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

Anonymous Author(s)
Affiliation
Address
email

Abstract

We present an impossibility result, called a theorem about facts and words, which pertains to a general communication system. The theorem states that the number of distinct words detectable in a finite text cannot be less than the number of independent elementary persistent facts described in the same text. In particular, this theorem can be related to Zipf's law, power-law scaling of mutual information, and power-law-tailed learning curves. The assumptions of the theorem are: a finite alphabet, linear sequence of symbols, complexity that does not decrease in time, entropy rate that can be estimated, and finiteness of the inverse complexity rate.

1 Introduction

In several recent large-scale computational experiments in statistical language modeling, there were observed power-law tails of learning curves [Takahira et al., 2016, Hestness et al., 2017, Hahn and Futrell, 2019, Braverman et al., 2020, Kaplan et al., 2020, Henighan et al., 2020, Hernandez et al., 2021, Tanaka-Ishii, 2021]. 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 respect to the amount of training data. Equivalently, this is tantamount to a power-law growth of mutual information between increasing blocks of text—the first observation thereof attributed to Hilberg [1990], see also [Crutchfield and Feldman, 2003]. This power-law growth occurs for languages typologically 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 $n^{0.8}$ [Takahira et al., 2016, Tanaka-Ishii, 2021].

In this paper, we would like to advertise some mathematical theory of this phenomenon, covering its potential causes and effects. We have been developing this theory for several years. Our results were resumed in the recently published book [citation suppressed] and the subsequent article [citation suppressed]. This paper supplies an updated brief overview for a venue of machine learning. The novel thing is a simple generalization to non-stationary processes with a growing complexity.

The basic theory of power-law-tailed learning curves can be simply stated as furnishing the proof of a general statement of form:

> The number of distinct words detectable in a finite text cannot be substantially less 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. In fact, the theorems about facts and words come into a few distinct flavors and can be proved relatively easily provided a certain attention is paid to the formal understanding of the concepts of a fact and of a word.

Submitted to Information-Theoretic Principles in Cognitive Systems Workshop at the 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

The theorems about facts and words can be regarded as an impossibility result that pertains to a general communication system. Simply speaking, one cannot communicate about a certain amount of independent facts in a repetitive fashion without effectively using at least as many distinct words. This result seems paradoxical since we might think that combining words we may express many more independent facts. Moreover, the above result links information theory and discrete stochastic processes with linguistics, semantics, and cognition. In principle, it applies to any kind of a communication system consisting of a finite number of discrete signs stringed into longer messages. Besides natural language, some obvious examples are computer programs, DNA, and music.

In particular, the statements and the proofs of the theorems about facts and words combine:

- Zipf's and Herdan-Heaps' laws for word frequency distributions [Zipf, 1935, Mandelbrot, 1954, Guiraud, 1954, Herdan, 1964, Heaps, 1978],
- universal coding based on grammars [de Marcken, 1996, Kieffer and Yang, 2000, Charikar et al., 2005] and on normalized maximum likelihood [Shtarkov, 1987, Ryabko, 1988, 2008],
- consistent (hidden) Markov order estimators [Merhav et al., 1989, Ziv and Merhav, 1992],
- the concept of infinite excess entropy [Hilberg, 1990, Ebeling and Nicolis, 1991, Ebeling and Pöschel, 1994, Bialek et al., 2001, Crutchfield and Feldman, 2003],
- the ergodic theorem and the ergodic decomposition [Birkhoff, 1932, Rokhlin, 1962, Gray and Davisson, 1974], and
- Kolmogorov complexity and algorithmic randomness [Kolmogorov, 1965, Martin-Löf, 1966, Li and Vitányi, 2008].

As we can see, there are many interacting mathematical concepts. There are also many open problems in the theory surrounding the theorems about facts and words. In the following, we will present some particular version of a theorem about facts and words, which pertains to algorithmic randomness [citation suppressed] and consistent Markov order estimation [citation suppressed].

