Global Convergence Rate of Deep Equilibrium Models with General Activations

Anonymous authors
Paper under double-blind review

Abstract

In a recent paper, Ling et al. investigated the over-parametrized Deep Equilibrium Model (DEQ) with ReLU activation. They proved that the gradient descent converges to a globally optimal solution at a linear convergence rate for the quadratic loss function. This paper shows that this fact still holds for DEQs with any general activation that has bounded first and second derivatives. Since the new activation function is generally non-homogeneous, bounding the least eigenvalue of the Gram matrix of the equilibrium point is particularly challenging. To accomplish this task, we need to create a novel population Gram matrix and develop a new form of dual activation with Hermite polynomial expansion.

1 Introduction

Deep learning is a class of machine learning algorithms that uses multiple layers to progressively extract higher-level features from the raw input. For example, in image processing, lower layers may identify edges, while higher layers may identify the concepts relevant to a human such as digits or letters or faces. Deep neural networks have underpinned state of the art empirical results in numerous applied machine learning tasks (Krizhevsky et al., 2012). Understanding neural network learning, particularly its recent successes, commonly decomposes into the two main themes: (i) studying generalization capacity of the deep neural networks and (ii) understanding why efficient algorithms, such as stochastic gradient, find good weights. Though still far from being complete, previous work provides some understanding on generalization capability of deep neural networks. However, question (ii) is rather poorly understood. While learning algorithms succeed in practice, theoretical analysis is overly pessimistic. Direct interpretation of theoretical results suggests that when going slightly deeper beyond single layer networks, e.g. to depth-two networks with very few hidden units, it is hard to predict even marginally better than random (Daniely et al., 2013; Kearns & Valiant, 1994).

The standard approach to develop generalization bounds on deep learning (and machine learning) was developed in seminal papers by (Vapnik, 1998), and it is based on bounding the difference between the generalization error and the training error. These bounds are expressed in terms of the so called VC-dimension of the class. However, these bounds are very loose when the VC-dimension of the class can be very large, or even infinite. In 1998, several authors (Bartlett & Shawe-Taylor, 1999; Bartlett et al., 1998) suggested another class of upper bounds on generalization error that are expressed in terms of the empirical distribution of the margin of the predictor (the classifier). Later, Koltchinskii and Panchenko proposed new probabilistic upper bounds on generalization error of the combination of many complex classifiers such as deep neural networks (Koltchinskii & Panchenko, 2002). These bounds were developed based on the general results of the theory of Gaussian, Rademacher, and empirical processes in terms of general functions of the margins, satisfying a Lipschitz condition. They improved previously known bounds on generalization error of convex combination of classifiers. (Truong, 2022a) and Truong (2022b) have recently provided generalization bounds for learning with Markov dataset based on Rademacher and Gaussian complexity functions. The development of new symmetrization inequalities and contraction lemmas in high-dimensional probability for Markov chains is a key element in these works. Several recent works have focused on gradient descent based PAC-Bayesian algorithms, aiming to minimise a generalisation bound for stochastic classifiers (Biggs & Guedj, 2021; Dziugaite & Roy., 2017). Most of these studies use a surrogate loss to avoid dealing with

the zero-gradient of the misclassification loss. There were some other works which use information-theoretic approach to find PAC-bounds on generalization errors for machine learning (Esposito et al., 2021; Xu & Raginsky, 2017) and deep learning (Jakubovitz et al., 2018).

Recently, deep equilibrium model (DEQ)(Bai et al., 2019) was introduced as a new approach to modelling sequential data. In many existing deep sequence models, the hidden layers converge toward some fixed points. DEQ directly finds these equilibrium points via root-finding of implicit equations. Such a model is equivalent to an infinite-depth weight-tied model with input-injection. DEQ has emerged as an important model in various aplications such as computer vision (Bai et al., 2020; Xie et al., 2022), natural language processing (Bai et al., 2019), and inverse problems (Gilton et al., 2021). This model has been shown to achieve performance competitive with the state-of-the-art deep networks while using significantly less memory. Despite of the empirical success of DEQ, theoretical understanding of this model is still limited. The effectiveness of over-parameterization in optimizing feedforward neural networks has been validated in many research literature (Arora et al., 2019; Du et al., 2018; Li & Liang, 2018). A recent work (Nguyen, 2021) showed that the convergence of gradient descent (GD) to a global optimum can be guaranteed when the width of the last hidden layer exceeds the number of training samples. The main idea is to investigate the property at initialization and bound the traveling distance of GD from the initialization.

However, it remains unknown whether the above results can be directly applied to DEQs. Due to the implicit weight-sharing, the initial random weights and features are dependent, which causes the standard concentration approaches in the existing research literature fail in DEQs. Recently, Ling et al. (2022) investigated the training dynamics of over-parameterized DEQs with ReLU activation. More specifically, they proposed 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, they proved that the gradient descent converges to a globally optimal solution at a linear convergence rate for the quadratic loss function. To achieve this target, they developed a lower bound on the least eigenvalue of the Gram matrix for the DEQs with ReLU activation. One interesting open question is whether the gradient descent algorithm still converge at a linear rate for DEQs with non-linear activation functions? In this paper, we show that this fact still holds for DEQs with a general activation function which has bounded first and second derivatives. Many popular activation functions such as $1/(1+e^{-x}), \mathrm{erf}(x), x/\sqrt{1+x^2}, \sin(x), \tanh(x)$ satisfy the boundedness requirements. In general, the new activation function does not have homogeneous property as ReLU, hence a novel population Gram matrix is designed for DEQs with general activations, and a new form of dual activation with Hermite polynomial expansion is developed in our work.

2 Problem settings

We consider the same model as Ling et al. (2022). However, different from Ling et al. (2022), we assume that the activation function, φ, satisfies some constraints in the first and second derivatives. These properties can be observed in many common activation functions. More specifically, we define a vanilla deep equilibrium model (DEQ) with the transform of the l-th layer as

$$\mathbf{T}^{(l)} = \varphi(\mathbf{W}\mathbf{T}^{(l-1)} + \mathbf{U}\mathbf{X}) \tag{1}$$

where $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n] \in \mathbb{R}^{d\times n}$ denotes the training inputs, $\mathbf{U} \in \mathbb{R}^{m\times d}$ and $\mathbf{W} \in \mathbb{R}^{m\times m}$ are trainable weight matrices, and $\mathbf{T}^{(l)} \in \mathbb{R}^{m\times n}$ is the output feature at the l-th hidden layer. If we were to repeat this update an infinite number of times, we would essentially be modeling an infinitely deep network of the form above. In practice, what we find is that for most "typical" deep layers the valued actually converge to a fixed point or equilibrium point (Bai et al., 2019). The output of the last hidden layer is defined by $\mathbf{T}^* := \lim_{l\to\infty} \mathbf{T}^{(l)}$. Therefore, instead of running infinitely deep layer-by-layer forward propagation, $\mathbf{T}^*$ can be calculated by directly solving the equilibrium point of the following equation

$$\mathbf{T}^* = \varphi(\mathbf{W}\mathbf{T}^* + \mathbf{U}\mathbf{X}). \tag{2}$$

Let $\mathbf{y} = [y_1, y_2, \cdots, y_n] \in \mathbb{R}^n$ denote the labels, and $\hat{\mathbf{y}}(\boldsymbol{\theta}) = \mathbf{a}^T\mathbf{T}^*$ be the prediction function with $\mathbf{a} \in \mathbb{R}^m$ being a trainable vector and $\theta = \mathrm{vec}(\mathbf{W}, \mathbf{U}, \mathbf{a})$. Our target is to minimize the empirical risk with the

quadratic loss function:

$$\Phi(\boldsymbol{\theta}) = \frac{1}{2}\|\hat{\mathbf{y}}(\boldsymbol{\theta}) - \mathbf{y}\|_2^2. \tag{3}$$

To optimize this loss function, we use the gradient descent update $\boldsymbol{\theta}(\tau+1) = \boldsymbol{\theta}(\tau) - \eta\nabla\Phi(\boldsymbol{\theta}(\tau))$, where η is the learning rate and $\boldsymbol{\theta}(\tau) = \mathrm{vec}(\mathbf{W}(\tau), \mathbf{U}(\tau), \mathbf{a}(\tau))$. For notational simplicity, we omit the superscript and denote $\mathbf{T}$ to be the equilibrium $\mathbf{T}^*$ when it is clear from the context. Moreover, the Gram matrix of the equilibrium point is defined by $\mathbf{G}(\tau) := \mathbf{T}^T(\tau)\mathbf{T}(\tau)$ and we define its least eigenvalue by $\lambda_\tau = \lambda_{\min}(\mathbf{G}(\tau))$. In this paper, for brevity we denote by $\mathbf{G} = \mathbf{G}(0)$.

Definition 1. *An activation $\varphi : \mathbb{R} \to \mathbb{R}$ is L-bounded if it is twice continuously differentiable and $\|\varphi\|_\infty, \|\varphi'\|_\infty, \|\varphi''\|_\infty \le L$.*

In this paper, we assume that $\varphi(\cdot)$ is L-bounded. In addition, the following holds:

$$q := \sqrt{\frac{2}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \varphi^2(z)\exp\left(-\frac{z^2}{2}\right)dz} > 0.$$

Many popular activation functions such as $1/(1+e^{-x}), \mathrm{erf}(x), x/\sqrt{1+x^2}, \sin(x), \tanh(x)$ satisfy the boundedness requirements.

Definition 2. *Two vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ are said to be parallel, denoted $\mathbf{a} \parallel \mathbf{b}$, if there exists a scalar $\kappa \in \mathbb{R}$ such that $\mathbf{a} = \kappa\mathbf{b}$. If $\mathbf{a}$ and $\mathbf{b}$ are not parallel, we write $\mathbf{a} \nparallel \mathbf{b}$.*

Besides, we use similar assumptions on the random initialisation and input data as Ling et al. (2022):

- **Assumption 1** (Random initialization). Assume that $\sigma_w^2 < \frac{1}{48L^2}$. In addition, $\mathbf{W}$ is initialized with an $m \times m$ matrix with i.i.d. entries $\mathbf{W}_{ij} \sim \mathcal{N}(0, 2\sigma_w^2/m)$, $\mathbf{U}$ is initialized with an $m \times d$ matrix with i.i.d. entries $\mathbf{U}_{ij} \sim \mathcal{N}(0, 2/m)$, and $\mathbf{a}$ is initialized with a random vector with i.i.d. entries $\sim \mathcal{N}(0, 1/m)$.
- **Assumption 2** (Input data). We assume that (i) $\|\mathbf{x}_i\|_2 = \sqrt{d}$ for all $i \in [n]$ and $\mathbf{x}_i \nparallel \mathbf{x}_j$ for all $i \neq j$; (ii) the labels satisfy $|y_i| = O(1)$ for all $i \in [n]$.

3 Main Results

In this paper, we show that if the learning rate is small enough, the loss converges to a global minimum at linear rate. The result is as follows.

Theorem 3. *Consider a DEQ. Let δ be a constant such that $\|\mathbf{W}(0)\| + \delta < 1/L$. Denote by $\bar{\rho}_w = \|\mathbf{W}(0)\|_2 + \delta, \bar{\rho}_u = \|\mathbf{U}(0)\|_2 + \delta, \bar{\rho}_a = \|\mathbf{a}(0)\|_2 + \delta$ and define*

$$c_a = \frac{L\bar{\rho}_u}{1 - L\bar{\rho}_w}, \qquad c_u = \frac{L\bar{\rho}_a}{1 - L\bar{\rho}_w}, \qquad c_m = \frac{|\sigma(0)|\sqrt{mn}}{1 - L\bar{\rho}_w}. \tag{4}$$

In addition, assume at initialization that

$$\lambda_0 \ge \frac{4}{\delta}\max\left\{c_u\big(c_a\|\mathbf{X}\|_F + c_m\big), c_u\|\mathbf{X}\|_F, c_a\|\mathbf{X}\|_F + c_m\right\}\|\hat{\mathbf{y}}(0) - \mathbf{y}\|, \tag{5}$$

$$\lambda_0^{3/2} \ge \frac{4(2+\sqrt{2})L}{(1 - L\bar{\rho}_w)}\left[c_u\big(c_a\|\mathbf{X}\|_F + c_m\big)^2 + c_u\|\mathbf{X}\|_F^2\right]\|\hat{\mathbf{y}}(0) - \mathbf{y}\|_2, \tag{6}$$

$$\lambda_0 \ge 8\left[c_u^2\big(c_a\|\mathbf{X}\|_F + c_m\big)^2 + c_u^2\|\mathbf{X}\|_F^2\right] \tag{7}$$

where λ_0 is the least eigenvalue of $\mathbf{G}(0) = \mathbf{T}(0)^T\mathbf{T}(0)$. Then, if the learning rate satisfies

$$\eta < \min\left(\frac{2}{\lambda_0}, \frac{2[c_u^2(c_a\|\mathbf{X}\|_F + c_m)^2 + c_u^2\|\mathbf{X}\|_F^2]}{c_u^2(c_a\|\mathbf{X}\|_F + c_m)^2 + c_u^2\|\mathbf{X}\|_F^2 + (c_a\|\mathbf{X}\|_F + c_m)^2}\right),$$

for every $\tau \ge 0$, the following hold:

- $\|\mathbf{W}(\tau)\|_2 \leq 1/L$, *i.e., the equilibrium points always exists,*
- $\lambda_\tau \geq \frac{1}{2}\lambda_0$, *and*

$$\|\nabla_\theta \Phi(\theta(\tau))\|_2^2 \geq \lambda_0 \Phi(\theta(\tau)). \tag{8}$$

- *The loss converges to a global minimum as*

$$\Phi(\theta(\tau)) \leq \left(1 - \eta\frac{\lambda_0}{2}\right)^\tau \Phi(\theta(0)). \tag{9}$$

The main challenge now is to find some initializations such that λ_0 satisfies all the conditions in Theorem 3. To lower bound λ_0, we need to design a population Gram matrix $\mathbf{K}$ and compare λ_0 with the least eigenvalue of $\mathbf{K}$ Ling et al. (2022). However, since the new activation function, φ, is non-linear in general, bounding λ_0 is more challenging than the ReLU network in Ling et al. (2022). The non-homegeneity of activation functions causes the techniques to design $\mathbf{K}$ in (Ling et al., 2022, Definition 1) can not be applied. For example, (Ling et al., 2022, Eq. 11) only holds for ReLU.

In Section 4, we propose a new method to create the population Gram matrix $\mathbf{K}$ for DEQs with general Lipschitz activation function. By using our new form of dual activation and Hermite polynomial expansion, we can prove that $\mathbf{K}$ is symmetric positive definite. In addition, we show that with probability at least $1-t$, $\lambda_0 \geq \frac{m}{2}\lambda_*$ provided that $m = \Omega\big(\frac{n^3}{\lambda_*^2}\log\frac{n}{t}\big)$ where λ_* is the least eigenvalue of $\mathbf{K}$ (cf. Section 7). This fact indicates that all the conditions of Theorem 3 at least hold for over-parametrized DEQs (or m sufficiently large) with $\varphi(0) = 0$. Hence, by (9) in Theorem 3, the gradient descent algorithm converges to a global optimum at a linear rate for the over-parametrized DEQs if the number of repetitions in (1) sufficiently large. This interesting fact is reaffirmed by our numerical experiments on real datasets such as MNIST and CFAR10 in Section 8.

4 A novel design of the population Gram matrix K

The key approach in lower bounding λ_0 is to design a population Gram matrix $\mathbf{K}$ in such a way that we can lower bound λ_0 by the least eigenvalue of $\mathbf{K}$ and that $\mathbf{K}$ is symmetric positive definite. This novel population Gram matrix is developed through our introduction of a new form of dual activation.

First, we define a new class of dual activation functions $\tilde{Q}_{\alpha,\beta} : [-1,1] \to \mathbb{R}$ for all pairs $(\alpha,\beta) \in \mathbb{R}_+^2$.

Definition 4. *Recall the definition of q in (4). For each pair (α,β), define*

$$\tilde{Q}_{\alpha,\beta}(x) := \frac{1}{\alpha\beta q^2}\mathbb{E}_{(a,b)^T\sim\mathcal{N}\left(0,\begin{bmatrix}1 & x\\ x & 1\end{bmatrix}\right)}\big[\varphi(\alpha a)\varphi(\beta b)\big], \quad \forall |x| \leq 1. \tag{10}$$

If $\varphi(x) = \max\{x, 0\}$ (ReLU), then $\tilde{Q}_{\alpha,\beta}(x) = \bar{Q}(x)$ for all $(\alpha,\beta) \in \mathbb{R}_+^2$, where

$$\bar{Q}(x) := \mathbb{E}_{(a,b)^T\sim\mathcal{N}\left(0,\begin{bmatrix}1 & x\\ x & 1\end{bmatrix}\right)}[\varphi(a)\varphi(b)]$$

is the dual activation defined in (Daniely et al., 2016, Sec. 3.2).

Now, we provide a novel design of the population Gram matrix $\mathbf{K}$ based on this new dual activation function.

Definition 5. *Given the training input $\mathbf{X} := [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n]$ satisfying Assumption 2. Let*

$$Q_{ij}(x) := \tilde{Q}_{\sqrt{2\left(\frac{\sigma_w^2}{m}\mathbb{E}[\mathbf{G}_{ii}]+1\right)},\sqrt{2\left(\frac{\sigma_w^2}{m}\mathbb{E}[\mathbf{G}_{jj}]+1\right)}}(x), \qquad \forall x \in \mathbb{R}. \tag{11}$$

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

$$\rho_{ij}^{(0)} = 0, \tag{12}$$

$$\rho_{ii}^{(l)} = 2q^2\sigma_w^2\rho_{ii}^{(l-1)}Q_{ii}(1) + 1, \tag{13}$$

$$\rho_{ij}^{(l)} = \sqrt{\rho_{ii}^{(l)}\rho_{jj}^{(l)}}, \quad i \neq j \tag{14}$$

$$\mathbf{K}^{(0)} = 0, \tag{15}$$

$$\nu_{ij}^{(l)} = \frac{\sigma_w^2\mathbf{K}_{ij}^{(l-1)} + d^{-1}\mathbf{x}_i^T\mathbf{x}_j}{\sqrt{\big(\sigma_w^2\mathbf{K}_{ii}^{(l-1)} + 1\big)\big(\sigma_w^2\mathbf{K}_{jj}^{(l-1)} + 1\big)}} \tag{16}$$

$$\mathbf{K}_{ij}^{(l)} = 2q^2\rho_{ij}^{(l)}Q_{ij}(\nu_{ij}^{(l)}) \tag{17}$$

for all $l \geq 1$ *and* $i, j \in [n] \times [n]$.

The next result shows that λ_0 can be lower bounded via the least eigenvalue of the population matrix $\mathbf{K}$.

Theorem 6. *If* $m = \Omega\big(\frac{n^2}{\lambda_*^2}\log\frac{n}{t}\big)$, *with probability at least* $1 - t$, *it holds that*

$$\lambda_0 \geq \frac{m}{2}\lambda_*. \tag{18}$$

Finally, the following result shows sufficient conditions such that $\mathbf{K}$ is strictly positive definite.

Theorem 7. *Assume that there exists a polynomial expansion of* $\tilde{Q}_{\alpha,\alpha}$ *satisfying:*

$$\tilde{Q}_{\alpha,\alpha}(x) = \sum_{r=0}^{\infty}\mu_{r,\alpha}^2(\varphi)x^r \tag{19}$$

for all $\alpha > 0$ *such that* $\sup\{r : \min_{\alpha\in[2,2(\sigma_w^2L^2+1)]}\mu_{r,\alpha}^2(\varphi) > 0\} = \infty$. *Then,* $\mathbf{K}$ *is strictly positive definite with the least eigenvalue satisfying* $\lambda_* \geq \lambda_0^*$ *for some* $\lambda_0^* > 0$ *which does not depend on* m.

5 Proof of Theorem 6

To prove Theorem 6, we first state some auxiliary results based on the population Gram matrix $\mathbf{K}$ in Definition 5. The proofs of these lemmas and prepositions can be found in Supplement Material.

