基于等效模型的随机 TRANSFORMER 上下文学习渐近研究

Samet Demir, Zafer Doğan

MLIP Research Group, KUIS AI Center & Department of EEE, Koç University
İstanbul, Turkey

ABSTRACT

我们研究了预训练的 Transformer 在非线性回归设置中的上下文学习（ICL）能力。具体来说，我们关注一个具有非线性 MLP 头部的随机 Transformer，在其中第一层是随机初始化并固定的，而第二层则是经过训练的。此外，我们在一种极限情况下进行考虑，即上下文长度、输入维度、隐藏维度、训练任务的数量和训练样本的数量共同增长。在这种设置下，我们展示了随机 Transformer 在 ICL 误差方面等同于一个有限阶 Hermite 多项式模型的行为。这种等价性通过不同激活函数、上下文长度、隐藏层宽度（揭示了双重下降现象）以及正则化设置下的模拟得到了验证。我们的结果提供了理论和经验上的洞察，说明 MLP 层在何时何地增强 ICL，以及非线性和过参数化如何影响模型性能。

Index Terms— 上下文学习、变换器、深度学习理论、高维渐近性。

1. 介绍

变换器 [1] 已成为近年来深度学习的基石。它们新兴的一项能力是所谓的上下文内学习（ICL），这使得任务能够通过提示进行适应，而无需修改内部参数 [2]。鉴于变换器在 ICL 方面的受欢迎程度和有效性，对其 ICL 能力的理论理解已经成为一个重要的研究课题。

由于全面分析 ICL 的复杂性，现有研究通常考虑简化的设置。特别是，许多工作专注于使用仅包含注意力机制架构的 Transformer 模型在线性回归或分类任务中的 ICL，完全省略了多层感知器 (MLPs)[3, 4, 5, 6]。尽管最近的研究开始探索基于 Transformer 的 ICL 中非线性 MLP 的作用 [7, 8, 9]，但仍存在一些限制。例如，[7] 特别关注使用 ReLU 基础 MLP 的分类任务，没有考虑更广泛的激活函数或回归任务。同时，[8, 9] 分析了在 MLP 层位于注意力机制之前的 Transformer 变体——这与原始 Transformer 架构 [1] 相反，在后者中，输入嵌入首先通过注意力块，然后再由 MLPs 处理。因此，对于标准 Transformer 架构中 MLPs 如何影响 ICL 性能的全面理解仍然不完整。

为了解决这一差距，我们通过渐近分析研究了带有非线性 MLP 的 Transformers 在非线性回归任务中的 ICL 性能。我们的方法连接了两个研究领域：Transformer 的 ICL 和 MLP 的渐近理论。对于后者，我们引用了高斯通用性结果 [10, 11, 12]，这些结果显示某些类别的 MLP 的行为等同于具有匹配矩的高斯模型。这种等价性使我们能够描述非线性 MLP 对 Transformers ICL 性能的影响。

具体而言，我们考虑一个具有线性注意力机制和非线性 MLP 的随机 Transformer 模型，在该模型中，MLP 的第一层权重被随机初始化并固定，而第二层则完全训练。在这种设置下，我们展示了 Transformer 在上下文学习误差方面与有限阶多项式模型渐近等价。我们还确定了非线性 MLP 的存在如何改善无 MLPs 的 Transformer 的 ICL 性能的条件。具体来说，我们

We acknowledge that this work is supported partially by TÜBİTAK under project 124E063 in ARDEB 1001 program. S.D. is supported by an AI Fellowship provided by KUIS AI Research Center and a PhD Scholarship (BİDEB 2211) from TÜBİTAK. The corresponding author is Zafer Doğan (zdogan@ku.edu.tr).

