

Outliers, Dynamics, and the Independence Postulate

Samuel Epstein*

July 7, 2022

Abstract

We show that outliers occur almost surely in computable dynamics over infinite sequences. Ever greater outliers can be found as the number of visited states increases. We show the Independence Postulate explains how outliers are found in the physical world. We generalize the outliers theorem to uncomputable sampling methods.

1 Introduction

An outlier is an observation whose value lies outside the set of values considered likely according to some hypothesis (usually one based on other observations); an isolated point. In the realm of algorithmic information theory, outliers are modeled using the deficiency of randomness. The deficiency of randomness of a number $a \in \mathbb{N}$ with respect to a computable probability measure p over \mathbb{N} is $\mathbf{d}(a|p) = -\log p(a) - \mathbf{K}(a)$. the randomness deficiency of an infinite sequence $\alpha \in \Sigma^\infty$ with respect to a computable probability measure P is $\mathbf{D}(\alpha|P) = \sup_n -\log P(\alpha[0..n]) - \mathbf{K}(\alpha[0..n])$. The term \mathbf{K} is the prefix free Kolmogorov complexity.

The definition of a randomness deficiency is a very useful mathematical definition, and can be used to show the ubiquity of outliers in different constructs. A sampling method is a probabilistic program that when given n , outputs with probability 1, n unique elements, either numbers or infinite sequences. In [Eps21], it was proven that sampling methods produce outliers. In [Eps22], it was proven that for ergodic dynamical systems, ever greater outlying states occur with diminishing measure. In this paper, we show that arbitrary (potentially non-ergodic) computable dynamical systems over Σ^∞ produce outliers, where as more states are visited, more and greater outliers are guaranteed to occur.

A computable dynamical system (λ, δ) consists of a computable starting state probability λ over Σ^∞ and a computable transition function $\delta : \Sigma^\infty \rightarrow \Sigma^\infty$. 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. *There exists $d \in \mathbb{N}$, where for computable probability μ over Σ^∞ and computable dynamics (λ, δ) , for λ -a.e. starting states $\alpha \in \Sigma^\infty$, 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 , $\mathbf{E}_{\alpha \sim \lambda} [s_\alpha - O(\log s_\alpha)] <^+ \mathbf{K}(\lambda, \delta) + \mathbf{K}(\mu)$.*

The proof technique is two stages. First is to prove that certain finite sets of natural numbers or infinite sequences have high mutual information, \mathbf{I} , with the halting sequence. This is represented in Theorems 1 and 2. The second step is to use conservation properties of \mathbf{I} , shown in Theorem 3, to achieve the main theorem of the paper. In fact, as shown in Section 6, a stronger version of the above theorem is proven, generalizing to arbitrary (i.e. uncomputable) dynamics. The above theorem holds if the (potentially infinite) encoding of the dynamical system has finite mutual

*JP Theory Group. samepst@jpththeorygroup.org

information with the halting sequence. This generalization is made possible due the two step process described above. Another generalization of the above theorem, detailed in the Discussion, is for the probability measure μ to be arbitrary, i.e. uncomputable.

The above approach is compatible with the Independence Postulate. The Independence Postulate (**IP**), [Lev84, Lev13], is an unprovable inequality on the information content shared between two sequences. **IP** is a finitary Church Turing Thesis, 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 main theorem of this paper and [Eps21, Eps22]. While the latter shows that computable constructs produce outliers with high probability, the former states that individual sequences without outliers have high addresses, i.e. are hard to find nature. Both methods start with proving certain sequences have mutual information with the halting sequence. In Section 7 the following statement is derived from **IP**.

Statement. *For computable probability p over \mathbb{N} and $\tau \in \mathbb{N}^{\mathbb{N}}$, let s_τ be the smallest number such that for $\tau(n)$ equal to the first 2^n unique numbers of τ , $\max_{a \in \tau(n)} \mathbf{d}(a|p) > n - 4\mathbf{K}(n) - s_\tau$. Then for every physical address k of τ , $s_\tau <^{\log} k + \mathbf{K}(p)$.*

The sequence τ can be seen to be a sequence of observations, each encoded by a number, $\tau[i] \in \mathbb{N}$. Furthermore, τ is assumed to have an infinite number of unique numbers. Section 7 also details known relationships between **IP** and the halting problem, Peano Arithmetic, and inductive inference [Lev13]. We recommend reading Section 7 first. Section 8 improves upon [Eps21], showing that (potentially uncomputable) sampling methods that have finite mutual information with the halting sequence will produce outliers.

2 Conventions

We use \mathbb{N} , \mathbb{Q} , \mathbb{R} , Σ , Σ^* , and Σ^∞ to represent natural numbers, rational numbers, reals, bits, finite strings, and infinite strings. The removal of the last bit of a string is denoted by $(p0^-) = (p1^-) = p$, for $p \in \Sigma^*$. We use $\Sigma^{*\infty}$ to denote $\Sigma^* \cup \Sigma^\infty$, the set of finite and infinite strings. For $x \in \Sigma^{*\infty}$, $y \in \Sigma^{*\infty}$, we say $x \sqsubseteq y$ if $x = y$ or $x \in \Sigma^*$ and $y = xz$ for some $z \in \Sigma^{*\infty}$. Also $x \sqsubset y$ if $x \sqsubseteq y$ and $x \neq y$. The indicator function of a mathematical statement A is denoted by $[A]$, where if A is true then $[A] = 1$, otherwise $[A] = 0$. For sets Z of infinite strings, $Z_n = \{\alpha[0..n] : \alpha \in Z\}$ and $\langle Z \rangle = \langle Z_1 \rangle \langle Z_2 \rangle \langle Z_3 \rangle \dots$.

