Precise expressions for the algorithmic information distance

Bruno Bauwens.

National Research University Higher School of Economics (HSE), Moscow.

10/07/2022, PLS13, Volos

Important contributions were made by Alexander (Sasha) Shen.

I am also grateful to
– Mikhail Andreev,
– Artem Grachev,
– the participants of the Kolmogorov seminar in Moscow,
for useful discussions.

Informal definitions

- The Kolmogorov complexity $C(x)$ of a bitstring x is

  "The minimal length of a program that outputs x and halts."

Informal definitions

- The Kolmogorov complexity $C(x)$ of a bitstring x is

  "The minimal length of a program that outputs x and halts."

- The conditional complexity $C(x|y)$ is

  "The minimal length of a program that maps y to x."

Informal definitions

- The Kolmogorov complexity $C(x)$ of a bitstring x is

  "The minimal length of a program that outputs x and halts."

- The conditional complexity $C(x|y)$ is

  "The minimal length of a program that maps y to x."

- The algorithmic information distance $E(x, y)$ is

  "The minimal length of a program that computes x from y as well as computing y from x."

Bennett, Gács, Li, Vitányi, and Zurek, in 1998.

Informal definitions

- The Kolmogorov complexity $C(x)$ of a bitstring x is

  "The minimal length of a program that outputs x and halts."

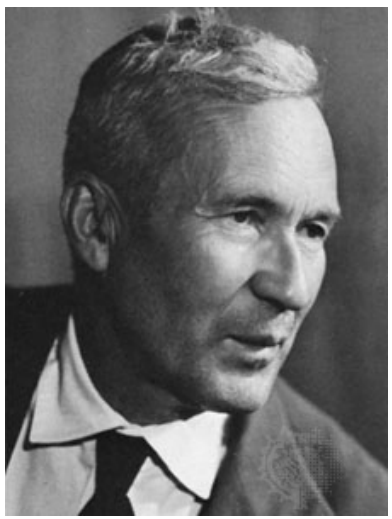
- The conditional complexity $C(x|y)$ is

  "The minimal length of a program that maps y to x."

- The algorithmic information distance $E(x, y)$ is

  "The minimal length of a program that computes x from y as well as computing y from x."

Bennett, Gács, Li, Vitányi, and Zurek, in 1998.



A. Kolmogorov

P. Gács

P. Vitányi

The algorithmic information distance $E(x, y)$ is

"The minimal length of a program that computes x from y as well as computing y from x."

Properties.

- The distance is close to satisfying the properties of a metric.

The algorithmic information distance $E(x, y)$ is

"The minimal length of a program that computes x from y as well as computing y from x."

Properties.

- The distance is close to satisfying the properties of a metric.
- Takes all possible type of regularities into account.

  Up to $O(1)$ additive terms, it is the minimal symmetric function D
  (i) that is approximable from above, and
  (ii) has balls $\{y : D(x, y) \leq k\}$ of size at most 2^k.

The algorithmic information distance $E(x, y)$ is

"The minimal length of a program that computes x from y as well as computing y from x."

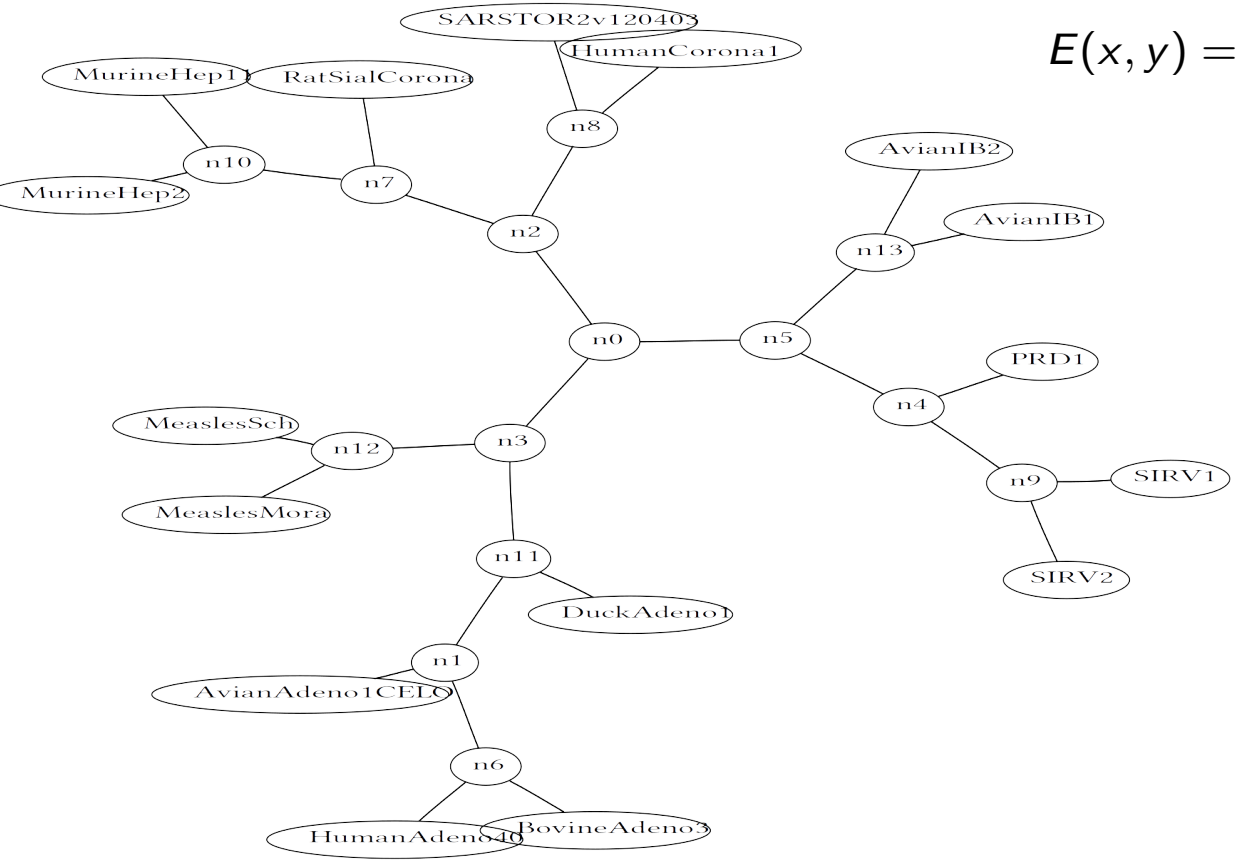
Properties.

