A review of the article *Gradient Descent Provably Optimizes Over-parametrized Neural Networks* by Du et al.

Elliot Epstein[1,2]

[1]School of Engineering Sciences, KTH Royal Institute of Technology, Sweden, e-mail: *elliote@kth.se*
[2]Department of Mathematics, ETH Zurich, Switzerland, e-mail: *eepstein@student.ethz.ch*

May 27, 2023

Abstract

Training a two layer neural network with gradient descent ensures, with high probability over the initialisation of the network weights, that the mean squared error training loss converges to zero, given that the neural network is polynomially over-parametrized. The convergence rate is bounded by the step size and the least eigenvalue of a matrix which depends on the input data but not on the initialisation of the network weights. Convergence is proved as long as this matrix in positive definite, which is the case as long as no input data are parallel. This is used to directly analyse the dynamics of the predictions, as opposed to the weights. The problem is first analysed with gradient flow, for training the first layer and subsequently for training both layers. Using the results from the continuous setting, the case with gradient descent is proved. Numerical simulations on synthetic data supports the findings. The analysis will closely follow [Simon S. Du et al. *Gradient Descent Provably Optimizes Over-parameterized Neural Networks*. 2018. arXiv: `1810.02054 [cs.LG]`].

Acknowledgement

I want to acknowledge my supervisor Prof. Arnulf Jentzen for his helpful assistance in writing this bachelor's thesis.

Contents

1 Introduction

Recent developments, such as the ImageNet project [5], has made large sets of high quality data readily available. GPUs has also seen large improvements in computing power recently [18]. Within this framework, deep learning has seen success in many areas, such as image classification and pattern recognition, where deep learning models have had a performance comparable to experts [10], highlighting the importance of these advances. Empirically it has been clear that the training loss (2) is converging to zero when trained by gradient descent (65) for certain neural network configurations. However, it has not been well understood theoretically as to why this happens. The goal of this paper is to showcase why a neural network with one hidden layer (1) can achieve a linear convergence rate for the loss function, by training with gradient descent, given that the over-parametrization of the network is large enough. This is considered a stepping stone towards a deeper understanding of the success of deep neural networks, which in the future might provide theoretical hints as how to design neural networks in an effective way.

First, the problem is analysed with an infinitesimal step size, i.e. gradient flow (26). This provides key insights into the structure of the matrices that are critical for the dynamics of the predictions. These insights are later used in the analysis for the gradient descent algorithm (65). Furthermore, a section on the gradient flow for training both the output layer and the hidden layer is provided to showcase the ability of the method to generalise. Several numerical experiments, with the full code available, are also made to give empirical insight into the proved theorems.

The analysis will closely follow the one made in [7]. Even stronger results has very recently been published [8], which furthers the analysis done in [7] for deep neural networks, by applying a similar, albeit more advanced, analysis technique.

1.1 Setting

Throughout this thesis the following setting will often be assumed.

Setting 1.1. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $d, m, n \in \mathbb{N} = \{1, 2, 3, \dots\}$, $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n \in \mathbb{R}^d$, $y = (y_1, y_2, \dots, y_n) \in \mathbb{R}^n$, let $a = (a_1, a_2, \dots, a_m) \colon [0, \infty) \times \Omega \to \mathbb{R}^m$, let for all $r \in \{1, 2, \dots, m\}$ $\mathbf{w}_r \colon [0, \infty) \times \Omega \to \mathbb{R}^d$, let $f = (f_1, f_2, \dots, f_n) \colon [0, \infty) \times \Omega \to \mathbb{R}^n$ satisfy for all $i \in \{1, 2, \dots, n\}$, $t \in [0, \infty)$, $\omega \in \Omega$ that*

$$f_i(t, \omega) = \frac{1}{\sqrt{m}} \sum_{r=1}^{m} a_r(t, \omega) \sigma((\mathbf{w}_r(t, \omega))^* \mathbf{x}_i), \tag{1}$$

let $L \colon [0, \infty) \times \Omega \to \mathbb{R}$ satisfy for all $t \in [0, \infty)$, $\omega \in \Omega$ that

$$L(t, \omega) = \frac{1}{2} \sum_{i=1}^{n} (f_i(t, \omega) - y_i)^2, \tag{2}$$

let $\mathbf{v}\colon \Omega \to \mathbb{R}^d$ *be a standard normal distributed random vector, let* $\mathfrak{H} = (\mathfrak{h}_{i,j})_{i,j\in\{1,2,\ldots,n\}} \in \mathbb{R}^{n\times n}$ *satisfy for all* $i,j \in \{1,2,\ldots,n\}$ *that*

$$\mathfrak{h}_{i,j} = \mathbb{E}[\mathbf{x}_i^*\mathbf{x}_j \mathbb{1}_{\{\mathbf{v}^*\mathbf{x}_i\geq 0, \mathbf{v}^*\mathbf{x}_j\geq 0\}}] \in \mathbb{R}, \tag{3}$$

let $\|\cdot\|_2\colon \mathbb{R}^n \to \mathbb{R}$ *be the Euclidean norm on* $\mathbb{R}^n$*, let* $\|\cdot\|_1\colon \mathbb{R}^n \to \mathbb{R}$ *satisfy for all* $b = (b_1, b_2, \ldots, b_n) \in \mathbb{R}^n$ *that* $\|b\|_1 = \sum_{i=1}^n |b_i|$*, let* λ_0 *be the smallest eigenvalue of* $\mathfrak{H}$*, assume for all* $j, i \in \{0,1,2,\ldots,n\}$*,* $i \neq j$ *that* $|\mathbf{x}_i^*\mathbf{x}_j| < \|\mathbf{x}_i\|_2\|\mathbf{x}_j\|_2$*, let* $\sigma\colon \mathbb{R}\to\mathbb{R}$ *satisfy for all* $z \in \mathbb{R}$ *that*

$$\sigma(z) = \max(0, z), \tag{4}$$

and let $\mathbf{H} = (\mathbf{h}_{i,j})_{i,j\in\{1,2,\ldots,n\}}\colon [0,\infty)\times\Omega \to \mathbb{R}^{n\times n}$ *satisfy for all* $i,j \in \{0,1,\ldots,n\}$*,* $t \in [0,\infty)$*,* $\omega \in \Omega$ *that*

$$\mathbf{h}_{i,j}(t,\omega) = \frac{1}{m}\mathbf{x}_i^*\mathbf{x}_j \sum_{r=1}^m (a_r(t,\omega))^2 \mathbb{1}_{\{(\mathbf{w}_r(t,.))^*\mathbf{x}_i\geq 0, (\mathbf{w}_r(t,.))^*\mathbf{x}_j\geq 0\}}(\omega). \tag{5}$$

1.2 Related articles

A related article is [15]. There, multiclass classification for a two layer neural network is considered, trained by stochastic gradient descent (SGD). The loss function considered is the Cross Entropy loss with soft-max activation. There, the major assumption on the input data is that the over-parametrization depends polynomially on $\frac{1}{\delta'}$, where δ' is the minimal distance between each of the inputs. The article show that if $m \geq M = poly(\frac{1}{\epsilon}, \frac{1}{\delta'}, B, T)$, with B being the batch size and T the number of iterations with SGD, then on an event with high probability it will hold that after T' iterations the training loss will be smaller than a positive constant. However, the article never shows that the training loss will converge to zero for an infinite time. One strength with this article is that the analysis is based on SGD which is an algorithm often used in practice. Other articles which has considered the dynamics of the predictions directly include [19] and [17]. Similar articles that assume a normal input distribution include [2].

Articles that has analysed the optimisation landscape include [11] and [6] .

In [21] matrices related to $\mathfrak{H}$ were analysed to guarantee convergence of the training loss. Related articles where the considered activation function σ was different include [16], [14], and [20].

Similar results in the setting of deep networks include [12]. In [3] the problem was considered using tools from optimal transport. Other articles establishing guarantees for gradient descent training neural networks include [4].

2 Continuous time analysis

In this section the gradient flow algorithm (26) for training the hidden weights of the neural network 1 will be considered. We will first prove that if no two input data are parallel, the matrix $\mathfrak{H}$ (3) will have a positive least eigenvalue. This quantity will be central in determining a bound on the convergence rate. We also prove an upper bound on this least eigenvalue, this will simplify many expressions related to the convergence rate of the training loss. Then, we show that $\mathbf{H}(t,\omega)$, which controls the dynamics of the predictions $f(t,\omega)$, will remain close to $\mathfrak{H}$ and the weights will remain close to the initialisation. This will ensure that under certain conditions, the training loss converges

to zero, i.e., for all times $t \in [0, \infty)$, and all ω in an event which hold true with high probability it holds that

$$L(t, \omega) \leq L(0, \omega) \exp(-t\lambda_0). \tag{6}$$

In this section we will follow the article [7] .

Remark 2.1. *The major assumption from Setting 1.1 that is used to prove Theorem 2.2, is that no two input vectors are parallel.*

Theorem 2.2. *Assume Setting 1.1. Then it holds that*

$$\lambda_0 > 0. \tag{7}$$

Proof of Theorem 2.2. Let $\mathbf{c} \in \mathbb{R}^d$,
let $\phi_{\mathbf{x}} \colon \mathbb{R}^d \to \mathbb{R}^d$ satisfy for all $\mathbf{x} \in \mathbb{R}^d$, $\mathbf{c} \in \mathbb{R}^d$ that

$$\phi_{\mathbf{x}}(\mathbf{c}) = \begin{cases} \mathbf{x} & \mathbf{x}^* \mathbf{c} \geq 0 \\ \mathbf{0} & \mathbf{x}^* \mathbf{c} < 0, \end{cases} \tag{8}$$

note that

$$\mathfrak{h}_{i,j} = \mathbb{E}[(\phi_{\mathbf{x}_i}(\mathbf{v}))^* \phi_{\mathbf{x}_j}(\mathbf{v})], \tag{9}$$

let $\mathcal{H}$ be the real vector space of integrable functions $f \colon \mathbb{R}^d \to \mathbb{R}^d$ with the inner product $\langle f, g \rangle_{\mathcal{H}} = \mathbb{E}[(f(\mathbf{v}))^* g(\mathbf{v})]$, let $\alpha_1, \alpha_2, \ldots, \alpha_n \in \mathbb{R}$, and assume that

$$\left\| \sum_{j=1}^{n} \phi_{\mathbf{x}_j}(\mathbf{w}) \alpha_j \right\|_{\mathcal{H}}^2 = 0. \tag{10}$$

Lemma 2.3. *Let $i \in \{1, 2, \ldots, n\}$ and let $D_i = \{\mathbf{c} \in \mathbb{R}^d \colon \mathbf{c}^* \mathbf{x}_i = 0\}$. Then it holds that $D_i \not\subset \bigcup_{j \in \{1, \ldots, n\} \colon j \neq i} D_j$.*

Proof of Lemma 2.3. Let $i \in \{0, 1, 2, \ldots, n\}$, let μ_i be the canonical Lebesgue measure on D_i, and let for all $j \in \{0, 1, 2, \ldots, n\}$, $j \neq i$, $A_{i,j}$ be the $2 \times d$ matrix given by

$$A_{i,j} = \begin{bmatrix} — & \mathbf{x}_i & — \\ — & \mathbf{x}_j & — \end{bmatrix}. \tag{11}$$

Note that $D_i \bigcap D_j = \{\mathbf{c} \in \mathbb{R}^d \colon \mathbf{c} \in \ker(A_{i,j})\}$. This, and the assumption for all $j \in \{0, 1, 2, \ldots, n\}$, $j \neq i$ that $|\mathbf{x}_i^* \mathbf{x}_j| < \|\mathbf{x}_i\|_2 \|\mathbf{x}_j\|_2$ ensures that $\dim(D_i \bigcap D_j) = d - 2$. Combining this, $\dim(D_i) = d - 1$, and $D_i \bigcap D_j \subset D_i$ ensure that for all $j \in \{0, 1, 2, \ldots, n\}$, $j \neq i$ that $\mu_i(D_i \bigcap D_j) = 0$. Hence, we obtain that

$$\sum_{i \neq j} \mu(D_i \cap D_j) = 0. \tag{12}$$

This and Theorem 7.5 ensures that

$$\begin{aligned} \mu_i\Big(D_i \bigcap (\cup_{j \in \{1, \ldots, n\} \colon j \neq i} D_j)\Big) &= \mu_i\left(\bigcup_{j \in \{1, \ldots, n\} \colon j \neq i} (D_i \cap D_j) \right) \\ &\leq \sum_{j \in \{1, \ldots, n\} \colon j \neq i} \mu_i(D_i \cap D_j) = 0. \end{aligned} \tag{13}$$

Hence, we obtain that $D_i \not\subset \cup_{j \in \{1, \ldots, n\} \colon j \neq i} D_j$. The proof of Lemma 2.3 is thus completed. □

Let $i \in \{1, 2, \ldots, n\}$. Lemma 2.3 now ensures that the set $D_i \setminus \cup_{j \in \{1,\ldots,n\}\colon j \neq i} D_j$ is not empty. Hence, let $\mathbf{z} \in D_i \setminus \cup_{j \in \{1,\ldots,n\}\colon j \neq i} D_j$, let $B(\mathbf{z}, r) = \{\mathbf{q} \in \mathbb{R}^d \colon |\mathbf{z} - \mathbf{q}| < r\}$, let $B_i(\mathbf{z}, r)^+ = B(\mathbf{z}, r) \cap \{\mathbf{c} \in \mathbb{R}^d \colon \mathbf{c}^* \mathbf{x}_i \geq 0\}$, let $B_i(\mathbf{z}, r)^- = B(\mathbf{z}, r) \cap \{\mathbf{c} \in \mathbb{R}^d \colon \mathbf{c}^* \mathbf{x}_i < 0\}$, let μ be the canonical Lebesgue measure on $\mathbb{R}^d$, and let $r > 0$ small enough to ensure that $B(\mathbf{z}, r) \cap D_j = \varnothing$. Moreover, note that for $j \neq i$ it hols that $\phi_{\mathbf{x_j}}(\mathbf{c})$ is continuous with respect to $\mathbf{c}$ in an neighbourhood of $\mathbf{z}$. This implies for all $j \in \{1, \ldots, n\}$, : $j \neq i$, a given $\epsilon > 0$ and $r(\epsilon) \in (0, \infty)$ sufficiently small that

$$\left| \frac{1}{\mu(B_i(\mathbf{z}, r)^+)} \int_{B_i(\mathbf{z},r)^+} \phi_{\mathbf{x}_j}(\mathbf{c}) - \phi_{\mathbf{x}_j}(\mathbf{z}) d\mathbf{c} \right| \leq \frac{1}{\mu(B_i(\mathbf{z}, r)^+)} \int_{B_i(\mathbf{z},r)^+} \left| \phi_{\mathbf{x}_j}(\mathbf{c}) - \phi_{\mathbf{x}_j}(\mathbf{z}) \right| d\mathbf{c} \leq \epsilon, \tag{14}$$

and

$$\left| \frac{1}{\mu(B_i(\mathbf{z}, r)^-)} \int_{B_i(\mathbf{z},r)^-} \phi_{\mathbf{x}_j}(\mathbf{c}) - \phi_{\mathbf{x}_j}(\mathbf{z}) d\mathbf{c} \right| \leq \frac{1}{\mu(B_i(\mathbf{z}, r)^-)} \int_{B_i(\mathbf{z},r)^-} \left| \phi_{\mathbf{x}_j}(\mathbf{c}) - \phi_{\mathbf{x}_j}(\mathbf{z}) \right| d\mathbf{c} \leq \epsilon. \tag{15}$$