As is typical of the field of algorithmic information theory, the theorems in this paper are relative to a fixed universal machine, and therefore their statements are only relative up to additive and logarithmic precision. For positive real functions f the terms $<^+ f$, $>^+ f$, $=^+ f$ represent $< f + O(1)$, $> f - O(1)$, and $= f \pm O(1)$, respectively. In addition $<^* f$, $>^* f$ denote $< f/O(1)$, $> f/O(1)$. The terms $=^* f$ denotes $<^* f$ and $>^* f$. For nonnegative real function f , the terms $<^{\log} f$, $>^{\log} f$, $=^{\log} f$ represent the terms $< f + O(\log(f+1))$, $> f - O(\log(f+1))$, and $= f \pm O(\log(f+1))$, respectively. A discrete measure is a nonnegative function $Q : \mathbb{N} \rightarrow \mathbb{R}_{\geq 0}$ over natural numbers. The support of a measure Q is the set of all elements whose Q value is positive, with $\text{Supp}(Q) = \{a : Q(a) > 0\}$. A measure is elementary if its support is finite and its range is a subset of \mathbb{Q} . We say that Q is a probability measure if $\sum_a Q(a) = 1$.

$T_y(x)$ is the output of algorithm T (or \perp if it does not halt) on input $x \in \Sigma^*$ and auxiliary input $y \in \Sigma^{*\infty}$. T is prefix-free if for all $x, s \in \Sigma^*$ with $s \neq \emptyset$, and $y \in \Sigma^{*\infty}$, either $T_y(x) = \perp$ or $T_y(xs) = \perp$. The complexity of $x \in \Sigma^*$ with respect to T_y is $\mathbf{K}_T(x|y) = \min\{\|p\| : T_y(p) = x\}$.

There exists optimal for \mathbf{K} prefix-free algorithm U , meaning that for all prefix-free algorithms T , there exists $c_T \in \mathbb{N}$, where $\mathbf{K}_U(x|y) \leq \mathbf{K}_T(x|y) + c_T$ for all $x \in \Sigma^*$ and $y \in \Sigma^{*\infty}$. For example, one can take a universal prefix-free algorithm U , where for each prefix-free algorithm T , there exists $t \in \Sigma^*$, with $U_y(tx) = T_y(x)$ for all $x \in \Sigma^*$ and $y \in \Sigma^{*\infty}$. The function $\mathbf{K}(x|y)$, defined to be $\mathbf{K}_U(x|y)$, is the Kolmogorov complexity of $x \in \Sigma^*$ relative to $y \in \Sigma^{*\infty}$. When we say that a universal Turing machine is relativized to an object, this means that an encoding of the object is provided to the universal Turing machine on an auxiliary tape. The universal probability of a number $a \in \mathbb{N}$ is $\mathbf{m}(a|y) = \sum_z [U_y(z) = a] 2^{-\|z\|}$. The coding theorem states $-\log \mathbf{m}(a|y) = {}^+ \mathbf{K}(a|y)$.

The chain rule for Kolmogorov complexity is $\mathbf{K}(x, y) = {}^+ \mathbf{K}(x) + \mathbf{K}(y|\langle x, \mathbf{K}(x) \rangle)$. The mutual information in finite strings x and y relative to $z \in \Sigma^*$ is $\mathbf{I}(x : y | z) = \mathbf{K}(x|z) + \mathbf{K}(y|z) - \mathbf{K}(\langle x, y \rangle | z) = {}^+ \mathbf{K}(x|z) - \mathbf{K}(x|\langle y, \mathbf{K}(y|z), z \rangle)$.

For computable probability measure λ over Σ^∞ , $\mathbf{K}(\lambda)$ is the size of the shortest program that can compute $\lambda(x\Sigma^\infty)$, for all $x \in \Sigma^*$. For computable function $\delta : \Sigma^\infty \rightarrow \Sigma^\infty$, $\mathbf{K}(\delta)$ is the size of the shortest program on the input tape that outputs $\delta(\alpha)$ when α is on the auxiliary tape. The halting sequence $\mathcal{H} \in \Sigma^\infty$ is the infinite string where $\mathcal{H}[i] = [U(i) \neq \perp]$ for all $i \in \mathbb{N}$. The amount of information that $a \in \mathbb{N}$ has with \mathcal{H} is denoted by $\mathbf{I}(a; \mathcal{H}) = \mathbf{K}(a) - \mathbf{K}(a|\mathcal{H})$.

The deficiency of randomness of x with respect to elementary measure Q and $v \in \mathbb{N}$ is $\mathbf{d}(x|Q, v) = \lfloor -\log Q(x) \rfloor - \mathbf{K}(x|\langle Q \rangle, v)$. The stochasticity of $a \in \mathbb{N}$, conditional to $b \in \mathbb{N}$, is measured by $\mathbf{Ks}(a|b) = \min\{\mathbf{K}(Q|b) + 3 \log \max\{\mathbf{d}(a|Q, b), 1\} : Q \text{ is an elementary probability measure}\}$. We have $\mathbf{Ks}(a) = \mathbf{Ks}(a|\emptyset)$, with $\mathbf{Ks}(a|b) < \mathbf{Ks}(a) + O(\log \mathbf{K}(b))$.

3 Sets of Numbers

Theorem 1 *For computable probability p over \mathbb{N} and for $D \subset \Sigma^*$, $|D| = 2^s$, $m \in [0, s - 1]$, there are 2^m elements $a \in D$, with $s - m < \mathbf{d}(a|p) + \mathbf{Ks}(D) + \mathbf{K}(p) + O(\log s + \log \mathbf{K}(p))$.*

Proof. We relativize the universal Turing machine U to p and s for the duration of the proof, which can be done as the theorem has precision $O(\log s)$. Let Q be an elementary probability distribution that realizes $\mathbf{Ks}(D)$. Let $d = \mathbf{d}(D|Q)$ be the deficiency of randomness of D with respect to Q . Let V be the combined elements of encoded sets of size 2^s in the support of Q . We create an algorithm, that given Q , s and p produces 2^{s-1} sets $F_i \subseteq V$. We start with the first round. Suppose each element of V is selected independently with probability $cd2^{-s}$, where c is a constant to be chosen later. The selected set is F_1 , and $\mathbf{E}[p(F_1)] \leq cd2^{-s}$. Furthermore

$$\mathbf{E}[Q(\{G : |G| = 2^s, G \cap F_1 = \emptyset\})] \leq \sum_G Q(G)(1 - cd2^{-s})^{2^s} < e^{-cd}.$$

