A GLOBAL CONVERGENCE THEORY FOR DEEP IMPLICIT LEARNING VIA OVER-PARAMETERIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Implicit deep learning is now receiving more attention recently since it generalizes the recursive prediction rule of many commonly used neural network architectures. Its prediction rule is provided implicitly based on the solution of an equilibrium equation. Although a line of recent studies demonstrate its superior performances, the theoretically understanding to implicit models is still limited. In general, the equilibrium equation may not well posed during the training. As a result, there is no guarantee a vanilla (stochastic) gradient descent training a nonlinear implicit model is capable to converge. This paper takes a step towards filling the gap by analyzing the gradient flow of ReLU activated implicit neural networks. For a m width implicit model with ReLU activation and n training data, we show that randomly initialized gradient flow converges to a global minimum at linear rate for the square loss function, as long as the implicit model is *over-parameterized*. A number of previous theoretical works have shown the global convergence of the (stochastic) gradient descent if the width m is polynomial large in the size of training data and the depth of the finite-layer neural network. However, the results do not apply in the implicit models since the implicit models are considered having *infinitely* many layers. Fortunately, in this paper we can show the global convergence is still hold, even the neutral networks have infinitely many layers.

1 INTRODUCTION

In this paper, we consider a new class of neural network models called *implicit deep learning* El Ghaoui et al. (2019) that is based on the implicit prediction rule. Such prediction rule is not obtained via a forward and recursive propagation through a finite number of hidden layers. Instead, it is provided implicitly by solving an equilibrium equation that contains the hidden features. Specifically, let $\boldsymbol{x} \in \mathbb{R}^d$ be an input vector, the prediction rule is obtained in the form of

$$\hat{y} = \boldsymbol{u}^T \boldsymbol{z}^* + \boldsymbol{v}^T \boldsymbol{\phi}, \tag{1}$$

$$\boldsymbol{z}^* = \sigma(\boldsymbol{A}\boldsymbol{z}^* + \boldsymbol{\phi}), \tag{2}$$

where $\sigma(\cdot)$ is the activation function, the vector $\boldsymbol{z}^* \in \mathbb{R}^m$ is the *equilibrium point* of the equilibrium equation 5, the function $\phi : \mathbb{R}^d \to \mathbb{R}^m$ is a feature map that is chosen by the designer to map the input vector $\boldsymbol{x}$ to the desired feature vector $\boldsymbol{\phi} \triangleq \phi(\boldsymbol{x})$, and $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{A}$ are the trainable parameters.

The popularity of implicit deep learning is partially due to it generalizes recursive rule of many commonly used neural network architecture. A line of recent works Bai et al. (2019); El Ghaoui et al. (2019); Bai et al. (2020) have shown that the implicit models include a number of start-of-the-art neural network architectures as special cases such as fully connected feed-forward neural network, Convolutional layers, residual nets, and recurrent units. Moreover, the implicit deep learning has also been shown to achieve performance competitive with the widely used deep networks while using significantly less parameters and memory Dabre & Fujita (2019); Dehghani et al. (2018); Bai et al. (2018).

Although the superior performance of the implicit model has been demonstrated numerically in many recent works, the theoretically understanding of the implicit model is still limited. In general, the equilibrium equation 5 cannot be solved in closed-form, and is not necessarily well-posed

in the sense that it may have no or multiple solutions. To handle the well-posedness, El Ghaoui et al. (2019) suggests to use projected gradient descent to train the model, but it needs to solve a projection subproblem at each iteration and the convergence result does not provide. A global linear convergence rate is obtained in Kawaguchi (2021) by using extra softmax layer to deal with the well-posedness issue but the result only holds for *linear* activation. In this paper, we try to fill the gap by establishing the global convergence result for the implicit model with *nonlinear* ReLU activation function. In particular, we will show the gradient flow converges to a global minimum at a linear convergence rate when the implicit model is *overparameterized*.

One of the mysteries of deep learning is randomly initialized first-order method such as (stochastic) gradient descent achieves *zero training loss*, even the labels are assigned arbitrarily Zhang et al. (2021). A widely believed explanation for this phenomena is *over-parameterization*. Despite the nonsmooth and nonconvexity of the objective function, several most recent studies Nguyen & Mondelli (2020); Oymak & Soltanolkotabi (2020); Allen-Zhu et al. (2019); Zou & Gu (2019); Du et al. (2019) have proven that a neural network with finite depth p and ReLU activation (or its approximation) can fit all the n training data perfectly if the width m of each layer is large enough, *i.e.*, $m = \mathcal{O}\left(\text{poly}(n, p)\right)$. However, their analyses does not hold for the implicit model since it has infinitely many hidden layers. Consequently, one has to ensure the unique equilibrium point always exists, and the width w is independent of the depth p.

The main contributions of this paper are listed as follows

(i) By scaling the weight matrix $\boldsymbol{A}$, we can show the unique equilibrium point $\boldsymbol{z}^*$ for each $\boldsymbol{x}$ always exists during the training if the parameters are randomly initialized, even for the nonlinear ReLU activation function.

(ii) We analyze the gradient flow of implicit model. Despite the nonsmooth and nonconvexity of the objective function, the convergence to global minimum at linear rate is guaranteed if the implicit model is overparameterized and the data is nondegenerate.

