Embedding Surfaces by Optimizing Neural Networks with Prescribed Riemannian Metric and Beyond

Yi Feng [*1] **Sizhe Li** [*2] **Ioannis Panageas** [*3] **Xiao Wang** [*1]

Abstract

From a machine learning perspective, the problem of solving partial differential equations (PDEs) can be formulated into a least square minimization problem, where neural networks are used to parametrized PDE solutions. Ideally a global minimizer of the square loss corresponds to a solution of the PDE. In this paper we start with a special type of nonlinear PDE arising from differential geometry, the isometric embedding equation, which relates to many long-standing open questions in geometry and analysis. We show that the gradient descent method can identify a global minimizer of the least-square loss function with two-layer neural networks under the assumption of over-parametrization. As a consequence, this solves the surface embedding locally with a prescribed Riemannian metric. We also extend the convergence analysis for gradient descent to higher order linear PDEs with over-parametrization assumption.

1. Introduction

In recent years deep learning has revolutionized many fields of science and engineering, including a variety of applications of deep learning in applied mathematics. There have been many breakthroughs in solving partial differential equations (PDEs) (M.W.M.G.Dissanayake & Phan-Thien, 1994; I.E.Lagaris et al., 1998; Rudd & Ferrari, 2015; G.Carleo & M.Troyer, 2017; J.Han et al., 2018; E et al., 2017; Raissi et al., 2019; Huang et al., 2020; Luo & Yang, 2020). The main idea of these approaches is to reformulate the PDE solution into a global minimizer of an expectation minimization problem, where deep neural networks (DNNs) are applied for discretization and the stochastic gradient descent (SGD) is adopted to solve the minimization problem. As a case of recent successes, physics informed neural networks (PINNs) have been widely used for robust and accurate approximate of PDEs. Deep neural networks possess the so-called universal approximation property (A.R.Barron, 1993; G.Cybenko, 1989; Hornik et al., 1989), namely any continuous, even measurable, function can be approximated by DNNs, (D.Yarotsky, 2017) has provided a precise descriptions of the required neural network architecture for functions with sufficient Sobolev regularity. Based on this result, it is natural to use deep neural networks for the space of solutions of PDEs.

In deep learning, the fact that first order methods like gradient descent can achieve zero training loss for non-convex objective functions remained mysteries until the merging of analysis based on over-parametrization. Two-layer fully connected ReLU activated neural networks were proven to achieve zero training error with high probability (Du et al., 2019). Their analysis relies on over-parametrization and random initialization jointly restrict every weight vector to be close to its initialization for all iterations, which allow one to exploit a strong convexity-like property to prove that gradient descent converges at a global linear to the global optimum. Extending the analysis of function approximation with over-parametrization to linear PDEs, authors of (Luo & Yang, 2020) provide optimization and generalization analysis for second order linear PDEs.

Despite the remarkable empirical successes in solving PDEs with neural networks and in theoretical analysis in optimizing loss function in function approximation and solving linear PDEs, it is less understood how gradient descent works in solving non-linear PDEs in general, even with over-parametrization assumptions. Orthogonal to linear PDEs (or semi-linear PDEs) which has a relatively uniform structure, non-linear PDEs diverse case by case and it is almost impossible to formulate a simple structure. For example, Monge-Ampere equation $\det(D^2u) - f(x, u, Du) = 0$ and KdV equation $u_t + u_{xxx} - 6uu_x = 0$ has completely different algebraic forms regarding them as functions of partial derivatives. Due to the complexity introduced by the non-linearity of all kinds of PDEs, the loss function usually contains highly non-linear expressions of parameters and

[*] Equal contribution [1] Shanghai University of Finance and Economics [2] Huazhong University of Science and Technology [3] University of California, Irvine. Correspondence to: Xiao Wang <wangxiao@sufe.edu.cn>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

activation functions. This is the main challenge in analyzing convergence property of first order methods like gradient descent in optimizing the loss functions induced by non-linear PDEs.

In order to leverage the power of over-parametrization in obtaining provable convergence of gradient descent in solving linear PDEs, we step out into a classic non-linear PDEs problems arising from differential geometry, i.e., the systems of isometric embedding of Riemannian manifolds. This type of PDEs are of interests in community of geometry and analysis. We start the convergence analysis for non-linear PDEs with isometric embedding systems because they are both mildly and sufficiently complicated in the sense that the algebraic function on the partial derivatives are quadratic functions (non-linear but not out of control), they are first order PDEs (the partial derivatives of the neural networks will not be too complicated to be handled), there is no explicit solutions in general for isometric embedding (Nash's embedding theorem is essentially an algorithm of approaching the exact solution of these PDEs), and embedding a Riemannian manifold into Euclidean space has important application in manifold learning and dimensionality reduction for high dimensional data sets. Moreover, we extend the over-parametrization based arguments of function approximation of (Du et al., 2019) to higher order PDEs which only contain partial derivatives of the same order.

2. Preliminaries

2.1. Over-parametrized Neural Networks

A widely believed explanation on why a neural network can fit all training labels in that the neural network is over-parametrized. A theoretical analysis on the convergence of gradient descent in optimizing two-layer neural networks is given by (Du et al., 2019). Formally, we consider a neural network of the following form.

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

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 (for convenience we assume $\mathbf{w}_r$ are row vectors), $a_r \in \mathbb{R}$ is the output weight and $\sigma(\cdot)$ is the activation function. Especially, in the surface embedding problems considered in this paper, $\mathbf{x}$ is a two-dimensional vector and $\sigma(\cdot) = \mathrm{ReLU}^2$ which is sufficient for smoothness requirement of the PDE of isometric embedding.

2.2. Surfaces in Low Dimensional Euclidean Spaces

In this section, we review briefly the classical theory of differentiable surfaces. In our notation $\Omega \subset \mathbb{R}^2$ is an open set in the plane and points of Ω are denoted by $\mathbf{x} = (x_1, x_2) \in \mathbb{R}^2$. A differentiable map $\mathbf{r} : \Omega \to \mathbb{R}^3$ induces a linear transform $d\mathbf{r} : \mathbb{R}^2 \to \mathbb{R}^3$ for each $\mathbf{x} \in \Omega$. Then $\mathbf{r}$ is call a regular surface if $d\mathbf{r}(\mathbf{x})$ is injective for each $\mathbf{x} \in \Omega$. The inner product on $\mathbb{R}^3$ composed with the linear map $d\mathbf{r}(\mathbf{x}) : \mathbb{R}^2 \to \mathbb{R}^3$ induces a quadratic differential form on $\mathbb{R}^2$, which is called the *first fundamental form* and is denoted by $g(\mathbf{x})$, i.e., a Riemannian metric induced by the ambient space $\mathbb{R}^3$. We also use $I(\mathbf{x})$ or I as the first fundamental form. Precisely, we have

$$I(\mathbf{x})(X, Y) = d\mathbf{r}(\mathbf{x})X \cdot d\mathbf{r}(\mathbf{x})Y \text{ for any } X, Y \in \mathbb{R}^2.$$

We usually write this, using the summation convention, as

$$I = d\mathbf{r} \cdot d\mathbf{r} = g_{ij} dx_i dx_j$$

where

$$g_{ij} = \partial_i \mathbf{r} \cdot \partial_j \mathbf{r}, \ \ i, j = 1, 2.$$

We call (g_{ij}) the coefficients of the first fundamental forms.

3. Main Results

3.1. Gradient Descent for Surface Embedding

The local isometric embedding problem, which is mainly considered in this paper, can be reduced to a constrained approximation problem by implementing the embedding map $\mathbf{r}$ as a function defined on Ω. Suppose the surface is the graph of certain function $h : \Omega \to \mathbb{R}$, and $\mathbf{r}(x_1, x_2)$ is defined as follows.

$$\mathbf{r}(x_1, x_2) = (x_1, x_2, h(x_1, x_2)) \in \mathbb{R}^3.$$

Then finding $\mathbf{r}$ such that $g_{ij} = \partial_i \mathbf{r} \cdot \partial_j \mathbf{r}$ is equivalent to finding a function h such that the following system of partial differential equations is satisfied,

$$\begin{aligned} 1 + \left(\frac{\partial h}{\partial x_1}\right)^2 &= g_{11} \\ 1 + \left(\frac{\partial h}{\partial x_2}\right)^2 &= g_{22} \\ \frac{\partial h}{\partial x_1} \cdot \frac{\partial h}{\partial x_2} &= g_{12} \end{aligned} \tag{1}$$

where g_{ij} are prescribed differentiable functions defined on Ω.

Let $\sigma(\cdot)$ be ReLU activation, and assume the neural network has form $f(\mathbf{x}, t) = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} a_i \sigma^2(\mathbf{w}_i(t)^\top \mathbf{x})$, where $\mathbf{w}_i, \mathbf{x} \in \mathbb{R}^2$, and there are N sample points $\{\mathbf{x}_i\}_{i=1}^{N}$, and we also assume $\|x_i\|= 1$ for convenience. In the following, we will also write $\mathbf{w}_i(t)$ by $\mathbf{w}_i$ when there is no confusion. Now we have

$$\frac{\partial f}{\partial x_1}(\mathbf{x}) = \frac{2}{\sqrt{m}} \sum_{i=1}^{m} a_i w_{i,1} \sigma(\mathbf{w}_i^\top \mathbf{x}) \mathbb{I}_{\{\mathbf{w}_i^\top \mathbf{x} \geq 0\}}$$

$$\frac{\partial f}{\partial x_2}(\mathbf{x}) = \frac{2}{\sqrt{m}}\sum_{i=1}^{m} a_i w_{i,2}\sigma(\mathbf{w}_i^\top \mathbf{x})\mathbb{I}_{\{\mathbf{w}_i^\top \mathbf{x}\geq 0\}}$$

The loss function is

$$L(\mathbf{w}) = \sum_{k=1}^{N}\left((\frac{\partial f}{\partial x_1}(\mathbf{x}_k))^2 + 1 - g_{11}(\mathbf{x}_k)\right)^2 \qquad \text{(Loss)}$$
$$+ \left((\frac{\partial f}{\partial x_2}(\mathbf{x}_k))^2 + 1 - g_{22}(\mathbf{x}_k)\right)^2 \qquad (2)$$

In the following we will denote $\frac{\partial f}{\partial x_1}(\mathbf{x}_k)$ and $\frac{\partial f}{\partial x_2}(\mathbf{x}_k)$ by $u_{1,k}(t)$ and $u_{2,k}(t)$ respectively, $\{u_{1,k}\}_{k=1}^N, \{u_{2,k}\}_{k=1}^N$ are functions of t because their values are depended on $\mathbf{w}(t)$. We also denote $y_{1,k} = g_{11}(\mathbf{x}_k) - 1$, $y_{2,k} = g_{22}(\mathbf{x}_k) - 1$, thus $\{y_{1,k}\}_{k=1}^N, \{y_{2,k}\}_{k=1}^N$ are constant numbers. With these notations, (Loss) can also be written as

$$L(\mathbf{w}) = \sum_{k=1}^{N}\left(((u_{1,k})^2 - y_{1,k})^2 + (u_{2,k})^2 - y_{2,k})^2\right). \qquad (3)$$

Note that there is a zero point of $L(\mathbf{w})$, $u_k^i = \sqrt{y_{i,k}}$ for $i = 1, 2$ and $k = 1, 2, ..., N$. Since the local embedding is nothing but a graph of a function on a small neighborhood, we can further assume that the values of $y_{i,j}$ are non-negative, i.e.,

$$\mathbf{y} = (\sqrt{y_{1,1}}, \sqrt{y_{1,2}}, ..., \sqrt{y_{1,N}}, \sqrt{y_{2,1}}, \sqrt{y_{2,2}}, ..., \sqrt{y_{2,N}})^\top.$$

Before stating the main result of surface embedding, we introduce the matrix $\mathbf{G}^\infty$ as follows,

$$\mathbf{G}^\infty_{k,b} = \begin{cases} \mathbb{E}_{\mathbf{w}\sim N(\mathbf{0},\mathbf{I})}[\mathbb{I}_{\{\mathbf{w}\mathbf{x}_k^\top \geq 0,\ \mathbf{w}\mathbf{x}_b^\top \geq 0\}} h_{11}(\mathbf{w}, \mathbf{x}_k, \mathbf{x}_b)], \\ \mathbb{E}_{\mathbf{w}\sim N(\mathbf{0},\mathbf{I})}[\mathbb{I}_{\{\mathbf{w}\mathbf{x}_k^\top \geq 0,\ \mathbf{w}\mathbf{x}_b^\top \geq 0\}} h_{12}(\mathbf{w}, \mathbf{x}_k, \mathbf{x}_b)], \\ \mathbb{E}_{\mathbf{w}\sim N(\mathbf{0},\mathbf{I})}[\mathbb{I}_{\{\mathbf{w}\mathbf{x}_k^\top \geq 0,\ \mathbf{w}\mathbf{x}_b^\top \geq 0\}} h_{21}(\mathbf{w}, \mathbf{x}_k, \mathbf{x}_b)], \\ \mathbb{E}_{\mathbf{w}\sim N(\mathbf{0},\mathbf{I})}[\mathbb{I}_{\{\mathbf{w}\mathbf{x}_k^\top \geq 0,\ \mathbf{w}\mathbf{x}_b^\top \geq 0\}} h_{22}(\mathbf{w}, \mathbf{x}_k, \mathbf{x}_b)]. \end{cases} \qquad (4)$$

where from the first to last expression, the k and b are taken from the following set respectively,

$$\begin{aligned} 1 \leq k \leq N,\ \ 1 \leq b \leq N \\ 1 \leq k \leq N,\ \ N \leq b \leq 2N \\ N \leq k \leq 2N,\ \ 1 \leq b \leq N \\ N \leq k \leq 2N,\ \ N \leq b \leq 2N \end{aligned}$$

The main result is:

Theorem 3.1. *There is a neighbourhood U of $\mathbf{y}$ such that if the neural network is over-parameterized with $m = \mathcal{O}(\frac{N^8}{\delta^4\lambda_0^6})$ and $(u_k^i(0)) \in U$, then $(u_k^i(t))$ will converge to $\mathbf{y}$ with rate $\mathcal{O}(e^{-\frac{\lambda_0 t}{4}})$where $\lambda_0 > 0$ is the least eigenvalue of $\mathbf{G}^\infty$ defined by (14).*

Similar to (Du et al., 2019), the convergence analysis relies on the understanding of the training dynamics of the gradient flow $\frac{dL(\mathbf{w})}{dt} = -\nabla L(\mathbf{w})$. Our notations enable us to write the trajectories of predictions in the following way,

$$\mathbf{u}(t) - \mathbf{y} = \begin{pmatrix} (u_{1,1}(t))^2 - y_{1,1} \\ \vdots \\ (u_{1,N}(t))^2 - y_{1,N} \\ (u_{2,1}(t))^2 - y_{2,1} \\ \vdots \\ (u_{2,N}(t))^2 - y_{2,N} \end{pmatrix}$$

In the proof of Theorem 3.1, Proposition A.1 and its corollary assert that the evolution of prediction governed by by training dynamics has a linear structure and the coefficient matrix is always positive definite such that the least eigenvalue of the coefficient matrix is uniformly lower bounded by a positive number γ. Consequently, we have that $\|\mathbf{u}(t) - \mathbf{y}\| \leq e^{-\gamma t}\|\mathbf{u}(0) - \mathbf{y}\|$. On top of this, the positive definite property of the coefficient matrix is derived by a perturbation argument of the matrix $\mathbf{G}^\infty$. The proof details are left in Appendix.

3.2. Gradient Descent for Homogeneous Linear PDEs

We proceed considering an important type of linear PDEs, the PDE that has partial derivatives of the same order. Amongst this class of PDEs, Laplace ($\sum \frac{\partial^2 f}{\partial x_i} = 0$) and Poisson ($\sum \frac{\partial^2 f}{\partial x_i} = h$) equations might be of highest interest in applications. To see that gradient descent can solve homogeneous linear PDEs, we compute the partial derivatives of the neural network.

$$\begin{aligned} \frac{\partial f}{\partial x_i} &= \frac{\partial}{\partial x_i}\left(\frac{1}{\sqrt{m}}\sum_{k=1}^{m} a_k\sigma(\mathbf{w}_k\mathbf{x})\right) = \frac{1}{\sqrt{m}}\sum_{k=1}^{m} a_k\frac{\partial\sigma(\mathbf{w}_k\mathbf{x})}{\partial x_i} \\ &= \frac{1}{\sqrt{m}}\sum_{k=1}^{m} a_k\sigma'(\mathbf{w}_k\mathbf{x})w_{ki}, \end{aligned}$$

and the most fundamental homogeneous linear PDE is of the following form,

$$\frac{\partial f}{\partial x_1} + ... + \frac{\partial f}{\partial x_d} = \frac{1}{\sqrt{m}}\sum_{k=1}^{m} a_k\left(\sum_{s=1}^{d} w_{ks}\right)\sigma'(\mathbf{w}_k\mathbf{x}) = h(\mathbf{x}). \qquad (5)$$

Naturally, solving the above PDE boils down to optimization of the loss function defined by a set of sample points

$\{\mathbf{x}_i, y_i\}_{i=1}^n$ such that $h(\mathbf{x}_i) = y_i$ for all $i \in [n]$. We claim that the convergence analysis of the gradient flow of the corresponding loss function can be reduced to that we have just obtained from last section, i.e., the convergence analysis of RePU networks. In (5), we can let $b_k = a_k \left(\sum_{s=1}^d w_{ks}\right)$ and then the PDE (5) is of the following form:

$$\frac{\partial f}{\partial x_1} + ... + \frac{\partial f}{\partial x_d} = \frac{1}{\sqrt{m}} \sum_{k=1}^m b_k \sigma'(\mathbf{w}_k \mathbf{x}) = h(\mathbf{x}).$$

The advantage of considering the above PDE is obvious. We can solve the approximation problem $\frac{1}{\sqrt{m}} \sum_{k=1}^m b_k \sigma'(\mathbf{w}_k \mathbf{x}) = h(\mathbf{x})$ by considering $\mathbf{b} = (b_1, ..., b_m)$ as independent variables with respect to $\mathbf{W}$, with the convergence rate that is just obtained in last section. Some extra effort is necessary to solve algebraic equations $b_k = a_k \left(\sum_{s=1}^d w_{ks}\right)$ to obtain the actual parameters for $\mathbf{a}$. Some cares should be taken in solving these algebraic equations since $\sum_{s=1}^d w_{ks}$ is likely to be 0. But this will not cause the process collapse in general for two reasons. Firstly, the set of parameters $\{\mathbf{w}_k\}_{k=1}^m$ such that $\sum_{s=1}^d w_{ks} = 0$ for some k is of measure zero, which means the chance that $a_k = b_k / (\sum_{s=1}^d w_{ks})$ blows is ignorable. Secondly, even if $\sum_{s=1}^d w_{ks} = 0$ for some k, we can add a small perturbation on this set of $\{w_{ks}\}_{s=1}^d$ such that the sum is not zero. Formally, the main convergence results for homogeneous linear equation of arbitrary order can be stated below, where we assume the coefficient of each partial derivative to be 1 without loss of generality.

