AIT Blog 2022

Samuel Epstein
samepst@jptheorygroup.org

February 2, 2023

This article contains 19 posts in 2022 to the "AIT Blog". The contents include algorithmic physics, machine learning, derandomization, clusters, left-total machines, and new proofs.

Contents

1 September 28th, 2022: AIT and Machine Learning

This is a math blog focusing on Algorithmic Information Theory. The main focus will be on strings $x \in \{0,1\}^*$ that have low mutual information with the halting sequence H, with $\mathbf{I}(x;H) = \mathbf{K}(x) - \mathbf{K}(x|H)$, being low. $\mathbf{K}$ is the prefix free Kolmogorov complexity. There are many properties that can be proven about elementary objects that have low $\mathbf{I}(x;H)$. We say an object is (non)exotic if it has (low)high mutual information with the halting sequence. Exotic objects cannot be found in the physical world. Furthermore, $\mathbf{I}(x;H)$ enjoys conservation laws, in that deterministic and random processing cannot increase information.

- For partial computable f, $\mathbf{I}(f(a):H) < \mathbf{I}(a;H) + \mathbf{K}(f) + O(1)$.
- For program q that computes probability p over $\mathbb{N}$, $\mathbf{E}_{a\sim p}[2^{\mathbf{I}(\langle q,a\rangle;H)}] < O(1)2^{\mathbf{I}(q;H)}$.

This entry deals with the relationship between Algorithithmic Information Theory and Machine Learning. Classification is the task of learning a binary function c from $\mathbb{N}$ to bits $\{0,1\}$. The learner is given a sample consisting of pairs (x,b) for string x and bit b and outputs a binary classifier $h : \mathbb{N} \to \{0,1\}$ that should match c as much as possible. Occam's razor says that "the simplest explanation is usually the best one." Simple hypothesis are resilient against overfitting to the sample data. With certain probabilistic assumptions, learning algorithms that produce hypotheses of low Kolmogorov complexity are likely to correctly predict the target function [BEHW89]. The following theorem shows that the samples can be compressed to their count.

Theorem 1 *Given a set of samples $\{(x_i,b_i)\}$, $i = 1,\ldots,n$, there is a total function $f : \mathbb{N} \to \{0,1\}$ such that $f(x_i) = b_i$ for $i = 1,\ldots,n$ and $\mathbf{K}(f) <^{\log} n + \mathbf{I}(\{(x_i,b_i)\};H)$.*

However, usually the samples can be modeled as coming from a probabilistic model. The target concept is modeled by a random variable X with distribution p over ordered lists of natural numbers. The random variable Y models the labels, and has a distribution over lists of bits, where the distribution of $X \times Y$ is $p(x,y)$ with conditional probability requirement $p(y|x) = \prod_{i=1..|x|} p(y_i|x_i)$. Each such (x_i,y_i) is a labeled sample. A binary classifier f is consistent with labelled samples (x,y), if for all i, $f(x_i) = y_i$. Let $\Gamma(x,y)$ be the minimum Kolmogorov complexity of a classifier consistent with (x,y). $Entropy(Y|X)$ is the conditional entropy of Y given X.

Theorem 2 $Entropy(Y|X) \leq \mathbf{E}[\Gamma(X,Y)] <^{\log} Entropy(Y|X) + \mathbf{K}(p)$.

Another area of machine learning is regression, in which one is give a set of pairs $\{(x_i,y_i)\}$, $i = 1\ldots n$, and the goal is to find a function f, such that $f(x_i) = y_i$. Usually each x_i and y_i represents a point in Euclidean space, but for our purpose they are natural numbers. As in classification, the goal is to use Occam's razor to find the simpliest function, to prevent overfitting to the random noise inherent in the sample data. The following theorem provides bounds on the simplest total computable function completely consistent with the data.

Theorem 3 *For $\{(x_i,y_i)\}$, $i = 1,\ldots,n$, there exists $f : \mathbb{N} \to \mathbb{N}$ with $f(x_i) = y_i$ for $i \in \{1,\ldots,n\}$ and $\mathbf{K}(f) <^{\log} \sum_{i=1}^{n} \mathbf{K}(y_i|x_i) + \mathbf{I}(\{(x_i,y_i)\};H)$.*

This theorem can be proved using Theorem 8 in [Eps22d]. However, this theorem is over computable probability measures, whereas the lower semi-computable $\mathbf{m}$ is needed. By using so-called left-total machines, $\mathbf{m}$ can be converted into a computable measure. In fact one of the benefits of using left-total machines and having $\mathbf{I}(,;H)$ as an error term, is that semi-computable functions can be converted into computable ones.

2 September 30th, 2022: Outliers from Algorithms and Dynamics

2.1 Randomness Deficiencies

Outliers or anomalies are an ubiquitous phenomenon. In this entry, we trace the occurrence of anomalies in algorithmic sampling methods, measurements procured from the real world, and dynamics over the Cantor space and computable metric spaces. In algorithmic information theory, outliers are measured by the randomness deficiency function. The deficiency of randomness $\mathbf{d}$ of a string $x \in \{0,1\}^*$ with respect a computable probability measure p over strings is

$$\mathbf{d}(x|p) = \lfloor -\log p(x) \rfloor - \mathbf{K}(x|p).$$

The condition prefix complexity is $\mathbf{K}$. Randomness deficiency can also be defined over infinite sequences $\alpha \in \{0,1\}^\infty$ and computable probability measures P over $\{0,1\}^\infty$.

$$\mathbf{D}(\alpha|P) = \sup_n -\log P(\alpha[0..n]) - \mathbf{K}(\alpha[0..n]|P).$$

For more information about randomness deficiency, we refer readers to [GÍ3].

2.2 Sampling Methods

A discrete sampling method A is a probabilistic function that maps an integer N with probability 1 to a set containing N different strings.

Theorem 4 *Let p be a computable probability measure over $\mathbb{N}$. Let A be a sampling method. There exists $c \in \mathbb{N}$ such that for all n and k:*

$$\Pr(\max_{a \in A(2^n)} \mathbf{d}(a|p) > n - k - c\log n) \geq 1 - 2e^{-2^k}.$$

There is also emergent outliers in continuous sampling methods. A continuous sampling method C is a probabilistic function that maps, with probability 1, an integer N to an infinite encoding of N different inftinite sequences.

Theorem 5 *Let P be a computable probability measure over $\{0,1\}^\infty$. Let C be a continuous sampling method. There exists $c \in \mathbb{N}$ such that for all n and k,*

$$\Pr(\max_{\alpha \in C(2^n)} \mathbf{D}(\alpha|P) > n - k - c\log n) \geq 1 - 2.5e^{-2^k}.$$

2.3 Probabilistic Algorithms

Outliers can be found in more general constructs than sampling methods. Given a computable measure μ over $\{0,1\}^\infty$ Any probabilistic algorithm that outputs an infinite sequence (with no individual sequence with positive probability) is guaranteed to produce ever larger μ-outlier sequences with diminishing measure. More formally, the following theorem encodes this fact.

Theorem 6 *For computable measures μ and non-atomic λ over $\{0,1\}^\infty$ and $n \in \mathbb{N}$, $\lambda\{\alpha : \mathbf{D}(\alpha|\mu) > n\} > 2^{-n-\mathbf{K}(n,\mu,\lambda)-O(1)}$.*

2.4 Outliers in the Physical World

In the prvious sections, it is proven that algorithmic sampling methods have to produce anomalies. However some sampling methods are too complex to be considered algorithmic. One example is your local weather forecast. Using the Independence Postulate, which is a finitary Church-Turing thesis, this open issue is addressed. Outliers must occur in the physical world.

The Independence Postulate (**IP**), [Lev84, Lev13], is an unprovable inequality on the information content shared between two sequences, postulating that certain infinite and finite sequences cannot be found in nature, a.k.a. have high "physical addresses". In this paper we show that **IP** explains why outliers are found in the physical world. The approach in **IP** is different from that of the previous sections. While the latter shows that computable consructs produce outliers with high probability, the former states that individual sequences without outliers have high addresses, i.e. are hard to find nature.

The information between two strings x, y is $\mathbf{I}(x:y) = \mathbf{K}(x) + \mathbf{K}(y) - \mathbf{K}(x,y)$. The algorithmic probability is $\mathbf{m}$. The information between two infinite sequences $\alpha, \beta \in \{0,1\}^\infty$ is defined to be [Lev74]

$$\mathbf{I}(\alpha:\beta) = \log \sum_{x,y \in \{0,1\}^*} \mathbf{m}(x|\alpha)\mathbf{m}(y|\beta)2^{\mathbf{I}(x:y)}.$$

The Independence Postulate states [Lev13]:

IP: *Let α be a sequence defined with an n-bit mathematical statement, and a sequence β can be located in the physical world with a k-bit instruction set. Then $\mathbf{I}(\alpha:\beta) < k + n + c$ for some small absolute constant c.*

Let $\tau \in \mathbb{N}^\mathbb{N}$ represent a series of observations. In reality, observed information is finite. But observations can be considered to be potentially infinite, and represented by never-ending sequences. τ is assumed to have an infinite number of unique observations. $\tau(n)$ is the first 2^n unique numbers of τ. For a probability p over $\mathbb{N}$, let $s_{\tau,p} = \sup_n \left(n - 3\mathbf{K}(n) - \max_{a \in \tau(n)} \mathbf{d}(a|p)\right)$. It is a score of the level of outliers in τ. If $s_{\tau,p}$ is large then τ can be considered to have low level of outliers. if $s_{\tau,p}$ is infinite, then τ has bounded level of outliers.

Theorem 7 $s_{\tau,p} <^{\log} \mathbf{I}(\langle\tau\rangle : \mathcal{H}) + O(\log \mathbf{K}(p))$.

Let k be a physical address of τ. The halting sequence $\mathcal{H}$ can be described by a small mathematical statement. By Theorem 7 and IP,

$$s_{\tau,p} <^{\log} \mathbf{I}(\langle\tau\rangle : \mathcal{H}) + O(\log \mathbf{K}(p)) <^{\log} k + c + O(\log \mathbf{K}(p)).$$

It's hard to find observations with small anomalies and impossible to find observations with no anomalies.

2.5 Dynamics in the Cantor Space

Sampling algorithms and probabilistic algorithms in Section 2.3 have outliers that are guaranteed to appear. In the next two sections, we show that outliers are emergent in dynamics. This appears in the natural setting of dynamics over the Cantor space as well as the more general setting of dynamics in computable metric spaces.

We define a metric g on $\{0,1\}^\infty$ with $g(\alpha,\beta) = 1/2^k$, where k is the first place where α and β disagree. Let $\mathfrak{F}$ be the topology induced by g on $\{0,1\}^\infty$; $\mathcal{B}(\mathfrak{F})$ be the Borel σ-algebra on $\{0,1\}^\infty$;

λ and μ be computable measures over $\{0,1\}^\infty$ and λ be nonatomic; and $(\{0,1\}^\infty, \mathcal{B}(\mathfrak{F}), \lambda)$ be a measure space and $T : \{0,1\}^\infty \to \{0,1\}^\infty$ be an ergodic measure preserving transformation. By the Birkoff theorem,

Theorem 8 *Starting λ-almost everywhere, $\overset{*}{>} \mathbf{m}(n, \mu, \lambda)2^{-n}$ states α visited by iterations of T have $\mathbf{D}(\alpha|\mu) > n$.*

A computable dynamical system (λ, δ) consists of a computable starting state probability λ over $\{0,1\}^\infty$ and a computable transition function $\delta : \{0,1\}^\infty \to \{0,1\}^\infty$. We assume that the dynamical system is non-degenerate, in that for λ-a.e. starting states α, an infinite number of states is visited using δ.

Theorem 9 *There exists $d \in \mathbb{N}$, where for computable probability μ over $\{0,1\}^\infty$ and computable dynamics (λ, δ), for λ-a.e. starting states $\alpha \in \{0,1\}^\infty$, there exists $s_\alpha \in \mathbb{N}$, where among the first 2^m states visited, for any $n < m$, there are at least 2^n states β with $\mathbf{D}(\beta|\mu) > m - n - d \log m - s_\alpha$. Furthermore, for the smallest such s_α, $\mathbf{E}_{\alpha\sim\lambda}[s_\alpha - O(\log s_\alpha)] < \mathbf{K}(\lambda, \delta) + \mathbf{K}(\mu)$.*

2.6 Dynamics in Computable Metric Spaces

Results on outliers can be applied to more general spaces than the Cantor space. In this section, we show how outliers are emergent in dynamics over computable metric spaces.

Definition 1 (Computable Metric Space) *A computable metric space is a triple $\mathfrak{X} = (X, d, \mathfrak{I})$, where*

- *(X, d) is a separable complete metric space.*
- *$\mathfrak{I} = \{s_i : i \in \mathbb{N}\}$ is a countable dense subset of X.*
- *The real numbers $d(s_i, s_j)$ are all computable, uniformly in $\langle i, j\rangle$.*

Definition 2 (Ideal Balls) *Let $B(x, r)$ be the metric ball $\{y \in X, d(x, y) < r\}$. The numbered sets $\mathfrak{I}$, and $\mathbb{Q}_{>0}$ induced the numbered set of ideal balls $\mathfrak{B} = \{B(s_i, q_j) : s_i \in \mathfrak{I}, q_j \in \mathbb{Q}_{>0}\}$. We write $B_{\langle i,j\rangle}$ for $B(s_i, q_j)$.*

Definition 3 (Computable Measures) *A measure μ over $\mathfrak{X}$ is computable if $\mu(B_{i_1} \cup \cdots \cup B_{i_k})$ is lower semi-computanle uniformly in $\langle i_1, \ldots, i_k\rangle$.*

Definition 4 (Computable Probability Space) *A computable probability space is a pair $(\mathfrak{X}, \mu)$ where $\mathfrak{X}$ is a computable metric space and μ a computable Borel probability measure space on X.*

Definition 5 (Uniform Tests of Randomness) *A uniform test t over computable probability space $(\mathfrak{X}, \mu)$ is a lower semi-computable function over $\mathfrak{X}$ and a description of the probability measure μ such that $\int_X t(\mu, x)dx \leq 1$. It is beyond the scope of this blog to describe how probability measures can be sent as input to a function. We refer readers to [HR09, GÍ3]. In general, it is a fast Cauchy sequence converging to a point in a space where every point is a probability measure. There exists a universal uniform test of randomness $\mathbf{t}$, where for every uniform test of randomness t, there is a constant $c \in \mathbb{N}$ where $c\mathbf{t} > t$. We denote the universal uniform test by $\mathbf{t}_\mu(x)$. The deficiency of randomness is $\mathbf{d}_\mu(x) = \log t_\mu(x)$.*

Definition 6 (Transformation Group) *Dynamics are represented by one-dimensional transformation groups. For computable metric space $\mathfrak{X}$, a topological group G is defined such that each element is a homeomorphism of $\mathfrak{X}$ onto itself:*

$$f(g;x) = g(x) = x' \in X; g \in G, x \in X.$$

The symbol G will be called a topological transformation group if for every pair of elements $g1, g2$ of G, and every $x \in X$, $g1(g2(x)) = (g1g2)(x)$ and if

$$x' = g(x) = f(g;x)$$

is continuous simultaneously in $x \in X$ and $g \in G$.

Theorem 10 (Dynamics over Computable Measure Spaces) *Let L be the Lebesgue measure over $\mathbb{R}$. For one dimensional computable topological transformation group G^t acting on computable probability space $(\mathfrak{X}, \mu)$, for all $\alpha \in X$, $L\{t \in [0,1] : \mathbf{d}_\mu(G^t\alpha) > n\} > 2^{-n-O(\log n)-c(\alpha,\mathfrak{X},\mu)}$. The term $c(\alpha, \mathfrak{X}, \mu)$ is a constant dependent solely on α and $(\mathfrak{X}, \mu)$.*

3 October 5th, 2022: Derandomization, Lovasz Local Lemma, and Classical Channels

Over time, this blog will present unpublished material and then public material, with the eventual goal of writing a manuscript covering everything presented. This blog will also focus on the connection between AIT and Quantum Mechanics.

This blog entry deals with *derandomization*, [Eps22d, Eps22a], first with a general discussion and then with some new examples. This blog entry will showcase a new application of derandomization to compressing codebooks. In classical information theory, parties that communicate over channels share auxiliary information such as codebooks. Derandomization can be used to show the more bits used to describe the codebook results in a greater rate of communication, similar to the Kolmogorov structure function.

A problem is a collection of instances and the goal is to determine whether the instance satisfies a certain property, i.e. has a solution. An example is SAT, where instances are formulas and the goal is to find an assignment of the variables which satisfies it. Through the recently introduced method of *derandomization*, this approach can be aligned with Algorithmic Information Theory in a new way: bounds on the Kolmogorov complexity of the simplest solutions can be proven. The notion of a "solution" is intentionally vague, for example in MAXSAT, a solution could entail any assignment that satisfies 6/7 the optimal number of possible satisfiable clauses. Derandomization represents a good research topic for students because its framework is flexible and it can be applied to many fields of study.

The procedure for derandomization is as follows. The first step is to prove solutions to certain (or all!) instances of problems occur with probability at least p, with respect to a simple probability measure P over the solution candidate space. This is done by employing standard techniques (to be discussed later). Then, by applying Lemma 1 and Theorem 12, bounds on the Kolmogorov complexity of the simpliest solution can be proven. More specifically there exists some solution encoded into $x \in \{0,1\}^*$, with

$$\mathbf{K}(x) <^{\log} \mathbf{K}(P) - \log p + \mathbf{I}(\langle \text{description of the instance} \rangle ; \mathcal{H}).$$

$\mathbf{K}$ is the prefix-free Kolmogorov complexity function. $\mathbf{I}(x;\mathcal{H}) = \mathbf{K}(x) - \mathbf{K}(x|\mathcal{H})$ is the asymmetric mutual information term between x and the halting sequence $\mathcal{H}$. The instance itself can be incredibly complex (for example a formula with exponential number of clauses to variables), but for all non-exotic instances, the term $\mathbf{I}(\langle \text{description of the instance} \rangle ; \mathcal{H})$ will be negligible.

Lovasz Local Lemma. One of the main tools to lower bounding the probability of a solution to an instance is the Lovasz Local Lemma. The Lovasz Local Lemma is traditionally used to show that some property is true for a random object with positive probability, and thus objects with this property exists. In fact, there is an exact lower bound on the probability of the property occurring and thus when LLL is applied to instances of problems, it produces a lower bound on the probability that a solution exists. Examples of this can be seen in [Eps22a]. Thus, going forward, one can look to where LLL is applied in the literature as well as apply LLL to new cases where LLL provides no traditional benefit, such as when the desired object property is trivially present.

3.1 Compressing Codebooks

There are deep connections between classical information theory and algorithmic information theory, with many theorems of the former appearing in an algorithmic form in the latter. In this section we revisit this connection. In particular we prove properties about the compression size of shared codebooks. A standard setup in information theory is two parties Alice and Bob who want to communicate over a noisy channel and share a codebook over a noiseless channel. However one might ask is how many bits did it take to communicate the codebook? By using derandomization, the tradeoff between codebook complexity and communication capacity can be proven.

Definition 7 (Discrete Memoryless Channel) *The input and output alphabets $\mathcal{X}$ and $\mathcal{Y}$ are finite. The channel $(\mathcal{X}, p(y|x), \mathcal{Y})$ is represented by a conditional probability distribution $p(y|x)$. To send multiple symbols, we have $p(y^n|x^n) = \prod_{i=1}^n p(y_i|x_i)$. The capacity of channel with respect to a distribution Q over $\mathcal{X}$ is*

$$C_Q = I(X:Y) \text{ where random variables } (X,Y) \text{ are distributed according to } Q(x)p(y|x).$$

The term I is the mutual information between random variables. The capacity of a channel is

$$C = \max_{Q(x)} C_Q.$$

Definition 8 (Codebook) *A (M,n) codebook for channel $(\mathcal{X}, p(y|x), \mathcal{Y})$ contains the following:*

1. *An encoder* $\mathrm{Enc}_n : \{1,\dots,M\} \to \mathcal{X}^n$.
2. *A decoder* $\mathrm{Dec}_n : \mathcal{Y}^n \to \{1,\dots,M\}$.

