GRADIENT DESCENT PROVABLY OPTIMIZES OVER-PARAMETERIZED NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

One of the mystery in the success of neural networks is randomly initialized first order methods like gradient descent can achieve zero training loss even though the objective function is non-convex and non-smooth. This paper demystifies this surprising phenomenon for two-layer fully connected ReLU activated neural networks. For an m hidden node shallow neural network with ReLU activation and n training data, we show as long as m is large enough and the data is non-degenerate, randomly initialized gradient descent converges a *globally* optimal solution with a *linear* convergence rate for the quadratic loss function.

Our analysis is based on the following observation: over-parameterization and random initialization jointly restrict every weight vector to be close to its initialization for all iterations, which allows us to exploit a strong convexity-like property to show that gradient descent converges at a global linear rate to the global optimum. We believe these insights are also useful in analyzing deep models and other first order methods.

1 INTRODUCTION

Neural networks trained by first order methods have achieved a remarkable impact on many applications, but their theoretical properties are still mysteries. One of the empirical observation is even though the optimization objective function is non-convex and non-smooth, randomly initialized first order methods like stochastic gradient descent can still find a global minimum. Surprisingly, this property is not correlated with labels. In Zhang et al. (2016), authors replaced the true labels with randomly generated labels, but still found randomly initialized first order methods can always achieve zero training loss.

A widely believed explanation on why a neural network can fit all training labels is because the neural network is over-parameterized. For example, Wide ResNet (Zagoruyko & Komodakis) uses 100x parameters than the number of training data and thus there must exist one such neural network of this architecture that can fit all training data. However, the existence does not imply why the network found by a randomly initialized first order method can fit all the data. The objective function is neither smooth nor convex, which makes traditional analysis technique from convex optimization not useful in this setting. To our knowledge, only the convergence to a stationary point is known (Davis et al., 2018).

In this paper we demystify this surprising phenomenon on two-layer neural networks with rectified linear unit (ReLU) activation. Formally, we consider a neural network of the following form.

$$f(\mathbf{W}, \mathbf{a}, \mathbf{x}) = \frac{1}{\sqrt{m}} \sum_{r=1}^{m} a_r \sigma\left(\mathbf{w}_r^\top \mathbf{x}\right) \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^d$ is the input, $\mathbf{w}_r \in \mathbb{R}^d$ is the weight vector of the first layer, $a_r \in \mathbb{R}$ is the output weight and $\sigma(\cdot)$ is the ReLU activation function: $\sigma(z) = z$ if $z \geq 0$ and $\sigma(z) = 0$ if $z < 0$.

We focus on the empirical risk minimization problem with a quadratic loss. Given a training data set $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, we want to minimize

$$\min_{\mathbf{W} \in \mathbb{R}^{d \times m}} L(\mathbf{W}) = \sum_{i=1}^{n} \frac{1}{2} \left(f(\mathbf{W}, \mathbf{a}, \mathbf{x}_i) - y_i\right)^2 . \tag{2}$$

To do this, we fix the second layer and apply gradient descent (GD) on the first layer weights matrix

$$\mathbf{W}(k+1) = \mathbf{W}(k) - \eta \frac{\partial L(\mathbf{W}(k))}{\partial \mathbf{W}(k)}. \tag{3}$$

where $\eta > 0$ is the step size. Here the gradient formula for each weight vector is [1]

$$\frac{\partial L(\mathbf{W})}{\partial \mathbf{w}_r} = \frac{1}{\sqrt{m}} \sum_{i=1}^{n} (f(\mathbf{W}, \mathbf{a}, \mathbf{x}_i) - y_i) a_r \mathbf{x}_i \mathbb{I}\left\{\mathbf{w}_r^\top \mathbf{x}_i \geq 0\right\} \tag{4}$$

Though this is only a shallow fully connected neural network, the objective function is still non-smooth and non-convex due to the use of ReLU activation function. Even for this simple function, why randomly initialized first order method can achieve zero training error is not known. In fact, many previous work have tried to answer this question or similar ones. Attempts include landscape analysis (Soudry & Carmon, 2016), partial differential equations (Mei et al.), analysis of the dynamics of the algorithm (Li & Yuan, 2017), optimal transport theory (Chizat & Bach, 2018), to name a few. These results often rely strong assumptions on the labels and input distributions, or do not imply why randomly initialized first order method can achieve zero training loss. See Section 2 for detailed comparisons between our result and previous ones.

In this paper, we rigorously prove that as long as the data set is not degenerate and m is large enough, with properly randomly initialized $\mathbf{a}$ and $\mathbf{W}(0)$, GD achieves zero training loss at a linear convergence rate, i.e., it finds a solution $\mathbf{W}(K)$ with $L(\mathbf{W}(K)) \leq \epsilon$ in $K = O(\log(1/\epsilon))$ iterations.[2] Thus, our theoretical result not only shows the global convergence but also gives a quantitative convergence rate in terms of the desired accuracy.

Analysis Technique Overview Our proof relies on the following insights. First we directly analyze the dynamics of each individual prediction $f(\mathbf{W}, \mathbf{a}, \mathbf{x}_i)$ for $i = 1, \ldots, n$. This is different from many previous work (Du et al., 2017b; Li & Yuan, 2017) which tried to analyze the dynamics of the parameter (W) we are optimizing. Note because the objective function is non-smooth and non-convex, analysis of the parameter space dynamics is very difficult. In contrast, we find the dynamics of prediction space is governed by the spectral property of a Gram matrix (which can vary in each iteration, c.f. equation 5) and as long as this Gram matrix's least eigenvalue is lower bounded, gradient descent enjoys a linear rate. Furthermore, previous work has shown in the initialization phase this Gram matrix does has lower bounded least eigenvalue as long as the data is not degenerate (Xie et al., 2017). Thus the problem reduces to showing the Gram matrix at later iterations is close to that in the initialization phase. Our second observation is this Gram matrix is only related to the activation patterns ($\mathbb{I}\left\{\mathbf{w}_r^\top \mathbf{x}_i \geq 0\right\}$) and we can use matrix perturbation analysis to show if most of the patterns do not change, then this Gram matrix is close to its initialization. Our third observation is we find over-parameterization, random initialization and the linear convergence jointly restrict every weight vector $\mathbf{w}_r$ is close to its initialization. Then we can use this property to show most of the patterns do not change. Combining these insights we prove the first global quantitative convergence result of gradient descent on ReLU activated neural networks for the empirical risk minimization problem. Notably, our proof only uses linear algebra and standard probability bounds so we believe it can be easily generalized to analyze deep neural networks.

