Spring, 2009

Kolmogorov complexity and its applications

Paul Vitanyi
Computer Science
University of Amsterdam
http://www.cwi.nl/~paulv/course-kc

We live in an information society. Information science is our profession. But do you know what is “information”, mathematically, and how to use it to prove theorems?

You will, by the end of the term.

Examples

- Average case analysis of Shellsort. Open since 1959.
- What is the distance between two pieces of information carrying entities? For example, distance from an internet query to an answer.

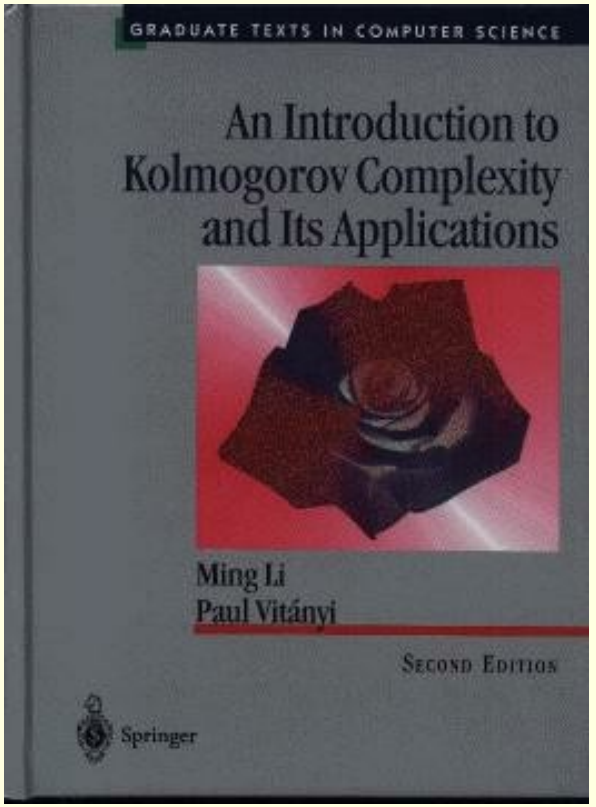
Lecture 1. History and Definition

- History
  - Intuition and ideas in the past
  - Inventors
- Basic mathematical theory

- The course.
  - Textbook: Li-Vitanyi: An introduction to Kolmogorov complexity and its applications,

preferably the *third* edition!

  - Homework every week about last lecture, mid-term and final exam (or possibly individual project and presentation).



