In-Context Learning by Linear Attention

Yue M. Lu [1] Mary Letey [1] Jacob Zavatone-Veth [1] Anindita Maiti [2] Cengiz Pehlevan [1]

[1]Harvard University [2]Perimeter Institute

In-Context Learning

Neural networks exhibit ability to **execute and learn tasks based only on examples seen in input**, without needing explicit training.

	Hello	nuqneH
	Help	QaH
	Thank you	tlho'
	How much	ar
	Warp	vlH
	Federation	Dlvl'
query token	Earth	??? (tera)

input-output examples

test label

Figure 1. ICL translation example [1]

- **When** does such an ability emerge?
- What **algorithm** is learned ICL for solving a task?
- What properties of **data** affect ICL in transformers?

A Toy Model

Study **linear features**, namely

$$\text{context} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_\ell, y_\ell)\}$$
$$\text{for } y_i = \boldsymbol{w}^\top \boldsymbol{x}_i + \epsilon_i$$

with

- **tokens** $\boldsymbol{x}_i \in \mathbb{R}^d$
- **label noise** $\epsilon_i \in \mathbb{R}$
- context-dependent **task vectors** $\boldsymbol{w} \in \mathbb{R}$

To prepare it for the attention model, embed each context as

$$Z = \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_\ell & \boldsymbol{x}_{\ell+1} \\ y_1 & y_2 & \cdots & y_\ell & 0 \end{bmatrix} \in \mathbb{R}^{(d+1)\times(\ell+1)}$$

where this structure includes a **query token** $\boldsymbol{x}_{\ell+1}$, and the 0-entry is a placeholder for the corresponding **answer label** $y_{\ell+1}$ [2, 3].

We consider a **linear attention mechanism** [4]

$$A(Z) = Z + (VZ)(KZ)^\top (QZ)/\ell\,.$$

We show this simplifies to a predictor

$$\hat{y}_{\ell+1} \equiv A(Z)_{\text{bottom right}} = \langle \Gamma, H_Z \rangle$$

for

$$\textbf{parameters } \Gamma \in \mathbb{R}^{d\times(d+1)}$$
$$\textbf{features } H_Z = \boldsymbol{x}_{\ell+1} \left[\tfrac{d}{\ell} \textstyle\sum_{i=1}^{\ell} y_i \boldsymbol{x}_i^\top \quad \tfrac{1}{\ell} \textstyle\sum_{i=1}^{\ell} y_i^2 \right].$$

Data Model

We will pretrain the linear transform on n **different contexts** of length ℓ. Each context has a corresponding task $\boldsymbol{w}^\mu$ for $\mu = 1, \ldots, n$.

We choose **Gaussian** tokens and noise

$$\boldsymbol{x}_i \sim \mathcal{N}(0, \mathbb{I}_d/d), \qquad \epsilon_i \sim \mathcal{N}(0, \rho)$$

Task Structure We limit the **task diversity** within the contexts by sampling each $\boldsymbol{w}^\mu$ from a finite set of k possible tasks $\{\boldsymbol{w}_1, \cdots, \boldsymbol{w}_k\}$ uniformly. Each of these k tasks is **gaussian**

$$\boldsymbol{w}_j \sim \mathcal{N}(0, \mathbb{I}_d) \qquad \text{for } j = 1, \ldots, k.$$

Evaluation

We study **two** different testing regimes.

1. **ICL test.** Generate tokens and noise as before. Sample a *fresh* task from the true task distribution $\boldsymbol{w}_{\text{test}} \sim \mathcal{N}(0, \mathbb{I}_d)$.
2. **IDG (In-Distribution Generalization) test.** Sample $\boldsymbol{w}_{\text{test}}$ uniformly from the training pool $\{\boldsymbol{w}_1, \cdots, \boldsymbol{w}_k\}$.

Key Parameters and Joint Scaling

The model parameters are

token/task **dimension** d context **length** ℓ
number of contexts n task **diversity** k

We introduce a scaling limit with rich behaviour given by

$$\alpha \equiv \ell/d\,, \quad \kappa \equiv k/d\,, \quad \tau \equiv n/d^2$$

We will solve the model in an asymptotic limit $d, \ell, n, k \to \infty$ holding $\alpha, \kappa, \tau = \mathcal{O}(1)$.

