Extracting Randomness and Evenly Distributed Hypergraphs

Matthew Harrison-Trainor

University of Michigan

Midwest Computability Seminar, February 2022

Suppose that σ is a string of length n, which is "one-quarter random". Can I produce a string τ, possibly shorter than σ, which is "half-random"?

We need to add some reasonable rules:

1. The procedure which produces τ from σ should be effective and work on all strings σ which are "one-quarter random".
2. The strings τ should not all be small; for each length m, there should be an $n = f(m)$ such that the procedure turns strings σ of length n into strings τ of length m.

The answer is NO.

Suppose that σ is a string of length n, which is “one-quarter random”. Can I produce a k strings $\tau_1, \ldots, \tau_k$, possibly shorter than σ, at least one of which is “half-random”?

The answer is YES (for the right k).

In this talk, we'll learn a littly bit about why, and how big k has to be to get a particular increase in randomness.

Other goals:

1. Learn something about Kolmogorov complexity.
2. A cool connection with graph theory.

This is joint work with Bienvenu and Csima.

Outline

1. Kolmogorov Complexity
2. Kolmogorov Extractors
3. Connections with Graphs
4. Edge Distribution in Graphs

Kolmogorov Complexity

Formalizing Randomness

Informal Definition

The Kolmogorov complexity of a string σ is the length shortest description of σ.

Definition

Let M be a (Turing) machine which takes finite strings as input and output. The (plain) Kolmogorov complexity of a string σ relative to M is

$$C_M(\sigma) = \min\{|\tau| : M(\tau) = \sigma\}.$$

Call a string τ such that $M(\tau) = \sigma$ an M-description of σ.

Example

Let U be the machine which, on input $0^n 1\tau$, finds the nth machine M_n and computes and outputs $M_n(\tau)$.

For any given machine M_n, U gives descriptions which are of length only $n + 1$ (a constant) worse than M_n.

Definition

A machine U is called universal or optimal if, for every machine M, there is a constant c_M such that:

$$(\forall \sigma \in \{0,1\}^*) \qquad U(\sigma) \leq M(\sigma) + c_M.$$

Fix, forever, a universal machine U.

Definition

For $\sigma \in \{0,1\}^*$, the (plain) Kolmogorov complexity of σ is

$$C(\sigma) = C_U(\sigma) = \min\{|\tau| : U(\tau) = \sigma\}.$$

We could have chosen a different universal machine, but this would be the same up to a constant.

Warning: Asking about the Kolmogorov complexity of a particular string σ does not really make sense, because the value depends on the universal machine. It only makes sense to ask questions which do not depend on the constant.

- A string like

  $$100100100100100100100100100100100100$$

  would have low Kolmogorov complexity.
- The strings of binary digits of π would have low Kolmogorov complexity.
- A string like

  $$110101101010001011101011101000011010$$

  would have high Kolmogorov complexity.

Theorem

There is a constant c such that, for all $\sigma \in \{0,1\}^*$,

$$C(\sigma) < |\sigma| + c.$$

Proof.

There is a machine M which is just the identity:

$$M(\sigma) = \sigma.$$

Then

$$C_M(\sigma) = |\sigma|.$$

But the universal machine U is as good as M up to a constant c_M,so

$$C(\sigma) = C_U(\sigma) \leq C_M(\sigma) + c_M \leq |\sigma| + c_M.$$

□

Fact

There are at most $2^{r+1} - 1$ strings σ with $C(\sigma) \leq r$.
(Incompressible strings must exist.)

Proof.

For each string σ with $C(\sigma) \leq r$, there is a τ of length $\leq r$ such that $U(\tau) = \sigma$.

There are only so many strings τ of length at most r to go around; exactly

$$2^0 + 2^1 + 2^2 + \cdots + 2^r = 2^{r+1} - 1$$

of them. □

Theorem

C is not a computable function.

