Ultimate Intelligence

Eray Özkural

Gök Us Sibernetik Ar&Ge Ltd. Şti.

Abstract. We argue that Solomonoff induction is universal and complete in the physical sense via several strong physical arguments. We argue that Solomonoff induction is fully applicable to quantum mechanics. We show how to choose an objective reference machine for universal induction by defining a physical message complexity for the presently considered ultimate intelligence research program. We introduce physical measures and limits of intelligence that are rigorously objective. We propose volume and energy measures that are more appropriate for physical computations. We extend logical depth and conceptual jump size measures in AIT to stochastic domains and physical measures that involve volume and energy. We show the relations between energy, logical depth and volume of computation for AI. We also introduce a highly relaxed, general physical model of computation that we believe to be more appropriate for our universe than other low level automata models previously considered. We briefly apply our ideas to the physical limits of computation in our universe to show the relation to ultimate intelligence.

1 Introduction

Ray Solomonoff has discovered algorithmic probability and introduced the universal induction method which is the foundation of general-purpose Artificial Intelligence (AI) theory [17,18,21]. In the present paper, we investigate the ultimate limits of intelligence in our physical universe. Although the theory of Solomonoff induction is independent of a physical theory, we interpret it physically and try to refine the understanding of the theory given constraints of physical law. First, we argue that its completeness and universality are compatible with contemporary physical theory, for which we give arguments from quantum mechanics and general relativity that show Solomonoff induction to be satisfactory for all possible physical prediction problems. Second, we define a physical message complexity measure, and show this to be the ultimate choice of a reference machine which gives us the advantage of objective predictions for physical problems and the typical disadvantages of using low-level reference machines. However, in theory, we show that setting the reference machine to the universe does have benefits, eliminating many constants from algorithmic

1 This working paper was initially submitted to ALT-2014. We are going to respond to reviewers on the arXiv.
information theory (AIT). We frame the question of ultimate intelligence in a
physical setting, for this we give a general definition of an intelligent system and
a physical performance criterion, which turns out to be a relation of physical
quantities and information. We generalize logical depth and conceptual jump
size to stochastic sources and consider the influence of volume, space and en-
ergy. We consider the energy efficiency of computing as an important parameter
for an intelligent system, forgoing other details of a universal induction approx-
imation. We thus relate the ultimate limits of intelligence to physical limits of
computation that were derived.

2 Background

Let us recall Solomonoff’s universal distribution. Let \( U \) be a universal computer
which runs programs with a prefix-free encoding like LISP. The algorithmic
probability that a bit string \( x \in \{0,1\}^+ \) is generated by a random program
\( \pi \in \{0,1\}^+ \) of \( U \) is:

\[
P_U(x) = \sum_{U(\pi) = x(01)^+} 2^{-|\pi|}
\]

which conforms to Kolmogorov’s axioms [10]. \( P_U(x) \) considers any continuation
of \( x \), taking into account non-terminating programs. \( P_U \) is also called the uni-
versal prior for it may be used as the prior in Bayesian inference, for any data
can be encoded as a bit string.

We also give the basic definition of Algorithmic Information Theory (AIT),
where the algorithmic entropy, or complexity of a bit string \( x \in \{0,1\}^+ \) is

\[
H_U(x) = \min(|\pi| \mid U(\pi) = x)
\]

We shall now briefly recall the well-known Solomonoff induction method [17][18]. Universal sequence induction method of Solomonoff works on bit strings
\( x \) drawn from a stochastic source \( \mu \). Equation 1 is a semi-measure, but that is
easily overcome as we can normalize it. We merely normalize sequence probabil-
ities

\[
P'(x_0) = \frac{P(x_0).P'(x)}{P(x_0) + P(x_1)}
\]

\[
P'(x_1) = \frac{P(x_1).P'(x)}{P(x_0) + P(x_1)}
\]

eliminating irrelevant programs and ensuring that the probabilities sum to 1,
from which point on \( P'_U(x_0|x) = P'_U(x_0)/P'_U(x) \) yields an accurate prediction.
The error bound for this method is the best known for any such induction
method. The total expected squared error between \( P'_U(x) \) and \( \mu \) is

\[
E_P \left[ \sum_{m=1}^{n} (P'_U(a_{m+1} = 1|a_1a_2...a_m) - \mu(a_{m+1} = 1|a_1a_2...a_m))^2 \right] \leq \frac{1}{2} \ln P_U(P_1)
\]
which is less than \(-1/2 \ln P'_U(\mu)\) according to the convergence theorem proven in [20], and it is roughly \(K_U(\mu) \ln 2\) [23]. Naturally, this method can only work if the algorithmic entropy of the stochastic source \(K_U(\mu)\) is finite, i.e., the source has a computable probability distribution. The convergence theorem is extremely important, because it shows that Solomonoff induction has the best generalization performance among all prediction methods. In particular, the total error is expected to be a constant independent of the input, and the error rate will rapidly decrease with increasing input size.

