Kolmogorov complexity, prime numbers, and complexity lower bounds

LMS Computer Science Colloquium

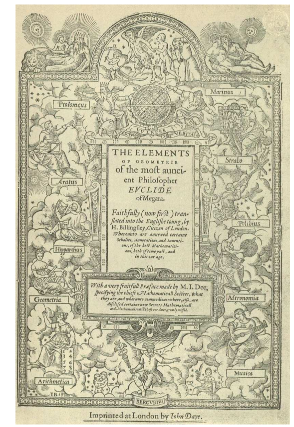
November/2020

Igor Carboni Oliveira

University of Warwick

Research funded by a Royal Society University Research Fellowship

Overview



Euclid's *Elements*

Wikipedia

Maths 1. Are there infinitely many prime numbers with "simple" descriptions?

CS 2. Is it hard to detect patterns in data?

Quanta

Maths/CS 3. Is there a fast deterministic algorithm that, given n, outputs an n-bit prime?

This talk:

New insights using a **probabilistic** extension of (time-bounded) Kolmogorov complexity

Background and Motivation

1. Number Theory: Mersenne Primes

Primes of the form $M_n = 2^n - 1$.

51 Mersenne primes are known. The largest known prime is $2^{82589933} - 1$.

In **binary representation**, a Mersenne prime is of the form $111111 \ldots\ldots\ldots\ldots\ldots\ldots 11111$.

Mersenne primes admit a <u>short</u> and <u>effective</u> representation.

"Simplest" possible representation of an n-bit prime.

Q. Are there infinitely many **Mersenne primes**? $\approx$ ***Q.*** Are there infinitely many **primes of "minimum description length"**?

Time-bounded Kolmogorov Complexity

Mersenne primes admit a short and effective representation. *Kolmogorov Complexity?*

$$0010101001001001001001001010000100101100101010101010 1$$

Levin (1984) proposed the following notion of complexity for strings.

$$\mathsf{Kt}(x) \stackrel{\text{def}}{=} \min_{\substack{\text{TM } M,\text{ time } t\\ M \text{ prints } x \text{ in time } t}} \underbrace{|M|}_{\text{short}} + \underbrace{\log t}_{\text{effective}}$$

For every n-bit string x: $\log n \;\leq\; \mathsf{Kt}(x) \;\leq\; n + O(\log n)$

$$111\ldots111$$

2. Complexity Theory: Intractability

Problems about the **complexity of strings** play a significant role in theory of computing.

(e.g. *learning* & *cryptography*)

Undecidable

Kolmogorov complexity

Given x, estimate C(x)

$\log t$ function is a threshold

Exponential Time vs Polynomial Time

Levin's Kt complexity

$$\mathsf{Kt}(x) \stackrel{\text{def}}{=} \min_{\substack{\text{TM M, time t} \\ \text{M prints x in time t}}} |M| + \log t$$

Given x, estimate Kt(x)

Q. ***Is it in polynomial time?***

e.g. [ABKvMR'06]

3. Algorithms: Deterministic constructions

POLYMATH 4

$$[\underbrace{100\ldots000}_{n \text{ bits}}, 111\ldots111]_2 = [2^{n-1}, 2^n - 1]_{10}$$

Challenge of deterministically generating primes: Given n, output an n-bit prime.

Best known deterministic algorithm runs in time $2^{n/2}$. [Lagarias-Odlyzko'87]

$\Longrightarrow$ For every large n, there is an n-bit prime p_n with $\mathsf{Kt}(p_n) \leq \frac{n}{2} + o(n)$.

"Simple objects are easier to find"

If there is a sequence p_n of n-bit primes with $\mathsf{Kt}(p_n) \leq \gamma_n$ then primes can be deterministically generated in time $\approx 2^{\gamma_n}$.

Summary

"Simplicity" as bounded Kt complexity (e.g. Mersenne primes).

Connections to basic questions in Maths/CS.

Q. Are there n-bit primes of Kt complexity $o(n)$?

Q. It is hard to estimate $\mathsf{Kt}(x)$ of a given string x?

Q. Deterministic prime generation in time $2^{o(n)}$?

These remain longstanding problems relevant to number theory, algorithms, and complexity.

A theory of probabilistic representations

[O-Santhanam'17] Pseudodeterministic constructions in subexponential time.

