Reconciling Grokking with Statistical Learning Theory

Luca Oneto[1], Sandro Ridella[1], Andrea Coraddu[2], and Davide Anguita[1] *

1 - University of Genoa, Genova, Italy
2 - Delft University of Technology, Delft, Netherlands

Abstract. In recent years, Artificial Intelligence, particularly Machine Learning (ML), has demonstrated remarkable success in addressing complex problems. However, this progress has been accompanied by the emergence of unexpected, poorly understood, and elusive phenomena that characterize the behavior of machine intelligence and learning processes. Researchers are often challenged to interpret these phenomena within the existing theoretical frameworks of ML, fostering a search for more complex or technical explanations. One such phenomenon, known as "grokking", occurs when an ML model, after a long period of stagnant or even regressive learning, suddenly exhibits rapid and substantial improvement. In this paper, we argue that grokking can be explained with the theoretical foundations of ML by leveraging Statistical Learning Theory, i.e., Algorithmic Stability theory. We provide insights into how this theory can reconcile grokking with established principles of learning and generalization.

1 Introduction

In recent years, Artificial Intelligence (AI), especially Machine Learning (ML), has significantly transformed society, industry, and science. From mastering games [1] to powering large language models [2] and solving complex problems like protein folding [3], the capabilities of this new generation of intelligent machines appear boundless.

Despite significant advancements, we are increasingly encountering unexpected and poorly understood phenomena that complicate the behavior of machine intelligence and learning processes. Catastrophic forgetting [4], where models lose previously acquired knowledge when learning new data, and bias amplification [5], which reinforces historical societal biases inherent in the data, are notable challenges. Other concerns include physical implausibility [6], lack of explainability [5], and privacy violations [5]. Adversarial vulnerability [5] exposes models to manipulation through subtle input alterations, while shortcut learning [7] leads them to exploit superficial patterns rather than deeper insights. Issues such as double descent [8], where performance temporarily dips before improving, and benign overfitting [8], where models fit noise without compromising generalization, challenge conventional wisdom. Over-parameterization [8] defies expectations by enhancing performance despite excessive model complexity, while grokking [9] refers to the delayed achievement of effective generalization. These phenomena underscore the growing complexity and unpredictability of ML, raising critical challenges for its reliable and responsible application.

In this work, we explore the phenomenon of "grokking", where, following a prolonged phase of stagnation or even regression, a model suddenly experiences

*This work is partially supported by (i) project SERICS (PE00000014) under the NRRP MUR program funded by the EU - NGEU, and (ii) project FAIR (PE00000013) under the NRRP MUR program funded by the EU - NGEU.

a rapid and substantial improvement in task performance [9]. This abrupt shift resembles the moment when a person achieves a breakthrough in understanding after struggling with a concept. What makes grokking particularly fascinating is its divergence from the traditional expectations of statistical learning, which generally follows a pattern of steady, incremental progress [10]. Instead, grokking suggests a dynamic in which models may initially show little or no improvement, or even a decline in performance, before unexpectedly surging in capability. Originally observed in supervised ML on algorithmic datasets [9, 11], this phenomenon has since been identified in real-world data [12–14] and other learning contexts [15, 16]. This unpredictability has posed significant challenges for researchers attempting to explain grokking within existing ML theoretical frameworks, spurring the search for more sophisticated interpretations and explanations [17–23].

In this paper, we argue that grokking can be understood through the lens of ML's theoretical foundations, specifically leveraging Statistical Learning Theory (SLT), and more precisely, Algorithmic Stability theory. We offer insights into how this framework reconciles grokking with established principles of learning and generalization. To this end, we will conduct a theoretical analysis of the phenomenon in Section 2. In Section 3, practical examples will be used to illustrate and clarify the theoretical concepts. Finally, Section 4 will summarize our findings and provide closing remarks.

