Journal of Machine Learning Research 25 (2024) 1-52 Submitted 10/22; Revised 6/24; Published 7/24

Grokking phase transitions in learning local rules with gradient descent

Bojan Žunkovič BOJAN.ZUNKOVIC@FRI.UNI-LJ.SI
University of Ljubljana
Faculty of Computer and Information Science
1000 Ljubljana, Slovenia

Enej Ilievski ENEJ.ILIEVSKI@FMF.UNI-LJ.SI
University of Ljubljana
Faculty of Mathematics and Physics
1000 Ljubljana, Slovenia

Editor: Maxim Raginsky

Abstract

We discuss two solvable grokking (generalisation beyond overfitting) models in a rule-learning scenario. We show that grokking is a phase transition and find exact analytic expressions for the critical exponents, grokking probability, and grokking time distribution. Further, we introduce a tensor network map that connects the proposed grokking setup with the standard (perceptron) statistical learning theory and provide evidence that grokking is a consequence of the locality of the teacher model. We analyze the rule-30 cellular automaton learning task, numerically determine the critical exponent and the grokking time distribution, and compare them with the prediction of the proposed grokking model. Finally, we numerically study the connection between structure formation and grokking.

Keywords: rule learning, tensor networks, phase transitions, structure formation, teacher-student model

Contents

1 Introduction **3**

2 Related work **4**

3 Perceptron grokking **6**

3.1 General setting . . . 7
3.2 1D exponential model . . . 8
3.2.1 Test error dynamics . . . 9
3.2.2 Grokking probability . . . 10
3.2.3 Grokking time . . . 12
3.3 D-dimensional uniform ball model . . . 13
3.3.1 Test error dynamics . . . 15
3.3.2 Grokking probability . . . 16
3.3.3 Grokking time . . . 19
3.3.4 Critical exponents for a general isotropic data PDF . . . 22

4 Learning local rules with shallow tensor networks 23
4.1 Local teacher model . . . 23
4.2 Uniform tensor-network student model . . . 25
4.2.1 Short introduction to tensor network methods . . . 25
4.2.2 Tensor-network attention model . . . 26
4.2.3 Tensor network map . . . 27
4.3 Simulation details and results . . . 28
4.3.1 Constant attention tensors . . . 28
4.3.2 Full model training and structure formation . . . 34

5 Summary and discussion 37

A Grokking time in the 1D model 44

B Grokking probability in the D-dimensional ball model 45
B.1 **Case $\lambda_1 = 0$** . . . 46
B.2 **Case $1 \gg \lambda_1 > 0$** . . . 47

C Fixed attention **47**

D Additional results for the 2–local and the 3–local rules **48**

1. Introduction

Despite recent progress in understanding the double descent phenomena Belkin et al. (2019); Nakkiran et al. (2021); Krogh and Hertz (1991); Pezeshki et al. (2021) we still do not have a complete theory of generalisation in over-parameterised models. Two recent empirical observations, neural collapse Papyan et al. (2020) and grokking (generalisation beyond over-fitting) Power et al. (2022), can help us understand the training and generalisation properties of over-parameterised models.

Neural collapse occurs in the terminal phase of training, i.e. the phase with zero train error. It refers to the collapse of the $N-$dimensional, last-layer features (input to the last/classification layer) Papyan et al. (2020) to a $(C-1)$-dimensional equiangular tight frame (ETF) structure, where C is the number of classes. The feature vectors converge towards the vertices of the ETF structure such that features for each class are close to one vertex. Also, the distance between the vertices is much larger than all intraclass feature variances. We can partially understand neural collapse within the unconstrained features and local elasticity models Kothapalli et al. (2022). However, its role in generalisation, relation to grokking, and appearance of different latent space structures are still not completely understood.

Grokking also occurs during the final phase of training. When training on algorithmic datasets past the zero train error, a sudden decrease of the test error from approximately one to zero is observed Power et al. (2022). The grokking phenomenon has been discussed within an effective theory approach Liu et al. (2022), where an empirical connection between representation/structure formation and generalisation has been made. An empirical study Thilak et al. (2022) established a relation between grokking and spikes in training loss and an increase in weight norm. However, no exactly solvable model exhibiting the grokking phenomenon has been discussed so far. Furthermore, it is unclear how to reconcile grokking with standard generalisation theory based on statistical learning methods Engel and Van den Broeck (2001). The statistical learning theory predicts (in a teacher-student setting) an algebraic (as $t^{-\nu}$, where $\nu = 1$ for most learning rules) decrease of the generalisation error with training time t (or the number of training samples) Engel and Van den Broeck (2001).

Grokking and neural collapse (or latent-space structure formation in general) have been observed in over-parametrised models. However, we do not know the minimal framework within which we can understand these phenomena or if they are genuinely deep-network phenomena. We aim to formulate a simple solvable model of grokking and relate it to latent-space structure formation and other common deep-network training features, e.g. spikes in the training loss.

Main contributions– We have four main contributions:

- We propose a simple learning scenario that exhibits grokking (Section 3). We study two solvable models where grokking is a *phase transition* to zero test error and calculate exact critical exponents and grokking-time distributions.
- We discuss the teacher-student model by using the tensor network approach. We map the sequence prediction model in the thermodynamic limit to the proposed grokking setup (Section 4) and provide evidence that grokking is a consequence of the locality of the teacher model.
- We numerically study grokking and structure formation on the example of learning a 1D rule-30 cellular automaton (Section 4). We show that sudden spikes in the training loss correspond to structural changes in the latent space representation of the data.
- Our analytical results and numerical experiments show a significant difference between L_1 and L_2 regularisations. The L_1 regularised models have a higher probability of grokking, shorter generalisation time, and smaller effective dimensions compared to L_2 regularised models.

Broader impact– The proposed exactly solvable grokking models are a step towards theoretical understanding of the late learning phase and generalisation benefits of the terminal phase of training. The introduced tensor-network map connects the standard teacher-student setup in the thermodynamic limit with the proposed grokking setup. It offers a new tool for studying generalisation properties of local rules (local teacher-student models), which could lead to more complex learning dynamics (compared to the standard infinite-range rules).

Although based on simple models, our results can also be relevant for deep learning training practice. We conjecture that good generalisation is more likely in models with latent space data distributions with small effective dimensions. Our results hint that L_1 regularisation can improve the generalisation properties of deep models compared to L_2 regularisation. Further, we show that spikes in the loss (which often occur during the training of deep neural networks) correspond to latent space structural changes that can be beneficial or detrimental for generalisation. Assuming this is also the case in deep networks, we can use the information about the effective dimension of latent space to revert the model to a state before the spike or continue training with the current model. Finally, our empirical results show that regularisation is beneficial in the early stages of training. In the late stages of training, regularization often leads to detrimental latent space structural changes that increase the test error.

2. Related work

A sudden transition from zero to 100% accuracy in algorithmic datasets in over-fitted transformer models has been first described in Power et al. (2022) and named grokking. In the grokking phase, a formation of simple structures reflecting the properties of the problem has been observed. This finding contradicts the common practice of early stopping and supports recent observations on the benefits of the terminal phase of training Soudry et al. (2018); Belkin et al. (2019); Merrill and Tsilivis (2022) and the double descent phenomenon Belkin et al. (2019); d'Ascoli et al. (2020); Nakkiran et al. (2021); Pezeshki et al. (2022). In Liu et al. (2022), an effective theory of grokking has been proposed. Within the effective theory, one can access the critical training size to observe grokking. The authors relate grokking with a good representation (or structure formation) and introduce it as

a phase between generalisation and memorisation. We go beyond these findings since we obtain exact solutions for the proposed setup and even calculate the grokking-time probability density function (PDF). A systematic experimental study of the grokking phenomena has been presented in Thilak et al. (2022). A slingshot mechanism (related to the edge of stability Cohen et al. (2021)) has been proposed as a necessary condition for grokking. The slingshot mechanism refers to the occurrence of cyclic spikes in training loss and steps in the weight norms during training. The slingshot behaviour is not restricted to algorithmic datasets but is also present in various common classification tasks Thilak et al. (2022). We find a similar behaviour, i.e. that the grokking coincides with train loss spikes. Moreover, we connect training loss spikes with a discontinuous step-like evolution of the effective dimension of the latent space representation of the data, which indicates structural changes in the latent space representation.

A particular structure formation common in deep classification neural networks is neural collapse (NC). It refers to four empirical/numerical observations in training deep neural network classifiers Papyan et al. (2020):

- **(NC1) Variability collapse** – variations of within class features become negligible
- **(NC2) Convergence to equiangular tight frame (ETF)**– class mean vectors form equal-sized angles between any given pair
- **(NC3) Convergence to self-duality**– the class means and linear classifiers converge to each other, up to rescaling
- **(NC4) Simplification to nearest class centre**– the network classifier converges to a classifier that selects the class with the nearest train class mean.

The role of the loss function, the regularisation, the batch normalisation, and the optimizer have been discussed within the unconstrained features model Mixon et al. (2020); Fang et al. (2021); Kothapalli et al. (2022) and the local elasticity assumption Kothapalli et al. (2022). The relation of NC to generalisation properties has been discussed in Zhu et al. (2021); Hui et al. (2022); Kothapalli et al. (2022). However, no relation to grokking has been discussed so far. Although we do not observe NC as defined above, our findings regarding the spikes in the training loss and latent-space data structure might also be relevant for the NC dynamics.

Our main technical tools are tensor networks, which are models obtained by contracting many low-dimensional tensors. Tensor networks have been very successful in modelling many-body quantum systems. Recently, they have also been applied to machine learning tasks. In particular to classification problems Stoudenmire and Schwab (2016); Stoudenmire (2018); Efthymiou et al. (2019); Liu et al. (2019); Martyn et al. (2020); Meng et al. (2020); Chen et al. (2021); Kong et al. (2021), generative modelling Cheng et al. (2019); Stokes and Terilla (2019); Sun et al. (2020); Liu et al. (2021), sequence and language modelling Pestun and Vlassopoulos (2017); Guo et al. (2018); Bradley et al. (2020); Bradley and Vlassopoulos (2020); Žunkovič (2022), anomaly detection Wang et al. (2020); Streit et al. (2020). Besides, tensor networks have been used as tools to advance machine learning theory by a derivation of interesting generalisation bounds Bradley et al. (2020), information theoretical insights Cohen et al. (2016); Deng et al. (2017); Levine et al. (2017); Glasser et al. (2019), and new connections between machine learning and physics Chen et al. (2018); Dymarsky and Pavlenko (2021); Adhikary et al. (2021). Particularly relevant for latent space structure formation is the connection between recurrent neural networks

(RNN) and matrix product states Wu et al. (2022). In Merrill and Tsilivis (2022) benefits of the terminal phase of training for the extraction of state automata from RNNs (and hence matrix product state tensor networks) have been discussed. The authors find internal state space compression and increased extraction in the terminal phase of training. This is similar to our findings of reduced effective dimension in the latent (internal) space. In contrast to Merrill and Tsilivis (2022), we introduce a new tensor network, similar to the tensor-network attention model Žunkovič (2022), and study grokking and structure formation in a teacher-student learning setup.

Finally, our work is related to the statistical-mechanics theory of supervised learning Engel and Van den Broeck (2001). In the case of the supervised teacher-student perceptron, an algebraic decrease in generalisation error with the size of the training set (training time) has been predicted Engel and Van den Broeck (2001). A first-order phase transition has been derived only in a restricted setting of discrete weights Engel and Van den Broeck (2001). These results are typically based on the replica method Gardner and Derrida (1989) which requires the thermodynamic limit, where both the number of samples and the dimension are large and their ratio is fixed. Recent outstanding results in this direction concern the analysis of optimal generalisation errors of generalised linear models Barbier et al. (2019); Carleo et al. (2019). We study the same teacher-student scenario but with a restriction to a local teacher (still within the thermodynamic limit). The locality of the teacher/rule enables us to map the problem to a finite-dimensional latent space where we discuss grokking (a second-order phase transition) and latent-space structure formation.

3. Perceptron grokking

In this section, we present a simplified model of grokking in a binary classification scenario. To illustrate our model, a trained deep neural network is divided into two components, as depicted in Fig. 1. The first component transforms the input probability distribution $P_{\mathcal{D}}$ of the training data, into two distributions that are linearly separable - one for positive samples and one for negative samples. The second component is a linear classifier. We avoid the feature learning part by replacing it with separable probability distributions $P^{\pm}$. We then examine the training dynamics of the linear model using gradient descent optimization, given a training dataset containing N positive and N negative samples from the distributions $P^{\pm}$. This training dynamics does not account for feature learning, which is a crucial aspect of the grokking process, as discussed in Section 4. Nonetheless, we show that even this simplified model captures certain aspects of grokking and allows for an analytical understanding of two probability distributions of interest: the probability of achieving zero test error ($P_{E(\infty)=0}$), and the probability density $P(t_{\text{G}})$, where t_{G} represents the time difference between the instances of zero test error and training error.

The training error is determined by the fraction of incorrectly classified training samples. The test error is calculated as a statistical average of incorrectly classified samples over the probability distributions $P^{\pm}$. This is analogous to the statistical average of an observable, e.g. magnetisation, over all possible configurations of the system. In our case, the number of test samples is infinite, in analogy to the thermodynamic limit in statistical models. The models thus permit non-analytic behaviour of the order parameter. Specifically, in our case, the order parameter is the test error showing a training phase transition. The controlled parameter of the transition is the training time t, in analogy with the inverse temperature in a thermodynamic setting.

The described setting encompasses behaviours that range from the common training dynamics with minimal t_{G} to the recently observed generalisation (zero test error) long after the training error vanishes, i.e. long t_{G}. Analytical results for $P_{E(\infty)=0}$ and P_{G} provide an important insight into how probable generalisation beyond overfitting is for a certain class of probability distributions and how training dynamics are affected by different choices of regularization schemes.

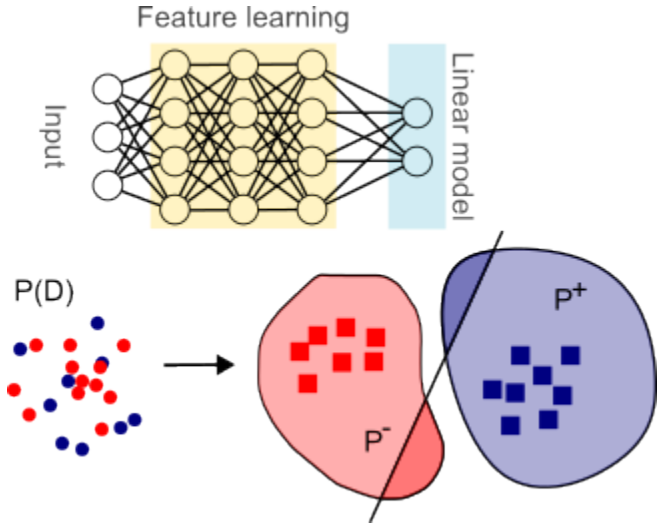


Figure 1: A schematic separation of the deep neural network into a feature learning part and a linear classification part. The feature learning part transforms the input distribution $P_{\mathcal{D}}$ into latent space distributions P^+ and P^-. Blue and red dots represent the training data in the input space and the blue and red squares represent the training data in the latent space, e.g., penultimum activations in a deep neural network. Our main assumption is the linear separation of the probability distributions in the latent space P^+ and P^-, represented by the red and blue shaded regions. The linear model represented by the black line does not separate all the test data while correctly classifying the training data (blue and red squares).

3.1 General setting

We consider a simple binary classification problem that exhibits the grokking phenomena. Let us assume that we have a dataset $\mathcal{D}$ consisting of $(\tilde{x}^i, y^i) \in \mathcal{D}$, with two linearly separable classes ($y^i \in \{-1, 1\}$) and $D-$dimensional features $\tilde{x}^i \in \mathbb{R}^D$. More precisely, the probability densities for the positive class (P^+) and the negative (P^-) class are linearly separable in $\mathbb{R}^D$. Our model class is a simple perceptron in D dimensions, namely

$$f(\tilde{x}) = \mathrm{sgn}(\hat{y}), \quad \hat{y} = w \cdot \tilde{x} + b, \tag{1}$$

where $w, x \in \mathbb{R}^D$ and $b \in \mathbb{R}$. We sample N positive and N negative samples and then train the model with gradient descent

$$\begin{aligned} \frac{\partial \theta}{\partial t} &= -\frac{\partial \mathcal{R}}{\partial \theta}, \\ \mathcal{R} &= \frac{1}{2N}\sum_{i=1}^{2N} \frac{1}{2}|\hat{y}^i - y^i|^2 + \lambda_1 ||\theta||_1 + \frac{\lambda_2}{2}||\theta||_2, \end{aligned} \tag{2}$$

where λ_1, λ_2 denote regularisation parameters, θ denotes the collection of all model parameters w, b, and $||\bullet||_{1,2}$ denote the one and two norm. By construction, the setup captures datasets that enable training the model to zero test error. We aim to study how the choice of latent space probabilities $P^{\pm}$ and various training parameters influence the training dynamics and the appearence of the grokking phenomenon. In particular, the probability of grokking, the grokking time, and the critical exponent.

The suggested setup is relevant in the transfer learning scenario, where we initially train a model on one task and then retrain only the final classification layer on another task. Additionally, in Section 4 we construct a tensor network map that connects the standard teacher-student statistical learning scenario in the thermodynamic limit to the setup described above.

In the following, we will explicitly calculate the dynamics of the model parameters, the dynamics of the test error, the critical exponent, the probability of grokking, and the probability density function (PDF) of grokking time for particular choices of the dimension D and probability densities $P^{\pm}$.

3.2 1D exponential model

We start by considering a simple, one-dimensional model where we obtain all results in closed form. Although the model does not apply to the real-world scenario it captures several qualitative features and provides a starting point to study more realistic models.

The distribution of the data set is shown in Fig. 2. Positive and negative samples follow the same probability, i.e. $P^+(\tilde{x}) = P^-(-\tilde{x})$. The minimal distance between the positive and negative samples is 2ϵ, therefore $P^{\pm}(|\tilde{x}| \leq \epsilon) = 0$.

Since the input x is one dimensional Eq. 1 reduces to

$$f(x) = \mathrm{sgn}(x - b), \tag{3}$$

where b is the sole model parameter (we fix the weight $w = 1$). As described above, we train the model with gradient descent and loss

$$\mathcal{R} = \frac{1}{2N}\sum_{i=1}^{2N} \frac{1}{2}((\tilde{x}^i - b) - y^i)^2 + \frac{\lambda_2 b^2}{2} + \lambda_1 |w|. \tag{4}$$

We also assume that the training dataset is balanced, i.e. $\sum_{i=1}^{2N} y^i = 0$.

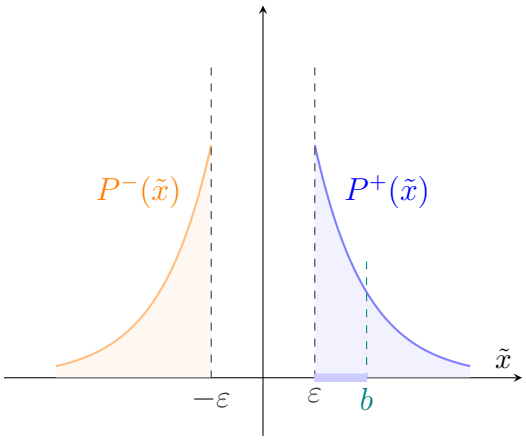


Figure 2: A schematic representation of the linearly separable random 1D model. The model is represented by b. The samples between ϵ and b (marked by a thick blue line) are incorrectly classified.

3.2.1 TEST ERROR DYNAMICS

First, we calculate the model parameter dynamics governed by the negative gradient of the loss function

$$\frac{\partial b}{\partial t} = -\frac{\partial \mathcal{R}}{\partial b} = \frac{1}{2N}\sum_{i=1}^{2N}(\tilde{x}^i - b - y^i) - \mathrm{sgn}(b)\lambda_1 - \lambda_2 b = \bar{x} - \mathrm{sgn}(b)\lambda_1 - (1+\lambda_2)b, \tag{5}$$

where $\bar{x} = \frac{1}{2N}\sum_{i=1}^{2N}\tilde{x}^i$ denotes the average over training inputs. The solution to Eq. 5 with the initial condition $b(0)$ is

$$b(t) = \bar{x}_\lambda - (\bar{x}_\lambda - b(0))\,\mathrm{e}^{-(1+\lambda_2)t}, \quad \bar{x}_\lambda = \begin{cases} \frac{\bar{x}-\lambda_1}{1+\lambda_2}, & b(t) \geq 0 \\ \frac{\bar{x}+\lambda_1}{1+\lambda_2}, & b(t) < 0 \end{cases}. \tag{6}$$

In the following, we assume that $b(0) > 0$. To have a nontrivial train- and test-error dynamics, we also assume that $x_{\min} < b(0)$, where $x_{\min}$ denotes the minimum of the positive samples in the training dataset.

