Grokking Finite-Dimensional Algebra

Pascal Jr Tikeng Notsawo [1,2] Guillaume Dumas [1,2,3] Guillaume Rabusseau[1,2,4]

[1]Université de Montréal [2]Mila - Quebec AI Institut [3]CHU Sainte-Justine Research Center [4]CIFAR AI Chair

Question and Contributions

Grokking is usually studied on finite groups. What changes when the operation is a general finite-dimensional algebra (FDA), possibly non-associative, non-commutative, or non-unital?

- We encode every n-dimensional FDA over a field $\mathbb{F}$ by a tensor $\mathcal{C} \in \mathbb{F}^{n\times n\times n}$
- We show group-operation grokking is a special case of FDA grokking.
- We test how algebraic properties and tensor complexity affect grokking
- We probe whether generalization coincides with the emergence of algebra-consistent representations in model feature space

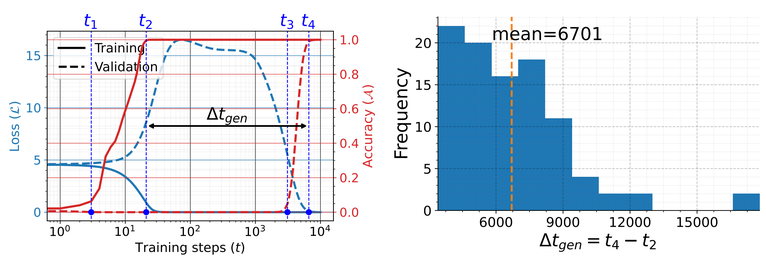

Figure 1. Grokking on the algebra of complex numbers over the field $\mathbb{F}_7 = \mathbb{Z}/7\mathbb{Z}$ and sensitivity to multiplication-rule perturbations.

Tensor View of FDAs

- $n\mathbb{F}$-FDA : n-dimensional FDA over a filed $\mathbb{F}$
- Fix a basis $(\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(n)})$ of a $n\mathbb{F}$-FDA $(\mathfrak{A}, \cdot)$: $\mathbf{a}^{(i)} \cdot \mathbf{a}^{(j)} = \sum_{k=1}^{n} \mathcal{C}_{ijk}\mathbf{a}^{(k)}$

Any $\mathcal{C} \in \mathbb{F}^{n\times n\times n}$ defines a $n\mathbb{F}$-FDA, unique up to isomorphism.

- Commutative iff $\mathcal{C}_{ijk} = \mathcal{C}_{jik}$.
- Unital iff $\exists \lambda \in \mathbb{F}^n$ such that $\sum_i \lambda_i \mathcal{C}_{i,:,:} = \sum_i \lambda_i \mathcal{C}_{:,i,:} = \mathbb{I}_n$.
- Associative iff $\sum_k \mathcal{C}_{ijk}\mathcal{C}_{klm} = \sum_k \mathcal{C}_{ikm}\mathcal{C}_{jlk}$.

Groups Are Embedded

For a finite group $(G, \circ)$ with n elements $\{g_i\}_{i\in[n]}$, the Cayley table becomes the structure tensor of the group algebra:

$$\mathcal{C}_{ijk} = \mathbf{1}\{g_i \circ g_j = g_k\}$$

Thus classical algorithmic grokking on groups lives inside the FDA framework.

Learning Multiplication over $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$

- We have $\mathfrak{A} = \mathbb{F}_p^n$ and $\mathbf{u} \cdot \mathbf{v} = \mathcal{C} \times_1 \mathbf{u} \times_2 \mathbf{v} = \mathcal{C}_{(3)}(\mathbf{v} \otimes \mathbf{u})$ for all $\mathbf{u}, \mathbf{v} \in \mathfrak{A}$.
- Each algebra element is a token with a learned embedding.
- MLP/LSTM/Transformer encoders a pair $(\mathbf{u}, \mathbf{v})$ as $\mathbf{f}(\mathbf{u}, \mathbf{v}) \in \mathbb{R}^d$
- A linear classifier takes $\mathbf{f}(\mathbf{u}, \mathbf{v})$ and predicts the product $\mathbf{u} \cdot \mathbf{v}$.

Representation Mechanism

A matrix representation is a family $\{\mathcal{R}_i\}_{i=1}^n \subset \mathbb{F}^{m\times m}$ satisfying

$$\mathcal{R}_i\mathcal{R}_j = \sum_{k=1}^{n} \mathcal{C}_{ijk}\mathcal{R}_k \; \forall i, j \in [n] \tag{1}$$

If features of a neural networks implement this multiplication up to a linear change of coordinates (bilinear probe), a linear classifier can use these features to predict $\mathbf{u} \cdot \mathbf{v}$ given $\mathbf{u}$ and $\mathbf{v}$ in $\mathfrak{A}$. This matches the representation-learning view of grokking.

