Critical Review: Implicit Bias in Leaky ReLU Networks trained on High-Dimensional Data

Aniruddha Deshpande
Technology Policy Program and EECS
MIT
ani0203@mit.edu

Abstract

This document is a critical review of the paper Implicit Bias in Leaky ReLU Networks trained on High-Dimensional Data Frei et al. [2022]. It is a work that investigates the implicit bias introduced by gradient based methods (gradient flow and gradient descent) in a 2-layer leaky ReLU neural network trained for a binary classification problem nearly orthogonal input training data. I summarize the key results from the paper, proof ideas as well as make some connections with the material studied in class.

1 Introduction

Neural networks trained with gradient-descent based optimization have been found to perform really well on classification tasks with high-dimensional input data such as classifying images, medical diagnosis, audio classification and numerous other applications Rawat and Wang [2017] Shen et al. [2017]. It is well-known that the parameter space of neural networks is highly non-convex and in general has several minima. We studied in class that gradient-based optimization techniques applied to these complex networks introduce some kind of bias in the kind of solutions it leads to, and there have been many efforts to better understand and characterize this bias Galanti and Poggio [2022] Banburski et al. [2019]. This paper studies this implicit bias introduced by gradient descent in a specific context; for a 2-layer fully-connected neural network with Leaky ReLU activation, where only the first layer is trainable and the second layer has fixed weights. The analysis is done under the assumption that input data is high-dimensional, ie. the input dimension d is much higher than the number of training samples n, which implies that different training samples are nearly uncorrelated. It is analyzed for the binary classification problem.

The paper has two distinct sections: in the first it analyses the above network's training dynamics under gradient flow for both exponential $l(z) = e^{-z}$ and logistic loss functions $l(z) = log(1 + e^{-z})$ and a leaky ReLU $\phi(z) = max(\gamma z, z)$ activation function.

In the second part, they study the implicit bias introduced by gradient descent in this network, under certain assumptions on the step-size and weight initialization variance, which is essentially that they are both sufficiently small. The analysis is done for a smoothed version of the leaky ReLU activation function, or more specifically any function that is γ-leaky and H-smooth. Only the logistic loss function is considered here.

The key takeaways for each part are as follows; in the first the authors first leverage a result from Lyu and Li [2019] which shows that when homogenous neural networks trained for binary classification under exponential and logistic losses are trained such that the empirical loss goes below a certain value, then under gradient flow the network converges to a KKT point of the max-margin classification problem in the neural network's parameter space. The authors of this paper extend this to show that the above theorem can be applied in their specific case and this KKT point has additional special

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

properties in the considered case; namely that the weight matrix W^* it converges to has a rank of at-most 2, is the global optimum of the max-margin problem and the trained network fully interpolates the training data, ie. $y_i f(x_i; W^*) = 1 \forall i \in [n]$. They also go on to derive a related linear classifier from the trained network weights, which correctly classifies the data, and is closely related to the max-margin linear classifier.

There are 3 key takeaways in the second part: the first is that under gradient descent with upper bounded step-size and low initialization variance, the empirical loss eventually goes to zero, even though the problem is non-convex. The second is that the weight matrix W of the trained network is low-rank (Specifically, its StableRank is upper bounded, Section 4) after the first step of gradient descent is taken and the l_2 norm of its rows grows to infinity.

The structure of this review is as follows; in Section 2 I describe the general problem setup, definitions and notations that are used throughout the paper. Section 3 contains details about the gradient flow analysis as well as a summary of the proof ideas used by the authors. Section 4 does a similar treatment of the gradient flow analysis. At the end of each of these sections, there is a subsection where I discuss my thoughts on the content of that section how it relates to material we've studied in class, other relevant literature as well as potential directions that could be further explored. Section 5 summarizes the verification experiments performed by the authors, as well as a discussion on the same and some comments on the implications of the results of this paper.

2 General problem setup

In this section I will detail some of the key parts of the problem setup that are common across both the asymptotic gradient flow analysis as well as the non-asymptotic gradient descent analysis.