2 Preliminaries and notations

We begin with an introduction to the notations. Given a vector $\boldsymbol{x}$, $\|\boldsymbol{x}\|$ is the Euclidean norm of $\boldsymbol{x}$. Given a matrix $\boldsymbol{A}$, $\|\boldsymbol{A}\|$ is the operator norm of $\boldsymbol{A}$. If $\boldsymbol{A}$ is square matrix, then $\lambda_{\min}(\boldsymbol{A})$ and $\lambda_{\max}(\boldsymbol{A})$ denote the smallest and largest eigenvalue of $\boldsymbol{A}$, respectively, and $\lambda_{\max}(\boldsymbol{A}) = \|\boldsymbol{A}\|$. We define $[n] \triangleq \{1, 2, \cdots, n\}$.

Let $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ be a training dataset with n samples, where $\boldsymbol{x}_i \in \mathbb{R}^d$ is the input vector, and $y_i \in \mathbb{R}$ is the output. In this paper, we consider an implicit model of the following form:

$$\boldsymbol{z}^\ell = \sigma(\tilde{\gamma}\boldsymbol{A}\boldsymbol{z}^{\ell-1} + \boldsymbol{\phi}), \tag{3}$$

where $\boldsymbol{z}^\ell \in \mathbb{R}^m$ is the output of the ℓ-th layer, $\boldsymbol{A} \in \mathbb{R}^{m\times m}$ is the trainable parameter matrix, the positive number $\tilde{\gamma} \triangleq \gamma/\sqrt{m}$ for some $\gamma \in (0, 1)$ is utilized to scale the matrix $\boldsymbol{A}$ to ensure the existence of the equilibrium point $\boldsymbol{z}^* = \lim_{\ell\to\infty} \boldsymbol{z}^\ell$, $\sigma(u) = \max\{0, u\}$ is the ReLU activation function, and $\boldsymbol{\phi} \triangleq \phi(\boldsymbol{x})$ is the desired feature vector obtained by the feature map $\phi : \mathbb{R}^d \to \mathbb{R}^m$. Here we consider a simple nonlinear feature map ϕ given by

$$\phi(\boldsymbol{x}) \triangleq \frac{1}{\sqrt{m}}\sigma(\boldsymbol{W}\boldsymbol{x}), \tag{4}$$

where $\boldsymbol{W} \in \mathbb{R}^{m\times d}$ is another trainable parameter matrix. Since $\boldsymbol{z}^*$ is the limit of the sequence $\{\boldsymbol{z}^\ell\}_{\ell=1}^\infty$ with $\boldsymbol{z}_0 = \boldsymbol{0}$, it is also the equilibrium point of the equilibrium equation as follows

$$\boldsymbol{z}^* = \sigma(\tilde{\gamma}\boldsymbol{A}\boldsymbol{z}^* + \boldsymbol{\phi}). \tag{5}$$

Along with the square loss as the loss function, *i.e.*, $\ell(\hat{y}, y) = \frac{1}{2}\ell(\hat{y} - y)^2$, the objective function or the total loss $L(\boldsymbol{\theta})$ is given by

$$L(\boldsymbol{\theta}) = \sum_{i=1}^n \ell(\hat{y}_i, y_i) = \frac{1}{2}\|\hat{\boldsymbol{y}} - \boldsymbol{y}\|^2, \tag{6}$$

where $\hat{\boldsymbol{y}}$ and $\boldsymbol{y}$ are the vectors of stacking all the prediction and output, respectively, and $\boldsymbol{\theta} = \text{vec}\,(\boldsymbol{A}, \boldsymbol{W}, \boldsymbol{u}, \boldsymbol{v})$ denotes all training parameters.

The gradient flow is equivalent to the gradient descent with infinitesimal step size. Thus, the analysis of the gradient flow is considered a stepping stone towards understanding discrete algorithms. The gradient flow of the implicit model is given by the following ordinary differential equations

$$\frac{d\text{vec}\,(\boldsymbol{A})}{dt} = -\frac{\partial L(t)}{\partial \boldsymbol{A}}, \quad \frac{d\text{vec}\,(\boldsymbol{W})}{dt} = -\frac{\partial L(t)}{\partial \boldsymbol{W}}, \quad \frac{d\boldsymbol{u}}{dt} = -\frac{\partial L(t)}{\partial \boldsymbol{u}}, \quad \frac{d\boldsymbol{v}}{dt} = -\frac{\partial L(t)}{\partial \boldsymbol{v}}, \tag{7}$$

where $L(t)$ represents the value of the objective function at time t.

3 Continuous time analysis

This section includes the main results of this paper. In particular, we show that the equilibrium equation 5 is well-posed in Section 3.1 in the sense that the unique equilibrium point exists. In Section **??**, we derive the dynamics of the objective value $L(t)$ and show the convergence to a stationary point. Under a mild condition, we prove the linear convergence to global minimum in Section **??**.

We make the following assumption on initialization of matrices $\boldsymbol{A}$ and $\boldsymbol{W}$ throughout the paper.

Assumption 1 (Random initialization)**.** The entries $\boldsymbol{A}_{ij}$ and $\boldsymbol{W}_{ij}$ are randomly initialized by standard Gaussian distribution $\mathcal{N}(0, 1)$, and $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ are randomly initialized by symmetric Bernoulli random variable with values $\{\pm 1/\sqrt{m}\}$.

Remark 3.1. This is one of the broadly used initialization in practice **?**. The result obtained in this paper can be easily extend to the case where the distribution used for $\boldsymbol{A}_{ij}$, $\boldsymbol{W}_{ij}$, $\boldsymbol{u}_i$, and $\boldsymbol{v}_i$ are be replaced by *sub-Gaussian random variable* with zero mean and unit variance. To maintain a clear expression, we choose to use standard Gaussian distribution and symmetric Bernoulli distribution.