[O'19] Randomness and intractability in Kolmogorov complexity.

[Lu-O'20] An efficient coding theorem via probabilistic representations and its applications.

Definition of rKt complexity

[O'19] A randomized analogue of Levin's Kt complexity:

$$\mathsf{rKt}(x) \stackrel{\text{def}}{=} \min_{\substack{\underline{\text{randomized}}\text{ TM } M\text{, time } t \\ \Pr_M[\,M \text{ prints } x \text{ in time } t\,] \geq 2/3}} |M| + \log t$$

A **short** and **effective** **probabilistic** procedure that is likely to generate the observed data.

Basic properties of rKt

$$\mathsf{rKt}(x) \stackrel{\text{def}}{=} \min_{\substack{\text{randomized TM } M, \text{ time } t \\ \Pr_M[\,M \text{ prints } x \text{ in time } t\,] \geq 2/3}} |M| + \log t$$

$$\mathsf{Kt}(x) \stackrel{\text{def}}{=} \min_{\substack{\text{TM } M, \text{ time } t \\ M \text{ prints } x \text{ in time } t}} |M| + \log t$$

For every string x, $\mathsf{rKt}(x) \leq \mathsf{Kt}(x)$.

Q. ***Are there strings that admit a more succinct representation using randomness?***

[O'19] If $\mathsf{E} \not\subseteq \text{i.o.}\mathsf{SIZE}[2^{\varepsilon n}]$, for every string x we have $\mathsf{rKt}(x) = \Theta(\mathsf{Kt}(x))$.

As far as we know, gap between rKt and Kt could be maximum.

Proxy measure to investigate Kt?

A theory of probabilistic representations

[O-Santhanam'17] Pseudodeterministic constructions in subexponential time.

[O'19] Randomness and intractability in Kolmogorov complexity.

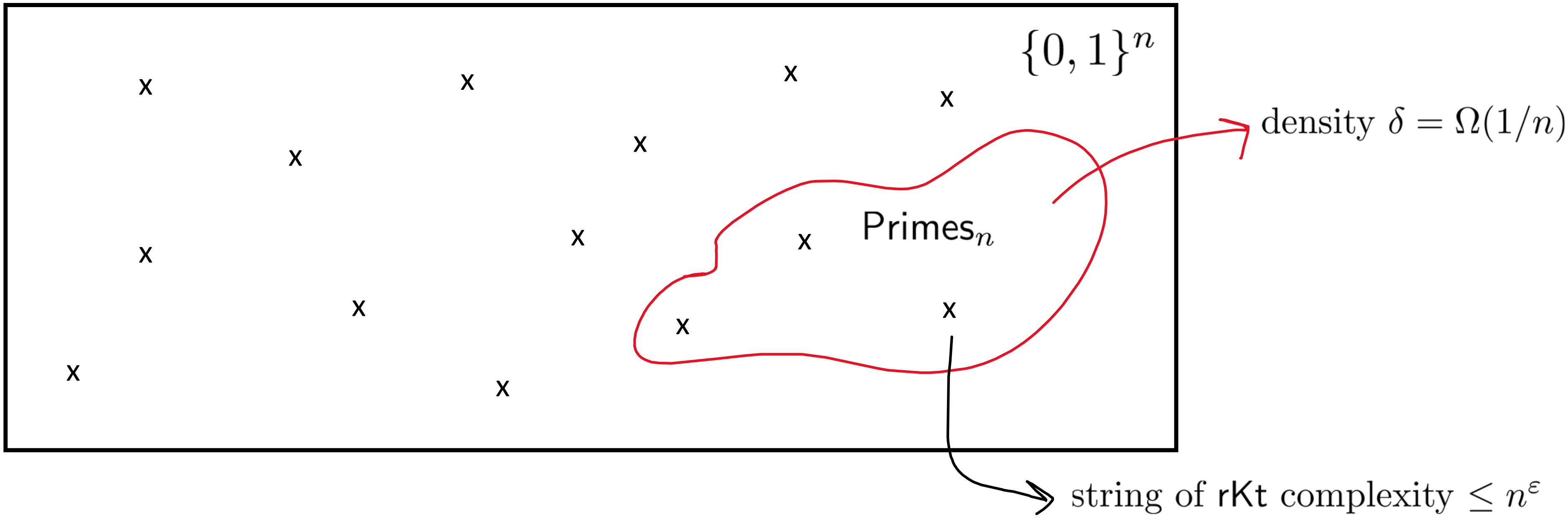
[Lu-O'20] An efficient coding theorem via probabilistic representations and its applications.