Notations We use bold-faced letters for vectors and matrices. We Let $[n] = \{1, 2, \ldots, n\}$. Given a set S, we use $\mathrm{unif}\{S\}$ to denote the uniform distribution over S. Given an event E, we use $\mathbb{I}\{A\}$ to be the indicator on whether this event happens. We use $N(\mathbf{0}, \mathbf{I})$ to denote the standard Gaussian distribution. For a matrix $\mathbf{A}$, we use $\mathbf{A}_{ij}$ to denote its (i, j)-th entry. We use $\|\cdot\|_2$ to denote the Euclidean norm of a vector, and use $\|\cdot\|_F$ to denote the Frobenius norm of a matrix. If a matrix $\mathbf{A}$ is positive semi-definite, we use $\lambda_{\min}(\mathbf{A})$ to denote its smallest eigenvalue. We use $\langle \cdot, \cdot \rangle$ to denote the standard Euclidean inner product between two vectors. Lastly, let $O(\cdot)$ and $\Omega(\cdot)$ denote standard Big-O and Big-Omega notations, only hiding absolute constants.

[1] Note ReLU is not continuously differentiable. One can view $\frac{\partial L(\mathbf{W})}{\partial \mathbf{w}_r}$ as a convenient notation for the right hand side of equation 4 and this is indeed the update rule used in practice.

[2] Here we omit the polynomial dependency on n and other data dependent quantities.

2 COMPARISON WITH PREVIOUS RESULTS

In this section we survey an incomplete list of previous attempts in analyzing why first order methods can find a global minimum.

Landscape Analysis A popular way to analyze non-convex optimization problems is to identify whether the optimization landscape has some benign geometric properties. Recently, researchers found if the objective function is smooth and satisfies (1) all local minima are global and (2) for every saddle point, there exists a negative curvature, then the noise-injected (stochastic) gradient descent (Jin et al., 2017; Ge et al., 2015; Du et al., 2017a) can find a global minimum in polynomial time. This algorithmic finding encouraged researchers to study whether the deep neural networks also admit these properties.

For the objective function define in equation 2, some partial results were obtained. Soudry & Carmon (2016) showed if $md \geq n$, then at every differentiable local minimum, the training error is zero. However, since the objective is non-smooth, it is hard to show gradient descent actually convergences to a differentiable local minimum. Xie et al. (2017) studied the same problem and related the loss to the gradient norm through the least singular value of the "extended feature matrix" $\mathbf{D}$ at the stationary points. However, they did not prove the convergence rate of gradient norm. Interestingly, our analysis relies on the Gram matrix which is actually $\mathbf{D}\mathbf{D}^\top$.

Landscape analyses of ReLU activated neural networks for other settings have also been studied in many previous work (Ge et al., 2017; Safran & Shamir, 2018; 2016; Zhou & Liang, 2017; Freeman & Bruna, 2016; Hardt & Ma, 2016). These works establish favorable landscape properties such as all local minimizers are global, but do not ensure that gradient descent converges to a global minimizer of the empirical risk. For other activation functions, some previous work show the landscape does have the desired geometric properties (Du & Lee, 2018; Soltanolkotabi et al., 2018; Nguyen & Hein, 2017; Kawaguchi, 2016; Haeffele & Vidal, 2015; Andoni et al., 2014; Venturi et al., 2018). However, it is unclear how to extend their analyses in our setting.

Analysis of Algorithm Dynamics Another way to prove convergence result is to directly analyze the dynamics of first order methods. Our paper also belongs to this category. Many previous work assumed (1) the input distribution is Gaussian and (2) the label is generated according to a planted neural network. Based on these two (unrealistic) conditions, it can be shown that randomly initialized (stochastic) gradient descent can learn a ReLU (Tian, 2017; Soltanolkotabi, 2017), a single convolutional filter (Brutzkus & Globerson, 2017), a convolutional neural network with one filter and one output layer (Du et al., 2018b) and residual network with small spectral norm weight matrix (Li & Yuan, 2017).[3] Beyond Gaussian input distribution, Du et al. (2017b) showed for learning a convolutional filter, the Gaussian input distribution assumption can be relaxed but they still require the label is generated from an underlying true filter. Comparing with these work, our paper does not try to recover the underlying true neural network. Instead, we focus on providing theoretical justification on why randomly initialized gradient descent can achieve zero training loss, which is what we can observe and verify in practice.

The most related paper is by Li & Liang (2018) who observed that when training a two-layer full connected neural network, most of the patterns ($\mathbb{I}\left\{\mathbf{w}_r^\top \mathbf{x}_i \geq 0\right\}$) do not change over iterations, which we also use to show the stability of the Gram matrix. They used this observation to obtain the convergence rate of GD on a two-layer over-parameterized neural network for the cross-entropy loss. They need number of hidden nodes m scales with $\mathrm{poly}(1/\epsilon)$ where ϵ is the desired accuracy. Thus unless the number of hidden nodes $m \to \infty$, they result does not imply GD can achieve zero training loss. We improve by allowing the amount of over-parameterization to be independent of the desired accuracy and show GD can achieve zero training loss. Furthermore, our proof is much simpler and more transparent so we believe it can be easily generalized to analyze other neural network architectures.

Other Analysis Approaches Chizat & Bach (2018) used optimal transport theory to analyze continuous time gradient descent on over-parameterized models. However, their results on ReLU acti-

[3] Since these work assume the label is realizable, converging to global minimum is equivalent to recover the underlying model.

vated neural network is only at the formal level. Mei et al. showed the dynamics of SGD can be captured by a partial differential equation in the suitable scaling limit. They listed some specific examples on input distributions including mixture of Gaussians. However, it is still unclear whether this framework can explain why first order methods can minimize the empirical risk. Daniely (2017) built connection between neural networks with kernel methods and showed stochastic gradient descent can learn a function that is competitive with the best function in the conjugate kernel space of the network. Again this work does not imply why first order method can achieve zero training loss.

3 Continuous Time Analysis

In this section, we present our result for gradient flow, i.e., gradient descent with infinitesimal step size. In the next section, we will modify the proof and give a quantitative bound for gradient descent with positive step size. Formally, we consider the ordinary differential equation[4] defined by :

$$\frac{d\mathbf{w}_r(t)}{dt} = -\frac{\partial L(\mathbf{W}(t))}{\partial \mathbf{w}_r(t)}$$

for $r \in [m]$. We denote $u_i(t) = f(\mathbf{W}(t), \mathbf{a}, \mathbf{x}_i)$ the prediction on input $\mathbf{x}_i$ at time t and we let $\mathbf{u}(t) = (u_1(t), \ldots, u_n(t)) \in \mathbb{R}^n$ be the prediction vector at time t. Our main result in this section is the following theorem.

