not express himself confident that even structure-symmetry was very compelling (1950, 564):

> No doubt, to the way of thinking which was customary in the classical period of the theory of probability, [structure-symmetry] would appear as validated as [state-symmetry], by the principle of indifference. However, to modern, more critical thought, this mode of reasoning appears as invalid because the structure-descriptions (in contradistinction to the individual constants) are by no means alike in their logical features but show very conspicuous differences ... It seems to me that the function $\mathfrak{c}^*$ cannot be justified by any features of the definition which are immediately recognizable, but only by studying the consequences to which the definition leads.

This he promised to do in a second volume; instead, in the 1952 monograph, $\mathfrak{c}^*$ took its place as just one particular choice within a *continuum* of confirmation functions.

1.2.3.5. *The λ-continuum.* The "valid part of the principle of indifference" (1963a, 975) now includes state-symmetry and *attribute-symmetry*: $\mathfrak{c}$-values are invariant under permutations of predicates (1952, 14). In our setting this comes down to stipulating that $\mathfrak{c}(x \mid \boldsymbol{x}^t) = \mathfrak{c}(y \mid \boldsymbol{y}^t)$ for all $x \neq y$ and $\boldsymbol{x}^t, \boldsymbol{y}^t$ such that $\boldsymbol{x}^t(i) \neq \boldsymbol{y}^t(i)$ for all $i < t$. Moreover, Carnap, like Johnson, adopts the *sufficientness postulate* (terminology due to Good, 1965) that states that the values $\mathfrak{c}(h_M, e^t_M)$ for fixed t are given by a *characteristic function* $G_{M,t}$ that only depends on the number of individuals satisfying M in e^t_M (1952, 15ff). Actually, in our case of only two predicates, this postulate vacuously holds true (Good, 1965, 26); but in the general case one can prove that G must be *linear* in the number of individuals satisfying M (1959), and in the case of two predicates we can make *this* stipulation instead (1963a, 976; see Zabell, 1982). That is, we stipulate that

$$\mathfrak{c}(x \mid \boldsymbol{x}^t) = G_{M,t}(\#_x \boldsymbol{x}^t) = a + b \cdot \#_x \boldsymbol{x}^t$$

for some $a, b \geq 0$. It turns out (1952, 27ff; 1963a, 976; Kemeny, 1963, 724ff) that each such confirmation function is of the form

$$\mathfrak{c}_\lambda(x \mid \boldsymbol{x}^t) := \frac{\#_x \boldsymbol{x}^t + \lambda/2}{t + \lambda} \tag{9}$$

for some $\lambda \in [0, \infty]$. This is the Johnson-Carnap continuum of inductive methods. Note that for $\lambda = 2$ we retrieve $\mathfrak{c}^*$, and for $\lambda \to \infty$ we retrieve $\mathfrak{c}\dagger$.

1.2.3.6. *The straight rule.* The value $\lambda = 0$ gives Reichenbach's *straight rule*, (2) in I.5 above. This rule "leads to quite implausible results" (1950, 568), having to do with the fact that it gives extreme predictive probabilities 0 and 1 if we have observed a sequence of only 0's or only 1's. More details on the technical difficulties this raises are given in 3.1.2.2.

1.2.3.7. *The λ-γ-continuum.* Carnap finally also dropped the requirement of attribute-symmetry (Carnap, 1980). Retaining only state-symmetry and the linearity of the characteristic function, we obtain confirmation functions of the form

$$\mathfrak{c}(x \mid \boldsymbol{x}^t) = G_{M,t}(\#_x \boldsymbol{x}^t) = a_x + b \cdot \#_x \boldsymbol{x}^t,$$

with a_x that depends on x. This introduces x-dependent terms γ_x (with $\gamma_0 + \gamma_1 = 1$) in the continuum of methods

$$\mathfrak{c}_{\lambda,\gamma}(x \mid \boldsymbol{x}^t) := \frac{\#_x \boldsymbol{x}^t + \gamma_x \lambda}{t + \lambda}. \tag{10}$$

1.2.3.8. *Interpretation: the prior and empirical factor.* The γ_x terms have an obvious interpretation as the initial weights we assign to both symbols: if $t = 0$ then $\mathfrak{c}_{\lambda,\gamma}(x \mid \varnothing) = \gamma_x$. In fact, a function $\mathfrak{c}_{\lambda,\gamma}$ in the λ-γ continuum—and this already holds for a function $\mathfrak{c}_\lambda$ in the λ-continuum—can be explicitly decomposed in an *empirical* and a *prior* element ("the empirical and the logical factor," 1952, 22):

$$\mathfrak{c}_{\lambda,\gamma}(x \mid \boldsymbol{x}^t) = \frac{t}{t + \lambda} \cdot \frac{\#_x \boldsymbol{x}^t}{t} + \frac{\lambda}{t + \lambda} \cdot \gamma_x. \tag{11}$$

The term $\frac{\#_x \boldsymbol{x}^t}{t}$ is the empirical frequency of x's and γ_x the initial or prior probability we assign to x. The terms $\frac{t}{t+\lambda}$ and $\frac{\lambda}{t+\lambda}$, that sum to 1, weigh the empirical and the prior element, respectively: the greater λ, the greater the weight to the prior element. (See Zabell, 2011, 276f.)

1.2.3.9. *The class of i.i.d. measures.* The previous interpretation shows how the Johnson-Carnap predictors are directed at the relative frequency of the data, and thereby already suggests a presupposition that the data show no interesting structure beyond a limiting relative frequency. This is more clear still in another pleasing pleasing interpretation of the Johnson-Carnap predictors, namely as maximum likelihood estimators for the class of *i.i.d. measures* (see 2.1.1.4), with various amounts of "virtual" or "a priori" symbols preceding the actual data: see Grünwald (2007, 258f). A further interpretation of the Johnson-Carnap predictors that I will discuss in chapter 3 is in fact as *mixtures* over the class of i.i.d. measures, with the explicit interpretation of incorporating an inductive assumption of an i.i.d. data-generating source.

* * *

1.3. Solomonoff's new start

Solomonoff's objective is clear (1964, 2):

> The problem dealt with will be the extrapolation of a long sequence of symbols—these symbols being drawn from some finite alphabet. More specifically, given a long sequence, represented by T, what is the probability that it will be followed by the subsequence represented by 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. This corresponds to Carnap's [logical probability].

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 (1.2.2.5, 1.2.2.6 above). Carnap's *requirement of total evidence* (see 1950, 211ff; 1963a, 972) returns in Solomonoff's remark that "the corpus that we will extrapolate ... *must contain all of the information that we want to use in the induction*" (ibid., 8). And Carnap's discussion under the header "Are Laws Needed for Making Predictions?" (ibid., 574f)—conclusion: "the use of laws is not indispensable"—is easily read as informing Solomonoff's statement 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).

The later Carnap would probably not have put this last point quite so boldly, though; and in general Solomonoff's is a purely syntactical perspective that is somewhat at odds with Carnap's views on the role of the language (1.2.1.6 above). This also comes to the fore in the following passage by Solomonoff, where he, looking back, confidently assesses Carnap's work and the relation to his own (1997, 76):

> Carnap's model of probability started with a long sequence of symbols that was a description of the entire universe. Through his own formal linguistic analysis, he was able to assign a priori probabilities to any possible string of symbols that might represent the universe. He derived his [confirmation function] from this a priori distribution using Bayes' theorem.
>
> I liked his function that went directly from data to probability distribution without explicitly considering various theories or "explanations" of the data. ... I also liked his idea of [degree of confirmation] and the idea of representing the universe by a digital string, but his method of computing the a priori probability distribution seemed unreasonable to me. The distribution depended very much on just what language was used to describe the universe. Furthermore, as one made the describing language larger and more complete, predictions were less and less contingent on the data. Carnap admitted these difficulties, but he felt that his theory nonetheless had redeeming qualities and that we would eventually find a way out of these difficulties.
>
> Algorithmic probability is close to Carnap's model, and it does overcome the difficulties described.

Solomonoff's suggestion that his explication of logical probability—what he here calls "algorithmic probability"—is not dependent on a choice of language is overly optimistic: again, this presupposes data are already presented in well-differentiated discrete form (1.2.1.6 above). But what is relevant to us is that the picture of Carnap's inductive logic that Solomonoff paints here, if not doing full justice to Carnap, closely resembles the picture that *Putnam* painted of Carnap's inductive logic, in order to challenge it.

(Did Carnap ever read Solomonoff's work? I have found no evidence for that: the only reference of Carnap to Solomonoff I know of are four pages of technical notes dating back already to March 1951, when Solomonoff "studied the logical basis of probability with Carnap" while majoring in physics in Chicago (Solomonoff, 1997, 74). These notes, entitled "R.J. Solomonoff" and now in the Carnap archive in Pittsburgh (Carnap, 1951), mention the "th. of information" and further consist of calculations on a measure function defined in terms of Shannon entropy.[6])

Solomonoff and Putnam both infer from Carnap the picture of purely syntactical universal prediction: Solomonoff as starting point for his theory, Putnam as starting point for his impossibility argument. Indeed, as we saw in 1.2.2.7 above, Solomonoff's problem setting of sequence extrapolation is readily translatable from the formal set-up that Putnam presupposes in his paper. (In particular, we identify individuals with positions in the sequence, as Putnam does, 1963a, 766, thus introducing an *ordering* on individuals.) This means that Solomonoff's setting is fully within the scope of Putnam's argument.

Carnap could still resort to the defense that 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). This is, however, a somewhat weak defense: Carnap does acknowledge at various places the need for taking into consideration the order of individuals in explicating logical probability (e.g., 1950, 62ff; 1963c, 225f); 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" (1963b, 297) is too weak to be adequate. Note that the adequacy of a confirmation function or a prediction method is now assessed by the regularities in the data that it is able to learn: the regularities that it is able to extrapolate in predictive probabilities. From this perspective, the difference in opinion 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; 1963c, 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]" (1963c, 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. This notion of effective computability is, of course, also the fundamental ingredient in Solomonoff's proposal. It is this aspect that genuinely sets Solomonoff's approach apart from Carnap's. The confirmation functions that Solomonoff proposed in (1964), and that evolved in the modern definition of the Solomonoff-Levin measure that we will investigate in the next chapter, were explicitly defined in terms of the inputs to a universal Turing machine. Moreover, one can show that the instance confirmation via the Solomonoff-Levin predictor of *any true computable hypothesis* will converge to 1, thus fulfilling (I).

But does that mean Solomonoff has somehow evaded Putnam's argument? This will be the topic of the next chapters, beginning with the precise definition of the Solomonoff-Levin measure.

*

CHAPTER 2

The Solomonoff-Levin measure

This chapter discusses the definition of the Solomonoff-Levin measure. The crucial move is the expansion of the class of computable or Δ_1 measures to the class of Σ_1 measures. This class, as opposed to the Δ_1 measures, cannot be diagonalized, which is to say that it has universal elements. The Solomonoff-Levin measure is such a universal element, and this accounts for its convergence to any true Δ_1 measure.

This is a mainly technical chapter that the reader who is more interested in the conceptual story may prefer to pass over. Many observations in the following chapters rely on the groundwork done here; but in those instances I will explictly refer back to the relevant places in this chapter and the reader can choose to only first familiarize herself with the technicalities, if she so desires, at those points.

In 2.1, I work towards the definition of the Solomonoff-Levin measure: in 2.1.1, I introduce the Δ_1 measures; in 2.1.2, I introduce the superclass of Σ_1 measures; in 2.1.3, I introduce the universal Σ_1 measures; in 2.1.4, I finally introduce the Solomonoff-Levin measure. In 2.2, I give alternative characterizations of the Solomonoff-Levin measure that will be put to work in the following chapters.

Innovations. Section 2.1, while containing no new mathematical results, synthesizes different and superficially disparate presentations of the Solomonoff-Levin measure in the literature. (This section is based on part of Sterkenburg, 201x.) Section 2.2 presents a number of new results, of which theorem 2.13 is the most important and which largely come together in the generalized representation theorem 2.19 that concludes the section. (This section is based on Sterkenburg, 2017.[7])

2.1. The definition

2.1.1. The Δ_1 measures.

2.1.1.1. *Measures on Cantor space.* We consider measures on the class $\mathbb{B}^\omega$ of infinite sequences, also known as the *Cantor space*. More accurately, a measure on Cantor space is defined on a tuple $(\mathbb{B}^\omega, \mathfrak{F})$, with $\mathfrak{F}$ a σ-algebra on $\mathbb{B}^\omega$. Then a probability measure on $(\mathbb{B}^\omega, \mathfrak{F})$ is a countably additive function $\mu : \mathfrak{F} \to [0, 1]$ with $\mu(\mathbb{B}^\omega) = \mu([\![\varnothing]\!]) = 1$. 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* or *cones* $[\![\boldsymbol{x}]\!] = \{\boldsymbol{x}^\omega : \boldsymbol{x} \prec \boldsymbol{x}^\omega\}$ for all $\boldsymbol{x} \in \mathbb{B}^*$. 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 $m(\varnothing) = 1$ and $m(\boldsymbol{x}0) + m(\boldsymbol{x}1) = m(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{B}^*$. The extension theorem due to Carathéodory (see Tao, 2011, 148ff) then gives a σ-algebra $\mathfrak{F}$ over $\mathbb{B}^\omega$ (which includes all Borel classes) and unique measure μ_m on $\mathfrak{F}$ with $\mu_m([\![\boldsymbol{x}]\!]) = m(\boldsymbol{x})$. (For more details see Reimann, 2009, 249ff; Nies, 2009, 68ff; Li and Vitányi, 2008, 262ff; Calude, 1994, 6ff.) For the purposes of this thesis it is mostly unnecessary to be very strict about the difference between a measure and its pre-measure, and I will often simply write '$\mu(\boldsymbol{x})$' for $\mu([\![\boldsymbol{x}]\!])$.

2.1.1.2. *The uniform measure.* The most basic measure on Cantor space is the *uniform* or *Lebesgue measure* λ. It is given by $\lambda(\boldsymbol{x}^t) = 2^{-t}$ for all $\boldsymbol{x}^t$ (or more proper, it is generated from the pre-measure with $m(\boldsymbol{x}^t) = 2^{-t}$ for all $\boldsymbol{x}^t$).

2.1.1.3. *Conditional measures.* In sequential prediction the only conditional measures we ever encounter are those conditional on single finite sequences $\boldsymbol{x}$ (or more proper, cones $[\![\boldsymbol{x}]\!]$). A measure μ conditional on $\boldsymbol{x}$ is defined by

$$\mu(\cdot \mid [\![\boldsymbol{x}]\!]) := \frac{\mu(\cdot)}{\mu([\![\boldsymbol{x}]\!])}, \tag{12}$$

provided $\mu([\![\boldsymbol{x}]\!]) \neq 0$. In fact, I will adopt the convention, very natural in our setting, of simply writing '$\mu(\boldsymbol{y} \mid \boldsymbol{x})$' for what is actually $\mu(\boldsymbol{xy} \mid \boldsymbol{x})$, or more proper still, $\mu([\![\boldsymbol{xy}]\!] \mid [\![\boldsymbol{x}]\!])$. The *one-step* conditional measure $\mu^1(\cdot \mid \boldsymbol{x})$ is the distribution $(\mu(0 \mid \boldsymbol{x}), \mu(1 \mid \boldsymbol{x}))$.

2.1.1.4. *I.i.d. measures.* An independent and identically distributed (i.i.d.) probabilistic process is modeled by an *i.i.d.* (or *Bernoulli*) measure: a μ such that there is a distribution p on $\mathbb{B}$ with $\mu^1(\cdot \mid \boldsymbol{x}) = p(\cdot)$ for every $\boldsymbol{x} \in \mathbb{B}^*$. I denote the i.i.d. measure with $p = (\theta, 1-\theta)$ by 'μ_θ.' Thus the Lebesgue measure λ is the i.i.d. measure $\mu_{1/2}$.

2.1.1.5. *Strictly positive measures.* A measure is *strictly positive* if it assigns positive probability to every finite sequence (more proper, to every cone; it is generated from a pre-measure that assigns a positive value to every element in its range).

2.1.1.6. *Deterministic measures.* In contrast, a *deterministic* measure gives probability 1 to all the initial segments of a single infinite sequence $\boldsymbol{x}^\omega$, so $\mu(\boldsymbol{x}^t) = 1$ for all $t \in \mathbb{N}$. Or more proper: it assigns probabilility 1 to this infinite sequence, and it is generated from a pre-measure that assigns the value 1 to all its initial segments.

2.1.1.7. *Continuous measures.* An *atomic* measure gives positive probability to one (or more than one) particular infinite sequence, an *atom*. (So a deterministic measure is a special case of an atomic measure where the one atom receives probability 1.) A *non-atomic* or *continuous* measure has no atoms: no infinite sequence is assigned positive probability. (See Downey and Hirschfeldt, 2010, 265. In the algorithmic information theory literature, Li and

Vitányi, 2008, 265ff, 294ff, often the term *continuous* measure is used to refer to measures on the "continuous" Cantor space; this in contradistinction to the *discrete* measures by which are then meant the *distributions* on all *finite* sequences, see A.1.2. Due to the obvious risk of confusion I will avoid this usage here.)

2.1.1.8. *Computable measures.* 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 $f : \mathbb{B}^* \times \mathbb{N} \to \mathbb{Q}$ such that $|f(\boldsymbol{x}, s) - m(\boldsymbol{x})| < 2^{-s}$ for all $\boldsymbol{x} \in \mathbb{B}^*, s \in \mathbb{N}$ (see Downey and Hirschfeldt, 2010, 202f). For example, the Lebesgue measure is obviously computable; and a computable deterministic measure assigns probability 1 to a computable infinite sequence.

2.1.1.9. Δ_1 *measures.* I will employ the nomenclature of the *arithmetical hierarchy* of levels of effective computability (Kleene, 1943; Mostovski, 1947; see Soare, 2016, 79ff) and henceforth refer to the computable measures as the Δ_1 ('delta-one') measures. Thus we have

Definition 2.1. A Δ_1 measure μ on $\mathbb{B}^\omega$ is defined by $\mu([\![\boldsymbol{x}]\!]) = m(\boldsymbol{x})$ for a computable $m : \mathbb{B}^* \to [0, 1]$ that satisfies $m(\varnothing) = 1$ and $m(\boldsymbol{x}0) + m(\boldsymbol{x}1) = m(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{B}^*$.

2.1.1.10. *The Solomonoff-Levin measure.* 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 $Q_U(x_{t+1} \mid \boldsymbol{x}^t)$ for $x_{t+1} \in \mathbb{B}, \boldsymbol{x}^t \in \mathbb{B}^t$ converge to the values $\mu(x_{t+1} \mid \boldsymbol{x}^t)$ as t goes to infinity. That is, Q_U satisfies the following condition on an inductive method M:

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

This is an instance of condition (I*) on a measure, that at the same time generalizes from deterministic computable hypotheses or single infinite computable sequences to probability measures on infinite sequences.

Moreover, we can rephrase condition (II*) on a measure as

(II: Δ_1) M is Δ_1.

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

2.1.2. The Σ_1 measures. I proceed with the notion of a *semi-computable* or Σ_1 ('sigma-one') measure on the extended space $\mathbb{B}^\omega \cup \mathbb{B}^*$ of infinite *and finite* sequences. This notion will strike those who see it for the first time as rather involved, if not downright awkward: I will try to explain in what sense it is both natural and important. First, in the current subsection, 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, in 2.1.3 below, 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.

2.1.2.1. *Transformations.* Let a *transformation* λ_F of the uniform measure by Borel function $F : \mathbb{B}^\omega \to \mathbb{B}^\omega$ be defined by $\lambda_F(A) = \lambda(F^{-1}(A))$. Every Borel measure μ on Cantor space can be obtained as a transformation of λ by some Borel function. (See Reimann, 2009, 252f.)

2.1.2.2. *Monotone mappings.* We will now consider transformations by functions that are *effectively computable.* In order to impose the restriction of computability, we need to downscale the transformations to functions on *finite* sequences. To that end we introduce (partial) mappings $\psi : \mathbb{B}^* \to \mathbb{B}^*$, that have to satisfy a condition of *monotonicity*:

$$\text{if } \boldsymbol{x} \preccurlyeq \boldsymbol{y} \text{ and } \psi(\boldsymbol{y})\downarrow \text{ then also } \psi(\boldsymbol{x})\downarrow \preccurlyeq \psi(\boldsymbol{y}). \tag{13}$$

That means that by taking the ψ-image of increasingly large initial segments of some infinite sequence $\boldsymbol{x}^\omega$, we construct a new (possibly but not necessarily infinite) sequence. Formally, ψ induces the function $\Phi_\psi : \boldsymbol{x}^\omega \mapsto \sup_\preccurlyeq\{\psi(\boldsymbol{x}) : \boldsymbol{x} \prec \boldsymbol{x}^\omega\}$. If $\sup_\preccurlyeq\{\psi(\boldsymbol{x}) : \boldsymbol{x} \prec \boldsymbol{x}^\omega\}$ is indeed an infinite sequence for all infinite $\boldsymbol{x}^\omega$, then Φ_ψ gives a total function $F : \mathbb{B}^\omega \to \mathbb{B}^\omega$. If not, then we have to restrict the domain and Φ_ψ is a partial function on $\mathbb{B}^\omega$. Alternatively, we can treat Φ_ψ as a total function $\mathbb{B}^\omega \cup \mathbb{B}^* \to \mathbb{B}^\omega \cup \mathbb{B}^*$ on the collection of infinite *and* finite sequences. (Cf. Reimann, 2009, 253; Shen et al., 20xx.)

2.1.2.3. *Computable monotone mappings: monotone machines.* 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 (see Li and Vitányi, 2008, 298f; Shen et al., 20xx). Originally dubbed an *algorithmic process* (Zvonkin and Levin, 1970, 99), this type of machine is now better known as a *monotone* machine.

2.1.2.4. *Monotone machines: the definition.* It is mathematically convenient to represent a monotone mapping ψ by the set M_ψ of pairs of sequences $(\boldsymbol{x}, \boldsymbol{y})$ such that $\psi(\boldsymbol{x}) \succcurlyeq \boldsymbol{y}$. The latter says that ψ when given $\boldsymbol{x}$ produces at least $\boldsymbol{y}$: therefore the interpretation of $(\boldsymbol{x}, \boldsymbol{y}) \in M_\psi$ is that $\boldsymbol{x}$ is a ψ*-description* for $\boldsymbol{y}$. A monotone machine is then given by a c.e. such set of pairs. (Cf. Reimann, 2009, 253f; Shen et al., 20xx.) Actually, following Levin (1973, 1413), I will employ the following definition, somewhat more abstract still, that leaves the associated mapping ψ fully implicit. (The economy of this definition will be useful in some of the proofs to follow.)

Definition 2.2 (Levin). A monotone machine is a c.e. set $M \subseteq \mathbb{B}^* \times \mathbb{B}^*$ that satisfies

$$\text{if } (\boldsymbol{x}_1, \boldsymbol{y}_1), (\boldsymbol{x}_2, \boldsymbol{y}_2) \in M \text{ and } \boldsymbol{x}_1 \preccurlyeq \boldsymbol{x}_2 \text{ then } \boldsymbol{y}_1 \sim \boldsymbol{y}_2. \tag{14}$$

The associated function $\Phi_M : \mathbb{B}^\omega \cup \mathbb{B}^* \to \mathbb{B}^\omega \cup \mathbb{B}^*$ is induced by

$$\Phi_M(\boldsymbol{x}) = \sup_{\preccurlyeq}\{\boldsymbol{y} \in \mathbb{B}^* : \exists \boldsymbol{x}' \preccurlyeq \boldsymbol{x}\, ((\boldsymbol{x}', \boldsymbol{y}) \in M)\} \tag{15}$$

(also see Gács 2016, 2f). We again call sequence $\boldsymbol{x}$ an *M-description* of sequence $\boldsymbol{y}$ if Φ_M on $\boldsymbol{x}$ produces at least $\boldsymbol{y}$, that is, if $(\boldsymbol{x}', \boldsymbol{y}') \in M$ for some $\boldsymbol{x}' \preccurlyeq \boldsymbol{x}$ and $\boldsymbol{y}' \succcurlyeq \boldsymbol{y}$.

2.1.2.5. **Monotone machines: different models.* I will not go into the exact Turing machine model (i.e., specification of the input and output tapes and allowed operations on them) that corresponds to the above definition. For further details, as well as discussion of similar models and definitions in the literature (including those of Solomonoff, 1964, Zvonkin and Levin, 1970, and Schnorr, 1973, 1977), see Downey and Hirschfeldt (2010, 145ff); Day (2009, 215f); Li and Vitányi (2008, 335ff).

2.1.2.6. *Effective transformations.* Now consider the transformation of the uniform measure λ by a computable monotone mapping ψ. This transformation is given by the pre-measure $m : \boldsymbol{y} \mapsto \lambda([\![\{\boldsymbol{x} : \psi(\boldsymbol{x}) \succcurlyeq \boldsymbol{y}\}]\!])$, mapping to each sequence $\boldsymbol{y}$ the uniform measure of the input sequences $\boldsymbol{x}$ that lead ψ to produce it (Zvonkin and Levin, 1970, 100). Putting things in terms of monotone machines, we let the transformation of λ by M be given by the pre-measure that maps each $\boldsymbol{y}$ to the uniform measure of its M-descriptions. I will also call λ_M the *uniform* transformation by M. The value $\lambda_M(\boldsymbol{y})$ can be interpreted as the probability that sequence $\boldsymbol{y}$ is produced by monotone Turing machine M when given uniformly random input.

Definition 2.3. The uniform transformation λ_M by M is given by

$$\begin{aligned}\lambda_M(\boldsymbol{y}) :&= \lambda(\{\boldsymbol{x} : \Phi_M(\boldsymbol{x}) \succcurlyeq \boldsymbol{y}\})\\ &= \lambda(\{\boldsymbol{x} : \exists \boldsymbol{x}' \preccurlyeq \boldsymbol{x}. \exists \boldsymbol{y}' \succcurlyeq \boldsymbol{y}.\ (\boldsymbol{x}', \boldsymbol{y}') \in M\})\\ &= \lambda(\{\boldsymbol{x} : \exists \boldsymbol{y}' \succcurlyeq \boldsymbol{y}.\ (\boldsymbol{x}, \boldsymbol{y}') \in M\}).\end{aligned}$$

The first equality in the definition follows from writing out $\Phi_M(\boldsymbol{x})$ per (15); the second equality follows because we only need to take into account the measure of the *minimal* M-descriptions of $\boldsymbol{y}$, those $\boldsymbol{x}$ that have no prefixes $\boldsymbol{x}' \prec \boldsymbol{x}$ that are already M-descriptions of $\boldsymbol{y}$. Indeed, we can rewrite

$$\lambda_M(\boldsymbol{y}) = \sum_{\boldsymbol{x} \in D_M(\boldsymbol{y})} \lambda(\boldsymbol{x}), \tag{16}$$

with $D_M(\boldsymbol{y}) = \lfloor \{\boldsymbol{x} : \Phi_M(\boldsymbol{x}) \succcurlyeq \boldsymbol{y}\} \rfloor$ the set of minimal M-descriptions of $\boldsymbol{y}$, i.e., the bottom of the set of M-descriptions of $\boldsymbol{y}$ (see page v).

2.1.2.7. **Probabilistic machines.* A slightly different alternative interpretation views a monotone machine on random input as a probabilistic machine that has an internal random-bit generator to aid its computations. In particular, a monotone machine as defined above is a probabilistic machine that computes *without* any input, but with the possibility of using self-generated random bits. The value $\lambda_M(\boldsymbol{y})$ is then interpretated as the probability that M in the course of its computation generates $\boldsymbol{y}$. (See Shen et al., 20xx.)

2.1.2.8. *Effective transformations and the Δ_1 measures.* If monotone machine M produces an infinite sequence with uniform probability 1 (i.e., the class of $\boldsymbol{x}^\omega$ with infinite $\Phi_M(\boldsymbol{x}^\omega)$ has uniform measure 1), then the uniform transformation λ_M is a pre-measure that again generates a Δ_1 measure on $\mathbb{B}^\omega$. We then indeed have an effective analogue to the statement of 2.1.2.1 above: 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, 100f).

2.1.2.9. *Measures on* $\mathbb{B}^\omega \cup \mathbb{B}^*$. The monotone machines leading to the Δ_1 measures thus have the special property that they are 'almost total,' meaning that they produce an unending sequence on λ-almost all infinite input streams. In general a monotone machine M can fail to do so. This is the case when there is some finite $\boldsymbol{y}$ such that with positive uniform probability machine M stops producing more symbols after $\boldsymbol{y}$ (that is, the class of $\boldsymbol{x}^\omega$ with *finite* $\Phi_M(\boldsymbol{x}^\omega)$ has positive uniform probability), and this implies that $\lambda_M(\boldsymbol{y})$ is *strictly greater* than $\lambda_M(\boldsymbol{y}0) + \lambda_M(\boldsymbol{y}1)$. In that case we can say that λ_M assigns positive probability to the *finite* sequence $\boldsymbol{y}$. A function λ_M can thus be interpreted as (a pre-measure to) a measure on the collection $\mathbb{B}^\omega \cup \mathbb{B}^*$ of infinite *and* finite sequences.

2.1.2.10. *Semi-measures.* Alternatively, one can interpret such a function as a "semi-measure" on $\mathbb{B}^\omega$ (Levin and V'yugin, 1977, 360), a "defective" probability measure. See Li and Vitányi (2008, 264, 331f).

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

Definition 2.4. A Σ_1 measure ν on $\mathbb{B}^\omega \cup \mathbb{B}^*$ is defined by $\nu([\![\boldsymbol{x}]\!]) = m(\boldsymbol{x})$ for a lower semi-computable $m : \mathbb{B}^* \to [0,1]$ that satisfies $m(\varnothing) \leq 1$ and $m(\boldsymbol{x}0) + m(\boldsymbol{x}1) \leq m(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{B}^*$.

Let us denote the class of all Σ_1 measures by $\mathcal{M}$. Every uniform transformation λ_M is a Σ_1 measure; and conversely, every Σ_1 measure is given by some uniform transformation.

Proposition 2.5 (Levin). $\mathcal{M} = \{\lambda_M\}_M$, where the M range over all monotone machines.

Proof. See the proof of the more general proposition 2.10 in B.1.1.

2.1.3. Universal measures. 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. (Cf. I.6 above.) 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. 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. 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 listing of all elements that is itself computable: *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.* "This fact is one of the reasons for introducing the concept of semi-computable measure" (Zvonkin and Levin, 1970)—we may take it as the main reason.

(The analogy between the Σ_1 measures and the p.c. functions is indeed an *equivalence* in the sense that an effective enumeration of all Σ_1 measures is obtained from an effective enumeration of all p.c. functions: see Li and Vitányi, 2008, 261, 267f.)

2.1.3.1. *Dominance and universality.* Informally, a universal Σ_1 measure "is 'larger' than any other measure, and is concentrated on the widest subset of $\mathbb{B}^\omega \cup \mathbb{B}^*$" (Zvonkin and Levin, 1970, 104, notation mine). Formally, a universal Σ_1 measure *majorizes* or *dominates* every other Σ_1 measure (ibid., 103f):

Definition 2.6 (Levin). A universal Σ_1 measure $\mathring{\nu}$ is a Σ_1 measure such that for every $\nu \in \Sigma_1$ we have

$$\mathring{\nu} \geq^\times \nu,$$

meaning that there is a constant $c_\nu \in \mathbb{N}$, that depends on $\mathring{\nu}$ and ν, such that for all $\boldsymbol{x} \in \mathbb{B}^*$ it holds that

$$\mathring{\nu}(\boldsymbol{x}) \geq c_\nu^{-1}\nu(\boldsymbol{x}).$$

2.1.3.2. *Mutual dominance.* Note that any two universal Σ_1 measures $\mathring{\nu}_1$ and $\mathring{\nu}_2$ by definition dominate *each other*: $\mathring{\nu}_1 \geq^\times \mathring{\nu}_2$ and $\mathring{\nu}_2 \geq^\times \mathring{\nu}_1$, that is,

$$\mathring{\nu}_1 =^\times \mathring{\nu}_2.$$

Thus every two universal Σ_1 measures are equivalent up to a multiplicative constant.

2.1.3.3. **Diagonalizing the Δ_1 measures (1).* So why, exactly, cannot there already exist a universal Δ_1 measure: a $\mathring{\mu} \in \Delta_1$ such that for every $\mu \in \Delta_1$ there is a c with $\mathring{\mu}(\boldsymbol{x}) \geq c^{-1}\mu(\boldsymbol{x})$? First of all, it is easy to see that there cannot be a computable enumeration $\{\mu_i\}_{i\in\mathbb{N}}$ of all Δ_1 measures. Namely, if there were, we could construct a diagonal *deterministic* Δ_1 measure μ—that is, an infinite sequence $\boldsymbol{x}^\omega$—as follows: for each i, let $\boldsymbol{x}^\omega(i) := 0$ (so $\mu(0 \mid \boldsymbol{x}^{i-1}) := 1$) if $\mu_i(0 \mid \boldsymbol{x}^{i-1}) < 0.5$, and $\boldsymbol{x}^\omega(i) := 1$ otherwise.

2.1.3.4. **Diagonalizing the Δ_1 measures (2).* This, however, does not yet immediately entail that there cannot be a universal Δ_1 measure. To decidedly prove so, suppose for a contradiction that there is such a Δ_1 measure $\mathring{\mu}$. Now consider some infinite computable prefix-free list $\{\boldsymbol{x}_i\}_{i \in \mathbb{N}}$ of finite sequences. Since the set is prefix-free, the probabilities assigned by $\mathring{\mu}$ to these finite sequences must sum to 1 (by Kraft's inequality, A.2.2.1). But then the probability that $\mathring{\mu}$ assigns to $\boldsymbol{x}_i$ goes to 0 as i goes to infinity; and we can compute an infinite sublist $\{\boldsymbol{x}_j\}_j$ of finite sequences with $\mathring{\mu}(\boldsymbol{x}_j) < 2^{-j}/j$. Now define, computably, a μ with $\mu(\boldsymbol{x}_j) = 2^{-j}$ (and $\mu(\boldsymbol{x}_i) = 0$ for those $\boldsymbol{x}_i$ not in the sublist of $\boldsymbol{x}_j$'s). Then for every c, there is an $\boldsymbol{x}_j$ with $\mathring{\mu}(\boldsymbol{x}_j) < c^{-1}\mu(\boldsymbol{x}_j)$, so $\mathring{\mu}$ does not dominate μ, contrary to assumption. (Also see Li and Vitányi, 2008, 270, 298.)

2.1.3.5. **Diagonalizing the Δ_1 measures (3).* Another way of seeing this is via Putnam's argument. This showed that no Δ_1 measure can converge on any true Δ_1 measure: no measure satisfying (II: Δ_1) can satisfy (I: Δ_1). But we will see below that it is enough for satisfying (I: Δ_1) to dominate all Δ_1 measures, hence to be a universal Δ_1 measure. A universal Δ_1 measure would satisfy both conditions: therefore it cannot exist.

2.1.3.6. *Incomputability of universal Σ_1 measures.* A universal Σ_1 measure $\mathring{\nu}$ clearly cannot be computable or Δ_1: otherwise it would already be a universal Δ_1 measure.

2.1.3.7. *Universal monotone machines.* A universal Σ_1 measure can be obtained as a transformation of a *universal monotone machine.* Since the monotone machines can be effectively enumerated (simply enumerate the c.e. sets on $\mathbb{B}^* \times \mathbb{B}^*$ with the appropriate restriction (14) in definition 2.2), we can define, for some prefix-free computable list $\{\boldsymbol{z}_i\}_{i \in \mathbb{N}}$ of sequences, a c.e. set U that contains the pair $(\boldsymbol{z}_i\boldsymbol{x}, \boldsymbol{y})$ if the i-th monotone machine in this enumeration contains the pair $(\boldsymbol{x}, \boldsymbol{y})$. The interpretation is that this universal monotone machine U can emulate any other monotone machine on receiving the corresponding code sequence.

Definition 2.7. A universal monotone machine U is defined by

$$(\boldsymbol{z}_e\boldsymbol{x}, \boldsymbol{y}) \in U :\Leftrightarrow (\boldsymbol{x}, \boldsymbol{y}) \in M_e \tag{17}$$

for some computable prefix-free encoding $\{\boldsymbol{z}_e\}_{e \in \mathbb{N}}$ of all monotone machines M_e.

2.1.3.8. **Weakly universal machines.* The above property (17) is sometimes referred to as universality *by adjunction*, to distinguish it from a strictly more general universality property (see Downey and Hirschfeldt, 2010, 111; Barmpalias and Dowe, 2012, 3492f). Namely, a monotone machine U is *weakly universal* if for all M there is a c_M such that

$$\text{if } (\boldsymbol{x}, \boldsymbol{y}) \in M \text{ then } \exists \boldsymbol{x}' \left(|\boldsymbol{x}'| < |\boldsymbol{x}| + c_M \;\&\; (\boldsymbol{x}', \boldsymbol{y}) \in U \right). \tag{18}$$

2.1.3.9. *Universal transformations.* We will call a transformation λ_U of λ by a universal monotone machine U a *universal* (*uniform*) *transformation.* A universal transformation of λ is a universal Σ_1 measure: for every $\nu \in \Sigma_1$ there is a constant c such that $\lambda_U(\boldsymbol{y}) \geq c^{-1}\nu(\boldsymbol{y})$ for all $\boldsymbol{y} \in \mathbb{B}^*$. This follows from the fact that we can write out (see the proof of theorem 2.16 below in B.1.4) that

$$\lambda_U(\cdot) = \sum_e \lambda(\boldsymbol{z}_e)\lambda_{M_e}(\cdot).$$

Now by proposition 2.5, we know that for any given $\nu \in \Sigma_1$ there is some M_e with $\lambda_{M_e} = \nu$. This means that for $c^{-1} = \lambda(\boldsymbol{z}_e)$, for all $\boldsymbol{y} \in \mathbb{B}^*$ we have $\lambda_U(\boldsymbol{y}) \geq c^{-1}\nu(\boldsymbol{y})$.

2.1.3.10. **From computability to universality.* The expansion to semi-computable 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*, employs the class of all Σ_1 *randomness tests*: a sequence $\boldsymbol{x}^\omega$ is random if it passes a universal such test (see A.4). We will see in chapter 6 that Vovk (1998; 2001b), in defining his notion of *predictive complexity* and indeed inspired by Levin, employs the class of Π_1 *loss processes*: the predictive complexity of $\boldsymbol{x}^\omega$ is the loss incurred by a universal such process.

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

Definition 2.8 (Solomonoff, Levin). $Q_U := \lambda_U$.

So there are in fact infinitely many such measures Q_U, one for each choice of universal monotone machine U. Let us denote the class of Solomonoff-Levin measures by

$$\mathcal{SL} := \{Q_U\}_U = \{\lambda_U\}_U,$$

where the U range over all universal monotone machines. Since, as universal transformations, they are all universal Σ_1 elements, any two Solomonoff-Levin measures are equivalent up to a multiplicative constant.

Solomonoff would later (e.g., 1986; 1997; 2009) also employ the term "algorithmic probability" for the values given by the Solomonoff-Levin measure. (Terminology in the field is not very stable; Li and Vitányi, 2008, 272f, for instance, reserve this label for the related but still importantly different function in A.3.1.7.) The usual interpretation is that $Q_U(\boldsymbol{y})$ gives the probability that $\boldsymbol{y}$ is produced by universal monotone machine U when given uniformly random input. We can again write this in the form (16), as

$$Q_U(\boldsymbol{y}) = \sum_{\boldsymbol{x} \in D_U(\boldsymbol{y})} \lambda(\boldsymbol{x}), \tag{19}$$

with $D_U(\boldsymbol{y}) = \lfloor \{\boldsymbol{x} : \Phi_U(\boldsymbol{x}) \succcurlyeq \boldsymbol{y}\} \rfloor$ the set of minimal U-descriptions of $\boldsymbol{y}$. The definition is commonly associated with data compression and a preference for simplicity, as I will discuss in various places (particularly, chapter 5) below.

But the crucial property of the Solomonoff-Levin measures is their universality: in particular, their dominance over the Δ_1 measures. It is this property that is exploited in the adequacy result.

THEOREM 2.9 (Solomonoff). *Q_U fulfills* (I: Δ_1).

PROOF. The fact that Q_U dominates given Δ_1 measure μ entails that μ is *absolutely continuous with respect to* Q_U, meaning that $\mu(A) > 0$ implies $Q_U(A) > 0$ for all A in the σ-algebra $\mathcal{B}$. This entails by the classical result of Blackwell and Dubins (1962) that μ-a.s. the variational distance

$$\sup_{A \in \mathcal{B}} |\mu(A \mid \boldsymbol{x}^t) - Q_U(A \mid \boldsymbol{x}^t)|$$

goes to 0 as t goes to infinity (see, e.g., Huttegger, 2015, 617f). In particular, μ-a.s.,

$$Q_U(x \mid \boldsymbol{x}^t) \xrightarrow{t \to \infty} \mu(x \mid \boldsymbol{x}^t).$$

(For more discussion and Solomonoff's original 1978 proof, see B.2.1.) □

* * *

2.2. Alternative definitions

This section provides alternative definitions of the Solomonoff-Levin measure. I employ these alternative definitions to support interpretative observations throughout this thesis.

In 2.2.1, I give a generalized characterization of the Solomonoff-Levin measures as universal transformations. In 2.2.2, I consider the characterization of the Solomonoff-Levin measures as mixtures over all Σ_1 measures, and again give a generalization.

2.2.1. Generalized transformations.

2.2.1.1. *Transformations of* Δ_1 *measures.* Recall definition 2.3 of an effective transformation of the uniform measure λ. This uniform transformation is an instance of the general definition of a transformation of a Δ_1 measure μ, viz.

$$\mu_M(\boldsymbol{y}) := \mu([\![\{\boldsymbol{x} : \exists \boldsymbol{y}' \succcurlyeq \boldsymbol{y}.\ (\boldsymbol{x}, \boldsymbol{y}') \in M\}]\!]). \tag{20}$$

This value is the probability of obtaining a M-description for $\boldsymbol{y}$ when sampling random bits from μ. It is the sum of the μ-probabilities of the minimal M-descriptions for $\boldsymbol{y}$:

$$\mu_M(\boldsymbol{y}) = \sum_{\boldsymbol{x} \in D_M(\boldsymbol{y})} \mu(\boldsymbol{x}),$$

with $D_M(\boldsymbol{y}) = \lfloor \{\boldsymbol{x} : \Phi_U(\boldsymbol{x}) \succcurlyeq \boldsymbol{y}\} \rfloor$ the set of minimal U-descriptions of $\boldsymbol{y}$.

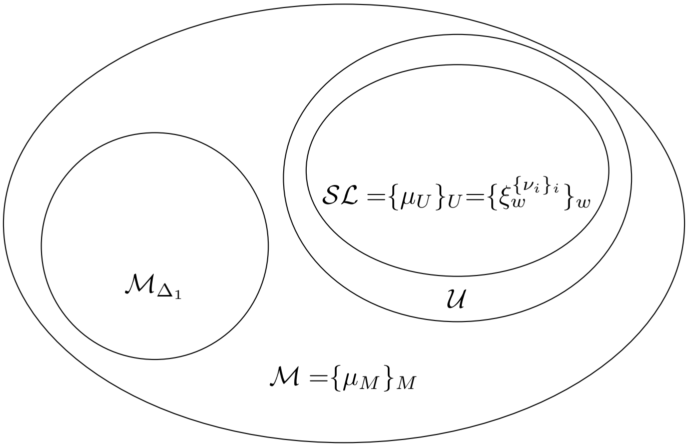

$\mathcal{SL} = \{\mu_U\}_U = \{\xi_w^{\{\nu_i\}_i}\}_w$
$\mathcal{M}_{\Delta_1}$
$\mathcal{U}$
$\mathcal{M} = \{\mu_M\}_M$

FIGURE 1. The class $\mathcal{M}$ of Σ_1 measures (2.1.2.11), that coincides with the class $\{\lambda_M\}_M$ of uniform transformations via all monotone machines (proposition 2.5), indeed, coincides with the class $\{\mu_M\}_M$ of all transformations of μ for any continuous Δ_1 measure μ (proposition 2.10). Contained in it the subclass $\mathcal{U}$ of universal Σ_1 measures (2.1.3.1), that is disjoint from the class $\mathcal{M}_{\Delta_1}$ of Δ_1 measures (2.1.3.6), and strictly contains (2.2.1.3) the class of Solomonoff-Levin measures $\mathcal{SL}$ (2.1.4), that coincides with all classes of universal transformations of a continuous Δ_1 measure μ (theorem 2.13) and with the class of universal Σ_1 mixtures (theorems 2.16 and 2.19).

2.2.1.2. *Generalized characterization of $\mathcal{M}$.* The class of Σ_1 measures can be characterized as the class of all effective transformations of—in place of λ—any chosen Δ_1 measure μ that is *continuous* (2.1.1.7 above). Thus proposition 2.5 is an instance of

PROPOSITION 2.10 (Levin). *For every continuous Δ_1 measure μ,*

$$\{\mu_M\}_M = \mathcal{M}.$$

PROOF. See B.1.1.

2.2.1.3. *Universal elements and Solomonoff-Levin measures.* We have seen that a universal uniform transformation λ_U, i.e., a Solomonoff-Levin measure Q_U, is a universal Σ_1 element (2.1.3.9 above). The converse is not true: not every universal element is also a Solomonoff-Levin measure. For instance, whereas one can easily define a universal Σ_1 measure $\mathring{\nu}$ such that $\sum_{\boldsymbol{x}^t \in \mathbb{B}^t} \mathring{\nu}(\boldsymbol{x}^t) = 1$ for all t up to a particular s, it must be the case that for any length t the sum of $\lambda_U(\boldsymbol{x}^t)$ for all sequences $\boldsymbol{x}^t$ of this length must fall short of 1. (The latter is easiest to see from the *mixture* representation of the Solomonoff-Levin measure, 2.2.2 below, and the fact that for *some* Σ_1 measures this will be the case. Also see Wood et al., 2013, 423f for a different proof.)

2.2.1.4. *Universal transformations of continuous Δ_1 measures.* A natural question is whether a universal transformation μ_U of a given continuous Δ_1 measure μ is also a universal element. This is not obvious: but the answer turns out to be positive (subject to a minimal compatibility condition on the universal machine U, as explained in B.1.5.2). In fact, any such universal transformation μ_U is of the form $\lambda_{U'}$ for some universal machine U', i.e., is a Solomonoff-Levin measure.

Proposition 2.11. For every continuous Δ_1 measure μ, and universal monotone machine U compatible with μ,

$$\mu_U \in \mathcal{SL}.$$

Proof. See B.1.5.

2.2.1.5. **A diagonal argument that fails.* The crucial step in the proof of proposition 2.11 is an application of Kleene's (*second*) *recursion theorem* (or *fixed point theorem*; 1938), that (applied to monotone machines) states that for given acceptable enumeration $\{M_e\}_e$ of all monotone machines, for every computable function f there is a fixed point $\hat{e}$ with $M_{f(\hat{e})} = M_{\hat{e}}$. Kleene actually arrived at this result in an attempt to diagonalize the class of p.c. functions (see Soare, 2016, 29), and the most instructive proof presents it as a "diagonal argument that fails" (Owings, 1973; Soare, 2016, 29f).

2.2.1.6. **Universal elements via non-universal transformations.* Universal transformations yield universal elements, but universal elements can also be obtained via transformations that are not universal. Indeed, every Σ_1 measure can be obtained as a transformation by a machine that is not even weakly universal.

Proposition 2.12. For every continuous Δ_1 measure μ, there is for every Σ_1 measure ν a non-(weakly) universal M such that $\nu = \mu_M$.

Proof. See B.1.3.

2.2.1.7. *Generalized characterization of $\mathcal{SL}$.* As the converse to proposition 2.11 we have the result that, for every continuous Δ_1 measure μ, any given Solomonoff-Levin measure can also be obtained as a universal transformation of *this* measure. Taken together we have the following generalized characterization of the class of Solomonoff-Levin measures.

Theorem 2.13. *For every continuous Δ_1 measure μ,*

$$\{\mu_U\}_U = \mathcal{SL}.$$

Proof. See B.1.6.

Thus, for any given continuous Δ_1 measure, a Solomonoff-Levin measure can also be interpreted as giving the probabilities for finite sequences being generated by some universal machine that is presented with a stream of bits sampled from *this* measure.[8]

2.2.2. The Σ_1 mixtures.

2.2.2.1. *Mixtures.* Let $\mathcal{H}$ be a countable class of probability measures over Cantor space. A *mixture* over $\mathcal{H}$ is simply a measure that is a weighted average over all measures in $\mathcal{H}$. To make this more precise, let a *weight function* $w : I \to [0,1]$ be a distribution over an index set I, so $\sum_i w(i) = 1$, that is everywhere positive, so $w(i) > 0$ for all $i \in I$. Then the mixture via weight function w over a particular enumeration $\{\mu_i\}_{i \in I} = \mathcal{H}$ is the w-weighted average over this enumeration of $\mathcal{H}$.

Definition 2.14. The mixture $\xi_w^{\{\mu_i\}_i}$ with weight function w over enumeration $\{\mu_i\}_{i \in I}$ of measures is given by

$$\xi_w^{\{\mu_i\}_i}(\boldsymbol{x}) = \sum_{i \in I} w(i)\mu_i(\boldsymbol{x}).$$

It is often convenient to leave the particular enumeration implicit and simply write '$\xi_w^{\mathcal{H}}$' for $\xi_w^{\{\mu_i\}_i}$.

2.2.2.2. *Universality.* A mixture over $\mathcal{H}$ clearly *dominates* every element of $\mathcal{H}$ (2.1.3.1 above): for every $\mu_i \in \mathcal{H}$ it holds that for every $\boldsymbol{x}$

$$\xi_w^{\{\mu_i\}_i}(\boldsymbol{x}) \geq w(i)\mu_i(\boldsymbol{x}). \tag{21}$$

Thus, whenever a mixture over $\mathcal{H}$ is itself still an element of $\mathcal{H}$, it is a *universal* element of $\mathcal{H}$.

2.2.2.3. Σ_1 *mixtures.* Perhaps the most straightforward example of a universal Σ_1 measure is an effective mixture over $\mathcal{M}$ (see Li and Vitányi, 2008, 294ff). Such a mixture is defined by an effective enumeration $\{\nu_i\}_{i \in \mathbb{N}}$ of $\mathcal{M}$ and a weight function over index set $I = \mathbb{N}$ that is also Σ_1 or semi-computable—so this mixture is itself Σ_1 and therefore a universal Σ_1 element. It is actually customary (see Hutter, 2007, 35) to allow the weight function here to be *defective*: it is an everywhere positive *semi-distribution* v that merely satisfies $\sum_{i \in \mathbb{N}} v(i) \leq 1$ (also see 2.2.2.6 below). I will simply call such an effective mixture over all of $\mathcal{M}$ a Σ_1 *mixture.*

Definition 2.15. The Σ_1 mixture $\xi_v^{\{\nu_i\}_i}$ with (defective) Σ_1 weight function v over enumeration $\{\nu_i\}_{i \in \mathbb{N}}$ of $\mathcal{M}$ is given by

$$\xi_v^{\{\nu_i\}_i}(\boldsymbol{x}) = \sum_{i \in \mathbb{N}} v(i)\nu_i(\boldsymbol{x}).$$

2.2.2.4. *The representation theorem.* The Solomonoff-Levin measures and the Σ_1 mixtures are all universal elements, hence every Solomonoff-Levin measure is equivalent to every Σ_1 mixture up to a multiplicative constant (2.1.3.2 above). Wood et al. (2013) have shown that the Solomonoff-Levin measures are in fact *precisely* the Σ_1 mixtures. I call this result a *representation theorem*, for reasons laid out more carefully in 3.2.4 below.

THEOREM 2.16 (Representation theorem, Wood et al.). $\mathcal{SL} = \{\xi_v^{\mathcal{M}}\}_v$.

PROOF. See B.1.4.

2.2.2.5. *Some refinements.* The statement of theorem 2.16 leaves room for improvement in two respects. First, it is not made explicit what effective enumerations of $\mathcal{M}$ are assumed to be included. It turns out that we can state the same result for every single fixed enumeration of $\mathcal{M}$ that is *acceptable* (a notion analogous to acceptable numberings of the p.c. functions, Rogers, 1967, 41, see B.1.8). Second, there is the funny notion of defective weight function in the definition of Σ_1 mixtures. This we can do away with: we can state the same result with mixtures using non-defective Δ_1 weight functions only. In short, the strengthened statement follows from

PROPOSITION 2.17. For every acceptable enumeration $\{\nu_i\}_i$ of $\mathcal{M}$, every $Q_U \in \mathcal{SL}$ equals $\xi_w^{\{\nu_i\}_i}$ for some Δ_1 weight function w.

PROOF. See B.1.8.

2.2.2.6. **Universal weight functions.* The admission of defective weight functions, i.e., Σ_1 semi-distributions, opens the way for elements $\mathring{v}$ that are *universal*: for every (defective) Σ_1 weight function v it holds that $\exists c \forall i.\ \mathring{v}(i) \geq c^{-1} v(i)$. Perhaps surprisingly, the following result shows that *every* Σ_1 mixture can be represented so as to have a universal weight function.

PROPOSITION 2.18. For every acceptable enumeration $\{\nu_i\}_i$ of $\mathcal{M}$, every $Q_U \in \mathcal{SL}$ is equal to $\xi_{\mathring{v}}^{\{\nu_i\}_i}$ for some universal Σ_1 weight function $\mathring{v}$.

PROOF. See B.1.9.

2.2.2.7. *The generalized representation theorem.* Taking the main results of 2.2.1 and 2.2.2 together, we obtain the following general characterization of the Solomonoff-Levin measures:

THEOREM 2.19. *For every continuous Δ_1 measure μ and every acceptable enumeration $\{\nu_i\}_i$ of $\mathcal{M}$,*

$$\mathcal{SL} = \{\mu_U\}_U = \{\xi_w^{\{\nu_i\}_i}\}_w,$$

with the U ranging over those universal machines compatible with μ and the w ranging over the Δ_1 weight functions.

*

CHAPTER 3

Perspectives on prediction methods

This chapter lays out different formal perspectives on prediction methods. These formal perspectives are associated with different conceptual interpretations of prediction methods. The systematic arrangement of these different perspectives serves to clarify different approaches and goals in sequential prediction. Specifically, it facilitates the appraisal of Solomonoff's theory of universal prediction in this and the next chapters.

In 3.1, I discuss the formal equivalence between the perspectives of prediction methods and a priori measures. This shows that prediction methods are associated with a Bayesian a priori assessment of all possible data sequences. In 3.2, I discuss the perspective of mixtures over measures. This is associated with the classical Bayesian interpretation of an epistemic prior over a model of the possible data-generating mechanisms. In 3.3, I discuss the perspective of mixtures over prediction methods. This is associated with the view of prediction methods as aggregating strategies over a pool of competing predictors.

Innovations. None of the discussed perspectives on methods for sequential prediction is new, and I rely on various sources in my presentation of them, but I believe this chapter for the first time brings all of these together. In particular, a main contribution of this thesis is a detailed analysis of the different possible interpretations of the Solomonoff-Levin predictor. This allows for a novel appraisal of several aspects of the Solomonoff-Levin predictor, including its universality and its objectivity. An important strand here is again the parallels drawn between the Solomonoff-Levin proposal and Carnap's inductive logic, which among other things prompts the interpretation of theorem 2.16 as a representation theorem. This in turn points at the interpretation of the Solomonoff-Levin predictor as operating under a particular inductive assumption, and the observation that the choice of universal machine in the definition is precisely the choice of effective Bayesian prior. The chapter contains one mathematical result, theorem 3.2 on the non-convergence of different Solomonoff-Levin predictors, that—although derivable from a minor modification of an existent construction—appears to be novel. (Parts of this chapter are based on parts of Sterkenburg, 2016.)

3.1. Prediction methods and a priori measures

This section is on the formal equivalence between methods for sequential prediction and measures on Cantor space.

In 3.1.1, I discuss the equivalence between prediction methods and probability measures, and the Bayesian interpretation of the latter. In 3.1.2, I illustrate this correspondence by the Johnson-Carnap predictors. In 3.1.3, I distinguish the Solomonoff-Levin measures and predictors. In 3.1.4, I discuss the correspondence between *effective* prediction methods and probability measures. In 3.1.5, I assess the Solomonoff-Levin measures as a priori measures.

3.1.1. Conditional and joint distributions. There is a straightforward formal equivalence between prediction rules and full measures on all possible data sequences. In essence, this is the equivalence between conditional and joint distributions. (Also see Dawid, 1984, 279; Merhav and Feder, 1998, 2127; Cesa-Bianchi and Lugosi, 2006, 247f.)

3.1.1.1. *From prediction rules to measures.* Recall that a prediction method is a function $\mathsf{p} : \mathbb{B}^* \to \mathcal{P}$ from past data sequences to *predictions*, distributions over $\mathbb{B}$. (I also use the shorthand $\mathsf{p}(x, \boldsymbol{x}) := \mathsf{p}(\boldsymbol{x})(x)$, and $p = (a_0, a_1)$ for $p(0) = a_0$ and $p(1) = a_1$.) A prediction rule p directly defines a *one-step conditional measure* $\mu^1_{\mathsf{p}}(\cdot \mid \cdot)$ by

$$\mu^1_{\mathsf{p}}(\cdot \mid \boldsymbol{x}) := \mathsf{p}(\boldsymbol{x}), \tag{22}$$

i.e., $\mu^1_{\mathsf{p}}(x \mid \boldsymbol{x}) = \mathsf{p}(x, \boldsymbol{x})$ for $x \in \mathbb{B}$. (Also see 2.1.1.3 above on notation of (one-step) conditional measures in sequential prediction.) The one-step conditional measure $\mu^1_{\mathsf{p}}(\cdot \mid \cdot)$ defines the measure μ_{p} on finite sequences by the product rule for conditional probabilities,

$$\mu_{\mathsf{p}}(\boldsymbol{x}^t) := \prod_{s=0}^{t-1} \mu^1_{\mathsf{p}}(x_{s+1} \mid \boldsymbol{x}^s). \tag{23}$$

Thus a prediction rule determines a probability assignment to all finite data sequences (cf. Zabell, 2011, 276). (More properly speaking, the one-step conditional measure defines via (23) a pre-measure m_{p} on the basic cylinders $[\![\boldsymbol{x}]\!] = \{\boldsymbol{x}^\omega : \boldsymbol{x} \prec \boldsymbol{x}^\omega\}$ for all $\boldsymbol{x} \in \mathbb{B}$, that extends to an actual measure μ_{p} on the Cantor space $\mathbb{B}^\omega$ with $\mu_{\mathsf{p}}([\![\boldsymbol{x}]\!]) = m_{\mathsf{p}}(\boldsymbol{x})$ as in 2.1.1.1 above.)

