Learning curves theory of hierarchically compositional data with power-law distributed features

author names withheld

Under Review for the Workshop on High-dimensional Learning Dynamics, 2025

Abstract

We study how two fundamental properties of natural data—hierarchical compositionality and Zipf-distributed features—affect the scaling of test performance with the number of training examples. Using synthetic datasets generated by probabilistic context-free grammars, we derive learning curves for classification and next-token prediction tasks in the data-limited regime. For classification, we show that introducing a Zipf distribution over production rules leads to a power-law learning curve with an exponent controlled by the Zipf distribution. By contrast, in next-token prediction, the exponent is determined by the hierarchical structure alone and is unaffected by Zipf statistics. These results are supported empirically by experiments with convolutional and transformer models, and highlight how different aspects of the data structure shape neural scaling laws.

1. Introduction

Scaling laws have emerged as a unifying framework for understanding generalisation in large deep learning models [9, 17]. In particular, the *learning curve*, which quantifies the decay of test error with the number of training examples P, often follows a power-law: $\epsilon(P) \sim P^{-\alpha}$. While such scaling is well-understood in linear [10, 16] and kernel-based models [1–3, 5, 8, 13, 14, 19], its origin in deep feature-learning networks remains poorly understood. Here we ask: what properties of the data control α in the data-limited regime, and how do they interact with the architecture? In particular, we focus on two features common to real-world data: hierarchical structure and power-law feature statistics. To this end, we consider synthetic datasets generated by the *Random Hierarchy Model* (RHM) [6]—an ensemble of hierarchically compositional generative processes corresponding to simple context-free grammars. The RHM models the hierarchical structure of data via a tree of production rules that recursively expand hidden symbols into sequences of hidden symbols at the next level of the hierarchy, and eventually visible tokens. The probabilities of these rules control the distribution of the fundamental features of the data, such as tokens and combinations thereof: when the rule probabilities follow Zipf's law, these distributions also follow a power law, mirroring the statistics of words in natural language corpora.

2. The Random Hierarchy Model with Zipf-distributed rules

Generally, Probabilistic Context-Free Grammars (PCFGs) consist of a vocabulary of hidden (*non-terminal*) symbols, a vocabulary of observable (*terminal*) symbols and *production rules* that quantify the probability that one hidden symbol generates tuples of either hidden or observable symbols. PCFGs provide a natural formalism for describing hierarchical structures found in the syn-

tax [11, 18] and semantics [12] of natural language, and images [20]. Here, for the sake of analytical tractability, we consider a restricted class of PCFGs, where production rules are sampled uniformly at random, compatibly with the following constraints:

C1. **(Fixed tree topology)** All grammars share a common tree structure: a fixed regular tree of arity s and depth L. This tree serves as the shared backbone for all the generated data, which consists of sequences of $d = s^L$ symbols. Because of the fixed topology constraint, exact inference on data generated by the RHM can be performed using the Belief Propagation (BP) algorithm [15].

C2. **(Unambiguity)** Production rules are chosen so that no two distinct hidden symbols are allowed to generate the same sequence of children. This constraint ensures that the observable datum uniquely determines the hidden structure of its derivation.

C3. **(Vocabulary size v)** Hidden symbols are split into L vocabularies $\mathcal{V}_\ell$, with $\ell = 0, \dots, L-1$ and $\mathcal{V}_L \equiv \mathcal{V}$ denoting the vocabulary of observable symbols. All vocabularies have the same cardinality v.

C4. **(m production rules per symbol)** Each hidden symbol is associated with m distinct and equiprobable production rules that yield symbols of the next level. Each rule can be picked with probability $f_k^{(\ell)}$, with $k = 1, \dots, m$ and $\sum_k f_k^{(\ell)} = 1$ for all ℓ.

To mimick the power-law distribution of word frequencies [7], we set the production rule distribution to be uniform in all but one layer ℓ, where it follows a Zipf law [10, 16], $f_k^{(\ell)} \propto k^{-(1+a)}$.

