Noisy Deductive Reasoning: How Humans Construct Math, and How Math Constructs Universes

David H. Wolpert
Santa Fe Institute, Santa Fe, New Mexico
Complexity Science Hub, Vienna
Arizona State University, Tempe, Arizona
http://davidwolpert.weebly.com

David Kinney
Santa Fe Institute, Santa Fe, New Mexico
http://davidbkinney.com

April 25, 2020

1 Introduction

Humans are imperfect reasoners. In particular, humans are imperfect *mathematical* reasoners. They are fallible, with a non-zero probability of making a mistake in any step of their reasoning. This means that there is a nonzero probability that any conclusion that they come to is mistaken. This is true no matter how convinced they are of that conclusion. Even brilliant mathematicians behave in this way; Poincaré wrote that he was "absolutely incapable of adding without mistakes" (1910, p. 323).

The banter of Poincaré aside, such unavoidable noise in human mathematical reasoning has some far-reaching consequences. An argument that goes back (at least) to Hume points out that since individual mathematicians are imperfect reasoners, the entire community of working mathematicians must also be one big, imperfect reasoner. This implies that there must be nonzero probability of a mistake in every conclusion that mathematicians have ever reached (Hume 2012, Viteri and DeDeo 2020). This noise in the output of communal mathematical research is *unavoidable*, inherent to any physical system (like a collection of human brains) that engages in mathematical reasoning. Indeed, one might argue that there will also be unavoidable noise in the mathematics constructed by any far-future, post-singularity hive of AI mathematicians, or by any society of demi-God aliens whose civilization is a billion years old. After all, awe-inspiring as those minds might be, they are still physical systems, subject to nonzero noise in the physical processes that underlie their reasoning.

Traditionally, almost all work on the foundations and philosophy of mathematics has presumed that mathematics is the result of noise-*free* deterministic reasoning; Hilbert (1928) famously said that "mathematical existence is merely freedom from contradiction". As just pointed out though, this cannot describe any body of mathematics that will ever be produced in our universe.

In light of this discrepancy, suppose we make a small leap, and identify what might be produced by any community of far-future, galaxy-spanning mathematicians with mathematics itself. What are the implications of the view that *mathematics itself*, in an idealized sense, abstracted from any particular set of physical reasoners, is a stochastic system? In other words, what if Hilbert engaged in a category error when he described mathematical existence? What if we ought to represent mathematics not only as subject to instances of undecidability and uncomputability, as Gödel (1934) showed, but also inescapably *unpredictable* in its conclusions, since it is actually stochastic?

In fact, if you just ask them, many practicing human mathematicians *will tell you* that there is a broad probability distribution over mathematical truths. For example, if you ask them about any

Clay prize question, most practicing mathematicians would say that any of the possible answers has nonzero probability of being correct. In fact, human mathematicians act somewhat like Bayesian learners; as mathematicians learn more by investigating open mathematical questions — as their data set of mathematical conclusions grows larger — they update their probability distributions over those open questions. For example, modern computer scientists assign a greater probability to the claim $NP \neq P$ than did computer scientists of several decades ago. What if mathematicians are right to behave as though there were a broad prior distribution over mathematical truths, which changes as they gather more and more mathematical data?

In this paper we present a model of mathematical reasoning as a fundamentally stochastic enterprise, and therefore of mathematics itself as fundamentally stochastic. Our model has the following advantages:

- It allows us to formalize the process by which actual mathematical researchers select questions to investigate.
- It provides a Bayesian justification for the role that *abductive* reasoning plays in actual mathematical research.
- It provides a Bayesian justification of the idea that a mathematical claim warrants a higher degree of belief if there are multiple lines of reasoning supporting that claim.
- It can be used to investigate the mathematical multiverse hypothesis (i.e., the hypothesis that there are multiple physical realities, each of which is isomorphic to a formal system) thereby integrating the analysis of the inherent uncertainty in the laws of physics with analysis of the inherent uncertainty in the laws of mathematics.
- It shows that if working mathematicians are even remotely Bayes rational, then their prior distribution over *mathematical universes* must assign nonzero probability to the possibility that the laws of mathematics are noisy, not mistake-free.

