On the Existence of Outliers

Sam Epstein

March 2025

Abstract

Outliers are datapoints that are outside a dataset. This paper reviews work that outliers have to become emergent in datasets. This is true for computable and uncomputable sampling methods. Non-algorithmic phenomena, such as one's local weather forecast, are proven to exhibit outlying behavior. This is proven for numbers, infinite sequences, and metric spaces. In addition, we give upper bounds on the complexity of classifiers completely consistent with training samples.

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. Although there are known partial indicators, the determination of an outlier remains a subjective endeavor.

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

Outliers can be precisely defined algorithmically with respect to computable probability measures over natural numbers, infinite sequences, or computable metric spaces. The probability measure represents the model, and numbers, infinite sequences, and points in metric spaces, 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 is given by the *randomness deficiency* function. It is defined by

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

where $x \in \mathbb{N}$ is the data point and p is the probability measure. The term $\mathbf{K}$ is the Kolmogorov complexity of a string. The term $\mathbf{d}(x|p)$ is the difference between the length of a string's p-code and its optimal description. The function $2^{\mathbf{d}(\cdot|p)}$ is a universal lower-computable p-test.

Outliers can be defined algorithmically with respect to infinite sequences. The randomness deficiency of an infinite sequence $\alpha \in \{0,1\}^\infty$ with respect to a computable probability P is

$$\mathbf{D}(\alpha|P) = \sup_n \lfloor -\log P(\alpha[1..n] \rfloor - \mathbf{K}(\alpha[1..n]|P).$$

The function $2^{\mathbf{D}(\cdot|P)}$ is a universal lower-computable test. Outliers can be defined with respect to (computable) metric spaces $\mathcal{X}$ and arbitrary (not necessarily computable) Borel probabilities μ over $\mathcal{X}$. The randomness deficiency of a point $\alpha\mathcal{X}$ with respect to $(\mathcal{X}, \mu)$ is

$$\mathbf{D}(\alpha|\mu) = \log \mathbf{t}_\mu(\alpha).$$

The term $\mathbf{t}_\mu$ is a universal lower-computable μ-test over $\mathcal{X}$. The paper shows that methods that accumulate data points will have outliers in their outputs. The larger the sample drawn, the larger the outlier score of a data point in the sample. In Section 3, it is shown that all computable and uncomputable sampling methods will produce outliers over numbers and infinite sequences. In Section 4, it is shown that outliers are present in all physical observations, whether algorithmic or not.

1.1 Binary Predicates

In Section 6, we also prove the upper bounds on the size of the smallest program that computes a complete extension of a given binary predicate γ. We prove that for non-exotic predicates, this size is not greater than the number of elements of γ. Exotic predicates have high mutual information with the halting sequence, and thus no algorithm can generate such predicates.

2 Conventions

We use $x <^+ y$, $x >^+ y$ and $x =^+ y$ to denote $x < y + O(1)$, $x + O(1) > y$ and $x = y \pm O(1)$, respectively. Furthermore, $\overset{*}{<} f$, $\overset{*}{>} f$ denotes $< O(1)f$ and $> f/O(1)$. In addition, $x <^{\log} y$ and $x >^{\log} y$ denote $x < y + O(\log y)$ and $x + O(\log x) > y$, respectively. The information between two finite strings is $\mathbf{I}(x : y) = \mathbf{K}(x) + \mathbf{K}(y) - \mathbf{K}(x, y)$. The universal semi-measure is $\mathbf{m}$. The halting sequence is $\mathcal{H}$. The information that $\mathcal{H}$ has about $x \in \{0, 1\}^*$ is $\mathbf{I}(x; \mathcal{H}) = \mathbf{K}(x) - \mathbf{K}(x|\mathcal{H})$. The mutual information between two infinite sequences is $\mathbf{I}(\alpha : \beta) = \log \sum_{x,y \in \mathbb{N}} 2^{\mathbf{I}(x:y)} \mathbf{m}(x|\alpha)\mathbf{m}(y|\beta)$ [Lev74].

3 Sampling Methods

A discrete sampling method A is a probabilistic function that maps an integer n with probability 1 to a set containing 2^n different strings.

Theorem 1. *For computable probability P over $\mathbb{N}$, for computable discrete sampling method A, there is a constant $c \in \mathbb{N}$, where for all $n, k \in \mathbb{N}$,*

$$\Pr[n - \max_{a \in A(n)} \mathbf{d}(a|p) > k] < 2^{-k+O(\mathbf{K},n,k))+c}.$$

A continuous sampling method B takes in a parameter n and with probability 1, outputs 2^n unique infinite sequences.

