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

Jingzhi Sun
Department of Statistics
University of California, Davis
Davis, CA 95616

Chau Tran
Department of Statistics
University of California, Davis
Davis, CA 95616

Abstract

Implicit bias of gradient-based optimization for deep neural networks is a well-known phenomenon. In this work, we study the implicit bias of gradient flow for two-layer ReLU networks trained on nearly-orthogonal data. In particular, we study the properties of homogeneous ReLU networks trained by minimizing exponentially-tailed classification losses. We provide numerical experiments to show the implicit bias of gradient flow toward low-rank networks.

1 Introduction

Gradient-based optimization methods is the workhorse for training deep neural networks. In particular, the neural network is trained by using gradient descent (and its variants) to minimize a loss function. It is well-known that gradient-based methods exhibit implicit bias toward solutions that has desirable properties. For example, gradient descent often finds simple models with low-complexity that can generalize well on unseen data. Recently, implicit bias of gradient-based optimization methods has been a focus of many research studies [Lyu and Li, 2020, Ji and Telgarsky, 2020, Vardi, 2023, Frei et al., 2023a,b].

In this work, we study the implicit bias of gradient flow for two-layer ReLU networks trained on nearly-orthogonal data. In particular, we consider the fully-connected two-layer ReLU network

$$f(x;W)=\sum_{j=1}^{m} a_j\phi(\langle x;w_j\rangle+b_j),$$

where $m\in\mathbb{N}$ is the number of neurons, $\phi(x)=\max\{0,x\}$, $W\in\mathbb{R}^{m,d}$ is the first-layer weight matrix with rows $w_j^\top$. We study implicit bias of gradient flow, i.e. gradient descent with learning rate tends to zero, for the two-layer ReLU networks above. We consider the setting when training data $\{x_i\}_{i=1}^n$ are nearly-orthogonal, i.e. $\|x_i\|_2^2\geq Cn\max_{i\neq j}|\langle x_i,x_j\rangle|$ for some large absolute constant C. For theoretical analysis, we leverage the directional convergence of gradient flow on homogeneous networks, such as two-layer ReLU networks, to a the Karush–Kuhn–Tucker (KKT) points of the margin-maximization problem [Lyu and Li, 2020, Ji and Telgarsky, 2020]. We provide numerical experiments to show the implicit bias of gradient flow toward low-rank networks.

2 Related Work

Implicit bias in neural networks. Implicit bias of gradient flow has been a focus of many recently studies. Seminal works by Lyu and Li [2020] and Ji and Telgarsky [2020] show the directional convergence of homogeneous neural networks trained with trained with exponentially-tailed classification losses to the KKT point of the margin-maximization problem. Frei et al. [2023b] show that

Preprint. Under review.

two-layer Leaky ReLU networks trained with gradient flow on nearly-orthogonal data converge to a KKT point that has rank at most two and has a linear decision boundary. Vardi et al. [2022] show that two-layer ReLU network trained with gradient flow on nearly-orthogonal data converge to a network that is non-robust even though robust network exists. Frei et al. [2023a] study the implications of the implicit bias of gradient flow for two-layer ReLU networks on generalization and adversarial robustness. Min et al. [2023] study the neuron alignment in two-layer ReLU networks trained with gradient flow when the training data with the same label are positively correlated and show that gradient flow converges to a networks that is low-rank.

Linearly separable and nearly-orthogonal data. The nearly-orthogonal condition on training data implies that data are linearly separable. Vardi et al. [2022] consider the two-layer ReLU networks trained by gradient flow on nearly-orthogonal data. This is the same setting that we consider for this work. Properties of KKT point of the margin-maximization problem when training data are nearly-orthogonal are also previously studied in Frei et al. [2023b,c,a]. Kou et al. [2023] also consider the setting where training data are nearly-orthogonal, but they study the implicit bias of gradient descent instead of gradient flow.

3 Preliminaries

Notation. For a vector x, we denote $\|x\|$ as the Euclidean norm. For a matrix A, we denote $\|A\|_F$ as the Frobenius norm, and we denote $\|A\|_2$ as the spectral norm. For an integer n, we denote $[n] = \{1, \dots, n\}$.

