Sparse Fourier Regularization for Modular Arithmetic Models

Amogh Dalal, Adityasinh Rathod, Akshay Rangamani

Department of Data Science, New Jersey Institute of Technology, Newark, NJ, USA

Background/Motivation

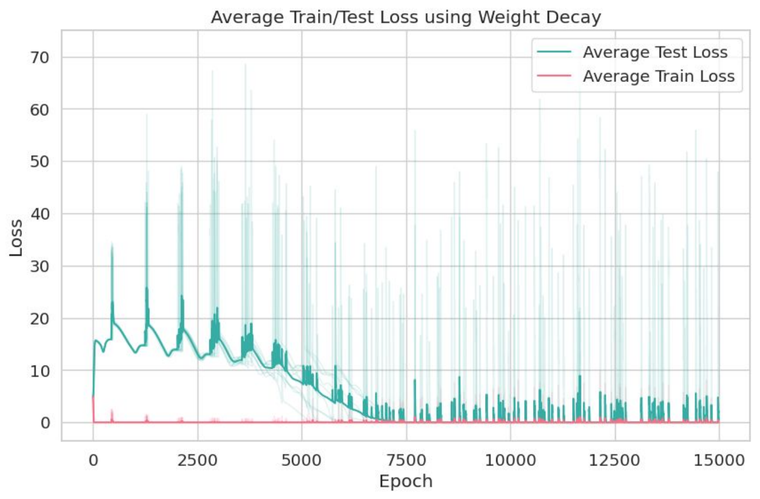
- **How do AI models learn useful abstractions?** Prior work in Mechanistic Interpretability has shown that models can learn interesting representations and circuits for various kinds of "dummy" problems.
- Using the task of **modular arithmetic**, we train sequence models on single and multiple group operations to study the learned circuits.
- We introduce l_1 regularization in the Fourier space of the (un)embedding layers to bypass grokking, letting models train up to 3× faster
- We also study how l_1 Fourier regularization can be used to perform targeted inventions on models trained on multiple operations across groups (ex. addition & multiplication), disentangling the embedding geometry across circuits.

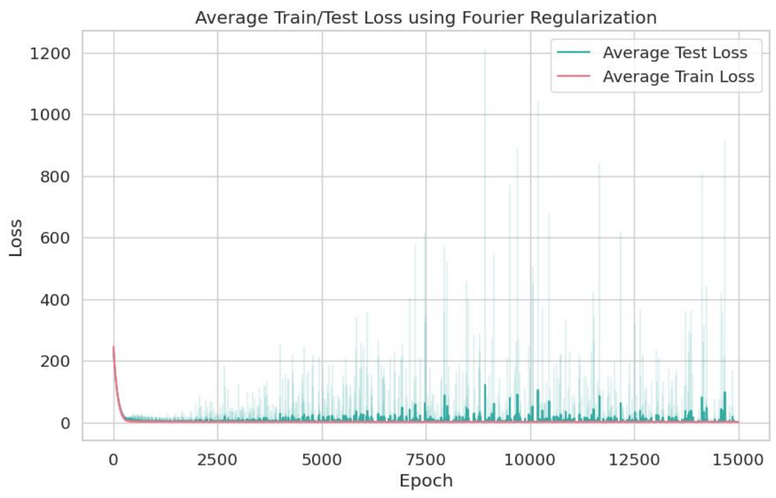
Results – Fourier Regularization

$$\mathcal{L}_F = \lambda_F \left(\|F_p W_E\|_1 + \|F_p W_{fc}\|_1 \right)$$

l_1 Fourier Regularization

- **~3× Faster convergence** on modular arithmetic tasks
- Applied to **Embedding** and **Unembedding** of sequence models
- Group-regularization can **disentangle circuits** across groups (addition & subtraction)