- The distance is close to satisfying the properties of a metric.
- Takes all possible type of regularities into account.

  Up to $O(1)$ additive terms, it is the minimal symmetric function D
  (i) that is approximable from above, and
  (ii) has balls $\{y : D(x, y) \leq k\}$ of size at most 2^k.
- Characterization: $E(x, y) = \max\{\mathrm{C}(x|y), \mathrm{C}(y|x)\} + O(\log n)$.
  Applications: approximate complexity with real compressors.

Applications [Cilibrasi and Vitanyi, 2005]

Below, some SARS-related viruses (including some Coronaviruses from 2003).



$$E(x, y) = \max\{\mathrm{C}(x|y), \mathrm{C}(y|x)\} + O(\log n).$$

When the approximated distance is applied to tree-cluster algorithm, it can recover non-obvious information:

- Group music files by genre.
- Group Russian books by authors.
- Recover phylogenetic trees in biology.
- ...

Disclaimer: a normalized variant of the distance was approximated.

15 virus genomes were downloaded from The Universal Virus Database of the International Committee on Taxonomy of Viruses and Canadas Michael Smith Genome Sciences Centre when they were sequenced. The algorithm recovered the evolutionary tree as was published in the in the New England Journal of Medicine in April 2003. (Submitted to ArXiv in Dec 2003.)

Outline

2 variants of the distance:
with plain and prefix complexity

	plain	prefix
triangle inequality	$\times$	?
$=$ max complexity	$\checkmark$	?

The plain information distance

$$E_U(x,y) = \min\{|p|\colon U(p,x) = y \text{ and } U(p,y) = x\},$$

We use machines U that minimize this distance up to $O(1)$ and drop index U.

Properties.

The plain information distance

$$E_U(x, y) = \min\{|p| : U(p, x) = y \text{ and } U(p, y) = x\},$$

We use machines U that minimize this distance up to $O(1)$ and drop index U.

Properties.

(1) Symmetric and non-negative.
Almost satisfies the axioms of a metric:

– $E(x, x) \leq O(1)$

– $E(x, z) \leq E(x, y) + E(y, z) + O(\log E(x, y))$,

The plain information distance

$$E_U(x,y) = \min\{|p|\colon U(p,x) = y \text{ and } U(p,y) = x\},$$

We use machines U that minimize this distance up to $O(1)$ and drop index U.

Properties.

(1) Symmetric and non-negative.
Almost satisfies the axioms of a metric:

– $E(x,x) \leq O(1)$

– $E(x,z) \leq E(x,y) + E(y,z) + O(\log E(x,y))$,

Indeed, similar to

$$\mathrm{C}(x|z) \leq \mathrm{C}(x|y) + \mathrm{C}(y|z) + O(\log \mathrm{C}(x|y))$$

The plain information distance

$$E_U(x,y) = \min\{|p|\colon U(p,x) = y \text{ and } U(p,y) = x\},$$

We use machines U that minimize this distance up to $O(1)$ and drop index U.

Properties.

(1) Symmetric and non-negative.
Almost satisfies the axioms of a metric:

– $E(x,x) \leq O(1)$

– $E(x,z) \leq E(x,y) + E(y,z) + O(\log E(x,y))$,

Indeed, similar to

$$\mathrm{C}(x|z) \leq \mathrm{C}(x|y) + \mathrm{C}(y|z) + O(\log \mathrm{C}(x|y))$$

Note: the $+O(\log E(x,y))$ can not be omitted...

The plain information distance

$$E_U(x,y) = \min\{|p|\colon U(p,x) = y \text{ and } U(p,y) = x\},$$

We use machines U that minimize this distance up to $O(1)$ and drop index U.

Properties.

(1) Symmetric and non-negative.
Almost satisfies the axioms of a metric:

– $E(x,x) \leq O(1)$

– $E(x,z) \leq E(x,y) + E(y,z) + O(\log E(x,y))$,

Indeed, similar to

$$\mathrm{C}(x|z) \leq \mathrm{C}(x|y) + \mathrm{C}(y|z) + O(\log \mathrm{C}(x|y))$$

Note: the $+O(\log E(x,y))$ can not be omitted...

(2) Optimality. Up to $O(1)$ additive terms, it is the minimal symmetric function D

(i) that is approximable from above,

(ii) has balls $\{y : D(x,y) \leq k\}$ of size at most 2^k.

The plain information distance

$$E_U(x,y) = \min\{|p|\colon U(p,x) = y \text{ and } U(p,y) = x\},$$

$$E(x,y) = \max\{\mathrm{C}(x|y), \mathrm{C}(y|x)\} + O(1).$$

We use machines U that minimize this distance up to $O(1)$ and drop index U.

Properties.

(1) Symmetric and non-negative.
Almost satisfies the axioms of a metric:

– $E(x,x) \leq O(1)$

– $E(x,z) \leq E(x,y) + E(y,z) + O(\log E(x,y))$,

Indeed, similar to

$$\mathrm{C}(x|z) \leq \mathrm{C}(x|y) + \mathrm{C}(y|z) + O(\log \mathrm{C}(x|y))$$

Note: the $+O(\log E(x,y))$ can not be omitted...

(2) Optimality. Up to $O(1)$ additive terms, it is the minimal symmetric function D

(i) that is approximable from above,

(ii) has balls $\{y : D(x,y) \leq k\}$ of size at most 2^k.

The plain information distance

$$E_U(x,y) = \min\{|p|: U(p,x) = y \text{ and } U(p,y) = x\},$$

$$E(x,y) = \max\{\mathrm{C}(x|y), \mathrm{C}(y|x)\} + O(1).$$

We use machines U that minimize this distance up to $O(1)$ and drop index U.

Properties.