The rate of the codebook is $R = \frac{\log M}{n}$. The error rate of the codebook with respect to a fixed channel is the probability that, given the uniform distribution over $\{1,\dots,M\}$ for the sending symbols, the receiver decodes a symbol different from the encoded one.

Imagine the following setup: there is a sender Alice and a receiver Bob that communicate through a noisy memoryless discrete channel and Alice can send a codebook to Bob once on a side noiseless channel. Bob has oracle acess to the channel function $p(y|x)$ but Alice does not. Given a computable distribution Q over the input alphabet, and assuming the channel is non-exotic, Alice can hypothetically send $\sim\mathbf{K}(Q)$ bits plus some encoded parameters describing a codebook to Bob on the side channel. Then Alice and Bob can communicate with any rate R less than the capacity C_Q over the noisy channel.

Theorem 11 *For channel $\mathfrak{C} = (\mathcal{X}, p(y|x), \mathcal{Y})$ and every computable distribution Q over $\mathcal{X}$, for every rate $R < C_Q$, there is a $(2^{nR}, n)$ codebook $(\mathrm{Enc}_n, \mathrm{Dec}_n)$ with rate R and error rate $o(1)$ and*

$$\mathbf{K}(\mathrm{Enc}_n) <^{\log} \mathbf{K}(n,R,Q) + \mathbf{I}((n,R,Q,\mathfrak{C});\mathcal{H}),$$
$$\mathbf{K}(\mathrm{Dec}_n|\mathfrak{C}) <^{\log} \mathbf{K}(n,R,Q) + \mathbf{I}((n,R,Q,\mathfrak{C});\mathcal{H}).$$

The proof comes from copying down pages of the original proof of the capacity of memoryless discrete channels, and noting which construct is computing from which. Future work entails looking at auxiliary information used in other types of channels, such as ones with feedback. Future work also involves proving properties of the following function, which is defined with respect to a channel, with for $0 \leq n \leq \mathbf{K}(\mathfrak{C})$,

$$\Gamma(n) = \max\{C_Q : \mathbf{K}(Q) \leq n\}.$$

3.2 Connected Subgraphs

This was an example of derandomization that I was interested in, and thought it might be of some independent interest. Conservation of information has been proven in all standard definitions of information. For example, in the symmetric definition, $\mathbf{I}(x : y) = \mathbf{K}(x) + \mathbf{K}(y) - \mathbf{K}(x, y)$, it has been proven that $\mathbf{I}(f(x) : y) <^+ \mathbf{I}(x : y)$, [Lev84]. The following lemma shows that conservation holds in the asymmetric form of information, as long as the halting sequence $\mathcal{H}$ is used.

Lemma 1 ([Eps22d]) *For partial computable $f : \{0,1\}^* \to \{0,1\}^*$, for all $a \in \{0,1\}^*$, $\mathbf{I}(f(a); \mathcal{H}) <^+ \mathbf{I}(a; \mathcal{H}) + \mathbf{K}(f)$.*

The following theorem has been called the Sets Have Simple Members theorem. For a set $D \subseteq \{0,1\}^*$, $\mathbf{m}(D) = \sum_{x \in D} \mathbf{m}(x)$, where $\mathbf{m}$ is the algorithmic probability. For the purposes of derandomization, D encodes all solutions to the instance. If $P(D)$ is large for some simple probability measure P, then $-\log \mathbf{m}(D)$ will be small. The error term is $\mathbf{I}(D; \mathcal{H})$, but using Lemma 1, this can be bounded by $\mathbf{I}(\langle \text{description of the instance} \rangle; \mathcal{H})$. This is because given a description of an instance (such as a MaxSat formula) one can define a simple function f that outputs all encoded solutions D (such as assignments that are 6/7 optimal).

Theorem 12 ([Lev16, Eps19c])
For finite $D \subset \{0,1\}^*$, $-\log \max_{x \in D} \mathbf{m}(x) <^{\log} -\log \mathbf{m}(D) + \mathbf{I}(D; \mathcal{H})$.

Theorem 13 *Let $G = (V, E)$ be a connected graph with n vertices and m edges. Then there is a set $S \subseteq E$ with $|S| \leq (m+n)/2$ such that when E restricted to S, a connected subgraph of G is created, and $\mathbf{K}(S) <^{\log} \mathbf{K}(m) + 2n + \mathbf{I}(G; \mathcal{H})$.*

Proof. We order the edges E from 1 to m. Let $T \subseteq E$ be a spanning tree of G of size $n-1$. T can be encoded as a set of natural numbers, each number representing an edge of E. We use strings $x \in \{0,1\}^m$ to represent subgraphs, where $x[i] = 1$ if edge i is included in the graph. We define the probability P over strings $x \in \{0,1\}^m$ of length m, with $P(x) = \prod_{i=1}^m (3/4)^{1-x[i]}(1/4)^{x[i]}$. The complexity of P is $\mathbf{K}(P) <^+ \mathbf{K}(m)$. Let $D \subset \{0,1\}^m$, such that if $x \in D$, then $x[i] = 1$ for all $i \in T$ and $\sum_{i \in \{1,\ldots,m\}-T} x[i] \leq (m-(n-1))/2$. Thus by the Markov inequality, $P(D) \geq (1/4)^{n-1} \times (1/2)$. The set D can be produced from a description of the graph, with $\mathbf{K}(D|G) = O(1)$. This is because T is simple relative to G. Thus, using Lemma 1 and Theorem 12, there is a set of $\leq (m+n)/2$ edges $x \in D$, with

$$\mathbf{K}(x) <^{\log} \mathbf{K}(P) - \log P(D) + \mathbf{I}(D; \mathcal{H}) <^{\log} \mathbf{K}(m) + 2n + \mathbf{I}(G; \mathcal{H}).$$

□

4 October 9th, 2022: A New Proof to the Sets Have Simple Members Theorem

This blog entry presents a new proof to the Sets Have Simple Members Theorem [Lev16, Eps19c]. The first game theoretic proof, and currently the most accessible proof in the literature, can be found in [She12]. The proof in this entry follows analogously to the proof in [Lev16] except left-total machines are not used. We also provide a proof that non-stochastic elements have high mutual information with the halting sequence, a well known result in the literature. This proof also does not rely on left-total machines. However in a later blog, I'll show the utility of left-total machines. The quest for more proofs is in part motivated by the question of whether there is a resource bounded version of the Sets Have Simple Members Theorem.

Theorem. *For finite* $D \subset \{0,1\}^*$, $\min_{x \in D} \mathbf{K}(x) <^{\log} -\log \mathbf{m}(D) + \mathbf{I}(D; \mathcal{H})$.

The term $\mathbf{m}(x) = 2^{-\mathbf{K}(x)}$ is the algorithmic probability, where $\mathbf{K}$ is the prefix complexity. $\mathbf{m}(D) = \sum_{x \in D} \mathbf{m}(x)$. $\mathbf{I}(x|\mathcal{H}) = \mathbf{K}(x) - \mathbf{K}(x|\mathcal{H})$ is the mutual information of x with the halting sequence $\mathcal{H}$.

4.1 Motivation

We revisit the original motivation of the theorem. One common goal in computer science is to find the hidden part of the environment, this task has been called Inductive Inference, Extrapolation, Passive Learning, etc. The complete environment can represented as a huge string $x \in \{0,1\}^*$. The known observations restrict it to a set $D \subset \{0,1\}^*$. For example in thermodynamics, the environment x can be seen as a record of every particle's position and velocity in a closed box. An observation of some macro parameters, such as pressure and temperature, restrict the possible environments to a set D of hypotheses consistent with the observation.

One method used to select a hypothesis (i.e. environment) is to leverage an *apriori* distribution over the environment space. This distribution p encodes any knowledge about the environment known before the observation is made. Then selection of the hypothesis is

$$\arg\max_{x \in D} p(x).$$

Note in AIT, for enumerable distributions (i.e. generatable as outputs of randomized algorithms), there is a universal apriori distribution $\mathbf{m}(x)$. This is because $O(1)\mathbf{m} > p$, for all enumerable p. Furthermore, for all $x \in \{0,1\}^*$, $\mathbf{d}(x|\mathbf{m}) = O(1)$, where $\mathbf{d}$ is deficiency of randomness; so there is no lower computable refutation to the statement: "x is generated from $\mathbf{m}$". Thus when the universal prior is used, inductive inference becomes an exercise of Occam's razor:

$$\arg\min_{x \in D} \mathbf{K}(x).$$

However there exists a potential complication. It could be there is a collection $G \subset D$ of hypotheses representing a concept (such as a more detailed description of particles) where its combined apriori measure is greater than that of the simpliest element x, with $\mathbf{m}(x) \ll \mathbf{m}(G)$. Or, making the endeavor more murkier, it could be that G is just the set of all complicated hypothesis and G has greater combined apriori measure than the simpliest element. In this case, which explanation does one choose?

The Sets Have Simple Members Theorem shows that this dilemma is purely a mathematical construction. All the universal apriori measure of an observation D is concentrated on its simpliest member. This is true for all non-exotic set D with low mutual information with the halting sequence. There are no (randomized) algorithmic means of creating D with arbitrarily high $\mathbf{I}(D;\mathcal{H})$.

4.2 New Proof

A probability is *elementary*, if it has finite support and rational values. The deficiency of randomness of x relative to a elementary probability measure Q is $\mathbf{d}(x|Q) = -\log Q(x) - \mathbf{K}(x|Q)$.

Definition 9 (Stochasticisty) *A string x is (α,β)-stochastic if there exists an elementary probability measure Q such that*

$$\mathbf{K}(Q) \leq \alpha \text{ and } \mathbf{d}(x|Q) \leq \beta.$$

Theorem 14 ([Lev16, Eps19c]) *Let P be a lower-semicomputable semimeasure and c be a large constant. Every (α,β)-stochastic set D with $s = \lceil -\log P(D) \rceil$ contains an element x with*

$$\mathbf{K}(x) < s + \alpha + 2\log\beta + \mathbf{K}(s) + 2\log\mathbf{K}(s) + c.$$

The theorem is directly implied by the following proposition.

Proposition 1 *Let P be a lower-semicomputable semimeasure and c be a large constant. If a set D is (α,β)-stochastic relative to an integer $s = \lceil -\log P(D) \rceil$, then D contains an element x with*

$$\mathbf{K}(x) < s + \alpha + \log\beta + \mathbf{K}(\log\beta) + \mathbf{K}(s) + c.$$

Note that if y is (α,β)-stochastic relative to s, then it is $(\alpha,\beta+\mathbf{K}(s))$-stochastic. Hence the proposition implies the theorem.

Lemma 2 *Let P be a discrete mesure and Q be a measure on sets. There exists a set S of size $\lceil \beta/\gamma \rceil$ such that*

$$Q(\{D : P(D) \geq \gamma \text{ and } D \text{ is disjoint from } S\}) \leq \exp(-\beta).$$

Proof. We use the probabilistic method, and show that if we draw $\lceil \beta/\gamma \rceil$ elements according to the distribution P, then the obtained set S satisfies the inequality with positive probability. The probability that a fixed set D with $P(D) \geq \gamma$ is disjoint from S is

$$\leq (1-\gamma)^{\beta/\gamma} \leq \exp(-\beta).$$

Hence the expected Q-measure of such a D is at most $\exp(-\beta)$ and the required set S exists. □

Proof of Proposition 1 for computable P. Let Q be an elementary probability measure with $\mathbf{K}(Q) \leq \alpha$ and $\mathbf{d}(D|Q,s) \leq \beta$. Without loss of generality, we assume that β is large positive power of 2. Fix a search procedure that on input Q, β, and $\gamma = 2^{-s}$ finds a set satisfying the conditions of Lemma 4.

For large β, the set D must intersect the obtained set S. Indeed, consider the Q-test $g(X|Q,s)$ that is equal to $\exp(\beta)$ if X is disjoint from S, and is zero otherwise. This is indeed a test, because the above lemma implies that its expected value for $X \sim Q$ is bounded by 1. Since the test is also computable, it is a lower bound to the optimal test $\mathbf{t}(X|Q,s)$, up to a constant factor. By

stochasticity of the set D, $g(D|Q,s) < O(1)\mathbf{t}(D|Q,s) < O(2^\beta)$, because $2^{\mathbf{d}(X|Q,s)}$ is an optimal Q test relative to s. Thus for large enough β, D intersects Q.

It remains to construct a description of each element in S of the size given in the proposition. We construct a special decompressor that assigns short description to each element in S. On input of a string, the decompressor interprets the string as a concatenation of 4 parts:

1. A prefix-free description of Q of size at most α.
2. A prefix-free description of $\log\beta$ of size $\mathbf{K}(\log\beta)$.
3. A prefix-free description of s of size $\mathbf{K}(s)$.
4. An integer of bitsize $\log(\beta/\gamma) = s + \log\beta$.

It interprets the last integer as the index of an element in the set S of size $\lceil\beta/\gamma\rceil$ that is computed by the search procedure on input Q, β, and γ. The element is the output of the decompressor. The proposition is proven for computable P. □

Remark 1 *If P is computable, a set S satisfying the conditions of the lemma can be easily searched. But if P is not computable, then the collection of sets D with $P(D) \geq \gamma$ grows over time. Thus after constructing a good S, it can happen that a large Q-measure of sets D appears that does not contain an element from S, and that new elements to S need to be added. This type of interactive construction leads to an equivalent characterization of the problem in terms of a game which is shown in [She12]. Below, another proof is presented. It might turn out to be useful to derive resource bounded versions of the theorem.*

Proof of Proposition 1 for lower-semicomputable P. We still assume that β is a large power of 2. Let $\gamma = 2^{-s}/2$. We can rewrite $P = \frac{\gamma}{\beta}(P_1 + \cdots + P_f + P_*)$, with $f \leq \beta/\gamma$, such that $P_1, \ldots P_f$ are probability measures with finite support obtained by a lower semi-computable approximation of P, and P_* is a lower-semicomputable semimeasure.

Construction of a lower-semicomputable test g over sets. We first construct tests $g_1, \ldots g_f$ together with a list of strings $z_1, \ldots, z_f$. Let $g_0(X) = 1$. Assume we already constructed $z_1, \ldots, z_{i-1}$ and g_{i-1} for some $i = 1, \ldots, f$. Choose z_i such that the test

$$g_i(X) = \begin{cases} g_{i-1}(X) & \text{if } g_{i-1}(X) \geq \exp(\beta) \\ \exp(P_i(X))g_{i-1}(X) & \text{if } g_{i-1}(X) < \exp(\beta) \text{ and } X \text{ is disjoint from } \{z_1, \ldots, z_i\} \\ 0 & \text{otherwise.} \end{cases}$$

satisfies $\mathbf{E}g_i(X) \leq \mathbf{E}g_{i-1}(X)$ where the expectations are taken for $X \sim Q$. Let $g(X)$ be equal to $\exp(\beta)$ if there exists an i such that $g_i(X) \geq \exp\beta$, otherwise let $g(X) = 0$. *End of construction*

We first show that each required string z_i in the construction exists. Suppose $z_1, \ldots, z_{i-1}$ and g_{i-1} have already been constructed. We show the existence of z_i using the probabilistic method. If we draw z_i according to P_i, then for each set X for which the second condition of g_i is satisfied, we have

$$\mathbf{E}_{z_i \sim P_i} g_i(X) \leq (1 - P_i(X))g_{i-1}(X)\exp P_i(X) \leq g_{i-1}(X),$$

because of the inequality $1 + r \leq \exp(r)$ for all reals r. If X satisfies the first or third condition, then $\mathbf{E}g_i(X) \leq \mathbf{E}g_{i-1}(X)$ is trivially true. So

$$\mathbf{E}_{X\sim Q}\mathbf{E}_{z_i\sim P_i}g_i(X) \leq \mathbf{E}_{X\sim Q}g_{i-1}(X),$$
$$\mathbf{E}_{z_i\sim P_i}\mathbf{E}_{X\sim Q}g_i(X) \leq \mathbf{E}_{X\sim Q}g_{i-1}(X),$$

and the required z_i exists.

We have $G(x) \leq O(\mathbf{t}(X|Q,(\gamma,\beta)))$, where $\mathbf{t}$ is the optimal test because the construction implies $\mathbf{E}g \leq 1$ and is effective, thus g is lower semicomputable. Every set X with $P(X) \geq 2^{-s} = 2\gamma$ satisfies $P_1(X) + \cdots + P_f(X) \geq \frac{\beta}{\gamma}P(D) - 1 \geq 2\beta - 1 \geq \beta$ by choice of P_i. Any such X that is disjoint from the set $\{z_1, \ldots, z_f\}$ satisfies

$$g_f(X) = \exp(P_1(X))\exp(P_2(X))\ldots\exp(P_f(X)) \geq \exp(\beta).$$

This implies $\mathbf{d}(X|Q,s) > \beta$ for large β, because up to $O(1)$ constants, we have

$$1.44\beta \leq \log g(X) \leq \mathbf{d}(X|Q,(\beta,\gamma)) \leq \mathbf{d}(X|Q,s) + 2\log\beta.$$

By the assumption on (α,β)-stochasticity of D, we have $\mathbf{d}(D|Q,s) \leq \beta$ and hence D must contain some z_j. The theorem follows by constructing a description for each string z_i of bitsize $s + \alpha + \log\beta + \mathbf{K}(\log\beta) + \mathbf{K}(s)$ in a similar way as above. □

4.3 Non-Stochastic Objects

The stochasticity of an object can be measured by

$$\mathbf{Ks}(x) = \min\{\mathbf{K}(P) + O(\log\max\{\mathbf{d}(x|P), 1\}) : P \text{ is an elementary probability measure}\}.$$

This term combines the complexity of the model P with how well it fits x, i.e. the randomness deficiency $\mathbf{d}$. It is well known in the literature that non-stochastic objects have high mutual information with the halting sequence [VS17]. In the following lemma, we reprove this fact, without using left-total machines, which was used in the original proof.

Lemma 3 $\mathbf{Ks}(x) <^{\log} \mathbf{I}(x;\mathcal{H})$.

Proof. We dovetail all programs to the universal Turing machine U. For $p \in \text{Domain}(U)$, $n(p) \in \mathbb{N}$ is the position in which the program $p \in \{0,1\}^*$ terminates. Let $\Omega^n = \sum_{p:n(p)<n} 2^{-\|p\|}$ and $\Omega = \Omega^\infty$ be Chaitin's Omega. Let Ω^n_t be Ω^n restricted to the first t digits. Let $x^* \in \{0,1\}^{\mathbf{K}(x)}$, with $U(x^*) = x$ with minimum $n(x^*)$. Let $k(p) = \max\{\ell : \Omega^{n(p)}_\ell = \Omega_\ell\}$ and $k = k(x^*)$. We define the elementary probability measure $Q(x) = \max\{2^{-\|p\|+k} : k(p) = k, U(p) = x\}$, $Q(\emptyset) = 1 - Q(\{0,1\}^* \setminus \{\emptyset\})$.

$$\begin{aligned}
\mathbf{d}(x|Q) &= -\log Q(x) - \mathbf{K}(x|Q) <^+ (\mathbf{K}(x) - k) - \mathbf{K}(x|\Omega_k) \\
&<^+ (\mathbf{K}(x|\Omega_k) + \mathbf{K}(\Omega_k) - k) - \mathbf{K}(x|\Omega_k) <^+ (k + \mathbf{K}(k)) - k \\
&<^+ \mathbf{K}(k).
\end{aligned}$$

$$\begin{aligned}
\mathbf{K}(x|\mathcal{H}) &<^+ \mathbf{K}(x|Q) + \mathbf{K}(Q|\mathcal{H}) <^+ \mathbf{K}(x|Q) + \mathbf{K}(\Omega_k|\mathcal{H}) \\
&<^+ -\log Q(x) + \mathbf{K}(k) <^+ (\mathbf{K}(x) - k) + \mathbf{K}(k) \\
k &<^{\log} \mathbf{K}(x) - \mathbf{K}(x|\mathcal{H})
\end{aligned}$$

$$\mathbf{Ks}(x) <^+ \mathbf{K}(Q) + O(\log\max\{\mathbf{d}(x|P), 1\}) <^+ k + O(\mathbf{K}(k)) <^{\log} \mathbf{I}(x;\mathcal{H}).$$

□

The following corollary comes from Theorem 14 and Lemma 3

Corollary 1 (Sets Have Simple Members)
For finite $D \subset \{0,1\}^*$, $\min_{x\in D} \mathbf{K}(x) <^{\log} -\log \mathbf{m}(D) + \mathbf{I}(D;\mathcal{H})$.