3.1 Forward propagation and well-posedness

In a general deep equilibrium model, the equilibrium equation 5 is not necessarily well-posed in the sense that it may admit no or multiple solutions. Fortunately, along with random initialization, the scalar $\tilde{\gamma} = \gamma/\sqrt{m}$ utilized on scaling the matrix $\boldsymbol{A}$ guarantees the existence and uniqueness of the equilibrium point $\boldsymbol{z}^*$. The result is based on a significant result from the field of *random matrix theory* (Vershynin, Theorem 4.4.5), which is included in the following lemma.

Lemma 3.1. Let $\boldsymbol{M}$ be an $m \times n$ random matrix with $m \geq n$ whose entries $\boldsymbol{M}_{ij}$ are independent, mean zero, unit variance, and sub-Gaussian random variable. Then *with exponentially high probability* (*i.e.*, $1 - \mathcal{O}\left(e^{-Cm}\right)$ for some absolute constant $C > 0$), one has $\|\boldsymbol{M}\| \leq c\sqrt{m}$ for some absolute constant $c > 0$.

Note that Lemma 3.1 still holds if one chooses larger constants c and C. Under Assumption 1, Lemma 3.1 implies with exponentially high probability $\|\boldsymbol{A}\| \leq c\sqrt{m}$ for some constant $c > 0$. By scaling $\boldsymbol{A}$ by the positive number $\tilde{\gamma}$, we can show the transition h in equation 3 is a contraction mapping, so that the unique equilibrium point exists.

Lemma 3.2. If $\|\boldsymbol{A}\| \leq c\sqrt{m}$ for some absolute constant $c > 0$, then for any $\gamma_0 \in (0, 1)$, the scalar $\gamma \triangleq \min\{\gamma_0, \gamma_0/c\}$ uniquely determines the existence of the equilibrium $\boldsymbol{z}^*$ for every $\boldsymbol{x}$. Moreover, $\|\boldsymbol{z}^\ell\| \leq \frac{1}{1-\gamma_0}\|\boldsymbol{\phi}\|$ for all ℓ.

Proof. Since $\tilde{\gamma}\|\boldsymbol{A}\| \leq \gamma_0 < 1$, we have $\|\boldsymbol{z}^{\ell+1} - \boldsymbol{z}^\ell\| \leq \gamma_0^\ell \|\boldsymbol{\phi}\|$. Therefore, we can show $\{\boldsymbol{z}^\ell\}_{\ell=1}^\infty$ is a Cauchy sequence, so that the unique equilibrium point $\boldsymbol{z}^*$ exists. The details of he proof is included in Appendix A.1. □

Lemma 3.2 tells us that the equilibrium points always exist if we can maintain $\|\boldsymbol{A}(t)\| \leq c\sqrt{m}$ for a constant $c > 0$ during the entire training and use a fixed scalar $\gamma = \min\{\gamma_0, \gamma_0/c\}$. In Section 3.3, we will show $\|\boldsymbol{A}(t)\| \leq 2c\sqrt{m}$ for all $t \geq 0$ if $\|\boldsymbol{A}(0)\| \leq c\sqrt{m}$ and the wide m is large enough. Thus, by using a scalar $\gamma = \min\{\gamma_0, \gamma_0/(2c)\}$, the equilibrium points always exists and no extra works (such as solving a subproblem at each iteration) are needed to ensure the equilibrium equation is well posed.

3.2 BACKWARD GRADIENT COMPUTING AND DYNAMICS OF PREDICTION $\hat{\boldsymbol{y}}(t)$

Given a general neural network with finite layers, one needs to store all intermediate parameters and apply backpropagation through all the intermediate iterations of forward pass to compute the gradients. For an implicit neural network, instead, we can derive the formula of the gradients with respect to the trainable parameters by using the *implicit function theorem*, since the equilibrium point $\boldsymbol{z}^*$ is a root of the function f given by

$$f(\boldsymbol{z}, \boldsymbol{A}, \boldsymbol{W}) = \boldsymbol{z} - \sigma\left(\tilde{\gamma}\boldsymbol{A}\boldsymbol{z} + \boldsymbol{\phi}\right). \tag{8}$$

The gradient derivations for each trainable parameters are provided in the following lemma and the proof is included in Appendix A.2.

Lemma 3.3 (Gradients of the equilibrium model)**.** If $\|\boldsymbol{A}\| \leq c\sqrt{m}$ for an absolute constant $c > 0$, then for any $\gamma_0 \in (0, 1)$, we define the scalar $\gamma \triangleq \min\{\gamma_0, \gamma_0/c\}$, and the followings hold:

(i) The partial derivatives of f with respect to $\boldsymbol{z}$, $\boldsymbol{A}$, and $\boldsymbol{W}$ are given by

$$\frac{\partial f}{\partial \boldsymbol{z}} = \left[\boldsymbol{I}_m - \tilde{\gamma}\,\mathbf{diag}(\sigma'(\tilde{\gamma}\boldsymbol{A}\boldsymbol{z} + \boldsymbol{\phi}))\boldsymbol{A}\right]^T, \tag{9}$$

$$\frac{\partial f}{\partial \boldsymbol{A}} = -\tilde{\gamma}\left[\boldsymbol{z}^T \otimes \mathbf{diag}(\sigma'(\tilde{\gamma}\boldsymbol{A}\boldsymbol{z} + \boldsymbol{\phi}))\right]^T, \tag{10}$$

$$\frac{\partial f}{\partial \boldsymbol{W}} = -\frac{1}{\sqrt{m}}\left[\boldsymbol{x}^T \otimes \mathbf{diag}(\sigma'(\boldsymbol{W}\boldsymbol{x}))\right]^T \mathbf{diag}(\sigma'(\tilde{\gamma}\boldsymbol{A}\boldsymbol{z} + \boldsymbol{\phi}))^T, \tag{11}$$

where $\tilde{\gamma} \triangleq \gamma/\sqrt{m}$.

