Journal of Machine Learning Research 25 (2024) 1-83 Submitted 6/23; Revised 1/24; Published 3/24

Overparametrized Multi-layer Neural Networks: Uniform Concentration of Neural Tangent Kernel and Convergence of Stochastic Gradient Descent*

Jiaming Xu JX77@DUKE.EDU
The Fuqua School of Business
Duke University
Durham, NC 27708, USA

Hanjing Zhu HZ176@DUKE.EDU
The Fuqua School of Business
Duke University
Durham, NC 27708, USA

Editor: Joan Bruna

Abstract

There have been exciting progresses in understanding the convergence of gradient descent (GD) and stochastic gradient descent (SGD) in overparameterized neural networks through the lens of neural tangent kernel (NTK). However, there remain two significant gaps between theory and practice. First, the existing convergence theory only takes into account the contribution of the NTK from the last hidden layer, while in practice the intermediate layers also play an instrumental role. Second, most existing works assume that the training data are provided a priori in a batch, while less attention has been paid to the important setting where the training data arrive in a stream. In this paper, we close these two gaps. We first show that with random initialization, the NTK function converges to some deterministic function uniformly for all layers as the number of neurons tends to infinity. Then we apply the uniform convergence result to further prove that the prediction error of multi-layer neural networks under SGD converges in expectation in the streaming data setting. A key ingredient in our proof is to show the number of activation patterns of an L-layer neural network with width m is only polynomial in m although there are mL neurons in total.

Keywords: neural network, neural tangent kernel, stochastic gradient descent, uniform concentration

1. Introduction

Deep Learning is proven to be successful in many real-life applications, while the underpinning of its success remains elusive. Recently, researchers are interested in understanding the success of neural networks from the optimization perspective. A neural network with Rectified Linear Units (ReLU) activation leads to a non-convex and non-smooth objective function, which is usually hard to optimize by gradient descent methods. However, surpris-

*. This work was presented in part at *the The 24th International Conference on Artificial Intelligence and Statistics*, 2021 Xu and Zhu (2021).

ingly, gradient descent (GD) or stochastic gradient descent (SGD) on neural networks with ReLU activation is observed to perform well not only in training but also in generalization (Krizhevsky et al., 2012). To demystify this phenomenon, an extensive amount of research has been done recently. For instance, the mean-field theory is used in Chen et al. (2020); Mei et al. (2018, 2019); Rotskoff and Vanden-Eijnden (2018); Sirignano and Spiliopoulos (2022); Tzen and Raginsky (2020) to analyze the SGD of infinite-width feed-forward neural networks. Optimal transport theory is employed in Chizat and Bach (2018) to study the gradient flow of neural networks and to show that the training error converges to the global optimum under some mild conditions. In addition, Hu et al. (2019) connects the SGD of neural networks in training to the diffusion process.

A different line of research focuses on understanding the gradient descent of neural networks through kernels, in particular the neural tangent kernel (NTK). Specifically, given an L-layer neural network $f(x;\mathbf{W})$ with input x and parameter $\mathbf{W}$, we define NTK for a sequence of weights $\{\mathbf{W}(t)\}$ as

$$H_t(x,x') \triangleq \left\langle \frac{\partial f(x;\mathbf{W}(t))}{\partial \mathbf{W}}, \frac{\partial f(x';\mathbf{W}(t))}{\partial \mathbf{W}} \right\rangle = \sum_{\ell=1}^{L} H_t^{(\ell)}(x,x')\,, \tag{1}$$

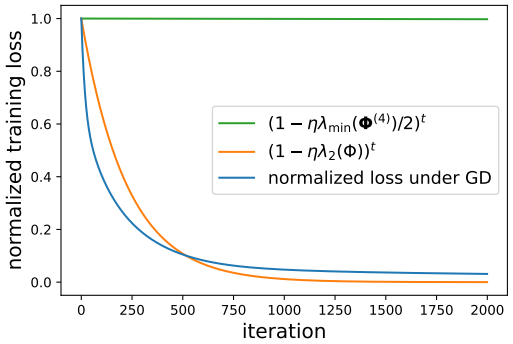
where

$$H_t^{(\ell)}(x,x') \triangleq \left\langle \frac{\partial f(x;\mathbf{W}(t))}{\partial \mathbf{W}^{(\ell)}}, \frac{\partial f(x';\mathbf{W}(t))}{\partial \mathbf{W}^{(\ell)}} \right\rangle \tag{2}$$

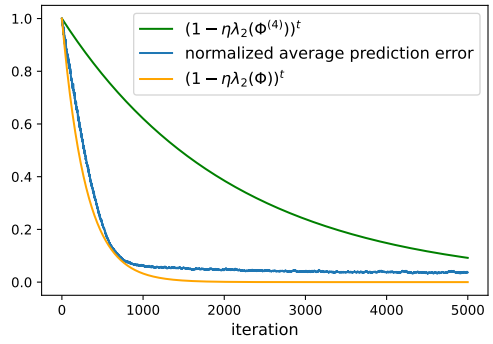
is the NTK from the ℓ-th hidden layer. It is first introduced by Jacot et al. (2018), which shows that gradient descent on infinite width neural networks can be viewed as learning through the NTK. Subsequent works (Allen-Zhu et al., 2019a; Du et al., 2019b; Su and Yang, 2019; Arora et al., 2019a; Du et al., 2019a; Zou et al., 2020) connect GD and SGD with the NTK, and show that with overparameterization and random initialization, the training error converges to 0. Similar convergence results are also established in other types of neural networks beyond feed-forward neural networks (Arora et al., 2019b; Allen-Zhu and Li, 2020, 2019a,b; Allen-Zhu et al., 2019b; Du et al., 2018; Li et al., 2019; Tirer et al., 2022), such as convolutional neural networks (CNN) and residual neural networks (ResNet).

Despite these remarkable progresses, there remain two significant gaps between theory and practice. Firstly, the existing theory does not accurately characterize the convergence rate of GD. Specifically, given a batch $\{(x_i,y_i)\}_{i=1}^n$, Du et al. (2019a) first shows that the NTK matrix from the last hidden layer $\mathbf{H}_t^{(L)} = \left(\frac{1}{n}H_t^{(L)}(x_i,x_j)\right)$ is close to some deterministic kernel matrix $\mathbf{\Phi}^{(L)}$. Based on this, the authors further show the training loss converges at a linear rate $\left(1-\frac{\eta}{2}\lambda_{\min}(\mathbf{\Phi}^{(L)})\right)^t$ where η is the step size. However, such a characterization based on the last hidden layer is very loose for two reasons. First of all, the characterization only captures the contribution from the last hidden layer. Secondly, as pointed out by Su and Yang (2019), $\lambda_{\min}(\mathbf{\Phi}^{(L)})$ goes to 0 as the batch size n goes to infinity. In contrast, as illustrated in Figure 1a, the actual GD dynamic converges much faster and can be more accurately characterized via the spectrum of the integral operator Φ associated with some deterministic kernel function $\Phi = \sum_{\ell=1}^{L}\Phi^{(\ell)}$, which captures the contribution from all layers.

Secondly, most existing works study GD in the batch setting where the training data is provided a priori in a batch. It remains unclear how SGD performs in the streaming setting



(a) Training loss under GD



(b) Average prediction error under SGD

Figure 1: Comparison of GD/SGD dynamic versus different characterizations. The actual estimation errors are shown in blue. The characterization based on the last layer is shown in green while the characterization based on all layers is shown in orange. The data is generated according to $y = f^*(x) + u$, where f^* is a linear function, $u \sim \mathcal{N}(0, 0.01)$, and x is generated uniformly over the unit sphere. We learn a 4-layer neural network with $m = 1000$ neurons in each hidden layer using step size $\eta = 0.2$. We use the symmetric initialization introduced in Section 4. For both GD and SGD, we normalize the error by the error at initialization. According to Corollary 4, under the symmetric initialization, $\lambda_2(\Phi)$ provides a good characterization for linear f^*.

where the data arrives in a stream. The streaming data arises in a variety of fields such as finance, news organization, and information technology (O'callaghan et al., 2002; Allen-Zhu and Li, 2019a; Ikonomovska et al., 2007). Such streaming data is usually inspected once and archived afterwards immediately without being examined again. Apart from vast sources of naturally generated streaming data, there are ubiquitous situations where the streaming data is preferred even though batches of samples can be obtained. For instance, O'callaghan et al. (2002) points out that in medical or marketing data mining, the volume of data is so large that only one pass over data is allowed due to computational constraints. Moreover, Feigenbaum et al. (2001); Muthukrishnan (2005) argues that the streaming data is useful in privacy-preserving data mining, where the data is kept confidential by users and analyzed via a single pass.

It is challenging to close these two gaps. The analysis of the NTK from intermediate layers is significantly harder than that from only the last hidden layer. To see this, the analysis of the last hidden layer reduces to the one hidden layer analysis by treating the output from the second-to-last hidden layer as the input (Du et al., 2019b). In particular, conditioning on the output from the second-to-last hidden layer, the NTK from the last hidden layer can be written as a sum of independent random variables. In contrast, the NTKs from intermediate layers not only depend on weights from previous layers but also subsequent layers. Thus, there is no similar conditional independence structure like the last hidden layer for us to utilize. As a result, a completely new method is needed to analyze the intermediate layers.

To see why it is hard to study SGD in the streaming setting, note that a critical step in existing analysis of GD on the batch setting (Du et al., 2019a) is to obtain the concentration of the finite-dimensional NTK matrix. There, pointwise concentration and union bounds suffice to obtain the desired convergence. However, in the analysis of SGD under the streaming setting, we need to show the uniform concentration of the infinite-dimensional kernel function. Existing analysis techniques tailored to finite-dimensional kernel matrices are not enough to obtain the uniform convergence.

In this paper, we overcome the above challenges and close the two gaps. In particular,

- For an L-layer fully connected feed-forward neural network $f(x; \mathbf{W})$ with m neurons in each layer, we show that under Gaussian initialization, with high probability as $m \to \infty$, the NTK function from the ℓ-th hidden layer $H_0^{(\ell)}(x, x')$ defined in (2) at initialization concentrates on some deterministic function $\Phi^{(\ell)}(x, x')$ uniformly for all $x, x' \in \mathbb{S}^{d-1}$ and all layers $1 \le \ell \le L$;

- We further apply the uniform concentration of NTK to show that with high probability as $m \to \infty$, the average prediction error under SGD at iteration T in the streaming setting is upper bounded by

$$\inf_{\ell \ge 1} \left\{ \prod_{t=0}^{T-1} (1 - \eta_t \lambda_\ell) \left\| \Delta_0 \right\|_2 + \mathcal{R}(\Delta_0, \ell) \right\} + \mathfrak{Y},$$

  where η_t is the step size at iteration t, λ_ℓ is the ℓ-th eigenvalue of the integral operator Φ associated with function $\Phi = \sum_{\ell=1}^{L} \Phi^{(\ell)}$, $\left\| \Delta_0 \right\|_2$ is the prediction error at initialization, and $\mathfrak{Y}$ is the error term capturing the approximation error from the non-linearity of ReLU function and the noise from stochastic gradients. Particularly, for an arbitrary small but fixed constant $\epsilon > 0$, by choosing an appropriate step size, we have $\mathfrak{Y} < \epsilon$, yielding a small average prediction error. In contrast to the characterization based on the NTK from only the last hidden layer, our analysis captures the contribution from all layers and provides a much tighter characterization of the average prediction error under SGD, as depicted in Figure 1b;

- On the technical front, to prove the convergence of the infinite-dimensional kernel function, one key step is to bound the number of activation patterns, that is, the sign patterns of the ReLU units in all layers when varying the network input x while fixing the weights $\mathbf{W}$. Leveraging the recursive structure of the network in layers, we show that the number of activation patterns grows multiplicatively by a factor of m^d in every layer. This immediately implies that there are at most m^{dL} different activation patterns, despite that the network has mL neurons in total.

Notation Let $(\mathcal{X}, \mu)$ denote a measurable space and $L^2(\mathcal{X}, \mu)$ denote the space of functions $f : \mathcal{X} \to \mathbb{R}$ that are integrable, i.e., $\left\| f \right\|_2 \triangleq \sqrt{\int_{\mathcal{X}} f^2(x) d\mu(x)} < \infty$. When $\mathcal{X}$ is the unit sphere $\mathbb{S}^{d-1}$ in $\mathbb{R}^d$, we abbreviate $L^2(\mathbb{S}^{d-1}, \mu)$ as $L^2(\mu)$ for simplicity. Define the L-infinite norm $\left\| f \right\|_\infty \triangleq \sup_{x \in \mathcal{X}} |f(x)|$. Given $f, g \in L^2(\mathcal{X}, \mu)$, define their inner product as $\langle f, g \rangle \triangleq \int_{\mathcal{X}} f(x) g(x) d\mu(x)$ with $\langle f, f \rangle = \left\| f \right\|_2^2$. Given a kernel function $K \in L^2(\mathcal{X} \times \mathcal{X}, \mu \otimes \mu)$, define the associated integral operator $\mathsf{K} : L^2(\mathcal{X}, \mu) \to L^2(\mathcal{X}, \mu)$ as $\mathsf{K} f(x) = \int_{\mathcal{X}} K(x, y) f(y) d\mu(y)$.

The operator norm of K is defined as $\|\mathsf{K}\|_2 \triangleq \sup_{\|f\|_2 \le 1} \|\mathsf{K}f\|_2$. Denote the composition of operators $\mathsf{K}_m \circ \mathsf{K}_{m-1} \circ \cdots \circ \mathsf{K}_1$ as $\prod_{i=1}^{m} \mathsf{K}_i$ with $\prod_{i=n+1}^{n} \mathsf{K}_i$ treated as the identity operator.

2. Related Work

There is a vast literature on overparametrized neural networks, and here we can only hope to cover a fraction of them we see the most relevant. A summary of the mostly related works on NTK is given in Table 1.

Table 1: Summary of related works

Literature	Error	Setting	Layer	Activation	Problem
Du et al. (2019b) Su and Yang (2019)	Training	Batch+GD	Single	ReLU	Regression
Du et al. (2019a)			Multi	Analytic	
Arora et al. (2019a)	Generalization	Batch+GD	Single	ReLU	Regression
Cao and Gu (2019)		Stream+SGD	Multi		Classification
this paper					Regression

To facilitate the discussion and better differentiate the algorithms, we use GD to denote the gradient descent algorithm where the entire batch is used to compute the gradient at each iteration, i.e., for the given batch $\{(x_i, y_i)\}_{i=1}^n$ and a loss function $\mathcal{L}(\cdot, \cdot)$,

$$\mathbf{W}(t+1) = \mathbf{W}(t) - \frac{\eta_t}{n} \sum_{i=1}^{n} \nabla_{\mathbf{W}} \mathcal{L}(f(x_i; \mathbf{W}(t)), y_i),$$

where $\mathbf{W}(t)$ is the weight matrix at iteration t, $f(x; \mathbf{W}(t))$ is the neural network with parameter $\mathbf{W}(t)$. In contrast, our study focuses on the one-pass SGD, abbreviated as SGD, which draws a *single* fresh sample from the true data distribution to compute the gradient at each iteration. In particular,

$$\mathbf{W}(t+1) = \mathbf{W}(t) - \eta_t \nabla_{\mathbf{W}} \mathcal{L}(f(x_t; \mathbf{W}(t)), y_t), \tag{3}$$

where (x_t, y_t) is a freshly drawn sample at the t-th iteration from some unknown distribution μ. The drawn sample (x_t, y_t) is then archived and not used any more.

Training error with batch learning For single-layer neural networks with a given batch $\{(x_i, y_i)\}_{i=1}^n$, it is shown in Du et al. (2019b) that the NTK matrix $\mathbf{H}_t = \left(\frac{1}{n} H_t(x_i, x_j)\right)$, concentrates on the deterministic matrix $\mathbf{\Phi} = \left(\frac{1}{n}\Phi(x_i, x_j)\right)$ as the number of neurons goes to infinity. Then they utilize the spectrum of $\mathbf{\Phi}$ to prove that the training error of over-parametrized neural networks under GD converges at a linear rate $\left[1 - \frac{\eta}{2}\lambda_{\min}(\mathbf{\Phi})\right]^t$, where t is the number of iterations and η is the step size. The follow-up work Su and Yang (2019) proves that as the sample size n grows, $\lambda_{\min}(\mathbf{\Phi})$ decreases to 0 and hence the convergence rate can be very close to 0. Instead, they provide a different characterization showing that the training error under GD is upper bounded by

$$\left[1 - \frac{3\eta}{4}\lambda_r(\Phi)\right]^t + 2\sqrt{2}\mathcal{R}\left(\Delta_0, r\right) + \Theta\left(\frac{1}{\sqrt{n}}\right),$$

where $\lambda_r(\Phi)$ is the r-th largest eigenvalue of the integral operator Φ associated with the kernel function $\Phi(x, x')$, and $\mathcal{R}(\Delta_0, r)$ is the L_2 norm of the projection of $\Delta_0 = f^*(x) - f(x; \mathbf{W}(0))$ onto the eigenspaces of kernel Φ associated with $\{\lambda_i(\Phi)\}_{i=r+1}^{\infty}$. In addition, Du et al. (2019a) extends the result of Du et al. (2019b) to multi-layer neural networks with analytic activation functions. In particular, they first show the NTK matrix from the last hidden layer concentrates on some deterministic matrix $\mathbf{\Phi}^{(L)}$ and then characterize the GD dynamic utilizing the spectrum of $\mathbf{\Phi}^{(L)}$. However, their analysis crucially utilizes the analytic property of the activation function, which does not cover the widely used ReLU-activated neural networks.

Generalization error with batch learning Apart from the training error, the generalization error which measures the accuracy of the model's prediction on unseen data is also of wide interest. Following Du et al. (2019b), Arora et al. (2019a) derives an upper bound of the generalization error of over-parameterized single layer neural networks under GD as

$$\mathbb{E}_{(X,y)\sim\mu}\left[\mathcal{L}(f(X; \mathbf{W}(t)), y)\right] \leq \frac{2y^\top \mathbf{\Phi}^{-1} y}{n} + O\left(\sqrt{\frac{\log(n/\lambda_{\min}(\mathbf{\Phi}))}{n}}\right),$$

where $y = (y_1, y_2, \cdots, y_n)^\top \in \mathbb{R}^n$ is the label of the i.i.d. sample $\{(X_i, y_i)\}_{i=1}^n$ drawn from the distribution μ. As mentioned above, $\lambda_{\min}(\mathbf{\Phi})$ decreases to 0 and hence the generalization error can potentially blow up to infinity as n grows.

Generalization error with streaming data To learn a neural network in streaming setting, one way is to use SGD shown in (3). One work studying SGD with streaming data is Cao and Gu (2019) which focuses on the classification problem with the hinge loss function. Technically this work applies the online-to-batch conversion proposed in Cesa-Bianchi et al. (2004) to bound the generalization error $\frac{1}{T}\sum_{s=1}^{T}\mathbb{E}_{(X,y)}\left[\mathbf{1}_{\{yf(X;W(s))<0\}}\right]$ from above by the empirical loss $\frac{1}{T}\sum_{s=1}^{T}\mathcal{L}\left(y_s f(x_s; W(s))\right.$ with the hinge loss function $\mathcal{L}(z) = \log(1+\exp(-z))$. Note that the online-to-batch conversion follows from an application of martingale concentration inequalities. It does not fully resolve the problem of bounding the generalization error as one still needs to bound the empirical loss. Indeed the authors bound the cumulative loss following a similar analysis of Du et al. (2019b) and obtain an upper bound of the generalization error as

$$\frac{1}{T}\sum_{s=1}^{T}\mathbb{E}_{(X,y)}\left[\mathbf{1}_{\{yf(X;W(s))<0\}}\right] = O\left(\sqrt{\frac{y^\top\left(\mathbf{\Phi}^{(L)}\right)^{-1}y}{T}}\right) + O\left(\sqrt{1/T}\right),$$

where $y = (y_1, \cdots, y_T)^\top$. However, as T increases, $\lambda_{\min}(\mathbf{\Phi}^{(L)})$ decreases to 0 and hence the upper bound may blow up.

3. Problem Setup

Suppose the data (X, y) is given by $y = f^*(X) + u$, where f^* is the underlying true function, $X \in \mathbb{R}^d$ is the feature vector generated according to some distribution μ on the unit sphere $\mathbb{S}^{d-1}$, and u is the bounded noise independent of X with mean 0, variance τ^2. Denote $\gamma \triangleq \max\{\|f^*\|_\infty, |u|\}$ which is independent of m.

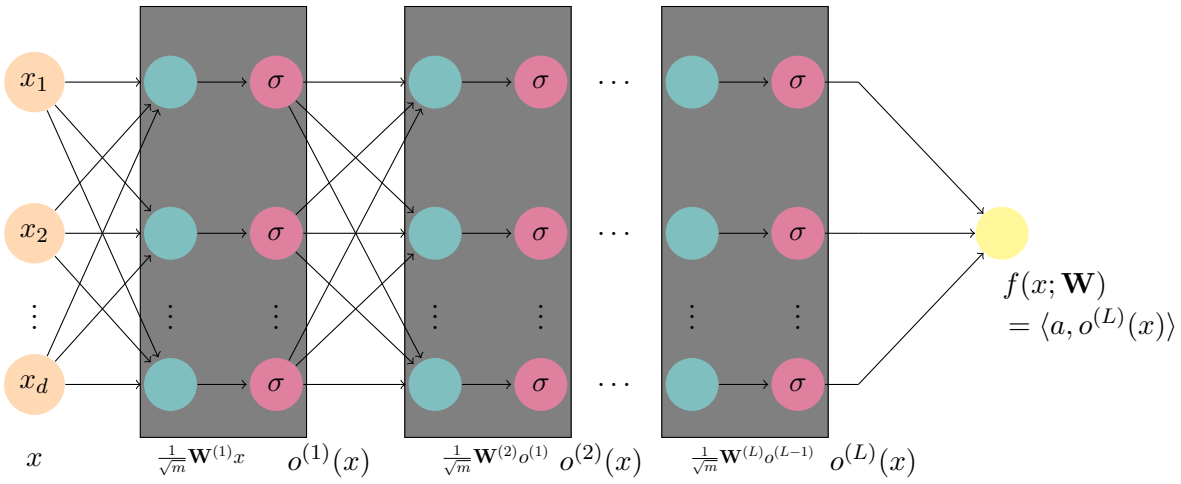


Figure 2: Illustration of Multi-Layer Neural Network $f(x;\mathbf{W})$

We consider the following L-layer neural network, as illustrated in Fig. 2:

$$f(x;\mathbf{W}) = a^\top \frac{1}{\sqrt{m}}\mathbf{D}^{(L)}(x)\mathbf{W}^{(L)} \cdots \frac{1}{\sqrt{m}}\mathbf{D}^{(1)}(x)\mathbf{W}^{(1)}x, \tag{4}$$

where $a \in \mathbb{R}^{n_L}$ is the outer weight, $\mathbf{W}^{(\ell)} \in \mathbb{R}^{n_\ell \times n_{\ell-1}}$ is the weight of the ℓ-th hidden layer whose i-th row is denoted as $w_i^{(\ell)}$,

$$\mathbf{D}^{(\ell)}(x) = \mathsf{diag}\left\{\mathbf{1}_{\left\{\langle w_i^{(\ell)}, o^{(\ell-1)}(x)\rangle \geq 0\right\}}\right\} \in \mathbb{R}^{n_\ell \times n_\ell}, \tag{5}$$

with n_ℓ as the number of neurons in the hidden layer ℓ, and $o^{(\ell)}(x)$ is the output of the ℓ-th layer given by

$$o^{(\ell)}(x) = \frac{1}{\sqrt{m}}\mathbf{D}^{(\ell)}(x)\mathbf{W}^{(\ell)} \cdots \frac{1}{\sqrt{m}}\mathbf{D}^{(1)}(x)\mathbf{W}^{(1)}x \tag{6}$$

with $o^{(0)}(x) = x$.

The neural network is trained by running the stochastic gradient descent on the streaming data in one pass. In particular, given the initialization $\left\{\mathbf{W}^{(\ell)}(0)\right\}_{\ell=1}^{L}$ and outer weight a, the ℓ-th layer weight matrix at the t-th iteration is updated as

$$\begin{aligned}\mathbf{W}^{(\ell)}(t+1) &= \mathbf{W}^{(\ell)}(t) - \eta_t \frac{\partial \mathcal{L}(y_t, f(X_t;\mathbf{W}(t)))}{\partial \mathbf{W}^{(\ell)}} \\ &= \mathbf{W}^{(\ell)}(t) + \eta_t \left(y_t - f\left(X_t;\mathbf{W}(t)\right)\right) \frac{\partial f(X_t;\mathbf{W}(t))}{\partial \mathbf{W}^{(\ell)}},\end{aligned} \tag{7}$$

where η_t is the step size, $\mathcal{L}(y,\widehat{y}) = \frac{1}{2}(y-\widehat{y})^2$ is the quadratic loss function, and (X_t, y_t) is the freshly drawn data that is independent and identically distributed as (X, y).

To derive $\frac{\partial f(x;\mathbf{W})}{\partial \mathbf{W}^{(\ell)}}$, recall from (4) and (6) that

$$\begin{aligned} f(x;\mathbf{W}) &= a^\top \frac{1}{\sqrt{m}}\mathbf{D}^{(L)}(x)\mathbf{W}^{(L)}\cdots\frac{1}{\sqrt{m}}\mathbf{D}^{(\ell)}(x)\mathbf{W}^{(\ell)}o^{(\ell-1)}(x) \\ &= a^\top\left[\mathbf{V}_L^{(\ell)}(x)\right]^\top\mathbf{W}^{(\ell)}o^{(\ell-1)}(x) \\ &= \left\langle \mathbf{V}_L^{(\ell)}(x)a\left[o^{(\ell-1)}(x)\right]^\top, \mathbf{W}^{(\ell)}\right\rangle, \end{aligned}$$

where

$$\left[\mathbf{V}_L^{(\ell)}(x)\right]^\top \triangleq \frac{1}{\sqrt{m}}\mathbf{D}^{(L)}(x)\mathbf{W}^{(L)}\cdots\frac{1}{\sqrt{m}}\mathbf{D}^{(\ell+1)}(x)\mathbf{W}^{(\ell+1)}\frac{1}{\sqrt{m}}\mathbf{D}^{(\ell)}(x). \tag{8}$$

Thus, we get[1]

$$\frac{\partial f(x;\mathbf{W})}{\partial \mathbf{W}^{(\ell)}} = \mathbf{V}_L^{(\ell)}(x)a\left[o^{(\ell-1)}(x)\right]^\top. \tag{9}$$

Plugging (9) into (7), we have

$$\mathbf{W}^{(\ell)}(t+1) = \mathbf{W}^{(\ell)}(t) + \eta_t\left(y_t - f(X_t;\mathbf{W}(t))\right)\mathbf{V}_{L,t}^{(\ell)}(x)a\left[o_t^{(\ell-1)}(x)\right]^\top, \tag{10}$$

where $\mathbf{V}_{L,t}^{(\ell)}(x)$ is defined as $\mathbf{V}_L^{(\ell)}(x)$ with $\mathbf{W}$ replaced by $\mathbf{W}(t)$.

At $t=0$, we initialize each weight matrix $\mathbf{W}^{(\ell)}(0)$ as Gaussian random matrix with *i.i.d.* standard normal entry. We also generate outer weight a to be Rademacher (symmetric Bernoulli) random variable with equal probability to be -1 or 1 which will be fixed throughout the training. This initialization is widely used in existing literature such as Du et al. (2019b); Arora et al. (2019a); Su and Yang (2019). Furthermore, it has been shown in Jacot et al. (2018) that the training dynamic of gradient descent method under this initialization is governed by the NTK defined in (1).

For ease of presentation, we assume $\gamma = O(1)$, the step size $\eta_t \le \frac{\theta}{t+1}$ for some θ independent of d and m and $n_1 = n_2 = \cdots = n_L = m$, i.e., all hidden layers have the same width, and consider the overparameterizd regime where m tends to ∞. Such overparameterized neural networks have been the focus in the literature of NTK (Allen-Zhu et al., 2019a; Du et al., 2019a).

4. Main Result

In Section 4.1, we show the uniform concentration of NTK. In Section 4.2, we apply the uniform concentration to derive an upper bound of the average prediction error under one-pass SGD.

1. Note that $\{\mathbf{D}^{(k)}, k \ge \ell\}$ all depend on $\mathbf{W}^{(\ell)}$. However, each entry of $\mathbf{D}^{(k)}$ only takes value 0 or 1, and hence does not change with $\mathbf{W}^{(\ell)}$ once its value is fixed to be either 0 and 1.

4.1 Concentration of NTK at Initialization

In this section, we show the concentration of NTK at initialization. For notation simplicity, we abbreviate H_0 as H, $H_0^{(\ell)}$ as $H^{(\ell)}$, $\mathbf{W}^{(\ell)}(0)$ as $\mathbf{W}^{(\ell)}$, $\mathbf{D}_0^{(\ell)}$ as $\mathbf{D}^{(\ell)}$ and $o_0^{(\ell)}$ as $o^{(\ell)}$ for all ℓ throughout this section and Section 5.

Note that the kernel function H is a sum of L kernel functions where $H^{(\ell)}$ represents the contribution from the ℓ-th hidden layer. To show the concentration of the kernel function H, it is sufficient to show the concentration of $H^{(\ell)}$ for each $1 \le \ell \le L$.

To obtain the closed-form expression of $H^{(\ell)}$, we plug (9) into (2) and get

$$H^{(\ell)}(x,x') = \underbrace{\langle o^{(\ell-1)}(x), o^{(\ell-1)}(x')\rangle}_{\text{(I)}} \times \underbrace{a^\top \mathbf{G}_L^{(\ell)}(x,x')\, a}_{\text{(II)}}\,, \tag{11}$$

where

$$\mathbf{G}_L^{(\ell)}(x,x') \triangleq \left[\mathbf{V}_L^{(\ell)}(x)\right]^\top \mathbf{V}_L^{(\ell)}(x') \tag{12}$$

with $\mathbf{V}_L^{(\ell)}(x)$ defined in (8).

Here, we provide a heuristic on obtaining the limiting function Φ. Consider $H^{(\ell)}$ in (11). For term (I), by the definition of $o^{(\ell)}$, we have the following recursion:

$$\langle o^{(\ell-1)}(x), o^{(\ell-1)}(x')\rangle = \frac{1}{m}\sum_{i=1}^m \sigma(\langle w_i^{(\ell-1)}, o^{(\ell-2)}(x)\rangle)\sigma(\langle w_i^{(\ell-1)}, o^{(\ell-2)}(x')\rangle). \tag{13}$$

Conditioning on $o^{(\ell-2)}$, since $w_i^{(\ell-1)}$ are i.i.d. Gaussian random vectors across i, we expect $\langle o^{(\ell-1)}(x), o^{(\ell-1)}(x')\rangle$ concentrates on its conditional mean, i.e.,

$$\langle o^{(\ell-1)}(x), o^{(\ell-1)}(x')\rangle \to \mathbb{E}_{w\sim\mathcal{N}(0,\mathbf{I})}\left[\sigma\left(\langle w, o^{(\ell-2)}(x)\rangle\right)\sigma\left(\langle w, o^{(\ell-2)}(x')\rangle\right)\right] \tag{14}$$

where

$$\left(\langle w, o^{(\ell-2)}(x)\rangle, \langle w, o^{(\ell-2)}(x')\rangle\right) \sim \mathcal{N}\left(0, \begin{pmatrix} \left\|o^{(\ell-2)}(x)\right\|_2^2 & \langle o^{(\ell-2)}(x), o^{(\ell-2)}(x')\rangle \\ \langle o^{(\ell-2)}(x), o^{(\ell-2)}(x')\rangle & \left\|o^{(\ell-2)}(x')\right\|_2^2 \end{pmatrix}\right).$$

Analogous to (14), we show the covariance matrix on the right hand side of the above displayed equation concentrates on

$$\begin{pmatrix} \mathbb{E}\left[\sigma^2\left(\langle w, o^{(\ell-3)}(x)\rangle\right)\right] & \mathbb{E}\left[\sigma\left(\langle w, o^{(\ell-3)}(x)\rangle\right)\sigma\left(\langle w, o^{(\ell-3)}(x')\rangle\right)\right] \\ \mathbb{E}\left[\sigma\left(\langle w, o^{(\ell-3)}(x)\rangle\right)\sigma\left(\langle w, o^{(\ell-3)}(x')\rangle\right)\right] & \mathbb{E}\left[\sigma^2\left(\langle w, o^{(\ell-3)}(x)\rangle\right)\right] \end{pmatrix}.$$

In view of this recursive relation of $\left(\langle w, o^{(\ell-2)}(x)\rangle, \langle w, o^{(\ell-2)}(x')\rangle\right)$, we can approximate $\left(\langle w, o^{(\ell-2)}(x)\rangle, \langle w, o^{(\ell-2)}(x')\rangle\right)$ by a pair of bivariate normal random variables. In particular, we define $(U^{(\ell-1)}(x), U^{(\ell-1)}(x'))$ such that

$$\begin{aligned} &(U^{(\ell-1)}(x), U^{(\ell-1)}(x')) \sim \mathcal{N}\left(0, \Sigma^{(\ell-2)}(x,x')\right) \\ &\Sigma^{(\ell-2)}(x,x') \triangleq \begin{pmatrix} \mathbb{E}\left[\sigma^2(U^{(\ell-2)}(x))\right] & \mathbb{E}\left[\sigma(U^{(\ell-2)}(x))\sigma(U^{(\ell-2)}(x'))\right] \\ \mathbb{E}\left[\sigma(U^{(\ell-2)}(x))\sigma(U^{(\ell-2)}(x'))\right] & \mathbb{E}\left[\sigma^2(U^{(\ell-2)}(x'))\right] \end{pmatrix} \end{aligned} \tag{15}$$

with $\Sigma^{(0)}(x,x') = \begin{pmatrix} 1 & \langle x,x'\rangle \\ \langle x,x'\rangle & 1 \end{pmatrix}$, and show that

$$\text{(I)} \to \mathbb{E}\left[\sigma\left(U^{(\ell-1)}(x)\right)\sigma\left(U^{(\ell-1)}(x')\right)\right]. \tag{16}$$

For (II), conditioning on weight matrices $\left\{\mathbf{W}^{(k)}\right\}_{k=1}^{L}$, we have

$$\text{(II)} \to \mathbb{E}_a\left[a^\top \mathbf{G}_L^{(\ell)} a\right] = \mathrm{Tr}(\mathbf{G}_L^{(\ell)}).$$

Moreover, crucially $\mathrm{Tr}(\mathbf{G}_L^{(\ell)})$ approximately satisfies a recursion. In particular, by the definition of $\mathbf{G}_L^{(\ell)}$, for any fixed $\ell \le L$,

$$\begin{aligned}
&\mathrm{Tr}\left(\mathbf{G}_{L+1}^{(\ell)}(x,x')\right) \\
&= \mathrm{Tr}\left(\mathbf{D}^{(L+1)}(x)\mathbf{W}^{(L+1)}\mathbf{G}_L^{(\ell)}(x,x')\left[\mathbf{W}^{(L+1)}\right]^\top \mathbf{D}^{(L+1)}(x')\right) \\
&= \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}_{\left\{\langle w_i^{(L+1)}, o^{(L)}(x)\rangle \ge 0\right\}} \mathbf{1}_{\left\{\langle w_i^{(L+1)}, o^{(L)}(x')\rangle \ge 0\right\}} \left[w_i^{(L+1)\top} \mathbf{G}_L^{(\ell)}(x,x') w_i^{(L+1)}\right] \\
&\to \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}_{\left\{\langle w_i^{(L+1)}, o^{(L)}(x)\rangle \ge 0\right\}} \mathbf{1}_{\left\{\langle w_i^{(L+1)}, o^{(L)}(x')\rangle \ge 0\right\}} \mathrm{Tr}\left(\mathbf{G}_L^{(\ell)}(x,x')\right),
\end{aligned} \tag{17}$$

where the last assertion holds because $w^\top \mathbf{G}_L^{(\ell)}(x,x')w$ concentrates on its mean $\mathrm{Tr}\left(\mathbf{G}_L^{(\ell)}(x,x')\right)$.