979-8-3503-2411-2/25/$31.00 ©2025 IEEE

的研究结果表明，在满足以下条件时，MLP 可以增强 ICL：(i) 激活函数选择得当，(ii) 上下文长度足够大，以及 (iii) 隐藏维度适当选择或模型得到恰当的正则化。此外，我们的分析揭示了具有非线性 MLP 的 Transformer 的 ICL 误差可能表现出非单调行为——通常被称为“双下降现象”——作为模型复杂性的函数。这扩展了先前在 MLPs [13, 14] 中观察到的双下降现象至 ICL 设置。

总体而言，我们的贡献如下：

1. 我们发现一个多项式模型在指令调用学习 (ICL) 中表现与带有 MLP 的随机 Transformer 等效。
2. 我们刻画了在何种条件下，带有非线性 MLPs 的 Transformer 优于不带 MLPs 的 Transformer。
3. 我们证明了通过适当的正则化可以缓解 Transformer 与 MLP 中 ICL 误差出现的非单调行为。

2. 相关工作：带有变换器的 ICL

开创性的工作 [2] 首次强调了 Transformer 在上下文学习中的能力，激发了大量的实证和理论研究。在实证方面，诸如 [15]、[16] 和 [17] 等研究表明，随着模型规模的增大，ICL（上下文学习）的能力会更加显著地显现出来，突出了其在大规模 AI 系统中的重要性。为了更好地理解 ICL 背后的机制，研究人员采用了受控的合成基准测试，最著名的是使用 Transformers 进行线性回归任务，如 [6]、[18] 和 [19] 所示。这些基准测试通过隔离特定的架构和数据驱动效应，使得精确分析成为可能。

理论上，许多研究表明，在预训练过程中，Transformer 隐式地获取了算法结构，这些结构随后在 ICL[20, 21, 3, 22, 4, 23, 24, 6, 25, 26] 中被应用。然而，这些学习到的算法的确切性质仍然是一个开放的问题。为了获得可处理性，许多研究分析了简化的 Transformer 模型——特别是那些具有线性化自注意力机制的模型。特别地，[6] 和 [27, 28] 为基于线性注意力机制的 Transformer 在合成 ICL 任务中的泛化能力提供了详细分析。

最近，将非线性 MLP 组件纳入 Transformer 架构引起了关注。例如，[7] 在分类设置中使用基于 ReLU 的 MLP 研究了 ICL，而 [8] 和 [9] 则研究了其中 MLP 层位于注意力机制之前的模型。然而，这些工作要么限制了激活函数和任务类型的选取，要么偏离了标准 Transformer 架构 [1]，在该架构中，注意力层位于 MLP 之前。

因此，关于在 Transformer 中具有非线性 MLP 头的 ICL 的全面理论处理——遵循原始架构并适用于回归等一般任务设置——仍然缺乏文献支持。这是我们的工作旨在解决的问题。

3. 设置

在这项工作中，我们研究了预训练的 Transformer 架构在非线性回归问题上的 ICL 能力。
具体来说，给定一系列配对示例

$$(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_\ell, y_\ell), (\boldsymbol{x}_{\ell+1}, ?),$$

其中每个输入 $\boldsymbol{x}_i \in \mathbb{R}^d$ 和对应的标量响应 $y_i \in \mathbb{R}$ 独立地从一个未知的联合分布中抽取，模型必须推断出前 ℓ 个示例的基本映射，并预测新输入 $\boldsymbol{x}_{\ell+1}$ 的标签 $y_{\ell+1}$。这里，ℓ 表示上下文长度。

我们假设 $\boldsymbol{x}$ 和 y 之间存在由特定上下文参数向量 $\boldsymbol{\xi} \in \mathbb{R}^d$ 治理的非线性关系，受加性高斯噪声影响。尽管在给定的上下文中 $\boldsymbol{\xi}$ 保持不变，但它在不同的上下文间重新采样，这就要求模型从观察到的成对数据中估计出 $\boldsymbol{\xi}$，然后再推广到新的输入。形式上，我们根据

$$\boldsymbol{x}_i \sim \mathcal{N}\big(\mathbf{0}, \boldsymbol{I}_d/d\big), \quad y_i = \sigma_*\big(\boldsymbol{\xi}^T \boldsymbol{x}_i\big) + \epsilon_i, \qquad (1)$$

生成数据，其中 $\sigma_* : \mathbb{R} \to \mathbb{R}$ 是一个非线性函数，并且 $\epsilon_i \sim \mathcal{N}(0, \rho)$ 对于 $\rho > 0$ 成立，而任务向量本身被绘制为

$$\boldsymbol{\xi} \sim \mathcal{N}\big(\mathbf{0}, \boldsymbol{I}_d\big). \qquad (2)$$

这种概率表示捕捉了回归任务中的非线性和固有变异性。然后我们考虑使用 k 个随机任务向量 $\boldsymbol{\xi}$ 来生成训练样本，而在上下文学习误差则通过所有可能的任务向量的期望值进行测量。