Two-layer ReLU network. In this work, we consider the fully-connected two-layer ReLU network

$$f(x; W) = \sum_{j=1}^{m} a_j \phi(\langle x; w_j \rangle + b_j),$$

where $m \in \mathbb{N}$ is the number of neurons, $\phi(x) = \max\{0, x\}$ is the ReLU activation function, $W \in \mathbb{R}^{m,d}$ is the first-layer weight matrix with rows $w_j^\top \in \mathbb{R}^d$, $a_j \in \{\pm 1/\sqrt{m}\}$ is the second-layer weights, $b_j \in R$ is the bias.

Gradien flow. Let $S = \{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \{\pm 1\}$ be the binary classification training data. Let $f(\cdot, W) : \mathbb{R}^d \to \mathbb{R}$ be the neural network parameterized by W. Let $l : \mathbb{R} \to \mathbb{R}$ be a loss function. Then, the empirical loss of $f(\cdot, W)$ on S is

$$\hat{L}(W) = \frac{1}{n} \sum_{i=1}^{n} l(y_i f(x_i, W)).$$

We consider exponentially-tailed losses such that the exponential loss $l(q) = \exp(-q)$ and the logistic loss $l(q) = \log(1 + \exp(-q))$. For a learning rate $\alpha > 0$, each gradient descent update at time t has the form

$$W^{(t+1)} = W^{(t)} - \alpha \nabla_W \hat{L}(W^{(t)}).$$

Gradient flow is gradient descent with learning rate tends to zero. From an initial point $W^{(0)}$, the dynamic of $W^{(t)}$ is given by $\dfrac{dW^{(t)}}{dt} = -\nabla_W \hat{L}(W^{(t)})$.

4 Main Results

In this work, we study the implicit bias of gradient flow for two-layer ReLU networks when the number of gradient steps t tends to infinity. We leverage the directional convergence of gradient flow on homogeneous networks, such as two-layer ReLU networks, to the Karush–Kuhn–Tucker (KKT) points of the margin-maximization problem [Lyu and Li, 2020, Ji and Telgarsky, 2020]. A network is *homogeneous* if there exists $L > 0$ such that for all $\eta > 0$, we have $f(x; \eta\theta) = \eta^L f(x; \theta)$. The two-layer ReLU networks that we consider in this work are homogeneous with $L = 2$. With the trajectory of $\theta^{(t)}$ as defined above, we say that trajectory $\theta^{(t)}$ converges in direction to $\tilde{\theta}$ if $\lim_{t\to\infty} \dfrac{\theta^{(t)}}{\left\|\theta^{(t)}\right\|} = \dfrac{\tilde{\theta}}{\left\|\tilde{\theta}\right\|}$. Note that for our two-layer ReLU network, $\theta = \text{vec}(W)$, where $\text{vec}(\cdot)$ is

the vectorize operator. Then, the theorem on directional convergence of homogeneous networks of Lyu and Li [2020], Ji and Telgarsky [2020] can be stated as follow

Theorem 4.1. *(Paraphrased from Lyu and Li [2020], Ji and Telgarsky [2020]) Let $f(x;\theta)$ be a homogeneous ReLU network parameterized by θ. Consider minimizing exponentially-tailed losses such as the exponential loss and the logistic loss over the binary classification training dataset $S = \{(x_i, y_i)\}_{i=1}^n$ using gradient flow. Suppose there exists a time t such that $\hat{L}(\theta^{(t)}) < \frac{\log(2)}{n}$. Then, gradient flow converges in direction to a KKT point of the following maximum-margin problem:*

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

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

Since the two-layer ReLU networks are homogeneous, we can show the implicit bias of gradient flow by studying the properties of KKT point of the maximum-margin problem. Let W be a KKT point of Problem (1), we attempt to show the following properties of W:

1. $y_i f(x_i; W) = 1$ for all $i \in [n]$.
2. $\limsup_{t\to\infty} \mathsf{StableRank}(W) \leq 2$,

where $\mathsf{StableRank}(W^{(t)}) = \left\|W^{(t)}\right\|_F^2 / \left\|W^{(t)}\right\|_2^2$ [Rudelson and Vershynin, 2007].

