Global Convergence of Over-parameterized Deep Equilibrium Models

Anonymous Author
Anonymous Institution

Abstract

A deep equilibrium model (DEQ) is implicitly defined through an equilibrium point of an infinite-depth weight-tied model with an input-injection. Instead of infinite computations, it solves an equilibrium point directly with root-finding and computes gradients with implicit differentiation. In this paper, the training dynamics of over-parameterized DEQs are investigated, and we propose a novel probabilistic framework to overcome the challenge arising from the weight-sharing and the infinite depth. By supposing a condition on the initial equilibrium point, we prove that the gradient descent converges to a globally optimal solution at a linear convergence rate for the quadratic loss function. We further perform a fine-grained non-asymptotic analysis about random DEQs and the corresponding weight-untied models, and show that the required initial condition is satisfied via mild over-parameterization. Moreover, we show that the unique equilibrium point always exists during the training.

1 Introduction

Deep equilibrium models (DEQs) [1] have recently emerged as a new neural network design paradigm. A DEQ is equivalent to an infinite-depth weight-tied model with input-injection. Different from conventional (explicit) neural networks, DEQs generate features by directly solving equilibrium points of implicit equations. DEQs also have the remarkable advantage that the gradients can be computed analytically by backpropagation only through the equilibrium point with implicit differentiation. Therefore, training a DEQ only requires constant memory.

DEQs have achieved impressive performance in various applications such as computer vision [2, 3], natural language processing [1], and inverse problems [4]. Although the empirical success of DEQs has been observed in many recent studies, theoretical understandings of DEQs are still limited compared to conventional models. In this paper, we aim to establish the global convergence of the gradient descent (GD) associated with an over-parameterized DEQ, as a step towards understanding general DEQs.

A large body of work [5, 6, 7] has validated the effectiveness of over-parameterization in optimizing feedforward neural networks. The main idea is to investigate the property at initialization and bound the traveling distance of GD from the initialization [8]. However, it remains unclear whether these results can be directly applied to DEQs. The implicit weight-sharing is the key challenge. Most standard concentration tools used in previous studies fail in DEQs. This is because these analyses rely on the independence of initial random weights and features, which is no longer the case in DEQs. Moreover, it remains unknown whether over-parameterization is sufficient to guarantee the well-posedness [9, 10] of the implicit mapping of a DEQ, which is crucial to the stability of the training process, *e.g.* [11] uses an extra softmax layer to resolve the well-posedness issue and achieves a global linear convergent rate. However, this result holds only for linear DEQs and it is difficult to be extended to nonlinear activations.

We start with the gradient analysis. We observe that, in the case of DEQs, the least singular value of the equilibrium points plays a key role in the gradient dynamic. Specifically, if the least singular values of the equilibrium points at all iterations can be lower bounded by a positive constant, then one can establish a corresponding version of the Polyak-Lojasiewicz inequality for DEQs, and thus the global convergence of GD can be obtained.

Our main results are based on the following observations. Firstly, we prove the global convergence of GD by supposing an initial condition on the lower bound of the least singular value of the initial equilibrium points. The perturbation of the weight matrices is small enough to ensure that the Lipschitz constant of the implicit layer transformation is smaller than 1. This means that, the unique equilibrium point always exists throughout the training. Our second observation is that, the required initial condition holds with

Preliminary work. Under review by AISTATS 2023. Do not distribute.

a high probability for general Gaussian initialization. Note that the weight of a DEQ is implicitly re-used across layers. Thus, standard concentration inequalities, which are based on the independence of weight matrices and features, cannot be employed directly in this scenario of random DEQs analysis. In order to to overcome the technical difficulty, we propose a novel probabilistic framework to approximate the empirical Gram matrix of the equilibrium point with a population Gram matrix induced from a weight-untied random network with infinite depth.

1.1 Related Work

Finite-depth Over-parameterized Feedforward Networks. Recently, over-parameterization has attracted much attention due to its effectiveness in optimizing finite-depth neural networks. [12] shows that, for smooth activation and infinite wide neural networks, the trajectory of the gradient descent (GD) method could be well-captured by a kernel called the neural tangent kernel (NTK). For a finite-width feed-forward neural network with smooth activation, [5, 6, 7] prove that the neural networks' dynamics are strongly related to a Gram matrix. For a finite-width feed-forward neural network with ReLU activation, [13, 14, 15, 16, 17] estimate the changes of the activation patterns, and show that GD can converge to a global minimum despite the non-smoothness of activation and non-convexity of the objective function. The only condition that needs to satisfy is that the width of each layer is a polynomial of the number of training samples and the number of layers. In particular, [8] provides an alternative framework that only requires tracking the evolution of the last hidden layer rather than the activation pattern. In all previous works, a non-asymptotic analysis at initialization plays a fundamental role. Their random analysis relies on the independence between the initial random weights and features. However, the weight matrix is implicitly shared in a DEQ. Thus, previous results do not directly apply to DEQs.

Finite-depth Over-parameterized Weight-tied Neural Networks. For over-parameterized weight-tied models, [18, 19, 20] investigate NTKs of the recurrent neural networks (RNNs) with infinite width by leveraging the "Gaussian Conditioning Trick" [21, 22]. The similar technique is also used in the study on infinitely wide weight-tied autoencoders [23]. These works reveal that infinitely wide Gaussian weight-tied neural networks are essentially Gaussian processes. However, their results do not apply to the regime of finite width. [24, 25] show that, for a RNN with finite width, GD can converge to a global minimum if the width of each layer is a polynomial of the number of training samples and the number of layers. However, their results do not apply to DEQs. This is because most upper bounds for the norm of the hidden units fail as the depth approaches infinity. Thus, a new analysis framework for the convergence of the implicit models is urgently needed.

Over-parameterized DEQs. The investigation of over-parameterized DEQs is still in the initial stage. [26] considers the NTK of DEQs with infinite width. This study claims that the NTK of DEQs is equivalent to the corresponding weight-untied models in the regime of infinite width. This study also shows that DEQs have non-degenerate NTKs even in the infinite depth. This observation is similar with ours. However, the analysis on finite-width DEQs is not studied in [26]. To conduct non-asymptotic analysis on finite-width DEQs, one of the key challenges is to estimate the least singular value of the equilibrium points. To the best of our knowledge, this problem has not been addressed in general settings. Two very recent works [27] and [28] consider simple cases in which the problem can be transferred into the estimation of the least singular value of features obtained from a single layer explicit network. Their results hold for specific predictions or special initialization. In contrast, we propose a novel probabilistic framework to estimate the least singular value of the equilibrium points in general settings. Please see detailed comparisons in the discussion of Section 4.

1.2 Contributions

Our contributions are summarized as follows.

(1) We propose a novel probabilistic framework to analyze DEQs. Our framework addresses the technical challenges arising from the weight-sharing and the infinite depth. To the best of our knowledge, this is the first time a fine-grained non-asymptotic analysis on DEQs has been performed in a general setting.

(2) We analyze the gradient dynamics of DEQs with the quadratic loss function. Under an initial condition on the least singular value of the initial equilibrium points, we prove that the gradient descent converges to a global optimum at a linear rate. Based on our initial analysis, we show that the required initial condition is satisfied via mild over-parameterization.

(3) We show that the unique equilibrium point of an over-parameterized DEQ always exists throughout the training process, even without using any normalization or re-parameterization method.

2 Preliminaries

Notations. We use $\mathcal{N}(0, \boldsymbol{I})$ to denote the standard Gaussian distribution. We let $[n] \triangleq [1, \cdots, n]$. For a vector $\boldsymbol{v}$, $\|\boldsymbol{v}\|_2$ is the Euclidean norm of $\boldsymbol{v}$. For a matrix $\boldsymbol{A}$, we use $\boldsymbol{A}_{ij}$ denote its (i, j)-th entry. We use $\|\boldsymbol{A}\|_F$ to denote the Frobenius norm and $\|\boldsymbol{A}\|_2$ to denote the operator norm. If a matrix is positive semi-definite, we use $\lambda_{\min}(\boldsymbol{A})$

and $\sigma_{\min}(\boldsymbol{A})$ to denote its least eigenvalue and singular value, respectively. We let $\mathcal{O}(\cdot)$, $\Theta(\cdot)$ and $\Omega(\cdot)$ denote standard Big-O, Big-Theta, and Big-Omega notations, respectively. We use $\phi(\cdot)$ to denote the ReLU function, namely $\phi(x) = \max(x, 0)$.

