On Outliers

Samuel Epstein
samepst@jptheorygroup.org

October 25, 2023

Abstract

An outlier is an observation that is set apart from a population. There are many reasons that such anomalies occur, including measurement error and human error. This manuscript proves that outliers are ingrained into the nature of algorithms and dynamics. By modelling outliers using the discrete and continuous randomness paper, in this manuscript, it is proven that computable sampling methods have to produce outliers. But what about measurements of systems that are too complex to be considered algorithmic? One example is the global weather system. One can attest to the fact that there are many strange formations that occur!

To show that anomalies occur, one can use the Independence Postulate. The Independence Postulate is a finitary Church-Turing thesis, postulating that certain finite and infinite sequences cannot be easily be found with a short "physical address". In this manuscript the Independence Postulate is used to show that observations, a.k.a. infinite sequences of natural numbers, that do not have outliers have high physical addresses. In other words, observations with no outliers cannot be found in nature. In addition we show that outliers occur almost surely with respect to computable dynamics.

Contents

Chapter 1

Introduction

An outlier is a data point that varies noticeably from other data points in a sample or collection. There is no exact mathematical definition of what constitutes an outlier. Though there are known partial indicators, the determination of an outlier remains a subjective endeavor.

Outliers can have many causes, such as due to variability in system performance, human mistakes, instrument malfunctions, contamination from elements outside the population or by inherent standard deviations in populations.

In algorithmic information theory, outliers are precisely defined algorithmically with respect to computable probability measures over either natural numbers or infinite sequences. The probability measure represents the model, and natural numbers and infinite sequences are assumed to be data points with respect to these models. The level or score to which a data point is an outlier to a model (probability measure) is given by the *deficiency of randomness* function. It is defined by $\mathbf{d}(x|P) = \lfloor -\log P(x) \rfloor - \mathbf{K}(x|P)$, where x is the data point and P is the probability measure. The term $\mathbf{K}$ is the Kolmogorov complexity of a string. $\mathbf{d}(x|P)$ is the difference between length a string's P-code and its optimal description. If x is not in the support of P, then $\mathbf{d}(x|P) = \infty$. The function $\mathbf{d}$ is optimal, in the following manner.

Given a computable probability measure P over $\mathbb{N}$, an expectation bounded test is a function $d : \mathbb{N} \to \mathbb{R}_{\geq 0}$ that is lower semi-computable and

$$\sum_{x\in\mathbb{N}} P(x)2^{d(x)} \leq 1.$$

Typical numbers x of P will have a low test score. An expectation bounded test d is universal if for every expectation bounded test d', there is a $c_{d'} \in \mathbb{N}$, such that for all $x \in \mathbb{N}$, $d(x) + c_{d'} > d'(x)$.

It can be shown that the deficiency of randomness, $\mathbf{d}$, is a universal expectation test, in that there is a constant $c \in \mathbb{N}$, where for any expectation bounded test d, for any $x \in \mathbb{N}$,

$$\mathbf{d}(x|P) + \mathbf{K}(d|P) > d(x) - c.$$

In this manuscript, we show that outliers are emergent in sampling methods, probabilistic algorithms, dynamics, and the physical world.

Chapter 2

Conventions

We use $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, $\{0,1\}$, $\{0,1\}^*$, and $\{0,1\}^\infty$ to represent natural numbers, integers, rational numbers, reals, bits, finite strings, and infinite strings. Let $X_{\geq 0}$ and $X_{>0}$ be the sets of non-negative and of positive elements of X. The length of a string $x{\in}\{0,1\}^n$ is denoted by $\|x\| = n$. The removal of the last bit of a string is denoted by $(p0^-){=}(p1^-){=}p$, for $p \in \{0,1\}^*$. For the empty string $\emptyset$, $(\emptyset^-)$ is undefined. We use $\{0,1\}^{*\infty}$ to denote $\{0,1\}^*{\cup}\{0,1\}^\infty$, the set of finite and infinite strings. For $x \in \{0,1\}^{*\infty}$, $y \in \{0,1\}^{*\infty}$, we say $x \sqsubseteq y$ if $x = y$ or $x \in \{0,1\}^*$ and $y = xz$ for some $z \in \{0,1\}^{*\infty}$. Also $x \sqsubset y$ if $x \sqsubseteq y$ and $x \neq y$. The ith bit of a string $x \in \{0,1\}^{*\infty}$ is denoted by $x[i]$. The first n bits of a string $x \in \{0,1\}^{*\infty}$ is denoted by $x[0..n]$. The indicator function of a mathematical statement A is denoted by $[A]$, where if A is true then $[A] = 1$, otherwise $[A] = 0$. The size of a finite set S is denoted to be $|S|$. We use $\langle x\rangle$ to represent a self delimiting code for $x \in \{0,1\}^*$, such as $1^{\|x\|}0x$. The self delimiting code for a finite set of strings $\{a_1,\ldots,a_n\}$ is $\langle\{a_1,\ldots,a_n\}\rangle = \langle n\rangle\langle a_1\rangle\langle a_2\rangle\ldots\langle a_n\rangle$. For two infinite strings α and β, $\langle\alpha,\beta\rangle = \alpha_1\beta_1\alpha_2\beta_2\ldots$. For sets Z of infinite strings, $Z_{\leq n} = \{\alpha[0..n] : \alpha\in Z\}$ and $\langle Z\rangle = \langle Z_{\leq 1}\rangle\langle Z_{\leq 2}\rangle\langle Z_{\leq 3}\rangle\ldots$.

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 $\overset{*}{<}f$, $\overset{*}{>}f$ denote $< f/O(1)$, $> f/O(1)$. The terms $\overset{*}{=}f$ denotes $\overset{*}{<}f$ and $\overset{*}{>}f$. 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 $Q : \mathbb{N} \to \mathbb{R}_{\geq 0}$ over natural numbers. The support of a measure Q is the set of all elements whose Q value is positive, with $\mathrm{Supp}(Q) = \{a : Q(a) > 0\}$. A measure is elementary if its support is finite and its range is a subset of $\mathbb{Q}$. We say Q is a semi-measure if $\sum_a Q(a) \leq 1$. We say that Q is 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 $x \in \{0,1\}^*$ and auxiliary input $y \in \{0,1\}^{*\infty}$. T is prefix-free if for all $x, s \in \{0,1\}^*$ with $s \neq \emptyset$, and $y \in \{0,1\}^{*\infty}$, either $T_y(x) = \perp$ or $T_y(xs) = \perp$. The complexity of $x \in \{0,1\}^*$ 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 $x \in \{0,1\}^*$ and $y \in \{0,1\}^{*\infty}$. For example, one can take a universal prefix-free algorithm U, where for each prefix-free algorithm T, there exists $t \in \{0,1\}^*$, with $U_y(tx) = T_y(x)$ for all $x \in \{0,1\}^*$ and $y \in \{0,1\}^{*\infty}$. The function $\mathbf{K}(x|y)$, defined to be $\mathbf{K}_U(x|y)$, is the Kolmogorov complexity of $x \in \{0,1\}^*$ relative to $y \in \{0,1\}^{*\infty}$. 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.

A function $f:\mathbb{N}\to\mathbb{R}$ is computable if there is a total recursive function $g(x,n)$ over all $x\in\mathbb{N}$ and $n\in\mathbb{N}$ where $|f(x)-g(x,n)|<1/n$. The complexity of such a computable function f, is $\mathbf{K}(f)$, the minimal length of a U-program to compute f. A function $f:\mathbb{N}\to\mathbb{R}$ is lower semi-computable if the set $S=\{(x,r):x\in\mathbb{N},r\in Q,r<f(x)\}$ is recursively enumerable. If f is not computable but lower semi-computable, then its complexity $\mathbf{K}(f)$ is equal to the size of smallest U-program that on input x, enumerates $\{r:f(x)>r\}$.

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 $z\in\{0,1\}^*$ is $\mathbf{I}(x:y\,|\,z)=\mathbf{K}(x|z)+\mathbf{K}(y|z)-\mathbf{K}(\langle x,y\rangle|z)=^+\mathbf{K}(x|z)-\mathbf{K}(x|\langle y,\mathbf{K}(y|z),z\rangle)$. 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 halting sequence $\mathcal{H}\in\{0,1\}^\infty$ is the infinite string where $\mathcal{H}[i]=[U(i)\neq\perp]$ for all $i\in\mathbb{N}$. As mentioned in the introduction, the amount of information that $a\in\mathbb{N}$ has with $\mathcal{H}$ is denoted by $\mathbf{I}(a:\mathcal{H})=\mathbf{K}(a)-\mathbf{K}(a|\mathcal{H})$. Stochasticity is $\mathbf{Ks}(a|b)=\min\{\mathbf{K}(Q|b)+3\log\max\{\mathbf{d}(a|Q,b),1\}$.

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

Proof.

$$\mathbf{I}(a;\mathcal{H})=\mathbf{K}(a)-\mathbf{K}(a|\mathcal{H})>^+\mathbf{K}(a,f(a))-\mathbf{K}(a,f(a)|\mathcal{H})-\mathbf{K}(f).$$

The chain rule $(\mathbf{K}(x,y)=^+\mathbf{K}(x)+\mathbf{K}(y|x,\mathbf{K}(x)))$ applied twice results in

$$\begin{aligned}\mathbf{I}(a;\mathcal{H})+\mathbf{K}(f)&>^+\mathbf{K}(f(a))+\mathbf{K}(a|f(a),\mathbf{K}(f(a)))-(\mathbf{K}(f(a)|\mathcal{H})+\mathbf{K}(a|f(a),\mathbf{K}(f(a)|\mathcal{H}),\mathcal{H})\\&=^+\mathbf{I}(f(a);\mathcal{H})+\mathbf{K}(a|f(a),\mathbf{K}(f(a)))-\mathbf{K}(a|f(a),\mathbf{K}(f(a)|\mathcal{H}),\mathcal{H})\\&=^+\mathbf{I}(f(a);\mathcal{H})+\mathbf{K}(a|f(a),\mathbf{K}(f(a)))-\mathbf{K}(a|f(a),\mathbf{K}(f(a)),\mathbf{K}(f(a)|\mathcal{H}),\mathcal{H})\\&>^+\mathbf{I}(f(a);\mathcal{H}).\end{aligned}$$

□

2.1 Infinite Sequences

In this section, we define the deficiency of randomness $\mathbf{D}$ of infinite sequence. This notion will be used in the no-go sampling theorems over infinite sequences.

Theorem. ([GÍ3]) *For computable probability measure P over $\{0,1\}^\infty$, there exists a universal integrable test $\mathbf{D}:\{0,1\}^\infty\to\mathbb{R}_{\geq 0}\cup\infty$, where for all other integrable tests D,*

$$D(\alpha)<^+\mathbf{D}(\alpha|P)+\mathbf{K}(D|P).$$

As shown in the following theorem, any such universal integrable test $\mathbf{D}$ is equal, up to an additive constant, to a supremum of a term that uses the finite prefix of an infinite sequence.

Theorem. ([GÍ3]) *For universal integrable test $\mathbf{D}$ for computable probability measure P over $\{0,1\}^\infty$,*

$$\mathbf{D}(\alpha|P)=^+\sup_{n\in\mathbb{N}}-\log P(\alpha[0..n])-\mathbf{K}(\alpha[0..n]|P),$$

where the constant depends on P.

This justifies the following definition.

Definition 1 (Deficiency of Randomness of an Infinite Sequence)
$\mathbf{D}(\alpha|P) = \sup_{n\in\mathbb{N}} -\log P(\alpha[0..n]) - \mathbf{K}(\alpha[0..n]|P)$.

As we look at sampling with respect to infinite sequences, we will need an information function between infinite sequences, and more specifically the amount of information that a specific sequence α has with the halting sequence $\mathcal{H}$. We use the symmetric function $\mathbf{I} : \{0,1\}^\infty \times \{0,1\}^\infty \to \mathbb{R}$, where

Definition 2 (Information of Infinite Sequences) *For $\alpha, \beta \in \{0,1\}^\infty$, and $c \in \{0,1\}^*$,*
$\mathbf{I}(\alpha : \beta|c) = \log \sum_{x,y\in\{0,1\}^*} \mathbf{m}(x|c,\alpha)\mathbf{m}(y|c,\beta)2^{\mathbf{I}(x:y|c)}$.

This function was introduced in [Lev74]. The following theorem was stated in [Lev74], and a proof of it can be found in [Ver21].

Theorem 1 *Assume that a family P_ρ, $\rho \in \Omega$, of probability distributions on Ω is fixed. Assume that there is a Turing machine T that for all ρ computes P_ρ having oracle access to ρ. Then for all α, $\rho \in \Omega$, there is a probability bounded (and even expectation) P_ρ-test $t_{\alpha,\rho,T}$ such that*

$$\mathbf{I}(\langle\rho,\omega\rangle : \alpha) \leq \mathbf{I}(\rho : \alpha) + t_{\alpha,\rho,T}(\omega) + c_T,$$

for all $\omega \in \Omega$, where c_T does not depend on ρ, α, ω.

In [Gei12], it is shown that the above theorem implies the following.