When $\ell = L+1$, we know $\mathbf{V}_{L+1}^{(L+1)}(x) = \frac{1}{\sqrt{m}}\mathbf{D}^{(L+1)}(x)$ in view of (8). From (12), we get $\mathbf{G}_{L+1}^{(L+1)}(x,x') = \frac{1}{m}\mathbf{D}^{(L+1)}(x)\mathbf{D}^{(L+1)}(x')$ and

$$\mathrm{Tr}\left(\mathbf{G}_{L+1}^{(L+1)}(x,x')\right) = \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}_{\left\{\langle w_i^{(L+1)}, o^{(L)}(x)\rangle \ge 0\right\}} \mathbf{1}_{\left\{\langle w_i^{(L+1)}, o^{(L)}(x')\rangle \ge 0\right\}}.$$

Furthermore,

$$\begin{aligned}
&\frac{1}{m}\sum_{i=1}^{m} \mathbf{1}_{\left\{\langle w_i^{(L+1)}, o^{(L)}(x)\rangle \ge 0\right\}} \mathbf{1}_{\left\{\langle w_i^{(L+1)}, o^{(L)}(x')\rangle \ge 0\right\}} \\
&\to \mathbb{E}_{w\sim\mathcal{N}(0,\mathbf{I})}\left[\mathbf{1}_{\left\{\langle w, o^{(L)}(x)\rangle \ge 0\right\}} \mathbf{1}_{\left\{\langle w, o^{(L)}(x')\rangle \ge 0\right\}}\right] \\
&\to \frac{\pi - \arccos \rho^{(L)}(x,x')}{2\pi},
\end{aligned} \tag{18}$$

where the first step holds by conditioning on $o^{(L)}$, and the last line follows as

$$\left\langle \frac{o^{(L)}(x)}{\left\|o^{(L)}(x)\right\|_2}, \frac{o^{(L)}(x')}{\left\|o^{(L)}(x')\right\|_2} \right\rangle \to \frac{\mathbb{E}\left[\sigma(U^{(L)}(x))\sigma(U^{(L)}(x'))\right]}{\sqrt{\mathbb{E}\left[\sigma^2(U^{(L)}(x))\right]}\sqrt{\mathbb{E}\left[\sigma^2(U^{(L)}(x'))\right]}} \triangleq \rho^{(L)}(x,x').$$

Therefore, by defining

$$q_{L+1}^{(\ell)}(x,x') = \frac{\pi - \arccos\rho^{(L)}(x,x')}{2\pi} q_L^{(\ell)}(x,x'), \quad \forall \ell \le L,$$
$$q_{L+1}^{(L+1)}(x,x') = \frac{\pi - \arccos\rho^{(L)}(x,x')}{2\pi}, \tag{19}$$

we get that

$$\text{(II)} \to q_L^{(\ell)}(x,x'). \tag{20}$$

Combining (16) and (20), we get that

$$H^{(\ell)}(x,x') \to \mathbb{E}\left[\sigma\left(U^{(\ell-1)}(x)\right)\sigma\left(U^{(\ell-1)}(x')\right)\right] q_L^{(\ell)}(x,x') \triangleq \Phi^{(\ell)}(x,x'). \tag{21}$$

It has been shown in Jacot et al. (2018) that for fixed (x,x') and fixed ℓ, $H^{(\ell)}(x,x')$ converges to $\Phi^{(\ell)}(x,x')$ in probability. The following theorem strengthens their result, showing the uniform convergence of $H^{(\ell)}$ to $\Phi^{(\ell)}$ for all ℓ and characterizing the rate of the convergence.

Theorem 1 *Under Gaussian initialization, For $m \ge Cd^2\exp\left(L^2\right)$ for some constant C, there exist constants C_1, C_2 and C_3 such that, with probability at least $1-\exp\left(-C_1 m^{1/3}\right)$,*

$$\left\|H^{(\ell)} - \Phi^{(\ell)}\right\|_\infty \le C_2\left(\frac{C_3^L}{m^{1/6}} + \sqrt{\frac{dL\log m}{m}}\right), \quad \forall 1 \le \ell \le L. \tag{22}$$

Remark 2 *Theorem 1 significantly improves the concentration bounds in Du et al. (2019a). Specifically, Du et al. (2019a) only establishes the concentration of the last hidden layer $H^{(L)}(x_i,x_j)$ for a bounded number of data points $\{x_i\}_{i=1}^n$. In contrast, Theorem 1 establishes the concentration uniformly over all $x \in \mathbb{S}^{d-1}$ and for all layers $\ell \in [L]$, which is much stronger and more challenging to obtain. To see why, note that a simple pointwise control and union bounds fall short of proving the uniform concentration over all $x \in \mathbb{S}^{d-1}$. More importantly, from the definition (12), we know*

$$H^{(L)} = \langle o^{(L-1)}(x), o^{(L-1)}(x')\rangle \frac{1}{m}\sum_{i=1}^m \mathbf{1}_{\left\{\langle w_i^{(L)}, o^{(L-1)}(x)\rangle \ge 0\right\}} \mathbf{1}_{\left\{\langle w_i^{(L)}, o^{(L-1)}(x')\rangle \ge 0\right\}}.$$

is a sum of independent random variables conditioning on $o^{(L-1)}$ for which a simple concentration inequality can be applied to. In contrast, the intermediate layer $H^{(\ell)}$ with $\ell < L$ depends on not only previous hidden layers but also weight matrices and activation patterns of subsequent layers through $\mathbf{G}_L^{(\ell)}$. To overcome these challenges, we fix $\mathbf{G}_L^{(\ell)}$ and view $a^\top \mathbf{G}_L^{(\ell)}(x,x')a$ as a quadratic term. This allows us to apply Hanson-Wright inequality (Vershynin, 2019, Theorem 6.2.1) and obtain the concentration of $a^\top \mathbf{G}_L^{(\ell)}(x,x')a$ for any fixed x,x'. To upgrade this point-wise concentration to the uniform one, we utilize the critical observation that the number of different $\mathbf{G}_L^{(\ell)}(x,x')$ for fixed $\left\{\mathbf{W}^{(\ell)}\right\}_{\ell=1}^L$ depends on the number of activation patterns $|\mathcal{D}_L|$ where $\mathcal{D}_L \triangleq \left\{\left(\mathbf{D}^{(1)}(x),\cdots,\mathbf{D}^{(L)}(x)\right) : x \in \mathbb{S}^{d-1}\right\}$. We then show $|\mathcal{D}_L| \le m^{dL}$ through showing $|\mathcal{D}_k| \le m^d|\mathcal{D}_{k-1}|$. See Lemma 7 for more details.

Implications in batch setting Theorem 1 implies that if $m = \Omega\left(\exp(L^2)\mathrm{poly}(d, \frac{1}{\epsilon})\right)$, then $\|H - \Phi\|_\infty < \epsilon$ with high probability. Interestingly, our uniform bounds enable us to derive a sufficient condition on the over-parameterization m in the batch setting that is independent of the batch size. Specifically, in the batch setting with data points $\{x_i\}_{i=1}^n$, by defining kernel matrices $\mathbf{H} = \left(\frac{1}{n}H(x_i, x_j)\right) \in \mathbb{R}^{n\times n}$ and $\mathbf{\Phi} = \left(\frac{1}{n}\Phi(x_i, x_j)\right) \in \mathbb{R}^{n\times n}$, we can deduce that $\|\mathbf{H} - \mathbf{\Phi}\|_{\mathrm{F}} \le \|H - \Phi\|_\infty < \epsilon$. In contrast, the previous works in the batch setting Du et al. (2019b,a); Su and Yang (2019) require m to grow sufficiently fast in n to ensure $\|\mathbf{H} - \mathbf{\Phi}\|_{\mathrm{F}} \le \epsilon$. For example, Du et al. (2019b) requires that $m = \Omega(n^6)$.

In addition to the above application, Theorem 1 also plays an important role in the analysis of gradient descent dynamic. Existing work (Du et al., 2019a) shows the training error under GD decays at the rate of $(1 - \eta\lambda_{\min}(\mathbf{\Phi}^{(L)})/2)^t$ where $\mathbf{\Phi}^{(L)} = \left(\frac{1}{n}\Phi^{(L)}(x_i, x_j)\right)$ is the limit of the NTK matrix from the last hidden layer as the number of neurons goes to infinity. With Theorem 1, we are able to show a tighter rate $(1 - \eta\lambda_{\min}(\mathbf{\Phi})/2)^t$.

Beyond the application in batch setting, Theorem 1 further enables us to characterize the convergence of the prediction error under SGD in the streaming data setting, as we shall present next.

4.2 Average Prediction Error under SGD

Define the prediction error $\Delta_t(x) \triangleq f^*(x) - f(x; \mathbf{W}(t))$. We aim to characterize the convergence of the average prediction error $\|\Delta_t\|_2 \triangleq \sqrt{\mathbb{E}_X\left[\Delta_t^2(X)\right]}$.

To analyze $\|\Delta_t\|_2$, we first show a linear approximation of Δ_t:

$$\Delta_{t+1} = (\mathsf{I} - \eta_t \mathsf{H}_t)\Delta_t + v_t + \epsilon_t, \tag{23}$$

where I is the identity operator, H_t is the integral operator associated with the kernel function $H_t(x, x')$, v_t is the noise from the stochastic gradient, and ϵ_t is the approximation error.

Note that H_t depends on $\{\mathbf{W}(s), s \le t\}$ and hence further depends on the sample path $\{X_s, y_s\}_{s=0}^{t-1}$. To circumvent this dependency, we first show $\mathbf{W}(t)$ stays relatively close to $\mathbf{W}(0)$ in operator norm under the over-parameterized regime with large m. This further allows us to show $\|H_t - H\|_\infty$ is small. Applying the triangle inequality together with Theorem 1, we deduce that $\|H_t - \Phi\|_\infty$ is small. It then follows from (23) that the prediction error under SGD can be approximated by a linear dynamic governed by Φ for any sample path $\{X_t, y_t\}$:

$$\Delta_{t+1} = (\mathsf{I} - \eta_t \mathsf{\Phi})\Delta_t + \eta_t\left(\mathsf{\Phi} - \mathsf{H}_t\right)\Delta_t + v_t + \epsilon_t, \tag{24}$$

where $\mathsf{\Phi}$ is the integral operator associated with Φ defined in (21). This recursion reveals that the evolution of Δ_t is governed by the spectrum of $\mathsf{\Phi}$.

More specifically, denote the eigenvalues of $\mathsf{\Phi}$ as $\{\lambda_i\}_{i=1}^\infty$ with $\lambda_1 \ge \lambda_2 \ge \cdots$ and the corresponding eigen-functions ϕ_i. For any function $g \in L_2(\mu)$, denote the residual projection error $\mathcal{R}(g, r)$ as the L_2 norm of the projection of g onto the space spanned by eigen-functions $\{\phi_i\}_{i=r+1}^\infty$, i.e.,

$$\mathcal{R}(g, r) = \sqrt{\sum_{i=r+1}^{\infty} \langle g, \phi_i\rangle^2}. \tag{25}$$

Theorem 3 *Given* $m \geq C_3 d^7 \exp(\theta C^L \log T)$ *and* $\eta_t = \frac{\theta}{t+1}$ *for* $\theta < \frac{9}{2\sqrt{44L}}$, *with probability at least* $1 - \exp\left(-C_4^{-L} m^{1/36}\right)$ *over the initialization, we have*

$$\mathbb{E}\left[\|\Delta_t\|_2\right] \leq \inf_{\ell} \left\{ \left(\prod_{s=0}^{t-1} (1 - \eta_s \lambda_\ell) \right) \|\Delta_0\|_2 + \mathcal{R}(\Delta_0, \ell) \right\} + 2c_2 \|\Delta_0\|_2 + 2c_2 \tau, \tag{26}$$

where $c_2 = \theta L e^{\sqrt{44L}\theta/9} \sqrt{\frac{1}{1 - 2\sqrt{44L}\theta/9} + 1}$.

Here, the first term on the right hand side of (26) comes from the linear approximation $\Delta_{t+1} \approx (\mathsf{I} - \eta_t \Phi) \Delta_t \approx \prod_{s=0}^{t} (\mathsf{I} - \eta_s \Phi) \Delta_0$ in view of (24). The term $2c_2 \|\Delta_0\|_2 + 2c_2\tau$ is the sum of three errors. One is the accumulation of the perturbation error v_t from the stochastic gradients. Another is the accumulation of the approximation error ϵ_t from the use of the linear approximation. The last one is the accumulation of the approximation error $\eta_t (\Phi - \mathsf{H}_t) \Delta_t$.

From Theorem 3, we see that an early stopping time T, which is commonly used (Su and Yang, 2019; Allen-Zhu et al., 2019a), is needed to ensure the condition on the number of neurons per layer m is satisfied. Intuitively, this dependency on T comes from two aspects. Firstly, to ensure the linear approximation holds, we crucially require $\mathbf{W}(t)$ to be close to $\mathbf{W}(0)$, resulting in an upper bound on the number of SGD iterations T. Secondly, the accumulation of the approximation error ϵ_t, albeit vanishing in m, grows in the number of iterations T. Thus to ensure the final approximation error is small, we need m to be sufficiently large compared with T.

Our result sheds light on the trade-off between the convergence rate and the accumulation of approximation errors. The trade-off is two-fold. One is between $\prod_{s=0}^{t} (1 - \eta_s \lambda_\ell)$ and $\mathcal{R}(\Delta_0, \ell)$ through ℓ. Intuitively, on one hand, a larger ℓ implies a larger principal space which yields a smaller $\mathcal{R}(\Delta_0, \ell)$. On the other hand, a larger ℓ also implies a smaller λ_ℓ. Thus, the contraction factor $\prod_{s=0}^{t} (1 - \eta_s \lambda_\ell)$ is smaller, indicating slower convergence. The other trade-off is between the contraction factor $\prod_{s=0}^{t} \left(1 - \frac{\theta \lambda_\ell}{s+1}\right)$ and the accumulation of approximation error and noise c_2 through θ. To make sure c_2 is small, we need small θ, thus yielding a small contraction factor. In return, we need more iterations to converge.

Now we present an application of Theorem 3 when f^* is a polynomial. Consider SGD under a symmetric initialization scheme of the last layer, i.e., $\mathbf{W}^{(L)}(0) = \begin{pmatrix} \mathbf{W} \\ \mathbf{W} \end{pmatrix}$ where $\mathbf{W} \in \mathbb{R}^{\frac{m}{2} \times m}$ is a random matrix with i.i.d. standard normal entries and $a = (b, -b)^\top$ where $b \in \mathbb{R}^{m/2}$ has i.i.d. Rademacher entries.

Corollary 4 *Assume* f^* *is a degree* ℓ^* *polynomial and the input data follows the uniform distribution over* $\mathbb{S}^{d-1}$. *Under the same condition as Theorem 3, we have with probability at least* $1 - \exp\left(-\Omega(C_4^{-L} m^{1/36})\right)$,

$$\mathbb{E}\left[\|\Delta_{t+1}\|_2 \,|\mathbf{W}(0), a\right] \leq \prod_{s=0}^{t} (1 - \eta_s \lambda_{\ell^*+1}) \|f^*\|_2 + 2c_2 \|f^*\|_2 + 2c_2 \tau. \tag{27}$$

The proof is deferred to Appendix E.

Remark 5 *From Corollary 4, for arbitrarily small constant ϵ, by choosing small step sizes, a sufficiently long horizon and a sufficient wide neural network, we ensure that the average prediction error under SGD is smaller than ϵ. To be more specific, for any $0 < \epsilon < \|f^*\|_2 + \tau$, by choosing $T \geq \left(\frac{\epsilon}{6\|f^*\|_2}\right)^{-1/(\theta\lambda_{\ell^*+1})}$ and $\theta \leq \frac{9\epsilon}{8\sqrt{44}(\|f^*\|_2+\tau)L}$, we ensure $\mathbb{E}\left[\|\Delta_{t+1}\|_2 | \mathbf{W}(0), a\right] \leq \epsilon$. To see why, note that*

$$\prod_{s=0}^{t}(1-\eta_s\lambda_{\ell^*+1}) \leq \exp\left(-\theta\lambda_{\ell^*+1}\log T\right) = T^{-\theta\lambda_{\ell^*+1}} \leq \frac{\epsilon}{6\|f^*\|_2}, \tag{28}$$

and

$$\begin{aligned} c_2\left(\|f^*\|_2+\tau\right) &= \theta L e^{\sqrt{44}L\theta/9}\sqrt{\frac{1}{1-2\sqrt{44}L\theta/9}+1}\left(\|f^*\|_2+\tau\right) \\ &\overset{(a)}{\leq} \frac{9\epsilon}{8\sqrt{44}}e^{1/8}\sqrt{7/3} \leq \frac{5}{12}\epsilon, \end{aligned} \tag{29}$$

where (a) holds since $\frac{2\sqrt{44}L\theta}{9} \leq \frac{\epsilon}{4\left(\|f^*\|_2+\tau\right)} < \frac{1}{4}$. The result follows by plugging (28) and (29) into (27).

5. Proof of Theorem 1

In this section, we present the proof of the main results.

Additional notation Define $\mathsf{VC}(\mathcal{F})$ as the VC dimension of Boolean function class $\mathcal{F}$. For any matrix $\mathbf{C} \in \mathbb{R}^{n\times m}$, we define $\|\mathbf{C}\|_\infty \triangleq \max_{1\leq i\leq n, 1\leq j\leq m}|\mathbf{C}_{ij}|$. Throughout the remaining paper, we use C to denote absolute constant whose value may vary in lines.

We present several key lemmas that will be used in the proof of Theorem 1. First, we show that $\langle o^{(\ell)}(x), o^{(\ell)}(x')\rangle$ concentrates on $\mathbb{E}\left[\sigma\left(U^{(\ell)}(x)\right)\sigma\left(U^{(\ell)}(x')\right)\right]$ uniformly over all $x, x' \in \mathbb{S}^{d-1}$ and all $\ell \in [L]$.

Lemma 6 *With probability at least $1 - L\exp\left(O\left(d\log m\right) - \Omega(m^{1/3})\right)$, for any $1 \leq \ell \leq L$,*

$$\sup_{x,x'}\left|\langle o^{(\ell)}(x), o^{(\ell)}(x')\rangle - \mathbb{E}\left[\sigma\left(U^{(\ell)}(x)\right)\sigma\left(U^{(\ell)}(x')\right)\right]\right| = O\left(\frac{\ell C^{2\ell}}{m^{1/3}}\right), \tag{30}$$

where $(U^{(\ell)}(x), U^{(\ell)}(x'))$ is defined in (15).

To prove Lemma 6, we follow the aforementioned heuristic in Section 4.1 to show that $\langle o^{(\ell)}(x), o^{(\ell)}(x')\rangle$ concentrates on $\mathbb{E}\left[\sigma(U^{(\ell)}(x))\sigma(U^{(\ell)}(x'))\right]$ for any fixed (x, x'). Then we establish that $o^{(\ell)}(x)$ is Lipschitz in x with high probability. This enables us to apply an ϵ-net argument to upgrade the pointwise concentration to the uniform one.

The next two lemmas together show that $a^\top \mathbf{G}_L^{(\ell)}(x, x')a$ uniformly concentrates on $q_L^{(\ell)}(x, x')$.

Lemma 7 *With probability at least $1 - \exp\left(O(dL\log m) - \Omega(m^{1/3})\right)$, for $\ell = 1, 2, \cdots, L$,*

$$\sup_{x,x'}\left|a^\top \mathbf{G}_L^{(\ell)}(x, x')a - \mathrm{Tr}\left(\mathbf{G}_L^{(\ell)}(x, x')\right)\right| = O\left(\frac{c_0^{2L-2\ell}}{m^{1/3}}\right). \tag{31}$$

The above lemma shows the uniform concentration of $a^\top \mathbf{G}_L^{(\ell)}(x,x')a$ on $\mathrm{Tr}(\mathbf{G}_L^{(\ell)}(x,x'))$. However, unlike the previous case, an ϵ-net argument cannot be applied here, as x influences $\mathbf{G}_L^{(\ell)}(x,x')$ through non-Lipschitz indicator functions $\mathbf{1}_{\left\{\langle w^{(k+1)}, o^{(k)}(x)\rangle \geq 0\right\}}$ for $k \geq \ell$. As mentioned in Remark 2, the key to overcome this challenge lies on the following crucial observation. Although there are infinite number of different matrices $\mathbf{G}_L^{(\ell)}(x,x')$ when varying x, x', conditioning on $\left\{\mathbf{W}^{(k)}\right\}_{k=1}^L$ the size of $\mathcal{G}_L^{(\ell)} \triangleq \left\{\mathbf{G}_L^{(\ell)}(x,x') : x, x' \in \mathbb{S}^{d-1}\right\}$ depends only on the size of

$$\mathcal{D}_L \triangleq \left\{\left(\mathbf{D}^{(1)}(x), \cdots, \mathbf{D}^{(L)}(x)\right) : x \in \mathbb{S}^{d-1}\right\}.$$

Since $\mathbf{D}^{(k)} \in \mathbb{R}^{m\times m}$ is diagonal with binary entries, one can directly bound $|\mathcal{D}_L|$ by 2^{mL}. Unfortunately, such naive bound is too loose to obtain a tight concentration. Instead, we show a much tighter bound $|\mathcal{D}_L| \leq m^{dL}$ utilizing the recursive relation $|\mathcal{D}_k| \leq m^d |\mathcal{D}_{k-1}|$ for all k. To obtain such recursive relation, a critical step is to decompose $\mathbb{S}^{d-1}$ into disjoint regions $\{V_j, j = 1, 2, \cdots, |\mathcal{D}_{k-1}|\}$ so that for any x within the same V_j, $\left(\mathbf{D}^{(1)}(x), \cdots, \mathbf{D}^{(k-1)}(x)\right)$ is the same. With such decomposition, we can get

$$|\mathcal{D}_k| \leq \sum_{j=1}^{|\mathcal{D}_{k-1}|} \left|\left\{\mathbf{D}^{(k)}(x) : x \in V_j\right\}\right|.$$

To further bound $\left|\left\{\mathbf{D}^{(k)}(x) : x \in V_j\right\}\right|$, we crucially utilize the fact that for any fixed j, $o^{(k-1)}(x) = P_j x$ for all $x \in V_j$ with some deterministic matrix $P_j \in \mathbb{R}^{m\times d}$ independent of x. Hence, $\left|\left\{\mathbf{D}^{(k)}(x) : x \in V_j\right\}\right| \leq m^d$ follows by applying Hajek and Raginsky (2019, Proposition 7.1) and Sauer-Shelah Lemma (Lemma 22). With this tighter bound in hand, we deduce the uniform concentration of $a^\top \mathbf{G}_L^{(\ell)}(x,x')a$ on its mean $\mathrm{Tr}\left(\mathbf{G}_L^{(\ell)}(x,x')\right)$ by combining Hanson-Wright inequality with a union bound over $\mathcal{G}_L^{(\ell)}$.

It remains to show the uniform concentration of $\mathrm{Tr}\left(\mathbf{G}_L^{(\ell)}(x,x')\right)$ on $q_L^{(\ell)}(x,x')$.

Lemma 8 *With probability at least* $1 - \exp\left(O(dL\log m) - \Omega(m^{1/3})\right)$, *for* $\ell = 1, 2, \cdots, L$,

$$\sup_{x,x'} \left|\mathrm{Tr}\left(\mathbf{G}_L^{(\ell)}(x,x')\right) - q_L^{(\ell)}(x,x')\right| = O\left(\frac{\sqrt{L}C^L}{m^{1/6}} + \sqrt{\frac{d\left(1 + (L-1)\log m\right)}{m}}\right) \tag{32}$$

for some universal constant C.

To prove Lemma 8, we follow the heuristic argument in Section 4.1 to prove (17) and (18). The proof of (17) follows similarly as that of Lemma 7. To prove the first step of (18), we utilize the following observation. Conditioning on $\mathbf{W}^{(1)}, \cdots, \mathbf{W}^{(L-1)}$, the change of $\sup_{x,x'} h^{(L)}(x,x')$ from the change of any single coordinate is bounded by $\frac{1}{m}$, where

$$h^{(L)}(x,x') \triangleq \left|\frac{1}{m}\sum_{i=1}^m \mathbf{1}_{\left\{\langle w_i^{(L)}, o^{(L-1)}(x)\rangle \geq 0\right\}} \mathbf{1}_{\left\{\langle w_i^{(L)}, o^{(L-1)}(x')\rangle \geq 0\right\}} - \mathbb{E}_w\left[\mathbf{1}_{\left\{\langle w, o^{(L-1)}(x)\rangle \geq 0\right\}} \mathbf{1}_{\left\{\langle w, o^{(L-1)}(x)\rangle \geq 0\right\}}\right]\right|.$$

This allows us to apply McDiarmid's inequality to show with high probability over the randomness of $\left\{w_i^{(L)}\right\}_{i=1}^m$, $\sup_{x,x'} h^{(L)}(x,x')$ concentrates on its mean. We then apply Lemma 21 to bound $\mathbb{E}\left[\sup_{x,x'} h^{(L)}(x,x')\right]$ by $O\left(\sqrt{\frac{\mathsf{VC}(\mathcal{H}^{(L)})}{m}}\right)$ where

$$\mathcal{H}^{(L)} \triangleq \left\{f_{x,x'}(w) = \mathbf{1}_{\left\{\langle w, o^{(L-1)}(x)\rangle \geq 0\right\}} \mathbf{1}_{\left\{\langle w, o^{(L-1)}(x')\rangle \geq 0\right\}} : x, x' \in \mathbb{S}^{d-1}\right\}.$$

Afterwards, we apply Lemma 20 to show $\mathsf{VC}(\mathcal{H}^{(L)}) = O\left(\mathsf{VC}(\mathcal{F}^{(L)})\right)$ where

$$\mathcal{F}^{(L)} \triangleq \left\{f_x(w) = \mathbf{1}_{\left\{\langle w, o^{(L-1)}(x)\rangle \geq 0\right\}} : x \in \mathbb{S}^{d-1}\right\}.$$

To bound $\mathsf{VC}(\mathcal{F}^{(L)})$, we follow a similar decomposition strategy as Lemma 7 and show

$$\mathsf{VC}(\mathcal{F}^{(L)}) = O(d(1 + (L-1)\log m)).$$

To prove the second step of (18), we crucially establish that the arccos function is Hölder continuous of order 1/2 despite that it is non-Lipschitz.

With the above lemmas, we now present the proof of Theorem 1. The full proofs of Lemma 6– 8 are deferred to Appendix B.

Proof [Proof of Theorem 1] Throughout the proof, we condition on the event such that (30), (31) and (32) hold simultaneously. By Lemma 6–Lemma 8, we get such event occurs with probability at least $1 - \exp\left(-\Omega(m^{1/3})\right)$ for sufficiently large m.

For any $1 \leq \ell \leq L$, by the triangle inequality, we have

$$\begin{aligned}
&\left\|H^{(\ell)} - \Phi^{(\ell)}\right\|_\infty \\
&= \sup_{x,x'} \left|\langle o^{(\ell-1)}(x), o^{(\ell-1)}(x')\rangle a^\top \mathbf{G}_L^{(\ell)}(x,x') a - \mathbb{E}\left[\sigma\left(U^{(\ell-1)}(x)\right)\sigma\left(U^{(\ell-1)}(x')\right)\right] q_L^{(\ell)}(x,x')\right| \\
&\leq \sup_{x,x'} \left|\left(\langle o^{(\ell-1)}(x), o^{(\ell-1)}(x')\rangle - \mathbb{E}\left[\sigma\left(U^{(\ell-1)}(x)\right)\sigma\left(U^{(\ell-1)}(x')\right)\right]\right) a^\top \mathbf{G}_L^{(\ell)}(x,x') a\right| \\
&+ \sup_{x,x'} \left|\mathbb{E}\left[\sigma\left(U^{(\ell-1)}(x)\right)\sigma\left(U^{(\ell-1)}(x')\right)\right]\left(a^\top \mathbf{G}_L^{(\ell)}(x,x') a - q_L^{(\ell)}(x,x')\right)\right|. \qquad (33)
\end{aligned}$$

Here, we claim that

$$\sup_{x,x'} \left|a^\top \mathbf{G}_L^{(\ell)}(x,x') a\right| \leq 1, \qquad (34)$$

and

$$\sup_{x,x'} \mathbb{E}\left[\sigma(U^{(\ell)}(x))\sigma(U^{(\ell)}(x'))\right] \leq \sup_x \sqrt{\mathbb{E}\left[\sigma^2(U^{(\ell)}(x))\right]} \sup_{x'} \sqrt{\mathbb{E}\left[\sigma^2(U^{(\ell)}(x'))\right]} = 2^{-\ell} \leq 1. \qquad (35)$$

Plugging the above two claims into (33), we have

$$\begin{aligned}
&\left\|H^{(\ell)}-\Phi^{(\ell)}\right\|_{\infty} \\
&\leq \sup_{x,x'}\left|\langle o^{(\ell-1)}(x), o^{(\ell-1)}(x')\rangle - \mathbb{E}\left[\sigma\left(U^{(\ell-1)}(x)\right)\sigma\left(U^{(\ell-1)}(x')\right)\right]\right| + \sup_{x,x'}\left|a^{\top}\mathbf{G}_L^{(\ell)}(x,x')a - q_L^{(\ell)}(x,x')\right| \\
&\overset{(a)}{=} O\left(\frac{\ell C^{2\ell}}{m^{1/3}}\right) + O\left(\frac{C^L}{m^{1/6}} + \sqrt{\frac{d\left(1+(L-1)\log m\right)}{m}}\right) \\
&= O\left(\frac{C^L}{m^{1/6}} + \sqrt{\frac{d\left(1+(L-1)\log m\right)}{m}}\right),
\end{aligned}$$

where (a) holds by (30), (31), and (32); and the last equality holds since $m = \Omega(\exp\left(L^2\right))$.

It remains to prove (34) and (35). To prove (34), by definition (19), we have

$$0 \leq q_L^{(\ell)}(x,x') \leq 1/2. \tag{36}$$

Therefore, by the triangle inequality, we have

$$\sup_{x,x'}\left|a^{\top}\mathbf{G}_L^{(\ell)}(x,x')a\right| \leq \sup_{x,x'}\left|a^{\top}\mathbf{G}_L^{(\ell)}(x,x')a - q_L^{(\ell)}(x,x')\right| + \sup_{x,x'}\left|q_L^{(\ell)}(x,x')\right| \leq 1,$$

where the last inequality holds since $O\left(\frac{\sqrt{L}C^L}{m^{1/6}} + \sqrt{\frac{d(1+(L-1)\log m)}{m}}\right) \leq \frac{1}{2}$ given $m = \Omega(d^2\exp(L^2)))$.

Now we prove (35). Since $U^{(1)}(x) = \langle w, x\rangle \sim \mathcal{N}(0,1)$ for any x, we have

$$\mathbb{E}\left[\sigma^2(U^{(1)}(x))\right] = \mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left[Z^2\mathbf{1}_{\{Z\geq 0\}}\right] = \frac{1}{2}, \quad \forall x. \tag{37}$$

By the definition of $\Sigma^{(\ell)}$, it follows that

$$\begin{aligned}
\mathbb{E}\left[\sigma^2(U^{(\ell)}(x))\right] &= \mathbb{E}_{U^{(\ell-1)}(x)}\left[\mathbb{E}_{Z\sim\mathcal{N}\left(0,\sigma^2(U^{(\ell-1)})(x)\right)}\left[Z^2\mathbf{1}_{\{Z\geq 0\}}|U^{(\ell-1)}(x)\right]\right] \\
&= \frac{1}{2}\mathbb{E}\left[\sigma^2(U^{(\ell-1)}(x))\right], \quad \forall x.
\end{aligned}$$

Recursively applying the above equality and noting (37), we get that

$$\mathbb{E}\left[\sigma^2(U^{(\ell)}(x))\right] = 2^{-\ell}, \quad \forall x. \tag{38}$$

By Cauchy-Schwartz inequality, we have

$$\sup_{x,x'}\mathbb{E}\left[\sigma(U^{(\ell)}(x))\sigma(U^{(\ell)}(x'))\right] \leq \sup_{x}\sqrt{\mathbb{E}\left[\sigma^2(U^{(\ell)}(x))\right]}\sup_{x'}\sqrt{\mathbb{E}\left[\sigma^2(U^{(\ell)}(x'))\right]} = 2^{-\ell} \leq 1.$$

■

6. Bounding $\|H_t - H_0\|_\infty$

In this section, we prove that with high probability, for any sample path $\{x_s, y_s\}_{s=0}^{T-1}$, $\|H_t - H_0\|_\infty$ is small. As discussed in Section 4.2, this is crucial to the analysis of the average prediction error under SGD in the streaming data setup.

Recall from (1) that $H_t = \sum_{\ell=1}^{L} H_t^{(\ell)}$ and

$$H_t^{(\ell)}(x, x') = \left\langle \frac{\partial f(x; \mathbf{W}(t))}{\partial \mathbf{W}^{(\ell)}}, \frac{\partial f(x'; \mathbf{W}(t))}{\partial \mathbf{W}^{(\ell)}} \right\rangle.$$

where

$$\frac{\partial f(x; \mathbf{W}(t))}{\partial \mathbf{W}^{(\ell)}} = \frac{1}{\sqrt{m}} \mathbf{D}_t^{(\ell)}(x) z_t^{(\ell)}(x) \left[o_t^{(\ell-1)}(x) \right]^\top, \tag{39}$$

and $z_t^{(\ell)}(x)$ measures the sensitivity of the output from the ℓ-th hidden layer defined as

$$\left[z_t^{(\ell)}(x) \right]^\top \triangleq \left[\frac{\partial f(x; \mathbf{W}(t))}{\partial o^{(\ell)}(x)} \right]^\top = a^\top \frac{1}{\sqrt{m}} \mathbf{D}_t^{(L)}(x) \mathbf{W}^{(L)}(t) \cdots \frac{1}{\sqrt{m}} \mathbf{D}_t^{(\ell+1)}(x) \mathbf{W}^{(\ell+1)}(t). \tag{40}$$

Throughout the section, we assume the width of each hidden layer m satisfies

$$m \geq d^9 \exp\left(\Omega(\theta L C^L \log T) \right) \tag{41}$$

for some absolute constant C. Also, recall from Section 3 that we assume $\gamma \triangleq \max\{\|f^*\|_\infty, |u|\}$ is independent of m and choose step size $\eta_t \leq \frac{\theta}{t+1}$.

Proposition 9 *Assume (41) holds. With probability* $1 - \exp\left(-\Omega(C^{-L} m^{1/36})\right)$*, for any sample path* $\{x_s, y_s\}_{s=0}^{T-1}$*, all* $t \leq T$*, and all* $1 \leq \ell \leq L$*, we have*

$$\sup_x \left\| \frac{\partial f(x; \mathbf{W}(t))}{\partial \mathbf{W}^{(\ell)}} - \frac{\partial f(x; \mathbf{W}(0))}{\partial \mathbf{W}^{(\ell)}} \right\|_2 = O\left(\frac{C^L}{m^{1/36}} \right),$$

and hence,

$$\left\| H_t^{(\ell)} - H_0^{(\ell)} \right\|_\infty = O\left(\frac{C^L}{m^{1/36}} \right). \tag{42}$$

To prove Proposition 9, in view of (39), the key is to control the deviations of $\mathbf{D}_t^{(\ell)}(x)$, $z_t^{(\ell)}(x)$ and $o_t^{(\ell-1)}(x)$ uniformly, which will be done in the following Lemma 10–12. The detailed proof of Proposition 9 and Lemma 10–12 are deferred to Appendix C.

We begin with bounding the deviation of $o_t^{(\ell-1)}(x)$. Define a sequence of real numbers:

$$\begin{aligned} R_0 &\triangleq m^{5/18}, \\ R_{t+1} &\triangleq R_0 + L C^{2L-2} \sum_{s=0}^{t} \eta_s (R_s + \gamma), t \geq 1. \end{aligned} \tag{43}$$

Lemma 10 *Assume (41) holds. Then we have $R_t \le m^{1/3}$ for all $t \le T$. Moreover, with probability at least $1 - \exp\left(-\Omega(C_1^{-L} m^{1/9})\right)$, for any $1 \le \ell \le L$, $t \le T$ and sample path $\{x_s, y_s\}_{s=0}^{T-1}$, the following holds:*

$$\left\|\mathbf{W}^{(\ell)}(t) - \mathbf{W}^{(\ell)}(0)\right\|_2 \le C_2^{L-1} \sum_{s=0}^{t-1} \eta_s \left(R_s + \gamma\right) \le R_t \tag{44}$$

$$\sup_x \left\|o_t^{(\ell)}(x) - o_0^{(\ell)}(x)\right\|_2 \le \frac{C_3^\ell}{m^{1/6}} \tag{45}$$

$$\sup_x |\Delta_t(x)| \le R_t, \tag{46}$$

for some absolute constant C_1, C_2 and C_3.

