Nonlinear transformers can perform inference-time feature learning

Naoki Nishikawa, Yujin Song, Kazusato Oko, Denny Wu, Taiji Suzuki

Learning Single-Index Models In-Context

In-Context Learning (ICL): Transformers can construct various regression algorithms at inference time.

$$\textbf{Prompt}: \underbrace{\boldsymbol{x}_1 \to \boldsymbol{y}_1, \ldots, \boldsymbol{x}_N \to \boldsymbol{y}_N}_{\text{in-context examples}}, \underbrace{\boldsymbol{x} \to ?}_{\text{query}}, \quad \text{where } y_i = f_*(\boldsymbol{x}_i) + (\text{noise})$$

- Models need to estimate f_* without any parameter updates.
  - Limited statistical and optimization guarantees, especially for nonlinear targets

Can transformers leverage **statistically efficient** mechanisms like feature learning, even at inference time?

Problem Setting: Single-Index Target Functions

$$f_*(\boldsymbol{x}_i) = \sigma_*(\langle \boldsymbol{x}_i, \boldsymbol{\beta} \rangle), \ \boldsymbol{\beta} \in \mathbb{R}^d; \quad \boldsymbol{x}_i \sim \mathcal{N}(0, \boldsymbol{I}_d).$$

$$\sigma_*(z) = \sum_{j=\mathrm{ie}(\sigma_*)}^{\deg(\sigma_*)} \alpha_j^* \mathrm{He}_j(z) \ (\mathrm{He}_j : j\text{-th } \textit{Hermite polynomial}).$$

$\to$ degree $\deg(\sigma_*)$ and information exponent $\mathrm{ie}(\sigma_*)$ ($\mathrm{ie} \le \deg$)

- f_* changes only along the feature vector $\boldsymbol{\beta}$: A **statistical testbed for feature learning**

Sample complexity lower bound for regression on test prompt
Kernel methods: $N \gtrsim d^{\Theta(\deg(\sigma_*))}$ (feature learning not available) [GMMM21,DWY21]
CSQ algorithms: $N \gtrsim d^{\Theta(\mathrm{ie}(\sigma_*))}$ (e.g. one-step GD on NN [DLS22])
SQ algorithms: $N \gtrsim d^{\Theta(\mathrm{ge}(\sigma_*))}$ (e.g. batch-reuse SGD on NN [ADK+24,LOSW24])

- (S)GD on NNs enables feature learning: Advantage over non-adaptive kernel methods
- **Generative exponent** $\mathrm{ge}(\sigma_*) := \min_{h \in L^2} \mathrm{ie}(h \circ \sigma_*) \le \mathrm{ie}(\sigma_*)$: Nonlinear label transformations introduce non-CSQ terms, enabling further efficiency.

- **Prior work on ICL**: Pretrained transformers can implement **kernel methods** over test prompts, requiring $N \gtrsim r^{\Theta(\deg(\sigma_*))}$ in-context examples (assuming $\boldsymbol{\beta}$ is drawn from a rank-r linear subspace) [OSSW24].
  - Suboptimal inference-time sample complexity

Question 1: Can pretrained transformers **perform feature learning at inference time**, adapting to task-specific $\boldsymbol{\beta}$?

Question 2: How **statistically efficient** can transformers be at inference time? Can they surpass the barriers of kernel and CSQ?

Main Result: Feature Learning In-Context

Model: One-layer transformer (**softmax attention + nonlinear MLP**)

$$f_{\mathrm{TF}}(\boldsymbol{X}_{1:N}, \boldsymbol{y}_{1:N}, \boldsymbol{x}; \boldsymbol{W}^{KQ}, \boldsymbol{W}^{FV}, \boldsymbol{a}, \boldsymbol{b})$$
$$= \boldsymbol{a}^\top \mathrm{ReLU}(\boldsymbol{W}^{FV} \boldsymbol{E} \cdot \mathrm{Softmax}(\mathrm{Mask}(\rho^{-1} \cdot \boldsymbol{E}^\top \boldsymbol{W}^{KQ} \boldsymbol{E})) + \boldsymbol{b})$$
$$\text{where } \boldsymbol{E} = \begin{bmatrix} x_1 & \cdots & x_N & x \\ y_1 & \cdots & y_N & 1 \end{bmatrix}$$

Pretraining Algorithm: (under parametrization $\boldsymbol{W}^{KQ} = \begin{bmatrix} \Gamma & 0 \\ 0^\top & 1 \end{bmatrix}$, $\boldsymbol{W}^{FV} = [O \ v]$)

I. One gradient descent step on the attention matrix $\boldsymbol{W}^{KQ}$.
$$\Gamma \leftarrow \Gamma - \eta_1 \nabla_\Gamma \Big[(2T_1)^{-1} \textstyle\sum_{t=1}^{T_1} (f_{\mathrm{TF}}(\boldsymbol{X}^t_{1:N_1}, \boldsymbol{y}^t_{1:N_1}, \boldsymbol{x}^t) - y^t)^2 + \frac{\lambda_1}{2} \|\Gamma\|_F^2 \Big]$$