5 October 11th, 2022: Clusters, Information Distances, and Left-Total Machines

5.1 Clustering

In this blog entry, I talk about clustering using the information distance. I'll also take this opportunity to showcase the benefit of left-total machines.

In cluster analysis, the goal is grouping objects in such a way that objects that are "similar" in some way are in the same group. It is a ubiquitous technique found in many research fields including computer vision, machine learning, information retrieval, and bioinformatics. Clustering is not a specific algorithms, but a collection of different approaches. However, in general there are two components, a distance (or similarity) measure and a dataset containing elements of a similar kind, such as songs or DNA sequences.

A question one might ask is whether there is a "core" to each cluster. That is, is there realizable object distilling the common components of the members? This makes more sense in clusters of DNA sequences rather than clusters of songs. To prove the existence of such cores, we need to prove it with respect to a distance measure. Just like in the previous blog, where I used a universal prior, in this blog, I'll use a universal distance. In AIT, there is such a notion, and it is known as the information distance [BGL+98],

$$\mathbf{E}_1(x,y) = \max\{\mathbf{K}(y|x), \mathbf{K}(x|y)\}.$$

The term $\mathbf{K}$ is the prefix free Kolmogorov complexity. $\mathbf{E}_1$ is universal over so called *admissible* distances. An admissible distance $\mathbf{D}$, is symmetric, satisfies the triangle inequality, is upper-semicomputable and normalized, that is

$$\sum_{y:y\neq x} 2^{-\mathbf{D}(x,y)} \leq 1.$$

Theorem ([BGL+98]) *For an appropriate constant c, let $\mathbf{E}(x,y) = \mathbf{E}_1(x,y) + c$ if $x \neq y$ and 0 otherwise. Then $\mathbf{E}(x,y)$ is a universal admissible metric in that every admissible distance $D(x,y)$ we have*

$$\mathbf{E}(x,y) <^+ \mathbf{D}(x,y).$$

In fact there is a large literature about the real world applications of information distances. This involves replacing $\mathbf{E}_1$ with computable measures such as compression algorithms [CV05] or using the Google search engine [CV07].

5.2 Information Cores

Let $\mathbf{I}(x;\mathcal{H}) = \mathbf{K}(x) - \mathbf{K}(x|\mathcal{H})$ be the amount of information that x has with the halting sequene $\mathcal{H}$. We say $X \subset \{0,1\}^*$ is an (m,ℓ)-*cluster* if for all $x, y \in X$, $\mathbf{E}_1(x,y) \leq m$ and $\log|X| \geq \ell$. The following recent theorem shows that all clusters of a certain size with low information distance will contain an information core.

Theorem 15 ([Rom22]) *Let X be an (m,ℓ) cluster. There exists $z \in \{0,1\}^*$ where for all $x \in X$, $\mathbf{K}(z|x) <^+ O(m-\ell) + \mathbf{K}(m)$ and $\mathbf{K}(x|z) <^+ m + O(m-\ell) + \mathbf{K}(m)$.*

In fact, for non-exotic clusters, the information core will be a member of the cluster. One canonical cluster is as follows. Let z be a random string of length n, and X be all strings of length $2n$ that start with z. Then the string $z0^n$ can be seen as an information core that of X.

Theorem 16 ([Eps21a]) *Let X be an (m, ℓ) cluster. There exists $z \in X$ where for all $x \in X$, $\mathbf{K}(z|x) <^{\log} 2(m-\ell) + \mathbf{I}(X; \mathcal{H})$.*

The last result states that non-exotic clusters will contain members that "clump" together with respect to the information distance. Thus there will exist two members of a cluster that are very close together, provided that the cluster is a large enough size and the information distance between members is small enough.

Theorem 17 ([Eps22c]) *Let X be an (m, ℓ) cluster. There exists $x, y \in X$ with $\mathbf{K}(y|x) <^{\log} \lceil m - 2\ell \rceil^+ + \mathbf{I}(X; \mathcal{H}) + \mathbf{K}(m, \ell)$.*

5.3 Left-Total Machines

There are many proofs in the literature that use the bits Chaitin's Omega, $\Omega \in \mathbb{R}$. This includes finite prefixes of Ω, sometimes denoted $\Omega(1 \dots t)$, as well as a lower approximation $\Omega^t < \Omega$ given some time resource t. Take for example, Theorem 3.3.1, in [GÍ3], Proposition 20 in [VS17], or the previous blog entry. However as manipulations become more involved, there is utility in denoting with these concepts in another way using so-called teft-total machines, first appearing in [EL11], also appearing in [Lev16]. In this blog, I provide a concise description of left-total machines. For more details on left-total machines, we refer readers to the self-contained Section 7 in [Eps22c].

We say $x \in \{0,1\}^*$ is total with respect to a machine if the machine halts on all sufficiently long extensions of x. More formally, x is total with respect to T_y for some $y \in \{0,1\}^{*\infty}$ iff there exists a finite prefix-free set of strings $Z \subset \{0,1\}^*$ where $\sum_{z \in Z} 2^{-\|z\|} = 1$ and $T_y(xz) \neq \perp$ for all $z \in Z$. We say (finite or infinite) string $\alpha \in \{0,1\}^{*\infty}$ is to the "left" of $\beta \in \{0,1\}^{*\infty}$ and use the notation $\alpha \lhd \beta$ if there exists an $x \in \{0,1\}^*$ such that $x0 \sqsubseteq \alpha$ and $x1 \sqsubseteq \beta$. A machine T is left-total if for all auxiliary strings $\alpha \in \{0,1\}^{*\infty}$ and for all $x, y \in \{0,1\}^*$ with $x \lhd y$, one has that $T_\alpha(y) \neq \perp$ implies that x is total with respect to T_α. For the rest of the paper, we assume that the universal Turing machine U is left-total. Let $(p0)^- = (p1)^- = p$.

Proposition 2 *There exists a unique infinite sequence $\mathcal{B}$ with the following properties.*

1. *All the finite prefixes of $\mathcal{B}$ have total and nontotal extensions.*
2. *If a finite string has total and nontotal extensions, then it is a prefix of $\mathcal{B}$.*
3. *If a string b is total and b^- is not, then $b^- \sqsubset \mathcal{B}$.*

We call this infinite sequence $\mathcal{B}$, "border" because for any string $x \in \{0,1\}^*$, $x \lhd \mathcal{B}$ implies that x is total with respect to U and $\mathcal{B} \lhd x$ implies that U will never halt when given x as an initial input. It is equal to the binary expansion of Chaitin's Omega. Figure 1 shows the domain of left-total U with respect to $\mathcal{B}$. For total string b, we define the busy beaver function,

$$\mathbf{bb}(b) = \max\{\|x\| : U(p) = x, p \lhd b \text{ or } p \sqsupseteq b\}.$$

Using Figure 1, $\mathbf{bb}(v0) \leq \mathbf{bb}(v1)$ and $\mathbf{bb}(v0) \leq \mathbf{bb}(v)$.

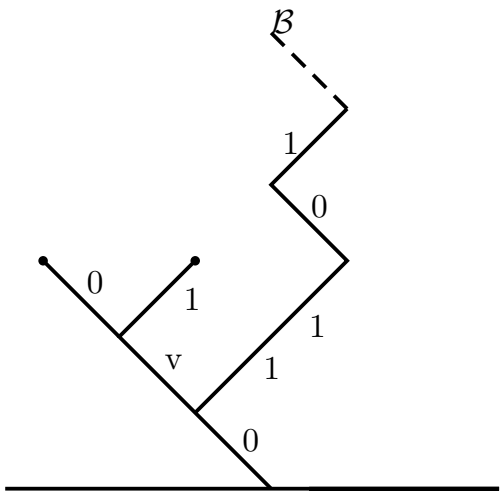

Figure 1: The above diagram represents the domain of the universal left-total algorithm U, with 0 bits branching to the left and 1 bits branching to the right. The strings in the above diagram, $0v0$ and $0v1$, are halting inputs to U with $U(0v0) \neq \perp$ and $U(0v1) \neq \perp$. Therefore, $0v$ is a total string. The infinite border sequence $\mathcal{B} \in \{0,1\}^\infty$ represents the unique infinite sequence such that all its finite prefixes have total and nontotal extensions. All finite strings branching to the right of $\mathcal{B}$ will cause U to diverge.

5.4 Total Computable Conversion Function

Left-total machines can be used in proofs to change partial recursive functions into total computable ones. Similarly, for more advanced proofs, it can turn lower-computable semi-measures (namely the algorithmic probability $\mathbf{m}$) into recursive ones. This was done in the proofs of all the theorems of the first blog post.

Definition 10 *We define the following two functions, representing the conversion distance between two strings.* $\mathbf{E}_0$ *was introduced in [BGL+98].*

- $\mathbf{E}_0(x,y) = \min\{\mathbf{K}(p) : p(x) = y, p(y) = x\}$.
- $\mathbf{F}_0(x,y) = \min\{\mathbf{K}(f) : f \text{ is total computable}, f(x) = y, f(y) = x\}$.

We start by recreating the proof showing that the min conversion distance is equal, up to a logarithmic error term, to $\mathbf{E}_1$.

Theorem 18 ([BGL+98])

$$\mathbf{E}_0(x,y) =^{\log} \mathbf{E}_1(x,y).$$

Proof. The first inequality is straitforward. Let $k = \mathbf{E}_1(x,y)+1$. We describe the MATCHING(k) algorithm. Let $G = (V,E)$ be an infinite labelled graph where every string has a corresponding vertex and undirected edges are enumerated and labeled with 2^{k+1} colors. The algorithm enumerates undirected edge (x,y) if $\mathbf{K}(y|x) < k$ and $\mathbf{K}(x|y) < k$. For any vertex x, there are at most 2^k edges (x,y) enumerated, because an edge implies $\mathbf{K}(y|x) < k$. Thus one can dynamically color the graph with 2^{k+1} colors such that no two edges with a common vertex share a color.

Let $c \in \{0,1\}^{k+1}$ be the string corrosponding to the color of of the edge (x,y). There is a function f that runs MATCHING(k) and on input x follows c and outputs y. Similarly, on

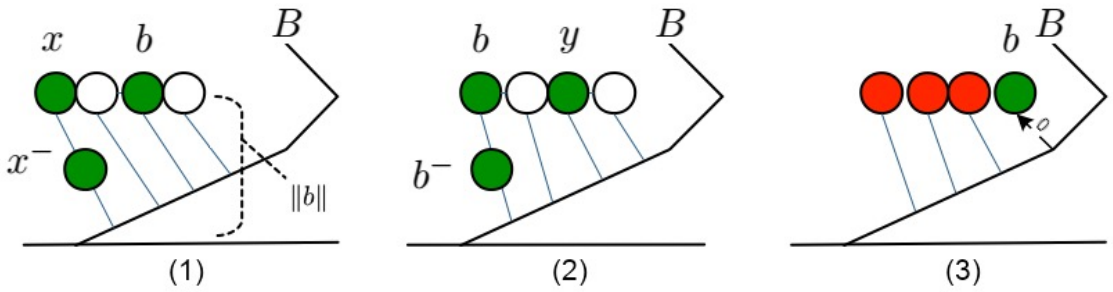

Figure 2: The above diagram represents the domain of the universal left-total Turing machine U and uses the same conventions as Figure 1, with 0s branching to the left and 1s branching to the right. It shows all the total strings of length $\|b\|$, including b. The large diagonal line is the border sequence, B. A string c is marked green if $(x, y) \in$ MATCHING(k, c). By definition, b is a shortest green string. If x is green and total, and $x \lhd y$, and y is total, then y is green, since $\mathbf{bb}(x) \leq \mathbf{bb}(y)$. Furthermore, if x is green and total and x^- is total, then x^- is green, as $\mathbf{bb}(x) \leq \mathbf{bb}(x^-)$. It cannot be that there is a green $x \lhd b$ with $\|x\| = \|b\|$. Otherwise, x^- is total, and thus, it is green, causing a contradiction because it is shorter than b. This is shown in part (1). Furthermore, there cannot be a green y, with $b \lhd y$ and $\|y\| = \|b\|$. Otherwise, b^- is total and thus green, contradicting the definition of b. This is shown in part (2). Thus, b is unique, and since b^- is not total, by Proposition 2, b^- is a prefix of the border, as shown in part (3). Thus, an algorithm returning a green string of length $\|b\|$ will return b.

input y, the function follows c and outputs y. Thus $\mathbf{K}(f) <^+ \mathbf{K}(c, k) <^+ k + \mathbf{K}(k)$, implying $\mathbf{E}_0(x, y) <^{\log} \mathbf{E}_1(x, y)$. □

Theorem 19

$$\mathbf{E}_1(x, y) <^+ \mathbf{F}_0(x, y) <^{\log} \mathbf{E}_1(x, y) + \mathbf{I}((x, y); \mathcal{H}).$$

Proof. The first inequality is straitforward. Let $k = \mathbf{E}_1(x, y) + 1$. Let MATCHING$(k, d)$ be the matching algorithm described above but it halts after $\mathbf{bb}(d)$ steps for some total string d. Let b be a shortest total string such that MATCHING(k, b) enumerates the link (x, y) with the color c. Let $f(z)$ be the function defined as follows. It runs the MATCHING(k, b) algorithm then, starting at z, it follows the edge marked c and returns the label of the adjacent vertex. If there is no edge marked c, then f returns 0. So f is total computable, $f(x) = y$, and $f(y) = x$. Thus $\mathbf{F}_0(x, y) <^+ \mathbf{K}(f) <^+ \mathbf{K}(k) + k + \mathbf{K}(b) <^{\log} \mathbf{E}_1(x, y) + \mathbf{K}(b)$. It remains to prove that $\mathbf{K}(b) <^{\log} \mathbf{I}((x, y); \mathcal{H}) + \mathbf{K}(k)$.

From Lemma 2 in [Eps22d], we have $\mathbf{I}(b; \mathcal{H}) <^+ \mathbf{I}((x, y); \mathcal{H}) + \mathbf{K}(b|(x, y))$. Furthermore since b is total and b^- is not, by Proposition 2, b^- is a prefix of border, which is the binary expansion of Chaitin's Omega. Thus $\mathbf{K}(b) <^{\log} \mathbf{I}(b; \mathcal{H})$. Furthermore $\mathbf{K}(b|(x, y)) <^+ \mathbf{K}(\|b\|, k)$, because there is a program, that enumerates total strings of length $\|b\|$ from left to right and return the first string d such that $(x, y) \in$ MATCHING(k, d). This returned string is b, as shown by Figure 3. Thus $\mathbf{K}(b) <^{\log} \mathbf{I}(b; \mathcal{H}) <^{\log} \mathbf{I}((x, y); \mathcal{H}) + \mathbf{K}(\|b\|, k) <^{\log} \mathbf{I}((x, y); \mathcal{H}) + \mathbf{K}(k)$. This is because since b is random, $\mathbf{K}(\|b\|) = O(\log \mathbf{K}(b))$. □

6 October 14th, 2022: Resolving Four Open Problems in Quantum Information Theory

In this blog entry, I talk about the resolution of 4 open problems in Quantum Information Theory. The problems were posed as questions in [GÓ1]. Quantum information theory studies ultimate capabilities of noisy physical systems, governed by the laws of quantum mechanics, to preserve information and correlations. Quantum Information Theory encompasses subjects as diverse as quantum computation, quantum algorithms, quantum complexity theory, quantum communication complexity, entanglement theory, quantum key distribution, quantum error correction, and even the experimental implementation of quantum protocols.

Quantum description complexity, started in [**?**, GÓ1], measures the algorithmic complexity of quantum states. In [**?**], complexity was defined as the shortest program to a universal quantum Turing machine that produces the target state, which we denote as **QC**. The measure **QC** has been refered to as Quantum Kolmogorov complexity. In [GÓ1], the complexity of a state was measured using a universal lower semicomputable mixed state, which we will call **H**. As stated in [GÓ1], one of the possible applications of **H** is to gain new insights into von Neumann entropy. In [GÓ1], four open problems were posed as questions.

1. Does smallness of **H** imply smallness of **QC**?
2. What is the proper quantum generalization of classical information non-growth laws?
3. What addition theorems apply to **H**?
4. Does **H** obey strong superadditivity?

In the recent literature [Eps19a, Eps20] problems (1-3) are resolved, which I will briefly review. In this blog entry, I resolve the 4th open question in the negative: **H** is neither strongly superadditive or strongly subadditive. A function **L** from quantum mixed states to whole numbers is strongly subadditive if there exists a constant $c \in \mathbb{N}$ such that for all mixed states ρ_{123}, $\mathbf{L}(\rho_{123}) + \mathbf{L}(\rho_2) < \mathbf{L}(\rho_{12}) + \mathbf{L}(\rho_{23}) + c$. Similarly **L** is strongly superadditive if there exists a constant $c \in \mathbb{N}$ such that for all mixed states ρ_{123}, $\mathbf{L}(\rho_{12}) + \mathbf{L}(\rho_{23}) < \mathbf{L}(\rho_{123}) + \mathbf{L}(\rho_2) + c$. I'll also pose some new questions about the intersection of AIT and physics.

6.1 Conventions

We use $\mathcal{H}_n$ to denote a Hilbert space with n dimensions, spanned by bases $|\beta_1\rangle, \ldots, |\beta_n\rangle$. A qubit is a unit vector in the Hilbert space $\mathcal{G} = \mathcal{H}_2$, spanned by vectors $|0\rangle$, $|1\rangle$.

To model n qubits, we use a unit vector in $\mathcal{H}_{2^n}$, spanned by basis vectors $|i\rangle$, where $i \in [1..2^n]$. A pure quantum state $|\psi\rangle$ of length n is a unit vector in $\mathcal{H}_{2^n}$. Its corresponding element in the dual space is denoted by $\langle\psi|$.

The conjugate transpose of a matrix A is A^*. The tensor product of two matrices A and B is $A \otimes B$. Tr is used to denote the trace of a matrix.

Let $H_A \otimes H_B$ be a Hilbert space and O be an operator acting on this composite space. Then $O = \sum_{i,j} c_{ij} M_i \otimes N_j$, where M_i and N_j are operators acting on H_A and H_B respectively. The partial trace over H_A is $\mathrm{Tr}_{H_A}(O) = \sum_{i,j} c_{ij} \mathrm{Tr}(M_i) N_j$. Furthermore $\mathrm{Tr}(I \otimes M)\rho_{12} = \mathrm{Tr} M \rho_2$. Density matrices are used to represent mixed states, and are self-adjoint, positive definite matrices with trace equal to 1. Semi-density matrices are density matrices except they may have a trace in [0,1].

The maximally mixed state is $2^{-n}I$.

Pure quantum states are elementary if their values are complex numbers with rational coefficients, and thus they can be represented with finite strings. Thus elementary quantum states $|\phi\rangle$ can be encoded as strings and assigned Kolmogorov complexities $\mathbf{K}(|\phi\rangle)$, and algorithmic probabilities $\mathbf{m}(|\phi\rangle)$. They are equal to the complexity (and algorithmic probability) of the strings that encodes the states.

In [GÓ1], a universal lower computable semi-density matrix, $\boldsymbol{\mu}$ was introduced. It can be defined (up to a multiplicative constant) by $\boldsymbol{\mu}^n = \sum_{\text{elementary } n \text{ qubit } |\phi\rangle} \mathbf{m}(|\phi\rangle / n) |\phi\rangle \langle\phi|$, where the summation is over all n qubit elementary pure quantum states, and $\mathbf{m}$ is the algorithmic probability. The quantum algorithmic entropy of an n qubit mixed state ρ is $\mathbf{H}(\rho) = \lceil -\log \mathrm{Tr}\boldsymbol{\mu}^n \rho \rceil$. This definition of algorithmic entropy generalizes the definition of $\underline{H}$ in [GÓ1] to mixed states.