2 Preliminaries

Consider a string $x_j^k := (x_j, x_{j+1}, ..., x_k)$ over a countable alphabet. Its prefix-free Kolmogorov complexity is denoted $K(x_j^k)$ [Li and Vitányi, 2008]. The algorithmic mutual information between strings u and v is is $J(u,v) := K(u) + K(v) - K(u,v)$. The expectation of random variable X is denoted $\mathbf{E}\,X$. The Shannon entropy of X is $H(X) := \mathbf{E}\,[-\log P(X)]$ [Cover and Thomas, 2006].

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

(*) The entropy rate $h := \inf_{n\in\mathbb{N}} H(X_1^n)/n$ can be estimated as $\lim_{n\to\infty} \mathbf{E}\,K(X_1^n)/n = h$.

(**) The complexity does not decrease in time: $\mathbf{E}\,K(X_j^k) \leq \mathbf{E}\,K(X_{j+1}^{k+1})$.

(***) The inverse complexity rate is finite, $H := \limsup_{n\to\infty} \mathbf{E}\,\dfrac{n}{K(X_1^n)} < \infty$ (thus $h > 0$ for (*)).

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

In particular, conditions (*) and (**) are satisfied by any stationary process that satisfies (****).

As in [citation suppressed], we will use the Hilberg exponent of a sequence, defined as

$$\operatorname*{hilb}_{n\to\infty} S(n) := \left[\limsup_{n\to\infty} \frac{\log S(n)}{\log n}\right]_+, \quad r_+ := r\,\mathbf{1}\{r \geq 0\}, \tag{1}$$

so that $\mathrm{hilb}_{n\to\infty}\, n^\beta = \beta$ for $\beta \geq 0$. We recall that if limit $s = \lim_{n\to\infty} S(n)/n$ exists then

$$\operatorname*{hilb}_{n\to\infty} [S(n) - sn] \leq \operatorname*{hilb}_{n\to\infty} [2S(n) - S(2n)]\,. \tag{2}$$

In particular, for conditions (*) and (**), we obtain that the power-law rate of redundancy is dominated by the power-law rate of mutual information,

$$\operatorname*{hilb}_{n\to\infty} [\mathbf{E}\,K(X_1^n) - hn] \leq \operatorname*{hilb}_{n\to\infty} \mathbf{E}\,J(X_1^n; X_{n+1}^n). \tag{3}$$

Condition $\operatorname*{hilb}_{n\to\infty} \mathbf{E}\,J(X_1^n; X_{n+1}^n) > 0$ is called the Hilberg condition, after Hilberg [1990].

3 Facts

The concept of independent elementary persistent 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 [citation suppressed]. Let $(K_i)_{i\in\mathbb{N}}$ be an IID process in natural numbers with the Zipfian marginal distribution

$$P(K_i = k) = \frac{k^{-\alpha}}{\zeta(\alpha)}, \tag{4}$$

where $k \in \mathbb{N}$, $\alpha > 1$ is a fixed parameter, and $\zeta(\alpha) := \sum_{k=1}^{\infty} k^{-\alpha}$ is the Riemann zeta function. Distribution (4) is a formal model of Zipf's law from quantitative linguistics [Zipf, 1935, Mandelbrot, 1954]. Moreover, let $(z_k)_{k\in\mathbb{N}}$ be an algorithmically random sequence, i.e., a sequence of particular (= fixed) bits (= coin flips) such that the Kolmogorov complexity of any string z_1^k is the highest possible, $K(z_1^k) \geq k - c$ for a certain constant $c < \infty$ and all lengths $k \in \mathbb{N}$ [Li and Vitányi, 2008]. Then the Santa Fe process $(X_i)_{i\in\mathbb{N}}$ is a sequence of pairs

$$X_i = (K_i, z_{K_i}). \tag{5}$$

In the following, bits z_k will be called facts.