The authors consider the 2-layer neural network $f(x; W)$ with activation $\phi(.)$, trainable first layer weights W and second layer weights fixed at $a_j = \pm 1/\sqrt{m}$. It is given by :-

$$f(x; W) = \sum_{j=1}^{m} a_j \phi(w_j^T x) = \sum_{j=1}^{m_1} \frac{1}{\sqrt{m}} \phi(v_j^T x) - \sum_{j=1}^{m_2} \frac{1}{\sqrt{m}} \phi(u_j^T x) \tag{1}$$

The authors conveniently re-label the rows of the matrix $W \in \mathbb{R}^{m \times d}$ corresponding to positive second layer weights to be v_j and those with negative second layer weights to be u_j and the number of such rows to m_1 and m_2 and $m = m_1 + m_2$. The activation function used is a leaky ReLU activation $\phi(z) = max(\gamma z, z)$ for the gradient flow analysis. For the gradient descent analysis they consider activation functions that are γ-leaky and H-smooth , which means it satisfies the following 2 properties :-

$$\begin{aligned} 0 \le \gamma \le \phi'(z) \le 1 \\ |\phi''(z)| \le H \end{aligned} \tag{2}$$

an example of this, as mentioned in the paper is the smooth approximation of the ReLU given by $\phi(z) = \gamma z + (1 - \gamma) log(1 + e^z)$.

The binary classification problem is considered in both cases, where the data set is $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^n \subseteq \mathbb{R}^d \times \{\pm 1\}$. The data set is indexed by $I := [n]$, $I_+ := \{i \in I : y_i = 1\}$ and $I_- := \{i \in I : y_i = -1\}$ to distinguish data-points with positive and negative labels.

The analysis for gradient flow hold for both exponential $l(z) = e^{-z}$ and logistic loss functions $l(z) = log(1 + e^{-z})$. The gradient descent analysis is done only assuming a logistic loss function, however there is scope to extend it for the exponential loss function as well. The empirical loss in both cases is given as:

$$\hat{L}(W) = \frac{1}{n} \sum_{i=1}^{n} l(y_i f(x_i; W)) \tag{3}$$

Some other notation that is useful going forward is:-

- $R_{min} = \min_{i \in I} ||x_i||_2$
- $R_{max} = \max_{i \in I} ||x_i||_2$
- $R = \frac{R_{max}}{R_{min}}$
- Sigmoid loss function: $g(z) = -l'(z) = \frac{1}{1+e^z}$
- Empirical sigmoid loss: $\hat{G}(W) = \frac{1}{n}\sum_{i=1}^{n} g(y_i f(x_i; W))$

Now, In the following 2 sections I will detail the findings for both the cases and summarize some of the key proof ideas used.

3 Gradient flow analysis

As a reminder, in this section the neural network weights evolve according to the gradient flow dynamics as given by:

$$\frac{dW(t)}{dt} = -\nabla_W \hat{L}(W(t)) \tag{4}$$

3.1 Key theorems

There are two main theorems in this section, theorems 3.2 and 3.4 and a lemma 3.3. For brevity I will not rewrite the entire theorems here, but explain their key ideas and rewrite only those parts that I deem critical for understanding the their essence. This section also heavily leverages the theorem 4.4 from Lyu and Li [2019] which I will describe first before the theorems from this paper.

3.1.1 Theorem 4.4 Lyu and Li [2019]

The theorem in Lyu and Li [2019] is one that applies generally for homogenous neural networks, ie. networks such that $f(x; \beta\theta) = \beta^L f(x; \theta)$ for some $L > 0$. It says the following: as we are minimizing the exponential or logistic empirical loss over a binary classification dataset under gradient flow, assuming that the empirical loss $\hat{L}(\theta(t))$ becomes smaller than $\frac{log(2)}{n}$ at some point, then gradient flow converges to a KKT point of the following max-margin problem in parameter space:

$$\min_{\theta} \quad \frac{1}{2}||\theta||^2 \qquad \text{s.t.} \quad \forall i \in [n] \quad y_i f(x_i; \theta) \geq 1 \tag{5}$$

This theorem assumes an exponential or logistic loss function.