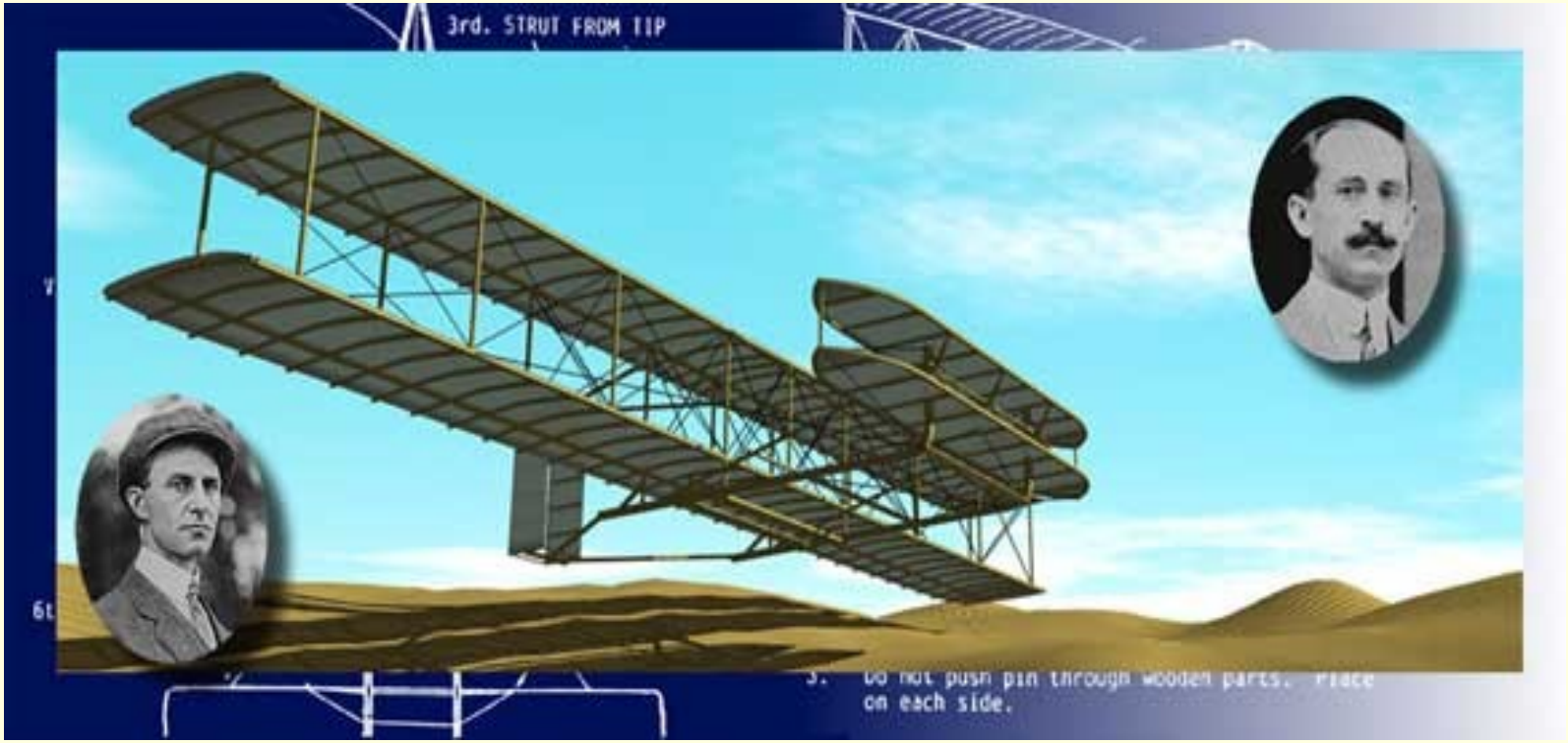
1. Intuition & history

- What is the information content of an individual string?
  - 111 …. 1 (n 1's)
  - π = 3.1415926 …
  - n = 2^{1024}
  - Champernowne's number:

    0.12345678910111213 14 …

    is normal in scale 10 (every block has same frequency)
  - All these numbers share one commonality: there are "small" programs to generate them.
- Shannon's information theory does not help here.