(ii) For any vector $\boldsymbol{v}$, the following inequality holds

$$\lambda_{\min}\left\{\boldsymbol{I}_m - \tilde{\gamma}\,\mathbf{diag}(\sigma'(\boldsymbol{v}))\boldsymbol{A}\right\} > 1 - \gamma_0 > 0, \tag{12}$$

which further implies $\partial f/\partial \boldsymbol{z}$ is invertible at the equilibrium point $\boldsymbol{z}^*$.

(iii) The gradient of the objective function L with respect to $\boldsymbol{A}$ and $\boldsymbol{W}$ are given by

$$\nabla_{\boldsymbol{A}} L = \sum_{i=1}^{n} \tilde{\gamma}(\hat{y}_i - y_i)\boldsymbol{D}_i^T\left[\boldsymbol{I}_m - \tilde{\gamma}\boldsymbol{D}_i\boldsymbol{A}\right]^{-T}\boldsymbol{u}\boldsymbol{z}_i^T, \tag{13}$$

$$\nabla_{\boldsymbol{W}} L = \sum_{i=1}^{n} \frac{1}{\sqrt{m}}(\hat{y}_i - y_i)\boldsymbol{E}_i^T\left\{\boldsymbol{D}_i^T\left[\boldsymbol{I}_m - \tilde{\gamma}\boldsymbol{D}_i\boldsymbol{A}\right]^{-T}\boldsymbol{u} + \boldsymbol{v}\right\}\boldsymbol{x}_i^T, \tag{14}$$

$$\nabla_{\boldsymbol{u}} L = \sum_{i=1}^{n}(\hat{y}_i - y_i)\boldsymbol{z}_i, \tag{15}$$

$$\nabla_{\boldsymbol{v}} L = \sum_{i=1}^{n}(\hat{y}_i - y_i)\boldsymbol{\phi}_i, \tag{16}$$

where $\boldsymbol{z}_i$ is the equilibrium point for the training data $(\boldsymbol{x}_i, y_i)$, and

$$\boldsymbol{D}_i \triangleq \mathbf{diag}(\sigma'(\tilde{\gamma}\boldsymbol{A}\boldsymbol{z}_i + \boldsymbol{\phi}_i)), \quad \text{and} \quad \boldsymbol{E}_i \triangleq \mathbf{diag}(\sigma'(\boldsymbol{W}\boldsymbol{x}_i)). \tag{17}$$

Let $L(t)$ denote the objective value $L(\boldsymbol{\theta})$ at time t. The gradient flows are given by

$$\frac{d\mathrm{vec}\left(\boldsymbol{A}\right)}{dt} = -\frac{\partial L(t)}{\partial \boldsymbol{A}}, \qquad \frac{d\mathrm{vec}\left(\boldsymbol{W}\right)}{dt} = -\frac{\partial L(t)}{\partial \boldsymbol{W}}, \qquad \frac{d\boldsymbol{u}}{dt} = -\frac{\partial L(t)}{\partial \boldsymbol{u}}, \qquad \frac{d\boldsymbol{v}}{dt} = -\frac{\partial L(t)}{\partial \boldsymbol{v}}.$$

To show the global convergence, we establish the dynamics of the prediction $\hat{\boldsymbol{y}}(t) \in \mathbb{R}^n$ in the following lemma.

Lemma 3.4 (Dynamics of prediction $\hat{\boldsymbol{y}}(t)$)**.** Suppose $\|\boldsymbol{A}(s)\| \leq c\sqrt{m}$ for all $0 \leq s \leq t$. Let $\boldsymbol{X} \in \mathbb{R}^{n\times d}$, $\boldsymbol{\Phi}(s) \in \mathbb{R}^{n\times m}$, and $\boldsymbol{Z}(s) \in \mathbb{R}^{n\times m}$ be the matrices whose rows are the training data $\boldsymbol{x}_i$, feature vectors $\boldsymbol{\phi}_i$, and equilibrium points $\boldsymbol{z}_i$ at time s, respectively. Along with scalar $\gamma \triangleq \min\{\gamma_0, \gamma_0/c\}$ for any $\gamma_0 \in (0, 1)$, the dynamics of prediction $\hat{\boldsymbol{y}}(t)$ is given by

$$\begin{aligned}\frac{d\hat{\boldsymbol{y}}}{dt} &= -\left[(\gamma^2\boldsymbol{M}(t) + \boldsymbol{I}_n) \circ \boldsymbol{Z}(t)\boldsymbol{Z}(t)^T + \boldsymbol{Q}(t) \circ \boldsymbol{X}\boldsymbol{X}^T + \boldsymbol{\Phi}(t)\boldsymbol{\Phi}(t)^T\right](\hat{\boldsymbol{y}} - \boldsymbol{y}),\\ &\triangleq -\boldsymbol{H}(t)(\hat{\boldsymbol{y}} - \boldsymbol{y}),\end{aligned} \tag{18}$$

where $\circ$ is *Hadamard product* (or elementwise product), and matrices $\boldsymbol{M}(t) \in \mathbb{R}^{n\times n}$ and $\boldsymbol{Q}(t) \in \mathbb{R}^{n\times n}$ are defined as follows

$$\boldsymbol{M}(t)_{ij} \triangleq \frac{1}{m}\boldsymbol{u}^T(I_m - \tilde{\gamma}\boldsymbol{D}_i\boldsymbol{A})^{-1}\boldsymbol{D}_i\boldsymbol{D}_j^T(I_m - \tilde{\gamma}\boldsymbol{D}_j\boldsymbol{A})^{-T}\boldsymbol{u}, \tag{19}$$

$$\boldsymbol{Q}(t)_{ij} \triangleq \frac{1}{m}\left(\boldsymbol{D}_i^T(I_m - \tilde{\gamma}\boldsymbol{D}_i\boldsymbol{A})^{-1}\boldsymbol{u} + \boldsymbol{v}\right)^T \boldsymbol{E}_i\boldsymbol{E}_j^T\left(\boldsymbol{D}_j^T(I_m - \tilde{\gamma}\boldsymbol{D}_j\boldsymbol{A})^{-T}\boldsymbol{u} + \boldsymbol{v}\right). \tag{20}$$