To obtain explicit expressions for the test error we choose the exponential distribution of the samples $P^+(x) = \mathrm{e}^{-(x-\epsilon)}\Theta(x-\epsilon)$, where $\Theta(x)$ denotes the Heaviside step function. In this case, the cumulative probability to get a sample with $b < x$ is given by $P(x > b) = \mathrm{e}^{\epsilon - b}$ and the test error is

$$E(t) = \begin{cases} \frac{1}{2}(1 - \mathrm{e}^{\epsilon - b(t)}) & b(t) > \epsilon \\ 0 & \text{else} \end{cases}, \tag{7}$$

where $b(t)$ is determined by Eq. 6. We are interested in the behaviour of the test error just before the test error drops to zero. The time at which the test error becomes zero (t_ϵ) is given by

$$t_\epsilon = \log\left(\frac{b(0) - \bar{x}_\lambda}{\epsilon - \bar{x}_\lambda}\right). \tag{8}$$

By expanding the test error (Eq. 7) around t_ϵ we get

$$E(t < t_\epsilon) \approx \frac{(\epsilon - \bar{x}_\lambda)}{2}(1 + \lambda_2)(t_\epsilon - t). \tag{9}$$

Interestingly, the first-order coefficient, as well as the critical exponent, do not depend on the initial condition (assuming $b(0) > \epsilon$). We now calculate the average test error over different initial conditions by aligning the phase-transition points t_ϵ. Since $P(-\bar{x}) = P(\bar{x})$, we find the test error

$$\langle\langle E(t)\rangle\rangle \approx \frac{\epsilon_\lambda}{2}(t_\epsilon - t), \quad \epsilon_\lambda = \epsilon(1 + \lambda_2) + \lambda_1, \tag{10}$$

where $\langle\langle \bullet \rangle\rangle$ denotes the average over all valid initial conditions $b(0)$ and training input averages $\bar{x}$.

Grokking in the considered 1D exponential model is a second-order phase transition with the test-error critical exponent equal to one. The regularisation parameters and the distance between positive and negative class distributions change only the prefactor. In general, we expect that the critical exponent depends on the distribution as well as the training parameters, e.g. regularisation strength.

An example for a particular dataset with $N = 10$ and $\bar{x} = 0.7$ is shown in Fig. 3. We set $\lambda_{1,2} = 0$, $\epsilon = 1$ and the initial value of the model parameter $b(0) = 21$. This example illustrates that a simple 1D exponential model already captures the phenomenon of generalisation beyond overfitting (grokking). However, due to the concentration of the exponential probability near the boundary, the typical training datasets sampled from the probability distribution considered in this section will not have a large separation between the times at which zero test and training errors are achieved and would hence not be regarded as an example of grokking. Still, our model captures datasets that exhibit grokking. We obtain the non-typical dataset shown in Fig. 3 by first sampling the positive and negative samples according to the exponential distributions and then shifting the positive samples to the right and the negative samples to the left by the same amount. In the right panel of Fig. 3 we shift by five.

In the next sections, we also calculate the probability of obtaining zero final test error and the probability density of the time separation between the zero test and training errors.

3.2.2 Grokking probability

We are also interested in the probability of sampling a training dataset with which we can train the model to zero test error. We name this probability the *grokking probability*. In the considered 1D case the final test error vanishes only if $|\bar{x}_\lambda| < \epsilon$. We express this condition for zero test/generalisation error of the trained model as

$$|\bar{x}| < \epsilon_\lambda, \tag{11}$$

where ϵ_λ is given by Eq. 10. Since our training dataset has an equal number of positive and negative samples, we need to consider the distribution of the mean of N independent exponentially distributed variables, which is given by the gamma distribution

$$P_N^{exp}(\bar{x}) = \frac{N^N}{\Gamma(N)}\bar{x}^{N-1}\mathrm{e}^{-N\bar{x}}\Theta(\bar{x}), \tag{12}$$

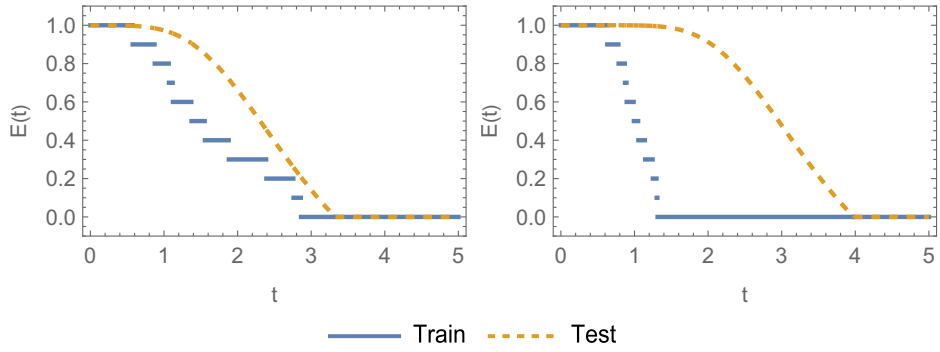


Figure 3: An example of train and test error dynamics with a specific dataset sampled from the 1D model with exponential probabilities. We show the train and test error on the positive samples since all negative samples are correctly classified from the start. The left panel corresponds to a typical dataset, where no grokking is observed. The right panel corresponds to a training dataset far from the classification boundary obtained by shifting the positive and negative samples away from the origin by five. In the non-typical (shifted) case, we find a region with a large test error while the training error vanishes. In both cases $N = 10$, $\bar{x} = 0.7$, $\epsilon = 1$, $\lambda_{1,2} = 0$.

where $\Gamma(N)$ denotes the gamma function. First, we calculate the probability $P_N(\bar{x})$ to get the average $\bar{x}$

$$\begin{aligned} P_N(\bar{x}) =& \int_{\bar{x}_+=0}^{\infty} \mathrm{d}\bar{x}_+ P_N^{\exp}(\bar{x}_+) \int_{\bar{x}_-=0}^{\infty} \mathrm{d}\bar{x}_- P_N^{\exp}(\bar{x}_-)\delta\left(\bar{x} - (\bar{x}_+ - \bar{x}_-)/2\right) \\ =& \frac{2N^{N+\frac{1}{2}}\bar{x}^{N-\frac{1}{2}}K_{N-\frac{1}{2}}(2N\bar{x})}{\sqrt{\pi}\Gamma(N)}, \end{aligned} \tag{13}$$

where $K_n(z)$ denotes a modified Bessel function of the second kind. The probability to get the dataset with zero test error is then given by

$$\begin{aligned} P_{E(\infty)=0}(\epsilon_\lambda, N) =& 2\int_{\bar{x}=0}^{\epsilon_\lambda} P_N(\bar{x})\mathrm{d}\bar{x} \\ =& \sqrt{\pi}(-1)^N (B\epsilon_\lambda)^{2N}\, {}_1\tilde{F}_2\left(N; N+\frac{1}{2}, N+1; N^2\epsilon_\lambda^2\right) \\ &+ \frac{\pi(-1)^{N+1}N\epsilon_\lambda\, {}_1\tilde{F}_2\left(\frac{1}{2}; \frac{3}{2}, \frac{3}{2}-N; N^2\epsilon_\lambda^2\right)}{\Gamma(Nd)}, \end{aligned} \tag{14}$$

where ${}_p\tilde{F}_q\left(a; b; z\right)$ is the regularized generalized hypergeometric function. The above expression (Eq. 14) simplifies for a particular choice of N, e.g. for $N = 2$ we get

$$P_{E(\infty)=0}(\epsilon_\lambda, N = 2) = 1 - (1 + 2\epsilon_\lambda)\mathrm{e}^{-4\epsilon_\lambda}. \tag{15}$$

In Fig. 4 we show several numerically exact grokking probabilities. As expected, the grokking probability increases with the number of training samples and the effective separation between the two classes determined by ϵ_λ.

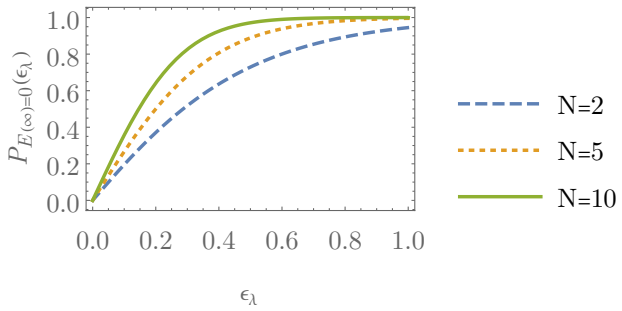


Figure 4: Exact numerical calculation for grokking probabilities for a different number of training samples $N = 2$ (dashed blue line), 5 (dotted orange line), and 10 (full green line). Grokking probability increases with N and ϵ_λ.

The effect of the L_1 and L_2 regularisations on the trained model is different. The L_2 regularisation is multiplicative, and the L_1 regularisation is additive concerning the gap between the positive and negative samples ϵ. Hence, in the case of a small gap, the L_1 regularisation becomes much more effective. In other words, for an infinitesimal gap ($\epsilon \ll 1$) and finite N, the L_1 regularisation ensures that the probability of zero test error is finite. This is not the case when using the L_2 regularisation.

It would be interesting to see if L_1 regularisation is preferred to L_2 regularisation also in more realistic scenarios. We find a similar distinction between the L_1 and L_2 normalized models also in the more general grokking scenario discussed in Section 3.3.2.

3.2.3 Grokking time

We define the *grokking time* as the difference between t_ϵ (zero-test-error time) and the time when the training error becomes zero. In our simple case, we have

$$t_G = \frac{1}{1+\lambda_2} \log\left(\frac{\epsilon + x_{\min} - \bar{x}_\lambda}{\epsilon - \bar{x}_\lambda}\right). \tag{16}$$

The grokking time does not depend on the initial condition as long as $b(0) > x_{\min}$. To find the grokking-time PDF, we need to calculate the distribution $P_N(\bar{x}, x_{\min})$ and then consider only the part $|\bar{x}| \leq \epsilon_\lambda$. We provide the details of the calculation in the Appendix A. For a finite N it is possible to obtain a closed-form expression which, however, is not very instructive. Here we provide the unnormalised grokking-time PDF for $N = 2$

$$\begin{aligned} P^{\text{unnorm}}_{N=2,\epsilon,\lambda_1}(t) = &\frac{1}{8} e^{-\frac{4e^t(2e^t+5)\epsilon_\lambda}{e^t+1}-2t}\Bigg(\exp\left(\frac{4e^t\epsilon_\lambda(3\sinh(t)+\cosh(t)+4)}{e^t+1}+3t\right) \\ &-64\epsilon_\lambda^2 e^{4\left(4-\frac{3}{e^t+1}\right)\epsilon_\lambda+3t} - 24\epsilon_\lambda e^{2\left(\left(8-\frac{6}{e^t+1}\right)\epsilon_\lambda+t\right)} - 8\epsilon_\lambda e^{4\left(4-\frac{3}{e^t+1}\right)\epsilon_\lambda+t} \\ &-2e^{\frac{4(4e^t+1)\epsilon_\lambda}{e^t+1}} - 3e^{4\left(4-\frac{3}{e^t+1}\right)\epsilon_\lambda+t} - \left(e^t+1\right)e^{4\left(2e^t-\frac{1}{e^t+1}+2\right)\epsilon_\lambda}\left(e^t\left(e^t-8\epsilon_\lambda-1\right)-2\right)\Bigg), \end{aligned} \tag{17}$$

which we normalise by dividing with the appropriate grokking probability, see Eq. 14.

In Fig. 5 we show several numerically exact grokking-time PDFs. The expected grokking time is smaller with increasing training size N and effective class separation ϵ_λ. This is consistent with the observations of Power et al. (2022); Liu et al. (2022) where a shorter grokking time has been reported for an increased number of training samples and a larger weight decay.

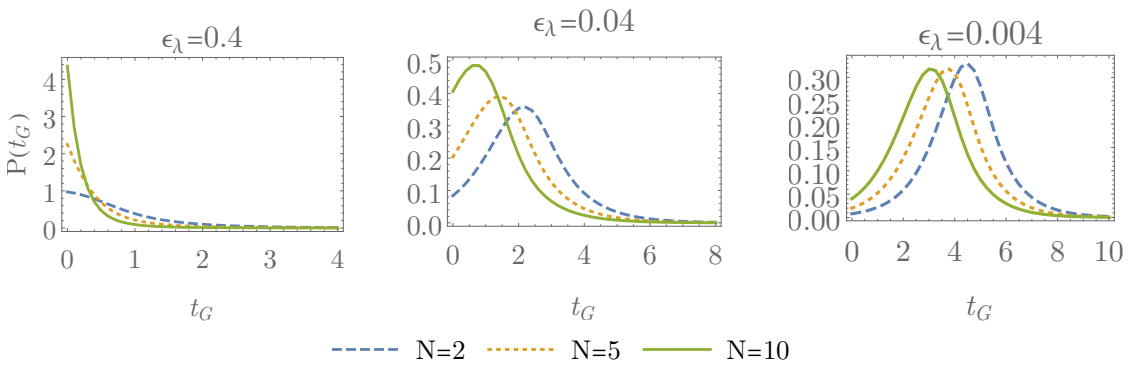


Figure 5: Grokking-time PDF for a different number of training samples N =2 (dashed blue line), 5 (dotted orange line), 10 (full green line). Left, middle, and right panels correspond to ϵ_λ =0.4 (left), 0.04 (middle), 0.004 (right). Grokking time is shorter with increasing N and ϵ_λ.

3.3 D-dimensional uniform ball model

The second model we consider is shown in Fig. 6. The inputs are D dimensional real vectors $\tilde{x} \in \mathbb{R}^D$. Positive and negative samples are distributed uniformly in unit balls shifted from the origin by vectors $\pm\epsilon \in \mathbb{R}^D$. We will assume that the shift is along the first dimension, i.e. $\epsilon_1 = \varepsilon$, and $\epsilon_{j>1} = 0$. The student model is a simple perceptron determined by the vector $w \in \mathbb{R}^D$ (we set the bias to zero $b = 0$)

$$f(\tilde{x}) = \mathrm{sgn}(\tilde{x} \cdot w). \tag{18}$$

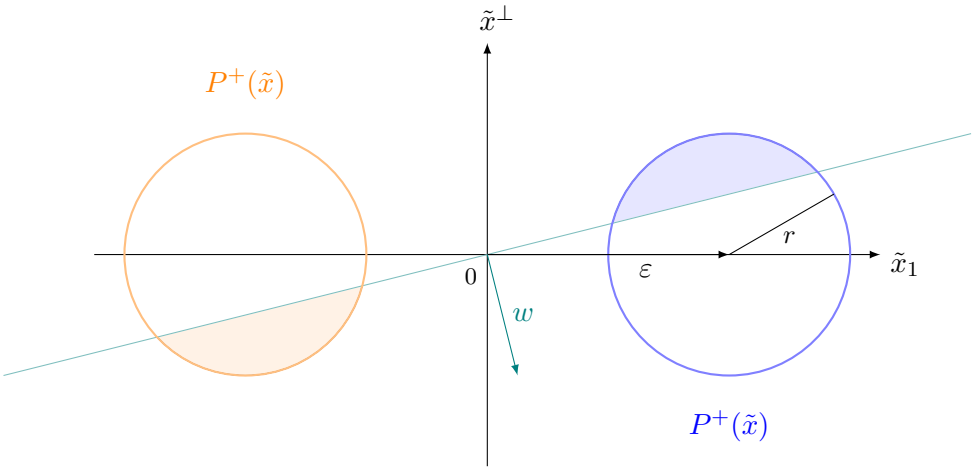


Figure 6: A two-dimensional projection of the $D-$dimensional uniform-ball model on the plane defined by the shift vector ϵ and the model vector w. The positive and negative samples are uniformly distributed in unit balls shifted away from the origin along the $\tilde{x}_1$ axis by $\pm\epsilon$, respectively. The model used to separate the classes is a linear model (determined by the vector w) going through the origin (green line). The volume of the shaded red and blue regions determines the test error.

We write the training loss with L_1 and L_2 regularisation as

$$\begin{aligned}
\mathcal{R} =& \frac{1}{2N}\sum_{i=1}^{2N}\frac{1}{2}(\tilde{x}^i\cdot w - y^i)^2 + \frac{\lambda_2||w||_2^2}{2} + \lambda_1|w|_1. \\
=& \frac{1}{2}w\cdot\left(\frac{1}{2N}\sum_{i=1}^{2N}\tilde{x}^i\otimes\tilde{x}^i + \lambda_2\mathbb{1}_D\right)w - w\cdot\left(\frac{1}{2N}\sum_{i=1}^{2N}y^i\tilde{x}^i - \lambda_1\text{sgn}(w)\right) + \frac{1}{2} \\
=& \frac{1}{2}w\cdot Gw - w\cdot a + \frac{1}{2}, \\
G =& \frac{1}{2N}\sum_{i=1}^{2N}\tilde{x}^i\otimes\tilde{x}^i + \lambda_2\mathbb{1}_D, \\
a =& \frac{1}{2N}\sum_{i=1}^{2N}y^i\tilde{x}^i - \lambda_1\text{sgn}(w).
\end{aligned} \tag{19}$$

Again we assume that the training dataset is balanced, i.e. $y^{i\leq N}=1$, $y^{i>N}=-1$, $\tilde{x}^{i\leq N}=x^i+\epsilon$, and $\tilde{x}^{i>N}=x^i-\epsilon$. In our model, x^i are distributed uniformly in a $D-$ dimensional ball centred at the origin.

The presented model is relevant in the transfer learning setting Weiss et al. (2016), if only the last layer of a network is retrained, and with sigmoid activation functions in the penultimate layer. If the model transfers well to a new classification task, the latent space distributions of the new classes are linearly separable and can be bounded by a D-dimensional ball. Since we do not know the details of the distributions, we assume the uniform dis-

tribution in the ball. Further, the positive and negative feature distributions might be embedded in a higher dimensional latent space. In this case, D corresponds to the effective dimension of the data, which can be calculated from the covariance matrix. As we will see in the next section, the introduced D-dimensional ball model qualitatively reproduces the critical exponent and the grokking-time PDF in a local-rule learning problem.

3.3.1 Test error dynamics

To determine the test error dynamics we first specify the dynamics of the model parameters which are determined by the negative gradient of the loss function

$$\frac{\partial w}{\partial t} = -\frac{\partial \mathcal{R}}{\partial w} = -Gw + a. \tag{20}$$

The solution to Eq. 20 with the initial condition $w(0)$ is

$$w(t) = w^\lambda - \left(w^\lambda - w(0)\right) \mathrm{e}^{-Gt}, \quad w^\lambda = G^{-1}a. \tag{21}$$

If $N < D/2$ and $\lambda_2 = 0$, the matrix G is not invertible. In this case, we use the pseudo-inverse.