Thus a finite set F_1 can be chosen such that $p(F_1) \leq 2cd2^{-s}$ and $Q(\{G : |G| = 2^s, G \cap F_1 = \emptyset\}) \leq e^{1-cd}$. Now it must be that $D \cap F_1 \neq \emptyset$. Otherwise, using the Q -test, $t(G) = [|G| = 2^s, G \cap F_1 = \emptyset]e^{cd-1}$, we have

$$\begin{aligned} \mathbf{K}(D|Q, d, c) &< {}^+ -\log Q(D) - (\log e)cd \\ (\log e)cd &< {}^+ -\log Q(D) - \mathbf{K}(D|Q) + \mathbf{K}(d, c) \\ (\log e)cd &< {}^+ d + \mathbf{K}(d, c), \end{aligned}$$

which is a contradiction for large enough c solely dependent on the universal Turing machine U . Thus there is an $a \in D \cap F_1$, where

$$\begin{aligned}\mathbf{K}(a) &<^+ -\log p(a) + \log d - s + \mathbf{K}(d) + \mathbf{K}(Q) \\ s &<^+ \mathbf{d}(a|p) + \mathbf{Ks}(D).\end{aligned}$$

Removing the relativization of p and s for just the following 2 equations,

$$\begin{aligned}s &< -\log p(a) - \mathbf{K}(a|p) + \mathbf{Ks}(D|p) + O(\log s), \\ s &< \mathbf{d}(a|p) + \mathbf{Ks}(D) + \mathbf{K}(p) + O(\log s + \log \mathbf{K}(p)).\end{aligned}$$

We define the construction of set F_i given that the first $i - 1$ rounds have already occurred. Let $F_{<i} = \bigcup_{j=1}^{i-1} F_j$. A set G is eligible if $|G| = 2^s$, and $|G \setminus F_{<i}| \geq 2^s - (i - 1)$. Set F_i is selected at random from V , with each element selected at random with probability cd_i2^{-s} , with $d_i = d \log i$. $\mathbf{E}[p(F_i)] \leq cd_i2^{-s}$.

$$\begin{aligned}\mathbf{E}[Q(\{G : G \text{ is eligible}, (G \setminus F_{<i}) \cap F_i = \emptyset\})] \\ \leq \sum_{\text{eligible } G} Q(G)(1 - cd_i2^{-s})^{2^s - (i-1)} \\ \leq e^{-cd_i2^{-s}(2^s - (i-1))} \leq e^{-cd_i/2}.\end{aligned}$$

Thus a finite set F_i can be chosen such that $p(F_i) \leq 2cd_i2^{-s}$ and $Q(\{G : G \text{ is eligible}, (G \setminus F_{<i}) \cap F_i = \emptyset\}) \leq e^{-cd_i/2+1}$. It must be that on the rounds i that D is eligible, $(D \setminus F_{<i}) \cap F_i \neq \emptyset$. Otherwise one can create a Q -test $t_i(G) = [G \text{ is eligible}, (G \setminus F_{<i}) \cap F_i = \emptyset]e^{cd_i/2-1}$. Thus $t_i(D) \neq 0$ and

$$\begin{aligned}\mathbf{K}(D|Q, d_i, i, c) &<^+ -\log Q(D) - (\log e)cd_i/2 \\ .5(\log e)cd \log i &<^+ -\log Q(D) - \mathbf{K}(D|Q) + \mathbf{K}(d_i, i, c) \\ .5(\log e)cd \log i &<^+ d + \mathbf{K}(d, i, c).\end{aligned}$$

This is a contradiction for large enough c dependent solely on the universal Turing machine U . Thus on rounds i where D is eligible, there exist an $a \in (D \setminus F_{<i}) \cap F_i$, with

$$\begin{aligned}\mathbf{K}(a) &<^+ -\log p(a) + \log d_i - s + \mathbf{K}(d_i) + \mathbf{K}(i) + \mathbf{K}(Q) \\ s &< \mathbf{d}(a|p) + \log i + O(\log \log i) + \log d + \mathbf{K}(d) + \mathbf{K}(Q) \\ s - \log i &< \mathbf{d}(a|p) + \mathbf{Ks}(D) + O(\log s).\end{aligned}$$

Removing the relativization of p (and s) results in

$$\begin{aligned}s - \log i &< -\log p(a) - \mathbf{K}(a|p) + \mathbf{Ks}(D|p) + O(\log s), \\ s - \log i &< \mathbf{d}(a|p) + \mathbf{Ks}(D) + \mathbf{K}(p) + O(\log s + \log \mathbf{K}(p)).\end{aligned}\tag{1}$$

On rounds i in which D is not eligible, then there exist a round $j < i$ where $|(D \setminus F_{<j}) \cap F_j| > 1$. And for each such element in the intersection, a bound on their deficiency of randomness similar to Equation 1 can be proven. \square

Corollary 1 *For computable probability p over \mathbb{N} and for $D \subset \Sigma^*$, $|D| = 2^s$, $s < \max_{a \in D} \mathbf{d}(a|p) + \mathbf{Ks}(D) + \mathbf{K}(p) + 2\mathbf{K}(s) + O(\log \mathbf{K}(p))$.*

4 Left-Total Machines

We say $x \in \Sigma^*$ 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 $y \in \Sigma^{*\infty}$ if there exists a finite prefix free set of strings $Z \subset \Sigma^*$ where $\sum_{z \in Z} 2^{-\|z\|} = 1$ and $T_y(xz) \neq \perp$ for all $z \in Z$. We say $\alpha \in \Sigma^{*\infty}$ is to the “left” of $\beta \in \Sigma^{*\infty}$, and use the notation $\alpha \triangleleft \beta$, if there exists $x \in \Sigma^*$ such that $x0 \sqsubseteq \alpha$ and $x1 \sqsubseteq \beta$. A machine T is left-total if for all auxiliary strings $\alpha \in \Sigma^{*\infty}$ and for all $x, y \in \Sigma^*$ with $x \triangleleft y$, one has that $T_\alpha(y) \neq \perp$ implies that x is total with respect to T_α . An example left-total machine can be seen in Figure 1.