3 Completeess of Universal Induction

Solomonoff induction model is known to be complete and incomputable. This is a consequence of the undecidability of the halting problem. Equation 1 enumerates a non-trivial property of all programs (the identity of a program’s output to a regular language), which makes it an incomputable function. It is more properly construed as a semi-computable function that may be approximated arbitrarily well in the limit. Solomonoff has shown that the incomputability of algorithmic probability does not inhibit its application in any way whatsoever, and emphasized this often misunderstood point in a number of publications.

The only remaining assumptions for Solomonoff induction to work in general, for any \(\mu\) are a) that we have picked a universal reference machine (which we assume in this paper), and b) that \(\mu\) has a computable probability density function (PDF).

The second assumption warrants our attention when we consider modern physical theory. We formalize the computability of \(\mu\) as follows:

\[
H_U(\mu) \leq k, \exists k \in \mathbb{Z}
\]  

which suggests that the PDF \(\mu(x)\) may be simulated on a computer, while \(x\) are stochastic.

3.1 Evidence from physics

There is an exact correspondence of such a construct in physics, which is the quantum wave function. The wave function of a finite quantum system is defined by a finite number of parameters, although its product with its conjugate is a PDF from which we sample stochastic observations. Since, it is irrational to consider an infinite quantum system in the finite observable universe, therefore, \(\mu\) can model the statistical behavior of matter for any quantum mechanical source. This is the first evidence of true, physical universality of Solomonoff induction we will consider, which is complemented by the definition of von Neumann entropy of a quantum system, which we recall for the sake of completeness. The von Neumann entropy of a quantum system described by a density matrix \(\rho\) is

\[
S = - \text{tr}(\rho \ln \rho) = - \sum_j \eta_j \ln \eta_j
\]
where tr is the trace of a matrix and \( \rho = \sum_j \eta_j \langle j | j \rangle \) is decomposed into its eigenvectors. Apparently, von Neumann entropy is equivalent to classical entropy and suggests a computable PDF, which is expected since we took \( \rho \) to be a finite matrix. Furthermore, the dynamic time evolution of a wave function is known to be unitary, which entails that if \( \mu \) is a quantum system, it will remain computable dynamically. Therefore, if \( \mu \) is a quantum system with a finite density matrix, Equation 5 holds.

The second piece of evidence from physical theory is that of universal quantum computer, which shows that any local quantum system may be simulated by a universal quantum computer [13]. Since a universal quantum computer is Turing-equivalent, this means that any local quantum system may therefore be simulated on a classical computer. This fact has been interpreted as a physical version of Church-Turing thesis by the quantum computing pioneer David Deutsch, in that 'every finitely realizable physical system can be perfectly simulated by a universal model computing machine operating by finite means' [3]. As a quantum computer may be construed as a probabilistic computer, whose outputs are probabilistic after decoherence, these two facts together entail that the PDF of a local quantum system is always computable. Which yields our second conclusion. If \( \mu \) is a local quantum system, Equation 5 holds.

The third piece of evidence from physics is that of the famous Bekenstein bound and the holographic principle. Bekenstein bound was originally conceived for black holes, however, it applies to any physical system, and states that any finite energy system enclosed within a finite volume of space will have finite entropy:

\[
S \leq \frac{2\pi k R E}{\hbar c}
\]  

(8)

where \( S \) is entropy, and \( R \) is the radius of the sphere that encloses the system, \( E \) is the total energy of the system including masses, and the rest are familiar physical constants. Such a finite entropy readily transforms into Shannon entropy, and corresponds to a computable PDF. The inequality is merely a physical elucidation of Equation 6. Therefore, if \( \mu \) is a finite-size and finite-energy physical system, Equation 5 holds.

Contemporary cosmology also affirms this observation, as the entropy of the observable universe has been estimated, and is naturally known to be finite [5]. Therefore, if contemporary cosmological models are true, any physical system in the observable universe must have finite entropy, thus validating Equation 5 holds.

Thus, since we have shown wide-reaching evidence for the computability of PDF of \( \mu \) from quantum mechanics, general relativity, and cosmology, we conclude that contemporary physical science strongly and directly supports the universal applicability of the convergence theorem. In other words, it has been physically proven.

