Provably Transformers Harness Multi-Concept Word Semantics for Efficient In-Context Learning

Dake Bu, Wei Huang*, Andi Han, Atsushi Nitanda, Taiji Suzuki, Qingfu Zhang, Hau-San Wong*

Contents

Motivation & Background

Japan: Sakura; China:

Japan: Sakura; China: Peony.

The peony is often considered a national flower of China and symbolizes prosperity, honor, and beauty. It holds cultural and historical significance, much like the cherry blossom does in Japan.

Japan: Sakura; Australia: Kangaroo; China:

Japan: Sakura; Australia: Kangaroo; China: Panda.

The giant panda is widely recognized as a symbol of China, representing peace, friendship, and the country's commitment to wildlife conservation.

Motivation & Background

Japan: Sakura; China:

Japan: Sakura; China: Peony.

The peony is often considered a national flower of China and symbolizes prosperity, honor, and beauty. It holds cultural and historical significance, much like the cherry blossom does in Japan.

Japan: Sakura; Australia: Kangaroo; China:

Japan: Sakura; Australia: Kangaroo; China: Panda.

The giant panda is widely recognized as a symbol of China, representing peace, friendship, and the country's commitment to wildlife conservation.

- **Observation**: Different *task concepts* are identified from the same words in different prompt.

Motivation & Background

Japan: Sakura; China:

Japan: Sakura; China: Peony.

The peony is often considered a national flower of China and symbolizes prosperity, honor, and beauty. It holds cultural and historical significance, much like the cherry blossom does in Japan.

Japan: Sakura; Australia: Kangaroo; China:

Japan: Sakura; Australia: Kangaroo; China: Panda.

The giant panda is widely recognized as a symbol of China, representing peace, friendship, and the country's commitment to wildlife conservation.

- **Observation**: Different *task concepts* are identified from the same words in different prompt.
- **Question**: Why can an additional demo-pair influence the outcome of ICL greatly?

Motivation & Background

➢ Observed Multi-Concept Latent Geometric Linearity of LLM.

Existing studies [1-4] suggest the multi-concepts are encoded linearly in the latent representation of LLM.

- Representations *within-concepts (topics)* have positive inner products
- Representations *cross-concepts (topics)* exhibit near-orthogonal relationships
- ICA is more suitable than PCA when extracting meaningful concepts

[1] Yamagiwa et al. Discovering universal geometry in embeddings with ICA. EMNLP 2023
[2] Li et al. How Do Transformers Learn Topic Structure: Towards a Mechanistic Understanding. ICML 2023
[3] Park et al. 2023: The linear representation hypothesis and the geometry of large language models. ICML 2024
[4] Jiang et. al. On the origins of linear representations in large language models. ICML 2024
[5] Reizinger et al. Position: Understanding LLMs Requires More Than Statistical Generalization. ICML 2024

Motivation & Background

➢ Observed Multi-Concept Latent Geometric Linearity of LLM.

Existing studies [1-4] suggest the multi-concepts are encoded linearly in the latent representation of LLM.

- Representations *within-concepts (topics)* have positive inner products
- Representations *cross-concepts (topics)* exhibit near-orthogonal relationships
- ICA is more suitable than PCA when extracting meaningful concepts

Essential Question

Whether and how do the observed latent geometry facilitate transformer in ICL, especially in OOD scenario?

Remark: This question is also raised as a research question **Question 5.1.4** in [5], available after our submission.

[1] Yamagiwa et al. Discovering universal geometry in embeddings with ICA. EMNLP 2023
[2] Li et al. How Do Transformers Learn Topic Structure: Towards a Mechanistic Understanding. ICML 2023
[3] Park et al. 2023: The linear representation hypothesis and the geometry of large language models. ICML 2024
[4] Jiang et. al. On the origins of linear representations in large language models. ICML 2024
[5] Reizinger et al. Position: Understanding LLMs Requires More Than Statistical Generalization. ICML 2024

Motivation & Background

➢ Observed Multi-Concept Latent Geometric Linearity of LLM.