6.2 Results

Problem 1. Quantum Kolmogorov complexity, defined in [**?**], uses a universal quantum Turing machine to define the complexity of a pure quantum state. The input and output tape of this machine consists of symbols of the type 0, 1, and #. The input is an ensemble $\{p_i\}$ of pure states $|\psi_i\rangle$ of the same length n, where $p_i \geq 0$ and $\sum_i p_i = 1$. This ensemble can be represented as a mixed state of n qubits. If, during the operation of the quantum Turing machine, all computational branches halt at a time t, then the contents on the output tape are considered the output of the quantum Turing machine. The quantum Kolmogorov complexity of a pure state, $\mathbf{QC}[\epsilon](|\psi\rangle)$ is the size of the smallest (possibly mixed state) input to a universal quantum Turing machine such that fidelity between the output and $|\psi\rangle$ is at least ϵ. The fidelity between a mixed state output σ and $|\psi\rangle$ is $\langle\psi|\,\sigma\,|\psi\rangle$.

Theorem. [Eps20] $\mathbf{QC}[\langle\psi|\,\boldsymbol{\mu}\,|\psi\rangle\, \mathbf{H}^{-O(1)}(|\psi\rangle)](|\psi\rangle) <^{\log} \mathbf{H}(|\psi\rangle)$.

Problem 2. In AIT, Information has been shown to be conserved with respect to randomized transformations for a large number of information terms over strings or infinite sequences. This property can be shown to be true for quantum states as well. In [Eps19a], an information term $\mathbf{I}$ between quantum mixed states was defined. This term is the summation of so called quantum tests, each weighted by their complexity. Using this definition, conservation was proven over unitary transform. In fact, this result can be strengthed to any computable quantum operation.

Theorem. [Eps19a] *For density matrices σ and ρ, relativized to elementary unitary transform A,* $\mathbf{I}(A\sigma A^* : \rho) =^+ \mathbf{I}(\sigma : \rho)$.

Problem 3. The addition theorem for classical entropy asserts that the joint entropy for a pair of random variables is equal to the entropy of one plus the conditional entropy of the other, with $\mathcal{H}(\mathcal{X}) + \mathcal{H}(\mathcal{Y}/\mathcal{X}) = \mathcal{H}(\mathcal{X}, \mathcal{Y})$. For algorithmic entropy, the chain rule is slightly more nuanced, with $\mathbf{K}(x) + \mathbf{K}(y/x, \mathbf{K}(x)) =^+ \mathbf{K}(x, y)$. An analogous relationship cannot be true for $\mathbf{H}$ since as shown in Theorem 15 of [GÓ1], there exists elementary $|\phi\rangle$ where $\mathbf{H}(|\phi\rangle\,|\phi\rangle) - \mathbf{H}(|\phi\rangle)$ can be arbitrarily large, and $\mathbf{H}(|\phi\rangle\,/\,|\phi\rangle) =^+ 0$. However, the following theorem shows that a chain rule inequality does hold for $\mathbf{H}$.

Theorem. [Eps19a] *For semi-density matrices σ, ρ, elementary ρ,*

$\mathbf{H}(\rho) + \mathbf{H}(\sigma/\langle \rho, \mathbf{H}(\rho)\rangle) <^+ \mathbf{H}(\sigma \otimes \rho)$.

Problem 4. The following two theorems address the last open problem in [GÓ1]. It is still an open question as to whether $\mathbf{H}$ is subadditive, that is $\mathbf{H}(\rho_{12}) <^+ \mathbf{H}(\rho_1) + \mathbf{H}(\rho_2)$. This property holds when $\rho_{12} = \rho_1\rho_2$ but it is unknown whether this property holds for arbitrary mixed states.

Theorem. $\mathbf{H}$ *is not strongly subadditive.*

Proof. We fix the number of qubits n, and for $i \in [1..2^n]$, $|i\rangle$ is the ith basis state of the n qubit space. Let $|\psi\rangle = \sum_{i=1}^{2^n} 2^{-n/2} |i\rangle |i\rangle$. The pure state $|\psi\rangle$ is elementary, with $\mathbf{K}(|\psi\rangle /2n) =^+ 0$. We define the the $3n$ qubit mixed state $\rho_{123} = .5 |\psi\rangle\langle\psi| \otimes |1\rangle\langle 1| + .5 |1\rangle\langle 1| \otimes |\psi\rangle\langle\psi|$. $\rho_{12} = .5 |\psi\rangle\langle\psi| + .5 |1\rangle\langle 1| \otimes 2^{-n} I$. $\rho_{23} = .5 * 2^{-n} I \otimes |1\rangle\langle 1| + .5 |\psi\rangle\langle\psi|$. $\rho_2 = 2^{-n} I$. $\mathbf{H}(\rho_{12}) =^+ -\log \mathrm{Tr} \boldsymbol{\mu}^{2n} \rho_{12} <^+ -\log \mathrm{Tr} \boldsymbol{\mu}^{2n} |\psi\rangle\langle\psi| <^+ -\log \mathbf{m}(|\psi\rangle /2n) |\langle\psi|\psi\rangle|^2 <^+ 0$. Similarly, $\mathbf{H}(\rho_{23}) =^+ 0$. $\mathbf{H}(\rho_2) =^+ n$. So $\mathbf{H}(\rho_{123}) + \mathbf{H}(\rho_2) >^+ n$ and $\mathbf{H}(\rho_{12}) + \mathbf{H}(\rho_{23}) =^+ 0$, proving that $\mathbf{H}$ is not strongly subadditive.

Theorem. $\mathbf{H}$ *is not strongly superadditive.*

Proof. We fix the number of qubits n, and for $i \in [1..2^n]$, $|i\rangle$ is the ith basis state of the n qubit space. Let $|\phi\rangle = \sum_{i=1}^{2^n} 2^{-n/2} |i\rangle |i\rangle |i\rangle$, with $\mathbf{K}(|\phi\rangle /3n) = 0$. Let $\sigma_{123} = |\phi\rangle\langle\phi|$. $\sigma_{12} = \sigma_{23} = \sum_{i=1}^{2^n} 2^{-n} |i\rangle\langle i| \otimes |i\rangle\langle i|$. $\mathbf{H}(\sigma_{123}) =^+ -\log \mathrm{Tr} \sigma_{123} \boldsymbol{\mu}^{3n} <^+ -\log \mathrm{Tr}\, \mathbf{m}(|\phi\rangle /3n) |\langle\phi|\phi\rangle|^2 <^+ 0$. Let D be a unitary transform where $D |i\rangle |i\rangle = |i\rangle |1\rangle$ and $\mathbf{K}(D/2n) =^+ 0$. By Theorem 11 in [GÓ1], $\mathbf{H}(\sigma_{12}) =^+ \mathbf{H}(D\sigma_{12}D^*) =^+ \mathbf{H}(2^{-n} I \otimes |1\rangle\langle 1|) =^+ n - \log \mathrm{Tr}(I \otimes |1\rangle\langle 1|) \boldsymbol{\mu}^{2n}$. By Theorem 4 of [GÓ1] and properties of partial trace, $\mathbf{H}(2^{-n} I \otimes |1\rangle\langle 1|) =^+ n - \log \mathrm{Tr} |1\rangle\langle 1| \boldsymbol{\mu}^n =^+ n$. So $\mathbf{H}(\sigma_{12}) = \mathbf{H}(\sigma_{23}) =^+ n$. So $\mathbf{H}(\sigma_{123}) + \mathbf{H}(\sigma_2) <^+ n$, and $\mathbf{H}(\sigma_{12}) + \mathbf{H}(\sigma_{23}) >^+ 2n$, proving that $\mathbf{H}$ is not strongly superadditive.

6.3 New Questions on AIT and Physics

Currently there is relatively little overlap in the literature of Algorithmic Information Theory and physics. This combination represents new opportunities for applying properties and relationships found in AIT to the foundations of our physical laws. Here are some new questions.

1. In [**?**], the notion of Quantum Martin Löf Random States was introduced for infinite quantum spin chains. One question is, how do such infinite quantum states transform? Is Quantum Martin Löf Randomness preserved over such transformations?

2. In [Eps19a], the notion of a typical quantum state was defined. Is there such a thing as a typical particle? i.e. a typical wave function? Certaintly one can define typical measurements of a wave function.

3. In [Gac94], two notions of algorithmic thermodynamic entropy were introduced. Can the Theorem 7 of my September 30th blog post be applied to prove fluctuations of thermodynamic entropy? In [Gac94], a reformulation of Maxwell's demon was achieved. Can recent results of fluctuations of randomness be used to revisit Maxwell's demon?

7 October 18th, 2022: A Complication for the Many Worlds Theory

The Many Worlds Theory (**MWT**) was formulated by Hugh Everett [Eve57] as a solution to the measurement problem of Quantum Mechanics. Branching (a.k.a splitting of worlds) occurs during any process that magnifies microscopic superpositions to the macro-scale. This occurs in events including human measurements such as the double slit experiments, or natural processes such as radiation resulting in cell mutations.

One question is if **MWT** causes issues with the foundations of computer science. The physical Church Turing Thesis (**PCTT**) states that any functions computed by a physical system can be simulated by a Turing machine. A straw man argument for showing **MWT** and **PCTT** are in conflict is an experiment that measures the spin of an unending number of electrons, with each measurement bifurcating the current branch into two sub-branches. This results in a single branch in which the halting sequence is outputted. However this branch has Born probability converging to 0, and can be seen as a deviant, atypical branch.

In fact, conflicts do emerge between **MWT** and Algorithmic Information Theory. In particular, the Independence Postulate (**IP**) is a finitary Church-Turing thesis, postulating that certain infinite and *finite* sequences cannot be found in nature, a.k.a. have high "addresses". If a forbidden sequence is found in nature, an information leak will occur. However **MWT** represents a theory in which such information leaks can occur. This blog entry covers the main arguments of this conflict.

7.1 Many Worlds Theory

Some researchers believe there is an inherent problem in quantum mechanics. On one hand, the dynamics of quantum states is prescribed by unitary evolution. This evolution is deterministic and linear. On the other hand, measurements result in the collapse of the wavefunction. This evolution is non-linear and nondeterministic. This conflict is called the measurement problem of quantum mechanics.

The time of the collapse is undefined and the criteria for the kind of collapse are strange. The Born rule assigns probabilities to macroscopic outcomes. The projection postulate assigns new microscopic states to the system measured, depending on the the macroscopic outcome. One could argue that the apparatus itself should be modeled in quantum mechanics. However it's dynamics is deterministic. Probabilities only enter the conventional theory with the measurement postulates.

MWT was proposed by Everett as a way to remove the measurement postulate from quantum mechanics. The theory consists of unitary evolutions of quantum states without measurement collapses. For **MWT**, the collapse of the wave function is the change in dynamical influence of one part of the wavefunction over another, the decoherence of one part from the other. The result is a branching structure of the wavefunction and a collapse only in the phenomenological sense.

7.1.1 Branching Worlds

An example of a branching of universes can be seen in an idealized Stern-Gerlarch experiment with a single electron with spin $|\phi_\uparrow\rangle$ and $|\phi_\downarrow\rangle$. This description can be found in [SBKW10]. There is a measuring apparatus $\mathcal{A}$, which is in an initial state of $|\psi^{\mathcal{A}}_{\text{ready}}\rangle$. After $\mathcal{A}$ reads spin-up or spin-down

then it is in state $|\psi^{\mathcal{A}}_{\text{reads spin }\uparrow}\rangle$ or $|\psi^{\mathcal{A}}_{\text{reads spin }\downarrow}\rangle$, respectively. The evolution for when the electron is solely spin-up or spin-down is

$$|\phi_\uparrow\rangle \otimes |\psi^{\mathcal{A}}_{\text{ready}}\rangle \overset{\text{unitary}}{\longrightarrow} |\phi_{\text{absorbed}}\rangle \otimes |\psi^{\mathcal{A}}_{\text{reads spin }\uparrow}\rangle$$
$$|\phi_\downarrow\rangle \otimes |\psi^{\mathcal{A}}_{\text{ready}}\rangle \overset{\text{unitary}}{\longrightarrow} |\phi_{\text{absorbed}}\rangle \otimes |\psi^{\mathcal{A}}_{\text{reads spin }\downarrow}\rangle .$$

Furthermore, one can model the entire quantum state of an observer $\mathcal{O}$ of the apparatus, with

$$\begin{aligned} &|\phi_\uparrow\rangle \otimes |\psi^{\mathcal{A}}_{\text{ready}}\rangle \otimes |\xi^{\mathcal{O}}_{\text{ready}}\rangle \\ \overset{\text{unitary}}{\longrightarrow} &|\phi_{\text{absorbed}}\rangle \otimes |\psi^{\mathcal{A}}_{\text{reads spin }\uparrow}\rangle \otimes |\xi^{\mathcal{O}}_{\text{ready}}\rangle \\ \overset{\text{unitary}}{\longrightarrow} &|\phi_{\text{absorbed}}\rangle \otimes |\psi^{\mathcal{A}}_{\text{reads spin }\uparrow}\rangle \otimes |\xi^{\mathcal{O}}_{\text{reads spin }\uparrow}\rangle \end{aligned}$$

$$\begin{aligned} &|\phi_\downarrow\rangle \otimes |\psi^{\mathcal{A}}_{\text{ready}}\rangle \otimes |\xi^{\mathcal{O}}_{\text{ready}}\rangle \\ \overset{\text{unitary}}{\longrightarrow} &|\phi_{\text{absorbed}}\rangle \otimes |\psi^{\mathcal{A}}_{\text{reads spin }\downarrow}\rangle \otimes |\xi^{\mathcal{O}}_{\text{ready}}\rangle \\ \overset{\text{unitary}}{\longrightarrow} &|\phi_{\text{absorbed}}\rangle \otimes |\psi^{\mathcal{A}}_{\text{reads spin }\downarrow}\rangle \otimes |\xi^{\mathcal{O}}_{\text{reads spin }\downarrow}\rangle . \end{aligned}$$

For the general case, the electron is in a state $|\phi\rangle = a\,|\phi_\uparrow\rangle + b\,|\phi_\downarrow\rangle$, where $|a|^2 + |b|^2 = 1$. In this case, the final superposition would be of the form:

$$\begin{aligned} &a\,|\phi_{\text{absorbed}}\rangle \otimes |\psi^{\mathcal{A}}_{\text{reads spin }\uparrow}\rangle \otimes |\xi^{\mathcal{O}}_{\text{reads spin }\uparrow}\rangle \\ +&b\,|\phi_{\text{absorbed}}\rangle \otimes |\psi^{\mathcal{A}}_{\text{reads spin }\downarrow}\rangle \otimes |\xi^{\mathcal{O}}_{\text{reads spin }\downarrow}\rangle . \end{aligned}$$

This is a superposition of two branches, each of which describes a perfectly reasonable physical story. This bifurcation is one method on how the quantum state of universe bifurcates into two branches.

7.1.2 Deriving the Born Rule

In my opinion, one of the glaring problem of **MWT** is its reconciliation of the Born rule, for which no proposed solution has universal consensus. In standard quantum mechanics, measurements are probabilistic operations. Measurements on a state vector $|\psi\rangle$, which is a unit vector over Hilbert space $\mathcal{H}$, are self-adjoint operators $\mathcal{O}$ on $\mathcal{H}$. Observables are real numbers that are the spectrum $\text{Sp}(\mathcal{O})$ of $\mathcal{O}$. A measurement outcome is a subset $E \subseteq \text{Sp}(\mathcal{O})$ with associated projector P_E on $\mathcal{H}$. Outcome E is observed on measurement of $\mathcal{O}$ on $|\psi\rangle$ with probability $P(E) = \langle\psi|\,P_E\,|\psi\rangle$. This is known as the Born rule. After this measurement, the new state becomes $P_E\,|\psi\rangle\,/\sqrt{\langle\psi|\,P_E\,|\psi\rangle}$. This is known as the projection postulate.

However, the Born rule and the projection postulate are not assumed by **MWT**. The dynamics are totally deterministic. Each branch is equally real to the observers in it. To address these issues, Everett first derived a typicality-measure that weights each branch of a state's superposition. Assuming a set of desirable constraints, Everett derived the typicality-measure to be equal to the norm-squared of the coefficients of each branch, i.e. the Born probability of each branch. Everett then drew a distinction between typical branches that have high typicality-measure and

exotic atypical branches of decreasing typicality-measure. For the repeated measurements of the spin of an electron $|\phi\rangle = a\,|\phi_\uparrow\rangle + b\,|\phi_\downarrow\rangle$, the relative frequencies of up and down spin measurements in a typical branch converge to $|a|^2$ and $|b|^2$, respectively. The notion of typicality can be extended to measurements with many observables.

In a more recent resolution to the relation between **MWT** and probability, Deutsch introduced a decision theoretic interpretation [Deu99] that obtains the Born rule from the non-probabilistic axioms of quantum theory and non-probabilistic axioms of decision theory. Deutsch proved that rational actors are compelled to adopt the Born rule as the probability measure associated with their available actions. This approach is highly controversial, as some critics say the idea has circular logic.

Another attempt uses subjective probability [Vai98]. The experimenter puts on a blindfold before he finishes performing the experiment. After he finishes the experiment, he has uncertainty about what world he is in. This uncertainty is the foundation of a probability measure over the measurements. However, the actual form of the probability measure needs to be postulated:

Probability Postulate. *An observer should set his subjective probability of the outcome of a quantum experiment in proportion to the total measure of existence of all worlds with that outcome.*

Whichever explanation of the Born rule one adopts, the following section shows there is an issue with **MWT** and **IP**. There exist branches of substantial Born probability where information leaks occurs.

7.2 Violating the Independence Postulate

In [Lev84, Lev13], the Independence Postulate, **IP**, was introduced:

> *Let $\alpha \in \{0,1\}^{*\infty}$ be a sequence defined with an n-bit mathematical statement (e.g., in Peano Arithmetic or Set Theory), and a sequence $\beta \in \{0,1\}^{*\infty}$ can be located in the physical world with a k-bit instruction set (e.g., ip-address). Then $\mathbf{I}(\alpha : \beta) < k+n+c_{\mathrm{IP}}$, for some small absolute constant c_{IP}.*

The $\mathbf{I}$ term is an information measure in Algorithmic Information Theory. For this blog, the information term we use is $\mathbf{I}(x : y) = \mathbf{K}(x)+\mathbf{K}(y)-\mathbf{K}(x, y)$, where $\mathbf{K}$ is the prefix-free Kolmogorov complexity. We can use this definition of $\mathbf{I}$ because we only deal with finite sequences.

IP can be violated in the following idealized Stern-Gerlarch experiment measuring the spin $|\phi_\uparrow\rangle$ and $|\phi_\downarrow\rangle$ of N isolated electrons. We denote $|\phi_0\rangle$ for $|\phi_\uparrow\rangle$ and $|\phi_1\rangle$ for $|\phi_\downarrow\rangle$. The "address" (in the sense of **IP**) of this experiment is $< O(\log n)$. There is a measuring apparatus $\mathcal{A}$ with initial state of $|\psi^{\mathcal{A}}\rangle$, and after reading N spins of N electrons, it is in the state $|\psi^{\mathcal{A}}[x]\rangle$, where $x \in \{0,1\}^N$, whose ith bit is 1 iff the ith measurement returns $|\phi_1\rangle$. The experiment evolves according to the following unitary transformation:

$$\bigotimes_{i=1}^{N} |\phi\rangle \otimes |\psi^{\mathcal{A}}\rangle \otimes \stackrel{\text{unitary}}{\longrightarrow} \sum_{a_1,\ldots,a_N \in \{0,1\}^N} 2^{-N/2} \bigotimes_{i=1}^{N} |\phi_{a_i}\rangle \otimes |\psi^{\mathcal{A}}[a_1 a_2 \ldots a_n]\rangle\,.$$

If the bits returned are the first N bits of Chaitin's Omega, then a memory leak of size $n - O(\log n)$ has occurred. Thus

$$\text{Born-Probability(a memory leak of size } n - O(\log n) \text{ occurred)} \geq 2^{-n}.$$

7.3 Conclusion

There are multiple variations of **MWT** when it comes to consistency across universes. In one formulation, all universes conform to the same physical laws. In another model, each universe has its own laws, for example different values of gravity, etc. However, the experiment in the previous section shows that mathematics itself is different between universes, regardless of which model is used. In some universes, **IP** holds and there is no way to create information leaks. In other universes information leaks occur, and there are tasks where randomized algorithms fail but non-algorithmic physical methods succeeds. One such task is finding new axioms of mathematics. This was envisioned as a possibility by Gödel [Gö1], but there is a universal consensus of the impossibility of this task. Not any more! In addition, because information leaks are finite events, the Born probability of worlds containing them is not insignificant. In such worlds, **IP** cannot be formulated, and the the foundations of Algorithmic Information Theory itself become detached from reality.