3.1.1.2. *From measures to prediction rules.* A measure μ defines a one-step conditional measure by

$$\mu^1(x \mid \boldsymbol{x}) := \frac{\mu(\boldsymbol{x}x)}{\mu(\boldsymbol{x})}, \tag{24}$$

hence a predictor p_μ by

$$\mathsf{p}_\mu(\boldsymbol{x}) := \mu^1(\cdot \mid \boldsymbol{x}). \tag{25}$$

Note, however, that (24) and hence (25) is undefined where $\mu(\boldsymbol{x}) = 0$. (Partially defined predictor p_μ "goes into a coma" when such $\boldsymbol{x}$ obtains, Kelly, 1996, 305.) That means that, strictly speaking, the formal equivalence only holds for measures μ that assign positive probability to all cones: the strictly positive measures (2.1.1.5 above; also see 3.1.2.2 below).

3.1.1.3. *Semi-measures.* The formal correspondence easily generalizes for semi-measures, or measures on the extended space $\mathbb{B}^\omega \cup \mathbb{B}^*$ (2.1.2.9, 2.1.2.10 above). Such a measure corresponds to a prediction method that maps finite sequences to semi-distributions, i.e., q with $\sum_{x \in \mathbb{B}} q(x) \leq 1$.

3.1.1.4. *Two perspectives.* The formal correspondence shows that every prediction strategy comes with an *a priori* measure that models all possible outcomes (Dawid and Vovk, 1999, 128), or, in explicitly Bayesian terms, gives a full specification of our beliefs over all possibilities. In yet different terms, an a priori distribution represents our *inductive assumption*, that is implemented by the corresponding prediction method (Howson, 2000; 3.2.1.2 below). We thus have a correspondence between, on the one hand, the "operational" perspective of direct prediction rules, on the other, the "metaphysical" perspective of a priori measures over all data sequences (Skyrms, 1996, 323).

3.1.1.5. *Bayes's rule.* As a Bayesian, you start out with an a priori measure μ that expresses your beliefs over all possibilities (which I could elicit, for instance, by testing what bets on the outcomes you are willing to take). How do you make sense of the induced prediction rule p_μ? It is the strategy you would follow if you were to adhere to *Bayes's rule* for adjusting beliefs in the light of evidence. (And, of course, if your predictions are in fact honest reports of your beliefs.) Namely, if we denote by μ_t the measure that expresses your beliefs at the start of trial $t+1$, after having observed $\boldsymbol{x}^t$ (in particular, $\mu_0 = \mu$), and x_{t+1} is revealed at $t+1$, then Bayes's rule says that

$$\mu_{t+1} := \mu_t(\cdot \mid x_{t+1}). \tag{26}$$

Equivalently,

$$\mu_{t+1} := \mu_0(\cdot \mid \boldsymbol{x}^{t+1}). \tag{27}$$

In words, Bayes's rule says that your beliefs after seeing a particular data sequence should be equal to your original beliefs conditional on this data sequence. This is in accordance with p_μ because by definition $\mathsf{p}_\mu(x, \boldsymbol{x}^t) = \mu_0(x \mid \boldsymbol{x}^t)$.

3.1.2. The Johnson-Carnap predictors. In Carnap's terminology, the formal correspondence between prediction methods p and a priori measures μ is the formal correspondence between confirmation functions $\mathfrak{c}$ and measure functions $\mathfrak{m}$.

In the *Logical Foundations* (1950), Carnap presents confirmation functions as induced from measure functions. The value $\mathfrak{m}(h)$ equals the *null confirmation* $\mathfrak{c}_0(h)$ of h, "the degree of confirmation of h before any factual information is available" (ibid., 308), which he allows might be called the "initial probability" or "the probability a priori" of the sentence. The full confirmation function $\mathfrak{c}$ follows from conditionalizing $\mathfrak{m}$ as in 3.1.1.2 above. In the *Continuum* (1952) and later (see 1963a; 1971b), Carnap defines the confirmation functions $\mathfrak{c}_\alpha$ directly, only noting that "the $\mathfrak{c}$-values for any sentences are reducible ... to $\mathfrak{c}$-values with respect to the tautological evidence 't' ('probability a priori' in the classical terminology). For the latter values we introduce the notation '$\mathfrak{m}$'" (1952, 16).

We find Carnap's last and most detailed account of the interpretation of $\mathfrak{m}$ and $\mathfrak{c}$ in (1971a). Briefly: the "purely logical" functions $\mathfrak{m}$ and $\mathfrak{c}$ correspond to the "quasi-psychological" concepts of a rational *initial credence function* Cr_0 and a rational *credibility function* Cred, respectively; these "are assigned to an imaginary subject X supposed to be equipped with perfect rationality and an unfailing memory" (ibid., 25)—or indeed "a robot" (ibid., 17). The *credence function* Cr_t gives X's beliefs at time t, and "evolved from Cr_0 by the effect of the data" (or Cr_0 is "the credence function we originally build in" our robot and that "he transforms step by step ... into the later credence function," ibid., 18). The credence is indeed the *conditional initial credence* $\mathrm{Cr}_0(\cdot \mid \cdot)$—here we have Bayes's rule (27). Alternatively (ibid.), we can put things in terms of the credibility function $\mathrm{Cred}(\cdot, \cdot)$ that equals the conditional initial credence $\mathrm{Cr}_0(\cdot \mid \cdot)$. ("While Cr_t characterizes the *momentary state* of X at time t with respect to his beliefs, his function Cred is *a trait of his underlying permanent intellectual character*, namely his permant disposition for forming beliefs on the basis of his observations," ibid., 19, slight change of notation.) Finally, on the formal correspondence: "Since each of the two functions Cr_0 and Cred is definable on the basis of the other, there are two alternative procedures for specifying a basic belief-forming disposition, namely, either by Cr_0 or by Cred" (ibid., 21).

That concludes Carnap's own view. In the rest of this section, I revisit the Johnson-Carnap functions for the purpose of giving some further illustration of the formal correspondence between the two perspectives of prediction rules and a priori measures. I also switch from Carnap's notation to the standard notation in this thesis.

3.1.2.1. *The indifferent predictor.* The most straightforward example of the correspondence is the *indifferent* predictor defined by $\mathsf{p}(\boldsymbol{x}) = (\frac{1}{2}, \frac{1}{2})$ for every $\boldsymbol{x}$ (Carnap's function $\mathfrak{c}\dagger$, 1.2.3.3 above) and the uniform measure λ. On a more conceptual level, the prediction rule that always expresses indifference between the next two possible symbols indeed corresponds to the a priori measure that expresses indifference between all same-length sequences. Recall that Carnap objects to this rule for the reason that it never learns from data: this is a defect that is shared by all of the prediction rules that correspond to an i.i.d. measure μ_θ.

3.1.2.2. *The straight rule.* To see the breakdown of the correspondence when probabilities 0 are involved, consider again the straight rule (1.2.3.6 above),

$$\mathsf{p}(\boldsymbol{x}) = \left(\frac{\#_0 \boldsymbol{x}^t}{t}, \frac{\#_1 \boldsymbol{x}^t}{t} \right).$$

This rule is actually only partially defined from the start, because $\mathsf{p}(x, \varnothing)$ involves division over $t = 0$; but we can just stipulate $\mathsf{p}(\varnothing) = (\frac{1}{2}, \frac{1}{2})$, say. Moreover, on the first outcome it immediately gives extreme predictions, namely

$$\mathsf{p}(0) = (1, 0); \ \mathsf{p}(1) = (0, 1),$$

which means that the corresponding conditional measure induces via (23) the measure μ with

$$\mu(\boldsymbol{x}^t) = \begin{cases} \frac{1}{2} & \text{if } \boldsymbol{x}^t = 0^t \text{ or } \boldsymbol{x}^t = 1^t; \\ 0 & \text{otherwise.} \end{cases} \tag{28}$$

But if we translate *this* measure back into a conditional measure and hence a predictor, it will be undefined on every input that is not a sequence of identical symbols. Carnap (1952, 42) diagnoses the "inadequacy" of the straight rule as its failure to be a *regular* confirmation function: a confirmation function that (is derived from a measure function that) gives positive probability to every logical possibility (Carnap, 1950, 294f; Carnap, 1963a, 974f; 1.2.2.2 above). In our setting, the requirement that every possibility should have positive probability (*Cromwell's rule*, see Lindley, 2006, 91) is satisfied by (a predictor corresponding to) a strictly positive measure (see 2.1.1.5 above).

3.1.2.3. *Exchangeability.* As we saw in 1.2.3.1 above, exchangeability is a property of measures that straightforwardly translates into a similar property of prediction rules. A measure μ is exchangeable if the probability values are invariant under finite permutations of outcomes, i.e., $\mu(\boldsymbol{x}^t) = \mu(\boldsymbol{y}^t)$ if $\#_0\boldsymbol{x}^t = \#_0\boldsymbol{y}^t$. Equivalently, the predictive probability $\mathsf{p}_\mu(x, \boldsymbol{x})$ only depends on $\#_x\boldsymbol{x}$. Thus the assessment that the predictive probability of x only depends on the number of earlier occurrences of x is equivalent to the assessment that all data sequences with the same number of 0's and 1's are equally probable. (Also see Skyrms, 1996, 324.)

3.1.2.4. *Exchangeability and structure-symmetry.* Carnap's stipulation of structure-symmetry (1.2.3.2 above) says that for each t, each number of 1's—a total of $t + 1$ possibilities—is equally likely. The number of sequences $\boldsymbol{x}^t$ with t_1 1's is

$$(t, t_1) = \frac{t!}{t_1!(t - t_1)!} = \frac{t}{t_0!t_1!},$$

which under the stipulation of exchangeability are also all equally likely. Thus the probability of a particular sequence $\boldsymbol{x}^t$ under both exchangeability and structure-symmetry (the probability according to Carnap's $\mathfrak{m}^*$) is

$$\begin{aligned} \mu(\boldsymbol{x}^t) &= \frac{1}{(t+1)} \frac{1}{\frac{t!}{(\#_0\boldsymbol{x}^t)!(\#_1\boldsymbol{x}^t)!}} \\ &= \frac{(\#_0\boldsymbol{x}^t)!(\#_1\boldsymbol{x}^t)!}{(t+1)!}. \end{aligned} \tag{29}$$

(Also see Zabell, 2011, 273f.) We saw earlier that Carnap's $\mathfrak{c}^*$ is the prediction method

$$\mathsf{p}(\boldsymbol{x}^t) = \left(\frac{\#_0\boldsymbol{x}^t + 1}{t+2}, \frac{\#_1\boldsymbol{x}^t + 1}{t+2} \right). \tag{30}$$

The measure (29) and predictor (30), while equivalent in the sense I have been discussing, take forms that, on a first glance, are quite different; this is the case for the λ-γ continuum in general.

3.1.2.5. *The λ-γ predictors and measures.* As we saw in 1.2.3.7 above, exchangeability with the further assessment that the predictive probability of x is linear in the number of earlier occurrences of x, so that $\mathsf{p}(x, \boldsymbol{x}) = a_x + b\#_x\boldsymbol{x}$ for some a_x, b, gives the rules of succession of the pleasing form

$$\mathsf{p}_{\lambda,\gamma}(\boldsymbol{x}) = \left(\frac{\#_0\boldsymbol{x}^t + \gamma_0\lambda}{t+\lambda}, \frac{\#_1\boldsymbol{x}^t + \gamma_1\lambda}{t+\lambda} \right). \tag{31}$$

Using the product rule for conditional probabilities (23), we can calculate that these predictors correspond to measures of the form (cf. Zabell, 1982, 1095)

$$\begin{aligned}
\mu_{\lambda,\gamma}(\boldsymbol{x}^t) &= \frac{\prod_{s=0}^{t-1}(\#_{x_{s+1}}\boldsymbol{x}^s + \gamma_{x_{s+1}}\lambda)}{\prod_{s=0}^{t-1}(s+\lambda)} \\
&= \frac{\prod_{j=0}^{\#_0\boldsymbol{x}^t-1}(j+\gamma_0\lambda)\prod_{j=0}^{\#_1\boldsymbol{x}^t-1}(j+\gamma_1\lambda)}{\prod_{s=0}^{t-1}(s+\lambda)} \\
&= \frac{\Gamma(\lambda)}{\Gamma(t+\lambda)} \frac{\Gamma(\#_0\boldsymbol{x}^t + \gamma_0\lambda)}{\Gamma(\gamma_0\lambda)} \frac{\Gamma(\#_1\boldsymbol{x}^t + \gamma_1\lambda)}{\Gamma(\gamma_1\lambda)} \\
&= \frac{\Gamma(\lambda)}{\Gamma(\gamma_0\lambda)\Gamma(\gamma_1\lambda)} \frac{\Gamma(\#_0\boldsymbol{x}^t + \gamma_0\lambda)\Gamma(\#_1\boldsymbol{x}^t + \gamma_1\lambda)}{\Gamma(t+\lambda)}. \qquad (32)
\end{aligned}$$

(Here the gamma function Γ is the generalization of the factorial function to real numbers, in the sense that $\Gamma(n) = (n-1)!$ for $n \in \mathbb{N}$.) This is a rather more complicated expression than the direct expression (31) as a prediction rule. However, the measure does admit of a different natural expression (traces of which we can already recognize here), namely as a *mixture* over the class of i.i.d. measures (where the first fraction in (32) returns in the *Beta prior* (38) that weighs the mixture): see 3.2.3 below.

3.1.3. The Solomonoff-Levin predictors.

3.1.3.1. *Solomonoff.* He writes (1964, 7):

> It is possible to devise a complete theory of inductive inference using Bayes's Theorem, if we are able to assign an a priori probability to every conceivable sequence of symbols. In accord with this approach, it is felt that sequences should be given high a priori probabilities if they have short descriptions and/or many different descriptions.

The use of Bayes's theorem that Solomonoff mentions is that

$$(33) \qquad Q_U(x \mid \boldsymbol{x}) = Q_U([\![\boldsymbol{x}x]\!] \mid [\![\boldsymbol{x}]\!]) = \frac{Q_U([\![\boldsymbol{x}]\!] \mid [\![\boldsymbol{x}x]\!]) \cdot Q_U([\![\boldsymbol{x}x]\!])}{Q_U([\![\boldsymbol{x}]\!])} = \frac{Q_U(\boldsymbol{x}x)}{Q_U(\boldsymbol{x})}$$

—or so goes the reconstruction of Li and Vitányi (2008, 349f). It comes down to the use of conditionalization (25) to turn an a priori measure into a predictor: the use of Bayes's rule (27) in a Bayesian interpretation. I will call a one-step conditional Solomonoff-Levin measure $Q^1_U(\cdot \mid \cdot)$ defined as in (33) a *Solomonoff-Levin predictor* p_{Q_U}.

3.1.3.2. *The Solomonoff-Levin measures and predictors.* In modern presentations of Solomonoff's theory the same route is taken: first the introduction of the "universal a priori measure" or the "universal a priori probability" (ibid., 302; the Solomonoff-Levin measure of definition 2.8 in 2.1.4 above), and next the observation that we can use this measure for prediction by conditionalization (ibid., 349ff; the Solomonoff-Levin predictor (33)). (Strictly speaking, of course, the correspondence is as in 3.1.1.3 above: Q_U is a Σ_1 measure on $\mathbb{B}^\omega \cup \mathbb{B}^*$ or semi-measure on $\mathbb{B}^\omega$, hence p_{Q_U} is a method that issues semi-distributions for predictions.) The Solomonoff-Levin *measure* Q_U has a natural definition in terms of the inputs to a universal Turing machine or a universal effective transformation; the Solomonoff-Levin *predictor* p_{Q_U}, in contrast, does not appear to admit of a natural direct definition as a prediction rule, other than simply as a fraction of a priori probabilities. (To qualify: there is an alternative representation as a mixture predictor, 3.2.4.1 below, but this still consists in taking a mixture of fractions of a priori probabilities.)

3.1.3.3. *Levin.* He was interested in formalizing complexity and randomness (I.6 above), and not so much in prediction (see Levin, 1984, 23; Solomonoff, 1997, 83, 2003, 598). The randomness branch—the main branch—of algorithmic information theory is based on versions of the Solomonoff-Levin measure to quantify complexity or information content (more precise, the central notion is the *negative logarithm* of versions of the Solomonoff-Levin measure, see A.3). It has no need of (versions of) the *conditional* Solomonoff-Levin measure that is central in Solomonoff's predictive branch. As should become more clear later in this thesis (see, in particular, 5.2.1.8 and 6.2.2.8 below), it has a certain elegance that the predictive theory fails to retain. This has much to do with the breakdown of the *effective* correspondence between measures and predictors, 3.1.4 below.

3.1.3.4. *An a priori measure.* When he does briefly indicate the relation to inductive inference, Levin positions the Solomonoff-Levin measure as follows (1970, 104, my notation):

> In mathematical statistics the following problem arises: to clarify with respect to what measure a given sequence can be obtained "randomly". If nothing is known in advance about the properties of the sequence, then the only (weakest) assertion we can make regarding it is that it can be obtained randomly with respect to

> Q_U. Thus, Q_U corresponds to what we intuitively understand by the words "a priori probability".

I explain this more fully in A.4.2.6 on randomness: basically, it is an appeal to the *universality* of the Solomonoff-Levin measure. I come back to this shortly, in 3.1.5 below.

3.1.4. Effective measures and predictors. The correspondence of 3.1.1 persists when effectivized to the Δ_1 or computable elements: the Δ_1 measures correspond precisely to the Δ_1 predictors. If a predictor p, hence the corresponding one-step conditional measure (22), is Δ_1, then so is the product (23), that gives the measure μ_{p}. Conversely, if a measure μ is Δ_1, then so is the fraction (24) that gives the one-step conditional measure (25), hence so is the corresponding predictor p.

Importantly, however, the correspondence of 3.1.1 does *not* persist when restricted to the Σ_1 elements: it is *not* the case that the Σ_1 measures correspond to the Σ_1 predictors. All is well in one direction: if a predictor p is Σ_1, then so is the product and hence the measure μ_{p}. But a *fraction* $m(\cdot) = \frac{m_1(\cdot)}{m_2(\cdot)}$ of two functions m_1 and m_2 that are both Σ_1 need not itself be Σ_1. The two approximating functions g_1 and g_2 for m_1 and m_2, respectively, give rise to an approximating function $g(\boldsymbol{x}, s) = \frac{g_1(\boldsymbol{x},s)}{g_2(\boldsymbol{x},s)}$; but this function, while it satisfies $\lim_{s\to\infty} g(\boldsymbol{x}, s) = m(\boldsymbol{x})$, does *not* need to satisfy $g(\boldsymbol{x}, s) \leq g(\boldsymbol{x}, s+1)$ for all $s \in \mathbb{N}$. For such a function we do not even know whether any given approximation is a lower aproximation, and whether the approximation at $s + 1$ will be at least as accurate as the one at s. In technical terms, the function m is only *limit-computable* or Δ_2 (see Soare, 2016, 63ff). Thus, a predictor corresponding to (i.e., a fraction of terms of) a Σ_1 measure is Δ_2, but need no longer be Σ_1. In particular, the Solomonoff-Levin *predictor* is no longer Σ_1. A proof of this fact is deferred to the next chapter (specifically, 4.3.3), where I will discuss its conceptual implications in detail.

3.1.5. The Solomonoff-Levin measures as a priori measures. The Solomonoff-Levin measures, like Carnap's admissable measure functions or initial confirmation functions, are proposed as measures that are 'a priori' in the sense of a starting point that precedes any input from observation data. Carnap talks about the distribution of an "idealized human baby" that is subsequently updated through contingent life experience (1971a, 17; 3.1.2 above); a similar view is expressed by Solomonoff (1997, 75). Not just any measure qualifies as a proper such starting point: Carnap sought to isolate a logical hence rational and objective a priori measure. Why does the Solomonoff-Levin measure make for a proper starting point?

3.1.5.1. *"The general intuitive basis."* Carnap proceeded by formulating invariance axioms that served as constraints on rational measure functions, where both the axioms themselves and the values given by the resulting confirmation functions are continuously tested against our inductive intuitions.

Solomonoff likewise states that the "validity" of his proposed definition is both tested by "the application of [the resulting method] to specific problems and comparison of the results with intuitive evaluations," and—although "of much less importance"—supported by the "general intuitive basis" for the definition itself (1964, 5). As for the latter, this starts with the intuition of assigning a priori probabilities to finite sequences by "examin[ing] the manner in which these strings might be produced by a universal Turing machine" (ibid., 3).

3.1.5.2. *Objectivity through computation.* At least part of the motivation appears to be an intuition about objectivity deriving from *computation*, as illustrated by Solomonoff's recollection of a conversation he had with McCarthy at the Dartmouth conference (Solomonoff, 1997, 76):

> I asked him about the induction problem: "Suppose you are given a long sequence of symbols describing events in the real world. How can you extrapolate that sequence?"
>
> The next day he said, "Suppose we were wandering about in an old house, and we suddenly opened a door to a room and in that room was a computer that was printing out your sequence. Eventually it came to the end of the sequence and was about to print the next symbol. Wouldn't you bet that it would be correct?"

This is a vague intuition, that does not begin to answer the basic problems of prediction (e.g., "There may be a large number of inputs to the machine that give the same initial output but extrapolate differently. Which should we use?," ibid.). But it is this intuition that is subsequently made more precise in the idea of a distribution on output sequences based on the length of the required input sequences. (It does not seem out of place here to note that a same basic intuition of objectivity through computation—objective algorithms!—is what drives popular conceptions about the reach of *big data* and purely data-driven inference today: see Mayer-Schönberger and Cukier, 2012—"let the data speak!"—for a representative account.)

3.1.5.3. *"The general intuitive basis," cont.* The inputs to the data-generating machine are sometimes (and without much context) characterized as the "explanations" or "causes" of the data (see Solomonoff, 1964, 19; Li and Vitányi, 1992a, 5; 2008, 260), with the likelier causes those that are shortest or *simplest.* The role of lengths of inputs in the definition is generally associated with data-compression and simplicity (*Occam's razor*): this is the topic of chapter 5. Moreover, a "suggested point of support" (Solomonoff, 1964, 4) is a version of the principle that was central to Carnap's approach, the principle of indifference (1.2.1.3 above).

3.1.5.4. *The principle of indifference.* Solomonoff (ibid., 19) writes that

> If we consider the input sequence to be the "cause" of the observed output sequence, and we consider all input sequences of a given length to be equiprobable (since we have no a priori reason

> to prefer one rather than any other) then we obtain the present model of induction.

(Also see Fine, 1973, 146ff; Li and Vitányi, 1992a, 5; Rathmanner and Hutter, 2011, 1119, "Epicurus's principle.") But this fact that "all inputs to a Turing machine that are of a given fixed length, are assigned '"indifferently equal a priori' likelihoods" (Solomonoff, 1964, 4) rests on the unique feature of the uniform measure λ that equal-length sequences are assigned equal probability. Theorem 2.13 shows that the choice of λ in defining the Solomonoff-Levin measure is only circumstantial, undermining this association.

3.1.5.5. *Universality.* In another place in his paper, Solomonoff says that "supposing that the string was created as the output of a universal machine on random input" is the "*optimum* manner" to account for a sequence (1964, 14, 16; also see 1966, 1191f):

> By "optimum manner" it is meant that the model we are discussing is at least as good as any other model of the universe in accounting for the sequence in question.

Here we encounter the idea of *universality.* As I hope to make clear in this thesis, the best way of looking at the Solomonoff-Levin measure is not so much as an objective starting point in the sense of a (or indeed *the*) rational Carnapian measure function, where the "validity" of the definition rests on considerations of symmetry or simplicity. Rather, the Solomonoff-Levin measure should be seen as an attempt at a *universal* starting point, a "universal a priori probability" (Li and Vitányi, 2008, 302) that in some sense captures all possibilities (Li and Vitányi, 1992a, 5):

> we call [the Solomonoff-Levin measure] the *a priori* probability, since it assigns maximal probability to all hypotheses in absence of any knowledge about them.

* * *

3.2. Mixtures over measures

This section discusses measures and predictors as mixtures over a class of measures. This perspective relates to the classical Bayesian interpretation of a prior over a class of hypotheses.

In 3.2.1, I introduce mixture measures and the classical Bayesian interpretation. In 3.2.2, I introduce mixture predictors. In 3.2.3, I discuss de Finetti's representation theorem that shows that the exchangeable predictors (in particular, the Johnson-Carnap predictors) are the mixture predictors over the i.i.d. measures (with a particular form of prior). In 3.2.4, I discuss the representation theorem 2.16 that shows that the Solomonoff-Levin predictors are the Σ_1 mixture predictors. In 3.2.5, I discuss the element of subjectivity in the choice of universal machine or effective prior.

3.2.1. Mixture measures. Recall from definition 2.14 in 2.2.2.1 above that a mixture measure $\xi_w^{\{\mu_i\}_i}$ is the w-weighted mean

$$\xi_w^{\{\mu_i\}_i}(\cdot) = \sum_{i \in I} w(i)\mu_i(\cdot) \tag{34}$$

over a class $\mathcal{H} = \{\mu_i\}_{i \in I}$ of measures. Here the weight function w is an everywhere positive distribution over the indices I for the enumeration $\{\mu_i\}_{i \in I}$ of $\mathcal{H}$.

It was necessary to be careful about the enumeration in the statement of effective mixtures, but in general we can forget about these details and talk about w as a function that gives weights directly to the measures in $\mathcal{H}$ (even though strictly speaking w is only a function on indices that requires the specification of an enumeration of $\mathcal{H}$ to define (34)). Accordingly I write '$\xi_w^{\mathcal{H}}$' for a w-weighted mixture measure over $\mathcal{H}$. Moreover, in the following I will often talk about the weight function w as actually *determining* the class $\mathcal{H}$: the idea is that a class $\mathcal{H}$ is *induced* by w as the class of those measures that receive positive weight from w. Accordingly I write 'ξ_w' for a mixture measure over some $\mathcal{H}$ that I understand to be implicity given by w. The reason for approaching things in this way is that it is convenient to present a mixture measure as fully determined by a weight function or *prior* in the classical Bayesian interpretation below.

(Also for simplicity and with an eye to the class most relevant to us, the class of Σ_1 measures, I will stick in this exposition to *countable* classes $\mathcal{H}$.)

3.2.1.1. *The classical Bayesian perspective.* In 3.1.1.4 above, I introduced the Bayesian perspective of an epistemic a priori measure, as expressing the degree of belief we attach to every possible outcome sequence. This corresponds to what Diaconis and Freedman (1986, 11); Skyrms (1996, 323ff) call the *subjective* Bayesian perspective; though I hestitate to adopt that label here, because in my presentation it includes Carnap's interpretation of epistemic yet *objective* a priori measures. However, I will follow these authors in distinguishing this perspective from the *classical* Bayesian perspective, that posits a model or class of hypotheses about the actual data-generating mechanism, and an epistemic prior probability assignment to its members. This is the interpretation that naturally applies to mixture measures: the class $\mathcal{H}$ is the class of hypotheses, and the weight function w is a prior distribution over the hypotheses.

3.2.1.2. *The inductive assumption.* A mixture measure is an a priori measure that, under an Bayesian interpretation, expresses an inductive assumption (3.1.1.4 above). In the case of a mixture measure, under the classical Bayesian interpretation, there is additional structure in our inductive assumption: we believe that one of a number of hypotheses must actually govern the data (cf. Romeijn, 2004). We believe that one of the μ in $\mathcal{H}$ is the correct or *true* data-generating measure, to various degrees that are given by the prior w. To put it economically, in this interpretation, the prior distribution w encodes our inductive assumption. We also say that a mixture measure with prior w over class $\mathcal{H}$ expresses the inductive assumption of $\mathcal{H}$. The predictor corresponding to a mixture measure—a *mixture predictor*—is a prediction strategy that operates under the corresponding inductive assumption.

3.2.2. Mixture predictors. Given again a weight function w over a class $\mathcal{H} = \{\mu_i\}_i$ of measures. The *mixture strategy* $\mathsf{p}_{\text{mix}(w)}$ predicts at each trial by an appropriately updated mean of conditional probabilities, as follows. At each trial $t+1$, having observed sequence $\boldsymbol{x}^t$, we replace w with (or *update* w to) the weight function w_t defined by

$$w_{t+1}(i) := \frac{w(i)\mu_i(\boldsymbol{x}^t)}{Z}, \tag{35}$$

with normalizing term $Z = \sum_i w(i)\mu_i(\boldsymbol{x}^t) = \xi_w(\boldsymbol{x}^t)$. An equivalent expression of w_{t+1}, as an update of the previous weight function w_t, is

$$w_{t+1}(i) = \frac{w_t(i)\mu_i(x_t \mid \boldsymbol{x}^{t-1})}{Z}, \tag{36}$$

with normalizing term $Z = \sum_i w_t(i)\mu_i(x_t \mid \boldsymbol{x}^{t-1})$, and $w_0 = w$. The prediction at trial $t+1$, and so the definition of the mixture predictor p_{mix}, is given by the thus updated weighted mean of the conditional measures:

$$\mathsf{p}_{\text{mix}}(\boldsymbol{x}^t) := \sum_i w_{t+1}(i)\mu_i(\cdot \mid \boldsymbol{x}^t). \tag{37}$$

In the classical Bayesian interpretation, the mixture strategy predicts by a posterior-weighted mixture over objective hypotheses conditional on the data.

3.2.2.1. *The classical Bayesian model.* A perfectly precise classical Bayesian account calls for the specification of a full probability model over the hypotheses and the data sequences (cf. Grünwald, 2007, 74ff.). That is, we define a joint probability measure μ_{bayes} over the space $I \times \mathbb{B}^*$, the Cartesian product of the set I that parametrizes $\mathcal{H}$ and the set of finite outcome sequences. (Strictly speaking, the marginal measure on $\mathbb{B}^*$ defined as such is really a pre-measure on Cantor space.) We set the marginal probability of i to that given by the prior:

$$\mu_{\mathsf{bayes}}(i) = \mu_{\mathsf{bayes}}(i, \varnothing) := w(i),$$

and we set the conditional probability of $\boldsymbol{x}$ given i to the likelihood of μ_i:

$$\mu_{\mathsf{bayes}}(\boldsymbol{x} \mid i) := \mu_i(\boldsymbol{x}).$$

This suffices to define the joint distribution, for $\mu_{\mathsf{bayes}}(i, \boldsymbol{x}) = \mu_{\mathsf{bayes}}(\boldsymbol{x} \mid i) \cdot \mu_{\mathsf{bayes}}(i)$. Now, by Bayes's theorem, the *conditional prior* $w(\cdot \mid \boldsymbol{x}) = \mu_{\mathsf{bayes}}(\cdot \mid \boldsymbol{x})$ is given by

$$\begin{aligned} w(i \mid \boldsymbol{x}) = \mu_{\mathsf{bayes}}(i \mid \boldsymbol{x}) &= \frac{\mu_{\mathsf{bayes}}(\boldsymbol{x} \mid i)\mu_{\mathsf{bayes}}(i)}{\mu_{\mathsf{bayes}}(\boldsymbol{x})} \\ &= \frac{\mu_i(\boldsymbol{x})w(i)}{\sum_j \mu_j(\boldsymbol{x})w(j)}. \end{aligned}$$

This is the belief we attach to μ_i *conditional on* $\boldsymbol{x}$. Here we can again invoke *Bayes's rule*, that says that the conditional prior belief $w(\cdot \mid \boldsymbol{x}^t)$ should be our *posterior* belief in i after having seen $\boldsymbol{x}^t$ (cf. 3.1.1.5 above). We model this by employing time-indexed measures $\mu_{\mathsf{bayes},t}$ to express our beliefs at trial t, with $\mu_{\mathsf{bayes},t+1}(\boldsymbol{y} \mid i) := \mu_i(\boldsymbol{y} \mid \boldsymbol{x}^t)$ and

$$w_{t+1}(i) := w(i \mid \boldsymbol{x}^t).$$

This is the motivation for the update rule (35).

3.2.2.2. *The mixture predictor and mixture measure.* The update rule (35) is such that the mixture predictor $\mathsf{p}_{\mathrm{mix}(w)}$ indeed corresponds to the mixture *measure* ξ_w, in the sense of the correspondence between predictors and a priori measures in 3.1.1 above. That is, $\mathsf{p}_{\mathrm{mix}(w)}$ coincides with p_{ξ_w}, the predictor defined by the one-step conditional measure $\xi_w(\cdot \mid \cdot)$. Namely (see Grünwald, 2007, 77; Merhav and Feder, 1998, 2128f),

$$\begin{aligned}
\mathsf{p}_{\mathrm{mix}(w)}(x_{t+1}, \boldsymbol{x}^t) &= \sum_i w(i \mid \boldsymbol{x}^t)\mu_i(x_{t+1} \mid \boldsymbol{x}^t) \\
&= \sum_i \frac{w(i)\mu_i(\boldsymbol{x}^t)}{\sum_{i'} w(i')\mu_{i'}(\boldsymbol{x}^t)} \mu_i(x_{t+1} \mid \boldsymbol{x}^t) \\
&= \frac{\sum_i w(i)\mu_i(\boldsymbol{x}^{t+1})}{\sum_{i'} w(i')\mu_{i'}(\boldsymbol{x}^t)} \\
&= \frac{\xi_w(\boldsymbol{x}^{t+1})}{\xi_w(\boldsymbol{x}^t)} \\
&= \xi_w(x_{t+1} \mid \boldsymbol{x}^t) \\
&= \mathsf{p}_{\xi_w}(x_{t+1}, \boldsymbol{x}^t).
\end{aligned}$$

3.2.2.3. *The inductive assumption.* Again, in the classical Bayesian interpretation, we can say that the prior w encodes our inductive assumption. Consequently, the mixture predictor $\mathsf{p}_{\mathrm{mix}(w)}$ is a predictor that operates under this inductive assumption. In general, an a priori measure answers Goodman's riddle (what patterns should be extrapolated?) by stipulation: it induces a prediction method that, to various numerical degrees, extrapolates certain patterns. Thus an inductive assumption dictates what patterns are extrapolatable, or *projectible*, and to what extent. In the classical Bayesian interpretation of the mixture predictors, there is additional structure in the form of different possible hypotheses that regulate the data. Each of these hypotheses are themselves a priori measures that stipulate what patterns are deemed projectible and to what extent. The prior w thus encodes what patterns are extrapolatable according to the associated hypotheses; and the mixture predictor with this prior proceeds precisely by attempting to extrapolate the patterns encoded in the prior. (Cf. Romeijn, 2004, 358.)

3.2.2.4. *Consistency.* It only sounds reasonable that if we operate under a particular inductive assumption, and this assumption is in fact *correct*, then we should be able to predict well. More precisely: if we assign positive prior probability to a particular hypothesis μ^* that is actually *true*—the data is in fact generated by μ^*—then we should converge to good or indeed the best possible predictions. This expectation is warranted, to an important extent: classical Bayesian inference (at least, with a countable hypothesis class) has the important characteristic that it is *consistent*: if the data is in fact generated by a measure μ^* in the hypothesis class $\mathcal{H}$, then, as t goes to infinity, the marginal $\mu_{\text{Bayes},t}(\cdot)$ will converge almost surely to the conditional $\mu^*(\cdot \mid \boldsymbol{x}^t)$. I also refer to the latter property of almost-sure convergence to the true conditional probabilities as *reliability*: a mixture predictor's property of consistency thus means that it is reliable when the inductive assumption in fact holds true.

Theorem 3.1 (Consistency). *For mixture predictor $\mathsf{p}_{\text{mix}(w)}$ over $\mathcal{H} \ni \mu^*$, with μ^*-probability 1,*

$$\mathsf{p}_{\text{mix}(w)}(\boldsymbol{x}^t) \xrightarrow{t\to\infty} \mu^*(\cdot \mid \boldsymbol{x}^t).$$

Proof. This follows directly from the fact that the mixture ξ_w dominates $\mu^* \in \mathcal{H}$, and the Blackwell-Dubins theorem (1962). See Dawid (1984, 284) and the proof of convergence theorem 2.9. Also see B.2.2 for more details. □

3.2.2.5. *A mistaken inductive assumption.* The other side of the coin is that a predictor operating under a particular induction assumption can do very bad (or much worse than other methods would do) if its inductive assumption does *not* match the actual situation. An instance is the use of an exchangeable predictor (the adoption of exchangeable degrees of belief), that, as discussed next, corresponds to the inductive assumption of an i.i.d. source, in case there are in fact significant order effects in the generation of the data (cf. Gillies, 2001b, 369ff). However, this does not *have* to be so: even if the inductive assumption represented by a certain class of measures is off the mark, there might still be measures in the class that give conditional probabilities that are reasonably aligned with the sequence that is being generated. This is one motivation for the later interpretation of an aggegating predictor over a pool of *predictors*, see 3.3.1 below.

3.2.3. De Finetti's representation theorem.

De Finetti's celebrated theorem (1937) states that every exchangeable measure equals a mixture over the hypothesis class $\mathcal{H} = \{\mu_\theta\}_{\theta\in[0,1]}$ of all i.i.d. measures. That is, every exchangeable μ can be expressed as

$$\mu(\boldsymbol{x}) = \int_0^1 \mu_\theta(\boldsymbol{x}) dW(\theta),$$

for some prior probability measure W over the interval $[0,1]$.

In the special case of the exchangeable measures corresponding to the Johnson-Carnap predictors $\mathsf{p}_{\lambda,\gamma}$ (3.1.2.5 above), the prior measure w takes the special form of a Beta(β_0, β_1) distribution with parameters $\beta_0 = \gamma_0\lambda, \beta_1 = \gamma_1\lambda$, with density (relative to the uniform measure) given by

$$w_{\mathrm{Beta}(\gamma_0\lambda,\gamma_1\lambda)}(\theta) = \frac{\Gamma(\lambda)}{\Gamma(\gamma_0\lambda)\Gamma(\gamma_1\lambda)}(1-\theta)^{\gamma_0\lambda-1}\theta^{\gamma_1\lambda-1}. \tag{38}$$

3.2.3.1. *Reading the representation theorem ($\Leftarrow$).* For de Finetti, the significance of his result was that "the nebulous and unsatisfactory definition of 'independent events with fixed but unknown probability'" (1937, 142), i.e., the notion of an i.i.d. probabilistic source, could be abandoned for a "simple condition of 'symmetry' in relation to our judgments of probability" (ibid.), i.e., a subjective judgment of invariance expressed in our degrees of belief. (See Galavotti, 2001, 163ff.) In the interpretation of Hintikka (1971; following Braithwaite, 1957), talk of general hypotheses, problematic from a strictly empiricist point of view, could be abandoned for constraints on methods of prediction (also see Romeijn, 2004, 336f).

3.2.3.2. *Reading the representation theorem ($\Rightarrow$).* However, one could also reason the other way around (cf. ibid.). On this reading, the representation theorem shows that an exchangeable a priori measure actually comes down to a particular inductive assumption: in this case the assumption of an i.i.d. data-generating source. An exchangeable predictor (in particular, a Johnson-Carnap predictor) actually operates under the inductive assumption of an i.i.d. data-generating source (with a particular prior allocation over the possible parameters). Likewise, looser symmetry constraints like *Markov* and *partial* exchangeability (see, e.g., Diaconis and Freedman, 1980) correspond via representation theorems to looser assumptions on the data. We can say that in general, a representation theorem that relates a particular class of mixture measures and a particular class of predictors shows that this particular class of predictors operates under a particular inductive assumption.

3.2.3.3. *Consistency.* Though they fall outside the scope of consistency theorem 3.1 (that is restricted to countable $\mathcal{H}$), it is not hard to show that the Johnson-Carnap predictors are, in fact, consistent. Namely, it is obvious from the definition that a predictor $\mathsf{p}_{\lambda,\gamma}$ satisfies the *axiom of convergence* (or *Reichenbach's axiom*, Carnap, 1963a, 976): it will always converge to the data's limiting relative frequency, and with true μ_{θ^*}-probability 1 the generated sequence will actually have limiting relative frequency θ^*. (See Skyrms, 1996, 324.)

3.2.4. The representation theorem 2.16. This, recall from 2.2.2.4, is the statement that the Solomonoff-Levin measures Q_U are precisely the universal Σ_1 mixtures ξ_w. I call it a representation theorem because, like de Finetti's theorem, it corresponds a particular type of a priori measure—the Solomonoff-Levin measures Q_U that give an a priori probability assignment to all sequences based, in the standard view, on their compressibility—to a particular type of mixture—the Σ_1 mixtures $\xi_w^{\mathcal{M}}$ over all Σ_1 hypotheses.

3.2.4.1. *The Σ_1 mixture predictors.* These are the methods $\mathsf{p}^{\mathcal{M}}_{\mathrm{mix}(w)}$ corresponding to the Σ_1 mixture measures, in the sense of 3.2.2.2 above. The representation theorem 2.16 thus shows the equivalence of the Solomonoff-Levin predictors p_{Q_U} and the Σ_1 mixture predictors $\mathsf{p}^{\mathcal{M}}_{\mathrm{mix}(w)}$.

3.2.4.2. *Reading the representation theorem.* If we interpret theorem 2.16 like we interpreted de Finetti's theorem in 3.2.3.2 above, then it says that the Solomonoff-Levin measures correspond to a particular inductive assumption: the assumption that the data is generated from a Σ_1 source. The Solomonoff-Levin predictors thus operate under a particular inductive assumption: the *inductive assumption of* Σ_1 *effectiveness.*

3.2.4.3. *The notion of* Σ_1 *source.* The notion of a data-generating measure on $\mathbb{B}^\omega \cup \mathbb{B}^*$ or semi-measure on $\mathbb{B}^\omega$ leaves, on a charitable reading, some options for interpretation (what does it mean for there to be a positive probability of *no* symbol?); on a less charitable reading, it is an odd notion. In particular, it is not fully clear what it should mean to *converge* on such a measure (see B.2.2.4). Barring the occasional evocation of the picture of (probabilistic) machines actually generating the data under investigation (e.g., Solomonoff, 1964, 14; Li and Vitányi, 2008, 350; in which case we do have to do with Σ_1 measures, 2.1.2 above), it is more in line, anyway, with how Solomonoff's theory is usually presented to simply take the relevant part of the inductive assumption of a Σ_1 source to be the inductive assumption of a data-generating source that is a computable or Δ_1 measure. When, therefore, in the following I talk about the inductive assumption of Σ_1 effectiveness, this may be read as 'the inductive assumption of Δ_1 effectiveness.'

3.2.4.4. **Disanalogies.* In my presentation here I seek an analogy between de Finetti's theorem and the representation theorem 2.16, but I should also stress that this analogy only goes so far.[9] There are several formal aspects of de Finetti's theorem and related representation theorems (that also account for their usefulness in many domains; see, e.g., Paris and Vencovská, 2015) that are not shared by theorem 2.16. An important reformulation of de Finetti's theorem is that the exchangeable measures form a convex class with the i.i.d. measures as extremal points (see Hewitt and Savage, 1955); there does not seem to be an interesting analogous statement of theorem 2.16. We do not have here a statement that a given Solomonoff-Levin measure, a transformation via some monotone machine, is uniquely expressible as a mixture of elements of some strict subclass of transformations. There is really one class at play here, the class $\mathcal{M}$ of Σ_1 measures, and by the central property of universality the Solomonoff-Levin measures are contained in it. The statement here is rather that there is an equivalence between transformations and Σ_1 measures (this was proposition 2.5), and that this equivalence holds in a specific form for a subclass of universal elements: the universal transformations are equivalent to the universal mixtures over all Σ_1 measures. The definition of a universal transformation is in a sense already quite close to a mixture representation, which shows in the proof of theorem 2.16 (see B.1.4): one direction is a trivial rewriting, and the harder direction is not too hard, either. (Wood et al. themselves present their theorem as only a minor improvement over the equivalence up to a multiplicative constant that directly follows from both definitions' universality.) Nevertheless, it remains an important conceptual observation about the Solomonoff-Levin proposal that the definition as a universal transformation (with the association of simplicity qua data-compression interpretation) is in fact equivalent to the definition as a universal mixture (with the interpretation of a particular inductive assumption). Thus, even if the analogy to *de Finetti's* representation theorem only persists at this conceptual level, it is still appropriate to refer to theorem 2.16, as a result establishing the equivalence of two importantly different definitions, as a *representation theorem* (cf. Suppes, 2002).

3.2.4.5. *Consistency.* The convergence theorem 2.9 states that the Solomonoff-Levin predictors converge almost surely to any true Δ_1 measure μ^*. The observation that the Solomonoff-Levin measures are precisely the universal Σ_1 mixtures reveals that the convergence theorem 2.9 is really an instance of the general Bayesian consistency theorem 3.1, applied to the Σ_1 mixture predictors (restricted to Δ_1 sources).

3.2.5. The element of subjectivity. In 3.1.5 above, we encountered the view of the Solomonoff-Levin measure as an a priori measure that is in some sense an *objective* starting point. Objectivity, one might feel, should come with a certain *uniqueness*: *the* objective starting point. Certainly in the common view of the Solomonoff-Levin measure as an objective a priori measure, it has therefore often been seen as problematic that the Solomonoff-Levin measure is *not* uniquely defined (e.g., Solomonoff, 1986, 477; Hutter, 2007, 44f). The fact is that the definition of Q_U retains an element of arbitrariness or subjectivity in the choice of universal machine U.

3.2.5.1. *Carnap and objectivity.* To Carnap, objectivity lies in the purely logical character of probability (1950):

> That [logical probability] is an objective concept means this: if a certain [logical probability] value holds for a certain hypothesis with respect to a certain evidence, then this value is entirely independent of what any person may happen to think about these sentences, just as the relation of logical consequence is independent in this respect.

This, to him, does not necessarily entail a *unique* rational measure function. In a letter to de Finetti (1963b),[10] Carnap summarizes his perspective:

> It is true that in the time before 1950 I thought sometimes that in the course of time an intuitive insight might yield (or at least approach as an ideal) requirements of rationality of such a strength that they would lead to a unique function. But I don't think that I published this thought as an assertion. In the Appendix to my probability book of 1950 (pp. 562f.) I said that the reasons for my

> choice of the function c* at that time were mainly of a negative nature and that "it is not claimed that c* is perfectly adequate, let alone that it is the only adequate function". Soon thereafter, in "The Continuum of Inductive Methods" (1952) I replaced the choice of one function c* by that of a system of an infinite number of c-functions (which I called the lambda-system). In section 25 of "Replies and Systematic Expositions" [(1963a, 971)] ... I said ...: "Let us now consider rational credibility functions. We shall not assume that there can be only one such function, but rather leave open the possibility that two reasonable persons, or one reasonable person during different periods of his life may have different credibility functions."

He gives one more example, taken from (Carnap, 1962b). An updated version of this work is (Carnap, 1971a), containing his—careful, yet still hopeful—last words on the matter (ibid., 27):

> Even on the basis of all axioms that I would accept at the present time for a simple quantitative language ..., the number of admissable [measure functions] ... is still infinite; but their class is immensely smaller than that of all coherent [measure functions]. There will presumably be further axioms, justified in the same way by considerations of rationality. We do now know today whether in this future development the number of admissable [measure functions] will always remain infinite or will become finite and possibly even be reduced to one. Therefore, at the present time I do not assert that there is only one rational [initial credence function].

3.2.5.2. *Latitudinarianism.* Thus Carnap allows for the possibility of multiple rational or *objective* measure functions, and he seems to take this to be compatible with the *subjective* choice of a particular one among those. This is, in any case, the interpretation of Jeffrey (1973), who talks about Carnap's "latitudinarianism." Jeffrey states that one part of what it means for a confirmation function to be logical is that "[i]ts values will be in agreement with our inductive intuitions *as they will exist at the time when the program* [of inductive logic] *has been carried out*" (ibid., 301), and then writes that

> Carnap was prepared to admit the possibility that different people might have somewhat different inductive intuitions, e.g. when, ca. 1951, he thought the right c-function might be found somewhere in the continuum of inductive methods, he thought that different people might discover that they had somewhat different values of λ and hence that their inductive intuitions were describable by somewhat different functions c_λ. He thought it possible that these differences were irreducible, so that his program ... might fail in the mild sense that there might be no such thing as *the* c-function which represents *our* inductive intuitions.

We have, in the choice of λ, still the choice how we weigh the logical and the empirical factor (1.2.3.8 above).

3.2.5.3. *Asymptotic equivalence.* Nevertheless, all functions p_λ with finite λ (indeed all Johnson-Carnap functions $\mathsf{p}_{\lambda,\gamma}$) adapt to the observed empirical frequency. Thus, Jeffrey writes, this "failure would be mild – so mild as not to deserve the name, failure – if the various inductive intuitions are sufficiently similar so that their differences are swamped out by experience" (ibid.). Since the λ-terms in the definition of the function $\mathfrak{c}_\lambda$ (indeed the λ and γ terms in the function $\mathfrak{c}_{\lambda,\gamma}$) will vanish as t grows, all such functions will converge to the relative frequency in the data, and any two such functions will converge to the same predictions. There is a strong invariance between different choices of λ (and γ): all Johnson-Carnap predictors are *asymptotically equivalent.*

3.2.5.4. *Invariance.* In the case of the Solomonoff-Levin measures Q_U, there also exist a specific invariance between different choices of universal machine U. This stems from the fact that every universal machine U_1 can emulate every other universal machine U_2: it follows from definition 2.7 of universal monotone machines that there are $\boldsymbol{z}_1, \boldsymbol{z}_2$ such that for all $\boldsymbol{x}$

$$U_1(\boldsymbol{z}_2\boldsymbol{x}) = U_2(\boldsymbol{x}) \text{ and } U_2(\boldsymbol{z}_1\boldsymbol{x}) = U_1(\boldsymbol{x}).$$

This implies that the shortest descriptions via one universal machine do not differ more than a constant length from those via the other, a fact known as *the invariance theorem* (Li and Vitányi, 2008, 104ff, 200ff). This also means that the probability assignments of two Solomonoff-Levin measures via different machines U_1 and U_2 never differ more than a fixed factor, which is generally taken to grant the definition of the Solomonoff-Levin measure a crucial robustness. I discuss the significance of the invariance theorem in more detail in 5.2.2 below: but I will already note here the obvious main weakness of this notion of invariance, which is that the constant factor that binds two different measures can still be arbitrarily large.

3.2.5.5. *The failure of asymptotic equivalence.* The invariance in the case of the Solomonoff-Levin definition is indeed strictly weaker than in the case of the Johnson-Carnap definition. It is *not* the case that any two Solomonoff-Levin predictors are asymptotically equivalent: their predictions will *not* converge on every infinite sequence.

Theorem 3.2 (No asymptotic equivalence). *For every universal machine U_1, there is a sequence $\boldsymbol{x}^\omega$ and another universal machine U_2 such that*

$$Q_{U_1}(x_{t+1} \mid \boldsymbol{x}^t) \overset{t\to\infty}{\not\longrightarrow} Q_{U_2}(x_{t+1} \mid \boldsymbol{x}^t).$$

Proof. It is not trivial to exhibit such a sequence,[11] because invariance still means that any two Solomonoff-Levin measures are very similar—as exhibited most clearly by the constant bound on the difference in their cumulative log-losses, proposition 3.4 below. The key is the construction by Hutter and Muchnik (2007, also see Lattimore and Hutter, 2015) of a Martin-Löf random sequence on which some Solomonoff-Levin measure does not converge.[12] See B.2.3. □

3.2.5.6. **Erratum.* In Sterkenburg (2016, 473), I wrote rather carelessly that invariance implies that two different "Q_U and $Q_{U'}$ are *asymptotically equivalent.*" (In addition, in the type-setting process the word "machines" was erroneously added to refer to these measures.) Now there is some justification for calling the *complexity measures* $-\log Q_U$ corresponding to the Solomonoff-Levin measures (A.3.2.4) asymptotically equivalent, at least *on average*, because the constant difference will always wash out in the sense that

$$\frac{-\log Q_{U_1}(\boldsymbol{x}^t)}{t} \xrightarrow{t\to\infty} \frac{-\log Q_{U_2}(\boldsymbol{x}^t)}{t}.$$

Similarly, the original measures Q_{U_1} and Q_{U_2} are asymptotically equivalent *on average* in the sense that

$$\sqrt[t]{Q_{U_1}(\boldsymbol{x}^t)} \xrightarrow{t\to\infty} \sqrt[t]{Q_{U_1}(\boldsymbol{x}^t)}.$$

That said, it is easy to refute asymptotical equivalence simpliciter by exhibiting sequences $\boldsymbol{x}^\omega$ with

$$Q_{U_1}(\boldsymbol{x}^t) \not\xrightarrow{t\to\infty} Q_{U_2}(\boldsymbol{x}^t).$$

For instance, consider two Solomonoff-Levin measures corresponding to universal Σ_1 measures that each give a weight of 0.9 to a different deterministic Δ_1 hypothesis or computable infinite sequence. Clearly, their respective probability assignments to the initial segments of either of these sequences cannot converge. Thus it is not the case that any two Solomonoff-Levin measures are asymptotically equivalent. Nor is it the case, by theorem 3.2, that any two Solomonoff-Levin *predictors* are asymptotically equivalent.

3.2.5.7. *Solomonoff and objectivity.* Acceptance of an element of subjectivity in the definition of the Solomonoff-Levin measure has been slow in the field (Solomonoff, 2009, 9f):

> For quite some time I felt that the dependence of [algorithmic probability] on the reference machine was a serious flaw in the concept, and I tried to find some "objective" universal device, free from the arbitrariness of choosing a particular universal machine.

This goal has indeed been elusive: there does not seem to be a principled way to single out a 'most natural' or objective universal machine with which to define the Solomonoff-Levin measure. Müller (2010) presents an interesting attempt to isolate a machine-invariant version of algorithmic probability: his idea is to derive the stationary distribution of the Markov process of universal machines emulating each other. But this distribution does not exist, and he concludes that "*there is no way to get completely rid of machine-dependence*, neither in the approach of this paper nor in any similar but different approach" (ibid., 126).

3.2.5.8. *Carnap's shift to the subjective.* Jeffrey (1973, 301) remarks, "Carnap's latitudinarianism is suggestive: perhaps he is describable as a special sort of subjectivist." The different Johnson-Carnap predictors are importantly similar, but among them, in the choice of λ (and γ), they still leave some room for "individual psychology" or subjectivity. Undeniably, "the selection of a particular value of λ to uniquely determine a measure seems in the grand tradition of subjective theories of probability" (Suppes, 2002, 198). As discussed by Zabell (2011, 302ff), the evolution in Carnap's program has been widely perceived as a "shift to the subjective," which, Zabell points out, is corroborated by Carnap himself when he says that the difference between "the objectivist point of view and the subjectivist or personalist point of view" is merely one of "attitude or emphasis between the subjectivist tendency to emphasize the existing freedom of choice, and the objectivist tendency to stress the existence of limitations" (1980, 119). Zabell, however, does not agree with Carnap's way of putting things: "the issue is not one of favoring 'limitation' versus 'choice'; it is one of *whether or not you think the postulate accurately captures the epistemic situation at hand*" (Zabell, 2011, 303). In accordance with 3.2.3.2 above, the Johnson-Carnap predictors should not be seen as resulting from particular rationality constraints: they are the predictors that operate under a particular inductive assumption, in this case, of an i.i.d. data source. Still, there remains a clear element of choice when we have formulated an inductive assumption like the assumption that the data is i.i.d., namely the specific prior distribution we put over the elements of the class of i.i.d. hypotheses.

3.2.5.9. *Solomonoff's shift to the subjective.* In the end, Solomonoff, too, turned away from the idea of a single most objective choice, and came to embrace the selection of a universal machine as an inevitable and essentially subjective element of prior information in the definition of his prediction method (2009, 9ff; also see 2003, 600). By the representation theorem 2.16, we know that this choice of machine corresponds to a choice of effective weight function over the Σ_1 measures. Like the Johnson-Carnap predictors and the inductive assumption of i.i.d. data, the Solomonoff-Levin predictors can be seen to operate under the inductive assumption of Σ_1 effectiveness; and this inductive assumption still leaves an element of choice in the specific effective prior over the hypothesis class, that is, the choice of universal machine. That there is indeed a correspondence between the choice of universal machine and the choice of a prior over effective hypotheses has been noted before, for instance by Wallace (2005, 401ff). The representation theorem 2.16 tells us that the analogy between the choice of universal machine and effective prior over the Σ_1 measures is in fact an *exact* correspondence.

* * *

3.3. Mixtures over prediction methods

This section discusses prediction methods that are mixtures over other prediction methods. The mixture predictors are reinterpreted as aggregating strategies over a pool of competing predictors.

In 3.3.1, I present the mixture predictors as aggregating predictors over a pool of prediction methods. In 3.3.2, I discuss the optimality of aggregating predictors. In 3.3.3, I present the Solomonoff-Levin predictors as aggregating predictors.

3.3.1. Aggregrating predictors. Let w be a weight function over a class $\mathcal{H} = \{\mu_i\}_{i \in I}$ of measures. Recall the definition (37) of the mixture predictor

$$\mathsf{p}_{\mathrm{mix}(w)}(\boldsymbol{x}^t) := \sum_i w_{t+1}(i)\mu_i(\cdot \mid \boldsymbol{x}^t), \tag{39}$$

where the updated weight function at trial t is given by

$$w_{t+1}(i) = \frac{w_t(i)\mu_i(x_t \mid \boldsymbol{x}^{t-1})}{Z},$$

with $Z = \sum_i w_t(i)\mu_i(x_t \mid \boldsymbol{x}^{t-1})$ and $w_0 = w$. In the above classical Bayesian interpretation, the weight function w is a *prior distribution* over the *hypothesis class* $\mathcal{H}$.