Our observation for the structure of linear PDEs also holds for higher order partial derivatives. In general, the higher order partial derivatives of function f is written as

$$D^\alpha f = \frac{\partial^r f}{\partial x_1^{\alpha_1}, ..., \partial x_n^{\alpha_n}}$$

where $\alpha = (\alpha_1, ..., \alpha_n)$, such that $|\alpha| = \alpha_1 + ... + \alpha_n \leq r$. With these notations, we are ready to state the convergence result of gradient descent using in solving homogeneous linear PDEs of arbitrary order. In the rest of this paper, we focus on two-layer neural networks with rectified power unit (RePU) activation, i.e.,

$$\sigma(x) = x^\ell \text{ if } x \geq 0 \text{ and } \sigma(x) = 0 \text{ if } x < 0.$$

Theorem 3.1. *Let $\alpha = (\alpha_1, ..., \alpha_d)$ be a partition of integer p, and $\sigma(x)$ be a RePU activation function of smoothness higher than p. Then the loss function defined by following PDE reaches 0 by running gradient descent.*

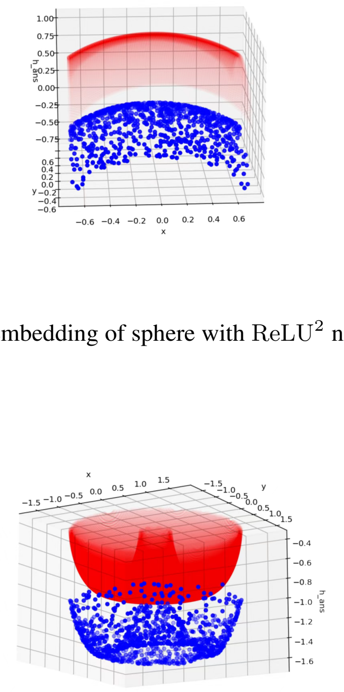
$$\sum_\alpha D^\alpha f = h(\mathbf{x}).$$

where α runs over all partitions of $|\alpha|$.

The proof is left in Appendix due to space constraint.

4. Experiments

In this section we illustrate that the gradient descent actually finds isometric embeddings of sphere and torus in $\mathbb{R}^3$. The graph of hemisphere is given by function $h(x_1, x_2) = \sqrt{1 - x_1^2 - x_2^2}$, and the graph of torus is given by function $h(x_1, x_2) = \sqrt{r^2 - (\sqrt{x_1^2 + x_2^2} - R)^2}$.



(a) Isometric embedding of sphere with ReLU^2 neural networks.

(b) Isometric embedding of torus with ReLU^2 neural networks.

Figure 1. Solving PDEs of sphere and torus embedding with gradient descent.

5. Conclusion

In this paper, we investigate local isometric embedding of surfaces into Euclidean space from the perspective of PDE solving with neural networks. We show that over-parametrization is a condition that guarantees convergence result of gradient flow in solving such non-linear PDEs. As an extension of the arguments, we generalize the over-parametrization based analysis to higher oder linear PDEs whose partial derivatives are of the same order.

References

A.R.Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. Inform. Theory.*, 39(3):930945, 1993.

Du, S. S., Zhai, X., Poczós, B., and Singh, A. Gradient descent provably optimizes over-parameterized neural networks. In *ICLR*, 2019.

D.Yarotsky. Error bounds for approximations with deep relu networks. *Neural Networks*, 94:103114, 2017.

E, W., Han, J., and Jentzen, A. Deep learning-based numerical methods for high-dimensional parabolic partial differential equations and backward stochastic differential equations. *Communications in Mathematics and Statistics*, 5(4):349380, 2017.

G.Carleo and M.Troyer. Solving the quantum many-body problem with artificial neural networks. *Science*, 355:602606, 2017.

G.Cybenko. Approximations by superpositions of sigmoidal functions. *Approximation theory and its applications.*, 9(3):1728, 1989.

Hornik, K., Stinchcombe, M., and White, H. Multilayer feedforward networks are universal approxi- mators. *Neural networks*, 2(5):359366, 1989.

Huang, J., Wang, H., and Yang, H. Int-deep: A deep learning initialized iter- ative method for nonlinear problems. *Journal of Computational Physics*, page 109675, 2020.

I.E.Lagaris, Likas, A., and D.I.Fotiadis. Artificial neural networks for solving ordinary and partial differential equations. *IEEE Trans. Neural Networks*, 9:987-1000, 1998.

J.Han, Jentzen, A., and W.E. Solving high-dimensional partial differential equations using deep learning. *Proc. Natl. Acad. Sci. USA*, 115:85058510, 2018.

Luo, T. and Yang, H. Two-layer neural networks for partial differential equations: Optimization and generalization theory. *https://arxiv.org/abs/2006.15733*, 2020.

M.W.M.G.Dissanayake and Phan-Thien, N. Neural-network-based approximation for solving partial differential equations. *Comm. Numer. Methods Engrg*, 10:195-201, 1994.

Raissi, M., Perdikaris, P., and Karniadakis, G. Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.*, 378:686 707, 2019.

Rudd, K. and Ferrari, S. A constrained integration (cint) approach to solving partial differential equations using artificial neural networks. *Neurocomputing*, 155:277-285, 2015.

A. Proof of Theorem 3.1

Proposition A.1. *Denote*

$$\mathbf{u}(t)-\mathbf{y}=((u_{1,1}(t))^2-y_{1,1},...,(u_{1,N}(t))^2-y_{1,N},(u_{2,1}(t))^2-y_{2,1},...,(u_{2,N}(t))^2-y_{2,N}), \tag{6}$$

then

$$\begin{aligned}\frac{dL(\mathbf{w})}{dt}&=\frac{d\|\mathbf{u}(t)-\mathbf{y}\|^2}{dt}\\&=-16(\mathbf{u}(t)-\mathbf{y})\cdot\\&\begin{bmatrix}u_{1,1}(t)&&&&&\\&\ddots&&&&\\&&u_{1,N}(t)&&&\\&&&u_{2,1}(t)&&\\&&&&\ddots&\\&&&&&u_{2,N}(t)\end{bmatrix}\cdot\mathbf{G}(t)\cdot\begin{bmatrix}u_{1,1}(t)&&&&&\\&\ddots&&&&\\&&u_{1,N}(t)&&&\\&&&u_{2,1}(t)&&\\&&&&\ddots&\\&&&&&u_{2,N}(t)\end{bmatrix}\\&\cdot(\mathbf{u}(t)-\mathbf{y})^\top\end{aligned} \tag{7}$$

where $\mathbf{G}(t)\in\mathbb{R}^{2N\times 2N}$ *and*

$$\mathbf{G}_{k,b}(t)=\begin{cases}\sum_{i=1}^m(\nabla_{\mathbf{w}_i}u_{1,b})^\top\nabla_{\mathbf{w}_i}u_{1,k}, & \textit{when } 1\le k\le N, 1\le b\le N\,.\\ \sum_{i=1}^m(\nabla_{\mathbf{w}_i}u_{2,b-N})^\top\nabla_{\mathbf{w}_i}u_{1,k}, & \textit{when } 1\le k\le N, N<b\le 2N.\\ \sum_{i=1}^m(\nabla_{\mathbf{w}_i}u_{1,b})^\top\nabla_{\mathbf{w}_i}u_{2,k-N}, & \textit{when } N<k\le 2N, 1\le b\le N.\\ \sum_{i=1}^m(\nabla_{\mathbf{w}_i}u_{2,b-N})^\top\nabla_{\mathbf{w}_i}u_{2,k-N}, & \textit{when } N<k\le 2N, N<b\le 2N.\end{cases}$$

Our aim is to prove when $(u_{1,1}(t),...,u_{1,N}(t),u_{2,1}(t),...,u_{2,N}(t))$ lies in some neighbourhood of $\mathbf{y}$, then

$$\begin{bmatrix}u_{1,1}(t)&&&&&\\&\ddots&&&&\\&&u_{1,N}(t)&&&\\&&&u_{2,1}(t)&&\\&&&&\ddots&\\&&&&&u_{2,N}(t)\end{bmatrix}\cdot\mathbf{G}(t)\cdot\begin{bmatrix}u_{1,1}(t)&&&&&\\&\ddots&&&&\\&&u_{1,N}(t)&&&\\&&&u_{2,1}(t)&&\\&&&&\ddots&\\&&&&&u_{2,N}(t)\end{bmatrix} \tag{8}$$

has a uniform positive lower bound of eigenvalues. If this is true, we will have following proposition about the convergence rate of $\|\mathbf{u}(t)-\mathbf{y}\|$:

Corollary A.2. *If the matrix in* (8) *is always positive definite and the smallest eigenvalue of* (8)*has a uniform lower bound* $\gamma>0$, *then we have*

$$\|\mathbf{u}(t)-\mathbf{y}\|\le e^{-\gamma t}\|\mathbf{u}(0)-\mathbf{y}\|. \tag{9}$$

Proof. This is a simple integral of (7). □

The proof of Proposition A.1 is completed by combining the following calculations.

Recall the loss function is

$$L(\mathbf{w}(t))=\sum_{k=1}^N\left(((u_{1,k}(t))^2-y_{1,k})^2+((u_{2,k}(t))^2-y_{2,k})^2\right)=\|\mathbf{u}(t)-\mathbf{y}\|^2$$

Then we have

$$\frac{dL(\mathbf{w}(t))}{dt} = \sum_{i=1}^{2}\sum_{j=1}^{N} \frac{\partial L(\mathbf{w})}{\partial (u_j^i)^2} \cdot \frac{d(u_j^i)^2}{dt} \tag{10}$$

$$= 4 \cdot ((u_{1,1})^2 - y_{1,1}, ..., (u_{1,N})^2 - y_{1,N}, (u_{2,1})^2 - y_{2,1}, ..., (u_{2,N})^2 - y_{2,N}) \cdot \begin{bmatrix} \frac{du_{1,1}}{dt} \cdot u_{1,1} \\ \vdots \\ \frac{du_{1,N}}{dt} \cdot u_{1,N} \\ \frac{du_{2,1}}{dt} \cdot u_{2,1} \\ \vdots \\ \frac{du_{2,N}}{dt} \cdot u_{2,N} \end{bmatrix} \tag{11}$$

$$= 4 \cdot (\mathbf{u}(t) - \mathbf{y}) \cdot \begin{bmatrix} \frac{du_{1,1}}{dt} \cdot u_{1,1} \\ \vdots \\ \frac{du_{1,N}}{dt} \cdot u_{1,N} \\ \frac{du_{2,1}}{dt} \cdot u_{2,1} \\ \vdots \\ \frac{du_{2,N}}{dt} \cdot u_{2,N} \end{bmatrix} \tag{12}$$

$$= 4 \cdot (\mathbf{u}(t) - \mathbf{y}) \cdot \begin{bmatrix} u_{1,1}(t) & & & & & \\ & \ddots & & & & \\ & & u_{1,N}(t) & & & \\ & & & u_{2,1}(t) & & \\ & & & & \ddots & \\ & & & & & u_{2,N}(t) \end{bmatrix} \cdot \begin{bmatrix} \frac{du_{1,1}}{dt} \\ \vdots \\ \frac{du_{1,N}}{dt} \\ \frac{du_{2,1}}{dt} \\ \vdots \\ \frac{du_{2,N}}{dt} \end{bmatrix} \tag{13}$$

Now we can calculate $\frac{du_{1,k}}{dt}$ for $k = 1, 2, ..., N$ straightforwardly.

$$\frac{du_{1,k}}{dt} = \sum_{i=1}^{m} (\nabla_{\mathbf{w}_i} u_{1,k})^\top \cdot \frac{d\mathbf{w}_i}{dt},$$

where

$$\nabla_{\mathbf{w}_i} u_{1,k} = (\frac{\partial u_{1,k}}{\partial \mathbf{w}_{i,1}}, \frac{\partial u_{1,k}}{\partial \mathbf{w}_{i,2}})^\top \in \mathbb{R}^2,$$

and

$$\frac{d\mathbf{w}_i}{dt} = (\frac{d\mathbf{w}_{i,1}}{dt}, \frac{d\mathbf{w}_{i,2}}{dt})^\top \in \mathbb{R}^2.$$

Using gradient descent, we have

$$\frac{d\mathbf{w}_i}{dt} = -\frac{\partial L(\mathbf{w})}{\partial \mathbf{w}_i},$$

and

$$\frac{\partial L(\mathbf{w})}{\partial \mathbf{w}_i} = 4 \sum_{s=1}^{N} \left([(u_{1,s})^2 - y_{1,s}] \cdot u_{1,s} \cdot \nabla_{\mathbf{w}_i} u_{1,s} + [(u_{2,s})^2 - y_{2,s}] \cdot u_{2,s} \cdot \nabla_{\mathbf{w}_i} u_{2,s} \right)$$

and moreover, combining above expressions that is calculated, we can furthermore conclude

$$
\begin{aligned}
\frac{du_{1,k}}{dt} &= \sum_{i=1}^{m} (\nabla_{\mathbf{w}_i} u_{1,k})^\top \cdot \frac{d\mathbf{w}_i}{dt} \\
&= -\sum_{i=1}^{m} (\nabla_{\mathbf{w}_i} u_{1,k})^\top \cdot \frac{\partial L(\mathbf{w})}{\partial \mathbf{w}_i} \\
&= -4 \sum_{i=1}^{m} \sum_{s=1}^{N} \left([(u_{1,s})^2 - y_{1,s}] \cdot u_{1,s} \cdot \nabla_{\mathbf{w}_i} u_{1,s} \cdot (\nabla_{\mathbf{w}_i} u_{1,k})^\top + [(u_{2,s})^2 - y_{2,s}] \cdot u_{2,s} \cdot \nabla_{\mathbf{w}_i} u_{2,s} \cdot (\nabla_{\mathbf{w}_i} u_{1,k})^\top \right) \\
&= -4 \cdot \left(\sum_{i=1}^{m} \nabla_{\mathbf{w}_i} u_{1,1} \cdot (\nabla_{\mathbf{w}_i} u_{1,k})^\top, ..., \sum_{i=1}^{m} \nabla_{\mathbf{w}_i} u_{1,N} \cdot (\nabla_{\mathbf{w}_i} u_{1,k})^\top, \sum_{i=1}^{m} \nabla_{\mathbf{w}_i} u_{2,1} \cdot (\nabla_{\mathbf{w}_i} u_{1,k})^\top, ..., \sum_{i=1}^{m} \nabla_{\mathbf{w}_i} u_{2,N} \cdot (\nabla_{\mathbf{w}_i} u_{1,k})^\top \right) \\
&\quad \cdot \begin{bmatrix} u_{1,1}(t) & & & & & \\ & \ddots & & & & \\ & & u_{1,N}(t) & & & \\ & & & u_{2,1}(t) & & \\ & & & & \ddots & \\ & & & & & u_{2,N}(t) \end{bmatrix} \cdot (\mathbf{u}(t) - \mathbf{y})^\top .
\end{aligned}
$$

Same calculation for $\frac{du_{2,k}}{dt}$ gives

$$
\begin{aligned}
\frac{du_{2,k}}{dt} &= -4 \cdot \left(\sum_{i=1}^{m} \nabla_{\mathbf{w}_i} u_{1,1} \cdot (\nabla_{\mathbf{w}_i} f_{2,k})^\top, ..., \sum_{i=1}^{m} \nabla_{\mathbf{w}_i} u_{1,N} \cdot (\nabla_{\mathbf{w}_i} f_{2,k})^\top, \sum_{i=1}^{m} \nabla_{\mathbf{w}_i} u_{2,1} \cdot (\nabla_{\mathbf{w}_i} f_{2,k})^\top, ..., \sum_{i=1}^{m} \nabla_{\mathbf{w}_i} u_{2,N} \cdot (\nabla_{\mathbf{w}_i} f_{2,k})^\top \right) \\
&\quad \cdot \begin{bmatrix} u_{1,1}(t) & & & & & \\ & \ddots & & & & \\ & & u_{1,N}(t) & & & \\ & & & u_{2,1}(t) & & \\ & & & & \ddots & \\ & & & & & u_{2,N}(t) \end{bmatrix} \cdot (\mathbf{u}(t) - \mathbf{y})^\top .
\end{aligned}
$$

Combine above with equation 13, we finish the proof of Proposition A.1. The following notations and results will be used for convenience in later proof. Let

$$
\mathbf{S} = \begin{bmatrix}
(\nabla_{\mathbf{w}_1} u_{1,1})^\top & (\nabla_{\mathbf{w}_2} u_{1,1})^\top & ... & (\nabla_{\mathbf{w}_m} u_{1,1})^\top \\
\vdots & \vdots & \vdots & \vdots \\
(\nabla_{\mathbf{w}_1} u_{1,N})^\top & (\nabla_{\mathbf{w}_2} u_{1,N})^\top & ... & (\nabla_{\mathbf{w}_m} u_{1,N})^\top \\
(\nabla_{\mathbf{w}_1} u_{2,1})^\top & (\nabla_{\mathbf{w}_2} u_{2,1})^\top & ... & (\nabla_{\mathbf{w}_m} u_{2,1})^\top \\
\vdots & \vdots & \vdots & \vdots \\
(\nabla_{\mathbf{w}_1} u_{2,N})^\top & (\nabla_{\mathbf{w}_2} u_{2,N})^\top & ... & (\nabla_{\mathbf{w}_m} u_{2,N})^\top
\end{bmatrix} \in \mathbb{R}^{2N \times 2m},
$$

then $\mathbf{G} = \mathbf{S} \cdot \mathbf{S}^\top$, thus $\mathbf{G}$ is a symmetric matrix.

We next establish the positiveness of matrix $\mathbf{G}^\infty$. Let

$$
\mathbf{G}^\infty_{k,b} = \begin{cases}
\mathbb{E}_{\mathbf{w} \sim N(\mathbf{0},\mathbf{I})} [\mathbb{I}_{\{\mathbf{w}\mathbf{x}_k^\top \geq 0,\ \mathbf{w}\mathbf{x}_b^\top \geq 0\}} h_{11}(\mathbf{w}, \mathbf{x}_k, \mathbf{x}_b)], & \text{when } 1 \leq k \leq N, 1 \leq b \leq N\ . \\
\mathbb{E}_{\mathbf{w} \sim N(\mathbf{0},\mathbf{I})} [\mathbb{I}_{\{\mathbf{w}\mathbf{x}_k^\top \geq 0,\ \mathbf{w}\mathbf{x}_b^\top \geq 0\}} h_{12}(\mathbf{w}, \mathbf{x}_k, \mathbf{x}_b)], & \text{when } 1 \leq k \leq N, N < b \leq 2N. \\
\mathbb{E}_{\mathbf{w} \sim N(\mathbf{0},\mathbf{I})} [\mathbb{I}_{\{\mathbf{w}\mathbf{x}_k^\top \geq 0,\ \mathbf{w}\mathbf{x}_b^\top \geq 0\}} h_{21}(\mathbf{w}, \mathbf{x}_k, \mathbf{x}_b)], & \text{when } N < k \leq 2N, 1 \leq b \leq N. \\
\mathbb{E}_{\mathbf{w} \sim N(\mathbf{0},\mathbf{I})} [\mathbb{I}_{\{\mathbf{w}\mathbf{x}_k^\top \geq 0,\ \mathbf{w}\mathbf{x}_b^\top \geq 0\}} h_{22}(\mathbf{w}, \mathbf{x}_k, \mathbf{x}_b)], & \text{when } N < k \leq 2N, N < b \leq 2N.
\end{cases} \tag{14}
$$

From 8, we can already see $\|\mathbf{u}(t) - \mathbf{y}\|^2$ will decrease with time since 8 is a symmetry matrix, thus all its eigenvalues are non negative real numbers.