为了将这些示例输入到 Transformer 中，我们遵循文献 [6] 并通过在其标量标签上方堆叠特征向量（最

终标签处用零占位符）形成一个嵌入矩阵 $\boldsymbol{Z}$，即

$$\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_\ell & \boldsymbol{x}_{\ell+1} \\ y_1 & y_2 & \cdots & y_\ell & 0 \end{bmatrix} \in \mathbb{R}^{(d+1)\times(\ell+1)}. \tag{3}$$

Transformer 必须利用这种表示来推断底层的非线性映射，并准确预测 $y_{\ell+1}$。然后，Transformer 中线性注意力的输出可以按照以下方式计算：

$$\boldsymbol{A} := \boldsymbol{Z} + \frac{1}{\ell}\boldsymbol{V}\boldsymbol{Z}(\boldsymbol{K}\boldsymbol{Z})^T(\boldsymbol{Q}\boldsymbol{Z}), \tag{4}$$

其中 $\boldsymbol{K}, \boldsymbol{Q}, \boldsymbol{V}$ 分别是适当大小的关键矩阵、查询矩阵和值矩阵。当预测 $y_{\ell+1}$ 时，相关的注意力输出是 $A_{d+1,\ell+1}$。因此，我们考虑线性 Transformer（无 MLPs）的预测为

$$\hat{y}_{linear} := A_{d+1,\ell+1}. \tag{5}$$

此预测可以使用由 [27, 28] 所用的重新参数化技术简化为以下形式：

$$\hat{y}_{linear} = \mathrm{vec}(\boldsymbol{\Gamma})^T \mathrm{vec}(\boldsymbol{H}_{\boldsymbol{Z}}), \tag{6}$$

其中 $\mathrm{vec}(\cdot)$ 表示向量化操作，$\boldsymbol{\Gamma} \in \mathbb{R}^{d\times(d+1)}$ 是使用 $\boldsymbol{V}, \boldsymbol{K}, \boldsymbol{Q}$ 矩阵的元素构成的参数矩阵，而 $\boldsymbol{H}_{\boldsymbol{Z}}$ 定义为

$$\boldsymbol{H}_{\boldsymbol{Z}} := \boldsymbol{x}_{\ell+1}\begin{bmatrix} \frac{d}{\ell}\sum_{i\le\ell} y_i \boldsymbol{x}_i^T & \frac{1}{\ell}\sum_{i\le\ell} y_i^2 \end{bmatrix} \in \mathbb{R}^{d\times(d+1)}.$$

请注意，线性 Transformer 的情况下的训练 $\boldsymbol{\Gamma}$ 为

$$\arg\min_{\boldsymbol{\Gamma}} \sum_{j=1}^{n} \left(y_{\ell+1}^j - \mathrm{vec}(\boldsymbol{\Gamma})^T \mathrm{vec}(\boldsymbol{H}_{\boldsymbol{Z}^j})\right)^2 + \lambda \frac{n}{d}\|\boldsymbol{\Gamma}\|_F^2,$$

其中 λ 是正则化常数，与 λ 相邻的 n/d 因子用于在我们考虑的渐近环境中保持正则化的意义。这里，$\{(\boldsymbol{Z}^j, y_{\ell+1}^j)\}_{j=1}^n$ 表示用于训练的 n 样本，其中 $\boldsymbol{Z}^j$ 由 (3) 构成。

对于带有非线性 MLP 的 Transformer，模型的预测可以类似地写为

$$\hat{y}_{nonlinear} := \boldsymbol{w}^T \sigma(\boldsymbol{F}^T \mathrm{vec}(\boldsymbol{H}_{\boldsymbol{Z}})), \tag{7}$$

其中 $\sigma : \mathbb{R} \to \mathbb{R}$ 是非线性 MLP 的激活函数。受到随机特征模型 [29, 10] 的启发，我们考虑随机初始化并固定的 $\boldsymbol{F} \in \mathbb{R}^{d(d+1)\times m}$，而类似于上述训练过程，参数向量 $\boldsymbol{w} \in \mathbb{R}^m$ 通过

$$\arg\min_{\boldsymbol{w}} \sum_{j=1}^{n} \left(y_{\ell+1}^j - \boldsymbol{w}^T \sigma(\boldsymbol{F}^T \mathrm{vec}(\boldsymbol{H}_{\boldsymbol{Z}^j}))\right)^2 + \lambda \frac{n}{d}\|\boldsymbol{w}\|_2^2.$$