Theorem 3.1 (Convergence Rate of Gradient Flow)**.** *Assume for all $i \in [n]$, $\|\mathbf{x}_i\|_2 = 1$ and $|y_i| \leq C$ for some constant C and the matrix $\mathbf{H}^\infty \in \mathbb{R}^{n\times n}$ with $\mathbf{H}^\infty_{ij} = \mathbb{E}_{\mathbf{w}\sim N(\mathbf{0},\mathbf{I})}\left[\mathbf{x}_i^\top \mathbf{x}_j \mathbb{I}\left\{\mathbf{w}^\top \mathbf{x}_i \geq 0, \mathbf{w}^\top \mathbf{x}_j \geq 0\right\}\right]$ satisfies $\lambda_{\min}(\mathbf{H}^\infty) \triangleq \lambda_0 > 0$. Then if we initialize $\mathbf{w}_r \sim N(\mathbf{0}, \mathbf{I})$, $a_r \sim \mathrm{unif}\left[\{-1, 1\}\right]$ for $r \in [m]$ and set the number of hidden nodes $m = \Omega\left(\frac{n^6}{\lambda_0^4}\right)$, with high probability over random initialization we have*

$$\|\mathbf{u}(t) - \mathbf{y}\|_2^2 \leq \exp(-\lambda_0 t)\, \|\mathbf{u}(0) - \mathbf{y}\|_2^2 .$$

We first discuss our assumptions. We assume $\|\mathbf{x}_i\|_2 = 1$ and $|y_i| \leq C$ for simplicity. We can easily modify the proof by properly scaling the initialization. The key assumption is the least eigenvalue of the matrix $\mathbf{H}^\infty$ is strictly positive. Interestingly, various properties of this $\mathbf{H}^\infty$ matrix has been thoroughly studied in previous work (Xie et al., 2017; Tsuchida et al., 2017). In general, unless the data is degenerate, the smallest eigenvalue of $\mathbf{H}^\infty$ is strictly positive. We refer readers to Xie et al. (2017); Tsuchida et al. (2017) for a detailed characterization of this matrix.

The number of hidden nodes m required is $\Omega\left(\frac{n^6}{\lambda_0^4}\right)$, which depends on the number of samples n and λ_0. As will be apparent in the proof, over-parameterization, i.e., the fact $m = \mathrm{poly}(n, 1/\lambda_0)$, plays a crucial role in guaranteeing gradient descent to find the global minimum. We believe using a more refined analysis, this dependency can be further improved. Lastly, note the convergence rate is linear because $\|\mathbf{u}(t) - \mathbf{y}\|_2^2$ decreases to 0 exponentially fast. The specific rate also depends on λ_0 but independent of number of hidden nodes m.

3.1 Proof of Theorem 3.1

Our first step is to calculate the dynamics of each prediction.

$$\begin{aligned}\frac{d}{dt}u_i(t) &= \sum_{r=1}^{m} \langle \frac{\partial f(\mathbf{W}(t), \mathbf{a}, \mathbf{x}_i)}{\partial \mathbf{w}_r(t)}, \frac{d\mathbf{w}_r(t)}{dt} \rangle \\ &= \sum_{j=1}^{n} (y_j - u_j) \langle \sum_{r=1}^{m} \frac{\partial f(\mathbf{W}(t), \mathbf{a}, \mathbf{x}_i)}{\partial \mathbf{w}_r(t)}, \frac{\partial f(\mathbf{W}(t), \mathbf{a}, \mathbf{x}_j)}{\partial \mathbf{w}_r(t)} \rangle \triangleq \sum_{j=1}^{n} (y_j - u_j) \mathbf{H}_{ij}(t)\end{aligned}$$

where $\mathbf{H}(t)$ is an $n \times n$ matrix with (i, j)-th entry

$$\mathbf{H}_{ij}(t) = \frac{1}{m} \mathbf{x}_i^\top \mathbf{x}_j \sum_{r=1}^{m} \mathbb{I}\left\{\mathbf{x}_i^\top \mathbf{w}_r(t) \geq 0, \mathbf{x}_j^\top \mathbf{w}_r(t) \geq 0\right\}. \tag{5}$$

[4] Strictly speaking, this should be differential inclusion (Davis et al., 2018)

With this $\mathbf{H}$ matrix, we can write the dynamics of prediction in a compact way:

$$\frac{d}{dt}\mathbf{u}(t) = \mathbf{H}(t)(\mathbf{y} - \mathbf{u}(t)).$$

Note $\mathbf{H}(t)$ is a time-dependent symmetric matrix. We first analyze its property when $t = 0$. The following lemma shows if m is large then $\mathbf{H}(0)$ has lower bounded least eigenvalue with high probability. The proof is by standard concentration bound so we defer it to appendix.

Lemma 3.1. *If $m = \Omega\left(\frac{n^2}{\lambda_0^2}\log^2\left(\frac{n}{\delta}\right)\right)$, we have with probability at least $1-\delta$, $\lambda_{\min}(\mathbf{H}(0)) \geq \frac{3}{4}\lambda_0$.*

Our second step is to show $\mathbf{H}(t)$ is stable in terms of $\mathbf{W}$. Formally, the following lemma shows if $\mathbf{W}(t)$ is close to the initialization $\mathbf{W}(0)$, $\mathbf{H}(t)$ is close $\mathbf{H}(0)$ and has a lower bounded least eigenvalue.

Lemma 3.2. *Suppose for a given time t, for all $r \in [m]$, $\|\mathbf{w}_r(t) - \mathbf{w}_r(0)\|_2 \leq \frac{c\lambda_0}{n^2} \triangleq R$ for some small positive constant c. Then we have with high probability over initialization, $\lambda_{\min}(\mathbf{H}(t)) > \frac{\lambda_0}{2}$.*

This lemma plays a crucial in our analysis so we give the proof below.