Existing studies [1-4] suggest the multi-concepts are encoded linearly in the latent representation of LLM.

- Representations *within-concepts (topics)* have positive inner products
- Representations *cross-concepts (topics)* exhibit near-orthogonal relationships
- ICA is more suitable than PCA when extracting meaningful concepts

➢ Existing transformer theories suffer from unrealistic settings.

- Prior theories are conducted on unrealistic settings such as linear or ReLU transformers, MLP-free attention-only models, QK-combined softmax attention and impractical loss functions like square / hinge loss.
- Due to their technical limitation, they only obtain linear or sub-linear convergence rates.

[1] Yamagiwa et al. Discovering universal geometry in embeddings with ICA. EMNLP 2023
[2] Li et al. How Do Transformers Learn Topic Structure: Towards a Mechanistic Understanding. ICML 2023
[3] Park et al. 2023: The linear representation hypothesis and the geometry of large language models. ICML 2024
[4] Jiang et. al. On the origins of linear representations in large language models. ICML 2024
[5] Reizinger et al. Position: Understanding LLMs Requires More Than Statistical Generalization. ICML 2024

Contents

Introduction

Grounded in the studies of the LLM linear concept representation, we conduct theoretical analysis on a concept-specific sparse coding prompt distribution for ICL bi-classification tasks. Our main contributions are highlighted as below.

- We are the first to analyze the realistic setting: *softmax* attention + *ReLU* MLP transformer, which is trained using the *cross-entropy loss* via stochastic gradient descent

Introduction

Grounded in the studies of the LLM linear concept representation, we conduct theoretical analysis on a concept-specific sparse coding prompt distribution for ICL bi-classification tasks. Our main contributions are highlighted as below.

- We are the first to analyze the realistic setting: *softmax* attention + *ReLU* MLP transformer, which is trained using the *cross-entropy loss* via stochastic gradient descent
- We are the first to showcase the *exponential 0-1 loss convergence* over the highly non-convex training dynamics in ICL theory

Introduction

Grounded in the studies of the LLM linear concept representation, we conduct theoretical analysis on a concept-specific sparse coding prompt distribution for ICL bi-classification tasks. Our main contributions are highlighted as below.

- We are the first to analyze the realistic setting: *softmax* attention + *ReLU* MLP transformer, which is trained using the *cross-entropy loss* via stochastic gradient descent
- We are the first to showcase the *exponential 0-1 loss convergence* over the highly non-convex training dynamics in ICL theory
- We provably show that transformers can perform *certain OOD ICL tasks* by leveraging the multi-concept semantic linearity after training, highlighting their *innovative potential* for large models.

Contents

Problem & Model Formulation

- **Polysemous Word Model.** $(\mathcal{D}_{\boldsymbol{x}}, \mathcal{D}_{\boldsymbol{y}}, \mathcal{D}_{\boldsymbol{z}}, \mathcal{D}_{\xi_{\boldsymbol{x}}}, \mathcal{D}_{\xi_{\boldsymbol{y}}})$

Define the feature and label dictionaries:

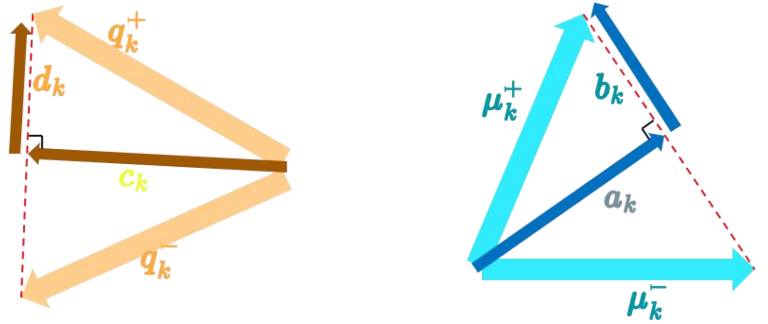
$$\mathbf{M} = [\boldsymbol{\mu}_1^+, \boldsymbol{\mu}_1^-, \cdots, \boldsymbol{\mu}_{K_1}^+, \boldsymbol{\mu}_{K_1}^-, \boldsymbol{\nu}_1, \cdots, \boldsymbol{\nu}_{K_2}]$$
$$\mathbf{Q} = [\boldsymbol{q}_1^+, \boldsymbol{q}_1^-, \cdots, \boldsymbol{q}_{K_1}^+, \boldsymbol{q}_{K_1}^-, 0, \cdots 0]$$

satisfying *within-concepts positive inner product* and *cross-concepts orthogonal* relationships.

