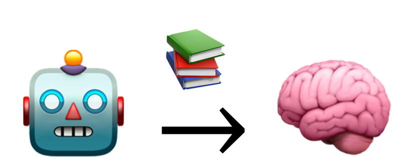
A phase transition between positional & semantic learning

in a solvable model of dot-product attention

Hugo Cui
Freya Behrens
Florent Krzakala
Lenka Zdeborová

Motivation Why and how do algorithmic abilities emerge in learned neural networks? How does the network understand the semantics of the inputs? Is this emergence a fast but smooth change of performance or a sharp boundary between different regimes of learning?

Phase Transition in Physics: Properties of a system of many interacting particles change apruptly as you change environmental parameters.



"Emergence" in ML: Capabilities/solution strategies of the learned algorithm change abruptly as more parameters/samples are available.

Positional & Semantic Learning To understand a sentence we use both two types of information.

The meaning of the tokens (semantics) ...

We sanitize a face ambition between rational and acrylic baking

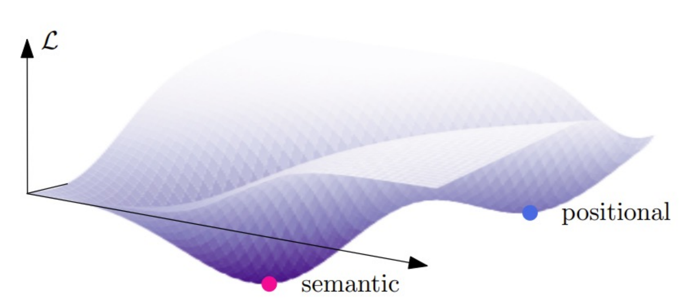
... and their ordering (positions)

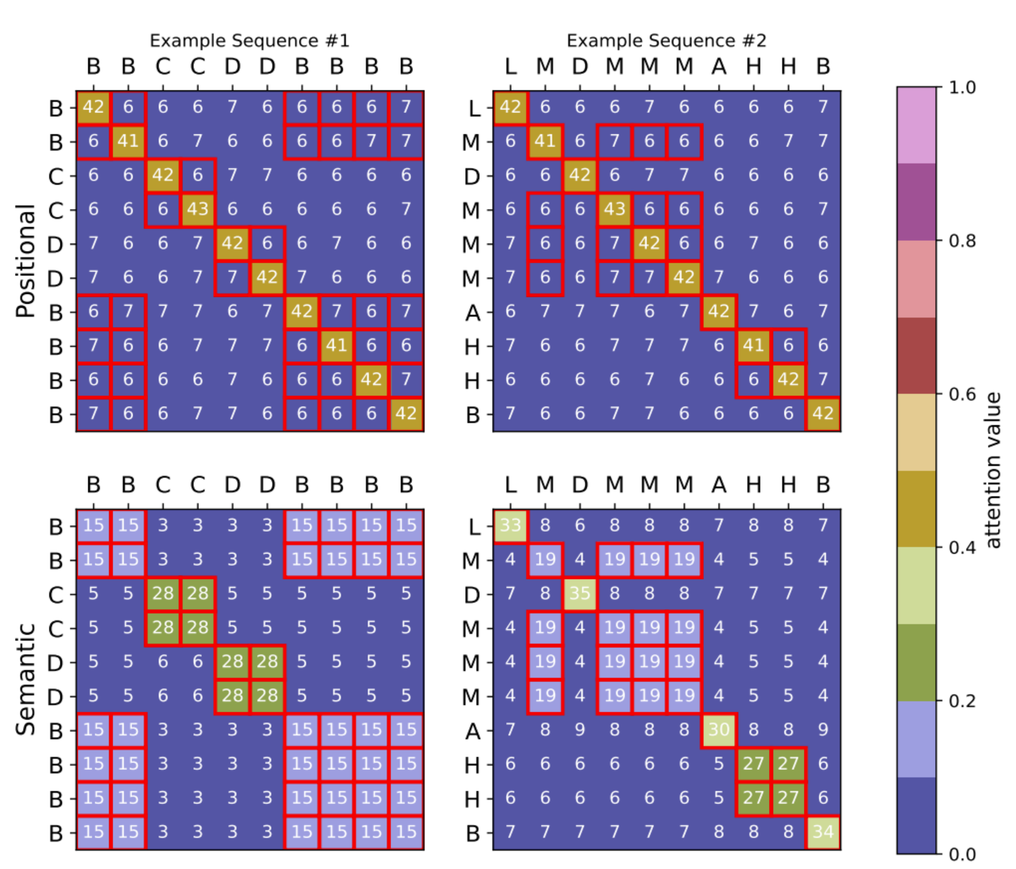
A between a phase semantic learning and positional analyze transition