Corollary 1 *Assume that a family P_ρ, $\rho \in \Omega$, of probability distributions on Ω is fixed. Assume that there is a Turing machine T that for all ρ computes P_ρ having oracle access to ρ. Then for all α, $\rho \in \Omega$, there is a probability bounded (and even expectation) P_ρ-test $t_{\alpha,\rho,T}$ such that*

$$\mathbf{I}(\langle\rho,\omega\rangle : \alpha) \leq \mathbf{I}(\rho : \alpha) + t_{\alpha,\rho,T}(\omega) + c_T,$$

for all $\omega \in \Omega$, where c_T does not depend on ρ, α, ω.

Theorem 2 *Let $(P_\alpha)_{\alpha\in\{0,1\}^\infty}$ be a family of uniformly α-computable continuous probability measures. Then for all $\alpha, \beta \in \{0,1\}^\infty$ we have*

$$P_\alpha(\{\gamma \in \{0,1\}^\infty : \mathbf{I}(\langle\alpha,\gamma\rangle : \beta) - \mathbf{I}(\alpha : \beta) > m\}) \leq 2^{-m+c_{\alpha,\beta}},$$

where $c_{\alpha,\beta}$ is a positive constant dependent solely on α and β.

Corollary 2 $\Pr_\mu(\mathbf{I}(\alpha : \mathcal{H}) > n) \overset{*}{<} 2^{-n+\mathbf{K}(\mu)}$.

In addition [Gei12] contains a short proof for the following theorem.

Theorem 3 *For partial recursive $f : \{0,1\}^\infty \to \{0,1\}^\infty$, $\alpha, \beta \in \{0,1\}^\infty$, $\mathbf{I}(f(\alpha) : \beta) <^+ \mathbf{I}(\alpha : \beta) + \mathbf{K}(f)$.*

Chapter 3

All Sampling Methods Produce Outliers

3.1 Discrete Sampling Theorem

A sampling method A is a probabilistic function that maps an integer N with probability 1 to a set containing N different strings. Let $P = P_1, P_2, \dots$ be a sequence of measures over strings. For example, one may choose $P_1 = P_2 \dots$ or choose P_n to be the uniform measure over n-bit strings. A conditional probability bounded P-test is a function $t : \{0,1\}^* \times \mathbb{N} \to \mathbb{R}_{\geq 0}$ such that for all $n \in \mathbb{N}$ and positive real number r, we have $P_n(\{x : t(x|n) \geq r\}) \leq 1/r$. If $P_1, P_2, \dots$ is uniformly computable, then there exists a lower-semicomputable such P-test t that is "maximal" (i.e., for which $t' \leq O(t)$ for every other such test t'). We fix such a t, and let $\overline{\mathbf{d}}_n(x|P) = \log t(x|n)$.

Lemma 2 *Let P be a computable measure on strings and let A be a sampling method. For all integers M and N, there exists a finite set $S \subset \{0,1\}^*$ such that $P(S) \leq 2M/N$, and with probability strictly more than $1 - 2e^{-M}$: $A(N)$ intersects S.*

Proof. We show that some possibly infinite set S satisfies the conditions, and thus, some finite subset also satisfies the conditions due to the strict inequality. We use the probabilistic method: we select each string to be in S with probability M/N and show that 2 conditions are satisfied with positive probability. The expected value of $P(S)$ is M/N. By the Markov inequality, the probability that $P(S) > 2M/N$ is at most 1/2. For any set D containing N strings, the probability that S is disjoint from D is

$$(1 - M/N)^N < e^{-M}.$$

Let Q be the measure over N-element sets of strings generated by the sampling algorithm $A(N)$. The left-hand side above is equal to the expected value of

$$Q(\{D : D \text{ is disjoint from } S\}).$$

Again by the Markov inequality, with probability greater than 1/2, this measure is less than $2e^{-M}$. By the union bound, the probability that at least one of the conditions is violated is less than $1/2 + 1/2$. Thus, with positive probability a required set is generated, and thus such a set exists.□

Theorem 4 *Let $P = P_1, P_2 \dots$ be a uniformly computable sequence of measures on strings and let A be a sampling method. There exists $c \in \mathbb{N}$ such that for all n and k:*

$$\Pr\left(\max_{a\in A(2^n)} \overline{\mathbf{d}}_n(a|P) > n-k-c\right) \geq 1-2e^{-2^k}.$$

Proof. We now fix a search procedure that on input N and M finds a set $S_{N,M}$ that satisfies the conditions of Lemma 2. Let $t'(a|n)$ be the maximal value of $2^n/2^{k+2}$ such that $a \in S_{2^n,2^k}$ for some integer k. By construction, t' is a computable probability bound test, because $P(\{x : t'(x|n) = 2^\ell\}) \leq 2^{-\ell-1}$, and thus $P(t'(x|n) \geq 2^\ell) \leq 2^{-\ell-1} + 2^{-\ell-2} + \dots$ With the given probability, the set $A(2^n)$ intersects $S_{2^n,2^k}$. For any number a in the intersection, we have $t'(x|n) \geq 2^{n-k-2}$, thus by the optimality of t and definition of $\overline{\mathbf{d}}$, we have $\overline{\mathbf{d}}_n(a|P) > n-k-O(1)$. □

An incomplete sampling method A takes in a natural number N and outputs, with probability $f(N)$, a set of N numbers. Otherwise A outputs $\perp$. f is computable.

Corollary 3 *Let $P = P_1, P_2 \dots$ be a uniformly computable sequence of measures on strings and let A be an incomplete sampling method. There exists $c \in \mathbb{N}$ such that for all n and k:*

$$\Pr_{D=A(n)}\left(D \neq \perp \textit{ and } \max_{a\in D} \overline{\mathbf{d}}_n(a|P) \leq n-k-c\right) < 2e^{-2^k}.$$

3.2 Continuous Sampling Method

Let $\mu = \mu_1, \mu_2, \dots$ be a uniformly computable sequence of measures over infinite sequences. Similar way as for strings in the introduction, the randomness deficiency $\overline{\mathbf{D}}_n(\omega|\mu)$ for sequences ω is defined using lower-semicomputable functions $\{0,1\}^\infty \times \mathbb{N} \to \mathbb{R}_{\geq 0}$. A continuous sampling method C is a probabilistic function that maps, with probability 1, an integer N to an infinite encoding of N different sequences.

Theorem 5 *There exists $c \in \mathbb{N}$ where for all n:*

$$\Pr\left(\max_{\alpha\in C(2^n)} \overline{\mathbf{D}}_n(\alpha|\mu) > n-k-c\right) \geq 1-2.5e^{-2^k}.$$

Proof. For $D \subseteq \{0,1\}^\infty$, $D_m = \{\omega[0..m] : \omega \in D\}$. Let $g(n) = \arg\min_m \Pr_{D=C(n)}(|D_m| < n) < 0.5e^{-2^n}$ be the smallest number m such that the initial m-segment of $C(n)$ are sets of n strings with very high probability. g is computable, because C outputs a set of distinct infinite sequences with probability 1. For probability ψ over $\{0,1\}^\infty$, let $\psi^m(x) = [|x| = m]\psi(\{\omega : x \sqsubset \omega\})$. Let $\mu^g = \mu_1^{g(1)}, \mu_2^{g(2)}, \dots$ be a uniformly computable sequence of discrete probability measures and let A be a discrete incomplete sampling method, where for random seed $\omega \in \{0,1\}^\infty$, $A(n,\omega) = C(n,\omega)_{g(n)}$

if $|C(n,\omega)_{g(n)}| = n$; otherwise $A(n,\omega) =\perp$. So $\Pr[A(n) =\perp] < 0.5e^{-2^n}$.

$$\begin{aligned}
&\Pr\left(\max_{\alpha\in C(2^n)} \overline{\mathbf{D}}_n(\alpha|\mu) \leq n-k-O(1)\right)\\
\leq &\Pr_{Z=C(2^n)}\left((|Z_{g(n)}| < 2^n) \text{ or } (|Z_{g(n)}| = 2^n \text{ and } \max_{\alpha\in Z} \overline{\mathbf{D}}_n(\alpha|\mu) \leq n-k-O(1)\right)\\
\leq &\Pr_{D=A(2^n)}\left(D=\perp \text{ or } (D\neq\perp \text{ and } \max_{x\in D} \overline{\mathbf{d}}_n(x|\mu^g) \leq n-k-O(1))\right)\\
< &0.5e^{-2^n} + 2e^{-2^k} \qquad (3.1)\\
\leq &2.5e^{-2^k},
\end{aligned}$$

where Equation 3.1 is due to Corollary 3. □

3.3 Necessity of Double Exponential

Theorem 4 showed that the probability that $A(2^n)$ contains no strings of randomness deficiency less than $n-k$ decreases double exponentially in k. We show that at least a double exponential probability is required for $k = n - O(1)$. Let P_n be the uniform measure on $(n+2)$-bit strings. The algorithm A that on input 2^n generates a random set of 2^n strings of length $n+2$ satisfies

$$\Pr\left(\forall x \in A(2^n) : \overline{\mathbf{d}}_n(x|P) \leq 2\right) \geq 2^{-2^n}.$$

The reasoning for this is as follows. For at most a quarter of the $(n+2)$-bit strings, we have $\overline{\mathbf{d}}_n(x|P) \geq 3$, by definition of a probability bounded test t. A random selection of $N = 2^n$ different $(n+2)$-bit strings, contains no such string with a probability of at least 2^{-N}. We consider the following situation. In a bag with $4N$ balls, N balls are marked. One selects N balls one by one. We consider the probability that no marked ball is drawn if previously no marked ball was drawn. The smallest probability appears at the last draw when there are $T = 4N - (N-1)$ balls in the bag. This probability is $(T-N)/T \geq 1/2$.

3.4 Partial Sampling Methods

A partial sampling method is a sampling method that can output with probability less than 1. Theorem 4 does not hold for partial sampling methods B. Let P_n be the uniform measure on $(n+1)$-bit strings. Let $\#B(N)$ represent the event that B halts and outputs a set of size N. We present a partial sampling method B for which

$$\Pr\left(\#B(2^n) \text{ and } \forall x \in B(2^n) : \overline{\mathbf{d}}_n(x|P) \leq 1\right) \geq 2^{-n}.$$

For at most half of the $(n+1)$-bit strings, we have $\overline{\mathbf{d}}_n(x|P) \geq 2$. On input 2^n,the partial sampling method B generates a random natural number s bounded by 2^n, searches for s strings x of length $n+1$ with $\overline{\mathbf{d}}_n(x|P) \geq 2$, and outputs 2^n other $(n+1)$-bit strings. For some s, this search may never terminate. If A chooses to be precisely equal to the number of strings satisfying the condition, then it outputs only strings with deficiency at most 1, and the claim is proven. However partial sampling methods do exhibit the following properties

Theorem 6 *Let $P = P_1, P_2, \ldots$ be a uniformly computable sequence of measures and B be a partial sampling method, where $\#B(N)$ represents the event that $B(N)$ terminates and outputs a set of N strings.*

$$\Pr\left(\#B(N) \text{ and } \forall x \in B(2^n) : \overline{\mathbf{d}}_n(x|P) \leq n-k\right) \leq O(k2^{-k}).$$

Proof. Let Q be the lower-semicomputable semimeasure over sets of size 2^n such that $Q(D)$ equals the probability that $B(N) = D$. We show that

$$\Pr\left(\#B(N) \text{ and } \forall x \in B(2^n) : \overline{\mathbf{d}}_n(x|P) \leq n - k + \log k + O(1)\right) \leq O(2^{-k}).$$

This result is followed by a redefinition of k. We write Q as a uniform mixture over at most 2^k measures Q_i with finite support, and one lower semi-computable semimeasure Q_*:

$$Q = 2^{-k}\left(Q_1 + Q_2 + \dots Q_f + Q_*\right).$$

With $f \leq 2^k$, we assume that the finite descriptions of $Q_1, \dots, Q_f$ are enumerated one by one by a program (that may never terminate). For each enumerated measure Q, we search for a set S_i that satisfies the conditions of Lemma 2 for $M = k$. Let $S = \bigcup_{i \leq f} S_i$. Also, $P(S) \leq k2^{k+1-n}$; thus every element in S satisfies $\overline{\mathbf{d}}_n(x|P) \geq n - k + \log k + O(1)$.

The probability that $A(2^n)$ produces a set that does not contain such an element is at most $2^{-k} + 2e^{-k}$ because we can equivalently generate a set D by randomly selecting j from the list $[1, \dots, f, *, \infty]$ with probabilities $[2^{-k}, \dots, 2^{-k}, 2^{-k}r, 1 - (f + r)2^{-k}]$ and generating a random set D from Q_j if $j \neq \infty$ and letting D be undefined otherwise. The probability that D is defined and does not contain an element from S is at most the probability $j = *$, which is $\leq 2^{-k}$, plus the probability that $j \in \{1, \dots, f\}$ times $2e^{-k}$. □

Chapter 4

The Independence Postulate

In this chapter, we revisit the celebrated Church-Turing thesis (CT) and define the Independence Postulate (IP), introduced in [Lev84, Lev13]. CT relates mechanical methods to functions computed from Turing machines. A method, M, for achieving some desired result is "effective" or "mechanical" if it can be carried out by a human with a pencil and paper. More formally,

1. M is set out in terms of a finite number of exact instructions (each instruction being expressed by means of a finite number of symbols).
2. M will, if carried out without error, produce the desired result in a finite number of steps.
3. M can (in practice or in principle) be carried out by a human being unaided by any machinery except paper and pencil.
4. M demands no insight, intuition, or ingenuity, on the part of the human being carrying out the method.

