arXiv:2302.05972v1 [cs.CC] 12 Feb 2023

论异常的存在

Samuel Epstein*

2023年2月14日

摘要

独立性公设（IP）是一种有限形式的丘奇-图灵论题，指出数学序列独立于物理序列。独立性公设意味着异常的存在。

1 引言

异常是指一种测量结果，它与其他观测结果的偏差如此之大，以至于令人怀疑它是由不同机制产生的。在此前的研究中，[Eps21], 已经证明，算法抽样方法必然会产生异常。然而，有些抽样方法过于复杂，不能视为算法。例如，你当地的天气预报。本文运用独立性公设——一种有限形式的丘奇-图灵论题——来探讨这一尚未解决的问题。离群值必然会出现在物理世界中。

2 相关工作

柯尔莫哥洛夫复杂度的研究源于 [Kol65]的工作。柯尔莫哥洛夫复杂度的规范自定界形式由 [ZL70]和 [Cha75]提出。独立性公设由[Lev84、Lev13] 提出，最初用于证明皮亚诺算术的一致扩张不可能存在，之后在 [BP16]中得到了推广。本文采用了与停机序列的互信息这一概念，关于该术语的更多信息可参见 [Eps21、VV02]。

3 约定

函数 $\mathrm{K}(x|y)$ 是条件前缀柯尔莫哥洛夫复杂度。两个字符串 x、$y \in \{0,1\}^*$ 之间的互信息为 $\mathrm{I}(x:y) = \mathrm{K}(x) + \mathrm{K}(y) - \mathrm{K}(x,y)$。对于关于 $\mathbb{N}$ 的概率 p，随机性亏损为 $\mathrm{d}(a|p,b) = \lfloor -\log p(a) \rfloor - \mathrm{K}(a|\langle p \rangle, b)$，它衡量在得到结果 a [Gác21]后，对假设 p 的反驳程度。$\mathrm{d}(a|p) = \mathrm{d}(a|p,\emptyset)$。停机序列 $\mathcal{H} \in \{0,1\}^\infty$ 关于 $a \in \{0,1\}^*$ 所含的信息量为 $\mathrm{I}(a;\mathcal{H}) = \mathrm{K}(a) - \mathrm{K}(a|\mathcal{H})$。我们用 $<^+ f$ 表示 $<f+O(1)$，用 $<^{\log} f$ 表示 $<f+O(\log(f+1))$。随机性为 $\Lambda(a|b) = \min\{\mathrm{K}(Q|b) + 3\log \max\{\mathrm{d}(a|Q,b)\ 1\} : Q$ 具有有限支撑，且值域包含于 $\mathbb{Q}\}$。$\Lambda(a|b) < \Lambda(a) + O(\log \mathrm{K}(b))$。以下定义引自 [Lev74] 。

定义 1（信息）对于无限序列 α、$\beta \in \{0,1\}^\infty$，其互信息定义为 $\mathrm{I}(\alpha:\beta) = \log \sum_{x,y \in \{0,1\}^*} 2^{\mathbf{I}(x:y) - \mathbf{K}(x|\alpha) - \mathbf{K}(y|\beta)}$。

*JP Theory Group。samepst@jptheorygroup.org

独立性公设（**IP**），[Lev84，Lev13], 是关于两个序列之间共享的信息量的一个不可证不等式。**IP**是一种有限形式的丘奇–图灵论题，假定自然界中不存在某些无限序列和有限序列，也就是说，它们具有很高的“物理地址”。

IP：设 α 是由一个 n 位数学陈述定义的序列，而序列 β 可以通过一个 k 位指令集在物理世界中找到。则对于某个较小的绝对常数 c，有 $\mathrm{I}(\alpha:\beta) < k+n+c$。

4 结果

文献中有许多证明表明，非随机数与停机序列具有较高的互信息。其中一个详细证明见 [Eps21]。

引理 1 $\Lambda(x) <^{\log} \mathrm{I}(x;\mathcal{H})$。

引理 2 对于 $\mathbb{N}$ 上的概率 p，$D\subset\mathbb{N}$，$|D| = 2^s$，$s < \max_{a\in D}\mathrm{d}(a|p) + \Lambda(D) + \mathrm{K}(s) + O(\log \mathrm{K}(s,p))$。

证明。我们将通用图灵机相对于 $\langle s,\ p\rangle$ 进行相对化。设 Q 是一个实现 $\Lambda(D)$ 的概率测度，满足 $d = \max\{\mathrm{d}(D|Q),\ 1\}$。设 $F\subseteq\mathbb{N}$ 是一个随机集合，其中每个元素 $a\in\mathbb{N}$ 都独立地以概率 $cd2^{-s}$ 被选中，$c\in\mathbb{N}$ 将在稍后选定。$\mathrm{E}[p(F)] \le cd2^{-s}$。此外

$$\mathbf{E}[Q(\{G : |G| = 2^s, G\cap F = \emptyset\})] \le \sum_G Q(G)(1-cd2^{-s})^{2^s} < e^{-cd}.$$

因此，可以选取有限的 $W\subset\mathbb{N}$，使得 $p(W) \le 2cd2^{-s}$ 且 $Q(\{G:\ |G| = 2^s, G\cap W = \emptyset\}) \le e^{1-cd}$。$D\cap W\neq\emptyset$，否则，使用 Q-检验，若 $(|G| = 2^s, G\cap W = \emptyset)$，则 $t(G) = e^{cd-1}$；否则 $t(G) = 0$，我们得到

$$\begin{aligned}\mathbf{K}(D|Q,d,c) &<^+ -\log Q(D) - (\log e)cd\\ (\log e)cd &<^+ -\log Q(D) - \mathbf{K}(D|Q) + \mathbf{K}(d,c)\\ (\log e)cd &<^+ d + \mathbf{K}(d,c),\end{aligned}$$