Example: Histogram Task* For each input token count the number of identical tokens in the input sequence.

```
Input             -> Output
[B,A,A,D,E] -> [1,2,2,1,1]
[A,C,C,A,A] -> [3,2,2,3,3]
[C,C,C,C,D] -> [4,4,4,4,1]
```

We train a 1-layer transformer and find two minima of the loss:





Low-rank Tied Dot-Product Attention We use sentences of uncorrelated (1-gram) words as $x \in \mathbb{R}^{L\times d}$ with L tokens $\{x_l\}_{l=1\dots L}$ independently drawn from a Gaussian distribution $x_l \sim N(0,\Sigma_l)$ with covariance $\Sigma_l \in \mathbb{R}^{d\times d}$, and n data samples. The goal is to learn the target using the student, by optimizing the empirical risk:

Target/Teacher
$$y(x) = \mathtt{T}\left[\frac{1}{\sqrt{d}}\boldsymbol{x}\boldsymbol{Q}_\star\right]\boldsymbol{x}$$

Learned Student
$$f_{\boldsymbol{Q}}(x) = \mathtt{S}\left[\frac{1}{\sqrt{d}}(\boldsymbol{x}+\boldsymbol{p})\boldsymbol{Q}\right](\boldsymbol{x}+\boldsymbol{p})$$

ERM
$$\hat{\boldsymbol{Q}} = \underset{\boldsymbol{Q}\in\mathbb{R}^{d\times r}}{\operatorname{argmin}}\left[\sum_{\mu=1}^{n}\frac{1}{2d}\|y(\boldsymbol{x}^\mu)-f_{\boldsymbol{Q}}(\boldsymbol{x}^\mu)\|^2+\frac{\lambda}{2}\|\boldsymbol{Q}\|^2\right]$$

Test Error
$$\epsilon_g \equiv \frac{1}{dL}\mathbb{E}_{\boldsymbol{x}\sim p_x}\left\|y(\boldsymbol{x})-f_{\hat{\boldsymbol{Q}}}(\boldsymbol{x})\right\|^2$$

Main Technical Result Our result holds for teacher T and student S functions in the infinite sample and parameter limit where the sample complexity $\alpha = \frac{n}{d}$ is constant. We provide a closed-formed characterization of the test MSE and training loss. Our derivation exploits a mapping of the student to a (variant of) a Generalized Linear Model [NW72,M19] . Then, summary statistics characterized by self-consistent state evolution equations [JM13] asymptotically describe the fixed points of a Generalized Approximate Message Passing algorithm [RSR+16]. The fixed points of GAMP in turn correspond to *critical points of the non-convex empirical loss landscape*, so we can use them to describe the local minima and saddles of the loss. This limit has be considered before for similar models (e.g. [EPR+20]), but for attention only by [RGL+23] and without an emergent phenomenology.

Phenomenology For a concrete teacher T and student S we find a positional and a semantic minimum in the training loss landscape. There is a phase transition in terms of sample complexity α and the teacher mix ω.

data

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_L \end{pmatrix} \in \mathbb{R}^{L\times d} \qquad x_\ell \sim \mathcal{N}(0,\Sigma_\ell) \in \mathbb{R}^d$$

target

$$y(x) = \left[(\mathbf{1}-\boldsymbol{\omega})\mathrm{softmax}\left(\frac{x\,Q_\star Q_\star^\top x^\top}{d}\right) + \boldsymbol{\omega}\boldsymbol{A}\right]\cdot x$$