Figure 1: The above diagram represents the domain of a left total machine T with the 0 bits branching to the left and the 1 bits branching to the right. For $i \in \{1, \dots, 5\}$, $x_i \triangleleft x_{i+1}$ and $x_i \triangleleft y$. Assuming $T(y)$ halts, each x_i is total. This also implies each x_i^- is total as well.

For the remaining part of this paper, we can and will change the universal self delimiting machine U into an optimal left-total machine U' . For a detailed explanation on how to construct a left-total universal Turing machine, we refer readers to [Eps21].

Without loss of generality, the complexity terms of this paper are defined with respect to the optimal left total machine U .

5 Sets of Infinite Strings

For total string b , let $\mathbf{bb}(b) = \max\{U(p) : p \triangleleft b \text{ or } p \sqsubseteq b\}$ be the largest number produced by a program that extends b or is to the left of b .

Definition 1 ([Lev74]) *The mutual information between two infinite sequences α and β is*

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

Lemma 1 ([EL11]) *For $x \in \mathbb{N}$, $\mathbf{Ks}(x) <^{\log} \mathbf{I}(x; \mathcal{H})$.*

Lemma 2 ([Eps21]) *If $b \in \Sigma^*$ is total and b^- is not, and $x \in \Sigma^*$, then $\mathbf{K}(b) + \mathbf{I}(x; \mathcal{H}|b) <^{\log} \mathbf{I}(x; \mathcal{H}) + \mathbf{K}(b|\langle x, \|b\| \rangle)$.*

Lemma 3 ([Eps21]) *If $b \in \Sigma^*$ is total and b^- is not, and for $x \in \Sigma^*$, $\mathbf{K}(b|\langle x, |b| \rangle) = O(1)$, then $\mathbf{K}(\|b\|) <^{\log} 2 \log \mathbf{I}(x; \mathcal{H})$.*

Theorem 2 *For computable probability P over Σ^∞ and $Z \subset \Sigma^\infty, |Z| = 2^s$, $m \in [0, s-1]$, there are 2^m elements $\alpha \in Z$, with $s-m < \mathbf{D}(\alpha|P) + \mathbf{I}(Z : \mathcal{H}) + \mathbf{K}(P) + O(\log s + \log \mathbf{I}(Z : \mathcal{H}) + \log \mathbf{K}(P))$.*

Proof. The proof of this theorem follows closely in form to the proof of Theorem 5 in [Eps21], except Theorem 1 is referenced. Fix $m \in [0, s-1]$. Let b be the shortest total string such that $|Z_{\mathbf{bb}(b)}| = 2^s$. Set $D = Z_{\mathbf{bb}(b)}$. Let $p(x) = [\|x\| = \mathbf{bb}(b)]P(\{\alpha : x \sqsubset \alpha\})$. Using Theorem 1, relativized to b , produces 2^m elements $F \subseteq D$ such that for $x \in F$,

$$\begin{aligned} \mathbf{K}(x|b) &< -\log p(x) - s + m + \mathbf{K}s(D|b) + \mathbf{K}(p|b) + O(\log s + \log \mathbf{K}(p|b)), \\ \mathbf{K}(x|b) &< -\log p(x) - s + m + \mathbf{K}s(D|b) + \mathbf{K}(P) + O(\log s + \log \mathbf{K}(P)). \end{aligned}$$

Using Lemma 1, relativized to b ,

$$\begin{aligned} \mathbf{K}(x|b) &< -\log p(x) - s + m + \mathbf{I}(D; \mathcal{H}|b) + \mathbf{K}(P) + O(\log s + \log \mathbf{I}(D; \mathcal{H}|b) + \log \mathbf{K}(P)) \\ s-m &< \log(\mathbf{m}(x)/p(x)) + \mathbf{K}(b) + \mathbf{I}(D; \mathcal{H}|b) + \mathbf{K}(P) + O(\log s + \log(\mathbf{I}(D; \mathcal{H}|b) + \mathbf{K}(b)) + \log \mathbf{K}(P)). \end{aligned}$$

By Lemma 2,

$$\begin{aligned} s-m &< \log(\mathbf{m}(x)/p(x)) + \mathbf{I}(D; \mathcal{H}) + \mathbf{K}(b|D, \|b\|) + \mathbf{K}(P) \\ &\quad + O(\log s + \log(\mathbf{I}(D; \mathcal{H}) + \mathbf{K}(b|D, \|b\|)) + \log \mathbf{K}(P)). \end{aligned}$$

Since $D \subseteq \Sigma^{\mathbf{bb}(b)}$, $\mathbf{K}(b|D, \|b\|) = O(1)$, as a program can output the leftmost total string y of length $\|b\|$ such that $\mathbf{bb}(y)$ is the length of the strings in D . So

$$s-m < \log(\mathbf{m}(x)/p(x)) + \mathbf{I}(D; \mathcal{H}) + \mathbf{K}(P) + O(\log s + \log \mathbf{I}(D; \mathcal{H}) + \log \mathbf{K}(P)).$$

We have that $\mathbf{K}(D|\langle Z \rangle) <^+ \mathbf{K}(\|b\|) + \mathbf{K}(s)$, as D is computable from $\langle Z \rangle$, $\|b\|$, and s . This is because b is computable from its length, s , and $\langle Z \rangle$, and thus so is $D = Z_{\mathbf{bb}(b)}$. By the Definition 1 of the mutual information between infinite sequences,

$$\begin{aligned} \mathbf{I}(D; \mathcal{H}) &<^+ \mathbf{I}(\langle Z \rangle : \mathcal{H}) + \mathbf{K}(D|\langle Z \rangle) \\ &<^+ \mathbf{I}(\langle Z \rangle : \mathcal{H}) + \mathbf{K}(\|b\|) + \mathbf{K}(s) \\ &<^+ \mathbf{I}(\langle Z \rangle : \mathcal{H}) + 2 \log \mathbf{I}(D; \mathcal{H}) + \mathbf{K}(s) \\ &<^{\log} \mathbf{I}(\langle Z \rangle : \mathcal{H}) + \mathbf{K}(s), \end{aligned} \tag{2}$$

