Probabilities are Balanced

Samuel Epstein*

September 29, 2023

Abstract

Non-exotic probabilities will have balanced distributions with respect to Kolmogorov complexity. We prove a lower bound on their measure over simple strings.

It has been proven that large sets of strings are exotic if they all have similar complexities. By exotic, we mean their encoding has high mutual information with the halting sequence. Similarly if one probability over infinite strings gives large measure to sequences with low deficiency of randomness with respect to a second probability, then it is exotic. In this paper, we look at probabilities over strings of length n, and prove that they must give measure to simple strings. We first prove a simple bound. The main result is the tighter bound. This result also appears in the black holes section of the Algorithmic Physics manuscript at http://www.jptheorygroup.org.

Proposition 1 (Simple Bound) *There is a c where for probability p over $\{0,1\}^n$, for all $m > \mathbf{K}(p) + c$, $p\{x : \mathbf{K}(x) < m\} > 2^{m-2\mathbf{K}(m,p)-n-c}$.*

Proof. Order strings x of size n by $p(x)$ value, with largest values first, and breaking ties through any simple ordering on $\{0,1\}^n$. It must be the first 2^ℓ strings X has $p(X) \geq 2^{\ell-n-1}$ Otherwise the average value of $p(x)$, $x \in X$, is less than 2^{-n-1}. Thus for the remaining $2^n - 2^\ell$ strings Y, $p(y) < 2^{-n-1}$, So

$$\begin{aligned} p(\{0,1\}^n) &= p(X) + P(Y) \\ &< 2^{\ell-n-1} + (2^n - 2^\ell)(2^{-n-1}) \\ &= 2^{\ell-n-1} + 2^{-1} - 2^{\ell-n-1} \\ &= 1/2, \end{aligned}$$

which is a contradiction. Furthermore, the first 2^ℓ elements x have complexity $\mathbf{K}(x|p) <^+ \ell + \mathbf{K}(\ell)$ or $\mathbf{K}(x) <^+ \mathbf{K}(p,\ell) + \ell$. Let $m = \ell + \mathbf{K}(\ell,p) + O(1)$. By Proposition 3, $m - 2\mathbf{K}(m,p) <^+ \ell$. □

1 Tools

U is the reference universal Turing machine. $\mathbf{K}$ is the prefix free Kolmogorov complexity. $\mathbf{m}$ is the algorithmic probability. The information that the halting sequence $\mathcal{H}$ has about a string $x \in \{0,1\}^*$ is $\mathbf{I}(x;\mathcal{H}) = \mathbf{K}(x) - \mathbf{K}(x|\mathcal{H})$. The deficiency of randomness of a string x with respect to a probability p, is $\mathbf{d}(a|p) = -\log p(a) - \mathbf{K}(a|p)$.

*JP Theory Group. samepst@jptheorygroup.org

Theorem 1 ([Eps22]) $\mathbf{I}(f(a);\mathcal{H}) <^+ \mathbf{I}(a;\mathcal{H}) + \mathbf{K}(f)$.

Theorem 2 ([Eps23a]) *For probability p over $\{0,1\}^*$, $D \subset \mathbb{N}$, $|D| = 2^s$, $s < \max_{a\in D} \mathbf{d}(a|p) + \mathbf{I}(D;\mathcal{H}) + O(\log \mathbf{I}(D;\mathcal{H})) + \mathbf{K}(s) + O(\log \mathbf{K}(s,p))$.*

Theorem 3 ([Eps23b]) *For probability p over $\mathbb{N}$, computed by program q, $\mathbf{E}_{a\sim p}[2^{\mathbf{I}(a;\mathcal{H})}] \stackrel{*}{<} 2^{\mathbf{I}(q;\mathcal{H})}$.*

2 Results

Theorem 4 (Tighter Bound) *There is a $c \in \mathbb{N}$ where for probability p over $\{0,1\}^n$, for $m > \mathbf{K}(p) + c$, $p\{x : \mathbf{K}(x) < m\} > 2^{m-n+2\mathbf{I}(p;\mathcal{H})+3\mathbf{K}(n,m)}$.*