Probabilistic representations can see patterns in prime numbers

Recall: Open to show $\exists$ primes of Kt complexity $< n/2$.

Theorem [O-Santhanam'17, O'19]. $\forall \varepsilon > 0$, for infinitely many values of n, $\exists$ n-bit prime p_n such that $\mathsf{rKt}(p_n) \leq n^{\varepsilon}$.

Informally, some primes are structured enough to admit "short" and "effective" probabilistic representations.

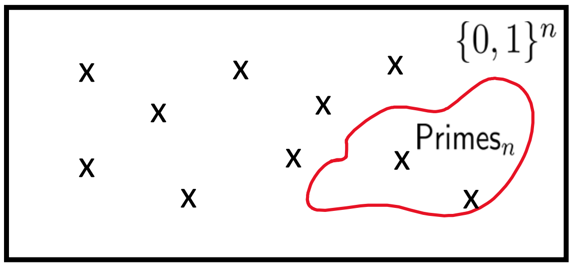


Intuition: A random n-bit string "hits" Primes_n with probability δ.

Constuct a pseudorandom distribution $\mathcal{D}$ supported over $\mathcal{R}^{\mathsf{rKt}}_{\leq n^{\varepsilon}}$.

"Tests": $\mathsf{Primes}_n \in \mathsf{DTIME}[n^k]$.

Pseudorandomness



$\mathcal{D}$ ε-fools a class of tests $\mathcal{T}$ if:

For every function $f\colon \{0,1\}^n \to \{0,1\}$ in $\mathcal{T}$,

$$\left| \Pr_{x\sim\{0,1\}^n}[f(x)=1] - \Pr_{y\sim\mathcal{D}}[f(y)=1] \right| \leq \varepsilon.$$

$G\colon \{0,1\}^s \to \{0,1\}^n$, $s \ll n$, generates a distribution $\mathcal{D} \equiv G(\mathcal{U}_s)$.

Example. $\mathcal{T} = \{x_1, \ldots, x_n\}$

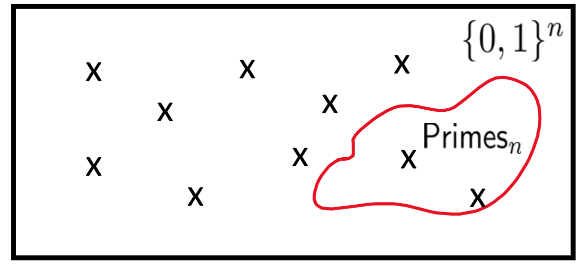
$G\colon \{0,1\}^1 \to \{0,1\}^n$

$$G(0) = 000\ldots000$$
$$G(1) = 111\ldots111$$

Fact: G ε-fools $\mathcal{T}$ for $\varepsilon = 0$.

$G\colon \{0,1\}^s \to \{0,1\}^n$, $s \ll n$, generates a distribution $\mathcal{D} \equiv G(\mathcal{U}_s)$.

"Fools" a class of tests $\mathcal{T}$.



Crucial: n-bit outputs of G have low rKt complexity: "explained" by a seed of length s.

Long list of works on PRGs in TCS:

- **Unconditional** PRGs against **"weak"** tests.
- **Conditional** PRGs against **"expressive"** tests.

[**IW'97**] $\mathsf{E} \not\subseteq$ i.o.SIZE$[2^{\lambda n}]$ $\longrightarrow$ $\exists\, G\colon \{0,1\}^{O(\log n)} \to \{0,1\}^n$ that fools SIZE$[n^k]$.

$(\star)$ Hardness Assumption (Pseudo)Randomness

$(\star)$

Hardness Assumption $\Longrightarrow$ (Pseudo)Randomness $\Longrightarrow$ Some primes have low rKt

NOT $(\star) \Longrightarrow$ "Easiness"

Time-bounded computations are more powerful.

Certain "patterns" in primes might become evident.

Example: The lexicographic first n-bit prime.

Need to find a "sweet spot" $(\star)$ that is useful.

