International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:14, No:6, 2020

Convergence Analysis of Training Two-Hidden-Layer Partially Over-Parameterized ReLU Networks via Gradient Descent

Zhifeng Kong

Abstract—Over-parameterized neural networks have attracted a great deal of attention in recent deep learning theory research, as they challenge the classic perspective of over-fitting when the model has excessive parameters and have gained empirical success in various settings. While a number of theoretical works have been presented to demystify properties of such models, the convergence properties of such models are still far from being thoroughly understood. In this work, we study the convergence properties of training two-hidden-layer partially over-parameterized fully connected networks with the Rectified Linear Unit activation via gradient descent. To our knowledge, this is the first theoretical work to understand convergence properties of deep over-parameterized networks without the *equally-wide-hidden-layer* assumption and other unrealistic assumptions. We provide a probabilistic lower bound of the widths of hidden layers and proved linear convergence rate of gradient descent. We also conducted experiments on synthetic and real-world datasets to validate our theory.

Keywords—Over-parameterization, Rectified Linear Units (ReLU), convergence, gradient descent, neural networks.

I. Introduction

THERE is a recent breakthrough in deep neural networks theory: the expressivity and generalization of neural networks become surprisingly good as the number of parameters exceeds the number of training samples [1]. This setting, namely over-parameterization, challenges the classic view that excessive number of parameters leads to over-fitting. This observation was further supported by empirical results in various settings [2]–[4].

Despite its empirical success, the theoretical properties of over-parameterized models remain ill-understood. As one of the first steps in understanding over-parameterized models, the training process of fully connected over-parameterized neural networks with the Rectified Linear Unit (ReLU) activation and (stochastic) gradient descent optimizer is only partially understood. For shallow networks where there is only one hidden layer, the convergence of the training process has been proven by a few prior works [5]–[9]. For deep networks there there are ≥ 2 hidden layers, the convergence was endorsed under the assumption that all hidden layers are equally wide [10]–[13]. We denote it as the *equally-wide-hidden-layer* assumption.

In this work, we provide convergence analysis to training two-hidden-layer partially over-parameterized ReLU networks using gradient descent with infinitely small learning rates. Different from prior works on deep architectures, we do not require the *equally-wide-hidden-layer* assumption in this work. Particularly, we show convergence when the second hidden layer is over-parameterized and much wider than the first hidden layer. Therefore, previous analysis on deep ReLU networks do not apply to our setting.

The technique in this work is to generalize the minimum-eigenvalue framework in [6] to the two-hidden-layer case. We follow the intuition from [5] and [6] that during training process, the weights and activation patterns do not change much, and the least eigenvalue λ_0 of the Gram matrix is lower bounded. However, it is not trivial to extend the one-hidden-layer analysis because the expressions are much more complicated for two hidden layers. To resolve this problem, we applied a more fine-grained statistical and numerical analysis to achieve applicable probabilistic bounds for certain parameters, and finally provided a lower bound for the widths of hidden layers.

We make two contributions in this work. First, we show that when the second hidden layer is over-parameterized and much wider than the first hidden layer (which could be either over-parameterized or $\Theta(1)$), gradient descent with infinitely small learning rates converges with high probability. We provided a lower bound on the widths of hidden layers that leads to convergence. Second, we conducted experiments on synthetic, MNIST [14] and Fashion-MNIST [15] datasets. Experimental results on the loss, λ_0, weight change, and pattern change rates support our theoretical analysis.

II. Related Works

It was addressed that expressivity and generalization of neural networks become surprisingly good under the over-parameterized setting [1]. As the first fatal property to understand over-parameterized neural networks, convergence was empirically studied in various settings [2]–[4].

Specifically, since the ReLU activation by [16] gained huge success in deep learning, many previous works were presented to understand the convergence properties of fully connected over-parameterized ReLU networks trained by (stochastic) gradient descent. These works can be classified into two groups based on their network architecture: shallow networks (where there is only one hidden layer) and deep networks (where there are multiple hidden layers).

Shallow networks. Widely believed, understanding the one-hidden-layer (namely two-layer in many papers)

Zhifeng Kong is with the Department of Computer Science and Engineering, University of California, San Diego, La Jolla, CA 92093, USA (e-mail: z4kong@eng.ucsd.edu).

over-parameterized architecture is crucial to understanding deeper networks: if there is only one hidden layer, then it is possible to examine every parameter and every node, so that the dynamics during training process can be explicitly analyzed. An early work under this setting by [17] showed that when data is generated by a linearly separable function, then stochastic gradient descent finds a global minimum of the Hinge loss. Later on, a surprising phenomenon was discovered by [5]: the activation patterns do not change much during training process. Based on this observation, they proved that for cross-entropy loss, gradient descent converges to ϵ loss if the width of the hidden layer is polynomial in $1/\epsilon$. [6], the closest to our work, proved that gradient descent converges to zero squared error loss under a more practical setting where the width is independent of the accuracy. The core idea is that the convergence rate is controlled by the least eigenvalue of the Gram matrix [18]–[20] when we analyze the output dynamics (instead of the weight dynamics). This matrix form was further analyzed by [7] and proved to approximate an integral operator. If the target function has a low-rank approximation w.r.t. the eigenspaces of this integral operator, then gradient descent approximately applies the powers of this integral operator on the target function, which yields a linear convergence rate. In other directions, [8] generalized the convergence results to adaptive gradient methods, and [9] improved the $\mathcal{O}(n^6)$ bound on width in [6] to $\mathcal{O}(n^2)$, where n is the number of training samples.

Deep networks: As a first step to analyze deep over-parameterized ReLU networks, deep linear networks were proved to enjoy the global-minima convergence property under various settings [21]–[26]. Then, a number of theoretical works were developed to demystify the optimization landscape of deep over-parameterized ReLU networks trained with (stochastic) gradient descent [10]–[13]. In detail, [10] and [12] proved convergence of (stochastic) gradient descent trained on deep over-parameterized ReLU networks for classification and regression problems respectively. In both papers, global minima is found by (stochastic) gradient descent in polynomial time with a polynomial bound on width. Their approaches are different in that [10] followed the observation in [5] and showed that the parameters are closed to their Gaussian initialization, while [12] proved equivalence between deep over-parameterized neural networks and finite-width neural tangent kernel (NTK) in [18]. Then, [13] improved the $\tilde{\Omega}(n^{24})$ bound on width and $\mathcal{O}(n^6)$ iteration complexity in [12] to $\tilde{\Omega}(n^8)$ and $\mathcal{O}(n^2)$. Besides analysis on general deep networks, [11] made a specific analysis to three-layer (that is, two-hidden-layer) over-parameterized ReLU network, which was proved able to learn three-layer neural networks equipped with smooth activation functions efficiently via stochastic gradient descent.

Interestingly, in [6], an approach based on the Gram matrices was discussed to generalize the convergence analysis to deep over-parameterized ReLU networks. However, due to the Hoeffding's inequality, every hidden layer has to be equally wide (or at least over-parameterized) to guarantee that the Gram matrices are closed to initial. In fact, such *equally-wide-hidden-layer* assumption exists in *all* the works mentioned above regarding convergence properties of training deep ReLU networks with (stochastic) gradient descent and squared error loss (see Assumption 3.6 in [10], Theorem 2 in [11], Section 2 and Section 3 in [12], and Section 2 in [13]). However, there is one recent work without the *equally-wide-hidden-layer* assumption. Instead of proving specifically for deep neural networks, [27] showed that for a general class of over-parameterized nonlinear learning problems, (stochastic) gradient descent takes a surprisingly short path i.e. the parameters follow an almost straight route towards optima. As a cost to its generality, a few strong assumptions (Jacobian is Lipschitz with bounded least singular value and spectrum norm) are made, which rules out the popular deep ReLU networks. In addition, the bounds in the assumptions could be extreme for deep neural networks, so their conclusions may be weakened.

In this paper, we follow the idea from [6] and generalized the minimum-eigenvalue framework to the two-hidden-layer architecture. By applying a more fine grained analysis, we do not need the *equally-wide-hidden-layer* assumption any more: we proved that if the second hidden layer is over-parameterized and the much wider than the first hidden layer (which could be either over-parameterized or $\Theta(1)$), gradient descent with infinitely small step size converges to zero training loss with high probability. To our knowledge, this is the first work that proves the convergence of gradient descent on a kind of *deep* ReLU network without the *equally-wide-hidden-layer* assumption and other unrealistic assumptions.

III. Preliminaries

In this paper, we consider the class of over-parameterized neural networks with two hidden layers and ReLU activation. The network is fully connected, taking inputs in $\mathbb{R}^d$ and outputs in $\mathbb{R}$. Particularly, we consider the case when the network is partially over-parameterized: the second hidden layer has a large amount of neurons and is much wider than the first hidden layer. We do not constrain the width of the first hidden layer; it could be either over-parameterized or $\Theta(1)$.