There exists 0 < κ_x, κ_y < 1 such that

$$0 < \cos\langle \boldsymbol{\mu}_{k_1}^+, \boldsymbol{\mu}_{k_1}^- \rangle \leq \kappa_{\boldsymbol{x}},\ 0 < \cos\langle \boldsymbol{q}_{k_1}^+, \boldsymbol{q}_{k_1}^- \rangle \leq \kappa_{\boldsymbol{y}}$$

We can *naturally* define the high-level concept features $\boldsymbol{a}_k := (\boldsymbol{\mu}_k^+ + \boldsymbol{\mu}_k^-)/2$ and the low-level semantic label features $\boldsymbol{b}_k := (\boldsymbol{\mu}_k^+ - \boldsymbol{\mu}_k^-)/2$ Also we define $\boldsymbol{c}_k := (\boldsymbol{q}_k^+ + \boldsymbol{q}_k^-)/2$, $\boldsymbol{d}_k := (\boldsymbol{q}_k^+ - \boldsymbol{q}_k^-)/2$.



Problem & Model Formulation

➢ **Polysemous Word Model.** $(\mathcal{D}_{\boldsymbol{x}}, \mathcal{D}_{\boldsymbol{y}}, \mathcal{D}_{\boldsymbol{z}}, \mathcal{D}_{\xi_{\boldsymbol{x}}}, \mathcal{D}_{\xi_{\boldsymbol{y}}})$

Then, the $\boldsymbol{z}$, $\xi_{\boldsymbol{x}}$, $\xi_{\boldsymbol{y}}$ are generated from $\mathcal{D}_{\boldsymbol{z}}$, and Gaussian distributions $\mathcal{D}_{\xi_{\boldsymbol{x}}}, \mathcal{D}_{\xi_{\boldsymbol{y}}}$ independently.

By reparameterization we define

$$\boldsymbol{x} := \mathbf{M}\boldsymbol{z} + \xi_{\boldsymbol{x}} \sim \mathcal{D}_{\boldsymbol{x}}, \quad \boldsymbol{y} := \mathbf{Q}\boldsymbol{z} + \xi_{\boldsymbol{y}} \sim \mathcal{D}_{\boldsymbol{y}}$$

➢ **Concept-specific Prompt Distribution.** $(\mathcal{D}_S)$

$$\mathcal{D}_S = \sum_{k=1}^{K_1} \left(\pi_k^+ \mathcal{P}_{k,L+1}^+ + \pi_k^- \mathcal{P}_{k,L+1}^- \right)$$

$$\begin{array}{|c|c|c|c|} \hline \mathbf{x}_1 & \cdots & \mathbf{x}_L & \mathbf{x}_{L+1} \\ \hline \mathbf{y}_1 & \cdots & \mathbf{y}_L & \mathbf{0} \\ \hline \end{array} = \begin{pmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_L & \boldsymbol{x}_{\text{query}} \\ \boldsymbol{y}_1 & \boldsymbol{y}_2 & \cdots & \boldsymbol{y}_L & \mathbf{0} \end{pmatrix}$$

where $\pi_k^\pm = (2K_1)^{-1}$ denotes the equal chance of $\mathcal{P}_{k,L+1}^\pm$; $\mathcal{P}_{k,L+1}^{y_{S_n}}$ represents the k-th concept-specific prompt distribution; $y_{S_n} \in [\pm 1]$ is the true label of a prompt. Each demo-pair $(\boldsymbol{x}_l^n, \boldsymbol{y}_l^n)$ in $\mathcal{P}_{k,L+1}^e$ includes either $(\boldsymbol{\mu}_k^+, \boldsymbol{q}_k^+)$ or $(\boldsymbol{\mu}_k^-, \boldsymbol{q}_k^-)$ with equal chance. Furthermore, $\forall l \in [L+1]$, $\mathbb{P}(z_{l,\neg(2k-1 \vee 2k)}^n = 1) = K^{-1}$, indicating an equal chance of diverse task-irrelevant feature presence.

