Vocabulary In-Context Learning in Transformers: Benefits of Positional Encoding

Qian Ma†, Ruoxiang Xu†, Yongqiang Cai† ‡

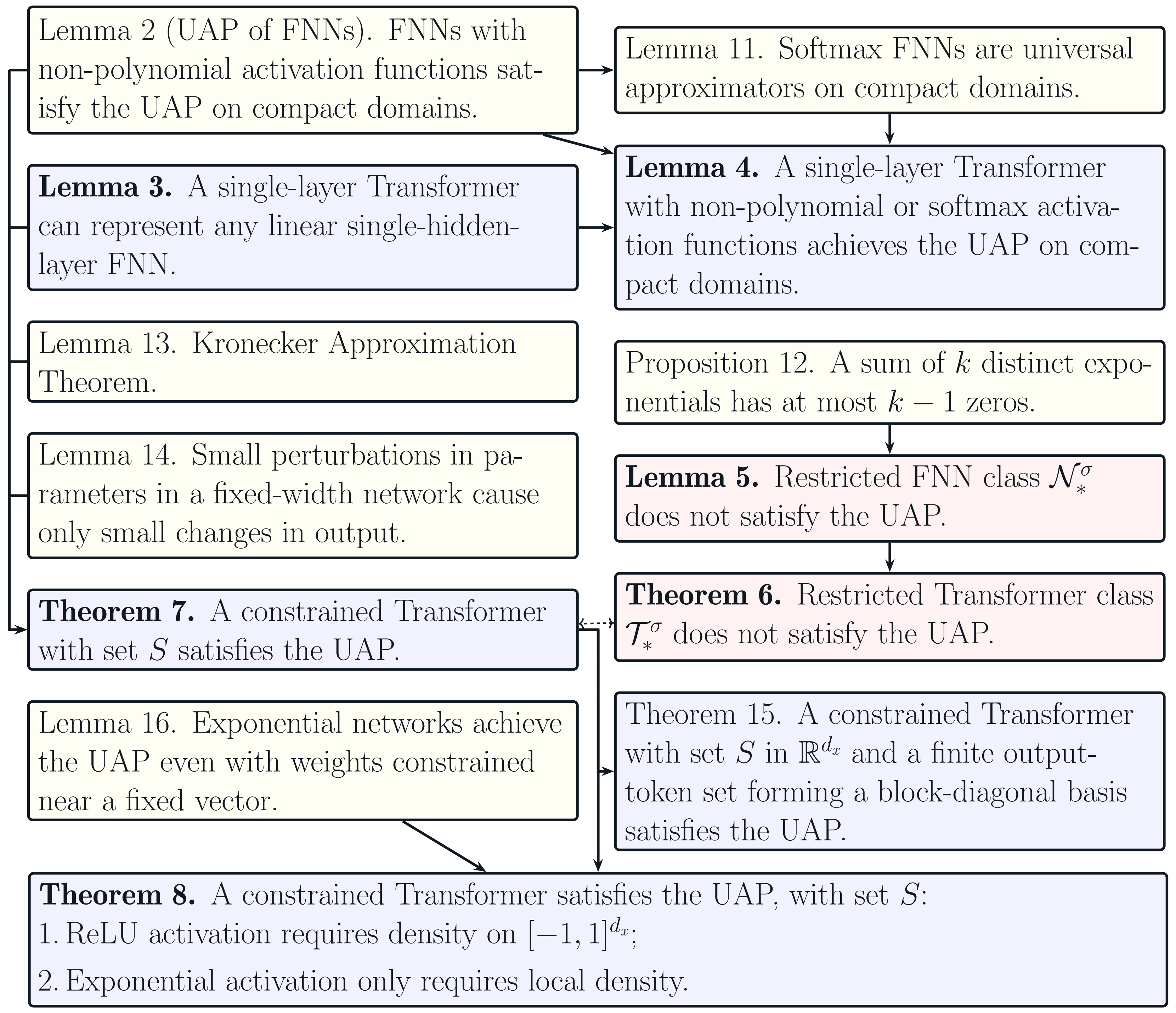
†School of Mathematical Sciences, Beijing Normal University, Beijing, China
‡caiyq.math@bnu.edu.cn

Abstract

We show that single-layer Transformers without positional encoding lack the universal approximation property (UAP), while suitable positional encodings enable it. We further provide sufficient conditions on positional encodings and reveal their role in establishing the UAP from an approximation-theoretic perspective.