Lemma A.3. $\|\mathbf{u}(t) - \mathbf{y}\|^2$ *is a non-increasing function of time.*

Proof. In 7, we have

$$\frac{d\|\mathbf{u}(t) - \mathbf{y}\|^2}{dt} = -16(\mathbf{u}(t) - \mathbf{y}) \cdot \mathbf{M}(t) \cdot (\mathbf{u}(t) - \mathbf{y})^\top$$

where $\mathbf{M}$ is symmetry matrix of form $\mathbf{U}(t) \cdot \mathbf{U}(t)^\top$, where

$$\mathbf{U}(t) = \begin{bmatrix} u_{1,1}(t) & & & & & \\ & \ddots & & & & \\ & & u_{1,N}(t) & & & \\ & & & u_{2,1}(t) & & \\ & & & & \ddots & \\ & & & & & u_{2,N}(t) \end{bmatrix} \cdot \mathbf{S}.$$

Thus all eigenvalues of M is non-negative, and we have

$$\frac{d\|\mathbf{u}(t) - \mathbf{y}\|^2}{dt} \leq 0.$$

□

Corollary A.4. *During the training process,* $\|\mathbf{w}(t)\|$ *is a bounded function.*

Proof. If $\|\mathbf{w}(t)\|$ is unbounded, then $\mathbf{u}(t)$ will also be unbounded, that is a contradiction with the fact $\|\mathbf{u}(t) - \mathbf{y}\|^2$ is a bounded function from proposition A.3. □

Lemma A.5. *For convenience, when $N < k, b \leq 2N$, we will use b and k to represent $b - N$ and $k - N$, this will not make confuse because k, b are indexes in $1, ..., N$. Then we have*

$$\mathbf{G}_{k,b}(t) = \begin{cases} \frac{4}{m}\sum_{i=1}^m a_i^2 \mathbb{I}_{\{\mathbf{w}_i(t)\mathbf{x}_k^\top \geq 0,\ \mathbf{w}_i(t)\mathbf{x}_b^\top \geq 0\}} h_{11}(\mathbf{w}_i(t), \mathbf{x}_k, \mathbf{x}_b), & \textit{when } 1 \leq k \leq N, 1 \leq b \leq N. \\ \frac{4}{m}\sum_{i=1}^m a_i^2 \mathbb{I}_{\{\mathbf{w}_i(t)\mathbf{x}_k^\top \geq 0,\ \mathbf{w}_i(t)\mathbf{x}_b^\top \geq 0\}} h_{12}(\mathbf{w}_i(t), \mathbf{x}_k, \mathbf{x}_b), & \textit{when } 1 \leq k \leq N, N < b \leq 2N. \\ \frac{4}{m}\sum_{i=1}^m a_i^2 \mathbb{I}_{\{\mathbf{w}_i(t)\mathbf{x}_k^\top \geq 0,\ \mathbf{w}_i(t)\mathbf{x}_b^\top \geq 0\}} h_{21}(\mathbf{w}_i(t), \mathbf{x}_k, \mathbf{x}_b), & \textit{when } N < k \leq 2N, 1 \leq b \leq N. \\ \frac{4}{m}\sum_{i=1}^m a_i^2 \mathbb{I}_{\{\mathbf{w}_i(t)\mathbf{x}_k^\top \geq 0,\ \mathbf{w}_i(t)\mathbf{x}_b^\top \geq 0\}} h_{22}(\mathbf{w}_i(t), \mathbf{x}_k, \mathbf{x}_b), & \textit{when } N < k \leq 2N, N < b \leq 2N. \end{cases} \tag{15}$$

where

$$h_{11}(\mathbf{w}_i, \mathbf{x}_k, \mathbf{x}_b) = (\mathbf{w}_i\mathbf{x}_k^\top + w_{i,1}x_{k,1})(\mathbf{w}_i\mathbf{x}_b^\top + w_{i,1}x_{b,1}) + w_{i,1}^2 x_{k,2}x_{b,2}, \tag{16}$$

$$h_{12}(\mathbf{w}_i, \mathbf{x}_k, \mathbf{x}_b) = (\mathbf{w}_i\mathbf{x}_k^\top + w_{i,1}x_{k,1})(w_{i,2}x_{b,1}) + (\mathbf{w}_i\mathbf{x}_b^\top + w_{i,2}x_{b,2})(w_{i,1}x_{k,2}), \tag{17}$$

$$h_{21}(\mathbf{w}_i, \mathbf{x}_k, \mathbf{x}_b) = (\mathbf{w}_i\mathbf{x}_k^\top + w_{i,1}x_{b,1})(w_{i,2}x_{k,1}) + (\mathbf{w}_i\mathbf{x}_k^\top + w_{i,2}x_{k,2})(w_{i,1}x_{b,2}), \tag{18}$$

$$h_{22}(\mathbf{w}_i, \mathbf{x}_k, \mathbf{x}_b) = (\mathbf{w}_i\mathbf{x}_k^\top + w_{i,2}x_{k,2})(\mathbf{w}_i\mathbf{x}_b^\top + w_{i,2}x_{b,2}) + w_{i,2}^2 x_{k,1}x_{b,1}. \tag{19}$$

Remark A.6. *Since during the training process, a_i is always a constant, thus we will assume $a_i = 1$ for $i = 1, ..., m$ form now on.*

If at time 0, $w_i \in \mathbb{R}^2$ is chosen from a $N(\mathbf{0}, \mathbf{I})$ distribution independently, then $\mathbf{G}(0)$ is an average and $\mathbf{G}^\infty$ is an expectation.

Our aim is to prove $\mathbf{G}^\infty$ is a positive definite matrix.

Definition A.7. $\phi_k(\mathbf{w}) = \begin{cases} \mathbb{I}_{\{\mathbf{w}\cdot\mathbf{x}_k^\top \geq 0\}}(\mathbf{w}\cdot\mathbf{x}_k^\top + w_1 x_{k,1}, w_1 x_{k,2}), & \textit{when } 1 \leq k \leq N\ . \\ \\ \mathbb{I}_{\{\mathbf{w}\cdot\mathbf{x}_{k-N}^\top \geq 0\}}(w_2 \cdot x_{k,1}, \mathbf{w}\cdot\mathbf{x}_k^\top + w_2 x_{k,2}), & \textit{when } N < k \leq 2N. \end{cases}$

$\mathbf{G}^\infty_{i,j} = \langle \phi_i(\mathbf{w}), \phi_j(\mathbf{w})\rangle$, where the inner product $\langle\cdot,\cdot\rangle$ is with respect to the expectation to $\mathbf{w}$, thus $\mathbf{G}^\infty$ is a Gram matrix.

Proposition A.8. $\mathbf{G}^\infty$ *is positive definite, and we denote the smallest eigenvalue of* $\mathbf{G}^\infty$ *by* λ_0*, thus* $\lambda_0 > 0$.

Proof. Since a Gram matrix is positive definite if and only if it is constructed from a set of linear independent vectors, thus we only need to proof $\{\phi_s\}_{s=1}^{2N}$ is a linearly independent set, that is, if

$$\sum_{s=1}^N a_s\phi_s + \sum_{s=1}^N b_{N+s}\phi_{N+s} = 0, \tag{20}$$

then

$$a_s, b_{N+s} = 0. \tag{21}$$

Fix an i and let $D_i = \{\mathbf{w}\in\mathbb{R}^2 | \mathbf{w}\cdot\mathbf{x}_i = 0\}$ for $i = 1, 2, ..., N$. Then we have $D_i \nsubseteq \cup_{j\neq i} D_j$, and we can choose a $\mathbf{z}\in D_i - \cup_{j\neq i} D_j$ and a $r > 0$ such that

$$B(\mathbf{z}, r) \cap D_j = \emptyset \tag{22}$$

for $j \neq i$.

Denote $B_r^+ = B(\mathbf{z}, r) \cap \{\mathbf{w}\in\mathbb{R}^2 | \mathbf{w}\cdot\mathbf{x}_i \geq 0\}$ and $B_r^- = B(\mathbf{z}, r) - B_r^+$.

Now consider the integral

$$\frac{1}{\mu(B_r^+)}\int_{B_r^+}\phi_j(\mathbf{w})d\mathbf{w} - \frac{1}{\mu(B_r^-)}\int_{B_r^-}\phi_j(\mathbf{w})d\mathbf{w}$$

when $r \to 0$ and $j \neq i$ this equals to 0 since both of these two terms are equal to $\phi_j(\mathbf{z})$.

Thus we have

$$0 = \lim_{r\to 0}\frac{1}{\mu(B_r)}\int_{B_r}\left(\sum_{s=1}^N a_s\phi_s(\mathbf{w}) + \sum_{s=1}^N b_{N+s}\phi_{s+N}(\mathbf{w})\right)d\mathbf{w} \tag{23}$$

$$= \lim_{r\to 0}\frac{1}{\mu(B_r^+)}\int_{B_r^+}\left(\sum_{s=1}^N a_s\phi_s(\mathbf{w}) + \sum_{s=1}^N b_{N+s}\phi_{s+N}(\mathbf{w})\right)d\mathbf{w} - \lim_{r\to 0}\frac{1}{\mu(B_r^-)}\int_{B_r^-}\left(\sum_{s=1}^N a_s\phi_s(\mathbf{w}) + \sum_{s=1}^N b_{N+s}\phi_{s+N}(\mathbf{w})\right)d\mathbf{w} \tag{24}$$

$$= \lim_{r\to 0}\frac{1}{\mu(B_r^+)}\int_{B_r^+} a_i\phi_i(\mathbf{w}) + b_{N+i}\phi_{i+N}(\mathbf{w})d\mathbf{w} - \lim_{r\to 0}\frac{1}{\mu(B_r^-)}\int_{B_r^-} a_i\phi_i(\mathbf{w}) + b_{N+i}\phi_{N+i}(\mathbf{w})d\mathbf{w} \tag{25}$$

$$= \lim_{r\to 0}\frac{1}{\mu(B_r^+)}\int_{B_r^+} a_i\phi_i(\mathbf{w}) + b_{N+i}\phi_{N+i}(\mathbf{w})d\mathbf{w} \tag{26}$$

where the last equality is because in B_r^-, $\phi_i(\mathbf{w})$ and $\phi_{N+i}(\mathbf{w})$ are equal to 0.

Since

$$\lim_{r\to 0}\frac{1}{\mu(B_r^+)}\int_{B_r^+} a_i\phi_i(\mathbf{w}) + b_{N+i}\phi_{N+i}(\mathbf{w})d\mathbf{w} = a_i\phi_i(\mathbf{z}) + b_{N+i}\phi_{N+i}(\mathbf{z}), \tag{27}$$

and

$$\phi_i(\mathbf{z}) = z_1\mathbf{x}_i^\top, \phi_{N+i}(\mathbf{z}) = z_2\mathbf{x}_i^\top, \tag{28}$$

thus

$$0 = \lim_{r \to 0} \frac{1}{\mu(B_r^+)} \int_{B_r^+} a_i \phi_i(\mathbf{w}) + b_{N+i}\phi_{N+i}(\mathbf{w}) d\mathbf{w} \tag{29}$$

$$= (a_i z_1 + b_{N+i} z_2)\mathbf{x}_i^\top. \tag{30}$$

which means $a_i z_1 + b_{N+i} z_2 = 0$. If $a_i, b_{N+i} \neq 0$, then we have $(a_i, b_{N+i})//\mathbf{x}_i \in \mathbb{R}^2$.

From above we have shown if

$$\sum_{s=1}^N a_s \phi_s + \sum_{s=1}^N b_{N+s}\phi_{N+s} = 0,$$

then for each i, whether $(a_i, b_{N+i}) = 0$ or $(a_i, b_{N+i})//\mathbf{x}_i$. We want to proof $(a_i, b_{N+i})//\mathbf{x}_i$ is impossible.

Assume $(a_i, b_{N+1}) = k_i \mathbf{x}_i$ for some constant k_i, and if $(a_i, b_{N+1}) = 0$ then $k_i = 0$.

Thus since $\sum_{s=1}^N a_s \phi_s + \sum_{s=1}^N b_{N+s}\phi_{N+s} = 0$, we have

$$\sum_{s=1}^N k_s x_{s,1}\phi_s(\mathbf{w}) + \sum_{s=1}^N k_s x_{s,2}\phi_{N+s}(\mathbf{w}) = 0 \tag{31}$$

for almost all $\mathbf{w} \in \mathbb{R}^2$. Our aim is to show $k_s = 0$ for $s = 1, 2, ..., N$.

Write down the definition of $\phi_s(\mathbf{w}), \phi_{N+s}(\mathbf{w})$, we have

$$\sum_{s=1}^N k_s x_{s,1}\phi_s(\mathbf{w}) + \sum_{s=1}^N k_i x_{s,2}\phi_{N+s}(\mathbf{w}) = \tag{32}$$

$$\sum_{s=1}^N k_s x_{s,1}\mathbb{I}_{\{\mathbf{w}\mathbf{x}_i^\top \geq 0\}}(\mathbf{w}\mathbf{x}_i^\top + w_1 x_{i,1}, w_1 x_{i,2}) + \sum_{s=1}^N k_s x_{s,2}\mathbb{I}_{\{\mathbf{w}\mathbf{x}_i^\top \geq 0\}}(w_2 x_{i,1}, \mathbf{w}\mathbf{x}_i^\top + w_2 x_{i,2}) \tag{33}$$

$$= (\sum_{i=1}^N k_i \mathbb{I}_{\{\mathbf{w}\mathbf{x}_i^\top \geq 0\}} \left(x_{i,1}(\mathbf{w}\mathbf{x}_i^\top + w_1 x_{i,1}) + x_{i,2} w_2 x_{i,1}\right), \sum_{i=1}^N k_i \mathbb{I}_{\{\mathbf{w}\mathbf{x}_i^\top \geq 0\}} \left(x_{i,2}(\mathbf{w}\mathbf{x}_i^\top + w_2 x_{i,2}) + x_{i,2} w_1 x_{i,1}\right)) \tag{34}$$

$$= 0 \tag{35}$$

for almost all $\mathbf{w}$.

Note that for every w, above equality give us a linear equation system for $\{k_1, ..., k_N\}$. Since this should hold for almost all w, this will make $k_i = 0$ for $i = 1, 2, ..., N$. □

$\mathbf{G}^\infty$ and $\mathbf{G}(0)$ is close with over-parametrization. The next fact which is essential in understanding the training dynamics is the perturbation of $\mathbf{G}(0)$ from $\mathbb{G}^\infty$. We start the argument with the following proposition.

Proposition A.9. *If $m \geq \frac{16N^2\ell^2 \ln(\frac{1}{\delta})}{\lambda_0^2}$ for some uniform constant ℓ and $\|\mathbf{x}_i\| = 1$ for all N sample points, then with probability larger than $1 - \delta$, we have*

$$\|\mathbf{G}^\infty - \mathbf{G}(0)\|_2 \leq \frac{\lambda_0}{4},$$

where $\lambda_0 > 0$ is the smallest eigenvalue of $\mathbf{G}^\infty$.

Proof. Firstly, we have

$$\|\mathbf{G}^\infty - \mathbf{G}(0)\|_2^2 \leq \|\mathbf{G}^\infty - \mathbf{G}(0)\|_F^2 = \sum_{k,b} |\mathbf{G}_{k,b}^\infty - \mathbf{G}_{k,b}(0)|^2. \tag{36}$$

Since $\mathbf{G}_{k,b}(0)$ is an average of samples $\{\mathbf{w}_i(0)\}_i$ and $G_{k,b}^\infty$ is the expectation $\{\mathbf{w}_i\}_i$ (see (15) and (14)), we will use Hoeffding's inequality to compare $G_{k,b}^\infty$ and $\mathbf{G}_{k,b}(0)$.

Let $\mathbf{w} \sim N(\mathbf{0}, \mathbf{I})$, and denote the random variable $\mathbb{I}_{\{\mathbf{w}\mathbf{x}_k^\top \geq 0, \mathbf{w}\mathbf{x}_b^\top \geq 0\}} h(\mathbf{w}, \mathbf{x}_k, \mathbf{x}_b)$ by $\mathbf{X}$, then we claim $\mathbf{X}$ satisfies a sub-gaussian distribution, that is,

$$\mathbb{P}(|\mathbf{X}|\geq t) \leq e^{-\alpha t^2}$$

for some constant α.

We prove this claim in the case $h = h_{11}$, other cases are similar. Since we have

$$|\mathbb{I}_{\{\mathbf{w}\mathbf{x}_k \geq 0, \mathbf{w}\mathbf{x}_b \geq 0\}} h(\mathbf{w}, \mathbf{x}_k, \mathbf{x}_b)|\geq t \Leftrightarrow |(\mathbf{w}\mathbf{x}_k + w_1 x_{k,1})(\mathbf{w}\mathbf{x}_b + w_1 x_{b,1}) + w_1^2 x_{k,2} x_{b,2}|\geq t, \tag{37}$$

and $\|x_i\|\leq 1$, thus this can be bounded by the norm of w, and a calculate shows

$$|(\mathbf{w}\mathbf{x}_k^\top + w_1 x_{k,1})(\mathbf{w}\mathbf{x}_b^\top + w_1 x_{b,1}) + w_1^2 x_{k,2} x_{b,2}|\leq 5\|\mathbf{w}\|^2. \tag{38}$$

Thus we have $\mathbb{P}(|\mathbf{X}|\geq t) \leq \mathbb{P}(5\|\mathbf{w}\|^2\geq t) \leq e^{-kt^2}$ since $\mathbf{w} \sim N(\mathbf{0}, \mathbf{I})$.

Then by the Hoeffding's inequality for sub-gaussian distribution, we have

$$\mathbb{P}(|G_{k,b}^\infty - G_{k,b}(0)|\geq t) \leq e^{-\frac{mt^2}{\ell^2}}$$

where ℓ is the sub-gaussian norm of $G_{k,b}(0)$. Then if we choose $m \geq \frac{16N^2\ell^2 \ln(\frac{1}{\delta})}{\lambda_0^2}$, we will get with at least probability $1-\delta$

$$\|\mathbf{G}^\infty - \mathbf{G}(0)\|_2\leq \frac{\lambda_0}{4}$$

□

The difference between $\mathbf{G}(0)$ and $\mathbf{G}(t)$ is characterized in the following way.

