Weekly Report (AUG 29, 2023 - Sep 3, 2023)

Zixun Xiong
Manning College of Information and Computer Sciences, UMASS Amherst
140 Governors Dr, Amherst, MA 01002
zixunxiong@umass.edu

Abstract

This week, I continued to read a paper that focused on the critical learning period (CLP) in a deep linear network. To understand the idea of the paper better, I reproduced some of the experiments in the paper. Then I read a paper on RLR attacks in federated learning (FL).

1 Critical Learning Periods Emerge Even in Deep Linear Networks

By applying a multi-path framework for a deep linear network, [1] demonstrated the critical learning period depends on the depth of the model and the structure of the target dataset. This paper eliminates the influence of factors like the learning rate scheduler by using a linear network as the research target. What's more, this paper also found that pre-training on specific tasks may damage the performance of the transferred model in a new task. This week, I continued to read this paper about the details of the experiments and the left part on CLP for matrix completion.

1.1 Learning Dynamics in reduced scalar differential equation highlight effect of competition

As shown in equation (1), the contribution of path a in the linear deep network $K_{a\alpha}$ can be updated by $q_{a\alpha}$ and $p_{a\alpha}$ (Since the relation of them in equation (1, we only need to update $p_{a\alpha}$). $q_{a\alpha}$ reflects the diagonal entries of the intermediary matrices ($d < D_a$), while $p_{a\alpha}$ represents the scale of diagonal entries of the final matrix ($d = D_a$). And $p_{a\alpha}$ can be updated by the differential equation (2). S_α is a fixed value of S from the SVD decomposition of $\Sigma_{xy} = \frac{1}{N}\sum_{i=1}^{N} y^{(i)}{x^{(i)}}^T$. According to the definition, $\overline{\Omega_\alpha} = p_{a\alpha} + p_{b\alpha}$. What's more $p_{a\alpha}$ is initialized as $p \sim \mathbf{N}(0, \delta)$.

$$\begin{aligned} K_{a\alpha} &= p_{a\alpha} q_{a\alpha}^{D_a - 1} \\ q_{a\alpha}^2 &= p_{a\alpha}^2 + 1 \end{aligned} \tag{1}$$

$$\tau \frac{d}{dt} p_{a\alpha} = q_{a\alpha}^{D_a - 1}[S_\alpha - \overline{\Omega_\alpha}] \tag{2}$$

According to the process mentioned above, I plotted Fig 1 for $S_\alpha = 10$. It's not hard to see that the deeper the network and δ are, the more extreme values are where one or the other pathway dominates for a fixed initial condition.

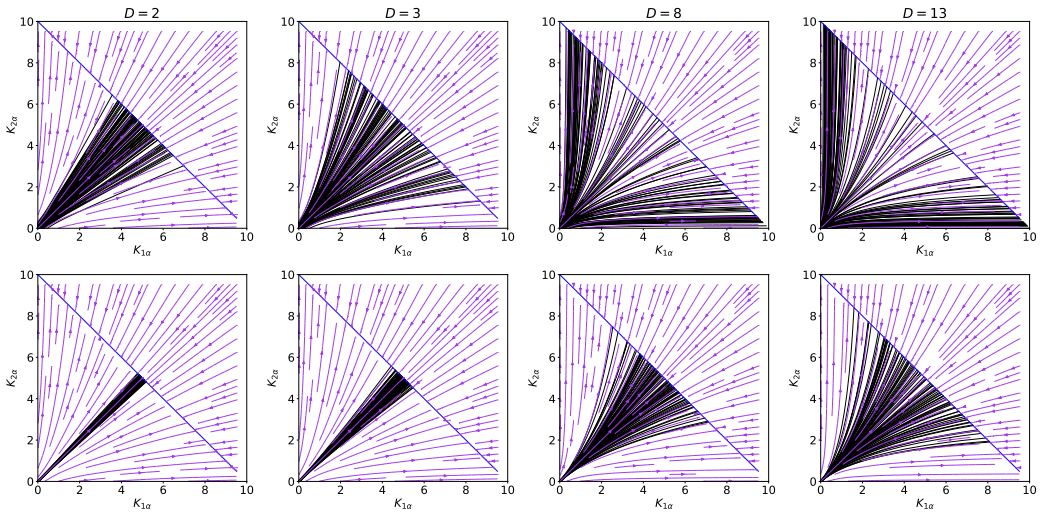

Figure 1: **Phase portraits for two pathway networks at different depths for a single** α. $D_a = D_b$. Initial conditions are normal distribution with zero means and $\delta = 0.1$ (top) and $\delta = 0.01$ (bottom).

Then I introduced deficits as mentioned in this paper for path b at different stages. Following the settings of this paper, I made $\delta = 0.01$, the learning rate as 0.001, and the max epochs as 1000. As indicated in Fig. 2, we can see that an early deficit will lead to the other path (path a in this case) to learn more about the feature. The deeper the network is, the more evident the phenomenon is. However, for later deficits, the phenomenon disappeared. This observation shed light on the possible explanation for the CLP as the similarities of the behaviors by varying deficits.

