Loss Landscape Geometry Reveals Stagewise Development of Transformers

George Wang george@timaeus.co Timaeus

Matthew Farrugia-Roberts matthew@timaeus.co Timaeus

Jesse Hoogland jesse@timaeus.co Timaeus

Liam Carroll lemmykc@gmail.com Independent

Susan Wei susan.wei@unimelb.edu.au The University of Melbourne

Daniel Murfet d.murfet@unimelb.edu.au The University of Melbourne

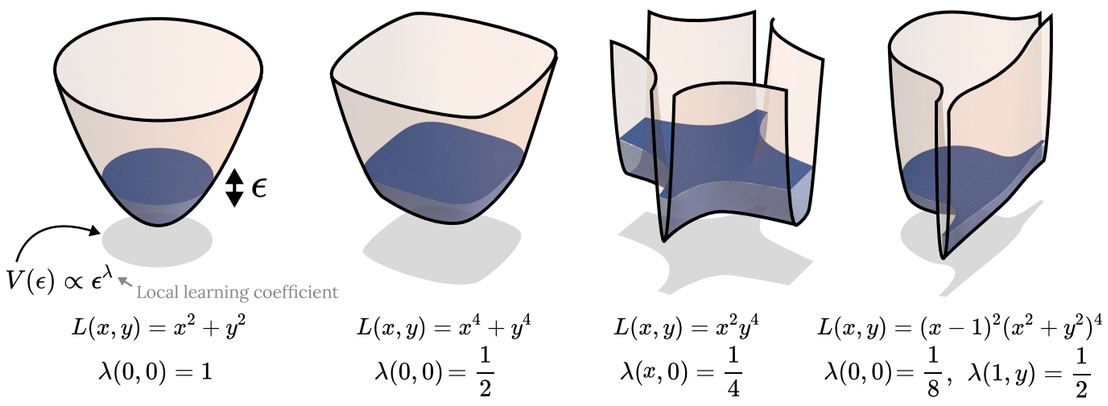
Overview

Problem: How does stagewise development arise in NNs?

Hypothesis: We propose that the local geometry of the population loss holds the key to understanding stagewise development in neural networks. This relates neural network development to **singular learning theory (SLT).**

Approach:

1. We **track the geometry** of the loss landscape over the course of training for 2-layer attention-only language transformers **using the local learning coefficient (LLC)**, a principled measure of *model complexity* and *geometric degeneracy*.
2. We automatically **divide the learning process** into distinct developmental stages **by locating critical points** in the LLC curve.
3. We **validate developmental stages** with a variety of **behavioral & structural metrics.** Interpretable changes in these metrics coincide with stages identified by the LLC.



The loss landscape is highly *degenerate*: changing weights does not mean changing the loss.

1. Tracking Geometry

The LLC measures *geometric degeneracy* of the population loss ℓ at a local minimum w^* and can be estimated using the following formula [1, 2, 3]:

$$\hat{\lambda}(w^*) = n\beta\Big(\mathbb{E}_{w|D_n,w^*,\gamma,\beta}[\ell_n(w)] - \ell_n(w^*)\Big),$$

(n: # of samples; D_n: Dataset; $\ell_n(w)$: Empirical loss)

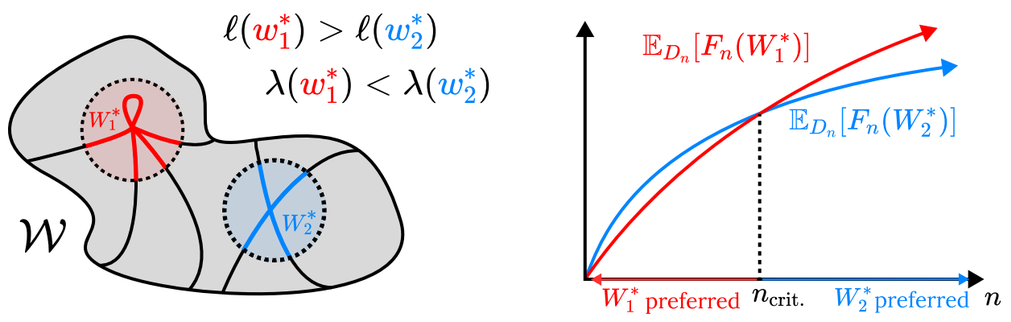
where $\mathbb{E}$ is over the *tempered, localized* Gibbs posterior,

$$p(w; w^*, \beta, \gamma) \propto \exp\Big\{-n\beta\ell_n(w) - \frac{\gamma}{2}\|w - w^*\|_2^2\Big\}.$$