进行训练。

然后，我们用

$$\mathbb{E}_{(\boldsymbol{\xi}, \boldsymbol{Z}, y_{\ell+1})}\left[(y_{\ell+1} - \hat{y})^2\right], \tag{8}$$

来测量 ICL 误差，其中 $\hat{y}$ 指的是线性 Transformer 的 (6)，或者表示具有非线性 MLP 的 Transformer 的 (7)。

4. 主要结果

在这项工作中，我们刻画了带有非线性 MLP 的 Transformer 在渐近区域中的 ICL 误差（8）。具体来说，我们假设输入维度 d、训练样本数量 n、用于训练的任务向量的数量 k、上下文长度 ℓ 以及隐藏层维度 m 共同发散，而 $\ell/d, k/d, n/d^2, m/n$ 保持常数。这里，前三种缩放已由 [27, 28] 对于使用线性 Transformer 的 ICL 进行了识别，而最后一种缩放确保了模型复杂度和训练样本是可比较的，这是通常在文献中假设的情况 [10]。

我们从以下引理中随机特征映射 $\boldsymbol{F}^T \mathrm{vec}(\boldsymbol{H}_{\boldsymbol{Z}})$ 的等价统计表示开始分析。稍后我们将利用它来分析带有非线性 MLP（7）的 Transformer。

引理 1 ($\boldsymbol{F}^T \mathrm{vec}(\boldsymbol{H}_{\boldsymbol{Z}})$ 的渐近分布).
假设 $\boldsymbol{F}$ 的元素独立同分布为 $\mathcal{N}(0, 1/\mathrm{tr}(\mathrm{Cov}(\mathrm{vec}(\boldsymbol{H}_{\boldsymbol{Z}}))))$，使得 $\boldsymbol{F}^T \mathrm{vec}(\boldsymbol{H}_{\boldsymbol{Z}})$ 的元素具有单位方差，其中 $\mathrm{tr}(.)$ 和 $\mathrm{Cov}(.)$ 分别表示迹算子和协方差算子。令 $\boldsymbol{f}_i$ 表示 i 列的 $\boldsymbol{F}$。类似于 $l, d \to \infty$ 中的 $l/d \in \mathbb{R}^+$，我们有

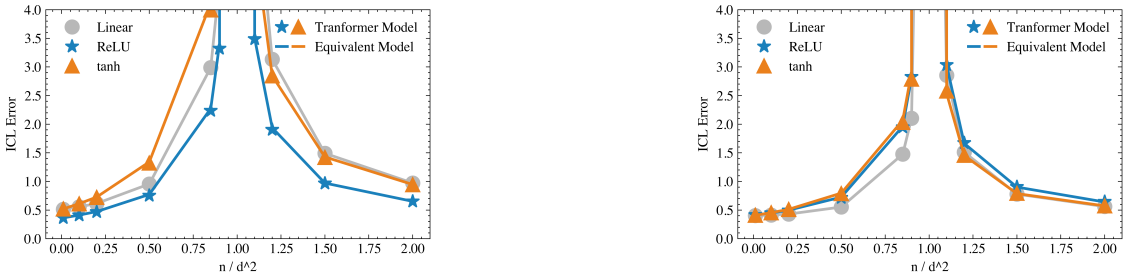
$$\boldsymbol{f}_i^T \mathrm{vec}(\boldsymbol{H}_{\boldsymbol{Z}}) \to \mathcal{N}(0, 1) \text{ almost surely}, \tag{9}$$

对于所有 $i \in \{1, \ldots, m\}$。

证明. 令 $t := \mathrm{tr}(\mathrm{Cov}(\mathrm{vec}(\boldsymbol{H}_{\boldsymbol{Z}})))$. 对于给定的 $\boldsymbol{H}_{\boldsymbol{Z}}$，$\boldsymbol{f}_i^T \mathrm{vec}(\boldsymbol{H}_{\boldsymbol{Z}})$ 的条件分布是

$$\boldsymbol{f}_i^T \mathrm{vec}(\boldsymbol{H}_{\boldsymbol{Z}}) \mid \boldsymbol{H}_{\boldsymbol{Z}} \sim \mathcal{N}(0, \|\mathrm{vec}(\boldsymbol{H}_{\boldsymbol{Z}})\|_2^2/t). \tag{10}$$