A. Notations

First, we formally define our network structure. The widths of the input layer, first hidden layer, and the second hidden layer are d, h_1, and h_2, respectively. The output is a real value. The weight matrix connecting the input layer and the first hidden layer is $A = \left(A_1^\top, \cdots, A_{h_1}^\top\right)^\top \in \mathbb{R}^{h_1 \times d}$, and the weight matrix connecting the hidden layers is $B = \left(B_1^\top, \cdots, B_{h_2}^\top\right)^\top \in \mathbb{R}^{h_2 \times h_1}$. Each hidden layer has activation function $\sigma(x) = ReLU(x) = \max(x, 0)$. The output is computed by the dot product of weight vector $w \in \{-1, 1\}^{h_2}$ and the output of the second hidden layer with a normalizing factor $1/\sqrt{h_1 h_2}$. Therefore, given an input $x \in \mathbb{R}^d$, the output is

$$f(x; A, B, w) = \frac{1}{\sqrt{h_1 h_2}} w^\top \sigma(B\sigma(Ax)) \tag{1}$$

Suppose we have n different training samples $\{(x_i, y_i)\}_{i=1}^n$. Then, the prediction of each sample x_i is

$$u_i = f(x_i; A, B, w),\ 1 \le i \le n \tag{2}$$

We initialize the parameters in the following way: each element in w is drawn $i.i.d.$ from the uniform distribution over $\{-1, +1\}$, and each element in A and B is drawn $i.i.d.$ from the standard normal distribution $\mathcal{N}(0, 1)$. Then, each row of A and B are normalized to have ℓ_2 norm ≤ 1. For the training data, we assume that each sample is normalized and its label is bounded by some constant C: for any $1 \le i \le n$, $\|x_i\|_2 = 1$, $|y_i| \le C$. We use the ℓ_2 norm as our loss function, defined by

$$L = \frac{1}{2}\sum_{i=1}^n (u_i - y_i)^2 = \frac{1}{2}\|u - y\|_2^2 \tag{3}$$

To minimize this loss, we apply gradient descent over A and B and constantly update these two weight matrices. In order to make rigorous theoretical analysis, we assume that the learning rate is infinitely small so that we can formulate the update rule in a differential form:

$$\frac{dA_r}{dt} = -\frac{\partial L}{\partial A_r},\ 1 \le r \le h_1 \tag{4}$$

$$\frac{dB_p}{dt} = -\frac{\partial L}{\partial B_p},\ 1 \le p \le h_2 \tag{5}$$

We define $1\{\text{statement}\}$ to be 1 if the statement is true and 0 if it is not true. Finally, the following variables take time t as the parameter since they update with t:

$$A = A(t), B = B(t), u = u(t), L = L(t) \tag{6}$$

In the rest of the paper, we omit the time parameter t for conciseness when there is no ambiguity.

B. Problem Definition

The dynamics of the output vector $u(t)$ at time t can be calculated through the chain rule in the multivariable case:

$$\begin{aligned}\frac{d}{dt}u_i(t) &= \sum_{p=1}^{h_2}\left\langle \frac{\partial u_i}{\partial B_p}, \frac{dB_p}{dt}\right\rangle + \sum_{r=1}^{h_1}\left\langle \frac{\partial u_i}{\partial A_r}, \frac{dA_r}{dt}\right\rangle \\ &= \sum_{p=1}^{h_2}\left\langle \frac{\partial u_i}{\partial B_p}, -\frac{\partial L}{\partial B_p}\right\rangle + \sum_{r=1}^{h_1}\left\langle \frac{\partial u_i}{\partial A_r}, -\frac{\partial L}{\partial A_r}\right\rangle \\ &= \sum_{j=1}^{n}\left(\sum_{p=1}^{h_2}\left\langle \frac{\partial u_i}{\partial B_p}, \frac{\partial u_j}{\partial B_p}\right\rangle + \sum_{r=1}^{h_1}\left\langle \frac{\partial u_i}{\partial A_r}, \frac{\partial u_j}{\partial A_r}\right\rangle\right) \\ &\quad \times (y_j - u_j)\end{aligned} \tag{7}$$

Therefore, the dynamics of $u(t)$ can be expressed in the following closed form:

$$\frac{d}{dt}u(t) = H(t)(y - u(t)) \tag{8}$$

where H is an $n \times n$ matrix with the (i, j)-th element

$$H_{ij} = \sum_{p=1}^{h_2}\left\langle \frac{\partial u_i}{\partial B_p}, \frac{\partial u_j}{\partial B_p}\right\rangle + \sum_{r=1}^{h_1}\left\langle \frac{\partial u_i}{\partial A_r}, \frac{\partial u_j}{\partial A_r}\right\rangle \tag{9}$$

In Section IV, we prove convergence of the dynamics (8) under the following partial over-parameterization assumption (Assumption 1) and the minimal eigenvalue assumption (Assumption 2).

C. The Partial Over-Parameterization Assumption

Assumption 1. *The network is partially over-parameterized:* $h_2 \gg h_1$.

D. The Minimal Eigenvalue Assumption

In order to design a vital tool to analyze H, we first compute the explicit expression of H.

Proposition 1. *$\forall i, j \in \{1, \cdots, n\}$, $h_1 h_2 H_{ij}$ is equal to*

$$\begin{aligned}&\sum_{p=1}^{h_2}\sum_{r=1}^{h_1}\left(x_i^\top A_r A_r^\top x_j + x_i^\top x_j B_{pr}^2\right) \\ &\quad \times 1\left\{B_p^\top y_i > 0, B_p^\top y_j > 0, A_r^\top x_i > 0, A_r^\top x_j > 0\right\} \\ &+ x_i^\top x_j \cdot \sum_{r=1}^{h_1}\sum_{p=1}^{h_2}\sum_{\substack{q=1\\ q\ne p}}^{h_2} w_p w_q B_{pr} B_{qr} \\ &\quad \times 1\left\{B_p^\top y_i > 0, B_q^\top y_j > 0, A_r^\top x_i > 0, A_r^\top x_j > 0\right\}\end{aligned} \tag{10}$$

where $y_i = \sigma(Ax_i)$.

Then, we define $H^\infty = \mathbb{E}_{A(0)}\mathbb{E}_{B(0)}\mathbb{E}_w H(0)$, a fatal bridge to analyze H. This is because the $\sum_r \sum_{p\ne q}$ term in H_{ij} is intractable but will vanish as we take the expectation over w, and terms are easily bounded as we take the expectation over $A(0)$ and $B(0)$. Given the concise and symmetric expression of H^∞, it is not hard to prove that H^∞ is positive semi-definite (PSD) in **Proposition** 2.

Proposition 2. *H^∞ is PSD.*

Therefore, it is reasonable for us to make the following assumption in this paper:

Assumption 2. $\lambda_0 = \lambda_{min}(H^\infty) > 0$.

In the main result (Theorem 3), λ_0 appears to be a key factor in the convergence rate, and $\lambda_0/2$ is used to lower bound the smallest eigenvalue of $H(t)$ throughout the proof in Section IV-B.

IV. Main Result

In this section, we prove convergence of the dynamics (8). In detail, we prove that if we use gradient descent to train a two-hidden-layer ReLU network with an infinitely small learning rate, and if the second hidden layer is over-parameterized and much wider than the first one, then we can expect the training loss converges to zero exponentially with high probability. This result is a theoretical extension to the convergence analysis in [6], where the one-hidden-layer setting was proved. In addition, our result does not overlap with analysis for the multiple-layer setting in [10]–[13] because they require each hidden layer should be equally wide (or lower bounded), which does not include situations where one hidden layer might be not wide enough. Our result does provide a positive answer to such situations. Therefore, our result is a complimentary work of their results on the convergence properties of training general two-hidden-layer over-parameterized ReLU networks with gradient descent.

A. Theorem of Convergence

Our formal result is shown in the following theorem.

Theorem 3. *Under Assumption 1 and Assumption 2, if the number of nodes of the second hidden layer satisfies* $h_2 = \Omega\left(\frac{n^{9/2}L(0)^{3/2}h_1^{3/2}}{\lambda_0^6\delta^3}\right)$*, then with probability* $\geq 1-\delta$ *over random initialization, the training error is bounded by* $L(t) \leq \exp(-\lambda_0 t)L(0)$.

In short, a lower bounded convergence rate is provided in Theorem 3 given the partially over-parameterized architecture. To keep our conclusion concise, we use the smallest eigenvalue to bound the convergence rate. Intuitively, a tighter bound induced by the mean of eigenvalues of H^∞ is likely to exist, at a cost of having a much more complicated probabilistic analysis.