Proof of Lemma 3.2 We define the event

$$A_{ir} = \left\{\exists \mathbf{w} : \|\mathbf{w} - \mathbf{w}_r(0)\| \leq R, \mathbb{I}\left\{\mathbf{x}_i^\top \mathbf{w}_r(0) \geq 0\right\} \neq \mathbb{I}\left\{\mathbf{x}_i^\top \mathbf{w} \geq 0\right\}\right\}.$$

Note this event happens if and only if $\left|\mathbf{w}_r(0)^\top \mathbf{x}_0\right| < R$. Recall $\mathbf{w}_r(0) \sim N(\mathbf{0}, \mathbf{I})$. By anti-concentration inequality of Gaussian, we have $P(A_{ir}) = P_{z\sim N(0,1)}\left(|z| < R\right) \leq \frac{2R}{\sqrt{2\pi}}$. Therefore, we can bound the entry-wise deviation on $\mathbf{H}(t)$ in expectation:

$$\begin{aligned}
&\mathbb{E}\left[|H_{ij}(t) - H_{ij}(0)|\right] \\
=&\mathbb{E}\left[\frac{1}{m}\left|\mathbf{x}_i^\top \mathbf{x}_j \sum_{r=1}^{m}\left(\mathbb{I}\left\{\mathbf{w}_r(0)^\top \mathbf{x}_i \geq 0, \mathbf{w}_r(0)^\top \mathbf{x}_j \geq 0\right\} - \mathbb{I}\left\{\mathbf{w}_r(t)^\top \mathbf{x}_i \geq 0, \mathbf{w}_r(t)^\top \mathbf{x}_j \geq 0\right\}\right)\right|\right] \\
\leq&\frac{1}{m}\sum_{r=1}^{m}\mathbb{E}\left[\mathbb{I}\left\{A_{ir} \cup A_{jr}\right\}\right] \leq \frac{4R}{\sqrt{2\pi}}.
\end{aligned}$$

Summing over (i,j), we have $\mathbb{E}\left[\sum_{(i,j)=(1,1)}^{(n,n)}|H_{ij}(t) - H_{ij}(0)|\right] \leq \frac{4n^2R}{\sqrt{2\pi}}$. Thus by Markov inequality, with high probability, we have $\sum_{(i,j)=(1,1)}^{(n,n)}|H_{ij}(t) - H_{ij}(0)| \leq Cn^2R$ where C is a large absolute constant. Next, we use matrix perturbation theory to bound the deviation from the initialization

$$\|\mathbf{H}(t) - \mathbf{H}(0)\|_2 \leq \|\mathbf{H}(t) - \mathbf{H}(0)\|_F \leq \sum_{(i,j)=(1,1)}^{(n,n)}|H_{ij}(t) - H_{ij}(0)| \leq Cn^2R.$$

Lastly, we lower bound the smallest eigenvalue by plugging in R

$$\lambda_{\min}(\mathbf{H}(t)) \geq \lambda_{\min}(\mathbf{H}(0)) - Cn^2R \geq \frac{\lambda_0}{2}. \quad \square$$

The next lemma shows two facts if the least eigenvalue of H is lower bounded. First, the loss converges to 0 at a linear convergence rate. Second, $\mathbf{w}_r(t)$ is close to the initialization for every $r \in [m]$. This lemma clearly demonstrates the power of over-parameterization.

Lemma 3.3. *Suppose for $0 \leq s \leq t$, $\lambda_{\min}\left(\mathbf{H}(s)\right) \geq \frac{\lambda_0}{2}$. Then we have $\|\mathbf{y} - \mathbf{u}(t)\|_2^2 \leq \exp(-\lambda_0 t)\|\mathbf{y} - \mathbf{u}(0)\|_2^2$ and for any $r \in [m]$, $\|\mathbf{w}_r(t) - \mathbf{w}_r(0)\|_2 \leq \frac{2\sqrt{n}\|\mathbf{y}-\mathbf{u}(0)\|_2}{\sqrt{m}\lambda_0} \triangleq R'$.*

Proof of Lemma 3.3 Recall we can write the dynamics of predictions as $\frac{d}{dt}\mathbf{u}(t) = \mathbf{H}(\mathbf{y} - \mathbf{u}(t))$. We can calculate the loss function dynamics

$$\begin{aligned}
\frac{d}{dt}\|\mathbf{y} - \mathbf{u}(t)\|_2^2 =& -2\left(\mathbf{y} - \mathbf{u}(t)\right)^\top \mathbf{H}(t)\left(\mathbf{y} - \mathbf{u}(t)\right) \\
\leq& -\lambda_0\|\mathbf{y} - \mathbf{u}(t)\|_2^2.
\end{aligned}$$

Thus we have $\frac{d}{dt}\left(\exp(\lambda_0 t)\left\|\mathbf{y}-\mathbf{u}(t)\right\|_2^2\right) \leq 0$, $\exp(\lambda_0 t)\left\|\mathbf{y}-\mathbf{u}(t)\right\|_2^2$ is a decreasing function with respect to t. Using this fact we can bound the loss

$$\left\|\mathbf{y}-\mathbf{u}(t)\right\|_2^2 \leq \exp(-\lambda_0 t)\left\|\mathbf{y}-\mathbf{u}(0)\right\|_2^2.$$

Therefore, $\mathbf{u}(t) \to \mathbf{y}$ exponentially fast. Now we bound the gradient. Recall for $0 \leq s \leq t$,

$$\begin{aligned}\left\|\frac{d}{dt}\mathbf{w}_r(s)\right\|_2 &= \left\|\sum_{i=1}^{n}(y_i-u_i)\frac{1}{\sqrt{m}}a_r\mathbf{x}_i\mathbb{I}\left\{\mathbf{w}_r(t)^\top\mathbf{x}_i \geq 0\right\}\right\|_2 \\ &\leq \frac{1}{\sqrt{m}}\sum_{i=1}^{n}|y_i-u_i(s)| \leq \frac{\sqrt{n}}{\sqrt{m}}\left\|\mathbf{y}-\mathbf{u}(s)\right\|_2 \leq \frac{\sqrt{n}}{\sqrt{m}}\exp(-\lambda_0 s/2)\left\|\mathbf{y}-\mathbf{u}(0)\right\|_2.\end{aligned}$$

Integrating the gradient, we can bound the distance from the initialization

$$\left\|\mathbf{w}_r(t)-\mathbf{w}_r(0)\right\|_2 \leq \int_0^t \left\|\frac{d}{ds}\mathbf{w}_r(s)\right\|_2 ds \leq \frac{2\sqrt{n}\left\|\mathbf{y}-\mathbf{u}(0)\right\|_2}{\sqrt{m}\lambda_0}. \quad \square$$

The next lemma shows we show if $R' < R$, the conditions in Lemma 3.2 and 3.3 hold for all $t \geq 0$. The proof is by contradiction and we defer it to appendix.