The Santa Fe process can be understood as a model of an infinite text that consists of random statements of form „the k-th fact equals z_k". Importantly, 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, we observe that facts z_k are in some sense independent (the Kolmogorov complexity of their concatenation is the highest possible), elementary (they assume only two distinct values), and persistent (described faithfully at any time instant).

We will say that a finite text x_1^n describes exactly first l facts of a fixed sequence $(z_k)_{k\in\mathbb{N}}$ by means of a computable function g if $l = U_g(x_1^n; z_1^\infty) - 1$, where

$$U_g(x_1^n; z_1^\infty) := \min\left\{k \in \mathbb{N} : g(k, x_1^n) \neq z_k\right\}. \tag{6}$$

For a Santa Fe process we can easily construct a function g such that $U_g(X_1^n; z_1^\infty) = \min\{k \in \mathbb{N} : k \notin K_1^n\}$ by reading off the values of z_k for all $k \in K_1^n$. Hence it can be proved that the expected number of initial facts described by a random text X_1^n grows as a power law

$$\operatorname*{hilb}_{n\to\infty} \mathbf{E}\, U_g(X_1^n; z_1^\infty) = 1/\alpha \in (0, 1). \tag{7}$$

Power laws (4) and (7) are related to each other as Zipf's law is related to Herdan-Heaps' law [Guiraud, 1954, Herdan, 1964, Heaps, 1978]. Processes $(X_i)_{i\in\mathbb{N}}$ such that $\operatorname*{hilb}_{n\to\infty} \mathbf{E}\, U_g(X_1^n; z_1^\infty) > 0$ for a certain algorithmically random sequence $(z_k)_{k\in\mathbb{N}}$ and a certain computable function g are called perigraphic. All perigraphic processes have incomputable probability distributions. We notice that Santa Fe processes are perigraphic.

These results can be linked to the redundancy rate. As shown in [citation suppressed], for any discrete stochastic process $(X_i)_{i\in\mathbb{N}}$, any algorithmically random sequence $(z_k)_{k\in\mathbb{N}}$, and any computable function g, we obtain inequality

$$\mathbf{E}\, U_g(X_1^n; z_1^\infty) - 6 \log \mathbf{E}\, U_g(X_1^n; z_1^\infty) - c_g \leq \sup_{k\in\mathbb{N}} \mathbf{E}\, J(X_1^n; z_1^k) \leq \mathbf{E}\, K(X_1^n) - H(X_1^n). \tag{8}$$

where constant $c_g < \infty$ depends on function g. Hence, the power-law rate of the number of facts is dominated by the power-law rate of redundancy,

$$\operatorname*{hilb}_{n\to\infty} \mathbf{E}\, U_g(X_1^n; z_1^\infty) \leq \operatorname*{hilb}_{n\to\infty} \left[\mathbf{E}\, K(X_1^n) - hn\right]. \tag{9}$$

The above inequality can be chained with inequality (3).

4 Words

In the remaining move, we will relate the algorithmic mutual information to the number of words detectable in a given text. This can be done in several ways. One way, pursued by [citation suppressed], is to apply the minimal grammar-based coding [Kieffer and Yang, 2000, Charikar et al., 2005], which

applies a mathematical concept of a word that resembles words in a linguistic sense [de Marcken, 1996] but this method is a bit lengthy to describe formally. Therefore, here, we will apply a different approach which is based on consistent Markov order estimation [Merhav et al., 1989, Ziv and Merhav, 1992]—that approach was pursued by [citation suppressed].

There is a function called subword complexity that counts how many distinct substrings of a given length there are in a given string, namely,

$$V(k|x_1^n) := \#\left\{x_{i+1}^{i+k} : 0 \le i \le n-k\right\}. \tag{10}$$

Function $V(k|x_1^n)$ will be a proxy for the number of distinct words in text x_1^n. The only problem is to choose a motivated length k of the substrings. In principle, this k may depend on string x_1^n.