Lemma 10 is proved via induction over t in Appendix C.2. A key underlying idea is as follows. While the weight vectors for some individual neurons may exhibit large deviations, collectively $\mathbf{W}^{(\ell)}(t)$ is close to $\mathbf{W}^{(\ell)}(0)$ in terms of the spectral norm, or equivalently the Frobenius norm as $\mathbf{W}^{(\ell)}(t) - \mathbf{W}^{(\ell)}(0)$ is of rank no more than t, which is much smaller than m following (41). This allows us to further control the deviation of $o_t^{(\ell)}(x)$ and $|\Delta_t(x)|$ uniformly over all x, which in turn results in a small deviation of $\mathbf{W}^{(\ell)}(t+1)$ in the next iteration. Departing from the previous work (e.g. Du et al. (2019a, Lemma B.5)), here to control the deviation of $\mathbf{W}^{(\ell)}(t+1)$, it is crucial to bound $|\Delta_t(x)|$ uniformly over all x. A critical intermediate step is to bound $|f(x; \mathbf{W}(0))|$. To this end, by observing that $f^2(x; \mathbf{W}(0)) \le a^\top \mathbf{Q}(x) a$ for some matrix $\mathbf{Q}(x)$ independent of a, we apply the Hanson-Wright inequality for a fixed $\mathbf{Q}(x)$ and then apply a union bound over $\left\{\mathbf{Q}(x) : x \in \mathbb{S}^{d-1}\right\}$, analogous to the proof of Lemma 7.

Next, we show $\mathbf{D}_t^{(\ell)}(x)$ is close to $\mathbf{D}_0^{(\ell)}(x)$ for any x. As such, define

$$S_t^{(\ell)}(x) \triangleq \|\mathbf{D}_t^{(\ell)}(x) - \mathbf{D}_0^{(\ell)}(x)\|_{\mathrm{F}}.$$

Equivalently, $S_t^{(\ell)}(x)$ measures the number of sign flips of the neurons at the ℓ-th layer.

Lemma 11 *Assume (41) holds. Then with probability $1 - \exp\left(-\Omega\left(C_1^{-L} m^{1/9}\right)\right)$ for any $1 \le \ell \le L$, $t \le T$ and sample path $\{x_s, y_s\}_{s=0}^{t-1}$,*

$$\sup_x S_t^{(\ell)}(x) \le C_2^\ell m^{8/9}, \tag{47}$$

for some absolute constant C_1 and C_2.

Note that the previous work Du et al. (2019b) has obtained bounds to the number of sign flips in the batch setting with one hidden layer. They crucially require every individual weight vector $w_i^{(1)}$ not to change much and hence only the neurons with small $|\langle w_i^{(1)}(0), x\rangle|$ can have sign flips. However, in our setting, we need to further bound the number of neurons with relatively large deviations based on our bound of $\left\|\mathbf{W}^{(\ell)}(t) - \mathbf{W}^{(\ell)}(0)\right\|_2$.

Finally, we bound the deviation of the sensitivity $z_t^{(\ell)}(x)$.

Lemma 12 *Assume (41) holds. With probability at least* $1-\exp\left(-\Omega(C_3^{-L+\ell}m^{1/36})\right)$*, for layer* ℓ *and* $t\le T$ *and any sample path* $\{x_s,y_s\}_{s=0}^{T-1}$*, we have*

$$\sup_x \left\|z_0^{(\ell)}(x)\right\|_\infty \le m^{1/36}, \tag{48}$$

and

$$\sup_x \left\|z_t^{(\ell)}(x)-z_0^{(\ell)}(x)\right\|_2 = O\left(C_4^{2L-\ell}m^{17/36}\right), \tag{49}$$

for some absolute constant C_3 *and* C_4*.*

Lemma 12 is proved via a backward induction over ℓ in Appendix C.3. In particular, we crucially utilize the following layer-wise recursive relation

$$z_t^{(\ell)}(x)=\frac{1}{\sqrt{m}}\left[\mathbf{W}^{(\ell+1)}(t)\right]^\top \mathbf{D}_t^{(\ell+1)}(x)z_t^{(\ell+1)}(x) \tag{50}$$

and apply the aforementioned deviation bounds of $\mathbf{W}^{(\ell+1)}(t)$ and $\mathbf{D}_t^{(\ell+1)}(x)$. Note that even if there is only a single sign flip at some r-th neuron, an enormous value of the r-th coordinate of $z_0^{(\ell+1)}(x)$ can possibly induce a large change of $z_t^{(\ell)}(x)$. To circumvent this issue, we derive a uniform bound to $\|z_0^{(\ell)}(x)\|_\infty$. Specifically, we observe that the r-th coordinate of $z_0^{(\ell)}(x)$ equals $\langle a, v_r^{(\ell)}(x)\rangle$ where $v_r^{(\ell)}(x)$ is the r-th column of matrix $\frac{1}{\sqrt{m}}\mathbf{D}^{(L)}(x)\mathbf{W}^{(L)}(0)\cdots\frac{1}{\sqrt{m}}\mathbf{D}^{(\ell+1)}(x)\mathbf{W}^{(\ell+1)}(0)$ in view of (40). Analogous to the proof of Lemma 7, by conditioning on $\left\{\mathbf{W}^{(k)}\right\}_{k=1}^L$, we first show the concentration of $\langle a, v_r^{(\ell)}(x)\rangle$ for a fixed $v_r^{(\ell)}(x)$ and then apply a union bound by counting the number of $v_r^{(\ell)}(x)$.

7. Proof of Theorem 3

Recall from Section 4.2 that the recursion (24) plays a key role in showing Theorem 3. In the following lemma, we prove the recursion (24). Denote operators as

$$\mathsf{K}_t=\mathsf{I}-\eta_t\Phi,\quad \mathsf{Q}_t=\mathsf{I}-\eta_t\mathsf{H}_t,\quad \mathsf{D}_t=\mathsf{Q}_t-\mathsf{K}_t, \tag{51}$$

where Φ is the integral operator associated with Φ, H_t is the integral operator associated with H_t defined in (1), and $\Phi\triangleq\sum_{\ell=1}^L\Phi^{(\ell)}$ with $\Phi^{(\ell)}$ defined in (21).

Lemma 13 *For any* t*, we have*

$$\Delta_{t+1}=\mathsf{K}_t\circ\Delta_t+\mathsf{D}_t\circ\Delta_t+v_t+\epsilon_t, \tag{52}$$

and hence,

$$\begin{aligned}\mathbb{E}\left[\|\Delta_{t+1}\|_2\,|\mathbf{W}(0),a\right]\le{}&\left\|\prod_{s=0}^t\mathsf{K}_s\circ\Delta_0\right\|_2+\sum_{r=0}^t\mathbb{E}\left[\left\|\prod_{i=r+1}^t\mathsf{Q}_i\mathsf{D}_r\prod_{j=0}^{r-1}\mathsf{K}_j\circ\Delta_0\right\|_2\Big|\mathbf{W}(0),a\right]\\&+\sum_{s=0}^t\mathbb{E}\left[\left\|\prod_{r=s+1}^t\mathsf{Q}_r\circ\epsilon_s\right\|_2\Big|\mathbf{W}(0),a\right]+\mathbb{E}\left[\left\|\sum_{s=0}^t\prod_{r=s+1}^t\mathsf{Q}_r\circ v_s\right\|_2\Big|\mathbf{W}(0),a\right].\end{aligned} \tag{53}$$

where $\epsilon_t \equiv \epsilon_t(x, X_t; \mathbf{W}(t), \mathbf{W}(t+1))$ *with*

$$\epsilon_t(x, x'; \mathbf{W}(t), \mathbf{W}(t+1)) \triangleq f(x; \mathbf{W}(t)) - f(x; \mathbf{W}(t+1)) + \eta_t H_t(x, x') \left(f^*(x') + u_t - f(x'; \mathbf{W}(t))\right)$$

and

$$v_t \equiv v_t(x, X_t) = -\eta_t \left[(\Delta_t(X_t) + u_t) H_t(x, X_t) - \mathbb{E}_{X_t}\left[\Delta_t(X_t) H_t(x, X_t) | \mathbf{W}(0), a\right]\right]. \quad (54)$$

Proof [Proof of Lemma 13] By the definition of ϵ_t, we have

$$\begin{aligned}\Delta_{t+1}(x) &= \Delta_t(x) - \eta_t H_t(x, X_t) \left(\Delta_t(X_t) + u_t\right) + \epsilon_t(x, X_t) \\ &= \Delta_t - \eta_t \mathbb{E}_{X_t}\left[H_t(x, X_t)\Delta_t(X_t) | \mathbf{W}(0), a\right] + \epsilon_t(x, X_t) + v_t(x, X_t).\end{aligned}$$

Using the notation in (51), we get the first equality of the lemma

$$\begin{aligned}\Delta_t &= \mathsf{Q}_t \circ \Delta_t + v_t + \epsilon_t \\ &= \mathsf{K}_t \circ \Delta_t + \mathsf{D}_t \circ \Delta_t + \epsilon_t + v_t,\end{aligned}$$

where the last equality holds since $\mathsf{Q}_t = \mathsf{D}_t + \mathsf{K}_t$.

Unrolling the above equality, we have

$$\Delta_{t+1} \leq \prod_{s=0}^{t} \mathsf{K}_s \circ \Delta_0 + \sum_{r=0}^{t} \prod_{i=r+1}^{t} \mathsf{Q}_i \mathsf{D}_r \prod_{j=0}^{r-1} \mathsf{K}_j \circ \Delta_0 + \sum_{s=0}^{t} \prod_{r=s+1}^{t} \mathsf{Q}_r \circ \epsilon_s + \sum_{s=0}^{t} \prod_{r=s+1}^{t} \mathsf{Q}_r \circ v_s.$$

Taking L_2 norm and conditional expectation, following the triangle inequality, we obtain the second inequality of the lemma. ■

To bound the average prediction error $\mathbb{E}\left[\|\Delta_{t+1}\|_2 | \mathbf{W}(0), a\right]$, it suffices to bound the right hand side of (53). The first term can be bounded using the eigen-decomposition of K_t. As an intermediate step, we prove both Φ and H_t are positive semi-definite with a bounded spectral norm in Lemma 14 below. This will be useful in bounding the spectrum of K_t and Q_t.

Lemma 14 Φ *is positive semi-definite with* $\|\Phi\|_2 \leq \|\Phi\|_\infty \leq \frac{L}{2}$. *Hence, for* $\eta_t \leq \frac{2}{L}$, *we have*

$$0 \leq \lambda_i(\mathsf{K}_t) \leq 1,$$

for all i *where* $\lambda_i(\mathsf{K}_t)$ *is the* i*-th largest eigen-value of* K_t.

Assume (41) holds. With probability at least $1 - \exp\left(-\Omega\left(C^{-L} m^{1/36}\right)\right)$, H_t *are positive semi-definite for all* $t \leq T$ *with* $\|\mathsf{H}_t\|_2 \leq \frac{2L}{3}$, *and hence for* $\eta_t \leq \frac{3}{2L}$,

$$0 \leq \lambda_i(\mathsf{Q}_t) \leq 1,$$

where $\lambda_i(\mathsf{Q}_t)$ *is the* i*-th largest eigen-value of* Q_t.

The second term of (53) is the approximation error of using Φ instead of H_t. To bound the second term, we apply Lemma 14 which bounds $\|\mathsf{Q}_t\|_2$ and $\|\mathsf{K}_t\|_2$ for all t. Then we apply Proposition 9 as well as Theorem 1 to bound $\|\mathsf{D}_t\|_2$.

It remains to bound the last two terms. Intuitively, the third term is the accumulation of ϵ_t and the last term is the accumulation of the noise from the stochastic gradients v_t.

The following lemma bounds the approximation error.

Lemma 15 *Assume (41) holds. With probability at least* $1-\exp\left(-\Omega(C^{-L+1}m^{1/36})\right)$*, we have*

$$\mathbb{E}\left[\|\epsilon_t\|_2 \,|\mathbf{W}(0), a\right] = O\left(\frac{\eta_t C^L \sigma_t}{m^{1/36}}\right), \tag{55}$$

where

$$\sigma_t^2 = \mathbb{E}\left[\|\Delta_t\|_2^2 \,|\mathbf{W}(0), a\right] + \tau^2. \tag{56}$$

and τ *is the variance of the noise* u*.*

Next, we bound the noise from the stochastic gradients in expectation.

Lemma 16 *Assume (41) holds. With probability at least* $1-\exp\left(-\Omega(C^{-L}m^{1/36})\right)$*, we have*

$$\mathbb{E}\left[\left\|\sum_{s=0}^{t}\prod_{r=s+1}^{t}\mathsf{Q}_r\circ v_s\right\|_2 \Big|\mathbf{W}(0), a\right] \le c_2\sigma_0, \tag{57}$$

where $c_2 = L\theta e^{\sqrt{44}L\theta/9}\sqrt{\frac{1}{1-2\sqrt{44}L\theta/9}+1}$*.*

To prove Lemma 16, we first utilize $\|\mathsf{Q}_t\|_2 \le 1$ and the observation

$$\mathbb{E}\left[v_s|\left\{X_r, y_r\right\}_{r=0}^{s-1}, \mathbf{W}(0), a\right] = 0$$

to show

$$\mathbb{E}\left[\left\|\sum_{s=0}^{t}\prod_{r=s+1}^{t}\mathsf{Q}_r\circ v_s\right\|_2^2 \Big|\mathbf{W}(0), a\right] \le \sum_{s=0}^{t}\mathbb{E}\left[\|v_s\|_2^2 \,|\mathbf{W}(0), a\right].$$

Then following the definition of v_t and the upper bound of $\|H_t\|_\infty$, we have

$$\mathbb{E}\left[\|v_s\|_2^2 \,|\mathbf{W}(0), a\right] \le \frac{4L\eta_s^2}{9}\sigma_s^2$$

where σ_s^2 is defined in (56).

Finally we bound $\sum_{s=0}^{T}\eta_s^2\sigma_s^2$. Note that $\eta_s \le \frac{\theta}{s+1}$. Therefore, by showing σ_t does not grow too fast in t, i.e., $\sigma_{t+1} \le \left(1+\frac{\sqrt{44}L\eta_t}{9}\right)\sigma_t$, we guarantee $\sum_{s=0}^{\infty}\eta_s^2\sigma_s^2$ converges and hence obtain the upper bound of $\sum_{s=0}^{T}\eta_s^2\sigma_s^2$.

Proof [Proof of Theorem 3] Throughout the proof, we condition on $\mathbf{W}(0)$ and a such that (22), (42), (55), (57) hold. This can be guaranteed with probability $1-\exp\left(-\Omega(C^{-L}m^{1/36})\right)$ following Theorem 1, Proposition 9, Lemma 15 and Lemma 16. For simplicity, we abbreviate the conditional expectation $\mathbb{E}\left[\cdot|\mathbf{W}(0), a\right]$ as $\mathbb{E}\left[\cdot\right]$. We now prove the theorem by induction.

When $t=0$, clearly $\|\Delta_0\|_2 \le \|\Delta_0\|_2 + 2c_2\|\Delta_0\|_2 + 2c_2\tau$.

Suppose (26) holds at all time $s \le t$, now we show it also holds at time $t+1$. Following (53) in Lemma 13, we have

$$\begin{aligned}\mathbb{E}\left[\|\Delta_{t+1}\|_2\right] &\le \left\|\prod_{s=0}^{t}\mathsf{K}_s\circ\Delta_0\right\|_2 + \sum_{r=0}^{t}\mathbb{E}\left[\left\|\prod_{i=r+1}^{t}\mathsf{Q}_i\mathsf{D}_r\prod_{j=0}^{r-1}\mathsf{K}_j\circ\Delta_0\right\|_2\right] \\ &\quad + \sum_{s=0}^{t}\mathbb{E}\left[\left\|\prod_{r=s+1}^{t}\mathsf{Q}_r\circ\epsilon_s\right\|_2\right] + \mathbb{E}\left[\left\|\sum_{s=0}^{t}\prod_{r=s+1}^{t}\mathsf{Q}_r\circ v_s\right\|_2\right] \\ &\le \left\|\prod_{s=0}^{t}\mathsf{K}_s\circ\Delta_0\right\|_2 + \sum_{r=0}^{t}\mathbb{E}\left[\|\mathsf{D}_r\|_2\right]\|\Delta_0\|_2 + \sum_{s=0}^{t}\mathbb{E}\left[\|\epsilon_s\|_2\right] + \mathbb{E}\left[\left\|\sum_{s=0}^{t}\prod_{r=s+1}^{t}\mathsf{Q}_r\circ v_s\right\|_2\right],\end{aligned} \tag{58}$$

where the last inequality holds by Lemma 14 that gives $\|\mathsf{Q}_t\|_2 \le 1$ and $\|\mathsf{K}_t\|_2 \le 1$ for all $t \le T$.

Now we bound each term on the right hand side above.

Denote $\rho_i(t) \triangleq \prod_{s=0}^{t}(1-\eta_s\lambda_i)$ where $\{\lambda_i\}_{i=1}^{\infty}$ are eigenvalues of $\mathsf{\Phi}$. Since $\lambda_1 \ge \lambda_2 \ge \cdots \ge 0$ and $\sup_i(1-\eta_s\lambda_i) \le 1$ for all s, we know $\rho_i(t)$ is bounded above by 1 and is increasing in i. To bound the first term on the right hand side of (58), here we use an induction to prove that

$$\prod_{s=0}^{t}\mathsf{K}_s\circ g = \sum_{i=1}^{\infty}\rho_i(t)\langle g,\phi_i\rangle\phi_i, \tag{59}$$

where ϕ_i is the eigenfunction of $\mathsf{\Phi}$ associated with eigenvalue λ_i.

When $t=0$, since K_0 is positive semi-definite, by Lemma 27, we have

$$\mathsf{K}_0\circ g = \sum_{i=1}^{\infty}\rho_i(0)\langle g,\phi_i\rangle\phi_i,$$

where ϕ_i is the eigenfunction of $\mathsf{\Phi}$ associated with eigenvalue λ_i.

Suppose (59) holds for some time t. Since K_{t+1} is PSD, by Lemma 27, we have

$$\begin{aligned}\mathsf{K}_{t+1}\circ\left(\prod_{s=0}^{t}\mathsf{K}_s\circ g\right) &= \sum_{i=1}^{\infty}(1-\eta_{t+1}\lambda_i)\left\langle\sum_{j=1}^{\infty}\rho_j(t)\langle g,\phi_j\rangle\phi_j,\phi_i\right\rangle\phi_i \\ &\overset{(a)}{=} \sum_{i=1}^{\infty}(1-\eta_{t+1}\lambda_i)\left\langle\rho_i(t)\langle g,\phi_i\rangle\phi_i,\phi_i\right\rangle\phi_i \\ &\overset{(b)}{=} \sum_{i=1}^{\infty}\rho_i(t+1)\langle g,\phi_i\rangle\phi_i,\end{aligned}$$

where (a) holds by orthogonality of $\{\phi_i\}$ and (b) holds by the definition of ρ_i and the normality of ϕ_i. Therefore, taking L_2 norm square on both hand sides of (59), for any

$r \in \mathbb{N}$, we get

$$\begin{aligned}\left\|\prod_{s=0}^{t} \mathsf{K}_s \circ \Delta_0\right\|_2^2 &= \sum_{i=0}^{\infty} \rho_i^2(t)\langle \Delta_0, \phi_i\rangle^2 \\ &\overset{(a)}{\le} \sum_{i=0}^{r} \rho_i^2(t)\langle \Delta_0, \phi_i\rangle^2 + \sum_{i=r+1}^{\infty} \langle \Delta_0, \phi_i\rangle^2 \\ &\overset{(b)}{\le} \rho_r^2(t) \sum_{i=0}^{r} \langle \Delta_0, \phi_i\rangle^2 + \mathcal{R}^2(\Delta_0, r) \\ &\le \rho_r^2(t) \left\|\Delta_0\right\|_2^2 + \mathcal{R}^2(\Delta_0, r),\end{aligned}$$

where the residual projection error $\mathcal{R}$ is defined in (25); (a) holds since $\rho_i(t) \le 1$ for all i and t; (b) holds since $\rho_i(t)$ is monotonic increasing in i. Hence, we have for any $r \in \mathbb{N}$,

$$\left\|\prod_{s=0}^{t} \mathsf{K}_s \circ \Delta_0\right\|_2 \le \prod_{s=0}^{t} (1 - \eta_s \lambda_r) \left\|\Delta_0\right\|_2 + \mathcal{R}(\Delta_0, r). \tag{60}$$

To bound $\sum_{r=0}^{t} \mathbb{E}\left[\|\mathsf{D}_r\|_2\right]$, note that

$$\|\mathsf{D}_r\|_2 \le \eta_r \left\|H_t - \Phi\right\|_\infty = O\left(\frac{\eta_r C^L}{m^{1/36}}\right).$$

Thus, we get

$$\sum_{r=0}^{t} \mathbb{E}\left[\|\mathsf{D}_r\|_2\right] = O\left(\frac{C^L \sum_{r=0}^{t} \eta_r}{m^{1/36}}\right) \le C_1 \frac{\theta C^L \log T}{m^{1/36}}, \tag{61}$$

for some absolute constant C_1 where the last inequality holds by plugging in $\eta_r \le \frac{\theta}{r+1}$.

By (55), we have

$$\sum_{s=0}^{t} \mathbb{E}\left[\|\epsilon_s\|_2\right] = O\left(\frac{C_2^L}{m^{1/36}} \sum_{s=0}^{t} \eta_s \sigma_s\right).$$

By definition of σ_s, we have

$$\sigma_s = \sqrt{\mathbb{E}\left[\|\Delta_s\|_2^2\right] + \tau^2} \le \mathbb{E}\left[\|\Delta_s\|_2\right] + \tau.$$

Now we prove that

$$\mathbb{E}\left[\|\Delta_s\|_2\right] \le (1 + 2c_2) \left\|\Delta_0\right\|_2 + 2c_2\tau.$$

and hence

$$\sigma_s \le (1 + 2c_2) \left\|\Delta_0\right\|_2 + (1 + 2c_2)\tau.$$

To see this, note that for any $\epsilon > 0$, $\mathcal{R}(\Delta_0, \ell) < \epsilon$ for sufficiently large ℓ. Therefore, since (26) holds for all $s \le t$, we have

$$\mathbb{E}\left[\|\Delta_s\|_2\right] \le \prod_{r=0}^{s} (1 - \eta_r \lambda_\ell) \left\|\Delta_0\right\|_2 + \epsilon + 2c_2 \left\|\Delta_0\right\|_2 + 2c_2\tau \le (1 + 2c_2) \left\|\Delta_0\right\|_2 + \epsilon + 2c_2\tau.$$

Since ϵ can be arbitrary, we have $\mathbb{E}\left[\|\Delta_s\|_2\right] \leq (1+2c_2)\|\Delta_0\|_2 + 2c_2\tau$. Plugging the bound to σ_s and $\eta_s \leq \frac{\theta}{s+1}$, we get

$$\sum_{s=0}^{t} \mathbb{E}\left[\|\epsilon_s\|_2\right] \leq \frac{C_3\theta C_2^L \log T}{m^{1/36}}(1+2c_2)\left(\|\Delta_0\|_2 + \tau\right). \tag{62}$$

for some absolute constant C_3 where we use the fact that $\sum_{t=0}^{T}\eta_t \leq \sum_{t=0}^{T}\frac{\theta}{t+1} \leq C'\theta\log T$.

Lastly, from (57), we have

$$\mathbb{E}\left[\left\|\sum_{s=0}^{t}\prod_{r=s+1}^{t}\mathsf{Q}_r\circ v_s\right\|_2\right] \leq c_2\left(\|\Delta_0\|_2+\tau\right).$$

Plugging the above bound as well as (60), (61) and (62) into (58), we have

$$\begin{aligned}
&\mathbb{E}\left[\|\Delta_{t+1}\|_2\right] \\
&\leq \prod_{s=0}^{t}(1-\eta_s\lambda_r)\|\Delta_0\|_2 + \mathcal{R}(\Delta_0, r) + C_1\frac{\theta C^L\log T}{m^{1/36}}\|\Delta_0\|_2 \\
&\quad + \frac{C_3\theta C_2^L\log T}{m^{1/36}}(1+2c_2)\left(\|\Delta_0\|_2+\tau\right) + c_2\left(\|\Delta_0\|_2+\tau\right) \\
&= \left\{\prod_{s=0}^{t}(1-\eta_s\lambda_r)\right\}\|\Delta_0\|_2 + \mathcal{R}(\Delta_0, r) \\
&\quad + \left[(C_1 + C_3(1+2c_2))\frac{\theta C_4^L\log T}{m^{1/36}} + c_2\right]\|\Delta_0\|_2 + \left(\frac{(1+2c_2)C_3\theta C_2^L\log T}{m^{1/36}} + c_2\right)\tau,
\end{aligned} \tag{63}$$

where $C_4 = \max\{C, C_2\}$.

When $m = \Omega\left((C_1 + C_3(1+c_2))^{36}\theta^{36}c_2^{36}C_4^{36L}\log^{36}T\right)$, we have

$$(C_1 + C_3(1+2c_2))\frac{\theta C_4^L\log T}{m^{1/36}} \leq c_2,$$

and

$$\frac{(1+2c_2)C_3\theta C_2^L\log T}{m^{1/36}} \leq c_2.$$

As a result, we have

$$\mathbb{E}\left[\|\Delta_{t+1}\|_2\right] \leq \prod_{s=0}^{t}(1-\eta_s\lambda_r)\|\Delta_0\|_2 + \mathcal{R}(\Delta_0, r) + 2c_2\|\Delta_0\|_2 + 2c_2\tau,$$

which completes the induction.

■

8. Numerical Study

In this section, we provide some numerical studies.

8.1 Synthetic data

We consider the following different choices of f^*:

- Linear: $f^*(x) = \langle b, x\rangle$ with parameter $b \in \mathbb{R}^d$;
- Quadratic: $f^*(x) = x^\top Ax + \langle b, x\rangle$, where $A \in \mathbb{R}^{d\times d}$ and $b \in \mathbb{R}^d$;
- Teacher neural network: $f^*(x) = \sum_{i=1}^3 b_i \psi(\langle v_i, x\rangle)$, where $\psi(z) = \frac{1}{1+e^{-z}}$ is the sigmoid function, $b_i \in \{-1, 1\}$, and $v_i \in \mathbb{R}^d$;
- Random label: $f^*(x)$ is an i.i.d. Bernoulli random variable with parameter $1/2$.

We use the symmetric initialization introduced in Section 4 as Corollary 4 suggests a zero residual projection error $\mathcal{R}(\Delta_0, \ell^* + 1)$ for a degree ℓ^* polynomial. We run the stochastic gradient descent algorithm (10) on the streaming data with constant step size $\eta = 0.3$ to train a four-layer neural network. At each iteration, we randomly draw data X uniformly from $\mathbb{S}^{d-1}$ and u from $\mathcal{N}(0, \tau^2)$ to obtain (X, y) where $y = f^*(X) + u$. The average prediction error is estimated using freshly drawn 200 data points, and the resulting error is further averaged over 20 independent runs.

In Section 4, we prove that the average prediction error converges under SGD in the streaming setting. Here we show the numerical performance of SGD. We study the normalized average prediction error $\mathbb{E}\left[\|\Delta_t\|_2\right] / \mathbb{E}\left[\|\Delta_0\|_2\right]$ for different f^* with $d = 5$, $m = 1000$, and $\tau = 0.1$. For linear, quadratic and teacher neural network f^*, the best achievable value of the normalized error equals 0. For random label f^*, since $f^*(x)$ is an i.i.d. Bernoulli random variable with parameter $1/2$ for any x, we get

$$\|\Delta_t\|_2^2 = \mathbb{E}_X\left[(f^*(X) - f(X; \mathbf{W}(t)))^2\right] = \frac{1}{2}\left[(f(X; \mathbf{W}(t)) - 1)^2 + f^2(X; \mathbf{W}(t))\right] \geq \frac{1}{4}, \forall t.$$

Hence, the best achievable value of the normalized average prediction error equals $\frac{1/2}{\mathbb{E}\left[\|\Delta_0\|_2\right]}$, which is represented by the horizontal dashed line in Figure 3. From Figure 3, we clearly see that SGD learns f^* efficiently for all four choices: the normalized average prediction error converges to the best achievable values.

As discussed in Section 4, our result which captures the contribution of NTK from all hidden layers, characterizes the average prediction error better than existing works (Du et al., 2019a) . Here, we provide numerical studies to verify this statement. Figure 4 plots the evolution of the average prediction error normalized by the error at initialization and the characterizations utilizing the spectrum of $\mathbf{\Phi}$ and $\mathbf{\Phi}^{(4)}$. It can be seen that our characterization based on $\mathbf{\Phi}$ is close to the actual SGD dynamic when f^* is linear, quadratic or teacher neural network function. Note that under the symmetric initialization, $\Delta_0 = f^*$. According to Corollary 4, we choose $\lambda_2(\mathbf{\Phi})$ for linear and $\lambda_3(\mathbf{\Phi})$ for quadratic f^* since the residual projection error equals 0. For teacher neural network f^* which is not polynomial, we cannot find some ℓ^* such that the residual projection error $\mathcal{R}(f^*, \ell^*) = 0$. Instead, we choose $\lambda_4(\mathbf{\Phi})$ as it provides the best fit among all $\lambda_i(\mathbf{\Phi}), i \geq 1$.

As shown in Section 6, to ensure H_t is close to H_0, we crucially prove that $\mathbf{W}^{(\ell)}(t)$ stays relatively close to $\mathbf{W}^{(\ell)}(0)$ in Lemma 10 and bound the number of sign changes for each hidden layer in Lemma 11. Here we study both phenomena numerically. Figure 5 studies

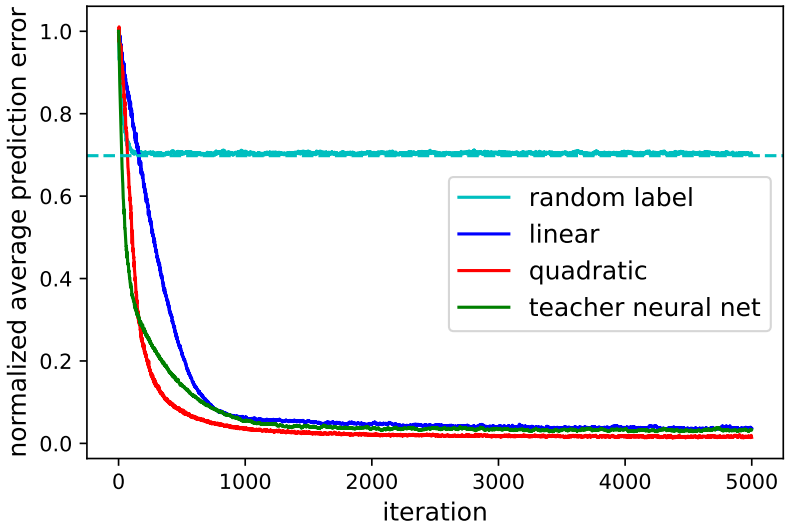
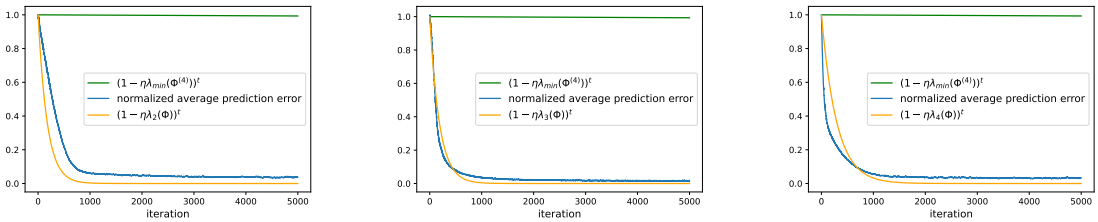


Figure 3: Normalized prediction error for different f^*



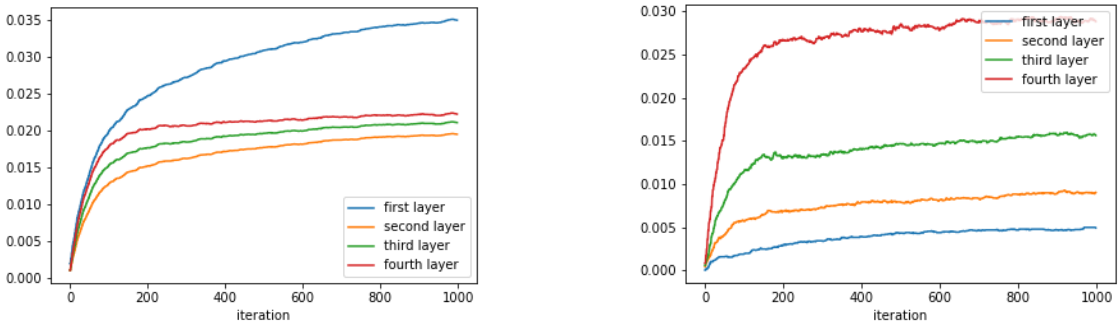
(a) Linear f^* (b) Quadratic f^* (c) Teacher neural net f^*

Figure 4: Average prediction error and the characterization based on Φ

teacher neural network f^*. Figure 5a shows that the relative deviation of weight matrix $\left\|\mathbf{W}^{(\ell)}(t)-\mathbf{W}^{(\ell)}(0)\right\|_2 / \left\|\mathbf{W}^{(\ell)}(0)\right\|_2$ for each ℓ-th hidden layer is small. This is consistent with the trend implied by Lemma 10 which shows that with high probability, the numerator $\left\|\mathbf{W}^{(\ell)}(t)-\mathbf{W}^{(\ell)}(0)\right\|_2$ is small compared to the denominator $\left\|\mathbf{W}^{(\ell)}(0)\right\|_2$. In Figure 5b, we observe only a small fraction of sign changes in each hidden layer throughout the training. Furthermore, we see the proportion of sign changes increase when the layer index ℓ increases. Both are consistent with the trend indicated by Lemma 11.

8.2 Real data experiment

To illustrate the characterization from our theoretical result on real data, we run a numerical experiment on MNIST dataset. For simplicity, we only use the data corresponding to digit 0 and digit 1. We randomly draw 1500 images with 28×28 pixels from each digit and treat the empirical distribution of these 3000 images as the underlying true data distribution. We reshape the data to have $x_i \in \mathbb{R}^{784}$. For each $x_i \in \mathbb{R}^{784}$ in the dataset, we assign $y_i = 1$ if the corresponding image is digit 1 and $y_i = -1$ if the image is digit 0. We



(a) Relative deviation of weight matrices

(b) Proportion of sign flips

Figure 5: Evolution of weight and sign flips for a 4-layer teacher neural network f^*

then normalize x_i to have $\|x_i\|_2 = 1$. We run the mini-batch SGD with mini-batch size 100 on streaming data with step size $\eta = 0.7$ using a 4-layer neural network with 10000 neurons in each hidden layer. The reason for the use of mini-batch is to limit the noise from the stochastic gradient. Figure 6 shows the training loss normalized by the loss at initialization and two characterizations from the spectrum of NTK. We use error_0 to denote the average prediction error at initialization. It can be seen that the characterization from our result provides a much tighter bound than the characterization from existing works (Du et al., 2019a) using only the spectrum of the NTK from the last hidden layer. In addition, we clearly see two elbow points on our characterization. The first 100 iterations correspond to $(1 - \eta\lambda_1(\Phi))^t + \mathcal{R}(\Delta_0, 1)/\text{error}_0$ while the next 400 iterations correspond to $(1 - \eta\lambda_2(\Phi))^t + \mathcal{R}(\Delta_0, 2)/\text{error}_0$.

