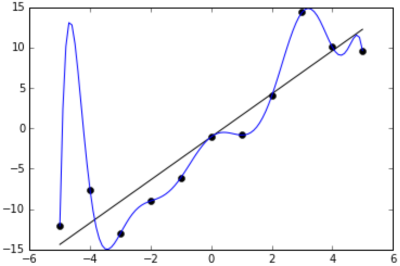
Implicit Bias of Gradient Flow for Two-layer ReLU Networks Trained on Nearly-orthogonal Data

Jingzhi Sun and Chau Tran

March 13, 2023

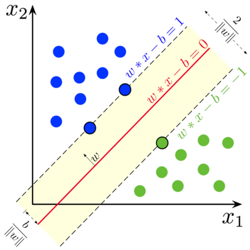
Implicit bias

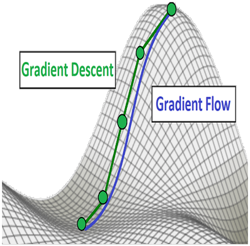
- Most neural networks are overparameterized, i.e. have more parameters than the number of examples
- Can interpolate the training data
- Still generalize well to unseen test data
- *Implicit bias*



Implicit bias

- Implicit bias of gradient-based optimization has been a focus of many recent studies
- In this presentation:
  - Two-layer ReLU networks: $f(x; W) = \sum_{j=1}^{m} a_j \phi(\langle x; w_j \rangle + b_j)$
  - Trained by Gradient Flow (GF) on binary-classification dataset
  - Exponentially-tailed loss (logistic loss, exponential loss)





Implicit bias

Theorem: Lyu and Li (2020), Ji and Telgarsky (2020)

For neural networks, under some conditions, GF converges in direction to the a KKT point of the maximum-margin problem:

$$\min_{\theta} \frac{1}{2}\|\theta\| \qquad \text{s.t} \qquad \forall i \in [n], y_i f(x_i, \theta) \geq 1.$$

Moreover, $\hat{L}(\theta^{(t)}) \to 0$ and $\left\|\theta^{(t)}\right\| \to \infty$ as $t \to \infty$.

- $\theta^{(t)}$ converges in direction to $\tilde{\theta}$ if $\lim_{t\to\infty} \frac{\theta^{(t)}}{\left\|\theta^{(t)}\right\|} = \frac{\tilde{\theta}}{\left\|\tilde{\theta}\right\|}$.

Implicit bias

- KKT Conditions: there exist $\lambda_1, \ldots, \lambda_n \geq 0$ such that

$$\theta = \sum_{i=1}^{n} \lambda_i y_i \nabla_\theta f(x_i; \theta),$$
$$\forall i \in [n], y_i f(x_i; \theta) \geq 1,$$
$$\forall i \in [n], \lambda_i = 0 \text{ if } y_i f(x_i; \theta) > 1.$$

- To show the implicit bias of GF in the limit $t \to \infty$, we study the properties of KKT point of the two-layer ReLU trained with GF.
- Let W be a KKT point of the max-margin problem, we attempt to show the following properties of W:
  1. $y_i f(x_i; W) = 1$ for all $i \in [n]$.
  2. $\limsup_{t \to \infty} \mathrm{StableRank}(W) \leq 2$.

Proof idea

- Since W is a KKT point, then there exist $\lambda_i \geq 0$ for $i \in [n]$ such that for $j \in [m]$

$$w_j = \sum_{i=1}^{n} \lambda_i \nabla_{w_j}(y_i f(x_i; W)) = a_j \sum_{i=1}^{n} \lambda_i y_i \phi'_{i,j} x_i,$$
$$b_j = \sum_{i=1}^{n} \lambda_i \nabla_{w_j}(y_i f(x_i; W)) = a_j \sum_{i=1}^{n} \lambda_i y_i \phi'_{i,j}.$$

Follow the framework in Vardi et al. (2022) [1] to prove the strictly positive lower bounds for λ_i. Since Vardi et al. (2022) assume nearly-orthogonality of training data, the proof should be similar.

[1] Vardi, Gal, Gilad Yehudai, and Ohad Shamir. "Gradient methods provably converge to non-robust networks." NeurIPS 2022.

Proof idea

- To prove the low-rank bias, we want to show that at time step t,

$$\left\|W^{(t)}\right\|_F^2 \leq 2\left\|W^{(t)}\right\|_2^2 + \left\|\nabla_W \hat{L}(W^{(t)})\right\|_F^2.$$

  We have $\hat{L}(W^{(t)}) \to 0$ and $\left\|W^{(t)}\right\|_F \to \infty$ as $t \to \infty$. Therefore, we get $\limsup_{t\to\infty} \mathrm{StableRank}(W) \leq 2$. With that said, it is unclear to us how to proceed with the proof.