Quite a natural choice of k is the estimator of the Markov order of the process defined as

$$M(x_1^n) := \min\left\{k \ge 0 : (n-k)h_k(x_1^n) \le K(x_1^n)\right\}, \tag{11}$$

where $K(x_1^n)$ is the Kolmogorov complexity and $h_k(x_1^n)$ is the empirical entropy of order k,

$$h_k(x_1^n) := \frac{1}{n-k}\sum_{i=k+1}^{n} \log \frac{N(x_{i-k}^{i-1}|x_1^{n-1})}{N(x_{i-k}^{i}|x_1^n)}, \qquad N(w_1^k|x_1^n) := \sum_{i=0}^{n-k} \mathbf{1}\{x_{i+1}^{i+k} = w_1^k\}. \tag{12}$$

Function $M(x_1^n)$ is a strongly consistent and asymptotically unbiased estimator of the Markov order. Namely, for any stationary ergodic process $(X_i)_{i\in\mathbb{N}}$ over a finite alphabet we have

$$\lim_{n\to\infty} M(X_1^n) = M \text{ almost surely}, \qquad \lim_{n\to\infty} \mathbf{E}\, M(X_1^n) = M. \tag{13}$$

where we denote the Markov order of the process as

$$M := \inf\left\{k \ge 0 : P(X_{k+1}^n|X_1^k) = \prod_{i=k+1}^{n} P(X_i|X_{i-k}^{i-1}) \text{ for all } n > k\right\}, \quad \inf\emptyset := \infty. \tag{14}$$

Let us write succinctly the number of substrings of the supposedly optimal length as

$$V(x_1^n) := V(M(x_1^n)|x_1^n). \tag{15}$$

Then, using the universal code by Ryabko [1988], we can prove inequality

$$J(x_1^n; x_{n+1}^{2n}) \le 2\left[DV(x_1^{2n}) + \frac{4n\log D}{K(x_1^{2n})} + c_1\right](\log n + c_2), \tag{16}$$

where D is the cardinality of the alphabet and $c_i < \infty$ are certain constants [citation suppressed]. Applying Hilberg exponents and expectations, we obtain that the power-law rate of mutual information is dominated by the power-law rate of the number of words,

$$\operatorname*{hilb}_{n\to\infty} \mathbf{E}\, J(X_1^n; X_{n+1}^n) \le \operatorname*{hilb}_{n\to\infty} \mathbf{E}\, V(X_1^n), \tag{17}$$

if conditions (***) and (****) are satisfied. The above inequality can be chained with inequalities (3) and (9). The asymptotic power law $\operatorname*{hilb}_{n\to\infty} \mathbf{E}\, V(X_1^n) > 0$ resembles Herdan-Heaps' law for words in the linguistic sense [Guiraud, 1954, Herdan, 1964, Heaps, 1978].

5 Conclusion

Chaining inequalities (3), (9), and (17) under conditions (*)–(****), we obtain the sandwich bound

$$\operatorname*{hilb}_{n\to\infty} \mathbf{E}\, U_g(X_1^n; z_1^\infty) \le \operatorname*{hilb}_{n\to\infty} [\mathbf{E}\, K(X_1^n) - hn] \le \operatorname*{hilb}_{n\to\infty} \mathbf{E}\, J(X_1^n; X_{n+1}^n) \le \operatorname*{hilb}_{n\to\infty} \mathbf{E}\, V(X_1^n), \tag{18}$$

which yields a formal statement of a certain theorem about facts and words. In particular, we can infer from inequalities (18) that no perigraphic process can be a Markov process, i.e., a process with a finite Markov order, $M < \infty$. These two classes of processes are disjoint.

Of course, the above inequalities rest on a repeated use of Kolmogorov complexity. Therefore they are ineffective in some sense. We note that there are other statements of theorems about facts and words that apply effective notions. However, they are more complicated to formulate.