### 3.2 Randomness, computability and quantum mechanics

Recall that the induction operator is an infinite mixture of all matching computations, another way to view it is as a mixture of PDF’s. Wood et. al wrote of
this as a “universal mixture” [26], which is essentially an infinite mixture of all possible computations that match the input. This entails that it should model even random events, due to Chaitin’s strong definitions of algorithmic randomness [3]. More expansive definitions of randomness are not empirically justifiable since they refer to uncountable sets. It is therefore impossible to imagine what more is needed. Our analysis is that stronger definitions of randomness are not needed as they would be referring to oracle machines, which would be truly incomputable, and by our arguments in this paper, have no physical relevance. Note that the halting probability is semi-computable.

The computable pdf model is ultimately a perfect abstraction of the wave function in quantum mechanics. In quantum mechanics, too, the wave function itself has finite description (finite entropy), with unitary evolution, while the observations are stochastic. That is to say, Solomonoff induction is complete with respect to quantum mechanics, as well, even when we assume the reality of non-determinism (which many interpretations of QM do admit). In other words, such claims that Solomonoff induction is not complete could only be true if and only if either physical Church-Turing thesis could be false, or if hypercomputers were possible (which seem like equivalent statements).

4 On The Existence of an Optimal $U$

4.1 Invariance theorem and choice of reference machines

The universal induction model is seen to be subjective, as the expected generalization error depends on the choice of a universal computer $U$ as the convergence theorem shows. This choice is natural according to a Bayesian interpretation of learning as $U$ may be considered to encode the subjective knowledge of the observer.

Furthermore, invariance theorem may be interpreted to imply that the choice of a reference machine is irrelevant, since any universal computer $U_1$ may simulate any other universal computer $U_2$. Recall that the invariance theorem is that the algorithmic complexity changes at most by a constant when a universal reference machine changes.

$$H_{U_1}(x) \leq H_{U_2}(x) + H_{U_1}(s)$$

where the simulation program $s$ simulates $U_2$ on $U_1$.

$$\forall x, U_2(x) = U_1(s, x)$$

Nevertheless, in the real world, there is a sense in which LISP is a more natural reference machine than C++ in at least two ways:

1. The constants in AIT are smaller for LISP than C++
2. The programs in C++ are much longer for many “natural” functions, i.e., LISP captures our intuitive notion of complexity better.
And even though LISP seems better for AIT than C++ in practice, simulating C++ in LISP requires a long program, which diminishes the usefulness of Equation 9. Therefore, it is reasonable to ask in which well-defined sense LISP is better than C++, and whether we can define an optimal $U$.

4.2 The universe as the reference machine

In the following, we shall examine a sense which we may consider the best choice for $U$. Solomonoff himself mentioned such a choice [24]:

For quite some time I felt that the dependence of ALP on the reference machine was a serious flaw in the concept, and I tried to find some “objective” universal device, free from the arbitrariness of choosing a particular universal machine. When I thought I finally found a device of this sort, I realized that I really didn’t want it - that I had no use for it at all! Let me explain:

In doing inductive inference, one begins with two kinds of information: First, the data itself, and second, the a priori data - the information one had before seeing the data. It is possible to do prediction without data, but one cannot do prediction without a priori information. In choosing a reference machine we are given the opportunity to insert into the a priori probability distribution any information about the data that we know before we see it.

If the reference machine were somehow objectively chosen for all induction problems, we would have no way to make use of our prior information. This lack of an objective prior distribution makes ALP very subjective - as are all Bayesian systems.

We proposed a philosophical solution to this problem in a previous article where we made a physical interpretation of algorithmic complexity, by setting $U$ to the universe itself [15]. This was achieved by adopting a physical definition of complexity, wherein program length was interpreted as physical length. The correspondence between spatial extension and program length directly follows from the proper physicalist account of information, for every bit extends in space. Which naturally gives rise to the definition of physical message complexity as the volume of the smallest machine that can compute a message, eliminating the requirement of a reference machine. There are a few difficulties with such a definition of complexity whose analysis is in order. Contrast also with thermodynamic entropy and Bennett’s work on physical complexity [27,1].

In the present article, we support the above philosophical solution to the choice of the reference machine with mathematical observations. Let us define physical message complexity:

\[ C(x) = \min \{ V(M) \mid M \to x \} \]  

(11)

where $x \in D^+$ is any d-ary message written in an alphabet $D$, $M$ is any physical machine that emits the message $x$ (denoted $M \to x$), and $V(M)$ is the volume
of machine $M$. $M$ is supposed to contain all physical computers that can emit message $x$.

Equation 11 is too abstract and it would have to be connected to physical law to be useful. However, it allows us to reason about the constraints we wish to put on physical complexity. If we imagine what sort of device $M$ would be, $M$ is supposed to contain every possible physical computer that can emit a message. For this definition to be useful, the concept of emission would have to be determined. Imagine for now that the device emits photons that can be detected by a sensor, interpreting the presence of a photon as 1 and absence 0. One would first have to build a device that can output a single bit and stops before outputting any messages. It might be hard or impossible for us to build the *smallest* device that can do this. However, let us assume that such a device can exist. It is likely that this minimal hardware would occupy quite a large volume compared to the output it emits. With every added “bit” of message complexity, the smallest device would have to get larger.