B. Steps of Proof

In this section, we present four steps to prove Theorem 3. In detail, we use the minimum-eigenvalue framework, which was initially introduced in [6]. In Lemma 4, we show that $H(0)$ is closed to H^∞ and the least eigenvalue of $H(0)$ is bounded. In Lemma 5, we prove that the minimum eigenvalue of H at time t is also lower bounded. In Lemma 6, we first demonstrate that the minimum eigenvalue of H provides a convergence guarantee to the gradient descent algorithm, and then show that the change of parameters A and B is bounded. Finally, Lemma 7 ties the previous lemmas into a cycle, showing that the conditions always hold and the probability that the cycle breaks is tiny, thus finalizing the proof. The proofs of these lemmas are given in Appendix A. Compared to the analysis in [6], their bounds (especially R and R' in their paper) cannot be naturally extended to the problem in this work, we use various mathematical tools to provide applicable probabilistic bounds in the analysis.

Lemma 4. *If* $h_2 = \Omega\left(\frac{n^2}{\lambda_0^2}\log\frac{2n^2}{\delta}\right)$*, then we have*

$$\lambda_{min}(H) \geq \frac{3}{4}\lambda_0 \tag{11}$$

with probability at least $1-\delta$ *at time* $t = 0$.

In the Lemma 4, we show that with high probability, the minimum eigenvalue of $H = H(0)$ at time 0 is bounded by a constant times λ_0. In the proof, we first bound $|H_{ij} - H_{ij}^\infty|$ using the Hoeffding's inequality and the Chebyshev's inequality. Then, we are able to bound the Frobenius norm of $H - H^\infty$. Finally, the difference between their least eigenvalues can be bounded by this amount.

Lemma 5. *Suppose at time* t*, there exists a small constant* $c > 0$ *such that* $\|A_r(t) - A_r(0)\|_2 \leq R_A\ \forall r$*,* $\|B_p(t) - B_p(0)\|_2 \leq R_B\ \forall p$*, where* $R_A = \mathcal{O}\left(\frac{c\delta\lambda_0}{nh_1}\right)$*,* $R_B = \mathcal{O}\left(\frac{c\delta\lambda_0}{n}\right)$*. Then with probability at least* $1-\delta$ *over initialization,* $\lambda_{min}(H(t)) \geq \lambda_0/2$.

In Lemma 5, we show that if at time t, the weight matrices A and B are close to initialization, then the minimum eigenvalue of $H = H(t)$ is also bounded by a constant times λ_0. The core idea in the proof is that when we calculate the expectation of $|H_{ij}(0) - H_{ij}(t)|$, we divide it into two cases: one with activation pattern change and the other one with no activation pattern change. If no activation pattern changes, then slight modification to A and B must yield slight change of H_{ij}. On the other side, the probability that some activation pattern changes can be bounded by a small constant. Therefore, we obtain our desired result.

Lemma 6. *Suppose* $\lambda_{min}(H(s)) \geq \lambda_0/2\ \forall s \in [0,t]$*. Then the training error is bounded as* $L(t) \leq \exp(-\lambda_0 t)L(0)$*. The modification of* A *and* B *from initialization can be bounded as* $\|A_r(t) - A_r(0)\|_2 \leq R'_A\ \forall 1 \leq r \leq h_1$ *and* $\|B_p(t) - B_p(0)\|_2 \leq R'_B\ \forall 1 \leq p \leq h_2$*, where* $R'_A = \frac{2}{\lambda_0}\sqrt{\frac{2nL(0)}{h_1 h_2^{2/3}}}$ *and* $R'_B = \frac{2}{\lambda_0}\sqrt{\frac{2nL(0)}{h_2}}$.

In the Lemma 6, we show that if at time s, the minimum eigenvalue of $H = H(s)$ is bounded, then the training loss converges to 0 with an exponential rate, and the weight matrices are closed to initialization with high probability. In the proof, we first calculate the derivative of the loss function w.r.t. time t, and then bound the loss with an exponential rate. To prove that A and B are closed to initialization, we also bound their derivatives first, and then do the integration.

Lemma 7. *If* $h_2 = \Omega\left(\frac{n^{9/2}L(0)^{3/2}h_1^{3/2}}{\lambda_0^6\delta^3}\right)$*, then with high probability the following statements hold: for any* $t > 0$*,* $\|A_r(t) - A_r(0)\|_2 \leq R'_A\ \forall 1 \leq r \leq h_1$*,* $\|B_p(t) - B_p(0)\|_2 \leq R'_B\ \forall 1 \leq p \leq h_2$*, and* $L(t) \leq \exp(-\lambda_0 t)L(0)$.

In the Lemma 7, we show that if $R'_A < R_A, R'_B < R_B$, then Lemma 5 and Lemma 6 are able to form a cycle: the conditions consistently hold as time t goes to infinity. The proof is finished by drawing a contradiction.

C. Finalizing the Proof

Proof: According to **Lemma**s 4-7, the theorem holds if $R'_A \leq R_A$ and $R'_B \leq R_B$. This is equivalent to $h_2^{2/3} = \Omega\left(\frac{n^3L(0)h_1}{\lambda_0^4\delta^2}\right)$. Finally, notice that

$$\mathbb{E}L(0) = \|y\|_2^2 + n = \Theta(n) \tag{12}$$

Thus, by Markov's inequality, with probability $\geq 1-\delta$, $L(0) = \mathcal{O}(n/\delta)$. Therefore, the conclusion holds with probability at least $1 - 2\delta - h_2^{-1/3} = 1 - 2\delta - \mathcal{O}(\delta)$. Since there is only a constant in front of δ, we finish the proof. ■

V. Experiments

In this section, we present experiments to validate our theory.

A. Settings

We conducted experiments on several datasets including synthetic datasets, the MNIST dataset [14], and the Fashion-MNIST dataset [15]. To construct the synthetic dataset, we randomly select 50/100 points on the unit ball in $\mathbb{R}^{100}$ and generate their labels in $[0,1]$ by random. These

two datasets are denoted as synthetic-50 and synthetic-100, respectively. To construct real-world datasets, we randomly select 50 samples with label c_1 and 50 samples with label c_2 from the MNIST dataset or the Fashion-MNIST dataset, denoted as MNIST/Fashion-c_1c_2. In this paper, we select $(c_1, c_2) = (0, 1), (1, 7)$ for MNIST and $(c_1, c_2) = (0, 1), (7, 9)$ for Fashion-MNIST. For both MNIST and Fashion-MNIST datasets, samples with label 0 and label 1 are visually quite distinct, so we also look at other cases where training samples with different labels also share some similarities. This encourages us to conduct experiments on MNIST-17 (where both 1 and 7 have a long straight line in the main body) and Fashion-MNIST-79 (where both sneakers and ankle boots are in the shoe-shape). Some samples are shown in Fig. 1.

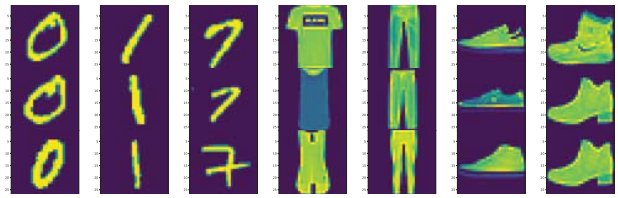
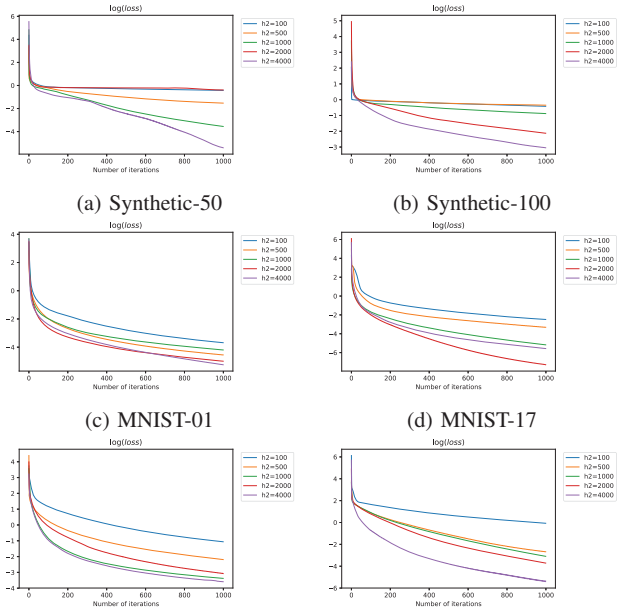


Fig. 1 Samples of training data. Each column shows three samples from the selected training datasets. The datasets are (from left to right,) MNIST-0, MNIST-1, MNIST-7, Fashion-0 (T-shirt/top), Fashion-1 (Trouser), Fashion-7 (Sneaker), Fashion-9 (Ankle boot).



(a) Synthetic-50 (b) Synthetic-100

(c) MNIST-01 (d) MNIST-17

(e) Fashion-01 (f) Fashion-79

Fig. 2 The logarithm of loss function $\log L(t)$ with 10 network structures on different training datasets.