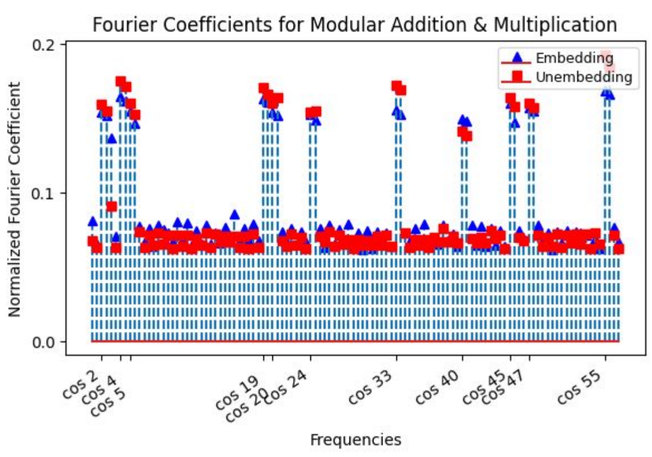


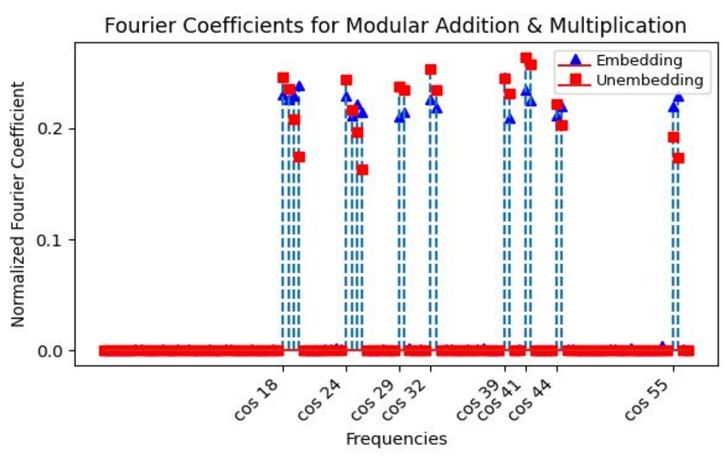
Results – Transformer Circuits Across Group Operations

Weight Decay (Normal)

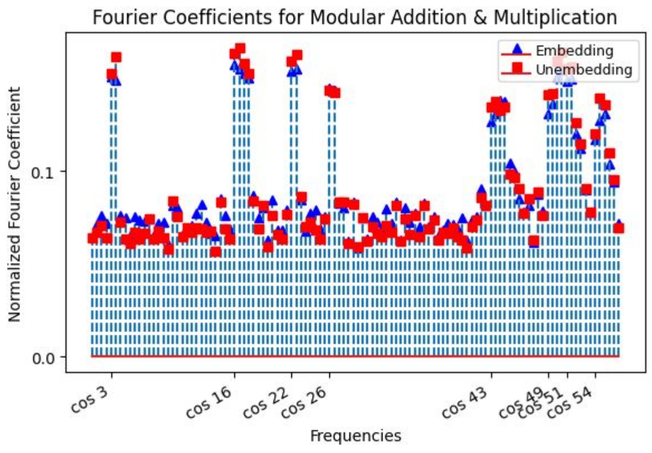
Group l_1 Fourier Regularization

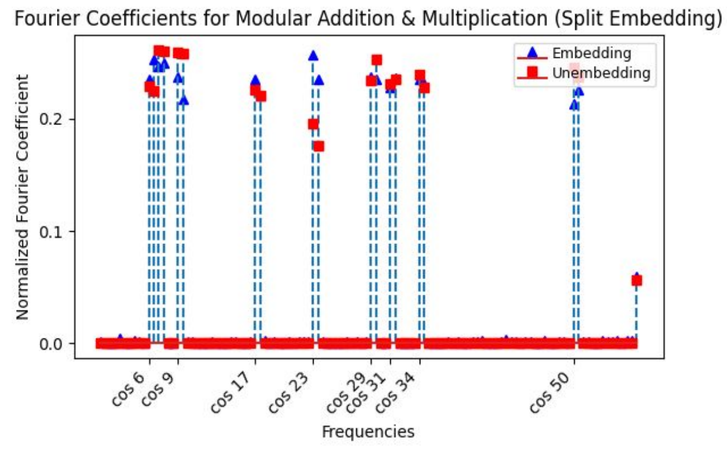
Natural Token Ordering





Discrete-log permuted



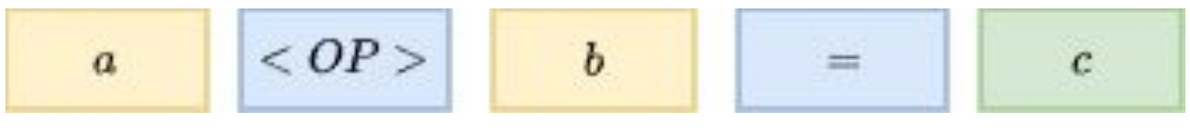


Addition & Multiplication models learn **overlapping circuits**

Group-sparse Fourier Regularization disentangles the embedding geometry

- Addition/Subtraction (same group) utilize one shared circuit
- Addition/Multiplication (different group) utilize different circuits, with overlapping frequencies
  - Addition model embeddings have sparse fourier spectra
  - Multiplication model embeddings have sparse discrete-log permuted fourier spectra
- Different group circuits exhibit a "noise floor" instead of perfect sparsity, indicating entanglement in the two fourier bases

Experimental Setup



- We train single layer RNNs and Transformers on the following single and multi-op modular arithmetic tasks:
  - **Addition - Subtraction - Multiplication - Division**
- The model learns to predict the value c = a <OP> b (mod p), for varying p = 79, 113, 151
- Each token is passed through a learned embedding, and into a hidden state size of d_h = 256, 8 heads, 1024 MLP width for Transformer. The outputs are passed through an embedding to classify the final output.
- We train for 15-30k epochs with a learning rate of 0.1 and various weight decays of 10^{-5} to 10^{-3}.

Discussion & Future Work

- Identifying circuits can help efficient model editing, turning implicitly learned circuits into explicitly learned algorithms
- Faster/efficient training of models by identifying the right biases, for harder problems involving multiple circuits and opaque representations
- Using known structure of circuits and embedding geometry on easier tasks can assist in training more interpretable models on harder tasks

CODE

Contact: {asd33, ar2575, akshay.rangamani}@njit.edu

Mechanistic Interpretability Workshop, ICML 2026

PAPER