where Equation 2 is due to Lemma 3, noting $\mathbf{K}(b|D, \|b\|) = O(1)$. So there is an $\alpha \in Z$, $x \sqsubset \alpha$, with

$$\begin{aligned} s-m &< \log(\mathbf{m}(x)/p(x)) + \mathbf{I}(D; \mathcal{H}) + \mathbf{K}(P) + O(\log s + \log \mathbf{I}(D; \mathcal{H}) + \log \mathbf{K}(P)) \\ s-m &< \mathbf{D}(\alpha|P) + \mathbf{I}(\langle Z \rangle : \mathcal{H}) + \mathbf{K}(P) + O(\log s + \mathbf{I}(\langle Z \rangle : \mathcal{H}) + \log \mathbf{K}(P)). \end{aligned}$$

□

6 Outliers in Dynamics

This section contains the main result of the paper, that dynamical systems will exhibit outliers. To prove this fact, Theorem 3 will be leveraged, which details the conservation properties of the halting sequence \mathcal{H} .

For a continuous function $\delta : \Sigma^\infty \rightarrow \Sigma^\infty$, $\langle \delta \rangle$ is any infinite sequence $\delta' \in \Sigma^\infty$, such that if $\alpha \in \Sigma^\infty$ is on auxilliary tape of U and δ' is on the input tape, U outputs $\delta(\alpha)$ on the output tape, without halting. Similarly for arbitrary (i.e. uncomputable) probability measure λ over Σ^∞ , $\langle \lambda \rangle$ is any infinite sequence λ' such that if $x \in \Sigma^*$ is on the auxillary tape and λ' is on the input tape of U , then U outputs $\lambda(x\Sigma^\infty)$. For probability λ , $\mathbf{I}(\lambda : \mathcal{H}) = \inf_{\langle \lambda \rangle} \mathbf{I}(\langle \lambda \rangle : \mathcal{H})$. We say that for dynamical system (λ, δ) , $\mathbf{I}((\lambda, \delta) : \mathcal{H}) = \inf_{\langle \lambda \rangle, \langle \delta \rangle} \mathbf{I}(\langle \lambda \rangle, \langle \delta \rangle : \mathcal{H})$.

Theorem 3 ([Ver21, Lev74, Gei12])

- $\mathbf{E}_{\alpha \sim \lambda} [2^{\mathbf{I}(\alpha : \mathcal{H})}] \overset{*}{<} 2^{\mathbf{I}(\lambda : \mathcal{H})}$.
- $\mathbf{I}(f(\alpha) : \mathcal{H}) <^+ \mathbf{I}(\alpha : \mathcal{H}) + \mathbf{K}(f)$.

Theorem 4 (Outliers in Dynamics) *There exists $d \in \mathbb{N}$, where for computable probability μ over Σ^∞ and dynamics (λ, δ) , with $\mathbf{I}((\lambda, \delta) : \mathcal{H}) \neq \infty$, for λ -a.e. starting states $\alpha \in \Sigma^\infty$, 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 , $\mathbf{E}_{\alpha \sim \lambda} [s_\alpha - O(\log s_\alpha)] <^+ \mathbf{I}((\lambda, \delta) : \mathcal{H}) + \mathbf{K}(\mu)$.*

Proof. Fix a starting state $\alpha \in \Sigma^\infty$ and fix $d \in \mathbb{N}$. Assume α has property A , in which for all $s_\alpha \in \mathbb{N}$, there exists $m, n \in \mathbb{N}$, $m < n$, where the first 2^n states $Z \subset \Sigma^\infty$ visited has less than 2^m states $\beta \in Z$, with

$$\mathbf{D}(\beta|\mu) > n - m - d \log n - s_\alpha.$$

Therefore, due to Theorem 2 there exists a state $\beta \in Z$, with

$$\mathbf{D}(\beta|\mu) \leq n - m - d \log n - s_\alpha$$

and

$$n - m < \mathbf{D}(\beta|\mu) + \mathbf{I}(Z : \mathcal{H}) + \mathbf{K}(\mu) + O(\log \mathbf{K}(\mu) + \log n + \log \mathbf{I}(Z : \mathcal{H})).$$

Due to Theorem 3, we have

$$\mathbf{I}(Z : \mathcal{H}) <^+ \mathbf{I}(\langle \alpha, \delta \rangle : \mathcal{H}) + \mathbf{K}(n),$$

so

$$n - m < \mathbf{D}(\beta|\mu) + \mathbf{I}(\langle \alpha, \delta \rangle : \mathcal{H}) + \mathbf{K}(\mu) + O(\log \mathbf{K}(\mu) + \log n + \log \mathbf{I}(\langle \alpha, \delta \rangle : \mathcal{H})).$$

So

$$n - m < n - m - d \log n - s_\alpha + \mathbf{I}(\langle \alpha, \delta \rangle : \mathcal{H}) + \mathbf{K}(\mu) + O(\log \mathbf{K}(\mu) + \log n + \log \mathbf{I}(\langle \alpha, \delta \rangle : \mathcal{H})),$$

implying

$$d \log n + s_\alpha < \mathbf{I}(\langle \alpha, \delta \rangle : \mathcal{H}) + \mathbf{K}(\mu) + O(\log \mathbf{K}(\mu) + \log n + \log \mathbf{I}(\langle \alpha, \delta \rangle : \mathcal{H})).$$

Thus for large enough d , dependent solely on the universal Turing machine U , $\mathbf{I}(\langle \alpha, \delta \rangle : \mathcal{H}) = \infty$. Let $\gamma(\langle \xi, \zeta \rangle) = \lambda(\xi)[\zeta = \langle \delta \rangle]$. By Theorem 3, $\mathbf{E}_{\alpha \sim \lambda} [2^{\mathbf{I}(\langle \alpha, \delta \rangle : \mathcal{H})}] = \mathbf{E}_{\xi \sim \gamma} [2^{\mathbf{I}(\xi : \mathcal{H})}] \overset{*}{<} 2^{\mathbf{I}(\gamma : \mathcal{H})} \overset{*}{<} 2^{\mathbf{I}((\lambda, \delta) : \mathcal{H})} < \infty$. Thus by Theorem 3, λ -a.e. states α do not have the property A .

