Journal of Machine Learning Research 27 (2026) 1-35 Submitted 12/25; Revised 6/26; Published 8/26

From Zipf's Law to Neural Scaling through Heaps' Law and Hilberg's Hypothesis

Łukasz Dębowski LDEBOWSK@IPIPAN.WAW.PL
Institute of Computer Science
Polish Academy of Sciences
01-248 Warszawa, Poland

Editor: Maya Gupta

Abstract

We inspect the deductive connection between the neural scaling law and Zipf's law—two statements discussed in machine learning and quantitative linguistics. The neural scaling law describes how the cross entropy rate of a foundation model—such as a large language model—changes with respect to the amount of training tokens, parameters, and compute. By contrast, Zipf's law posits that the distribution of tokens exhibits a power law tail. Whereas similar claims have been made in more specific settings, we show that the neural scaling law is a consequence of Zipf's law under certain broad assumptions that we reveal systematically. The derivation steps are as follows: We derive Heaps' law on the vocabulary growth from Zipf's law, Hilberg's hypothesis on the entropy scaling from Heaps' law, and the neural scaling from Hilberg's hypothesis. We illustrate these inference steps by a toy example of the Santa Fe process that satisfies all four statistical laws.

Keywords: neural scaling law, Zipf's law, Heaps' law, Hilberg's hypothesis, Santa Fe processes

1. Introduction

It has been increasingly recognized in the machine learning literature (Hutter, 2021; Maloney et al., 2022; Michaud et al., 2023; Cabannes et al., 2024; Neumann and Gros, 2025; Pan et al., 2025) that the *neural scaling law* observed for contemporary foundation models—such as large language models—may arise as a consequence of *Zipf's law* or similar distributional regularities of natural language. This emerging perspective suggests that some remarkable empirical regularities in large-scale deep learning need not be explained solely by architectural, optimization, or hardware considerations, but instead reflect intrinsic statistical constraints of *natural texts*—human linguistic data.

In the present paper, we aim to provide an actually simple but systematic derivation of this chain of implications that extends earlier results in information theory, probability theory, and quantitative linguistics. We will prove formal theorems about as general stochastic processes as possible rather than experiment with particular empirical data. Our goal is to consolidate knowledge of several scientific disciplines by means of mathematical deduction. We try to avoid a too abstract formalism to make this paper more accessible.

The conceptual trajectory is as follows. Beginning with Zipf's law for word frequencies, we derive *Heaps' law* for vocabulary growth. From Heaps' law we extract *Hilberg's hypoth-*

esis on the sublinear growth of block entropy. Finally, we show that Hilberg's hypothesis leads to the *neural scaling* that ties model performance to the amounts of training data, model parameters, and training compute. In parallel, we discuss *Santa Fe processes*—toy stochastic sources that exhibit these statistical laws. Each link in this chain introduces its own assumptions. By isolating these steps, we hope to illuminate where the derivations might be strengthened in the future.

Further organization of the article. In Section 2, we map preliminaries and the research context, including basic notation, power laws at large, and prior theoretical connections. Section 3 contains the formal statement of the main theorems, preceded by a discussion of their novelty and followed by a consideration of their relevance. Section 4 concludes the article. The paper is followed by two appendices. Appendix A contains general results about counting types and bits of information for diverse stochastic sources. Proofs of the main theorems can be found in Appendix B.

2. Preliminaries and Context

This section begins with an explanation of mathematical notation, to be followed by a discussion of empirical power laws in quantitative linguistics and machine learning, to be concluded by the attempts of relating them formally.

2.1 Basic Notations

Consider a countable alphabet $\mathbb{X}$, which need not be finite. All considered random variables live on the measurable space of doubly infinite sequences $(\Omega, \mathcal{J})$, where $\Omega = \mathbb{X}^{\mathbb{Z}}$ and $\mathcal{J} \subset 2^{\Omega}$ is the smallest σ-field generated by cylinder sets. Function $P : \mathcal{J} \to [0,1]$ is the "true" probability measure that need not have a finite description or be computable.

Tokens and types. Notation $(X_t)_{t\in\mathbb{Z}}$ denotes an integer-time stochastic process consisting of countable-alphabet random variables $X_j : \Omega \ni (x_t)_{t\in\mathbb{Z}} \mapsto x_j \in \mathbb{X}$. As in quantitative linguistic research, we distinguish tokens and types. That is, string

$$X_j^k := (X_j, X_{j+1}, \ldots, X_k)$$

consists of random elements called tokens, whereas calligraphically written set

$$\mathcal{X}_j^k := \{X_j, X_{j+1}, \ldots, X_k\}$$

consists of random elements called types (i.e., unique tokens). We adopt that X_1^0 is the empty string and $\mathcal{X}_1^0$ is the empty set. Analogous notations are applied to processes denoted by other letters such as $(K_t)_{t\in\mathbb{Z}}$ or $(Z_k)_{k\in\mathbb{N}}$.

Information measures. Expectation with respect to the "true" probability measure P is denoted $\mathbf{E}\, X := \int X dP$. Random variable $P(X)$ for a random variable X is defined as

$$P(X)(\omega) := P(X = x) \iff X(\omega) = x.$$

This convention allows us to shorten formulas by writing $P(X)$ instead of $P(X = x)$. In particular, notation

$$H(X) := \mathbf{E}(-\log P(X))$$

denotes the Shannon entropy of a simple random variable X. The conditional entropy is $H(X|Z) := H(X,Z) - H(Z)$, the mutual information is $I(X;Y) := H(X) - H(X|Y)$, and the conditional mutual information is $I(X;Y|Z) := H(X|Z) - H(X|Y,Z)$.

Extension to sigma-fields. For random variables that are not simple, i.e., assume an infinite number of values, we use a convenient generalization of the above Shannon information measures to arbitrary σ-fields. For a textbook treatment of their calculus, see Dębowski (2021a, Sections 5.3 and 5.4). The original contributions were made by Gelfand et al. (1956), Dobrushin (1959), Pinsker (1964), Wyner (1978), and Dębowski (2009, 2020).

Random measures. Symbol Q most often denotes a random probability measure on the measurable space $(\Omega, \mathcal{J})$, which is "learned" from a finite sample such as X_j^k, i.e., $Q = f(X_j^k)$ for a certain (stochastic) function f. Mind that the entropy of Q equals

$$H(Q) = -\sum_q P(Q=q) \log P(Q=q),$$

where q are distinct values (concrete measures) that Q assumes. In particular, it is not the entropy rate of process $(X_t)_{t\in\mathbb{Z}}$ with respect to Q. In contrast to P, values of measure Q are often computable objects, expressed via a finite number of finite-resolution parameters. In particular, $H(Q) \le H(X_j^k)$ for $Q = f(X_j^k)$ and a deterministic function f.

Expected cardinalities. An important analogue of Shannon entropy in our considerations is the expected cardinality of a random finite set $\mathcal{X}$ denoted

$$V(\mathcal{X}) := \mathbf{E}\,\#\,\mathcal{X}.$$

Respectively, the analogue of conditional entropy $H(X|Z) = H(X,Z) - H(Z)$ is the expected cardinality of set difference $V(\mathcal{X} \setminus \mathcal{Z}) = V(\mathcal{X} \cup \mathcal{Z}) - V(\mathcal{Z})$ and the counterpart of mutual information $I(X;Y) = H(X)+H(Y)-H(X,Y)$ is the expected cardinality of intersection $V(\mathcal{X} \cap \mathcal{Y}) = V(\mathcal{X}) + V(\mathcal{Y}) - V(\mathcal{X} \cup \mathcal{Y})$, cf. the idea of I-measure (Hu, 1962; Yeung, 2002). However, this is an incomplete analogy since in general there is no random variable $Z = f(X,Y)$ such that $H(Z) = I(X;Y)$, cf. Gács and Körner (1973); Wyner (1975).

2.2 Power Laws at Large

Let us briefly present four power laws to be related in our formal reasoning—in the chronological order of their discovery.

Zipf's law. The oldest known of quantitative linguistic laws, Zipf's law asserts that, for texts in natural language, the frequency of the k-th most common type of word decays approximately as $k^{-\alpha}$, where $\alpha \approx 1$. This regularity was noticed over one century ago (Estoup, 1916; Condon, 1928; Zipf, 1935). An empirical study of this law across one hundred languages can be found in the work by Mehri and Jamaati (2017). Similar power-law distributions are observed also in ecology, sociology, economics, and physics (Zipf, 1949), being a hallmark of complex systems. The literature on Zipf's law is scattered over diverse venues. Historical references have been compiled by Li (2021).

It has been known that Zipf's law is a stylized fact rather than follows simple exact formulas. Despite these complications, we will assume that stationary processes $(K_t)_{t\in\mathbb{Z}}$ over natural numbers with the approximate Zipf probability distribution

$$A_Z^- k^{-\alpha_Z} \le P(K_t = k) \le A_Z^+ k^{-\alpha_Z}, \qquad \alpha_Z > 1, \qquad A_Z^\pm > 0, \tag{1}$$

provide a compact description of heavy-tailed data distributions. Statistical law (1), called also the Zipf-Mandelbrot law after Mandelbrot (1954), will be our departure point.

Heaps' law. A closely related law, Heaps' law, also known as Herdan's law, describes the vocabulary growth of natural texts (Kuraszkiewicz and Łukaszewicz, 1951; Guiraud, 1954; Herdan, 1964; Heaps, 1978). Heaps' law is also studied for large language models (Tao et al., 2024; Lai et al., 2023). According to this law, the expected number of unique types in the first t tokens of the process $(K_t)_{t\in\mathbb{Z}}$ considered in (1) increases like a sublinear power-law function,

$$A_V^- t^{\beta_V} \leq V(\mathcal{K}_1^t) \leq A_V^+ t^{\beta_V}, \qquad 0<\beta_V<1, \qquad A_V^\pm > 0. \tag{2}$$

Heaps' law (2) is widely viewed as a direct consequence of Zipf's law (1) with $\beta_V = 1/\alpha_Z$ but there are some complications, which we will inspect further. A naive estimation often yields $\alpha_Z \approx 1$ and $\beta_V \approx 0.8$ for million-token corpora (novel-sized texts). There are also problems with languages that use the ideographic script (Tanaka-Ishii, 2021).

Hilberg's hypothesis. A reinterpretation of Shannon's early findings from 1950's (Shannon, 1948, 1951), Hilberg's hypothesis, also called Hilberg's law, was developed around 1990's, mostly in physics of complex systems (Hilberg, 1990; Ebeling and Nicolis, 1992; Bialek et al., 2001; Crutchfield and Feldman, 2003; Dębowski, 2006, 2011). According to this hypothesis, for a stationary process $(X_t)_{t\in\mathbb{Z}}$ that models text in natural language, the block entropy of the first t word tokens contains a sublinear power-law component,

$$A_H^- t^{\beta_H} \leq H(X_1^t) - ht \leq A_H^+ t^{\beta_H}, \qquad 0<\beta_H<1, \qquad A_H^\pm > 0. \tag{3}$$

where we apply the entropy rate

$$h := \inf_{t\in\mathbb{N}} \frac{H(X_1^t)}{t}. \tag{4}$$

The original rough estimate by Hilberg (1990) was $\beta_H = 0.5$ and was based on a meager evidence for $t \leq 100$. If we estimate the entropy by the prediction-by-partial matching (PPM) algorithm (Ryabko, 1984, 1988; Cleary and Witten, 1984), then law (3) holds more universally and uniformly than Zipf's or Heaps' laws. The empirical estimate $\beta_H \approx 0.8$ obtained for the PPM algorithm is quite stable for billion-token corpora and does not differ significantly across typologically diverse languages, using either alphabetic or ideographic scripts (Takahira et al., 2016; Tanaka-Ishii, 2021). Thus Hilberg's law seems a plausible candidate for a statistical language universal (Tanaka-Ishii, 2021).

Neural scaling. The universal power-law behavior of information measures, when applied to natural big data, can be further confirmed by experiments with large foundation models—language or multimodal models in particular, developed in the beginning of 2020's (Devlin et al., 2019; Radford et al., 2019; Brown et al., 2020; Thoppilan et al., 2022; Chowdhery et al., 2023). These advanced statistical models build upon deep neural networks (Bengio et al., 2021), word embeddings (Mikolov et al., 2013), and the transformer architecture (Vaswani et al., 2017). Foundation models can be regarded as a game-changer in the research of language and cognition, as they allow to test probabilistic hypotheses about human language on an unprecedented scale and detail (Futrell and Mahowald, 2026; Futrell and Hahn, 2026).

A particularly salient empirical finding is that the predictive performance of these models improves with scale in a power-law fashion. The neural scaling law characterizes how the loss function of a foundation model—typically measured by the cross entropy rate on a test dataset—decreases as training data t, model size n, and compute c increase (Hestness et al., 2017; Kaplan et al., 2020; Henighan et al., 2020; Hernandez et al., 2021; Hoffmann et al., 2022; Pearce and Song, 2024; Li et al., 2026). Simplifying particular empirical observations and ignoring complex interactions among t, n, and c, this power-law relationship can be approximated as

$$A_T^- t^{-\gamma_T} \leq h(s,t,\infty,\infty) - h \leq A_T^+ t^{-\gamma_T}, \qquad 0 < \gamma_T < 1, \qquad A_T^{\pm} > 0, \tag{5}$$

$$A_N^- n^{-\gamma_N} \leq h(s,\infty,n,\infty) - h \leq A_N^+ n^{-\gamma_N}, \qquad 0 < \gamma_N < 1, \qquad A_N^{\pm} > 0, \tag{6}$$

$$A_C^- n^{-\gamma_C} \leq h(s,\infty,\infty,c) - h \leq A_C^+ c^{-\gamma_C}, \qquad 0 < \gamma_C < 1, \qquad A_C^{\pm} > 0, \tag{7}$$

for a fixed $s < \infty$, where h is the entropy rate and $h(s,t,n,c)$ is the cross entropy rate of a foundation model trained on t tokens and tested on s tokens with the amount of parameters n and the amount of compute c. Laws (5)–(7) hold for trillion-token corpora.

2.3 Prior Connections and Explanations

The history of theoretical explanations of power laws is as ancient as empirical observations of these regularities (Li, 2021). To explain statistical regularities observed in natural language is a long-standing goal of quantitative linguistics. We envision that a similarly systematic approach to Hilberg's law and neural scaling may succeed as well. A desired unified theory of language and large language models should predict the functional forms of all these laws simultaneously and predict the values of their parameters. Let us comment briefly on the progress made so far.

Quantitative linguistics. In spite of sheer cross-disciplinary literature coverage (or maybe because of that), there is no single explanation of Zipf's law. The law can be explained by diverse mechanisms ranging from monkey-typing (Mandelbrot, 1954; Miller, 1957), through preferential attachment (Simon, 1955), also known as the Chinese restaurant process (Aldous, 1985, page 92), to potential links with game theory (Harremoës and Topsøe, 2005), semantics, and information theory (Dębowski, 2021a).

Explaining Zipf's and Heaps' laws is not so straightforward because of multiple phenomena that break the simple picture of power laws in natural language. These challenges include two-regime rank-frequency plots (Ferrer-i-Cancho and Solé, 2001; Montemurro and Zanette, 2002; Petersen et al., 2012; Gerlach and Altmann, 2013; Williams et al., 2015; Chacoma and Zanette, 2020), log-log convexity of the vocabulary growth (Font-Clos and Corral, 2015), monotone decreasing or U-shaped hapax rates (Fan, 2010; Dębowski, 2025), and finite-size effects (Lü et al., 2010). One should also note that expected values do not explain all observed phenomena since variances are large for language data, due to Taylor's law (Kobayashi and Tanaka-Ishii, 2018; Tanaka-Ishii, 2021). One also observes burstiness of words, i.e., departure of recurrence times from the Poisson process (Altmann et al., 2009) and long-range correlations, i.e., slow decay of mutual information between two word tokens separated by a given lag (Wieczyński and Dębowski, 2025).

In spite of these challenges, it is fascinating that the empirical marginal distribution of words, summarized in Zipf's and Heaps' laws, can be quite well modeled by non-parametric

and parametric urn models—IID sources of words (Baayen, 2001; Milička, 2009, 2013; Davis, 2018; Dębowski, 2025). For this reason, it is sound to investigate Zipf's law and Heaps' law mathematically as if the generating process were a memoryless source (Karlin, 1967; Khmaladze, 1988; Pitman and Yor, 1997; Baayen, 2001) or more generally a stationary process—as it will be assumed in this paper. Such a theory is capable of deriving Heaps' law (2) from Zipf's law (1) under broad conditions.

Santa Fe processes. Since Hilberg's law (3) does not hold for IID and finite-state sources (Crutchfield and Feldman, 2003; Dębowski, 2021b), this observation might be considered a witness to large memory and complex structure of natural texts. This view is somewhat inaccurate, however. Large memory does not necessitate complex structure in an intuitive sense. A simple stochastic source that satisfies condition (3) is the Santa Fe process described by Dębowski (2005, 2009, 2011) and later rediscovered by Hutter (2021) in the context of machine learning.

The idea of the Santa Fe process $(X_t)_{t\in\mathbb{Z}}$ is to decompose each token X_t as a pair of a natural number K_t and an additional bit—which is copied from a certain sequence of bits $(Z_k)_{k\in\mathbb{N}}$ by taking the bit at position K_t. Thus, each text token X_t may be written as a pair

$$X_t = (K_t, Z_{K_t}), \tag{8}$$

where $(Z_k)_{k\in\mathbb{N}}$ is the sequence of bits, called the knowledge, and $(K_t)_{t\in\mathbb{Z}}$ is the sequence of natural numbers, called the narration. To draw attention to a important detail, the extra bit Z_k has the property of making token (k, Z_k) one bit more unpredictable the first time it is observed and its effect vanishes later.

The terms "knowledge" and "narration" were chosen because of a semantic interpretation of Santa Fe processes, discussed by Dębowski (2011, 2021a,b). We may interpret that pairs (K_t, Z_{K_t}) are statements that describe bits of sequence $(Z_k)_{k\in\mathbb{N}}$ in an arbitrary order but in a non-contradictory way. Namely, if statements (k, Z_k) and $(k', Z_{k'})$ describe the same bit $(k = k')$ then they assert the same value $(Z_k = Z_{k'})$. Thus, we may interpret that sequence $(Z_k)_{k\in\mathbb{N}}$ expresses some unbounded immutable general knowledge which is only partially accessed and described in finite texts.

Contrary to intuitions about complex structures, Hilberg's law arises also in the highly simplified setting of Santa Fe processes. In particular, it suffices to assume that narration $(K_t)_{t\in\mathbb{Z}}$ is an IID source with the Zipf distribution (1) and knowledge $(Z_k)_{k\in\mathbb{N}}$ is a sequence of independent fair coin flips, independent of narration $(K_t)_{t\in\mathbb{Z}}$. Under these conditions, process $(X_t)_{t\in\mathbb{Z}}$ is exchangeable and we obtain Heaps' law (2) and Hilberg's law (3) with $\beta_V = \beta_H = 1/\alpha_Z$, similar results having been obtained by Dębowski (2011, 2021a).

Machine learning. There is a rapidly growing body of literature that seeks to explain the neural scaling law and similar quantitative observations concerning foundation models (Belkin et al., 2019; Louart et al., 2018; Bartlett et al., 2020; Bahri et al., 2024; Zhang et al., 2021; Belkin, 2021; Hutter, 2021; Maloney et al., 2022; Roberts et al., 2022; Wei et al., 2022; Sharma and Kaplan, 2022; Michaud et al., 2023; Achille and Soatto, 2026; Barkeshli et al., 2026; Cagnetta et al., 2026). Many of these works involve sophisticated mathematical frameworks, including applications of random matrix theory and techniques from theoretical physics such as Feynman diagrams.

From this perspective, we think that it is important to appreciate works which derive neural scaling from facts as simple as Zipf's law (Hutter, 2021; Maloney et al., 2022; Michaud et al., 2023; Cabannes et al., 2024; Neumann and Gros, 2025; Pan et al., 2025). In the following, we compare our results with several most related works in this area, namely those by Hutter (2021), Michaud et al. (2023), Achille and Soatto (2026), and Cagnetta et al. (2026).

Chronologically, in the first work, Hutter (2021) derived the neural scaling for sample length (5) for the specific model of a Santa Fe process (8) with the IID narration $(K_t)_{t\in\mathbb{Z}}$ distributed according to the exact Zipf-Mandelbrot law $P(K_t = k) \propto k^{-\alpha_Z}$ and the fair-coin-flip knowledge $(Z_k)_{k\in\mathbb{N}}$. Staying in the IID regime, Hutter (2021) was also able to develop laws not only for expectations but also variances.

In the second work, Michaud et al. (2023) considered the same specific Santa Fe process as considered by Hutter (2021). Independently, they assigned it a more semantic interpretation like in the works by Dębowski (2011, 2021a,b), and derived the neural scaling laws (5)–(6) with exponents

$$\gamma_T = \frac{\alpha_Z - 1}{\alpha_Z} = 1 - \beta_V, \qquad \gamma_N = \alpha_Z - 1 = \frac{1 - \beta_V}{\beta_V}. \tag{9}$$

for the Heaps law exponent $\beta = 1/\alpha_Z$.

Identifying the Heaps law exponent β_V with the Hilberg law exponent β_H, for estimate $\beta_H \approx 0.8$ reported by Takahira et al. (2016), formulas (9) yield $\gamma_T = 0.2$ and $\gamma_N = 0.25$. This is quite a difference to the values $\gamma_T \approx 0.095$ and $\gamma_N \approx 0.076$ reported by Kaplan et al. (2020). We note that these estimates were obtained for corpora of a different magnitude—billions of tokens in the case of the work by Takahira et al. (2016) to be contrasted with trillions of tokens in the case of the work by Kaplan et al. (2020).

Formulas (9) imply inequality $\gamma_T < \gamma_N$, which is not confirmed by Kaplan et al. (2020) but corroborated by later experiments (Henighan et al., 2020; Hernandez et al., 2021; Hoffmann et al., 2022; Pearce and Song, 2024; Li et al., 2026). Figure 15 by Michaud et al. (2023) depicts a huge variation in the empirical estimates of parameters γ_T and γ_N across different studies, depending on the cited experiment or even within the particular experimental paper. Thus theoretical derivations of the neural scaling law probably cannot fully explain the diversity of empirically obtained parameter values.

Moreover, a theoretical derivation of the neural scaling law for compute (7) seems to have not been achieved yet. From this perspective, it may be important to mention the work by Achille and Soatto (2026), who sought to link Hilberg's law with time-bounded Kolmogorov complexity (Levin, 1973, 1984). This work does not concern neural scaling (5)–(7) but rather intelligent agents that find themselves under pressure to memorize patterns if they are rewarded for saving time. Still, it may be inspiring for future derivations of the compute law (7).

In the fourth, more recent work, Cagnetta et al. (2026) developed a neural scaling formula from Hilberg's law, which they have hypothesized without any explicit reference to the earlier work in this area. Their reasoning is different than one to be presented in this paper. In particular, Cagnetta et al. (2026) assume additionally a power-law decay of token-token correlation strength, which we also found empirically in language data (Wieczyński and Dębowski, 2025). However, this condition is not necessary in our theoretical derivation

of neural scaling from Hilberg's law. Our assumption (15) is the closest to this power-law decay of token-token correlation out of all our formulas. This assumption, however, is only used to derive Heaps' law from Zipf's for long-range dependent data.

As we can see, there is a lot of partial insight in the machine learning literature. Besides a succinct review of references to much earlier works in quantitative linguistics, mathematics, information theory, and complex systems, which we tried to sketch, a clear mathematical message is due.

3. Statement of Results

In this section, we present the core of our theoretical results. We begin with explaining what is a novel development, then we state the theorems formally, and we end up with discussing the relevance of assumptions.

3.1 Novel Ideas

As we have mentioned in the first sentence of this article, it is increasingly more frequently acknowledged that the neural scaling law observed for foundation models may be caused by the inherent power laws of natural language. Whereas there are some formal results in this in vein discussed in Section 2.3, the present paper develops and strengthens the ideas from an earlier unpublished attempt to derive neural scaling from Hilberg's hypothesis (Dębowski, 2023). The present paper supplies a simple-minded baseline that, in contrast to the work by Dębowski (2023), takes into account also the effect of limiting the amount of training compute. Particular Theorems 1–5 to be discussed in Section 3.2 contribute to an effort of understanding large-scale statistical regularities of language and language models within a common theoretical framework. Their novelty lies in five points listed below.