(β: Inverse temperature; γ: Localization strength)

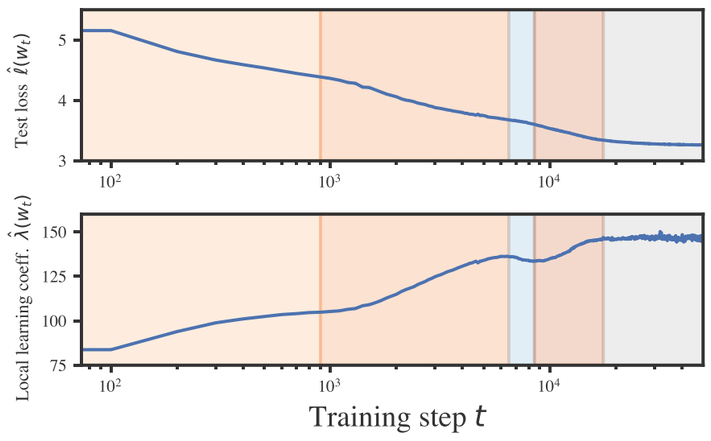
Singular learning theory predicts stagewise learning: Bayesian inference is equivalent to minimizing the *free energy* F_n over regions of parameter space $W^* \subseteq \mathcal{W}$, which involves a **data-dependent tradeoff between loss and complexity** [1]:

$$\mathbb{E}_{D_n}[F_n(W^*)] = n\ell(w^*) + \lambda(w^*)\log n + O(\log\log n).$$



2. Dividing Stages

By looking for plateaus in the LLC with respect to (log) training time, we automatically identify candidate stage boundaries.



Key	Stage	End t	$\Delta\hat{\ell}$	$\Delta\hat{\lambda}$
	LM1	900	−2.33	+26.4
	LM2	6.5k	−1.22	+22.5
	LM3	8.5k	−0.18	−1.57
	LM4	17k	−0.40	+8.62
	LM5	50k	−0.34	+1.77

Hidden stagewise development in 2-layer transformers: Even when the loss decreases smoothly, the LLC can reveal a hidden succession of stages separated by critical points.

3. Validating Stages

We confirm the validity of these stages by measuring various **behavioral metrics** that measure changes in input-output behavior, such as the *bigram score* (cross entropy between model predictions and the empirical bigram distribution) and the *n-gram score* (average final loss on a set of n-grams).

We also track **structural metrics** that measure changes in weights or activations, including the *previous token score*, *prefix-matching score*, and *ICL score* introduced in [4].

(a) LM1 (0 - 900)