By the reasoning above, the smallest such s_α has $s_\alpha <^{\log} \mathbf{I}(\langle \alpha, \delta \rangle : \mathcal{H}) + \mathbf{K}(\mu)$. So $s_\alpha - O(\log s_\alpha) < \mathbf{I}(\langle \alpha, \delta \rangle : \mathcal{H}) + \mathbf{K}(\mu)$. So by Theorem 3, $\mathbf{E}_{\alpha \sim \lambda} \left[2^{s_\alpha - O(1)} \right] <^* \mathbf{E}_{\alpha \sim \lambda} \left[2^{\mathbf{I}(\langle \alpha, \delta \rangle : \mathcal{H}) + \mathbf{K}(\mu)} \right] <^* \mathbf{E}_{\xi \sim \gamma} \left[2^{\mathbf{I}(\xi : \mathcal{H}) + \mathbf{K}(\mu)} \right]$ implies $\mathbf{E}_{\alpha \sim \lambda} [s_\alpha - O(\log s_\alpha)] <^+ \mathbf{I}(\gamma : \mathcal{H}) + \mathbf{K}(\mu) <^+ \mathbf{I}((\lambda, \delta) : \mathcal{H}) + \mathbf{K}(\mu)$. \square

Corollary 2 (Computable Dynamics) *There exists $d \in \mathbb{N}$, where for computable probability μ over Σ^∞ and computable dynamics (λ, δ) , for λ -a.e. starting states $\alpha \in \Sigma^\infty$, 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 , $\mathbf{E}_{\alpha \sim \lambda} [s_\alpha - O(\log s_\alpha)] <^+ \mathbf{K}(\lambda, \delta) + \mathbf{K}(\mu)$.*

This follows from Theorem 4 and Theorem 3, where $\mathbf{I}((\lambda, \delta) : \mathcal{H}) <^+ \mathbf{I}(0^\infty : \mathcal{H}) + \mathbf{K}(\lambda, \delta) <^+ \mathbf{K}(\lambda, \delta)$.

7 Independence Postulate

In this section, we show how the Independence Postulate can be used to explain outliers found in the physical world. The Independence Postulate (**IP**), [Lev84, Lev13], is an unprovable inequality on the information measure of two sequences. **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”. One such example is finite prefixes of the halting problem. If a forbidden sequence can be found with a low address, then an “information leak” occurs. L. A. Levin states [Lev13]

The toolkit of our models may change (e.g., quantum amplitudes work somewhat differently than probabilities) but it is hard to expect new realistic primitives allowing such “information leaks”.

The statement of the **IP** is as follows.

IP: *Let X be a sequence defined with an n -bit mathematical statement (e.g., in PA or set theory), and a sequence Y can be located in the physical world with a k -bit instruction set (e.g., ip-address). Then $\mathbf{I}(X : Y) < k + n + c$ for some small absolute constant c .*

We take **I** be the information term in Definition 1. One consequence of **IP** is a finite version of the (physical) Church-Turing Thesis. **IP** says that the only finite sequences that can be found in nature (i.e. have short physical addresses) will have non-recursive descriptions that are equal in length to their recursive descriptions. This can be seen when **IP** is applied to the case when $\alpha = \beta$ is a finite sequence which has a non-recursive description of length n that is much shorter than its recursive description $\mathbf{K}(\alpha)$, with $n \ll \mathbf{K}(\alpha)$. Let k be the shortest physical address of α . Then by **IP**,

$$\begin{aligned} \mathbf{K}(\alpha) &<^+ \mathbf{I}(\alpha : \alpha) + c <^+ k + n + c \\ \mathbf{K}(\alpha) - n - c &<^+ k. \end{aligned} \tag{3}$$

Thus k is large and α cannot be easily located in the physical world. The only sequences α with short physical addresses must have $n \approx \mathbf{K}(\alpha)$.

7.1 Halting Sequence

An example of the inequality in Equation 3 is prefixes of the halting sequence. Let H_n be the finite sequence that is the prefix of size 2^n of \mathcal{H} . It is well known that there is some small constant c_{HK}

independent of n such that $\mathbf{K}(H_n) \in (n - O(1), n + \mathbf{K}(n) + O(1))$. The entire halting sequence \mathcal{H} can be described in a mathematical statement of size equal to some small constant c_{HM} . Each n can be described using a program of size $\mathbf{K}(n)$. Therefore each H_n can be defined by a mathematical statement of size $c_{\text{HM}} + \mathbf{K}(n)$. So by **IP** applied to $\alpha = \beta = H_n$ where k_n is the smallest physical address of H_n ,

$$\begin{aligned} n <^+ \mathbf{K}(H_n) <^+ \mathbf{I}(H_n : H_n) <^+ c_{\text{HM}} + \mathbf{K}(n) + c + k_n \\ n - \mathbf{K}(n) - (c_{\text{HM}} + c) <^+ k_n. \end{aligned}$$

Therefore the prefixes of \mathcal{H} do not exist in reality because, up to a small additive constant ($c_{\text{HM}} + c + O(1)$), the information content of H_n is in the range $(n - O(1), n + \mathbf{K}(n) + O(1))$ but by **IP** the smallest physical address to reach H_n is more than $n - \mathbf{K}(n)$.

