Emergent Low-Rank Training Dynamics in MLPs with Smooth Activations

postertitlebg.pdf

Alec Xu, Can Yaras, Matthew Asato, Qing Qu, Laura Balzano

University of Michigan, Ann Arbor

Background & Motivation

Low-Rank Training in Neural Networks

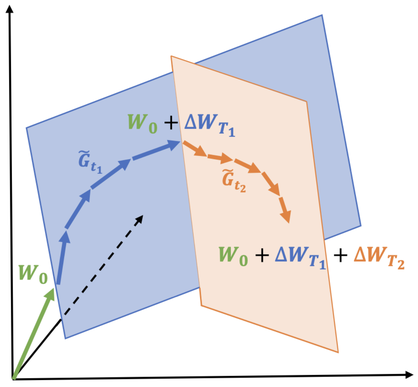


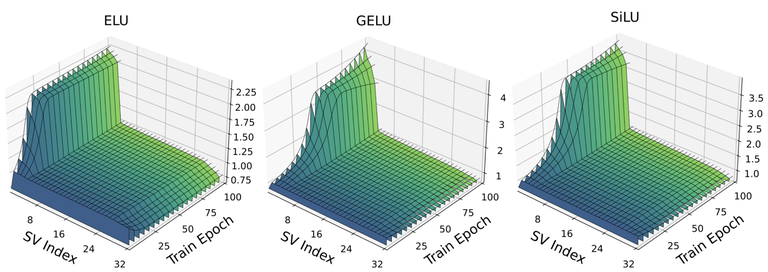
Figure: Taken from [1].

- Neural network training and fine-tuning mostly occur within low-dimensional subspaces [2-4].
- Previous work [5, 6] identified training subspaces in *linear* networks.
- In practice, neural networks are highly nonlinear!

Open question: in neural networks with nonlinearities, what low-dimensional subspaces does training largely occur within?

We focus our study on multi-layer perceptrons (MLPs).

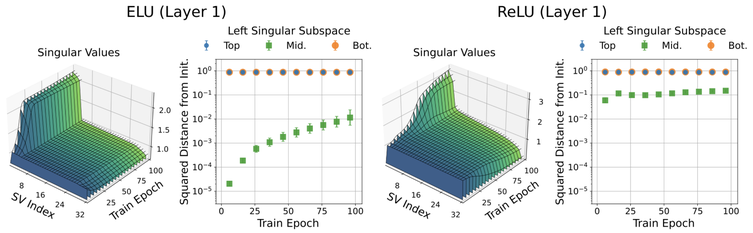
Contributions



- Show MLPs with *smooth* activations (e.g., GELU) exhibit stronger low-rank training dynamics compared to MLPs with nonsmooth activations (e.g., ReLU).
- Precisely characterize the low-dimensional subspaces in which the weights are trained.
- Leverage observations to develop low-rank MLP parameterization that mimics original model performance throughout training.

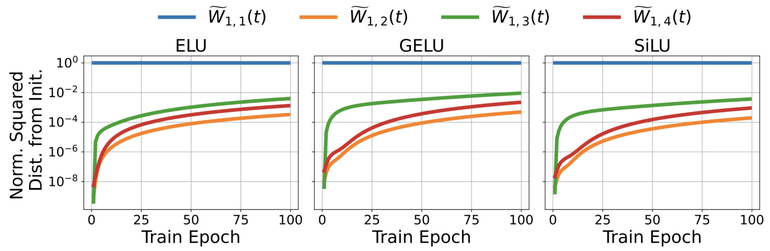
Emergent Low-Rank Training

Case study: *when do low-rank training dynamics emerge?*



Smooth activations encourage lower-rank training dynamics!

Theoretical Results: Two-Layer Networks



Data: Data matrix $\boldsymbol{X} \in \mathbb{R}^{d\times N}$ and labels $\boldsymbol{Y} \in \mathbb{R}^{K\times N}$ with $K \ll d$. Define $p := d - 2K$.