(14) hence implies that

$$\lim_{r \to 0^+} \frac{1}{\mu(B_i(\mathbf{z}, r)^+)} \int_{B_i(\mathbf{z},r)^+} \phi_{\mathbf{x}_j}(\mathbf{c}) d\mathbf{c} = \phi_{\mathbf{x}_j}(\mathbf{z}), \tag{16}$$

and (15) implies that

$$\lim_{r \to 0^+} \frac{1}{\mu(B_i(\mathbf{z}, r)^-)} \int_{B_i(\mathbf{z},r)^-} \phi_{\mathbf{x}_j}(\mathbf{c}) d\mathbf{c} = \phi_{\mathbf{x}_j}(\mathbf{z}). \tag{17}$$

Next, observe that it holds that

$$\lim_{r \to 0^+} \frac{1}{\mu(B_i(\mathbf{z}, r)^+)} \int_{B_i(\mathbf{z},r)^+} \phi_{\mathbf{x}_i}(\mathbf{c}) d\mathbf{c} = \mathbf{x}_i, \tag{18}$$

and that

$$\lim_{r \to 0^+} \frac{1}{\mu(B_i(\mathbf{z}, r)^-)} \int_{B_i(\mathbf{z},r)^-} \phi_{\mathbf{x}_i}(\mathbf{c}) d\mathbf{c} = 0. \tag{19}$$

Note that (10) ensures that $\mathbb{P}$-a.s. it holds that

$$\sum_{j=1}^{n} \alpha_j \phi_{\mathbf{x}_j}(\mathbf{w}) = 0. \tag{20}$$

This implies that $\sum_{j=1}^{n} \alpha_j \phi_{\mathbf{x}_j}(\mathbf{c}) = 0$ for almost all $\mathbf{c} \in \mathbb{R}^d$. Combining this (16), (17), (18), and (19) ensures that

$$\begin{aligned} 0 = \lim_{r \to 0^+} \frac{1}{\mu(B_i(\mathbf{z}, r)^+)} \int_{B_i(\mathbf{z},r)^+} \sum_{j=1}^{n} \alpha_j \phi_{\mathbf{x}_j}(\mathbf{c}) d\mathbf{c} \\ - \lim_{r \to 0^+} \frac{1}{\mu(B_i(\mathbf{z}, r)^-)} \int_{B_i(\mathbf{z},r)^-} \sum_{j=1}^{n} \alpha_j \phi_{\mathbf{x}_j}(\mathbf{c}) d\mathbf{c} = \mathbf{x}_i \alpha_i. \end{aligned} \tag{21}$$

This, the fact that $i \in \{0, 1, 2, \ldots, n\}$ was arbitrarily picked, and for all $i \in \{1, 2, \ldots, n\}$ it holds that $\mathbf{x}_i \neq \mathbf{0}$ ensures that

$$\alpha_1 = ... = \alpha_n = 0. \tag{22}$$

This implies for all $\mathbf{z} \in \mathbb{R}^d \backslash \{\mathbf{0}\}$ that

$$\begin{aligned} \mathbf{z}^* \mathfrak{H} \mathbf{z} &= \langle \sum_{i=1}^{n} \phi_{\mathbf{x}_i} z_i, \sum_{j=1}^{n} \phi_{\mathbf{x}_j} z_j \rangle_{\mathcal{H}} \\ &= \| \sum_{j=1}^{n} \phi_{\mathbf{x}_j} z_j \|_{\mathcal{H}}^2 > 0. \end{aligned} \tag{23}$$

This establishes that $\mathfrak{H}$ is positive definite. This implies that $\lambda_0 = \lambda_{min}(\mathfrak{H}) > 0$. The proof of Theorem 2.2 is thus completed.

□

Remark 2.4. *Note that the assumption in Setting 1.1 that no two input vectors are parallel will hold in many real world data-sets. When the Setting 1.1 is assumed, it is understood that $\lambda_0 > 0$.*

Theorem 2.5. *Assume Setting 1.1 and assume for all $i \in \{0, 1, 2, \ldots, n\}$ that $\|\mathbf{x}_i\|_2 = 1$. Then it holds that*

$$\lambda_0 \leq n. \tag{24}$$

Proof of Theorem 2.5 . Let $\mathbf{z} \in \mathbb{R}^n$, and note that the assumptions in the theorem, Lemma 7.6, and Lemma 7.9 ensures that

$$\begin{aligned} \lambda_0 \|\mathbf{z}\|_2^2 &\leq \sum_{i=1}^{n} \sum_{j=1}^{n} \big| \mathbb{E}[(\phi_{\mathbf{x}_i}(\mathbf{v}))^* \phi_{\mathbf{x}_j}(\mathbf{v})] \big| |z_i| |z_j| \\ &\leq \|\mathbf{z}\|_1^2 \leq n \|\mathbf{z}\|_2^2. \end{aligned} \tag{25}$$

The proof of Theorem 2.5 is thus completed. □

Theorem 2.6 (Convergence Rate of Gradient Flow)**.** *Assume Setting 1.1, let $C \in (0, \infty)$, assume for all $i \in \{0, 1, 2, \ldots, n\}$ that $\|\mathbf{x}_i\|_2 = 1$, assume for all $i \in \{0, 1, 2, \ldots, n\}$ that $|y_i| \leq C$, assume for all $r \in \{0, 1, 2, \ldots, m\}$ that $\mathbf{w}_r(0, .) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $a_r(0, .) \sim \text{Unif}[\{-1, 1\}]$ are i.i.d. random vectors, assume for all $t \in (0, \infty)$, $\omega \in \Omega$ that $a(t, \omega) = a(0, \omega)$, let $K = \frac{2^2 16^2 (C^2+1)^3}{2\pi}$, assume that $m \geq \frac{K n^6}{\lambda_0^3 \delta^3}$, let $\delta \in (0, 1)$, assume for all $r \in \{1, 2, \ldots, m\}$, $t \in [0, \infty)$, $\omega \in \Omega$ that*

$$\frac{d\mathbf{w}_r(t, \omega)}{dt} = -\frac{1}{\sqrt{m}} \sum_{i=1}^{n} (f_i(t, \omega) - y_i) a_r \mathbf{x}_i \mathbb{1}_{\{(\mathbf{w}_r(t,.))^* \mathbf{x}_i \geq 0\}}(\omega), \tag{26}$$

let $\delta_1 = \delta_2 = \delta_3 = \frac{\delta}{3}$, let $R = \frac{\sqrt{2\pi} \delta_2 \lambda_0}{16 n^2}$, let $B_3 = \{\omega \in \Omega \colon \|y - f(0, \omega)\|_2^2 \leq \frac{n(C^2 + \frac{1}{2})}{\delta_3}\} \in \mathcal{F}$, let $B_0(\omega) = \big\{\hat{\mathbf{w}}_1, \hat{\mathbf{w}}_2, \ldots, \hat{\mathbf{w}}_m \in \mathbb{R}^d \colon \max_{r \in \{1,2,\ldots,m\}} \|\mathbf{w}_r(0, \omega) - \hat{\mathbf{w}}_r(\omega)\|_2 \leq R\big\}$, let $\hat{\mathbf{h}}_{i,j}$ satisfy for all $i, j \in \{1, 2, \ldots, n\}$, $\hat{\mathbf{w}}_1, \hat{\mathbf{w}}_2, \ldots, \hat{\mathbf{w}}_n \in \mathbb{R}^d$ that $\hat{\mathbf{h}}_{i,j} = \frac{1}{m} \mathbf{x}_i^* \mathbf{x}_j \sum_{r=1}^{m} \mathbb{1}_{\{(\hat{\mathbf{w}}_r)^* \mathbf{x}_i \geq 0, (\hat{\mathbf{w}}_r)^* \mathbf{x}_j \geq 0\}}$, let $B_2 = \Big\{\omega \in \Omega \colon \sum_{j,i=1}^{n,n} \sup_{\hat{\mathbf{w}}_1, \hat{\mathbf{w}}_2, \ldots, \hat{\mathbf{w}}_m \in B_0(\omega)} \Big\{ \Big| \hat{\mathbf{h}}_{i,j} - \mathbf{h}_{i,j}(0, \omega) \Big| \Big\} \leq \frac{4Rn^2}{\delta_2 \sqrt{2\pi}} \Big\} \in \mathcal{F}$, and let $B_1 = \bigcap_{i,j=1}^{n,n} \Big\{ \omega \in \Omega \colon |\mathbf{h}_{i,j}(0, \omega) - \mathfrak{h}_{i,j}| \leq \sqrt{\frac{1}{2m} \log\Big(\frac{2n^2}{\delta_1}\Big)} \Big\} \in \mathcal{F}$. Then

(i) it holds that $\mathbb{P}(B_1 \cap B_2 \cap B_3) \geq 1 - \delta$

(ii) it holds for all $t \in [0, \infty)$, $\omega \in B_1 \cap B_2 \cap B_3$ that

$$\|f(t, \omega) - y\|_2^2 \leq \exp(-\lambda_0 t) \|f(0, \omega) - y\|_2^2. \tag{27}$$

Proof of Theorem 2.6. Note that the assumptions of Theorem 2.6, simplifies $\mathbf{H} = (\mathbf{h}_{i,j})_{i,j\in\{1,2,\dots,n\}}$ from Setting 1.1 for all $i,j \in \{0,1,2,\dots,n\}$, $t \in [0,\infty)$, $\omega \in \Omega$ to

$$\mathbf{h}_{i,j}(t,w) = \sum_{r=1}^{m} \mathbf{x}_i^* \mathbf{x}_j \frac{1}{m} \mathbb{1}_{\{(\mathbf{w}_r(t,.))^*\mathbf{x}_i \geq 0, (\mathbf{w}_r(t,.))^*\mathbf{x}_j \geq 0\}}(\omega). \tag{28}$$

Next observe that for all $t \in [0,\infty)$, $\omega \in \Omega$ it holds that

$$\begin{aligned}
\frac{d}{dt} f_i(t,\omega) &= \frac{1}{\sqrt{m}} \sum_{r=1}^{m} a_r \frac{d}{dt} \sigma((\mathbf{w}_r(t,\omega))^*\mathbf{x}) \\
&= \frac{1}{\sqrt{m}} \sum_{r=1}^{m} a_r \mathbf{x}_i^* \mathbb{1}_{\{(\mathbf{w}_r(t,.))^*\mathbf{x} \geq 0\}}(\omega) \frac{d}{dt} \mathbf{w}_r(t,\omega) \\
&= \mathbf{x}_i^* \frac{1}{m} \sum_{r=1}^{m} a_r \mathbb{1}_{\{(\mathbf{w}_r(t,.))^*\mathbf{x}_i \geq 0\}}(\omega) \sum_{j=1}^{n} (y_j - f_j(t,\omega)) a_r \mathbf{x}_j \mathbb{1}_{\{(\mathbf{w}_r(t,.))^*\mathbf{x}_j \geq 0\}} \\
&= \sum_{j=1}^{n} (y_j - f_j(t,\omega)) \sum_{r=1}^{m} \mathbf{x}_i^* \mathbf{x}_j \frac{1}{m} \mathbb{1}_{\{(\mathbf{w}_r(t,.))^*\mathbf{x}_i \geq 0, (\mathbf{w}_r(t,.))^*\mathbf{x}_j \geq 0\}}(\omega) \\
&= \sum_{j=1}^{n} (y_j - f_j(t,\omega)) \mathbf{h}_{i,j}.
\end{aligned} \tag{29}$$

Hence, we obtain that

$$\frac{d}{dt} f(t,\omega) = \mathbf{H}(t,\omega)(y - f(t,\omega)). \tag{30}$$

Lemma 2.7. *Assume Setting 1.1, assume for all $i \in \{0,1,2,\dots,n\}$ that $\|\mathbf{x}_i\|_2 = 1$, assume for all $r \in \{0,1,2,\dots,m\}$ that $\mathbf{w}_r(0,.) \sim \mathcal{N}(\mathbf{0},\mathbf{I})$ and $a_r(0,.) \sim \mathrm{Unif}[\{-1,1\}]$ are i.i.d. random vectors, let $\delta_1 \in (0,1)$, and let*

$$B_1 = \bigcap_{i,j=1}^{n,n} \left\{ \omega \in \Omega \colon |\mathbf{h}_{i,j}(0,\omega) - \mathfrak{h}_{i,j}| \leq \sqrt{\frac{1}{2m} \log\left(\frac{2n^2}{\delta_1}\right)} \right\} \in \mathcal{F}. \tag{31}$$

Then $\mathbb{P}(B_1) \geq 1 - \delta_1$, for all $\omega \in B_1$ it holds that $\|\mathbf{H}(0,\omega) - \mathfrak{H}\|_2 \leq \frac{\lambda_0}{4}$, and $\lambda_{min}(\mathbf{H}(0,\omega)) \geq \frac{3}{4}\lambda_0$.

Proof of Lemma 2.7. Let $\delta' = \frac{\delta_1}{n^2}$, let $A_{i,j} = \left\{ \omega \in \Omega \colon |\mathbf{h}_{i,j}(0,\omega) - \mathfrak{h}_{i,j}| \leq \sqrt{\frac{1}{2m} \log\left(\frac{2}{\delta'}\right)} \right\}$, let $X_{i,j}^r(t,\omega)$ satisfy for all $i,j \in \{1,2,\dots,n\}$, $r \in \{1,2,\dots,m\}$, $t \in [0,\infty)$, $\omega \in \Omega$ that $X_{i,j}^r(t,\omega) = \mathbf{x}_i^* \mathbf{x}_j \mathbb{1}_{\{(\mathbf{w}_r(t,.))^*\mathbf{x}_i \geq 0, (\mathbf{w}_r(t,.))^*\mathbf{x}_j \geq 0\}}(\omega)$. First, observe that

$$\begin{aligned}
|\mathbf{h}_{i,j}(0,\omega) - \mathfrak{h}_{i,j}| &= \Bigg| \frac{1}{m} \sum_{r=1}^{m} \mathbf{x}_i^* \mathbf{x}_j \mathbb{1}_{\{(\mathbf{w}_r(0,.))^*\mathbf{x}_i \geq 0, (\mathbf{w}_r(0,.))^*\mathbf{x}_j \geq 0\}}(\omega) \\
&\qquad - \mathbb{E}\Big[\mathbf{x}_i^* \mathbf{x}_j \mathbb{1}_{\{(\mathbf{w}_r(0,.))^*\mathbf{x}_i \geq 0, (\mathbf{w}_r(0,.))^*\mathbf{x}_j \geq 0\}} \Big] \Bigg| \\
&= \left| \frac{1}{m} \sum_{r=1}^{m} \big(X_{i,j}^r(0,\omega) - \mathbb{E}\big[X_{i,j}^r(0,.) \big] \big) \right|.
\end{aligned} \tag{32}$$

Note for all $i \in \{1, 2, \ldots, n\}$ that $\{(\mathbf{w}_r)^* \mathbf{x}_i\}_{r=1}^m$ are independent random variables. This implies for all $i, j \in \{1, 2, \ldots, n\}$ it holds that $\{X_{i,j}^r\}_{r=1}^m$ are independent random variables. Combining this, (32), and Theorem 7.4 implies that

$$\mathbb{P}\left(\left\{\omega \in \Omega \colon |\mathbf{h}_{i,j}(0,\omega) - \mathfrak{h}_{i,j}| \leq \sqrt{\frac{1}{2m}\log\left(\frac{2}{\delta'}\right)}\right\}\right) \geq 1 - \delta'. \tag{33}$$

