Limits of Solomonoff Induction

MSc Thesis *(Afstudeerscriptie)*

written by

Joel Artturi Saarinen
(born July 15, 1998 in Helsinki, Finland)

under the supervision of **Dr. Francesca Zaffora Blando** and **Dr. Aybüke Özgün**, and submitted to the Examinations Board in partial fulfillment of the requirements for the degree of

MSc in Logic

at the *Universiteit van Amsterdam*.

Date of the public defense:	**Members of the Thesis Committee:**
June 25th, 2025	Dr. Benno van den Berg (chair)
	Dr. Francesca Zaffora Blando (co-supervisor)
	Dr. Aybüke Özgün (co-supervisor)
	Prof. Dr. Johan van Benthem
	Dr. Tom Sterkenburg

INSTITUTE FOR LOGIC, LANGUAGE AND COMPUTATION

Abstract

Solomonoff's theory of inductive inference is often regarded as a gold standard for formal theories of learning. However, several results have shown that Solomonoff's predictor fails to converge in a greater variety of circumstances than originally thought, exemplified by Hutter's and Muchnik's result that this learning failure occurs for a specific type of data string that Martin-Löf random relative to the uniform Lebesgue measure, when the latter is assumed to be the underlying distribution generating the data, and Hutter's and Lattimore's result that for any predictor, there exists a certain type of data string relative to the uniform Lebesgue measure for which convergence fails. This thesis aims to expand upon these results by showing that the argument given by Hutter and Muchnik holds for an arbitrary computable measure satisfying a certain boundedness assumption, and that one of the arguments given by Hutter and Lattimore holds for this same measure, delivering an additional blow to the Solomonoff inductor as being a gold standard for learning. In light of these results, the thesis also offers a brief philosophical discussion on how, in light of further failures of such seemingly optimal learners, and the skepticism of being able to find an optimal learner in the first place, whether there might be an alternative standard by which to evaluate optimality.

Acknowledgements

Those close to me know that the Master of Logic has been exceptionally challenging, for reasons not having only to do with the rigor of the program, but many factors outside of it, which is precisely because of this that I feel an overwhelming sense of gratitude for everyone with whom I've shared this phase of my journey in Amsterdam in different ways. This thesis not only marks the end of a certain academic chapter, but much more — what feels like a lifetime packed into a short span of years.

Firstly, there are my thesis supervisors. Francesca, thank you so much for your patience, flexibility, and most importantly, your enthusiasm for the subject of algorithmic randomness itself, which made the research experience all the more rewarding for me. I really appreciated the time you took in the beginning to help me identify a meaningful project in this area, based on the questions I wanted to ask, and balancing the setting up relevant steps for me so that it truly felt like I was making my own discoveries, while also never making me feel like I was lost (while also, not to mention, doing this all in a remote capacity). I have definitely felt humbled in that I came to understand far better what constitutes good, rigorous research, but at the same time, I feel I have also come out from the other end even more confident in my potential as a scholar. While I'm not sure exactly how much my PhD will draw upon the topics we have worked with, this was a special experience, and I very much look forward to pursuing these topics and aligned ones in the future in some form.

To Aybüke, thank you also for your support for this thesis, and also for the general guidance through the years. You were consistently encouraging when it came to my many zany project ideas, even if you didn't end up being directly involved in them, which meant a great deal to me. The Bayesian epistemology course you taught during my first year also served as a major inspiration for this thesis, which I'm grateful for as well. I also hope to expand the results in this thesis to a more logical setting with you in the future.

To Dean McHugh, Eline de Jong, and Daira Pinto Prieto, I appreciated discussions with each of you at different points of my experience in the MoL about logic and academia in general, which all helped orient my thinking about the future and enjoy my time during this degree.

Then, there are my friends. First, to my 91B flatmates. Paulius, thank you for all the burrito nights, our adventures across the European continent, our conversations about most topics under the sun, our fun logic papers and side quests, and much more. I sincerely hope (yet somehow, still know for certain) that these will continue into the future. Hannah, thank you for always being down to go on a beach trip when I needed a break from the intensity of our program, and for always being down to teach me about any of your numerous new hobbies on a whim.

Then, there is the Bijlmer gang at large. Sarah, thank you for always managing to help me see the world through a more curious, playful angle, especially when I was in need of a shift in headspace, and for helping me embrace my quirks more than ever. Ștefan, thank you for all the gains you helped me make — physical, intellectual, and spiritual. I will miss passing you going in the opposite direction on our bikes just steps from our building (while seemingly never bumping into each other anywhere more convenient). Mădălina, thank you for the many fun film outings and edifying discussions that followed. Horia, thank you for all of the whimsical,

contemplative talks over the years, and for always being ready push my views to their limit. Lamia, thank you for showing me a new side of Amsterdam, and for bringing my attention to issues and perspectives that I didn't have the opportunity to think about in connection with my studies as much. To those in the world outside the Bijlmer, to Alyssa, Martijn, Rodrigo, Andor, Alex Lund, Alex Stan, Liam, Brendan, and Andrea: thank you for your love, support, and presence during these past few years as well. To Max and Zach, even though neither of you lived physically near me during the past few years, I'm incredibly grateful for our semi-regular outings to see each other, which were always tremendously rejuvenating, and for maintaining our friendships over time.

To Adrian, and the rest of the guys at Gracie Jiu Jitsu, thank you for contributing to my growth in more ways than I can count, and for serving as a refuge where I could always feel good about myself no matter what was going on in other parts of my life. And of course, I will always continue to respect the tap.

Next, my family. Suuret kiitokset Nanalle ja Mikolle taloudellisesta sekä henkisestä tuesta viime vuosien aikana. Ilman teitä, en olisi pystynyt kokemaan niin antoisaa kokemusta maisteriopintojen muodossa joka on nautinnollisuuden lisäksi samalla luonut pohjan pidemässä tähtäimessä juuri sellaiselle elämälle, josta voisin vain haaveilla. Thank you also to Mom and Oliver for everything – I could not have done this without you guys.

Finally, to Lucrezia: thank you for all of the countless, spontaneous conversations that go on for hours, for the stickers that never fail to cheer me up, and for your love. I'm very excited for what the future holds.

Contents

1 Introduction

Truth is ever to be found in simplicity, and not in the multiplicity and confusion of things.

Isaac Newton

1.1 General Introduction

The problem of induction is arguably one of the oldest, most intriguing, and yet most pernicious riddles in all of philosophy. Hume, to whom the first mention of the problem in 1748 is often attributed, describes it thus:

> *All reasonings concerning matter of fact seem to be founded on the relation of cause and effect. [...] But when we look beyond the present testimony of our senses, or the records of our memory, the conclusions are not founded on reasoning or any process of the understanding.* [Hume (1748), pg. 35]

In more colloquial language, there is no rational principle that justifies the seemingly common process of making a general, universal claim – meant to hold across time – on the basis of finite observations. Even if we have seen the sun rise in the east every day of our lives, there is nothing wrong with saying that it *could* rise in the west tomorrow. Scientific theories that supply explanations for the various physical phenomena of the universe can always be proven wrong, as has been seen plenty of times throughout history. To some, such as Popper, this falsifiablity is in fact a necessary condition for a theory to be worth taking seriously, to be considered scientific [Popper (1959)], illustrating further skepticism about induction.

At the same time, to not use induction on the basis of it not being rationally justified would make it impossible to get through the day. Imagine, for instance, someone deciding to simply not reason about whether it will rain when there are dark clouds in the sky because any such correlations were formed from (faulty) inductive inferences, only for them to eventually go on a walk and get drenched within minutes of leaving the house. Imagine someone not making any plans about the future because they think they could lose their job tomorrow and have their financial situation destabilized, in spite of them receiving continual good praise at work from higher-ups. Such intuitions beg for some kind of justification of inductive reasoning.

A natural solution is to posit that there are different types of rationality, as we conceive of it in the ordinary sense of the word, of which induction violates only one. To ground our discussion in something concrete, we might broadly term an individual as *rational* to the extent that they engage in behavior that promotes their general well-being, though indeed, it seems philosophers have shied away from these kinds of general definitions in favor of ones specifying what it means to be rational with respect to certain activities [Rysiew (2020)]. For instance, defining rationality directly in either a *theoretical* and *practical* context seems to be the more historically common practice, with theoretical rationality being understood as rationality as pertaining to the logical consistency of one's beliefs, and practical rationality as being the embodiment of practical attitudes – intentions, desires, and preferences, that are executed upon in action – promoting an individual's well-being [Wedgwood (2021), pg. 138]. To be clear, Hume's concerns about induction go beyond simply the suspect move of forming a general conclusion from finitely many observations, but dividing rationality into types may assuage at least this part. To the extent that theoretical rationality is concerned with this kind of traditional, logical consistency, induction may be seen as theoretically irrational but still practically rational, for failing to use it would pose difficulties for an individual in their daily lives, as we have seen.

Yet, even if one is sympathetic to one such possible (practical) justification of induction, there are still many other questions that surround this process. Chief among these questions would be: what sorts of inductive inferences are permissible? Intuitively, we might find it justifiable for someone to infer from dark clouds that it will soon rain, but less justifiable to infer that these dark clouds portend the thunderstorm of the century, even if dark clouds have always preceded both the case of mere rain and a violent thunderstorm.

This is perhaps one angle of a broader question: what is an ideal theory of inductive inference? Addressing this would not only help us separate permissible inferences from overly faulty ones, but also reveal to us just how fine-grained we must be about induction in the first place. Must such a theory specify, numerically, how much one's beliefs should shift given a certain piece of evidence, or must it only specify a set of heuristics, good enough for an agent to employ in everyday situations?

All these questions demonstrate the rich set of considerations surrounding the phenomenon of inductive inference and learning more broadly. As desirable as such an investigation done properly might be, this thesis aims not to address all of these in one fell swoop, but rather zooms in on one aspect of this both meta and object-level discussion about induction. Here, we focus on one such aforementioned theory of inductive inference and learning more broadly, and modestly resume certain previous efforts to push it to its limit. Specifically, we examine Solomonoff's formal theory of inductive inference, and show that predictors implementing the theory – Solomonoff inductors – fail to perform as expected in an even wider variety of settings than previously thought [Hutter and Muchnik (2007), Lattimore and Hutter (2015)]. As

these Solomonoff inductors are considered in some sense to embody an ideal inductive learner, such additional failures prompt a continuation of a wider discussion concerning the prospect of modeling an ideal (inductive) learners, and what this might represent in the first place. After showing these main negative results, we contribute a bit more to this discussion by introducing an alternative notion of ideality and seeing whether the Solomomonoff inductor adheres to it.

To begin, we will pick up from where the previously introduced questions about ideal theories of induction left off. Specifically, we will briefly examine some of the different formal models for induction that have been proposed across time, in order to both contextually situate and motivate the introduction of Solomonoff's theory and the main results to follow.

1.2 Specific Background and Problems

The history of formal modeling of inductive inference may have a few different origins, depending on one's vantage point. Despite having introduced the problem of induction into philosophical circles in his time, Hume did not attempt to craft a framework that could model the accuracy of inductive inferences – perhaps predictably, given his overall skepticism about induction. Only in the next century were seemingly greater steps taken to model inductive inference. An arguably important step is to identify inductive inferences with probabilities. Intuitively, to believe that it is *likely* to rain given that there are dark clouds in the sky (a probabilistic statement) is akin to believing that it will rain, given that there are dark clouds in the sky and that it has rained every finite amount of times, or most of them, that there have been dark clouds in the sky (an inductive inference) [1]. Early such sentiments are reflected in the writings of Laplace in 1812, assuming one interprets the following "common sense" inferences as ones that aren't quite certain but almost ought to be, with him claiming:

> *One sees in this essay that the theory of probabilities is basically only common sense reduced to a calculus. It makes one estimate accurately what right-minded people feel by a sort of instinct, often without being able to give a reason for it.* [Laplace (1995), pg. 124]

Later on in the decade, the identification of inductive inferences with probabilistic inferences started to truly take hold. Peirce was among the first, in 1867, to propose a basic formal schema for induction, along with a justification for induction in the long run based on the idea that even an incremental improvement in the accuracy of statements over time would lead to more accurate premises to be used in later inferences [Jessup (1974), pg. 226] – a tone strikingly different to Hume's a century prior.

[1] To be clear, the study and development of probability itself predates Hume, but connecting probability to inductive inference more explicitly came after.

A more extensive model of inductive inference was invented by Carnap almost 80 years later [Skyrms (1996), pg. 321]. Carnap's system of inductive logic sought to not only expand upon Peirce's initial effort to formally specify the conditions necessary for an inductive inference to be made, but also the precise degree to which evidence entails a hypothesis. More specifically, Carnap introduced confirmation functions which, taking into account the number of observations of relevant evidence, how quickly evidence overrides priors, and the number of different types of possible evidence, gave a probability of how much the total evidence supports a hypothesis [Skyrms (1996), pg. 322]. Additionally, although the confirmation functions were based on probabilities, confirmation itself was intended by Carnap to be a *logical* relation between statements [Skyrms (1996), pg. 334], with a key difference between the greater inductive logic system of which these functions were a part and deductive logics being that entailment of one statement from another was not certain, but rather a matter of degree expressed by the confirmation function.

