On the Computable Members of Open Sets

Samuel Epstein*

July 22, 2022

Abstract

In this paper we prove that all open set of the Cantor space with large uniform measure will either have a simple computable member of high mutual information with the halting sequence. The first result characterizes clopen sets and the second result characterizes arbitrary open sets.

1 Introduction

This paper introduces results about the relationship between complexity and measure theory. The first result shows if a clopen set of the Cantor space has large uniform measure then it will have a simple computable member. Otherwise, the clopen set is exotic, in that it has high mutual information with the halting sequence. This result is then generalized to arbitrary open sets of the Cantor space.

For $x \in \Sigma^*$ let $\Gamma_x = \{x\beta : \beta \in \Sigma^\infty\}$ be the interval of x. For open set $S \subseteq \Sigma^\infty$, let its encoding be $\langle S \rangle = \langle \{x : \Gamma_x \text{ is maximal in } S\} \rangle$. For clopen sets C, $\langle C \rangle$ is a finite string. Arbitrary open sets $S \subset \Sigma^\infty$ can have infinite $\langle S \rangle$. The mutual information of a string x with the halting sequence is $\mathbf{I}(x;\mathcal{H}) = \mathbf{K}(x) - \mathbf{K}(x|\mathcal{H})$, where $\mathbf{K}$ is Kolmogorov complexity and $\mathcal{H} \in \Sigma^\infty$ is the halting sequence. The Kolmogorov complexity of an infinite sequence $\alpha \in \Sigma^\infty$ is $\mathbf{K}(\alpha)$, the size of the smallest program to a universal Turing machine that will output, without halting, α on the output tape. Let μ be the uniform measure of the Cantor space. The information term between infinite sequences is $\mathbf{I}(\alpha : \beta) = \log \sum_{x,y \in \Sigma^*} \mathbf{m}(x|\alpha)\mathbf{m}(y|\beta)2^{\mathbf{I}(x:y)}$.

Theorem.

1. *For clopen* $C \subseteq \Sigma^\infty$, $\min_{\alpha \in C} \mathbf{K}(\alpha) <^{\log} -\log \mu(C) + \mathbf{I}(\langle C \rangle; \mathcal{H})$.
2. *For open* $S \subseteq \Sigma^\infty$, $\min_{\alpha \in S} \mathbf{K}(\alpha) <^{\log} -\log \mu(S) + \mathbf{I}(\langle S \rangle : \mathcal{H})$.

For information on algorithmic information theory, we refer readers to [LV08]. The information function $\mathbf{I}$ was introduced in [Lev74].

2 Conventions

Let $\mathbb{N}$, $\mathbb{R}$, Σ, Σ^*, and Σ^∞ be the sets of natural numbers, real numbers, bits, finite strings, and infinite strings. We use $\langle x \rangle$ to represent a self delimiting code for $x \in \Sigma^*$, such as $1^{\|x\|}0x$. The self delimiting code for a finite set of strings $\{a_i\}_{i=1}^n$ is $\langle \{a_i\}_{i=1}^n \rangle = \langle n \rangle \langle a_1 \rangle \langle a_2 \rangle \dots \langle a_n \rangle$. Similarly for an

*JP Theory Group. samepst@jptheorygroup.org

infinite set of strings, $\{a_i\}_{i=1}^{\infty}$, its encoding is an infinite sequence $\langle\{a_i\}_{i=1}^{\infty}\rangle = \langle a_1\rangle\langle a_2\rangle\ldots$. For sets $S \subseteq \Sigma^{\infty}$ and $D \subseteq \Sigma^*$, $S \trianglelefteq D = \{x : x \in D, \Gamma_x \subseteq S\}$. For open set S, $\langle S\rangle = \langle\{x : \Gamma_x \text{ is maximal in } S\}\rangle$. This string can be infinite but it is finite for clopen sets.