2.1 Problem Setup

We define a vanilla deep equilibrium model (DEQ) with the transform at the l-th layer as

$$\boldsymbol{Z}^{(l)} = \phi(\boldsymbol{W}\boldsymbol{Z}^{(l-1)} + \boldsymbol{U}\boldsymbol{X}), \tag{1}$$

where $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n] \in \mathbb{R}^{d\times n}$ denotes the training inputs, $\boldsymbol{U} \in \mathbb{R}^{m\times d}$ and $\boldsymbol{W} \in \mathbb{R}^{m\times m}$ are trainable weight matrices, and $\boldsymbol{Z}^{(l)} \in \mathbb{R}^{m\times n}$ is the output feature at the l-th hidden layer. The output of the last hidden layer is defined by $\boldsymbol{Z}^* \triangleq \lim_{l\to\infty} \boldsymbol{Z}^{(l)}$. Therefore, instead of running the infinitely deep layer-by-layer forward propagation, $\boldsymbol{Z}^*$ can be calculated by directly solving the equilibrium point of the following equation

$$\boldsymbol{Z}^* = \phi(\boldsymbol{W}\boldsymbol{Z}^* + \boldsymbol{U}\boldsymbol{X}). \tag{2}$$

Let $\boldsymbol{y} = [y_1, \cdots, y_n] \in \mathbb{R}^n$ denote the labels, and $\hat{\boldsymbol{y}}(\boldsymbol{\theta}) = \boldsymbol{a}^\top \boldsymbol{Z}^*$ be the prediction function with $\boldsymbol{a} \in \mathbb{R}^m$ being a trainable vector and $\boldsymbol{\theta} = \mathrm{vec}\,(\boldsymbol{W}, \boldsymbol{U}, \boldsymbol{a})$. The object of our interest is the empirical risk minimization problem with the quadratic loss function

$$\Phi(\boldsymbol{\theta}) = \frac{1}{2}\left\|\hat{\boldsymbol{y}}(\boldsymbol{\theta}) - \boldsymbol{y}\right\|_2^2.$$

To do so, we consider the gradient descent (GD) update $\boldsymbol{\theta}_{\tau+1} = \boldsymbol{\theta}_\tau - \eta\nabla\Phi\left(\boldsymbol{\theta}_\tau\right)$, where η is the learning rate and $\boldsymbol{\theta}_\tau = \mathrm{vec}\left(\boldsymbol{W}(\tau), \boldsymbol{U}(\tau), \boldsymbol{a}(\tau)\right)$ is the parameter we optimize over at step τ. For notational simplicity, we omit the superscribe and denote $\boldsymbol{Z}$ to be the equilibrium $\boldsymbol{Z}^*$ when it is clear from the context. Moreover, the Gram matrix of the equilibrium point is defined by $\boldsymbol{G}(\tau) \triangleq \boldsymbol{Z}(\tau)^\top \boldsymbol{Z}(\tau)$ and we denote its least eigenvalue as $\lambda_\tau = \lambda_{\min}\left(\boldsymbol{G}(\tau)\right)$.

In this paper, we make the following assumptions on the random initialization and the input data.

Assumption 1 (Random initialization)**.** *Take $\sigma_w^2 < 1/8$. We assume that $\boldsymbol{W}$ is initialized with an $m \times m$ matrix with* i.i.d. *entries $\boldsymbol{W}_{ij} \sim \mathcal{N}(0, 2\sigma_w^2/m)$, $\boldsymbol{U}$ is initialized with an $m \times d$ matrix with* i.i.d. *entries $\boldsymbol{U}_{ij} \sim \mathcal{N}(0, 2/d)$, $\boldsymbol{a}$ is initialized with a random vector with* i.i.d. *entries $\boldsymbol{a} \sim \mathcal{N}(0, 1/m)$.*

Assumption 2 (Input data)**.** *We assume that (i) $\|\boldsymbol{x}_i\|_2 = \sqrt{d}$, for all $i \in [n]$, and $\boldsymbol{x}_i \nparallel \boldsymbol{x}_j$, for each pair $i \neq j \in [n]$, (ii) the labels satisfy $|y_i| = \mathcal{O}(1)$ for all $i \in [n]$.*

2.2 Well-Posedness and Gradients

Well-Posedness. For the stability of the training of the DEQs, it is crucial to guarantee the existence and uniqueness of the equilibrium points [9, 29]. It is equivalent to guarantee the well-posedness of the transformation defined in Eq. (1). In order to ensure the well-posedness, it suffices to take $\|\boldsymbol{W}(\tau)\|_2 < 1$, for all $\tau \geq 0$, with which Eq. (1) becomes a *contractive mapping*. From Lemma 1, we know that $\|\boldsymbol{W}(0)\|_2 < 1$ holds with a high probability under Assumption 1. Lemma 1 is a consequence of standard bounds concerning the singular values of Gaussian random matrices [30].

Lemma 1. *Let $\boldsymbol{W}$ be an $m \times m$ random matrix with* i.i.d. *entries $\boldsymbol{W}_{ij} \sim \mathcal{N}(0, \frac{2\sigma_w^2}{m})$. With probability at least $1 - \exp\{-\Omega(m)\}$, it holds that $\|\boldsymbol{W}\|_2 \leq 2\sqrt{2}\sigma_w$.*

In Section 3, we show that the condition of $\|\boldsymbol{W}(\tau)\|_2 < 1$ always holds for $\tau \geq 0$. It is worth mentioning that the constraint on the spectral norm of $\boldsymbol{W}(\tau)$ can be lightened by that on the spectral radius of $\boldsymbol{W}(\tau)$ through special re-parameterization methods [9, 10]. However, in this paper, we do not make extra assumptions on specific structures of weight matrices, and thus our constraint on the spectral norm of $\boldsymbol{W}(\tau)$ is in general mild.

Gradients. The gradients of conventional neural networks are usually computed via backpropagation through all the intermediate layers. On the contrary, the gradients *w.r.t.* parameters of a DEQ are computed analytically via backpropagation only through the equilibrium point $\boldsymbol{Z}$ by applying the *implicit function theorem*. Specifically, note that the equilibrium point of Eq. (2) is the root of the function $F(\tau) \triangleq \boldsymbol{Z}(\tau) - \phi(\boldsymbol{W}(\tau)\boldsymbol{Z}(\tau) + \boldsymbol{U}(\tau)\boldsymbol{X}(\tau))$. Let $\boldsymbol{J}(\tau) \triangleq \partial\mathrm{vec}(F(\tau))/\partial\mathrm{vec}(\boldsymbol{Z}(\tau))$ denote the Jacobian matrix. One can derive that

$$\boldsymbol{J}(\tau) = \boldsymbol{I}_{mn} - \boldsymbol{D}(\tau)\left(\boldsymbol{I}_n \otimes \boldsymbol{W}(\tau)\right),$$

where $\boldsymbol{D}(\tau) \triangleq \mathrm{diag}[\mathrm{vec}(\phi'(\boldsymbol{W}(\tau)\boldsymbol{Z}(\tau) + \boldsymbol{U}(\tau)\boldsymbol{X}(\tau)))]$. Using the Lipschitz property of ReLU, it is easy to check that $\boldsymbol{J}(\tau)$ is invertible if $\|\boldsymbol{W}(\tau)\|_2 < 1$. The gradient of each trainable parameter is given by the following lemma[1].