We also surveyed various network structures. The width of the second hidden layer is selected from $\mathcal{W} = \{100, 500, 1000, 2000, 4000\}$, and the first hidden layer is set to have five neurons. Therefore, for each dataset, we did five sub-experiments on networks with $h_1 = 5, h_2 \in \mathcal{W}$. We set the learning rate as small as 0.0002 to simulate the continuous dynamics. At each iteration t ($t \leq 1000$), we recorded the following values:

- Weight change: the Frobenius norm of the difference between the weight matrices at time t and the initial ones, computed as $\|A(t) - A(0)\|_F$ (first hidden layer) and $\|B(t) - B(0)\|_F$ (second hidden layer);
- Pattern change rate: the average percentage of activation pattern changes on the whole training set for each hidden layer. This is calculated by #total activation pattern changes for i-th hidden layer divided by $n \times h_i$, $i = 1, 2$;
- The minimum eigenvalue $\lambda_{min}(H(t))$;
- The ℓ_2 loss $L(t)$.

B. Results

In Fig. 2, we present the ℓ_2 losses on the six datasets. The full results can be found in Appendix B.

These results are consistent with our theory according to the following observations. First, the logarithm of loss decreases at a bounded rate as t increases, showing that the gradient descent algorithm converges at a bounded rate. This directly substantiates our main result in **Theorem** 3. In addition, the smallest eigenvalue $\lambda_{min}(H(t))$ remains stable during the training process, and a larger $\lambda_{min}(H(t))$ typically yields a faster convergence rate. This is especially shown on real-world datasets. Finally, both the weight change rate and pattern change rate increase as t increases, but are still below a low threshold. As the networks become wider the threshold generally gets lower. These observations comply with the minimum-eigenvalue framework in our proof.

VI. Conclusion

In this paper, we surveyed the convergence properties of two-hidden-layer ReLU networks trained by gradient descent with ℓ_2 loss, where the network is partially over-parameterized. In Section IV, we proved that if the width of the second hidden layer is large enough, then with high probability the gradient descent algorithm with infinitely small step size converges to an optimal solution that achieves zero training loss. The main part of the proof is divided into 4 lemmas in Section IV-B, which form a cycle and reveal that the minimum eigenvalue of H is lower bounded both initially and during the training process. Since the ℓ_2 loss can be written as a quadratic form in H, this minimum eigenvalue thus provides guarantee of the convergence rate. We also conducted experiments on synthetic dataset and real-world datasets in Section V. The experimental results are consistent with our theory in that convergence of dynamics, stability of the minimum eigenvalues, small weight change and small pattern change rates are verified.

Based our theory and experiments, there are several further directions worth deeper exploitation.

- In this paper we proved convergence for the *continuous* dynamics, while it is widely acknowledged in the area of dynamical systems [28] that discrete dynamics are generally more complicated and thus harder to analyze. Therefore, it is significant to prove the convergence theory with a constant step size.

- Unlike the one-hidden-layer over-parameterized ReLU network in [6], in which λ_0 can be easily calculated since it has an explicit form, λ_0 for the two-hidden-layer network remains mysterious. Since λ_0 is a vital term in the convergence rate, its strict positiveness and relationship to data as well as network structure are crucial in understanding the convergence properties. To achieve this, a method that either provides an analytic form or an approximation algorithm of λ_0 is needed.
- It is observed that the logarithm of loss does not decrease at a constant rate in experiments. Instead, it decreases much faster during the initial training status, and achieves a stable rate after many iterations. However, the minimum eigenvalue $\lambda_{min}(H(t))$ appears to be stable during the whole training process. Therefore, it is likely that the dynamics are controlled by some other indicators that dominate the training procedure at the very first beginning.
- Finally, we believe that using more advanced inequalities in our proof will help decrease the bound of the width in our main theorem. We used classical bounds in most of our proof, where it is possible to introduce more refined concentration and tail inequalities, especially when the dimension of data is high.

Appendix A. Omitted Proofs

Since we have many sums in this paper, we make some shorthand here:

$$\sum_r = \sum_{r=1}^{h_1}\ ;\ \sum_{p(q)=1} = \sum_{p(q)=1}^{h_2}\ ;\ \sum_{q:q\neq p} = \sum_{\substack{q=1\\ q\neq p}}^{h_2}\ ;\ \sum_{p\neq q} = \sum_{p=1}^{h_2}\sum_{q:q\neq p}\ ;\ \sum_{i(j)} = \sum_{i(j)=1}^{n} \tag{13}$$

Also, the activation patterns of each neuron have to be decided, so we make the following shorthand:

$$1_i\{r\} = 1\{A_r^\top x_i > 0\}\ ;\ 1_i\{p(q)\} = 1\{B_{p(q)}^\top \sigma(Ax_i) > 0\} \tag{14}$$

Proof of ***Proposition*** *1*

Proof: First, we specify the calculations of $\frac{\partial f(x)}{\partial B_p}$, $\frac{\partial f(x)}{\partial A_r}$:

$$\begin{aligned}
\frac{\partial f(x)}{\partial B_p} &= \frac{1}{\sqrt{h_1h_2}}\cdot\frac{\partial}{\partial B_p}(w_p\sigma(B_p^\top\sigma(Ax)))\\
&= \frac{1}{\sqrt{h_1h_2}}\cdot w_p\sigma(Ax)\cdot 1\{B_p^\top\sigma(Ax)>0\}\\
\frac{\partial f(x)}{\partial A_r} &= \frac{1}{\sqrt{h_1h_2}}\cdot\frac{\partial}{\partial A_r}\sum_p w_p\sigma\left(\sum_r B_{pr}\sigma(A_r^\top x)\right)\\
&= \frac{1}{\sqrt{h_1h_2}}\cdot\sum_p w_p\cdot 1\{B_p^\top\sigma(Ax)>0\}\cdot\frac{\partial}{\partial A_r}(B_{pr}\sigma(A_r^\top x))\\
&= \frac{1}{\sqrt{h_1h_2}}\cdot\sum_p w_pB_{pr}x\cdot 1\{B_p^\top\sigma(Ax)>0, A_r^\top x>0\}
\end{aligned} \tag{15}$$

Therefore, we obtain

$$\begin{aligned}
h_1h_2\cdot H_{ij} = &\ y_i^\top y_j\cdot\sum_p 1\left\{B_p^\top y_i>0, B_p^\top y_j>0\right\}\\
&+x_i^\top x_j\cdot\sum_r\left(\sum_p w_pB_{pr}\cdot 1\{B_p^\top y_i>0, A_r^\top x_i>0\}\right)\\
&\quad\cdot\left(\sum_p w_pB_{pr}\cdot 1\{B_p^\top y_j>0, A_r^\top x_j>0\}\right)\\
= &\ y_i^\top y_j\cdot\sum_p 1\left\{B_p^\top y_i>0, B_p^\top y_j>0\right\}\\
&+x_i^\top x_j\cdot\\
&\sum_{p,r} B_{pr}^2\cdot 1\left\{B_p^\top y_i>0, B_p^\top y_j>0, A_r^\top x_i>0, A_r^\top x_j>0\right\}\\
&+x_i^\top x_j\cdot\\
&\sum_r\sum_{p\neq q} w_pw_qB_{pr}B_{qr}\cdot 1\left\{B_p^\top y_i>0, B_q^\top y_j>0, A_r^\top x_i>0, A_r^\top x_j>0\right\}\\
&\sum_{p,r}\left(x_i^\top A_rA_r^\top x_j + x_i^\top x_jB_{pr}^2\right)\cdot\\
&1\left\{B_p^\top y_i>0, B_p^\top y_j>0, A_r^\top x_i>0, A_r^\top x_j>0\right\}\\
&+x_i^\top x_j\cdot\sum_r\sum_{p\neq q} w_pw_qB_{pr}B_{qr}\cdot\\
&1\left\{B_p^\top y_i>0, B_q^\top y_j>0, A_r^\top x_i>0, A_r^\top x_j>0\right\}
\end{aligned} \tag{16}$$

where $y_i = \sigma(Ax_i)$. ■

Proof of ***Proposition*** *2*

Proof: Since $\mathbb{E}_w w_pw_q = 0$, we have

$$h_1h_2\cdot H_{ij}^\infty = \mathbb{E}\sum_{p,r}\left(x_i^\top A_rA_r^\top x_j + x_i^\top x_jB_{pr}^2\right)\cdot 1_{i,j}\{p,r\} \tag{17}$$

Therefore, we only need to prove that for any non-zero vector $v\in\mathbb{R}^n$,

$$\sum_{i=1}^n\sum_{j=1}^n\sum_{p,r} v_iv_j\left(x_i^\top A_rA_r^\top x_j + x_i^\top x_jB_{pr}^2\right)\cdot 1_{i,j}\{p,r\}\geq 0 \tag{18}$$

We change the summing order into $\sum_{p,r}\sum_{i,j}$ and for each pair (p,r), we show the sums of two parts (the $x_i^\top A_rA_r^\top x_j$ term and the $x_i^\top x_jB_{pr}^2$ term) over i,j are non-negative. For the first part,

$$\sum_{i,j} v_iv_jx_i^\top A_rA_r^\top x_j\cdot 1_{i,j}\{p,r\} = \left(\sum_i v_iA_r^\top x_i\cdot 1_i\{p,r\}\right)^2 \tag{19}$$