Lemma 3.4. *If $R' < R$, we have for all $t \geq 0$, $\lambda_{\min}(\mathbf{H}(t)) \geq \frac{1}{2}\lambda_0$, for all $r \in [m]$, $\left\|\mathbf{w}_r(t)-\mathbf{w}_r(0)\right\|_2 \leq R'$ and $\left\|\mathbf{y}-\mathbf{u}(t)\right\|_2^2 \leq \exp(-\lambda_0 t)\left\|\mathbf{y}-\mathbf{u}(0)\right\|_2^2$.*

Thus it is sufficient to show $R' < R$ which is equivalent to $m = \Omega\left(\frac{n^5\|\mathbf{y}-\mathbf{u}(0)\|_2^2}{\lambda_0^4}\right)$. We bound

$$\mathbb{E}\left[\left\|\mathbf{y}-\mathbf{u}(0)\right\|_2^2\right] = \sum_{i=1}^{n}(y_i^2 + y_i\mathbb{E}\left[f(\mathbf{W}(0),\mathbf{a},\mathbf{x}_i)\right] + \mathbb{E}\left[f(\mathbf{W}(0),\mathbf{a},\mathbf{x}_i)^2\right]) = \sum_{i=1}^{n}(y_i^2+1) = O(n).$$

Thus by Markov's inequality, we have with high probability $\left\|\mathbf{y}-\mathbf{u}(0)\right\|_2^2 = O(n)$. Plugging in this bound we prove the theorem. $\square$

4 Discrete Time Analysis

In this section, we show randomly initialized gradient descent with a constant positive step size converges to the global minimum at a linear rate. We first present our main theorem.

Theorem 4.1 (Convergence Rate of Gradient Descent)**.** *Assume for all $i \in [n]$, $\left\|\mathbf{x}_i\right\|_2 = 1$, $|y_i| \leq C$ for some constant C and the matrix $\mathbf{H}^\infty \in \mathbb{R}^{n\times n}$ with $\mathbf{H}_{ij}^\infty = \mathbb{E}_{\mathbf{w}\sim N(\mathbf{0},\mathbf{I})}\left[\mathbf{x}_i^\top\mathbf{x}_j\mathbb{I}\left\{\mathbf{w}^\top\mathbf{x}_i \geq 0, \mathbf{w}^\top\mathbf{x}_j \geq 0\right\}\right]$ satisfies $\lambda_{\min}(\mathbf{H}^\infty) \triangleq \lambda_0 > 0$. If we initialize $\mathbf{w}_r \sim N(\mathbf{0},\mathbf{I})$, $a_r \sim \mathrm{unif}\left[\{-1,1\}\right]$ for $r \in [m]$ and the number of hidden nodes $m = \Omega\left(\frac{n^6}{\lambda_0^4}\right)$ and we set the step size $\eta = O\left(\frac{\lambda_0}{n^2}\right)$ then with high probability over the random initialization we have for $k = 0, 1, 2, \ldots$*

$$\left\|\mathbf{u}(k)-\mathbf{y}\right\|_2^2 \leq \left(1-\frac{\eta\lambda_0}{2}\right)^k\left\|\mathbf{u}(0)-\mathbf{y}\right\|_2^2.$$

Theorem 4.1 shows even though the objective function is non-smooth and non-convex, gradient descent with a constant positive step size still enjoys a linear convergence rate. Our assumptions on the least eigenvalue and the number of hidden nodes are exactly the same as the theorem for gradient flow. Notably, our choice of step size is independent of number of hidden nodes m in contrast to the previous work (Li & Liang, 2018).

4.1 Proof of Theorem 4.1

We prove Theorem 4.1 by induction. Our induction hypothesis is just the following convergence rate of empirical loss.

Condition 4.1. *At the k-th iteration, we have* $\|\mathbf{y}-\mathbf{u}(k)\|_2^2 \leq (1-\frac{\eta\lambda_0}{2})^k \|\mathbf{y}-\mathbf{u}(k)\|_2^2$.

A directly corollary of this condition is the following bound of deviation from the initialization. The proof is similar to that of Lemma 3.3 so we defer it to appendix.

Corollary 4.1. *If Condition 4.1 holds for $k'=0,\ldots,k$, then we have for every $r\in[m]$*

$$\|\mathbf{w}_r(k+1)-\mathbf{w}_r(0)\|_2 \leq \frac{4\sqrt{n}\,\|\mathbf{y}-\mathbf{u}(0)\|_2}{\sqrt{m}\lambda_0} \triangleq R'. \tag{6}$$

Now we show Condition 4.1 for every $k=0,1,\ldots$. For the base case $k=0$, by definition Condition 4.1 holds. Suppose for $k'=0,\ldots,k$, Condition 4.1 holds and we want to show Condition 4.1 holds for $k'=k+1$.

Our strategy is similar to the proof of Theorem 3.1. We define the event

$$A_{ir}=\left\{\exists\mathbf{w}:\|\mathbf{w}-\mathbf{w}_r(0)\|\leq R, \mathbb{I}\left\{\mathbf{x}_i^\top\mathbf{w}_r(0)\geq 0\right\}\neq\mathbb{I}\left\{\mathbf{x}_i^\top\mathbf{w}\geq 0\right\}\right\}.$$

where $R=\frac{c\lambda_0}{n^2}$ for some small positive constant c. Different from gradient flow, for gradient descent we need a more refined analysis. We let $S_i=\{r\in[m]:\mathbb{I}\{A_{ir}\}=1\}$ and $S_i^\perp=[m]\setminus S_i$. The following Lemma bounds the sum of sizes of $S_i^\perp$. The proof is similar to the analysis used in Lemma 3.2. See Section A for the whole proof.