2 Formal Systems

The concept of a "mathematical structure" can be formalized in several equivalent ways, e.g., in terms of model theory, Turing machines, formal systems, etc. Here we will follow Tegmark (1998) and use formal systems. Specifically, a **(recursive) formal system** can be summarized as any triple of the form

1. A finite collection of symbols, (called an **alphabet**), which can be concatenated into **strings**.
2. A (recursive) set of rules for determining which strings are **well-formed formulas** (WFFs).
3. A (recursive) set of rules for determining which WFFs are **theorems**.

For simplicity, we can assume that there is some large set of symbols that contains the alphabets of all formal systems we consider. Strictly speaking, formal systems are equivalence classes, defined by all possible automorphisms of the symbols in the alphabet (Tegmark, 1998, 2008). For current purposes, we do not need to formalize what we mean by the term "rule" in (2, 3); it covers both what are called "inference rules" and "axioms" in the literature. Note that the set of all possible recursive formal systems is countably infinite, and so any formal system can be represented by an integer. As an example, '1 + 1 = 2' is a concatenation of five arithmetic symbols into a string. In the conventional formal system representing standard arithmetic, '1 + 1 = 2' is a WFF and a

theorem. However, '+4−' is not a WFF in that formal system, despite being a string of symbols from its alphabet.

The community of real-world mathematicians does not spend their days only generating theorems in various formal systems. Rather, as mentioned in the introduction, they pose "open questions" in various formal systems, which they try to "answer". To model this, here we restrict attention to formal systems that contain the Boolean $\sim$ (NOT) symbol, with its usual meaning. If in a given such formal system a particular WFF φ is not a theorem, but $\sim \varphi$ is a theorem, we say that φ is an **antitheorem**. For example, '$1 + 1 = 3$' is an antitheorem in standard arithmetic. Loosely speaking, the "open questions" of current mathematics are pairs of a formal system $\mathcal{S}$ together with a WFF in $\mathcal{S}$, φ, and mathematicians would like to conclude that φ is either a theorem or an antitheorem. Sometimes, φ will be a WFF in $\mathcal{S}$ but neither a theorem nor an antitheorem. We call such strings φ **undecidable**. As an example, Gödel (1934) showed that any formal system strong enough to axiomatize arithmetic must contain undecidable WFFs.

To capture this focus of mathematicians on "open questions", in this essay we re-express formal systems as pairs rather than triples:

1. An alphabet;

2. A set of rules for assigning one of four **syntactic valences** to all possible strings of symbols in that alphabet: 'theorem (t)', 'antitheorem (a)', 'not a WFF (n)', or 'undecidable (u)'.

It will be convenient to refer to any pair $(\mathcal{S}, \varphi)$ where $\mathcal{S}$ is a formal system and φ is a string in the alphabet of $\mathcal{S}$ as a **question**, and write it generically as q. We will also refer to any pair (q, v) where v is a valence as a **claim**.

3 A Noisy Mathematical Computer

The *physical Church-Turing thesis* states that the set of functions computable by Turing machines (TMs) include all those functions "that are computable using mechanical algorithmic procedures admissible by the laws of physics" (Wolpert 2019, p. 17). If we assume that any mathematician's brain is bound by the laws of physics, and so their reasoning is also so bound, it follows that any reasoning by a mathematician may be emulated by a TM. In light of our discussion in the introduction, we amend this to suppose that any reasoning by a mathematicians may be emulated by a *probabilistic* Turing Machine (PTM) (see appendix for discussion of TMs and PTMs).

In this paper we exploit this version of the physical Church-Turing thesis, and model any mathematical reasoner – human or otherwise — as a special type of PTM, which we call a **noisy deterministic reasoning machine** (NDR machine). An NDR machine has several tapes. The **questions tape** always contains a finite sequence of questions. We write such a sequence as Q, and interpret it as the set of all "open questions currently being considered by the community of mathematicians" at any iteration of the NDR machine. The separate **claims tape** of the NDR machine always contains a finite sequence of claims, which we refer to as a **claims list**. We write the claims list as C, and interpret it as the set of all claims "currently accepted by the community of mathematicians" at any iteration of the NDR machine.