Proposition A.10. *Let* $\mathbf{w}_1(0), ..., \mathbf{w}_m(0) \sim N(\mathbf{0}, \mathbf{I})$*, then if* $\|\mathbf{w}_i(0) - \mathbf{w}_i(t)\|\leq R = (\frac{\delta\lambda_0}{16N^2c})^2$ *for some constant c, then with probability* $1-\delta$*, we have*

$$\|\mathbf{G}(0) - \mathbf{G}(t)\|_2\leq \frac{\lambda_0}{4}$$

Proof.

$$\mathbb{E}[|\mathbf{G}_{k,b}(0) - \mathbf{G}_{k,b}(t)|] \tag{39}$$

$$= \mathbb{E}[\frac{1}{m}|\sum_{i=1}^m \mathbb{I}_{\{\mathbf{w}_i(0)\mathbf{x}_k^\top, \mathbf{w}_i(0)\mathbf{x}_b^\top \geq 0\}} h(\mathbf{w}_i(0), \mathbf{x}_k^\top, \mathbf{x}_b^\top) - \mathbb{I}_{\{\mathbf{w}_i(t)\mathbf{x}_k^\top, \mathbf{w}_i(t)\mathbf{x}_b^\top \geq 0\}} h(\mathbf{w}_i(t), \mathbf{x}_k, \mathbf{x}_b)|] \tag{40}$$

$$= \mathbb{E}[\frac{1}{m}|\sum_{i=1}^m \mathbb{I}_{\{i,k,b,0\}} h_{i,k,b}(0) - \mathbb{I}_{\{i,k,b,t\}} h_{i,k,b}(t)|] \tag{41}$$

$$\leq \mathbb{E}[\frac{1}{m}\sum_{i=1}^m |\mathbb{I}_{\{i,k,b,0\}} h_{i,k,b}(0) - \mathbb{I}_{\{i,k,b,t\}} h_{i,k,b}(t)|] \tag{42}$$

$$= \mathbb{E}[\frac{1}{m}\sum_{i=1}^m |\mathbb{I}_{\{i,k,b,0\}} h_{i,k,b}(0) - \mathbb{I}_{\{i,k,b,t\}} h_{i,k,b}(0) + \mathbb{I}_{\{i,k,b,t\}} h_{i,k,b}(0) - \mathbb{I}_{\{i,k,b,t\}} h_{i,k,b}(t)|] \tag{43}$$

$$\leq \frac{1}{m}\sum_{i=1}^m \mathbb{E}[|(\mathbb{I}_{\{i,k,b,0\}} - \mathbb{I}_{\{i,k,b,t\}}) h_{i,k,b}(0)|] + \frac{1}{m}\sum_{i=1}^m \mathbb{E}(|\mathbb{I}_{\{i,k,b,t\}}(h_{i,k,b}(0) - h_{i,k,b}(t))|] \tag{44}$$

By Cauchy-Schwartz inequality, we have

$$(44) \leq \frac{1}{m}\sum_{i=1}^m \left(\mathbb{E}[\mathbb{I}_{\{i,k,b,0\}} - \mathbb{I}_{\{i,k,b,t\}}]^2\right)^{\frac{1}{2}} \left(\mathbb{E}[h_{i,k,b}(0)^2]\right)^{\frac{1}{2}} + \frac{1}{m}\sum_{i=1}^m \left(\mathbb{E}[h_{i,k,b}(0) - h_{i,k,b}(t)]^2\right)^{\frac{1}{2}} \left(\mathbb{E}[\mathbb{I}_{\{i,k,b,t\}}^2]\right)^{\frac{1}{2}}. \tag{45}$$

Since $\mathbf{w} \sim N(\mathbf{0}, \mathbf{I})$ and $\|\mathbf{x}\|$ are bounded, the terms $\left(\mathbb{E}[h_{i,k,b}(0)^2]\right)^{\frac{1}{2}} \leq \alpha$ for some constant α, which only relies on the expectation of w and bound of $\|x\|$, thus independent of N. We also have $\left(\mathbb{E}(\mathbb{I}^2_{\{i,k,b,t\}})\right)^{\frac{1}{2}} \leq 1$. Thus we have

$$(45) \leq \frac{\alpha}{m}\sum_{i=1}^{m}\left(\mathbb{E}[\mathbb{I}_{\{i,k,b,0\}} - \mathbb{I}_{\{i,k,b,t\}}]^2\right)^{\frac{1}{2}} + \frac{1}{m}\sum_{i=1}^{m}\left(\mathbb{E}[h_{i,k,b}(0) - h_{i,k,b}(t)]^2\right)^{\frac{1}{2}}. \tag{46}$$

We firstly consider the term $\frac{\alpha}{m}\sum_{i=1}^{m}[\mathbb{E}(\mathbb{I}_{\{i,k,b,0\}} - \mathbb{I}_{\{i,k,b,t\}})^2]^{\frac{1}{2}}$. Let $R \in \mathbb{R}$ and $A_{i,k}$ be an event defined by

$$A_{i,k} = \{\mathbf{w}_i \mid \|\mathbf{w}_i - \mathbf{w}_i(0)\| < R, \mathbb{I}_{\{\mathbf{w}_i\mathbf{x}_k^\top \geq 0\}} \neq \mathbb{I}_{\{\mathbf{w}_i(0)\mathbf{x}_k^\top \geq 0\}}\}. \tag{47}$$

Note that $A_{i,k}$ happens if and only if $|\mathbf{w}_i(0)\mathbf{x}_k^\top| < R$, and since $\mathbf{w}_i(0) \sim N(\mathbf{0}, \mathbf{I}),\ \|x_k\| = 1$, by anti-concentration inequality of Gaussian distribution we have

$$\mathbb{P}(A_{i,k}) \leq \frac{2R}{\sqrt{2\pi}}. \tag{48}$$

Then we have

$$\frac{\alpha}{m}\sum_{i=1}^{m}\left(\mathbb{E}[\mathbb{I}_{\{i,k,b,0\}} - \mathbb{I}_{\{i,k,b,t\}}]^2\right)^{\frac{1}{2}} = \frac{\alpha}{m}\sum_{i=1}^{m}\left(\mathbb{E}[\ |\mathbb{I}_{\{i,k,b,0\}} - \mathbb{I}_{\{i,k,b,t\}}|\ ]\right)^{\frac{1}{2}} \tag{49}$$

$$= \frac{\alpha}{m}\sum_{i=1}^{m}\left(\mathbb{E}[\mathbb{I}_{\{A_{i,k} \cup A_{i,b}\}}]\right)^{\frac{1}{2}} \tag{50}$$

$$\leq \frac{2\alpha}{\sqrt[4]{2\pi}}\sqrt{R}. \tag{51}$$

For the term $\frac{1}{m}\sum_{i=1}^{m}[\mathbb{E}(h_{i,k,b}(0) - h_{i,k,b}(t))^2]^{\frac{1}{2}}$, we have

$$(h_{i,k,b}(0) - h_{i,k,b}(t))^2 \leq CR^2(\|\mathbf{w}_i(0)\| + \|\mathbf{w}_i(t)\|)^2 \tag{52}$$

$$\leq CR^2(2\|\mathbf{w}_i(0)\| + R)^2, \tag{53}$$

for some constant C. Thus we have

$$(\mathbb{E}[h_{i,k,b}(0) - h_{i,k,b}(t)]^2)^{\frac{1}{2}} \leq \sqrt{C}R\sqrt{\mathbb{E}[(2\|\mathbf{w}_i(0)\| + R)^2]}. \tag{54}$$

Since we assume $R \ll 1$, thus combine (51) and (54) we have

$$\mathbb{E}[|\mathbf{G}_{k,b}(0) - \mathbf{G}_{k,b}(t)|] \leq \frac{2\alpha}{\sqrt[4]{2\pi}}\sqrt{R} + \sqrt{C}R\sqrt{\mathbb{E}[(2\|w_i(0)\| + R)^2]} \leq c\sqrt{R}$$

for some constant c. By Markov inequality, we have with probability $1 - \delta$

$$\|\mathbf{G}(0) - \mathbf{G}(t)\|_2 \leq \sum_{k,b=1}^{2N}|\mathbf{G}_{k,b}(0) - \mathbf{G}_{k,b}(t)| \leq \frac{4N^2c\sqrt{R}}{\delta}$$

Thus if $R = (\frac{\delta\lambda_0}{16N^2c})^2$, we will have

$$\|\mathbf{G}(0) - \mathbf{G}(t)\|_2 \leq \frac{\lambda_0}{4}. \tag{55}$$

□

Now the only question is : under what condition, during the training process $\|\mathbf{w}_i(0) - \mathbf{w}_i(t)\|$ can satisfy the condition given in proposition A.10? We will show this can be done by choose the number of neurons sufficient large and this will provide an exponentially convergence rate of the distance the distance between values of neural network and zero of the loss function $\|\mathbf{u}(t) - \mathbf{y}\|$.

Proposition A.11. *For $m \geq \frac{LN^8}{\delta^4\lambda_0^6}$ where constant L is independent of N, δ and λ_0, we have $\|\mathbf{w}_i(0) - \mathbf{w}_i(t)\|$ satisfies the condition given in proposition A.10.*

Proof. We only need to prove the last part of the statement.

$$\|\frac{d\mathbf{w}_i(s)}{ds}\| = \|\nabla_{\mathbf{w}_i} L(\mathbf{w})\| \tag{56}$$

$$\leq 4\sum_{j=1}^{N} |(u_{1,j})^2 - y_{1,j}| \cdot |u_{1,j}| \cdot \|\frac{\partial u_{1,j}}{\partial \mathbf{w}_i}\| + 4\sum_{j=1}^{N} |(u_{2,j})^2 - y_{2,j}| \cdot |u_{2,j}| \cdot \|\frac{\partial u_{2,j}}{\partial \mathbf{w}_i}\| \tag{57}$$

Here $u_{1,j}$ and $u_{2,j}$ denote their value at time s.

Since we assume $\{u_{1,j}\}_j, \{u_{2,j}\}_j$ lie in a neighbourhood of $\mathbf{y}$ and they are bounded by some constant M, thus

$$(57) \leq 4M\sum_{j=1}^{N}(|(u_{1,j})^2 - y_{1,j}| \cdot \|\frac{\partial u_{1,j}}{\partial \mathbf{w}_i}\| + |(u_{2,j})^2 - y_{2,j}| \cdot \|\frac{\partial u_{2,j}}{\partial \mathbf{w}_i}\|). \tag{58}$$

Since we have

$$\frac{\partial u_{1,j}}{\partial \mathbf{w}_i} = \frac{2}{\sqrt{m}} a_r \mathbb{I}_{\{\mathbf{w}_i \mathbf{x}_j^\top \geq 0\}} (\mathbf{w}_i \mathbf{x}_j^\top + w_{i,1}x_{j,1} \ , \ w_{i,1}x_{j,2}) \tag{59}$$

$$\frac{\partial u_{2,j}}{\partial \mathbf{w}_i} = \frac{2}{\sqrt{m}} a_r \mathbb{I}_{\{\mathbf{w}_i \mathbf{x}_j^\top \geq 0\}} (w_{i,2}x_{j,1} \ , \ \mathbf{w}_i \cdot \mathbf{x}_j^\top + w_{i,2}x_{j,2}) \tag{60}$$

and there exists a constant C makes $\|\mathbf{w}_i(t)\| \leq C$ due to corollary A.4, thus there exists $\tilde{C}$ such that

$$\frac{\partial u_{1,j}}{\partial \mathbf{w}_i} \leq \frac{\tilde{C}}{\sqrt{m}}, \ \frac{\partial u_{2,j}}{\partial \mathbf{w}_r} \leq \frac{\tilde{C}}{\sqrt{m}}. \tag{61}$$

This gives

$$(58) \leq 4M\frac{\tilde{C}}{\sqrt{m}}\sum_{j=1}^{N}(|(u_{1,j})^2 - y_{1,j}| + |(u_{2,j})^2 - y_{2,j}|). \tag{62}$$

And since

$$\sum_{j=1}^{N}(|(u_{1,j}(s))^2 - y_{1,j}| + |(u_{2,j}(s))^2 - y_{2,j}|) \leq e^{-\lambda_0 s}\sum_{j=1}^{N}(|(u_{1,j}(0))^2 - y_{1,j}| + |(u_{2,j}(0))^2 - y_{2,j}|) \tag{63}$$

$$\leq e^{-\lambda_0 s}\|\mathbf{u}(0) - \mathbf{y}\|. \tag{64}$$

Thus combine above we have

$$\|\frac{d\mathbf{w}_i(s)}{ds}\| \leq \frac{4M\tilde{C}}{\sqrt{m}} e^{-\lambda_0 s}\|\mathbf{u}(0) - \mathbf{y}\|$$

Do an integral, we have $\|\mathbf{w}_i(0) - \mathbf{w}_i(t)\| \leq \frac{4M\tilde{C}}{\sqrt{m}\lambda_0}\|\mathbf{u}(0) - \mathbf{y}\| \leq \frac{\tilde{M}}{\sqrt{m}\lambda_0}$ Thus choose $m \geq \frac{LN^8}{\delta^4\lambda_0^6}$ for some constant L independent of N, δ, λ_0 will makes $\|w_r(0) - w_r(t)\|$ satisfies condition in proposition A.10. □

B. Proof of Theorem 3.1

The proof is a direct consequence of that gradient descent optimizes over-parametrized RePU neural networks. We focus on the shallow neural network as follows,

$$f(\mathbf{W}, \mathbf{a}, \mathbf{x}) = \frac{1}{\sqrt{m}} \sum_{k=1}^{m} a_k \sigma(\mathbf{w}_k \mathbf{x}),$$

then the partial derivative with respect to $\mathbf{w}_r$ is

$$\frac{\partial f}{\partial \mathbf{w}_r} = \frac{1}{\sqrt{m}} \sum_{k=1}^{m} a_k \frac{\partial \sigma(\mathbf{w}_k \mathbf{x})}{\partial \mathbf{w}_r} = \frac{1}{\sqrt{m}} a_r \frac{\partial \sigma(\mathbf{w}_r \mathbf{x})}{\partial \mathbf{w}_r}.$$

The loss function with training set $\{\mathbf{x}_i, y_i\}_{i=1}^n$ is defined to be

$$L(\mathbf{W}, \mathbf{a}) = \sum_{i=1}^{n} \frac{1}{2} \left(f(\mathbf{W}, \mathbf{a}, \mathbf{x}_i) - y_i\right)^2,$$

and the partial derivative with respect to $\mathbf{W}$ is

$$\frac{\partial L}{\partial \mathbf{W}} = \sum_{i=1}^{n} (f(\mathbf{W}, \mathbf{a}, \mathbf{x}_i) - y_i) \frac{\partial f}{\partial \mathbf{W}} = \sum_{i=1}^{n} (f(\mathbf{W}, \mathbf{a}, \mathbf{x}_i) - y_i) \sum_{r=1}^{m} \frac{\partial f}{\partial \mathbf{w}_r}.$$

For the vector $\mathbf{w}_r$, we can compute the partial derivatives of L

$$\frac{\partial L}{\partial \mathbf{w}_r} = \sum_{i=1}^{n} \frac{\partial}{\partial \mathbf{w}_r} \frac{1}{2} (f(\mathbf{W}, \mathbf{a}, \mathbf{x}_i) - y_i)^2 \tag{65}$$

$$= \sum_{i=1}^{n} (f(\mathbf{W}, \mathbf{a}, \mathbf{x}_i) - y_i) \frac{\partial}{\partial \mathbf{w}_r} f(\mathbf{W}, \mathbf{a}, \mathbf{x}_i), \tag{66}$$

where

$$\frac{\partial}{\partial \mathbf{w}_r} f(\mathbf{W}, \mathbf{a}, \mathbf{x}_i) = \frac{1}{\sqrt{m}} a_r \frac{\partial}{\partial \mathbf{w}_r} \sigma(\mathbf{w}_r \mathbf{x}_i).$$

If we denote $\zeta(\cdot) = p\text{RePU}^{p-1}(\cdot)$, the partials can be denoted as

$$\frac{\partial}{\partial \mathbf{w}_r} f(\mathbf{W}, \mathbf{a}, \mathbf{x}_i) = \frac{1}{\sqrt{m}} a_r \zeta(\mathbf{w}_r \mathbf{x}_i) \mathbf{x}_i.$$

Thus

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

Next we compute the ode of evolution for predictions,

$$u_i(t) = f(\mathbf{W}, \mathbf{a}, \mathbf{x}_i),$$

and then

$$\frac{du_i}{dt} = \sum_{r=1}^{m} \langle \frac{\partial f(\mathbf{x}_i)}{\partial \mathbf{w}_r}, \frac{d\mathbf{w}_r}{dt} \rangle \tag{67}$$

$$= \sum_{r=1}^{m} \langle \frac{\partial f(\mathbf{x}_i)}{\partial \mathbf{w}_r}, -\frac{\partial L}{\partial \mathbf{w}_r} \rangle, \tag{68}$$

where

$$\langle \frac{\partial f(\mathbf{x}_i)}{\partial \mathbf{w}_r}, -\frac{\partial L}{\partial \mathbf{w}_r} \rangle = \langle \frac{\partial f(\mathbf{x}_i)}{\partial \mathbf{w}_r}, -\sum_{j=1}^{n} (u_j - y_j) \frac{\partial f(\mathbf{x}_j)}{\partial \mathbf{w}_r} \rangle \tag{69}$$

$$= \sum_{j=1}^{n} (y_j - u_j) \langle \frac{\partial f(\mathbf{x}_i)}{\partial \mathbf{w}_r}, \frac{\partial f(\mathbf{x}_j)}{\partial \mathbf{w}_r} \rangle. \tag{70}$$

So

$$\frac{du_i}{dt} = \sum_{r=1}^{m}\sum_{j=1}^{n}(y_j - u_j)\langle\frac{\partial f(\mathbf{x}_i)}{\partial \mathbf{w}_r}, \frac{\partial f(\mathbf{x}_j)}{\partial \mathbf{w}_r}\rangle \tag{71}$$

$$= \sum_{j=1}^{n}(y_j - u_j)\sum_{r=1}^{m}\langle\frac{\partial f(\mathbf{x}_i)}{\partial \mathbf{w}_r}, \frac{\partial f(\mathbf{x}_j)}{\partial \mathbf{w}_r}\rangle \tag{72}$$

$$= \sum_{j=1}^{n}(y_j - u_j)H_{ij}(t) \tag{73}$$

where

$$H_{ij}(t) = \sum_{r=1}^{m}\langle\frac{\partial f(\mathbf{x}_i)}{\partial \mathbf{w}_r}, \frac{\partial f(\mathbf{x}_j)}{\partial \mathbf{w}_r}\rangle$$

and

$$\frac{\partial f(\mathbf{x}_i)}{\partial \mathbf{w}_r} = \frac{1}{\sqrt{m}}a_r\zeta(\mathbf{w}_r\mathbf{x}_i)\mathbf{x}_i \tag{74}$$

$$\frac{\partial f(\mathbf{x}_j)}{\partial \mathbf{w}_r} = \frac{1}{\sqrt{m}}a_r\zeta(\mathbf{w}_r\mathbf{x}_j)\mathbf{x}_j \tag{75}$$

Thus

$$H_{ij}(t) = \sum_{r=1}^{m}\langle\frac{1}{\sqrt{m}}a_r\zeta(\mathbf{w}_r\mathbf{x}_i)\mathbf{x}_i, \frac{1}{\sqrt{m}}a_r\zeta(\mathbf{w}_r\mathbf{x}_j)\mathbf{x}_j\rangle \tag{76}$$

$$= \sum_{r=1}^{m}\frac{1}{m}\langle\mathbf{x}_i, \mathbf{x}_j\rangle a_r^2\zeta(\mathbf{w}_r\mathbf{x}_i)\zeta(\mathbf{w}_r\mathbf{x}_j) \tag{77}$$

$$= \langle\mathbf{x}_i, \mathbf{x}_j\rangle\left(\frac{1}{m}\sum_{r=1}^{m}a_r^2\zeta(\mathbf{w}_r\mathbf{x}_i)\zeta(\mathbf{w}_r\mathbf{x}_j)\right) \tag{78}$$

Without loss of generality, we can assume $a_r = 1$ for all $r \in [m]$, and the matrix H_{ij} becomes

$$H_{ij} = \langle\mathbf{x}_i, \mathbf{x}_j\rangle\left(\frac{1}{m}\sum_{r=1}^{m}\zeta(\mathbf{w}_r\mathbf{x}_i)\zeta(\mathbf{w}_r\mathbf{x}_j)\right)$$

respectively, the corresponding $H_{ij}(t)$ at $t = 0$ is denoted

$$H_{ij}(0) = \langle\mathbf{x}_i, \mathbf{x}_j\rangle\left(\frac{1}{m}\sum_{r=1}^{m}\zeta(\mathbf{w}_r(0)\mathbf{x}_i)\zeta(\mathbf{w}_r(0)\mathbf{x}_j)\right)$$

Before stating our first result, we give the following assumption.

Assumption 1. $\|\mathbf{w}_r\|$ is bounded by κ during the whole training process.

For each fix pair (i, j), $\left|H_{ij}(0) - H_{ij}^{\infty}\right|$ can be bounded using Hoeffding inequality, i.e., with probability at least $1 - \delta'$, we have

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

Furthermore, setting $\delta' = n^2\delta$, we have

$$\left|H_{ij}(0) - H_{ij}^{\infty}\right| \leq \frac{4\sqrt{\log(n/\delta)}}{\sqrt{m}},$$

and then $\|H(0) - H^{\infty}\|_2^2 \leq \frac{16n^2\log(n/\delta)}{m}$. Formally the optimization result for function approximation with RePU activation functions is stated into the following theorem.

Theorem B.1. *Suppose Assumption 1 holds, $\|\mathbf{x}_i\| = 1$ and the sample values y_i are bounded. Then if we set the number of hidden nodes $m = \Omega\left(\frac{n^2 \log(n/\delta)}{\lambda_0^2}\right)$ and initializing $\mathbf{w}_r \sim \mathcal{N}(\mathbf{0}, I)$, $a_r = 1$, then with probability $1-\delta$ over the initialization, we have*

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

B.1. Proof of Theorem B.1

The proof is consisted of a few lemmas.