However, by the correspondence between measures and prediction methods of 3.1 above, we can also interpret $\mathcal{H}$ as (corresponding to) a *pool of prediction methods* $\{\mathsf{p}_i\}_{i \in I}$, with $\mathsf{p}_i := \mathsf{p}_{\mu_i}$ the predictor corresponding to μ_i. Then the predictor (39), given by

$$\mathsf{p}_{\mathrm{mix}(w)}(\boldsymbol{x}^t) = \sum_i w_{t+1}(i)\mathsf{p}_i(\boldsymbol{x}^t), \tag{40}$$

with

$$w_{t+1}(i) = \frac{w_t(i)\mathsf{p}_i(x_t, \boldsymbol{x}^{t-1})}{Z}, \tag{41}$$

where $Z = \sum_i w_t(i)\mathsf{p}_i(x_t, \boldsymbol{x}^{t-1})$ and $w_0 = w$, is a prediction method that mixes the predictions given by the predictors in pool $\mathcal{H}$. (Here and below I use 'predictors in pool $\mathcal{H}$' as shorthand for 'predictors corresponding to the measures in $\mathcal{H}$.') Predictor $\mathsf{p}_{\mathrm{mix}(w)}$ takes a weighted mean of all the predictions issued by the elements in pool $\mathcal{H}$, where the weights are updated by (41) to correct for how well each predictor has done in the past. To use a different term, $\mathsf{p}_{\mathrm{mix}(w)}$ *aggregates* the predictions of all p_i into a single prediction. In order to distinguish this interpretation from the classical Bayesian interpretation of a predictor that mixes over measures, I will refer to a predictor given by (40) as an *aggregating predictor*.

The important difference of this interpretation from the classical Bayesian interpretation is that it frees us from the commitment to a belief that one of the measures actually generates the data. This operationalist interpretation is arguably more natural in our setting of sequential prediction, where we are not so much interested in finding the true generating source as in simply predicting well. It is the operationalist interpretation that very much goes together with the information-theoretic tradition in prediction (see, e.g., Cesa-Bianchi et al., 1997, 431, Barron, 1998, 29f; Grünwald, 2007, 26ff), that can indeed be seen to go back to Solomonoff (3.3.3 below). Here we are interested in predicting well *in every case*, not just when an inductive assumption we make on the world happens to be correct—although, lest we set ourselves an utterly impossible goal, with an important qualification. Namely, we start out with some natural pool of competing prediction strategies, which may or may not be interpreted as giving predictions in accordance with beliefs about the true data-generating mechanism, and our goal is to predict well, in every case, *relative to this pool of prediction methods.* This problem setting of predicting well relative to a given pool of experts has been studied in machine learning under various headers, including *universal prediction* (Merhav and Feder, 1998, 2126ff), *individual sequence prediction* (Grünwald, 2007, 575ff), and *prediction with expert advice* (Cesa-Bianchi and Lugosi, 2006).

As an illustration, consider the pool of predictors that correspond to all i.i.d. measures μ_θ for $\theta \in [0, 1]$. In the previous interpretation a mixture over this pool operates under the inductive assumption of an i.i.d. source, and consistency guarantees we will do well if this assumption is in fact true. However, we might be in a situation where the inductive assumption is flat wrong, while some of the predictors in the pool still perform quite well: for instance, the predictor that issues the stationary distribution of the Markov process that actually governs the data. In that case it is of value that we can indeed design strategies that will predict not much worse than any of the predictors in the pool, without relying on any inductive assumptions.

I discuss the framework of prediction with expert advice more fully in chapter 6; in the remainder of the current chapter I will focus on the aggregating predictor and its optimality. To start with: why is the aggregating predictor suited for the above problem—in particular, why do we retain the update rule (41) in accordance with Bayes's rule, rather than updating the weights in some other way? (It is not clear that the motivation for Bayes's rule in 3.1.1.5 above, explicitly stated as a condition on beliefs, still makes sense in this new interpretation.) A justification is given by the fact that this method is indeed guaranteed to be never much worse, in a specific sense, than the predictors it aggregates over: it is provably *optimal*, in a sense I explain next.

3.3.2. Optimality. Here I point out in what sense an aggregating predictor over a pool of predictors is provably *optimal* relative to (the predictors in) this pool.

In 3.2.2.4 above I discussed that the mixture predictors in the classical Bayesian interpretation are consistent, or *reliable* in the sense that they converge with probability 1 to the true predictive probabilities—*if* these are given by a measure μ^* that is actually in the hypothesis class $\mathcal{H}$. In contrast, *optimality* of an aggregating predictor is to mean that it does well relative to any prediction method in $\mathcal{H}$, *on every possible data stream.* In particular, we make no assumptions at all on the data-generating mechanism, which also means that we have no way of expressing a predictor's performance in terms of convergence to actual probabilities. For that reason we need to express a predictor's performance directly in terms of how far its predictions diverged from the data obtained, and the way we do this is by means of a *loss function.*

3.3.2.1. **Convergence of predictions.* What do we get, though, if we simply reinterpret the original consistency theorem 3.1? If we reinterpret the true measure μ_i as a prediction method, then the theorem would take the meaning that every p_i in the predictor pool initially *anticipates with certainty*, per the corresponding a priori measure μ_i, that the mixture's predictions converge to its own. I briefly discuss this in B.2.2.3, but I will not have use for this interpretation in the main text: a problem is again that for *semi-measures* the notion of almost-surety is ambiguous.

3.3.2.2. *Losses.* A loss function $\ell : \mathcal{P} \times \mathbb{B} \to \mathbb{R}^{\geq 0}$ measures how bad a prediction $p \in \mathcal{P}$ was in light of the outcome $x \in \mathbb{B}$. Thus the *instantaneous loss* of prediction p on outcome x is given by $\ell(p, x)$; I will also write '$\ell_p(x)$.' The *cumulative loss* of a prediction method p on a sequence $\boldsymbol{x}$ is the sum of instantaneous losses of p's predictions in the course of the generation of $\boldsymbol{x}$:

$$L_{\mathsf{p}}(\boldsymbol{x}^s) := \sum_{t=0}^{s-1} \ell_{\mathsf{p}(\boldsymbol{x}^t)}(x_{t+1}). \tag{42}$$

3.3.2.3. *Loss bounds.* Our goal in designing prediction methods is to incur low and ideally quickly decreasing instantaneous losses: this prompts us to try and establish *bounds* on a method's cumulative loss. To gain some intuition, let me relate bounds on the cumulative loss to losses per outcome. First of all, a *linear* bound on the cumulative loss on a data stream $\boldsymbol{x}^\omega$, so $L(\boldsymbol{x}^t) = O(t)$, comes down to a same positive amount of loss on this stream every single round: in this case our instantaneous losses do not decrease at all, which means we are not really learning anything from the data. (We can always achieve this with the indifferent predictor.) A bound of order $O(\sqrt{t})$ already translates in instantaneous losses that decrease at a rate $O(1/\sqrt{t})$. (This follows from the approximate equality $\sum_{s=1}^{t} 1/\sqrt{s} \approx \sqrt{t}$, which can be seen from evaluating $\int_{s=1}^{t} s^{-1/2}$.) Better yet is a bound of order $O(\log t)$ that translates in instantaneous losses that decrease at a rate $O(1/t)$. (Which follows from the approximate equality $\sum_{s=1}^{t} 1/s \approx \log t$.) Still superior to this is a *constant* bound, where $L(\boldsymbol{x}^t) = O(1)$ or $L(\boldsymbol{x}^t) \leq c$ for some constant c, all t: in this case the instantaneous loss we incur each round on this data stream must decrease at a rate *faster* than $O(1/t)$. (The series $\sum_{s=1}^{\infty} 1/s^a$ is convergent as soon as $a > 1$.)

3.3.2.4. *Regrets.* Given the previously mentioned goal of predicting well relative to a pool of competing methods, we are rather interested in how much *more* loss we incur then these other methods: we are interested in our surplus loss or *regret* relative to these methods (and on *all* data streams). Formally, the cumulative regret $R_{\mathsf{p}_1,\mathsf{p}_2}(\boldsymbol{x}) \in \mathbb{R}$ of one prediction method p_1 relative to another method p_2 on sequence $\boldsymbol{x}$ is the surplus loss that p_1 incurred in comparison to p_2,

$$R_{\mathsf{p}_1,\mathsf{p}_2}(\boldsymbol{x}) := L_{\mathsf{p}_1}(\boldsymbol{x}) - L_{\mathsf{p}_2}(\boldsymbol{x}). \tag{43}$$

As in the case of losses, we can establish bounds on the cumulative regret a method incurs relative to another on a given data stream, where a linear bound still allows for the first method to keep incurring at least a same amount of loss more than the other, while a constant bound translates in the first method, if not actually incurring less loss than the other, at least reducing the gap at a rate of $1/t$.

3.3.2.5. *Logarithmic losses and regrets.* In 6.1 below I will discuss requirements on loss functions and several standard such functions: here I will assume one important loss function, the *logarithmic loss function.* The logarithmic loss or simply *log-loss* of prediction p on outcome x is (Good, 1952)

$$\ell(p, x) := -\log p(x). \tag{44}$$

Thus the instantaneous log-loss of a predictor p on outcome x_{t+1} after $\boldsymbol{x}^t$ is given by

$$\begin{aligned} \ell_{\mathsf{p}(\boldsymbol{x}^t)}(x_{t+1}) &= -\log \mathsf{p}(x_{t+1}, \boldsymbol{x}^t) \\ &= -\log \mu_{\mathsf{p}}(x_{t+1} \mid \boldsymbol{x}^t), \end{aligned}$$

where μ_{p} is the measure corresponding to p. The latter representation has the advantage that the chain rule for conditional probabilities transfers to a chain rule for sums of losses, allowing us to write the cumulative log-loss as

$$\begin{aligned} L_{\mathsf{p}}(\boldsymbol{x}^s) &= \sum_{t=0}^{s-1} -\log \mu_{\mathsf{p}}(x_{t+1} \mid \boldsymbol{x}^t) \\ &= -\log \prod_{t=0}^{s-1} \mu_{\mathsf{p}}(x_{t+1} \mid \boldsymbol{x}^t) \\ &= -\log \mu_{\mathsf{p}}(\boldsymbol{x}^s). \end{aligned} \tag{45}$$

This very useful effect is also known as *telescoping.* The log-*regret* can now also be simply written as

$$\begin{aligned} R_{\mathsf{p}_1,\mathsf{p}_2}(\boldsymbol{x}) &= -\log \mu_{\mathsf{p}_1}(\boldsymbol{x}) - \big(-\log \mu_{\mathsf{p}_2}(\boldsymbol{x})\big) \\ &= -\log \frac{\mu_{\mathsf{p}_1}(\boldsymbol{x})}{\mu_{\mathsf{p}_2}(\boldsymbol{x})}. \end{aligned} \tag{46}$$

For simplicity I also write 'L_μ' for 'L_{p_μ}' and 'L_i' for 'L_{p_i},' and likewise for the regret.

3.3.2.6. *Optimality.* It is essentially the above telescoping property that allows us to translate the fact that a mixture measure majorizes every $\mu_i \in \mathcal{H}$ into the fact that the cumulative log-loss of an aggregating predictor *minorizes* the cumulative log-loss of each p_i in the pool $\mathcal{H}$,

$$L_{\mathrm{mix}(w)} \leq^{+} L_i.$$

In other words, we can actually derive a *constant* bound on the cumulative regret $R_{\mathrm{mix}(w),i}$ of the aggregating predictor relative to *every* method p_i in $\mathcal{H}$, on *every* data stream $\boldsymbol{x}^{\omega}$. This I call the *optimality* of the aggregating predictor. The relevant constant is actually a direct expression of the weight attributed to p_i, so a higher weight results in a stronger bound:

THEOREM 3.3 (Optimality). *For aggregating predictor $\mathsf{p}_{\mathrm{mix}(w)}$ over $\mathcal{H}$, every p_i in the pool $\mathcal{H}$, and every finite sequence $\boldsymbol{x}$,*

$$R_{\mathrm{mix}(w),i}(\boldsymbol{x}) \leq -\log w(i).$$

PROOF. For the mixture measure ξ_w corresponding to $\mathsf{p}_{\mathrm{mix}(w)}$ we have, for any $\mu_i \in \mathcal{H}$,

$$\xi_w(\boldsymbol{x}) = \sum_j w(j)\mu_j(\boldsymbol{x}) \geq w(i)\mu_i(\boldsymbol{x}),$$

hence for every single $\boldsymbol{x} \in \mathbb{B}^*$

$$\begin{aligned} R_{\mathrm{mix}(w),i}(\boldsymbol{x}) &= -\log \frac{\xi_w(\boldsymbol{x})}{\mu_i(\boldsymbol{x})} \\ &\leq -\log \frac{w(i)\mu_i(\boldsymbol{x})}{\mu(\boldsymbol{x})} \\ &= -\log w(i). \end{aligned}$$

□

3.3.2.7. *Semi-measures.* Note that theorem 3.3 is equally valid for mixtures over a pool $\mathcal{H}$ that contains measures over $\mathbb{B}^{\omega} \cup \mathbb{B}^*$ or *semi-measures*: all that is needed in the proof is the dominance of the mixture.

3.3.2.8. *The update rule.* The optimality result could be seen as a justification for the update rule (41)—at least in the case of the logarithmic loss function. As it turns out, for different loss functions optimality, if attainable at all, requires a modified aggregating predictor with a different update rule that in a precise sense generalizes (41). I explain this in 6.1 below.

3.3.3. The Solomonoff-Levin predictors. Solomonoff in (1964) proposed a number of different definitions for prediction methods, of which the first few can be seen as predecessors of the modern Solomonoff-Levin definition 2.8. The last proposal is different: it is a method "that makes probability evaluations by using a weighted mean of the evaluations given by all possible probability evaluation methods," i.e., all possible prediction methods (ibid., 19). It is an aggregating predictor—though Solomonoff here still identifies "all possible probability evaluation methods" with the pool of computable or Δ_1 methods.

Solomonoff writes that "[t]here are some arguments that make it plausible" that the last definition is equivalent to one or more of his other proposals (ibid., 21), but at least some of these are too imprecisely stated to prove him right or wrong (also see Solomonoff, 1997, 85f). For the modern definition, the representation theorem 2.16 shows that the Solomonoff-Levin predictors p_{Q_U} are precisely the aggregating predictors $\mathsf{p}^{\mathcal{M}}_{\mathrm{mix}(w)}$ over the pool of all predictors corresponding to the Σ_1 measures.

Solomonoff further speculates that (1964, 21, notation mine)

> It would seem, then, that if $\boldsymbol{x}$ is a very long string, [the aggregating predictor] will make an evaluation based largely on the [prediction method] of greatest weight ... This suggests that for very long $\boldsymbol{x}$, [the aggregating predictor] gives almost all of the weight to the single "best" [prediction method] ...
>
> This suggests that for very long $\boldsymbol{x}$'s, [the aggregating predictor] gives at least about as good predictions as any other [prediction method], and is much better than most of them.

While theorem 3.2 above shows that two different Solomonoff-Levin predictors do not always converge to the *same* predictions (which at least shows that the situation is not always as simple as Solomonoff sketched here), there is a precise sense in which one can say that a Solomonoff-Levin predictor must always give at least as good predictions as any other prediction method. Namely, as an aggregating predictor over the pool of predictors corresponding to the Σ_1 measures, a Solomonoff-Levin predictor's cumulative log-loss will never exceed that of any such predictor by more than a fixed constant: for every Σ_1 measure ν_i,

$$L_{Q_U} \leq^{+} L_i.$$

More precisely,

Proposition 3.4 (Optimality). For Σ_1 mixture predictor $\mathsf{p}^{\mathcal{M}}_{\mathrm{mix}(w)}$, every p_i corresponding to a $\mu \in \Sigma_1$, and every finite sequence $\boldsymbol{x}$,

$$R_{\mathrm{mix}(w),i}(\boldsymbol{x}) \leq w(i).$$

Proof. This is an instance of theorem 3.3. Also see B.2.1.1–B.2.1.1. □

Thus the Solomonoff-Levin predictors are optimal for the pool of prediction methods corresponding to the Σ_1 measures. The question whether we can call this truly *universal* optimality I take up in the next chapter.

*

CHAPTER 4

A universal prediction method

This chapter concludes the evaluation of the Solomonoff-Levin predictor as a universal prediction method. A promising interpretation of the Solomonoff-Levin predictor is the Reichenbachian interpretation as a universally optimal prediction method. The negative conclusion of this chapter, however, is that this interpretation ultimately cannot be maintained, and the reason for this already lies in Putnam's original diagonal argument.

In 4.1, I discuss and dimiss the second part of Putnam's charge against Carnap, regarding the special status of the hypothetico-deductive method; this clears the way for a final appraisal of the Solomonoff-Levin predictor as a universal prediction method. In 4.2, I discuss and dismiss the interpretation of the Solomonoff-Levin predictor as a universally reliable prediction method. In 4.3, I discuss and dismiss the interpretation of the Solomonoff-Levin predictor as a universally optimal prediction method. I conclude this chapter and part II of the thesis in 4.4.

Innovations. Section 4.1 significantly expands on a critique due to Kelly et al. (1994), and relates Putnam's argument to the problem of theory change and the fixity of methods. A main contribution of this thesis is the assessment, in sections 4.2 and 4.3, of the relevance of the Solomonoff-Levin proposal to the problem of induction, including the suggestion of a Reichenbachian vindication. This also includes the discussion of the problems with the notion of prediction method stemming from a Σ_1 measure (in particular, proposition 4.1), which ultimately leads to the negative main conclusion that the Solomonoff-Levin predictor cannot be construed as a universal prediction method. While proposition 4.1 was proven before by Leike and Hutter (2015), the proof given here exhibits it as a direct consequence of Putnam's original diagonal argument and has the additional advantage of being much simpler. (This chapter is based on parts of Sterkenburg, 201x.[13])

4.1. Back to Putnam

Can the Solomonoff-Levin definition escape Putnam's diagonal argument? As we saw in 2.1 above, the very motivation for the expansion to the class of Σ_1 measures is to evade diagonalization—to obtain universal elements. The Solomonoff-Levin measure is a universal Σ_1 element; as such, it tracks every Δ_1

measure in the sense of convergence condition (I: Δ_1). The downside is that, as a universal Σ_1 element, the Solomonoff-Levin measure 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 prediction method can satisfy both condition (I*) and condition (II*), and the Solomonoff-Levin proposal is no exception. The Solomonoff-Levin definition is powerful enough to avoid diagonalization and fulfill convergence condition (I: Δ_1), but the price to pay is that it might be said to be *too* powerful. It is no longer effective in the sense of condition (II: Δ_1). Does this invalidate the Solomonoff-Levin predictor as a prediction method—let alone a universal one?

One reply is that we cannot hold this against the Solomonoff-Levin definition, since, after all, Putnam has shown that incomputability is really a *necessary condition* for a policy to be optimal in the sense of convergence condition (I*): "an optimal strategy, if such a strategy should exist, cannot be computable ... any optimal inductive strategy must exhibit recursive undecidability" (Hintikka, 1965, 283; also see Solomonoff, 1986, 474; 2009, 8). However, this reply seems to miss the second component of Putnam's charge. This is the claim that, while no Carnapian *confirmation function* can fulfill both adequacy conditions, *other methods* could—in particular, the hypothetico-deductive method HD.

In the current section I consider this claim. As discussed already in some detail by Kelly et al. (1994, 99ff), it actually turns out to be the weak spot in Putnam's argument. With this claim out of the way, we can, in the next section, follow up on the above reply and consider the question of the Solomonoff-Levin definition's adequacy afresh.

4.1.1. The HD and Bayes architectures. Recall that I formulated (I*) and (II*) as conditions on inductive methods in general, not just confirmation 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 might be seen as a rather moot charge that indeed no Carnapian *confirmation function* can satisfy them in tandem. No confirmation function can satisfy both—conditions (I: Δ_1) and (II: Δ_1) are mutually exclusive—but other methods can: and the method HD that Putnam describes is to be the case in point.

Crucially, however, Putnam's method HD depends on the hypotheses that are actually proposed in the course of time. The method HD fulfills convergence condition (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 method HD 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 method HD 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 $\mathcal{S}$ as *part of the method itself*; but if any such method HD-with-particular-hypothesis-stream-$\mathcal{S}$—or simply 'HD$^{\mathcal{S}}$'—is powerful enough to satisfy (I*), then the hypothesis stream and hence the method HD$^{\mathcal{S}}$ 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 1.1 readily applies to any method M: simply construct a computable sequence that goes against M's computable predictions at each point in time (also see Kelly et al., 1994, 102f).

In short, the HD$^{\mathcal{S}}$ methods are in exactly the same predicament as Carnap's confirmation 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 method HD provides an "architecture," a recipe for building particular methods (in our above terminology, methods HD$^{\mathcal{S}}$), that is "universal" in the sense that for every computable hypothesis, there is a particular computable instantiation of the architecture (a particular computable method HD$^{\mathcal{S}}$) that will come to accept (and forever stick to) the hypothesis if it is 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., 110f; 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 method HD were not open to confirmation functions. That is, it would only succeed if confirmation 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.

By this I mean the architecture that goes together with the classical Bayesian interpretation of the mixture predictor (3.2 above). This architecture BAYES is the general form of the mixture predictor p_{mix}, that is instantiated by a particular prior w over a hypothesis class $\mathcal{H}$. The corresponding BAYES-with-particular-hypothesis-class-$\mathcal{H}$ method—the method 'BAYES$^{\mathcal{H}}$'—is the prediction method $\mathsf{p}^{\mathcal{H}}_{\text{mix}(w)}$.

The BAYES architecture is a universal architecture in the sense of Kelly et al. because for every (computable) deterministic hypothesis, there is a particular (computable) instantiation of the architecture (a method BAYES$^{\mathcal{H}}$ 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 $\mathcal{S}$, so the BAYES architecture is guaranteed to converge on every true deterministic hypothesis, *whenever* it is included in the hypothesis class $\mathcal{H}$. The latter is guaranteed by Bayesian consistency, theorem 3.1 in 3.2.2.4 above. And this, of course, extends to *probabilistic* hypotheses. Putnam (1963a, 774) also sketches how to adapt the method HD to deal with "statistical hypotheses" (in this case an hypothesis is rejected as being inconsistent with the data if its likelihood is sufficiently low; a further important difference is that an hypothesis may later "rule itself back in"). Now it is impossible to *guarantee* convergence to a true hypothesis, but "the *probability* that one will stick to the true hypothesis, once it has been accepted, converges to 1" (ibid.). Likewise, Bayesian consistency says that a method BAYES$^{\mathcal{H}}$ will converge with probability 1 to a true μ^* in $\mathcal{H}$.

In conclusion of this discussion, there is a strong analogy between the situation for the HD method and for the BAYES method. No *particular* confirmation function—BAYES$^{\mathcal{H}}$ method—can satisfy both (I*) and (II*). But, similarly, no *particular* HD$^{\mathcal{S}}$ method can satisfy both (I*) and (II*). Nevertheless, the HD *architecture* is universal. But, similarly, the BAYES *architecture* is universal.

4.1.2. The fixity of methods. Still, there remains a conspicuous disanalogy between the HD and the BAYES approach. This difference is *not* the use of theory per se, even though Putnam took that to be the salient characteristic of the method HD. After all, the BAYES approach uses theory, in the form of the hypothesis class $\mathcal{H}$.

Rather, this difference seems to lie in the use of *new* theory. What is somewhat shrouded in the above analogy between the 'oracles' $\mathcal{S}$ and $\mathcal{H}$ is that the method HD is conceived to operate dynamically, with hypotheses that come to it on the fly (and that are presumably informed by the data), whereas a Bayes method must do with a class of hypotheses that is fixed from the start. The latter is the well-known Bayesian problem of new theory (see Earman, 1992, 195ff) or the "fixity of the theoretical framework" (Gillies, 2001b): the Bayesian procedure, in its standard form, can only be run after we have fixed the model, and no matter how seriously at odds with the data this model will come to be, the procedure does not allow us to take a step back and adjust it.

But for our purposes this is really just an instance of the general fact of the *fixity of a prediction method.* A prediction method is a *fixed* method, a function p that for every possible finite data sequence has fixed a prediction. And as highlighted before, any such fixed method falls prey to Putnam's diagonal argument.

This issue is actually quite independent even of the role of new theory. We could modify the method Bayes$^{\mathcal{H}}$ to evaluate its own performance at certain points, and, if called for, derive from the data new hypotheses and insert those in $\mathcal{H}$—but in the end this more complicated procedure again specifies a single fixed prediction method (also see Dawid, 1985a, 1255). Likewise, an algorithm that implements the method HD, *plus* an automated search for and discovery of new hypotheses, in the end again fully specifies a particular algorithm for extrapolating data (cf. Gillies, 2001a). These are all fixed methods (that, recall from 3.1, also correspond to particular a priori measures on all possible data sequences), and the relevant difference from Putnam's HD architecture is that the latter is an *architecture*, a method that is not fully specified. (Incidentally, the modified method Bayes$^{\mathcal{H}}$ could also be seen as instantiating a modified Bayes architecture that *is* capable of incorporating new theory—like the model proposed by Wenmackers and Romeijn, 2016.)

In conclusion, Putnam's argument, purporting to show that confirmation functions have fundamental shortcomings that other methods do not, fails. If there is a shortcoming, it is being a fixed prediction method at all. If Putnam wants to maintain that it is possible for some procedure to satisfy both of his conditions, then this cannot be a fixed procedure. It needs to leave things unspecified, as the HD architecture does, and as the (modified) Bayes architecture does, too. And, again, that what is left unspecified needs to be filled in by something incomputable. Putnam would need to say that the scientific process of coming up with hypotheses is an incomputable process.

What Putnam has shown, at the end of the day, is that we are stuck with a dilemma between two possibilities that both sound dubious: either science is fundamentally unable to discover some computable patterns, or science is itself fundamentally incomputable.

4.1.3. *Simplicity orderings.

The classical Bayesian architecture naturally accommodates a *simplicity ordering* of hypotheses that Putnam (inspired by Kemeny, 1953) envisages a refined HD method to employ (1963a, 775ff), and that in (1963b, 301f) he proposes as a line of further investigation for inductive logic (ibid., 302):

> 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.

The solution to this problem is the method $\text{BAYES}^{\mathcal{H}}$ 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.

Note, however, that a prior distribution (that, I silently assumed, must satisify countable additivity) imposes some constraints on the type of ordering. It is impossible, for instance, to have infinitely many hypotheses that are *equally* simple (it is impossible to assign the same positive amount of prior probability to infinitely many hypotheses). More generally, it is impossible to have an hypothesis such that infinitely many other hypotheses are *at least as simple* (it is impossible, for any given hypothesis with positive prior probability, to assign at least as much prior probability to infinitely many other hypotheses).

Interestingly, it is exactly this kind of ordering that Carnap (1963a, 983ff) describes as a counterexample against the plausibility of Putnam's convergence condition ($\mathrm{I}^\dagger$) for the HD procedure—and consequently also against convergence condition (I). Thus Carnap dismisses Putnam's argument on the grounds that (ibid., 986)

> his result shows only that the two requirements, in spite of their prima facie appearance of plausibility, are logically incompatible and that therefore at least one of them must be abandoned. I find [effectiveness condition (II)] fairly plausible, but not [convergence condition (I)].

Instead of loosening the effectiveness condition (II), the route taken in the Solomonoff-Levin proposal, for Carnap it is convergence condition (I) that has to go!

So why does Carnap find convergence condition ($\mathrm{I}^\dagger$) implausible? Carnap (ibid., 984) first insists that Putnam's "rule of tenacity" (stating that an hypothesis, once accepted by method HD, is not later abandoned unless it becomes inconsistent with the data) should be replaced by a rule that takes simplicity into account (an hypothesis, once accepted, is not abandoned unless it becomes inconsistent with the data *or* an equally consistent but *simpler* hypothesis is proposed). Here Carnap is right that this is more in line with what Putnam writes about a simplicity ordering in a refined HD procedure. Next, Carnap describes a situation where there is a true hypothesis h that at each point in time t must give way to another hypothesis that is equally compatible with the data but simpler than h: hence h is never accepted. "Since [the refined rule of tenacity] seems to me more plausible than [the original rule of tenacity], the requirement [convergence condition (I†)] appears implausible" (ibid.).

This point of contention between Putnam and Carnap thus also depends on the required structure of a simplicity ordering, or in general on what a "satisfactory criterion of simplicity" (ibid., 985) might turn out to look like. In the next chapter I investigate what the Solomonoff-Levin theory has to say on the issue of simplicity.

* * *

4.2. Universal reliability

We have observed that conditions (I*) and (II*) are mutually exclusive: no fixed prediction 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 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 with a fresh look at the question: could the Solomonoff-Levin definition be an adequate characterization of a *universal prediction method*?

We can still, with Putnam, divide this question into two parts. First, in the spirit of convergence condition (I*), will a Solomonoff-Levin predictor be able to convergence on every reasonable hypothesis, if it is true—is it *universal* in this sense? Second, in the spirit of effectiveness condition (II*), is a Solomonoff-Levin predictor itself still a reasonably effective method—a proper *prediction method*?

To start with the first. We know that the Solomonoff-Levin predictor is able to track every computable or Δ_1 measure μ^*: this is the convergence theorem 2.9. The Solomonoff-Levin predictor is *reliable* under the assumption of a Δ_1 data-generating measure. But does this suffice to call the Solomonoff-Levin predictor a *universally reliable* prediction method?

This is the place to finally squarely confront the problem that lurked in the background to everything that I have said so far: the problem of induction. And for that, let me start, one last time, with the view of Carnap. This is the view that inductive reasoning can attain justification from some objective or rational starting point. It is in this spirit that Carnap (1962b, 317; 1971a, 30) writes that against agent X's credences that are derived from a rational initial credence function, "Hume's objection does not hold, because X can give rational reasons for it": the rationality requirements that are codified as axioms constraining the a priori measure. It also seems in this spirit that Li and Vitányi (2008), presenting the Solomonoff-Levin measure 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 difference from Carnap's view, as pointed out in 3.1.5 above, is that the Solomonoff-Levin proposal is best seen as aiming at a *universal* starting point. This is made precise in the representation of the Solomonoff-Levin measures as the Σ_1 mixture measures, with the classical Bayesian interpretation that these measures express a particular inductive assumption, the assumption of a Σ_1 source (which, of course, covers the case of a Δ_1 source). The Solomonoff-Levin measures express a universal starting point, a truly universal inductive assumption, insofar the class $\mathcal{M}$ of Σ_1 hypotheses (or already the class $\mathcal{M}_{\Delta_1}$ of Δ_1 hypotheses) is an all-inclusive or truly universal class of hypotheses.

Howson concludes his book on the problem of induction declaring: "Hume was right" (2000, 240). The way the Bayesian framework answers Hume's problem is (ibid., 239)

> about the only way it could be solved: by divorcing the justification of inductive reasoning from a justification of its consequences. *Inductive reasoning is justified to the extent that it is sound, given appropriate premises.* These consist of initial assignments of positive probability that cannot themselves be justified in any absolute sense.

A prediction method's forecasts are sound because they are consistent with the "Humean inductive assumptions" originally encoded in the corresponding a priori measure; but Hume's argument stands because the question of the justification for the premises, the inductive assumption, falls entirely outside the framework. This "logical solution to the problem of induction," similarly defended by Romeijn (2004), is in sharp contrast with Carnap's logical approach, that sought to pin down and justify the starting point itself (see Howson, 2001, 2011). It is likewise important for the compatibility of the logical solution with Hume's argument that there is no *universal* starting point. Inductive assumptions must be *restrictive*: it is impossible to have a prior over *everything* that could be true (Howson, 2000, 61ff, Romeijn, 2004, 357ff). 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 truly universal.

Could $\mathcal{M}$, then, escape Hume's argument—is $\mathcal{M}$ truly universal? Naturally, it is not. (Recall I.4 above.) As a restriction on what hypotheses could ever be *true*, a genuinely *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.

* * *

4.3. Universal optimality

Schervish (1985a, 1274) puts it more tersely:

> Nature is not (to my knowledge) hampered by the same computability restrictions as statisticians are.

This passing remark draws our attention to something important: Nature might not be constrained by computability, but it sounds plausible that *we* necessarily do "view the world through the rose-colored glasses of computable forecasting systems" (ibid.). Plausibly, we *are* constrained by computability in our methods of prediction.

Consequently, if we interpret, per 3.1 above, the elements of $\mathcal{M}$ as corresponding to *prediction methods* rather than as hypotheses (a priori measures), then $\mathcal{M}$ might be interpreted as containing *all possible* prediction methods. Wherefore the Solomonoff-Levin predictor, interpreted, per 3.3, as an aggregating predictor, is an aggregating predictor over the pool of *all possible* prediction methods.

4.3.1. Towards a universally optimal prediction method. On this interpretation, proposition 3.4 states that a Solomonoff-Levin predictor's cumulative regret with respect to *any* other given prediction method p is *always* bounded by a constant that only depends on this predictor p. We can say, on this interpretation, that a Solomonoff-Levin predictor is a universally *optimal* prediction method: *it is a prediction method that compared to* any *other prediction method will* always *come to perform at least as well.*

A Solomonoff-Levin predictor might not do well if Nature generates—incomputably—adversarial data, but as suggested in I.5 above, there is a sense in which this is not so interesting. *No* prediction method would do well in that case. More interesting is the case when at least *some* prediction method would do well. And in a precise sense, on the proposed interpretation, a Solomonoff-Levin predictor will do well in such a case: it will do well if *any* predictor does. As such, a Solomonoff-Levin predictor is *vindicated* in the sense of Reichenbach.

This interpretation is actually in line—more so than the above reliability interpretation—with Putnam's demand that the cleverest possible inductive rule should be able to eventually pick up any pattern *that our actual inductive methods would.* It is also in line with Solomonoff's stated 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).

If we accept this interpretation, then the Solomonoff-Levin definition *does* give a universal prediction method—defying the lesson taken from Putnam that there can be no such thing (Dawid, 1985b; recall again I.5). As we have seen, the crucial move to unlock this possibility after all, hence the crucial precondition to this optimality interpretation, is the expansion to the nondiagonalizable class of Σ_1 elements. The special property of this class is that it is undiagonalizable, that it contains universal elements—thus defying Dawid's observation that "in great generality" an aggregating element is "more complicated" than the elements in the pool (ibid., 340). We must now answer the question whether the expansion to this pool is reasonable at all. Analogous to convergence condition (I*) about the identification of all reasonable hypotheses with the Δ_1 measures: is it reasonable to identify all possible prediction methods with those corresponding to the Σ_1 measures?

4.3.2. Towards a universal pool of predictors. Most importantly, is the class of Σ_1 measures not *too* wide—does a Σ_1 measure that fails to be Δ_1 still induce a proper prediction method? In particular, we have returned to the second question that started 4.2 above: in the spirit of effectiveness condition (II*), does the Solomonoff-Levin predictor itself constitute a reasonable (reasonably effective) method?

With the Solomonoff-Levin definition, we do embark, in Putnam's words, on the "doubtful project of investigating measure functions which are not effectively computable" (Putnam, 1963a, 778; also see Putnam, 1985, 146). Now an incomputable measure is certainly "impractical" (Cover et al., 1989, 863), or indeed "of no use to anybody" (Putnam, 1963a, 768) in any practical sense—but that already goes for any measure that *is* computable but not in some way *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 implementable, in principle, as an explicit method—computability is the minimal requirement to be a method at all (see I.4, I.5 above). 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.")

But even if this is so, the property of mere semi-computability is still not easy to make sense of in an actual prediction game. To illustrate: are we really much better of with semi-computable functions than with *partial computable* (p.c.) functions, as suggested in I.6 above? A p.c. function (say for categorical prediction: either 0 or 1) does not seem very suited for prediction, for the following reason (cf. Kelly et al., 1994, 104). At each trial the function might not be defined, and we either have to be prepared to wait forever (in which case, if the function is indeed not defined at that trial, the prediction game is put on hold indefinitely), or we wait until at some point we decide to break the spell and just issue a default prediction (in which case we actually use a method that reduces to a *total* computable method, or, if this decision is somehow incomputable, a method that is not computable at all). In all cases, we end up with a function that is either not universal or not computable. Now a semi-computable function is at least defined on all trials, which makes it *look* less problematic: but the situation is still fundamentally the same. At each trial we can only compute lower approximations of unknown accuracy, and we either have to be prepared to wait forever to reach the actual value (and unless the probability values sum to 1, in which case we will reach surety about the value up to any accuracy, the game indeed freezes forever), or we (incomputably?) decide at some point to just go with the current approximation. In all cases the actual prediction function is either not universal or not computable.

This is already a serious problem—but there is actually another problem that precedes it, a crucial detail that decisely invalidates the universal optimality interpretation.

4.3.3. Diagonalization strikes again. This crucial detail is the fact that for the purpose of prediction, we are, of course, not so much interested in the probabilities issued by the measure functions, but by the *conditional* probabilities that give the corresponding predictors' outputs. We are not so much interested in the a priori measures as in the induced prediction methods. 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. As noted in 3.1.4 above, if a measure is Δ_1, then (and only then) the corresponding prediction method is Δ_1 as well. However, as noted in 3.1.4, too, for the Σ_1 measures this *does* make a difference. In particular, the Solomonoff-Levin predictor p_{Q_U} 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 I introduced back in 1.1. In particular, recall the statement of Putnam's original effectiveness condition, that in our setting of sequential prediction reads

(II) For every $\boldsymbol{x}^t$, it must be possible to compute an s such that $\mathsf{p}(1, \boldsymbol{x}^t 1^s) > 0.5$.

If p_{Q_U}, i.e., the one-step conditional measure $Q_U^1(\cdot \mid \cdot)$, were Σ_1, then p_{Q_U} would also satisfy effectiveness condition (II): for any given $\boldsymbol{x}^t$, by computing lower approximations of $Q_U(x_{t'+1} \mid \boldsymbol{x}^t 1^{t'})$ for increasing $t' > t$ we will effectively discover an s with $Q_U(1 \mid \boldsymbol{x}^t 1^s) > 0.5$ (note that here we also used the convergence condition in relying on the *existence* of an s where Q_U gives sufficiently high instance confirmation—Putnam's original (II) is thus not completely independent of his (I)!). This would mean that p_{Q_U} satisfies both (I) and (II), which is shown impossible by the diagonal argument. For completeness, the following proof recounts the details of this diagonalization. (See Putnam, 1963a, 768f, Putnam, 1963b, 299 for the original. A different proof has been given by Leike and Hutter, 2015, 370f.[14])

Proposition 4.1. $\mathsf{p}_{Q_U} \notin \Sigma_1$.

Proof. Suppose towards a contradiction that $\mathsf{p}_{Q_U} = Q_U^1(\cdot \mid \cdot)$ is Σ_1. We can now construct a computable infinite sequence $\boldsymbol{x}^\omega$ as follows. Start calculating $Q_U(1 \mid 1^t)$ from below in dovetailing fashion for increasing $t \in$

$\mathbb{N}$, until an t_0 such that $Q_U(1 \mid 1^{t_0}) > 0.5$ is found (since 1^ω is obviously computable, and Q_U satisfies convergence condition (I), such t_0 must exist). Next, calculate $Q_U(1 \mid 1^{t_0}01^t)$ for increasing t until an t_1 with $Q_U(0 \mid 1^{t_0}01^{t_1}) > 0.5$ is found (again, t_1 must exist because $1^{t_0}01^\omega$ is computable). Continuing like this, we obtain a list $t_0, t_1, t_2, \ldots$ of positions; let $\boldsymbol{x}^\omega := 1^{t_0}01^{t_1}01^{t_2}1\ldots$. Sequence $\boldsymbol{x}^\omega$ is computable, but by construction the instance confirmation of $\boldsymbol{x}^\omega$ will never remain above 0.5, contradicting convergence condition (I). □

Now one could try to argue that p_{Q_U} 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 (see Leike and Hutter, 2015, 365). But the problem runs deeper. The problem is that we cannot recover the optimality interpretation for conditional measures, or prediction methods.

Namely, if we would accept that a Δ_2 prediction method (i.e., a Δ_2 conditional measure) still counts as a possible prediction method, then we should identify the possible prediction methods with the class of Δ_2 prediction methods (rather than the original class of prediction methods with underlying Σ_1 measures). That means that the sought-for optimality would have to be relative to *this* class. But the Solomonoff-Levin predictor is not optimal among the Δ_2 prediction methods—*no* Δ_2 prediction method is. This is because the class of Δ_2 *measures*, that precisely induces the class of Δ_2 *prediction methods*, *is* diagonalizable: just like in the Δ_1 case, one can, for any given Δ_2 measure, construct a Δ_2 sequence that it will never converge on. The easiest way to infer this is to realize that the Δ_2 measures are precisely the Δ_1 measures that have access to the halting problem $\emptyset'$ as an oracle: in the diagonal proof we can simply replace all occurances of 'Δ_1' with 'Δ_1 in $\emptyset'$.' In computability-theoretic jargon, the diagonal argument can be *relativized* to $\emptyset'$, thus applying to the Δ_2 measures.

Nor can we take a step back and settle for the class of Σ_1 prediction methods. Once again it follows from Putnam's argument above that there cannot exist universal elements in the class of measures that induce the Σ_1 prediction methods: in the exact same way as above, one can for any given Σ_1 conditional measure construct a Σ_1 conditional measure (namely, a computable sequence) it will never converge on.

All of this easily relativizes to any jump $\emptyset^{(n)}$ of the Halting problem, showing that the diagonal argument works for the class of Δ_{n+1} prediction methods and the class of Σ_{n+1} prediction methods, for any $n \in \mathbb{N}$. The strategy for optimality cannot work on any level in the arithmetical hierarchy.

* * *

4.4. Conclusion

Thus we conclude our story on an unhappy note. We have discussed how Putnam's diagonal argument shows that no fixed method whatsoever can satisfy at the same time two conditions to qualify as a universal prediction method: 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 prediction methods definitions that only satisfy a weaker pair of conditions. Specifically, we considered the Solomonoff-Levin definition. The overarching strategy we identified to bring versions of the two conditions together is to locate a natural class of effective functions that cannot be diagonalized, i.e., that contains universal elements. If one could reasonably identify this class of functions with all possible prediction methods, then the universal elements would be vindicated as universally optimal prediction methods: methods that are in a precise sense at least as good as any other prediction method. 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 prediction methods. Unfortunately, we found a fatal flaw in this strategy. We are interested in prediction methods rather than the a priori measures they are induced from, and there is a mismatch between the two when it comes to their effectiveness properties. Specifically, while there exist undiagonalizable classes of all measures of a particular level of effectiveness (the Σ_1 measures being the case in point), Putnam's original argument reveals that there do not exist undiagonalizable classes of all prediction methods of a particular level of effectiveness, not at any level of the arithmetical hierarchy. This conclusively blocks the possibility of a prediction method that is optimal among all possible prediction methods, on any identification of the latter with a particular level of effectiveness.

The Solomonoff-Levin definition does not give a universal prediction method. No definition does. Putnam was right.

*

Part III

Complexity

CHAPTER 5

Datacompression and Occam's razor

This chapter investigates the association of the Solomonoff-Levin predictor with Occam's razor, the principle of preferring simplicity. The standard view of algorithmic information theory is that at its basis lies a general and objective notion of simplicity qua compressibility of data sequences. In particular, the Solomonoff-Levin predictor is often presented as a formalization of Occam's razor, and it is even suggested at times that this leads the way to a justification of this principle.

In 5.1, I discuss the argument to justify Occam's razor based on the Solomonoff-Levin predictor. In 5.1.1, I spell out the argument. In 5.1.2, I expose why the argument fails. In 5.1.3, I indicate some leeway to resurrect it. In 5.2, also in response to this hope of resurrection, I take up the question whether we actually have to do with a convincing notion of complexity.

Innovations. A justification of a simplicity preference via the Solomonoff-Levin definition is hinted at in many places, but it has, as far as I know, never been articulated— let alone criticized—in detail, as done in section 5.1. (This section is based on Sterkenburg, 2016.[15][16][17][18][19][20]) Likewise, the supposed formalization of a notion of simplicity qua compressibility has before, as far as I know, neither been spelled out nor evaluated in detail, as done in section 5.2.

5.1. A justification of Occam's razor?

As explained in I.7 above, the challenge of the justification for the epistemic version of Occam's razor is to show that we have epistemic grounds for a simplicity preference, while avoiding any metaphysical simplicity assumption on the world. This challenge is still preceded by the challenge of actually giving a precise definition of simplicity.

As already mentioned in I.7 above, too, the notion of Kolmogorov complexity is generally presented as giving a precise definition of simplicity, and the Solomonoff-Levin predictor is generally presented as implementing this measure to give a formalization of Occam's razor. Moreover, the convergence theorem 2.9 is generally taken as showing that the Solomonoff-Levin predictor has the epistemic virtue of a powerful reliability, a powerful truth-convergence.

Here emerges the argument that is suggested in many writings on the subject. The argument concludes from (1) the definition a type of predictor with a preference for simplicity and (2) a formal proof that predictors of this type are reliable that (per Occam's razor) a preference for simplicity helps us in finding the truth. Thus it is an argument to justify Occam's razor.

5.1.1. The argument. Li and Vitányi (2008, 347f; 1997, 14) write,

> It is widely believed that the better a theory compresses the data concerning some phenomenon under investigation, the better we have learned and generalized, and the better the theory predicts unknown data, following the Occam's razor paradigm about simplicity. This belief is vindicated in practice but apparently has not been rigorously proved before . . . We . . . show that compression is almost always the best strategy . . . in prediction methods in the style of R.J. Solomonoff.

The general form of the argument that we can distill from these words is as follows. First, we identify a maximal class $\mathcal{PR}$ of prediction methods that have a preference for simplicity (those prediction methods "following the Occam's razor paradigm"). Second, we prove that these predictors are *reliable* ("almost always the best strategy"). Taken together, the two steps yield the statement that *prediction methods that possess a simplicity bias are reliable.* In short, the argument is as follows:

1. The predictors in class $\mathcal{PR}$ are those with a simplicity bias.
2. The predictors in class $\mathcal{PR}$ are reliable.

∴ The predictors with a simplicity bias are reliable.

The force of the argument is that it establishes a connection between two seemingly distinct properties of a predictor: a preference for simplicity on the one hand, and a general reliability on the other. Occam's razor, in our setting of sequential prediction, is the principle that *a predictor should possess a simplicity bias*; the established connection provides an epistemic justification for this principle. A predictor should possess a simplicity bias *because* that guarantees its reliability. I proceed below to make precise the two steps of the argument, including the relevant notions of simplicity and reliability.

5.1.1.1. **A sufficient or necessary condition?* Let me first note, though, that this argument would only show that a simplicity preference is *sufficient* for reliability. This leaves open the possibility of predictors that do not have a simplicity preference but that are likewise reliable. One might object that a true justification for Occam's razor would need to show that simplicity is in fact *necessary* for reliability.[21] I will proceed with the original argument, but my refutation of the argument in 5.1.2 below can indeed be cast as exposing that the below simplicity property is *not necessary* for reliability, 5.1.2.5 below.

5.1.1.2. *Step 1: the class of predictors.* The relevant class of predictors is the class of Solomonoff-Levin predictors. To a first approximation (I take up a more detailed discussion in 5.2 below), the identification of a simplicity bias in these predictors proceeds as follows. We first observe from the definition of a

Solomonoff-Levin *measure* that the value

$$Q_U(\boldsymbol{y}) = \sum 2^{-|\boldsymbol{x}|} [\![\boldsymbol{x} \in \lfloor \{ \boldsymbol{x} : \Phi_U(\boldsymbol{x}) \succcurlyeq \boldsymbol{y} \} \rfloor]\!], \tag{47}$$

for $\boldsymbol{y}$ is higher as its *descriptions* $\boldsymbol{x}$ via universal machine U have greater uniform measure $\lambda(\boldsymbol{x}) = 2^{-|\boldsymbol{x}|}$. That is, $\boldsymbol{y}$ has a greater algorithmic probability as its descriptions are *shorter.* The accompanying interpretation is that $\boldsymbol{y}$ has greater algorithmic probability as it is more *compressible.* If we further interpret this measure of compressibility as a measure of *simplicity* of finite sequences, then the statement becomes: a sequence has greater algorithmic probability as it is simpler. This transfers to a simplicity preference in sequential prediction with the Solomonoff-Levin *predictor* as follows. The one-symbol extension of $\boldsymbol{y}$ with the greatest probability $Q_U(\boldsymbol{y}y)$ among the two possibilities $\boldsymbol{y}0$ and $\boldsymbol{y}1$ is the one that is the simpler; consequently, $Q_U(y \mid \boldsymbol{y}) = Q_U(\boldsymbol{y}y)/Q_U(\boldsymbol{y})$ is greatest for the y such that $\boldsymbol{y}y$ is the simpler. Hence, the Solomonoff-Levin predictor $Q_U(\cdot \mid \cdot)$ will predict with higher probability the y that renders the complete sequence $\boldsymbol{y}y$ more simple. This is, in the words of Ortner and Leitgeb (2011, 734), "evidently an implementation of Occam's razor that identifies simplicity with compressibility."

5.1.1.3. **Other simplicity properties?* Note, though, that this is only one specific kind of a preference for simplicity: there might be other and very different properties of prediction methods that can also be interpreted as simplicity biases. If these are *not* likewise connected to reliability, then, strictly speaking, the argument would not even establish that a simplicity bias is sufficient for reliability. Strictly speaking, it would only establish this for the above specific kind of simplicity-qua-compressibility bias; and it would only justify a specific form of Occam's razor about this particular simplicity-qua-compressibility. Having noted this worry here, I will again just proceed with the argument; but the peculiarity of this particular simplicity bias does play a role in my refutation in 5.1.2, and will in fact be the topic of 5.2 below.

5.1.1.4. *Step 2: the reliability.* By 'reliability' I mean the property of almost-sure convergence to the truth. The relevant result is the familiar convergence theorem 2.9, asserting the Solomonoff-Levin predictors' reliability under the assumption of a Σ_1 (or in particular, a Δ_1) source. For the sake of the argument I will ignore the earlier discussion in 4.2 on the universal reliability of the Solomonoff-Levin predictor, and charitably phrase this as reliability "in essentially every case."

5.1.1.5. **Reliability or optimality?* The phrasing "almost always the best strategy" in the passage at the start of this section perhaps suggests the property of optimality—convergence to predictions that are at least as good as those of any other prediction method—rather than reliability. However, when theorem 2.9 is invoked in the literature to demonstrate the Solomonoff-Levin predictors' performance it is invariably interpreted as a reliability result (also recall the quote on page 42). A link with finding the *truth* would indeed be the most obvious epistemic virtue to try and connect to simplicity in a justification of Occam's razor. Still, optimality can also count as an important epistemic virtue, or such is the line I have taken in this thesis: and, arguably, it would likewise be able to support the sought-for justification. This, in any case, does not matter for the refutation of the argument in 5.1.2 below, see 5.1.2.4.

5.1.1.6. *The argument, again.* With the details provided by 5.1.1.2 and 5.1.1.4 we can restate the argument as follows.

1. The Solomonoff-Levin predictors are those with a simplicity-qua-compressibility bias.
2. The Solomonoff-Levin predictors are reliable in essentially every case.

∴ Predictors that have a simplicity-qua-compressibility bias are reliable in essentially every case.

Again, this connection between a simplicity preference and a general reliability would seem to justify the principle that a predictor should prefer simplicity, the principle of Occam's razor.

5.1.2. The argument refuted.

5.1.2.1. *Step 1 translated.* By the representation theorem 2.16, the Solomonoff-Levin predictors, those predictors that possess the relevant simplicity-qua-compressibility bias, are exactly the Σ_1 mixture predictors, i.e., those mixture predictors that operate under the inductive assumption of Σ_1 effectiveness (3.2.4 above). Hence the following two formulations of step 1 of the argument are equivalent.

1. The Solomonoff-Levin predictors are those that have a simplicity-qua-compressibility bias.
1. The Σ_1 mixture predictors are those that operate under the inductive assumption of Σ_1 effectiveness.

5.1.2.2. *Step 2 translated.* Reliability 'in essentially every case' actually means reliability under the assumption of Σ_1 effectiveness. This same reliability for the Σ_1 mixtures is just the property of Bayesian consistency (3.2.4.5 above). Hence the following two formulations of step 2 are equivalent.

2. The Solomonoff-Levin predictors are reliable in essentially every case.
2. The Σ_1 mixture predictors are consistent.

5.1.2.3. *The argument translated.* If we make the property of consistency in step 2 explicit, the two steps of the argument look as follows.

1. The Σ_1 mixture predictors are those that operate under the inductive assumption of Σ_1 effectiveness.
2. The Σ_1 mixture predictors are reliable under the assumption of Σ_1 effectiveness.

Taken together, the two steps yield the conclusion that *predictors that operate under the inductive assumption of Σ_1 effectiveness are reliable under the assumption of Σ_1 effectiveness.*

5.1.2.4. **Optimality.* On the optimality interpretation of the Solomonoff-Levin predictors, we take the content of the representation theorem 2.16 to be that the Solomonoff-Levin predictors are exactly the Σ_1 aggregating predictors. Accordingly, the two steps of the argument can be translated as

1. The Σ_1 aggregating predictors are those that aggregate over the Σ_1 predictors.
2. The Σ_1 mixture predictors are optimal among the Σ_1 predictors.

Taken together, the two steps yield the conclusion that *predictors that aggregate over the Σ_1 predictors are optimal among the Σ_1 predictors.* My conclusion below applies as in the case of reliability.

5.1.2.5. **A necessary condition.* The reliability of step 2 holds for every prediction method corresponding to an a priori measure that is a universal Σ_1 measure. Since there are universal Σ_1 measures that are not Solomonoff-Levin measures (2.2.1.3 above), this means that the specific simplicity-qua-compressibility bias associated with the Solomonoff-Levin predictors is not necessary for this reliability. This strengthens my conclusion below that there is no special role for simplicity here.

5.1.2.6. *Conclusion.* In the original formulation, we define a class of predictors with a characterizing simplicity bias that we can subsequently prove to be reliable 'in essentially every case.' This formulation suggests that we have established a connection between two properties of a predictor that are quite distinct. We got out a general reliability, whereas we put in a specific preference for simplicity. This link between a simplicity bias and reliability would provide an epistemic justification of Occam's razor, the principle that a predictor should have a simplicity bias. The more explicit reformulation, however, shows that the original formulation is misleading. We got out what we put in, after all. We define a class of predictors that operate under the inductive assumption of Σ_1 effectiveness, that we can subsequently prove to be reliable *under the very same assumption* of Σ_1 effectiveness. Even if we want to stick to the interpretation of a simplicity bias rather than a specific inductive assumption, this clearly fails to count as a demonstration that a simplicity preference is good *without* an assumption that reality itself is simple. To the extent that a predictor's inductive assumption of Σ_1 effectiveness embodies a simplicity preference we have to make the exact same simplicity assumption on the world in order to prove the predictor's good performance. Thus the argument fails to justify Occam's razor.

5.1.3. The room for reply. In a way, the revelation of 5.1.2 that "we got out what we put in" is just what was to be expected.[22] On a *formal* level, obviously the two properties of simplicity and reliability cannot be very distinct: or it would not be possible to deductively derive the one from the other. Indeed, we might say that *any* manner of defining a particular property that provably guarantees another desired property will be "essentially circular, in effect assuming what one wishes to prove"—in the words of Zabell (1988, 6), when he discusses the property of exchangeability. "Of course, in one sense this must obviously be the case. All mathematics is essentially tautologous, and any implication is contained in its premises." Nevertheless, Zabell continues, "mathematics has its uses"; enabling us to "translate certain assumptions into others more palatable" (ibid.). I have, indeed, discussed how the inductive assumption of Σ_1 effectiveness can be translated into two properties that do seem *conceptually* different: a property of reliability and a property of simplicity preference. Now, to rescue the justificatory force of the argument, one could attempt to reassert the *conceptual* distinctness of the two notions; most importantly, one could try to make the case that we are dealing with a natural notion of simplicity, rather than just a peculiar interpretation of the assumption of Σ_1 effectiveness. This is the matter I take up in the remainer of the chapter.

* * *

5.2. A formalization of Occam's razor?

The driving force for the argument to justify Occam's razor, and the essential component in any conceivable attempt to rescue it from the above critique, is the idea that the Solomonoff-Levin measure implements a general and objective measure of complexity of data sequences. The aim of this section is to shake that idea.

In 5.2.1, I spell out how the relevant simplicity-as-compressibility interpretation comes about. In 5.2.2, I discuss the greatest challenge to this compressibility notion, the well-known issue of variance. This issue is a symptom of the great permissiveness of the notion of Σ_1 universality, which, I argue, presents a problem for the complexity interpretation.[23]

5.2.1. A quantitative notion of compressibility. Here I review the information-theoretic foundation for the data-compression interpretation of the Solomonoff-Levin measure. I discuss the notion of system of descriptions and its equivalence with probability functions, leading up to the notion of universal effective description length function and its equivalence with the Solomonoff-Levin measure. A fuller account of the important concepts of this section is provided in A.2.

5.2.1.1. *Description systems and code systems.* A description system is a set $D \subseteq \mathbb{B}^* \times \mathbb{B}^*$ of pairs of *source sequences* and their *description sequences*, so that $D(\boldsymbol{y}, \boldsymbol{x})$ means that $\boldsymbol{x}$ is a description of $\boldsymbol{y}$. A *coding system* or simply *code* is a description system that is a function itself, meaning that each source sequence has a unique description. The usual type of codes in information theory are the *prefix* codes (A.2.2); a more general type that is of relevance to us are the *sequential* description systems (A.2.4).

5.2.1.2. *Description length functions.* A description system comes with a *description length function* $L_D : \mathbb{B}^* \to \mathbb{N}$ that returns an expression of the length(s) of a given source sequence's shortest description(s). (In the simple case of a code, this is just the length of the given source sequence's unique description.)