Proof. The proof is provided in Appendix A.3. □

Note that the matrix $\boldsymbol{H}(t)$ is at least positive semidefinite since it is the sum of three positive semidefinite matrices. If there exists a constant $\lambda > 0$ such that $\lambda_{\min}\{\boldsymbol{H}(t)\} \geq \lambda > 0$ for all t, then the dynamics of the loss function $L(t)$ can be shown to satisfies the following inequality

$$L(t) \leq \exp\{-\lambda t\}L(0).$$

It immediately indicates that the objective value $L(t)$ is consistently decreasing to zero at geometric rate. Along with random initialization, we will show it is indeed the case (with exponentially highly probability) as long as the number of parameters m is large enough and no two data points are parallel to each other.

3.3 Lower bound of the smallest eigenvalue of $\boldsymbol{H}(t)$ and global convergence

By using the nonlinearity of the feature map ϕ, we will show the smallest eigenvalue of $\boldsymbol{\Phi}(t)\boldsymbol{\Phi}(t)^T$ is strictly positive overall time t with high probability. As a result, the smallest eigenvalue of $\boldsymbol{H}(t)$ is also lower bounded. To simplify the notation, we define the matrix $\boldsymbol{G}(t) \in \mathbb{R}^{n\times n}$ as follows

$$\boldsymbol{G}(t) \triangleq \boldsymbol{\Phi}(t)\boldsymbol{\Phi}(t)^T. \tag{21}$$

The matrix $\boldsymbol{G}(t)$ is a time-dependent matrix. We first analyze its property at its initialization. When $t = 0$, it follows from the definition of feature vector $\boldsymbol{\phi}$ in equation 4 that

$$\boldsymbol{G}(0) = \boldsymbol{\Phi}(0)\boldsymbol{\Phi}(0)^T = \frac{1}{m}\sigma(\boldsymbol{X}\boldsymbol{W}(0)^T)\sigma(\boldsymbol{X}\boldsymbol{W}(0)^T)^T = \frac{1}{m}\sum_{r=1}^{m}\sigma(\boldsymbol{X}\boldsymbol{w}_r(0))\sigma(\boldsymbol{X}\boldsymbol{w}_r(0))^T.$$

By Assumption 1, each vector $\boldsymbol{w}_r(0)$ follows the standard multivariate normal distribution, *i.e.*, $\boldsymbol{w}_r(0) \sim N(\boldsymbol{0}, \boldsymbol{I}_d)$. Let $m \to \infty$, then we obtain the covariance matrix $\boldsymbol{G}^\infty \in \mathbb{R}^{n\times n}$ give by

$$\boldsymbol{G}^\infty \triangleq \mathbb{E}_{\boldsymbol{w}\sim N(\boldsymbol{0},\boldsymbol{I}_d)}[\sigma(\boldsymbol{X}\boldsymbol{w})\sigma(\boldsymbol{X}\boldsymbol{w})^T]. \tag{22}$$

The matrix $\boldsymbol{G}^\infty$ is a Gram matrix induced by the ReLU activation function and the random initialization. The following lemma will show the smallest eigenvalue of $\boldsymbol{G}^\infty$ is strictly positive as long as no two data points are parallel. Moreover, later in Lemma 3.6 and 3.7, we will show that during the training, the property of $\boldsymbol{G}^\infty$ is preserved in the Gram matrices $\boldsymbol{G}(0)$ and $\boldsymbol{G}(t)$ as long as the number of parameter m is large enough.

Lemma 3.5. Assume $\|\boldsymbol{x}_i\| = 1$ for all $i \in [n]$. If $\boldsymbol{x}_i \neq \boldsymbol{x}_j$ for all $i \neq j$, then $\lambda_0 \triangleq \lambda_{\min}\{\boldsymbol{G}^\infty\} > 0$.

Proof. The proof is based on the *Hermite Expansions* of the matrix $\boldsymbol{G}^\infty$ and the entire proof is given in Appendix A.4. □

Assumption 2 (Training data)**.** Without loss generality, we can assume each $\boldsymbol{x}_i$ is normalized to have unit norm, *i.e.*, $\|\boldsymbol{x}_i\| = 1$, for all $i \in [n]$. Moreover, we assume $\boldsymbol{x}_i \nparallel \boldsymbol{x}_j$ for all $i \neq j$.

In most of realistic datasets, it is extremely rare two data are parallel. If so, one can add a random noise to each data. Then Assumption 2 is satisfied.

The next step is to show that at the initialization, the property of $\boldsymbol{G}^\infty$ is preserved in $\boldsymbol{G}(0)$ if m is large enough. In particular, the following lemma shows that if $m = \tilde{\Omega}(n^2)$, then $\boldsymbol{G}(0)$ has a lower bounded smallest eigenvalue with (exponentially) high probability. The proof is by using the standard concentration bound for Gaussian random variables so we defer the proof to Appendix A.5.