Formulated another way, let us suppose the Born probability is derived from the probability postulate. We have a "blindfolded mathematician" who performs the experiment described above. Before the mathematician takes off her blindfold, she states the Independence Postulate. By the probability postulate, with measure 2^{-n} over all worlds, there is a memory leak of size $n - O(\log n)$ and **IP** statement by the mathematician is in error.

This is the crux of the argument. Even if there is a trivial resolution to this issue, I believe it is worth mentioning, as it relates two concepts that haven't been reconciled before.

8 October 20th, 2022: Two Resource Bounded EL Theorems

In my blog on October 9th, I mentioned that one open problem is a resource bounded version of the Sets Have Simple Members Theorem [Eps19c, Lev16]. In fact there are two such resource bounded EL theorems that can be derived almost directly from the literature. The first result is a corollary to Theorem 30 in [LOZ22]. The second result is a corollary to Theorem 4.1 in [AF09]. I'll also show how the first result can be applied to derandomization, in the sense of [Eps22d, Eps22a], to produce a resource bounded derandomization. The following definition is known as the probabilistic t-bounded Kolmogorov complexity.

8.1 Result One

Definition 11 ([LOZ22])

$$\mathrm{pK}^t(x) = \min\left\{k \;\middle|\; \Pr_{w\sim\{0,1\}^{t(|x|)}}\left[\exists \mathcal{M}\in\{0,1\}^k, \mathcal{M}(w) \text{ outputs } x \text{ within } t(|x|) \text{ steps}\right] \geq \frac{2}{3}\right\}.$$

Using this definition, the following theorem was proven.

Theorem 20 ([LOZ22]) *Suppose there is a randomized algorithm A for sampling strings such that $A(1^n)$ runs in time $T(n) \geq n$ and outputs a string $x \in \{0,1\}^n$ with probability at least $\delta > 0$. Then*

$$\mathrm{pK}^t(x) = \log(1/\delta) + O(\log T(n)),$$

where $t(n) = \mathrm{poly}(T(n))$ and the the constant $O(\cdot)$ depends on $|A|$ and is independent of the remaining parameters.

The advantage to the above theorem is that is unconditional, not requiring crypographic assumptions. If $k = \mathrm{pK}^t(x)$, if two parties share a typical random string w, x can be transmitted with k bits and decompressed in time $\mathrm{poly}(|x|)$. The proof of the above theorem can be readily extended to sets of strings. The above definition can be reformulated into the resources bounded complexity of sets.

Definition 12 (Probabilistic t-bounded Kolmogorov complexity(Sets))

$$\mathrm{pK}^t(D) = \min\left\{k : \Pr_{w\sim\{0,1\}^{t(|x|)}}\left[\exists \mathcal{M}\in\{0,1\}^k, \mathcal{M}(w) \text{ outputs } x\in D \text{ within } t(|x|) \text{ steps}\right] \geq \frac{2}{3}\right\}.$$

Corollary 2 *Suppose there is a randomized algorithm A for sampling strings such that $A(1^n)$ runs in time $T(n)$ and outputs a string $x \in D \subseteq \{0,1\}^n$ with probability at least $\delta > 0$. Then*

$$\mathrm{pK}^t(D) = \log(1/\delta) + O(\log T(n)),$$

where $t(n) = \mathrm{poly}(T(n))$ and the constants depends on $|A|$ and is independent of the rest of the parameters.

8.2 Result Two

Another avenue to explore can be found in [AF09], using the *t-time-bounded Kolmogorov complexity.*

Definition 13

$$\mathrm{K}^t(x) = \min_{\mathrm{TM},\mathcal{M}}\{|\mathcal{M}|+ : \mathcal{M} \text{ outputs } x \text{ in at most } t(|x|) \text{ steps.}\}$$

In [AF09], in the proof of Theorem 4,1, a coding theorem was used using resourced bounded complexity. The implications of this was shown in Theorem 20 of [LOZ22]. Let **Crypto** be the assumption that **E** is not contained in $\mathbf{DSPACE}(2^{\epsilon n})$ for some $\epsilon > 0$ and infinitely many n.

Theorem 21 ([AF09]) *Assume **Crypto**. Suppose there is an polynomial time algorithm A such that $A(1^n)$ outputs a string $x \in \{0,1\}^n$ with probability at least $\delta > 0$. Then for some polynomial p dependent on A,*

$$\mathrm{K}^p(x) \leq \log(1/\delta) + O(\log n).$$

To generalize to sets, an extra assumption needs to be made, because otherwise, let $D \subset \{0,1\}^n$ be all random strings of size n. We define the sampler A produce the uniform distribution over $\{0,1\}^n$. Thus $\min_{x \in D} \mathrm{K}^p(x) \leq -\log 1/2 + O(\log n)$, which is incorrect. The assumption made is that there exists a polynommial time algorithm that can test membership to the set D in question. The following corollary follows almost directly from the above theorem. Generally speaking, it states that sets that are efficently decidable and easily sampleable have efficiently compressible members. It remains to be seen if the **Crypto** assumption is needed.

Corollary 3 *Assume **Crypto**. Let $L \in \mathrm{P}$. Suppose there is a polynomial time algorithm A such that $A(1^n)$ outputs a member of L_n with probability $\delta_n > 0$. Then for some polynomial p,*

$$\min_{x \in L_n} \mathrm{K}^p(x) \leq \log(1/\delta_n) + O(\log n).$$

8.3 Resource Bounded Derandomization

Corollary 2 is immediately compatible with derandomization, in the sense of [Eps22d, Eps22a]. This proves the existence of a means to succinctly transmit solutions of instances of problems that can be efficiently decompressed.

This property can be described using a protocol between Alice and Bob. They share a typical random string of size poly(n). Alice has access to an instance of the Vertex Coloring, which is the number of colors k and an undirected graph $G = (V, E)$, $|V| = n$, of max degree $d < k/2$. The probability that a random coloring (under uniform randomness) is correct is $\geq (1 - d/k)^n$. Thus Alice gives Bob a string of size $2nd/k + O(\log nk)$, for which Bob can run in time poly(n) and produce a graph coloring with probability $2/3$. Put another way, using G defined earlier,

$$\mathrm{pK}^t(\{x : x \text{ is a } k \text{ coloring of } G\}) = 2nd/k + O(\log nk).$$

9 October 25th, 2022: A Theorem in Algorithmic Rate Distortion Theory

9.1 Classical Rate Distortion Theory

We provide a well known classical rate-distortion theory result and then prove one Algorithmic Information Theoretic version of the theorem. The source produces a sequence $X_1, X_2, \ldots, X_n$, i.i.d. $p(x)$, over the input alphabet $x \in \mathcal{X}$. The encoder is of the form $f_n(X^n) \in \{1, 2, \ldots, 2^{nR}\}$ and the decoder produces an estimate $\hat{X}^n \in \hat{\mathcal{X}}^n$. This is a $(2^{nr}, n)$ rate distortion code. A distortion function is $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$. The expected distortion is $D = \sum_{x^n} p(x^n) d(x^n, g_n(f_n(x^n)))$. A rate distortion pair (R, D) is said to be achievable if there exists a sequence of $(2^{nR}, n)$ distortion codes (f_n, g_n) with $\lim_{n\to\infty} \mathbf{E} d(X^n, g_n(f_n(X^n))) \leq D$. The rate distortion region for a source is the closure of the set of achievable rate distortion pairs (R, D).

Definition 14 (Rate Distortion Function) *The rate distortion function $R(D)$ is the infinimum of rates R such that (R, D) is in the rate distortion region of the source for a given distortion D.*

Definition 15 (Information Rate Distortion Function) *The information rate distortion function $R^{(I)}(D)$ for a source X with distortion function $d(x, \hat{x})$ is*

$$R^{(I)}(D) = \min_{p(\hat{x}|x):\sum_{(x,\hat{x})} p(x)p(\hat{x}|x)d(x,\hat{x}) \leq D} \mathbf{I}(X; \hat{X}).$$

Theorem 22 *The rate distortion function for an i.i.d. source X with distribution $p(x)$ and bounded distortion function $d(x, \hat{x})$ is equal to the associated information function.*

$$R(D) = R^{(I)}(D).$$

9.2 Distortion of Individual Codewords

This section contains a theorem reworking Theorem 2 in [VV10]. The difference is that we explicitly use a distortion function that is upper semi-computable, whereas in [VV10] it is generalized into distortion families. Furthermore the $O(\log n)$ error term in [VV10] is transformed into $\mathbf{I}(\cdot; \mathcal{H})$, where $\mathbf{I}(x; y) = \mathbf{K}(x) - \mathbf{K}(x|y)$, and $\mathcal{H}$ is the halting sequence. $\mathbf{K}$ is the prefix-free Kolmogorov complexity.

One algorithmic version of rate distortion theory is as follows. Alice wants to communicate a single message $\mathbf{y}$ to Bob, and they both share the same reference universal Turing machine U. Alice sends a program p to Bob, who decompresses it to a codeword $\mathbf{x} = U(p)$ and this message has distortion $d(\mathbf{x}, \mathbf{y})$. A distortion function $d : \{0,1\}^* \times \{0,1\}^* \to \mathbb{R}^{\infty}_{\geq 0}$ is a non-negative upper semi-computable function. Let $\mathbf{x}$ be a message sent to Bob. The following theorem shows that if $\mathbf{y}$ is non-exotic there exists a message $\mathbf{x}'$ such that $\mathbf{K}(\mathbf{x}') <^{\log} \mathbf{I}(\mathbf{x}; \mathbf{y})$, with $d(\mathbf{x}', \mathbf{y}) \leq d(\mathbf{x}, \mathbf{y})$.

Theorem 23 *Relativized to upper semi-computable distortion function d and $R \in \mathbb{R}_{>0}$,*

$$\min_{\mathbf{x}: d(\mathbf{x},\mathbf{y}) < R} \mathbf{K}(\mathbf{x}) <^{\log} \min_{\mathbf{x}: d(\mathbf{x},\mathbf{y}) < R} \mathbf{I}(\mathbf{x}; \mathbf{y}) + \mathbf{I}(\mathbf{y}; \mathcal{H}).$$

Proof. We assume the universal Turing machine U is left-total. For more details on left-total machines, the reader is referred to my October 11th blog post or [Eps22c]. We recall that relativization to elementary objects means that the universal Turing machine has access to their encodings on auxilliary tapes and the complexity terms implicity have the encoded objects in the conditional

terms. Let $D_\infty = \{\mathbf{x} : d(\mathbf{x}, \mathbf{y}) < R\}$ be the finite or infinite set of codewords that have distortion measure less than R with $\mathbf{y}$. The set D_∞ is enumerable given y, and for total string $b \in \{0,1\}^*$, let D_b be the finite subset of D_∞ enumerated in $\mathbf{bb}(b)$ steps. We recall the following busy beaver function on total b is

$$\mathbf{bb}(b) = \max\{\|x\| : U(p) = x, p \lhd b \text{ or } p \sqsupseteq b\}.$$

Let $i = 1 + \lceil -\log \mathbf{m}(D_\infty) \rceil$ and b be the shortest total string where $i \geq -\log \mathbf{m}(D_b)$. Arguments similar to those used in my October 11th blog post show $\mathbf{K}(b|\mathbf{y}, \|b\|) <^+ \mathbf{K}(i)$. Theorem 1 of my October 9th blog post, relativized to b results in $\mathbf{x}' \in \{0,1\}^*$, with

$$\mathbf{K}(\mathbf{x}'|b) <^{\log} i + \mathbf{Ks}(D_b|b).$$

Thes stochasticity function is

$$\mathbf{Ks}(a|b) = \min\{\mathbf{K}(P|b) + 3\log \mathbf{d}(a|P,b) : \ P \text{ is an elementary probability measure}\}.$$

An elementary probability measure has finite support and a range in $\mathbb{Q}_{\geq 0}$. The deficiency of randomness function is $\mathbf{d}(a|P,b) = \lfloor -\log P(a) \rfloor - \mathbf{K}(a|P,b)$. By Lemma 2 of my October 9th blog post, where $D = D_b$, relativized to b,

$$\mathbf{K}(\mathbf{x}'|b) <^{\log} i + \mathbf{I}(D_b; \mathcal{H}|b).$$

Using Lemma 2 from [Eps22d], which states $\mathbf{I}(f(a); \mathcal{H}) <^+ \mathbf{I}(a; \mathcal{H}) + \mathbf{K}(f)$,

$$\mathbf{K}(\mathbf{x}'|b) <^{\log} i + \mathbf{I}(\mathbf{y}; \mathcal{H}|b).$$

This is because given y and b, one can produce D_b. We can apply Lemma 4 to this equation, which results in

$$\begin{aligned}
\mathbf{K}(\mathbf{x}') &<^{\log} i + \mathbf{K}(b) + \mathbf{I}(\mathbf{y}; \mathcal{H}|b) \\
&<^{\log} i + \mathbf{I}(\mathbf{y}; \mathcal{H}) + \mathbf{K}(b|x, \|b\|) \\
&<^{\log} i + \mathbf{I}(\mathbf{y}; \mathcal{H}) + \mathbf{K}(i) \\
&<^{\log} i + \mathbf{I}(\mathbf{y}; \mathcal{H}). \qquad (1)
\end{aligned}$$

Let $\tau(x) = 2^{i-2}\mathbf{m}(x)[x \in D_\infty]$, where $[A] = 1$ if A is true, otherwise $[A] = 0$. This semi-measure is lower computable, and if $\mathbf{x} \in D_\infty$, then

$$\begin{aligned}
\mathbf{K}(\mathbf{x}|\mathbf{y}) &<^+ -\log \tau(\mathbf{x}) + \mathbf{K}(\tau|\mathbf{y}) \\
\mathbf{K}(\mathbf{x}|\mathbf{y}) &<^+ \mathbf{K}(\mathbf{x}) - i + \mathbf{K}(i) \\
i &<^{\log} \mathbf{I}(\mathbf{x}; \mathbf{y}). \qquad (2)
\end{aligned}$$

Combining Equations 1 and 2, results in the theorem statemebt, that is there exists a $\mathbf{x}' \in D_\infty$ where for all $\mathbf{x} \in D_\infty$,

$$\mathbf{K}(\mathbf{x}') <^{\log} \mathbf{I}(\mathbf{x}; \mathbf{y}) + \mathbf{I}(\mathbf{y}; \mathcal{H}).$$

□

Lemma 4 ([Eps21b]) *If b is total and b^- is not, then $\mathbf{I}(x; \mathcal{H}|b) + \mathbf{K}(b) <^{\log} \mathbf{I}(x; \mathcal{H}) + \mathbf{K}(b|x, \|b\|)$.*

10 October 27th, 2022: Two Modest Lemmas

This blog post contains two small lemmas that might be of independent interest. In general, the blog posting will slow down as I intend to write a survey over the material covered. I still intend to post blogs of papers of interest, but with a slower rate. As of today, the survey will contain the following contents.

1. Outliers in strings, sequences, and general spaces
2. Machine Learning and AIT
3. Clusters
4. Sets Have Simple Members Theorem
5. Resource Bounded EL Theorems
6. Derandomization (resourse free and resourse bounded) in particular its connection with Classical Information Theory and also parameterized instances
7. Quantum Information Theory, Many Worlds Theory

10.1 Computable Probability

In my October 11th blog post, I demonstrated the utility of so-called left-total machines. In this section, we show how to make an semi-computable semi-measure, $\mathbf{m}$, computable by using left-total machines. This enables a greater range of flexibility in proving results when $\mathbf{m}$ is computable, as shown in my September 28th blog post. Let $\mathbf{K}$ be the prefix-free Kolmogorov complexity. Let $\mathbf{I}(a;\mathcal{H}) = \mathbf{K}(a) - \mathbf{K}(a|\mathcal{H})$, where $\mathcal{H}$ is the halting sequence.

Definition 16 *For $D \subseteq \{0,1\}^*$, $\overline{\mathbf{m}}(D) = \min\{\mathbf{m}(P)P(D) : \text{probability } P \text{ is total computable}\}$.*

Lemma 5 $-\log \overline{\mathbf{m}}(D) <^{\log} -\log \mathbf{m}(D) + \mathbf{I}(D;\mathcal{H})$.

Proof. Let $(p0)^- = (p1)^- = p$. We define the following computable semi-measure, with $\mathbf{m}_b(x) = \sum\{2^{-\|p\|} : U(p) = x, p \lhd b \text{ or } p \sqsupseteq b\}$. If b and b^- are total then $\mathbf{m}_b(x) \leq \mathbf{m}_{b^-}(x)$. Let $s = \lceil -\log \mathbf{m}(D) \rceil + 1$. Let b be the shortest total string such that $\mathbf{m}_b(D) \geq 2^{-s}$. Thus b^- is not total. Thus $-\log \overline{\mathbf{m}}(D) <^+ -\log \mathbf{m}(\mathbf{m}_b)\mathbf{m}_b(D) <^+ s + \mathbf{K}(b)$. We show that $\mathbf{K}(b) <^{\log} \mathbf{I}(D;\mathcal{H}) + \mathbf{K}(s)$. From Lemma 2 in [Eps22d], we have that $\mathbf{I}(f(a);\mathcal{H}) <^+ \mathbf{I}(a;\mathcal{H}) + \mathbf{K}(f)$ and so $\mathbf{I}(b;\mathcal{H}) <^+ \mathbf{I}(D;\mathcal{H}) + \mathbf{K}(b|D)$. Now since b is total and b^- is not, b^- is a prefix of border, the binary expansion of Chaitin's Omega, and thus b is random. Furthermore b is simple relative to the halting sequence, with $\mathbf{K}(b|\mathcal{H}) <^+ \mathbf{K}(\|b\|)$. Thus $\mathbf{K}(b) <^{\log} \mathbf{I}(b;\mathcal{H})$. Now we prove that $\mathbf{K}(b|D) <^+ \mathbf{K}(\|b\|) + \mathbf{K}(s)$. There is an algorithm that can enumerate total strings of length $\|b\|$ and return the first string c such that $\mathbf{m}_c(D) \geq 2^{-s-1}$. This string is indeed b, as shown in Figure 3. □

10.2 Mutual Information with the Halting Sequence

The following lemma presents a non intuitive inequality about the mutual information with the halting sequence.

Lemma 6 $\mathbf{I}(x;\mathcal{H}/y) <^{\log} \mathbf{I}(\langle x,y\rangle;\mathcal{H})$.

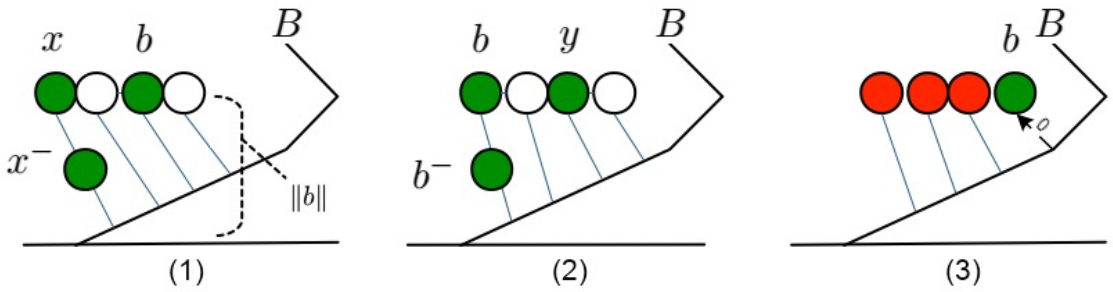

Figure 3: The above diagram represents the domain of the universal left-total Turing machine U with 0s branching to the left and 1s branching to the right. It shows all the total strings of length $\|b\|$, including b. The large diagonal line is the border sequence, B. A string c is marked green if $\mathbf{m}_c(D) \geq 2^{-s-1}$. By definition, b is a shortest green string. If x is green and total, and $x \lhd y$, and y is total, then y is green, since $\mathbf{bb}(x) \leq \mathbf{bb}(y)$. Furthermore, if x is green and total and x^- is total, then x^- is green, as $\mathbf{bb}(x) \leq \mathbf{bb}(x^-)$. It cannot be that there is a green $x \lhd b$ with $\|x\| = \|b\|$. Otherwise, x^- is total, and thus, it is green, causing a contradiction because it is shorter than b. This is shown in part (1). Furthermore, there cannot be a green y, with $b \lhd y$ and $\|y\| = \|b\|$. Otherwise, b^- is total and thus green, contradicting the definition of b. This is shown in part (2). Thus, b is unique, and since b^- is not total, b^- is a prefix of the border, as shown in part (3). Thus, an algorithm returning a green string of length $\|b\|$ will return b.

