Provable Scaling Laws of Feature Emergence from Learning Dynamics of Grokking

Yuandong Tian (yuandong.tian@gmail.com)

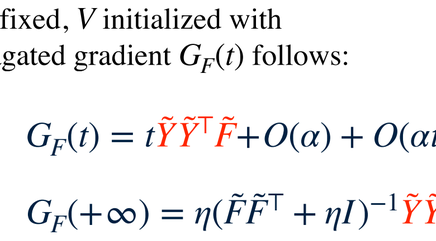
Paper link: https://arxiv.org/abs/2509.21519

Code link: https://github.com/yuandong-tian/understanding

Motivation

Representation
- (Traditional) Symbolic representation
- Neural Representation 🤔

$$\nabla\cdot\mathbf{E}=\frac{\rho_v}{\varepsilon} \quad \text{(Gauss' Law)}$$
$$\nabla\cdot\mathbf{H}=0 \quad \text{(Gauss' Law for Magnetism)}$$
$$\nabla\times\mathbf{E}=-\mu\frac{\partial\mathbf{H}}{\partial t} \quad \text{(Faraday's Law)}$$
$$\nabla\times\mathbf{H}=\mathbf{J}+\varepsilon\frac{\partial\mathbf{E}}{\partial t} \quad \text{(Ampere's Law)}$$



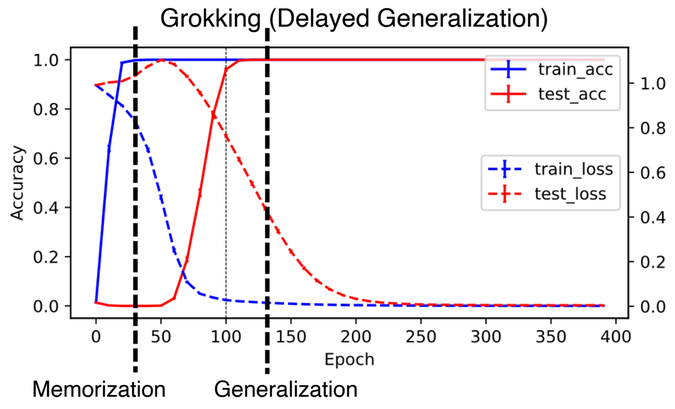
4 Conclusions

We have extended the GLU family of layers and proposed their use in Transformer. In a transfer-learning setup, the new variants seem to produce better perplexities for the de-noising objective used in pre-training, as well as better results on many downstream language-understanding tasks. These architectures are simple to implement, and have no apparent computational drawbacks. We offer no explanation as to why these architectures seem to work; we attribute their success, as all else, to divine benevolence.

[N. Shazeer, *GLU Variants Improve Transformer*]

*Shall we just acknowledge neural representations as "**divine benevolence**"?*

Problem Setting



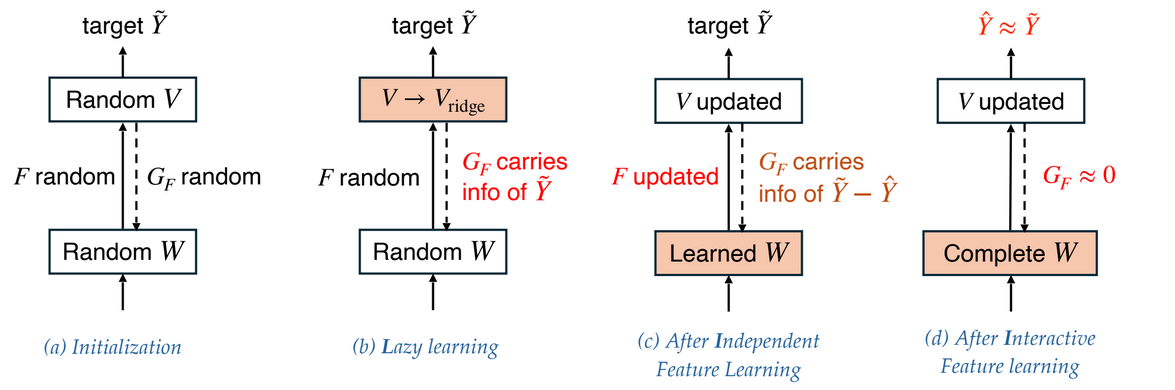
The $\mathbf{Li}_2$ framework

- Memorization (Sharp Optima)
  - (I) Ridge Regression on Random Features
- Generalization (Flat Optima)
  - (II) Feature Learning on Independent Neurons
  - (III) Neuron Interactions