5.2.1.3. *Descriptions and probabilities.* It is a central information-theoretic fact that description systems and probability assignments can be treated as interchangeable. Namely, for every description system D the function 2^{-L_D} gives a probability assignment; conversely, for every probability assignment there is some description system that thus corresponds to it. To put it more precisely, in the relevant case of sequential description systems: for every such description system D the function

$$n_D(\cdot) := 2^{-L_D(\cdot)} \tag{48}$$

is a pre-measure to a measure ν_D on $\mathbb{B}^* \cup \mathbb{B}^\omega$; conversely, for every measure ν on $\mathbb{B}^* \cup \mathbb{B}^\omega$ there is a sequential description system D_ν with

$$L_{D_\nu}(\cdot) = -\log \nu(\llbracket \cdot \rrbracket). \tag{49}$$

As such, probabilities and description lengths are formally interchangeable. A high probability $\nu(\boldsymbol{y})$ is equivalent to a small description length $L_D(\boldsymbol{y})$ and vice versa.

5.2.1.4. *Compressibility.* If $\boldsymbol{y}$ has a small description length $L_D(\boldsymbol{y})$ then one can say that D *compresses* $\boldsymbol{y}$ well, or even that $\boldsymbol{y}$ is *simple* to D. But it should be clear that this formal notion of simplicity-to-D is a relative notion, and really an expression of how well the sequence $\boldsymbol{y}$ is *fit* by D: it is equivalent to ν_D's likelihood given $\boldsymbol{y}$, or the goodness-of-fit of ν_D for $\boldsymbol{y}$.

5.2.1.5. *Universal description systems.* Let $\mathcal{D}$ be some class of description systems. A *universal* description system $D^{\mathcal{D}}$ for this class is "almost as good" as any description system in $\mathcal{D}$: for every $D \in \mathcal{D}$ there is an *overhead constant* c_D such that, for every source sequence $\boldsymbol{y}$, the universal description length of $\boldsymbol{y}$ via $D^{\mathcal{D}}$ does not exceed the description length $L_D(\boldsymbol{y})$ more than this overhead. A universal code for $\mathcal{D}$ represents the full class $\mathcal{D}$ in the sense that if some $D \in \mathcal{D}$ assigns a particular sequence a short description, then the universal code does too (up to the overhead c_D).

5.2.1.6. *Universal compressibility.* In this sense, the description lengths of a universal system for $\mathcal{D}$ can be said to reflect how well the class $\mathcal{D}$ compresses sequences, or how simple sequences are to $\mathcal{D}$. But we have to be careful here. First and foremost, there is still the choice of overhead constants, that introduces an element of arbitrariness or subjectivity. I delegate this issue to the next section 5.2.2. Second, this notion of universal compressibility is again really a measure of how well sequences are fit by the universal description system, equivalent to the goodness-of-fit of the corresponding universal distribution over the class $\mathcal{P}$ of distributions corresponding to $\mathcal{D}$. Finally, the term "universal" is slightly deceptive because a universal description system and the accompanying universal compressibility measure are, of course, relative notions still: relative to a class $\mathcal{D}$ of description systems.

5.2.1.7. *Universal effective compressibility.* The last point would be addressed by the proposal of a class $\mathcal{D}$ that is sufficiently wide to render the notion of compressibility-to-$\mathcal{D}$ a *truly* universal, fully general notion of compressibility. Interestingly, this case could be made for the *effective* description systems, as follows. To start, we could plausibly assert that it is part of the very notion of description system that it comes with a decoding *algorithm*; hence, by the Church-Turing thesis, that the decoding D^{-1} is given by a Turing machine. In this spirit, we define the *effective* (sequential) description systems as those with decoders given by (monotone) machines. Further, a *universal* machine gives the decoder for an effective description method that is universal for the class of effective description methods—i.e., we assert, the class of *all* description systems. Consequently, compressibility relative to this class, as instantiated by a universal element in this class, is a universal notion of compressibility.

5.2.1.8. *Universal description systems and Solomonoff-Levin measures.* Let me pause a little longer on how natural the constraint of effectiveness on description systems is. The effective descriptions systems are precisely those with decoders given by Turing machines. As conveyed, hopefully, by the discussion in A.2–A.3, it is for a large part this identification of effective decoders with machines that accounts for the elegance of algorithmic information theory as a branch of information theory proper. Contrast this to the notion of a Σ_1 measure on $\mathbb{B}^\omega \cup \mathbb{B}^*$, which already gives a notion that is somewhat hard to digest (certainly under the interpretation of a data-generating source, 3.2.4.3 above), and which leads to a class of prediction methods that is really not very natural (4.3.2–4.3.3 above). *Unless*, perhaps, one motivates this class precisely by falling back on the notion of effective description system. Namely, *formally* the effective sequential description systems and the Σ_1 measures (hence the predictors corresponding to those) are again equivalent, in the sense of 5.2.1.3 above. Indeed, the definition of a universal effective sequential description length function is precisely the negative logarithm of a transformation λ_U, i.e., a Solomonoff-Levin measure (A.2.4.6). So perhaps one can argue that the naturalness of this notion transfers to the Solomonoff-Levin measures. In particular, one might argue that a Solomonoff-Levin measure inherits a general and objective notion of compressibility.

5.2.1.9. **Codes and Kolmogorov complexity.* Instead of effective description methods, we could have considered effective *codes*. One might feel that codes lead to a more proper notion of compressibility because the description length function just gives the length of a single shortest description, rather than an expression of multiple descriptions. (For instance, in the latter case, it is conceivable that the function returns a low value if there is not necessarily a short description but *a large number of* (long) *descriptions.*) A description length function for a universal effective code is precisely the Kolmogorov complexity via a universal machine (see A.3); and indeed in the literature one encounters the idea that the Solomonoff-Levin measure inherits a simplicity notion insofar as it formally resembles monotone or even prefix-free Kolmogorov complexity: "[f]rom $Q_U(\boldsymbol{x}) \approx 2^{-K(\boldsymbol{x})}$ we see that Q_U assigns high probability to simple strings (Occam)" (Hutter, 2005, 47, notation mine; also see Li and Vitányi, 2008, 272f). This raises the question as to the extent of this resemblance. In the case of *prefix-free* Kolmogorov complexity K_U and the universal *prefix* description length functions there is the *coding theorem* that states their equivalence up to an additive constant (see A.3.1.7); but in the relevant case of monotone Kolmogorov complexity Km_U and the universal sequential description length functions $KM_U = -\log \lambda_U$ the situation is more messy (see A.3.2.5). Moreover, even if the function Km_U is "closer to the spirit of Occam's razor" (Hutter, 2006, 96), the resulting prediction method is in some respects actually worse than Q_U (ibid.). I will not pursue this theme further, and only note that Kolmogorov complexity as a complexity measure must also still face the challenge of variance discussed in 5.2.2 below.

5.2.1.10. *Taking stock.* Does the notion of universal effective compressibility constitute a natural notion of complexity? On the one hand, the formal correspondence between probability assignments and description systems means that a universal description length is a direct translation of a Solomonoff-Levin measure's goodness-of-fit; as such, the corresponding notion of simplicity-qua-compressibility does seem little more than a rewording of the particular assumption of Σ_1 effectiveness. On the other hand, there is clearly some intuitive appeal to the notion of compressibility associated with a description system; and we saw that a universal effective description system is arguably universal relative to *all* description systems, perhaps yielding a respectable notion of truly universal compressibility. However, I have suspended discussion of one important aspect, that perhaps presents the greatest challenge to this notion. This is the issue of the arbitrariness or subjectivity in the introduction of overhead constants, that I will call the issue of *variance*.

***Related approaches.** The equivalence between description length functions and probability functions is the cornerstone of the *minimum description length* (MDL) approach to statistical inference (see the textbook Grünwald, 2007 and the overview paper De Rooij and Grünwald, 2011). The development of MDL was influenced by Solomonoff's ideas, and the two approaches, including the role they attribute to Occam's razor, are often not clearly distinguished. This is a mistake: MDL and Solomonoff's approach part ways in crucial respects. While Rissanen, the founder of MDL, acknowledges that the "main source of inspiration in developing the *MDL* principle for general statistical problems has been the theory of *algorithmic* complexity [algorithmic information theory]," he is quick to add that "the role of the algorithmic complexity theory is inspirational, only, for almost everything about it, such as the idea of a model and even the very notion of complexity, must be altered to make the ideas practicable" (1989, 10).

One can say that in Solomonoff's theory there is only one statistical model: the class of all Σ_1 hypotheses; and the prediction method that is universal for this class induces a complexity measure on individual data sequences. Now one could use this model to define a data sample's minimum description length: this is the sample's Kolmogorov complexity, and this approach also goes by the name of "idealized MDL" (Grünwald, 2007). But this is only an idealized relative of "practical MDL," that is intended to apply to models one normally encounters in statistics. To emphasize, in practical MDL there is no longer a special role for effective computability, whereas this was the key ingredient in Solomonoff's proposal; and this leads to further important differences. The modern or "refined" variant of MDL (first summarized in Barron et al., 1998) rests on the design of universal codes for given statistical models, and, crucially, the central notion of complexity pertains to these models rather than to individual data sequences or single hypotheses (Grünwald, 2007, 30ff). Consequently, the resulting simplicity bias is more akin to that in model selection approaches like the *Bayes factor* method (ibid., 539ff; see Kass and Raftery, 1995). This also means that my discussion in this chapter of Occam's razor in Solomonoff's approach has no direct ramifications for practical MDL.

Idealized MDL is subsumed by the general approach of "nonprobabilistic" or "algorithmic statistics" based on algorithmic information theory, that originated in the *structure function* that Kolmogorov proposed in two talks in 1974 (see Vitányi, 2005; Li and Vitányi, 2008, 401ff). The starting idea is to decompose a data sample's shortest description into two parts: one describing a hypothesis that covers its structure, and one describing the remaining noise. (For instance, a sequence's shortest description via universal machine U, that gives its Kolmogorov complexity via U, can be decomposed into a part giving an index for a machine—the structural part—and a part giving the input to this machine—the noise.) The length of the structural part gives the data sample's 'structural complexity' or *sophistication* (Koppel and Atlan, 1991). However, it appears to be hard to specify this in a way that avoids triviality (where the structural part is either the complete description or is always given by the same universal hypothesis), while retaining $O(1)$-invariance under different enumerations of the hypotheses; and recently Bloem et al. (2015) have argued that this is in fact impossible.[24] This apparent impossibility—in particular, of a principled decomposition of universal description lengths into separate quantities for a particular machine and an input for that machine—ties in with other results that exhibit the great flexibility we have in characterizing universal elements, which I turn to below (particularly, 5.2.2.7).

5.2.2. The issue of variance. If any choice of overhead constants for the effective description systems gives a universal description system that is as valid as the next one, does this not make such a choice fully arbitrary? And if this is so, do we not end up with a notion of universal compressibility that leaves way too much variance, that is way too loose to be meaningful?

5.2.2.1. *The invariance theorem.* The standard reply to this issue is the invariance theorem, that states that any two choices of overhead constants are equivalent up to an additive constant (Li and Vitányi, 2008, 104ff, 200ff; also recall 3.2.5.4 above). This is Kolmogorov (1993, 221; also see Shiryaev, 1989, 921):

> The intuitive difference between "simple" and "complicated" objects has apparently been perceived a long time ago. On the way to its formalization, an obvious difficulty arises: something that can be described in one language may not have a simple description in another and it is not clear what method of description should be chosen. The main discovery, made by myself and simultaneously by R. Solomonoff, is that by means of the theory of algorithms it is possible to limit this arbitrary choice, defining complexity in an almost invariant way (the replacement of one method of description by another only leads to the addition of a bounded summand).

The invariance theorem, hinted at by Solomonoff in (1964, 11ff) and independently stated by Kolmogorov (1965, 5f) and Chaitin (1969, 156f), marks the birth of algorithmic information theory (see Li and Vitányi, 2008, 95ff, 192).

5.2.2.2. *Invariance and universality.* To be precise, the invariance theorem says that for any two universal effective description systems D_1 and D_2, there are constants c_1, c_2 such that for every sequence $\boldsymbol{y}$ we have $L_{D_1}(\boldsymbol{y}) \leq L_{D_2}(\boldsymbol{y}) + c_1$ and $L_{D_2}(\boldsymbol{y}) \leq L_{D_1}(\boldsymbol{y}) + c_2$. It is important to note that invariance does not necessarily follow for two description systems that are just universal for some $\mathcal{D}$ in the sense of 5.2.1.6 above; a mutual divergence that remains within a single overhead constant is only guaranteed if the universal systems are in $\mathcal{D}$ themselves. Again, it is this property of including universal elements that makes the class of effective description systems special.

5.2.2.3. *Two perspectives.* The bearing of the invariance theorem is that "from an asymptotic perspective, the complexity ... does not depend on accidental peculiarities of the chosen optimal method" (Kolmogorov, 1983, 33). Given a universal description system, as we investigate increasingly long sequences, that will have increasing universal description lengths, the effect of the arbitrary element in the description lengths becomes "disappearingly small" (Kolmogorov, 1969, 207). More concretely: given two different universal description methods, as we investigate increasingly long sequences, the difference in the two description lengths will become increasingly insignificant. Note, however, that this presupposes a particular perspective: I fix some universal description system, you fix another; then for any sequence we investigate the description lengths will not differ more than a constant. An alternative perspective is this: I fix some universal description system, and for any sequence I investigate, you can choose another universal description system such that the two description lengths for this sequence *diverge arbitrarily much.* From this perspective, my (and your) universal description lengths as a quantitative measure of complexity of finite sequences do look arbitrary (Kelly, 2008, 324f) or even "meaningless" (Muchnik et al., 1998, 315).

5.2.2.4. *The order of complexity.* A view of Kolmogorov complexity that is in line with the first perspective, and that indeed seems to be how many in the field think about the notion, is that it allows us to give an expression of the *growth* or the *order* of the complexity of a data stream. We can, within a variance of order $O(1)$, still distinguish, for instance, data streams for which the Kolmogorov complexity is of order $O(\log t)$ from those of order $O(1)$ or of order $O(t)$. This does constitute a certain coarse-graining that shifts focus to data streams or *infinite* sequences, though; it is consistent with the view that the complexity value of a single given finite sequence is "meaningless."

5.2.2.5. *Privileged systems.* An intuition that does try to revive the meaningfulness of the complexity of individual finite sequences, one that goes beyond the invariance theorem, is that there will be a small number of 'natural' or 'reasonable' choices of universal description methods, that do not differ much among each other (Kolmogorov, 1965, 6; Muchnik et al., 1998, 315). This is an instance of the hope for natural or objective universal machines, that I mentioned and questioned in 3.2.5.7 above. But even if it were possible to isolate a subclass of privileged universal description methods, that *do* induce a clear-cut notion of complexity, this would necessitate a serious modification of the argument for the justification of Occam's razor in 5.1 above. Namely, we would need to have some stronger reliability result that is restricted to the subclass of predictors that use *this* notion—or it would not be a simplicity preference that drives the reliability.

5.2.2.6. *Privileged priors.* Consider the proposal of Li and Vitányi (2008, 295) of a Solomonoff-Levin measure (Σ_1 mixture) with the particular weight function $v : i \mapsto 2^{-K(i)}$, a "mixture of hypotheses" that assigns "greater weight to the simpler ones" and can thus "be viewed as a mathematical form of Occam's razor" (ibid., 358). One obvious flaw in this proposal is that the problem of subjectivity just reappears at the level of choosing the universal machine to define the prefix-free Kolmogorov complexity K. (We can push the problem further and further away, but that will not make it disappear.) But, again, if the Solomonoff-Levin predictors with this particular prior are to figure in a justification of Occam's razor, we would also have to demonstrate that they have some special reliability properties. There is the argument by Hutter (2003b, 990f; 2005, 102f) that these predictors are *optimal* among all Σ_1 mixtures, for the reason that this weight function is *universal* (see 2.2.2.6 above). This would imply that the constant for this weight function in proposition 3.4 about the mixture's regret bound dominates those for other weight functions. Proposition 2.18, however, shows that this optimality is meaningless: *every* Σ_1 mixture can be represented so as to have a universal weight function.

5.2.2.7. *The permissiveness of universality.* The previous example brings me back to the fact that Σ_1 universality, despite the invariance theorem, *is an extremely permissive notion.* We have a great amount of freedom in characterizing Σ_1 universal elements; and this casts doubt on any interpretation that is grounded in the peculiarities of just a single one of those characterizations. For instance, even if a particular Solomonoff-Levin can be represented as a mixture with a *universal* weight function (with the interpretation of 5.2.2.6: an optimal element!), this is just one representation among many: we can represent the very same element with many other nonuniversal weight functions, even including (proposition 2.17) computable ones. None of those other representations admits of this optimality interpretation (save via the blunt fact of formal equivalence to the original representation), so this seems an artifact of a particular choice of characterization rather than an indicative property of these elements. (This was confirmed by proposition 2.18 that indeed every Solomonoff-Levin measure can be represented as such.) Something similar could be said about the compressibility interpretation that I traced in 5.2.1. I introduced description systems and interpreted the accompanying description length functions as measures of compressibility, to arrive at the universal description systems that are given by the universal monotone machines. I noted that the definition of a universal description length function is precisely the negative logarithm of a universal transformation λ_U of the uniform measure (i.e., a Solomonoff-Levin measure)—essentially because a sequence's length is the negative logarithm of its uniform probability. But, again, a Solomonoff-Levin measure's representation as a *uniform* transformation (interpretation: a notion of compressibility!) is just one among many: by theorem 2.13, we can represent the very same element as a transformation of any continuous computable measure. None of those other representations admits of this compressibility interpretation (save via the blunt fact of their formal equivalence to the original representation), so, again, this would rather seem an artifact of a particular choice of characterization.

5.2.2.8. *Predictive complexity.* The robust property of the Solomonoff-Levin measures is their universality within the class of Σ_1 measures. The claim that these elements also incorporate a natural notion of simplicity of data sequences, a notion that is more than a rephrasing of the inductive assumption of Σ_1 effectiveness, proceeds by the compressibility interpretation of a particular representation. But this interpretation is confronted with the fact that Σ_1 universality is such a permissive notion that the relevant representation is just one among an endless number of representations, in none of which this interpretation is manifest. As it stands, then, the case for a natural notion of simplicity remains inconclusive, and the argument for the justification of Occam's razor cannot be made to work. Nevertheless, this leaves us free to talk about a specialized complexity notion that is quite explicitly an expression of a Solomonoff-Levin predictor's success. We can still say that we have a specialized notion of the complexity of a data sequence that is an explicit expression of how difficult it is for a Solomonoff-Levin method to predict it. This brings us to Vovk's notion of *predictive complexity*, the topic of the next chapter.

*

CHAPTER 6

Predictive complexity

This chapter examines Vovk's notion of predictive complexity of data sequences. Predictive complexity, via a given loss function, is an expression of cumulative loss that is again universal in a precise sense. The main instance of predictive complexity, via the logarithmic loss function, is precisely the cumulative loss of the Solomonoff-Levin predictor. In general, different loss functions specify different games in the theory of prediction with expert advice, and the predictive complexity for each such game is given by the loss of an aggregating algorithm that appropriately generalizes the standard Bayesian updating employed by the Solomonoff-Levin predictor.

In 6.1, I discuss the relevant part of the theory of prediction with expert advice. In 6.1.1, I specify the general framework of a game of prediction with expert advice. In 6.1.2, I introduce the log-loss game and critically discuss various motivations for the logarithmic loss as the preferred loss function in sequential prediction. In 6.1.3, I introduce the Brier game and the absolute-loss game. In 6.1.4, I discuss Vovk's aggregating (pseudo) algorithm for a given game, and the important concept of mixability: for games that are mixable, this aggregating algorithm defines a predictive complexity. In 6.2, I explain and criticize the notion of predictive complexity.[25]

Innovations. The largest part of section 6.1—specifically, 6.1.1, 6.1.3, and 6.1.4—summarizes existent theory, but I hope to have succeeded at providing a concise yet clear presentation of some ideas that are not so easily accessible (particularly, the aggregating (pseudo) algorithm and the definition of mixloss). I know of no earlier critical discussion of the different motivations for the log-loss function in sequential prediction, as given in 6.1.2. In particular, the identification of the property of sequential locality, and the fact that this is a property of the log-loss function only, proposition 6.1, appears to be novel. I also know of no earlier critical discussion of Vovk's notion of predictive complexity, as given in 6.2. Proposition 6.2, that constitutes an important component of my critique, follows directly from older results, but has, I think, not been employed in this context before.

6.1. Games and loss functions

6.1.1. A game of prediction with expert advice. I repeat the specification of a prediction game, that I first described slightly more informally in I.1, including the notions of cumulative loss and regret, that I first specified in 3.3.2 above.

A *game* $\mathfrak{G} = (\Omega, \Gamma, \ell)$ consists of an *outcome space* Ω, a *prediction space* Γ, and a *loss function* $\ell : \Gamma \times \Omega \to [0, \infty]$. We restrict ourselves here to the familiar outcome space $\Omega = \mathbb{B}$, and the prediction space $\Gamma = \mathcal{P}$ of distributions over $\mathbb{B}$. Then a prediction strategy p is as before a function from finite outcome sequences in $\mathbb{B}^*$ to distributions over $\mathbb{B}$, the possible next symbols. At each trial $t + 1$, having observed sequence $\boldsymbol{x}^t$, the strategy issues prediction $\mathsf{p}(\boldsymbol{x}^t)$, after which the outcome $x_{t+1} \in \mathbb{B}$ is revealed and the strategy suffers an *instantaneous loss* given by $\ell(p, x_{t+1})$ with $p = \mathsf{p}(\boldsymbol{x}^t)$. To slightly economize notation, I will write this as '$\ell_{\mathsf{p}}(x_{t+1})$,' thus also leaving the preceding sequence $\boldsymbol{x}^t$ implicit. Always assuming a countable pool of prediction strategies indexed by $k \in \mathbb{N}$, I will further simply write '$\ell_k(x_{t+1})$' for the instantaneous loss of predictor p_k at $t + 1$. The *cumulative* loss suffered by strategy p on a sequence $\boldsymbol{x}^s$ is the sum

$$L_{\mathsf{p}}(\boldsymbol{x}^s) := \sum_{t=0}^{s-1} \ell_{\mathsf{p}}(x_{t+1}) \tag{50}$$

of instantaneous losses. (Likewise 'L_k' stands for the sum of instantaneous losses of p_k.)

Given a game (that is, a loss function), we seek to formulate a prediction strategy that, having available at each trial $t + 1$ all the predictions of some countable pool of prediction strategies or *experts*, incurs a cumulative loss that is *never* very bad compared to *any* of these experts. Defining the *regret* $R_{\mathsf{p},k}(\boldsymbol{x})$ of predictor p relative to predictor p_k on finite sequence $\boldsymbol{x}$ as the excess loss incurred by p on this sequence, i.e.,

$$R_{\mathsf{p},k}(\boldsymbol{x}) := L_{\mathsf{p}}(\boldsymbol{x}) - L_k(\boldsymbol{x}),$$

we seek to formulate a strategy p that manages a low regret relative to any other expert p_k on *every* data stream.

One way this goal might be construed is to find a bound on the worst-case regret

$$\max_{\boldsymbol{x} \in \mathbb{B}^*, \mathsf{p}_k \in \mathcal{H}} R_{\mathsf{p},k}(\boldsymbol{x}), \tag{51}$$

i.e., a single bound on the regret relative to any expert, on every single finite sequence. This, however, is too ambitious in the general case of a pool of infinitely many experts, as I will always assume here. Consider the pool that contains all computable experts: for every finite sequence there will be a predictor that has predicted perfectly and hence incurred loss 0, so (51) actually reduces to the cumulative *loss* of p. Bounding (51) would come down to bounding p's loss on all possible data sequences.

However, we *can* derive meaningful bounds on the regret $R_{\mathsf{p},k}(\boldsymbol{x})$ that depend on p_k. Without any assumptions whatsoever on the experts and the origin of the data, we can still, it turns out, formulate general strategies such that for every p_k we have a meaningful bound on the worst-case regret

$$\max_{\boldsymbol{x} \in \mathbb{B}^*} R_{\mathsf{p},k}(\boldsymbol{x}).$$

It is in this sense that we can design strategies that, compared to *any* given expert in the pool, will *never* do much worse than this expert.

In fact, we are already familiar with the relevant strategy for one type of game, a strategy, it turns out, that we can generalize for other games: the Bayesian mixture over all experts, for the *log-loss game*.

6.1.2. The log-loss game. Indeed, in 3.3.2 above I already described the game for the logarithmic or simply log-loss function, defined by $\ell(p, x) := -\ln p(x)$. (In the current context it is customary to use the base e rather than the base 2 logarithm; this, of course, only makes for a difference of a multiplicative constant.) To repeat, the instantaneous log-loss of a predictor p_k is then given by

$$\begin{aligned} \ell_k(x_{t+1}) &= -\ln \mathsf{p}_k(x_{t+1}, \boldsymbol{x}^t) \\ &= -\ln \mu_k(x_{t+1} \mid \boldsymbol{x}^t), \end{aligned}$$

where μ_k is the measure corresponding to p_k; and by the telescoping effect the cumulative log-loss simplifies to

$$\begin{aligned} L_k(\boldsymbol{x}^s) &= \sum_{t=0}^{s-1} - \ln \mu_k(x_{t+1} \mid \boldsymbol{x}^t) \\ &= -\ln \prod_{t=0}^{s-1} \mu_k(x_{t+1} \mid \boldsymbol{x}^t) \\ &= -\ln \mu_k(\boldsymbol{x}^s). \end{aligned} \tag{52}$$

The game with the log-loss function is called the *log-loss game*.

6.1.2.1. *Conditions on loss functions.* A *scoring rule* is a function that is to express how good a probabilistic prediction was in light of an actual outcome (what the "value" or "utility" of a particular prediction was, in a context of "pure inference" where our only goal is to predict accurately, Bernardo and Smith, 1994, 69ff). Its counterpart, a loss function, is to express how *bad* a prediction was in light of an actual outcome (what the *cost* or *loss* of a particular prediction was if our goal is to predict accurately). There are a couple of basic requirements one wants to impose on a such a function $\ell : \mathcal{P} \times \mathbb{B} \to [0, \infty]$. It is natural to require that $\ell(p, x) = 0$ if (and only if) $p(x) = 1$; and that $\ell(p, x)$ is monotonically increasing in $p(x)$. It is natural to require a condition of *smoothness* (say, continuous differentiability in $p(x)$), with the motivation, apart from being mathematically convenient, that small differences in predictions should only lead to small differences in loss. It is natural, in the presupposed context of pure inference (as opposed to contexts where, for instance, one unexpected outcome has more serious consequences than another), to require a *symmetry* condition to the effect that $\ell(p_1, 0) = \ell(p_2, 1)$ if and only if $p_1(0) = p_2(1)$. Moreover, it is natural to require that the loss function should induce a predictor to be *honest*, by precluding that the p-expected loss is actually minimized by a prediction different from p. Accordingly, the requirement of *propriety* (Murphy and Epstein, 1967; Good, 1952, 112; also see Gneiting and Raftery, 2007) is that the p-expected loss should always be minimized at prediction p, i.e.,

$$\arg\min_{p'} \mathbf{E}_{X \sim p}\, \ell(p', X) = p. \tag{53}$$

The property of propriety extends to sequences of outcomes, a property that I will call *sequential propriety* (see B.2.4.1 for the definition and proof).

6.1.2.2. *Motivation for the log-loss function.* The log-loss function satisfies the preceding requirements—but so do other functions, including the *Brier* loss function discussed in 6.1.3.1 below. Are there reasons why we should still prefer the log-loss function? One can find several reasons in the literature, both of a technical and of a more conceptual nature (see Merhav and Feder, 1998, 2127f).

6.1.2.3. *Motivation for the log-loss function: telescoping.* What I should stress, once more, as a significant technical benefit of working with the log-loss function, is its telescoping property (52)—which will in fact assume an important role even in the analysis of games *other* than the log-loss game (6.1.4.2 below).

6.1.2.4. *Motivation for the log-loss function: sequential locality.* In the case of more than two possible outcomes, the log-loss function is the only proper loss function that is also *local*, meaning that $\ell(p, x)$ only depends on $p(x)$, the probability assigned to the outcome that actually obtained (Bernardo, 1979; it is actually more accurate to say 'loss functions of logarithmic form,' see B.2.4.2 below). In our case of distributions over two outcomes, however, locality is vacuously satisified; and many other loss functions are possible. (Also see Bernardo and Smith, 1994, 72ff.) Nevertheless, if we look at the loss over *sequences* of outcomes, i.e., the cumulative loss, there is a corresponding form of *sequential locality*, that is again only satisfied by loss functions of logarithmic form. Namely, the cumulative log-loss $L_{\mathsf{p}}(\boldsymbol{x}^s)$ over sequence $\boldsymbol{x}^s$ of outcomes, by the telescoping property, is a function of $\mu_{\mathsf{p}}(\boldsymbol{x}^s)$ for the measure μ_{p} corresponding to p; whereas in general the cumulative loss is a function of all the conditional probabilities $\mu_{\mathsf{p}}(x_{t+1} \mid \boldsymbol{x}^t)$ corresponding to the $\mathsf{p}(x_{t+1}, \boldsymbol{x}^t)$'s for $t < s$.

Proposition 6.1. Loss functions of logarithmic form, and only those loss functions, are sequentially local.

Proof. See B.2.4.3. □

6.1.2.5. **Sequential locality: example.* Here is a simple illustration of the failure of locality, starring the absolute-loss function defined in 6.1.3.2 below. Take the sequence 00 and two predictors p_1 and p_2 with corresponding μ_1, μ_2 such that $\mu_1(0) = \mu_1(0 \mid 0) = \frac{1}{2}$ and $\mu_2(0) = \frac{1}{4}$, $\mu_2(0 \mid 0) = 1$. Then the probability assignments for the complete sequence 00 are identical, $\mu_1(00) = \mu_2(00) = \frac{1}{4}$, yet (the more cautious) predictor p_1 is penalized more severely, $L_{\mathsf{p}_1}(00) = 1 \neq L_{\mathsf{p}_2}(00) = \frac{3}{4}$.

6.1.2.6. *Motivation for the log-loss function: sequential locality, cont.* Does sequential locality make for a property that we should impose on a loss function? The original locality property rests on a very minimal demand: the accuracy of two predictions p_1 and p_2 in light of a particular outcome x should be judged the same if $p_1(x) = p_2(x)$, regardless of possibly differing probability values assigned to counterfactual outcomes that did not actually materialize. The property of sequential locality rests on a stronger demand: the accuracy of two predictors p_1 and p_2 in light of a sequence of outcomes $\boldsymbol{x}$ should be judged the same if $\mu_{\mathsf{p}_1}(\boldsymbol{x}) = \mu_{\mathsf{p}_2}(\boldsymbol{x})$ for the corresponding total probability assignments to $\boldsymbol{x}$, regardless of how these are built from the one-step predictions or conditional probabilities on $\boldsymbol{x}$. To restate, the difference is that sequential locality is not a matter of disregarding probabilities of counterfactuals, but of disregarding conditional probabilities that make up the total probability. To further bring out the contrast, consider the likely motivation for *rejecting* each property. An opening for rejecting locality and making accuracy dependent on all probabilities would be the desire to take into account how *cautious* the prediction p is (how flat the distribution p is)—but here it still needs explaining why for $p_1(x) = p_2(x)$ a difference in flatness elsewhere should impact the accuracy on x. A reason for rejecting sequential locality could be the desire to take into account how cautious the individual probability assignments to what turned out to be the actual outcomes were (how close to one half the probabilities were)—a desire that sounds sensible even in contexts of pure inference (Popper's methodology of making daring predictions comes to mind).

6.1.2.7. *Motivation for the log-loss function: information.* On a more conceptual level, there is the strong link of the log-loss function to *information theory.* The log-loss function is also called the *self-information* loss function: "[a]s is well known, the self-information manifests the degree of uncertainty, or the amount of information treasured in the occurrence of an event" (Merhav and Feder, 1998, 2127). What is alluded to here is the interpretation of the term

$$h(x) := -\log p(x)$$

as the *Shannon information content* of the outcome x, which ties in with the interpretation of a distribution's *Shannon entropy* (Shannon, 1948; also see A.2.3 below)

$$H(p) = \mathbf{E}_{X \sim p}\left[h(X)\right] \tag{54}$$

as its expected information content (see MacKay, 2003, 67ff). The idea is that $h(x)$ represents the amount of information we gain when outcome x obtains, as the extent of our *surprisal*: a low-probability outcome is highly surprising so very informative when it occurs, whereas a high-probability outcome is unsurprising and we gain little from observing it. Consequently, the entropy $H(p)$ is a measure of the *expected* amount of information to be gained from the next outcome; and as such, it is a measure of the *uncertainty* expressed by p: if we are quite uncertain about the next outcome (i.e., $p(0) \approx p(1) \approx \frac{1}{2}$) it will be informative to see it, but if we are very certain about the next outcome ($p(0) \approx 1$ or $p(1) \approx 1$) we do not expect to learn much from it. Now it is reasonable to choose a measure of information-as-surprisal for a loss function: the more suprised we are by an outcome, the less accurate our prediction turned out to be. What is left unmotivated, however, is why this analysis had to start with the function h—other than the bare fact that it is the function used in information theory. That is, going the other direction, could we not take any other proper loss function and likewise interpret it as a measure of information-as-surprisal? All we have done now is to reduce the problem of justifying the log-loss as the preferred loss function to the problem of justifying it as the preferred surprisal function, but it is actually not clear that this task is any easier. The apparent reason why it might be, again, is its status within information theory, featuring in the definition of entropy that is uniquely characterized by means of a number of axioms (1948, 392f). However, while it is standard to interpret the Shannon entropy as a measure of uncertainty (see, for instance, Cover and Thomas, 2006, 13ff), it is forcefully argued by Uffink (1990, 65ff; also see Timpson, 2013, 25ff) that *for the purpose of a measure of uncertainty* the last of Shannon's axioms is not well-motivated, and so the function H is still only one among a multitude of feasible definitions of a distribution's uncertainty. Indeed, as discussed in detail by Grünwald and Dawid (2004), we can define a notion of entropy for any given loss function by replacing h in (54) for *this* function. This blocks the strategy of justifying the choice of the log-loss function as the uniquely preferred choice of a measure of information-as-surprisal.

6.1.2.8. *Motivation for the log-loss function: datacompression.* The Shannon entropy *is* unique in a different sense. It gives a unique measure of the optimal compression of the elements generated from a given probability distribution, as follows (see A.2.3 for more details). The entropy $H(p)$ of distribution p over countable outcome space Ω, defined as the p-expected code length of the idealized prefix coding sytem $L_p(\cdot) = -\log p(\cdot)$ corresponding to p, is a lower bound on the p-expected code length of any distribution q's corresponding idealized coding sytem,

$$H(p) = \mathbf{E}_{X \sim p} \left[-\log p(X) \right] \leq \mathbf{E}_{X \sim p} \left[-\log q(X) \right], \tag{55}$$

with equality precisely if $p = q$; which implies it is a lower bound on *any* idealized prefix coding system's p-expected code length,

$$H(p) \leq \mathbf{E}_{X\sim p}\left[L_C(X)\right].$$

This is the *information inequality* (proposition A.10, A.2.3.1 below); further results show that the Shannon entropy likewise gives both a bound and a guarantee on the p-expected code length achievable with actual (nonidealized) coding systems (the source coding theorems, A.2.3.3–A.2.3.4 below). Interesting as this cluster of theory is, however, it does not yet amount to an argument in favour of the log-loss function. (The information inequality (55) as a property of the log-loss function, saying that the p-expected loss is minimized by prediction p, is certainly nice—but we already know that it is satisfied by other functions, too: it is the property (53) of propriety!) Nevertheless, this connection to information theory points at a particular interpretation of the log-loss function, that at least provides a further handle on understanding the importance it is attributed. This is the interpretation of log-losses as description lengths, and the goal of minimizing log-loss as *learning by datacompression*, the philosophy that underlies MDL, the principle of minimum description length. (See Rissanen, 1989 for the inventor's spirited defense of this conception of statistical inference, and Grünwald, 2007 for a somewhat more pragmatic presentation.) The fundament of MDL is the equivalence between probabilities and description lengths (ibid., 90ff, "The Most Fundamental Section of This Book"; also see A.2), and this strong structural similarity between probabilities and log-losses can be seen to explain such benign properties as the telescoping, the propriety, and the sequential locality of the log-loss function. Resting on this equivalence, the general intuition that both drives and is reinforced by work in MDL is that expressed in Grünwald's "Concluding Remark on the MDL Philosophy" (ibid., 595): "*if one has learned something of interest, one has implicitly also compressed the data*," which we can recast as: if one has learned something, one has also achieved a low log-loss. That this is not a trivial statement is supported by combinatorial observations to the effect that there is only a tiny fraction of possible outcomes that can be significantly compressed, indicating that it is hard to achieve low log-losses (see A.2.3.5 on the *no-hypercompression inequality*). Further, the statement itself, or rather the more precise version "if one achieves low loss via *any* reasonable loss function, then one achieves low log-loss" can also be grounded in formal results. Namely, Vovk (2001b) showed for the Brier loss function, and in (2015) extended this to a general class of loss-functions, that if an infinite outcome sequence is random under the log-loss function, it is random under the other loss function (i.e., the contrapositive statement that if it is not possible to achieve low loss under the log-loss function, it is not possible to achieve low loss under the other function; see A.4 for details on randomness).

6.1.2.9. *Motivation for the log-loss function: conclusion.* The previous collected various reasons in support of the log-loss function for measuring the accuracy of predictions. Most if not all of these reasons appear to be connected to the strong structural correspondence between log-losses and probabilities. We might say that the log-loss function provides the most faithful mapping between probabilities and losses. Still, this does nothing to discount the simple fact that there are other loss functions that are widely used (Vovk, 2015, 317), and specific contexts can indeed dictate different loss functions (e.g., 6.1.3.2 below).

6.1.3. Other games.

6.1.3.1. *The Brier game.* A prominent alternative loss function is the *square-loss* or *Brier* loss function (Brier, 1950; de Finetti, 1962), that, like the theory of proper loss functions in general, originated in meteorology and in parallel assumed an important role in foundational work in probability (see Dawid, 1986, 2008). In the case of two outcomes, we define it by

$$\ell(p, x) := (1 - p(x))^2,$$

i.e., $\ell(p, x) = p(\bar{x})^2$. A notable difference from the log-loss function is that the range of the Brier function is the interval $[0, 1]$, whereas the log-loss function has a range that is *unbounded* (indeed giving an infinite loss in the case of a probability assignment of 0 to the actual outcome; this is an aspect one may hold *against* the log-loss function, see Vovk, 2015, 317). The game with the Brier loss function we call the *Brier game.*

6.1.3.2. *The absolute-loss game.* The setting I always assume is *probabilistic* prediction: methods of prediction return probability distributions over the possible outcomes. This is arguably the obvious setting if we want to allow for the expression of uncertainty (Dawid, 1984, 278); at the same time, ordinary prediction is rather done in a *categorical* manner—either 0 or 1. The corresponding *simple game* is the game with prediction space $\Gamma = \Omega = \mathbb{B}$, and the *0/1-loss* ('zero-one loss') function given by

$$\ell(y, x) := \mathbb{1}_{x \neq y}.$$

The simple game is actually hard when it comes to the goal formulated in 6.1.1. An adversarial data sequence can make our strategy fail at every single point in time, while already the inclusion in the pool of experts of the two constant strategies ('always 0' and 'always 1') guarantees that the best expert fails no more than half of the time, leading to a regret relative to this expert at least as bad as $t/2$ (cf. Cesa-Bianchi and Lugosi, 2006, 67). A solution is to *randomize*: to use a probability distribution in $\mathcal{P}$ to decide whether to say 0 or 1. Formally, this gives the game where again $\Gamma = \mathcal{P}$, and the loss function is the *absolute loss* given by

$$\ell(p, x) := p(\bar{x}).$$

Unlike the log-loss and the Brier loss function, the absolute-loss function is *not* proper (for p with $p(1) > \frac{1}{2}$ the p-expected loss is minimized for p' with

$p'(1) = 1$); nor is it sequentially local (see the example in 6.1.2.4 above). Moreover, the *absolute-loss game* is still harder than the other two games, in a sense made precise in 6.1.4.5 and 6.1.4.6 below.

6.1.4. Aggregating strategies.

6.1.4.1. *The log-loss game: the mixture predictor.* Recall from 3.3.1 that the aggregating predictor $\mathsf{p}_{\mathrm{mix}(w)}$ is given by

$$\mathsf{p}_{\mathrm{mix}(w)}(\boldsymbol{x}^t) = \sum_k w(k \mid \boldsymbol{x}^t)\mathsf{p}_k(\boldsymbol{x}^t), \tag{56}$$

where w is a weight function over all strategies that is updated at each trial $t+1$ by

$$w(k \mid \boldsymbol{x}^t) = \frac{w(k \mid \boldsymbol{x}^{t-1})\mathsf{p}_k(x_t, \boldsymbol{x}^{t-1})}{Z}, \tag{57}$$

with $Z = \sum_k w(k \mid \boldsymbol{x}^{t-1})\mathsf{p}_k(x_t, \boldsymbol{x}^{t-1})$ a normalizing term. Recall, too, the optimality theorem 3.3 that says that the cumulative log-regret is bounded by a constant,

$$R_{\mathrm{mix}(w),k}(\boldsymbol{x}^t) \leq -\ln w(k) \tag{58}$$

for any k and $\boldsymbol{x}^t$, and w given by the mixture measure ξ_w that corresponds to $\mathsf{p}_{\mathrm{mix}(w)}$. So we have a strategy that, compared to any strategy in the given pool, never accumulates a loss that exceeds this strategy's cumulative loss by a fixed constant—a constant that is in fact an expression of the weight this aggregating predictor assigned to this strategy. (Cesa-Bianchi and Lugosi, 2006, 47, 55f refer to bounds like (58) as *oracle inequalities.*) To put it another way, for any strategy in the pool, our aggregating strategy's *average loss per outcome*,

$$\frac{L_{\mathrm{mix}(w)}(\boldsymbol{x}^t)}{t},$$

converges to this strategy's average loss at a rate faster than $O(1/t)$, if it does not even become smaller.

6.1.4.2. *Towards other games.* The pleasant property of the logaritmic loss function is that it retains the telescoping effect of the conditional measures corresponding to the predictors, as shown in (52); this is what allows us to directly infer the regret bound (58) on the aggregating predictor from the dominance of the mixture measure. Unfortunately, we cannot rely on this property in the case of other loss functions. We can, however, try to mimic this effect. What we can do is formulate a mixture directly over the predictors' losses (according to the given loss function), in such a way that this mixture's loss will benefit from a telescoping effect and thus achieve a bound similar to (58). This *mix-loss* (terminology De Rooij et al., 2014) is the loss corresponding to Vovk's *aggregating (pseudo) algorithm* (Vovk, 1990; also see Vovk, 1998, 2001a; Cesa-Bianchi and Lugosi, 2006, 52ff; Grünwald, 2007, 573ff), which I will now present as a generalization of the original mixture strategy.

6.1.4.3. *The generalized update rule.* Note that we can rewrite the update rule (57) directly in terms of the log-loss function, as

$$w(k \mid \boldsymbol{x}^t) = \frac{w(k \mid \boldsymbol{x}^{t-1})e^{-\ell_k(x_t)}}{Z}. \tag{59}$$

This we can generalize to any given loss function. Introducing, in addition, a parameter $\eta > 0$ that we call the *learning rate*, we define, for any ℓ, the generalized update rule

$$w_\eta(k \mid \boldsymbol{x}^t) := \frac{w_\eta(k \mid \boldsymbol{x}^{t-1})e^{-\eta\ell_k(x_t)}}{Z}, \tag{60}$$

with normalizing term $Z = \sum_k w_\eta(k \mid \boldsymbol{x}^{t-1})e^{-\eta\ell_k(x_t)}$. (Thus (57) is (60) with the log-loss function and $\eta = 1$.) Compared to the base case of $\eta = 1$, the larger a choice of $\eta > 1$, the more the inflation of differences in instantaneous losses, the greater the change in weights, and the more aggressive the learning, while the smaller a choice of $\eta < 1$, the more the deflation of those differences, and the more conservative the learning; hence the name, learning rate.

6.1.4.4. *The mix-loss.* Generalizing the definition of the log-loss of the original mixture predictor $\mathsf{p}_{\text{mix}(w)}$, i.e.,

$$\ell_{\text{mix}(w)}(x_{t+1}) = -\ln \sum_k w(k \mid \boldsymbol{x}^t)e^{-\ell_k(x_{t+1})},$$

we define the *instantaneous mix-loss* as

$$\ell_{\text{mix}(\eta,w)}(x_{t+1}) := -\eta^{-1} \ln \sum_k w_\eta(k \mid \boldsymbol{x}^t)e^{-\eta\ell_k(x_{t+1})}. \tag{61}$$

It is important to note that the mix-loss is different from the original loss function ℓ; I spell out the relation below. Crucially, the *cumulative mix-loss* again incorporates a telescoping effect. Observing that as before

$$w_\eta(k \mid \boldsymbol{x}^t) = \frac{w_\eta(k)e^{-\eta L_k(\boldsymbol{x}^t)}}{Z}$$

with $Z = \sum_k w_\eta(k) e^{-\eta L_k(\boldsymbol{x}^t)}$, we have

$$\begin{aligned} L_{\mathrm{mix}(\eta,w)}(\boldsymbol{x}^s) &= \sum_{t=0}^{s-1} \ell_{\mathrm{mix}(\eta,w)}(x_{t+1}) \\ &= \sum_{t=0}^{s-1} -\eta^{-1} \ln \sum_k w_\eta(k \mid \boldsymbol{x}^t) e^{-\eta \ell_k(x_{t+1})} \\ &= -\eta^{-1} \ln \prod_{t=0}^{s-1} \sum_k w_\eta(k \mid \boldsymbol{x}^t) e^{-\eta \ell_k(x_{t+1})} \\ &= -\eta^{-1} \ln \prod_{t=0}^{s-1} \sum_k \frac{w_\eta(k) e^{-\eta L_k(\boldsymbol{x}^t)}}{\sum_{k'} w_\eta(k') e^{-\eta L_{k'}(\boldsymbol{x}^t)}} e^{-\eta \ell_k(x_{t+1})} \\ &= -\eta^{-1} \ln \prod_{t=0}^{s-1} \frac{\sum_k w_\eta(k) e^{-\eta L_k(\boldsymbol{x}^{t+1})}}{\sum_{k'} w_\eta(k') e^{-\eta L_{k'}(\boldsymbol{x}^t)}} \\ &= -\eta^{-1} \ln \sum_k w_\eta(k) e^{-\eta L_k(\boldsymbol{x}^s)}. \end{aligned}$$

That means we can readily infer the bound

$$\begin{aligned} R_{\mathrm{mix}(\eta,w),k}(\boldsymbol{x}^s) &\leq -\eta^{-1} \ln \left[w_\eta(k) e^{-\eta L_k(\boldsymbol{x}^s)} \right] - L_k \\ &= -\eta^{-1} \ln w_\eta(k) \end{aligned} \tag{62}$$

for any predictor p_k. Again, this means that, for any strategy p_k, if the mean mix-loss does not already drop below p_k's mean loss, it will at least converge to that of p_k at a rate $O(1/t)$.

6.1.4.5. *Mixable games.* The catch here is that the mix-loss (61) might be *too* good: it might not correspond to a possible prediction! (Note that we defined the mix-loss $\ell_{\mathrm{mix}(\eta,w)}$ as a direct generalization of the log-loss of a mixture predictor, without first defining an actual generalized mixture predictor $\mathsf{p}_{\mathsf{mix}(\eta,w)}$. For ℓ again the log-loss, and $\eta = 1$, this generalized mixture does exist, but for other loss functions, it might not.) This is why Vovk used the term "aggregating *pseudo* strategy" (APA). The derived *aggregating strategy* (AA) sets this straight by inserting the step of picking an actual prediction with a loss that matches the mix-loss as closely as possible. To be precise, let $\mathcal{G}$ be the class of *generalized predictions*, or instantaneous loss functions of the form

$$g_w : x \mapsto -\eta^{-1} \ln \sum_k w(k) e^{-\eta \ell_k(x)};$$

then the AA employs a *substitution function* $\Sigma : \mathcal{G} \to \mathcal{P}$ that maps a generalized prediction g to an actual prediction p. Now the game might be such that for given η we can actually find a *perfect* substitution function Σ that satisfies

$$\ell_{\Sigma(g)}(x) \leq g(x) \tag{63}$$

for all g and x. That means that there actually always exist allowed predictions that suffer a loss no worse than the mix-loss. Hence the AA, employing such a perfect substitution function, satisfies

$$L_{\mathrm{AA}(\eta,w)}(\boldsymbol{x}^t) \leq L_{\mathrm{mix}(\eta,w)}(\boldsymbol{x}^t),$$

and therefore gives an actual strategy that satisfies the regret bound (62). A game for which this holds for given η we call *η-mixable*; we call a game *mixable* if it is η-mixable for some $\eta > 0$. (The mixability of a game also has a neat characterization as the convexity of the set of pairs $(e^{-\eta a}, e^{-\eta b)})$ with $\ell(p, 0) \leq a, \ell(p, 1) \leq b$ for some prediction p, see Vovk, 1998; Cesa-Bianchi and Lugosi, 2006, 54.) The log-loss game is obviously mixable, for $\eta \leq 1$ (see Vovk, 1998, 156): for $\eta = 1$ the AA coincides with the aggregating predictor. Less obviously, the Brier game is mixable, too, and it is in fact so for $\eta \leq 2$ (Vovk, 1990; Haussler et al., 1995; see Vovk, 1998, 156; note that the bound (62) is stronger as η is larger). It is for the mixable games that we can define a notion of *predictive complexity* (6.2.1 below).

6.1.4.6. *Non-mixable games.* But not all games are mixable—the absolute-loss game is a case in point. What we can say in full generality is that there always exists a substitution function Σ with

$$\ell_{\Sigma(g)}(x) \leq c(\eta) \cdot g(x) \tag{64}$$

where

$$c(\eta) := \inf\{c : \forall g \in \mathcal{G}\ \exists p \in \mathcal{P}\ \forall x \in \mathbb{B}\ \ell(p, x) \leq c \cdot g(x)\}. \tag{65}$$

(Note that $c(\eta) \geq 1$, with equality if the game is in fact mixable.) That means that the AA always satisfies

$$L_{\mathrm{AA}(\eta,w)}(\boldsymbol{x}^t) \leq c(\eta) L_{\mathrm{mix}(\eta,w)}(\boldsymbol{x}^t),$$

so for any k we have

$$L_{\mathrm{AA}(\eta,w)}(\boldsymbol{x}^t) \leq c(\eta) L_k(\boldsymbol{x}^t) - \frac{c(\eta)}{\eta} \ln w_\eta(k),$$

or

$$R_{\mathrm{AA}(\eta,w),k}(\boldsymbol{x}^t) \leq (c(\eta) - 1)\, L_k(\boldsymbol{x}^t) - \frac{c(\eta)}{\eta} \ln w_\eta(k).$$

So the AA's mean loss converges to at most a factor $c(\eta)$ of the (best) strategy p_k's mean loss. Again, if $c(\eta) = 1$, this means that the mean loss per outcome will at least be as good as that of p_k (6.1.4.4 above); but if $\eta > 1$ this bound is rather less interesting, as it is still consistent with the AA suffering every single round more than a positive ϵ loss than the best expert (with ϵ depending on the magnitude of $c(\eta) - 1$ and the minimal loss per round of the best expert). Nevertheless, even if $c(\eta) > 1$ for all η, it might still be the case, as it indeed is for the absolute-loss game, that $c(\eta) \to 1$ as $\eta \to 0$. That also still allows us to define a notion of *weak predictive complexity* (6.3.1 below).

* * *

6.2. Predictive complexity

This section investigates the notion of predictive complexity that arises from the theory of prediction with expert advice.

In 6.2.1, I lay out the definition of predictive complexity. In 6.2.2, I discuss the question whether the definition succeeds to give a natural notion of the difficulty of prediction of data sequences.

6.2.1. The definition.

6.2.1.1. *The pool of all experts.* The theory of prediction with expert advice is about designing aggregating strategies that never perform much worse than the best in the given pool of prediction methods or experts they aggregate over. This leaves open the question of how we have selected the pool of experts in the first place (see Vovk and Watkins, 1998, 16ff). An important part of the appeal of the theory resides in it explicitly involving no assumptions whatsoever on the origin of the data (recall 3.3.1 above); but such assumptions do threaten to come in through the back door when our selection of a particular pool is guided by the expectation that it contains experts that do well. A choice of pool of experts is only truly assumptionless—and a strategy satisfying the constant regret bound (62) for this pool is only truly *universally* good—if it is the pool of *all possible prediction methods.* Note that I have begun to retrace a familiar story: the next step is to assert that there is an obvious constraint on the possible prediction methods, namely *effective computability.* This is indeed the step taken by Vovk and Watkins (1998, 16):

> instead of imposing restrictions on the Environment part we can impose restrictions on the Learner part. Indeed, such Learner-side limitations are very natural: we know that she must *compute* her strategy . . . Instead of pools reflecting our beliefs we can use pools reflecting Learner's limitations such as the pool of all computable strategies . . .

We proceed the analysis again directly on the level of the effective predictors' *losses.*

6.2.1.2. *The class of superloss processes.* For a given game (loss function ℓ), I shall now refer to the cumulative loss function L_{p} of a prediction method p as a *loss process.* Assuming that ℓ is computable (as the usual loss functions are), the computable or Δ_1 loss processes are precisely those corresponding to the computable prediction methods. A loss process with a constant regret bound relative to each computable loss process would then make for a truly universal loss process (i.e., correspond to a truly universal prediction method)—were it not for the fact, all too familiar by now, that such a universal element must itself fall outside the class of computable elements (ibid., 17):

> It would be ideal if the class of computable loss processes contained a smallest (say, to within an additive constant) element. Unfortunately, for the loss functions in our games such a smallest element does not exist: given a computable prediction strategy S, it is easy to construct a computable prediction strategy that greatly outperforms S on at least one ... outcome sequence.

The proposed way to deal with this is familiar now, too. "Levin suggested (for a particular game, the log-loss game; see below) a very natural solution to the problem of the non-existence of a smallest computable loss process" (ibid.): we extend the class of Δ_1 loss processes to the class of Π_1 *superloss processes*, functions $L : \mathbb{B}^* \to [0, \infty]$ that

(a) are upper semi-computable, or Π_1 (this is the complementary property to lower semi-computability or Σ_1: function f is upper semi-computable precisely if $-f$ is lower semi-computable);

(b) satisfy $L(\varnothing) = 0$, and for all $\boldsymbol{x} \in \mathbb{B}^*$, for some $p \in \mathcal{P}$, for both $x \in \mathbb{B}$, $L(\boldsymbol{x}x) \geq L(\boldsymbol{x}) + \ell(p, x)$.

That is, we expand a diagonalizable class of computable elements to a non-diagonalizable class of semi-computable elements. To prevent the collapse to full computability we needed to weaken the notion of a loss process to the definition (b) above; we can also interpret such a function as the loss process corresponding to a method that makes *superpredictions* p' that are no better than actual predictions, i.e., that satisfy $\ell(p', x) \geq \ell(p, x)$ on both $x \in \mathbb{B}$ for some actual prediction p. (Perhaps '*sub*prediction' would have been a better term.) The result is again that we have opened up the possibility of universal elements—for the *mixable* games, anyway.

6.2.1.3. *Predictive complexity.* For the mixable games, our earlier work in 6.1.4.2–6.1.4.5 above immediately rewards us with a universal Π_1 superloss process: the mix-loss! To put it more precisely: for an η-mixable game, we have that the mix-loss $L_{\text{mix}(\eta,w)}$, for any semi-computable w over all Π_1 superloss processes, is itself a Π_1 superloss process (one can easily verify (a) and (b), see ibid., 22), and by design fulfills (62) or

$$L_{\text{mix}(\eta,w)} \leq^{+} L_k$$

for any Π_1 superloss process L_k. This, then, gives Vovk's measure of an η-mixable game's *predictive complexity* $K_w : \mathbb{B}^* \to \mathbb{R}$ of data sequences (ibid., 17). It is defined to be the mix-loss $L_{\text{mix}(\eta,w)}$ with semi-computable w over all Π_1 superloss processes.

6.2.1.4. *The log-loss game.* Consider again the particular case of the log-loss game. Here the loss processes are the cumulative log-loss functions, so the functions $L(\cdot) = -\ln \mu(\cdot)$ corresponding to all measures μ (6.1.2 above); and the *computable* loss processes are clearly those corresponding to the *computable* or Δ_1 measures. What about the Π_1 superloss processes? Starting with condition (b), the functions L satisfying, for all $\boldsymbol{x}$,

$$L(\boldsymbol{x}x) \geq L(\boldsymbol{x}) - \ln p(x) \text{ for } x \in \mathbb{B}, \text{ for some distribution } p \text{ on } \mathbb{B},$$

i.e., for all $\boldsymbol{x}$,

$$L(\boldsymbol{x}x) = L(\boldsymbol{x}) - \ln p(x) + c_x \text{ for } x \in \mathbb{B}, \text{ for some } c_x \in \mathbb{R}^{\geq 0}, p \text{ on } \mathbb{B},$$

are (by moving the constants into the scope of the logarithm) precisely those satisfying, for all $\boldsymbol{x}$,

$$L(\boldsymbol{x}x) = L(\boldsymbol{x}) - \ln p'(x) \text{ for } x \in \mathbb{B}, \text{ for some } \textit{semi}\text{-distribution } p' : x \mapsto r_x p(x).$$

Thus the superpredictions are the semi-distributions; and the functions L satisfying (b) are precisely the loss processes corresponding to the semi-measures ν. Furthermore, the condition (a) of upper semi-computability of L translates precisely in the lower semi-computability of the correspoding ν. The Π_1 superloss processes are therefore precisely those loss processess corresponding to the Σ_1 measures. What about the universal superloss processes given by the mix-loss? The log-loss game is mixable for $\eta = 1$, so the mix-loss is the loss corresponding to the standard Bayesian mixture strategy; hence the mix-loss with semi-computable prior over all Π_1 superloss processes corresponds to the Bayesian mixture predictor with semi-computable prior over all Σ_1 measures. That is, these universal Π_1 superloss processes are precisely the loss processes corresponding to the Solomonoff-Levin predictors. Wherefore the log-loss game's predictive complexity of a sequence is defined as the log-loss incurred on it by the Solomonoff-Levin predictor.