The test error is given by the volume of an ϵ-shifted, unit ball that is cut out by the plane defined by the vector w. The relevant parameter determining this volume is the distance h between the plane and the origin of the ball. We find that $h = \varepsilon \frac{w_1}{||w||_2}$, where $\varepsilon = ||\epsilon||_2$ and assuming $w_1 > 0$. The critical angle is given by $w_1/||w||_2 = \frac{1}{\varepsilon}$. For larger $w_1/||w||_2$ the error is zero. For smaller values of $w_1/||w||_2$ the error is given by

$$E_D(h) = \begin{cases} \frac{1}{2} - \frac{D\Gamma\left(\frac{D}{2}\right)}{2\sqrt{\pi}\Gamma\left(\frac{D+1}{2}\right)} \, {}_2F_1\left(\frac{1}{2}, \frac{1-D}{2}; \frac{3}{2}; h^2\right) h & , \quad h \leq 1 \\ 0 & , \quad h > 1 \end{cases}, \tag{22}$$

where ${}_2F_1(a, b; c, z)$ represents the Gaussian hypergeometric function. For $h < 1$ and close to the critical point $h \approx 1$ we find the following expression for the test error

$$\begin{aligned} E_D(t) \approx & \frac{D 2^{\frac{D-3}{2}} \Gamma\left(\frac{D}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{D+3}{2}\right)} (1 - h(t))^{\frac{D+1}{2}} \\ = & \frac{D 2^{\frac{D-3}{2}} \Gamma\left(\frac{D}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{D+3}{2}\right)} \left(1 - \epsilon \frac{w_1(t)}{||w(t)||_2}\right)^{\frac{D+1}{2}} \\ = & \frac{D 2^{\frac{D-3}{2}} \Gamma\left(\frac{D}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{D+3}{2}\right)} (k_G(t - t_\epsilon)))^{\frac{D+1}{2}}, \end{aligned} \tag{23}$$

where t_ϵ is defined as the time at which the test error vanishes, and the coefficient k_G is given by the linear expansion of $w(t)$ around t_ϵ. The critical exponent is hence determined only by the dimensionality of the feature distribution.

As in the previous 1D exponential model, the typical dataset sampled from the initial probability distribution does not display the grokking behaviour. Still, we find a finite, albeit very small, probability to sample a training dataset with grokking dynamics. We obtain the non-typical dataset by sampling according to the following procedure. First,

we sample a D-dimensional normal distribution with zero mean and unit variance. Then, we shift the positive samples in the direction of the first coordinate (parallel to ε) and the second coordinate (perpendicular to ε) for the same amount. We shift the negative samples in the opposite direction to the positive samples. After the shift, we normalize the samples and multiply them with a square root of a random number in the interval $[r_0, 1]$. If the shift and r_0 are zero, the procedure corresponds to a random sample from the original uniform D-dimensional ball distribution. In Fig. 7 we show a simple example with $D = 128$, $\epsilon = 1.2$, $\lambda_1 = 0$, $\lambda_2 = 0.1$ and the initial condition with maximum test and train error. We use the shift equal to 10 and $r_0 = 0.8$ to obtain the non-typical dataset in the right panel of Fig. 7.

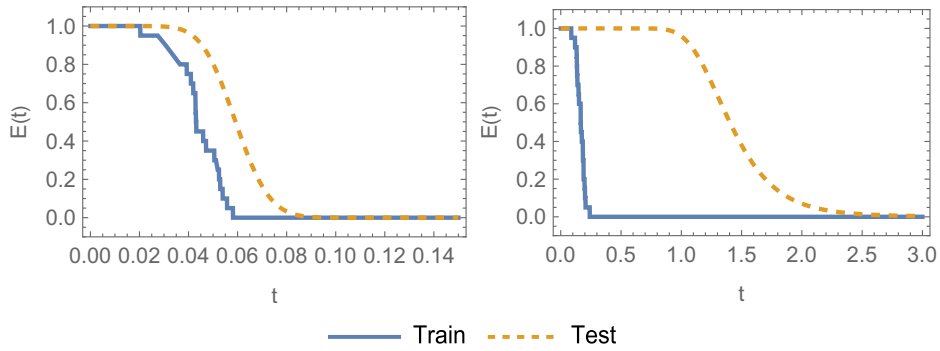


Figure 7: An example of train and test error dynamics with a specific dataset sampled from the D-dimensional ball model with uniform distributions. We show the train and test error as a function of training time. The left panel corresponds to a typical dataset obtained by random sampling according to the given uniform distribution, where no grokking is observed. The right panel corresponds to a shifted training dataset far from the typical region where most samples are concentrated (see main text for details). In the non-typical dataset case, we find a region with a large test error while the training error vanishes. As expected the region is larger with larger shifts. In both cases $N = 10$, $D = 128$, $\epsilon = 1.2$, $\lambda_1 = 0$, and $\lambda_2 = 0.1$.

In the setting discussed here, the main condition to observe generalisation beyond overfitting is, that the positive and negative training samples $\tilde{x}^i$ are far from the concentration of the probabilities $P^{\pm}(\tilde{x}^i)$. In the 1D model, the largest probability density was close to the boundary at ϵ. In the uniform D-dimensional ball distribution considered here, the concentration of the probability density is in the middle of the ball as $D \gg 1$. Non-typical training dataset in combination with measure concentration partly explains the absence of grokking in more realistic scenarios, where training samples are expected to be typical and closer to the maximum probability density.

3.3.2 Grokking probability

Next, we will calculate the probability of training a model with zero test error (*grokking probability*) for a given number of positive/negative samples N. The condition for the final

test error to be zero is given by

$$\frac{w_1^\lambda}{||w^\lambda||_2} \geq \frac{1}{\epsilon}. \tag{24}$$

It will be useful to write the zero test error condition in terms of components of w^λ

$$(\epsilon^2 - 1)\left(w_1^\lambda\right)^2 \geq \sum_{j=2}^{D}\left(w_j^\lambda\right)^2 = r, \tag{25}$$

where r denotes the 2-norm squared of the final weights vector w^λ with the first component equal to zero.

A general calculation of the grokking probability and the grokking-time PDF is not feasible since we would have to invert a random matrix G. Therefore, we consider the limit of many training samples $N \gg 1$, where the matrix G decomposes into a diagonal part proportional to $\lambda_{2,D} = \lambda_2 + \frac{1}{D+2}$ and an off-diagonal part proportional to $N^{-1/2}$. We provide the full derivation of the grokking probability in this limit in Appendix B. Here we consider a simpler case, where we additionally assume that $N \gg \lambda_{2,D} \gg 1$. In this case, G is approximately proportional to the identity

$$G \approx \lambda_{2,D}\mathbb{1}_D + \epsilon \otimes \epsilon. \tag{26}$$

The inverse is

$$[G^{-1}]_{i,j} \approx \begin{cases} \left(\lambda_{2,D} + \varepsilon^2\right)^{-1} & , \ i = j = 1 \\ \lambda_{2,D}^{-1} & , \ i = j \neq 1 \\ 0 & , \ i \neq j \end{cases}. \tag{27}$$

Finally, we get

$$w_{i=1}^\lambda \approx \frac{a_1}{\varepsilon^2 + \lambda_{2,D}}, \quad w_{i>1}^\lambda \approx \frac{a_i}{\lambda_{2,D}}. \tag{28}$$

In the limit $N \gg 1$, the probability of the mean of $2N$ random vectors distributed uniformly in a D-dimensional ball is well approximated by the normal distribution with zero mean and variance $\mathbb{1}_D/2N(D+2)$,

$$P_{D,2N}(\bar{x}) \approx \mathcal{N}_{0,\mathbb{1}_D/2N(D+2)}(\bar{x}). \tag{29}$$

In the following, we separately describe the grokking probability in the case $\lambda_1 = 0$ and the case $\lambda_1 > 0$.

Case $\lambda_1 = 0$ – Let us first consider the case without the L_1 regularisation, i.e. $\lambda_1 = 0$. The grokking probability is given by (see Appendix B)

$$\begin{aligned} P_{E(\infty)=0} &= \int \mathrm{d}\bar{x}\, P_{D,2N}(\bar{x})\Theta\left(\frac{w_1^\lambda(\bar{x})}{||w^\lambda(\bar{x})||_2} - \frac{1}{\varepsilon}\right) \\ &\approx \int_{-\varepsilon}^{\infty} \mathrm{d}\bar{x}_1\, \mathcal{N}_{0,1/2N(D+2)} \int_0^{2N(D+2)(\varepsilon^2-1)\left(\frac{\bar{x}_1+\varepsilon}{1+\varepsilon^2/\lambda_{2,D}}\right)^2} \mathrm{d}r\, \chi^2_{D-1}(r) \\ &= \int_{-\varepsilon}^{\infty} \mathrm{d}\bar{x}_1\, \mathcal{N}_{0,1/2N(D+2)}\, P\left(\frac{D-1}{2}, \frac{N(D+2)\left(\varepsilon^2-1\right)(\bar{x}_1+\varepsilon)^2}{\left(\frac{\varepsilon^2}{\lambda_{2,D}}+1\right)^2}\right), \end{aligned} \tag{30}$$

where $\chi^2_{D-1}(r)$ is the standard Chi-square distribution and $P(s,t)$ is the regularized gamma function. We also introduced the sample average $\bar{x} = \frac{1}{2N}\sum_{i=1}^{2N} y^i x^i$. The full grokking probability without the additional assumption has essentially the same structure with more complicated expressions for the parameters of the distributions (see Appendix B).

For a given set of parameters λ_2, D, and ε we can efficiently numerically evaluate the integral in Eq. 30 (and the full formula reported in Appendix B). In Fig. 8 we show the full grokking probability as a function of D, λ_2, N, and ε. As expected, the grokking probability is larger with increasing distance ε and number of samples N. We also observe that the grokking probability exponentially decreases with the dimensionality of the latent-space data distribution D. Therefore, latent-space distribution with a low effective dimension is preferred for better generalisation. This result partially explains the observation in Liu et al. (2022) which relates grokking to structure formation and effective dimension decrease at the transition. We expect that low effective dimension in the latent space increases generalisation in a more general setting, beyond the simple grokking scenario described in this section. In other words, models with latent space distributions with small effective dimensions will more likely lead to good generalisation. Finally, by increasing the L_2 regularisation strength λ_2 the grokking probability increases up to a maximum that depends on the remaining parameters. These results provide, some justification of the numerical observation in Power et al. (2022); Liu et al. (2022); Thilak et al. (2022) that weight decay increases the parameter region where grokking is observed.

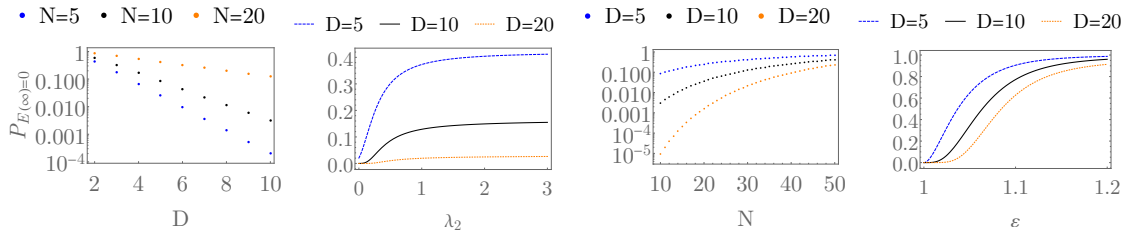


Figure 8: Grokking probability as a function of D, λ_2, N, and ε. We find that larger dimension D decreases the grokking probability. In contrast, larger regularisation strength λ_2 increases the grokking probability. As expected, increased class separation ε and number of training samples N also increases the grokking probability. If not specified in the panels, additional parameters are set to $D = 10$, $\lambda_2 = 0.1$, $N = 10$, and $\varepsilon = 1.01$.

Case $\lambda_1 > 0$ – Let us again consider the limit $N \gg \lambda_2 \gg 1$. If $\lambda_1 > 0$ the stationary solution w^λ_j depends on the sign of $|\bar{x}_j| - \lambda_1$, where $\bar{x} = \frac{1}{2N}\sum_{i=1}^{2N} x^i$. The $j-$th component of the stationary vector is

$$w^\lambda_1 = \begin{cases} 0 & , \ \lambda_1 \geq |\bar{x}_1 + \varepsilon| \\ \frac{\bar{x}_1 + \varepsilon - \lambda_1 \mathrm{sgn}(\bar{x}_1)}{\lambda_{2,D} + \varepsilon^2} & , \ \text{else} \end{cases}, \quad w^\lambda_{j\neq 1} = \begin{cases} 0 & , \ \lambda_1 \geq |\bar{x}_j| \\ \frac{\bar{x}_j - \lambda_1 \mathrm{sgn}(\bar{x}_j)}{\lambda_{2,D}} & , \ \text{else} \end{cases}. \tag{31}$$

The number of non-vanishing components of the stationary solution w^λ depends on the value of λ_1. Therefore, we get (in the $N \gg 1$ limit) an additional sum over the number of

non-zero elements in the w^λ,

$$P_{E(\infty)=0} = \int_{\lambda_1-\varepsilon}^{\infty} \mathrm{d}\bar{x}_1 \mathcal{N}_{0,1/2N(D+2)} \Bigg[(1-p_\lambda)^{D-1} + \sum_{k=1}^{D-1} \binom{D-1}{k} p_\lambda^k (1-p_\lambda)^{D-1-k} \int_0^{2N(D+2)(\varepsilon^2-1)\left(\frac{\bar{x}_1+\varepsilon-\lambda_1}{1+(\varepsilon/\lambda_{2,D})^2}\right)^2} \mathrm{d}r R_k(r) \Bigg], \tag{32}$$

where $p_\lambda = 1 - \mathrm{erf}\left(\sqrt{N(D+2)}\lambda_1\right)$ is the probability of the variable $|\bar{x}_i|$ (for $i > 1$) to be larger than λ_1 and $R_k(r)$ is the PDF of the sum of squares of k random variables sampled from the truncated normal distribution. Since the half-Gaussian distribution has a longer tail than the truncated Gaussian at λ_1, we can lower bound (or estimate) the grokking probability by using the Chi-squared distribution instead of $R(r)$. In the limit $\lambda_1 = 0$ we recover Eq. 30. In the case $\lambda_1 > 0$, we can approximate the inner integrals in Eq. 30 by the regularised gamma function and efficiently numerically evaluate Eq. 30.

Further, by discarding the sum over $k \geq 1$ in Eq. 32 we obtain a lower bound on the grokking probability

$$\begin{aligned} P_{E(\infty)=0} \geq & \int_{\lambda_1-\varepsilon}^{\infty} \mathrm{d}\bar{x}_1 \mathcal{N}_{0,1/2N(D+2)} (1-p_\lambda)^{D-1} \\ = & \frac{1}{2}\left(1 + \mathrm{erf}\left(\sqrt{N(D+2)}(\varepsilon-\lambda_1)\right)\right)\left(\mathrm{erf}\left(\sqrt{N(D+2)}\lambda_1\right)\right)^{D-1}. \end{aligned} \tag{33}$$

We find a similar distinction between the L_1 and L_2 regularisations as in the simple 1D case. At $\varepsilon = 1$ and $\lambda_1 = 0$ the grokking probability vanishes for any value of λ_2. In contrast, for $\lambda_1 > 0$ the grokking probability can increase even above 90% for any $D \geq 2$. Interestingly, the grokking probability increases with the dimensionality of the data distribution D. In fact, if we send $D \to \infty$ the grokking probability becomes 100% if $0 < \lambda_1 < \epsilon$. This result is a consequence of the concentration of measure of the uniform distribution. Similarly, by using the lower bound Eq. 33 we estimate the best value of λ_1 for any ϵ, D and N and find that the grokking probability maximum is always larger than 0.915. In contrast, in the $\lambda_1 = 0$ case, the grokking probability becomes exponentially small with D, independent of the remaining parameter values. We make similar observations also if we relax the condition $\lambda_2 \gg 1$ (see Appendix B).

The discussed results could be applicable more generally. It would be interesting to check if L_1 weight regularisation in the last (classification) layer significantly improves the generalisation of deep models compared to the L_2 regularisation. The works Power et al. (2022); Liu et al. (2022) do not study the differences between L_1 and L_2 regularisations. In Liu et al. (2022) a consistent observation has been made, namely larger weight decay leads in most cases to a larger parameter region where grokking is observed. We confirm this expectation on a simple model discussed in Section 4.3.1 and Section 4.3.2.

3.3.3 Grokking time

To calculate the grokking time, we first determine the condition for the zero train error. In contrast to the simple 1D case, this condition depends non-trivially on the training dataset and the initial condition $w(0)$. To simplify the calculation, we calculate the distribution of

the upper bound on the grokking time in the limit $N \gg 1$. We obtain the most conservative estimate for zero train error by selecting the training sample $\tilde{x}$ that forms the smallest angle with the shift vector ϵ. We write this condition in terms of the cosine of the angle as

$$\frac{w_1}{||w||_2} \geq \xi_{\text{train}} = \max_i \frac{\sqrt{||x^i+\varepsilon||_2^2 - (x_1^i+\varepsilon)^2}}{||x^i+\varepsilon||_2}, \tag{34}$$

where ξ_{train} denotes the cosine of the smallest angle between the plane defined by ϵ and any training sample $\tilde{x}^i$. We will consider only the zeroth-order solution in $1/\sqrt{N}$, where the grokking probability becomes 100%. Namely, we also discard terms proportional to $1/\sqrt{N}$. In this limit the stationary solution is proportional to ϵ, i.e. $w^\lambda \approx \frac{\epsilon}{\lambda_D+\varepsilon^2}$. The time-dependent model parameters are simplified to

$$w_1(t) \approx \frac{\varepsilon}{\lambda_D+\varepsilon^2} + \left(w_1(0) - \frac{\varepsilon}{\lambda_D+\varepsilon^2}\right) \mathrm{e}^{-(\lambda_{2,D}+\varepsilon^2)t}, \tag{35}$$

$$w_j(t) \approx w_j(0)\mathrm{e}^{-\lambda_{2,D}t}, \quad j>1. \tag{36}$$

Since we consider only the leading (zeroth) order in $\frac{1}{\sqrt{N}}$, the value of λ_1 does not have such a dramatic effect as in Section 3.3.2. Therefore, we will study only the case $\lambda_1 = 0$. To further simplify the calculation we will also assume $\lambda_{2,D} = \lambda_2 + \frac{1}{D+2} \ll 1$. In this limit we find

$$\xi_{\text{train}} \approx \max_i \sqrt{1 - \left(\frac{x_1^i+\varepsilon}{||x^i+\epsilon||_2}\right)^2} \approx \frac{x_{\max}}{\sqrt{x_{\max}^2+\varepsilon^2}} \approx \frac{1}{\sqrt{1+\varepsilon^2}}, \tag{37}$$

where $x_{\max} = \max_i ||x^i||_2$. Similarly, Eq. 35 and Eq. 36 simplify to

$$w_1(t) \approx \frac{1}{\varepsilon} + \left(w_1(0) - \frac{1}{\varepsilon}\right) \mathrm{e}^{-\varepsilon^2 t}, \tag{38}$$

$$w_j(t) \approx w_j(0)\mathrm{e}^{-\lambda_{2,D}t}, \quad j>1. \tag{39}$$

We find that the first component of w relaxes much faster than the remaining components. Therefore, the parameter path can be approximated by two straight lines/paths. Along the first path, $w_1(t)$ quickly relaxes towards the stationary value $w_1^\lambda \approx 1/\varepsilon$. Then, along the second path, the remaining parameters slowly relax towards the stationary value $w_{j>1}^\lambda \approx 0$. This leads to two different zero train/test error conditions.

First, we consider the case when grokking occurs during the fast relaxation (first path). In this case, the condition for grokking to occur reads

$$1 \geq \frac{1}{\varepsilon^2} + ||w^\perp(0)||_2^2, \tag{40}$$

where $w^\perp(t)$ is obtained from $w(t)$ by setting w_1 to zero. The zero train/test error is achieved after time

$$t = \frac{1}{\varepsilon^2} \ln\left(\frac{\frac{1}{\varepsilon} - w_1(0)}{\frac{1}{\varepsilon} - \frac{\xi}{\sqrt{1-\xi^2}}||w^\perp(0)||_2}\right). \tag{41}$$

We assume that $w_1(0) < w_1^\lambda \approx \frac{1}{\varepsilon}$ and obtain the final expression for the grokking time

$$t_{\text{G}} = \frac{1}{\varepsilon^2} \ln\left(\frac{\frac{1}{\varepsilon} - \frac{\xi_{\text{train}}}{\sqrt{1-\xi_{\text{train}}^2}}||w^\perp(0)||_2}{\frac{1}{\varepsilon} - \frac{\xi_{\text{test}}}{\sqrt{1-\xi_{\text{test}}^2}}||w^\perp(0)||_2}\right) \approx \frac{1}{\varepsilon^2} \ln\left(\frac{1 - ||w^\perp(0)||_2}{1 - \frac{\varepsilon}{\sqrt{\varepsilon^2-1}}||w^\perp(0)||_2}\right). \tag{42}$$

The grokking time in the considered limit depends only on the initial condition $w^{\perp}(0)$, i.e. on the initial distribution of the classifier weights. We assume that the initial model weights are sampled independently from a normal distribution with zero mean and unit variance. Setting $r = ||w^{\perp}(0)||_2^2$, the variable r follows the χ^2_{D-1} distribution. Therefore, we express the grokking-time PDF as

$$P_{\text{fast}}(t_{\text{G}}) \approx \chi^2_{D-1}(r(t_{\text{G}}))\frac{\partial r(t_{\text{G}})}{\partial t_{\text{G}}}, \tag{43}$$

where

$$r(t_{\text{G}}) = \frac{\left(\varepsilon^2-1\right)\left(e^{t_{\text{G}}\varepsilon^2}-1\right)^2\left(\varepsilon\left(\varepsilon e^{2t_{\text{G}}\varepsilon^2}+2\sqrt{\varepsilon^2-1}e^{t_{\text{G}}\varepsilon^2}+\varepsilon\right)-1\right)}{\left(\varepsilon^2\left(e^{2t_{\text{G}}\varepsilon^2}-1\right)+1\right)^2}. \tag{44}$$

The above result represents only one part of the grokking probability and hence the distribution $P_{\text{fast}}(t_{\text{G}})$ is not normalised. Integrating $P_{\text{fast}}(t)$ over the whole domain we obtain the probability to start with the initial condition where grokking occurs during the fast relaxation

$$p_{\text{fast}} = \int_w P(w)\Theta\left(1-\frac{1}{\varepsilon^2}-|w^{\perp}|^2\right)\mathrm{d}w = \int_0^{1-\frac{1}{\varepsilon^2}} \chi^2_{D-1}(r)\mathrm{d}r, \tag{45}$$

where χ^2_{D-1} is the standard Chi-squared distribution.

