The Outliers Theorem Revisited

Samuel Epstein*

May 13, 2022

Abstract

An outlier is a datapoint set apart from a sample population. The outliers theorem in algorithmic information theory states given a computable sampling method, outliers have to appear. We present a simple proof to the outliers theorem, with exponentially improved bounds. We extend the outliers theorem to dynamical systems. Dynamical systems are guaranteed to hit ever larger outlier states with diminishing measure.

1 Introduction

The deficiency of randomness of an infinite sequence $\alpha \in \{0,1\}^\infty$ with respect to a computable measure P over $\{0,1\}^\infty$ is defined to be $\mathbf{D}(\alpha|P) = \log \sup_n \mathbf{m}(\alpha[0..n])/P(\alpha[0..n])$. The $\mathbf{m}$ term is the algorithmic probability. In the following theorem, we show that ever larger outlying sequences occur with diminishing probability.

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

This has special meaning when λ is the stationary measure of a dynamical system. The theorem was proven using a general template consistent with the Independence Postulate, [Lev13, Lev84]. This involves first proving that an object has mutual information with the halting sequence. The second step involves removing the mutual information term from the inequality. The removal of the information term can be done in a number of ways, and the dynamical systems theorem represents one such example.

In addition, we present a simple proof of the outliers theorem in [Eps21] with exponentially improved bounds. 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 uniformaly 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'. Fix such a t, and let $\overline{\mathbf{d}}_n(x|P) = \log t(x|n)$.

Theorem. *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:*

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

*JP Theory Group. samepst@jptheorygroup.org

In addition, we discuss related aspects of sampling methods. This includes sampling methods over infinite sequences and sampling methods that can not halt with positive probability.

2 Dynamical Systems

In this section, we prove that dynamical systems will hit ever larger outliers with diminishing probability. To do so, we use properties of the mutual information of an infinite sequence with the halting problem. Prefix free Kolmogorov complexity is $\mathbf{K}(x|y)$. The algorithmic probability is $\mathbf{m}(x|y)$. The mutual information between two finite strings is defined to be $\mathbf{I}(x:y) = \mathbf{K}(x) + \mathbf{K}(y) - \mathbf{K}(x,y)$. The halting sequence $\mathcal{H} \in \{0,1\}^\infty$ is the infinite sequence where the ith bit is 1 iff the ith program to the universal Turing machine halts. The mutual information between two infinite sequences α and β is

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

This definition originated from [Lev74]. The deficiency of randomness of an infinite sequence $\alpha \in \{0,1\}^\infty$ with respect to a computable probability measure P over $\{0,1\}^\infty$ is defined to be

$$\mathbf{D}(\alpha|P,x) = \log \sup_n \mathbf{m}(\alpha[0..n]|x)/P(\alpha[0..n]).$$

We have $\mathbf{D}(\alpha|P) = \mathbf{D}(\alpha|P,\emptyset)$. We require the following two theorems for the main proof of this section.

Theorem 1 ([Ver21, Lev74, Gei12]) $\Pr_\mu(\mathbf{I}(\alpha:\mathcal{H}) > n) \stackrel{*}{<} 2^{-n+\mathbf{K}(\mu)}$.

Theorem 2 ([Eps21]) *For 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 $s < \sup_{\alpha\in Z} \mathbf{D}(\alpha|P) + \mathbf{I}(\langle Z\rangle : \mathcal{H}) + O(\mathbf{K}(s) + \log \mathbf{I}(\langle Z\rangle : \mathcal{H}) + \mathbf{K}(P))$.*

Theorem 3 (Dynamical Systems) *For non-atomic computable measures λ and μ over $\{0,1\}^\infty$, for all $n \in \mathbb{N}$, $\lambda\{\alpha : \mathbf{D}(\alpha|\mu) > n\} > 2^{-n-\mathbf{K}(n,\mu,\lambda)-O(1)}$.*

Proof. Assume not. For all $c \in \mathbb{N}$, there exist computable non-atomic 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 $\beta \in D$, has $\mathbf{D}(\beta|\mu) \leq n$ 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 a 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 $\langle n,\mu,\lambda\rangle$. L By Theorem 1, with the universal Turing machine relativized to v, $\lambda^{n,c}(\{\gamma : \mathbf{I}(\gamma:\mathcal{H}|v) > m\}) \stackrel{*}{<} 2^{-m+\mathbf{K}(n,\mathbf{K}(n,\mu,\lambda),c,\lambda|v)} \stackrel{*}{<} 2^{-m+\mathbf{K}(c)}$. So there is a constant $f \in \mathbb{N}$, with

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

So, by probabilistic arguments, there exists $\alpha \in \{0,1\}^\infty$, such that α 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 2, relativized to v, there are constants $d, f \in \mathbb{N}$ where

$$\begin{aligned} m = \log |D| &< \max_{\beta \in D} \mathbf{D}(\beta|\mu, v) + 2\mathbf{I}(D : \mathcal{H}|v) + d\mathbf{K}(m|v) + f\mathbf{K}(\mu|v) \\ &<^+ \max_{\beta \in D} \mathbf{D}(\beta|\mu) + \mathbf{K}(n,\mu,\lambda) + 2\mathbf{K}(c) + d\mathbf{K}(m|v) + f\mathbf{K}(\mu|v) \\ &<^+ n + \mathbf{K}(n,\mu,\lambda) + d\mathbf{K}(m|v) + 2\mathbf{K}(c). \end{aligned} \tag{1}$$

So

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

Plugging the inequality for $\mathbf{K}(m|v)$ back into Equation 1 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 is a contradiction for large enough 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 deficency of randomness to be $\mathbf{D}_n(\alpha|Q) = \log T(\alpha|n)$.

