Introduction to singular learning theory

Zach Furman

THE UNIVERSITY OF MELBOURNE

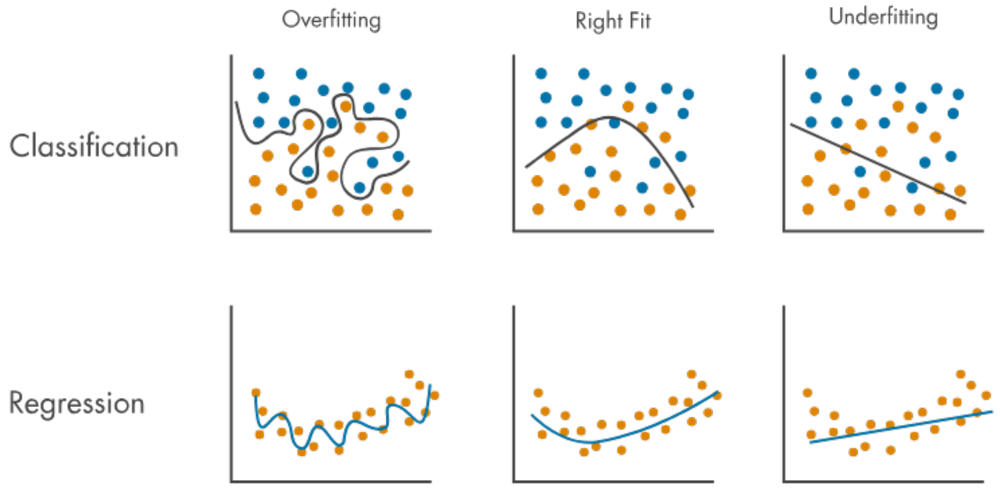
Background & Motivation

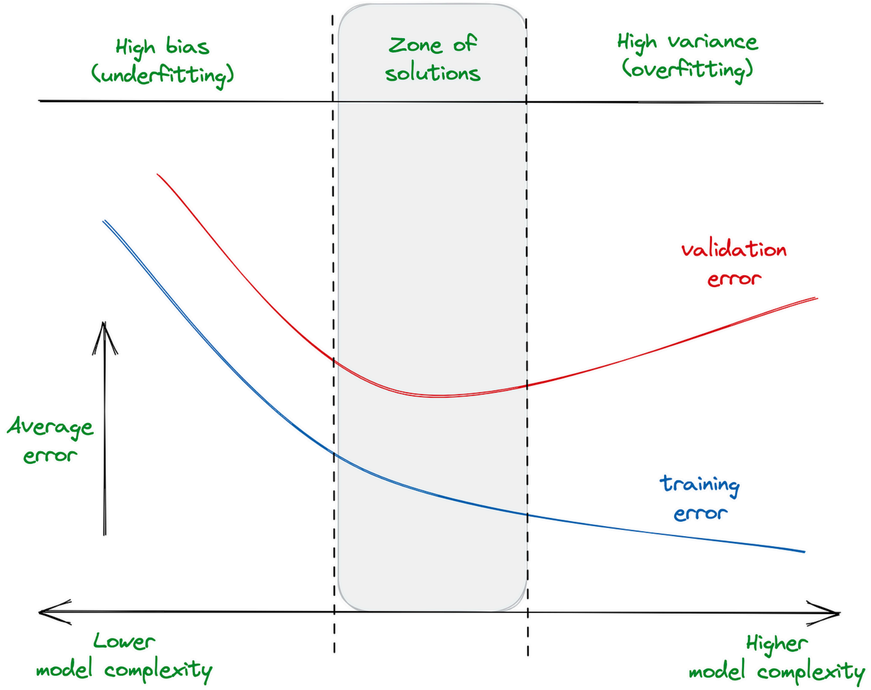
Background & Motivation

- Why does deep learning work? Can we understand it?
  - Classical statistical learning theory actively predicted that deep learning *couldn't* work
- Two mysteries:
  - Generalization
  - Learned structure

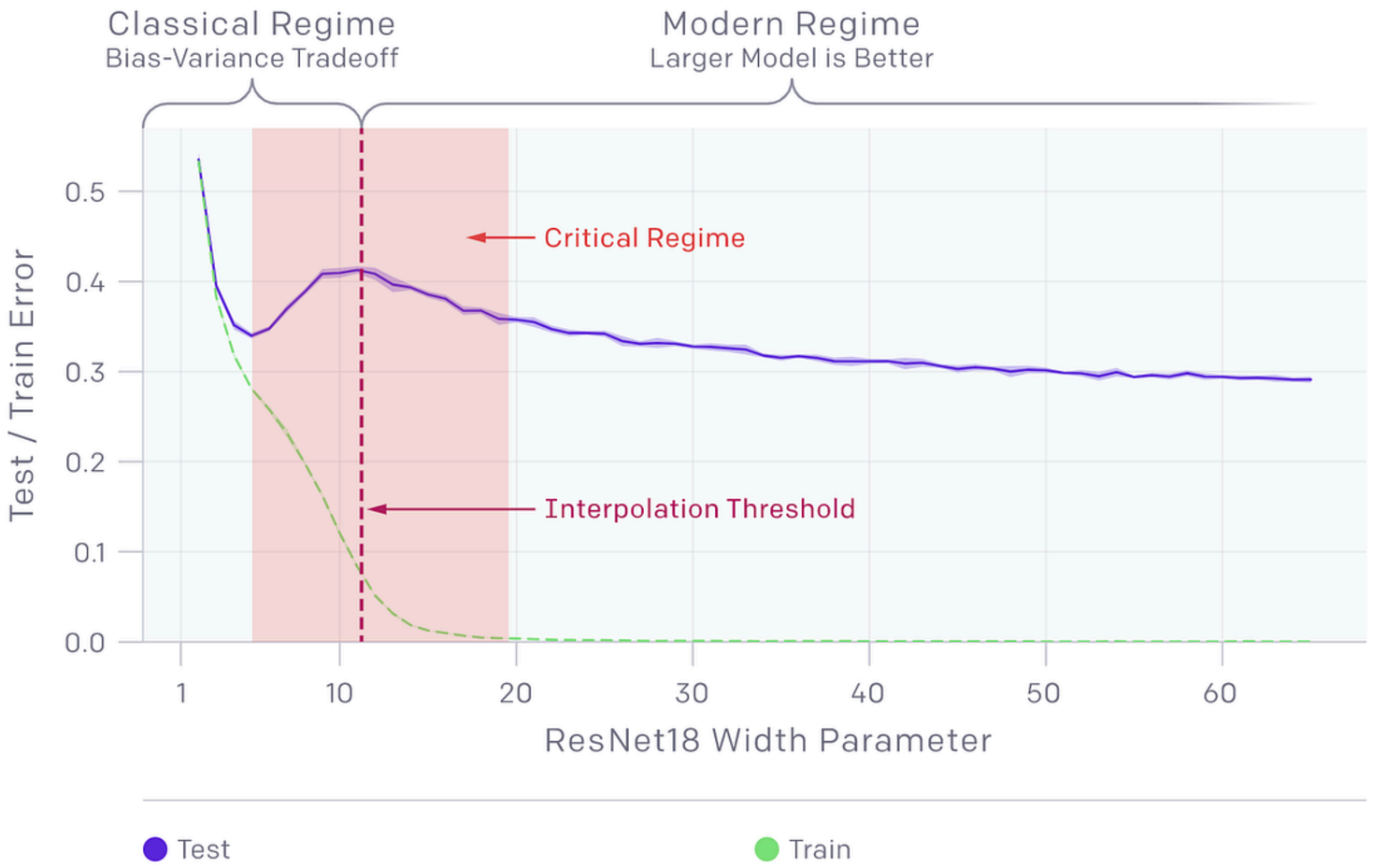
Generalization

What the classical theory said:





Generalization



Generalization

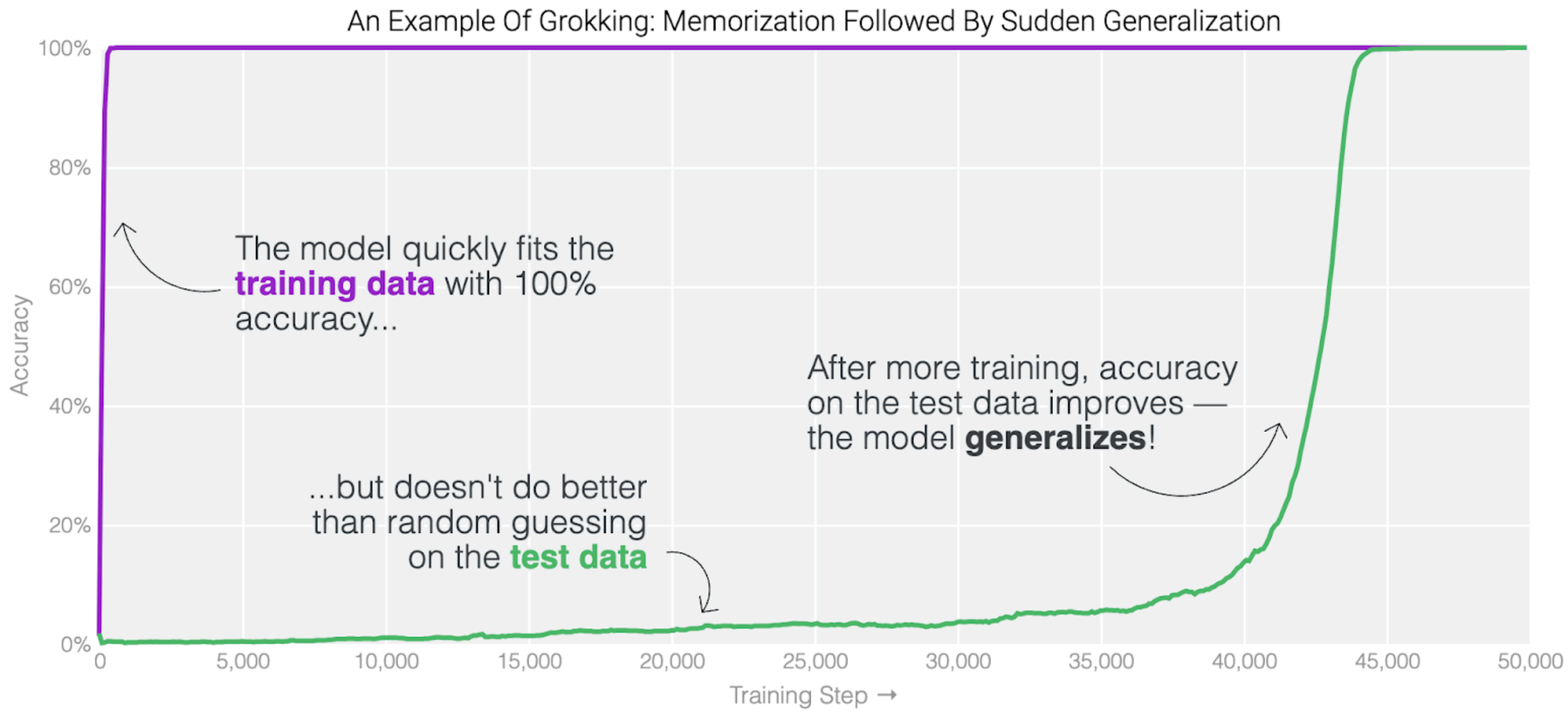
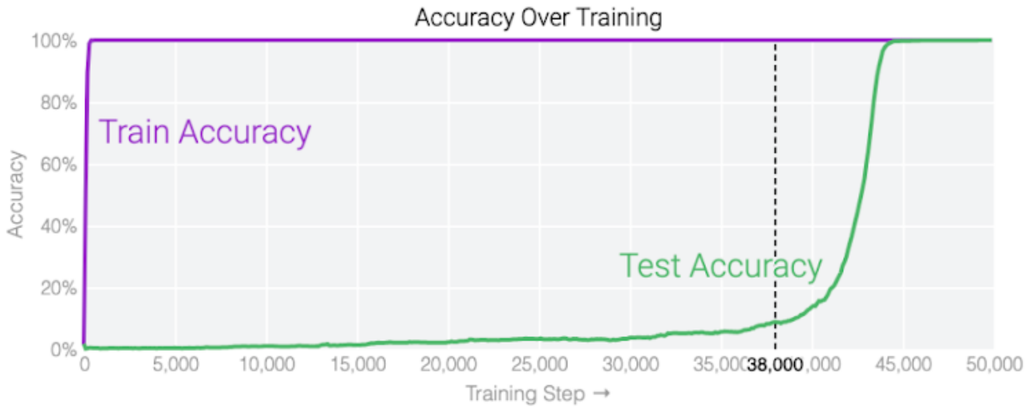
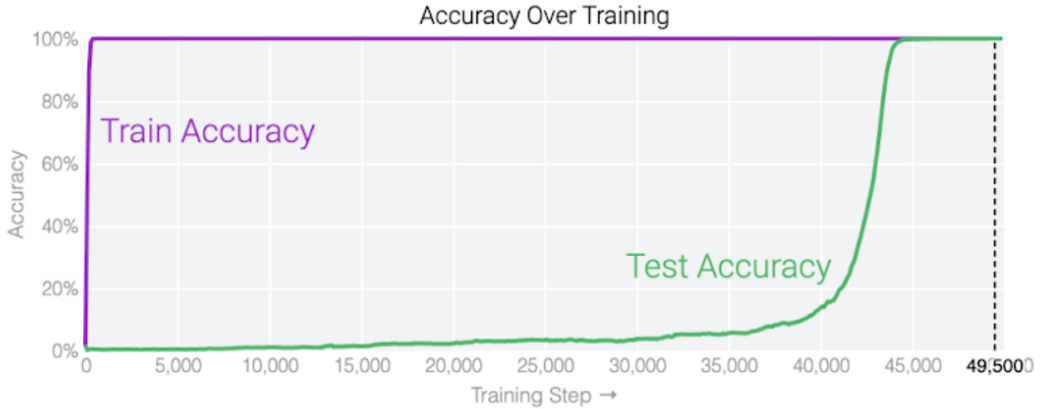
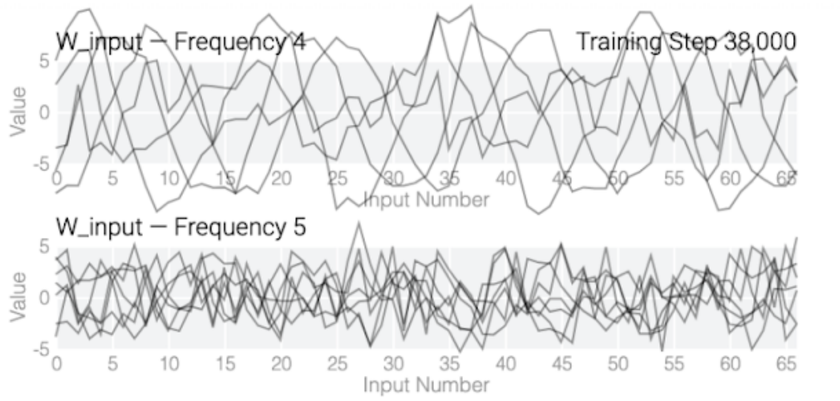


Illustration by Pearce et al. (2023)