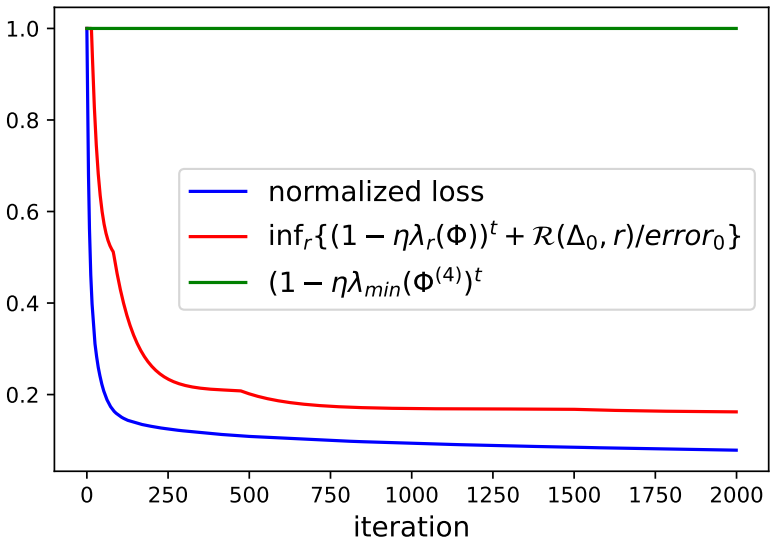


Figure 6: Normalized training loss for the first 2000 iterations

9. Conclusion

In this paper, we show the uniform concentration of NTK from all hidden layers of neural networks which allows us to capture the contribution of intermediate layers in the characterization of GD/SGD dynamics. Furthermore, in the streaming setting, we show the average prediction error under SGD converges in expectation. Our analysis opens the door for several interesting future directions. For example, it is of great interest to extend our study to Markovian data arising in the reinforcement learning. It is also useful to extend our uniform concentration result to other neural network architectures such as convolutional neural network (CNN).

Acknowledgments.

The research is supported in part by the NSF Grant CCF-1856424 and an NSF CAREER award CCF-2144593.

References

M. Abramowitz, I. A. Stegun, and R. H. Romer. Handbook of mathematical functions with formulas, graphs, and mathematical tables, 1988.

Z. Allen-Zhu and Y. Li. What can resnet learn efficiently, going beyond kernels? In *Advances in Neural Information Processing Systems*, pages 9017–9028, 2019a.

Z. Allen-Zhu and Y. Li. Can sgd learn recurrent neural networks with provable generalization? In *Advances in Neural Information Processing Systems*, pages 10331–10341, 2019b.

Z. Allen-Zhu and Y. Li. Backward feature correction: How deep learning performs deep learning. *arXiv preprint arXiv:2001.04413*, 2020.

Z. Allen-Zhu, Y. Li, and Z. Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pages 242–252, 2019a.

Z. Allen-Zhu, Y. Li, and Z. Song. On the convergence rate of training recurrent neural networks. In *Advances in neural information processing systems*, pages 6676–6688, 2019b.

S. Arora, S. Du, W. Hu, Z. Li, and R. Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pages 322–332. PMLR, 2019a.

S. Arora, S. S. Du, W. Hu, Z. Li, R. R. Salakhutdinov, and R. Wang. On exact computation with an infinitely wide neural net. *Advances in neural information processing systems*, 32, 2019b.

M. J. Cantero and A. Iserles. On rapid computation of expansions in ultraspherical polynomials. *SIAM Journal on Numerical Analysis*, 50(1):307–327, 2012.

Y. Cao and Q. Gu. Generalization bounds of stochastic gradient descent for wide and deep neural networks. In *Advances in Neural Information Processing Systems*, pages 10836–10846, 2019.

N. Cesa-Bianchi, A. Conconi, and C. Gentile. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057, 2004.

Z. Chen, Y. Cao, Q. Gu, and T. Zhang. Mean-field analysis of two-layer neural networks: Non-asymptotic rates and generalization bounds. *arXiv preprint arXiv:2002.04026*, 2020.

L. Chizat and F. Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in neural information processing systems*, pages 3036–3046, 2018.

Y. Cho and L. Saul. Kernel methods for deep learning. *Advances in neural information processing systems*, 22, 2009.

F. Dai and Y. Xu. *Approximation theory and harmonic analysis on spheres and balls*, volume 23. Springer, 2013.

S. Du, J. Lee, H. Li, L. Wang, and X. Zhai. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*, pages 1675–1685, 2019a.

S. S. Du, Y. Wang, X. Zhai, S. Balakrishnan, R. R. Salakhutdinov, and A. Singh. How many samples are needed to estimate a convolutional neural network? In *Advances in Neural Information Processing Systems*, pages 373–383, 2018.

S. S. Du, X. Zhai, B. Poczos, and A. Singh. Gradient descent provably optimizes over-parameterized neural networks. *ICLR 2019*, 2019b.

J. Feigenbaum, Y. Ishai, T. Malkin, K. Nissim, M. J. Strauss, and R. N. Wright. Secure multiparty computation of approximations. In *International Colloquium on Automata, Languages, and Programming*, pages 927–938. Springer, 2001.

B. Hajek and M. Raginsky. Statistical learning theory. *Lecture Notes*, 387, 2019.

W. Hu, C. J. Li, L. Li, and J.-G. Liu. On the diffusion approximation of nonconvex stochastic gradient descent. *Annals of Mathematical Sciences and Applications*, 4(1), 2019.

E. Ikonomovska, S. Loskovska, and D. Gjorgjevik. A survey of stream data mining. In *Proceedings of 8th National Conference with International participation, ETAI*, pages 19–21, 2007.

L. Isserlis. On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. *Biometrika*, 12(1/2):134–139, 1918.

A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580, 2018.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

Z. Li, R. Wang, D. Yu, S. S. Du, W. Hu, R. Salakhutdinov, and S. Arora. Enhanced convolutional neural tangent kernels. *arXiv preprint arXiv:1911.00809*, 2019.

S. Mei, A. Montanari, and P.-M. Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

S. Mei, T. Misiakiewicz, and A. Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *Conference on Learning Theory*, pages 2388–2464. PMLR, 2019.

J. Mercer. Xvi. functions of positive and negative type, and their connection the theory of integral equations. *Philosophical transactions of the royal society of London. Series A, containing papers of a mathematical or physical character*, 209(441-458):415–446, 1909.

S. Muthukrishnan. *Data streams: Algorithms and applications*. Now Publishers Inc, 2005.

L. O'callaghan, N. Mishra, A. Meyerson, S. Guha, and R. Motwani. Streaming-data algorithms for high-quality clustering. In *Proceedings 18th International Conference on Data Engineering*, pages 685–694. IEEE, 2002.

G. Rotskoff and E. Vanden-Eijnden. Parameters as interacting particles: long time convergence and asymptotic error scaling of neural networks. *Advances in neural information processing systems*, 31, 2018.

J. Shawe-Taylor, N. Cristianini, et al. *Kernel methods for pattern analysis*. Cambridge university press, 2004.

J. Sirignano and K. Spiliopoulos. Mean field analysis of deep neural networks. *Mathematics of Operations Research*, 47(1):120–152, 2022.

L. Su and P. Yang. On learning over-parameterized neural networks: A functional approximation perspective. In *Advances in Neural Information Processing Systems*, pages 2641–2650, 2019.

T. Tirer, J. Bruna, and R. Giryes. Kernel-based smoothness analysis of residual networks. In *Mathematical and Scientific Machine Learning*, pages 921–954. PMLR, 2022.

B. Tzen and M. Raginsky. A mean-field theory of lazy training in two-layer neural nets: entropic regularization and controlled mckean-vlasov dynamics. *arXiv preprint arXiv:2002.01987*, 2020.

A. Van Der Vaart and J. A. Wellner. A note on bounds for vc dimensions. *Institute of Mathematical Statistics collections*, 5:103, 2009.

R. Vershynin. *High-dimensional probability*. Cambridge, UK: Cambridge University Press, 2019.

Y. Wang. Harmonic analysis and isoperimetric inequalities. *Lecture Notes*, 2010.

J. Xu and H. Zhu. One-pass stochastic gradient descent in overparametrized two-layer neural networks. In *International Conference on Artificial Intelligence and Statistics*, pages 3673–3681. PMLR, 2021.

D. Zou, Y. Cao, D. Zhou, and Q. Gu. Gradient descent optimizes over-parameterized deep relu networks. *Machine Learning*, 109(3):467–492, 2020.

Appendix A. Auxiliary Results

A.1 Concentration Inequalities

In this section, we provide the concentration inequalities used in this paper. First of all, we present McDiarmid's inequality.

Lemma 17 *(Hajek and Raginsky, 2019, Theorem 2.3) Let $X = (X_1, \cdots, X_m) \in \mathcal{X}^m$ be an n-tuple of $\mathcal{X}$-valued independent random variables and $f : \mathcal{X}^m \to \mathbb{R}$ be a measurable function. Assume the value of $f(x)$ can change by at most $c_i > 0$ under an arbitrary change of the i-th coordinate. Then for any $t > 0$,*

$$\mathbb{P}\left[f(X) - \mathbb{E}\left[f(X)\right] \geq t\right] \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^m c_i^2}\right).$$

The following lemma is Bernstein inequality which shows the concentration of the sum of i.i.d. sub-exponential random variables.

Lemma 18 *(Vershynin, 2019, Theorem 2.8.1) Let $X_1, \cdots, X_n$ be i.i.d., sub-exponential random variables with sub-exponential norm $\|X_i\|_{\psi_1} \leq K$. Then for any $t > 0$, we have*

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}\left[X_1\right]\right| > t\right] \leq 2\exp\left(-C\min\left(\frac{nt^2}{K^2}, \frac{nt}{K}\right)\right).$$

Finally, we present Hanson-Wright inequality. For any random variable X, define $\|X\|_{\psi_2} \triangleq \inf\left\{t > 0 : \mathbb{E}\left[\exp\left(X^2/t^2\right)\right] \leq 2\right\}$.

Lemma 19 *(Vershynin, 2019, Theorem 6.2.1) Let $X = (X_1, \cdots, X_m) \in \mathbb{R}^m$ be a random vector with independent, mean zero, sub-Gaussian coordinates. Let $\mathbf{A}$ be an $m \times m$ matrix. Then for any $t \geq 0$,*

$$\mathbb{P}\left[\left|X^\top \mathbf{A} X - \mathbb{E}\left[X^\top \mathbf{A} X\right]\right| > t\right] \leq 2\exp\left(-C\min\left\{\frac{t^2}{K^4\|\mathbf{A}\|_{\mathrm{F}}^2}, \frac{t}{K^2\|\mathbf{A}\|_2}\right\}\right),$$

where $K = \max_i \|X_i\|_{\psi_2}$.

A.2 VC Dimension

Let $\mathcal{C}$ be a collection of subsets in $\mathbb{R}^p$. For any set A consisting of finite points in $\mathbb{R}^p$, we denote $\mathcal{C}_A = \{C \cap A : C \in \mathcal{C}\}$. We say $\mathcal{C}_A$ shatters set A if $|\mathcal{C}_A| = 2^{|A|}$. Let $\mathcal{M}_{\mathcal{C}}(n) = \max\{|\mathcal{C}_F| : F \subset \mathbb{R}^p, |F| = n\}$ and $\mathcal{P}(\mathcal{C}) = \sup\{n : \mathcal{M}_{\mathcal{C}}(n) = 2^n\}$ which is the largest cardinality of a set that can be shattered by $\mathcal{C}$.

Consider a Boolean function class $\mathcal{F}$ on $\mathbb{R}^p$. For each $f \in \mathcal{F}$, we denote $D_f = \{x \in \mathbb{R}^p : f(x) = 1\}$. As a result, the collection $\mathcal{C}_{\mathcal{F}} \triangleq \{D_f : f \in \mathcal{F}\}$ forms a collection of subsets of $\mathbb{R}^p$. Define $\mathcal{C}_{\mathcal{F}}(A) = \{D_f \cap A : f \in \mathcal{F}\}$. The VC dimension of $\mathcal{F}$ is then defined as

$$\mathsf{VC}(\mathcal{F}) \triangleq \mathcal{P}\left(\mathcal{C}_{\mathcal{F}}\right) = \sup\left\{n : \max_A |\mathcal{C}_{\mathcal{F}}(A)| = 2^n : |A| = n, A \subset \mathbb{R}^p\right\}.$$

Now we provide the auxiliary results in this paper regarding VC dimension.

The following lemma can be used to obtain the bound of VC dimension of the function class consisting of functions with the form of a product of Boolean functions.

Lemma 20 *(Van Der Vaart and Wellner, 2009, Theorem 1.1) For Boolean function classes $\mathcal{H}$ and $\{\mathcal{F}_i\}_{i=1}^N$, if for any $h \in \mathcal{H}$, there exists functions $f_1 \in \mathcal{F}_1, \cdots, f_N \in \mathcal{F}_N$ such that $h = \prod_{i=1}^N f_i$, then we have*

$$\mathsf{VC}(\mathcal{H}) \le \frac{5}{2} \log(4N) \sum_{i=1}^N \mathsf{VC}(\mathcal{F}_i).$$

The next lemma bounds the expectation of the largest deviation of an average of some Boolean function through VC dimension.

Lemma 21 *(Vershynin, 2019, Theorem 8.3.23) Let $\mathcal{F}$ be a class of Boolean functions on a probability space (Ω, Σ, μ) with finite VC dimension. Let $X_1, X_2, \cdots, X_n$ be independent random points in Ω. Then*

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}_X\left[f(X)\right] \right| \right] \le C\sqrt{\frac{\mathsf{VC}(\mathcal{F})}{n}}$$

for some constant C.

Next, we present Sauer-Shelah Lemma which bounds the cardinality of $\mathcal{C}_{\mathcal{F}}(A)$ with the VC dimension of $\mathcal{F}$.

Lemma 22 *(Vershynin, 2019, Theorem 8.3.16) Let $\mathcal{F}$ be a Boolean function class and $A = \{a_1, \cdots, a_n\}$ be a set of n points in the space. Then for any $n \ge \mathsf{VC}(\mathcal{F})$,*

$$|\left\{(f(a_1), \cdots, f(a_n)) : f \in \mathcal{F}\right\}| = |\mathcal{C}_{\mathcal{F}}(A)| \le \left(\frac{en}{\mathsf{VC}(\mathcal{F})}\right)^{\mathsf{VC}(\mathcal{F})}.$$

Lemma 23 *(Hajek and Raginsky, 2019, Proposition 7.1) Let $\mathcal{F} = \left\{f_\theta(y) = \mathbf{1}_{\{\langle y,\theta\rangle \ge 0\}} : \theta \in \Theta\right\}$ where $y \in \mathbb{R}^p$ and Θ is some q-dimensional subspace of $\mathbb{R}^p$. Then*

$$\mathsf{VC}(\mathcal{F}) = q.$$

Lastly, we provide a lemma that bounds the VC dimension of union of function classes when the number of function classes is much larger than their VC dimension.

Lemma 24 *Suppose $\mathcal{F} = \cup_{i=1}^N \mathcal{F}_i$ where $\mathsf{VC}(\mathcal{F}_i) = d$ for all i, then*

$$\mathsf{VC}(\mathcal{F}) \le C \max\left(d \log d, \log N\right).$$

Proof [Proof of Lemma 24] Fix arbitrary set $A = \{y_1, \cdots, y_n\}$ of size n. Since $\mathcal{F} = \cup_{i=1}^N \mathcal{F}_i$, we have $\mathcal{C}_{\mathcal{F}}(A) = \cup_{j=1}^N \mathcal{C}_{\mathcal{F}_j}(A)$.

Thus, we have

$$|\mathcal{C}_{\mathcal{F}}(A)| \le \sum_{j=1}^N |\mathcal{C}_{\mathcal{F}_j}(A)| \overset{(a)}{\le} \sum_{j=1}^N n^{\mathsf{VC}(\mathcal{F}_j)} \le Nn^d.$$

where (a) holds by Lemma 22.

By the definition of VC dimension, if $n^d N < 2^n$, then $\mathsf{VC}(\mathcal{F}) < n$.

Taking logarithm on both hand sides, we have $d \log n + \log N < n \log 2$.

Note that when $\frac{n}{2} > d \log n$ and $\frac{n}{2} > \log N$, i.e., $n \geq \max(Cd \log d, 2 \log N)$, the above inequality clearly holds.

Therefore, we get

$$\mathsf{VC}(\mathcal{F}) \leq C \max(Cd \log d, \log N)$$

for some universal constant C. ■

A.3 Kernel

Here, we provide some intermediate results used in this paper regarding kernel operator.

Lemma 25 *(Shawe-Taylor et al., 2004, Proposition 3.22) For any positive semi-definite kernel κ_1 and κ_2, any function ϕ, we have κ_3, κ_4 and κ_5 are positive semi-definite kernels where*

$$\kappa_3(x, y) \triangleq \kappa_1(x, y) + \kappa_2(x, y), \tag{64}$$

$$\kappa_4(x, y) \triangleq \kappa_1(x, y)\kappa_2(x, y), \tag{65}$$

and

$$\kappa_5(x, y) \triangleq \kappa_1(\phi(x), \phi(y)). \tag{66}$$

Lemma 26 *(Shawe-Taylor et al., 2004, Theorem 3.13) Suppose $f(x, y)$ is a kernel function. If for any $g \in L_2(\mu)$,*

$$\int\int f(x, y)g(x)g(y)d\mu(x)d\mu(y) \geq 0,$$

then f is positive semi-definite.

Lemma 27 *(Mercer, 1909) Suppose κ is a positive semi-definite kernel. Then there exists non-negative eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots$ and orthonormal eigenfunctions $\{\phi_i\}$ such that*

$$\kappa(x, y) = \sum_{j=1}^{\infty} \lambda_j \phi_j(x)\phi_j(y).$$

Lemma 28 *For any positive semi-definite kernel operator J associated with function J, we have*

$$\|\mathsf{J}\|_2 \leq \|J\|_\infty.$$

Proof [Proof of Lemma 28] By Cauchy-Schwartz inequality, we have

$$\begin{aligned}
\|\mathsf{J}\|_2^2 &= \sup_{\|g\|_2=1} \int \left(\int \mathsf{J}(x, y)g(y)d\mu(y) \right)^2 d\mu(x) \\
&\leq \sup_{\|g\|_2=1} \int\int \mathsf{J}^2(x, y)d\mu(y)d\mu(x) \int g^2(y)d\mu(y) \\
&\leq \|J\|_\infty^2
\end{aligned}$$

where $\|g\|_2 = \sqrt{\int g^2(x) d\mu(x)}$ is the L_2 norm for function g. ■

A.4 Probability of the intersection of events

Lemma 29 *Let $\{A_i\}$ and $\{B_i\}$ be two sequences of events, where A_0 and B_0 are the whole probability spaces. Then we have for any $n \geq 1$,*

$$\mathbb{P}\left[\cap_{i=1}^n (A_i \cap B_i)\right] \geq 1 - \sum_{i=1}^n \mathbb{P}\left[B_i^c \mid \cap_{k=1}^{i-1}(A_k \cap B_k)\right] - \sum_{i=1}^n \mathbb{P}\left[A_i^c\right]$$

and

$$\mathbb{P}\left[B_n \cap (\cap_{i=1}^n A_i)\right] \geq 1 - \sum_{i=1}^n \mathbb{P}\left[B_i^c | B_{i-1} \cap (\cap_{k=1}^{i-1} A_k)\right] - \sum_{i=1}^n \mathbb{P}\left[A_i^c\right],$$

where $\cap_{i=1}^0 F_i$ for any event F_i is understood as the whole probability space.

Proof [Proof of Lemma 29] Note that for any event E and F, we have

$$\mathbb{P}[F] \leq \mathbb{P}[E \cap F] + \mathbb{P}[E^c \cap F] \leq \mathbb{P}[E] + \mathbb{P}[E^c \cap F]. \tag{67}$$

Taking $E = \cap_{i=1}^n (A_i \cap B_i)$ and $F = \cap_{i=1}^{n-1} (A_i \cap B_i)$, we have

$$\mathbb{P}\left[\cap_{i=1}^{n-1} (A_i \cap B_i)\right] \leq \mathbb{P}\left[\cap_{i=1}^n (A_i \cap B_i)\right] + \mathbb{P}\left[(\cap_{i=1}^n (A_i \cap B_i))^c \cap \left(\cap_{i=1}^{n-1} (A_i \cap B_i)\right)\right]. \tag{68}$$

Now we bound $\mathbb{P}\left[(\cap_{i=1}^n (A_i \cap B_i))^c \cap \left(\cap_{i=1}^{n-1} (A_i \cap B_i)\right)\right]$.
Since $(\cap_{i=1}^n (A_i \cap B_i))^c = \left[\cap_{i=1}^{n-1} (A_i \cap B_i)\right]^c \cup A_n^c \cup B_n^c$, we have

$$\begin{aligned}
\mathbb{P}\left[(\cap_{i=1}^n (A_i \cap B_i))^c \cap \left(\cap_{i=1}^{n-1} (A_i \cap B_i)\right)\right] &\leq \mathbb{P}\left[(A_n^c \cup B_n^c) \cap \left(\cap_{i=1}^{n-1} (A_i \cap B_i)\right)\right] \\
&\leq \mathbb{P}[A_n^c] + \mathbb{P}\left[B_n^c \cap \left(\cap_{i=1}^{n-1} (A_i \cap B_i)\right)\right] \\
&\leq \mathbb{P}[A_n^c] + \mathbb{P}\left[B_n^c | \cap_{i=1}^{n-1} (A_i \cap B_i)\right].
\end{aligned}$$

Plugging the aboved displayed equation into (68), we have

$$\mathbb{P}\left[\cap_{i=1}^n (A_i \cap B_i)\right] \geq \mathbb{P}\left[\cap_{i=1}^{n-1} (A_i \cap B_i)\right] - \mathbb{P}[A_n^c] - \mathbb{P}\left[B_n^c | \cap_{i=1}^{n-1} (A_i \cap B_i)\right].$$

Recursively replacing $\mathbb{P}\left[\cap_{i=1}^{n-1} (A_i \cap B_i)\right]$ on the right hand side of the above inequality, we obtain the first inequality of Lemma 29.

Similarly, we prove the second inequality of Lemma 29 . From (67), taking $E = B_n \cap (\cap_{i=1}^n A_i)$ and $F = B_{n-1} \cap \left(\cap_{i=1}^{n-1} A_i\right)$, we have

$$\mathbb{P}\left[B_{n-1} \cap \left(\cap_{i=1}^{n-1} A_i\right)\right] \leq \mathbb{P}\left[B_n \cap (\cap_{i=1}^n A_i)\right] + \mathbb{P}\left[(B_n \cap (\cap_{i=1}^n A_i))^c \cap \left(B_{n-1} \cap \left(\cap_{i=1}^{n-1} A_i\right)\right)\right].$$

Since $(B_n \cap (\cap_{i=1}^n A_i))^c = B_n^c \cup A_n^c \cup \left(\cap_{i=1}^{n-1} A_i\right)^c$, we have

$$\begin{aligned}
&\mathbb{P}\left[(B_n \cap (\cap_{i=1}^n A_i))^c \cap \left(B_{n-1} \cap \left(\cap_{i=1}^{n-1} A_i\right)\right)\right] \\
&\leq \mathbb{P}[A_n^c] + \mathbb{P}\left[B_n^c \cap B_{n-1} \cap \left(\cap_{i=1}^{n-1} A_i\right)\right] \\
&\leq \mathbb{P}[A_n^c] + \mathbb{P}\left[B_n^c | B_{n-1} \cap \left(\cap_{i=1}^{n-1} A_i\right)\right].
\end{aligned}$$

Thus, $\mathbb{P}\left[B_n \cap (\cap_{i=1}^n A_i)\right] \geq \mathbb{P}\left[B_{n-1} \cap \left(\cap_{i=1}^{n-1} A_i\right)\right] - \mathbb{P}\left[A_n^c\right] - \mathbb{P}\left[B_n^c | B_{n-1} \cap \left(\cap_{i=1}^{n-1} A_i\right)\right]$.

Recursively applying the above inequality, we obtain the second inequality of Lemma 29. ■

Appendix B. Proofs in Section 5

Recall from Section 5 that the proof of Theorem 1 consists of Lemma 6–8. Here, we present the full proofs of these lemmas. Since the proofs involve several key intermediate results, to ease the reading, we present the following diagram to illustrate the proof structure.

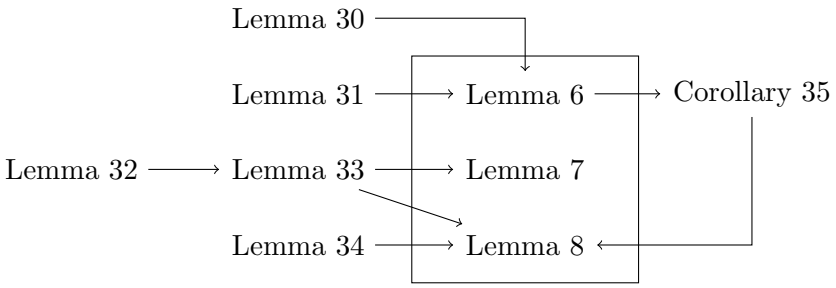


Figure 7: Diagram of the proof structure in Appendix B

B.1 Proof of Lemma 6: Concentration of $\langle o^{(\ell)}(x), o^{(\ell)}(x')\rangle$

As mentioned in Section 5, to prove Lemma 6, we need to establish that $o^{(\ell)}(x)$ is Lipschitz in x. This is done in the following lemma.

Define

$$\mathcal{E}_0^{(k)} = \left\{ \left\| \mathbf{W}^{(k)} \right\|_2 \leq c_0 \sqrt{m} \right\}. \tag{69}$$

where $\mathbf{W}^{(k)}$ is the weight matrix of the k-th hidden layer.

Lemma 30 *For any $0 \leq \ell \leq L$, under event $\cap_{k=1}^{\ell} \mathcal{E}_0^{(k)}$,*

- $\sup_x \left\| o^{(\ell)}(x) \right\|_2 \leq c_0^{\ell}$;
- $\left\| o^{(\ell)}(x) - o^{(\ell)}(z) \right\|_2 \leq c_0^{\ell} \left\| x - z \right\|_2$.

Proof [Proof of Lemma 30] Recall that $o^{(0)}(x) = x$ and

$$o^{(\ell)}(x) = \frac{1}{\sqrt{m}} \mathbf{D}^{(\ell)}(x) \mathbf{W}^{(\ell)} \cdots \frac{1}{\sqrt{m}} \mathbf{D}^{(1)}(x) \mathbf{W}^{(1)} x, \quad \forall \ell \geq 1$$

where $\mathbf{D}^{(\ell)}(x) = \mathsf{diag}\left\{ \mathbf{1}_{\left\{ \langle w_i^{(\ell)}, o^{(\ell-1)}(x) \rangle \geq 0 \right\}} \right\}$ and $w_i^{(\ell)}$ is the i-th row of $\mathbf{W}^{(\ell)}$.

When $\ell = 0$, since $o^{(0)}(x) = x$, both inequalities of Lemma 30 hold directly.

Now consider the case for $\ell \geq 1$. Under $\cap_{k=1}^{\ell} \mathcal{E}_0^{(k)}$, we know $\left\|\mathbf{W}^{(k)}\right\|_2 \leq c_0\sqrt{m}$ for all $k = 1, 2, \cdots, \ell$. Therefore, for any x, $\left\|\frac{1}{\sqrt{m}}\mathbf{D}^{(k)}(x)\mathbf{W}^{(k)}\right\|_2 \leq c_0$ for all $k = 1, 2, \cdots, \ell$. Thus, we have

$$\begin{aligned}
\sup_x \left\|o^{(\ell)}(x)\right\|_2 &\leq \sup_x \left\|\frac{1}{\sqrt{m}}\mathbf{D}^{(\ell)}(x)\mathbf{W}^{(\ell)} \cdots \frac{1}{\sqrt{m}}\mathbf{D}^{(1)}(x)\mathbf{W}^{(1)}\right\|_2 \\
&\leq \sup_x \left\|\frac{1}{\sqrt{m}}\mathbf{D}^{(\ell)}(x)\mathbf{W}^{(\ell)}\right\|_2 \cdots \sup_x \left\|\frac{1}{\sqrt{m}}\mathbf{D}^{(1)}(x)\mathbf{W}^{(1)}\right\|_2 \\
&\leq c_0^{\ell}.
\end{aligned}$$

This completes the proof of the first inequality of Lemma 30.

Now we prove the second inequality of Lemma 30. By the definition of $o^{(\ell)}(x)$, we know

$$\begin{aligned}
\left[o^{(\ell)}(x)\right]_i &= \left[\frac{1}{\sqrt{m}}\mathbf{D}^{(\ell)}(x)\mathbf{W}^{(\ell)}o^{(\ell-1)}(x)\right]_i = \frac{1}{\sqrt{m}}\mathbf{1}_{\left\{\langle w_i^{(\ell)}, o^{(\ell-1)}(x)\rangle \geq 0\right\}}\langle w_i^{(\ell)}, o^{(\ell-1)}(x)\rangle \\
&= \frac{1}{\sqrt{m}}\sigma\left(\langle w_i^{(\ell)}, o^{(\ell-1)}(x)\rangle\right),
\end{aligned}$$

where $\left[o^{(\ell)}(x)\right]_i$ is the i-th coordinate of $o^{(\ell)}(x)$.

As a result, for any x and z, we have

$$\begin{aligned}
\left\|o^{(\ell)}(x) - o^{(\ell)}(z)\right\|_2^2 &= \frac{1}{m}\sum_{i=1}^{m}\left(\sigma(\langle w_i^{(\ell)}, o^{(\ell-1)}(x)\rangle) - \sigma(\langle w_i^{(\ell)}, o^{(\ell-1)}(z)\rangle)\right)^2 \\
&\overset{\text{(i)}}{\leq} \frac{1}{m}\sum_{i=1}^{m}\left(\langle w_i^{(\ell)}, o^{(\ell-1)}(x)\rangle - \langle w_i^{(\ell)}, o^{(\ell-1)}(z)\rangle\right)^2 \\
&= \frac{1}{m}\left\|\mathbf{W}^{(\ell)}\left(o^{(\ell-1)}(x) - o^{(\ell-1)}(z)\right)\right\|_2^2 \\
&\leq c_0^2\left\|o^{(\ell-1)}(x) - o^{(\ell-1)}(z)\right\|_2^2.
\end{aligned}$$

where (i) holds since ReLU function is 1-Lipchitz and the last inequality holds under $\cap_{k=1}^{\ell} \mathcal{E}_0^{(k)}$.

Recursively applying the above displayed equation, we obtain the second inequality of Lemma 30. ■

The following lemma from Du et al. (2019a, Lemma G.4) shows that if the covariance matrices of two pairs of bivariate normal random variables are close entrywise, then the expectation of some function F on these two pairs are also close.

Lemma 31 *Let*

$$\mathbf{A} = \begin{pmatrix} a_1^2 & \rho_1 a_1 b_1 \\ \rho_1 a_1 b_1 & b_1^2 \end{pmatrix} \quad \text{and} \quad \mathbf{B} = \begin{pmatrix} a_2^2 & \rho_2 a_2 b_2 \\ \rho_2 a_2 b_2 & b_2^2 \end{pmatrix}.$$

Suppose there exists some constant $C > 0$ such that $\frac{1}{C} \leq \min(a_1, b_1, a_2, b_2) \leq \max(a_1, b_1, a_2, b_2) \leq C$. Define $F(\mathbf{X}) = \mathbb{E}_{(U,V)\sim\mathcal{N}(0,\mathbf{X})}\left[\sigma(U)\sigma(V)\right]$ for any positive definite matrix X. Then we have

$$|F(\mathbf{A}) - F(\mathbf{B})| = O\left(\|\mathbf{A} - \mathbf{B}\|_{\infty}\right).$$

Proof of Lemma 6 Denote V_0 as a $\frac{1}{m^2}$-net of $\mathbb{S}^{d-1}$. By Vershynin (2019, Corollary 4.2.13), we have V_0 is of size $O\left(m^{2d}\right)$. Define event $\mathcal{E}_1^{(k)}$ such that the following holds for any $x_0, x_0' \in V_0$:

$$\left|\frac{1}{m}\sum_{i=1}^{m}\sigma\left(\langle w_i^{(k)}, o^{(k-1)}(x_0)\rangle\right)\sigma\left(\langle w_i^{(k)}, o^{(k-1)}(x_0')\rangle\right) - \mathbb{E}_w\left[\sigma\left(\langle w, o^{(k-1)}(x_0)\rangle\right)\sigma\left(\langle w, o^{(k-1)}(x_0')\rangle\right)\right]\right| \leq C\frac{c_0^{2(k-1)}}{m^{1/3}}. \tag{70}$$

Denote $\mathcal{E}_1 = \cap_{k=1}^{L}\mathcal{E}_1^{(k)}$ and $\mathcal{E}_0 = \cap_{k=1}^{L}\mathcal{E}_0^{(k)}$.

To prove the lemma, we first bound $\mathbb{P}\left[\mathcal{E}_0 \cap \mathcal{E}_1\right]$ and then show (30) holds under $\mathcal{E}_0 \cap \mathcal{E}_1$. By Lemma 29, we have

$$\mathbb{P}\left[\mathcal{E}_0 \cap \mathcal{E}_1\right] \geq 1 - \sum_{\ell=1}^{L}\mathbb{P}\left[\left(\mathcal{E}_1^{(\ell)}\right)^c \mid \left(\cap_{k=1}^{\ell-1}\mathcal{E}_1^{(k)}\right) \cap \left(\cap_{k=1}^{\ell-1}\mathcal{E}_0^{(k)}\right)\right] - \sum_{\ell=1}^{L}\mathbb{P}\left[\left(\mathcal{E}_0^{(\ell)}\right)^c\right]. \tag{71}$$

For $1 \leq \ell \leq L$, since $\mathbf{W}^{(\ell)}$ has i.i.d. standard Gaussian entries, by Vershynin (2019, Theorem 4.4.5),

$$\mathbb{P}\left[\left(\mathcal{E}_0^{(\ell)}\right)^c\right] \leq \exp\left(-\Omega(m)\right). \tag{72}$$

Next we condition on $\left\{\mathbf{W}^{(k)}\right\}_{k=1}^{\ell-1}$ such that $\left(\cap_{k=1}^{\ell-1}\mathcal{E}_1^{(k)}\right) \cap \left(\cap_{k=1}^{\ell-1}\mathcal{E}_0^{(k)}\right)$ holds. Since $w_i^{(\ell)}$'s are independent of $\left\{\mathbf{W}^{(k)}\right\}_{k=1}^{\ell-1}$, $\langle w_i^{(\ell)}, o^{(\ell-1)}(x)\rangle \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \left\|o^{(\ell-1)}(x)\right\|_2^2)$. Therefore,

$$\begin{aligned}&\|\sigma(\langle w_i^{(\ell)}, o^{(\ell-1)}(x_0)\rangle)\sigma(\langle w_i^{(\ell)}, o^{(\ell-1)}(x_0')\rangle)\|_{\psi_1}\\ &\leq \|\sigma(\langle w_i^{(\ell)}, o^{(\ell-1)}(x_0)\rangle)\|_{\psi_2}\|\sigma(\langle w_i^{(\ell)}, o^{(\ell-1)}(x_0')\rangle)\|_{\psi_2} \leq c_0^{2\ell-2},\end{aligned}$$

where $\|X\|_{\psi_2} \triangleq \inf\left\{t > 0 : \mathbb{E}\left[\exp\left(X^2/t^2\right)\right] \leq 2\right\}$, and $\|X\|_{\psi_1} \triangleq \inf\left\{t > 0 : \mathbb{E}\left[\exp\left(|X|/t\right)\right] \leq 2\right\}$ for any random variable X; the first inequality holds by Vershynin (2019, Lemma 2.7.7) and the second inequality holds by Lemma 30 under $\cap_{k=1}^{\ell-1}\mathcal{E}_0^{(k)}$.

