Selecting events is not rewriting the history of events (in a continuous probability space)

Leonardo Pedro

June 11, 2023

Abstract

Using the simple fact described in the title, we prove the existence of a computational problem with implications to Machine Learning, Quantum Mechanics and Complexity Theory. We also prove P!=NP (the solution can be verified in time polynomial in the number of bits of the input and output (NP) but the problem cannot be solved in time polynomial in the number of bits of the input and output (P)), but this claim still needs to be reviewed by experts in Complexity Theory.

1. Introduction

In this article we will be using the words "real" as in $\mathbb{R}$, "real-world" and "random" to avoid misunderstandings in contexts where we could (and perhaps should) just use the word "real" instead of "real-world" and the word "non-deterministic" (in the Physics sense, not in the Complexity Theory sense) instead of "random".

We can always select events with some feature without rewriting the history of events, in a standard probability space. In fact, in a continuous probability space we can select events such that a random real variable $y \in [0, 1]$ (with probability given by the Lebesgue measure) verifies $y = 0$, but there is no complete history of events where $y = 0$ (always as would be required, or even just once) because the probability space is continuous by assumption and thus the event $y = 0$ has null probability (only an interval would have non-null probability).

In this article we will show that this simple but non-trivial fact has profound implications not only to Complexity Theory[1][2][3][4] but also to Machine Learning[2] and Quantum Mechanics (independently of the implications to the P vs. NP problem). Almost all (in the sense we will define in this article) real functions of a real variable cannot be computed for all practical purposes, not even approximately. But a random selection allows computations in polynomial-time complexity

involving the incomplete knowledge about a real function that cannot be computed in polynomial-time complexity. This is a fundamental reason why we cannot exclude a random time-evolution: a deterministic time-evolution may exist, but it has so much complexity that it cannot be calculated for all practical purposes, not even approximately (since L^∞ is non-separable).

Note that whenever we deal with a non-separable space, there are issues with computability[5] because some elements of a non-separable space cannot be approximated by a finite set of elements, up to an arbitrarily small error. For instance, $L^\infty([0,1])$ and its dual space are both non-separable[6]. While there are separable spaces of real functions of real variables, whenever we add uncertainties/probabilities to such spaces (which is often required when doing approximations) we tend to create non-separable spaces[6], unless equivalence relations change the space of functions. For instance, the set $[0,1]$ is separable but the Lebesgue measure imposes that the rational numbers in $[0,1]$ can be discarded, despite that the sets $[0,1]$ including /excluding rational numbers are different. In the same way, the smooth functions (or functions computable in a reasonable time) may be discarded from a space of functions for particular uncertainties/probabilities.

Using the simple fact mentioned above, we will also prove the existence of a computational problem (defined by a continuous probability space) whose solution can be verified in time polynomial in the numbers of bits of the input and output (NP) but cannot be solved in time polynomial in the numbers of bits of the input and output (P), when using only a deterministic Turing machine[7][8]. That is, P$\neq$NP.

The goal of this article is to define the specific problem unambiguously, and the level of mathematical details will be adjusted to that: too much mathematical detail would shift the focus from the specific problem. Less detail does not always imply less mathematical rigor. Since the present author is an expert in Physics but not an expert in Complexity Theory, we will also try to prove the P vs. NP as much as it is possible, but only as a secondary goal knowing that much work by experts in complexity theory is still required because it is likely that: 1) something went wrong in the relation described here between the specific problem and the P vs. NP problem; and/or 2) the specific problem indeed can be used to solve the P vs. NP problem, but the proof presented here is incomplete.

2. Complexity Theory in the context of probability theory

Complexity Theory can be studied in the context of probability theory[1][2][3][4], because many real-world problems require approximations and uncertainties not only due to the limitations of any computer (already accounted for by Complexity Theory when using only finite numbers of bits, although there is also room for improvement here) but also due to the limitations of the measuring

devices of physical phenomena and limitations of the mathematical models used to approximate the real-world, for instance when dealing with real variables. Most uncertainties are not related to the computer used and do not get smaller when increasing the number of bits in the computer. In fact, Physics as a science tries to be independent of Computer Science and vice-versa, as much as possible. Probability theory is a language (or interface) that allows us to transfer a problem between two sciences (these two or others).

There are two possible approaches to errors or uncertainties[1][2][3][4]: average error (with respect to a probability measure) and maximal (except in sets of null measure with respect to a probability measure) error. It turns out that both approaches can be defined using Hilbert spaces: the average error (that is, L^2 norm) is defined by a normalized wave-function (an element of the Hilbert space) being the square-root of a probability density function; and the maximal error (that is, L^∞ norm) is defined by an element of the abelian von Neumann algebra of operators on the Hilbert space.

The average error is relevant because even if P!=NP it could still make no difference with respect to the scenario P=NP for many practical purposes, if every NP problem with a reasonable probability distribution on its inputs could be solved in polynomial time-complexity on average on a deterministic Turing machine[9][10].

