Training Dynamics of Multi-Head Softmax Attention for In-Context Learning: Emergence, Convergence, and Optimality

Siyu Chen SIYU.CHEN.SC3226@YALE.EDU
Heejune Sheen HEEJUNE.SHEEN@YALE.EDU
Tianhao Wang TIANHAO.WANG@YALE.EDU
Zhuoran Yang ZHUORAN.YANG@YALE.EDU
Yale University

Editors: Shipra Agrawal and Aaron Roth

Abstract

Recent studies demonstrate the **I**n-**C**ontext **L**earning (ICL) capability of large pre-trained models, where by feeding the model with a few input-output pairs as the context sequence and quering a new input, the model is able to correctly infer the output of the new query. Current results indicate that models with single-layer, one-head linear attention and combined weights can be trained for one-step pre-conditioned gradient descent on the context sequence. We investigate the training dynamics of a single-layer **M**ulti-head **S**oftmax **A**ttention (`MS-Attn`) for ICL in high-dimensional multi-task linear regression, where the input dimension scales with the sequence length, extending current results to a more practical setting. We establish global convergence of gradient flow under population squared loss for suitable choices of initialization and model parameters, and also prove the optimality of the convergence point within the `MS-Attn` architecture.

When trained on split attention weights, we identify three training phases and observe "task allocation" and "emergence" phenomena in the `MS-Attn` model. In the following theorem, a head is said to select a task if its output contributes most to the model's output for that task.

Theorem 1.[Gradient flow for `MS-Attn`, informal] *Under proper initialization of the gradient flow, assuming that the number of heads H is larger than the number of tasks I, then the following holds:*

Warm-up: *There exists a threshold time $T_0 > 0$, before which the loss value decreases slowly, and I of H attention heads gradually adjust their weights to each select an individual task.*

Emergence: *For a short period around T_0, the loss undergoes a sudden decrease, and these I attention heads rapidly focuse on their own task. Cross-head interference vanishes.*

Convergence: *Finally, gradient flow converges to a point where each task is predominantly handled by a single attention head, in the sense that the output from this head dominates the output of the entire model for this task. The unused heads have diminishing output.*

Furthermore, each of these I attention heads acts as a softmax kernel-smooth estimator on its selected task upon convergence. We prove the optimality of this model and show a strict separation between the multi-head and single-head attention models in terms of the prediction loss of ICL.

Theorem 2.[Optimality for `MS-Attn`, informal] *Assume that the number of heads H is equal to the number of tasks I. Under certain symmetric conditions, the ICL loss $\mathcal{L}_H$ achieved by the convergence model is within a constant factor of the best possible* `MS-Attn` *model with the same number of heads. Furthermore, considering the minimal loss $\mathcal{L}_1^\star$ for a single-head softmax attention model for the same ICL task, we have $\mathcal{L}_1^\star \gtrsim H \cdot \mathcal{L}_H$.*

The key technique for our convergence analysis is to map the gradient flow dynamics in the parameter space to a set of ordinary differential equations in the spectral domain, together with a rigorous characterization of the nonlinearity of the softmax function. To our best knowledge, our work provides the first convergence result for the multi-head softmax attention model.

Keywords: Transformer, attention, gradient flow, in-context learning.

. Extended abstract. Full version appears as [arXiv:2402.19442, v2]