There are fewer facts than words: Power laws for stationary processes

Łukasz Dębowski
Institute of Computer Science PAS
e-mail: ldebowsk@ipipan.waw.pl

Several recent large-scale computational experiments in statistical language modeling reported power-law tails of learning curves [5, 7, 8, 9]. 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 [6], 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 $n^{0.8}$ [8, 9].

We advertise a mathematical theory of this phenomenon that we have been developing for several years. Most of our results were resumed in monograph [3] and articles [2, 4]. The focal point of the theory of power-law-tailed learning curves is a theorem of form:

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

We call this sort of a statement the theorem about facts and words. The theorem about facts and words is an impossibility result that pertains to a general stationary process and it associates ergodic decomposition with semantics and statistics.

This result seems paradoxical since we might think that combining words we could express many more independent facts. However, this theorem can be proved easily, by adopting quite natural definitions of facts and words. The ideas are as follows:

Preliminaries: Consider a discrete stationary process $(X_i)_{i\in\mathbb{Z}}$. Blocks of random variables are denoted $X_j^k := (X_j, X_{j+1}, ..., X_k)$. The *Shannon entropy* of a discrete random variable X is $H(X) := \mathbf{E}\,[-\log P(X)]$, $\mathbf{E}\,Y$ denoting the expectation of Y. There exists the limit called the *entropy rate* $h := \lim_{n\to\infty} H(X_1^n)/n$. Sublinear effects can be investigated using *Hilberg exponent* $\operatorname*{hilb}_{n\to\infty} S(n) := \left[\limsup_{n\to\infty} \frac{\log S(n)}{\log n}\right]_+$. The *redundancy exponent* is

$$\beta := \operatorname*{hilb}_{n\to\infty} \left[H(X_1^n) - hn\right]. \tag{1}$$

Condition $\beta > 0$ is called *Hilberg's law*, in honor of [6].

Facts and redundancy exponent: A stationary process $(X_i)_{i\in\mathbb{Z}}$ with a non-atomic invariant σ-field is called *strongly non-ergodic*. In this case, Let $(Z_k)_{k\in\mathbb{N}}$ be a Bernoulli($\frac{1}{2}$) process measurable with respect to this invariant σ-field. Variables Z_k are called *facts* because they do not depend on time. We say that a finite text x_1^n describes l initial facts by means of a function g if $l + 1 = U_g(x_1^n) := \min\{k \in \mathbb{N} : g(k, x_1^n) \neq Z_k\}$. For a strongly non-ergodic process $(X_i)_{i\in\mathbb{Z}}$ and any function g, we have a lower bound for the redundancy exponent of form

$$\operatorname*{hilb}_{n\to\infty} \mathbf{E}\, U_g(X_1^n) \leq \beta. \tag{2}$$

Santa Fe processes: The concept of facts can be illustrated by a simple example of a strongly non-ergodic process called the Santa Fe process [2]. Let $(K_i)_{i\in\mathbb{Z}}$ be an IID process in natural numbers with *Zipf's distribution* $P(K_i = k) \sim k^{-\alpha}$, where $\alpha > 1$. Let process $(Z_k)_{k\in\mathbb{N}}$ be Bernoulli($\frac{1}{2}$). The *Santa Fe process* $(X_i)_{i\in\mathbb{Z}}$ is a sequence of pairs

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

The Santa Fe process models 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 ($Z_{K_i} = Z_{K_j}$). Putting $g(k, x_1^n) := z$ if $(k, z) \in x_1^n$ and $(k, 1-z) \notin x_1^n$, whereas $g(k, x_1^n) := 2$ for other (k, x_1^n), we obtain power law $\operatorname*{hilb}_{n\to\infty} \mathbf{E}\, U_g(X_1^n) = 1/\alpha \in (0, 1)$.

Words and mutual information: Consider processes over a D-ary alphabet. The *maximum likelihood* in the class of Markov measures is

$$\hat{\mathbb{P}}(k|x_1^n) := \max_Q \prod_{i=k+1}^{n} Q(x_i|x_{i-k}^{i-1}), \qquad Q(x_i|x_{i-k}^{i-1}) \geq 0, \qquad \sum_{x_i} Q(x_i|x_{i-k}^{i-1}) = 1. \tag{4}$$

Consider also the *subword complexity* $V(k|x_1^n) := \#\left\{x_{i+1}^{i+k} : 0 \leq i \leq n-k\right\}$. The *penalized maximum likelihood* (PML) is $\mathbb{P}(x_1^n) := w_n \max_{k\geq 0} \frac{w_k \hat{\mathbb{P}}(k|x_1^n)}{Z(k|x_1^n)}$, where $w_k := \frac{1}{k+1} - \frac{1}{k+2}$ and $\log Z(k|x_1^n) := k \log D + DV(k|x_1^n)(\log n + 6)$. Since $\sum_{n\geq 0} \sum_{x_1^n} \mathbb{P}(x_1^n) \leq 1$, we have $\mathbf{E}\, K(X_1^n) \geq H(X_1^n)$ for the *PML entropy* $K(u) := -\log \mathbb{P}(u)$. We also have weak and strong universality, $\lim_{n\to\infty} K(X_1^n)/n = h$ a.s. and in L^1. Hence the redundancy exponent is bounded by the *PML mutual information* $J(u, v) := K(u) + K(v) - K(u, v)$ as

$$\beta \leq \operatorname*{hilb}_{n\to\infty} [\mathbf{E}\, K(X_1^n) - hn] = \operatorname*{hilb}_{n\to\infty} \mathbf{E}\, J(X_1^n; X_{n+1}^{2n}). \tag{5}$$

Statistic $M(x_1^n) := \min\left\{k \geq 0 : \hat{\mathbb{P}}(k|x_1^n) \geq w_n \mathbb{P}(x_1^n)\right\}$ is a consistent and asymptotically unbiased estimator of the *Markov order*. We have $\lim_{n\to\infty} M(X_1^n) = M$ a.s. and in L^1, where $M := \inf\left\{k \geq 0 : H(X_i|X_{i-k}^{i-1}) = h\right\}$ and $H(X|Y) := H(X, Y) - H(Y)$. Because of inequality $M(x_1^n)K(x_1^n) \leq n \log n$, we also have bound

$$\operatorname*{hilb}_{n\to\infty} \mathbf{E}\, J(X_1^n; X_{n+1}^{2n}) \leq \operatorname*{hilb}_{n\to\infty} \mathbf{E}\left[DV(X_1^n) + \frac{n \log D}{K(X_1^n)}\right], \tag{6}$$

where the *number of Markov suwords* is defined as $V(x_1^n) := V(M(x_1^n)|x_1^n)$. The power-law growth of $V(x_1^n) \sim n^{0.8}$ is observed for natural language.

Chaining inequalities (2), (5) i (6), we obtain the theorem about facts and words.

[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, *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] 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, *preprint*, `https://arxiv.org/abs/1712.00409`, 2017

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

[7] 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, *preprint*, `https://arxiv.org/abs/2001.08361`, 2020

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

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