KOLMOGOROV COMPLEXITY AS ENERGY: COMPUTER SCIENCE vs PHYSICS

Yuri I. Manin

PLAN

PART I: ZIPF's LAW

PART II: ERROR–CORRECTING CODES AND THEIR ASYMPTOTIC BOUNDS

PART III: KOLMOGOROV COMPLEXITY AS ENERGY

PART IV: COMPUTABILITY AND MATHEMATICS OF KOLMOGOROV COMPLEXITY

PART V: CAN WE COMPUTE ASYMPTOTIC BOUNDS?

PART I: ZIPF's LAW

• **Consider a corpus of texts in a given language, make the list of words occurring in them and the numbers of occurrences. Range these words in the order of *diminishing* frequencies. Define the Zipf rank of a word as its number in this ordering.**

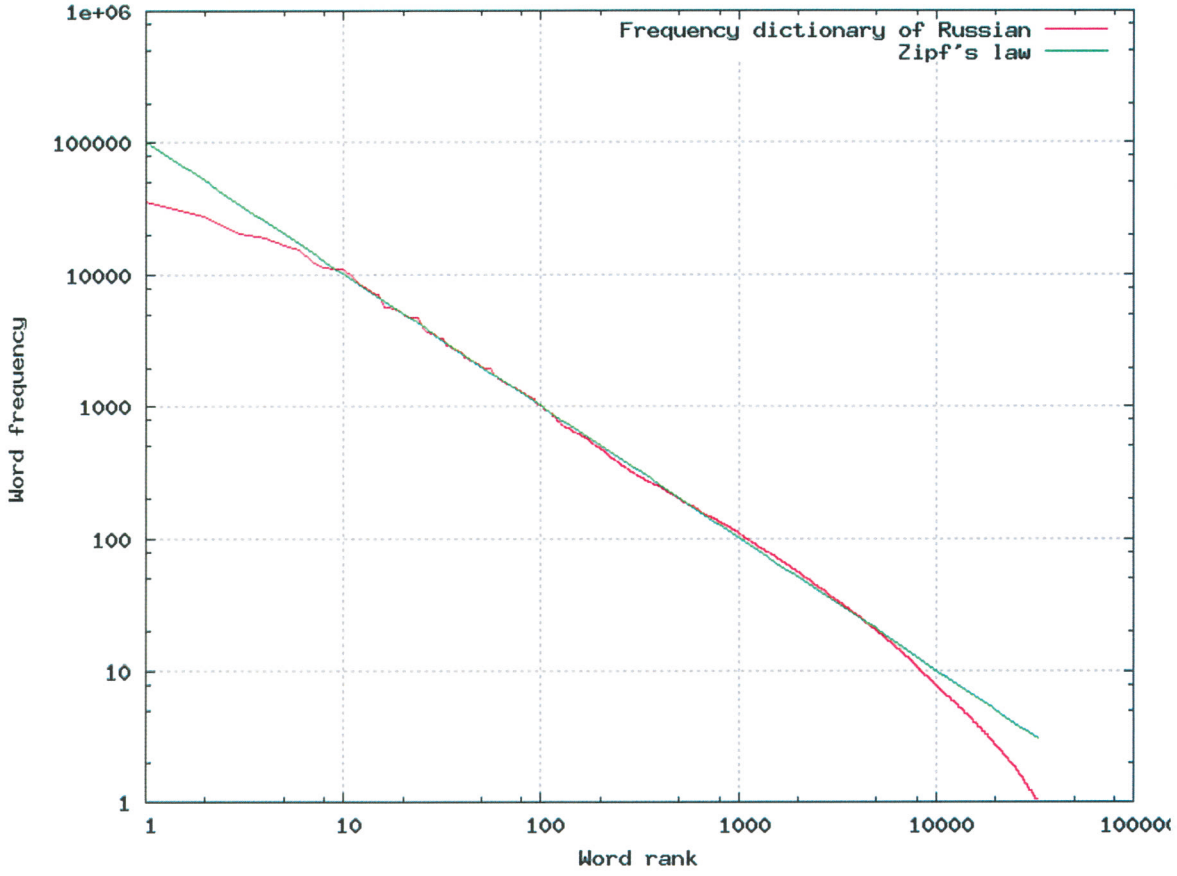
• *Zipf's Law* :