1903: An interesting year



This and the next two pages were stolen from Lance Fortnow

1903: An interesting year

Kolmogorov

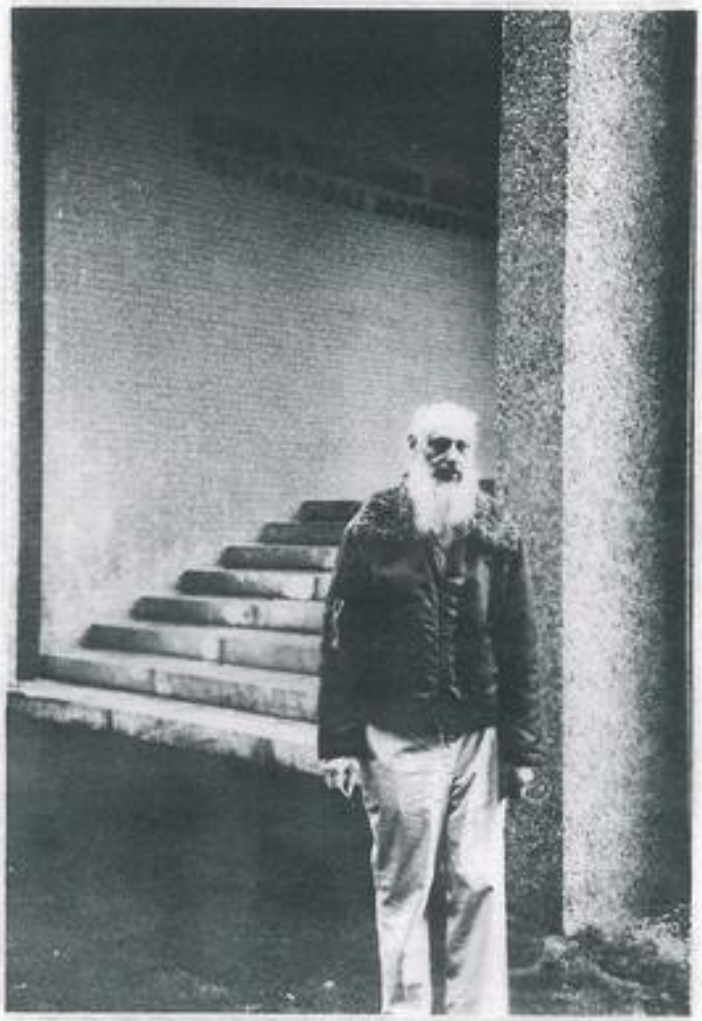
Church

von Neumann

Andrey Nikolaevich Kolmogorov

(1903, Tambov, Russia—1987 Moscow)

- Measure Theory
- Probability
- Analysis
- Intuitionistic Logic
- Cohomology
- Dynamical Systems
- Hydrodynamics
- Kolmogorov complexity



R. J. SOLOMONOFF

1960, 1964

2. INVENTORS

AFOSR TN-60-1459

ZTB-138

A PRELIMINARY REPORT ON A GENERAL THEORY OF INDUCTIVE INFERENCE

$$\frac{\sum_{i_1}^{r''}\sum_{i}^{\infty}\left(\frac{1-\epsilon}{2}\right)^{N(S_{T_a C_{n,k}})_i}}{\sum_{i_1}^{r''}\sum_{i}^{\infty}\left(\frac{1-\epsilon}{2}\right)^{N(S_{T_a C_{n,k}})_i}}$$

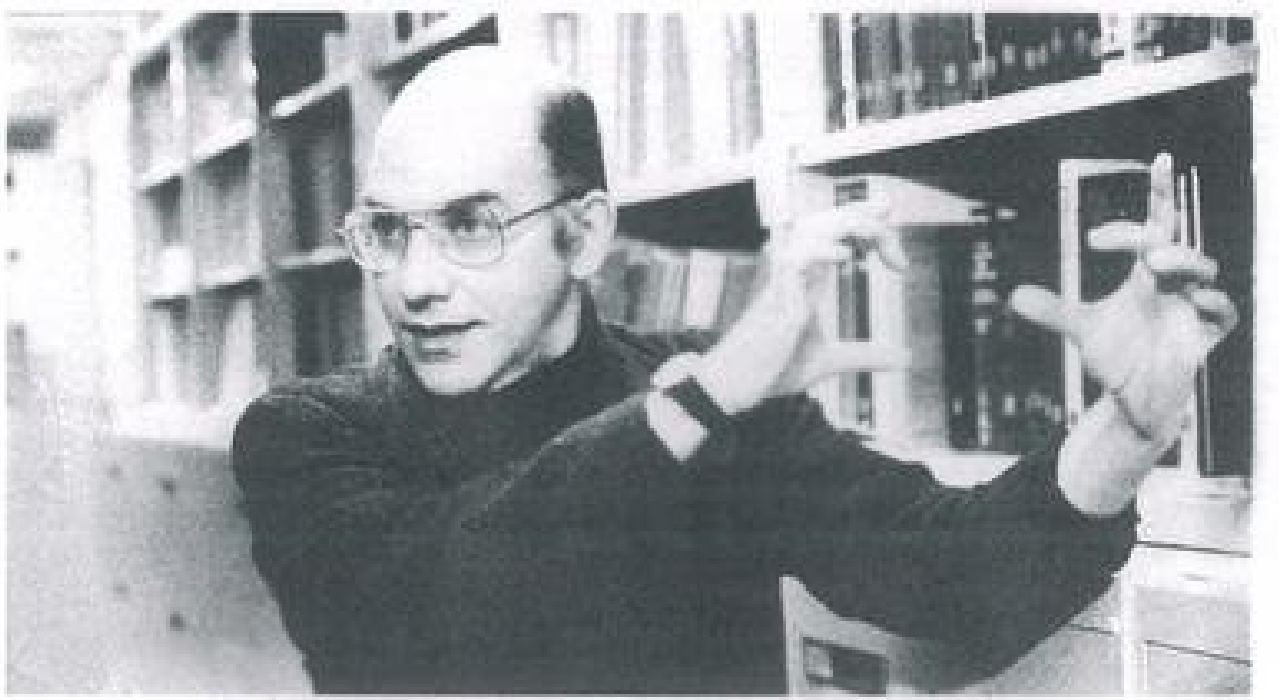
R. J. Solomonoff

November 1960
(Revision of Report V-131, February 1960)

CONTRACT AF 49(638)-376

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH
AIR RESEARCH AND DEVELOPMENT COMMAND
UNITED STATES AIR FORCE
WASHINGTON 25, D. C.