We now discuss our proof idea. Since W is a KKT point of Problem (1), then there exist $\lambda_i \geq 0$ for $i \in [n]$ such that for $j \in [m]$

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

where $\phi'_{i,j}$ is the subgradient of ϕ at $(\langle x_i; w_j\rangle + b_j)$, i.e., $\phi'(a) = \mathrm{sign}(a)$ if $a \neq 0$, and $\phi'(a) \in [0, 1]$ if $a = 0$. Note that $\lambda_i = 0$ if $y_i f(x_i; W) \neq 1$. We also have

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

In order to show the first property that $y_i f(x_i; W) = 1$ for all $i \in [n]$, it is sufficient to show that $\lambda_i > 0$ for all i. Then, one can follow the framework in Vardi et al. [2022] 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.

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

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

By Theorem 4.1, we have $\hat{L}(W^{(t)}) \to 0$ and $\left\|W^{(t)}\right\|_F \to \infty$ as $t \to \infty$. Therefore, we get $\limsup_{t\to\infty} \mathsf{StableRank}(W) \leq 2$. With that said, it is unclear to us how to proceed with the proof.

5 Numerical Experiments

In this section, we present numerical experiments of both synthetic and real datasets.

Synthetic-data experiments. We consider generate a mixture of Gaussian data as previously described in Kou et al. [2023]: Let $\mu \in \mathbb{R}^d$ be a fixed vector. We generate a dataset $S = \{(x_i, y_i)\}_{i=1}^n$, with $x_i \in \mathbb{R}^d$ and $y_i \in \{\pm 1\}$, as follows

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 $z_i \sim N(0, \sigma_e^2 I_d)$.

We set the sample size $n = 10$ and dimension $d = 784$. We set μ to be a random sample from $N(0, \sigma_p^2 I_d)$ where $\sigma_p = 0.01$. We then generate z_i from $N(0, \sigma_e^2 I_d)$ where $\sigma_e = 1$. We

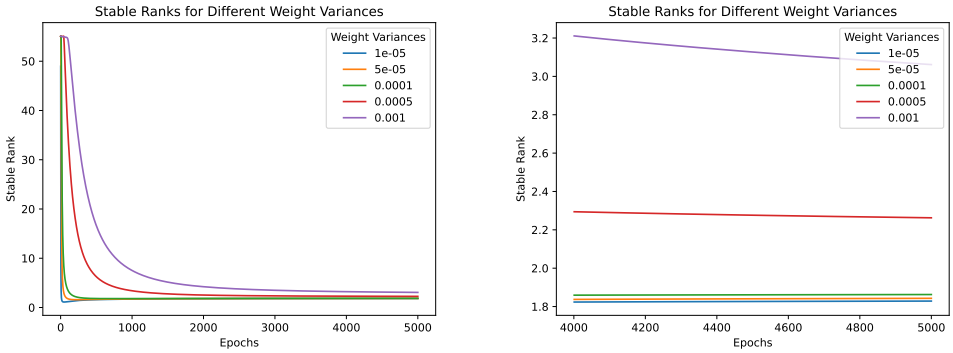

Figure 1: **Left**: Stable rank of two-layer ReLU networks with different weight initialization variances. **Right**: Stable rank from the last 1000 epochs.

consider the two-layer ReLU networks with biases, and the number of neurons is $m = 100$. We initialize the first-layer weights with i.i.d. mean zero Gaussians with standard deviation $\sigma \in \{0.00001, 0.00005, 0.0001, 0.0005, 0.001\}$. We train the model with gradient descent with step size $\alpha = 0.0001$ for 5000 epochs. Figure 1 shows the stable rank of the first-layer weights. The figures suggests that the stable rank decreases to a small value when number of epochs tends to infinity. In our simulations, with a small weight variances, the stable ranks stay below 2.

Real Dataset Experiments on High-Dimension & Non-Orthogonal Dataset. In our study, we analyze a dataset comprising 251 patients from the German Neuroblastoma Trials NB90-NB2004, spanning diagnoses made between 1989 and 2004. The age range of these patients is from 0 to 296 months, with a median age of 15 months. Each patient's data includes an oligonucleotide microarray encompassing $10,707$ data points. The specifics of these trials are meticulously documented by Berthold et al. [2017]. The dataset presents a sample size of $n = 251$ and a dimension of $d = 10,707$. Our objective is to predict the binary indicator (0-1) reflecting whether a patient survives beyond the 3-year trial period. Given the selection of oligonucleotide microarrays from proximate genes that exhibit a pronounced intrinsic correlation structure, this dataset is characterized as non-orthogonal.