Learned structure







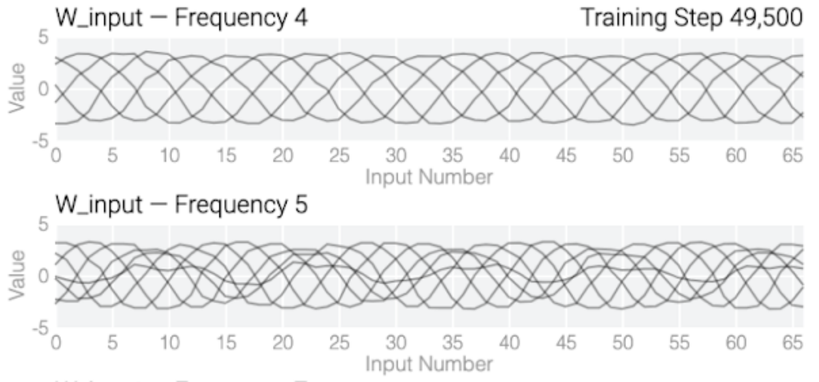
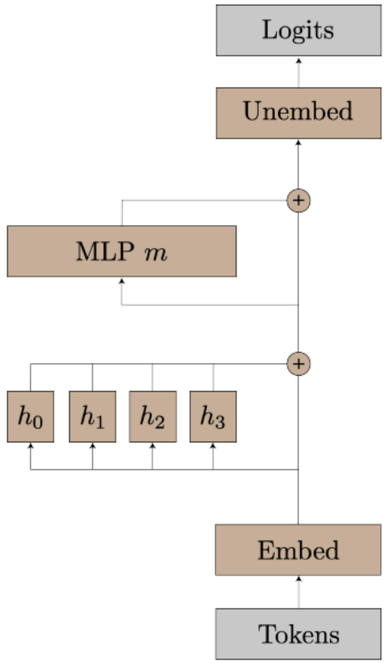


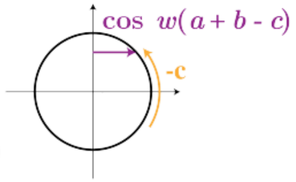
Illustration by Pearce et al. (2023)

Learned structure



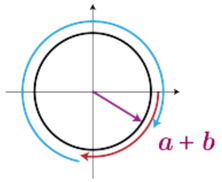
Computes logits using further trig identities:

$$\text{Logit}(c) \propto \cos(w(a+b-c))$$
$$= \cos(w(a+b))\cos(wc) + \sin(w(a+b))\sin(wc)$$



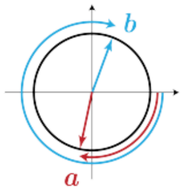
Calculates sine and cosine of $a+b$ using trig identities:

$$\sin(w(a+b)) = \sin(wa)\cos(wb) + \cos(wa)\sin(wb)$$
$$\cos(w(a+b)) = \cos(wa)\cos(wb) - \sin(wa)\sin(wb)$$



Translates one-hot a, b to Fourier basis:

$$a \to \sin(wa), \cos(wa)$$
$$b \to \sin(wb), \cos(wb)$$

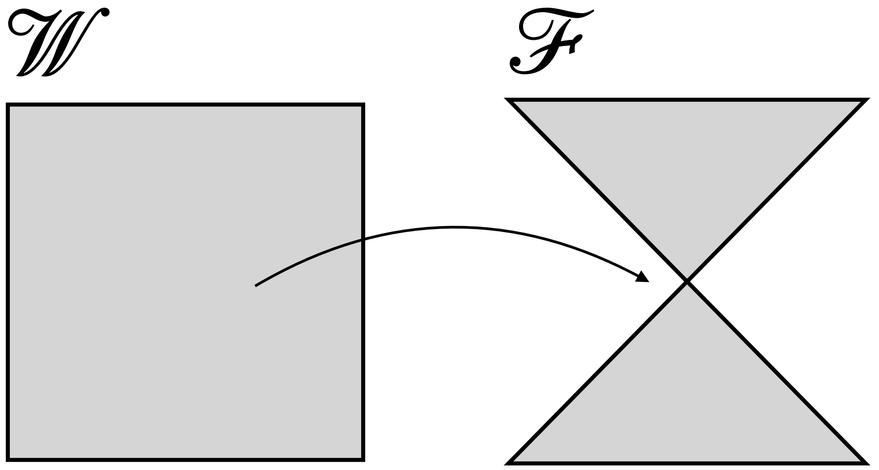


Nanda et al. (2023)

Singular learning theory: overview

Singular learning theory: overview

- Created 25 years ago by Sumio Watanabe, by dropping a single assumption from statistical learning theory
- Where did statistical learning theory go wrong?
  - It assumed **non-degeneracy**: that the parameter-to-function map is locally injective
  - When this assumption doesn't hold, we call the model *singular*, and much of classical statistical learning theory falls apart



When the parameter-function map isn't one-to-one, the choice of parameterization can't be swept under the rug.

Singular learning theory: overview

- Points in parameter space which are degenerate are called *singularities*
  - "How singular" a model is *changes depending on where you are in parameter space*
  - **More singular points act as if they have a lower parameter count and generalize better**
- In many examples, **singularities correspond with learned structure**
  - Studying singularities can potentially allow reverse-engineering a model's learned algorithm

What exactly is a singular model?

Singular model: definition

We say that a model $p(x \mid w)$ is *singular* at a parameter $w \in W$ if there exists a non-zero direction $v \in TW$ such that the directional derivative $\nabla_v p(x \mid w) = 0$ for all x in the domain of the model*. It is *regular* otherwise.

A *singular model* is a model which is singular for at least one w. It is a *regular model* otherwise.

*Equivalently, the Fisher information matrix is not full rank.

Example

Regular model: Gaussian

$$p(x \mid w) = \frac{1}{\sqrt{\pi}} e^{-(x-w)^2}$$

$$\frac{d}{dw} p(x \mid w) = \frac{1}{\sqrt{\pi}} \, (x - w) \, \mathrm{e}^{-(x-w)^2}$$

$$\left. \frac{d}{dw} \right|_{w=0} p(x \mid w) = \frac{1}{\sqrt{\pi}} \, x \, e^{-x^2}$$

Singular model: Cubically-parameterized Gaussian

$$p(x \mid w) = \frac{1}{\sqrt{\pi}} e^{-(x-w^3)^2}$$

$$\frac{d}{dw} p(x \mid w) = \frac{1}{\sqrt{\pi}} \, 6w^2 \left(x - w^3\right) \mathrm{e}^{-\left(x-w^3\right)^2}$$

$$\left. \frac{d}{dw} \right|_{w=0} p(x \mid w) = 0$$

What models are singular?

Examples of singular models:

- Layered neural networks
- Radial basis functions
- Normal mixtures
- Binomial and multinomial mixtures
- Mixtures of statistical models
- Reduced rank regressions
- Boltzmann machines
- Bayes networks
- Hidden Markov models
- Stochastic context-free grammars

In general, singular models often:

- Have hierarchical structures.
- Contain hidden variables.
- Consist of several information processing modules.
- Are designed to obtain hidden knowledge from random samples.
- Estimate probabilistic grammars.
- Are made by superposition of parametric functions.

Lists from Watanabe (2009)

Singular models break statistics

Much of parametric statistics and learning theory fails for singular models:

- Generalization error is not necessarily proportional to parameter count
- The Cramer-Rao inequality does not hold
- The BIC is incorrect
- Maximum-likelihood estimation is not efficient
- The Bayesian posterior distribution is not asymptotically Gaussian
- Their information geometry is not Riemannian

How do you deal with singular models?

The learning coefficient

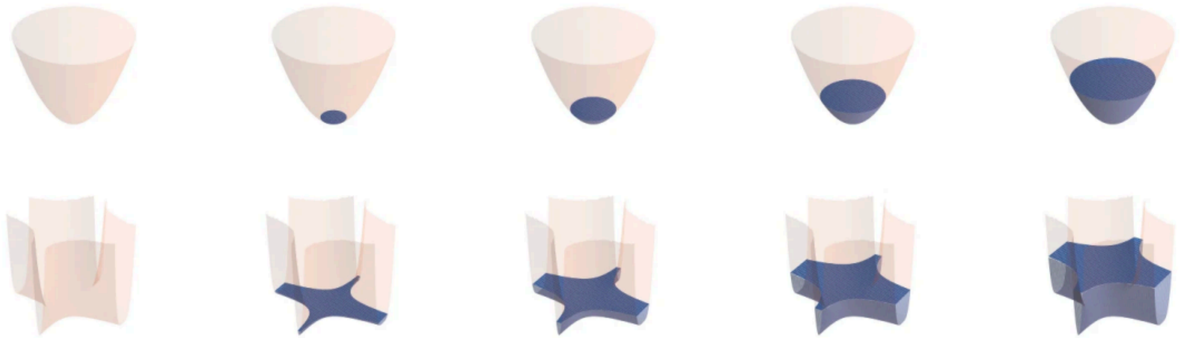
- We can quantify the "effective size" of a singular model for some parameter w with the *learning coefficient*, λ
- The learning coefficient is a geometric invariant of the loss function which captures **volume scaling**

The learning coefficient

Define $V(\epsilon)$ as the Lebesgue measure of all parameters with population loss $L(w)$ within ϵ of the minimum. Then, asymptotically as $\epsilon \to 0$, this has the following functional form:

$$V(\epsilon) \approx c\epsilon^{\lambda}$$

where c is some constant and λ is the learning coefficient.