Lemma 2. *If $\boldsymbol{J}(\tau)$ is invertible, the gradient of the objective function $\Phi(\tau)$* w.r.t. *each trainable parameters is given by*

$$\begin{cases} \mathrm{vec}\left(\nabla_{\boldsymbol{W}}\Phi(\tau)\right) = \left(\boldsymbol{Z}(\tau) \otimes \boldsymbol{I}_m\right)\boldsymbol{R}(\tau)^\top\left(\hat{\boldsymbol{y}}(\tau) - \boldsymbol{y}\right) \\ \mathrm{vec}\left(\nabla_{\boldsymbol{U}}\Phi(\tau)\right) = \left(\boldsymbol{X}(\tau) \otimes \boldsymbol{I}_m\right)\boldsymbol{R}(\tau)^\top\left(\hat{\boldsymbol{y}}(\tau) - \boldsymbol{y}\right) \\ \nabla_{\boldsymbol{a}}\Phi(\tau) = \boldsymbol{Z}(\tau)\left(\hat{\boldsymbol{y}}(\tau) - \boldsymbol{y}\right) \end{cases}, \tag{3}$$

where $\boldsymbol{R}(\tau) = \left(\boldsymbol{a}(\tau) \otimes \boldsymbol{I}_n\right)\boldsymbol{J}(\tau)^{-1}\boldsymbol{D}(\tau)$.

2.3 Polyak-Lojasiewicz Inequalities

Polyak-Lojasiewicz (PL) inequality [31] is a commonly used recipe to prove linear convergence of GD algorithms [17, 8]. In order to obtain a corresponding version

[1] To simplify the notation, we omit the parameter $\boldsymbol{\theta}$ and write just $\Phi(\tau)$ and $\hat{\boldsymbol{y}}(\tau)$.

of PL inequalities for DEQs, our starting observation is that

$$\|\nabla_{\boldsymbol{\theta}}\Phi(\tau)\|_2^2 \geq 2\lambda_{\min}\left(\boldsymbol{H}(\tau)\right)\Phi(\tau), \tag{4}$$

where $\boldsymbol{H}(\tau)=\boldsymbol{H}_1(\tau)+\boldsymbol{H}_2(\tau)+\boldsymbol{H}_3(\tau)$ is a sum of three positive semi-definite matrices defined as

$$\begin{cases} \boldsymbol{H}_1(\tau)=\boldsymbol{G}(\tau) \\ \boldsymbol{H}_2(\tau)=\boldsymbol{R}(\tau)\left(\boldsymbol{G}(\tau)\otimes\boldsymbol{I}_m\right)\boldsymbol{R}(\tau)^\top \\ \boldsymbol{H}_3(\tau)=\boldsymbol{R}(\tau)\left(\boldsymbol{X}^\top\boldsymbol{X}\otimes\boldsymbol{I}_m\right)\boldsymbol{R}(\tau)^\top \end{cases},$$

Eq. (4) is a direct application of Lemma 2. It suggests that if $\lambda_{\min}\left(\boldsymbol{H}(\tau)\right)$ can be lower bounded away from zero, both at initialization and throughout the training, then one can establish a PL inequality that holds for the loss function, and thus GD converges to a global minimum. However, it is technically difficult to directly estimate the lower bound of $\lambda_{\min}\left(\boldsymbol{H}(\tau)\right)$ because (1) for $\tau=0$, $\boldsymbol{H}(0)$ involves both the sum and multiplication of random matrices with complex structures, and (2) for $\tau>0$, $\boldsymbol{H}(\tau)$ involves the derivatives of ReLU at activation neurons which requires the estimation related to the changes of the activation patterns [32, 33].

To make the problem tractable, we further observe that $\lambda_{\min}\left(\boldsymbol{H}(\tau)\right)\geq\lambda_\tau$, *i.e.* the least eigenvalue of the Gram matrix of the equilibrium point. Applying this observation to Eq. (4), one obtains

$$\|\nabla_{\boldsymbol{\theta}}\Phi(\tau)\|_2^2 \geq 2\lambda_\tau\Phi(\tau). \tag{5}$$

This means that, in order to obtain a PL-like inequality for DEQs, it suffices to bound the changes of $\boldsymbol{G}(\tau)$ throughout the training if λ_0 is bounded away from zero at initialization. In Section 3, we show that it holds $\lambda_\tau\geq\frac{1}{2}\lambda_0$ for every $\tau>0$, and we have a PL inequality for DEQs: $\|\nabla_{\boldsymbol{\theta}}\Phi(\tau)\|_2^2\geq\lambda_0\Phi(\tau)$. Based on this, we prove that GD converges to a global optimum at a linear rate. The result of Section 3 is built upon an initial condition on λ_0, and we further demonstrate that such an initial condition can be satisfied with mild over-parameterization.

2.4 Challenges in initial analysis

The initial condition on the lower bound of λ_0 plays a fundamental role in our convergence result. It is hard to estimate λ_0 directly. A common way is to build a concentration inequality between the initial *empirical* Gram matrix $\boldsymbol{G}$ and the corresponding *population* Gram matrix with easily estimated least eigenvalue. In the case of DEQs, we consider a population Gram matrix $\boldsymbol{K}$ defined as follows.

Definition 1. *We define the population Gram matrices* $\boldsymbol{K}^{(l)}$ *of each layer recursively as*

$$\begin{aligned} &\boldsymbol{K}^{(0)}=0 \\ &\boldsymbol{\Lambda}_{ij}^{(l)}=\begin{bmatrix} \sigma_w^2\boldsymbol{K}_{ii}^{(l-1)}+1 & \sigma_w^2\boldsymbol{K}_{ij}^{(l-1)}+\frac{1}{d}\boldsymbol{x}_i^\top\boldsymbol{x}_j \\ \sigma_w^2\boldsymbol{K}_{ji}^{(l-1)}+\frac{1}{d}\boldsymbol{x}_j^\top\boldsymbol{x}_i & \sigma_w^2\boldsymbol{K}_{jj}^{(l-1)}+1 \end{bmatrix}, \\ &\boldsymbol{K}_{ij}^{(l)}=2\mathbb{E}_{(\mathrm{u},\mathrm{v})^\top\sim\mathcal{N}(0,\boldsymbol{\Lambda}_{ij}^{(l)})}\left[\phi(\mathrm{u})\phi(\mathrm{v})\right] \end{aligned} \tag{6}$$

for $l\geq1$ *and* $(i,j)\in[n]\times[n]$. *Letting* $l\to\infty$, *we define* $\boldsymbol{K}\triangleq\boldsymbol{K}^{(\infty)}$ *and* $\lambda_*\triangleq\lambda_{\min}(\boldsymbol{K})$.

The population Gram matrix $\boldsymbol{K}$ is induced from an *infinite-depth weight-untied* model. The convergence of $\boldsymbol{K}^{(l)}$ for $l\to\infty$ and the positive definiteness of $\boldsymbol{K}$ are deferred to Section 4.1. The least eigenvalue of $\boldsymbol{K}$ is the fundamental quantity that determines the lower bound of λ_0.

Non-asymptotic analysis on DEQs is more difficult than explicit models. The weight-sharing is the key technical challenge, and one cannot resort to standard concentration tools. Specifically, note that initial $\boldsymbol{G}_{ij}$ is implicitly defined as $\boldsymbol{z}_i^\top\boldsymbol{z}_j=\phi([\boldsymbol{W},\boldsymbol{U}][\boldsymbol{z}_i^\top,\boldsymbol{x}_i^\top]^\top)\phi([\boldsymbol{W},\boldsymbol{U}][\boldsymbol{z}_j^\top,\boldsymbol{x}_j^\top])$[2]. On one hand, one cannot directly apply the standard inequality like previous works. This is because they rely on the independence between initial weight matrices and features, which is no longer the case in DEQs. On the other hand, one cannot directly use the standard ε-net argument [30]. Note that $[\boldsymbol{W},\boldsymbol{U}]$ is a *short* matrix, *i.e.* $[\boldsymbol{W},\boldsymbol{U}]\in\mathbb{R}^{m\times(m+d)}$, and the size of ε-net for $[\boldsymbol{z}_i^\top,\boldsymbol{x}_i^\top]$ is too large for us to derive a mild over-parameterization condition. In order to overcome the technical challenge, we propose a novel probabilistic framework in Section 4 by introducing the "new fresh randomness" [24] and perform a fine-grained non-asymptotic analysis on random DEQs.