然后，通过 $\|\mathrm{vec}(\boldsymbol{H}_{\boldsymbol{Z}})\|_2^2/t$ 在 1 附近的集中性，我们得到了引理的陈述。 □



(a) $\sigma_* = ReLU$ (b) $\sigma_* = tanh$

Fig. 1: 上下文学习误差与训练样本数量的关系：线性注意力模型、Transformer 和等效模型。使用了两种不同的激活函数的 Transformer 模型（用形状和颜色表示）：ReLU 和 tanh。绘制了 20 次蒙特卡洛运行的平均值。n 是样本数量，$d = 80, l = d, k = 0.5d, m = d^2, \rho = 0.01, \lambda = 10^{-8}$。

该引理使我们能够简化项 $\boldsymbol{F}^T \mathrm{vec}(\boldsymbol{H_Z})$，从而促进其在 Transformer 的上下文学习分析中的使用。因此，我们将重点转移到 $(\boldsymbol{F}^T \mathrm{vec}(\boldsymbol{H_Z}), \boldsymbol{\xi}^T \boldsymbol{x}_{\ell+1})$ 的联合分布，而不是 $(\boldsymbol{H_Z}, y_{\ell+1})$，因为 $y_{\ell+1} = \sigma_*(\boldsymbol{\xi}^T \boldsymbol{x}_{\ell+1}) + \epsilon_{\ell+1}$ 由 (1) 给出。以下证明了在较高维度的极限情况下，对 $(\boldsymbol{F}^T \mathrm{vec}(\boldsymbol{H_Z}), \boldsymbol{\xi}^T \boldsymbol{x}_{\ell+1})$ 这一对收敛于联合高斯随机向量。

推论 2 ($(\boldsymbol{F}^T \mathrm{vec}(\boldsymbol{H_Z}), \boldsymbol{\xi}^T \boldsymbol{x}_{\ell+1})$ 的联合分布). *假设任务向量 $\boldsymbol{\xi}$ 是给定的。在引理 1的假设下，当 $l, d \to \infty$ 与 $l/d \in \mathbb{R}^+$ 关联时，$(\boldsymbol{F}^T \mathrm{vec}(\boldsymbol{H_Z}), \boldsymbol{\xi}^T \boldsymbol{x}_{\ell+1})$ 成为具有零均值和某些协方差的联合高斯分布。这里，联合分布出现是因为 $\mathrm{vec}(\boldsymbol{H_Z})$ 的构造包括了 $\boldsymbol{x}_{\ell+1}$。*

引理 1 和推论 2 共同刻画了非线性 MLP 的输入 $\mathrm{vec}(\boldsymbol{H_Z})$ 及其标签 $y_{\ell+1}$ 的分布。在这些组件建立之后，我们现在可以利用前面的结果结合两层神经网络 [11, 12] 的现有渐近分析来研究定义于 (8) 中的配备非线性 MLP 的 Transformer 的上下文学习误差。以下定理确定了一个与这种 Transformer 架构在渐近等价意义上的模型，其中“渐近等价”被定义为两个模型在本工作中考虑的渐近极限下达到相同的 ICL 误差 (8)。

定理 3 (等价多项式模型). *考虑引理 1中的设置，其中 d, n, m, ℓ, k 与 $\ell/d, k/d, n/d^2, m/n \in \mathbb{R}^+$ 共同发散。进一步假设 σ 和 σ_* 是满足 $\mathbb{E}_{x\sim\mathcal{N}(0,1)}[\sigma(x)^2] < \infty$ 的 Lipschitz 函数，这确保了 Hermite 展开的存在。然后，具有非线性 MLP 的 Transformer 在 (7) 中渐近等价（在 ICL 误差方面）于以下模型：*

$$\boldsymbol{w}^T \hat{\sigma}_r(\boldsymbol{F}^T \mathrm{vec}(\boldsymbol{H_Z})). \tag{11}$$

