Regular Hilberg Processes: Nonexistence of Universal Redundancy Ratios

Łukasz Dębowski
ldebowsk@ipipan.waw.pl

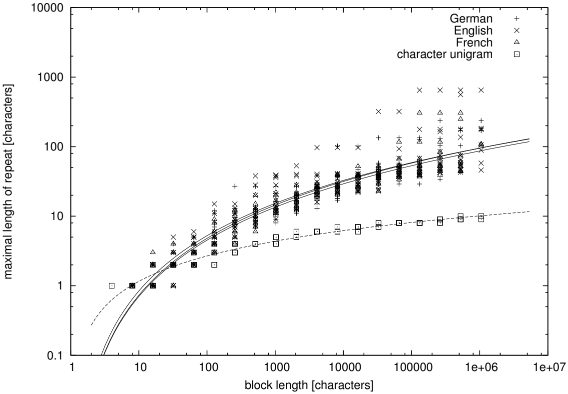
Institute of Computer Science
Polish Academy of Sciences

WITMSE 2015, København

Experimental data for natural language (Dębowski 2015)

Maximal repetition:

$$L(\xi_{1:k}) := \max\{m : x_{1:m} \text{ is repeated in } \xi_{1:k}\}.$$



$$L(\xi_{1:m}) \propto (\log m)^{\alpha}, \qquad \alpha \approx 2.7$$

Maximal repetition and topological entropy

Maximal repetition:

$$\mathbf{L}(\xi_{1:k}) := \max \{ m : x_{1:m} \text{ is repeated in } \xi_{1:k} \} .$$

Topological entropy:

$$\mathbf{H}_{top}(m|\xi_{1:\infty}) := \log \operatorname{card} \{ x_{1:m} : x_{1:m} \text{ is a subsequence of } \xi_{1:\infty} \} .$$

Theorem

If $\mathbf{H}_{top}(m|\xi_{1:\infty}) < \log(k - m + 1)$ then $\mathbf{L}(\xi_{1:k}) \geq m$.

In particular:

$$\mathbf{H}_{top}(m|\xi_{1:\infty}) = O\left(m^{\beta}\right) \implies \mathbf{L}(\xi_{1:m}) = \Omega\left((\log m)^{1/\beta}\right),$$

$$\mathbf{L}(\xi_{1:m}) = O\left((\log m)^{1/\beta}\right) \implies \mathbf{H}_{top}(m|\xi_{1:\infty}) = \Omega\left(m^{\beta}\right).$$

Regular Hilberg processes

We have a hypothesis that for texts in natural language

$$L(\xi_{1:m}) = \Theta\left((\log m)^{1/\beta}\right), \quad (1)$$

$$H_{top}(m|\xi_{1:\infty}) = \Theta\left(m^{\beta}\right), \quad (2)$$

where $\beta \approx 0.37$ (Hilberg 1990, Dębowski 2015).

Definition

A stationary measure μ is called a regular Hilberg process with an exponent $\beta \in (0,1)$ if it satisfies conditions (1)–(2) μ-almost surely, where the lower bound for $L(\xi_{1:m})$ and the upper bound for $H_{top}(m|\xi_{1:\infty})$ are uniform in $\xi_{1:\infty}$.

Regular Hilberg processes have zero entropy rate

The block entropy of measure μ is

$$H_\mu(m) := \mathbb{E}_\mu \left[-\log \mu(\xi_{1:m}) \right],$$

and the entropy rate of μ is the limit

$$h_\mu := \inf_{m \in \mathbb{N}} \frac{H_\mu(m)}{m} = \lim_{m \to \infty} \frac{H_\mu(m)}{m}.$$

Regular Hilberg processes have the entropy rate $h_\mu = 0$:

For the random ergodic measure $F = \mu(\cdot|\mathcal{I})$, where $\mathcal{I}$ is the shift-invariant algebra, by the ergodic theorem, we have

$$H_F(m) \le H_{top}(m|\xi_{1:\infty})$$

μ-almost surely, so $h_F = 0$, whereas we have

$$h_\mu = \mathbb{E}_\mu h_F,$$

from which $h_\mu = 0$ follows.

Towards nonexistence of universal redundancy ratios

Ergodic regular Hilberg processes have this property:

We have $\mu = F = \mu(\cdot|\mathcal{I})$, so the block entropy satisfies

$$H_\mu(m) = H_F(m) \le H_{top}(m|\xi_{1:\infty}) = O\left(m^\beta\right),$$

whereas the length of the Lempel-Ziv code μ-almost surely satisfies

$$|C(\xi_{1:m})| \ge \frac{m}{L(\xi_{1:m}) + 1} \log \frac{m}{L(\xi_{1:m}) + 1} = \Omega\left(\frac{m}{(\log m)^{1/\beta - 1}}\right).$$

In other words the length of the universal LZ code is orders of magnitude larger than the block entropy of the process!

We cannot estimate block entropy as length of the LZ code!