Moreover, since some functions are real constant functions, the definition of a real function must be consistent with the definition of a real number. Certainly, a natural definition of a real number in the context of Complexity Theory uses a standard probability space. Non-standard probability spaces are rarely (or never) used in Experimental Physics, so it is not obvious how useful a "real number" (or "real function") defined in a non-standard probability space could be in real-world applications. There are only countable or continuous measures (or mixed) in a standard probability space, then we can define exactly only a countable number of real numbers (usually the rationals, but not necessarily), the remaining real numbers can only be constrained to be inside an interval with a finite width, eventually very small but never zero. We should also define all real functions using a standard probability space, unless we find a fundamental reason not to do it (we will not find it in this article).

We require a continuum standard probability space, since we can always define a regular conditional probability density which implements a selection of events in such probability space[11]. For instance, this is what we do when we neglect the intrinsic computation error (due to cosmic rays and many other reasons), we define a deterministic function by selecting only certain events from a complete history of random events. It is well known since many decades that in an infinite-dimensional sphere of radius 1 (subset of a real Hilbert space) there is a uniform prior measure induced by the L^2-distance in the Hilbert space[12]. Every point in the sphere has null measure, only regions of the sphere with non-null distance between some of its points may have non-null measure (compatible

with the uniform prior measure). This implies that any knowledge (compatible with the uniform prior measure on the sphere) about a real normalized wave-function has necessarily uncertainties, defined by a connected region in the sphere with non-null maximum L^2 distance to some wave-function, however small it might be.

The discrete nature of the Turing machine is certainly compatible with a continuous probability space: the number of bits of the input or output can be arbitrarily large, and it is proportional to the logarithm of the resolution of the partition of the interval $[0, 1]$, with each disjoint set of the partition corresponding to a different binary number. Excluding a continuous cumulative prior would be unjustified, for many reasons including: no prior is better for all cases[13], there are many problems where a step cumulative prior would not fit well (for instance there is no uniform measure for the rationals in $[0, 1]$ only for the reals); it is hard to formulate any real-world problem where only a step cumulative prior is used (think about the numbers π or $\sqrt{2}$ in numerical approximations, for instance), we usually use a mixture of step and continuous cumulative priors; we can map an ensemble of discrete random variables one-to-one to the real numbers, for instance an ensemble of fair coins corresponds to the uniform real measure in the interval $[0, 1]$; also any real-world computer has an intrinsic computation error (due to cosmic rays and many other reasons) which is usually very small, but it cannot be eliminated. Thus, while we can formulate a new unsolved version of the P vs. NP problem where only a step cumulative prior is accepted, such version of the problem has little to do with real-world computers and real-world problems.

Note that there are 2^{2^n} different boolean[1] functions on n boolean variables, Shannon proved that almost all Boolean functions on n variables require circuits of size $\mathcal{O}(2^n/n)$ [14][2], thus the time complexity of almost all Boolean functions on n variables for a Turing machine is at least (and at most, for all Boolean functions) $\mathcal{O}(2^n/(n\log(n)))$[15] which is not polynomial in n. Thus, almost all numerical functions are not in the complexity class P, according to the uniform prior measure for any resolution of the partition. Moreover, the uniform prior measure is compatible with a prior measure which excludes (both in the maximal and in the average error approaches) all real functions which are approximated by numerical functions with complexity class P, for all resolutions bigger than some resolution of the partition.

3. Definition of the problem

Consider three measure spaces $X = Y = [0, 1] \in \mathbb{R}$ and $X \otimes X$ with the Lebesgue measure, corresponding to inputs (X or $X \otimes X$) and an output (Y) of real functions. Given an input in X, we

[1] That is, a function with $m = 1$m=1 boolean outputs

[2] See also: https://math.stackexchange.com/questions/756813/do-there-exist-polynomials-not-computable-in-polynomial-time

define a regular conditional probability density which is a function from $X \otimes X \to Y$, given by the probability that a function for an input in X some constant output $y \in Y$. But there is always also a marginal probability density for Y, and we cannot say without uncertainty which is the output, because the corresponding prior probability density would be incompatible with the prior Lebesgue measure (by the Radon–Nikodym theorem). Thus, the regular conditional probability density is a deterministic selection of events which cannot be a complete history of events.

In a standard measure space it is always possible to define regular conditional probabilities[11] and to choose the probability density $p(x) = p_0(x) > 0$ for all $x \in X$, except in sets with null measure. Thus, we will define $p(x \otimes y) = p(y|x)p_0(x)$ the joint probability density for the tensor product $X \otimes Y$ for a particular $p(x) = p_0(x) > 0$ for all $x \in X$, except in sets with null measure. Then, we can obtain any other joint probability density $p(x \otimes y)$ from the expression $p(y|x)p(x) = \frac{p(x \otimes y)}{p_0(x)} p(x)$.

The following results are valid for a random input in the interval $[0, 1]$ (which is a standard probability space) and also for an input (or output) without uncertainties up to sets with null measure with respect to the prior marginal measure of the input (or output), because we use regular conditional probabilities (which always exist in standard probability spaces[11]) for fully known inputs (or outputs). This is crucial, since the input includes two samples from a uniform distribution in $[0, 1]$ which may generate numerical functions in P when the sample is in a set of null measure.