Methodology

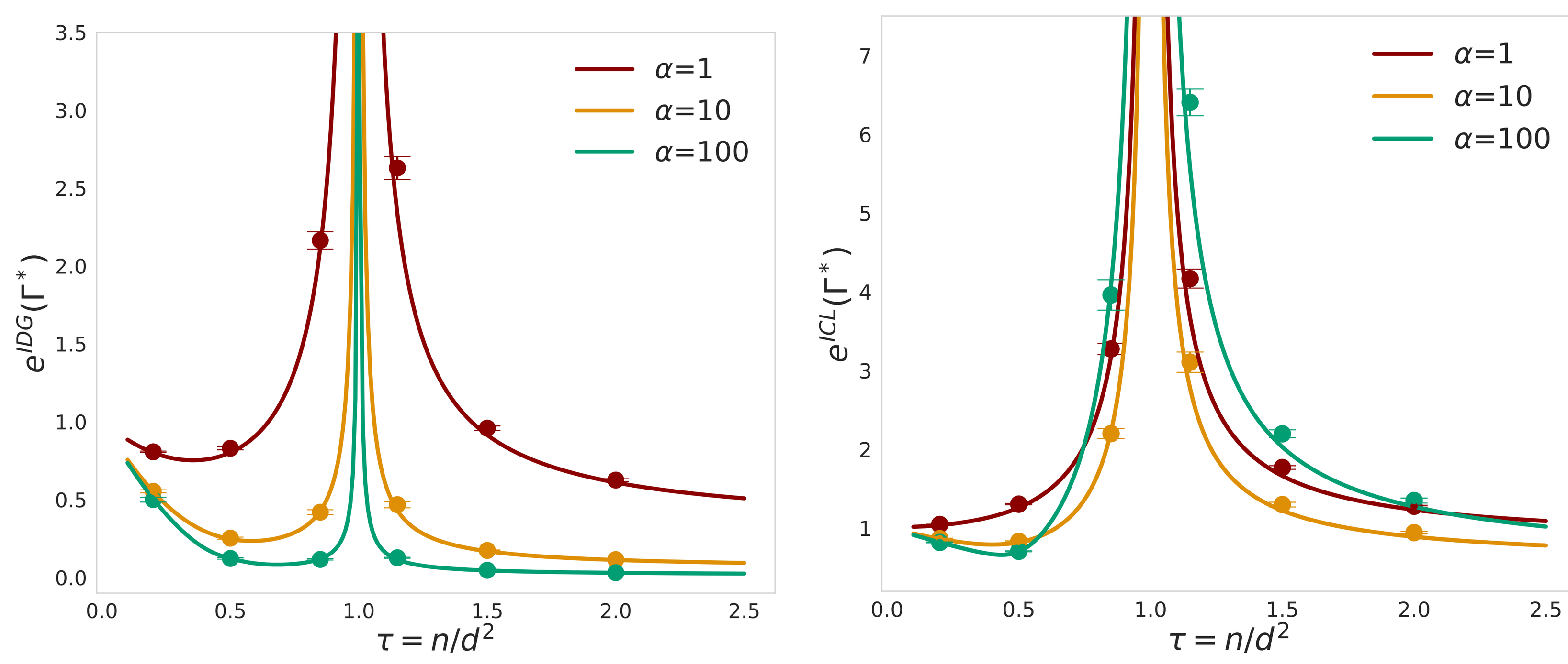
Under a square loss, we can analytically solve for the optimal paramter matrix

$$\text{vec}(\Gamma^*) = \frac{\sum_{\mu=1}^{n} y_{\ell+1}^{\mu} \text{vec}(H_{Z^\mu})}{\left(\frac{n}{d}\lambda I + \sum_{\mu=1}^{n} \text{vec}(H_{Z^\mu})\,\text{vec}(H_{Z^\mu})^\top\right)}$$

Using **random matrix theory** we can find a deterministic equivalent for Γ^* which we use to find exact ICL and IDG error curves.