The Church-Turing thesis states

> **CT:** *A method is effective if and only if it can be computed by a Turing machine.*

One well known variant of CT is the physical Church-Turing thesis, which states *all physically computable functions are Turing-computable.* However there are several drawbacks associated with CT. The notion of an "effective method" is vague, admitting multiple different interpretations. On such early assessment of this fact can be found in [Kle52],

> *Since our original notion of effective calculability of a function ... is a somewhat vague intuitive one, the thesis cannot be proved. ... While we cannot prove Church's thesis, since its role is to delimit precisely an hitherto vaguely conceived totality, we require evidence.*

Turing himself had reservations about his thesis, [Tur36]

> *...fundamentally, appeals to intuition, and for this reason rather unsatisfactory mathematically.*

IP is an unprovable inequality on the information measure of two sequences. Among other applications, 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". IP provides a solution to the concerns of the somewhat vague formulation of CT. The statement of the IP is as follows [Lev13].

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

We take $\mathbf{I}$ to be the information term of Definition 2. Whereas IP is simpler, CT is more abstract. IP is supported by the so-called Independence Conservation Inequalities (Section 4.3). IP was succinctly described in a single page. In this paper, we expand upon the arguments in [Lev13] to make them accessible for a general audience. Section 4.1 details how IP is applied to finite sequences. Section 4.2 describes the applications of IP to logic. IP is a statement in the field of algorithmic information theory (AIT), but no prior knowledge of AIT is required by the readers.

4.1 Non-Recursive Finite Sequences

One consequence of IP is a finite version of the Church-Turing Thesis (CT). This advantage was mentioned by L. A. Levin in [Lev13],

> *IP is simpler, CT more abstract. All sequences we ever see are computable just by being finite: CT is useless for them! IP works equally well for finite and infinite sequences.*

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 $\alpha=\beta\in\{0,1\}^*$ is a finite sequence which has a non-recursive description of length $\mathbf{NR}(\alpha)$ that is much shorter than its recursive description $\mathbf{K}(\alpha)$, with $\mathbf{NR}(\alpha)\ll\mathbf{K}(\alpha)$. Let k be the shortest physical address of α. Then by IP, with $\beta=\alpha$,

$$\begin{aligned}\mathbf{K}(\alpha) &<^+ \mathbf{I}(\alpha:\alpha) <^+ k+\mathbf{NR}(\alpha)+c\\ \mathbf{K}(\alpha)-\mathbf{NR}(\alpha)-c &<^+ k. \qquad (4.1)\end{aligned}$$

Thus k is large and α cannot be easily located in the physical world. The only sequences α with short physical addresses must have $\mathbf{NR}(\alpha)\approx\mathbf{K}(\alpha)$. Thus if a sequence $x\in\{0,1\}^*$ is mathematical, with $\mathbf{NR}(x)\ll\|x\|$, then it must be algorithmic to be physically obtainable, that is, produced from a simple program, with $\mathbf{K}(x)\approx\mathbf{NR}(x)\ll\|x\|$. For example a string represention of 9999^{9999} is mathematical, algorithmic, and physically obtainable. In general, not all sequences generated are algorithmic, take any typical outcome of the rolling of random dice.

4.1.1 Prefixes of the Halting Sequence

A canonical example of the inequality in Equation 4.1 is prefixes of the halting sequence. The halting sequence $\mathcal{H}\in\{0,1\}^\infty$ is the unique sequence defined by $\mathcal{H}[i]=1$ iff $U(i)$ halts. Let

$\mathcal{H}_n \in \{0,1\}^*$ be the finite sequence that is the prefix of size 2^n of $\mathcal{H}$. It is well known that

$$\mathbf{K}(\mathcal{H}_n) \in (n - O(1), n + \mathbf{K}(n) + O(1)).$$

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 $\mathcal{H}_n$ can be defined by a mathematical statement of size $<^+ c_{\mathrm{HM}} + \mathbf{K}(n)$. So by IP applied to $\alpha = \beta = \mathcal{H}_n$ where k_n is the smallest physical address of $\mathcal{H}_n$,

$$\begin{aligned} n <^+ \mathbf{K}(\mathcal{H}_n) <^+ \mathbf{I}(\mathcal{H}_n : \mathcal{H}_n) &<^+ c_{\mathrm{HM}} + \mathbf{K}(n) + c + k_n \\ n - \mathbf{K}(n) - c_{\mathrm{HM}} - c &<^+ k_n. \end{aligned}$$

Therefore the prefixes of $\mathcal{H}$ of size 2^n have physical address of size at least $n - O(\log n)$, and thus are not physically obtainable.

4.2 Logic

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 ith bit is 1 iff the ith 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 [G̈61]:

> *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):

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

More generally, in [Lev13], it was shown that every consistent completion β of PA has $\mathbf{I}(\beta : \mathcal{H}) = \infty$. Then by IP, since $\mathcal{H}$ is represented by a formula of size c_{HM}, no consistent completion β has a finite physical address.

4.3 Information Conservation Inequalities

IP is an upper bound on the information between two sequences. The set of all true statements in arithmetic has no information about the stock market. Or the halting problem has nothing to say about any easily accessible series of physical of measurements.

This leaves open the possibility of deterministic or randomized processing to increase the amount of information that the sequences have. For example, one such method is to select statements of arithmetic with probability .5, in hopes of gleaning information about the next stock market crash. However the door to such circumventions is closed due to Independence Conservatism Inequalities (ICI) [Lev74, Lev84], which complements IP.

Whereas IP is an unprovable postulate, ICI are provable statements in the field of algorithmic information theory that says target information cannot be increased. The origins of ICI are in data processing inequalities in classical information theory, detailed in [CT91]. For two random variables $\mathcal{X}$ and $\mathcal{Y}$,

$$\mathcal{I}(\mathcal{X} : T(\mathcal{Y})) \leq \mathcal{I}(\mathcal{X} : \mathcal{Y}).$$

The term $\mathcal{I}$ is the classical information measure between two variables and T is any local processing done on the random variable $\mathcal{Y}$. Theorems in classical information theory often have equivalents in algorithmic information theory, and this is the case for the data processing inequality. The ICI for deterministic processing is

$$\text{For sequences } \alpha \text{ and } \beta\text{, computable function } f\text{, } \mathbf{I}(f(\alpha) : \beta) <^{+} \mathbf{I}(\alpha : \beta) + \mathbf{K}(f).$$

A function can add some mutual information between two sequences, but no more than the complexity of the function. There is also a randomized ICI. There are several forms, with the following inequality being one such instance [Lev74, Gei12, Ver21]. Let f be a function that transforms a sequence β using a random seed ω. Let $\mathcal{U}$ be the uniform measure over $\{0,1\}^{\infty}$.

$$\mathbf{E}_{\omega \sim \mathcal{U}}[\mathbf{I}(f(\beta, \omega) : \alpha)] <^{+} \mathbf{I}(\beta : \alpha) + \mathbf{K}(f).$$

Thus ICI prevents the processing of data to gain more target. As L. A. Levin states,

torturing an uniformed witness cannot give information about the crime.

Another application of ICI discussed in [Lev13] is to processes, represented by infinite sequences. A process ω_1 is "explained" by a simpler process ω_2 if there is some computable function f_1 such that $f_1(\omega_2) = \omega_1$. In CS, we say ω_1 reduces to ω_2. Say $\mathbf{I}(\omega_1 : \beta) = \infty$, where β is specified by a finite mathematical sequence, and ω_1 is a process. Assume ω_1 is a complicated process, and can be explained by a series of reductions to simpler process

$$\omega_1 \leftarrow_{f_1} \omega_2 \leftarrow_{f_2} \omega_3 \cdots \leftarrow_{f_{n-1}} \omega_{f_n}.$$

Assume ω_n is a "simple" process, admitting no further meaningful reduction. This notion of a simple, unexplainable process is a subjective one, as a process can always be reduced to another one. Thus by ICI, $\mathbf{I}(\omega_n : \beta) = \infty$, and by IP, ω_n cannot be found in nature, as all its physical addresses have infinite length. Thus to recap, using ICI, complicated processes with unlimited target information can be reduced to unexplainable simple processes and by using IP can be shown to not exist in nature.

4.4 Discussion

Further work can be done by determining more constructs that have high mutual information with the halting sequence. Also an in depth analysis could be done to more rigorously define the notion of a "physical address". For example, physical addresses must be sufficiently global. Take for example 2^n locations around the world, each given a number from 1 to 2^n. Then one such location has at least $\sim n$ bits of mutual information with the halting sequence, which can be reached by that location with a physical address of size $O(1)$, causing an information leak.

Chapter 5

Outliers in the Physical World

In Chapter 3, it has been proved 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, detailed in Chapter 4, this open issue is addressed. Outliers must occur in the physical world. In this section, we model observations as infinite sequences of natural numbers or reals.

5.1 Observations as Natural Numbers

Lemma 3 *For probability p over $\mathbb{N}$, $D\subset\mathbb{N}$, $|D| = 2^s$, $s < \max_{a\in D}\mathbf{d}(a|p) + \mathbf{Ks}(D) + \mathbf{K}(s) + O(\log\mathbf{K}(s,p))$.*

Proof. We relativize the universal Turing machine to $\langle s,p\rangle$. Let Q be a probability measure that realizes $\mathbf{Ks}(D)$, with $d = \max\{\mathbf{d}(D|Q),1\}$. Let $F\subseteq\mathbb{N}$ be a random set where each element $a\in\mathbb{N}$ is selected independently with probability $cd2^{-s}$, where $c\in\mathbb{N}$ is chosen later. $\mathbf{E}[p(F)]\leq cd2^{-s}$. Furthermore

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

Thus finite $W\subset\mathbb{N}$ can be chosen such that $p(W)\leq 2cd2^{-s}$ and $Q(\{G : |G| = 2^s, G\cap W=\emptyset\})\leq e^{1-cd}$. $D\cap W\neq\emptyset$, otherwise, using the Q-test, $t(G) = e^{cd-1}$ if $(|G| = 2^s, G\cap W=\emptyset)$ and $t(G) = 0$ otherwise, 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. Thus there is an $a\in D\cap W$, where

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

Making the relativization of $\langle s,p\rangle$ explicit,

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

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. Assuming τ has an infinite amount of unique numbers, $\tau(n)$ is the first 2^n unique numbers of τ.

Theorem 7 *For probability p over $\mathbb{N}$, $\tau\in\mathbb{N}^{\mathbb{N}}$, let $s_{\tau,p} = \sup_n\left(n - 3\mathbf{K}(n) - \max_{a\in\tau(n)}\mathbf{d}(a|p)\right)$. Then $s_{\tau,p} <^{\log} \mathbf{I}(\langle\tau\rangle : \mathcal{H}) + O(\log\mathbf{K}(p))$.*

Proof. By Lemmas 3 and 15, and the fact that $\mathbf{I}(x;\mathcal{H}) <^{+} \mathbf{I}(\alpha : \mathcal{H}) + \mathbf{K}(x|\alpha)$,

$$\begin{aligned}
n &< \max_{a\in\tau(n)}\mathbf{d}(a|p) + \mathbf{I}(\tau(n);\mathcal{H}) + +\mathbf{K}(n) + O(\log\mathbf{I}(\tau(n);\mathcal{H}) + \log\mathbf{K}(p) + \log\mathbf{K}(n)),\\
n &< \max_{a\in\tau(n)}\mathbf{d}(a|p) + 2\mathbf{K}(n) + \mathbf{I}(\langle\tau\rangle : \mathcal{H}) + O(\log\mathbf{I}(\langle\tau\rangle : \mathcal{H}) + \log\mathbf{K}(p) + \log\mathbf{K}(n)),\\
n - 3\mathbf{K}(n) - &\max_{a\in\tau(n)}\mathbf{d}(a|p) <^{\log} \mathbf{I}(\langle\tau\rangle : \mathcal{H}) + O(\log\mathbf{K}(p)).\square
\end{aligned}$$

Let k be a physical address of τ. $\mathcal{H}$ can be described by a small mathematical statement. By Theorem 7 and **IP**, there is a small constant c where

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

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

5.2 Outliers as Reals

Let $\Omega = \sum\{2^{-\|p\|} : U(p) \text{ halts}\}$ be Chaitin's Omega, $\Omega_n \in \mathbb{Q}_{\geq 0}$ be be the rational formed from the first n bits of Ω, and $\Omega^t = \sum\{2^{-\|p\|} : U(p) \text{ halts in time } t\}$. For $n\in\mathbb{N}$, let $\mathbf{bb}(n) = \min\{t : \Omega_n < \Omega^t\}$. $\mathbf{bb}^{-1}(m) = \arg\min_n\{\mathbf{bb}(n-1) < m \leq \mathbf{bb}(n)\}$. Let $\Omega[n] \in \{0,1\}^*$ be the first n bits of Ω.