The indicator function of a mathematical statement A is denoted by $[A]$, where if A is true then $[A] = 1$, otherwise $[A] = 0$. For positive real functions f the terms $<^+ f$, $>^+ f$, and $=^+ f$ represent $< f + O(1)$, $> f - O(1)$, and $= f \pm O(1)$, respectively. For nonnegative real function f the terms $<^{\log} f$, $>^{\log} f$ and $=^{\log} f$ represent $< f + O(\log(f+1))$, $> f - O(\log(f+1))$, and $= \pm O(\log(f+1))$, respectively.

A semi measure is a function $Q : \mathbb{N} \to \mathbb{R}_{\geq 0}$ such that $\sum_{a\in\mathbb{N}} Q(a) \leq 1$. A probability measure is a semi measure such that $\sum_{a\in\mathbb{N}} Q(a) = 1$. A probability measure Q is elementary if $|\{a : Q(a) > 0\}| < \infty$ and Range(Q) consists of all rationals. Elementary measures Q can be encoded into finite strings $\langle Q\rangle$.

For $x \in \Sigma^*$, $y \in \Sigma^* \cup \Sigma^{\infty}$, the output of algorithm T on input x and auxiliary input y is denoted $T_y(x)$. An algorithm T is prefix free if for strings $x, s \in \Sigma^*$, $y \in \Sigma^* \cup \Sigma^{\infty}$, $s \neq \varnothing$, if $T_y(x)$ halts then $T_y(xs)$ does not halt. There exists an optimal prefix-free algorithm U, meaning that for all prefix-free algorithms T, there exists $t \in \Sigma^*$ where $U_y(tx) = T_y(x)$ The function $\mathbf{K}(x|y) = \min\{\|p\| : U_y(p) = x\}$, is the Kolmogorov complexity of $x \in \Sigma^*$ conditional to $y \in \Sigma^* \cup \Sigma^{\infty}$. The Kolmogorov complexity of an infinite sequence $\alpha \in \Sigma^{\infty}$ is the size of the smallest input to U which will output, without halting, α on the output tape. The mutual information of finite strings $x, y \in \Sigma^*$ is $\mathbf{I}(x : y) = \mathbf{K}(x) + \mathbf{K}(y) - \mathbf{K}(x, y)$. The universal probability of a string $x \in \Sigma^*$, conditional to $y \in \Sigma^* \cup \Sigma^{\infty}$, is $\mathbf{m}(x|y) = \sum\{2^{-\|p\|} : U_y(p) = x\}$. The coding theorem states $-\log \mathbf{m}(x|y) =^+ \mathbf{K}(x|y)$. The halting sequence $\mathcal{H} \in \Sigma^{\infty}$ is the infinite string where $\mathcal{H}[i] = 1$ iff $U(i)$ halts. The amount of information that $\mathcal{H}$ has about $x \in \Sigma^*$ is $\mathbf{I}(x; \mathcal{H}) = \mathbf{K}(x) - \mathbf{K}(x|\mathcal{H})$. The mutual information between two sequences $\alpha, \beta \in \Sigma^{\infty}$, is $\mathbf{I}(\alpha : \beta) = \log \sum_{x,y\in\Sigma^*} \mathbf{m}(x|\alpha)\mathbf{m}(y|\beta)2^{\mathbf{I}(\alpha:\beta)}$.

This paper uses notions of stochasticity in the field of algorithmic statistics [VS17]. A string x is stochastic, i.e. has a low $\Lambda(x)$ score if it is typical of a simple probability distribution. The deficiency of randomness function of a string x with respect to an elementary probability measure P conditional to $y \in \Sigma^*$, is $\mathbf{d}(x|P, y) = \lfloor -\log P(x)\rfloor - \mathbf{K}(x|\langle P\rangle, y)$.

Definition 1 (Stochasticity)**.** *For $x, y \in \Sigma^*$,*
$\Lambda(x|y) = \min\{\mathbf{K}(P|y) + 3\log\max\{\mathbf{d}(x|P, y), 1\} : P$ *is an elementary probability measure*$\}$.

