# Scaling & Emergence Projects

**Main Doc for All Projects**



---

## Basic Info

1. Call in link
2. **Discussions:** **AGI discord**, **public channel #grokking-emergence** (you are welcome to send me a friend request to irina#1789 so I can add you to private project channels)
3. List of relevant papers
4. Reading group webpage
5. Working document

## Minutes

### Mon June 17, 2024

Papers discussed:
Jared Kaplan's et al papers on manifold dimension to predict the exponent (edited)

Sharma et. al.: https://arxiv.org/abs/2004.10802
Bahri et. al.: https://arxiv.org/abs/2102.06701
Recent Pennington's paper on scaling laws (4+3....): https://arxiv.org/abs/2405.15074

### Mon Apr 8, 2024

- Saturating small models with data. What is the best possible perplexity one can achieve with a model of a fixed size, given that we have infinite data and compute?
    a. Data quality and de-duplication (e.g. SlimPajama). Note: also might be worth considering multiple epochs here.
    b. Data selection (e.g. paper from Surya Ganguli's group, 2024 review from AI2)
    c. Distillation
    d. Application to continual learning
- Grokking survey: Try to distill the literature on grokking – possibly with commentary on useful future directions

- Scaling laws for SSM and hybrid models (e.g. Mamba and Jambda)
    a. Chinchilla-type analysis with perplexity
    b. Downstream evaluation

### Mon Apr 1, 2024

1. Possible topic 1: optimal hyperparameter extrapolation and scaling laws? (Based on Tensor Programs series of papers).
2. Possible topic 2: comparison of scaling laws (with data, mode, size, compute a la Chinchilla paper) for SSM-based models (including various hybrids) and transformers. This can include both “upstream” metrics (standard test loss scaling laws) and “downstream”, task-dependent scaling laws. Go smol or go home. (Look at Mamba and Jambda papers.)
   Also, comparing performance on specific tasks that are hard for transformers, and vice versa (are their tasks hard for SSM but easy for transformers).
3. Combination of closed-box scaling laws for non-linear scaling (e.g. Broken Neural Scaling Laws) and open-box predictors of scaling/grokking (e.g. Predicting Grokking Long Before it Happens and many others - basically, use the knowledge of the functional form and try to estimate its parameters from the learning dynamics (loss, activation, weight properties).
   What can be done in terms of deriving the parameters of BNSL from data, model, and learning dynamics properties? Similar motivation to Scaling Laws from the Data Manifold Dimension
4. Links to Paul Bogdan’s work that I mentioned - potentially interesting for scaling and emergence:
   - Quantifying emergence in LLMs is described in https://arxiv.org/pdf/2402.09099.pdf but the formalism is general and we can extend it to any ML algorithm and architecture
   - Distributed neural networks redescribed in this paper
   https://openreview.net/pdf?id=Ys3uPmZGOR
   - Inferring network generators from partially observable weighted graphs
   https://www.nature.com/articles/s42005-021-00701-5
   - learning PDEs, forecasting PDEs, and learning + solving coupled PDEs
   https://proceedings.neurips.cc/paper_files/paper/2021/file/c9e5c2b59d98488fe1070e744041ea0e-Paper.pdf https://openreview.net/pdf?id=d2TT6gK9qZn
   https://openreview.net/pdf?id=kIo_C6QmMOM
   - inferring nano-particle to protein interactions
   https://www.nature.com/articles/s43588-022-00229-w
   - learning the community structure in weighted networks
   https://www.nature.com/articles/s41598-019-46079-x
   https://www.nature.com/articles/s41598-022-14631-x and detecting phase transitions complex networks including the learning phases in DNNs
   https://www.nature.com/articles/s41598-023-44791-3
   - quantifying trust/trustworthiness in ML algorithms
   https://www.frontiersin.org/articles/10.3389/frai.2020.00054/full or

https://www.ifaamas.org/Proceedings/aamas2021/pdfs/p332.pdf or https://proceedings.mlr.press/v199/cheng22a/cheng22a.pdf

We also worked on learning non-Markovian dynamic networks or fractional order dynamical systems from partially observed data https://ieeexplore.ieee.org/document/8815074, controlling non-Markovian dynamic networks https://www.nature.com/articles/s41598-023-46349-9, learning hyperbolic representations https://openreview.net/pdf?id=yqPnIRhHtZv, simplicial complexes https://proceedings.mlr.press/v198/yang22a/yang22a.pdf
We also showed that we can detect and exploit the multifractal properties for system optimization in https://www.nature.com/articles/s44172-023-00127-7

## Mon March 18, 2024

1. Relationship between Tensor Programs/Greg Yang/optimal hyperparameters at scale and scaling laws?
2. Topic: use Yuhai's 2 metrics from Nature paper to predict parameters of BNSL?
   Activity–weight duality in feed-forward neural networks reveals two co-determinants for generalization, Y Feng, W Zhang, Y Tu, Nature Machine Intelligence 5 (8), 908-918

3. Different settings of different levels of data "complexity"/"information content" and simulated environments with controllable complexity (e.g. sparsity, or matrix rank)
4. Catch22 multivariate scaling laws?
5. Predicting scaling from Yuhai's 2 metrics?
6. Multiple measures of complexity of data (information content) as scaling predictors
7. Different simulated data with complexity control (sparsity, matrix rank)
8. Greg Yang's mup and scaling
9. Different architecture scaling and grokking phenomena with SSMs?