It follows from the sub-exponential concentration inequality (Lemma 18) that for any fixed $(x_0, x_0') \in V_0$, (70) holds with probability at least $1 - \exp\left(-\Omega(m^{1/3})\right)$. Further taking union bounds over V_0, we have

$$\mathbb{P}\left[\mathcal{E}_1^{(\ell)} \,\middle|\, \left(\cap_{k=1}^{\ell-1}\mathcal{E}_1^{(k)}\right) \cap \left(\cap_{k=1}^{\ell-1}\mathcal{E}_0^{(k)}\right)\right] \geq 1 - \exp\left(O(d\log m) - \Omega(m^{1/3})\right). \tag{73}$$

Plugging (73) and (72) into (71), we have

$$\begin{aligned}\mathbb{P}\left[\mathcal{E}_0 \cap \mathcal{E}_1\right] &\geq 1 - L\exp\left(O(d\log m) - \Omega(m^{1/3})\right) - L\exp\left(-\Omega(m)\right)\\ &\geq 1 - L\exp\left(O(d\log m) - \Omega(m^{1/3})\right).\end{aligned} \tag{74}$$

It remains to show (30) under $\mathcal{E}_0 \cap \mathcal{E}_1$. Fix any (x, x') and denote $(x_0, x_0') \in V_0 \times V_0$ such that $\|x - x_0\|_2 \leq \frac{1}{m^2}$ and $\|x' - x_0'\|_2 \leq \frac{1}{m^2}$. For any $0 \leq \ell \leq L-1$, by the triangle

inequality,

$$\begin{aligned}
&\left|\langle o^{(\ell+1)}(x), o^{(\ell+1)}(x')\rangle - \mathbb{E}\left[\sigma(U^{(\ell+1)}(x))\sigma(U^{(\ell+1)}(x'))\right]\right| \\
&\le \underbrace{\left|\langle o^{(\ell+1)}(x), o^{(\ell+1)}(x')\rangle - \langle o^{(\ell+1)}(x_0), o^{(\ell+1)}(x_0')\rangle\right|}_{\text{(I)}} \\
&+ \underbrace{\left|\langle o^{(\ell+1)}(x_0), o^{(\ell+1)}(x_0')\rangle - \mathbb{E}_w\left[\sigma(\langle w, o^{(\ell)}(x_0)\rangle)\sigma(\langle w, o^{(\ell)}(x_0')\rangle)\right]\right|}_{\text{(II)}} \\
&+ \underbrace{\left|\mathbb{E}_w\left[\sigma(\langle w, o^{(\ell)}(x_0)\rangle)\sigma(\langle w, o^{(\ell)}(x_0')\rangle)\right] - \mathbb{E}\left[\sigma(U^{(\ell+1)}(x))\sigma(U^{(\ell+1)}(x'))\right]\right|}_{\text{(III)}}, \qquad (75)
\end{aligned}$$

where

$$\begin{aligned}
&(U^{(\ell+1)}(x), U^{(\ell+1)}(x')) \sim \mathcal{N}\left(0, \Sigma^{(\ell)}(x, x')\right) \\
&\Sigma^{(\ell)}(x, x') \triangleq \begin{pmatrix} \mathbb{E}\left[\sigma^2(U^{(\ell)}(x))\right] & \mathbb{E}\left[\sigma(U^{(\ell)}(x))\sigma(U^{(\ell)}(x'))\right] \\ \mathbb{E}\left[\sigma(U^{(\ell)}(x))\sigma(U^{(\ell)}(x'))\right] & \mathbb{E}\left[\sigma^2(U^{(\ell)}(x'))\right] \end{pmatrix} \qquad (76)
\end{aligned}$$

with $\Sigma^{(0)}(x, x') = \begin{pmatrix} 1 & \langle x, x'\rangle \\ \langle x, x'\rangle & 1 \end{pmatrix}$.

To bound (I), note that for any y, z, y', z', by the triangle inequality and Cauchy-Schwartz inequality, we have

$$\left|\langle y, y'\rangle - \langle z, z'\rangle\right| \le \|y - z\| \left\|y'\right\| + \left\|y' - z'\right\| \|z\|. \qquad (77)$$

Thus, we get

$$\begin{aligned}
\text{(I)} &\le \sup_{x,x'} \left(\left\|o^{(\ell+1)}(x) - o^{(\ell+1)}(x_0)\right\|_2 \left\|o^{(\ell+1)}(x')\right\|_2 + \left\|o^{(\ell+1)}(x') - o^{(\ell+1)}(x_0')\right\|_2 \left\|o^{(\ell+1)}(x_0)\right\|_2\right) \\
&\le \frac{2c_0^{2\ell+2}}{m^2}. \qquad (78)
\end{aligned}$$

where the last inequality holds under $\mathcal{E}_0$ by Lemma 30.

For term (II), recall by (13) that

$$\langle o^{(\ell+1)}(x_0), o^{(\ell+1)}(x_0')\rangle = \frac{1}{m}\sum_{i=1}^{m} \sigma(\langle w_i^{(\ell+1)}, o^{(\ell)}(x_0)\rangle)\sigma(\langle w_i^{(\ell+1)}, o^{(\ell)}(x_0')\rangle).$$

Thus, under $\mathcal{E}_1$,

$$\text{(II)} \le C\frac{c_0^{2\ell}}{m^{1/3}}. \qquad (79)$$

To bound (III), note that conditioning on $o^{(\ell)}$, $\left(\langle w, o^{(\ell)}(x_0)\rangle, \langle w, o^{(\ell)}(x_0')\rangle\right)$ is a bivariate normal random vector with mean 0 and covariance

$$\mathbf{A}^{(\ell)}(x_0, x_0') = \begin{pmatrix} \left\|o^{(\ell)}(x_0)\right\|_2^2 & \langle o^{(\ell)}(x_0), o^{(\ell)}(x_0')\rangle \\ \langle o^{(\ell)}(x_0), o^{(\ell)}(x_0')\rangle & \left\|o^{(\ell)}(x_0')\right\|_2^2 \end{pmatrix}.$$

Thus, by Lemma 31, we have

$$\begin{aligned}(\text{III}) &= O\left(\left\|\mathbf{A}^{(\ell)}(x_0,x_0')-\Sigma^{(\ell)}(x,x')\right\|_\infty\right)\\ &\le O\left(\left\|\mathbf{A}^{(\ell)}(x_0,x_0')-\mathbf{A}^{(\ell)}(x,x')\right\|_\infty\right)+O\left(\left\|\mathbf{A}^{(\ell)}(x,x')-\Sigma^{(\ell)}(x,x')\right\|_\infty\right)\\ &= O\left(\frac{c_0^{2\ell}}{m^2}\right)+O\left(\left\|\mathbf{A}^{(\ell)}(x,x')-\Sigma^{(\ell)}(x,x')\right\|_\infty\right) \end{aligned} \tag{80}$$

where the last equality holds by (77).

Plugging (78), (79) and (80) into (75) and taking supremum over (x,x'), we get

$$\begin{aligned}&\sup_{x,x'}\left|\langle o^{(\ell+1)}(x),o^{(\ell+1)}(x')\rangle-\mathbb{E}\left[\sigma(U^{(\ell+1)}(x))\sigma(U^{(\ell+1)}(x')\right]\right|\\ &\le O\left(\frac{c_0^{2\ell}}{m^{1/3}}\right)+O\left(\sup_{x,x'}\left\|\mathbf{A}^{(\ell)}(x,x')-\Sigma^{(\ell)}(x,x')\right\|_\infty\right).\end{aligned} \tag{81}$$

By definition of $\mathbf{A}^{(\ell)}$ and $\Sigma^{(\ell)}$, for $\ell\ge 1$,

$$\sup_{x,x'}\left\|\mathbf{A}^{(\ell)}(x,x')-\Sigma^{(\ell)}(x,x')\right\|_\infty \le O\left(\sup_{x,x'}\left|\langle o^{(\ell)}(x),o^{(\ell)}(x')\rangle-\mathbb{E}\left[\sigma(U^{(\ell)}(x))\sigma(U^{(\ell)}(x'))\right]\right|\right). \tag{82}$$

Recursively applying (81) and (82), and noting $\sup_{x,x'}\left\|\mathbf{A}^{(0)}(x,x')-\Sigma^{(0)}(x,x')\right\|_\infty=0$, we complete the proof of (30).

B.2 Proof of Lemma 7: Concentration of $a^\top\mathbf{G}_L^{(\ell)}(x,x')a$ on $\mathrm{Tr}\left(\mathbf{G}_L^{(\ell)}(x,x')\right)$

Recall from the discussion in Section 5 that a crucial step in the proof of Lemma 7 is to bound the number of different matrices $\mathbf{G}_k^{(\ell)}(x,x')$ by varying x,x' through bounding the cardinality of $\mathcal{D}_k$ where $\mathbf{G}_k^{(\ell)}(x,x')=\left[\mathbf{V}_k^{(\ell)}(x)\right]^\top\mathbf{V}_k^{(\ell)}(x')$ from (12),

$$\left[\mathbf{V}_k^{(\ell)}(x)\right]^\top\triangleq\frac{1}{\sqrt{m}}\mathbf{D}^{(k)}(x)\mathbf{W}^{(k)}\cdots\frac{1}{\sqrt{m}}\mathbf{D}^{(\ell+1)}(x)\mathbf{W}^{(\ell+1)}\frac{1}{\sqrt{m}}\mathbf{D}^{(\ell)}(x)$$

from (8) and $\mathcal{D}_k=\left\{(\mathbf{D}^{(1)}(x),\cdots,\mathbf{D}^{(k)}(x)):x\in\mathbb{S}^{d-1}\right\}$.

Lemma 32 *Fix any $k>0$ and $\ell\le k$. For any fixed $\left\{\mathbf{W}^{(r)}\right\}_{r=1}^k$, we have $|\mathcal{D}_k|\le m^{dk}$, and hence $|\mathcal{G}_k^{(\ell)}|\le m^{2dk}$ where $\mathcal{G}_k^{(\ell)}(\mathbf{W}^{(1)},\cdots,\mathbf{W}^{(k)})\triangleq\left\{\mathbf{G}_k^{(\ell)}(x,x'):x,x'\in\mathbb{S}^{d-1}\right\}$.*

Intuitively, Lemma 32 implies that while there are infinite many different choices of x,x' on the unit sphere $\mathbb{S}^{d-1}$, the number of different matrices $\mathbf{G}_k^{(\ell)}(x,x')$ is finite for any fixed $\left\{\mathbf{W}^{(r)}\right\}_{r=1}^k$.

Before presenting the proof of Lemma 32, we provide a proof sketch for the ease of reading. To prove Lemma 32, note that given fixed $\left\{\mathbf{W}^{(r)}\right\}_{r=1}^k$, by the definition of $\mathcal{G}_k^{(\ell)}$, we

have $|\mathcal{G}_k^{(\ell)}| \le |\mathcal{D}_k|^2$. The proof is then completed by showing $|\mathcal{D}_k| \le m^{dk}$. To obtain this, we show $|\mathcal{D}_1| \le m^d$ and $|\mathcal{D}_k| \le m^d |\mathcal{D}_{k-1}|$ for all k. The key of proving $|\mathcal{D}_k| \le m^d |\mathcal{D}_{k-1}|$ lies on a refinement idea illustrated in Figure 8. In particular, we partition $\mathbb{S}^{d-1}$ into disjoint V_j for $j = 1, 2, \cdots, |\mathcal{D}_{k-1}|$ such that $\cup_j V_j = \mathbb{S}^{d-1}$ and $V_j \cap V_{j'} = \emptyset$ for all $j \neq j'$, and that for any x within the same V_j, $(\mathbf{D}^{(1)}(x), \cdots \mathbf{D}^{(k-1)}(x))$ is the same. We then refine V_j so that within each subregion after refinement, $\mathbf{D}^{(k)}(x)$ is the same. Here, we crucially show $|\left\{\mathbf{D}^{(k)}(x) : x \in V_j\right\}| \le m^d$ for all j, i.e., the refinement within each V_j cannot exceed m^d and hence conclude $|\mathcal{D}_k| \le m^d |\mathcal{D}_{k-1}|$.

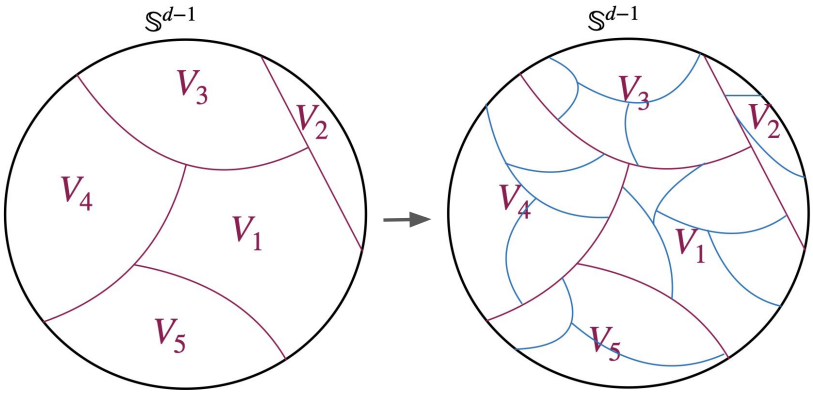


Figure 8: Illustration of the key idea in showing $|\mathcal{D}_k| \le m^d |\mathcal{D}_{k-1}|$. Left hand side shows the partition of $\mathbb{S}^{d-1}$ into disjoint $\{V_j\}_{j=1}^{|\mathcal{D}_{k-1}|}$ so that for any x within V_j, $(\mathbf{D}^{(1)}(x), \cdots \mathbf{D}^{(k-1)}(x))$ is the same. We then refine each V_j to obtain the right hand side so that within each refined sub-region in V_j, $\mathbf{D}^{(k)}(x)$ is also the same. We then show the number of such sub-region cannot exceed m^d for any V_j, which leads to $|\mathcal{D}_k| \le m^d |\mathcal{D}_{k-1}|$.

Proof [Proof of Lemma 32] Throughout the proof, we fix $\left\{\mathbf{W}^{(r)}\right\}_{r=1}^k$. Since $\left\{\mathbf{W}^{(r)}\right\}_{r=1}^k$ is fixed, we have

$$\mathcal{G}_k^{(\ell)} \subset \Big\{ \mathbf{V}^\top \widetilde{\mathbf{V}} : \mathbf{V} = \mathbf{D}^{(k)} \mathbf{W}^{(k)} \cdots \mathbf{D}^{(\ell+1)} \mathbf{W}^{(\ell+1)} \mathbf{D}^{(\ell)}, \widetilde{\mathbf{V}} = \widetilde{\mathbf{D}}^{(k)} \mathbf{W}^{(k)} \cdots \widetilde{\mathbf{D}}^{(\ell+1)} \mathbf{W}^{(\ell+1)} \widetilde{\mathbf{D}}^{(\ell)},$$
$$\left(\mathbf{D}^{(1)}, \cdots, \mathbf{D}^{(k)}\right) \in \mathcal{D}_k, \left(\widetilde{\mathbf{D}}^{(1)}, \cdots, \widetilde{\mathbf{D}}^{(k)}\right) \in \mathcal{D}_k \Big\}.$$

Thus $|\mathcal{G}_k^{(\ell)}| \le |\mathcal{D}_k|^2$, and the proof is completed by the following claim:

$$|\mathcal{D}_k| \le m^{dk}. \tag{83}$$

To prove this claim, we first show $|\mathcal{D}_1| \le m^d$ and then show the recursion $|\mathcal{D}_k| \le m^d |\mathcal{D}_{k-1}|$ for all $k \ge 2$.

Step 1 bounding $|\mathcal{D}_1|$: Note that $\mathbf{D}^{(1)}(x)$ is diagonal whose i-th diagonal element equals $f_x(w_i^{(1)})$, where $f_x(w) = \mathbf{1}_{\{\langle w, x \rangle \ge 0\}}$. Thus, letting $\mathcal{F}^{(1)} = \left\{f_x(w) = \mathbf{1}_{\{\langle w, x \rangle \ge 0\}} : x \in \mathbb{S}^{d-1}\right\}$, we have

$$\left|\left\{\mathbf{D}^{(1)}(x) : x \in \mathbb{S}^{d-1}\right\}\right| = \left|\left\{\left(f(w_1^{(1)}), \cdots, f(w_m^{(1)})\right) : f \in \mathcal{F}^{(1)}\right\}\right|.$$

It follows from Lemma 22 that $|\left\{\mathbf{D}^{(1)}(x) : x \in \mathbb{S}^{d-1}\right\}| \leq m^{\mathsf{VC}(\mathcal{F}^{(1)})}$. By Hajek and Raginsky (2019, Proposition 7.1),

$$\mathsf{VC}(\mathcal{F}^{(1)}) = d. \tag{84}$$

As a result, we get $|\mathcal{D}_1| = \left|\left\{\mathbf{D}^{(1)}(x) : x \in \mathbb{S}^{d-1}\right\}\right| \leq m^d$.

Step 2 showing $|\mathcal{D}_k| \leq m^d|\mathcal{D}_{k-1}|$ for any $k \geq 2$: Partition $\mathbb{S}^{d-1}$ into disjoint V_j for $j = 1, 2, \cdots, |\mathcal{D}_{k-1}|$ such that for any x and x' within the same V_j,

$$\left(\mathbf{D}^{(1)}(x), \mathbf{D}^{(2)}(x), \cdots, \mathbf{D}^{(k-1)}(x)\right) = \left(\mathbf{D}^{(1)}(x'), \mathbf{D}^{(2)}(x'), \cdots, \mathbf{D}^{(k-1)}(x')\right).$$

Note that $\mathcal{D}_k = \cup_{j=1}^{|\mathcal{D}_{k-1}|} \left\{\left(\mathbf{D}^{(1)}(x), \cdots, \mathbf{D}^{(k)}(x)\right) : x \in V_j\right\}$. Thus,

$$|\mathcal{D}_k| \leq \sum_{j=1}^{|\mathcal{D}_{k-1}|} \left|\left\{\left(\mathbf{D}^{(1)}(x), \cdots, \mathbf{D}^{(k)}(x)\right) : x \in V_j\right\}\right| = \sum_{j=1}^{|\mathcal{D}_{k-1}|} \left|\left\{\mathbf{D}^{(k)}(x) : x \in V_j\right\}\right|, \tag{85}$$

where the last equality holds because $(\mathbf{D}^{(1)}(x), \cdots, \mathbf{D}^{(k-1)}(x))$ is the same for all $x \in V_j$.

It remains to bound $\left|\left\{\mathbf{D}^{(k)}(x) : x \in V_j\right\}\right|$. The i-th diagonal element of $\mathbf{D}^{(k)}(x)$ equals $f_x(w_i^{(k)})$, where $f_x(w) = \mathbf{1}_{\left\{\langle w, o^{(k-1)}(x)\rangle \geq 0\right\}}$. Therefore, by letting

$$\mathcal{F}_j^{(k)}(\mathbf{W}^{(1)}, \cdots, \mathbf{W}^{(k-1)}) \triangleq \left\{f_x(w) = \mathbf{1}_{\left\{\langle w, o^{(k-1)}(x)\rangle \geq 0\right\}} : x \in V_j\right\}, \tag{86}$$

we have

$$\left|\left\{\mathbf{D}^{(k)}(x) : x \in V_j\right\}\right| = \left|\left\{\left(f(w_1^{(k)}), \cdots, f(w_m^{(k)})\right) : f \in \mathcal{F}_j^{(k)}\right\}\right| \leq m^{\mathsf{VC}(\mathcal{F}_j^{(k)})}, \tag{87}$$

where the last inequality holds by Lemma 22.

Now we bound $\mathsf{VC}(\mathcal{F}_j^{(k)})$. Since $\left(\mathbf{D}^{(1)}(x), \ldots, \mathbf{D}^{(k-1)}(x)\right)$ is the same across all $x \in V_j$ and $\left\{\mathbf{W}^{(r)}\right\}_{r=1}^{k-1}$ are fixed, by definition of $o^{(k-1)}$, we have $o^{(k-1)}(x) = P_j x$, for all $x \in V_j$, where $P_j = \frac{1}{\sqrt{m}}\mathbf{D}^{(k-1)}(x)\mathbf{W}^{(k-1)} \cdots \frac{1}{\sqrt{m}}\mathbf{D}^{(1)}(x)\mathbf{W}^{(1)} \in \mathbb{R}^{m\times d}$ is some matrix independent of x. Therefore, $o^{(k-1)}(x)$ lies on the same d-dimensional subspace of $\mathbb{R}^m$ for all $x \in V_j$. By Hajek and Raginsky (2019, Proposition 7.1),

$$\mathsf{VC}\left(\mathcal{F}_j^{(k)}\right) = d. \tag{88}$$

It then follows from (87) that $\left|\left\{\mathbf{D}^{(k)}(x) : x \in V_j\right\}\right| \leq m^d$ for all $j = 1, 2, \cdots, |\mathcal{D}_{k-1}|$. Further plugging this bound into (85) yields that $|\mathcal{D}_k| \leq m^d|\mathcal{D}_{k-1}|$. ■

With Lemma 32, we prove the following key intermediate result by applying Hanson-Wright inequality with a union bound on $\mathbf{G}_k^{(\ell)}(x, x')$.

Lemma 33 *Let $Y = (Y_1, Y_2, \cdots, Y_m) \in \mathbb{R}^m$ be a random vector with mean zero, independent, sub-Gaussian coordinates with $\|Y_i\|_{\psi_2} \le C$. Assume Y is independent of $\left\{\mathbf{W}^{(r)}\right\}_{r=1}^k$. For any $\ell = 1, 2, \cdots, k$, we have,*

$$\begin{aligned}&\mathbb{P}\left[\sup_{\ell\in[k]}\sup_{x,x'}\left|Y^\top \mathbf{G}_k^{(\ell)}(x,x')Y - \mathrm{Tr}\left(\mathbf{G}_k^{(\ell)}(x,x')\right)\right| \ge \frac{c_0^{2k-2}}{m^{1/3}}\,\middle|\, \cap_{r=0}^k \mathcal{E}_0^{(r)}\right]\\ &\le k\exp\left(O(dk\log m) - \Omega(m^{1/3})\right).\end{aligned} \tag{89}$$

In the above lemma, by taking $Y = a$ and $k = L$, we obtain the uniform concentration of $a^\top \mathbf{G}_L^{(\ell)}(x,x')a$ on $\mathrm{Tr}\left(\mathbf{G}_L^{(\ell)}(x,x')\right)$ conditional on $\cap_{r=1}^L \mathcal{E}_0^{(r)}$, which readily implies Lemma 7. Furthermore, by taking $Y = w_i^{(k+1)}$, we obtain the uniform concentration of $\left[w_i^{(k+1)}\right]^\top \mathbf{G}_k^{(\ell)}(x,x')w_i^{(k+1)}$ on $\mathrm{Tr}\left(\mathbf{G}_k^{(\ell)}(x,x')\right)$ for any $i \in [m]$, where $w_i^{(k+1)}$ is the i-th row of $\mathbf{W}^{(k+1)}$. That turns out to be instrumental in the proof of Lemma 8.

Proof [Proof of Lemma 33] Fix arbitrary $\ell \le k$. We condition on $\left\{\mathbf{W}^{(r)}\right\}_{r=1}^k$ such that $\cap_{r=1}^k \mathcal{E}_0^{(r)}$ holds. Under $\cap_{r=1}^k \mathcal{E}_0^{(r)}$, for any $x \in \mathbb{S}^{d-1}$, we have

$$\begin{aligned}\left\|\mathbf{V}_k^{(\ell)}(x)\right\|_2 &= \left\|\left[\frac{1}{\sqrt{m}}\mathbf{D}^{(k)}(x)\mathbf{W}^{(k)}\cdots\frac{1}{\sqrt{m}}\mathbf{D}^{(\ell+1)}(x)\mathbf{W}^{(\ell+1)}\frac{1}{\sqrt{m}}\mathbf{D}^{(\ell)}(x)\right]^\top\right\|_2\\ &\le \frac{c_0^{k-\ell}}{\sqrt{m}}.\end{aligned}$$

By definition of $\mathbf{G}_k^{(\ell)}$, we get

$$\left\|\mathbf{G}_k^{(\ell)}(x,x')\right\|_2 = \left\|\left[\mathbf{V}_k^{(\ell)}(x)\right]^\top \mathbf{V}_k^{(\ell)}(x')\right\|_2 \le \left\|\mathbf{V}_k^{(\ell)}(x)\right\|_2\left\|\mathbf{V}_k^{(\ell)}(x')\right\|_2 \le \frac{c_0^{2k-2\ell}}{m}. \tag{90}$$

and

$$\|\mathbf{G}_k^{(\ell)}(x,x')\|_{\mathrm{F}} \le \sqrt{m}\left\|\mathbf{G}_k^{(\ell)}(x,x')\right\|_2 \le \frac{c_0^{2k-2\ell}}{\sqrt{m}}. \tag{91}$$

Since Y has mean zero and is independent of $\left\{\mathbf{W}^{(r)}\right\}_{r=1}^k$, we have

$$\mathbb{E}\left[Y^\top \mathbf{G}_k^{(\ell)}(x,x')Y\,\middle|\,\left\{\mathbf{W}^{(r)}\right\}_{r=1}^k\right] = \mathrm{Tr}\left(\mathbf{G}_k^{(\ell)}(x,x')\right).$$

Thus, under event $\cap_{r=1}^k \mathcal{E}_0^{(r)}$, by Hanson-Wright inequality, we have for any fixed x, x',

$$\begin{aligned}&\mathbb{P}\left[\left|Y^\top \mathbf{G}_k^{(\ell)}(x,x')Y - \mathrm{Tr}\left(\mathbf{G}_k^{(\ell)}(x,x')\right)\right| > \frac{c_0^{2k-2\ell}}{m^{1/3}}\,\middle|\,\left\{\mathbf{W}^{(r)}\right\}_{r=1}^k\right]\\ &\le 2\exp\left(-C\min\left(\frac{c_0^{4k-4\ell}m^{-2/3}}{\|\mathbf{G}_k^{(\ell)}(x,x')\|_{\mathrm{F}}^2}, \frac{c_0^{2k-2\ell}m^{-1/3}}{\left\|\mathbf{G}_k^{(\ell)}(x,x')\right\|_2}\right)\right) = \exp\left(-\Omega(m^{1/3})\right).\end{aligned}$$

By Lemma 32, we have $|\mathcal{G}_k^{(\ell)}(\mathbf{W}^{(1)}, \cdots, \mathbf{W}^{(k)})| \leq m^{2dk}$. Taking union bounds over all possible $\mathbf{G}_k^{(\ell)}$, we have under event $\cap_{r=1}^k \mathcal{E}_0^{(r)}$,

$$\begin{aligned}
&\mathbb{P}\left[\sup_{x,x'}\left|Y^\top \mathbf{G}_k^{(\ell)}(x,x')Y - \operatorname{Tr}\left(\mathbf{G}_k^{(\ell)}(x,x')\right)\right| > \frac{c_0^{2k-2\ell}}{m^{1/3}}\middle|\left\{\mathbf{W}^{(r)}\right\}_{r=1}^k\right] \\
&= \mathbb{P}\left[\sup_{\mathbf{G}\in\mathcal{G}_k^{(\ell)}}\left|Y^\top \mathbf{G} Y - \operatorname{Tr}\left(\mathbf{G}\right)\right| > \frac{c_0^{2k-2\ell}}{m^{1/3}}\middle|\left\{\mathbf{W}^{(r)}\right\}_{r=1}^k\right] \\
&= m^{2dk}\exp\left(-\Omega(m^{1/3})\right) = \exp\left(O\left(dk\log m\right) - \Omega(m^{1/3})\right).
\end{aligned}$$

Further take union bounds over ℓ, we obtain that

$$\begin{aligned}
&\mathbb{P}\left[\sup_{\ell\in[k]}\sup_{x,x'}\left|Y^\top \mathbf{G}_k^{(\ell)}(x,x')Y - \operatorname{Tr}\left(\mathbf{G}_k^{(\ell)}(x,x')\right)\right| > \frac{c_0^{2k-2\ell}}{m^{1/3}}\middle|\left\{\mathbf{W}^{(r)}\right\}_{r=1}^k\right] \\
&= k\exp\left(O(dk\log m) - \Omega(m^{1/3})\right).
\end{aligned}$$

Taking the average of $\left\{\mathbf{W}^{(r)}\right\}_{r=1}^k$ on the event $\cap_{r=1}^k \mathcal{E}_0^{(r)}$, we get the desired conclusion. ■

Proof of Lemma 7: Denote

$$\mathcal{E}_2 = \left\{\sup_{\ell\in[L]}\sup_{x,x'}\left|a^\top \mathbf{G}_L^{(\ell)}(x,x')a - \operatorname{Tr}\left(\mathbf{G}_L^{(\ell)}(x,x')\right)\right| \leq C\frac{c_0^{2L-2}}{m^{1/3}}\right\}.$$

Note that a is mean zero, sub-Gaussian and is independent of $\left\{\mathbf{W}^{(k)}\right\}_{k=1}^L$. Thus, by Lemma 33, we have

$$\mathbb{P}\left[\mathcal{E}_2|\cap_{k=1}^L \mathcal{E}_0^{(k)}\right] \geq 1 - L\exp\left(O(dL\log m) - \Omega(m^{1/3})\right)$$

From (72), we have

$$\mathbb{P}\left[\cap_{k=1}^L \mathcal{E}_0^{(k)}\right] \geq 1 - L\exp\left(-\Omega(m)\right). \tag{92}$$

Therefore,

$$\begin{aligned}
\mathbb{P}\left[\mathcal{E}_2\right] &\geq \mathbb{P}\left[\mathcal{E}_2|\cap_{k=1}^L \mathcal{E}_0^{(k)}\right]\mathbb{P}\left[\cap_{k=1}^L \mathcal{E}_0^{(k)}\right] \\
&\geq \left(1 - L\exp\left(O(dL\log m) - \Omega(m^{1/3})\right)\right)\left(1 - L\exp(-\Omega(m))\right) \\
&\geq 1 - \exp\left(O(dL\log m) - \Omega(m^{1/3})\right).
\end{aligned}$$

B.3 Proof of Lemma 8: Concentration of $\mathrm{Tr}\left(\mathbf{G}_L^{(\ell)}(x,x')\right)$ on $q_L^{(\ell)}(x,x')$

Recall from (19) that

$$q_L^{(\ell)}(x,x') = \frac{\pi - \arccos \rho^{(L-1)}(x,x')}{2\pi} q_{L-1}^{(\ell)}(x,x'), \quad \forall \ell \le L,$$
$$q_L^{(L)}(x,x') = \frac{\pi - \arccos \rho^{(L-1)}(x,x')}{2\pi},$$

where

$$\rho^{(L-1)}(x,x') = \frac{\mathbb{E}\left[\sigma(U^{(L-1)}(x))\sigma(U^{(L-1)}(x'))\right]}{\sqrt{\mathbb{E}\left[\sigma^2(U^{(L-1)}(x))\right]}\sqrt{\mathbb{E}\left[\sigma^2(U^{(L-1)}(x'))\right]}}, \tag{93}$$

and $U^{(L-1)}$ is defined in (76).

To prove Lemma 8, we crucially show the concentration of $\mathrm{Tr}\left(\mathbf{G}_L^{(\ell)}(x,x')\right)$ on $q_{L-1}^{(L-1)}(x,x')\,\mathrm{Tr}\left(\mathbf{G}_{L-1}^{(\ell)}(x,x')\right)$. Then, by repeatedly applying this recursive relation of $\mathrm{Tr}\left(\mathbf{G}_{L-1}^{(\ell)}(x,x')\right)$, we obtain the concentration of $\mathrm{Tr}\left(\mathbf{G}_L^{(\ell)}(x,x')\right)$ on $q_L^{(\ell)}(x,x')$.

Proving the concentration of $\mathrm{Tr}\left(\mathbf{G}_L^{(\ell)}(x,x')\right)$ on $q_{L-1}^{(L-1)}(x,x')\,\mathrm{Tr}\left(\mathbf{G}_{L-1}^{(\ell)}(x,x')\right)$ consists of the following three steps.

Step 1: As the first step, we show the concentration of $\mathrm{Tr}\left(\mathbf{G}_L^{(\ell)}(x,x')\right)$ on

$$\frac{1}{m}\sum_{i=1}^m \mathbf{1}_{\left\{\langle w_i^{(L)}, o^{(L-1)}(x)\rangle \ge 0\right\}} \mathbf{1}_{\left\{\langle w_i^{(L)}, o^{(L-1)}(x')\rangle \ge 0\right\}} \mathrm{Tr}\left(\mathbf{G}_{L-1}^{(\ell)}(x,x')\right).$$

This is achieved by applying Lemma 33.

In the next two steps, we show the concentration of $\frac{1}{m}\sum_{i=1}^m \mathbf{1}_{\left\{\langle w_i^{(L)}, o^{(L-1)}(x)\rangle \ge 0\right\}} \mathbf{1}_{\left\{\langle w_i^{(L)}, o^{(L-1)}(x')\rangle \ge 0\right\}}$ on $q_{L-1}^{(\ell-1)}(x,x')$.

Step 2: Here, we show the concentration of $\frac{1}{m}\sum_{i=1}^m \mathbf{1}_{\left\{\langle w_i^{(L)}, o^{(L-1)}(x)\rangle \ge 0\right\}} \mathbf{1}_{\left\{\langle w_i^{(L)}, o^{(L-1)}(x')\rangle \ge 0\right\}}$ on $\mathbb{E}_{w\sim\mathcal{N}(0,\mathbf{I})}\left[\mathbf{1}_{\left\{\langle w, o^{(L-1)}(x)\rangle \ge 0\right\}} \mathbf{1}_{\left\{\langle w, o^{(L-1)}(x')\rangle \ge 0\right\}}\right]$ by the following lemma.

Lemma 34 *Let $\{w_i\}_{i=1}^m \in \mathbb{R}^d$ be i.i.d. Gaussian random vectors with standard normal entries. Define for $0 \le \ell \le L-1$,*

$$h_{x,x'}^{(\ell+1)}(z_1,\cdots,z_m) \triangleq \left|\frac{1}{m}\sum_{i=1}^m \mathbf{1}_{\left\{\langle z_i, o^{(\ell)}(x)\rangle \ge 0\right\}} \mathbf{1}_{\left\{\langle z_i, o^{(\ell)}(x')\rangle \ge 0\right\}} - \mathbb{E}_{w\sim\mathcal{N}(0,\mathbf{I})}\left[\mathbf{1}_{\left\{\langle w, o^{(\ell)}(x)\rangle \ge 0\right\}} \mathbf{1}_{\left\{\langle w, o^{(\ell)}(x')\rangle \ge 0\right\}}\right]\right|.$$

Conditioning on $\left\{\mathbf{W}^{(k)}\right\}_{k=1}^{\ell}$, with probability at least $1-\exp(-2m^{1/3})$,

$$\sup_{x,x'} h_{x,x'}^{(\ell+1)}(w_1,\cdots,w_m) \le C\sqrt{\frac{d\,(1+\ell\log m)}{m}} + \frac{1}{m^{1/3}}.$$

Proof [Proof of Lemma 34] Throughout the proof, we condition on $\left\{\mathbf{W}^{(k)}\right\}_{k=1}^{\ell}$. We first show that

$$\begin{aligned}&\mathbb{P}\left[\sup_{x,x'} h_{x,x'}^{(\ell+1)}(w_1,\cdots,w_m) \le \mathbb{E}\left[\sup_{x,x'} h_{x,x'}^{(\ell+1)}(w_1,\cdots,w_m)\right] + \frac{1}{m^{1/3}}\right]\\ &\ge 1-\exp\left(-2m^{1/3}\right).\end{aligned} \tag{94}$$

To prove this, note that by the triangle inequality, for arbitrary i, $\exists (x_0,x_0')\in\mathbb{S}^{d-1}$ such that

$$\begin{aligned}&\left|\sup_{x,x'} h^{(\ell+1)}(z_1,\cdots,z_{i-1},z_i,z_{i+1},\cdots,z_m) - \sup_{x,x'} h^{(\ell+1)}(z_1,\cdots,z_{i-1},z_i',z_{i+1},\cdots,z_m)\right|\\ &\le \frac{1}{m}\left|\mathbf{1}_{\left\{z_i,o^{(\ell)}(x_0)\rangle\ge 0\right\}}\mathbf{1}_{\left\{z_i,o^{(\ell)}(x_0')\rangle\ge 0\right\}} - \mathbf{1}_{\left\{z_i',o^{(\ell)}(x_0)\rangle\ge 0\right\}}\mathbf{1}_{\left\{z_i',o^{(\ell)}(x_0')\rangle\ge 0\right\}}\right|\\ &\le \frac{1}{m}.\end{aligned} \tag{95}$$

Therefore, (94) follows by applying McDiarmid's inequality (Lemma 17).