Lemma 4.1. *With high probability over initialization, we have $\sum_{i=1}^n\left|S_i^\perp\right|\leq CmnR$ for some positive constant $C>0$.*

Next, we calculate the difference of predictions between two consecutive iterations, analogue to $\frac{du_i(t)}{dt}$ term in Section 3.

$$\begin{aligned}u_i(k+1)-u(k)=&\frac{1}{\sqrt{m}}\sum_{r=1}^m a_r\left(\sigma\left(\mathbf{w}_r(k+1)^\top\mathbf{x}_i\right)-\sigma\left(\mathbf{w}_r(k)^\top\mathbf{x}_i\right)\right)\\=&\frac{1}{\sqrt{m}}\sum_{r=1}^m a_r\left(\sigma\left(\left(\mathbf{w}_r(k)-\eta\frac{\partial L(\mathbf{W}(k))}{\partial\mathbf{w}_r(k)}\right)^\top\mathbf{x}_i\right)-\sigma\left(\mathbf{w}_r(k)^\top\mathbf{x}_i\right)\right).\end{aligned}$$

Here we divide the right hand side into two parts. I_1^i accounts for terms that the pattern does not change and I_2^i accounts for terms that pattern may change.

$$\begin{aligned}I_1^i\triangleq&\frac{1}{\sqrt{m}}\sum_{r\in S_i}a_r\left(\sigma\left(\left(\mathbf{w}_r(k)-\eta\frac{\partial L(\mathbf{W}(k))}{\partial\mathbf{w}_r(k)}\right)^\top\mathbf{x}_i\right)-\sigma\left(\mathbf{w}_r(k)^\top\mathbf{x}_i\right)\right)\\I_2^i\triangleq&\frac{1}{\sqrt{m}}\sum_{r\in S_i^\perp}a_r\left(\sigma\left(\left(\mathbf{w}_r(k)-\eta\frac{\partial L(\mathbf{W}(k))}{\partial\mathbf{w}_r(k)}\right)^\top\mathbf{x}_i\right)-\sigma\left(\mathbf{w}_r(k)^\top\mathbf{x}_i\right)\right)\end{aligned}$$

We view I_2^i as a perturbation and bound its magnitude. Because ReLU is a 1-Lipschitz function and $|a_r|=1$, we have

$$\left|I_2^i\right|\leq\frac{\eta}{\sqrt{m}}\sum_{r\in S_i^\perp}\left|\left(\frac{\partial L(\mathbf{W}(k))}{\partial\mathbf{w}_r(k)}\right)^\top\mathbf{x}_i\right|\leq\frac{\eta\left|S_i^\perp\right|}{\sqrt{m}}\max_{r\in[m]}\left\|\frac{\partial L(\mathbf{W}(k))}{\partial\mathbf{w}_r(k)}\right\|_2\leq\frac{\eta\left|S_i\right|^\perp\sqrt{n}\,\|\mathbf{u}(k)-\mathbf{y}\|_2}{m}.$$

To analyze I_1^i, by Corollary 4.1, we know $\|\mathbf{w}_r(k)-\mathbf{w}_r(0)\|\leq R'$ and $\|\mathbf{w}_r(k)-\mathbf{w}_r(0)\|\leq R'$ for all $r\in[m]$. Furthermore, because $R'<R$, we know $\mathbb{I}\left\{\mathbf{w}_r(k+1)^\top\mathbf{x}_i\geq 0\right\}=\mathbb{I}\left\{\mathbf{w}_r(k)^\top\mathbf{x}_i\geq 0\right\}$. Thus we can find a more convenient expression of I_1^i for analysis

$$\begin{aligned}I_1^i=&-\frac{\eta}{m}\sum_{j=1}^n\mathbf{x}_i^\top\mathbf{x}_j\left(u_j-y_j\right)\sum_{r\in S_i}\mathbb{I}\left\{\mathbf{w}_r(k)^\top\mathbf{x}_i\geq 0,\mathbf{w}_r(k)^\top\mathbf{x}_j\geq 0\right\}\\=&-\eta\sum_{j=1}^n(u_j-y_j)(\mathbf{H}_{ij}(k)-\mathbf{H}_{ij}^\perp(k))\end{aligned}$$

where $\mathbf{H}_{ij}(k) = \frac{1}{m}\sum_{r=1}^{m}\mathbf{x}_i^\top\mathbf{x}_j\mathbb{I}\left\{\mathbf{w}_r(k)^\top\mathbf{x}_i \geq 0, \mathbf{w}_r(k)^\top\mathbf{x}_j \geq 0\right\}$ is just the (i,j)-th entry of a discrete version of Gram matrix defined in Section 3 and $\mathbf{H}_{ij}^{\perp}(k) = \frac{1}{m}\sum_{r\in S_i^\perp}\mathbf{x}_i^\top\mathbf{x}_j\mathbb{I}\left\{\mathbf{w}_r(k)^\top\mathbf{x}_i \geq 0, \mathbf{w}_r(k)^\top\mathbf{x}_j \geq 0\right\}$ is a perturbation matrix. Let $\mathbf{H}^\perp(k)$ be the $n\times n$ matrix with (i,j)-th entry being $H_{ij}^\perp(k)$. Using Lemma 4.1, with high probability we obtain an upper bound of the operator norm

$$\left\|\mathbf{H}^\perp(k)\right\|_2 \leq \sum_{(i,j)=(1,1)}^{(n,n)}\left|\mathbf{H}_{ij}^\perp(k)\right| \leq \frac{n\sum_{i=1}^{n}\left|S_i^\perp\right|}{m} \leq \frac{n^2mR}{m} \leq n^2R.$$

Similar to the classical analysis of gradient, we also need bound the quadratic term.

$$\|\mathbf{u}(k+1)-\mathbf{u}(k)\|_2^2 \leq \eta^2\sum_{i=1}^{n}\frac{1}{m}\left(\sum_{r=1}^{m}\left\|\frac{\partial L(\mathbf{W}(k))}{\partial \mathbf{w}_r(k)}\right\|_2\right)^2 \leq \eta^2n^2\left\|\mathbf{y}-\mathbf{u}(k)\right\|_2^2.$$

With these estimates at hand, we are ready to prove the induction hypothesis.

$$\begin{aligned}
\|\mathbf{y}-\mathbf{u}(k+1)\|_2^2 =& \|\mathbf{y}-\mathbf{u}(k)-(\mathbf{u}(k+1)-\mathbf{u}(k))\|_2^2\\
=& \|\mathbf{y}-\mathbf{u}(k)\|_2^2 - 2\left(\mathbf{y}-\mathbf{u}(k)\right)^\top\left(\mathbf{u}(k+1)-\mathbf{u}(k)\right) + \|\mathbf{u}(k+1)-\mathbf{u}(k)\|_2^2\\
=& \|\mathbf{y}-\mathbf{u}(k)\|_2^2 - 2\eta\left(\mathbf{y}-\mathbf{u}(k)\right)^\top\mathbf{H}(k)\left(\mathbf{y}-\mathbf{u}(k)\right)\\
&+ 2\eta\left(\mathbf{y}-\mathbf{u}(k)\right)^\top\mathbf{H}(k)^\perp\left(\mathbf{y}-\mathbf{u}(k)\right) - 2\eta\left(\mathbf{y}-\mathbf{u}(k)\right)^\top\mathbf{I}_2\\
&+ \|\mathbf{u}(k+1)-\mathbf{u}(k)\|_2^2\\
\leq& (1-\eta\lambda_0+2\eta n^2R+2\eta n^{3/2}R+\eta^2n^2)\left\|\mathbf{y}-\mathbf{u}(k)\right\|_2^2\\
\leq& (1-\frac{\eta\lambda_0}{2})\left\|\mathbf{y}-\mathbf{u}(k)\right\|_2^2.
\end{aligned}$$