Lemma B.2. *Let $\mathbf{w}_1, ..., \mathbf{w}_r$ be i.i.d. sample of $\mathcal{N}(\mathbf{0}, I)$, then with probability at least $1 - \delta$, the following holds. For any set of vectors $\mathbf{w}_1, ..., \mathbf{w}_r$ that satisfy for any $r \in [m]$,*

$$\|\mathbf{w}_r(0) - \mathbf{w}_r\|_2 \leq R < \frac{\delta\lambda_0}{4n^2\ell^2 K}$$

for some constant K, then the matrix H defined by

$$H_{ij} = \frac{1}{m}\langle \mathbf{x}_i, \mathbf{x}_j\rangle \sum_{r=1}^{m} \ell^2 (\mathbf{w}_r\mathbf{x}_i)^{\ell-1}(\mathbf{w}_r\mathbf{x}_j)^{\ell-1}\mathbb{I}\{\mathbf{w}_r\mathbf{x}_i \geq 0, \mathbf{w}_r\mathbf{x}_j \geq 0\}$$

satisfies $\|H - H(0)\|_2 < \frac{\lambda_0}{4}$, and furthermore, $\lambda_{\min}(H) > \frac{\lambda_0}{2}$.

Proof.

$$\begin{aligned} |H_{ij}(0) - H_{ij}| = \Bigg| &\frac{1}{m}\langle \mathbf{x}_i, \mathbf{x}_j\rangle \sum_{r=1}^{m} \ell^2 (\mathbf{w}_r(0)\mathbf{x}_i)^{\ell-1}(\mathbf{w}_r(0)\mathbf{x}_j)^{\ell-1}\mathbb{I}\{\mathbf{w}_r(0)\mathbf{x}_i \geq 0, \mathbf{w}_r(0)\mathbf{x}_j \geq 0\} \\ &- \frac{1}{m}\langle \mathbf{x}_i, \mathbf{x}_j\rangle \sum_{r=1}^{m} \ell^2 (\mathbf{w}_r\mathbf{x}_i)^{\ell-1}(\mathbf{w}_r\mathbf{x}_j)^{\ell-1}\mathbb{I}\{\mathbf{w}_r\mathbf{x}_i \geq 0, \mathbf{w}_r\mathbf{x}_j \geq 0\}\Bigg| \end{aligned} \tag{79}$$

Since $\mathbf{x}_i$ are sampled so that $\|\mathbf{x}_i\| \leq 1$, we have

$$|\langle \mathbf{x}_i, \mathbf{x}_j\rangle| \leq \|\mathbf{x}_i\| \cdot \|\mathbf{x}_j\| = 1,$$

and then

$$\begin{aligned} |H_{ij}(0) - H_{ij}| \leq \frac{1}{m}\Bigg| &\sum_{r=1}^{m} \ell^2 (\mathbf{w}_r(0)\mathbf{x}_i)^{\ell-1}(\mathbf{w}_r(0)\mathbf{x}_j)^{\ell-1}\mathbb{I}\{\mathbf{w}_r(0)\mathbf{x}_i \geq 0, \mathbf{w}_r(0)\mathbf{x}_j \geq 0\} \\ &- \sum_{r=1}^{m} \ell^2 (\mathbf{w}_r\mathbf{x}_i)^{\ell-1}(\mathbf{w}_r\mathbf{x}_j)^{\ell-1}\mathbb{I}\{\mathbf{w}_r\mathbf{x}_i \geq 0, \mathbf{w}_r\mathbf{x}_j \geq 0\}\Bigg| \\ \leq \frac{1}{m}\sum_{r=1}^{m} \ell^2 \big| &(\mathbf{w}_r(0)\mathbf{x}_i)^{\ell-1}(\mathbf{w}_r(0)\mathbf{x}_j)^{\ell-1}\mathbb{I}\{\mathbf{w}_r(0)\mathbf{x}_i \geq 0, \mathbf{w}_r(0)\mathbf{x}_j \geq 0\} \\ &-(\mathbf{w}_r\mathbf{x}_i)^{\ell-1}(\mathbf{w}_r\mathbf{x}_j)^{\ell-1}\mathbb{I}\{\mathbf{w}_r\mathbf{x}_i \geq 0, \mathbf{w}_r\mathbf{x}_j \geq 0\}\big| . \end{aligned} \tag{80}$$

For a chosen set of vectors $\mathbf{w}_r(0) \sim \mathcal{N}(\mathbf{0}, \sigma^2 I)$, we take expectation of $\mathbf{w}_r$'s, the expectation is the following

$$\begin{aligned} \mathbb{E}\left[|H_{ij}(0) - H_{ij}|\right] \leq \ell^2 \frac{1}{m}\sum_{r=1}^{m} \mathbb{E}\big| &(\mathbf{w}_r(0)\mathbf{x}_i)^{\ell-1}(\mathbf{w}_r(0)\mathbf{x}_j)^{\ell-1}\mathbb{I}\{\mathbf{w}_r(0)\mathbf{x}_i \geq 0, \mathbf{w}_r(0)\mathbf{x}_j \geq 0\} \\ &-(\mathbf{w}_r\mathbf{x}_i)^{\ell-1}(\mathbf{w}_r\mathbf{x}_j)^{\ell-1}\mathbb{I}\{\mathbf{w}_r\mathbf{x}_i \geq 0, \mathbf{w}_r\mathbf{x}_j\}\big| . \end{aligned} \tag{81}$$

We next focus on the expectation

$$\mathbb{E}\left|(\mathbf{w}_r(0)\mathbf{x}_i)^{\ell-1}(\mathbf{w}_r(0)\mathbf{x}_j)^{\ell-1}\mathbb{I}\{\mathbf{w}_r(0)\mathbf{x}_i \geq 0, \mathbf{w}_r(0)\mathbf{x}_j \geq 0\} - (\mathbf{w}_r\mathbf{x}_i)^{\ell-1}(\mathbf{w}_r\mathbf{x}_j)^{\ell-1}\mathbb{I}\{\mathbf{w}_r\mathbf{x}_i \geq 0, \mathbf{w}_r\mathbf{x}_j \geq 0\}\right|$$

where $\mathbf{w}_r(0) \sim \mathcal{N}(\mathbf{0}, \sigma^2 I)$ and all $\mathbf{w}_r$ satisfying $\|\mathbf{w}_r(0) - \mathbf{w}_r\|_2 \leq R$. The position of $\mathbf{w}_r(0)$ determines the form of the expectation. We introduce some notations.

$$S_i^+ = \{\mathbf{v} : \langle \mathbf{v} - R\frac{\mathbf{x}_i}{\|\mathbf{x}_i\|}, \mathbf{x}_i\rangle > 0\}$$

and

$$S_j^+ = \{\mathbf{v} : \langle \mathbf{v} - R\frac{\mathbf{x}_j}{\|\mathbf{x}_j\|}, \mathbf{x}_j\rangle > 0\},$$

it is obvious that S_i^- and S_j^- are obtained from the set

$$\{\mathbf{v} : \langle \mathbf{v}, \mathbf{x}_i\rangle > 0, \langle \mathbf{v}, \mathbf{x}_j\rangle > 0\}$$

through translation by vectors $R\frac{\mathbf{x}_i}{\|\mathbf{x}_i\|}$ and $R\frac{\mathbf{x}_j}{\|\mathbf{x}_j\|}$, where the length of the translated vectors are R.

Once $\mathbf{w}_r(0) \in S_i^+ \cap S_j^+$, it holds that

$$\mathbb{I}\{\mathbf{w}_r(0)\mathbf{x}_i \geq 0, \mathbf{w}_r(0)\mathbf{x}_j \geq 0\} = 1$$

and

$$\mathbb{I}\{\mathbf{w}_r\mathbf{x}_i \geq 0, \mathbf{w}_r\mathbf{x}_j \geq 0\} = 1$$

for all $\mathbf{w}_r$ satisfying $\|\mathbf{w}_r(0) - \mathbf{w}_r\|_2 \leq R$. As a consequence, the expectation has the following form:

$$\left|(\mathbf{w}_r(0)\mathbf{x}_i)^{\ell-1}(\mathbf{w}_r(0)\mathbf{x}_j)^{\ell-1} - (\mathbf{w}_r\mathbf{x}_i)^{\ell-1}(\mathbf{w}_r\mathbf{x}_j)^{\ell-1}\right| \text{ if } \mathbf{w}_r(0) \in S_i^+ \cap S_j^+.$$

On the other hand, we further introduce two sets:

$$S_i^- = \{\mathbf{v} : \langle \mathbf{v} + R\frac{\mathbf{x}_i}{\|\mathbf{x}_i\|}, \mathbf{x}_i\rangle < 0\}$$

and

$$S_j^- = \{\mathbf{v} : \langle \mathbf{v} + R\frac{\mathbf{x}_j}{\|\mathbf{x}_j\|}, \mathbf{x}_j\rangle < 0\}.$$

Once $\mathbf{w}_r(0) \in S_i^- \cup S_j^-$, it holds that

$$\mathbb{I}\{\mathbf{w}_r(0)\mathbf{x}_i \geq 0, \mathbf{w}_r(0)\mathbf{x}_j \geq 0\} = 0$$

and

$$\mathbb{I}\{\mathbf{w}_r\mathbf{x}_i \geq 0, \mathbf{w}_r\mathbf{x}_j \geq 0\} = 0$$

and then the expression of expectation equals to 0. In the rest of the proof, we will denote the following for convenience:

$$\begin{aligned} F_{ij}(\mathbf{w}_r(0)) \overset{\text{def}}{=} & \big|(\mathbf{w}_r(0)\mathbf{x}_i)^{\ell-1}(\mathbf{w}_r(0)\mathbf{x}_j)^{\ell-1}\mathbb{I}\{\mathbf{w}_r(0)\mathbf{x}_i \geq 0, \mathbf{w}_r(0)\mathbf{x}_j \geq 0\} \\ & -(\mathbf{w}_r\mathbf{x}_i)^{\ell-1}(\mathbf{w}_r\mathbf{x}_j)^{\ell-1}\mathbb{I}\{\mathbf{w}_r\mathbf{x}_i \geq 0, \mathbf{w}_r\mathbf{x}_j \geq 0\}\big| \end{aligned} \tag{82}$$

and let $\rho(\mathbf{w}_r(0))$ be the probability density function of Gaussian variable in $\mathbb{R}^d$ with 0 the mean and $\sigma^2 I$ the covariance matrix.

From the above analysis, we have the following expression for the expectation:

$$\begin{aligned} \mathbb{E}\, & \big|(\mathbf{w}_r(0)\mathbf{x}_i)^{\ell-1}(\mathbf{w}_r(0)\mathbf{x}_j)^{\ell-1}\mathbb{I}\{\mathbf{w}_r(0)\mathbf{x}_i \geq 0, \mathbf{w}_r(0)\mathbf{x}_j \geq 0\} \\ & -(\mathbf{w}_r\mathbf{x}_i)^{\ell-1}(\mathbf{w}_r\mathbf{x}_j)^{\ell-1}\mathbb{I}\{\mathbf{w}_r\mathbf{x}_i \geq 0, \mathbf{w}_r\mathbf{x}_j \geq 0\}\big| \\ & = \mathbb{E}[F_{ij}(\mathbf{w}_r(0))] \\ & = \int_{\mathbb{R}^d} F_{ij}(\mathbf{w}_r(0))\rho(\mathbf{w}_r(0))d\mathbf{w}_r(0) \\ & = \int_{S_i^+ \cap S_j^+} F_{ij}(\mathbf{w}_r(0))\rho(\mathbf{w}_r(0))d\mathbf{w}_r(0) \\ & + \int_{S_i^- \cup S_j^-} F_{ij}(\mathbf{w}_r(0))\rho(\mathbf{w}_r(0))d\mathbf{w}_r(0) \\ & + \int_{\mathbb{R}^d - (S_i^+ \cap S_j^+) - (S_i^- \cup S_j^-)} F_{ij}(\mathbf{w}_r(0))\rho(\mathbf{w}_r(0))d\mathbf{w}_r(0). \end{aligned} \tag{83}$$

Note that the third part tends to 0 if $R \to 0$, and the second part, i.e., the integral over $S_i^- \cup S_j^-$ is identically zero since both indicator functions are zero. Thus we only have to estimate the integral over $S_i^+ \cap S_j^+$. Since the integral is for the expression

$$\left|(\mathbf{w}_r(0)\mathbf{x}_i)^{\ell-1}(\mathbf{w}_r(0)\mathbf{x}_j)^{\ell-1} - (\mathbf{w}_r\mathbf{x}_i)^{\ell-1}(\mathbf{w}_r\mathbf{x}_j)^{\ell-1}\right|$$

can be bounded by function of $\mathbf{w}_r(0)$ and R, we first obtain this bound. Lemma B.3 enables us to perform the estimate on the difference $|F(\mathbf{w}_r(0)) - F(\mathbf{w}_r)|$ where $F(\cdot)$ is following Lemma B.3. To be precise, we have the following:

$$|F(\mathbf{w}_r(0)) - F(\mathbf{w}_r)| = \|\nabla F(\mathbf{w}_r(0) + \boldsymbol{\theta})\| \cdot \|\mathbf{w}_r(0) - \mathbf{w}_r\| \tag{84}$$

where $\boldsymbol{\theta}$ lies on the line segment connecting $\mathbf{w}_r(0)$ and $\mathbf{w}_r$. Use Lemma B.3 and the condition $\|\mathbf{w}_r(0) - \mathbf{w}_r\| \le R$, we have

$$\begin{aligned}
|F(\mathbf{w}_r(0)) - F(\mathbf{w}_r)| &\le C \|\mathbf{w}_r(0) + \boldsymbol{\theta}\|^{2\ell-3} \cdot R && (85)\\
&\le C \cdot 2^{2\ell-4} \left(\|\mathbf{w}_r(0)\|^{2\ell-3} + \|\boldsymbol{\theta}\|^{2\ell-3}\right) \cdot R && (86)\\
&\le C2^{2\ell-4}R\|\mathbf{w}_r(0)\|^{2\ell-3} + C2^{2\ell-4}R\|\boldsymbol{\theta}\|^{2\ell-3} && (87)\\
&\le C2^{2\ell-4}R\|\mathbf{w}_r(0)\|^{2\ell-3} + C2^{2\ell-4}R^{2\ell-2}. && (88)
\end{aligned}$$

The last inequality holds for $\|\boldsymbol{\theta}\| \le R$. The integral of $F_{ij}\rho$ over the set $S_i^+ \cap S_j^+$ becomes

$$\begin{aligned}
&\int_{S_i^+ \cap S_j^+} F_{ij}(\mathbf{w}_r(0))\rho(\mathbf{w}_r(0))d\mathbf{w}_r(0) && (89)\\
&= \int_{S_i^+ \cap S_j^+} |F(\mathbf{w}_r(0)) - F(\mathbf{w}_r)|\,\rho(\mathbf{w}_r(0))d\mathbf{w}_r(0) && (90)\\
&\le \int_{S_i^+ \cap S_j^+} \left(C2^{2\ell-4}R\|\mathbf{w}_r(0)\|^{2\ell-3} + C2^{2\ell-4}R^{2\ell-2}\right)\rho(\mathbf{w}_r(0))d\mathbf{w}_r(0) && (91)\\
&= \int_{S_i^+ \cap S_j^+} C2^{2\ell-4}R\|\mathbf{w}_r(0)\|^{2\ell-3}\rho(\mathbf{w}_r(0))d\mathbf{w}_r(0) && (92)\\
&+ \int_{S_i^+ \cap S_j^+} C2^{2\ell-4}R^{2\ell-2}\rho(\mathbf{w}_r(0))d\mathbf{w}_r(0) && (93)\\
&\le C2^{2\ell-4}R\int_{S_i^+ \cap S_j^+} \|\mathbf{w}_r(0)\|^{2\ell-3}\rho(\mathbf{w}_r(0))d\mathbf{w}_r(0) && (94)\\
&+ \int_{\mathbb{R}^d} C2^{2\ell-4}R^{2\ell-2}\rho(\mathbf{w}_r(0))d\mathbf{w}_r(0) && (95)\\
&= C2^{2\ell-4}R\int_{S_i^+ \cap S_j^+} \|\mathbf{w}_r(0)\|^{2\ell-3}\rho(\mathbf{w}_r(0))d\mathbf{w}_r(0) + C2^{2\ell-4}R^{2\ell-2} && (96)
\end{aligned}$$

It is immediate that the integral over $S_i^+ \cap S_j^+$ is bounded above by $\frac{1}{2}$ of the integral over the whole $\mathbb{R}^d$,

$$\int_{S_i^+ \cap S_j^+} \|\mathbf{w}_r(0)\|^{2\ell-3}\rho(\mathbf{w}_r(0))d\mathbf{w}_r(0) \le \frac{1}{2}\int_{\mathbb{R}^d} \|\mathbf{w}_r(0)\|^{2\ell-3}\rho(\mathbf{w}_r(0))d\mathbf{w}_r(0).$$

Note that the integral on the right hand side can be computed with spherical coordinate, where $\|\mathbf{w}_r(0)\| = r$ and the

Gaussian density of $\mathcal{N}(\mathbf{0}, \sigma^2 I)$ is also a function of r.

$$\begin{aligned}\int_{\mathbb{R}^d} \|\mathbf{w}_r(0)\|^{2\ell-3} \rho(\mathbf{w}_r(0)) d\mathbf{w}_r(0) &= \int_0^\infty \left(\int_{\partial B(r)} r^{2\ell-3} \frac{e^{-\frac{r^2}{2\sigma^2}}}{\sqrt{(2\pi)^d}\sigma^d} dS \right) dr && (97)\\ &= \int_0^\infty r^{2\ell-3} \frac{e^{-\frac{r^2}{2\sigma^2}}}{\sqrt{(2\pi)^d}\sigma^d} \left(\int_{\partial B(r)} dS \right) dr && (98)\\ &= \int_0^\infty r^{2\ell-3} \frac{e^{-\frac{r^2}{2\sigma^2}}}{\sqrt{(2\pi)^d}\sigma^d} \text{Area}(\partial B(r)) dr && (99)\\ &= \frac{1}{(2\pi)^{\frac{d}{2}}\sigma^d} \int_0^\infty r^{2\ell-3} e^{-\frac{r^2}{2\sigma^2}} \frac{2\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2})} r^{d-1} dr && (100)\\ &= \frac{2}{2^{\frac{d}{2}}\sigma^d\Gamma(\frac{d}{2})} \int_0^\infty r^{2\ell+d-2} e^{-\frac{r^2}{2\sigma^2}} dr && (101)\end{aligned}$$

Let $\frac{r^2}{2\sigma^2} = t$, it holds that

$$\begin{aligned}\int_0^\infty r^{2\ell+d-2} e^{-\frac{r^2}{2\sigma^2}} dr &= \int_0^\infty (\sqrt{2}\sigma t^{\frac{1}{2}})^{2\ell+d-2} e^{-t} \sqrt{2}\sigma \frac{1}{2} t^{-\frac{1}{2}} dt && (102)\\ &= \int_0^\infty 2^{\frac{2\ell+d-2}{2}} \sigma^{2\ell+d-2} t^{\frac{2\ell+d-2}{2}} 2^{\frac{1}{2}} \sigma 2^{-1} t^{-\frac{1}{2}} e^{-t} dt && (103)\\ &= 2^{\ell+\frac{d}{2}-\frac{3}{2}} \sigma^{2\ell+d-1} \int_0^\infty t^{\ell+\frac{d}{2}-\frac{1}{2}-1} e^{-t} dt && (104)\\ &= 2^{\ell+\frac{d}{2}-\frac{3}{2}} \sigma^{2\ell+d-1} \Gamma\left(\ell + \frac{d}{2} - \frac{1}{2}\right) && (105)\end{aligned}$$