Lemma 4 *For $n = \mathbf{bb}^{-1}(m)$, $\mathbf{K}(\Omega[n]|m,n) = O(1)$.*

Proof. For a string x, let $BB(x) = \inf\{t : \Omega^t > 0.x\}$. Enumerate strings of length n, starting with 0^n, and return the first string x such that $BB(x) \geq m$. This string x is equal to $\Omega[n]$, otherwise let y be the largest common prefix of x and $\Omega[n]$. Thus $BB(y) = \mathbf{bb}(\|y\|) \geq BB(x) \geq m$, which means $\mathbf{bb}^{-1}(m) \leq \|y\| < n$, causing a contradiction. □ The following lemma, while lengthy, is a series of straightforward application of inequalities.

Lemma 5 *For continuous probability P over $\{0,1\}^\infty$, $Z \subset \{0,1\}^\infty$, $|Z| = 2^s$, $s <^{\log} \max_{\alpha\in Z}\mathbf{D}(\alpha|P) + \mathbf{I}(\langle Z\rangle : \mathcal{H}) + O(\log\mathbf{K}(P))$.*

Proof. We relativize the universal Turing machine to s, which can be done due to the precision of the theorem. Let $Z_n = \{\alpha[0..n] : \alpha \in Z\}$ and $m = \arg\min_m |Z_m| = |Z|$. Let $n = \mathbf{bb}^{-1}(m)$ and $k = \mathbf{bb}(n)$. Let p be a probability over $\{0,1\}^*$, where $p(x) = [\|x\| = k]P(x)$ and $\langle p\rangle = \langle k, P\rangle$. Using $D = Z_k$, Lemmas 3 and 15 relativized to k produces $x \in Z_k$, where

$$\begin{aligned}
s &<^{\log} -\log P(x) - \mathbf{K}(x|k,P) + \mathbf{I}(Z_k;\mathcal{H}|k) + O(\log\mathbf{K}(P,k|k))\\
&<^{\log} -\log P(x) - \mathbf{K}(x|P) + \mathbf{K}(Z_k|k) + \mathbf{K}(k) - \mathbf{K}(Z_k|k,\mathcal{H}) + O(\log\mathbf{K}(P)).
\end{aligned}$$

Since $\mathbf{K}(k) <^+ n+\mathbf{K}(n)$, by the chain rule,

$$\begin{aligned}&\mathbf{K}(Z_k|k)+\mathbf{K}(k)\\ <^+&\mathbf{K}(Z_k|k,\mathbf{K}(k))+\mathbf{K}(\mathbf{K}(k)|k)+\mathbf{K}(k)\\ <&\mathbf{K}(Z_k,k)+O(\log n)\\ <&\mathbf{K}(Z_k)+O(\log n).\end{aligned}$$

So

$$s <^{\log} -\log P(x)-\mathbf{K}(x|P)+\mathbf{K}(Z_k)-\mathbf{K}(Z_k|k,\mathcal{H})+O(\log n+\log\mathbf{K}(P)).$$

Since $\mathbf{K}(k|n,\mathcal{H})=O(1)$, $\mathbf{K}(Z_k|\mathcal{H}) <^+ \mathbf{K}(Z_k|k,\mathcal{H})+\mathbf{K}(n)$. So

$$s <^{\log} -\log P(x)-\mathbf{K}(x|P)+\mathbf{I}(Z_k;\mathcal{H})+O(\log n+\log\mathbf{K}(P)).$$

By Lemma 4, $\mathbf{K}(\Omega[n]|Z_k) <^+ \mathbf{K}(n)$ so by Lemma 1,

$$n <^{\log} \mathbf{I}(\Omega[n];\mathcal{H}) <^{\log} \mathbf{I}(Z_k;\mathcal{H})+\mathbf{K}(n) <^{\log} \mathbf{I}(Z_k;\mathcal{H}).$$

The above equation used the common fact that the first n bits of Ω has $n-O(\log n)$ bits of mutual information with $\mathcal{H}$. So

$$s <^{\log} -\log P(x)-\mathbf{K}(x|P)+\mathbf{I}(Z_k;\mathcal{H})+O(\log\mathbf{K}(P)).$$

By the definition of mutual information $\mathbf{I}$ between infinite sequences

$$\mathbf{I}(Z_k;\mathcal{H}) <^+ \mathbf{I}(Z:\mathcal{H})+\mathbf{K}(Z_k|Z) <^{\log} \mathbf{I}(Z:\mathcal{H})+\mathbf{K}(k|Z).$$

Now m is simple relative to Z and by Lemma 4, $\Omega[n]$ is simple relative to m and n. Furthermore k is simple relative to $\Omega[n]$. Therefore $\mathbf{K}(Z_k|Z) <^+ \mathbf{K}(n)$. So

$$\begin{aligned}s &<^{\log} -\log P(x)-\mathbf{K}(x|P)+\mathbf{I}(Z:\mathcal{H})+\mathbf{K}(n)+O(\log\mathbf{K}(P))\\ s &<^{\log} \max_{\alpha\in Z}\mathbf{D}(\alpha|P)+\mathbf{I}(Z:\mathcal{H}))+O(\log\mathbf{K}(P)).\end{aligned}$$

□

Through careful observation, the above lemma can even be tightened to the following corollary

Corollary 4 *For continuous probability P over $\{0,1\}^\infty$, $Z\subset\{0,1\}^\infty$, $|Z|=2^s$, $s<\max_{\alpha\in Z}\mathbf{D}(\alpha|P)+\mathbf{I}(\langle Z\rangle:\mathcal{H})+O(\log\mathbf{I}(\langle Z\rangle:\mathcal{H})+\log\mathbf{K}(P))$.*

5.3 Observations as Reals

We model observations as infinite sequences of reals in the interval $[0,1]$, or equivalently infinite sequences γ of infinite sequences $\gamma_i\in\{0,1\}^\infty$, where each γ_i is unique. Of course, in the real world, infinite sequences of observations do not exist. But infinite sequences model processes that are potentially never ending. Let $\langle\gamma\rangle\in\{0,1\}^\infty$ be a standard encoding of γ. Let $\gamma(n)\subset\{0,1\}^\infty$be the first 2^n infinite sequences of γ. The following theorem uses the simple fact that $\mathbf{I}(f(\alpha):\mathcal{H}) <^+ \mathbf{I}(\alpha:\mathcal{H})+\mathbf{K}(f)$, for $\alpha\in\{0,1\}^\infty$.

Theorem 8
For probability P over $\{0,1\}^\infty$, $\gamma\in\{0,1\}^{\infty\mathbb{N}}$, let $t_{\gamma,P}=\sup_n(n-\mathbf{K}(n)-\max_{\alpha\in\gamma(n)}\mathbf{D}(\alpha|P))$. Then $t_{\gamma,P} <^{\log} \mathbf{I}(\langle\gamma\rangle:\mathcal{H})+O(\log\mathbf{K}(P))$.

Proof. By Corollary 4 applied to $\gamma(n)$,

$$\begin{aligned} n &< \max_{\alpha\in\gamma(n)} \mathbf{D}(\alpha|P) + \mathbf{I}(\gamma(n):\mathcal{H}) + O(\log \mathbf{I}(\gamma(n):\mathcal{H}) + \log \mathbf{K}(P)) \\ n - \max_{\alpha\in\gamma(n)} \mathbf{D}(\alpha|P) &<^{\log} +\mathbf{I}(\gamma(n):\mathcal{H}) + O(\log \mathbf{K}(P)) \\ n - \max_{\alpha\in\gamma(n)} \mathbf{D}(\alpha|P) - \mathbf{K}(n) &<^{\log} +\mathbf{I}(\langle\gamma\rangle:\mathcal{H}) + O(\log \mathbf{K}(P)) \\ t_{\gamma,P} &<^{\log} +\mathbf{I}(\langle\gamma\rangle:\mathcal{H}) + O(\log \mathbf{K}(P)). \end{aligned}$$

□

Let k be a physical address of γ. $\mathcal{H}$ can be described by a small mathematical statement. By Theorem 8 and **IP**, there is a small constant c where

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

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

Chapter 6

Better Bounds on Outlier Scores

In this chapter, better bounds are achieved than Lemmas 3 and 7 in Chapter 5. We first start with a discussion about algorithmic statistics. Algorithmic Statistics is the study of the separation of information, i.e. a string $x \in \{0,1\}^*$, into two parts. The first part is the model containing the "denoised" information of x. The second part is the data-to-model code representing the remaining randomness in x. The algorithmic statistics that we use in this paper are computable semi-measures P which have x in their support. Other models studied in the literature are finite setspf numbers and total recursive functions. For semi-measures, the model is an encoding or Turing number of an algorithm that computes P. The data-to-model code is the Shannon Fano encoding of length $=^+ -\log P(x)$ of x with respect to P. If x is typical of a model then it has a low deficiency of randomness $\mathbf{d}(x|P) = \lfloor -\log P(x) \rfloor - \mathbf{K}(x|P)$.

The field of algorithmic statistics studies properties of algorithmic sufficient statistics, i.e. statistics whose sum of the model complexity and data-to-model code length is equal (up to a small error term) to $\mathbf{K}(x)$. For probability distributions, these are such P where $\mathbf{K}(P) - \log P(x) \approx \mathbf{K}(x)$. A minimal sufficient statistic is an algorithmic sufficient statistic with the smallest model complexity, i.e. one that minimizes $\mathbf{K}(P)$. According to Occam's razor, out of all the algorithmic sufficient statistics, the minimal ones summarize the relevant information of x in the most concise manner.

This paper is connected to algorithmic statistics in two ways. First, the main theorem is a result about deficiencies of randomness, $\mathbf{d}$. The deficiency function $\mathbf{d}$ and its relation to models are one of the central areas of study in algorithmic statistics. Second, Lemma 7 is a statement about the stochasticity measure of a finite set of strings. The stochasticity term is related to those used in algorithmic statistics in that it measures whether a string is typical of a simple probability measure. The extended 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

Definition 3 (Stochasticity)
$\mathbf{Ks}(a|b) = \min\{\mathbf{K}(Q|b) + 3\log\max\{\mathbf{d}(a|Q,b),1\} : Q \textit{ is an elementary probability measure}\}$.

We have $\mathbf{Ks}(a) = \mathbf{Ks}(a|\emptyset)$, with $\mathbf{Ks}(a|b) <^+ \mathbf{Ks}(a) + \mathbf{K}(b)$. Thus if a has low $\mathbf{Ks}(a)$, then it is typical for a simple probability measure. Stochasticity is an important area of research because the stochasticity measure of an elementary object lower bounds the amount of information that the object has with the halting sequence, as shown in Section 8.2. Objects with high mutual information with the halting sequence are exotic in that there is no (randomized) method to produce them, due to information nongrowth laws. Thus the study of stochasticity yields insight into the properties of objects that can and cannot be produced by algorithms.

6.1 Games

In this section we introduce a generalization to the so-called "Epstein-Levin" game, introduced in [She12]. This new generalized game consists of a finite bipartite graph $E \subseteq L \times R$, with $L \subset \mathbb{N}$ and $R \subset \mathbb{N}$. There is a computable probability distribution P over the right vertices. The game is between Alice and Bob and is defined by four additional parameters.

1. An integer k.
2. A positive rational l.
3. A positive rational number δ.
4. A computable function $W : \mathbb{N} \to \mathbb{R}_{\geq 0}$.

The rules of the game are as follows. Alice assigns increasing rational nonnegative *weights* to vertices on L, which are all initially 0. The sum $\sum_{a \in L} W(a) \cdot \text{weight}(a)$ cannot exceed 1. After each turn by Alice, Bob can mark vertices on L and R. Once a vertex is marked, it will stay marked. There are restrictions on how Bob can mark the left and right vertices. The sum $W(a)$ over all marked left vertices cannot exceed l. Furthermore, the total P-probability of marked vertices on the right is at most δ.

Bob wins if every R vertex whose combined weight of its L-neighbors is equal to or greater than 2^{-k} either has a marked neighbor or is marked itself. Note that this is a generalization of the "Epstein-Levin" game in [She12], whose instantiation is equivalent to setting $W(a) = 1$ for all $a \in \mathbb{N}$.

Lemma 6 *For $l = O(2^k \log(1/\delta))$, Bob has a computable winning strategy.*

Note that the game can be made finite by making the weights restricted to the form 2^{-m} for $m \in \mathbb{Z}$. Since this new game changes the weights by a factor of at most 2, Bob can compensate by changing k by 1. In addition, the minimal weight is changed to be an $m \in \mathbb{Z}$ where $2^{-m} \max_{a \in L} W(a)|L| < 1$ so the sum $\sum_{a \in L} W(a) \cdot \text{weight}(a) \leq 2$, which is a constant factor. Thus this game is a finite game with full information so either Alice or Bob has a winning strategy. We prove that Bob has a probabilistic strategy that has a non-zero chance of winning. Thus Alice can't have a winning strategy, otherwise Bob's strategy would succeed with probability 0. Bob's simple probabilistic strategy is unchanged from that in [She12]:

- If Alice increases the weight on a vertex $a \in L$, by some value $\varepsilon \in (0, 1]$, then Bob marks that vertex with probability $c2^k\varepsilon$, where $c > 1$ is a constant to be chosen later. If $c2^k\varepsilon > 1$, then Bob marks the vertex.
- If a vertex on R has neighbors in L with total weight not less than 2^{-k} but no marked neighbors, then Bob immediately marks this vertex.

To prove that Bob has a non-zero chance of succeeding, we prove the following two events each have probability less than 1/2.