We assume that the NDR machine is **non-repeating**, which means that C cannot ever contain two claims that have the same question. Intuitively, this means that while there might be hidden contradictions lurking in the set of all claims currently accepted by mathematicians, there are not *explicit* contradictions. In addition to the questions and claims tapes, any NDR machine that models the community of real human mathematicians in any detail will have many work tapes, but we do not need to consider such tapes here.

The NDR machine starts with the questions and claims tapes blank. Then the NDR machine iterates a sequence of three steps. In the first step, it adds new questions to Q. In the second step the NDR machine "tries" to determine the valences of the questions in Q. In the third step, if the valence v of one or more questions q has been found, then the pair (q, v) is added to the end of C, and q is removed from Q. The NDR machine iterates this sequence of three steps forever, i.e., it never halts.

Write $|C|$ for the number of claims in C, and for each counting number n, let $\mathcal{C}_n$ be the set of all sequences of n claims. For any current C and any $n \leq |C|$, define $C(n)$ to be the sequence of the first n claims in C. As an illustration, for any NDR machine that accurately models the real community of practicing mathematicians, the precise sequence of questions in $C(n)$ must be somewhat random, reflecting randomness in which questions the community of mathematicians happened to consider first. The NDR machine models that randomness in the update distribution of the underlying PTM.

We say that a finite claims list C is **mistake-free** if for every claim $(q, v) \in C$, v is either t, a, n, u, depending on whether the question q is t, a, n or u, respectively. In other words, a claims list is mistake-free if every claim is such that, if $q = (\mathcal{S}, \varphi)$, then v is the syntactic valence assigned to φ by $\mathcal{S}$. Intuitively, the "current body of mathematics", as traditionally conceived, is a mistake-free claims list. However, even if it so happened that the current claims list actually were mistake-free, we do *not* assume that humans can determine that fact; in fact, we presume that humans cannot make that determination in many instances. We say that an NDR machine is mistake-free if for all finite n, the probability is 1 that the claims list $C(n)$ produced by the NDR machine will be mistake-free.

We want to analyze the stochastic properties of the claims list, in the limit that the mathematical reasoner has been running for very many iterations. To do that, we require that for any n, the probability distribution of sequences $C(n) \in \mathcal{C}_n$ converges in probability in the limit of an infinite number of iterations of the NDR machine. We further require that for all $n > 0$, the limiting distribution over $\mathcal{C}_n$ is given by marginalizing the last pair in the limiting distribution over $\mathcal{C}_{n+1}$. This is equivalent to requiring that an NDR machine is a "sequential information source" (Grunwald and Vitányi 2004). We write those limiting distributions as $P^{\infty}(C(n))$, one such distribution for each n.

Any sequence of n claims – any $C(n)$ – specifies an associated (unordered) set of claims, which we write as $U(C(n))$. So for each n, $P^{\infty}(C(n))$ defines an associated distribution over all possible (unordered) sets of n claims, which we write as $P^{\infty}(U(C(n)))$. Under the assumptions of this paper, the $n \to \infty$ limit of this distribution over claims lists specifies an associated distribution over all possible lists of claims, i.e., $\lim_{n\to\infty} P^{\infty}(U(C(n)))$ is well-defined. We refer to this as the **claims distribution** of the underlying NDR machine. Intuitively, the claims distribution is the probability distribution over all possible bodies of mathematics that could end up being produced if current mathematicians kept working forever.

Given a claims distribution of an NDR machine, we refer to the associated conditional distribution $P(v|q)$, defined for all q that have nonzero probability of being on the claims tape of the NDR machine at some iteration, as the **answer distribution** of the NDR machine. We will sometimes abuse terminology and refer to the "answer distribution" even if we are implicitly considering $P(v|q)$ restricted to a proper subset of the questions q that can be produced by the NDR machine. We write an answer distribution as $\mathcal{A}$. A **mistake-free answer distribution** is one produced by a mistake-free NDR machine.

Suppose we have a claims distribution which is a delta function about some formal system $\mathcal{S}$, where any string φ which is a WFF under $\mathcal{S}$ has nonzero probability of being on the claims tape of the NDR machine at some iteration (the reason for this second condition is to ensure that the

answer distribution, $\mathcal{A} = P(v|\mathcal{S}, \varphi)$, is well-defined for any φ which is a WFF under $\mathcal{S}$). We refer to the associated pair $(\mathcal{S}, \mathcal{A})$ of any such claims distribution as a **(NDR) world**. Intuitively, it is the combination of a formal system and the set of answers that some NDR machine would provide to questions formulated in terms of that formal system, without specifying a distribution over such questions.