Lemma 8. *Recall the definition of* $\tilde{Q}_{\alpha,\beta}$ *in Definition 4. Then, the following hold for all* $\alpha \geq 1, \beta \geq 1$ *and* $x \in \mathbb{R}$*:*

$$\big|\tilde{Q}_{\alpha,\beta}(x)\big| \leq \sqrt{\tilde{Q}_{\alpha,\alpha}(1)\tilde{Q}_{\beta,\beta}(1)}, \tag{20}$$

$$\big|\tilde{Q}_{\alpha,\beta}(x)\big| \leq \frac{4L^2}{q^2}, \qquad \forall|x| \leq 1. \tag{21}$$

In addition, $\tilde{Q}_{\alpha,\beta}(\cdot)$ *is* $\frac{2L^2}{q^2}$*-Lipchitz for any fixed positive pair* (α, β).

Lemma 9. *(Ling et al., 2022, Proof of Lemma 4) For* $l \geq 1$, $\mathbf{G}_{ij}^{(l+1)}$ *can be reconstructed as* $\mathbf{G}_{ij}^{(l+1)} = \varphi(\mathbf{M}\mathbf{h}_{l+1})^T\varphi(\mathbf{M}\mathbf{h}'_{l+1})$ *such that*

- *(i)* $\mathbf{h}_{l+1}^T\mathbf{h}'_{l+1} = \frac{\sigma_w^2}{m}\mathbf{G}_{ij}^{(l)} + \frac{1}{d}\mathbf{x}_i^T\mathbf{x}_j$,
- *(ii)* $\mathbf{M} \in \mathbb{R}^{m\times(2l+d+2)}$ *is a rectangle matrix, and the entries of* $\mathbf{M}$ *are i.i.d. from* $\mathcal{N}(0, 2)$ *conditioning on previous layers.*

Lemma 10. *For the given setting, we have*

$$\rho_{ii}^{(l)} = \sigma_w^2 \mathbf{K}_{ii}^{(l-1)} + 1, \tag{22}$$

$$\rho_{ij}^{(l)} \nu_{ij}^{(l)} = \sigma_w^2 \mathbf{K}_{ij}^{(l-1)} + d^{-1} \mathbf{x}_i^T \mathbf{x}_j, \qquad \forall i, j, \tag{23}$$

and

$$\nu_{ij}^{(l)} = \begin{cases} \frac{Q_{ij}\left(\nu_{ij}^{(l-1)}\right)/\sqrt{Q_{ii}(1)Q_{jj}(1)}\sqrt{(\rho_{ii}^{(l)}-1)(\rho_{jj}^{(l)}-1)}+d^{-1}\mathbf{x}_i^T\mathbf{x}_j}{\sqrt{\rho_{ii}^{(l)}\rho_{jj}^{(l)}}}, & i \neq j \\ 1, & i = j \end{cases}. \tag{24}$$

In addition, we also have

$$\left|\nu_{ij}^{(l)}\right| \leq 1 \tag{25}$$

for all $i, j \in [n] \times [n]$ and $l \geq 0$.

Proposition 11. *Under the Assumptions 1 and 2 we have $\|\mathbf{K} - \mathbf{K}^{(l)}\|_F = O\big(n\big(8L^2\sigma_w^2\big)^l\big)$ which implies that, for $l \to \infty$, $\mathbf{K}^{(l)} \to \mathbf{K}$ with entries*

$$\mathbf{K}_{ij} = 2q^2 Q_{ij}(\nu_{ij})\sqrt{\rho_{ii}\rho_{jj}} \tag{26}$$

where

$$\nu_{ij} = \begin{cases} \frac{Q_{ij}\left(\nu_{ij}\right)/\sqrt{Q_{ii}(1)Q_{jj}(1)}\sqrt{(\rho_{ii}-1)(\rho_{jj}-1)}+d^{-1}\mathbf{x}_i^T\mathbf{x}_j}{\sqrt{\rho_{ii}\rho_{jj}}}, & i \neq j \\ 1, & i = j \end{cases}. \tag{27}$$

Here,

$$\rho_{ii} = \frac{1}{1 - 2q^2\sigma_w^2 Q_{ii}(1)}. \tag{28}$$

Proposition 12. *Under Assumptions 1 and 2 with probability at least $1 - n^2\exp(-\Omega(m))$, it holds that*

$$\frac{1}{m}\left\|\mathbf{G} - \mathbf{G}^{(l)}\right\|_F = O\left(n\big(2L\sqrt{2}\sigma_w\big)^l\right). \tag{29}$$

Proposition 13. *Under Assumptions 1 and 2, with probability at least $1 - n^2 l\exp\big\{-\Omega(8^l L^{2l}\sigma_w^{2l}mnL^2) + O(l^2)\big\}$, it holds that*

$$\left\|\frac{1}{m}\mathbf{G}^{(l)} - \mathbf{K}^{(l)}\right\|_F = O\left(n(2L\sqrt{2}\sigma_w)^l\right). \tag{30}$$

By combining Propositions 11–13, we can bound λ_0 via the least eigenvalue of the population matrix $\mathbf{K}$ as follows.

Proof of Theorem 6. From Propositions 11–13, with probability at least $1-n^2\exp\big(-\Omega(m8^l L^{2l}\sigma_w^{2l}) + O(l^2)\big)$, it holds that

$$\begin{aligned} \left\|\frac{1}{m}\mathbf{G} - \mathbf{K}\right\|_F &\leq \frac{1}{m}\left\|\mathbf{G} - \mathbf{G}^{(l)}\right\|_F + \left\|\frac{1}{m}\mathbf{G}^{(l)} - \mathbf{K}^{(l)}\right\| + \left\|\mathbf{K} - \mathbf{K}^{(l)}\right\|_F \\ &= O\left(n\left(2L\sqrt{2}\sigma_w\right)^l\right) + O\left(n\left(2L\sqrt{2}\sigma_w\right)^l\right) + O\left(n(8L^2\sigma_w^2)^l\right) \\ &= O\left(n\left(2L\sqrt{2}\sigma_w\right)^l\right), \end{aligned} \tag{31}$$

where (31) follows from $\sigma_w^2 < 1/(8L^2)$.

Next, we fix l to omit the explicit dependence on l. Specifically, let

$$l = \Theta(\log(2\lambda_*^{-1}n)/\log(\sqrt{2}/(4L\sigma_w)),$$

then from (31), we have

$$\left\|\frac{1}{m}\mathbf{G} - \mathbf{K}\right\|_F \leq \frac{\lambda_*}{2}.$$

It is easy to prove by induction that $\mathbf{K}$ is symmetric. Therefore, by Weyl's inequality (Ling et al., 2022, Lemma 5), it holds that

$$\max_{i\in[r]}\left|\lambda_i\left(\frac{1}{m}\mathbf{G}\right) - \lambda_i(\mathbf{K})\right| \leq \left\|\frac{1}{m}\mathbf{G} - \mathbf{K}\right\|_2 \leq \left\|\frac{1}{m}\mathbf{G} - \mathbf{K}\right\|_F \leq \frac{\lambda_*}{2}.$$

Now, by choosing $i_0 := \arg\min_i \lambda_i(\mathbf{K})$, we have

$$\lambda_{i_0}(\mathbf{K}) = \lambda_* \tag{32}$$

and

$$\left|\frac{1}{m}\lambda_{\min}(\mathbf{G}) - \lambda_*\right| \leq \frac{\lambda_*}{2}. \tag{33}$$

It follows from (32) and (33) that

$$\lambda_0 = \lambda_{\min}(\mathbf{G}) \geq \frac{m}{2}\lambda_*.$$

Consequently, w.p. $\geq 1-t$, we have $\lambda_0 \geq \frac{m}{2}\lambda_*$ provided that $m = \Omega\big(\frac{n^2}{\lambda_*^2}\log\frac{n}{t}\big)$. □

6 Checking the conditions of Theorem 7

In this section, we will show how the condition in Theorem 7 holds for some common activation functions. We first recall the definition of a traditional dual activation function, say $\hat{\varphi}$, associate with φ in (Daniely et al., 2016, Sect. 4.2):

$$\hat{\varphi}(x) = \mathbb{E}_{(u,v)\sim\mathcal{N}\left(0,\begin{bmatrix}1 & x\\ x & 1\end{bmatrix}\right)}[\varphi(u)\varphi(v)].$$

Then, by using a similar proof as (Daniely et al., 2016, Lemma 11), it can be shown that the new activation function (see Definition 4) satisfies

$$\tilde{Q}_{\alpha,\alpha}(x) = \frac{1}{q^2\alpha^2}\sum_{n=1}^{\infty} a_n^2\alpha^{2n}x^n \tag{34}$$

if $\varphi(x) = \sum_{n=1}^{\infty} a_n h_n(x)$ (Hermite polynomial expansion) or $\hat{\varphi}(x) = \sum_{n=1}^{\infty} a_n^2 x^n$.

In the following, we apply (34) and show how the condition in Theorem 7 is fulfilled.

Example 14. *Consider the sine activation, $\varphi(x) = \sin(ax)$. By (Daniely et al., 2016, Sect. 8), we have*

$$\hat{\varphi}(x) = e^{-a^2}\sinh(a^2x).$$

By Taylor's expansion of sinh *function, i.e.,*

$$\sinh(x) = \sum_{r=0}^{\infty}\frac{1}{(2r+1)!}x^{2r+1}.$$

Hence, from (34) *we have*

$$\tilde{Q}_{\alpha,\alpha}(x) = \frac{1}{q^2\alpha^2} e^{-a^2} \sum_{r=0}^{\infty} \frac{a^{4r+2}\alpha^{4r+2}}{(2r+1)!} x^{2r+1},$$

which leads to

$$\mu^2_{r,\alpha}(\varphi) = \begin{cases} \frac{1}{q^2\alpha^2} e^{-a^2} \frac{a^{2r}\alpha^{2r}}{r!} & r \mod 2 = 1 \\ 0 & \textit{otherwise} \end{cases}.$$

This means that the condition in Theorem 7 is satisfied.

Example 15. *Consider the tanh activation function,* $\varphi(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$. *By (Szego, 1959, Eq. 8.23.4),* $\varphi(x)$ *can be uniquely described in the basis of Hermite polynomials,*

$$\varphi(x) = \sum_{n=1}^{\infty} a_n h_n(x)$$

where

$$|a_n| = \frac{1}{\sqrt{\pi} 2^n n!} \frac{\Gamma\left(\frac{n}{2}+1\right)}{\Gamma(n+1)} \exp\left(-\frac{\pi\sqrt{2n}}{2}\right).$$

Hence, from (34)*, we obtain*

$$\tilde{Q}_{\alpha,\alpha}(x) = \frac{1}{q^2\alpha^2} \sum_{n=1}^{\infty} a_n^2 \alpha^{2n} x^n,$$

so we have

$$\mu^2_{r,\alpha}(\varphi) = \frac{1}{q^2\alpha^2} a_n^2 \alpha^{2n}$$

This means that the condition in Theorem 7 is satisfied.

Example 16. *Consider the sigmoid activation function* $\varphi(x) = \frac{1}{1+e^{-x}}$. *It is known that*

$$\varphi(x) = \frac{1 + \tanh(x/2)}{2}.$$

Hence, by using similar arguments as Example 15, we can prove that the condition in Theorem 7 is also satisfied.

7 Weight Initialisation Algorithm

Before proposing an algorithm to initialise weights, we introduce some initial results.

Lemma 17. *(Vershynin, 2018, Theorem 4.4.5) For a random matrix* $\mathbf{A} \in \mathbb{R}^{n\times m}$ *with* $\mathbf{A}_{ij} \sim \mathcal{N}(0,1)$*, it holds that*

$$\|\mathbf{A}\|_2 \leq C(\sqrt{m} + \sqrt{n} + t) \tag{35}$$

with probability $1 - 2e^{-t^2}$*, where* C *is some constant.*

Lemma 18. *For any fixed* $t \in \mathbb{R}_+$*, it holds that*

$$\|\hat{\mathbf{y}}(0) - \mathbf{y}\| = O(\sqrt{n}) \tag{36}$$

with probability at least $1 - t$.

A weight initialisation algorithm (WIALG) is as follows.

- **Initialise:** $m = 1000, \sigma_w^2 = \frac{1}{96L^2}$.
- **Step 1:**
  - Generate a matrix $\mathbf{W} \in \mathbb{R}^{m\times m}$ where $\mathbf{W}_{ij} \sim \mathcal{N}\big(0, \frac{2\sigma_w^2}{m}\big)$.
  - Generate a matrix $\mathbf{U} \in \mathbb{R}^{m\times d}$ where $\mathbf{U}_{ij} \sim \mathcal{N}\big(0, \frac{2}{m}\big)$.
  - Generate a vector $\mathbf{a} \in \mathbb{R}^m$ where $\mathbf{a}_i \sim \mathcal{N}\big(0, \frac{1}{m}\big)$.
- **Step 2:**
  - Find a fixed-point $\mathbf{T}$ of the equation $\mathbf{T} = \varphi(\mathbf{WT}+\mathbf{UX})$ by using Anderson acceleration method Walker & Ni (2011).
  - Estimate $\frac{\mathbb{E}[\mathbf{G}_{ii}]}{m}$ by using the Monte-Carlo method. Note that by our Assumption 2, $\mathbb{E}[\mathbf{G}_{ii}]$ does not depend on i, so we only need to estimate $\frac{\mathbb{E}[\mathbf{G}_{11}]}{m}$.
  - Set $\hat{\mathbf{y}}(0) = \mathbf{a}^T\mathbf{T}$.
- **Step 3:**
  - Recursively construct a sequence $\mathbf{K}^{(l)}$ by using (12)–(17) until $\|\mathbf{K}^{(l)} - \mathbf{K}^{(l-1)}\|_F \le \varepsilon$ for some small value $\varepsilon > 0$.
  - Estimate the least eigenvalue λ_* of $\mathbf{K}^{(l)}$.
- **Step 4:** Check the following conditions:

$$\frac{m}{2}\lambda_* \ge \frac{4}{\delta}\max\Big\{c_u\big(c_a\|\mathbf{X}\|_F + c_m\big), c_u\|\mathbf{X}\|_F, c_a\|\mathbf{X}\|_F + c_m\Big\}\|\hat{\mathbf{y}}(0) - \mathbf{y}\|, \tag{37}$$

$$\Big(\frac{m}{2}\lambda_*\Big)^{3/2} \ge \frac{4(2+\sqrt{2})L}{(1-L\bar{\rho}_w)}\Big[c_u\big(c_a\|\mathbf{X}\|_F + c_m\big)^2 + c_u\|\mathbf{X}\|_F^2\Big]\|\hat{\mathbf{y}}(0) - \mathbf{y}\|_2, \tag{38}$$

$$\frac{m}{2}\lambda_* \ge 8\Big[c_u^2\big(c_a\|\mathbf{X}\|_F + c_m\big)^2 + c_u^2\|\mathbf{X}\|_F^2\Big], \tag{39}$$

  where $c_u, c_a, c_m, \bar{\rho}_w$ are defined in Theorem 3.
- **Step 5:** If all the conditions (37)–(39) hold, we STOP the initialisation. Otherwise, we increase $m = m + 10$ and REPEAT Step 1.

Theorem 19. *For DEQs with $\sigma(0) = 0$, WIALG will STOP with probability $1 - t$ at $m = \Omega\big(\frac{n^3}{(\lambda_*)^2}\log\frac{n}{t}\big)$.*

Proof. By using Theorem 6 and Theorem 7, it holds that $\lambda_* \ge \lambda_0^* > 0$ where λ_0^* is a function of n only, which does not depend on m. On the other hand, by Lemmas 17, it holds with probability $1 - \exp(-\Omega(m))$ that

$$\bar{\rho}_w = O(1), \qquad \bar{\rho}_u = O(1), \qquad \bar{\rho}_a = O(1),$$

which implies that

$$c_a = O(1), \qquad c_u = O(1), \qquad c_m = 0.$$

Now, by Theorem 6 and Theorem 7, it holds with probability $1 - t$ that

$$0 < \lambda_* \le \frac{2}{m}\lambda_0 \le \frac{2}{m}\mathrm{tr}(\mathbf{G}) = \frac{2}{m}\mathrm{tr}(\mathbf{T}(0)^T\mathbf{T}(0)) \le 2nL^2, \tag{40}$$

where (40) follows from the fact that $\mathbf{T}(0) = \varphi(\mathbf{WT}(0) + \mathbf{UX})$, so all elements of $\mathbf{T}(0)$ is bounded by L. In addition, by Assumption 2 we have

$$\|\mathbf{X}\|_F = \sqrt{nd}. \tag{41}$$

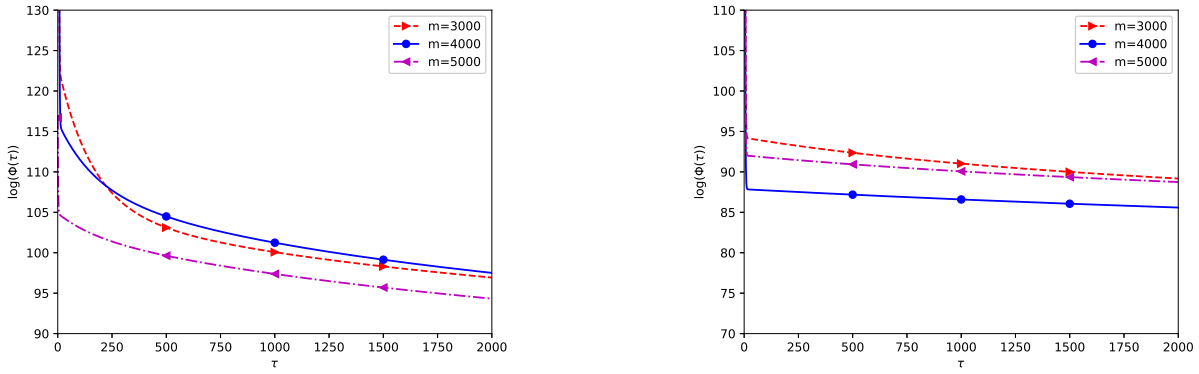
Hence, by combining with Lemma 18, i.e., $\|\hat{\mathbf{y}}(0) - \mathbf{y}\| = O(\sqrt{n})$, the inequalities (37)-(39) hold with probability at least $1 - t$ if $m = \Omega\big(\frac{n^3}{(\lambda_*)^2}\log\frac{n}{t}\big)$. □

Finally, by combining with Theorem 6, it hold that if $m = \Omega\big(\frac{n^3}{(\lambda_0^*)^2}\log\frac{n}{t}\big)$ and n sufficiently large, with probability at least $1-t$, all the conditions (5)-(7) in Theorem 3 hold.

8 Numerical Results

In this section, we implement some experiments to verify Theorem 3. We evaluate the DEQ model on MNIST and CIFAR-10 datasets. For each dataset, the training dataset is generated by randomly sampling 500 images from the first and second classes. We use Gaussian initialization as Assumption 1 and normalize each data point as Assumption 2.

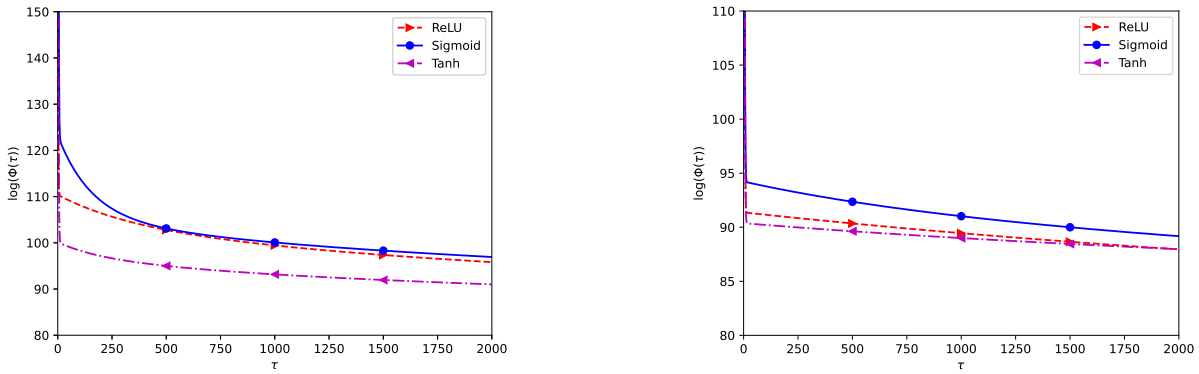
In the first experiment, we variate m and plot the training dynamic for MNIST and CIFAR-10 when φ is the sigmoid function ($L = 1$). It can be seen from Fig. 1 that as m big enough and τ sufficient large, the curves become straight lines. This fact re-affirms that (9) holds.



(a) MNIST (b) CIFAR-10

Figure 1: Training dynamics at different values of m.

In the second experiment, we variate the activation function and plot the training dynamic for MNIST and CIFAR-10 at $m = 3000$. It can be seen from Fig. 2 that as m big enough and τ sufficient large, the tanh network converges faster than the sigmoid or ReLU one for both datasets.



(a) MNIST (b) CIFAR-10

Figure 2: Training dynamics for different activation functions.

9 Conclusion

In this paper, we proved that the gradient descent converges to a globally optimal solution at a linear convergence rate for the quadratic loss function for the over-parametrized DEQ with L-bounded activation functions. This fascinating fact is also re-affirmed by our numerical experiments on MNIST and CFAR-10 datasets. To overcome new technical challenges caused by the non-linearity of activation functions, a novel population Gram matrix is introduced and a new form of dual activation with Hermite polynomial expansion is developed. An interesting future research direction is to study whether the linear convergence rate property still holds for other classes of activation functions.

References

Sanjeev Arora, Simon Shaolei Du, Wei Hu, Zhiyuan Li, Ruslan Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. In *Neural Information Processing Systems*, 2019.

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. Deep equilibrium models. *ArXiv*, abs/1909.01377, 2019.

Shaojie Bai, Vladlen Koltun, and J. Zico Kolter. Multiscale deep equilibrium models. *ArXiv*, abs/2006.08656, 2020.

Peter Bartlett and John Shawe-Taylor. *Generalization Performance of Support Vector Machines and Other Pattern Classifiers*, pp. 43–54. MIT Press, 1999.

Peter Bartlett, Yoav Freund, Wee Sun Lee, and Robert E. Schapire. Boosting the margin: a new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26(5):1651 – 1686, 1998.

F. Biggs and B. Guedj. Differentiable PAC-Bayes objectives with partially aggregated neural networks. *Entropy*, 23, 2021.

Amit Daniely, Nathan Linial, and Shai Shalev-Shwartz. From average case complexity to improper learning complexity. *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, 2013.

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 (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

Simon Shaolei Du, J. Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*, 2018.

G. K. Dziugaite and D. M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Uncertainty in Artificial Intelligence (UAI)*, 2017.

Amedeo Roberto Esposito, Michael Gastpar, and Ibrahim Issa. Generalization error bounds via Rényi-f-divergences and maximal leakage. *IEEE Transactions on Information Theory*, 67(8):4986–5004, 2021.

Davis Gilton, Greg Ongie, and Rebecca M. Willett. Deep equilibrium architectures for inverse problems in imaging. *IEEE Transactions on Computational Imaging*, 7:1123–1133, 2021.

R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1985.

D. Jakubovitz, R. Giryes, and M. R. D. Rodrigues. Generalization Error in Deep Learning. *Arxiv: 1808.01174*, 30, 2018.

Michael Kearns and Leslie G. Valiant. Cryptographic limitations on learning boolean formulae and finite automata. In *JACM*, 1994.

V. Koltchinskii and D. Panchenko. Empirical Margin Distributions and Bounding the Generalization Error of Combined Classifiers. *The Annals of Statistics*, 30(1):1 – 50, 2002.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60:84 – 90, 2012.

Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. *ArXiv*, abs/1808.01204, 2018.

Zenan Ling, Xingyu Xie, Qiuhao Wang, Zongpeng Zhang, and Zhouchen Lin. Global convergence of over-parameterized deep equilibrium models. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022.

Quynh N. Nguyen. On the proof of global convergence of gradient descent for deep relu networks with linear widths. *ArXiv*, abs/2101.09612, 2021.

G. Szego. Orthogonal polynomials. *American Mathematical Society*, 1959.

T. Tao. *Topics in Random Matrix Theory*. American Mathematical Society, 2012.

Lan V. Truong. Generalization error bounds on deep learning with markov datasets. *Thirty-Sixth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2022a.

Lan V. Truong. On rademacher complexity-based generalization bounds for deep learning. *ArXiv*, abs/2208.04284, 2022b.

V. N. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics, 2018.

Homer F. Walker and Peng Ni. Anderson acceleration for fixed-point iterations. *SIAM J. Numer. Anal.*, 49: 1715–1735, 2011.

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*, 45:3604–3616, 2022.

A. Xu and M. Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. In *Advances of Neural Information Processing Systems (NIPS)*, 2017.

A Appendix

B Proof of Lemma 8

Observe that

$$\begin{aligned}
\big|\tilde{Q}_{\alpha,\beta}(x)\big| &\le \frac{1}{\alpha\beta q^2}\mathbb{E}_{(a,b)^T\sim\mathcal{N}\left(0,\begin{bmatrix}1 & x\\ x & 1\end{bmatrix}\right)}\big|\varphi(\alpha a)\varphi(\beta b)\big| \\
&= \frac{1}{\alpha\beta q^2}\mathbb{E}_{(u,v)^T\sim\mathcal{N}\left(0,\begin{bmatrix}\alpha^2 & x\alpha\beta\\ x\alpha\beta & \beta^2\end{bmatrix}\right)}\big|\varphi(u)\varphi(v)\big| \\
&\le \frac{1}{\alpha\beta}\sqrt{\frac{1}{q^2}\mathbb{E}_{a\sim\mathcal{N}(0,\alpha^2)}[\varphi^2(a)]}\sqrt{\frac{1}{q^2}\mathbb{E}_{b\sim\mathcal{N}(0,\beta^2)}[\varphi^2(b)]} \qquad (42)\\
&= \sqrt{\tilde{Q}_{\alpha,\alpha}(1)\tilde{Q}_{\beta,\beta}(1)}, \qquad (43)
\end{aligned}$$

where (42) follows from Cauchy–Schwarz inequality.

In addition, by the L-bounded property of φ, we also have

$$|\varphi(\alpha z) - \varphi(0)| \le L|\alpha z|. \qquad (44)$$

Hence, for any $\alpha \ge 1$, it holds that

$$\begin{aligned}
|\varphi(\alpha z)| &\le |\varphi(0)| + L|\alpha||z| \\
&\le L\big(1 + |\alpha||z|\big) \\
&\le L|\alpha|\sqrt{2(1+z^2)}. \qquad (45)
\end{aligned}$$

From (45), we obtain

$$\begin{aligned}\mathbb{E}_{a\sim\mathcal{N}(0,\alpha^2)}[\varphi^2(a)] &= \int_{-\infty}^{\infty}\frac{1}{\alpha\sqrt{2\pi}}\varphi^2(z)\exp\Big(-\frac{z^2}{2\alpha^2}\Big)dz\\ &= \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}\varphi^2(\alpha z)\exp\Big(-\frac{z^2}{2}\Big)dz\\ &\le 2L^2\alpha^2\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}\big(1+z^2\big)\exp\Big(-\frac{z^2}{2}\Big)dz\\ &= 4L^2\alpha^2. \end{aligned} \tag{46}$$

Similarly, we also have

$$\mathbb{E}_{b\sim\mathcal{N}(0,\beta^2)}[\varphi^2(b)] \le 4L^2\beta^2. \tag{47}$$

From (42), (46) and (47), we obtain $|\tilde{Q}_{\alpha,\beta}(x)| \le 4L^2/q^2$ for all $\alpha \ge 1$, $\beta \ge 1$, and $x \in \mathbb{R}$.

Now, for a fixed pair $(\alpha \ge 1, \beta \ge 1)$, define $z := (u, v)$, $\phi(z) := \varphi(u)\varphi(v)$, and

$$\Sigma_x := \begin{bmatrix}\alpha^2 & x\alpha\beta\\ x\alpha\beta & \beta^2\end{bmatrix}. \tag{48}$$

Then, by (Daniely et al., 2016, Lemma 12) we have

$$\frac{\partial\tilde{Q}_{\alpha,\beta}}{\partial\Sigma_x} = -\frac{1}{2q^2\alpha\beta}\mathbb{E}_{(u,v)\sim\mathcal{N}(0,\Sigma_x)}\bigg[\frac{\partial\phi^2(z)}{\partial^2 z}(u,v)\bigg]. \tag{49}$$

On the other hand, we note that

$$\frac{\partial\phi^2(z)}{\partial^2 z}(u,v) = \begin{bmatrix}\frac{\partial^2\varphi(u)}{\partial u^2}\varphi(v) & \frac{\partial\varphi(u)}{\partial u}\frac{\partial\varphi(v)}{\partial v}\\ \frac{\partial\varphi(u)}{\partial u}\frac{\partial\varphi(v)}{\partial v} & \frac{\partial^2\varphi(v)}{\partial v^2}\varphi(u)\end{bmatrix}. \tag{50}$$

Hence, from (49) and (50) we have

$$\begin{aligned}\bigg\|\mathrm{vec}\Big(\frac{\partial\tilde{Q}_{\alpha,\beta}}{\partial\Sigma_x}\Big)\bigg\|_\infty \le \frac{1}{2q^2\alpha\beta}\max\bigg\{&\mathbb{E}_{(u,v)\sim\mathcal{N}(0,\Sigma_x)}\bigg[\bigg|\frac{\partial^2\varphi(u)}{\partial u^2}\varphi(v)\bigg|\bigg], \mathbb{E}_{(u,v)\sim\mathcal{N}(0,\Sigma_x)}\bigg[\bigg|\frac{\partial\varphi(u)}{\partial u}\frac{\partial\varphi(v)}{\partial v}\bigg|\bigg],\\ &\mathbb{E}_{(u,v)\sim\mathcal{N}(0,\Sigma_x)}\bigg[\bigg|\frac{\partial^2\varphi(v)}{\partial v^2}\varphi(u)\bigg|\bigg]\bigg\}.\end{aligned} \tag{51}$$

Hence, by the assumption that $\|\varphi\|_\infty \le L, \|\varphi''\|_\infty \le L$, from (51) we obtain

$$\bigg\|\mathrm{vec}\Big(\frac{\partial\tilde{Q}_{\alpha,\beta}}{\partial\Sigma_x}\Big)\bigg\|_\infty \le \frac{L^2}{2q^2\alpha\beta}. \tag{52}$$

It follows that

$$\begin{aligned}\big|\tilde{Q}_{\alpha,\beta}(y) - \tilde{Q}_{\alpha,\beta}(x)\big| &= \bigg|\int_x^y \frac{d\tilde{Q}_{\alpha,\beta}}{dt}dt\bigg|\\ &= \bigg|\int_x^y \mathrm{tr}\bigg(\Big(\frac{\partial\tilde{\mathrm{Q}}_{\alpha,\beta}}{\partial\Sigma_\mathrm{t}}\Big)^\mathrm{T}\frac{\partial\Sigma_\mathrm{t}}{\mathrm{dt}}\bigg)\mathrm{dt}\bigg|\\ &\le \int_x^y \bigg|\mathrm{tr}\bigg(\Big(\frac{\partial\tilde{\mathrm{Q}}_{\alpha,\beta}}{\partial\Sigma_\mathrm{t}}\Big)^\mathrm{T}\frac{\partial\Sigma_\mathrm{t}}{\mathrm{dt}}\bigg)\bigg|\mathrm{dt}\\ &= \int_x^y \bigg|\mathrm{vec}\Big(\frac{\partial\tilde{\mathrm{Q}}_{\alpha,\beta}}{\partial\Sigma_\mathrm{t}}\Big)^\mathrm{T}\mathrm{vec}\Big(\frac{\partial\Sigma_\mathrm{t}}{\mathrm{dt}}\Big)\bigg|\mathrm{dt}\end{aligned}$$

$$
\begin{aligned}
&\leq 4\int_x^y \left\|\mathrm{vec}\left(\frac{\partial \tilde{\mathrm{Q}}_{\alpha,\beta}}{\partial \Sigma_{\mathrm{t}}}\right)\right\|_\infty \left\|\mathrm{vec}\left(\frac{\partial \Sigma_{\mathrm{t}}}{\mathrm{dt}}\right)\right\|_\infty \mathrm{dt} \\
&\leq \frac{4L^2}{2q^2\alpha\beta}\int_x^y \left\|\mathrm{vec}\left(\frac{\partial \Sigma_{\mathrm{t}}}{\mathrm{dt}}\right)\right\|_\infty \mathrm{dt} \\
&= \frac{4L^2}{2q^2\alpha\beta}\alpha\beta|y-x| \\
&= \frac{2L^2}{q^2}|y-x|. \qquad (53)
\end{aligned}
$$

C Proof of Lemma 10

From (16) in Definition 5, we have

$$
\begin{aligned}
\nu_{ii}^{(l)} &= \frac{\sigma_w^2 \mathbf{K}_{ii}^{(l-1)} + d^{-1}\mathbf{x}_i^T\mathbf{x}_i}{\sigma_w^2 \mathbf{K}_{ii}^{(l-1)} + 1} \\
&= 1. \qquad (54)
\end{aligned}
$$

From (13) and (17) in Definition 5 and (54), we have

$$
\rho_{ii}^{(l)} = \sigma_w^2 \mathbf{K}_{ii}^{(l-1)} + 1. \qquad (55)
$$

In addition, from (14) and (16) in Definition 5 and (55), we also have

$$
\rho_{ij}^{(l)}\nu_{ij}^{(l)} = \sigma_w^2 \mathbf{K}_{ij}^{(l-1)} + d^{-1}\mathbf{x}_i^T\mathbf{x}_j, \qquad \forall i,j. \qquad (56)
$$

Replacing (17) in Definition 5 and (55) to (16) in Definition 5, we obtain for $i \neq j$,

$$
\begin{aligned}
|\nu_{ij}^{(l)}| &= \frac{\left|\sigma_w^2 \mathbf{K}_{ij}^{(l-1)} + d^{-1}\mathbf{x}_i^T\mathbf{x}_j\right|}{\sqrt{\left(\sigma_w^2 \mathbf{K}_{ii}^{(l-1)} + 1\right)\left(\sigma_w^2 \mathbf{K}_{jj}^{(l-1)} + 1\right)}} \\
&= \frac{\left|2q^2\sigma_w^2\rho_{ij}^{(l-1)}Q_{ij}\left(\nu_{ij}^{(l-1)}\right) + d^{-1}\mathbf{x}_i^T\mathbf{x}_j\right|}{\sqrt{\rho_{ii}^{(l)}\rho_{jj}^{(l)}}} \\
&= \frac{\left|Q_{ij}\left(\nu_{ij}^{(l-1)}\right)/\sqrt{Q_{ii}(1)Q_{jj}(1)}\sqrt{(2q^2\sigma_w^2\rho_{ii}^{(l-1)}Q_{ii}(1))(2q^2\sigma_w^2\rho_{jj}^{(l-1)}Q_{jj}(1))} + d^{-1}\mathbf{x}_i^T\mathbf{x}_j\right|}{\sqrt{\rho_{ii}^{(l)}\rho_{jj}^{(l)}}} \\
&= \frac{\left|Q_{ij}\left(\nu_{ij}^{(l-1)}\right)/\sqrt{Q_{ii}(1)Q_{jj}(1)}\sqrt{(\rho_{ii}^{(l)}-1)(\rho_{jj}^{(l)}-1)} + d^{-1}\mathbf{x}_i^T\mathbf{x}_j\right|}{\sqrt{\rho_{ii}^{(l)}\rho_{jj}^{(l)}}} \\
&\leq \frac{\sqrt{(\rho_{ii}^{(l)}-1)(\rho_{jj}^{(l)}-1)} + \left|d^{-1}\mathbf{x}_i^T\mathbf{x}_j\right|}{\sqrt{\rho_{ii}^{(l)}\rho_{jj}^{(l)}}} \qquad (57) \\
&\leq \frac{\sqrt{(\rho_{ii}^{(l)}-1)(\rho_{jj}^{(l)}-1)} + 1}{\sqrt{\rho_{ii}^{(l)}\rho_{jj}^{(l)}}} \qquad (58) \\
&\leq 1, \qquad (59)
\end{aligned}
$$

where (57) follows from Lemma 8, and (58) follows from $d^{-1}|\mathbf{x}_i^T\mathbf{x}_j| \leq d^{-1}\|\mathbf{x}_i\|_2\|\mathbf{x}_j\|_2 = 1$.

D Proof of Proposition 11

For all $i, j \in [n] \times [n]$, observe that

$$\begin{aligned}
&\big|\mathbf{K}_{ij}^{(l+1)} - \mathbf{K}_{ij}^{(l)}\big| \\
&\quad = 2q^2\big|\rho_{ij}^{(l+1)} Q_{ij}(\nu_{ij}^{(l+1)}) - \rho_{ij}^{(l)} Q_{ij}(\nu_{ij}^{(l)})\big| \\
&\quad \le 2q^2\big|\rho_{ij}^{(l+1)} Q_{ij}(\nu_{ij}^{(l+1)}) - \rho_{ij}^{(l+1)} Q_{ij}\big(\nu_{ij}^{(l)}\big)\big| + 2q^2\big|\rho_{ij}^{(l+1)} Q_{ij}\big(\nu_{ij}^{(l)}\big) - \rho_{ij}^{(l)} Q_{ij}(\nu_{ij}^{(l)})\big|, \qquad (60)
\end{aligned}$$

where (60) follows from the triangle inequality.

Now, we bound each term in (60). First, from Assumption 1 and Lemma 8, we have

$$2q^2\sigma_w^2 Q_{ii}(1) \le 8L^2\sigma_w^2 < 1. \qquad (61)$$

Therefore, from (13) we have

$$\rho_{ii}^{(l)} = \frac{1 - (2q^2\sigma_w^2 Q_{ii}(1))^{l+1}}{1 - 2q^2\sigma_w^2 Q_{ii}(1)}, \qquad \forall i. \qquad (62)$$

It follows that

$$\big|\rho_{ii}^{(l)} - \rho_{ii}^{(l+1)}\big| \le O\big((2q^2\sigma_w^2 Q_{ii}(1))^l\big). \qquad (63)$$

Hence, for $i \ne j$, we have

$$\begin{aligned}
\big|\rho_{ij}^{(l+1)} - \rho_{ij}^{(l)}\big| &= \left|\sqrt{\rho_{ii}^{(l+1)}\rho_{jj}^{(l+1)}} - \sqrt{\rho_{ii}^{(l)}\rho_{jj}^{(l)}}\right| \\
&\le \sqrt{\rho_{ii}^{(l+1)}}\left|\sqrt{\rho_{jj}^{(l+1)}} - \sqrt{\rho_{jj}^{(l)}}\right| + \sqrt{\rho_{jj}^{(l)}}\left|\sqrt{\rho_{ii}^{(l+1)}} - \sqrt{\rho_{ii}^{(l)}}\right| \\
&\le O\big((2q^2\sigma_w^2 Q_{ii}(1))^l\big) + O\big((2q^2\sigma_w^2 Q_{jj}(1))^l\big), \qquad (64)
\end{aligned}$$

where (64) follows from (62) and (63).

From (61) and (64), we obtain

$$\big|\rho_{ij}^{(l)} - \rho_{ij}^{(l+1)}\big| \le O\big((8L^2\sigma_w^2)^l\big), \qquad \forall i, j. \qquad (65)$$

Now, we have

$$\begin{aligned}
&\big|\rho_{ij}^{(l+1)} Q_{ij}(\nu_{ij}^{(l+1)}) - \rho_{ij}^{(l+1)} Q_{ij}\big(\nu_{ij}^{(l)}\big)\big| \\
&\quad = \Bigg|\rho_{ij}^{(l+1)}\tilde{Q}_{\sqrt{2\left(\frac{\sigma_w^2}{m}\mathbb{E}[\mathbf{G}_{ii}]+1\right)},\sqrt{2\left(\frac{\sigma_w^2}{m}\mathbb{E}[\mathbf{G}_{jj}]+1\right)}}(\nu_{ij}^{(l+1)}) \\
&\qquad\quad - \rho_{ij}^{(l+1)}\tilde{Q}_{\sqrt{2\left(\frac{\sigma_w^2}{m}\mathbb{E}[\mathbf{G}_{ii}]+1\right)},\sqrt{2\left(\frac{\sigma_w^2}{m}\mathbb{E}[\mathbf{G}_{jj}]+1\right)}}\big(\nu_{ij}^{(l)}\big)\Bigg| \\
&\quad \le \frac{2L^2}{q^2}\big|\rho_{ij}^{(l+1)}\nu_{ij}^{(l+1)} - \rho_{ij}^{(l+1)}\nu_{ij}^{(l)}\big| \qquad (66) \\
&\quad \le \frac{2L^2}{q^2}\big|\rho_{ij}^{(l+1)}\nu_{ij}^{(l+1)} - \rho_{ij}^{(l)}\nu_{ij}^{(l)}\big| + \frac{2L^2}{q^2}\big|\rho_{ij}^{(l)} - \rho_{ij}^{(l+1)}\big||\nu_{ij}^{(l)}| \\
&\quad \le \frac{2L^2}{q^2}\big|\rho_{ij}^{(l+1)}\nu_{ij}^{(l+1)} - \rho_{ij}^{(l)}\nu_{ij}^{(l)}\big| + \frac{2L^2}{q^2}\big|\rho_{ij}^{(l)} - \rho_{ij}^{(l+1)}\big| \qquad (67) \\
&\quad = \frac{2L^2}{q^2}\sigma_w^2\big|\mathbf{K}_{ij}^{(l)} - \mathbf{K}_{ij}^{(l-1)}\big| + \frac{2L^2}{q^2}O\big((8L^2\sigma_w^2)^l\big), \qquad (68)
\end{aligned}$$

where (66) follows from Lemma 8, (67) follows from Lemma 10, (68) follows from (23) in Lemma 10 and (65).