(assuming $m = 1$)

Some remarks

The learning coefficient is a sort of **fractal dimension**: if I have a length ϵ hypercube in λ dimensions, its volume is ϵ^λ.

Regular models: $\lambda = \frac{d}{2}$ where d is the parameter count.

Singular models: $\lambda \leq \frac{d}{2}$.

Generalization in singular models

The Bayesian generalization error in a singular model is given to leading order by

$$G(n) \approx \frac{\lambda}{n} + o\left(\frac{1}{n}\right)$$

where n is the number of training samples. For singular models, the learning coefficient **replaces the parameter count** as the dominant contribution to generalization error.

Free energy and phase transitions

Free energy

The Bayesian free energy is defined as

$$F_n(W_\alpha) = -\log \int_{W_\alpha} e^{-\beta n L_n(w)}\, dw,$$

where W_α is a subset of parameter space, and $L_n(w)$ is the negative log-likelihood across the entire dataset.

Bayesian inference may be characterized as a **variational problem** of **minimizing the free energy**.

The free energy formula

Watanabe (1999)

$$F_n(W_\alpha) \approx nL_n(w_\alpha^*) + \lambda(w_\alpha^*) \log n$$

The free energy formula

Watanabe (1999)

$$F_n(W_\alpha) \approx nL_n(w^*_\alpha) + \lambda(w^*_\alpha)\ \log n$$

Tradeoff between **accuracy** (energy) and **complexity** (entropy)

The free energy formula

Watanabe (1999)

$$F_n(W_\alpha) \approx nL_n(w^*_\alpha) + \lambda(w^*_\alpha)\ \log n$$

This tradeoff **depends on the dataset size** n

Internal model selection

Suppose a model's parameter space W can be partitioned into two subspaces, W_1 and W_2. Then

$$F_n(W) = -\log(e^{-F_n(W_1)} + e^{-F_n(W_2)})$$
$$\approx \min\{F_n(W_1), F_n(W_2)\}$$

Bayesian inference **implicitly** selects between submodels via free energy minimization: it is automatically performing **model selection**.

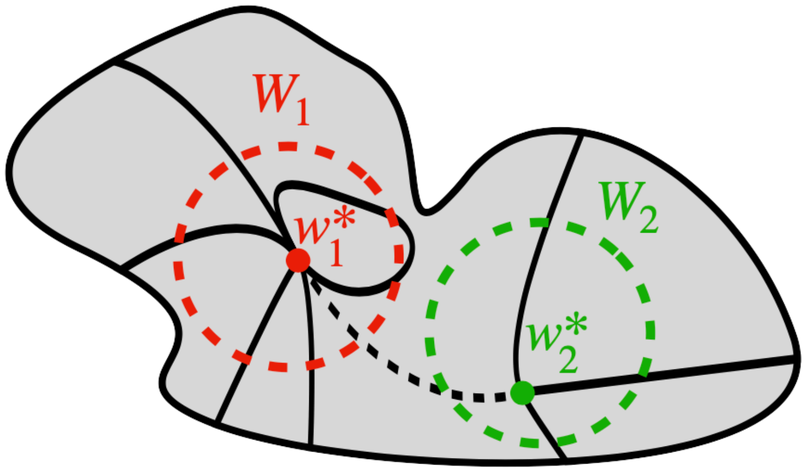
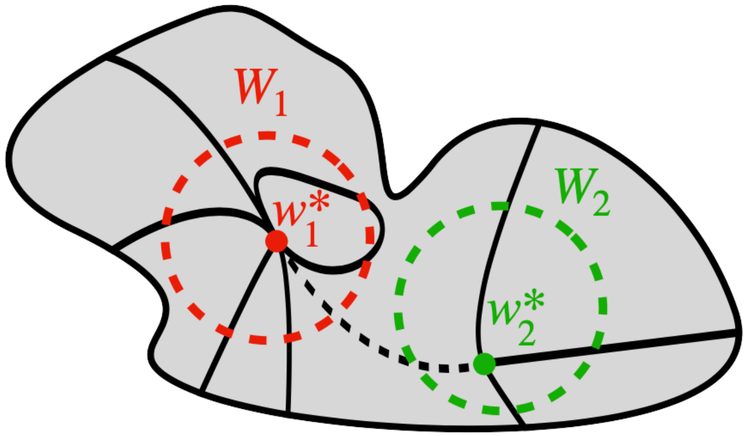