4 Connections to Actual Mathematical Practice

In this section we show how NDR machines can be used to quantify and investigate some of the specific features of the behavior of human mathematicians (see also Viteri and DeDeo (2020)).

4.1 Generating New Research Questions

Given our supposition that the community of practicing mathematicians can be modeled as an NDR machine, what is the precise stochastic process that that NDR machine uses in each iteration, in the step where it adds new questions to Q? Phrased differently, what are the goals that guide how the community of mathematicians decides which open questions to investigate at any given moment?

This is obviously an extremely complicated issue, ultimately involving elements of sociology and human psychology. Nonetheless, it is possible to make some high-level comments. First, most obviously, one goal of human mathematicians is that there be high probability that they generate questions whose valence is either t, a or u. Human mathematicians don't want to "waste their time" considering questions $(\mathcal{S}, \varphi)$ where it turns out that φ is not a WFF under $\mathcal{S}$. So we would expect there to be low probability that any such question is added to Q. Another goal is that mathematicians prefer to consider questions whose answer would be a "breakthrough", leading to many fruitful "insights". One way to formalize this second goal is that human mathematicians want to add questions q to Q such that, if q could be answered (i.e., if the valence v of q could be determined), then after C was augmented with that question-answer pair, the NDR machine would rapidly produce answers to many of the *other* open questions $q \in Q$.

4.2 A Bayesian Model of Mathematical Abduction

While the community of mathematicians can be modeled as an NDR machine, the members of that community don't know the answer distribution of that NDR machine. By definition, the current community of mathematicians only knows the finite set of claims that they have already placed in C, without having an agreed answer v to any question q that is currently in Q (or more generally, without having the answer to any question that is not currently in one of the claims on C). In other words, the current community of mathematicians is uncertain about the distribution $P(v|q)$ to which their far-future intellectual descendants will converge. Like any other kind of uncertainty, this uncertainty can be formalized as a probability distribution, i.e., a posterior distribution $P(\mathcal{A}|C)$. The uncertainty represented by such a distribution is analogous to the uncertainty faced by a person deliberating over whether, at a given moment in time, they currently have cancer; there is a matter of fact as to whether the person has cancer or not, but due to their uncertainty, there is a non-degenerate probability distribution that represents their degree of belief over those two possibilities.

It can be illuminating to consider whether the community of mathematicians can be seen as Bayesian reasoners, exploiting the distribution $P(\mathcal{A}|C)$ to make inferences. For example, the common use of abductive reasoning by mathematicians can be justified on Bayesian grounds. To see

this, let $q = (\mathcal{S}, \varphi), q' = (\mathcal{S}, \varphi')$ be two distinct open questions, not contained in C, which share the same formal system $\mathcal{S}$. For simplicity, assume that mathematicians are quite confident that under $\mathcal{S}$, both φ and φ' are WFFs and are decidable. Formally, using generalized integrals, this means that both point distributions $P(v|q, C) = \int d\mathcal{A}\, P(\mathcal{A}|C)\mathcal{A}(v|q)$ and $P(v|q', C) = \int d\mathcal{A}\, P(\mathcal{A}|C)\mathcal{A}(v|q')$ are vanishingly small if evaluated for the valences $v = u, n$. Suppose as well that if q' were a theorem under $\mathcal{S}$, that would make it more likely that q was also a theorem, i.e., suppose that

$$P(v = t|q, C \cup \{(q', t)\}) > P(v = t|q, C)$$

(where "(q', t)" is shorthand for the event that the valence of q' turns out to be t if $\mathcal{A}(v|q')$ is sampled). Then by Bayes' theorem, *no matter what the distribution* $P(\mathcal{A}|C)$ *is*, the probability that q' is true goes up if q is true, i.e.,

$$P(v = t|q', C \cup \{(q, t)\}) > P(v = t|q', C)$$

Stripped down, this inference pattern can be explicated in two simple steps. First, suppose that mathematicians believe that some hypothesis H would be more likely to be true if a different hypothesis H' were true. Next, upon finding out that H actually is true, they assign higher probability to H' also being true. This general pattern of reasoning, in which we adopt a greater degree of belief in one hypothesis because it would lend credence to some other hypothesis that we already believe to be true, is known as "abduction" (Peirce, 1960), and plays a prominent role in actual mathematical practice (Viteri and DeDeo, 2020). As we have just shown, it is exactly the kind of reasoning one would expect mathematicians to use if they were Bayesian reasoners making inferences about their own answer distribution $\mathcal{A}$.