To complete the proof of Lemma 34, it remains to show

$$\mathbb{E}\left[\sup_{x,x'} h_{x,x'}^{(\ell+1)}(w_1,\cdots,w_m)\right] = O\left(\sqrt{\frac{d\,(1+\ell\log m)}{m}}\right). \tag{96}$$

Since $\{w_i\}_{i=1}^m$ are i.i.d. conditional on $\left\{\mathbf{W}^{(r)}\right\}_{r=1}^{\ell}$, by Lemma 21,

$$\mathbb{E}\left[\sup_{x,x'} h_{x,x'}^{(\ell+1)}(w_1,\cdots,w_m)\right] \le C\sqrt{\frac{\mathsf{VC}(\mathcal{H}^{(\ell+1)})}{m}}, \tag{97}$$

where

$$\mathcal{H}^{(\ell+1)}(\mathbf{W}^{(1)},\cdots,\mathbf{W}^{(\ell)}) = \left\{\alpha_{x,x'}^{(\ell)}(w) : x,x'\in\mathbb{S}^{d-1}\right\},$$

with $\alpha_{x,x'}^{(\ell)}(w) = \mathbf{1}_{\left\{\langle w,o^{(\ell)}(x)\rangle\ge 0\right\}}\mathbf{1}_{\left\{\langle w,o^{(\ell)}(x')\rangle\ge 0\right\}}$. Now we bound $\mathsf{VC}\left(\mathcal{H}^{(\ell+1)}\right)$. Let $\mathcal{F}^{(\ell+1)} = \left\{f_x(w) = \mathbf{1}_{\left\{\langle w,o^{(\ell)}(x)\rangle\ge 0\right\}} : x\in\mathbb{S}^{d-1}\right\}$. Then for any $\alpha(w)\in\mathcal{H}^{(\ell+1)}$, we can always find $f(w)$ and $g(w)$ in $\mathcal{F}^{(\ell+1)}$ such that $\alpha = f\times g$. Thus, by Lemma 20, $\mathsf{VC}(\mathcal{H}^{(\ell+1)}) \le C\mathsf{VC}(\mathcal{F}^{(\ell+1)})$ for some universal constant C. We claim that $\mathsf{VC}\left(\mathcal{F}^{(\ell+1)}\right) = O\left(d(1+\ell\log m)\right)$. Plugging this bound into the above displayed equation and combining it with (97), we obtain (96).

Finally, we prove the claim. When $\ell = 0$, by (84), we have $\mathsf{VC}\left(\mathcal{F}^{(1)}\right) = d$.

Now consider the case when $\ell\ge 1$. Similar to the proof of Lemma 32, we decompose $\mathbb{S}^{d-1}$ into $\mathcal{V} = \{V_j, j=1,2,\cdots,|\mathcal{D}_\ell|\}$ where $\cup V_j = \mathbb{S}^{d-1}$ and $V_j\cap V_{j'} = \emptyset$ whenever $j\ne j'$ such that for any x,x' within the same V_j,

$$\left(\mathbf{D}^{(1)}(x),\cdots,\mathbf{D}^{(\ell)}(x)\right) = \left(\mathbf{D}^{(1)}(x'),\cdots,\mathbf{D}^{(\ell)}(x')\right).$$

Recall from (86) that $\mathcal{F}_j^{(\ell+1)} \triangleq \left\{ f_x(w) = \mathbf{1}_{\left\{\langle w, o^{(\ell)}(x)\rangle \geq 0\right\}} : x \in V_j \right\}$. Since $\cup V_j = \mathbb{S}^{d-1}$, we have $\mathcal{F}^{(\ell+1)} = \cup_{j=1}^{|\mathcal{D}_\ell|} \mathcal{F}_j^{(\ell+1)}$. From (88), we have $\mathsf{VC}(\mathcal{F}_j^{(\ell+1)}) \leq d$ for all j. Thus, by Lemma 24, we have

$$\mathsf{VC}(\mathcal{F}^{(\ell+1)}) = O\left(\max\left\{d \log d, \log |\mathcal{D}_\ell|\right\}\right) = O\left(\max\left\{d \log d, d\ell \log m\right\}\right) = O\left(d\ell \log m\right),$$

where the second equality holds by (83) which gives $|\mathcal{D}_\ell| \leq m^{d\ell}$ and the last equality holds since $\log d \leq \ell \log m$ as $m \geq d$. ■

Step 3: Note that $\mathbb{E}_{w\sim\mathcal{N}(0,\mathbf{I})}\left[\mathbf{1}_{\left\{\langle w, o^{(L-1)}(x)\rangle \geq 0\right\}} \mathbf{1}_{\left\{\langle w, o^{(L-1)}(x')\rangle \geq 0\right\}}\right] = \frac{\pi - \arccos \widehat{\rho}^{(L-1)}(x,x')}{2\pi}$ where

$$\widehat{\rho}^{(L-1)}(x,x') \triangleq \left\langle \frac{o^{(L-1)}(x)}{\left\|o^{(L-1)}(x)\right\|_2}, \frac{o^{(L-1)}(x')}{\left\|o^{(L-1)}(x')\right\|_2} \right\rangle. \tag{98}$$

To show the the concentration of $\mathbb{E}_{w\sim\mathcal{N}(0,\mathbf{I})}\left[\mathbf{1}_{\left\{\langle w, o^{(L-1)}(x)\rangle \geq 0\right\}} \mathbf{1}_{\left\{\langle w, o^{(L-1)}(x')\rangle \geq 0\right\}}\right]$ on $q_{L-1}^{(L-1)}(x,x')$, we show the concentration of $\arccos \widehat{\rho}^{(L-1)}(x,x')$ on $\arccos \rho^{(L-1)}(x,x')$ through the following corollary.

Corollary 35 *Fix any $\ell \leq L$. Under $\left(\cap_{k=1}^{\ell} \mathcal{E}_0^{(k)}\right) \cap \left(\cap_{k=1}^{\ell} \mathcal{E}_1^{(k)}\right)$ where $\mathcal{E}_0^{(k)}$ is defined in (69) and $\mathcal{E}_1^{(k)}$ is defined in (70),*

$$\sup_{x,x'} \left|\widehat{\rho}^{(\ell)}(x,x') - \rho^{(\ell)}(x,x')\right| = O\left(\frac{\ell C^\ell}{m^{1/3}}\right).$$

and hence

$$\sup_{x,x'} \left|\arccos \rho^{(\ell)}(x,x') - \arccos \widehat{\rho}^{(\ell)}(x,x')\right| = O\left(\frac{\sqrt{\ell} C^\ell}{m^{1/6}}\right).$$

To see why Corollary 35 holds, note that Lemma 6 implies both the numerator and denominator of $\widehat{\rho}^{(\ell)}$ are close to those of $\rho^{(\ell)}$. To obtain the second bound of Corollary 35, we prove that the arccos function is Hölder continuous of order $1/2$, that is,

$$\arccos z - \arccos y \leq \arccos(1 - (y - z)) \leq 3\sqrt{y - z}, \ \forall 0 \leq z \leq y \leq 1. \tag{99}$$

Combining the above with the first bound of Corollary 35 finishes the proof.

Proof [Proof of Corollary 35] We first prove $\widehat{\rho}^{(\ell)}(x,x')$ is close to $\rho^{(\ell)}(x,x')$. Note that for any y, y', z, z', by the triangle inequality we have

$$\left|\frac{y}{z} - \frac{y'}{z'}\right| \leq \left|\frac{y - y'}{z}\right| + \left|\frac{y'}{z} - \frac{y'}{z'}\right| = \left|\frac{y - y'}{z}\right| + \left|\frac{y'(z' - z)}{zz'}\right| \leq \left|\frac{y - y'}{z}\right| + \left|\frac{z' - z}{z}\right|,$$

where the last inequality holds under the assumption that $|y'/z'| \leq 1$. Taking $y = \mathbb{E}\left[\sigma(U^{(\ell)}(x))\sigma(U^{(\ell)}(x'))\right]$, $y' = \langle o^{(\ell)}(x), o^{(\ell)}(x')\rangle$, $z = \sqrt{\mathbb{E}\left[\sigma^2\left(U^{(\ell)}(x)\right)\right]}\sqrt{\mathbb{E}\left[\sigma^2\left(U^{(\ell)}(x')\right)\right]}$, and $z' = \left\|o^{(\ell)}(x)\right\|_2 \left\|o^{(\ell)}(x')\right\|_2$, by definition (93) and (98), we have

$$\widehat{\rho}^{(\ell)}(x,x') = \frac{y}{z}, \ \rho^{(\ell)}(x,x') = \frac{y'}{z'},$$

By Cauchy Schwartz inequality, $|\rho^{(\ell)}(x,x')| \le 1$. As a result, we have

$$\left|\widehat{\rho}^{(\ell)}(x,x') - \rho^{(\ell)}(x,x')\right| \le \underbrace{\left|\frac{\langle o^{(\ell)}(x), o^{(\ell)}(x')\rangle - \mathbb{E}\left[\sigma(U^{(\ell)}(x))\sigma(U^{(\ell)}(x'))\right]}{\sqrt{\mathbb{E}\left[\sigma^2\left(U^{(\ell)}(x)\right)\right]}\sqrt{\mathbb{E}\left[\sigma^2\left(U^{(\ell)}(x')\right)\right]}}\right|}_{\text{(I)}}$$
$$+ \underbrace{\left|\frac{\left\|o^{(\ell)}(x)\right\|_2\left\|o^{(\ell)}(x')\right\|_2 - \sqrt{\mathbb{E}\left[\sigma^2\left(U^{(\ell)}(x)\right)\right]}\sqrt{\mathbb{E}\left[\sigma^2\left(U^{(\ell)}(x')\right)\right]}}{\sqrt{\mathbb{E}\left[\sigma^2\left(U^{(\ell)}(x)\right)\right]}\sqrt{\mathbb{E}\left[\sigma^2\left(U^{(\ell)}(x')\right)\right]}}\right|}_{\text{(II)}}.$$

Note that

$$\text{(I)} = 2^\ell\left|\langle o^{(\ell)}(x), o^{(\ell)}(x')\rangle - \mathbb{E}\left[\sigma(U^{(\ell)}(x))\sigma(U^{(\ell)}(x'))\right]\right| = O\left(\frac{\ell C^\ell}{m^{1/3}}\right)$$

where the first equality holds by (38) which gives $\mathbb{E}\left[\sigma^2(U^{(\ell)}(x))\right] = \mathbb{E}\left[\sigma^2(U^{(\ell)}(x'))\right] = 2^{-\ell}$, $\forall x, x'$ and the last equality holds by Lemma 6.

To bound (II), by (38), we have

$$\text{(II)} = 2^\ell\left|\left\|o^{(\ell)}(x)\right\|_2\left\|o^{(\ell)}(x')\right\|_2 - \sqrt{\mathbb{E}\left[\sigma^2\left(U^{(\ell)}(x)\right)\right]}\sqrt{\mathbb{E}\left[\sigma^2\left(U^{(\ell)}(x')\right)\right]}\right|. \tag{100}$$

Note that for any $y, \widetilde{y}, z, \widetilde{z} \ge 0$,

$$\begin{aligned}|y\widetilde{y} - z\widetilde{z}| &\le \widetilde{y}\,|y-z| + z\,|\widetilde{y}-\widetilde{z}| \le \widetilde{y}\frac{\left|y^2-z^2\right|}{y+z} + z\frac{\left|\widetilde{y}^2-\widetilde{z}^2\right|}{\widetilde{y}+\widetilde{z}}\\ &\le \widetilde{y}\frac{\left|y^2-z^2\right|}{z} + z\frac{\left|\widetilde{y}^2-\widetilde{z}^2\right|}{\widetilde{z}}.\end{aligned}$$

Taking $y = \left\|o^{(\ell)}(x)\right\|_2$, $\widetilde{y} = \left\|o^{(\ell)}(x')\right\|_2$, $z = \sqrt{\mathbb{E}\left[\sigma^2\left(U^{(\ell)}(x)\right)\right]}$, and $z' = \sqrt{\mathbb{E}\left[\sigma^2\left(U^{(\ell)}(x')\right)\right]}$, we have $\widetilde{y} \le c_0^\ell$ by Lemma 30 under event $\cap_{k=1}^\ell \mathcal{E}_0^{(k)}$, $z, \widetilde{z} = 2^{-\ell/2}$ by (38), and $|y^2 - z^2|$ and $|\widetilde{y}^2 - \widetilde{z}^2|$ are upper bounded by $\ell C^\ell/m^{1/3}$ by Lemma 6. Therefore,

$$\begin{aligned}&\left|\left\|o^{(\ell)}(x)\right\|_2\left\|o^{(\ell)}(x')\right\|_2 - \sqrt{\mathbb{E}\left[\sigma^2\left(U^{(\ell)}(x)\right)\right]}\sqrt{\mathbb{E}\left[\sigma^2\left(U^{(\ell)}(x')\right)\right]}\right|\\ &= O\left(c_0^\ell\frac{\ell C^\ell/m^{1/3}}{2^{-\ell/2}}\right) + O\left(\frac{\ell C^\ell}{m^{1/3}}\right) = O\left(\frac{\ell C^\ell}{m^{1/3}}\right).\end{aligned}$$

Plugging the above bound into (100), we have $\text{(II)} = O\left(\ell\frac{C^\ell}{m^{1/3}}\right)$.

Combining the bound of (I) and (II), we have for any x and x',

$$|\widehat{\rho}^{(\ell)}(x,x') - \rho^{(\ell)}(x,x')| = O\left(\frac{\ell C^\ell}{m^{1/3}}\right). \tag{101}$$

Next we prove $\arccos \widehat{\rho}^{(\ell)}(x,x')$ is close to $\arccos \rho^{(\ell)}(x,x')$ for any x and x' on $\mathbb{S}^{d-1}$. For notation simplicity, in the remaining part of the proof, we denote ρ as $\rho^{(\ell)}$ and $\widehat{\rho}$ as $\widehat{\rho}^{(\ell)}$. Here, we claim for any y and z, $|\arccos y - \arccos z| \le 3\sqrt{|y-z|}$. Given the claim, taking $y=\widehat{\rho}$ and $z=\rho$, we complete the proof since $|\arccos\rho - \arccos\widehat{\rho}| \le 3\sqrt{|\rho-\widehat{\rho}|} = O\left(\sqrt{\ell}C^{\ell}m^{-1/6}\right)$.

Now we prove the claim. WLOG, assume $\rho \le \widehat{\rho} \le 1$, so $\arccos\rho \ge \arccos\widehat{\rho}$. By Abramowitz et al. (1988, 4.4.33), we have $\arccos\rho - \arccos\widehat{\rho} = \arccos\left(\rho\widehat{\rho} + \sqrt{1-\rho^2}\sqrt{1-\widehat{\rho}^2}\right) \triangleq \arccos\xi$. Define $\delta \triangleq \widehat{\rho}-\rho$. Note that $\xi = \rho\widehat{\rho} + \sqrt{1-\rho^2}\sqrt{1-\widehat{\rho}^2} \ge \widehat{\rho}^2 - \delta\widehat{\rho} + 1 - \widehat{\rho}^2 \ge 1-\delta$, where the second inequality holds by $1-\rho^2 \ge 1-\widehat{\rho}^2$ and the last inequality holds by $\widehat{\rho}\le 1$.

Since arccos function is monotonic decreasing, it remains to show $\arccos(1-\delta) \le 3\sqrt{\delta}$. Denote $h(x) = 3\sqrt{x} - \arccos(1-x), x\in(0,1]$. Since $\frac{dh}{dx} = \frac{1}{\sqrt{x}}\left(3 - \frac{2}{\sqrt{2-x}}\right) > 0$ for any $x\in(0,1]$ and $h(0)=0$, we have $\arccos(1-x) \le 3\sqrt{x}$ for any $x\in[0,1]$. ∎

Proof of Lemma 8: Denote for any k and all $\ell \le k$,

$$\mathcal{E}_{d,k}^{(\ell)} = \left\{ \sup_{x,x'} \left| \mathrm{Tr}\left(\mathbf{G}_k^{(\ell)}(x,x')\right) - q_k^{(\ell)}(x,x') \right| = O\left(\frac{\sqrt{k}C^k}{m^{1/6}} + \sqrt{\frac{d\log^{k-1} m}{m}} \right) \right\},$$

and $\mathcal{E}_{d,k} = \cap_{\ell=1}^k \mathcal{E}_{d,k}^{(\ell)}$.

Note that under $\mathcal{E}_{d,L}$, (32) holds directly. Thus, it suffices to prove

$$\mathbb{P}\left[\mathcal{E}_{d,L}\cap\mathcal{E}_0\cap\mathcal{E}_1\right] = 1 - \exp\left(O(dL\log m) - \Omega(m^{1/3})\right).$$

where $\mathcal{E}_0 = \cap_{k=1}^L \mathcal{E}_0^{(k)}$ and $\mathcal{E}_1 = \cap_{k=1}^L \mathcal{E}_1^{(k)}$ with $\mathcal{E}_0^{(k)}$ defined in (69) and $\mathcal{E}_1^{(k)}$ defined in (70). For notation simplicity, denote $\mathcal{V}^{(k)} = \cap_{r=1}^k \left(\mathcal{E}_0^{(r)} \cap \mathcal{E}_1^{(r)}\right)$. By the second inequality of Lemma 29, we have

$$\mathbb{P}\left[\mathcal{E}_{d,L}\cap\mathcal{E}_0\cap\mathcal{E}_1\right] \ge 1 - \sum_{k=1}^L \mathbb{P}\left[\mathcal{E}_{d,k}^c | \mathcal{E}_{d,k-1}\cap\mathcal{V}^{(k-1)}\right] - \sum_{k=1}^L \mathbb{P}\left[\left(\mathcal{V}^{(k)}\right)^c\right]. \tag{102}$$

By (74), we have

$$\mathbb{P}\left[\mathcal{V}^{(k)}\right] \ge 1 - k\exp\left(O\left(d\log m\right) - \Omega(m^{1/3})\right). \tag{103}$$

Next we bound $\mathbb{P}\left[\mathcal{E}_{d,k}^c | \mathcal{E}_{d,k-1}\cap\mathcal{V}^{(k-1)}\right]$. By definition of $\mathcal{E}_{d,k}$,

$$\mathbb{P}\left[\mathcal{E}_{d,k}^c | \mathcal{E}_{d,k-1}\cap\mathcal{V}^{(k-1)}\right] \le \sum_{\ell=1}^k \mathbb{P}\left[\left(\mathcal{E}_{d,k}^{(\ell)}\right)^c | \mathcal{E}_{d,k-1}\cap\mathcal{V}^{(k-1)}\right]. \tag{104}$$

In the remaining proof, we condition on $\left\{\mathbf{W}^{(r)}\right\}_{r=1}^{k-1}$ such that $\mathcal{E}_{d,k-1}\cap\mathcal{V}^{(k-1)}$ holds.

Case 1 $\ell = k$**:** By definition,

$$\mathrm{Tr}\left(\mathbf{G}_k^{(k)}(x,x')\right) = \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}_{\left\{\langle w_i^{(k)}, o^{(k-1)}(x)\rangle \geq 0\right\}} \mathbf{1}_{\left\{\langle w_i^{(k)}, o^{(k-1)}(x')\rangle \geq 0\right\}}, \tag{105}$$

and

$$q_k^{(k)}(x,x') = \frac{\pi - \arccos \rho^{(k-1)}(x,x')}{2\pi},$$

where $\rho^{(k-1)}$ is defined in (93).

By the triangle inequality, we have

$$\begin{aligned}
&\left|\mathrm{Tr}\left(\mathbf{G}_k^{(k)}(x,x')\right) - q_k^{(k)}(x,x')\right| \\
&= \left|\frac{1}{m}\sum_{i=1}^{m} \mathbf{1}_{\left\{\langle w_i^{(k)}, o^{(k-1)}(x)\rangle \geq 0\right\}} \mathbf{1}_{\left\{\langle w_i^{(k)}, o^{(k-1)}(x')\rangle \geq 0\right\}} - \frac{\pi - \arccos \rho^{(k-1)}(x,x')}{2\pi}\right| \\
&\leq \left|\frac{1}{m}\sum_{i=1}^{m} \mathbf{1}_{\left\{\langle w_i^{(k)}, o^{(k-1)}(x)\rangle \geq 0\right\}} \mathbf{1}_{\left\{\langle w_i^{(k)}, o^{(k-1)}(x')\rangle \geq 0\right\}} - \frac{\pi - \arccos \widehat{\rho}^{(k-1)}(x,x')}{2\pi}\right| \\
&+ \left|\frac{1}{2\pi}\left(\arccos \widehat{\rho}^{(k-1)}(x,x') - \arccos \rho^{(k-1)}(x,x')\right)\right|,
\end{aligned} \tag{106}$$

where $\widehat{\rho}^{(k-1)}$ is defined in (98).

Under $\mathcal{V}^{(k-1)}$, by Corollary 35, we have

$$\sup_{x,x'}\left|\frac{1}{2\pi}\left(\arccos \widehat{\rho}^{(k-1)}(x,x') - \arccos \rho^{(k-1)}(x,x')\right)\right| = O\left(\frac{\sqrt{k-1}C^{k-1}}{m^{1/6}}\right). \tag{107}$$

Now we bound the first term on the RHS of (106). Note that $\left\{w_i^{(k)}\right\}_{i=1}^m$ is independent of $\left\{\mathbf{W}^{(\ell)}\right\}_{\ell=1}^{k-1}$ and

$$\mathbb{E}_{w_i^{(k)}}\left[\mathbf{1}_{\left\{\langle w_i^{(k)}, o^{(k-1)}(x)\rangle \geq 0\right\}} \mathbf{1}_{\left\{\langle w_i^{(k)}, o^{(k-1)}(x')\rangle \geq 0\right\}}\right] = \frac{\pi - \arccos \widehat{\rho}^{(k-1)}(x,x')}{2\pi}.$$

Thus, by Lemma 34, with probability at least $1 - \exp(-2m^{1/3})$,

$$\begin{aligned}
&\sup_{x,x'}\left|\frac{1}{m}\sum_{i=1}^{m} \mathbf{1}_{\left\{\langle w_i^{(k)}, o^{(k-1)}(x)\rangle \geq 0\right\}} \mathbf{1}_{\left\{\langle w_i^{(k)}, o^{(k-1)}(x')\rangle \geq 0\right\}} - \frac{\pi - \arccos \widehat{\rho}^{(k-1)}(x,x')}{2\pi}\right| \\
&= O\left(\sqrt{\frac{d\left(1 + (k-1)\log m\right)}{m}} + \frac{1}{m^{1/3}}\right),
\end{aligned}$$

Combining the above displayed equation with (107), we have with probability at least $1 - \exp\left(-2m^{1/3}\right)$,

$$\begin{aligned}
\sup_{x,x'}\left|\mathrm{Tr}\left(\mathbf{G}_k^{(k)}(x,x')\right) - q_k^{(k)}(x,x')\right| &= O\left(\frac{\sqrt{k-1}C^{k-1}}{m^{1/6}} + \sqrt{\frac{d\left(1 + (k-1)\log m\right)}{m}} + \frac{1}{m^{1/3}}\right) \\
&= O\left(\frac{\sqrt{k-1}C^{k-1}}{m^{1/6}} + \sqrt{\frac{d\left(1 + (k-1)\log m\right)}{m}}\right).
\end{aligned}$$

Thus,

$$\mathbb{P}\left[\mathcal{E}_{d,k}^{(k)}|\mathcal{E}_{d,k-1}\cap\mathcal{V}^{(k-1)}\right]\geq 1-\exp(-2m^{1/3}). \tag{108}$$

Case 2 $\ell<k$: By the definition of $\mathbf{G}_k^{(\ell)}$, we have

$$\begin{aligned}\mathrm{Tr}\left(\mathbf{G}_k^{(\ell)}(x,x')\right)&=\mathrm{Tr}\left(\mathbf{D}^{(k)}(x)\mathbf{W}^{(k)}\mathbf{G}_{k-1}^{(\ell)}(x,x')\left[\mathbf{W}^{(k)}\right]^\top\mathbf{D}^{(k)}(x')\right)\\&=\frac{1}{m}\sum_{i=1}^m\mathbf{1}_{\left\{\langle w_i^{(k)},o^{(k-1)}(x)\rangle\geq 0\right\}}\mathbf{1}_{\left\{\langle w_i^{(k)},o^{(k-1)}(x')\rangle\geq 0\right\}}\left[w_i^{(k)}\right]^\top\mathbf{G}_{k-1}^{(\ell)}(x,x')w_i^{(k)}.\end{aligned}$$

Thus, by the triangle inequality, we have

$$\begin{aligned}&\sup_{x,x'}\left|\mathrm{Tr}\left(\mathbf{G}_k^{(\ell)}(x,x')\right)-q_k^{(\ell)}(x,x')\right|\\&\leq\sup_{x,x'}\left|\frac{1}{m}\sum_{i=1}^m\mathbf{1}_{\left\{\langle w_i^{(k)},o^{(k-1)}(x)\rangle\geq 0\right\}}\mathbf{1}_{\left\{\langle w_i^{(k)},o^{(k-1)}(x')\rangle\geq 0\right\}}\left[w_i^{(k)}\right]^\top\mathbf{G}_{k-1}^{(\ell)}(x,x')w_i^{(k)}\right.\\&\left.-\frac{1}{m}\sum_{i=1}^m\mathbf{1}_{\left\{\langle w_i^{(k)},o^{(k-1)}(x)\rangle\geq 0\right\}}\mathbf{1}_{\left\{\langle w_i^{(k)},o^{(k-1)}(x')\rangle\geq 0\right\}}q_{k-1}^{(\ell)}(x,x')\right|\\&+\sup_{x,x'}\left|\frac{1}{m}\sum_{i=1}^m\mathbf{1}_{\left\{\langle w_i^{(k)},o^{(k-1)}(x)\rangle\geq 0\right\}}\mathbf{1}_{\left\{\langle w_i^{(k)},o^{(k-1)}(x')\rangle\geq 0\right\}}q_{k-1}^{(\ell)}(x,x')-q_k^{(\ell)}(x,x')\right|\\&\leq\underbrace{\sup_{x,x',i}\left|\left[w_i^{(k)}\right]^\top\mathbf{G}_{k-1}^{(\ell)}(x,x')w_i^{(k)}-q_{k-1}^{(\ell)}(x,x')\right|}_{\text{(I)}}\\&\underbrace{+\sup_{x,x'}\left|\frac{1}{m}\sum_{i=1}^m\mathbf{1}_{\left\{\langle w_i^{(k)},o^{(k-1)}(x)\rangle\geq 0\right\}}\mathbf{1}_{\left\{\langle w_i^{(k)},o^{(k-1)}(x')\rangle\geq 0\right\}}q_{k-1}^{(\ell)}(x,x')-q_k^{(\ell)}(x,x')\right|}_{\text{(II)}}.\end{aligned}$$

By the triangle inequality, we have

$$\begin{aligned}\text{(I)}&\leq\sup_{x,x',1\leq i\leq m}\left|\left[w_i^{(k)}\right]^\top\mathbf{G}_{k-1}^{(\ell)}(x,x')w_i^{(k)}-\mathrm{Tr}\left(\mathbf{G}_{k-1}^{(\ell)}(x,x')\right)\right|+\sup_{x,x'}\left|\mathrm{Tr}\left(\mathbf{G}_{k-1}^{(\ell)}(x,x')\right)-q_{k-1}^{(\ell)}(x,x')\right|\\&\leq\sup_{x,x',1\leq i\leq m}\left|\left[w_i^{(k)}\right]^\top\mathbf{G}_{k-1}^{(\ell)}(x,x')w_i^{(L)}-\mathrm{Tr}\left(\mathbf{G}_{k-1}^{(\ell)}(x,x')\right)\right|\\&\quad+O\left(\frac{\sqrt{k-1}C^{k-1}}{m^{1/6}}+\sqrt{\frac{d\left(1+(k-2)\log m\right)}{m}}\right),\end{aligned} \tag{109}$$

where the last equality holds under $\mathcal{E}_{d,k-1}^{(\ell)}$.

We next bound the first term in the RHS of (109). Since $\left\{w_i^{(k)}\right\}_{i=1}^m$ are i.i.d. $\mathcal{N}(0, \mathbf{I}_m)$ and are independent of $\left\{\mathbf{W}^{(r)}\right\}_{r=1}^{k-1}$, by Lemma 33, for any $i \in [m]$,

$$\begin{aligned}
&\mathbb{P}_{w_i^{(k)}}\left[\sup_{x,x'}\left|\left[w_i^{(k)}\right]^\top \mathbf{G}_{k-1}^{(\ell)}(x,x')w_i^{(k)} - \mathrm{Tr}\left(\mathbf{G}_{k-1}^{(\ell)}(x,x')\right)\right| > \frac{c_0^{2k-2-2\ell}}{m^{1/3}}\middle|\mathcal{E}_{d,k-1}\cap\mathcal{V}^{(k-1)}\right] \\
&= \exp\left(O\left(d(k-1)\log m\right) - \Omega(m^{1/3})\right).
\end{aligned}$$

Further taking union bounds over i, we have

$$\begin{aligned}
&\mathbb{P}\left[\sup_{x,x',i\in[m]}\left|\left[w_i^{(k)}\right]^\top \mathbf{G}_{k-1}^{(\ell)}(x,x')w_i^{(k)} - \mathrm{Tr}(\mathbf{G}_{k-1}^{(\ell)}(x,x'))\right| > \frac{c_0^{2k-2-2\ell}}{m^{1/3}}\middle|\mathcal{E}_{d,k-1}\cap\mathcal{V}^{(k-1)}\right] \\
&\le m\exp\left[O\left(d(k-1)\log m\right) - \Omega(m^{1/3})\right] = \exp\left[O\left(d(k-1)\log m\right) - \Omega(m^{1/3})\right]. \qquad (110)
\end{aligned}$$

Plugging (110) into (109), we get

$$\begin{aligned}
&\mathbb{P}\left[(\mathrm{I}) = O\left(\frac{\sqrt{k-1}C^{k-1}}{m^{1/6}} + \sqrt{\frac{d\left(1+(k-2)\log m\right)}{m}}\right) + \frac{c_0^{2k-2-2\ell}}{m^{1/3}}\middle|\mathcal{E}_{d,k-1}\cap\mathcal{V}^{(k-1)}\right] \\
&= \mathbb{P}\left[(\mathrm{I}) = O\left(\frac{\sqrt{k-1}C^{k-1}}{m^{1/6}} + \sqrt{\frac{d\left(1+(k-2)\log m\right)}{m}}\right)\middle|\mathcal{E}_{d,k-1}\cap\mathcal{V}^{(k-1)}\right] \\
&\ge 1 - \exp\left(O\left(d(k-1)\log m\right) - \Omega(m^{1/3})\right).
\end{aligned}$$

To bound (II), we have with probability at least $1-\exp(-2m^{1/3})$,

$$\begin{aligned}
(\mathrm{II}) &\overset{(a)}{=} \sup_{x,x'}\left|q_{k-1}^{(\ell)}(x,x')\left(\mathrm{Tr}\left(\mathbf{G}_k^{(k)}\right) - q_k^{(k)}(x,x')\right)\right| \\
&\overset{(b)}{\le} \sup_{x,x'}\left|\mathrm{Tr}\left(\mathbf{G}_k^{(k)}(x,x')\right) - q_k^{(k)}(x,x')\right| \\
&= O\left(\frac{\sqrt{k}C^k}{m^{1/6}} + \sqrt{\frac{d\left(1+(k-1)\log m\right)}{m}}\right).
\end{aligned}$$

where (a) holds by (19), that is $q_k^{(\ell)}(x,x') = q_{k-1}^{(\ell)}(x,x')q_k^{(k)}(x,x')$, and (105); (b) holds as $\sup_{x,x'}\left|q_{k-1}^{(\ell)}(x,x')\right| \le 1$; and the last equality holds from (108).

Combining the above bounds on (I) and (II), we have for any $\ell < k$,

$$\mathbb{P}\left[\left[\mathcal{E}_{d,k}^{(\ell)}\right]^c|\mathcal{E}_{d,k-1}\cap\mathcal{V}_0^{(k-1)}\cap\mathcal{V}_1^{(k-1)}\right] \le \exp\left[O\left(d(k-1)\log m\right) - \Omega(m^{1/3})\right] + \exp\left(-2m^{1/3}\right).$$

Combining the last displayed equation with (104) and (108) yields that

$$\mathbb{P}\left[\mathcal{E}_{d,k}^c|\mathcal{E}_{d,k-1}\cap\mathcal{V}_0^{(k-1)}\cap\mathcal{V}_1^{(k-1)}\right] \le (k-1)\exp\left[O\left(d(k-1)\log m\right) - \Omega(m^{1/3})\right] + k\exp\left(-2m^{1/3}\right). \qquad (111)$$

Plugging (111) and (103) into (102), we get

$$\begin{aligned}
&\mathbb{P}\left[\mathcal{E}_{d,k} \cap \mathcal{E}_0 \cap \mathcal{E}_1\right] \\
&\geq 1 - \sum_{k=1}^{L} \left[(k-1)\exp\left[O\left(d(k-1)\log m\right) - \Omega(m^{1/3})\right] + k\exp\left(-2m^{1/3}\right)\right] \\
&\quad - \sum_{k=1}^{L} \exp\left(O\left(d\log m\right) - \Omega(m^{1/3})\right) \\
&= 1 - \exp\left[O\left(dL\log m\right) - \Omega(m^{1/3})\right].
\end{aligned}$$

Appendix C. Proofs in Section 6

Recall $\mathcal{E}_0 = \cap_{\ell=1}^{L} \mathcal{E}_0^{(\ell)}$ where $\mathcal{E}_0^{(\ell)}$ is defined in (69). By (92), we have

$$\mathbb{P}\left[\mathcal{E}_0\right] \geq 1 - L\exp\left(-\Omega(m)\right). \tag{112}$$

C.1 Proof of Proposition 9

Throughout the proof, we assume $\mathcal{E}_0$, all three conclusions of Lemma 10, conclusion of Lemma 11, and conclusions of Lemma 12 hold. These altogether can be guaranteed with probability at least $1 - \exp\left(\Omega(C^{-L}m^{1/36})\right)$ by union bounds following (112) and Lemma 10–12.

Recall from (39) that

$$\frac{\partial f(x;\mathbf{W}(t))}{\partial \mathbf{W}^{(\ell)}} = \frac{1}{\sqrt{m}}\mathbf{D}_t^{(\ell)}(x) z_t^{(\ell)}(x)\left[o_t^{(\ell-1)}(x)\right]^\top,$$

where

$$\left[z_t^{(\ell)}(x)\right]^\top = a^\top \frac{1}{\sqrt{m}}\mathbf{D}_t^{(L)}(x)\mathbf{W}^{(L)}(t)\cdots\frac{1}{\sqrt{m}}\mathbf{D}_t^{(\ell+1)}(x)\mathbf{W}^{(\ell+1)}(t). \tag{113}$$

Therefore, by the triangle inequality, we have

$$\begin{aligned}
&\left\|\frac{\partial f(x;\mathbf{W}(t))}{\partial \mathbf{W}^{(\ell)}} - \frac{\partial f(x;\mathbf{W}(0))}{\partial \mathbf{W}^{(\ell)}}\right\|_2 \\
&= \frac{1}{\sqrt{m}}\left\|\mathbf{D}_t^{(\ell)}(x) z_t^{(\ell)}(x)\left[o_t^{(\ell-1)}(x)\right]^\top - \mathbf{D}_0^{(\ell)}(x) z_0^{(\ell)}(x)\left[o_0^{(\ell-1)}(x)\right]^\top\right\|_2 \\
&\leq \frac{1}{\sqrt{m}}\left\|\mathbf{D}_t^{(\ell)}(x)\left(z_t^{(\ell)}(x) - z_0^{(\ell)}(x)\right)\left[o_t^{(\ell-1)}(x)\right]^\top\right\|_2 + \frac{1}{\sqrt{m}}\left\|\left(\mathbf{D}_t^{(\ell)}(x) - \mathbf{D}_0^{(\ell)}(x)\right) z_0^{(\ell)}(x)\left[o_t^{(\ell-1)}(x)\right]^\top\right\|_2 \\
&\quad + \frac{1}{\sqrt{m}}\left\|\mathbf{D}_0^{(\ell)}(x) z_0^{(\ell)}(x)\left(o_t^{(\ell-1)}(x) - o_0^{(\ell-1)}(x)\right)^\top\right\|_2.
\end{aligned} \tag{114}$$

Now we bound the first term on the right hand side of (114). Note that

$$\sup_x \left\|o_t^{(\ell-1)}(x)\right\|_2 \leq \sup_x \left\|o_t^{(\ell-1)}(x) - o_0^{(\ell-1)}(x)\right\|_2 + \sup_x \left\|o_0^{(\ell-1)}(x)\right\|_2 \leq \frac{C^{\ell-1}}{m^{1/6}} + c_0^{\ell-1} \leq C^{\ell-1}, \tag{115}$$

where the second inequality holds by (45) in Lemma 10, i.e., $\sup_x \left\|o_t^{(\ell)}(x) - o_0^{(\ell)}(x)\right\|_2 \leq \frac{C^\ell}{m^{1/6}}$, and Lemma 30 under $\mathcal{E}_0$. Since $\left\|\mathbf{D}_t^{(\ell)}(x)\right\|_2 \leq 1$ for all x and t, we have

$$\frac{1}{\sqrt{m}}\left\|\mathbf{D}_t^{(\ell)}(x)\left(z_t^{(\ell)}(x) - z_0^{(\ell)}(x)\right)\left[o_t^{(\ell-1)}(x)\right]^\top\right\|_2 \leq \frac{1}{\sqrt{m}}\left\|z_t^{(\ell)}(x) - z_0^{(\ell)}(x)\right\|_2\left\|o_t^{(\ell-1)}(x)\right\|_2 = O\left(C^{2L-1}m^{-1/36}\right) \tag{116}$$

where the last inequality holds by (49) from Lemma 12, i.e., $\sup_x \left\|z_t^{(\ell)}(x) - z_0^{(\ell)}(x)\right\|_2 = O\left(C^{2L-\ell}m^{17/36}\right)$.