2 Theoretical Analysis

Consider the supervised learning setting [10, 24]. Given a random observation $X \in \mathcal{X}$, the goal is to estimate the corresponding $Y \in \mathcal{Y}$, sampled according to an unknown distribution μ over $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. This is achieved by selecting an appropriate function $f : \mathcal{X} \to \hat{\mathcal{Y}}$ from a hypothesis set $\mathcal{F}$. A learning algorithm $\mathscr{A}_{\mathcal{H}}$, characterized by hyperparameters $\mathcal{H}$, outputs a function $\hat{f} \in \mathcal{F}$ based on a labeled dataset of n samples, $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\} = \{Z_1, \ldots, Z_n\}$, where each $Z_i \in \mathcal{Z}$ is sampled independently from μ. Thus, $\hat{f}$ is defined as $\hat{f} = \mathscr{A}_{\mathcal{H}}(\mathcal{D}_n)$. The generalization error of $\hat{f}$, representing its performance in approximating $\mathbb{P}\{Y|X\}$, is given by $\mathsf{L}(\hat{f}) = \mathbb{E}_Z\{\ell(\mathscr{A}_{\mathcal{H}}(\mathcal{D}_n), Z)\}$, where $\ell : \mathcal{F} \times \mathcal{Z} \to [0, 1]$ is a loss function that assesses the accuracy of the approximation for each sample point. Since $\mathsf{L}(\hat{f})$ is unknown, we estimate it using the empirical error, defined as $\mathsf{L}_{\text{emp}}(\mathscr{A}_{\mathcal{H}}(\mathcal{D}_n)) = {}^1\!/_n \sum_{Z \in \mathcal{D}_n} \ell(\mathscr{A}_{\mathcal{H}}(\mathcal{D}_n), Z)$. Ideally, a learning algorithm would return an oracle function $f^* = \arg\min \mathsf{L}(f)$, minimizing the generalization error across all possible models. However, since $\mathsf{L}(f)$ is unknown, we need to frame the problem as $\hat{f} = \arg\min_{f \in \mathcal{F}} \mathsf{L}_{\text{emp}}(f)$. This approach is known as Empirical Risk Minimization (ERM) [10, 24]. To address the issue of ERM's overfitting, hyperparameters $\mathcal{H}$ must be carefully tuned, often by introducing regularization techniques, which can be either explicit or implicit[10, 24]. Explicit regularization typically involves constraints on the parameter norms of f, such as an their p-norm. Implicit regularization, on the other hand, can include modifying the functional form of f (e.g., selecting a linear model, kernel method, or neural network architecture with specific layers or activations), altering the optimization algorithm (e.g., by applying early stopping or dropout), or even overparameterization. Moreover,

the model is usually expressed as $f(X) = g_\omega(\Phi_\theta(X))$, where $g : \mathbb{R}^D \to \hat{\mathcal{Y}}$ is a task specific function, $\Phi : \mathcal{X} \to \mathbb{R}^D$ a representation function, and ω and θ are the parameters of g and Φ respectively allowing to formulate ERM as $\hat{\omega}, \hat{\theta} = \arg\min_{\omega,\theta\in\mathcal{F}} \mathsf{L}_{\text{emp}}(\omega, \theta)$, where $\mathcal{F}$ becomes the, implicit or explicit , search space of the parameters [10, 24]. An alternative estimator to the empirical error is the leave-one-out error, $\mathsf{L}_{\text{loo}}(\mathscr{A}_\mathcal{H}(\mathcal{D}_n)) = {}^1\!/_n \sum_{Z\in\mathcal{D}_n} \ell(\mathscr{A}_\mathcal{H}(\mathcal{D}_n \setminus Z), Z)$, which calculates the average error on individual samples from $\mathcal{D}_n$ left out during training.

In this context [8], it can be shown that $\mathbb{P}\{\mathsf{L}(\mathscr{A}_\mathcal{H}(\mathcal{D}_n)) \leq \mathsf{L}_*(\mathscr{A}_\mathcal{H}(\mathcal{D}_n)) + \mathsf{M}(\mathscr{A}_\mathcal{H}) + \Delta(n, \delta)\} \geq 1-\delta$, indicating that the generalization error of $\hat{f}$ is bounded by an empirical estimate, $\mathsf{L}_* \in \{\mathsf{L}_{\text{emp}}, \mathsf{L}_{\text{loo}}\}$, along with two additional terms. The term $\mathsf{M}(\mathscr{A}_\mathcal{H})$ reflects the risk associated with the chosen algorithm and its hyperparameters; this term increases when the algorithm prioritizes memorization or overfitting rather than effective learning from the data. The second term[1], $\Delta(n, \delta)$, represents a confidence measure related to the sample; this term grows larger as the sample size decreases or as higher confidence is required.