FREQUENCY

IS INVERSELY PROPORTIONAL TO THE RANK

PICTURE:

Zipf's distribution of Russian words(logarithmic scale)



• **Universality of Zipf's law : the law is empirically observed in very different databases, that allow one to calculate frequency of occurrence of certain** *patterns* **("words") in certain** *massifs of data.*

• **Example on the next page: patterns in** *financial audit data.*

Decision Support Systems 46 (2008) 70–83

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

An investigation of Zipf's Law for fraud detection (DSS#06-10-1826R(2))

Shi-Ming Huang [a], David C. Yen [b,*], Luen-Wei Yang [a], Jing-Shiuan Hua [c]

[a] Department of Accounting & Information Technology, National Chung-Cheng University, Chia-Yi, Taiwan, ROC
[b] Department of DSC & MIS, Miami University, Oxford, OH 45056, United States
[c] Department of Information Management, National Chung-Cheng University, Chia-Yi, Taiwan, ROC

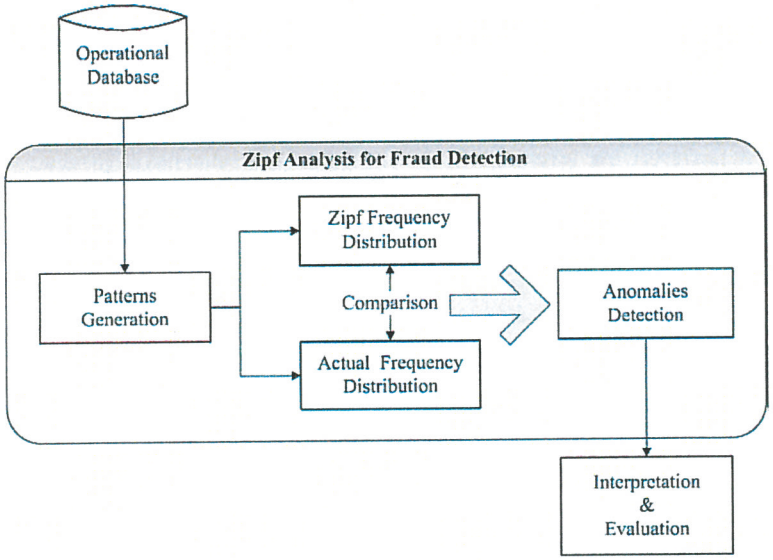


Fig. 1. Fraud detection model of Zipf Analysis.

PART II
ERROR–CORRECTING CODES
AND THEIR ASYMPTOTIC BOUNDS

• BASIC NOTATION:

Alphabet A:= a finite set of cardinality $q \geq 2$.

Code C $\subset A^n$:= a subset of words of length n.

Hamming distance between two words:

$$d((a_i),(b_i)) := \mathrm{card}\{i \in (1,\ldots,n) \mid a_i \neq b_i\}.$$

Code parameters: cardinality of the alphabet q and

$$n(C) := n, \quad k(C) := k := [\log_q \mathrm{card}(C)],$$
$$d(C) := d = \min\{d(a,b) \mid a, b \in C, a \neq b\}.$$

Relative distance and Transmission rate :

$$\delta(C) := \frac{d(C)}{n(C)}, \quad R(C) = \frac{k(C)}{n(C)}.$$

Briefly, C is an $[n,k,d]_q$–code.