这里，$\hat{\sigma}_r : \mathbb{R} \to \mathbb{R}$ 是一个带有残差项的 r 次多项式函数，定义为

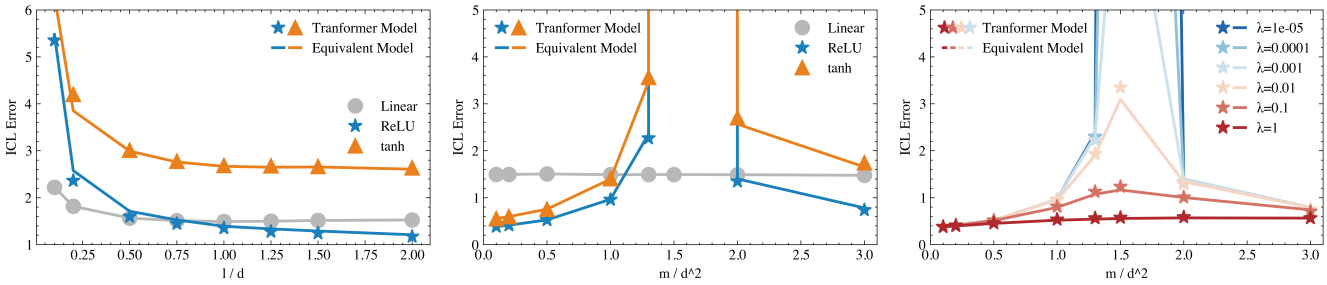
$$\hat{\sigma}_r(x) := \sum_{i=0}^{r} \frac{1}{i!} c_i He_i(x) + c_r^* z, \quad z \sim \mathcal{N}(0,1), \tag{12}$$

其中 $He_i : \mathbb{R} \to \mathbb{R}$ 表示第 i 个（概率论者的）埃尔米特多项式，c_i 是相应的埃尔米特系数，c_r^* 是一个残差项，使得 $\mathbb{E}_{x\sim\mathcal{N}(0,1)}[\hat{\sigma}_r(x)^2] = \mathbb{E}_{x\sim\mathcal{N}(0,1)}[\sigma(x)^2]$。次数 r 取决于 σ 和 $(\boldsymbol{F}^T \mathrm{vec}(\boldsymbol{H_Z}), \boldsymbol{\xi}^T \boldsymbol{x}_{\ell+1})$ 的联合分布，但经验表明一个较小的有限 r 就足够了。

证明. 埃尔米特多项式关于高斯测度的正交性可以与推论 2结合使用来证明该定理，这里省略了证明。对于此类证明的一个示例，我们参考 [12]。 □

定理 3建立了用于研究配备非线性 MLP 的随机 Transformer 的等效且可分析模型。利用这一等价形式，我们系统地比较了配备非线性 MLP 的 Transformer 与线性 Transformer 的性能。此外，我们通过实证验证了非线性 MLP Transformer 与其多项式替代模型之间的等价性，强调了在各种条件下这两种模型的预测一致性。

图 1描述了 ICL 误差如何随着训练样本数量 n 的变化而变化。首先，由于正则化常数相对较小，观察到了一个非单调趋势。值得注意的是，等效多项式模型



(a) 影响（与 ℓ（与 $m = d^2$ 和 $\lambda = 10^{-8}$））(b) m 的影响（与 $\ell = d$ 和 $\lambda = 10^{-8}$ 相关）(c) 影响（带有 λ（与 $\sigma = ReLU$ 和 $\ell = d$））

Fig. 2: 上下文长度（ℓ）、隐藏维度（m）和正则化常数（λ）对 ICL 误差的影响。在 (a) 中，随着我们增加上下文长度 ℓ，ICL 误差减少。在 (b) 中，模型大小的 ICL 误差表现出双下降趋势，在 (c) 中通过正则化得到缓解。对于 (a)-(b)，使用了两种不同的激活函数（用形状和颜色表示），而在 (c) 中仅使用 ReLU 激活函数。我们展示了 20 次蒙特卡罗运行的平均值。这里，σ_* 是 ReLU，$d = 80$，$n = 1.5d^2$，$k = 0.5d$，$\rho = 0.01$。