A given instance of the RHM corresponds to a probability distribution over sequences of tokens and the whole tree of hidden symbols. We consider both classification tasks, whose objective is the conditional probability of the root symbol (label) given the leaves (visible tokens),

$$\mathbb{P}\left\{X_1 = x_1, \dots, X_d = x_d | Y = y\right\} \tag{1}$$

and next-token prediction tasks, where the target is the conditional probability of the last observable token given the others,

$$\mathbb{P}\left\{X_d = x_d | X_1 = x_1, \dots, X_{d-1} = x_{d-1}\right\}. \tag{2}$$

3. Correlation-based learning theory

The statistics of the sequences generated by the RHM carry a signature of the hierarchical structure of the generative process. In particular, as shown in [5, 6], the correlations between tuples of visible tokens generated by the same hidden variable and other symbols in the tree are equal for all the tuples generated by the same hidden variable. As a result, these correlations serve as the primary cues a learner can exploit to reconstruct the hidden structure of the data. For classification tasks, for instance, the hidden variables can be inferred from the correlations between the root label Y and s-tuples of contiguous input tokens $\boldsymbol{X}_j = (X_{(j-1)s+1}, \dots, X_{js})$

$$C_j(y, \boldsymbol{\mu}) \coloneqq \mathbb{P}\left\{Y = y; X_{(j-1)s+1} = \mu_1, \dots, X_{js} = \mu_s\right\} - \mathbb{P}\left\{Y = y\right\} \mathbb{P}\left\{\boldsymbol{X}_j = \boldsymbol{\mu}\right\}. \tag{3}$$

Due to the unambiguity of the production rules, each s-tuple μ is generated by a unique nonterminal symbol at the level above. As a result, the correlations $C_j(y, \boldsymbol{\mu})$ are equal for all tuples $\boldsymbol{\mu}$ generated by the same hidden variable. In other words, tuples with identical ancestry exhibit identical correlations. This observation implies that, given enough data to estimate the correlations reliably, a learner can cluster tuples by their generating nonterminal symbol, effectively inverting the corresponding production rules. Once the first level of hidden variables is recovered, the same strategy can be applied recursively to higher levels, enabling a bottom-up reconstruction of the entire tree. Motivated by this property, we adopt the following assumption:

> **Assumption 1.** A production rule can be used by the learner once the correlations it induces on observable variables can be resolved from the training data.

In the case of the RHM, **Assumption 1** corresponds to requiring that the variance of the empirical correlations due to sampling noise is small compared to the variance between correlations associated with distinct production rules. The latter quantifies how distinguishable different production rules are in expectation, and can be estimated as the variance of the correlation function over the RHM ensemble, $\left\langle C_j(y, \boldsymbol{\mu})^2 \right\rangle_{\text{RHM}}$.

We then introduce a second assumption to translate the set of learned production rules into generalisation performance:

> **Assumption 2.** The learner achieves the same performance as the optimal predictor conditioned on the subset of available production rules.

In other words, once a production rule becomes accessible via its associated correlations, we assume the learner can exploit it as if it were directly observing the associated hidden variable. We estimate the resulting performance as the average over the RHM ensemble of the performance of the optimal conditioned predictor.

4. Classification: Zipf's law controls the exponent

In the presence of Zipf-distributed level-1 production rules, the probability of observing a given tuple μ is proportional to the probability of the unique production rule that generates it. Denoting the rank of this rule with $k(\boldsymbol{\mu})$, the probability is $f_{k(\boldsymbol{\mu})}$. As a consequence, the correlations $C_j(y, \mu)$ are themselves proportional to the rule probability f_k, implying that the variance of these correlations over the RHM ensemble scales as f_k^2. The remaining factor, depending on the higher hidden levels of the data structure, can be determined with the techniques developed in [6] for the uniform RHM. The variance due to sampling noise, instead, is proportional to the average occurrence of the tuple $\boldsymbol{\mu}$ in the training data, scaling as f_k/P if P denotes the size of the training set.