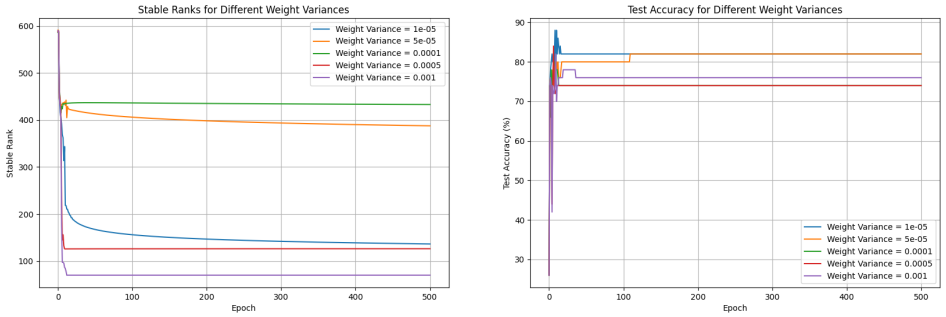

Figure 2: Stable rank of two-layer ReLU networks with different weight initialization variances on Neuroblastoma data.

We consider the two-layer ReLU and Leaky ReLU (with a slope of $\gamma = 0.01$) networks with biases, and the number of neurons is $m = 1000$. We initialize the first-layer weights with i.i.d. mean zero Gaussians with standard deviation $\sigma \in \{0.00001, 0.00005, 0.0001, 0.0005, 0.001\}$. We train the model with full-batch gradient descent and learning rate $\alpha = 0.22$ for 50 epochs.

Figure 2 illustrates that within the ReLU activation function framework, the configuration with the highest initial standard deviation ($\sigma = 0.001$) achieves convergence to the lowest stable rank. This contrasts with the other four settings, which also reach convergence, albeit at higher stable rank

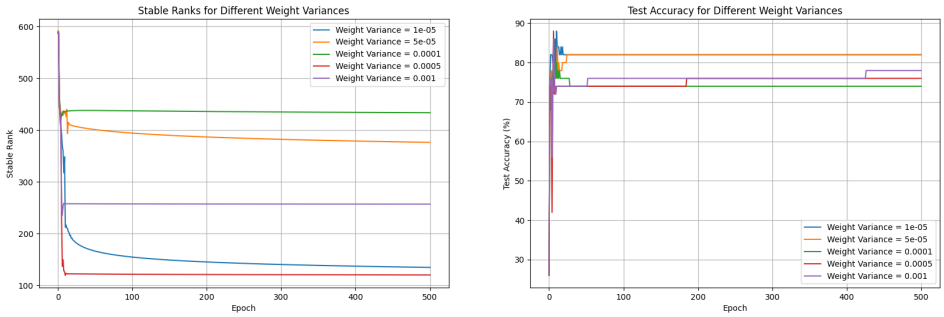

Figure 3: Stable rank of two-layer Leaky-ReLU networks with different weight initialization variances on Neuroblastoma data.

values. Furthermore, once the stable rank reduction begins, it converges to the value at a rapid fast speed.

Figure 3 demonstrates that, when utilizing the Leaky-ReLU activation function, the setup characterized by the second highest initial standard deviation ($\sigma = 0.0005$) successfully attains convergence at the minimal stable rank observed within this study. The lowest stable rank we can achieve through this initial settings is lower on ReLU than Leaky-ReLU.

Real Dataset Experiments on Low-Dimension & Orthogonal Dataset. Autism Dataset are measured among 104 samples: 47 autisms and 57 healthy controls [Gupta et al., 2014]. We select expressions from the top 5 differently expressed genes, identified from an extensive dataset of over 60,000 gene expression profiles $(n = 104, d = 5)$. These selected genes are regarded as orthogonal in nature, based on the analysis presented by Fan et al. [2017]. Our primary objective is to employ this dataset to distinguish between autism and healthy conditions.