Next we combine (33) and Theorem 7.5 to obtain that

$$\begin{aligned} \mathbb{P}\left(\bigcup_{i,j=1}^{n,n} A_{i,j}^c\right) &\leq \sum_{i,j=1}^{n,n} \mathbb{P}\big(A_{i,j}^c\big) \\ &\leq n^2 \delta' \\ &= \delta_1. \end{aligned} \tag{34}$$

This implies that

$$\mathbb{P}(B_1) \geq 1 - \delta_1. \tag{35}$$

Hence, for all $\omega \in B_1$ it holds that

$$\begin{aligned} \|\mathbf{H}(0,\omega) - \mathfrak{H}\|_2^2 &\leq \|\mathbf{H}(0,\omega) - \mathfrak{H}\|_F^2 \\ &= \sum_{i,j=1}^{n,n} |\mathbf{h}_{i,j}(0,\omega) - \mathfrak{h}_{i,j}|^2 \\ &\leq \frac{n^2}{2m}\log\left(\frac{2n^2}{\delta_1}\right) \\ &\leq \frac{2n^2}{m}\log\left(\frac{n}{\delta_1}\right). \end{aligned} \tag{36}$$

Next, note that the assumption on m and Theorem 2.5 ensures that

$$m \geq \frac{Kn^6}{2\lambda_0^3\delta^3} \geq \frac{n^2 8\log\big(\frac{3n}{\delta}\big)}{\lambda_0}. \tag{37}$$

Next we combine this and (36) to obtain for all $\omega \in B_1$ that

$$\|\mathbf{H}(0,\omega) - \mathfrak{H}\|_2 \leq \frac{\lambda_0}{4}. \tag{38}$$

This, combined with Lemma 7.8 ensures for all $\omega \in B_1$ that

$$\|\mathbf{H}(0,\omega)\|_2 \geq \frac{3\lambda_0}{4}. \tag{39}$$

The proof of Lemma 2.7 is thus completed. □

Lemma 2.8. *Assume Setting 1.1, assume for all $r \in \{0, 1, 2, \ldots, m\}$ that $\mathbf{w}_r(0,.) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $a_r(0,.) \sim \mathrm{Unif}[\{-1, 1\}]$ are i.i.d. random vectors, let R be given by $R = \frac{\sqrt{2\pi}\delta_2\lambda_0}{16n^2}$, let $\delta_1, \delta_2 \in (0,1)$ let for all $\omega \in \Omega$*
$B_0(\omega) = \big\{\hat{\mathbf{w}}_1, \hat{\mathbf{w}}_2, \ldots, \hat{\mathbf{w}}_m \in \mathbb{R}^d \colon \max_{r\in\{1,2,\ldots,m\}} \|\mathbf{w}_r(0,\omega) - \hat{\mathbf{w}}_r\|_2 \leq R\big\}$, let $\hat{\mathbf{H}} = (\hat{\mathbf{h}}_{i,j})_{i,j\in\{1,2,\ldots,n\}} \in \mathbb{R}^{n\times n}$ satisfy for all $\hat{\mathbf{w}}_1, \hat{\mathbf{w}}_2, \ldots, \hat{\mathbf{w}}_m \in \mathbb{R}^d$ that

$$\hat{\mathbf{h}}_{i,j} = \frac{1}{m}\mathbf{x}_i^* \mathbf{x}_j \sum_{r=1}^m \mathbb{1}_{\{(\hat{\mathbf{w}}_r)^*\mathbf{x}_i \geq 0, (\hat{\mathbf{w}}_r)^*\mathbf{x}_j \geq 0\}}, \tag{40}$$

let $B_2 = \left\{\omega \in \Omega\colon \sum_{j,i=1}^{n,n} \sup_{\hat{\mathbf{w}}_1,\hat{\mathbf{w}}_2,\ldots,\hat{\mathbf{w}}_m \in B_0(\omega)} \left\{\left|\hat{\mathbf{h}}_{i,j} - \mathbf{h}_{i,j}(0,\omega)\right|\right\} \leq \frac{4Rn^2}{\delta_2\sqrt{2\pi}}\right\} \in \mathcal{F}$, *and let* $B_1 = \bigcap_{i,j=1}^{n,n}\left\{\omega \in \Omega\colon |\mathbf{h}_{i,j}(0,\omega) - \mathfrak{h}_{i,j}| \leq \sqrt{\frac{1}{2m}\log\left(\frac{2n^2}{\delta_1}\right)}\right\} \in \mathcal{F}$. *Then*

(i) $\Omega \ni \omega \mapsto \sup_{\hat{\mathbf{w}}_1,\hat{\mathbf{w}}_2,\ldots,\hat{\mathbf{w}}_m \in B_0(\omega)} \left\{\left|\hat{\mathbf{h}}_{i,j} - \mathbf{h}_{i,j}(0,\omega)\right|\right\} \in \mathbb{R}$ *is measurable.*

(ii) it holds that $\mathbb{P}(B_2) \geq 1 - \delta_2$

(iii) let $\omega \in B_1 \cap B_2$ *and let* $\hat{\mathbf{w}}_1, ..., \hat{\mathbf{w}}_m \in \mathbb{R}^d$ *satisfy for all* $r \in \{1, 2, \ldots, m\}$ *that* $\|\mathbf{w}_r(0,\omega) - \hat{\mathbf{w}}_r\|_2 \leq R$. *Then it holds that* $\|\mathbf{H}(0,\omega) - \hat{\mathbf{H}}\|_2 \leq \frac{\lambda_0}{4}$ *and* $\lambda_{min}(\hat{\mathbf{H}}) \geq \frac{\lambda_0}{2}$.

Proof of Lemma 2.8. First, note that $\mathbf{w}_r(0,.)$ is a random vector ensures that $\mathbb{1}_{\{(\mathbf{w}_r(0,.))^*\mathbf{x}_i \geq 0, (\mathbf{w}_r(0,.))^*\mathbf{x}_j \geq 0\}}$ is measurable. This combined with the fact that sums, supremums, and addition by constants to measurable functions are measurable functions establishes item (i). Note that Theorem 7.1 ensures that

$$\mathbb{P}(\{|(\mathbf{w}_r(0,.))^*\mathbf{x}_i| \leq R\}) \leq \frac{2R}{\sqrt{2\pi}}. \tag{41}$$

Obseve for all $\omega \in B$, $j \in \{1, 2, \ldots, n\}$ that $\left|\mathbf{x}_j^*(\mathbf{w}_r(0,\omega) - \hat{\mathbf{w}}_r)\right| < R$ This, combined with (41) and the fact that $\mathbf{w}_1(0,.), \mathbf{w}_2(0,.), \ldots, \mathbf{w}_m(0,.)$ are independent ensures for all

$i, j \in \{1, 2 \ldots, n\}$ that

$$
\begin{aligned}
&\mathbb{E}\left[\sup_{\hat{\mathbf{w}}_1,\hat{\mathbf{w}}_2,\ldots,\hat{\mathbf{w}}_m\in B(.)}\left\{\left|\hat{\mathbf{h}}_{i,j}-\mathbf{h}_{i,j}(0,.)\right|\right\}\right]\\
&=\mathbb{E}\Bigg[\sup_{\hat{\mathbf{w}}_1,\hat{\mathbf{w}}_2,\ldots,\hat{\mathbf{w}}_m\in B(.)}\Bigg\{\frac{1}{m}\Bigg|\mathbf{x}_i^*\mathbf{x}_j\sum_{r=1}^{m}\mathbb{1}_{\{(\mathbf{w}_r(0,.))^*\mathbf{x}_i\geq 0,(\mathbf{w}_r(0,.))^*\mathbf{x}_j\geq 0\}}(\omega)\\
&\qquad\qquad -\mathbb{1}_{\{(\hat{\mathbf{w}}_r)^*\mathbf{x}_i\geq 0,(\hat{\mathbf{w}}_r)^*\mathbf{x}_j\geq 0\}}\Bigg|\Bigg\}\Bigg]\\
&=\mathbb{E}\Bigg[\sup_{\hat{\mathbf{w}}_1,\hat{\mathbf{w}}_2,\ldots,\hat{\mathbf{w}}_m\in B(.)}\Bigg\{\frac{1}{m}\Bigg|\mathbf{x}_i^*\mathbf{x}_j\sum_{r=1}^{m}\mathbb{1}_{\{(\mathbf{w}_r(0,.))^*\mathbf{x}_i\geq 0,(\mathbf{w}_r(0,.))^*\mathbf{x}_j\geq 0\}}(\omega)\\
&\qquad -\mathbb{1}_{\{(\hat{\mathbf{w}}_r)^*\mathbf{x}_i\geq 0,(\hat{\mathbf{w}}_r)^*\mathbf{x}_j\geq 0\}}+\mathbb{1}_{\{(\mathbf{w}_r(0,.))^*\mathbf{x}_i\geq 0,(\hat{\mathbf{w}}_r)^*\mathbf{x}_j\geq 0\}}\\
&\qquad\qquad -\mathbb{1}_{\{(\mathbf{w}_r(0,.))^*\mathbf{x}_i\geq 0,(\hat{\mathbf{w}}_r)^*\mathbf{x}_j\geq 0\}}\Bigg|\Bigg\}\Bigg]\\
&=\mathbb{E}\Bigg[\sup_{\hat{\mathbf{w}}_1,\hat{\mathbf{w}}_2,\ldots,\hat{\mathbf{w}}_m\in B(.)}\Bigg\{\frac{1}{m}\Bigg|\mathbf{x}_i^*\mathbf{x}_j\sum_{r=1}^{m}\mathbb{1}_{\{(\mathbf{w}_r(0,.))^*\mathbf{x}_i\geq 0\}}\big(\mathbb{1}_{\{(\mathbf{w}_r(0,.))^*\mathbf{x}_j\geq 0\}}(\omega)\\
&\qquad -\mathbb{1}_{\{(\hat{\mathbf{w}}_r)^*\mathbf{x}_j\geq 0\}}\big)+\mathbb{1}_{\{(\hat{\mathbf{w}}_r)^*\mathbf{x}_j\geq 0\}}\big(\mathbb{1}_{\{(\mathbf{w}_r(0,.))^*\mathbf{x}_i\geq 0\}}(\omega)-\mathbb{1}_{\{(\hat{\mathbf{w}}_r)^*\mathbf{x}_i\geq 0\}}\big)\Bigg|\Bigg\}\Bigg]\\
&\leq\frac{1}{m}\sum_{r=1}^{m}\mathbb{E}\Bigg[\sup_{\hat{\mathbf{w}}_1,\hat{\mathbf{w}}_2,\ldots,\hat{\mathbf{w}}_m\in B(.)}\left\{\left|\mathbb{1}_{\{(\mathbf{w}_r(0,.))^*\mathbf{x}_i\geq 0\}}(\omega)-\mathbb{1}_{\{(\hat{\mathbf{w}}_r)^*\mathbf{x}_i\geq 0\}}\right|\right\}\\
&\qquad +\sup_{\hat{\mathbf{w}}_1,\hat{\mathbf{w}}_2,\ldots,\hat{\mathbf{w}}_m\in B(.)}\left\{\left|\mathbb{1}_{\{(\mathbf{w}_r(0,.))^*\mathbf{x}_j\geq 0\}}(\omega)-\mathbb{1}_{\{(\hat{\mathbf{w}}_r)^*\mathbf{x}_j\geq 0\}}\right|\right\}\Bigg]\\
&\leq\frac{1}{m}\sum_{r=1}^{m}\mathbb{P}(\{|(\mathbf{w}_r(0,.))^*\mathbf{x}_i|<R\})+\mathbb{P}(\{|(\mathbf{w}_r(0,.))^*\mathbf{x}_j|<R\})\\
&\leq\frac{4R}{\sqrt{2\pi}}.
\end{aligned}
\tag{42}
$$

This and the fact that the expectation is linear ensures that

$$
\mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{n}\sup_{\hat{\mathbf{w}}_1,\hat{\mathbf{w}}_2,\ldots,\hat{\mathbf{w}}_m\in B(.)}\left\{\left|\hat{\mathbf{h}}_{i,j}-\mathbf{h}_{i,j}(0,.)\right|\right\}\right]\leq\frac{4Rn^2}{\sqrt{2\pi}}
\tag{43}
$$

This, item (i), and Markov's inequality 7.2 ensures that $\mathbb{P}(B_2)\geq 1-\delta_2$. This establishes item (ii). Let $\hat{\omega}\in B_1\cap B_2$ and let $\hat{\mathbf{w}}_1,...,\hat{\mathbf{w}}_m\in\mathbb{R}^d$ satisfy for all $r\in 1,2,...,m$, that $\|\mathbf{w}_r(0,\hat{\omega})-\hat{\mathbf{w}}_r\|_2\leq R$. Lemma 7.6 and Theorem 7.7 thus ensures that

$$
\begin{aligned}
\left\|\hat{\mathbf{H}}-\mathbf{H}(0,\hat{\omega})\right\|_2&\leq\left\|\hat{\mathbf{H}}-\mathbf{H}(0,\hat{\omega})\right\|_F\\
&\leq\sum_{i=1}^{n}\sum_{j=1}^{n}\left|\hat{\mathbf{h}}_{i,j}-\mathbf{h}_{i,j}(0,\hat{\omega})\right|\\
&\leq\sum_{i=1}^{n}\sum_{j=1}^{n}\sup_{\hat{\mathbf{w}}_1,\hat{\mathbf{w}}_2,\ldots,\hat{\mathbf{w}}_m\in B(\hat{\omega})}\left\{\left|\hat{\mathbf{h}}_{i,j}-\mathbf{h}_{i,j}(0,\hat{\omega})\right|\right\}\\
&\leq\frac{4Rn^2}{\sqrt{2\pi}\delta_2}.
\end{aligned}
\tag{44}
$$

This, combined with the fact that $R = \frac{\sqrt{2\pi}\delta_2\lambda_0}{n^2 16}$ ensures that

$$\left\|\hat{\mathbf{H}} - \mathbf{H}(0,\hat{\omega})\right\|_2 \leq \frac{\lambda_0}{4}. \tag{45}$$

Next we combine this, Lemma 7.8, and Lemma 2.7 to establish that

$$\lambda_{min}(\hat{\mathbf{H}}) \geq \lambda_{min}(\mathbf{H}(0,\hat{\omega})) - \left\|\hat{\mathbf{H}} - \mathbf{H}(0,\hat{\omega})\right\|_2 \geq \frac{\lambda_0}{2}. \tag{46}$$

