Information Theory and Power Laws

for Language Models and Stochastic Processes

Łukasz Dębowski
ldebowsk@ipipan.waw.pl

Institute of Computer Science
Polish Academy of Sciences

Agents, AI & Alignment Workshop, Oxford, March 3–9, 2024

The dawn of large language models

- BERT, GPT, LaMDA, Palm, ... — statistical language models based on transformers and trained on internet-sized corpora...
- ... excel at solving traditionally understood tasks in natural language processing (machine translation, etc.),
- ... can also generate quite arbitrary texts on demand:
  - rhymed poems,
  - school essays,
  - simple computer programs,
  - formalized proofs for automated theorem provers.

Euclid's proof

`https://arxiv.org/abs/2303.12712`

Can you write a proof that there are infinitely many primes,
with every line that rhymes?

Yes, I think I can,
Though it might take a clever plan.
I'll start by noting Euclid's proof,
Which shows that primes aren't just aloof.
Assume we have a finite list,
Of primes, and that none have been missed.
Multiply them all together,
And add one, just to be clever.
The result will be a number,
That has no prime factors, oh wonder!
But every number has a prime,
That divides it, it's just a matter of time.
So we've found a contradiction,
And our finite list needs eviction.
There must be infinitely many primes,
And that's the end of my rhyming lines.

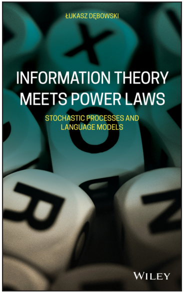
How come and what next?

- Large language models appeared quite suddenly...
- ... made a huge progress within a few years,
- ... exhibit curious emergent behaviors.
- You may chat with them and get a feeling that you converse with something kind of human...
- **Not exactly! They prefer to sound likely rather than intellectually honest. They hallucinate!**

Besides programming, we need theoretical insight:
neuroscience, linguistics, mathematics, philosophy, physics, ...

My own interests in abstract language models

- I did my masters (1999) in theoretical physics (statistical mechanics).
- Later I worked in statistical natural language processing (visiting Fred Jelinek during his sabbatical in Prague in 2001, part-of-speech tagging of the IPI PAN corpus of Polish).
- But my heart was taken by power laws and information theory (Zipf's law, Hilberg's hypothesis, refutation of finite-state models, visiting Jim Crutchfield in Santa Fe Institute in 2002).
- I did my PhD (2005) in information theory and Gaussian processes with long memory, then worked with Peter Grünwald and Peter Harremoës in CWI.
- Ever since then I have been working on mathematical foundations of statistical language modeling (measure theory, ergodic decomposition, excess entropy, Kolmogorov complexity, universal coding and universal prediction).
- **Quite a lot of apparently abstract math...**



Outline of this lecture

1. **Power laws:**
   - Neural scaling law vs. Zipf's and Heaps' laws.
2. **Information theory:**
   - Shannon entropy and Kolmogorov complexity.
   - Hilberg's hypothesis, entropy rate, and excess entropy.
3. **Santa Fe processes:**
   - Knowledge-narration decomposition.
   - IID and multiperiodic narration.

---

4. **Universal coding:**
   - PPM and grammar-based universal codes.
   - Vocabulary growth and Hilberg exponents.
5. **Ergodic decomposition:**
   - Decomposition of excess entropy.
   - Knowledge growth and Hilberg exponents.
6. **Theoretical challenges:**
   - Stretched exponential growth of repetition time.
   - Power-law decay of word embedding correlations.

Language models — Cross entropy

Let us write text $(x_1, x_2, \dots, x_T)$ as x_1^T.

A language model is a (probability) measure on tokens:

$$Q(x_t|x_{t-M}^{t-1}) \geq 0, \quad \sum_{x_t} Q(x_t|x_{t-M}^{t-1}) = 1.$$

The cross entropy of the model is the mean minus log-probability:

$$\mathcal{H}(Q) := -\frac{1}{T}\sum_{t=1}^{T} \log Q(x_t|x_{t-M}^{t-1}) \geq 0.$$

$\mathcal{H}(Q)$ is the average surprisal of model Q on text x_1^T.

We seek for Q that is a computable function of training data x_1^T and minimizes cross entropy on different data, called the test data.

Language models — Embeddings and transformers

In language models based on transformers, probabilities $Q(x_t|x_{t-M}^{t-1})$ are computed by stacking two mechanisms:

- embeddings — vectors x_t corresponding to words/concepts,
- attention — a nonlinear operation on embeddings

$$y_t = \sum_{s=t-M}^{t-1} \frac{\exp(x_t \cdot x_s)}{\sum_{r=t-M}^{t-1} \exp(x_t \cdot x_r)} x_s.$$

The GPT-3 language model:

- Number of parameters: $N = 175$ billions (800 GB RAM).
- Context length: $M = 2048$ words.
- Training data: Common Crawl (410 bln, 60%), WebText2 (19 bln, 22%), books (67 bln, 16%), Wikipedia (3 bln, 3%).