7.2 Peano Arithmetic

IP can also be used in instances where $\alpha \neq \beta$, and one canonical example is to logic, and in particular Peano Arithmetic (PA). PA is a logic system that encodes statements of arithmetic through a set of initial axioms and a deduction system. Gödel proved that PA is incomplete, in that there are well formed formulas in the language of PA which are true but are unprovable in PA. Suppose we order every well formed formula of PA and let the infinite sequence L be defined such that its i th bit is 1 iff the i th formula of PA is true. Then L is undecidable, in that there is no algorithm that can compute it. However Gödel himself thought that there can be other means to produce true axioms of mathematics [Gö1]:

Namely, it turns out that in the systematic establishment of the axioms of mathematics, new axioms, which do not follow by formal logic from those previously established, again and again become evident. It is not at all excluded by the negative results mentioned earlier that nevertheless every clearly posed mathematical yes-or-no question is solvable in this way. For it is just this becoming evident of more and more new axioms on the basis of the meaning of the primitive notions that a machine cannot imitate.

However, as detailed in [Lev13], **IP** forbids such information leaks. The sequence L can be defined by a small mathematical formula of size n . Let β be any source of information with a reasonably small physical address of size k , such as the contents of an entire mathematical library. Then by **IP**, with $\alpha = L$, this information source will have negligible shared information with L (which encodes PA):

$$\mathbf{I}(\beta : L) < k + n + c.$$

7.3 Inductive Inference

The deficiency of randomness $\mathbf{d}(a|p)$ is a score that refutes the hypothesis that $a \in \mathbb{N}$ was generated by the model p , which is a probability measure over \mathbb{N} . If $p = \mathbf{m}$, defined in Section 2, then for all $a \in \mathbb{N}$, $\mathbf{d}(a|\mathbf{m}) = O(1)$. Thus there is no refutation to the statement: “ a was generated by \mathbf{m} ”. This implies \mathbf{m} acts as a universal “a-priori” distribution over \mathbb{N} .

In inductive inference, the goal is to describe the environment from a set of candidates, D . In its essence, this is the process of selecting an element $x \in \Sigma^*$ from a set $D \subseteq \Sigma^*$. The set D can be seen as the group of (very long) descriptions of the environment consistent with an observation. Let $\mathbf{m}(D) = \sum_{x \in D} \mathbf{m}(x)$, be the total a-priori probability of the members of D . In [EL11], it

was shown that the apriori probability of a non-exotic set of candidates D is focused on a single observation $x \in D$, with

$$\min_{x \in D} -\log \mathbf{m}(x) <^{\log} -\log \mathbf{m}(D) + \mathbf{I}(D; H).$$

We can apply **IP** to this inequality, and for this section only, we use $\mathbf{I}(x; \mathcal{H}) = \mathbf{K}(x) - \mathbf{K}(x|\mathcal{H})$ as the information term in **IP**. Let k be the size of a physical address for the candidates D . Let c_{HM} be the smallest mathematical description of the halting problem. By **IP**,

$$\left(\min_{x \in D} -\log \mathbf{m}(x)\right) + \log \mathbf{m}(D) \lesssim k + c_{\text{HM}} + c.$$

Thus observations in the physical world, (i.e. sets with small addresses) will have explanations (i.e. elements of the set) with a-priori probability concentrated on a single explanation. Due to the coding theorem, this element is equal to $\arg \min_{x \in D} \mathbf{K}(x)$, the simplest explanation of the given observation.

7.4 Outliers

The current paper represents another application of **IP**. This section shows that observable sequences in the physical world will have emergent outliers. The results of Section 3, along with Lemma 1 show us that finite sets of numbers will have outliers or large mutual information with the halting sequence. This can easily applied to individual infinite sequences of numbers, as shown in Theorem 5. For $\tau \in \mathbb{N}^{\mathbb{N}}$, $\langle \tau \rangle = \langle \tau[1] \rangle \langle \tau[2] \rangle \langle \tau[3] \rangle \dots$. Let $\tau(n)$ be the first 2^n unique numbers found in τ . The sequence τ is assumed to have an infinite amount of unique numbers, and represents a series of observations.

Theorem 5 *For probability p over \mathbb{N} and $\tau \in \mathbb{N}^{\mathbb{N}}$, let s_τ be the smallest number such that $\max_{a \in \tau(n)} \mathbf{d}(a|p) > n - 4\mathbf{K}(n) - s_\tau$. Then $s_\tau <^{\log} \mathbf{I}(\langle \tau \rangle : \mathcal{H}) + \mathbf{K}(p)$.*

Proof. By Corollary 1, Lemma 1, and the fact that $\mathbf{I}(x; \mathcal{H}) <^+ \mathbf{I}(\alpha : \mathcal{H}) + \mathbf{K}(x|\alpha)$,

$$\begin{aligned} n &< \max_{a \in \tau(n)} \mathbf{d}(a|p) + \mathbf{I}(\tau(n); \mathcal{H}) + \mathbf{K}(p) + 2\mathbf{K}(n) + O(\log(\mathbf{I}(\tau(n); \mathcal{H}) + \mathbf{K}(p))) \\ n &< \max_{a \in \tau(n)} \mathbf{d}(a|p) + 4\mathbf{K}(n) + \mathbf{I}(\langle \tau \rangle : \mathcal{H}) + \mathbf{K}(p) + O(\log(\mathbf{I}(\langle \tau \rangle : \mathcal{H}) + \mathbf{K}(p))) \\ n - 4\mathbf{K}(n) - (\mathbf{I}(\langle \tau \rangle : \mathcal{H}) + \mathbf{K}(p) + O(\log(\mathbf{I}(\langle \tau \rangle : \mathcal{H}) + \mathbf{K}(p)))) &< \max_{a \in \tau(n)} \mathbf{d}(a|p). \end{aligned}$$

□

Let k be the physical address of an infinite sequence of numbers τ . As defined above, the halting sequence \mathcal{H} can be described in a mathematical statement of size equal to some small constant c_{HM} . Then by Theorem 5, for some sequence $\tau \in \mathbb{N}^{\mathbb{N}}$ and probability p , **IP** states

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

Thus sequences τ with large s_τ , as defined in Theorem 5, will have large physical address. Thus it is hard to find physical sequences which do not have large outliers, and completely impossible to find sequences with no outliers. As the complexity of the probability p in the randomness deficiency term increases, the bounds loosen. Theorem 5 can be strengthened to characterize subsets of initial subsequences, in similar form to Theorem 4. One can also use **IP** to characterize sequences of infinite sequences or reals.