The third equality we used the decomposition of $\mathbf{u}(k+1)-\mathbf{u}(k)$. The first inequality we used the Lemma 3.2, the bound on the step size, the bound on $\mathbf{I}_2$, the bound on $\left\|\mathbf{H}(k)^\perp\right\|_2$ and the bound on $\|\mathbf{u}(k+1)-\mathbf{u}(k)\|_2^2$. The last inequality we used the bound of the step size and the bound of R. Therefore Condition 4.1 holds for $k' = k+1$. Now by induction, we prove Theorem 4.1. □

5 Conclusion and Discussion

In this paper we show with over-parameterization, gradient descent provable converges to the global minimum of the empirical loss at a linear convergence rate. The key proof idea is to show the over-parameterization makes Gram matrix remain positive definite for all iterations, which in turn guarantees the linear convergence. Here we list some future directions.

First, we believe the number of hidden nodes m required can be reduced. For example, previous work (Soudry & Carmon, 2016) showed $m \geq \frac{n}{d}$ is enough to make all differentiable local minima global. In our setting, using advanced tools from probability and matrix perturbation theory to analyze $\mathbf{H}(t)$, we may able to tighten the bound.

Another direction is to prove the global convergence of gradient descent on multiple layer neural networks and convolutional neural networks. We believe our approach is still applicable because for a fixed a neural network architecture, when the number of hidden nodes is large and the initialization scheme is random Gaussian with proper scaling, the Gram matrix is also positive definite, which ensures the linear convergence of the empirical loss (at least at the begining). We believe combing our proof idea with the discovery of balancedness between layers in Du et al. (2018a) is a promising approach.

Lastly, in our paper we used the empirical loss as a potential function to measure the progress. If we use another potential function, we may able to prove the convergence rate of accelerated methods. This technique has been exploited in Wilson et al. (2016) for analyzing convex optimization. It would be interesting to bring their idea to analyze other first order methods.

REFERENCES

Alexandr Andoni, Rina Panigrahy, Gregory Valiant, and Li Zhang. Learning polynomials with neural networks. In *International Conference on Machine Learning*, pp. 1908–1916, 2014.

Alon Brutzkus and Amir Globerson. Globally optimal gradient descent for a ConvNet with gaussian inputs. In *International Conference on Machine Learning*, pp. 605–614, 2017.

Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *arXiv preprint arXiv:1805.09545*, 2018.

Amit Daniely. SGD learns the conjugate kernel class of the network. In *Advances in Neural Information Processing Systems*, pp. 2422–2430, 2017.

Damek Davis, Dmitriy Drusvyatskiy, Sham Kakade, and Jason D Lee. Stochastic subgradient method converges on tame functions. *arXiv preprint arXiv:1804.07795*, 2018.

Simon S Du and Jason D Lee. On the power of over-parametrization in neural networks with quadratic activation. *Proceedings of the 35th International Conference on Machine Learning*, pp. 1329–1338, 2018.

Simon S Du, Chi Jin, Jason D Lee, Michael I Jordan, Aarti Singh, and Barnabas Poczos. Gradient descent can take exponential time to escape saddle points. In *Advances in Neural Information Processing Systems*, pp. 1067–1077, 2017a.

Simon S Du, Jason D Lee, and Yuandong Tian. When is a convolutional filter easy to learn? *arXiv preprint arXiv:1709.06129*, 2017b.

Simon S Du, Wei Hu, and Jason D Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. *arXiv preprint arXiv:1806.00900*, 2018a.

Simon S Du, Jason D Lee, Yuandong Tian, Barnabas Poczos, and Aarti Singh. Gradient descent learns one-hidden-layer cnn: Don't be afraid of spurious local minima. *Proceedings of the 35th International Conference on Machine Learning*, pp. 1339–1348, 2018b.

C Daniel Freeman and Joan Bruna. Topology and geometry of half-rectified network optimization. *arXiv preprint arXiv:1611.01540*, 2016.

Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points − online stochastic gradient for tensor decomposition. In *Proceedings of The 28th Conference on Learning Theory*, pp. 797–842, 2015.

Rong Ge, Jason D Lee, and Tengyu Ma. Learning one-hidden-layer neural networks with landscape design. *arXiv preprint arXiv:1711.00501*, 2017.

Benjamin D Haeffele and René Vidal. Global optimality in tensor factorization, deep learning, and beyond. *arXiv preprint arXiv:1506.07540*, 2015.

Moritz Hardt and Tengyu Ma. Identity matters in deep learning. *arXiv preprint arXiv:1611.04231*, 2016.

Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M. Kakade, and Michael I. Jordan. How to escape saddle points efficiently. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1724–1732, 2017.

Kenji Kawaguchi. Deep learning without poor local minima. In *Advances In Neural Information Processing Systems*, pp. 586–594, 2016.

Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. *arXiv preprint arXiv:1808.01204*, 2018.

Yuanzhi Li and Yang Yuan. Convergence analysis of two-layer neural networks with relu activation. In *Advances in Neural Information Processing Systems*, pp. 597–607, 2017.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layers neural networks. *Proceedings of the National Academy of Sciences*.

Quynh Nguyen and Matthias Hein. The loss surface of deep and wide neural networks. In *International Conference on Machine Learning*, pp. 2603–2612, 2017.

Itay Safran and Ohad Shamir. On the quality of the initial basin in overspecified neural networks. In *International Conference on Machine Learning*, pp. 774–782, 2016.

Itay Safran and Ohad Shamir. Spurious local minima are common in two-layer ReLU neural networks. *In International Conference on Machine Learning*, pp. 4433–4441, 2018.

Mahdi Soltanolkotabi. Learning ReLus via gradient descent. In *Advances in Neural Information Processing Systems*, pp. 2007–2017, 2017.

Mahdi Soltanolkotabi, Adel Javanmard, and Jason D Lee. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Transactions on Information Theory*, 2018.