with $A \in \mathbb{R}^{L\times L}$, $Q_* \in \mathbb{R}^d$

student

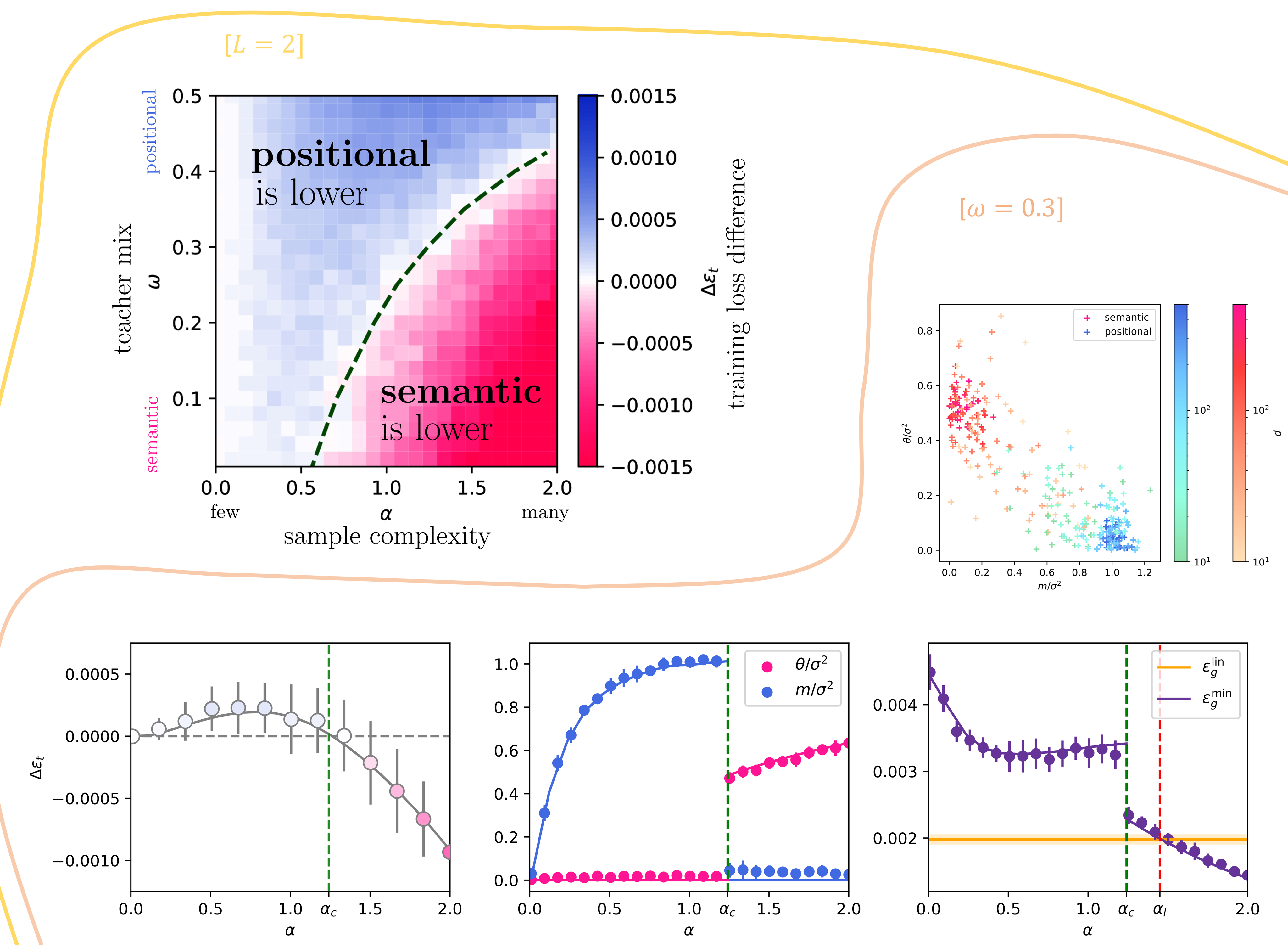
$$f_Q(x) = \mathrm{softmax}\left(\frac{(x+p)QQ^\top(x+p)^\top}{d}\right)\cdot x$$

with , $p = \begin{pmatrix}\mu \\ -\mu\end{pmatrix} Q \in \mathbb{R}^d$

observables

$$\theta_\ell \equiv \frac{\hat{Q}^\top \Sigma_\ell Q_\star}{d} \in \mathbb{R}^{r_s\times r_t} \qquad m_\ell \equiv \frac{\hat{Q}^\top p_\ell}{d} \in \mathbb{R}^{r_s}$$

semanticity positionality



[SMK23] Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? NeurIPS 2023
[NW72] John Ashworth Nelder and Robert WM Wedderburn. Generalized linear models. Journal of the Royal Statistical Society Series A, 1972.
[M19] Peter McCullagh. Generalized linear models. Routledge, 2019

* MORE INFO → ON HISTOGRAM

[JM13] Adel Javanmard and Andrea Montanari. State evolution for general approximate message passing algorithms, with applications to spatial coupling. Information and Inference, 2013
[RGL+23] Riccardo Rende, Federica Gerace, Alessandro Laio, and Sebastian Goldt. Optimal inference of a generalised Potts model by single-layer transformers with factored attention. arXiv:2304.07235, 2023
[EPR+20] Melikasadat Emami, Mojtaba Sahraee-Ardakan, Parthe Pandit, Sundeep Rangan, and Alyson Fletcher. Generalization error of generalized linear models in high dimensions. ICML, 2020
[RSR+16] Sundeep Rangan, Philip Schniter, Erwin Riegler, Alyson K Fletcher, and Volkan Cevher. Fixed points of generalized approximate message passing with arbitrary matrices. IEEE Transactions on Information Theory, 2016