4.3 The Epistemic Value of Multiple Proof Paths

Note that real human mathematicians often have higher confidence that some question q is a theorem if many independent paths of reasoning suggest that is the case. To understand why this might be Bayes-rational, for any question q, and any claims list size n, expand the prior probability

$$\begin{aligned} P(v|q) &= \int d\mathcal{A}\, \mathcal{A}(v|q)P(\mathcal{A}) \\ &= \int d\mathcal{A} dC(n) dC(n-1) \ldots\, \mathcal{A}(v|q)P(\mathcal{A}|C(n))P(C(n)|C(n-1))P(C(n-1)|C(n-2)) \ldots \end{aligned}$$

Examining the integrand, we see that if many sequences $C(1), C(2), \ldots, C(n)$ with high joint prior probability all result in a value $\mathcal{A}(v|q)P(\mathcal{A}|C(n))$ that is peaked about $v = t$, then the probability that q is a theorem is high. This simple result can be seen as a justification of *why* real human mathematicians should have higher confidence that q is a theorem if "many independent paths of reasoning" – many sequences $C(1), C(2), \ldots, C(n)$ – all suggest that q is a theorem. This shows how, as we mentioned in the introduction, the NDR machine model of human mathematicians lends formal justification to the idea that, everything else being equal, a mathematical claim should be believed more if there are multiple distinct lines of reasoning supporting that claim.

5 Measures over Multiverses

The mathematical universe hypothesis (MUH) argues that our physical universe is just one particular formal system, namely, the one that expresses the laws of physics of our universe (Schmidhuber,

1997; Tegmark, 1998; Hut et al., 2006; Tegmark, 2008, 2009, 2014). Similar ideas are advocated by Barrow (1991, 2011), who uses the phrase "pi in the sky" to describe this view. Somewhat more precisely, the MUH is the hypothesis that any physical world (i.e., any world bound by the laws of physics) is isomorphic to a formal system. A key advantage of the MUH is that it allows for a straightforward explanation of why it is the case that, to use Wigner's (1960) phrase, mathematics is "unreasonably effective" in describing the natural world. If the natural world is, by definition, isometric to mathematical structures, then the isometry between nature and mathematics is no mystery; rather, it is a tautology. While the MUH is accepted (implicitly or otherwise) by many theoretical physicists working on cosmology, some disagree with various aspects of it; for an overview of the controversy, see Hut et al. (2006).

Rephrased in terms of the NDR machine framework, previous versions of the MUH hold that our physical universe is a mistake-free NDR world. That is, the physical universe is isomorphic to a particular formal system $\mathcal{S}$ which in turn assigns, with certainty, a specific syntactic valence to each possible string in the alphabet of $\mathcal{S}$. Our approach allows for an additional possibility; namely, we allow for the possibility that the physical world is isomorphic to an NDR world that is not mistake-free. In such a world, some strings have their syntactic valence not because of the perfect application of the rules of some formal system, but rather because of the stochastic application of those very rules. Thus, our augmented version of the MUH allows for the possibility that mathematical and physical reality are both fundamentally stochastic.

An idea closely related to the MUH as just defined is the mathematical multiverse hypothesis (MMH). The MMH says that some non-singleton subset of formal systems is such that there is a physical universe that is isomorphic to each element of that subset. Each of these possible physical universes is taken to be perfectly *real*, in the sense that the formal system to which that universe is isomorphic is not just the fictitious invention of a mathematician, but rather a description of a physical universe. In this view, the world that we happen to live in is unique not because it is uniquely real, but because it is our *actual* world. Following Lewis (1973), defenders of the MMH understand claims about 'the actual universe' as **indexical** expressions, i.e. expressions whose meaning can shift depending on contingent properties of their speaker (pp. 85-86).