Language models — Neural scaling law

$Q_{N,T}$ — neural model with N parameters trained on T tokens.

$\mathcal{H}(Q)$ — cross entropy of Q on the test data.

Kaplan et al. (2020) observed empirically that

$$\mathcal{H}(Q_{N,T}) \approx \left(\frac{N_0}{N}\right)^{\gamma_N} + \left(\frac{T_0}{T}\right)^{\gamma_T}$$

for $N_0 = 6.4 \times 10^{13}$, $T_0 = 1.8 \times 10^{13}$, $\gamma_N = 0.076$, $\gamma_T = 0.103$.

The more data and parameters, the better is the model:

$$\mathcal{H}(Q_{\infty,T}) \approx \left(\frac{T_0}{T}\right)^{\gamma_T}, \quad \mathcal{H}(Q_{N,\infty}) \approx \left(\frac{N_0}{N}\right)^{\gamma_N}, \quad \mathcal{H}(Q_{\infty,\infty}) \approx 0.$$

For each T there is roughly an optimal $N = N_0(T/T_0)^{\gamma_T/\gamma_N}$.

Zipf-Mandelbrot's and Herdan-Heaps' laws

Shakespeare's
First Folio/35 Plays:

rank $r(w)$	freq $f(w)$	word w
1	21557	I
2	19059	and
3	16571	to
4	14921	of
5	14491	a
6	12077	my
7	10463	you
8	9789	in
9	8754	is
...	...	...

Numbers of tokens and types:

$$N = \sum_w f(w), \quad V = \sum_w 1.$$

Zipf-Mandelbrot's law:

$$r(w) \approx \frac{V}{f(w)^\beta}, \quad \beta \in (0,1).$$

Herdan-Heaps' law:

$$V \propto N^\beta, \quad \beta \in (0,1).$$

[Put $r(w) = 1$ and $f(w) \propto N$.]

Is there a link between the neural scaling law and Zipf's law?

Shannon entropy and Kolmogorov complexity

The Shannon entropy of a random variable W is

$$H(W) := \mathbb{E}\left(-\log_2 P(W)\right) = -\sum_w p(w)\log_2 p(w).$$

The (prefix-free) Kolmogorov complexity of a string w is

$$C(w) := \min\left\{|x| : x \in \{0,1\}^*, \mathcal{U}(x) = w\right\}.$$

We also define the mutual information

$$\begin{aligned} I(W;Z) &:= H(W) + H(Z) - H(W,Z), && \text{(Shannon)} \\ J(w;z) &:= C(w) + C(z) - C(w,z). && \text{(algorithmic)} \end{aligned}$$

The source coding inequality links these quantities,

$$0 \le \mathbb{E}\, C(W) - H(W) \le C(p), \quad \mathbb{E}\, C(W) = \sum_w p(w)C(w).$$

We have $C(p) < \infty$ iff distribution p is computable.

Fair-coin process and algorithmically random sequences

The entropy and the Kolmogorov complexity sometimes coincide:

1. A fair-coin process $(Z_k)_{k\in\mathbb{N}}$ is a sequence of independent uniformly distributed binary random variables:

$$P(Z_1^k = z_1^k) = 2^{-k}, \qquad \mathbb{E}\, C(Z_1^k) \approx H(Z_1^k) = k.$$

2. An algorithmically random sequence $(z_k)_{k\in\mathbb{N}}$ is a fixed binary sequence that has the maximal Kolmogorov complexity:

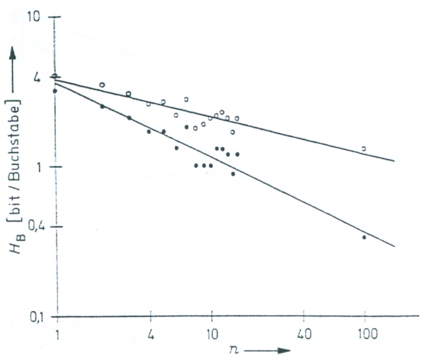
$$C(z_1^k) \geq k - c.$$

For the fair-coin process $(Z_k)_{k\in\mathbb{N}}$, almost every realization is algorithmically random,

$$P((Z_k)_{k\in\mathbb{N}} \text{ is algorithmically random}) = 1.$$

Hilberg's plot of Shannon's data for English

In 1990, German telecommunication engineer Wolfgang Hilberg published a claim that $H(X_1^n) \propto \sqrt{n}$ holds for Claude Shannon's guessing data from 1951.



$$H(X_n|X_1^{n-1}) = H(X_1^n) - H(X_1^{n-1}) \propto \frac{1}{\sqrt{n}}, \quad n \leq 100$$

Entropy rate and excess entropy

Let $(X_i)_{i\in\mathbb{Z}}$ be a discrete stationary process, namely

$$P(X_{t+1}^{t+k} = x_1^k) = p(x_1^k) \text{ for all } t \in \mathbb{Z}.$$

The entropy rate is the limit

$$h := \lim_{n\to\infty} \frac{H(X_1^n)}{n} = \lim_{n\to\infty} \frac{\mathbb{E}\, C(X_1^n)}{n}.$$