For the second part,

$$\sum_{i,j} v_iv_jx_i^\top x_jB_{pr}^2\cdot 1_{i,j}\{p,r\} = B_{pr}^2\left\|\sum_i v_ix_i\cdot 1_i\{p,r\}\right\|_2^2 \tag{20}$$

As a result, H^∞ is PSD. ■

Proof of ***Lemma*** *4*

Proof: We rewrite H_{ij} in the following way:

$$H_{ij} = \frac{1}{h_1h_2}\sum_{p,r} X_{pr} \tag{21}$$

where

$$X_{pr} = \left(x_i^\top A_rA_r^\top x_j + x_i^\top x_jB_{pr}^2\right)\cdot 1_{i,j}\{p,r\} + \sum_{q:q\neq p} x_i^\top x_jw_pw_qB_{pr}B_{qr}\cdot 1_i\{p,r\}1_j\{q,r\} \tag{22}$$

Let $\bar{B}_{:,r} = \sum_q w_qB_{qr}$. By the Chebyshev's inequality, for any $\tau>0$,

$$P(|\bar{B}_{:,r}|\geq\tau)\leq\frac{\sum_q B_{qr}^2}{\tau^2}\leq\frac{1}{\tau^2} \tag{23}$$

Thus, with probability at least $1-\frac{h_1}{\tau^2}$, we have $|X_{pr}| \leq \tau+2$ for any p, r and $\tau > 0$. Applying this bound to Hoeffding's inequality, we have

$$P\left(\left|H_{ij}-H_{ij}^{\infty}\right| \geq \delta\right) \leq 2\left(1-\frac{h_1}{\tau^2}\right)\exp\left(-\frac{h_1 h_2 \delta^2}{(t+2)^2}\right) \tag{24}$$

If we let $\tau+2 = h_1^{1/2}h_2^{k/2}$ with $k \in (0,1)$, we can simplify the inequality as

$$P\left(\left|H_{ij}-H_{ij}^{\infty}\right| \geq \delta\right) \leq 2\left(1-\frac{1}{h_2^k}\right)\exp\left(-\frac{h_2^{1-k}\delta^2}{2}\right) \tag{25}$$

This is equivalent to: with probability $\geq 1-\delta$, for any $0 < k < 1$,

$$\left|H_{ij}-H_{ij}^{\infty}\right| \leq \sqrt{\frac{2}{h_2^{1-k}}\log\frac{2(h_2^k-1)}{\delta h_2^k}} \tag{26}$$

Now, we set $k = \left(\log\frac{\sqrt{2}}{\sqrt{2}-\sqrt{\delta}}\right)/\log h_2$ and $\delta \leq 1/2$. The above inequality can be simplified as

$$\left|H_{ij}-H_{ij}^{\infty}\right| \leq \sqrt{\frac{2}{h_2}\frac{\sqrt{2}}{\sqrt{2}-\sqrt{\delta}}\log\sqrt{2/\delta}} \leq \sqrt{\frac{2}{h_2}\log\frac{2}{\delta}} \tag{27}$$

Then, we apply this bound for all (i,j) pairs, and obtain that with probability at least $1-\delta$

$$\left|H_{ij}-H_{ij}^{\infty}\right| \leq \sqrt{\frac{2}{h_2}\log\frac{2n^2}{\delta}} \tag{28}$$

which gives that

$$\|H-H^{\infty}\|_2^2 \leq \frac{2n^2}{h_2}\log\frac{2n^2}{\delta} \tag{29}$$

Similar to the previous analysis, if $h_2 \geq \frac{32n^2}{\lambda_0^2}\log\frac{2n^2}{\delta}$, we have the desired result. ∎

*Proof of **Lemma** 5*

Proof: By the definition of H, we have

$$\begin{aligned}|H_{ij}(t)-H_{ij}(0)| \leq\ & \frac{1}{h_1h_2}\{\sum_{p,r}\left|x_i^\top A_r(t)A_r(t)^\top x_j \cdot 1_{ij}^{(t)}\{p,r\} - x_i^\top A_r(0)A_r(0)^\top x_j \cdot 1_{ij}^{(0)}\{p,r\}\right|\cdot \\ & +|x_i^\top x_j|\sum_{p,r}|B_{pr}(t)^2\cdot 1_{ij}^{(t)}\{p,r\} - B_{pr}(0)^2\cdot 1_{ij}^{(0)}\{p,r\}| \\ & +|x_i^\top x_j|\sum_{p,r}|w_p|\Big|\sum_{q:q\neq p} w_q B_{pr}(t)B_{qr}(t)\cdot 1_i^{(t)}\{p,r\}1_j^{(t)}\{q,r\} \\ & - \sum_{q:q\neq p} w_q B_{pr}(0)B_{qr}(0)\cdot 1_i^{(0)}\{p,r\}1_j^{(0)}\{q,r\}\Big|\Big\}\end{aligned} \tag{30}$$

According to Lemma 3.2 in [6], we have

$$P\left(1_i^{(t)}\{r\} \neq 1_i^{(0)}\{r\}\right) \leq \frac{2R_A}{\sqrt{2\pi}} \tag{31}$$

and

$$\begin{aligned}P\left(1_j^{(t)}\{p\} = 1_j^{(0)}\{p\}\right) \geq\ & P\left(1_j^{(t)}\{p\} = 1_j^{(0)}\{p\} \,\middle|\, 1_i^{(t)}\{r\} = 1_i^{(0)}\{r\}\ \forall i\right) \\ & P\left(1_i^{(t)}\{r\} = 1_i^{(0)}\{r\}\right)^{h_1} \\ \geq\ & \left(1-\frac{2R_A}{\sqrt{2\pi}}\right)^{h_1}\left(1-\frac{2R_B}{\sqrt{2\pi}}\right)\end{aligned} \tag{32}$$

which indicates that

$$P\left(1_j^{(t)}\{p\} \neq 1_j^{(0)}\{p\}\right) \leq \frac{2(h_1R_A+R_B)}{\sqrt{2\pi}} \tag{33}$$

Thus, we have

$$P\left(1_i^{(t)}\{p,r\} \neq 1_i^{(0)}\{p,r\}\right) \leq \frac{2(h_1R_A+R_A+R_B)}{\sqrt{2\pi}} \tag{34}$$

and as a result,

$$P\left(1_{ij}^{(t)}\{p,r\} \neq 1_{ij}^{(0)}\{p,r\}\right) \leq \frac{4(h_1R_A+R_A+R_B)}{\sqrt{2\pi}} \tag{35}$$

Therefore,

$$\begin{aligned}\mathbb{E}|H_{ij}(t)-H_{ij}(0)| =\ & P(\text{activation does not change})\cdot\mathbb{E}\left(|H_{ij}(t)-H_{ij}(0)| \,\middle|\, \text{activation does not change}\right) \\ & +P(\text{activation changes})\cdot\mathbb{E}\left(|H_{ij}(t)-H_{ij}(0)| \,\middle|\, \text{activation changes}\right) \\ \leq\ & \mathbb{E}\left(|H_{ij}(t)-H_{ij}(0)| \,\middle|\, \text{activation does not change}\right) \\ & +\frac{4(h_1R_A+R_A+R_B)}{\sqrt{2\pi}}\cdot\mathbb{E}\left(|H_{ij}(t)-H_{ij}(0)| \,\middle|\, \text{activation changes}\right)\end{aligned} \tag{36}$$

When there some activation pattern changes, we apply the simple bound

$$\mathbb{E}\left(|H_{ij}(t)-H_{ij}(0)| \mid \text{activation changes}\right) \leq \mathbb{E}|H_{ij}(t)| + \mathbb{E}|H_{ij}(0)| \tag{37}$$

where

$$\begin{aligned}\mathbb{E}|H_{ij}(t)| \leq\ & \frac{1}{h_1h_2}\sum_{p,r}\mathbb{E}\left(|x_i^\top A_r(t)A_r^\top(t)x_j| + |x_i^\top x_j B_{pr}(t)^2|\right) \\ \leq\ & \frac{1}{h_1h_2}\sum_{p,r}\mathbb{E}\left(\|A_r(t)\|_2^2 + B_{pr}^2(t)\right) \\ \leq\ & \frac{2}{h_1h_2}\sum_{p,r}\mathbb{E}\left(R_A^2 + \|A_r(0)\|_2^2 + B_{pr}^2(0) + (B_{pr}(t)-B_{pr}(0))^2\right) \\ \leq\ & 2\left(1+R_A^2+\frac{1}{h_1h_2}\sum_r\left(\|B_p(0)\|_2^2+R_B^2\right)\right) \\ \leq\ & 2\left(1+R_A^2+\frac{1+R_B^2}{h_2}\right)\end{aligned} \tag{38}$$

and similarly,

$$\mathbb{E}|H_{ij}(0)| \leq 2\left(1+\frac{1}{h_2}\right) \tag{39}$$

Since δ can be arbitrarily small, both R_A and R_B can be $\mathcal{O}(1)$. Therefore,

$$\mathbb{E}|H_{ij}(t)-H_{ij}(0)| = \mathcal{O}\left(\max\left(\mathbb{E}\left(|H_{ij}(t)-H_{ij}(0)| \,\middle|\, \text{activation does not change}\right), h_1R_A, R_B\right)\right) \tag{40}$$