In addition, by using the fact that $|Q_{\alpha,\beta}(x)| \leq \frac{4L^2}{q^2}$ for all $\alpha \geq 1, \beta \geq 1$ in Lemma 8, we have

$$\begin{aligned}\big|\rho_{ij}^{(l+1)} Q_{ij}\big(\nu_{ij}^{(l)}\big) - \rho_{ij}^{(l)} Q_{ij}(\nu_{ij}^{(l)})\big| &\leq \frac{4L^2}{q^2}\big|\rho_{ij}^{(l+1)} - \rho_{ij}^{(l)}\big| \\ &= \frac{4L^2}{q^2} O\big((8L^2\sigma_w^2)^l\big), \end{aligned} \tag{69}$$

where (69) follows from (65).

From (17), (68), and (69) we have

$$\begin{aligned}&\big|\mathbf{K}_{ij}^{(l+1)} - \mathbf{K}_{ij}^{(l)}\big| \\ &\quad= 2q^2\big|\rho_{ij}^{(l+1)} Q_{ij}(\nu_{ij}^{(l+1)}) - \rho_{ij}^{(l)} Q_{ij}(\nu_{ij}^{(l)})\big| \\ &\quad\leq 2q^2\big|\rho_{ij}^{(l+1)} Q_{ij}(\nu_{ij}^{(l+1)}) - \rho_{ij}^{(l+1)} Q_{ij}(\nu_{ij}^{(l)})\big| + 2q^2\big|\rho_{ij}^{(l+1)} Q_{ij}(\nu_{ij}^{(l)}) - \rho_{ij}^{(l)} Q_{ij}(\nu_{ij}^{(l)})\big| \\ &\quad\leq 2q^2\left[\frac{2L^2}{q^2}\sigma_w^2\big|\mathbf{K}_{ij}^{(l)} - \mathbf{K}_{ij}^{(l-1)}\big| + \frac{2L^2}{q^2} O\big((8L^2\sigma_w^2)^l\big)\right] + 2q^2 \times \frac{4L^2}{q^2} O\big((8L^2\sigma_w^2)^l\big). \end{aligned} \tag{70}$$

By using induction, from (70) we have

$$\big|\mathbf{K}_{ij}^{(l+1)} - \mathbf{K}_{ij}^{(l)}\big| = O\big(\big(4L^2\sigma_w^2\big)^l\big). \tag{71}$$

Since $\sigma_w^2 < 1/(8L^2)$, $\{\mathbf{K}_{ij}^{(l)}\}_{l=1}^{\infty}$ can be easily shown to be a Cauchy sequence. From the completeness of $\mathbb{R}$, it holds that

$$\mathbf{K}_{ij}^{(l)} \to \mathbf{K}_{ij} \tag{72}$$

uniformly in $i, j \in [n] \times [n]$ as $l \to \infty$ for some matrix $\mathbf{K}$. By using the triangle inequality, we have

$$\big|\mathbf{K}_{ij}^{(l+1)} - \mathbf{K}_{ij}^{(l)}\big| \geq \big|\mathbf{K}_{ij}^{(l)} - \mathbf{K}_{ij}\big| - \big|\mathbf{K}_{ij}^{(l+1)} - \mathbf{K}_{ij}\big|. \tag{73}$$

From (71) and (73), we obtain

$$\big|\mathbf{K}_{ij}^{(l)} - \mathbf{K}_{ij}\big| = O\big(\big(8L^2\sigma_w^2\big)^l\big). \tag{74}$$

From (74), we obtain

$$\big\|\mathbf{K}^{(l)} - \mathbf{K}\big\|_F = O\big(n\big(8L^2\sigma_w^2\big)^l\big). \tag{75}$$

Now, by (17) and (72) we have

$$\mathbf{K}_{ij}^{(l)} = 2q^2\rho_{ij}^{(l)} Q_{ij}(\nu_{ij}^{(l)}) \tag{76}$$

and $\mathbf{K}_{ij}^{(l)} \to \mathbf{K}_{ij}$. On the other hand, by (61) we have $2q^2\sigma_w^2 Q_{ii}(1) < 1$. It follows from (62) that

$$\rho_{ii}^{(l)} \to \frac{1}{1 - 2q^2\sigma_w^2 Q_{ii}(1)} \tag{77}$$

as $l \to \infty$. Hence, it holds that $\nu_{ij}^{(l)} \to \nu_{ij}$ uniformly in $i, j \in [n] \times [n]$.

Hence, by Lemma 10, we have

$$\nu_{ij} = \begin{cases} \frac{Q_{ij}\big(\nu_{ij}\big)/\sqrt{Q_{ii}(1)Q_{jj}(1)}\sqrt{(\rho_{ii}-1)(\rho_{jj}-1)}+d^{-1}\mathbf{x}_i^T\mathbf{x}_j}{\sqrt{\rho_{ii}\rho_{jj}}}, & i \neq j \\ 1, & i = j \end{cases}, \tag{78}$$

where

$$\rho_{ii} = \frac{1}{1 - 2q^2\sigma_w^2 Q_{ii}(1)}. \tag{79}$$

E Proof of Proposition 12

Assume that $\mathbf{T}^{(l)} = [\mathbf{v}_1^{(l)}, \mathbf{v}_2^{(l)}, \cdots, \mathbf{v}_n^{(l)}]$ where $\mathbf{v}_i^{(l)} \in \mathbb{R}^m$ for all $i \in [n]$. By (1), we have

$$\mathbf{v}_i^{(l)} = \varphi\big(\mathbf{W}\mathbf{v}_i^{(l-1)} + \mathbf{U}\mathbf{x}_i\big), \qquad \forall i \in [n]. \tag{80}$$

Hence, with probability at least $1 - \exp\big(-\Omega(m)\big)$, we have

$$\begin{aligned}
\big\|\mathbf{v}_i^{(l+1)} - \mathbf{v}_i^{(l)}\big\|_2 &= \big\|\varphi\big(\mathbf{W}\mathbf{v}_i^{(l)} + \mathbf{U}\mathbf{x}_i\big) - \varphi\big(\mathbf{W}\mathbf{v}_i^{(l-1)} + \mathbf{U}\mathbf{x}_i\big)\big\|_2 \\
&\le L\big\|\mathbf{W}\big(\mathbf{v}_i^{(l)} - \mathbf{v}_i^{(l-1)}\big)\big\|_2 \qquad (81)\\
&\le L\big\|\mathbf{W}\big\|_2\big\|\mathbf{v}_i^{(l)} - \mathbf{v}_i^{(l-1)}\big\|_2 \\
&\le 2L\sqrt{2}\sigma_w\big\|\mathbf{v}_i^{(l)} - \mathbf{v}_i^{(l-1)}\big\|_2 \qquad (82)
\end{aligned}$$

where (81) is a consequence of the assumption that φ is L-bounded, and (82) follows from (Tao, 2012, Sect. (2.3)) .

Therefore, for all $l \ge 2$, it holds that

$$\begin{aligned}
\big\|\mathbf{v}_i^{(l)} - \mathbf{v}_i^{(l-1)}\big\|_2 &\le \big(2L\sqrt{2}\sigma_w\big)^l\big\|\mathbf{v}_i^{(1)} - \mathbf{v}_i^{(0)}\big\|_2 \\
&= \big(2L\sqrt{2}\sigma_w\big)^l\big\|\mathbf{v}_i^{(1)}\big\|_2 \\
&\le \big(2L\sqrt{2}\sigma_w\big)^l\sqrt{mL}, \qquad (83)
\end{aligned}$$

where (83) follows from the fact that

$$\big\|\mathbf{v}_i^{(1)}\big\|_2 = \big\|\varphi(\mathbf{U}\mathbf{x}_i)\big\|_2 \le \sqrt{mL}$$

by the L-boundedness of φ.

Then, for all $r > s$, with probability at least $1 - \exp(-\Omega(m))$, we have

$$\begin{aligned}
\big\|\mathbf{v}_i^{(r)} - \mathbf{v}_i^{(s)}\big\| &\le \sum_{l=s+1}^{r}\big\|\mathbf{v}_i^{(l)} - \mathbf{v}_i^{(l-1)}\big\|_2 \\
&\le \sqrt{mL}\sum_{l=s+1}^{r}\big(2L\sqrt{2}\sigma_w\big)^l \\
&\le \sqrt{mL}\big(2L\sqrt{2}\sigma_w\big)^{s+1}\frac{1}{1 - 2L\sqrt{2}\sigma_w} \to 0 \qquad (84)
\end{aligned}$$

as $s \to \infty$ since $2L\sqrt{2}\sigma_w < 1$. It follows that $\{\mathbf{v}_i^{(l)}\}_{l=1}^{\infty}$ is a Cauchy sequence. Since $\mathbb{R}$ is complete, hence we have

$$\|\mathbf{v}_i^{(l)} - \mathbf{v}_i\| \to 0 \tag{85}$$

for some vector $\mathbf{v}_i$.

Therefore, we have

$$\begin{aligned}
\big\|\mathbf{v}_i^{(l-1)} - \mathbf{v}_i\big\| - \big\|\mathbf{v}_i^{(l)} - \mathbf{v}_i\big\| &\le \big\|\mathbf{v}_i^{(l)} - \mathbf{v}_i^{(l-1)}\big\| \\
&\le \sqrt{mL}\big(2L\sqrt{2}\sigma_w\big)^l, \qquad \forall l \ge 2. \qquad (86)
\end{aligned}$$

From (86), with probability at least $1-\exp(-\Omega(m))$ we have

$$\begin{aligned}\big\|\mathbf{v}_i^{(l)}-\mathbf{v}_i\big\| &\le \sqrt{mL}\big\|\sum_{k=l+1}^{\infty}\big(2L\sqrt{2}\sigma_w\big)^k\\ &= \sqrt{mL}\frac{\big(2L\sqrt{2}\sigma_w\big)^{l+1}}{1-2L\sqrt{2}\sigma_w}.\end{aligned} \tag{87}$$

Consequently, we have

$$\begin{aligned}\big|\mathbf{G}_{ij}-\mathbf{G}_{ij}^{(l)}\big| &= \big|\mathbf{v}_i^T\mathbf{v}_j-\big(\mathbf{v}_i^{(l)}\big)^T\big(\mathbf{v}_j^{(l)}\big)\big|\\ &\le \big|\mathbf{v}_i^T\mathbf{v}_j-\mathbf{v}_i^T\big(\mathbf{v}_j^{(l)}\big)\big|+\big|\mathbf{v}_i^T\big(\mathbf{v}_j^{(l)}\big)-\big(\mathbf{v}_i^{(l)}\big)^T\big(\mathbf{v}_j^{(l)}\big)\big|\\ &\le \big\|\mathbf{v}_i\big\|\big\|\mathbf{v}_j-\mathbf{v}_j^{(l)}\big\|+\big\|\mathbf{v}_j^{(l)}\big\|\big\|\mathbf{v}_i-\mathbf{v}_i^{(l)}\big\|\\ &\le \big\|\mathbf{v}_i\big\|\sqrt{mL}\frac{\big(2L\sqrt{2}\sigma_w\big)^{l+1}}{1-2L\sqrt{2}\sigma_w}\\ &\quad+\big\|\mathbf{v}_j^{(l)}\big\|\sqrt{mL}\frac{\big(2L\sqrt{2}\sigma_w\big)^{l+1}}{1-2L\sqrt{2}\sigma_w}\\ &\le 2mL\frac{\big(2L\sqrt{2}\sigma_w\big)^{l+1}}{1-2L\sqrt{2}\sigma_w},\end{aligned} \tag{88}$$

where (88) follows from the fact that $\|\mathbf{v}_i\|\le\sqrt{mL}$ and $\|\mathbf{v}_j^{(l)}\|\le\sqrt{mL}$ by the L-boundedness of φ. From (88) we obtain

$$\frac{1}{m}\big|\mathbf{G}_{ij}-\mathbf{G}_{ij}^{(l)}\big|\le 2L\frac{\big(2L\sqrt{2}\sigma_w\big)^{l+1}}{1-2L\sqrt{2}\sigma_w}. \tag{89}$$

Finally, we obtain (29) from (89).

F Proof of Proposition 13

Define

$$\hat{\mathbf{G}}_{ij}^{(l)} := \mathbb{E}\bigg[\frac{1}{m}\mathbf{G}_{ij}^{(l)}\bigg|\mathbf{h}_l,\mathbf{h}_l'\bigg]. \tag{90}$$

Then, by Lemma 9, we have

$$\begin{aligned}\hat{\mathbf{G}}_{ij}^{(l)} &= \mathbb{E}\bigg[\frac{1}{m}\varphi(\mathbf{M}\mathbf{h}_l)^T\varphi(\mathbf{M}\mathbf{h}_l')\bigg|\mathbf{h}_l,\mathbf{h}_l'\bigg]\\ &= \mathbb{E}_{\mathbf{w}\sim\mathcal{N}(0,2\mathbf{I})}\big[\varphi(\mathbf{w}^T\mathbf{h}_l)\varphi(\mathbf{w}^T\mathbf{h}_l')\big].\end{aligned} \tag{91}$$

Let

$$\hat{\mathbf{A}}_{ij}^{(l)} := \mathbf{h}_l^T\mathbf{h}_l', \qquad \hat{\mathbf{A}}_{ii}^{(l)} := \|\mathbf{h}_l\|_2^2, \qquad \hat{\mathbf{A}}_{jj}^{(l)} := \|\mathbf{h}_l'\|_2^2, \tag{92}$$

and define

$$\hat{\nu}_{ij}^{(l)} := \frac{\hat{\mathbf{A}}_{ij}^{(l)}}{\sqrt{\hat{\mathbf{A}}_{ii}^{(l)}\hat{\mathbf{A}}_{jj}^{(l)}}}. \tag{93}$$

Then, we have

$$\begin{aligned}\hat{\mathbf{G}}_{ij}^{(l)} &= \mathbb{E}_{(u,v)\sim\mathcal{N}\left(0,2\begin{bmatrix}\|\mathbf{h}_l\|^2 & \mathbf{h}_l^T\mathbf{h}_l' \\ \mathbf{h}_l^T\mathbf{h}_l' & \|\mathbf{h}_l'\|^2\end{bmatrix}\right)}\big[\varphi(u)\varphi(v)\big] \\ &= \mathbb{E}_{(u,v)\sim\mathcal{N}\left(0,\begin{bmatrix}1 & \frac{\mathbf{h}_l^T\mathbf{h}_l'}{\|\mathbf{h}_l\|\|\mathbf{h}_l'\|} \\ \frac{\mathbf{h}_l^T\mathbf{h}_l'}{\|\mathbf{h}_l\|\|\mathbf{h}_l'\|} & 1\end{bmatrix}\right)}\big[\varphi(\sqrt{2}\|\mathbf{h}_l\|u)\varphi(\sqrt{2}\|\mathbf{h}_l'\|v)\big] \\ &= 2q^2\|\mathbf{h}_l\|\|\mathbf{h}_l'\|\tilde{Q}_{\sqrt{2}\|\mathbf{h}_l\|,\sqrt{2}\|\mathbf{h}_l'\|}(\hat{\nu}_{ij}^{(l)}) \\ &= 2q^2\sqrt{\hat{\mathbf{A}}_{ii}^{(l)}\hat{\mathbf{A}}_{jj}^{(l)}}\tilde{Q}_{\sqrt{2}\|\mathbf{h}_l\|,\sqrt{2}\|\mathbf{h}_l'\|}(\hat{\nu}_{ij}^{(l)}). \end{aligned} \tag{94}$$

Now, we consider two cases:

- **Case 1:** $i = j$**.**

By Lemma 9, we have

$$\mathbf{G}_{ii}^{(l+1)} = \varphi(\mathbf{M}\mathbf{h}_{l+1})^T\varphi(\mathbf{M}\mathbf{h}_{l+1}), \tag{95}$$

where

$$\|\mathbf{h}_{l+1}\|^2 = \frac{\sigma_w^2}{m}\mathbf{G}_{ii}^{(l)} + 1. \tag{96}$$

Now, for a fixed $\mathbf{h}_{l+1}$, by Beinstein's inequality and (95), it holds with probability $1 - \exp(-\Omega(m\varepsilon^2))$ that

$$\left|\frac{1}{m}\mathbf{G}_{ii}^{(l+1)} - \hat{\mathbf{G}}_{ii}^{(l+1)}\right| \le \varepsilon/2. \tag{97}$$

On the other hand, since φ is L-bounded, it holds from (95) that $\mathbf{G}_{ii}^{(l)} \in [0, mL^2]$ for all $l \ge 1$. Hence, from (96) we have

$$1 \le \|\mathbf{h}_{l+1}\|^2 \le \sigma_w^2 L^2 + 1. \tag{98}$$

This means that the ε-net size for $\mathbf{h}_{l+1}$ is at most $\exp\big\{O\big(l\log\frac{1}{\varepsilon}\big)\big\}$. Therefore, with probability at least $1 - n^2\exp\big(-\Omega(m\varepsilon^2) + O(l\log\frac{1}{\varepsilon})\big)$ we have

$$\left|\frac{1}{m}\mathbf{G}_{ii}^{(l+1)} - \hat{\mathbf{G}}_{ii}^{(l+1)}\right| \le \varepsilon/2. \tag{99}$$

Now, observe that

$$\begin{aligned}\hat{\mathbf{G}}_{ii}^{(l+1)} &= \mathbb{E}_{\mathbf{w}\sim\mathcal{N}(0,2\mathbf{I})}\big[\varphi^2(\mathbf{w}^T\mathbf{h}_{l+1})\big] \\ &= \mathbb{E}_{u\sim\mathcal{N}(0,2\|\mathbf{h}_{l+1}\|_2^2)}[\varphi^2(u)] \\ &= \mathbb{E}_{u\sim\mathcal{N}(0,1)}[\varphi^2(\sqrt{2}\|\mathbf{h}_{l+1}\|_2 u)] \\ &= 2q^2\|\mathbf{h}_{l+1}\|_2^2\tilde{Q}_{\sqrt{2}\|\mathbf{h}_{l+1}\|,\sqrt{2}\|\mathbf{h}_{l+1}\|}(1). \end{aligned} \tag{100}$$

On the other hand, we also have

$$\begin{aligned}\mathbf{K}_{ii}^{(l+1)} &= 2q^2\rho_{ii}^{(l+1)}Q_{ii}(1) \\ &= 2q^2(\sigma_w^2\mathbf{K}_{ii}^{(l)} + 1)Q_{ii}(1), \end{aligned} \tag{101}$$

where (101) follows from (17) and Lemma 10, and (101) follows from Lemma 10.