Daniel Soudry and Yair Carmon. No bad local minima: Data independent training error guarantees for multilayer neural networks. *arXiv preprint arXiv:1605.08361*, 2016.

Yuandong Tian. An analytical formula of population gradient for two-layered ReLU network and its applications in convergence and critical point analysis. In *International Conference on Machine Learning*, pp. 3404–3413, 2017.

Russell Tsuchida, Farbod Roosta-Khorasani, and Marcus Gallagher. Invariance of weight distributions in rectified mlps. *arXiv preprint arXiv:1711.09090*, 2017.

Luca Venturi, Afonso Bandeira, and Joan Bruna. Neural networks with finite intrinsic dimension have no spurious valleys. *arXiv preprint arXiv:1802.06384*, 2018.

Ashia C Wilson, Benjamin Recht, and Michael I Jordan. A Lyapunov analysis of momentum methods in optimization. *arXiv preprint arXiv:1611.02635*, 2016.

Bo Xie, Yingyu Liang, and Le Song. Diverse neural network learns true target functions. In *Artificial Intelligence and Statistics*, pp. 1216–1224, 2017.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *NIN*, 8:35–67.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

Yi Zhou and Yingbin Liang. Critical points of neural networks: Analytical forms and landscape properties. *arXiv preprint arXiv:1710.11205*, 2017.

A TECHNICAL PROOFS

Proof of Lemma 3.1. For note every fixed (i,j) pair, $\mathbf{H}_{ij}(0)$ is an average of independent random variables. Therefore, by Hoeffding inequality, we have with probability $1-\delta'$,

$$\left|\mathbf{H}_{ij}(0)-\mathbf{H}_{ij}^{\infty}\right| \leq \frac{2\log(1/\delta')}{\sqrt{m}}.$$

Setting $\delta' = n^2\delta$ and applying union bound over (i,j) pairs, we have for every (i,j) pair with probability at least $1-\delta$

$$\left|\mathbf{H}_{ij}(0)-\mathbf{H}_{ij}^{\infty}\right| \leq \frac{4\log(n/\delta)}{\sqrt{m}}.$$

Thus we have

$$\|\mathbf{H}(0)-\mathbf{H}^{\infty}\|_2^2 \leq \|\mathbf{H}(0)-\mathbf{H}^{\infty}\|_F^2 \leq \sum_{i,j}\left|\mathbf{H}_{ij}(0)-\mathbf{H}_{ij}^{\infty}\right|^2 \leq \frac{16n^2\log^2(n/\delta)}{m}.$$

Thus if $m=\Omega\left(\frac{n^2\log^2(n/\delta)}{\lambda_0^2}\right)$ we have the desired result. □

Proof of Lemma 3.4. Suppose the conclusion does not hold at time t. If there exists $r\in[m]$, $\|\mathbf{w}_r(t)-\mathbf{w}_r(0)\| \geq R'$ or $\|\mathbf{y}-\mathbf{u}(t)\|_2^2 > \exp(-\lambda_0 t)\|\mathbf{y}-\mathbf{u}(0)\|_2^2$, then by Lemma 3.3 we know there exists $s\leq t$ such that $\lambda_{\min}(\mathbf{H}(s)) < \frac{1}{2}\lambda_0$. By Lemma 3.2 we know there exists

$$t_0 = \inf\left\{t>0 : \max_{r\in[m]}\|\mathbf{w}_r(t)-\mathbf{w}_r(0)\|_2^2 \geq R\right\}.$$

Thus at t_0, there exists $r\in[m]$, $\|\mathbf{w}_r(t_0)-\mathbf{w}_r(0)\|_2^2 = R$. Now by Lemma 3.2, we know $\mathbf{H}(t_0)\geq \frac{1}{2}\lambda_0$ for $t'\leq t_0$. However, by Lemma 3.3, we know $\|\mathbf{w}_r(t_0)-\mathbf{w}_r(0)\|_2 < R' < R$. Contradiction.

For the other case, at time t, $\lambda_{\min}(\mathbf{H}(t)) < \frac{1}{2}\lambda_0$ we know we know there exists

$$t_0 = \inf\left\{t\geq 0 : \max_{r\in[m]}\|\mathbf{w}_r(t)-\mathbf{w}_r(0)\|_2^2 \geq R\right\}.$$

The rest proof is the same as the previous case. □

Proof of Corollary 4.1. We use the norm of gradient to bound this distance.

$$\begin{aligned}
\|\mathbf{w}_r(k+1)-\mathbf{w}_r(0)\|_2 \leq& \eta\sum_{k'=0}^{k}\left\|\frac{\partial L(\mathbf{W}(k'))}{\partial \mathbf{w}_r(k')}\right\|_2 \\
\leq& \eta\sum_{k'=0}^{k}\frac{\sqrt{n}\,\|\mathbf{y}-\mathbf{u}(k')\|_2}{\sqrt{m}} \\
\leq& \eta\sum_{k'=0}^{k}\frac{\sqrt{n}(1-\frac{\eta\lambda}{2})^{k'/2}}{\sqrt{m}}\|\mathbf{y}-\mathbf{u}(k')\|_2 \\
\leq& \eta\sum_{k'=0}^{\infty}\frac{\sqrt{n}(1-\frac{\eta\lambda_0}{2})^{k'/2}}{\sqrt{m}}\|\mathbf{y}-\mathbf{u}(k')\|_2 \\
=& \frac{4\sqrt{n}\,\|\mathbf{y}-\mathbf{u}(0)\|_2}{\sqrt{m}\lambda_0}.
\end{aligned}$$

□

Proof of Lemma 4.1. For a fixed $i\in[n]$ and $r\in[m]$, by anti-concentration inequality, we know $\mathbb{P}(A_{ir})\leq \frac{2R}{\sqrt{2\pi}}$. Thus we can bound the size of $S_i^{\perp}$ in expectation.

$$\mathbb{E}\left[\left|S_i^{\perp}\right|\right] = \sum_{r=1}^{m}\mathbb{P}(A_{ir}) \leq \frac{2mR}{\sqrt{2\pi}}. \tag{7}$$

Summing over $i = 1, \ldots, n$, we have

$$\mathbb{E}\left[\sum_{i=1}^{n} \left|S_i^{\perp}\right|\right] \leq \frac{2mnR}{\sqrt{2\pi}}.$$

Thus by Markov's inequality, we have with high probability

$$\sum_{i=1}^{n} \left|S_i^{\perp}\right| \geq CmnR. \tag{8}$$

for some large positive constant $C > 0$. □