Proof.

$$\begin{aligned}\mathbf{I}(x;\mathcal{H}/y) &= \mathbf{K}(x/y) - \mathbf{K}(x/y,\mathcal{H}) \\ &<^+ \mathbf{K}(x,y) - \mathbf{K}(y) + \mathbf{K}(\mathbf{K}(y)/y) - \mathbf{K}(x/y,\mathcal{H}).\end{aligned}$$

Due to Theorem 3.3.1 in [Gác21], $\mathbf{K}(\mathbf{K}(y)/y) <^{\log} \mathbf{I}(y;\mathcal{H})$, so

$$\begin{aligned}\mathbf{I}(x;\mathcal{H}/y) &< \mathbf{K}(x.y) - \mathbf{K}(y) + \mathbf{I}(y;\mathcal{H}) - \mathbf{K}(x/y,\mathcal{H}) + O(\log \mathbf{I}(y;\mathcal{H})) \\ &< \mathbf{K}(x,y) - \mathbf{K}(y/\mathcal{H}) - \mathbf{K}(x/y,\mathcal{H}) + O(\log \mathbf{I}(y;\mathcal{H})) \\ &< \mathbf{K}(x,y) - \mathbf{K}(x,y/\mathcal{H}) + O(\log \mathbf{I}(y;\mathcal{H})) \\ &<^{\log} \mathbf{I}(\langle x,y\rangle;\mathcal{H}).\end{aligned}$$

The last inequality is due Lemma 2 in [Eps22d], which states that $\mathbf{I}(y;\mathcal{H}) <^+ \mathbf{I}(\langle x,y\rangle;\mathcal{H})$.

11 November 7th, 2022: Certificates and Inverting Hash Functions

This blog entry covers applications of a resource bounded version of the Sets Have Simple Members Theorem. Sufficient conditions for the efficient compression of proofs is given. We show that the pre-image of a simple element with respect to a hash function contains an element that is simple with respect to resource bounded Kolmogorov complexity. For my survey, I've decided to change it to a paper, focusing on the Sets Have Simple Members Theorem and its applications. More specifically, the application is derandomization in the context of bounded and unbounded resources. The contents are as followed.

1. New Proof to Sets Have Simple Members Theorem
2. Resource Bounded EL Theorem (From [AF09])
3. Resource Free Derandomization - Codebook Compression Size vs. Channel Capacity Tradeoff
4. Resource Bounded Derandomization - Certificate Compression

There are several variants of resouce bounded complexity. In the following definition, the running time of universal Turing machine is taken into account.

Definition 17 *For function t, the time-bounded Kolmogorov complexity is* $\mathbf{K}^t(x) = \min\{\|p\| : U(p) = x \text{ in } t(\|x\|) \text{ steps}\}$.

In my October 25th blog, I stated the following result, which follows almost directly from the proof of Theorem 4.1 in [AF09]. Let **Crypto** be the assumption that **E** is not contained in $\mathbf{DSPACE}(2^{\epsilon n})$ for some $\epsilon > 0$ and infinitely many n.

Theorem 24 *Assume **Crypto**. Let $L \in$ P. Suppose there is a polynomial time algorithm A such that $A(1^n)$ outputs a member of L_n with probability $\delta_n > 0$. Then for some polynomial p,* $\min_{x \in L_n} \mathbf{K}^p(x) < \log(1/\delta_n) + O(\log n)$.

Generally speaking, the following corollary states that if an efficiently generatable string has a good change of producing a random certificate with respect to an NP language, then the string has a simple proof.

Corollary 4 *Assume **Crypto**. Let $\{x_n\}$ be uniformly computable in polynomial time, where $\|x_n\| = n$. Fix a language in* NP. *There is a polynomial p where if a random proof for x_n has success rate γ_n,*

$$\min_{y \in \text{Proofs}(x_n)} \mathbf{K}^p(y) < -\log \gamma_n + O(\log \|y\|).$$

The following corollary shows that every efficiently computable sequence of strings have hash function pre-images that contain an efficiently compressible member.

Corollary 5 *Assume **Crypto**. Let $\{x_n\}$ be a uniformly polynomial time sequence and $\|x_n\| = n$. Let f be a polynomial time hash function, where $\|f(x)\| = \|x\| - k$. There is a polynomial p where for $D = f^{-1}(x_n)$,*

$$\min_{y \in D} \mathbf{K}^p(y) = n + k - \log |D| + O(\log(n + k)).$$

It would be of interest to see if the generatable sequence requirement in the above corollary could be removed. This would entail revisiting Theorem 4.1 in [AF09], and changing the sampling function to a hash function. If it is true it means that to invert x with f, one can find a secret key π of size approximately equal to x that efficiently expands to an element in $f^{-1}(x)$.

Conjecture 1 *Assume **Crypto**. Let f be a polynomial time function, where $f(\{0,1\}^n) \subseteq \{0,1\}^{n-k}$. There is a polynomial p where for $\{0,1\}^n \supseteq D = f^{-1}(x)$,*

$$\min_{y \in D} \mathbf{K}^p(y) = n - \log |D| + O(\log n).$$

The following corollary is a resource bounded version of SAT derandomization, which is Theorem 4 in [Eps22a]. This Theorem 4 uses Lovasz Local Lemma to achieve its bounds. The following corollary states that if simple SAT formulas have variables that do not appear in too many clauses, then they will admit efficiently compressible solutions.

Corollary 6 *Assume **Crypto**. Let Φ_n be a $k(n)$-SAT formula, using n variables, $m(n)$ clauses, uniformly polynomial time computable in n. Furthermore, each variable occurs in at most $2^{k(n)}/e-1$ clauses. There is a polynomial p where*

$$\min_{x \ satisfies \ \Phi_n} \mathbf{K}^p(x) < 2m(n)e2^{-k(n)} + O(\log n).$$

In fact many of the derandomization examples in [Eps22a] can be converted to resource bounded versions. So far, resource bounded games only derandomize to trivial examples. Thus it is an open question if given a non-trivial polynomial time environment, there is a player with low resource bounded Kolmogorov complexity that could do well against it. We show two more examples of resource bounded derandomization.

Corollary 7 *Assume **Crypto.** Let $\{G_n\}$ be a uniformly polynomial time computable sequence of k-regular graphs. There is a polynomial p where for each G_n, there is a partition x of $\lfloor \frac{k}{3 \ln k} \rfloor$ components each containing a cycle with complexity*

$$\mathbf{K}^p(x) < 2n/k^2 + O(\log n).$$

Corollary 8 *For vector v, $\|v\|_\infty = \max_i |v_i|$. A binary matrix M has entries of 0s or 1s. Let $\{M_n\}$ be a uniformly polynomial time computable sequence of $n \times n$ binary matrices. There is a polynomial p where for each M_n there is a vector $b \in \{-1,1\}^n$ such that $\|M_n b\|_\infty \leq 4\sqrt{n \ln n}$ and*

$$\mathbf{K}^p(b) = O(\log n).$$

12 November 8th, 2022: Conservation of Information

In [Lee06], four pillars of Kolmogorov Complexity were stated:

1. Coding Theorem.
2. Incompressibility.
3. Language Compression.
4. Symmetry of Information.

In my opinion, the fifth pillar should be conservation of information, which states that deterministic or randomized processing cannot increase target information. Thus for some partial computable function, A, and information term $\mathbf{I}$ over finite or infinite sequences, the deterministic non growth law is

$$\mathbf{I}(A(a) : b) <^{+} \mathbf{I}(a : b).$$

Given a probability γ over finite or infinite sequences computed by program p, the randomized non-growth law is

$$\mathbf{E}_{a\sim\gamma}[2^{\mathbf{I}((a,p):b)}] \stackrel{*}{<} 2^{\mathbf{I}(p:b)}.$$

Note that there several other formulations of randomized non-growth laws, some which use tests. These inequalities have been proven for several different information measures and it proves the impossibility of exotic situations such as accurately guessing bits of the halting sequence. Conservation of information is an essential component of derandomization, in the sense of [Eps22a].

In this blog I'll introduce a new information measure and show that this information term can be used to produce better bounds of the main result in [Lev13]. Before hand, I'll go over known information non growth laws. This is not an exhaustice list; there are several notions of information which I do not go over.

12.1 Symmetric Information over Strings

Symmetric information is defined for $x, y \in \{0,1\}^*$, is $\mathbf{I}(x : y) = \mathbf{K}(x) + \mathbf{K}(y) - \mathbf{K}(x, y)$, where $\mathbf{K}$ is the prefix-free Kolmogorov complexity. From [Lev84], one has the following non-growth laws.

Theorem 25

- *For partial computable* $f : \{0,1\}^* \to \{0,1\}^*$, $\mathbf{I}(f(a) : b) <^+ \mathbf{I}(a : b) + \mathbf{K}(f)$.
- *For probability* q *over* $\{0,1\}^*$ *computed by program* p, $\mathbf{E}_{a\sim q}[2^{\mathbf{I}((a,p):b)}] \stackrel{*}{<} 2^{\mathbf{I}(p:b)}$.

12.2 Mutual Information with the Halting Sequence

The amount of information that $x \in \{0,1\}^*$ has with the halting sequence $\mathcal{H} \in \{0,1\}^\infty$ is defined to be $\mathbf{I}(x;\mathcal{H}) = \mathbf{K}(x) - \mathbf{K}(x|\mathcal{H})$. From [Eps22d, Eps22c], the non-growth laws are as follows.

Theorem 26

- *For partial computable* $f : \{0,1\}^* \to \{0,1\}^*$, $\mathbf{I}(f(a);\mathcal{H}) <^+ \mathbf{I}(a;\mathcal{H}) + \mathbf{K}(f)$.
- *For probability* q *over* $\{0,1\}^*$ *computed by program* p, $\mathbf{E}_{a\sim q}[2^{\mathbf{I}((a,p);\mathcal{H})}] <^+ 2^{\mathbf{I}(a;\mathcal{H})}$.

12.3 Information over Infinite Sequences

The information between sequences $\alpha, \beta \in \{0,1\}^\infty$ is $\mathbf{I}(\alpha : \beta) = \log \sum_{x,y\in\{0,1\}^*} \mathbf{m}(x|\alpha)\mathbf{m}(y|\beta)2^{\mathbf{I}(x:y)}$, where $\mathbf{m}$ is the algorithmic probability [Lev74]. The non-growth laws (and later explicit proof) are due to [Lev74, Ver21].

Theorem 27

- *For partial computable* $f : \{0,1\}^\infty \to \{0,1\}^\infty$, $\mathbf{I}(f(\alpha) : \beta) <^+ \mathbf{I}(\alpha : \beta) + \mathbf{K}(f)$.
- *Let* $\gamma \in \{0,1\}^\infty$ *compute* $\mu(x\{0,1\}^\infty)$ *for some probability* μ *over* $\{0,1\}^\infty$*, over all* $x \in \{0,1\}^*$*. Then* $\mathbf{E}_{\alpha\sim\mu}\left[2^{\mathbf{I}((\alpha,\gamma):\beta)}\right] \stackrel{*}{<} 2^{\mathbf{I}(\gamma:\beta)}$.

12.4 New Information Term

This new term $\mathbf{I}(\alpha;\mathcal{H})$, measures the amount of information between a finite or infinite sequences $\alpha \in \{0,1\}^* \cup \{0,1\}^\infty$ and the halting sequence $\mathcal{H}$. This term is larger than the information term between infinite sequences defined earlier, I will show how it can be used achieve better bounds in theorems. We will prove information non-growth for deterministic processing.

A function $Q : \{0,1\}^* \to \mathbb{R}_{\geq 0}$ is a semi measure if $Q(\emptyset) \leq 1$ and $Q(x) \geq Q(x0) + Q(x1)$. We use $\mathbf{M}$ to denote a majorant (up to a multiplicative constant) lower semi-computable semi measure. For a prefix free set $D \subset \{0,1\}^*$, we have that $Q(D) = \sum_{x\in D} Q(x)$. Similarly, for an open set $S \subseteq \{0,1\}^\infty$, $Q(S) = Q(\{x : \Gamma_x \text{ is a maximal interval in } S\})$.

$\sqsubseteq$-sup is the supremum under the partial order of $\sqsubseteq$ on $\{0,1\}^* \cup \{0,1\}^\infty$. A function $\nu : \{0,1\}^* \to \{0,1\}^*$ is prefix-monotone iff for all $p, q \in \{0,1\}^*$, $\nu(p) \sqsubseteq \nu(pq)$. Then $\overline{\nu} : \{0,1\}^{*\infty} \to \{0,1\}^{*\infty}$ denotes the unique extension of ν, where $\overline{\nu}(p) = \sqsubseteq\text{-sup}\{\nu(p_{\leq n}) : n \leq \|p\|, n \in \mathbb{N}\}$ for all $p \in \{0,1\}^{*\infty}$. The set of all extensions $\overline{\nu}$, of prefix-monotone functions ν that are computable relative to $\alpha \in \{0,1\}^\infty$, is $\mathcal{D}^\alpha$. $\mathcal{D} = \mathcal{D}^\emptyset$. For $x \in \{0,1\}^*$, $\xi \in \mathcal{D}^\alpha$, $\xi^{-1}(x) = \{y : y \in \{0,1\}^*, x \sqsubseteq \xi(y), x \not\sqsubseteq \xi(y^-)\}$. For semi measure Q, $\xi \in \mathcal{D}^\alpha$, let $\xi Q(x) = Q(\xi^{-1}(x))$.

Let $(x0)^- = (x1)^- = x$. For semi measure Q, we say that $t : \{0,1\}^* \to \mathbb{R}_{\geq 0}$ is a Q test if for $x, y \in \{0,1\}^*$, $t(x) \leq t(xy)$, and for each $n \in \mathbb{N}$ where $D_{t,n} = \{x : t(x) > 2^n, t(x^-) \leq 2^n\}$,

$\sum_{x\in D_{t,n}} Q(x) < 2^{-n}$. The domain of tests are extended to infinite sequences $\alpha \in \{0,1\}^\infty$ by $t(\alpha) = \sup\{t(x) : x \sqsubset \alpha, x \in \{0,1\}^*\}$. For a group of tests T, we say $t \in T$ is majorant if for all $g \in T$, there exists $c \in \mathbb{R}_{>0}$ such that for all $x \in \{0,1\}^*$, $t(x) > cg(x)$.

Theorem 28 *For computable semi measure Q, there exists a majorant lower semicomputable Q test.*

This can be proved by enumerating all lower semicomputable Q tests and then summing them up in the standard way. Let $\mathbf{h}$ be a majorant, lower semicomputable relative to $\mathcal{H}$, $\mathbf{M}$ test.

Definition 18 *The term $\mathbf{I}(\alpha;\mathcal{H}) = \log \mathbf{h}(\alpha)$ represents the information $\mathcal{H}$ has about α.*

Note that for any r.e., relative to $\mathcal{H}$, a $\mathbf{M}$ test, t, has $\log t(\alpha) <^+ \mathbf{I}(\alpha;\mathcal{H}) + \mathbf{K}(t/\mathcal{H})$. For $A \in \mathcal{D}^{\mathcal{H}}$, let $\mathbf{h}_A$ be a majorant, lower semicomputable relative to $\mathcal{H}$, $A\mathbf{M}$ test.

Theorem 29 *For $A \in \mathcal{D}^{\mathcal{H}}$, $\mathbf{h}_A(A\cdot)$ is a lower semicomputable, relative to $\mathcal{H}$, $\mathbf{M}$ test.*

Proof. Otherwise there exists an $n \in \mathbb{N}$ such that $\sum_{x\in D_{\mathbf{h}_A(A\cdot),n}} \mathbf{M}(x) \geq 2^{-n}$. So for each $x \in D_{\mathbf{h}_A(A\cdot),n}$, $\mathbf{h}_A(Ax) > 2^n$ and $\mathbf{h}_A(Ax^-) \leq 2^n$. So

$$\begin{aligned}
&\sum_{x\in D_{\mathbf{h}_A,n}} A\mathbf{M}(x)\\
=&\sum_{x\in D_{\mathbf{h}_A,n}} \mathbf{M}(A^{-1}x)\\
=&\sum_{x\in D_{\mathbf{h}_A,n}} \sum_{x\sqsubseteq Ay, x\not\sqsubseteq Ay^-} \mathbf{M}(y)\\
=&\sum_{y:\mathbf{h}_A(Ay)>2^n, \mathbf{h}_A(Ay^-)\leq 2^n} \mathbf{M}(y)\\
=&\sum_{x\in D_{\mathbf{h}_A(A\cdot),n}} \mathbf{M}(x)\\
\geq& 2^{-n},
\end{aligned}$$

causing a contradiction, because $\mathbf{h}_A$ is a $A\mathbf{M}$ test.

Corollary 9 *For $A \in \mathcal{D}^{\mathcal{H}}$, $\mathbf{h}_A(A\alpha) <^+ \mathbf{h}(\alpha) + \mathbf{K}(A|\mathcal{H})$.*

Theorem 30 (Information Conservation) *For $A \in \mathcal{D}$ and $\alpha \in \{0,1\}^\infty$, $\mathbf{I}(A\alpha;\mathcal{H}) <^+ \mathbf{I}(\alpha;\mathcal{H}) + 3\mathbf{K}(A)$.*

Proof. Since $\mathbf{M} \overset{*}{>} \mathbf{m}(A)A\mathbf{M}$, $O(1)\mathbf{m}(A)\mathbf{h}(\alpha)$ is a lower semicomputable, relative to $\mathcal{H}$, $A\mathbf{M}$ test. So $\log 2^{-\mathbf{K}(A)}\mathbf{h}(\alpha) <^+ \log \mathbf{h}_A(\alpha) + \mathbf{K}(A)$. Putting this inequality and Corollary 9 together results in $\mathbf{I}(A\alpha;\mathcal{H}) <^+ \mathbf{I}(\alpha;\mathcal{H}) + 3\mathbf{K}(A)$.

12.5 New Bounds On Universal Partial Predicate Theorem

A partial predicate p is a finite or infinite set of pairs (x, b) consisting of indices $x \in \{0,1\}^*$ and bits $b \in \{0,1\}$. If $(x, b) \in p$, then we say $p(x) = b$, otherwise $p(x)$ is undefined. Let $\mathrm{Enc}(p) \in \{0,1\}^\infty$ be some standard encoding of partial predicate p. We use u to represent a universal partial recursive predicate, where for each partial recursive predicate p, there exists $z \in B^*$ such that $u(zx) = p(x)$, for all $x \in B^*$. The following theorem updates Theorem 1 from [Lev13], using the new information term $\mathbf{I}(\alpha; \mathcal{H})$, and achieves better bounds. The motivation for this theorem can also be found in [Lev13].

Theorem 31 *Let r be a partial predicate that on $\{0,1\}^n$ is a total extension of u.*
Then $\mathbf{I}(\mathrm{Enc}(r); \mathcal{H}) >^+ n$.

This theorem can be used to produce better bounds to Corollary 1 in [Lev13]. In a footnote of [Lev13] it is stated this bound can also be achieved by using the information term in [Lev84].

Corollary 10 *For randomized algorithm A, the probability that A computes on $\{0,1\}^n$ a total extension of u is at most $O(2^{-n})$.*

13 November 23rd, 2022: On Creating Pairs of Derandomization Theorems

I've uploaded a new paper to my site, titled *Derandomization under Different Resource Constraints* with the intention of eventually uploading it to arXiv and submission for publication.

http://www.jptheorygroup.org/doc/Resource.pdf

The main contribution is a resource bounded EL Theorem and a general formula for resource bounded derandomization, in the sense of [Eps22a]. In this blog, I show an example of taking a theorem produced from a non-constructive probabilistic proof and produce a pair of derandomization theorems, one that is resource free and one that is resource bounded. This methodology supports the following claim.