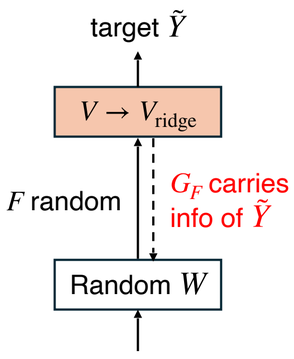
Weight Initialization

The back-propagated gradient G_F matters

MSE loss $J(W,V)=\frac{1}{2}\|P_1^{\perp}(Y-\sigma(XW)V)\|_2^2$ Activation $F=\sigma(XW)$ Zero-mean activation $\tilde{F}=P_1^{\perp}F$ Zero-mean target $\tilde{Y}=P_1^{\perp}Y$



Stage I: Lazy Learning (Overfitting)

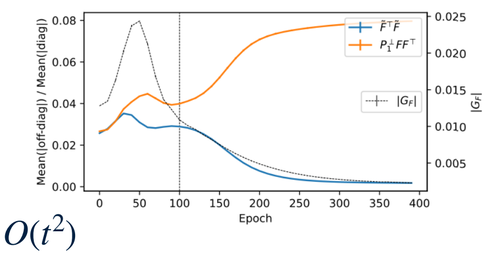


(Proposition 1) Assume $\tilde{F}$ is fixed, V initialized with $N(0,\alpha^2)$, then the back propagated gradient $G_F(t)$ follows:

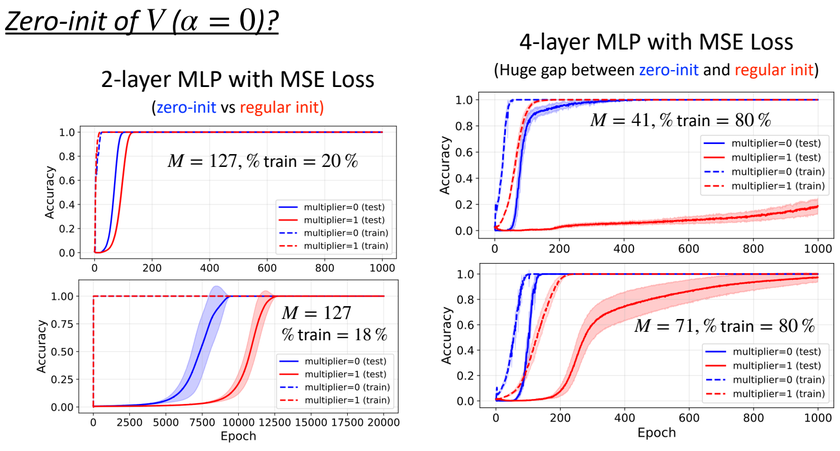
$$\begin{cases} \text{Initial stage } (t \text{ small}) & G_F(t)=t\tilde{Y}\tilde{Y}^\top\tilde{F}+O(\alpha)+O(\alpha t)+O(t^2) \\ \text{Final stage} & G_F(+\infty)=\eta(\tilde{F}\tilde{F}^\top+\eta I)^{-1}\tilde{Y}\tilde{Y}^\top\tilde{F}(\tilde{F}^\top\tilde{F}+\eta I)^{-1} \end{cases}$$

(Lemma 1) $G_F(+\infty)=\frac{\eta}{(Kc_1+\eta)(nc_2+\eta)}\tilde{Y}\tilde{Y}^\top\tilde{F}+O(K^{-1}\epsilon)$

- If weight decay $\eta=0$, then $G_F=0$ (no feature learning)
- If weight decay is large, then $G_F\to 0$
- If number of hidden nodes $K\to+\infty$, then $G_F\to 0$ (NTK regime)



Zero-init of V ($\alpha=0$)?