A hierarchy of statistical laws. The principal conceptual contribution of this paper is the derivation of the chain of implications

$$\text{Zipf} \xRightarrow{\text{Theorems 1–3}} \text{Heaps} \xRightarrow{\text{Theorem 4}} \text{Hilberg} \xRightarrow{\text{Theorem 5}} \text{neural scaling}, \tag{10}$$

which link four regularities that are usually investigated in different research communities. In particular, Zipf's and Heaps' laws originated in quantitative linguistics, Hilberg's law arose in information-theoretic studies of long-range dependence in complex systems, whereas neural scaling laws emerged from empirical investigations in machine learning. Our results suggest that these phenomena need not be viewed as independent observations but may instead form a chain of progressively more general implications.

Increasing generality of assumptions. Indeed, the individual propositions are established under assumptions of different strength. Theorems 2 and 3 concern IID and some pseudo-mixing processes, respectively. Theorem 4 is proved for stationary Santa Fe processes, which provide a mathematically tractable model exhibiting both Zipfian and Hilberg-type behavior. Finally, Theorem 5 is formulated in an information-theoretic framework that can accommodate also certain non-stationary sources. Although no single theorem covers all settings of practical interest, the sequence of results demonstrates that more global statistical laws can be deduced from increasingly weaker assumptions.

Differential forms of Heaps' and Hilberg's laws. A conceptual novelty of this paper is the use of differential versions of Heaps' and Hilberg's laws. Classical formulations describe the growth of cumulative quantities such as vocabulary size or excess entropy. Deriving neural scaling, however, requires controlling the marginal gains obtained from additional data. This motivates the introduction of differential laws that quantify incremental growth rates. These laws constitute the natural intermediate quantities linking linguistic statistics to learning curves.

Implications for neural scaling. Theorem 5 suggests that neural scaling may be understood as a consequence of the statistical structure of the data, rather than solely as a property of particular optimization algorithms, architectures, or parameterizations. The theorem does not predict the exact scaling exponents observed in practice. Instead, it provides information-theoretic lower bounds on the achievable rates of improvement with increasing data, model complexity, and computational resources. In this sense, the result may be viewed as a structural constraint on learning curves that holds independently of many implementation details.

Information-theoretic modeling of resources. The assumptions concerning model size (22) and training compute (23) are stated in terms of entropy bounds rather than specific computational architectures. This choice reflects the idea that any resource constraint ultimately limits the amount of information that can be stored, processed, or inferred. By expressing the assumptions at this level of abstraction, the resulting bounds apply simultaneously to a wide range of learning systems.

3.2 Formal Propositions

This paper strives to identify mathematical implications in the most general setting and to make interdisciplinary connections. The main contributions of the present paper are organized as the chain of derivations (10). The proofs of the respective Theorems 1–5 can be found in Appendix B, whereas the discussion of the strength of these formal propositions is deferred to Section 3.3.

Now let us announce these results formally. First of all, Zipf's law with the Mandelbrot correction (1) implies structural inequalities (12) and (13) that matter in further developments.

Theorem 1 *Let a sequence $(p_k)_{k\in\mathbb{N}}$, where $p_k \geq 0$ and $\sum_{k\in\mathbb{N}} p_k = 1$. Suppose that we have Zipf's law with the Mandelbrot correction*

$$A_Z^- k^{-1/\beta} \leq p_k \leq A_Z^+ k^{-1/\beta} \tag{11}$$

for some constants $\beta \in (0,1)$, $A_Z^\pm > 0$, and all $k \in \mathbb{N}$. Then for

$$A_S^+ := \left(\frac{1}{A_Z^+}\right)^{1-\beta}, \qquad A_S^- := \frac{A_Z^-}{1/\beta - 1}\left(\frac{1}{A_Z^+}\right)^{1-\beta},$$

and all $p > 0$, we have two structural inequalities

$$\sum_{k\in\mathbb{N}} \mathbf{1}\{p_k \geq p\} \leq A_S^+ p^{-\beta}, \tag{12}$$

$$\sum_{k\in\mathbb{N}} p_k \mathbf{1}\{p_k \leq p\} \geq A_S^- p^{1-\beta}. \tag{13}$$

Bounds similar to (12) and (13) were discussed by Karlin (1967).

Having the first structural bound for sequences of independent identically distributed random variables, we derive an upper bound for the differential Heaps law.

Theorem 2 *For an IID process $(K_t)_{t\in\mathbb{Z}}$ over natural numbers, let $p_k := P(K_0 = k)$. Suppose that the structural inequality (12) holds for some constants $\beta \in (0,1)$, $A_S^+ > 0$, and all $p > 0$. Then for*

$$A_V^+ = \Gamma(1-\beta)A_S^+,$$

we have an upper bound for the differential Heaps law

$$\frac{V(\mathcal{K}_{t+1}^{t+s} \setminus \mathcal{K}_1^t)}{s} \leq \frac{A_V^+}{t^{1-\beta}}. \tag{14}$$

By contrast, if the process is stationary with the second structural bound and appropriately bounded conditional probabilities, then a lower bound for the differential Heaps law is satisfied.

Theorem 3 *For a stationary process $(K_t)_{t\in\mathbb{Z}}$ over natural numbers, let $p_k := P(K_0 = k)$ and $p_k(t) := P(K_t = k|K_0 = k)$. Suppose that the structural inequality (13) and condition*

$$p_k(t) \leq A_M p_k^\delta \tag{15}$$

hold for some constants $\beta \in (0,1)$, $\delta \in (0,1]$, $A_S^-, A_M > 0$, all $k, t \in \mathbb{N}$, and all $p > 0$. Then for

$$A_V^- := \frac{A_S^-}{2}\left(\frac{1}{4A_M}\right)^{\frac{1-\beta}{\delta}},$$

we have a lower bound for the differential Heaps law

$$\frac{V(\mathcal{K}_{t+1}^{t+s} \setminus \mathcal{K}_1^t)}{s} \geq \frac{A_V^-}{(t+s)^{\frac{1-\beta}{\delta}}}. \tag{16}$$

For IID processes, condition (15) holds with $A_M = 1$ and $\delta = 1$. For finite-state processes, it holds with $A_M < \infty$ and also $\delta = 1$, as it will be discussed further.

The lower bound for the differential Heaps law implies the analogous bound for the differential Hilberg law in case of some stationary Santa Fe processes.

Theorem 4 *Consider a process $(X_t)_{t\in\mathbb{Z}}$ and a one-to-one mapping ϕ such that there holds a Santa Fe decomposition*

$$\phi(X_t) = (K_t, Z_{K_t}), \tag{17}$$

where narration $(K_t)_{t\in\mathbb{Z}}$ is a stationary ergodic process over natural numbers and knowledge $(Z_k)_{k\in\mathbb{N}}$ is a sequence of random variables independent of narration $(K_t)_{t\in\mathbb{Z}}$ with conditional entropies

$$H(Z_k|Z_1^{k-1}) \geq A_K. \tag{18}$$

for a constant $A_K > 0$. If the marginal entropy satisfies

$$H(X_t) < \infty \tag{19}$$

then we have inequality

$$\frac{H(X_{k+1}^{k+s}|X_1^t)}{s} - h \geq \frac{A_K V(\mathcal{K}_{k+1}^{k+s} \setminus \mathcal{K}_1^t)}{s}, \tag{20}$$

where h is the entropy rate defined as (4).

Inequality (19) holds for all stationary processes over a finite alphabet.

Finally, using the lower bound for the differential Hilberg law, we derive a special lower bound for the neural scaling law.

Theorem 5 *Let $(X_t)_{t\in\mathbb{Z}}$ be an arbitrary process such that we have a lower bound for the differential Hilberg law*

$$\sup_{k\geq t} \frac{H(X_{k+1}^{k+s}|X_1^t)}{s} - h \geq \frac{A_H^-}{(t+s)^{1-\beta}} \tag{21}$$

for some constants $h \in (0,\infty)$, $\beta \in (0,1)$, and $A_H^- > 0$. Let Q be a random probability measure that satisfies bounds

$$H(Q) \leq A_H^- n, \tag{22}$$

$$H(Q|X_1^t) \leq A_H^- c. \tag{23}$$

Define also the maximal test sample length

$$s_{\max}(t,n,c) := \min\left\{ \frac{t\sqrt{\frac{ct^{-\beta}}{1-\beta}}}{1-\sqrt{\frac{ct^{-\beta}}{1-\beta}}}, \left(\frac{n}{1-\beta}\right)^{\frac{1}{\beta}} \right\}. \tag{24}$$

Then for test sample lengths $s \leq s_{\max}(t,n,c)$ we have the lower bound for the neural scaling

$$\begin{aligned} &\sup_{k\geq t} \frac{\mathbf{E}(-\log Q(X_{k+1}^{k+s}))}{s} - h \\ &\geq A_H^- \max\left\{ \frac{1}{t^{1-\beta}} \left(\frac{1-\sqrt{\frac{ct^{-\beta}}{1-\beta}}}{1+\sqrt{\frac{ct^{-\beta}}{1-\beta}}}\right)^{1-\beta} - \frac{c}{t}\left(\frac{1-\sqrt{\frac{ct^{-\beta}}{1-\beta}}}{2\sqrt{\frac{ct^{-\beta}}{1-\beta}}}\right), \frac{2^\beta - 1 + \beta}{2}\left(\frac{1-\beta}{n}\right)^{\frac{1-\beta}{\beta}} \right\}. \end{aligned} \tag{25}$$

3.3 Relevance of the Assumptions

Overall, our results support a view in which neural scaling laws are not isolated empirical phenomena but the final link in a broader chain of statistical regularities that connect token frequencies, vocabulary growth, long-range predictability, and learnability. Let us discuss the relevance of assumptions of Theorems 1–5.

Long-range dependence. Theorems 1–3 deal with a subclass of stationary processes. We notice that for $\delta = 1$, condition (15) reduces to a uniform upper bound on the pointwise mutual information between any two tokens. That condition is satisfied for IID and finite-state processes (Dębowski, 2021a, Theorem 11.3). It is implied by condition $\psi^*(1) < \infty$ in the terminology of Bradley (2005). We also note that any stationary strongly mixing process $(K_t)_{t\in\mathbb{Z}}$ satisfies identity

$$\lim_{t\to\infty} p_k(t) = p_k.$$

By contrast, the proof of Theorem 3 allows for taking also $\delta < 1$, which may be useful for long-range dependent processes. In particular, natural language may require $\delta < 1$ because of Taylor's law (Kobayashi and Tanaka-Ishii, 2018; Tanaka-Ishii, 2021) and burstiness of words (Altmann et al., 2009).

Validity of the Santa Fe decomposition. Theorem 4 deals with another subclass of stationary processes and is a proxy for more general results. Its assumptions are motivated by the intuition that language and other information-bearing signals communicate discrete units of knowledge. In the Santa Fe decomposition, the random variables $(Z_k)_{k\in\mathbb{N}}$ represent such units, whereas the process $(K_t)_{t\in\mathbb{Z}}$ determines which unit is being referred to at a given position. The assumption (18) states that each unit contributes a non-negligible amount of novel information and therefore cannot be reconstructed from previous units.

Although this representation is highly idealized, it captures a natural distinction between references and facts. For example, words, phrases, or textual passages often refer repeatedly to the same underlying entities, events, or concepts. In this interpretation, narration $(K_t)_{t\in\mathbb{Z}}$ determines the sequence of references, whereas knowledge $(Z_k)_{k\in\mathbb{N}}$ describes the latent facts being referenced. The growth of predictive information then reflects the gradual discovery of previously unknown facts rather than merely the repetition of surface forms.

The Santa Fe decomposition should therefore be viewed not as a literal generative model of language but as a mathematical abstraction of semantic structure. Establishing analogous results under less restrictive assumptions is an important open problem. One would like to replace the explicit decomposition (17) by more intrinsic information-theoretic conditions that characterize the existence of persistent latent variables or reusable knowledge fragments. Such a development can extend the connection between Heaps' and Hilberg's laws far beyond the specific class of Santa Fe processes.