An effective method to estimate $\mathsf{M}(\mathscr{A}_\mathcal{H})$ is based on Algorithmic Stability (AS) [25–27]. AS has frequently provided valuable insights into generalization [25, 26] and complex phenomena [8, 28]. The core idea of AS is intuitive: the more consistently an algorithm performs when the training data are slightly modified the more it generalizes. There are various types of AS [8, 25, 26], including uniform stability, Hypothesis Stability (HS), cross-validation and leave-one-out stability, error stability, and pointwise HS. Among these, HS has proven particularly insightful for understanding complex behaviors. HS can be estimated either practically or theoretically [8, 27] and is sensitive to the properties of the algorithm itself [8, 25–27]. HS can be defined as, for example, $\beta_{\text{loo}} = \mathbb{E}_{\mathcal{D}n,Z'} |\ell(\mathscr{A}_\mathcal{H}(\mathcal{D}_n), Z') - \ell(\mathscr{A}_\mathcal{H}(\mathcal{D}_n \setminus Z_i), Z')|$, where Z' is a sample from μ. Using L_{loo} as the empirical estimator, we find $\mathsf{M}(\mathscr{A}_\mathcal{H}) \propto \beta_{\text{loo}}$. Additionally, if ℓ is Lipschitz continuous with respect to a distance $\mathtt{d}(\cdot, \cdot)$ (where $\mathtt{d} : \hat{\mathcal{Y}} \times \hat{\mathcal{Y}} \to \mathbb{R}$), then $\beta_{\text{loo}} \propto \mathbb{E}_{\mathcal{D}_n,X'} \mathtt{d}(\mathscr{A}_\mathcal{H}(\mathcal{D}_n)(X'), \mathscr{A}_\mathcal{H}(\mathcal{D}_n \setminus Z_i)(X'))$. This HS formulation is particularly useful because it can be fully estimated once models are trained, allowing for stability analysis not only of the overall model but also of individual layers in deep architectures [8]. Specifically, after training $\hat{f} = \mathscr{A}_\mathcal{H}(\mathcal{D}_n)$, one can adjust the target $\hat{\mathcal{Y}}$ to any internal representation layer and, using a suitable distance $\mathtt{d}(\cdot, \cdot)$, compute β_{loo} for that layer. Furthermore, it can be shown that $\beta_{\text{loo}} \propto \mathsf{Cond}(\mathbf{HH}')$, where $\mathbf{H} = [\Phi_\theta(X_1), \cdots, \Phi_\theta(X_n)]'$, $\mathbf{HH}'$ is the Gram matrix, and Cond represents the condition number (i.e., the ratio between the largest and smallest singular values).

In this work, we argue that HS, as seen in other instances [8], can offer insights into the grokking phenomenon. When ERM is approached using gradient-based methods [10, 24], results from Statistical Learning Theory (SLT) suggest that, as a first approximation, $\mathsf{M}(\mathscr{A}_\mathcal{H}) \propto \iota$, where ι represents the number of gradient iterations [10, 24]. This implies that excessive iterations may harm generalization. Interestingly, in certain situations, increasing ι might show minimal or no benefit, or even a decrease, in generalization. However, after an extensive number of iterations, a sudden improvement in capability, grokking, can occur [9, 11]. This observation appears to challenge SLT principles, prompting researchers to seek more nuanced explanations [17–23]. A refined interpretation of HS, as presented

[1] We will not delve into this term as it is independent of $\mathscr{A}_\mathcal{H}$.

in [8], reveals that the approximation $\mathsf{M}(\mathscr{A}_{\mathcal{H}}) \propto \iota$ in SLT, similar to SLT's interpretation of overparameterization, is overly simplistic. HS can provide more immediate insights into grokking, as we explore in the next section. Specifically, a high ι can lead to a lower β_{loo}, which corresponds to improved generalization.

3 Empirical Evidences

In this section we will consider two examples: a toy[2] example (TOY) and the seminal example of [9] (MOD97).