However, a (continuous cumulative) probability distribution does not contain enough information to unambiguously define a function. On the other hand, a real wave-function whose square is the joint probability distribution allows the definition of a unitary operator on a separable Hilbert space. A unitary operator is a random generalization of a deterministic symmetry transformation of a (countable or continuous) sample space. Any unitary operator defined by a wave-function of two continuous variables cannot be a deterministic symmetry transformation (for similar reasons that a continuous probability distribution cannot unambiguously define a function).

Since $p(x \otimes y) \geq 0$ then there is always a normalized wave-function $\Psi \in L^2(X \otimes Y)$ such that $|\Psi(x \otimes y)|^2 = p(x \otimes y)$. Note that the Koopman-von Neumann version of classical statistical mechanics[16] defines classical statistical mechanics as a particular case of quantum mechanics where the algebra of observable operators is necessarily commutative (because the time-evolution is deterministic). In an infinite-dimensional sphere of radius 1 (subset of a real Hilbert space) there is a uniform prior measure induced by the L^2-distance in the Hilbert space. We choose a prior measure (compatible with the uniform prior measure) which excludes all real functions which are approximated by numerical functions with complexity class P, for a high enough resolution of the partition.

Given an input in $([0, 1])^2$ (the input consists of two samples from a uniform distribution in the

interval $[0,1]$, imported from ANU QRNG [17][3] for instance) and a candidate output in $[0,1]$ the wave-function uniquely defines a random symmetry transformation. Such symmetry transformation is not lacking information since it can be inverted (the non-unitary isometries have null prior measure). The cumulative probability distribution is given by the integral of the modulus squared of the wave-function in the corresponding region of the sample space.

From the cumulative marginal probability distribution, such that Y is fully integrated we determine $x \in X$ using the first sample from the uniform distribution in $[0,1]$. From the cumulative conditional probability distribution with the condition that $x \in X$ is what we determined previously, we determine $y \in X$ using the second sample from the uniform distribution in $[0,1]$. We apply the inverse-transform sampling method[18], that is check in the interval of the partition defined by the bits corresponding to X and Y, whether the cumulative distribution crosses the sample from the uniform distribution. This defines the deterministic verification of the candidate output corresponding to the input, in agreement with the Born's rule of Quantum Mechanics.

Note that the resulting deterministic function is not necessarily invertible, because the collapse of the wave-function is irreversible (unless the symmetry transformation would be deterministic, which is excluded in this case because we have a continuous probability space of functions).

4. The classical Turing machine defined as a Quantum computer

The Turing machine can be equivalently defined as the set of general recursive functions, which are partial functions from non-negative integers to non-negative integers[19]. But the set of all functions from non-negative integers to non-negative integers is not suitable to define a measure, since they form an uncountable set, in a context where the continuum is not defined. Moreover, the general recursive functions are based in the notion of computability (that the Turing machine halts in a finite time), but computability does not hold in the limit of an infinite number of input bits, thus to study such limit we need to define uncomputable functions somehow (we will use complete spaces, where Cauchy sequences always converge to an element inside the space).

On the other hand, it is widely believed (and we will show in the following) that any computational problem that can be solved by a classical computer can also be solved by a quantum computer and vice-versa. That is, quantum computers obey the Church–Turing thesis. Note that it is well known that some circuits (classical hardware) provide exponential speedups when compared with some other circuits in some functions (because the input bits can be reparametrized, this is why the time complexity of a function has an upper bound, but it is not known how to establish a lower bound; it is also consistent with the fact that the halting problem is undecidable, that is, given an arbitrary

[3] See also: https://qrng.anu.edu.au/random-binary

function from integers to integers and an arbitrary input, we cannot determine if the output of such function is computable or not), thus the fact that a classical Turing machine can be defined as a Quantum computer is compatible with the fact that quantum computers provide exponential speedups when compared with some classical computers in some functions.

We start by noticing that the domain of a general recursive function can be defined by a dense countable basis of a particular Hilbert space which is the (Guichardet) L^2 completion of the set of all finite linear combinations of simple tensor products of elements of a countable basis of a base Hilbert space, where all but finitely many of the factors equal the vacuum state[20] (like in a Fock-space, but without the symmetrization). But the unitary linear transformations on a normalized wave-function are not necessarily the most general transformations of the probability measure corresponding to the wave-function. Because of that, we build a Fock-space where the base Hilbert space is the previous Guichardet-space, then the unitary transformations on this Fock-Guichardet-space allow us to implement the most general transformations of a probability measure, corresponding to a normalized wave-function in the base Guichardet-space. Note that a countable basis of the Guichardet-space is already made of simple tensor products, and the simple tensor product is associative, thus the Fock-Guichardet-space is isomorphic to the Guichardet-space, but we still prefer to use the Fock-Guichardet-space due to the existence of standard tools for Fock-spaces.