In view of so called theorems about facts and words (Dębowski, 2021a, Section 8.4), we suppose that one can replace decomposition (17) in Theorem 4 by a more general assumption such as strong non-ergodicity. According to (Dębowski, 2021a, Definition 5.16), a stationary process is called strongly non-ergodic if its shift-invariant σ-field is non-atomic. Equivalently, it means that there is a continuous real random variable that can be estimated arbitrarily well from any sufficiently long sample (Dębowski, 2021a, Theorem 8.12). This is a natural generalization since Santa Fe processes are strongly non-ergodic. We also note that non-ergodicity is not a necessary condition to obtain Hilberg's law. There are simple modifications of Santa Fe processes and other constructive toy examples that allow for Hilberg's law in mixing or non-mixing ergodic settings, see Dębowski (2012, 2020, 2021b, 2023). We need more examples and results in this vein, also for numerical experiments with artificial data, but their detailed discussion exceeds a reasonable page limit of this article.

Hallucinated worlds vs. arbitrary words. As we can see, after discerning a general implication from Hilberg's law to neural scaling, the burden of explanations shifts to stating why Hilberg's hypothesis may be sound. In some settings, Hilberg's hypothesis can be viewed as Zipf's law for tokens that are internally random enough like in the Santa Fe processes (8) or their lossless encoding (17). The concrete interpretation of these tokens can be varied. In Section 2.2, we suggested a semantic interpretation, where random variables (8) are logically consistent propositions that describe a certain random state of world. This world can be "coherently hallucinated", so as to speak in modern terms. Such interpretation dates back to Dębowski (2011) and was defended independently by Michaud et al. (2023).

Yet, there is a more simple-minded interpretation. Decomposition (17), somewhat more abstract than (8), suggests that Hilberg's law is equivalent to Zipf's law for orthographic words if those have sufficiently arbitrary shapes. Arbitrariness of word shapes is one of the classical tenets of linguistics (Dingemanse et al., 2015). However, proving that Hilberg's law is tantamount to Zipf's law for empirically detectable and partly random word-like structures requires an inequality opposite to (20). Turning (20) into a sandwich bound necessitates a longer excursion to universal coding and involves results that exceed our present statement of so called theorems about facts and words (Dębowski, 2021a, Section 8.4), see also Dębowski (2011, 2021b). For this reason, we postpone this theoretical development to another article. We want to stress, however, that the explanation of Hilberg's hypothesis by arbitrariness of word shapes does not exclude a semantic interpretation of this law if the information content of the vocabulary and the information content of other more abstract knowledge are of a similar magnitude. A one-to-one correspondence between words and concepts may be the desired optimum.

Bound on the parameter count. The entropy of a discrete object cannot be essentially larger than the description length of this object. Thus limiting the amount of a certain resource implies constraining the respective entropy. In view of this, bound (22) on the entropy of parameters seems uncontroversial. However, the reality may be more complicated. Observation $\gamma_N > \gamma_T$ for the simple parameter count reported by Kaplan et al. (2020) seems to contradict Theorem 5 in contrast to inequality $\gamma_N < \gamma_T$ reported by Hoffmann et al. (2022). We note that these inconsistent observations might be potentially explained out because the number of trainable parameters is an imperfect proxy for the amount of information represented by a model. A real-valued parameter can in principle encode an arbitrarily large number of bits and different foundation models may use a varying amount of digits of the binary expansion per real-valued parameter (Liu et al., 2025). Thus a naive parameter count may overestimate or underestimate the effective capacity of a real-valued neural network and one should rather use universally comparable measures of model complexity. Determining the true number of degrees of freedom of a foundation model may not be so straightforward.

Bound on compute. Bound (23) is also formulated in a purely information-theoretic sense but its justification is more fragile. We model the time complexity of computing model Q from training data X_1^t by conditional entropy, whereas using time-bounded Kolmogorov complexity (Levin, 1973, 1984) should be more proper for this goal. Inequality (23) may make sense if Q is a stochastic function of X_1^t, which is the less predictable from X_1^t the more compute is applied. Some ideas of Achille and Soatto (2026) may be useful to rectify this issue, at the expense of significantly complicating the proof of the time-bounded analogue

of Theorem 5. To keep this paper focused on purely Shannon information measures, we relegate this problem to future research.

The worst-case expected loss. Theorem 5 suggests that the natural theoretical definition of the model cross entropy for the neural scaling laws (5)–(7) in long-range dependent processes is

$$h(s,t,n,c) := \sup_{k \geq t} \frac{\mathbf{E}(-\log Q(X_{k+1}^{k+s}))}{s}. \tag{26}$$

where the foundation model Q has been tested on data X_{k+1}^{k+s} and trained on data X_1^t with the amount of parameters n and the amount of compute c. In this setting, $h(s,t,n,c)$ becomes a sort of the worst-case expected cross entropy rate. Actually, setting (26) may model desirable testing conditions theoretically better than using

$$h(s,t,n,c) := \frac{\mathbf{E}(-\log Q(X_{t+1}^{t+s}))}{s}, \tag{27}$$

since one usually wants the test data to be maximally different from the training data while still coming from the same generative source. We notice that the difference between (26) and (27) does not arise for memoryless sources but it can be quite important for general long-range dependent stationary or non-stationary data.

Inequalities and interactions. Because a more appropriate information-theoretic bound on compute is needed, determining a more precise relationship between the Hilberg exponent and the neural scaling exponents remains an important empirical and theoretical problem. Still, let us make the following remark. Suppose that conditions (5)–(7) hold with a finite compute bound $c \ll t^\beta$ rather than $c = \infty$. If additionally the neural scaling law bound (25) is true then we obtain asymptotic inequalities

$$\gamma_T \leq 1 - \beta, \qquad\qquad \gamma_N \leq \frac{1-\beta}{\beta},$$

which generalize equalities (9) known for exchangeable Santa Fe processes (those with an IID narration). Besides these inequalities, which hold in the general setting of arbitrary processes, Theorem 5 predicts quite complicated interactions between parameters t, n, and c and the cross entropy of the foundation model. Having such a theoretical baseline may be more informative for the research of neural scaling laws than fitting theory-free regressions to empirical data.

This remark ends our discussion. After going through the paper conclusion, the reader is invited to consult Appendices A and B. The appendices contain the main mathematical derivations as well as some general results of quantitative linguistic importance.

4. Conclusion

Extending prior results in machine learning such as those by Hutter (2021) and Michaud et al. (2023) and earlier results in information theory and quantitative linguistics (Dębowski, 2006, 2011, 2021a, 2023), we have supplied and inspected a deductive chain that connects Zipf's law to neural scaling. By isolating the discrete steps that go through Heaps' law

and Hilberg's hypothesis and by giving explicit assumptions needed for each step, we have attempted to clarify which aspects of natural language are responsible for the power-law behavior observed in foundation models. Our results show that once vocabulary growth exhibits a power-law growth and once block entropy inherits this scaling, then the constraints imposed by limited data, parameters, and compute produce the neural scaling law.

The direct reduction of the neural scaling law to Hilberg's hypothesis shifts attention to the deeper question of why power-law scaling of entropy arises in natural language at all. Broadly speaking, the results of this paper suggest that statistical laws of language and scaling laws of learning should not be studied in isolation. A comprehensive theory would explain not only individual laws but also the relationships among them, the numerical values of their exponents, and their variation. We hope that the framework developed here provides a baseline for such investigation and that future work will refine the idealizations which we have identified.

Acknowledgments

Several paragraphs of the article were drafted in a dialog between the author and ChatGPT (`https://chatgpt.com/`) and critically post-edited by the author. A prior version of the article was reviewed by the Stanford Agentic Reviewer (`https://paperreview.ai/`) and its suggestions were partially followed by the author.

This work received no external funding.

Appendix A. Remarks on Counting Bits and Types

This appendix is a gentle mathematical introduction that precedes the proper proofs of Theorems 1–5. It is split into four parts. Part A.1 treats general analogies for Shannon entropy and expected cardinality. Part A.2 discusses the rates of entropy and expected cardinality for arbitrary (non-stationary) processes. Part A.3 handles the block entropy and the expected block cardinality for stationary processes. It also introduces spectrum elements and bounds the rate of hapaxes. Part A.4 analyzes the spectrum elements for IID processes.

A.1 Fundamental Inequalities

We will be developing parallel results for the Shannon entropy and the expected number of types. Our reasoning is based on algebraically simple but systematically refined information-theoretic considerations. The general spirit of the I-measure will be accompanying them (Hu, 1962). For the textbook treatment and the background, we refer to Cover and Thomas (2006) and Yeung (2002).

In general, both Shannon entropy and expected cardinality are subadditive and enjoy the triangle inequality. The formulas for Shannon entropy are well known.

Proposition 6 (subadditivity) *For random variables X, Y, Q, we have*

$$H(X, Y|Q) \leq H(X|Q) + H(Y|Q). \tag{28}$$

Proof The claim follows by the chain rule

$$H(X,Y|Q) = H(X|Q) + H(Y|Q,X)$$

and inequality $H(Y|Q,X) \leq H(Y|Q)$. ■

Proposition 7 (triangle inequality) *For random variables X, Y, Q, we have*

$$H(X|Y) \leq H(X|Q) + H(Q|Y). \tag{29}$$

Proof The claim follows by the chain rule

$$H(X,Q|Y) = H(X|Y,Q) + H(Q|Y)$$

and inequalities $H(X|Y) \leq H(X,Q|Y)$ and $H(X|Y,Q) \leq H(X|Q)$. ■

The above familiar formulas for Shannon entropy have their counterparts for expected cardinality if we replace pairing of variables with the union of sets and conditioning with the set difference.

Proposition 8 (subadditivity) *For random sets $\mathcal{X}, \mathcal{Y}, \mathcal{Q}$, we have*

$$V(\mathcal{X} \cup \mathcal{Y} \setminus \mathcal{Q}) \leq V(\mathcal{X} \setminus \mathcal{Q}) + V(\mathcal{Y} \setminus \mathcal{Q}). \tag{30}$$

Proof The claim follows by the chain rule

$$V(\mathcal{X} \cup \mathcal{Y} \setminus \mathcal{Q}) = V(\mathcal{X} \setminus \mathcal{Q}) + V(\mathcal{Y} \setminus \mathcal{Q} \cup \mathcal{X})$$

and inequality $V(\mathcal{Y} \setminus \mathcal{Q} \cup \mathcal{X}) \leq V(\mathcal{Y} \setminus \mathcal{Q})$. ■

Proposition 9 (triangle inequality) *For random sets $\mathcal{X}, \mathcal{Y}, \mathcal{Q}$, we have*

$$V(\mathcal{X} \setminus \mathcal{Y}) \leq V(\mathcal{X} \setminus \mathcal{Q}) + V(\mathcal{Q} \setminus \mathcal{Y}). \tag{31}$$

Proof The claim follows by the chain rule

$$V(\mathcal{X} \cup \mathcal{Q} \setminus \mathcal{Y}) = V(\mathcal{X} \setminus \mathcal{Y} \cup \mathcal{Q}) + V(\mathcal{Q} \setminus \mathcal{Y})$$

and inequalities $V(\mathcal{X} \setminus \mathcal{Y}) \leq V(\mathcal{X} \cup \mathcal{Q} \setminus \mathcal{Y})$ and $V(\mathcal{X} \setminus \mathcal{Y} \cup \mathcal{Q}) \leq V(\mathcal{X} \setminus \mathcal{Q})$. ■

There is also an important bridging inequality for cross entropy of random measures.

Proposition 10 (source coding) *For a random probability measure Q applied to another random variable X, we have*

$$\mathbf{E}(-\log Q(X)) \geq H(X|Q). \tag{32}$$

Proof We have

$$\begin{aligned}\mathbf{E}\left(-\log Q(X)\right) &= \mathbf{E}\,\mathbf{E}\left(-\log Q(X)|Q\right)\\ &= \mathbf{E}\,\mathbf{E}\left(-\log P(X|Q)|Q\right) + \mathbf{E}\,\mathbf{E}\left(\log\frac{P(X|Q)}{Q(X)}\middle|Q\right)\\ &\geq \mathbf{E}\,\mathbf{E}\left(-\log P(X|Q)|Q\right)\\ &= \mathbf{E}\left(-\log P(X|Q)\right) = H(X|Q),\end{aligned}$$

where we used the law of total expectation and the fact that the Kullback-Leibler divergence is non-negative. ■

A.2 Arbitrary Processes

In the following, we consider arbitrary processes (possibly non-stationary) over a countable alphabet. Let us inspect the rates of the Shannon entropy and the expected cardinality for these objects. We may define the rates as follows.