The second part of the grokking-time PDF comes from the initial conditions where the zero test/train error is obtained during the slow relaxation process. In this case, we assume that the value $w_1(t)$ is stationary, i.e. $w_1(t) \approx w_1^{\lambda} \approx \frac{1}{\varepsilon}$. The remaining model parameters evolve according to Eq. 39. The time at which the train/test error vanishes reads

$$t_{\text{train/test}} = \frac{1}{\lambda_{2,D}}\ln\left(\frac{||w^{\perp}(0)||_2}{\sqrt{1-\xi^2_{\text{train/test}}}w_1^{\lambda}}\right). \tag{46}$$

After simplification, we find the grokking time in the slow relaxation regime

$$t_{\text{G}} = \frac{1}{2\lambda_{2,D}}\ln\left(\frac{\varepsilon^4}{\varepsilon^4-1}\right). \tag{47}$$

Interestingly, the grokking time is independent of the initial condition. Therefore, the distribution of the slow-relaxation grokking time is trivial, i.e. proportional to a Dirac delta distribution with the weight $1-p_{\text{fast}}$, where p_{fast} is the probability of initialising the parameters with grokking during the fast relaxation given in Eq. 45.

By combining the grokking-time PDFs for the fast and the slow relaxation we obtain the grokking-time PDF in the limit $\lambda_2 \ll \varepsilon^2$. In Fig. 9 we show the grokking-time PDFs for several parameter sets in the considered limit. Increasing the input size D reduces the probability of fast-relaxation grokking times and increases the slow-relaxation grokking time. While the fast-relaxation grokking time does not depend on the regularisation strength λ_2, smaller regularisation leads to increased slow-relaxation grokking time. On the contrary, larger class separation decreases both fast- and slow-relaxation grokking times. We do not expect the analytically obtained grokking-time PDF to quantitatively describe real experiments, particularly because it is a zeroth-order large N solution. However, the bimodal structure and the qualitative parameter dependence should also be present in more realistic scenarios. We will discuss one such example in Section 4.3.

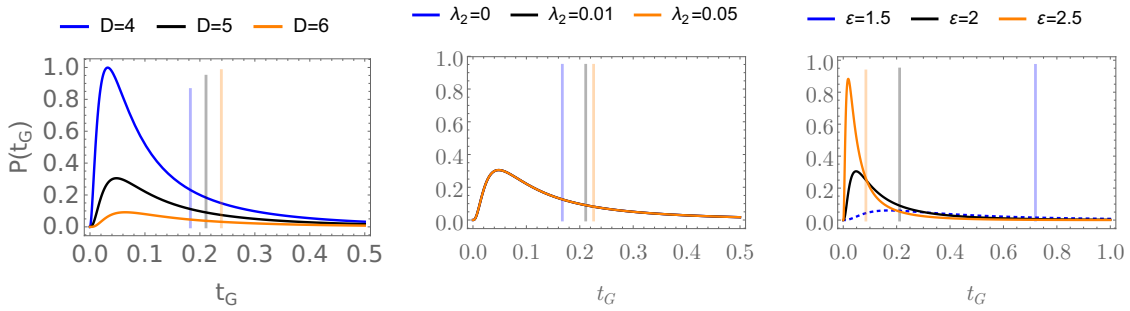


Figure 9: Grokking-time PDF for several values of D, ε, and λ_2. The short relaxation grokking-time PDFs are represented by full lines. The Dirac-delta long-relaxation grokking time is represented by vertical bars. The position of the bar is the position of the Dirac-delta function and the height of the bar represents the weight of the Dirac-delta part of the distribution. If not specified in the panels, additional parameters are set to $D = 5$, $\lambda_2 = 0.01$, and $\varepsilon = 2$.

3.3.4 Critical exponents for a general isotropic data PDF

By assuming isotropic probability densities on a compact domain $P^{\pm}$ in $\mathbb{R}^D$ we can relate the data PDF close to the domain boundary ($D-$dimensional sphere) with the critical exponent. For an isotropic data probability density we can write the test error close to the grokking transition as

$$E_{\text{test}}(\delta h) \approx \frac{1}{2} I_{2\delta h - \delta h^2}\left(\frac{D-1}{2}, \frac{1}{2}\right) \int_0^{\delta h} \mathrm{d}r \rho(r), \tag{48}$$

where, $I_z(a,b)$ is the regularized incomplete beta function, $\rho(r)$ is a probability distribution to find a sample with $||x||_2 = 1 - r$, and $\delta h \approx k(t_\epsilon - t)$, and coefficient k is determined by expanding Eq. 21 around t_ϵ. If the density $\rho(r)$ admits a Taylor expansion around zero, i.e. $\rho(r) = \rho_0 + \rho_1 r + \mathcal{O}(r^2)$, we find

$$\begin{aligned} E_{\text{test}}(\delta h) \approx & \frac{2^{\frac{D-1}{2}} \delta h^{\frac{D-1}{2}}}{(D-1)B\left(\frac{D-1}{2}, \frac{1}{2}\right)} \left(\rho_0 \delta h + \frac{1}{2}\rho_1 \delta h^2\right) \\ \propto & \rho_0 \delta h^{\frac{D+1}{2}} + \frac{1}{2}\rho_1 \delta h^{\frac{D+2}{2}} \\ \propto & (t - t_\epsilon)^{\frac{D+1}{2}}, \end{aligned} \tag{49}$$

where $B(a,b)$ is the Euler beta function. Obtained critical exponent $\nu = \frac{D+1}{2}$ is universal for isotropic probability densities that do not vanish at the ball boundary and is consistent with the result in Eq. 23. If in addition the density $\rho(\delta h)$ has an algebraic divergence, e.g. $\rho(\delta h) \approx \rho_\xi \delta h^{-\xi}$ where $0 < \xi < 1$ we get

$$E_{\text{test}}(t) \propto (t - t_\epsilon)^{\frac{D+1-2\xi}{2}}. \tag{50}$$

The critical exponent of the test error reveals the behaviour of the sample density at the boundary of the sample domain. While both the grokking probability and the grokking-time PDF depend on the details of the model's initial parameters and the evolution, the

critical exponent ν depends only on the data distribution at the boundary of the domain. Therefore, we expect that Eq. 50 describes the critical exponent quantitatively also in a more general setting. In this case, we might have to relax the condition $0 < \xi < 1$ to accommodate a more general divergence of the data distribution at the sample domain boundary.

4. Learning local rules with shallow tensor networks

In this section we go beyond the simple perception grokking model described previously in Section 3 and discuss grokking in the teacher-student model. In standard rule-learning theory, the teacher-student model describes a setting where the student model has to learn a rule given by the teacher model, see Engel and Van den Broeck (2001). In the simplest scenario where the teacher and the student models are perceptrons of the form Eq. 18, we use statistical mechanics methods to calculate the expected generalisation error for a given number of training samples (or training time). This is achieved in the thermodynamic limit where the input size M, and the number of training samples N go to infinity such that $N = \alpha M$. The teacher and student weights are sampled uniformly on an M-sphere. In this setup, one can use the replica trick Gardner and Derrida (1989) to calculate the test-error behaviour as a function of α. One finds $E_{\text{test}} \propto \frac{1}{\alpha}$ when $\alpha \to \infty$. Although sudden transitions to zero generalisation error are possible, they are a consequence of a restriction on the phase space of parameters, e.g. in the Ising perceptron, the parameters can take only values ± 1.

In summary, the standard rule-learning theory does not describe the grokking phenomenon and it is not clear how to reconcile the standard algebraic decay to zero test error with the grokking phase transition observed in deep models and presented in Section 3.

In this section, we fill this gap by introducing a local-rule learning scenario and a tensor-network map, allowing us to interpolate between the standard mean-field-like theory and the local, grokking setup. In particular, we introduce a local teacher and a tensor-network student setup which displays the grokking behaviour without any restriction on the values of the student model parameters. The tensor-network techniques will provide a correspondence between the standard teacher-student setup in the thermodynamic limit and the setup described in the Section 3. Therefore, the grokking phase transition is enabled by the locality of the learned rule. In contrast to the previous section, we will also study the feature learning part of the problem and establish a connection between the structure formation and grokking based on statistical averages of the test error and the effective dimension of the latent space representations.

4.1 Local teacher model

In the standard statistical-learning scenario, we determine the output of the teacher model (the rule) by Eq. 18 (see Engel and Van den Broeck (2001)). In this case, all values of the input contribute to the final result. In the thermodynamic limit this leads to a mean-field-like behaviour, i.e. the value of the input at any particular position has only infinitesimal influence on the result/rule.

We will study the opposite, local scenario $x \to y$, where $x, y \in \{-1, 1\}^M$. The i-th component of the output vector y_i will depend only on a K-neighborhood of the input at position i

$$y_i = \text{rule}(x_{i-K}, \dots, x_i, \dots x_{i+K}). \tag{51}$$

We call such model a $K-$local model. The Eq. 51 describes a well-known cellular automata computational paradigm. Cellular automata are universal discrete space-time dynamical systems with a finite set of possible states at each position Wolfram (1983); Wolfram et al. (2002). We define a cellular automaton by a set of rules which transform one configuration of states into another configuration. We will consider the rule-30 one-dimensional automaton ($K = 1$) Wolfram (1983); Wolfram et al. (2002), which exhibits chaotic behaviour and is defined by the rule $y_i = \text{rule30}(x_{i-1}, x_i, x_{i+1})$. The next state of the cell i, i.e. y_i, is determined by the current configuration at cells $i-1$, i, and $i+1$, i.e. x_{i-1}, x_i, x_{i+1}, as follows

$\mathbf{x_{i-1}, x_i, x_{i+1}}$	-1,-1,-1	-1,-1,1	-1,1,-1	-1,1,1	1,-1,-1	1,-1,1	1,1,-1	1,1,1
$\mathbf{y_i} = \text{rule30}(\mathbf{x_{i-1}, x_i, x_{i+1}})$	-1	1	1	1	1	-1	-1	-1

. (52)

We show an example time evolution of the rule-30 cellular automaton in Fig. 10. The initial condition is represented by the first line, black cells represent the value 1, and white cells represent the value -1. We will aim to learn one step of this evolution.

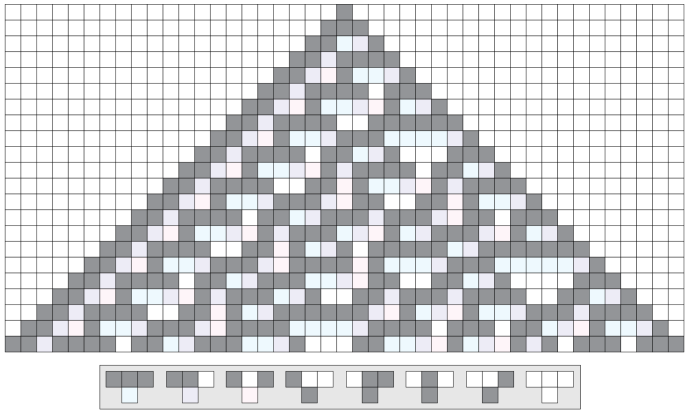


Figure 10: Diagramatic representation of the rule-30 cellular automaton. In our convention, the black cells represent 1 and the white cells represent -1. The diagram is taken from Wikipedia Wikipedia (2022).

The rule-30 automaton has already been discussed in the context of sequence-to-sequence prediction with tensor networks Guo et al. (2018); Efthymiou et al. (2019); Žunkovič (2022), however, no grokking phenomena have been reported. To study the effect of the neighbourhood size K, we shall consider a rule defined by K consecutive applications of rule 30. We will refer to such a rule as a K–local rule.

In summary, we modify the standard perceptron teacher-student setup by restricting the teacher model to local instead of global rules. The teacher will be modelled by a local map

transforming a sequence x into the sequence y. The task will be to approximate the chosen map by training on a finite set of input samples of length M. For a finite M we can choose open and closed boundary conditions. Open boundary conditions refer to the case when y_1 and y_M are calculated as if $x_0 = x_{M+1} = -1$. In the case of closed boundary conditions, we have $x_0 = x_M$ and $x_{M+1} = x_1$. The test set will include all possible input sizes from $M_{\text{test}} = 3, 4, \ldots, \infty$. We will determine the error as the ratio of incorrectly predicted values y_i.

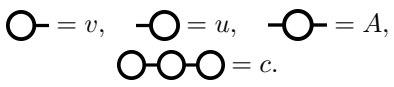
Besides the change from a global to a local rule, we will also modify the student model. Instead of the standard perceptron student model, we will use the uniform tensor-network attention model.

4.2 Uniform tensor-network student model

The simplest student model discussed in the literature is a perceptron model, which does not apply to our problem since we will discuss inputs of different sizes. The standard architectures applicable to variable-size inputs are the recurrent neural network (RNN) and convolutional neural network models. However, we found it convenient to use a tensor network approach, which enables us to construct a bridge between the teacher-student rule learning scenario and the grokking model discussed in the previous section.[1] Before introducing the tensor-network attention layer and the student model we will summarise the basic properties of tensor networks applied to machine learning Stoudenmire and Schwab (2016); Stoudenmire (2018).

4.2.1 Short introduction to tensor network methods

A tensor network is a tensor that is represented as a contraction of two or more tensors. The tensors that are contracted typically have a much smaller number of dimensions (indices) and hence fewer parameters. A trivial example of a tensor network is a scalar product of two vectors, where the second vector is a result of a matrix-vector multiplication, $c = u \cdot Av$



(53)

We introduced a diagrammatic notation, which makes longer tensor contractions more transparent. A tensor in this notation is represented as a circle with legs. The direction of the legs is typically not important. The number of legs determines the dimensionality of the tensor, e.g. a number has zero legs, a vector has one leg, a matrix has two legs etc. The most prominent tensor network, related to RNNs, is the matrix product state (MPS) obtained by contracting 3-dimensional tensors

$\psi =$  .

(54)

To use a tensor network as a machine learning model, we have to transform the inputs such that they can be contracted with the tensor network model to produce a scalar output. We

1. Due to a connection between RNNs and tensor networks Wu et al. (2022) we expect that one can rephrase our tensor network model in the language of RNNs.

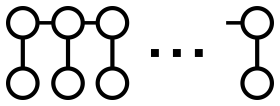
do that by using an embedding function and transforming the elements of the input vector with a vector transformation

$$\phi(x_j) = \tag{55}$$

The entire input vector is then transformed as

$$\Phi(x) = \phi(x_1) \otimes \phi(x_2) \otimes \ldots \otimes \phi(x_M) = \tag{56}$$ 

Formally, $\Phi(x)$ is an exponentially large vector with a compact MPS representation and will never be used directly. The output of the MPS model is then a contraction of the embedded input elements with the MPS tensor-network model

$$\psi \cdot \Phi(x) = \tag{57}$$ 

We can produce a vector output by adding one dimension to one of the MPS tensors. The presented setup has all the main parts of the typical tensor network model. It is differentiable with respect to tensor-network parameters and applicable to the standard training methods based on gradient descent.

4.2.2 Tensor-network attention model

In this section, we will introduce a simplified version of the tensor network proposed in Žunkovič (2022). As in the introductory example above, the entire model has two parts: an embedding layer and a tensor-network attention layer. Since the input is binary, we define the embedding layer with a local embedding function $\phi(x_i) : \{-1, 1\} \to \mathbb{R}^2$ as

$$\phi(-1) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \phi(1) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}. \tag{58}$$

After the embedding, we apply the tensor-network attention determined by two parameter tensors $A, B \in \mathbb{R}^{d\times d\times 2}$. We call the tensor A the attention tensor and the tensor B the classification tensor. The names of the tensors A and B reflect their role in the tensor-network attention layer. As we will describe below (see also Section 4.2.3), for a given position the tensor A determines the context of the input which is then linearly classified by the tensor B.

First, we construct matrices $\mathcal{A}(i)$ by contracting the attention tensor A with the local embedding vectors $\phi(x_i)$

$$\mathcal{A}_{\mu,\nu}(i) = \sum_{j=1}^{2} A_{\mu,\nu,j}\phi(x_i)_j. \tag{59}$$

Then, we use the matrices $\mathcal{A}(i)$ to construct the left and right context matrices $H^{\mathrm{L,R}}(i)$

$$H^{\mathrm{L}}(1) = \mathbb{1}_d, \qquad H^{\mathrm{L}}(i) = H^{\mathrm{L}}(i-1)\mathcal{A}(i-1), \tag{60}$$

$$H^{\mathrm{R}}(M) = G, \qquad H^{\mathrm{R}}(i) = \mathcal{A}(i+1)H^{\mathrm{R}}(i+1). \tag{61}$$

The matrix G determines the boundary conditions of the model. In the case of closed boundary conditions $G = \mathbb{1}_d$. In the case of open boundary conditions $G = v^{\mathrm{L}} \otimes v^{\mathrm{R}}$, where the boundary vectors $v^{\mathrm{L,R}} \in \mathbb{R}^d$ are additional model parameters. Alternatively, the boundary vectors $v^{\mathrm{L,R}}$ can be determined as left and right eigenvectors of the matrix A_0 corresponding to the largest eigenvalue. The final local weight vector $w(i)$ is then obtained by contracting the tensor B with the normalised left and right context matrices $H^{\mathrm{L,R}}_{\mathrm{N}} = H^{\mathrm{L,R}}/||H^{\mathrm{L,R}}||_2$,

$$w(i)_j = \mathrm{Tr}\left(H^{\mathrm{L}}_{\mathrm{N}}(i) B_j H^{\mathrm{R}}_{\mathrm{N}}(i)\right), \quad j = 1, 2, \quad i = 1, \ldots, M, \tag{62}$$

where B_j denotes the matrix with elements $[B_j]_{\mu,\nu} = B_{\mu,\nu,j}$. We calculate the attention layer output at position i as

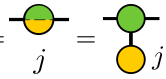
$$\hat{y}_i = w(i) \cdot \phi(x_i). \tag{63}$$

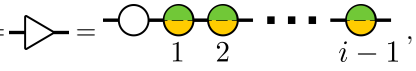
The final model output is then obtained by using the sign nonlinearity $f(x) = \mathrm{sgn}(\hat{y})$. The described tensor-network layer is a generalisation of the linear-dot attention mechanism (see Žunkovič (2022)). Therefore, we refer to it as a tensor-network attention.

It is instructive to present the tensor-network attention layer in a diagrammatic form by using the following definitions

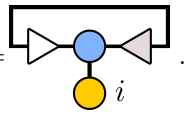
$\phi(x_i) =$  i, $A =$  , $B =$  , $\mathbb{I}_D =$  , $G =$  . (64)

We compactly write the entire transformation of an input at the position i as

$\mathcal{A}(j) =$  , (65)

$H^{\mathrm{L}}(i) =$  ,

$H^{\mathrm{R}}(i) =$  ,

$\hat{y}_i =$  .