Model: Two-layer network with smooth ϕ and fixed second layer [7, 8]:

$$f_{\boldsymbol{W}_1}(\boldsymbol{X}) = \boldsymbol{W}_2\phi(\boldsymbol{W}_1\boldsymbol{X}), \tag{1}$$

where $\boldsymbol{W}_2 \in \mathbb{R}^{K\times m}$ and $\boldsymbol{W}_1 \in \mathbb{R}^{m\times d}$, with $m \geq d$.

Training: minimize squared-error loss via gradient descent

$$\min_{\boldsymbol{W}_1} \mathcal{L}(\boldsymbol{W}_1) := \frac{1}{2}\left\|f_{\boldsymbol{W}_1}(\boldsymbol{X}) - \boldsymbol{Y}\right\|_F^2, \quad \boldsymbol{W}_1(t+1) = \boldsymbol{W}_1(t) - \eta \boldsymbol{G}_1(t),$$

where $\boldsymbol{G}_1(t) := \nabla_{\boldsymbol{W}_1}\mathcal{L}(\boldsymbol{W}_1(t))$.

Theorem 1 (Informal). There exist orthogonal matrices $\boldsymbol{U} \in \mathbb{R}^{m\times m}$ and $\boldsymbol{V} \in \mathbb{R}^{d\times d}$ that depend only on $\boldsymbol{W}_1(0)$ and $\boldsymbol{G}_1(0)$ such that, for all $t \geq 0$:

$$\boldsymbol{W}_1(t) = \boldsymbol{U}\widetilde{\boldsymbol{W}}_1(t)\boldsymbol{V}^\top = \boldsymbol{U}\begin{bmatrix}\widetilde{\boldsymbol{W}}_{1,1}(t) & \widetilde{\boldsymbol{W}}_{1,2}(t)\\ \widetilde{\boldsymbol{W}}_{1,3}(t) & \widetilde{\boldsymbol{W}}_{1,4}(t)\end{bmatrix}\boldsymbol{V}^\top, \tag{2}$$

where $\widetilde{\boldsymbol{W}}_{1,1}(t) \in \mathbb{R}^{(m-p)\times 2K}$, $\widetilde{\boldsymbol{W}}_{1,2}(t) \in \mathbb{R}^{(m-p)\times p}$, $\widetilde{\boldsymbol{W}}_{1,3}(t) \in \mathbb{R}^{p\times 2K}$, and $\widetilde{\boldsymbol{W}}_{1,4}(t) \in \mathbb{R}^{p\times p}$, with

$$\frac{\|\widetilde{\boldsymbol{W}}_{1,i}(t+1) - \widetilde{\boldsymbol{W}}_{1,i}(t)\|_F}{\sqrt{d-2K}} \lesssim \eta \cdot \rho(t),$$

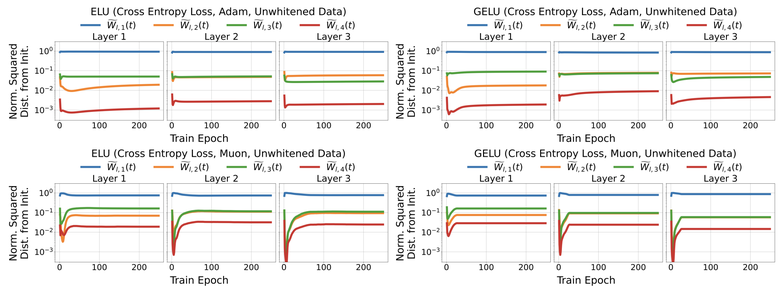
for all $i \in \{2,3,4\}$, where $\rho(t)$ depends on the **alignment of the singular subspaces between $\boldsymbol{G}_1(t)$ and $\boldsymbol{G}_1(0)$.**

Experimental Results

Now consider

$$f_{\boldsymbol{\Theta}}(\boldsymbol{X}) = \boldsymbol{W}_L\phi\left(\boldsymbol{W}_{L-1}\phi\left(\cdots \boldsymbol{W}_2\phi\left(\boldsymbol{W}_1\boldsymbol{X}\right)\cdots\right)\right), \tag{3}$$

where $\boldsymbol{W}_1 \in \mathbb{R}^{m\times d}$, $\boldsymbol{W}_l \in \mathbb{R}^{m\times m}$ for $l = 2, \ldots, L-1$, and $\boldsymbol{W}_L \in \mathbb{R}^{K\times m}$.