Definition 11 *Let $(X_t)_{t\in\mathbb{N}}$ be an arbitrary process. We define the entropy rate*

$$h := \inf_{s\in\mathbb{N}} \sup_{k\geq 0} \frac{H(X_{k+1}^{k+s})}{s}. \tag{33}$$

Definition 12 *Let $(K_t)_{t\in\mathbb{N}}$ be an arbitrary process. We define the expected cardinality rate*

$$v := \inf_{s\in\mathbb{N}} \sup_{k\geq 0} \frac{V(\mathcal{K}_{k+1}^{k+s})}{s}. \tag{34}$$

The Fekete lemma (Fekete, 1923) states that any sequence $(a_s)_{s\in\mathbb{N}}$ that satisfies the condition of subadditivity $a_{s+r} \leq a_s + a_r$ has the rate that can be equivalently defined as

$$\lim_{s\to\infty} \frac{a_s}{s} = \inf_{s\in\mathbb{N}} \frac{a_s}{s}. \tag{35}$$

Hence having subadditivity (28) and (30), we can show that rates (33) and (34) can be equivalently expressed somewhat differently.

Proposition 13 *Let $(X_t)_{t\in\mathbb{N}}$ be a process such that $h < \infty$ for the entropy rate (33). For an arbitrary $t \geq 0$, we have*

$$h = \lim_{s\to\infty} \sup_{k\geq t} \frac{H(X_{k+1}^{k+s}|X_1^t)}{s} = \inf_{s\in\mathbb{N}} \sup_{k\geq t} \frac{H(X_{k+1}^{k+s}|X_1^t)}{s}.$$

Proof By inequality (28), we notice subadditivity

$$\begin{aligned}\sup_{k\geq t} H(X_{k+1}^{k+s+r}|X_1^t) &\leq \sup_{k\geq t}\left[H(X_{k+1}^{k+s}|X_1^t) + H(X_{k+s+1}^{k+s+r}|X_1^t)\right]\\ &\leq \sup_{k\geq t} H(X_{k+1}^{k+s}|X_1^t) + \sup_{k\geq t} H(X_{k+1}^{k+r}|X_1^t).\end{aligned}$$

Hence by the Fekete lemma (Fekete, 1923), we obtain

$$h(t) := \lim_{s\to\infty} \sup_{k\geq t} \frac{H(X_{k+1}^{k+s}|X_1^t)}{s} = \inf_{s\in\mathbb{N}} \sup_{k\geq t} \frac{H(X_{k+1}^{k+s}|X_1^t)}{s}.$$

It suffices to show $h(t) = h$. Since $h < \infty$, we have $\sup_{t\geq 0} H(X_t) < \infty$. Hence, by the chain rule and a simple calculation, we obtain a uniform bound

$$\left|H(X_{k+1}^{k+s}) - H(X_{k+t+1}^{k+t+s})\right| \leq B(t) < \infty.$$

For the same reason, we also have

$$\left|H(X_{k+t+1}^{k+t+s}) - H(X_{k+t+1}^{k+t+s}|X_1^t)\right| \leq D(t) < \infty.$$

Chaining these two sandwich bounds and taking the supremums over k and infimums over s, we infer $h(t) = h$. ■

Proposition 14 *Let $(K_t)_{t\in\mathbb{N}}$ be an arbitrary process. For an arbitrary $t \geq 0$, we have*

$$v = \lim_{s\to\infty} \sup_{k\geq t} \frac{V(\mathcal{K}_{k+1}^{k+s} \setminus \mathcal{K}_1^t)}{s} = \inf_{s\in\mathbb{N}} \sup_{k\geq t} \frac{V(\mathcal{K}_{k+1}^{k+s} \setminus \mathcal{K}_1^t)}{s}.$$

Proof Mutatis mutandis, the same as the proof of Proposition 13. ■

In view of the above two propositions, we have

$$\sup_{k\geq t} H(X_{k+1}^{k+s}|X_1^t) \geq hs,$$
$$\sup_{k\geq t} V(\mathcal{K}_{k+1}^{k+s} \setminus \mathcal{K}_1^t) \geq vs.$$

A.3 Stationary Processes

In case of stationary processes, the translation symmetry yields some further results. By a general result, any stationary process with natural number indices can be uniquely extended to the stationary process with integer indices (Kallenberg, 1997, Lemma 9.2). Thus, for stationary processes $(X_t)_{t\in\mathbb{Z}}$ and $(K_t)_{t\in\mathbb{Z}}$, we denote the the block Shannon entropy and the expected number of types

$$H(t) := H(X_1^t),$$
$$V(t) := V(\mathcal{K}_1^t).$$

Let us define the finite difference operator $\Delta f(t) := f(t+1) - f(t)$. Function $f(t)$ is called positive, growing, and concave if $f(t) \geq 0$, $\Delta f(t) \geq 0$, and $\Delta^2 f(t) \leq 0$, respectively.

We have two analogous statements. We recall the results for entropy to apply them by analogy to the expected cardinality.

Proposition 15 *Let a stationary process $(X_t)_{t\in\mathbb{Z}}$. For $t \in \mathbb{N}$, we claim that:*

1. *Function $t \mapsto H(t)$ is positive, growing, and concave and $H(0) = 0$.*
2. *Function $t \mapsto H(t)/t$ is decreasing.*
3. *We have $h = \lim_{t\to\infty} H(t)/t \geq 0$.*

Proof See Dębowski (2021a, Section 5.2). In general, we have

$$\begin{aligned} H(t) &= H(X_1^t) \geq 0, \\ \Delta H(t) &= H(X_{t+1}|X_1^t) \geq 0, \\ \Delta^2 H(t) &= -I(X_0; X_{t+1}|X_1^t) \leq 0. \end{aligned}$$

Function $t \mapsto \Delta H(t)$ is decreasing since $\Delta^2 H(t) \leq 0$. Hence

$$\begin{aligned} \frac{H(t+1)}{t+1} = \frac{\sum_{i=0}^{t} \Delta H(i)}{t+1} &\leq \frac{\sum_{i=0}^{t-1} \Delta H(i) + \frac{\sum_{i=0}^{t-1} \Delta H(i)}{t}}{t+1} \\ &= \frac{\sum_{i=0}^{t-1} \Delta H(i)}{t} = \frac{H(t)}{t}. \end{aligned}$$

Thus function $t \mapsto H(t)/t$ is decreasing and limit $\lim_{t\to\infty} H(t)/t$ exists. By stationarity, it equals h defined in (33). ■

Proposition 16 *Let a stationary process $(K_t)_{t\in\mathbb{Z}}$. For $t \in \mathbb{N}$, we claim that:*

1. *Function $t \mapsto V(t)$ is positive, growing, and concave and $V(0) = 0$.*
2. *Function $t \mapsto V(t)/t$ is decreasing and $V(t)/t \leq 1$.*
3. *We have $v = \lim_{t\to\infty} V(t)/t = 0$.*

Proof The proof is analogous to the proof of Proposition 15 except for the last claim. In particular, we have

$$\begin{aligned} V(t) &= V(\mathcal{K}_1^t) \geq 0, \\ \Delta V(t) &= V(\mathcal{K}_{t+1} \setminus \mathcal{K}_1^t) \geq 0, \\ \Delta^2 V(t) &= -V(\mathcal{K}_0 \cap \mathcal{K}_{t+1} \setminus \mathcal{K}_1^t) \leq 0. \end{aligned}$$

Inequality $V(t)/t \leq 1$ is obvious since the number of type is not greater than the number tokens. The proof of $\lim_{t\to\infty} V(t)/t = 0$ is as follows. Without loss of generality, we assume that the alphabet is the set of natural numbers. Consider a real function $g(t) \geq 0$. Generalizing an idea used by Khmaladze (1988) for IID processes, we observe

$$V(t) = \mathbf{E} \sum_{k=1}^{\infty} \mathbf{1}\{k \in \mathcal{K}_1^t\} = \sum_{k=1}^{\infty} P(k \in \mathcal{K}_1^t) \leq g(t) + \sum_{k>g(t)} P(k \in \mathcal{K}_1^t),$$

whereas the union bound and stationarity yield

$$P(k \in \mathcal{K}_1^t) = P(K_1 = k \vee \ldots \vee K_t = k) \leq tP(K_0 = k).$$

Thus we have bound

$$V(t) \leq g(t) + tP(K_0 > g(t))$$

that holds for any function $g(t) \geq 0$. In particular, for an $\epsilon > 0$, we may take $g(t) = \epsilon t/2$. For all sufficiently large t, we observe $P(K_0 > g(t)) \leq \epsilon/2$. Hence, for these t, we derive $V(t)/t \leq \epsilon$. By arbitrariness of ϵ, we derive $\lim_{t\to\infty} V(t)/t = 0$. ■

Denote the set of types that occur exactly m times in sequence K_1^t as

$$\mathcal{K}_1^t(m) := \left\{ k : \sum_{i=1}^t \mathbf{1}\{K_i = k\} = m \right\}.$$

Besides the expected total number of types $V(t)$, let us introduce the spectrum elements

$$V(t|m) = V(\mathcal{K}_1^t(m)), \tag{36}$$

being the expected numbers of types that occur exactly m times (Karlin, 1967; Khmaladze, 1988; Baayen, 2001). The bar in notation $V(t|m)$ is a separator of arguments, not related to conditioning. In particular, we have $V(t|m) \geq 0$ and $\sum_{m\in\mathbb{N}} V(t|m) = V(t)$ so the relative spectrum $m \mapsto V(t|m)/V(t)$ is a probability distribution.

The first spectrum element $V(t|1)$ is the expected number of types that occur once. These one-time elements are called *hapax legomena* in Greek or, succinctly, hapaxes in English. In the stationary case, there is a general upper bound for the expected number of hapaxes in terms of the difference of the total number of types.

Proposition 17 *For a stationary process $(K_t)_{t\in\mathbb{Z}}$, we have*

$$\frac{V(t|1)}{t} \leq \Delta V\left(\left\lceil \frac{t}{2} \right\rceil\right). \tag{37}$$

Proof For a stationary process, we observe that

$$\begin{aligned} V(t|1) &= \sum_{i=1}^t P\left(K_i \notin \mathcal{K}_1^{i-1} \cup \mathcal{K}_{i+1}^t\right) \\ &\leq \sum_{i=1}^t \min\left\{P\left(K_i \notin \mathcal{K}_1^{i-1}\right), P\left(K_i \notin \mathcal{K}_{i+1}^t\right)\right\} \\ &= \sum_{i=1}^t \min\left\{\Delta V(i-1), \Delta V(t-i)\right\} \leq t\Delta V\left(\left\lceil \frac{t}{2} \right\rceil\right) \end{aligned}$$

since $t \mapsto \Delta V(t)$ is decreasing. ■

A.4 IID Processes

So far, the statements for Shannon entropy and expected cardinality were mirror-like. However, for more specific processes, the analogy between these two functionals is rather as follows: The expected cardinality applies to IID processes in a similar fashion as the Shannon entropy applies to exchangeable Santa Fe processes (those with an IID narration). This analogy will be applied in Appendix B.4. Now we consider the expected cardinality for IID processes, being simpler to analyze. We notice that the expected cardinality for IID sources enjoys further special properties. Namely, it is a Hausdorff sequence—a discrete-time analogue of a Bernstein function. This observation nicely complements the results by Karlin (1967) for IID and Poisson cases.

The development is as follows.

Definition 18 *A sequence $v : \mathbb{N} \cup \{0\} \to \mathbb{R}$ is called a Hausdorff sequence if $v(t) \geq 0$ and $(-1)^{m+1}\Delta^m v(t) \geq 0$ for all $m \in \mathbb{N}$ and $n \in \mathbb{N} \cup \{0\}$. A sequence $u : \mathbb{N} \cup \{0\} \to \mathbb{R}$ is called completely alternating if $u(t) = \Delta v(t)$ for a certain Hausdorff sequence $v(t)$. We call a sequence $v(t)$ standard if $v(0) = 0$ and $\Delta v(0) = 1$.*

Remark: The name *Hausdorff sequence* is non-standard itself. We have coined it by an analogy to the standard term *Bernstein function*, which is the continuous time-analog, applying derivatives rather than differences (Schilling et al., 2010).