Carnap's work also led to a more widespread acceptance of Bayesian reasoning, a framework whose seeds were sown almost two centuries earlier, but became more relevant during the early 20th century quest of automating scientific reasoning (to which Carnap also contributed) [Zabell (2004), pg. 305, Sterkenburg (2018), pg. 4]. More precisely, Thomas Bayes, a clergyman doubling as a mathematician, came up with a theorem that expressed the probability of a hypothesis being true, given some evidence. This rule takes into account an agent's initial probability of the hypothesis being true (termed the *prior* probability) before seeing the evidence, which means that the output of the rule can be interpreted as an "update" in the probability of the hypothesis. Together with Lewis' Principal Principle, which asserts that an agent is rational to the extent that their credences of propositions are aligned with their objective probabilities, Bayes' rule can be seen as prescribing inference rules for a rational agent that has degrees of belief in propositions, as opposed to all-or-nothing belief [Lin (2022)]. From this, the field of Bayesian epistemology was born, which looks at the broader consequences of Bayes' rule in an effort to describe a more complete formal theory of inductive reasoning [Hartmann and Sprenger (2010)]. Of course, the framework is not completely airtight. For instance, it is not at all obvious how an agent should choose the aforementioned prior probabilities used in the update rule [Titelbaum (2022), pg. 447]. Regardless, Bayesianism remains a popular, unified account of scientific reasoning [Sterkenburg (2018), pg. 4], and thus serves as a sound framework for inductive inference as well.

Finally, even though they may not be traditionally conceived of as such, the advent of artificial intelligence also marked an important step forward in modeling inductive learning – if not in a novel, more mechanized way. Inspired primarily by the brain's circuitry and the works of Turing rather than the aforementioned more philosophical leaps, Pitts and McCulloch developed the first mathematical model of a neural network in 1942, which laid the foundation for the more

advanced, multi-layered networks that came later and power much of the digital infrastructure in our world today [Christian (2020), Prologue]. As AI systems are fundamentally tasked with performing the same inferential learning and optimization under uncertainty that humans are, triumphs in creating capable systems could well be seen as successes of accurately modeling inductive inference [2].

Indeed, the development of statistical learning theory, in response to the growing popularity of neural networks originating from Pitts' and McCulloch's work, aimed to provide sound theoretical foundations for the learning properties of more advanced models, and ended up becoming more reminiscent of the Carnapian and Bayesian frameworks previously mentioned. In the simplest sense, this paradigm involves a learner (the inductive agent, represented as a learning algorithm) that is given a training data set (a set of points $(x, y) \in \mathbb{R}^2$ with $x \in X, y \in Y$ and $X, Y \subseteq \mathbb{R}$) sampled from some unknown underlying distribution, and tasked with outputting a "predictor" that comes up with a hypothesis to explain the data (a function $h : X \to Y$), as accurate as possible with respect to the underlying distribution [Shalev-Shwartz and Ben-David (2014), pg. 35]. The resulting predictor function could be interpreted as giving probabilities for different outcomes given the training data it has seen, in much the same way that Carnap's confirmation functions provide a degree of support for a hypothesis given evidence. Thus, AI systems could well be regarded as inductive learners, and the combination of algorithms and programs that underlie them could be regarded as the underlying inductive inference or learning frameworks.

Having better acquainted ourselves with different models for inductive learning, we are inevitably again tugged by the question of what model of induction seems most optimal. Of course, to address such a question properly, one must dig a bit deeper into the notion of optimality. Intuitively, it depends on at least a few different factors, one of which is context. If one's motivation is to be able to implement some sort of formal inference framework for a task that requires computational efficiency and easy practical implementation, to the extent that one considers AI systems learning frameworks, they seem best suited for the purpose. However, as AI systems are known to be black boxes, in the sense that it is difficult to understand specifically why they form the conclusions they do [Doshi-Velez and Kim (2017)], in contexts where it is important to know exactly how an inference was made, frameworks with clear rules for what is required for a given inference, or degree of support for some hypothesis (such as Carnap's theory), may be more desirable.

Yet, undoubtedly, ideality also has something to do with a more theoretical accuracy as well. By theoretical accuracy, we are referring to a learner's ability to make accurate predictions,

[2] Of course, it may not be that the successful execution of a task was *caused* by an agent's accurate inferential machinery, since it could also be influenced by the agent's decision theory under uncertainty. This is also, of course, merely one way by which to judge the accuracy of an inferential framework – which will become an important discussion later on.

independent of any specific concerns about implementability. One proposed definition for ideality by Sterkenburg focuses precisely on this accuracy criterion, phrased in terms of accuracy as relative to other learners. Specifically, Sterkenburg posits an optimal method as one that will come to predict at least as successfully as any other one, no matter what the world may do to interfere with the learner's predictions [Sterkenburg (2018), pg. 30]. For now, in discussing optimality, we adopt Sterkenburg's definition of optimality, for it seems broad enough to cover many possible prediction methods and settings (as it does not make any assumptions about underlying distibutions generating learning data).

In settings where we believe that the likelihoods of events follow some underlying distribution, there is one additional learning framework that may be considered optimal. Namely, Solomonoff's learner, first introduced in 1964, is a learner whose predictions were proved rigorously to converge to an assumed underlying distribution for a wide variety of sequences sampled from this distribution. Although this does not guarantee convergence for any particular sequence, this broadly good performance is naturally appealing, and the framework itself has even been considered a "gold standard" of inductive learning [Rathmanner and Hutter (2011), pg. 63].

Besides this convergence property, Solomonoff's predictor is also claimed to have other advantages. In our brief exploration of Bayesianism, we saw that an update rule is dependent on the choice of a specific prior probability, for which there is no obvious rule for choosing. Solomonoff's predictor has an in-built way of choosing a prior, electing to assign weight to hypotheses roughly in accordance with their simplicity, while still considering all possible hypotheses in making a prediction. Furthermore, its predictions purportedly are good as those of any other method in the long run [3].

Still, there are disadvantages to Solomonoff induction. For one, it is uncomputable, and not implementable in a more practical setting. Computable approximations of Solomonoff inductors still exist and partially remedy this concern, as approximations have been used to derive useful performance bounds for certain prediction problems [Leike and Hutter (2015)]. In spite of this, there are other concerns that hamper these learners. For instance, results in Hutter and Muchnik (2007) (stated here) show that Solomonoff's predictors fail to converge for a certain type of data sequence when the underlying measure is a uniform Lebesgue measure, while Lattimore and Hutter (2015) (stated here) prove an almost more extensive result that in fact for every possible type of Solomonoff predictor, there is a type of data sequence for which it fails to converge to the same underlying uniform Lebesgue measure. This may be a meaningful downside depending on how ubiquitous and useful prediction problems involving these specific types of sequences are. Moreover, Sterkenburg (2018) questions and eventually

[3] For those inclined to skip ahead for the technical reasons for this claim, we refer the reader to the definition of a universal semimeasure.

argues against the claim that Solomonoff's universal learner is optimal, in the sense that we have been considering, in the first place [4].

With appropriate context in mind, we now define our plan more precisely.

1.3 Roadmap

This investigation will seek to take a new step in addressing these tensions about the optimality of Solomonoff predictors. We first extend the results of Hutter, Muchnik, and Lattimore, further challenging the theoretical reach of Solomonoff predictors. Then, we introduce an alternative standard of optimality by which to evaluate inductive learning frameworks more generally, and then perform a brief analysis of Solomonoff predictors under this standard as a test case, in an effort to see whether this assessment might differ from that under Sterkenburg's definition of (universal) optimality. We begin by introducing relevant notation for understanding the technical results that follow. We then introduce the results that our main results are based on – that is, the main results of Solomonoff (1978), Hutter and Muchnik (2007), and Lattimore and Hutter (2015). Then, in Section 3 and Section 4, we present the novel main results, and in Section 5, we analyze the optimality of the results, before concluding the investigation in Section 6.

[4] In fact, Sterkenburg (2018) goes further and argues that no such universally optimal learner can exist in the first place. However, this argument does not affect our results here.

2 Preliminaries

Before diving into the main technical material, we provide some relevant technical context, notation and definitions. We first clarify that the type of inductive learning we work with involves learning to correctly predict the next symbol of an infinitely long sequence of symbols, given a finitely long string symbols of from the sequence that the learner has seen so far. Formally, for a sequence of symbols $x = x_1x_2 \dots x_{n-1}$, a learner is to predict the next symbol x_n in the sequence, and learning competence is intuitively measured by whether it correctly predicts the next symbol. This task is termed more canonically as a *sequential prediction task*.

To be even more fine-grained, we specifically assume that the string is composed of 0s and 1s. We aim to assess the degree to which a learner's probability of a given digit being the next one differs from the probability of that digit being the next according to some other computable probability measure which is assumed to be the "true" distribution generating the binary sequence – that is, where the probability of a given symbol being next equals the probability assigned to this event by said computable probability measure. This is the context in which Solomonoff's original, seemingly ideal learner is presented, as a learner that whose conditional probabilities of a given digit appearing next are eventually equal to those of this true, underlying distribution, for many binary sequences. We soon clarify the structure of this learner more precisely and present the original theorem, after introducing notation and additional definitions.

As mentioned, Solomonoff's results do not establish that it achieves this convergence for all possible sequences, and thus makes no claims about convergence holding for any *particular* sequence. One class of sequences for which it is worthwhile to assess this convergence are sequences that are *Martin-Löf random* relative to the true underlying measure. We define this notion more rigorously soon, but for now, we can interpret this as simply meaning that a sequence being Martin-Löf random relative to some measure means that it passes all conceivable tests for randomness – for example, the law of large numbers, the law of iterated logarithm – and so on. Intuitively, a sequence passing these tests and thus being Martin-Löf random means that it is "typical" according to the underlying measure, a random sample drawn from it. As it turns out, under a typical way that Solomonoff predictors are formally represented, Martin-Löf randomness can be defined in terms of the Solomonoff predictor itself, as opposed to solely relative to the underlying measure (Hutter and Muchnik (2007), pg. 2). This lends a more objective notion to randomness. Again, we defer discussion of further details to the presentation of the results, but for now, we note that our results aim to show non-convergence specifically for

certain Martin-Löf random sequences relative to the underlying true measure. As such, we note that we work with notions seen in algorithmic randomness and algorithmic information theory, which are concerned with the study of such Martin-Löf random sequences.

Overviews of algorithmic information theory, relevant for many of the results presented, may be found in Li and Vitányi (2008) and R. G. Downey and Hirschfeldt (2010), and a table of notation may be found in Appendix B. Appendix A contains relevant, non-novel auxiliary results, which are referenced as needed in the main results sections. For general formatting of these preliminaries and notation, we draw upon Hutter and Muchnik (2007) and Lattimore and Hutter (2015).

2.1 Notation

General. The natural and real numbers are denoted by $\mathbb{N}, \mathbb{R}$. Logarithms are taken with base 2. A real $\theta \in (0,1)$ has entropy $H(\theta) := -\theta\log\theta - (1-\theta)\log(1-\theta)$.

Strings. A finite binary string is a finite sequence $x = x_1x_2x_3\ldots x_n$ with $x_i \in \{0,1\}$, $1 \leqslant i \leqslant n$, with its length denoted as $\ell(x)$. The set of all finite binary strings is denoted as $\{0,1\}^*$. An infinite binary string ω is an infinite sequence $\omega = \omega_1\omega_2\omega_3\ldots$, and the set of all infinite binary strings is denoted as $\{0,1\}^\infty$. We use $\{0,1\}^n$ to represent the set of all binary strings of length n. The empty string of length zero is denoted by ε (distinct from $1 > \epsilon > 0 \in \mathbb{R}$, which will be used later). Substrings of $x \in \{0,1\}^* \cup \{0,1\}^\infty$ are denoted by $x_{s:t} := x_s x_{s+1}\ldots x_{t-1}x_t$, where $s, t \in \mathbb{N}$ and $s \leqslant t$. That is, we interpret $x_{s:t}$ as the string that exists from index s in x to index t in x, inclusive. If $s > t$, then $x_{s:t} := \varepsilon$. We also employ the shorthand $x_{<t} := x_{1:t-1}$ to represent the initial segment of x of length $t-1$. For $x, y \in \{0,1\}^*$, $xy \in \{0,1\}^*$ is their concatenation. We use $[\![x]\!] := \{\omega \in \{0,1\}^\infty : \omega_1\ldots\omega_{\ell(x)} = x\}$ to denote the cylinder set of x: the set of all infinite binary strings that begin with x.