(1) Symmetric and non-negative.
Almost satisfies the axioms of a metric:

– $E(x,x) \leq O(1)$

– $E(x,z) \leq E(x,y) + E(y,z) + O(\log E(x,y))$,

Indeed, similar to

$$\mathrm{C}(x|z) \leq \mathrm{C}(x|y) + \mathrm{C}(y|z) + O(\log \mathrm{C}(x|y))$$

Note: the $+O(\log E(x,y))$ can not be omitted...

(2) Optimality. Up to $O(1)$ additive terms, it is the minimal symmetric function D

(i) that is approximable from above,

(ii) has balls $\{y : D(x,y) \leq k\}$ of size at most 2^k.

Proof. $E_U(x,y) \geq \mathrm{C}(x|y) + O(1)$, is obvious. The $\geq$-inequality holds.

The plain information distance

$$E_U(x,y) = \min\{|p|: U(p,x) = y \text{ and } U(p,y) = x\},$$

We use machines U that minimize this distance up to $O(1)$ and drop index U.

Properties.

(1) Symmetric and non-negative.
Almost satisfies the axioms of a metric:

– $E(x,x) \leq O(1)$

– $E(x,z) \leq E(x,y) + E(y,z) + O(\log E(x,y))$,

Indeed, similar to

$$\mathrm{C}(x|z) \leq \mathrm{C}(x|y) + \mathrm{C}(y|z) + O(\log \mathrm{C}(x|y))$$

Note: the $+O(\log E(x,y))$ can not be omitted...

(2) Optimality. Up to $O(1)$ additive terms, it is the minimal symmetric function D

(i) that is approximable from above,

(ii) has balls $\{y : D(x,y) \leq k\}$ of size at most 2^k.

$$E(x,y) = \max\{\mathrm{C}(x|y), \mathrm{C}(y|x)\} + O(1).$$

Proof. $E_U(x,y) \geq \mathrm{C}(x|y) + O(1)$, is obvious. The $\geq$-inequality holds.

The $\leq$-inequality. Consider for each k the binary relation on strings (of all lengths)

$$R_k(x,y) \iff \mathrm{C}(x|y) < k \text{ and } \mathrm{C}(y|x) < k.$$

The plain information distance

$$E_U(x,y) = \min\{|p|\colon U(p,x) = y \text{ and } U(p,y) = x\},$$

We use machines U that minimize this distance up to $O(1)$ and drop index U.

Properties.

(1) Symmetric and non-negative.
Almost satisfies the axioms of a metric:

– $E(x,x) \leq O(1)$

– $E(x,z) \leq E(x,y) + E(y,z) + O(\log E(x,y))$,

Indeed, similar to

$$\mathrm{C}(x|z) \leq \mathrm{C}(x|y) + \mathrm{C}(y|z) + O(\log \mathrm{C}(x|y))$$

Note: the $+O(\log E(x,y))$ can not be omitted...

(2) Optimality. Up to $O(1)$ additive terms, it is the minimal symmetric function D

(i) that is approximable from above,

(ii) has balls $\{y : D(x,y) \leq k\}$ of size at most 2^k.

$$E(x,y) = \max\{\mathrm{C}(x|y), \mathrm{C}(y|x)\} + O(1).$$

Proof. $E_U(x,y) \geq \mathrm{C}(x|y) + O(1)$, is obvious. The $\geq$-inequality holds.

The $\leq$-inequality. Consider for each k the binary relation on strings (of all lengths)

$$R_k(x,y) \iff \mathrm{C}(x|y) < k \text{ and } \mathrm{C}(y|x) < k.$$

R_k is (computably) enumerable uniformly in k.

The plain information distance

$$E_U(x,y) = \min\{|p|: U(p,x) = y \text{ and } U(p,y) = x\},$$

We use machines U that minimize this distance up to $O(1)$ and drop index U.

Properties.

(1) Symmetric and non-negative.
Almost satisfies the axioms of a metric:
– $E(x,x) \leq O(1)$
– $E(x,z) \leq E(x,y) + E(y,z) + O(\log E(x,y))$,

Indeed, similar to

$$\mathrm{C}(x|z) \leq \mathrm{C}(x|y) + \mathrm{C}(y|z) + O(\log \mathrm{C}(x|y))$$

Note: the $+O(\log E(x,y))$ can not be omitted...

(2) Optimality. Up to $O(1)$ additive terms, it is the minimal symmetric function D

(i) that is approximable from above,

(ii) has balls $\{y : D(x,y) \leq k\}$ of size at most 2^k.

$$E(x,y) = \max\{\mathrm{C}(x|y), \mathrm{C}(y|x)\} + O(1).$$

Proof. $E_U(x,y) \geq \mathrm{C}(x|y) + O(1)$, is obvious. The $\geq$-inequality holds.

The $\leq$-inequality. Consider for each k the binary relation on strings (of all lengths)

$$R_k(x,y) \iff \mathrm{C}(x|y) < k \text{ and } \mathrm{C}(y|x) < k.$$

R_k is (computably) enumerable uniformly in k.

Graph.

- Vertices are strings. Edges are pairs in R_k.

The plain information distance

$$E_U(x,y) = \min\{|p|\colon U(p,x) = y \text{ and } U(p,y) = x\},$$

We use machines U that minimize this distance up to $O(1)$ and drop index U.

Properties.

(1) Symmetric and non-negative.
Almost satisfies the axioms of a metric:

– $E(x,x) \leq O(1)$

– $E(x,z) \leq E(x,y) + E(y,z) + O(\log E(x,y))$,

Indeed, similar to

$$\mathrm{C}(x|z) \leq \mathrm{C}(x|y) + \mathrm{C}(y|z) + O(\log \mathrm{C}(x|y))$$

Note: the $+O(\log E(x,y))$ can not be omitted...

(2) Optimality. Up to $O(1)$ additive terms, it is the minimal symmetric function D

(i) that is approximable from above,

(ii) has balls $\{y : D(x,y) \leq k\}$ of size at most 2^k.

$$E(x,y) = \max\{\mathrm{C}(x|y), \mathrm{C}(y|x)\} + O(1).$$

Proof. $E_U(x,y) \geq \mathrm{C}(x|y) + O(1)$, is obvious. The $\geq$-inequality holds.