It follows that

$$\begin{aligned}
&\left|\hat{\mathbf{G}}_{ii}^{(l+1)} - \mathbf{K}_{ii}^{(l+1)}\right| \\
&\quad= 2q^2\left|\|\mathbf{h}_{l+1}\|_2^2\tilde{Q}_{\sqrt{2}\|\mathbf{h}_{l+1}\|_2,\sqrt{2}\|\mathbf{h}_{l+1}\|_2}(1) - \big(\sigma_w^2\mathbf{K}_{ii}^{(l)}+1\big)Q_{ii}(1)\right| \\
&\quad= 2q^2\left|\left(\frac{\sigma_w^2}{m}\mathbf{G}_{ii}^{(l)}+1\right)\tilde{Q}_{\sqrt{2}\|\mathbf{h}_{l+1}\|_2,\sqrt{2}\|\mathbf{h}_{l+1}\|_2}(1) - \big(\sigma_w^2\mathbf{K}_{ii}^{(l)}+1\big)Q_{ii}(1)\right| \\
&\quad\le 2q^2\left|\left(\frac{\sigma_w^2}{m}\mathbf{G}_{ii}^{(l)}+1\right)\tilde{Q}_{\sqrt{2}\|\mathbf{h}_{l+1}\|_2,\sqrt{2}\|\mathbf{h}_{l+1}\|_2}(1) - \big(\sigma_w^2\mathbf{K}_{ii}^{(l)}+1\big)\tilde{Q}_{\sqrt{2}\|\mathbf{h}_{l+1}\|,\sqrt{2}\|\mathbf{h}_{l+1}\|}(1)\right| \\
&\qquad\qquad + 2q^2\big(\sigma_w^2\mathbf{K}_{ii}^{(l)}+1\big)\left|\tilde{Q}_{\sqrt{2}\|\mathbf{h}_{l+1}\|_2,\sqrt{2}\|\mathbf{h}_{l+1}\|_2}(1) - Q_{ii}(1)\right| \\
&\quad\le 2q^2\sigma_w^2\left|\frac{\mathbf{G}_{ii}^{(l)}}{m} - \mathbf{K}_{ii}^{(l)}\right|\left|\tilde{Q}_{\sqrt{2}\|\mathbf{h}_{l+1}\|_2,\sqrt{2}\|\mathbf{h}_{l+1}\|_2}(1)\right| \\
&\qquad\qquad + 2q^2\big(\sigma_w^2\mathbf{K}_{ii}^{(l)}+1\big)\left|\tilde{Q}_{\sqrt{2}\|\mathbf{h}_{l+1}\|_2,\sqrt{2}\|\mathbf{h}_{l+1}\|_2}(1) - Q_{ii}(1)\right| \\
&\quad\le 8L^2\sigma_w^2\left|\frac{\mathbf{G}_{ii}^{(l)}}{m} - \mathbf{K}_{ii}^{(l)}\right| + 2q^2\big(\sigma_w^2\mathbf{K}_{ii}^{(l)}+1\big)\left|\tilde{Q}_{\sqrt{2}\|\mathbf{h}_{l+1}\|,\sqrt{2}\|\mathbf{h}_{l+1}\|}(1) - Q_{ii}(1)\right|, \qquad (102)
\end{aligned}$$

where (102) follows from Lemma 8.

Now, let

$$\|\mathbf{h}\|_2^2 := \frac{\sigma_w^2}{m}\mathbf{G}_{ii} + 1. \qquad (103)$$

Then, we have

$$\begin{aligned}
\big|\|\mathbf{h}_{l+1}\|_2^2 - \|\mathbf{h}\|_2^2\big| &= \frac{\sigma_w^2}{m}\big|\mathbf{G}_{ii}^{(l)} - \mathbf{G}_{ii}\big| \qquad (104)\\
&= O\Big(\big(2L\sqrt{2}\sigma_w\big)^l\Big) \qquad (105)
\end{aligned}$$

where (104) follows from (96) and (103), and (105) follows from (89).

Since $\|\mathbf{h}\|, \|\mathbf{h}_{l+1}\| \in [1, \sigma_w^2L^2+1]$, from (105) we obtain

$$\big|\|\mathbf{h}_{l+1}\| - \|\mathbf{h}\|\big| = O\Big(n\big(2L\sqrt{2}\sigma_w\big)^l\Big). \qquad (106)$$

On the other hand, by (89) it holds with probability at least $1-\exp(-\Omega(m)-\Omega(m\varepsilon^2))$ that

$$\left|\frac{1}{m}\mathbf{G}_{ii}^{(l+1)} - \mathbf{G}_{ii}\right| = O\Big(\big(2L\sqrt{2}\sigma_w\big)^{l+1}\Big). \qquad (107)$$

Hence, from (97) and (107) with probability at least $1-\exp(-\Omega(m)-\Omega(m\varepsilon^2))$, we have

$$\left|\frac{\mathbf{G}_{ii}}{m} - \mathbb{E}\left[\frac{\mathbf{G}_{ii}^{(l+1)}}{m}\middle|\mathbf{h}_{l+1}\right]\right| \le \frac{\varepsilon}{2} + O\Big(\big(2L\sqrt{2}\sigma_w\big)^{l+1}\Big) \qquad (108)$$

for any fixed $\mathbf{h}_{l+1}$. Now, observe that

$$\mathbb{E}\left[\frac{\mathbf{G}_{ii}^{(l+1)}}{m}\middle|\mathbf{h}_{l+1}\right] = \mathbb{E}\left[\frac{\big\|\varphi(\mathbf{M}\mathbf{h}_{l+1})\big\|^2}{m}\middle|\mathbf{h}_{l+1}\right]$$

is a fixed function of $\mathbf{h}_{l+1}$. Hence, there exists, $\hat{\mathbf{h}}_{l+1}$ such that

$$\hat{\mathbf{h}}_{l+1} = \arg\max_{\mathbf{h}_{l+1}} \left| \frac{\mathbf{G}_{ii}}{m} - \mathbb{E}\left[\frac{\mathbf{G}_{ii}^{(l+1)}}{m} \middle| \mathbf{h}_{l+1} \right] \right|. \tag{109}$$

Then, with probability at least $1 - \exp(-\Omega(m) - \Omega(m\varepsilon^2))$ it holds that

$$\begin{aligned} \left| \frac{\mathbf{G}_{ii}}{m} - \mathbb{E}\left[\frac{\mathbf{G}_{ii}^{(l+1)}}{m} \right] \right| &\le \mathbb{E}\left[\left| \frac{\mathbf{G}_{ii}}{m} - \mathbb{E}\left[\frac{\mathbf{G}_{ii}^{(l+1)}}{m} \middle| \mathbf{h}_{l+1} \right] \right| \right] \\ &\le \left| \frac{\mathbf{G}_{ii}}{m} - \mathbb{E}\left[\frac{\mathbf{G}_{ii}^{(l+1)}}{m} \middle| \hat{\mathbf{h}}_{l+1} \right] \right| \\ &\le \frac{\varepsilon}{2} + O\Big(\big(2L\sqrt{2}\sigma_w \big)^{l+1} \Big). \end{aligned} \tag{110}$$

Hence, by taking $l \to \infty$, with probability $1 - \exp(-\Omega(m) - \Omega(m\varepsilon^2))$, it holds that

$$\big| \|\mathbf{h}\|^2 - \mathbb{E}[\|\mathbf{h}\|^2] \big| \le \varepsilon, \tag{111}$$

$$\big| \|\mathbf{h}\| - \mathbb{E}[\|\mathbf{h}\|] \big| \le \varepsilon, \tag{112}$$

where (112) follows from (111) and the fact that $\|\mathbf{h}\|^2 \in [1, \sigma_w^2 L^2 + 1]$ for any $\mathbf{h}$.

From (105), (106), (111), and (112), with probability at least $1 - \exp(-\Omega(m) - \Omega(m\varepsilon^2))$ it holds that

$$\big| \|\mathbf{h}_{l+1}\|^2 - \mathbb{E}[\|\mathbf{h}\|^2] \big| = \varepsilon + O\Big(\big(2L\sqrt{2}\sigma_w \big)^{l} \Big), \tag{113}$$

$$\big| \|\mathbf{h}_{l+1}\| - \mathbb{E}[\|\mathbf{h}\|] \big| = \varepsilon + O\Big(\big(2L\sqrt{2}\sigma_w \big)^{l} \Big). \tag{114}$$

Now, for any $a \in \mathbb{R}$ note that

$$\begin{aligned} &\Big| \varphi^2(\sqrt{2}\|\mathbf{h}_{l+1}\|a) - \varphi^2(\sqrt{2}\|\mathbf{h}\|a) \Big| \\ &\quad = \Big| \varphi(\sqrt{2}\|\mathbf{h}_{l+1}\|a) - \varphi(\sqrt{2}\|\mathbf{h}\|a) \Big| \Big| \sigma(\sqrt{2}\|\mathbf{h}_{l+1}\|a) + \sigma(\sqrt{2}\|\mathbf{h}\|a) \Big|. \end{aligned} \tag{115}$$

On the other hand, we have

$$\Big| \varphi(\sqrt{2}\|\mathbf{h}_{l+1}\|a) - \varphi(\sqrt{2}\|\mathbf{h}\|a) \Big| \le L\sqrt{2}|a| \big| \|\mathbf{h}_{l+1}\| - \|\mathbf{h}\| \big|, \tag{116}$$

$$\Big| \varphi(\sqrt{2}\|\mathbf{h}_{l+1}\|a) + \varphi(\sqrt{2}\|\mathbf{h}\|a) \Big| \le 2L, \tag{117}$$

where we use the assumption that φ is L-bounded on (117).

From (115), (116), and (117), we obtain

$$\begin{aligned} \Big| \varphi^2(\sqrt{2}\|\mathbf{h}_{l+1}\|a) - \varphi^2(\sqrt{2}\|\mathbf{h}\|a) \Big| &\le 2L^2\sqrt{2}|a| \big| \|\mathbf{h}_{l+1}\| - \|\mathbf{h}\| \big| \\ &= 2L^2\sqrt{2}|a| \Big[\varepsilon + O\Big(\big(2L\sqrt{2}\sigma_w \big)^{l} \Big) \Big], \end{aligned} \tag{118}$$

where (118) follows from (105) and (106).

From (118), we obtain

$$\begin{aligned}
&\left|\mathbb{E}_{a\sim\mathcal{N}(0,1)}\left[\varphi^2(\sqrt{2}\|\mathbf{h}_{l+1}\|a)\right]-\mathbb{E}_{a\sim\mathcal{N}(0,1)}\left[\varphi^2(\sqrt{2}\|\mathbf{h}\|a)\right]\right| \\
&\quad\le 2L^2\sqrt{2}\mathbb{E}_{a\sim\mathcal{N}(0,1)}[|a|]\big[\varepsilon+O\Big(\big(2L\sqrt{2}\sigma_w\big)^l\Big)\big] \\
&\quad= 2L^2\sqrt{2}O\Big(\varepsilon+\big(2L\sqrt{2}\sigma_w\big)^l\Big). \qquad (119)
\end{aligned}$$

Similarly, by the assumption that φ is L-bounded, we also have

$$\mathbb{E}_{a\sim\mathcal{N}(0,1)}\left[\varphi^2(\sqrt{2}\|\mathbf{h}\|a)\right]\le L^2. \qquad (120)$$

It follows that

$$\begin{aligned}
&\left|\tilde{Q}_{\sqrt{2}\|\mathbf{h}_{l+1}\|,\sqrt{2}\|\mathbf{h}_{l+1}\|}(1)-Q_{ii}(1)\right| \\
&\quad=\left|\frac{1}{2q^2\|\mathbf{h}_{l+1}\|^2}\mathbb{E}_{a\sim\mathcal{N}(0,1)}\left[\varphi^2(\sqrt{2}\|\mathbf{h}_{l+1}\|a)\right]-\frac{1}{2q^2\mathbb{E}[\|\mathbf{h}\|^2]}\mathbb{E}_{a\sim\mathcal{N}(0,1)}\left[\varphi^2(\sqrt{2}\mathbb{E}[\|\mathbf{h}\|]a)\right]\right| \\
&\quad\le\left|\frac{1}{2q^2\|\mathbf{h}_{l+1}\|^2}\mathbb{E}_{a\sim\mathcal{N}(0,1)}\left[\varphi^2(\sqrt{2}\|\mathbf{h}_{l+1}\|a)\right]-\frac{1}{2q^2\|\mathbf{h}_{l+1}\|^2}\mathbb{E}_{a\sim\mathcal{N}(0,1)}\left[\varphi^2(\sqrt{2}\mathbb{E}[\|\mathbf{h}\|]a)\right]\right| \\
&\quad+\left|\frac{1}{2q^2\|\mathbf{h}_{l+1}\|^2}\mathbb{E}_{a\sim\mathcal{N}(0,1)}\left[\varphi^2(\sqrt{2}\mathbb{E}[\|\mathbf{h}\|]a)\right]-\frac{1}{2q^2\mathbb{E}[\|\mathbf{h}\|^2]}\mathbb{E}_{a\sim\mathcal{N}(0,1)}\left[\varphi^2(\sqrt{2}\mathbb{E}[\|\mathbf{h}\|]a)\right]\right| \\
&\quad\le\frac{1}{2q^2\|\mathbf{h}_{l+1}\|^2}\left|\mathbb{E}_{a\sim\mathcal{N}(0,1)}\left[\varphi^2(\sqrt{2}\|\mathbf{h}_{l+1}\|a)\right]-\mathbb{E}_{a\sim\mathcal{N}(0,1)}\left[\varphi^2(\sqrt{2}\mathbb{E}[\|\mathbf{h}\|]a)\right]\right| \\
&\quad+\frac{1}{2q^2}\left|\frac{1}{\|\mathbf{h}_{l+1}\|^2}-\frac{1}{\mathbb{E}[\|\mathbf{h}\|^2]}\right|\mathbb{E}_{a\sim\mathcal{N}(0,1)}\left[\varphi^2(\sqrt{2}\mathbb{E}[\|\mathbf{h}\|]a)\right]. \qquad (121)
\end{aligned}$$

By combining (105), (119), and (120), from (121), we obtain

$$\left|\tilde{Q}_{\sqrt{2}\|\mathbf{h}_{l+1}\|,\sqrt{2}\|\mathbf{h}_{l+1}\|}(1)-Q_{i,i}(1)\right|=2L^2O\Big(\varepsilon+\big(2L\sqrt{2}\sigma_w\big)^l\Big) \qquad (122)$$

since $\|\mathbf{h}_{l+1}\|,\mathbb{E}[\|\mathbf{h}\|]\in[1,\sigma_w^2L^2+1]$.

On the other hand, by (74) and the assumption $2L\sqrt{2}\sigma_w<1$, we have

$$\|\mathbf{K}_{ii}^{(l+1)}-\mathbf{K}_{ii}\|=O\Big((2L\sqrt{2}\sigma_w)^{l+1}\Big). \qquad (123)$$

From (122), (123), by setting

$$\varepsilon:=O\Big(\big(2L\sqrt{2}\sigma_w\big)^{l+1}\Big) \qquad (124)$$

from (102), we obtain

$$\left|\hat{\mathbf{G}}_{ii}^{(l+1)}-\mathbf{K}_{ii}^{(l+1)}\right|\le 8L^2\sigma_w^2\left|\frac{\mathbf{G}_{ii}^{(l)}}{m}-\mathbf{K}_{ii}^{(l)}\right|+2L^2O\Big(\big(2L\sqrt{2}\sigma_w\big)^{l+1}\Big). \qquad (125)$$

It follows from (99) and (125) that with probability at least $1-\exp\big\{-\Omega(8^lL^{2l}\sigma_w^{2l}m)+O(l^2)\big\}$,

$$\begin{aligned}
\left|\frac{1}{m}\mathbf{G}_{ii}^{(l+1)}-\mathbf{K}_{ii}^{(l+1)}\right|&\le\left|\frac{1}{m}\mathbf{G}_{ii}^{(l+1)}-\hat{\mathbf{G}}_{ii}^{(l+1)}\right|+\left|\hat{\mathbf{G}}_{ii}^{(l+1)}-\mathbf{K}_{ii}^{(l+1)}\right| \\
&\le 8L^2\sigma_w^2\left|\frac{1}{m}\mathbf{G}_{ii}^{(l)}-\mathbf{K}_{ii}^{(l)}\right|+2L^2O\Big(\big(2L\sqrt{2}\sigma_w\big)^{l+1}\Big),
\end{aligned}$$

which implies that with probability at least $1 - l\exp\big\{-\Omega(8^l L^{2l}\sigma_w^{2l} m) + O(l^2)\big\}$, we have

$$\left|\frac{1}{m}\mathbf{G}_{ii}^{(l)} - \mathbf{K}_{ii}^{(l)}\right| = O\bigg(\Big(2L\sqrt{2}\sigma_w\Big)^{l+1}\bigg). \tag{126}$$

- **Case 2:** $i \neq j$.

For this case, let

$$\|\mathbf{h}\|^2 := \frac{\sigma_w^2}{m}\mathbf{G}_{ii} + 1, \tag{127}$$

$$\|\mathbf{h}'\|^2 := \frac{\sigma_w^2}{m}\mathbf{G}_{jj} + 1. \tag{128}$$

By (89), with probability at least $1 - \exp(-\Omega(m))$, we have

$$\frac{1}{m}\big|\mathbf{G}_{ii} - \mathbf{G}_{ii}^{(l)}\big| = O\bigg(\Big(2L\sqrt{2}\sigma_w\Big)^{l}\bigg). \tag{129}$$

In addition, we also have

$$\|\mathbf{h}_{l+1}\|^2 = \frac{\sigma_w^2}{m}\mathbf{G}_{ii}^{(l)} + 1 \geq 1, \tag{130}$$

$$\|\mathbf{h}'_{l+1}\|^2 = \frac{\sigma_w^2}{m}\mathbf{G}_{jj}^{(l)} + 1 \geq 1. \tag{131}$$

Hence, we have

$$\begin{aligned}
\big|\|\mathbf{h}_{l+1}\| - \|\mathbf{h}\|\big| &= O\Big(\big|\|\mathbf{h}_{l+1}\|^2 - \|\mathbf{h}\|^2\big|\Big) \\
&= \frac{\sigma_w^2}{m}\left\|\mathbf{G}_{ii}^{(l)} - \mathbf{G}_{ii}\right\| \\
&= O\bigg(\Big(2L\sqrt{2}\sigma_w\Big)^{l}\bigg).
\end{aligned} \tag{132}$$

Then, it holds that

$$\begin{aligned}
&\left|\hat{\mathbf{G}}_{ij}^{(l+1)} - \mathbf{K}_{ij}^{(l+1)}\right| \\
&\quad= 2q^2\left|\sqrt{\hat{\mathbf{A}}_{ii}^{(l+1)}\hat{\mathbf{A}}_{jj}^{(l+1)}}\tilde{Q}_{\sqrt{2}\|\mathbf{h}_{l+1}\|,\sqrt{2}\|\mathbf{h}'_{l+1}\|}(\hat{\nu}_{ij}^{(l+1)}) - \rho_{ij}^{(l+1)}Q_{ij}(\nu_{ij}^{(l+1)})\right| \\
&\quad\leq 2q^2\left|\sqrt{\hat{\mathbf{A}}_{ii}^{(l+1)}\hat{\mathbf{A}}_{jj}^{(l+1)}}\tilde{Q}_{\sqrt{2}\|\mathbf{h}_{l+1}\|,\sqrt{2}\|\mathbf{h}'_{l+1}\|}(\hat{\nu}_{ij}^{(l+1)}) - \rho_{ij}^{(l+1)}\tilde{Q}_{\sqrt{2}\|\mathbf{h}_{l+1}\|,\sqrt{2}\|\mathbf{h}'_{l+1}\|}(\nu_{ij}^{(l+1)})\right| \\
&\qquad+ 2q^2\rho_{ij}^{(l+1)}\left|\tilde{Q}_{\sqrt{2}\|\mathbf{h}_{l+1}\|,\sqrt{2}\|\mathbf{h}'_{l+1}\|}(\nu_{ij}^{(l+1)}) - Q_{ij}(\nu_{ij}^{(l+1)})\right|.
\end{aligned} \tag{133}$$

Now, for all $|x| \leq 1$, we have

$$\begin{aligned}
\left|\tilde{Q}_{\sqrt{2}\|\mathbf{h}_{l+1}\|,\sqrt{2}\|\mathbf{h}'_{l+1}\|}(x) - Q_{ij}(x)\right| &\leq \left|\tilde{Q}_{\sqrt{2}\|\mathbf{h}_{l+1}\|,\sqrt{2}\|\mathbf{h}'_{l+1}\|}(x) - \tilde{Q}_{\sqrt{2}\mathbb{E}[\|\mathbf{h}\|],\sqrt{2}\|\mathbf{h}'_{l+1}\|}(x)\right| \\
&\quad+ \left|\tilde{Q}_{\sqrt{2}\mathbb{E}[\|\mathbf{h}\|],\sqrt{2}\|\mathbf{h}'_{l+1}\|}(x) - Q_{ij}(x)\right|.
\end{aligned} \tag{134}$$