Measures & Semimeasures. We let $\mathcal{B}$ denote the Borel σ-algebra generated by $[\![x]\!]$, the cylinder sets of x. We define a semimeasure μ on $(\{0,1\}^\infty, \mathcal{B})$ as $\mu : \mathcal{B} \to [0,1]$ which assigns values in $[0,1]$ to cylinder sets of every string $x \in \{0,1\}^*$, satisfying $\mu(\varepsilon) \leqslant 1$ and $\mu(x) \geqslant \mu(x0) + \mu(x1)$ for all $x \in \{0,1\}^*$. As the collection of cylinder sets above forms a ring that generates $\mathcal{B}$, by Carathéodory's extension theorem, μ can be extended to a semimeasure on $\mathcal{B}$, so more precisely we consider μ a semimeasure $\mu : \mathcal{B} \to [0,1]$ satisfying the aforementioned inequalities. It is a measure if both conditions hold by strict equality. For $b \in \{0,1\}$ and $x \in \{0,1\}^*$, $\mu(b|x) := \frac{\mu(xb)}{\mu(x)}$ is the μ-probability that x is followed by b (also known as the conditional probability of x given b), for $\mu(x) > 0$.

Computability. Furthermore, we call a function f computable if there is an effective procedure that returns the exact value of $f(x)$ in a finite time for every input $x \in \{0,1\}^*$. We call a function f

lower semicomputable if there exists a computable sequence of rational-valued functions $\{f^t\}_{t\in\mathbb{N}}$ such that $f^t(x) \leqslant f^{t+1}(x)$ for all x, t (non-decreasing in t) and $\lim_{t\to\infty} f^t(x) = f(x)$, ie. that f can be approximated from below by a series of computable approximations.

2.2 Theory of Algorithmic Randomness

Given our aim to show non-convergence for certain types of aforementioned Martin-Löf random sequences, we now both formally define this notion and provide some context concerning the theory of algorithmic randomness, the framework devoted to the study of randomness.

For the sequential prediction task discussed earlier in this section, there are intuitively some finite strings for which it is more difficult to predict the next digit than others. For example, given the initial segment of a string:

$$0101010101010101010101...$$

one could seemingly be justified in guessing the next digit to be 0, given that the segment up to the end is generated by alternating 0 and 1, starting from 0. We may also be tempted to say that the sequence is not very *random*, in the common sense of the word, since the rule to generate the string seems to jump out quite clearly. On the other hand, for the string:

$$1101010101111101010001...$$

it may be more difficult to guess the next digit, as there does not seem to be as discernible of a pattern generating the appearance of 0s and 1s. One may also be tempted to term the latter sequence to be *more* random than the first, as it seems to be generated on a whim, adhering to no obvious rule. However, these are still merely intuitions about randomness – not proper, mathematical judgments about one sequence being more random than the other.

The theory of algorithmic randomness aims to make precisely such considerations more concrete. It aims to address questions concerning what counts as a formally random sequence, whether there are degrees of randomness, among others. A more extensive overview of the development of the theory of algorithmic randomness can be found in R. Downey and Hirschfeldt (2018), but for our purposes, we continue with a more modest discussion.

Namely, we note that arguably the first robust definition of randomness for such sequences came from Martin-Löf in 1966, based on a sequence passing a series of tests meant to gauge a sequence's randomness. Earlier tests (such as those by von Mises) existed before Martin-Löf's as well, and focused mainly on narrowing down which sequences could be considered random relative to the uniform measure [R. Downey and Hirschfeldt (2018), pg. 3]. Intuitively, one way that randomness might be measured is by trying to measure how many patterns are contained in the string. For instance, one pattern would be that a zero occurs at every third index of the

string, starting at index 0. One could test for this by first looking at whether a zero occurs at index 0, then whether a zero occurs at index 3, then at 6, and so on. The test at each index would correspond to a test at a given level, and a sequence that fails the test at all levels would not be considered random. Yet, Martin-Löf's extension of this idea was to generalize this process, and defined a sequence as random only if it passes *all* such tests [R. Downey and Hirschfeldt (2018), pg. 3]. These tests may be defined as follows:

Definition 2.1. (Martin-Löf Randomness).

1. Let $\{U_n\}_{n\in\mathbb{N}}$ be a sequence of uniformly Σ^0_1 classes satisfying $\mu(U_n) \leqslant 2^{-n}$ for all $n \in \mathbb{N}$. Such a sequence is called a *sequential* μ*-Martin-Löf test*.

2. A sequence $\omega \in 2^{\mathbb{N}}$ is μ-Martin-Löf random *if and only if* there is no sequential μ-Martin-Löf test $\{U_n\}_{n\in\mathbb{N}}$ such that $\omega \in \bigcap_{n\in\mathbb{N}} U_n$.

We now introduce the rest of the definitions to motivate the main results to come, including the notion of a Martin-Löf random sequence, which follows from the definition above.

2.3 Other Definitions

As we will see, we interpret Solomonoff's learner as a type of universal semimeasure, which we define in a more general form below.

Definition 2.2 (Universal Semimeasure). A semimeasure M is called *universal* in a class of semimeasures $\mathcal{M}$ if, for all semimeasures $\nu \in \mathcal{M}$, there exists a constant $c_\nu > 0$ such that $M(x) \geqslant c_\nu \cdot \nu(x)$ for all $x \in \{0,1\}^*$.

This universality thus comes from the fact that M "dominates" all the other measures in the respective class, in that it is greater than or equal to every other element of that class multiplied by some constant. As implied in the definition, the concept of a universal semimeasure can be modified based on the type, or class of semimeasures that we want to discuss. For example, for universal lower semicomputable semimeasures, the ones which are used to represent Solomonoff's learner in the original theorem, these would have to dominate every other lower semicomputable semimeasure multiplied by a constant. Sticking with this theme, in one of the results we extend, the universal lower semicomputable semimeasure is represented as a "mixture", or weighted sum, of all these other lower semicomputable semimeasures it dominates. This equivalence follows from a result in Zvonkin and Levin (1970) establishing that the set of all lower semicomputable semimeasures is recursively enumerable. This mixture is defined as follows.

Definition 2.3 (Universal Mixture). For a universal lower semicomputable semimeasure M, we may define it as a *universal mixture* of the other lower semicomputable semimeasures it dominates in the following way:

$$M(x) := \sum_{i \in \mathbb{N}} w_i \nu_i(x)$$

Furthermore, $w_i \in [0, 1]$ represents the initial weight assigned to the ith lower semicomputable semimeasure ν_i, reflecting its plausibility before observing any data. We define the prior of this ith lower semicomputable semimeasure as:

$$w_i(x) := \frac{w_i \nu_i(x)}{M(x)}$$

It determines how much influence ν_i has in the overall universal mixture $M(x) = \sum_{i \in \mathbb{N}} w_i \nu_i(x)$. This prior is useful because it allows us to define the conditional probability of a given bit according to M, based on what has been seen so far, by substituting the definition of the prior into the definition of the mixture above:

$$M(b|x) = \sum_{i \in \mathbb{N}} w_i(x) \nu_i(b|x)$$

Having defined the concept of a universal semimeasure, we can now define the notion of a Martin-Löf random sequence, appealing to an alternative, non-test-based definition of Martin-Löf randomness given in [Levin (1973)]:

Definition 2.4. (Martin-Löf random sequence) A string $\omega = \omega_{1:\infty}$ is μ-*Martin-Löf random* (μ.M.L.) if and only if there is a constant $c < \infty$ such that $M(\omega_{1:n}) \leqslant c \cdot \mu(\omega_{1:n})$ for all n.

We also use the concept of supermartingales in some of the non-convergence results to come.

Definition 2.5. (Supermartingale) Let μ be a computable probability measure. A μ-*supermartingale* is a function $m : \{0,1\}^* \to \mathbb{R}^{\geqslant 0}$ such that $m(x)\mu(x) \geqslant m(x0)\mu(x0) + m(x1)\mu(x1)$ for all $x \in \{0,1\}^*$.

More intuitively, supermartingales are similar to semimeasures in that they also assign probabilities or weights to binary strings, but instead of interpreting them as predictors, we might think of them as capturing a that the expected value of a given function (in the case of the definition, μ) cannot increase. Yet, most relevant for our purposes is noting that supermartingales can be constructed from semimeasures, which will be done later on in order to show that supermartingales do not exhibit a certain convergence property, which ultimately shows the non-convergence of our Solomonoff predictor in a certain context.

The first set of non-convergence results center on non-convergence for a specific type of bounded arbitrary computable measure, which we introduce below.

Definition 2.6 (ϵ**-boundedness**). Let $\epsilon > 0$. A measure μ is ϵ-*bounded* if $\epsilon < \mu(b|x) < 1 - \epsilon$ for all $x \in \{0,1\}^*$ and $b \in \{0,1\}$.

Some further definitions will be presented later on, in the context of the main results, based on their usage in only these sections and thus clarity of the arguments.

2.4 Relevant Previous Results

Finally, with all the relevant contextual and notational context, we state the original convergence result from Solomonoff in 1978.

Theorem 2.7 (Solomonoff's Universal Convergence Result). *Let M be a universal lower semicomputable semimeasure, μ be a computable measure, and $x \in \{0,1\}^\infty$. Then it holds that:*

$$M(x_n|x_{<n}) \to \mu(x_n|x_{<n})$$

with μ-probability 1 as $n \to \infty$.

In other words, convergence occurs for sequences that are randomly sampled from μ. However, as hinted at before, this still does not say anything about convergence for a specific sequence. The aims of Hutter and Muchnik (2007) and Lattimore and Hutter (2015), which we seek to expand upon, are precisely to see whether this convergence applies for specific individual sequences – namely, the aforementioned Martin-Löf sequences of a certain form. We present one of the main results of the first mentioned work, showing that, when the underlying true distribution is the uniform Lebesgue measure λ, there is at least one M.L-random sequence relative to λ for which convergence fails.

Theorem 2.8 (Existence of Specific Sequence for Universal Semimeasure Non-Convergence [5]). *There exists a universal semimeasure M and a λ M.L-random sequence α such that:*

$$M(\alpha_n|\alpha_{<n}) \not\to \lambda(\alpha_n|\alpha_{<n}) \text{ for } n \to \infty$$

A later result by Lattimore and Hutter (2015) shows that, not only is there a specific sequence λ-random sequence for which the predictions of a universal semimeasure fail to converge to λ, but that for every universal semimeasure M, there is a λ M.L random sequence on which M fails to converge. They do this for both off-sequence and on-sequence convergence. More specifically, we say that a universal lower semicomputable semimeasure M converges *on-sequence* to an underlying computable measure μ on a μ M.L-random string α if:

$$\lim_{n\to\infty} M(\alpha_n|\alpha_{<n}) = \mu(\alpha_n|\alpha_{<n})$$

[5] The result may be found on Hutter and Muchnik (2007), pg. 8.

Meanwhile, we say that M converges *off-sequence* if:

$$\lim_{n\to\infty} \sum_{b\in\{0,1\}} \left[M(b|\alpha_{<n}) - \mu(b|\alpha_{<n})\right]^2 = 0$$

We present the off-sequence version first.

Theorem 2.9 (Universal Semimeasure Off-Sequence Non-Convergence [6])**.** *Let* M *be a universal mixture. Then there exists a* λ *M.L-random* α *such that it is not the case that:*

$$\lim_{n\to\infty} \sum_{b\in\{0,1\}} \left(M(b|\alpha_{<n}) - \frac{1}{2}\right)^2 = 0.$$

Having presented the relevant works we draw upon, we now present the main results. Namely, we seek to generalize the results of both Hutter and Muchnik (2007) and Lattimore and Hutter (2015) to M.L-random sequences of different underlying measures, namely ϵ-bounded computable measures. Specifically, we extend Theorem 2.9 and Theorem 2.9 to an arbitrary ϵ-bounded computable measures. Section 3 will contain the extensions of Theorem 2.8 and Section 4 will contain the extensions of Theorem 2.9.

[6]The result may be found on Lattimore and Hutter (2015), pg. 6.

3 First Generalization Results

The main focus of this section is on proving the following claim: the extension of Theorem 2.11 to an arbitrary ϵ-bounded computable measure.

Theorem 3.1. *There exists a universal semimeasure* M *such that, for any computable* ϵ*-bounded measure* μ*, there exists a* μ*-M.L random sequence* α *such that:*

$$M(\alpha_n|\alpha_{<n}) \not\to \mu(\alpha_n|\alpha_{<n}), \text{ for } n \to \infty$$

Both constructive and non-constructive proofs will be presented. We will begin by proving a series of auxiliary lemmas, primarily to aid in the non-constructive proof, before moving to the main proofs of Theorem 3.1 themselves.

3.1 Auxiliary Results

We first begin by defining the μ M.L-random sequence α along which convergence will be disrupted. Formally, we define α, inductively in $n = 1, 2, 3, ...$ by:

$$\alpha_n = \begin{cases} 0 & \text{if } M(\alpha_{<n}0) \leqslant \mu(\alpha_{<n}0) \\ 1 & \text{if } M(\alpha_{<n}0) > \mu(\alpha_{<n}0) \end{cases}$$

We verify that α is μ.M.L-random in the first place. Appealing to Definition 2.2, we show that, for all n, $M(\alpha_{<n}) \leqslant \mu(\alpha_{<n})$ (that is, we are treating $c = 1$).