Problem & Model Formulation

➢ **Transformer Model.** $\Psi' := \{\mathbf{W}_Q^{\boldsymbol{x}}, \mathbf{W}_K^{\boldsymbol{x}}, \mathbf{W}_O^{\boldsymbol{y}}\}$

$$\mathbf{H} = \mathbf{E}(S) = \begin{pmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_L & \boldsymbol{x}_{\text{query}} \\ \boldsymbol{y}_1 & \boldsymbol{y}_2 & \cdots & \boldsymbol{y}_L & \mathbf{0} \end{pmatrix} := (\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_{\text{query}}) \in \mathbb{R}^{(d_{\mathcal{X}}+d_{\mathcal{Y}})\times(L+1)}$$

$$f(\mathbf{H}; \Psi) = \mathbf{r}^\top \sigma_R\left(\mathbf{W}_O \operatorname{attn}(\mathbf{H}; \Psi)\right),$$

$$\operatorname{attn}(\mathbf{H}; \Psi) = \sum_{l=1}^{L} \mathbf{W}_V \mathbf{h}_l \sigma_S\left((\mathbf{W}_K \mathbf{h}_l)^\top \mathbf{W}_Q \mathbf{h}_{\text{query}}\right),$$

$$\mathbf{W}_Q = \begin{pmatrix} \mathbf{W}_Q^{\boldsymbol{x}} & * \\ * & * \end{pmatrix}, \quad \mathbf{W}_K = \begin{pmatrix} \mathbf{W}_K^{\boldsymbol{x}} & * \\ * & * \end{pmatrix}, \quad \mathbf{W}_V = \begin{pmatrix} * & * \\ * & \mathbf{W}_V^{\boldsymbol{y}} \end{pmatrix} \quad \mathbf{W}_O = (* \quad \mathbf{W}_O^{\boldsymbol{y}}),$$

where $\mathbf{W}_Q^{\boldsymbol{x}}, \mathbf{W}_K^{\boldsymbol{x}} \in \mathbb{R}^{d_{\mathcal{X}}\times d_{\mathcal{X}}}$, $\mathbf{W}_V^{\boldsymbol{y}} \in \mathbb{R}^{(m_v - d_{\mathcal{X}})\times d_{\mathcal{Y}}}$, $\mathbf{W}_O^{\boldsymbol{y}} \in \mathbb{R}^{m\times d_{\mathcal{Y}}}$. Here, we set the elements other than $\mathbf{W}_Q^{\boldsymbol{x}}, \mathbf{W}_K^{\boldsymbol{x}}, \mathbf{W}_V^{\boldsymbol{y}}$ and $\mathbf{W}_O^{\boldsymbol{y}}$ to be zero. Besides, we fix $\mathbf{W}_V^{\boldsymbol{y}}$ to be $\mathbf{I}_{(m_v - d_{\mathcal{X}})\times d_{\mathcal{Y}}}$. We sample $\mathbf{r}_i$ from a uniform distribution $\text{Unif}\{-1, 1\}$ and fixed during the training process. Based on this setting, the trainable part we need to consider is actually $\Psi' := \{\mathbf{W}_Q^{\boldsymbol{x}}, \mathbf{W}_K^{\boldsymbol{x}}, \mathbf{W}_O^{\boldsymbol{y}}\}$. This problem remains highly non-convex and challenging.

