Straight Lines and Algorithmic Statistics

Sam Epstein*

May 29, 2024

Abstract

In this paper we show that all strings with low mutual information with the halting sequence will have flat structure functions, in the context of algorithmic statistics.

1 Introduction

In statistics, one tries to determine a model (such as a parameter for a distribution) from data which is assumed to have noise. In the Minimum Description Principle [Gru07], the model that describes information with the shortest code is assumed to be the best model. The data is described as a two part code, where the first part is the model and the second part is the noise. In one of his last works, Kolmogorov suggested a two part code for individual strings $x \in \{0,1\}^*$ based off Kolmogorov Complexity. The first part (the model) is a set D containing x, the second part (the noise) is the code of x given D, of size $\lceil \log |D| \rceil$. Other works examined probabilities and also total computable functions as models [Vit02]. Kolmogorov suggested the following function at the Tallinn conference in Estonia, 1973.

$$\mathbf{H}_k(x) = \min_{p, \|p\| \leq k} \log |S|,$$

where $U(p) = S$ and $x \in S$. This definition is used for the *Structure Function*, a central definition of *Algorithmic Statistics* [VS15, VS17, VV04],

$$k \mapsto k + \mathbf{H}_k(x) - \mathbf{K}(x)$$

The function $\mathbf{K}$ is the prefix Kolmogorov complexity. The structure function is known to be equivalent to several other definitions [SSV24]. Furthermore, Theorem 1 of [VS17] showed that any shape of the structure function is possible.

This paper shows that the structure function is flat for all strings with low mutual information with the halting sequence. Assuming the *Independence Postulate*, [Lev84, Lev13], strings with non-negligible mutual information with the halting sequence are exotic, in that they cannot be found in nature.

*samepst@jptheorygroup.org

2 Results

The amount of information that the halting sequence $\mathcal{H} \in \{0,1\}^\infty$ has about $x \in \{0,1\}^*$ is $\mathbf{I}(x;\mathcal{H}) = \mathbf{K}(x) - \mathbf{K}(x|\mathcal{H})$. 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. In addition, $x <^{\log} y$ and $x >^{\log} y$ denote $x < y + O(\log y)$ and $x + O(\log x) > y$, respectively.

Lemma 1 ([Eps22]). *For partial computable function f, $\mathbf{I}(f(x);\mathcal{H}) <^+ \mathbf{I}(x;\mathcal{H}) + \mathbf{K}(f)$.*

Theorem 1. *For $x \in \{0,1\}^*$, $n = \mathbf{K}(x)$, for all $m \le n$, $m \in \mathbb{W}$, there is a set $S \ni x$ such that $\log |S| = 2^m$ and $\mathbf{K}(S) + m <^{\log} n + \mathbf{I}(x;\mathcal{H})$.*

Proof. We relativize the proof to $\langle m, n\rangle$, which can be done because of the precision of the theorem. Let $\Omega = \sum\{2^{-\|p\|} : U(p) \text{ halts}\}$ be Chaitin's Omega and $\Omega^t = \sum\{2^{-\|p\|} : U(p) \text{ halts in time } t\}$. For a string x, let $BB(x) = \min\{t : \Omega^t > 0.x + 2^{-\|x\|}\}$. Note that $BB(x)$ is undefined if $0.x + 2^{-\|x\|} > \Omega$. For $n \in \mathbb{N}$, let $\mathbf{bb}(n) = \max\{BB(x) : \|x\| \le n\}$. $\mathbf{bb}^{-1}(m) = \arg\min_n\{\mathbf{bb}(n-1) < m \le \mathbf{bb}(n)\}$. Let $bb(n) = \arg\max_x\{BB(x) : \|x\| \le n\}$.

Lemma 2. *For $n = \mathbf{bb}^{-1}(m)$, $\mathbf{K}(bb(n)|m,n) = O(1)$.*

Proof. Enumerate strings of length n, starting with 0^n, and return the first string y such that $BB(y) \ge m$. This string y is equal to $bb(n)$, otherwise $BB(y^-)$ is defined and $BB(y^-) \ge BB(y) \ge m$. Thus $\mathbf{bb}(n-1) \ge m$, causing a contradiction. □

Proposition 1.

1. $\mathbf{K}(bb(n)) >^+ n$.

2. $\mathbf{K}(bb(n)|\mathcal{H}) <^+ \mathbf{K}(n)$.

Let $\mathbf{K}^t(x) = \inf\{\|p\| : U(p) = x \text{ in } t \text{ steps}\}$. Let $t = \min\{t : \mathbf{K}^t(x) = n\}$. Let $s = \mathbf{bb}^{-1}(t)$ and $r = \mathbf{bb}(s)$. Let $b = bb(r)$. So by Lemma 2 and Proposition 1, $\mathbf{K}(b|x) = \mathbf{K}(s) = O(\log \|b\|)$.

Let $R = \{y : \mathbf{K}^r(y) = n\}$. Pad R with enough unique strings so that $|R| = 2^n$. Thus $\mathbf{K}(S|b) = O(1)$. Segment R into subsets S of size 2^m. Each such subset $S \subseteq R$ has Kolmogorov complexity

$$\begin{aligned}
\mathbf{K}(S) &<^+ n - m + \mathbf{K}(b) \\
&<^{\log} n - m + \mathbf{K}(b) - \mathbf{K}(b|\mathcal{H}) && (1)\\
&<^{\log} n - m + \mathbf{I}(x;\mathcal{H}) + O(\log b) && (2)\\
&<^{\log} n - m + \mathbf{I}(x;\mathcal{H}).
\end{aligned}$$

Equation 1 is due to Proposition 1. Equation 2 is due to Lemma 1. One such set S contains x and has the properties required by the theorem. □

References

[Eps22] S. Epstein. 22 examples of solution compression via derandomization. *CoRR*, abs/2208.11562, 2022.

[Gru07] P. Grunwald. *The Minimum Description Length Principle*. The MIT Press, 2007.

[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.

[SSV24] A. Semenov, A. Shen, and N. Vereshchagin. Kolmogorov's Last Discovery? (Kolmogorov and Algorithmic Statistics). *Theory of Probability & Its Applications*, 68(4):582–606, 2024.

[Vit02] P. Vitányi. Meaningful information. In *Algorithms and Computation*, pages 588–599, Berlin, Heidelberg, 2002. Springer Berlin Heidelberg.

[VS15] N. Vereshchagin and A. Shen. *Algorithmic Statistics Revisited*, pages 235–252. Springer International Publishing, 2015.

[VS17] N. Vereshchagin and A. Shen. *Algorithmic Statistics: Forty Years Later*, pages 669–737. Springer International Publishing, 2017.

[VV04] N. Vereshchagin and P. Vitanyi. Kolmogorov's structure functions and model selection. *IEEE Transactions on Information Theory*, 50(12):3265–3290, 2004.