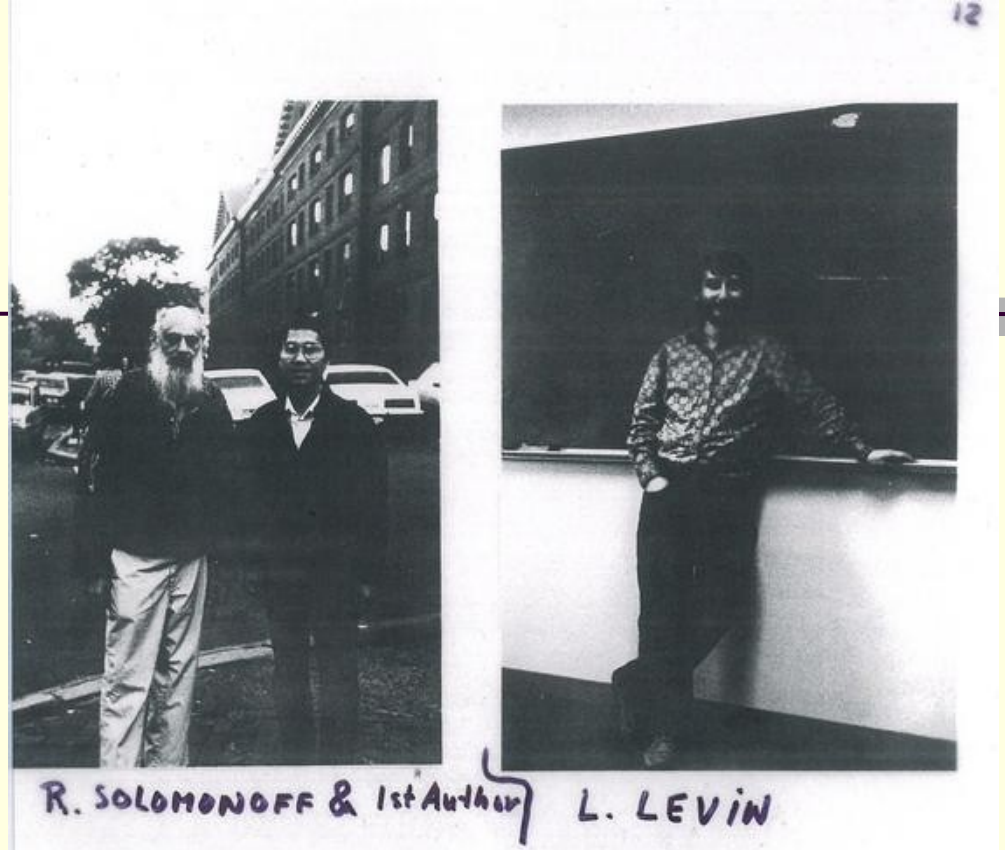
ZATOR COMPANY
140½ MOUNT AUBURN STREET, CAMBRIDGE 38, MASS.



G. J. CHAITIN

1966, 1969

When there were no digital cameras (1987).



A story of Dr. Samuel Johnson

… Dr. Beattie observed, as something remarkable which had happened to him, that he chanced to see both No.1 and No.1000 hackney-coaches. “Why sir,” said Johnson “there is an equal chance for one’s seeing those two numbers as any other two.”

Boswell’s *Life of Johnson*

The case of cheating casino

Bob proposes to flip a coin with Alice:

- Alice wins a dollar if Heads;
- Bob wins a dollar if Tails

Result: TTTTTT …. 100 Tails in a roll.

Alice lost $100. She feels being cheated.

Alice goes to the court

- Alice complains: T^{100} is not random.
- Bob asks Alice to produce a random coin flip sequence.
- Alice flipped her coin and got

  THTTHHTHTHHHTTTTH …
- But Bob claims Alice's sequence has probability 2^{-100}, and so does his.
- How do we define randomness?