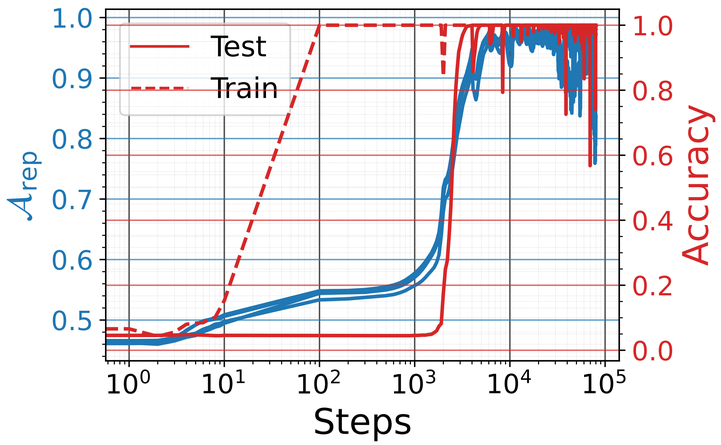

Figure 2. Before grokking the representation quality $\mathcal{A}_{\text{rep}} \in [0, 1]$ remains relatively low and then rises sharply at grokking.

References

[1] Pascal Jr Tikeng Notsawo et al. Grokking beyond the euclidean norm of model parameters, 2025.

[2] Power et al. Grokking: Generalization beyond overfitting on small algorithmic datasets. 2022.

Algebraic Properties Matter

For $(n, p) = (2, 7)$, we stratify all $p^{n^3} = 7^8$ tensors by associativity (a), commutativity (c), and unitality (u), then vary the training data fraction $r \in (0, 1)$.

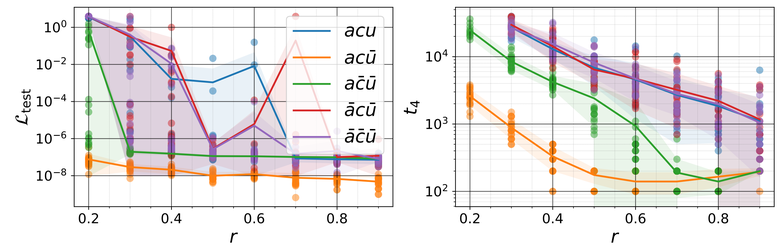

Figure 3. Test loss and grokking step across algebraic regimes.

- $ac\bar{u}$ algebras are easiest: earlier grokking and lower test loss.
- Non-unital associative algebras admit accessible representation solutions.
- The unit constraint $\sum_i \lambda_i \mathcal{R}_i = \mathbb{I}_m$ shrinks the feasible representation set.
- Commutativity reduces the number of constraints.

Structure Tensor Complexity: Rank and Sparsity

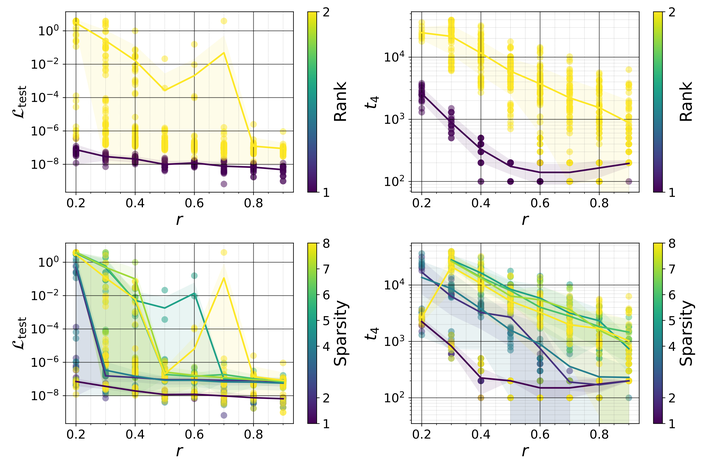

Figure 4. Higher rank/sparsity gives larger test loss and longer delays. Rank is basis-invariant and tracks multilinear complexity.

Contact: pascalnotsawo@gmail.com Paper: https://arxiv.org/abs/2602.19533 Code: github.com/Tikquuss/grokking_fda