Note that the input of a general recursive function is a finite number of integers, but its output is only one integer. However, any function which outputs several integers is a direct sum of functions which output one integer. The other way around is also true, once we define a vacuum state (included in the Fock-Guichardet-space), that is, a function which outputs one integer is a particular case of a function that outputs several integers where all outputs except one correspond to the vacuum state. Thus, we can consider only unitary automorphisms of the Fock-Guichardet-space.

To be able to define a measure, we make the integers correspond to the (countable) step functions with rational endpoints in the interval $[0, 1]$ and weights which are plus or minus the square root of a rational number[21]. The vacuum state is the constant positive function with norm 1, and it corresponds to the integer 0. We eliminate duplicated step functions in the correspondence with the integers, for instance if two neighbor intervals have the same weight then they are fused.

Then, the limit of infinitesimal intervals is well-defined, and it is defined by an element of $L^2([0, 1])$. Since the general recursive functions are partial functions, then they are a particular case of partially-defined linear operators $L^2([0, 1]) \to L^2([0, 1])$, and we can define the base Hilbert space of the Fock-Guichardet-space as $L^2([0, 1])$.

5. Worst case prior measure, rational functions and radical determinism

In a standard probability space, there are only continuous and/or countable measures. However, these may be mixed in an arbitrary way. For a theory of Physics we could choose the best case prior measure (as we did in the previous sections), since we just want to find a prior which is consistent with the experimental data, without many concerns about alternative priors. However, in Cryptography we need guarantees that our limits are robust under arbitrary choices, so we need to assume the worst case prior measure.

The previous sections could also be made compatible with the worst case prior measure, if we had a computer capable of comparing real numbers not rational. That would be acceptable for a theory of Physics, but it would make it difficult to obtain guarantees for Cryptography.

It is also difficult to guarantee true randomness is real-world applications of Cryptography. Since probabilities only mean incomplete information, we can use Probability Theory in the context of radical determinism (where nothing is random). For Cryptography, we need the worst case prior measure, rational functions and radical determinism.

So we start by eliminating a non-standard probability measure: any probability theory is a universal language (like English or mathematical logic) to define abstract models of the objects we want to study. A standard probability theory is universal and irreducible, meaning that it has the minimal content to be considered a probability theory (in agreement with Quantum Mechanics and Experimental Physics, for instance). If the non-standard probability theory is also irreducible, then the corresponding models are equivalent, and we can use the standard version without loss of generality. This allows us to transfer models between different sciences. But often the non-standard probability theory is reducible, this means that the boundary between model and probability theory is not where it would be in the standard case and there are properties that we are attributing to the probability theory that in fact belong to the model.

Thus, we should assume a standard probability theory and leave some flexibility in the definition of the computer model and not the other way around, as it often happens in Complexity Theory where there are strict axioms for different computer models, while asymptotic limits are taken without defining the probability space, which is recipe to end up with mathematical results and questions which are hard to transfer to experimental physics and many other sciences.

In the context of radical determinism, the history of events is a non-random countable sequence of events. Thus, some events with null measure might happen, due to radical determinism. But only a countable number of such events. That means that a continuous measure is only truly continuous

up to a countable number of points, this possibility is already considered by the worst case prior measure.

Consider now a boolean function of a countable infinite number of bits. The time complexity of almost all Boolean functions on n variables for a Turing machine is at least (and at most, for all Boolean functions) $\mathcal{O}(2^n/(n\log(n)))$[14] [15] which is not polynomial in n, but the same boolean function has different time complexities in different circuits[14] (because the input bits can be reparametrized). Thus, the time complexity of a function depends on the circuit (computer design). Also, an algorithm with polynomial time-complexity is not guaranteed to be fast (due to large order and coefficients of the polynomial), thus only the asymptotic behavior is fast, and so we cannot put an upper bound on the number of bits of the input. But the arbitrarily large number of bits of the input introduces another ambiguity: given any problem with any time complexity (exponential $\mathcal{O}(2^n)$, for instance) there is a problem with linear time-complexity in the number of bits that takes the same amount of time to run. Thus, the polynomial time-complexity is faster than exponential time-complexity in asymptotic behavior, only for the same number of bits of the input. This is a condition without much meaning when the input has a countable infinite number of bits.

What we can say is that the countable (or mixed) measure allows defining functions that eventually have polynomial time complexity (that is, not necessarily non-polynomial time complexity). This contrasts with the necessarily non-polynomial time complexity of the functions defined by the continuous measure. In the following we will show that, given any mixed prior measure, we can always redefine the problem to have a continuous prior measure such that its results cannot be reproduced by a mixed prior measure, effectively converting the worst case prior measure into the best case prior measure.

Converting the worst case measure to best case measure

The prior measure must be mixed or continuous, to allow for the limit of an infinite number of bits of all computable functions. But we must prove that $P \neq NP$ in a single deterministic Turing machine (one with a continuous prior measure, for instance), not in the set of all possible deterministic Turing machines. That would not be possible, since for any computable function f (including any function in NP) there is a deterministic Turing machine where f is in P, by reparametrizing the input bits.