3.1.2 Theorem 3.5 Frei et al. [2022]

Theorem 3.5 from Frei et al. [2022] just shows that the above theorem from Lyu and Li [2019] can be applied in the setting of this paper and that the network under consideration does converge to a KKT point of the max-margin problem. In order to show this, all that is needed is to show that the main assumption of the theorem above, which is that the empirical loss $\hat{L}(W(t))$ goes below $\frac{log(2)}{n}$ at some point , is satisfied in this setting. In order to prove this the authors first set up a proxy PL inequality Frei and Gu [2021] ,which is essentially a lower-bound on the Frobenius norm of the empirical loss with respect to the weight, ie. $||\nabla_W \hat{L}(W(t))||_F$. This allows them to upper bound the rate of change of the empirical loss under gradient flow using the chain-rule in the following way:

$$\begin{aligned} \frac{d}{dt}\hat{L}(W(t)) &= -||\nabla \hat{L}(W(t))||_F^2 \\ &\leq -\left(\frac{\sqrt{2}R_{min}\gamma}{3R\sqrt{n}} G(\hat{W}(t))\right)^2 \end{aligned} \tag{6}$$

where, $G(\hat{W}(t))$ is the empirical loss with respect to the sigmoid loss function $g(z) = \frac{1}{1+e^z}$. $R_{min} = \min_i ||x_i||$, $R_{max} = \max_i ||x_i||$ and $R = \frac{R_{max}}{R_{min}}$. One more thing to note here is that, while setting up the proxy-PL inequality they use (and prove) the notion that the vector $\hat{\mu} = \sum_{i=1}^{n} y_i x_i$ correctly classifies the dataset and that its inner product with $y_k x_k$ is lower bounded $\forall k \in [n]$.

Using the above differential equation the authors then go on to show that the empirical loss does go below $\frac{log(2)}{n}$ at some time τ, stays below it $\forall t \geq \tau$ and that τ is upper-bounded, ie. that time is reached. This completes the proof.

3.1.3 Theorem 3.2 Frei et al. [2022]

Theorem 3.3 builds on theorem 3.5 in saying that, for the setting under consideration, gradient flow not only converges to a KKT point of the max-margin problem in parameter space, but that this KKT point, which I will call W^* for ease has additional special properties under the assumption that the input datapoints x_i are near-orthogonal , ie. the inner product between any two input data points is small

$$\max_{i \neq j} |\langle x_i, x_j \rangle| \leq \frac{R_{min}^2 \gamma^3}{3R^2 n} \tag{7}$$

The special properties are the following:-

1. The network fully interpolates the training data; $y_i f(x_i; W^*) = 1 \quad \forall i \in [n]$
2. The two kinds of rows in W^* as shown in equation 1 are such that $u_j^* = u \quad \forall j \in [m_1]$ and $v_j^* = v \quad \forall j \in [m_2]$. This implies that the rank of W^* is at most 2. The authors also derive analytical expressions for these rows.
3. W^* is the global optimum of the max-margin problem 5. To show this, the authors also prove that the the rows v, u from item 2 are the global optima of a related constrained convex optimization problem
4. A linear decision boundary made with a linear combination of u AND v , $z = \frac{m_1}{\sqrt{m}} v - \frac{m_2}{\sqrt{m}} u$ classifies the data in the exact same way as $f(x; W^*)$, ie. $sign(z^T x) = sign(f(x;)W^*) \quad \forall x \in \mathbb{R}^d$. The implies that at convergence, the neural network has a linear decision boundary. The authors also show that while z is not the linear max-margin classifier, it is very close to it in a sense.

The proof of this theorem is very long and has many moving parts. I will do my best here to summarize the key ideas and flow of the proof. The proof starts with the assumption that we are currently at a KKT point of the max-margin problem. This assumption is true given theorem 3.5. Using this, they apply the KKT conditions on the problem to get the following analytical expressions for the two kinds of rows v_j and u_j in the matrix W^*:

$$\begin{aligned} v_j &= \frac{1}{\sqrt{m}} \sum_{i \in I} \lambda_i y_i \phi'(v_j^T x_i) x_i \\ u_j &= \frac{1}{\sqrt{m}} \sum_{i \in I} \lambda_i (-y_i) \phi'(v_j^T x_i) x_i \end{aligned} \tag{8}$$

where λ_i are the lagrange multipliers. One thing to note here, is that since the leaky ReLU activation is not smooth, $\phi'(.)$ is the sub-gradient of the activation function at a given point and is denoted by a set, rather than a single value. For a smooth-differentiable function this set has only 1 element which is the derivative. Applying the complementary slackness from the KKT conditions on 5 also gives the following:

$$\lambda_i (1 - y_i f(x_i; W^*)) = 0 \quad \forall i \in [n] \tag{9}$$

They then go on to prove that the λ_is are upper and lower-bounded as $\lambda_i \in \left(\frac{1}{2R_{max}^2}, \frac{3}{2\gamma^2 R_{min}^2}\right)$. This combined with equation 9 proves item 1, that $y_i f(x_i; W^*) = 1 \quad \forall i \in [n]$.