8 Uncomputable Sampling Methods

As mentioned in the introduction, a sampling method A , takes in a number n , and outputs, with probability 1, n unique numbers or infinite sequences. In [Eps21], it was shown that computable sampling methods produce outliers. In [Eps22], the bounds were improved and results were proven about sampling methods that can possibly not halt with positive probability. In this section, we show that discrete and continuous sampling methods that are uncomputable but whose encodings has finite information with the halting sequence will produce outliers.

An encoding $\langle A \rangle$ of a discrete sampling method A , is any infinite sequence, such that if it is on the input tape of the universal Turing machine U , and $\langle n, \omega \rangle$ is on the auxiliary tape, U outputs n elements using random seed ω and then halts. Halting will occur with uniform probability 1 over the random seeds. $\mathbf{I}(A : \mathcal{H}) = \inf_{\langle A \rangle} \mathbf{I}(\langle A \rangle : \mathcal{H})$.

Theorem 6 *For discrete (possibly uncomputable) sampling method A , computable probability p over \mathbb{N} , $\Pr_{D \sim A(2^n)}[n - k > \max_{a \in D} \mathbf{d}(a|p)] < 2^{-k + \mathbf{I}(A : \mathcal{H}) + O(\log n) + c_p}$.*

Proof. Let $d_D = \max_{a \in D} \mathbf{d}(a|p)$. For any $D \subset \mathbb{N}$, $|D| = 2^n$, by Corollary 1 and Lemma 1, we have

$$n < d_D + \mathbf{I}(D; \mathcal{H}) + O(\log(n) + \log(\mathbf{I}(D; \mathcal{H}))) + c_p.$$

So $n - d_D - O(\log n) < \mathbf{I}(D; \mathcal{H}) + c_p$. So

$$\mathbf{E}_{D \sim A(2^n)} \left[2^{n - d_D - O(\log n)} \right] < \mathbf{E}_{D \sim A(2^n)} \left[2^{\mathbf{I}(D; \mathcal{H})} \right] 2^{c_p}.$$

Let γ_n be a probability measure over Σ^∞ , where $\gamma_n(\langle D \rangle 0^\infty) = \Pr(A(2^n) = D)$. From Theorem 3,

$$\begin{aligned} & \mathbf{E}_{D \sim A(2^n)} \left[2^{n - d_D} \right] \\ & < \mathbf{E}_{D \sim A(2^n)} \left[2^{\mathbf{I}(\langle D \rangle 0^\infty : \mathcal{H})} \right] 2^{O(\log n) + c_p} \\ & < \mathbf{E}_{\alpha \sim \gamma_n} \left[2^{\mathbf{I}(\alpha : \mathcal{H})} \right] 2^{O(\log n) + c_p} \\ & < 2^{\mathbf{I}(\langle \gamma_n \rangle : \mathcal{H}) + O(\log n) + c_p} \\ & < 2^{\mathbf{I}(\langle n, A \rangle : \mathcal{H}) + O(\log n) + c_p} \\ & < 2^{\mathbf{I}(A : \mathcal{H}) + O(\log n) + c_p}. \end{aligned}$$

Thus we get the theorem statement, with

$$\Pr_{D \sim A(2^n)} \left[n - k > \max_{a \in D} \mathbf{d}(a|p) \right] < 2^{-k + \mathbf{I}(A : \mathcal{H}) + O(\log n) + c_p}.$$

□

Better bounds than $O(\log n)$ can be achieved at the cost of complicating the proof. A continuous sampling method, A , takes in a number n , and outputs n unique sequences, encoded as:

$$\omega_1[1]\omega_2[1] \dots \omega_n[1]\omega_1[2] \dots$$

The encoding of a continuous sampling method A , and its information with the halting sequence, $\mathbf{I}(A : \mathcal{H})$, follows analogously to the discrete case.

Theorem 7 *For (possibly uncomputable) continuous sampling method A , computable probability P over Σ^∞ , $\Pr_{D \sim A(2^n)}[n - k > \max_{\alpha \in D} \mathbf{D}(\alpha|P)] < 2^{-k + \mathbf{I}(A : \mathcal{H}) + O(\log n) + c_P}$.*

The proof follows almost identically to that of Theorem 6. We leave the details to the reader.

9 Discussion

Corollary 2 can also be generalized to arbitrary probability measures μ if the deficiency of randomness definition is changed to $\mathbf{D}(\alpha|P) = \sup_n -\log P(\alpha[0..n]) - \mathbf{K}(\alpha[0..n]|\langle P \rangle)$. This is the same for Theorem 4, with some minor technicalities. The paper proves results for discrete time dynamical systems. An open problem is how outliers manifest in continuous time dynamical systems. Another avenue of research involves the connections of outliers in dynamical systems to thermodynamics. A good starting point for research into this area would be the paper [Gac94], and the general research direction would be to prove properties of algorithmic fine grained entropy instead of randomness deficiency. The results in the paper is over the Cantor space. This sets the stage to characterize properties of computable dynamics over general spaces. Another application to **IP** is complete extensions of partial predicates and their relation to statistical learning theory, as discussed in [Eps21].

References

- [EL11] Samuel Epstein and Leonid Levin. On sets of high complexity strings. *CoRR*, abs/1107.1458, 2011.
- [Eps21] Samuel Epstein. All sampling methods produce outliers. *IEEE Transactions on Information Theory*, 67(11):7568–7578, 2021.
- [Eps22] Samuel Epstein. The outliers theorem revisited. *CoRR*, abs/2203.08733, 2022.
- [G61] 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.
- [Gac94] P. Gacs. The boltzmann entropy and randomness tests. In *Proceedings Workshop on Physics and Computation. PhysComp '94*, pages 209–216, 1994.
- [Gei12] Philipp Geiger. *Mutual information and Gödel incompleteness*. PhD thesis, Heidelberg University, 10 2012.
- [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.
- [Ver21] N. Vereshchagin. Proofs of conservation inequalities for levin’s notion of mutual information of 1974. *Theoretical Computer Science*, 856, 2021.