1. The total P-measure of marked R-vertices exceeds δ.
2. The sum of $W(a)$ over all marked left vertices $a \in L$ is more than l.

For (1), for each $y \in R$, with left neighbors with weights increasing $\varepsilon_1, \dots, \varepsilon_m$, with $\sum \varepsilon_i \geq 2^{-k}$, the probability that all its neighbors are unmarked is not more than

$$(1 - c2^k\varepsilon_1)\dots(1 - c2^k\varepsilon_m) \leq e^{-c2^k(\varepsilon_1+\dots+\varepsilon_m)} \leq e^{-c}.$$

For every P-measure, the expected P-measure of marked vertices in R does not exceed e^{-c}. For (1) to be less than $1/2$, it suffices for $c = \ln(1/\delta) + O(1)$.

For (2), the requirement that $\sum_{a \in L} W(a) \cdot \mathrm{weight}(a) \leq 1$ guarantees the following bound on the expectation

$$\begin{aligned} &E\left[\sum \{W(a) : a \text{ is a left marked vertex}\}\right] \\ &\leq \sum_{a \in L} W(a) \cdot \mathrm{weight}(a) c2^k \\ &\leq c2^k. \end{aligned}$$

Thus (2) is satisfied for $l = c2^{k+2} = O(2^k \log(1/\delta))$, thus proving the lemma.

6.1.1 Stochasticity

The above game can be applied to the following statement about the stochasticity of finite sets of natural numbers.

Lemma 7 *Let $\eta : \mathbb{N} \to \mathbb{R}_{\geq 0}$ be a lower semi-computable function, $W : \mathbb{N} \to \mathbb{R}_{\geq 0}$ be a computable function with $\sum_{a \in \mathbb{N}} W(a)\eta(a) \leq 1$. Then for every finite $D \subset \mathbb{N}$ with $\log \sum_{a \in D} \eta(a) \geq s \in \mathbb{Z}$ there is $a \in D$ with $\mathbf{K}(a) <^+ -\log W(a) - s + \Lambda(D) + 2\mathbf{K}(s)$. Note the above is true relative to any oracle α.*

Proof. Let Q be any elementary probability measure witnessing $\mathbf{Ks}(\langle D \rangle | s)$. The randomness deficiency of $\langle D \rangle$ with respect to Q, conditional to s, is $d = \max\{\mathbf{d}(\langle D \rangle | Q, s), 1\}$. From Q we create the following generalized Epstein-Levin game. The bipartite graph $E \subseteq L \times R$ is created by having R be the encoded sets $\langle G \rangle$ in the support of Q. The combined members of encoded sets in R are set to L and there is a connection between a vertex $a \in L$ and an encoded set $\langle G \rangle \in R$, if and only if $a \in G$. Alice approximates the weights η from below. At each round, Alice increases the weight of a vertex in L by the amount specified in the corresponding round of the lower enumeration of η. We set the parameters $k = -s$ and $\delta = 2^{-cd}$, for a constant $c \in \mathbb{N}$ solely dependent on the universal Turing machine to be determined later. The elementary probability is $P = Q$. By Lemma 6, Bob has a winning strategy where the sum of all $W(a)$ over left vertices marked by Bob is at most

$$l = O(2^k \log(1/\delta)) = O(cd2^{-s}).$$

The right vertex $\langle D \rangle$ is not marked. Otherwise, since the Q measure of vertices that are marked is not more than 2^{-cd}, and right vertices are marked during the course of the game, the function $Q' = (Q \cdot 2^{cd})$ restricted to marked right vertices is a lower semi-computable semi-measure. This semi-measure can be lower computed using Q, d, s, and c. Hence the Q' code of D would have the size $=^+ -\log(Q(\langle D \rangle)2^{cd})$. Thus the following contradiction occurs for large enough $c \in \mathbb{N}$ dependent solely on the universal Turing machine U,

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

Therefore, since $\langle D\rangle$ is not marked, and since $\sum_{a\in D}\eta(a)\geq 2^s = 2^{-k}$, by the rules of the game, D has a marked $a \in L$. The semi-measure $p(a) = W(a)/l$ for Bob's marked L vertices is lower semi-computable relative to Q, s, and d, so

$$\begin{aligned}
\mathbf{K}(a|Q,s,d) &<^+ -\log p(a)\\
&<^+ -\log W(a) + \log l\\
&<^+ -\log W(a) - s + \log d\\
\mathbf{K}(a) &<^+ -\log W(a) - s + \mathbf{K}(d) + \log d + \mathbf{K}(Q|s) + \mathbf{K}(s)\\
\mathbf{K}(a) &<^+ -\log W(a) - s + \Lambda(\langle D\rangle|s) + \mathbf{K}(s)\\
\mathbf{K}(a) &<^+ -\log W(a) - s + \Lambda(\langle D\rangle) + 2\mathbf{K}(s).
\end{aligned}$$

□

6.1.2 Stochastic Sets

The above lemma can be applied to the following result showing that large sets of numbers with low randomness deficiencies are exotic.

Theorem 9 *Relativized to computable semi-measure P over $\mathbb{N}$, for any finite set $D \subset \mathbb{N}$, if $\mathbb{N} \ni s < \log\sum_{a\in D}\mathbf{m}(a)/P(a)$, then $s <^+ \log\max_{a\in D}\mathbf{m}(a)/P(a) + \Lambda(D) + 2\mathbf{K}(s)$.*

Proof. We invoke Lemma 7. $W(a)$ is set to $P(a)$. $\eta(a)$ is set to $\mathbf{m}(a)/P(a)$ and is thus lower semi-computable. In addition $\sum_{a\in\mathbb{N}}W(a)\eta(a)\leq 1$ and $\sum_{a\in D}\eta(a)\geq 2^s$. The lemma produces an $a\in\mathbb{N}$ such that $\mathbf{K}(a) <^+ -\log P(a) - s + \Lambda(D) + 2\mathbf{K}(s)$. Some reworking proves the theorem. □

Corollary 5 *Relativized to computable semi-measure P over $\mathbb{N}$, for any finite set $D \subset \mathbb{N}$, if $\mathbb{N} \ni s < \log|D|$, then $s <^+ \log\max_{a\in D}\mathbf{m}(a)/P(a) + \Lambda(D) + 2\mathbf{K}(s)$.*

6.2 Helper Lemmas

The following elementary lemmas are used, in a helper capacity, throughout the paper. The terminology $O(f)$ for some function $f:\mathbb{N}\to\mathbb{N}$ signifies Big Oh notation of f with the parameters solely dependent on the choice of the universal Turing machine U. This holds also for the $f <^+ g$ inequality, which is equal to $f < g + O(1)$, for functions f, g between $\mathbb{N}$.

Lemma 8 *For every $c, n \in \mathbb{N}$ there exists $c' \in \mathbb{N}$ where if $x < y + c$ for some $x, y \in \mathbb{N}$ then $x + n\mathbf{K}(x) < y + n\mathbf{K}(y) + c'$.*

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

Lemma 9 *For $d, d', n, m \in \mathbb{N}$ there exists $d'' \in \mathbb{N}$ where for any $f, g, h \in \mathbb{N}$, if $f < g + n\mathbf{K}(g) + d$ and $g < h + m\mathbf{K}(h) + d'$, then $f < h + (m+2n)\mathbf{K}(h) + d''$.*

Proof. If $g < h + d'$, then due to Lemma 8 applied to $x = g$, $y = h$, n, and $c = d'$, there exists c' dependent on d' and n where $g + n\mathbf{K}(g) < h + n\mathbf{K}(h) + c'$ and thus $f < h + n\mathbf{K}(h) + d + c'$, proving the lemma. Thus $h + d' \leq g$ and $g - h < m\mathbf{K}(h) + d'$, which implies $\mathbf{K}(g-h) <^+ 2\log m\mathbf{K}(h) + 2\log d'$. Therefore $\mathbf{K}(g) <^+ \mathbf{K}(h) + \mathbf{K}(g-h) <^+ \mathbf{K}(h) + 2\log m + 2\log\mathbf{K}(h) + 2\log d'$. So,

$$\begin{aligned}
f &< g + n\mathbf{K}(g) + d \\
&<^+ h + m\mathbf{K}(h) + n\mathbf{K}(g) + d + d' \\
&<^+ h + m\mathbf{K}(h) + n(\mathbf{K}(h) + \mathbf{K}(g-h) + O(1)) + d + d' \\
&<^+ h + m\mathbf{K}(h) + n(\mathbf{K}(h) + 2\log m + 2\log\mathbf{K}(h) + 2\log d' + O(1)) + d + d' \\
&< h + (m+2n)\mathbf{K}(h) + O(n\log m) + 2n\log d' + d + d' \\
&< h + (m+2n)\mathbf{K}(h) + d'',
\end{aligned}$$

where $d'' = O(n\log m) + 2n\log d' + d + d'$. □

Lemma 10 *For every $c, n \in \mathbb{N}$, there exists $c' \in \mathbb{N}$ where for all $x, y \in \mathbb{N}$, if $x < y + n\log x + c$ then $x < y + 2n\log y + c'$.*

Proof.

$$\begin{aligned}
\log x &< \log y + \log\log x + \log cn \\
2\log x - 2\log\log x &< 2\log y + 2\log cn \\
\log x &< 2\log y + 2\log cn.
\end{aligned}$$

Combining with the original inequality

$$\begin{aligned}
x &< y + n\log x + c \\
x &< y + n(2\log y + 2\log cn) + c \\
&= y + 2n\log y + c',
\end{aligned}$$

where $c' = 2n\log cn + c$. □

Lemma 11 *For every $d \in \mathbb{N}$ there exists $d' \in \mathbb{N}$ where if $x < y + \mathbf{K}(x) + d$ then $x < y + 2\mathbf{K}(y) + d'$.*

Proof. It must be that $y + d < x$, otherwise the lemma is trivially solved. Thus $x - y < \mathbf{K}(x) + d$, so $\mathbf{K}(x-y) <^+ 2\log\mathbf{K}(x) + 2\log d$. So $\mathbf{K}(x) <^+ \mathbf{K}(y) + \mathbf{K}(x-y) <^+ \mathbf{K}(y) + 2\log\mathbf{K}(x) + 2\log d$. By Lemma 10, where $n = 2$ and $c = 2\log d + O(1)$, there is a $c' \in \mathbb{N}$, where $\mathbf{K}(x) < \mathbf{K}(y) + 4\log\mathbf{K}(y) + c' < 2\mathbf{K}(y) + c' + O(1)$. So

$$\begin{aligned}
x &< y + \mathbf{K}(x) + d \\
&< y + 2\mathbf{K}(y) + c' + d + O(1) \\
&= y + 2\mathbf{K}(y) + d',
\end{aligned}$$

where $d' = c' + d + O(1)$. □

Lemma 12 *For all $d, m \in \mathbb{N}$ there is a $d' \in \mathbb{N}$ where if $x + m\mathbf{K}(x) + d > y$ then $x + d' > y - 2m\mathbf{K}(y)$.*

Proof. If $x+d>y$, then the lemma is satisfied, so $x+d\leq y$. Thus $y-x<m\mathbf{K}(x)+d$ implies $\mathbf{K}(y-x)<^+ 2\log\mathbf{K}(x)+2\log dm$. Thus $\mathbf{K}(x)<^+\mathbf{K}(y)+\mathbf{K}(y-x)<^+\mathbf{K}(y)+2\log\mathbf{K}(x)+2\log dm$. Applying Lemma 10 where $c=2\log dm+O(1)$ and $n=2$, we get a c' dependent on c and n where $\mathbf{K}(x)<\mathbf{K}(y)+4\log\mathbf{K}(y)+c'<2\mathbf{K}(y)+c'+O(1)$. So

$$\begin{aligned} x+m\mathbf{K}(x)+d &> y\\ x+m(2\mathbf{K}(y)+c'+O(1))+d &> y\\ x+d' &> y-2m\mathbf{K}(y), \end{aligned}$$

where $d'=m(c'+O(1))+d$. □

6.3 Left-Total Machines

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 $y\in\{0,1\}^{*\infty}$ if 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 $\alpha\in\{0,1\}^{*\infty}$ is to the "left" of $\beta\in\{0,1\}^{*\infty}$, and use the notation $\alpha\lhd\beta$, if there exists $x\in\{0,1\}^*$ 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. An example left-total machine can be seen in Figure 6.1.

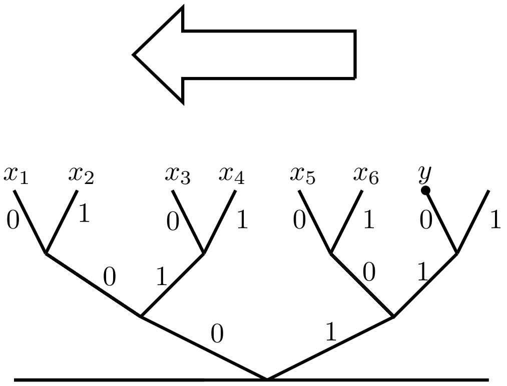

Figure 6.1: The above diagram represents the domain of a left total machine T with the 0 bits branching to the left and the 1 bits branching to the right. For $i\in\{1,\dots,5\}$, $x_i\lhd x_{i+1}$ and $x_i\lhd y$. Assuming $T(y)$ halts, each x_i is total. This also implies each x_i^- is total as well.

