On Mesa-Optimization in Autoregressively Trained Transformers: Emergence and Capability

Chenyu Zheng[1] Wei Huang[2] Rongzhen Wang[1] Guoqiang Wu[3] Jun Zhu[4] Chongxuan Li[1]

[1]Renmin University of China, [2]RIKEN AIP

[3]Shandong University, [4]Tsinghua University

Nov. 2024

Table of Contents

Table of Contents

1.1 In-context learning

Foundation models have revolutionized the AI community in lots of fields.

- The crux behind these large models is a very simple yet profound strategy named autoregressive (AR) pretraining with transformers.
- One of their most intriguing properties is the in-context learning (ICL) ability.

Unfortunate fact

However, the reason behind the emergence of ICL ability is still poorly understood.

1.2 Mesa-optimization hypothesis

Nowadays, the mesa-optimization has become a popular hypothesis for explaining ICL.

Mesa-optimization hypothesis

Transformers learn some algorithms during the AR pretraining. In other words, the forward pass of the trained transformers is equivalent to optimizing some inner objective functions on the in-context data.

Our questions

1. *When do mesa-optimization algorithms emerge in autoregressively trained transformers?*
2. *What is the capability limitation of the mesa-optimizer if it does emerge?*

1.3 Our contributions

Our contributions can be summarized as follows.

Our contributions

- We propose a theoretical baseline to study the properties of the AR transformer.
- We verify the empirical mesa-optimization hypothesis in such setup.

Table of Contents

2.1 Data distribution

We want to generate sequence $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T) \in \mathbb{C}^{d \times T}$ according to the true distribution.

- The start point $\boldsymbol{x}_1$ is sampled from some distribution $\mathcal{D}_{\boldsymbol{x}_1}$.
- A unitary matrix $\boldsymbol{W} \in \mathbb{C}^{d \times d}$ is sampled uniformly from $\mathcal{P}_{\boldsymbol{W}} = \{\mathrm{diag}(\lambda_1, \ldots, \lambda_d) \mid |\lambda_i| = 1, \forall i \in [d]\}$.
- Subsequent elements are generated as $\boldsymbol{x}_{t+1} = \boldsymbol{W}\boldsymbol{x}_t$ for $t \in [T-1]$.

Why this distribution?

Given $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{t-1})$ sampled from this distribution, the optimal algorithm to predict $\boldsymbol{x}_t$ is optimizing the ordinary least squares (OLS) problem over $\{(\boldsymbol{x}_1, \boldsymbol{x}_2), \ldots, (\boldsymbol{x}_{t-2}, \boldsymbol{x}_{t-1})\}$, and then using the estimated $\widehat{\boldsymbol{W}}$ to predict $\widehat{\boldsymbol{x}}_{t+1} = \widehat{\boldsymbol{W}}\boldsymbol{x}_t$. We want to examine whether the trained transformers can learn this optimal algorithm.

2.2 Model

We study the one-layer linear casual attention with residual connection as follows.

- Model computation:

$$\boldsymbol{f}_t(\boldsymbol{E}_t;\boldsymbol{\theta}) = \boldsymbol{e}_t + \boldsymbol{W}^{PV}\boldsymbol{E}_t \cdot \frac{\boldsymbol{E}_t^* \boldsymbol{W}^{KQ}\boldsymbol{e}_t}{\rho_t}.$$

- Embedding:

$$\boldsymbol{E}_t = (\boldsymbol{e}_1, \dots, \boldsymbol{e}_t) = \begin{pmatrix} \boldsymbol{0}_d & \boldsymbol{0}_d & \cdots & \boldsymbol{0}_d \\ \boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_t \\ \boldsymbol{x}_0 & \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_{t-1} \end{pmatrix} \in \mathbb{C}^{3d\times t}.$$

- Model output:

$$\widehat{\boldsymbol{y}}_t(\boldsymbol{E}_t;\boldsymbol{\theta}) = [\boldsymbol{f}_t(\boldsymbol{E}_t;\boldsymbol{\theta})]_{1:d}.$$

2.3 Training algorithm

We consider the next-token prediction loss and its gradient flow.

$$L(\boldsymbol{\theta}) = \sum_{t=2}^{T-1} L_t(\boldsymbol{\theta}) = \sum_{t=2}^{T-1} \mathbb{E}_{\boldsymbol{x}_1, \boldsymbol{W}} \left[\frac{1}{2} \|\widehat{\boldsymbol{y}}_t - \boldsymbol{x}_{t+1}\|_2^2 \right], \quad \frac{\mathrm{d}}{\mathrm{d}\tau}\boldsymbol{\theta} = -\nabla L(\boldsymbol{\theta}).$$

Assumption 1 (Diagonal initialization)

At the initial time $\tau = 0$, we assume that

$$\boldsymbol{W}^{KQ}(0) = \begin{pmatrix} \mathbf{0}_{d\times d} & \mathbf{0}_{d\times d} & \mathbf{0}_{d\times d} \\ \mathbf{0}_{d\times d} & \color{red}{\mathbf{0}_{d\times d}} & \color{red}{\mathbf{0}_{d\times d}} \\ \mathbf{0}_{d\times d} & \color{red}{a_0 \boldsymbol{I}_d} & \color{red}{\mathbf{0}_{d\times d}} \end{pmatrix}, \boldsymbol{W}^{PV}(0) = \begin{pmatrix} \mathbf{0}_{d\times d} & \color{red}{b_0 \boldsymbol{I}_d} & \color{red}{\mathbf{0}_{d\times d}} \\ \mathbf{0}_{d\times d} & \mathbf{0}_{d\times d} & \mathbf{0}_{d\times d} \\ \mathbf{0}_{d\times d} & \mathbf{0}_{d\times d} & \mathbf{0}_{d\times d} \end{pmatrix},$$

where the red submatrices are related to the $\widehat{y}_t$ and changed during the training process.