Thus we have

$$\int_{\mathbb{R}^d} \|\mathbf{w}_r(0)\|^{2\ell-3} \rho(\mathbf{w}_r(0)) d\mathbf{w}_r(0) = 2^{\ell-\frac{1}{2}} \sigma^{2\ell-1} \frac{\Gamma\left(\ell + \frac{d}{2} - \frac{1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)},$$

and then the bound

$$\begin{aligned}\int_{S_i^+ \cap S_j^+} F_{ij}(\mathbf{w}_r(0)) \rho(\mathbf{w}_r(0)) d\mathbf{w}_r(0) &\le C 2^{2\ell-4} R \int_{S_i^+ \cap S_j^+} \|\mathbf{w}_r(0)\|^{2\ell-3} \rho(\mathbf{w}_r(0)) d\mathbf{w}_r(0) + C 2^{2\ell-4} R^{2\ell-2} && (106)\\ &\le C 2^{2\ell-4} R \cdot \frac{1}{2} \left(2^{\ell-\frac{1}{2}} \sigma^{2\ell-1} \frac{\Gamma\left(\ell + \frac{d}{2} - \frac{1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)} \right) + C 2^{2\ell-4} R^{2\ell-2} && (107)\\ &= C R 2^{3\ell-\frac{11}{2}} \sigma^{2\ell-1} \frac{\Gamma\left(\ell + \frac{d}{2} - \frac{1}{2}\right)}{\Gamma(\frac{d}{2})} + C 2^{2\ell-4} R^{2\ell-2} && (108)\end{aligned}$$

Combining with the result of Lemma B.4, we conclude that if $\|\mathbf{w}_r(0) - \mathbf{w}_r\| \le R < 1$, $\ell \ge 1$, then there exists a constant K, such that

$$\begin{aligned}&\mathbb{E} \left| (\mathbf{w}_r(0)\mathbf{x}_i)^{\ell-1} (\mathbf{w}_r(0)\mathbf{x}_j)^{\ell-1} \mathbb{I}\{\mathbf{w}_r(0)\mathbf{x}_i \ge 0, \mathbf{w}_r(0)\mathbf{x}_j \ge 0\} \right.\\ &\quad \left. -(\mathbf{w}_r\mathbf{x}_i)^{\ell-1} (\mathbf{w}_r\mathbf{x}_j)^{\ell-1} \mathbb{I}\{\mathbf{w}_r\mathbf{x}_i \ge 0, \mathbf{w}_r\mathbf{x}_j \ge 0\} \right| \qquad (109)\\ &\quad \le KR.\end{aligned}$$

According to the expectation of $\mathbb{E}\left|H_{ij}(0) - H_{ij}\right|$, i.e.,

$$\begin{aligned}\mathbb{E}\left|H_{ij}(0) - H_{ij}\right| = \ell^2 \frac{1}{m} \sum_{r=1}^m &\mathbb{E} \left| (\mathbf{w}_r(0)\mathbf{x}_i)^{\ell-1} (\mathbf{w}_r(0)\mathbf{x}_j)^{\ell-1} \mathbb{I}\{\mathbf{w}_r(0)\mathbf{x}_i \ge 0, \mathbf{w}_r(0)\mathbf{x}_j \ge 0\} \right.\\ &\left. -(\mathbf{w}_r\mathbf{x}_i)^{\ell-1} (\mathbf{w}_r\mathbf{x}_j)^{\ell-1} \mathbb{I}\{\mathbf{w}_r\mathbf{x}_i \ge 0, \mathbf{w}_r\mathbf{x}_j \ge 0\} \right| \qquad (110)\\ &\le \ell^2 KR,\end{aligned}$$

and then

$$\mathbb{E}\sum_{i,j}|H_{ij}(0)-H_{ij}| = \sum_{i,j}\mathbb{E}\,|H_{ij}(0)-H_{ij}| \leq n^2\ell^2 KR.$$

Using inequality of $\|\cdot\|_2$, $\|\cdot\|_F$ and $|\cdot|_1$, we have

$$\|H(0)-H\|_2 \leq \|H(0)-H\|_F \leq \sum_{ij}|H_{ij}(0)-H_{ij}| \leq \frac{n^2\ell^2 KR}{\delta}$$

with probability at least $1-\delta$. So choose R such that

$$R < \frac{\delta\lambda_0}{4n^2\ell^2 K},$$

we can have

$$\|H(0)-H\|_2 \leq \frac{\lambda_0}{4}$$

with probability $1-\delta$. □

Lemma B.3. *Let*

$$F(\mathbf{w}_r) \overset{def}{=} (\mathbf{w}_r\mathbf{x}_i)^{\ell-1}(\mathbf{w}_r\mathbf{x}_j)^{\ell-1}.$$

Then there exists some constant C such that

$$\|\nabla F(\mathbf{w}_r)\| \leq C\,\|\mathbf{w}_r\|^{2\ell-3}.$$

Proof. Direct calculation gives the gradient:

$$\nabla F(\mathbf{w}_r) = \left(\frac{\partial F}{\partial w_{r1}}, ..., \frac{\partial F}{\partial w_{rd}}\right),$$

where

$$\frac{\partial F}{\partial w_{rk}} = (\ell-1)(\mathbf{w}_r\mathbf{x}_i)^{\ell-2}x_{ik}(\mathbf{w}_r\mathbf{x}_j)^{\ell-1} + (\mathbf{w}_r\mathbf{x}_i)^{\ell-1}(\ell-1)(\mathbf{w}_r\mathbf{x}_j)^{\ell-2}x_{jk} \tag{111}$$

$$= (\ell-1)x_{ik}(\mathbf{w}_r\mathbf{x}_i)^{\ell-2}(\mathbf{w}_r\mathbf{x}_j)^{\ell-1} + (\ell-1)x_{jk}(\mathbf{w}_r\mathbf{x}_i)^{\ell-1}(\mathbf{w}_r\mathbf{x}_j)^{\ell-2} \tag{112}$$

$$= (\ell-1)(\mathbf{w}_r\mathbf{x}_i)^{\ell-2}(\mathbf{w}_r\mathbf{x}_j)^{\ell-2}(x_{ik}(\mathbf{w}_r\mathbf{x}_j) + x_{jk}(\mathbf{w}_r\mathbf{x}_i)). \tag{113}$$

Then the ℓ_2-norm of $\nabla F(\mathbf{w}_r)$ can be estimated as

$$\left|\frac{\partial F}{\partial w_{rk}}\right| = (\ell-1)\left|(\mathbf{w}_r\mathbf{x}_i)^{\ell-2}(\mathbf{w}_r\mathbf{x}_j)^{\ell-2}(x_{ik}(\mathbf{w}_r\mathbf{x}_j) + x_{jk}(\mathbf{w}_r\mathbf{x}_i))\right| \tag{114}$$

$$\leq (\ell-1)\,|\mathbf{w}_r\mathbf{x}_i|^{\ell-2}\,|\mathbf{w}_r\mathbf{x}_j|^{\ell-2}\,|x_{ik}(\mathbf{w}_r\mathbf{x}_j) + x_{jk}(\mathbf{w}_r\mathbf{x}_i)| \tag{115}$$

$$\leq (\ell-1)(\|\mathbf{w}_r\|\cdot\|\mathbf{x}_i\|)^{\ell-2}(\|\mathbf{w}_r\|\cdot\|\mathbf{x}_j\|)^{\ell-2}\left(|x_{ik}(\mathbf{w}_r\mathbf{x}_j)| + |x_{jk}(\mathbf{w}_r\mathbf{x}_i)|\right) \tag{116}$$

$$\leq (\ell-1)\,\|\mathbf{w}_r\|^{\ell-2}\,\|\mathbf{w}_r\|^{\ell-2}\left(|x_{ik}|\cdot|\mathbf{w}_r\mathbf{x}_j| + |x_{jk}|\cdot|\mathbf{w}_r\mathbf{x}_i|\right) \tag{117}$$

$$\leq (\ell-1)\,\|\mathbf{w}_r\|^{2\ell-4}\left(\sqrt{d}\,|\mathbf{w}_r\mathbf{x}_j| + \sqrt{d}\,|\mathbf{w}_r\mathbf{x}_i|\right) \tag{118}$$

$$\leq (\ell-1)\,\|\mathbf{w}_r\|^{2\ell-4}\left(\sqrt{d}\,\|\mathbf{w}_r\|\cdot\|\mathbf{x}_j\| + \sqrt{d}\,\|\mathbf{w}_r\|\cdot\|\mathbf{x}_i\|\right) \tag{119}$$

$$\leq 2\sqrt{d}(\ell-1)\,\|\mathbf{w}_r\|^{2\ell-3} \tag{120}$$

where we use the inequality of ℓ_1-ℓ_2 norms for $\mathbf{x}_i$:

$$\|\mathbf{x}\|_2 \leq \|\mathbf{x}\|_1 \leq \sqrt{d}\,\|\mathbf{x}\|_2$$

which implies that $|x_{ik}|$ and $|x_{jk}|$ are less than $\sqrt{d}$. The same inequality implies that

$$\|\nabla F(\mathbf{w}_r)\| \leq \sum_{k=1}^{d}\left|\frac{\partial F}{\partial w_{rk}}\right| \leq 2d^{\frac{3}{2}}(\ell-1)\,\|\mathbf{w}_r\|^{2\ell-3}.$$

The proof completes. □

Lemma B.4. *Following the notation above, let*

$$F_{ij} = \left|(\mathbf{w}_r(0)\mathbf{x}_i)^{\ell-1}(\mathbf{w}_r(0)\mathbf{x}_j)^{\ell-1}\mathbb{I}\{\mathbf{w}_r(0)\mathbf{x}_i \geq 0, \mathbf{w}_r(0)\mathbf{x}_j \geq 0\} - (\mathbf{w}_r\mathbf{x}_i)^{\ell-1}(\mathbf{w}_r\mathbf{x}_j)^{\ell-1}\mathbb{I}\{\mathbf{w}_r\mathbf{x}_i \geq 0, \mathbf{w}_r\mathbf{x}_j \geq 0\}\right|$$

and $A = \mathbb{R}^d - (S_i^+ \cap S_j^+) - (S_i^- \cup S_j^-)$*, then*

$$\int_A F_{ij}\rho(\mathbf{w}_r(0))d\mathbf{w}_r(0) \leq C_1\sigma^{2\ell-3}R + C_2\sigma^{2\ell-4}R^2 + C_3\sigma^{-1}R^{2\ell-1}$$

for some constants C_1*,* C_2 *and* C_3*.*

Remark B.5. If we further assume that $\sigma = 1$, it is immediate that

$$\int_A F_{ij}\rho(\mathbf{w}_r(0))d\mathbf{w}_r(0) \leq C_1R + C_2R^2 + C_3R^{2\ell-1},$$

and this can be simplified to a bound of KR for some constant K.

Proof. Note that F_{ij} might have four different forms if $\mathbf{w}_r(0) \in A$, i.e.,

1. 0,
2. $\left|(\mathbf{w}_r(0)\mathbf{x}_i)^{\ell-1}(\mathbf{w}_r(0)\mathbf{x}_j)^{\ell} - (\mathbf{w}_r\mathbf{x}_i)^{\ell-1}(\mathbf{w}_r\mathbf{x}_j)^{\ell-1}\right|$,
3. $\left|(\mathbf{w}_r(0)\mathbf{x}_i)^{\ell-1}(\mathbf{w}_r(0)\mathbf{x}_j)^{\ell-1}\right|$,
4. $\left|(\mathbf{w}_r\mathbf{x}_i)^{\ell-1}(\mathbf{w}_r\mathbf{x}_j)^{\ell-1}\right|$

Using previous notation, i.e.,

$$F(\mathbf{w}_r) = (\mathbf{w}_r\mathbf{x}_i)^{\ell-1}(\mathbf{w}_r\mathbf{x}_j)^{\ell-1},$$

if $\mathbf{w}_r$ is in the neighborhood of $\mathbf{w}_r(0)$, i.e., $\|\mathbf{w}_r(0) - \mathbf{w}_r\| \leq R$, the Taylor expansion with $\mathbf{w}_r = \mathbf{w}_r(0) + \boldsymbol{\theta}$,

$$F(\mathbf{w}_r(0) + \boldsymbol{\theta}) = F(\mathbf{w}_r(0)) + \nabla F(\mathbf{w}_r(0)) \cdot \boldsymbol{\theta} + O(\|\boldsymbol{\theta}\|^2),$$

gives

$$\begin{aligned} |F(\mathbf{w}_r)| &\leq |F(\mathbf{w}_r(0))| + \|\nabla F(\mathbf{w}_r(0))\| \cdot \|\boldsymbol{\theta}\| + O(\|\boldsymbol{\theta}\|^2) \qquad (121)\\ &\leq |F(\mathbf{w}_r(0))| + CR\,\|\mathbf{w}_r(0)\|^{2\ell-3} + O(\|\boldsymbol{\theta}\|^2). \qquad (122) \end{aligned}$$

As we have proved,

$$|F(\mathbf{w}_r(0)) - F(\mathbf{w}_r)| \leq C2^{2\ell-4}R\,\|\mathbf{w}_r(0)\|^{2\ell-3} + C2^{2\ell-4}R^{2\ell-2}.$$

So a global upper bound of F_{ij} can be the following

$$|F(\mathbf{w}_r(0))| + C2^{2\ell-4}R\,\|\mathbf{w}_r(0)\|^{2\ell-3} + O(\|\boldsymbol{\theta}\|^2) + C2^{2\ell-4}R^{2\ell-2}.$$

Since $\|\boldsymbol{\theta}\| \leq R$, and ℓ in our setting is at least 2, the above expression can be simplified as

$$|F(\mathbf{w}_r(0))| + C2^{2\ell-4}R\,\|\mathbf{w}_r(0)\|^{2\ell-3} + C_1R^2$$

for some constant C_1. We denote

$$I_1 = \int_A F(\mathbf{w}_r(0))\rho(\mathbf{w}_r(0))d\mathbf{w}_r(0)$$

$$I_2 = \int_A C2^{2\ell-4}R\,\|\mathbf{w}_r(0)\|^{2\ell-3}\rho(\mathbf{w}_r(0))d\mathbf{w}_r(0)$$

and

$$I_3 = \int_A C_1R^2\rho(\mathbf{w}_r(0))d\mathbf{w}_r(0).$$

Further estimation can be obtained.

$$
\begin{aligned}
I_1 &= \int_A \left|(\mathbf{w}_r(0)\mathbf{x}_i)^{\ell-1}(\mathbf{w}_r(0)\mathbf{x}_j)^{\ell-1}\right| \rho(\mathbf{w}_r(0))d\mathbf{w}_r(0) && (123)\\
&\leq \int_A \|\mathbf{w}_r(0)\|^{2\ell-2} \rho(\mathbf{w}_r(0))d\mathbf{w}_r(0) && (124)\\
&= \int_A \|\mathbf{w}_r(0)\|^{2\ell-2} \frac{e^{-\frac{\|\mathbf{w}_r(0)\|^2}{2\sigma^2}}}{\sqrt{(2\pi)^d}\sigma^d} d\mathbf{w}_r(0) && (125)\\
&\leq 2\cdot 2R \int_{\mathbb{R}^{d-1}} \|\mathbf{x}\|^{2\ell-2} \frac{e^{-\frac{\|\mathbf{x}\|^2}{2\sigma^2}}}{\sqrt{(2\pi)^d}\sigma^d} d\mathbf{x} && (126)
\end{aligned}
$$

where $\mathbf{x}$ is the first $d-1$ component of $\mathbf{w}_r(0) \in \mathbb{R}^d$. And then

$$
\begin{aligned}
I_1 &\leq \frac{4R}{\sqrt{(2\pi)^d}\sigma^d} \int_{\mathbb{R}^{d-1}} \|\mathbf{x}\|^{2\ell-2} e^{-\frac{\|\mathbf{x}\|^2}{2\sigma^2}} d\mathbf{x} && (127)\\
&= \frac{4R}{\sqrt{(2\pi)^d}\sigma^d} \int_0^\infty \left(\int_{S^{d-2}(r)} r^{2\ell-2} e^{-\frac{r^2}{2\sigma^2}} dS \right) dr && (128)\\
&= \frac{4R}{\sqrt{(2\pi)^d}\sigma^d} \int_0^\infty r^{2\ell-2} e^{-\frac{r^2}{2\sigma^2}} \text{Area}(S^{d-2}(r)) dr && (129)\\
&= \frac{4R}{\sqrt{(2\pi)^d}\sigma^d} \int_0^\infty r^{2\ell-2} e^{-\frac{r^2}{2\sigma^2}} \frac{2\pi^{\frac{d-1}{2}}}{\Gamma\left(\frac{d-1}{2}\right)} r^{d-2} dr && (130)\\
&= \frac{4R}{\sqrt{(2\pi)^d}\sigma^d} \frac{2\pi^{\frac{d-1}{2}}}{\Gamma\left(\frac{d-1}{2}\right)} \int_0^\infty r^{2\ell+d-4} e^{-\frac{r^2}{2\sigma^2}} dr && (131)
\end{aligned}
$$

where

$$
\begin{aligned}
\int_0^\infty r^{2\ell+d-4} e^{-\frac{r^2}{2\sigma^2}} dr &= 2^{\ell+\frac{d}{2}-\frac{5}{2}} \sigma^{2\ell+d-3} \int_0^\infty t^{\ell+\frac{d}{2}-\frac{3}{2}-1} e^{-t} dt && (132)\\
&= 2^{\ell+\frac{d}{2}-\frac{5}{2}} \sigma^{2\ell+d-3} \Gamma\left(\ell+\frac{d}{2}-\frac{3}{2}\right) && (133)
\end{aligned}
$$

Thus we have

$$
I_1 \leq R \cdot 2^{\ell+\frac{1}{2}} \pi^{\frac{d}{2}-1} \sigma^{2\ell-3} \frac{\Gamma\left(\ell+\frac{d}{2}-\frac{3}{2}\right)}{\Gamma\left(\frac{d-1}{2}\right)}.
$$

$$
I_2 = C 2^{2\ell-4} R \int_A \|\mathbf{w}_r(0)\|^{2\ell-3} \rho(\mathbf{w}_r(0)) d\mathbf{w}_r(0) \qquad (134)
$$

where

$$\int_A \|\mathbf{w}_r(0)\|^{2\ell-3}\rho(\mathbf{w}_r(0))d\mathbf{w}_r(0) \leq 2\cdot 2R\int_{\mathbb{R}^{d-1}}\|\mathbf{x}\|^{2\ell-3}\frac{e^{-\frac{\|\mathbf{x}\|^2}{2\sigma^2}}}{\sqrt{(2\pi)^d}\sigma^d}d\mathbf{x} \tag{135}$$

$$= \frac{4R}{\sqrt{(2\pi)^d}\sigma^d}\int_0^\infty r^{2\ell-3}e^{-\frac{r^2}{2\sigma^2}}\text{Area}(S^{d-2}(r))dr \tag{136}$$

$$= \frac{4R}{\sqrt{(2\pi)^d}\sigma^d}\frac{2\pi^{\frac{d-1}{2}}}{\Gamma\left(\frac{d-1}{2}\right)}\int_0^\infty r^{2\ell+d-5}e^{-\frac{r^2}{2\sigma^2}}dr \tag{137}$$

$$= \frac{4R}{\sqrt{(2\pi)^d}\sigma^d}\frac{2\pi^{\frac{d-1}{2}}}{\Gamma\left(\frac{d-1}{2}\right)}2^{\ell+\frac{d}{2}-3}\sigma^{2\ell+d-4}\int_0^\infty t^{\ell+\frac{d}{2}-2-1}e^{-t}dt \tag{138}$$

$$= \frac{4R}{\sqrt{(2\pi)^d}\sigma^d}\frac{2\pi^{\frac{d-1}{2}}}{\Gamma\left(\frac{d-1}{2}\right)}2^{\ell+\frac{d}{2}-3}\sigma^{2\ell+d-4}\Gamma\left(\ell+\frac{d}{2}-2\right) \tag{139}$$

$$= R\cdot\frac{2^\ell\sigma^{2\ell-4}}{\sqrt{\pi}}\frac{\Gamma\left(\ell+\frac{d}{2}-2\right)}{\Gamma\left(\frac{d-1}{2}\right)}. \tag{140}$$

And then we end up with estimate of I_2 to be

$$I_2 \leq C2^{2\ell-4}R\cdot R\cdot\frac{2^\ell\sigma^{2\ell-4}}{\sqrt{\pi}}\frac{\Gamma\left(\ell+\frac{d}{2}-2\right)}{\Gamma\left(\frac{d-1}{2}\right)} \tag{141}$$

$$= R^2\cdot\frac{C2^{3\ell-4}\sigma^{2\ell-4}}{\sqrt{\pi}}\frac{\Gamma\left(\ell+\frac{d}{2}-2\right)}{\Gamma\left(\frac{d-1}{2}\right)} \tag{142}$$