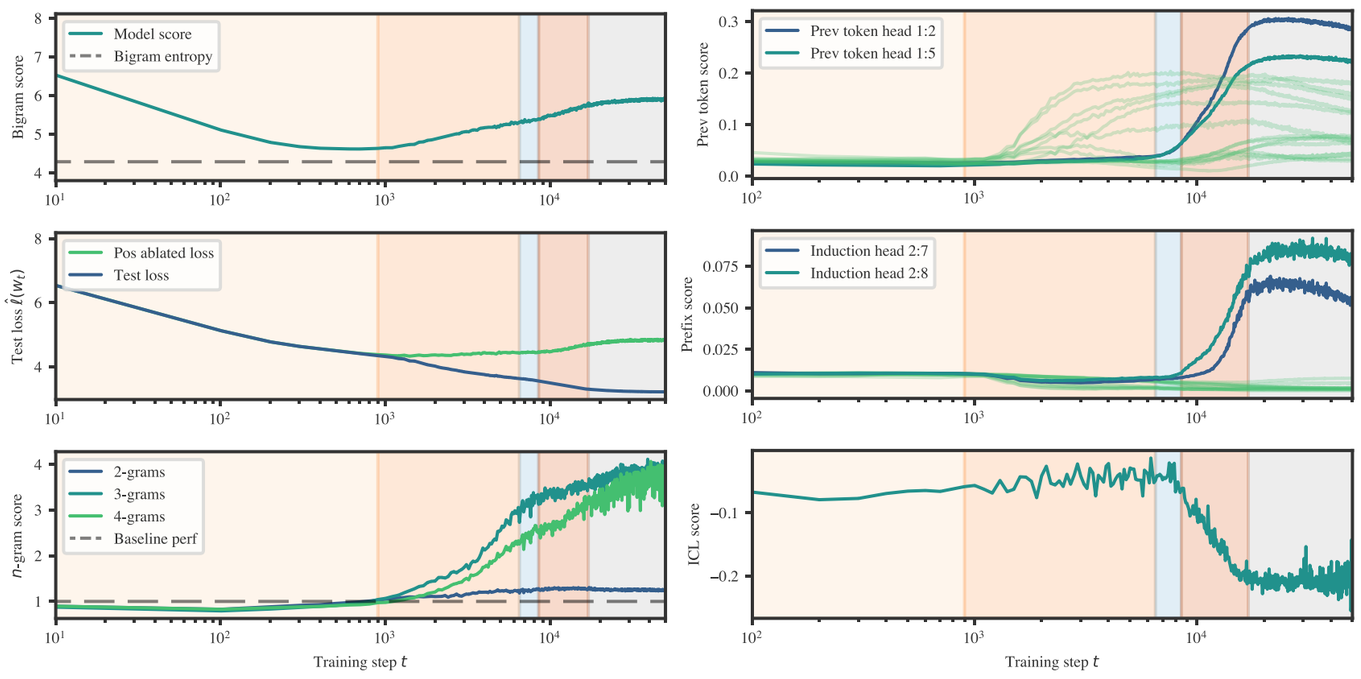
<lendoftext|>I should like, before proceeding further, to tell you how I feel about the State which we have described. I might compare myself to a person who, on beholding beautiful animals either created by the painter's art, or, better still, alive but at rest, is seized with a desire of seeing them in motion or engaged in some struggle or conflict to which their forms appear suited;

(b) LM2 (900 - 6,500)

<lendoftext|>In the midst of unexpected circumstances with Linux and Python, the honorable Supreme Court in Boston delivered a ruling emphasizing a crazy database framework last week.

(c) LM3 + LM4 (6,500 - 17,000)

<lendoftext|>Mr. and Mrs. Dursley, of number four, Privet Drive, were proud to say that they were perfectly normal, thank you very much. They were the last people you'd expect to be involved in anything strange or mysterious, because they just didn't hold with such nonsense. Mr. Dursley was the director of a firm called Grunnings, which made drills. He was a big, beefy man with hardly any neck, although he did have a very large mustache. Mrs. Dursley was thin and blonde and had nearly twice the usual amount of neck, which came in very useful as she spent so much of her time craning over garden fences, spying on the neighbors. The Dursleys had a small son called Dudley and in their opinion there was no finer boy anywhere.



Developmental stages are interpretable: Simple language models learn (LM1) bigrams, (LM2) n-grams, (LM3-4) previous-token heads, and (LM4) induction heads in order before (LM5) converging.

[1]: Sumio Watanabe. Algebraic Geometry and Statistical Learning Theory. Cambridge University Press, 2009.
[2]: Edmund Lau, Daniel Murfet, and Susan Wei. Quantifying degeneracy in singular models via the learning coefficient. Preprint arXiv:2308.12108 [stat.ML], 2023.
[3]: Zach Furman and Edmund Lau. Estimating the local learning coefficient at scale. Preprint arXiv:2402.03698 [cs.LG], 2024.
[4]: Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. Transformer Circuits Thread, 2022.

Developmental Interpretability & Singular Learning Theory

Full Paper: