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]Gaoling School of AI, Renmin University of China [2]RIKEN AIP [3]Shandong University [4]Tsinghua University

Abstract

Autoregressively trained transformers have brought a profound revolution to the world, especially with their in-context learning (ICL) ability to address downstream tasks. Recently, several studies suggest that transformers learn a mesa-optimizer during autoregressive (AR) pretraining to implement ICL. Namely, the forward pass of the trained transformer is equivalent to optimizing an inner objective function in-context. However, whether the practical non-convex training dynamics will converge to the ideal mesa-optimizer is still unclear. Towards filling this gap, we investigate the non-convex dynamics of a one-layer linear causal self-attention model autoregressively trained by gradient flow, where the sequences are generated by an AR process $\boldsymbol{x}_{t+1} = \boldsymbol{W}\boldsymbol{x}_t$. First, under a certain condition of data distribution, we prove that *an autoregressively trained transformer learns $\boldsymbol{W}$ by implementing one step of gradient descent to minimize an ordinary least squares (OLS) problem in-context. It then applies the learned $\widehat{\boldsymbol{W}}$ for next-token prediction*, thereby verifying the mesa-optimization hypothesis. Next, under the same data conditions, we explore the capability limitations of the obtained mesa-optimizer. We show that a stronger assumption related to the moments of data is the sufficient and necessary condition that the learned mesa-optimizer recovers the distribution. Besides, we conduct exploratory analyses beyond the first data condition and prove that generally, the trained transformer will not perform vanilla gradient descent for the OLS problem. Finally, our simulation results verify the theoretical results, and the code is available at *https://github.com/ML-GSAI/MesaOpt-AR-Transformer*.

Highlights

- We propose a **theoretical baseline** to study the properties of the AR transformer.
- We **verify the empirical mesa-optimization hypothesis** in such setup.
- We conduct simulation to **validate our theoretical findings**.

Problem setup

We introduce the proposed theoretical setting.

- 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.
  - 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 $t \in [T-1]$.
- Model: one-layer linear casual attention
  - 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:

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

  - Model output:

$$\widehat{\boldsymbol{y}}_t(\boldsymbol{E}_t; \boldsymbol{\theta}) = [\boldsymbol{f}_t(\boldsymbol{E}_t; \boldsymbol{\theta})]_{1:d} = \begin{pmatrix} \boldsymbol{W}_{12}^{PV} & \boldsymbol{W}_{13}^{PV} \end{pmatrix} \frac{\boldsymbol{E}_t^{\boldsymbol{x}} \boldsymbol{E}_t^{\boldsymbol{x}*}}{\rho_t} \begin{pmatrix} \boldsymbol{W}_{22}^{KQ} & \boldsymbol{W}_{23}^{KQ} \\ \boldsymbol{W}_{32}^{KQ} & \boldsymbol{W}_{33}^{KQ} \end{pmatrix} \boldsymbol{e}_t^{\boldsymbol{x}}.$$

- Training algorithm
  - Next-token prediction loss:

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

  - **Assumption 1**(Diagonal initialization) At the initial time $\tau = 0$, we assume that

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

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

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\boldsymbol{\theta} = -\nabla L(\boldsymbol{\theta}).$$

Trained transformer is a mesa-optimizer

Assumption 2(Sufficient condition for the emergence of mesa-optimizer) 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 $\{i_1, \ldots, i_n \mid n \le 4\}$ 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 $j \in [d]$.

Theorem 1(Convergence of the gradient flow) 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

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

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

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

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 $L_{\mathrm{OLS},t}(\boldsymbol{W}) = \frac{1}{2}\sum_{i=1}^{t-1} \|\boldsymbol{x}_{i+1} - \boldsymbol{W}\boldsymbol{x}_i\|^2$, starting from the initialization $\boldsymbol{W} = \boldsymbol{0}_{d\times d}$ with a step size $\frac{\widetilde{a}\widetilde{b}}{t-1}$.

Capability limitation of the trained transformer

First, we give a "simple" distribution that can not be recovered by the trained transformer.

Proposition 1(AR process with normal distributed initial token can not be learned) Let $\mathcal{D}_{\boldsymbol{x}_1}$ be the multivariate normal distribution $\mathcal{N}(\boldsymbol{0}_d, \sigma^2 \boldsymbol{I}_d)$ 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. Formally, when the training sequence length T_{tr} is large enough, for any test context length T_{te} and dimension $j \in [d]$, the prediction from the trained transformer satisfies

$$E_{x_1, W}[\frac{(\widehat{y}_{T_{te}})_j}{(W x_{T_{te}})_j}] \to \frac{1}{5}.$$

Therefore, the prediction $\widehat{\boldsymbol{y}}_{T_{te}}$ will not converges to the true next token $\boldsymbol{W}\boldsymbol{x}_{T_{te}}$.

