Kolmogorov Complexity

CS 3120 - April 8, 2025

Recall

A *Turing Machine,* is defined by (Σ, k, δ):

$\boldsymbol{k} \in \mathbb{N}$: a finite number of states

Σ : alphabet $-$ finite set of symbols

$$\Sigma \supseteq \{0, 1, \rhd, \emptyset\}$$

$\boldsymbol{\delta}$: transition function

$$\boldsymbol{\delta}: [k] \times \Sigma \rightarrow [k] \times \Sigma \times \{\mathbf{L}, \mathbf{R}, \mathbf{S}, \mathbf{H}\}$$

Recall

A *Turing Machine,* is defined by (Σ, k, δ):

$\boldsymbol{k} \in \mathbb{N}$: a finite number of states

Σ : alphabet $-$ finite set of symbols

$$\Sigma \supseteq \{0, 1, \rhd, \emptyset\}$$

$\boldsymbol{\delta}$: transition function

$$\boldsymbol{\delta}: [k] \times \Sigma \rightarrow [k] \times \Sigma \times \{\mathbf{L}, \mathbf{R}, \mathbf{S}, \mathbf{H}\}$$

$$|\, TMs \,| = |\, \text{finite binary strings} \,| = |\, \mathbb{N} \,|$$

What is information?

What is information?

00000000000000000000

10101111010101111100

What is a computable random number?

What is a computable random number?

0.1010010001000010000010000001

0.15264839763039644951841857640

What is an incompressible string?

What is an incompressible string?

1010101010101010

0110011101010001

How much information is in a string?

How much information is in a string?

100100100100100100100100100

100100100100100100100100100

100100100100100100100100100

How much information is in a string?

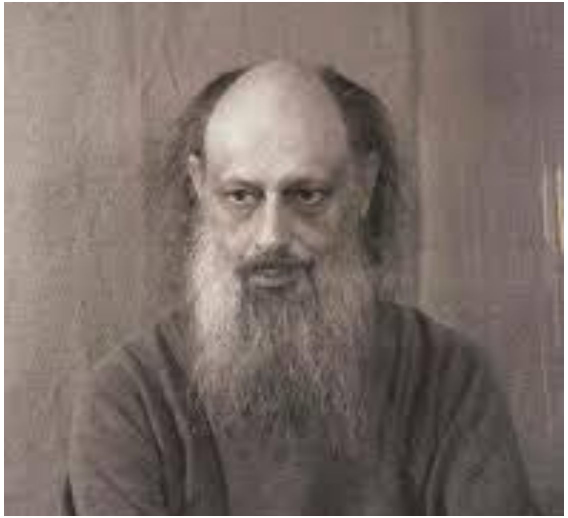
23571113171923293137414347 53
596167717379838997

How much information is in a string?

383862748878325473579680183
468291898745981708710670140
958198041893735035359221176

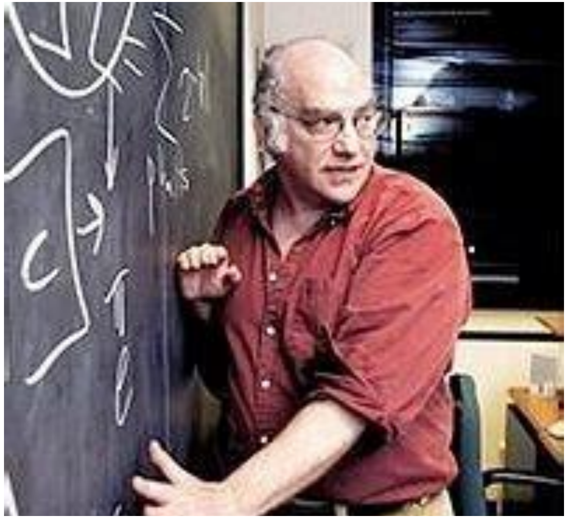
Kolmogorov Complexity

Origins of Kolmogorov Complexity



Ray Solomonoff (1964)

Andrey Kolmogorov (1965)



Gregory Chaitin (1969)

Kolmogorov Complexity

The information in a string x is the size of the smallest description <M> of any TM M generating x.

Kolmogorov Complexity

The information in a string x is the size of the smallest description <M> of any TM M generating x.

Let U be a universal TM that takes as input the description p=<M> of a TM M and produces as output U(p).

Kolmogorov Complexity

The information in a string x is the size of the smallest description <M> of any TM M generating x.

Let U be a universal TM that takes as input the description p=<M> of a TM M and produces as output U(p).

Kolmogorov Complexity of x, denoted $K_U(x)$

$$K_U(x) = \min_n \{ | < M_n > | : U \text{ simulates } M_n \text{ and outputs } x \}$$

Kolmogorov Complexity

The information in a string x is the size of the smallest description <M> of any TM M generating x.

Let U be a universal TM that takes as input the description p=<M> of a TM M and produces as output U(p).