We show this via induction, beginning with the base case where $\alpha = \varepsilon$, the empty string. Then $M(\varepsilon) \leqslant 1$ and $\mu(\varepsilon) = 1$, so $M(\varepsilon) \leqslant 1 = \mu(\varepsilon)$ holds as desired. Now for our inductive hypothesis we suppose that $M(\alpha_{<n}) \leqslant \mu(\alpha_{<n})$. There are two cases to consider, based on the continuations of $\alpha_{<n}$, as follows.

Case 1: $\alpha_{<n+1} = \alpha_{<n}0$

Then, by construction:

$$\begin{aligned} M(\alpha_{<n+1}) &= M(\alpha_{<n}0) \\ &\leqslant \mu(\alpha_{<n}0) \\ &= \mu(\alpha_{<n+1}) \end{aligned}$$

Case 2: $\alpha_{<n+1} = \alpha_{<n}1$

Then, by construction:

$$M(\alpha_{<n+1}) = M(\alpha_{<n}1)$$

So,

$$\begin{aligned}\mu(\alpha_{<n}) &\overset{(a)}{\geqslant} M(\alpha_{<n}) \\ &\overset{(b)}{\geqslant} M(\alpha_{<n}0) + M(\alpha_{<n}1) \\ &= M(\alpha_{<n}0) + M(\alpha_{<n+1}) \\ &\overset{(c)}{>} \mu(\alpha_{<n}0) + M(\alpha_{<n+1})\end{aligned}$$

where (a) follows by the inductive hypothesis, (b) follows by the definition of a semimeasure, and (c) follows by the definition of α. Then,

$$\begin{aligned}M(\alpha_{<n+1}) &\leqslant \mu(\alpha_{<n}) - \mu(\alpha_{<n}0) \\ &= \mu(\alpha_{<n}1) \\ &= \mu(\alpha_{<n+1})\end{aligned}$$

Therefore, we indeed have that, for all n:

$$M(\alpha_{<n}) \leqslant \mu(\alpha_{<n}) \tag{3.1}$$

Next, we present a set of results about the frequency of 0s and 1s in μ M.L-random sequences (and thus apply to α) which are used in later auxiliary lemmas. The following results depend on some new definitions, which we will introduce before diving in. We first consider atoms, which are singleton sets that μ gives positive weight to.

Definition 3.2 (Atoms). Let μ be a probability measure and $\omega \in \{0,1\}^\infty$. ω is called an *atom* of μ if and only if $\mu(\{\omega\}) > 0$.

We employ the definition of atoms in the next two results below.

Lemma 3.3. *Let μ be a computable probability measure. If $\omega \in \{0,1\}^\infty$ is computable and μ M.L-random, then ω is an atom of μ.*

Proof. Suppose for the sake of contradiction that $\mu(\{\omega\}) = 0$. Let $U_1 := [\![\omega_{<n_1}]\!]$, where n_1 is the smallest number such that $\mu([\![\omega_{<n_1}]\!]) \leqslant \frac{1}{2} = 2^{-1}$. For $k > 1$, suppose n_{k-1} and $U_{k-1} := [\![\omega_{<n_{k-1}}]\!]$ have already been defined, and let $U_k := [\![\omega_{<n_k}]\!]$, where n_k is the smallest number

strictly greater than n_{k-1} such that $\mu([\![\omega_{<n_k}]\!]) \leqslant 2^{-k}$. Then, as each U_k is a cylinder set, the sequence $\{U_k\}_{k\in\mathbb{N}}$ is a sequence of cylinder sets, and given that $\mu(U_k) = \mu([\![\omega_{<n_{k-1}}]\!]) \leqslant 2^{-k}$ for all $k \in \mathbb{N}$, by the definition of a M.L test relative to μ, $\{U_k\}_{k\in\mathbb{N}}$ is a M.L test relative to μ. But $\omega \in \bigcap_k U_k$. So ω is not μ M.L-random, which is a contradiction. Therefore, it must be that $\mu(\{\omega\}) > 0$, or equivalently that μ is an atom of μ.

□

Lemma 3.4. *Let μ be a computable ϵ-bounded measure for some $\epsilon > 0$. Let $\omega \in \{0,1\}^\infty$ be a μ M.L-random sequence. Then it is not the case that $\omega = b^\infty$ for $b \in \{0,1\}$, ie. in ω there cannot be a point after which all subsequent digits are zeros or all subsequent digits are ones.*

Proof. Assume for contradiction that ω has the structure specified in the lemma statement, ie. that $\omega = b^\infty$, for $b \in \{0,1\}$. Then ω is computable. Since ω is computable and μ M.L-random, by Lemma 3.4, ω is an atom of μ. Since μ is ϵ-bounded, it must be that for all n, $\mu(\omega_n|\omega_{<n}) < 1-\epsilon$. But $\mu(\{\omega\}) = \prod_{n\in\mathbb{N}} \mu(\omega_n|\omega_{<n}) < \prod_{n\in\mathbb{N}} (1-\epsilon) = 0$, so $\mu(\{\omega\}) = 0$. But this contradicts ω being an atom of μ. Therefore, it cannot be the case that $\omega = b^\infty$ for $b \in \{0,1\}$. □

Naturally, if there is never an index in a μ.M.L-random sequence after which either 0 or 1 stops occurring, alternations of those digits will occur an infinite amount of times as well. This is spelled out in the following corollary, which is used in proving the main convergence result.

Corollary 3.5. *If $\omega \in \{0,1\}^\infty$ is μ.M.L-random, then it will contain an infinite amount of indices $n, n+1$ where $\omega_n\omega_{n+1} = 01$.*

We also show that any ϵ-bounded μ.M.L-random sequence will equal 0 at infinitely many even indices – which is necessary for showing Lemma 3.3. For this argument, we employ the concept of bi-immunity, which, conveniently for our purposes, states that certain types of sequences may not contain any computable subsequences consisting of all 0s or all 1s. To use bi-immunity in our desired way, we first define strong atomlessness, a strengthening of the concept of atomlessness, where μ being atomless meaning it has no atoms.

Definition 3.6 (Strong atomlessness[7]). A probability measure μ is *strongly atomless* if, for all $\omega \in \{0,1\}^\infty$ and all computable infinite strictly increasing sequences $n_0, n_1, n_2, \ldots$ of natural numbers, $\lim_{i\to\infty} \prod_{k=0}^{i} \frac{\mu([\![\omega_{<n_{k+1}}]\!])}{\mu([\![\omega_{<n_k}]\!])} = 0$ (where the ratio $\frac{\mu([\![\omega_{<n_{k+1}}]\!])}{\mu([\![\omega_{<n_k}]\!])}$ is taken to be 0 if the denominator is 0).

Then, we define bi-immunity.

Definition 3.7 (Bi-immunity). A string $\omega \in \{0,1\}^*$ is *bi-immune* if and only if ω contains no computable infinite list of positions consisting of all 0's or all 1's.

[7]This notion is taken from Persiau and Zaffora Blando (2025).

Now, we move to our result showing that 0 must occur infinitely often at even indices, leveraging the two previously introduced definitions.

Lemma 3.8. *Let μ be a computable ϵ-bounded measure for some $\epsilon > 0$. Then if ω is a μ.M.L-random sequence, then there is no n such that, for all $m \geqslant n$, $\omega_{2m} = 1$.*

Proof. We first assume for contradiction that there is an n such that, for all $m \geqslant n$, $\omega_{2n} = 1$.

We first show that μ is strongly atomless. That is, we must show that, for all $\omega' \in \{0,1\}^\infty$ and all computably infinite strictly increasing sequences $n_0, n_1, n_2, \ldots$ of natural numbers, $\lim_{i\to\infty} \Pi_{k=0}^{i} \frac{\mu([\![\omega'_{<n_k+1}]\!])}{\mu([\![\omega'_{<n_k}]\!])} = 0$. We note that because $[\![\omega'_{<n_k+1}]\!] \subseteq [\![\omega'_{<n_k}]\!]$, $[\![\omega'_{<n_k+1}]\!] = [\![\omega'_{<n_k+1}]\!] \cap [\![\omega'_{<n_k}]\!]$. So what we must show is $\lim_{i\to\infty} \Pi_{k=0}^{i} \mu([\![\omega'_{<n_k+1}]\!] | [\![\omega'_{<n_k}]\!]) = 0$. But because μ is ϵ-bounded, we have that $\lim_{i\to\infty} \Pi_{k=0}^{i} \mu([\![\omega_{<n_k+1}]\!] | [\![\omega_{<n_k}]\!]) < \lim_{i\to\infty} \Pi_{k=0}^{i} (1-\epsilon) = 0$. Therefore, μ is strongly atomless. By Persiau and Zaffora Blando (2025) (pg. 23), given that ω is μ M.L-random and μ is strongly atomless, then ω is bi-immune.

Then consider the list of positions formed from adding ω_{2m} to the list, for all $m \geqslant n$. Then this list would contain only 1s. However, this contradicts the bi-immunity of ω. Therefore, there cannot be an even index e at and after which at all subsequent even indices α is 1.

□

Given this result, we obtain the following corollary.

Corollary 3.9. *Let μ be a computable ϵ-bounded measure, for some $\epsilon > 0$. If ω is μ M.L-random, there are infinitely many even indices e at which $\omega = 0$.*

Next, we present our first auxiliary result, defining an enumerable supermartingale which will play a key role in showing non-convergence of M in Theorem 3.1.

Lemma 3.10. *Let μ be a computable ϵ-bounded measure, for some $\epsilon > 0$. Let $M^1(x), M^2(x), \ldots$ be a computable sequence of rationals, uniformly in x, that approximates $M(x)$ from below (we have one such sequence for each x).[8] Similarly, let $\mu^1(x), \mu^2(x), \ldots$ be a computable sequence of rationals, also uniformly in x, that approximates $\mu(x)$ from above (where we also have one such sequence for each x). For each t, we define recursively a sequence α^t similar to α as follows:*

$$\alpha^t_n = \begin{cases} 0 & \text{if } M^t(\alpha^t_{<n}0) \leqslant \mu^t(\alpha_{<n}0) \\ 1 & \text{otherwise} \end{cases}$$

Next, for even $\ell(x)$ define $r(x) = 1$ if there are t, n such that $x = \alpha^t_{<n}$ and $r(x) = 0$ otherwise. For odd $\ell(x)$ define $r(x) = \frac{r(x0)\mu(x0) + r(x1)\mu(x1)}{\mu(x)}$. Then r is an enumerable supermartingale relative to μ with $r(\alpha_{1:n})$ being 1 and $\mu(0|\alpha_{<n})$ for infinitely many n's, where $\alpha = lim_{t\to\infty} \alpha^t$ (with $\alpha^t \nearrow \alpha$ lexicographically increasing).

[8] The existence of such a sequence is guaranteed by the enumerability of M.

Proof. It will first be shown that r is an enumerable supermartingale relative to μ. Note that α^t is computable, and therefore that r is enumerable. For odd $\ell(x)$, using the definition of r we note that:

$$r(x)\mu(x) = \frac{r(x0)\mu(x0) + r(x1)\mu(x1)}{\mu(x)} \cdot \mu(x) = r(x0)\mu(x0) + r(x1)\mu(x1)$$

ie. the semimeasure property is satisfied. We note that all terms here are well-defined since the ϵ-boundedness of μ guarantees that $\mu(x) > 0$ for all $x \in \{0,1\}^*$.

For even $\ell(x)$ and $x = \alpha^t_{<n}$, we note that with $\ell(x0), \ell(x1)$ odd:

$$r(x0) = \frac{r(x00)\mu(x00) + r(x01)\mu(x01)}{\mu(x0)}, \quad r(x1) = \frac{r(x10)\mu(x10) + r(x11)\mu(x11)}{\mu(x1)}$$

Therefore:

$$\begin{aligned} r(x0)\mu(x0) &= r(x00)\mu(x00) + r(x01)\mu(x01) \\ r(x1)\mu(x1) &= r(x10)\mu(x10) + r(x11)\mu(x11) \end{aligned}$$

And adding these together we obtain:

$$\begin{aligned} &r(x0)\mu(x0) + r(x1)\mu(x1) \\ &= r(x00)\mu(x00) + r(x01)\mu(x01) + r(x10)\mu(x10) + r(x11)\mu(x11) \end{aligned}$$

We note that $r(x) = 1$ so then $r(x)\mu(x) = \mu(x)$. Furthermore, note that each of $r(x00), r(x01), r(x10), r(x11)$ all equal at most 1, since $r(y)$ for $\ell(y)$ even evaluates to either 0 or 1, per our definition of r. So if each of $r(x00), r(x01), r(x10), r(x11)$ equal at most 1:

$$\begin{aligned} &r(x0)\mu(x0) + r(x1)\mu(x1) \\ &= r(x00)\mu(x00) + r(x01)\mu(x01) + r(x10)\mu(x10) + r(x11)\mu(x11) \\ &\leqslant \mu(x00) + \mu(x01) + \mu(x10) + \mu(x11) \\ &\overset{(a)}{=} \mu(x0) + \mu(x1) \\ &\overset{(b)}{=} \mu(x) \\ &= r(x)\mu(x) \end{aligned}$$

where (a) and (b) follow from μ being a measure. Therefore, $r(x)\mu(x) \geqslant r(x0)\mu(x0) + r(x1)\mu(x1)$.