（来自定理 3）的 ICL 误差与 Transformer 模型的 ICL 误差相匹配，证实了我们的理论发现。此外，类似于具有两层神经网络的监督学习的情况 [11]，Transformer 模型的 ICL 性能受到激活函数 σ 和目标函数 σ_* 之间关系的影响，这基于它们的 Hermite 展开（系数），如图 1a-b 所示。在图 1a中，对于所有值的 n，增强的 Transformer（使用 ReLU 激活的非线性 MLP 头）始终比线性 Transformer 基线实现更低的 ICL 错误。这种性能差距强调了在这种情况下非线性特征处理的好处。相比之下，带有 tanh 激活的 MLP 提供的改进有限或没有超过线性模型，这表明非线性头的有效性对激活函数的选择非常敏感。当底层目标函数本身是 tanh（图 1b）时，使用 MLP 的性能提升减少。带有 tanh 激活的 Transformer 只有轻微改进优于线性基线，这表明 MLP 激活与真实任务非线性的匹配在起关键作用。这些结果表明，当 MLP 层的非线性与目标函数的结构相匹配或互补时最为有益。

图 2说明了各种参数对 ICL 误差的影响，并确认多项式模型的 ICL 性能与随机 Transformer 模型相匹配。首先，在图 2a中，我们研究上下文长度 ℓ 对 ICL 误差的影响，表明增加上下文长度可以均匀减少所有模型的 ICL 误差。关键的是，一旦上下文长度超过一个阈值（我们在设置中的 $\ell \approx d$），带有 MLP 头的 Transformer 开始显著优于线性 Transformer。这一观察突出了提示丰富性的的重要性：足够长的上下文对于非线性 MLP 组件有效提取和利用高阶统计结构是必要的。接下来，在图 2b中，性能对隐藏维度 m 的依赖揭示了一种特征性的非单调行为，通常称为“双重下降现象”：ICL 误差在欠参数化区间减少，在插值阈值（$m/n \approx 1$）附近达到峰值，然后随着模型进入过参数化区间再次下降。这一模式强调了谨慎选择 MLP 宽度的重要性。最优结果是在适度过参数化的区间中实现的，在这个区间内，模型可以充分利用非线性适应能力。最后，正则化（在图 2c中）通过平滑插值阈值附近的尖峰来缓解这种非单调行为。随着正则化强度 λ 的增加，曲线变得更加稳定且误差峰值被减弱。这些结果证实了适当的正则化对于缓解双下降现象 [13] 是必不可少的。

5. 结论

在本文中，我们展示了对配备非线性 MLP 头的 Transformer 进行上下文学习（ICL）的高维分析。通过利用来自 Gaussian universality 和 Hermite 多项式展开的工具，我们证明了一个随机初始化且带有 MLP 头的 Transformer 在渐近上等价于一个有限度数的多项式模型。我们的理论发现精确地界定了非线性特征处理相对于纯线性注意力机制提供显著改进的范围。广泛的模拟支持了我们的理论，表明：(i) 当目标函数是非线性时，MLP 头显著减少了 ICL 误差；(ii) MLP 的好处仅在上下文长度超过某些与维度相关的阈值时才显现出来；以及 (iii) 模型的 ICL 误差表现出一种双

下降行为，这种行为可以通过适当的正则化有效缓解。

这些结果不仅深化了我们对非线性和过参数化如何影响 ICL 的理论理解，还为设计带有 MLP 层的 Transformer 架构提供了实用指导。未来的研究方向包括将分析扩展到多头注意力机制，探索 MLP 块的深层堆叠，并开发适应插值制度的自适应正则化方案。

6. REFERENCES

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," Advances in neural information processing systems, 2017.

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al., "Language models are few-shot learners," Advances in neural information processing systems, 2020.

[3] Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou, "What learning algorithm is in-context learning? investigations with linear models," in International Conference on Learning Representations, 2023.

[4] Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, Joao Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov, "Transformers learn in-context by gradient descent," in International Conference on Machine Learning, 2023.

[5] Jingfeng Wu, Difan Zou, Zixiang Chen, Vladimir Braverman, Quanquan Gu, and Peter Bartlett, "How many pretraining tasks are needed for in-context learning of linear regression?," in International Conference on Learning Representations, 2024.

[6] Ruiqi Zhang, Spencer Frei, and Peter L. Bartlett, "Trained transformers learn linear models in-context," Journal of Machine Learning Research, vol. 25, no. 49, pp. 1–55, 2024.