Stage II: Independent Feature Learning

Local maxima of $\mathscr{E}$ are the **emergent features**

Existence of energy function $\mathscr{E}$

G_F carries info of $\tilde{Y}$

Random W

$G_F \propto \tilde{Y}\tilde{Y}^\top F$ → *Each nodes learn independently* → $\begin{cases}\dot{\mathbf{w}}_j=X^\top\mathrm{diag}(\sigma'(X\mathbf{w}_j))\mathbf{g}_j \\ \mathbf{g}_j\propto\tilde{Y}\tilde{Y}^\top\sigma(X\mathbf{w}_j)\end{cases}$ $\dot{\mathbf{w}}_j=\nabla_{\mathbf{w}}\mathscr{E}$ $\mathscr{E}(\mathbf{w})=\frac{1}{2}\|\tilde{Y}^\top\sigma(X\mathbf{w})\|_2^2$

Properties of local maxima of $\mathscr{E}$

Group Arithmetic Tasks: Given $h_1,h_2\in H$, predict h_1h_2. Here H is a group with size $|H|=M$.

Analytic form of energy $\mathscr{E}(\mathbf{w})=\frac{1}{2}\sum_h\langle\tilde{R}_h,S\rangle_F^2=\frac{M}{2}\sum_{k\neq 0}\frac{1}{d_k}\left|\sum_r \mathrm{tr}(\hat{S}_{k,r})\right|^2$ → ***Strict*** local maxima for each frequency k

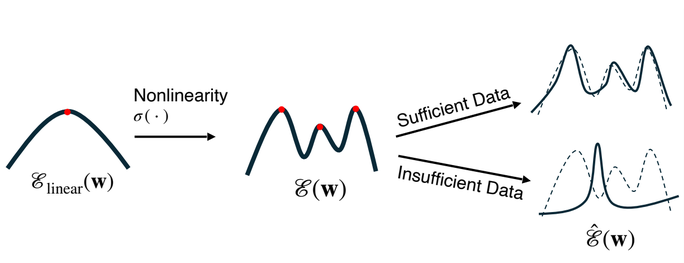
($\hat{S}_{k,r}$: Fourier transform of S at frequency k)

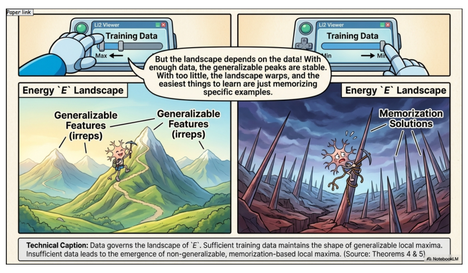
$\mathbf{w}=[\mathbf{w}_a,\mathbf{w}_b]$ $\quad S_{ij}:=\sigma(w_{ai}+w_{bj})$

$\tilde{R}_h=Q\left(\bigoplus_{k\neq 0}\bigoplus_{r=1}^{m_k}C_k(h)\right)Q^*$

(Theorem 3) For Abelian group H, it suffices to use $2M-2$ hidden nodes to reconstruct $\hat{Y}$, compared to M^2 hidden nodes with memorization.

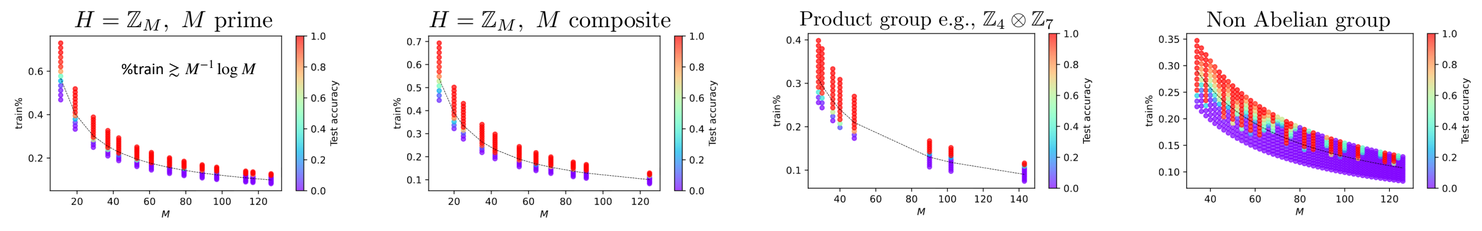
Provable Scaling Laws: Memorization 👉 Generalization





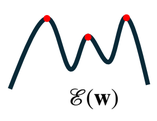
(Theorem 4) If $\%\text{train}=\frac{n}{M^2}\gtrsim\frac{\log M}{M}$, then the local maxima stay.