6.2.1.5. *Examples.* Let me give some examples of predictive complexity on infinite sequences in the log-loss game. Intuitively, the easiest to predict is a computable 'deterministic' or single infinite sequence. Indeed, since there is a Δ_1 predictor (the one corresponding to the deterministic measure generating the sequence) that predicts it perfectly, incurring loss 0, the Solomonoff-Levin predictor will incur not more than a single constant amount of loss and the predictive complexity is of order $O(1)$. This means that the *mean* loss per outcome goes to 0. What about a Martin-Löf random sequence—a sequence that is intuitively very hard to predict? Since by definition for such a sequence $\boldsymbol{x}^\omega$

$$-\log Q_U(\boldsymbol{x}^t) =^+ -\log \lambda(\boldsymbol{x}^t) =^+ t,$$

the predictive complexity of such a sequence is of order $t + O(1)$. In other words, if the data is generated by the i.i.d. measure μ_θ with $\theta = \frac{1}{2}$, then almost surely the predictive complexity of the data stream is of order $t + O(1)$, and the mean loss per outcome goes to 1. Intuitively, moreover, the data should be easier to predict if it is generated from an i.i.d. measure with a greater bias. At the extreme points 0 and 1 the data is indeed again deterministic and the predictive complexity of order $O(1)$; but in general a data stream generated from μ_θ will almost surely be a μ_θ-ML random sequence with a limiting relative frequency of $1-\theta$ 0's and θ 1's wherefore

$$-\log Q_U(\boldsymbol{x}^t) =^+ -\log \mu_\theta(\boldsymbol{x}^t) =^+ -\log\left((1-\theta)^{(1-\theta)t}\theta^{\theta t}\right) = tH(p_\theta),$$

with H the Shannon entropy that indeed has its maximum at $\theta = \frac{1}{2}$. So, for instance, with $\theta = \frac{1}{4}$ we have $H(p_\theta) \approx 0.8$ hence the predictive complexity is almost surely of order $0.8t + O(1)$.

6.2.2. Discussion. Vovk (2001b, 66, notation mine) writes:

> The intuition behind the universal [Π_1 superloss process] L is that $L(\boldsymbol{x})$ is an intrinsic measure of difficulty of a sequence $\boldsymbol{x}$ ...

Can the definition of predictive complexity actually support this intuition?

6.2.2.1. *From the intuition to the definition.* A natural quantitative measure of the difficulty of a sequence, *to a particular predictor* p, is the predictor's cumulative loss on this sequence. (This measure is, of course, also relative to the loss function of choice; and so is the definition of predictive complexity.) To turn it into an *intrinsic* measure of a sequence's difficulty (per a given loss function), we have to somehow lift it beyond its formulation relative to a particular predictor, to be relative to *any* predictor or the universal pool of *all* predictors. Again, this class is initially equated with the pool of *computable* predictors. At this point we might, naively, attempt the following definition: the predictive complexity of a sequence is the minimal loss that any (computable) predictor incurs on it. But this leads to a trivial notion: for every finite sequence there is a computable predictor that predicts it perfectly and so incurs no loss at all. In order to avoid such trivialization, we can do the following: we still let the measure be relative to a particular predictor, but we choose a predictor that in a way *represents* the universal pool of all predictors. Namely, a *universal* element in the universal pool of all predictors never does much worse than *any* other predictor, while it is not *too* good because it is still a legitimate predictor itself.

6.2.2.2. *From the intuition to the definition: the final step.* This is the point where we realize that the pool of predictors corresponding to the Δ_1 measures does not contain universal elements; and we are obliged to expand to a nondiagonalizable class of Σ_1 elements. There are at least two distinct ways, however, in which we can make this step. The familiar way is to identify the universal pool of possible predictors with those corresponding to the Σ_1 measures, and to choose a universal element to represent the pool. This interpretation turns out to be not just questionable for familiar reasons (6.2.2.3 below), but in fact problematic for novel reasons (6.2.2.6). A different interpretation is suggested by the actual route that Vovk takes towards the definition of predictive complexity. As we saw in 6.2.1 above, the analysis proceeds directly at the level of loss processes instead of prediction methods: we thus expand the class of Δ_1 loss processes to the nondiagonalizable class of Π_1 superloss processes. We might, rather pragmatically, try to present the prediction methods corresponding to the Π_1 superloss processes as the universal pool of all predictors; alternatively, we move right ahead to the fully pragmatic view that the expansion to Σ_1 elements is simply a device to obtain an approximation to the losses of the original Δ_1 predictors (6.2.2.7 below). Resuming the passage I started this section with (ibid.),

> The intuition behind the universal [superloss process] L is that $L(\boldsymbol{x})$ is an intrinsic measure of difficulty of a sequence $\boldsymbol{x}$: the loss of no computable prediction strategy is much less than $L(\boldsymbol{x})$, but the latter can be obtained "in the limit".

That concludes my reconstruction of the route from the intuition to the definition. Let me now discuss the main moves in some more detail, starting from the basic case of the log-loss game.

6.2.2.3. *The Solomonoff-Levin definition as a measure of complexity.* Recall the discussion in 5.2 about the Solomonoff-Levin predictor as implementing a measure of complexity of data sequences. I conluded this discussion in 5.2.2.8 saying that while its interpretation as a general and objective measure of simplicity-as-compressibility fails to be convincing, there always remains one sense in which a Solomonoff-Levin predictor gives an expression of a sequence's complexity: simply as the difficulty it has predicting it. This is an instance of the first step in 6.2.2.1 above: a quantification of the complexity of a sequence, to a particular predictor, is the predictor's loss on the sequence. But of course any such particular complexity measure can only make a claim to generality and objectivity insofar the corresponding predictor does. That means that, if we seek to present the predictive complexity (in the log-loss game) as the difficulty to (i.e., the log-loss incurred by) a particular Solomonoff-Levin predictor, it inherits the issues that stand in the way of a particular Solomonoff-Levin predictor p_{Q_U}'s claim to generality and objectivity, the familiar issues of variance and the choice of universal pool of predictors.

6.2.2.4. *Variance.* The issue of variance is that there is still a choice of particular Solomonoff-Levin predictor (particular universal machine) to define predictive complexity with (see 5.2.2 above); in general, there is still the choice of particular universal superloss process. This is certainly a problem for the view, suggested in the passage taken from Vovk, of any particular Solomonoff-Levin predictor's log-loss function (in general, any particular universal superloss process) as quantifiying a given finite sequence's intrinsic complexity. Nevertheless, by the invariance theorem every two choices only differ by an additive constant; and we can still retain the view of predictive complexity as at least measuring the growth or the order of the complexity of a data stream (5.2.2.4 above). We can distinguish between, for instance, data streams of order of predictive complexity $O(1)$ and $t+O(1)$, or of $t+O(1)$ and $\frac{t}{2}+O(1)$—also see the examples in 6.2.1.5 above. Thus, we alleviate the problem of variance by taking a view of predictive complexity as tracking the order of complexity of data streams or infinite sequences. There are, however, more serious problems to come.

6.2.2.5. *The pool of predictors.* The issue of the choice of pool of predictors in the log-loss game I discussed in 4.3 above. The interpretation of the Solomonoff-Levin predictor as a universal prediction method is barred by the fact that the class of Σ_1 measures does not correspond to a natural universal pool of predictors or one-step *conditional* measures.

6.2.2.6. *Superloss processes and predictors.* While, therefore, the correspondence between superloss processes and predictors in the case of the log-loss is already less than satisfying, in the case of other games this correspondence does not even seem to persist. In the log-loss case we have the nice equality $L_\nu = -\log \nu$ due to sequential locality, so that $L_\nu \in \Pi_1$ precisely if $\nu \in \Sigma_1$. But for any proper loss function that is not sequentially local (i.e., any proper loss function that is not of logarithmic shape, 6.1.2.4 above) the cumulative loss L_ν depends on the individual instantenous losses ℓ_ν, that is, the individual conditional probabilities $\nu(\cdot \mid \cdot)$. Since these may *not* be Σ_1—as exemplified by the Solomonoff-Levin measures, proposition 4.1—the instantaneous losses may not be Π_1, in which case it seems unlikely in general that their sum, the cumulative loss, is Π_1 again. (To repeat, in the case of the log-loss the sum on non-Π_1 instantaneous losses *is* Π_1 again, but this follows from the special property of the log-loss that this sum reduces to a function of the unconditional measure.) So a Σ_1 measure may not correspond to a Π_1 superloss process. (Although I think this *likely*, I found it suprisingly tricky to actually *prove* it and I have to admit that I have been unable to. I give some more details in B.2.5.) Conversely, it seems that a Π_1 cumulative loss (a Σ_1 sum of predictions) is compatible with a product of predictions that is *not* Σ_1; which gives Π_1 superloss processes that do not correspond to Σ_1 measures. (See again B.2.5.) Moreover, it is certainly the case, by definition so, that for any game that is not 1-mixable the *universal* Π_1 superloss processes do not correspond to the universal Σ_1 measures. Note that this further implies that the pool of (universal) predictors that corresponds to the (universal) Π_1 superloss processes can differ for different games. In short, there looms a serious mismatch between effective prediction methods and effective superloss processes, which is probably why Vovk moves away from the former and puts things directly in terms of the latter.

6.2.2.7. *Approximation of computable losses.* The problem with setting things up purely in terms of superloss processes is that it becomes unclear how exactly we must make sense of predictive complexity as a measure of *predictive complexity*, a measure of the difficulty presented to prediction methods. For that, we really do need a connection to a pool of prediction methods. (This is certainly not solved by simply branding the superloss processes themselves "superstrategies," Kalnishkan, 2015, 118, or "strategies in a generalized sense," see Vovk, 2001a, 237.) But, again, the pool of prediction methods corresponding to the Π_1 superloss processes is not a very natural one. The more appealing alternative is to forget about the interpretation of the superloss processes altogether, and simply view the predictive complexity as an approximation to the losses of the original pool of Δ_1 predictors. By design this certainly works one way: the loss of no computable method is much less than that of the universal Π_1 superloss process. But we also need the universal Π_1 superloss process not to be too far removed from the computable methods from the other direction. (The loss of no computable method is less than 'always zero,' either.) Both directions are addressed in the passage I took from Vovk, where he says that

> . . . the loss of no computable prediction strategy is much less than $L(\boldsymbol{x})$, but the latter can be obtained "in the limit".

He continues:

> The universal [Π_1 superloss process] when applied to $\boldsymbol{x}^t \in \mathbb{B}^*$ eventually learns all regularities in $\boldsymbol{x}^t$ relevant to the on-line prediction of x_1, then $x_2, \dots$, finally x_t, but this process of learning never ends: the upper semicomputability (but not computability) of L means that there always remains [the] possibility of discovering new regularities in $\boldsymbol{x}^t$ which will decrease L's estimate of the loss attainable on $\boldsymbol{x}^t$.

But just the observation that computability "in the limit" is in some sense close to computability is not enough: what we really need is that the universal Π_1 superloss process is in some more precise sense *not too good.* In the above words, might the universal Π_1 superloss process not discover *too many* regularities, regularities that no computable method can discover? As a matter of fact, it does. There exist infinite sequences such that the log-loss of any computable predictor is of order $O(t)$, while the universal Π_1 superloss is of order $O(\log t)$.

Proposition 6.2. There are sequences $\boldsymbol{x}^\omega \in \mathbb{B}^\omega$ such that

$$L_\mu(\boldsymbol{x}^t) \geq^+ t$$

for every $\mu \in \Delta_1$, while

$$L_{Q_U}(\boldsymbol{x}^t) \leq^+ 3 \log t$$

for any Solomonoff-Levin measure Q_U.

Proof. This follows from the existence of sequences that are *computably random* yet *ultracompressible* (Lahtrop and Lutz, 1999). See B.2.6. □

That means that, for the purpose of an approximation to the Δ_1 predictors' losses, the universal Π_1 superloss process is really too good.

6.2.2.8. *Predictive complexity and descriptive complexity.* Universal *descriptive* complexity, or Kolmogorov complexity via a universal machine U, gives a natural notion of minimum description length because, as I emphasized before in 5.2.1.8, the class of description systems that have a decoding algorithm (with the Church-Turing thesis, a decoder given by a Turing machine) is a natural class of description systems. The universal class of effective decoders is (via the Church-Turing thesis) precisely the class of Turing machines, and the universal elements among the latter, the universal Turing machines, give universal effective decoders that in turn give a notion of universal minimum description length. Importantly, this is a natural notion in spite of the fact that the function *itself*, the Kolmogorov complexity via universal U *itself*, is not computable, but only semi-computable or Σ_1. This is an aspect in which descriptive complexity is markedly different from *predictive* complexity. Namely, here the relevant class is the class of effective predictors (or effective loss processes) themselves, and such effective prediction methods, methods that come with an algorithm to calculate their actual predictions, must really be identified (via the Church-Turing thesis) with the total computable or Δ_1 predictors. We saw that a method that only corresponds to a Σ_1 measure strains a natural conception of effective prediction method (4.3.2 above); and the situation is worse still for the fact that the predictor itself need not even be Σ_1 (4.3.3 above). (Again, for the log-loss, the superloss processes are still Σ_1; but this follows from the log-loss function's sequential locality and might not continue to hold for other loss functions, 6.2.2.6 above.) Thus, while it is consistently highlighted in papers on the subject that there is a formal coincidence between *descriptive* complexity (monotone Kolmogorov complexity) and *predictive* complexity (at least for the log-loss game), on an interpretational level there is an important difference between the two, to the detriment of the latter.

6.2.2.9. *Conclusion.* Vovk's notion of predictive complexity of data sequences must give an expression of the difficulty according to a particular pool of predictors. In order for it to be an expression of 'intrinsic' difficulty, this pool of predictors needs to be interpreted as the universal pool of all possible predictors. But then it seems unable to escape the following dilemma. Either it is to express the difficulty according to the particular pool corresponding to the Π_1 superloss processes, but this pool fails to be a natural one and can hardly be interpreted as required. Or it is to express the predictive difficulty according to the pool of Δ_1 prediction methods, which *is* a natural choice, but then the values it gives are too low.

* * *

6.3. Further topics

This section collects a number of topics that arose from the work in this thesis, but that I have not been able to develop in sufficient depth to report here in completed form. As such, this list of topics presents suggestions for—if not promises of—future research.

6.3.1. A weaker predictive complexity. As mentioned in 6.1.4.5 above, we can define predictive complexity for the mixable games. That is, for a game such that for some η the constant $c(\eta)$ as in (65) equals 1, there are universal Π_1 superloss processes that are invariant up to an additive constant. However, as mentioned in 6.1.4.6 above, not all games are mixable: the absolute-loss game is not, and for this game we cannot define a predictive complexity (Kalnishkan and Vyugin, 2002a). It is indeed the case that *only* the mixable games have a predictive complexity (Kalnishkan et al., 2004).

Nevertheless, some games that are not mixable still have the property that $c(\eta) \to 1$ as $\eta \to 0$: this again includes the absolute-loss game (Kalnishkan and Vyugin, 2002b). For such games, we can still define a *weak* predictive complexity: a predictive complexity that is not invariant up to $O(1)$, but up to $O(\sqrt{t})$.

Kalnishkan and Vyugin (ibid., 108) write that the definition of "*predictive complexity up to* $f(n) \ldots$ makes sense if $f(n) = o(n)$ as $n \to \infty$," i.e., if $f(n)/n$ goes to 0 as n grows. In other words, if the *mean* predictive complexity per outcome is asymptotically invariant.

However, a more loose invariance means that the definition is less well-equipped to register differences in the growth or order of complexity (6.2.2.4 above). As a simple example, compare the trivial computable sequence 0^ω with the sequence $\boldsymbol{x}^\omega$ that is defined with the help of a Martin-Löf random sequence $\boldsymbol{y}^\omega$ by

$$\boldsymbol{x}^\omega(t) = \begin{cases} \boldsymbol{y}^\omega(\sqrt{t}) & \text{if} \sqrt{t} \in \mathbb{N}; \\ 0 & \text{otherwise.} \end{cases}$$

In words, $\boldsymbol{x}^\omega$ is the sequence of all 0's interpersed with one next bit from a Martin-Löf random sequence at each position that is a square number—a sequence that is intuitively quite unpredictable! A predictive complexity that is invariant up to $O(1)$ can indeed distinguish 0^ω of complexity order $O(1)$ from $\boldsymbol{x}^\omega$ of complexity order $O(1) + \sqrt{t}$, but a predictive complexity that is invariant up to $O(\sqrt{t})$ is too coarse-grained to register the difference between the two.

Given that predictive complexity with invariance up to $O(1)$ is impossible for non-mixable games like the absolute-loss game, we cannot improve on this definition in full generality; but we might still be able to strengthen it in special cases. Namely, we might be able to define a notion of predictive complexity that has a stronger invariance in the case of *easy data*. We might be able to define a notion that is invariant up to $O(\sqrt{t})$ in general, but more fine-grained in the natural cases where it is possible to predict succesfully; for instance, when the data is in fact generated by an i.i.d. source with a strong bias, or is a deterministic sequence with a clear structure.

An interesting case of easy data is when there is at least one prediction method (superloss process) that does very well or indeed near perfect on the data. This case was suggested in this context by Wouter Koolen: and he described a definition of a Π_1 superloss process, as a mixture with changing learning rate (a version of the *Squint* algorithm of Koolen and Van Erven, 2015), that is invariant up to order $\min_k \sqrt{2L_k(\boldsymbol{x}^t)(-\ln w(k) + \ln\ln t)} + O(1)$, hence invariant up to $O(\sqrt{t})$ in general but more fine-grained whenever there is a superloss process that does well.

6.3.2. Prediction and randomness. In A.4, I explain the notion of Martin-Löf randomness and its predictive characterization. In this characterization, a sequence $\boldsymbol{x}^\omega$ is ML-random relative to computable μ (or μ-ML-random) precisely if $Q_U(\boldsymbol{x}^t) =^\times \mu(\boldsymbol{x}^t)$, or equivalently, $L_{Q_U}(\boldsymbol{x}^t) =^+ L_\mu(\boldsymbol{x}^t)$ with L the cumulative log-loss.

This is a direct characterization of ML-randomness in terms of Vovk's predictive complexity for the log-loss game (given by the Π_1 superloss process L_{Q_U}), which invites a straightforward generalization to randomness for other games, like the Brier game, that is similarly put in terms of the game's predictive complexity (Vovk, 2001b). However, the interpretation of this notion of randomness is susceptible to the same critique, set out in 6.2 above, that applies to the notion of predictive complexity, and that is rooted in the unconvincing link between the class of Π_1 superloss processes and a natural class of prediction strategies.

Perhaps a more natural road is to explicitly define a notion of randomness relative to a given pool of prediction methods, in accordance with the above characterization. Thus, given a pool of $\mathcal{H}$ of (measures corresponding to) prediction methods, and $\mu \in \mathcal{H}$, we say that a sequence $\boldsymbol{x}^\omega$ is μ-$\mathcal{H}$-random if for all $\mu' \in \mathcal{H}$ it holds that

$$\mu'(\boldsymbol{x}^t) \leq^\times \mu(\boldsymbol{x}^t), \tag{66}$$

or equivalently (with L the cumulative log-loss),

$$L'_\mu(\boldsymbol{x}^t) \geq^+ L_\mu(\boldsymbol{x}^t). \tag{67}$$

In words, a sequence is μ-random relative to a pool of (measures corresponding to) prediction methods, if no prediction method in the pool is more successful on this sequence than (the predictor corresponding to) μ is.

From this perspective, Martin-Löf randomness is not a terribly natural notion, because, as argued extensively in this thesis, the corresponding pool $\mathcal{M}$ of (prediction methods corresponding to) Σ_1 measures is not a terribly natural class. Things are actually worse for the fact that the original test characterization breaks down for randomness relative to elements of $\mathcal{M}$ that are not Δ_1 measures (A.4.2.7).

Superficially, this analysis connects to the occasional discussion in the field of algorithmic randomness on the 'right' notion of randomness within the "randomness zoo" of existing notions (a recent example is Porter, 2016). While Martin-Löf randomness is often presented as the most convincing notion (e.g., Dasgupta, 2011), it also has properties that strain intuition. For one thing, Chaitin is (in)famous for making the claim that the halting probability $\Omega = \sum_{\boldsymbol{x}:T(\boldsymbol{x})\downarrow} \lambda(\boldsymbol{x})$ of a universal prefix-free machine, which gives an infinite sequence that is λ-ML-random (and that encodes the solutions to Diophantine equations), illustrates how number theory is beset by randomness; but a more sober observation is that a sequence that still has so much structure, that is indeed left-c.e., hardly qualifies as random (cf. Van Lambalgen, 1989).

In fact, the Martin-Löf random sequences still have so much structure that *every* sequence is computable from some ML-random sequence (Gács, 1986). In response to this, some have argued for stronger (more restrictive) notions of randomness; in particular (Osherson and Weinstein, 2008), for the notion of *weak 2-randomness* (Kurtz, 1981; Gaifman and Snir, 1982; see Downey and Hirschfeldt, 2010, 287f). Others have actually argued for *weaker* notions of randomness: notably, Schnorr (1971a) felt that the ML-test concept is at too high a level of effectiveness (also see A.4.1.3) and proposed stricter ones.

From our earlier predictive perspective, it is in fact natural to advance a weaker notion of randomness. Namely, the pool of *computable* prediction methods is a natural pool of strategies, thus leading to a natural notion of a sequence being μ-random if there is no strategy that beats μ on this sequence. This notion is known as *computable randomness* in the literature: the predictive (or martingale) characterization goes back to Schnorr (1971a; 1971b), and it also allows for characterizations in terms of tests and complexity (Downey et al., 2004; Merkle et al., 2006; see Downey and Hirschfeldt, 2010, 279ff). It is often taken as a downside of this and other notions that, unlike for Martin-Löf's definition, there are no *universal* tests (or universal superloss processes, or universal prediction methods, . . .) available (cf. ibid., 275f). However, while it is mathematically convenient to have a characterization in terms of a single universal element (and while this *was* important in obtaining a complexity measure), this is not at all necessary for a notion of randomness along the lines set out here.

A further question is whether other important randomness notions also admit of a natural predictive (martingale) interpretation; in particular, what pools of prediction methods these would specify. Interestingly, for instance, it is still an open question what kind of martingale characterization could be given, if one can be given at all, for the notion of weak 2-randomness.[26]

6.3.3. Prediction with expert advice and meta-induction. Schurz (2008) proposes a *meta-inductive justification of induction* in the setting of sequential prediction, based on results from prediction with expert advice.

The naive reply to the problem of induction is that induction is justified because it has been successful in the past. Naive, because this suggested justification is either circular (it requires induction itself), or, on a more refined perspective, must rest on a higher principle of induction (the justification of *object*-induction on the level of data requires *meta*-induction on the level of methods), which itself begs for justification and so sets in motion an infinite regress (recall I.1; also see Skyrms, 2000). Now Schurz's idea is essentially that this regress may be halted directly at the second level: it is possible to give a purely *analytic* justification of meta-induction. If we tie in this analytic justification with the *empirical* observation that object-induction has in fact been successful in the past, then meta-induction instructs us to proceed object-inductively; and this, Schurz argues, comes down to an a posteriori justification of object-induction.

The crucial analytic justification of meta-induction is grounded in results in prediction with expert advice on the optimality of aggregating predictors. In particular, Schurz bases much of his analysis on the weighted average forecaster described in Cesa-Bianchi and Lugosi (2006, 12ff), that weighs a finite pool of predictors directly based on their past performance, and that can be shown, for all convex loss functions, to incur a mean loss converging to that of the best predictor in the pool. The weighted average forecaster, or meta-inductive method, follows those predictors that have been successful in the past; and this strategy is provably optimal in the above sense.

By clearly separating object-induction on the level of the observation data and meta-induction on the level of prediction strategies, Schurz manages to give a precise form to Reichenbach's fundamental observation that induction picks up the past success of alternative methods. Does Schurz's proposal succeed in giving a justification of induction?

Of course, there is the basic concern that the framework of sequential prediction falls short of capturing the essence of inductive inference (I.1 above), and so a justification within this framework falls short of the real goal: to justify inductive or scientific reasoning (cf. Arnold, 2010, 591). Another possible concern is that Schurz assumes that there is something like *the* object-inductive method, thus simply side-stepping Goodman's new riddle (2008, 279). However, as also noted in I.1, the problem of induction derives its bite from there being something we can identify as scientific or inductive reasoning; and therefore the right approach here would be to consider things from a very general perspective, where we distinguish 'the scientific method' from a number of other 'methods' like accepting the predictions of politicians or consulting horoscopes. Then, since as a matter of empirical fact science has been the most successful of known methods up to this point in human history, we would be meta-inductively justified to keep following it: this is the proposed justification of induction. Note, though, that this does not give a full-blown justification of the scientific or object-inductive method: it might be the case that in the future the scientific method will become less successful than some other method, and then we will be meta-inductively pressed to use *this* method. Thus this would only be a justification for sticking with the scientific method *for now*.

Arnold (2010) argues that the analytic justification of meta-induction does not extend to aggregating predictors over *infinitely* many predictors, and that this is a problem for Schurz's proposal. As for the first point, in case of an infinite pool of predictors there are some differences: most importantly, an aggregating predictor must use weights that not only express the predictors' success but also some prior loading. This means that *uniform* convergence is no longer possible in general, but at least for the mixable loss functions an aggregating predictor is still optimal in the sense of always converging to not more than any given predictor' losses (the notion of optimality I studied in this thesis). As for the second point, what Arnold seems to drive at here is a universally optimal method, that is optimal relative to the infinite pool of all possible predictors. This is, of course, what I discussed in this thesis, and argued to be impossible. However, in Schurz's proposed justification, as I reconstructed it above, all that needs to be taken into account are the necessarily finitely many alternatives to the scientific method that have been suggested to us so far.

The latter does point at a different aspect that is of interest. Namely, there is the possibility of *new* alternative methods being suggested over the course of time, and this needs to be taken into account if Schurz's justification is to work. This asks for spelling out how exactly new methods are dynamically incorporated by the meta-inductive strategy, and whether and in what sense the analytic optimality is preserved. Certainly in case this can be made to work, it promises to be very instructive to relate this to the discussion in 4.1 above on the fixity of the methods and the structurally similar issue of Bayesian theory change.[27]

*

Part IV

Appendices

APPENDIX A

Supplementary material

This appendix gathers and explains concepts, mostly from algorithmic information theory, that are peripheral to the main story, but still used throughout the thesis.

In A.1, I present the Σ_1 semi-distributions as stemming from the restriction of monotone machines to prefix-free machines. In A.2, I present the notions of description and coding systems from information theory, with an eye to their role in algorithmic information theory. In A.3, I present notions of Kolmogorov complexity (particularly, the prefix-free and the monotone variant) as the universal description length functions for effective description systems. Here I also present a novel generalization of prefix-free Kolmogorov complexity, including a generalized coding theorem, theorem A.16. In A.4, I present the notion of Martin-Löf randomness, including a predictive interpretation.

A.1. The Σ_1 semi-distributions

A.1.1. Prefix-free machines. I will present prefix-free machines here as a variant of monotone machines (introduced in 2.1.2.4 above). The basic idea is to limit monotone machines in such a way that for each infinite input sequence at most one finite output sequence is produced. There are at least two ways to make this conception perfectly precise (Shen et al., 20xx, 91ff).[28]

A.1.1.1. *Prefix-stable machines.* Consider the c.e. sets M of pairs of sequences that satisfy the following strengthening of characterization (14) in 2.1.2.4 above:

$$\text{if } (\boldsymbol{x}_1, \boldsymbol{y}_1), (\boldsymbol{x}_2, \boldsymbol{y}_2) \in M \text{ and } \boldsymbol{x}_1 \preccurlyeq \boldsymbol{x}_2 \text{ then } \boldsymbol{y}_1 = \boldsymbol{y}_2. \tag{68}$$

The induced functions Φ_M that are defined as before by (15) are such that if $\Phi_M(\boldsymbol{x})$ is defined at all then $\Phi(\boldsymbol{x}') = \Phi(\boldsymbol{x})$ for all extensions $\boldsymbol{x}' \succcurlyeq \boldsymbol{x}$. The monotone machines that adhere to (68) are called the *prefix-stable* machines (Levin, 1974, 207; Gács, 1974, 1477; see Shen et al., 20xx, 91; Gács, 2016, 3).

A.1.1.2. *Prefix-free machines.* But notice that (68) implies that M is already the graph of a function itself. These functions are exactly the p.c. functions with prefix-free domain, which is the usual characterization of *prefix-free* or *self-delimiting* machines (Chaitin, 1975, 330f; see Shen et al., 20xx, 91; Li and Vitányi, 2008, 200f; Nies, 2009, 83; Downey and Hirschfeldt, 2010, 122).

Definition A.1 (Levin, Gács, Chaitin). A prefix-free machine is a p.c. function $T : \mathbb{B}^* \to \mathbb{B}^*$ with prefix-free domain.

We then call $\boldsymbol{x}$ a (prefix-free) *T-description* of $\boldsymbol{y}$ precisely if $T(\boldsymbol{x}) = \boldsymbol{y}$.

A.1.1.3. *Universal prefix-free machines.* A *universal* prefix-free machine is given by

$$U(\boldsymbol{z}_e\boldsymbol{x}) = \boldsymbol{y} :\Leftrightarrow T_e(\boldsymbol{x}) = \boldsymbol{y}$$

for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{B}^*$ and some computable prefix-free list $\{\boldsymbol{z}_e\}_e$ of sequences that serves as an encoding of an enumeration $\{T_e\}_{e\in\mathbb{N}}$ of all prefix-free machines.

A.1.2. The Σ_1 semi-distributions. A transformation of λ by a prefix-*stable* machine (a type of monotone machine) is, of course, a Σ_1 measure on $\mathbb{B}^\omega \cup \mathbb{B}^*$ (see 2.1.2 above). But now consider just the *minimal* descriptions of M, those $\boldsymbol{x}$ with $(\boldsymbol{x}, \boldsymbol{y}) \in M$ for some $\boldsymbol{y}$ but not $(\boldsymbol{x}', \boldsymbol{y}) \in M$ for any prefix $\boldsymbol{x}' \prec \boldsymbol{x}$. The set of all minimal $\boldsymbol{x}$ is prefix-free, by (68). This means, by the Kraft inquality (see A.2.2.1), that the sum of the uniform probabilities of these descriptions is bounded by 1. That in turn means that the uniform probabilities of the descriptions induce a probability (semi-)*distribution* on $\mathbb{B}^*$ (with positive probability for those sequences that actually have a description). It is in this spirit that we define a transformation on prefix-*free* machines.

A.1.2.1. *Transformations.* Analogous to definition 2.3 above for monotone machines, we define the transformation of λ by prefix-free machine T by

$$\begin{aligned}\lambda_T(\boldsymbol{y}) &:= \lambda(\{\boldsymbol{x} : (\boldsymbol{x}, \boldsymbol{y}) \in T\}) \\ &= \sum_{\boldsymbol{x}:\ T(\boldsymbol{x})=\boldsymbol{y}} \lambda(\boldsymbol{x}).\end{aligned}$$

Like in 2.2.1.1 above, this definition can be generalized to any Δ_1 measure.

Definition A.2. The transformation μ_T of μ by prefix-free T is given by

$$\mu_T(\boldsymbol{y}) = \sum_{\boldsymbol{x}:\ T(\boldsymbol{x})=\boldsymbol{y}} \mu(\boldsymbol{x}).$$

A.1.2.2. *The Δ_1 distributions.* If the prefix-free set of T-descriptions is *complete*, meaning that the sum of their uniform (or in general, μ-) probabilities *equals* 1, then also $\sum_{\boldsymbol{y}\in\mathbb{B}^*} \mu(\boldsymbol{y}) = 1$. Thus μ_T is a *probability distribution* p over the finite sequences $\mathbb{B}^*$. Morever, since for $\mu \in \Delta_1$ we can computably approximate $\mu_T(\boldsymbol{y})$ to any desired accuracy (we can both computably approximate $\mu_T(\boldsymbol{y})$ and $1 - \mu_T(\boldsymbol{y}) = \sum_{\boldsymbol{y}'\in\mathbb{B}^*\setminus\{\boldsymbol{y}\}} \mu_T(\boldsymbol{y}')$ from below), μ_T is a computable or Δ_1 distribution over $\mathbb{B}^*$.

A.1.2.3. *The Σ_1 semi-distributions.* In general, a transformation μ_T is a *semi-distribution* over $\mathbb{B}^*$, meaning that $\sum_{\boldsymbol{y}\in\mathbb{B}^*} \mu_T(\boldsymbol{y}) \leq 1$. Moreover, if μ is Δ_1, then μ_T is *semi-computable* or Σ_1. Conversely, for any continuous Δ_1 measure μ, we have that every Σ_1 semi-distribution over $\mathbb{B}^*$ equals the transformation of μ by some prefix-free T. Let $\mathcal{Q}$ denote the class of all Σ_1 semi-distributions over $\mathbb{B}^*$. We thus have the following analogue to proposition 2.10 about the Σ_1 measures.

Proposition A.3. For every continuous Δ_1 measure μ,

$$\{\mu_T\}_T = \mathcal{Q},$$

where the T range over all prefix-free machines.

Proof. See B.1.2.

A.1.2.4. *Universal Σ_1 semi-distributions.* The notions of dominance and universality are also inherited from the Σ_1 measures. A semi-distribution q_1 on $\mathbb{B}^*$ *dominates* another semi-distribution q_2 if there is a constant c such that for every $\boldsymbol{x} \in \mathbb{B}^*$ it holds that

$$q_1(\boldsymbol{x}) \geq c^{-1} q_2(\boldsymbol{x}),$$

also written '$q_1 \geq^\times q_2$.' A *universal* Σ_1 semi-distribution dominates every Σ_1 semi-distribution:

Definition A.4. A universal Σ_1 semi-distribution $\mathring{q}$ is such that for every Σ_1 semi-distribution q we have

$$\mathring{q} \geq^\times q.$$

A.1.2.5. *Universal transformations: the Levin-Chaitin semi-distributions.* A universal Σ_1 semi-distribution is given by a *universal* transformation of λ by a universal prefix-free machine. I will refer to these universal transformations as the *Levin-Chaitin semi-distributions*, as the definition was first independently described by Levin (1974, 207) and Chaitin (1975, 332).

Definition A.5 (Levin, Chaitin). The Levin-Chaitin semi-distribution q_U via universal prefix-free machine is the universal transformation λ_U.

The class $\mathcal{LC}$ of Levin-Chaitin semi-distributions coincides with the class of universal transformations of any continuous Δ_1 measure μ.

Proposition A.6. For every continuous Δ_1 measure μ,

$$\{\mu_U\}_U = \mathcal{LC},$$

where the U range over all universal prefix-free machines.

Proof. This can be derived in a manner identical to the proof of theorem 2.13 about the Solomonoff-Levin measures: see B.1.5, B.1.6.

* * *

A.2. Description and coding systems

A.2.1. The basic definitions. I repeat the notions of description system and code from 5.2.1.1 above. (Terminology varies in the literature; I follow Grünwald, 2007, 79ff.) Let Ω be a countable set of *source elements.* A *description system* for Ω is a set $D \subseteq \Omega \times \mathbb{B}^*$ of pairs of source elements and their *description sequences* in $\mathbb{B}^*$, so that $D(a, \boldsymbol{x})$ means that $\boldsymbol{x}$ is a description of a. A *coding system* is a description system that is a function itself, meaning that each source sequence has a unique description or *code.*

A.2.2. Prefix description and coding systems. A description system is *prefix-free* or simply *prefix* if no description is an initial segment of another. We further stipulate that a prefix description system is not *lossy*: each description corresponds to a unique source sequence. Then each prefix description system comes with a *decoding function* D^{-1} that maps descriptions to their source sequences.

A.2.2.1. *The Kraft inequality.* The basic fact underlying the link between description lengths and probabilities is the Kraft inequality (1949; McMillan, 1956).

Proposition A.7 (Kraft inequality). For every prefix-free set $A \subseteq \mathbb{B}^*$ it holds that $\sum_{\boldsymbol{x} \in A} 2^{-|\boldsymbol{x}|} \leq 1$. Conversely, for every (possibly infinite) sequence $l_1, l_2, \dots$ of lengths in $\mathbb{N}$ with $\sum_i 2^{-l_i} \leq 1$, there is a prefix-free set $A = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots\}$ with $|\boldsymbol{x}_i| = l_i$ for all i.

A.2.2.2. *Request sets.* A way of giving meaning to the converse direction is the following. (See Nies, 2009, 86ff.) Let a *request set* $\{(a_i, l_i)\}_i$ be a list of pairs of source elements and integer lengths. Then for any request set that we specify, as long as $\sum_i 2^{-l_i} \leq 1$ (we say that the request set is *bounded*), we are guaranteed the existence of a prefix description system D with descriptions $\boldsymbol{x}_i$ such that $D^{-1}(\boldsymbol{x}_i) = a_i$ and $|\boldsymbol{x}_i| = l_i$. If, in addition, $a_i \neq a_j$ for $i \neq j$, there exists a prefix coding system C with descriptions $\boldsymbol{x}_i$ such that $C(a_i) = \boldsymbol{x}_i$ and $|\boldsymbol{x}_i| = l_i$ for all i.

A.2.2.3. *Prefix code length functions and semi-distributions.* Define $L_C : a \mapsto |C(a)|$ to be the *code length function* for coding system C. The first part of the Kraft inequality tells us that for prefix coding system C we have that

$$P_C(\cdot) := 2^{-L_C(\cdot)} \tag{69}$$

satisfies $\sum_{a \in \Omega} P_C(a) \leq 1$, i.e., is a semi-distribution on Ω. Conversely, for any semi-distribution P, the set of requests $(a, \lceil -\log P(a) \rceil)$ for all $a \in \Omega$ is bounded, hence there is a prefix coding system C that satisfies $L_C(\cdot) = \lceil -\log P(\cdot) \rceil$. Writing '$L_P$' for the unique code length function of any such C, we can rephrase: for every distribution P we have a prefix code length function L_P with

$$L_P(\cdot) = \lceil -\log P(\cdot) \rceil. \tag{70}$$

(Also see Cover and Thomas, 2006, 107ff; Grünwald, 2007, 91ff.)

A.2.2.4. *Example.* Consider the set of source elemens $\Omega = \{a_1, a_2, a_3, a_4\}$. A coding system for Ω is $C = \{(a_1, 0), (a_2, 11), (a_3, 100), (a_4, 101)\}$. The corresponding code length function L_C assigns to (a_1, a_2, a_3, a_4) the values $(1, 2, 3, 3)$; and the probabilities via (69) are $(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8})$. Conversely, a semi-distribution P on Ω is given by $(\frac{1}{3}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6})$. This results via (70) in the code length function L_P that gives lengths $(2, 3, 3, 3)$, instantiated by (for instance) the codes $(00, 010, 011, 100)$.

A.2.2.5. *Complete prefix code length functions and distributions.* The last example started with a strict semi-distribution (i.e., the probabilities did not sum to 1), which gave a code length function that is *incomplete.* That is, there are coding systems C' that are more *efficient*: $L_{C'}(a) \leq L_C(a)$ for all source elements $a \in \Omega$, with strict inequality for at least one element (for example, the last code 100 could safely be replaced by the code 1 to give a more efficient C'). In general, proper probability distributions (like the first example above) correspond to prefix coding systems C that are complete—at least in the *idealized* case.

A.2.2.6. *Idealized prefix codes.* The correspondence between semi-distributions and prefix code length functions is not 1-1 because code lengths are integers, which calls for the rounding-off in (70). In general, however, it is an unneccessary inconvenience to worry about the minor influence of this correction, and indeed the approach taken in MDL is to simply assert that there is a 1-1 correspondence between semi-distributions and *idealized* code length functions that return real-valued code lengths (Grünwald, 2007, 95f). That is, every semi-distribution P corresponds uniquely to an idealized code length function L_P given by

$$L_P(\cdot) = -\log P(\cdot). \tag{71}$$

A.2.2.7. *Prefix description systems and semi-distributions.* If we permit source sequences to have multiple descriptions, there are at least two ways we can define a *description length function.* We can take the length of the *shortest* description,

$$L_D(a) := \min\{|\boldsymbol{x}| : (a, \boldsymbol{x}) \in D\}. \tag{72}$$

In that case, L_D is just the description length function of the code obtained from D by deleting the pairs that do not change (72), and we have the correspondence to semi-distributions as in A.2.2.3 above. Alternatively, we can take into account *all* descriptions and define L_D by

$$L_D(a) := -\log \sum_{\boldsymbol{x}\colon\ (a,\boldsymbol{x}) \in D} 2^{-|\boldsymbol{x}|}, \tag{73}$$

which is motivated by the following correspondence to semi-distributions. The first part of the Kraft inequality tells us that for prefix description system D

we have that

$$\begin{aligned} P_D(\cdot) :&= \sum_{\boldsymbol{x}:\ (\cdot,\boldsymbol{x})\in D} 2^{-|\boldsymbol{x}|} \\ &= 2^{-L_D(\cdot)} \end{aligned}$$

is again a semi-distribution on Ω. Conversely, for any semi-distribution P, there exists a bounded request set $R = \{(a_i, l_i)\}$ with $\sum_{l:(a,l)\in R} 2^{-l} = P(a)$ for each a (every real can be expressed as a sum of binary powers—viz. the binary expansion), hence there is a prefix-free description length function L_P that satisfies

$$L_P(\cdot) = -\log P(\cdot). \tag{74}$$

Note that this gives a 1-1 correspondence between semi-distributions and description length functions, in contrast to (70).

A.2.2.8. *The KC theorem.* In the following, I will take the set Ω of source elements to be the set $\mathbb{B}^*$ of finite sequences, too. Let an *effective* prefix description system D be such that the decoding function D^{-1} is given by a prefix-free machine (see A.1.1.1 above). Effective prefix description systems and *computably enumerable* bounded request sets are linked by the following effective version of the Kraft inequality, that I shall call the *KC theorem*, following Downey and Hirschfeldt (2010, 125) who write

> This result is usually known as the *Kraft-Chaitin Theorem*, as it appears in [Chaitin, 1975, 333f], but it appeared earlier in [Levin, 1974] ... There is also a version of it in [Schnorr, 1973, 380]. In [Chaitin, 1975], where the first proof explicitly done for prefix-free complexity seems to appear, the key idea of that proof is attributed to Nick Pippinger. Thus perhaps we should refer to the theorem by the rather unwieldy name of Kraft-Levin-Schnorr-Pippinger-Chaitin Theorem. Instead, we will refer to it as the KC Theorem. Since it is an effectivization of Kraft's inequality, one should feel free if one wishes to regard the initials as coming from "Kraft's inequality (Computable version)".

THEOREM A.8 (KC theorem). *For every prefix-free machine T, the set $\{(T(\boldsymbol{x}), |\boldsymbol{x}|) : T(\boldsymbol{x}) \downarrow\}$ is a c.e. bounded request set. Conversely, for every c.e. bounded request set $\{(\boldsymbol{y}_i, l_i)\}_i$, we can (effectively) construct a prefix-free machine T with domain $\{\boldsymbol{x}_i\}_i$ such that $T(\boldsymbol{x}_i) = \boldsymbol{y}_i$ and $\boldsymbol{x}_i = l_i$.*

PROOF. See Downey and Hirschfeldt (2010, 125f); Nies (2009, 88f). Alternatively, see the proof in B.1.2 below of proposition A.1.2.1 above.

A.2.2.9. *Effective prefix description systems and Σ_1 semi-distributions.* Given effective description system D, consider the length function L_D defined as in (73) above. (If we define L_D as in (72) we are led to the notion of Kolmogorov complexity, see A.3.1 below.) From the KC Theorem it follows that for our D we have that

$$\begin{aligned} P_D(\cdot) &= 2^{-L_D(\cdot)} \\ &= \sum_{\boldsymbol{x}:\ T(\boldsymbol{x})=\cdot} 2^{-|\boldsymbol{x}|} \end{aligned}$$

is a semi-distribution, that is Σ_1 because the set $\{(T(\boldsymbol{x}), |\boldsymbol{x}|) : T(\boldsymbol{x}) \downarrow\}$ is c.e. (we can effectively enumerate the lengths of descriptions that lead T to produce given $\boldsymbol{y}$). Conversely, for every Σ_1 semi-distribution P, we can effectively construct a bounded request set $R = \{(\boldsymbol{y}_i, l_i)\}$ with $\sum_{l:(\boldsymbol{y},l)\in R} 2^{-l} = P(\boldsymbol{y})$ for each $\boldsymbol{y}$, hence we can (effectively) construct an effective prefix-free description system D that satisfies $L_D(\cdot) = -\log P(\cdot)$. Thus every Σ_1 semi-distribution corresponds to a Σ_1 description length function L_P with

$$L_P(\cdot) = -\log P(\cdot). \tag{75}$$

A.2.2.10. *Transformations and Σ_1 semi-distributions.* Note that we can rewrite P_D above as

$$P_D(\cdot) = \sum_{\boldsymbol{x}:\ T(\boldsymbol{x})=\cdot} \lambda(\boldsymbol{x}).$$

Consequently, the correspondence between Σ_1 semi-distributions and effective prefix description systems is precisely the correspondence between Σ_1 semi-distributions and transformations of λ by prefix-free machines (A.1.2.1 above), i.e., an instance of proposition A.3 above.

A.2.3. The Shannon entropy. Let p be a probability distribution over a countable outcome space Ω (like $\mathbb{B}^*$, or $\mathbb{B}$). The *Shannon entropy* of a distribution p is the p-expected code length of the corresponding (idealized, see A.2.2.6 above) prefix code length function L_p.

Definition A.9. The Shannon entropy of distribution p is

$$\begin{aligned} H(p) &:= \mathbf{E}_{X\sim p}\left[-\log p(X)\right] \\ &= \mathbf{E}_{X\sim p}\left[L_p(X)\right]. \end{aligned}$$

A.2.3.1. *The information inequality.* The *information inequality* or *Gibb's inequality* (see MacKay, 2003, 34; Cover and Thomas, 2006, 28) says that the optimal p-expected code length is given by the idealized code length function L_p: any idealized L_q for q different from p will result in a greater p-expected code length.

Proposition A.10 (Information inequality). For all distributions p, q,

$$\begin{aligned} H(p) &\leq \mathbf{E}_{X\sim p}\left[-\log q(X)\right] \\ &= \mathbf{E}_{X\sim p}\left[L_q(X)\right], \end{aligned}$$

with equality iff $p = q$.

PROOF. We derive

$$\begin{aligned}\mathbf{E}_{X\sim p}\left[-\log p(X)\right] &= \mathbf{E}_{X\sim p}\left[-\log q(X)\right] + \mathbf{E}_{X\sim p}\left[\log \frac{q(x)}{p(x)}\right] \\ &\leq \mathbf{E}_{X\sim p}\left[-\log q(X)\right] + \log \mathbf{E}_{X\sim p}\left[\frac{q(x)}{p(x)}\right] \\ &= \mathbf{E}_{X\sim p}\left[-\log q(X)\right] + \log \sum_{x\in\Omega} q(x) \\ &= \mathbf{E}_{X\sim p}\left[-\log q(X)\right],\end{aligned}$$

where the inequality follows from Jensen's inequality, that says that for a convex function f, like $f(\cdot) = -\log\cdot$, it holds that $f(\mathbf{E}[\cdot]) \leq \mathbf{E}[f(\cdot)]$; hence for a concave function, like $f(\cdot) = \log\cdot$, the reverse inequality. □

Since *every* idealized prefix code length function for Ω corresponds to some distribution on Ω, the information inequality indeed states that

$$H(p) \leq \mathbf{E}\left[L_C(X)\right]$$

for *every* idealized coding system C, with, again, equality iff $L_C = L_p$.

A.2.3.2. *The Kullback-Leibler divergence.* The extent to which the p-expected code length given by L_q is worse than that given by L_p is expressed by the *relative entropy of p with respect to q* or the *Kullback-Leibler divergence from q to p.*

DEFINITION A.11. The Kullback-Leibler divergence from q to p is

$$\begin{aligned}D(p \parallel q) &:= \mathbf{E}_{X\sim p}\left[\frac{-\log q(X)}{-\log p(X)}\right] \\ &= \mathbf{E}_{X\sim p}\left[L_q(X) - L_p(X)\right]\end{aligned}$$

The information inequality, proposition A.10 above, says that the Kullback-Leibler divergence is nonnegative,

$$D(p \parallel q) \geq 0,$$

with equality iff $p = q$.

A.2.3.3. *The source coding theorem for symbol codes.* From all of the previous it quickly follows that, first, the Shannon entropy of p puts a lower bound on the p-expected codelength of any actual (i.e., nonidealized) coding system, and, second, that there always exists a coding system with p-expected codelength within one bit of p's entropy.

THEOREM A.12 (Source coding theorem for symbol codes, Shannon). *For every distribution p, for every prefix coding system C,*

$$\mathbf{E}_{X\sim p}\left[L_C(X)\right] \geq H(p),$$

and there exists a code C with

$$\mathbf{E}_{X\sim p}\left[L_C(X)\right] < H(p) + 1.$$

PROOF. The first inequality can be derived from the information inequality and the Kraft inequality, propositions A.10 and A.7 above (cf. MacKay, 2003, 97). The second inequality is fulfilled by the codelength function $L(\cdot) = \lceil -\log p(\cdot) \rceil$, which corresponds to a valid coding system by the Kraft inequality (A.2.2.3 above). □

A.2.3.4. *The source coding theorem and the AEP.* Shannon's source coding theorem for symbol codes must be distinguished from Shannon's *source coding theorem* simpliciter. This is rather a statement about lossy compression and equivalent to what is known as the *asymptotic equipartition property*; it is basically a variant of the weak law of large numbers. For details, see MacKay (2003, 74ff); Cover and Thomas (2006, 57ff).

A.2.3.5. *Competitive optimality and no-hypercompression.* The Shannon entropy also gives optimal codelengths in a competitive sense, as follows (see Cover and Thomas, 2006, 130ff). For distribution p on outcome space Ω and any prefix-code C on Ω, we have that the probability of an outcome x such that $L_C(x) \leq H(p) - c$ is no greater than 2^{-c}. This can be restated as what Grünwald (2007, 103) calls the *no-hypercompression inequality*, and what gives content to the assertion in 6.1.2.8 that it is hard to achieve low log-loss: for every measure μ^*, and second measure μ, on $\mathbb{B}^\omega$, the μ^*-probability of a finite sequence $\boldsymbol{x}$ such that $-\log \mu(\boldsymbol{x}) \leq -\log \mu^*(\boldsymbol{x}) - c$ is no greater than 2^{-c}. For instance, this directly implies that a sequence generated from the i.i.d. measure $\mu_{1/2}$ is only significantly compressible with vanishing probability. In fact, this is already quite easy to see from a purely combinatoric point of view: the fraction of data sequences of length t that can be mapped to a unique sequence that is c bits shorter decreases exponentially in c. The same combinatorial fact returns in the property that most sequences have a Kolmogorov complexity that is close to their length (Li and Vitányi, 2008, 116ff), and indeed in the property that the preditive complexity of most sequences is close to that given by the indifferent strategy (Kalnishkan et al., 2005).

A.2.4. Sequential description systems. In the same way that prefix-free machines are the decoding functions for a certain type of effective description system, that stands in direct correspondence to the Σ_1 semi-distributions, we can define a type of description system that the *monotone* machines are the decoding functions for, and that corresponds to the Σ_1 *measures.* In the following I will again take the class of source elements $\Omega = \mathbb{B}^*$.

A.2.4.1. *Definition.* A *sequential* description system D has the following properties. First, if $\boldsymbol{x}$ is a description of $\boldsymbol{y}$, so $(\boldsymbol{y}, \boldsymbol{x}) \in D$, then all its extensions are also descriptions of all initial segments of $\boldsymbol{y}$, so $(\boldsymbol{y}', \boldsymbol{x}') \in D$ for all $\boldsymbol{x}' \succcurlyeq \boldsymbol{x}$ and $\boldsymbol{y}' \preccurlyeq \boldsymbol{y}$. Second, if $(\boldsymbol{x}_1, \boldsymbol{y}_1), (\boldsymbol{x}_2, \boldsymbol{y}_2) \in D$ for two descriptions $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ that are compatible, then source sequences $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ must be compatible too. (Cf. Shen et al., 20xx, 140.) In order to define a description length function for such a sequential description system D, it is easier to work with the induced *minimal* description system $\hat{D}$, with $(\boldsymbol{y}, \boldsymbol{x}) \in \hat{D}$ if and only if $(\boldsymbol{y}, \boldsymbol{x}) \in D$ and not $(\boldsymbol{y}, \boldsymbol{x}') \in D$ for any shorter $\boldsymbol{x}' \prec \boldsymbol{x}$.

A.2.4.2. *Examples.* Consider the sequential description system D with, for every $\boldsymbol{y}$, pair $(\boldsymbol{y}, \boldsymbol{x}) \in D$ for $\boldsymbol{x} = \boldsymbol{y}$. That is, each source sequence has itself as a description. But then also $(\boldsymbol{y}', \boldsymbol{x}') \in D$ for all $\boldsymbol{x}' \succcurlyeq \boldsymbol{y}$ and $\boldsymbol{y}' \preccurlyeq \boldsymbol{y}$. That is, for every $\boldsymbol{y}$, every extension $\boldsymbol{x}'$ of $\boldsymbol{y}$ is also a description for every initial segment $\boldsymbol{y}'$ of $\boldsymbol{y}$. Now the corresponding *minimal* description system $\hat{D}$ is such that for every $\boldsymbol{y}$, $(\boldsymbol{y}, \boldsymbol{x})$ precisely if $\boldsymbol{y} = \boldsymbol{x}$: that is, the minimal description for $\boldsymbol{y}$ is again $\boldsymbol{y}$ itself. This description system leads to the uniform measure λ, via the correspondence explained below. As another example, take some infinite sequence $\boldsymbol{y}^\omega \in \mathbb{B}^\omega$. Let us now define two different D_1 and D_2. The first, D_1, is such that for every $t \in \mathbb{N}$, $(\boldsymbol{y}^t, \boldsymbol{x}^t) \in D_1$ for every $\boldsymbol{x}^t$; that is, every sequence of length t is a description for $\boldsymbol{y}^\omega$'s initial segment of length t. The corresponding minimal description system $\hat{D}_1$ also has those and only those pairs. The second is such that for every t, $(\boldsymbol{y}^t, \boldsymbol{x})$ for *every* $\boldsymbol{x} \in \mathbb{B}^*$: one can say that every sequence is a description for the whole of $\boldsymbol{y}^\omega$. The minimal description system contains for every t *only* the pair $(\boldsymbol{y}^t, \varnothing)$: one can say that the empty sequence is already a description for the whole of $\boldsymbol{y}^\omega$. Both description systems lead to the deterministic measure that assigns probability 1 to $\boldsymbol{y}^\omega$, via the correspondence explained below. (But note the difference between the two when we use the description length function given by (76) rather than (77) below!)

A.2.4.3. *Sequential description systems and measures on $\mathbb{B}^\omega \cup \mathbb{B}^*$.* To define a description length function L_D for sequential description system D, we consider the induced *minimal* description system $\hat{D}$, with $\hat{D}(\boldsymbol{y}, \boldsymbol{x})$ if and only if $D(\boldsymbol{y}, \boldsymbol{x})$ and not $\hat{D}(\boldsymbol{y}, \boldsymbol{x}')$ for any shorter $\boldsymbol{x}' \prec \boldsymbol{x}$. Then we can again define the description length function either by

$$L_D(\boldsymbol{y}) := \min\{|\boldsymbol{x}| : \hat{D}(\boldsymbol{y}, \boldsymbol{x})\} \tag{76}$$

or by

$$L_D(\boldsymbol{y}) := -\log \sum_{\boldsymbol{x}:\ \hat{D}(\boldsymbol{y}, \boldsymbol{x})} 2^{-|\boldsymbol{x}|}. \tag{77}$$

For the latter definition we have that for any sequential D the function

$$\begin{aligned} n_D(\cdot) &= 2^{-L_D(\cdot)} \\ &= \sum_{\boldsymbol{x}:\ \hat{D}(\cdot, \boldsymbol{x})} 2^{-|\boldsymbol{x}|} \end{aligned}$$

satisfies $n_D(\varnothing) \leq 1$ and $n_D(\boldsymbol{x}0) + n_D(\boldsymbol{x}1) \leq n_D(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{B}^*$, hence is a pre-measure to a measure on $\mathbb{B}^\omega \cup \mathbb{B}^*$. Conversely, for every measure ν on $\mathbb{B}^\omega \cup \mathbb{B}^*$ there is a sequential description system D_ν with

$$L_{D_\nu}(\boldsymbol{y}) = -\log \nu([\![\boldsymbol{y}]\!]) \tag{78}$$

for all $\boldsymbol{y} \in \mathbb{B}^*$. The proof of this correspondence is covered by the proof of the correspondence between *effective* sequential description systems and Σ_1 measures.

A.2.4.4. *Effective sequential description systems and* Σ_1 *measures.* A monotone machine M gives a decoding for a sequential description system D in the sense that $D(\boldsymbol{y}, \boldsymbol{x})$ precisely if $\boldsymbol{x}$ is an M-description for $\boldsymbol{y}$, i.e., $\Phi_M(\boldsymbol{x}) \succcurlyeq \boldsymbol{y}$. (This is no surprise given that the definition in A.2.4.1 above is designed to fit the definition of monotone machine in 2.1.2.4 above.) We call the D that are thus given by monotone machines the *effective* sequential description systems. Now consider the description length functions of the effective sequential descriptions systems, defined as in (77). (If we define L_D as in (76) we are again led to a variant of Kolmogorov complexity, see A.3.2 below.) Then for any effective sequential D with decoding given by monotone M we have that the function

$$\begin{aligned} n_D(\cdot) &= 2^{-L_D(\cdot)} \\ &= \sum 2^{-|\boldsymbol{x}|} \; [\![\boldsymbol{x} : \; \Phi_M(\boldsymbol{x}) \succcurlyeq \cdot \; \& \; \neg \exists \boldsymbol{x}' \prec \boldsymbol{x}. \; \Phi_M(\boldsymbol{x}') \succcurlyeq \cdot]\!] \end{aligned}$$

is a pre-measure to a measure on $\mathbb{B}^\omega \cup \mathbb{B}^*$, that is Σ_1 because we can effectively approximate from above the minimal M-description lengths for given $\boldsymbol{y}$. Conversely, for every Σ_1 measure ν on $\mathbb{B}^\omega \cup \mathbb{B}^*$, we can (effectively) construct a monotone machine that is the decoding function to a description system D_ν with

$$L_{D_\nu}(\boldsymbol{x}) = -\log \nu([\![\boldsymbol{x}]\!]) \tag{79}$$

for all $\boldsymbol{x} \in \mathbb{B}^*$.