A related concern of people working on the MMH (e.g. Schmidhuber 1997 and Tegmark 2014) is how to specify a probability measure over the set of all universes, which we will refer to as an **MMH measure**. The goal, loosely specified, is to treat such a measure as a prior distribution over universes, take the associated data to be what we happen to know about our specific physical universe, and then use Bayes' theorem to specify the posterior distribution of which physical universe we inhabit, given what we know about our universe. In existing approaches to MMH, it is assumed that the nature of physical reality is completely described by a set of recursive rules that assign, with certainty, a particular syntactic valence to any string. This amounts to the assumption that all physical universes are mistake-free NDR worlds. So the conventional conception of an MMH measure is a distribution over mistake-free physical universes, i.e., a distribution over NDR machines restricted to only allow those that produce mistake-free physical universes. A natural extension, of course, is to have the MMH measure be a distribution over *all* NDR worlds, not just those that are mistake-free. Thus, the probability measure over mathematical universes is a probability distribution over all possible NDR worlds.

6 Do Practicing Mathematicians Believe that Math is Mistake-Free?

In this section we show that the behavior of working mathematicians can only be Bayes rational if they have a prior distribution $P(\mathcal{A})$ that assigns positive probability to answer distributions that are not mistake-free. In other words, either mathematicians are irrational, or they actually believe it is possible that the mathematical universe is inherently noisy.

To see this, first fix some set of formal systems, $\sigma = \{\mathcal{S}\}$, and a set of questions $\mathscr{Q}(\sigma)$ formulated in terms of the alphabets of those formal systems. Every mistake-free NDR machine whose claims distribution ascribes nonzero probability to every subset of σ must have the same (delta function) answer distribution $P(v|q)$ for every string $q \in \mathscr{Q}(\sigma)$. Therefore, any prior distribution over NDR machines that only allows mistake-free ones must induce a delta function over the space of all possible answer distributions $\mathcal{A}$ (implicitly restricted to the questions in any subset of $\mathscr{Q}(\sigma)$). Moreover, that unique allowed answer distribution must be a single-valued map from questions to answers, which we can write as $V(q)$. Note that such a "mistakes-free" prior does not only mean that $P(\mathcal{A})$ is restricted to single-valued functions, allowing any of several such functions; it means that $P(\mathcal{A})$ is restricted to a *unique* single-valued function. Intuitively, $V(q)$ is just the "omniscient" function that assigns to any question $q = (\mathcal{S}, \varphi)$ the actual valence of φ under the formal system $\mathcal{S}$.

Under that mistakes-free prior over NDR machines, $P(\mathcal{A}|C)$ would be undefined unless the claims list C of the current community of mathematicians were mistake-free. In other words, adopting that prior would mean assuming there is zero probability of a mistake in C. On the other hand, we have just shown that under the mistakes-free prior, even if C actually were mistake-free, so that $P(\mathcal{A}|C)$ were well-defined, that posterior probability $P(\mathcal{A}|C)$ would have to equal the prior $P(\mathcal{A})$. So under the mistakes-free prior, $P(\mathcal{A}|C)$ would be a delta function about $V(q)$. In other words, loosely speaking, if $P(\mathcal{A})$ is restricted to mistake-free answer distributions, then what is currently known by mathematicians forces a single-valued answer to all open questions mathematicians are currently considering, along with any open questions they might consider in the future.

Recall though that as described in the introduction, a typical mathematician would respond to the question, "what is the probability that $P = NP$?" with an answer far from the extremal values of 0 and 1. In other words, if you ask them, they will say that $P(\mathcal{A}|C)$ is very far from being a delta function about some single-valued function $V(q)$. In this sense, the behavior of human mathematicians is only consistent with a prior $P(\mathcal{A})$ that assigns positive probability to NDR machines in which strings are assigned their syntactic valences stochastically.

Mathematicians are, of course, free to admit to forming beliefs in way that is (very) far from Bayes rational. But if they wish to maintain Bayes rationality, then they must admit that they ascribe strictly positive prior probability to the possibility that the ultimate laws of mathematics that humans will come up with (i.e., the answer distribution of the NDR machine of the community of human mathematicians) are not even close to mistake-free. For those familiar with epistemic logic, assigning zero prior to a mistake free answer distribution provides a novel way to resolve *the problem of logical omniscience*, which arises if we assume that the prior over answer distributions only assigns positive probability to mistake-free answer distributions.