SOURCE DATA

Encoding: ↓

A sequence of code words

Noisy channel: ↓

Sequence of (corrupted) code words

Error correction: ↓

(Ideally) sequence of initial code words

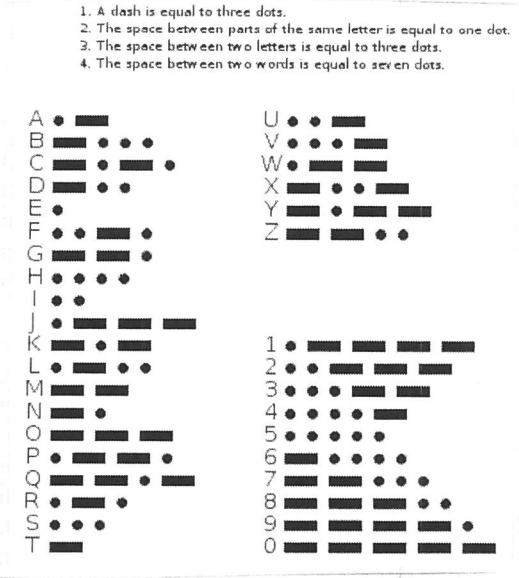
Decoding: ↓

TRANSMITTED DATA

Examples: Morse code and Barcodes

Samuel F. B. Morse
(from 1836 on)

Alphabet: $\{dash,\ dot,\ space\}$,
$q = 3$.
Block length: $n = 7$
$d = ?$: (*Exercise*)



Norman J. Woodland
(from 1949 on):
"His [...]inspiration came from Morse code, and he formed his first barcode from sand on the beach.

I just extended the dots and dashes downwards and made narrow lines and wide lines out of them"

A 13-digit ISBN, 978-3-16-148410-0, as represented by an EAN-13 bar code.

- **Explaining terms:**

Minimal relative distance and Transmission rate :

$$\delta(C) := \frac{d(C)}{n(C)}, \quad R(C) = \frac{[k(C)]}{n(C)}.$$

Minimal relative distance **must match channel's** *noisiness*: **probability of corruption of one letter.**

Transmission rate **is the share of** *meaningful* **(code) words; their number must be maximized for any given relative distance.**

• **A good code must maximize minimal relative distance when the transmission rate is chosen.**

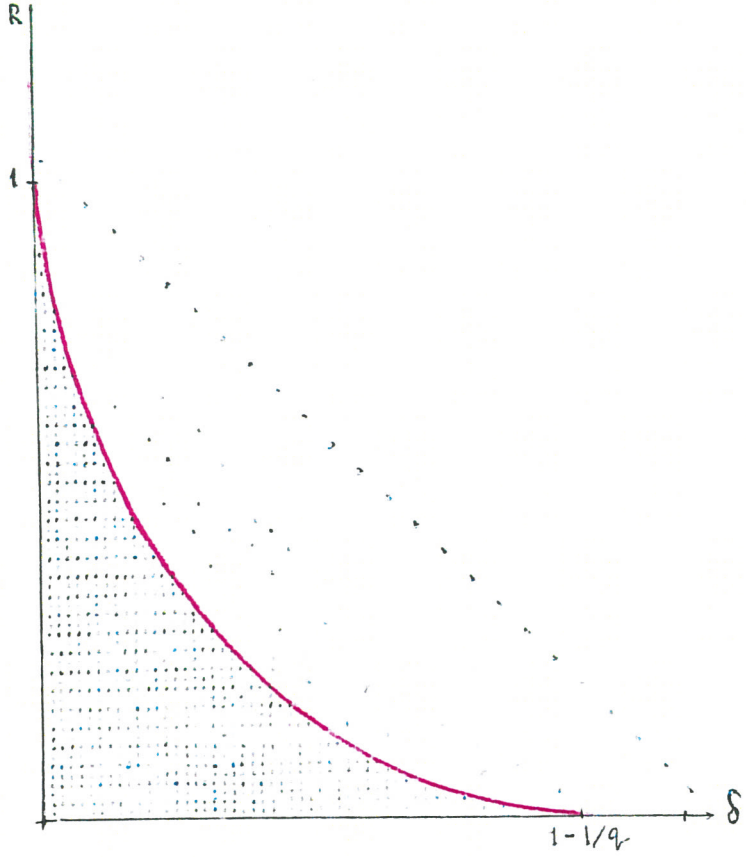
• **One more property of good codes: they must admit** *efficient algorithms of encoding, decoding and error–correction.*

How this can be achieved: consider structured codes. Typical choice:

• **Linear codes := linear subspaces of $\mathbf{F}_q^n$.**

- **Code points:**

$$[n,k,d]_q - code\ C \mapsto P_C := (R(C), \delta(C)) = \left(\frac{[k(C)]}{n(C)}, \frac{d(C)}{n(C)}\right)$$



How a finite pixel plot of all code points might look (q fixed)

Explanations to the picture :

• **DEFINITION.** *Multiplicity* **of a code point is the number of codes that project onto it.**

• **THEOREM (Yu.M.,1981 + 2011). There exists such a continuous function** $\alpha_q(\delta)$, $\delta \in [0,1]$, **that**

(i) The set of code points of infinite multiplicity is exactly the set of rational points $(R,\delta) \in [0,1]^2$ **satisfying** $R \leq \alpha_q(\delta)$.