Now, we bound $\mathbb{E}\left(|H_{ij}(t)-H_{ij}(0)| \mid \text{activation does not change}\right)$, the expected difference between $H_{ij}(t)$ and $H_{ij}(0)$ when no activation pattern changes. This expected difference is upper bounded by the sum of three terms:

$$\begin{aligned}& \mathbb{E}\left(|H_{ij}(t)-H_{ij}(0)| \,\middle|\, \text{activation does not change}\right) \\ \leq\ & \frac{1}{h_1h_2}\sum_{p,r}\mathbb{E}\left|x_i^\top(A_r(t)A_r^\top(t)-A_r(0)A_r^\top(0))x_j\right| \\ & +\frac{1}{h_1h_2}\sum_{p,r}\mathbb{E}|x_i^\top x_j(B_{pr}^2(t)-B_{pr}^2(0))| \\ & +\frac{1}{h_1h_2}\sum_r\sum_{p\neq q}\mathbb{E}\,|B_{pr}(t)B_{qr}(t)-B_{pr}(0)B_{qr}(0)| \\ \leq\ & \frac{1}{h_1h_2}\sum_{p,r}\mathbb{E}\left(|x_i^\top(A_r(t)-A_r(0))|\right) \\ & +\left(|x_j^\top(A_r(t)-A_r(0))| + |x_i^\top(A_r(t)-A_r(0))x_j^\top(A_r(t)-A_r(0))|\right) \\ & +\frac{1}{h_1h_2}\sum_p\mathbb{E}\sum_r\left(2|B_{pr}(0)(B_{pr}(t)-B_{pr}(0))| + (B_{pr}(t)-B_{pr}(0))^2\right) \\ & +\frac{1}{h_1h_2}\sum_{p,q}\mathbb{E}\sum_r\left(|B_{pr}(t)(B_{qr}(t)-B_{qr}(0))| + |B_{qr}(0)(B_{pr}(t)-B_{pr}(0))|\right) \\ \leq\ & \frac{1}{h_1h_2}\sum_{p,r}\|A_r(t)-A_r(0)\|_2 + \|A_r(t)-A_r(0)\|_2 + \|A_r(t)-A_r(0)\|_2^2 \\ & +\frac{1}{h_1h_2}\sum_p 2\|B_p(0)\|_2\|B_p(t)-B_p(0)\|_2 + \|B_p(t)-B_p(0)\|_2^2 \\ & +\frac{1}{h_1h_2}\sum_{p,q}(\|B_p(0)\|_2 + \|B_q(0)\|_2 + \|B_q(t)-B_q(0)\|_2)\|B_p(t)-B_p(0)\|_2 \\ \leq\ & \frac{1}{h_1h_2}\left(\sum_{p,r}(2R_A+R_A^2)+\sum_p(2R_B+R_B^2)+\sum_{p,q}(2R_B+R_B^2)\right) \\ =\ & \mathcal{O}\left(\max\left(R_A, R_B\right)\right)\end{aligned} \tag{41}$$

Therefore,

$$\mathbb{E}|H_{ij}(t)-H_{ij}(0)| = \mathcal{O}\left(\max\left(h_1R_A, R_B\right)\right) \tag{42}$$

In order to bound $\|H(t)-H(0)\|_2 \leq \frac{\lambda_0}{4}$, we only need to set $\|H(t)-H(0)\|_F \leq \frac{\lambda_0}{4}$. By Markov inequality, this is satisfied with probability $\geq 1-\delta$ when

$$\max\left(h_1 R_A, R_B\right) = \mathcal{O}\left(\frac{c\delta\lambda_0}{n}\right) \tag{43}$$

where c is a small positive number. Finally, we obtain that when

$$R_A = \mathcal{O}\left(\frac{c\delta\lambda_0}{nh_1}\right),\ R_B = \mathcal{O}\left(\frac{c\delta\lambda_0}{n}\right) \tag{44}$$

we have with high probability over initialization, $\lambda_{min}(H(t)) \geq \lambda_0/2$. ∎

Proof of ***Lemma*** *6*

Proof: The derivative of loss is bounded by

$$L'(t) = -(y-u(t))^\top \frac{du(t)}{dt} = -(y-u(t))^\top H(t)(y-u(t)) \leq -\lambda_0 L(t) \tag{45}$$

Therefore, we know $\frac{d}{dt}\left(\exp(\lambda_0 t)L(t)\right) \leq 0$, indicating that $L(t) \leq \exp(-\lambda_0 t)L(0)$.

Next, we bound the difference between A, B and their initialization. This is achieved by integrating the gradient of A and B. Therefore, we bound their gradient first.

We set $\tau = h_2^{1/6}$ in the proof of **Lemma** 4. Then, we obtain with probability at least $1-1/h_2^{1/3}$, the sum $|\sum_q w_q B_{qr}|$ can be up bounded by $h_2^{1/6}$. Then, for $s \in [0,t]$ and $\forall 1 \leq r \leq h_1$,

$$\begin{aligned}
\left\|\frac{d}{ds}A_r(s)\right\|_2 &= \left\|\sum_{i=1}^n (y_i - u_i(s))\frac{\partial f(x_i)}{\partial A_r(s)}\right\|_2 \\
&= \left\|\sum_{i=1}^n (y_i - u_i(s))\sum_p w_p B_{pr}(s) x_i \cdot 1_i^{(s)}\{p,r\}\right\|_2 / \sqrt{h_1 h_2} \\
&\leq \frac{\|y-u(s)\|_1 h_2^{1/6}}{\sqrt{h_1 h_2}} \\
&\leq \|y-u(s)\|_2 \sqrt{\frac{n}{h_1 h_2^{2/3}}} \\
&\leq \sqrt{\frac{2n\exp(-\lambda_0 s)L(0)}{h_1 h_2^{2/3}}}
\end{aligned} \tag{46}$$

Similarly, $\forall 1 \leq p \leq h_2$,

$$\begin{aligned}
\left\|\frac{d}{ds}B_p(s)\right\|_2 &= \left\|\sum_{i=1}^n (y_i - u_i(s))\frac{\partial f(x_i)}{\partial B_p(s)}\right\|_2 \\
&= \left\|\sum_{i=1}^n (y_i - u_i(s)) w_p \sigma(A(s)x_i) \cdot 1_i^{(s)}\{p\}\right\|_2 / \sqrt{h_1 h_2} \\
&= \left\|\sum_{i=1}^n (y_i - u_i(s)) w_p A(s) x_i \cdot 1_i^{(s)}\{p,r\}\right\|_2 / \sqrt{h_1 h_2} \\
&\leq \|y-u(s)\|_1 \|A(s)\|_2 / \sqrt{h_1 h_2} \\
&\leq \|y-u(s)\|_1 \|A(s)\|_F / \sqrt{h_1 h_2} \\
&\leq \sqrt{n/h_2}\|y-u(s)\|_2 \\
&\leq \sqrt{2n\exp(-\lambda_0 s)L(0)/h_2}
\end{aligned} \tag{47}$$

Then, we are able to bound the modification through integration

$$\|A_r(t) - A_r(0)\|_2 \leq \int_0^t \left\|\frac{d}{ds}A_r(s)\right\|_2 ds \leq \frac{2}{\lambda_0}\sqrt{\frac{2nL(0)}{h_1 h_2^{2/3}}} \tag{48}$$

$$\|B_p(t) - B_p(0)\|_2 \leq \int_0^t \left\|\frac{d}{ds}B_p(s)\right\|_2 ds \leq \frac{2}{\lambda_0}\sqrt{\frac{2nL(0)}{h_2}} \tag{49}$$

∎

Proof of ***Lemma*** *7*

Proof: Given the bounds of h_1 or h_2 above, we are able to show that $R'_A < R_A$ and $R'_B < R_B$ using basic arithmetic. Suppose the conclusion does not hold at time t. According to **Lemma** 6, with high probability there exists $s \leq t$ such that $\lambda_{min}(H(s)) < \lambda_0/2$. Then, according to **Lemma** 5, with high probability there exists r such that $\|A_r(t) - A_r(0)\|_2 > R_A$, or there exists p such that $\|B_p(t) - B_p(0)\|_2 > R_B$. This indicates that the following infimum exists:

$$t_0 = \inf_{t>0}\{\exists r : \|A_r(t) - A_r(0)\|_2 \geq R_A \text{ or } \exists p : \|B_p(t) - B_p(0)\|_2 \geq R_B\} \tag{50}$$