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' by the following definition. The algorithm U' enumerates all strings $p\in\{0,1\}^*$ in order of their convergence time of $U(p)$ and successively assigns them consecutive intervals $i_p\subset[0,1]$ of width $2^{-\|p\|}$. Then U' outputs $U(p)$ on input p' if the open interval corresponding to p' and not that of $(p')^-$ is strictly contained in i_p. The open interval in [0,1] corresponding with p' is $((p')2^{-\|p'\|},((p')+1)2^{-\|p'\|})$ where (p) is the value of p in binary. For example, the value of both strings 011 and 0011 is 3. The value of 0100 is 4. The same definition applies for the machines U'_α and U_α, over all $\alpha\in\{0,1\}^{*\infty}$. We now set U to equal U'.

Without loss of generality, the complexity terms of this paper are defined with respect to the optimal left total machine U. The infinite border sequence $\mathcal{B}\in\{0,1\}^\infty$ represents the unique

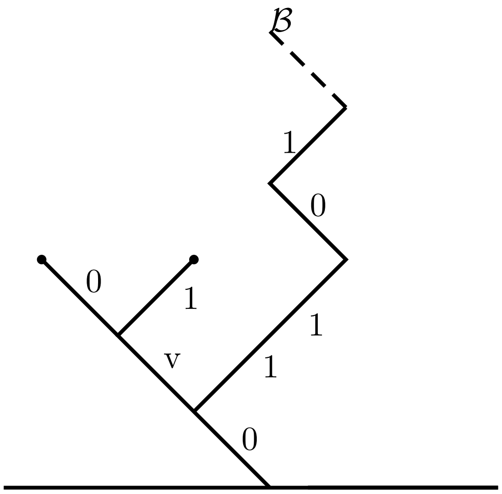

Figure 6.2: The above diagram represents the domain of the optimal left-total algorithm U, with the 0 bits branching to the left and the 1 bits branching to the right. The strings in the above diagram, $0v0$ and $0v1$, are halting inputs to U with $U(0v0) \neq\perp$ and $U(0v1) \neq\perp$. So $0v$ is 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 non total extensions. All finite strings branching to the right of $\mathcal{B}$ will cause U to diverge.

infinite sequence such that all its finite prefixes have total and non total extensions. The term "border" is used because for any string $x \in \{0,1\}^*$, $x \lhd \mathcal{B}$ implies that x total with respect to U and $\mathcal{B} \lhd x$ implies that U will never halt when given x as an initial input. Figure 6.2 shows the domain of U with respect to $\mathcal{B}$.

6.3.1 Properties of Total Strings

This section uses the notion of a Martin Löf random infinite sequence. An infinite sequence $\alpha \in \{0,1\}^\infty$ is Martin Löf random if there is a constant $c \in \mathbb{N}$ such that for all $n \in \mathbb{N}$, $\mathbf{K}(\alpha[0..n]) > n-c$. Let $\Omega = \sum_x \mathbf{m}(x)$ be Chaitin's Omega, the probability that U will halt. It is well known that the binary expansion of Ω is Martin Löf random.

Proposition 1 *The border sequence $\mathcal{B}$ is Martin Löf random. Furthermore if $b \in \{0,1\}^*$ is total and b^- is not, then $b^- \sqsubset \mathcal{B}$.*

Proof. The border sequence is the binary expansion of Chaitin's Omega for machine U, because the probability that a random infinite sequence contains a prefix that is a halting program is precisely the probability that the random sequence is at the left of the border sequence. If $b \in \{0,1\}^*$ is total and b^- is not, then b^- has a total extension b^-0 and a non total extension b^-1, thus by the definition of the border sequence, $b^- \sqsubset \mathcal{B}$. □

Lemma 13 *If $b \in \{0,1\}^*$ is total and b^- is not, and $x \in \{0,1\}^*$,*
then $\mathbf{K}(b) + \mathbf{I}(x : \mathcal{H}|b) <^{\log} \mathbf{I}(x:\mathcal{H}) + \mathbf{K}(b|\langle x, \|b\|\rangle)$.

Proof. By Proposition 1, $b^- \sqsubset \mathcal{B}$ is a prefix of the border sequence and thus $\|b\| <^+ \mathbf{K}(b)$. Since $\mathcal{B}$ is computable from the halting sequence $\mathcal{H}$, we have that b is computable from $\|b\|$ and $\mathcal{H}$, with $\mathbf{K}(b|\mathcal{H}) <^+ \mathbf{K}(\|b\|)$.

The chain rule gives the equality $\mathbf{K}(b)+\mathbf{K}(x|b,\mathbf{K}(b)) =^+ \mathbf{K}(x)+\mathbf{K}(b|x,\mathbf{K}(x))$. Combined with the inequalities $\mathbf{K}(x|b) <^+ \mathbf{K}(x|b,\mathbf{K}(b))+\mathbf{K}(\mathbf{K}(b))$ and $\mathbf{K}(b|x,\mathbf{K}(x)) <^+ \mathbf{K}(b|x)$, we get

$$\mathbf{K}(b)+\mathbf{K}(x|b) <^+ \mathbf{K}(x)+\mathbf{K}(b|x)+\mathbf{K}(\mathbf{K}(b)).$$

Subtracting $\mathbf{K}(x|b,\mathcal{H})$ from both sides results in

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

So $\mathbf{K}(b)+\mathbf{I}(x:\mathcal{H}|b) <^{\log} \mathbf{I}(x:\mathcal{H})+\mathbf{K}(b|\langle x,\|b\|\rangle)$. □

Lemma 14 *If $b\in\{0,1\}^*$ is total and b^- is not, and for $x\in\{0,1\}^*$, $\mathbf{K}(b|\langle x,|b\|\rangle)=O(1)$, then $\mathbf{K}(\|b\|) <^{\log} 2\log\mathbf{I}(x:\mathcal{H})$.*

Proof. Due to Proposition 1, by the definition of b, b is total and b^- is not, so $b^-\sqsubset\mathcal{B}$ is a prefix of border, and is thus a random string, with $\|b\| <^{\log} \mathbf{K}(b)$. Due to Lemma 13, with the second term removed,

$$\begin{aligned}\|b\| <^{\log} \mathbf{K}(b) &<^{\log} \mathbf{I}(x:\mathcal{H})+\mathbf{K}(b|\langle x,\|b\|\rangle)\\ \|b\| &<^{\log} \mathbf{I}(x:\mathcal{H})\\ \mathbf{K}(\|b\|) &<^+ 2\log\mathbf{I}(x:\mathcal{H}).\end{aligned}$$

□

6.3.2 Stochasticity and the Halting Sequence

Left-total machines can be used to prove properties of stochasticity. As mentioned earlier, the stochasticity of a string lower bounds the amount of mutual information it has with the halting sequence. The following lemma was first introduced in [EL11].

Lemma 15 *For $x\in\mathbb{N}$, $\mathbf{Ks}(x) < \mathbf{I}(x:\mathcal{H})+6\mathbf{K}(\mathbf{I}(x:\mathcal{H}))$.*

Proof. Using the optimal left-total Turing machine, let $U(x^*)=x$, $\|x^*\|=\mathbf{K}(x)$, and v be the shortest total prefix of x^*. We define the elementary probability measure Q such that $Q(a)=\sum_w 2^{-\|w\|}[U(vw)=a]$. A graphical depiction of these definitions can be seen in Figure 6.3. Thus Q is computable relative to v. In addition, since $v\sqsubseteq x^*$, one has the lower bound $Q(x)\geq 2^{-\|x^*\|+\|v\|}=2^{-\mathbf{K}(x)+\|v\|}$. Therefore

$$\begin{aligned}\mathbf{d}(x|Q,v) &= \lfloor -\log Q(x)\rfloor - \mathbf{K}(x|\langle Q\rangle,v)\\ &=^+ -\log Q(x)-\mathbf{K}(x|v)\\ &<^+ \mathbf{K}(x)-\|v\|-\mathbf{K}(x|v)\\ &<^+ (\mathbf{K}(v)+\mathbf{K}(x|v))-\|v\|-\mathbf{K}(x|v)\\ &<^+ (\|v\|+\mathbf{K}(\|v\|)+\mathbf{K}(x|v))-\|v\|-\mathbf{K}(x|v)\\ \mathbf{d}(x|Q,v) &<^+ \mathbf{K}(\|v\|). \end{aligned} \tag{6.1}$$

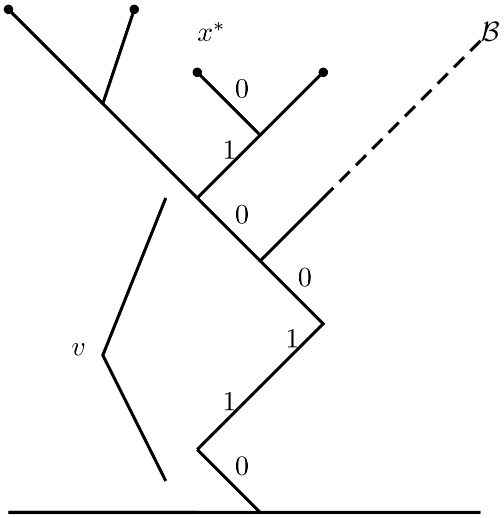

Figure 6.3: A graphical depiction of the terms used in Lemma 15. The shortest program for $x \in \mathbb{N}$ is $x^* = 0110010$, with $U(x^*) = x$. The shortest total prefix of x^* is $v = 01100$, with $v^- = 0110$ being a prefix of border $\mathcal{B}$. Assuming x^* is the only extension of v that is a program for x, then $Q(x) = 2^{-\|x^*\|+\|v\|} = 2^{-2}$.

Since v is total and v^- is not total, by Proposition 1, v^- is a prefix of the border sequence $\mathcal{B}$. In addition, Q is computable from v. Therefore

$$\begin{aligned}
\mathbf{K}(x|\mathcal{H}) &<^+ \mathbf{K}(x|Q) + \mathbf{K}(Q|\mathcal{H}) \\
&<^+ \mathbf{K}(x|Q) + \mathbf{K}(v|\mathcal{H}) \\
&<^+ -\log Q(x) + \mathbf{K}(\|v\|) \qquad (6.2)\\
&<^+ \mathbf{K}(x) - \|v\| + \mathbf{K}(\|v\|) \\
\|v\| &<^+ \mathbf{K}(x) - \mathbf{K}(x|\mathcal{H}) + \mathbf{K}(\|v\|) \\
\|v\| &<^+ \mathbf{I}(x:\mathcal{H}) + 2\mathbf{K}(\mathbf{I}(x:\mathcal{H})). \qquad (6.3)
\end{aligned}$$

Equation (6.2) is due to $\mathcal{B}$ being computable from $\mathcal{H}$, therefore $v^- \sqsubset \mathcal{B}$ is simple relative to $\mathcal{H}$ and $\|v\|$. Equation (6.3) is from Lemma 11. Since Q is computable from v, one gets

$$\begin{aligned}
\mathbf{Ks}(x) &<^+ \mathbf{K}(v) + 3\log(\max\{\mathbf{d}(x|Q,v),1\}) \\
&<^+ \|v\| + \mathbf{K}(\|v\|) + 3\log(\max\{\mathbf{d}(x|Q,v),1\}) \\
&<^+ \|v\| + \mathbf{K}(\|v\|) + 3\log\mathbf{K}(\|v\|) \\
&<^+ \|v\| + 2\mathbf{K}(\|v\|).
\end{aligned}$$

Applying Lemma 9 to $f \equiv \mathbf{Ks}(x)$, $g \equiv \|v\|$, and $h \equiv \mathbf{I}(x:\mathcal{H})$, with $n = 2$ and $m = 2$, gets $\mathbf{Ks}(x) <^+ \mathbf{I}(x:\mathcal{H}) + 6\mathbf{K}(\mathbf{I}(x:\mathcal{H}))$. □

6.4 Infinite Sequences

This section we prove that large sets of infinite sequences with low $\mathbf{D}$ scores are exotic, in that they have high mutual information with the halting sequence. We recall that $x \sqsubseteq y$ for $x, y \in$

$\{0,1\}^*$ implies that x is a prefix of y or equal to y. For a string $x \in \{0,1\}^*$, let $\mathbf{D}(x|P) = \max_{y \sqsubseteq x}(\log(\mathbf{m}(y|P)/P(y)))$. Let $\mathbf{bb}(b) = \max\{U(p) : p \lhd b, \text{ or } p \sqsupseteq b\}$ be the largest number produced by a program that extends b or is to the left of b.