2. Roots of Kolmogorov complexity and preliminaries

(1) Foundations of Probability

- P. Laplace: … a sequence is extraordinary (nonrandom) because it contains regularity (which is rare).
- 1919. von Mises' notion of a random sequence S:
  - $\lim_{n\to\infty}\{$ #(1) in n-prefix of S$\}/n = p$, $0<p<1$
  - The above holds for any subsequence of S selected by an "admissible" selection rule.
- If `admissible rule' is any partial function then there are no random sequences.
- A. Wald: countably many admissible selection rules. Then there are "random sequences.
- A. Church: recursive selection functions
- J. Ville: von Mises-Wald-Church random sequence does not satisfy all laws of randomness.

Roots …

(2) Information Theory. Shannon theory is on an ensemble. But what is information in an individual object?

(3) Inductive inference. Bayesian approach using universal prior distribution

(4) Shannon's State x Symbol (Turing machine) complexity.

Preliminaries and Notations

- Strings: x, y, z. Usually binary.
  - $x=x_1x_2$ ... an infinite binary sequence
  - $x_{i:j} = x_i\, x_{i+1}\, \ldots\, x_j$
  - |x| is number of bits in x. Textbook uses l(x).
- Sets, A, B, C …
  - |A|, number of elements in set A. Textbook uses d(A).
- K-complexity vs C-complexity, names etc.
- I assume you know Turing machines, universal TM's, basic facts ...

3. Mathematical Theory

Solomonoff (1960)-Kolmogorov (1965)-Chaitin (1969): The amount of information in a string is the size of the smallest program of an optimal Universal TM U generating that string.

$$C_U(x) = \min_p \{|p|: U(p) = x\}$$

Invariance Theorem: It does not matter which optimal universal Turing machine U we choose. I.e. all "universal encoding methods" are ok.

Proof of the Invariance theorem

- Fix an effective enumeration of all Turing machines (TM's): T_1, T_2, … Define $C_T = \min \{|p|: T(p) = x\}$ $(T \quad p)$
- U is an optimal universal TM such that (p produces x)
  $U(1^n0p) = T_n(p)$
- Then for all x: $C_U(x) \leq C_{Tn}(x) + n+1$, and $|C_U(x) - C_{U'}(x)| \leq c$.
- Fixing U, we write C(x) instead of $C_U(x)$. QED

Formal statement of the Invariance Theorem: There exists a computable function S_0 such that for all computable functions S, there is a constant c_S such that for all strings $x \in \{0,1\}^*$

$C_{S0}(x) \leq C_S(x) + c_S$

It has many applications

- Mathematics --- probability theory, logic, statistics.
- Physics --- chaos, thermodynamics.
- Computer Science – average case analysis, inductive inference and learning, shared information between documents, data mining and clustering, incompressibility method -- examples:
  - Prime number theorem
  - Goedel's incompleteness
  - Shellsort average case
  - Heapsort average case
  - Circuit complexity
  - Lower bounds on combinatorics, graphs,Turing machine computations, formal languages, communication complexity, routing
- Philosophy, biology, cognition, etc – randomness, inference, learning, complex systems, sequence similarity
- Information theory – information in individual objects, information distance
  - Classifying objects: documents, genomes
  - Query Answering systems

Mathematical Theory cont.

- Intuitively*: C(x)*= length of shortest description of *x*
- Define conditional Kolmogorov complexity similarly, with C*(x|y)*=length of shortest description of *x* given *y*.
- Examples
  - C(xx) = C(x) + O(1)
  - C(xy) ≤ C(x) + C(y) + O(log(min{C(x),C(y)})
  - $C(1^n) \leq O(\log n)$
  - $C(\pi_{1:n}) \leq O(\log n)$; $C(\pi_{1:n} | n) \leq O(1)$
  - For all x, C(x) ≤ |x|+O(1)
  - C(x|x) = O(1)
  - C(x|ε) = C(x); C(ε|x)=O(1)

3.1 Basics

- Incompressibility: For constant c>0, a string x ε $\{0,1\}^*$ is c-incompressible if C(x) ≥ |x|-c. For constant c, we often simply say that x is incompressible. (We will call incompressible strings random strings.)