Next, the authors further evaluate the expressions in 10. The key part in doing this is noting that $\phi'(.)$ can take only one of 2 possible values, γ or 1. To figure out which one, an analysis is done to find out the sign of the term inside the activation. By splitting up the data set into positive and negative labels $I = I_+ \cup I_-$ this analysis follows and it is found that item 2 is true and the analytical expressions for the two kinds of rows are given by :-

$$\begin{aligned} v &= \frac{1}{\sqrt{m}} \sum_{i \in I_+} \lambda_i x_i - \frac{\gamma}{\sqrt{m}} \sum_{i \in I_-} \lambda_i x_i \\ u &= \frac{1}{\sqrt{m}} \sum_{i \in I_-} \lambda_i x_i - \frac{\gamma}{\sqrt{m}} \sum_{i \in I_+} \lambda_i x_i \end{aligned} \tag{10}$$

Next, to prove that W^*, the KKT point is also the global optimum of the max-margin problem, the authors first show that the vectors u and v from 10 are the unique global optima of a related linearly constrained convex optimization problem:

$$\begin{aligned} \min_{v,u \in \mathbb{R}^d} \quad & \frac{m_1}{2}||v||^2 + \frac{m_2}{2}||u||^2 \\ \text{s.t.} \quad & \forall i \in I_+ : \quad \frac{m_1}{\sqrt{m}} v^T x_i - \gamma \frac{m_2}{\sqrt{m}} u^T x_i \geq 1 \\ & \forall i \in I_- : \quad \frac{m_2}{\sqrt{m}} u^T x_i - \gamma \frac{m_1}{\sqrt{m}} v^T x_i \geq 1 \end{aligned} \tag{11}$$

This is pretty straightforward to prove as for a linearly constrained strictly convex optimization problem, the KKT conditions are sufficient to prove global optimality.

Now, to show that W^* is a global optimum of the max-margin problem 5 it is sufficient to show that W^* is the only KKT point for this problem, and since the KKT conditions are necessary for global optimalty, it is the only choice for being the global optimum. The uniqueness of W^* comes from combining the findings of items 2 as well as the fact that the rows u and v derived are unique optimizers of the related problem, thereby implying that the weight matrix W^* is a unique KKT point. This completes the proof.

Item 4 is proved pretty much by just writing out the expressions and evaluating the signs of $z^t x$ and $f(x; W^*)$. More details of it can be found in the appendix of the paper.

3.1.4 Lemma 3.3 Frei et al. [2022]

The last part of this section is lemma 3.3. This lemma says that the key assumption of theorem 3.2 , the near-orthogonality of input data-points is a valid one. It just says that when the number of data points n is sufficiently smaller than the input dimension d, and if the input data is assumed to come from a i.i.d. d-dimensional zero mean Gaussian, then the input samples will be nearly orthogonal with very high probability.

3.2 Discussion

I believe that the analysis in this section is intimately linked with some of the material and results we saw in class 10. There, we studied the dynamics of the evolution of weights of a deep fully connected ReLU neural net with L layers for the binary classification problem, under the square-loss. We re-parametrized the network there using the homogeneity of the ReLU network to be:

$$f_W(x) = \rho_L V_L \phi(\rho_{L-1} ... \phi(\rho_1 V_1 x)...) = \rho f_V(x)$$

Note, that this re-parametrization can also be done for the γ-leaky ReLU activation function as the only requirement is homogeneity of the network. Given that the network we considered in class is homogenous, we might be able apply the theorem from Lyu and Li [2019] (section 3)in this case as well to show that gradient flow converges to a KKT point of the max-margin problem 5. This is not straightforward though because of some key differences. First, that the theorem from Lyu and Li [2019] assumed exponential or logistic loss. There may be scope for expanding this finding for the square loss as well. Another slight difference is that the analysis in class also considered weight decay and some thought will have to be put into how to account for that. We did however show in lecture (slide 37, lec 10) that for no weight normalization there are solutions that interpolate all data points, which is similar to item 1 of Theorem 3.2.

