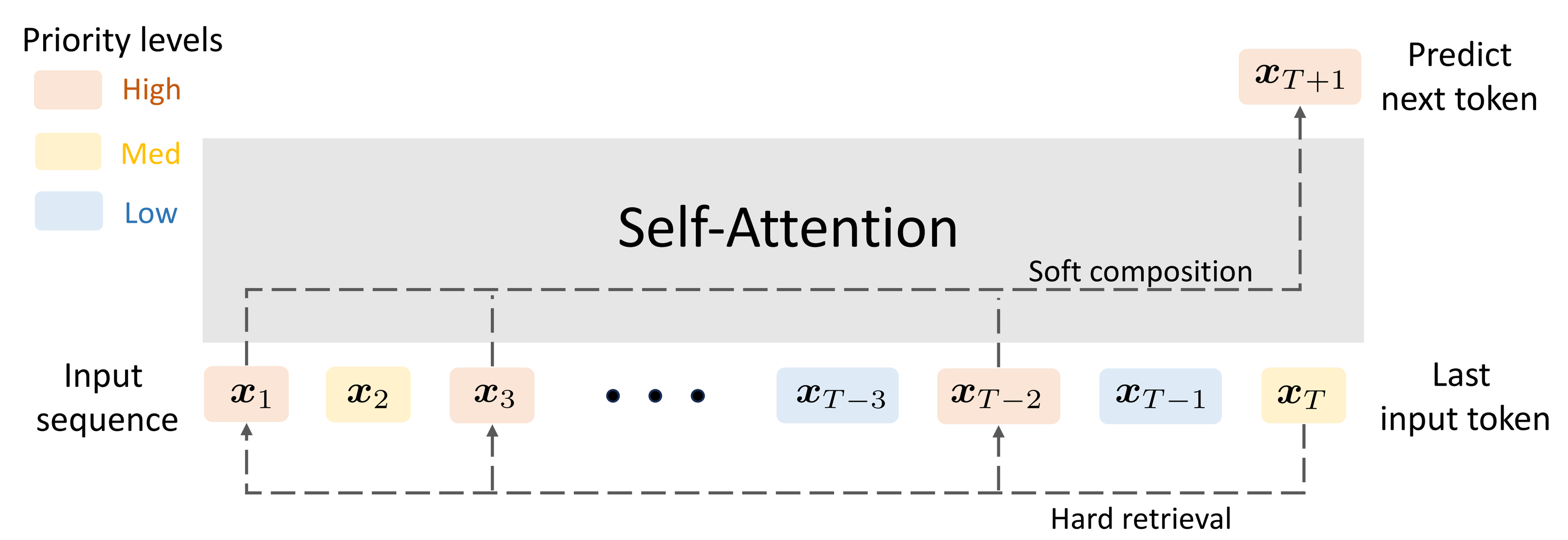
Mechanics of Next Token Prediction with Self-Attention

Yingcong Li[1,†] Yixiao Huang[1,†] M. Emrullah Ildiz[1] Ankit Singh Rawat[2] Samet Oymak[1]

University of Michigan, Ann Arbor[1] Google Research NYC[2] Equal contribution[†]

Motivation



Question

What relationships in the training data are captured by the single-layer self-attention model?

Motivation

Exploring implicit bias is a key step in unraveling the generalization of the (softmax-)attention mechanism.

Optimization Methods

- **Gradient descent:** Given starting point $\boldsymbol{W}(0)$ and step size η,

$$\boldsymbol{W}(\tau+1) = \boldsymbol{W}(\tau) - \eta\nabla\mathcal{L}(\boldsymbol{W}(\tau)). \qquad \text{(Algo-GD)}$$

- **Regularization path:** Given radius $R > 0$, $\boldsymbol{W} \in \mathbb{R}^{d\times d}$,

$$\bar{\boldsymbol{W}}_R = \arg\min_{\|\boldsymbol{W}\|_F \le R} \mathcal{L}(\boldsymbol{W}). \qquad \text{(Algo-RP)}$$

Theorem (informal)

The combined attention weights $\boldsymbol{W} := \boldsymbol{W}_K\boldsymbol{W}_Q^\top$ evolve as

$$\boldsymbol{W}_{\text{GD}} \approx C \cdot \boldsymbol{W}_{\text{hard}} + \boldsymbol{W}_{\text{soft}},$$

where $C \cdot \boldsymbol{W}_{\text{hard}}$ is the hard retrieval component and $\boldsymbol{W}_{\text{soft}}$ is the soft composition component.