3 Clopen Sets

Lemma 1. *For clopen set $C \subseteq \Sigma^{\infty}$, $s = \lceil -\log\mu(C)\rceil$, $\min_{\alpha\in C} \mathbf{K}(\alpha) < s + \Lambda(\langle C\rangle) + O(\mathbf{K}(s))$.*

Proof. Let P be an elementary probabilty measure that realizes $\Lambda(\langle C\rangle|s)$. Let $n = \max\{\|x\| : x \in W \subset \Sigma^*, \langle W\rangle \in \text{Supp}(P)\}$. Let $d = \max\{\mathbf{d}(\langle C\rangle|P, s), 1\}$ and $c \in \mathbb{N}$ be a constant to be chosen later. Let κ be the uniform probability measure over lists L of $cd2^{s+1}$ strings of length n, where $\kappa(L) = 2^{-ncd2^{s+1}}$. Let $t_L(\langle W\rangle)$ be a function, parameterized by a list $L \subseteq \Sigma^n$, over encoded clopen

sets $W \subseteq \Sigma^\infty$, with $t_L(\langle W\rangle) = [\mu(W) \geq 2^{-s}, W \trianglelefteq L = \varnothing]e^{cd}$.

$$\begin{aligned}\mathbf{E}_{L\sim\kappa}\mathbf{E}_{\langle W\rangle\sim P}[t_L(\langle W\rangle)] &\leq \sum_{\text{clopen } W\subseteq\Sigma^\infty} P(\langle W\rangle)\,(1-2^{-s})^{cd2^{s+1}}\,e^{cd}\\ &\leq e^{-2^{-s}cd2^{s+1}}e^{cd} = e^{-cd}\\ &< 1.\end{aligned}$$

Thus there exists a list L of $cd2^{s+1}$ strings such that $\mathbf{E}_{\langle W\rangle\sim P}[t_L(\langle W\rangle)] < 1$. This L can be found with brute force search, with $\mathbf{K}(L|c,d,s,P) = O(1)$. It must be that $C \trianglelefteq L \neq \varnothing$. Otherwise $t_L(\langle C\rangle) = e^{cd}$ and since $t_L(\cdot)P(\cdot)$ is a semi-measure, for large enough c solely dependent on the universal Turing machine U, a contradiction occurs, with

$$\begin{aligned}\mathbf{K}(C|c,d,s,\langle P\rangle) &< -\log t_L(\langle C\rangle)P(\langle C\rangle) + O(1)\\ \mathbf{K}(C|c,d,s,\langle P\rangle) &< -\log P(\langle C\rangle) - (\lg e)cd + O(1)\\ (\lg e)cd &< -\log P(\langle C\rangle) - \mathbf{K}(C|s,\langle P\rangle) + \mathbf{K}(d,c) + O(1)\\ (\lg e)cd &< d + \mathbf{K}(d,c) + O(1).\end{aligned}$$

So there exists $x \in C \trianglelefteq L$, with

$$\begin{aligned}\mathbf{K}(x) &<^+ \log\left|L\right| + \mathbf{K}(L)\\ &<^+ \log\left|L\right| + \mathbf{K}(d,s,P)\\ &<^+ \log d + s + \mathbf{K}(d) + \mathbf{K}(s) + \mathbf{K}(P|s)\\ &< s + \Lambda(\langle C\rangle|s) + \mathbf{K}(s)\\ &< s + \Lambda(\langle C\rangle) + O(\mathbf{K}(s)).\end{aligned}$$

Since $x \in C \trianglelefteq L$, $\Gamma_x \subseteq C$. Thus there is a program g that outputs x and then an infinite sequence of 0's. Since $x0^\infty \in C$ and $\|g\| <^+ \mathbf{K}(x)$,

$$\min_{\alpha\in C}\mathbf{K}(\alpha) \leq \|g\| <^+ \mathbf{K}(x) < s + \Lambda(\langle C\rangle) + O(\mathbf{K}(s)).$$

Theorem 1.
For clopen set $C \subseteq \Sigma^\infty$, $s = \lceil -\log\mu(C)\rceil$, $h = \mathbf{I}(\langle C\rangle;\mathcal{H})$, $\min_{\alpha\in C}\mathbf{K}(\alpha) < s + h + O(\mathbf{K}(s,h))$.

Proof. This follows from Lemma 10 in [Eps21], which states $\Lambda(x) < \mathbf{I}(x;\mathcal{H}) + O(\mathbf{K}(\mathbf{I}(x;\mathcal{H})))$.