Lemma. There are at least $2^n - 2^{n-c} + 1$ c-incompressible strings of length n.

Proof. There are only $\sum_{k=0,\ldots,n-c-1} 2^k = 2^{n-c} - 1$ programs with length less than n-c. Hence only that many strings (out of total 2^n strings of length n) can have shorter programs (descriptions) than n-c.
QED.

Facts

- If x=uvw is incompressible, then

  $C(v) \geq |v| - O(\log |x|)$. Proof. $C(uvw) = |uvw| \leq |uw| + C(v) + O(\log |u|) + O(\log C(v))$.

- If p is the shortest program for x, then

  $C(p) \geq |p| - O(1)$

- $C(x|p) = O(1)$ but $C(p|x) \leq C(|p|)+O(1)$ (optimal because of the Halting Problem!)
- If a subset A of $\{0,1\}^*$ is recursively enumerable (r.e.) (the elements of A can be listed by a Turing machine), and A is sparse ($|A^{=n}| \leq p(n)$ for some polynomial p), then for all x in A, |x|=n,

  $C(x) \leq O(\log p(n)) + O(C(n)) + O(|A|) = O(\log n)$.

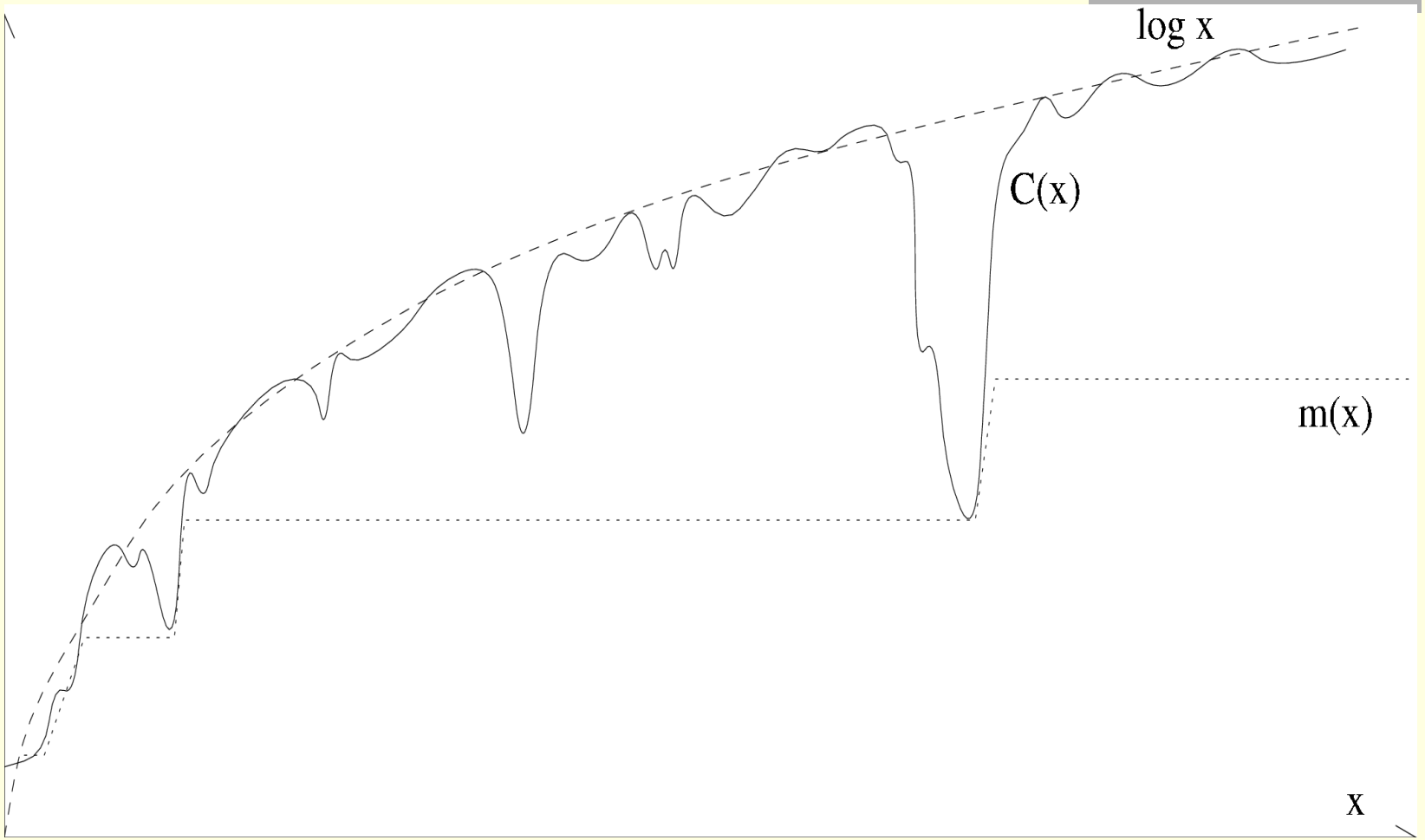
3.2 Asymptotics

- Enumeration of binary strings: 0,1,00,01,10, mapping to natural numbers 0, 1, 2, 3, …
- $C(x) \to \infty$ as $x \to \infty$
- Define m(x) to be the monotonic lower bound of C(x) curve (as natural number $x \to \infty$). Then

  $m(x) \to \infty$, as $x \to \infty$, and

  $m(x) < Q(x)$ for all unbounded computable Q.
- Nonmonotonicity: for x=yz, it does not imply that $C(y) \leq C(x)+O(1)$.

Graph of C(x) for integer x. Function m(x) is greatest monotonic non-decreasing lower bound.



Length-conditional complexity.

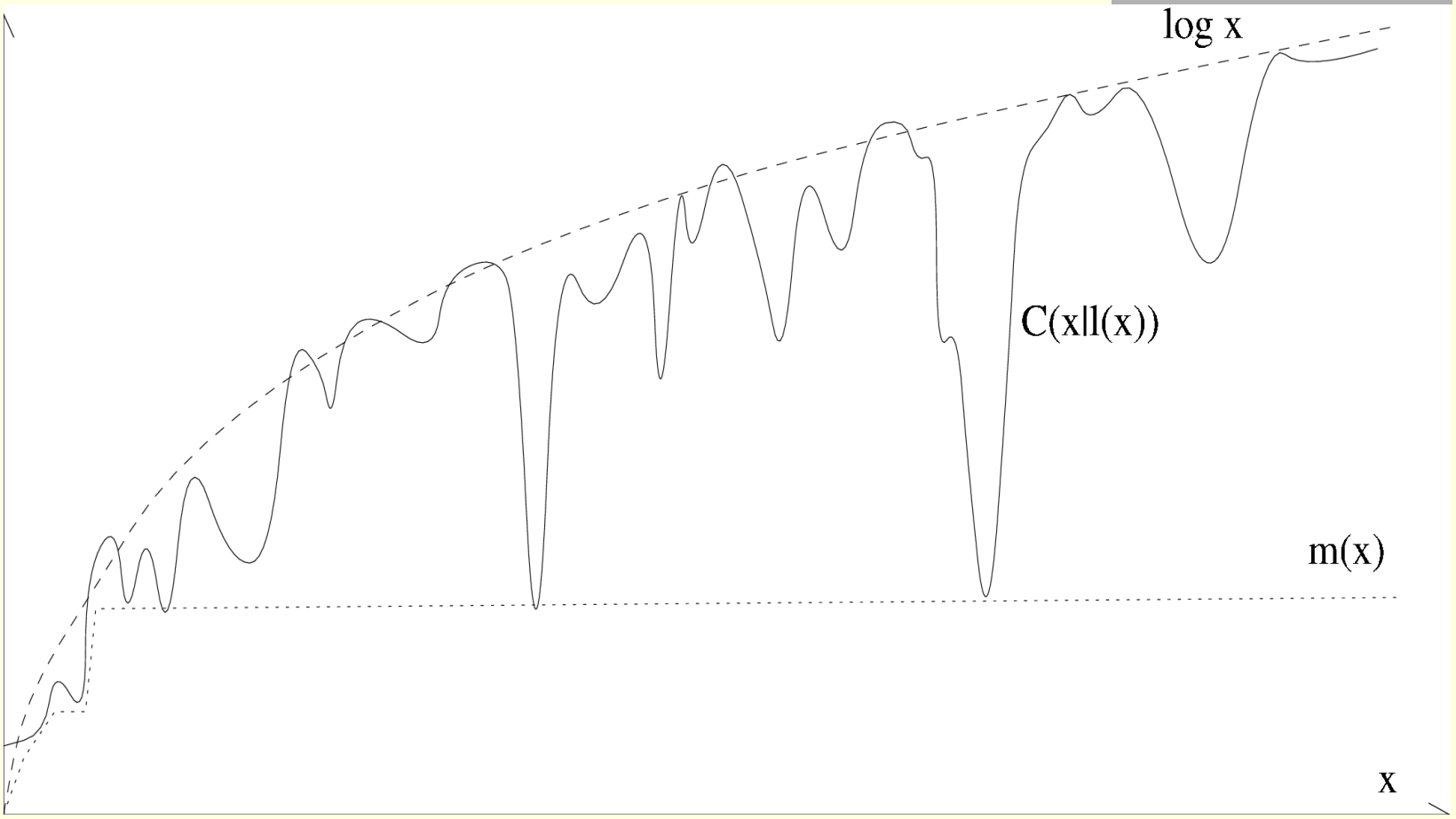
- Let x=x_1...x_n.
- Self-delimiting codes are
- x'=1^n 0x with |x'|=2n+1 , and
- x''= 1^{|n|}0nx with |x''| = n+2|n|+1 (|n|=log n).