II. Ridge regression on the MLP parameters.
$$\boldsymbol{a} \leftarrow \arg\min_{\boldsymbol{a}} (2T_2)^{-1} \textstyle\sum_{t=T_1}^{T_1+T_2} (f_{\mathrm{TF}}(\boldsymbol{X}^t_{1:N_2}, \boldsymbol{y}^t_{1:N_2}, \boldsymbol{x}^t) - y^t)^2 + \frac{\lambda_2}{2} \|\boldsymbol{a}\|^2$$

Proposition (inference-time feature learning)

For a transformer (pretrained on $T \gtrsim d^{\Theta(\mathrm{ie}(\sigma_*))}$ tasks), the softmax attention output is approximated by $C_a + C_b \left(\frac{\langle \boldsymbol{x}, \boldsymbol{\beta} \rangle}{\sqrt{r}} \right)^{\mathrm{ge}(\sigma_*)}$, where $\boldsymbol{\beta}$ is randomly drawn from a rank-$r \le d$ linear subspace.

*the approximation holds without any parameter updates for each test prompt

Result 1: Softmax attention modules can identify the task-specific feature vector $\boldsymbol{\beta}$ (**inference-time feature learning**)

Theorem (sample complexity bound)

The pretrained transformer achieves $o_d(1)$ estimation error at inference time for f_*, provided that the number of in-context examples N satisfies $N_{\mathrm{test}} \gtrsim r^{\Theta(\mathrm{ge}(\sigma_*))}$.

Result 2: Nonlinear transformers can **surpass the statistical barriers** of both kernel methods and CSQ algorithms

Sample Complexity Rates	
[OSSW24]	This work
Pretraining: $d^{\Theta(\mathrm{ie}(\sigma_*))}$	*Pretraining:* $d^{\Theta(\mathrm{ie}(\sigma_*))}$
Inference: $r^{\Theta(\deg(\sigma_*))}$	*Inference:* $r^{\Theta(\mathrm{ge}(\sigma_*))}$

- End-to-end statistical and optimization guarantees
- Sample complexity depends only on the inner dimension $r \to$ **adaptive to low-dimensional prior** [OSSW24].

Technical Overview

Fact: Correlational signal $\mathbb{E}[y\phi(\boldsymbol{x})]$ in CSQ gets weaker as $\mathrm{ie}(\sigma_*)$ increases...

Obs: Nonlinear label transformations in softmax **reduce the information exponent**.

- Model output is $f_{\mathrm{TF}} = \sum_{j=1}^m a_j \sigma(v_j g(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{x}; \boldsymbol{\Gamma}) + b_j)$ where the attention output g is

$$g(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{x}; \boldsymbol{\Gamma}) := \underbrace{\frac{N^{-1} \sum_{i=1}^N y_i \exp(y_i/\rho) \exp(\boldsymbol{x}_i^\top \boldsymbol{\Gamma} \boldsymbol{x}/\rho)}{N^{-1} \sum_{i=1}^N \exp(y_i/\rho) \exp(\boldsymbol{x}_i^\top \boldsymbol{\Gamma} \boldsymbol{x}/\rho)}}_{\text{Correlation between exponentially-transformed labels}}.$$

Proposition

The IE of (the clipped version of) $\sigma_* \exp(\sigma_*/\rho)$ and $\exp(\sigma_*/\rho)$ is equal to $\mathrm{ge}(\sigma_*)$.

Technical Challenges to derive $\mathrm{ge}(\sigma_*)$-dependent sample complexity

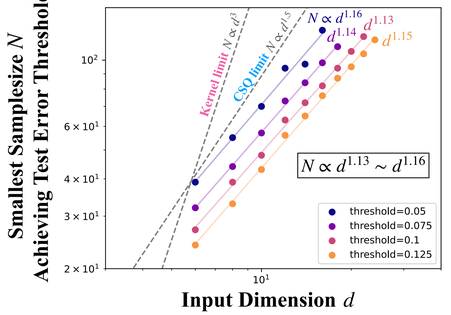
- ***Softmax correlations (fractions):*** Challenges in handling correlations between numerator and denominator
- ***Optimization guarantee:*** (One GD step on $\boldsymbol{\Gamma}$) $\simeq \mathbb{E}[\boldsymbol{\beta}\boldsymbol{\beta}^\top]$ (rank-r) $\to$ inference-time sample complexity $r^{\Theta(\mathrm{ge}(\sigma_*))}$ free from d [OSSW24].

Experiment: Inference-Time Feature Learning of GPT-2

Measuring the **inference-time sample complexity** for $f_*(\boldsymbol{x}) = \mathrm{He}_3(\langle \boldsymbol{\beta}, \boldsymbol{x} \rangle)$

Model: 6-layer GPT-2 with 4 attention heads, trained using the Adam optimizer

GPT-2 Inference-time Sample Complexity



Result: The GPT-2 model achieves inference-time sample complexity scaling, **surpassing both the kernel lower bound $\tilde{O}(d^3)$ and the CSQ lower bound $\tilde{O}(d^{1.5})$.**