Since $A_r(t)$ and $B_p(t)$ change continuously, we have that $\|A_r(t_0) - A_r(0)\|_2 \leq R_A$, $\|B_p(t_0) - B_p(0)\|_2 \leq R_B$, and equality holds for at least one of them. Then by **Lemma** 5, with high probability $\lambda_{min}(H(t')) \geq \lambda_0/2\ \forall t' \leq t_0$, and thus $\|A_r(t_0) - A_r(0)\|_2 \leq R'_A\ \forall r$ and $\|B_p(t_0) - B_p(0)\|_2 \leq R'_B\ \forall p$ by **Lemma** 6. Recall that $R'_A < R_A$ and $R'_B < R_B$, we draw a contradiction here. ∎

Appendix B. Omitted Experiments

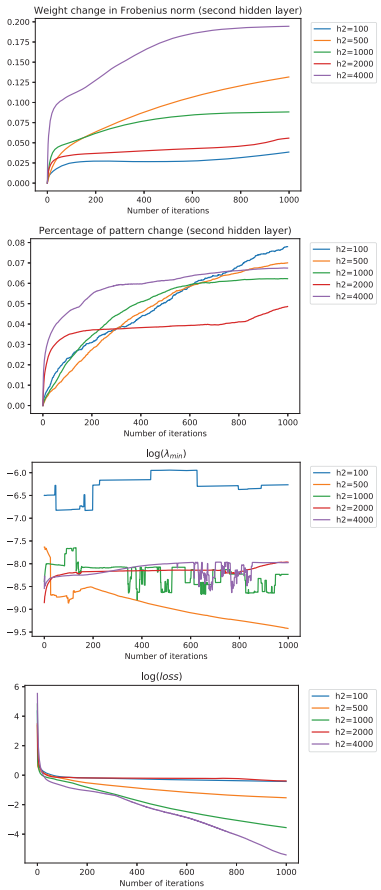


Fig. 3 Results on synthetic-50 data

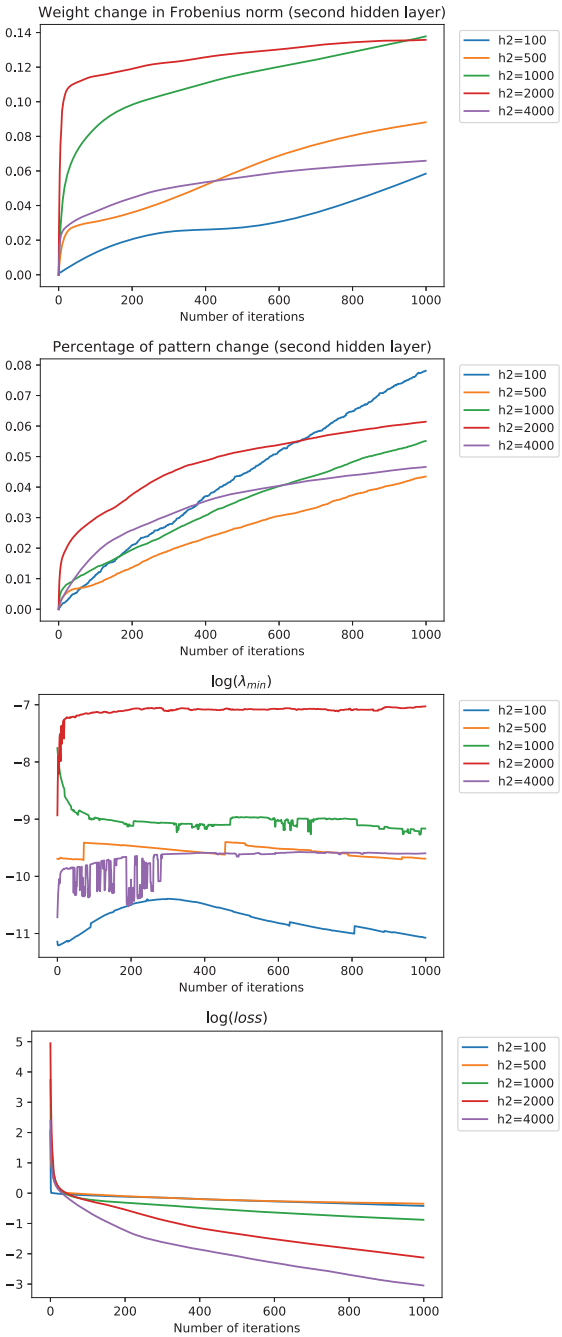


Fig. 4 Results on synthetic-100 data

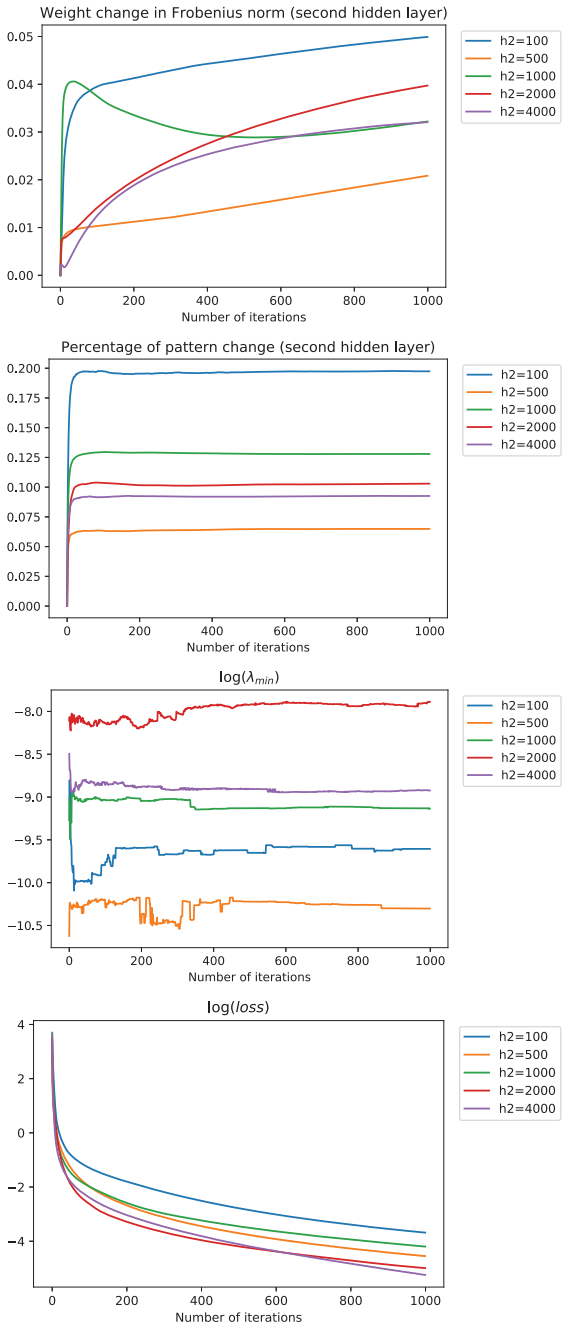


Fig. 5 Results on MNIST-01

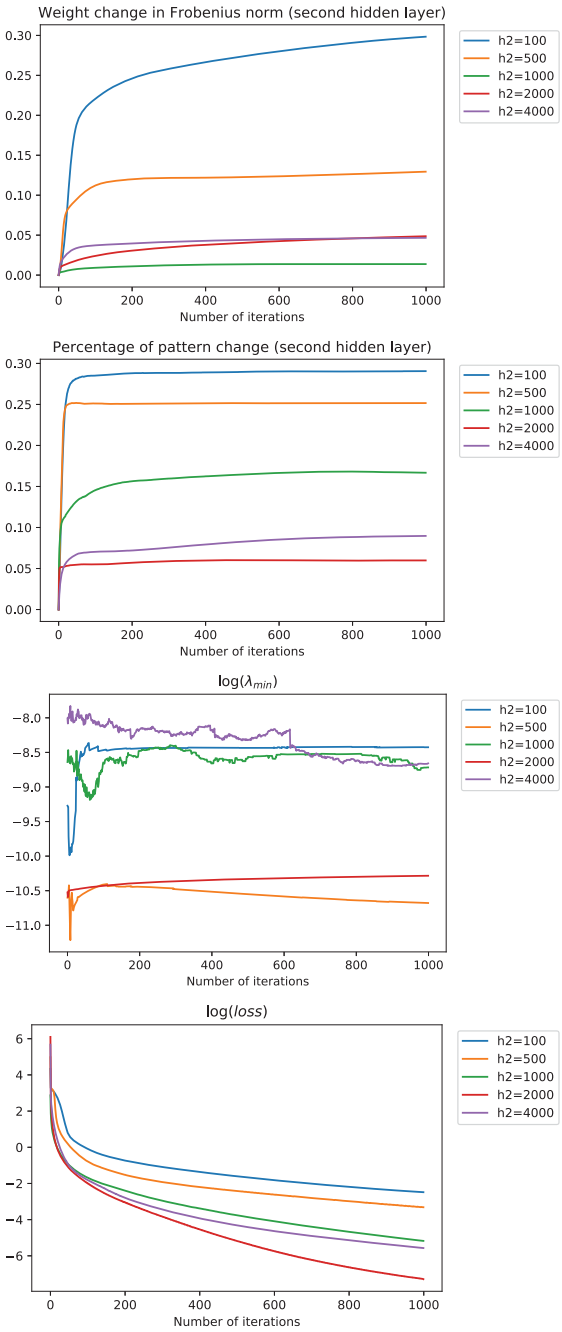


Fig. 6 Results on MNIST-17

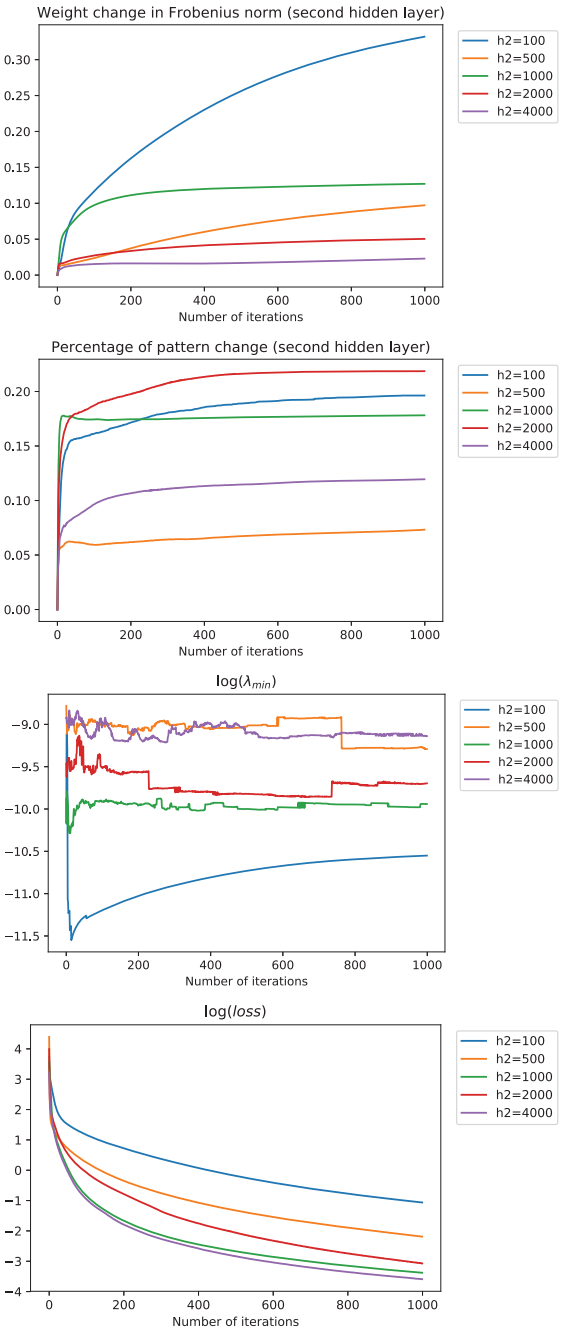


Fig. 7 Results on Fashion-01

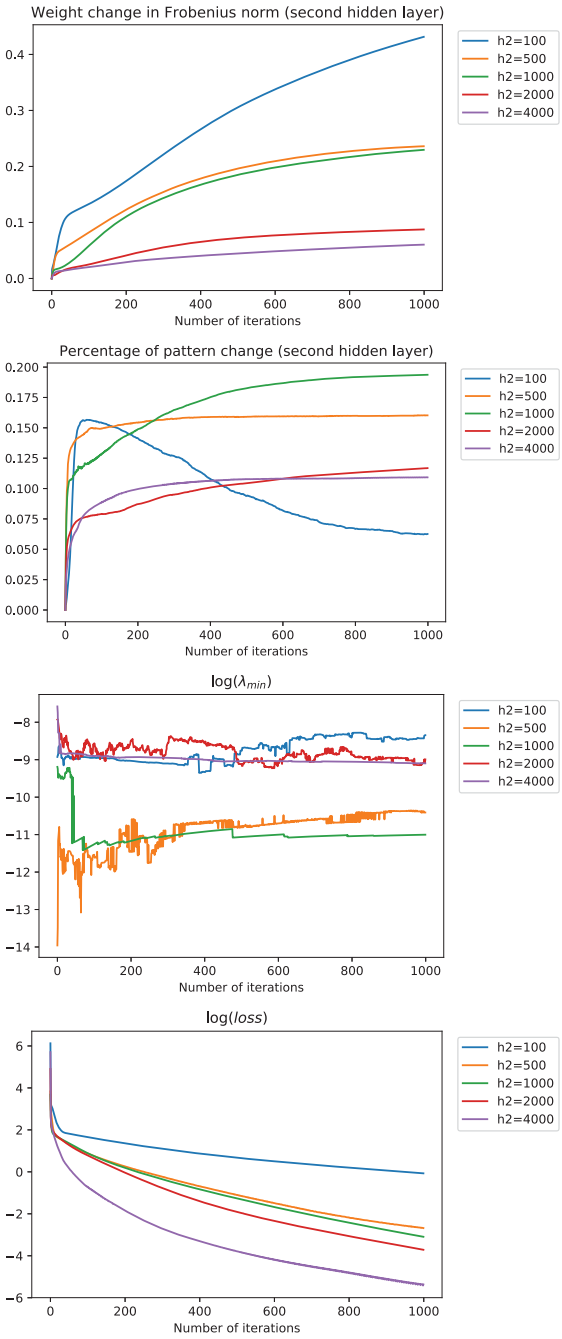


Fig. 8 Results on Fashion-79

REFERENCES

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

[2] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

[3] Levent Sagun, Utku Evci, V Ugur Guney, Yann Dauphin, and Leon Bottou. Empirical analysis of the hessian of over-parametrized neural networks. *arXiv preprint arXiv:1706.04454*, 2017.

[4] Ji Xu, Daniel J Hsu, and Arian Maleki. Benefits of over-parameterization with em. In *Advances in Neural Information Processing Systems*, pages 10662–10672, 2018.

[5] Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in Neural Information Processing Systems*, pages 8157–8166, 2018.

[6] Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*, 2018.

[7] Lili Su and Pengkun Yang. On learning over-parameterized neural networks: A functional approximation prospective. *arXiv preprint arXiv:1905.10826*, 2019.

[8] Xiaoxia Wu, Simon S Du, and Rachel Ward. Global convergence of adaptive gradient methods for an over-parameterized neural network. *arXiv preprint arXiv:1902.07111*, 2019.

[9] Samet Oymak and Mahdi Soltanolkotabi. Towards moderate overparameterization: global convergence guarantees for training shallow neural networks. *arXiv preprint arXiv:1902.04674*, 2019.

