On the Role of Transformer Feed-Forward Layers in Nonlinear In-Context Learning

Haoyuan Sun HAOYUANS@MIT.EDU
Ali Jadbabaie JADBABAI@MIT.EDU
Navid Azizan AZIZAN@MIT.EDU
Massachusetts Institute of Technology

Editors: Matus Telgarsky and Jonathan Ullman

Abstract

Large language models (LLM) based on the Transformer architecture (Vaswani et al., 2017) have demonstrated the extraordinary ability to perform *in-context learning* (ICL) (Brown et al., 2020; Wei et al., 2022; Min et al., 2022), where a model adapts to new tasks at inference time through a *prompt* containing a few examples, without requiring updates to its parameters. Formally, given a prompt composed of input-label pairs and a query, $(x_1, f(x_1), x_2, f(x_2), \ldots, x_n, f(x_n), x_{\text{query}})$, a Transformer model is said to learn a function class $\mathcal{F}$ *in context* if it can accurately predict $f(x_{\text{query}})$ for previously unseen functions $f \in \mathcal{F}$.

The study of in-context learning has attracted significant attention in recent literature. Empirical studies by Garg et al. (2022) demonstrated that Transformers can perform ICL for various function classes, including linear functions, decision trees, and ReLU neural networks. Building on these insights, Akyürek et al. (2023) and Von Oswald et al. (2023) showed theoretically that Transformers can learn linear functions in context by implicitly implementing gradient descent over a linear regression objective using only the attention mechanism. Notably, this mechanism can be facilitated by a simple parameter configuration of *linear self-attention* (LSA), which is a variant of self-attention that omits the softmax activation function. Subsequent analysis by Ahn et al. (2024); Zhang et al. (2024); Mahankali et al. (2024) showed that under suitable distributions, the optimal single-layer LSA can effectively implement one step of gradient descent, providing strong theoretical justification for linear ICL.

Despite these advances, the theoretical understanding of *nonlinear in-context learning* remains underdeveloped. A key bottleneck lies in the expressive power of linear self-attention, where we prove that LSA is inherently incapable of outperforming linear predictors on nonlinear tasks, thereby highlighting a hard expressivity barrier for attention-only models. To overcome this limitation, we demonstrate that *feed-forward layers* — a core component of Transformer blocks — play a crucial and previously underappreciated role in enabling nonlinear ICL.

To illustrate this, we analyze a Transformer block consisting of LSA and feed-forward layers inspired by the *gated linear units* (GLU), which is a standard component in modern Transformer architectures. We show that such a block achieves nonlinear ICL by implementing kernel regression, where the feed-forward layers compute a kernel function that generates nonlinear features, while the attention mechanism performs gradient updates over this feature space. Furthermore, our analysis reveals that the expressivity of a single such block is inherently limited by its dimensions. We then show that a deep Transformer can overcome this bottleneck by distributing the computation of richer kernel functions across multiple blocks, effectively performing block-coordinate descent in a high-dimensional feature space that cannot be represented by just one Transformer block. Our findings highlight that the feed-forward layers provide a crucial, scalable, and interpretable mechanism by which Transformers can express nonlinear representations for in-context learning.

Keywords: Transformers, in-context learning, neural networks, kernel regression, gated linear unit

[1] Extended abstract. Full version appears as Sun et al. (2025).

Our Contributions

The key contributions of this work are organized as follows:

- We establish the essential role of feed-forward layers in enabling nonlinear ICL. We prove that for nonlinear target functions, *no deep linear self-attention (LSA) network* can achieve lower in-context learning loss than a linear least-squares predictor. Then, we draw inspiration from modern Transformer architectures and show that a Transformer block consisting of a feed-forward layer similar to the gated linear unit (GLU) and an LSA layer can implement one step of gradient descent with respect to a quadratic kernel.
- Next, we analyze the Transformer model featuring this LSA-GLU mechanism. We derive optimal conditions for the weights within these layers. In particular, we show that the optimal feed-forward layer must compute all possible quadratic features — effectively implementing a quadratic kernel. And the optimal attention layer implements gradient descent in the feature space with a preconditioner approximately equal to the negative inverse of the features' covariance matrix. This confirms the optimality of our construction.
- Finally, we identify several challenges with the initial two-layer construction and study a deeper Transformer model that can effectively perform complex nonlinear ICL tasks, including learning higher-order polynomial functions in context, by distributing the computation of kernel functions across multiple feed-forward layers. Crucially, our analysis extends to other forms of feed-forward layers: similar results hold for other feed-forward architectures provided they can express the required nonlinear feature mappings. These results illustrate the expressive power of deep Transformer architectures, where their ICL capabilities scale with depth.

Acknowledgment

This work was supported in part by MathWorks, the MIT-IBM Watson AI Lab, the MIT-Amazon Science Hub, the MIT-Google Program for Computing Innovation, the DARPA Artificial Intelligence Quantified program, and the ONR grant #N00014-23-1-2299. The authors thank Amirhossein Reisizadeh, Mehrdad Ghadiri, and Juan Cervino for their insightful discussions and valuable feedback.

References

Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. Transformers learn to implement preconditioned gradient descent for in-context learning. *Advances in Neural Information Processing Systems*, 36, 2024.

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.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 2020.

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*, 35:30583–30598, 2022.

Arvind Mahankali, Tatsunori B 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.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. Metaicl: Learning to learn in context. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2791–2809, 2022.

Haoyuan Sun, Ali Jadbabaie, and Navid Azizan. On the role of transformer feed-forward layers in nonlinear in-context learning. *arXiv preprint arXiv:2501.18187*, 2025. URL https://arxiv.org/abs/2501.18187.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, et al. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, et al. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*. PMLR, 2023.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022.

Ruiqi Zhang, Spencer Frei, and Peter L Bartlett. Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, 25(49):1–55, 2024.