Problem & Model Formulation

➢ **Stochastic Gradient Descent**.

$$L_{\mathcal{B}_t}(\Psi) = L_{\mathcal{B}_t}(\Psi') := \frac{1}{B}\sum_{n\in\mathcal{B}_t} \ell\left(y_{S_n} \cdot f(\mathbf{H};\Psi)\right) + \frac{\lambda}{2}\|\Psi'\|_F^2,$$

where $\ell(z) = \log(1+\exp(-z))$ $\|\Psi'\|_F^2$ represents $\|\mathbf{W}_Q^{\boldsymbol{x}}\|_F^2 + \|\mathbf{W}_K^{\boldsymbol{x}}\|_F^2 + \|\mathbf{W}_O^{\boldsymbol{y}}\|_F^2$ $\eta_t = \frac{2}{\lambda(\gamma+t)}$

Initialization Setting. All initial values of $\mathbf{W}_O^{\boldsymbol{y}}$ are sampled from a i.i.d. Gaussian distributions with mean 0 and variance σ_1^2. The initialization of $\mathbf{W}_Q^{\boldsymbol{x}}$ and $\mathbf{W}_K^{\boldsymbol{x}}$ are diagonal matrices $\sigma_0\mathbb{I}$

Algorithm 1 Training algorithm

Input: Training distribution $\mathcal{D}_S$, Test distribution $\mathcal{D}^*$, Batch size B, step size $\eta_t = \frac{2}{\lambda(\gamma+t)}$, stopping criterion ε and total epochs T.

Initialize model parameters $\Psi'^{(0)}$.

for $t = 0, 1, \ldots, T-1$ **do**

  If $L_{\mathcal{D}^*}^{0-1}(\Psi^{(t)}) \le \varepsilon$ stop else continue.

  Randomly sample mini batches $\mathcal{B}_t$ of size B from $\mathcal{D}_S$.

  Update model parameters: $\Psi'^{(t+1)} = \Psi'^{(t)} - \eta_t \nabla_{\Psi'} L_{\mathcal{B}_t}(\Psi'^{(t)})$.

end for

Contents

Main Result

- **Exponential Convergence of 0-1 loss under low-noise condition**

Theorem 1. Under Condition 1, for $\forall\varepsilon > 0$, $\exists$ C1,C2 > 0, with probability no less than $1 - \delta$, for $T \geq \hat{T}$, we have

$$L_{\mathcal{D}^*}^{0-1}(\Psi^{(T)}) \leq \exp\left(-\frac{C_2\nu^2 m\lambda^2(\gamma + T)}{K_1\|\mathbf{q}\|^2((L-1)\|\mathbf{u}\|^2 + 1)}\right).$$

- Thus after $T_\varepsilon = \frac{K_1\|\mathbf{q}\|^2((L-1)\|\mathbf{u}\|^2+1)}{C_2\nu^2 m\lambda^2}\log(\frac{1}{\varepsilon})$ iterations, we have

$$L_{\mathcal{D}^*}^{0-1}(\Psi^{(T)}) \leq \varepsilon$$

- Importantly, $\hat{T}$ is independent of ε and does not affect the convergence rate as $\varepsilon \to 0$.

Main Result

- **Out-of-Distribution-Generalization.**

Proposition 1. Under Condition 1, for $\forall \varepsilon > 0$, The learned model satisfies $L^{0-1}_{\mathcal{D}^*_S}(\Psi^{(T^*)}) \leq \varepsilon$ for $T^* \geq T_\varepsilon$, where the $\mathcal{D}^*_S$ can enjoy the following distribution shifts.

- The prompt length can be any positive integer.
- $\mathcal{D}^*_{\boldsymbol{z}}$ can enjoy any shift, with each prompt sharing ≥ 1 co-concept, and equal chance to be ± 1.
- $\mathcal{D}^*_{\boldsymbol{x}} \times \mathcal{D}^*_{\boldsymbol{y}}$ can enjoy great shift. The new M^* and Q^* satisfying that

$$\boldsymbol{\mu}_k^{\pm *} = \boldsymbol{a}_k^* \pm \boldsymbol{b}_k^*, \quad \boldsymbol{q}_k^{\pm *} = \boldsymbol{c}_k^* \pm \boldsymbol{d}_k^*, \quad \boldsymbol{\nu}_{k_2} = \boldsymbol{\nu}_{k_2}^*$$