Claim. *If the existence of an object can be proven with the probabilistic method, then bounds on its Kolmogorov complexity can be proven as well.*

In my paper, I leveraged the fact that codebooks are defined probabilistically to prove a tradeof between their complexity and communication capacity. We show another example using hypergraphs. A *hypergraph* is a pair $J = (V, E)$ of vertices V and edges $E \subseteq \mathcal{P}(V)$. Thus each edge can connect ≥ 2 vertices. A hypergraph is *k-regular* of the size $|e| = k$ for all edges $e \in E$. A 2-regular hypergraph is just a simple graph. A valid *C-coloring* of a hypergraph (V, E) is a mapping $f : V \to \{1, \ldots, C\}$ where every edge $e \in E$ is not *monochromatic* $|\{f(v) : v \in e\}| > 1$. The following classic result [**?**] is the first proved consequence of Lovász Local Lemma.

Theorem. **[Probabilistic Method]** *Let $G = (V, E)$ be a k-regular hypergraph. If for each edge f, there are at most $2^{k-1}/e-1$ edges $h \in E$ such that $h \cap f \neq \emptyset$, then there exists a valid 2-coloring of G.*

We can now use derandomization, in the sense of [Eps22a], to produce bounds on the Kolmogorov complexity of the simpliest such 2-coloring of G.

Theorem. **[Derandomization]** *Let $G = (V, E)$ be a k-regular hypergraph with $|E| = m$. If, for each edge f, there are at most $2^{k-1}/e - 1$ edges $h \in E$ such that $h \cap f \neq \emptyset$, then there exists a valid 2-coloring x of G with*

$$\mathbf{K}(x) <^{\log} \mathbf{K}(n) + 4me/2^k + \mathbf{I}(G; \mathcal{H}).$$

The function $\mathbf{K}$ is the prefix free Kolmogorov complexity. $\mathbf{I}(G; \mathcal{H}) = \mathbf{K}(G) - \mathbf{K}(G|\mathcal{H})$ is the amount of asymmetric information the halting sequence $\mathcal{H} \in \{0, 1\}^\infty$ has about the graph G. We can now use resource derandomization, introduced in http://www.jptheorygroup.org/doc/Resource.pdf, to achieve bounds for the smallest time-bounded Kolmogorov complexity $\mathbf{K}^t(x) = \min\{p : U(p) = x \text{ in } t(\|x\|) \text{ steps}\}$ of a 2-coloring of G. **Crypto** is the assumption that there exists a language in $\mathbf{DTIME}(2^{O(n)})$ that does not have size $2^{o(n)}$ circuits with Σ_2^p gates.

Theorem. **[Resource Bounded Derandomization]** *Assume **Crypto**. Let $G_n = (V, E)$ be a $k(n)$-regular hypergraph where $|V| = n$ and $|E| = m(n)$, uniformly polynomial time computable in n. Furthermore, for each edge f in G_n there are at most $2^{k(n)-1}/e - 1$ edges $h \in E$ such that $h \cap f \neq \emptyset$. Then there is a polynomial p, and a valid 2-coloring x of G_n with*

$$\mathbf{K}^p(x) < 4m(n)e/2^{k(n)} + O(\log n).$$

The conjecture is that one can produce a suite of derandomization theorems, each one mapping to Kolmogorov complexity with different time and space constraints, and access to a certain number of random bits. In my uploaded paper, I used derandomization to show the tradeoff between codebook compression rate and channel capacity, so I believe there are a lot of applications of derandomization. However the codebook is of exponential size, so it is not suitable for resource-bounded derandomization. In [Eps22a], derandomization was used on games, where probabilistic players can be turned into winning deterministic ones. So far, resource bounded derandomization does not lend itself to games. This is because the environment must be polynomial time computable which means the agent can efficiently simulate it, making the results trivial.

14 December 15th, 2022: A Quantum EL Theorem

In this blog entry, we introduce a Quantum EL theorem: non exotic projections of large rank must have simple quantum pure states in their images. Simplicity is measured according to the classical information content of a pure state. It is similar to the definition in [Vit00] except a classical Turing machine is used instead of a quantum Turing machine.

Definition 19 (Complexity of a Quantum Pure State)
For n qubit state $|\phi\rangle$, $\mathbf{K}(|\phi\rangle |n) = \min\{\mathbf{K}(|\psi\rangle |n) - \log |\langle\phi\rangle \psi|^2 : |\psi\rangle$ is an elementary pure state$\}$.

Definition 20 (Computable Operators) *For computable operator A, $\mathbf{I}(A; \mathcal{H}|y) = \min\{\mathbf{K}(p|y) - \mathbf{K}(p|y, \mathcal{H}) : p$ is a program that computes $A\}$. $\mathcal{H} \in \{0,1\}^\infty$ is the halting sequence.*

Theorem 32 (Quantum EL Theorem) *Fix an n qubit Hilbert space. Let P be a computable projection of rank $> 2^m$. Then, $\min_{|\phi\rangle \in \text{Image}(P)} \mathbf{K}(|\phi\rangle |n) <^{\log} 3(n-m) + \mathbf{I}(P; \mathcal{H}|n)$.*

Corollary 11 *Fix an n qubit Hilbert space. Let ρ be a density matrix of rank $> 2^m$. Then, $\min_{|\phi\rangle \in \text{Image}(\rho)} \mathbf{K}(|\phi\rangle |n) <^{\log} 3(n-m) + \mathbf{I}(\rho; \mathcal{H}|n)$.*

The corollary is due to conservation of information. More specifically, if operator P is the projection onto the image of density matrix ρ, then $\mathbf{K}(P|\rho) = O(1)$ and also $\mathbf{I}(P; \mathcal{H}) <^+ \mathbf{I}(\rho; \mathcal{H})$. Thus the theorem applies to any quantum operator. This also applies to algorithmic quantum entropy [GÓ1] since it is less than $\mathbf{K}(|\phi\rangle)$. Another application is that if a quantum measurement can (even approximately) detect quantum algorithmic complexity, then it is exotic, in that it has high mutual information with the halting sequence. Put another way, if you receive a measurement, you cannot use it to infer the algorithmic complexity of the collapsed state. One note is that the theorem can be generalized to arbitrary (i.e. uncomputable) operators A. The error term is $\inf_{\langle A\rangle} \mathbf{I}(\langle A\rangle : \mathcal{H})$, where $\langle A\rangle$ is any appropriate infinite encoding of the operator and $\mathbf{I}$ is the mutual information term between infinite sequences.

A quantum source is a set of pure states $\{|a_i\rangle\}$ along with their probabiities $p(a_i)$. Let $\rho = \sum_i p(a_i) |a_i\rangle \langle a_i|$ be the density operator associated with the source. By Schumacher encoding, one can losslessly compress and communicate this source using $S(\rho)$ qubits, where S is the von Neumann entropy of ρ. In [GÓ1] the following remark was made.

Remark 2 ([GÓ1]) *Maybe the study of the problem for quantum description complexity helps with the understanding of the problem for von Neumann entropy, and its relation to coding tasks of quantum information theory.*

The theorem displayed in this blog helps address this remark.

Claim 1 *As the von Neumann entropy associated with the quantum source increases, the lossless quantum coding projectors have larger rank and thus must have simpler (in the algorithmic quantum complexity sense) pure states in their images.*

This blog post is another example of a result in the intersection of AIT and physics. Future work in algorithmic physics involves proving that algorithmic thermodynamic entropy must oscillate in the presence of dynamics. Another avenue of future work is conservation of algorithmic randomness and information with respect to the most general transformation of a qubit, a quantum operation.

15 December 24th, 2022: Algorithmic Thermodynamic Entropy

For definitions in this post, we use [HR09]. A computable metric space X is a metric space with a dense set of ideal points on which the distance function is computable. A computable probability is defined by a computable sequence of converging points in the corresponding space of Borel probability measures, $\mathcal{M}(X)$, over X. For measure μ, a μ-test is a μ-constructive lower computable function f, such that $\int_X f d\mu \leq 1$. There exists a universal test, $\mathbf{t}_\mu$. We extend the result from [Eps22d] to computable metric spaces.

Theorem 33 *Given computable non-atomic probability measures μ and λ over a computable metric space X and universal test $\mathbf{t}_\mu(\cdot)$. For all n, $\lambda(\{\alpha : \mathbf{t}_\mu(\alpha) > 2^n\}) \overset{*}{>} 2^{-n-\mathbf{K}(n)}$.*

Reworking the above theorem, one can get a result in algorithmic physics. To define algorithmic fine-grain entropy, we use a slightly modified version of the definition in [Gac94], and I refer to that paper for the motivation of the definition. First, note that all the results [HR09] can be easily extended to arbitrary nonnegative measures which are used to represent volume in the space. This can be achieved by defining the product space of $\mathcal{M}(X)$ and $\mathbb{R}_{\geq 0}$, where the second metric space defines the size of the measure. We also include the special case of $X = \mathbb{N}$, with the counting measure $\#$. This means we don't have to define a space of measures which $\#$ is a member of. Given a measure $\mu \in \mathcal{M}(X) \times \mathbb{R}_{\geq 0}$, the algorithmic fine grained entropy of a point $\alpha \in X$ is as follows.

Definition 21 (Algorithmic Fine-Grained Entropy) $\mathbf{H}_\mu(\alpha) = -\log \mathbf{t}_\mu(\alpha)$.

One can then prove that this term will oscillate in the presence of dynamics. Dynamics can be defined using group theory.

Definition 22 (Transformation Group) *Let M denote a computable metric space and G a topological group each element of which is a homeomorphism of M onto itself:*

$$f(g; x) = g(x) = x' \in M; g \in G, x \in M.$$

The pair (G, M) will be called a topological transformation group if for every pair of elements $g1, g2$ of G, and every $x \in M$, $g1(g2(x)) = (g1g2)(x)$ and if

$$x' = g(x) = f(g; x)$$

is continuous simultaneously in $x \in M$ and $g \in G$.

We will be restricting dynamics to one dimensional transformation groups G^t, which is either continuous with $t \in \mathbb{R}$ or discrete, with $t \in \mathbb{I}$. Also the transform is μ-measure preserving over all measurable subsets of X.

Theorem 34 (Oscillation of Thermodynamic Entropy) *Let L be the Lebesgue measure over $\mathbb{R}$. For continuous transformation group (G^t, X) acting on computable metric space X, for all $\alpha \in X$, $L\{t \in [0, 1] : \mathbf{H}_\mu(G^t\alpha) < \log \mu(X) - n\} \overset{*}{>} 2^{-n-\mathbf{K}(n)}$.*

Obviously, the above theorem does not hold in the case of static dynamics. I also will include some other secondary results that will hopefully round out the forthcoming paper. The Stability Theorem 5 in [Gac94] can be updated with the integration results in [HR09]. Let $\Pi(\cdot)$ be a set of disjoint uniformly enumerable open sets in the metric space X.

Definition 23 (Algorithmic Coarse Grained Entropy) $\mathbf{H}_\mu(\Pi_i) = \mathbf{K}(i|\mu) + \log \mu(\Pi_i)$.

Coarse grained entropy is an excellent approximation of fine grained entropy, as shown by the following two results.

Proposition 3 *If $\mu(\Pi_i)$ is uniformly computable and $\alpha \in \Pi_i$ then* $\mathbf{H}_\mu(\alpha) <^+ \mathbf{H}_\mu(\Pi_i) + \mathbf{K}(\Pi)$.

Lemma 7 (Stability) $\mu\{\alpha \in \Pi_i : \mathbf{H}_\mu(\alpha) < \mathbf{H}_\mu(\Pi_i) - \mathbf{K}(\Pi) - m\} \stackrel{*}{<} 2^{-m}\mu(\Pi_i)$.

We revisit an rather interesting result in [Gac94] that translates directly with the slightly new definitions of this blog entry. It states that if dynamics are used to increase or decrease algorithmic thermodynamic entropy by a non trivial amount, then the encoded dynamics shares algorithmic information with the ending or starting state, respectively. Put another way,

> *if you want to increase the entropy of a state, you need information about its ending state and if you want to decrease the entropy of a state, you need information about its starting state.*

The following definition introduces information between a point $\alpha \in X$ of the metric space and a sequence t. The term $\mathbf{H}_\mu(\alpha|t)$ is the fine grained algorithmic entropy of α when the universal Turing machine is relativized to the sequence t.

Definition 24 (Information) *For $\alpha \in X_1$, $\beta \in X_2$ and $t \in \{0,1\}^* \cup \{0,1\}^\infty$,*

- $\mathbf{I}(\alpha; t) = \mathbf{H}_\mu(\alpha) - \mathbf{H}_\mu(\alpha|t)$.
- $\mathbf{I}(\alpha : \beta) = \mathbf{H}_{\mu_1}(\alpha) + \mathbf{H}_{\mu_2}(\beta) - \mathbf{H}_{\mu_1 \times \mu_2}((\alpha, \beta))$.

Proposition 4 ([Gac94]) $-\mathbf{I}(\alpha; \langle t \rangle) <^+ \mathbf{H}_\mu(G^t\alpha) - \mathbf{H}_\mu(\alpha) <^+ \mathbf{I}(G^t\alpha; \langle t \rangle)$.

Proposition 5 (Conservation of Information) $\mathbf{I}(G^t\alpha : \beta) <^+ \mathbf{I}(\alpha : \beta) + 2\mathbf{K}(t)$.

We revisit Maxwell's demon, providing yet another interpretation. This is done by reworking the Entropy Balance Theorem 9 in [Gac94]to the specific case of finite sequences. Let X be a computable metric space and μ its corresponding computable measure. We use the finite space of $\{0,1\}^*$, finite sequences, and we use the counting measure, $\#$, with it. Thus if $x \in \{0,1\}^*$, $\mathbf{H}_\#(x) =^+ \mathbf{K}(x)$. For a starting point $\alpha \in X$, we couple it with an empty sequence 0, with discrete dynamics G^t producing $(x, \alpha') = G^1(0, \alpha)$.

Proposition 6 (Maxwell's Demon) $\mathbf{H}_\mu(\alpha) - \mathbf{H}_\mu(\alpha') <^+ \mathbf{K}(x)$.

Thus after α decreases in thermodynamic entropy, the contents of the register fills up. This shows that one benefit of an algorithmic formulation of thermodynamics is that pure algorithmic information and thermodynamic entropy are interchangeable. The original theorem is more general, and is over general spaces rather than sequences, and is over arbitrary time. Thus, putting Propositions 4 and 6 together, if one wants to lower the thermodynamic entropy of a state, the information of the state must be encoded into the dynamics or an independent environment can be coupled with the system which will absorb the entropy.

Its interesting to note that the proof for the above theorem follows from first using a combinatorial argument about finite strings and then applying this result to prove a property of randomness deficiencies of sequences and then transfering this result to universal uniform tests and then finally

algorithmic thermodynamic entropy. An open question is whether other such transfers can be proven, resulting in further characterization of $\mathbf{H}_\mu$.

I have one more result in physics, which is conservation of algorithmic quantum randomness deficiency and information with respect to quantum operations. All these results will go into two papers: one containing the Quantum EL Theorem and the conservation inequalities, and another containing all the results of AIT and thermodynamics. The goal is make headway into the intersection of AIT and physics, dubbed *algorithmic physics*. Another area to look into would be the connection of AIT with special and general relativity, and black hole entropy.

16 December 24th, 2022: On The Curious Lack of Algorithmic Information In Quantum States

For classical algorithmic information theory, random strings have a high amount of self information, with $\mathbf{K}(x) =^+ \mathbf{I}(x : x)$. We can generalize from strings to arbitrary signals, formalized by probability measure over strings.

Definition 25 (Information, Signals)
For semi-measures p and q over $\{0,1\}^*$, $\mathbf{I}_P(p : q) = \log \sum_{x,y \in \{0,1\}^*} 2^{\mathbf{I}(x:y)} p(x)q(y)$.

As shown in http://www.jptheorygroup.org/doc/InfoProb.pdf, this measure observes conservation inequalities over deteministic or randomnized processing. Thus processing cannot increase information between two signals. In addition, information of probabilities can be extended to infinite sequences or general spaces. If the the probability measure is concentrated at a single point, then it contains self-information equal to the complexity of that point. If the probability measure is spread out, then it is white noise, and contains no self-information. Some examples are as follows.

Example 1

- *In general, a probability p, will have low $\mathbf{I}_P(p : p)$ if it has large measure on simple strings, or low measure on a large number of complex strings, or some combination of the two.*
- *If probability p is concentrated on a single string x, then $\mathbf{I}_P(p : p) = \mathbf{K}(x)$.*
- *The uniform distribution U_n over strings of length n has self information equal to (up to an additive constant) $\mathbf{K}(n)$.*
- *There are semi-measures that have infinite self information. Let α_n be the n bit prefix of a Martin Löf random sequence α and $n \in [2, \infty)$. Semi-measure $p(x) = [x = \alpha_n]n^{-2}$ has $\mathbf{I}_P(p : p) = \infty$.*
- *The universal semi-measure $\mathbf{m}$ has no self information.*

This blog explains a curious fact: most quantum states have negligible algorithmic self-information and given a measurement, the overwhelming majority of pure quantum states will produce random noise. For algorithmic information $\mathbf{I}_Q$ between quantum states we refer the reader to the definition $\mathbf{I}$ in [Eps19c]. The following theorem shows that self information of states is negligible.

Given a quantum state $|\psi\rangle$, a measurement, or POVM, E produces a probability measure $E|\psi\rangle$ over strings. This probability represents the classical information, or *signal* produced from the measurement. We refer readers to https://en.wikipedia.org/wiki/POVM for an introduction to quantum measurements. Given a measurement E, for an overwhelming majority of quantum states $|\psi\rangle$, the signal (probability) produced will be white noise or the empty signal, i.e. have no meaningful information, i.e. $\mathbf{I}_P(E|\psi\rangle : E|\psi\rangle)$ is negligible.

Theorem 35 ([Eps19c]) *Let Λ be the uniform distribution on the n qubit space.*

- $\int 2^{\mathbf{I}_Q(|\psi\rangle : |\psi\rangle)} d\Lambda = O(1)$.
- *Relativized to POVM E, $\int 2^{\mathbf{I}_P(E|\psi\rangle : E|\psi\rangle)} d\Lambda = O(1)$.*

This is result is in constrast to the fact that most pure quantum states will have a very large *algorithmic quantum entropy*, using any definition from [GÓ1, Vit00, **?**]. Thus most quantum pure states contain high quantum algorithmic entropy, low self algorithmic information, and will most likely produce random noise (or the empty signal) given a quantum measurement. One interesting note is that the measurement theorem is derived from the strangest proof I've ever written, leveraging *upper computable* tests!

17 December 27th, 2022: Conservation Inequalities over Quantum Operations

In [Eps19b], algorithmic notions of randomness and information between two quantum mixed states were introduced. These notions were shown to satisfy conservation inequalities with respect to unitary transforms and partial traces. This blog entry generalizes these results by proving conservation of randomness and information with respect to quantum operations. Quantum operations model not only reversible unitary transforms of isolated systems, but also transient interactions with the environment and the effects of measurements. Thus, quantum operations are the most general physically realizable transform that can be applied to a quantum state. We show a computable quantum operation cannot increase the deficiency of randomness of one state with respect to another. Similarly, a quantum operation cannot increase the algorithmic mutual information shared between two states.

18 Conventions

We use $\mathcal{H}_n$ to denote a Hilbert space with n dimensions, spanned by bases $|\beta_1\rangle, \ldots, |\beta_n\rangle$. A qubit is a unit vector in the Hilbert space $\mathcal{G} = \mathcal{H}_2$, spanned by vectors $|0\rangle$, $|1\rangle$. To model n qubits, we use a unit vector in $\mathcal{H}_{2^n}$, spanned by basis vectors $|x\rangle$, where x is a string of size n.

A pure quantum state $|\psi\rangle$ of length n is a unit vector in $\mathcal{H}_{2^n}$. Its corresponding element in the dual space is denoted by $\langle\phi|$. The conjugate transpose of a a matrix A is A^*. The tensor product of two matrices A and B is $A \otimes B$. Tr is used to denote the trace of a matrix, and for Hilbert space $\mathcal{H}_X \otimes \mathcal{Y}$, the partial trace with respect to Y is Tr_Y.

For positive semi-definite matrices A and B, we say $B \preceq A$, iff $A - B$ is positive semi-definite. For functions f whose range are Hermitian matrices, we use $\overset{*}{<} f$ and $\overset{*}{>} f$ to denote $\preceq f/O(1)$ and $\succeq f/O(1)$. We use $\overset{*}{=} f$ to denote $\overset{*}{<} f$ and $\overset{*}{>} f$.