The proof of Lemma 2.8 is thus completed. □

Lemma 2.9. *Assume that the conditions of Theorem 2.6 hold, let $\omega \in \Omega$, let $t \in (0,\infty)$, assume for all $s \in [0,t]$ that $\lambda_{min}(\mathbf{H}(s,\omega)) \geq \frac{\lambda_0}{2}$, and let R_1 be given by $R_1 = \frac{2\sqrt{n}\|y - f(0,\omega)\|_2}{\sqrt{m}\lambda_0}$. Then for all $s \in [0,t]$ it holds that*

$$\|y - f(s,\omega)\|_2^2 \leq \exp(-\lambda_0 s)\|y - f(0,\omega)\|_2^2, \tag{47}$$

and for all $r \in \{0,1,2,\ldots,m\}$, $s \in [0,t]$ that

$$\|\mathbf{w}_r(s,\omega) - \mathbf{w}_r(0,\omega)\|_2 \leq R_1. \tag{48}$$

Proof of Lemma 2.9. First, note that (30) ensures for all $s \in [0,t]$ that

$$\begin{aligned} \frac{d}{ds}\|y - f(s,\omega)\|_2^2 &= -2\|y - f(s,\omega)\|_2 \frac{df(s,\omega)}{ds} \\ &= -2(y - f(s,\omega))^* \mathbf{H}(t,\omega)(y - f(s,\omega)) \\ &\leq -\lambda_0 \|y - f(s,\omega)\|_2^2. \end{aligned} \tag{49}$$

This ensures for all $s \in [0,t]$ that

$$\frac{d}{ds}\Big(\exp(\lambda_0 s)\|y - f(s,\omega)\|_2^2\Big) \leq 0. \tag{50}$$

This implies for all $s \in [0,t]$ that

$$\|y - f(s,\omega)\|_2^2 \leq \exp(-\lambda_0 s)\|y - f(0,\omega)\|_2^2. \tag{51}$$

Next, note that (26), the fact that $\|\mathbf{x}_i\|_2 = 1$, (51), and Lemma 7.6 ensures for all $s \in [0,t]$ that

$$\begin{aligned} \left\|\frac{d\mathbf{w}_r(s,\omega)}{ds}\right\|_2 &= \frac{1}{\sqrt{m}}\left\|\sum_{i=1}^n (f_i(s,\omega) - y_i) a_r \mathbf{x}_i \mathbb{1}_{\{(\mathbf{w}_r(s,.))^* \mathbf{x}_i \geq 0\}}(\omega)\right\|_2 \\ &\leq \frac{1}{\sqrt{m}} \sum_{i=1}^n \|f_i(s,\omega) - y_i\|_2 \\ &\leq \frac{1}{\sqrt{m}} \sum_{i=1}^n \|f_i(s,\omega) - y_i\|_1 \\ &= \frac{\sqrt{n}}{\sqrt{m}} \|f(s,\omega) - y\|_2 \\ &\leq \frac{\sqrt{n}}{\sqrt{m}} \|f(0,\omega) - y\|_2 \exp\left(\frac{-\lambda_0 s}{2}\right) \end{aligned} \tag{52}$$

This, combined with Theorem 7.3 implies that

$$\begin{aligned}\|\mathbf{w}_r(t,\omega)-\mathbf{w}_r(0,\omega)\|_2 &= \left\|\int_0^t \frac{d\mathbf{w}_r(s,\omega)}{ds}ds\right\|_2\\ &\leq \int_0^t \left\|\frac{d\mathbf{w}_r(s,\omega)}{ds}\right\|_2 ds\\ &\leq \int_0^t \frac{\sqrt{n}}{\sqrt{m}}\|f(0,\omega)-y\|_2 \exp\!\left(-\frac{\lambda_0 s}{2}\right)ds\\ &\leq \frac{2\sqrt{n}\|f(0,\omega)-y\|_2}{\sqrt{m}\lambda_0} = R_1.\end{aligned} \tag{53}$$

The proof of Lemma 2.9 is thus completed. □

Note that Theorem 7.1, $\mathbb{E}[a_r^2]=1$, and for all $r\in\{1,2,\ldots,m\}$, $i\in\{1,2,\ldots,n\}$ it holds that $\mathbb{E}[(\sigma((\mathbf{w}_r)^*\mathbf{x}_i))^2]=\frac{1}{2}$, ensures for all $i\in\{1,2,\ldots,n\}$ that

$$\mathbb{E}[f_i^2(0,.)]=\frac{1}{2}. \tag{54}$$

Next, combining for all $r\in\{1,2,\ldots,m\}$ that $a_r(0,.)$ is independent of $\mathbf{w}_r(0,.)$, $\mathbb{E}[a_r(0,.)]=0$, (54), for all $i\in\{1,2,\ldots,n\}$ that $|y_i|<C$, and the linearity of expectations ensures that

$$\begin{aligned}\mathbb{E}[\|y-f(0,\omega)\|_2^2] &= \sum_{i=1}^n y_i^2 - 2y_i\mathbb{E}[f_i(0,.)]+\mathbb{E}[f_i^2(0,.)]\\ &= \sum_{i=1}^n y_i^2+\frac{1}{2}\\ &\leq n\left(C^2+\frac{1}{2}\right).\end{aligned} \tag{55}$$

Let $B_3=\left\{\omega\in\Omega\colon \|y-f(0,\omega)\|_2^2\leq \frac{n\left(C^2+\frac{1}{2}\right)}{\delta_3}\right\}$ and let $R_1=\frac{2\sqrt{n}\|f(0,\omega)-y\|_2}{\sqrt{m}\lambda_0}$. Note that $\|y-f(0,.)\|_2^2$ is a non negative random variable. This, (55), and Markov's inequality 7.2 ensures that

$$\mathbb{P}(B_3)\geq 1-\delta_3. \tag{56}$$

This combined with fact that $m\geq \frac{n^6 2^2 16^2 (C^2+1)}{2\pi\lambda_0^4\delta_2^2\delta_3}$ ensures for all $\omega\in B_3$ that

$$R_1<R. \tag{57}$$

Lemma 2.10. *Assume Setting 1.1, assume that the conditions of Theorem 2.6 hold, let* $B_3=\left\{\omega\in\Omega\colon \|y-f(0,\omega)\|_2^2\leq\frac{n\left(C^2+\frac{1}{2}\right)}{\delta_3}\right\}\in\mathcal{F}$, *let* $B_0(\omega)=\left\{\hat{\mathbf{w}}_1,\hat{\mathbf{w}}_2,\ldots,\hat{\mathbf{w}}_m\in\mathbb{R}^d\colon \max_{r\in\{1,2,\ldots,m\}}\|\mathbf{w}_r(0,\omega)-\hat{\mathbf{w}}_r(\omega)\|_2\leq R\right\}$, *let* $B_2=\left\{\omega\in\Omega\colon \sum_{j,i=1}^{n,n}\sup_{\hat{\mathbf{w}}_1,\hat{\mathbf{w}}_2,\ldots,\hat{\mathbf{w}}_m\in B_0(\omega)}\left\{\left|\hat{\mathbf{h}}_{i,j}-\mathbf{h}_{i,j}(0,\omega)\right|\right\}\leq\frac{4Rn^2}{\delta_2\sqrt{2\pi}}\right\}$, *and let* $B_1=\bigcap_{i,j=1}^{n,n}\left\{\omega\in\Omega\colon |\mathbf{h}_{i,j}(0,\omega)-\mathfrak{h}_{i,j}|\leq\sqrt{\frac{1}{2m}\log\!\left(\frac{2n^2}{\delta_1}\right)}\right\}\in\mathcal{F}$. *Then*

(i) $\mathbb{P}(B_1\cap B_2\cap B_3)\geq 1-\delta_1-\delta_2-\delta_3=1-\delta$

(ii) for all $r \in \{0, 1, 2, \ldots, m\}$, $t \in [0, \infty)$, $\omega \in B_1 \cap B_2 \cap B_3$, $t \in [0, \infty)$ *it holds that*

$$\|\mathbf{w}_r(t, \omega) - \mathbf{w}_r(0, \omega)\|_2 \leq R_1 \tag{58}$$

(iii) for all $\omega \in B_1 \cap B_2 \cap B_3$, $t \in [0, \infty)$ *it holds that*

$$\lambda_{min}(\mathbf{H}(t, \omega)) \geq \frac{\lambda_0}{2} \tag{59}$$

(iv) for all $t \in [0, \infty)$, $\omega \in B_1 \cap B_2 \cap B_3$ *it holds that that*

$$\|y - f(t, \omega)\|_2^2 \leq \|y - f(0, \omega)\|_2^2 \exp(-\lambda_0 t) \tag{60}$$

.

Proof of Lemma 2.10. First note that Lemma 2.7, Lemma 2.8 and (56), establishes item (i). We prove item (ii) by a contradiction argument.

Assume there exist $r \in \{0, 1, 2, \ldots, m\}$, $\omega_0 \in B_1 \cap B_2 \cap B_3$, $t^{'} \in [0, \infty)$ such that (58) does not hold. Then by Lemma 2.9 there exist a $s_0 \leq t^{'}$ such that $\lambda_{min}(\mathbf{H}(s_0, \omega_0)) < \frac{\lambda_0}{2}$. This and Lemma 2.8 ensures that there exist a finite time $t_0 < s_0$ such that

$$t_0 = \inf\left\{ t > 0 \colon \max_{r \in \{0,1,2,\ldots,m\}} \|\mathbf{w}_r(t, \omega_0) - \mathbf{w}_r(0, \omega_0)\|_2 \geq R \right\} \tag{61}$$

This, combined with the continuity of $\mathbf{W}(t)$ ensures that there exist a $r_0 \in \{0, 1, 2, \ldots, m\}$ such that

$$\|\mathbf{w}_{r_0}(t_0, \omega_0) - \mathbf{w}_{r_0}(0, \omega_0)\|_2 = R. \tag{62}$$

This and Lemma 2.8 implies for all $t \leq t_0$ that $\lambda_{min}(\mathbf{H}(t, \omega_0)) \geq \frac{\lambda_0}{2}$. This combined with Lemma 2.9 ensures for all $r \in \{0, 1, 2, \ldots, m\}$ that

$$\|\mathbf{w}_r(t_0, \omega_0) - \mathbf{w}_r(0, \omega_0)\|_2 \leq R_1. \tag{63}$$

This, combined with the fact that $R_1 < R$ and (62) gives a contradiction. This establishes item (ii). Lemma 2.8 and item (ii) establishes item (iii). Lemma 2.9 and item (iii) establishes item (iv). The proof of Lemma 2.10 is thus completed. □

The proof of Theorem 2.6 is thus completed. □

Remark 2.11. *This result is an improvement on the result in [7] in that it is possible to see at the moment of initialization if* $\omega \in B_1 \cap B_2 \cap B_3$*, and hence if convergence is guaranteed. If convergence is not guaranteed it is possible to re-initialise until convergence is guaranteed.*

3 Discrete time analysis

Throughout this section, we will consider the convergence of the training loss (2) for a neural network (1) trained by the gradient descent algorithm (65). The proof idea is to develop several inequalities that allows us to prove that if m is large enough, and η is small enough, with a high probability over the initialisation it holds for all $k \in \mathbb{N}_0$ that

$$\|y - f(k + 1, \omega)\|_2 \leq (1 - \frac{\eta \lambda_0}{2}) \|y - f(k, \omega)\|_2. \tag{64}$$

From the continuous time analysis we will need Lemma 2.8 and hence also indirectly Lemma 2.7 to bound the least eigenvalue of $\mathbf{H}(k,\omega)$. Both of which are valid independently of the optimisation algorithm used. In this section we will follow the article [7].

Theorem 3.1 (Convergence Rate of Gradient Descent)**.** *Assume Setting 1.1, assume for all* $i \in \{0,1,2,\dots,n\}$ *that* $\|\mathbf{x}_i\|_2 = 1$*, let* $C \in [0,\infty)$*, assume* $n > 3$*, assume for all* $i \in \{0,1,2,\dots,n\}$ *that* $|y_i| \leq C$*, assume for all* $r \in \{0,1,2,\dots,m\}$ *that* $\mathbf{w}_r(0,.) \sim \mathcal{N}(\mathbf{0},\mathbf{I})$ *and* $a_r(0,.) \sim \text{Unif}[\{-1,1\}]$ *are i.i.d. random vectors, assume for all* $t \in (0,\infty)$*,* $\omega \in \Omega$ *that* $a(t,\omega) = a(0,\omega)$*, let* $\eta \leq \frac{\lambda_0}{8n^2}$*, let* $K_2 = \frac{4^5(C^2+1)16^2}{2\pi}$*, let* $\delta \in (0,1)$*, let* $m \geq \frac{K_2 n^6}{\lambda_0^4 \delta^3}$*, assume for all* $k \in \mathbb{N}_0$*,* $r \in \{1,2,\dots,m\}$*,* $\omega \in \Omega$ *that*

$$\mathbf{w}_r(k+1,\omega) = \mathbf{w}_r(k,\omega) - \eta\frac{1}{\sqrt{m}}\sum_{i=1}^{n}(f_i(k,\omega)-y_i)a_r\mathbf{x}_i\mathbb{1}_{\{(\mathbf{w}_r(t,.))^*\mathbf{x}_i\geq 0\}}(\omega), \tag{65}$$

let $\delta_1 = \delta_2 = \delta_3 = \delta_4 = \frac{\delta}{4}$*,* $c = \frac{\sqrt{2\pi}}{16}$*,* $R_3 = \frac{c\lambda_0\delta_4}{n^2}$*, let* $\hat{\mathbf{h}}_{i,j}$ *satisfy for all* $i,j \in \{1,2,\dots,n\}$*,* $\hat{\mathbf{w}}_1, \hat{\mathbf{w}}_2, \dots, \hat{\mathbf{w}}_n \in \mathbb{R}^d$ *that* $\hat{\mathbf{h}}_{i,j} = \frac{1}{m}\mathbf{x}_i^*\mathbf{x}_j\sum_{r=1}^{m}\mathbb{1}_{\{(\hat{\mathbf{w}}_r)^*\mathbf{x}_i\geq 0,(\hat{\mathbf{w}}_r)^*\mathbf{x}_j\geq 0\}}$*, let* $R = \frac{\sqrt{2\pi}\delta_2\lambda_0}{16n^2}$*, let*
$A_{i,r} = \left\{\exists \mathbf{w}\colon \|\mathbf{w}-\mathbf{w}_r(0,\omega)\|_2 \leq R_3, \mathbb{1}_{\{\mathbf{x}_i^*\mathbf{w}_r(0,.)\geq 0\}}(\omega) \neq \mathbb{1}_{\{\mathbf{x}_i^*\mathbf{w}\geq 0\}}\right\} \in \mathcal{F}$*, let*
$B_4 = \left\{\omega\in\Omega\colon \sum_{i=1}^{n}\left|\left\{r\in\{1,2,\dots,m\}\colon \mathbb{1}_{A_{i,r}}(\omega)=1\right\}\right| \leq \frac{C_2 mnR_3}{\delta_4}\right\} \in \mathcal{F}$*, let*
$B_0(\omega) = \left\{\hat{\mathbf{w}}_1,\hat{\mathbf{w}}_2,\dots,\hat{\mathbf{w}}_m \in \mathbb{R}^d\colon \max_{r\in\{1,2,\dots,m\}}\|\mathbf{w}_r(0,\omega)-\hat{\mathbf{w}}_r\|_2 \leq R\right\}$*, let*
$B_2 = \left\{\omega\in\Omega\colon \sum_{j,i=1}^{n,n}\sup_{\hat{\mathbf{w}}_1,\hat{\mathbf{w}}_2,\dots,\hat{\mathbf{w}}_m\in B_0(\omega)}\left\{\left|\hat{\mathbf{h}}_{i,j}-\mathbf{h}_{i,j}(0,\omega)\right|\right\} \leq \frac{4Rn^2}{\delta_2\sqrt{2\pi}}\right\} \in \mathcal{F}$*, let*
$B_1 = \bigcap_{i,j=1}^{n,n}\left\{\omega\in\Omega\colon |\mathbf{h}_{i,j}(0,\omega)-\mathfrak{h}_{i,j}| \leq \sqrt{\frac{1}{2m}\log\left(\frac{2n^2}{\delta_1}\right)}\right\} \in \mathcal{F}$ *and let*
$B_3 = \left\{\omega\in\Omega\colon \|y-f(0,\omega)\|_2^2 \leq \frac{n\left(C^2+\frac{1}{2}\right)}{\delta_3}\right\} \in \mathcal{F}.$
Then

(i) it holds for all $k \in \mathbb{N}_0$*,* $\omega \in B_1\cap B_2\cap B_3\cap B_4$ *that*

$$\|f(k,\omega)-y\|_2^2 \leq (1-\frac{\eta\lambda_0}{2})^k\|f(0,\omega)-y\|_2^2. \tag{66}$$

(ii) it holds that $\mathbb{P}(B_1\cap B_2\cap B_3\cap B_4) \geq 1-\delta$

Proof of Theorem 3.1.