On the other hand, we have

$$
\begin{aligned}
&\left|\tilde{Q}_{\sqrt{2}\mathbb{E}[\|\mathbf{h}\|],\sqrt{2}\|\mathbf{h}'_{l+1}\|}(x) - Q_{ij}(x)\right| \\
&= \left|\frac{1}{2q^2\mathbb{E}[\|\mathbf{h}\|]\|\mathbf{h}'_{l+1}\|}\mathbb{E}_{(a,b)^T\sim\mathcal{N}\left(0,\begin{bmatrix}1 & x\\ x & 1\end{bmatrix}\right)}\varphi(\sqrt{2}\mathbb{E}[\|\mathbf{h}\|]a)\varphi(\sqrt{2}\|\mathbf{h}'_{l+1}\|b)\right. \\
&\quad \left. - \frac{1}{2q^2\mathbb{E}[\|\mathbf{h}\|]\mathbb{E}[\|\mathbf{h}'\|]}\mathbb{E}_{(a,b)^T\sim\mathcal{N}\left(0,\begin{bmatrix}1 & x\\ x & 1\end{bmatrix}\right)}\varphi(\sqrt{2}\mathbb{E}[\|\mathbf{h}\|]a)\varphi(\sqrt{2}\mathbb{E}[\|\mathbf{h}'\|]b)\right| \\
&\le \left|\frac{1}{2q^2\mathbb{E}[\|\mathbf{h}\|]\|\mathbf{h}'_{l+1}\|}\mathbb{E}_{(a,b)^T\sim\mathcal{N}\left(0,\begin{bmatrix}1 & x\\ x & 1\end{bmatrix}\right)}\varphi(\sqrt{2}\mathbb{E}[\|\mathbf{h}\|]a)\varphi(\sqrt{2}\|\mathbf{h}'_{l+1}\|b)\right. \\
&\quad \left. - \frac{1}{2q^2\mathbb{E}[\|\mathbf{h}\|]\|\mathbf{h}'_{l+1}\|}\mathbb{E}_{(a,b)^T\sim\mathcal{N}\left(0,\begin{bmatrix}1 & x\\ x & 1\end{bmatrix}\right)}\varphi(\sqrt{2}\mathbb{E}[\|\mathbf{h}\|]a)\varphi(\sqrt{2}\mathbb{E}[\|\mathbf{h}'\|]b)\right| \\
&\quad + \left|\frac{1}{2q^2\mathbb{E}[\|\mathbf{h}\|]\|\mathbf{h}'_{l+1}\|}\mathbb{E}_{(a,b)^T\sim\mathcal{N}\left(0,\begin{bmatrix}1 & x\\ x & 1\end{bmatrix}\right)}\varphi(\sqrt{2}\mathbb{E}[\|\mathbf{h}\|]a)\varphi(\sqrt{2}\mathbb{E}[\|\mathbf{h}'\|]b)\right. \\
&\quad \left. - \frac{1}{2q^2\mathbb{E}[\|\mathbf{h}\|]\mathbb{E}[\|\mathbf{h}'\|]}\mathbb{E}_{(a,b)^T\sim\mathcal{N}\left(0,\begin{bmatrix}1 & x\\ x & 1\end{bmatrix}\right)}\varphi(\sqrt{2}\mathbb{E}[\|\mathbf{h}\|]a)\varphi(\sqrt{2}\mathbb{E}[\|\mathbf{h}'\|]b)\right| \\
&\le \frac{1}{2q^2\mathbb{E}[\|\mathbf{h}\|]\|\mathbf{h}'_{l+1}\|}\mathbb{E}_{(a,b)^T\sim\mathcal{N}\left(0,\begin{bmatrix}1 & x\\ x & 1\end{bmatrix}\right)}\left|\varphi(\sqrt{2}\mathbb{E}[\|\mathbf{h}\|]a)\varphi(\sqrt{2}\|\mathbf{h}'_{l+1}\|b)\right. \\
&\qquad \left. - \varphi(\sqrt{2}\mathbb{E}[\|\mathbf{h}\|]a)\varphi(\sqrt{2}\mathbb{E}[\|\mathbf{h}'\|]b)\right| \\
&\quad + \frac{1}{2q^2\mathbb{E}[\|\mathbf{h}\|]}\left|\frac{1}{\|\mathbf{h}'_{l+1}\|} - \frac{1}{\mathbb{E}[\|\mathbf{h}'\|]}\right|\mathbb{E}_{(a,b)^T\sim\mathcal{N}\left(0,\begin{bmatrix}1 & x\\ x & 1\end{bmatrix}\right)}\left|\varphi(\sqrt{2}\mathbb{E}[\|\mathbf{h}\|]a)\varphi(\sqrt{2}\mathbb{E}[\|\mathbf{h}'\|]b)\right|. \qquad (135)
\end{aligned}
$$

In addition, by the assumption that φ is L-bounded we have

$$|\varphi(\sqrt{2}\mathbb{E}[\|\mathbf{h}\|]a)| \le L \qquad (136)$$

$$|\varphi(\sqrt{2}\mathbb{E}[\|\mathbf{h}'\|]b)| \le L. \qquad (137)$$

It follows that

$$
\begin{aligned}
&\left|\varphi(\sqrt{2}\mathbb{E}[\|\mathbf{h}\|]a)\varphi(\sqrt{2}\|\mathbf{h}'_{l+1}\|b) - \varphi(\sqrt{2}\mathbb{E}[\|\mathbf{h}\|]a)\varphi(\sqrt{2}\mathbb{E}[\|\mathbf{h}'\|]b)\right| \\
&\quad = \left|\varphi(\sqrt{2}\mathbb{E}[\|\mathbf{h}\|]a)\right|\left|\varphi(\sqrt{2}\|\mathbf{h}'_{l+1}\|b) - \varphi(\sqrt{2}\mathbb{E}[\|\mathbf{h}'\|]b)\right| \\
&\quad \le L\left|\varphi(\sqrt{2}\|\mathbf{h}'_{l+1}\|b) - \varphi(\sqrt{2}\mathbb{E}[\|\mathbf{h}'\|]b)\right| \\
&\quad \le L^2\sqrt{2}|b|\left|\|\mathbf{h}'_{l+1}\| - \mathbb{E}[\|\mathbf{h}'\|]\right|. \qquad (138)
\end{aligned}
$$

On the other hand, by (112), with probability at least $1-\exp(-\Omega(m)-\Omega(m\varepsilon^2))$, it holds that

$$\left|\|\mathbf{h}'\| - \mathbb{E}[\|\mathbf{h}'\|]\right| \le \varepsilon. \qquad (139)$$

From (132) and (139), we have

$$\big|\|\mathbf{h}'_{l+1}\| - \mathbb{E}[\|\mathbf{h}'\|]\big| \le \big|\|\mathbf{h}'_{l+1}\| - \|\mathbf{h}'\|\big| + \big|\|\mathbf{h}'\| - \mathbb{E}[\|\mathbf{h}'\|]\big| \tag{140}$$

$$\le \varepsilon + O\bigg(\Big(2L\sqrt{2}\sigma_w\Big)^l\bigg). \tag{141}$$

Now, by setting

$$\varepsilon := O\bigg(\Big(2\sqrt{2}\sigma_w\Big)^l\bigg), \tag{142}$$

from (141), we obtain

$$\big|\|\mathbf{h}'_{l+1}\| - \mathbb{E}[\|\mathbf{h}'\|]\big| = O\bigg(\Big(2L\sqrt{2}\sigma_w\Big)^l\bigg). \tag{143}$$

Similarly, we also have

$$\big|\|\mathbf{h}_{l+1}\| - \mathbb{E}[\|\mathbf{h}\|]\big| = O\bigg(\Big(2L\sqrt{2}\sigma_w\Big)^l\bigg). \tag{144}$$

From (135), (138), (143) and (144), we obtain

$$\bigg|\tilde{Q}_{\sqrt{2}\mathbb{E}[\|\mathbf{h}\|],\sqrt{2}\|\mathbf{h}'_{l+1}\|}(x) - Q_{ij}(x)\bigg| = O\bigg(\Big(2L\sqrt{2}\sigma_w\Big)^l\bigg), \qquad \forall x : |x| \le 1. \tag{145}$$

Similarly, we can prove that

$$\bigg|\tilde{Q}_{\sqrt{2}\|\mathbf{h}_{l+1}\|,\sqrt{2}\|\mathbf{h}'_{l+1}\|}(x) - \tilde{Q}_{\sqrt{2}\mathbb{E}[\|\mathbf{h}\|],\sqrt{2}\|\mathbf{h}'_{l+1}\|}(x)\bigg| = O\bigg(\Big(2L\sqrt{2}\sigma_w\Big)^l\bigg), \qquad \forall x : |x| \le 1. \tag{146}$$

From (134), (145), and (146), we obtain

$$\bigg|\tilde{Q}_{\sqrt{2}\|\mathbf{h}_{l+1}\|,\sqrt{2}\|\mathbf{h}'_{l+1}\|}(x) - Q_{ij}(x)\bigg| \le O\bigg(\Big(2L\sqrt{2}\sigma_w\Big)^l\bigg), \qquad \forall x : |x| \le 1. \tag{147}$$

Next, we aim to upper bound

$$2q^2\big|\sqrt{\hat{\mathbf{A}}^{(l+1)}_{ii}\hat{\mathbf{A}}^{(l+1)}_{jj}}\tilde{Q}_{\sqrt{2}\|\mathbf{h}_{l+1}\|,\sqrt{2}\|\mathbf{h}'_{l+1}\|}(\hat{\nu}^{(l+1)}_{ij}) - \rho^{(l+1)}_{ij}\tilde{Q}_{\sqrt{2}\|\mathbf{h}_{l+1}\|,\sqrt{2}\|\mathbf{h}'_{l+1}\|}(\nu^{(l+1)}_{ij})\big|.$$

Observe that with probability at least $1 - n^2\exp(-\Omega(m))$, it holds for all l sufficiently large that

$$\begin{aligned}
&\big|\sqrt{\hat{\mathbf{A}}^{(l+1)}_{ii}\hat{\mathbf{A}}^{(l+1)}_{jj}}\tilde{Q}_{\sqrt{2}\|\mathbf{h}_{l+1}\|,\sqrt{2}\|\mathbf{h}'_{l+1}\|}(\hat{\nu}^{(l+1)}_{ij}) - \rho^{(l+1)}_{ij}\tilde{Q}_{\sqrt{2}\|\mathbf{h}_{l+1}\|,\sqrt{2}\|\mathbf{h}'_{l+1}\|}(\nu^{(l+1)}_{ij})\big| \\
&\le \bigg|\bigg(\sqrt{\hat{\mathbf{A}}^{(l+1)}_{ii}\hat{\mathbf{A}}^{(l+1)}_{jj}} - \rho^{(l+1)}_{ij}\bigg)\tilde{Q}_{\sqrt{2}\|\mathbf{h}_{l+1}\|,\sqrt{2}\|\mathbf{h}'_{l+1}\|}(\hat{\nu}^{(l+1)}_{ij})\bigg| \\
&\quad + \bigg|\rho^{(l+1)}_{ij}\bigg(\tilde{Q}_{\sqrt{2}\|\mathbf{h}_{l+1}\|,\sqrt{2}\|\mathbf{h}'_{l+1}\|}(\hat{\nu}^{(l+1)}_{ij}) - \tilde{Q}_{\sqrt{2}\|\mathbf{h}_{l+1}\|,\sqrt{2}\|\mathbf{h}'_{l+1}\|}(\nu^{(l+1)}_{ij})\bigg)\bigg| \\
&\le \frac{4L^2}{q^2}\bigg|\sqrt{\hat{\mathbf{A}}^{(l+1)}_{ii}\hat{\mathbf{A}}^{(l+1)}_{jj}} - \rho^{(l+1)}_{ij}\bigg| + \rho^{(l+1)}_{ij}\frac{2L^2}{q^2}\big|\hat{\nu}^{(l+1)}_{ij} - \nu^{(l+1)}_{ij}\big| \qquad (148) \\
&\le \frac{4L^2}{q^2}\bigg[\bigg|\sqrt{\hat{\mathbf{A}}^{(l+1)}_{ii}\hat{\mathbf{A}}^{(l+1)}_{jj}} - \rho^{(l+1)}_{ij}\bigg| + \rho^{(l+1)}_{ij}\big|\hat{\nu}^{(l+1)}_{ij} - \nu^{(l+1)}_{ij}\big|\bigg], \qquad (149)
\end{aligned}$$

where (148) follows from Lemma 8.

On the other hand, we have

$$\begin{aligned}
&\left|\sqrt{\hat{\mathbf{A}}_{ii}^{(l+1)}\hat{\mathbf{A}}_{jj}^{(l+1)}}-\rho_{ij}^{(l+1)}\right|+\rho_{ij}^{(l+1)}\left|\hat{\nu}_{ij}^{(l+1)}-\nu_{ij}^{(l+1)}\right| \\
&\quad=\left|\sqrt{\hat{\mathbf{A}}_{ii}^{(l+1)}\hat{\mathbf{A}}_{jj}^{(l+1)}}-\rho_{ij}^{(l+1)}\right| \\
&\qquad+\left|\left(\sqrt{\hat{\mathbf{A}}_{ii}^{(l+1)}\hat{\mathbf{A}}_{jj}^{(l+1)}}+\rho_{ij}^{(l+1)}-\sqrt{\hat{\mathbf{A}}_{ii}^{(l+1)}\hat{\mathbf{A}}_{jj}^{(l+1)}}\right)\hat{\nu}_{ij}^{(l+1)}-\rho_{ij}^{(l+1)}\nu_{ij}^{(l+1)}\right| \\
&\quad\leq 2\left|\sqrt{\hat{\mathbf{A}}_{ii}^{(l+1)}\hat{\mathbf{A}}_{jj}^{(l+1)}}-\rho_{ij}^{(l+1)}\right|+\left|\sqrt{\hat{\mathbf{A}}_{ii}^{(l+1)}\hat{\mathbf{A}}_{jj}^{(l+1)}}\hat{\nu}_{ij}^{(l+1)}-\rho_{ij}^{(l+1)}\nu_{ij}^{(l+1)}\right|, \qquad (150)
\end{aligned}$$

where (150) follows from $|\hat{\nu}_{ij}^{(l)}|\leq 1$.

On the other hand, since $\rho_{ij}^{(l+1)}=\sqrt{\rho_{ii}^{(l+1)}\rho_{jj}^{(l+1)}}$, we also have

$$\begin{aligned}
&\big|\sqrt{\hat{\mathbf{A}}_{ii}^{(l+1)}\hat{\mathbf{A}}_{jj}^{(l+1)}}-\rho_{ij}^{(l+1)}\big| \\
&\quad=\left|\sqrt{\left(\frac{\sigma_w^2}{m}\mathbf{G}_{ii}^{(l)}+1\right)\left(\frac{\sigma_w^2}{m}\mathbf{G}_{jj}^{(l)}+1\right)}-\sqrt{\big(\sigma_w^2\mathbf{K}_{ii}^{(l)}+1\big)\big(\sigma_w^2\mathbf{K}_{jj}^{(l)}+1\big)}\right| \\
&\quad=O\Big(\big(2L\sqrt{2}\sigma_w\big)^l\Big), \qquad (151)
\end{aligned}$$

where (151) follows from (126) and the fact that $\frac{\mathbf{G}_{ii}^{(l)}}{m}\in[0,L^2]$.

Moreover, note that

$$\begin{aligned}
\sqrt{\hat{\mathbf{A}}_{ii}^{(l+1)}\hat{\mathbf{A}}_{jj}^{(l+1)}}\hat{\nu}_{ij}^{(l+1)}&=\hat{\mathbf{A}}_{ij}^{(l+1)} \\
&=\mathbf{h}_{l+1}^T\mathbf{h}_{l+1}' \\
&=\frac{\sigma_w^2}{m}\mathbf{G}_{ij}^{(l)}+\frac{1}{d}\mathbf{x}_i^T\mathbf{x}_j \qquad (152)
\end{aligned}$$

and

$$\begin{aligned}
\rho_{ij}^{(l+1)}\nu_{ij}^{(l+1)}&=\nu_{ij}^{(l+1)}\sqrt{\rho_{ii}^{(l+1)}\rho_{jj}^{(l+1)}} \\
&=\nu_{ij}^{(l+1)}\sqrt{\big(\sigma_w^2\mathbf{K}_{ii}^{(l)}+1\big)\big(\sigma_w^2\mathbf{K}_{jj}^{(l)}+1\big)} \\
&=\sigma_w^2\mathbf{K}_{ij}^{(l)}+\frac{1}{d}\mathbf{x}_i^T\mathbf{x}_j. \qquad (153)
\end{aligned}$$

Thus, it holds that

$$\left|\sqrt{\hat{\mathbf{A}}_{ii}^{(l+1)}\hat{\mathbf{A}}_{jj}^{(l+1)}}\hat{\nu}_{ij}^{(l+1)}-\rho_{ij}^{(l+1)}\nu_{ij}^{(l+1)}\right|=\sigma_w^2\left|\frac{1}{m}\mathbf{G}_{ij}^{(l)}-\mathbf{K}_{ij}^{(l)}\right|. \qquad (154)$$

Thus, with probability at least $1-l\exp\big(-\Omega(m\varepsilon^2)+O\big(l\log 1/\varepsilon\big)\big)$, it holds that

$$\left|\hat{\mathbf{G}}_{ij}^{(l+1)}-\mathbf{K}_{ij}^{(l+1)}\right|\leq 8L^2\sigma_w^2\left|\frac{1}{m}\mathbf{G}_{ij}^{(l)}-\mathbf{K}_{ij}^{(l)}\right|+O\Big(\big(2L\sqrt{2}\sigma_w\big)^{l+1}\Big). \qquad (155)$$

On the other hand, by Lemma 9, we have

$$\mathbf{G}_{ij}^{(l+1)}=\varphi(\mathbf{M}\mathbf{h}_{l+1})^T\varphi(\mathbf{M}\mathbf{h}_{l+1}'). \qquad (156)$$

Hence, for a fixed vector pair $\mathbf{h}_{l+1}, \mathbf{h}'_{l+1}$, by Beinstein's inequality, with probability at least $1-\exp(-\Omega(m\varepsilon^2))$ it holds that

$$\left|\frac{1}{m}\mathbf{G}_{ij}^{(l+1)} - \hat{\mathbf{G}}_{ij}^{(l+1)}\right| \leq \varepsilon. \tag{157}$$

Then, by using ε-net arguments as in Case 1, with probability at least $1 - l\exp\big(-\Omega(m\varepsilon^2) + O\big(l\log 1/\varepsilon\big)\big)$, we have

$$\left|\frac{1}{m}\mathbf{G}_{ij}^{(l+1)} - \hat{\mathbf{G}}_{ij}^{(l+1)}\right| \leq \varepsilon. \tag{158}$$

Consequently, we have

$$\begin{aligned}\left|\frac{1}{m}\mathbf{G}_{ij}^{(l+1)} - \mathbf{K}_{ij}^{(l+1)}\right| &\leq \left|\frac{1}{m}\mathbf{G}_{ij}^{(l+1)} - \hat{\mathbf{G}}_{ij}^{(l+1)}\right| + \left|\hat{\mathbf{G}}_{ij}^{(l+1)} - \mathbf{K}_{ij}^{(l+1)}\right| \\ &\leq 2O\bigg(\big(2L\sqrt{2}\sigma_w\big)^{l+1}\bigg) + 8L^2\sigma_w^2\left|\frac{1}{m}\mathbf{G}_{ij}^{(l)} - \mathbf{K}_{ij}^{(l)}\right| \end{aligned} \tag{159}$$

where (159) follows from (155) and (158) and the choice of ε in (142).