We examine neural networks utilizing both ReLU and Leaky-ReLU (with a slope of $\gamma = 0.01$) activation functions, incorporating biases, and configuring the hidden layer width to $m = 128$ neurons. The initial weights of the input-to-hidden layer are Gaussian distributed, with a mean of 0 and standard deviations $\sigma \in \{0.00001, 0.00005, 0.0001, 0.0005, 0.001\}$. We train the network with full-batch gradient descent with a learning rate of $\alpha = 0.01$ for 5000 epochs.

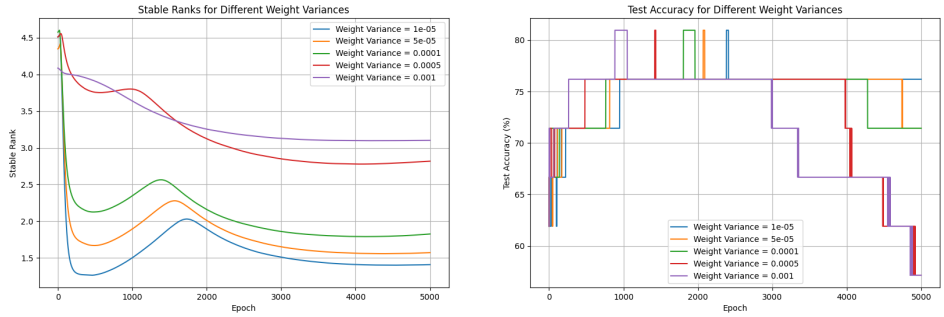

Figure 4: Stable rank of two-layer ReLU networks with different weight initialization variances on Autism data.

The curves representing the variation of stable rank with the increase in epoch number, as depicted in Figure 4 (ReLU)and Figure 5 (Leaky-ReLU), exhibit remarkably similar patterns. For identical initial σ settings, convergence to equivalent values was observed in both cases. Each curve demonstrates a trend of initial decrease, followed by an increase, and then a subsequent decrease. As in our previous experiments, it was observed that once convergence of the stable rank initiates, it proceeds at a rapid pace. Despite observing significant overfitting beyond 3000 epochs, characterized by a marked

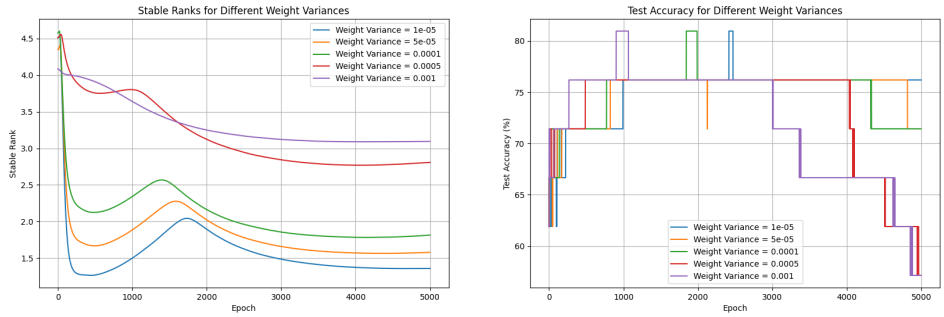

Figure 5: Stable rank of two-layer Leaky-ReLU networks with different weight initialization variances on Autism data.

decline in the test set accuracy, the stable rank of our model continued to exhibit a convergence trend with only a marginal increase.

Real-data Experiments on MNIST dataset. In this study, we focus on the training of a two-layer feed-forward neural network, leveraging both ReLU and Leaky-ReLU ($\gamma = 0.01$) activation functions on MNIST. The network's hidden layer width is set to $m = 1000$. The initialization of input-to-hidden layer weights employs a Gaussian distribution, with a mean of 0 and varying standard deviations $\sigma \in \{0.00001, 0.00005, 0.0001, 0.0005, 0.001\}$.Training is conducted using stochastic gradient descent (SGD), with a batch size of 64 and a learning rate of 0.1, over the course of 10 epochs. This experiment is designed by Kou et al. [2023].