3 Main Result

3.1 Global Convergence under an Initial Condition

Let δ be any positive constant such that $\rho_w(0)+\delta<1$. We define the following quantities:

$$\bar{\rho}_w=\|\boldsymbol{W}(0)\|_2+\delta,\ \bar{\rho}_u=\|\boldsymbol{U}(0)\|_2+\delta,\ \bar{\rho}_a=\|\boldsymbol{a}(0)\|_2+\delta.$$

We first present the global convergence of GD by supposing the following condition on the least eigenvalue λ_0 of the initial Gram matrix $\boldsymbol{G}(0)$.

Condition 1. *At initialization, it holds that*

$$\lambda_0\geq\frac{4}{\delta}\max\left(c_w,c_u,c_a\right)\|\boldsymbol{X}\|_F\|\hat{\boldsymbol{y}}(0)-\boldsymbol{y}\|_2, \tag{7}$$

$$\lambda_0^{\frac{3}{2}}\geq4(2+\sqrt{2})\bar{\rho}_a^{-1}\left(c_w^2+c_u^2\right)\|\boldsymbol{X}\|_F^2\|\hat{\boldsymbol{y}}(0)-\boldsymbol{y}\|_2, \tag{8}$$

$$\lambda_0\geq4\left(c_w^2+c_u^2\right)\|\boldsymbol{X}\|_F^2, \tag{9}$$

where $c_w=\frac{\bar{\rho}_u\bar{\rho}_a}{(1-\bar{\rho}_w)^2}$, $c_u=\frac{\bar{\rho}_a}{1-\bar{\rho}_w}$, $c_a=\frac{\bar{\rho}_u}{1-\bar{\rho}_w}$.

[2]For notational simplicity, we denote $\boldsymbol{W}(0)$, $\boldsymbol{U}(0)$ and $\boldsymbol{G}(0)$ by $\boldsymbol{W}$, $\boldsymbol{U}$ and $\boldsymbol{G}$, respectively.

The convergence result under Condition 1 is presented as follows.

Theorem 1. *Consider a DEQ defined in Eq. (2). Suppose that Condition 1 holds at initialization. If the learning rate satisfies*

$$\eta < \min\left(\frac{2}{\lambda_0}, \frac{2(c_w^2 + c_u^2)}{(c_w^2 + c_u^2 + c_a^2)^2 \|\boldsymbol{X}\|_F^2}\right), \tag{10}$$

for every $\tau \geq 0$, the following holds

(i) $\|\boldsymbol{W}(\tau)\|_2 < 1$, i.e. *the equilibrium points always exist,*

(ii) $\lambda_\tau > \frac{1}{2}\lambda_0$, *and thus the PL condition holds as* $\|\nabla_{\boldsymbol{\theta}}\Phi(\tau)\|_2^2 \geq \lambda_0\Phi(\tau)$,

(iii) the loss converges to a global minimum as

$$\Phi(\tau) \leq \left(1 - \eta\frac{\lambda_0}{2}\right)^\tau \Phi(0).$$

Theorem 1 shows that the unique equilibrium point always exists during the training, and GD converges to a global optimum under Condition 1. The proof of this part is inspired by the framework proposed in [8, Theorem 2.2]. The complete proof is deferred to the supplementary material. Next, we discuss how these initial conditions can be fulfilled via over-parameterization.

3.2 Initial Condition

In this section, we aim to show that, Condition 1 holds under Assumptions 1 and 2 via over-parameterization. To do so, it suffices to derive a lower bound on λ_0, and upper bounds on $\max(c_w, c_u, c_a)$, $\bar{\rho}_a^{-1}$ and the initial loss $\Phi(0)$, and put these bounds into Eq. (7)-Eq. (9) to obtain a specific condition on width m.

Firstly, we present the lower bound of λ_0 as follows.

Theorem 2. *If $m = \Omega\left(\frac{n^2}{\lambda_*^2}\left(\log\frac{n}{\lambda_* t}\right)\right)$, with probability at least $1 - t$, it holds that*

$$\lambda_0 \geq \frac{m}{2}\lambda_*.$$

The proof of Theorem 2 is based on a concentration inequality between the empirical Gram matrix $\boldsymbol{G}$ and the population Gram matrix $\boldsymbol{K}$. Since that the weight matrices are implicitly reused in DEQs for infinite times (see Eq. (1)), standard concentration tools are not applicable. In order to overcome the challenge arising from the weight-sharing and the infinite depth. We present a novel probabilistic framework. It is one of our core contributions. Detailed analysis is presented in Section 4.

Secondly, by standard bounds on the operator norm of Gaussian matrices [30], we have *w.p.* $\geq 1 - \exp\{-\Omega(m)\}$ that, $\|\boldsymbol{W}(0)\|_2 = \mathcal{O}(1)$, $\|\boldsymbol{U}(0)\|_2 = \mathcal{O}\left(\sqrt{m/d}\right)$ under Assumption 1. Thus, *w.p.* $\geq 1 - \exp\{-\Omega(m)\}$, it holds that

$$\bar{\rho}_w = \mathcal{O}(1), \quad \bar{\rho}_u = \mathcal{O}\left(\sqrt{\frac{m}{d}}\right)$$

which implies that

$$\max(c_w, c_u, c_a) = \mathcal{O}\left(\sqrt{\frac{m}{d}}\right).$$

By standard bounds on the norm of Gaussian vector [30], we have *w.p.* $\geq 1 - \exp\{-\Omega(m)\}$ that, $\bar{\rho}_a^{-1} = \mathcal{O}(1)$.

Thirdly, by using the property of the contractive mapping Eq. (1) and using the standard concentration argument, it is easy to show that *w.p.* $\geq 1 - t$, it holds that

$$\Phi(0) = \mathcal{O}(n).$$

Putting all these bounds into Eq. (7)-Eq. (9), one can show that *w.p.* $\geq 1 - t$, Condition 1 is satisfied for $m = \Omega\left(\frac{n^2}{\lambda_*^2}\left(\log\frac{n}{\lambda_* t}\right)\right)$.

4 Analysis at Initialization

For notational simplicity, we denote $\boldsymbol{W}(0)$ and $\boldsymbol{U}(0)$ by $\boldsymbol{W}$ and $\boldsymbol{U}$ in this section. The Gram matrices of the equilibrium point $\boldsymbol{Z}^{(l)}$ of the l-th layer are defined as $\boldsymbol{G}^{(l)} \triangleq \left(\boldsymbol{Z}^{(l)}\right)^\top \boldsymbol{Z}^{(l)}$, for $l \geq 1$. Without loss of generality, we assume that $\boldsymbol{Z}^{(0)} = 0$.

Our main idea is to establish the concentration inequality between the empirical Gram matrix $\boldsymbol{G}$ and the population Gram matrix $\boldsymbol{K}$. A simple case is to initialize DEQs as single-layer explicit models. Specifically, take $\sigma_w^2 = 0$ and $\boldsymbol{Z} = \phi(\boldsymbol{U}\boldsymbol{X})$. Using the standard Bernstein inequality, one can show that *w.p.*$\geq 1 - t$, $\lambda_0 \geq m\tilde{\lambda}/2$, as long as $m = \Omega\left(\tilde{\lambda}^{-2}n^2\log(n/t)\right)$ where $\tilde{\lambda} = \lambda_{\min}\left(\boldsymbol{K}^{(1)}\right)$. This case has been studied in many previous works [17, 16].