1.2 Deep Multi-Pathway Linear Neural Network Simulations

To verify that this phenomenon exists for all α in a learning process, the author plotted K_α for all α of both paths.

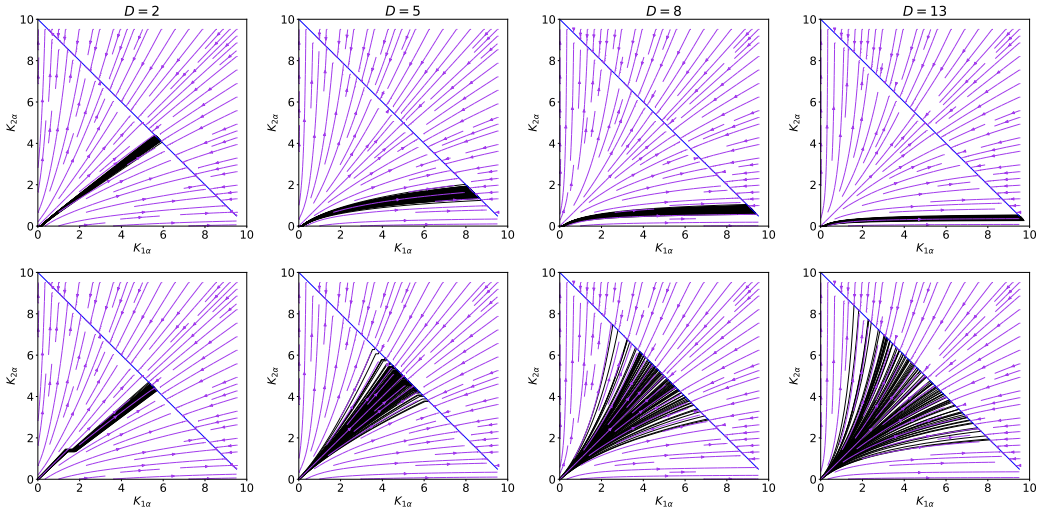

Figure 2: **Phase portrait for deficits at different stages and networks with different depths**. Deficits at an early stage (top: first 15 epochs) and deficits at a later stage (bottom: epoch 100 to 115).

I reproduced the experiment in this section and analyzed other metrics for further understanding.

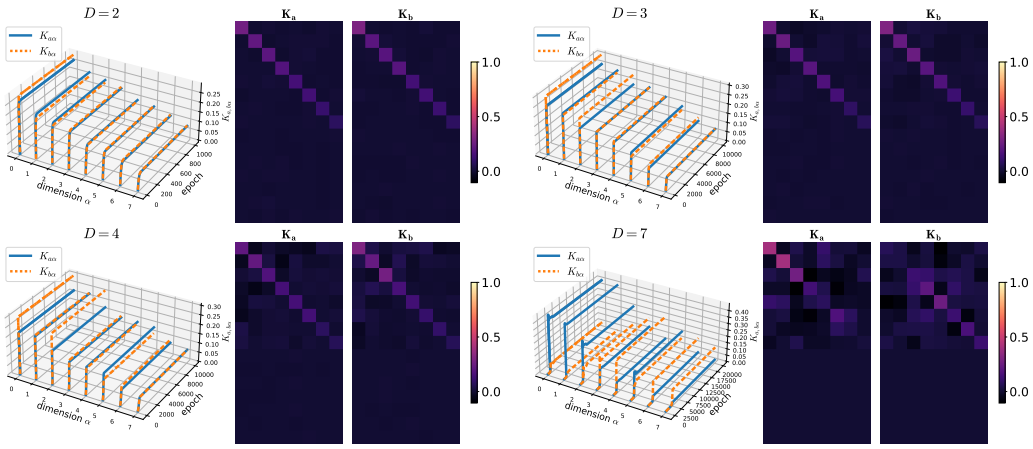

Figure 3: **Learning dynamics in two pathway networks with increasing depth**. From left to right, top to bottom, are learning dynamics for this deep linear network with different depths. Each part consists of 3 plots: (1) the leftist of three K values for different singular values for the SVD decomposition of 2 paths; the middle and the right are the final matrix of 2 paths.

Since in this paper, the training data are very special (identity matrix for X), I plotted Fig.3 with random training data (after whitening pre-processing) from a uniform distribution $U(0, 1)$ using **torch.rand**. As shown in Fig. 3, different paths learn different features (which is indicated by different K_a and K_b values). This observation is consistent with the last subsection. Then, I reproduced the experiment in this paper of gating to explore the influences of deficit in this dynamic system.

2 Future Plan

References

[1] Kleinman, Michael, Alessandro Achille, and Stefano Soatto. "Critical Learning Periods Emerge Even in Deep Linear Networks." arXiv preprint arXiv:2308.12221 (2023).