The $\leq$-inequality. Consider for each k the binary relation on strings (of all lengths)

$$R_k(x,y) \iff \mathrm{C}(x|y) < k \text{ and } \mathrm{C}(y|x) < k.$$

R_k is (computably) enumerable uniformly in k.

Graph.

- Vertices are strings. Edges are pairs in R_k.
- The degree of each vertex x is at most $N = 2^k$, because less than 2^k strings y satisfy $\mathrm{C}_U(y|x) < k$.

The plain information distance

$$E_U(x,y) = \min\{|p|\colon U(p,x) = y \text{ and } U(p,y) = x\},$$

We use machines U that minimize this distance up to $O(1)$ and drop index U.

Properties.

(1) Symmetric and non-negative.
Almost satisfies the axioms of a metric:

– $E(x,x) \leq O(1)$

– $E(x,z) \leq E(x,y) + E(y,z) + O(\log E(x,y))$,

Indeed, similar to

$$\mathrm{C}(x|z) \leq \mathrm{C}(x|y) + \mathrm{C}(y|z) + O(\log \mathrm{C}(x|y))$$

Note: the $+O(\log E(x,y))$ can not be omitted...

(2) Optimality. Up to $O(1)$ additive terms, it is the minimal symmetric function D

(i) that is approximable from above,

(ii) has balls $\{y : D(x,y) \leq k\}$ of size at most 2^k.

$$E(x,y) = \max\{\mathrm{C}(x|y), \mathrm{C}(y|x)\} + O(1).$$

Proof. $E_U(x,y) \geq \mathrm{C}(x|y) + O(1)$, is obvious. The $\geq$-inequality holds.

The $\leq$-inequality. Consider for each k the binary relation on strings (of all lengths)

$$R_k(x,y) \iff \mathrm{C}(x|y) < k \text{ and } \mathrm{C}(y|x) < k.$$

R_k is (computably) enumerable uniformly in k.

Graph.

- Vertices are strings. Edges are pairs in R_k.
- The degree of each vertex x is at most $N = 2^k$, because less than 2^k strings y satisfy $\mathrm{C}_U(y|x) < k$.
- We enumerate R_k and color edges with $2N - 1$ colors. Edges adjacent to the same vertex must have $\neq$ colors.

The plain information distance

$$E_U(x,y) = \min\{|p|\colon U(p,x) = y \text{ and } U(p,y) = x\},$$

We use machines U that minimize this distance up to $O(1)$ and drop index U.

Properties.

(1) Symmetric and non-negative.
Almost satisfies the axioms of a metric:

– $E(x,x) \leq O(1)$

– $E(x,z) \leq E(x,y) + E(y,z) + O(\log E(x,y))$,

Indeed, similar to

$$\mathrm{C}(x|z) \leq \mathrm{C}(x|y) + \mathrm{C}(y|z) + O(\log \mathrm{C}(x|y))$$

Note: the $+O(\log E(x,y))$ can not be omitted...

(2) Optimality. Up to $O(1)$ additive terms, it is the minimal symmetric function D

(i) that is approximable from above,

(ii) has balls $\{y : D(x,y) \leq k\}$ of size at most 2^k.

$$E(x,y) = \max\{\mathrm{C}(x|y), \mathrm{C}(y|x)\} + O(1).$$

Proof. $E_U(x,y) \geq \mathrm{C}(x|y) + O(1)$, is obvious. The $\geq$-inequality holds.

The $\leq$-inequality. Consider for each k the binary relation on strings (of all lengths)

$$R_k(x,y) \iff \mathrm{C}(x|y) < k \text{ and } \mathrm{C}(y|x) < k.$$

R_k is (computably) enumerable uniformly in k.

Graph.

- Vertices are strings. Edges are pairs in R_k.
- The degree of each vertex x is at most $N = 2^k$, because less than 2^k strings y satisfy $\mathrm{C}_U(y|x) < k$.
- We enumerate R_k and color edges with $2N - 1$ colors. Edges adjacent to the same vertex must have $\neq$ colors.

The plain information distance

$$E_U(x,y) = \min\{|p|\colon U(p,x) = y \text{ and } U(p,y) = x\},$$

We use machines U that minimize this distance up to $O(1)$ and drop index U.

Properties.

(1) Symmetric and non-negative.
Almost satisfies the axioms of a metric:

– $E(x,x) \leq O(1)$

– $E(x,z) \leq E(x,y) + E(y,z) + O(\log E(x,y))$,

Indeed, similar to

$$\mathrm{C}(x|z) \leq \mathrm{C}(x|y) + \mathrm{C}(y|z) + O(\log \mathrm{C}(x|y))$$

Note: the $+O(\log E(x,y))$ can not be omitted...

(2) Optimality. Up to $O(1)$ additive terms, it is the minimal symmetric function D

(i) that is approximable from above,

(ii) has balls $\{y : D(x,y) \leq k\}$ of size at most 2^k.

$$E(x,y) = \max\{\mathrm{C}(x|y), \mathrm{C}(y|x)\} + O(1).$$

Proof. $E_U(x,y) \geq \mathrm{C}(x|y) + O(1)$, is obvious. The $\geq$-inequality holds.

The $\leq$-inequality. Consider for each k the binary relation on strings (of all lengths)

$$R_k(x,y) \iff \mathrm{C}(x|y) < k \text{ and } \mathrm{C}(y|x) < k.$$

R_k is (computably) enumerable uniformly in k.

Graph.

- Vertices are strings. Edges are pairs in R_k.
- The degree of each vertex x is at most $N = 2^k$, because less than 2^k strings y satisfy $\mathrm{C}_U(y|x) < k$.
- We enumerate R_k and color edges with $2N-1$ colors. Edges adjacent to the same vertex must have $\neq$ colors.

$2N-1$ colors is enough: when adding an edge, at most $\leq 2(N-1)$ colors are used in the 2 adjacent vertices.

