There Are Fewer Facts Than Words: Communication With A Growing Complexity

Łukasz Dębowski (ldebowsk@ipipan.waw.pl)

Institute of Computer Science
Polish Academy of Sciences
Warsaw, Poland

Key words: Zipf's law, mutual information, Markov order estimation, Kolmogorov complexity

Introduction

Several recent large-scale computational experiments in statistical language modeling reported power-law tails of learning curves [10, 7, 9, 5, 6, 11]. Namely, the difference between the cross entropy rate of the statistical language model and the entropy rate of natural language decays as a power law with the amount of training data. This is equivalent to a power-law growth of mutual information between increasing blocks of text—the first observation thereof attributed to Hilberg [8], see also [1]. This power-law growth occurs for languages as diverse as English, French, Russian, Chinese, Korean, and Japanese. Moreover, we observe a universal language-independent value of the power-law exponent: the mutual information between two blocks of length n is proportional to $n^{0.8}$.

We advertise some mathematical theory of this phenomenon that we have been developing for several years. Our results were resumed in the recently published book [3] and the subsequent article [4].

The basic theory of power-law-tailed learning curves consists in furnishing the proof of a theorem of form:

The number of distinct words used in a finite text is roughly greater than the number of independent elementary persistent facts described in this text.

We call this sort of a statement a theorem about facts and words. These theorems come into a few distinct flavors and can be proved easily, paying a certain attention to the formal understanding of the concepts of a fact and of a word.

Theorems about facts and words are an impossibility result that pertains to a general communication system. This result seems paradoxical since we might think that combining words we may express many more independent facts.

From a mathematical point of view, theorems about facts and words combine:

- Zipf's and Herdan-Heaps' laws for word frequency distributions,
- universal coding based on grammars and on normalized maximum likelihood,
- consistent (hidden) Markov order estimators,
- the concept of infinite excess entropy,
- the ergodic theorem and the ergodic decomposition,
- Kolmogorov complexity and algorithmic randomness.

Can theorems about facts and words be applied to natural language, computer programs, DNA, or music? Are there other perigraphic processes besides Santa Fe processes? Several related open questions were stated in [3] and [4].

References

[1] J. P. Crutchfield and D. P. Feldman. Regularities unseen, randomness observed: The entropy convergence hierarchy. *Chaos*, 15:25–54, 2003.

[2] Ł. Dębowski. On the vocabulary of grammar-based codes and the logical consistency of texts. *IEEE Transactions on Information Theory*, 57:4589–4599, 2011.

[3] Ł. Dębowski. *Information Theory Meets Power Laws: Stochastic Processes and Language Models*. New York: Wiley & Sons, 2021.

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

[5] T. Henighan, J. Kaplan, M. Katz, M. Chen, C. Hesse, J. Jackson, H. Jun, T. B. Brown, P. Dhariwal, S. Gray, et al. Scaling laws for autoregressive generative modeling. https://arxiv.org/abs/2010.14701, 2020.

[6] D. Hernandez, J. Kaplan, T. Henighan, and S. McCandlish. Scaling laws for transfer. https://arxiv.org/abs/2102.01293, 2021.

[7] J. Hestness, S. Narang, N. Ardalani, G. Diamos, H. Jun, H. Kianinejad, M. Patwary, M. Ali, Y. Yang, and Y. Zhou. Deep learning scaling is predictable, empirically. https://arxiv.org/abs/1712.00409, 2017.

[8] W. Hilberg. Der bekannte Grenzwert der redundanzfreien Information in Texten — eine Fehlinterpretation der Shannonschen Experimente? *Frequenz*, 44:243–248, 1990.

[9] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models. https://arxiv.org/abs/2001.08361, 2020.

[10] R. Takahira, K. Tanaka-Ishii, and Ł. Dębowski. Entropy rate estimates for natural language—a new extrapolation of compressed large-scale corpora. *Entropy*, 18(10):364, 2016.

[11] K. Tanaka-Ishii. *Statistical Universals of Language: Mathematical Chance vs. Human Choice*. New York: Springer, 2021.

REDUNDANCY AND MUTUAL INFORMATION

A string is denoted $x_j^k := (x_j, x_{j+1}, ..., x_k)$. The prefix-free Kolmogorov complexity is $K(u)$ and algorithmic mutual information is $J(u,v) := K(u) + K(v) - K(u,v)$. The Shannon entropy is $H(X) := \mathbf{E}\,[-\log P(X)]$, quantity $\mathbf{E}\,X$ being the expectation of X.

The **Hilberg exponent** is $\operatorname*{hilb}_{n\to\infty} S(n) := \left[\limsup_{n\to\infty} \frac{\log S(n)}{\log n}\right]_+$, so $\operatorname*{hilb}_{n\to\infty} n^\beta = \beta$ if $\beta \geq 0$.

For a discrete one-sided stochastic process $(X_i)_{i\in\mathbb{N}}$, we consider conditions:

(A) The complexity rate $h := \lim_{n\to\infty} \mathbf{E}\,K(X_1^n)/n$ exists and $\operatorname*{hilb}_{n\to\infty} \left[hn - H(X_1^n)\right] = 0$.

(B) The complexity does not decrease in time: $\mathbf{E}\,K(X_1^n) \leq \mathbf{E}\,K(X_{n+1}^{2n})$.

(C) The inverse complexity rate is finite, $\limsup_{n\to\infty} \mathbf{E}\,\frac{n}{K(X_1^n)} < \infty$. Thus $h > 0$ for (A).