Another key difference between the setup of this paper and the material from class is that in this paper the authors show that the problem has only one global minima, which also happens to be the KKT point. The network considered in class is deeper and therefor has several global minima, which would complicate the analysis further.

Another related idea that we briefly saw in class from Banburski et al. [2019] is that both the analysis there as well as in the paper I am reviewing shows that once full separability is achieved, then gradient flow converges in some way to some stationary point. Banburski et al. [2019] in section 4.4 of the paper also shows that the stationary point in that case corresponds to the solution of the max-margin problem in the neural network's parameter space. This is very similar to item 4 in theorem 3.3 of the paper being reviewed. I believe that there is some scope to make formal connections between these two settings as they are saying very similar thing, for slightly different but related setups.

4 Non-asymptotic Gradient Descent analysis

In this section, the paper studies the evolution of the weight matrix W, under gradient descent and the implicit bias induced it. It is given by:-

$$W^{(t+1)} \leftarrow W^t - \alpha \nabla_W \hat{L}(W^{(t)}) \tag{12}$$

The analysis also uses the notion of the stable rank of matrix W as opposed to the regular rank. The stable rank is given by:-

$$StableRank(W) = \frac{||W||_F^2}{||W||_2^2} \tag{13}$$

The reason for this choice is that the stable rank is smoother and more well-behaved than the regular rank and can take non-integer values. It is at most equal to the rank of the matrix and is not as sharply affected by small singular values as the regular rank.

4.1 Key Theorems

4.1.1 Theorem 4.2 Frei and Gu [2021]

There is just one main theorem in this section, but it makes 3 distinct points. It says that for the described Neural network architecture with a γ-leaky, H-smooth activation function trained using gradient descent to minimize the empirical logistic loss on a binary classification data set, under certain assumptions on the step-size α, the weight initialization variance ω_{init} and near-orthogonality (similar to that in the asymptotic case) of the input data, then with high probability ($> 1 - \delta$) the trained network will satisfy the following:

1. The empirical logistic loss goes to 0.

$$\forall t \geq 1 \quad \hat{L}(W^{(t)}) \leq \sqrt{\frac{C_1 n}{R_{min}^2 \alpha t}}$$

2. The l_2-norm of each row of the weight matrix goes to infinity

$$\forall j \in [m] \quad ||w_j^{(t)}||_2 \to \infty$$

3. The stable rank of the weight matrix is upper bounded throughout the training process (after the first step of gradient descent)

$$\sup_{t\geq 1}\{StableRank(W^{(t)})\} \leq C_2$$

Note, the probability comes through the random initialization of the weight matrix. The conditions on the step-size α and the initialization variance ω_{init} are just upper bounds and are explicitly given as :-

$$\alpha \leq \gamma^2(5nR_{max}^2 R^2 C_R max(1,H))^{-1}$$
$$\omega_{init} \leq \alpha\gamma^2 R_{min}(72RC_R n\sqrt{mdlog(\frac{4m}{\delta})})^{-1}$$

The proof for this theorem is also very long and involved. I will do my best here as before to summarize the key ideas and flow of the proof.

To prove item 1, the empirical loss going to 0, the authors set up a proxy-PL inequality Frei and Gu [2021], similar to the one in proving theorem 3.4 in the previous section. This is again, a lower bound on the Frobenius norm of the gradient of the empirical loss with respect to the weight matrix W under the near-orthogonality assumption:

$$||\nabla_W \hat{L}(W^{(t)})|| \geq \frac{\gamma R_{min}}{2\sqrt{2}R\sqrt{n}}\hat{G}(W^{(t)}) \tag{14}$$

The proof for setting up this inequality is very similar to the one used on proving theorem 3.4, and uses the idea that the vector $\hat{\mu} = \sum_{i=1}^n y_i x_i$ correctly classifies the data and is highly correlated with $y_k x_k \quad \forall k \in [n]$ (the norm of their inner products is large).