The excess entropy is the limit

$$\begin{aligned} E &:= \lim_{n\to\infty} \left(H(X_1^n) - nh\right) = \lim_{n\to\infty} I(X_1^n; X_{n+1}^{2n}) \\ &\le \limsup_{n\to\infty} \left(\mathbb{E}\, C(X_1^n) - nh\right) \le \limsup_{n\to\infty} \mathbb{E}\, J(X_1^n; X_{n+1}^{2n}). \end{aligned}$$

Hilberg exponent of a sequence

To measure power-law growth, we introduce Hilberg exponent

$$\operatorname*{hilb}_{n\to\infty} S(n) := \left[\limsup_{n\to\infty} \frac{\log S(n)}{\log n}\right]_+ .$$

In particular, we obtain

$$\operatorname*{hilb}_{n\to\infty} n^\beta = \beta \text{ if } \beta \geq 0.$$

Theorem

If $\lim_{n\to\infty} S(n)/n = s$ *then*

$$\operatorname*{hilb}_{n\to\infty} (S(n) - ns) \leq \operatorname*{hilb}_{n\to\infty} (2S(n) - S(2n)) .$$

with an equality for $S(n) \geq ns$.

Hilberg's law

For a stationary process, we have two distincts exponents:

$$\beta_P := \operatorname*{hilb}_{n\to\infty} \left(H(X_1^n) - nh\right) = \operatorname*{hilb}_{n\to\infty} I(X_1^n; X_{n+1}^{2n}) \in [0,1]$$

$$|\wedge$$

$$\beta_C := \operatorname*{hilb}_{n\to\infty} \left(\mathbb{E}\, C(X_1^n) - nh\right) = \operatorname*{hilb}_{n\to\infty} \mathbb{E}\, J(X_1^n; X_{n+1}^{2n}) \in [0,1]$$

Exponents β_P and β_C can be different if p is uncomputable.

Relationship $\beta_C > 0$ will be called Hilberg's law.

K-state D-symbol unifilar processes satisfy $\beta_P = \beta_C = 0$ since $I(X_1^n; X_{n+1}^{2n}) \le \log K$ and $\mathbb{E}\, J(X_1^n; X_{n+1}^{2n}) \lesssim 2DK \log n$.

We will construct some simple processes that enjoy $\beta_C > 0$.

Abstract semantics — Knowledge

In our approach, the knowledge is a sequence of binary digits that describe a model of reality that is referred to by texts.

Consider a row of cinema chairs that are vacant or occupied:

The state of this row can be described by a collection of facts $k \mapsto Z_k$, indexed by numbers $k = 1, 2, 3, \dots$, where

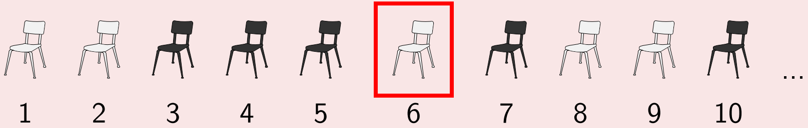
$$Z_k := \begin{cases} 0 & \text{if } k\text{-th chair is vacant,} \\ 1 & \text{if } k\text{-th chair is occupied.} \end{cases}$$

Mapping $\mathbb{N} \ni k \mapsto Z_k \in \{0, 1\}$ will be called the knowledge.

Abstract semantics — Narration

By contrast, the narration is a process of selecting which facts are described at a particular position of a text.

Suppose that in the 5-th proposition of a phone call with a friend, we are communicating that the 6-th chair is vacant:



The process of selecting facts can be described by a sequence of topics $i \mapsto K_i$, indexed by numbers $i = \ldots, -1, 0, 1, \ldots$.
Here, we posit that the 5-th topic is 6 and the 6-th fact is 0,

$$K_5 = 6 \text{ and } Z_6 = 0.$$

Mapping $\mathbb{Z} \ni i \mapsto K_i \in \mathbb{N}$ will be called the narration.

Santa Fe process — A logically consistent text (2002)

- The knowledge $(Z_k)_{k\in\mathbb{N}}$ is a collection of facts (bits).
- The narration $(K_i)_{i\in\mathbb{Z}}$ is a sequence of topics (numbers).
- The text $(X_i)_{i\in\mathbb{Z}}$ is a sequence of propositions (pairs):

$$X_i := (K_i, Z_{K_i}).$$

A semantic interpretation

Process $(X_i)_{i\in\mathbb{Z}}$ is a sequence of random propositions consistently describing knowledge $(Z_k)_{k\in\mathbb{N}}$:

- Proposition $X_i = (k, z)$ asserts that the k-th chair in the row has state z, in such way that one can determine both k and z.
- For $X_i = (k, z)$ and $X_j = (k', z')$ we do not know in advance which chairs they describe but $k = k' \implies z = z'$.