Theorem 10 *Relativized to computable probability measure P over $\{0,1\}^\infty$, for $Z \subseteq \{0,1\}^\infty$, if $\mathbb{N} \ni s < \log \sum_{\alpha \in Z} 2^{\mathbf{D}(\alpha|P)}$, then $s < \sup_{\alpha \in Z} \mathbf{D}(\alpha|P) + \mathbf{I}(\langle Z\rangle : \mathcal{H}) + O(\mathbf{K}(s) + \log \mathbf{I}(\langle Z\rangle : \mathcal{H}))$.*

Informal Proof. The proof starts off by determining an $N \in \mathbb{N}$, such that $\sum_{x \in Z_{\leq N}} 2^{\mathbf{D}(x|P)} > 2^s$. This N is equal to $\mathbf{bb}(b)$ for some total string b. Then Lemma 7 is invoked with $W(x) = P(x)$, $\eta(x) = [x \in \{0,1\}^N]2^{\mathbf{D}(x|P)}$, $D = Z_{\leq N}$, relativized to b. This produces $x \in D$ where $\mathbf{K}(x|b) <^+ -\log P(x) + \mathbf{Ks}(D|b) + O(\mathbf{K}(s))$. Using Lemma 15, the $\mathbf{Ks}(D|b)$ term is replaced with $\mathbf{I}(D : \mathcal{H}|b)$. The conditioning on b is removed using Lemma 13. Finally the $\mathbf{I}(D : \mathcal{H})$ term is replaced with $\mathbf{I}(\langle Z\rangle : \mathcal{H})$ to achieve the theorem.

Proof.

1. Determination of N.

For a total $b \in \{0,1\}^*$, let $\mathbf{m}_b(x|y) = \sum\{2^{-\|z\|} : U_y(z) = x, U_y(z) \text{ halts in } \mathbf{bb}(b) \text{ time}\}$ be the algorithmic weight of x using solely programs that are running in $\mathbf{bb}(b)$ time. For $x \in \{0,1\}^*$, let $\mathbf{D}_b(x|P) = \max_{y \sqsubseteq x}(\log(\mathbf{m}_b(y|P)/P(y)))$, with $\mathbf{D}_b \leq \mathbf{D}$. We set b to be the shortest total string with

1. $N = \mathbf{bb}(b)$.
2. $\sum_{x \in Z_{\leq N}} 2^{\mathbf{D}_b(x|P)} > 2^s$.

2. Invocation of Lemma 7.

We let $W(x) = P(x)$, $\eta(x) = 2^{\mathbf{D}(x|P)}[x \in \{0,1\}^N]$, and $D = Z_{\leq N}$. Since the universal Turing machine is relativized to P, it must be that $\mathbf{K}(\langle W, \eta\rangle|b) = O(1)$, $\log \sum_{x \in D} \eta(x) > s$, and

$$\begin{aligned}\sum_{x \in \{0,1\}^*} W(x)\eta(x) &= \sum_{x \in \{0,1\}^N} P(x)2^{\mathbf{D}(x|P)}\\ &= \int_\alpha 2^{\mathbf{D}(\alpha[0..N]|P)} dP(\alpha)\\ &\leq \int_\alpha 2^{\mathbf{D}(\alpha|P)} dP(\alpha) \leq 1.\end{aligned}$$

Lemma 7, relativized to b, gives $x \in D$ with

$$\mathbf{K}(x|b) < -\log P(x) - s + \mathbf{Ks}(D|b) + O(\mathbf{K}(s)).$$

3. Replace $\mathbf{Ks}(D|b)$ with $\mathbf{I}(D : \mathcal{H}|b)$.

Due to Lemma 15,

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

4. Remove conditioning of b.
By Lemma 13,

$$\mathbf{K}(b) + \mathbf{I}(D : \mathcal{H}|b) <^{\log} \mathbf{I}(D : \mathcal{H}) + \mathbf{K}(b|\langle D, \|b\|\rangle).$$

Therefore

$$s \leq \log(\mathbf{m}(x)/P(x)) + \mathbf{I}(D : \mathcal{H}) + \mathbf{K}(b|\langle D, \|b\|\rangle) + O(\mathbf{K}(s) + \log(\mathbf{I}(D : \mathcal{H}) + \mathbf{K}(b|\langle D, \|b\|\rangle))).$$

Since $D \subseteq \{0,1\}^{\mathbf{bb}(b)}$, $\mathbf{K}(b|\langle D, \|b\|\rangle) = O(1)$, 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. Hence

$$s \leq \log(\mathbf{m}(x)/P(x)) + \mathbf{I}(D : \mathcal{H}) + O(\mathbf{K}(s) + \log \mathbf{I}(D : \mathcal{H})).$$

5. Replace $\mathbf{I}(D : \mathcal{H})$ with $\mathbf{I}(\langle Z\rangle : \mathcal{H})$.
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 $\|b\|$, s, and $\langle Z\rangle$ and thus so is $D = Z_{\leq \mathbf{bb}(b)}$. By Definition 2 of mutual information between infinite sequences,

$$\begin{aligned}
\mathbf{I}(D : \mathcal{H}) &<^+ \mathbf{I}(\langle Z\rangle : \mathcal{H}) + \mathbf{K}(D|\langle Z\rangle) \\
&<^+ \mathbf{I}(\langle Z\rangle : \mathcal{H}) + \mathbf{K}(\|b\|) + \mathbf{K}(s) \\
&<^+ \mathbf{I}(\langle Z\rangle : \mathcal{H}) + 2\log \mathbf{I}(D : \mathcal{H}) + \mathbf{K}(s) \qquad (6.4) \\
&<^{\log} \mathbf{I}(\langle Z\rangle : \mathcal{H}) + \mathbf{K}(s). \qquad (6.5)
\end{aligned}$$

Where Equation 6.4 is due to the application of Lemma 14, noting $\mathbf{K}(b|\langle D, \|b\|\rangle) = O(1)$. Equation 6.5 is due to Lemma 10. So

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

□

Corollary 6 *Relativized to computable probability measure P over $\{0,1\}^\infty$, for any set $Z \subseteq \{0,1\}^\infty$ with $n < \log|Z|$, $n < \sup_{\alpha\in Z} \mathbf{D}(\alpha|P) + \mathbf{I}(\langle Z\rangle : \mathcal{H}) + O(\mathbf{K}(n) + \log \mathbf{I}(\langle Z\rangle : \mathcal{H}))$.*

Proof. This follows from the fact that for any $\alpha \in \{0,1\}^\infty$, $\mathbf{D}(\alpha|P) >^+ 0$ because using continuous Shannon-Fano encoding, there is a prefix $x \sqsubset \alpha$ that can be identified by a code of $=^+ -\log P(x)$. This implies $\mathbf{K}(x|P) <^+ -\log P(x)$ and thus $\mathbf{D}(\alpha|P) \geq -\log P(x) - \mathbf{K}(x|P) >^+ 0$. Therefore there is some $c \in \mathbb{N}$ solely dependent on the universal Turing machine U, such that $\log \sum_{\alpha\in Z} 2^{\mathbf{D}(\alpha|P)} > \log \sum_{\alpha \in Z} 2^{-c} > n - c$. □

A continuous sampling method A takes in a parameter n, a infinite source of random bits and outputs 2^n unique infinite sequences encoded in the form

$$\alpha_1[1]\alpha_2[1]...\alpha_{2^n}[1]\alpha_1[2]\alpha_2[2]...\alpha_{2^n}[2]\ldots$$

We get the following continuous sampling corollary which is analogous the discrete case.

Corollary 7 *For computable measure P over $\{0,1\}^\infty$, for continuous sampling method A, there exists $c_{P,A} \in \mathbb{N}$, where for all $n, k \in \mathbb{N}$, $\Pr(n - \max_{\alpha \in A(n)} \mathbf{D}(\alpha|P) > k) < 2^{-k+O(\log k + \mathbf{K}(n)) + c_{P,A}}$.*

Proof. We use $\gamma \sim \mathcal{U}$ to represent infinite sequences distributed according to the uniform distribution.

$$\begin{aligned}
&\Pr_{\gamma\sim\mathcal{U}}\left(n - \max_{\alpha\in A(n,\gamma)} \mathbf{D}(\alpha|P) > k\right) \\
&< \Pr_{\gamma\sim\mathcal{U}} (c_P + \mathbf{I}(A(n,\gamma) : \mathcal{H}) + O(\log \mathbf{I}(A(n,\gamma) : \mathcal{H})) > k - O(\mathbf{K}(n))) && (6.6)\\
&< \Pr_{\gamma\sim\mathcal{U}} (c_P + \mathbf{I}(A(n,\gamma) : \mathcal{H}) > k - O(\mathbf{K}(n) + \log k)) && (6.7)\\
&< \Pr_{\gamma\sim\mathcal{U}} (\mathbf{I}(\gamma : \mathcal{H}) > k - O(\mathbf{K}(n) + \log k) - c_P - c_A) && (6.8)\\
&< 2^{-k+O(\log k+\mathbf{K}(n))+c_P+c_A}. && (6.9)
\end{aligned}$$

Equation 6.6 comes from Corollary 6, where $c_P \in \mathbb{N}$ is a constant solely dependent on P and the universal Turing machine U. Equation 6.7 comes from the fact that $a < i + O(\log i)$ implies that either $a < i$ or then $O(\log a) > O(\log i)$ and then $a - O(\log a) < i$. Equation 6.8 comes from Theorem 3, where $f = A(n, \cdot)$, with $\mathbf{K}(f) = O(\mathbf{K}(n))$. Thus $c_A \in \mathbb{N}$ is a constant solely dependent on A and the universal Turing machine U. Equation 6.9 comes from the application of Theorem 2, where $\alpha = 0^\infty$, $\beta = \mathcal{H}$, and $P_\alpha = \mathcal{U}$. □

Chapter 7

Probabilistic Algorithms

In this section, we prove that the non-automatic output of randomized algorithms are guaranteed to have high $\mathbf{D}$ scores, i.e. be outliers.

Theorem 11 *For computable measures μ and nonatomic λ over $\{0,1\}^\infty$ and $n \in \mathbb{N}$, $\lambda\{\alpha : \mathbf{D}(\alpha|\mu) > n\} > 2^{-n-\mathbf{K}(n,\mu,\lambda)-O(1)}$.*

Proof. We first assume not. For all $c \in \mathbb{N}$, there exist computable nonatomic measures μ, λ, and there exists n, where $\lambda\{\alpha : \mathbf{D}(\alpha|\mu) > n\} \leq 2^{-n-\mathbf{K}(n,\mu,\lambda)-c}$. Sample $2^{n+\mathbf{K}(n,\mu,\lambda)+c-1}$ elements $D \subset \{0,1\}^\infty$ according to λ. The probability that all samples $\beta \in D$ have $\mathbf{D}(\beta|\mu) \leq n$ is

$$\prod_{\beta \in D} \lambda\{\mathbf{D}(\beta|\mu) \leq n\} \geq (1 - |D|2^{-n-\mathbf{K}(n,\mu,\lambda)-c}) \geq (1 - 2^{n+\mathbf{K}(n,\mu,\lambda)+c-1}2^{-n-\mathbf{K}(n,\mu,\lambda)-c}) \geq 1/2.$$

Let $\lambda^{n,c}$ be the probability of an encoding of $2^{n+\mathbf{K}(n,\mu,\lambda)+c-1}$ elements each distributed according to λ. Thus

$$\lambda^{n,c}(\text{Encoding of } 2^{n+\mathbf{K}(n,\mu,\lambda)+c-1} \text{ elements } \beta\text{, each having } \mathbf{D}(\beta|\mu) \leq n) \geq 1/2.$$

Let v be a shortest program to compute $\langle n, \mu, \lambda \rangle$. By Corollary 2, with the universal Turing machine relativized to v,

$$\lambda^{n,c}(\{\gamma : \mathbf{I}(\gamma : \mathcal{H}|v) > m\}) \overset{*}{<} 2^{-m+\mathbf{K}(\lambda^{n,c}|v)} \overset{*}{<} 2^{-m+\mathbf{K}(n,\mathbf{K}(n,\mu,\lambda),c,\lambda|v)} \overset{*}{<} 2^{-m+\mathbf{K}(c)}.$$

Therefore,

$$\lambda^{n,c}(\{\gamma : \mathbf{I}(\gamma : \mathcal{H}|v) > \mathbf{K}(c) + O(1)\}) \leq 1/4.$$

Thus, by probabilistic arguments, there exists $\alpha \in \{0,1\}^\infty$, such that $\alpha = \langle D \rangle$ is an encoding of $2^{n+\mathbf{K}(n,\mu,\lambda)+c-1}$ elements $\beta \in D \subset \{0,1\}^\infty$, where each β has $\mathbf{D}(\beta|\mu) \leq n$ and $\mathbf{I}(\alpha : \mathcal{H}|v) <^+ \mathbf{K}(c)$. By Theorem 10, relativized to v, there are constants $d, f \in \mathbb{N}$ where

$$\begin{aligned}
m = \log|D| &< \max_{\beta \in D} \mathbf{D}(\beta|\mu, v) + 2\mathbf{I}(D : \mathcal{H}|v) + d\mathbf{K}(m|v) + f\mathbf{K}(\mu|v) \\
m &< \max_{\beta \in D} \mathbf{D}(\beta|\mu) + \mathbf{K}(v) + 2\mathbf{I}(D : \mathcal{H}|v) + d\mathbf{K}(m|v) + f\mathbf{K}(\mu|v) \qquad (7.1) \\
&<^+ \max_{\beta \in D} \mathbf{D}(\beta|\mu) + \mathbf{K}(n,\mu,\lambda) + 2\mathbf{K}(c) + d\mathbf{K}(m|v) + f\mathbf{K}(\mu|v) \\
&<^+ n + \mathbf{K}(n,\mu,\lambda) + d\mathbf{K}(m|v) + 2\mathbf{K}(c). \qquad (7.2)
\end{aligned}$$

Therefore:

$$\begin{aligned} m &= n + \mathbf{K}(n,\mu,\lambda) + c - 1 \\ \mathbf{K}(m|v) &<^+ \mathbf{K}(c). \end{aligned}$$

Plugging the inequality for $\mathbf{K}(m|v)$ back into Equation 7.2 results in

$$\begin{aligned} n + \mathbf{K}(n,\mu,\lambda) + c &<^+ n + \mathbf{K}(n,\mu,\lambda) + 2\mathbf{K}(c) + d\mathbf{K}(c) \\ c &<^+ (2+d)\mathbf{K}(c). \end{aligned}$$

This result is a contradiction for sufficiently large c solely dependent on the universal Turing machine. □

Similar to the construction in the introduction, we can define a universal conditional lower computable integral test $T(\alpha|n)$ over a sequence of uniformly computable measures Q_1, Q_2, ... over $\{0,1\}^\infty$. We can also define the randomness deficiency to be $\mathbf{D}_n(\alpha|Q) = \log T(\alpha|n)$. The following corollary is derived from the fact that $\mathbf{D}_n(\alpha|\mu,n) = \mathbf{D}_n(\alpha|\mu)$.