Theorem 2. *For computable probability P over $\{0,1\}^\infty$, for computable continuous sampling method B, there exists a constant $c \in \mathbb{N}$, where for all $n, k \in \mathbb{N}$,*

$$\Pr[n - \max_{\alpha \in B(n)} \mathbf{D}(\alpha|P) > k] < 2^{-k+O(\log k + \mathbf{K}(n))+c}.$$

3.1 Uncomputable Sampling Methods

Sampling methods do not need to be computable to produce outliers. The only criterion is that their infinite encodings need to have low mutual information with the halting sequence. An encoding $\langle A \rangle$ of a discrete sampling method A is any infinite sequence such that it is on the input tape of a universal Turing machine U, and $\langle n, \omega \rangle$ is on the auxiliary tape, with $n \in \mathbb{N}$ and $\omega \in \{0,1\}^\infty$, U outputs 2^n unique 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 3. *For a discrete uncomputable sampling method A, computable probability p over $\mathbb{N}$, there is a constant $c_p \in \mathbb{N}$, where for all $n, k \in \mathbb{N}$,*

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

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

Theorem 4. *For a computable probability P over $\{0,1\}^\infty$, for uncomputable continuous sampling method B, there exists a constant $c_P \in \mathbb{N}$, where for all $n, k \in \mathbb{N}$,*

$$\Pr[n - \max_{\alpha \in B(n)} \mathbf{D}(\alpha|P) > k] < 2^{-k+\mathbf{I}(B:\mathcal{H})+O(\log n)+c_P}.$$

For non-atomic probabilities over infinite sequences that are the outputs of randomized algorithms, outliers are guaranteed to be asymptotically present.

Theorem 5. *For computable probabilities Q and non-atomic P over $\{0,1\}^\infty$ and $n \in \mathbb{N}$,*

$$P\{\alpha : \mathbf{D}(\alpha|Q) > n\} 2^{-n-\mathbf{K}(n,P,Q)-O(1)}.$$

4 Physical Observations

Section 3 showed that anomalies emerge from probabilistic computable and uncomputable methods. 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 in non-algorithmic phenomena, one can

use the Independence Postulate, **IP** [Lev84, Lev13]. IP is a finitary Church-Turing thesis, postulating that certain finite and infinite sequences cannot be easily be found with a short "physical address":

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

Thus, any construct α with high $\mathbf{I}(\alpha : \mathcal{H})$ cannot be found in the physical world. The following theorem shows that large sets of numbers which do not have outliers are exotic, in that they have high mutual information with the halting sequence.

Theorem 6. *For a computable probability p over $\mathbb{N}$, $D \subset N$, $|D| = 2^s$,*

$$s <^{\log} \max_{a \in D} \mathbf{d}(a|P) + \mathbf{I}(D; \mathcal{H}) + O(\log \mathbf{K}(p)).$$

Theorem 7. *For a computable probability P over $\{0,1\}^\infty$, $Z \subset \{0,1\}^\infty$, $|Z| = 2^s$,*

$$s <^{\log} \max_{\alpha \in Z} \mathbf{D}(\alpha|P) + \mathbf{I}(\langle Z \rangle : \mathcal{H}) + O(\log \mathbf{K}(P)).$$

Theorem 8. *For computable metric space $\mathcal{X}$, Borel probability μ, there is a constant c where for all $Z \subseteq \mathcal{X}$, $|Z| = 2^s$,*

$$s <^{\log} \max_{\alpha \in Z} \mathbf{D}(\alpha|\mu) + \mathbf{I}(\langle Z \rangle : \mathcal{H}) + c.$$

5 Dynamics

Outliers become emergent in continuous dynamics in computable metric spaces $\mathcal{X}$. For $\alpha \in \mathcal{X}$, $\mathbf{I}(\alpha : \mathcal{H})$ is the infinum of all encodings (in the standard way for computable metric spaces) of α with $\mathcal{H}$. A transformation group is a computable one-dimensional group $G^t : \mathcal{X} \to \mathcal{X}$ indexed by $t \in \mathbb{R}$.

Theorem 9. *Let L be the Lebesue measure over $\mathbb{R}$ and $(\mathcal{X}, \mu)$ be a computable metric space and Borel probability measure. Let $\alpha \in \mathcal{X}$ with finite $\mathbf{I}(\alpha : \mathcal{H})$. For transformation group G^t, there is a constant c with*

$$2^{-\mathbf{K}(n)-c} < L\{t \in [0,1] : \mathbf{D}(G^t\alpha|\mu) > n\} < 2^{-n}.$$

6 Binary Predicates

A binary predicate is defined to be a function of the form $f : D \to \{0,1\}$, where $D \subseteq \mathbb{N}$. We say that binary predicate $\lambda : \mathbb{N} \to \{0,1\}$ is a complete extension of γ, if for all $i \in \mathrm{Dom}(\gamma) = \lambda(i) = \gamma(i)$. Complete extensions are encoded as infinite sequences.

Theorem 10. *For binary predicate γ and the set Γ of complete extensions of γ, $\min_{g \in \Gamma} \mathbf{K}(g) <^{\log} |\mathrm{Dom}(\gamma)| + \mathbf{I}(\langle \gamma \rangle; \mathcal{H})$.*

References

[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.