Proof. Without loss of generality, p can be assumed to have a range in powers of 2. Assume not, then there exist $\ell \in (\mathbf{K}(p)+c, n)$ such that $p\{x : \mathbf{K}(x) \le \ell\} < 2^{-k}$, where $k = n-\ell-c-2\mathbf{I}(p;\mathcal{H}) - 4\mathbf{K}(n,\ell)$ and c solely depends on the universal Turing machine. $\mathbf{K}(k) <^+ \$K(n,\ell,c,\mathbf{I}(p;\mathcal{H}))$. Suppose $\max\{p(x) : \mathbf{K}(x) > \ell\} \ge 2^{-k}$. Then

$$\mathbf{K}(p) + O(1) > \mathbf{K}\left(\arg\max_x p(x)\right) > \ell > \mathbf{K}(p) + c,$$

causing a contradiction, for choice of c dependent on U. Sample 2^{k-2} elements D without replacement. p^* is the probability of D, where $\mathbf{K}(p^*) <^+ \mathbf{K}(p,\ell,\mathbf{K}(\ell),c)$. Even if every element x chosen has $p(x) = 2^{-k-1}$, the total p mass sampled is not greater than

$$2^{k-1}2^{k-2} \le 2^{-3}.$$

The probability q that all $x \in D$ has $\mathbf{K}(x) > \ell$ is

$$q > \left(1 - 2^{-k}/(1-2^{-3})\right)^{2^{k-2}} > \left(1 - 2^{k+1}\right)^{2^{k-2}} = 1/2.$$

Thus, by Theorems 3 and 1,

$$\Pr_{S\sim p^*}[\mathbf{I}(S;\mathcal{H}) > \mathbf{I}(p^*;\mathcal{H}) + m] \stackrel{*}{<} 2^{-m},$$
$$\Pr_{S\sim p^*}[\mathbf{I}(S;\mathcal{H}) > \mathbf{I}((p,\ell,\mathbf{K}(\ell));\mathcal{H}) + m] \stackrel{*}{<} 2^{-m}.$$

So by probabilistic arguments, there exists $D \subset \{0,1\}^n$, where for all $x \in D$, $\mathbf{K}(x) > \ell$ and $\mathbf{I}(D;\mathcal{H}) <^+<^+ \mathbf{I}(p^*;\mathcal{H}) <^+ \mathbf{I}((p,\ell,\mathbf{K}(\ell));\mathcal{H})$. So by Theorem 2, applied to D and the uniform measure U_n over strings of length n,

$$\begin{aligned} k < \max_{a\in D} \mathbf{d}(a|U_n) + \mathbf{I}(D;\mathcal{H}) + O(\log \mathbf{I}(D;\mathcal{H})) + \mathbf{K}(k) + O(\log \mathbf{K}(U_n,k)) \\ n - \ell + \mathbf{K}(n) + c + 3\mathbf{K}(\ell,n) + 2\mathbf{I}(p;\mathcal{H}) < n - \ell + \mathbf{I}(p;\mathcal{H}) + \mathbf{K}(\ell) + O(\log(\mathbf{I}(p;\mathcal{H}) + \mathbf{K}(\ell))) \\ + \mathbf{K}(n) + \mathbf{K}(n,c,\ell,\mathbf{I}(p,\mathcal{H})) + O(\log \mathbf{K}(n,c,\ell,\mathbf{I}(p,\mathcal{H})) \\ c < \mathbf{K}(c) + O(\log \mathbf{K}(c)). \end{aligned}$$

which is a contradiction for c dependent solely on the universal Turing machine U. □

References

[Eps22] S. Epstein. The outlier theorem revisited. *CoRR*, abs/2203.08733, 2022.

[Eps23a] Samuel Epstein. On the Existence of Anomalies. *CoRR*, abs/2302.05972, 2023.

[Eps23b] Samuel Epstein. The Kolmogorov Birthday Paradox. *Theoretical Computer Science*, 963, 2023.

A Helper Propositions

Proposition 2 *For every $c, n \in \mathbb{N}$, there exists $c' \in \mathbb{N}$ where for all $a, b \in \mathbb{N}$, if $a < b + n \log a + c$ then $a < b + 2n \log b + c'$.*

Proof.

$$\begin{aligned} \log a &< \log b + \log\log a + \log cn \\ 2\log a - 2\log\log a &< 2\log b + 2\log cn \\ \log a &< 2\log b + 2\log dn. \end{aligned}$$

Combining with the original inequality

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

where $c' = 2n \log cn + c$. □

Proposition 3 *For all $d \in \mathbb{N}$ there is a $d' \in \mathbb{N}$ where if $x + \mathbf{K}(x, z) + d > y$ then $x + d' > y - 2\mathbf{K}(y, z)$.*

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

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

where $d'' = d' + O(1) + d$. □