Sufficient conditions for Hilberg's law

Suppose that knowledge $(Z_k)_{k\in\mathbb{N}}$ is algorithmically random when sampled by narration $(K_i)_{i\in\mathbb{Z}}$, i.e., for $X_i = (K_i, Z_{K_i})$, we have

$$C(\{X_1, ..., X_n\} \,|K_1^n) \geq \#\{K_1, ..., K_n\} - c.$$

By the chain rule, we obtain:

$$C(X_1^n) \approx C(K_1^n) + C(\{X_1^n\} \,|K_1^n) \approx C(K_1^n) + \#\{K_1^n\}.$$

As a result, whenever narration $(K_i)_{i\in\mathbb{Z}}$ is stationary then

$$\lim_{n\to\infty} \frac{\mathbb{E}\,\#\{K_1^n\}}{n} = 0,$$

$$\beta_C := \operatorname*{hilb}_{n\to\infty} \left(\mathbb{E}\, C(X_1^n) - nh\right) \geq \operatorname*{hilb}_{n\to\infty} \mathbb{E}\,\#\{K_1^n\}$$

with an equality if

$$\operatorname*{hilb}_{n\to\infty} \left(\mathbb{E}\, C(K_1^n) - nh\right) \leq \operatorname*{hilb}_{n\to\infty} \mathbb{E}\,\#\{K_1^n\}.$$

IID and multiperiodic narration

As for the narration, we have two simple choices:

1. **An IID Zipfian process:**

$$P(K_i = k) := \frac{k^{-\alpha}}{\zeta(\alpha)}, \quad \zeta(\alpha) := \sum_{k=1}^{\infty} k^{-\alpha}, \quad \alpha > 1$$

$$h > 0, \quad \mathbb{E}\, C(K_1^n) \approx nh, \quad \operatorname*{hilb}_{n\to\infty} \mathbb{E}\, \#\left\{K_1^n\right\} = \frac{1}{\alpha}$$

2. **A deterministic multiperiodic process:**

If we delete $K_i < r$, value r appears every $\lceil 1 + cr \rceil$ positions.

For example, for $c = 1$ we obtain:

$$(K_i)_{i\in\mathbb{Z}} = (..., 1, 2, 1, 3, 1, 4, 1, 2, 1, 5, 1, 6, 1, 2, 1, 3, 1, ...)$$

$$h = 0, \quad \operatorname*{hilb}_{n\to\infty} \mathbb{E}\, C(K_1^n) \le \operatorname*{hilb}_{n\to\infty} \mathbb{E}\, \#\left\{K_1^n\right\} = \frac{c}{c+1}$$

Partial recapitulation

1. We presented Santa Fe processes that exhibit Hilberg's law — power-law growth of algorithmic mutual information.
2. These processes are motivated by an abstract semantic model which decomposes a text into knowledge and narration.
3. Hilberg's law is implied by Zipf's law for knowledge.
4. It is a matter of further research whether similar ideas can be applied to natural language and neural language models.
5. Anyway, it seems quite unsurprising that excess entropy of natural language may be very large — potentially unbounded — like the number of distinct words in a given language.

We will show that Hilberg's law implies Zipf's law for words.

The main result of the second part

Theorem about facts and words:

The number of distinct words in a finite text is roughly greater than the number of independent facts described by the text.

- The above proposition is a general result in information theory connected to Hilberg's and Zipf's laws.
- It's an impossibility result that pertains to a general stationary communication system.
- This result is paradoxical since we might think that combining words we may express many more independent facts.
- The paradox is less surprising if we realize that repeated facts are expressed via fixed phrases.

Universal codes — efficient data compression

A computable prefix-free code $B : \mathbb{X}^* \to \{0,1\}^*$ is called universal if for every stationary ergodic process $(X_i)_{i\in\mathbb{Z}}$, we have

$$\lim_{n\to\infty} \frac{\left|B(X_1^n)\right|}{n} = h \text{ almost surely.}$$

There are many different universal codes:

- Lempel-Ziv code,
- prediction by partial matching (PPM),
- grammar-based codes.

Since $|B(w)| \geq C(w) - C(B)$ then

$$\operatorname*{hilb}_{n\to\infty} \left(\mathbb{E}\left|B(X_1^n)\right| - nh\right) = \operatorname*{hilb}_{n\to\infty} \mathbb{E}\, J_B(X_1^n; X_{n+1}^{2n}) \geq \beta_C \geq \beta_P,$$

where $J_B(w;z) := |B(w)| + |B(z)| - |B(wz)|$.