By applying the induction argument, one can show that for $l \geq 1$, it holds with probability at least $1 - l\exp\big(-\Omega(m\varepsilon^2) + O\big(l\log 1/\varepsilon\big)\big)$, we have

$$\left|\frac{1}{m}\mathbf{G}_{ij}^{(l)} - \mathbf{K}_{ij}^{(l)}\right| \leq \frac{2O\bigg(\big(2L\sqrt{2}\sigma_w\big)^{l+1}\bigg)}{1 - 8L^2\sigma_w^2}. \tag{160}$$

By the choice of ε in (142), it holds that with probability at least $1 - l\exp\big\{-\Omega(8^l L^{2l}\sigma_w^{2l}m) + O(l^2)\big\}$, we have

$$\left|\frac{1}{m}\mathbf{G}_{ij}^{(l)} - \mathbf{K}_{ij}^{(l)}\right| = O\bigg((2L\sqrt{2}\sigma_w)^l\bigg). \tag{161}$$

Finally, from (126) and (162) with probability at least $1 - n^2 l\exp\big\{-\Omega(8^l L^{2l}\sigma_w^{2l}m) + O(l^2)\big\}$, it holds that

$$\left\|\frac{1}{m}\mathbf{G}^{(l)} - \mathbf{K}^{(l)}\right\|_F = O\bigg(n(2L\sqrt{2}\sigma_w)^l\bigg). \tag{162}$$

G Proof of Theorem 7

Since $\mathbf{U}\mathbf{x}_i$ is a Gaussian vector with zero-mean and variance depending on $\|\mathbf{x}_i\|^2$. On the other hand, by the Assumption 2, $\|\mathbf{x}_i\| = \sqrt{d}$. Hence, from $\mathbf{t}_i = \varphi(\mathbf{W}\mathbf{t}_i + \mathbf{U}\mathbf{x}_i)$, it is easy to see that $\mathbb{E}[\mathbf{G}_{ii}] = \mathbb{E}[\|\mathbf{t}_i\|^2]$ does not depend on $i \in [n]$. This means that $\mathbb{E}[\mathbf{G}_{ii}] = \mathbb{E}[\mathbf{G}_{jj}]$ for all $i, j \in [n] \times [n]$. On the other hand, we have $\mathbb{E}[\|\mathbf{t}_i\|^2] = \mathbb{E}[\|\varphi(\mathbf{W}\mathbf{t}_i + \mathbf{U}\mathbf{x}_i)\|_2^2] \in [0, mL^2]$ by the L-boundedness of the function φ. Hence, we have

$$0 \leq \frac{\mathbb{E}[\mathbf{G}_{ii}]}{m} \leq L^2, \qquad \forall i \in [m]. \tag{163}$$

It follows that $Q_{ij}(x)$ has the form $\tilde{Q}_{\alpha,\alpha}(x)$ for some $\alpha \in [2, 2(\sigma_w^2 L^2 + 1)]$.

Thanks to this fact, from Proposition 11 and the assumption on this theorem, for all $(i, j) \in [n] \times [n]$, it holds that

$$\begin{aligned}\mathbf{K}_{ij} &= 2q^2 Q_{ij}(\nu_{ij})\sqrt{\rho_{ii}\rho_{jj}} \\ &= 2q^2\sqrt{\rho_{ii}\rho_{jj}}\sum_{r=0}^{\infty}\mu_{r,\alpha}^2(\varphi)\nu_{ij}^r,\end{aligned}$$

where

$$\nu_{ij} = \frac{Q_{ij}\big(\nu_{ij}\big)/\sqrt{Q_{ii}(1)Q_{jj}(1)}\sqrt{(\rho_{ii}-1)(\rho_{jj}-1)} + d^{-1}\mathbf{x}_i^T\mathbf{x}_j}{\sqrt{\rho_{ii}\rho_{jj}}}. \tag{164}$$

Here,

$$\rho_{ii} = \frac{1}{1-2q^2\sigma_w^2 Q_{ii}(1)}. \tag{165}$$

Now, by Lemma 10, we have $|\nu_{ij}| \leq 1$ for all $(i,j) \in [n]\times[n]$. Let $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_n]$ where $\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_n$ be unit vectors such that $\nu_{ij} = \mathbf{h}_i^T\mathbf{h}_j$ for all $(i,j) \in [n]\times[n]$. It is easy to check that $[(\mathbf{H}^T\mathbf{H})^{\odot r}]_{ij} = (\mathbf{h}_i^T\mathbf{h}_j)^r$ holds for all $(i,j) \in [n]\times[n]$. Let $\tilde{\mathbf{K}}$ be a $n\times n$ matrix such that

$$\tilde{\mathbf{K}}_{ij} = \mathbf{K}_{ij}/\sqrt{\rho_{ii}\rho_{jj}}, \qquad \forall i,j \in [n]\times[n]. \tag{166}$$

Then, $\tilde{\mathbf{K}}$ can be written as

$$\tilde{\mathbf{K}} = 2q^2\sum_{r=0}^{\infty}\mu_{r,\alpha}^2(\varphi)\big(\mathbf{H}^T\mathbf{H}\big)^{(\odot r)}. \tag{167}$$

Now, for any unit vector $\mathbf{u} = [u_1, u_2, \cdots, u_n]^T \in \mathbb{R}^n$, it holds that

$$\begin{aligned}\mathbf{u}^T\big(\mathbf{H}^T\mathbf{H}\big)^{(\odot r)}\mathbf{u} &= \sum_{i,j} u_i u_j (\mathbf{h}_i^T\mathbf{h}_j)^r \\ &= \sum_i u_i^2 + \sum_{i\neq j} u_i u_j \nu_{ij}^r \\ &= 1 + \sum_{i\neq j} u_i u_j \nu_{ij}^r. \end{aligned} \tag{168}$$

Next, we show that $|\nu_{ij}| < 1$ if $i \neq j$. Indeed, assume that there exists $i \neq j$ such that $|\nu_{ij}| \geq 1$. Then, from (30) in Lemma 10, we have

$$\begin{aligned} 1 &\leq |\nu_{ij}| \\ &= \left|\frac{Q_{ij}\big(\nu_{ij}\big)/\sqrt{Q_{ii}(1)Q_{jj}(1)}\sqrt{(\rho_{ii}-1)(\rho_{jj}-1)} + d^{-1}\mathbf{x}_i^T\mathbf{x}_j}{\sqrt{\rho_{ii}\rho_{jj}}}\right| \\ &\leq \frac{\sqrt{(\rho_{ii}-1)(\rho_{jj}-1)} + \big|d^{-1}\mathbf{x}_i^T\mathbf{x}_j\big|}{\sqrt{\rho_{ii}\rho_{jj}}} \qquad (169) \\ &< \frac{\sqrt{(\rho_{ii}-1)(\rho_{jj}-1)} + 1}{\sqrt{\rho_{ii}\rho_{jj}}} \qquad (170) \\ &\leq 1, \qquad (171) \end{aligned}$$

where (169) follows from Lemma 8, and (170) follows by the fact that since $\mathbf{x}_i \nparallel \mathbf{x}_j$, from Cauchy–Schwarz inequality and Assumption 2, we have $\mathbf{x}_i^T\mathbf{x}_j < \|\mathbf{x}_i\|_2\|\mathbf{x}_j\| = d$. This is a contradiction. Hence, we have $|\beta| < 1$ where

$$\beta := \max_{i\neq j}|\nu_{ij}|. \tag{172}$$

Now, by taking $r > -\frac{\log(2n)}{\log \beta}$, we have

$$\begin{aligned}\left|\sum_{i\neq j} u_i u_j \nu_{ij}^r\right| &\leq \sum_{i\neq j} |u_i||u_j|\beta^r \\ &\leq \left(\sum_i |\nu_i|\right)^2 \beta^r k \\ &\leq n\beta^r \\ &< \frac{1}{2}. \end{aligned} \tag{173}$$

From (168) and (173), we obtain

$$\mathbf{u}^T \left(\mathbf{H}^T\mathbf{H}\right)^{(\odot r)} \mathbf{u} > \frac{1}{2}, \qquad \forall \mathbf{u},$$

so $\left(\mathbf{H}^T\mathbf{H}\right)^{(\odot r)}$ is positive definite. Following Theorem 7, it holds that $\min_{\alpha\in[2,2(\sigma_w^2 L^2+1)]} \mu_{r,\alpha}^2(\varphi) > 0$ for infinitely many values of r. Hence, $\tilde{\mathbf{K}}$ is positive definite for all initialisations since $0 \leq \frac{\mathbb{E}[\mathbf{G}_{ii}]}{m} \leq L^2$.

Now, let $\mathbf{\Gamma} = \{\sqrt{\rho_{ii}\rho_{jj}}\}_{i,j}$ be an $n \times n$ matrix where the (i,j) element is $\sqrt{\rho_{ii}\rho_{jj}}$. Then, we have

$$\mathbf{K} = \tilde{\mathbf{K}} \odot \mathbf{\Gamma}. \tag{174}$$

Now, for any vector $\mathbf{u} = [u_1, u_2, \cdots, u_n]^T$, we have

$$\begin{aligned}\mathbf{u}^T\mathbf{\Gamma}\mathbf{u} &= \sum_{i,j} u_i u_j \sqrt{\rho_{ii}\rho_{jj}} \\ &= \left(\sum_i u_i \sqrt{\rho_{ii}}\right)^2 \\ &\geq 0. \end{aligned} \tag{175}$$

Hence, $\mathbf{\Gamma}$ is positive semi-definite. Now, by applying (Ling et al., 2022, Lemma 6), we have

$$\begin{aligned}\lambda_{\min}(\mathbf{K}) &\geq \left(\min_i \rho_{ii}\right) \lambda_{\min}(\tilde{\mathbf{K}}) \\ &\geq \lambda_{\min}(\tilde{\mathbf{K}}) \\ &\geq \min_{\frac{\mathbb{E}[\mathbf{G}_{ii}]}{m} \in [0, L^2], \forall i} \lambda_{\min}(\tilde{\mathbf{K}}) := \lambda_0^* > 0, \end{aligned}$$

so $\mathbf{K}$ is positive definite with the smallest eigenvalue $\lambda_* \geq \lambda_0^* > 0$, where λ_0^* is some constant which does not depend on m.

H Proof of Theorem 3

The following proof follows the same steps as (Ling et al., 2022, Proof of Theorem 1). There are some changes caused by the new activation function. First, we recall the two important auxiliary lemmas:

Lemma 20. *(Horn & Johnson, 1985, Sect. 5.8) Let $\mathbf{\Delta} = \mathbf{B} - \mathbf{A}$ where $\mathbf{A}$ and $\mathbf{B}$ are square complex matrices. Then, it holds that*

$$\|\mathbf{B}^{-1}\| \leq \frac{\|\mathbf{A}^{-1}\|}{1 - \|\mathbf{A}^{-1}\mathbf{\Delta}\|}. \tag{176}$$

Lemma 21. *(Weyl's inequality)(Ling et al., 2022, Lemma 5) Let $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m\times n}$ with their singular values satisfying $\sigma_1(\mathbf{A}) \geq \sigma_2(\mathbf{A}) \geq \cdots \geq \sigma_r(\mathbf{A})$ and $\sigma_1(\mathbf{B}) \geq \sigma_2(\mathbf{B}) \geq \cdots \geq \sigma_r(\mathbf{B})$ and $r = \min(m,n)$. Then,*

$$\max_{i\in[r]} \left|\sigma_i(\mathbf{A}) - \sigma_i(\mathbf{B})\right\| \leq \|\mathbf{A} - \mathbf{B}\|. \tag{177}$$

The equilibrium point of Eq. (2) is the root of the function $F(\tau) := \mathbf{T}(\tau) - \varphi(\mathbf{W}(\tau)\mathbf{T}(\tau) + \mathbf{U}(\tau)\mathbf{X}) = 0$. Let $\mathbf{J}(\tau) := \partial \mathrm{vec}(\mathrm{F}(\tau))/\partial \mathrm{vec}(\mathbf{T}(\tau))$ denote the Jacobian matrix. Then, it is easy to see that

$$\mathbf{J}(\tau) = \mathbf{I}_{mn} - \mathbf{D}(\tau)\big(\mathbf{I}_n \otimes \mathbf{W}(\tau)\big),$$

where $\mathbf{D}(\tau) := \mathrm{diag}[\mathrm{vec}(\varphi'(\mathbf{W}(\tau)\mathbf{T}(\tau) + \mathbf{U}(\tau)\mathbf{X}))]$. Using the Lipschitz property of activation function, it is easy to check that $\mathbf{J}(\tau)$ is invertible if $\|\mathbf{W}(\tau)\| < 1/L$. The gradient of each trainable parameter is given by the following lemma.

Lemma 22. *(Ling et al., 2022, Lemma 2) If $\mathbf{J}(\tau)$ is invertible, the gradient of the objective function $\Phi(\tau)$ w.r.t. each trainable parameters is given by*

$$\begin{aligned}
\mathrm{vec}(\nabla_{\mathbf{W}}\Phi(\tau)) &= (\mathbf{T}(\tau) \otimes \mathbf{I}_m)\mathbf{R}(\tau)^T(\hat{\mathbf{y}}(\tau) - \mathbf{y}) \\
\mathrm{vec}(\nabla_{\mathbf{U}}\Phi(\tau)) &= (\mathbf{X} \otimes \mathbf{I}_m)\mathbf{R}(\tau)^T(\hat{\mathbf{y}}(\tau) - \mathbf{y}), \\
\nabla_{\mathbf{a}}\Phi(\tau) &= \mathbf{T}(\tau)(\hat{\mathbf{y}}(\tau) - \mathbf{y})
\end{aligned}$$

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

Based on these three lemmas, we can prove the following result:

Lemma 23. *For each $s \in [0, \tau]$, suppose that $\|\mathbf{W}(s)\|_2 \le \bar{\rho}_w$, $\|\mathbf{U}(s)\|_2 \le \bar{\rho}_u$ and $\|\mathbf{a}(s)\|_2 \le \bar{\rho}_a$. It holds that*

$$\|\mathbf{T}(s)\|_F \le c_a\|\mathbf{X}\|_F + c_m \tag{178}$$

and

$$\|\nabla_{\mathbf{W}}\Phi(s)\|_F \le c_u\big(c_a\|\mathbf{X}\|_F + c_m\big)\|\hat{\mathbf{y}}(s) - \mathbf{y}\|_2, \tag{179}$$

$$\|\nabla_U\Phi(s)\|_F \le c_u\|\mathbf{X}\|_F\|\hat{\mathbf{y}}(s) - \mathbf{y}\|_2, \tag{180}$$

$$\|\nabla_a\Phi(s)\|_F \le \big(c_a\|\mathbf{X}\|_F + c_m\big)\|\hat{\mathbf{y}}(s) - \mathbf{y}\|_2. \tag{181}$$

Furthermore, for each $k, s \in [0, \tau]$, it holds that

$$\begin{aligned}
\|\mathbf{T}(k) - \mathbf{T}(s)\| \le{}& \frac{L}{1 - L\bar{\rho}_w}\big(c_a\|\mathbf{X}\|_F + c_m\big)\|\mathbf{W}(k) - \mathbf{W}(s)\|_2 \\
&+ \frac{L}{1 - L\bar{\rho}_w}\|\mathbf{U}(k) - \mathbf{U}(s)\|_2\|\mathbf{X}\|_F
\end{aligned} \tag{182}$$

and

$$\begin{aligned}
&\|\hat{\mathbf{y}}(k) - \hat{\mathbf{y}}(s)\|_2 \\
&\quad\le \bar{\rho}_a\bigg[\frac{L}{1 - L\bar{\rho}_w}\big(c_a\|\mathbf{X}\|_F + c_m\big)\|\mathbf{W}(k) - \mathbf{W}(s)\|_2 \\
&\qquad + \frac{L}{1 - L\bar{\rho}_w}\|\mathbf{U}(k) - \mathbf{U}(s)\|_2\|\mathbf{X}\|_F\bigg] + \big(c_a\|\mathbf{X}\|_F + c_m\big)\|\mathbf{a}(k) - \mathbf{a}(s)\|_2.
\end{aligned} \tag{183}$$

Proof. Observe that $\mathbf{T}(s) = \varphi(\mathbf{W}(s)\mathbf{T}(s) + \mathbf{U}(s)\mathbf{X})$. Using the fact that $|\varphi(x) - \varphi(0)| \le L|x|$ (Lipschitz condition of φ), we have

$$\begin{aligned}
\|\mathbf{T}(s) - \varphi(\underline{0})\|_F &= \big\|\varphi(\mathbf{W}(s)\mathbf{T}(s) + \mathbf{U}(s)\mathbf{X}) - \varphi(\underline{0})\big\|_F \\
&\le L\|\mathbf{W}(s)\mathbf{T}(s) + \mathbf{U}(s)\mathbf{X}\|_F \\
&\le L\bigg(\|\mathbf{W}(s)\|_2\|\mathbf{T}(s)\|_F + \|\mathbf{U}(s)\|_2\|\mathbf{X}\|_F\bigg) \\
&\le L\bar{\rho}_w\|\mathbf{T}(s)\|_F + L\bar{\rho}_u\|\mathbf{X}\|_F.
\end{aligned} \tag{184}$$

From (184), we have

$$\begin{aligned}
\|\mathbf{T}(s)\|_F &\le \|\varphi(\underline{0})\|_F + L\bar{\rho}_w\|\mathbf{T}(s)\|_F + L\bar{\rho}_u\|\mathbf{X}\|_F \\
&= |\varphi(0)|\sqrt{mn} + L\bar{\rho}_w\|\mathbf{T}(s)\|_F + L\bar{\rho}_u\|\mathbf{X}\|_F.
\end{aligned} \tag{185}$$

Since $\bar{\rho}_w < 1/L$, from (185), we obtain

$$\|\mathbf{T}(s)\|_F \leq c_a\|\mathbf{X}\|_F + c_m. \tag{186}$$

Now, we prove (179)-(181). By using Lemma 20 with $\mathbf{A} = \mathbf{I}_{mn}, \mathbf{B} = \mathbf{J}(s), \boldsymbol{\Delta} = -\mathbf{D}(s)(\mathbf{I}_n \otimes \mathbf{W}(s))$, we have

$$\begin{aligned}\|\mathbf{J}(s)^{-1}\|_2 &\leq \frac{1}{1 - \|\mathbf{D}(\tau)(\mathbf{I}_n \otimes \mathbf{W}(s))\|_2} \\ &\leq \frac{1}{1 - \|\mathbf{D}(s)\|_2\|\mathbf{W}(s)\|_2}.\end{aligned} \tag{187}$$

On the other hand since $\|\varphi'\|_\infty \leq L$, we have

$$\|\mathbf{D}(s)\|_2 \leq L. \tag{188}$$

Hence, from (187), we have

$$\|\mathbf{J}(s)^{-1}\|_2 \leq \frac{1}{1 - L\bar{\rho}_w},$$

and thus it holds that

$$\begin{aligned}\|\mathbf{R}(s)\|_2 &\leq \|\mathbf{a}(s)\|_2\|\mathbf{J}(s)^{-1}\|_2\|\mathbf{D}(s)\|_2 \\ &\leq \frac{L\bar{\rho}_a}{1 - L\bar{\rho}_w}.\end{aligned} \tag{189}$$

Then, we have

$$\begin{aligned}\|\nabla_{\mathbf{W}}\Phi(s)\|_F &= \|\mathrm{vec}(\nabla_{\mathbf{W}}\Phi(\mathrm{s}))\|_2 \\ &= \|(\mathbf{T}(s) \otimes \mathbf{I}_m)\mathbf{R}(s)^T(\hat{\mathbf{y}}(s) - \mathbf{y})\|_2 \\ &\leq \|\mathbf{T}(s)\|_2\|\mathbf{R}(s)\|_2\|\hat{\mathbf{y}}(s) - \mathbf{y}\|_2 \\ &\leq \frac{L\bar{\rho}_a}{1 - L\bar{\rho}_w}(c_a\|\mathbf{X}\|_F + c_m)\|\hat{\mathbf{y}}(s) - \mathbf{y}\|_2,\end{aligned} \tag{190}$$

$$\begin{aligned}\|\nabla_{\mathbf{U}}\Phi(s)\|_F &= \|\mathrm{vec}(\nabla_{\mathbf{U}}\Phi(\mathrm{s}))\|_2 \\ &= \|(\mathbf{X} \otimes \mathbf{I}_m)\mathbf{R}(s)^T(\hat{\mathbf{y}}(s) - \mathbf{y})\|_2 \\ &\leq \frac{L\bar{\rho}_a}{1 - L\bar{\rho}_w}\|\mathbf{X}\|_F\|\hat{\mathbf{y}}(s) - \mathbf{y}\|_2,\end{aligned} \tag{191}$$

$$\begin{aligned}\|\nabla_{\mathbf{a}}\Phi(s)\|_F &= \|\mathbf{T}(s)(\hat{\mathbf{y}}(s) - \mathbf{y})\| \\ &\leq \big(c_a\|\mathbf{X}\|_F + c_m\big)\|\hat{\mathbf{y}}(s) - \mathbf{y}\|_2.\end{aligned} \tag{192}$$