The plain information distance

$$E_U(x,y) = \min\{|p|\colon U(p,x) = y \text{ and } U(p,y) = x\},$$

We use machines U that minimize this distance up to $O(1)$ and drop index U.

Properties.

(1) Symmetric and non-negative.
Almost satisfies the axioms of a metric:

– $E(x,x) \leq O(1)$

– $E(x,z) \leq E(x,y) + E(y,z) + O(\log E(x,y))$,

Indeed, similar to

$$\mathrm{C}(x|z) \leq \mathrm{C}(x|y) + \mathrm{C}(y|z) + O(\log \mathrm{C}(x|y))$$

Note: the $+O(\log E(x,y))$ can not be omitted...

(2) Optimality. Up to $O(1)$ additive terms, it is the minimal symmetric function D

(i) that is approximable from above,

(ii) has balls $\{y : D(x,y) \leq k\}$ of size at most 2^k.

$$E(x,y) = \max\{\mathrm{C}(x|y), \mathrm{C}(y|x)\} + O(1).$$

Proof. $E_U(x,y) \geq \mathrm{C}(x|y) + O(1)$, is obvious. The $\geq$-inequality holds.

The $\leq$-inequality. Consider for each k the binary relation on strings (of all lengths)

$$R_k(x,y) \iff \mathrm{C}(x|y) < k \text{ and } \mathrm{C}(y|x) < k.$$

R_k is (computably) enumerable uniformly in k.

Graph.

- Vertices are strings. Edges are pairs in R_k.
- The degree of each vertex x is at most $N = 2^k$, because less than 2^k strings y satisfy $\mathrm{C}_U(y|x) < k$.
- We enumerate R_k and color edges with $2N-1$ colors. Edges adjacent to the same vertex must have $\neq$ colors.

$2N-1$ colors is enough: when adding an edge, at most $\leq 2(N-1)$ colors are used in the 2 adjacent vertices.

We use $(k+1)$-bit strings as colors.

The plain information distance

$$E_U(x,y) = \min\{|p|\colon U(p,x)=y \text{ and } U(p,y)=x\},$$

We use machines U that minimize this distance up to $O(1)$ and drop index U.

Properties.

(1) Symmetric and non-negative.
Almost satisfies the axioms of a metric:

– $E(x,x) \leq O(1)$

– $E(x,z) \leq E(x,y) + E(y,z) + O(\log E(x,y))$,

Indeed, similar to

$$\mathrm{C}(x|z) \leq \mathrm{C}(x|y) + \mathrm{C}(y|z) + O(\log \mathrm{C}(x|y))$$

Note: the $+O(\log E(x,y))$ can not be omitted...

(2) Optimality. Up to $O(1)$ additive terms, it is the minimal symmetric function D

(i) that is approximable from above,

(ii) has balls $\{y : D(x,y) \leq k\}$ of size at most 2^k.

$$E(x,y) = \max\{\mathrm{C}(x|y), \mathrm{C}(y|x)\} + O(1).$$

Proof. $E_U(x,y) \geq \mathrm{C}(x|y) + O(1)$, is obvious. The $\geq$-inequality holds.

The $\leq$-inequality. Consider for each k the binary relation on strings (of all lengths)

$$R_k(x,y) \iff \mathrm{C}(x|y) < k \text{ and } \mathrm{C}(y|x) < k.$$

R_k is (computably) enumerable uniformly in k.

Graph.

- Vertices are strings. Edges are pairs in R_k.
- The degree of each vertex x is at most $N = 2^k$, because less than 2^k strings y satisfy $\mathrm{C}_U(y|x) < k$.
- We enumerate R_k and color edges with $2N-1$ colors. Edges adjacent to the same vertex must have $\neq$ colors.

$2N-1$ colors is enough: when adding an edge, at most $\leq 2(N-1)$ colors are used in the 2 adjacent vertices.

We use $(k+1)$-bit strings as colors.
We construct a machine U for which $U(p,x) = y$ and $U(p,y) = x$ if $(x,y) \in R_k$. □

Can we satisfy the definition of a metric precisely? Prefix complexity.

- A machine U is *prefix-free* if for all strings y, p: if $U(p, y)$ is defined, then $U(q)$ is undefined for each strict prefix q of p.

Can we satisfy the definition of a metric precisely? Prefix complexity.

- A machine U is *prefix-free* if for all strings y, p: if $U(p, y)$ is defined, then $U(q)$ is undefined for each strict prefix q of p.
- There exists a prefix-free machine U for which $C_U(x|y)$ is optimal among all prefix-free machines. Fix such U and write Kolmogorov complexity as $K(x|y)$.

Can we satisfy the definition of a metric precisely? Prefix complexity.

- A machine U is *prefix-free* if for all strings y, p: if $U(p, y)$ is defined, then $U(q)$ is undefined for each strict prefix q of p.
- There exists a prefix-free machine U for which $C_U(x|y)$ is optimal among all prefix-free machines. Fix such U and write Kolmogorov complexity as $K(x|y)$.
- Plain and prefix complexity are close: $\mathrm{K}(x) \leq \mathrm{C}(x) + O(\log \mathrm{C}(x))$.

Can we satisfy the definition of a metric precisely? Prefix complexity.

- A machine U is *prefix-free* if for all strings y, p: if $U(p, y)$ is defined, then $U(q)$ is undefined for each strict prefix q of p.
- There exists a prefix-free machine U for which $C_U(x|y)$ is optimal among all prefix-free machines. Fix such U and write Kolmogorov complexity as $K(x|y)$.
- Plain and prefix complexity are close: $\mathrm{K}(x) \leq \mathrm{C}(x) + O(\log \mathrm{C}(x))$.
- Let

$$E'_U(x, y) = \max\{K_U(x|y), K_U(y|x)\}.$$

Can we satisfy the definition of a metric precisely? Prefix complexity.

- A machine U is *prefix-free* if for all strings y, p: if $U(p, y)$ is defined, then $U(q)$ is undefined for each strict prefix q of p.
- There exists a prefix-free machine U for which $C_U(x|y)$ is optimal among all prefix-free machines. Fix such U and write Kolmogorov complexity as $K(x|y)$.
- Plain and prefix complexity are close: $\mathrm{K}(x) \leq \mathrm{C}(x) + O(\log \mathrm{C}(x))$.
- Let

$$E'_U(x, y) = \max\{K_U(x|y), K_U(y|x)\}.$$

- $E'(x, z) \leq E'(x, y) + E'(y, z) + O(1)$.

