Eigendirections of the empirical NTK recover learned features in trained models

PRINCIPLES of INTELLIGENCE

Feature Identification via the Empirical NTK

Jennifer Lin arXiv:2510.00468

1 TL;DR

- We provide evidence that **eigendirections of the empirical NTK track learned features** in trained models.
- In models trained on modular arithmetic, **top eNTK eigenspaces align with Fourier features** used by the model to implement known ground-truth algorithms.
- In the language model Gemma-3-270M, **eNTK eigenanalysis beats a same-budget PCA baseline at surfacing grammar features**.

2 Background

- Motivated by deep learning theory, we test the hypothesis that top eigendirections of the empirical NTK (eNTK)

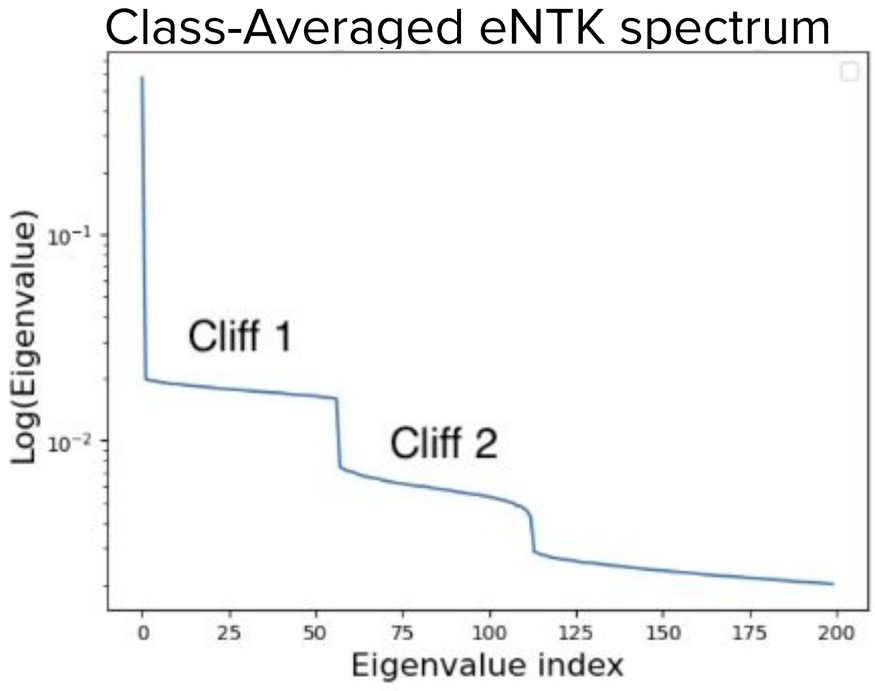
$$K_{ij}(x_1, x_2) = \sum_{\mu} \frac{df_i(x_1)}{dW_\mu} \frac{df_j(x_2)}{dW_\mu}$$

  align with feature directions in trained models. Here $f_i(x)$ is the model's ith output logit on input x, and W_μ runs over the model's parameters.

- On a size N dataset, eigenvectors of the eNTK are length N vectors that assign a number to each data point. We can therefore say that on a test set, an eNTK eigenvector represents a certain feature if whenever a test point exhibits that feature, the eigenvector assigns a high number to that test point and vice versa.

3 Modular Arithmetic results

- On an MLP trained on the task (a,b) ➔ a+b mod p, the eNTK eigenspectrum contains two cliffs.
- **Cliff 1 aligns with the a/b Fourier features at all distinct frequencies and Cliff 2 with the sum/difference Fourier features** used by the model to compute the known ground-truth solution.



- The first derivative of Cliff 2's overlap with the sum modes **peaks at the onset of grokking.**
- A Transformer trained on modular arithmetic exhibits similar results.

4 Language model results

- For Gemma-3-270M on a test set of 1024 length-64 context windows from TinyStories, we computed the top 25 eNTK eigendirections at each layer and compared with the top 25 PCA directions on model activations at each layer.
- Across 14 independently-generated grammar labels for the model-predicted next word after each context window (e.g. whether the next word was a proper noun or a past tense verb), **the best eNTK eigendirection computed had a higher AUROC score than the best PCA direction on 13/14 grammar labels.**

Limitations: Our largest experiments were nonetheless on a relatively small model and dataset.
Future work: Scale to larger models, datasets and feature sets; compare with SAEs; and theoretically analyze when and why eNTK eigendirections correspond to interpretable features.

Mechanistic Interpretability Workshop, ICML 2026
jylin04@gmail.com