Next, we prove (182). Observe that

$$\begin{aligned}&\|\mathbf{T}(k) - \mathbf{T}(s)\|_F \\ &\quad= \|\varphi(\mathbf{W}(k)\mathbf{T}(k) + \mathbf{U}(k)\mathbf{X}) - \varphi(\mathbf{W}(s)\mathbf{T}(s) + \mathbf{U}(s)\mathbf{X})\|_F \\ &\quad\leq L\|\mathbf{W}(k)\mathbf{T}(k) + \mathbf{U}(k)\mathbf{X} - \mathbf{W}(s)\mathbf{T}(s) - \mathbf{U}(s)\mathbf{X}\|_F \\ &\quad\leq L\big(\|\mathbf{W}(k)\mathbf{T}(k) - \mathbf{W}(k)\mathbf{T}(s)\|_F + \|\mathbf{W}(k)\mathbf{T}(s) - \mathbf{W}(s)\mathbf{T}(s)\|_F \\ &\qquad+ \|\mathbf{U}(k)\mathbf{X} - \mathbf{U}(s)\mathbf{X}\|_F\big) \\ &\quad\leq L\|\mathbf{W}(k)\|_2\|\mathbf{T}(k) - \mathbf{T}(s)\|_F + L\|\mathbf{W}(k) - \mathbf{W}(s)\|_2\|\mathbf{T}(s)\|_F \\ &\qquad+ L\|\mathbf{U}(k) - \mathbf{U}(s)\|_2\|\mathbf{X}\|_F \\ &\quad\leq L\bar{\rho}_w\|\mathbf{T}(k) - \mathbf{T}(s)\|_F + L\big(c_a\|\mathbf{X}\|_F + c_m\big)\|\mathbf{W}(k) - \mathbf{W}(s)\|_2 \\ &\qquad+ L\|\mathbf{U}(k) - \mathbf{U}(s)\|_2\|\mathbf{X}\|_F.\end{aligned} \tag{193}$$

From (193), we obtain

$$\|\mathbf{T}(k)-\mathbf{T}(s)\|_F \le \frac{L}{1-L\bar{\rho}_w}\big(c_a\|\mathbf{X}\|_F+c_m\big)\|\mathbf{W}(k)-\mathbf{W}(s)\|_2 + \frac{L}{1-L\bar{\rho}_w}\|\mathbf{U}(k)-\mathbf{U}(s)\|_2\|\mathbf{X}\|_F. \tag{194}$$

Finally, we prove (183). Observe that

$$\begin{aligned}
&\|\hat{\mathbf{y}}(k)-\hat{\mathbf{y}}(s)\|_F \\
&\quad= \|\mathbf{a}(k)\mathbf{T}(k)-\mathbf{a}(s)\mathbf{Z}(s)\|_F \\
&\quad\le \|\mathbf{a}(k)\mathbf{T}(k)-\mathbf{a}(k)\mathbf{T}(s)\|_F+\|\mathbf{a}(k)\mathbf{T}(s)-\mathbf{a}(s)\mathbf{T}(s)\|_F \\
&\quad\le \|\mathbf{a}(k)\|_2\|\mathbf{T}(k)-\mathbf{T}(s)\|_F+\|\mathbf{a}(k)-\mathbf{a}(s)\|_2\|\mathbf{T}(s)\|_F \\
&\quad\le \bar{\rho}_a\bigg[\frac{L}{1-L\bar{\rho}_w}\big(c_a\|\mathbf{X}\|_F+c_m\big)\|\mathbf{W}(k)-\mathbf{W}(s)\|_2 \\
&\qquad+\frac{L}{1-L\bar{\rho}_w}\|\mathbf{U}(k)-\mathbf{U}(s)\|_2\|\mathbf{X}\|_F\bigg]+\big(c_a\|\mathbf{X}\|_F+c_m\big)\|\mathbf{a}(k)-\mathbf{a}(s)\|_2.
\end{aligned} \tag{195}$$

□

Now, we return to prove Theorem 3. We prove by induction for every $\tau>0$,

$$\|\mathbf{W}(s)\|\le\bar{\rho}_w, \|\mathbf{U}(s)\|\le\bar{\rho}_u, \|\mathbf{a}(s)\|_2\le\bar{\rho}_a, s\in[0,\tau], \tag{196}$$

$$\lambda_s \ge \frac{\lambda_0}{2}, s\in[0,\tau], \tag{197}$$

$$\Phi(s)\le\left(1-\eta\frac{\lambda_0}{2}\right)^s\Phi(0), \qquad s\in[0,\tau]. \tag{198}$$

For $\tau=0$, it is clear that (196)-(198) hold. Assume that (196)-(198) holds up to τ iterations. Then, by using triangle inequality, we have

$$\begin{aligned}
\|\mathbf{W}(\tau+1)-\mathbf{W}(0)\|_F &\le \sum_{s=0}^{\tau}\|\mathbf{W}(s+1)-\mathbf{W}(s)\|_F \\
&= \sum_{s=0}^{\tau}\eta\|\nabla_{\mathbf{W}}\Phi(s)\|_F \\
&\le \eta\sum_{s=0}^{\tau}c_u\big(c_a\|\mathbf{X}\|_F+c_m\big)\|\hat{\mathbf{y}}(s)-\mathbf{y}\|_2 \qquad (199)\\
&= \eta c_u\big(c_a\|\mathbf{X}\|_F+c_m\big)\sum_{s=0}^{\tau}\left(1-\eta\frac{\lambda_0}{2}\right)^{s/2}\|\hat{\mathbf{y}}(0)-\hat{\mathbf{y}}\|_2, \qquad (200)
\end{aligned}$$

where (199) follows from Lemma 23. Let $u:=\sqrt{1-\eta\lambda_0/2}$. Then $\|\mathbf{W}(\tau+1)-\mathbf{W}(0)\|_F$ can be bounded with

$$\begin{aligned}
&\frac{2}{\lambda_0}(1-u^2)\frac{1-u^{\tau+1}}{1-u}c_u\big(c_a\|\mathbf{X}\|_F+c_m\big)\|\hat{\mathbf{y}}(0)-\mathbf{y}\| \\
&\le \frac{4}{\lambda_0}c_u\big(c_a\|\mathbf{X}\|_F+c_m\big)\|\hat{\mathbf{y}}(0)-\mathbf{y}\| \\
&\le \delta.
\end{aligned} \tag{201}$$

Then, we have

$$\|\mathbf{W}(\tau+1)\|\le\|\mathbf{W}(0)\|_2+\delta=\bar{\rho}_w<1/L. \tag{202}$$

Using the similar technique, one can show that

$$\begin{aligned}
\|\mathbf{U}(\tau+1)-\mathbf{U}(0)\|_F &\le \sum_{s=0}^{\tau}\|\mathbf{U}(s+1)-\mathbf{U}(s)\|_2 \\
&= \sum_{s=0}^{\tau}\eta\|\nabla_{\mathbf{U}}\Phi(s)\|_F \\
&\le \sum_{s=0}^{\tau}\eta c_u\|\mathbf{X}\|_F\|\hat{\mathbf{y}}(s)-\mathbf{y}\|_2 \\
&\le \eta c_u\|\mathbf{X}\|_F\sum_{s=0}^{\tau}\left(1-\eta\frac{\lambda_0}{2}\right)^{s/2}\|\hat{\mathbf{y}}(0)-\mathbf{y}\|_2 \\
&\le \frac{4}{\lambda_0}c_u\|\mathbf{X}\|_F\|\hat{\mathbf{y}}(0)-\mathbf{y}\|_2 \\
&\le \delta, \qquad (203)\\
\|\mathbf{a}(\tau+1)-\mathbf{a}(0)\|_F &\le \sum_{s=0}^{\tau}\|\mathbf{a}(s+1)-\mathbf{a}(s)\|_F \\
&= \sum_{s=0}^{\tau}\eta\|\nabla_{\mathbf{a}}\Phi(s)\|_F \\
&\le \eta\big(c_a\|\mathbf{X}\|_F+c_m\big)\sum_{s=0}^{\tau}\|\hat{\mathbf{y}}(s)-\mathbf{y}\|_2 \\
&\le \eta\big(c_a\|\mathbf{X}\|_F+c_m\big)\sum_{s=0}^{\tau}\left(1-\eta\frac{\lambda_0}{2}\right)^{s/2}\|\hat{\mathbf{y}}(0)-\mathbf{y}\|_2 \\
&\le \frac{4}{\lambda_0}\big(c_a\|\mathbf{X}\|_F+c_m\big)\|\hat{\mathbf{y}}(0)-\mathbf{y}\|_2 \\
&\le \delta. \qquad (204)
\end{aligned}$$

Finally, using (182), we have

$$\begin{aligned}
\|\mathbf{T}(\tau+1)-\mathbf{T}(0)\| &\le \frac{L}{1-L\bar{\rho}_w}\big(c_a\|\mathbf{X}\|_F+c_m\big)\|\mathbf{W}(\tau+1)-\mathbf{W}(0)\|_2 \\
&\quad + \frac{L}{1-L\bar{\rho}_w}\|\mathbf{U}(\tau+1)-\mathbf{U}(0)\|_2\|\mathbf{X}\|_F \\
&\le \frac{L}{1-L\bar{\rho}_w}\big(c_a\|\mathbf{X}\|_F+c_m\big)\frac{4}{\lambda_0}c_u\big(c_a\|\mathbf{X}\|_F+c_m\big)\|\hat{\mathbf{y}}(0)-\mathbf{y}\| \\
&\quad + \frac{L}{1-L\bar{\rho}_w}\frac{4}{\lambda_0}c_u\|\mathbf{X}\|_F\|\hat{\mathbf{y}}(0)-\mathbf{y}\|_2\|\mathbf{X}\|_F \\
&= \frac{4L}{(1-L\bar{\rho}_w)\lambda_0}\left[c_u\big(c_a\|\mathbf{X}\|_F+c_m\big)^2+c_u\|\mathbf{X}\|_F^2\right]\|\hat{\mathbf{y}}(0)-\mathbf{y}\|_2 \\
&\le \frac{2-\sqrt{2}}{2}\sqrt{\lambda_0} \qquad (205)
\end{aligned}$$

by (6).

By Wely's inequality, it implies that the least singular value of $\mathbf{T}(\tau+1)$ satisfies $\sigma_{\min}(\mathbf{T}(\tau+1)) \ge \sqrt{\frac{\lambda_0}{2}}$. Thus, it holds $\lambda_{\tau+1} \ge \frac{\lambda_0}{2}$.

Now, we define $\mathbf{g} := \mathbf{a}(\tau+1)^T\mathbf{T}(\tau)$ and note that

$$\Phi(\tau+1)-\Phi(\tau)$$

$$= \frac{1}{2}\|\hat{\mathbf{y}}(\tau+1) - \hat{\mathbf{y}}(\tau)\|_2^2 + (\hat{\mathbf{y}}(\tau+1) - \mathbf{g})^T(\hat{\mathbf{y}}(\tau) - \mathbf{y}) + (\mathbf{g} - \hat{\mathbf{y}}(\tau))^T(\hat{\mathbf{y}}(\tau) - \mathbf{y}). \tag{206}$$

We bound each term of the RHS of this equation individually. First, using (183), we have

$$\begin{aligned}
&\|\hat{\mathbf{y}}(\tau+1) - \hat{\mathbf{y}}(\tau)\|_2 \\
&\quad\le \bar{\rho}_a\Bigg[\frac{L}{1-L\bar{\rho}_w}\big(c_a\|\mathbf{X}\|_F + c_m\big)\|\mathbf{W}(\tau+1) - \mathbf{W}(\tau)\|_2 \\
&\qquad + \frac{L}{1-L\bar{\rho}_w}\|\mathbf{U}(\tau+1) - \mathbf{U}(\tau)\|_2\|\mathbf{X}\|_F\Bigg] + \big(c_a\|\mathbf{X}\|_F + c_m\big)\|\mathbf{a}(\tau+1) - \mathbf{a}(\tau)\|_2 \\
&\quad= \bar{\rho}_a\Bigg[\frac{L}{1-L\bar{\rho}_w}\big(c_a\|\mathbf{X}\|_F + c_m\big)\eta c_u(c_a\|\mathbf{X}\|_F + c_m)\|\hat{\mathbf{y}}(\tau) - \mathbf{y}\|_2 \\
&\qquad + \frac{L}{1-L\bar{\rho}_w}\eta c_u\|\mathbf{X}\|_F\|\hat{\mathbf{y}}(\tau) - \mathbf{y}\|_2\|\mathbf{X}\|_F\Bigg] \\
&\qquad + \big(c_a\|\mathbf{X}\|_F + c_m\big)\eta(c_a\|\mathbf{X}\|_F + c_m)\|\hat{\mathbf{y}}(\tau) - \mathbf{y}\|_2 \\
&\quad= \eta C_1\|\hat{\mathbf{y}}(\tau) - \mathbf{y}\|_2,
\end{aligned} \tag{207}$$

where $C_1 := c_u^2(c_a\|\mathbf{X}\|_F + c_m)^2 + c_u^2\|\mathbf{X}\|_F^2 + (c_a\|\mathbf{X}\|_F + c_m)^2$.

On the other hand, we have

$$\begin{aligned}
&(\hat{\mathbf{y}}(\tau+1) - \mathbf{g})^T(\hat{\mathbf{y}}(\tau) - \mathbf{y}) \\
&\quad= \mathbf{a}(\tau+1)^T(\mathbf{T}(\tau+1) - \mathbf{T}(\tau))(\hat{\mathbf{y}}(\tau) - \mathbf{y}) \\
&\quad\le \|\mathbf{a}(\tau+1)\|_2\|\mathbf{T}(\tau+1) - \mathbf{T}(\tau)\|_2\|\hat{\mathbf{y}}(\tau) - \mathbf{y}\|_2 \\
&\quad\le \|\mathbf{a}(\tau+1)\|_2\Bigg[\frac{L}{1-L\bar{\rho}_w}\big(c_a\|\mathbf{X}\|_F + c_m\big)\|\mathbf{W}(\tau+1) - \mathbf{W}(\tau)\|_2 \\
&\qquad + \frac{L}{1-L\bar{\rho}_w}\|\mathbf{U}(\tau+1) - \mathbf{U}(\tau)\|_2\|\mathbf{X}\|_F\Bigg]\|\hat{\mathbf{y}}(\tau) - \mathbf{y}\|_2 \\
&\quad\le \bar{\rho}_a\Bigg[\frac{L}{1-L\bar{\rho}_w}\big(c_a\|\mathbf{X}\|_F + c_m\big)\|\eta c_u\big(c_a\|\mathbf{X}\|_F + c_m\big)\|\hat{\mathbf{y}}(\tau) - \mathbf{y}\|_2 \\
&\qquad + \frac{L}{1-L\bar{\rho}_w}\eta c_u\|\mathbf{X}\|_F\|\hat{\mathbf{y}}(\tau) - \mathbf{y}\|_2\|\mathbf{X}\|_F\Bigg]\|\hat{\mathbf{y}}(\tau) - \mathbf{y}\|_2 \\
&\quad= \eta C_2\|\hat{\mathbf{y}}(\tau) - \mathbf{y}\|_2^2,
\end{aligned} \tag{208}$$

where $C_2 := c_u^2\big(c_a\|\mathbf{X}\|_F + c_m\big)^2 + c_u^2\|\mathbf{X}\|_F^2$.

Furthermore, we also have

$$\begin{aligned}
&(\mathbf{g} - \hat{\mathbf{y}}(\tau))^T(\hat{\mathbf{y}}(\tau) - \mathbf{y}) \\
&\quad= (\mathbf{a}(\tau+1) - \mathbf{a}(\tau))^T\mathbf{T}(\tau)(\hat{\mathbf{y}}(\tau) - \mathbf{y}) \\
&\quad= -\big(\eta\nabla_{\mathbf{a}}\Phi(\tau)\big)^T\mathbf{T}(\tau)(\hat{\mathbf{y}}(\tau) - \mathbf{y}) \\
&\quad= -\eta(\hat{\mathbf{y}}(\tau) - \mathbf{y})^T\mathbf{T}(\tau)^T\mathbf{T}(\tau)(\hat{\mathbf{y}}(\tau) - \mathbf{y}) \\
&\quad\le -\eta\frac{\lambda_0}{2}\|\hat{\mathbf{y}}(\tau) - \mathbf{y}\|_2^2
\end{aligned} \tag{209}$$

where we use induction $\lambda_\tau \ge \frac{\lambda_0}{2}$.

From (206)-(209), we obtain

$$\begin{aligned}
&\Phi(\tau+1)-\Phi(\tau)\\
&\quad\le \frac{1}{2}\eta^2 C_1^2\|\hat{\mathbf{y}}(\tau)-\mathbf{y}\|_2^2+\eta C_2\|\hat{\mathbf{y}}(\tau)-\mathbf{y}\|_2^2-\eta\frac{\lambda_0}{2}\|\hat{\mathbf{y}}(\tau)-\mathbf{y}\|_2^2\\
&\quad= 2\Phi(\tau)\left[\frac{1}{2}\eta^2 C_1^2+\eta C_2-\eta\frac{\lambda_0}{2}\right]\\
&\quad= \Phi(\tau)\Big[\eta^2 C_1^2+2\eta C_2-\eta\lambda_0\Big],
\end{aligned}$$

which leads to

$$\begin{aligned}
\Phi(\tau+1)&\le \Phi(\tau)\Big[1-\eta(\lambda_0-\eta C_1^2-2C_2)\Big]\\
&\le \big(1-\eta(\lambda_0-4C_2)\big)\Phi(\tau)\\
&\le \left(1-\eta\frac{\lambda_0}{2}\right)\Phi(\tau). \qquad (210)
\end{aligned}$$

I Proof of Lemma 18

Observe that

$$\|\hat{\mathbf{y}}(0)-\mathbf{y}\|\le\|\hat{\mathbf{y}}(0)\|+\|\mathbf{y}\|. \qquad (211)$$

On the other hand, let $\mathbf{v}_1,\mathbf{v}_2,\cdots,\mathbf{v}_n$ be n columns of $\mathbf{T}(0)$. Then, we have

$$\begin{aligned}
\|\hat{\mathbf{y}}(0)\|&=\big\|\mathbf{a}^T(0)\mathbf{T}^*(0)\big\|\\
&=\sqrt{\sum_{i=1}^n(\mathbf{a}^T(0)\mathbf{v}_i)^2}. \qquad (212)
\end{aligned}$$

Now, observe that

$$\begin{aligned}
\mathbb{P}\left[\sum_{i=1}^n(\mathbf{a}^T(0)\mathbf{v}_i)^2\ge n\frac{L^2}{t}\right]&\le\frac{\mathbb{E}\big[\sum_{i=1}^n(\mathbf{a}^T(0)\mathbf{v}_i)^2\big]}{n\frac{L^2}{t}}\\
&=\frac{t}{nL^2}\sum_{i=1}^n\mathbb{E}\big[(\mathbf{a}^T(0)\mathbf{v}_i)^2\big]\\
&=\frac{t}{nL^2}\sum_{i=1}^n\frac{\mathbb{E}[\|\mathbf{v}_i\|^2]}{m} \qquad (213)\\
&\le\frac{t}{nL^2}\sum_{i=1}^n\frac{L^2m}{m} \qquad (214)\\
&=t, \qquad (215)
\end{aligned}$$

where (213) follows from Assumption 1 that $\mathbf{a}$ is initialised with a random vector with i.i.d. entries $\mathcal{N}(0,1/m)$ and the fact that $\mathbf{a}(0)$ is independent of $\mathbf{v}_i$, (214) follows from the fact that $\mathbf{v}_i=\varphi(\mathbf{W}\mathbf{v}_i+\mathbf{U}\mathbf{x}_i)$, so $\|\mathbf{v}_i\|_\infty=\|\varphi(\mathbf{W}\mathbf{v}_i+\mathbf{U}\mathbf{x}_i)\|_\infty\le L$.

From (215), with probability at least $1-t$ it holds that

$$\sum_{i=1}^n(\mathbf{a}^T(0)\mathbf{v}_i)^2\le n\frac{L^2}{t}=O(n),$$

which leads to

$$\|\hat{\mathbf{y}}(0)\| = O(\sqrt{n}) \tag{216}$$

by (212).

In addition, we have

$$\begin{aligned} \|\mathbf{y}\| &= \sqrt{\sum_{i=1}^{n} y_i^2} \\ &= O(\sqrt{n}) \end{aligned} \tag{217}$$

by Assumption 2.

From (211), (216), and (217) we obtain

$$\|\hat{\mathbf{y}}(0) - \mathbf{y}\| = O(\sqrt{n}). \tag{218}$$