$(\star) \quad \equiv \quad \boxed{\textsf{PSPACE} \overset{?}{\subseteq} \textsf{BPP}}$

yes

no

Fact: PSPACE computations can detect first n-bit prime.

(probabilistic representation)

[IW98, TV07] PRG G under hardness assumption that fools $\mathsf{DTIME}[n^k]$ infinitely often.

$\mathsf{Support}(G) \subseteq \mathcal{R}^{\mathsf{rKt}}_{n^{\varepsilon}}$

$\mathsf{Support}(G)$ intersects Primes_n

Infinitely many primes of bounded rKt complexity

A theory of probabilistic representations

[O-Santhanam'17] Pseudodeterministic constructions in subexponential time.

[O'19] Randomness and intractability in Kolmogorov complexity.

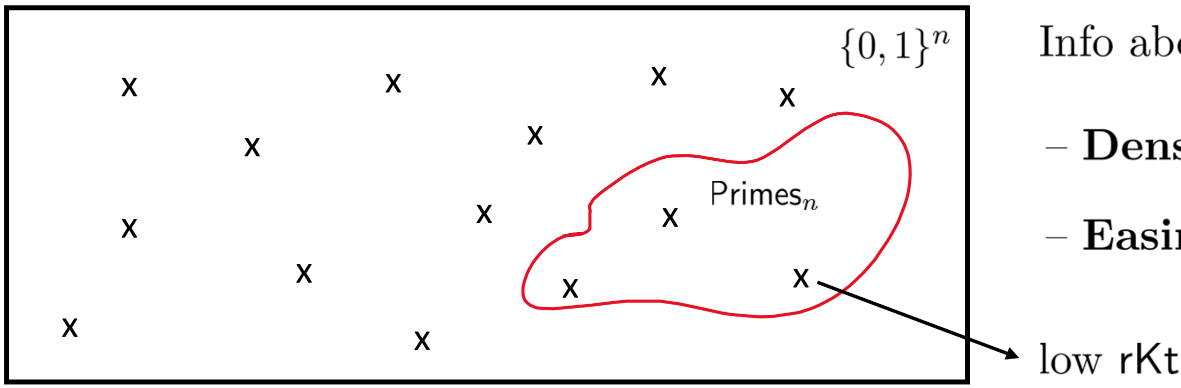
[Lu-O'20] An efficient coding theorem via probabilistic representations and its applications.

Is it hard to detect patterns?

Theorem [O'19]. $\forall \varepsilon > 0$, there is no randomised algorithm running in quasi-polynomial time that accepts strings in $\mathcal{R}^{\mathsf{rKt}}_{\leq n^{\varepsilon}}$ and rejects strings in $\mathcal{R}^{\mathsf{rKt}}_{\geq .99n}$

We cannot feasibly distinguish "*structured*" strings from "*random*" strings.

Proof of a weaker result: The set $\mathcal{R}^{\mathsf{rKt}}_{>n^{\varepsilon}}$ is not in P.



Info about Primes_n needed in previous proof:

- **Density:** $\delta \geq 1/\mathsf{poly(n)}$
- **Easiness:** $\mathsf{Primes}_n \in \mathsf{P}$

Lemma. Any **dense** and **easy** set contains, infinitely often, strings x with $\mathsf{rKt}(x) \leq n^{\varepsilon}$.

Proof of a weaker result: The set $\mathcal{R}^{\mathsf{rKt}}_{>n^{\varepsilon}}$ is not in P.

> **Lemma.** Any **dense** and **easy** set contains, infinitely often, strings x with $\mathsf{rKt}(x) \leq n^{\varepsilon}$.

$\mathcal{R}^{\mathsf{rKt}}_{>n^{\varepsilon}}$ is **dense**.

If $\mathcal{R}^{\mathsf{rKt}}_{>n^{\varepsilon}}$ is also **easy** (in P), then we contradict the lemma.

A more delicate argument is used for the **gap problem** and against **BPTIME[quasi-poly]**.

A theory of probabilistic representations

[O-Santhanam'17] Pseudodeterministic constructions in subexponential time.

[O'19] Randomness and intractability in Kolmogorov complexity.

[Lu-O'20] An efficient coding theorem via probabilistic representations and its applications.

Perspective

rKt has enabled results that remain intriguing questions for Levin's **Kt** complexity.

existence of shorter representations

intractability of estimating rKt

Q. Can we further advance **time-bounded Kolmogorov complexity** using probabilistic representations?

Pillars of Kolmogorov Complexity

Three essential results in Kolmogorov complexity:

Time-bounded version?

Language Compression Theorem —Hardness Assumption→

$$C(x) \leq |L^{=n}| + O(\log n)$$

See e.g. [Troy Lee, PhD Thesis]

Symmetry of Information —Hardness Assumption→

$$C(xy) \approx C(x) + C(y|x) \approx C(y) + C(x|y)$$

Source Coding Theorem →

$$C(x) \leq \log(1/\delta(x)) + O(1)$$

Coding Theorem

Shannon's Information Theory

Distributions, entropy, compression, etc.

$\boxed{\text{Coding Theorem in Kolmogorov Complexity}}$

Kolmogorov Complexity

Individual strings and their complexities.

$$\begin{array}{l}\text{An object } x \text{ can be sampled} \\ \text{with probability } \delta\end{array} \quad \Longrightarrow \quad \begin{array}{l}x \text{ admits a representation} \\ \text{of length} \approx \log(1/\delta)\end{array}$$

Interested in establishing an **unconditional time-bounded** version of the **Coding Theorem**.

An *Efficient* Coding Theorem for rKt

[Zhenjian Lu-O'20] "**Samplable** objects admit **short** and **effective** representations."

Randomized Algorithm $A(1^m)$

Runs in time $T(m)$

Outputs string x with probability $\geq \delta$

$$\Longrightarrow \quad \mathsf{rKt}(x) = O_A\big(\log(1/\delta) + \log(T) + \log m\big)$$

Efficient generation of representation: Given x, A, m, and δ, we can compute in time $\mathsf{poly}(|x|, |A|, \log m, \log(1/\delta))$ and with probability $\geq .99$ a valid rKt representation.

"Magic": Running time has no dependence on T.

Extremely useful in applications!

Application: Efficient universal compression

$x =$ 0010101000101111010010010111100010101010001010100101111010010100

"There is a way of compressing it to k bits" (in the sense of **rKt**).

$\mathsf{rKt}(x) \leq k$

$\Longrightarrow$ We can output in polynomial time with probability $\geq .99$ a valid rKt encoding of x of complexity $O(k)$.

Open Problem

Is the existence of succinct representations for primes a ***rare*** *phenomenon?*

Prove that for every large n, there is an n-bit prime of rKt complexity $\leq \varepsilon n$.

By the ***Coding Theorem for rKt****, enough to show that:*

Problem. Is there a probabilistic algorithm that, given n as input, runs in time $2^{\varepsilon n}$ and outputs some fixed n-bit prime with probability at least $1/2^{\varepsilon n}$?

This is a relaxation of the Polymath problem of deterministically generating primes.

Summary: Probabilistic Data Representations

$\mathsf{Kt} \longrightarrow \mathsf{rKt}$

(under assumptions)

$\mathsf{Kt} \approx \mathsf{rKt}$

Succinct Descriptions:

Infinitely many primes have rKt complexity $\leq n^{\varepsilon}$.

Computational Hardness:

It is intractable to estimate the rKt complexity of an input string.

Coding Theorem:

Samplable objects admit short and effective representations.

Main References

[O-Santhanam'17] Pseudodeterministic constructions in subexponential time (STOC'2017).

[O'19] Randomness and intractability in Kolmogorov complexity (ICALP'19).

[Lu-O'20] An efficient coding theorem via probabilistic representations and its applications (preprint).

[Levin'84] Randomness conservation inequalities; information and independence in mathematical theories. Information and Control, 61(1):15-37, 1984.

[ABKvMR'06] Eric Allender, Harry Buhrman, Michal Koucky, Dieter van Melkebeek, and Detlef Ronneburger. Power from random strings. SIAM J. Comput., 35(6):1467-1493, 2006.

[Polymath 4 - TCH12] Terence Tao, Ernest Croot, III, and Harald Helfgott. Deterministic methods to find primes. Math. Comp., 81(278):1233-1246, 2012.

[Troy Lee, PhD Thesis] Kolmogorov complexity and formula lower bounds. University of Amsterdam, 2006.

https://www.dcs.warwick.ac.uk/~igorcarb/complexity-meetings.html