7 Conclusion

Starting from the discovery of non-Euclidean geometry, mathematics has been greatly enriched whenever it has weakened its assumptions and expanded the range of formal possibilities that it

considers. Following in that spirit of weakening assumptions, here we have aimed to demonstrate the potential fruitfulness of weakening the assumption that mathematics itself is fully deterministic. We believe that this reveals a rich landscape of novel results and subtleties, many still waiting to be uncovered.

References

S. Arora and B. Barak. *Computational complexity: a modern approach*. Cambridge University Press, 2009.

J. D. Barrow. *Theories of everything: The quest for ultimate explanation*. Clarendon Press Oxford, 1991.

J. D. Barrow. Godel and physics. *Kurt Gödel and the Foundations of Mathematics: Horizons of Truth*, page 255, 2011.

K. Gödel. *On undecidable propositions of formal mathematics systems*. Institute for Advanced Study, 1934.

P. Grunwald and P. Vitányi. Shannon information and kolmogorov complexity. *arXiv preprint cs/0410002*, 2004.

D. Hilbert. Die grundlagen der mathematik. In *Die Grundlagen der Mathematik*, pages 1–21. Springer, 1928.

D. Hume. *A treatise of human nature*. Courier Corporation, 2012. Book 1, Part 4, Section 1.

P. Hut, M. Alford, and M. Tegmark. On math, matter and mind. *Foundations of Physics*, 36(6): 765–794, 2006.

D. Lewis. *Counterfactuals*. Oxford: Basil Blackwell, 1973.

C. S. Peirce. *Collected papers of charles sanders peirce*, volume 2. Harvard University Press, 1960.

H. Poincaré. Mathematical creation. *The Monist*, pages 321–335, 1910.

J. Schmidhuber. A computer scientist's view of life, the universe, and everything. In *Foundations of computer science*, pages 201–208. Springer, 1997.

M. Tegmark. Is "the theory of everything" merely the ultimate ensemble theory? *Annals of Physics*, 270(1):1–51, 1998.

M. Tegmark. The mathematical universe. *Foundations of physics*, 38(2):101–150, 2008.

M. Tegmark. The multiverse hierarchy. *arXiv preprint arXiv:0905.1283*, 2009.

M. Tegmark. *Our mathematical universe: My quest for the ultimate nature of reality*. Vintage, 2014.

S. Viteri and S. DeDeo. Explosive proofs of mathematical truths. *arXiv preprint arXiv:2004.00055*, 2020.

E. P. Wigner. The unreasonable effectiveness of mathematics in the natural sciences. 13:1–14, 1960.

D. H. Wolpert. The stochastic thermodynamics of computation. *Journal of Physics A: Mathematical and Theoretical*, 52(19):193001, 2019.

A Probabilistic Turing Machines

Perhaps the most famous class of computational machines are Turing machines. One reason for their fame is that it seems one can model any computational machine that is constructable by humans as a Turing machine. A bit more formally, the **Church-Turing thesis** states that "a function on the natural numbers is computable by a human being following an algorithm, ignoring resource limitations, if and only if it is computable by a Turing machine."

There are many different definitions of Turing machines (TMs) that are "computationally equivalent" to one another. For us, it will suffice to define a TM as a 7-tuple $(R, \Lambda, b, v, r^{\varnothing}, r^{A}, \rho)$ where:

1. R is a finite set of **computational states**;
2. Λ is a finite **alphabet** containing at least three symbols;
3. $b \in \Lambda$ is a special **blank** symbol;
4. $v \in \mathbb{Z}$ is a **pointer**;
5. $r^{\varnothing} \in R$ is the **start state**;
6. $r^{A} \in R$ is the **halt state**; and
7. $\rho : R \times \mathbb{Z} \times \Lambda^{\infty} \to R \times \mathbb{Z} \times \Lambda^{\infty}$ is the **update function**. It is required that for all triples (r, v, T), that if we write $(r', v', T') = \rho(r, v, T)$, then v' does not differ by more than 1 from v, and the vector T' is identical to the vectors T for all components with the possible exception of the component with index v;[1]

We sometimes refer to R as the states of the "head" of the TM, and refer to the third argument of ρ as a **tape**, writing a value of the tape (i.e., of the semi-infinite string of elements of the alphabet) as T.