Lemma 3.2. *Assume the conditions of Theorem 3.1, let* $k \in \mathbb{N}_0 = \{0,1,2\dots\}$*, let* R_2 *be given by*
$R_2 = \frac{4\sqrt{n}\|y-f(0,\omega)\|_2}{\sqrt{m}\lambda_0}$*, assume that* (66) *holds for all* $k' \in \{1,2,\dots,k\}$*,* $\omega \in B$*. Then it holds for all* $r \in \{0,1,2,\dots,m\}$*,* $\omega \in B$ *that*

$$\|\mathbf{w}_r(k+1,\omega)-\mathbf{w}_r(0,\omega)\|_2 \leq R_2. \tag{67}$$

Proof of Lemma 3.2. First note that $\eta \leq \frac{\lambda_0}{8n^2}$, Theorem 2.6 and Theorem 2.5 ensures that

$$0 < \eta < \frac{2}{\lambda_0}. \tag{68}$$

This ensures that

$$\frac{1}{1-\left(1-\frac{\eta\lambda_0}{2}\right)^{\frac{1}{2}}} \leq \frac{4}{\eta\lambda_0}. \tag{69}$$

Note that (65), the triangle inequality, (66), the formula for a geometric series, Lemma 7.6, and (69) ensures for all $\omega \in B$ that

$$\begin{aligned}
\|\mathbf{w}_r(k+1,\omega) - \mathbf{w}_r(0,\omega)\|_2 &\leq \sum_{j=0}^{k} \|\mathbf{w}_r(j+1,\omega) - \mathbf{w}_r(j,\omega)\|_2 \\
&= \sum_{j=0}^{k} \left\| \eta \frac{1}{\sqrt{m}} \sum_{i=1}^{n} (f_i(j,\omega) - y_i) a_r \mathbf{x}_i \mathbb{1}_{\{(\mathbf{w}_r(t,.))^* \mathbf{x}_i \geq 0\}}(\omega) \right\|_2 \\
&\leq \frac{\eta}{\sqrt{m}} \sum_{j=0}^{k} \sum_{i=1}^{n} \|f_i(k,\omega) - y_i\|_2 \\
&\leq \frac{\eta}{\sqrt{m}} \sum_{j=0}^{k} \sum_{i=1}^{n} \|f_i(k,\omega) - y_i\|_1 \\
&= \frac{\eta}{\sqrt{m}} \sum_{j=0}^{k} \|f(j,\omega) - y\|_1 \\
&\leq \frac{\eta\sqrt{n}}{\sqrt{m}} \sum_{j=0}^{k} \|f(j,\omega) - y\|_2 \\
&\leq \frac{\eta\sqrt{n}}{\sqrt{m}} \|y - f(0,\omega)\|_2 \sum_{j=0}^{k} \left(1 - \frac{\eta\lambda_0}{2}\right)^{\frac{j}{2}} \\
&\leq \frac{\eta\sqrt{n}}{\sqrt{m}} \frac{4}{\eta\lambda_0} \|y - f(0,\omega)\|_2 = R_2.
\end{aligned} \tag{70}$$

The proof of Lemma 3.2 is thus completed. □

Lemma 3.3. *Assume Setting 1.1, let* $C_2 = \frac{2}{\sqrt{2\pi}}$, $c = \frac{\sqrt{2\pi}}{16}$, $R_3 = \frac{c\lambda_0\delta_4}{n^2}$, *let* $A_{i,r} \in \mathcal{F}$ *be given by*

$$A_{i,r} = \Big\{ \exists \mathbf{w} \colon \|\mathbf{w} - \mathbf{w}_r(0,\omega)\|_2 \leq R_3, \mathbb{1}_{\{\mathbf{x}_i^* \mathbf{w}_r(0,.) \geq 0\}}(\omega) \neq \mathbb{1}_{\{\mathbf{x}_i^* \mathbf{w} \geq 0\}} \Big\}, \tag{71}$$

let $S_i \colon \Omega \to 2^{\{1,2,\ldots,m\}}$ *satisfy for all* $\omega \in \Omega$, $i \in \{1,2,\ldots,n\}$ *that*

$$S_i(\omega) = \big\{ r \in \{1,2,\ldots,m\} \colon \mathbb{1}_{A_{i,r}}(\omega) = 0 \big\}, \tag{72}$$

let $S_i^{\perp} \colon \Omega \to 2^{\{1,2,\ldots,m\}}$ *satisfy for all* $\omega \in \Omega$, $i \in \{1,2,\ldots,n\}$ *that* $S_i^{\perp}(w) = \{1,2,\ldots,m\} \setminus S_i(w)$, *let* $\big|S_i^{\perp}\big| \colon \Omega \to \{1,\ldots,m\}$ *satisfy for all* $\omega \in \Omega$, $i \in \{1,2,\ldots,n\}$ *that* $\big|S_i^{\perp}\big|(w) = \big|S_i^{\perp}(w)\big|$, *and let* $B_4 = \Big\{ \omega \in \Omega \colon \sum_{i=1}^{n} \big|S_i^{\perp}\big|(w) \leq \frac{C_2 m n R_3}{\delta_4} \Big\} \in \mathcal{F}$. *Then it holds that*

$$\mathbb{P}(B_4) \geq 1 - \delta_4. \tag{73}$$

Proof of Lemma 3.3. First, let $X \colon \Omega \to \mathbb{R}$ be a standard normal random variable, let $b \in [0,1]$ be given by

$$b = \mathbb{P}(\{|X| \leq R_3\}), \tag{74}$$

and let $Y \colon \Omega \to \mathbb{R}$ be a random variable that satisfies $Y \sim \mathrm{bin}(b,m)$. This, combined with the expectation of a binomial distributed random variable, (72), and Theorem 7.1 ensures that

$$\begin{aligned}\mathbb{E}\Big[\Big|S_i^{\perp}\Big|\Big] &= \mathbb{E}[\big|\{r \in \{0,1,2,\dots,m\}\colon \mathbb{1}_{A_{i,r}}(.) \neq 0\}\big|] \\ &= \mathbb{E}[|\{r \in \{0,1,2,\dots,m\}\colon w \in A_{i,r}\}|] \\ &= \mathbb{E}[|\{r \in \{0,1,2,\dots,m\}\colon |\mathbf{w}_r(0,.))^* \mathbf{x}_i| \leq R_3\}|] \\ &= \mathbb{E}[Y] \\ &= Km \\ &\leq \frac{2R_3 m}{\sqrt{2\pi}}.\end{aligned} \tag{75}$$

This, the linearity of expectation, the fact that $\big|S_i^{\perp}\big|$ is a non negative random variable, and Markov's inequality 7.2 ensures that

$$\mathbb{P}(B_4) \geq 1 - \delta_4. \tag{76}$$

The proof of Lemma 3.3 is thus completed. □

Let $S_i\colon \Omega \to 2^{\{1,2,\dots,m\}}$ satisfy for all $\omega \in \Omega$, $i \in \{1,2,\dots,n\}$ that

$$S_i(\omega) = \big\{r \in \{1,2,\dots,m\}\colon \mathbb{1}_{A_{i,r}}(\omega) = 0\big\}, \tag{77}$$

let $S_i^{\perp}\colon \Omega \to 2^{\{1,2,\dots,m\}}$ satisfy for all $\omega \in \Omega$, $i \in \{1,2,\dots,n\}$ that
$S_i^{\perp}(w) = \{1,2,\dots,m\} \setminus S_i(w)$, let $k \in \mathbb{N}_0$, assume that (66) hold for all $k^{'} \in \{0,1,\dots,k\}$, $\omega \in B$, let $I_1^{i,k}\colon \Omega \to \mathbb{R}$ satisfy for all $\omega \in \Omega$, $i \in \{0,1,\dots,n\}$, $k \in \mathbb{N}_0 = \{0,1,\dots\}$ that

$$\begin{aligned}&I_1^{i,k}(\omega) \\ &= \frac{1}{\sqrt{m}} \sum_{r \in S_i} a_r(\sigma((\mathbf{w}_r(k,\omega) \\ &\quad - \eta \frac{1}{\sqrt{m}} \sum_{j=1}^{n} (f_j(k,\omega) - y_j) a_r \mathbf{x}_j \mathbb{1}_{\{(\mathbf{w}_r(t,.))^* \mathbf{x}_j \geq 0\}}(\omega))^* \mathbf{x}_j) - \sigma((\mathbf{w}_r(k,\omega))^* \mathbf{x}_j)),\end{aligned} \tag{78}$$

let $I_2^{i,k}\colon \Omega \to \mathbb{R}$ satisfy for all $\omega \in \Omega$, $i \in \{0,1,\dots,n\}$, $k \in \mathbb{N}_0 = \{0,1,\dots\}$ that

$$\begin{aligned}&I_2^{i,k}(\omega) \\ &= \frac{1}{\sqrt{m}} \sum_{r \in S_i^{\perp}} a_r(\sigma((\mathbf{w}_r(k,\omega) \\ &\quad - \eta \frac{1}{\sqrt{m}} \sum_{j=1}^{n} (f_j(k,\omega) - y_j) a_r \mathbf{x}_j \mathbb{1}_{\{(\mathbf{w}_r(t,.))^* \mathbf{x}_j \geq 0\}}(\omega))^* \mathbf{x}_j) - \sigma((\mathbf{w}_r(k,\omega))^* \mathbf{x}_j)),\end{aligned} \tag{79}$$

let $\mathbf{H}^{\perp} = (\mathbf{h}_{i,j}^{\perp})_{i,j \in \{1,2,\dots,n\}}\colon \mathbb{N}_0 \times \Omega \to R^{n \times n}$ satisfy for all $i,j \in \{1,2,\dots,n\}$, $k \in \mathbb{N}_0$, $\omega \in \Omega$ that

$$\mathbf{h}_{i,j}^{\perp}(k,\omega) = \frac{1}{m} \sum_{r \in S_i^{\perp}} \mathbf{x}_i^* \mathbf{x}_j \mathbb{1}_{\{(\mathbf{w}_r(k,.))^* \mathbf{x}_i \geq 0, (\mathbf{w}_r(k,.))^* \mathbf{x}_j \geq 0\}}(\omega). \tag{80}$$

First, observe that

$$f_i(k+1,\omega) - f_i(k,\omega) = I_1^{i,k}(\omega) + I_2^{i,k}(\omega). \tag{81}$$

The fact that σ is a 1-Lipschitz function, (78), and Lemma 7.6 ensures for all $\omega \in \Omega$, $i \in \{1, 2, \ldots, n\}$, $k \in \mathbb{N}_0$ that

$$\begin{aligned}
&\left|I_2^{i,k}(\omega)\right| \\
&\leq \frac{\eta}{\sqrt{m}} \sum_{r \in S_i^\perp} \left| \left(\frac{1}{\sqrt{m}} \sum_{j=1}^{n} (f_j(k,\omega) - y_j) a_r \mathbf{x}_j \mathbb{1}_{\{(\mathbf{w}_r(t,.))^* \mathbf{x}_j \geq 0\}}(\omega) \right)^* \mathbf{x}_j \right| \\
&\leq \frac{\eta}{\sqrt{m}} \left|S_i^\perp\right|(\omega) \max_{r \in \{0,1,2,\ldots,m\}} \left| \left(\frac{1}{\sqrt{m}} \sum_{j=1}^{n} (f_j(k,\omega) - y_j) a_r \mathbf{x}_j \mathbb{1}_{\{(\mathbf{w}_r(t,.))^* \mathbf{x}_j \geq 0\}}(\omega) \right)^* \mathbf{x}_j \right| \\
&\leq \frac{\eta}{\sqrt{m}} \left|S_i^\perp\right|(\omega) \max_{r \in \{0,1,2,\ldots,m\}} \left\| \frac{1}{\sqrt{m}} \sum_{j=1}^{n} (f_j(k,\omega) - y_j) a_r \mathbf{x}_j \mathbb{1}_{\{(\mathbf{w}_r(t,.))^* \mathbf{x}_j \geq 0\}}(\omega) \right\|_1 \\
&\leq \left|S_i^\perp\right|(\omega) \frac{\eta}{m} \|f(k,\omega) - y\|_1 \\
&\leq \left|S_i^\perp\right|(\omega) \frac{\eta\sqrt{n}}{m} \|f(k,\omega) - y\|_2.
\end{aligned} \tag{82}$$

The assumption that for all $k^{'} \in \{0, ..., k\}$, $\omega \in B$ that (66) holds and Lemma 3.2 ensures for all $r \in \{0, 1, \ldots, m\}$, $\omega \in B$ that

$$\|\mathbf{w}(k+1,\omega) - \mathbf{w}(0,\omega)\|_2 \leq R_2. \tag{83}$$

Moreover, observe that $m \geq \frac{K_2 n^6}{\lambda_0^4 \delta^3}$, $R_2 = \frac{4\sqrt{n}\|y - f(0,\omega)\|_2}{\sqrt{m}\lambda_0}$, $R_3 = \frac{c\lambda_0\delta_4}{n^2}$ ensures for all $\omega \in B_3$ that

$$R_2 < R_3. \tag{84}$$

Note that (84), (55), and (83) ensures for all $\omega \in B \cap B_3$, $r \in S_i(\omega)$, $k \in \mathbb{N}_0$ that

$$\mathbb{1}_{\{\mathbf{x}_i^* \mathbf{w}_r(k+1,.) \geq 0\}}(\omega) = \mathbb{1}_{\{\mathbf{x}_i^* \mathbf{w}_r(k,.) \geq 0\}}(\omega). \tag{85}$$

Next, note that (80), (78), and (85) ensures for all $\omega \in B \cap B_3$, $i \in \{1, 2, \ldots, n\}$, $k \in \mathbb{N}_0$ that

$$\begin{aligned}
&I_1^{i,k}(\omega) \\
&= -\frac{\eta}{m} \sum_{r \in S_i} a_r^2 \left(\sum_{j=1}^{n} (f_j(k,\omega) - y_j) \mathbf{x}_j \mathbb{1}_{\{(\mathbf{w}_r(k,.))^* \mathbf{x}_j \geq 0\}}(\omega) \right)^* \mathbf{x}_i \mathbb{1}_{\{(\mathbf{w}_r(k,\omega))^* \mathbf{x}_i \geq 0\}}(\omega) \\
&= -\frac{\eta}{m} \sum_{r \in S_i} \sum_{j=1}^{n} \mathbf{x}_i^* \mathbf{x}_j \mathbb{1}_{\{(\mathbf{w}_r(k,.))^* \mathbf{x}_i \geq 0, (\mathbf{w}_r(k,.))^* \mathbf{x}_j \geq 0\}}(\omega) (f_j(k,\omega) - y_j) \\
&= -\eta \sum_{j=1}^{n} (f_j(k,\omega) - y_j) \Big(\mathbf{h}_{i,j}(k,\omega) - \mathbf{h}_{i,j}^\perp(k,\omega) \Big).
\end{aligned} \tag{86}$$