The $\boldsymbol{a}_k^*$, $\boldsymbol{b}_k^*$, $\boldsymbol{c}_k^*$, $\boldsymbol{d}_k^*$, $\boldsymbol{\nu}_{k_2}$ are any vectors in the conic hulls of

$$\{\boldsymbol{a}_k\}_{k=1}^{K_1}, \{\boldsymbol{b}_k\}_{k=1}^{K_1}, \{\boldsymbol{c}_k\}_{k=1}^{K_1}, \{\boldsymbol{d}_k\}_{k=1}^{K_1}, \{\pm\boldsymbol{\nu}_k\}_{k_2}^{K_2}$$

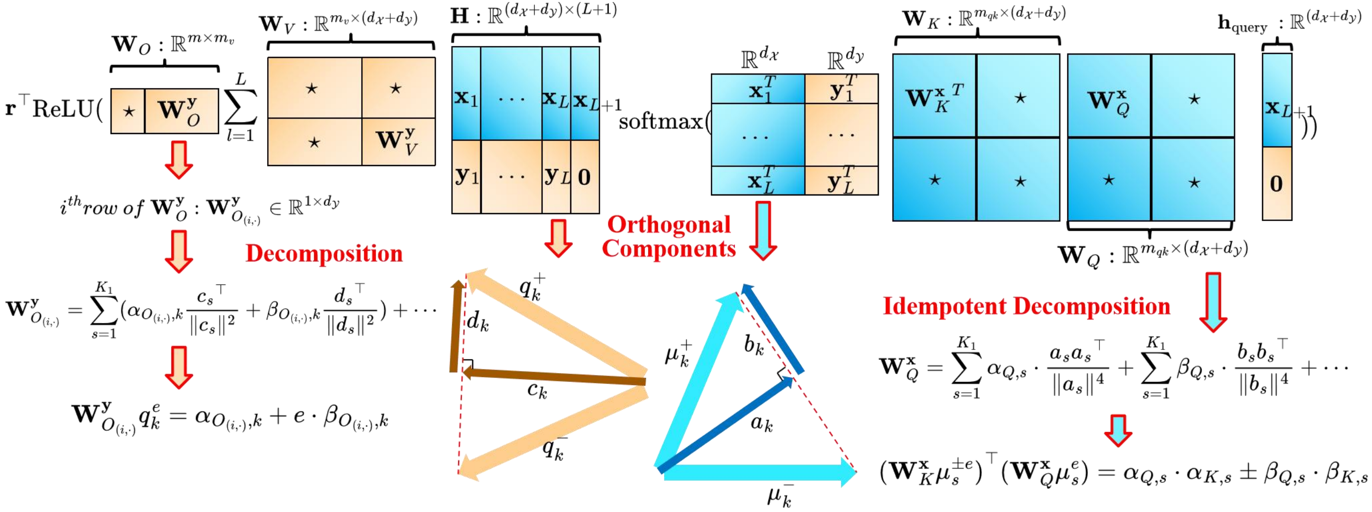
respectively. $\|\boldsymbol{b}_k^*\| \geq \|\boldsymbol{a}_k^*\| = \Theta(\|\mathbf{u}\|)$, $\|\boldsymbol{d}_k^*\| \geq \|\boldsymbol{c}_k^*\| = \Theta(\|\mathbf{q}\|)$ and $\boldsymbol{\nu}_{k_2}^* = \Theta(\|\mathbf{u}\|)$

Main Result

➢ **Proof Strategy: Convergence of Expectation - Exponential Variance Reduction [1]**

In a big picture, we **extend the standard techniques in SGD** [1] to our model under **strong low-noise condition**

(i) The expected estimator would fastly converge; (ii) The variance can connverge exponentially by the property of tails



With a good initialization and a symmetric low-noise prompt distribution, we can decompose the expected (over the stochastic batches) NN matrices along concept and semantic directions.

[1] Nitanda and Suzuki. Stochastic gradient descentwith exponential convergence rates of expected classification errors. In AISTATS, 2019.

Contents

Experiments

In-Distribution Test Distribution.