The curve $R = \alpha_q(\delta)$ **is called the asymptotic bound.**

(ii) Code points x of finite multiplicity all lie above the asymptotic bound and are called isolated ones:

for each such point there is an open neighborhood containing x as the only code point.

(iii) The same statements are true for linear codes, with, a possibly, different asymptotic bound $R = \alpha_q^{lin}(\delta)$.

PART III: KOLMOGOROV COMPLEXITY AS ENERGY

IIIa. AN INTUITIVE DESCRIPTION

• Logarithmic Kolmogorov complexity of a combinatorial object ω is defined as

---

the measure of compressibility of ω *:=*
the length of the shortest program that can generate ω

• A representative example:

Combinatorial object = an integer $N > 0$*, represented by* N *written dashes.*

Then the length of its compressed description is $\leq c \log N$**, and its exponential Kolmogorov complexity is** $\leq C\, N$**.**

IIIb. ZIPF's RANK AND ZIPF's LAW FROM COMPLEXITY

• I suggest that (at least in some situations) Zipf's law emerges as the combined effect of two factors:

(A) Rank ordering coincides with the ordering with respect to the growing (exponential) Kolmogorov complexity $K(w)$ up to a factor $exp\,(O(1))$.

(B) The probability distribution producing Zipf's law is (an approximation to) the L. Levin a priori distribution.

If we accept (A) and (B), then Zipf's law follows from two basic properties of Kolmogorov complexity:

(a) rank of w defined according to (A) is $exp\,(O(1)) \cdot K(w)$.

(b) L. Levin's a priori distribution assigns to an object w probability $\sim KP(w)^{-1}$ where KP is the exponentiated prefix Kolmogorov complexity, and we have, up to $exp\,(O(1))$–factors,

$$K(w) \preceq KP(w) \preceq K(w) \cdot \log^{1+\varepsilon} K(w)$$

with arbitrary $\varepsilon > 0$.

• NB A probability distribution on infinity of objects cannot be constructed directly from K: the series $\sum_m K(m)^{-1}$ diverges. However, on finite sets of data the small discrepancy is additionally masked by the dependence of both K and KP on the choice of an optimal encoding.

• **COMPLEXITY AS EFFORT. The picture described above agrees with Zipf's motto "minimization of effort", but reinterprets the notion of effort: its role is now played by the logarithm of the Kolmogorov complexity that is, by the length of the maximally compressed description of an object.**

Such a picture makes sense especially if the objects satisfying Zipf's distribution, are *generated* **rather than simply** *observed.*

IIIc. ASYMPTOTIC BOUNDS FROM COMPLEXITY

• **Oracle assisted approximate computation of the asymptotic bound.**

– **The set** $Codes_q$ **of all** q**–ary codes in a fixed alphabet** A **is a constructive world.**

– **CLAIM. If an oracle produces for us elements of** $Codes_q$ **in their Kolmogorov order, then we can write an oracle assisted algorithm that for each "pixel size"** N^{-1} **enumerates all code points of the form**

$$(k/N, d/N), \quad a, d \in \mathbf{Z}_+$$

CF. PICTURE ON PAGE 13

• **Partition function for codes involving complexity.**

– **The function $\alpha_q(\delta)$ is continuous and strictly decreasing for $\delta \in [1, 1 - q^{-1})$.**

Hence the limit points domain $R \leq \alpha_q(\delta)$ can be equally well described by the inequality $\delta \leq \beta_q(R)$ where β_q is the function inverse to α_q.

– **Fix an $R \in \mathbf{Q} \cap (0, 1)$. For $\Delta \in \mathbf{Q} \cap (0, 1)$, put**

$$Z(R, \Delta; \beta) := \sum_{C:\, R(C)=R, 1-\Delta \leq \delta(C) \leq 1} K_u(C)^{-\beta+\delta(C)-1},$$

where K_u is an exponential Kolmogorov complexity on $Codes_q$.