This ensures for all $\omega \in B \cap B_3$, $k \in \{0, 1, \ldots\}$ that

$$-\mathbf{I}_1^k = \Big(-\mathbf{H}(k,\omega) + \mathbf{H}(k,\omega)^\perp \Big)(y - f(k,\omega)). \tag{87}$$

Moreover, Lemma 3.3 ensures for all $\omega \in B_4$, $k \in \mathbb{N}_0$ that

$$\begin{aligned}\left\|\mathbf{H}^{\perp}(k,\omega)\right\|_2^2 &= \sum_{i=1}^{n}\sum_{j=1}^{n}\left|\mathbf{h}_{i,j}^{\perp}(k,\omega)\right|^2 \\ &\leq \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\left|S_i^{\perp}\right|^2(\omega)}{m^2} = \sum_{i=1}^{n}\frac{n\left|S_i^{\perp}\right|^2(\omega)}{m^2} \\ &\leq \frac{n}{m^2}\left(\sum_{i=1}^{n}\left|S_i^{\perp}\right|^2(\omega)\right) \\ &\leq \frac{nR_3^2}{m^2}.\end{aligned} \tag{88}$$

This ensures that

$$\left\|\mathbf{H}^{\perp}(k,\omega)\right\|_2 \leq \frac{n^{3/2}C_2R_3}{\delta_4}. \tag{89}$$

Next note that Lemma 7.6, σ is 1-Lipschitz, (65), and definition of f_i ensures for all $\omega \in \Omega$, $k \in \mathbb{N}_0$ that

$$\begin{aligned}|f_i(k+1,\omega) - f_i(k,\omega)| &\leq \frac{\eta}{\sqrt{m}}\sum_{r=1}^{m}\left|\left(\frac{1}{\sqrt{m}}\sum_{i=1}^{n}(f_i(k,\omega)-y_i)a_r\mathbf{x}_i\mathbb{1}_{\{(\mathbf{w}_r(t,.))^*\mathbf{x}_i\geq 0\}}(\omega)\right)^*\mathbf{x}_i\right| \\ &\leq \frac{\eta}{\sqrt{m}}\sum_{r=1}^{m}\left\|\frac{1}{\sqrt{m}}\sum_{i=1}^{n}(f_i(k,\omega)-y_i)a_r\mathbf{x}_i\mathbb{1}_{\{(\mathbf{w}_r(t,.))^*\mathbf{x}_i\geq 0\}}(\omega)\right\|_1 \\ &\leq \frac{\eta}{m}\sum_{r=1}^{m}\|f(k,\omega)-y\|_1 \\ &\leq \eta\sqrt{n}\|f(k,\omega)-y\|_2.\end{aligned} \tag{90}$$

This ensures for all $\omega \in \Omega$, $k \in \mathbb{N}_0$ that

$$\|f(k+1,\omega) - f(k,\omega)\|_2^2 \leq \eta^2 n^2 \|f(k,\omega)-y\|_2^2. \tag{91}$$

Furthermore, observe that Lemma 3.2 ensures that

$$\begin{aligned}-(y-f(k,\omega))^*\mathbf{I}_2^k(\omega) &\leq \sum_{i=1}^{n}|y_i - f_i(k,\omega)|\left|I_2^{i,k}(\omega)\right| \\ &\leq \frac{\eta\sqrt{n}}{m}\|y-f(k,\omega)\|_2\|y-f(k,\omega)\|_1\sum_{i=1}^{n}\left|S_i^{\perp}(\omega)\right| \\ &\leq \frac{\eta n}{m}\|y-f(k,\omega)\|_2^2\frac{R_3nmC_2}{\delta_4}.\end{aligned} \tag{92}$$

Moreover, Lemma 2.8, (83), the fact that $\delta_2 = \delta_4$, and (84) ensures for all $\omega \in B_1 \cap B_2 \cap B_3 \cap B$ that

$$-(y-f(k,\omega))^*\mathbf{H}(k,\omega)(y-f(k,\omega)) \leq -\frac{\lambda_0}{2}\|y-f(k,\omega)\|_2^2. \tag{93}$$

Combining this, (81), (82), (87), (89), (90), and (92) ensures for all $\omega \in B_1 \cap B_2 \cap B_3 \cap B_4 \cap B$ that

$$\begin{aligned}
&\|y - f(k+1,\omega)\|_2^2 \\
&= \|(y - f(k,\omega)) + (f(k,\omega) - f(k+1,\omega))\|_2^2 \\
&= \|y - f(k,\omega)\|_2^2 + \|f(k+1,\omega) - f(k,\omega)\|_2^2 \\
&\qquad - 2(y - f(k,\omega))^* \mathbf{I}_2^k - 2(y - f(k,\omega))^* \mathbf{I}_1^k \\
&= \|y - f(k,\omega)\|_2^2 + \|f(k+1,\omega) - f(k,\omega)\|_2^2 \\
&\quad - 2(y - f(k,\omega))^* \mathbf{I}_2^k - 2\eta(y - f(k+1,\omega))^* \mathbf{H}(k,\omega)(y - f(k,\omega)) \\
&\qquad\qquad + 2\eta(y - f(k+1,\omega))^* \mathbf{H}^{\perp}(k)(y - f(k,\omega)) \\
&\leq (1 + \eta^2 n^2 + 2\frac{\eta n^2 R_3 C_2}{\delta_4} - \eta\lambda_0 + \frac{2\eta n^{3/2} C_2 R_3}{\delta_4}) \|y - f(k,\omega)\|_2^2.
\end{aligned} \tag{94}$$

This, combined with $R_3 = \frac{c\lambda_0\delta_4}{n^2}$, $C = \frac{2}{\sqrt{2\pi}}$, $c = \frac{\sqrt{2\pi}}{16}$, and $\eta \leq \frac{\lambda_0}{8n^2}$ asserts for all $\omega \in B_1 \cap B_2 \cap B_3 \cap B_4 \cap B$ that

$$\begin{aligned}
\|y - f(k+1,\omega)\|_2^2 &\leq \|y - f(k,\omega)\|_2^2 \Big(1 + \lambda_0\eta\Big(-1 + \frac{\eta n^2}{\lambda_0} + \frac{1}{4} + \frac{1}{4\sqrt{n}}\Big)\Big) \\
&\leq \|y - f(k,\omega)\|_2^2 \Big(1 - \frac{n\eta}{2}\Big).
\end{aligned} \tag{95}$$

Combining this and induction on k ensures for all $\omega \in B_1 \cap B_2 \cap B_3 \cap B_4$, $k \in \mathbb{N}_0$ that

$$\|y - f(k+1,\omega)\|_2^2 \leq \|y - f(k,\omega)\|_2^2 \Big(1 - \frac{n\eta}{2}\Big). \tag{96}$$

Observe that Lemma 3.3 and Lemma 2.10 establishes that $\mathbb{P}(B_1 \cap B_2 \cap B_3 \cap B_3) \geq 1 - \delta$. The proof of Theorem 3.1 is thus completed.

□

4 Jointly training both layers

In this section we will consider the gradient flow algorithm for training both layers (97) of the neural network (1). The analysis will be similar as when we trained only the hidden layer, the bound on m will be of the same order of magnitude, and the bound on the convergence rate will be the same. In this section we will follow the article [7] .

Remark 4.1. *Note that the assumptions and resulting convergence rate in Theorem 4.2 are similar to the assumptions of Theorem 2.6.*

Theorem 4.2. *Assume Setting 1.1, let $C \in [0,\infty)$, assume for all $i \in \{1,2,\dots,n\}$ that $\|\mathbf{x}_i\|_2 = 1$, $|y_i| \leq C$, let $\delta \in (0,1)$, $K = \frac{2^2 16^2 (C^2+1) 3^3}{2\pi}$, $m \geq \frac{Kn^6}{\lambda_0^4 \delta} \max\Big(\frac{32^2}{3^3\delta^2}, 2\pi \log\big(\frac{4nm}{\delta}\big)\Big)$, assume for all $r \in \{0,1,2,\dots,m\}$ that, $\mathbf{w}_r(0,.) \sim \mathcal{N}(\mathbf{0},\mathbf{I})$ and $a_r(0,.) = \mathrm{Unif}[\{-1,1\}]$ are i.i.d. random vectors, assume for all $r \in \{1,2,\dots,m\}$, $\omega \in \Omega$, $t \in [0,\infty)$ that*

$$\begin{aligned}
\frac{d\mathbf{w}_r(t,\omega)}{dt} &= -\frac{1}{\sqrt{m}} \sum_{i=1}^{n} (f_i(t,\omega) - y_i) a_r \mathbf{x}_i \mathbb{1}_{\{(\mathbf{w}_r(t,.))^* \mathbf{x}_i \geq 0\}}(\omega) \\
\frac{da_r(t,\omega)}{dt} &= -\sum_{i=1}^{n} (f_i - y_i) \frac{1}{\sqrt{m}} \sigma((\mathbf{w}_r(t,\omega))^* \mathbf{x}_i),
\end{aligned} \tag{97}$$

let $\delta_1 = \delta_2 = \delta_3 = \delta_5 = \frac{\delta}{4}$, *let* $\hat{\mathbf{h}}_{i,j}$ *satisfy for all* $i, j \in \{1, 2, \ldots, n\}$, $\hat{\mathbf{w}}_1, \hat{\mathbf{w}}_2, \ldots, \hat{\mathbf{w}}_m \in \mathbb{R}^d$, $\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_m \in \mathbb{R}$ *that* $\hat{\mathbf{h}}_{i,j} = \frac{1}{m}\mathbf{x}_i^* \mathbf{x}_j \sum_{r=1}^m \hat{a}_r^2 \mathbb{1}_{\{(\hat{\mathbf{w}}_r)^* \mathbf{x}_i \geq 0, (\hat{\mathbf{w}}_r)^* \mathbf{x}_j \geq 0\}}$, *let* $R = \frac{\sqrt{2\pi}\delta_2 \lambda_0}{16 n^2}$, *let*
$B_5 = \bigcap_{i,r=1}^{n,m} \left\{ |(\mathbf{w}_r(0,\omega))^* \mathbf{x}_i| \leq \sqrt{\frac{1}{2}\log\left(\frac{2nm}{\delta_5}\right)} \right\} \in \mathcal{F}$, *let*
$B_0(\omega) = \left\{ \hat{\mathbf{w}}_1, \hat{\mathbf{w}}_2, \ldots, \hat{\mathbf{w}}_m \in \mathbb{R}^d \colon \max_{r \in \{1,2,\ldots,m\}} \|\mathbf{w}_r(0,\omega) - \hat{\mathbf{w}}_r\|_2 \leq R \right\}$, *let*
$B_2 = \left\{ \omega \in \Omega \colon \sum_{j,i=1}^{n,n} \sup_{\hat{\mathbf{w}}_1, \hat{\mathbf{w}}_2, \ldots, \hat{\mathbf{w}}_m \in B_0(\omega)} \left\{ \left| \hat{\mathbf{h}}_{i,j} - \mathbf{h}_{i,j}(0,\omega) \right| \right\} \leq \frac{4Rn^2}{\delta_2 \sqrt{2\pi}} \right\} \in \mathcal{F}$, *let*
$B_1 = \bigcap_{i,j=1}^{n,n} \left\{ \omega \in \Omega \colon |\mathbf{h}_{i,j}(0,\omega) - \mathfrak{h}_{i,j}| \leq \sqrt{\frac{1}{2m}\log\left(\frac{2n^2}{\delta_1}\right)} \right\} \in \mathcal{F}$ *and let*
$B_3 = \left\{ \omega \in \Omega \colon \|y - f(0,\omega)\|_2^2 \leq \frac{n(C^2 + \frac{1}{2})}{\delta_3} \right\} \in \mathcal{F}$. *Then*

(i) it holds that $\mathbb{P}(B_1 \cap B_2 \cap B_3 \cap B_5) \geq 1 - \delta$

(ii) it holds for all $t \in [0, \infty)$, $\omega \in B_1 \cap B_2 \cap B_3 \cap B_5$ *that*

$$\|f(t,\omega) - y\|_2^2 \leq \exp(-\lambda_0 t)\|f(0,\omega) - y\|_2^2. \tag{98}$$

Proof of Theorem 4.2. First, let R_w, R_a, R_w' and R_a' be given by $R_w = \frac{\sqrt{2\pi}\delta_2\lambda_0}{32n^2} = \frac{R}{2}$, $R_a = \frac{\lambda_0}{n^2 20}$, $R_w' = \frac{4\sqrt{n}}{\lambda_0\sqrt{m}}\|f(0,\omega) - y\|_2$, and $R_a' = \frac{4\sqrt{n}}{\lambda_0\sqrt{m}}\left(\sqrt{\log\left(\frac{nm}{\delta_5}\right)}\right)\|f(0,\omega) - y\|_2$, and let $\mathbf{G} = (\mathbf{g}_{i,j})_{i,j \in \{1,2,\ldots,n\}} \colon [0,\infty) \times \Omega \to \mathbb{R}^{n \times n}$ satisfy for all $i, j \in \{0, 1, 2, \ldots, n\}$, $\omega \in \Omega$, $t \in [0, \infty)$ that

$$\mathbf{g}_{i,j}(t,\omega) = \sum_{r=1}^m \frac{1}{m}\sigma((\mathbf{w}_r(t,\omega))^* \mathbf{x}_i)\sigma((\mathbf{w}_r(t,\omega))^* \mathbf{x}_j). \tag{99}$$

This, combined with

$$\frac{df_i(t,\omega)}{da_r} = \frac{1}{\sqrt{m}}\sigma((\mathbf{w}_r(t,\omega))^* \mathbf{x}_i), \tag{100}$$

(30), and (97) ensures that

$$\frac{df(t,\omega)}{dt} = (\mathbf{H}(t,\omega) + \mathbf{G}(t,\omega))(y - f(t,\omega)). \tag{101}$$

Lemma 4.3. *Assume Setting 1.1, assume that the conditions of Theorem 4.2 hold, let* $\omega \in \Omega$, $t \in (0, \infty)$, *assume for all* $s \in [0, t]$ *that* $\lambda_{min}(\mathbf{H}(s,\omega)) \geq \frac{\lambda_0}{2}$. *Then it holds that*

$$\|y - f(t,\omega)\|_2^2 \leq \|y - f(0,\omega)\|_2^2 \exp(-\lambda_0 t). \tag{102}$$

Proof of Lemma 4.3. First, note that Lemma 7.9 and the fact that $\mathbf{G}(t,\omega)$ is positive semi-definite ensures that

$$\frac{d}{dt}\|y - f(t,\omega)\|_2^2 = -2(y - f(t,\omega))^*(\mathbf{H}(t,\omega) + \mathbf{G}(t,\omega))(y - f(t,\omega)) \leq -\lambda_0\|y - f(t,\omega)\|_2^2. \tag{103}$$

This ensures that

$$\frac{d}{dt}\left(\exp(\lambda_0 t)\|y - f(t,\omega)\|_2^2\right) \leq 0. \tag{104}$$