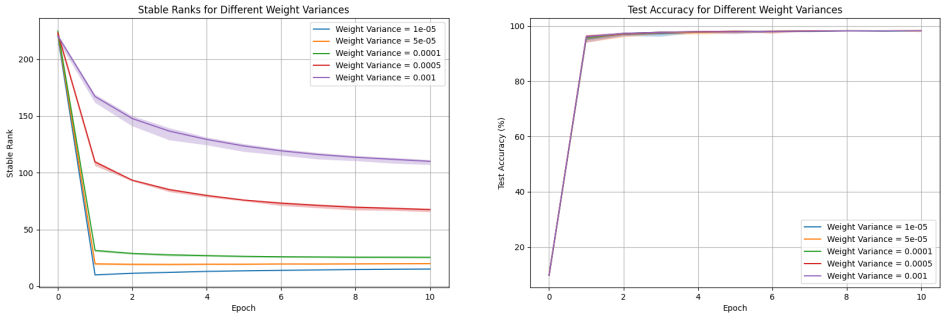

Figure 6: Stable rank of two-layer ReLU networks with different weight initialization variances on MNIST.

Derived from the insights provided by Figure 6 and Figure 7, it becomes evident that the stable rank of networks employing ReLU or Leaky-ReLU activation functions is significantly influenced by both the initialization parameters and the duration of training. Specifically, when the initialization is set to a sufficiently low threshold, we observe a rapid decrease in the stable rank, culminating in a value markedly lower than that of its initial state. This phenomenon underscores the critical role that the choice of initialization plays in the dynamics of network training, particularly in its capacity to facilitate a swift convergence to a more compact and efficient representational structure within the network.

Real-data Experiments on CIFAR-10 dataset. We also consider training the two-layer neural networks on the CIFAR-10 dataset. Frei et al. [2023b] use standard Dense layer initialization in Leaky-Relu using TensorFlow Keras(Glorot Uniform) for the initial settings. We consider the two-layer ReLU networks with biases, and the number of neurons is $m = 128$. We initialize the first-layer weights with i.i.d. mean zero Gaussians with standard deviation

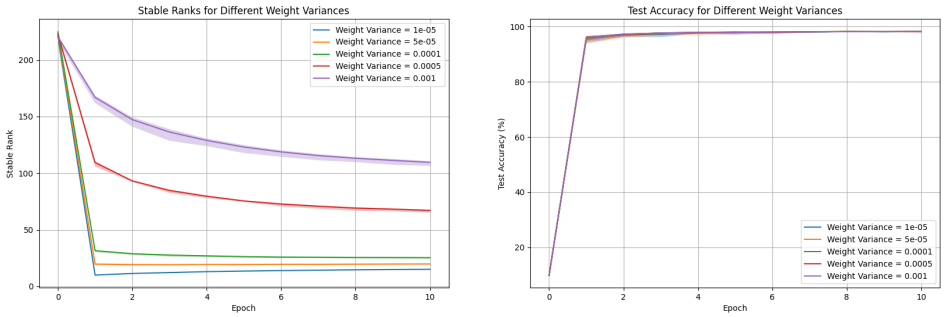

Figure 7: Stable rank of two-layer Leaky-ReLU networks with different weight initialization variances on MNIST.

$\sigma \in \{0.00001, 0.00005, 0.0001, 0.0005, 0.001\}$. We train NN with SGD with batch size 128 and a learning rate of $\alpha = 0.01$ for 100 epochs.

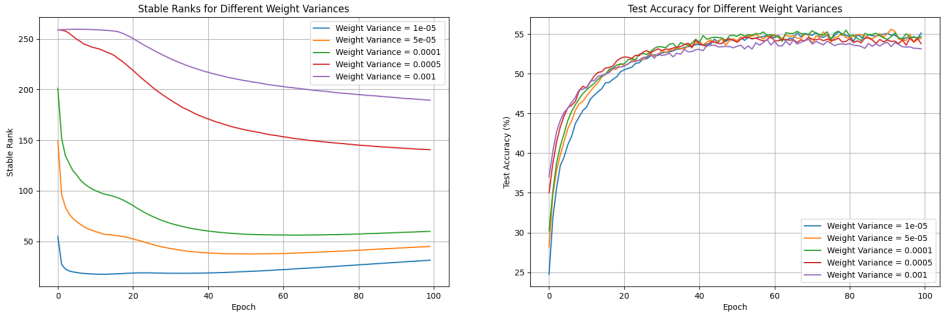

Figure 8: Stable rank of two-layer ReLU networks with different weight initialization variances on CIFAR10.