Numerical experiments

Synthetic-data experiments:

- We consider generate a mixture of Gaussian data as previously described in Kou et al. (2023) [2] as follow:
    1. y_i is generated from the Rademacher distribution, i.e. $\Pr(y_i = -1) = \Pr(y_i = 1) = 1/2$.
    2. x_i is generated by $x_i = y_i\mu + z_i$, where $z_i \sim N(0, \sigma_e^2 I_d)$.
- $n = 10$ and $d = 784$. $\mu \sim N(0, \sigma_p^2 I_d)$ where $\sigma_p = 0.01$.
- For $i = 1, \ldots, n, z_i \sim N(0, \sigma_e^2 I_d)$ where $\sigma_e = 1$.
- Number of neurons is $m = 100$.
- We train the model with gradient descent with step size $\alpha = 0.0001$ for 5000 epochs.

[2] Kou, Yiwen, Zixiang Chen, and Quanquan Gu. "Implicit Bias of Gradient Descent for Two-layer ReLU and Leaky ReLU Networks on Nearly-orthogonal Data." NeurIPS 2023.

Numerical experiments

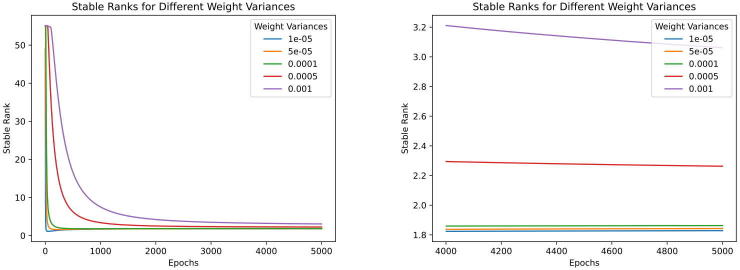


Figure 1: Left: Stable rank of two-layer ReLU networks with different weight initialization variances. **Right**: Stable rank from the last 1000 epochs.

High-Dimensional Non-Orthogonal Data

- Analysis of the German Neuroblastoma Trials dataset (NB90-NB2004).
- Dataset of 251 patients, age 0-296 months. ($n = 251$)
- Each patient's data includes 10,707 data points from oligonucleotide(DNA or RNA) microarrays. ($p = 10707$)
- Objective: Predict survival beyond a 3-year trial period (0-1).
- The selection of oligonucleotide microarrays from proximate genes that exhibit a pronounced correlation structure, this dataset is characterized as **non-orthogonal**.
- Number of neurons is $m = 1000$.
- We initialize the first-layer weights with i.i.d. mean zero Gaussians with standard deviation $\sigma \in \{0.00001, 0.00005, 0.0001, 0.0005, 0.001\}$.

Findings: Neuroblastoma Data

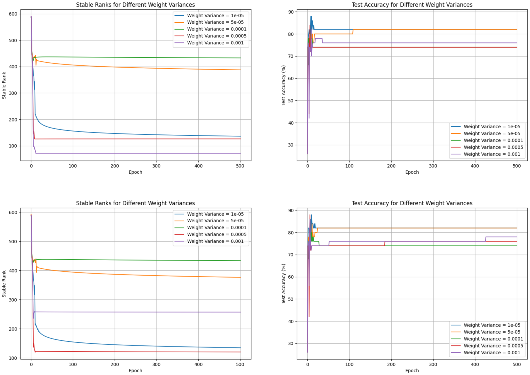


Figure 2: Above: ReLU **Below:** Leaky-ReLU

- Stable Rank do not follow the order of initial σ settings.
- ReLU can achieve lower stable rank.

Low-Dimensional Orthogonal Data

- Focus on the Autism Dataset from Next Generation Sequencing.
- Comprises 104 samples: 47 autisms and 57 healthy controls.
- Analysis of expressions from the top 5 differently expressed genes,identified from an extensive dataset of over 60,000 gene expression profiles. ($p = 5$)
- Objective: Distinguish between autism and healthy conditions. (0-1)
- These selected genes are regarded as orthogonal, based on the analysis presented by Fan et al. (2021) [3].

[3] Jianqing Fan and Weichen Wang and Ziwei Zhu, A Shrinkage Principle for Heavy-Tailed Data: High-Dimensional Robust Low-Rank Matrix Recovery. Annals of Statistics, 2021.