Finally, for even $\ell(x)$ and $x \neq \alpha^t_{<n}$ for any t, $xy \neq \alpha^t_{1:\ell(xy)}$ for all t, y, so $r(x)\mu(x) = 0 \cdot \mu(x) = 0$. Then, taking $r(x0)\mu(x0) + r(x1)\mu(x1)$ and applying the definition of r for odd length sequences, we obtain:

$$\begin{aligned} r(x0)\mu(x0) + r(x1)\mu(x1) &= r(x00)\mu(x00) + r(x01)\mu(x01) \\ &\quad + r(x10)\mu(x10) + r(x11)\mu(x11) \\ &= 0 \cdot \mu(x00) + 0 \cdot \mu(x01) \\ &\quad + 0 \cdot \mu(x10) + 0 \cdot \mu(x11) \\ &= 0 \end{aligned}$$

Therefore, in this case $r(x)\mu(x) \geqslant r(x0)\mu(x0) + r(x1)\mu(x1)$ also as desired. As the relevant property is satisfied at index of either parity, r is a supermartingale.

Next, we note that α^t monotonically converges to α. This is due to the following. Since for each x, $M^t(x)$ converges to x from below while $\mu^t(x)$ converges to $\mu(x)$ from above, by construction α^t is monotonically increasing with respect to the lexicographical ordering on $\{0,1\}^\infty$. Moreover, again by construction, for each n, there is some t_n such that, for all $t \geqslant t_n$, $\alpha^t_{1:n} = \alpha_{1:n}$. Hence, α^t monotonically converges to α with respect to the lexicographic order.

This implies that, for all odd n, $r(\alpha_{<n}) = r(\alpha^{t_n}_{<n}) = 1$. By Corollary 3.9, we know that $\alpha_n = 0$ for infinitely many even n, since α is μ-random. For each such n, $\alpha^t_n = 0$ for all t for each such n, so $r(\alpha_{<n}) = r(\alpha^{t_n}_{<n}) = \frac{r(\alpha^{t_n}_{<n}0)\mu(\alpha^{t_n}_{<n}0)+r(\alpha^{t_n}_{<n}1)\mu(\alpha^{t_n}_{<n}1)}{\mu(\alpha^{t_n}_{<n})} = \frac{r(\alpha^{t_n}_{<n}0)\mu(\alpha^{t_n}_{<n}0)}{\mu(\alpha^{t_n}_{<n})} \overset{(a)}{=} \frac{\mu(\alpha^{t_n}_{<n}0)}{\mu(\alpha^{t_n}_{<n})} = \mu(0|\alpha^{t_n}_{<n})$, where $\epsilon < \mu(0|\alpha^{t_n}_{<n}) < 1 - \epsilon$ and (a) follows by the fact that $\alpha^{t_n}_{<n}0 = \alpha_{<n}0$ and because, since n is even, by construction $r(\alpha_{<n}0) = 1$. This shows that $r(\alpha_{<n}) = 1$ infinitely often and that $r(\alpha_{<n}) = \mu(0|\alpha_{<n})$ infinitely often.

□

Now, with r defined, we leverage it to show that it will lead to a "non-convergence" of another supermartingale R', which will be directly used in the proof of Theorem 3.1.

Lemma 3.11. *Let μ be a computable ϵ-bounded measure, for some $\epsilon > 0$. For the M.L.-random sequence α as defined before and the enumerable supermartingale r defined in Lemma 3.10 and for any $\eta, \eta' \in \mathbb{R}$ and any on α bounded supermartingale R, ie. $0 < \zeta < R(\alpha_{1:n}) < c < \infty, \forall n$, one of the following two will hold:*

$$\left|\frac{R(\alpha_{1:n})}{R(\alpha_{<n})} - \eta\right| > \delta \quad \textbf{(1)}, \qquad \left|\frac{R'(\alpha_{1:n})}{R'(\alpha_{<n})} - \eta'\right| > \delta \quad \textbf{(2)}$$

(or both) for a non-vanishing fraction of n, where supermartingale $R' := R + r$ and some $\delta > 0$.

Proof. As can be inspected, this proof essentially proceeds from the same argument given by Hutter and Muchnik (2007) since it relies less on the specific structure of the uniform Lebesgue measure, though it will be restated here properly for the sake of completeness. We first define $R'(x) := R(x) + r(x)$ (with, again, r as defined in Lemma 3.10), with again R being a supermartingale bounded on α, and verify that it is a supermartingale by noting that:

$$\begin{aligned}
R'(x)\mu(x) &= (R(x) + r(x))\mu(x) \\
&= R(x)\mu(x) + r(x)\mu(x) \\
&\overset{(a)}{\geqslant} R(x0)\mu(x0) + R(x1)\mu(x1) + r(x0)\mu(x0) + r(x1)\mu(x1) \\
&= \mu(x0)[R(x0) + r(x0)] + \mu(x1)[R(x1) + r(x1)] \\
&= \mu(x0)R'(x0) + \mu(x1)R'(x1)
\end{aligned}$$

where (a) follows from R and r being supermartingales, yielding the desired result.

Now we show that the inequalities in the lemma statement hold for infinitely many n. We then refine the proof to a non-vanishing fraction of n's. We assume that $\frac{R(\alpha_{1:n})}{R(\alpha_{<n})} \to \eta$ for $n \to \infty$. Otherwise, we are done, as we would be in case **(1)**, so we show that with this assumption, case **(2)** will occur. $\eta > 1$ implies $R(\alpha_{1:n}) \to \infty$, $\eta < 1$ implies $R(\alpha_{1:n}) \to 0$. Per our assumption of R being bounded, η must be 1, hence for sufficiently large n_0 we have $|R(\alpha_{1:n}) - R(\alpha_{<n})| < \zeta$ for all $n \geqslant n_0$.

Let r be the supermartingale from Lemma 3.10. For the infinitely many $n \geqslant n_0$ where $r(\alpha_{<n}) = k \in (\epsilon, 1-\epsilon)$ (with k not fixed), $r(\alpha_{1:n}) = 1$, and noting again that $|R(\alpha_{1:n}) - R(\alpha_{<n})| < \zeta$ and $R(\alpha_{1:n}) < c$, we have that:

$$\begin{aligned}
\frac{R'(\alpha_{1:n})}{R'(\alpha_{<n})} - 1 &= \frac{R(\alpha_{1:n}) - R(\alpha_{<n}) + r(\alpha_{1:n}) - r(\alpha_{<n})}{R(\alpha_{<n}) + r(\alpha_{<n})} \\
&\geqslant \frac{-\zeta + (1-k)}{c+k} > \frac{-\zeta + \epsilon}{c + (1-\epsilon)} \geqslant \delta > 0
\end{aligned}$$

for sufficiently small ζ and δ. Similarly for infinitely many $n \geqslant n_0$ where $r(\alpha_{<n}) = 1$ and $r(\alpha_{1:n}) = h \in (\epsilon, 1-\epsilon)$, and using the same bounds for R as above, we have that:

$$\begin{aligned}
1 - \frac{R'(\alpha_{1:n})}{R'(\alpha_{<n})} &= \frac{R(\alpha_{<n}) - R(\alpha_{1:n}) + r(\alpha_{<n}) - r(\alpha_{1:n})}{R(\alpha_{<n}) + r(\alpha_{<n})} \\
&\geqslant \frac{-\zeta + (1-h)}{c+1} > \frac{-\zeta + \epsilon}{c+1} \geqslant \delta > 0
\end{aligned}$$

Thus, we have shown that the lemma statement holds for infinitely many n. We note that, if $r = 0$ for $x \neq \alpha_{1:\ell(x)}$, then r is a supermartingale but non-enumerable, as α is not computable.

So, we have shown that either R or R' does not converge (or possibly both). If R does not converge, there is no need to check whether R' converges, but if the former does converge, we can show that R' does not. The convergence of R depends on the properties of the specific R. This result will become especially useful in the forthcoming non-constructive proof.

□

3.2 Main Theorem Proofs

With the auxiliary results in hand, we now prove Theorem 3.1 for arbitrary computable ϵ-bounded measures. This can be proved both constructively and non-constructively, and we begin with the latter.

Non-Constructive Proof of Theorem 3.1

Proof. Let M be an arbitrary universal lower semicomputable semimeasure, μ a computable ϵ-bounded measure for some $\epsilon > 0$, and r as defined in Lemma 3.10.

We furthermore define $R := \frac{M}{\mu}$, and note that R is a supermartingale. We note that R is well-defined because μ is ϵ-bounded, thus it will always be positive. This is because, computing $R(x)\mu(x) = \frac{M(x)}{\mu(x)}\mu(x) = M(x)$. Then:

$$\begin{aligned} R(x0)\mu(x0) + R(x1)\mu(x1) &= \frac{M(x0)}{\mu(x0)} \cdot \mu(x0) + \frac{M(x1)}{\mu(x1)} \cdot \mu(x1) \\ &= M(x0) + M(x1) \end{aligned}$$

But as $M(x) \geqslant M(x0) + M(x1)$ by M being a semimeasure, then R is a supermartingale.

Next, we define $q(x) = r(x)\mu(x)$, and show that it is an enumerable semimeasure. We note that because q is the product of a lower semicomputable function (r) and a computable function (μ), q is also lower semicomputable. To show that q is a semimeasure, we note that $q(x) \geqslant q(x0) + q(x1)$. We note that $q(x) = r(x)\mu(x)$ and that $q(x0) + q(x1) = r(x0)\mu(x0) + r(x1)\mu(x1)$. But since r is a supermartingale with respect to μ, $r(x)\mu(x) \geqslant r(x0)\mu(x0) + r(x1)\mu(x1)$ or in other words $q(x) \geqslant q(x0) + q(x1)$ as desired.

Furthermore, we define $M'(x) = \frac{1}{2}(M(x) + q(x))$, and show that it is a universal semimeasure. We note that M is a (universal) semimeasure, so $M(x) \geqslant M(x0) + M(x1)$, and that q is a semimeasure, so $q(x) \geqslant q(x0) + q(x1)$. Then $M'(x) = \frac{1}{2}(M(x) + q(x)) \geqslant \frac{1}{2}(M(x0) + q(x0) + M(x1) + q(x1)) = M'(x0) + M'(x1)$, so the semimeasure inequality holds. We note further that $M'(\epsilon) \leqslant \frac{1}{2}(M(\epsilon) + q(\epsilon)) \leqslant \frac{1}{2} \cdot 2 = 1$, again by M and q being semimeasures, so the empty-string mass condition holds as well. Therefore, M' is indeed a semimeasure. Then, to show that M' is universal, we note that $M'(x) \geqslant M(x) + q(x) \geqslant M(x) \geqslant c_\nu \cdot \nu(x)$, where $c_\nu > 0$.

We next note that $R(\alpha_{1:n}) = \frac{M(\alpha_{1:n})}{\mu(\alpha_{1:n})} \leqslant 1$ from Equation (3.1). Because M is universal, $M(x) \geqslant c_\mu \cdot \mu(x)$ so $\frac{M(x)}{\mu(x)} \geqslant c_\mu$ so $R(x) \geqslant c_\mu > 0$. Finally, it follows that $R' := \frac{M'}{\mu}$. Then, we can apply Lemma 3.11 to R, R', r with c_μ being substituted for ζ, and obtain:

$$\begin{aligned}\frac{R'(\alpha_{1:n})}{R'(\alpha_{<n})} &= \frac{M'(\alpha_{1:n})}{\mu(\alpha_{1:n})} \cdot \frac{\mu(\alpha_{<n})}{M'(\alpha_{<n})} \\ &= \frac{M'(\alpha_{1:n})}{M'(\alpha_{<n})} \Big/ \frac{\mu(\alpha_{1:n})}{\mu(\alpha_{<n})} \\ &= \frac{M'(\alpha_n|\alpha_{<n})}{\mu(\alpha_n|\alpha_{<n})} \not\to 1 \\ &\Rightarrow M'(\alpha_n|\alpha_{<n}) \not\to \mu(\alpha_n|\alpha_{<n})\end{aligned}$$

as desired. □

As with the original version of Theorem 3.1 for the uniform Lebesgue measure Hutter and Muchnik (2007), the non-constructive proof above shows that either M or M' does not converge, but we do not know which one. This is because, as we show in Lemma 3.11, either R or R' does not converge (or possibly both, depending on the specific properties of R), and as we effectively substitute M and M' for R and R', respectively, above, this same property holds for M and M'. We would need to know more about M to determine whether it converges. The constructive proof below gets around this ambiguity.