Any Hausdorff sequence can be expressed as a convex combination of sequences $t \mapsto p^{-1}\left(1-(1-p)^t\right)$ for varying $p > 0$. This result is known as the Hausdorff moment theorem (Hausdorff, 1923). It is a discrete-time analogue of the Bernstein theorem on completely monotone functions (Bernstein, 1928), also known as the Lévy-Khintchine representation in probability (Lévy, 1948).

Proposition 19 *A sequence $v : \mathbb{N} \cup \{0\} \to \mathbb{R}$ is a Hausdorff sequence if and only if there exists a unique non-negative measure $\tilde{v}$ on $[0, 1]$ such that*

$$v(t) = t\tilde{v}(\{0\}) + \int_{(0,1)} \frac{1-(1-p)^t}{p} d\tilde{v}(p) + \tilde{v}(\{1\}), \tag{38}$$

$$\Delta v(t) = \tilde{v}(\{0\}) + \int_{(0,1)} (1-p)^t d\tilde{v}(p). \tag{39}$$

Remark: We call measure $\tilde{v}$ the Hausdorff measure of sequence $v(t)$. A Hausdorff sequence $v(t)$ is standard if and only if $\tilde{v}(\{1\}) = 0$ and the measure $\tilde{v}$ is a probability measure.

Proof See Hausdorff (1923) and Akhiezer (2021). ■

Now let us define another non-standard object. The bar in notation $v(t\|m)$ is a separator of arguments, unrelated to conditioning.

Definition 20 *For an arbitrary sequence $v : \mathbb{N} \cup \{0\} \to \mathbb{R}$, we define its Taylor elements*

$$v(t\|m) := (-1)^{m+1}\binom{t}{m}\Delta^m v(t-m), \qquad 1 \leq m \leq t. \tag{40}$$

We notice that $v(t\|m) \geq 0$ if and only if sequence $v(t)$ is a Hausdorff sequence.

Proposition 21 *The Taylor elements satisfy consistency conditions*

$$\sum_{m=1}^{\infty} v(t\|m) = v(t), \qquad\qquad \sum_{m=1}^{\infty} m v(t\|m) = t. \tag{41}$$

if and only if sequence $v(t)$ is standard.

Proof The claim follows by the Newton formula $(1-\Delta)^r = \sum_{m=0}^{r} \binom{r}{m}(-\Delta)^m$, written as

$$v(t-r) = \sum_{m=0}^{r} (-1)^m \binom{r}{m} \Delta^m v(t-m), \tag{42}$$

which resembles the Taylor expansion. It suffices to consider (42) for $r=t$ and $r=t-1$. In the first case, we obtain

$$v(t) - \sum_{m=1}^{t} v(t\|m) = \sum_{m=0}^{t} (-1)^m \binom{t}{m} \Delta^m v(t) = v(t-t) = v(0).$$

In the second case, we use identity $m\binom{t}{m} = t\binom{t-1}{m-1}$, which yields

$$\begin{aligned}\sum_{m=1}^{t} m v(t\|m) &= \sum_{m=1}^{t} (-1)^{m+1} m \binom{t}{m} \Delta^m v(t) = t \sum_{m=1}^{t} (-1)^{m+1} \binom{t-1}{m-1} \Delta^m v(t) \\ &= t \sum_{j=0}^{t-1} (-1)^j \binom{t-1}{j} \Delta^{j+1} v(t) = t \Delta v(t-t+1) = t \Delta v(1).\end{aligned}$$

Hence conditions (41) hold if and only if sequence $v(t)$ is standard. ■

Without loss of generality, let us assume that the alphabet of an IID process $(K_t)_{t\in\mathbb{Z}}$ is the set of natural numbers. By the Bernoulli scheme, the expected number of types and the spectrum elements defined via (36) equal

$$V(t) = \sum_{k\in\mathbb{N}} (1 - (1-p_k)^t), \tag{43}$$

$$V(t|m) = \sum_{k\in\mathbb{N}} \binom{t}{m} p_k^m (1-p_k)^{t-m}, \qquad\qquad 1 \le m \le t, \tag{44}$$

where $p_k := P(K_0 = k)$. By formula (44) and identity $\Delta(1-p_k)^t = -p_k(1-p_k)^t$, the spectrum elements $V(t|m)$ and the Taylor elements of $V(t)$ are equal,

$$V(t|m) = V(t\|m). \tag{45}$$

Consequently, sequence $V(t)$ is a Hausdorff sequence. Moreover, sequence $V(t)$ is standard and spectrum elements $V(t|m)$ satisfy consistency conditions analogous to (41). The Hausdorff measure of $V(t)$ is an atomic probability measure and assumes form

$$\tilde{V}(A) = \sum_{k\in\mathbb{N}} p_k \mathbf{1}\{p_k \in A\}. \tag{46}$$

By formulas (40) and (45), the expected rate of hapaxes for IID processes equals

$$\frac{V(t|1)}{t} = \Delta V(t-1), \tag{47}$$

which can be contrasted with the more general inequality (37) for stationary sources.

Appendix B. Proofs of the Main Results

It is high time to demonstrate the chain of implications (10). Respectively, in Appendices B.1–B.5, we demonstrate Theorems 1–5.

B.1 Proof of Theorem 1

Bounding the sums by integrals, may evaluate

$$\begin{aligned}
\sum_{k\in\mathbb{N}} \mathbf{1}\{p_k \geq p\} &\leq \int_0^\infty \mathbf{1}\Big\{A_Z^+ k^{-1/\beta} \geq p\Big\} dk \\
&= \int_0^\infty \mathbf{1}\left\{k \leq \left(\frac{p}{A_Z^+}\right)^{-\beta}\right\} dk = \left(\frac{p}{A_Z^+}\right)^{-\beta}, \\
\sum_{k\in\mathbb{N}} p_k \mathbf{1}\{p_k \leq p\} &\geq \int_1^\infty A_Z^- k^{-1/\beta} \mathbf{1}\Big\{A_Z^+ k^{-1/\beta} \leq p\Big\} dk \\
&= \int_1^\infty A_Z^- k^{-1/\beta} \mathbf{1}\left\{k \geq \left(\frac{p}{A_Z^+}\right)^{-\beta}\right\} dk = \frac{A_Z^-}{1/\beta - 1}\left(\frac{p}{A_Z^+}\right)^{1-\beta}.
\end{aligned}$$

B.2 Proof of Theorem 2

Since $t \mapsto \Delta V(t)$ is decreasing, we derive

$$\frac{V(\mathcal{K}_{t+1}^{t+s} \setminus \mathcal{K}_1^t)}{s} \leq \Delta V(t).$$

By (47) and (44), we may bound

$$\begin{aligned}
\Delta V(t) = \frac{V(t+1|1)}{t+1} &= \sum_{k\in\mathbb{N}} p_k (1-p_k)^t \\
&\leq \sum_{k\in\mathbb{N}} \int_0^{p_k} (1-p)^t dp = \int_0^1 \left(\sum_{k\in\mathbb{N}} \mathbf{1}\{p_k \geq p\}\right)(1-p)^t dp \\
&\leq A_S^+ \int_0^1 p^{-\beta}(1-p)^t dp.
\end{aligned}$$

Further evaluation yields

$$\int_0^1 (1-p)^t p^{-\beta} dp = \frac{\Gamma(t+1)\Gamma(1-\beta)}{\Gamma(t+2-\beta)} \leq \frac{\Gamma(1-\beta)}{t^{1-\beta}}.$$

B.3 Proof of Theorem 3

Since $t \mapsto \Delta V(t)$ is decreasing, we derive

$$\frac{V(\mathcal{K}_{t+1}^{t+s} \setminus \mathcal{K}_1^t)}{s} \geq \Delta V(t+s).$$

By (37), using the union bound, we may write

$$\begin{aligned}
\Delta V(t) &\geq \frac{V(2t|1)}{2t} = \frac{1}{2t}\sum_{i=1}^{2t} P\left(K_i \notin \mathcal{K}_1^{i-1} \cup \mathcal{K}_{i+1}^{2t}\right) \\
&= \sum_{k\in\mathbb{N}} \frac{1}{2t}\sum_{i=1}^{2t} P\left(K_i = k \notin \mathcal{K}_1^{i-1} \cup \mathcal{K}_{i+1}^{2t}\right) \\
&= \sum_{k\in\mathbb{N}} \max\left\{0, \frac{1}{2t}\sum_{i=1}^{2t} P\left(K_i = k \notin \mathcal{K}_1^{i-1} \cup \mathcal{K}_{i+1}^{2t}\right)\right\} \\
&\geq \sum_{k\in\mathbb{N}} \max\left\{0, 2p_k - \frac{1}{2t}\sum_{i=1}^{2t}\sum_{j=1}^{2t} P(K_i = k, K_j = k)\right\} \\
&\geq \sum_{k\in\mathbb{N}} p_k \max\left\{0, 1 - A_M(2t-1)p_k^\delta\right\} \geq \sum_{k\in\mathbb{N}} p_k \max\left\{0, 1 - 2A_M t p_k^\delta\right\} \\
&\geq \frac{1}{2}\sum_{k\in\mathbb{N}} p_k \mathbf{1}\left\{p_k^\delta \leq \frac{1}{4A_M t}\right\} \geq \frac{A_S^-}{2}\left(\frac{1}{4A_M}\right)^{\frac{1-\beta}{\delta}}.
\end{aligned}$$

B.4 Proof of Theorem 4

Without loss of generality, we may assume $P(K_t = k) > 0$ for all $k \in \mathbb{N}$ since we may re-index sequence $(Z_k)_{k\in\mathbb{N}}$ eliminating those variables Z_k for which $P(K_t = k) = 0$. Then we proceed as follows.

The main idea of the proof is contained in this paragraph. By independence of $(K_t)_{t\in\mathbb{Z}}$ and $(Z_k)_{k\in\mathbb{N}}$ and condition (18), we may decompose

$$\begin{aligned}
H(X_1^t) &= H(K_1^t) + H(X_1^t|K_1^t) \\
&= H(K_1^t) + H(\{(k, Z_k) : k \in \mathcal{K}_1^t\}|\mathcal{K}_1^t) \\
&\geq H(K_1^t) + A_K V(\mathcal{K}_1^t).
\end{aligned}$$

Analogously, for $k \geq t$, we also obtain the conditional statement

$$\begin{aligned}
H(X_{k+1}^{k+s}|X_1^t) &= H(K_{k+1}^{k+s}|X_1^t) + H(X_{k+1}^{k+s}|X_1^t, K_{k+1}^{k+s}) \\
&= H(K_{k+1}^{k+s}|K_1^t) + H\left(\left\{(k, Z_k) : k \in \mathcal{K}_{k+1}^{k+s} \setminus \mathcal{K}_1^t\right\}\middle|\mathcal{K}_{k+1}^{k+s} \setminus \mathcal{K}_1^t\right) \\
&\geq H(K_{k+1}^{k+s}|K_1^t) + A_K V(\mathcal{K}_{k+1}^{k+s} \setminus \mathcal{K}_1^t).
\end{aligned}$$

Since conditioning decreases entropy, we derive

$$H(K_{k+1}^{k+s}|K_1^t) \geq H(K_{k+1}^{k+s}|K_{-\infty}^k) = sH(K_{t+1}|K_{-\infty}^t)$$

by stationarity and the chain rule for conditional entropy generalized to σ-fields (Dębowski, 2021a, Section 5.3). Hence we derive (20) if we prove the equality of entropy rates

$$H(K_{t+1}|K_{-\infty}^t) = H(X_{t+1}|X_{-\infty}^t) = h. \tag{48}$$

Equality (48) seems intuitive but its formal proof is not so short since there may arise some problems if $H(X_t) = \infty$ and $H(X_t|X_{-\infty}^{t-1}) = 0$.

We prove (48) by applying the toolbox from Dębowski (2021a) and the classical results from Breiman (1957). First, we demonstrate the left-most equality in (48). For this goal, we observe the following facts:

- Since $P(K_t = k) > 0$ for any $k \in \mathbb{N}$ and $(K_t)_{t\in\mathbb{Z}}$ is stationary ergodic then $\lim_{t\to\infty} P(k \in \mathcal{K}_1^t) = 1$ by the Birkhoff ergodic theorem (Dębowski, 2021a, Theorem 4.6) and the Riesz theorem (Dębowski, 2021a, Theorem 3.27).