- n-strings are x's of the form x=n'0...0 with n=|x|.

Note that |n'|=2 log n +1.

So, for every n, C(x| n)=O(1) for all n-strings.

Graph of C(x|l(x)). Function m(x) is greatest monotonic non-decreasing lower bound.



3.3 Properties

Theorem (Kolmogorov) (i) C(x) is not partially recursive. That is, there is no Turing machine M s.t. M accepts (x,k) if C(x)≥k and undefined otherwise. (ii) However, there is H(t,x) such that H(t+1,x) ≤ H(t,x) and

$$\lim_{t\to\infty} H(t,x) = C(x)$$

where H(t,x) is total recursive.

Proof. (i) If such M exists, then design M' as follows. M' simulates M on input (x,n), for all |x|=n in "parallel" (one step each), and outputs the first x such that M says `yes.' Choose n >> |M'|. Thus we have a contradiction: C(x)≥n by M, but M' outputs x hence

$$|x|=n >> |M'| \geq C(x) \geq n.$$

(ii) TM with program for x running for t steps defines H(t,x). QED

3.4 Godel's Theorem

Theorem. The statement "x is random (=incompressible)" is undecidable for all but finitely many x.

Proof (J. Barzdins, G. Chaitin). Let F be an axiomatic theory (sound, consistent, containing PA). C(F)= C. If the theorem is false and statement "x is random" is provable in F, then we can enumerate all proofs in F to find a proof of "x is random". Consider x's with (1) $|x| >> C+O(\log |x|)$, and output (first) random such x. Then (2) $C(x) < C +O(\log |x|)$ But the proof for "x is random" implies that (3) $C(x) \geq |x|$. Now (1)+(2)+(3) yields a contradiction, $C+O(\log |x|) >> C+O(\log |x|)$. QED

3.5 Barzdin's Lemma

- A characteristic sequence of set A is an infinite binary sequence $\chi=\chi_1\chi_2 \ldots$, $\chi_i=1$ iff $i\varepsilon A$.

Theorem. (i) The characteristic sequence χ of an r.e. set A satisfies $C(\chi_{1:n}|n)\leq \log n+c_A$ for all n. (ii) There is an r.e. set such that $C(\chi_{1:n})\geq \log n$ for all n.

Proof. (i) Use the number m of 1's in the prefix $\chi_{1:n}$ as termination condition [$C(m) \leq \log n+O(1)$].

- (ii) By diagonalization. Let U be the universal TM. Define $\chi=\chi_1\chi_2 \ldots$, by $\chi_i=1$ if the i-th bit output by $U(i)<\infty$ equals 0, otherwise $\chi_i=0$. χ defines an r.e. set. Suppose, for some n, we have $C(\chi_{1:n})<\log n$. Then, there is a program p such that for all $i \leq n$ we have $U(p,i)= \chi_i$ and $|p|< \log n$, hence $p<n$. But $U(p,p)$ not equal χ_p by definition. QED