U is partial computable, so $U(\tau)$ may take a long time to compute, or may not halt at all.

Think of $C(\sigma)$ as being dynamically approximable as a decreasing sequence:

- Let $C_s(\sigma)$ be the length of the shortest U-description of σ we have found by time s.

Then $C_s(\sigma)$ is decreasing as s increases, and eventually stabilizes at $C(\sigma)$.

We can think of $C(\sigma)$ as the information content of σ, but we can also think of it as measuring the randomness of σ.

Definition

$C(\sigma)/|\sigma|$ is the rate of randomness or information densitity of σ.

Consider creating a binary string of length $3n$ by flipping n coins, and making every heads into 000 and every tails into 111:

$$HTHHTH \ldots \mapsto 000111000000111000 \ldots .$$

We expect this sequence to have rate of randomness $\approx 1/3$, and this is what happens with high probability.

Kolmogorov Extractors

Theorem (Fortnow, Hitchcock, Pavan, Vinodchandran, Wang)

Let $0 < \alpha < \beta < 1$.

There exist

- *polynomial-time functions* $\Gamma_1, \ldots, \Gamma_k : \{0,1\}^* \to \{0,1\}^*$ *and*
- *a linear function* $f : \mathbb{N} \to \mathbb{N}$

such that for every n *and for every* σ *of length* $f(n)$, *if* $C(\sigma) \geq \alpha|\sigma|$ *then for some* i, $\tau = \Gamma_i(\sigma)$ *has length* n *and* $C(\tau) \geq \beta|\tau|$.

Zimand called this a (single-source) Kolmogorov Extractor.

$$\begin{tikzcd}
 & & \sigma \arrow[dll] \arrow[dl] \arrow[d] \arrow[drr] & & & \in \{0,1\}^{f(n)} \\
\Gamma_1(\sigma) & \Gamma_2(\sigma) & \Gamma_3(\sigma) & \cdots & \Gamma_k(\sigma) & \in \{0,1\}^{n}
\end{tikzcd}$$

If σ has rate of randomness $\geq \alpha$, then one of the $\Gamma_i(\sigma)$ should have rate of randomness $\geq \beta$.

Theorem (Zimand, based on Vereshchagin and Vyugin)

Let $0 < \alpha < \beta < 1$.
Suppose there are

- *partial computable functions* $\Gamma_1, \ldots, \Gamma_k : \{0,1\}^* \to \{0,1\}^*$ *and*
- *a linear function* $f : \mathbb{N} \to \mathbb{N}$

as in the previous theorem.

Then

$$\beta \leq 1 - \frac{1 - \alpha}{2k - 1} + o(1).$$

Definition

For $k \geq 1$, let $\mathrm{EXT}(k)$ be the set of pairs of reals (α, β) such that $\alpha, \beta \in [0, 1]$ and for which there exist

- a total one-to-one computable function $f : \mathbb{N} \to \mathbb{N}$,
- k total computable functions $\Gamma_1, \dots, \Gamma_k : \{0,1\}^* \to \{0,1\}^*$, and
- a constant $d \in \mathbb{N}$,

such that

- for all n, and every string σ, if $|\sigma| = f(n)$, then $|\Gamma_i(\sigma)| = n$ for all $i \leq k$, and
- if $C(\sigma) \geq \alpha|\sigma| + d$, then for some i, $C(\Gamma_i(\sigma)) \geq \beta|\Gamma_i(\sigma)| - d$.

What can we say about f?