Constructive Proof of Theorem 3.1

Proof. Let μ be a computable ϵ-bounded measure, for some $\epsilon > 0$, and recall that α^t is as defined in Lemma 3.10. We ν as follows:

$$\nu^t(x) = \begin{cases} \mu(x) & \text{if } \ell(x) = t \text{ and } x < \alpha^t_{1:t} \\ 0 & \text{if } \ell(x) = t \text{ and } x \geqslant \alpha^t_{1:t} \\ 0 & \text{if } \ell(x) > t \\ \nu^t(x0) + \nu^t(x1) & \text{if } \ell(x) < t \end{cases}$$

where $<$ is the lexicographical ordering on sequences. We show by cases that ν^t is a semimeasure. First we consider $\ell(x) < t$, for $x \in \{0,1\}^*$. Then by definition of ν^t, we have that $\nu^t(x) = \nu^t(x0) + \nu^t(x1)$, so $\nu^t(x) \geqslant \nu^t(x0) + \nu^t(x1)$ as desired. Then we consider $\ell(x) = t$, which means that $\nu^t(x0) = \nu^t(x1) = 0$ because $\ell(x0), \ell(x1) > t$, so then $\nu^t(x0) + \nu^t(x1) = 0 \leqslant \nu^t(x)$. Finally, we consider $\ell(x) > t$. Then by the definition of ν^t, we have that $\nu^t(x) = 0$. We also have that then $\ell(x0), \ell(x1) > t$, so by the definition of ν^t again we have that $\nu^t(x0) = \nu^t(x1) = 0$, so $\nu^t(x) = \nu^t(x0) + \nu^t(x1)$. As the semimeasure property is satisfied in any case, ν^t is indeed a semimeasure.

So, as ν^t is a semimeasure and because α^t is computable, so is ν^t. Moreover, ν^t is monotone increasing in t, hence we can define ν as $\lim_{t\to\infty} \nu^t$. Next, we prove by induction that:

$$\nu^t(x) = \mu(x) \text{ if } x < \alpha^t_{1:\ell(x)} \text{ and } \ell(x) \leqslant t, \quad \nu^t(x) = 0 \text{ if } x > \alpha^t_{1:\ell(x)}$$

We begin with the base case, where $\ell(x) = t$. If $x < \alpha^t_{1:\ell(x)}$, then $\nu^t(x) = \mu(x)$ by definition of ν^t. If $x > \alpha^t_{1:\ell(x)}$, then $\nu^t(x) = 0$ by definition as well. Therefore for $\ell(x) = t$, we have that $\nu^t(x) = \mu(x)$ if $x < \alpha^t_{1:\ell(x)}$ and $\quad \nu^t(x) = 0$ if $x > \alpha^t_{1:\ell(x)}$ as desired.

We now take as our inductive hypothesis that the claim holds for all strings $h \in \{0,1\}^*$ of length $\ell(h) = k, 1 \leqslant k \leqslant t$.

We now show that it holds for $x \in \{0,1\}^*$ with $\ell(x) = k - 1 < t$. By definition of ν^t, this means that:

$$\nu^t(x) = \nu^t(x0) + \nu^t(x1)$$

We consider two cases, first where $x < \alpha^t_{1:k-1}$. Then $x0, x1 < \alpha^t_{1:k}$, so by the inductive hypothesis we obtain $\nu^t(x0) = \mu(x0)$ and $\nu^t(x1) = \mu(x1)$, so $\nu^t(x) = \mu(x0) + \mu(x1)$ because μ is a measure. In the second case, where $x > \alpha^t_{1:k-1}$, we have that $x0, x1 > \alpha^t_{1:k}$ as well, so applying the inductive hypothesis again we obtain that $\nu^t(x0) = 0$ and $\nu^t(x1) = 0$. Thus we have shown for $x \in \{0,1\}^*$ with $\ell(x) = k - 1 < t$ that $\nu^t(x) = \mu(x)$ if $x < \alpha^t_{1:\ell(x)}$ and $\nu^t(x) = 0$ if $x > \alpha^t_{1:\ell(x)}$ for $1 \leqslant k \leqslant t$, and thus $\nu^t(x) = \mu(x)$ if $x < \alpha^t_{1:\ell(x)}$ and $\ell(x) \leqslant t, \quad \nu^t(x) = 0$ if $x > \alpha^t_{1:\ell(x)}$.

Since sequence $\alpha := \lim_t \alpha^t$ is μ M.L-random, by Corollary 3.5, it contains 01 infinitely often, ie. $\alpha_n \alpha_{n+1} = 01$ infinitely often. In the following we fix such an n. For $t \geqslant n$ we get:

$$\begin{aligned}
\nu^t(\alpha_{<n}) &\overset{(a)}{=} \nu^t(\alpha_{<n}0) + \nu^t(\alpha_{<n}1) \\
&\overset{(b)}{=} \nu^t(\alpha_{<n}0) \\
&= \nu^t(\alpha_{1:n}) \\
&\Rightarrow \nu(\alpha_{<n}) = \nu(\alpha_{1:n})
\end{aligned}$$

where (a) follows by the definition of ν^t, since $\ell(\alpha_{<n}) = n - 1 < t$

follows from the fact that $\alpha_{<n}1 > \alpha_{1:n} \geqslant \alpha^t_{1:n}$ since $\alpha_n = 0$, making $\nu^t(\alpha_{<n}1) = 0$. This all means that $\nu(\alpha_n|\alpha_{<n}) = 1 > \mu(\alpha_n|\alpha_{<n})$. Now, for $t > n$ large enough such that $\alpha^t_{1:n+1} = \alpha_{1:n+1}$ we get:

$$\begin{aligned}
\nu^t(\alpha_{1:n}) &= \nu^t(\alpha^t_{1:n}) \\
&\overset{(a)}{\geqslant} \nu^t(\alpha^t_{1:n}0) \\
&\overset{(b)}{=} \mu^t(\alpha^t_{1:n}0) \\
&\overset{(c)}{\Rightarrow} \nu(\alpha_{1:n}) \geqslant \mu(\alpha_{1:n}0)
\end{aligned}$$

where (a) follows from ν being a semimeasure, (b) follows from the fact that $\alpha^t_{1:n}0 < \alpha^t_{1:n+1}$ since $\alpha_{n+1} = 1$, and that ν^t, and (c) follows by the fact that $\alpha_{1:n}0 = \alpha^t_{1:n}0$ and $\nu^t \nearrow \nu$.

This all ensures that $\nu(\alpha_{1:n}) \geqslant \mu(\alpha_{1:n}0) \geqslant \mu(0|\alpha_{1:n})M(\alpha_{1:n})$ (by our proof that α is μ. M.L.-random).

Now, let $\gamma > 0$. Define $M'(x) := (1-\gamma)\nu(x) + \gamma M(x)$ for all $x \in \{0,1\}^*$. Then, M' is also a universal semimeasure. We define an upper bound for γ such that a γ picked in the relevant range yields an M' that converges to a number greater than the greatest value in the range of the conditional probability of μ. That is, we aim to show that $M'(\alpha_n|\alpha_{<n}) > k > 1-\epsilon$ for any n.

We first simplify M':

$$\begin{aligned} M'(\alpha_n \mid \alpha_{<n}) &= \frac{(1-\gamma)\nu(\alpha_{1:n}) + \gamma M(\alpha_{<n})}{(1-\gamma)\nu(\alpha_{<n}) + \gamma M(\alpha_{<n})} \\ &\overset{(c)}{\geqslant} \frac{(1-\gamma)\nu(\alpha_{1:n})}{(1-\gamma)\nu(\alpha_{<n}) + \gamma\mu(\alpha_{<n})} \\ &\overset{(d)}{=} \frac{(1-\gamma)}{(1-\gamma) + \gamma\mu(\alpha_{<n})/\nu(\alpha_{1:n})} \\ &\overset{(e)}{\geqslant} \frac{(1-\gamma)}{(1-\gamma) + \frac{\gamma\mu(\alpha_{<n})}{\mu(\alpha_{1:n}0)}} \end{aligned}$$

where (c) follows from $M(\alpha_{<n}) \leqslant \mu(\alpha_{<n})$, $M(\alpha_{1:n}) \geqslant 0$, and $\gamma > 0$ where (d) follows from $\nu(\alpha_{<n}) = \nu(\alpha_{1:n})$, and where (e) follows from $\nu(\alpha_{1:n}) \geqslant \mu(\alpha_{1:n}0)$.

Because μ is always positive, we deduce that $\mu(\alpha_{<n}) \geqslant \mu(\alpha_{1:n}0)$. To complete the derivation of our upper bound for γ, we would seemingly need to consider two cases, one where $\mu(\alpha_{<n}) = \mu(\alpha_{1:n}0)$, and the other where $\mu(\alpha_{<n}) > \mu(\alpha_{1:n}0)$. However, we note that, if $\mu(\alpha_{<n}) = \mu(\alpha_{1:n}0)$ were to hold, this would mean that, $\mu(\alpha_{<n}) = \mu(\alpha_{1:n})$ and that $\mu(\alpha_{1:n}) = \mu(\alpha_{1:n}0)$, and thus that $\frac{\mu(\alpha_{1:n}0)}{\mu(\alpha_{1:n})} = \mu(0|\alpha_{1:n}) = 1$. But because $\epsilon < \mu(\alpha_n|\alpha_{<n}) < 1-\epsilon < 1$, this cannot be, so we need only consider the case where $\mu(\alpha_{<n}) > \mu(\alpha_{1:n}0)$.

If $\mu(\alpha_{<n}) > \mu(\alpha_{1:n}0)$, then $\frac{\mu(\alpha_{<n})}{\mu(\alpha_{1:n}0)} = \frac{\mu(\alpha_{<n})}{\mu(\alpha_{1:n})} \cdot \frac{\mu(\alpha_{1:n})}{\mu(\alpha_{1:n}0)} = \frac{1}{\mu(\alpha_n|\alpha_{<n})} \cdot \frac{1}{\mu(0|\alpha_{1:n})} > \frac{1}{(1-\epsilon)^2} = h > 1$. We then observe that $\frac{(1-\gamma)}{(1-\gamma)+h\gamma} = \frac{(1-\gamma)}{1+\gamma(h-1)} = \frac{(1-\gamma)}{1+\gamma q}$, where $q = h-1 > 0$. Then, for M' to not converge to k, we must have that $\frac{(1-\gamma)}{1+\gamma q} > k$, as then M' would always be slightly greater than k, and thus slightly greater than $1-\epsilon$.

Simplifying:

$$\begin{aligned} \frac{1-\gamma}{1+\gamma q} &> k \\ 1-\gamma &> k(1+\gamma q) \\ 1-\gamma &> k + k\gamma q \\ 1-k &> \gamma + k\gamma q \\ 1-k &> \gamma(1+kq) \\ \gamma &< \frac{1-k}{1+kq} \end{aligned}$$

We know that $\frac{1-k}{1+kq}$ is positive because $k < 1$. Therefore, we have a valid upper bound, and so for M' to not converge to μ, we must have that $0 < \gamma < \frac{1-k}{1+kq}$, dependent on the precise k and h.

Since we have come up with a M' that doesn't converge, we have shown that there exists a universal semimeasure that doesn't converge to μ as desired.

□

4 Second Generalization Results

Having shown that, for a wide variety of measures, we can find a universal lower semicomputable semimeasure that fails to mimic their behavior in the limit for a given M.L-random sequence relative to said measures, we now attempt the more ambitious generalization of showing that, for any universal lower semicomputable semimeasure there exists a M.L-random sequence relative to said measures on which it fails to mimic their behavior in the limit, for more measures beyond just the uniform Lebesgue measure.

This time around, we consider off-sequence end behavior. Recall, we say that a universal lower semicomputable semimeasure M converges *off-sequence* to an underlying computable measure μ on a string α if:

$$\lim_{n\to\infty} \sum_{b\in\{0,1\}} \left[M(b|\alpha_{<n}) - \mu(b|\alpha_{<n})\right]^2 = 0$$

More specifically, we show off-sequence non-convergence for an arbitrary, ϵ-bounded measure.

4.1 Off-Sequence Non-Convergence Results

Theorem 4.1. *Let M be a universal mixture and let μ be an ϵ-bounded measure for some $\epsilon > 0$. Then there exists a μ M.L-random α such that:*

$$\lim_{n\to\infty} \sum_{b\in\{0,1\}} \left[M(b|\alpha_{<n}) - \mu(b|\alpha_{<n})\right]^2 \neq 0$$

Proof. We define the same μ-random string α as before. That is, we define α inductively in $n = 1, 2, 3, ...$ by:

$$\alpha_n = \begin{cases} 0 & \text{if } M(\alpha_{<n}0) \leqslant \mu(\alpha_{<n}0) \\ 1 & \text{if } M(\alpha_{<n}0) > \mu(\alpha_{<n}0) \end{cases}$$

Then, we define $\nu : \{0,1\}^* \to [0,1]$ by:

$$\nu(x) := \begin{cases} M(x) & \text{if for all } n \leqslant \ell(x) : x_n = 0 \text{ or } M(x_{<n}0) > \mu(x_{<n}0) \\ 0 & \text{else} \end{cases}$$

We show that ν is both lower semicomputable and a semimeasure. We first show that it is lower semicomputable. We note that ν can be rewritten as follows:

$$\nu(x) = M(x) \cdot [\text{for all } n \leqslant \ell(x) : x_n = 0 \text{ or } M(x_{<n}0) > \mu(x_{<n}0)] \tag{4.1}$$

Let $P(x)$ be the predicate that says, precisely, for all $n \leqslant \ell(x)$, $n \leqslant \ell(x) : x_n = 0$ or $M(x_{<n}0) > \mu(x_{<n}0)$. We note that the first disjunct inside the universal quantifier is decidable because it can be checked in finite time for a finitely long string. The second disjunct is semi-decidable because M is lower semicomputable. That is, because there exists a series of approximations M^t of M after which some approximation M^{t_n}, for $k \geqslant t_n$, $M^k(x0) > \mu(x0)$ if $M(x0) > \mu(x0)$, so in any instance x that makes the second disjunct true, there is a Turing machine T that halts and accepts on the input, making the disjunct semi-decidable.