The theorems about facts and words, as an impossibility result, raise questions about their applicability to empirical data as well as questions about further examples of perigraphic processes that are more complex or more realistic than Santa Fe processes. Some of these questions were addressed or stated as open problems in book [citation suppressed] and article [citation suppressed].

References

W. Bialek, I. Nemenman, and N. Tishby. Complexity through nonextensivity. *Physica A*, 302:89–99, 2001.

G. D. Birkhoff. Proof of the ergodic theorem. *Proceedings of the National Academy of Sciences of the United States of America*, 17:656–660, 1932.

M. Braverman, X. Chen, S. M. Kakade, K. Narasimhan, C. Zhang, and Y. Zhang. Calibration, entropy rates, and memory in language models. In *2020 International Conference on Machine Learning (ICML)*, 2020.

M. Charikar, E. Lehman, A. Lehman, D. Liu, R. Panigrahy, M. Prabhakaran, A. Sahai, and A. Shelat. The smallest grammar problem. *IEEE Transactions on Information Theory*, 51:2554–2576, 2005.

T. M. Cover and J. A. Thomas. *Elements of Information Theory, 2nd ed.* New York: Wiley & Sons, 2006.

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

C. G. de Marcken. *Unsupervised Language Acquisition*. PhD thesis, Massachussetts Institute of Technology, 1996.

W. Ebeling and G. Nicolis. Entropy of symbolic sequences: the role of correlations. *Europhysics Letters*, 14:191–196, 1991.

W. Ebeling and T. Pöschel. Entropy and long-range correlations in literary English. *Europhysics Letters*, 26:241–246, 1994.

R. M. Gray and L. D. Davisson. Source coding theorems without the ergodic assumption. *IEEE Transactions on Information Theory*, 20:502–516, 1974.

P. Guiraud. *Les caractères statistiques du vocabulaire*. Paris: Presses Universitaires de France, 1954.

M. Hahn and R. Futrell. Estimating predictive rate-distortion curves via neural variational inference. *Entropy*, 21:640, 2019.

H. S. Heaps. *Information Retrieval—Computational and Theoretical Aspects*. New York: Academic Press, 1978.

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.

G. Herdan. *Quantitative Linguistics*. London: Butterworths, 1964.

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

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.

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

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.

J. C. Kieffer and E. Yang. Grammar-based codes: A new class of universal lossless source codes. *IEEE Transactions on Information Theory*, 46:737–754, 2000.

A. N. Kolmogorov. Three approaches to the quantitative definition of information. *Problems of Information Transmission*, 1(1):1–7, 1965.

M. Li and P. M. B. Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications, 3rd ed.* New York: Springer, 2008.

B. Mandelbrot. Structure formelle des textes et communication. *Word*, 10:1–27, 1954.

P. Martin-Löf. The definition of random sequences. *Information and Control*, 9:602–619, 1966.

N. Merhav, M. Gutman, and J. Ziv. On the estimation of the order of a Markov chain and universal data compression. *IEEE Transactions on Information Theory*, 35(5):1014–1019, 1989.

V. A. Rokhlin. On the fundamental ideas of measure theory. *American Mathematical Society Translations, series 1*, 10:1–54, 1962.

B. Ryabko. Compression-based methods for nonparametric density estimation, on-line prediction, regression and classification for time series. In *2008 IEEE Information Theory Workshop, Porto*, pages 271–275. Institute of Electrical and Electronics Engineers, 2008.

B. Y. Ryabko. Prediction of random sequences and universal coding. *Problems of Information Transmission*, 24(2):87–96, 1988.

Y. M. Shtarkov. Universal sequential coding of single messages. *Problems of Information Transmission*, 23(2):3–17, 1987.

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.

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

G. K. Zipf. *The Psycho-Biology of Language: An Introduction to Dynamic Philology*. Boston: Houghton Mifflin, 1935.

J. Ziv and N. Merhav. Estimating the number of states of a finite-state source. *IEEE Transactions on Information Theory*, 38(1):61–65, 1992.