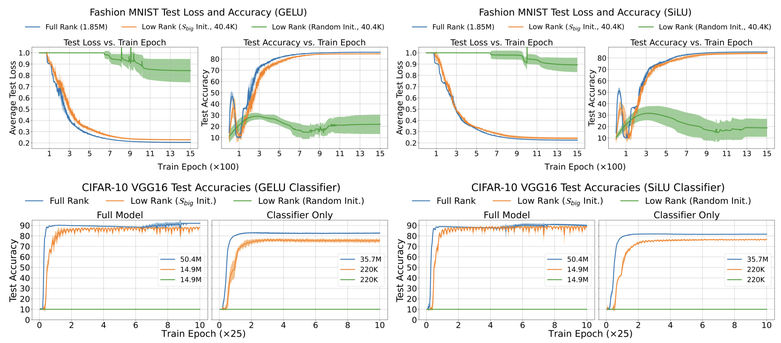


Low-rank training holds for deep networks, other loss functions, and other optimizers! For $l \in \{1, 2, \ldots, L-1\}$, each $\boldsymbol{W}_l$ has own orthogonal $\boldsymbol{U}_l$ and $\boldsymbol{V}_l$ analogous to those in Theorem 1.

There exists a **low-rank MLP** that mimics performance of (3):

$$\tilde{f}_{\tilde{\boldsymbol{\Theta}}}(\boldsymbol{X}) = \boldsymbol{W}_L\phi\left(\widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{W}}_{L-1}\phi\left(\widetilde{\boldsymbol{W}}_{L-2}\phi\left(\cdots\widetilde{\boldsymbol{W}}_2\phi\left(\widetilde{\boldsymbol{W}}_1\widetilde{\boldsymbol{V}}^\top\boldsymbol{X}\right)\cdots\right)\right)\right),$$

where $\widetilde{\boldsymbol{U}} \in \mathbb{R}^{m\times r}$, $\widetilde{\boldsymbol{W}}_l \in \mathbb{R}^{r\times r}$, and $\widetilde{\boldsymbol{V}} \in \mathbb{R}^{r\times d}$. From above, **initialize $\widetilde{\boldsymbol{U}}(0) := \boldsymbol{U}_{L-1}$ and $\widetilde{\boldsymbol{V}}(0) := \boldsymbol{V}_1$.**



Future Work

- Low-rank MLPs in Transformer fine-tuning and MLP-Mixer training.
- Explore low-rank training in other mechanisms, e.g., self-attention.

References

[1] Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, Yuandong Tian. "GaLore: Memory-Efficient LLM Training by Gradient Low-Rank Projection." ICML 2024.
[2] Guy Gur-Ari, Daniel A. Roberts, Ethan Dyer. "Gradient Descent Happens in a Tiny Subspace." arXiv:1812.04754.
[3] Chunyuan Li, Heerad Farkhoor, Rosanne Liu, Jason Yosinski. "Measuring the Intrinsic Dimension of Objective Landscapes." ICLR 2018.
[4] Zhong Zhang, Bang Liu, Junming Shao. "Fine-tuning Happens in Tiny Subspaces: Exploring Intrinsic Task-specific Subspaces of Pre-trained Language Models." ACL 2023.
[5] Can Yaras, Peng Wang, Laura Balzano, Qing Qu. "Compressible Dynamics in Deep Overparameterized Low-Rank Learning & Adaptation." ICML 2024.
[6] Soo Min Kwon, Zekai Zhang, Dogyoon Song, Laura Balzano, Qing Qu. "Efficient Low-Dimensional Compression of Overparameterized Models." AISTATS 2024.
[7] Spencer Frei, Gal Vardi, Peter Bartlett, Nathan Srebro, Wei Hu. "Implicit Bias in Leaky ReLU Networks Trained on High-Dimensional Data." ICLR 2023.
[8] Yiwen Kou, Zixiang Chen, Quanquan Gu. "Implicit Bias of Gradient Descent for Two-layer ReLU and Leaky ReLU Networks on Nearly-orthogonal Data." NeurIPS 2023.