• **Theorem.** **(i) If $\Delta > \beta_q(R)$, then $Z(R, \Delta; \beta)$ is a real analytic function of β.**

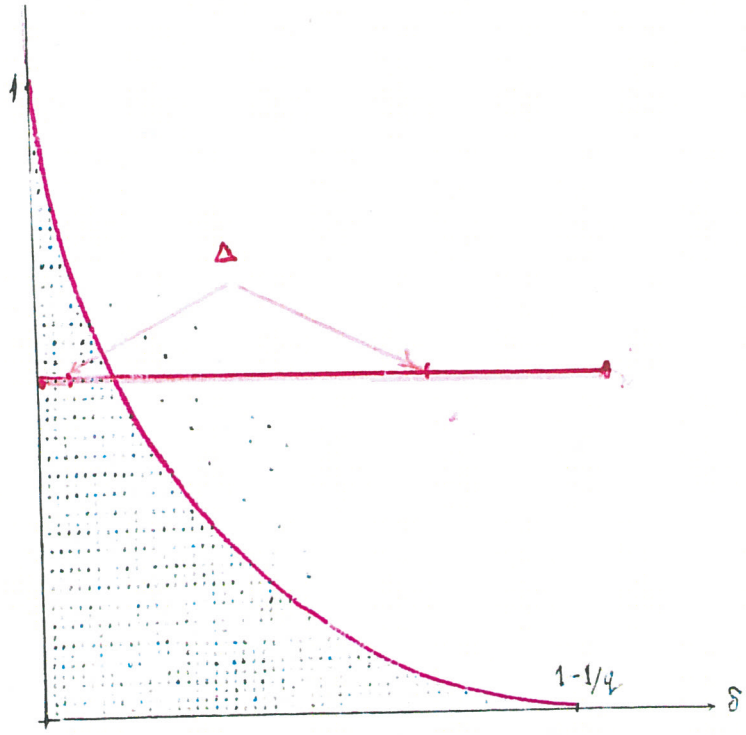
(ii) If $\Delta < \beta_q(R)$, then $Z(R, \Delta; \beta)$ is a real analytic function of β for $\beta > \beta_q(R)$ such that its limit for $\beta - \beta_q(R) \to +0$ does not exist.

• **Thermodynamical analogies.**

– The argument β of the partition function corresponds to the inverse temperature.

– The transmission rate R corresponds to the density ρ.

– Our asymptotic bound transported into $(T = \beta^{-1}, R)$–plane as $T = \beta_q(R)^{-1}$ becomes the phase transition boundary in the (temperature, density)–plane.


1
Δ
1-1/q
δ

PART IV: COMPUTABILITY

AND MATHEMATICS OF KOLMOGOROV COMPLEXITY

IVa. A BRIEF GUIDE TO COMPUTABILITY

- **Three descriptions of partial recursive functions.**

NB A "function", say, $f : X \to Y$, below always means a pair $(f, D(f))$, where $D(f) \subset X$ and $f : D(f) \to Y$. The definition domain is not always mentioned explicitly. If $D(f) = X$, the function will be called "total"; generally it may be called "partial" one.

• **Intuitive description.** **A function $f : \mathbf{Z}_+^m \to \mathbf{Z}_+^n$ is (partial) recursive iff it is "semi–computable" in the following sense:**

there exists an algorithm F accepting as inputs vectors $x = (x_1, \ldots, x_m) \in \mathbf{Z}_+$). with the following properties:

– if $x \in D(f)$, F produces as output $f(x)$.

– if $x \notin D(f)$, F either produces answers "NO", or works indefinitely long without producing any output.

• **Formal description (sketch). It starts with two lists:**

– An explicit list of "obviously" semi–computable elementary functions such as constant functions, projections onto i–th coordinate etc.

– An explicit list of elementary operations, such as an inductive definition, that can be applied to several semi–computable functions and "obviously" produces from them a new semi–computable function.

– After that, the set of partial recursive functions is defined as the minimal set of functions containing all elementary functions and closed wrt all elementary operations.