Lemma 3.6. Let $\lambda_0 = \lambda_{\min}(\boldsymbol{G}^\infty) > 0$. If $m \geq \frac{32n^2}{\lambda_0} \log\left(\frac{2n^2}{\delta}\right)$, then with probability at least $1-\delta$, we have $\|\boldsymbol{G}(0) - \boldsymbol{G}^\infty\|_2 \leq \frac{\lambda_0}{4}$, and $\lambda_{\min}(\boldsymbol{G}(0)) \geq \frac{3}{4}\lambda_0$.

During the training, the Gram matrix $\boldsymbol{G}(t)$ may change but it is still close to $\boldsymbol{G}(0)$ so as to preserve the property of $\boldsymbol{G}^\infty$ as long as the matrices $\boldsymbol{W}(t)$ has bounded operator norm and it is not far away from $\boldsymbol{W}(0)$. The result is critical but proof is not complicated as we only use the Lipschitz continuity of σ and the matrix norm inequality. The proof is included in Appendix A.6.

Lemma 3.7. Suppose $\|\boldsymbol{W}(0)\| \leq c\sqrt{m}$, and $\lambda_{\min}\{\boldsymbol{G}(0)\} \geq \frac{3}{4}\lambda_0$. For any matrix $\boldsymbol{W} \in \mathbb{R}^{m\times d}$ that satisfies $\|\boldsymbol{W}\| \leq 2c\sqrt{m}$ and $\|\boldsymbol{W} - \boldsymbol{W}(0)\| \leq \frac{\sqrt{m}\lambda_0}{16c\|\boldsymbol{X}\|^2} \triangleq R$, the matrix defined by

$$\boldsymbol{G} \triangleq \frac{1}{m}\sigma(\boldsymbol{X}\boldsymbol{W}^T)\sigma(\boldsymbol{X}\boldsymbol{W}^T)^T$$

satisfies $\|\boldsymbol{G} - \boldsymbol{G}(0)\| \leq \frac{\lambda_0}{4}$ and $\lambda_{\min}(\boldsymbol{G}) \geq \frac{\lambda_0}{2}$.

The next lemma shows three facts. First, the smallest eigenvalue of $\boldsymbol{G}(t)$ is lower bounded. Second, the objective value $L(t)$ converges to zero at a linear convergence rate. Third, the (operator) norms of all trainable parameters are upper bounded by some constants, which further implies the unique equilibrium point $\boldsymbol{z}^*(t)$ always exists. We prove the lemma by induction which is provided in Appendix A.7.

Lemma 3.8. Suppose $\|\boldsymbol{W}(0)\| \leq c\sqrt{m}$, $\|\boldsymbol{A}(0)\| \leq c\sqrt{m}$ for some absolute constant $c > 0$, and $\lambda_{\min}\{\boldsymbol{G}(0)\} \geq \frac{3}{4}\lambda_0 > 0$. If $m = \Omega\left(\frac{c^2 n\|\boldsymbol{X}\|^2}{\lambda_0^3}\|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\|^2\right)$ and choose $0 < \gamma \leq \min\{1/2, 1/4c\}$, then for any $t \geq 0$, we have

(i) $\lambda_{\min}(\boldsymbol{G}(t)) \geq \frac{\lambda_0}{2}$,

(ii) $\|\boldsymbol{u}(t)\| \leq \frac{16c\sqrt{n}}{\lambda_0}\|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\|$,

(iii) $\|\boldsymbol{v}(t)\| \leq \frac{8c\sqrt{n}}{\lambda_0}\|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\|$,

(iv) $\|\boldsymbol{W}(t)\| \leq 2c\sqrt{m}$,

(v) $\|\boldsymbol{A}(t)\| \leq 2c\sqrt{m}$,

(vi) $\|\hat{\boldsymbol{y}}(t) - \boldsymbol{y}\|^2 \leq \exp\{-\lambda_0 t\}\|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\|^2$.

By using simple union bounds, Lemma 3.1 and 3.8 immediately implies our main result.

Theorem 3.2 (Convergence rate of gradient flow)**.** Suppose Assumption 1 and 2 holds. Then if we set the number of parameter $m = \Omega\left(\frac{n^2}{\lambda_0}\log\left(\frac{n}{\delta}\right)\right)$, then with probability at least $1-\delta$ over the initialization, we have for any t

$$\|\hat{\boldsymbol{y}}(t) - \boldsymbol{y}\|^2 \leq \exp\{-\lambda_0 t\}\|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\|^2.$$

This theorem establish the global convergence of the gradient flow. Despite the nonconvexity of the objective function L, Theorem 3.2 shows that if m is large enough, then the objective value L is consistently decreasing to zero at the geometric rate. In particular, Theorem 3.2 requires $m = \tilde{\Omega}(n^2)$, which is similar or even better than the most recent results for the neural network with finite layers Nguyen & Mondelli (2020); Oymak & Soltanolkotabi (2020); Allen-Zhu et al. (2019); Zou & Gu (2019); Du et al. (2019). In particular, the previous result shows m is a polynomial of the number of training data n and the number of layers L. That does not work in our case since we have infinitely many layers. By taking advantage of the nonlinear feature maps, we are able to establish the global convergence while m is independent of depth.

4 Discrete time analysis

In this section, we show randomly initialized gradient descent with a fixed stepsize converges to a global minimum at a linear rate. The following lemma shows the global convergence is obtained if the operator norms of $\boldsymbol{A}$ and $\boldsymbol{W}$ are bounded and $\lambda_{\min}(\boldsymbol{G}(0)) > 0$.