Suppose that d, f, and (Γ_i) witness that $(\alpha, \beta) \in \mathrm{EXT}(k)$:

- $C(\Gamma_i(\sigma)) \leq C(\sigma) + O(1)$ for all i.
- If σ is such that $|\sigma| = f(n)$ and

  $$C(\sigma) = \alpha f(n) + O(1),$$

  then for some i,

  $$C(\Gamma_i(\sigma)) \geq \beta n - O(1).$$

Putting this all together,

$$\alpha f(n) \geq \beta n - O(1).$$

$$f(n) \geq (\beta/\alpha) n - O(1).$$

Theorem (Bienvenu, Csima, HT)

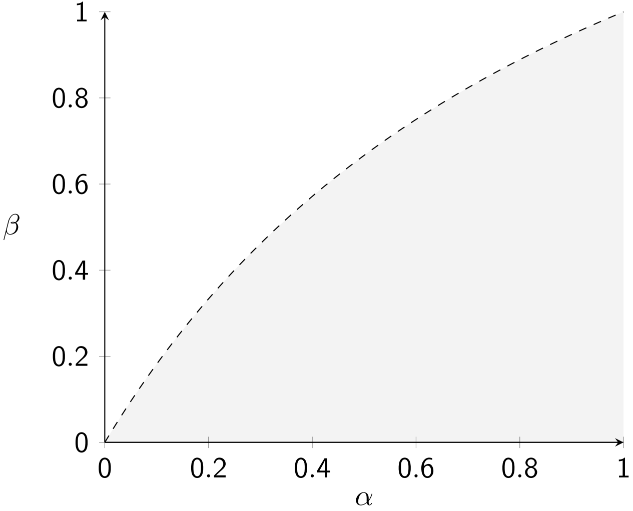
$(\alpha, \beta) \in \mathrm{EXT}(k)$ *if and only if one of the following holds:*

- $k = 1$ *and* $\beta \leq \alpha$, *or*
- $k \geq 2$ *and either* $\alpha = \beta = 0$, $\alpha = \beta = 1$, *or*

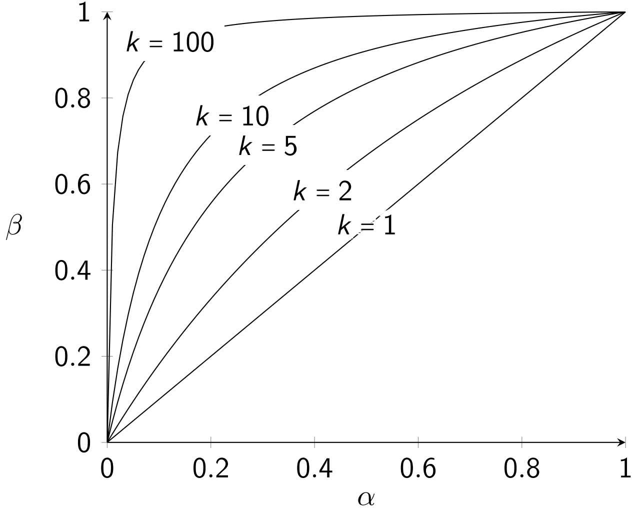
$$\beta < \frac{k\alpha}{1 + (k-1)\alpha}.$$

Moreover, these are witnessed by $f(n) = (\beta/\alpha)n + O(1)$.

$\beta < \frac{k\alpha}{1+(k-1)\alpha}$ for $k = 2$



$\beta = \frac{k\alpha}{1+(k-1)\alpha}$ for various values of k



Definition

For $k \geq 1$, let $\mathrm{EXT}^{\mathrm{p}}(k)$ be the set of pairs of reals (α, β) such that $\alpha, \beta \in [0, 1]$ and for which there exist

- a total one-to-one computable function $f : \mathbb{N} \to \mathbb{N}$,
- k partial computable functions $\Gamma_1, \dots, \Gamma_k : \{0,1\}^* \to \{0,1\}^*$, and
- a constant $d \in \mathbb{N}$,

such that

- for all n, and every string σ, if $|\sigma| = f(n)$, then $|\Gamma_i(\sigma)| = n$ for all $i \leq k$ for which $\Gamma_i(\sigma)$ is defined, and
- if $C(\sigma) \geq \alpha|\sigma| + d$, then for some i, $\Gamma_i(\sigma)$ is defined and $C(\Gamma_i(\sigma)) \geq \beta|\Gamma_i(\sigma)| - d$.