However, when $\sigma_w^2 > 0$, previous random analyses on explicit networks cannot be directly applied to DEQs. This is due to the fact that these analyses rely on the independence of initial random weights and features, which is no longer the case in DEQs. In order to overcome these technical difficulties, we propose a novel probabilistic framework based on the following observations: $\boldsymbol{K}_{ij}^{(l)}$, $\frac{1}{m}\boldsymbol{G}_{ij}^{(l)}$, and $\boldsymbol{G}_{ij}^{(l)}$ converge to $\boldsymbol{K}_{ij}$, $\boldsymbol{K}_{ij}^{(l)}$, and $\boldsymbol{G}_{ij}$ at an exponential rate, respectively. Moreover, $\boldsymbol{K}$ is strictly positive definite. These observations imply that it suffices to take sufficiently large l and m to ensure that $\|\frac{1}{m}\boldsymbol{G} - \boldsymbol{K}\|_F$ is smaller than λ_*. Thus, one can lower bound λ_0 by invoking Weyl's inequality. Detailed analysis is given as follows.

4.1 Bound between infinite-depth and finite-depth weight-untied models

In the case of $\phi = \mathrm{ReLU}$, given any positive definite matrix $\boldsymbol{A} = \begin{bmatrix} 1 & x \\ x & 1 \end{bmatrix}$ with $|x| \leq 1$, and two random variables $(\mathrm{u}, \mathrm{v})^\top \sim \mathcal{N}(0, \boldsymbol{A})$, as shown in [34], it holds that

$$\mathbb{E}\left[\phi(\mathrm{u})\phi(\mathrm{v})\right] = \frac{1}{2}Q(x), \tag{11}$$

where $Q(x) \triangleq \frac{\sqrt{1-x^2}+(\pi-\arccos x)x}{\pi}$.

Combining Eq. (11) with the homogeneity of ReLU, we can have more precise expressions of $\boldsymbol{K}_{ij}^{(l)}$ as follows.

Lemma 3. *Let* $\cos\theta_{ij}^{(l)} = \frac{\sigma_w^2 \boldsymbol{K}_{ij}^{(l-1)} + d^{-1}\boldsymbol{x}_i^\top \boldsymbol{x}_j}{\sigma_w^2 \boldsymbol{K}_{ij}^{(l-1)}+1}$, *and* $\rho^{(l)} = \boldsymbol{K}_{ii}^{(l)}$. *For* $l \geq 1$ *and* $(i,j) \in [n] \times [n]$, $\boldsymbol{K}_{ij}^{(l)}$ *defined in Definition 1 can be written as*

$$\boldsymbol{K}_{ij}^{(l)} = \rho^{(l)} Q\left(\cos\theta_{ij}^{(l)}\right), \tag{12}$$

with $\rho^{(l)} = \frac{1-\sigma_w^{2l}}{1-\sigma_w^2}$ *and*

$$\cos\theta_{ij}^{(l)} = \left(1 - \frac{1}{\rho^{(l-1)}}\right) Q\left(\cos\theta_{ij}^{(l-1)}\right) + \frac{1}{\rho^{(l-1)}} \frac{\boldsymbol{x}_i^\top \boldsymbol{x}_j}{d}. \tag{13}$$

Theorem 3. *Under Assumptions 1 and 2, it holds that*

(i) $\left\|\boldsymbol{K} - \boldsymbol{K}^{(l)}\right\|_F = \mathcal{O}\left(nl^{\frac{1}{2}}\sigma_w^l\right)$, *which implies that, for* $l \to \infty$, $\boldsymbol{K}^{(l)}$ *converges to* $\boldsymbol{K}$ *with each entry*

$$\boldsymbol{K}_{ij} = \frac{1}{1-\sigma_w^2} Q(\cos\theta_{ij}), \tag{14}$$

where $\cos\theta_{ij} = \sigma_w^2 Q(\cos\theta_{ij}) + (1-\sigma_w^2)\frac{\boldsymbol{x}_i \boldsymbol{x}_j}{d}$.

(ii) $\boldsymbol{K}$ *is strictly positive definite,* i.e. $\lambda_* > 0$.

Sketch of Proof. (i) By Eq. (6) and Eq. (12), one can show that $\left|\boldsymbol{K}_{ij}^{(l+1)} - \boldsymbol{K}_{ij}^{(l)}\right| \leq \sigma_w^2 \left|\boldsymbol{K}_{ij}^{(l)} - \boldsymbol{K}_{ij}^{(l-1)}\right| + 2\sigma_w^{2l}$, which implies that $\left|\boldsymbol{K}_{ij} - \boldsymbol{K}_{ij}^{(l)}\right| = \mathcal{O}\left(l\sigma_w^l\right)$.

Note that $\sigma_w < 1$. Thus $\boldsymbol{K}^{(l)}$ converges to a deterministic Gram matrix $\boldsymbol{K}$, and one can easily obtain Eq. (14) from Eq. (13) by letting $l \to \infty$.

(ii) The proof of this part is similar with [16, 17]. By performing the Hermite analysis on Eq. (14), one can show that $\boldsymbol{K}$ is strictly positive definite if $|\cos\theta_{ij}| < 1$ for all $i \neq j$. Using the fact that $|Q(x)| \leq 1$, Eq. (14) implies that $|\cos\theta_{ij}| < 1$ under Assumption 2. Please see the complete proof in the supplementary material. □

4.2 Bound between infinite-depth and finite-depth weight-tied models

By leveraging the contractility of the transformation defined in Eq. (1) and by invoking the standard Bernstein inequality, we establish a concentration inequality between the Gram matrix of the l-th layer's output and that of the initial equilibrium point as follows.

Theorem 4. *Under Assumptions 1 and 2, with probability at least* $1 - n^2\exp\{-\Omega(m)\}$, *it holds*

$$\frac{1}{m}\left\|\boldsymbol{G} - \boldsymbol{G}^{(l)}\right\|_F = \mathcal{O}\left(n\left(2\sqrt{2}\sigma_w\right)^l\right).$$

Sketch of Poof. Using the contractility of the implicit layer, it is easy to have *w.p.* $\geq 1 - \exp\{-\Omega(m)\}$, $\|\boldsymbol{z}_i^{(l)}\|_2 = \mathcal{O}\left(\|\boldsymbol{z}_i^{(1)}\|_2\right)$, and $\|\boldsymbol{z}_i - \boldsymbol{z}_i^{(l)}\|_2 = \mathcal{O}\left((2\sqrt{2}\sigma_w)^l \|\boldsymbol{z}_i^{(1)}\|_2\right)$. Moreover, using Bernstein inequality, we have *w.p.* $\geq 1 - \exp\left\{-\Omega\left(mt^2\right)\right\}$, $\left|\frac{1}{m}(\boldsymbol{z}_i^{(1)})^\top \boldsymbol{z}_i^{(1)} - 1\right| \leq t$. Let t be an absolute positive constant. Theorem 4 can be proved by applying the simple union bound. Please see the complete proof in the supplementary material. □

4.3 Bound between weight-tied and weight-untied models with finite-depth

Due to the implicit weight-tied structure of DEQs, standard concentration inequalities built upon the independence cannot be directly applied. In order to overcome technical difficulties, we build necessary probabilistic tools for DEQs. We first introduce Lemma 4 which provides the "fresh new randomness" for our analysis. The construction method in Lemma 4 is inspired by the previous work [24] on RNNs.