This implies that

$$\|y - f(t,\omega)\|_2^2 \leq \|y - f(0,\omega)\|_2^2 \exp(-\lambda_0 t). \tag{105}$$

The proof of Lemma 4.3 is thus completed. □

Lemma 4.4. *Assume the conditions of Theorem 4.2 hold, let* $\hat{\mathbf{w}}_1, \hat{\mathbf{w}}_2, \dots, \hat{\mathbf{w}}_m \in \mathbb{R}^d$, $\hat{a}_1, \hat{a}_2, \dots, \hat{a}_m \in \mathbb{R}$, *let* $\hat{\mathbf{H}} = (\hat{\mathbf{h}}_{i,j})_{i,j\in\{1,2,\dots,n\}} \in \mathbb{R}^{n\times n}$ *satisfy for all* $i, j \in \{1, 2, \dots, n\}$ *that*

$$\hat{\mathbf{h}}_{i,j} = \frac{1}{m}\mathbf{x}_i^* \mathbf{x}_j \sum_{r=1}^{m} \hat{a}_r^2 \mathbb{1}_{\{(\hat{\mathbf{w}}_r)^*\mathbf{x}_i \geq 0, (\hat{\mathbf{w}}_r)^*\mathbf{x}_j \geq 0\}}, \tag{106}$$

let $\delta_1, \delta_2 \in (0,1)$, *let* $B_1 = \bigcap_{i,j=1}^{n,n} \left\{\omega \in \Omega\colon |\mathbf{h}_{i,j}(0,\omega) - \mathfrak{h}_{i,j}| \leq \sqrt{\frac{1}{2m}\log\left(\frac{2n^2}{\delta_1}\right)}\right\} \in \mathcal{F}$, $B_2 = \left\{\omega \in \Omega\colon \sum_{j,i=1}^{n,n} \sup_{\hat{\mathbf{w}}_1, \hat{\mathbf{w}}_2, \dots, \hat{\mathbf{w}}_m \in B_0(\omega)} \left\{\left|\hat{\mathbf{h}}_{i,j} - \mathbf{h}_{i,j}(0,\omega)\right|\right\} \leq \frac{4Rn^2}{\delta_2\sqrt{2\pi}}\right\}$, *let* $\omega \in B_1 \cap B_2$ *and assume that for all* $r \in \{0, 1, 2, \dots, m\}$ *it holds that*

$$\|\hat{\mathbf{w}}_r - \mathbf{w}_r(0,\omega)\|_2 \leq R_w, \tag{107}$$

and

$$|\hat{a}_r - a_r(0,\omega)| \leq R_a. \tag{108}$$

Then

(i) it holds that

$$\left\|\hat{\mathbf{H}} - \mathbf{H}(0,\omega)\right\| \leq \frac{\lambda_0}{4} \tag{109}$$

(ii) it holds that

$$\lambda_{min}(\hat{\mathbf{H}}) \geq \frac{\lambda_0}{2}. \tag{110}$$

Proof of Lemma 4.4. Lemma 2.8 and (106) ensures that

$$\left\|\mathbf{H}' - \mathbf{H}(0,\omega)\right\|_2 \leq \frac{\lambda_0}{8} \tag{111}$$

and

$$\lambda_{min}(\mathbf{H}') \geq \frac{\lambda_0}{2}. \tag{112}$$

In the next step, observe that

$$\begin{aligned} \left|\hat{\mathbf{h}}_{i,j} - \mathbf{h}'_{i,j}\right| &\leq \frac{1}{m}\sum_{r=1}^{m} |\hat{a}_r^2 - 1| \\ &\leq R_a^2 + 2R_a. \end{aligned} \tag{113}$$

Note that Theorem 2.5 ensures that

$$\lambda_0^2 \leq n^2 10. \tag{114}$$

Combining this, (113), (111), Lemma 7.7 and Lemma 7.6, and the triangle inequality ensures that

$$\begin{aligned} \left\|\hat{\mathbf{H}} - \mathbf{H}(0,\omega)\right\|_2 &\leq \left\|\hat{\mathbf{H}} - \mathbf{H}'\right\|_2 + \left\|\mathbf{H}' - \mathbf{H}(0,\omega)\right\|_2 \\ &\leq n^2\big(R_a^2 + 2R_a\big) + \frac{\lambda_0}{8} \leq \frac{\lambda_0}{4}. \end{aligned} \tag{115}$$

This, Lemma 2.7 and Lemma 7.8 ensures that

$$\lambda_{min}(\hat{\mathbf{H}}) \geq \frac{\lambda_0}{2}. \tag{116}$$

The proof of Lemma 4.4 is thus completed. □

Lemma 4.5. *Let $\omega \in \Omega$, assume that for all $s \in [0,t]$ it holds that*

$$\lambda_{min}(\mathbf{H}(s,\omega)) \geq \frac{\lambda_0}{2}, \tag{117}$$

and assume that for all $r \in \{0,1,2,\ldots,m\}$, $s \in [0,t]$ it holds that $|a_r(s,\omega) - a_r(0,\omega)| \leq R_a$. Then it holds that

$$\|\mathbf{w}_r(t,\omega) - \mathbf{w}_r(0,\omega)\|_2 \leq R_w^{'}. \tag{118}$$

Proof of Lemma 4.5. First, note that Lemma 4.3 ensures that for all $s \in [0,t]$ it holds that

$$\begin{aligned}
\left\| \frac{d\mathbf{w}_r(s,\omega)}{ds} \right\|_2 &= \frac{1}{\sqrt{m}} \left\| \sum_{i=1}^{n} (f_i(s,\omega) - y_i) a_r(s,\omega) \mathbf{x}_i \mathbb{1}_{\{(\mathbf{w}_r(s,.))^* \mathbf{x}_i \geq 0\}}(\omega) \right\|_2 \\
&\leq \frac{1}{\sqrt{m}} \sum_{i=1}^{n} \|(f_i(s,\omega) - y_i) a_r(s,\omega)\|_2 \\
&\leq \frac{|a_r(s,\omega)|}{\sqrt{m}} \sum_{i=1}^{n} \|f_i(s,\omega) - y_i\|_1 \\
&= \frac{|a_r(s,\omega)|\sqrt{n}}{\sqrt{m}} \|f(s,\omega) - y\|_2 \\
&\leq \frac{|a_r(s,\omega)|\sqrt{n}}{\sqrt{m}} \|f(0,\omega) - y\|_2 \exp\left(\frac{-\lambda_0 s}{2} \right).
\end{aligned} \tag{119}$$

Note that Theorem 2.5 ensures that

$$R_a \leq 1. \tag{120}$$

This, combined with (119), Lemma 7.3, implies that

$$\begin{aligned}
\|\mathbf{w}_r(t,\omega) - \mathbf{w}_r(0,\omega)\|_2 &= \left\| \int_0^t \frac{d\mathbf{w}_r(s,\omega)}{ds} ds \right\|_2 \\
&\leq \int_0^t \left\| \frac{d\mathbf{w}_r(s,\omega)}{ds} \right\|_2 ds \\
&\leq \frac{\sqrt{n}}{\sqrt{m}} \|f(0,\omega) - y\|_2 \int_0^t |a_r(s,\omega)| \exp\left(-\frac{\lambda_0 s}{2} \right) ds \\
&\leq \frac{\sqrt{n}}{\sqrt{m}} \|f(0,\omega) - y\|_2 |R_a + 1| \frac{2}{\lambda_0} \\
&\leq \frac{4\sqrt{n}}{\lambda_0 \sqrt{m}} \|f(0,\omega) - y\|_2 = R_w^{'}.
\end{aligned} \tag{121}$$

The proof of Lemma 4.5 is thus completed.

□

Lemma 4.6. *Let $B_5 = \bigcap_{i,r=1}^{n,m} \left\{ |(\mathbf{w}_r(0,\omega))^* \mathbf{x}_i| \leq \sqrt{\frac{1}{2} \log\left(\frac{2nm}{\delta_5} \right)} \right\}$, let $\omega \in B_5$, $t \in (0,\infty)$, assume for all $s \in [0,t]$ that*

$$\lambda_{min}(\mathbf{H}(s,\omega)) \geq \frac{\lambda_0}{2}, \tag{122}$$

and assume for all $r \in \{0,1,2,\ldots,m\}$, $s \in [0,t]$ that $\|\mathbf{w}_r(s,\omega) - \mathbf{w}_r(0,\omega)\|_2 \leq R_w$. Then

(i) it holds for all $r \in \{0, 1, 2, \ldots, m\}$ *that*

$$|a_r(t,\omega) - a_r(0,\omega)| \leq R_a^{'} \tag{123}$$

(ii) $\mathbb{P}(B_5) \geq 1 - \delta_5$.

Proof of Lemma 4.6. First note that Lemma 7.1 ensures for all $r \in \{0, 1, 2, \ldots, m\}$, $i \in \{0, 1, 2, \ldots, n\}$ that $(\mathbf{w}_r(0,.))^* \mathbf{x}_i \sim \mathcal{N}(0, 1)$. This, Theorem 7.4, and Theorem 7.5 thus ensures that

$$\mathbb{P}(B_5) \geq 1 - \delta_5. \tag{124}$$

Let $\omega \in B_5$, $t \in (0, \infty)$, assume for all $s \in [0, t]$ that

$$\lambda_{min}(\mathbf{H}(s,\omega)) \geq \frac{\lambda_0}{2}, \tag{125}$$

and assume for all $r \in \{0, 1, 2, \ldots, m\}$, $s \in [0, t]$ that $\|\mathbf{w}_r(s,\omega) - \mathbf{w}_r(0,\omega)\|_2 \leq R_w$. This, combined with (97), Lemma 7.6, Lemma 4.3, and the fact that $|\mathbf{x}_i^* \mathbf{w}_r(s,\omega)| \leq |\mathbf{x}_i^* \mathbf{w}_r(0,\omega)| + \|\mathbf{w}_r(s,\omega) - \mathbf{w}_r(0,\omega)\|_2$ ensures for all $r \in \{0, 1, 2, \ldots, m\}$ that

$$\begin{aligned}
\left|\frac{da_r(s,\omega)}{ds}\right| &= \left|\sum_{i=1}^{n} (f_i(s,\omega) - y_i) \frac{1}{\sqrt{m}} \sigma((\mathbf{w}_r(s,\omega))^* \mathbf{x}_i)\right| \\
&\leq \frac{1}{\sqrt{m}} \sum_{i=1}^{n} |(f_i(s,\omega) - y_i)||(\mathbf{w}_r(s,\omega))^* \mathbf{x}_i| \\
&\leq \frac{1}{\sqrt{m}} \left(\sqrt{\frac{1}{2}\log\left(\frac{2nm}{\delta_5}\right)} + R_\omega\right) \|f(s,\omega) - y\|_1 \\
&\leq \frac{\sqrt{n}}{\sqrt{m}} \left(\sqrt{\frac{1}{2}\log\left(\frac{2nm}{\delta_5}\right)} + R_\omega\right) \|f(s,\omega) - y\|_2 \\
&\leq \frac{\sqrt{n}}{\sqrt{m}} \left(\sqrt{\frac{1}{2}\log\left(\frac{2nm}{\delta_5}\right)} + R_\omega\right) \|f(0,\omega) - y\|_2 \exp\left(-\frac{\lambda_0 s}{2}\right).
\end{aligned} \tag{126}$$

Note that Theorem 2.5 ensures that

$$R_w \leq 1. \tag{127}$$

This, (126), ensures for all $r \in \{0, 1, 2, \ldots, m\}$ that

$$\begin{aligned}
|a_r(t,\omega) - a_r(0,\omega)| &\leq \int_0^t \left|\frac{da_r(s)}{ds}\right| ds \\
&\leq \frac{2\sqrt{n}}{\lambda_0\sqrt{m}} \left(\sqrt{\log\left(\frac{nm}{\delta_5}\right)} + R_\omega\right) \|f(0,\omega) - y\|_2 \\
&\leq \frac{4\sqrt{n}}{\lambda_0\sqrt{m}} \left(\sqrt{\log\left(\frac{nm}{\delta_5}\right)}\right) \|f(0,\omega) - y\|_2 \\
&= R_a^{'}.
\end{aligned} \tag{128}$$

The proof of Lemma 4.6 is thus completed. □

Next, note that $m \geq \frac{Kn^6}{\lambda_0^4 \delta} \max\Big(\frac{32^2}{3^3\delta^2}, 2\pi \log\big(\frac{4nm}{\delta}\big)\Big)$ ensures for all $\omega \in B_3 = \Big\{\omega \in \Omega \colon \|y - f(0,\omega)\|_2^2 \leq \frac{n\left(C^2+\frac{1}{2}\right)}{\delta_3}\Big\} \in \mathcal{F}$ that $R_w^{'} < R_w$ and $R_a^{'} < R_a$.

Lemma 4.7. *Assume the conditions of Theorem 4.2 and let $\omega \in B_1 \cap B_2 \cap B_3 \cap B_5$, $t \in [0, \infty)$. Then*

(i) it holds that $\mathbb{P}(B_1 \cap B_2 \cap B_3 \cap B_5) \geq 1 - \delta$

(ii) it holds that

$$\lambda_{min}(\mathbf{H}(t,\omega)) \geq \frac{\lambda_0}{2} \tag{129}$$

(iii) for all $r \in \{0, 1, 2, \ldots, m\}$ it holds that

$$\|\mathbf{w}_r(t,\omega) - \mathbf{w}_r(0,\omega)\|_2 \leq R_w^{'} \tag{130}$$

(iv) it holds that

$$|a_r(t,\omega) - a_r(0,\omega)| \leq R_a^{'} \tag{131}$$

(v) it holds that

$$\|y - f(t,\omega)\|_2^2 \leq \exp(-t\lambda_0)\|y - f(0,\omega)\|_2^2. \tag{132}$$

Proof of Lemma 4.7. First note that (124) and Lemma 2.10 establishes item (i). Note that $R_w^{'} < R_w$ and that $R_a^{'} < R_a$. Assume that the theorem does not hold, let $t_1 \in [0, \infty)$ be the minimal such that the theorem does not hold for all $t \in [0, t_1]$, $\omega \in B_1 \cap B_2 \cap B_3 \cap B_5$.

Assume that there exist a $\omega_0 \in B_1 \cap B_2 \cap B_3 \cap B_5$ such that $\lambda_{min}(\mathbf{H}(t_1,\omega_0)) < \frac{\lambda_0}{2}$ then Lemma 4.4 ensures that there exist a $r \in \{1, 2, \ldots, m\}$ such that $\|\mathbf{w}_r(t_1,\omega_0) - \mathbf{w}_r(0,\omega_0)\| > R_w$ or/and $|a_r(t_1,\omega_0) - a_r(0,\omega_0)| > R_a$. The assumption that $R_w^{'} < R_w$ and $R_a^{'} < R_a$ ensures a contradiction to the minimality of t_1.

Assume that there exist $r \in \{1, 2, \ldots, m\}$, $\omega_0 \in B_1 \cap B_2 \cap B_3 \cap B_5$ such that $\|\mathbf{w}_r(t_1,\omega_0) - \mathbf{w}_r(0,\omega_0)\|_2 > R_w^{'}$ then Lemma 4.5 and the fact that $R_a^{'} < R_a$ ensures that $\|\mathbf{w}_r(t_1,\omega_0) - \mathbf{w}_r(0,\omega_0)\|_2 \leq R_w^{'}$, a contradiction.

