Deep Networks in Context — Rates, Algorithms and Datastructures

Anonymous Authors[1]

Abstract

Deep Networks have demonstrated unreasonably good in-context learning ability. In this paper we provide theoretical guarantees for this phenomenon in sequence learning. Moreover, we demonstrate related effects for images and graphs. Lastly, we show that this leads to a unified explanation of attention, RAG, and context compression algorithms. Experiments confirm our theoretical guarantees and point to a rich area of new sequence models.

1. Introduction

- In context learning is really popular. Very convenient alternative to training a specialized model from scratch.
- One of the first observations was the Open AI paper where they used this for GPT-2 (look up the paper).
- Pretty much the default for customizing LLMs for small numbers of training examples. Give some references.
- Turkish MIT thesis looks at the problem in a phenomenological manner. Still seems to mystify the community. Even works for time series prediction (see Andrew Wilson's paper for time series using Llama).
- Even recent ICL for audio models.

One of the key questions this paper aims to answer is why ICL works as well as it does and why the models exhibit ICL 'learning rates' that are very reminiscent of classical learning algorithms. We will see that this is not a coincidence but rather a consequence of the learning problems encoded in sequence models. Our work relies on observations that sequence models are regression estimators, as shown e.g. by (**????**).

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

Moreover, our work uses theoretical guarantees of regret minimization guarantees for kernel density estimators and convex optimization and applies them to sequence models. These show that in many cases the regret in sequence models reduces at the rate of $O\left(l^{-\frac{1}{2}}\right)$ in the length l of the context.

Furnished with this insight, we extend our reasoning to structured data, such as images and graphs. In particular, we show that in context learning improves as a function of image patch size. Moreover, we illustrate how attention-enabled graph neural networks generalize label propagation algorithms and how the set transformer (**?**) implements what can be considered the purest version of ICL.

Lastly, our insights aren't just analytical but also constructive - we use them to introduce a range of new algorithms that blend attention, sequence compression and RAG. While their detailed analysis is the subject of future work, it demonstrates the strength of our approach.

2. Related Work and Preliminaries

Maybe listing it all here is a bit too much but it'll set the context nicely for our work. It also serves the purpose of putting the Wang et al. 2025 paper into a bigger context.

Unified DL and classical models

- Kernel density estimator from the Smola and Zhang ICML tutorial as a first example.
- Longhorn paper
- Speed always wins setup
- Alex Wang's paper

Regret minimization for sequences Note that not all guarantees are for sequences. Some are for IID data.

- Bartlett, Hazan and Rakhlin results - https://www.stat.berkeley.edu/ bartlett/papers/bhr-aogd-07.pdf
- For IID data we have rates for kernels - https://proceedings.mlr.press/v70/jiang17b.html but note that the data isn't IID.

- For online convex programming (i.e. SGD) we have the old Zinkevich result from 2003, namely https://www.cs.cmu.edu/ maz/publications/techconvex.pdf

It's worth finding out whether and what similar rates exist for the following methods. Note that this isn't for the *learning algorithms* but rather for said models deployed on new data.

Other datatypes We should briefly mention

- Label propagation algorithms on graphs
- Graph neural networks that have vertex update functions (so that's like unnormalized versions)
- Deep Sets and Set Transformer
- Vision Transformer (local context / full context)
- TabPFN is a great example where they learn missing entries (tabular data AutoML).

Function classes There are actually a lot of different transformer variants:

- Vanilla
- Grouped Query Attention
- NSA (Native Sparse Attention)
- History compression (Jegelka's paper)
- RAG for history

3. In Context Learning

- Spell out the memorize and recall procedure
- Explain that since this is 'learning', it needs to follow the usual rules for learning.
- Explain how this works for sequences. Note that this is *not* IID, hence the regret bounds usually employed need to be used, rather than the IID convergence guarantees. This makes it hard for kernel density estimators but it will work for SSMs.

Then decide to ignore the fact that this is not IID and just suggest that this should hold for sequence models in general. The tricky part is that while the kernel density estimate will probably work for the first layer, it won't work for subsequent ones, as the embeddings there are very much a function of data that's arrived so far.

3.1. Experiments

Maybe we should pull the experiments for that aspect up here?

4. Datastructures

We now need to expand our reasoning to nontrivial datastructures. IID doesn't make much sense there any longer but we can still do something useful: use the reasoning anyway.

- For images check what happens if we try filling in the blank image patch and we increase the size of the context for that. We should see a nice performance improvement as the amount of context provided grows.
- For graphs, we can do the same thing - vertex attribute estimation with increasing graph context size.
- For TabPFN we should be able to get decent learning rates.
- Definitely good rates for Set Transformer models. Ideally let's download models from Huggingface and test it out.

5. Function Classes

Given that all of these sequence models are just there to predict what happens next, we can look at the various memory compression and approximation strategies from a fresh point of view:

- Grouped Query Attention — basically just approximates larger groups of (key,value) sets as a single key and value. This works if shorter context doesn't matter that much at a distance. Reference He He's PhD work for this.
- Low-dimensional Approximation — this is one of the tricks from one of the DeepSeek papers. They run attention lower-dimensional. Again, this works as an approximation.
- Retain (key,value) cache even though you swapped the model. E.g. this is used in ServiceNow Pipelined RL / Olmo 3.
- Compress long-range content (rather than all groups). This is the NSA paper.
- Compress long range overall — Look up Steffi Jegelka's paper on using the Blelloch sum for long history.
- Most extreme case is RAG.
- Multi-edit and condense for kNN. Maybe we can recycle this here: https://cgm.cs.mcgill.ca/ godfried/teaching/pr-notes/dasarathy.pdf (the paper is quite soft tissue).
- Can we design a hierarchical clustering / aggregation algorithm for the history? This should make lookups a lot cheaper (compressed prefill).

Acknowledgements

Do not include acknowledgements in the initial version of the paper submitted for blind review.

If a paper is accepted, the final camera-ready version can (and usually should) include acknowledgements. Such acknowledgements should be placed at the end of the section, in an unnumbered section that does not count towards the paper page limit. Typically, this will include thanks to reviewers who gave useful comments, to colleagues who contributed to the ideas, and to funding agencies and corporate sponsors that provided financial support.

Impact Statement

Authors are **required** to include a statement of the potential broader impact of their work, including its ethical aspects and future societal consequences. This statement should be in an unnumbered section at the end of the paper (co-located with Acknowledgements – the two may appear in either order, but both must be before References), and does not count toward the paper page limit. In many cases, where the ethical impacts and expected societal implications are those that are well established when advancing the field of Machine Learning, substantial discussion is not required, and a simple statement such as the following will suffice:

"This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here."

The above statement can be used verbatim in such cases, but we encourage authors to think about whether there is content which does warrant further discussion, as this statement will be apparent if the paper is later flagged for ethics review.

You should not write

"This section has been left intentionally blank."

The above statement would lead to blank faces and stares by the referees.

References

Langley, P. Crafting papers on machine learning. In *Proc. Intl. Conf. Machine Learning*, pp. 1207–1212. Morgan Kaufmann, San Francisco, CA, 2000.

A. You *can* have an appendix here.

You can have as much text here as you want. The main body must be at most 8 pages long. For the final version, one more page can be added. If you want, you can use an appendix like this one.

The `\onecolumn` command above can be kept in place if you prefer a one-column appendix, or can be removed if you prefer a two-column appendix. Apart from this possible change, the style (font size, spacing, margins, page numbering, etc.) should be kept the same as the main body.