Findings: Autism Data

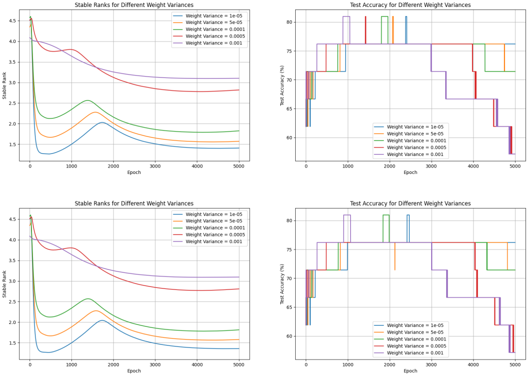


Figure 3: Above: ReLU **Below:** Leaky-ReLU

- For the same initial sigma (σ) settings, both cases converged to same stable rank values.
- Stable Rank : Decrease $\rightarrow$ Increase $\rightarrow$ Decrease.

MNIST Dataset: Findings

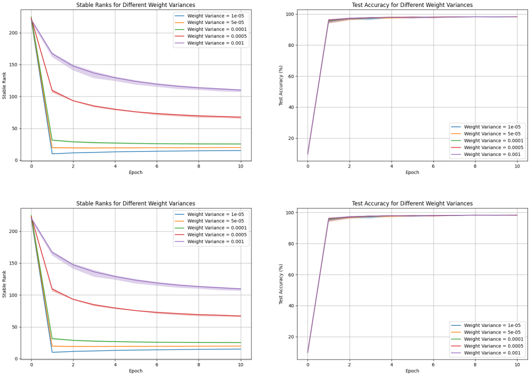


Figure 4: Above: ReLU **Below:** Leaky-ReLU

- The choice of initialization variance is important.
- The order of stable rank follows the order of initial σ.

High-Dimensional Nearly-Orthogonal Data: CIFAR10

- Frei et al.[4] established the experiment using 2 layers leaky-ReLU with Glorot Uniform initial settings.
- We initialize the first-layer weights with i.i.d. mean zero Gaussians with standard deviation $\sigma \in \{0.00001, 0.00005, 0.0001, 0.0005, 0.001\}$ and trained on 2-layer ReLu.
- We train NN with SGD with batch size 128 and a learning rate of $\alpha = 0.01$ for 100 epochs.

[4] Frei, S., Vardi, G., Bartlett, P., Srebro, N., & Hu, W. (2023). Implicit bias in Leaky ReLU networks trained on high-dimensional data. In The Eleventh International Conference on Learning Representations.

CIFAR10 Dataset: Findings

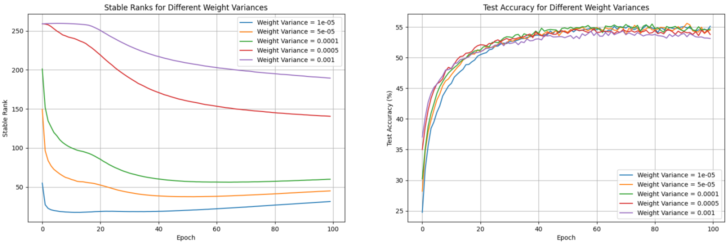


Figure 5: Stable rank of ReLU networks on CIFAR10 among different initial settings

Interesting Findings among Experiments

- There are two conditions of our dataset: Nearly-orthogonal v.s. Non-Orthogonal and High dimension v.s. Low dimension.
- If the dataset is not Nearly-orthogonal, then the order of stable ranks do not follow the order of our initial variance settings. (No matter ReLU or Leaky ReLU)

CIFAR10 Dataset: Findings

We choose $\sigma = 0.00001$, and trained for 2000 epochs on ReLU.

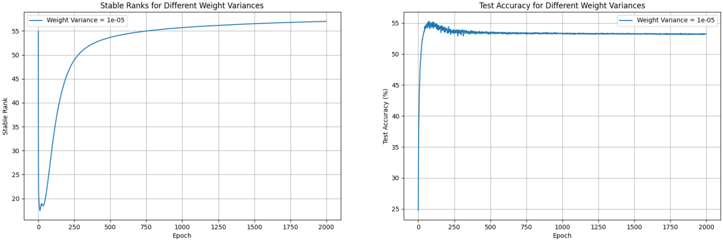


Figure 6: Stable rank of ReLU networks on CIFAR10

- A sever increase in stable rank after 100 epochs.

Conclusion and Future Work

- In this project, we study the implicit bias of gradient flow for two-layer ReLU networks trained on nearly-orthogonal data.
- We provide numerical experiments to show the implicit bias of gradient descent with small learning rate toward low-rank networks.
- Important future work is to provide the formal proof of the low-rank bias.
- Particularly, we will explore the relationship between the neuron alignment and neuron activation patterns in ReLU networks.