Kolmogorov Complexity of x, denoted $K_U(x)$

$$K_U(x) = \min_n \{ |< M_n >| : U \text{ simulates } M_n \text{ and outputs } x \}$$

K(x) is the length of the shortest description of x

Kolmogorov Theory Applications

Mathematics: probability theory, logic

Physics: chaos, thermodynamics

Computer Science: average case analysis, inductive inference and learning, shared information between documents, data mining and clustering, incompressibility method

Misc: randomness, inference, complex systems, sequence similarity

Information Theory: information in individual objects, information distance

Invariance Theorem

Kolmogorov Complexity is robust.

The choice of universal Turing machine only affects complexity by an additive constant.

→ All encoding methods are equivalent up to a constant.

Invariance Theorem

There exists a Turing machine U such that for all Turing machines M_n, there exists a constant c_n such that for all strings $x \in \{0,1\}^*$,

$$K_U(x) \leq K_{M_n}(x) + c_n$$

Proof

Proof

Fix an enumeration of all Turing machines (TMs): $M_1, M_2, \ldots$.

Proof

Fix an enumeration of all Turing machines (TMs): $M_1, M_2, \ldots$.

Let U be a universal Turing machine such that $U(0^n 1a) = M_n(a)$

Proof

Fix an enumeration of all Turing machines (TMs): $M_1, M_2, \ldots$.

Let U be a universal Turing machine such that $U(0^n 1a) = M_n(a)$

Let $x \in \{0,1\}^*$, and let p be the shortest program such that $M_n(p) = x$. Then $K_{M_n}(x) = |p|$.

Proof

Fix an enumeration of all Turing machines (TMs): $M_1, M_2, \ldots$.

Let U be a universal Turing machine such that $U(0^n 1a) = M_n(a)$

Let $x \in \{0,1\}^*$, and let p be the shortest program such that $M_n(p) = x$. Then $K_{M_n}(x) = |p|$.

Construct $p' = 0^n 1p$. Then: $U(p') = M_n(p) = x$

Proof

Fix an enumeration of all Turing machines (TMs): $M_1, M_2, \ldots$.

Let U be a universal Turing machine such that $U(0^n 1a) = M_n(a)$

Let $x \in \{0,1\}^*$, and let p be the shortest program such that $M_n(p) = x$. Then $K_{M_n}(x) = |p|$.

Construct $p' = 0^n 1p$. Then: $U(p') = M_n(p) = x$

Thus, $K_U(x) \leq |p'| = |0^n 1| + |p| = (n+1) + K_{M_n}(x)$

Proof

Fix an enumeration of all Turing machines (TMs): $M_1, M_2, \ldots$.

Let U be a universal Turing machine such that $U(0^n 1a) = M_n(a)$

Let $x \in \{0,1\}^*$, and let p be the shortest program such that $M_n(p) = x$. Then $K_{M_n}(x) = |p|$.

Construct $p' = 0^n 1p$. Then: $U(p') = M_n(p) = x$

Thus, $K_U(x) \leq |p'| = |0^n 1| + |p| = (n+1) + K_{M_n}(x)$

So, $K_U(x) \leq K_{M_n}(x) + c_n \quad$ where $c_n = n + 1$

Proof

Fix an enumeration of all Turing machines (TMs): $M_1, M_2, \ldots$.

Let U be a universal Turing machine such that $U(0^n 1a) = M_n(a)$

Let $x \in \{0,1\}^*$, and let p be the shortest program such that $M_n(p) = x$. Then $K_{M_n}(x) = |p|$.

Construct $p' = 0^n 1p$. Then: $U(p') = M_n(p) = x$

Thus, $K_U(x) \le |p'| = |0^n 1| + |p| = (n+1) + K_{M_n}(x)$

So, $K_U(x) \le K_{M_n}(x) + c_n \quad \text{where } c_n = n + 1$

The constant c_n depends only on n, not on x. $\square$

Kolmogorov Complexity Properties

1. $K(xx) = K(x) + O(1)$

Kolmogorov Complexity Properties

1. $K(xx) = K(x) + O(1)$
2. $K(1^n) \leq O(\log n)$

Kolmogorov Complexity Properties

1. $K(xx) = K(x) + O(1)$
2. $K(1^n) \leq O(\log n)$
3. $K(n!) \leq O(\log n)$

Kolmogorov Complexity Properties

1. $K(xx) = K(x) + O(1)$
2. $K(1^n) \leq O(\log n)$
3. $K(n!) \leq O(\log n)$
4. For all x, $K(x) \leq |x| + O(1)$

Kolmogorov Complexity Properties

1. $K(xx) = K(x) + O(1)$
2. $K(1^n) \leq O(\log n)$
3. $K(n!) \leq O(\log n)$
4. For all x, $K(x) \leq |x| + O(1)$
5. $K(xy) \leq K(x) + K(y) + O(\log(\min\{K(x), K(y)\})$

Incompressibility

Incompressibility: For constant $c > 0$, a string $x \in \{0,1\}^*$ is c-incompressible if $K(x) \geq |x| - c$.