It is not obvious if we can prove $P \neq NP$ in any single deterministic Turing machine, or just in one particular deterministic Turing machine. However, given a worst case prior measure (thus mixed measure), there is a subset of the input random sample which also implements a deterministic Turing machine and where the prior measure is continuous, where $P \neq NP$, thus it becomes legitimate to claim that this fact already shows that $P \neq NP$. Moreover, using subsets of the input random sample (thus, regular conditioned probability) we can create any other prior measure, because any

abelian von Neumann algebra of operators on a separable Hilbert space is *-isomorphic to exactly one of the following:

- $l^{\infty}(\{1,2,...,n\}), n \geq 1$
- $l^{\infty}(\mathbb{N})$
- $L^{\infty}([0,1])$
- $L^{\infty}([0,1] \cup \{1,2,...,n\}), n \geq 1$
- $L^{\infty}([0,1] \cup \mathbb{N})$.

Equivalently, a standard probability space is isomorphic (up to sets with null measure) to the interval $[0,1]$ with Lebesgue measure, a finite or countable set of atoms, or a combination (disjoint union) of both.

Thus, for the worst case prior measure if we include all subsets of the input random sample, then using two integers (which is countable) we include a countable set of deterministic Turing machines $\{k\}$ and countable functions $f_{k,n}$, one machine for each countable function $f_{k,1}$ such that $f_{k,1}$ is in P and $f_{k,n} = f_{n,k}$, then we cannot prove $P \neq NP$ (in fact we would conclude that $P = NP$, due to the possibility of reparametrizing the input bits).

Then, we can only prove $P \neq NP$ in one particular deterministic Turing machine and not in any single deterministic Turing machine.

Given any mixed prior measure, there is an interval of the input random sample where it is continuous. We rescale to $[0,1]$ all intervals corresponding to such interval where the mixed measure is continuous, using conditioned probability. The results in (the new interval of the input random sample) $[0,1]$ cannot be fully reproduced by any other measure which is mixed in $[0,1]$.

Given any other measure which is mixed in $[0,1]$, there is an interval with rational endpoints of the input random sample where there is a finite difference between the two cumulative probability distributions, otherwise both would be continuous. This translates into two different averages of y in an interval (for x) of a partition of $[0,1]$, separated by a finite difference.

Then the indicator function of a y corresponding to the continuous measure, in the interval (for x) of the partition of $[0,1]$, is in NP (more precisely, it can be extended to be in NP, since we only defined for x in a subset of $[0,1]$) but it cannot be reproduced by the mixed measure. It cannot be reproduced by the continuous prior measure either, since a function constant in x in a finite interval has null measure (for a continuous prior measure). Note that the function that corresponds to the indicator function above defined, is a function constant in x in a finite interval.

Note that a continuous prior measure admits a regular conditional probability density, which allows us to define a selection (verification) of a candidate output. The verification of a constant output is in NP and thus in a mixed measure, but it requires one more input (the candidate output) and thus it is compatible with a continuous measure of functions of x. That is, the measure is overall mixed, being continuous only for functions with one input.

6. P≠NP

When using a computer to solve the problem defined in the previous section, any disjoint set of the partition is an interval. This gives us two options and two options only, either the selection of events is fully deterministic or only approximately deterministic:

1. The selection of events is fully deterministic. Then we impose a condition on the output Y of any wave-function in the sphere, this defines a regular conditional probability density for the input X conditioned on a constant rational output y. As shown in the previous section, there is a rational y corresponding to the continuous measure, in an interval with rational endpoints (for x) of the partition of $[0, 1]$, which it cannot be reproduced by a countable prior measure. It cannot be reproduced by the continuous prior measure either, since a function constant in x in a finite interval has null measure (for a continuous prior measure). That is, there is no function in P corresponding to the indicator function for y, which is in NP (more precisely, it can be extended to be in NP, since we only defined it for x in a subset of $[0, 1]$). This implies $P \neq NP$.

2. The selection of events has some randomness, as small as we want. Then we do not impose a condition on the output (eventually we impose a condition on the input X, depending on whether we want fixed or averaged input). In a strict interpretation of the P vs. NP problem this option is excluded by definition since the official formulation assumes that both the verification and the solution of the problem are both fully deterministic. This already implies $P \neq NP$, in a strict interpretation.

Note that a complete history of events needs to be countable, so that we can convert it into a single event (mapping complete histories of events one-to-one to the real numbers in the interval $[0, 1]$, for instance). We could also define a density (that is, yet to be integrated, using the disintegration theorem[22]) of an event. Such density is a regular conditional probability, since regular conditional probabilities always exist in standard probability spaces[11]. But a density cannot correspond to a single event (by definition) and thus it cannot be considered a complete history of events.

This proof is dependent on the fact that the prior measure is continuous. If it in part continuous, and in part countable, then we can choose just the continuous part of the sample space (see the

previous section). While we can use a countable part of the sample space to approximately solve a continuous problem, and a continuous part of the sample space to solve a countable problem, we cannot change the prior measure from continuous to countable or vice-versa (by the Radon-Nikodym theorem), because there is no Radon-Nikodym derivative between the two measures, since the sets of null measure are disjoint between the two measures. The prior measure defines the physical world where the computer exists, thus it cannot be removed from any complete computer model related to a physical computer.