Then, we note further that a finite disjunction of semi-decidable predicates is semi-decidable. For each n, we can wait until either $x_n = 0$, or whether $M(x_{<n}0) > \mu(x_{<n}0)$. Given that there exists a Turing machine that halts only for inputs that make the second disjunct true, it will halt only when the overall predicate $P(x)$ is true, for any given input. Therefore, $P(x)$ is semi-decidable.

Then, we define the indicator function of $P(x)$ as follows:

$$1_{P(x)} := \begin{cases} 1 & \text{if } P(x) \text{ is true} \\ 0 & \text{else} \end{cases}$$

We show that $1_{P(x)}$ is lower semicomputable. That is, we must construct a total computable function $\phi(x, t)$ such that both $\phi(x, t) \leqslant \phi(x, t+1)$ and $\lim_{t\to\infty} \phi(x, t) = 1_{P(x)}$.

We define $\phi(x, t)$ by simulating $M(x)$ for t steps:

$$\phi(x, t) = \begin{cases} 1 & \text{if } M(x) \text{ accepts within } t \text{ steps} \\ 0 & \text{else} \end{cases}$$

Then we have that $\phi(x, t) \leqslant \phi(x, t+1)$. Furthermore, if $P(x)$ is true, then $\phi(x, t)$ eventually becomes 1 and stays there. If $P(x)$ is false, then $\phi(x, t) = 0$ for all t. Therefore, $\phi(x, t) \nearrow 1_{P(x)}$ for $t \to \infty$.

We note that the right expression in the product in Equation (4.1) is equal to $1_{P(x)}$. Now we show that ν as written in Equation (4.1) is is lower semicomputable. Let $M^t(x)$, $1_{P^t(x)}$ be a sequence of computable approximations for $M(x)$ and $1_{P(x)}$, respectively, with $1_{P^t(x)}$ defined as follows:

$$1_{P^t(x)} := \begin{cases} 1 & \text{if for all } n \leqslant \ell(x) : x_n = 0 \text{ or } M^t(x_{<n}0) > \mu^t(x_{<n}0) \\ 0 & \text{else} \end{cases}$$

where μ^t is defined as in Lemma 3.10 and, due to similar considerations as in Lemma 3.10, $M^t(x_{<n}0) > \mu^t(x_{<n}0)$ if $M(x_{<n}0) > \mu(x_{<n}0)$ as $t \to \infty$.

Then each ν-approximation $\nu^t(x) = M^t(x) \cdot 1_{P^t(x)}$ is computable. Furthermore, because M^t and $1_{P^t(x)}$ are non-decreasing, their product is also non-decreasing, ie. $M^t(x) \cdot 1_{P^t(x)} \leqslant M^{t+1}(x) \cdot 1_{P^{t+1}(x)}$. Furthermore, we have by continuity of multiplication for limits that $\lim_{t\to\infty}(M^t(x) \cdot 1_{P^t(x)}) = M(x) \cdot 1_{P(x)} = \nu(x)$. Combining these facts, we obtain that ν is lower semicomputable.

Now we show that ν is a semimeasure. That is, we must show that $\nu(x) \geqslant \nu(x0) + \nu(x1)$ and that $\nu(\epsilon) \leqslant 1$. We begin by showing that the first property holds. There are two possible cases, one where $\nu(x) = M(x)$ and one where $\nu(x) = 0$. If $\nu(x) = M(x)$, then either $x_n = 0$ for all n or $M(x_{<n}0) > \mu(x_{<n}0)$ for all n, or both. If $x_n = 0$ for all $n \leqslant \ell(x)$, then $\nu(x0) = M(x0)$, and $\nu(x1) \leqslant M(x1)$ (depending on whether the second condition of the disjunct is satisfied for $x1$) so then, $\nu(x) \geqslant \nu(x0) + \nu(x1)$ as desired. If $\nu(x) = 0$, then also for $x0$ and $x1$, there will be some $0 \leqslant n \leqslant \ell(x) + 1$ such that the disjunct it not satisfied as x_n (as is the case for x), thereby making it so that $\nu(x0) = \nu(x1) = 0$ and thus $\nu(x) = \nu(x0) + \nu(x1)$ as desired.

Next, we check that $\nu(\epsilon) \leqslant 1$. We note again that $\nu(x)$ can be written as $\nu(x) = M(x) \cdot 1_{P(x)}$, with $P(x) = [\text{for all } n \leqslant \ell(x) : x_n = 0 \text{ or } M(x_{<n}0) > \mu(x_{<n}0)]$. Since the empty string ϵ has length 0, $P(\epsilon) = 1$ vacuously, since the universal quantifier ranges over the empty set. Then we have that $\nu(\epsilon) = M(\epsilon)$. But since M is a semimeasure, we have that $M(\epsilon) \leqslant 1$, so $\nu(\epsilon) \leqslant 1$ as desired. Thus, we have shown ν is a lower semicomputable semimeasure.

Having shown this, we can claim that there exists a $j \in \mathbb{N}$ such that $\nu = \nu_j$ in the enumeration of all lower semicomputable semimeasures used by M. Now if $\alpha_n = 1$, then $M(\alpha_{<n}0) > \mu(\alpha_{<n}0)$ by the definition of α. Therefore $\alpha_n = 0$ or $M(\alpha_{<n}0) > \mu(\alpha_{<n}0)$ is true for all n and so by the definition of ν, we have that $\nu(\alpha_{1:n}) = M(\alpha_{1:n})$ for all n. Therefore $w_j(\alpha_{<n}) = w_j\nu(\alpha_{<n})/M(\alpha_{<n}) = w_j$. Furthermore,

$$\alpha_n = 0 \Rightarrow M(\alpha_{<n}0) \leqslant \mu(\alpha_{<n}0) \Rightarrow \nu(\alpha_{<n}1) = 0 \Rightarrow \nu(1|\alpha_{<n}) = 0$$

where we used the definitions of α, ν and the definition of conditional probability respectively. Therefore if $\alpha_n = 0$, then:

$$M(0|\alpha_{<n}) + M(1|\alpha_{<n}) \overset{(a)}{=} \sum_{i\in\mathbb{N}} w_i(\alpha_{<n})(\nu_i(0|\alpha_{<n}) + \nu_i(1|\alpha_{<n}))$$

$$\overset{(b)}{=} \Big[\sum_{i\neq j} w_i(\alpha_{<n})(\nu_i(0|\alpha_{<n})+\nu_i(1|\alpha_{<n}))\Big] + w_j(\alpha_{<n})(\nu_j(0|\alpha_{<n}) + \nu_j(1|\alpha_{<n}))$$

$$\overset{(c)}{\leqslant} \Big[\sum_{i\neq j} w_i(\alpha_{<n})\Big] + w_j(\alpha_{<n})M(0|\alpha_{<n})$$

$$= \Big[\sum_{i\in\mathbb{N}} w_i(\alpha_{<n})\Big] - w_j(\alpha_{<n}) + w_j(\alpha_{<n})M(0|\alpha_{<n})$$

$$= \Big[\sum_{i\in\mathbb{N}} w_i(\alpha_{<n})\Big] - w_j(\alpha_{<n})(1 - M(0|\alpha_{<n}))$$

$$\overset{(d)}{=} 1 - w_j(1 - M(0|\alpha_{<n}))$$

$$\overset{(e)}{\leqslant} 1 - w_j(M(1|\alpha_{<n})) \quad (\star)$$

where (a) follows directly from the definition of a universal mixture, (b) follows by extracting $w_j(\alpha_{<n})$ from the sum, (c) follows from using the facts that $\nu_j(0|\alpha_{<n}) + \nu_j(1|\alpha_{<n}) = M(0|\alpha_{<n})$ since $\nu_j(1|\alpha_{<n}) = \nu(1|\alpha_{<n}) = 0$ and $\nu_i(0|\alpha_{<n}) + \nu_i(1|\alpha_{<n}) \leqslant 1$ for all i, (d) follows because $\sum_i w_k(x) = 1$ and $w_j(\alpha_{<n}) = w_j$. For (e) we note that M is a semimeasure, which implies that $1 - M(0|\alpha_{<n}) \geqslant M(1|\alpha_{<n})$. Because α is μ M.L-random, and μ is a non-trivial measure, μ must contain infinitely many zeros. Let n_i be the location of the ith 0 in α and let $k \in \mathbb{N}$ be such that $\nu_k = \mu$. Therefore there exists a $c > 0$ such that:

$$M(1|\alpha_{<n_i}) \overset{(a)}{=} \sum_{l\in\mathbb{N}} w_l(\alpha_{<n_i})\nu_i(1|\alpha_{<n_i}) \overset{(b)}{\geqslant} w_k(\alpha_{<n_i})\mu(1|\alpha_{<n_i}) \overset{(c)}{>} c$$

where (a) follows given the definition of a universal semimeasure (b) follows by extracting the contribution of μ, (c) follows given that $\mu(1|\alpha_{<n}) > \epsilon > 0$ for all n, the fact that α is μ-random, and, in particular, given the definition of M.L-random sequences. Then by $(\star)$ above:

$$\liminf_{i\to\infty} \left[M(0|\alpha_{<n_i}) + M(1|\alpha_{<n_i})\right] \leqslant 1 - w_j c < 1$$

Therefore $\lim_{n\to\infty} \left[M(0|\alpha_{<n}) + M(1|\alpha_{<n})\right] \neq 1$. As $\mu(0|\alpha_{<n}) + \mu(1|\alpha_{<n}) = 1$, it is not the case that:

$$\lim_{n\to\infty} \sum (M(b|\alpha_{<n}) - \mu(b|\alpha_{<n}))^2 \neq 0$$

as required.

□

5 Reflections

Having shown that Solomonoff's supposedly optimal predictors fail to converge to the underlying true measure for an even wider variety of measures than the ones considered in Hutter and Muchnik (2007) and Lattimore and Hutter (2015), we now return to some of the themes brought up at the beginning of this investigation about the limits of predictors. A broader failure of convergence demonstrates further limits for Solomonoff predictors, decreasing the chances of this particular framework serving as an optimal one, according to Sterkenburg's definition in the beginning.

Yet, in spite of these additional negative results, it might also be sensible to come up with an additional notion of optimality, given our considerations in the beginning that optimality of a framework could plausibly also be judged by the framework's on various practical tasks. Sterkenburg's optimality criteria, based on an agent's relative performance to all possible others, and relatedly, the convergence to truth in the long run seem to represent an almost *theoretical* ideality. It refers to good relative performance as it pertains to prediction tasks for arbitrary sequences in $\{0,1\}^\infty$. Such a notion is useful, because it may inspire efforts to push the boundaries and expand our understanding about the theoretical limits of learners. However, Sterkenburg's criteria do not make any further demands about a predictor's optimality for real-world contexts, which intuitively require more than just accuracy to execute well. A method that could be shown to predict these sequences well but be difficult to translate into these sorts of more practical settings – whether because of computational cost, or

With these kinds of considerations in mind, it could also be useful to formulate a notion of optimality centered more around such practical concerns directly. If we were to arrive at such a notion, it may cast previous results in a new light, in the sense that their potential theoretical non-optimality (if an appropriate notion exists in the first place, in light of the results seen in Sterkenburg (2018)) may not impact this more practical optimality. Conversely, it is not obvious that induction methods that approximate the previously examined theoretically optimal ones would also be the most practically optimal under the definition provided. Next, I propose precisely such a new conception of optimality.

We may term an induction method as *practically optimal* if it is able to perform a wide range of practically relevant, real-world tasks not much worse than any other inductive method. In what follows, we will keep this notion mostly informal, due to the changing nature of what may be considered practically relevant, while still providing an example of how this ideality can be gauged.