[7] Hongkang Li, Meng Wang, Songtao Lu, Xiaodong Cui, and Pin-Yu Chen, "How do nonlinear transformers learn and generalize in in-context learning?," in International Conference on Machine Learning, 2024.

[8] Juno Kim and Taiji Suzuki, "Transformers learn nonlinear features in context: Nonconvex mean-field dynamics on the attention landscape," in International Conference on Machine Learning, 2024.

[9] Kazusato Oko, Yujin Song, Taiji Suzuki, and Denny Wu, "Pretrained transformer efficiently learns low-dimensional target functions in-context," Advances in Neural Information Processing Systems, 2024.

[10] Hong Hu and Yue M Lu, "Universality laws for high-dimensional learning with random features," IEEE Transactions on Information Theory, vol. 69, no. 3, pp. 1932–1964, Mar. 2023.

[11] Samet Demir and Zafer Dogan, "Random features outperform linear models: Effect of strong input-label correlation in spiked covariance data," arXiv preprint arXiv:2409.20250, 2024.

[12] Samet Demir and Zafer Dogan, "Asymptotic analysis of two-layer neural networks after one gradient step under gaussian mixtures data with structure," in International Conference on Learning Representations, 2025.

[13] Preetum Nakkiran, Prayaag Venkat, Sham M Kakade, and Tengyu Ma, "Optimal regularization can mitigate double descent," in Interna-

tional Conference on Learning Representations, 2021.

[14] Samet Demir and Zafer Doğan, "Optimal nonlinearities improve generalization performance of random features," in Asian Conference on Machine Learning, 2024.

[15] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus, "Emergent abilities of large language models," Transactions on Machine Learning Research, 2022.

[16] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah, "In-context learning and induction heads," Transformer Circuits Thread, 2022.

[17] Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo, "Are emergent abilities of large language models a mirage?," in Advances in Neural Information Processing Systems, 2023.

[18] Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant, "What can transformers learn in-context? a case study of simple function classes," Advances in Neural Information Processing Systems, 2022.

[19] Allan Raventós, Mansheej Paul, Feng Chen, and Surya Ganguli, "Pretraining task diversity and the emergence of non-bayesian in-context learning for regression," Advances in Neural Information Processing Systems, 2024.

[20] Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei, "Transformers as statisticians: Provable in-context learning with in-context algorithm selection," in Advances in Neural Information Processing Systems, 2023.

[21] Yingcong Li, Muhammed Emrullah Ildiz, Dimitris Papailiopoulos, and Samet Oymak, "Transformers as algorithms: Generalization and stability in in-context learning," in International Conference on Machine Learning, 2023.

[22] Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra, "Transformers learn to implement preconditioned gradient descent for in-context learning," in Advances in Neural Information Processing Systems, 2023.

[23] Arvind V. Mahankali, Tatsunori Hashimoto, and Tengyu Ma, "One step of gradient descent is provably the optimal in-context learner with one layer of linear self-attention," in International Conference on Learning Representations, 2024.

[24] Deqing Fu, Tianqi CHEN, Robin Jia, and Vatsal Sharan, "Transformers learn higher-order optimization methods for in-context learning: A study with linear models," 2024.

[25] Yingcong Li, Ankit Singh Rawat, and Samet Oymak, "Fine-grained analysis of in-context linear estimation: Data, architecture, and beyond," in Advances in Neural Information Processing Systems, 2024.

[26] Core Francisco Park, Ekdeep Singh Lubana, Itamar Pres, and Hidenori Tanaka, "Competition dynamics shape algorithmic phases of in-context learning," arXiv preprint arXiv:2412.01003, 2024.

[27] Yue M Lu, Mary I Letey, Jacob A Zavatone-Veth, Anindita Maiti, and Cengiz Pehlevan, "In-context learning by linear attention: Exact asymptotics and experiments," in NeurIPS 2024 Workshop on Mathematics of Modern Machine Learning, 2024.

[28] Yue M Lu, Mary Letey, Jacob A Zavatone-Veth, Anindita Maiti, and Cengiz Pehlevan, "Asymptotic theory of in-context learning by linear attention," Proceedings of the National Academy of Sciences, vol. 122, no. 28, pp. e2502599122, 2025.

[29] Ali Rahimi and Benjamin Recht, "Random features for large-scale kernel machines," in Advances in Neural Information Processing Systems, 2007.