[10] Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Stochastic gradient descent optimizes over-parameterized deep relu networks. *arXiv preprint arXiv:1811.08888*, 2018.

[11] Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. *arXiv preprint arXiv:1811.04918*, 2018.

[12] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. *arXiv preprint arXiv:1811.03962*, 2018.

[13] Difan Zou and Quanquan Gu. An improved analysis of training over-parameterized deep neural networks. *arXiv preprint arXiv:1906.04688*, 2019.

[14] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[15] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[17] Alon Brutzkus, Amir Globerson, Eran Malach, and Shai Shalev-Shwartz. Sgd learns over-parameterized networks that provably generalize on linearly separable data. *arXiv preprint arXiv:1710.10174*, 2017.

[18] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580, 2018.

[19] Bo Xie, Yingyu Liang, and Le Song. Diverse neural network learns true target functions. *arXiv preprint arXiv:1611.03131*, 2016.

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

[21] Kenji Kawaguchi. Deep learning without poor local minima. In *Advances in neural information processing systems*, pages 586–594, 2016.

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

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

[24] Sanjeev Arora, Nadav Cohen, Noah Golowich, and Wei Hu. A convergence analysis of gradient descent for deep linear neural networks. *arXiv preprint arXiv:1810.02281*, 2018.

[25] Sanjeev Arora, Nadav Cohen, and Elad Hazan. On the optimization of deep networks: Implicit acceleration by overparameterization. *arXiv preprint arXiv:1802.06509*, 2018.

[26] Peter L Bartlett, David P Helmbold, and Philip M Long. Gradient descent with identity initialization efficiently learns positive-definite linear transformations by deep residual networks. *Neural computation*, 31(3):477–502, 2019.
[27] Samet Oymak and Mahdi Soltanolkotabi. Overparameterized nonlinear learning: Gradient descent takes the shortest path? *arXiv preprint arXiv:1812.10004*, 2018.
[28] Robert Devaney. *An introduction to chaotic dynamical systems*. CRC Press, 2018.