Theorem (Bienvenu, Csima, HT)

$(\alpha, \beta) \in \mathrm{EXT}^{\mathrm{p}}(k)$ *if and only if one of the following holds:*

- $k = 1$ *and* $\alpha \leq \beta$*,*
- $k \geq 2$ *and* $\beta < \frac{k\alpha}{1+(k-1)\alpha}$*, or*
- $k \geq 2$*,* $\beta = \frac{k\alpha}{1+(k-1)\alpha}$*, and* α *and* β *are computable.*

Connections with Graphs

Suppose that d, f, and (Γ_i) witness that $(\alpha, \beta) \in \mathrm{EXT}(k)$.

We can build out of these, for each n, a k-hypergraph $G_n = (V_n, E_n)$ with $f(n)$ edges.

- The vertices V_n are $\{0,1\}^n$.
- The edges E_n are $\{0,1\}^{f(n)}$.
- An edge σ is incident on $\Gamma_1(\sigma), \ldots, \Gamma_n(\sigma)$.

We can also go the other way.

Given, for each n, a k-hypergraph $G_n = (V_n, E_n)$ with 2^n vertices and $2^{f(n)}$ edges:

- Fix a bijection between the vertices V_n and $\{0,1\}^n$.
- Fix a bijection between the edges E_n and $\{0,1\}^{f(n)}$.
- Given $\sigma \in \{0,1\}^{f(n)}$, thinking of σ as an edge, set $\Gamma_1(\sigma), \ldots, \Gamma_k(\sigma)$ to be the k vertices on which σ is incident.

Suppose that d, f, and (Γ_i) witness that $(\alpha, \beta) \in \mathrm{EXT}(k)$. Then:

- for each $\sigma \in \{0,1\}^{f(n)}$ with $C(\sigma) \geq \alpha f(n) + d$, there is i such that $C(\Gamma_i(\sigma)) \geq \beta n + d$.

Think of the corresponding hypergraphs G_n:

- for each edge σ of G_n with $C(\sigma) \geq \alpha f(n) + d$, there is vertex τ incident on σ such that $C(\tau) \geq \beta n + d$.

Or equivalently:

- for each edge σ of G_n, if every vertex τ incident on σ has $C(\tau) < \beta n + d$, then $C(\sigma) < \alpha f(n) + d$.

Think dynamically. We will play the part of a machine M assigning descriptions ρ to strings τ by setting $M(\rho) = \tau$.

There is a constant c such that

$$C(\tau) = C_U(\tau) \leq C_M(\tau) + c.$$

So when we assign a short M-description to τ, this makes $C_M(\tau)$ small, which guarantees that $C(\tau)$ is small.

Our restriction is that we cannot assign the same description to two different strings σ. We have a restricted number of descriptions of each size: 1 description of size 0, 2 descriptions of size 1, 4 descriptions of size 2, and so on.

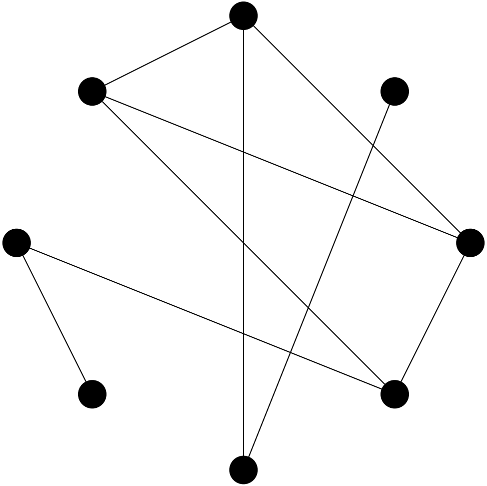
We think of ourselves as giving short descriptions to vertices $\tau \in \{0,1\}^*$ (the outputs of the Γ_i).