Once we have the proxy PL inequality, then they use that along with some results on the smoothness of the empirical logistic loss function (lemma E.3) to upper and lower bound the gradient norm as:-

$$\frac{\gamma^2 R_{min}^2}{8R^2 n}\hat{G}(W^{(t)})^2 \leq ||\nabla_W \hat{L}(W^{(t)})|| \leq \frac{2}{\alpha}[\hat{L}(W^{(t+1)}) - \hat{L}(W^{(t)})] \tag{15}$$

They then sum these inequalities over all times up T. Then using some additional analysis they upper bound the empirical sigmoid loss $\hat{G}(W^{(T-1)})$. The final step is then using the relation $\hat{L}(.) \leq 2\hat{G}(.)$ to complete the proof. Another way to look at this proof is using Theorem 3.1 from the paper Frei and Gu [2021] , which says that if a proxy-PL inequality of the kind $||\nabla f(w)||^{\alpha} \geq \frac{1}{2}\mu(g(w) - \xi)$, then we have:-

$$\min_{t<T} g(w^{(t)}) \leq \xi + \epsilon \tag{16}$$

substituting the PL inequality from this paper, we get a similar upper-bound on $\hat{G}$. I haven't had a chance to delve deeper into the exact details, but I believe both these approaches are identical, as the proxy PL framework was also introduced by the same author as this paper, and the proof used here would be similar to the one used in proving the theorem I just used.

To prove item 3, the upper bound on the StableRank, they break the problem into 2 parts; first is to upper bound the Frobenius norm $||W||_F^2$ and then to lower bound the spectral norm $||W||_2^2$.

First, to upper bound the Frobenius norm the authors start with using the triangle inequality:

$$||W^{(t)}||_F \leq ||W^{(0)}||_F + \sum_{s=0}^{t-1} ||W^{(s+1)} - W^{(s)}||_F = ||W^{(0)}||_F + \alpha \sum_{s=0}^{t-1} ||\nabla \hat{L}(W^{(s)})||_F$$

The standard approach of bounding the gradient norm would be to write out the gradient of the network and then use the bounds on $\phi'(.)$ and the datapoint norms, but this bound is too loose (is constant with respect to n). They instead upper-bound the gradient norm using the loss-ratio bound, which is also derived in the appendix of the paper. First note that for the logistic loss $l(z) = log(1 + e^{-z})$ we have $g(z) = -l'(z) = \frac{1}{1+e^z}$, which is the sigmoid loss. The loss-ratio bound says that the sigmoid loss between any two data points is (whp) bounded throughout the training procedure. This essentially shows that the sigmoid loss for all datapoints grows approximately equally fast through the training. This bound is given by :-

$$\sup_{t \geq 0} \left\{ \max_{i,j \in [n]} \frac{l'(y_i f(x_i; W^{(t)}))}{l'(y_j f(x_j; W^{(t)}))} \right\} \leq C_R$$

To use this bound to upper bound the Frobenius norm, they first bound the l-2 norm of the rows of W , $||w_j||_2$ and then combine these to bound the Frobenius norm. To bound $||w_j||_2$ they use some clever manipulation, and then apply bounds using the smoothness of the activation function, near-orthogonality of the input data as well as the loss ratio bound to get :

$$||w_j^{(t)}|| \leq ||w_j^{(0)}|| + \frac{\sqrt{2C_R}|a_j|R_{max}\alpha}{\sqrt{n}} \sum_{s=0}^{t-1} \hat{G}(W^{(s)}) \tag{17}$$

Putting this together for all rows, with $|a_j| = 1/\sqrt{m}$ we get:

$$||W^{(t)}||_F \leq ||W^{(0)}|| + \frac{\sqrt{2C_R}R_{max}\alpha}{\sqrt{n}} \sum_{s=0}^{t-1} \hat{G}(W^{(s)}) \tag{18}$$

The derivation of the loss ratio bound is very involved. I will briefly describe the main idea of how it is derived, but this description is in no way complete. Some ideas that are useful in this proof are:-