Can we satisfy the definition of a metric precisely? Prefix complexity.

- A machine U is *prefix-free* if for all strings y, p: if $U(p, y)$ is defined, then $U(q)$ is undefined for each strict prefix q of p.
- There exists a prefix-free machine U for which $C_U(x|y)$ is optimal among all prefix-free machines. Fix such U and write Kolmogorov complexity as $K(x|y)$.
- Plain and prefix complexity are close: $\mathrm{K}(x) \leq \mathrm{C}(x) + O(\log \mathrm{C}(x))$.
- Let

$$E'_U(x, y) = \max\{K_U(x|y), K_U(y|x)\}.$$

- $E'(x, z) \leq E'(x, y) + E'(y, z) + O(1)$.
- Let c be large. The function

$$D_U(x, y) = \begin{cases} E'_U(x, y) + c & \text{if } x \neq y \\ 0 & \text{otherwise} \end{cases}$$

satisfies the axioms of a metric precisely.

Can we satisfy the definition of a metric precisely? Prefix complexity.

- A machine U is *prefix-free* if for all strings y, p: if $U(p, y)$ is defined, then $U(q)$ is undefined for each strict prefix q of p.
- There exists a prefix-free machine U for which $C_U(x|y)$ is optimal among all prefix-free machines. Fix such U and write Kolmogorov complexity as $K(x|y)$.
- Plain and prefix complexity are close: $\mathrm{K}(x) \leq \mathrm{C}(x) + O(\log \mathrm{C}(x))$.
- Let

$$E'_U(x, y) = \max\{K_U(x|y), K_U(y|x)\}.$$

- $E'(x, z) \leq E'(x, y) + E'(y, z) + O(1)$.
- Let c be large. The function

$$D_U(x, y) = \begin{cases} E'_U(x, y) + c & \text{if } x \neq y \\ 0 & \text{otherwise} \end{cases}$$

  satisfies the axioms of a metric precisely.
- Optimality. Up to $O(1)$ additive terms, it is the minimal symmetrical function D

Can we satisfy the definition of a metric precisely? Prefix complexity.

- A machine U is *prefix-free* if for all strings y, p: if $U(p, y)$ is defined, then $U(q)$ is undefined for each strict prefix q of p.
- There exists a prefix-free machine U for which $C_U(x|y)$ is optimal among all prefix-free machines. Fix such U and write Kolmogorov complexity as $K(x|y)$.
- Plain and prefix complexity are close: $\mathrm{K}(x) \leq \mathrm{C}(x) + O(\log \mathrm{C}(x))$.
- Let

$$E'_U(x, y) = \max\{K_U(x|y), K_U(y|x)\}.$$

- $E'(x, z) \leq E'(x, y) + E'(y, z) + O(1)$.
- Let c be large. The function

$$D_U(x, y) = \begin{cases} E'_U(x, y) + c & \text{if } x \neq y \\ 0 & \text{otherwise} \end{cases}$$

  satisfies the axioms of a metric precisely.
- Optimality. Up to $O(1)$ additive terms, it is the minimal symmetrical function D
  (i) that is approximable from above,

Can we satisfy the definition of a metric precisely? Prefix complexity.

- A machine U is *prefix-free* if for all strings y, p: if $U(p, y)$ is defined, then $U(q)$ is undefined for each strict prefix q of p.
- There exists a prefix-free machine U for which $C_U(x|y)$ is optimal among all prefix-free machines. Fix such U and write Kolmogorov complexity as $K(x|y)$.
- Plain and prefix complexity are close: $\mathrm{K}(x) \leq \mathrm{C}(x) + O(\log \mathrm{C}(x))$.
- Let

$$E'_U(x, y) = \max\{K_U(x|y), K_U(y|x)\}.$$

- $E'(x, z) \leq E'(x, y) + E'(y, z) + O(1)$.
- Let c be large. The function

$$D_U(x, y) = \begin{cases} E'_U(x, y) + c & \text{if } x \neq y \\ 0 & \text{otherwise} \end{cases}$$

  satisfies the axioms of a metric precisely.
- Optimality. Up to $O(1)$ additive terms, it is the minimal symmetrical function D
  (i) that is approximable from above,
  (ii) for which $\sum_y 2^{-D(x,y)} \leq 1$.

Can we satisfy the definition of a metric precisely? Prefix complexity.

- A machine U is *prefix-free* if for all strings y, p: if $U(p, y)$ is defined, then $U(q)$ is undefined for each strict prefix q of p.
- There exists a prefix-free machine U for which $C_U(x|y)$ is optimal among all prefix-free machines. Fix such U and write Kolmogorov complexity as $K(x|y)$.
- Plain and prefix complexity are close: $\mathrm{K}(x) \leq \mathrm{C}(x) + O(\log \mathrm{C}(x))$.
- Let

$$E'_U(x, y) = \max\{K_U(x|y), K_U(y|x)\}.$$

- $E'(x, z) \leq E'(x, y) + E'(y, z) + O(1)$.
- Let c be large. The function

$$D_U(x, y) = \begin{cases} E'_U(x, y) + c & \text{if } x \neq y \\ 0 & \text{otherwise} \end{cases}$$

  satisfies the axioms of a metric precisely.
- Optimality. Up to $O(1)$ additive terms, it is the minimal symmetrical function D
  - (i) that is approximable from above,
  - (ii) for which $\sum_y 2^{-D(x,y)} \leq 1$.

- Consider E_U for a prefix-free machine U for which this distance is minimal up to $+O(1)$.

Can we satisfy the definition of a metric precisely? Prefix complexity.

