

---

# Deep Networks in Context — Rates, Algorithms and Datastructures

---

Anonymous Authors<sup>1</sup>

## 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 (????).

<sup>1</sup>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  $O(l^{-\frac{1}{2}})$  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 data-structures. 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.

## 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

110 which does warrant further discussion, as this statement will  
111 be apparent if the paper is later flagged for ethics review.

112 You should not write

113  
114 “This section has been left intentionally blank.”

115 The above statement would lead to blank faces and stares  
116 by the referees.  
117

118  
119 **References**

120 Langley, P. Crafting papers on machine learning. In *Proc.*  
121 *Intl. Conf. Machine Learning*, pp. 1207–1212. Morgan  
122 Kaufmann, San Francisco, CA, 2000.  
123

124

125

126

127

128

129

130

131

132

133

134

135

136

137

138

139

140

141

142

143

144

145

146

147

148

149

150

151

152

153

154

155

156

157

158

159

160

161

162

163

164

165 **A. You *can* have an appendix here.**166 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  
167 page can be added. If you want, you can use an appendix like this one.  
168169 The `\onecolumn` command above can be kept in place if you prefer a one-column appendix, or can be removed if you  
170 prefer a two-column appendix. Apart from this possible change, the style (font size, spacing, margins, page numbering, etc.)  
171 should be kept the same as the main body.  
172

173

174

175

176

177

178

179

180

181

182

183

184

185

186

187

188

189

190

191

192

193

194

195

196

197

198

199

200

201

202

203

204

205

206

207

208

209

210

211

212

213

214

215

216

217

218

219