We may consider additional complications. For instance, we may demand that these machines do not receive any physical input, i.e., supply their own energy. We note that time can also be put into this picture.

When we use $C(x)$ instead of $H_U(x)$, we do not only eliminate the need for a reference machine, but we also eliminate many constraints and constants in AIT. First of all, there is not the same worry of a self-delimiting program, because every physical machine that can be constructed will either emit a message or not in isolation, although its meaning slightly changes and will be considered in the following. Secondly, we expect all the basic theorems of AIT to hold, while the arbitrary constants that correspond to glue code to be eliminated. Recall that the constants in AIT usually correspond to such elementary operations as function composition and so forth. Let us consider the sub-additivity of information which represents a good example:

$$H(x, y) = H(x) + H(y|x) + O(1)$$  \hspace{1cm} (12)

The program that corresponds to the right-hand size, which generates bit string $xy$, may be considered to be $(\text{defun } p_1 \ldots) (\text{defun } p_2 \ldots) (\text{BIT-CONCAT } (p_1) (p_2\ p_1))$ for a LISP-based algorithmic complexity definition where $p_1$ generates $x$ and $p_2$ generates $y$ given an optimal program for $x$, and BIT-CONCAT would concatenate bit strings. The program schema illustrates the correctness of Equation 12: there is a small constant program that needs to be added to $p_1$ and $p_2$ in order to obtain a near-optimal code for $xy$. Such a definition is natural in a LISP-like language and is among the best that we can have among choices of reference machines (note that Chaitin’s definition of LISP-based complexity is different than this one). When we consider $C(x, y)$, however, the sub-additivity of information becomes exactly

$$C(x, y) = C(x) + C(y|x)$$  \hspace{1cm} (13)

since there does not need to be a gap between a machine emitting a photon and another sensing one. In the consideration of an underlying physical theory
of computing (like quantum computing), the relations may further change, and become clearer.

From the viewpoint of AI theory, however, what we are interested in is whether the elimination of a reference machine may improve the performance of machine learning. Recall that the convergence theorem is related to the algorithmic entropy of the stochastic source with respect to the reference machine. A reasonable concern in this case is that the choice of a “bad” reference machine may inflate the errors prohibitively for small data size, for which induction works best, i.e., as mentioned before the basically functional composition of a physical system may be poorly reflected in an artificial language such as C++, increasing generalization error. On the other hand, setting $U$ to the universe obtains an objective measurement, which does not depend on subjective choices, and furthermore, always corresponds well to the actual physical complexity of the stochastic source. First, let us derive the convergence theorem with respect to $C(x)$. We shall first need to re-define algorithmic probability for an alphabet of $D$.

$$P(x) = \frac{\sum_{M \rightarrow D} D^{-V(M)}}{\sum_{M \rightarrow D} D^{-V(M)}}$$

(14)

Here, it does not matter that program-encodings of $M$ are prefix-free, because infinity is not a valid concern in physical theory. Due to general relativity, there cannot be any influence from beyond the observable universe, i.e., there is not enough time for any message to arrive from beyond it, even if there is anything beyond the cosmic horizon. Therefore, the volume $V(M)$ of the largest machine is constrained by the volume of the observable universe, e.g., it is finite. Hence, the sum always converges.

Solomonoff’s observation that subjectivity is required to solve any problem of significant complexity is of paramount importance. Our proposal of using a physical measure of complexity for objective inference does not neglect that property of universal induction. Instead, we observe that a guiding PDF contains prior information in the form of a PDF. Let $U_1$ be a universal computer that contains much prior information about a problem domain, based on a universal computer $U$ that does not contain any significant information. Such prior information may always be split off to a memory bank.

$$P_{U_1}(x) = P_U(x|M)$$

(15)

Therefore, we can use a conditional physical message complexity given a memory bank to account for prior information, instead of modifying a PDF. Subjectivity is thus retained.