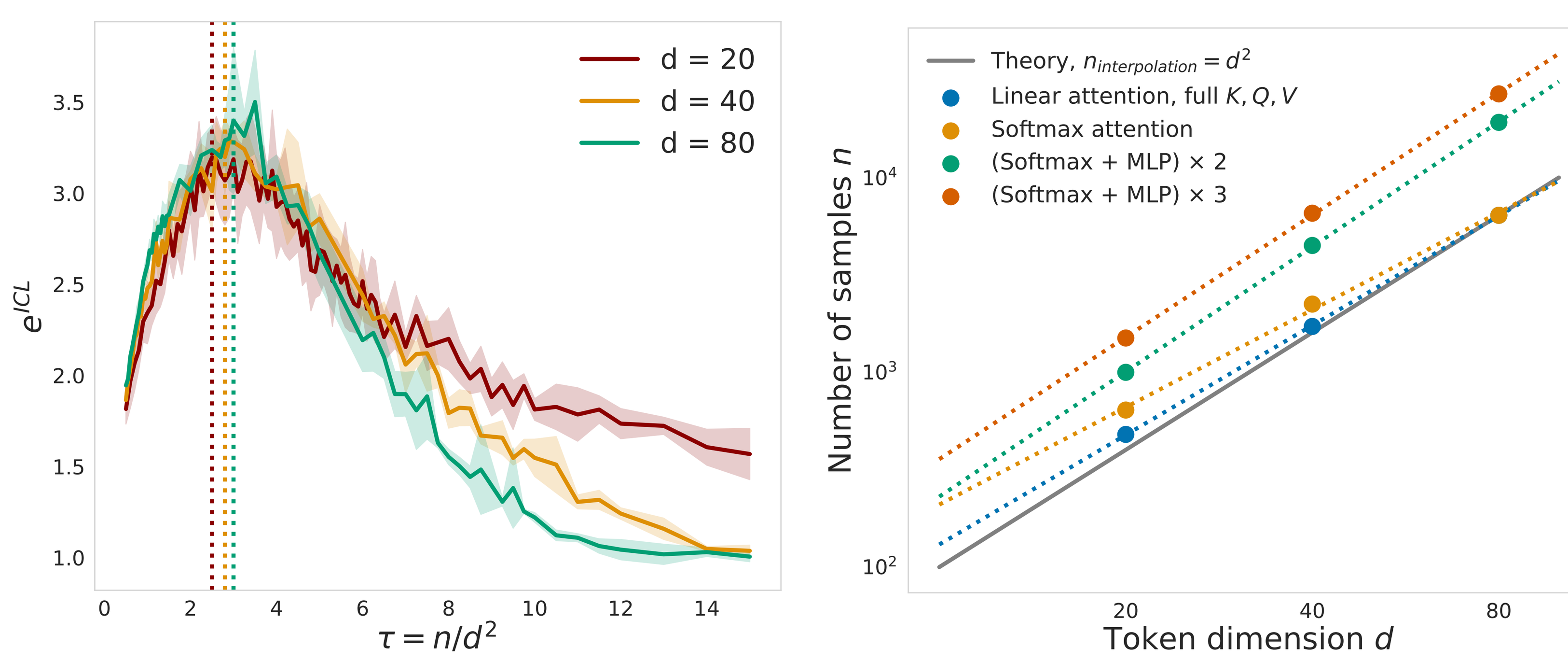
We present implications of these error curves.