Note that in the first paragraph of the official statement of the P vs. NP problem[7], it is stated:

> *To define the problem precisely it is necessary to give a formal model of a computer. The standard computer model in computability theory is the Turing machine, introduced by Alan Turing in 1936. Although the model was introduced before physical computers were built, it nevertheless continues to be accepted as the proper computer model for the purpose of defining the notion of computable function.*

As for any other proof, this proof is only as good as the axioms used (that is, assumptions). The computer model used for a solid proof of the P vs. NP problem should be widely accepted as a good approximation to a physical computer for the purpose of defining the notion of computable function. We believe our computer model is accepted by most experts in Physics (as argued in the previous section). We claim that our computer model makes no more assumptions than those required by the official statement[7] (including the deterministic Turing machine), and it is as close to a physical computer as possible, by today standards. Assuming a countable prior measure is not realistic (as argued in the previous sections, for instance it would exclude an ensemble of fair coins).

However, we believe that allowing a random selection of events is even more realistic (as discussed in the previous sections, also with implications to Machine Learning and Quantum Mechanics). In the next section, we will define a selection of events which has some randomness (as small as we want) and prove that even in that case, we still have P≠NP.

7. Realistic version of the problem (still P≠NP)

A selection of events which is only approximately deterministic can be approximated by a step function (step functions are dense in L^2) and thus there is a square with non-null constant measure. We rescale such square to $[0,1] \times [0,1]$. We then consider a real polynomial wave-function that is near the point in the sphere corresponding to a constant wave-function, up to an error in the L^2 norm which can be as small as we want because the polynomials are dense in L^2 (the corresponding numerical polynomial does not need to be in P).

The first sample from the uniform distribution defines directly $x \in X$. An approximation (in the L^2 norm) with polynomial time-complexity to the selection function, is defined by setting $y \in Y$ equal to the second sample from the uniform distribution.

Since the wave-function is polynomial non-constant, then the corresponding cumulative probability distribution minus the second sample is strictly crescent (except in sets of null measure). Thus, when we define the corresponding deterministic function we can choose the second sample which produces an output which is as far from zero as we want (in the interval $[0, 1]$), because we are using the L^∞ norm now. We cannot average over the random sample, otherwise we need a random computer (see next section). Thus, no approximation is possible, and it suffices that we define a partition for the output which has two disjoint sets (the measures of the sets are arbitrary, as long as they are non-null) and a numerical output with one bit. Then, almost all numerical functions are not in the P class, according to the prior measure.

8. Generation of random numbers has linear time-complexity

The two random samples from a uniform distribution in the interval $[0, 1]$ are inputs, in the deterministic Turing machine. However, in the real-world these samples need to be generated somewhere and in polynomial time-complexity, otherwise the time complexity of the random selection computed by the real-world random computer could be non-polynomial in the number of bits of the random samples.

Moreover, it would be better if the generation of random samples had linear time complexity, since then we could do a constant rate of experiments over time to validate the probability distribution of the random selection, otherwise it would be impractical to generate an infinite sequence of experiments.

We cannot prove this mathematically (since we would need more axioms). However, the implications of this article to Quantum Mechanics help to clarify the source of randomness of Quantum Mechanics (and thus of the random samples). It is relevant to verify empirically that the generation of random numbers with linear time complexity is possible, for all practical purposes. We can visually check on the website from ANU QRNG [4] that the number of bits of the random sample grows linearly in time and any complete history of events converges to a uniform probability distribution.

Moreover, the entropy is maximal, in the sense that the deterministic function needed to correlate the bits is not computable for all practical purposes, not even approximately (since L^∞ is non-separable) according to the prior measure.

[4] See also: https://qrng.anu.edu.au/random-binary

9. On the consequences to Machine Learning

In the introduction we discussed the implications (of the results of this article), which are common to Machine Learning and Quantum Mechanics. But Machine Learning (for instance Deep Neural Networks) is not firmly based in probability theory, unlike Quantum Mechanics, then there are more consequences.

In Machine Learning, methods inspired by probability theory are used often[2], but the formalism is based in approximations to deterministic functions, guided by a distance (or equivalently, an optimization problem) and not a measure. In fact, two of the main open problems are the alignment of models and the incorporation of prior knowledge[23], which could be both well solved by a prior measure if there would be any measure defined.

Our results imply that under reasonable assumptions, almost all functions are not computable not even approximately. Thus, Machine Learning works because the functions we are approximating are in fact probability distributions (eventually after some reparametrization[24]). This shouldn't be surprising, since Classical Information Theory shows (under reasonable assumptions) that probability is unavoidable when we are dealing with representations of knowledge/information[1][2]. But in Machine Learning the probability measure is not consistently defined (despite that many methods are inspired by probability theory), the probability measure emerges from the approximation[24] and often in an inconsistent way. The inconsistency is not due to a lack of computational power since modern neural networks can fit very complex deterministic functions and fail badly[25][26] in relatively simple probability distributions (e.g. catastrophic forgetting or the need of calibration to have some probabilistic guarantees[26]).