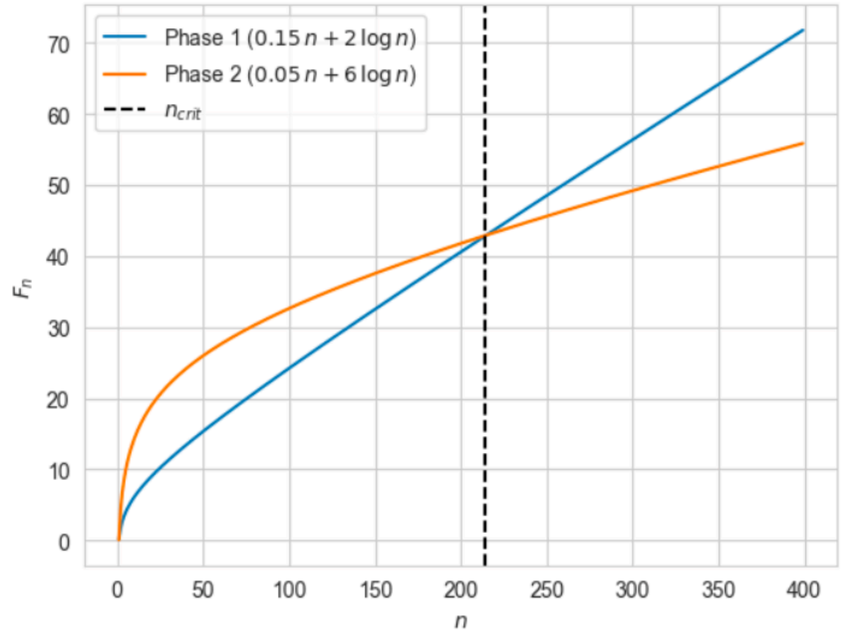
Image credit:
Jesse Hoogland

Phase transitions

$$F_n(W_\alpha) \approx nL_n(w_\alpha^*) + \lambda(w_\alpha^*)\ \log n$$

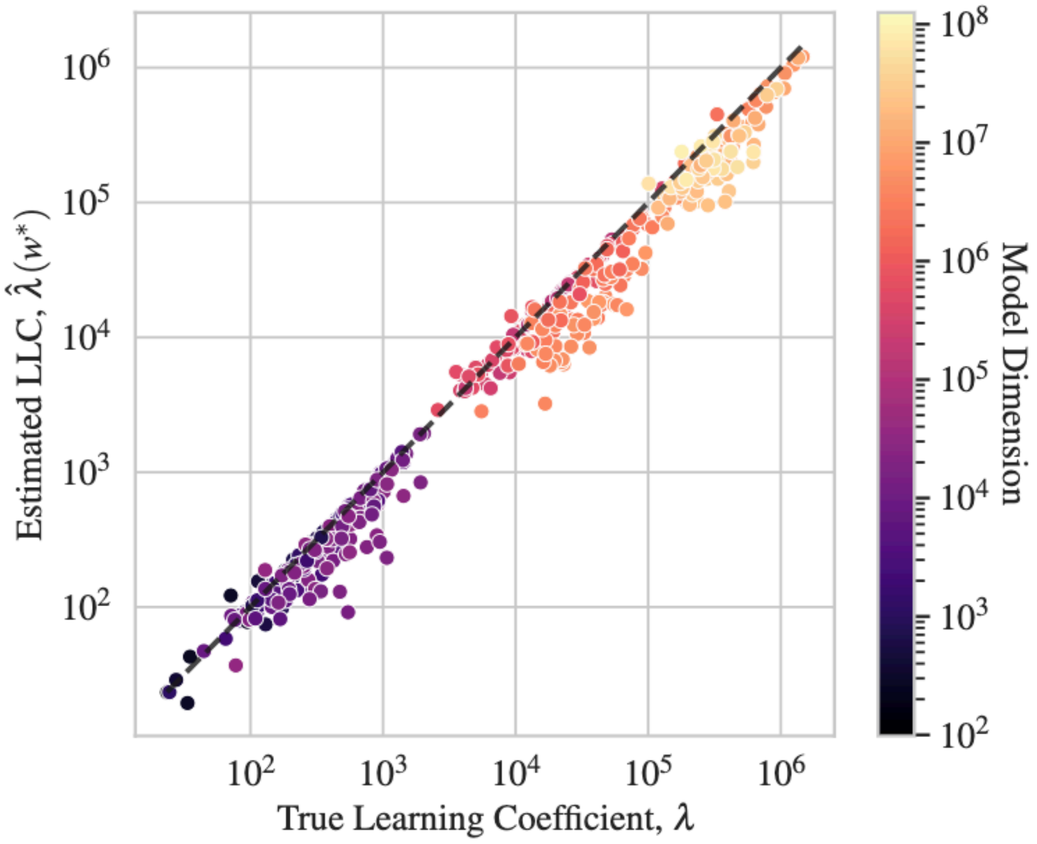
$$L_n(w_1^*) < L_n(w_2^*)$$
$$\lambda(w_1^*) > \lambda(w_2^*)$$



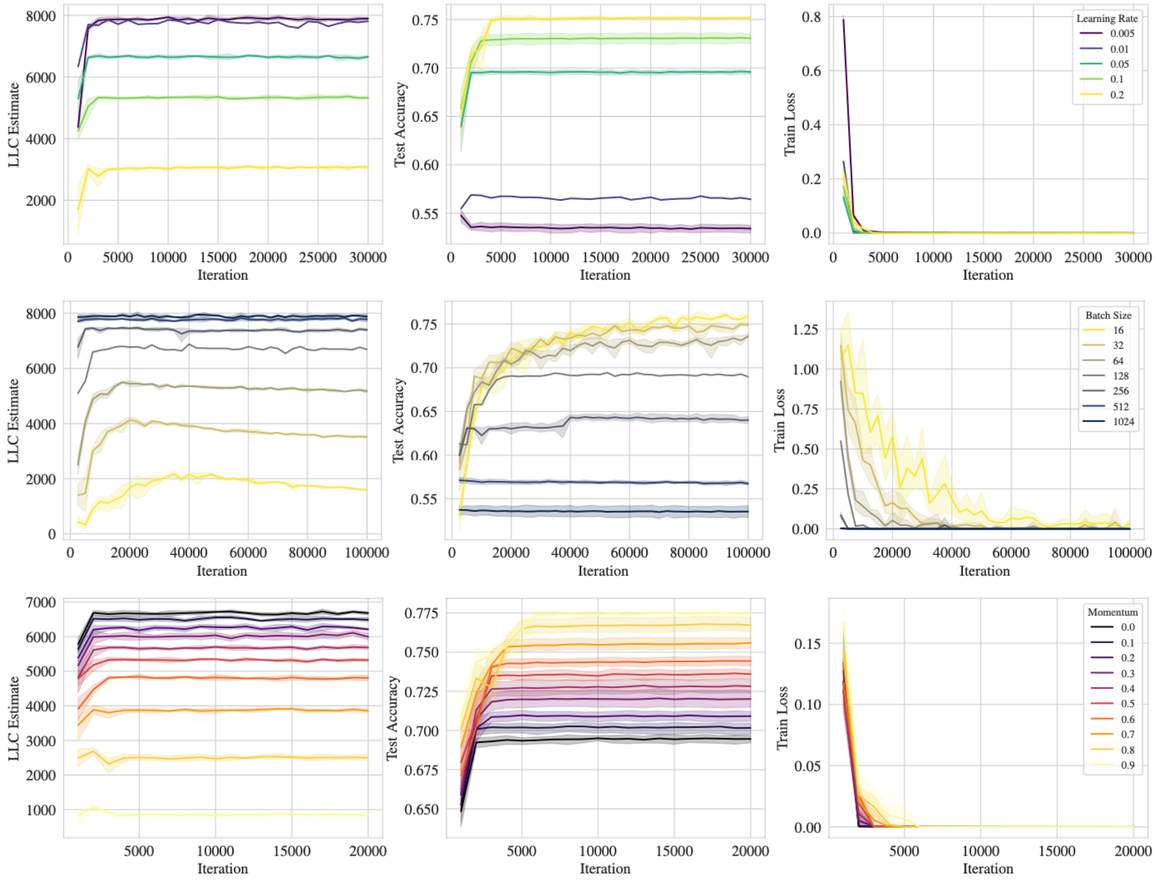


Applications and empirical validation

Estimating learning coefficients at scale



Furman & Lau (2024)

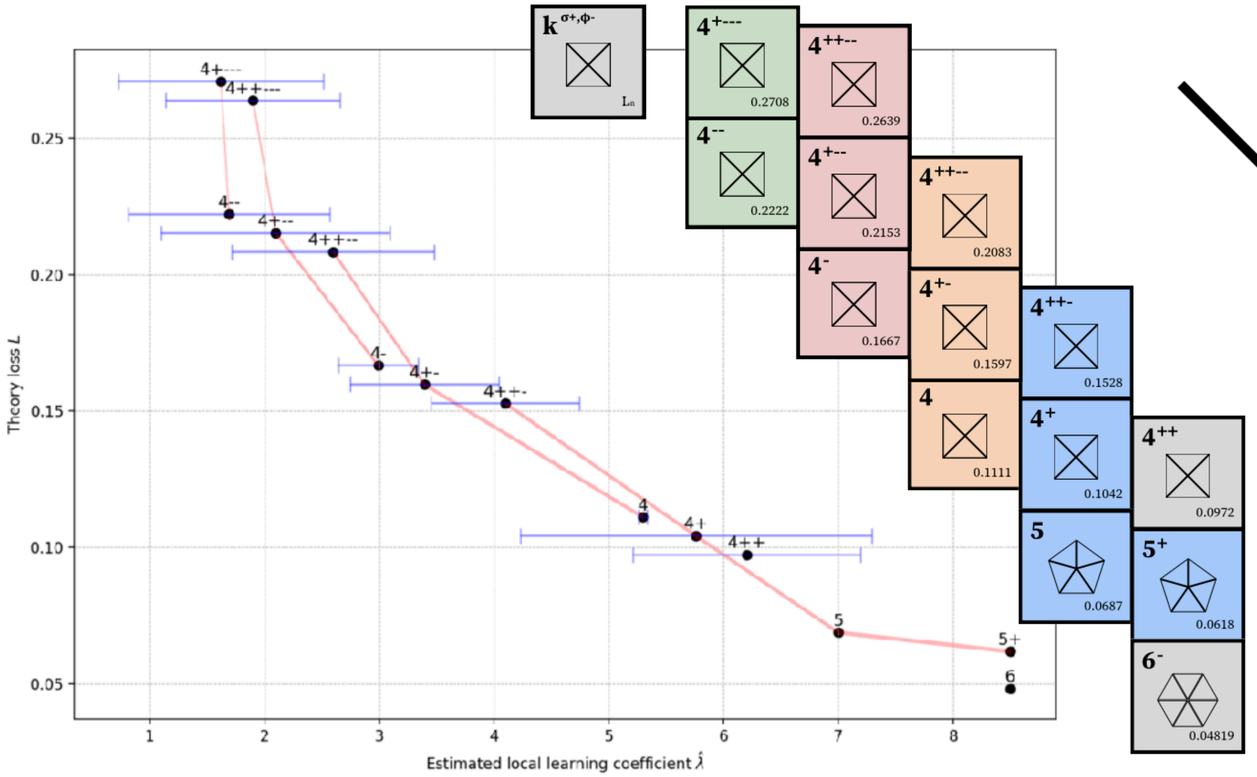


Lau et al. (2024)