To bound the second term of (114), note that by the definition of $\mathbf{D}_t^{(\ell)}$, we have

$$\begin{aligned}\left\|\left(\mathbf{D}_t^{(\ell)}(x) - \mathbf{D}_0^{(\ell)}(x)\right)z_0^{(\ell)}(x)\right\|_2^2 &= \sum_{i=1}^m \left(\mathbf{1}_{\left\{\langle w_i^{(\ell)}(t), o_t^{(\ell-1)}(x)\rangle \geq 0\right\}} - \mathbf{1}_{\left\{\langle w_i^{(\ell)}(0), o_0^{(\ell-1)}(x)\rangle \geq 0\right\}}\right)^2\left[z_0^{(\ell)}(x)\right]_i^2 \\ &\leq \left\|z_0^{(\ell)}(x)\right\|_\infty^2 \sum_{i=1}^m \left|\mathbf{1}_{\left\{\langle w_i^{(\ell)}(t), o_t^{(\ell-1)}(x)\rangle \geq 0\right\}} - \mathbf{1}_{\left\{\langle w_i^{(\ell)}(0), o_0^{(\ell-1)}(x)\rangle \geq 0\right\}}\right| \\ &\leq 4m^{1/18}S_t^{(\ell)}(x) = O\left(C^\ell m^{17/18}\right)\end{aligned}$$

where the last inequality holds by $\sup_x \left\|z_0^{(\ell)}(x)\right\|_\infty \leq m^{1/36}$ from (48) in Lemma 12, and the definition of $S_t(x)$ and the last equality holds by (47) from Lemma 11, i.e., $\sup_x S_t^{(\ell)}(x) \leq C_2^\ell m^{8/9}$. Thus, by (115), we have

$$\begin{aligned}&\frac{1}{\sqrt{m}}\left\|\left(\mathbf{D}_t^{(\ell)}(x) - \mathbf{D}_0^{(\ell)}(x)\right)z_0^{(\ell)}(x)\left[o_t^{(\ell-1)}(x)\right]^\top\right\|_2 \\ &\leq \frac{1}{\sqrt{m}}\left\|\left(\mathbf{D}_t^{(\ell)}(x) - \mathbf{D}_0^{(\ell)}(x)\right)z_0^{(\ell)}(x)\right\|_2\left\|o_t^{(\ell-1)}(x)\right\|_2 \\ &= \frac{1}{\sqrt{m}}O\left(C^{2\ell-1}m^{17/36}\right) = O\left(C^{2\ell-1}m^{-1/36}\right).\end{aligned} \tag{117}$$

Now we bound the last term on the right hand side of (114). By the definition of $z_t^{(\ell)}(x)$ in (113), since $\left\|\mathbf{D}_0^{(\ell)}(x)\right\|_2 \leq 1$ for all x, under $\mathcal{E}_0$, we have

$$\left\|z_0^{(\ell)}(x)\right\|_2 \leq \prod_{k=\ell+1}^L \left\|\frac{1}{\sqrt{m}}\mathbf{D}_0^{(k)}(x)\mathbf{W}^{(k)}(0)\right\|_2 \|a\|_2 \leq c_0^{L-\ell}\sqrt{m}. \tag{118}$$

Thus, we have

$$\begin{aligned}&\frac{1}{\sqrt{m}}\left\|\mathbf{D}_0^{(\ell)}(x)z_0^{(\ell)}(x)\left(o_t^{(\ell-1)}(x) - o_0^{(\ell-1)}(x)\right)^\top\right\|_2 \\ &\leq \frac{1}{\sqrt{m}}\left\|z_0^{(\ell)}(x)\right\|_2\left\|o_t^{(\ell-1)}(x) - o_0^{(\ell-1)}(x)\right\|_2 \leq \frac{C^L}{m^{1/6}},\end{aligned} \tag{119}$$

where the last inequality holds by (118) and (45) of Lemma 10.

Plugging (116), (117) and (119) back into (114), we get

$$\left\|\frac{\partial f(x;\mathbf{W}(t))}{\partial \mathbf{W}^{(\ell)}}-\frac{\partial f(x;\mathbf{W}(0))}{\partial \mathbf{W}^{(\ell)}}\right\|_2 = O\left(C^{2L-1}m^{-1/36}+C^{2\ell-1}m^{-1/36}+\frac{C^L}{m^{1/6}}\right)=O\left(C^{2L}m^{-1/36}\right). \tag{120}$$

Next, we prove $\|H_t-H_0\|_\infty = O\left(C^{2L}m^{-1/36}\right)$. By (77), we have

$$\begin{aligned}
&\left|H_t^{(\ell)}(x,x')-H^{(\ell)}(x,x')\right| \\
&=\left|\left\langle\frac{\partial f(x;\mathbf{W}(t))}{\partial \mathbf{W}^{(\ell)}},\frac{\partial f(x';\mathbf{W}(t))}{\partial \mathbf{W}^{(\ell)}}\right\rangle-\left\langle\frac{\partial f(x;\mathbf{W}(0))}{\partial \mathbf{W}^{(\ell)}},\frac{\partial f(x';\mathbf{W}(0))}{\partial \mathbf{W}^{(\ell)}}\right\rangle\right| \\
&\le\left\|\frac{\partial f(x;\mathbf{W}(t))}{\partial \mathbf{W}^{(\ell)}}-\frac{\partial f(x;\mathbf{W}(0))}{\partial \mathbf{W}^{(\ell)}}\right\|_2\left\|\frac{\partial f(x';\mathbf{W}(t))}{\partial \mathbf{W}^{(\ell)}}\right\|_2 \\
&\quad+\left\|\frac{\partial f(x';\mathbf{W}(t))}{\partial \mathbf{W}^{(\ell)}}-\frac{\partial f(x';\mathbf{W}(0))}{\partial \mathbf{W}^{(\ell)}}\right\|_2\left\|\frac{\partial f(x;\mathbf{W}(0))}{\partial \mathbf{W}^{(\ell)}}\right\|_2 \\
&=O\left(C^{2L}m^{-1/36}\right)\left(\left\|\frac{\partial f(x';\mathbf{W}(t))}{\partial \mathbf{W}^{(\ell)}}\right\|_2+\left\|\frac{\partial f(x;\mathbf{W}(0))}{\partial \mathbf{W}^{(\ell)}}\right\|_2\right),
\end{aligned} \tag{121}$$

where the last equality holds by (120).

From (118) and Lemma 30, we get under $\mathcal{E}_0$,

$$\left\|\frac{\partial f(x;\mathbf{W}(0))}{\partial \mathbf{W}^{(\ell)}}\right\|_2\le\frac{1}{\sqrt{m}}\left\|z_0^{(\ell)}(x)\right\|_2\left\|o_0^{(\ell-1)}(x)\right\|_2=O\left(c_0^L\right). \tag{122}$$

By the triangle inequality, we further have

$$\left\|\frac{\partial f(x;\mathbf{W}(t))}{\partial \mathbf{W}^{(\ell)}}\right\|_2\le\left\|\frac{\partial f(x;\mathbf{W}(0))}{\partial \mathbf{W}^{(\ell)}}\right\|_2+\left\|\frac{\partial f(x;\mathbf{W}(t))}{\partial \mathbf{W}^{(\ell)}}-\frac{\partial f(x;\mathbf{W}(0))}{\partial \mathbf{W}^{(\ell)}}\right\|_2=O(C^{2L})$$

where the last equality holds by plugging in (122) and (120).

Plugging the above bound and (122) into (121), we complete the proof.

C.2 Proof of Lemma 10

Step 1, showing $R_t\le m^{1/3}$: Recall that $R_0=m^{5/18}$ and $R_{t+1}=R_0+LC^{2L-2}\sum_{s=0}^t\eta_s(R_s+\gamma)$. Therefore $R_{t+1}+\gamma-(R_t+\gamma)=LC^{2L-2}\eta_t(R_t+\gamma)$, which is equivalent as $R_{t+1}+\gamma=\left(1+LC^{2L-2}\eta_t\right)(R_t+\gamma)$. Thus,

$$\begin{aligned}
R_t+\gamma&=\prod_{s=0}^{t-1}\left(1+LC^{2L-2}\eta_s\right)(R_0+\gamma) \\
&\le\exp\left(LC^{2L-2}\sum_{s=0}^{t-1}\eta_s\right)(R_0+\gamma)\le 2\exp\left(LC^{2L-2}\theta\log T\right)m^{5/18}
\end{aligned}$$

where the second inequality holds since $1+z\le e^z$ for any z and the last equality holds by plugging in $\eta_s\le\frac{\theta}{s+1}$ and the fact that $R_0+\gamma\le 2m^{5/18}$.

As a result, when $m = \exp\left(\Omega(LC^{2L-2}\theta \log T)\right)$, we have $R_t \leq R_t + \gamma \leq m^{1/3}$ for all $t \leq T$.

In the following Step 2, we show $\mathcal{E}_0 \cap \mathcal{E}_3$ occurs with high probability where

$$\mathcal{E}_3 \triangleq \left\{ \sup_x |\Delta_0(x)| \leq m^{5/18}, \forall t \leq T \right\},$$

with $\Delta_0(x) = f^*(x) - f(x; \mathbf{W}(t))$.

Then in Step 3, we use an inductive argument to show under $\mathcal{E}_0 \cap \mathcal{E}_3$, (44)–(46) in Lemma 10 hold for all $t \leq T$.

Step 2, bounding $\mathbb{P}[\mathcal{E}_0 \cap \mathcal{E}_3]$: Note that it suffices to show

$$\mathbb{P}[\mathcal{E}_3 | \mathcal{E}_0] \geq 1 - \exp\left(-\Omega(C^{-L} m^{1/9})\right). \tag{123}$$

With (123), by (112), we then have

$$\begin{aligned}
\mathbb{P}[\mathcal{E}_0 \cap \mathcal{E}_3] &\geq \left(1 - \exp\left(-\Omega(C^{-L} m^{1/9})\right)\right)(1 - L\exp(-\Omega(m))) \\
&= 1 - \exp\left(-\Omega(C_1^{-L} m^{1/9})\right),
\end{aligned} \tag{124}$$

for some constant C and C_1.

Now we prove 123. By the triangle inequality, we have $\sup_x |\Delta_0(x)| \leq \sup_x |f^*(x)| + \sup_x |f(x; \mathbf{W}(0))|$. Since $\sup_x |f^*(x)| \leq m^{5/18}$ by assumption, we have

$$\mathbb{P}[\mathcal{E}_3 | \mathcal{E}_0] \geq \mathbb{P}\left[\sup_x f^2(x; \mathbf{W}(0)) \leq m^{5/9} \Big| \mathcal{E}_0 \right].$$

Thus, it remains to show

$$\mathbb{P}\left[\sup_x f^2(x; \mathbf{W}(0)) \leq m^{5/9} \Big| \mathcal{E}_0 \right] = 1 - \exp\left(-\Omega(C^{-L} m^{1/9})\right).$$

Throughout the remaining proof of Step 2, we condition on $\left\{\mathbf{W}^{(k)}(0)\right\}_{k=1}^L$ such that $\mathcal{E}_0$ holds. Following the definition of f in (4),

$$\begin{aligned}
f^2(x; \mathbf{W}(0)) &= \left[a^\top \left(\frac{1}{\sqrt{m}} \mathbf{D}_0^{(L)}(x) \mathbf{W}^{(L)}(0) \cdots \frac{1}{\sqrt{m}} \mathbf{D}_0^{(1)}(x) \mathbf{W}^{(1)}(0) \right) x \right]^2 \\
&\leq \left\| a^\top \left(\frac{1}{\sqrt{m}} \mathbf{D}_0^{(L)}(x) \mathbf{W}^{(L)(0)} \cdots \frac{1}{\sqrt{m}} \mathbf{D}_0^{(1)}(x) \mathbf{W}^{(1)}(0) \right) \right\|_2^2 \\
&= a^\top \mathbf{Q}(x) a,
\end{aligned}$$

where

$$\begin{aligned}
&\mathbf{Q}(x) \\
&\triangleq \left(\frac{1}{\sqrt{m}} \mathbf{D}_0^{(L)}(x) \mathbf{W}^{(L)}(0) \cdots \frac{1}{\sqrt{m}} \mathbf{D}_0^{(1)}(x) \mathbf{W}^{(1)}(0) \right) \left(\frac{1}{\sqrt{m}} \mathbf{D}_0^{(L)}(x) \mathbf{W}^{(L)}(0) \cdots \frac{1}{\sqrt{m}} \mathbf{D}_0^{(1)}(x) \mathbf{W}^{(1)}(0) \right)^\top.
\end{aligned}$$

Under $\mathcal{E}_0$, we have $\|\mathbf{Q}(x)\|_2 \le c_0^{2L}$ and hence, $\|\mathbf{Q}(x)\|_{\mathrm{F}} \le \sqrt{m}\,\|\mathbf{Q}(x)\|_2 \le c_0^{2L}\sqrt{m}$. Since a is independent with $\left\{\mathbf{W}^{(k)}(0)\right\}_{k=1}^{L}$, by Hanson-Wright inequality, for any fixed x,

$$\begin{aligned}&\mathbb{P}\left[\left|a^\top \mathbf{Q}(x) a\right| \ge m^{5/9} \middle| \left\{\mathbf{W}^{(k)}(0)\right\}_{k=1}^{L} \text{ s.t. } \mathcal{E}_0 \text{ holds}\right]\\ &\le 2\exp\left(-C\min\left(\frac{m^{10/9}}{c_0^{4L}m}, \frac{m^{5/9}}{c_0^{2L}}\right)\right) = \exp\left(-\Omega(C^{-L}m^{1/9})\right).\end{aligned}$$

Denote $\mathcal{Q}\left(\mathbf{W}^{(1)}(0), \cdots, \mathbf{W}^{(L)}(0)\right) = \left\{\mathbf{Q}(x,x') : x, x' \in \mathbb{S}^{d-1}\right\}$. Note that for any given $\mathbf{W}^{(1)}(0), \cdots, \mathbf{W}^{(L)}(0)$, by the definition of $\mathcal{D}_k$ in Section B.2, we have

$$\mathcal{Q} \subset \left\{\mathbf{V}\mathbf{V}^\top : \mathbf{V} = \mathbf{D}_L\mathbf{W}^{(L)}(0)\cdots\mathbf{D}_1\mathbf{W}^{(1)}(0), (\mathbf{D}_1, \cdots, \mathbf{D}_L) \in \mathcal{D}_L\right\}.$$

Thus by (83), we have $|\mathcal{Q}| \le |\mathcal{D}_L| \le m^{dL}$. Taking union bounds over $\mathcal{Q}$, for sufficiently large m, we have

$$\begin{aligned}&\mathbb{P}\left[\sup_x \left|a^\top \mathbf{Q}(x) a\right| \ge m^{5/9} \middle| \left\{\mathbf{W}^{(k)}(0)\right\}_{k=1}^{L} \text{ s.t. } \mathcal{E}_0 \text{ holds}\right]\\ &= \mathbb{P}\left[\sup_{\mathbf{Q}\in\mathcal{Q}} \left|a^\top \mathbf{Q} a\right| \ge m^{5/9} \middle| \left\{\mathbf{W}^{(k)}(0)\right\}_{k=1}^{L} \text{ s.t. } \mathcal{E}_0 \text{ holds}\right]\\ &\le m^{dL}\exp\left(-\Omega(C^{-L}m^{1/9})\right) = \exp\left(-\Omega(C^{-L}m^{1/9})\right).\end{aligned}$$

Averaging over $\left\{\mathbf{W}^{(k)}(0)\right\}_{k=1}^{L}$, we get

$$\mathbb{P}\left[\sup_x f^2(x; \mathbf{W}(0)) \le m^{5/9} \middle| \mathcal{E}_0\right] = 1 - \exp\left(-\Omega(C^{-L}m^{1/9})\right).$$

Step 3, inductive argument to show small deviations of $\mathbf{W}^{(\ell)}$, $o^{(\ell)}$ and bounded Δ_t: Throughout step 3, we assume $\mathcal{E}_0 \cap \mathcal{E}_3$ holds. Here, we use an inductive argument to show that under $\mathcal{E}_0 \cap \mathcal{E}_3$, (44)–(46) hold for all $t \le T$.

When $t = 0$, (44) and (45) hold by definition. By the definition of Δ_0, we know (46) holds under $\mathcal{E}_3$.

Now suppose (44)–(46) hold for some t. We first show (44) holds at $t+1$ by showing

$$\left\|\mathbf{W}^{(\ell)}(t+1) - \mathbf{W}^{(\ell)}(t)\right\|_2 \le C^{L-1}\eta_t\left(R_s + \gamma\right) \tag{125}$$

for all ℓ.

By (10), we have

$$\begin{aligned}&\left\|\mathbf{W}^{(\ell)}(t+1) - \mathbf{W}^{(\ell)}(t)\right\|_2\\ &= \eta_t\left|\Delta_t(X_t) + u_t\right|\left\|\frac{1}{\sqrt{m}}\mathbf{D}_t^{(\ell)}(X_t)\left(\prod_{k=\ell+1}^{L}\left[\frac{1}{\sqrt{m}}\mathbf{W}^{(k)}(t)\right]^\top \mathbf{D}_t^{(k)}(X_t)\right) a\left[o_t^{(\ell-1)}(X_t)\right]^\top\right\|_2\\ &\le \eta_t(R_t+\gamma)\left\|\frac{1}{\sqrt{m}}\mathbf{D}_t^{(\ell)}(X_t)\left(\prod_{k=\ell+1}^{L}\left[\frac{1}{\sqrt{m}}\mathbf{W}^{(k)}(t)\right]^\top \mathbf{D}_t^{(k)}(X_t)\right) a\left[o_t^{(\ell-1)}(X_t)\right]^\top\right\|_2, \quad (126)\end{aligned}$$

where the last inequality holds since $|\Delta_t(X_t) + u_t| \leq \sup_x |\Delta_t(x)| + \sup |u_t| \leq R_t + \gamma$.

Now we show

$$\left\| \frac{1}{\sqrt{m}} \mathbf{D}_t^{(\ell)}(X_t) \left(\prod_{k=\ell+1}^{L} \left[\frac{1}{\sqrt{m}} \mathbf{W}^{(k)}(t) \right]^\top \mathbf{D}_t^{(k)}(X_t) \right) a \left[o_t^{(\ell-1)}(X_t) \right]^\top \right\|_2 \leq C^{L-1}.$$

Plugging the above inequality into (126), we obtain (125).

By the triangle inequality, under $\mathcal{E}_0 \cap \mathcal{E}_3$, for any k, we have

$$\left\| \mathbf{W}^{(k)}(t) \right\|_2 \leq \left\| \mathbf{W}^{(k)}(t) - \mathbf{W}^{(k)}(0) \right\|_2 + \left\| \mathbf{W}^{(k)}(0) \right\|_2 = O(\sqrt{m}), \tag{127}$$

where the last equality holds since under $\mathcal{E}_0 \cap \mathcal{E}_3$, $\left\| \mathbf{W}^{(k)}(0) \right\|_2 = O(\sqrt{m})$ and

$$\left\| \mathbf{W}^{(k)}(t) - \mathbf{W}^{(k)}(0) \right\|_2 \leq C^{L-1} \sum_{s=0}^{t-1} \eta_s \left(R_s + \gamma \right) \leq R_t \leq m^{1/3}, \forall 0 \leq t \leq T. \tag{128}$$

Similarly, by the triangle inequality, we have

$$\sup_x \left\| o_t^{(\ell-1)}(x) \right\|_2 \leq \sup_x \left\| o_t^{(\ell-1)}(x) - o_0^{(\ell-1)}(x) \right\|_2 + \sup_x \left\| o_0^{(\ell-1)}(x) \right\|_2 \leq \frac{\ell c_0^\ell}{m^{1/6}} + c_0^{\ell-1} \leq C^{\ell-1}, \tag{129}$$

where the last inequality holds by Lemma 30 under $\mathcal{E}_0$.

As a result, under $\mathcal{E}_0 \cap \mathcal{E}_3$, we have

$$\begin{aligned}
&\left\| \frac{1}{\sqrt{m}} \mathbf{D}_t^{(\ell)}(X_t) \left(\prod_{k=\ell+1}^{L} \left[\frac{1}{\sqrt{m}} \mathbf{W}^{(k)}(t) \right]^\top \mathbf{D}_t^{(k)}(X_t) \right) a \left[o_t^{(\ell-1)}(X_t) \right]^\top \right\|_2 \\
&\leq \left\| \frac{1}{\sqrt{m}} \mathbf{D}_t^{(\ell)}(X_t) \right\|_2 \left(\prod_{k=\ell+1}^{L} \left\| \left[\frac{1}{\sqrt{m}} \mathbf{W}^{(k)}(t) \right]^\top \mathbf{D}_t^{(k)}(X_t) \right\|_2 \right) \|a\|_2 \left\| o_t^{(\ell-1)}(X_t) \right\|_2 \\
&\leq C^{L-1},
\end{aligned}$$

where the last inequality holds by (127), (129) and the fact that $\left\| \mathbf{D}_t^{(k)}(x) \right\|_2 \leq 1$ for any k and x.

Next, we show (45) holds at $t+1$. When $\ell = 0$, since $o_{t+1}^{(0)}(x) = x$ for all t, we get

$$\sup_x \left\| o_{t+1}^{(0)}(x) - o_0^{(0)}(x) \right\|_2 = 0. \tag{130}$$

Fix arbitrary ℓ. By the definition of $o^{(\ell)}$, under $\mathcal{E}_0 \cap \mathcal{E}_3$, for any $x \in \mathbb{S}^{d-1}$, we have

$$\begin{aligned}
&\left\|o_{t+1}^{(\ell+1)}(x) - o_0^{(\ell+1)}(x)\right\|_2 \\
&= \frac{1}{\sqrt{m}} \left\|\sigma(\mathbf{W}^{(\ell+1)}(t+1)o_{t+1}^{(\ell)}(x)) - \sigma(\mathbf{W}^{(\ell+1)}(0)o_0^{(\ell)}(x))\right\|_2 \\
&\le \frac{1}{\sqrt{m}} \left\|\left(\mathbf{W}^{(\ell+1)}(t+1) - \mathbf{W}^{(\ell+1)}(0)\right) o_{t+1}^{(\ell)}(x)\right\|_2 + \frac{1}{\sqrt{m}} \left\|\mathbf{W}^{(\ell+1)}(0)\left(o_{t+1}^{(\ell)}(x) - o_0^{(\ell)}(x)\right)\right\|_2 \\
&\overset{(a)}{\le} m^{-1/6} \left\|o_{t+1}^{(\ell)}(x)\right\|_2 + c_0 \left\|o_{t+1}^{(\ell)}(x) - o_0^{(\ell)}(x)\right\|_2 \\
&\le m^{-1/6} \left(\left\|o_{t+1}^{(\ell)}(x) - o_0^{(\ell)}(x)\right\|_2 + \left\|o_0^{(\ell)}(x)\right\|_2\right) + c_0 \left\|o_{t+1}^{(\ell)}(x) - o_0^{(\ell)}(x)\right\|_2 \\
&\overset{(b)}{\le} \left(c_0 + m^{-1/6}\right) \left\|o_{t+1}^{(\ell)}(x) - o_0^{(\ell)}(x)\right\|_2 + c_0^{\ell} m^{-1/6},
\end{aligned}$$

where the first inequality holds by the triangle inequality, (a) holds by (128) and the definition of $\mathcal{E}_0$ which gives $\left\|\mathbf{W}^{(\ell+1)}(0)\right\|_2 \le c_0\sqrt{m}$, and (b) holds by Lemma 30 under $\mathcal{E}_0$.

Recursively applying the above inequality and being aware of (130), we get for any $x \in \mathbb{S}^{d-1}$, $\left\|o_{t+1}^{(\ell)}(x) - o_0^{(\ell)}(x)\right\|_2 \le c_0^{\ell} m^{-1/6} \sum_{k=0}^{\ell-1} \left(c_0 + m^{-1/6}\right)^k = O\left(C^{\ell} m^{-1/6}\right)$.

Finally, we show (46) holds at $t+1$. For notation simplicity, define $\mathbf{E}^{(k)}(t) \triangleq \mathbf{W}^{(k)}(t) - \mathbf{W}^{(k)}(0)$. By the triangle inequality, we have for any $x \in \mathbb{S}^{d-1}$,

$$\begin{aligned}
|\Delta_{t+1}(x)| &= |f^*(x) - f(x; \mathbf{W}(t+1)| \\
&\le |f^*(x) - f(x; \mathbf{W}(0))| + |f(x; \mathbf{W}(0) - f(x; \mathbf{W}(t+1))| \\
&= |\Delta_0(x)| + \left|\frac{1}{\sqrt{m}} a^\top \left(\sigma(\mathbf{W}^{(L)}(t+1)o_{t+1}^{(L-1)}(x)) - \sigma(\mathbf{W}^{(L)}(0)o_0^{(L-1)}(x))\right)\right| \\
&\overset{(a)}{\le} R_0 + \left\|\mathbf{W}^{(L)}(t+1)o_{t+1}^{(L-1)}(x) - \mathbf{W}^{(L)}(0)o_0^{(L-1)}(x)\right\|_2 \left\|\frac{1}{\sqrt{m}} a\right\|_2 \\
&\overset{(b)}{\le} R_0 + \left\|\mathbf{E}^{(L)}(t+1)o_{t+1}^{(L-1)}(x)\right\|_2 + \left\|\mathbf{W}^{(L)}(0)\left(o_{t+1}^{(L-1)}(x) - o_0^{(L-1)}(x)\right)\right\|_2 \\
&\overset{(c)}{\le} R_0 + \sum_{\ell=0}^{L-1} \left\|\mathbf{E}^{(L-\ell)}(t+1)o_t^{(L-\ell-1)}(x)\right\|_2
\end{aligned}$$

where (a) holds by Cauchy-Schwartz inequality under $\mathcal{E}_3$ and the fact that ReLU is 1-Lipschitz, (b) holds by the triangle inequality and (c) holds by recursively decomposing $o_{t+1}^{(L-1)}(x) - o_0^{(L-1)}(x)$.

Plugging (129) and the assumption that $\left\|\mathbf{W}^{(\ell)}(t+1) - \mathbf{W}^{(\ell)}(0)\right\|_2 \le C^{L-1} \sum_{s=0}^{t} \eta_s \left(R_s + \gamma\right)$ for any ℓ in the above displayed equation, we have for any x,

$$\begin{aligned}
|\Delta_{t+1}(x)| &\le R_0 + \sum_{\ell=1}^{L} C^{L-1} \left\|\mathbf{E}^{(\ell)}(t+1)\right\|_2 \\
&\le R_0 + LC^{2L-2} \sum_{s=0}^{t} \eta_s (R_s + \gamma) \\
&= R_{t+1},
\end{aligned}$$

where the last equality holds by the definition of R_{t+1}.

C.3 Proof of Lemma 11

Denote $O_t^{(\ell)}(x) \triangleq \left\{ i : \mathbf{1}_{\left\{\langle w_i^{(\ell)}(t), o_t^{(\ell)}(x)\rangle \geq 0\right\}} - \mathbf{1}_{\left\{\langle w_i^{(\ell)}(0), o_0^{(\ell)}(x)\rangle \geq 0\right\}} \neq 0 \right\}$. Therefore, we have $S_t^{(\ell)}(x) = |O_t^{(\ell)}(x)|$.

Note that if any neuron at layer ℓ has a sign flip, then it has either a small output value at initialization or a larger deviation than the initial output value. As such, we define

$$B^{(\ell)}(x) \triangleq \left\{ i : |\langle w_i^{(\ell)}(0), o_0^{(\ell-1)}(x)\rangle \leq \ell^{-1/3} C^{-1/3} m^{-1/9} \right\}, \tag{131}$$

as the set of neurons with small output values at initialization. Then

$$\sup_x S_t^{(\ell)}(x) \leq \sup_x |B^{(\ell)}(x)| + \sup_x \left| O_t^{(\ell)}(x) \cap \left[B^{(\ell)}(x) \right]^c \right|. \tag{132}$$

It remains to bound both $\sup_x |B^{(\ell)}(x)|$ and $\sup_x \left| O_t^{(\ell)}(x) \cap \left[B^{(\ell)}(x)\right]^c \right|$.

Define $\mathcal{E}_4 \triangleq \left\{ \sup_x |B^{(\ell)}(x)| = O\left(C^\ell m^{8/9}\right), \forall 1 \leq \ell \leq L \right\}$. It can be shown that $\mathcal{E}_4$ occurs with high probability. The proof is deferred to the end.

Step 1, bounding $\sup_x S_t^{(\ell)}(x)$: Throughout Step 1, we assume $\mathcal{E}_0 \cap \mathcal{E}_3 \cap \mathcal{E}_4$ and all conclusions of Lemma 10 hold.

Fix arbitrary ℓ. We first bound the deviation of the output value:

$$\sup_x \left\| \mathbf{W}^{(\ell)}(t) o_t^{(\ell-1)}(x) - \mathbf{W}^{(\ell)}(0) o_0^{(\ell-1)}(x) \right\|_2.$$

From (115), under $\mathcal{E}_0$, we have $\sup_x \left\| o_t^{(\ell)}(x) \right\|_2 \leq C^\ell$. Thus, by the triangle inequality, we have

$$\begin{aligned}
&\sup_x \left\| \mathbf{W}^{(\ell)}(t) o_t^{(\ell-1)}(x) - \mathbf{W}^{(\ell)}(0) o_0^{(\ell-1)}(x) \right\|_2 \\
&\leq \sup_x \left\| \left(\mathbf{W}^{(\ell)}(t) - \mathbf{W}^{(\ell)}(0) \right) o_t^{(\ell-1)}(x) \right\|_2 + \sup_x \left\| \mathbf{W}^{(\ell)}(0) \left(o_t^{(\ell-1)}(x) - o_0^{(\ell-1)}(x) \right) \right\|_2 \\
&\leq C^\ell m^{1/3} + c_0 \sqrt{m} \frac{C_1^\ell}{m^{1/6}} \\
&\leq C_2^\ell m^{1/3},
\end{aligned}$$

for some constant C_1 and C_2 where the second inequality holds by (45) from Lemma 10 under $\mathcal{E}_0$ and (128) under $\mathcal{E}_0 \cap \mathcal{E}_3$.

For neuron i in $\left[B^{(\ell)}(x)\right]^c$, we know $\left|\langle w_i^{(\ell)}(0), o_0^{(\ell-1)}(x)\rangle\right| > \ell^{-1/3}C^{-1/3}m^{-1/9}$. It follows that

$$\begin{aligned}\sup_x \left|O_t^{(\ell)}(x) \cap \left[B^{(\ell)}(x)\right]^c\right| &\le \frac{\sup_x \sum_{i=1}^m \left(\langle w_i^{(\ell)}(t), o_t^{(\ell-1)}(x)\rangle - \langle w_i^{(\ell)}(0), o_0^{(\ell-1)}(x)\rangle\right)^2}{\ell^{-2/3}C^{-2/3}m^{-2/9}} \\ &= \frac{\sup_x \left\|\mathbf{W}^{(\ell)}(t)o_t^{(\ell-1)}(x) - \mathbf{W}^{(\ell)}(0)o_0^{(\ell-1)}(x)\right\|_2^2}{\ell^{-2/3}C^{-2/3}m^{-2/9}} \\ &= O\left(C^\ell m^{8/9}\right).\end{aligned}$$

Plugging the above displayed equation into (132), under $\mathcal{E}_0 \cap \mathcal{E}_3 \cap \mathcal{E}_4$, we have $\sup_x S_t^{(\ell)}(x) = O\left(C^\ell m^{8/9}\right)$.

Step 2, $\mathcal{E}_4$ occurs with high probability: Here, we prove

$$\mathbb{P}\left[\mathcal{E}_4\right] = 1 - L\exp\left(O(d\log m) - \Omega(m^{1/3})\right). \tag{133}$$

Define the deviation for any $1 \le \ell \le L$

$$\begin{aligned}&\phi_x^{(\ell-1)}(z_1, \cdots, z_m) \equiv \phi_x^{(\ell-1)}\left(z_1, \cdots, z_m; \mathbf{W}^{(1)}(0), \cdots, \mathbf{W}^{(\ell-1)}(0)\right) \\ &= \left|\frac{1}{m}\sum_{i=1}^m \mathbf{1}_{\left\{|\langle z_i, o_0^{(\ell-1)}(x)\rangle| \le \ell^{-1/3}C^{-\ell/3}m^{-1/9}\right\}} - \mathbb{E}_w\left[\mathbf{1}_{\left\{|\langle w, o_0^{(\ell-1)}(x)\rangle| \le \ell^{-1/3}C^{-\ell/3}m^{-1/9}\right\}}\right]\right|\end{aligned}$$

where $\mathbb{E}_w[\cdot]$ is the expectation over w.

We first show $\phi^{(\ell-1)}(w_1^{(\ell)}, \cdots, w_m^{(\ell)})$ concentrates on its mean for any x. By the triangle inequality, we have

$$\begin{aligned}&\left|\sup_x \phi_x^{(\ell-1)}(z_1, \cdots, z_{i-1}, z_i, z_{i+1}, \cdots, z_m) - \sup_x \phi_x^{(\ell-1)}(z_1, \cdots, z_{i-1}, z_i', z_{i+1}, \cdots, z_m)\right| \\ &\le \frac{1}{m}\left|\mathbf{1}_{\left\{|\langle z_i, o_0^{(\ell-1)}(x)\rangle| \le \ell^{-1/3}C^{-\ell/3}m^{-1/9}\right\}} - \mathbf{1}_{\left\{|\langle z_i', o_0^{(\ell-1)}(x)\rangle| \le \ell^{-1/3}C^{-\ell/3}m^{-1/9}\right\}}\right| \le \frac{1}{m}.\end{aligned}$$

Thus, by McDiarmid's inequality (Lemma 17), we get

$$\begin{aligned}&\mathbb{P}\left[\sup_x \phi_x^{(\ell-1)}(w_1^{(\ell)}(0), \cdots, w_m^{(\ell)}(0)) \le \mathbb{E}\left[\sup_x \phi_x^{(\ell-1)}(w_1^{(\ell)}(0), \cdots, w_m^{(\ell)}(0))\right] + m^{-1/3}\right] \\ &= 1 - \exp\left(-m^{1/3}\right).\end{aligned} \tag{134}$$

Now we bound $\mathbb{E}\left[\sup_x \phi_x^{(\ell-1)}(w_1^{(\ell)}(0), \cdots, w_m^{(\ell)}(0))\right]$. By Lemma 21, we have

$$\mathbb{E}\left[\sup_x \phi_x^{(\ell-1)}(w_1^{(\ell)}(0), \cdots, w_m^{(\ell)}(0))\right] \le C\sqrt{\frac{\mathsf{VC}(\mathcal{Y})}{m}}, \tag{135}$$

where $\mathcal{Y} = \left\{f_x(w) = \mathbf{1}_{\left\{|\langle w, o_0^{(\ell-1)}(x)\rangle| \le \ell^{-1/3}C^{-\ell/3}m^{-1/9}\right\}} : x \in \mathbb{S}^{d-1}\right\}$.