Lemma 4.1. Suppose $\|\boldsymbol{W}(0)\| \leq c\sqrt{m}$, $\|\boldsymbol{A}(0)\| \leq c\sqrt{m}$ for some absolute constant $c > 0$, and $\lambda_{\min}\{\boldsymbol{G}(0)\} \geq \frac{3}{4}\lambda_0 > 0$. If $m = \Omega\left(\frac{c^2 n \|\boldsymbol{X}\|^2}{\lambda_0^3}\|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\|^2\right)$, $0 < \gamma \leq \min\{1/2, 1/4c\}$, and stepsize $\alpha = \mathcal{O}\left(\lambda_0/n^2\right)$, then for any $k \geq 0$, we have

(i) $\lambda_{\min}(\boldsymbol{G}(k)) \geq \frac{\lambda_0}{2}$,

(ii) $\|\boldsymbol{u}(k)\| \leq \frac{32c\sqrt{n}}{\lambda_0}\|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\|$,

(iii) $\|\boldsymbol{v}(k)\| \leq \frac{16c\sqrt{n}}{\lambda_0}\|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\|$,

(iv) $\|\boldsymbol{W}(k)\| \leq 2c\sqrt{m}$,

(v) $\|\boldsymbol{A}(k)\| \leq 2c\sqrt{m}$,

(vi) $\|\hat{\boldsymbol{y}}(k) - \boldsymbol{y}\|^2 \leq (1 - \alpha\lambda_0/2)^k\|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\|^2$.

By using simple union bounds of Lemma 3.1 and 4.1, we obtain the global convergence result for the gradient descent.

Theorem 4.1 (Convergence rate of gradient descent)**.** Suppose Assumption 1 and 2 holds. Then if we set the number of parameter $m = \Omega\left(\frac{n^2}{\lambda_0}\log\left(\frac{n}{\delta}\right)\right)$, $0 < \gamma \leq \min\{1/2, 1/4c\}$, and stepsize $\alpha = \mathcal{O}\left(\lambda_0/n^2\right)$, then with probability at least $1 - \delta$ over the initialization, we have for any $k \geq 0$

$$\|\hat{\boldsymbol{y}}(k) - \boldsymbol{y}\|^2 \leq (1 - \alpha\lambda_0/2)^k\|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\|^2.$$

Observe that if $\lambda_{\min}\{\boldsymbol{G}(0)\} > 0$, the linear global convergence result for gradient descent is obtained as long as $m \gtrsim n$. However, one needs to have $m \gtrsim n^2$ to achieve $\lambda_{\min}\{\boldsymbol{G}(0)\} > 0$.

5 Experiments

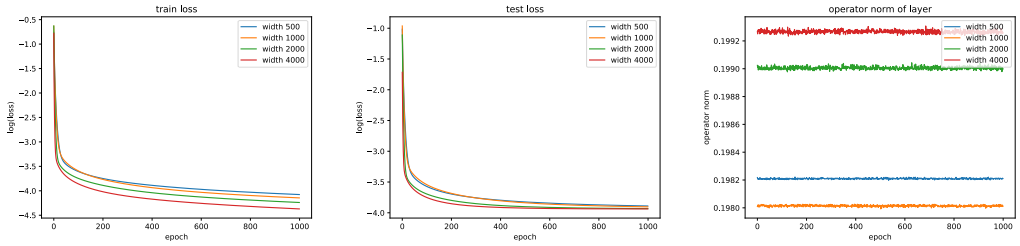
In this section, we use the real datasets MNST, FashionMNST, CIFAR10, CIFAR100, and SVHN to verify our theoretical findings. We initialize the entries of parameters $\boldsymbol{A}$, $\boldsymbol{W}$, $\boldsymbol{u}$, and $\boldsymbol{v}$ by standard Gaussian or symmetric Bernoulli distribution independently. Therefore, Assumption 1 are satisfied. For each real dataset, we only use classes 0 and 1. Moreover, 500 samples are randomly chosen from each class to generate the training dataset with $n = 1000$ samples. We normalize each data to have unit norm. Since no two duplicate data, Assumption 2 is also satisfied. We run 1000 epoch of gradient descent with fixed stepsize.

We test three different metrics with different widths m. First, we test how the amount of overparameterization affects the convergence rates. Second, we test the relation between the amount of overparameterization and the operator norms between matrix $\boldsymbol{A}(t)$ and its initialization. Note that the "operator norm" in the plots denotes the operator norm of the scaled matrix $\boldsymbol{A}(t)$ by the scalar $\gamma/\sqrt{m}$, *i.e.*, $(\gamma/\sqrt{m})\|\boldsymbol{A}(t)\|$. Third, we test the amount of overparameterization and the performance of the trained neural network on the unseen test data. Similar to the training dataset, the test dataset is generated by randomly selecting 500 samples from each class.

Figure 1d shows that as m becomes larger, we have better convergence rates. Figure 1f shows that the operator norms are slightly larger for larger m but overall the operator norms are approximately equal to its initialization. Figure 1e shows that as m becomes larger, the neural networks achieve lower test loss. The bell curve from the classical bias-variance trade-off does not appear. This open up a new research direction of the deep equilibrium model in the analyses of generalization error and bias-variance trade-off.