Following **Assumption 1** and balancing the two variances, we get that the sample complexity required for learning the production rules with rank k is $P_k^* = vm^{L-1}/(f_k)$. To estimate the learning curve we follow **Assumption 2**, which implies that, when $P > P_k^*$, the model can correctly classify data consisting of tuples with probability higher than f_k. In other words, the model classifies the input correctly if and only if all the s^{L-1} input patches are resolvable. The resulting test error reads

$$\varepsilon(P) = 1 - \left(\sum_{k | P_k^* < P} f_k \right)^{s^{L-1}} , \tag{4}$$

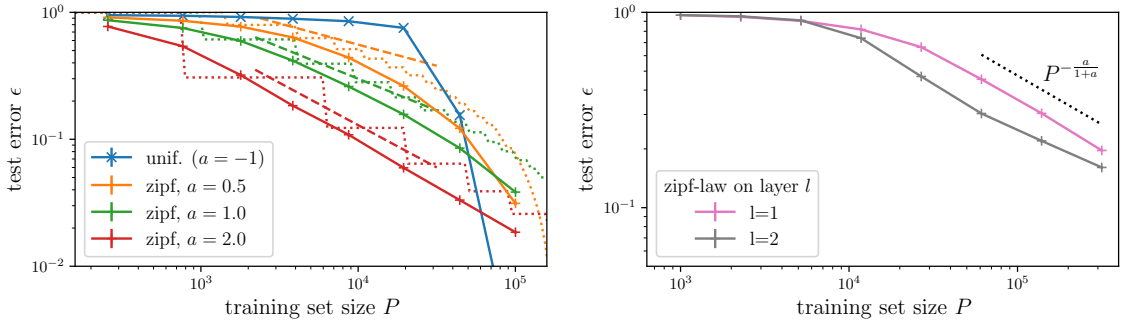

Figure 1: **Left:** Learning curves of 3-layers CNNs trained on RHM data with $L=2$, $s=2$, $v=m=25$ and Zipf exponent a indicated in caption. Solid lines are the empirical learning curves whereas dotted lines are predictions from Eq. (4). The dashed line represents the scaling law $\epsilon \sim P^{-a/(1+a)}$. **Right:** As in the left panel, but $v=m=100$. Here a is fixed and the layer where production rules are Zipf-distributed changes. The black dotted line represents the asymptotic scaling law $\epsilon \sim P^{-a/(1+a)}$.

which, when $P \gg P_1^* \simeq vm^{L-1}$, yields the asymptotic scaling $P^{-a/(1+a)}$. This result agrees with that of [10, 16] and is confirmed empirically in Figure 1, showing the learning curves of deep CNNs trained on RHM classification.

4.1. Next-token prediction: hierarchy controls the scaling

In next-token prediction, the relevant correlations are those between tuples of visible tokens and the last token X_d,

$$C_j(\boldsymbol{\mu}, \nu) \coloneqq \mathbb{P}\left\{\boldsymbol{X}_j = \boldsymbol{\mu}, X_d = \nu\right\} - \mathbb{P}\left\{\boldsymbol{X}_j = \boldsymbol{\mu}\right\}\mathbb{P}\left\{X_d = \nu\right\}.$$

In the uniform case, the correlations' variance decays with the tree distance ℓ between $\boldsymbol{X}_j$ and X_d as $\left\langle C_j^2 \right\rangle \propto m^{-2\ell}$. The variance due to sampling, instead, is independent of ℓ and inversely proportional to the dataset size P, resulting in a sequence of sample complexities $P_\ell \propto m^{2\ell}$ for inferring all production rules withing the depth-ℓ sub-tree above the last token X_d. When $P \gg P_\ell$, the model outputs the s^ℓ-gram approximation of the last-token probability,

$$\mathbb{P}\left\{X_d = x_d | X_{d-(s^\ell-1)} = x_{d-(s^\ell-1)}, \dots, X_{d-1} = x_{d-1}\right\}. \tag{5}$$

Combining the scaling of P_ℓ with ℓ with that of the average cross-entropy losses of the s^ℓ-grams, $\mathcal{L}_\ell \propto (m/v^{s-1})^\ell$, yields the *scaling law* [4]

$$\mathcal{L}(P) \sim P^{-\log(m/v^{s-1})/(2\log m)}, \tag{6}$$

where $\sim$ implies that P-independent factors are neglected.