A toy model of superposition

Chen et al. (2023)

Image credit: Jesse Hoogland



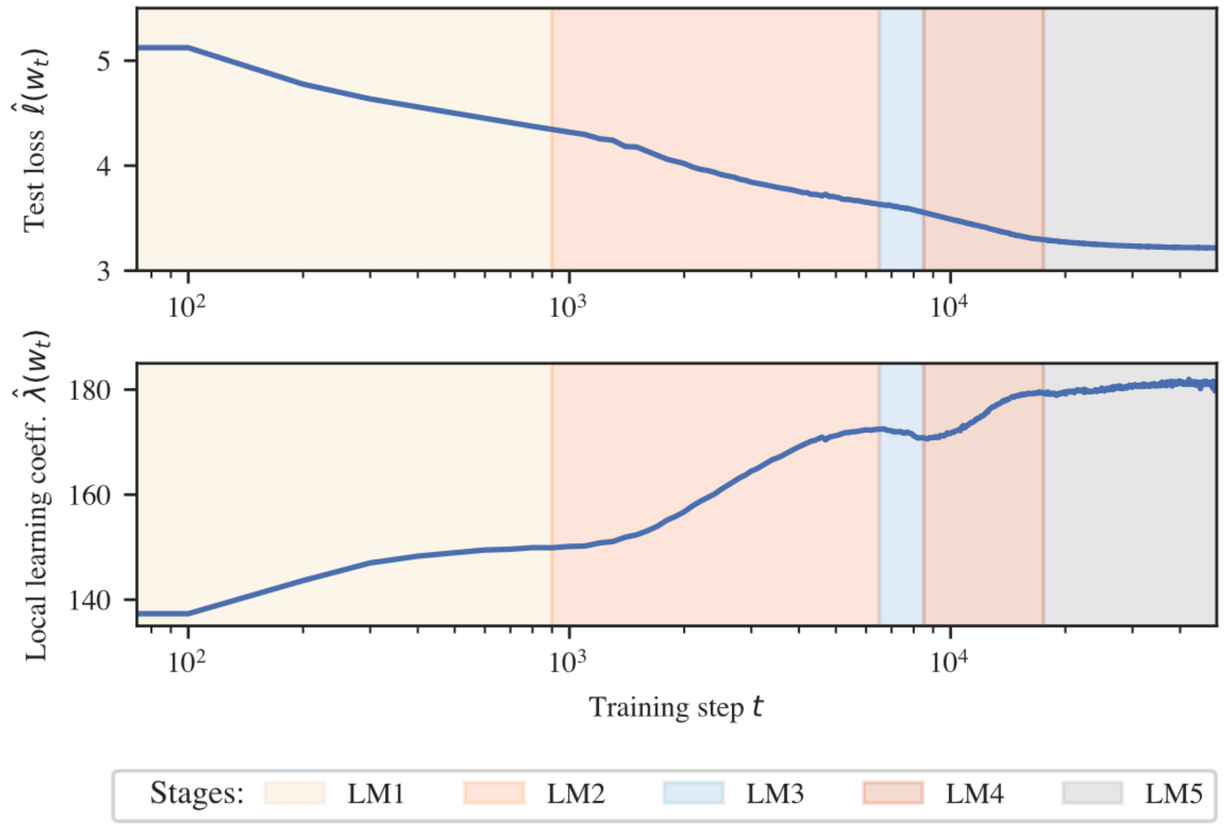
Increasing LLC

Increasing Structure

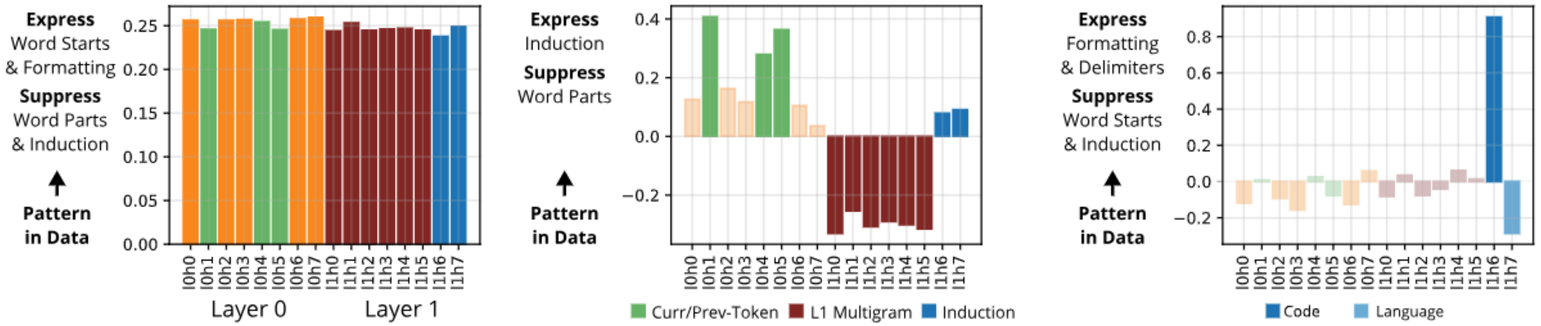
$k^{\sigma+,\phi-}$-gons: $k \leq c$ vertices, σ positive biases, and ϕ *large* negative biases.

Language models

Hoogland et al. (2024)



Beyond learning coefficients?



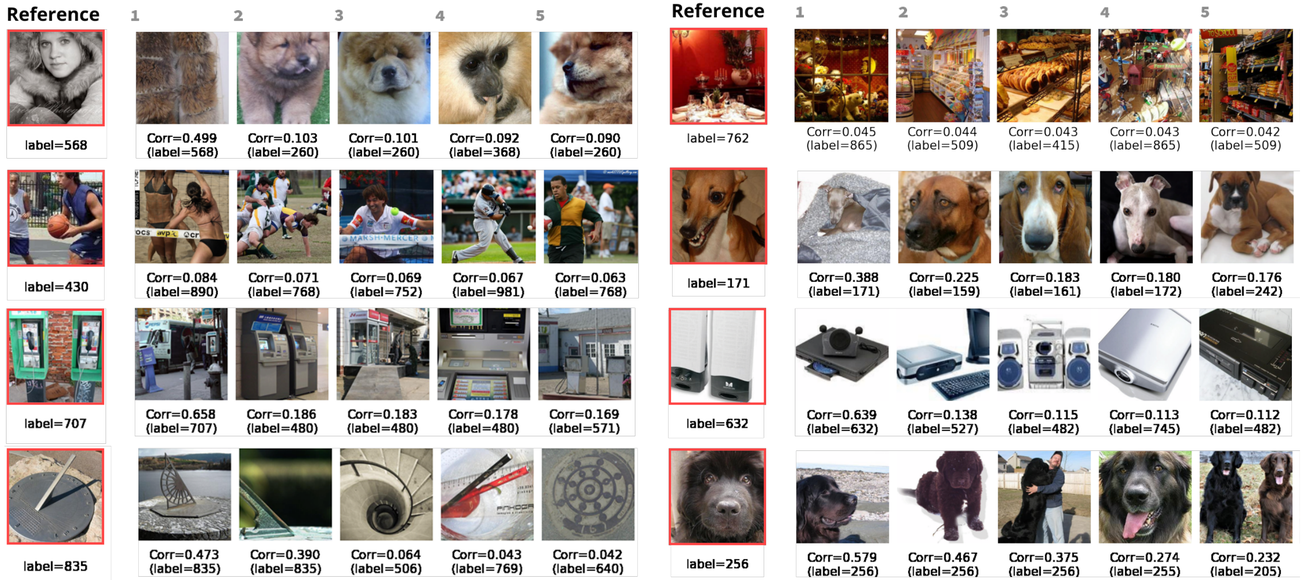
(a) **PC1: Uniform Mode**. PC1 shows that all heads learn to represent *word boundaries*.