Motivation and Contributions

The UAP has been a central topic in neural network theory, traditionally studied in the context of feed-forward networks. Recent works, such as Yun et al. (2020) and Petrov et al. (2024), extend this analysis to Transformers and explore the role of prompting. However, these studies overlook a practical constraint: language-model inputs come from a finite vocabulary embedded in a high-dimensional space. When investigating whether the UAP holds for in-context learning (ICL) under this constraint, we identify positional encoding as a critical factor.



Lemma 2 (UAP of FNNs). FNNs with non-polynomial activation functions satisfy the UAP on compact domains.

Lemma 11. Softmax FNNs are universal approximators on compact domains.

Lemma 3. A single-layer Transformer can represent any linear single-hidden-layer FNN.

Lemma 4. A single-layer Transformer with non-polynomial or softmax activation functions achieves the UAP on compact domains.

Lemma 13. Kronecker Approximation Theorem.

Proposition 12. A sum of k distinct exponentials has at most $k-1$ zeros.

Lemma 14. Small perturbations in parameters in a fixed-width network cause only small changes in output.

Lemma 5. Restricted FNN class $\mathcal{N}_*^\sigma$ does not satisfy the UAP.

Theorem 7. A constrained Transformer with set S satisfies the UAP.

Theorem 6. Restricted Transformer class $\mathcal{T}_*^\sigma$ does not satisfy the UAP.

Lemma 16. Exponential networks achieve the UAP even with weights constrained near a fixed vector.

Theorem 15. A constrained Transformer with set S in $\mathbb{R}^{d_x}$ and a finite output-token set forming a block-diagonal basis satisfies the UAP.

Theorem 8. A constrained Transformer satisfies the UAP, with set S:
1. ReLU activation requires density on $[-1,1]^{d_x}$;
2. Exponential activation only requires local density.

Notations

In the ICL task, when predicting the target at a query vector $x \in \mathbb{R}^{d_x}$ or $z=(x,0)\in\mathbb{R}^{d_x+d_y}$, we define the input matrix Z as follows:

$$Z = \left[z^{(1)}\ z^{(2)}\ \cdots\ z^{(n)}\ z\right] := \begin{bmatrix} x^{(1)} & x^{(2)} & \cdots & x^{(n)} & x \\ y^{(1)} & y^{(2)} & \cdots & y^{(n)} & 0 \end{bmatrix} \in \mathbb{R}^{(d_x+d_y)\times(n+1)}, \tag{1}$$

where n demonstrations are denoted as $z^{(i)} = (x^{(i)}, y^{(i)})$ and $x^{(i)} \in \mathbb{R}^{d_x}$, $y^{(i)} \in \mathbb{R}^{d_y}$, for $i = 1, 2, ..., n$. For neural networks, we denote a one-hidden-layer FNN with activation function σ as N^σ, and the set of all such networks is denoted as $\mathcal{N}^\sigma$, *i.e.*,

$$\begin{aligned}\mathcal{N}^\sigma &= \left\{\mathrm{N}^\sigma := A\,\sigma(Wx+b) \;\middle|\; A\in\mathbb{R}^{d_y\times k},\ W\in\mathbb{R}^{k\times d_x},\ b\in\mathbb{R}^k,\ k\in\mathbb{N}^+\right\} \\ &= \left\{\mathrm{N}^\sigma := \sum_{i=1}^{k} a_i\sigma(w_i\cdot x+b_i) \;\middle|\; (a_i,w_i,b_i)\in\mathbb{R}^{d_y}\times\mathbb{R}^{d_x}\times\mathbb{R},\ k\in\mathbb{N}^+\right\},\end{aligned} \tag{2}$$

and a single-layer Transformer without positional encoding is defined as:

$$\mathrm{T}^\sigma(x;X,Y) := (Z + VZM\sigma((QZ)^\top KZ))_{d_x+1:d_x+d_y,n+1}. \tag{3}$$

In the above notations, the tokens in context of ICL are general and unrestricted. When we refer to a "vocabulary", we mean that the tokens are drawn from a finite set. In this case, we use subscript $*$, *i.e.*, $\mathrm{T}_*^\sigma(x;X,Y)$, to represent the Transformer $\mathrm{T}^\sigma(x;X,Y)$ defined above, and denote the set of such Transformers as $\mathcal{T}_*^\sigma$:

$$\mathcal{T}_*^\sigma = \left\{\mathrm{T}_*^\sigma(x;X,Y) := \mathrm{T}^\sigma(x;X,Y) \;\middle|\; z^{(i)}\in\mathcal{V},\ i\in\{1,2,\cdots,n\},\ n\in\mathbb{N}^+\right\}, \tag{4}$$

where $\mathcal{V}$ denotes the finite vocabulary set.