Choosing the universe as $U$ has a particular disadvantage of using the lowest possible level computer architecture. We have not yet formulated complete descriptions of the computation at the lowest level of the universe, therefore further research is needed. However, for solving problems at macro-scale, and/or from artificial sources, algorithmic information pertaining to such domains must be encoded as prior information in $M$, since otherwise solution would be infeasible.
The argument from practical finiteness of the universe was mentioned briefly
by Solomonoff in [19]. Let us note, however, that the abstract theory of algo-
rithmic probability implies an infinite probabilistic universe, in which every
program may be generated, and each bit of each program is equiprobable. In
such an abstract universe, a Boltzmann Brain, with considerably more entropy
than our humble universe is even possible, although it has a very small proba-
bility. In a finite observable universe with finite resources, however, we obtain
a slightly different picture, for instance a Boltzmann Brain is improbable, and
a Boltzmann Brain with a much greater entropy than our universe would be
impossible. Obviously, in a sequence of universes with increasing volume of ob-
servable universe, the limit would be much like pure algorithmic probability.
However, for our definition of physical message complexity, a proper physical
framework is much more appropriate, and such considerations quickly veer into
the territory of metaphysics (since they truly consider universes with physical
law unlike our own). Thus firmly footed in contemporary physics, we gain a
better understanding of the limits of ultimate intelligence.

Note that it is well possible to extend the proposal in this section to a quan-
tum version of AIT by setting $U$ to a universal quantum computer. There are
likely other advantages of using a universal quantum computer, e.g., efficient
simulation of physical systems.

5 Physical Quantification of Intelligence

5.1 Universal measures of intelligence

There is much literature on the subject of defining a measure of intelligence.
Hutter has defined an intelligence order relation in the context of his reinforce-
ment learning (RL) model AIXI [6], which suggests that intelligence cor-
responds to the set of problems an agent can solve. Also notable is the universal
intelligence measure [8,9], which is again based on the AIXI model. Their unive-
sal intelligence measure is based on the following informal definition of inte-
lligence:

intelligence measures an agent’s ability to achieve goals in a wide range
of environments

therefore assuming that intelligence requires being an autonomous agent, based
on a thorough review of several definitions of intelligence in AI literature. The
intelligence measure of [8] is defined as

$$\Upsilon(\pi) = \sum_{\mu \in E} 2^{-H_U(\mu)} V_\mu^\pi$$ (16)

where $\mu$ is a computable reward bounded environment, And $V_\mu^\pi$ is the expected
sum of future rewards in the total interaction sequence of agent $\pi$.

Although this definition corresponds to any kind of reinforcement-learning
or goal-following agent in AI literature quite well, and can be adapted to solve
other kinds of problems, its definition of autonomy and rewards, as they are
originally biological concepts, slightly distract from the pure mathematical core
of intelligence, which may be defined as the ability to solve problems at all.
5.2 A Physical Measure of Operator Induction

Following Equation 16, we propose a metric by which operator induction performs well. Operator induction can be solved by finding in available time a set of operator models $O_j(\cdot|\cdot)$ such that the following goodness of fit is maximized

$$\Psi(\mu) = \sum_j a^j_n$$  \hspace{1cm} (17)

for a stochastic source $\mu$ where each term in the summation is defined as

$$a^j_n = P_U(O^j(\cdot|\cdot)) \prod_{i=1}^n O^j(A_i|Q_i).$$  \hspace{1cm} (18)

$Q_i$ and $A_i$ are question/answer pairs in the input dataset, and $O_j$ is a computable conditional PDF (CPDF) in Equation 13. We can use the found operators to predict unseen data

$$P_U(A_{n+1}|Q_{n+1}) = \sum_{j=1}^n a^j_n O^j(A_{n+1}|Q_{n+1})$$  \hspace{1cm} (19)

The goodness of fit in this case strikes a balance between high a priori probability and reproduction of data like in MML, yet uses a universal mixture like in sequence induction. The convergence theorem for operator induction was proven in [22].

Note that operator induction is insufficient to describe universal agents such as AIXI, because classical induction is inappropriate for solving optimization problems [6]. However, a modified Levin search procedure can solve such optimization problems as in finding an optimal control program [16]. In OOPS-RL, the perception module searches for the best world-model given the history of sensory input and actions in allotted time using OOPS, and the planning module searches for the best control program using the world-model of the perception module to determine the control program that maximizes cumulative reward likewise. In this paper, we consider the perception module of such a generic agent which must produce a world-model, given sensory input.

We cannot use the intelligence measure Equation 16 directly in a physical theory of intelligence because it contains terms that have not been reduced to the language of physics (i.e., we are looking for a reductive definition). We therefore attempt to obtain such a measure using the more benign goodness-of-fit (Equation 17). Let the universal measure of the fitness of operator induction be defined as

$$\Upsilon(\pi) = \sum_{\mu \in S} 2^{-H_U(\mu)} \psi(\mu).$$  \hspace{1cm} (20)

where $S$ is the set of stochastic sources in the observable universe $U$. This would obviously be maximal if we assume that operator induction were solved by an oracle machine. Note that $H_U(\mu) \leq H(\text{Universe})$ where $H(\text{Universe})$ denotes
the total entropy of the observable universe. $\psi(\mu)$ is likewise bounded by the amount of computation we can spend on approximating operator induction.