Recall that

$$F(\mathbf{w}_r) = F(\mathbf{w}_r(0)) + \nabla F(\mathbf{w}_r(0)+\boldsymbol{\theta})(\mathbf{w}_r - \mathbf{w}_r(0)) \tag{143}$$

so it is bounded as

$$|F(\mathbf{w}_r)| \leq |F(\mathbf{w}_r(0))| + C2^{2\ell-4}R\|\mathbf{w}_r(0)\|^{2\ell-3} + C2^{2\ell-4}R^{2\ell-2},$$

and then it remains to estimate the integral of the last term over A.

$$\int_A C2^{2\ell-4}R^{2\ell-2}\frac{e^{-\frac{\|\mathbf{x}\|^2}{2\sigma^2}}}{\sqrt{(2\pi)^d}\sigma^d}d\mathbf{x} \leq C2^{2\ell-4}R^{2\ell-2}4R\int_{\mathbb{R}^{d-1}}\frac{e^{-\frac{\|\mathbf{x}\|^2}{2\sigma^2}}}{\sqrt{(2\pi)^d}\sigma^d}d\mathbf{x} \tag{144}$$

$$= C2^{2\ell-4}R^{2\ell-2}\frac{4R}{\sqrt{2\pi}\sigma}\int_{\mathbb{R}^{d-1}}\frac{e^{-\frac{\|\mathbf{x}\|^2}{2\sigma^2}}}{\sqrt{(2\pi)^{d-1}}\sigma^{d-1}}d\mathbf{x} \tag{145}$$

$$= C2^{2\ell-4}R^{2\ell-2}\frac{4R}{\sqrt{2\pi}\sigma} \tag{146}$$

$$= \frac{C2^{2\ell-2}}{\sqrt{2\pi}\sigma}R^{2\ell-1} \tag{147}$$

Combining with the bounds of I_1 and I_2, we complete the proof. □

Lemma B.6. *Suppose for* $0 \leq s \leq t$, $\lambda_{\min}(H(s)) \geq \frac{\lambda_0}{2}$. *Then we have*

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

Proof. It has been calculated that for each sample $i \in [n]$, the evolution of prediction satisfies

$$\frac{du_i}{dt} = \sum_{j=1}^n (y_j - u_j(t))H_{ij}(t).$$

Thus the evolution of norm $\|\mathbf{y}-\mathbf{u}(t)\|_2^2$ satisfies

$$\begin{aligned}\frac{d}{dt}\left\|\mathbf{y}-\mathbf{u}(t)\right\|_2^2 &= 2\langle\frac{d}{dt}(\mathbf{y}-\mathbf{u}(t)),\mathbf{y}-\mathbf{u}(t)\rangle && (148)\\ &= -2\langle\frac{d}{dt}\mathbf{u}(t),\mathbf{y}-\mathbf{u}(t)\rangle && (149)\\ &= -2\langle\sum_{j=1}^{n}(y_j-u_j(t))H_{ij}(t),\mathbf{y}-\mathbf{u}(t)\rangle && (150)\\ &= -2(\mathbf{y}-\mathbf{u}(t))^\top H(t)(\mathbf{y}-\mathbf{u}(t)) && (151)\\ &\leq -\lambda_0\left\|\mathbf{y}-\mathbf{u}(t)\right\|_2^2. && (152)\end{aligned}$$

Then

$$\begin{aligned}\frac{d}{dt}\left(e^{\lambda_0 t}\left\|\mathbf{y}-\mathbf{u}(t)\right\|_2^2\right) &= \lambda_0 e^{\lambda_0 t}\left\|\mathbf{y}-\mathbf{u}(t)\right\|_2^2 + e^{\lambda_0 t}\frac{d}{dt}\left\|\mathbf{y}-\mathbf{u}(t)\right\|_2^2 && (153)\\ &\leq \lambda_0 e^{\lambda_0 t}\left\|\mathbf{y}-\mathbf{u}(t)\right\|_2^2 - \lambda_0 e^{\lambda_0 t}\left\|\mathbf{y}-\mathbf{u}(t)\right\|_2^2 && (154)\\ &= 0 && (155)\end{aligned}$$

which implies that $e^{\lambda_0 t}\left\|\mathbf{y}-\mathbf{u}(t)\right\|_2^2$ is decreasing in t and then it holds that

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

The previous calculation yields

$$\frac{d}{ds}\mathbf{w}_r(s) = -\frac{\partial L}{\partial \mathbf{w}_r} = -\frac{1}{\sqrt{m}}\sum_{i=1}^{n}(u_i-y_i)a_r\ell(\mathbf{w}_r\mathbf{x}_i)^{\ell}\mathbb{I}\{\mathbf{w}_r\mathbf{x}_i\geq 0\}\mathbf{x}_i \tag{156}$$

and then

$$\begin{aligned}\left\|\frac{d\mathbf{w}_r}{ds}\right\|_2 &= \left\|-\frac{1}{\sqrt{m}}\sum_{i=1}^{n}(u_i-y_i)a_r\ell(\mathbf{w}_r\mathbf{x}_i)^{\ell-1}\mathbb{I}\{\mathbf{w}_r\mathbf{x}_i\geq 0\}\mathbf{x}_i\right\|_2 && (157)\\ &\leq \frac{1}{\sqrt{m}}\sum_{i=1}^{n}\left\|(u_i-y_i)a_r\ell(\mathbf{w}_r\mathbf{x}_i)^{\ell-1}\mathbb{I}\{\mathbf{w}_r\mathbf{x}_i\geq 0\}\mathbf{x}_i\right\|_2 && (158)\\ &\leq \frac{1}{\sqrt{m}}\sum_{i=1}^{n}\ell\left|a_r\right|\cdot\left|u_i-y_i\right|\cdot\left|(\mathbf{w}_r\mathbf{x}_i)^{\ell-1}\right|\cdot\left\|\mathbf{x}_i\right\|_2 && (159)\end{aligned}$$

For the training of $\mathbf{W}$, we let $a_r=1$ and it is assumed that $\|\mathbf{x}_i\|=1$, we have

$$\begin{aligned}\left\|\frac{d\mathbf{w}_r}{ds}\right\|_2 &\leq \frac{\ell}{\sqrt{m}}\sum_{i=1}^{n}\left|u_i-y_i\right|\cdot\left\|\mathbf{w}_r\right\|_2^{\ell-1} && (160)\\ &\leq \frac{\ell\sqrt{n}}{\sqrt{m}}\left\|\mathbf{w}_r\right\|_2^{\ell-1}\cdot\left\|\mathbf{u}(s)-\mathbf{y}\right\|_2. && (161)\end{aligned}$$

Since it is assumed that during the whole training process, $\mathbf{w}_r$ is in a bounded region, i.e., $\left\|\mathbf{w}_r\right\|_2\leq\kappa$, and this implies

$$\left\|\frac{d\mathbf{w}_r}{ds}\right\|_2 \leq \frac{\ell\kappa^{\ell-1}\sqrt{n}}{\sqrt{m}}\left\|\mathbf{u}(s)-\mathbf{y}\right\|_2 \leq \frac{\ell\kappa^{\ell-1}\sqrt{n}}{\sqrt{m}}e^{-\frac{\lambda_0 s}{2}}\left\|\mathbf{u}(0)-\mathbf{y}\right\|_2.$$

Integration gives

$$\begin{aligned}
\left\|\mathbf{w}_r(t)-\mathbf{w}_r(0)\right\|_2 &\leq \int_0^t \left\|\frac{d\mathbf{w}_r}{ds}\right\|_2 ds \qquad (162)\\
&\leq \frac{\ell\kappa^{\ell-1}\sqrt{n}}{\sqrt{m}}\left\|\mathbf{u}(0)-\mathbf{y}\right\|_2 \int_0^t e^{-\frac{\lambda_0 s}{2}} ds \qquad (163)\\
&= \frac{\ell\kappa^{\ell-1}\sqrt{n}}{\sqrt{m}}\left\|\mathbf{u}(0)-\mathbf{y}\right\|_2 \cdot \frac{2}{\lambda_0}\left(1-e^{-\lambda_0 t}\right) \qquad (164)\\
&\leq \frac{2\ell\kappa^{\ell-1}\sqrt{n}}{\lambda_0\sqrt{m}}\left\|\mathbf{u}(0)-\mathbf{y}\right\|_2 \qquad (165)
\end{aligned}$$

□

The next theorem gives the convergence guarantee of joint training of $\mathbf{W}$ and $\mathbf{a}$.

Theorem B.7. *Consider the joint gradient descent on both layers. Suppose Assumption 1 holds, $\|\mathbf{x}_i\| = 1$ and the sample values y_i are bounded. Then if we set the number of hidden nodes $m = \Omega\left(\frac{n^2\log(n/\delta)}{\lambda_0}\right)$ and initializing $\mathbf{w}_r(0) \sim \mathcal{N}(\mathbf{0}, I)$, $a_r = 1$, then with probability $1-\delta$ over the initialization, we have*

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

Lemma B.8. *With probability at least $1-\delta$, if a set of weight vectors $\{\mathbf{w_r}\}_r^m$ and the output weight $\mathbf{a}$ satisfy for all $r \in [m]$, $\left\|\mathbf{w}_r - \mathbf{w}_r(0)\right\|_2 \leq R_w$ and $|a_r - a_r(0)| \leq R_a$, the the matrix H satisfies*

$$\left\|H - H(0)\right\|_2 \leq \frac{\lambda_0}{4} \text{ and } \lambda_{\min}(H) > \frac{\lambda_0}{2}.$$

Proof. Let

$$H'_{ij} = \langle \mathbf{x}_i, \mathbf{x}_j\rangle \left(\frac{1}{m}\sum_{r=1}^m a_r^2 \zeta(\mathbf{w}_r\mathbf{x}_i)\zeta(\mathbf{w}_r\mathbf{x}_j)\right),$$

from the proof of previous section, we have that

$$\left\|H' - H(0)\right\|_2 \leq \frac{n^2\ell^2 K R_w}{\delta}$$

for some constant K. On the other hand,

$$\begin{aligned}
H_{ij} - H'_{ij} &= \langle \mathbf{x}_i, \mathbf{x}_j\rangle \left(\frac{1}{m}\sum_{r=1}^m a_r^2 \zeta(\mathbf{w}_r\mathbf{x}_i)\zeta(\mathbf{w}_r\mathbf{x}_j)\right) \qquad (166)\\
&\quad - \langle \mathbf{x}_i, \mathbf{x}_j\rangle \left(\frac{1}{m}\sum_{r=1}^m \zeta(\mathbf{w}_r\mathbf{x}_i)\zeta(\mathbf{w}_r\mathbf{x}_j)\right) \qquad (167)\\
&= \langle \mathbf{x}_i, \mathbf{x}_j\rangle \left(\frac{1}{m}\sum_{r=1}^m (a_r^2-1) \zeta(\mathbf{w}_r\mathbf{x}_i)\zeta(\mathbf{w}_r\mathbf{x}_j)\right), \qquad (168)
\end{aligned}$$

and then

$$\begin{aligned}
\left|H_{ij} - H'_{ij}\right| &= \left|\langle \mathbf{x}_i, \mathbf{x}_j\rangle \left(\frac{1}{m}\sum_{r=1}^m (a_r^2-1) \zeta(\mathbf{w}_r\mathbf{x}_i)\zeta(\mathbf{w}_r\mathbf{x}_j)\right)\right| \qquad (169)\\
&\leq \left|\langle \mathbf{x}_i, \mathbf{x}_j\rangle\right| \cdot \left|\frac{1}{m}\sum_{r=1}^m (a_r^2-1) \zeta(\mathbf{w}_r\mathbf{x}_i)\zeta(\mathbf{w}_r\mathbf{x}_j)\right| \qquad (170)\\
&\leq \left(\max_{r\in[m]}\left|a_r^2-1\right|\right)\left|\frac{1}{m}\sum_{r=1}^m \zeta(\mathbf{w}_r\mathbf{x}_i)\zeta(\mathbf{w}_r\mathbf{x}_j)\right| \qquad (171)\\
&\leq \left(\max_{r\in[m]}\left|a_r^2-1\right|\right)\frac{1}{m}\sum_{r=1}^m \left|\zeta(\mathbf{w}_r\mathbf{x}_i)\zeta(\mathbf{w}_r\mathbf{x}_j)\right|. \qquad (172)
\end{aligned}$$

The expectation is then bounded by

$$\mathbb{E}\left|H_{ij}-H'_{ij}\right| \leq \left(\max_{r\in[m]}\left|a_r^2-1\right|\right)\frac{1}{m}\sum_{r=1}^{m}\mathbb{E}\left|\zeta(\mathbf{w}_r\mathbf{x}_i)\zeta(\mathbf{w}_r\mathbf{x}_j)\right| \tag{173}$$

$$= \left(\max_{r\in[m]}\left|a_r^2-1\right|\right)\mathbb{E}\left|\zeta(\mathbf{w}_r\mathbf{x}_i)\zeta(\mathbf{w}_r\mathbf{x}_j)\right|. \tag{174}$$

We next bound $\mathbb{E}\left|\zeta(\mathbf{w}_r\mathbf{x}_i)\zeta(\mathbf{w}_r\mathbf{x}_j)\right|$. Since the expression of $\zeta(\cdot)$ gives the following estimate:

$$\left|\zeta(\mathbf{w}_r\mathbf{x}_i)\zeta(\mathbf{w}_r\mathbf{x}_j)\right| = \left|(\mathbf{w}_r\mathbf{x}_i)^{\ell}(\mathbf{w}_r\mathbf{x}_j)^{\ell}\mathbb{I}\{\mathbf{w}_r\mathbf{x}_i \geq 0, \mathbf{w}_r\mathbf{x}_j \geq 0\}\right| \tag{175}$$

$$\leq \left|(\mathbf{w}_r\mathbf{x}_i)^{\ell}(\mathbf{w}_r\mathbf{x}_j)^{\ell}\right| \tag{176}$$

$$\leq \left|\mathbf{w}_r\mathbf{x}_i\right|^{\ell}\left|\mathbf{w}_r\mathbf{x}_j\right|^{\ell} \tag{177}$$

$$\leq \left\|\mathbf{w}_r\right\|^{2\ell}\left\|\mathbf{x}_i\right\|^{\ell}\left\|\mathbf{x}_j\right\|^{\ell} \tag{178}$$

$$\leq \left\|\mathbf{w}_r\right\|^{2\ell}, \tag{179}$$

it suffices to estimate $\left\|\mathbf{w}_r(0)+\boldsymbol{\theta}\right\|^{2\ell}$, given $\boldsymbol{\theta}=\mathbf{w}_r-\mathbf{w}_r(0)$ and $\|\boldsymbol{\theta}\| \leq R_w$.

$$\mathbb{E}\left|\zeta(\mathbf{w}_r\mathbf{x}_i)\zeta(\mathbf{w}_r\mathbf{x}_j)\right| \leq \mathbb{E}\left\|\mathbf{w}_r+\boldsymbol{\theta}\right\|^{2\ell} \tag{180}$$

$$\leq 2^{2\ell-1}\mathbb{E}\left(\left\|\mathbf{w}_r(0)\right\|^{2\ell}+\left\|\boldsymbol{\theta}\right\|^{2\ell}\right) \tag{181}$$

$$\leq 2^{2\ell-1}\mathbb{E}\left\|\mathbf{w}_r(0)\right\|^{2\ell}+2^{2\ell-1}\mathbb{E}R^{2\ell} \tag{182}$$

$$= 2^{2\ell-1}\int_{\mathbb{R}^d}\left\|\mathbf{w}_r(0)\right\|^{2\ell}\rho(\mathbf{w}_r(0))d\mathbf{w}_r(0)+2^{2\ell-1}R^{2\ell} \tag{183}$$

where $\rho(\cdot)$ is standard Gaussian distribution. The previous calculation has already given that

$$\int_{\mathbb{R}^d}\left\|\mathbf{w}_r(0)\right\|^{2\ell}\rho(\mathbf{w}_r(0))d\mathbf{w}_r(0) = 2^{\ell+1}\frac{\Gamma(\ell+\frac{d}{2}+1)}{\Gamma(\frac{d}{2})},$$

and then

$$\mathbb{E}\left|\zeta(\mathbf{w}_r\mathbf{x}_i)\zeta(\mathbf{w}_r\mathbf{x}_j)\right| \leq 2^{2\ell-1}2^{\ell+1}\frac{\Gamma(\ell+\frac{d}{2}+1)}{\Gamma(\frac{d}{2})}+2^{2\ell-1}R_w^{2\ell} \tag{184}$$

$$= 2^{3\ell}\frac{\Gamma(\ell+\frac{d}{2}+1)}{\Gamma(\frac{d}{2})}+2^{2\ell-1}R_w^{2\ell}. \tag{185}$$

If we assume R_w is less than 1, and then the above expression is bounded by some constant, we have shown that

$$\mathbb{E}\left|H_{ij}-H'_{ij}\right| \leq K_1R_a$$

for some constant K_1. Then we have

$$\mathbb{E}\left\|H-H'\right\|_2 \leq n^2\mathbb{E}\left|H_{ij}-H'_{ij}\right| = n^2K_1R_a,$$

and then

$$\mathbb{E}\left\|H-H(0)\right\|_2 \leq \mathbb{E}\left\|H-H'\right\|_2+\mathbb{E}\left\|H'-H(0)\right\|_2 \leq n^2K_1R_a+n^2\ell^2KR_w.$$

Thus, according to concentration inequality, the radius R_w and R_a can be chosen based on λ_0 and probability threshold δ, so that $\left\|H-H(0)\right\|_2 \leq \frac{\lambda_0}{4}$ with probability at least $1-\delta$. The proof completes. □

Lemma B.9. *Suppose for* $0 \leq s \leq t$, $\lambda_{\min}(H(s)) \geq \frac{\lambda_0}{2}$ *and* $|a_r(s)-a_r(0)| \leq R_a$. *Then we have* $\left\|\mathbf{w}_r(t)-\mathbf{w}_r(0)\right\|_2 \leq R'_w$.