• Diophantine description **(A DIFFICULT THEOREM). A function** $f : \mathbf{Z}_+^m \to \mathbf{Z}_+^n$ **is partial recursive iff there is a polynomial**

$$P(x_1, \ldots, x_m; y_1, \ldots, y_n; t_1, \ldots, t_q) \in \mathbf{Z}[x, y, t]$$

such that the graph

$$\Gamma_f := \{(x, f(x))\} \subset \mathbf{Z}_+^m \times \mathbf{Z}_+^n$$

is the projection of the subset $P = 0$ **in** $\mathbf{Z}_+^m \times \mathbf{Z}_+^n \times \mathbf{Z}_+^q$**.**

IVb. CONSTRUCTIVE WORLDS

• **An (infinite) constructive world is a countable set X (usually of some finite Bourbaki structures) given together with a class of structural numberings : intuitively computable bijections $\nu : \mathbf{Z}_+ \to X$ which form a principal homogeneous space over the group of totally recursive permutations of $\mathbf{Z}_+$.**

• **CHURCH'S THESIS :** **Let X, Y be two constructive worlds, $\nu_X : \mathbf{Z}_+ \to X$ $\nu_Y : \mathbf{Z}_+ \to X$ their structural numberings, and F an (intuitive) algorithm that takes as input an object $x \in X$ and produces an object $F(x) \in Y$ whenever x lies in the domain of definition of F.**

Then $f := \nu_Y^{-1} \circ F \circ \nu_X : \mathbf{Z}_+ \to \mathbf{Z}_+$ is a partial recursive function.

KOLMOGOROV COMPLEXITY

• **Kolmogorov complexity and Kolmogorov order.**

– Let X be a constructive world. For any (semi)–computable function $u : \mathbf{Z}_+ \to X$, the (exponential) complexity of an object $x \in X$ relative to u is

$$K_u(x) := \min \{m \in \mathbf{Z}_+ \mid u(m) = x\}.$$

If such m does not exist, we put $K_u(x) = \infty$.

– CLAIM: there exists such u ("an optimal Kolmogorov numbering", or "decompressor") that for each other v, some constant $c_{u,v} > 0$, and all $x \in X$,

$$K_u(x) \leq c_{u,v} K_v(x).$$

This $K_u(x)$ is called Kolmogorov complexity of x.

– A Kolmogorov order of a constructive world X is a bijection $\mathbf{K} = \mathbf{K}_u : X \to \mathbf{Z}$ arranging elements of X in the increasing order of their complexities K_u.

• **WARNINGS :**

– Any optimal numbering is only partial function, and its definition domain is not decidable.

– Kolmogorov complexity K_u itself is not computable. It is the lower bound of a sequence of computable functions.

– Kolmogorov order of $\mathbf{Z}_+$ cardinally differs from the natural order in the following sense: it puts in the initial segments very large numbers that can be at the same time Kolmogorov simple.