- For each $k \in \mathbb{N}$, there is a function g_k of strings such that for any $s \in \mathbb{Z}$ and $t \in \mathbb{N}$, we have $Z_k = g_k(X_{s+1}^{s+t})$ if $k \in \mathcal{K}_{s+1}^{s+t}$.

  In plain words, we know the correct value of $Z_k = z$ if we find a pair (k, z) in $\phi(X_{s+1}^{s+t})$.

- As a result of two above points, by the result of Dębowski (2021a, Theorem 8.10), for each $k \in \mathbb{N}$, the completion of the σ-field of Z_k is contained in the completion of the shift-invariant σ-field of $(X_t)_{t\in\mathbb{Z}}$.

  In plain words, it means that there is a function g_k^* of infinite sequences such that for any $s \in \mathbb{Z}$, we have $Z_k = g_k^*((X_{t+s})_{t\in\mathbb{Z}})$ almost surely.

- On the other hand, we invoke that the completion of the shift-invariant σ-field is contained in the intersection of the completions of the tail σ-field of past and the tail σ-field of future of process $(X_t)_{t\in\mathbb{Z}}$ (Dębowski, 2021a, Theorem 5.34).

  In plain words, it means that for any $t \in \mathbb{Z}$ there are functions g_k^+ and g_k^- of semi-infinite sequences such that $Z_k = g_k^-(X_{-\infty}^t)$ and $Z_k = g_k^+(X_t^\infty)$ almost surely.

Hence, the completion of the σ-field of $X_{-\infty}^t$ contains the completion of the σ-field of Z_1^∞.

The rest is a straightforward application of the calculus of Shannon information measures for σ-fields (Dębowski, 2021a, Sections 5.3 and 5.4). We make the following observations:

- As it has been shown, the completion of the σ-field of $X_{-\infty}^t$ contains the completion of the σ-field of Z_1^∞, so

$$H(X_{t+1}|X_{-\infty}^t) = H(X_{t+1}|X_{-\infty}^t, Z_1^\infty).$$

- Since the σ-fields of $K_{-\infty}^{t+1}$ and Z_1^∞ are independent, we also have

$$H(K_{t+1}|K_{-\infty}^t) = H(K_{t+1}|K_{-\infty}^t, Z_1^\infty).$$

- The σ-field of $(X_{-\infty}^t, Z_1^\infty)$ equals the the σ-field of $(K_{-\infty}^t, Z_1^\infty)$ by decomposition (17).

- Moreover, the σ-field of $Z_{K_{t+1}}$ is contained in the σ-field of $(K_{-\infty}^{t+1}, Z_1^\infty)$, so

$$H(Z_{K_{t+1}}|K_{-\infty}^{t+1}, Z_1^\infty) = 0.$$

Hence by the generalized chain rule, we may write

$$\begin{aligned} H(X_{t+1}|X_{-\infty}^t) &= H(X_{t+1}|X_{-\infty}^t, Z_1^\infty) \\ &= H(X_{t+1}|K_{-\infty}^t, Z_1^\infty) \\ &= H(K_{t+1}, Z_{K_{t+1}}|K_{-\infty}^t, Z_1^\infty) \\ &= H(K_{t+1}|K_{-\infty}^t, Z_1^\infty) + H(Z_{K_{t+1}}|K_{-\infty}^{t+1}, Z_1^\infty) = H(K_{t+1}|K_{-\infty}^t), \end{aligned}$$

which is the left-most equality in (48).

Now, we prove the right-most equality in (48). Applying the calculus of Shannon information measures for σ-fields (Dębowski, 2021a, Sections 5.3 and 5.4), we observe that

$$\begin{aligned} H(X_t|X_{t-k}^{t-1}) &= \mathbf{E}\left(-\log P(X_t|X_{t-k}^{t-1})\right), \\ H(X_t|X_{-\infty}^{t-1}) &= \mathbf{E}\left(-\log P(X_t|X_{-\infty}^{t-1})\right). \end{aligned}$$

We also notice that $\lim_{k\to\infty} P(X_t|X_{t-k}^{t-1}) = P(X_t|X_{-\infty}^{t-1})$ almost surely by the Lévy law and

$$\mathbf{E}\sup_{k\in\mathbb{N}}\left(-\log P(X_t|X_{t-k}^{t-1})\right) \leq H(X_t) + \log e < \infty$$

by a lemma of Breiman (1957) and the assumption $H(X_t) < \infty$. Hence by the Lebesgue dominated convergence, we obtain

$$\lim_{k\to\infty} H(X_t|X_{t-k}^{t-1}) = H(X_t|X_{-\infty}^{t-1}).$$

It remains to apply the Toeplitz lemma

$$\lim_{k\to\infty} a_k = a \implies \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^n a_k = a,$$

and the Fekete lemma (35) to derive the right-most equality in (48).

B.5 Proof of Theorem 5

We begin the proof with an analysis of a function. Fix a $\beta \in (0,1)$. We denote the function

$$f(s) = f(s,t,c) := (t+s)^{\beta-1} - \frac{c}{s}$$

for $t, c \geq 0$. Let $r = r(t,c) := \arg\max_{s>0} f(s,t,c)$. We have

$$0 = \left.\frac{df(s)}{ds}\right|_{s=r} = -(1-\beta)(t+r)^{\beta-2} + \frac{c}{r^2}.$$

Consequently, $(1-\beta)(t+r)^{\beta-2}r^2 = c$.

Assume first that $t > 0$. Then we may bound

$$(1-\beta)t^{\beta-2}r^2 \geq (1-\beta)(t+r)^{\beta-2}r^2 \geq (1-\beta)t^\beta\left(\frac{r}{t+r}\right)^2.$$

Hence we obtain the sandwich bound $r_0 \le r \le r_2$, where

$$r_0 = r_0(t,c) := ty, \qquad r_2 = r_2(t,c) := \frac{ty}{1-y}, \qquad y := \sqrt{\frac{ct^{-\beta}}{1-\beta}}.$$

Function $f(s)$ is increasing for $s \le r$ and decreasing for $s \ge r$. Suppose that $s \le r_2$. For $q = \lceil r_2/s \rceil$, we may bound

$$r \le sq \le s\left(\frac{r_2}{s} + 1\right) \le r_2 + s \le 2r_2$$

so we can also bound

$$f(sq,t,c) \ge f(2r_2) = t^{\beta-1}\left(\frac{1-y}{1+y}\right)^{1-\beta} - \frac{c}{t}\left(\frac{1-y}{2y}\right).$$

Now assume that $t = 0$ and $c = n$. Then

$$r = r(0,n) := \left(\frac{n}{1-\beta}\right)^{\frac{1}{\beta}}$$

Suppose that $s \le r$. For $q = \lceil r/s \rceil$, we may bound

$$r \le sq \le s\left(\frac{r}{s} + 1\right) \le r + s \le 2r$$

so we can also bound

$$f(sq,0,n) \ge f(2r) = \frac{2^\beta - 1 + \beta}{2}\left(\frac{n}{1-\beta}\right)^{1-\frac{1}{\beta}}.$$

This completes the analysis of function $f(s)$ that will be needed next.

Now we proceed to the main part of the proof. By an iterated application of inequality (28), for any $q \in \mathbb{N}$, we have inequality

$$\sup_{k \ge t} \frac{H(X_{k+1}^{k+s}|Q)}{s} \ge \sup_{k \ge t} \frac{H(X_{k+1}^{k+sq}|Q)}{sq}.$$

By contrast, by inequality (29), conditions (23) and (22) imply inequality

$$H(X_{k+1}^{k+s}|Q) \ge \max\left\{H(X_{k+1}^{k+s}|X_1^t) - A_H^- c, H(X_{k+1}^{k+s}) - A_H^- n\right\}.$$

Consequently, for $s \le r_2(t,c)$ and $q = \lceil r_2(t,c)/s \rceil$, condition (21) yields

$$\begin{aligned}\sup_{k \ge t} \frac{H(X_{k+1}^{k+s}|Q)}{s} - h &\ge \sup_{k \ge t} \frac{H(X_{k+1}^{k+sq}|Q)}{sq} - h \ge \sup_{k \ge t} \frac{H(X_{k+1}^{k+sq}|X_1^t) - A_H^- c}{sq} - h \\ &\ge A_H^- f(sq,t,c) \ge A_H^-\left(t^{\beta-1}\left(\frac{1-y}{1+y}\right)^{1-\beta} - \frac{c}{t}\left(\frac{1-y}{2y}\right)\right).\end{aligned}$$

Complementing the above inequality with the analogous development for conditions $s \leq r(0,n)$ and $q = \lceil r(0,n)/s \rceil$ yields

$$\sup_{k \geq t} \frac{H(X_{k+1}^{k+s}|Q)}{s} - h \geq \sup_{k \geq t} \frac{H(X_{k+1}^{k+sq}|Q)}{sq} - h \geq \sup_{k \geq t} \frac{H(X_{k+1}^{k+sq}) - A_H^- n}{sq} - h$$
$$\geq A_H^- f(sq,0,n) \geq A_H^- \frac{2^\beta - 1 + \beta}{2} \left(\frac{n}{1-\beta} \right)^{1-\frac{1}{\beta}}.$$

Hence, by the source coding inequality (32), we recover the claim (25).

References

A. Achille and S. Soatto. AI agents as universal task solvers. *Entropy*, 28(3):332, 2026.

N. I. Akhiezer. *The Classical Moment Problem and Some Related Questions in Analysis*. Philadelphia: Society for Industrial and Applied Mathematics, 2021.

D. J. Aldous. Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII — 1983*, volume 1117 of *Lecture Notes in Mathematics*, pages 1–198. New York: Springer, 1985.

E. G. Altmann, J. B. Pierrehumbert, and A. E. Motter. Beyond word frequency: Bursts, lulls, and scaling in the temporal distributions of words. *PLoS ONE*, 4:e7678, 2009.

R. H. Baayen. *Word frequency distributions*. Dordrecht: Kluwer Academic Publishers, 2001.

Y. Bahri, E. Dyer, J. Kaplan, J. Lee, and U. Sharma. Explaining neural scaling laws. *Proceedings of the National Academy of Sciences of the United States of America*, 121 (27):e2311878121, 2024.

M. Barkeshli, A. Alfarano, and A. Gromov. On the origin of neural scaling laws: from random graphs to natural language. In *2026 International Conference on Machine Learning (ICML)*, 2026.

P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences of the United States of America*, 117 (48):30063–30070, 2020.

M. Belkin. Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation. *Acta Numerica*, 30:203–248, 2021.

M. Belkin, D. Hsu, S. Ma, and S. Mandal. Reconciling modern machine-learning practice and the classical bias-variance trade-off. *Proceedings of the National Academy of Sciences of the United States of America*, 116(32):15849–15854, 2019.

Y. Bengio, Y. LeCun, and G. E. Hinton. Deep learning for AI. *Communications of the ACM*, 64(7):58–65, 2021.

S. N. Bernstein. Sur les fonctions absolument monotones. *Acta Mathematica*, 52:1–66, 1928.

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

R. C. Bradley. Basic properties of strong mixing conditions. a survey and some open questions. *Probability Surveys*, 2:107–144, 2005.

L. Breiman. The individual ergodic theorem of information theory. *The Annals of Mathematical Statistics*, 28:809–811, 1957.

T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, G. K. Ariel Herbert-Voss, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, M. L. Eric Sigler, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In *2020 Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

V. Cabannes, E. Dohmatob, and A. Bietti. Scaling laws for associative memories. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

F. Cagnetta, A. Raventós, S. Ganguli, and M. Wyart. Deriving neural scaling laws from the statistics of natural language. In *2026 International Conference on Machine Learning (ICML)*, 2026.

A. Chacoma and D. H. Zanette. Heaps' law and Heaps functions in tagged texts: Evidences of their linguistic relevance. *Royal Society Open Science*, 7(3):200008, 2020.

A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, and N. Fiedel. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24:240:1–240:113, 2023.

J. G. Cleary and I. H. Witten. Data compression using adaptive coding and partial string matching. *IEEE Transactions on Communications*, 32:396–402, 1984.