This is a very loose definition as it assumes that we have unrestricted access to any stochastic source $\mu$ which is not the case in physics, and ignores the constraints from distance, and distribution of cosmic sources. Surely, the probability distribution of programs in a highly evolved world like Earth and the midst of intergalactic space is not quite the same. It may be possible to relate such calculations to cosmic microwave background radiation and average power spectrum of the cosmos. Such astrophysical limits are considered to be open research subjects.

6 Physical Limits to Intelligence

6.1 Logical depth and conceptual jump size

The physical limits to OOPS based on Conceptual Jump Size were examined in \[10\]. Here, we give a more detailed treatment. Let $\pi^*$ be the computable CPDF that exactly simulates $\mu$ with respect to $U$, for operator induction.

$$
\pi^* = \arg\min_{\pi_j} \{|\pi_j| \mid \forall x, y \in \{0, 1\}^* : U(\pi_j, x, y) = P(\mu = x|y)\}
$$

(21)

Recall that conceptual jump size of inductive inference (CJS) can be defined with respect to the optimal solution program:

$$
\frac{t(\pi^*)}{P(\pi^*)} \leq t(\mu) \leq \frac{2t(\pi^*)}{P(\pi^*)}
$$

(22)

where $t(\mu)$ is the CJS of solving an induction problem from source $\mu$ with sufficient input complexity ($>> H_U(\mu)$), and $t(\cdot)$ is the running time of a program on $U$, since

$$
H_U(\pi^*) = -\log_2 P_U(\pi^*) = -\log_2 P_U(\mu)
$$

(23)

$$
t(\mu) \leq t(\pi^*)2^{H_U(\mu)+1}
$$

(24)

we observe that the asymptotic complexity is

$$
t(\mu) = O(2^{H_U(\mu)})
$$

(25)

for fixed $t(\pi^*)$. Note that $t(\pi^*)$ corresponds to the stochastic extension of Bennett’s logical depth \[2\], which was defined as:

the running time of the minimal program that computes $x$

Recall that the minimal program is essentially unique \[3\].

**Definition 1.** Stochastic logical depth is the running time of the minimal program that accurately simulates a stochastic source $\mu$.

$$
L_U(\mu) = t(\pi^*)
$$

(26)
which entails that

\[ t(\mu) \leq L_U(\mu).2^{H_U(\mu)+1} \]  

(27)

CJS is related to the expectation of the simulation time of the universal mixture.

\[ \frac{1}{2} t(\mu) \leq \sum_{U(\pi) = x(01)^*} t(\pi).2^{-|\pi|} = E_{P_U}[[t(\pi) \mid U(\pi) = x(01)^*]] \]  

(28)

where \( x \) is the input data to sequence induction, without loss of generality.

Proof. Observe that Levin search applies a watchdog policy that approximates a search which allocates running time in proportion to a program’s a priori probability.

6.2 A Graphical Analysis of Intelligent Computation

Traditional models of computation are supposed to resemble low-level mechanical architectures like the Turing Machine, or a mathematical construction like Lambda Calculus. Cellular Automata resemble some biological processes but we will assume an asynchronous, irregular lattice which is more general.

Let us introduce an asynchronous space-time graph-automata model of computation, which is unlike graph automata that assume Newtonian time.

Definition 2. Let the computation be represented by a graph \( G = (V, E) \) where vertices are labeled from an alphabet \( D \) via a labeling function \( l: V \rightarrow D \) (space), corresponding to memory states, and edges correspond to the arrow of time. \( I \subset V \) and \( O \subset V \) are input and output vertices interacting with the rest of the world. There is a path from \( I \) to every other vertex. Transformation rules may be specified via Cellular Automata like rules where we have a set of transformations \( T = \{t_i\} \) that transform a connected sub-graph to a new sub-graph. A rule is of the form \( t_i(g_s) = g'_s \) where \( g_s \subset G \) and \( g'_s \subset G \).

Definition 3. Let the volume of computation be defined as \( V_U(\pi) \) which measures the space-time volume of computation of \( \pi \) on \( U \) (in physical units).

Volume of computation may be taken as \( |V(G)| \) for the graphical model of computation above, and it is a valid for any measurable, metric space-time and is therefore compatible with General Relativity and any model of computation. For instance, space of a Turing Machine is the Instantaneous Description (ID) of it, and its time corresponds to \( Z^+ \). For instance, a Turing Machine derivation that has an ID of length \( 2i \) at time \( i \) and takes \( t \) steps to complete would have a volume of \( t.(t+1) \).