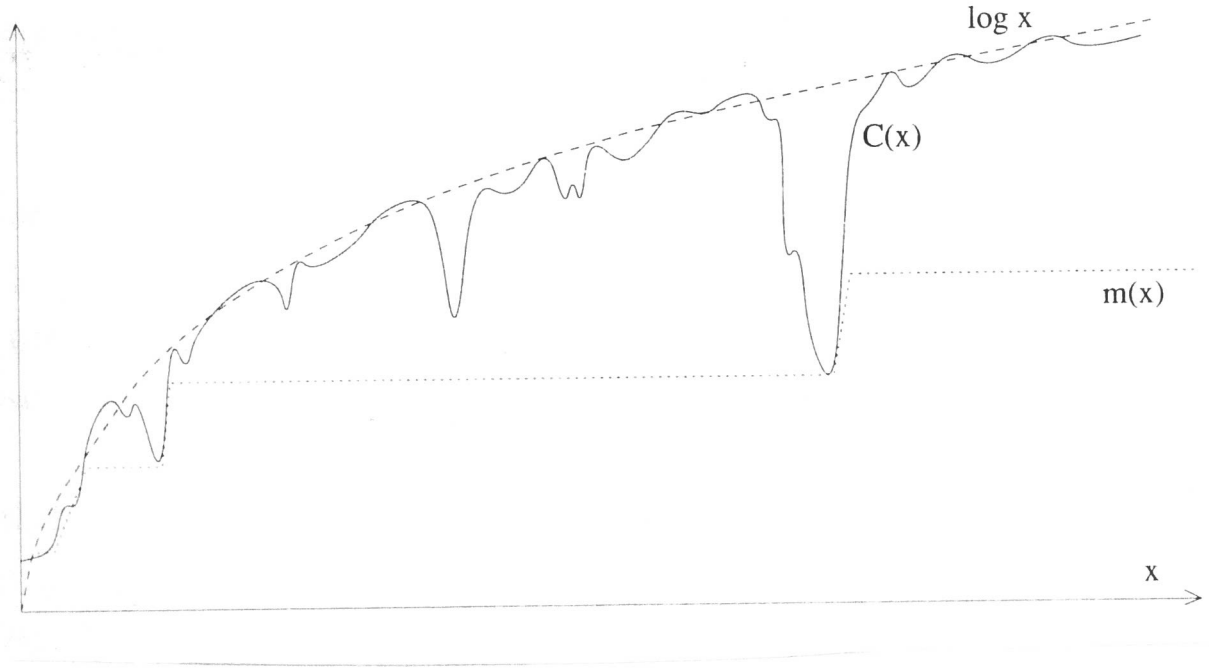
– Example: let $a_n := n^{n^{\cdot^{\cdot^{\cdot^{n}}}}}$ (n times).

Then $K_u(a_n) \leq cn$ for some $c > 0$.

• **FRACTALITY**

– The (graph of) logarithmic complexity has very rich self–similarity properties.

Arguably, they explain the universality of Zipf's Law coming from complexity.



THE GRAPH OF THE KOLMOGOROV COMPLEXITY

PART V:

CAN WE COMPUTE ASYMPTOTIC BOUNDS?

- **What is known :**

– Many upper and lower bounds for asymptotic bounds α_q and $\alpha_q^{lin} \leq \alpha_q$.

– Some isolated codes.

• **What is NOT known :**

– **any explicit formula for α_q or α_q^{lin}.**

– **any single value of $\alpha_q(\delta)$ or $\alpha_q^{lin}(\delta)$ for $0 < \delta < 1 - q^{-1}$.**

– **any method of approximate computation of $\alpha_q(\delta)$ or $\alpha_q^{lin}(\delta)$.**

– **Clearly, $\alpha_q^{lin} \leq \alpha_q$. Is this inequality strict somewhere?**

SOME ESTIMATES

• **Statistical lower estimates : Gilbert–Varshamov bounds.**

– **Most unstructured q–ary codes lie lower or only slightly above the curve**

$$R = \frac{1}{2}(1 - H_q(\delta)),$$

$$H_q(\delta) = \delta \log_q(q-1) - \delta \log_q \delta - (1-\delta)\log_q(1-\delta).$$

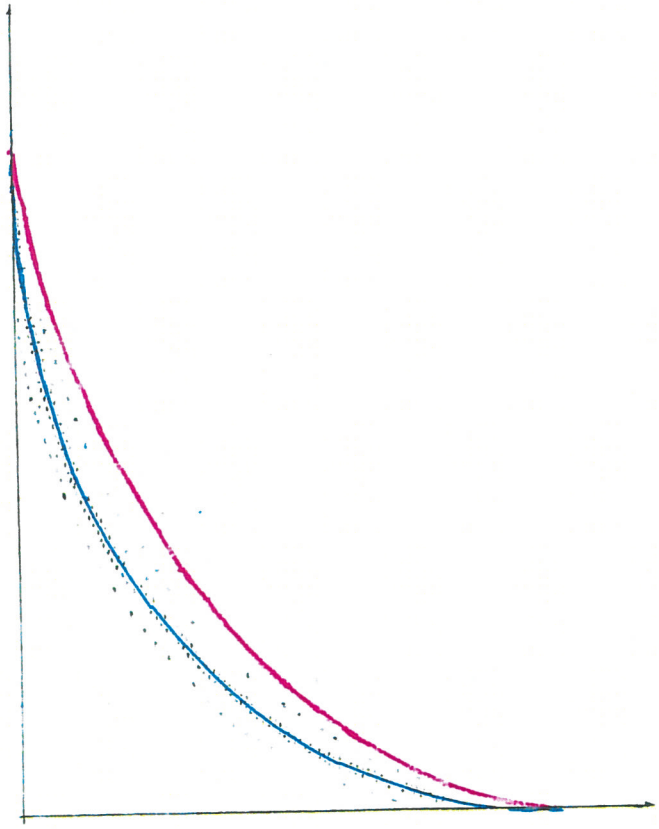
– **Most linear q–ary codes lie lower or only slightly above the Gilbert–Varshamov bound**