E. U. Condon. Statistics of vocabulary. *Science*, 67(1733):300–300, 1928.

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.

V. Davis. Types, tokens, and hapaxes: A new heap's law. *Glottotheory*, 9(2):113–129, 2018.

J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, and T. Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019.

M. Dingemanse, D. E. Blasi, G. Lupyan, M. H. Christiansen, and M. P. Arbitrariness, iconicity, and systematicity in language. *Trends in Cognitive Science*, 19(10):603–615, 2015.

Ł. Dębowski. *Excess entropy for stochastic processes over various alphabets*. PhD thesis, Institute of Computer Science, Polish Academy of Sciences, 2005. In Polish.

Ł. Dębowski. On Hilberg's law and its links with Guiraud's law. *Journal of Quantitative Linguistics*, 13:81–109, 2006.

Ł. Dębowski. A general definition of conditional information and its application to ergodic decomposition. *Statistics and Probability Letters*, 79:1260–1268, 2009.

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

Ł. Dębowski. Mixing, ergodic, and nonergodic processes with rapidly growing information between blocks. *IEEE Transactions on Information Theory*, 58:3392–3401, 2012.

Ł. Dębowski. Approximating information measures for fields. *Entropy*, 22(1):79, 2020.

Ł. Dębowski. *Information Theory Meets Power Laws: Stochastic Processes and Language Models*. New York: 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. A simplistic model of neural scaling laws: Multiperiodic Santa Fe processes. `https://arxiv.org/abs/2302.09049v1`, 2023.

Ł. Dębowski. Corrections of Zipf's and Heaps' laws derived from hapax rate models. *Journal of Quantitative Linguistics*, 32(2):128–165, 2025.

R. L. Dobrushin. A general formulation of the fundamental Shannon theorems in information theory. *Uspekhi Matematicheskikh Nauk*, 14(6):3–104, 1959. In Russian.

W. Ebeling and G. Nicolis. Word frequency and entropy of symbolic sequences: a dynamical perspective. *Chaos, Solitons and Fractals*, 2:635–650, 1992.

J. B. Estoup. *Gammes sténographiques*. Paris: Institut Stenographique de France, 1916.

F. Fan. An asymptotic model for the English hapax/vocabulary ratio. *Computational Linguistics*, 36(4):631–637, 2010.

M. Fekete. Über die Verteilung der Wurzeln bei gewissen algebraischen Gleichungen mit ganzzahligen Koeffizienten. *Mathematische Zeitschrift*, 17:228–249, 1923.

R. Ferrer-i-Cancho and R. V. Solé. Two regimes in the frequency of words and the origins of complex lexicons: Zipf's law revisited. *Journal of Quantitative Linguistics*, 8(3):165–173, 2001.

F. Font-Clos and A. Corral. Log-log convexity of type-token growth in Zipf's systems. *Physical Review Letters*, 114:238701, 2015.

R. Futrell and M. Hahn. Linguistic structure from a bottleneck on sequential information processing. *Nature Human Behaviour*, 10:589–600, 2026.

R. Futrell and K. Mahowald. How linguistics learned to stop worrying and love the language models. *Behavioral and Brain Sciences*, 49:e198, 2026.

P. Gács and J. Körner. Common information is far less than mutual information. *Problems of Control and Information Theory*, 2:119–162, 1973.

I. M. Gelfand, A. N. Kolmogorov, and A. M. Yaglom. Towards the general definition of the amount of information. *Doklady Akademii Nauk*, 111:745–748, 1956. In Russian.

M. Gerlach and E. G. Altmann. Stochastic model for the vocabulary growth in natural languages. *Physical Review X*, 3:021006, 2013.

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

P. Harremoës and F. Topsøe. Zipf's law, hyperbolic distributions and entropy loss. *Electronic Notes in Discrete Mathematics*, 21:315–318, 2005. General Theory of Information Transfer and Combinatorics.

F. Hausdorff. Momentprobleme für ein endliches Intervall. *Mathematische Zeitschrift*, 16: 220–248, 1923.

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. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. de Las Casas, L. A. Hendricks, J. Welbl, A. Clark, T. Hennigan, E. Noland, K. Millican, G. van den Driessche, B. Damoc, A. Guy, S. Osindero, K. Simonyan, E. Elsen, J. W. Rae, O. Vinyals, and L. Sifre. Training compute-optimal large language models. In *2022 Conference on Neural Information Processing Systems (NeurIPS)*, pages 30016–30030, 2022.

G. Hu. On the amount of information. *Teorija verojatnostej i ee primenenija*, 4:439–447, 1962.

M. Hutter. Learning curve theory. `https://arxiv.org/abs/2102.04074`, 2021.

O. Kallenberg. *Foundations of Modern Probability*. New York: Springer, 1997.

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.

S. Karlin. Central limit theorems for certain infinite urn schemes. *Journal of Mathematics and Mechanics*, 17(4):373–401, 1967.

E. Khmaladze. The statistical analysis of large number of rare events. Technical Report MS-R8804. Centrum voor Wiskunde en Informatica, Amsterdam, 1988.

T. Kobayashi and K. Tanaka-Ishii. Taylor's law for human linguistic sequences. In I. Gurevych and Y. Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1138–1148, Melbourne, Australia, 2018. Association for Computational Linguistics.

W. Kuraszkiewicz and J. Łukaszewicz. The number of different words as a function of text length. *Pamiętnik Literacki*, 42(1):168–182, 1951. In Polish.

U. Lai, G. S. Randhawa, and P. Sheridan. Heaps' law in GPT-Neo large language model emulated corpora. In *Proceedings of the Tenth International Workshop on Evaluating Information Access (EVIA)*, 2023.

L. A. Levin. Universal sequential search problems. *Problems of Information Transmission*, 9(3):265–266, 1973.

L. A. Levin. Randomness conservation inequalities: Information and independence in mathematical theories. *Information and Control*, 61:15–37, 1984.

P. Lévy. *Processus stochastiques et mouvement brownien*. Paris: Gauthier Villars, 1948.

H. Li, W. Zheng, Q. Wang, Z. Ding, H. Wang, Z. Wang, S. Xuyang, N. Ding, S. Zhou, X. Zhang, and D. Jiang. Farseer: A refined scaling law in large language models. In *2026 Conference on Neural Information Processing Systems (NeurIPS)*, 2026.

W. Li. References on Zipf's law. https://wli-zipf.upc.edu/, 2021.

Y. Liu, Z. Liu, and J. Gore. Superposition yields robust neural scaling. In *2025 Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

C. Louart, Z. Liao, and R. Couillet. A random matrix approach to neural networks. *The Annals of Applied Probability*, 28(2):1190–1248, 2018.

L. Lü, Z.-K. Zhang, and T. Zhou. Zipf's law leads to heaps' law: Analyzing their relation in finite-size systems. *PLoS ONE*, 5:1–11, 12 2010.

A. Maloney, D. A. Roberts, and J. Sully. A solvable model of neural scaling laws. https://arxiv.org/abs/2210.16859, 2022.

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

A. Mehri and M. Jamaati. Variation of Zipf's exponent in one hundred live languages: A study of the Holy Bible translations. *Physics Letters A*, 381(31):2470–2477, 2017.

E. J. Michaud, Z. Liu, U. Girit, and M. Tegmark. The quantization model of neural scaling. In *2023 Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013*, pages 3111–3119, 2013.

J. Milička. Type-token & hapax-token relation: A combinatorial model. *Glottotheory*, 2(1): 99–110, 2009.

J. Milička. Rank-frequency relation & type-token relation: Two sides of the same coin. In I. Obradović, E. Kelih, and R. Köhler, editors, *Methods and Applications of Quantitative Linguistics—Selected papers of the 8th International Conference on Quantitative Linguistics (QUALICO)*, pages 163–171. Belgrade: Academic Mind, 2013.

G. A. Miller. Some effects of intermittent silence. *American Journal of Psychology*, 70: 311–314, 1957.

M. A. Montemurro and D. H. Zanette. New perspectives on Zipf's law in linguistics: from single texts to large corpora. *Glottometrics*, 4:87–99, 2002.

O. Neumann and C. Gros. AlphaZero neural scaling and Zipf's law: A tale of board games and power laws. In *2025 Conference on Neural Information Processing Systems (NeurIPS)*, pages 65070–65100, 2025.

Z. Pan, S. Wang, P. Liao, and J. Li. Understanding LLM behaviors via compression: Data generation, knowledge acquisition and scaling laws. In *2025 Conference on Neural Information Processing Systems (NeurIPS)*, pages 186604–186654, 2025.

T. Pearce and J. Song. Reconciling Kaplan and Chinchilla scaling laws. *Transactions on Machine Learning Research*, 11, 2024.

A. Petersen, J. Tenenbaum, S. Havlin, H. E. Stanley, and M. Perc. Languages cool as they expand: Allometric scaling and the decreasing need for new words. *Scientific Reports*, 2: 943, 2012.

M. S. Pinsker. *Information and Information Stability of Random Variables and Processes*. San Francisco: Holden-Day, 1964.

J. Pitman and M. Yor. The two-parameter Poisson–Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, 25(2):855–900, 1997.

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

D. A. Roberts, S. Yaida, and B. Hanin. *The Principles of Deep Learning Theory*. Cambridge: Cambridge University Press, 2022.

B. Ryabko. Twice-universal coding. *Problems of Information Transmission*, 20(3):173–177, 1984.

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

R. L. Schilling, R. Song, and Z. Vondraček. *Bernstein Functions — Theory and Applications*. Berlin: Walter de Gruyter, 2010.

C. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 30: 379–423,623–656, 1948.

C. Shannon. Prediction and entropy of printed English. *Bell System Technical Journal*, 30: 50–64, 1951.

U. Sharma and J. Kaplan. Scaling laws from the data manifold dimension. *Journal of Machine Learning Research*, 23(9):1–34, 2022.

H. A. Simon. On a class of skew distribution functions. *Biometrika*, 42:425–440, 1955.

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.

C. Tao, Q. Liu, L. Dou, N. Muennighoff, Z. Wan, P. Luo, M. Lin, and N. Wong. Scaling laws with vocabulary: Larger models deserve larger vocabularies. In *2024 Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

R. Thoppilan, D. D. Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H.-T. Cheng, A. Jin, T. Bos, L. Baker, Y. Du, Y. Li, H. Lee, H. S. Zheng, A. Ghafouri, M. Menegali, Y. Huang, M. Krikun, D. Lepikhin, J. Qin, D. Chen, Y. Xu, Z. Chen, A. Roberts, M. Bosma, V. Zhao, Y. Zhou, C.-C. Chang, I. Krivokon, W. Rusch, M. Pickett, P. Srinivasan, L. Man, K. Meier-Hellstern, M. R. Morris, T. Doshi, R. D. Santos, T. Duke, J. Soraker, B. Zevenbergen, V. Prabhakaran, M. Diaz, B. Hutchinson, K. Olson, A. Molina, E. Hoffman-John, J. Lee, L. Aroyo, R. Rajakumar, A. Butryna, M. Lamm, V. Kuzmina, J. Fenton, A. Cohen, R. Bernstein, R. Kurzweil, B. Aguera-Arcas, C. Cui, M. Croak, E. Chi, and Q. Le. LaMDA: Language models for dialog applications. `https://arxiv.org/abs/2201.08239`, 2022.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, pages 5998–6008, 2017.

A. Wei, W. Hu, and J. Steinhardt. More than a toy: Random matrix models predict how real-world neural representations generalize. *Proceedings of Machine Learning Research*, 162:23549–23588, 2022.

P. Wieczyński and Ł. Dębowski. Long-range dependence in word time series: The cosine correlation of embeddings. *Entropy*, 27(6):613, 2025.

J. R. Williams, J. P. Bagrow, C. M. Danforth, and P. S. Dodds. Text mixing shapes the anatomy of rank-frequency distributions. *Physical Review E*, 91:052811, 2015.

A. D. Wyner. The common information of two dependent random variables. *IEEE Transactions on Information Theory*, IT-21:163–179, 1975.

A. D. Wyner. A definition of conditional mutual information for arbitrary ensembles. *Information and Control*, 38:51–59, 1978.

R. W. Yeung. *First Course in Information Theory*. Dordrecht: Kluwer Academic Publishers, 2002.

C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

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

G. K. Zipf. *Human Behavior and the Principle of Least Effort*. Reading: Addison-Wesley, 1949.