A.2.4.5. *Transformations and* Σ_1 *measures.* Note that the unwieldy expression for n_D above can be rewritten as

$$n_D(\cdot) = \lambda([\![\boldsymbol{x} : \exists \boldsymbol{y}' \succcurlyeq \cdot . \; (\boldsymbol{x}, \boldsymbol{y}') \in M]\!]).$$

This shows that, similar to the prefix case, the correspondence between Σ_1 measures and effective sequential description systems is precisely the correspondence between Σ_1 measures and transformations of λ by monotone machines (2.1.2.6 above), i.e., proposition 2.5 (an instance of proposition 2.10) above. For the proof of the correspondence between Σ_1 measures and prequential description systems I therefore refer to the proof of proposition 2.10, in B.1.1 below.

A.2.4.6. *The Solomonoff-Levin measures.* In particular, the Solomonoff-Levin measures were defined as the *universal* uniform transformations λ_U. The Solomonoff-Leven measures are thus precisely those measures corresponding to the sequential description systems with decoders that are *universal* monotone machines.

* * *

A.3. Kolmogorov complexity

Here I continue A.2 by presenting versions of Kolmogorov complexity as description length functions for different effective description systems.

A.3.1. Prefix-free Kolmogorov complexity. I first consider the case of prefix description systems, A.2.2 above.

A.3.1.1. *The definition.* Let D be an effective prefix description system, i.e, the decoding function D^{-1} is given by a prefix-free machine T (A.2.2.8 above). Now the *prefix-free Kolmogorov complexity* via T is given by the description length function defined as in (72),

$$\begin{aligned} L_D(\boldsymbol{y}) &= \min\{|\boldsymbol{x}| : (\boldsymbol{y}, \boldsymbol{x}) \in D\} \\ &= \min\{|\boldsymbol{x}| : T(\boldsymbol{x}) = \boldsymbol{y}\}. \end{aligned}$$

Definition A.13. The prefix-free Kolmogorov complexity via prefix-free machine T is the function

$$K_T(\boldsymbol{y}) := \min\{|\boldsymbol{x}| : T(\boldsymbol{x}) = \boldsymbol{y}\}.$$

A.3.1.2. *Invariance.* Let U be a universal prefix-free Turing machine, definition A.1. It follows right from the definition that the Kolmogorov complexity K_U *additively minorizes* every K_T: for every prefix-machine T, there is a constant c such that for all $\boldsymbol{x}$,

$$K_U(\boldsymbol{x}) \leq K_T(\boldsymbol{x}) + c,$$

also simply written

$$K_U \leq^+ K_T. \tag{80}$$

Of course, for every two universal prefix-free machines U_1 and U_2, the functions K_{U_1} and K_{U_2} minorize each other; they are equivalent up to an additive constant. This fact is known as the invariance theorem (Li and Vitányi, 2008, 104ff, 200ff). We can say that K_{U_1} and K_{U_2} are *asymptotically equivalent* (*on average*) in the sense that for every $\boldsymbol{y}^\omega$,

$$K_{U_1}(\boldsymbol{y}^t)/t \xrightarrow{t\to\infty} K_{U_2}(\boldsymbol{y}^t)/t.$$

A.3.1.3. **Kolmogorov complexity of the natural numbers.* In the literature, K_T is often applied to elements of $\mathbb{N}$. Here '$K_T(n)$' for $n \in \mathbb{N}$ should be read as '$K_T(\boldsymbol{x}_n)$' with $\boldsymbol{x}_n$ the n-th element in the lexicographical ordering of all finite sequences (Li and Vitányi, 2008, 12). An upper bound on the Kolmogorov complexity on natural numbers is given by

$$K_U(n) \leq^+ 2\log n, \tag{81}$$

because the convergence of the series $\sum_n n^{-2}$ means that for some constant c the set $\{(\boldsymbol{x}_n, cn^2)\}_n$ is a valid bounded request set, which via the KC theorem A.8 implies the existence of a prefix-free machine T with $K_T(n) \leq \log c + \log n^2$ (also see Shen et al., 20xx).

A.3.1.4. **Plain Kolmogorov complexity.* The variant of descriptive complexity originally proposed by Kolmogorov (1965) is defined in terms of all p.c. functions or Turing machines, rather than just the prefix-free machines. The original definition, nowadays known as the *plain* Kolmogorov complexity and denoted by the letter 'C,' is the more straightforward; but as a complexity measure it has certain drawbacks that the prefix-free variant overcomes (Li and Vitányi, 2008, 197ff; Nies, 2009, 82f). Since the sum $\sum_{\boldsymbol{y} \in \mathbb{B}^*} 2^{-C(\boldsymbol{y})}$ does not generally converge, there is also no correspondence between the description length functions C and probability functions.

A.3.1.5. *Kolmogorov complexity and Σ_1 distributions.* The prefix-free Kolmogorov complexity via machine T corresponds to the Σ_1 semi-distribution 2^{-K_T}. However, not every Σ_1 semi-distribution can be expressed in this form. The definition of Kolmogorov complexity, based on the definition (72) of a description length function that only takes into account the single shortest description, has a certain clunkiness that is reminiscient of that of the integer-valued length functions of nonidealized prefix *code* systems (A.2.2.6 above): in both cases there is a less than perfect correspondence with semi-distributions.

A.3.1.6. *Kolmogorov complexity and prefix-free transformations.* The definition (73) of a description length function, that for effective description systems gives the definition of a uniform transformation by a prefix-free machine (A.2.2.10 above), takes into account *all* descriptions. The latter definition, we saw, corresponds precisely to the Σ_1 semi-distributions (A.1.2.3, A.2.2.9 above). What can we say about the relation between this definition and the Kolmogorov complexity? Precisely: what is the relation, for given T, between the description length function K_T and the description length function given by (73), i.e., the negative logarithm of the transformation λ_U,

$$-\log \lambda_T(\boldsymbol{y}) = -\log \sum_{\boldsymbol{x}:T(\boldsymbol{x}=\boldsymbol{y})} 2^{-|\boldsymbol{x}|}?$$

One direction is obvious: since

$$2^{-\min\{|\boldsymbol{x}|:T(\boldsymbol{x}=\boldsymbol{y})\}} \leq \sum_{\boldsymbol{x}:T(\boldsymbol{x})=\boldsymbol{y}} 2^{-|\boldsymbol{x}|},$$

we have

$$-\log \lambda_T \leq K_T. \tag{82}$$

In the other direction, however, there can be a serious divergence. To illustrate, let $(\boldsymbol{y}_i)_{i \in \mathbb{N}}$ be a listing of all sequences, and consider the prefix-free machine T with for all $\boldsymbol{y}_i$, $\{(1^i 0 \boldsymbol{x}, \boldsymbol{y}_i) : \boldsymbol{x} \in \mathbb{B}^{2^i}\} \subseteq T$. That is, each $\boldsymbol{y}_i$ has a total number of 2^{2^i} descriptions, each of length $2^i + i + 1$. Hence $K_T(\boldsymbol{y}_i) = 2^i + i + 1$, but

$$-\log Q_T(\boldsymbol{y}_i) = -\log 2^{2^i} 2^{-(2^i+i+1)} = i + 1.$$

Thus the gap grows exponentially.

A.3.1.7. *The coding theorem.* Nevertheless, it is not hard to tweak the machine T of the previous example in such a way that K_T again coincides with $-\log \lambda_T$: simply replace the above by $(1^i, \boldsymbol{y}_i) \in T$ for all i. In general, for every given Σ_1 semi-distribution p it is possible to construct some prefix-free machine T such that $K_T \leq^+ -\log p$ does hold. Inserting the universal transformation λ_U for p here, as well as in (82), we then get that

$$K_T \leq^+ -\log \lambda_U \leq^+ K_U.$$

But then by the minorization property (80) of K_U we actually obtain equivalence of K_U and $-\log \lambda_U$ up to an additive constant; in particular, we have that 2^{-K_U} is a universal Σ_1 semi-distribution. This is the *coding theorem* (Levin, 1974; Chaitin, 1975). The interpretation is that (Li and Vitányi, 2008, 277)

> A priori, an outcome x may have high probability because it has many long descriptions. The coding theorem ... tells us that in that case it must have a short description too. In other words, the a priori probability of x is dominated by the shortest program for x.

THEOREM A.14 (Coding theorem, Levin, Chaitin). *For every universal prefix-free machine U,*

$$-\log \lambda_U =^+ K_U,$$

equivalently,

$$\lambda_U =^\times 2^{-K_U}.$$

PROOF. See the proof of theorem A.16 in B.1.7. □

A.3.1.8. *A generalized Kolmogorov complexity.* In the previous, we took the perspective of description length functions, comparing Kolmogorov complexity, that only takes the shortest description into account, to the description length function that takes all descriptions into account—the latter is the negative logarithm of a uniform transformation λ_T. The complementary perspective is that of *distributions*, where we compare a transformation

$$\lambda_T = \sum_{\boldsymbol{x}:T(\boldsymbol{x})=\cdot} \lambda(\boldsymbol{x}),$$

that takes all descriptions into account, to the semi-distribution that only takes the *maximal-probability* description into account:

$$\begin{aligned} 2^{-K_T} &= 2^{-\min\{|\boldsymbol{x}|:T(\boldsymbol{x})=\cdot\}} \\ &= \lambda(\arg\min_{\boldsymbol{x}:T(\boldsymbol{x})=\cdot} |\boldsymbol{x}|) \\ &= \max\{\lambda(\boldsymbol{x}) : T(\boldsymbol{x}) = \cdot\}. \end{aligned}$$

The uniform transformations are a particular case of the general definition of a transformation

$$\mu_T = \sum_{\boldsymbol{x}:T(\boldsymbol{x})=\cdot} \mu(\boldsymbol{x})$$

by a prefix-free machine of a computable measure μ. If we restrict this definition to the single maximal-probability description, we obtain the semi-measure

$$\max\{\mu(\boldsymbol{x}) : T(\boldsymbol{x}) = \cdot\}. \tag{83}$$

We can now infer a generalized notion of prefix-free Kolmogorov complexity K_T^μ, such that $2^{-K_T^\mu}$ equals (83), i.e.,

$$K_T^\mu(\boldsymbol{y}) := -\log\max\{\mu(\boldsymbol{x}) : T(\boldsymbol{x}) = \boldsymbol{y}\}.$$

(So in particular $K_T = K_T^\lambda$.)

A.3.1.9. *A generalized coding theorem.* Can we also generalize theorem A.14 to show that, for computable μ other than λ, Kolmogorov complexity K_U^μ and the negative logarithm of the universal transformation μ_U are equivalent up to an additive constant? We do obviously have the generalization of (82),

$$-\log\mu_T \leq K_T^\mu. \tag{84}$$

It is in fact not so obvious that the minorization property (80) also generalizes to $K_U^\mu \leq^+ K_T^\mu$—though in the end it does, at least for those computable measures μ that fulfill a condition that is still slightly stronger than being continuous. Namely, it holds for every μ that is *conditionally bounded away from 0*: there is a $d \in \mathbb{N}$ such that for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{B}^*$ it holds that $\mu(\boldsymbol{x} \mid \boldsymbol{y}) \geq 2^{-d}$. I state this as

LEMMA A.15. For every computable continuous measure μ that is conditionally bounded away from 0, for every universal prefix-free machine U and prefix-free machine T,

$$K_U^\mu \leq^+ K_T^\mu.$$

PROOF. See B.1.7.2. □

Conjoined with this lemma, a generalization of the construction to prove the original coding theorem then does yield the statement for K_U^μ, with μ any computable measure that is conditionally bounded away from 0.

THEOREM A.16 (Generalized coding theorem). *For every computable continuous measure μ that is conditionally bounded away from 0, for every universal prefix-free U,*

$$K_U^\mu =^+ -\log\mu_U.$$

PROOF. See B.1.7.

A.3.2. Monotone Kolmogorov complexity. As for the prefix description systems, there is a notion of complexity connected to the sequential description systems (A.2.4 above).

A.3.2.1. *The definition.* Let D be an effective sequential description system, i.e, the decoding function D^{-1} is given by a monotone machine M (A.2.2.8 above). Now the *monotone Kolmogorov complexity* via M is again given by a description length function for D that only takes into account the single shortest description:

$$\begin{aligned} L_D(\boldsymbol{y}) &= \min\{|\boldsymbol{x}| : \hat{D}(\boldsymbol{y}, \boldsymbol{x})\} \\ &= \min\{|\boldsymbol{x}| : \exists \boldsymbol{y}' \succcurlyeq \boldsymbol{y}. (\boldsymbol{x}, \boldsymbol{y}') \in M\}. \end{aligned}$$

Definition A.17 (Levin). The monotone Kolmogorov complexity via monotone machine M is the function

$$Km_M(\boldsymbol{y}) := \min\{|\boldsymbol{x}| : \exists \boldsymbol{y}' \succcurlyeq \boldsymbol{y}. (\boldsymbol{x}, \boldsymbol{y}') \in M\}.$$

A.3.2.2. *Invariance.* As before, for every universal monotone machine U and other monotone machine M it holds that

$$Km_U \leq^+ Km_M,$$

and hence that for every two universal monotone machines U_1 and U_2

$$Km_{U_1} =^+ Km_{U_2}.$$

A.3.2.3. *The relation with prefix-free Kolmogorov complexity.* Since prefix-free machines arise as a restriction of monotone machines (above), the monotone Kolmogorov complexity is smaller than the prefix-free Kolmogorov complexity in the sense that for universal monotone U and universal prefix-free U' we have (also see Shen et al., 20xx)

$$Km_U \leq^+ K_{U'}. \tag{85}$$

A.3.2.4. *The complexity KM.* The complexity measure that corresponds to the effective description length function (77), i.e., a uniform transformation by a monotone machine, is in the literature often denoted

$$KM_M := -\log \lambda_M.$$

A.3.2.5. *The failure of a coding theorem.* Like in the prefix-free case (82), it is obvious from the fact that

$$2^{-\min\{|\boldsymbol{x}|:\exists \boldsymbol{y}' \succcurlyeq \boldsymbol{y}.(\boldsymbol{x},\boldsymbol{y}') \in M\}} \leq \sum_{\boldsymbol{x} \in \lfloor \{\boldsymbol{x} : \exists \boldsymbol{y}' \succcurlyeq \boldsymbol{y}\} \rfloor} 2^{-|\boldsymbol{x}|}$$

that we have

$$KM_M \leq Km_M.$$

But in the other direction things are worse than in the prefix-free case. Specifically, there is no coding theorem that at least shows equivalence up to an additive constant for universal monotone machines. Day (2011), improving on work by Gács (1983), showed that for any real $r < 1$ there are infinitely many $\boldsymbol{x}$ with

$$Km_U(\boldsymbol{x}) > KM_U(\boldsymbol{x}) + r \log\log|\boldsymbol{x}|.$$

A.3.2.6. **Other complexities and relations.* An early overview of the different versions of Kolmogorov complexity and their relations, including notes on the history and terminology, is given by Uspensky (1992; a later version with proofs is Uspensky and Shen, 1996.)

* * *

A.4. Randomness

The motivation for the field of algoritmic randomness is the characterization of *random* infinite sequences, in particular, uniformly random sequences: data streams produced by the uniform measure λ. Here I discuss the main notion of algorithmic randomness, the one due to Martin-Löf (1966).

A.4.1. The definition of Martin-Löf randomness.

An intuition about a data stream generated from an i.i.d. uniform measure is that it will not have any properties that make it stand out from the majority of other possible data streams: it will be a *typical* sequence. A way to try and make this precise is to say that it will satisfy all probability-1 properties, or equivalently, that it will avoid all probability-0 properties. That is, it will avoid all *null sets*, all subsets of $\mathbb{B}^\omega$ that have uniform probability 0.

However, such a definition of a random sequence would render *no* infinite sequence random: the singleton $\{\boldsymbol{x}^\omega\}$ is always a null set. In other words, *no* infinite sequence can be typical in *all* respects, if these are identified with all properties of probability 1: the class $\mathbb{B}^\omega \setminus \{\boldsymbol{x}^\omega\}$ always has uniform probability 1. The basic idea due to Martin-Löf is therefore to restrict the typical properties to the *effective* ones.

More accurately, the idea is that possessing an exceptional property, or being contained in a null set, can be detected by means of a *statistical test*: and the latter are restricted to the effective ones. The definition of a statistical test in this context starts with the notion of an *open set*, which is a set of the form $[\![A]\!]$ for some $A \subseteq \mathbb{B}^*$. Now a set $\mathcal{A}$ is null precisely if there is a uniform sequence $(G_m)_{m\in\mathbb{N}}$ of open sets such that $\lim_{m\to\infty} \lambda(G_m) = 0$ and $\mathcal{A} \subseteq \bigcap_m G_m$. The sequence $(G_m)_{m\in\mathbb{N}}$ can be seen as a test for the exceptional property $\mathcal{A}$: Martin-Löf further put effectiveness constraints on both the sequence of tests and its convergence to measure 0. (See Nies, 2009, 102ff.)

A.4.1.1. *The definition.* Thus a *Martin-Löf (ML)-test* is defined to be a uniformly c.e. sequence $(G_m)_{m\in\mathbb{N}}$ of open sets, such that, moreover, the measure $\lambda(G_m) \leq 2^{-m}$ for all $m \in \mathbb{N}$. A sequence $\boldsymbol{x}^\omega$ is ML-random if it passes each ML-test: $\boldsymbol{x}^\omega \notin \bigcap_m G_m$ for each ML-test $(G_m)_{m\in\mathbb{N}}$. This definition generalizes to computable μ other than the uniform measure.

Definition A.18 (Martin-Löf randomness). A *μ-ML-test* is a uniformly c.e. sequence $(G_m)_{m\in\mathbb{N}}$ of open sets such that $\mu(G_m) \leq 2^{-m}$ for all $m \in \mathbb{N}$.

A sequence $\boldsymbol{x}^\omega$ is μ-ML-random if it passes each μ-ML-test: $\boldsymbol{x}^\omega \notin \bigcap_m G_m$ for each μ-ML-test $(G_m)_{m\in\mathbb{N}}$.

A.4.1.2. *Universal ML-tests.* A central fact is the existence of *universal* μ-ML-tests, that include all other tests. Precisely, a univeral μ-ML-test $(U_m)_{m\in\mathbb{N}}$ is a μ-ML-test such that $\bigcap_m U_m \supseteq \bigcap_m G_m$ for each other μ-ML-test $(G_m)_m$. That means that if $\boldsymbol{x}^\omega$ passes a universal μ-ML-test, it passes *all* μ-ML-tests, and it is μ-ML-random. Conversely, if $\boldsymbol{x}^\omega$ does not pass a universal μ-ML-test, it is not μ-ML-random. Hence a sequence is random if and only if it passes a universal μ-ML-test.

A.4.1.3. *Effectiveness.* A μ-ML-test is effective in the sense that it is uniformly Σ_1. More concretely, we can effectively verify that a sequence $\boldsymbol{x}^\omega$ fails the μ-ML test at a given significance level m: this is a Σ_1 property. Still, the actual failure to be μ-ML-random (i.e., failing a test at all significance levels) transcends effective verifiability: this is a Π_2 property.

A.4.1.4. *ML-randomness for semi-measures.* Note, too, that the definition is restricted to μ-ML-randomness for *computable* measures μ. I discuss the possibility of extending this definition to the Σ_1 measures in general in A.4.2.7 below.

A.4.2. Other characterizations of Martin-Löf randomness.

There is another intuition about the nature of a random sequence, that might actually appear somewhat in tension with the intuition of typicality. This intuition is that a random sequence is also highly *irregular* in the sense of lacking clear patterns. As such, a random sequence is highly *complex*: this invites its characterization in terms of Kolmogorov complexity; moreover, it is highly *unpredictable*: this invites its characterization in terms of predictive success. Interestingly, the different characterizations following from these intuitions lead to the same notion.

A.4.2.1. *Prefix-free Kolmogorov complexity.* A central result is *Schnorr's theorem* (1973; or the *Schnorr-Levin* theorem, see below), that characterizes Martin-Löf randomness in terms of prefix-free Kolmogorov complexity.

THEOREM A.19 (Schnorr). *A sequence $\boldsymbol{x}^\omega$ is μ-ML-random if and only if $K_U(\boldsymbol{x}^t) \geq^+ -\log\mu(\boldsymbol{x}^t)$. In particular, sequence $\boldsymbol{x}^\omega$ is ML-random if and only if $K_U(\boldsymbol{x}^t) \geq^+ t$.*

PROOF SKETCH. The intuition is that a universal ML-test can be seen as a uniformly c.e. sequence of prefix-free sets of low complexity, and so a sequence $\boldsymbol{x}^\omega$ is *not* random (is caught in a universal ML-test) precisely if infinitely many of its initial segments are of low complexity. Somewhat more precisely (for a detailed proof, see Nies, 2009, 108; or see the derivation in A.4.2.5 below), the left-to-right direction follows from the fact that the open sets $O_m = \{\boldsymbol{x}^\omega : \exists t . K_U(\boldsymbol{x}^t) < -\log\mu(\boldsymbol{x}^t) - m\}$ form a sequential test, that shows a sequence that is not complex (i.e., $K_U(\boldsymbol{x}^t) \not\geq^+ -\log\mu(\boldsymbol{x}^t)$) to be nonrandom; and the right-to-left direction follows from the fact that we can convert a universal μ-ML-test via the KC theorem A.8 into a machine that gives description lengths in accordance with the above, showing that a sequence that is not random (caught by the universal test) is not complex. □

A.4.2.2. *Monotone Kolmogorov complexity.* Van Lambalgen (1987a, 147) writes,

> for an infinite sequence to be random it is necessary and sufficient if it has no (except perhaps finitely many) initial segments of low complexity. In other words, any complexity measure C is able to characterise Martin-Löf randomness if the universal sequential test can be written in terms of C. Nothing more is necessary, but much more is possible.

We can indeed give an identical characterization in terms of *monotone* Kolmogorov complexity (A.3.2 above).

THEOREM A.20 (Levin, 1973). *A sequence $\boldsymbol{x}^\omega$ is μ-ML-random if and only if $Km_U(\boldsymbol{x}^t) \geq^+ -\log \mu(\boldsymbol{x}^t)$. In particular, sequence $\boldsymbol{x}^\omega$ is ML-random if and only if $Km_U(\boldsymbol{x}^t) \geq^+ t$.*

In fact, theorem A.20 still holds when we substitute the complexity measure KM_U for Km_U (A.3.2.4 above). Since KM_U is precisely the negative logarithm of the universal transformation λ_U, i.e., of the Solomonoff-Levin measure, that means that a sequence $\boldsymbol{x}^\omega$ is μ-random if and only if

$$-\log Q_U(\boldsymbol{x}^t) \geq^+ -\log \mu(\boldsymbol{x}^t), \tag{86}$$

or

$$Q_U(\boldsymbol{x}^t) \leq^\times \mu(\boldsymbol{x}^t). \tag{87}$$

(Also see Shen et al., 20xx.)

A.4.2.3. *Martin-Löf randomness and the Σ_1 measures.* We thus have that a sequence $\boldsymbol{x}^\omega$ is μ-ML-random precisely if $Q_U(\boldsymbol{x}^t) \leq^\times \mu(\boldsymbol{x}^t)$; or, since by the universality of the Solomonoff-Levin measure also $Q_U(\boldsymbol{x}^t) \geq^\times \mu(\boldsymbol{x}^t)$,

$$Q_U(\boldsymbol{x}^t) =^\times \mu(\boldsymbol{x}^t). \tag{88}$$

Hence a sequence $\boldsymbol{x}^\omega$ is μ-Martin-Löf random precisely if the Solomonoff-Levin predictor does not have higher likelihoods on $\boldsymbol{x}^\omega$ than μ does. Equivalently, a sequence $\boldsymbol{x}^\omega$ is μ-Martin-Löf random precisely if *no* $\nu \in \Sigma_1$ has higher likelihoods on $\boldsymbol{x}^\omega$ than μ does: for every $\nu \in \Sigma_1$,

$$\nu(\boldsymbol{x}^t) \leq^\times \mu(\boldsymbol{x}^t). \tag{89}$$

A related interpretation is that μ is in a precise sense *the best explanation* (among all Σ_1 measures) for $\boldsymbol{x}^\omega$.

A.4.2.4. *Prediction and predictive complexity.* With L the cumulative log-loss function, we have that a sequence $\boldsymbol{x}^\omega$ is μ-ML-random precisely if

$$L_{Q_U}(\boldsymbol{x}^t) =^+ L_\mu(\boldsymbol{x}^t). \tag{90}$$

This gives a predictive interpretation of randomness: a sequence $\boldsymbol{x}^\omega$ is μ-Martin-Löf random precisely if the Solomonoff-Levin predictor cannot do better on $\boldsymbol{x}^\omega$ than μ does. Equivalently, a sequence $\boldsymbol{x}^\omega$ is μ-Martin-Löf random precisely if *no* $\nu \in \Sigma_1$ can do better on $\boldsymbol{x}^\omega$ than μ does: for every $\nu \in \Sigma_1$,

$$L_\nu(\boldsymbol{x}^t) \geq^+ L_\mu(\boldsymbol{x}^t). \tag{91}$$

In terms of Vovk's notion of predictive complexity (6.2), a sequence is μ-ML-random precisely if its predictive complexity for the log-loss, that is by definition given by the universal Π_1 superloss process, is already given by the Δ_1 loss process corresponding to μ.

A.4.2.5. *A derivation of the equivalence.* The predictive interpretation is usually presented as a *gambling* interpretation (intuition: it is impossible to make money on a random sequence; recall von Mises's idea, I.6), and put in terms of c.e. *martingales*, a notion representing gambling strategies that is formally very close to the notion of (a predictor corresponding to) a Σ_1 measure. For completeness, and to provide some more intuition, I will here give a direct derivation of the equivalence of the characterization of Martin-Löf randomness in terms of tests and in terms of Σ_1 measures, that is a translation of the proof for martingales as given by Nies (2009, 265f).

Proof of $\boldsymbol{x}^t$ is μ-ML-random $\iff \forall \nu \in \Sigma_1.\ L_\nu(\boldsymbol{x}^t) \geq^+ L_\mu(\boldsymbol{x}^t)$. Assume without loss of generality that for all μ-ML-tests $(G_m)_m$ it holds that $G_m \supseteq G_{m+1}$ for all m and that $\sum_m \mu(G_m) \leq 1$.

First, suppose that $\boldsymbol{y}^\omega$ fails the test $(G_m)_{m\in\mathbb{N}}$. Consider the function $\nu_G : \boldsymbol{x}^t \mapsto \sum_{m\in\mathbb{N}} \mu(G_m \cap [\![\boldsymbol{x}^t]\!])$. It is clearly semi-computable, and it is a semi-measure on $\mathbb{B}^\omega$ because $\mu(G_m \cap [\![\boldsymbol{x}^t]\!]) = \mu(G_m \cap [\![\boldsymbol{x}^t 0]\!]) + \mu(G_m \cap [\![\boldsymbol{x}^t 1]\!])$ for all $\boldsymbol{x}^t$ and $\nu(\boldsymbol{\varnothing}) = \sum_{m\in\mathbb{N}} \mu(G_m) \leq 1$. Then

$$\begin{aligned}
\boldsymbol{y}^\omega \in \bigcap_m G_m &\Rightarrow \forall m.\ \boldsymbol{y}^\omega \in G_m \\
&\Rightarrow \forall m \exists t.\ [\![\boldsymbol{y}^n]\!] \subseteq G_m \\
&\Rightarrow \forall m \exists t \forall j \leq m.\ [\![\boldsymbol{y}^n]\!] \subseteq G_j \\
&\Rightarrow \forall m \exists t.\ \sum_j \mu(G_j \mid \boldsymbol{y}^t) > m \\
&\Rightarrow \forall m \exists t.\ \frac{\nu(\boldsymbol{y}^t)}{\mu^*(\boldsymbol{y}^t)} > m \\
&\Rightarrow \forall c \exists t.\ L_\mu(\boldsymbol{y}^t) > L_\nu(\boldsymbol{y}^t) + c.
\end{aligned}$$

Conversely, suppose that $\boldsymbol{y}^\omega$ is such that L_μ fails to stay below L_ν in the sense of (91), for some ν. Consider the classes $G_m = \{\boldsymbol{x}^\omega : \exists t.\ L_\mu(\boldsymbol{x}^t) > L_\nu(\boldsymbol{x}^t) + m\}$

for all $m \in \mathbb{N}$. This gives a sequence $(G_m)_{m\in\mathbb{N}}$ of open sets that is clearly uniformly c.e.; it is a μ-ML-test because moreover

$$\begin{aligned}\mu(G_m) &= \mu\left(\boldsymbol{x}^\omega : \exists t.\ \nu(\boldsymbol{x}^t) \geq 2^m \mu(\boldsymbol{x}^t)\right) \\ &= \sum \mu(\boldsymbol{z}) [\![\boldsymbol{z} \in \lfloor\{\boldsymbol{x} : \nu(\boldsymbol{x}) \geq 2^m \mu(\boldsymbol{x})\}\rfloor]\!] \\ &\leq \mu(\boldsymbol{z}) \frac{\nu(\boldsymbol{z})}{2^m \mu(\boldsymbol{z})} [\![\boldsymbol{z} \in \lfloor\{\boldsymbol{x} : \nu(\boldsymbol{x}) \geq 2^m \mu(\boldsymbol{x})\}\rfloor]\!] \\ &= 2^m.\end{aligned}$$

Then

$$\begin{aligned}\forall c \exists t.\ L_\mu(\boldsymbol{y}^t) > L_\nu(\boldsymbol{y}^t) + c \Rightarrow \forall m \exists t.\ \mu(\boldsymbol{y}^t) &< 2^{-m}\nu(\boldsymbol{y}^t) \\ \Rightarrow \boldsymbol{y}^\omega &\in \bigcap_m G_m. \qquad \square\end{aligned}$$

Thus the property of L_μ staying below the particular Π_1 superloss process L_ν on $\boldsymbol{y}^\omega$, up to some constant, is assessed by a particular μ-ML-test that attempts to falsify this for all constants or significance levels m. If $\boldsymbol{y}^\omega$ passes the test and so escapes falsification at some level m, i.e., $\boldsymbol{y}^\omega \notin G_m$, that means that L_μ stays below L_ν on $\boldsymbol{y}^\omega$ up to this m, and $\boldsymbol{y}^\omega$ can still be μ-ML-random; if it fails the test, i.e., $\boldsymbol{y}^\omega \in \bigcap_m G_m$, that means that L_μ fails to stay below L_ν on $\boldsymbol{y}^\omega$ up to *any* constant, and $\boldsymbol{y}^\omega$ is not μ-ML-random. Naturally, a *universal* μ-ML-test, that includes all other tests, assesses a *universal* Π_1 superloss process, that minorizes all other Π_1 superloss processes.

A.4.2.6. *The weakest of randomness assumptions.* Recall from 3.1.3.4 the interpretation by Levin of the Solomonoff-Levin measure as a universal a priori measure, with the motivation that "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" (Zvonkin and Levin, 1970, 104, my notation). Since we trivially have $Q_U(\boldsymbol{x}^t) =^\times Q_U(\boldsymbol{x}^t)$, by characterization (88) of Martin-Löf randomness it follows that *every single sequence* $\boldsymbol{x}^\omega$ is Q_U-ML-random.

A.4.2.7. *ML-randomness for semi-measures.* However, recall that the original defintion of μ-ML-randomness was restricted to *computable* measures μ. It is indeed the case that if ν is Σ_1 but not Δ_1, then the negation of (91) is no longer Σ_1, and the corresponding test is no longer uniformly Σ_1—so it is not a proper ML-test. This cannot be solved by only requiring a ν-ML-test to be uniformly $\Sigma_1(\nu)$ (so allowing ν as an oracle—for computable μ this reduces to the standard definition). Namely, as mentioned above, for a universal $\mathring{\nu}$ *every* sequence should be ν-ML-random according to (90), but it is not hard to define a uniformly $\Sigma_1(\mathring{\nu})$ sequence of classes $(G_m)_m$ with $\mathring{\nu}(G_m) \leq 2^{-m}$ for every m and $\bigcap_m G_m \neq \emptyset$. Hence the predictive characterization that naturally extends to the whole class $\mathcal{M}$ of Σ_1 measures leads to an unclear notion from the perspective of tests. (See Bienvenu et al., 2017 for a different attempt at extending the definition of Martin-Löf randomness to Σ_1 measures.)

*

APPENDIX B

Proofs

B.1. The Σ_1 measures and semi-distributions

B.1.1. The proof of proposition 2.10.

B.1.1.1. *The proof of ($\Leftarrow$).* The easy direction is that every effective transformation μ_M of a Δ_1 measure μ is a Σ_1 measure (see Day, 2011, theorem 4(i)).

Proof. Given Δ_1 measure μ and monotone machine M. We have to verify that $\mu_M : \mathbb{B}^* \to [0,1]$ is lower semi-computable and satisfies $\mu_M(\varnothing) \leq 1$ and $\mu_M(\boldsymbol{x}0) + \mu_M(\boldsymbol{y}1) \leq \mu_M(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{B}^*$.

First, let M_s denote the set of pairs enumerated in c.e. M by stage s, and let $g(\boldsymbol{y}, t) = \mu_{M_t}(\boldsymbol{y})$. Clearly, g is computable, nondecreasing in t, and $\lim_{t\to\infty} g(\boldsymbol{y}, t) = \mu_M(\boldsymbol{y})$, meaning that μ_M is semi-computable.

Furthermore, we have that for every $\boldsymbol{x}$ such that $\exists \boldsymbol{y}' \succcurlyeq \boldsymbol{y}y.\ (\boldsymbol{x}, \boldsymbol{y}') \in M$ certainly $\exists \boldsymbol{y}' \succcurlyeq \boldsymbol{y}.\ (\boldsymbol{x}, \boldsymbol{y}') \in M$, so

$$\llbracket \{\boldsymbol{x} : \exists \boldsymbol{y}' \succcurlyeq \boldsymbol{y}0.\ (\boldsymbol{x}, \boldsymbol{y}') \in M\} \rrbracket \cup \llbracket \{\boldsymbol{x} : \exists \boldsymbol{y}' \succcurlyeq \boldsymbol{y}1.\ (\boldsymbol{x}, \boldsymbol{y}') \in M\} \rrbracket \subseteq \llbracket \{\boldsymbol{x} : \exists \boldsymbol{y}' \succcurlyeq \boldsymbol{y}y.\ (\boldsymbol{x}, \boldsymbol{y}') \in M\} \rrbracket;$$

and we have that for no $\boldsymbol{x}$ both $\exists \boldsymbol{y}' \succcurlyeq \boldsymbol{y}0.\ (\boldsymbol{x}, \boldsymbol{y}') \in M$ and $\exists \boldsymbol{y}' \succcurlyeq \boldsymbol{y}1.\ (\boldsymbol{x}, \boldsymbol{y}') \in M$ (or we would get $\boldsymbol{y}0 \sim \boldsymbol{y}1$ from property (14) of M), so

$$\llbracket \{\boldsymbol{x} : \exists \boldsymbol{y}' \succcurlyeq \boldsymbol{y}0.\ (\boldsymbol{x}, \boldsymbol{y}') \in M\} \rrbracket \cap \llbracket \{\boldsymbol{x} : \exists \boldsymbol{y}' \succcurlyeq \boldsymbol{y}1.\ (\boldsymbol{x}, \boldsymbol{y}') \in M\} \rrbracket = \emptyset.$$

Hence $\mu_M(\boldsymbol{x}0) + \mu_M(\boldsymbol{y}1) \leq \mu_M(\boldsymbol{x})$, and also $\mu_M(\varnothing) \leq \mu(\mathbb{B}^*) = 1$. □

B.1.1.2. *The derivation from Levin's results.* Proposition 2.10 is attributed to Levin because the hard direction can already be derived from theorems 3.1(b) and 3.2 in Zvonkin and Levin (1970). Theorem 3.2 in that paper is the characterization $\mathcal{M} = \{\lambda_M\}_M$ in terms of uniform transformations, i.e., proposition 2.5 above; the hard part is that for every Σ_1 measure ν we can construct a machine M such that $\lambda_M = \nu$. Theorem 3.1(b) in that paper includes the assertion that for every continuous Δ_1 measure μ we can construct a machine M such that $\mu_M = \lambda$. Together, we have that for every $\nu \in \Sigma_1$, $\mu \in \Delta_1$ we can construct M_1, M_2 with $\nu = (\mu_{M_1})_{M_2}$, i.e., $\nu = \mu_M$ for machine M defined by

$$(\boldsymbol{x}, \boldsymbol{z}) \in M \quad \text{if and only if} \quad \exists \boldsymbol{y}, \boldsymbol{y}'.\ \boldsymbol{y}' \succcurlyeq \boldsymbol{y} \ \&\ (\boldsymbol{x}, \boldsymbol{y}') \in M_1 \ \&\ (\boldsymbol{y}, \boldsymbol{z}) \in M_2.$$

B.1.1.3. *The direct proof of* ($\Rightarrow$)*.* Levin's presentation (reproduced in Li and Vitányi, 2008, theorem 4.5.2) of a construction to transform λ into any given $\nu \in \Sigma_1$ is very quick; a more detailed construction was published by Day (2011, theorem 4(ii); also see Downey and Hirschfeldt, 2010, theorem 3.16.2(ii)).[29] The following direct proof of the case for any continuous Δ_1 measure is an adaptation of this construction.

Proof of ($\Rightarrow$). Given Δ_1 measure μ, and Σ_1 measure ν with computable approximation function g. We construct in stages $s = \langle \boldsymbol{y}, t\rangle$ a monotone machine $M = \bigcup_s M_s$ that transforms μ into ν. Let $D_s(\boldsymbol{y}) := \{\boldsymbol{x} \in \mathbb{B}^* : (\boldsymbol{x}, \boldsymbol{y}) \in M_s\}$.

Construction. Let $M_0 := \emptyset$.

At stage $s = \langle \boldsymbol{y}, t\rangle$, if $\mu(\llbracket D_{s-1}(\boldsymbol{y})\rrbracket) = g(\boldsymbol{y}, t)$ then let $M_s := M_{s-1}$.

Otherwise, first consider the case $\boldsymbol{y} \neq \boldsymbol{\varnothing}$. By lemma 1 in Day (2011) there is a set $R \subseteq \mathbb{B}^s$ of *available* sequences of length s such that $\llbracket R\rrbracket = \llbracket D_{s-1}(\boldsymbol{y}^-)\rrbracket \setminus (\llbracket D_{s-1}(\boldsymbol{y}^-0)\rrbracket \cup \llbracket D_{s-1}(\boldsymbol{y}^-1)\rrbracket)$. Denote $a := \mu(\llbracket R\rrbracket)$, the amount of measure available for descriptions for $\boldsymbol{y}$, which equals $\mu(\llbracket D_{s-1}(\boldsymbol{y}^-)\rrbracket) - \mu(\llbracket D_{s-1}(\boldsymbol{y}^-0)\rrbracket) - \mu(\llbracket D_{s-1}(\boldsymbol{y}^-1)\rrbracket)$ because we ensure by construction that $\llbracket D_{s-1}(\boldsymbol{y}^-)\rrbracket \supseteq \llbracket D_{s-1}(\boldsymbol{y}^-0)\rrbracket \cup \llbracket D_{s-1}(\boldsymbol{y}^-1)\rrbracket$ and $\llbracket D_{s-1}(\boldsymbol{y}^-0)\rrbracket \cap \llbracket D_{s-1}(\boldsymbol{y}^-1)\rrbracket = \emptyset$. Denote $b := g(\boldsymbol{y}, t) - \mu(\llbracket D_{s-1}(\boldsymbol{y})\rrbracket)$, the amount of measure the current descriptions fall short of the latest approximation of $\nu(\boldsymbol{y})$. We collect in the auxiliary set A_s a number of available sequences from R such that $\mu(\llbracket A_s\rrbracket)$ is maximal while still bounded by $\min\{a, b\}$.

If $\boldsymbol{y} = \boldsymbol{\varnothing}$, then denote $b := g(\boldsymbol{\varnothing}, t) - \mu(\llbracket D_{s-1}(\boldsymbol{\varnothing})\rrbracket)$. Collect in A_s a number of available sequences from $R \subseteq \mathbb{B}^s$ with $\llbracket R\rrbracket = \mathbb{B}^\omega \setminus \llbracket D_{s-1}(\boldsymbol{\varnothing})\rrbracket$ such that $\mu(\llbracket A_s\rrbracket)$ is maximal but bounded by b.

Put $M_s := M_{s-1} \cup \{(\boldsymbol{x}, \boldsymbol{y}) : \boldsymbol{x} \in A_s\}$.

Verification. The verification of the fact that M is a monotone machine is identical to that in Day (2011).

It remains to prove that $\mu_M(\boldsymbol{y}) = \nu(\boldsymbol{y})$ for all $\boldsymbol{y} \in \mathbb{B}^*$. Since by construction $\llbracket D_s(\boldsymbol{y}')\rrbracket \subseteq \llbracket D_s(\boldsymbol{y})\rrbracket$ for any $\boldsymbol{y}' \succcurlyeq \boldsymbol{y}$, we have that $\mu_{M_s}(\boldsymbol{y}) = \mu(\cup_{\boldsymbol{y}' \succcurlyeq \boldsymbol{y}} \llbracket D_s(\boldsymbol{y}')\rrbracket) = \mu(\llbracket D_s(\boldsymbol{y})\rrbracket)$. Hence $\mu_M(\boldsymbol{y}) = \lim_{s\to\infty} \mu(\llbracket D_s(\boldsymbol{y})\rrbracket)$, and our objective is to show that $\lim_{s\to\infty} \mu(\llbracket D_s(\boldsymbol{y})\rrbracket) = \nu(\boldsymbol{y})$. To that end it suffices to demonstrate that for every $\delta > 0$ there is some stage s_0 where $\mu(\llbracket D_{s_0}(\boldsymbol{y})\rrbracket) > \nu(\boldsymbol{y}) - \delta$. We prove this by induction on the finite sequences.

For the base step, let $\boldsymbol{y} = \boldsymbol{\varnothing}$. Choose positive $\delta' < \delta$. There will be a stage $s_0 = \langle \boldsymbol{\varnothing}, t_0\rangle$ where $g(\boldsymbol{\varnothing}, t_0) > \nu(\boldsymbol{\varnothing}) - \delta'$, and (since μ is continuous) $\mu(\llbracket \boldsymbol{x}\rrbracket) \leq \delta - \delta'$ for all $\boldsymbol{x} \in \mathbb{B}^{s_0}$. Then, if not already $\mu(\llbracket D_{s_0-1}(\boldsymbol{\varnothing})\rrbracket) > \nu(\boldsymbol{\varnothing}) - \delta$, the latter guarantees that the construction will select a number of available sequences in A_{s_0} such that $\nu(\boldsymbol{\varnothing}) - \delta < \mu(\llbracket D_{s_0-1}(\boldsymbol{\varnothing})\rrbracket) + \mu(\llbracket A_s\rrbracket) \leq g(\boldsymbol{\varnothing}, t_0)$. It follows that $\mu(\llbracket D_{s_0}(\boldsymbol{\varnothing})\rrbracket) = \mu(\llbracket D_{s_0-1}(\boldsymbol{\varnothing})\rrbracket) + \mu(\llbracket A_s\rrbracket) > \nu(\boldsymbol{\varnothing}) - \delta$ as required.

For the inductive step, let $\boldsymbol{y} \neq \varnothing$, and denote by $\boldsymbol{y}'$ the one-bit extension of $\boldsymbol{y}^-$ with $\boldsymbol{y}' \mid \boldsymbol{y}$. Choose positive $\delta' < \delta$. By induction hypothesis, there exists a stage s_0' such that $\mu(\llbracket D_{s_0'}(\boldsymbol{y}^-)\rrbracket) > \nu(\boldsymbol{y}^-) - \delta'$. At this stage s_0', we have

$$\begin{aligned} \mu(\llbracket D_{s_0'}(\boldsymbol{y}^-)\rrbracket) - \mu(\llbracket D_{s_0'}(\boldsymbol{y}')\rrbracket) &\geq \mu(\llbracket D_{s_0'}(\boldsymbol{y}^-)\rrbracket - \nu(\boldsymbol{y}') \\ &> \nu(\boldsymbol{y}^-) - \delta' - \nu(\boldsymbol{y}') \\ &\geq \nu(\boldsymbol{y}) - \delta', \end{aligned}$$

where the last inequality follows from the semi-measure property $\nu(\boldsymbol{y}^-) \geq \nu(\boldsymbol{y}) + \nu(\boldsymbol{y}')$. There will be a stage $s_0 = \langle \boldsymbol{y}, t_0 \rangle \geq s_0'$ with $g(\boldsymbol{y}, t_0) > \nu(\boldsymbol{y}) - \delta'$ and $\mu(\llbracket \boldsymbol{x} \rrbracket) \leq \delta - \delta'$ for all $\boldsymbol{x} \in \mathbb{B}^{s_0}$. Clearly, we have $\min\{\mu(\llbracket D_{s_0}(\boldsymbol{y}^-)\rrbracket) - \mu(\llbracket D_{s_0}(\boldsymbol{y}')\rrbracket), g(\boldsymbol{y}, t_0)\} > \nu(\boldsymbol{y}) - \delta'$. Then, as in the base case, if not already $\mu(\llbracket D_{s_0 - 1}(\boldsymbol{y})\rrbracket) > \nu(\boldsymbol{y}) - \delta$, the construction selects a number of available descriptions such that $\mu(\llbracket D_{s_0}(\boldsymbol{y})\rrbracket) > \nu(\boldsymbol{y}) - \delta$ as required. □

B.1.2. The proof of proposition A.3. The fact that every Σ_1 semi-distribution q can be obtained as a transformation of λ is usually inferred (e.g., Downey and Hirschfeldt, 2010, 130; Nies, 2009, 90) from the effective version of Kraft's inequality that is called the KC theorem (A.2.2.8 above). However, we can easily prove the general case in a direct manner by a much simplified version of the construction for proposition 2.10 in B.1.1.3 above.

Proof. Given Δ_1 measure μ. Every transformation μ_T for a prefix-free machine T is lower semi-computable: to calculate $\mu_T(\boldsymbol{y})$, enumerate all possible inputs to T and add $\mu(\llbracket \boldsymbol{x} \rrbracket)$ to the approximation as soon as $T(\boldsymbol{x}) \downarrow = \boldsymbol{y}$. Moreover, μ_T is a semi-distribution: the set $D_T = \{\boldsymbol{x} : T(\boldsymbol{x}) \downarrow\}$ of *all* T-descriptions $\boldsymbol{x}$ is by definition prefix-free, hence all corresponding cones $\llbracket \boldsymbol{x} \rrbracket$ are disjoint and $\sum_{\boldsymbol{x} \in D_T} \mu(\llbracket \boldsymbol{x} \rrbracket) \leq 1$, which entails that $\sum_{\boldsymbol{y} \in \mathbb{B}^*} \mu_T(\boldsymbol{y}) \leq 1$.

Conversely, let q be a Σ_1 semi-distribution with computable approximation function g. We construct a prefix-free machine $T = \bigcup_s T_s$ with $\mu_T = q$ in stages $s = \langle \boldsymbol{y}, t \rangle$. Let $D_s(\boldsymbol{y}) = \{\boldsymbol{x} \in \mathbb{B}^* : (\boldsymbol{x}, \boldsymbol{y}) \in T_s\}$.

Construction. Let $T_0 = \emptyset$.

At stage $s = \langle \boldsymbol{y}, t \rangle$, if $\mu(\llbracket D_{s-1}(\boldsymbol{y})\rrbracket) = g(\boldsymbol{y}, t)$ then let $T_s := T_{s-1}$.

Otherwise, let the set $R \subseteq \mathbb{B}^s$ of *available* sequences be such that $\llbracket R \rrbracket = \mathbb{B}^\omega \setminus \llbracket \cup_{\boldsymbol{z} \in \mathbb{B}^*} D_{s-1}(\boldsymbol{z})\rrbracket$. Collect in the auxiliary set A_s a number of available sequences $\boldsymbol{x}$ from R with $\sum_{\boldsymbol{x} \in A_s} \mu(\llbracket \boldsymbol{x} \rrbracket)$ maximal but bounded by $g(\boldsymbol{y}, t) - \mu(\llbracket D_{s-1}(\boldsymbol{y})\rrbracket)$, the amount of measure the current descriptions fall short of the latest approximation of $q(\boldsymbol{y})$. Put $T_s := T_{s-1} \cup \{(\boldsymbol{x}, \boldsymbol{y}) : \boldsymbol{x} \in A_s\}$.

Verification. It is immediate from the construction that $\cup_{\boldsymbol{y} \in \mathbb{B}^*} D_s(\boldsymbol{y})$ is prefix-free at all stages s, so $T = \lim_{s \to \infty} T_s$ is a prefix-free machine. To show that $\mu_T(\boldsymbol{y}) = \lim_{s \to \infty} \mu(\llbracket D_s(\boldsymbol{y})\rrbracket)$ equals $q(\boldsymbol{y})$ for all $\boldsymbol{y} \in \mathbb{B}^*$, it suffices to demonstrate that for every $\delta > 0$ there is some stage s_0 where $\mu(\llbracket D_{s_0}(\boldsymbol{y})\rrbracket) > q(\boldsymbol{y}) - \delta$.

Choose positive $\delta' < \delta$. Wait for a stage $s_0 = \langle \boldsymbol{y}, t_0 \rangle$ with $\mu([\![\boldsymbol{x}]\!]) \leq \delta - \delta'$ for all $\boldsymbol{x} \in \mathbb{B}^{s_0}$ and $g(\boldsymbol{y}, t_0) > q(\boldsymbol{y}) - \delta'$. Clearly, the available μ-measure

$$\begin{aligned} \mu([\![R]\!]) &= 1 - \sum_{\boldsymbol{z} \in \mathbb{B}^*} \mu([\![D_{s_0-1}(\boldsymbol{z})]\!]) \\ &\geq 1 - \mu([\![D_{s_0-1}(\boldsymbol{y})]\!]) - \sum_{\boldsymbol{z} \in \mathbb{B}^* \setminus \{\boldsymbol{y}\}} q(\boldsymbol{z}) \\ &\geq q(\boldsymbol{y}) - \mu([\![D_{s_0-1}(\boldsymbol{y})]\!]) \\ &\geq g(\boldsymbol{y}, t_0) - \mu([\![D_{s_0-1}(\boldsymbol{y})]\!]). \end{aligned}$$

Consequently, if not already $\mu([\![D_{s_0-1}(\boldsymbol{y})]\!]) > q(\boldsymbol{y}) - \delta$, then the construction collects in A_{s_0} a number of descriptions of length s_0 from R such that $\mu([\![D_{s_0}(\boldsymbol{y})]\!]) = \mu([\![D_{s_0-1}(\boldsymbol{y})]\!]) + \sum_{\boldsymbol{x} \in A_{s_0}} \mu([\![\boldsymbol{x}]\!]) > q(\boldsymbol{y}) - \delta$ as required. □

B.1.3. The proof of proposition 2.12. We adapt the construction for proposition 2.10 in B.1.1.3 above in such a way that the constructed machine M fails to be even weakly universal.

Proof. Let U be an arbitrary universal machine. We construct a machine M with $\mu_M = \nu$ such that for every constant $c \in \mathbb{N}$ there is a $\boldsymbol{y}$ such that for some $\boldsymbol{x}'$ with $(\boldsymbol{x}', \boldsymbol{y}) \in U$, we have that $|\boldsymbol{x}| > |\boldsymbol{x}'| + c$ for all $\boldsymbol{x}$ with $(\boldsymbol{x}, \boldsymbol{y}) \in M$. This ensures that M is not weakly universal.

Construction. The only change to the earlier construction is that at stage s we try to collect available sequences of length l_s, where l_s is defined as follows. Let $l_0 = 0$. For $s = \langle \boldsymbol{y}, t \rangle$ with $t > 0$, let $l_s = l_{s-1} + 1$. In case $s = \langle \boldsymbol{y}, 0 \rangle$, enumerate pairs in U until a pair $(\boldsymbol{x}', \boldsymbol{y})$ for some $\boldsymbol{x}'$ is found. Let $l_s := \max\{l_{s-1} + 1, |\boldsymbol{x}'| + s\}$.

Verification. The verification that $\mu_M = \nu$ proceeds as before. In addition, the construction guarantees that for every $c \in \mathbb{N}$, we have for $\boldsymbol{y}$ with $c = \langle \boldsymbol{y}, 0 \rangle$ that $|\boldsymbol{x}| > |\boldsymbol{x}'| + c$ for the first enumerated $\boldsymbol{x}'$ with $(\boldsymbol{x}', \boldsymbol{y}) \in U$ and all $\boldsymbol{x}$ with $(\boldsymbol{x}, \boldsymbol{y}) \in M$. □

B.1.4. The proof of theorem 2.16. The left-to-right direction is lemma 2 in Wood et al. (2013).

Proof of (⇒). Given universal U with associated encoding $\{\boldsymbol{x}_e\}_e$ of all monotone machines, we write out

$$\begin{aligned}\lambda_U(\boldsymbol{y}) &= \lambda(\{\boldsymbol{x} : \exists \boldsymbol{y}' \succcurlyeq \boldsymbol{y}((\boldsymbol{x},\boldsymbol{y}') \in U)\}) \\ &= \sum_e \lambda(\{\boldsymbol{x}_e\boldsymbol{x} : \exists \boldsymbol{y}' \succcurlyeq \boldsymbol{y}((\boldsymbol{x},\boldsymbol{y}') \in M_e)\}) \\ &= \sum_e \lambda(\boldsymbol{x}_e)\lambda(\{\boldsymbol{x} : \exists \boldsymbol{y}' \succcurlyeq \boldsymbol{y}((\boldsymbol{x},\boldsymbol{y}') \in M_e)\}) \\ &= \sum_e \lambda(\boldsymbol{x}_e)\lambda_{M_e}(\boldsymbol{y}).\end{aligned}$$

We know by proposition 2.5 above that the λ_{M_e} range over all elements in $\mathcal{M}$. Moreover, $w : e \mapsto \lambda(\boldsymbol{x}_e)$ is a weight function because $\{\boldsymbol{x}_e\}_e$ is prefix-free. That means that λ_U is a Σ_1 mixture. □

The right-to-left direction is lemma 4 in Wood et al. (2013). The following presentation streamlines their proof, to facilitate a generalized version for theorem 2.13 in B.1.5.2 below.

Proof of ($\Leftarrow$). Given Σ_1 mixture $\xi_v^{\{\nu_i\}_i}$ with (defective) weight function v. Let $\{\boldsymbol{x}_i\}_i$ be some effective prefix-free listing of finite sequences; then function $q : \mathbb{B}^\omega \to [0,1]$ defined by

$$q(\boldsymbol{y}) = \begin{cases} v(i) & \text{if } \boldsymbol{y} = \boldsymbol{x}_i \text{ for } i \in \mathbb{N} \\ 0 & \text{otherwise} \end{cases}$$

is a Σ_1 semi-distribution. By the proof of proposition A.3 above we can construct a prefix-free machine T that transforms λ into q: so $\lambda_T = q$, and

$$\sum_i v(i)\nu_i(\boldsymbol{y}) = \sum_i \lambda_T(\boldsymbol{x}_i)\nu_i(\boldsymbol{y}).$$

Denote by $n_i := \#\{\boldsymbol{z} : (\boldsymbol{z},\boldsymbol{x}_i) \in T\}$ the number of T-descriptions of $\boldsymbol{x}_i$, and let $\langle\cdot,\cdot\rangle : \mathbb{N}\times\mathbb{N} \to \mathbb{N}$ be a partial computable pairing function that maps the pairs (i,j) with $j < n_i$ onto $\mathbb{N}$. Let $\boldsymbol{z}_{\langle i,j\rangle}$ be the j-th enumerated T-description of $\boldsymbol{x}_i$. We then have

$$\sum_i \lambda_T(\boldsymbol{x}_i)\nu_i(\boldsymbol{y}) = \sum_i \sum_{j<n_i} \lambda(\boldsymbol{z}_{\langle i,j\rangle})\nu_i(\boldsymbol{y}).$$

Now for every $\langle i,j\rangle$ for which $\boldsymbol{z}_{\langle i,j\rangle}$ becomes defined we can run the construction of proposition 2.10 above on λ and ν_i to obtain a machine $M_{\langle i,j\rangle}$. In this way we obtain an enumeration of machines $\{M_e\}_e$ such that $\lambda_{M_{\langle i,j\rangle}} = \nu_i$ (with $j < n_i$) for all i, resulting in

$$\begin{aligned}\sum_i \sum_{j<n_i} \lambda(\boldsymbol{z}_{\langle i,j\rangle})\nu_i(\boldsymbol{y}) &= \sum_e \lambda(\boldsymbol{z}_e)\lambda_{M_e}(\boldsymbol{y}) \\ &= \lambda_U(\boldsymbol{y}),\end{aligned}$$

where we define U by $(\boldsymbol{z}_e\boldsymbol{x},\boldsymbol{y}) \in U :\Leftrightarrow (\boldsymbol{x},\boldsymbol{y}) \in M_e$.