Sample-wise Double Descent



(a) Ridgeless IDG error against τ (b) Ridgeless ICL error against τ

Figure 2. Double descent in $\tau = n/d^2$ for linear transformer



(a) Nonlinear model double descent (b) Interpolation threshold scaling

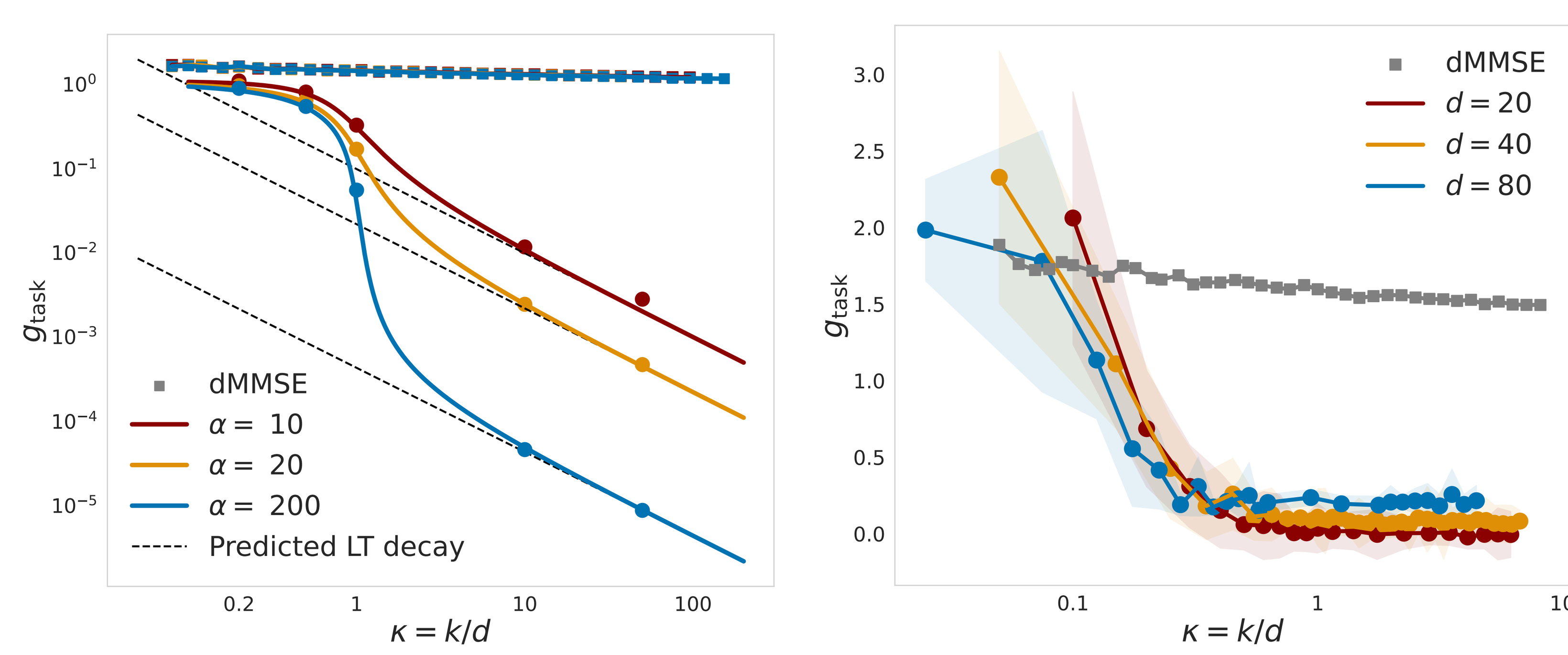
Figure 3. Verification of (4a) double descent and (4b) $n \sim d^2$ scaling in nonlinear models

Learning Transition in Task Diversity

When is a model actually learning in-context, i.e. solving a new regression problem by adapting to the specific structure of the task, rather than memorizing training task vectors? We refer to this as **task generalization** and model it as $g_{\text{task}} = e_{\text{ICL}} - e_{\text{IDG}}$.

- g_{task} large: model memorizes training tasks, has not learned the true task distribution.
- g_{task} small: the model is leveraging the underlying structure to generalize in task rather than memorize.

We compare g_{task} for the linear transformer against a **memorization prior** [5] called dMMSE.



(a) Linear transformer (b) Nonlinear architecture

Figure 4. Plot of task generalisation g_{task} against task diversity κ showing learning transition

References

1. Brown, T. B. *et al. Language Models are Few-Shot Learners.* 2020.
2. Zhang, R. *et al. Trained Transformers Learn Linear Models In-Context.* 2023.
3. Wu, J. *et al. How Many Pretraining Tasks Are Needed for In-Context Learning of Linear Regression?.* 2024.
4. Wang, S. *et al. Linformer: Self-Attention with Linear Complexity.* 2020.
5. Raventós, A. *et al. Pretraining task diversity and the emergence of non-Bayesian in-context learning for regression.* 2023.