A context-free grammar that generates one text

$$
\begin{aligned}
\boldsymbol{A}_1 &\rightarrow \boldsymbol{A}_2\boldsymbol{A}_2\boldsymbol{A}_4\boldsymbol{A}_5\textbf{dear\_children}\boldsymbol{A}_5\boldsymbol{A}_3\textbf{all.}\\
\boldsymbol{A}_2 &\rightarrow \boldsymbol{A}_3\textbf{you}\boldsymbol{A}_5\\
\boldsymbol{A}_3 &\rightarrow \boldsymbol{A}_4\textbf{\_to\_}\\
\boldsymbol{A}_4 &\rightarrow \textbf{Good\_morning}\\
\boldsymbol{A}_5 &\rightarrow \textbf{,\_}
\end{aligned}
$$

Good morning to you,
Good morning to you,
Good morning, dear children,
Good morning to all.

Minimal grammar-based codes

Grammar-based coding:

- a grammar transform $\Gamma : \mathbb{X}^* \to \mathcal{G}$ for each string $\boldsymbol{w} \in \mathbb{X}^*$ returns a grammar $\Gamma(\boldsymbol{w})$ that generates this string.
- a grammar encoder $\phi : \mathcal{G} \to \{0,1\}^*$ encodes the grammar.
- Transform Γ is called minimal if $|\phi(\Gamma(\boldsymbol{w}))| \le |\phi(\boldsymbol{G})|$ for any string $\boldsymbol{w}$ and any grammar $\boldsymbol{G}$ that generates $\boldsymbol{w}$.

Vocabulary bound of mutual information:

For local ϕ, minimal Γ, grammar-based code $\boldsymbol{B}(\boldsymbol{w}) := \phi(\Gamma(\boldsymbol{w}))$, and $\boldsymbol{L}(\boldsymbol{w})$ being the maximal repetition length:

$$\boldsymbol{J_B}(\boldsymbol{w};\boldsymbol{z}) \le \boldsymbol{c}\,\#\,\boldsymbol{V}(\Gamma(\boldsymbol{wz}))\boldsymbol{L}(\boldsymbol{wz}).$$

where $\boldsymbol{V}(\boldsymbol{G})$ is the set of nonterminals in grammar $\boldsymbol{G}$.

Some minimal grammar transforms are NP-hard to compute.

Markov order estimators

For a stationary process $(X_i)_{i\in\mathbb{Z}}$, the Markov order is

$$M := \inf\left\{k \geq 0 : P(X_{k+1}^n|X_1^k) = \prod_{i=k+1}^{n} P(X_i|X_{i-k}^{i-1})\right\}.$$

Function $\mathbb{M} : \mathbb{X}^* \to \mathbb{N}$ is called a consistent estimator of M if

$$\lim_{n\to\infty} \mathbb{M}(X_1^n) = M \text{ almost surely}$$

for any stationary ergodic process $(X_i)_{i\in\mathbb{Z}}$.

PPM Markov order

Empirical frequency:

$$\#(w_1^k|x_1^n) := \sum_{i=0}^{n-k} 1\left\{x_{i+1}^{i+k} = w_1^k\right\}.$$

Maximum likelihood and PPM distributions:

$$\mathrm{ML}_k(x_1^n) := \prod_{i=k+1}^{n} \frac{\#(x_{i-k}^{i}|x_1^n)}{\#(x_{i-k}^{i-1}|x_1^{n-1})}, \quad k \ge 0,$$

$$\mathrm{PPM}_k(x_1^n) := D^{-k} \prod_{i=k+1}^{n} \frac{\#(x_{i-k}^{i}|x_1^{i-1}) + 1}{\#(x_{i-k}^{i-1}|x_1^{i-2}) + D}, \quad k \ge 0,$$

$$\mathrm{PPM}(x_1^n) := \sum_{k=0}^{\infty} w_k \,\mathrm{PPM}_k(x_1^n), \quad w_k := \frac{1}{k+1} - \frac{1}{k+2}.$$

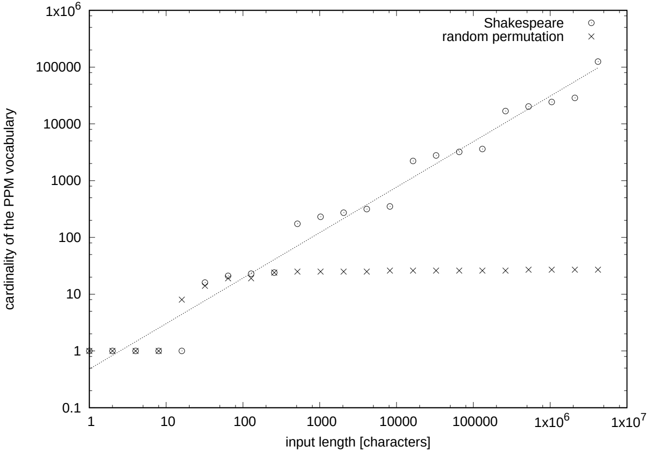
Some consistent estimator of M is the **PPM Markov order**:

$$\mathbb{M}(x_1^n) := \min\left\{k \ge 0 : \mathrm{ML}_k(x_1^n) \ge w_n \,\mathrm{PPM}(x_1^n)\right\}.$$

Heaps' law for the PPM vocabulary

Empirical vocabulary: $V_k(x_1^n) := \left\{ x_{t+1}^{t+k} : 0 \leq t \leq n-k \right\}.$

PPM vocabulary: $V_{\mathbb{M}}(x_1^n) := V_{\mathbb{M}(x_1^n)}(x_1^n).$



Vocabulary growth and Hilberg exponents

The Shannon-Fano code w.r.t. the PPM distribution is universal.