Problem Formulation

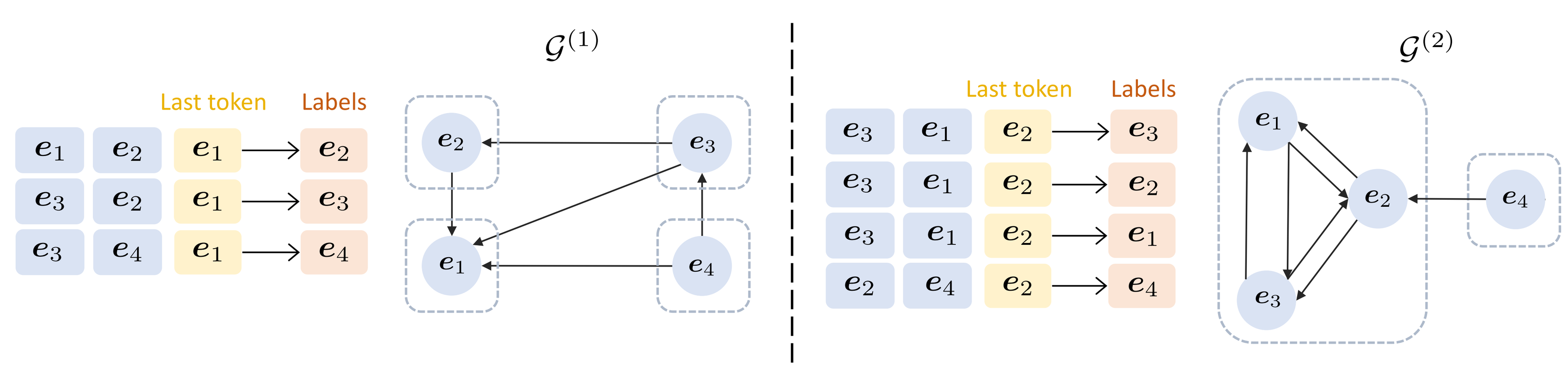
Next-token prediction

$$f_{\boldsymbol{W}}(\boldsymbol{X}) = \boldsymbol{X}^\top \mathbb{S}(\boldsymbol{X}\boldsymbol{W}\boldsymbol{x}_{\text{last}})$$

- $\mathbb{S}(\cdot)$: softmax function; $\boldsymbol{W} := \boldsymbol{W}_K\boldsymbol{W}_Q^\top$: attention weights.

Problem description: Given embedding matrix $\boldsymbol{E} = [\boldsymbol{e}_1 \cdots \boldsymbol{e}_K]^\top \in \mathbb{R}^{K\times d}$ and input $\boldsymbol{X} \in \mathbb{R}^{T\times d}$, where $\boldsymbol{x}_t \in \boldsymbol{E}$, the next-token prediction is to predict the next token $y \in [K]$. Then given training dataset $\{(\boldsymbol{X}_i, y_i)\}_{i=1}^n$, linear prediction head $\boldsymbol{c}_k, k \in [K]$ and loss ℓ, we consider ERM problem:

$$\mathcal{L}(\boldsymbol{W}) = \frac{1}{n}\sum_{i=1}^n \ell(\boldsymbol{c}_{y_i}^\top \boldsymbol{X}_i^\top \mathbb{S}(\boldsymbol{X}_i\boldsymbol{W}\boldsymbol{x}_{i,\text{last}})).$$



Token-Priority Graph (TPG)

Suppose $(\boldsymbol{X}, y)$ has query/last token k. For all (x, y) pairs in $(\boldsymbol{X}, y)$ where x is the token ID of $\boldsymbol{x}$, add a directed edge $(y \to x)$ to graph $\mathcal{G}^{(k)}$.

- $(i \Rightarrow j) \in \mathcal{G}^{(k)}$: $(i \to j)$ is present in $\mathcal{G}^{(k)}$ but $(j \to i)$ is not.
- $(i \asymp j) \in \mathcal{G}^{(k)}$: both nodes i, j are in the same SCC of $\mathcal{G}^{(k)}$.