Definition 4. Let the energy of computation be defined as \( E_U(\pi) \) which measures the total energy required by computation of \( \pi \) on \( U \) (in physical units).
Definition 5. Let the space of computation be defined as $S_U(\pi)$ which measures the maximum volume of a synchronous slice of the space-time of computation $\pi$ on $U$ (in physical units).

Definition 6. Self-contained physical computation is the physical correspondence of self-delimiting computation whereas all the physical resources required by computation (space-time, and energy) are required to be contained within the volume of computation.

Therefore, we do not allow a self-contained physical computation to send queries over the internet, or use a power cord, for instance.

Using these new more general concepts, we measure the conceptual jump size in space-time volume rather than time. Algorithmic complexity remains the same, as length of a program readily generalizes to volume of program at the input boundary of computation.

Let us generalize logical depth to the logical volume of a bit string $x$:

$$L_V^U(x) = V(\arg\min_{\pi} \{ U(\pi) = x(01^*) \})$$

(29)

Let us also generalize stochastic logical depth to stochastic logical volume:

$$L_V^U(\mu) = V(\pi^*)$$

(30)

which entails that

$$V(\mu) \leq L_V^U(\mu) \cdot 2^{H_U(\mu) + 1}$$

(31)

where left-hand-side corresponds to space-time generalization of CJS.

We now show an interesting relation which is the case for self-contained computations.

Lemma 1.

$$E(\pi^*) = O(V_U(\pi^*)) \quad E(\mu) = O(L_V^U(\mu))$$

One must spend energy to conserve a memory state, or to perform a basic operation. If all basic memory operations (refresh, load, store) and basic operators spend constant energy for a fixed space-time volume, we may assume the relation.

Let us also assume that the space complexity of a program is proportional to how much mass is required. Then, the energy from the resting mass of an optimal computation may be taken into account:

Lemma 2.

$$E(\pi^*) = O(V_U(\pi^*)) + O(S_U(\pi^*)c^2) \quad E(\mu) = O(L_V^U(\mu)) + O(S_U(\mu)c^2)$$

where $c$ is the speed of light. Which brings us to an energy based statement of conceptual jump size, which we term conceptual jump energy:

Theorem 1. $E_U(\mu) = [O(L_V^U(\mu)) + O(S_U(\mu)c^2)] \cdot 2^{H_U(\mu) + 1}$
As a simple consequence of the above lemma, we show an upper bound on minimum energy required, that is related to the volume, and space linearly, and algorithmic complexity of a stochastic source exponentially, for optimal induction.

Landauer’s limit is a thermodynamic limit of $kT\ln 2$ where $k$ is the Boltzmann constant and $T$ is the temperature [7]. The total number of bit-wise operations that a quantum computer can evolve is $2E/h$ operations where $E$ is average energy, and thus the physical limit to energy efficiency of computation is about $3.32 \times 10^{33}$ operations/J [14]. Note that the Margolus-Levitin limit supports our relation of the volume of computation (called $E.t$ “action volume” in their paper) with total energy. Lloyd [11] assumes all the mass may be converted to energy and calculates the maximum computation capacity of a 1 kilogram “black-hole computer”, performing $10^{51}$ operations over $10^{31}$ bits. According to an earlier paper, the whole universe may not have performed more than $10^{120}$ operations over $10^{90}$ bits [12].

**Theorem 2.** The universal limits to computation imply that $H(\mu) \leq 397.6$ for any $\mu$.

**Proof.** Even if $L_U^V(\mu) = 1$

\[
V(\mu) \leq L_U^V(\mu) \cdot 2^{H_U(\mu)+1} \leq 10^{120} \tag{32}
\]

\[
\log_2(2^{H_U(\mu)+1}) \leq 3.321 \times 120 \tag{33}
\]

\[
H(\mu) + 1 \leq 398.6 \tag{34}
\]

Therefore, if $\mu$ has a greater algorithmic complexity than about 400 bits, it would have been impossible to discover it without any a priori information. This limit shows the remarkable importance of incremental learning as both Solomonoff [25] and Schmidhuber [16] have emphasized, which is part of ongoing research. Optimizing energy efficiency of computation would also be an obviously useful goal for a self-improving AI.