Remark. Proposition 4.1 suggests that ICL by AR pretraining is different from ICL by few-shot pretraining [2].

Assumption 3(Condition for success of mesa-optimizer) Based on Assumption 2, we further suppose that $\frac{\kappa_1}{\kappa_2}\frac{\sum_{i=1}^{T_{te}-1} \boldsymbol{x}_i \boldsymbol{x}_i^*}{T_{te}-1} \boldsymbol{x}_{T_{te}} \to \boldsymbol{x}_{T_{te}}$ for any $\boldsymbol{x}_1$ and $\boldsymbol{W}$, when T_{te} is large enough.

Example 1(sparse vector) If the random vector $\boldsymbol{x}_1 \in R^d$ is uniformly sampled from the candidate set of size $2d$ $\{\pm(c, 0, \ldots, 0)^\top, \pm(0, c, \ldots, 0)^\top, \pm(0, \ldots, 0, c)^\top\}$ for any fixed $c \in \mathbb{R}$, then the distribution $\mathcal{D}_{\boldsymbol{x}_1}$ satisfies Assumption 3.

Theorem 2(Trained transformer succeed to learn the distribution satisfies Assumption 3) Suppose that Assumption 1 and 2 hold, then Assumption3 is the sufficient and necessary condition for the trained transformer to learn the AR process. Formally, when the training sequence length T_{tr} and test context length T_{te} are large enough, the prediction from the trained transformer satisfies

$$\widehat{\boldsymbol{y}}_{T_{te}} \to \boldsymbol{W}\boldsymbol{x}_{T_{te}}, \quad T_{tr}, T_{te} \to +\infty.$$

Go beyond the Assumption 1

We conduct exploratory analyses by adopting the setting in [1], where the initial token $\boldsymbol{x}_1$ is fixed as $\boldsymbol{1}_d$.

First, sharing the similar but weaker assumption of [1], we impose $\boldsymbol{W}_{32}^{KQ}$ and $\boldsymbol{W}_{12}^{PV}$ to stay diagonal during training by masking the non-diagonal gradients, then the trained transformer will perform one step of gradient descent.

Theorem 3(Trained transformer as mesa-optimizer with non-diagonal gradient masking) Suppose the initialization satisfies Assumption 1, the initial token is fixed as $\boldsymbol{1}_d$, and we clip non-diagonal gradients of $\boldsymbol{W}_{32}^{KQ}$ and $\boldsymbol{W}_{12}^{PV}$ during the training, then the gradient flow of the one-layer linear transformer over the population AR loss converges to the same structure as the result in Theorem 1, with

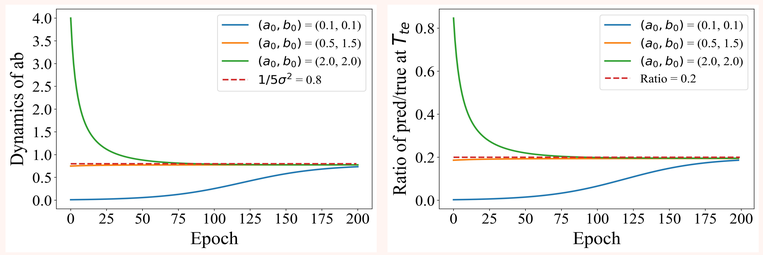
$$\widetilde{a}\widetilde{b} = \frac{1}{1 + \frac{d-1}{T-2}\sum_{t=2}^{T-1} \frac{1}{t-1}}.$$

Therefore, the obtained transformer performs one step of gradient descent in this case.

Next, we adopt some exploratory analyses for the gradient flow without additional non-diagonal gradient masking.

Proposition 2(Trained transformer does not perform on step of gradient descent) The limiting point found by the gradient does not share the same structure as that in Theorem 1, thus the trained transformer will not implement one step of vanilla gradient descent for minimizing the OLS problem $\frac{1}{2}\sum_{i=1}^{t-1} \|\boldsymbol{x}_{i+1} - W\boldsymbol{x}_i\|^2$. We suggest that it will perform some preconditioned gradient descent.

Simulation results



(a) Gaussian with $\sigma = 0.5$, dynamics of ab. (b) Gaussian with $\sigma = 0.5$, ratio of $\widehat{\boldsymbol{y}}_{T_{te}-1}/\boldsymbol{x}_{T_{te}}$.

Figure 1. For example, simulation results on Gaussian show that the convergence of ab satisfies Theorem 1 and verifies Proposition 1.

References

[1] Michael Eli Sander, Raja Giryes, Taiji Suzuki, Mathieu Blondel, and Gabriel Peyré. How do transformers perform in-context autoregressive learning ? In *ICML*, 2024.

[2] Ruiqi Zhang, Spencer Frei, and Peter L Bartlett. Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, 25(49):1–55, 2024.