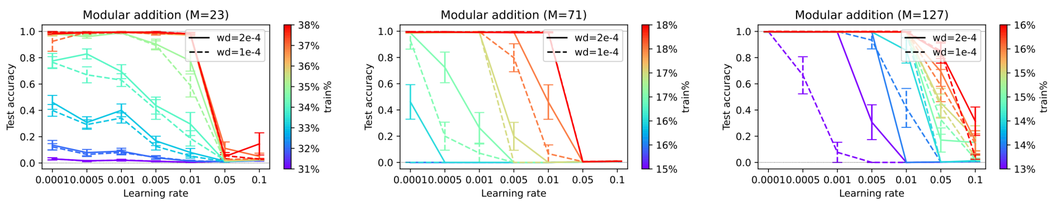
Strict local maxima matters.

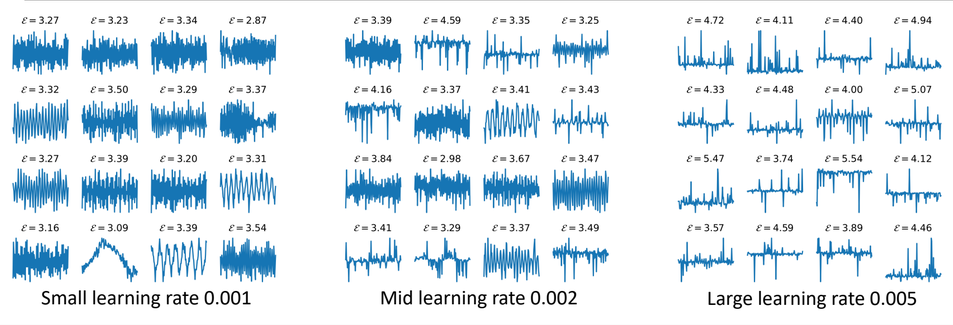


At the boundary of memorization / generalization

At the boundary, large Learning rate leads to memorization (higher $\mathscr{E}$)







When does Grokking happen?

Learning rate. Grokking can occur without regularization if a **large initial learning rate** is used [1]. In $\mathbf{Li}_2$'s view, this leads to increased $G_F\propto\tilde{Y}\tilde{Y}^\top F$ at the initial stage ⇒ hidden nodes receives sufficient feature learning signal.

Loss function. Using *stable softmax (linear form)* instead of softmax *(exponential form)* delays overfitting and maintains a nonzero G_F useful for feature learning [2].

Stay longer before overfitting. Grokking without regularization has also been observed with vanilla SGD [3]. In $\mathbf{Li}_2$'s view, slower convergence to V_{ridge} gives the hidden layer more time to accumulate G_F before it vanishes.

Weight initialization scale α. Small initialization promotes grokking regardless of weight decay [4]. In $\mathbf{Li}_2$'s view, $G_F(t)=O(\alpha)+t\tilde{Y}\tilde{Y}^\top F+O(\alpha t)+O(t^2)$. With small α, the clean term $t\tilde{Y}\tilde{Y}^\top F$ dominates, enabling grokking. With large α, $O(\alpha)$ dominates, the training needs to wait for weight decay η to expose the signal $\tilde{Y}\tilde{Y}^\top F$ later.

Output scaling factor β. Scaling with $\beta>1$ accelerates grokking [3,5]. In $\mathbf{Li}_2$'s view, this amplifies the initial gradient $G_F(t)=O(\alpha)+t[\beta\tilde{Y}\tilde{Y}^\top F+O(\alpha\beta^2)]+O(t^2)$. So large $\beta>1$ makes the term $t\beta\tilde{Y}\tilde{Y}^\top F$ dominate earlier and the transition occurs faster. β cannot be too large since it will destroy initial G_F and make $G_F(+\infty)$ much smaller.

Weight decay η. Because $G_F(+\infty)\propto\eta\,\tilde{Y}\tilde{Y}^\top F$, in $\mathbf{Li}_2$'s view, weight decay servers as the *learning rate* in the feature learning, consistent with empirical findings that grokking follows $t\sim 1/\eta$ scaling laws [4,6].

Data size n. $\mathbf{Li}_2$'s sample complexity analysis proves a threshold above which memorization becomes generalization, aligned with [1,8,9,10,12]. More samples ⇒ more stable local maxima ⇒ earlier grokking, consistent with [6,8,9,10].