当 c 足够大时，这会导致矛盾。因此存在一个 $a\in D\cap W$，其中

$$\begin{aligned}\mathbf{K}(a) &<^+ -\log p(a) + \log d - s + \mathbf{K}(d) + \mathbf{K}(Q)\\ s &<^+ \mathbf{d}(a|p) + \Lambda(D).\end{aligned}$$

明确写出 $\langle s,p\rangle$ 的相对化后，

$$\begin{aligned}s &< -\log p(a) - \mathbf{K}(a|s,p) + \Lambda(D|s,p)\\ s &< \max_{a\in D}\mathbf{d}(a|p) + \Lambda(D) + \mathbf{K}(s)\\ &\quad + O(\log \mathbf{K}(s,p)).\ \square\end{aligned}$$

令 $\tau\in\mathbb{N}^{\mathbb{N}}$ 表示一个观测序列。实际上，观测信息是有限的。但观测可以被视为潜在无限的，并以无穷序列表示。假设 τ 中有无限多个互异数，$\tau(n)$ 就是 τ 中前 2^n 个互异数。

定理 1 对于 $\mathbb{N}$ 上的概率 p， $\tau \in \mathbb{N}^{\mathbb{N}}$，令 $s_{\tau,\ p} = \sup_n \left(n - 3\mathrm{K}(n) - \max^{a\in\tau(n)} \mathrm{d}(a|p)\right)$。则 $s_{\tau,\ p} <^{\log} \mathrm{I}(\langle\tau\rangle :\ \mathcal{H}) + O(\log \mathrm{K}(p))$。

证明。根据引理1 和2，以及$\mathbf{I}(x;\mathcal{H}) <^{+} \mathbf{I}(\alpha :\ \mathcal{H}) + \mathbf{K}(x|\alpha)$,

$$
\begin{aligned}
n &< \max_{a\in\tau(n)} \mathbf{d}(a|p) + \mathbf{I}(\tau(n);\mathcal{H}) + +\mathbf{K}(n) + O(\log \mathbf{I}(\tau(n);\mathcal{H})\mathbf{K}(p)\mathbf{K}(n)),\\
n &< \max_{a\in\tau(n)} \mathbf{d}(a|p) + 2\mathbf{K}(n) + \mathbf{I}(\langle\tau\rangle : \mathcal{H}) + O(\log \mathbf{I}(\langle\tau\rangle : \mathcal{H})\mathbf{K}(p)\mathbf{K}(n)),\\
n - 3\mathbf{K}(n) &- \max_{a\in\tau(n)} \mathbf{d}(a|p) <^{\log} \mathbf{I}(\langle\tau\rangle:\mathcal{H}) + O(\log \mathbf{K}(p)).\square
\end{aligned}
$$

设 k 是 τ的一个物理地址。 $\mathcal{H}$ 可以用一个简短的数学陈述来描述。根据定理1 和**IP**,

$$
s_{\tau,p} <^{\log} \mathbf{I}(\langle\tau\rangle : \mathcal{H}) + O(\log \mathbf{K}(p)) <^{\log} k + c + O(\log \mathbf{K}(p)).
$$

很难找到异常较少的观测，而不可能找到没有异常的观测。

5 讨论

本文表明，观测序列 τ 将包含不断增加的异常 a，即对于某个概率 p，$\mathrm{d}(a|p)$ 很高；否则，它无法通过物理地址到达。本文将观测建模为自然数的无限序列。不过，经过一些工作，可以将其推广到实数的无限序列。进一步的研究可将“离群值”或“异常”的概念推广到更一般的拓扑，例如可计算度量空间。

参考文献

[BP16] L. Bienvenu 和 C. Porter。《深层类别》。《符号逻辑通报》，22(2):249–286，2016 年。[Cha75] G. J. Chaitin。与信息论形式上等同的程序大小理论。《ACM 学报》，22(3):329–340，1975 年。[Eps21] Samuel Epstein。所有抽样方法都会产生离群值。IEEE《信息论汇刊》，67(11):7568–7578，2021 年。[Gác21] Peter Gács。描述复杂性与随机性讲义。CoRR，abs/2105.04704，2021 年。[Kol65] A. N. Kolmogorov。信息定量定义的三种方法。《信息传输问题》，1:1–7，1965 年。[Lev74] L. A. Levin。信息守恒（不增长）定律与概率论基础的若干方面。《信息传输问题》，10(3):206–210，1974 年。[Lev84] L. A. Levin。随机性守恒不等式；数学理论中的信息与独立性。《信息与控制》，61(1):15–37，1984 年。

[Lev13] L. A. Levin。禁止信息。ACM《计算机协会会刊》，60(2)，2013 年。[VV02] N. Vereshchagin 和 P. Vitányi。柯尔莫哥洛夫结构函数与模型选择。IEEE《信息论汇刊》，50:3265–3290，2002 年。[ZL70] A. K. Zvonkin 和 L. A. Levin。有限对象的复杂性，以及通过算法理论发展信息与随机性概念。《俄罗斯数学调查》，第 11 页，1970 年。