Attention SVM

$$\boldsymbol{W}^{\text{svm}} = \arg\min_{\boldsymbol{W}} \|\boldsymbol{W}\|_F \qquad \text{(Graph-SVM)}$$

$$\text{s.t.} \quad (\boldsymbol{e}_i - \boldsymbol{e}_j)^\top \boldsymbol{W}\boldsymbol{e}_k \begin{cases} = 0 & \forall (i \asymp j) \in \mathcal{G}^{(k)} \\ \ge 1 & \forall (i \Rightarrow j) \in \mathcal{G}^{(k)} \end{cases}$$

Main Results

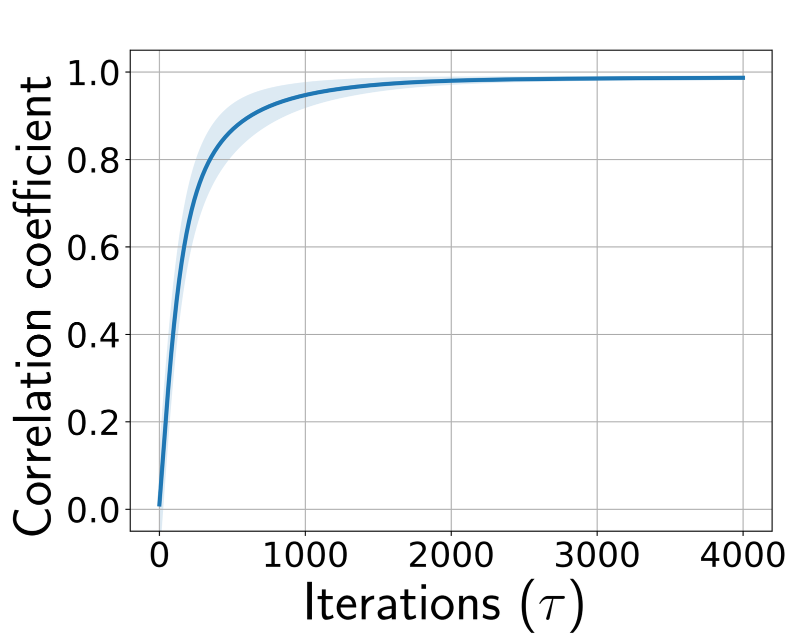
Definition:

Define cyclic subspace $\mathcal{S}_{\text{fin}}$ as the span of all matrices $(\boldsymbol{e}_i - \boldsymbol{e}_j)\boldsymbol{e}_k^\top$ for all $(i \asymp j) \in \mathcal{G}^{(k)}$ and $k \in [K]$.

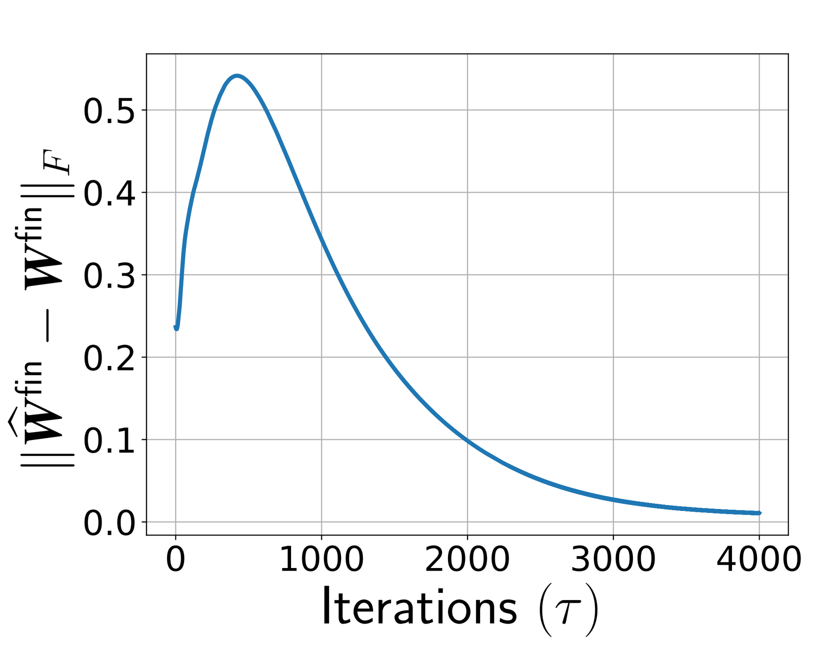
Assumptions:

1. For $\forall y, k \in [K]$, $k \neq y$, $\boldsymbol{c}_y^\top \boldsymbol{e}_y = 1$ and $\boldsymbol{c}_y^\top \boldsymbol{e}_k = 0$; and
2. For any $(\boldsymbol{X}, y)$, token $\boldsymbol{e}_y$ is contained in the input sequence $\boldsymbol{X}$.

Simulation Results:



(a) Evolution of $\frac{\boldsymbol{W}(\tau)}{\|\boldsymbol{W}(\tau)\|_F} \to \frac{\boldsymbol{W}^{\text{svm}}}{\|\boldsymbol{W}^{\text{svm}}\|_F}$



(b) Evolution of $\boldsymbol{\Pi}_{\mathcal{S}_{\text{fin}}}(\boldsymbol{W}(\tau)) \to \boldsymbol{W}^{\text{fin}}$

(a) shows the directional convergence of $\boldsymbol{W}(\tau)$;
(b) presents the convergence of $\boldsymbol{\Pi}_{\mathcal{S}_{\text{fin}}}(\boldsymbol{W}(\tau))$.

Theorem I: Convergence of Gradient Descent

Suppose Assumptions 1&2 hold and $\ell(u) = -\log(u)$. Let $\boldsymbol{W}^{\text{svm}} \in \mathcal{S}_{\text{fin}}^\perp$ be the solution of (Graph-SVM) and suppose $\boldsymbol{W}^{\text{svm}} \neq 0$. Starting from any $\boldsymbol{W}(0)$ with constant step size η, the algorithm Algo-GD satisfies $\lim_{\tau\to\infty} \|\boldsymbol{W}(\tau)\|_F = \infty$,

$$\lim_{\tau\to\infty} \frac{\boldsymbol{W}(\tau)}{\|\boldsymbol{W}(\tau)\|_F} = \frac{\boldsymbol{W}^{\text{svm}}}{\|\boldsymbol{W}^{\text{svm}}\|_F} \quad \text{and} \quad \lim_{\tau\to\infty} \boldsymbol{\Pi}_{\mathcal{S}_{\text{fin}}}(\boldsymbol{W}(\tau)) = \boldsymbol{W}^{\text{fin}}.$$

Here $\boldsymbol{W}^{\text{fin}}$ is the unique finite minima of the loss $\tilde{\mathcal{L}}(\boldsymbol{W}) := \lim_{R\to\infty} \mathcal{L}(\boldsymbol{W} + R \cdot \boldsymbol{W}^{\text{svm}})$ over $\mathcal{S}_{\text{fin}}$.

Theorem II: Convergence of Regularized Path

Suppose Assumptions 1&2 hold and loss $\ell : \mathbb{R} \to \mathbb{R}$ is strictly decreasing and $|\ell'|$ is bounded. Let $\boldsymbol{W}^{\text{svm}} \in \mathcal{S}_{\text{fin}}^\perp$ be the solution of (Graph-SVM) and suppose $\boldsymbol{W}^{\text{svm}} \neq 0$. Then the solution of regularization path Algo-RP obeys

$$\lim_{R\to\infty} \frac{\bar{\boldsymbol{W}}_R}{R} = \frac{\boldsymbol{W}^{\text{svm}}}{\|\boldsymbol{W}^{\text{svm}}\|_F} \quad \text{and} \quad \lim_{R\to\infty} \boldsymbol{\Pi}_{\mathcal{S}_{\text{fin}}}(\bar{\boldsymbol{W}}_R) \in \mathcal{W}^{\text{fin}}.$$

Here $\mathcal{W}^{\text{fin}} = \arg\min_{\boldsymbol{W}\in\mathcal{S}_{\text{fin}}} \lim_{R\to\infty} \mathcal{L}(\boldsymbol{W} + R \cdot \boldsymbol{W}^{\text{svm}})$.