Proof.

$$\left\|\frac{d\mathbf{w}_r(s)}{ds}\right\|_2 = \left\|-\frac{1}{\sqrt{m}}\sum_{i=1}^{n}(u_i - y_i)a_r(s)\ell(\mathbf{w}_r\mathbf{x}_i)^{\ell-1}\mathbb{I}\{\mathbf{w}_r\mathbf{x}_i \geq 0\}\mathbf{x}_i\right\|_2 \tag{186}$$

$$\leq \frac{1}{\sqrt{m}}\sum_{i=1}^{n}|u_i - y_i|\,|a_r(s)|\,\ell\,\|\mathbf{w}_r\|^{\ell-1}. \tag{187}$$

With the assumption that $\|\mathbf{w}_r\|$ is less than some constant κ, we can prove that

$$\|\mathbf{w}_r(t) - \mathbf{w}_r(0)\|_2 \leq \int_0^t \left\|\frac{d\mathbf{w}_r(s)}{ds}\right\|_2 ds \tag{188}$$

$$\leq C(\kappa)\int_0^t \|\mathbf{y} - \mathbf{u}(s)\|_2\, ds \tag{189}$$

$$\leq C(\kappa)\int_0^t e^{-\frac{\lambda_0 s}{2}}\|\mathbf{y} - \mathbf{u}(0)\|_2\, ds \tag{190}$$

$$\leq 2C(\kappa)\frac{\|\mathbf{y} - \mathbf{u}(0)\|_2}{\lambda_0}, \tag{191}$$

where $C(\kappa)$ is a constant depending on κ and involving ℓ, n and m. □

Lemma B.10. *With probability at least $1-\delta$ over initialization, suppose for $0 \leq s \leq t$, $\lambda_{\min}(H(s)) \geq \frac{\lambda_0}{2}$ and $\|\mathbf{w}_r(s) - \mathbf{w}_r(0)\|_2 \leq R_w$. Then we have $|a_r(t) - a_r(0)| \leq R'_a$ for all $r \in [m]$.*

Proof. Since $\mathbf{w}_r(0) \sim \mathcal{N}(0, I)$, we have with probability at least $1-\delta$, $|\mathbf{w}_r(0)\mathbf{x}_i| \leq 3\sqrt{\log\left(\frac{mn}{\delta}\right)}$. For $0 \leq s \leq t$, we have

$$\left|\frac{d}{ds}a_r(s)\right| = \left|\frac{1}{\sqrt{m}}\sum_{i=1}^{n}(f(\mathbf{W}(s), \mathbf{a}(s), \mathbf{x}_i) - y_i)\sigma(\mathbf{w}_r(s)\mathbf{x}_i)\right| \tag{192}$$

$$\leq \frac{1}{\sqrt{m}}\sum_{i=1}^{n}|f(\mathbf{W}(s), \mathbf{a}(s), \mathbf{x}_i) - y_i| \cdot |\mathbf{w}_r(s)\mathbf{x}_i|^{\ell}. \tag{193}$$

Note that

$$|\mathbf{w}_r(s)\mathbf{x}_i| = |\mathbf{w}_r(0)\mathbf{x}_i + (\mathbf{w}_r(s) - \mathbf{w}_r(0))\mathbf{x}_i| \tag{194}$$

$$\leq |\mathbf{w}_r(0)\mathbf{x}_i| + |(\mathbf{w}_r(s) - \mathbf{w}_r(0))\mathbf{x}_i| \tag{195}$$

$$\leq 3\sqrt{\log\left(\frac{mn}{\delta}\right)} + R_w, \tag{196}$$

and then

$$|\mathbf{w}_r(s)\mathbf{x}_i|^{\ell} \leq \left(3\sqrt{\log\left(\frac{mn}{\delta}\right)} + R_w\right)^{\ell}.$$

Therefore, the differential $\frac{da_r(s)}{ds}$ can be bounded as

$$\left|\frac{da_r(s)}{ds}\right| \leq \frac{\sqrt{n}}{\sqrt{m}}\|\mathbf{y} - \mathbf{u}(s)\|_2\left(3\sqrt{\log\left(\frac{mn}{\delta}\right)} + R_w\right)^{\ell} \tag{197}$$

So the bound of $\|\mathbf{a}(t) - \mathbf{a}(0)\|_2$ can be obtained by

$$\|\mathbf{a}(t) - \mathbf{a}(0)\|_2 \leq \int_0^t \left\|\frac{d\mathbf{a}(s)}{ds}\right\|_2 ds \leq C'(\delta, \ell, m, n)\,\|\mathbf{y} - \mathbf{u}(0)\|_2 \cdot \frac{1}{\lambda_0}.$$

□

B.2. Proof of Discrete Time Gradient Descent

Lemma B.11.

$$\|\mathbf{u}(t+1)-\mathbf{u}(t)\|_2^2 \leq \ell^4\kappa^{4\ell-4}\eta^2 n^2 \|\mathbf{u}(t)-\mathbf{y}\|_2^2$$

Proof. Recall that the neural network has the form of

$$f(\mathbf{W},\mathbf{a},\mathbf{x}) = \frac{1}{\sqrt{m}}\sum_{k=1}^{m} a_k\sigma(\mathbf{w}_k\mathbf{x}).$$

where $\sigma(\cdot)$ is the ℓ'th power of ReLU function. So the difference of predictions between two iterations has the following expression by a direct calculation,

$$\begin{aligned} u_i(t+1)-u_i(t) &= \frac{1}{\sqrt{m}}\sum_{k=1}^{m} a_k\sigma(\mathbf{w}_k(t+1)\mathbf{x}_i) - \frac{1}{\sqrt{m}}\sum_{k=1}^{m} a_k\sigma(\mathbf{w}_k(t)\mathbf{x}_i) \\ &= \frac{1}{\sqrt{m}}\sum_{k=1}^{m} a_k\left(\sigma(\mathbf{w}_k(t+1)\mathbf{x}_i)-\sigma(\mathbf{w}_k(t)\mathbf{x}_i)\right) \\ &= \frac{1}{\sqrt{m}}\sum_{r=1}^{m} a_r\left(\sigma\left(\left(\mathbf{w}_k(t)-\eta\frac{\partial L(\mathbf{W}(t))}{\partial \mathbf{w}_k(t)}\right)\mathbf{x}_i\right)-\sigma(\mathbf{w}_k(t)\mathbf{x}_i)\right) \end{aligned} \tag{198}$$

Since the range of $\|\mathbf{w}\|_2$ is assumed to be bounded by κ during the whole training process, the function $\sigma(\mathbf{w}\mathbf{x}_i) = (\mathbf{w}\mathbf{x}_i)^\ell$ is Lipschitz and the Lipschitz constant can be estimated as follows.

$$\nabla_{\mathbf{w}}\sigma(\mathbf{w}\mathbf{x}_i) = \left(\ell(\mathbf{w}\mathbf{x}_i)^{\ell-1}x_{i1},...,\ell(\mathbf{w}\mathbf{x}_i)^{\ell-1}x_{id}\right)$$

which implies

$$\begin{aligned} \|\nabla_{\mathbf{w}}\sigma(\mathbf{w}\mathbf{x}_i)\|_2 &= \ell(|\mathbf{w}\mathbf{x}_i|)^{\ell-1}\|\mathbf{x}_i\|_2 \\ &\leq \ell(\|\mathbf{w}\|_2\|\mathbf{x}_i\|_2)^{\ell-1} \\ &= \ell\|\mathbf{w}\|_2^{\ell-1} \\ &\leq \ell\kappa^{\ell-1}. \end{aligned} \tag{199}$$

Then we have

$$\begin{aligned} \left|\sigma\left(\left(\mathbf{w}_k(t)-\eta\frac{\partial L(\mathbf{W}(t))}{\partial \mathbf{w}_k(t)}\right)\mathbf{x}_i\right)-\sigma(\mathbf{w}_k(t)\mathbf{x}_i)\right| &\leq \sup_{\|\mathbf{w}\|_2\leq\kappa}\|\nabla_{\mathbf{w}}\sigma(\mathbf{w}\mathbf{x}_i)\|\cdot\left\|\eta\frac{\partial L(\mathbf{W}(t))}{\partial \mathbf{w}_k(t)}\right\| \\ &\leq \ell\kappa^{\ell-1}\left\|\eta\frac{\partial L(\mathbf{W}(t))}{\partial \mathbf{w}_k(t)}\right\| \\ &= \ell\kappa^{\ell-1}\eta\left\|\frac{\partial L(\mathbf{W}(t))}{\partial \mathbf{w}_k(t)}\right\|, \end{aligned} \tag{200}$$

and furthermore, the bound of $|u_i(t+1)-u_i(t)|$ can be estimated as

$$\begin{aligned} |u_i(t+1)-u_i(t)| &\leq \frac{1}{\sqrt{m}}\sum_{k=1}^{m}\left|\sigma\left(\left(\mathbf{w}_k(t)-\eta\frac{\partial L(\mathbf{W}(t))}{\partial \mathbf{w}_k(t)}\right)\mathbf{x}_i\right)-\sigma(\mathbf{w}_k(t)\mathbf{x}_i)\right| \\ &\leq \frac{\ell\kappa^{\ell-1}\eta}{\sqrt{m}}\sum_{k=1}^{m}\left\|\frac{\partial L(\mathbf{W}(t))}{\partial \mathbf{w}_k(t)}\right\|. \end{aligned} \tag{201}$$

Recall that the partial of L with respect to $\mathbf{w}_k(t)$ is

$$\frac{\partial L(\mathbf{W}(t))}{\partial \mathbf{w}_k(t)} = \frac{1}{\sqrt{m}}\sum_{i=1}^{n}(u_i(t)-y_i)a_k\zeta(\mathbf{w}_k(t)\mathbf{x}_i)\mathbf{x}_i$$

where

$$\zeta(\mathbf{w}_k(t)\mathbf{x}_i) = \ell(\mathbf{w}_k(t)\mathbf{x}_i)^{\ell-1}\mathbb{I}\{\mathbf{w}_k(t)\mathbf{x}_i \geq 0\},$$

and then the norm of $\frac{\partial L(\mathbf{W}(t))}{\partial \mathbf{w}_k(t)}$ can be bounded by

$$\begin{aligned}
\left\|\frac{\partial L(\mathbf{W}(t))}{\partial \mathbf{w}_k(t)}\right\| &= \left\|\frac{1}{\sqrt{m}}\sum_{i=1}^{n}(u_i(t)-y_i)a_k\zeta(\mathbf{w}_k(t)\mathbf{x}_i)\mathbf{x}_i\right\| && (202)\\
&\leq \frac{1}{\sqrt{m}}\sum_{i=1}^{n}|u_i(t)-y_i|\cdot|\zeta(\mathbf{w}_k(t)\mathbf{x}_i)|\cdot\|\mathbf{x}_i\| && (203)\\
&\leq \frac{1}{\sqrt{m}}\sum_{i=1}^{n}|u_i(t)-y_i|\,\ell\kappa^{\ell-1} && (204)\\
&= \frac{\ell\kappa^{\ell-1}}{\sqrt{m}}\sum_{i=1}^{n}|u_i(t)-y_i|\,. && (205)
\end{aligned}$$

Therefore,

$$\begin{aligned}
\|\mathbf{u}(t+1)-\mathbf{u}(t)\|_2^2 &= \sum_{i=1}^{n}|u_i(t+1)-u_i(t)|^2 && (206)\\
&\leq \sum_{i=1}^{n}\left(\frac{\ell\kappa^{\ell-1}\eta}{\sqrt{m}}\sum_{k=1}^{m}\left\|\frac{\partial L(\mathbf{W}(t))}{\partial \mathbf{w}_k(t)}\right\|\right)^2 && (207)\\
&= \frac{\ell^2\kappa^{2\ell-2}\eta^2}{m}\sum_{i=1}^{n}\left(\sum_{k=1}^{m}\frac{\ell\kappa^{\ell-1}}{\sqrt{m}}\sum_{i=1}^{n}|u_i(t)-y_i|\right)^2 && (208)\\
&\leq \frac{\ell^2\kappa^{2\ell-2}\eta^2}{m}\sum_{i=1}^{n}\left(\sum_{k=1}^{m}\frac{\ell\kappa^{\ell-1}}{\sqrt{m}}\sqrt{n}\,\|\mathbf{u}(t)-\mathbf{y}\|_2\right)^2 && (209)\\
&= \ell^4\kappa^{4\ell-4}\eta^2 n^2\,\|\mathbf{u}(t)-\mathbf{y}\|_2^2\,. && (210)
\end{aligned}$$

The proof completes. □

For the rest of the proof of the theorem, we denote

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

and

$$S_i = \{r \in [m] : \mathbb{I}\{A_{ir} = 0\}\}$$
$$S_i^{\perp} = [m] \setminus S_i$$

where R is the radius chosen based on the argument of continuous time gradient flow.

Lemma B.12. *Given $\delta \in (0,1)$, it holds that*

$$\left\|H_{ij}^{\perp}(t)\right\|_2 \leq \frac{\ell^2\kappa^{2\ell-2}n^2R}{\delta}$$

with probability at least $1-\delta$.

Proof. Since $H_{ij}^{\perp}(t)$ has the following form,

$$H_{ij}^{\perp}(t) = \frac{1}{m}\langle\mathbf{x}_i,\mathbf{x}_j\rangle\sum_{k\in S_i^{\perp}}\ell^2(\mathbf{w}_k(t)\mathbf{x}_i)^{\ell-1}(\mathbf{w}_k(t)\mathbf{x}_j)^{\ell-1}\mathbb{I}\{\mathbf{w}_k(t)\mathbf{x}_i \geq 0, \mathbf{w}_k(t)\mathbf{x}_j \geq 0\}$$

and then

$$\left|H_{ij}^{\perp}(t)\right| \leq \frac{\ell^2\kappa^{2\ell-2}}{m}\left|S_i^{\perp}\right|.$$

Since $\mathbb{E}[\left|S_i^{\perp}\right|] \leq \frac{2mR}{\sqrt{2\pi}}$, and it holds that

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

by Markov inequality, for a given $\delta \in (0,1)$, as long as $a > \frac{2mnR}{\sqrt{2\pi}\delta}$, it holds that

$$P\left(\sum_{i=1}^{n}\left|S_i^{\perp}\right| < a\right) \geq 1-\delta,$$

and for convenience we can let $a = \frac{mnR}{\delta}$ for each given δ. Moreover,

$$\left\|H^{\perp}(t)\right\|_2 \leq \sum_{j=1}^{n}\sum_{i=1}^{n}\left|H_{ij}^{\perp}(t)\right| \leq \frac{\ell^2\kappa^{2\ell-2}}{m}\sum_{j=1}^{n}\sum_{i=1}^{n}\left|S_i^{\perp}\right| \leq \frac{\ell^2\kappa^{2\ell-2}n}{m}\cdot\frac{mnR}{\delta} = \frac{\ell^2\kappa^{2\ell-2}n^2R}{\delta}$$

with probability at least $1-\delta$. □

Denote I_2^i as follows,

$$I_2^i = \frac{1}{\sqrt{m}}\sum_{k\in S_i^{\perp}} a_k\left(\sigma\left(\left(\mathbf{w}_k(t) - \eta\frac{\partial L(\mathbf{W}(t))}{\partial \mathbf{w}_k(t)}\right)\mathbf{x}_i\right) - \sigma(\mathbf{w}_k(t)\mathbf{x}_i)\right)$$

and we can have the estimate of I_2^i below,

$$\left|I_2^i\right| \leq \frac{1}{\sqrt{m}}\sum_{k\in S_i^{\perp}}\ell\kappa^{\ell-1}\eta\max_{k\in[m]}\left\|\frac{\partial L(\mathbf{W}(t))}{\partial \mathbf{w}_k(t)}\right\| = \frac{\ell\kappa^{\ell-1}\eta\left|S_i^{\perp}\right|}{\sqrt{m}}\max_{k\in[m]}\left\|\frac{\partial L(\mathbf{W}(t))}{\partial \mathbf{w}_k(t)}\right\|$$

Since

$$\left\|\frac{\partial L(\mathbf{W}(t))}{\partial \mathbf{w}_k(t)}\right\| \leq \frac{\ell\kappa^{\ell-1}}{\sqrt{m}}\sum_{i}^{n}\left|u_i(t)-y_i\right| \leq \frac{\ell\kappa^{\ell-1}\sqrt{n}}{\sqrt{m}}\left\|\mathbf{u}(t)-\mathbf{y}\right\|_2,$$

we can have furthermore the bound of I_2^i as follows,

$$\left|I_2^i\right| \leq \frac{\ell^2\kappa^{2\ell-2}\eta\left|S_i^{\perp}\right|\sqrt{n}}{m}\left\|\mathbf{u}(t)-\mathbf{y}\right\|_2.$$

To finish the proof of the theorem, we combine the bounds all together,

$$\begin{aligned}\left\|\mathbf{u}(t+1)-\mathbf{y}\right\|_2^2 = &\left\|\mathbf{y}-\mathbf{u}(t)\right\|_2^2 - 2\eta(\mathbf{y}-\mathbf{u}(t))^{\top}H(t)(\mathbf{y}-\mathbf{u}(t)) &(211)\\ &+ 2\eta(\mathbf{y}-\mathbf{u}(t))^{\top}H(t)^{\perp}(\mathbf{y}-\mathbf{u}(t)) &(212)\\ &- 2(\mathbf{y}-\mathbf{u}(t))^{\top}\mathbf{I}_2 + \left\|\mathbf{u}(t+1)-\mathbf{u}(t)\right\|_2^2. &(213)\end{aligned}$$

Note that

$$(\mathbf{y}-\mathbf{u}(t))^{\top}\mathbf{I}_2 \leq \left\|\mathbf{y}-\mathbf{u}(t)\right\|_2\left\|\mathbf{I}_2\right\|_2 \leq \left\|\mathbf{y}-\mathbf{u}(t)\right\|_2\left\|\mathbf{I}_2\right\|_1 = \left\|\mathbf{y}-\mathbf{u}(t)\right\|_2\sum_{i=1}^{n}\left|I_2^i\right|$$

and

$$\sum_{i=1}^{n}\left|I_2^i\right| \leq \ell^2\kappa^{2\ell-2}\eta\sqrt{n}\left\|\mathbf{y}-\mathbf{u}(t)\right\|_2\frac{\sum_{i=1}^{n}\left|S_i^{\perp}\right|}{m}$$

where

$$\sum_{i=1}^{n}\left|S_i^{\perp}\right| \leq \frac{mnR}{\delta}$$

with probability at least $1-\delta$. Therefore, it holds that (with probability $1-\delta$)

$$(\mathbf{y}-\mathbf{u}(t))^{\top}\mathbf{I}_2 \leq \frac{\ell^2\kappa^{2\ell-2}\eta n^{\frac{3}{2}}R}{\delta}\left\|\mathbf{y}-\mathbf{u}(t)\right\|_2^2.$$

Thus we have

$$\begin{aligned}\left\|\mathbf{u}(t+1)-\mathbf{y}\right\|_2^2 &\leq \left(1-\eta\lambda_0+\frac{2\ell^2\kappa^{2\ell-2}\eta n^2 R}{\delta}+\frac{2\ell^2\kappa^{2\ell-2}\eta n^{3/2}R}{\delta}+\ell^4\kappa^{4\ell-4}\eta^2 n^2\right)\left\|\mathbf{y}-\mathbf{u}(t)\right\|_2^2 && (214)\\ &\leq \left(1-\frac{\eta\lambda_0}{2}\right)\left\|\mathbf{y}-\mathbf{u}(t)\right\|_2^2 && (215)\end{aligned}$$

for properly chosen R and η.

Having proven that gradient descent provably optimizes over-parametrized neural networks with RePU activation functions, the optimization theory for higher order linear PDE follows almost immediately. We complete the proof of Theorem 3.1 as follows.

Proof. The proof is almost immediate. For a specific partition α, direct calculation on the partial derivative $\frac{\partial^{|\alpha|}f}{\partial x_1^{\alpha_1}\partial x_2^{\alpha_2}...\partial x_d^{\alpha_d}}$ gives following

$$D^{\alpha}f=\frac{1}{\sqrt{m}}\sum_{k=1}^{m}a_k w_{k1}^{\alpha_1}w_{k2}^{\alpha_2}...w_{kd}^{\alpha_d}\sigma^{(|\alpha|)}(\mathbf{w}_k\mathbf{x})$$

where $\sigma^{(|\alpha|)}$ denotes the derivative of activation function till the order of $|\alpha|$. In the following context, we use notation

$$W_{k\alpha}:=w_{k1}^{\alpha_1}w_{k2}^{\alpha_2}...w_{kd}^{\alpha_d}$$

Therefore, summing over all α with $|\alpha|=r$, we have that

$$\begin{aligned}\sum_{|\alpha|=r}D^{\alpha}f &= \sum_{|\alpha|=r}\frac{1}{\sqrt{m}}\sum_{k=1}^{m}a_k W_{k\alpha}\sigma^{(|\alpha|)}(\mathbf{w}_k\mathbf{x})\\ &= \frac{1}{\sqrt{m}}\sum_{k=1}^{m}\sum_{|\alpha|=r}a_k W_{k\alpha}\sigma^{(|\alpha|)}(\mathbf{w}_k\mathbf{x}) && (216)\\ &= \frac{1}{\sqrt{m}}\sum_{k=1}^{m}a_k\left(\sum_{|\alpha|=r}W_{k\alpha}\right)\sigma^{(|\alpha|)}(\mathbf{w}_k\mathbf{x}).\end{aligned}$$

Considering $b_k:=a_k\sum_{|\alpha|=r}W_{k\alpha}$ as a whole parameter, we can obtain the convergence result from Theorem B.1. □