# On Kolmogorov Structure Functions

Sam Epstein[*]

June 1, 2024

**Abstract**

All strings with low mutual information with the halting sequence will have flat Kolmogorov Structure Functions, in the context of Algorithmic Statistics. Assuming the Independence Postulate, strings with non-negligible information with the halting sequence are purely mathematical constructions, and cannot be found in the physical world. We also discuss issues with set-restricted Kolmogorov Structure Functions.

## 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 *structure function* at the Tallinn conference in Estonia, 1973.

$$\mathbf{H}_k(x) = \min\{\log |S| : x \in S, \mathbf{K}(S) \leq k\}.$$

The function $\mathbf{K}$ is the prefix Kolmogorov complexity. This definition is used for the following function, which is a central definition of *Algorithmic Statistics* [VS15, VS17, VV04],

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

This function's equivalence to several other definitions is the main theorem of Algorithmic Statistics [SSV24]. Furthermore, Theorem 1 of [VS17] showed that any shape of the structure function is possible.

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. Such strings are purely mathematical constructions.

[*]samepst@jptheorygroup.org

## 2 Bounds

We review the results of [GTV01], in particular Theorem of III.24, which I don't think is widely known. $\mathbf{m}(x)$ is the algorithmic probability. 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. For $x, y \in \{0,1\}^*$, $x \sqsubseteq y$ if $y = xz$ for some $z \in \{0,1\}^*$. $[A] = 1$ if mathematical statement $A$ is true, and $[A] = 0$ otherwise.

Let $S_k = \{x : \mathbf{K}(x) \leq k\}$. Let $N_k = |S_k|$ where $\log N_k =^+ k - \mathbf{K}(k)$, due to [GTV01]. Let $I_k^x$ be the index of $x$ in an enumeration of $S_k$. For $\mathbf{K}(x) = k$, let $m_x$ be the longest joint prefix of $I_k^x$ and $N_k$. So $m_x 0 \sqsubseteq I_k^x$ and $m_x 1 \sqsubseteq N_k$. Let $S_x = \{y : m_x 0 \sqsubseteq I_k^y\}$. So

$$\log |S_x| =^+ k - \mathbf{K}(k) - \|m_x\|$$
$$\mathbf{K}(S_x) <^+ \mathbf{K}(k) + \mathbf{K}(m_x) <^+ \mathbf{K}(k) + \|m_x\| + \mathbf{K}(\|m_x\|).$$

**Theorem 1** ([GTV01])**.**

$$\|m_x\| < \mathbf{K}(\mathbf{K}(x)) + \mathbf{I}(x;\mathcal{H}) + O(\log \mathbf{I}(x;\mathcal{H})).$$

*Proof.* Let $\nu(y) = c[\mathbf{K}(y) \leq k]\mathbf{m}(y)2^{\|m_y\|}/(\|m_y\|^2)$. For proper choice of $c$, $\nu$ is a semimeasure and computable relative to $\mathcal{H}$ and $k$. So $\mathbf{K}(x|\mathcal{H}, k) <^+ -\log \nu(x) =^+ \mathbf{K}(x) - \|m_x\| + 2\log \|m_x\|$. □

Note that with some additional effort, the $\mathbf{K}(\mathbf{K}(x))$ term can be eliminated.

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

**Claim 1.** *Thus there exists a set $S \ni x$ such that $\mathbf{K}(S) <^{\log} 2\mathbf{K}(\mathbf{K}(x)) + \mathbf{I}(x;\mathcal{H})$ and $\mathbf{K}(S) + \log |S| <^+ \mathbf{K}(x) + \mathbf{K}(\mathbf{K}(x)) + O(\log(\mathbf{I}(x;\mathcal{H}) + \mathbf{K}(\mathbf{K}(x))))$. This fact combined with the following proposition characterizes the Kolmogorov Structure Function.*

**Proposition 1.** *Let $S \ni x$. For all $s < \log |S|$ there exists a set $S' \ni x$ such that $|S'| \leq |S|2^{-s}$ and $\mathbf{K}(S') <^+ \mathbf{K}(S) + s + \mathbf{K}(s)$.*

## 3 Restricted Structure Functions

One potential method to create strings with non-simple Kolmogorov Structure Functions is to restrict the sets under consideration. Thus for a set of sets $\mathcal{S}$ such that

$$\mathbf{H}_k^{\mathcal{S}}(x) = \min\{\log|S| : x \in S \in \mathcal{S}, \mathbf{K}(S) \leq k\}.$$

This would banish the pesky set $S_x$ defined in the last section. This was studied in Section 6 of [VS15]. However there is an inherent obstacle to proving such functions can have any shape. Proofs to statements (such as Theorem 10 in [VS15]) of such effect use a shape function $R$ to (non-recursively) construct a string $x$ whose structure function has that shape $R$ (up to a degree of precision depending on $\mathcal{S}$). Thus the proof can be thought of as a program to produce $x$ given $R$ and $\mathcal{H}$, with $\mathbf{K}(x|\mathcal{H}) <^+ \mathbf{K}(R)$. Thus proofs saying that for every shape $R$ there is a set $x$ such that $\mathbf{H}_k^{\mathcal{S}}(x)$ has shape $R$ (up to a certain precision) also implies that $\mathbf{I}(x;\mathcal{H}) >^+ \mathbf{K}(x) - \mathbf{K}(R)$.

## References


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

[GTV01] P. Gács, J. Tromp, and P. Vitányi. Algorithmic statistics. *Information Theory, IEEE Transactions on*, 47:2443 – 2463, 2001.

[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.