Any TM $(R, \Sigma, b, v, r^{\varnothing}, r^{A}, \rho)$ starts with $r = r^{\varnothing}$, the counter set to a specific initial value (e.g, 0), and with T consisting of a finite contiguous set of non-blank symbols, with all other symbols equal to b. The TM operates by iteratively applying ρ, until the computational state falls in r^{A}, at which time it stops, i.e., any ID with the head in the halt state is a fixed point of ρ.

If running a TM on a given initial state of the tape results in the TM eventually halting, the largest blank-delimited string that contains the position of the pointer when the TM halts is called the TM's **output**. The initial state of T (excluding the blanks) is sometimes called the associated **input**, or **program**. (However, the reader should be warned that the term "program" has been used by some physicists to mean specifically the shortest input to a TM that results in it computing a given output.) We also say that the TM **computes** an output from an input. In general, there will be inputs for which the TM never halts. The set of all those inputs to a TM that cause it to eventually halt is called its **halting set**.

The set of triples that are possible arguments to the update function of a given TM are sometimes called the set of **instantaneous descriptions** (IDs) of the TM. Note that as an alternative to the definition in (7) above, we could define the update function of any TM as a map over an associated space of IDs.

In one particularly popular variant of this definition of TMs the single tape is replaced by multiple tapes. Typically one of those tapes contains the input, one contains the TM's output (if

[1] Technically the update function only needs to be defined on the "finitary" subset of $\mathbb{R} \times \mathbb{Z} \times \Lambda^{\infty}$, namely, those elements of $\mathbb{R} \times \mathbb{Z} \times \Lambda^{\infty}$ for which the tape contents has a non-blank value in only finitely many positions.

and) when the TM halts, and there are one or more intermediate "work tapes" that are in essence used as scratch pads. The advantage of using this more complicated variant of TMs is that it is often easier to prove theorems for such machines than for single-tape TMs. However, there is no difference in their computational power. More precisely, one can transform any single-tape TM into an equivalent multi-tape TM (i.e., one that computes the same partial function), as shown by Arora and Barak (2009).

A **universal Turing machine** (UTM), M, is one that can be used to emulate any other TM. More precisely, in terms of the single-tape variant of TMs, a UTM M has the property that for any other TM M', there is an invertible map f from the set of possible states of the tape of M' into the set of possible states of the tape of M, such that if we:

1. apply f to an input string σ' of M' to fix an input string σ of M;
2. run M on σ until it halts;
3. apply f^{-1} to the resultant output of M;

then we get exactly the output computed by M' if it is run directly on σ'.

An important theorem of computer science is that there exist universal TMs (UTMs). Intuitively, this just means that there exists programming languages which are "universal", in that we can use them to implement any desired program in any other language, after appropriate translation of that program from that other language. The physical CT thesis considers UTMs, and we implicitly restrict attention to them as well.

Suppose we have two strings s^1 and s^2 where s^1 is a proper prefix of s^2. If we run the TM on s^1, it can detect when it gets to the end of its input, by noting that the following symbol on the tape is a blank. Therefore, it can behave differently after having reached the end of s^1 from how it behaves when it reaches the end of the first $\ell(s^1)$ bits in s^2. As a result, it may be that both of those input strings are in its halting set, but result in different outputs. A **prefix (free) TM** is one in which this can never happen: there is no string in its halting set that is a proper prefix of another string in its halting set. For technical reasons, it is conventional in the physics literature to focus on prefix TMs, and we do so here.

The **coin-flipping distribution** of a prefix TM M is the probability distribution over the strings in M's halting set generated by IID "tossing a coin" to generate those strings, in a Bernoulli process, and then normalizing. So any string σ in the halting set has probability $2^{-|\sigma|}/\Omega$ under the coin-flipping prior, where Ω is the normalization constant for the TM in question.

Finally, for our purposes, a **Probabilistic Turing Machine** (PTM) is a conventional TM as defined by conditions (1)-(7), except that the update function ρ is generalized to be a conditional distribution. In particular, we typically require that there is zero probability that applying such an update conditional distribution violates condition (7). Depending on how we use a PTM to model NDR, we may introduce other requirements as well.