- A machine U is *prefix-free* if for all strings y, p: if $U(p, y)$ is defined, then $U(q)$ is undefined for each strict prefix q of p.
- There exists a prefix-free machine U for which $C_U(x|y)$ is optimal among all prefix-free machines. Fix such U and write Kolmogorov complexity as $K(x|y)$.
- Plain and prefix complexity are close: $\mathrm{K}(x) \leq \mathrm{C}(x) + O(\log \mathrm{C}(x))$.
- Let

$$E'_U(x, y) = \max\{K_U(x|y), K_U(y|x)\}.$$

- $E'(x, z) \leq E'(x, y) + E'(y, z) + O(1)$.
- Let c be large. The function

$$D_U(x, y) = \begin{cases} E'_U(x, y) + c & \text{if } x \neq y \\ 0 & \text{otherwise} \end{cases}$$

  satisfies the axioms of a metric precisely.
- Optimality. Up to $O(1)$ additive terms, it is the minimal symmetrical function D
  (i) that is approximable from above,
  (ii) for which $\sum_y 2^{-D(x,y)} \leq 1$.

- Consider E_U for a prefix-free machine U for which this distance is minimal up to $+O(1)$.
- We have
  $E(x, y) = E'(x, y) + O(\log E'(x, y))$.

Can we satisfy the definition of a metric precisely? Prefix complexity.

- A machine U is *prefix-free* if for all strings y, p: if $U(p, y)$ is defined, then $U(q)$ is undefined for each strict prefix q of p.
- There exists a prefix-free machine U for which $C_U(x|y)$ is optimal among all prefix-free machines. Fix such U and write Kolmogorov complexity as $K(x|y)$.
- Plain and prefix complexity are close: $\mathrm{K}(x) \leq \mathrm{C}(x) + O(\log \mathrm{C}(x))$.
- Let

$$E'_U(x, y) = \max\{K_U(x|y), K_U(y|x)\}.$$

- $E'(x, z) \leq E'(x, y) + E'(y, z) + O(1)$.
- Let c be large. The function

$$D_U(x, y) = \begin{cases} E'_U(x, y) + c & \text{if } x \neq y \\ 0 & \text{otherwise} \end{cases}$$

  satisfies the axioms of a metric precisely.
- Optimality. Up to $O(1)$ additive terms, it is the minimal symmetrical function D
  - (i) that is approximable from above,
  - (ii) for which $\sum_y 2^{-D(x,y)} \leq 1$.

- Consider E_U for a prefix-free machine U for which this distance is minimal up to $+O(1)$.
- We have $E(x, y) = E'(x, y) + O(\log E'(x, y))$.
- Is it true that $E(x, y) = E'(x, y) + O(1)$?

Can we satisfy the definition of a metric precisely? Prefix complexity.

- A machine U is *prefix-free* if for all strings y, p: if $U(p, y)$ is defined, then $U(q)$ is undefined for each strict prefix q of p.
- There exists a prefix-free machine U for which $C_U(x|y)$ is optimal among all prefix-free machines. Fix such U and write Kolmogorov complexity as $K(x|y)$.
- Plain and prefix complexity are close: $\mathrm{K}(x) \leq \mathrm{C}(x) + O(\log \mathrm{C}(x))$.
- Let

$$E'_U(x, y) = \max\{K_U(x|y), K_U(y|x)\}.$$

- $E'(x, z) \leq E'(x, y) + E'(y, z) + O(1)$.
- Let c be large. The function

$$D_U(x, y) = \begin{cases} E'_U(x, y) + c & \text{if } x \neq y \\ 0 & \text{otherwise} \end{cases}$$

  satisfies the axioms of a metric precisely.
- Optimality. Up to $O(1)$ additive terms, it is the minimal symmetrical function D
  (i) that is approximable from above,
  (ii) for which $\sum_y 2^{-D(x,y)} \leq 1$.

- Consider E_U for a prefix-free machine U for which this distance is minimal up to $+O(1)$.
- We have
  $E(x, y) = E'(x, y) + O(\log E'(x, y))$.
- Is it true that $E(x, y) = E'(x, y) + O(1)$?
- Conjectured to be false in [BGLVZ 1998].

Can we satisfy the definition of a metric precisely? Prefix complexity.

- A machine U is *prefix-free* if for all strings y, p: if $U(p, y)$ is defined, then $U(q)$ is undefined for each strict prefix q of p.
- There exists a prefix-free machine U for which $C_U(x|y)$ is optimal among all prefix-free machines. Fix such U and write Kolmogorov complexity as $K(x|y)$.
- Plain and prefix complexity are close: $\mathrm{K}(x) \leq \mathrm{C}(x) + O(\log \mathrm{C}(x))$.
- Let

$$E'_U(x, y) = \max\{K_U(x|y), K_U(y|x)\}.$$

- $E'(x, z) \leq E'(x, y) + E'(y, z) + O(1)$.
- Let c be large. The function

$$D_U(x, y) = \begin{cases} E'_U(x, y) + c & \text{if } x \neq y \\ 0 & \text{otherwise} \end{cases}$$

  satisfies the axioms of a metric precisely.
- Optimality. Up to $O(1)$ additive terms, it is the minimal symmetrical function D
  (i) that is approximable from above,
  (ii) for which $\sum_y 2^{-D(x,y)} \leq 1$.

- Consider E_U for a prefix-free machine U for which this distance is minimal up to $+O(1)$.
- We have
  $E(x, y) = E'(x, y) + O(\log E'(x, y))$.
- Is it true that $E(x, y) = E'(x, y) + O(1)$?
- Conjectured to be false in [BGLVZ 1998].
- 2010-2016: 3 papers claimed this was true, with incorrect proofs.

Can we satisfy the definition of a metric precisely? Prefix complexity.

- A machine U is *prefix-free* if for all strings y, p: if $U(p, y)$ is defined, then $U(q)$ is undefined for each strict prefix q of p.
- There exists a prefix-free machine U for which $C_U(x|y)$ is optimal among all prefix-free machines. Fix such U and write Kolmogorov complexity as $K(x|y)$.
- Plain and prefix complexity are close: $\mathrm{K}(x) \leq \mathrm{C}(x) + O(\log \mathrm{C}(x))$.
- Let

$$E'_U(x, y) = \max\{K_U(x|y), K_U(y|x)\}.$$

- $E'(x, z) \leq E'(x, y) + E'(y, z) + O(1)$.
- Let c be large. The function

$$D_U(x, y) = \begin{cases} E'_U(x, y) + c & \text{if } x \neq y \\ 0 & \text{otherwise} \end{cases}$$

  satisfies the axioms of a metric precisely.
- Optimality. Up to $O(1)$ additive terms, it is the minimal symmetrical function D
  (i) that is approximable from above,
  (ii) for which $\sum_y 2^{-D(x,y)} \leq 1$.