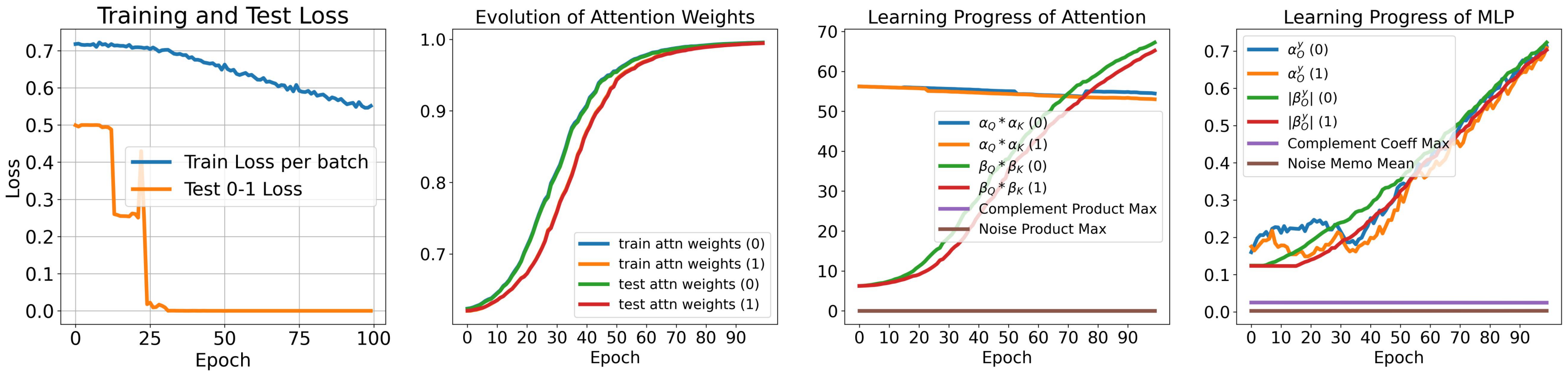
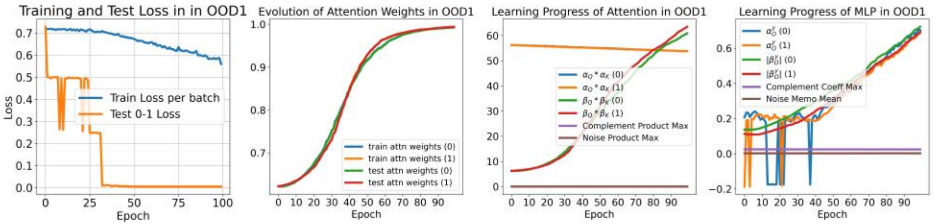


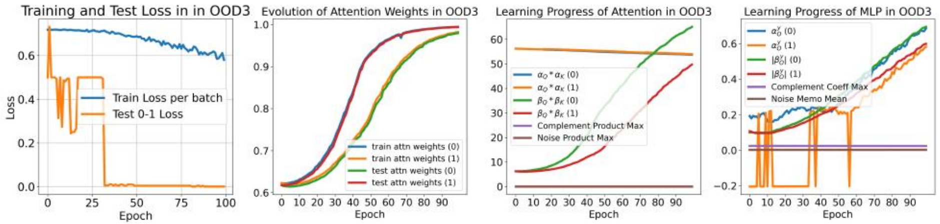
Figure 2: Learning dynamics: (i) training and test loss; (ii) correct attention weight; (iii) maximum values of $\alpha_{Q,s} \cdot \alpha_{K,s}$, $\beta_{Q,s} \cdot \beta_{K,s}$, maximum values of the complement products $\tau_{Q,r} \cdot \tau_{K,r}$ or $\rho_{Q,2} \cdot \rho_{K,2}$, and maximum values of product-with-noise $(\mathbf{W}_K^{\boldsymbol{x}} \xi_{\boldsymbol{x}})^\top \mathbf{W}_Q^{\boldsymbol{x}} \xi_{\boldsymbol{x}}$; (iv) maximum values of $\alpha_{O_{(i,\cdot)},k}$ and $|\beta_{O_{(i,\cdot)},k}|$, maximum values of the complement coefficients $\rho_{O_{(i,\cdot)},w}$ and maximum values of product-with-noise $\mathbf{W}_{O_{(i,\cdot)}}^{\boldsymbol{y}} \xi_{\boldsymbol{y}}$. The parameter settings are: $L = 4$, $K_1 = 2$, $K = 104$, $n_{\text{test}} = 5000$, $d_{\mathcal{X}} = d_{\mathcal{Y}} = 1000$, $m = 50$, $\|\mathbf{u}\| = \|\mathbf{q}\| = 10$, $\forall k \in [K_1]$, $\langle \boldsymbol{\mu}_k^+, \boldsymbol{\mu}_k^- \rangle / \|\mathbf{u}\|^2 = \langle \boldsymbol{q}_k^+, \boldsymbol{q}_k^- \rangle / \|\mathbf{q}\|^2 = 0.5$, $\sigma_0 = 0.1$, $\sigma_1 = 0.01$, $\sigma_\xi = 0.01$, $\lambda = 0.002$, $B = 16$, $\gamma = 10000$, and the total training epochs is 100.