In the case of next-token prediction, as for classification tasks, the Zipf distribution induces a proportional dependence of the correlation magnitude on the probability of the corresponding production rule. However, this dependence does not alter the scaling of the correlations with the tree distance ℓ. Moreover, although the nonuniform rule probabilities affect the specific values of the s^ℓ-gram cross-entropies—modifying the asymptotic loss—they do not change the rate at which

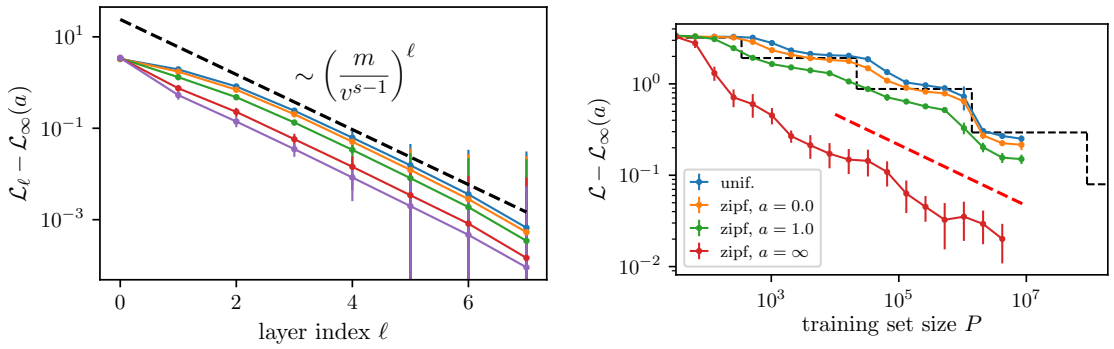

Figure 2: **Left:** Average cross-entropies of the s^ℓ-grams versus ℓ, for RHM datasets with $s = 2$, $v = 32$, $m = 8$, with the colour denoting the Zipf exponent. The points are obtained by averaging the cross-entropies over 32 independent realisations of the RHM. The cross-entropies of the uniform production rules case are shown in blue for comparison. For all a's, the cross-entropies $\mathcal{L}_\ell$ decay with ℓ towards some a-dependent value $\mathcal{L}_\infty(a)$. However, the approach to $\mathcal{L}_\infty(a)$ is independent of a and follows the behaviour of the test loss bound derived in [4] in the uniform case (black dashed line). **Right:** Empirical scaling laws of depth-4 transformers trained on RHM next-token prediction with $L = 4$, $s = 2$, $v = 32$, $m = 8$ and varying a. The limit $a \to \infty$ corresponds to having only one production rule per level-1 nonterminal symbol. The red dashed line is a guide to the eye for the asymptotic decay of Equation 6.

these entropies decay with ℓ. This property is illustrated in the left panel of Figure 2, displaying the approach of the$\mathcal{L}_\ell$'s, computed exactly for fixed realisations of the RHM via belief propagation and averaged over independent realisations, to the limiting residual cross-entropy $\mathcal{L}_\infty(a)$. As a result, the local details of the learning curve, such as the plateau heights and transition sharpness, may vary, but the global scaling law with respect to the number of training examples remains unchanged from the uniform RHM case.

5. Discussion and outlook

We studied how the scaling laws of deep networks trained in a feature-learning and data-limited regime are affected by two ubiquitous properties of natural data: hierarchical compositionality and Zipfian distribution of features. Remarkably, the effects of these two structural properties on learning greatly differ between classification and next token prediction tasks.