Lemma 4. *For* $l \geq 1$, $\boldsymbol{G}_{ij}^{(l+1)}$ *can be reconstructed as* $\boldsymbol{G}_{ij}^{(l+1)} = \phi(\boldsymbol{M}\boldsymbol{h})^\top \phi(\boldsymbol{M}\boldsymbol{h}')$ *such that*

(i) $\boldsymbol{h}^\top \boldsymbol{h}' = \frac{\sigma_w^2}{m}\boldsymbol{G}_{ij}^{(l)} + \frac{1}{d}\boldsymbol{x}_i^\top \boldsymbol{x}_j$,

(ii) $\boldsymbol{M} \in \mathbb{R}^{m\times(2l+d+2)}$ *is a* rectangle matrix, *and the entries of* $\boldsymbol{M}$ *are* i.i.d. *from* $\mathcal{N}(0,2)$ *conditioning on previous layers.*

Proof. (i) Let $\boldsymbol{V}_l \in \mathbb{R}^{m\times 2l}$ denote a column orthonormal matrix using Gram-Schmidt as

$$\boldsymbol{V}_l = \mathrm{GS}\left(\boldsymbol{z}_i^{(1)}, \cdots, \boldsymbol{z}_i^{(l)}, \boldsymbol{z}_j^{(1)}, \cdots, \boldsymbol{z}_j^{(l)}\right).$$

For each i, j, we define $\boldsymbol{p} \triangleq (\boldsymbol{I} - \boldsymbol{V}_{l-1}\boldsymbol{V}_{l-1}^\top)\boldsymbol{z}_i^{(l)}$ and $\boldsymbol{q} \triangleq (\boldsymbol{I} - \boldsymbol{V}_{l-1}\boldsymbol{V}_{l-1}^\top)\boldsymbol{z}_j^{(l)}$. We split $\boldsymbol{q}$ into two parts $\boldsymbol{q} = \boldsymbol{q}^\parallel + \boldsymbol{q}^\perp$, where $\boldsymbol{q}^\parallel$ is parallel to $\boldsymbol{p}$ and $\boldsymbol{q}^\perp$ is orthogonal to $\boldsymbol{p}$ as

$$\boldsymbol{q}^\parallel = \frac{\boldsymbol{p}^\top \boldsymbol{q}}{\|\boldsymbol{p}\|_2^2}\boldsymbol{p}, \quad \boldsymbol{q}^\perp = \left(\boldsymbol{I} - \frac{\boldsymbol{p}\boldsymbol{p}^\top}{\|\boldsymbol{p}\|_2^2}\right)\boldsymbol{q}.$$

First, we construct $\boldsymbol{M}$ as $\boldsymbol{M} = [\boldsymbol{M}_1, \boldsymbol{M}_2, \boldsymbol{M}_3, \boldsymbol{M}_4]$ with

$$\boldsymbol{M}_1 = \sigma_w^{-1}\sqrt{m}\boldsymbol{W}\boldsymbol{V}_{l-1}, \quad \boldsymbol{M}_2 = \sigma_w^{-1}\sqrt{m}\frac{\boldsymbol{W}\boldsymbol{p}}{\|\boldsymbol{p}\|_2},$$
$$\boldsymbol{M}_3 = \sigma_w^{-1}\sqrt{m}\frac{\boldsymbol{W}\boldsymbol{q}}{\|\boldsymbol{q}\|_2}, \qquad \boldsymbol{M}_4 = \sqrt{d}\boldsymbol{U}.$$

Then, we construct $\boldsymbol{h} = [\boldsymbol{h}_1^\top, \boldsymbol{h}_2^\top, \boldsymbol{h}_3^\top, \boldsymbol{h}_4^\top]^\top$ with

$$\boldsymbol{h}_1 = \tfrac{\sigma_w}{\sqrt{m}}\boldsymbol{V}_{l-1}^\top \boldsymbol{z}_i^{(l)}, \quad \boldsymbol{h}_2 = \tfrac{\sigma_w}{\sqrt{m}}\|\boldsymbol{p}\|_2,$$
$$\boldsymbol{h}_3 = 0, \qquad \boldsymbol{h}_4 = \tfrac{1}{\sqrt{d}}\boldsymbol{x}_i,$$

and $\boldsymbol{h}' = [\boldsymbol{h}_1'^\top, \boldsymbol{h}_2'^\top, \boldsymbol{h}_3'^\top, \boldsymbol{h}_4'^\top]^\top$ with

$$\boldsymbol{h}_1' = \tfrac{\sigma_w}{\sqrt{m}}\boldsymbol{V}_{l-1}^\top \boldsymbol{z}_j^{(l)}, \quad \boldsymbol{h}_2' = \tfrac{\sigma_w \boldsymbol{p}^\top \boldsymbol{q}}{\sqrt{m}\|\boldsymbol{p}\|_2},$$
$$\boldsymbol{h}_3' = \tfrac{\sigma_w}{\sqrt{m}}\|\boldsymbol{q}\|_2, \qquad \boldsymbol{h}_4' = \tfrac{1}{\sqrt{d}}\boldsymbol{x}_j.$$

It is easy to check that $\boldsymbol{G}_{ij}^{(l+1)} = \phi(\boldsymbol{M}\boldsymbol{h})^\top \phi(\boldsymbol{M}\boldsymbol{h}')$ with $\boldsymbol{h}^\top \boldsymbol{h}' = \frac{\sigma_w^2}{m}\boldsymbol{G}_{ij}^{(l)} + \frac{1}{d}\boldsymbol{x}_i^\top \boldsymbol{x}_j$.

(ii) Let $\boldsymbol{V}_{l-1} \triangleq \big[\boldsymbol{v}_1, \cdots, \boldsymbol{v}_{2(l-1)}\big]$. Note that $\boldsymbol{v}_j$ only depends on the randomness of $\boldsymbol{U}$ and $\boldsymbol{W}\,[\boldsymbol{v}_1, \cdots, \boldsymbol{v}_{j-1}]$. This means that, conditioning on $\boldsymbol{U}$ and $\boldsymbol{W}\,[\boldsymbol{v}_1, \cdots, \boldsymbol{v}_{j-1}]$, $\boldsymbol{W}\boldsymbol{v}_j$ is still an independent Gaussian vector. Similarly, we have $\boldsymbol{W}\boldsymbol{p}$ and $\boldsymbol{W}\boldsymbol{q}$ are also independent Gaussian vectors. Consequently, we prove that the entries of $\boldsymbol{M}$ are *i.i.d.* from $\mathcal{N}(0, 2)$. □

We stress that although $\boldsymbol{M}$ constructed in Lemma 4 has *i.i.d.* entries, it still depends on $\boldsymbol{h}$ and $\boldsymbol{h}'$. Thus, the standard Bernstein inequality cannot be directly applied. To address this issue, we perform the standard ε-argument [30] and leverage the "fresh new randomness" [24] provided in Lemma 4. The concentration inequality between $\frac{1}{m}\boldsymbol{G}^{(l)}$ and $\boldsymbol{K}^{(l)}$ is established as follows.

Theorem 5. *Under Assumptions 1 and 2, with probability at least* $1 - n^2 \exp\{-\Omega\left(m8^l\sigma_w^{2l}\right) + \mathcal{O}(l^2)\}$*, it holds that*

$$\left\|\frac{1}{m}\boldsymbol{G}^{(l)} - \boldsymbol{K}^{(l)}\right\|_F \leq n\left(2\sqrt{2}\sigma_w\right)^l.$$

Sketch of Proof. We give the main ideas of the proof.

It holds that $\big|\frac{1}{m}\boldsymbol{G}_{ij}^{(l)} - \boldsymbol{K}_{ij}^{(l)}\big| \leq \big|\frac{1}{m}\boldsymbol{G}_{ij}^{(l)} - \mathbb{E}\big[\frac{1}{m}\boldsymbol{G}_{ij}^{(l)}\big]\big| + \big|\mathbb{E}\big[\frac{1}{m}\boldsymbol{G}_{ij}^{(l)}\big] - \boldsymbol{K}_{ij}^{(l)}\big|$ by the triangle inequality. Following Lemma 4, we reconstruct $\boldsymbol{G}_{ij}^{(l)} = \phi(\boldsymbol{M}\boldsymbol{h})^\top \phi(\boldsymbol{M}\boldsymbol{h}')$.