6 Conclusion and future works

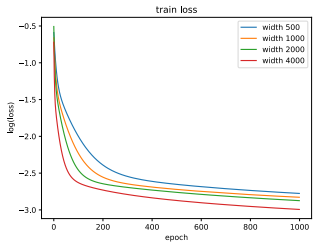
In this paper, we provide a convergence theory of the implicit models with ReLU activation in the overparameterization regime. In particular, we show that the gradient flow converges to global minimum of the loss function at a linear rate if the width m is large enough. The key idea is to using



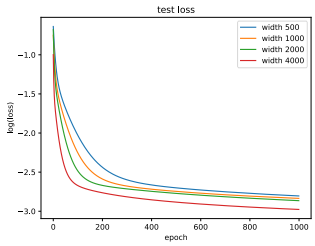
(a) Training loss

(b) Test loss

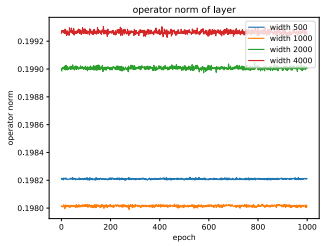
(c) Operator norms



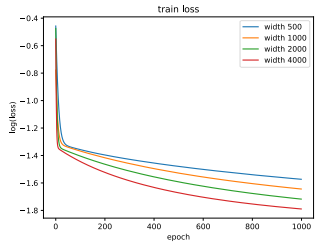
(d) Training loss



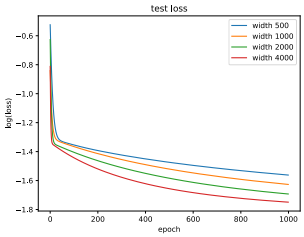
(e) Test loss



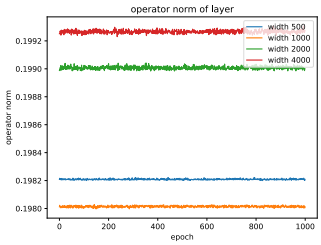
(f) Operator norms



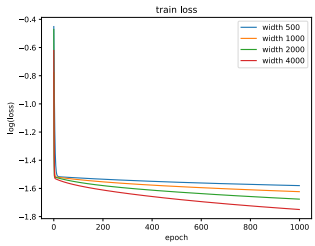
(g) Training loss



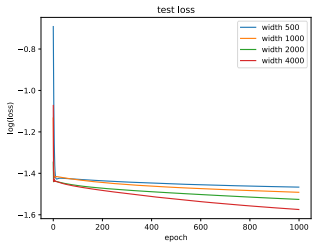
(h) Test loss



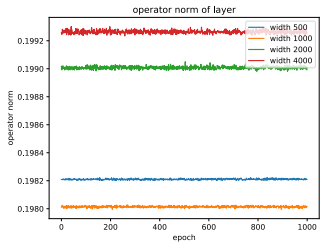
(i) Operator norms



(j) Training loss



(k) Test loss



(l) Operator norms

Figure 1: Results on MNIST, FashionMNIST, CIFAR10, CIFAR100, SVHN: we testify the impact of the width m on the training loss, test loss, and operator norm of the scaled matrix $(\gamma/\sqrt{m})\boldsymbol{A}(t)$ for each real dataset.

the scalar $\gamma \in (0, 1)$ to scale the random initialized weight matrix $\boldsymbol{A}$ so that the equilibrium equation is always well posed during the training. By analyzing the dynamics flow, we observe the dynamics of the prediction vector is controlled by a Gram matrix whose smallest eigenvalue is lower bounded by a strictly positive constant as long as the width m is about square of the size of training data n. We list a several potential future works based on the observations made in the experiments.

First, we believe our analysis can be generalized to implicit models with other scaling techniques and initialization. Here we use the scalar $\gamma/\sqrt{m}$ to ensure the existence of the equilibrium point $\boldsymbol{z}^*$ with random initialization. With an appropriate normalization technique, the global convergence for identity initialization should be obtained.

Second, we believe the width m is not only improving the convergence rates but also improving the generalization performance. In particular, Figure 1e shows that with larger m, the test loss is reduced while the classical bell curve of bias-variance trade-off is not observed.

REFERENCES

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pp. 242–252. PMLR, 2019.

Shaojie Bai, J Zico Kolter, and Vladlen Koltun. Trellis networks for sequence modeling. *arXiv preprint arXiv:1810.06682*, 2018.

Shaojie Bai, J Zico Kolter, and Vladlen Koltun. Deep equilibrium models. *arXiv preprint arXiv:1909.01377*, 2019.

Shaojie Bai, Vladlen Koltun, and J Zico Kolter. Multiscale deep equilibrium models. *arXiv preprint arXiv:2006.08656*, 2020.

Raj Dabre and Atsushi Fujita. Recurrent stacking of layers for compact neural machine translation models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 6292–6299, 2019.

Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Łukasz Kaiser. Universal transformers. *arXiv preprint arXiv:1807.03819*, 2018.

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*, pp. 1675–1685. PMLR, 2019.

Laurent El Ghaoui, Fangda Gu, Bertrand Travacca, Armin Askari, and Alicia Y Tsai. Implicit deep learning. *arXiv preprint arXiv:1908.06315*, 2, 2019.

Kenji Kawaguchi. On the theory of implicit deep learning: Global convergence with implicit layers. *arXiv preprint arXiv:2102.07346*, 2021.

Erwin Kreyszig. *Introductory functional analysis with applications*, volume 1. wiley New York, 1978.

Barbara MacCluer. *Elementary functional analysis*, volume 253. Springer Science & Business Media, 2008.

Quynh Nguyen and Marco Mondelli. Global convergence of deep networks with one wide layer followed by pyramidal topology. *arXiv preprint arXiv:2002.07867*, 2020.

Samet Oymak and Mahdi Soltanolkotabi. Toward moderate overparameterization: Global convergence guarantees for training shallow neural networks. *IEEE Journal on Selected Areas in Information Theory*, 1(1):84–105, 2020.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Number 47 in Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press. ISBN 978-1-108-41519-4.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

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