This unavoidable emergence of a probability measure should be investigated as a potential source of inefficiency, inconsistency and even danger. If the emergence of a probability measure is unavoidable, why don't we just define a probability measure in the formalism consistently? Many people say "it is how our brain works", so mathematics should step aside when there is empirical evidence.

But the empirical evidence is: oversized deep neural networks still generalize well, apparently because often the learning process converges to a local maximum (of the optimization problem) near the point where the learning begun[27]. This implies that if we repeat the learning process with a random initialization (as we do when we consider ensembles of neural networks[25][28]), then we do not expect the new parameters to be near any particular value, regardless of the result of the first learning process. This expectation is justified by the fact that every three layers of a wide enough neural network is a universal approximator of a function[29], so any deviation introduced by three layers can be fully corrected in the next three layers, when composing dozens or hundreds of layers as we do in a deep neural network. Then the correlation between the parameters corresponding to

different local maximums converges to zero, when the number of layers increases.

Thus, there is empirical evidence that oversized deep neural networks still generalize well, precisely because a prior measure emerges: deep learning does not converge to the global maximum and instead to one of the local maximums chosen randomly, effectively sampling from a prior measure in the sample space defined by all local maximums. This is consistent with the good results achieved by ensembles of neural networks[25][28], which mimic many samples. However, it is a prior measure which we cannot easily modify or even understand, because the measure space is the set of all local maximums of the optimization problem. But, since we expect the parameters to be fully uncorrelated between different local maximums, then many other prior measures (which we can modify and understand, such as the uniform measure) should achieve the same level of generalization.

This is not a surprise, since oversized statistical models that still generalize well were already found many decades ago by many people[12]: a standard probability space with a uniform probability measure can be infinite-dimensional (the sphere[12] studied in this article, for instance).

More empirical evidence: no one looks to a blurred photo of a gorilla and says with certainty that it is not a man in a gorilla suit. We all have many doubts, when we are not sure about a subject we usually express doubts through the absence of an action (not just us, but also many animals), for instance we don't write a book about the subject we don't know about.

There is no empirical evidence that our brain tries to create content which is a short distance from content (books, conversations, etc.) created under exceptional circumstances (when doubts are minimal). When we are driving, and we do not know what is in front of us, we usually just slow down or stop the car. But what content defines "not knowing"? Is there empirical evidence about the unknown? The unknown can only be an abstract concept, expressed through probability theory or a logical equivalent. Is there empirical evidence that probabilities are reducible, that there is a simpler logical equivalent? No, quite the opposite.

The only trade-off seems to be between costs (time complexity, etc.) and understanding/control. A prior measure which we understand and/or control may mean much more costs than an emergent (thus, inconsistent and uncontrollable) prior measure which just minimizes some distance. But this trade-off is not new, and it is already present in all industries which deal with some safety risk (which is essentially all industries). Distances are efficient for proof of concepts (pilot projects), when the goal is to show that we are a short distance from where we want to be. But safety (as most features) is not being at a short distance from being safe[5]. "We were at a short distance from avoiding nuclear annihilation" is completely different from "we avoided nuclear annihilation". To avoid nuclear annihilation we need (probability) measures, not only distances.

[5]See for instance https://edition.cnn.com/2023/04/29/us/ai-scam-calls-kidnapping-cec

References

[1] Novak, E, and H Wozniakowski. 2008. "Tractability of Multivariate Problems, Volume I: Linear Information, European Math." *Soc., Zürich* 2 (3).

[2] Hennig, Philipp, Michael A Osborne, and Hans P Kersting. 2022. *Probabilistic Numerics: Computation as Machine Learning*. Cambridge University Press. https://www.probabilistic-numerics.org/assets/ProbabilisticNumerics.pdf.

[3] Ritter, K. 2000. *Average-Case Analysis of Numerical Problems*. Nr. 1733. Springer.

[4] ———. 2020. "Bayesian Numerical Analysis." Encyclopedia of Mathematics. http://encyclopediaofmath.org/index.php?title=Bayesian_numerical_analysis&oldid=50648.

[5] Weihrauch, Klaus, and Ning Zhong. 2002. "Is Wave Propagation Computable or Can Wave Computers Beat the Turing Machine? Proc." *Proceedings of the London Mathematical Society* 85 (September). https://doi.org/10.1112/S0024611502013643.

[6] Toland, John. 2020. *The Dual of $L^{\infty}(X,L,\lambda)$, Finitely Additive Measures and Weak Convergence: A Primer*. Springer International Publishing. https://doi.org/10.1007/978-3-030-34732-1.

[7] Cook, Stephen. 2000. "The P Versus NP Problem." *Clay Mathematics Institute*. https://www.claymath.org/sites/default/files/pvsnp.pdf.

[8] Cook, Stephen A. 1972. "A Hierarchy for Nondeterministic Time Complexity." In *Proceedings of the Fourth Annual ACM Symposium on Theory of Computing*, 187–92. STOC '72. Denver, Colorado, USA: Association for Computing Machinery. https://doi.org/10.1145/800152.804913.