Assume that there exist $r \in \{1, 2, \ldots, m\}$, $\omega_0 \in B_1 \cap B_2 \cap B_3 \cap B_5$ such that $|a_r(t_1,\omega_0) - a_r(0,\omega_0)| > R_a^{'}$. Then Lemma 4.6, and the fact that $R_w^{'} < R_w$ ensures for all $r \in \{0, 1, 2, \ldots, m\}$ that $|a_r(t_1,\omega_0) - a_r(0,\omega_0)| \leq R_a^{'}$, a contradiction.

Assume that there exist $\omega_0 \in B_1 \cap B_2 \cap B_3 \cap B_5$ such that $\|y - f(t_1,\omega_0))\|_2^2 > \exp(-t_1\lambda_0)\|y - f(0,\omega_0)\|_2^2$, then by Lemma 4.3 there exist an $s \in [0, t_1]$ such that $\lambda_{min}(\mathbf{H}(s,\omega)) < \frac{\lambda_0}{2}$, a contradiction.

The proof of Lemma 4.7 is thus completed. □

The proof of Theorem 4.2 is thus completed.

□

5 Numerical Experiments

Following [7], we test some of the results by applying gradient descent to synthetic data. We generated $n = 100$ data points $\{\mathbf{x}_i\}_{i=1}^{100}$ such that $\mathbf{x}_i \in \text{Unif}[(-1,1)]^d$ for $d = 1000$, and $y \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Note that one epoch corresponds to one step with gradient descent. In Figure 1(A) we compare the convergence rate of the training loss for different values of m when trained with gradient descent. We normalise the training loss for the cases where

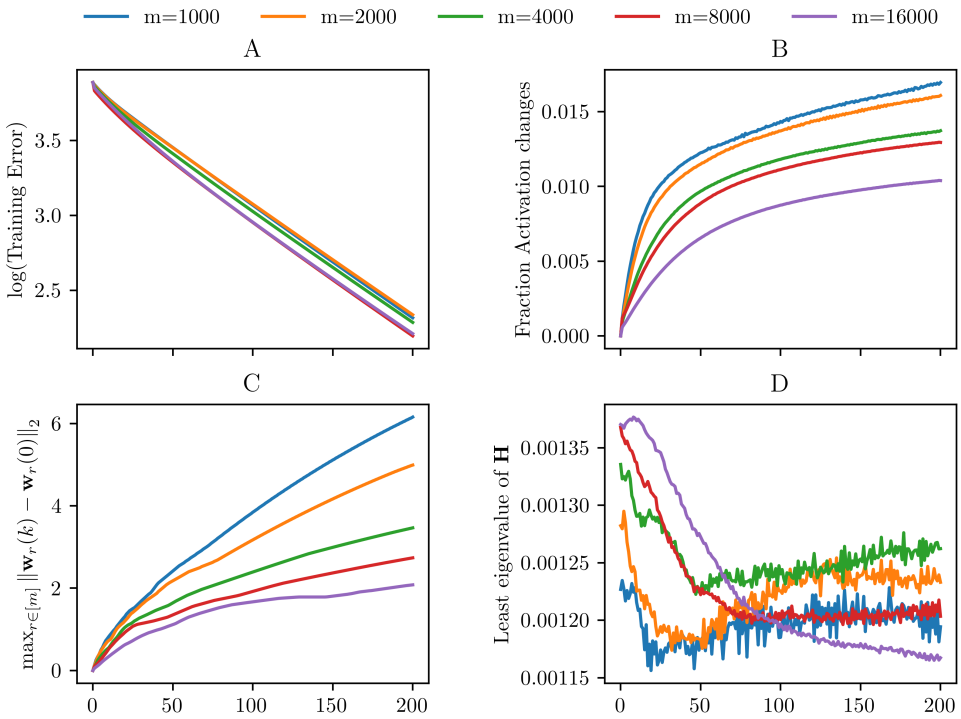

Figure 1: Results for synthetic data when training the neural network with gradient descent for $n = 100$, $d = 1000$.

$m \neq 1000$ so that all training losses are the same as the training loss for $m = 1000$ in the first epoch to more clearly see the difference in convergence rate for different values of m. In Figure 1(B) we plot the percentage of all $r \in \{0, 1, 2, \ldots, m\}$, $i \in \{0, 1, 2, \ldots, n\}$ such that the activation pattern changes sign, i.e., such that it holds that $\text{sign}((\mathbf{w}_r(k, \omega))^* \mathbf{x}_i) \neq \text{sign}((\mathbf{w}_r(0, \omega))^* \mathbf{x}_i)$. This illustrates the result of Lemma 2.8.

In Figure 1(C) we plot the maximum distance the weights has from their initialisation, aiming to illustrate the result in Lemma 3.2.

In Figure 1(D) we consider how the least eigenvalue of $\mathbf{H}$, an important measure to understand the convergence properties of the training loss, changes during the iterations and for different values of m. This illustrates Theorem 2.5, and showcases an estimate on an bound on the convergence rate.

The full code to generate the results shown in the figures, written in Python, is available on GitHub: https://github.com/Elliotepsteino/Gradient-Descent.git.

6 Discussion

The analysis focus only on the training loss and not the test loss, which is a limitation on the strength of the result. Also, a change of the proof to allow for accelerated methods to optimise the training loss would make the result more practically useful. The dependence

on n^6 for the over-parametrization of the number of weights limits the practical scope to use the result to theoretically ensure convergence for algorithms used in practice. Further, most neural networks used in practice are deep, which limits the scope of the result. This weakness is addressed in [8], a result that requires more involved arguments.

Despite this, the result is novel in that with small limitations on the input data, convergence to zero training loss in polynomial time is guaranteed with a high probability over the random initialisation of the weights.

6.1 Further directions

In order to extend the analysis to deep neural networks, i.e., if f has the form

$$f = (a)^* \sigma\Big(\mathbf{W}^{(H)} \sigma(\mathbf{W}^{(H-1)} ... \sigma(\mathbf{W}^{(1)}))\Big), \tag{133}$$

were $\mathbf{W}^1, \mathbf{W}^2, \dots \mathbf{W}^H$ are matrices and H is the depth of the network. One can follow the analysis made in [8]. When the depth of the network, H, is equal to one, the dependence on n for the over-parametrization is also improved from $\Omega(n^6)$ to $\Omega(n^4)$. In a similar way as in [7], it is possible to write the prediction dynamics as $\frac{df(t,\omega)}{dt} = \Big(\sum_{j=1}^{H} \mathbf{G}^j\Big)(y - f(t,\omega))$. It can be shown that all $\mathbf{G}^j$ are positive semi-definite, thus, the analysis focuses on showing that $\mathbf{G}^{(H)}$ is positive definite. To show this, it is used that at each step $k = 0, 1, ...,$ $\mathbf{G}(k)$ is close in 2-norm to a recursively defined matrix $\mathbf{K}^{(H)}$ that is not dependant on the weights and that is positive definite if no two inputs are parallel, in analogy to the present treatment of $\mathfrak{H}$.

7 Appendix

In the appendix a number of basic theorems that are used repeatedly are listed for an easy reference. Most theorems will not be proved, but a reference will be given.

Theorem 7.1. *Let* $\mathbf{w}$ *be a standard normal random vector and let* $\|\mathbf{x}\|_2^2 = 1$*. Then it holds that*

$$\mathbf{w}^* \mathbf{x} \sim \mathcal{N}(0, 1) \tag{134}$$

Proof of Theorem 7.1. Using that all components of a standard normal random vector is independent standard normal random variables, and that $\|\mathbf{x}\|_2^2 = 1$ ensures that

$$\mathbf{w}^* \mathbf{x} = \sum_{i=1}^{d} w_i x_i \sim \mathcal{N}\left(0, \sum_{i=1}^{d} x_i^2\right) = \mathcal{N}(0, 1). \tag{135}$$

The proof of Theorem 7.1 is thus completed.

□

Theorem 7.2 (Markov)**.** *If* X *is a nonnegative random variable and* $a > 0$ *then it holds that*

$$P(X \geq a) \leq \frac{\mathbb{E}[X]}{a}. \tag{136}$$

Proof of Theorem 7.2. See the book [9]. □

Theorem 7.3. *Let* $X \colon \mathbb{R} \to \mathbb{R}^d$ *be a vector valued function. Then it holds that*

$$\left\| \int_0^t X(s) ds \right\|_2 \leq \int_0^t \|X(s)\|_2 ds \tag{137}$$

Proof of Theorem 7.3. See the book [9]. □

Theorem 7.4 (Hoeffding)**.** *Let $(\Omega, F, \mathbb{P})$ be a probability space, let $N \in \mathbb{N}$ and let $X_n \colon \Omega \to [0,1]$, be independent random variables. Then*

$$\mathbb{P}\left(\frac{1}{N}\left|\sum_{n=1}^{N}(X_n - \mathbb{E}[X_n])\right| \geq \epsilon\right) \leq 2\exp(-2\epsilon^2 N). \tag{138}$$

Proof. See the article [13]. □

Theorem 7.5 (Union Bound)**.** *Let $\mathbb{P}$ be a probability measure on $(\Omega, \mathcal{F})$. Then it holds for all $A_i \in \mathcal{F}$, $i \in \mathbb{N}$ that*

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{n} \mathbb{P}(A_i). \tag{139}$$

Proof of Theorem 7.5. See the book [9]. □

Lemma 7.6. *Let $\mathbf{x} \in R^n$, let $\|.\|_2 \colon \mathbb{R}^n \to \mathbb{R}$ be the Euclidean norm on $\mathbb{R}^n$, and let $\|.\|_1 \colon \mathbb{R}^n \to \mathbb{R}$ satisfy for all $b = (b_1, b_2, \ldots, b_n) \in \mathbb{R}^n$ that $\|b\|_1 = \sum_{i=1}^{n} |b_i|$. Then it holds that*

$$\|\mathbf{x}\|_2 \leq \|\mathbf{x}\|_1 \leq \sqrt{n}\|\mathbf{x}\|_2 \tag{140}$$

Proof of Lemma 7.6. See the book [1]. □

Lemma 7.7. *Let A be a square matrix, let $\|A\|_2$ be the matrix 2-norm and let $\|A\|_F$ be the frobenius norm. Then it holds that*

$$\|A\|_2 \leq \|A\|_F. \tag{141}$$

Proof of Lemma 7.7. See the book [1]. □

Lemma 7.8. *For all square, symmetric matrices A and B it holds that*

$$|\lambda_{min}(A) - \lambda_{min}(B)| \leq \|A - B\|_2 \tag{142}$$

Proof of Lemma 7.8. See the book [1]. □

Lemma 7.9. *Let $n \in \mathbb{N}$, assume that $\mathbf{H} \in \mathbb{R}^{n \times n}$ is a square matrix. Then for all $\mathbf{x} \in \mathbb{R}^n$ it holds that*

$$\mathbf{x}^* \mathbf{H} \mathbf{x} \geq \lambda_{min}(\mathbf{H})\|\mathbf{x}\|_2^2. \tag{143}$$

Proof of Lemma 7.9. See the book [1]. □

References

[1] Larisa Beilina, Evgenii Karchevskii, and Mikhail Karchevskii. *Numerical Linear Algebra: Theory and Applications*. Springer, 2017.

[2] Alon Brutzkus and Amir Globerson. "Globally Optimal Gradient Descent for a ConvNet with Gaussian Inputs". In: *CoRR* abs/1702.07966 (2017). arXiv: 1702.07966. URL: http://arxiv.org/abs/1702.07966.

[3] Lenaic Chizat and Francis Bach. *On the Global Convergence of Gradient Descent for Over-parameterized Models using Optimal Transport*. 2018. arXiv: 1805.09545 [math.OC].

[4] Amit Daniely. "SGD Learns the Conjugate Kernel Class of the Network". In: *CoRR* abs/1702.08503 (2017). arXiv: 1702.08503. URL: http://arxiv.org/abs/1702.08503.

[5] J. Deng et al. "ImageNet: A Large-Scale Hierarchical Image Database". In: *CVPR09*. 2009.

[6] Simon S. Du et al. "Gradient Descent Learns One-hidden-layer CNN: Don't be Afraid of Spurious Local Minima". In: *CoRR* abs/1712.00779 (2017). arXiv: 1712.00779. URL: http://arxiv.org/abs/1712.00779.

[7] Simon S. Du et al. *Gradient Descent Provably Optimizes Over-parameterized Neural Networks*. 2018. arXiv: 1810.02054 [cs.LG].

[8] Simon Du et al. "Gradient Descent Finds Global Minima of Deep Neural Networks". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. Long Beach, California, USA: PMLR, 2019, pp. 1675–1685. URL: http://proceedings.mlr.press/v97/du19c.html.

[9] Rick Durrett. *Probability: Theory and Examples*. 5th ed. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019. DOI: 10.1017/9781108591034.

[10] Elisabeth Epstein, Filip Christiansen, and Elliot Epstein. "Ultrasound Image Analysis using Deep Neural Networks to Discriminate Benign and Malignant Ovarian Tumours - A Comparision to Subjective Expert Assessment". In:

[11] Rong Ge, Jason D. Lee, and Tengyu Ma. "Learning One-hidden-layer Neural Networks with Landscape Design". In: *CoRR* abs/1711.00501 (2017). arXiv: 1711.00501. URL: http://arxiv.org/abs/1711.00501.

[12] Benjamin D. Haeffele and René Vidal. "Global Optimality in Tensor Factorization, Deep Learning, and Beyond". In: *CoRR* abs/1506.07540 (2015). arXiv: 1506.07540. URL: http://arxiv.org/abs/1506.07540.

[13] Wassily Hoeffding. "Probability Inequalities for Sums of Bounded Random Variables". In: *Journal of the American Statistical Association* (1963), pp. 13–30. URL: https://doi.org/10.1080/01621459.1963.10500830.

[14] Kenji Kawaguchi. *Deep Learning without Poor Local Minima*. 2016. arXiv: 1605.07110 [stat.ML].

[15] Yuanzhi Li and Yingyu Liang. *Learning Overparameterized Neural Networks via Stochastic Gradient Descent on Structured Data*. 2018. arXiv: 1808.01204 [cs.LG].

[16] Quynh N. Nguyen and Matthias Hein. "The loss surface of deep and wide neural networks". In: *CoRR* abs/1704.08045 (2017). arXiv: 1704.08045. URL: http://arxiv.org/abs/1704.08045.

[17] Mahdi Soltanolkotabi. "Learning ReLUs via Gradient Descent". In: *CoRR* abs/1705.04591 (2017). arXiv: 1705.04591. URL: http://arxiv.org/abs/1705.04591.

[18] Yifan Sun et al. *Summarizing CPU and GPU Design Trends with Product Data*. 2019. arXiv: 1911.11313 [cs.DC].

[19] Yuandong Tian. "An Analytical Formula of Population Gradient for two-layered ReLU network and its Applications in Convergence and Critical Point Analysis". In: *CoRR* abs/1703.00560 (2017). arXiv: 1703.00560. URL: http://arxiv.org/abs/1703.00560.

[20] Luca Venturi, Afonso S. Bandeira, and Joan Bruna. *Spurious Valleys in Two-layer Neural Network Optimization Landscapes*. 2018. arXiv: 1802.06384 [math.OC].

[21] Bo Xie, Yingyu Liang, and Le Song. "Diversity Leads to Generalization in Neural Networks". In: *CoRR* abs/1611.03131 (2016). arXiv: 1611.03131. URL: http://arxiv.org/abs/1611.03131.