Universal Approximation Property of In-context Learning

Before we dive into our results, we first state the following assumption, which is inspired by Cheng et al. (2024), and is used to simplify our analysis.

Assumption 1. *The matrices $Q,\ K,\ V \in \mathbb{R}^{(d_x+d_y)\times(d_x+d_y)}$ have the following sparse partition:*

$$Q = \begin{bmatrix} B & 0 \\ 0 & 0 \end{bmatrix}, \quad K = \begin{bmatrix} C & 0 \\ 0 & 0 \end{bmatrix}, \quad V = \begin{bmatrix} D & E \\ F & U \end{bmatrix}, \tag{5}$$

where $B,\ C,\ D \in \mathbb{R}^{d_x\times d_x}$, $E\in\mathbb{R}^{d_x\times d_y}$, $F\in\mathbb{R}^{d_y\times d_x}$ and $U\in\mathbb{R}^{d_y\times d_y}$. Furthermore, the matrices $B,\ C$ and U are non-singular, and the matrix $F=0$.

The following lemmas emphasize the connection between FNNs and Transformers, which suggest that the context in Transformers can act as a control parameter for the model.

Lemma 3. *Let $\sigma:\mathbb{R}\to\mathbb{R}$ be a non-polynomial, locally bounded, piecewise continuous activation function, and T^σ be a single-layer Transformer satisfying Assumption 1. For any one-hidden-layer network $\mathrm{N}^\sigma:\mathbb{R}^{d_x-1}\to\mathbb{R}^{d_y}\in\mathcal{N}^\sigma$ with n hidden neurons, there exist matrices $X\in\mathbb{R}^{d_x\times n}$ and $Y\in\mathbb{R}^{d_y\times n}$ such that*

$$\mathrm{T}^\sigma(\tilde{x};X,Y) = \mathrm{N}^\sigma(x), \quad \forall x\in\mathbb{R}^{d_x-1}. \tag{6}$$

Lemma 4. *Let $\sigma:\mathbb{R}\to\mathbb{R}$ be a non-polynomial, locally bounded, piecewise continuous activation function or softmax function, and T^σ be a single-layer Transformer satisfying Assumption 1, and $\mathcal{K}$ be a compact domain in $\mathbb{R}^{d_x-1}$. Then for any continuous function $f:\mathcal{K}\to\mathbb{R}^{d_y}$ and any $\varepsilon>0$, there exist matrices $X\in\mathbb{R}^{d_x\times n}$ and $Y\in\mathbb{R}^{d_y\times n}$ such that*

$$\left\|\mathrm{T}^\sigma(\tilde{x};X,Y) - f(x)\right\| < \varepsilon, \quad \forall x\in\mathcal{K}. \tag{7}$$

Non-Universal Approximation Property of $\mathcal{T}_*^\sigma$

Theorem 6. *The function class $\mathcal{T}_*^\sigma$, with a non-polynomial, locally bounded, piecewise continuous element-wise activation function or softmax activation function σ and every $\mathrm{T}^\sigma\in\mathcal{T}_*^\sigma$ satisfies Assumption 1, cannot achieve the UAP. Specifically, there exist a compact domain $\mathcal{K}\subset\mathbb{R}^{d_x}$, a continuous function $f:\mathcal{K}\to\mathbb{R}^{d_y}$, and $\varepsilon_0>0$ such that*

$$\max_{x\in\mathcal{K}}\left\|f(x)-\mathrm{T}_*^\sigma(\tilde{x})\right\| \geq \varepsilon_0, \quad \forall\ \mathrm{T}_*^\sigma\in\mathcal{T}_*^\sigma. \tag{8}$$

The element-wise activation case follows directly from Lemma 3 and Lemma 5. The softmax case requires additional arguments for normalization, with the proof (using Proposition 12) provided in Appendix C. Importantly, Theorem 6 holds without any constraints on V, Q, or K in Eq.(5); see Appendix F.

Universal Approximation Property of $\mathcal{T}_{*,\mathcal{P}}^\sigma$

After establishing that $\mathcal{T}_*^\sigma$ can not achieve the UAP, we aim to leverage a key feature of Transformers: their ability to incorporate APEs during token input. This motivates us to investigate whether $\mathcal{T}_{*,\mathcal{P}}^\sigma$ can realize the UAP.