In particular, the remarkable independence of the next-token prediction scaling law from the production rule probabilities suggests the hierarchical structure of the data—rather than the statistics of individual features—as a prime candidate for explaining the emergence of scaling behaviour in modern machine learning.

References

[1] Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *arXiv preprint arXiv:2102.06701*, 2021.

[2] Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. In *International Conference on Machine Learning*, pages 1024–1034. PMLR, 2020.

[3] Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. A dynamical model of neural scaling laws. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=nbOY1OmtRc.

[4] Francesco Cagnetta and Matthieu Wyart. Towards a theory of how the structure of language is acquired by deep neural networks. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=NaCXcUKihH.

[5] Francesco Cagnetta, Alessandro Favero, and Matthieu Wyart. What can be learnt with wide convolutional neural networks? In *International Conference on Machine Learning*, pages 3347–3379. PMLR, 2023.

[6] Francesco Cagnetta, Leonardo Petrini, Umberto M. Tomasini, Alessandro Favero, and Matthieu Wyart. How deep neural networks learn compositional data: The random hierarchy model. *Phys. Rev. X*, 14:031001, Jul 2024. doi: 10.1103/PhysRevX.14.031001. URL https://link.aps.org/doi/10.1103/PhysRevX.14.031001.

[7] Álvaro Corral, Gemma Boleda, and Ramon Ferrer-i Cancho. Zipf's law for word frequencies: Word forms versus lemmas in long texts. *PLOS ONE*, 10(7):e0129031, 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0129031. URL http://dx.doi.org/10.1371/journal.pone.0129031.

[8] Alessandro Favero, Francesco Cagnetta, and Matthieu Wyart. Locality defeats the curse of dimensionality in convolutional teacher-student scenarios. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=sBBnfOFtPc.

[9] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack William Rae, and Laurent Sifre. An empirical analysis of compute-optimal large language model training. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=iBBcRUlOAPR.

[10] Marcus Hutter. Learning curve theory. *arXiv preprint arXiv:2102.04074*, 2021.

[11] Aravind K. Joshi. Tree adjoining grammars: How much context-sensitivity is required to provide reasonable structural descriptions? In David R. Dowty, Lauri Karttunen, and Arnold M. Zwicky, editors, *Natural Language Parsing: Psychological, Computational, and Theoretical Perspectives*, pages 206–250. Cambridge Univ. Press, Cambridge, UK, 1985.

[12] Donald E Knuth. Semantics of context-free languages. *Mathematical systems theory*, 2(2): 127–145, 1968.

[13] Licong Lin, Jingfeng Wu, Sham M. Kakade, Peter Bartlett, and Jason D. Lee. Scaling laws in linear regression: Compute, parameters, and data. In *The Thirty-eighth Annual Conference*

on Neural Information Processing Systems, 2024. URL https://openreview.net/forum?id=PH7sdEanXP.

[14] Alexander Maloney, Daniel A Roberts, and James Sully. A solvable model of neural scaling laws. *arXiv:2210.16859*, 2022.

[15] Marc Mezard and Andrea Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

[16] Eric J Michaud, Ziming Liu, Uzay Girit, and Max Tegmark. The quantization model of neural scaling. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=3tbTw2ga8K.

[17] OpenAI. Gpt-4 technical report. 2023.

[18] Geoffrey K. Pullum and Gerald Gazdar. Natural languages and context-free languages. *Linguist. Philos.*, 4(4):471–504, 1982. doi: 10.1007/BF00360802.

[19] Stefano Spigler, Mario Geiger, and Matthieu Wyart. Asymptotic learning curves of kernel methods: empirical data versus teacher–student paradigm. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(12), 2020.

[20] Song-Chun Zhu and David Mumford. A stochastic grammar of images. *Found. Trends Comput. Graph. Vis.*, 2(4):259–362, 2006. doi: 10.1561/0600000017.