Corollary 8 *For uniformly computable measures $\{\mu_i\}$ and nonatomic $\{\lambda_i\}$ over $\{0,1\}^\infty$, for all n, $\lambda_n\{\alpha : \mathbf{D}_n(\alpha|\mu) > n\} > 2^{-n-\mathbf{K}(\mu,\lambda)-O(1)}$.*

Theorem 11 can be extended to incomputable λ, which can be accomplished using Theorem 2. The term $\langle\lambda\rangle \in \{0,1\}^\infty$ in the following corollary represents any encoding of λ that can compute $\lambda(x\{0,1\}^\infty)$ for $x \in \{0,1\}^*$ up to arbitrary precision. Let $\mathbf{I}(\lambda : \mathcal{H}) = \inf_{\langle\lambda\rangle} \mathbf{I}(\langle\lambda\rangle; \mathcal{H})$.

Corollary 9

- *For measures μ and λ over $\{0,1\}^\infty$, nonatomic λ, computable μ, for all n, $\lambda\{\alpha : \mathbf{D}(\alpha|\mu) > n\} > 2^{-n-O(\mathbf{K}(n,\mu)+\mathbf{I}(\lambda:\mathcal{H}))}$.*
- *For measures μ and λ over $\{0,1\}^\infty$, nonatomic λ, computable μ, if for every $c \in \mathbb{N}$, there is an $n \in \mathbb{N}$, where $\lambda\{\alpha : \mathbf{D}(\alpha|\mu) > n\} < 2^{-n-O(\mathbf{K}(n))-c}$, then $\mathbf{I}(\lambda : \mathcal{H}) = \infty$.*

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,

Corollary 10 *Starting λ-almost everywhere, $\overset{*}{>} \mathbf{m}(n,\mu,\lambda)2^{-n}$ states α visited by iterations of T have $\mathbf{D}(\alpha|\mu) > n$.*

7.1 Alternative Proof

Using Theorem 5 of Chapter 3, one can produce a simple proof to a variant of Theorem 11. The longer proof was included due to of its tight error terms, its use of the standard randomness deficiency function $\mathbf{D}$, and its corollaries extending the results to incomputable measures. Let $\lambda = \lambda_1, \lambda_2, \ldots$ and $\mu = \mu_1, \mu_2, \ldots$ be uniformly computable sequences of measures over infinite sequences. Each λ_n is non-atomic.

Theorem 12 *There are constants $b, c \in \mathbb{N}$, dependent on μ and λ, where for all $n \in \mathbb{N}$, $\lambda_n\left\{\alpha : \overline{\mathbf{D}}_n(\alpha|\mu) > n - b\right\} > 2^{-n-c}$.*

Proof. We define the continuous sampling method C, where on input n, randomly samples n elements from λ_n. Let $d_n = \lambda_n\{\alpha : \overline{\mathbf{D}}_n(\alpha|\mu) > n - b\}$, where b is the constant in 5. Evoking this theorem, with $k = 0$,

$$\begin{aligned}
\Pr\left(\max_{\alpha \in C(2^n)} \overline{\mathbf{D}}_n(\alpha|\mu) > n - b\right) >& 1 - 2.5e^{-1} \\
1 - (1 - d_n)^{2^n} >& 1 - 2.5e^{-1} \\
1 - 2^n d_n <& 2.5/e \\
d_n >& (1 - 2.5/e)2^{-n} \\
\lambda_n\{\alpha : \overline{\mathbf{D}}_n(\alpha|\mu) > n - b\} >& 2^{-n-c}.
\end{aligned}$$

□

Chapter 8

Outliers in Dynamics

In this chapter, 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. 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.

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 13 and 14. The second step is to use conservation properties of $\mathbf{I}$, shown in Theorem 15, to achieve the main theorem of the chapter. In fact, as shown in Section 8.4, 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 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.

8.1 Sets of Numbers

Theorem 13 *For computable probability p over $\mathbb{N}$ and for $D \subset \{0,1\}^*$, $|D| = 2^s$, $m \in [0, s-1]$, 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 occured. 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 $cd_i 2^{-s}$, with $d_i = d \log i$. $\mathbf{E}[p(F_i)] \leq cd_i 2^{-s}$.

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

Thus a finite set F_i can be chosen such that $p(F_i) \leq 2cd_i 2^{-s}$ 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

$$\begin{aligned}s - \log i &< -\log p(a) - \mathbf{K}(a|p) + \mathbf{Ks}(D|p) + O(\log s),\\ s - \log i &< \mathbf{d}(a|p) + \mathbf{Ks}(D) + \mathbf{K}(p) + O(\log s + \log \mathbf{K}(p)). \qquad (8.1)\end{aligned}$$

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 8.1 can be proven. □

Corollary 11 *For computable probability p over $\mathbb{N}$ and for $D \subset \{0,1\}^*$, $|D| = 2^s$, $s < \max_{a \in D} \mathbf{d}(a|p) + \mathbf{Ks}(D) + \mathbf{K}(p) + 2\mathbf{K}(s) + O(\log \mathbf{K}(p))$.*

8.2 Left-Total Machines

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 $y \in \{0,1\}^{*\infty}$ if 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 $\alpha \in \{0,1\}^{*\infty}$ is to the "left" of $\beta \in \{0,1\}^{*\infty}$, and use the notation $\alpha \lhd \beta$, if there exists $x \in \{0,1\}^*$ 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 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]. More information on left-total machines can be found in Chapter 6. Without loss of generality, the complexity terms of this paper are defined with respect to the optimal left total machine U.

8.3 Sets of Infinite Strings

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

Theorem 14 *For computable probability P over $\{0,1\}^\infty$ and $Z \subset \{0,1\}^\infty$,$|Z| = 2^s$, $m \in [0, s-1]$, 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 13 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 13, relativized to b, produces 2^m elements $F \subseteq D$ such that for $x \in F$,

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

Using Lemma 15, relativized to b,

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

By Lemma 13,

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

Since $D \subseteq \{0,1\}^{\mathbf{bb}(b)}$, $\mathbf{K}(b|D, \|b\|) = O(1)$, 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

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

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 2 of the mutual information between infinite sequences,

$$\begin{aligned}
\mathbf{I}(D;\mathcal{H}) &<^+ \mathbf{I}(\langle Z\rangle : \mathcal{H}) + \mathbf{K}(D|\langle Z\rangle) \\
&<^+ \mathbf{I}(\langle Z\rangle : \mathcal{H}) + \mathbf{K}(\|b\|) + \mathbf{K}(s) \\
&<^+ \mathbf{I}(\langle Z\rangle : \mathcal{H}) + 2\log \mathbf{I}(D;\mathcal{H}) + \mathbf{K}(s) \qquad (8.2) \\
&<^{\log} \mathbf{I}(\langle Z\rangle : \mathcal{H}) + \mathbf{K}(s),
\end{aligned}$$

where Equation 8.2 is due to Lemma 14, noting $\mathbf{K}(b|D,\|b\|) = O(1)$. So there is an $\alpha \in Z$, $x \sqsubset \alpha$, with

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

□

8.4 Outliers in Dynamics

This section contains the main result of the chapter, that dynamical systems will exhibit outliers. To prove this fact, Theorem 15 will be leveraged, which details the conservation properties of the halting sequence $\mathcal{H}$.

For a continuous function $\delta : \{0,1\}^\infty \to \{0,1\}^\infty$, $\langle\delta\rangle$ is any infinite sequence $\delta' \in \{0,1\}^\infty$, such that if $\alpha \in \{0,1\}^\infty$ 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 $\{0,1\}^\infty$, $\langle\lambda\rangle$ is any infinite sequence λ' such that if $x \in \{0,1\}^*$ is on the auxillary tape and λ' is on the input tape of U, then U outputs $\lambda(x\{0,1\}^\infty)$. 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), $\mathbf{I}((\lambda,\delta) : \mathcal{H}) = \inf_{\langle\lambda\rangle,\langle\delta\rangle} \mathbf{I}(\langle\langle\lambda\rangle,\langle\delta\rangle\rangle : \mathcal{H})$.

Theorem 15 ([Ver21, Lev74, Gei12])

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

Theorem 16 (Outliers in Dynamics) *There exists $d \in \mathbb{N}$, where for computable probability μ over $\{0,1\}^\infty$ and dynamics (λ,δ), with $\mathbf{I}((\lambda,\delta) : \mathcal{H}) \neq \infty$, 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, $\mathbf{E}_{\alpha\sim\lambda}[s_\alpha - O(\log s_\alpha)] < \mathbf{I}((\lambda,\delta) : \mathcal{H}) + \mathbf{K}(\mu)$.*

Proof. Fix a starting state $\alpha \in \{0,1\}^\infty$ 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 $Z \subset \{0,1\}^\infty$ 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 14 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 15, 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 15, $\mathbf{E}_{\alpha \sim \lambda}[2^{\mathbf{I}(\langle \alpha, \delta \rangle : \mathcal{H})}] = \mathbf{E}_{\xi \sim \gamma}[2^{\mathbf{I}(\xi : \mathcal{H})}] \overset{*}{<} 2^{\mathbf{I}(\gamma : \mathcal{H})} \overset{*}{<} 2^{\mathbf{I}((\lambda, \delta) : \mathcal{H})} < \infty$. Thus by Theorem 15, 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 15, $\mathbf{E}_{\alpha \sim \lambda}\left[2^{s_\alpha} s_\alpha^{-O(1)}\right] \overset{*}{<} \mathbf{E}_{\alpha \sim \lambda}\left[2^{\mathbf{I}(\langle \alpha, \delta \rangle : \mathcal{H}) + \mathbf{K}(\mu)}\right] \overset{*}{<} \mathbf{E}_{\xi \sim \gamma}\left[2^{\mathbf{I}(\xi : \mathcal{H}) + \mathbf{K}(\mu)}\right]$ implies $\mathbf{E}_{\alpha \sim \lambda}\left[s_\alpha - O(\log s_\alpha)\right] <^+ \mathbf{I}(\gamma : \mathcal{H}) + \mathbf{K}(\mu) <^+ \mathbf{I}((\lambda, \delta) : \mathcal{H}) + \mathbf{K}(\mu)$. □

Corollary 12 (Computable Dynamics) *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, $\mathbf{E}_{\alpha \sim \lambda}\left[s_\alpha - O(\log s_\alpha)\right] < \mathbf{K}(\lambda, \delta) + \mathbf{K}(\mu)$.*

This follows from Theorem 16 and Theorem 15, where $\mathbf{I}((\lambda, \delta) : \mathcal{H}) <^+ \mathbf{I}(0^\infty : \mathcal{H}) + \mathbf{K}(\lambda, \delta) <^+ \mathbf{K}(\lambda, \delta)$.

8.5 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 Chapter 3, it was shown that computable sampling methods produce outliers. 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 17 *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 11 and Lemma 15, 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 $\{0,1\}^\infty$, where $\gamma_n(\langle D\rangle 0^\infty) = \Pr(A(2^n) = D)$. From Theorem 15,

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

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:

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

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 18 *For (possibly uncomputable) continuous sampling method A, computable probability P over $\{0,1\}^\infty$, $\Pr_{D\sim A(2^n)}[n-k > \max_{\alpha\in D}\mathbf{D}(\alpha|P)] < 2^{-k+\mathbf{I}(A:\mathcal{H})+O(\log n)+c_P}$.*

The proof follows almost identically to that of Theorem 17. We leave the details to the reader. Corollary 12 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 16, with some minor technicalities.

Bibliography

[CT91] T. Cover and J. Thomas. *Elements of Information Theory.* Wiley-Interscience, New York, NY, USA, 1991.

[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.

[G\"61] 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.

[G\'13] P. Gács. Lecture notes on descriptional complexity and randomness, 2013.

[Gei12] Philipp Geiger. *Mutual information and Gödel incompleteness.* PhD thesis, Heidelberg University, 10 2012.

[Kle52] S. C. Kleene. *Introduction to Metamathematics.*, page 318. North-Holland Publishing., Amsterdam, 1952.

[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.

[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.

[Tur36] M. Turing, A. On Computable Numbers, with an Application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, 2(42):230–265, 1936.

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