Density matrices are used to represent mixed states, and are self-adjoint, positive definite matrices with trace equal to 1. Semi-density matrices are used in this paper, and they are density matrices except they may have a trace in [0,1].

Pure and mixed quantum states are elementary if their values are complex numbers with rational coefficients, and thus they can be represented with finite strings. Thus elementary quantum states $|\phi\rangle$ and ρ can be enncoded as strings, $\langle|\phi\rangle\rangle$ and $\langle\rho\rangle$, and assigned Kolmogorov complexities $\mathbf{K}(|\phi\rangle)$, $\mathbf{K}(\rho)$ and algorithmic probabilities $\mathbf{m}(|\phi\rangle)$ and $\mathbf{m}(\rho)$. They are equal to the complexity (and algorithmic probability) of the strings that encodes the states.

More generally, a complex matrix A is elementary if its entries are complex numbers with rational coefficients and can be encoded as $\langle A\rangle$, and has a Kolmogorov complexity $\mathbf{K}(A)$ and algorithmic probability $\mathbf{m}(A)$.

In [GÓ1], a universal lower computable semi-density matrix, $\boldsymbol{\mu}$ was introduced. It is the quantum analogy to $\mathbf{m}$. It can be defined (up to a multiplicative constant) by

$$\boldsymbol{\mu}_{/x} = \sum_{\text{elementary } |\phi\rangle} \mathbf{m}(|\phi\rangle\,|x,n)\,|\phi\rangle\,\langle\phi|\,,$$

where the summation is over all n qubit elementary pure quantum states. We use $\boldsymbol{\mu}$ to denote $\boldsymbol{\mu}_{/\emptyset}$.

A matrix is computable if its entries can be computed to any degree of precision. We say a semi-density matrix ρ is lower computable if there a program $p \in \{0,1\}^*$ such that when given to the universal Turing machine U, outputs, with or without halting, a finite or infinite sequence

of elementary matrices ρ_i such that $\rho_i \preceq \rho_{i+1}$ and $\lim_{i\to\infty} \rho_i = \rho$. If U reads $\leq \|p\|$ bits on the input tape, then we say p lower computes ρ. From [GÓ1] Theorem 2, if q lower computes ρ, when $\mathbf{m}(q|n)\rho \overset{*}{<} \boldsymbol{\mu}$.

18.1 Quantum Operations

A map transforming a quantum state σ to $\varepsilon(\sigma)$ is a quantum operation if it satisfies the following three requirements

1. The map of ε is positive and trace preserving, with $\mathrm{Tr}(\sigma) = \mathrm{Tr}(\varepsilon(\sigma))$.
2. The map is linear with $\varepsilon(\sum_i p_i\sigma_i) = \sum_i p_i\varepsilon(\sigma_i)$.
3. The map is completely positive, were any map of the form $\varepsilon \otimes \mathbf{1}$ acting on the extended Hilbert space is also positive.

The operator $\mathbf{1}$ is the identity matrix. Another means to describe quantum operations is through a series of operators. A quantum operation ε on mixed state σ_A can be seen as the appending of an ancillia state σ_b, applying a unitary transform U, then tracing out the ancillia system with

$$\varepsilon(\sigma_A) = \mathrm{Tr}_B\left(U(\sigma_A \otimes \sigma_B)U^*\right). \tag{3}$$

A third way to characterize a quantum operation is through Kraus operators, which can be derived using an algebraic reworking of Equation 3. Map ε is a quantum operation iff it can be represented (not necessarily uniquely) using a set of matrices $\{M_i\}$ where $\varepsilon(\sigma) = \sum_i M_i\varepsilon M_i^*$ and $\sum_i M_i^* M_i \leq \mathbf{1}$.

A quantum operation ε is computable if it admits a represented of the form in Equation 3 where B, U, and σ_B are each computable, in that they each can be computable to arbtirary precision with a program. Each computable quantum operation admits an computable Kraus operator representation $\{M_i\}$, in that each M_i is an computable matrix.

18.2 Conservation of Randomness and Information

In [GÓ1], the deficiency of randomness of a mixed state σ with respect to computable mixed state ρ was introduced. A positive semi-definite matrix ν is a ρ-test if it is lower computable and $\mathrm{Tr}\rho\nu \leq 1$. Since ρ is computable, the set of ρ-tests, $\{\nu_i\}$, is enumerable. Thus the deficiency of randomness of σ with respect to ρ was defined to be $\mathrm{Tr}\sigma\sum_i \mathbf{m}(i)\nu_i$. Like the classical variant, this measured the level of typicality of σ with respect to ρ.

In [Eps19b], the deficiency of a randomness of a mixed state σ with respect to an arbitrary (not necessarily computable) matrix ρ was introduced. Like [GÓ1], ν is a ρ-test, $\nu \in \mathcal{T}_\rho$, if it is positive semi-definite and lower computable and $\mathrm{Tr}\rho\nu \leq 1$. The lower probability of a lower computable mixed state was defined, with $\underline{\mathbf{m}}(\nu|x) = \sum\{\mathbf{m}(q|x) : q \text{ lower computes } \nu\}$. The deficiency of randomness of σ with respect to ρ is defined as follows.

Definition 26 ([Eps19b])
For n qubit semi-density matrices σ and ρ, $\mathbf{d}(\sigma|\rho) = \log \mathrm{Tr}\sigma\sum_{\nu\in\mathcal{T}_\rho} \underline{\mathbf{m}}(\nu|n)\nu$.

In [Eps19b], the algorithmic information of two mixed states σ and ρ was introduced, using notions of quantum tests seen in the deficiency of randomness definition. Let $\mathcal{C}_{C\otimes D}$ be the set of all lower computable matrices of the form $A \otimes B$, where $\mathrm{Tr}(A \otimes B)(C \otimes D) \leq 1$. Let $\mathfrak{C}_{C\otimes D} = \sum_{A\otimes B\in\mathcal{C}_{C\otimes D}} \underline{\mathbf{m}}(A \otimes B|n)A \otimes B$ be an aggregation of $C \otimes D$ tests of the form $A \otimes B$, weighted by their lower probability. Using $\mathfrak{C}$, we get the following definition of information.

Definition 27 ([Eps19b])
For semi-density matrices σ and ρ, $\mathbf{I}(\sigma : \rho) = \log \mathrm{Tr}\mathfrak{C}_{\boldsymbol{\mu}\otimes\boldsymbol{\mu}}(\sigma \otimes \rho)$.

The following theorem shows conservation of randomness with respect to elementary quantum operations. It generalizes Theorems 2 and 3 from [Eps19b].

Theorem 36 (Randomness Conservation) *Relativized to computable quantum operation ε, for semi-density matrices ρ, σ, $\mathbf{d}(\varepsilon(\rho)|\varepsilon(\sigma)) <^+ \mathbf{d}(\rho|\sigma)$.*

The following theorem shows information nongrowth with respect to elementary quantum operations. It generalizes Theorems 5 and 10 from [Eps19b].

Theorem 37 (Information Conservation) *Relativized to computable quantum operation ε, for semi-density matrices ρ, σ, $\mathbf{I}(\varepsilon(\rho) : \sigma) <^+ \mathbf{I}(\rho : \sigma)$.*

19 December 28th, 2022: A New Proof to the Outliers Theorem

In this blog entry, I present a new proof to the Outliers Theorem in [Eps21b]. The proof to this theorem is similar to the proof of Proposition 1 in [Epsb] which is analogously similar to the proof in [Lev16]. It is different from the proof in [Lev16] in that it doesn't use left-total machines. It implies that large sets of strings with low deficiency of randomness will be non-stochastic, which implies that they have high mutual information with the halting sequence. In fact this works for any computable pairing function, not just probability measures. There are more direct proofs that show that algorithms must produce sets which contain elements that have high randomness deficiencies. These proofs can be found in [Eps22d] and achieve better bounds than if you used stochasticity theorems to produce the algorithmic no-go theorems. However the theorem in this blog and the theorems in [Eps21b] are beneficial because one can use conservation properties of the mutual information function to prove new and interesting facts. This can be seen in [Epsa] which contains a result that thermodynamic entropy must oscillate in the presence of dynamics. The proof to this theorem uses properties of the mutual information with the halting sequence.

The hope was that the new proof in [Epsb] could lead to a Resource Bounded EL Theorem. However it turns out that such a theorem basically already existed in the literature, which can be found in [AF09]. The results in this paper could be quickly reworked into a Resource EL Theorem, with applications in Resource Bounded Derandomization [**?**]. However it could be that the proof in this blog entry and [Epsb] will have other interesting applications.

New Proof

A probability is *elementary*, if it has finite support and rational values. The deficiency of randomness of x relative to a elementary probability measure Q is $\mathbf{d}(x|Q) = -\log Q(x) - \mathbf{K}(x|Q)$.

Definition 28 (Stochasticisty) *A string x is (α, β)-stochastic if there exists an elementary probability measure Q such that*

$$\mathbf{K}(Q) \leq \alpha \text{ and } \mathbf{d}(x|Q) \leq \beta.$$

Theorem 38 *Let W be a positive computable function on strings and be c be a large constant. If D is (α, β)-stochastic relative to s and $\sum_{x \in D} \mathbf{m}(x)/W(x) \geq 2^s$ then there exists $x \in D$ with*

$$\log \frac{\mathbf{m}(x)}{W(x)} > s - \alpha - \log \beta - \mathbf{K}(\log \beta, s) - c.$$

This implies a slightly weaker form of Theorem 1 in [Epsb], which is when $W(a) = 1$.

Proof. Let Q be a probability measure that realizes the stochasticity of D. Assume β is large. Let $\gamma = 2^{s-1}$ and $f(x) = \mathbf{m}(x)/W(x)$. Let f_* be a function on strings and $f_1, f_2, \ldots$ be a finite or infinite series of function on strings such that

- In the infinite case, we have $f_* = 0$.
- The sequence $f_1, f_2, \ldots$ can be uniformly computed given a program for W.
- All functions are nonnegative, and each function f_1, f_2 is non-zero for finitely many strings.
- $\sum_x f_i(x) = \gamma/\beta$ for all integers i and $\sum_x f_*(x) \leq \gamma/\beta$.

- $f = f_* + \sum_i f_i$.

We consider the infinite case. The finite case is similar.

Construction of a lower-semicomputable probability bounded test g on sets X of strings. Let $g_0 = 1$. We construct g_1, g_2, together with a sequence $z_1, z_2, \ldots$ Assume that for some $i \geq 0$, we already defined g_i and $z_1, \ldots, z_i$. Let z_{i+1} be chosen such that the function

$$g_{i+1}(X) = \begin{cases} g_i(X) & \text{if } g_i(X) \geq \exp(\beta) \\ g_i(X)\exp(\frac{\beta}{\gamma} f_i(X)) & \text{if } g_i(X) < \exp(\beta) \text{ and } X \text{ is disjoint from } \{z_1, \ldots, z_i\} \\ 0 & \text{otherwise.} \end{cases}$$

satisfies

$$\frac{\gamma}{\beta} W(z_i) + \mathbb{E}_{X\sim Q} g_i(X) <^+ \sum_x f_i(x) W(x) + \mathbb{E}_{X\sim Q} g_{i-1}(X). \tag{4}$$

Let $g(X) = \exp(\beta)$ if some $g_i(X) \geq \exp(\beta)$ and 0 otherwise. *End of construction.*

We first prove that there always exists a z_i that satisfies the condition. Assume we select z_i randomly with probability $\frac{\beta}{\gamma} f_i(x)$. This is possible because these probabilities sum up to one by the definition of f_i. We show that the expected value of the left hand side in 4 equals the right hand side. Indeed $\mathbb{E}W(x) = \frac{\beta}{\gamma} \sum_x f_i(x) W(x)$. We also have

$$\begin{aligned} \mathbb{E}_{z_i \sim (\beta/\gamma) f_i} g_i(X) &\leq (1 - (\beta/\gamma) f_i(X)) \cdot g_{i-1}(X) \exp((\beta/\gamma) f_i(X)) \leq g_{i-1}(X) \\ \mathbb{E}_{z_i \sim (\beta/\gamma) f_i, X\sim Q} g_i(X) &\leq \mathbb{E}_{X\sim Q} g_{i-1}(X). \end{aligned}$$

Hence the required z_i exists. We sum Equation 4 for all i, and obtain

$$\frac{\gamma}{\beta} \sum_i W(z_i) + \lim_i \mathbb{E}_{X\sim Q} g_i(X) \leq \sum_x f(x) W(x) + \mathbb{E}_{X\sim Q} g_0(X) \leq 1 + 1.$$

This implies that $\sum_i W(z_i) \leq 2\frac{\beta}{\gamma}$ and $\mathbb{E}g \leq 2$. Hence $g/2$ is an expection bounded test and $\mathbf{d}(X|\beta,\gamma) >^+ \log g(X)$. Note that if there is some set X with $\sum_{x\in X} f(x) \geq 2^s = 2\gamma$ and X does not contain any element z_i, then $g_i(X) = \exp \sum_{j=1}^i \frac{\beta}{\gamma} f_j(X)$ and thus there is some i where $g_i(X) > \exp\beta$. Thus $g(X) = \exp\beta$ and hence X is not (α,β)-stochastic relative to s. So

$$1.44\beta \leq \log g(X) \leq \mathbf{d}(X|\beta,\gamma) \leq \mathbf{d}(X|s) + 2\log\beta.$$

Thus any D that satisfies the conditions of the theorem (i.e. has large β) must contain some element z_i, otherwise $1.44\beta \leq \mathbf{d}(D|s) + 2\log\beta <^{\log} \beta$, causing a contradiction. Since $\sum_i \frac{\gamma}{\beta} W(z_i) \leq 1$, and the list $z_1, z_2, \ldots$ can be enumerated, we have that $\frac{\gamma}{\beta} W(z_i) \stackrel{*}{<} \mathbf{m}(z_i|\beta,s)$, and thus

$$\log \frac{\mathbf{m}(z_i)}{W(z_i)} >^+ \log \frac{\gamma}{\beta} - \mathbf{K}(\beta, s),$$

which proves the theorem. □

References

[AF09] L. Antunes and L. Fortnow. Worst-case running times for average-case algorithms. In *2009 24th Annual IEEE Conference on Computational Complexity*, pages 298–303, 2009.

[BEHW89] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36(4):929–965, 1989.

[BGL+98] C. Bennett, P. Gacs, M. Li, P. Vitanyi, and W. Zurek. Information distance. *IEEE Transactions on Information Theory*, 44(4):1407–1423, 1998.

[CV05] R. Cilibrasi and P. Vitanyi. Clustering by compression. *IEEE Transactions on Information Theory*, 51(4):1523–1545, 2005.

[CV07] R. Cilibrasi and P. Vitanyi. The google similarity distance. *IEEE Transactions on Knowledge and Data Engineering*, 19(3):370–383, 2007.

[Deu99] David Deutsch. Quantum theory of probability and decisions. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 455(1988), 1999.

[EL11] Samuel Epstein and Leonid Levin. On sets of high complexity strings. *CoRR*, abs/1107.1458, 2011.

[Epsa] S. Epstein. December 24th: Algorithmic thermodynamic entropy. AIT Blog.

[Epsb] S. Epstein. Otcober 9th: A new proof to the sets have simple members theorem. AIT Blog.

[Epsc] S. Epstein. Uniform Tests and Algorithmic Thermodynamic Entropy. JP Theory Group Website. http://www.jptheorygroup.org/doc/Oscillation.pdf.

[Eps19a] S. Epstein. Algorithmic no-cloning theorem. *IEEE Transactions on Information Theory*, 65(9):5925–5930, 2019.

[Eps19b] S. Epstein. Algorithmic no-cloning theorem. *IEEE Transactions on Information Theory*, 65(9), 2019.

[Eps19c] S. Epstein. On the algorithmic probability of sets. *CoRR*, abs/1907.04776, 2019.

[Eps20] Samuel Epstein. An extended coding theorem with application to quantum complexities. *Information and Computation*, 275, 2020.

[Eps21a] S. Epstein. On the conditional complexity of sets of strings. *CoRR*, abs/1907.01018, 2021.

[Eps21b] Samuel Epstein. All sampling methods produce outliers. *IEEE Transactions on Information Theory*, 67(11):7568–7578, 2021.

[Eps22a] S. Epstein. 22 examples of solution compression via derandomization. *CoRR*, abs/2208.11562, 2022.

[Eps22b] S. Epstein. Derandomization under different resource constraints. *CoRR*, abs/2211.14640, 2022.

[Eps22c] S. Epstein. The kolmogorov birthday paradox. *CoRR*, abs/2208.11237, 2022.

[Eps22d] S. Epstein. The outlier theorem revisited. *CoRR*, abs/2203.08733, 2022.

[Eps23] S. Epstein. A Quantum EL Theorem. *CoRR*, abs/2301.08348, 2023.

[Eve57] Hugh Everett. "relative state" formulation of quantum mechanics. *Rev. Mod. Phys.*, 29, 1957.

[G̈61] Kurt Gödel. The modern development of the foundations of mathematics in the light of philosophy. *In: Kurt Gödel. Collected Works. Volume III. Oxford University Press.*, 1961.

[Gác01] P. Gács. Quantum Algorithmic Entropy. *Journal of Physics A Mathematical General*, 34(35), 2001.

[Gác13] P. Gács. Lecture notes on descriptional complexity and randomness, 2013.

[Gác21] Peter Gács. Lecture notes on descriptional complexity and randomness. *CoRR*, abs/2105.04704, 2021.

[Gac94] P. Gacs. The boltzmann entropy and randomness tests. In *Proceedings Workshop on Physics and Computation. PhysComp '94*, pages 209–216, 1994.

[HR09] M. Hoyrup and C. Rojas. Computability of probability measures and martin-löf randomness over metric spaces. *Information and Computation*, 207(7):830–847, 2009.

[Lee06] T. Lee. *Kolmogorov complexity and formula lower bounds.* PhD thesis, Amsterdam, 2006.

[Lev74] L. A. Levin. Laws of Information Conservation (Non-growth) and Aspects of the Foundations of Probability Theory. *Problemy Peredachi Informatsii*, 10(3):206–210, 1974.

[Lev84] L. A. Levin. Randomness conservation inequalities; information and independence in mathematical theories. *Information and Control*, 61(1):15–37, 1984.

[Lev13] L. A. Levin. Forbidden information. *J. ACM*, 60(2), 2013.

[Lev16] L. A. Levin. Occam bound on lowest complexity of elements. *Annals of Pure and Applied Logic*, 167(10):897–900, 2016. And also: S. Epstein and L.A. Levin, Sets have simple members, arXiv preprint arXiv:1107.1458, 2011.

[LOZ22] Z. Lu, I. Oliveira, and M. Zimand. Optimal coding theorems in time-bounded kolmogorov complexity. *CoRR*, abs/2204.08312, 2022.

[Rom22] A. Romashchenk. Clustering with respect to the information distance. *Theoretical Computer Science*, 929:164–171, 2022.

[SBKW10] S. Saunders, J. Barrett, A. Kent, and D. Wallace. *Many Worlds?: Everett, Quantum Theory, & Reality.* OUP Oxford, 2010.

[She12] A. Shen. Game Arguments in Computability Theory and Algorithmic Information Theory. In *Proceedings of 8th Conference on Computability in Europe* , volume 7318 of *LNCS*, pages 655–666, 2012.

[Vai98] L Vaidman. On schizophrenic experiences of the neutron or why we should believe in the many-worlds interpretation of quantum theory. *International Studies in the Philosophy of Science*, 12(3):245–261, 1998.

[Ver21] N. Vereshchagin. Proofs of conservation inequalities for levin's notion of mutual information of 1974. *Theoretical Computer Science*, 856, 2021.

[Vit00] P Vitányi. Three Approaches to the Quantitative Definition of Information in an Individual Pure Quantum State. In *Proceedings of the 15th Annual IEEE Conference on Computational Complexity*, COCO '00, page 263. IEEE Computer Society, 2000.

[VS17] Nikolay K. Vereshchagin and Alexander Shen. Algorithmic statistics: Forty years later. In *Computability and Complexity*, pages 669–737, 2017.

[VV10] N. Vereshchagin and P. Vitányi. Rate Distortion and Denoising of Individual Data using Kolmogorov Complexity. *IEEE Transactions on Information Theory*, 56, 2010.