Recall:

- for each edge σ of G_n, if every vertex τ incident on σ has $C(\tau) < \beta n + d$, then $C(\sigma) < \alpha f(n) + d$.

So whenever we give a short description (size $< \alpha f(n)$) of *every* vertex of an edge σ, the universal machine must give a short description (length $< \beta n$ to σ).

$k = 2, \ n = 8.$



Given a set U of vertices, let $E(U)$ be the number of hyperedges contained entirely inside U.

If we assign short descriptions to all the vertices in a set U, the universal machine must assign short descriptions to all of the edges in $E(U)$.

We are limited in the number of short descriptions we have to assign to edges, and the universal machine is limited in the number of short descriptions it can assign to vertices.

- The bigger α is, the more short descriptions we have available.
- The bigger β is, the more short descriptions the universal machine has available.

The more edges in $E(U)$, the harder it is for the universal machine; if there are too many edges, then the universal machine runs out of short descriptions.

After counting precisely, we get:

Theorem

Fix $k \geq 2$ and let (α, β) be a pair of computable reals in $[0, 1]$. The following are equivalent

(a) $(\alpha, \beta) \in \mathrm{EXT}(k)$

(b) *There is a constant d and computable function f with*

$$f(n) \geq (\beta/\alpha)n - O(1)$$

and such that for all n there is a k-hypergraph G_n with:

- 2^n *vertices,*
- $2^{f(n)}$ *hyperedges, and*
- *for every $U \subseteq G_n$ with $|U| \leq 2^{\beta n - d}$, $|E(U)| < 2^{\alpha f(n) + d}$.*

Edge Distribution in Graphs

We have reduced the question to asking which (α, β) have a constant d and computable function f with

$$f(n) \geq (\beta/\alpha)n - O(1)$$

and such that for all n there is a k-hypergraph G_n with:

- 2^n vertices,
- $2^{f(n)}$ hyperedges, and
- for every $U \subseteq G_n$ with $|U| \leq 2^{\beta n - d}$, $|E(U)| < 2^{\alpha f(n) + d}$.

In a graph, the edge density is the ratio of edges to potential edges.

Definition

Let $G = (V, E)$ be a k-hypergraph. The edge density p of G is

$$p = \frac{|E|}{\binom{|V|}{k}}.$$

If the edges were completely evenly distributed, we would expect a set of vertices U to contain about $p\binom{|U|}{k}$ edges.

But this is not always the case! Any graph will have some sets U which have slightly more vertices.

Ramsey's theorem says that if U is very small, then U could contain every possible edge, or no edges at all.

Erös, Goldberg, Pach, and Spencer study the case when U is about half the vertices of the graph.

Our case is when $|U| = |V|^{\alpha}$.

In our case, we can get them to be very evenly-distributed.

1. It is always possible to find a set U with slightly more than the expected number of edges.
2. There are graphs where this is not too much more.

The main tool for solving these problems is the probabilistic method.

Theorem (Bienvenu, Csima, HT)

$(\alpha, \beta) \in \mathrm{EXT}(k)$ *if and only if one of the following holds:*

- $k = 1$ *and* $\beta \leq \alpha$, *or*
- $k \geq 2$ *and either* $\alpha = \beta = 0$, $\alpha = \beta = 1$, *or*

$$\beta < \frac{k\alpha}{1 + (k-1)\alpha}.$$

Theorem (Bienvenu, Csima, HT)

$(\alpha, \beta) \in \mathrm{EXT}^{\mathrm{p}}(k)$ *if and only if one of the following holds:*

- $k = 1$ *and* $\alpha \leq \beta$,
- $k \geq 2$ *and* $\beta < \frac{k\alpha}{1+(k-1)\alpha}$, *or*
- $k \geq 2$, $\beta = \frac{k\alpha}{1+(k-1)\alpha}$, *and* α *and* β *are computable.*