(D) The alphabet is finite: $X_i : \Omega \to \{a_1, a_2, ..., a_D\}$, where $D \in \mathbb{N}$.

Conditions (A) and (B) are satisfied by any **stationary process** with (D). For (A) and (B),

$$\operatorname*{hilb}_{n\to\infty} \left[\mathbf{E}\,K(X_1^n) - hn\right] \leq \operatorname*{hilb}_{n\to\infty} \mathbf{E}\,J(X_1^n; X_{n+1}^n). \tag{1}$$

Condition $\operatorname*{hilb}_{n\to\infty} \mathbf{E}\,J(X_1^n; X_{n+1}^n) > 0$ is called the **Hilberg condition**, after Hilberg [8].

SANTA FE PROCESSES

The formal concept of facts can be most easily understood on the example of a certain stationary ergodic process over a countably infinite alphabet called a Santa Fe process [2]. Let $(K_i)_{i\in\mathbb{N}}$ be an IID process in natural numbers with **Zipf's distribution**

$$P(K_i = k) = \frac{k^{-\alpha}}{\zeta(\alpha)}, \quad k \in \mathbb{N}, \quad \alpha > 1, \quad \zeta(\alpha) := \sum_{k=1}^{\infty} k^{-\alpha}. \tag{2}$$

Moreover, let $(z_k)_{k\in\mathbb{N}}$ be an **algorithmically random sequence**, i.e., $K(z_1^k) \geq k - c$ for a certain constant $c < \infty$ and all lengths $k \in \mathbb{N}$. In the following, bits z_k will be called **facts**. Then the **Santa Fe process** $(X_i)_{i\in\mathbb{N}}$ is a sequence of pairs

$$X_i = (K_i, z_{K_i}). \tag{3}$$

The Santa Fe process is a model of a text that consists of random statements of form „the k-th fact equals z_k". These statements are **non-contradictory**, namely, if statements X_i and X_j describe the same fact ($K_i = K_j$) then they assert the same value of this fact ($z_{K_i} = z_{K_j}$). Moreover, facts z_k are **independent** (the complexity of their concatenation is the highest), **elementary** (they assume only two distinct values), and **persistent** (described faithfully at any time instant i).

FACTS AND REDUNDANCY

We say that a finite text x_1^n **describes** first l facts of a fixed sequence $(z_k)_{k\in\mathbb{N}}$ by means of a function g if $l+1 = U_g(x_1^n; z_1^\infty) := \min\left\{k \in \mathbb{N} : g(k, x_1^n) \neq z_k\right\}$. For a Santa Fe process, the expected number of initial facts described by a random text X_1^n grows as a power law

$$\operatorname*{hilb}_{n\to\infty} \mathbf{E}\,U_g(X_1^n; z_1^\infty) = 1/\alpha \in (0,1). \tag{4}$$

In general, for any stochastic process $(X_i)_{i\in\mathbb{N}}$ with (A), any algorithmically random sequence $(z_k)_{k\in\mathbb{N}}$, and any **computable function** g, we have

$$\operatorname*{hilb}_{n\to\infty} \mathbf{E}\,U_g(X_1^n; z_1^\infty) \leq \operatorname*{hilb}_{n\to\infty} \left[\mathbf{E}\,K(X_1^n) - hn\right]. \tag{5}$$

Processes $(X_i)_{i\in\mathbb{N}}$ with $\operatorname*{hilb}_{n\to\infty} \mathbf{E}\,U_g(X_1^n; z_1^\infty) > 0$, called **perigraphic**, are incomputable.

WORDS AND MUTUAL INFORMATION

Consider the **estimator** of the Markov order of the process defined as

$$M(x_1^n) := \min\left\{k \geq 0 : -\log L_k(x_1^n) \leq K(x_1^n)\right\}, \tag{6}$$

where $K(x_1^n)$ is the **Kolmogorov complexity** and $L_k(x_1^n)$ is the **maximum likelihood**,

$$L_k(x_1^n) := \max_Q \textstyle\prod_{i=k+1}^n Q(x_i|x_{i-k}^{i-1}), \quad Q(x_i|x_{i-k}^{i-1}) \geq 0, \quad \textstyle\sum_{x_i} Q(x_i|x_{i-k}^{i-1}) = 1. \tag{7}$$

Function $M(x_1^n)$ is a consistent estimator of the **Markov order**. Namely, for any stationary ergodic process $(X_i)_{i\in\mathbb{N}}$ with (D) we have $\lim_{n\to\infty} M(X_1^n) = M$ almost surely, where

$$M := \inf\left\{k \geq 0 : P(X_{k+1}^n|X_1^k) = \textstyle\prod_{i=k+1}^n P(X_i|X_{i-k}^{i-1}) \text{ for all } n > k\right\}. \tag{8}$$

A proxy for the number of words is the **subword complexity** of the Markov order estimate

$$V(x_1^n) := V(M(x_1^n)|x_1^n), \quad V(k|x_1^n) := \#\left\{x_{i+1}^{i+k} : 0 \leq i \leq n-k\right\}. \tag{9}$$

For any stochastic process $(X_i)_{i\in\mathbb{N}}$ that satisfies conditions (C) and (D), we have

$$\operatorname*{hilb}_{n\to\infty} \mathbf{E}\,J(X_1^n; X_{n+1}^n) \leq \operatorname*{hilb}_{n\to\infty} \mathbf{E}\,V(X_1^n). \tag{10}$$

Condition $\operatorname*{hilb}_{n\to\infty} \mathbf{E}\,V(X_1^n) > 0$ resembles **Herdan-Heaps' law** for words.