4.2.3 TENSOR NETWORK MAP

The described tensor-network attention model also implements a map from inputs of variable length M to vectors of length $2d^2$. In the case of fixed attention tensors A, all possible infinitely many inputs define a PDF of vectors $z_i(x) \in \mathbb{R}^{2d^2}$, where

$$z_i(x) = H^{\mathrm{R}}_{\mathrm{N}}(i) H^{\mathrm{L}}_{\mathrm{N}}(i) \otimes \phi(x_i). \tag{66}$$

We show a schematic representation of the map in Fig. 11. By considering $z_i(x)$ as input features we can interpret the model defined by Eq. 63 as a perceptron defined by the weight tensor B, namely

$$\hat{y} = z_i(x) \cdot \vec{B}. \tag{67}$$

In the above formula $\vec{B}$ denotes the vectorised classification tensor B. By setting $D = 2d^2$, we have mapped the local-rule learning problem in the thermodynamic limit to a (grokking)

classification problem of the form discussed in Section 3. Interestingly, for a $K-$local rule, we can find 4^K-dimensional matrices A for which the transformed problem is solvable by a simple perceptron model and exhibits the grokking phenomena. Therefore, the $1/\alpha$ dependence on the training set size obtained from the standard rule-learning theory seems to be a consequence of the mean-field type infinite-range rule. For any local rule, we will observe grokking.

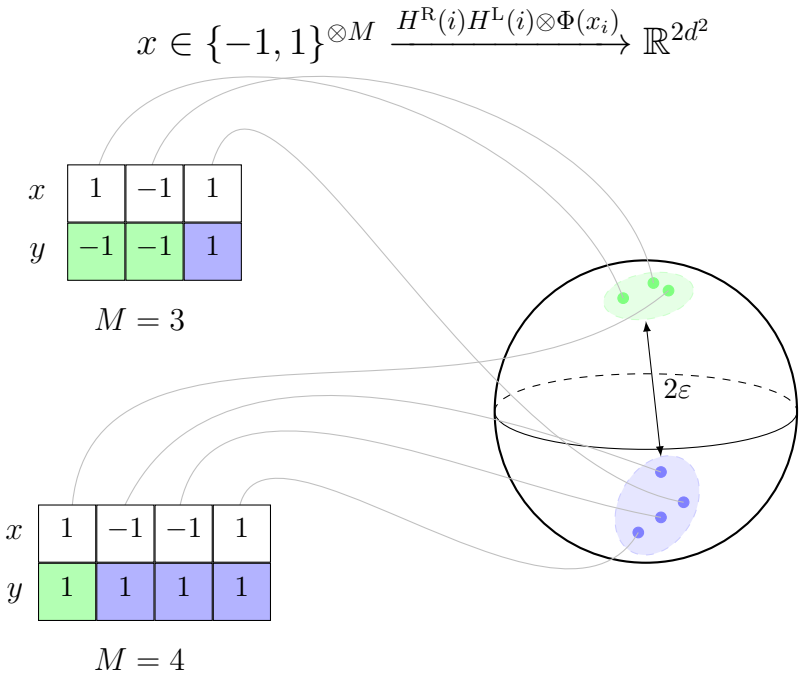


Figure 11: The tensor network map from $\{-1,1\}^M$ to $\mathbb{R}^{2d^2}$ implemented with Eq. 66, with 2ε denoting the distance between the closest positive and negative samples.

4.3 Simulation details and results

In this section, we present the results of training the uniform tensor-network student model on local algorithmic datasets. First, we consider the setting where the tensors A are fixed and map the inputs to separable distributions as shown in Fig. 11. We compare the numerical results with the predictions of the grokking model presented in Section 3. Second, we train the entire student model, i.e. the tensor A and the tensor B. In this case, we also discuss structure formation.

4.3.1 Constant attention tensors

We now discuss the simulation results obtained by fixing the attention tensors A. Namely, we use the proposed tensor-network model as a map from $\{-1,1\}^M$ to $\mathbb{R}^{2d^2}$ as discussed

in Section 4.2.3 and shown in Fig. 11. We choose the left and the right boundary vectors $v^{\mathrm{L,R}}$ to be the eigenvectors of A_0 corresponding to the largest eigenvalue. We also fix the bond dimension $d = 2$ and study the 1–local rule, which facilitates the comparison with the results discussed in Section 3.

We determine the attention tensors A by independently sampling each element according to the normal distribution with zero mean and unit variance. Since not all attention vectors lead to solvable problems, we perform rejection sampling by checking if the final model parameters given by Eq. 21 have zero test error. Once we obtain a solvable instance of the attention tensor A, we do not change its parameters during training. The rejection sampling procedure works well for $D \leq 3$. In higher dimensions, the probability of sampling a correct tensor A is very small, therefore, we evaluate the grokking probability and grokking time distributions only for small $d = 2$.

An example evolution of training and test errors for bond dimension $d = 6$ is shown in Fig. 12. The context tensors A have been trained in advance to ensure that the model admits zero final test error. Grokking is not as pronounced since we are training only a linear classifier. Nevertheless, we observe in the left panel of Fig. 12 a plateau of test error which eventually goes to zero, while the test error remains constant, or in some cases even slightly increases, the training error vanishes. The test error is much smaller than in standard cases of grokking due to a significant overlap between the train and test distributions and in the latent space. The middle and right panels are shown for completeness to demonstrate a representative range of training behaviour which includes cases without a test error plateau. The cases shown in Fig. 12 differ only in the initial condition of the linear classifier.

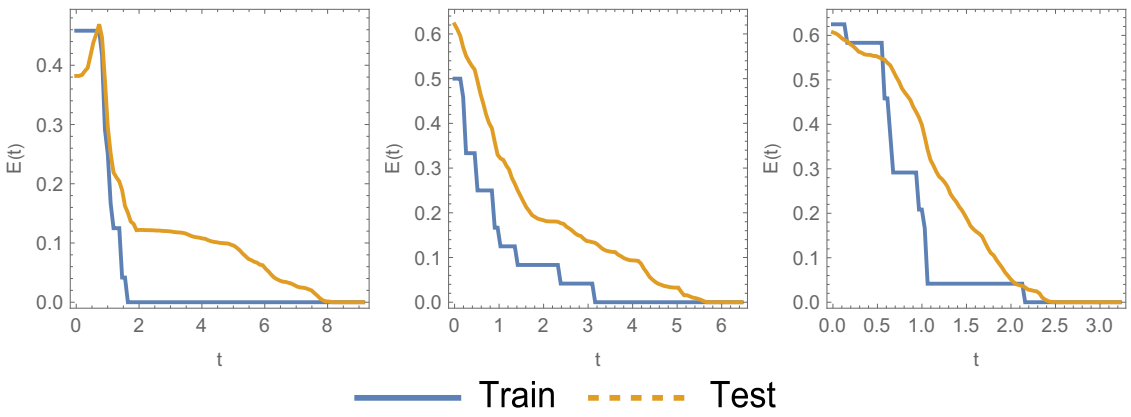


Figure 12: Examples of test and train error evolution for a fixed attention tensor with $d = 6$. We show three representative situations. In the left panel, the test error has a pleau while the training error is zero. Finally, the test error drops to zero. The plateau is significantly smaller compared to the standard cases of grokking due to the simplicity of the problem. The middle and the right panels do not show a plateau in the test error and are more representative of the standard training dynamics. The training parameters are $\lambda_1 = 0$, $\lambda_2 = 0.01$.

Exact 1–local attention We first consider the learning dynamics in the case of exact attention tensors

$$A_0 = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad A_1 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix}. \tag{68}$$

The minimal bond dimension of the exact solution can be reduced to 2 if we generalise the model and train different left and right attention tensors A. In the exact, 1-local attention case, the vectors $z_i(x)$ contain only information about the state of the neighbouring positions. Since the smallest size $M = 3$ contains all eight possible inputs, the larger training set size M does not change the results. After averaging over many initialisations of the classifier part of the network B we obtain the average test error shown in Fig. 13. We observe a first-order transition with a jump of 1/4 in the test/train error. Here the factor 1/4 comes from the fact that the neighbourhood of any given position has four possible different values.

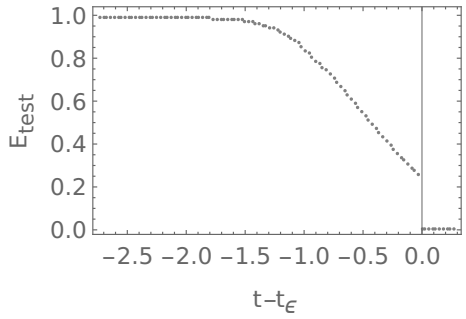


Figure 13: The first-order phase transition when learning with the exact attention tensors A given in Eq. 68. The jump in the transition is 1/4.

In the following, we discuss results obtained by randomly sampling the attention tensors A. We show results for three different attention tensors A, namely `Example 1`, `Example 2`, and `Example 3` reported in Appendix C.

Grokking probability We estimate the grokking probability as the fraction of the sampled attention tensors A that leads to linearly separable latent space data for the studied rule. In contrast to the grokking probabilities discussed in Section 3, we fix the training set to contain all possible samples of length $M = 3$. In Fig. 14 we show the dependence of the grokking probability on regularisation strengths $\lambda_{1,2}$. We observe that the L_2 regularisation decreases the grokking probability while the L_1 regularisation first slightly increases the grokking probability and then decreases compared to models without regularisation. In all cases, the L_1 regularised model has a larger grokking probability than the L_2 regularised model with the same regularisation strength. A larger grokking probability for L_1 regularised models is another indicator that L_1 regularisation could lead to better generalisation compared to the L_2 regularisation.

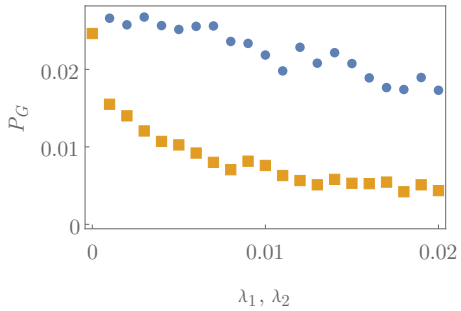


Figure 14: Grokking probability (P_{G}), representing the fraction of the attention tensors A that map the 1–local rule to linearly separable data. We show the dependence of P_{G} on the L_1 regularisation strength (λ_1 blue circles) and the L_2 regularisation strength (λ_2 orange squares). The L_1 regularised models have larger grokking probability compared to the L_2 regularised models. We used 20k random attention tensors to estimate the grokking probability.

Critical exponent ν Sampled attention vectors $H^{\mathrm{L,R}}(i)$ also contain information about the input beyond only the neighbouring sites. Moreover, information about the neighbours is not complete. Therefore, we observe a second-order transition, as discussed in Section 3. In Fig. 15 we show the average test error for three different but fixed attention tensors obtained by rejection sampling. The exact values of the attention vectors are reported in Appendix C. We find that the critical exponent ν does not depend on the regularisation strengths $\lambda_{1,2}$ and is in all cases smaller than one, which is in agreement with the predictions of the simple grokking model discussed in Section 3.

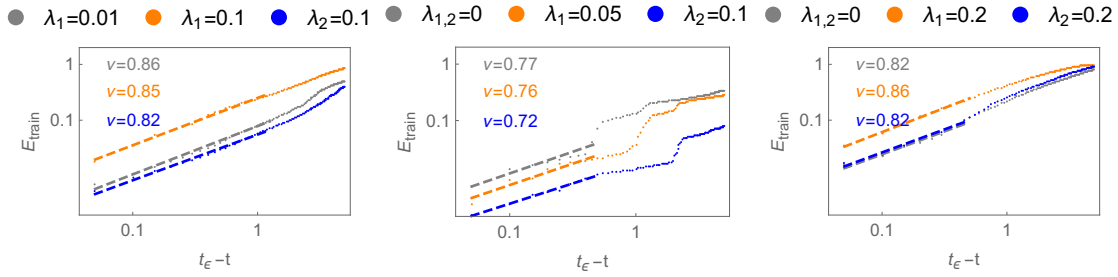


Figure 15: Average test error during training with fixed attention vectors (log-log plot). From left to right we report results for `Example 1`, `Example 2`, and `Example 3` attention tensors A given in Appendix C. The fitted critical exponents ν are shown in the plots and only mildly depend on $\lambda_{1,2}$.

Besides the test-error critical exponent, we estimate several properties of the feature distributions. In particular, we calculate the effective dimension D_{eff}, the divergence exponent ξ of the sample PDF at the boundary of the domain, and the distance between positive and negative samples ε. These quantities are calculated from the training-dataset features $z_i(x)$. To calculate the effective dimension D_{eff} we first find σ_k defined as the fraction

of the variance explained by the kth principal component of the training dataset features $z_i(x)$. Then we calculate the entropy S of the ratios σ_k defined as $S = -\sum_k \sigma_k \log \sigma_k$. Finally, the effective dimension is obtained as the exponent of the entropy, i.e. $D_{\text{eff}} = \mathrm{e}^S$. We report the effective dimensions for the considered `Examples 1-3` in Table 1.

We also use the vectors $z_i(x)$ to estimate the divergence exponent of the sample PDF at the boundary of the domain. In the considered case, the vectors $z_i(x)$ are a tensor product of three vectors. Therefore, we estimate the divergence in the PDF by focusing separately on each of the components of the vector $z_i(x)$. One of the vectors is a constant vector determined by the embedding function and does not contribute to the divergence exponent. The remaining, important parts are the left and the right context vectors, namely $v^{\mathrm{L}} H^{\mathrm{L}}_N(i)$ and $H^{\mathrm{R}}_N(i) v^{\mathrm{R}}$. We consider the normalised context vectors which, in addition, have size two (since we fix $d = 2$). Therefore, they are uniquely determined by the angle with the first component and we can accordingly estimate the divergence at the boundary of the domain by studying the PDF of the angle. We estimate the divergence exponent by looking at the behaviour of the estimated PDF at the boundary with an increasing number of bins. The final exponent is obtained as a sum of the exponents obtained from the left and the right-attention part of the feature vector $z_i(x)$. We find (see Fig. 16) that the PDF diverges algebraically with the powers reported in Table 1.

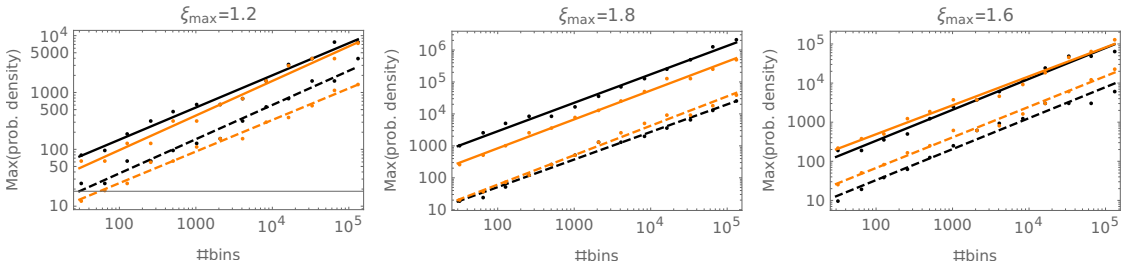


Figure 16: The estimated PDF maximum of the positive (black) and negative (orange) samples. The dashed line fits correspond to the left context vectors $v^{\mathrm{L}} H^{\mathrm{L}}_{\mathrm{N}}(i)$ and the full line fits the right context vectors $H^{\mathrm{R}}_{\mathrm{N}}(i) v^{\mathrm{R}}$. The final value reported in the panel title is obtained as a maximum sum of the left and right divergence exponents. From left to right we report results for `Example 1`, `Example 2`, and `Example 3` attention tensors A given in Appendix C.

In Section 3.3.4 we derived a relation between the exponents ν, ξ and the effective dimension for a simple $D-$dimensional ball model. Interestingly, we find that the relation given by Eq. 50 obtained from a simple spherically symmetric model is reasonably close in two out of the three considered cases (see Table 1).

We also estimate the class separation from the actual latent space distribution and report it in the units of the intra-class variance (see Table 1).

Grokking time Finally, we estimate grokking-time PDF, see Fig. 17. We do not expect that the prediction of Section 3.3.3 will quantitatively describe the estimated grokking-time PDF. Besides the "worst-case" initial condition assumption, the condition $N \gg 1$ is

	D_{eff}	ν	ξ^*	$\xi = \frac{1}{2}(D_{\text{eff}} - 2\nu + 1)$ (Eq. 50)	ϵ
`Example 1`	3.0	0.85	1.2	1.15	1.45
`Example 2`	3.8	0.75	1.8	1.65	1.46
`Example 3`	3.0	0.84	1.6	1.16	1.6

Table 1: Critical exponent ν and numerically calculated characteristic parameters of the feature vector $z_i(x)$ distribution. We also compare the numerically estimated divergence of the sample PDF at the boundary ξ^* with the prediction of the spherical model (Eq. 50).

not valid. Hence, the actual value of G is far from identity. However, some qualitative predictions of the D-dimensional ball model can still be observed. First, we notice, the bimodal structure of the estimated grokking-time PDF. In the spherical model, the two peaks are a consequence of the separation between the slow and fast modes, where the dynamics of the slow modes were essentially determined by the regularisation strength λ_2. Similarly, in all three considered cases (i.e. `Example 1-3`) we can separate the eigenvalues of G by size in two sets. In one set, the eigenvalues are larger by one order of magnitude than in the other. However, increasing the regularisation strength λ_2 often leads to increased grokking time, which is not the case in the simple uniform ball model. The discrepancy is a consequence of the non-diagonal matrix G, which mixes different components of the vector $z_i(x)$. We also observe that a larger effective dimension D_{eff} (see Table 1) leads to longer grokking times and a larger class separation ε to smaller grokking times. The last two observations are in agreement with the $D-$dimensional ball model discussed in Section 3.3.3.

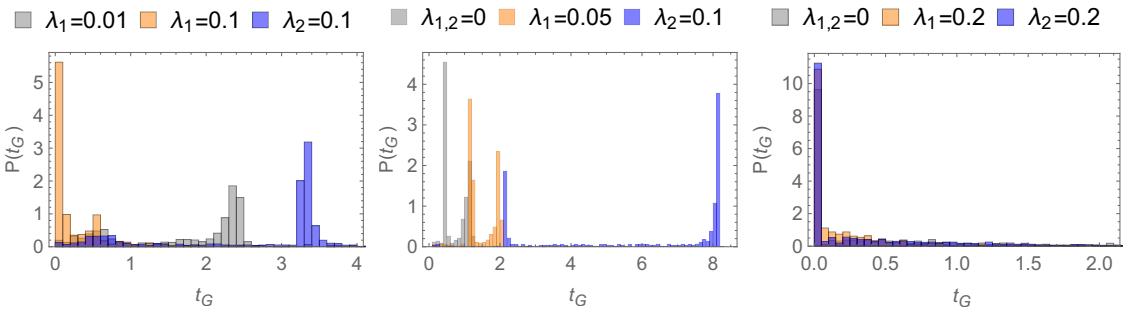


Figure 17: We show the estimated grokking-time PDFs for three fixed attention tensors: `Example 1` (left), `Example 2` (middle), and `Example 3` (right). In most cases, the grokking-time PDF is bimodal, which is in agreement with the prediction of the simple grokking model discussed in Section 3.3.3.

The presented results are obtained by averaging over many initialisations of the classification tensor B. We sample the initial elements of B from a normal distribution with zero mean and unit variance. Changing the initial distribution can impact the results. Determining the effect of the initial distribution of B on the grokking-time PDF and the critical exponent is left for future research.

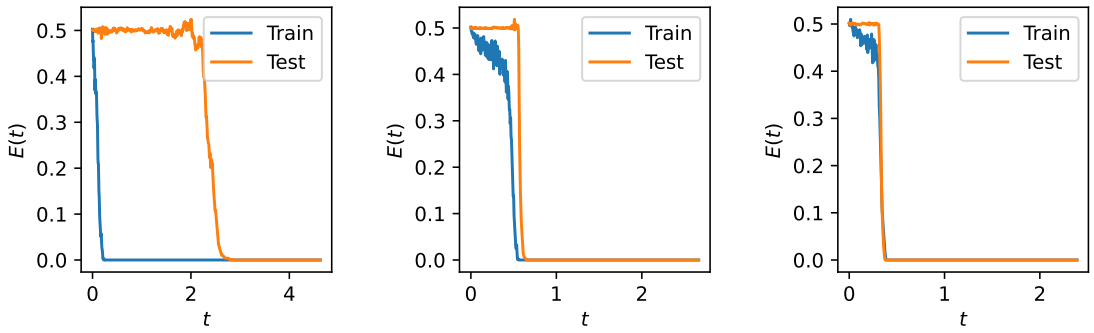


Figure 18: We show typical examples of train and test error dynamics while learning the rule-30 cellular automaton on all inputs of size 8. We show three training runs with different initial conditions. We set $d = 10$, $lr = 0.005$, and $\lambda_{1,2} = 0$. The three panels correspond to three different initial conditions.