**References**

1. Bennett, C.H.: How to define complexity in physics, and why. In: Zurek, W.H. (ed.) Complexity, Entropy, and the Physics of Information. vol. VIII, pp. 137–148 (1980)
2. Bennett, C.H.: Logical depth and physical complexity. In: Herkin, R. (ed.) The Universal Turing Machine: A Half-Century Survey, pp. 227–257. Oxford University Press, Oxford (1988), [citeseer.ist.psu.edu/bennet88logical.html](citeseer.ist.psu.edu/bennet88logical.html)
3. Chaitin, G.J.: Algorithmic Information Theory. Cambridge University Press (2004)
4. Deutsch, D.: Quantum theory, the church-turing principle and the universal quantum computer. Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences 400(1818), 97–117 (1985), [http://rspa.royalsocietypublishing.org/content/400/1818/97.abstract](http://rspa.royalsocietypublishing.org/content/400/1818/97.abstract)
5. Frampton, P.H., Hsu, S.D.H., Kephart, T.W., Reeb, D.: What is the entropy of the universe? Classical and Quantum Gravity 26(14), 145005 (Jul 2009)
6. Hutter, M.: Universal algorithmic intelligence: A mathematical top-down approach. In: Goertzel, B., Pennachin, C. (eds.) Artificial General Intelligence, pp. 227–290. Cognitive Technologies, Springer, Berlin (2007), http://www.hutter1.net/ai/aixigentle.htm

7. Landauer, R.: Irreversibility and heat generation in the computing process. IBM J. Res. Dev. 5(3), 183–191 (Jul 1961), http://dx.doi.org/10.1147/rd.53.0183

8. Legg, S., Hutter, M.: Universal intelligence: A definition of machine intelligence. Minds Mach. 17(4), 391–444 (Dec 2007), http://dx.doi.org/10.1007/s11023-007-9079-x

9. Legg, S., Veness, J.: An approximation of the universal intelligence measure. In: Dowe, D. (ed.) Algorithmic Probability and Friends. Bayesian Prediction and Artificial Intelligence, Lecture Notes in Computer Science, vol. 7070, pp. 236–249. Springer Berlin Heidelberg (2013), http://dx.doi.org/10.1007/978-3-642-44958-1_18

10. Levin, L.A.: Some theorems on the algorithmic approach to probability theory and information theory. CoRR abs/1009.5894 (2010)

11. Lloyd, S.: Ultimate physical limits to computation. Nature406, 1047–1054 (Aug 2000)

12. Lloyd, S.: Computational Capacity of the Universe. Physical Review Letters 88(23), 237901 (Jun 2002)

13. Lloyd, S.: Universal quantum simulators. Science 273(5278), 1073–1078 (1996), http://www.sciencemag.org/content/273/5278/1073.abstract

14. Margolus, N., Levitin, L.B.: The maximum speed of dynamical evolution. Physica D Nonlinear Phenomena 120, 188–195 (Sep 1998)

15. ¨Ozkural, E.: Worldviews, Science and Us: Philosophy and Complexity, chap. A compromise between reductionism and non-reductionism. World Scientific Books (2007)

16. Schmidhuber, J.: Optimal ordered problem solver. Machine Learning 54, 211–256 (2004)

17. Solomonoff, R.J.: A formal theory of inductive inference, part i. Information and Control 7(1), 1–22 (March 1964)

18. Solomonoff, R.J.: A formal theory of inductive inference, part ii. Information and Control 7(2), 224–254 (June 1964)

19. Solomonoff, R.J.: Inductive inference research status spring 1967. Tech. Rep. RTB 154, Rockford Research, Inc. (1967)

20. Solomonoff, R.J.: Complexity-based induction systems: Comparisons and convergence theorems. IEEE Trans. on Information Theory IT-24(4), 422–432 (July 1978)

21. Solomonoff, R.J.: The discovery of algorithmic probability. Journal of Computer and System Sciences 55(1), 73–88 (August 1997)

22. Solomonoff, R.J.: Progress in incremental machine learning. Tech. Rep. IDSIA-16-03, IDSIA, Lugano, Switzerland (2003)

23. Solomonoff, R.J.: Three kinds of probabilistic induction: Universal distributions and convergence theorems. The Computer Journal 51(5), 566–570 (2008), christopher Stewart Wallace (1933-2004) memorial special issue

24. Solomonoff, R.J.: Algorithmic probability: Theory and applications. In: Dehmer, M., Emmert-Streib, F. (eds.) Information Theory and Statistical Learning, Springer Science+Business Media, pp. 1–23. N.Y. (2009)

25. Solomonoff, R.J.: Algorithmic probability, heuristic programming and agi. In: Third Conference on Artificial General Intelligence. pp. 251–157 (2010)
26. Wood, I., Sunehag, P., Hutter, M.: (non-)equivalence of universal priors. In: Dowe, D. (ed.) Algorithmic Probability and Friends. Bayesian Prediction and Artificial Intelligence, Lecture Notes in Computer Science, vol. 7070, pp. 417–425. Springer Berlin Heidelberg (2013), \url{http://dx.doi.org/10.1007/978-3-642-44958-1_33}

27. Zurek, W.H.: Algorithmic randomness, physical entropy, measurements, and the demon of choice. In: Hey, A.J.G. (ed.) Feynman and computation: exploring the limits of computers. Perseus Books (1998)