(b) **PC2: Induction vs. Multigram.** PC2 distinguishes heads in the *induction circuit* from *L1 multigram heads*.

(b) **PC3: Induction Specialization.** PC3 distinguishes specialization between two *types of induction head*.

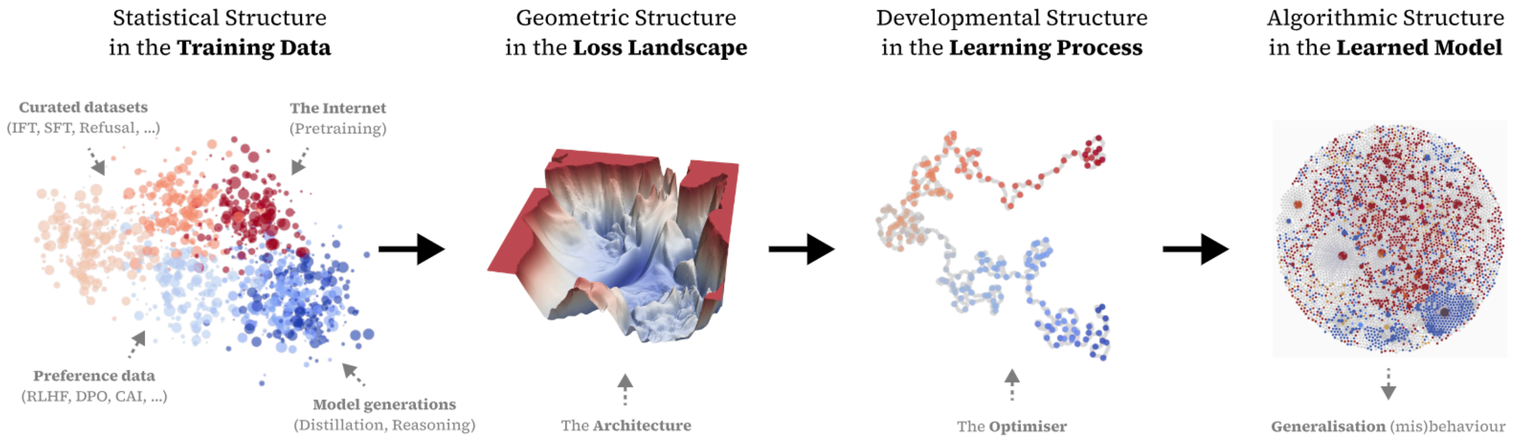
Baker et al. (2025)

Figure 4: **Susceptibilities decompose into interpretable loadings over components.** The loadings of the top three principal components for per-token susceptibility PCA on attention heads.



Adam et al. (2025)

Deeper theory?



Lehalleur et al. (2024)

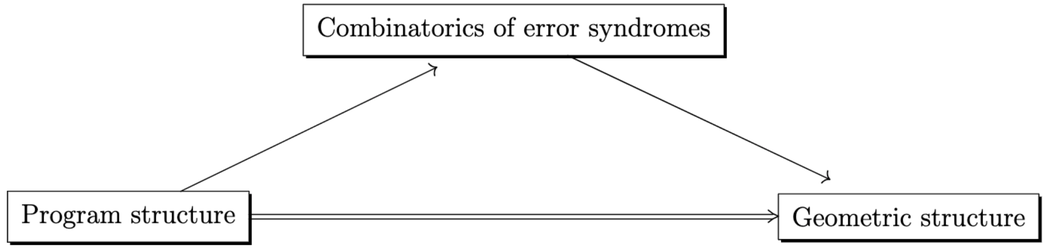


Figure 2: The relationship between program structure and geometry we establish works by relating both to the combinatorics of error syndromes, which are patterns of flips in bits on the description tape of a UTM which affect the output of the simulated machine.

Murfet & Troiani (2025)

More information

- **Learn more about SLT**: devinterp.com/resources
- **Read the latest papers**: devinterp.com/publications
- **Join the Discord**: devinterp.com/discord
- **Try the tools**: github.com/timaeus-research/devinterp/

Timaeus

A toy model of superposition

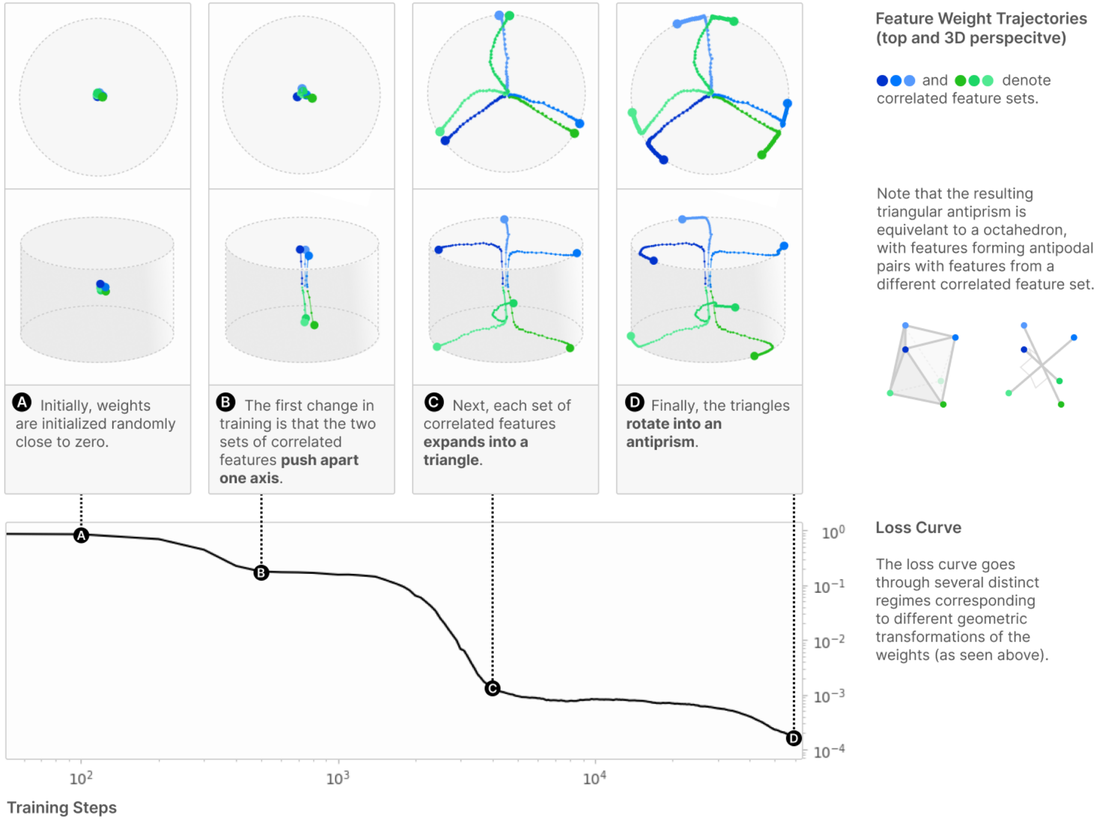
- Setup: Anthropic's toy model of superposition (Elhage et al. 2022)
- A very simple model:

$$f(x, w) = \text{ReLU}(W^T W x + b)$$

$$L_n(w) = \sum_{x_i} ||x_i - f(x_i, w)||^2$$

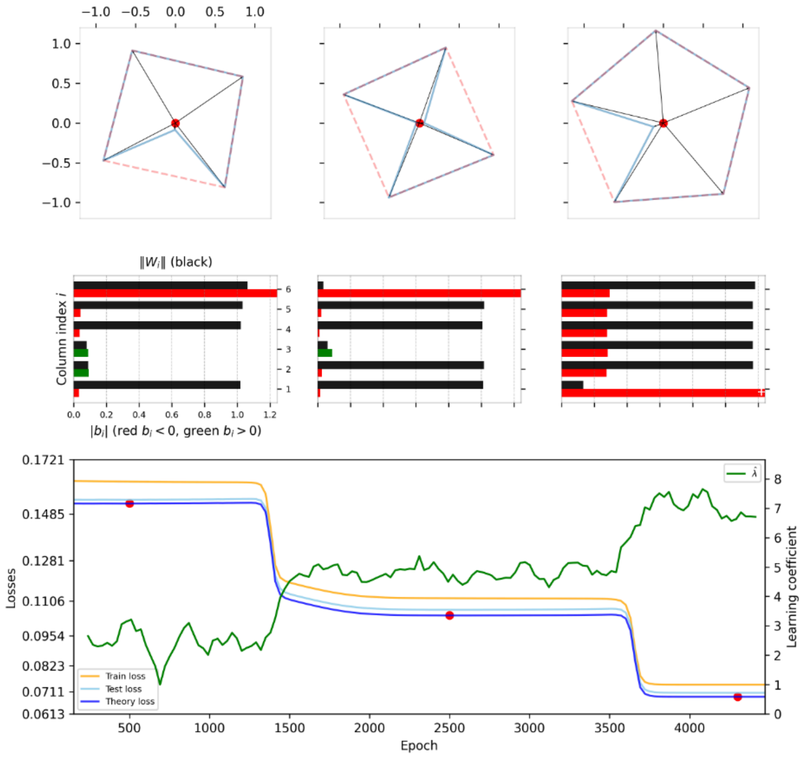
A toy model of superposition

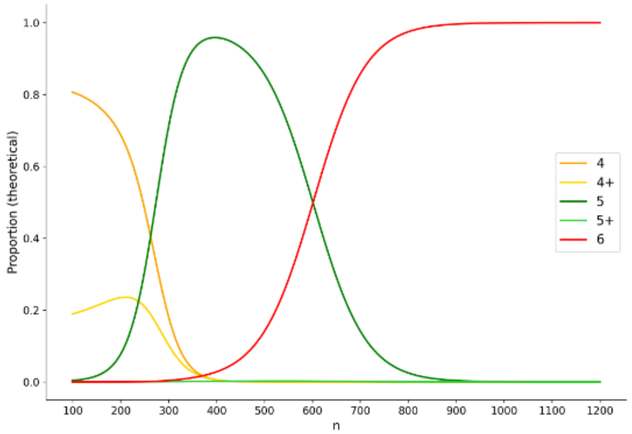
Elhage et al. (2022)

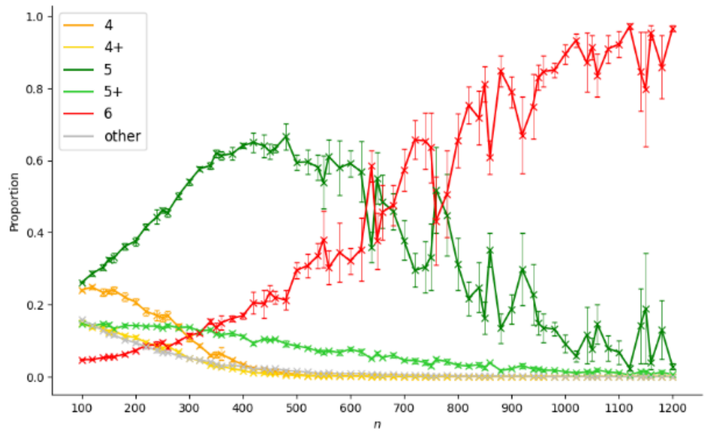


A toy model of superposition

Chen et al. (2023)

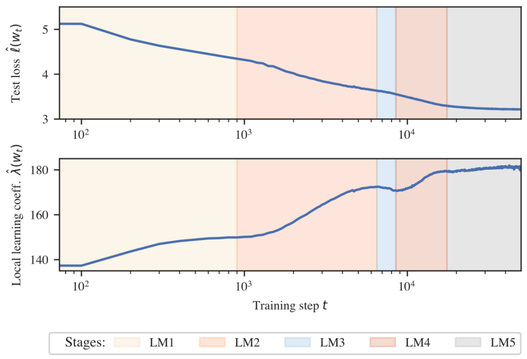


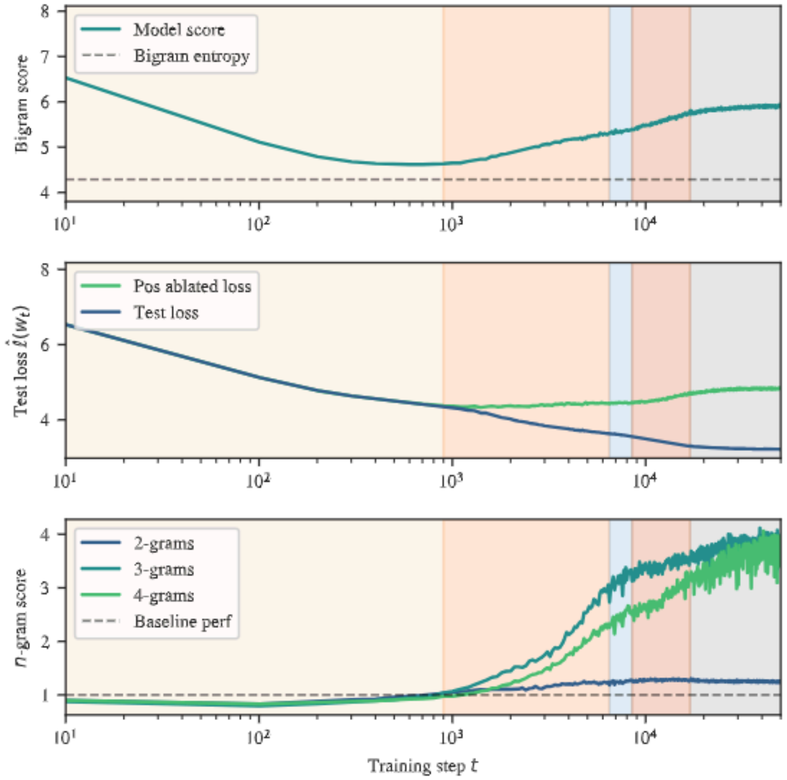


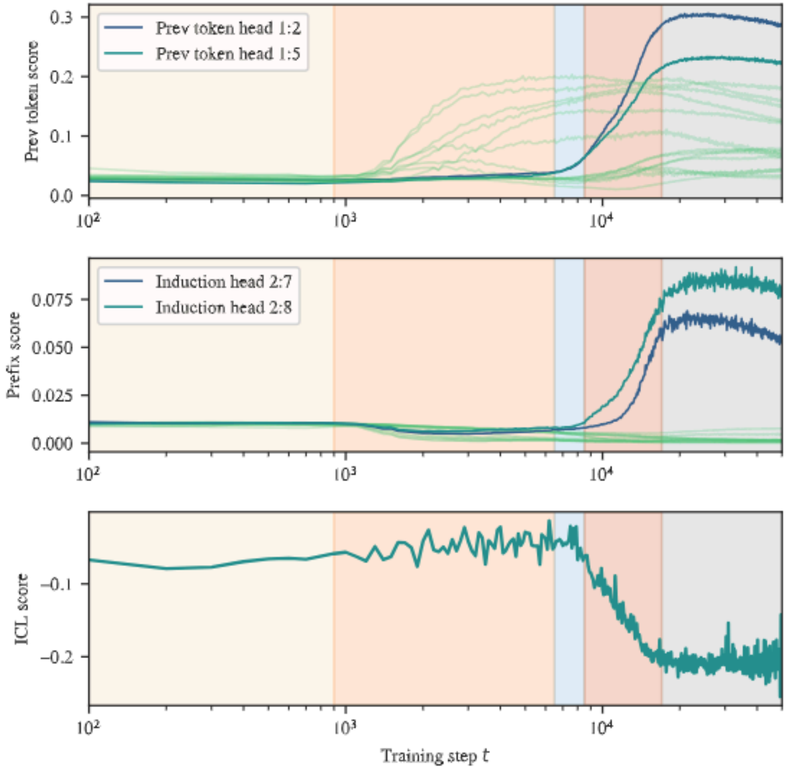


Language models

Hoogland et al. (2024)







Algorithmic tasks

Carroll et al. (in prep)