It has length $|\boldsymbol{B}(\boldsymbol{w})| \approx -\log \mathrm{PPM}(\boldsymbol{w})$.

Moreover the respective mutual information is bounded by

$$\boldsymbol{J}_{\boldsymbol{B}}(\boldsymbol{w};\boldsymbol{z}) \le \boldsymbol{c}\,\#\,\boldsymbol{V}_{\mathbb{M}}(\boldsymbol{wz}) \log |\boldsymbol{wz}|\,.$$

Hence, the number of PPM words bounds the Hilberg exponent

$$\operatorname*{hilb}_{\boldsymbol{n}\to\infty} \mathbb{E}\,\#\,\boldsymbol{V}_{\mathbb{M}}(\boldsymbol{X}_1^{\boldsymbol{n}}) \ge \operatorname*{hilb}_{\boldsymbol{n}\to\infty} \mathbb{E}\,\boldsymbol{J}_{\boldsymbol{B}}(\boldsymbol{X}_1^{\boldsymbol{n}};\boldsymbol{X}_{\boldsymbol{n}+1}^{2\boldsymbol{n}}) \ge \beta_{\boldsymbol{C}} \ge \beta_{\boldsymbol{P}}.$$

Hilberg's law implies Heaps' law for PPM words.

Back to mathematical foundations

- Hilberg's and Zipf's laws may arise more generally.
- We would like to argue that processes that are strongly nonergodic may resemble Santa Fe processes.
- For this goal, we need a more careful inspection of maths.

Stationary and ergodic processes

A stochastic process $(X_i)_{i\in\mathbb{Z}}$ is called stationary if for all $t \in \mathbb{Z}$, all $k \in \mathbb{N}$ and all strings x_1^k, we have

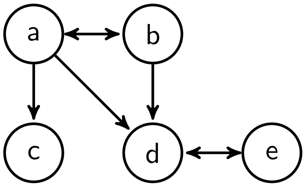
$$P(X_{t+1}^{t+k} = x_1^k) = P(X_1^k = x_1^k).$$

A stationary process $(X_i)_{i\in\mathbb{Z}}$ is called ergodic if for all $k \in \mathbb{N}$ and all strings x_1^k, we have

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} 1\left\{X_{i+1}^{i+k} = x_1^k\right\} = P(X_1^k = x_1^k) \text{ a.s.}$$

A non-ergodic Markov process

	a	***b***	***c***	***d***	***e***
a	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	0
b	$\frac{1}{6}$	0	0	$\frac{5}{6}$	0
c	0	0	1	0	0
d	0	0	0	$\frac{1}{2}$	$\frac{1}{2}$
e	0	0	0	$\frac{1}{2}$	$\frac{1}{2}$



Ergodic decomposition: Analogies and differences

- Just like any stationary Markov process can be decomposed into ergodic Markov processes, any stationary process can be decomposed into ergodic processes.
- The important difference is that a stationary Markov process can decompose into countably many ergodic components, whereas a general stationary process can decompose into uncountably many ergodic components.
- For example, non-ergodic Santa Fe process $\boldsymbol{X_i} = (\boldsymbol{K_i}, \boldsymbol{Z_{K_i}})$ where $(\boldsymbol{Z_k})_{\boldsymbol{k}\in\mathbb{N}}$ is an IID process decomposes into uncountably many ergodic Santa Fe processes $\boldsymbol{X_i} = (\boldsymbol{K_i}, \boldsymbol{z_{K_i}})$ where $(\boldsymbol{z_k})_{\boldsymbol{k}\in\mathbb{N}}$ are realizations of process $(\boldsymbol{Z_k})_{\boldsymbol{k}\in\mathbb{N}}$.

Falling into a particular ergodic component can last infinitely long.
This looks like an accumulation of frozen randomness.

Ergodic decomposition of excess entropy

Entropy of a σ-field:

$$H(\mathcal{J}) := \sup_{\alpha \subset \mathcal{J}} \left(- \sum_{A \in \alpha} P(A) \log_2 P(A) \right).$$

We have $H(\mathcal{J}) = \infty$ if σ-field $\mathcal{J}$ is non-atomic.

Shift-invariant σ-field:

$$\mathcal{I} := \left\{ A \in \mathcal{X}^{\mathbb{Z}} : A = T^{-1} A \right\}.$$

A process is non-ergodic iff $H(\mathcal{I}) > 0$.