It does still remain to verify that U is universal. Namely, we cannot take for granted that $\{M_e\}_e$ is an enumeration of *all* machines, whence it is not clear that U is universal. (This issue is overlooked in the original proof by Wood et al.) Note that it is enough if there were a single universal machine U' in $\{M_e\}_e$, but even that is not obvious (by proposition 2.12 we know that there are for any universal U *non*-universal M such that $\lambda_M = \lambda_U$). However, there is a simple patch to the enumeration that guarantees this fact. Namely, given an arbitrary universal machine U', we may simply put $M_e := U'$ at some $e = \langle i, j \rangle$ where it so happens that $\lambda_{U'} = \lambda_{M_e}$ (this is finite information so the existence of such e implies the existence of the patched enumeration). □

B.1.5. The proof of proposition 2.11. We generalize the left-to-right direction of the proof of theorem 2.16, that is given in B.1.4 above. For this we require a lemma that is a refined version of proposition 2.10.[30]

B.1.5.1. *A fixed-point lemma.* Write $\mu^{\boldsymbol{x}}(\cdot) := \mu(\cdot \mid \boldsymbol{x})$ for measure μ conditional on $\boldsymbol{x} \in \mathbb{B}^*$. Let $\mu^{\boldsymbol{x}}_M$ denote the transformation of $\mu^{\boldsymbol{x}}$ by M.

Lemma B.1. Given effective enumeration $\{M_e\}_e$ of the monotone machines with computable prefix-free encoding $\{\boldsymbol{x}_e\}_{e \in \mathbb{N}}$. For every continuous Δ_1 measure μ with $\mu(\boldsymbol{x}_e) > 0$ for every e,

$$\{\mu^{\boldsymbol{x}_e}_{M_e}\}_e = \mathcal{M}.$$

Proof. Let ν be any Σ_1 measure. Since $\mu^{\boldsymbol{x}_e}$ is obviously a continuous Δ_1 measure for every $e \in \mathbb{N}$, by the construction in B.1.1.3 above we obtain for every e a monotone machine M with $\nu = \mu^{\boldsymbol{x}_e}_M$. Indeed, there is a total computable function $g : \mathbb{N} \to \mathbb{N}$ that for given e retrieves an index $g(e)$ in the given enumeration $\{M_e\}_{e \in \mathbb{N}}$ such that $\nu = \mu^{\boldsymbol{x}_e}_{M_{g(e)}}$. But by Kleene's recursion theorem (see Soare, 2016, 29), there must be a fixed point $\hat{e}$ such that $M_{g(\hat{e})} = M_{\hat{e}}$, hence $\mu^{\boldsymbol{x}_{\hat{e}}}_{M_{\hat{e}}} = \mu^{\boldsymbol{x}_{\hat{e}}}_{M_{g(\hat{e})}}$.

This shows that for every ν there is an index e such that $\nu = \mu^{\boldsymbol{x}_e}_{M_e}$. Conversely, by the proof in B.1.1 above the function $\mu^{\boldsymbol{x}_e}_{M_e}$ gives a Σ_1 measure for every e. □

B.1.5.2. *The proof of proposition 2.11.* We can show that a universal transformation μ_U of a continuous Δ_1 measure is a Solomonoff-Levin measure, as long as universal monotone U has an associated encoding $\{\boldsymbol{x}_e\}_e$ that does not receive measure 0 from μ: so $\mu(\boldsymbol{x}_e) > 0$ for every e. Call (the associated encodings of) such machines *compatible* with μ. (This is clearly no restriction for those μ that give positive measure to all sequences, which includes all non-deterministic i.i.d. measures.)

Proof. Given continuous Δ_1 measure μ and μ-compatible universal machine U. We write out

$$\begin{aligned}\mu_U(\boldsymbol{y}) &= \mu(\{\boldsymbol{x} : \exists \boldsymbol{y}' \succcurlyeq \boldsymbol{y}((\boldsymbol{x},\boldsymbol{y}') \in U)\}) \\ &= \sum_e \mu(\{\boldsymbol{x}_e\boldsymbol{x} : \exists \boldsymbol{y}' \succcurlyeq \boldsymbol{y}((\boldsymbol{x},\boldsymbol{y}') \in M_e)\}) \\ &= \sum_e \mu(\boldsymbol{x}_e)\mu(\{\boldsymbol{x} : \exists \boldsymbol{y}' \succcurlyeq \boldsymbol{y}((\boldsymbol{x},\boldsymbol{y}') \in M_e)\} \mid \boldsymbol{x}_e) \\ &= \sum_e \mu(\boldsymbol{x}_e)\mu^{\boldsymbol{x}_e}_{M_e}(\boldsymbol{y}).\end{aligned}$$

Our lemma B.1 tells us that the $\mu^{\boldsymbol{x}_e}_{M_e}$ range over all elements in $\mathcal{M}$. Moreover, $w : e \mapsto \mu(\boldsymbol{x}_e)$ is a (defective) weight function because $\{\boldsymbol{x}_e\}_e$ is prefix-free and U is compatible with μ. So μ_U is a Σ_1 mixture, hence, by theorem 2.16, a Solomonoff-Levin measure. □

B.1.6. The proof of theorem 2.13. We show that for every two continuous Δ_1 measures μ and $\bar{\mu}$, for any universal monotone machine U that is μ-compatible, there is a universal monotone machine V such that $\mu_U = \bar{\mu}_V$. This implies (since λ is itself a continuous Δ_1 measure) that every Solomonoff-Levin measure can be written as a universal transformation μ_U of any continuous Δ_1 measure μ. The proof is a generalization of that of theorem 2.16 in B.1.4 above.

Proof. Given continuous computable μ and $\bar{\mu}$, and universal U compatible with μ. Write out as in B.1.5.2 above

$$\mu_U(\boldsymbol{y}) = \sum_e \mu(\boldsymbol{x}_e)\mu^{\boldsymbol{x}_e}_{M_e}(\boldsymbol{y}).$$

The function

$$q(\boldsymbol{y}) = \begin{cases}\mu(\boldsymbol{y}) & \text{if } \boldsymbol{y} = \boldsymbol{x}_e \text{ for } e \in \mathbb{N} \\ 0 & \text{otherwise}\end{cases}$$

is a Σ_1 semi-distribution; hence by proposition A.3 we can construct a prefix-free machine T such that $q = \bar{\mu}_T$. Let $n_e := \#\{\boldsymbol{z} : (\boldsymbol{z},\boldsymbol{x}_e) \in T\}$, and let p.c. $\langle\cdot,\cdot\rangle : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$ map the pairs (e,j) with $j < n_e$ onto $\mathbb{N}$. Let $\boldsymbol{z}_{\langle e,j\rangle}$ be the j-th enumerated T-description of $\boldsymbol{x}_e$. We then have

$$\begin{aligned}\sum_e \mu(\boldsymbol{x}_e)\mu^{\boldsymbol{x}_e}_{M_e}(\boldsymbol{y}) &= \sum_e \bar{\mu}_T(\boldsymbol{x}_e)\mu^{\boldsymbol{x}_e}_{M_e}(\boldsymbol{y}) \\ &= \sum_e \sum_{j<n_e} \bar{\mu}(\boldsymbol{z}_{\langle e,j\rangle})\mu^{\boldsymbol{x}_e}_{M_e}(\boldsymbol{y}).\end{aligned}$$

For every $\langle e,j\rangle$ for which $\boldsymbol{z}_{\langle e,j\rangle}$ becomes defined we run the construction of proposition 2.10 above on $\bar{\mu}^{\boldsymbol{z}_{\langle e,j\rangle}}$ and $\mu^{\boldsymbol{x}_e}_{M_e}$ to obtain an enumeration of machines $\{N_d\}_d$ such that $\bar{\mu}^{\boldsymbol{z}_{\langle e,j\rangle}}_{N_{\langle e,j\rangle}} = \mu^{\boldsymbol{x}_e}_{M_e}$ (with $j < n_e$) for all e. Then

$$\sum_{e}\sum_{j<n_e}\bar{\mu}(\boldsymbol{z}_{\langle e,j\rangle})\mu^{\boldsymbol{x}_e}_{M_e}(\boldsymbol{y})=\sum_{d}\bar{\mu}(\boldsymbol{z}_d)\bar{\mu}^{\boldsymbol{z}_d}_{N_d}(\boldsymbol{y}),$$

which we can rewrite to $\bar{\mu}_V(\boldsymbol{y})$, defining V by $(\boldsymbol{z}_d\boldsymbol{x},\boldsymbol{y})\in V :\Leftrightarrow (\boldsymbol{x},\boldsymbol{y})\in N_d$.

We then still need to enforce that V is in fact universal by putting $N_d := V'$ for some universal V' and some $d=\langle e,j\rangle$ with $\bar{\mu}^{\boldsymbol{z}_{\langle e,j\rangle}}_{V'}=\mu^{\boldsymbol{x}_e}_{M_e}$. Our final objective is thus to show that $\bar{\mu}^{\boldsymbol{z}_{\langle e,j\rangle}}_{V'}=\mu^{\boldsymbol{x}_e}_{M_e}$ for some e,j. To that end, define computable function $g:\mathbb{N}\to\mathbb{N}$ by $\mu^{\boldsymbol{x}_e}_{M_{g(e)}}=\bar{\mu}^{\boldsymbol{z}_{\langle e,0\rangle}}_{V'}$. Since $\bar{\mu}_T(\boldsymbol{x}_e)>0$ for each e, the sequence $\boldsymbol{z}_{\langle e,0\rangle}$ is defined for each e. Hence $\bar{\mu}^{\boldsymbol{z}_{\langle e,0\rangle}}_{V'}$ is defined, and g, that retrieves the index $g(e)$ of a machine that transforms $\mu^{\boldsymbol{x}_e}$ to this Σ_1 measure, is total. Then by the recursion theorem there is an index $\hat{e}$ such that $M_{\hat{e}}=M_{g(\hat{e})}$, so $\mu^{\boldsymbol{x}_{\hat{e}}}_{M_{\hat{e}}}=\mu^{\boldsymbol{x}_{\hat{e}}}_{M_{g(\hat{e})}}=\bar{\mu}^{\boldsymbol{z}_{\langle \hat{e},0\rangle}}_{V'}$. □

B.1.7. The proof of theorem A.16. The generalized coding theorem follows from a generalization of the construction for the original coding theorem (see Li and Vitányi, 2008, 273ff; Shen et al., 20xx), together with lemma A.15.

B.1.7.1. *The main construction*[31]. This is

Proposition B.2. For every computable measure μ that is conditionally bounded away from 0, there is for every Σ_1 semi-distribution p a prefix-free machine T such that $K^{\mu}_T \leq^+ -\log p$.

Proof. Let p be any Σ_1 semi-distribution, with uniformly computable approximation function g. Let μ be conditionally bounded away from 0 by d. We will construct a prefix-free machine T such that $K^{\mu}_T \leq^+ -\log p$ in stages $s=\langle \boldsymbol{y},t\rangle$. Let $D_s(\boldsymbol{y})=\{\boldsymbol{x}\in\mathbb{B}^*:(\boldsymbol{x},\boldsymbol{y})\in T_s\}$.

Construction. Let $T_0=\emptyset$, and also $R_0=\emptyset$.

At stage $s=\langle \boldsymbol{y},t\rangle$, let $k\in\mathbb{N}$ be such that $2^{-k}\leq g(\boldsymbol{y},t)<2^{-k+1}$. If $T_{\langle \boldsymbol{y},r\rangle}=T_{\langle \boldsymbol{y},r\rangle-1}$ for those $r<t$ where $g(\boldsymbol{y},r)\geq 2^{-k}$ already (in particular, if $r=t$ is the first r where $g(\boldsymbol{y},r)\geq 2^{-k}$), then we proceed as follows.

Let $\boldsymbol{z}_0$ be the left-most sequence of length s with $[\![\boldsymbol{z}_0]\!]\cap[\![\cup_{s'<s}R_{s'}]\!]=\emptyset$, i.e., $[\![\boldsymbol{z}]\!]\subseteq[\![\cup_{s'<s}R_{s'}]\!]$ for all $\boldsymbol{z}\in\mathbb{B}^s$ with $\boldsymbol{z}<_L\boldsymbol{z}_0$. Starting with this $\boldsymbol{z}_0$, construct set R_s by iteratively adding the next left-most sequence $\boldsymbol{z}_i>_L\boldsymbol{z}_{i-1}$ of length s, until $\sum_{\boldsymbol{z}\in R_s}\mu(\boldsymbol{z})\geq 2^{-k-2}$. If already $\sum_{\boldsymbol{z}\in R_s}\mu(\boldsymbol{z})\geq 2^{-k-1}$, then reset $R_s:=\emptyset$, put $T_s:=T_{s-1}$, and proceed to next stage. Otherwise, select a string $\boldsymbol{x}$ of maximal measure $\mu(\boldsymbol{x})$ that satisfies $[\![\boldsymbol{x}]\!]\subseteq[\![R_s]\!]$, and put $T_s:=T_{s-1}\cup\{(\boldsymbol{x},\boldsymbol{y})\}$.

Verification. For given $\boldsymbol{y}$, let $k\in\mathbb{N}$ minimal such that $p(\boldsymbol{y})>2^{-k}$; so $p(\boldsymbol{y})\leq 2^{-k+1}$. We will show that the construction enumerates a pair $(\boldsymbol{x},\boldsymbol{y})$

with $\mu(\boldsymbol{x}) \geq 2^{-k-d-3}$ in T. This guarantees that, for all $\boldsymbol{y} \in \mathbb{B}^*$,

$$\begin{aligned} K^\mu_T(\boldsymbol{y}) &= -\log \max\{\mu(\boldsymbol{x}) : (\rho, \boldsymbol{y}) \in T\} \\ &\leq -\log(2^{-d-3} p(\boldsymbol{y})) \\ &= -\log p(\boldsymbol{y}) + d + 3. \end{aligned}$$

Let $s_0 = \langle \boldsymbol{y}, t_0 \rangle$ be large enough to satisfy $g(\boldsymbol{y}, t_0) \geq 2^{-k}$ and, by nonatomicity of μ, $\mu(\boldsymbol{z}) < 2^{-k-2}$ for all $\boldsymbol{z} \in \mathbb{B}^s$. If still $D_{\langle \boldsymbol{y}, r\rangle}(\boldsymbol{y}) = D_{\langle \boldsymbol{y}, r\rangle - 1}(\boldsymbol{y})$ for those $r < t_0$ where already $g(\boldsymbol{y}, r) \geq 2^{-k}$ (otherwise we are done), then from the fact that $\sum_{l : 2^{-l} \leq p(\boldsymbol{z})} 2^{-l} \leq 2 \cdot p(\boldsymbol{z})$, we also have that

$$\begin{aligned} \sum_{s < s_0} \mu([\![R_s]\!]) &< \left(\sum_{\boldsymbol{z} \in \mathbb{B}^*} \sum_{l : 2^{-l} \leq p(\boldsymbol{z})} 2^{-l-1} \right) - 2^{-k-1} \\ &\leq \left(\sum_{\boldsymbol{z} \in \mathbb{B}^*} p(\boldsymbol{z}) \right) - 2^{-k-1} \\ &\leq 1 - 2^{-k-1}. \end{aligned}$$

Thus the construction has room to select in R_{s_0} a number of sequences from $\mathbb{B}^{s_0}$ such that $2^{-k-2} \leq \sum_{\boldsymbol{z} \in R_{s_0}} \mu(\boldsymbol{z}) \leq 2^{-k-1}$. All that remains to show is that there is $\boldsymbol{x}$ with $[\![\boldsymbol{x}]\!] \subseteq [\![R_{s_0}]\!]$ such that $\mu(\boldsymbol{x}) \geq 2^{-k-d-3}$.

To that end, consider the sequence $\boldsymbol{x}_0$ with $[\![\boldsymbol{x}_0]\!] \subseteq [\![R_{s_0}]\!]$ of smallest length. Suppose without loss of generality that $\boldsymbol{x}_0 = \boldsymbol{x}_0^- 0$. Then for $\boldsymbol{x}_0' := \boldsymbol{x}_0^- 1$ with possibly $[\![\boldsymbol{x}_0']\!] \cap [\![R_{s_0}]\!] \neq \emptyset$ (but not $[\![\boldsymbol{x}_0']\!] \subseteq [\![R_{s_0}]\!]$ or $\boldsymbol{x}_0^-$ would be shortest) we have that $\mu(\boldsymbol{x}_0) \geq 2^{-d} \mu(\boldsymbol{x}_0')$. Moreover, let $\boldsymbol{x}_1$ the shortest sequence with $\boldsymbol{x}_1 <_L \boldsymbol{x}_0$ and $[\![\boldsymbol{x}_1]\!] \subseteq [\![R_{s_0}]\!]$ (if such sequence exists). Again, for $\boldsymbol{x}_1' := \rho_1^- 0$ with possibly $[\![\rho_1']\!] \cap [\![R_{s_0}]\!] \neq \emptyset$ but not $[\![\boldsymbol{x}_1']\!] \subseteq [\![R_{s_0}]\!]$ we have that $\mu(\boldsymbol{x}_0) \geq 2^{-d} \mu(\boldsymbol{x}_1')$. It follows that for $\boldsymbol{x}$ equal to $\boldsymbol{x}_0$ or to $\boldsymbol{x}_1$ it must hold that $\mu(\boldsymbol{x}) \geq 2^{-d-1} \mu([\![R_{s_0}]\!])$, and we are done. □

B.1.7.2. *The proof of lemma A.15.* It does not follow immediately from proposition B.2 that for μ conditionally bounded away from 0 we have $K^\mu_U \leq^+ -\log p$ for every universal prefix-free machine U and Σ_1 semi-distribution p. While we have that for any universal U and all $e \in \mathbb{N}$,

$$\begin{aligned} K^\mu_U(\boldsymbol{y}) &= -\log \max\{\mu(\boldsymbol{x}) : (\boldsymbol{x}, \boldsymbol{y}) \in U\} \\ &\leq -\log \mu(\boldsymbol{x}_e) - \log \max\{\mu(\boldsymbol{x} \mid \boldsymbol{x}_e) : (\boldsymbol{x}, \boldsymbol{y}) \in T_e\} \qquad (92) \\ &= K^{\mu^e}_{T_e}(\boldsymbol{y}) + O(1), \end{aligned}$$

we have to do a little more work to show that this implies that also $K^\mu_U \leq^+ K^\mu_{T_e}$.

Proof. For given prefix-free T, we can employ the construction of proposition B.2 to define a computable $g : \mathbb{N} \to \mathbb{N}$ such that $K^{\mu^e}_{T_{g(e)}} \leq^+ K^\mu_T$. Now by the Kleene recursion theorem there is a fixed point $\hat{e}$ such that $T_{g(\hat{e})} = T_{\hat{e}}$, so we have $K^{\mu^{\hat{e}}}_{T_{\hat{e}}} \leq^{+} K^{\mu}_{T}$, and by (92) also $K^{\mu}_{U} \leq^{+} K^{\mu^{\hat{e}}}_{T_{\hat{e}}}$, hence $K^{\mu}_{U} \leq^{+} K^{\mu}_{T}$. □

B.1.7.3. *The proof of theorem A.16.*

Proof. By proposition B.2 we can exhibit a prefix-free T such that $K^{\mu}_{T} \leq^{+} -\log \mu_U$. Moreover, by lemma A.15 we have $K^{\mu}_{U} \leq^{+} K^{\mu}_{T}$ and by (84) also $-\log \mu_U \leq^{+} K^{\mu}_{U}$, hence $-\log \mu_U =^{+} K^{\mu}_{U}$ as required. □

B.1.8. The proof of proposition 2.17. Call an enumeration $\{\nu_i\}_{i \in \mathbb{N}}$ of all Σ_1 measures *acceptable* if it is generated from an enumeration $\{M_i\}_i$ of all monotone Turing machines by the procedure in B.1.1.3 above, i.e., $\nu_i = \lambda_{M_i}$. This terminology matches that of the definition of *acceptable numberings* of the p.c. functions (Rogers, 1967, 41; Soare, 2016, 21). Every effective listing of all Turing machines yields an acceptable numbering. Importantly, any two acceptable numberings differ only by a computable permutation (Rogers, 1958); in our case, for any two acceptable enumerations $\{\nu_i\}_i$ and $\{\bar{\nu}_i\}_i$ there is a computable permutation $f : \mathbb{N} \to \mathbb{N}$ of indices such that $\bar{\nu}_i = \nu_{f(i)}$.

Proof. Given $\lambda_U \in \mathcal{SL}$, with enumeration $\{M_i\}_i$ of all monotone machines corresponding to U. We know that λ_U is equal to $\xi^{\{\bar{\nu}_i\}_i}_{v}$ for some acceptable enumeration $\{\bar{\nu}_i\}_i = \{\lambda_{M_i}\}_i$ of $\mathcal{M}$ and (defective) weight function v. First we show that $\xi^{\{\bar{\nu}_i\}_i}_{v}$ is equal to $\xi^{\{\nu_i\}_i}_{v'}$ for given acceptable enumeration $\{\nu_i\}_i$ and (defective) weight function v'; then we show that it is also equal to $\xi^{\{\nu_i\}_i}_{w}$ for proper weight function w.

Since enumerations $\{\nu_i\}_i$ and $\{\bar{\nu}_e\}_e$ are both acceptable, there is a 1-1 computable f such that $\bar{\nu}_i = \nu_{f(i)}$. Then

$$\begin{aligned}\sum_i v(i)\bar{\nu}_i(\cdot) &= \sum_i v(i)\nu_{f(i)}(\cdot) \\ &= \sum_i v(f^{-1}(i))\nu_i(\cdot) \\ &= \sum_i v'(i)\nu_i(\cdot),\end{aligned}$$

with $v' : i \mapsto v(f^{-1}(i))$.

We proceed with the description of a proper weight function w. The idea is to have w assign to each i a positive computable weight that does not exceed $v'(i)$, additional computable weight to the index of a single suitably defined Σ_1 measure in order to regain the original mixture, and all of the remaining weight to an "empty" Σ_1 measure.

Let $a \in \mathbb{Q}$ be such that $\xi^{\{\nu_i\}_i}_{v'}(\varnothing) < a < 1$, and let c be such that $\sum_i 2^{-i-c} < 1 - a$. Let $v'_0(i)$ denote the first approximation of semi-computable

$v'(i)$ that is positive. We now define computable $g : \mathbb{N} \to \mathbb{Q}$ by

$$g(i) = \min\{2^{-i-c}, v'_0(i)\}.$$

Clearly, $\sum_i g(i) < 1 - a$. Moreover, $\sum_i g(i)$ is computable because for any $\delta > 0$ we have a $j \in \mathbb{N}$ with $\sum_{i>j} 2^{-i-c} < \delta$, hence $\sum_{i \leq j} g(i) < \sum_i g(i) < \sum_{i \leq j} g(i) + \delta$.

Next, define $\pi(\cdot) := a^{-1} \sum_i (v'(i) - g(i))\, \nu_i(\cdot)$. This is a semi-measure because $\pi(\varnothing) \leq a^{-1} \xi_{v'}^{\{\nu_i\}_i}(\varnothing) < a^{-1}a = 1$. Let k be such that $\nu_k = \pi$, and let l be such that ν_l is the "empty" Σ_1 measure with $\nu(\boldsymbol{x}) = 0$ for all $\boldsymbol{x} \in \mathbb{B}^*$ (both indices exist even if we cannot effectively find them).

Finally, we define w by

$$w(i) = \begin{cases} g(i) & \text{if } i \neq k, l \\ g(i) + a & \text{if } i = k \\ 1 - a - \sum_{j \neq l} g(j) & \text{if } i = l \end{cases}.$$

Function w is computable and indeed a proper weight function, and

$$\begin{aligned} \sum_i w(i)\nu_i(\cdot) &= \sum_i g(i)\nu_i(\cdot) + a\nu_k(\cdot) + 0 \\ &= \sum_i g(i)\nu_i(\cdot) + \sum_i (v'(i) - g(i))\, \nu_i(\cdot) \\ &= \sum_i v'(i)\nu_i(\cdot). \end{aligned}$$ □

B.1.9. The proof of proposition 2.18.

Proof. By proposition 2.17 we know that any given element in $\mathcal{SL}$ equals $\xi_w^{\{\nu_i\}_i}(\cdot)$ for some computable weight function w over given $\{\nu_i\}_i$. Let k be such that $\nu_k(\cdot) = \sum_i 2^{-K(i)}\nu_i(\cdot)$, with $K(i)$ the prefix-free Kolmogorov complexity (via some universal prefix-free machine $\mathring{T}$) of the i-th lexicographically ordered string; $2^{-K(\cdot)}$ is a universal weight function. Define

$$\mathring{v}(i) = \begin{cases} w(i) + w(k) \cdot 2^{-K(i)} & \text{if } i \neq k \\ w(k) \cdot 2^{-K(i)} & \text{if } i = k \end{cases},$$

which is a weight function because $\sum_i \mathring{v}(i) < \sum_{i \neq k} w(i) + w(k) = \sum_i w(i)$. Moreover, $\mathring{v}$ is universal because $2^{-K(\cdot)}$ is, and

$$\begin{aligned} \sum_i \mathring{v}(i)\nu_i(\cdot) &= \sum_{i \neq k} w(i)\nu_i(\cdot) + w(k) \sum_i 2^{-K(i)}\nu_i(\cdot) \\ &= \sum_i w(i)\nu_i(\cdot). \end{aligned}$$ □

* * *

B.2. Sequential prediction

B.2.1. The proof of convergence theorem 2.9. In the main text I presented this result as a direct corollory of the Blackwell-Dubins theorem (1962). However, in the literature (Li and Vitányi, 2008, 352ff; Hutter, 2003a, 2062; Poland and Hutter, 2005, 3781) it is usually presented as a consequence of (variations of) the following stronger result, first shown by Solomonoff (1978, 426f; already presented in 1975).

Let us introduce as a measure of the divergence between two distributions p_1 and p_2 over $\mathbb{B}$ the squared *Hellinger distance*

$$H(p_1, p_2) := \sum_{x \in \mathbb{B}} \left(\sqrt{p_1(x)} - \sqrt{p_2(x)} \right)^2 . \tag{93}$$

Then:

THEOREM B.3 (Solomonoff). *For every $\mu \in \Delta_1$, the expected infinite sum of squared Hellinger distances between one-step conditional Q_U and μ*

$$\mathbf{E}_{X^\omega \sim \mu} \left[\sum_{t=0}^{\infty} H\left(\mu^1(\cdot \mid X^t), Q_U^1(\cdot \mid X^t) \right) \right] \tag{94}$$

is bounded by a constant.

PROOF OF THEOREM 2.9. To see how theorem 2.9 follows from theorem B.3, 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 $|\mu(x_{t+1} \mid \boldsymbol{x}^t) - Q_U(x_{t+1} \mid \boldsymbol{x}^t)| > \delta$ infinitely often. But that means that with positive probability the infinite sum of squared Hellinger distances is infinite, and the expectation (94) cannot be bounded by a constant. □

B.2.1.1. *Constant bound on log-regret.* Theorem B.3, in turn, rests on the following simple but crucial consequence of Q_U's universality.

PROPOSITION B.4. For every $\mu \in \Delta_1$, the cumulative log-regret of p_{Q_U} relative to p_μ is bounded by a constant.

PROOF. Recall from 3.3.2.5 or 6.1.2 above that the cumulative log-regret of p_{Q_U} relative to p_μ is given by

$$R_{Q_U,\mu}(\boldsymbol{x}^s) = -\ln \frac{Q_U(\boldsymbol{x}^s)}{\mu(\boldsymbol{x}^s)} .$$

By the universality of Q_U in the class of Σ_1 measures we know that Q_U dominates μ: there is a constant c such that for every finite $\boldsymbol{x}$ we have $Q_U(\boldsymbol{x}) \geq$

$c^{-1}\mu(\boldsymbol{x})$. This fact allows us to derive that for every single sequence $\boldsymbol{x}$

$$\begin{aligned} R_{Q_U,\mu}(\boldsymbol{x}) &= -\ln \frac{Q_U(\boldsymbol{x})}{\mu(\boldsymbol{x})} \\ &\leq -\ln \frac{c^{-1}\mu(\boldsymbol{x})}{\mu(\boldsymbol{x})} \\ &= \ln c. \end{aligned}$$

□

B.2.1.2. **Constant bound on log-regret: Σ_1 measures.* Proposition B.4 actually directly generalizes to the Σ_1 measures on $\mathbb{B}^\omega \cup \mathbb{B}^*$, or *semi-measures.*

Proposition B.5. For every $\nu \in \Sigma_1$, the cumulative log-regret of p_{Q_U} relative to p_ν is bounded by a constant.

Proof. As in proposition B.4, because all that is required is the dominance of Q_U over ν. □

B.2.1.3. *Squared Hellinger distance and Kullback-Leibler divergence.* Another lemma that is needed for the proof of theorem B.3 is the fact that the squared Hellinger distance is bounded by the Kullback-Leibler divergence, which, recall (A.2.3.2 above), is defined by

$$D(p_1 \parallel p_2) = \mathbf{E}_{X\sim p_1}\left[-\ln \frac{p_2(X)}{p_1(X)}\right], \tag{95}$$

i.e., as the p_1-expected log-regret of p_2 relative to p_1.

Lemma B.6. For all non-zero semi-distributions p_1 and p_2,

$$H(p_1, p_2) \leq D(p_1 \parallel p_2).$$

Proof. We derive

$$
\begin{aligned}
H(p_1, p_2) &= \sum_{x\in\mathbb{B}} \left(\sqrt{p_1(x)} - \sqrt{p_2(x)}\right)^2 \\
&= \sum_{x\in\mathbb{B}} \left(p_1(x) + p_2(x) - 2\sqrt{p_1(x)p_2(x)}\right) \\
&\leq 2 - 2\sum_{x\in\mathbb{B}} \left(p_1(x)\sqrt{\frac{p_2(x)}{p_1(x)}}\right) \qquad (96) \\
&= -2\left(\mathbf{E}_{X\sim p_1}\left[\sqrt{\frac{p_2(X)}{p_1(X)}}\right] - 1\right) \\
&\leq -2\ln \mathbf{E}_{X\sim p_1}\left[\sqrt{\frac{p_2(X)}{p_1(X)}}\right] \qquad (97) \\
&\leq \mathbf{E}_{X\sim p_1}\left[-2\ln\sqrt{\frac{p_2(X)}{p_1(X)}}\right] \qquad (98) \\
&\leq \mathbf{E}_{X\sim p_1}\left[-\ln\frac{p_2(X)}{p_1(X)}\right].
\end{aligned}
$$

Here (96) follows from the semi-distribution property that $p(0) + p(1) \leq 1$; (97) follows from the inequality $\ln y \leq y - 1$ for $y > 0$; and (98) follows from Jensen's inequality. □

B.2.1.4. *The proof of theorem B.3.* The result now follows easily from the previous.

Proof of theorem B.3. We use lemma B.6 to bound every expected *finite* sum of the first s divergences between one-step conditional Q_U and μ,

$$
\mathbf{E}_{X^{s-1}\sim\mu}\left[\sum_{t=0}^{s-1} H\left(\mu^1(\cdot \mid X^t), Q_U^1(\cdot \mid X^t)\right)\right], \qquad (99)
$$

by

$$
\mathbf{E}_{X^{s-1}\sim\mu}\left[\sum_{t=0}^{s-1} \mathbf{E}_{X_{t+1}\sim\mu(\cdot|X^t)}\left[-\ln\frac{Q_U(X_{t+1} \mid X^t)}{\mu(X_{t+1} \mid X^t)}\right]\right]. \qquad (100)
$$

Moving around the sums and the expectations, we work out that (100) equals

$$\sum_{t=0}^{s-1} \mathbf{E}_{X^{s-1}\sim\mu} \left[\mathbf{E}_{X_{t+1}\sim\mu(\cdot|X^t)} \left[-\ln \frac{Q_U(X_{t+1} \mid X^t)}{\mu(X_{t+1} \mid X^t)} \right] \right]$$

$$= \sum_{t=0}^{s-1} \mathbf{E}_{X^{t}\sim\mu} \left[\mathbf{E}_{X_{t+1}\sim\mu(\cdot|X^t)} \left[-\ln \frac{Q_U(X_{n+1} \mid X^t)}{\mu(X_{t+1} \mid X^t)} \right] \right] \tag{101}$$

$$= \sum_{t=0}^{s-1} \mathbf{E}_{X^{t+1}\sim\mu} \left[-\ln \frac{Q_U(X_{t+1} \mid X^t)}{\mu(X_{t+1} \mid X^t)} \right]$$

$$= \mathbf{E}_{X^{s}\sim\mu} \left[\sum_{t=0}^{s-1} -\ln \frac{Q_U(X_{t+1} \mid X^t)}{\mu(X_{t+1} \mid X^t)} \right], \tag{102}$$

where step (101) follows from the fact that for $s \geq t$ measure μ satisfies $\mu(\boldsymbol{x}^t) = \sum \mu(\boldsymbol{x}^s)[\![\boldsymbol{x}^s \in \mathbb{B}^s : \boldsymbol{x}^s \succcurlyeq \boldsymbol{x}^t]\!]$. But (102) is the expected log-regret

$$\mathbf{E}_{X^s\sim\mu}\left[R_{Q_U,\mu}(\boldsymbol{x}^s)\right], \tag{103}$$

and since by proposition B.4 we know that for some constant c we have *for every sequence* $\boldsymbol{x}^s$ *of any length* s that

$$R_{Q_U,\mu}(\boldsymbol{x}^s) < c,$$

we also have the constant bound c on (103) and hence on (94). □

B.2.2. The proof of consistency theorem 3.1. Since Solomonoff's proof of theorem B.3 in B.2.1 above only depends on Q_U's property of dominance over the relevant class of measures, we can similarly derive the parallel result for the mixture predictors $\mathsf{p}_{\text{mix}(w)}$ in general.

B.2.2.1. *Constant bound on log-regret: optimality.* In particular, proposition B.4 holds for mixtures in general, and the constant that bounds a mixture predictor's regret relative to p_i is indeed given by the weight $w(i)$. That is, for every $\mathsf{p}_{\text{mix}(w)}$ over $\mathcal{H}$ and $\mu_i \in \mathcal{H}$, we have that $R_{\text{mix}(w),i} \leq w(i)$. I actually stated this fact as optimality theorem 3.3 above, and gave the proof in the main text.

B.2.2.2. *Bound on expected Hellinger distance.* From optimality theorem 3.3 we can as before derive the bound on the Hellinger distance.

THEOREM B.7. *For every* $\mu \in \mathcal{H}$*, the expected infinite sum of divergences between the predictions given by* $\mathsf{p}_{\text{mix}(w)}$ *and the one-step conditional probabilities given by* μ

$$\mathbf{E}_{X^\omega\sim\mu}\left[\sum_{t=0}^{\infty} H\left(\mu^1(\cdot \mid X^t), \mathsf{p}_{\text{mix}}(X^t)\right)\right] \tag{104}$$

is bounded by a constant.

This implies as before the Bayesian consistency theorem 3.1.

B.2.2.3. *Reinterpretation: anticipation of convergence.* As mentioned in 3.3.2.1, we might reinterpret consistency theorem 3.1 for aggregating mixtures as follows: for every such $\mathsf{p}_{\mathrm{mix}(w)}$, every p_i in the pool $\mathcal{H}$, p_i anticipates with almost-certainty (per p_i's a priori measure μ_i) that $\mathsf{p}_{\mathrm{mix}(w)}$'s predictions converge to p_i's own. In particular, two different aggregating predictors both anticipate with probability 1 that their predictions converge to each other. This is in fact close in spirit to the original statement of the Blackwell-Dubins theorem as a result about the *merging of opinions* of two Bayesian agents that are mutually continuous with respect to each other (also see Huttegger, 2015). (But note that this statement—merging of the full conditional measures—is stronger than what I discuss here: convergence of predictions or one-step conditional measures.) It is of interest to state that two Solomonoff-Levin predictors both expect with almost-certainty that their predictions will converge to each other (given the fact, theorem 3.2, that they do not necessarily do converge). However, it is not perfectly clear how to make sense of almost-certainty in the case of measures on $\mathbb{B}^\omega \cup \mathbb{B}^*$ or *semi-measures.*

B.2.2.4. *Convergence to semi-measures.* To make sense of the Bayesian consistency result applied to Σ_1 mixtures (both in the original interpretation and that of B.2.2.3 above), we need to make precise what 'almost surely' should mean for such 'semi-measures.' We might try to 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 $\boldsymbol{x} \in \mathbb{B}^*$. 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. But one can object to this way of setting things up: in particular the 'convergence' on appearence of an s, where both the measure and the predictor jump to forever following the deterministic sequence s^ω, is quite trivial. In case of a semi-measure ν with probability 1 of reaching an s (i.e., $\sum_{\boldsymbol{x}^t \in \mathbb{B}^t} \nu(\boldsymbol{x}^t) \xrightarrow{t\to\infty} 0$), for instance, even a modified predictor 'that never learns' (given by an i.i.d. strict semi-measure, i.e., with ν' that gives positive probability to s) would converge on ν with probability 1! In light of this, one might prefer to try to restrict attention to the infinite sequences $\mathbb{B}^\omega$; ostensibly, a way of doing this is to consider a *normalization* of ν that preserves relevant structure of ν. One possibility is the normalization $\mu_\nu^{\mathrm{Lev}}(\boldsymbol{x}^{t+1}) = \lim_s \sum_{\boldsymbol{y}^s \in \mathbb{B}^s} \nu(\boldsymbol{x}^{t+1}\boldsymbol{y}^s)/\mu_\nu^{\mathrm{Lev}}(\varnothing)$ with $\mu_\nu^{\mathrm{Lev}}(\varnothing) = \lim_s \sum_{\boldsymbol{x}^s} \nu(\boldsymbol{x}^s)$ (Levin and V'yugin, 1977, 360, also discussed by Bienvenu et al., 2017, 317 towards a notion of randomness for semi-measures); another is $\mu_\nu^{\mathrm{Sol}}(\boldsymbol{x}^{t+1}) = \mu_\nu^{\mathrm{Sol}}(\boldsymbol{x}^t)\nu(\boldsymbol{x}^{t+1})/(\nu(\boldsymbol{x}^t0) + \nu(\boldsymbol{x}^t1))$ with $\mu_\nu^{\mathrm{Sol}}(\varnothing) = 1$ (Solomonoff, 1978, 423ff; see Li and Vitányi, 2008, 302ff, though here the focus is on normalizing the Solomonoff-Levin predictor and not the possible sources ν). Note, however, that the normalization μ_ν^{Lev} is not even defined for the above ν with probability 0 of tracing a non-s infinite path. The normalization μ_ν^{Lev} is defined in this case, and it preserves important structure in the sense that it keeps the ratio between $\nu(\boldsymbol{x}0)$ and $\nu(\boldsymbol{x}1)$. Still, it necessarily fails to retain some structure (specifically the relationship between $\nu(\boldsymbol{x})$ and its two one-bit extensions); and what may also be lost here is the dominance of a universal Σ_1 measure (e.g., the Solomonoff-Levin measure), which complicates how we would proceed to make sense of the convergence on the normalized measure as a substitute for the original ν. In conclusion of this discussion, I have not been able to come up with a truly satisfying way of making sense of almost-sure convergence in the case of a semi-measure.[32]

B.2.3. The proof of theorem 3.2. The proof rests on the specification by Hutter and Muchnik (2007, 251ff) (also see Lattimore and Hutter, 2015, 5ff) of a particular Martin-Löf-random sequence $\boldsymbol{x}^\omega$ (which the authors denote 'α'), and a particular Σ_1 measure ν that is defined with the help of $\boldsymbol{x}^\omega$. The specification of $\boldsymbol{x}^\omega$ and ν is quite complicated; I refer for the details to the original paper and mention in the proof only the properties that are needed there.

Proof. We suppose $\boldsymbol{x}^\omega$ and ν as mentioned. Given some Solomonoff-Levin measure, or universal Σ_1 mixture ξ_w, we define a second universal Σ_1 mixture by

$$\xi_{w'}(\cdot) := \gamma\xi_w(\cdot) + (1-\gamma)\nu(\cdot)$$

for some $\gamma \in (0,1)$, and we show that ξ_w and $\xi_{w'}$ do not converge on $\boldsymbol{x}^\omega$.

First of all, since $\boldsymbol{x}^\omega$ is ML-random, we have that $\lambda(\boldsymbol{x}^t) \geq c_1^{-1}\xi_w(\boldsymbol{x}^t)$ for some c_1 and all t, and hence that

$$\begin{aligned}
\xi_w(x \mid \boldsymbol{x}^t) &= \sum_i w(i \mid \boldsymbol{x}^t)\nu_i(x \mid \boldsymbol{x}^t)\\
&= \sum_i \frac{w(i)\nu_i(\boldsymbol{x}^t)}{\xi_w(\boldsymbol{x}^t)}\nu_i(x \mid \boldsymbol{x}^t)\\
&\geq \frac{w(i_\lambda)\lambda(\boldsymbol{x}^t)}{\xi_w(\boldsymbol{x}^t)}\lambda(x \mid \boldsymbol{x}^t)\\
&\geq \frac{w(i_\lambda)\lambda(\boldsymbol{x}^t)}{c_1^{-1}\lambda(\boldsymbol{x}^t)}\lambda(x \mid \boldsymbol{x}^t)\\
&= c_1 w(i_\lambda)\frac{1}{2},
\end{aligned}$$

i.e., a constant. This means that there are $a_{\min}, a_{\max}$ strictly between 0 and 1 such that for all t, the probability $\xi_w(x_{t+1} \mid \boldsymbol{x}^t)$ that ξ_w assigns to the next element of $\boldsymbol{x}^\omega$ must lie within the interval $[a_{\min}, a_{\max}]$.

Another consequence of $\boldsymbol{x}^\omega$ being ML-random is that it must contain the subsequence 01 infinitely often. We focus on these occurences, employing the specific properties that ν (and $\boldsymbol{x}^\omega$) were specifically designed to fulfill.

Namely, ν is specified in such a way that for all $\boldsymbol{y} \in \mathbb{B}^* \setminus \{\varnothing\}$, $\nu(\boldsymbol{y}) = \nu(\boldsymbol{y}0) + \nu(\boldsymbol{y}1)$, and

- ○ if $x_t = 1$, then $\nu(\boldsymbol{x}^{t-1}0) = 2^{-t}$;

- ∘ if $x_t = 0$, then $\nu(\boldsymbol{x}^{t-1}1) = 0$.

Now consider the infinitely many t such that $\boldsymbol{x}^{t+2} = \boldsymbol{x}^t 01$. For such t, it holds, first, that $\nu(\boldsymbol{x}^t 1) = 0$, so that $\nu(\boldsymbol{x}^t) = \nu(\boldsymbol{x}^t 0) = \nu(\boldsymbol{x}^{t+1})$ hence $\nu(x_{t+1} \mid \boldsymbol{x}^t) = 1$. Second, for such t it holds that $\nu(\boldsymbol{x}^t) \geq \nu(\boldsymbol{x}^{t+1}0) = 2^{-t-2} = 2^{-2}\lambda(\boldsymbol{x}^t)$.

We then consider

$$\begin{aligned}\xi_{w'}(x \mid \boldsymbol{x}^t) &= \sum_j \frac{w'(j)\nu_j(\boldsymbol{x}^t)}{\xi_{w'}(\boldsymbol{x}^t)} \nu_j(x \mid \boldsymbol{x}^t) \\ &= \gamma \sum_i \frac{w(i)\nu_i(\boldsymbol{x}^t)}{\xi_{w'}(\boldsymbol{x}^t)} \nu_i(x \mid \boldsymbol{x}^t) + (1-\gamma)\frac{\nu(\boldsymbol{x}^t)}{\xi_{w'}(\boldsymbol{x}^t)} \nu(x \mid \boldsymbol{x}^t).\end{aligned}$$

Since

$$\frac{w(i)\nu_i(\boldsymbol{x}^t)}{\xi_{w'}(\boldsymbol{x}^t)} \geq^{\times} \frac{w(i)\nu_i(\boldsymbol{x}^t)}{\xi_{w}(\boldsymbol{x}^t)} = w(i \mid \boldsymbol{x}^t),$$

and, by the fact that infinitely often $\nu(\boldsymbol{x}^t) \geq^{\times} \lambda(\boldsymbol{x}^t)$, also infinitely often

$$\frac{\nu(\boldsymbol{x}^t)}{\xi_{w'}(\boldsymbol{x}^t)} \geq^{\times} \frac{\xi_{w'}(\boldsymbol{x}^t)}{\xi_{w'}(\boldsymbol{x}^t)} = 1,$$

we have infinitely often that $\xi_w(x \mid \boldsymbol{x}^t)$ is of the form

$$\xi_{w'}(x_{t+1} \mid \boldsymbol{x}^t) = \gamma' \xi_w(x_{t+1} \mid \boldsymbol{x}^t) + (1-\gamma')\nu(x_{t+1} \mid \boldsymbol{x}^t)$$

with γ' contained in an interval $[\gamma_{\min}, \gamma_{\max}]$ for $\gamma_{\min}, \gamma_{\max}$ strictly between 0 and 1. In fact, we saw that for these t it holds that $\nu(x_{t+1} \mid \boldsymbol{x}^t) = 1$, wherefore

$$\begin{aligned}\xi_{w'}(x_{t+1} \mid \boldsymbol{x}^t) &= \gamma' \xi_w(x_{t+1} \mid \boldsymbol{x}^t) + (1-\gamma') \\ &\geq \gamma_{\max} \xi_w(x_{t+1} \mid \boldsymbol{x}^t) + (1-\gamma_{\max}) \\ &= \xi_w(x_{t+1} \mid \boldsymbol{x}^t) + (1-\gamma_{\max})\left(1 - \xi_w(x_{t+1} \mid \boldsymbol{x}^t)\right) \\ &\geq \xi_w(x_{t+1} \mid \boldsymbol{x}^t) + (1-\gamma_{\max})(1 - a_{\max})\end{aligned}$$

infinitely often, and the predictions of $\xi_{w'}$ and ξ_w do not converge to each other on $\boldsymbol{x}^\omega$. □

B.2.4. Sequential propriety and locality.

B.2.4.1. *Sequential propriety.* Let us call a loss function ℓ *sequentially proper* if the cumulative loss function is proper in the sense that for all t,

$$\arg\min_{\mathsf{p}^t} \mathbf{E}_{X^t \sim \mu}\left[L_{\mathsf{p}}(X^t)\right] = \mathsf{p}^t_\mu,$$

meaning that a predictor that minimizes the μ-expected cumulative loss over a sequence of length t must correspond to μ up to t (i.e., must be a p with $\mathsf{p}(x_{s+1}, \boldsymbol{x}^s) = \mu(x_{s+1} \mid \boldsymbol{x}^s)$ for $s \leq t$). Sequential propriety carries over from propriety:

Proposition B.8. Every proper loss function is also sequentially proper.

Proof. We write out

$$\begin{aligned}\mathbf{E}_{X^t\sim\mu}\, L_{\mathsf{p}}(X^t) &= \mathbf{E}_{X^t\sim\mu}\left[\sum_{s=0}^{t-1}\ell(\mathsf{p}(X^s), X_{s+1})\right]\\ &= \sum_{s=0}^{t-1}\mathbf{E}_{X^{s+1}\sim\mu}\left[\ell(\mathsf{p}(X^s), X_{s+1})\right]\\ &= \sum_{s=0}^{t-1}\left(\sum_{\boldsymbol{x}^s\in\mathbb{B}^s}\mu(\boldsymbol{x}^s)\,\mathbf{E}_{X_{s+1}\sim\mu|\boldsymbol{x}^s}\left[\ell(p, X_{s+1})\right]\right).\end{aligned}$$

Then by the propriety of ℓ, every term

$$\mathbf{E}_{X_{s+1}\sim\mu|\boldsymbol{x}^s}\left[\ell(p, X_{s+1})\right]$$

is minimized by $p = \mu^1(\cdot \mid \boldsymbol{x}^s)$, hence the total sum of terms is minimized by the p with $\mathsf{p}(\boldsymbol{x}^s) = \mu^1(\cdot \mid \boldsymbol{x}^s)$ for all $s \leq t$, i.e., the p corresponding to μ up to t. □

B.2.4.2. *Locality.* Recall that ℓ is local if $\ell(p, x)$ is a function of $p(x)$. For distributions p over $\mathbb{B}$, this is a vacuous property: every loss function must satisfy it. For distributions p over Ω with $|\Omega| > 2$, a smooth proper loss function is local only if it is of the form

$$\ell(p, x) = -r\ln p(x) + c_x \tag{105}$$

for constants $r > 0$ and c_x for all $x \in \Omega$.

Proof. See Bernardo and Smith (1994, 73f). □

Note that we obtain the log-loss function if we set all $c_x = 0$ in (105).

B.2.4.3. *Sequential locality.* In the following, we restrict ourselves again to the outcome space $\mathbb{B}$. Let us call a loss function ℓ *sequentially local* if the cumulative loss $L_{\mathsf{p}}(\boldsymbol{x})$ of p on $\boldsymbol{x}$ is only a function of $\mu_{\mathsf{p}}(\boldsymbol{x})$ for μ_{p} corresponding to p. The log-loss function, with $L_{\mathsf{p}}(\boldsymbol{x}^t) = -\ln\mu_{\mathsf{p}}(\boldsymbol{x}^t)$, is obviously sequentially local. Conversely, every proper loss function that is sequentially local must again be of the logarithmic form (105).

Proof of proposition 6.1. Take any smooth (sequentially) proper loss function ℓ, and suppose it is sequentially local, so $L_{\mathsf{p}}(\boldsymbol{x})$ is a function of $\mu_{\mathsf{p}}(\boldsymbol{x})$. In particular, restricting the cumulative loss function to sequences of a fixed length t, $L_{\mathsf{p}}(\boldsymbol{x}^t)$ is a function of $\mu^t_{\mathsf{p}}(\boldsymbol{x}^t)$. But the measures restricted to sequences of length t are just the distributions p on $\mathbb{B}^t$; so we can view the function L^t restricted to sequences of length t as an *instantaneous* loss function ℓ^t for the distributions over $\mathbb{B}^t$, which is local because $\ell^t_p(\boldsymbol{x}) = L_{\mathsf{p}}(\boldsymbol{x})$ for $\boldsymbol{x} \in \mathbb{B}^t$ is a function of $p(\boldsymbol{x}) = \mu^t_{\mathsf{p}}(\boldsymbol{x})$. Then, since ℓ^t is also proper (because ℓ is proper and so sequentially proper by assumption), it must take the form (105), i.e.,

$$\ell^t(p, \boldsymbol{x}^t) = -r_t\ln p(\boldsymbol{x}^t) + c_{\boldsymbol{x}^t}.$$

Switching views again, it follows that for all t the function L must take the form

$$L_{\mathsf{p}}(\boldsymbol{x}^t) = -r_t \ln \mu_{\mathsf{p}}(\boldsymbol{x}^t) + c_{\boldsymbol{x}^t}. \tag{106}$$

But then the special case $t = 1$ implies that

$$\ell(\mathsf{p}(\varnothing), x) = -r \ln \mu_{\mathsf{p}}(x) + c_x,$$

for any p, $x \in \mathbb{B}$, for $r = r_1$ (note that this entails for (106) that all r_t equal the same factor r, and that $c_{\boldsymbol{x}^t}$ only depends on the number of 0's and 1's). That is,

$$\ell(p, x) = -r \ln p(x) + c_x,$$

for any $p \in \mathcal{P}$, $x \in \mathbb{B}$, which is what was to be shown. □

B.2.5. The semi-computability of cumulative losses. Let $\nu \in \Sigma_1$. Then $-\ln \nu(\cdot) \in \Pi_1$. However, it might be the case that $\nu(\cdot \mid \cdot) \notin \Sigma_1$; Q_U is a case in point, proposition 4.1. In such a case, also $-\ln \nu(\cdot \mid \cdot) \notin \Pi_1$.

Nevertheless, the sum $\sum_{t=0}^{s-1} - \ln \nu(x_{t+1} \mid \boldsymbol{x}^t)$ as a function of $\boldsymbol{x}^s$ *is* Π_1, because it is the function $-\ln \nu(\cdot)$. This is despite the fact that the individual terms in the sum are *not* Π_1. It is a direct reflection of the fact that ν is Σ_1 but conditional $\nu(\cdot \mid \cdot)$ is not.

In other words, the cumulative log-loss L_ν of a $\nu \in \Sigma_1$ is always Π_1, even if the instantaneous losses ℓ_ν might not be. The question is: does this fail to hold for other loss functions? Specifically, could it not be the case that the cumulative loss L_ν for $\nu \in \Sigma_1$ is not Π_1, if ℓ_ν is not?

As an obvious start, can we show that the sum $\sum_{t=0}^{s-1} \nu(x_{t+1} \mid \boldsymbol{x}^t)$ might *not* be Σ_1, if the conditional $\nu(\cdot \mid \cdot)$ itself is not? In particular, is the sum $\sum_{t=0}^{s-1} Q_U(x_{t+1} \mid \boldsymbol{x}^t)$ in Σ_1? I found this surprisingly hard to either prove or disprove, hence I must leave it here as an open question.

Question B.9. Is the function $S : \boldsymbol{x}^{t+1} \mapsto \sum_{t=0}^{s-1} Q_U(x_{t+1} \mid \boldsymbol{x}^t)$ in $\Delta_2 \setminus \Sigma_1$?

B.2.6. The proof of proposition 6.2.

Proof. This follows from the result of Lahtrop and Lutz (1999, also see Merkle, 2008; Nies, 2009, 271ff) that there exist computably random sequences that are at the same time *ultracompressible*. Namely, a computably random sequence $\boldsymbol{x}^\omega$ is such that for every Δ_1 measure μ it holds that

$$L_\mu(\boldsymbol{x}^t) = -\ln \mu(\boldsymbol{x}^t) \geq^+ -\ln \lambda(\boldsymbol{x}^t) = t,$$

while $\boldsymbol{x}^\omega$ is ultracompressible if for every computable unbounded non-decreasing function (every computable *order function*) $g : \mathbb{N} \to \mathbb{N}$, for almost all t,

$$K(\boldsymbol{x}^t) < K(t) + g(t).$$

Since (see A.3.2.3)

$$L_{Q_U} =^+ KM \leq^+ K$$

and (see A.3.1.3)

$$K(t) \leq^+ 2 \ln t,$$

by choice of order function $g : t \mapsto \ln t$ we have $L_{Q_U}(\boldsymbol{x}^t) \leq^+ 3 \ln t$. □

*

Notes of thanks

1. In particular, my colleagues of the Algorithms and Complexity group: both those that moved on to the newly formed Machine Learning group, and those so unfortunate as to be left behind. A special mention for my once-officemates Wouter Koolen, Nishant Mehta, and Thijs van Ommen, from whom I learned much; for my favorite foosball teammate and *paranimf* Erik Quaeghebeur; and for Rianne de Heide, who bravely proofread part of the thesis. Thanks, too, to many colleagues I had pleasant encounters with outside of the comfort zone of our own group; a special mention for team Prague.

2. In particular, my colleagues of TF, the Department of *Theoretische Filosofie*. The entire Groningen Faculty of Philosophy is a particularly friendly and constructive environment. A special mention for the dining and dancing ensemble consisting of my favorite Groningen host and *paranimf* Coos Engelsma, and Job de Grefte, Pieter van der Kolk, and Sander Verhaegh; and for my once and future favorite officemate, Marta Sznajder.

3. Thanks to the Leverhulme Trust and the University of Groningen for the financial support, and to Kevin Zollman, Jan-Willem Romeijn, Richard Pettigrew, and Laura Lanceley, the people who made it happen.

4. Ironically, right after I left, the CWI decided that a good part of the library must be cleared out to make office space for the Machine Learning group. Here I want to say thanks and good luck to the people of the library: Bikkie Aldeias, Wouter Mettrop, Rob van Rooijen, Vera Sarkol, and Lieke Schultze.

5. Thanks to Marta Sznajder for locating this document.

6. Thanks again to Marta Sznajder for locating this document.

7. Thanks to the anonymous reviewers of the paper (Sterkenburg, 2017) for comments on proofs of the results in this section.

8. Thanks to Jeanne Peijnenburg for asking the question whether this is the case, which prompted my work on this problem and which led to the answer in theorem 2.13, and ultimately to the paper (Sterkenburg, 2017).

9. Thanks to Teddy Seidenfeld for helpful discussion about this.

10. Thanks once again to Marta Sznajder for locating this document.

11. Thanks to Wouter Koolen and Tim van Erven for a helpful discussion about this problem.

12. Thanks to Tor Lattimore for confirming that this construction was relevant for this problem, and for sketching the proof.

13. Thanks to the anonymous reviewers of (Sterkenburg, 201x) for comments on the material in this chapter.

14. Thanks to Jan Leike for informing me about this proof.

15. Thanks to Leon Geerdink for commenting on a very early version of this paper.

16. Thanks to the participants of the Groningen PCCP seminar edition that was devoted to an early version of the this paper.

17. Thanks to Filippo Massari for comments on this paper.

18. Thanks to Hannes Leitgeb and Samuel Fletcher, among others, for feedback on my presentation of this work at the MCMP in Munich.

19. Thanks to Simon Huttegger for inviting me to present this work at his workshop on inductive logic at UC Irvine, and for his feedback both there and during an earlier meeting in Groningen.

20. Thanks to the anonymous reviewers of this paper for helpful comments.

21. Thanks to Konstantin Genin for insisting on this point.

22. Thanks to Ronnie Hermens for first emphasizing this point.

23. Thanks to Atze van der Ploeg for spirited discussions on this topic.

24. Thanks to Steven de Rooij and Peter Bloem for enjoyable discussions on this topic.

25. Thanks to Wouter Koolen for valuable help in understanding the theory covered in this chapter.

26. Thanks to George Barmpalias and Wolfgang Merkle for correspondence about this problem.

27. Thanks to Gerhard Schurz for helpful discussions about meta-induction.

28. Thanks to Peter Gács and to Denis Hirschfeldt for clarifying the presentation of the relation of prefix-free machines to monotone machines in Gács (2016) and Downey and Hirschfeldt (2010), respectively.