[9] Levin, Leonid A. 1986. "Average Case Complete Problems." *SIAM Journal on Computing* 15 (1): 285–86. https://www.gwern.net/docs/cs/algorithm/1986-levin.pdf.

[10] Impagliazzo, Russel. 1995. "A Personal View of Average-Case Complexity." In *Proceedings of Structure in Complexity Theory. Tenth Annual IEEE Conference*, 134–47. IEEE. https://www2.karlin.mff.cuni.cz/~krajicek/ri5svetu.pdf.

[11] Durrett, R. 2019. *Probability: Theory and Examples*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

[12] Peterson, Amy. 2019. "Gaussian Limits and Polynomials on High Dimensional Spheres." Doctoral Dissertations. Phdthesis, Storrs, CT USA: University of Connecticut. https://opencommons.uconn.edu/dissertations/2137.

[13] Eaton, Morris L., and David A. Freedman. 2004. "Dutch Book Against Some 'Objective' Priors." *Bernoulli* 10 (5): 861–72. https://doi.org/10.3150/bj/1099579159.

[14] Shannon, Claude E. 1949. "The Synthesis of Two-Terminal Switching Circuits." *The Bell System Technical Journal* 28 (1): 59–98.

[15] Pippenger, Nicholas, and Michael J. Fischer. 1979. "Relations Among Complexity Measures." *J. ACM* 26 (2): 361–81. https://doi.org/10.1145/322123.322138.

[16] Sudarshan, E. C. G. 1976. "Interaction Between Classical and Quantum Systems and the Measurement of Quantum Observables." *Pramana* 6 (3): 117–26. https://doi.org/10.1007/BF02847120.

[17] Haw, J. Y., S. M. Assad, A. M. Lance, N. H. Y. Ng, V. Sharma, P. K. Lam, and T. Symul. 2015. "Maximization of Extractable Randomness in a Quantum Random-Number Generator." *Phys. Rev. Appl.* 3 (5): 054004. https://doi.org/10.48550/arXiv.1411.4512.

[18] Rubinstein, Reuven Y, and Dirk P Kroese. 2016. *Simulation and the Monte Carlo Method; 3rd Ed.* Wiley Series in Probability and Statistics. Wiley.

[19] Dean, Walter. 2021. "Recursive Functions." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Winter 2021. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2021/entries/recursive-functions/.

[20] Bratteli, O., and D.W. Robinson. 1987. *Operator Algebras and Quantum Statistical Mechanics 1: C*- and W*-Algebras. Symmetry Groups. Decomposition of States*. 2nd edition. Operator Algebras and Quantum Statistical Mechanics. Springer.

[21] Royden, Halsey, and Patrick Fitzpatrick. 2010. *Real Analysis, Fourth Edition*. 4th ed. Prentice Hall.

[22] Chang, Joseph T, and David Pollard. 1997. "Conditioning as Disintegration." *Statistica Neerlandica* 51 (3): 287–317.

[23] Müller, Samuel, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. 2022. "Transformers Can Do Bayesian Inference." In *International Conference on Learning Representations*. https://openreview.net/forum?id=KSugKcbNf9.

[24] Anonymous. 2023. "Value-Probability Duality of Neural Networks." *OpenReview Preprint*. https://openreview.net/forum?id=nHGkRwmztoQ.

[25] Lakshminarayanan, Balaji, Alexander Pritzel, and Charles Blundell. 2017. "Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles." *Advances in Neural Information Processing Systems* 30. https://proceedings.neurips.cc/paper/7219-simple-and-scalable-predictive-uncertainty-estimation-using-deep-ensembles.pdf.

[26] Grancey, Florence de, Jean-Luc Adam, Lucian Alecu, Sébastien Gerchinovitz, Franck Mamalet, and David Vigouroux. 2022. "Object Detection With Probabilistic Guarantees." In *Fifth International Workshop on Artificial Intelligence Safety Engineering (WAISE 2022)*. SAFECOMP 2022, LNCS 13415. München, Germany. https://hal.archives-ouvertes.fr/hal-03769683.

[27] Li, Yuanzhi, and Yingyu Liang. 2018. "Learning Overparameterized Neural Networks via Stochastic Gradient Descent on Structured Data." *Advances in Neural Information Processing Systems* 31. https://proceedings.neurips.cc/paper/8038-learning-overparameterized-neural-networks-via-stochastic-gradient-descent-on-structured-data.pdf.

[28] Egele, Romain, Romit Maulik, Krishnan Raghavan, Bethany Lusch, Isabelle Guyon, and Prasanna Balaprakash. 2022. "Autodeuq: Automated Deep Ensemble with Uncertainty Quantification." In *2022 26th International Conference on Pattern Recognition (ICPR)*, 1908–14. IEEE. https://doi.org/10.48550/ARXIV.2110.13511.

[29] Hornik, Kurt, Maxwell Stinchcombe, and Halbert White. 1989. "Multilayer Feedforward Networks Are Universal Approximators." *Neural Networks* 2 (5): 359–66.