Number of hidden nodes K. $\mathbf{Li}_2$ requires sufficient K to capture diverse local maxima, and over-parameterization makes $G_F(+\infty)$ close to $\tilde{Y}\tilde{Y}^\top F$, facilitate feature learning. This is consistent with experiments. However, very large K ⇒ $G_F(+\infty)$ small + noisy initial stage $G_F(t)$, consistent with [5] and NTK [11] that feature learning does not happen.

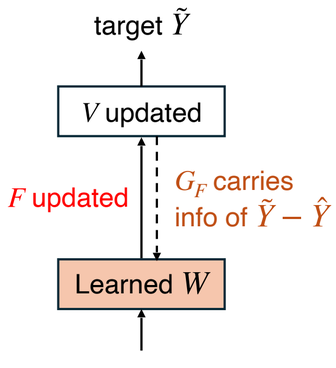
Early stopping. Model may overfit after generalization [13]. $\mathbf{Li}_2$ shows that if data are scarce, generalization solutions have lower energy than memorization ones. If trained long enough, weights will move to solutions with larger $\mathscr{E}$.

Muon optimizers. $\mathbf{Li}_2$ shows that Muon can lead to more diverse weights in the stage of interaction feature learning.

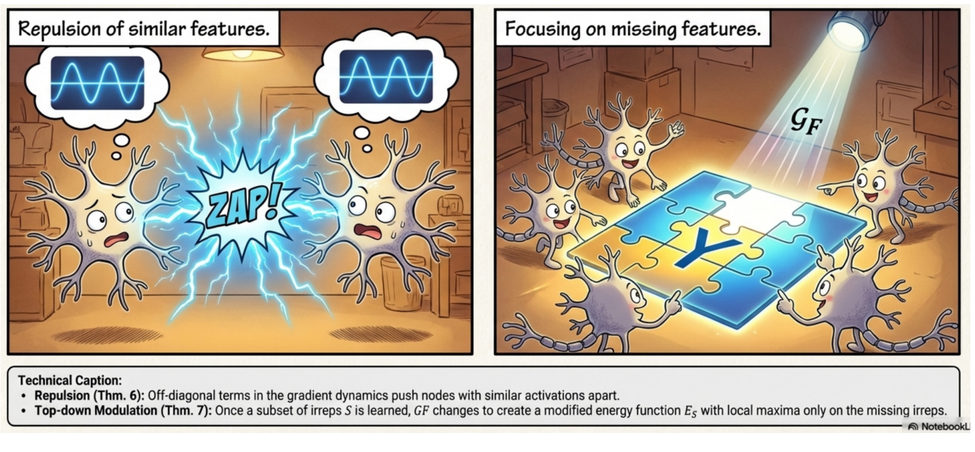
References

[1] A. Gromov (2023): Grokking modular arithmetic.
[2] L. Prieto et al. (2025): Grokking at the edge of numerical stability.
[3] T. Kumar et al. (2024): Grokking as the Transition from Lazy to Rich Training Dynamics
[4] Liu et al. (2023): Omnigrok: Grokking Beyond Algorithmic Data
[5] Chizat et al. (2019): On lazy training in differentiable programming
[6] Power et al. (2022): Grokking: Generalization beyond overfitting on small algorithmic datasets
[7] Clauw et al. (2024): Information-theoretic progress measures reveal grokking is an emergent phase transition
[8] N. Nanda et al. (2023): Progress measures for grokking via mechanistic interpretability
[9] B. Wang et al. (2024): Grokked Transformers are Implicit Reasoners: A Mechanistic Journey to the Edge of Generalization
[10] R. Abramov et al. (2025): Grokking in the Wild: Data Augmentation for Real-World Multi-Hop Reasoning with Transformers
[11] A. Jacot et al. (2018): Neural Tangent Kernel: Convergence and Generalization in Neural Networks
[12] Liu et al. (2022): Towards understanding grokking: An effective theory of representation learning
[13] Montanari et al. (2025) Dynamical Decoupling of Generalization and Overfitting in Large Two-Layer Networks

Stage III (Interactive Feature Learning)



Feature Repulsion (Theorem 6) Top-Down Modulation (Theorem 7)



Future Works

- ✅ Perform analysis of the gradient dynamics
- ✅ Analyze sample complexity and allow missing data
- ✅ Analyze grokking behaviors and hyperparameters related
- 🤔 Based on quadratic activations but can go beyond
- 🤔 Based on 2-layer networks, have a sketch on multilayers
- ❓ Modern Architectures (Attention, etc.)
- ❓ Generalized to tasks beyond group arithmetics