RHA processes, part I: Random selection of blocks

Let integers $(k_n)_{n\in\{0\}\cup\mathbb{N}}$, which we will call perplexities, satisfy

$$0 < k_{n-1} \le k_n \le k_{n-1}^2.$$

Next, for each $n \in \mathbb{N}$, let $(L_{nj}, R_{nj})_{j\in\{1,\dots,k_n\}}$ be an independent random combination of k_n pairs of numbers from the set $\{1,\dots,k_{n-1}\}$ drawn without repetition. That is,

$$P((L_{n1}, R_{n1}, \dots, L_{nk_n}, R_{nk_n}) = (l_{n1}, r_{n1}, \dots, l_{nk_n}, r_{nk_n})) = \binom{k_{n-1}^2}{k_n}^{-1}.$$

Subsequently we define random variables

$$Y_j^0 = j, \qquad j \in \{1,\dots,k_0\},$$

$$Y_j^n = Y_{L_{nj}}^{n-1} \times Y_{R_{nj}}^{n-1}, \qquad j \in \{1,\dots,k_n\}, n \in \mathbb{N},$$

where $a \times b$ denotes concatenation.

Hence Y_j^n are k_n distinct random blocks of 2^n numbers.

RHA processes, part II: The nonstationary process

Let $(C_n)_{n\in\{0\}\cup\mathbb{N}}$ be a sequence of independent random variables with uniform distribution

$$P(C_n = j) = 1/k_n, \qquad j \in \{1, ..., k_n\}. \tag{3}$$

Definition

The random hierarchical association (RHA) process $\mathcal{X}$ with perplexities $(k_n)_{n\in\{0\}\cup\mathbb{N}}$ is defined as

$$\mathcal{X} = Y^0_{C_0} \times Y^1_{C_1} \times Y^2_{C_2} \times \dots . \tag{4}$$

Sequence $\mathcal{X}$ will be parsed into a sequence of numbers X_j, where

$$\mathcal{X} = X_1 \times X_2 \times X_3 \times ... \tag{5}$$

RHA processes, part III: Stationary mean

A measure ν is called asymptotically mean stationary with respect to blocks (AMSB) if limits

$$\mu(x_{1:m}) := \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \nu(\xi_{i:i+m-1} = x_{1:m})$$

exist for every block $x_{1:m}$.

Theorem

The RHA process is AMSB. In particular, for $m \le 2^n$ and $k \in \mathbb{N}$, the stationary mean is

$$\mu(x_{1:m}) = \frac{1}{2^n} \sum_{j=0}^{2^n-1} P(X_{k2^n+j:k2^n+j+m-1} = x_{1:m}).$$

Main result

Theorem

For perplexities $k_n = \left\lfloor \exp\left(2^{\beta n}\right) \right\rfloor$, *where* $\beta \in (0, 1)$, *the stationary mean* μ *of the RHA process is nonergodic and satisfies:*

1. *The block entropy is sandwiched by*

$$C_1 m \left(\frac{1}{\log m}\right)^{1/\beta - 1} \le H_\mu(m) \le C_2 m \left(\frac{\log\log m}{\log m}\right)^{1/\beta - 1}.$$

2. *Measure* μ *is a regular Hilberg process, i.e.,*

$$L(\xi_{1:m}) = \Theta\left((\log m)^{1/\beta}\right),$$
$$H_{top}(m|\xi_{1:\infty}) = \Theta\left(m^\beta\right)$$

μ*-almost surely, where the lower bound for* $L(\xi_{1:m})$ *and the upper bound for* $H_{top}(m|\xi_{1:\infty})$ *are uniform in* $\xi_{1:\infty}$.

Conclusions: Information theory

1. Shields (1993) showed that for any uniquely decodable code C and any function $\rho(m) = o(m)$ there exists such an ergodic source F that

$$\limsup_{m\to\infty}[\mathbb{E}_F |C(\xi_{1:m})| - H_F(m) - \rho(m)] > 0.$$

2. Whereas Shields' result concerns nonexistence of a universal bound for $\mathbb{E}_F |C(\xi_{1:m})| - H_F(m)$, ours indicates nonexistence of a universal bound for $\mathbb{E}_F |C(\xi_{1:m})| / H_F(m)$.
3. For the RHA process and any uniquely decodable code C,

$$\mathbb{E}_\mu \frac{\mathbb{E}_F |C(\xi_{1:m})|}{H_F(m)} \ge \mathbb{E}_\mu \frac{H_\mu(m)}{H_{top}(m|\xi_{1:\infty})} = \Omega\left(\frac{m^{1-\beta}}{(\log m)^{1/\beta-1}}\right).$$

Conclusions: Linguistics

1. We have shown that regular Hilberg processes arise in a very simple setting of random sampling of texts from a restricted random hierarchical pool.
2. The pool of texts in natural language need not be so random.
3. Consequently, this might explain why the estimates of block entropy for natural language, obtained through text prediction experiments with human subjects, suggest $H_\mu(m) = \Theta(m^\beta)$ rather than $H_\mu(m) = \Omega(m/(\log m)^{1/\beta-1})$.

`www.ipipan.waw.pl/~ldebowsk`