The TOY dataset leverage the concept of implicit biases as a possible explanation for the phenomenon of grokking. Machine learning algorithms are typically designed to favor certain types of solutions, such as simpler or easier ones. However, in some datasets, this bias does not hold, and memorization may be required to achieve optimal performance. To illustrate this, we consider a binary classification task on a dataset of two-dimensional points. A training set of linearly separable points is constructed. A hard-margin Support Vector Machine is then applied to this data to identify the maximum-margin solution, which is subsequently used to generate a test set consisting of points positioned close to the decision boundary. Next, central points from the test set are removed, creating a scenario where test accuracy will not improve unless the learned solution aligns closely with the maximum-margin solution. This dataset is represented in Figure 1. For the TOY dataset, we perform binary classification with $\mathcal{Y} \in \{\pm 1\}$ on a dataset of 2-dimensional points $X \in \mathbb{R}^2$. We learn a model $f(X) = W \cdot X$, where the weights $W \in \mathbb{R}^2$ are optimized using ERM with gradient descent (learning rate $= 0.02$) and a loss function $\ell(f, Z) = \exp(-Y f(X))$. In Figure 2, we report the training and test

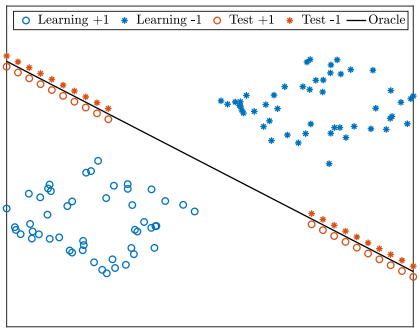

Fig. 1: TOY Dataset.

[2] https://xanderdavies.com/writing/toy_grok/toy_grok.html

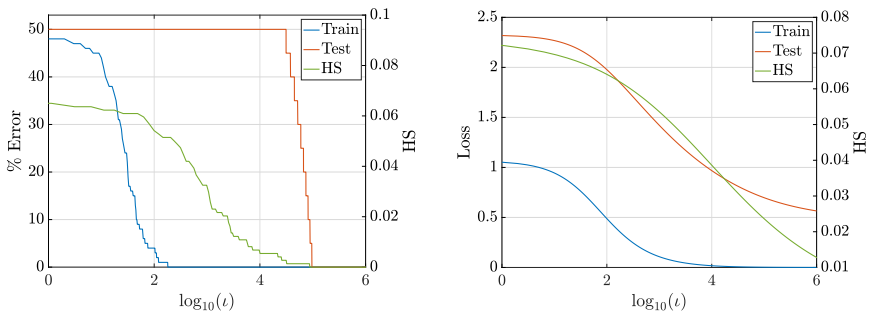

Fig. 2: Results for the TOY Dataset.

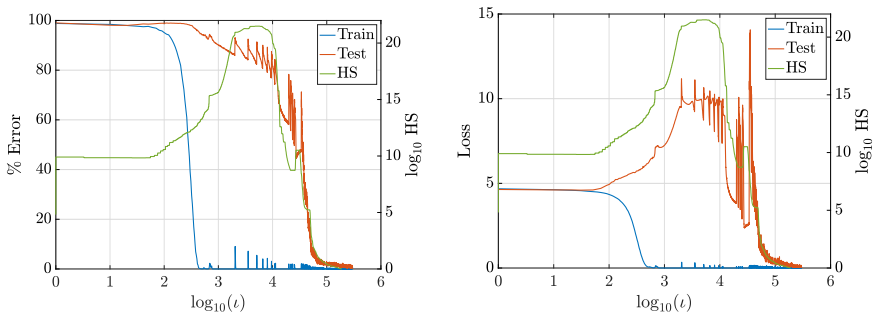

Fig. 3: Results for the MOD97 Dataset.

error percentages along with the HS, estimated on the data using the misclassification loss and the training loss [8, 27] as ι varies. As illustrated in Figure 2 and discussed in Section 2, the HS does not increase with ι. Instead, it decreases, which indicates potential improvements in test error that are indeed observed. Notably, even when the training curve appears to plateau, the HS significantly decreases, suggesting enhanced generalization performance, which is reflected in the test error reduction.

For the MOD97 dataset, we precisely replicate the experiment from [9] (specifically, Figures 1 and 4) and present the results in Figure 3, with the addition of HS. In this case, directly estimating HS was computationally infeasible; therefore, following the approach in [8], we approximate it using the condition number computed on the representation vector (as described in Section 2), while varying ι (keeping it consistent across both figures). Notably, the results in Figure 3 closely resemble those in [9]. Here, as well, HS serves as a robust indicator of the learned model's generalization ability. Even if the training curve suggests convergence or a slowdown in learning, the HS provides insight into the likely performance on the test set (assessed through the loss rather than the error, as we are estimating the HS of the representation).

Both the TOY and MOD97 dataset results highlight the potential of HS for understanding and predicting generalization behavior, thereby identifying early signs of grokking.