A key challenge in making such a notion more objective is pinning down what exactly, if anything, is shared by different tasks for which we need (implementable) methods of inductive learning. To start, it may be useful to try and list some of these. As discussed in the beginning, the problem of induction was picked up by the logical empiricists in their quest to make science a mechanistic process, and thus came to be seen as the problem of determining how a scientist ought to adjust their degrees of belief across all possible hypotheses for a phenomenon given evidence that they have seen thus far. Intuitively, an inductive model such as Bayesian reasoning, while perhaps not optimal in the theoretical sense, could suffice for this task. If a scientist were to sit down, write down their initial credences across possible hypotheses, apply the correct updating rules, and adopt the credences that come out on the other side, they may find that Bayesianism is ideal for this purpose as a learning and reasoning framework. [9]

For machine learning tasks, different inductive methods may be needed. For instance, one such type of problem is binary classification where, based on a training data set of vectors $(x_0, \ldots, x_n) \in \mathbb{R}^n$ of features, a learner must output either 0 or 1 (which is used in, for example, tumor classification). Other problems might include more general extrapolation tasks from structured data. Language models, for instance, have natural language sentences as input and training data, and must output other sentences in response, extrapolating the structure present in training sentences (word order, grammatical dependencies, etc.).

For now, as an initial test case, we will concern ourselves only with evaluating practical optimality with respect to these machine learning tasks – specifically, the optimality of Solomonoff inductors, given our extended negative results and those of Hutter, Lattimore, and Muchnik.

To evaluate the practical optimality of the Solomonoff predictor, in light of these new negative results, we will conjecture about the ways in which the predictors might fail on certain practical tasks, with these limitations. More specifically, we try to establish a modest correspondence between the types of ϵ-bounded computable μ random sequences that M fails to converge on, and the types of learning problems Solomonoff predictors may struggle with.

We begin by first zooming in on the ϵ-bounded μ M.L-random sequence α, defined properly as:

$$\alpha_n = \begin{cases} 0 & \text{if } M(\alpha_{<n}0) \leqslant \mu(\alpha_{<n}0) \\ 1 & \text{if } M(\alpha_{<n}0) > \mu(\alpha_{<n}0) \end{cases}$$

In other words, conceptually, α is a data sequence with a positive instance when M's estimates of the next digit are greater than those of μ, and a negative instance when the estimates are less

[9] To be clear, it is not as if a given learning framework is suited for only one purpose. The development of Bayesianism may have been motivated by the attempt to model the reasoning of a rational scientist (and to thereby employ by a scientist in their work), but Bayesianism has also been used in building AI systems (Bayesian neural networks, as at least one example). This could lend more credence to Bayesianism as a learning framework being more practically optimal, to the extent that Bayesianism-inspired AI systems can be employed in a variety of practical tasks.

than or equal to those of μ. When phrased in this way, it seems as if the sequence is intentionally constructed to fool M, and thus raise questions about whether such samples are actually seen typically enough in training data sets to warrant worry about non-convergence. However, since α is μ M.L-random, it is a sequence that would be seen as "typical" according to μ, so if training data is meant to be representative of an underlying distribution, such sequences could seemingly be a part of the training data.

Nevertheless, there is still way in which this sequence seems designed based off the structure of M and μ to prevent convergence. We might term such sequences *adversarial*, in the sense that it is as if it was constructed by an adversary against M to prevent its eventual convergence. Formal notions for adversaries (which also use this same term) exist in other fields already. In machine learning theory, for instance, for an agent to learn in an adversarial environment means that there is no probabilistic assumption about how the data is generated, and that the environment may adapt to the learner's past behavior in providing data [Shalev-Shwartz and Ben-David (2014), pg. 288]. In other words, in learning settings or tasks where we do not try to fool the learner in an adversarial manner and believe that the training data follows some (potentially unknown) distribution, perhaps it may be appropriate to use (approximations of) a Solomonoff predictor. If we judge such tasks to be commonplace in the space of possible across the space of possible, practically relevant machine learning tasks, we may term Solomonoff predictors practically ideal.

The purpose here, again, is not to provide an exhaustive treatment of practical optimality, but rather to introduce it as an alternative notion through which to consider the merits of various inductive learning methods.

6 Conclusion

The conclusions of our investigation have been at least two-fold. For one, we successfully extended two existing types of arguments demonstrating different types of non-convergence for Solomonoff's predictors. Namely, we found first that there exists a lower semicomputable semimeasure that, for a specific α that is M.L-random relative to an arbitrary ϵ-bounded computable measure μ, the semimeasure does not converge to μ, thus extending the argument of Hutter and Muchnik (2007). Then, we found that, for any lower semicomputable universal semimeasure M, and any computable ϵ-bounded measure μ, there is a μ M.L-random sequence α for which M fails to converge to μ off-sequence, extending the argument of Lattimore and Hutter (2015). These additional failures seem to bolster the Solomonoff predictor's inoptimality, as the notion is traditionally understood. However, if we define optimality differently, with a more practical motivation in mind, the predictor's optimality may be viewed anew. Indeed, we introduce a notion of practical optimality, and conclude that the Solomonoff predictor may meet this definition, depending on the precise number of tasks to which its performance would translate poorly, and we make some guesses at what such translations might look like.

All in all, this leaves plenty of room for further works, a few possible ones which will be spelled out here. For one, the restrictions made upon the measures we use in Sections 3 could be challenged, by seeing whether there is an α M.L-random relative to *any* computable measure for which M fails to converge. Similarly, in Section 4, it would be interesting to see whether the on-sequence non-convergence also holds when the underlying measure is an arbitrary computable one, or whether the on-sequence convergence holds for either an ϵ-bounded computable measure or an arbitrary one. We merely found that these results were hard to prove only trying to extend the ones from Hutter, Muchnik, and Lattimore, but given the successful push in algorithmic information theory of generalizing past results only given for the uniform Lebesgue measure, this seems a promising prospect.

It would also be interesting to further develop the notion of practical ideality to view learning frameworks in a new light. In particular, further mapping out correspondances between the types of data sequences that more theoretical predictors such as Solomonoff's struggle with and the types of real-world problems they correspond to would be a useful endeavor, especially in contexts where the reliability of the underlying learners matters, and for understanding what types of work should be explored further for different practical ends. For instance, if it turns out through further analysis that Solomonoff's predictor performs better or at the same level as any other learner in tasks that are not adversarial, it could be termed quite practically useful in

that domain, and those working on such problems ought to put most effort into implementing Solomonoff's predictors more practically or making the learner more efficient. Such a conclusion would be more difficult to make without some notion of practical ideality, or a framework for evaluating how performance in more theoretical contexts translates to real-world problems.

References

Christian, B. (2020). *The Alignment Problem: Machine Learning and Human Values*. New York: W. W. Norton & Company. ISBN: 978-0393635829 (cit. on p. 5).

Doshi-Velez, F. and B. Kim (2017). "Towards A Rigorous Science of Interpretable Machine Learning." In: *arXiv preprint arXiv:1702.08608*. Version 2. URL: https://arxiv.org/abs/1702.08608 (cit. on p. 5).

Downey, R. and D. R. Hirschfeldt (2018). "Algorithmic Randomness." In: *Proceedings of ACM Conference (Conference'17)*. New York, NY, USA: Association for Computing Machinery, pp. 1–8. DOI: 10.1145/nnnnnnn.nnnnnnn. URL: https://www.math.uchicago.edu/~drh/Papers/Papers/algrand.pdf (cit. on pp. 10, 11).

Downey, R. G. and D. R. Hirschfeldt (2010). *Algorithmic Randomness and Complexity*. Theory and Applications of Computability. New York, NY: Springer. ISBN: 978-0-387-95567-4. DOI: 10.1007/978-0-387-68441-3. URL: https://link.springer.com/book/10.1007/978-0-387-68441-3 (cit. on p. 9).

Hartmann, S. and J. Sprenger (2010). "Bayesian Epistemology." In: *The Routledge Companion to Epistemology*. Ed. by S. Bernecker and D. Pritchard. Routledge, pp. 609–620 (cit. on p. 4).

Hume, D. (1748). *An Enquiry Concerning Human Understanding*. Section IV, *Sceptical Doubts Concerning the Operations of the Understanding*, p. 35. London: A. Millar. URL: https://davidhume.org/texts/ehu (cit. on p. 1).

Hutter, M. and A. Muchnik (2007). "On semimeasures predicting Martin-Löf random sequences." In: *Theoretical Computer Science* 382.3, pp. 247–261. DOI: 10.1016/j.tcs.2007.03.040. URL: https://doi.org/10.1016/j.tcs.2007.03.040 (cit. on pp. 2, 6–9, 13, 14, 20, 23, 31, 34).

Jessup, J. A. (1974). "Peirce's Early Account of Induction." In: *Transactions of the Charles S. Peirce Society* 10.4, pp. 224–234. URL: https://www.jstor.org/stable/40319717 (cit. on p. 3).

Laplace, P.-S. (1995). *Philosophical Essay on Probabilities*. Trans. by A. I. Dale. Vol. 13. Sources in the History of Mathematics and Physical Sciences. Translated from the fifth French edition of 1825 with notes by the translator. New York: Springer (cit. on p. 3).

Lattimore, T. and M. Hutter (2015). "On Martin-Löf (non-)convergence of Solomonoff's universal mixture." In: *Theoretical Computer Science* 588, pp. 2–15. DOI: 10.1016/j.tcs.2014.12.004. URL: https://doi.org/10.1016/j.tcs.2014.12.004 (cit. on pp. 2, 6, 7, 9, 13, 14, 31, 34).

Leike, J. and M. Hutter (2015). "On the Computability of Solomonoff Induction and Knowledge-Seeking." In: *arXiv preprint arXiv:1507.04124*. URL: https://arxiv.org/abs/1507.04124 (cit. on p. 6).

Levin, L. A. (1973). "On the Notion of a Random Sequence." In: *Soviet Mathematics Doklady* 14.5, pp. 1413–1416 (cit. on p. 12).

Li, M. and P. Vitányi (2008). *An Introduction to Kolmogorov Complexity and Its Applications*. 3rd. Texts in Computer Science. New York: Springer. ISBN: 978-0-387-49820-1. DOI: 10.1007/978-0-387-49820-1. URL: https://doi.org/10.1007/978-0-387-49820-1 (cit. on p. 9).

Lin, H. (2022). "Bayesian Epistemology." In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta. URL: https://plato.stanford.edu/entries/epistemology-bayesian/ (cit. on p. 4).

Persiau, F. and F. Zaffora Blando (2025). "Randomness and Invariance." Manuscript, unpublished. (cit. on pp. 17, 18).

Popper, K. R. (1959). *The Logic of Scientific Discovery*. Originally published in German as *Logik der Forschung*, 1934. London: Hutchinson (cit. on p. 1).

Rathmanner, S. and M. Hutter (2011). "A Philosophical Treatise of Universal Induction." In: *arXiv preprint arXiv:1105.5721*. URL: https://arxiv.org/abs/1105.5721 (cit. on p. 6).

Rysiew, P. (2020). *Rationality*. Oxford Bibliographies. DOI: 10.1093/obo/9780195396577-0175. URL: https://www.oxfordbibliographies.com/abstract/document/obo-9780195396577/obo-9780195396577-0175.xml (cit. on p. 2).

Shalev-Shwartz, S. and S. Ben-David (2014). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press. ISBN: 978-1-107-05713-5. DOI: 10.1017/CBO9781107298019. URL: https://www.cambridge.org/core/books/understanding-machine-learning/3059695661405D25673058E43C8BE2A6 (cit. on pp. 5, 33).

Skyrms, B. (1996). "Carnapian Inductive Logic and Bayesian Statistics." In: *Philosophy of Science* 63.3, pp. 325–338. DOI: 10.1086/289922 (cit. on p. 4).

Solomonoff, R. J. (1978). "Complexity-based induction systems: Comparisons and convergence theorems." In: *IEEE Transactions on Information Theory* 24.4, pp. 422–432 (cit. on p. 7).

Sterkenburg, T. F. (2018). "Universal Prediction: A Philosophical Investigation." PhD thesis. University of Groningen. URL: https://philsci-archive.pitt.edu/14486/7/proeffin.pdf (cit. on pp. 4, 6, 7, 31).

Titelbaum, M. G. (2022). *Fundamentals of Bayesian Epistemology 2: Arguments, Challenges, Alternatives*. Oxford University Press. ISBN: 9780192863140. DOI: 10.1093/oso/9780192863140.001.0001 (cit. on p. 4).

Wedgwood, R. (2021). "Practical and Theoretical Rationality." In: *The Handbook of Rationality*. Ed. by M. Knauff and W. Spohn. MIT Press, pp. 137–145. ISBN: 9780262366175. DOI: 10.7551/

mitpress/11252.001.0001. URL: https://doi.org/10.7551/mitpress/11252.001.0001 (cit. on p. 2).

Zabell, S. L. (2004). "Carnap and the Logic of Inductive Inference." In: *Handbook of the History of Logic*. Ed. by D. M. Gabbay, J. Woods, and A. Kanamori. Vol. 10. Elsevier, pp. 265–309. DOI: 10.1016/B978-0-444-52936-7.50008-2. URL: https://www.sciencedirect.com/science/article/pii/B9780444529367500082 (cit. on p. 4).

Zvonkin, A. K. and L. A. Levin (1970). "The Complexity of Finite Objects and the Development of the Concepts of Information and Randomness by Means of the Theory of Algorithms." In: *Russian Mathematical Surveys* 25.6, pp. 83–124. DOI: 10.1070/RM1970v025n06ABEH001269 (cit. on p. 11).