4.3.2 Full model training and structure formation

In this section, we discuss grokking and structure formation properties of the complete student model introduced in Section 4.2. We initialise the model with a random initial condition, where all the entries of the tensors A, B are uncorrelated and sampled according to a normal distribution with zero mean and unit variance. We train the model with the Adam optimiser (with standard parameter setting) and a learning rate of 0.005. We use the same loss as in the previous sections Eq. 4 with $L_{1,2}$ regularisation strength $\lambda_{1,2} \in [0, 0.001]$. The regularisation strength is the same for the attention tensor A and the classifier tensor B. We add a sigmoid non-linearity before the final sign non-linearity to improve the training stability and reduce the training time. We consider only open boundary conditions and set $x_{-1} = x_{M+1} = -1$. Finally, in the main text, we consider only the 1-local rule. We discuss the 2– and 3–local rules in Appendix D. We perform tests in three situations, namely, without regularisation ($\lambda_{1,2} = 0$), with L_2 regularisation ($\lambda_1 = 0$, $\lambda_2 = 0.0001$), and with L_1 regularisation ($\lambda_1 = 0$, $\lambda_2 = 0.001$). We chose the regularisation strengths $\lambda_{1,2}$ to be the largest regularisation strengths with only a few spikes in the training loss after the grokking time. To obtain zero test error it is sufficient (in almost all cases) to train only the attention parameters A and fix the classification parameters B. This is a consequence of the gauge symmetry of the tensor attention layer Žunkovič (2022). However, we will always train all model parameters. Since the full tensor-attention model is non-linear, we do not expect the theory developed in Section 3 to be valid. On the other hand, we do observe phenomena related to neural collapse Papyan et al. (2020) and structure formation Liu et al. (2022).

In Fig. 18 we show a typical example of the train and test error dynamics. The grokking phenomenon is, in this case, a typical behaviour, in contrast to our effective distributions discussed in the previous section Section 3. The pronounced grokking, i.e. a very short transition of the test error from zero to one, is a consequence of a concentration of the latent space distribution close to the boundary of the domain and structure formation at the transition to zero test error as detailed in the following subsection.

Average test error and average effective dimension First, we investigate the dynamics of the average test error and calculate the critical exponent ν. In Fig. 19 we show that the critical exponent decreases upon increasing regularisation. Larger regularisation leads to a sharper transition to zero test error, in contrast with the linear case studied in the Section 3.3.2 and in the Section 4.3.1, where the critical exponent was found to be independent of the regularisation strengths $\lambda_{1,2}$. The test error drops to zero at the grokking transition. Following the grokking transition, the test error is non-zero and experiences fluctuations. These fluctuations can be detected as sharp increases in training loss and are more common in models with large regularisation. Therefore, the $L_{1,2}$ regularised models have larger average test error after the grokking transition.

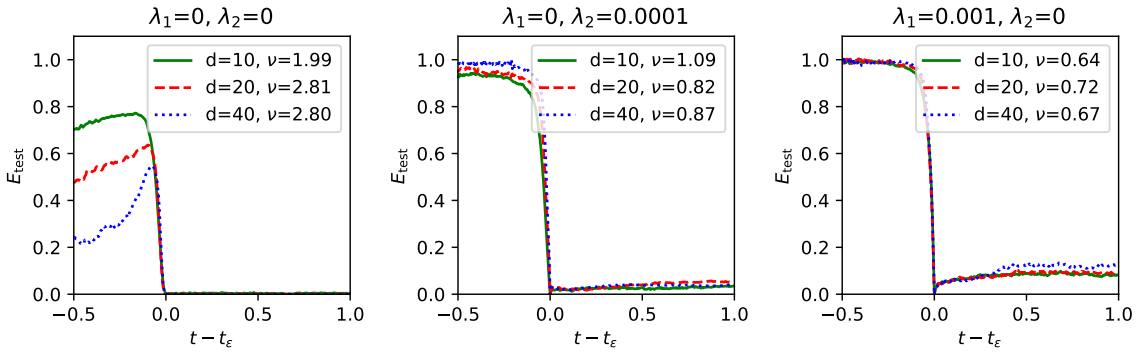


Figure 19: The average test error at the phase transition. We align the first point where the test error becomes zero (i.e. the time t_ϵ) and take the average over many (~ 1000) initialisations of the model parameters. Training without regularisation results in larger critical exponent $\nu \approx 2.5$ as training with L_1 ($\nu \approx 0.7$) or L_2 ($\nu \approx 0.9$) regularisation. In all experiments, we used a learning rate of 0.005. In the legends, we report the bond dimension of the trained models d and the corresponding fitted critical exponent ν.

The shape of the average test error close to transition point t_ϵ (or the critical exponent ν) depends only slightly on the model size (bond dimension d). This suggests that the effective dimension of the mapped data D_{eff} is independent of the model size. We confirm this by calculating the effective dimension of the features $z(i)$. Since we study only open boundary conditions, we consider only the effective dimension of the left context vectors $v^{\text{L}} H^{\text{L}}_{\text{N}}(i)$. The right context vectors $H^{\text{R}}_{\text{N}}(i) v^{\text{R}}$ have the same properties because of the model symmetry. As shown in Fig. 20, the average effective dimension drops significantly just before the grokking transition. We observe that regularisation significantly decreases the effective dimension of the mapped vectors $v^{\text{L}} H^{\text{L}}_{\text{N}}(i)$. The effective dimension is smallest with L_1 regularisation, which is expected since the L_1 regularisation enforces sparsity while the L_2 regularisation enforces smoothness.

Structure formation and grokking We relate the decrease of the effective dimension to structure formation. As can be seen in examples shown in Fig. 21 and Fig. 22, a small effective dimension signals an emergent latent space structure which, however, can be different in each example. Similarly, in Liu et al. (2022) the authors argue that the grokking

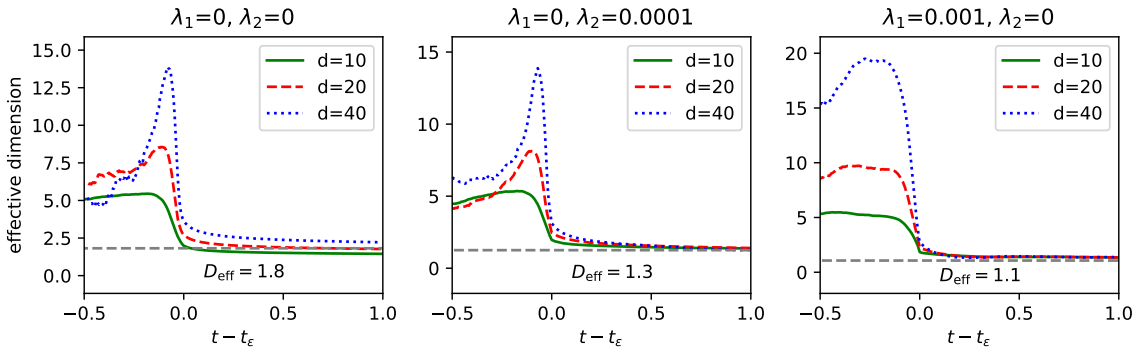


Figure 20: The average effective dimension corresponding to the error in Fig. 19. Larger regularisation results in a smaller effective dimension D_{eff} after the grokking transition. The horizontal dashed line corresponds to the average of the minimal effective dimension of data with zero test error (over all samples with fixed d).

in deep models is related to structure formation. Our findings differ from those of Liu et al. (2022) in that we discuss ensemble/average phenomena. The authors of Liu et al. (2022) discuss the connection between grokking and structure formation on the single-model level. In contrast, we argue that grokking and structure formation are related on average as shown in Fig. 19 and Fig. 20, and not for every training run individually. We call the structure formation and grokking for a single training of a model the model-wise structure formation and the model-wise grokking. We disentangle model-wise structure formation from model-wise grokking by observing specific training samples. We typically observe the appearance of a simple structure in data in the proximity of the grokking transition. This is consistent with a sharp drop of the average effective dimension at the transition (shown in Fig. 20). Additionally, we observe that latent space structures can be different for different model parameters and initialisations.

In Fig. 21 and Fig. 22 we show the structures appearing in the features ${}^{\text{L}}v H^{\text{L}}_{\text{n}}(i)$ with bond dimension $d = 3$. In Fig. 21 we see that the structure of the latent space data changes also during a single run. This can be detected as a spike in the training loss or as a step-like jump in the effective dimension (see Fig. 21). The structures can change from lower to higher dimensional and vice versa, e.f. see Fig. 21 – the transition between $t = 1$ and $t = 1.2$ increases the effective dimension $D_{\text{eff}.}$ of the mapped data. The appearance of geometric structures in the latent space does not necessarily lead to good generalisation, i.e. small test error (see Fig. 22 at time $t = 0.57$). Finally, we also show in Fig. 22 that we can have a small generalisation/test error with complex or not apparent latent space structures (see Fig. 22 at time $t = 1.19$). These empirical observations suggest that grokking and structure formation are not related model-wise. That structure formation and grokking are in general two distinct phenomena is further corroborated by our simple grokking model discussed in Section 3, which does not require any special geometric structure (aside from the condition of linear separability).

Grokking time Finally, we estimate the PDF of the grokking times, see Fig. 23 (top panels). Taking the non-regularised case as the baseline, we find that L_1 regularisation

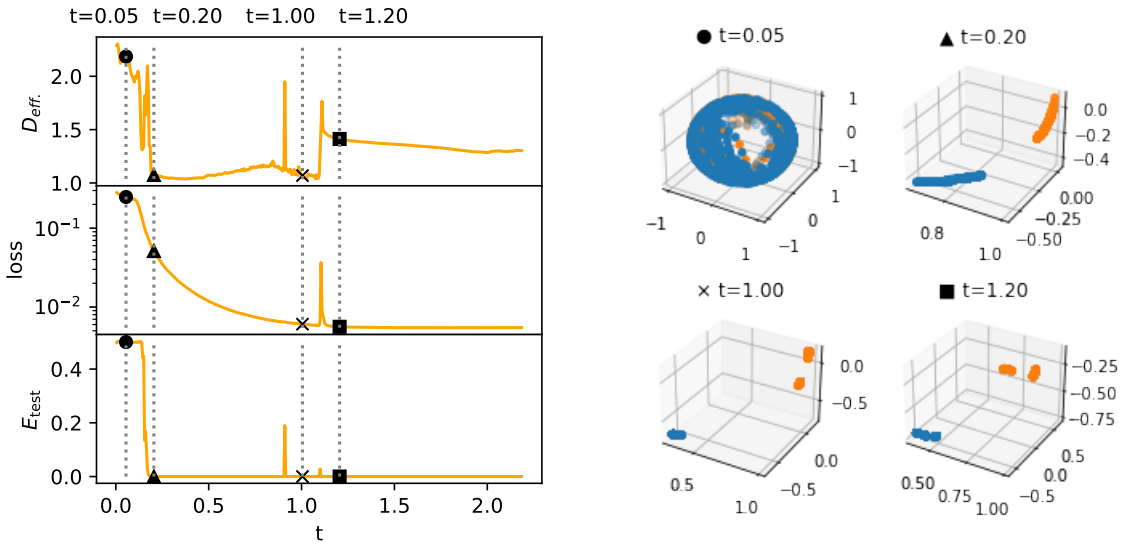


Figure 21: Several emergent structures in the latent space (Example 1). The left plots show the effective dimension D_{eff} (top), train loss (middle), and test error (bottom). The black markers show the value of the plotted quantities at specific times marked by vertical dotted lines and written on the top of the left panels. The right panels show the structure of the features at the marked times. We observe that an essentially one-dimensional feature distribution with two distinct islands of features splits into an almost 2D feature distribution with three isolated islands. We can detect this transition as a sharp peak in the loss and a step in the effective dimension D_{eff}.

decreases the average grokking time $\overline{t_{\text{G}}}$ significantly more than L_2 regularisation. Further, grokking times for L_2 regularised models increase with the model size. On the other hand, non-regularised and L_1 regularised models have roughly a model-size-independent grokking-time PDF, and hence the grokking-time average.

Since the grokking time is measured relative to the time at which the zero train error is achieved, we estimate also the PDF of times t_ϵ (zero-test-error time). We find that both L_1 and L_2 reduce the time at which the zero test error is attained. Therefore, both, the L_1 and the L_2 regularisation decrease the number of steps required for good generalisation. In addition, the L_1 generalisation seems to be more efficient, in the sense, that there is a shorter time interval with a large difference between training and test error.

5. Summary and discussion

We analyse grokking from two perspectives. First, we propose a simple grokking setup (*perceptron grokking*) and consider two solvable grokking models. Second, we introduce

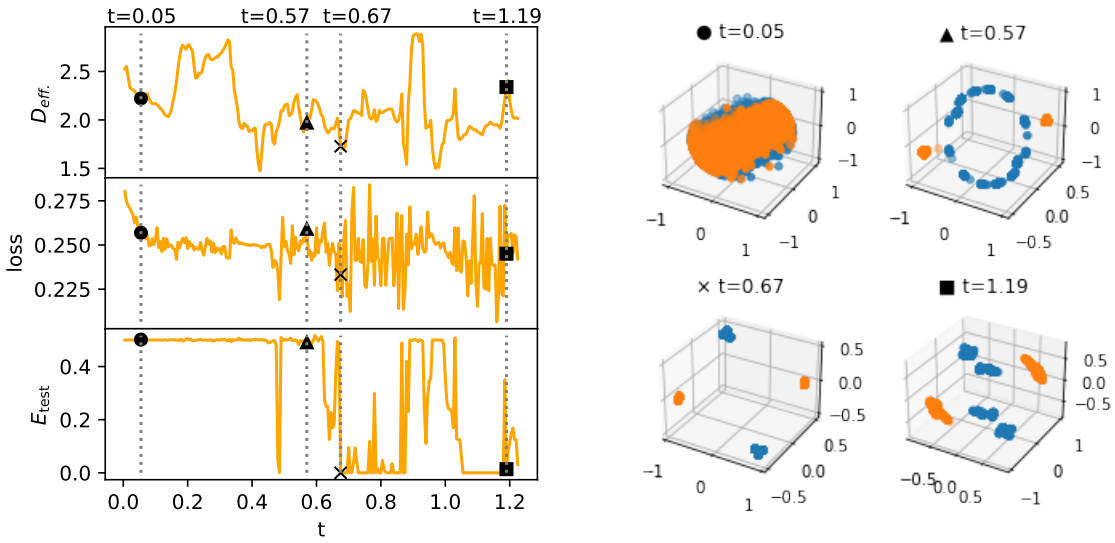


Figure 22: Several emergent structures in the latent space (Example 2). The left plots show the effective dimension D_{eff} (top), train loss (middle), and test error (bottom). Black markers show the value of the plotted quantities at times marked by vertical dotted lines and written on the top of the first plot. The right panels show the structure of the features at the marked times. We observe that structured data also appears in the case of high test error ($t = 0.57$). The zero-test-error structure ($t = 0.67$) is different compared to the example in Fig. 21. Here, we see a 2D structure with four isolated feature islands. Finally, at time $t = 1.19$ the structure starts to disappear while the test error is still considerably small (smaller than 1%).

a tensor-network attention map and connect the standard statistical mechanics of the teacher-student setup with the perceptron grokking setup.

Perceptron grokking By studying two solvable grokking models, we show that grokking is a phase transition and calculate the critical exponent, grokking probability, and grokking-time PDF. The obtained analytic expressions allow us to determine the effect of model and training parameters on the grokking probability and the grokking-time PDF. In particular, we find a stark difference between the L_1- and L_2-regularised models. The L_1-regularised models have a higher grokking probability and a shorter grokking time than the L_2-regularised models. We also obtain a universal expression for the test-error critical exponent of spherically symmetric models, which is relevant in the transfer learning setting, where only the last layer is retrained.

Learning local rules with shallow tensor networks We use the tensor-network attention model with fixed attention tensors A to test the predictions of the perceptron grokking setup on a 1D cellular-automaton rule-30 learning task. Our prediction of the critical exponent roughly agrees with the numerical estimation, thereby validating the grokking

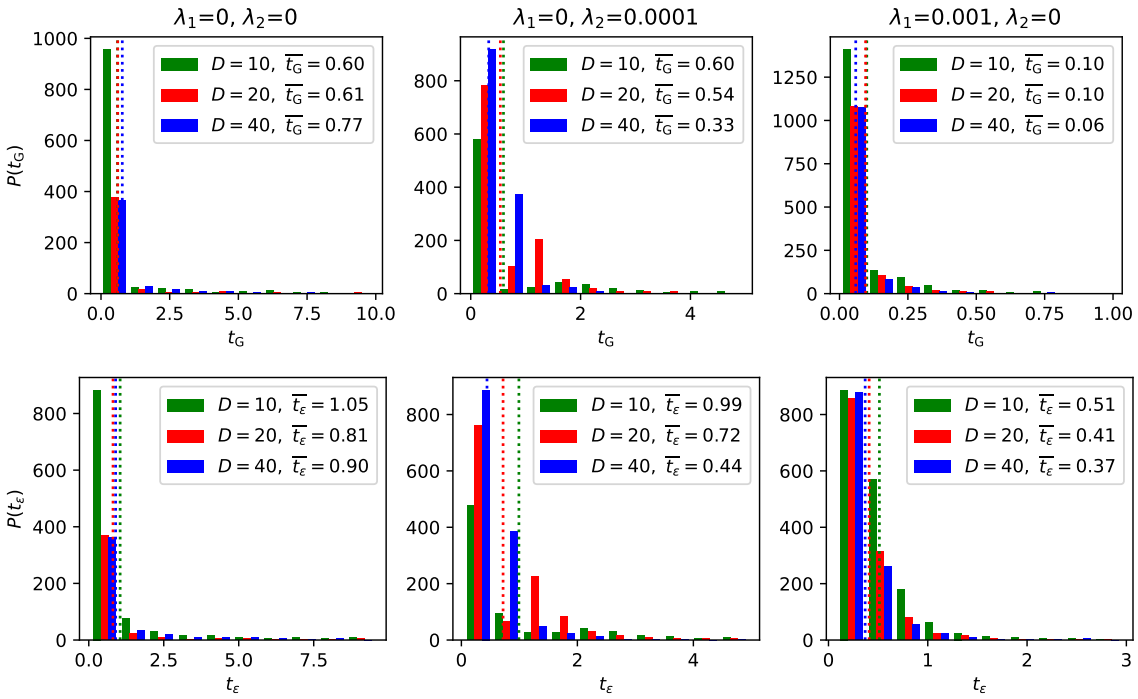


Figure 23: The estimated grokking-time PDF and t_ϵ PDF. The colours correspond to different model sizes, namely $d = 10$ (blue), $d = 20$ (orange), and $d = 40$ (green). The vertical lines correspond to the averages reported in the legends of the panels. We find that L_1 regularisation reduces t_G and t_ϵ. By contrast, the L_2 regularisation decreases only the absolute time t_ϵ. In the L_2 case and $d = 20$ we also find a clear bimodal grokking time distribution. In all cases, we set the learning rate to 0.005.

scenario on a simple problem. On the other hand, the grokking-time PDF approximation, which invokes strong assumptions, predicts the actual numerical estimate only qualitatively.

We also perform the training of the full tensor-network student model. Similarly to solvable perceptron grokking models, we observe a difference between the L_1-regularised and the L_2-regularised models. The former has a shorter grokking time and a lower effective dimension, which agrees with the analytic predictions of the perceptron grokking models. Therefore, we expect that L_1 regularisation leads to improved generalisation properties (e.g. smaller test error) compared to L_2 regularisation also in a more general classification setting.