Experiments

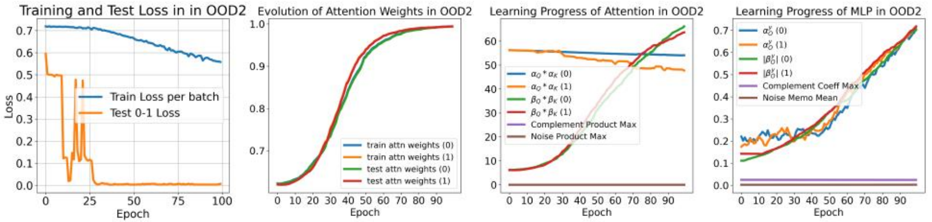
OOD Test Distribution.



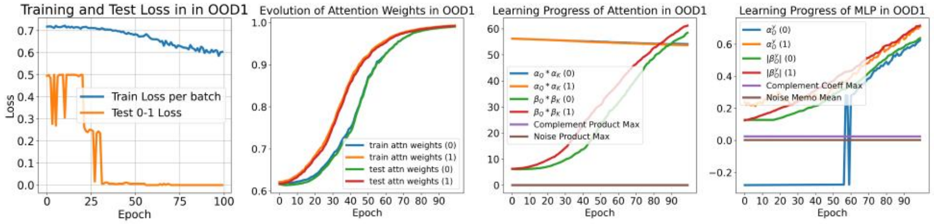
(a) OOD Scenario 1(i): $L^* = 5$ during testing.



(b) OOD Scenario 1(ii): $L^* = 2$ during testing.



(c) OOD Scenario 2: 0.8 fraction for the first and 0.2 fraction for the second concept during testing.



(d) OOD Scenario 3: Shift the data as $\boldsymbol{\mu}_1^{\pm *} = \boldsymbol{a}_1 \pm \boldsymbol{b}_2$ and $\boldsymbol{\mu}_2^{\pm *} = \boldsymbol{a}_2 \pm \boldsymbol{b}_1$ during testing.

Figure 3: Learning dynamic in three OOD scenarios. The training settings and plotting methods are identical to those used in Figure 2. The consistency of the results validates Proposition 1.

Contents

Conclusion

➢ Advancing the Theory of Transformers and ICL.

We provide a fine-grained analysis of the learning dynamics for a three-layer transformer model, comprising an **softmax** attention followed by a **ReLU**-activated feedforward network. We showcase the asymptotic properties governing the coupled learning of the attention and MLP layers.

➢ Exponential Convergence of 0-1 Loss.

Despite the highly non-convex nature of the problem, we are the first to prove an exponential convergence rate for the 0-1 loss utilizing techniques in stochastic optimization literature.

➢ Connecting Multi-Concept Semantics to Efficient ICL.

We provably show how the multi-concept encoded linear geometry of representations can enable transformer to conduct certain OOD ICL tasks.

Thanks for Listening