(i) For *fixed* $\boldsymbol{h}$ and $\boldsymbol{h}'$, using the Bernstein inequality, and on can show that *w.p.* $\geq 1 - \exp\{-\Omega(m\varepsilon^2)\}$, it holds that

$$\left|\frac{1}{m}\boldsymbol{G}_{ij}^{(l)} - \mathbb{E}[\frac{1}{m}\boldsymbol{G}_{ij}^{(l)}]\right| \leq \varepsilon.$$

(ii) For *all* $\boldsymbol{h}$ and $\boldsymbol{h}'$, we apply the standard ε-net argument. Note that the size of ε-net for $\boldsymbol{h}, \boldsymbol{h}'$ is at most $\exp\{\mathcal{O}\big(l\log\frac{1}{\varepsilon}\big)\}$. Thus, one can derive that *w.p.* $\geq 1 - \exp\{-\Omega(m\varepsilon^2) + \mathcal{O}\big(l\log\frac{1}{\varepsilon}\big)\}$, it holds that

$$\left|\frac{1}{m}\boldsymbol{G}_{ij}^{(l)} - \mathbb{E}\big[\frac{1}{m}\boldsymbol{G}_{ij}^{(l)}\big]\right| \leq \varepsilon.$$

(iii) Substitute the choice of $\boldsymbol{h}$ and $\boldsymbol{h}'$ such that $\boldsymbol{h}^\top \boldsymbol{h}' = \frac{\sigma_w^2}{m}\boldsymbol{G}_{ij}^{(l-1)} + \frac{1}{d}\boldsymbol{x}_i^\top \boldsymbol{x}_j$. Using the fact that $\mathbb{E}\big[\frac{1}{m}\boldsymbol{G}_{ij}^{(l)}\big]$ is determined by $\boldsymbol{h}^\top \boldsymbol{h}'$, one can further derive that *w.p.* $\geq 1 - l\exp\{-\Omega\left(m\varepsilon^2\right) + \mathcal{O}\big(l\log\frac{1}{\varepsilon}\big)\}$,

$$\left|\mathbb{E}\big[\frac{1}{m}\boldsymbol{G}_{ij}^{(l)}\big] - \boldsymbol{K}_{ij}^{(l)}\right| \leq \sigma_w^2\left|\frac{1}{m}\boldsymbol{G}_{ij}^{(l-1)} - \boldsymbol{K}_{ij}^{(l-1)}\right| + \varepsilon,$$

which implies that

$$\left|\frac{1}{m}\boldsymbol{G}_{ij}^{(l)} - \boldsymbol{K}_{ij}^{(l)}\right| \leq \sigma_w^2\left|\frac{1}{m}\boldsymbol{G}_{ij}^{(l-1)} - \boldsymbol{K}_{ij}^{(l-1)}\right| + 2\varepsilon.$$

Note that $\sigma_w^2 < 1/8$, and thus a simple induction argument works here.

Lastly, let $\varepsilon = (2\sqrt{2}\sigma_w)^l$, and Theorem 5 can be proved by using the simple union bound. Please see the complete proof in the supplementary material. □

4.4 Proof of Theorem 2

Firstly, combining Theorems 3, 4 and 5, one can show that *w.p.* $\geq 1 - n^2\exp\{-\Omega\left(m8^l\sigma_w^{2l}\right) + \mathcal{O}\big(l^2\big)\}$, it holds

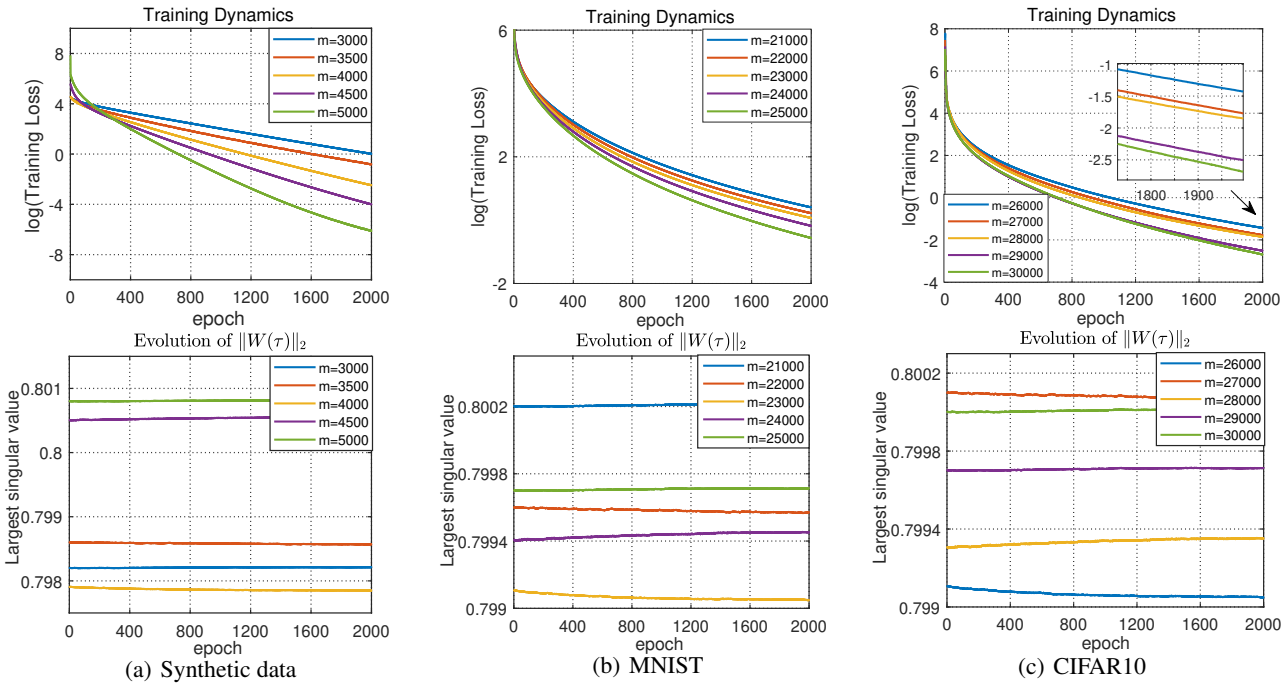
$$\begin{aligned}
&\left\|\frac{1}{m}\boldsymbol{G} - \boldsymbol{K}\right\|_F \\
\leq&\frac{1}{m}\left\|\boldsymbol{G} - \boldsymbol{G}^{(l)}\right\|_F + \left\|\frac{1}{m}\boldsymbol{G}^{(l)} - \boldsymbol{K}^{(l)}\right\|_F + \left\|\boldsymbol{K} - \boldsymbol{K}^{(l)}\right\|_F \\
=&\mathcal{O}\bigg(n\left(2\sqrt{2}\sigma_w\right)^l\bigg) + \mathcal{O}\bigg(n\left(2\sqrt{2}\sigma_w\right)^l\bigg) + \mathcal{O}\Big(nl^{\frac{1}{2}}\sigma_w^l\Big) \\
=&\mathcal{O}\bigg(n\left(2\sqrt{2}\sigma_w\right)^l\bigg),
\end{aligned}$$

where the last equality comes from the fact that $l^{\frac{1}{2}} \leq (2\sqrt{2})^l$, for $l \geq 1$.

Next, we fix l to omit the explicit dependence in l. Specifically, we take $l = \Theta\left(\log\big(\lambda_*^{-1}n\big)/\log\big(\sqrt{2}/4\sigma_w\big)\right)$. The lower bound of l is large enough to ensure that $\left\|\frac{1}{m}\boldsymbol{G} - \boldsymbol{K}\right\|_F \leq \frac{\lambda_*}{2}$. Therefore, by Weyl's inequality, it holds that $\lambda_0 > \frac{m}{2}\lambda_*$. Meanwhile, the upper bound of l guarantees that the probability does not decrease exponentially. Thus, one can use a mild over-parameterization condition on m to ensure the high probability.

Consequently, one can show that, it holds *w.p.* $\geq 1 - t$, $\lambda_0 \geq \frac{m}{2}\lambda_*$, as long as $m = \Omega\left(\frac{n^2}{\lambda_*^2}\left(\log\frac{n}{\lambda_* t}\right)\right)$.