Table of Contents

3.1 A data condition

We figure out a sufficient condition for the emergence of mesa-optimizer.

Assumption 2

We assume that the distribution $\mathcal{D}_{\boldsymbol{x}_1}$ of the initial token $\boldsymbol{x}_1 \in \mathbb{R}^d$ satisfies $\mathbb{E}_{\boldsymbol{x}_1 \sim \mathcal{D}_{\boldsymbol{x}_1}}[x_{1i_1} x_{1i_2}^{r_2} \cdots x_{1i_n}^{r_n}] = 0$ for any subset $\{i_1, \ldots, i_n \mid n \leq 4\}$ of $[d]$, and $r_2, \ldots r_n \in \mathbb{N}$. In addition, we assume that $\kappa_1 = \mathbb{E}[x_{1j}^4]$, $\kappa_2 = \mathbb{E}[x_{1j}^6]$ and $\kappa_3 = \sum_{r \neq j} \mathbb{E}[x_{1j}^2 x_{1r}^4]$ are finite constant for any $j \in [d]$.

Example

We note that any random vectors $\boldsymbol{x}_1$ whose coordinates x_{1i} are i.i.d. random variables with zero mean and finite moments satisfy this assumption. For example, it includes the normal distribution $\mathcal{N}(\mathbf{0}_d, \sigma^2 \boldsymbol{I}_d)$, which is a common setting in the learning theory field.

3.2 Convergence of the gradient flow

Theorem 1

Consider the gradient flow of the one-layer linear transformer over the population AR pretraining loss. Suppose the initialization satisfies Assumption 1, and the initial token's distribution $\mathcal{D}_{\boldsymbol{x}_1}$ *satisfies Assumption 2, then the gradient flow converges to*

$$\begin{pmatrix} \widetilde{\boldsymbol{W}_{22}^{KQ}} & \widetilde{\boldsymbol{W}_{23}^{KQ}} \\ \widetilde{\boldsymbol{W}_{32}^{KQ}} & \widetilde{\boldsymbol{W}_{33}^{KQ}} \end{pmatrix} = \begin{pmatrix} \mathbf{0}_{d\times d} & \mathbf{0}_{d\times d} \\ \widetilde{a}\boldsymbol{I}_d & \mathbf{0}_{d\times d} \end{pmatrix}, \begin{pmatrix} \widetilde{\boldsymbol{W}_{12}^{PV}} & \widetilde{\boldsymbol{W}_{13}^{PV}} \end{pmatrix} = \begin{pmatrix} \widetilde{b}\boldsymbol{I}_d & \mathbf{0}_{d\times d} \end{pmatrix}.$$

Though different initialization (a_0, b_0) *lead to different* $(\widetilde{a}, \widetilde{b})$*, the solutions' product* $\widetilde{a}\widetilde{b}$ *satisfies*

$$\widetilde{a}\widetilde{b} = \frac{\kappa_1}{\kappa_2 + \frac{\kappa_3}{T-2}\sum_{t=2}^{T-1}\frac{1}{t-1}}.$$

3.3 Trained transformer is a mesa-optimizer

Corollary 1

We suppose that the same precondition of Theorem 1 holds. When predicting the $(t+1)$-th token, the trained transformer obtains $\widehat{\boldsymbol{W}}$ by implementing one step of gradient descent for the OLS problem $L_{\text{OLS},t}(\boldsymbol{W}) = \frac{1}{2}\sum_{i=1}^{t-1} \|\boldsymbol{x}_{i+1} - \boldsymbol{W}\boldsymbol{x}_i\|^2$, starting from the initialization $\boldsymbol{W} = \mathbf{0}_{d\times d}$ with a step size $\frac{\widetilde{a}\widetilde{b}}{t-1}$.

Remark

The one-layer transformer learns to perform one step of GD to optimize the optimal objective.

3.4 Capability limitation of the mesa-optimizer

Theorem 2

Let $\mathcal{D}_{\boldsymbol{x}_1}$ be the multivariate normal distribution $\mathcal{N}(\mathbf{0}_d, \sigma^2 \boldsymbol{I}_d)$ with any $\sigma^2 > 0$, then the "simple" AR process can not be recovered by the trained transformer even in the ideal case with long training context.

Remark

This negative result shows that one-step GD learned by the AR transformer can not recover the distribution. Future works are suggested to study more complex transformer architecture.

Table of Contents

Conclusion

Our contributions can be summarized as follows.

Our contributions

- We propose a theoretical baseline to study the properties of the AR transformer.
- We verify the empirical mesa-optimization hypothesis in such setup.