## Mon Nov 06., 2023

Q1: how does the complexity of a problem (e.g., functional form) affects the amount of data necessary to achieve grokking

Q2: what measure of task complexity should we use?

15:57:56 From Darsh Kaushik To Everyone: can multitask learning help induce grokking in harder tasks (ones for which the openAI paper couldn't grok)?

Q3: same questions can be asked in the context of pretraining on a task of some complexity, with testing on the same task, versus pretraining on some fixed data but then applying the model in few shot way on different downstream tasks of varying complexity?

you said fraction of data gets smaller as p is increased (in mod p), is this true for p prime too?

The Riemann Hypothesis 🙂

Join discord: https://discord.gg/rUCguaW2
Go to discussion in #p11_grokking channel :
https://discord.com/channels/881148623155499038/1093192107331702874

Mon Oct 2
Open questions:

1. In general, we are interested in transitions in performance or other behaviors/metric, as a function of either compute (grokking), or data, or model size (or a combination of those). In this case, the last 3 vars are x-axis/"critical parameter" (a la temperature for solid to liquid etc, or probability of edge in a random graph, with y axes probability of being connected, etc etc).
2. Andrei: Transitions with data set size?

Topics and references:

Scaling & Emergence workshop - see talks by Andrey, Neel, Pascal et al:

Andrey Gromov: A solvable model for grokking modular arithmetic video, paper: Grokking modular arithmetic

Jr. Tikeng Notsawo (University of Montreal/Mila): Is grokking predictable? video
paper: Predicting Grokking Long Before it Happens: A look into the loss landscape of models which grok

Eric Michaud (MIT): The Quantization Model of Neural Scaling (remote talk) video
paper: The Quantization Model of Neural Scaling Interview with Eric AI Alignment forum

11:40 - 12:10 **Neel Nanda** (Google DeepMind): **Progress measures for grokking via mechanistic interpretability** (remote talk) **video**

**Rylan Schaeffer** (Stanford): **Are emergent abilities of Large Language Models a mirage?** **Video**
paper: **Are Emergent Abilities of Large Language Models a Mirage?**

1. Yasaman's paper: Explaining Scaling Laws of Neural Network Generalization
2. Dan's paper: A Solvable Model of Neural Scaling Laws
3. We saw non-monotonicity of test loss in grokking (MLP with MSE loss).
4. Fine-tuning scaling laws
5. Recent paper on grokking theory: https://x.com/VikrantVarma_/status/1699823229307699305?s=20

6. Yuhai's paper: Activity-weight duality in feed forward neural networks reveals two co-determinants for generalization

Wed Sept 20
To do:

Pascal, Irina et al: Survey on Grokking
Papers to read and discuss in the following weeks?
Presentations/speakers? Pascal? Eric Michaud? Others from ICML workshop? 5th Neural Scaling Laws workshop on Emergence and Phase Transitions /

**Open questions to Andrey:**

1. (by Vincent)You claim that his analytic solution for the weights of a net trained on modular addition works for other activations than the quadratic function. Can you elaborate on that?
2. Have you found a similar exact analytic solution for other groups?
3. Distinction between random features (NTK) and feature learning
4.

**General Open Questions**

1. Can we better understand/derive the parameters of scaling laws, especially less trivial cases such as Broken Neural Scaling Laws (e.g. a la Kaplan's attempt to characterize power law exponent using data manifold dimension: Scaling Laws from the Data Manifold Dimension). Like in physics, we would like to link the internal measurable properties of the data, the network training dynamics, etc to the different parameters such as power law exponents, the number of breaks, the locations of breaks, their sharpness, etc.
2. Can we combine "closed box" approaches to predicting neural net behaviors (performance like loss, accuracy, various metrics of downstream task performance, as well as alignment-related behaviors like truthfulness etc) such as classical scaling laws (power laws, and more generally smoothly broken power laws in BNSL) with "open-box" approaches - e.g. Pascal's paper on spectral properties of the training loss being predictive about the occurrence of grokking later in the training run, and many others that teams from MIT (Max Tegmark's lab), UC Berkeley (Neel Nanda et al) and others were investigating.

## Related work

* Benign Overfitting and Grokking in ReLU Networks for XOR Cluster Data
* Droplets of Good Representations: Grokking as a First Order Phase Transition in Two Layer Networks
* Grokking as Compression: A Nonlinear Complexity Perspective
* Grokking as the Transition from Lazy to Rich Training Dynamics
* GROKKING TICKETS: LOTTERY TICKETS ACCELERATE GROKKING