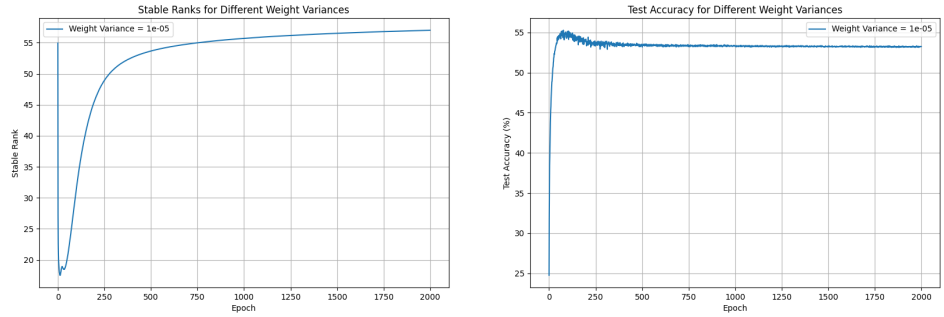

Figure 9: Stable rank of two-layer ReLU networks $\sigma = 0.00001$ on CIFAR10.

However, when we choose to $\sigma = 0.00001$ and setting the epochs to 2000, we observe a sever increase in stable rank after 100 epochs in Figure 9. This is very different from the graph of Leaky-ReLU trained on CIFAR10 by Frei et al. [2023b].

6 Conclusion and Future Work

In this work, we study the implicit bias of gradient flow for two-layer ReLU networks trained on nearly-orthogonal data. Through numerical experiments, we show that gradient descent converges to a network with small stable rank. An important future work is to provide the formal proof of the low-rank bias as we mentioned above. Particularly, we will explore the relationship between the neuron alignment and neuron activation patterns in ReLU networks.

References

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations*, 2020.

Ziwei Ji and Matus Telgarsky. Directional convergence and alignment in deep learning. *Advances in Neural Information Processing Systems*, 33:17176–17186, 2020.

Gal Vardi. On the implicit bias in deep-learning algorithms. *Communications of the ACM*, 66(6): 86–93, 2023.

Spencer Frei, Gal Vardi, Peter Bartlett, and Nathan Srebro. The double-edged sword of implicit bias: Generalization vs. robustness in reLU networks. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023a.

Spencer Frei, Gal Vardi, Peter Bartlett, Nathan Srebro, and Wei Hu. Implicit bias in leaky relu networks trained on high-dimensional data. In *The Eleventh International Conference on Learning Representations*, 2023b.

Gal Vardi, Gilad Yehudai, and Ohad Shamir. Gradient methods provably converge to non-robust networks. *Advances in Neural Information Processing Systems*, 35:20921–20932, 2022.

Hancheng Min, René Vidal, and Enrique Mallada. Early neuron alignment in two-layer relu networks with small initialization. *arXiv preprint arXiv:2307.12851*, 2023.

Spencer Frei, Gal Vardi, Peter Bartlett, and Nathan Srebro. Benign overfitting in linear classifiers and leaky relu networks from kkt conditions for margin maximization. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 3173–3228. PMLR, 2023c.

Yiwen Kou, Zixiang Chen, and Quanquan Gu. Implicit bias of gradient descent for two-layer relu and leaky relu networks on nearly-orthogonal data, 2023.

Mark Rudelson and Roman Vershynin. Sampling from large matrices: An approach through geometric functional analysis. *Journal of the ACM (JACM)*, 54(4):21–es, 2007.

Frank Berthold, Claudia Spix, Peter Kaatsch, and Fred Lampert. Incidence, survival, and treatment of localized and metastatic neuroblastoma in germany 1979-2015. *Paediatric Drugs*, 19(6):577–593, 2017. doi: 10.1007/s40272-017-0251-3.

Simone Gupta, Shannon E Ellis, Foram N Ashar, Anna Moes, Joel S Bader, Jianan Zhan, Andrew B West, and Dan E Arking. Transcriptome analysis reveals dysregulation of innate immune response genes and neuronal activity-dependent genes in autism. *Nature communications*, 5(1):5748, 2014.

Jianqing Fan, Weichen Wang, and Ziwei Zhu. A shrinkage principle for heavy-tailed data: High-dimensional robust low-rank matrix recovery, 2017.