Discussion on comparisons with [27] and [28]. In [27], the prediction of an implicit network is formulated as a weighted summation of equilibrium points and explicit features *i.e.* $\hat{\boldsymbol{y}} = \boldsymbol{a}^\top \boldsymbol{Z} + \boldsymbol{b}^\top \phi(\boldsymbol{U}\boldsymbol{X})$. Therefore, it is hard to measure the contribution of the equilibrium point to the capacity of implicit models in their case. In contrast, we only use equilibrium points for predictions, and our result illustrates that arbitrary training label can be fitted only using



(a) Synthetic data (b) MNIST (c) CIFAR10

Figure 1: Results of different widths on (a) Synthetic data; (b) MNIST; (c) CIFAR10.

equilibrium points. In a concurrent work [28] on ReLU implicit networks, they consider "a subset of initialization" which requires entries of $\boldsymbol{W}$ to be non-negative. In their case, one can show that $\boldsymbol{Z} = (\boldsymbol{I} - \boldsymbol{W})^{-1}\phi(\boldsymbol{W}\boldsymbol{X})$ at initialization. This implies that the initial equilibrium point is essentially a linear transformation of the explicit feature $\phi(\boldsymbol{U}\boldsymbol{X})$. In contrast, we consider general Gaussian initialization which is commonly used in practice. The results in [27, 28] can be built upon the lower bound of the least singular value of $\phi(\boldsymbol{U}\boldsymbol{X})$, which is given by previous works on explicit models. However, previous analyzes do not applied in our case. In order to address the technical problem, we propose a novel probabilistic framework. Therefore, our studies have distinct differences.

5 Numerical Experiments

In this section, we implement several numerical experiments to verify our main theoretical conclusions. We first evaluate our method on various datasets including synthetic data, MNIST, and CIFAR10. For constructing synthetic data, we uniformly generate $n = 1000$ data points from a $d = 1000$ dimensional sphere with radius $\sqrt{d}$, and labels are generated from a one-dimensional standard Gaussian distribution. For each dataset of MNIST and CIFAR10, we randomly sample 500 images from each of class 0 and class 1 to generate the training dataset with $n = 1000$ samples. We use Gaussian initialization as suggested in Assumption 1 and σ_w^2 is set as 0.08. We normalize each data point as suggested in Assumption 2.

In the first experiment, we test how the width affects the convergence rate. As shown in Figure 1, the convergence speed becomes faster as m increases, and the final training loss becomes smaller. We believe that the reason is as m increases, Gram matrices become more stable. The second experiment verifies that $\|\boldsymbol{W}(\tau)\|_2$ is smaller than 1 throughout the training, which implies that the unique equilibrium point always exists.

6 Conclusion

In this paper, we analyze the gradient dynamics of DEQs with the quadratic loss function. Under a specific initial condition, we prove that the GD converges to a global optimum at a linear rate. By performing a fine-grained analysis on Gaussian initialized DEQs, we further show that the initial conditions are satisfied via mild over-parameterization. Specifically, we present a new probabilistic framework to address the challenge arising from the weight-sharing and the infinite depth. To the best of our knowledge, it is the first time to analyze the equilibrium point of a random DEQ. Moreover, we show that the unique equilibrium points always exist during the training process. Our analysis is specific to ReLU DEQs. For future research, it would be interesting to generalize the result to other nonlinear activations. Moreover, it would be interesting to explore the generalization performance of over-parameterized DEQs based on our analysis at initialization.

References

[1] Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. Deep equilibrium models. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[2] Shaojie Bai, Vladlen Koltun, and J Zico Kolter. Multiscale deep equilibrium models. *Advances in Neural Information Processing Systems*, 2020.

[3] Xingyu Xie, Qiuhao Wang, Zenan Ling, Xia Li, Guangcan Liu, and Zhouchen Lin. Optimization induced equilibrium networks: An explicit optimization perspective for understanding equilibrium models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[4] Davis Gilton, Gregory Ongie, and Rebecca Willett. Deep equilibrium architectures for inverse problems in imaging. *arXiv preprint arXiv:2102.07944*, 2021.

[5] Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems*, pages 8141–8150, 2019.

[6] Simon S Du, Jason D Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*, pages 3003–3048, 2019.

[7] Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in neural information processing systems*, 2018.

[8] Quynh Nguyen. On the proof of global convergence of gradient descent for deep relu networks with linear widths. In *International Conference on Machine Learning*, pages 8056–8062. PMLR, 2021.

[9] Ezra Winston and J Zico Kolter. Monotone operator equilibrium networks. In *Advances in Neural Information Processing Systems*, 2020.

[10] Max Revay, Ruigang Wang, and Ian R Manchester. Lipschitz bounded equilibrium networks. *arXiv preprint arXiv:2010.01732*, 2020.

[11] Kenji Kawaguchi. On the theory of implicit deep learning: Global convergence with implicit layers. In *International Conference on Learning Representations*, 2020.

[12] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, pages 8580–8589, 2018.

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

[14] Difan Zou and Quanquan Gu. An improved analysis of training over-parameterized deep neural networks. In *Advances in neural information processing systems*, pages 2055–2064, 2019.

[15] Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Gradient descent optimizes over-parameterized deep relu networks. *Machine Learning*, 109(3):467–492, 2020.

[16] 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.

[17] Quynh N Nguyen and Marco Mondelli. Global convergence of deep networks with one wide layer followed by pyramidal topology. *Advances in Neural Information Processing Systems*, 33:11961–11972, 2020.

[18] Greg Yang. Wide feedforward or recurrent neural networks of any architecture are gaussian processes. *Advances in Neural Information Processing Systems*, 32, 2019.

[19] Greg Yang. Tensor programs ii: Neural tangent kernel for any architecture. *arXiv preprint arXiv:2006.14548*, 2020.

[20] Greg Yang. Tensor programs iii: Neural matrix laws. *arXiv preprint arXiv:2009.10685*, 2020.

[21] Erwin Bolthausen. An iterative construction of solutions of the tap equations for the sherrington–kirkpatrick model. *Communications in Mathematical Physics*, 325(1):333–366, 2014.

[22] Mohsen Bayati and Andrea Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57(2):764–785, 2011.

[23] Ping Li and Phan-Minh Nguyen. On random deep weight-tied autoencoders: Exact asymptotic analysis, phase transitions, and implications to training. In *International Conference on Learning Representations*, 2018.

[24] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. On the convergence rate of training recurrent neural networks. In *Advances in Neural Information Processing Systems*, pages 6676–6688, 2019.

[25] Lifu Wang, Bo Shen, Bo Hu, and Xing Cao. On the provable generalization of recurrent neural networks. *Advances in Neural Information Processing Systems*, 34, 2021.

[26] Zhili Feng and J Zico Kolter. On the neural tangent kernel of equilibrium models. *arxiv*, 2020.

[27] T. Gao, H. Liu, J. Liu, H. Rajan, and H. Gao. A global convergence theory for deep relu implicit networks via over-parameterization. *ICLR*, 2022.

[28] T. Gao and H. Gao. Gradient descent optimizes infinite-depth relu implicit networks with linear widths. *arxiv*, 2022.

[29] Laurent El Ghaoui, Fangda Gu, Bertrand Travacca, Armin Askari, and Alicia Tsai. Implicit deep learning. *SIAM Journal on Mathematics of Data Science*, 3(3):930–958, 2021.

[30] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[31] B. T. Polyak. Gradient methods for minimizing functionals. *Zh.vychisl.mat.mat.fiz*, 3:643–653, 1963.

[32] Simon S Du, X. Zhai, B. Poczos, and A. Singh. Gradient descent provably optimizes over-parameterized neural networks. *ICLR*, 2018.

[33] D. Zou and Q. Gu. An improved analysis of training over-parameterized deep neural networks. *NeurIPS*, 2019.

[34] Amit Daniely, Roy Frostig, and Yoram Singer. Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.