1. Near-Orthogonality of gradients: $||\nabla f(f_i; W^{(t)})||^2 \geq C'n \max_{k \neq i} |\langle \nabla f(x_i; W^{(t)}), \nabla f(x_k; W^{(t)}) \rangle|$
2. Gradient Persistence: $||\nabla f(x_i); W^{(t)}|| \geq c||x_i||^2$ This ensures that there is no chance of the gradient vanishing
3. For the given neural network under near-orthogonality constraints on the input training data, near-orthogonality of gradients and gradient persistence holds throughout training.
4. The sigmoid and exponential loss ratios are related as follows:

$$\frac{g(z_1)}{g(z_2)} \leq \max\left(2, 2\frac{e^{-z_1}}{e^{-z_2}}\right),$$

and for $z_1, z_2 > 0$

$$\frac{e^{-z_1}}{e^{-z_2}} \leq 2\frac{g(z_1)}{g(z_2)}$$

The proof then sequentially proves some bounds on how the the exponential loss ratio evolves during during training (Lemma E.8, E.10) and then uses the relation between the exponential and sigmoid loss ratios to prove the loss ratio bound.

To lower bound the spectral norm they start with the definition of the lower-bound and then conveniently choose the $\hat{\mu} = \sum_{i=1}^{n} y_i x_i$ vector to lower bound it.

$$||W^{(t)}||_2^2 = \max_{|x|_2 \neq 0} \frac{||Ax||_2^2}{||x||_2^2} \geq \frac{||W^{(t)}\hat{\mu}||_2^2}{||\hat{\mu}||_2^2}$$

This is then further lower-bounded by showing that every row w_j of the matrix W is highly correlated with $\hat{\mu}$ and also an upper-bound on $||\hat{\mu}||_2^2$. Showing this is also an involved proof, and details for it can be found in Lemma E.13 in section E.5 of the appendix of the paper. This results in the following lower bounds on the l_2 norms of the rows of the matrix and the spectral norm.

$$\begin{aligned} ||w_j^{(t)}||_2 &\geq \frac{\alpha |a_j| R_{min}^2}{4\sqrt{2} R_{max} \sqrt{n}} \sum_{s=0}^{t-1} \hat{G}(W^{(s)}) \\ ||W^{(t)}||_2 &\geq \frac{\alpha \gamma R_{min}^2}{4\sqrt{2} R \sqrt{n}} \sum_{s=0}^{t-1} \hat{G}(W^{(s)}) \end{aligned} \tag{19}$$

Now, to bound the stable rank, all that is needed is to combine the two equations 18 and 19 while taking proper care of the $||W^{(0}||_F$ term in 18. The authors deal with this by considering two cases (1) $||W^{(t)}||_F > 2||W^{(0)}||_F$ and (2) $||W^{(t)}||_F < 2||W^{(0)}||_F$ and the bounding the ratio for both cases in slightly different ways. This completes the proof.

To prove item 2 , it is sufficient to show that $\sum_{s=0}^{t-1} \hat{G}(W^{(s)}) \to \infty$ (see first equation 19). The authors prove this by contradiction, by showing that if this sum is upper-bounded, then the empirical logistic loss $\hat{L}(W^{(t)})$ is lower-bounded, which contradicts item 1 of this theorem and hence $\sum_{s=0}^{t-1} \hat{G}(W^{(s)}) \to \infty$.

4.2 Discussion

The analysis in this section is for vanilla gradient descent. It would be interesting to see how the results in this section are affected if we train the network with SGD instead of regular gradient descent. This would be of interest specifically because most modern networks use batch-wise updates as opposed to using the loss from the entire data set and SGD analysis of this setup would be a valuable addition. I believe the SGD convergence analysis for this can be done using ideas from stochastic approximation of a fixed point with ideas from Robbins [1951].