$$R = 1 - H_q(\delta).$$

– **In particular,**

$$\alpha_q(R) \geq \alpha_q^{lin}(R) \geq 1 - H_q(\delta)$$

Can we see the asymptotic bound
plotting the set of (linear) code points of bounded size?



NO, we will see a cloud of points
concentrating near the Varshamov–Gilbert bound

• **Combinatorial upper estimate : Singleton bound:**

$$R(C) + \delta(C) \leq 1 + \frac{1}{n(C)}$$

– **Corollary:**

$$\alpha_q(\delta) \leq 1 - \delta.$$

• **Some isolated (linear) codes (Reed – Solomon) :**

– **Parameters:** $1 \leq k \leq n \leq q, d = n + 1 - k$

– **Code** $C \in \mathbf{F}_q^n$**:**

choose pairwise distinct $x_1, \ldots, x_n \in \mathbf{F}_q$**,**

embed the space of polynomials $f(x) \in \mathbf{F}_q[x]$ **of degree** $\leq k-1$ **into** $\mathbf{F}_q^n$ **by**

$$f \mapsto (f(x_1), \ldots, f(x_n)) \in \mathbf{F}_q^n.$$

• **Generalization : algebraic geometric codes.**

– **Input data: a complete smooth algebraic curve X over $\mathbf{F}_q$, a line bundle L on this curve, and a choice of generating vector in each geometric fiber $L(x),\ x \in X(\mathbf{F}_q)$.**

– **The resulting linear code C: image of $\Gamma(X, L)$ wrt the map**

$$\Gamma(X, L) \to \oplus_{x \in X(\mathbf{F}_q)} L(x) \cong \mathbf{F}_q^{\mathrm{card} X(\mathbf{F}_q)}$$

– **Example : Reed–Solomon codes essentially correspond to $X = \mathbf{P}^1$, $L = \mathcal{O}(k-1)$.**

REFERENCES

[Lev] **L. A. Levin,** *Various measures of complexity for finite objects (axiomatic description)*, **Soviet Math. Dokl. Vol.17 (1976) N. 2, 522–526.**

[LiVi] **M. Li, P. M. B. Vitányi,** *An Introduction to Kolmogorov Complexity and its Applications*, **3rd edn. Springer, New York, 2008.**

[Man1] **Yu. I. Manin,** *What is the maximum number of points on a curve over* $\mathbf{F}_2$*?* **J. Fac. Sci. Tokyo, IA, Vol. 28 (1981), 715–720.**

[Man2] **Yu. I. Manin,** *Renormalization and computation I: motivation and background.* **Preprint math.QA/0904.4921**

[Man3] **Yu. I. Manin,** *A course in Mathematical Logic for Mathematicians*, **2nd ed., Springer, New York, 2010.**

[ManMar] **Yu. I. Manin, M. Marcolli.** *Error–correcting codes and phase transitions.* **Mathematics in Computer Science, Vol. 5 (2011) 133-170. arXiv:0910.5135**

[ManVla] **Yu. I. Manin. S.G. Vladut,** *Linear codes and modular curves.* **J. Soviet Math., Vol. 30 (1985), 2611–2643.**

[VlaNoTsfa] **S. G. Vladut, D. Yu. Nogin, M. A. Tsfasman.** *Algebraic geometric codes: basic notions.* **Mathematical Surveys and Monographs, 139. American Mathematical Society, Providence, RI, 2007.**

THANK YOU FOR YOUR ATTENTION!