29. Thanks to Daniël Noom for going through this proof together.

30. Thanks to Alexander Shen for valuable comments on this proof.

31. Thanks to Alexander Shen for sketching this proof.

32. Thanks to Jan Leike for a helpful e-mail exchange about this issue.

*

Bibliography

P. Achinstein. Confirmation theory, order, and periodicity. *Philosophy of Science*, 30:17–35, 1963. [p. 61]

P. W. Adriaans and J. F. A. K. van Benthem, editors. *Philosophy of Information*, volume 8 of *Handbook of the Philosophy of Science*. Elsevier, 2008. [pp. 213 and 215]

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*. Springer, 2016. [p. 215]

E. Arnold. Can the best-alternative justification solve Hume's problem? On the limits of a promising approach. *Philosophy of Science*, 77(4):584–593, 2010. [p. 158]

P. S. Bandyopadhyay and M. R. Forster, editors. *Philosophy of Statistics*, volume 7 of *Handbook of the Philosophy of Science*. Elsevier, 2011. [pp. 211, 212, and 214]

G. Barmpalias and D. L. Dowe. Universality probability of a prefix-free machine. *Philosophical Transactions of the Royal Society A*, 370(1971):3488–3511, 2012. [p. 70]

A. R. Barron. Information-theoretic characterization of Bayes performance and the choice of priors in parametric and nonparametric problems. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Proceedings of the Sixth Valencia International Meeting*, volume 6 of *Bayesian Statistics*, pages 27–52. Oxford University Press, Oxford, 1998. [p. 98]

A. R. Barron, J. J. Rissanen, and B. Yu. The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory*, 44(6):2743–2760, 1998. [p. 130]

J. M. Bernardo. Expected information as expected utility. *The Annals of Statistics*, 7(3): 686–690, 1979. [p. 138]

J. M. Bernardo and A. F. M. Smith. *Bayesian Theory*. Wiley Series in Probability and Statistics. John Wiley & Sons, Chichester, England, 1994. [pp. 137, 138, and 203]

L. Bienvenu, G. Shafer, and A. K. Shen. On the history of martingales in the study of randomness. *Journal Electronique d'Histoire des Probabilités et de la Statistique*, 5(1), 2009. [pp. 37 and 38]

L. Bienvenu, R. Hölzl, C. P. Porter, and P. Shafer. Randomness and semimeasures. *Notre Dame Journal of Formal Logic*, 58(3), 2017. [pp. 184 and 200]

D. Blackwell and L. Dubins. Merging of opinion with increasing information. *The Annals of Mathematical Statistics*, 33:882–886, 1962. [pp. 72, 90, and 196]

P. Bloem, S. de Rooij, and P. W. Adriaans. Two problems for sophistication. In K. Chaudhuri, C. Gentile, and S. Zilles, editors, *Algorithmic Learning Theory: Proceedings of the Twenty-Sixth International Conference (ALT 2015)*, volume 9355 of *Lecture Notes in Artificial Intelligence*, pages 379–394. Springer, 2015. [p. 130]

G. Boole. *An Investigation of the Laws of Thought On Which Are Founded the Mathematical Theories of Logic and Probabilities*. Macmillan, London, 1854. [p. 2]

G. S. Boolos, J. P. Burgess, and R. C. Jeffrey. *Computability and Logic*. Cambridge University Press, New York, fifth edition, 2007. [p. 27]

R. B. Braithwaite. On unknown probabilities. In S. Körner, editor, *Observation and Interpretation. A Symposium of Philosophers and Physicists. Proceedings of the Ninth Symposium of the Colston Research Society*, pages 3–11, London, 1957. Butterworths. [p. 91]

G. W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950. [p. 142]

C. Calude. *Information and Randomness: An Algorithmic Perspective*. Monographs in Theoretical Computer Science. An EATCS Series. Springer, 1994. [p. 64]

G. F. L. P. Cantor. Ueber eine elementare Frage der Mannigfaltigkeitslehre. *Jahresbericht der Deutsche Mathematiker-Vereinigung*, 1:75–78, 1891. [p. 35]

R. Carnap. On inductive logic. *Philosophy of Science*, 12:72–97, 1945. [p. 3]

R. Carnap. Probability as a guide in life. *Journal of Philosophy*, 44(6):141–148, 1947a. [p. 51]

R. Carnap. On the application of inductive logic. *Philosophy and Phenomenological Research*, 8(1):133–148, 1947b. [p. 53]

R. Carnap. Reply to Nelson Goodman. *Philosophy and Phenomenological Research*, 8(3): 461–462, 1948. [p. 53]

R. Carnap. *Logical Foundations of Probability*. The University of Chicago Press, Chicago, IL, 1950. [pp. 3, 49, 50, 51, 55, 56, 57, 58, 59, 61, 79, 81, and 93]

R. Carnap. Solomonoff, March 18, 1951. Rudolf Carnap Papers, 1905-1970, ASP.1974.01, Series XII: Notes and Correspondence: Probability, Mathematics, Publishers, UCLA Administrative, and Lecture Notes, 1927-1970, Subseries 1: Probability Authors, Box 84b, Folder 57, Special Collections Department, University of Pittsburgh. Available at `http://digital2.library.pitt.edu/islandora/object/pitt%3A31735061814673/`. [p. 60]

R. Carnap. *The Continuum of Inductive Methods*. The University of Chicago Press, Chicago, IL, 1952. [pp. 3, 50, 55, 56, 57, 58, 79, 80, and 81]

R. Carnap. Inductive logic and science. *Proceedings of the American Academy of Arts and Sciences*, 80(3):189–197, 1953. [p. 52]

R. Carnap. Letter to Hilary Putnam, February 9, 1958. Rudolf Carnap Papers, 1905-1970, ASP.1974.01, Series XIV: Correspondence with Philosophy Authors, 1930-1970, Box 88c, Folder 84, Special Collections Department, University of Pittsburgh. [p. 50]

R. Carnap. *Logical Foundations of Probability*. The University of Chicago Press, Chicago, IL, 2nd edition, 1962a. [p. 210]

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, 1962b. [pp. 51, 94, 111, and 210]

R. Carnap. Replies and systematic expositions. In Schilpp (1963), pages 859–1013. [pp. 50, 51, 54, 55, 56, 58, 59, 61, 80, 81, 91, 94, and 110]

R. Carnap. Letter to Bruno de Finetti, July 30, 1963b. Rudolf Carnap Papers, 1905-1970, ASP.1974.01, Series XII. Notes and Correspondence: Probability, Mathematics, Publishers, UCLA Administrative, and Lecture Notes, 1927-1970, 1930-1970, Subseries 1: Probability Authors, Box 84a, Folder 16, Special Collections Department, University of Pittsburgh. [p. 93]

R. Carnap. Variety, analogy, and periodicity in inductive logic. *Philosophy of Science*, 30 (3):222–227, 1963c. [p. 61]

R. Carnap. Remarks on probability. *Philosophical Studies*, 14(5):65–75, 1963d. Slightly modified version of the preface to Carnap (1962a). [p. 51]

R. Carnap. Inductive logic and rational decisions. In Carnap and Jeffrey (1971), pages 5–31. Modifed and expanded version of Carnap (1962b). [pp. 51, 80, 84, 94, and 111]

R. Carnap. A basic system of inductive logic: Part 1. In Carnap and Jeffrey (1971), pages 33–165. [pp. 54 and 80]

R. Carnap. Notes on probability and induction. *Synthese*, 25(3–4):269–298, 1973. Reprinted in Hintikka (1975), pages 293–324. [p. 50]

R. Carnap. A basic system of inductive logic: Part 2. In Jeffrey (1980), pages 7–155. [pp. 54, 58, and 97]

R. Carnap and M. Gardner, editor. *Philosophical Foundations of Physics: An Introduction to the Philosophy of Science*. Basic Books, New York, 1966. [pp. 4 and 50]

R. Carnap and R. C. Jeffrey, editors. *Studies in Inductive Logic and Probability*, volume 1. University of California Press, 1971. [pp. 210 and 214]

R. Carnap and W. Stegmüller. *Induktive Logik und Wahrscheinlichkeit*. Springer, Vienna, 1959. [pp. 50 and 58]

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning and Games*. Cambridge University Press, Cambridge, 2006. [pp. 42, 78, 99, 142, 143, 146, and 158]

N. Cesa-Bianchi, Y. Freund, D. Haussler, D. P. Helmbold, R. E. Schapire, and M. K. Warmuth. How to use expert advice. *Journal of the ACM*, 44(3):427–485, 1997. [pp. 42 and 98]

G. J. Chaitin. On the length of programs for computing finite binary sequences: Statistical considerations. *Journal of the Association for Computing Machinery*, 16:145–159, 1969. [p. 131]

G. J. Chaitin. A theory of program size formally identical to information theory. *Journal of the Association for Computing Machinery*, 22(3):329–340, 1975. [pp. 163, 165, 168, and 176]

A. Church. Review of Turing (1936). *Journal of Symbolic Logic*, 2(1):42–43, 1937. [p. 26]

A. Church. On the concept of a random sequence. *Bulletin of the American Mathematical Society*, 46:130–135, 1940. [p. 38]

B. J. Copeland. Narrow versus wide mechanism: Including a re-examination of Turing's views on the mind-machine issue. *Journal of Philosophy*, 97(1):5–32, 2000. [p. 27]

B. J. Copeland. Turing's thesis. In Olszewski et al. (2006), pages 147–174. [p. 27]

D. Corfield and J. Williamson, editors. *Foundations of Bayesianism*, volume 24 of *Applied Logic Series*. Kluwer, 2001. [pp. 212, 213, and 214]

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, Hoboken, NJ, second edition, 2006. [pp. 140, 166, 169, and 171]

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. [p. 114]

A. Dasgupta. Mathematical foundations of randomness. In Bandyopadhyay and Forster (2011), pages 641–710. [p. 156]

A. P. Dawid. The well-calibrated Bayesian. *Journal of the American Statistical Association*, 77(379):605–611, 1982. [p. 33]

A. P. Dawid. Present position and potential developments: Some personal views. Statistical theory. The prequential approach. *Journal of the Royal Statistical Society A*, 147:278–292, 1984. [pp. 22, 78, 90, and 142]

A. P. Dawid. Calibration-based empirical probability. *The Annals of Statistics*, 13(4):1251–1274, 1985a. [pp. 31, 33, 109, and 219]

A. P. Dawid. The impossibility of inductive inference. Comment on Oakes (1985). *Journal of the American Statistical Association*, 80(390):339, 1985b. [pp. 15, 31, 33, and 113]

A. P. Dawid. Probability forecasting. In S. Kotz and N. L. Johnson, editors, *Encyclopedia of Statistical Sciences*, volume 7, pages 210–218. Wiley-Interscience, New York, NY, 1986. [p. 142]

A. P. Dawid. Comments on: Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *Test*, 17:243–244, 2008. [p. 142]

A. P. Dawid and V. G. Vovk. Prequential probability: Principles and properties. *Bernoulli*, 5(1):125–162, 1999. [p. 79]

A. R. Day. On the computational power of random strings. *Annals of Pure and Applied Logic*, 160:214–228, 2009. [p. 67]

A. R. Day. Increasing the gap between descriptional complexity and algorithmic probability. *Transactions of the American Mathematical Society*, 363(10):5577–5604, 2011. [pp. 178, 185, and 186]

B. de Finetti. La prévision: Ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré*, 7:1–68, 1937. Translated by H.J. Kyburg, Jr. as: Foresight: its logical laws, its subjective sources. In H.J. Kyburg, Jr., H.E. Smokler, editors, *Studies in Subjective Probability*, pages 97-158. John Wiley & Sons, Inc., 1964. [pp. 90 and 91]

B. de Finetti. Does it make sense to speak of 'good probability appraisers'? In I. J. Good, editor, *The Scientist Speculates: An Anthology of Partly-Baked Ideas*, pages 357–364. Basic Books, New York, NY, 1962. [p. 142]

S. de Rooij and P. D. Grünwald. Luckiness and regret in minimum description length inference. In Bandyopadhyay and Forster (2011), pages 865–900. [p. 129]

S. de Rooij, T. van Erven, P. D. Grünwald, and W. M. Koolen. Follow the leader if you can, hedge if you must. *Journal of Machine Learning Research*, 15:1281–1316, 2014. [p. 143]

P. W. Diaconis and D. A. Freedman. De Finetti's generalizations of exchangeability. In Jeffrey (1980), pages 233–249. [p. 91]

P. W. Diaconis and D. A. Freedman. On the consistency of Bayes estimates. *The Annals of Statistics*, 14(1):1–26, 1986. [pp. 23 and 87]

P. Domingos. *The Master Algorithm: How the Quest For the Ultimate Learning Machine Will Remake Our World*. Basic Books, New York, NY, 2015. [p. 5]

R. G. Downey and D. R. Hirschfeldt. *Algorithmic Randomness and Complexity*, volume 1 of *Theory and Applications of Computability*. Springer, New York, 2010. [pp. vii, 39, 64, 65, 67, 68, 70, 157, 163, 168, 186, 187, and 208]

R. G. Downey, E. J. Griffiths, and G. LaForte. On Schnorr and computable randomness, martingales, and machines. *Mathematical Logic Quarterly*, 50(6):613–627, 2004. [p. 157]

J. Earman. *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory*. A Bradford Book. MIT Press, Cambridge, MA, 1992. [pp. 48 and 109]

H. Feigl. De principiis non disputandum . . . ? On the meaning and the limits of justification. In M. Black, editor, *Philosophical Analysis*, pages 119–156. Cornell University Press, New York, NY, 1950. [p. 29]

T. L. Fine. *Theories of Probability*. Academic Press, New York, NY, 1973. [p. 86]

C. E. Freer and D. M. Roy. Computable de Finetti measures. *Annals of Pure and Applied Logic*, 163(5):530–546, 2012. [p. 31]

D. M. Gabbay, S. Hartmann, and J. Woods, editors. *Inductive Logic*, volume 10 of *Handbook of the History of Logic*. Elsevier, 2011. [pp. 218 and 223]

P. Gács. On the symmetry of algorithmic information. *Soviet Mathematics Doklady*, 15(5): 1477–1480, 1974. Translation of the Russian original in *Doklady Akademii Nauk SSSR*, 218(6):1265-1267, 1974. [p. 163]

P. Gács. On the relation between descriptional complexity and algorithmic probability. *Theoretical Computer Science*, 22:71–93, 1983. [p. 178]

P. Gács. Every sequence is reducible to a random one. *Information and Control*, 70:186–192, 1986. [p. 157]

P. Gács. Expanded and improved proof of the relation between description complexity and algorithmic probability. Unpublished manuscript, available at `http://www.cs.bu.edu/~gacs/`, 2016. [pp. 66, 163, and 208]

H. Gaifman and M. Snir. Probabilities over rich languages, testing and randomness. *Journal of Symbolic Logic*, 47(3):495–548, 1982. [p. 157]

M. C. Galavotti. Subjectivism, objectivism and objectivity in Bruno de Finetti's Bayesianism. In Corfield and Williamson (2001), pages 161–174. [p. 91]

M. C. Galavotti. *Philosophical Introduction to Probability*, volume 167 of *CSLI Lecture Notes*. Center for the Study of Language and Information, Stanford, CA, 2005. [p. 52]

R. O. Gandy. Church's Thesis and principles for mechanisms. In J. Barwise, H. J. Keisler, and K. Kunen, editors, *Proceedings of the Kleene Symposium*, volume 101 of *Studies in Logic and the Foundations of Mathematics*, pages 123–148. North-Holland, 1980. [p. 27]

R. O. Gandy. The confluence of ideas in 1936. In R. Herken, editor, *The Universal Turing Machine: A Half-Century Survey*, pages 55–111. Oxford University Press, 1988. [p. 26]

H. G. Gauch, Jr. *Scientific Method in Practice*. Cambridge University Press, 2003. [p. 41]

A. Gelman and C. R. Shalizi. Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, 66:8–38, 2013. [p. 24]

D. A. Gillies. *Philosophical Theories of Probability*. Philosophical Issues in Science. Routledge, New York, NY, 2000. [pp. 51 and 52]

D. A. Gillies. Popper and computer induction. *BioEssays*, 23:859–860, 2001a. [pp. 5 and 109]

D. A. Gillies. Bayesianism and the fixity of the theoretical framework. In Corfield and Williamson (2001), pages 363–379. [pp. 90 and 109]

T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102:359–378, 2007. [p. 138]

I. J. Good. Rational decisions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 14(1):107–114, 1952. [pp. 101 and 138]

I. J. Good. *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*. MIT Press, Cambridge, MA, 1965. [p. 58]

N. Goodman. A query on confirmation. *The Journal of Philosophy*, 43(14):383–385, 1946. [pp. 17, 53, and 61]

N. Goodman. On infirmities of confirmation-theory. *Philosophy and Phenomenological Research*, 8(1):149–151, 1947. [p. 61]

N. Goodman. *Fact, Fiction, and Forecast*. Athlone Press, London, 1954. [pp. 16 and 17]

T. Groves. *Let's Reappraise Carnapian Inductive Logic!* PhD Dissertation, University of Kent, 2015. [p. 54]

P. D. Grünwald. *The Minimum Description Length Principle*. MIT Series in Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2007. [pp. vi, 28, 41, 59, 88, 89, 98, 99, 129, 130, 141, 143, 166, 167, and 171]

P. D. Grünwald and A. P. Dawid. Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. *The Annals of Statistics*, 32(4):1367–1433, 2004. [p. 140]

P. D. Grünwald and P. M. B. Vitányi. Algorithmic information theory. In Adriaans and van Benthem (2008), pages 289–317. [p. 42]

I. Hacking. *An Introduction to Probability and Inductive Logic*. Cambridge University Press, 2001. [p. 24]

I. Hacking. *The Emergence of Probability: A Philosophical Study of Early Ideas About Probability, Induction and Statistical Inference*. Cambridge University Press, 2nd edition, 2006. [pp. 3 and 54]

G. Harman and S. Kulkarni. *Reliable Reasoning: Induction and Statistical Learning Theory*. The Jean Nicod Lectures. A Bradford Book. MIT Press, Cambridge, MA, 2007. [p. 41]

D. Haussler, J. Kivinen, and M. K. Warmuth. Tight worst-case loss bounds for predicting with expert advice. In P. M. B. Vitányi, editor, *Computational Learning Theory: Proceedings of the Second European Conference (EuroCOLT '95)*, volume 744 of *Lecture Notes in Artificial Intelligence*, pages 69–83. Springer, 1995. [p. 146]

C. Hempel. A purely syntactical definition of confirmation. *Journal of Symbolic Logic*, 8(4): 122–143, 1943. [p. 53]

P. Herz. Kritische Bemerkungen zu Reichenbachs Behandlung des Humeschen Problems. *Erkenntnis*, 6:25–31, 1936. [p. 30]

E. Hewitt and L. J. Savage. Symmetric measures on cartesian products. *Transactions of the American Mathematical Society*, 80(2):470–501, 1955. [p. 92]

D. Hilbert and W. F. Ackermann. *Grundzüge der Theoretischen Logik*. Springer, Berlin, 1928. [p. 2]

R. Hilpinen. Carnap's new system of inductive logic. *Synthese*, 25(3–4):307–333, 1973. Reprinted in Hintikka (1975), pages 333-359. [p. 54]

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. [p. 106]

J. Hintikka. Unknown probabilities, Bayesianism, and de Finetti's representation theorem. In R. C. Buck and R. S. Cohen, editors, *Proceedings of the 1970 Biennial Meeting of the Philosophy of Science Association*, pages 325–341. Reidel, Dordrecht, 1971. [p. 91]

J. Hintikka, editor. *Rudolf Carnap, Logical Empiricist. Materials and Perspectives*, volume 73 of *Synthese Library*. Reidel, 1975. [pp. 210, 213, and 214]

A. Hodges. Did Church and Turing have a thesis about machines? In Olszewski et al. (2006), pages 242–252. [p. 27]

R. Hoffmann, V. I. Minkin, and B. K. Carpenter. Ockham's razor and chemistry. *Bulletin de la Société Chimique de France*, 133:117–130, 1996. [p. 41]

J. V. Howard. Computable explanations. *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik*, 21:215–224, 1975. [p. 37]

C. Howson. *Hume's Problem: Induction and the Justification of Belief*. Oxford University Press, New York, 2000. [pp. 18, 24, 28, 79, and 112]

C. Howson. The logic of Bayesian probability. In Corfield and Williamson (2001), pages 137–160. [p. 112]

C. Howson. Bayesianism as a pure logic of inference. In Bandyopadhyay and Forster (2011), pages 441–471. [p. 112]

D. Hume. *Philosophical Essays Concerning Human Understanding*. A. Millar, London, 1st edition, 1748. Published in 1758 and since under the title *An Enquiry Concerning Human Understanding*. [p. 17]

S. M. Huttegger. Merging of opinions and probability kinematics. *The Review of Symbolic Logic*, 8(4):611–648, 2015. [pp. 72 and 200]

S. M. Huttegger. Analogical predictive probabilities. Forthcoming in *Mind*, 201x. [p. 54]

M. Hutter. Convergence and loss bounds for Bayesian sequence prediction. *IEEE Transactions on Information Theory*, 49(8):2061–2067, 2003a. [p. 196]

M. Hutter. Optimality of universal Bayesian sequence prediction for general loss and alphabet. *Journal of Machine Learning Research*, 4:971–1000, 2003b. [pp. 15 and 132]

M. Hutter. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*. Texts in Theoretical Computer Science. An EATCS Series. Springer, Berlin, 2005. [pp. 129 and 132]

M. Hutter. Sequential predictions based on algorithmic complexity. *Journal of Computer and System Sciences*, 72:95–117, 2006. [p. 129]

M. Hutter. On universal prediction and Bayesian confirmation. *Theoretical Computer Science*, 384(1):33–48, 2007. [pp. 75 and 93]

M. Hutter and A. A. Muchnik. On semimeasures predicting Martin-Löf random sequences. *Theoretical Computer Science*, 382:247–261, 2007. [pp. 95 and 201]

R. C. Jeffrey. Goodman's query. *Journal of Philosophy*, 63(11):281–288, 1966. [p. 53]

R. C. Jeffrey. Probability measures and integrals. In Carnap and Jeffrey (1971), pages 167–223. [p. 54]

R. C. Jeffrey. Review of Schilpp (1963). *Journal of Symbolic Logic*, 37(3):631–633, 1972. [p. 53]

R. C. Jeffrey. Carnap's inductive logic. *Synthese*, 25(3–4):299–306, 1973. Reprinted in Hintikka (1975), pages 325–332. [pp. 53, 94, and 96]

R. C. Jeffrey, editor. *Studies in Inductive Logic and Probability*, volume 2. University of California Press, 1980. [pp. 210 and 212]

H. Jeffreys. *Theory of Probability*. Clarendon Press, Oxford, 1939. [pp. 17 and 40]

W. E. Johnson. *Logic, Part III: The Logical Foundations of Science*. Cambridge University Press, 1924. [p. 56]

W. E. Johnson. Probability: the deductive and inductive problems. *Mind*, 41:409–423, 1932. [pp. 3 and 56]

Y. Kalnishkan. Predictive complexity for games with finite outcome sequences. In V. G. Vovk, H. Papadopoulos, and A. Gammerman, editors, *Measures of Complexity: Festschrift for Alexey Chervonenkis*, pages 117–139. Springer, 2015. [p. 152]

Y. Kalnishkan and M. V. Vyugin. On the absence of predictive complexity for some games. In N. Cesa-Bianchi, M. Numao, and R. Reischuk, editors, *Algorithmic Learning Theory: Proceedings of the 13th International Conference (ALT 2002)*, volume 2533 of *Lecture Notes in Computer Science*, pages 164–172. Springer, 2002a. [p. 154]

Y. Kalnishkan and M. V. Vyugin. Mixability and the existence of weak complexities. In J. Kivinen and R. H. Sloan, editors, *Proceedings of the 15th Annual Conference on Computational Learning Theory (COLT 2002)*, volume 2375 of *Lecture Notes in Artificial Intelligence*, pages 105–120. Springer, 2002b. [p. 155]

Y. Kalnishkan, V. G. Vovk, and M. V. Vyugin. A criterion for the existence of predictive complexity for binary games. In S. Ben-David, J. Case, and A. Maruoka, editors, *Algorithmic Learning Theory: Proceedings of the 15th International Conference (ALT 2004)*, volume 3244 of *Lecture Notes in Computer Science*, pages 249–263. Springer, 2004. [p. 154]

Y. Kalnishkan, V. G. Vovk, and M. V. Vyugin. How many strings are easy to predict? *Information and Computation*, 201:55–71, 2005. [p. 171]

R. Kass and A. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90 (420):773–795, 1995. [p. 130]

K. T. Kelly. *The Logic of Reliable Inquiry*. Logic and Computation in Philosophy. Oxford University Press, New York, 1996. [pp. 15 and 78]

K. T. Kelly. Learning theory and epistemology. In I. Niiniluoto, M. Sintonen, and J. Woleński, editors, *Handbook of Epistemology*, pages 183–203. Kluwer, Dordrecht, 2004. Page numbers refer to reprint in Arló-Costa et al. (2016), 695-716. [pp. 32 and 48]

K. T. Kelly. Ockham's razor, truth, and information. In Adriaans and van Benthem (2008), pages 321–360. [p. 131]

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. [pp. 36, 105, 106, 107, and 114]

J. G. Kemeny. The use of simplicity in induction. *Philosophical Review*, 62(3):391–408, 1953. [pp. 39, 40, and 109]

J. G. Kemeny. Carnap's theory of probability and induction. In Schilpp (1963), pages 711–738. [pp. 55 and 58]

J. M. Keynes. *A Treatise on Probability*. Macmillan, London, 1921. [pp. 3, 51, and 52]

S. C. Kleene. On notation for ordinal numbers. *Journal of Symbolic Logic*, 3(4):150–155, 1938. [pp. 35 and 74]

S. C. Kleene. Recursive predicates and quantifiers. *Transactions of the American Mathematical Society*, 53:41–73, 1943. [p. 65]

S. C. Kleene. *Introduction to Metamathematics*. North-Holland, 1952. [p. 26]

S. C. Kleene. *Mathematical Logic*. Wiley, New York, NY, 1967. [p. 26]

A. N. Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer, Berlin, 1933. [p. 38]

A. N. Kolmogorov. On tables of random numbers. *Sankhyā: The Indian Journal of Statistics: Series A*, 25(4):369–376, 1963. Reprinted in *Theoretical Computer Science* 207:387-395, 1998. [p. 38]

A. N. Kolmogorov. Three approaches to the quantitive definition of information. *Problems of Information Transmission*, 1(1):1–7, 1965. Translation of the Russian original in *Problemy Peredachi Informatsii* 1(1):3-11, 1965. Reprinted in Shiryaev (1993), 184-193. [pp. 37, 38, 131, 132, and 175]

A. N. Kolmogorov. On the logical foundations of information theory and probability theory. *Problems of Information Transmission*, 5(3):1–4, 1969. Translation of the Russian original in *Problemy Peredachi Informatsii*, 5(3):3-7, 1969. Page numbers refer to reprint in Shiryaev (1993), 203-207. [p. 131]

A. N. Kolmogorov. Combinatiorial foundations of information theory and the calculus of probabilities. *Russian Mathematical Surveys*, 38(4):27–36, 1983. Translation of the Russian original in *Uspekhi Matematicheskikh Nauk*, 38(4):29-40, 1983. Reprinted in Shiryaev (1993), 208-218. [p. 131]

A. N. Kolmogorov. On works in information theory and some of its applications. In Shiryaev (1993), pages 219–221. [p. 131]

W. M. Koolen and T. van Erven. Second-order quantile methods for experts and combinatorial games. In N. Lawrence and M. Reid, editors, *Proceedings of the Twenty-Eighth Annual Conference on Computational Learning Theory (COLT 2015)*, volume 40 of *Journal of Machine Learning Research: Workshop and Conference Proceedings*, pages 1–21, 2015. [p. 155]

M. Koppel and H. Atlan. An almost machine-independent theory of program-length, complexity, sophistication, and induction. *Information Sciences*, 56:23–33, 1991. [p. 130]

L. G. Kraft. A device for quantizing, grouping, and coding amplitude modulated pulses. Master's thesis, Dept. of Electrical Engineering, MIT, Cambridge, MA, 1949. [p. 166]

T. S. Kuhn. Objectivity, value judgement, and theory choice. In *The Essential Tension: Selected Studies in Scientific Tradition and Change*, pages 320–339. University of Chicago Press, Chicago, IL, 1977. [p. 4]

S. A. Kurtz. *Randomness and genericity in the degrees of unsolvability.* PhD Dissertation, University of Illinois, 1981. [p. 157]

J. I. Lahtrop and J. H. Lutz. Recursive computational depth. *Information and Computation*, 153(2):139–172, 1999. [pp. 153 and 204]

I. Lakatos. Changes in the problem of inductive logic. In I. Lakatos, editor, *The Problem of Inductive Logic: Proceedings in the International Colloquium in the Philosophy of Science, volume 2*, Studies in Logic and the Foundations of Mathematics, pages 315–417. North-Holland, Amsterdam, 1968. [p. 54]

T. Lattimore and M. Hutter. On Martin-Löf (non-)convergence of Solomonoff's universal mixture. *Theoretical Computer Science*, 588:2–15, 2015. [pp. 95 and 201]

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 Twenty-Sixth International Conference (ALT 2015)*, volume 9355 of *Lecture Notes in Artificial Intelligence*, pages 364–378. Springer, 2015. [pp. 105, 115, and 116]

L. A. Levin. On the notion of a random sequence. *Soviet Mathematics Doklady*, 14(5): 1413–1416, 1973. Translation of the Russian original in *Doklady Akademii Nauk SSSR*, 212(3):548-550, 1973. [pp. 38, 66, and 181]

L. A. Levin. Laws of information conservation (nongrowth) and aspects of the foundation of probability theory. *Problems of Information Transmission*, 10:206–210, 1974. Translation of the Russian original in *Problemy Peredachi Informatsii* 10(3):30-35, 1974. [pp. 163, 165, 168, and 176]

L. A. Levin. Randomness conservation inequalities; information and independence in mathematical theories. *Information and Control*, 61(1):15–37, 1984. [p. 83]

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. [p. 39]

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. [pp. 68 and 200]

M. Li and P. M. B. Vitányi. Inductive reasoning and Kolmogorov complexity (preliminary version). In *Proceedings of the Fourth IEEE Annual Conference on Structure in Complexity Theory*, pages 165–185. IEEE, 1989. [p. 39]

M. Li and P. M. B. Vitányi. Philosophical issues in Kolmogorov complexity. In W. Kuich, editor, *Proceedings of the 19th International Colloquium on Automata, Languages and Programming*, volume 623 of *Lecture Notes in Computer Science*, pages 1–16. Springer, 1992a. [pp. 85 and 86]

M. Li and P. M. B. Vitányi. Inductive reasoning and Kolmogorov complexity. *Journal of Computer and System Sciences*, 44(2):343–384, 1992b. [p. 39]

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. [pp. 36, 37, 42, 64, 66, 67, 68, 69, 70, 71, 75, 83, 85, 86, 92, 95, 111, 122, 129, 130, 131, 132, 163, 171, 174, 175, 176, 186, 192, 196, and 200]

D. V. Lindley. *Understanding Uncertainty*. John Wiley & Sons, Hoboken, NJ, 2006. [p. 81]

N. Littlestone and M. K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108:212–261, 1994. [p. 42]

D. J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003. [pp. 140, 169, and 171]

P. Martin-Löf. The definition of random sequences. *Information and Control*, 9:602–619, 1966. [pp. 38, 71, and 179]

P. Martin-Löf. Algorithms and randomness. *Review of the International Statistical Institute*, 37(3):265–272, 1969. [p. 114]

V. Mayer-Schönberger and K. Cukier. *Big Data: The Revolution That Will Transform How We Live, Work, and Think*. Houghton Mifflin Harcourt, Boston, MA, 2012. [p. 85]

B. McMillan. Two inequalities implied by unique decipherability. *IRE Transactions on Information Theory*, 2(4):115–116, 1956. [p. 166]

N. Merhav and M. Feder. Universal prediction. *IEEE Transactions on Information Theory*, 44(8):2124–2147, 1998. [pp. 78, 89, 99, 138, and 139]

W. Merkle. The complexity of stochastic sequences. *Journal of Computer and System Sciences*, 74:350–357, 2008. [p. 204]

W. Merkle, N. Mihailović, and T. A. Slaman. Some results on effective randomness. *Theory of Computing Systems*, 39(5):707–721, 2006. [p. 157]

M. L. Minsky. Steps towards artificial intelligence. *Proceedings of the IRE*, 49(1):8–30, 1961. [p. 34]

A. Mostovski. On definable sets of positive integers. *Fundamenta Mathematicae*, 34:81–112, 1947. [p. 65]

A. A. Muchnik, A. L. Semonov, and V. A. Uspensky. Mathematical metaphysics of randomness. *Theoretical Computer Science*, 207:263–317, 1998. [pp. 131 and 132]

M. Müller. Stationary algorithmic probability. *Theoretical Computer Science*, 411(1):113–130, 2010. [p. 96]

A. H. Murphy and E. S. Epstein. A note on probability forecasts and "hedging". *Journal of Applied Meteorology*, 6:1002–1004, 1967. [p. 138]

E. Nagel. Carnap's theory of induction. In Schilpp (1963), pages 785–825. [p. 54]

A. Nies. *Computability and Randomness*, volume 51 of *Oxford Logic Guides*. Oxford University Press, 2009. [pp. 39, 64, 163, 166, 168, 175, 179, 180, 182, 187, and 204]

J. D. Norton. 'Nature is the realisation of the simplest conceivable mathematical ideas': Einstein and the canon of mathematical simplicity. *Studies in History and Philosophy of Modern Physics*, 31(2):135–170, 2000. [p. 41]

D. Oakes. Self-calibrating priors do not exist. *Journal of the American Statistical Association*, 80(390):340–341, 1985. [pp. 33, 211, and 219]

P. Odifreddi. *Classical Recursion Theory: The Theory of Functions and Sets of Natural Numbers*, volume 125 of *Studies in Logic and The Foundations of Mathematics*. North-Holland, Amsterdam, 1989. [pp. 25 and 35]

A. Olszewski, J. Woleński, and R. Janusz, editors. *Church's Thesis After 70 Years*, volume 1 of *ontos mathematical logic*. ontos, 2006. [pp. 211 and 214]

R. Ortner and H. Leitgeb. Mechanizing induction. In Gabbay et al. (2011), pages 719–772. [p. 123]

D. Osherson and S. Weinstein. Recognizing strong random reals. *The Review of Symbolic Logic*, 1(1):56–63, 2008. [p. 157]

J. C. Owings, Jr. Diagonalization and the recursion theorem. *Notre Dame Journal of Formal Logic*, 14(1):95–99, 1973. [p. 74]

J. B. Paris and A. Vencovská. *Pure Inductive Logic*. Perspectives in Logic. Cambridge University Press, 2015. [pp. 54 and 92]

G. Piccinini. The physical Church-Turing thesis: Modest or bold? *British Journal for the Philosophy of Science*, 62:733–769, 2011. [pp. 26 and 28]

H. Poincaré. *La Science et l'Hypothèse*. Flammarion, Paris, 1902. [pp. 17, 40, and 41]

J. Poland and M. Hutter. Asymptotics of discrete MDL for online prediction. *IEEE Transactions on Information Theory*, 51(11):3780–3795, 2005. [p. 196]

C. P. Porter. On analogues of the Church-Turing thesis in algorithmic randomness. *Review of Symbolic Logic*, 9(3), 2016. [p. 156]

H. Putnam. 'Degree of confirmation' and inductive logic. In Schilpp (1963), pages 761–783. Reprinted in Putnam (1975), pages 270–292. [pp. 5, 32, 47, 54, 61, 106, 107, 108, 109, 114, and 115]

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 Putnam (1975), pages 293–304. [pp. 5, 50, 61, 110, and 115]

H. Putnam. The 'corroboration' of theories. In P. A. Schilpp, editor, *The Philosophy of Karl Popper, Book I*, volume 14 of *The Library of Living Philosophers*, pages 221–240. Open Court, La Salle, IL, 1974. Reprinted in Putnam (1975), pages 250–269. [p. 107]

H. Putnam. *Mathematics, Matter, and Method*, volume 1. Cambridge University Press, 1975. [p. 218]

H. Putnam. Reflexive reflections. *Erkenntnis*, 22:143–153, 1985. [p. 114]

F. P. Ramsey. Truth and probability. In R. B. Braithwaite, editor, *The Foundations of Mathematics and other Logical Essays*, chapter 7, pages 156–198. Routledge and Kegan Paul, London, 1931. [p. 52]

S. Rathmanner and M. Hutter. A philosophical treatise of universal induction. *Entropy*, 13 (6):1076–1136, 2011. [pp. 28 and 86]

H. Reichenbach. Die logischen Grundlagen des Wahrscheinlichkeitsbegriffs. *Erkenntnis*, 3: 401–425, 1933. [p. 29]

H. Reichenbach. *Wahrscheinlichkeitslehre: eine Untersuchung Über die Logischen und Mathematischen Grundlagen der Wahrscheinlichkeitsrechnung*. Sijthoff, Leiden, 1935. [p. 29]

H. Reichenbach. *Experience and Prediction*. University of Chicago Press, Chicago, IL, 1938. [pp. 3, 29, and 30]

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. [pp. 64 and 66]

J. J. Rissanen. Stochastic complexity and modeling. *Annals of Statistics*, 14:1080–1100, 1986. [p. 41]

J. J. Rissanen. Stochastic complexity (with discussion). *Journal of the Royal Statistical Society, Series B*, 49:223–239, 252–265, 1987. [p. 41]

J. J. Rissanen. *Stochastic Complexity in Statistical Inquiry*, volume 15 of *Series in Computer Science*. World Scientific, Singapore, 1989. [pp. 129 and 141]

H. Rogers, Jr. Gödel numberings of partial recursive functions. *Journal of Symbolic Logic*, 23(3):331–341, 1958. [p. 194]

H. Rogers, Jr. *Theory of Recursive Functions and Effective Computability*. McGraw-Hill, New York, 1967. [pp. 25, 35, 76, and 194]

J.-W. Romeijn. Hypotheses and inductive predictions. *Synthese*, 141(3):333–364, 2004. [pp. 24, 87, 89, 91, and 112]

D. Ryabko. On finding predictors for arbitrary families of processes. *Journal of Machine Learning Research*, 11:581–602, 2010. [p. 37]

W. C. Salmon. *The Foundations of Scientific Inference*. University of Pittsburgh Press, 1967. [pp. 18, 29, and 52]

W. C. Salmon. The pragmatic justification of induction. Oxford Readings in Philosophy, pages 85–97. Oxford University Press, Oxford, 1974. [p. 29]

W. C. Salmon. Rationality and objectivity in science or Tom Kuhn meets Tom Bayes. In C. W. Savage, editor, *Scientific Theories*, volume 14 of *Minnesota Studies in the Philosophy of Science*, pages 175–204. University of Minnesota Press, Minneapolis, MN, 1990. [p. 4]

W. C. Salmon. Hans Reichenbach's vindication of induction. *Erkenntnis*, 35:99–122, 1991. [p. 29]

M. J. Schervish. Comment on Dawid (1985a). *The Annals of Statistics*, 13(4):1274–1282, 1985a. [p. 112]

M. J. Schervish. Comment on Oakes (1985). *Journal of the American Statistical Association*, 80(390):341–342, 1985b. [p. 19]

P. A. Schilpp, editor. *The Philosophy of Rudolf Carnap*, volume 11 of *The Library of Living Philosophers*. Open Court, La Salle, IL, 1963. [pp. 47, 210, 214, 215, 217, and 218]

M. Schmidt and H. Lipson. Distilling free-form natural laws from experimental data. *Science*, 324(5923):81–85, 2009. [p. 5]

C.-P. Schnorr. *Zufälligkeit und Wahrscheinlichkeit. Eine algorithmische Begründung der Wahrscheinlichkeitstheorie*, volume 218 of *Lecture Notes in Mathematics*. Springer, Berlin, 1971a. [pp. 38 and 157]

C.-P. Schnorr. A unified approach to the definition of random sequences. *Mathematical Systems Theory*, 5(3):246–258, 1971b. [pp. 38 and 157]

C.-P. Schnorr. Process complexity and effective random tests. *Journal of Computer and System Sciences*, 7(4):376–388, 1973. [pp. 38, 67, 168, and 180]

C.-P. Schnorr. A survey of the theory of random sequences. In R. E. Butts and J. Hintikka, editors, *Basic Problems in Methodology and Linguistics: Proceedings of the Fifth International Congress of Logic, Methodology, and Philosophy of Science*, volume 11 of *The University of Western Ontario Series in Philosophy of Science*, pages 193–211. Reidel, Dordrecht, 1977. [p. 67]

G. Schurz. The meta-inductivist's winning strategy in the prediction game: A new approach to Hume's problem. *Philosophy of Science*, 75(3):278–305, 2008. [pp. 18, 30, 157, and 158]

W. Sellars. Induction as vindication. *Philosophy of Science*, 31(3):197–231, 1964. [p. 30]

C. E. Shannon. The mathematical theory of communication. *Bell System Technical Journal*, 27(3, 4):379–423, 623–656, 1948. [pp. 139 and 140]

A. K. Shen, V. A. Uspensky, and N. K. Vereshchagin. Kolmogorov complexity and algorithmic randomness. Forthcoming translation of Russian edition, MCCME Publishing House, Moscow, Russia, 2013. Draft available at `http://www.lirmm.fr/~ashen/`, 20xx. [pp. 39, 66, 67, 163, 171, 174, 178, 181, and 192]

A. N. Shiryaev. Kolmogorov: Life and creative activities. *The Annals of Probability*, 17(3): 866–944, 1989. [p. 131]

A. N. Shiryaev, editor. *Selected Works of A.N. Kolmogorov. Volume III: Information Theory and the Theory of Algorithms*, volume 27 of *Mathematics and Its Applications (Soviet Series)*. Kluwer, Dordrecht, 1993. [pp. 215 and 216]

W. Sieg. Calculations by man and machine: Conceptual analysis. In W. Sieg, R. Sommer, and C. Talcot, editors, *Reflections on the Foundations of Mathematics: Essays in Honor of Solomon Feferman*, volume 15 of *Lecture Notes in Logic*, pages 390–409. Association for Symbolic Logic, 2002a. [p. 27]

W. Sieg. Calculations by man and machine: Mathematical presentation. In P. Gärdenfors, J. Woleński, and K. Kijania-Placek, editors, *In the Scope of Logic, Methodology and Philosophy of Science. Volume I of the 11th International Congress*, volume 315 of *Synthese Library*, pages 247–262. Kluwer, 2002b. [p. 27]

W. Sieg. On computability. In A. Irvine, editor, *Philosophy of Mathematics*, volume 4 of *Handbook of the Philosophy of Science*, pages 525–621. Elsevier, 2008. [pp. 26 and 27]

B. Skyrms. On failing to vindicate induction. *Philosophy of Science*, 32(3):253–268, 1965. [p. 30]

B. Skyrms. Carnapian inductive logic for Markov chains. *Erkenntnis*, 35:439–460, 1991. [p. 61]

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. [pp. 79, 81, 87, and 91]

B. Skyrms. *Choice and Chance: An Introduction to Inductive Logic*. Wadsworth, 4th edition, 2000. [pp. 30 and 157]

R. I. Soare. *Recursively Enumerable Sets and Degrees: A Study of Computable Functions and Computably Generated Sets*. Perspectives in Mathematical Logic. Springer, 1987. [p. 25]

R. I. Soare. Computability and recursion. *Bulletin of Symbolic Logic*, 2(3):284–321, 1996. [p. 25]

R. I. Soare. The history of the concept of computability. In E. R. Griffor, editor, *Handbook of Computability Theory*, volume 140 of *Studies in Logic and the Foundations of Mathematics*, pages 3–36. Elsevier, 1999. [p. 25]

R. I. Soare. *Turing Computability: Theory and Applications*, volume 4 of *Theory and Applications of Computability*. Springer, New York, 2016. [pp. vi, 25, 35, 65, 74, 84, 190, and 194]

E. Sober. The principle of parsimony. *British Journal for the Philosophy of Science*, 32: 145–156, 1981. [p. 40]

E. Sober. *Ockham's Razors: A User's Manual*. Cambridge University Press, 2015. [p. 41]

R. J. Solomonoff. A formal theory of inductive inference. Parts I and II. *Information and Control*, 7:1–22, 224–254, 1964. [pp. 2, 6, 34, 36, 42, 50, 59, 61, 67, 82, 85, 86, 92, 102, 103, 113, and 131]

R. J. Solomonoff. Some recent work in artificial intelligence. *Proceedings of the IEEE*, 54 (12):1687–1697, 1966. [p. 86]

R. J. Solomonoff. The adequacy of complexity models of induction, 1975. Presentation at the Fifth International Congress on Logic, Methodology, and Philosophy of Science. [p. 196]

R. J. Solomonoff. Complexity-based induction systems: Comparisons and convergence theorems. *IEEE Transactions on Information Theory*, IT-24(4):422–432, 1978. [pp. 72, 196, and 200]

R. J. Solomonoff. The application of algorithmic probability to problems in artificial intelligence. In L. Kanal and J. Lemmer, editors, *Uncertainty in Artificial Intelligence*, volume 4 of *Machine Intelligence and Pattern Recognition*, pages 473–491. Elsevier, 1986. [pp. 71, 93, and 106]

R. J. Solomonoff. The discovery of algorithmic probability. *Journal of Computer and System Sciences*, 55(1):73–88, 1997. [pp. 60, 71, 83, 84, 85, and 103]

R. J. Solomonoff. The Kolmogorov lecture: The universal distribution and machine learning. *The Computer Journal*, 46(6):598–601, 2003. [pp. 83 and 97]

R. J. Solomonoff. Algorithmic probability: theory and applications. In F. Emmert-Streib and M. Dehmer, editors, *Information Theory and Statistical Learning*, pages 1–23. Springer, 2009. [pp. 71, 96, 97, and 106]

P. V. Spade. Ockham's nominalist metaphysics: Some main themes. In P. V. Spade, editor, *The Cambridge Companion to Ockham*, pages 100–116. Cambridge University Press, 1999. [p. 40]

T. F. Sterkenburg. Solomonoff prediction and Occam's razor. *Philosophy of Science*, 83(4): 459–479, 2016. [pp. 77, 96, and 121]

T. F. Sterkenburg. A generalized characterization of algorithmic probability. *Theory of Computing Systems*, 61(4):1337–1352, 2017. [pp. 63 and 207]

T. F. Sterkenburg. Putnam's diagonal argument and the impossibility of a universal learning machine. Submitted, 201x. [pp. 47, 63, 105, and 207]

A. W. Sudbury. Could there exist a world which obeyed no scientific laws? *British Journal for the Philosophy of Science*, 24(1):39–40, 1973. [p. 19]

F. Suppe. The search for philosophical understanding of scientific theories. In F. Suppe, editor, *The Structure of Scientific Theories*, pages 1–241. University of Illinois Press, Urbana, IL, 2nd edition, 1977. [p. 3]

P. Suppes. *Representation and Invariance of Scientific Structures*. CSLI Publications, Center for the Study of Language and Information, Stanford, CA, 2002. [pp. 47, 52, 56, 93, and 97]

R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. A Bradford Book. MIT PRess, Cambridge, MA, 1998. [p. 15]

M. Sznajder. What conceptual spaces can do for Carnap's late inductive logic. *Studies in History and Philosophy of Science Part A*, 56:62–71, 2016. [p. 54]

T. Tao. *An Introduction to Measure Theory*, volume 126 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2011. [p. 64]

C. G. Timpson. *Quantum Information Theory and the Foundations of Quantum Mechanics*. Oxford University Press, 2013. [p. 140]

A. M. Turing. On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, 2(42):230–265, 1936. [pp. 3, 25, 26, 34, 68, and 211]

J. B. M. Uffink. *Measures of Uncertainty and the Uncertainty Principle*. PhD Dissertation, Utrecht University, 1990. [p. 140]

V. A. Uspensky. Complexity and entropy: An introduction to the theory of Kolmogorov complexity. In O. Watanabe, editor, *Kolmogorov Complexity and Computational Complexity*, Monographs in Theoretical Computer Science. An EATCS Series, pages 85–102. Springer-Verlag, 1992. [p. 179]

V. A. Uspensky and A. K. Shen. Relations between varieties of Kolmogorov complexity. *Mathematical Systems Theory*, 29:271–292, 1996. [p. 179]

B. C. van Fraassen. *Laws and Symmetry*. Clarendon Press, Oxford, 1989. [pp. 4 and 52]

B. C. van Fraassen. The false hopes of traditional epistemology. *Philosophy and Phenomenological Research*, 60(2):253–280, 2000. [p. 33]

M. van Lambalgen. *Random sequences*. PhD dissertation, Universiteit van Amsterdam, 1987a. [pp. 38 and 181]

M. van Lambalgen. Von Mises' definition of random sequences reconsidered. *Journal of Symbolic Logic*, 52(3):725–755, 1987b. [p. 38]

M. van Lambalgen. Algorithmic information theory. *Journal of Symbolic Logic*, 54:1389–1400, 1989. [p. 156]

V. N. Vapnik. *Statistical Learning Theory*. Wiley, New York, NY, 1998. [pp. 15 and 41]

V. N. Vapnik and A. J. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications*, 16(2):264–280, 1971. Translation of the Russian original in *Teoriya Veroyatnostei i Ee Primeneniya*, 16(2): 264–279, 1971. [p. 41]

J.-A. Ville. Sur la notion de collectif. *Comptes rendus*, 203:26–27, 1936. [p. 38]

J.-A. Ville. *Étude critique de la notion de collectif*. Monographies des probabilités. Gauthier-Villars, Paris, 1939. [p. 38]

P. M. B. Vitányi. Algorithmic statistics and Kolmogorov's structure functions. In P. D. Grünwald, I. J. Myung, and M. A. Pitt, editors, *Advances in Minimum Description Length: Theory and Applications*, Neural Information Processing Series, pages 151–174. MIT Press, Cambridge, MA, 2005. [p. 130]

P. M. B. Vitányi and M. Li. On prediction by data compression. In M. van Someren and G. Widmer, editors, *Proceedings of the 9th European Conference on Machine Learning (ECML-97)*, volume 1224 of *Lecture Notes in Computer Science*, pages 14–30. Springer, 1997. [p. 122]

P. M. B. Vitányi and M. Li. Simplicity, information, Kolmogorov complexity, and prediction. In A. Zellner, H. A. Keuzenkamp, and M. McAleer, editors, *Simplicity, Inference and Modelling: Keeping it Sophisticatedly Simple*, pages 135–155. Cambridge University Press, 2002. [p. 42]

V. G. Vovk. Aggregating strategies. In M. Fulk and J. Case, editors, *Proceedings of the Third Annual Workshop on Computational Learning Theory (COLT90)*, pages 371–383, San Mateo, CA, 1990. Morgan Kaufmann. [pp. 42, 143, and 146]

V. G. Vovk. Probability theory for the Brier game. In M. Li and A. Maruoka, editors, *Algorithmic Learning Theory: Proceedings of the Eight International Conference (ALT '97)*, volume 1316 of *Lecture Notes in Computer Science*, pages 323–338. Springer, 1997. [p. 222]

V. G. Vovk. A game of prediction with expert advice. *Journal of Computer and System Sciences*, 56:153–173, 1998. [pp. 37, 42, 143, and 146]

V. G. Vovk. Competitive on-line statistics. *International Statistical Review*, 69(2):213–248, 2001a. [pp. 143 and 152]

V. G. Vovk. Probability theory for the Brier game. *Theoretical Computer Science*, 261:57–79, 2001b. A preliminary version appeared as Vovk (1997). [pp. 43, 71, 141, 150, and 156]

V. G. Vovk. The fundamental nature of the log loss function. In L. D. Beklemishev, A. Blass, N. Dershowitz, B. Finkbeiner, and W. Schulte, editors, *Fields of Logic and Computation II: Essays Dedicated to Yuri Gurevich on the Occasion of this 75th Birthday*, volume 9300 of *Lecture Notes in Computer Science*, pages 307–318. Springer, 2015. [pp. 141 and 142]

V. G. Vovk and C. J. H. C. Watkins. Universal portfolio selection. In P. L. Bartlett and Y. Mansour, editors, *Proceedings of the Eleventh Annual Conference on Computational Learning Theory (COLT98)*, pages 12–23. ACM, 1998. [pp. 43, 71, and 147]

F. Waismann. Logische Analyse des Wahrscheinlichkeitsbegriffs. *Erkenntnis*, 1(1):228–248, 1930. [p. 3]

C. Wallace. *Statistical and Inductive Inference by Minimum Message Length*. Information Science and Statistics. Springer, 2005. [p. 97]

J. Watkins. *Science and Scepticism*. Princeton University Press, Princeton, NJ, 1984. [p. 40]

S. Wenmackers and J.-W. Romeijn. New theory about old evidence: A framework for open-minded Bayesianism. *Synthese*, 193(4):1225–1250, 2016. [p. 109]

D. Williams. *Probability with Martingales*. Cambridge University Press, 1991. [p. 1]

I. Wood, P. Sunehag, and M. Hutter. (Non-)equivalence of universal priors. In D. L. Dowe, editor, *Proceedings of the Solomonoff Memorial Conference*, volume 7070 of *Lecture Notes in Artificial Intelligence*, pages 417–425. Springer, 2013. [pp. 74, 75, 93, 188, and 189]

S. L. Zabell. Johnson's sufficientness postulate. *The Annals of Statistics*, 10(4):1091–1099, 1982. Reprinted in Zabell (2005), pages 84–95. [pp. 58 and 82]

S. L. Zabell. Symmetry and its discontents. In B. Skyrms and W. L. Harper, editors, *Causation, Chance, and Credence: Proceedings of the Irvine Conference on Probability and Causation*, volume 1, pages 155–190. Kluwer, 1988. Page numbers refer to reprint in Zabell (2005), pages 3–37. [p. 125]

S. L. Zabell. Predicting the unpredictable. *Synthese*, 90:205–232, 1992. Reprinted in Zabell (2005), pages 217–242. [p. 15]

S. L. Zabell. *Symmetry and Its Discontents*. Cambridge Studies in Probability, Induction, and Decision Theory. Cambridge University Press, 2005. [pp. 56, 222, and 223]

S. L. Zabell. Carnap and the logic of inductive inference. In Gabbay et al. (2011), pages 265–309. [pp. 4, 23, 47, 54, 56, 59, 78, 82, and 97]

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. [pp. 6, 39, 66, 67, 68, 69, 83, 183, and 185]

*

Nederlandse samenvatting

In dit proefschrift onderzoek ik de theoretische mogelijkheid van *universele voorspellers*: voorspelmethodes die in zekere zin succesvol zijn in alle mogelijke gevallen. Daartoe beschouw ik de formele specificatie van een voorspelmethode die teruggaat naar Solomonoff (1964) en Levin (1970), en die zich onderscheidt door de rol van *effectieve berekenbaarheid* en de associatie met een *eenvouds*voorkeur.

De eerste hoofdlijn van dit proefschrift is de interpretatie van de Solomonoff-Levindefinitie als een universele voorspelmethode. Mijn vertrekpunt is de relatie van dit voorstel met Carnaps onderzoeksprogramma van inductieve logica; in het bijzonder bestudeer ik een invloedrijk argument van Putnam tegen Carnaps programma, een wiskundig bewijs dat de onmogelijkheid van een universele voorspelmethode moet aantonen. Ik presenteer de Solomonoff-Levinvoorspellers als het natuurlijke resultaat van een uitdrukkelijke poging om een klasse van effectieve voorspellers bloot te leggen die immuun is voor Putnams bewijsprocedure. Dit maakt, zo lijkt het, de weg vrij voor een Reichenbachiaanse interpretatie van de Solomonoff-Levinvoorspellers als optimaal onder alle mogelijke voorspellers.

Mijn conclusie is echter negatief: deze interpretatie houdt geen stand, en de reden is een discrepantie tussen de effectiviteitsniveaus van de Solomonoff-Levin*maten* en de Solomonoff-Levin*voorspellers*. Deze laatste zijn niet vatbaar voor de gewenste interpretatie, en dit blijkt al te volgen uit Putnams oorspronkelijke bewijs.

De tweede hoofdlijn van dit proefschrift is de associatie van het Solomonoff-Levinvoorstel met het ongrijpbare concept van eenvoud. Ik analyseer het idee dat de Solomonoff-Levinvoorspellers niet slechts een *formalisering* geven van een eenvoudsvoorkeur in voorspelling, het principe van Occams scheermes, maar ook een *rechtvaardiging* daarvan. Ik bespreek de relevante notie van eenvoud als comprimeerbaarheid, en reconstrueer het geopperde argument voor een kentheoretische rechtvaardiging die circulariteit weet te vermijden. Tevens evalueer ik Vovks gerelateerde formele notie van voorspelcomplexiteit als een notie van de intrinsieke moeilijkheid van datareeksen.

Mijn conclusies zijn opnieuw negatief. De voorgestelde rechtvaardiging gaat niet op, juist omdat de relevante eenvoudsvoorkeur al een specifieke inductieve aanname vertegenwoordigt. Daarbij beargumenteer ik dat de betreffende definitie van eenvoud als comprimeerbaarheid niet kan overtuigen als een objectieve formalisering van een eenvoudsvoorkeur in voorspelling. Hoewel, tenslotte, de notie van voorspelcomplexiteit een meer rechtstreekse en daarom ogenschijnlijk minder problematische interpretatie kent, schiet deze toch tekort als intrinsieke-complexiteitsmaat.

*

Biography

The author was born more than thirty years ago in the town of Purmerend. Nevertheless, he completed a BSc in Artificial Intelligence (VU University Amsterdam, *cum laude*), a MSc in Logic (ILLC, University of Amsterdam, *cum laude*), a MSc in History and Philosophy of Science (Descartes Center, University Utrecht, *cum laude*), and a MA in Philosophy (University of Groningen, as part of the subsequent PhD project).

During his PhD project he was based at the Centrum Wiskunde & Informatica (CWI, the Dutch national research center for mathematics and computer science) in Amsterdam, initially in the Algorithms and Complexity group and later in the newly formed Machine Learning group, as well as the University of Groningen, in the Department of Theoretical Philosophy.

He continues his research as a postdoctoral fellow at the Munich Center for Mathematical Philosophy, LMU Munich.

*