MSc Project Proposal 2025

Title: In-Context Learning Capabilities of Transformers

Supervisors: Xiaoqi (Shirley) Liu, Patrick Rebeschini

Brief description: Consider this simple example: when you give GPT the sequence "sakne→snake, aplep→apple, mvoie→movie, moues→?", it correctly responds with "mouse". What's remarkable here is that GPT wasn't explicitly trained on unscrambling words. Instead, it learned the pattern from just the few examples provided in the prompt, without any parameter updates or fine-tuning. This fascinating emergent capability is called in-context learning (ICL) [1], which has attracted many recent research interest [2–8]. ICL challenges our traditional understanding that learning requires parameter updates. Instead, transformers adapt to new tasks through a single forward pass, using attention to learn from examples on the fly.

The goal of this project is to i) investigate when and why ICL emerges in transformers from both a theoretical and empirical perspective through simple learning tasks (e.g., linear regression), and ii) quantify the statistical efficiency of ICL of transformers compared to classical statistical methods (e.g., least squares).

Possible avenues of investigation: One promising entry point is the landmark paper [2], which studies whether one can train a transformer to in-context learn a certain function class $\mathcal{F}$. Concretely, the authors say that a model can in-context learn $\mathcal{F}$ if, for any function $f \in \mathcal{F}$, the model can approximate $f(x_{\mathsf{query}})$ for a new query input x_{query} by conditioning on a prompt sequence $P = (x_1, f(x_1), \ldots, x_k, f(x_k), x_{\mathsf{query}})$ containing in-context examples and the query input. Let $D_{\mathcal{X}}$ be a distribution over inputs, $D_{\mathcal{F}}$ be a distribution over functions in $\mathcal{F}$, and each prompt P be a sequence where the inputs are drawn i.i.d. from $D_{\mathcal{X}}$ and f is drawn from $D_{\mathcal{F}}$. Their proposed training procedure of a transformer TF (that hopefully in-context learns) involves minimising the following loss:

$$\min_{\theta} \frac{1}{B} \sum_{j=1}^{B} \frac{1}{k+1} \sum_{i=0}^{k} \ell\left(\mathsf{TF}_{\theta}(P_i^j), f(x_{i+1}^j)\right) \tag{1}$$

where $\ell(\cdot, \cdot)$ is an appropriately chosen loss function, P_i denotes the prompt prefix containing i in-context examples, and the superscript j indexes the j-th prompt in the set of B training prompts.

Building upon this framework, some initial explanations have been established about the emergence of ICL. We describe some of them and pose some open questions for investigation in this project:

- A popular hypothesis is that transformers perform ICL by simulating (proximal) gradient descent [3,5,6] in-context. For instance, in [6] the authors provide constructions of modest-sized transformers that can approximate gradient descent steps in-context on various empirical risk functions, including least squares, ridge regression, and Lasso regularization. They show that these models achieve near-Bayes-optimal risk. Can we extend these results to demonstrate transformers' capability to simulate more sophisticated optimization algorithms (such as accelerated gradient descent (like in [9]), adaptive methods, and other general first-order methods [10]) in-context, thereby establishing a broader theoretical foundation for their effectiveness on more complex ICL tasks?

- A complementary theoretical perspective emerges from studying the optimization landscape and training dynamics of transformers, rather than their test-time behaviour [4,11]. For example, [4] demonstrates that gradient flow, with appropriate random initialisation, converges to a global minimum of the (non-convex) training objective. The authors characterise the learning algorithm implicitly encoded by the transformer at convergence and provide bounds on its prediction error for ICL tasks. However, the transformer architecture considered in these works is fairly simple (i.e., a single-layer linear attention mechanism) and so are the learning tasks (i.e., linear regression). Can we establish similar convergence results for deeper architectures and generalised regression tasks with non-linear link functions, under mild forms of non-convexity such as the Polyak-Łojasiewicz condition?

Another interesting entry point, alternative to the analysis framework of [2], is the Bayesian inference framework proposed in [12]. There, the authors show that when the training data has long-range coherence (e.g., a mixture of Hidden Markov Models), ICL occurs when the transformer infers a shared latent concept between in-context examples in a prompt at test time. This perspective admits a more precise quantification of the sample-complexity and error tradeoff, see for instance the information-theoretic analysis in [7].

More broadly, training transformers to perform ICL can be seen as a meta-learning procedure, i.e., learning to learn. This motivates a more fundamental question: can our theoretical understanding of ICL's emergence inform the design of training procedures and prompting strategies to enhance other transformer capabilities, such as reasoning?

Number of students: Three students maximum. Students may specialize – some focusing on surveying and extending theoretical results while others develop experiments to generate new insights – or they may each contribute across both theoretical and experimental aspects of the work.

Prerequisite courses/ knowledge (listed as useful, recommended): Foundations of Statistical Inference, Statistical Machine Learning.

Data available: Synthetic data (and if time permits, freely available datasets such as GINC [12]) will be used to corroborate theoretical results through suitable experiments. A good entry point is to study and reproduce some key numerical results from [2].

Computing: The project includes the option to train a small transformer from scratch. The desktops in the department's IT Suite should be sufficient. We will build upon existing open-source implementations from the cited papers above.

References

[1] 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*, 33:1877–1901, 2020.

[2] 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.

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

[4] Ruiqi Zhang, Spencer Frei, and Peter Bartlett. Trained Transformers Learn Linear Models In-Context. In *R0-FoMo:Robustness of Few-shot and Zero-shot Learning in Large Foundation Models*, 2023.

[5] Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, Joao Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 35151–35174, 23–29 Jul 2023.

[6] Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. *Advances in neural information processing systems*, 36, 2024.

[7] Hong Jun Jeon, Jason D. Lee, Qi Lei, and Benjamin Van Roy. An Information-Theoretic Analysis of In-Context Learning. In *Forty-first International Conference on Machine Learning*, 2024.

[8] Ruiqi Zhang, Jingfeng Wu, and Peter Bartlett. In-context learning of a linear transformer block: benefits of the MLP component and one-step GD initialization. *Advances in Neural Information Processing Systems*, 37:18310–18361, 2024.

[9] Licong Lin, Yu Bai, and Song Mei. Transformers as decision makers: Provable in-context reinforcement learning via supervised pretraining. In *The Twelfth International Conference on Learning Representations*, 2024.

[10] Michael Celentano, Andrea Montanari, and Yuchen Wu. The estimation error of general first order methods. In *Conference on Learning Theory*, pages 1078–1141. PMLR, 2020.

[11] Arvind 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*, volume 2024, pages 32077–32096, 2024.

[12] Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*, 2022.