4 Conclusions

This study demonstrates that the phenomenon of "grokking", while initially perplexing, can be understood within the established framework of Statistical Learning Theory, particularly through the lens of Algorithmic Stability. By applying these theoretical principles, we clarify how grokking aligns with the core concepts of learning and generalization. Our findings suggest that rather than requiring novel frameworks, existing theories are sufficient to explain this abrupt transition in model performance. This reinforces the adaptability and depth of current theories in addressing emerging, complex behaviors in artificial intelligence.

References

[1] D. Silver, A. Huang, C. J. Maddison, A. Guez, and Others. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

[2] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[3] M. Varadi, D. Bertoni, P. Magana, U. Paramval, and Others. Alphafold protein structure database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic acids research*, 52(D1):D368–D375, 2024.

[4] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, and Others. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

[5] L. Oneto, S. Ridella, and D. Anguita. Towards algorithms and models that we can trust: A theoretical perspective. *Neurocomputing*, 592:127798, 2024.

[6] G. E. Karniadakis, I. G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, and L. Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021.

[7] R. Geirhos, J. H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[8] L. Oneto, S. Ridella, and D. Anguita. Do we really need a new theory to understand over-parameterization? *Neurocomputing*, 543:126227, 2023.

[9] A. Power, Y. Burda, H. Edwards, I. Babuschkin, and V. Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022.

[10] S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[11] A. Gromov. Grokking modular arithmetic. *arXiv preprint arXiv:2301.02679*, 2023.

[12] B. Žunkovič and E. Ilievski. Grokking phase transitions in learning local rules with gradient descent. *Journal of Machine Learning Research*, 25(199):1–52, 2024.

[13] Z. Liu, E. J. Michaud, and M. Tegmark. Omnigrok: Grokking beyond algorithmic data. In *International Conference on Learning Representations*, 2022.

[14] S. Murty, P. Sharma, J. Andreas, and C. D. Manning. Grokking of hierarchical structure in vanilla transformers. *arXiv preprint arXiv:2305.18741*, 2023.

[15] S. Samothrakis, A. Matran-Fernandez, U. Abdullahi, M. Fairbank, and M. Fasli. Grokking-like effects in counterfactual inference. In *International Joint Conference on Neural Networks*, 2022.

[16] Z. Xu, Y. Wang, S. Frei, G. Vardi, and W. Hu. Benign overfitting and grokking in relu networks for xor cluster data. *arXiv preprint arXiv:2310.02541*, 2023.

[17] Z. Liu, Z. Zhong, and M. Tegmark. Grokking as compression: A nonlinear complexity perspective. *arXiv preprint arXiv:2310.05918*, 2023.

[18] Z. Tan and W. Huang. Understanding grokking through a robustness viewpoint. *arXiv preprint arXiv:2311.06597*, 2023.

[19] Z. Liu, O. Kitouni, N. S. Nolte, E. Michaud, M. Tegmark, and M. Williams. Towards understanding grokking: An effective theory of representation learning. *Neural Information Processing Systems*, 2022.

[20] N. Nanda, L. Chan, T. Lieberum, J. Smith, and J. Steinhardt. Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217*, 2023.

[21] K. Lyu, J. Jin, Z. Li, S. S. Du, J. D. Lee, and W. Hu. Dichotomy of early and late phase implicit biases can provably induce grokking. In *International Conference on Learning Representations*, 2023.

[22] V. Thilak, E. Littwin, S. Zhai, O. Saremi, R. Paiss, and J. Susskind. The slingshot mechanism: An empirical study of adaptive optimizers and the grokking phenomenon. *arXiv preprint arXiv:2206.04817*, 2022.

[23] T. Kumar, B. Bordelon, S. J. Gershman, and C. Pehlevan. Grokking as the transition from lazy to rich training dynamics. *arXiv preprint arXiv:2310.06110*, 2023.

[24] C. M. Bishop and H. Bishop. *Deep learning: Foundations and concepts*. Springer Nature, 2023.

[25] O. Bousquet and A. Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

[26] A. Maurer. A second-order look at stability and generalization. In *Conference on learning theory*, 2017.

[27] L. Oneto, A. Ghio, S. Ridella, and D. Anguita. Fully empirical and data-dependent stability-based bounds. *IEEE transactions on cybernetics*, 45(9):1913–1926, 2014.

[28] L. Oneto, S. Ridella, and D. Anguita. Informed machine learning: Excess risk and generalization. In *Neurocomputing (submitted)*, 2024.