4.2.1 Proxy PL (Polyak-Lojasiewicz) Inequality Frei and Gu [2021]

Another idea I want to briefly discuss in the context of materials studied in class, is the proxy PL inequality , which the authors use for convergence analysis in both the theorems. I thought this notion was really interesting in that it serves a similar purpose as the strong-convexity assumption (in Lecture 4) did to analyse gradient descent convergence. In order to use this assumption, in lecture we forced it by adding a regularization term. Several other efforts have been also made to understand gradient descent convergence under weaker condition. One such condition is the Polyak-Lojasiewicz inequality. It is stated as:

$$\frac{1}{2}||\nabla f(x)||^2 \geq \mu(f(x) - f^*)$$

The paper Karimi et al. [2016] shows that under the PL-inequality condition on the objective function, several machine learning algorithms have linear convergence using Gradient descent and proximal gradient descent. The PL inequality essentially says the following three things:-

1. The gradient of the function grows faster than a quadratic function as we move away from the optimum value
2. Implies that every stationary point is a global minimum
3. It does not imply the existence of a unique solution. This is in contrast to strong convexity, making this a weaker assumption

However, a downside of using the PL inequality, specifically for analysis of neural networks is item 2 above, which implies that every stationary point is a global minimum. However, given the highly non-convex loss landscape of neural networks, this condition is not true thereby preventing us from applying the PL inequality for its convergence analysis. This is the gap that the proxy-PL inequality tries to bridge; satisfying the proxy-PL inequality does not imply stationary points are global minima, making it more amenable to use for analysis of neural network optimization. The function $f : \mathbb{R}^p \to \mathbb{R}$ is said to satisfy the g-proxy ξ-optimal PL inequality with parameters $\alpha > 0$, $\mu > 0$ if $\exists g : \mathbb{R}^p \to \mathbb{R}$ and scallars ξ, μ such that $\forall w \in \mathbb{R}^p$

$$||\nabla f(w)||^\alpha \geq \frac{1}{2}\mu(g(w) - \xi)$$

The authors in Frei and Gu [2021] claim that this approach provides a unified framework to analyse gradient descent convergence of neural networks for different setups and provide some convergence results under this assumption. It is also used extensively in this paper to prove the convergence of the empirical loss to zero both in gradient descent and gradient flow.

5 Implications and experiments

The authors also briefly discuss the implications of their findings and argue that the bias of the network towards having a linear classfication boundary is good in some cases and not helpful in others. This is a function of the kind of classification problem it tries to solve and may not work well in all cases, specifically where linear boundaries are not sufficient.

The authors also perform 2 experiments to investigate their theoretical findings. In the first they verify their claims by running classification training on a setup that matches the assumptions of their theorem and find that there is a bias towards lower rank weight matrices as suggested by their theory and they find that this bias increases with increase of input data dimension and decrease of initialization variance as is derived.

They also investigate whether these low-rank biases are found in a setting not covered by the theory, specifically that of classification on the CIFAR-10 data set (here $d = 1024$ and $n = 60000$) and find that reducing the initialization variance does bias the network to train towards a weight matrix with lower stable rank.

6 Conclusion

In this report I reviewed the work Frei et al. [2022]. I provided an overview of the key theorems introduced in this work, what they mean as well as a summary of the approach used to prove them. I believe that while this paper does analysis for quite a restrictive case, there is scope to make connections from this work to other ideas that we have studied in class and beyond.

References

Spencer Frei, Gal Vardi, Peter L Bartlett, Nathan Srebro, and Wei Hu. Implicit bias in leaky relu networks trained on high-dimensional data. *arXiv preprint arXiv:2210.07082*, 2022. 1, 3, 4, 5, 10

Waseem Rawat and Zenghui Wang. Deep convolutional neural networks for image classification: A comprehensive review. *Neural computation*, 29(9):2352–2449, 2017. 1

Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. *Annual review of biomedical engineering*, 19:221, 2017. 1

Tomer Galanti and Tomaso Poggio. Sgd noise and implicit low-rank bias in deep neural networks. Technical report, Center for Brains, Minds and Machines (CBMM), 2022. 1

Andrzej Banburski, Qianli Liao, Brando Miranda, Lorenzo Rosasco, Fernanda De La Torre, Jack Hidary, and Tomaso Poggio. Theory iii: Dynamics and generalization in deep networks. *arXiv preprint arXiv:1903.04991*, 2019. 1, 6

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. *arXiv preprint arXiv:1906.05890*, 2019. 1, 3, 6

Spencer Frei and Quanquan Gu. Proxy convexity: A unified framework for the analysis of neural networks trained by gradient descent. *Advances in Neural Information Processing Systems*, 34:7937–7949, 2021. 3, 6, 7, 9, 10

Herbert E. Robbins. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951. 9

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 795–811. Springer, 2016. 10