Decomposition of excess entropy:

$$E = I(X_{-\infty}^{t}; X_{t+1}^{\infty}) = H(\mathcal{I}) + \underbrace{I(X_{-\infty}^{t}; X_{t+1}^{\infty} | \mathcal{I})}_{\text{excess entropy of components}}.$$

Thus $E = \infty$ if $H(\mathcal{I}) = \infty$ even if $I(X_{-\infty}^{t}; X_{t+1}^{\infty} | \mathcal{I}) < \infty$.

Knowledge growth and Hilberg exponents

A process is called strongly non-ergodic if $\mathcal{I}$ is non-atomic.

For a strongly non-ergodic process:

1. We may partition $\mathcal{I}$ so as to carve out a fair-coin process $(Z_k)_{k\in\mathbb{N}}$ that is $\mathcal{I}$-measurable.
2. There exists a guessing function $g : \mathbb{N} \times \mathbb{X}^* \to \{0, 1, 2\}$ s.t.
$$\lim_{n\to\infty} g(k; X_{t+1}^{t+n}) = Z_k \text{ almost surely.}$$
3. We may define the set of facts described in text X_1^n as
$$U(X_1^n) := \left\{ l \in \mathbb{N} : g(k; X_1^n) = Z_k \text{ for all } k \le l \right\}.$$
4. The number of described facts bounds the Hilberg exponent
$$\beta_P := \operatorname*{hilb}_{n\to\infty} \left(H(X_1^n) - nh \right) \ge \operatorname*{hilb}_{n\to\infty} \mathbb{E}\, \# U(X_1^n).$$

Theorem about facts and words

For a strongly non-ergodic process, consider an ergodic component $(X_i)_{i\in\mathbb{Z}}$ with algorithmically random knowledge $(Z_k)_{k\in\mathbb{N}}$.

Assume a computable guessing function $g : \mathbb{N} \times \mathbb{X}^* \to \{0, 1, 2\}$.

The Hilberg exponent of this process

$$\beta_C := \underset{n\to\infty}{\text{hilb}} \left(\mathbb{E}\, C(X_1^n) - nh\right) = \underset{n\to\infty}{\text{hilb}}\ \mathbb{E}\, J(X_1^n; X_{n+1}^{2n})$$

is bounded by the numbers of described facts and of PPM words:

$$\underset{n\to\infty}{\text{hilb}}\ \mathbb{E}\,\#\, U(X_1^n) \le \beta_C \le \underset{n\to\infty}{\text{hilb}}\ \mathbb{E}\,\#\, V_{\mathbb{M}}(X_1^n).$$

The number of distinct words in a finite text is roughly greater than the number of independent facts described by the text.

The knowledge is an algorithmically random parameter of an uncomputable ergodic probability measure.

Mixing processes

A stationary process $(X_i)_{i\in\mathbb{Z}}$ is called mixing if for all $k \in \mathbb{N}$ and all strings x_1^k, y_1^k, we have

$$\lim_{n\to\infty} P\left(X_{i+1}^{i+k} = x_1^k | X_1^k = y_1^k\right) = P(X_1^k = x_1^k) \text{ a.s.}$$

All mixing processes are ergodic.

Mixing Santa Fe processes:

$$X_i = (K_i, Z_{i,K_i}),$$

where facts evolve in time: $P(Z_{i+1,k} = z | Z_{i,k} = z) < 1$.

$$\lim_{k\to\infty} \frac{P(Z_{i+1,k} = z | Z_{i,k} = z)}{P(K_i = k)} = 0 \implies \beta_C = \beta_P > 0.$$

Hilberg's law when facts are mentioned quicker than they evolve.

The repetition time

- The repetition time $R_k^{(2)}$ is the first position in the process on which a copy of any previously seen string X_{j+1}^{j+k} occurs,

$$R_k^{(2)} := \inf \left\{ i \ge 1 : X_{i+1}^{i+k} = X_{j+1}^{j+k} \text{ for some } 0 \le j < i \right\}.$$

- For natural language, we have stretched exponential growth:

$$\log R_k^{(2)} \propto k^\beta, \quad \beta \approx 1/3.$$

Assume short memory $\log \gamma_k \sim 0$ and $\log \delta_k \sim 0$, where

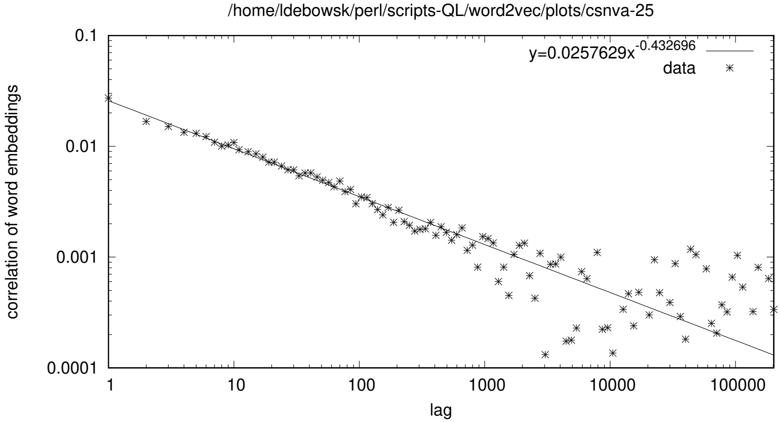
$$\gamma_k := \sup_{n \in \mathbb{N}} \max_{x_1^k} \frac{\operatorname{Var} F_n(x_1^k)}{\mathbb{E}\, F_n(x_1^k)}, \qquad F_n(x_1^k) := \sum_{i=0}^{n-1} 1\left\{ X_{i+1}^{i+k} = x_1^k \right\},$$