In the case of training the full tensor-network student model, we discuss the connection between grokking and structure formation. We determine the grokking transition by observing the average test error. We show that the average effective dimension of the feature-space data sharply decreases at the grokking transition. By observing specific models, we also find that small effective dimensions correspond to particular feature-space structures. We accordingly relate grokking with structure formation on the ensemble level. By contrast, we find several models with zero test error without apparent feature-space structures and

vice versa. This shows that in specific (though rare) cases, the test error drops to zero even if no structure is present in the data. Similarly, simple structures can appear during training also when the test error does not vanish. We thus separate the grokking and the structure formation on the level of individual training runs.

As a distinct feature of the full tensor-network training, we highlight the spikes in the training loss. We observe that spikes become more frequent with larger $L_{1,2}$ regularisation. We also relate the training-loss spikes to the changes in the feature-space structures, which may become less or more complex during training. We can assess the shape of the feature-space structures by observing the effective dimension, which shows a step-like behaviour whenever we observe a training-loss spike. Typically less complex structures correspond to a smaller effective dimension. These findings can be relevant for deep neural network training where training-loss spikes are also observed. Frequent training-loss spikes can be avoided by using a smaller regularisation. Moreover, we can determine whether the model parameters should be reverted by monitoring the feature-space effective dimension. For example, we can revert the model only if the training-loss spikes correspond to an increased effective dimension.

Finally, the proposed tensor network map connects the grokking phenomena, which have so far been observed only in deep models, with the standard teacher-student learning setup. The considered local tensor-network rule learning setup is an extreme example of a learning rule. The standard teacher-student mean-field setup is the opposite extreme. It would be interesting to study if the proposed grokking setup and the tensor-network map can be extended to study algebraically decaying rules which interpolate between the two extremes. Extending the presented theory to deep neural networks appears to be difficult within the proposed framework.

Acknowledgement

The authors received support from the ARIS project J1-2480 From Classical To Quantum Machine Learning With Tensor Networks. BZ was also supported by the ARIS research program P2-0209 Artificial Intelligence and Intelligent Systems (2022 – 2024). Computational resources were provided by SLING – Slovenian National Supercomputing Network. We thank Marko Robnik Šikonja for reading the first version of the draft and providing useful comments.

References

Sandesh Adhikary, Siddarth Srinivasan, Jacob Miller, Guillaume Rabusseau, and Byron Boots. Quantum tensor networks, stochastic processes, and weighted automata. In *International Conference on Artificial Intelligence and Statistics*, pages 2080–2088. PMLR, 2021.

Jean Barbier, Florent Krzakala, Nicolas Macris, Léo Miolane, and Lenka Zdeborová. Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460, 2019.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*,

116(32):15849–15854, 2019.

Tai-Danae Bradley and Yiannis Vlassopoulos. Language modeling with reduced densities. *arXiv preprint arXiv:2007.03834*, 2020.

Tai-Danae Bradley, E Miles Stoudenmire, and John Terilla. Modeling sequences with quantum states: a look under the hood. *Machine Learning: Science and Technology*, 1(3):035008, 2020.

Giuseppe Carleo, Ignacio Cirac, Kyle Cranmer, Laurent Daudet, Maria Schuld, Naftali Tishby, Leslie Vogt-Maranto, and Lenka Zdeborová. Machine learning and the physical sciences. *Reviews of Modern Physics*, 91(4):045002, 2019.

Jing Chen, Song Cheng, Haidong Xie, Lei Wang, and Tao Xiang. Equivalence of restricted boltzmann machines and tensor network states. *Physical Review B*, 97(8):085104, 2018.

Yiwei Chen, Yu Pan, and Daoyi Dong. Residual tensor train: a flexible and efficient approach for learning multiple multilinear correlations. *arXiv preprint arXiv:2108.08659*, 2021.

Song Cheng, Lei Wang, Tao Xiang, and Pan Zhang. Tree tensor networks for generative modeling. *Physical Review B*, 99(15):155131, 2019.

Jeremy M Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. *arXiv preprint arXiv:2103.00065*, 2021.

Nadav Cohen, Or Sharir, and Amnon Shashua. On the expressive power of deep learning: A tensor analysis. In *Conference on learning theory*, pages 698–728. PMLR, 2016.

Stéphane d'Ascoli, Levent Sagun, and Giulio Biroli. Triple descent and the two kinds of overfitting: Where & why do they appear? *Advances in Neural Information Processing Systems*, 33:3058–3069, 2020.

Dong-Ling Deng, Xiaopeng Li, and S Das Sarma. Quantum entanglement in neural network states. *Physical Review X*, 7(2):021021, 2017.

Anatoly Dymarsky and Kirill Pavlenko. Tensor network to learn the wavefunction of data. *arXiv preprint arXiv:2111.08014*, 2021.

Stavros Efthymiou, Jack Hidary, and Stefan Leichenauer. Tensornetwork for machine learning. *arXiv preprint arXiv:1906.06329*, 2019.

Andreas Engel and Christian Van den Broeck. *Statistical mechanics of learning*. Cambridge University Press, 2001.

Cong Fang, Hangfeng He, Qi Long, and Weijie J Su. Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training. *Proceedings of the National Academy of Sciences*, 118(43):e2103091118, 2021.

Elizabeth Gardner and Bernard Derrida. Three unfinished works on the optimal storage capacity of networks. *Journal of Physics A: Mathematical and General*, 22(12):1983, 1989.

Ivan Glasser, Ryan Sweke, Nicola Pancotti, Jens Eisert, and Ignacio Cirac. Expressive power of tensor-network factorizations for probabilistic modeling. *Advances in neural information processing systems*, 32, 2019.

Chu Guo, Zhanming Jie, Wei Lu, and Dario Poletti. Matrix product operators for sequence-to-sequence learning. *Physical Review E*, 98(4):042114, 2018.

Like Hui, Mikhail Belkin, and Preetum Nakkiran. Limitations of neural collapse for understanding generalization in deep learning. *arXiv preprint arXiv:2202.08384*, 2022.

Fanjie Kong, Xiao-yang Liu, and Ricardo Henao. Quantum tensor network in machine learning: An application to tiny object classification. *arXiv preprint arXiv:2101.03154*, 2021.

Vignesh Kothapalli, Ebrahim Rasromani, and Vasudev Awatramani. Neural collapse: A review on modelling principles and generalization. *arXiv preprint arXiv:2206.04041*, 2022.

Anders Krogh and John Hertz. A simple weight decay can improve generalization. *Advances in neural information processing systems*, 4, 1991.

Yoav Levine, David Yakira, Nadav Cohen, and Amnon Shashua. Deep learning and quantum entanglement: Fundamental connections with implications to network design. *arXiv preprint arXiv:1704.01552*, 2017.

Ding Liu, Shi-Ju Ran, Peter Wittek, Cheng Peng, Raul Blázquez García, Gang Su, and Maciej Lewenstein. Machine learning by unitary tensor network of hierarchical tree structure. *New Journal of Physics*, 21(7):073059, 2019.

Jing Liu, Sujie Li, Jiang Zhang, and Pan Zhang. Tensor networks for unsupervised machine learning. *arXiv preprint arXiv:2106.12974*, 2021.

Ziming Liu, Ouail Kitouni, Niklas Nolte, Eric J Michaud, Max Tegmark, and Mike Williams. Towards understanding grokking: An effective theory of representation learning. *arXiv preprint arXiv:2205.10343*, 2022.

John Martyn, Guifre Vidal, Chase Roberts, and Stefan Leichenauer. Entanglement and tensor networks for supervised image classification. *arXiv preprint arXiv:2007.06082*, 2020.

Ye-Ming Meng, Jing Zhang, Peng Zhang, Chao Gao, and Shi-Ju Ran. Residual matrix product state for machine learning. *arXiv preprint arXiv:2012.11841*, 2020.

William Merrill and Nikolaos Tsilivis. Extracting finite automata from rnns using state merging. *arXiv preprint arXiv:2201.12451*, 2022.

Dustin G Mixon, Hans Parshall, and Jianzong Pi. Neural collapse with unconstrained features. *arXiv preprint arXiv:2011.11619*, 2020.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003, 2021.

Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020.

Vasily Pestun and Yiannis Vlassopoulos. Tensor network language model. *arXiv preprint arXiv:1710.10248*, 2017.

Mohammad Pezeshki, Amartya Mitra, Yoshua Bengio, and Guillaume Lajoie. Multi-scale feature learning dynamics: Insights for double descent. *arXiv preprint arXiv:2112.03215*, 2021.

Mohammad Pezeshki, Amartya Mitra, Yoshua Bengio, and Guillaume Lajoie. Multi-scale feature learning dynamics: Insights for double descent. In *International Conference on Machine Learning*, pages 17669–17690. PMLR, 2022.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1): 2822–2878, 2018.

James Stokes and John Terilla. Probabilistic modeling with matrix product states. *Entropy*, 21(12): 1236, 2019.

E Miles Stoudenmire. Learning relevant features of data with multi-scale tensor networks. *Quantum Science and Technology*, 3(3):034003, 2018.

E Miles Stoudenmire and David J Schwab. Supervised learning with quantum-inspired tensor networks. *arXiv preprint arXiv:1605.05775*, 2016.

Ananda Streit, Gustavo Santos, Rosa Leão, Edmundo de Souza e Silva, Daniel Menasché, and Don Towsley. Network anomaly detection based on tensor decomposition. In *2020 Mediterranean Communication and Computer Networking Conference (MedComNet)*, pages 1–8. IEEE, 2020.

Zheng-Zhi Sun, Cheng Peng, Ding Liu, Shi-Ju Ran, and Gang Su. Generative tensor network classification model for supervised machine learning. *Physical Review B*, 101(7):075135, 2020.

Vimal Thilak, Etai Littwin, Shuangfei Zhai, Omid Saremi, Roni Paiss, and Joshua Susskind. The slingshot mechanism: An empirical study of adaptive optimizers and the grokking phenomenon. *arXiv preprint arXiv:2206.04817*, 2022.

Jinhui Wang, Chase Roberts, Guifre Vidal, and Stefan Leichenauer. Anomaly detection with tensor networks. *arXiv preprint arXiv:2006.02516*, 2020.

Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big data*, 3(1):1–40, 2016.

Wikipedia. Rule 30, 2022. URL `https://en.wikipedia.org/wiki/Rule_30`.

Stephen Wolfram. Statistical mechanics of cellular automata. *Reviews of modern physics*, 55(3):601, 1983.

Stephen Wolfram et al. *A new kind of science*, volume 5. Wolfram media Champaign, 2002.

Dian Wu, Riccardo Rossi, Filippo Vicentini, and Giuseppe Carleo. From tensor network quantum states to tensorial recurrent neural networks. *arXiv preprint arXiv:2206.12363*, 2022.

Zhihui Zhu, Tianyu Ding, Jinxin Zhou, Xiao Li, Chong You, Jeremias Sulam, and Qing Qu. A geometric analysis of neural collapse with unconstrained features. *Advances in Neural Information Processing Systems*, 34:29820–29834, 2021.

Bojan Žunkovič. Deep tensor networks with matrix product operators. *Quantum Machine Intelligence volume*, 4(21), 2022.

Appendix A. Grokking time in the 1D model

In this appendix, we provide the details of the grokking-time PDF calculation in the 1D model discussed in the main text.

First, we calculate the joint probability to find $x_{\min}$ and $\bar{x}$ for a given number of positive and negative training samples N,

$$P_N(x_{\min},\bar{x}) = \prod_{k=1}^{N}\left[\int_0^\infty \mathrm{d}\tilde{x}_k P(\tilde{x}_k)\int_0^\infty \mathrm{d}x_k P(x_k)\right]\delta\left(x_{\min}-\min(\{x_l\})\right)\delta\left(2\bar{x}-\frac{1}{N}\sum_{l=1}^{N}(x_l-\tilde{x}_l)\right) \tag{69}$$
$$= 2\prod_{k=1}^{N}\left[\int_0^\infty \mathrm{d}x_k P(x_k)\right]\delta\left(x_{\min}-\min(\{x_l\})\right)P_N^{\exp}\left(\bar{x}_+-2\bar{x}\right).$$

With $\bar{x}_+$ we denote the average value of the positive samples, $\bar{x}_+ = \frac{1}{N}\sum_{l=1}^{N} x_l$. We calculate the PDF of the minimum given in Eq. 69 by considering its cumulative density

$$\begin{aligned}
C_N(x_{\min};\bar{x}) =& 2\prod_{k=1}^{N}\left[\int_{x_{\min}}^\infty \mathrm{d}x_k P(x_k)\right]P_N^{\exp}\left(\bar{x}_+-2\bar{x}\right) \qquad (70)\\
=& 2\prod_{k=1}^{N}\left[\int_0^\infty \mathrm{d}x_k P(x_k+x_{\min})\right]P_N^{\exp}\left(\bar{x}_++x_{\min}-2\bar{x}\right)\\
=& 2\prod_{k=1}^{N}\left[\int_0^\infty \mathrm{d}x_k P(x_k)\right]P(Nx_{\min})P_N^{\exp}\left(\bar{x}_++x_{\min}-2\bar{x}\right)\\
=& 2\int_0^\infty \mathrm{d}\bar{x}_+ P_N(\bar{x}_+)P(Nx_{\min})P_N^{\exp}\left(\bar{x}_++x_{\min}-2\bar{x}\right)\\
=& \frac{2^{\frac{3}{2}-N}N^{N+\frac{1}{2}}e^{-Nx_{\min}}\Theta(2\bar{x}-x_{\min})(2\bar{x}-x_{\min})^{N-\frac{1}{2}}K_{N-\frac{1}{2}}(2N\bar{x}-Nx_{\min})}{\sqrt{\pi}\Gamma(n)}\\
&+\frac{\sqrt{\pi}2^{\frac{3}{2}-N}N^{N+\frac{1}{2}}e^{-Nx_{\min}}\csc(\pi N)\Theta(x_{\min}-2\bar{x})(x_{\min}-2\bar{x})^{N-\frac{1}{2}}K_{N-\frac{1}{2}}(N(x_{\min}-2\bar{x}))}{\Gamma(1-N)\Gamma(N)^2}.
\end{aligned}$$

For $N=2$ we find

$$\begin{aligned}
C_{N=2}(x_{\min};\bar{x}) =& -e^{-4\bar{x}}(4\bar{x}-2x_{\min}+1)(\Theta(2x_{\min}-4\bar{x})-1) \qquad (71)\\
&-e^{4\bar{x}-4x_{\min}}(4\bar{x}-2x_{\min}-1)\Theta(2x_{\min}-4\bar{x})
\end{aligned}$$

We obtain the PDF by taking the derivatives of the cumulative probabilities Eq. 70 with respect to $x_{\min}$, namely $P_N(\bar{x},x_{\min}) = -\frac{\partial C_N(x_{\min};\bar{x})}{\partial x_{\min}}$. For $N=2$ we get

$$P_{N=2}(\bar{x},x_{\min}) = 2e^{-4\bar{x}}\Theta(2\bar{x}-x_{\min}) - 2e^{4\bar{x}-4x_{\min}}(8\bar{x}-4x_{\min}-1)\Theta(x_{\min}-2\bar{x}). \tag{72}$$

Next, we calculate the joint probability for the grokking time t_{G} and the average $\bar{x}$

$$P_{N,\epsilon,\lambda}(t_{\mathrm{G}},\bar{x}) = \frac{\partial x_{\min}}{\partial t_{\mathrm{G}}}P_N(\bar{x},x_{\min}(t_{\mathrm{G}},\epsilon,\bar{x}_\lambda)). \tag{73}$$

While we can derive a closed-form expression for an arbitrary N, they are not particularly informative. Thus, we write here only the expressions for $N=2$ and $\lambda_2=0$ (we use t

instead of $t_{\rm G}$ to shorten the notation)

$$P_{N=2,\epsilon,\lambda_1}(t,\bar{x}) =2e^{t-4\bar{x}}(\epsilon_\lambda-\bar{x})\Theta\left(\bar{x}-\epsilon_\lambda\tanh\left(\frac{t}{2}\right)\right) \tag{74}$$
$$+2(\bar{x}-\epsilon_\lambda)e^{4\left(e^t(\bar{x}-\epsilon_\lambda)+\epsilon_\lambda\right)+t}\left(4e^t(\bar{x}-\epsilon_\lambda)+4\bar{x}+4\epsilon_\lambda-1\right)\Theta\left(\epsilon_\lambda\tanh\left(\frac{t}{2}\right)-\bar{x}\right).$$

Finally, we integrate out the average of the samples $\bar{x}$ and obtain the grokking-time PDF.

Appendix B. Grokking probability in the D-dimensional ball model

In this section, we derive the grokking probability in the $D-$dimensional ball model in the limit of many training samples, i.e. $N\gg 1$. The condition for zero test error is

$$\frac{w_1^\lambda}{||w^\lambda||_2}>\frac{1}{\varepsilon}, \tag{75}$$

and can be rewritten as

$$(\epsilon^2-1)(w_1^\lambda)^2\geq(w_2^\lambda)^2+(w_3^\lambda)^2+\ldots(w_D^\lambda)^2,\quad \text{and}\quad (w_1^\lambda)>0. \tag{76}$$

The stationary solution $w^\lambda=G^{-1}a$ is

$$G=\frac{1}{2N}\sum_{i=1}^{2N}\tilde{x}^i\otimes\tilde{x}^i+\lambda_2\mathbb{1}_D=\frac{1}{2N}\sum_{i=1}^{N}x^i\otimes x^i+\epsilon\otimes\epsilon+\epsilon\otimes\bar{x}+\bar{x}\otimes\epsilon+\lambda_2\mathbb{I}_D, \tag{77}$$
$$a=\frac{1}{2N}\sum_{i=1}^{2N}y^i\tilde{x}^i-\lambda_1\text{sgn}(w)=\bar{x}-\lambda_1\text{sgn}(w),$$

where $\bar{x}=\frac{1}{2N}\sum_{i=1}^{2N}y^ix^i$. In the limit $N\gg 1$ we can separate two contributions to the matrix $G=A+B$ where

$$A=\lambda_D\mathbb{I}_D+\epsilon\otimes\epsilon, \tag{78}$$
$$B=\frac{1}{2N}\sum_{i=1}^{2N}x^i\otimes x^i-\frac{1}{D+2}\mathbb{I}_D+\epsilon\otimes\bar{x}+\bar{x}\otimes\epsilon,$$

where $\lambda_D=\frac{1}{D+2}+\lambda_2$. In the limit $N\gg 1$ we have $||B||_F=\mathcal{O}(1/\sqrt{N})$, hence we can approximate the inverse of the matrix G as,

$$G^{-1}\approx A^{-1}-A^{-1}BA^{-1}. \tag{79}$$

The stationary solution can thus be approximated by

$$w^\lambda\approx A^{-1}a-A^{-1}BA^{-1}a \tag{80}$$
$$\approx A^{-1}(\bar{x}+\epsilon-\text{sgn}(w)\lambda_1)-A^{-1}BA^{-1}(\epsilon+\text{sgn}(w)\lambda_1),$$

where we have kept only the first nontrivial order in $1/\sqrt{N}$. We will separately consider the case $\lambda_1=0$ and the case $\lambda_1>0$.

B.1 Case $\lambda_1 = 0$

By explicitly evaluating the above expression, Eq. 80, and assuming $\lambda_1 = 0$, we find

$$w_1^\lambda = \beta + \alpha_1 \bar{x}_1 + \alpha_2 \overline{x_1^2}, \quad w_{j>1}^\lambda = \alpha_3 \bar{x}_j + \alpha_4 \overline{x_1 x_j}, \tag{81}$$

where

$$\begin{aligned} \beta =& \frac{\varepsilon}{\lambda_D + \varepsilon^2} + \frac{\varepsilon}{(\lambda_D + \varepsilon^2)^2 (D+2)}, \\ \alpha_1 =& \frac{1}{\lambda_D + \varepsilon^2} - \frac{2\varepsilon^2}{(\lambda_D + \varepsilon^2)^2}, \\ \alpha_2 =& -\frac{\varepsilon}{(\lambda_D + \varepsilon^2)^2}, \\ \alpha_3 =& \frac{1}{\lambda_D} - \frac{\varepsilon}{\lambda_D (\lambda_D + \varepsilon^2)}, \\ \alpha_4 =& -\frac{\varepsilon}{\lambda_D (\lambda_D + \varepsilon^2)}. \end{aligned} \tag{82}$$

The first few nontrivial moments of the uniform distribution in a D-dimensional ball are reported in Table 2. Considering the variances and the means in Table 2, we find that (in

Statistics	Mean	Second moment
x_j	0	$\frac{1}{D+2}$
x_j^2	$\frac{1}{D+2}$	$\frac{3}{8+6D+D^2}$
$x_i x_j,\ i \neq j$	0	$\frac{1}{8+6D+D^2}$

Table 2: First nontrivial moments of the uniform ball distribution. All odd moments vanish.

the limit $N \gg 1$) all random variables appearing in Eq. 81 to be normally distributed,

$$\begin{aligned} \bar{x}_1 \sim&\ \mathcal{N}\left(0, \frac{1}{2N(D+2)}\right), \\ \overline{x_1^2} \sim&\ \mathcal{N}\left(\frac{1}{D+2}, \frac{D+1}{N(D+2)^2(D+4)}\right), \\ \overline{x_1 x_{j>1}} \sim&\ \mathcal{N}\left(0, \frac{1}{2N(8+6D+D^2)}\right). \end{aligned} \tag{83}$$

The distributions are independent since all the necessary covariances vanish.