Theorem 7. *Let $\mathcal{T}_{*,\mathcal{P}}^\sigma$ be the class of functions $\mathrm{T}_{*,\mathcal{P}}^\sigma$ satisfying Assumption 1, with a non-polynomial, locally bounded, piecewise continuous element-wise activation function σ, the subscript refers to the finite vocabulary $\mathcal{V}=\mathcal{V}_x\times\mathcal{V}_y$, $\mathcal{P}=\mathcal{P}_x\times\mathcal{P}_y$ represents the positional encoding map, and denote a set S as:*

$$S := \mathcal{V}_x + \mathcal{P}_x = \left\{x_i + \mathcal{P}_x^{(j)} \;\middle|\; x_i\in\mathcal{V}_x,\ i,\ j\in\mathbb{N}^+\right\}. \tag{9}$$

If S is dense in $\mathbb{R}^{d_x}$, $\{1,\ -1,\ \sqrt{2},\ 0\}^{d_y}\subset\mathcal{V}_y$ and $\mathcal{P}_y=0$, then $\mathcal{T}_{*,\mathcal{P}}^\sigma$ can achieve the UAP. More specifically, given a network $\mathrm{T}_{*,\mathcal{P}}^\sigma$, then for any continuous function $f:\mathbb{R}^{d_x-1}\to\mathbb{R}^{d_y}$ defined on a compact domain $\mathcal{K}$ and $\varepsilon>0$, there always exist $X\in\mathbb{R}^{d_x\times n}$ and $Y\in\mathbb{R}^{d_y\times n}$ from the vocabulary $\mathcal{V}$, i.e., *$x^{(i)}\in\mathcal{V}_x, y^{(i)}\in\mathcal{V}_y$, with some length $n\in\mathbb{N}^+$ such that*

$$\left\|\mathrm{T}_{*,\mathcal{P}}^\sigma(\tilde{x};X,Y) - f(x)\right\| < \varepsilon, \quad \forall x\in\mathcal{K}. \tag{10}$$

We reduce a special case of the Transformer to an equivalent FNN (Lemma 3) and exploit the FNN's UAP (Lemma 2). The proof constructs context tokens whose embeddings and positional encodings emulate the target weights, ensuring approximation when the set S is dense. We later relax this density requirement by removing the unboundedness of $\mathcal{P}_x$, aligning the conditions with practical settings.

Theorem 8. *Let $\mathcal{T}_{*,\mathcal{P}}^\sigma$ be the class of functions $\mathrm{T}_{*,\mathcal{P}}^\sigma$ satisfying Assumption 1, with a non-polynomial, locally bounded, piecewise continuous element-wise activation function σ, the subscript refers to the finite vocabulary $\mathcal{V}=\mathcal{V}_x\times\mathcal{V}_y$, $\mathcal{P}=\mathcal{P}_x\times\mathcal{P}_y$ represents the positional encoding map, and denote a set S as:*

$$S := \mathcal{V}_x + \mathcal{P}_x = \left\{x_i + \mathcal{P}_x^{(j)} \;\middle|\; x_i\in\mathcal{V}_x,\ i,\ j\in\mathbb{N}^+\right\}. \tag{11}$$

If the set S is dense in $[-1,1]^{d_x}$, then $\mathcal{T}_{*,\mathcal{P}}^{\mathrm{ReLU}}$ is capable of achieving the UAP. Additionally, if S is only dense in a neighborhood $B(w^*,\delta)$ of a point $w^*\in\mathbb{R}^{d_x}$ with radius $\delta>0$, then the class of transformers with exponential activation, i.e., *$\mathcal{T}_{*,\mathcal{P}}^{\exp}$, is capable of achieving the UAP.*

This section refines the density condition on S by exploiting properties specific to different activations. For ReLU networks, positive homogeneity allows constraining weights to $[-1,1]$ without loss of expressivity. For exponential networks, a weaker condition suffices: derivatives of $\exp(w\cdot x)$ link exponential neurons to polynomial approximation, and finite-difference schemes implement these derivatives via small weight perturbations. Combined with the Stone–Weierstrass theorem, this establishes UAP under constrained weights. The result extends to standard ICL when $y^{(i)}=f(x^{(i)})$ and $\mathcal{V}_y$ satisfies mild conditions.

References

[1] Chulhee Yun et al. "Are Transformers Universal Approximators of Sequence-to-Sequence Functions?" In: *International Conference on Learning Representations*. 2020.

[2] Aleksandar Petrov et al. "Prompting a Pretrained Transformer Can Be a Universal Approximator". In: *International Conference on Machine Learning*. 2024.

[3] Xiang Cheng et al. "Transformers Implement Functional Gradient Descent to Learn Non-Linear Functions In Context". In: *International Conference on Machine Learning*. 2024.