$$\delta_k := \sup_{n \in \mathbb{N}} \frac{\mathbb{E} \max_{x_1^k} F_n(x_1^k)}{\max_{x_1^k} \mathbb{E}\, F_n(x_1^k)}.$$

Then $\log R_k^{(2)} \sim -\log \max_{x_1^k} P(X_1^k = x_1^k)$ almost surely.

Word embedding correlations

We convert a word time series $(X_i)_{i\in\mathbb{Z}}$ into vectors $Y_i = f(X_i)$.



Autocorrelation: $\rho(k) := \dfrac{\mathbb{E}\, Y_i \cdot Y_{i+k} - \mathbb{E}\, Y_i \cdot \mathbb{E}\, Y_i}{\mathbb{E}\, Y_i \cdot Y_i - \mathbb{E}\, Y_i \cdot \mathbb{E}\, Y_i}$.

Recapitulation — The main result of this talk

Theorem about facts and words:

The number of distinct words in a finite text is roughly greater than the number of independent facts described by the text.

- The above proposition is a general result in information theory connected to Hilberg's and Zipf's laws.
- It's an impossibility result that pertains to a general stationary communication system.
- This result is paradoxical since we might think that combining words we may express many more independent facts.
- The paradox is less surprising if we realize that repeated facts are expressed via fixed phrases.

An account of descriptive meaningfulness

- Meaningfulness of texts can be understood as:
  1. description of some knowledge (descriptive m-fulness);
  2. internal cohesion of narration (cohesive m-fulness);
  3. control of the reader toward some goal (telic m-fulness).
- Our results concern only descriptive meaningfulness.
- Knowledge can be both described and created by texts.
- Knowledge may evolve in time, which may cause $\boldsymbol{E} < \infty$.
- Complexity of knowledge is extended by technical tools created by humans over ages (like script or internet).

Toward cohesive and telic meaningfulness

- Here our understanding and modeling is less advanced.
- Cohesive meaningfulness:
  - stretched exponential growth of repetition time, power-law growth of Rényi entropies;
  - power-law decay of word embedding correlations, large scale context-free structures, hierarchical memes.
- Telic meaningfulness:
  - arrow of time, (un)bounded accumulation of knowledge, (no) point Omega (singularity), AMS processes;
  - control of a (non)random environment, (non)deterministic interpretation of texts, positive entropy rate.
- Does cohesive m-fulness imply descriptive & telic m-fulness?

Idealization in statistical language models

- Stochastic processes = idealized models of possible texts.
  - This idealization becomes clear upon a closer scrutiny of these models, which takes effort, time, and imagination.
  - Imagination is a skill constructed through examples.
  - Linguistic and math intuitions can help each other.

- Sorts of idealization in stochastic processes:
  - actual or potential infinities (unbounded texts),
  - unbounded sources of (algorithmic) randomness,
  - infinite precision,
  - infinite recursion,
  - (conditional) computability of distributions,
  - rigid structure of mathematical definitions,
  - plethora of processes that cannot be effectively defined...
  - ... but these processes can be theorized about.

It's time for a synthesis!

Entropy not only speaks the language of arithmetic;
it also speaks the language of language.

— *Warren Weaver (1949)*

It is an irony of 20th century linguistics that Shannon's theory of information, though explicitly linked to semantics, was deemed irrelevant by linguists, while Chomsky's formal syntax, though explicitly dissociated from semantics, was adopted as the default theory of natural language.

— *Christian Bentz (2018)*

Some works of mine

Ł. Dębowski. On the vocabulary of grammar-based codes and the logical consistency of texts. *IEEE Trans. Inform. Theory*, 57:4589–4599, 2011.

Ł. Dębowski. *Information Theory Meets Power Laws: Stochastic Processes and Language Models*. Wiley & Sons, 2021a.

Ł. Dębowski. A refutation of finite-state language models through Zipf's law for factual knowledge. *Entropy*, 23:1148, 2021b.

Ł. Dębowski. Local grammar-based coding revisited. `https://arxiv.org/abs/2209.13636`, 2022.

Ł. Dębowski. A simplistic model of neural scaling laws: Multiperiodic Santa Fe processes. `https://arxiv.org/abs/2302.09049`, 2023a.

Ł. Dębowski. Recurrence and repetition times in the case of a stretched exponential growth. `https://arxiv.org/abs/2306.14703`, 2023b.