The sum of independent normal distributions is again a normal distribution, leading to

$$w_1 \sim\ \mathcal{N}_1(w_1) = \mathcal{N}\left(\beta + \frac{\alpha_2}{D+2}, \frac{\alpha_1^2}{2N(D+2)} + \frac{\alpha_2^2 (D+1)}{N(D+2)^2(D+4)}\right), \tag{84}$$

$$w_{j>1} \sim\ \mathcal{N}\left(0, \frac{\alpha_3^2}{2N(D+2)} + \frac{\alpha_4^2}{2N(8+6D+D^2)}\right). \tag{85}$$

The grokking probability is then given by

$$P_{E(\infty)=0} = \int_0^\infty \mathrm{d}w_1 \mathcal{N}_1(w_1) \int_0^{(\epsilon^2-1)w_1^2 \left(\frac{\alpha_3^2}{2N(D+2)} + \frac{\alpha_4^2}{2N(8+6D+D^2)}\right)} \mathrm{d}r\, \chi_{D-1}^2(r). \tag{86}$$

In the limit $N \gg \lambda_D \gg 1$ we make the following approximations

$$\beta \approx \frac{\varepsilon}{\lambda_D + \epsilon^2}, \quad \alpha_1 \approx \frac{1}{\lambda_D + \epsilon^2}, \quad \alpha_2 = 0, \quad \alpha_3 \approx \frac{1}{\lambda_D}, \quad \alpha_4 = 0. \tag{87}$$

With these simplifications, Eq. 86 reduces to the grokking probability obtained in the main text, see Eq. 30.

B.2 Case $\mathbf{1 \gg \lambda_1 > 0}$

By explicitly evaluating the expression in Eq. 80 we find

$$w_1^\lambda = \begin{cases} \beta - \frac{\lambda_1 \mathrm{sgn}(w_1^\lambda)}{\lambda_D + \varepsilon^2} + \alpha_1 \bar{x}_1 + \alpha_2 \overline{x_1^2} & , \ (\lambda_D + \varepsilon^2)|\beta + \alpha_1 \bar{x}_1 + \alpha_2 \overline{x_1^2}| > \lambda_1 \\ 0 & , \text{ else} \end{cases}, \tag{88}$$

$$w_{j>1}^\lambda = \begin{cases} -\frac{\lambda_1 \mathrm{sgn}(w_j^\lambda)}{\lambda_D} + \alpha_3 \bar{x}_j + \alpha_4 \overline{x_1 x_j} & , \ \lambda_D |\alpha_3 \bar{x}_j + \alpha_4 \overline{x_1 x_j}| > \lambda_1 \\ 0 & , \text{ else} \end{cases},$$

with α_j and β given in Eq. 82. The number of non-vanishing components of the stationary solution w^λ depends on the value of λ_1. Therefore, we get (in the $N \gg 1$ limit) an additional sum over the number of non-zero elements in the w^λ,

$$P_{E(\infty)=0} = \int_{\frac{\lambda_1}{\lambda_{2,D}+\varepsilon^2}}^{\infty} \mathrm{d}w_1 \mathcal{N}_1(w_1) \Bigg[(1-p_\lambda)^{D-1} \tag{89}$$
$$+ \sum_{k=1}^{D-1} \binom{D-1}{k} p_\lambda^k (1-p_\lambda)^{D-1-k} \int_0^{(\epsilon^2-1)\left(w_1 - \frac{\lambda_1}{\lambda_{2,D}+\varepsilon^2}\right)^2 \left(\frac{\alpha_3^2}{2N(D+2)} + \frac{\alpha_4^2}{2N(8+6D+D^2)}\right)} \mathrm{d}r R_k(r) \Bigg],$$

where $p_\lambda = 1 - \mathrm{erf}\left(\frac{\lambda_1}{\lambda_D} / \sqrt{\frac{\alpha_3^2}{2N(D+2)} + \frac{\alpha_4^2}{2N(8+6D+D^2)}} \right)$ is the probability that $|w_{j>1}|$ (sampled from Eq. 85)is larger than λ_1/λ_D. With $R_k(r)$ we denote the PDF of the sum of squares of k random variables sampled from the truncated normal distribution. As in the main text (see Section 3.3.2), we can calculate a lower bound on the grokking probability by discarding the sum over $k > 1$,

$$P_{E(\infty)=0} \geq (1-p_\lambda)^{D-1} \int_{\frac{\lambda_1}{\lambda_{2,D}+\varepsilon^2}}^{\infty} \mathrm{d}w_1 \mathcal{N}_1(w_1) \tag{90}$$
$$\approx \frac{(1-p_\lambda)^{D-1}}{2} \left(1 + \mathrm{erf}\left(\left(\beta + \frac{\alpha_2}{D+2} - \frac{\lambda_1}{\lambda_{2,D}+\epsilon^2} \right) / \sqrt{\frac{\alpha_1^2}{2N(D+2)} + \frac{\alpha_2^2(D+1)}{N(D+2)^2(D+4)}} \right) \right).$$

Also in the more general case, we find the same difference between the L_1 and L_2 regularisations as discussed in the main text.

Appendix C. Fixed attention

Below we specify the attention tensors for the discussed examples, see Section 4.3.1 in the main text:

- `Example 1`:

$$A_0 = \begin{pmatrix} 0.782735 & 0.225481 \\ -0.21562 & 0.290028 \end{pmatrix}, \quad A_1 = \begin{pmatrix} 1.17554 & -0.275503 \\ 1.18283 & -0.157563 \end{pmatrix},$$

- `Example 2`:

$$A_0 = \begin{pmatrix} 1.6749 & -1.29059 \\ 0.285324 & -0.708621 \end{pmatrix}, \quad A_1 = \begin{pmatrix} 0.0462428 & -0.0797724 \\ -0.509457 & 0.922777 \end{pmatrix},$$

- `Example 3`:

$$A_0 = \begin{pmatrix} 1.37336 & -0.465853 \\ -1.10382 & 0.720113 \end{pmatrix}, \quad A_1 = \begin{pmatrix} 0.128517 & 0.166033 \\ 0.634426 & 1.13816 \end{pmatrix}.$$

Appendix D. Additional results for the 2–local and the 3–local rules

In this appendix, we present a similar analysis as the one in Section 4.3.2, but for the 2–local rule and the 3–local rule. In general, we observe a similar behaviour as in the 1–local case discussed in the main text.

First, we discuss the test error shown in Fig. 24. Again, we find that the critical exponent ν is roughly independent of the model size and is smaller for larger rule range K.

As in the 1–local case, the grokking transition corresponds to a sharp decrease in the effective dimension of the $H^{\mathrm{L}}(i)$ attention vectors, shown in Fig. 25. In the considered rule-learning scenario, the smallest effective dimension is determined by the locality of the rule and is expected to increase exponentially with K Žunkovič (2022). In contrast to the 1–local case, we find that larger regularisation does not necessarily correspond to a smaller effective dimension. However, this is only the case for smaller instances where the bond dimension is very close to the effective dimension (or the smallest possible bond dimension with zero test error). For the larger instances, we again find that larger regularisation corresponds to smaller effective dimension D_{eff}. In this case, we also find that models with L_1 regularisation have a slightly smaller average effective dimension compared to models with L_2 regularisation.

Finally, we estimate the grokking-time PDF and the generalisation-time (t_ϵ) PDF for the 2–local and the 3–local rule, shown in Fig. 26 and Fig. 27.

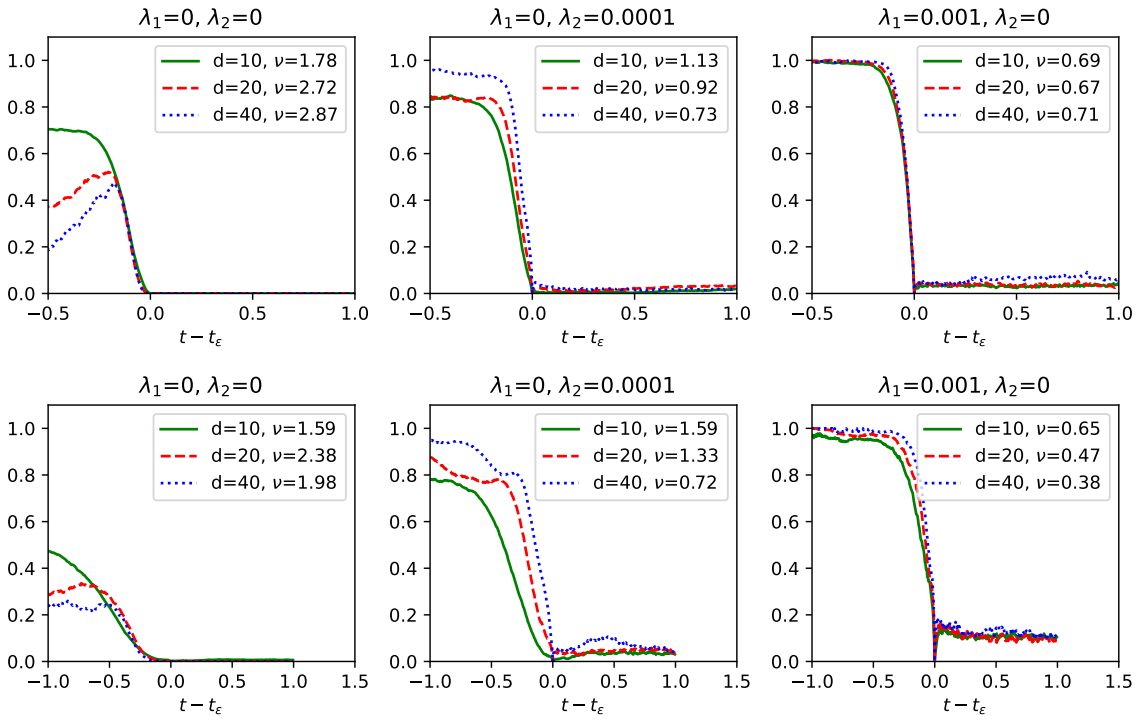


Figure 24: The average error at the phase transition for the 2–local rule (first row) and the 3–local rule (second row). We align the first point where the test error becomes zero (i.e. the time t_ϵ) and take the average over many (~ 1000) initialisations of the model parameters. As in the 1–local rule case the training without regularisation results in larger critical exponent $\nu \approx 2$ as training with L_1 or L_2 regularisation. In all experiments, we used a learning rate of 0.005.

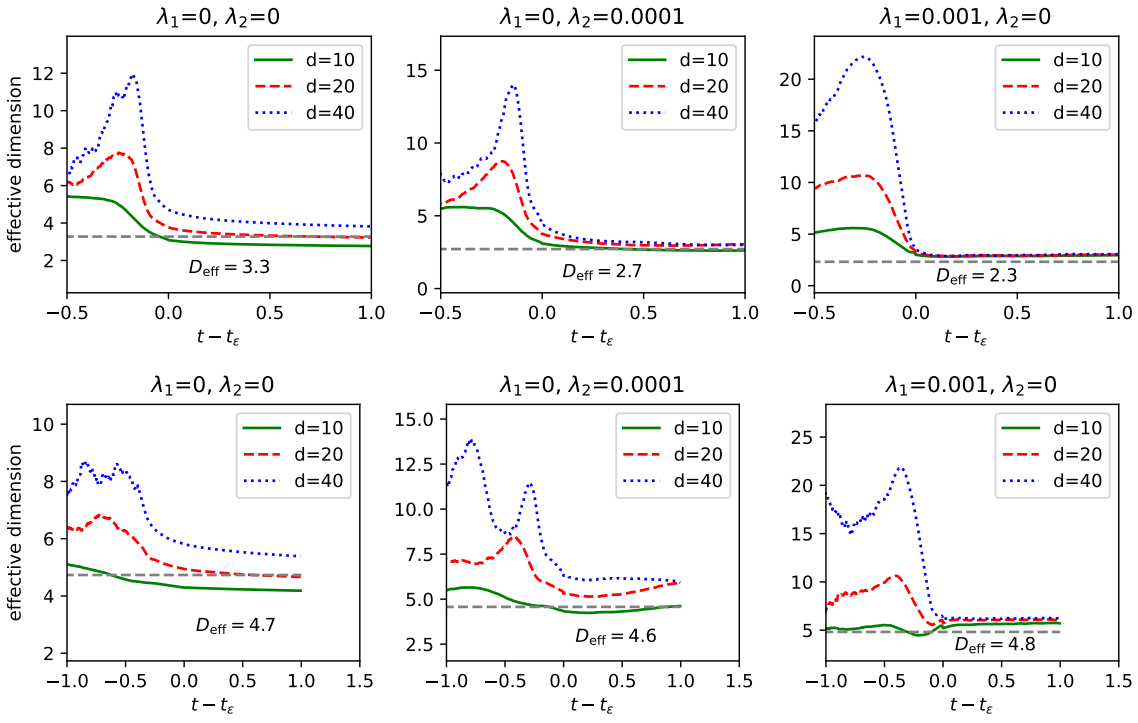


Figure 25: The average effective dimension of the 2–local rule (first row) and the 3–local rule (second row) corresponding to the error in Fig. 24. The horizontal dashed line corresponds to the average (over the bond dimension d) of the minimal effective dimension of data with zero test error (over all samples with fixed d).

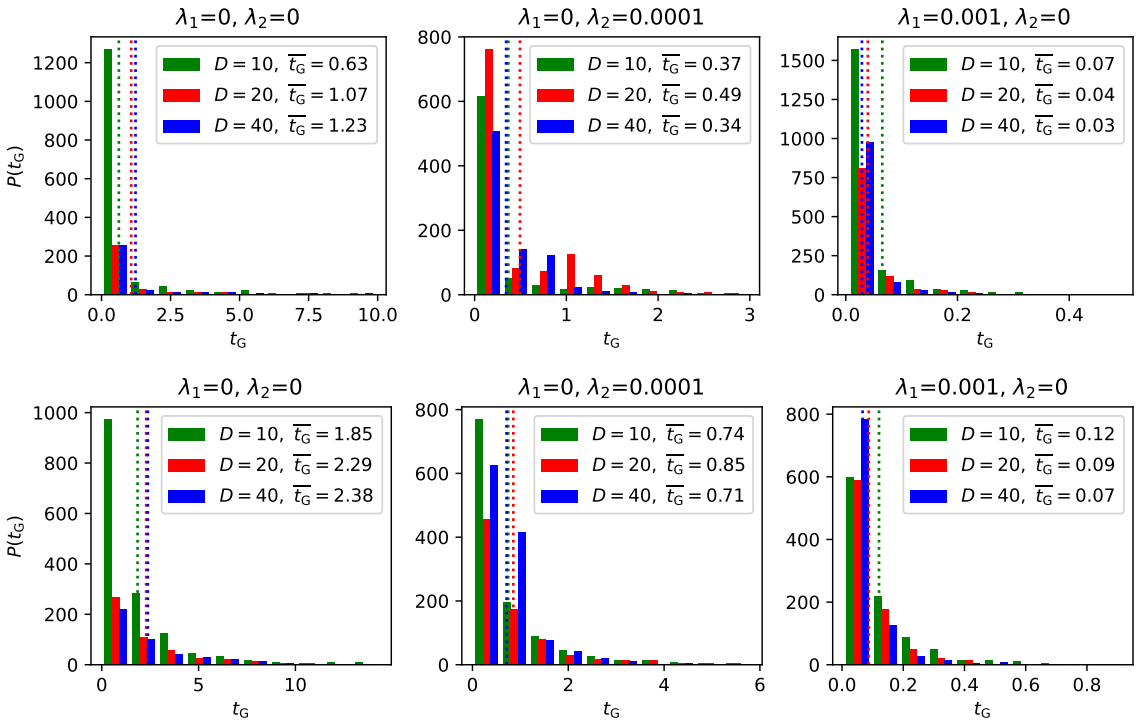


Figure 26: The estimated grokking-time PDF for the 2–local rule (first row) and the 3–local rule (second row). The colours correspond to different model sizes, namely $d = 10$ (blue), $d = 20$ (orange), and $d = 40$ (green). The vertical lines correspond to the averages reported in the legends of the panels. We find that grokking time increases with increased rule range K. As in the 1–local case, we find that L_1 regularisation reduces t_G in all cases. In contrast, the use of L_2 regularisation can in some cases increase the average grokking time. In the L_2 case, we also find a clear bimodal grokking time distribution. In all cases, we set the learning rate to 0.005.

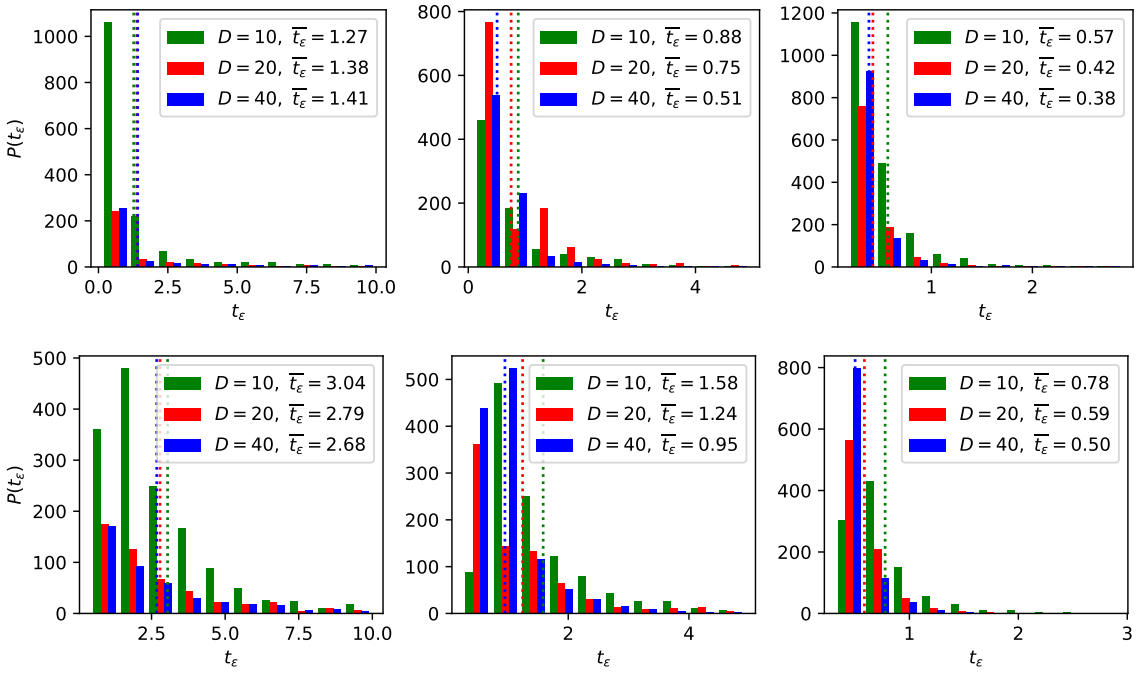


Figure 27: The estimated t_ϵ PDF for the 2–local rule (first row) and the 3–local rule (second row). The colours correspond to different model sizes, namely $d = 10$ (blue), $d = 20$ (orange), and $d = 40$ (green). The vertical lines correspond to the averages reported in the legends of the panels. In all cases, we set the learning rate to 0.005.