Corollary 1 *For non-atomic uniformly computable measures $\{\lambda_i\}$ and $\{\mu_i\}$ 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 3 can be extended to uncomputable λ. The term $\langle\lambda\rangle \in \{0,1\}^\infty$ represents any encoding of λ that can compute $\lambda(x\{0,1\}^\infty)$ for $x \in \{0,1\}^*$ up to arbitrary precision.

Corollary 2

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

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 $\{0,1\}^\infty$. Let $\mathcal{B}$ be the Borel σ-algebra on $\{0,1\}^\infty$. Let λ and μ be computable measures over $\{0,1\}^\infty$. Let $(\{0,1\}^\infty, \mathcal{B}, \lambda)$ be a measure space and $T : \{0,1\}^\infty \to \{0,1\}^\infty$ be an ergodic measure preseving transformation. By the Birkoff theorem,

Corollary 3 *Starting λ-almost everywhere, $\stackrel{*}{>} \mathbf{m}(n,\mu,\lambda)2^{-n}$ states α visited by iterations of T have $\mathbf{D}(\alpha|\mu) > n$.*

3 Outliers Theorem

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

Lemma 1 *Let $P = P_1, P_2 \ldots$ be a uniformly computable sequence of measures 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 hence some finite subset also satisfies the conditions because of 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. Indeed, 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 less than $1/2$, this measure is less than $2e^{-M}$. By the union bound, the probability that at least one of the conditions are 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 $P = P_1, P_2 \ldots$ 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. Fix a search procedure that on input N and M finds a set $S_{N,M}$ that satisfies the conditions of Lemma 1. 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, since $P(\{x : t'(x|n) = 2^\ell\}) \leq 2^{-\ell-1}$, and thus $P(t'(x|n) \geq 2^\ell) \leq 2^{-\ell-1} + 2^{-\ell-2} + \ldots$ 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 d, we have $\overline{\mathbf{d}}_n(a|P) > n - k - O(1)$. □

3.1 Continuous Sampling Method

Let $\mu = \mu_1, \mu_2, \ldots$ be a uniformly computable sequence of measures over infinite sequences. In a similar way as for strings, the 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:*

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

Proof. The encoding of N unique sequences $\omega_1, \ldots, \omega_N$ into a single sequence ω is

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

For such a sequence, let $\overline{\omega} \in \{0,1\}^*$ be the smallest prefix of ω where N unique finite strings are described. The prefix free set of all such finite sequences over all ω (encoding of N sequences) is $J \subset \{0,1\}^*$. We define a discrete measure P_N over J, where $P_N(x) = \mu_N(\{\omega : \overline{\omega} = x\})$. This implies

$$\overline{\mathbf{D}}_n(\alpha|\mu) > \overline{\mathbf{d}}_n(\overline{\alpha}|P) - O(1).$$

The result for infinite sequences follows from Theorem 4.

3.2 Necessity of Double Exponential

Theorem 4 showed that the probability that $A(2^n)$ contains no strings of defiency less than $n-k$, decreases double exponential in k. We show that at least a double exponential probability is needed 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

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

Indeed, 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 probability at least 2^{-N}. Indeed, imagine that in a bag with $4N$ balls, N balls are marked. One selects N balls one by one. 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 $(T-N)/T \geq 1/2$.

3.3 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

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

Note that for at most half of the $(n+1)$-bit strings, we have $\overline{\mathbf{d}}_n(x|P) \geq 2$. On input 2^n, 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 ouptputs only strings with deficiency at most 1, and the claim is proven. However partial sampling methods do exhibit the following properties

Proposition 1 *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 semi-measure over sets of size 2^n such that $Q(D)$ equals the probability that $B(N) = D$. We show that

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

The result follows 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 semicomputable semi-measure Q_*:

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

with $f \leq 2^k$. We assume that the finite descriptions of $Q_1, \ldots, Q_f$ 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 1 for $M = k$. Let $S = \bigcup_{i \leq f} S_i$. Note that $P(S) \leq k2^{k+1-n}$. Hence 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 $[1, \ldots, f, *, \infty]$ with probabilities $[2^{-k}, \ldots, 2^{-k}, 2^{-k}r, 1 - (f + r)2^{-k}]$ 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 $j \in \{1, \ldots, f\}$ times $2e^{-k}$. □

4 Discussion

In the proof Theorem 3, a relativization technique can be used to convert an $O(\mathbf{K}(x))$ error term to a $\mathbf{K}(x)$ error term. This enables the removal of quantifiers from the theorem statement. This technique can be done by first relativizing inequalities to a shortest program that computes all the relevant parameters μ, λ, and n. Then the next part is to reconfigure all terms that have the parameters as conditional information, in this case the deficiency of randomness $\mathbf{D}(\alpha|\mu)$. This technique was also used in [Eps22].

References

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

[Eps22] Samuel Epstein. On the kolmogorov complexity of binary classifiers. *CoRR*, abs/2201.12374, 2022.

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

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

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

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

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