Incompressibility

Incompressibility: For constant $c > 0$, a string $x \in \{0,1\}^*$ is c-incompressible if $K(x) \geq |x| - c$.

Incompressible strings have properties similar to random strings.

Incompressibility

Incompressibility: For constant $c > 0$, a string $x \in \{0,1\}^*$ is c-incompressible if $K(x) \geq |x|-c$.

Incompressible strings have properties similar to random strings.

There are infinitely many incompressible strings.

There are infinitely many incompressible binary strings. That is, for any constant c, there exist infinitely many strings $x \in \{0,1\}^*$ such that $K(x) \geq |x| - c$ where $K(x)$ denotes the Kolmogorov complexity of x.

Fix constant $c > 0$ and $n \in \mathbb{N}$ and consider the set of all length n binary strings. There are exactly 2^n such strings.

Fix constant $c > 0$ and $n \in \mathbb{N}$ and consider the set of all length n binary strings. There are exactly 2^n such strings.

Consider $S = \{x \in \{0,1\}^* : K(x) < n - c\}$.

Fix constant $c > 0$ and $n \in \mathbb{N}$ and consider the set of all length n binary strings. There are exactly 2^n such strings.

Consider $S = \{x \in \{0,1\}^* : K(x) < n - c\}$.

For all $x \in S$ there exists a binary program p (under some fixed universal Turing Machine) such that $|p| < n - c$.

Fix constant $c > 0$ and $n \in \mathbb{N}$ and consider the set of all length n binary strings. There are exactly 2^n such strings.

Consider $S = \{x \in \{0,1\}^* : K(x) < n - c\}$.

For all $x \in S$ there exists a binary program p (under some fixed universal Turing Machine) such that $|p| < n - c$.

The number of such binary programs of length less than $n - c$ is at most:
$\sum_{i=0}^{n-c-1} 2^i = 2^{n-c} - 1$.

Fix constant $c > 0$ and $n \in \mathbb{N}$ and consider the set of all length n binary strings. There are exactly 2^n such strings.

Consider $S = \{x \in \{0,1\}^* : K(x) < n - c\}$.

For all $x \in S$ there exists a binary program p (under some fixed universal Turing Machine) such that $|p| < n - c$.

The number of such binary programs of length less than $n - c$ is at most: $\sum_{i=0}^{n-c-1} 2^i = 2^{n-c} - 1$.

Therefore, $|S| \leq 2^{n-c} - 1$.

Fix constant $c > 0$ and $n \in \mathbb{N}$ and consider the set of all length n binary strings. There are exactly 2^n such strings.

Consider $S = \{x \in \{0,1\}^* : K(x) < n - c\}$.

For all $x \in S$ there exists a binary program p (under some fixed universal Turing Machine) such that $|p| < n - c$.

The number of such binary programs of length less than $n - c$ is at most: $\sum_{i=0}^{n-c-1} 2^i = 2^{n-c} - 1$.

Therefore, $|S| \leq 2^{n-c} - 1$.

So the number of strings x of length n such that $K(x) \geq n - c$ is at least:

$$2^n - |S| \geq 2^n - 2^{n-c} + 1 = 2^n \left(1 - \frac{1}{2^c}\right) + 1.$$

Fix constant $c > 0$ and $n \in \mathbb{N}$ and consider the set of all length n binary strings. There are exactly 2^n such strings.

Consider $S = \{x \in \{0,1\}^* : K(x) < n - c\}$.

For all $x \in S$ there exists a binary program p (under some fixed universal Turing Machine) such that $|p| < n - c$.

The number of such binary programs of length less than $n - c$ is at most: $\sum_{i=0}^{n-c-1} 2^i = 2^{n-c} - 1$.

Therefore, $|S| \leq 2^{n-c} - 1$.

So the number of strings x of length n such that $K(x) \geq n - c$ is at least:

$$2^n - |S| \geq 2^n - 2^{n-c} + 1 = 2^n \left(1 - \frac{1}{2^c}\right) + 1.$$

This quantity is positive for all n and fixed c, implying that for each n, there is at least one string of length n such that $K(x) \geq n - c$.

Fix constant $c > 0$ and $n \in \mathbb{N}$ and consider the set of all length n binary strings. There are exactly 2^n such strings.

Consider $S = \{x \in \{0,1\}^* : K(x) < n - c\}$.

For all $x \in S$ there exists a binary program p (under some fixed universal Turing Machine) such that $|p| < n - c$.

The number of such binary programs of length less than $n - c$ is at most: $\sum_{i=0}^{n-c-1} 2^i = 2^{n-c} - 1$.

Therefore, $|S| \leq 2^{n-c} - 1$.

So the number of strings x of length n such that $K(x) \geq n - c$ is at least:

$$2^n - |S| \geq 2^n - 2^{n-c} + 1 = 2^n \left(1 - \frac{1}{2^c}\right) + 1.$$

This quantity is positive for all n and fixed c, implying that for each n, there is at least one string of length n such that $K(x) \geq n - c$.

Thus, for any constant c, there are infinitely many incompressible strings. $\square$