4 Open Sets

Theorem 1 can be generalized to arbitrary open sets of the Cantor space. Such sets S can have encodings $\langle S\rangle$ that are infinite sequences. The Big Oh term O and the $<^+$ are dependent solely on the choice of universal Turing machine.

Proposition 1.
For every $c,n \in \mathbb{N}$, if $x < y + c$ for some $x,y \in \mathbb{N}$ then $x + n\mathbf{K}(x) < y + n\mathbf{K}(y) + O(n\log n) + 2c$.

Proof. $\mathbf{K}(x) <^+ \mathbf{K}(y) + \mathbf{K}(y - x)$ as x can be computed from y and $(y - x)$. So $n\mathbf{K}(x) - n\mathbf{K}(y) < n\mathbf{K}(y - x) + dn$, for some $d \in \mathbb{N}$ dependent on U. Assume not, then there exists $x, y, c \in \mathbb{N}$ where $x < y + c$, and $g \leq O(n \log n) + 2c$ where $y - x + g < n\mathbf{K}(x) - n\mathbf{K}(y) < n\mathbf{K}(y - x) + dn$, which is a contradiction for $g =^+ dn + 2c + \max_a\{2n \log a - a\} =^+ dn + 2c + 2n \log n$.

Theorem 2.
For open set $S \subseteq \Sigma^\infty$, $s = \lceil -\log \mu(S) \rceil$, $h = \mathbf{I}(\langle S \rangle : \mathcal{H})$, $\min_{\alpha \in S} \mathbf{K}(\alpha) < s + h + O(\mathbf{K}(s, h))$.

Proof. Let $\{x_i\}_{i=1}^n = \{x : \Gamma_x \text{ is maximal in } S\}$, with $n \in \mathbb{N} \cup \infty$. Let $N \in \mathbb{N}$ be the smallest number such that $\sum_{i=1}^N 2^{-\|x_i\|} > 2^{-s-1}$. Let $C = \bigcup_{i=1}^N \Gamma_{x_i}$ be a clopen set with $C \subseteq S$. By Theorem 1,

$$\min_{\alpha \in C} \mathbf{K}(\alpha) < s + \mathbf{I}(\langle C \rangle; \mathcal{H}) + O(\mathbf{K}(s)) + O(\mathbf{K}(\mathbf{I}(\langle C \rangle; \mathcal{H}))). \tag{1}$$

By the definition of $\mathbf{I}$,

$$\begin{aligned} \mathbf{I}(\langle C \rangle; \mathcal{H}) &<^+ \mathbf{I}(\langle S \rangle : \mathcal{H}) + \mathbf{K}(\langle C \rangle | \langle S \rangle) \\ &<^+ \mathbf{I}(\langle S \rangle : \mathcal{H}) + \mathbf{K}(s). \end{aligned}$$

By Proposition 1, where $x = \mathbf{I}(\langle C \rangle; \mathcal{H})$, $y = \mathbf{I}(\langle S \rangle : \mathcal{H})$, and $c = \mathbf{K}(s) + O(1)$,

$$\mathbf{I}(\langle C \rangle; \mathcal{H}) + O(\mathbf{K}(\mathbf{I}(\langle C \rangle; \mathcal{H}))) < \mathbf{I}(\langle S \rangle : \mathcal{H}) + O(\mathbf{K}(\mathbf{I}(\langle S \rangle : \mathcal{H}))) + O(\mathbf{K}(s)). \tag{2}$$

Putting equations 1 and 2 together results in

$$\min_{\alpha \in S} \mathbf{K}(\alpha) < s + \mathbf{I}(\langle S \rangle : \mathcal{H}) + O(\mathbf{K}(s, \mathbf{I}(\langle S \rangle : \mathcal{H}))).$$

References

[Eps21] Samuel Epstein. All sampling methods produce outliers. *IEEE Transactions on Information Theory*, 67(11):7568–7578, 2021.

[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.

[LV08] M. Li and P. Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer Publishing Company, Incorporated, 3 edition, 2008.

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