- Consider E_U for a prefix-free machine U for which this distance is minimal up to $+O(1)$.
- We have $E(x, y) = E'(x, y) + O(\log E'(x, y))$.
- Is it true that $E(x, y) = E'(x, y) + O(1)$?
- Conjectured to be false in [BGLVZ 1998].
- 2010-2016: 3 papers claimed this was true, with incorrect proofs.
- This work: we prove that it is false, but holds if $|x| = |y|$ and $E(x, y) \geq 6 \log |x|$.

Main results

Theorem

If $|x| = |y|$ *and* $E'(x, y) \geq 6 \log |x|$, *then* $E(x, y) = E'(x, y) + O(1)$.

Main results

Theorem

If $|x| = |y|$ *and* $E'(x, y) \geq 6 \log |x|$, *then* $E(x, y) = E'(x, y) + O(1)$.

Theorem

$E(x, y) - E'(x, y) \geq 0.99 \log \log n$ *on* n*-bit* x *and* y.

Main results

Theorem

If $|x| = |y|$ *and* $E'(x, y) \geq 6 \log |x|$, *then* $E(x, y) = E'(x, y) + O(1)$.

Theorem

$E(x, y) - E'(x, y) \geq 0.99 \log \log n$ *on* n*-bit* x *and* y.

Recent interest: extention to commen information in a tuple of strings.

Main results

Theorem

If $|x| = |y|$ *and* $E'(x,y) \geq 6 \log |x|$, *then* $E(x,y) = E'(x,y) + O(1)$.

Theorem

$E(x,y) - E'(x,y) \geq 0.99 \log \log n$ *on* n*-bit* x *and* y.

Recent interest: extention to commen information in a tuple of strings.

Open questions:

- Does E satisfy the triangle inequality up to $O(1)$?
  Recall that E' satisfies the triangle inequality.
- We can define the information distance with
  - 2 machines U and V: min $|p|$ such that $U(p,x) = y$ and $V(p,y) = x$,
  - with prefix-stable machines: if $U(p,y)$ and $U(pq,y)$ are defined, they must be the same.
  $\Rightarrow$ 4 variants of the prefix distance. Are they equal?

Main results

Theorem

If $|x| = |y|$ *and* $E'(x,y) \geq 6 \log |x|$, *then* $E(x,y) = E'(x,y) + O(1)$.

Theorem

$E(x,y) - E'(x,y) \geq 0.99 \log \log n$ *on* n*-bit* x *and* y.

Recent interest: extention to commen information in a tuple of strings.

Open questions:

- Does E satisfy the triangle inequality up to $O(1)$?
  Recall that E' satisfies the triangle inequality.
- We can define the information distance with
  - 2 machines U and V: min $|p|$ such that $U(p,x) = y$ and $V(p,y) = x$,
  - with prefix-stable machines: if $U(p,y)$ and $U(pq,y)$ are defined, they must be the same.
  $\Rightarrow$ 4 variants of the prefix distance. Are they equal?

Questions?

Textbooks

Ming Li and Paul M.B. Vitányi. An Introduction to Kolmogorov Complexity and Its Applications, 4th edition. Springer, 2019.

Alexander Shen, Vladimir A. Uspensky, and Nikolay Vereshchagin. Kolmogorov complexity and algorithmic randomness, volume 220. American Mathematical Soc., 2017.

New results

Bruno Bauwens.
Precise expression for the algorithmic information distance.
arXiv preprint arXiv:2009.00469, 2020.

Founding paper

Charles H. Bennett, Péter Gács, Ming Li, Paul M.B. Vitányi, and Wojciech H. Zurek.
Information distance.
IEEE Transactions on information theory, 44(4):1407–1423, 1998.

Applications

M. Li, X. Chen, X. Li, B. Ma, and B.M.B. Vitányi.
The similarity metric.
Information Theory, IEEE Transactions on, 50(12):3250–3264, 2004.

Rudi L Cilibrasi and Paul MB Vitanyi.
The google similarity distance.
IEEE Transactions on knowledge and data engineering, 19(3):370–383, 2007.

Gleb Filatov, Bruno Bauwens, and Attila Kertsz-Farkas.
LZW-Kernel: fast kernel utilizing variable length code blocks from LZW compressors for protein sequence classification.
Bioinformatics, 34(19):3281–3288, 05 2018.

Eamonn Keogh, Stefano Lonardi, Chotirat Ann Ratanamahatana, Li Wei, Sang-Hee Lee, and John Handley.
Compression-based data mining of sequential data.
Data Mining and Knowledge Discovery, 14(1):99–129, 2007.

On the characterization for plain complexity

Paul M.B. Vitányi. Exact expression for information distance.
IEEE Transactions on Information Theory, 2017.

Algorithmic set distance

Chong Long, Xiaoyan Zhu, Ming Li, and Bin Ma. Information shared by many objects.
In Proceedings of the 17th ACM conference on Information and knowledge management, pages 1213–1220. ACM, 2008.

Paul M.B. Vitányi. Information distance in multiples.
IEEE Transactions on Information Theory, 57(4):2451–2456, 2011.

Logic papers

Laurent Bienvenu, Andrei Romashchenko, Alexander Shen, Antoine Taveneaux, and Stijn Vermeeren. The axiomatic power of kolmogorov complexity.
Annals of Pure and Applied Logic, 165(9):1380–1402, 2014.

Shira Kritchman and Ran Raz. The surprise examination paradox and the second incompleteness theorem.
Notices of the AMS, 57(11):1454–1458, 2010.

Yong Cheng. Current research on gödels incompleteness theorems.
Bulletin of Symbolic Logic, 27(2):113–167, 2021.