Note that for any $f \in \mathcal{Y}(\mathbf{W}^{(1)}(0), \cdots, \mathbf{W}^{(\ell-1)}(0))$, we can always find $g \in \mathcal{W}$ and $h \in \mathcal{W}'$ such that $f(w) = g(w)h(w)$ where

$$\mathcal{W}(\mathbf{W}^{(1)}(0), \cdots, \mathbf{W}^{(\ell-1)}(0)) = \left\{ g_x(w) = \mathbf{1}_{\left\{\langle w, o_0^{(\ell-1)}(x)\rangle \le \ell^{-1/3} C^{-\ell/3} m^{-1/9}\right\}} : x \in \mathbb{S}^{d-1} \right\},$$

and

$$\mathcal{W}'(\mathbf{W}^{(1)}(0), \cdots, \mathbf{W}^{(\ell-1)}(0)) = \left\{ h_x(w) = \mathbf{1}_{\left\{\langle w, o_0^{(\ell-1)}(x)\rangle \ge -\ell^{-1/3} C^{-\ell/3} m^{-1/9}\right\}} : x \in \mathbb{S}^{d-1} \right\}.$$

Therefore, by Lemma 20, we have

$$\mathsf{VC}(\mathcal{Y}) = O\left(\mathsf{VC}(\mathcal{W}) + \mathsf{VC}(\mathcal{W}')\right).$$

Following the same procedure as bounding $\mathsf{VC}(\mathcal{F}^{(\ell+1)})$ in the proof of Lemma 34 from Appendix B.3, we can get

$$\mathsf{VC}(\mathcal{W}) = \mathsf{VC}(\mathcal{W}') = O\left(d\ell \log m\right).$$

As a result, $\mathsf{VC}(\mathcal{Y}) = O\left(d\ell \log m\right)$. Plugging the bound of $\mathsf{VC}(\mathcal{Y})$ into (135), we get

$$\mathbb{E}\left[\sup_x \phi_x^{(\ell-1)}(w_1^{(\ell)}(0), \cdots, w_m^{(\ell)}(0))\right] \le C\sqrt{\frac{d\ell \log m}{m}} \le m^{-1/3} \tag{136}$$

when m satisfies (41).

Plugging (136) into (134) and taking union bounds over ℓ, we get with probability at least $1 - L\exp\left(-m^{1/3}\right)$, for all x and ℓ,

$$\begin{aligned}
&\frac{1}{m}\sum_{i=1}^m \mathbf{1}_{\left\{|\langle w_i^{(\ell)}, o_0^{(\ell-1)}(x)\rangle| \le \ell^{-1/3} C^{-\ell/3} m^{-1/9}\right\}} \\
&\le \mathbb{E}_w\left[\mathbf{1}_{\left\{|\langle w, o_0^{(\ell-1)}(x)\rangle| \le \ell^{-1/3} C^{-\ell/3} m^{-1/9}\right\}}\right] + 2m^{-1/3}.
\end{aligned} \tag{137}$$

Next, we bound $\sup_{x,\ell} \mathbb{E}_w\left[\mathbf{1}_{\left\{|\langle w, o_0^{(\ell-1)}(x)\rangle| \le \ell^{-1/3} C^{-\ell/3} m^{-1/9}\right\}}\right]$. Note that conditioning on $\left\{\mathbf{W}^{(k)}(0)\right\}_{k=1}^L$, $\langle w, o_0^{(\ell-1)}(x)\rangle \sim \mathcal{N}\left(0, \left\|o_0^{(\ell-1)}(x)\right\|_2^2\right)$ as $w \sim \mathcal{N}(0, \mathbf{I})$. Therefore,

$$\begin{aligned}
&\sup_x \mathbb{E}_w\left[\mathbf{1}_{\left\{|\langle w, o_0^{(\ell-1)}(x)\rangle| \le \ell^{-1/3} C^{-\ell/3} m^{-1/9}\right\}}\right] \\
&= \sup_x \mathbb{P}_w\left[|\langle w, o_0^{(\ell-1)}(x)\rangle| \le \ell^{-1/3} C^{-\ell/3} m^{-1/9}\right] \le \frac{2\ell^{-1/3} C^{-\ell/3} m^{-1/9}}{\sqrt{2\pi}\left\|o_0^{(\ell-1)}(x)\right\|_2}.
\end{aligned} \tag{138}$$

Now we bound $\left\|o_0^{(\ell-1)}(x)\right\|_2$ from below. By Lemma 6, we have with probability at least $1 - L\exp\left(O(d\log m) - \Omega(m^{1/3})\right)$, for any x and ℓ,

$$\begin{aligned}\left\|o_0^{(\ell-1)}(x)\right\|_2^2 &= \langle o_0^{(\ell-1)}(x), o_0^{(\ell-1)}(x)\rangle \\ &\geq \mathbb{E}\left[\sigma^2\left(U^{(\ell-1)}(x)\right)\right] - O\left(\frac{\ell C_1^{2\ell}}{m^{1/3}}\right) \\ &\overset{\text{(a)}}{=} 2^{-\ell} - O\left(\frac{\ell C_1^{2\ell}}{m^{1/3}}\right) = \Omega(C_2^\ell), \end{aligned} \tag{139}$$

for some constant C_1 and C_2 where (a) holds by (38) which gives $\mathbb{E}\left[\sigma^2(U^{(\ell-1)}(x))\right] = 2^{-\ell+1}$.

Plugging (139) into (138), we have with probability at least $1-L\exp\left(O(d\log m) - \Omega(m^{1/3})\right)$, for any $\ell \leq L$,

$$\sup_x \mathbb{E}_w\left[\mathbf{1}_{\left\{|\langle w, o_0^{(\ell-1)}(x)\rangle| \leq \ell^{-1/3}C^{-\ell/3}m^{-1/9}\right\}}\right] \leq C_3^\ell m^{-1/9}$$

for some constant C_3. Combining the above inequality with (137), we have with probability $1 - L\exp\left(O(d\log m) - \Omega(m^{1/3})\right) - L\exp\left(-m^{-1/3}\right)$,

$$\frac{1}{m}\sum_{i=1}^m \mathbf{1}_{\left\{|\langle w_i^{(\ell)}, o_0^{(\ell-1)}(x)\rangle| \leq \ell^{-1/3}C^{-\ell/3}m^{-1/9}\right\}} \leq C_3^\ell m^{-1/9} + 2m^{-1/3} = O\left(C_3^\ell m^{-1/9}\right).$$

This completes the proof of Step 2.

C.4 Proof of Lemma 12

Step 1, bounding $\sup_x \left\|z_0^{(k)}(x)\right\|_\infty$: We begin with proving (48). In particular, we show with probability $1 - \exp\left(-\Omega(C^{-L+k+1}m^{1/36}\right)$,

$$\sup_x \left\|z_0^{(k)}(x)\right\|_\infty \leq m^{1/36}. \tag{140}$$

Note that

$$\mathbb{P}\left[\sup_k \left\|z_0^{(k)}\right\|_\infty \leq m^{1/36}\right] = \mathbb{P}\left[\sup_k \left\|z_0^{(k)}\right\|_\infty \leq m^{1/36}|\mathcal{E}_0\right]\mathbb{P}\left[\mathcal{E}_0\right]. \tag{141}$$

Now we show

$$\mathbb{P}\left[\sup_k \left\|z_0^{(k)}\right\|_\infty \leq m^{1/36}\Big|\mathcal{E}_0\right] \geq 1 - \exp\left(O(dL\log m) - \Omega(C^{-L+k+1}m^{1/36})\right).$$

Plugging the above bound on $\mathbb{P}\left[\sup_k \left\|z_0^{(k)}\right\|_\infty \leq m^{1/36}|\mathcal{E}_0\right]$ and (112) into (141), we complete the proof of step 1.

Throughout the remaining proof of step 1, we condition on $\left\{\mathbf{W}^{(k)}(0)\right\}_{k=1}^L$ such that $\mathcal{E}_0$ holds.

Denote

$$\mathbf{P}^{(k+1)}(x) \triangleq \frac{1}{\sqrt{m}}\left[\mathbf{W}^{(k+1)}(0)\right]^\top \mathbf{D}_0^{(k+1)}(x)\cdots\frac{1}{\sqrt{m}}\left[\mathbf{W}^{(L)}(0)\right]^\top \mathbf{D}_0^{(L)}(x), \tag{142}$$

and hence $z_0^{(k)} = \mathbf{P}^{(k+1)}a$ from (113).

Therefore,

$$\left[z_0^{(k)}(x)\right]_r = \left\langle a, \frac{1}{\sqrt{m}}p_r^{(k+1)}(x)\right\rangle,$$

where $\left[z_0^{(k)}(x)\right]_r$ is the r-th coordinate of $z_0^{(k)}(x)$ and $p_r^{(k+1)}(x)$ is the r-th row of $\mathbf{P}^{(k+1)}(x)$.

Under $\mathcal{E}_0$, for any k and x, we have

$$\left\|p_r^{(k+1)}(x)\right\|_2 = \left\|\mathbf{P}^{(k+1)}(x)e_1\right\|_2 \le \left\|\mathbf{P}^{(k+1)}(x)\right\|_2 \le c_0^{L-k}. \tag{143}$$

where $e_1 = (1,0,\cdots,0)^\top \in \mathbb{R}^m$.

Since a is independent with $p_r^{(k+1)}(x)$, by Hoeffding inequality, we have for any fixed $x \in \mathbb{S}^{d-1}$,

$$\begin{aligned}
&\mathbb{P}\left[\left|\left\langle a, \frac{1}{\sqrt{m}}p_r^{(k+1)}(x)\right\rangle\right| > m^{1/36}\,\middle|\, \left\{\mathbf{W}^{(k)}\right\}_{k=1}^L \text{ s.t. } \mathcal{E}_0 \text{ holds.}\right] \\
&\le \exp\left(-\frac{m^{1/18}}{2\left\|\frac{1}{\sqrt{m}}p_r^{(k+1)}(x)\right\|_2^2}\right) \\
&= \exp\left(-\Omega\left(C^{-L+k}m^{1/18}\right)\right).
\end{aligned}$$

Taking union bounds over r, we have for any fixed x,

$$\begin{aligned}
\mathbb{P}\left[\left\|z_0^{(k)}(x)\right\|_\infty > m^{1/36}\,\middle|\, \left\{\mathbf{W}^{(k)}(0)\right\}_{k=1}^L \text{ s.t. } \mathcal{E}_0 \text{ holds.}\right] &\le m\exp\left(-\Omega\left(C^{-L+k}m^{1/18}\right)\right) \\
&= \exp\left(\log m - \Omega\left(C^{-L+k}m^{1/18}\right)\right).
\end{aligned}$$

Thus, we have

$$\begin{aligned}
&\mathbb{P}\left[\sup_x\left\|z_0^{(k)}(x)\right\|_\infty > m^{1/36}\,\middle|\, \left\{\mathbf{W}^{(k)}(0)\right\}_{k=1}^L \text{ s.t. } \mathcal{E}_0 \text{ holds.}\right] \\
&= \mathbb{P}\left[\sup_{\mathbf{P}\in\mathcal{P}^{(k+1)}}\|\mathbf{P}a\|_\infty > m^{1/36}\,\middle|\, \left\{\mathbf{W}^{(k)}(0)\right\}_{k=1}^L \text{ s.t. } \mathcal{E}_0 \text{ holds.}\right] \\
&\le |\mathcal{P}^{(k+1)}|\exp\left(\log m - \Omega\left(C^{-L+k}m^{1/18}\right)\right),
\end{aligned} \tag{144}$$

where $\mathcal{P}^{(k+1)}(\mathbf{W}^{(1)}(0),\cdots,\mathbf{W}^{(L)}(0)) = \left\{\mathbf{P}^{(k+1)}(x) : x \in \mathbb{S}^{d-1}\right\}$ with $\mathbf{P}^{(k+1)}(x)$ defined in (142).

Now we bound $|\mathcal{P}^{(k+1)}|$. Recall the definition of $\mathcal{D}_L$ in Section 5. By definition, there is an injective mapping from $\mathcal{P}^{(k+1)}$ to $\mathcal{D}_L$. Therefore, we have $|\mathcal{P}^{(k+1)}| \le |\mathcal{D}_L| \le m^{dL}$, where the last inequality holds by (83).

Plugging this bound on $|\mathcal{P}^{(k+1)}|$ into (144), we get

$$\mathbb{P}\left[\sup_x \left\|z_0^{(k)}(x)\right\|_\infty > m^{1/36} \,\middle|\, \left\{\mathbf{W}^{(k)}(0)\right\}_{k=1}^{L} \text{ s.t. } \mathcal{E}_0 \text{ holds}\right] \le m^{dL}\exp\left(\log m - \Omega(C^{-L+k}m^{1/18})\right) = \exp\left(-\Omega(C^{-L+k}m^{1/18})\right).$$

Step 2, bounding $\left\|z_t^{(\ell)}(x) - z_0^{(\ell)}(x)\right\|_2$: Now we prove the second inequality (49) holds with high probability. Assume (140), all three conditions in Lemma 10 and (47) hold, which can be guaranteed with probability at least $1-\exp\left(-\Omega(C^{-L+k}m^{1/36})\right)$ following Lemma 10, Lemma 11 and Step 1 above.

Fix arbitrary t. We will use an inductive argument on layer to prove (49) holds for all $1 \le \ell \le L$.

By definition, $z_t^{(L)}(x) = a$ for any x. Therefore, (49) holds at $\ell = L$.

Now suppose (49) holds at some $\ell+1$, we are going to show (49) holds at ℓ as well. Note that

$$z_t^{(\ell)}(x) = \frac{1}{\sqrt{m}}\left[\mathbf{W}^{(\ell+1)}(t)\right]^\top \mathbf{D}_t^{(\ell+1)}(x) z_t^{(\ell+1)}(x).$$

Similar to (114), by the triangle inequality, we have

$$\begin{aligned}\left\|z_t^{(\ell)}(x) - z_0^{(\ell)}(x)\right\|_2 \le{}& \left\|\left[\mathbf{W}^{(\ell+1)}(t) - \mathbf{W}^{(\ell+1)}(0)\right]^\top \frac{1}{\sqrt{m}}\mathbf{D}_t^{(\ell+1)}(x) z_t^{(\ell+1)}(x)\right\|_2 \\ &+ \left\|\frac{1}{\sqrt{m}}\left[\mathbf{W}^{(\ell+1)}(0)\right]^\top \mathbf{D}_t^{(\ell+1)}(x)\left(z_t^{(\ell+1)}(x) - z_0^{(\ell+1)}(x)\right)\right\|_2 \\ &+ \left\|\frac{1}{\sqrt{m}}\left[\mathbf{W}^{(\ell)}(0)\right]^\top \left(\mathbf{D}_t^{(\ell+1)}(x) - \mathbf{D}_0^{(\ell+1)}(x)\right) z_0^{(\ell+1)}(x)\right\|_2. \end{aligned} \tag{145}$$

Now we bound the first term on the right hand side of (145). By the triangle inequality,

$$\begin{aligned}\left\|z_t^{(\ell+1)}(x)\right\|_2 &\le \left\|z_0^{(\ell+1)}(x)\right\|_2 + \left\|z_t^{(\ell+1)}(x) - z_0^{(\ell+1)}(x)\right\|_2 \\ &\le c_0^{L-\ell}\sqrt{m} + O\left(C^{2L-\ell-1}m^{17/36}\right) = O(C_1^{2L-\ell-1}\sqrt{m}).\end{aligned} \tag{146}$$

for some constant C and C_1 where the second inequality holds by (118) under $\mathcal{E}_0$ and the inductive hypothesis.

As a result,

$$\begin{aligned}&\left\|\left[\mathbf{W}^{(\ell+1)}(t) - \mathbf{W}^{(\ell+1)}(0)\right]^\top \frac{1}{\sqrt{m}}\mathbf{D}_t^{(\ell+1)}(x) z_t^{(\ell+1)}(x)\right\|_2 \\ &\le \left\|\mathbf{W}^{(\ell+1)}(t) - \mathbf{W}^{(\ell+1)}(0)\right\|_2 \frac{1}{\sqrt{m}}\left\|z_t^{(\ell+1)}(x)\right\|_2 \\ &\le C_1^{2L-\ell-1}m^{1/3},\end{aligned} \tag{147}$$

where the last inequality holds by (44) and $R_t \le m^{1/3}$ from Lemma 10, and the above bound of $\left\|z_t^{(\ell+1)}(x)\right\|_2$.

To bound the second term, note that under $\mathcal{E}_0$ and the inductive hypothesis, we have

$$\begin{aligned}
&\left\| \frac{1}{\sqrt{m}} \left[\mathbf{W}^{(\ell+1)}(0) \right]^\top \mathbf{D}_t^{(\ell+1)}(x) \left(z_t^{(\ell+1)}(x) - z_0^{(\ell+1)}(x) \right) \right\|_2 \\
&\le \frac{1}{\sqrt{m}} \left\| \mathbf{W}^{(\ell+1)}(0) \right\|_2 \left\| \mathbf{D}_t^{(\ell+1)}(x) \right\|_2 \left\| z_t^{(\ell+1)}(x) - z_0^{(\ell+1)}(x) \right\|_2 \\
&= O\left(C^{2L-\ell} m^{17/36} \right). \qquad (148)
\end{aligned}$$

To bound the last term on the right hand side of (145), note that under $\mathcal{E}_0$,

$$\begin{aligned}
&\left\| \frac{1}{\sqrt{m}} \left[\mathbf{W}^{(\ell)}(0) \right]^\top \left(\mathbf{D}_t^{(\ell+1)}(x) - \mathbf{D}_0^{(\ell+1)}(x) \right) z_0^{(\ell+1)}(x) \right\|_2 \\
&\le \frac{1}{\sqrt{m}} \left\| \mathbf{W}^{(\ell)}(0) \right\|_2 \left\| \left(\mathbf{D}_t^{(\ell+1)}(x) - \mathbf{D}_0^{(\ell+1)}(x) \right) z_0^{(\ell+1)}(x) \right\|_2 \\
&\le c_0 \left\| \left(\mathbf{D}_t^{(\ell+1)}(x) - \mathbf{D}_0^{(\ell+1)}(x) \right) z_0^{(\ell+1)}(x) \right\|_2 .
\end{aligned}$$

By definition, we have

$$\begin{aligned}
&\left\| \left(\mathbf{D}_t^{(\ell+1)}(x) - \mathbf{D}_0^{(\ell+1)}(x) \right) z_0^{(\ell+1)}(x) \right\|_2 \\
&\le \sqrt{ \sum_{i=1}^m \left(\mathbf{1}_{\left\{ \langle w_i^{(\ell+1)}(t), o_t^{(\ell)}(x) \rangle \ge 0 \right\}} - \mathbf{1}_{\left\{ \langle w_i^{(\ell+1)}(0), o_0^{(\ell)}(x) \rangle \ge 0 \right\}} \right)^2 \left[z_0^{(\ell+1)}(x) \right]_i^2 } \\
&\le \left\| z_0^{(\ell)}(x) \right\|_\infty \sqrt{ \left\| S_t^{(\ell+1)} \right\|_\infty } \\
&= O\left(C^{2L-\ell-1} m^{17/36} \right), \qquad (149)
\end{aligned}$$

where the last equality holds by (47) and the assumption $\sup_{x,k} \left\| z_0^{(k)}(x) \right\|_\infty \le m^{1/36}$.

Therefore,

$$\left\| \frac{1}{\sqrt{m}} \left[\mathbf{W}^{(\ell)}(0) \right]^\top \left(\mathbf{D}_t^{(\ell+1)}(x) - \mathbf{D}_0^{(\ell+1)}(x) \right) z_0^{(\ell+1)}(x) \right\|_2 = O\left(C^{2L-\ell} m^{17/36} \right).$$

Plugging (147), (148) and the above displayed equation into the right hand side of (145), we complete the proof of Step 2.

Appendix D. Proof of lemmas in Section 7

D.1 Proof of Lemma 14

Recall from (21) that $\Phi^{(\ell)}(x, x') \triangleq \mathbb{E}\left[\sigma\left(U^{(\ell-1)}(x) \right) \sigma\left(U^{(\ell-1)}(x') \right) \right] q_L^{(\ell)}(x, x')$ where $U^{(\ell)}(x)$ is defined in (15) and $q_L^{(\ell)}$ is defined in (20).

Step 1, Φ is positive semi-definite: We begin with showing Φ is positive semi-definite (PSD). Since $\Phi = \sum_{\ell=1}^{L} \Phi^{(\ell)}$, by (64) and (65) of Lemma 25, it suffices to show both $\mathbb{E}\left[\sigma(U^{(\ell)}(x))\sigma(U^{(\ell)}(x'))\right]$ and $q_L^{(\ell)}(x,x')$ are PSD kernels.

Denote $G(x,x') \triangleq \mathbb{E}\left[\sigma(U^{(\ell)}(x))\sigma(U^{(\ell)}(x'))\right]$. Here, we apply Lemma 26 to prove $G(x,x')$ is PSD. To begin with, we show

$$\mathbb{E}\left[\sigma^2(U^{(\ell)}(x))\sigma^2(U^{(\ell)}(x'))\right] < \infty. \tag{150}$$

By definition (15), we have

$$\mathbb{E}\left[\left(U^{(\ell)}(x)\right)^2\right] = \left[\Sigma^{(\ell-1)}\right]_{11} \le \mathbb{E}\left[\left(U^{(\ell-1)}(x)\right)^2\right]$$

where the last inequality holds since $\sigma^2(U^{(\ell-1)}(x)) \le \left(U^{(\ell-1)}(x)\right)^2$.

Since for any $x \in \mathbb{S}^{d-1}$, $\mathbb{E}\left[\left(U^{(0)}(x)\right)^2\right] = \|x\|_2^2 \le 1$, we get for any ℓ, $\mathbb{E}\left[\left(U^{(\ell)}(x)\right)^2\right] \le 1$. By Cauchy-Schwartz inequality, we have

$$\mathbb{E}\left[U^{(\ell)}(x)U^{(\ell)}(x')\right] \le \sqrt{\mathbb{E}\left[\left(U^{(\ell)}(x)\right)^2\right]\mathbb{E}\left[\left(U^{(\ell)}(x')\right)^2\right]} \le 1.$$

Thus, for any $x, y \in \mathbb{S}^{d-1}$,

$$\begin{aligned}\mathbb{E}\left[\sigma^2(U^{(\ell)}(x))\sigma^2(U^{(\ell)}(y))\right] &\le \mathbb{E}\left[\left(U^{(\ell)}(x)U^{(\ell)}(y)\right)^2\right] \\ &= \mathbb{E}\left[\left(U^{(\ell)}(x)\right)^2\right]\mathbb{E}\left[\left(U^{(\ell)}(y)\right)^2\right] + 2\mathbb{E}\left[U^{(\ell)}(x)U^{(\ell)}(y)\right] \le 3,\end{aligned}$$

where the equality holds by Isserlis' Theorem (Isserlis, 1918).

By Cauchy-Schwartz inequality, for any $g \in L_2(\mu)$, we get

$$\begin{aligned}&\int\int \mathbb{E}\left[\left|g(x)\sigma(U^{(\ell)}(x))\sigma(U^{(\ell)}(y))g(y)\right|\right] d\mu(x)d\mu(y) \\ &\le \int\int g^2(x)g^2(y)d\mu(x)d\mu(y) \int\int \mathbb{E}\left[\left(\sigma(U^{(\ell)}(x))\sigma(U^{(\ell)}(y))\right)^2\right] d\mu(x)d\mu(y) < \infty,\end{aligned}$$

where the last inequality holds by (150) and the fact that $g \in L_2(\mu)$.

As a result, by Fubini Theorem, we have

$$\begin{aligned}\int\int g(x)G(x,y)g(y)d\mu(x)d\mu(y) &= \int\int \mathbb{E}\left[g(x)\sigma(U^{(\ell)}(x))\sigma(U^{(\ell)}(y))g(y)\right] d\mu(x)d\mu(y) \\ &= \mathbb{E}\left[\int\int g(x)\sigma(U^{(\ell)}(x))\sigma(U^{(\ell)}(y))g(y)d\mu(x)d\mu(y)\right] \\ &= \mathbb{E}\left[\left(\int g(x)\sigma(U^{(\ell)}(x))d\mu(x)\right)^2\right] \ge 0.\end{aligned}$$

By Lemma 26, we get $G(x,x')$ is PSD.

Now we show $q_L^{(\ell)}(x,x')$ is PSD by induction. By definition, we know

$$q_L^{(\ell)}(x,x') = \prod_{k=\ell}^{L} \frac{\pi - \arccos \rho^{(k)}(x,x')}{2\pi}.$$

Following (65) of Lemma 25, it remains to show

$$F^{(k)}(x,x') \triangleq \frac{\pi - \arccos \rho^{(k)}(x,x')}{2\pi}$$

is PSD for any k where $\rho^{(k)}(x,x') = \frac{\mathbb{E}\left[\sigma\left(U^{(k)}(x)\right)\sigma\left(U^{(k)}(x')\right)\right]}{\sqrt{\mathbb{E}\left[\sigma^2\left(U^{(k)}(x)\right)\right]}\sqrt{\mathbb{E}\left[\sigma^2\left(U^{(k)}(x')\right)\right]}}$ is defined in (93).

Note that we have shown the numerator of $\rho^{(k)}(x,x')$ is PSD. From (38), we know the denominator of $\rho^{(k)}(x,x')$ is some constant independent of x and x'. Therefore, $\rho^{(k)}(x,x')$ is PSD and hence we have $\rho^{(k)}(x,x') = \langle \phi(x), \phi(x') \rangle$ for some function ϕ.

Since $\frac{\pi - \arccos(\langle x,x' \rangle)}{2\pi}$ is PSD (Cho and Saul, 2009), by (66) of Lemma 25, we get $F^{(k)}$ is PSD.

Step 2, $\|\Phi\|_2 \le \|\Phi\|_\infty \le \frac{L}{2}$ The inequality $\|\Phi\|_2 \le \|\Phi\|_\infty$ follows from Lemma 28.

To bound $\|\Phi\|_\infty$, we follow (36) and (35) and get $\left\|\Phi^{(\ell)}\right\|_\infty \le \frac{1}{2}$ for all ℓ.

Since $\Phi = \sum_{\ell=1}^{L} \Phi^{(\ell)}$, we have $\|\Phi\|_\infty \le \frac{L}{2}$.

Step 3, bounding $\|\mathsf{K}_t\|_2$ **and** $\|\mathsf{Q}_t\|_2$ By definition, the eigenvalues of K_t equals $1 - \eta_t \lambda_i, i = 1, 2, \cdots$ where λ_i is the i-th largest eigenvalue of Φ. Since $0 \le \lambda_i \le \frac{L}{2}$ for all i, with $\eta_t \le \frac{2}{L}$, we have $0 \le 1 - \eta_t \lambda_i \le 1$ for all i. This shows $\|\mathsf{K}_t\|_2 \le 1$ and K_t is PSD.

Similarly, we bound $\|\mathsf{Q}_t\|_2$. Following Theorem 1 and Proposition 9, by the triangle inequality, with probability at least $1 - \exp\left(-\Omega(C^{-L} m^{1/36})\right)$,

$$\|H_t - \Phi\|_\infty \le \|H_t - H_0\|_\infty + \|H_0 - \Phi\|_\infty = O\left(\frac{C^L}{m^{1/36}}\right).$$

Therefore, for $m = \exp\left(\Omega(L)\right)$ which is guaranteed by (41), we have

$$\|\mathsf{H}_t\|_2 \le \|H_t\|_\infty \le \|H_t - \Phi\|_\infty + \|\Phi\|_\infty \le \frac{2L}{3} \tag{151}$$

for all t. By the definition of H_t in (1), we know H_t is PSD. As a result, we have

$$0 \le \lambda_i(\mathsf{H}_t) \le \frac{2L}{3}.$$

For $\eta_t \le \frac{3}{2L}$, we have $0 \le 1 - \eta_t \lambda_i(\mathsf{H}_t) \le 1, \forall i$, where $\lambda_i(\mathsf{H}_t)$ is the i-th largest eigenvalue of H_t. This shows $\|\mathsf{Q}_t\|_2 \le 1$ and Q_t is PSD.

D.2 Proof of Lemma 15

Recall that

$$\epsilon_t(x,x') = f(x; \mathbf{W}(t)) - f(x; \mathbf{W}(t+1)) + \eta_t H_t(x,x') \left(f^*(x') + u_t - f(x'; \mathbf{W}(t))\right). \tag{152}$$

Here we provide a lower bound of ϵ_t. The upper bound of ϵ_t can be obtained analogously.

The proof consists of three steps. Firstly, we study the evolution of the prediction value $f(x;\mathbf{W}(t+1)) - f(x;\mathbf{W}(t))$. Since the change of the prediction value is driven by the update of weight $\mathbf{W}(t)$ in (10), intuitively we have

$$\begin{aligned} f(x;\mathbf{W}(t+1)) - f(x;\mathbf{W}(t)) &\approx \sum_{\ell=1}^{L} \left\langle \frac{\partial f(x;\mathbf{W}(t))}{\partial \mathbf{W}^{(\ell)}}, \mathbf{W}^{(\ell)}(t+1) - \mathbf{W}^{(\ell)}(t) \right\rangle \\ &= \eta_t \left(\Delta_t(X_t) + u_t\right) H_t(x, X_t). \end{aligned}$$

To justify the above approximation, we first show

$$f(x;\mathbf{W}(t+1)) - f(x;\mathbf{W}(t)) \le \sum_{\ell=1}^{L} \Delta_{\mathbf{W}^{(\ell)}(t)}(x) + \sum_{\ell=1}^{L-1} \mathfrak{A}_t^{(\ell)}(x),$$

for some $\Delta_{\mathbf{W}^{(\ell)}(t)}(x)$ and $\mathfrak{A}_t^{(\ell)}(x)$ defined in (155) and (159).

Then we prove that for each ℓ,

$$\Delta_{\mathbf{W}^{(\ell)}(t)}(x) = \eta_t(\Delta_t(X_t) + u_t) \left(H_t^{(\ell)}(x, X_t) + \mathfrak{B}_t^{(\ell)}(x, X_t) + \mathfrak{R}_t^{(\ell)}(x, X_t) \right) \tag{153}$$

where $\mathfrak{B}_t^{(\ell)}(x, X_t)$ and $\mathfrak{R}_t^{(\ell)}(x, X_t)$ are some error terms defined in (164) and (165).

Following the definition of ϵ_t, we have

$$\begin{aligned} \epsilon_t(x) &= f(x;\mathbf{W}(t)) - f(x;\mathbf{W}(t+1)) + \eta_t H_t(x, X_t) \left(\Delta_t(X_t) + u_t\right) \\ &\ge -\eta_t \left(\Delta_t(X_t) + u_t\right) \sum_{\ell=1}^{L} \left(\mathfrak{B}_t(x, X_t) + \mathfrak{R}_t^{(\ell)}(x, X_t) \right) - \sum_{r=1}^{L-1} \mathfrak{A}_t^{(r)}. \end{aligned} \tag{154}$$

To bound ϵ_t, it suffices to bound $\mathfrak{A}_t^{(\ell)}$, $\mathfrak{B}_t^{(\ell)}$ and $\mathfrak{R}_t^{(\ell)}$. In short, we show these terms depend on either the change of output $o_{t+1}^{(\ell)} - o_t^{(\ell)}$ or the change of activation pattern $\mathbf{D}_{t+1}^{(\ell)}(x) - \mathbf{D}_t^{(\ell)}(x)$ which are shown to be small with high probability in Section 6.

D.2.1 Analysis of the evolution of $f(x;\mathbf{W}(t))$

For all $0 \le \ell \le L$, define

$$\mathbf{Z}_t^{(\ell)+}(x) \triangleq \mathsf{diag}\left\{ \mathbf{1}_{\left\{ \left[z_t^{(\ell)}(x) \right]_i \ge 0 \right\}} \right\}, \quad \text{and} \quad \mathbf{Z}_t^{(\ell)-}(x) \triangleq \mathsf{diag}\left\{ \mathbf{1}_{\left\{ \left[z_t^{(\ell)}(x) \right]_i < 0 \right\}} \right\}.$$

When $\ell = L$, since $z_t^{(L)}(x) = a$ does not change over time,

$$\mathbf{Z}_t^{(L)+}(x) = \mathsf{diag}\left\{a_i = 1\right\}, \ \mathbf{Z}_t^{(L)-}(x) = \mathsf{diag}\left\{a_i = -1\right\}, \forall t.$$

Denote

$$\begin{aligned} \Delta_{\mathbf{W}^{(\ell)}(t)}(x) \triangleq & \left[z_t^{(\ell)}(x) \right]^\top \left[\mathbf{Z}_t^{(\ell)+}(x) \mathbf{D}_{t+1}^{(\ell)}(x) + \mathbf{Z}_t^{(\ell)-}(x) \mathbf{D}_t^{(\ell)}(x) \right] \\ & \frac{1}{\sqrt{m}} \left(\mathbf{W}^{(\ell)}(t+1) - \mathbf{W}^{(\ell)}(t) \right) o_{t+1}^{(\ell-1)}(x), \end{aligned} \tag{155}$$

and

$$\Delta_{o_t^{(\ell)}}(x) \triangleq \left[z_t^{(\ell+1)}(x)\right]^\top \left[\mathbf{Z}_t^{(\ell+1)+}(x)\mathbf{D}_{t+1}^{(\ell+1)}(x) + \mathbf{Z}_t^{(\ell+1)-}(x)\mathbf{D}_t^{(\ell+1)}(x)\right] \frac{1}{\sqrt{m}}\mathbf{W}^{(\ell+1)}(t)\left(o_{t+1}^{(\ell)}(x) - o_t^{(\ell)}(x)\right).$$

Intuitively, $\Delta_{\mathbf{W}^{(\ell)}(t)}(x)$ and $\Delta_{o_t^{(\ell)}}(x)$ capture the change of prediction value $f(x;\mathbf{W}(t))$ by the change of $\mathbf{W}^{(\ell)}(t)$ and $o_t^{(\ell)}(x)$, respectively.

Note that for any vector $p, b, e \in \mathbb{R}^m$, we have

$$\begin{aligned} p^\top(\sigma(b)-\sigma(e)) &= \sum_{i:p_i\geq 0} p_i(\sigma(b_i)-\sigma(e_i)) + \sum_{i:p_i<0}(-p_i)(\sigma(e_i)-\sigma(b_i)) \\ &\leq \sum_{i:p_i\geq 0} p_i\mathbf{1}_{\{b_i\geq 0\}}(b_i-e_i) + \sum_{i:p_i<0}(-p_i)\mathbf{1}_{\{e_i\geq 0\}}(e_i-b_i), \end{aligned} \tag{156}$$

where the last inequality holds by the fact that $\sigma(y)-\sigma(x) \leq \mathbf{1}_{\{y\geq 0\}}(y-x)$.

Therefore, we have

$$\begin{aligned} &f(x;\mathbf{W}(t+1)) - f(x;\mathbf{W}(t)) \\ &= \frac{1}{\sqrt{m}}a^\top\left(\sigma(\mathbf{W}^{(L)}(t+1)o_{t+1}^{(L-1)}(x)) - \sigma(\mathbf{W}^{(L)}(t)o_t^{(L-1)}(x))\right) \\ &\leq \frac{1}{\sqrt{m}}a^\top\left(\mathbf{Z}_t^{(L)+}\mathbf{D}_{t+1}^{(L)}(x) + \mathbf{Z}_t^{(L)-}\mathbf{D}_t^{(L)}(x)\right)\left(\mathbf{W}^{(L)}(t+1)o_{t+1}^{(L-1)}(x) - \mathbf{W}^{(L)}(t)o_t^{(L-1)}(x)\right), \\ &= \Delta_{\mathbf{W}^{(L)}(t)}(x) + \Delta_{o_t^{(L-1)}}(x), \end{aligned} \tag{157}$$

where the last equality holds since

$$\begin{aligned} &\mathbf{W}^{(L)}(t+1)o_{t+1}^{(L-1)}(x) - \mathbf{W}^{(L)}(t)o_t^{(L-1)}(x) \\ &= \left(\mathbf{W}^{(L)}(t+1) - \mathbf{W}^{(L)}(t)\right)o_{t+1}^{(L-1)}(x) + \mathbf{W}^{(L)}(t)\left(o_{t+1}^{(L-1)}(x) - o_t^{(L-1)}(x)\right), \end{aligned}$$

$$\Delta_{\mathbf{W}^{(L)}(t)}(x) = a^\top\left[\mathbf{Z}_t^{(L)+}(x)\mathbf{D}_{t+1}^{(L)}(x) + \mathbf{Z}_t^{(L)-}(x)\mathbf{D}_t^{(L)}(x)\right]\frac{1}{\sqrt{m}}\left(\mathbf{W}^{(L)}(t+1) - \mathbf{W}^{(L)}(t)\right)o_{t+1}^{(L-1)}(x),$$

and

$$\Delta_{o_t^{(L-1)}}(x) = a^\top\left[\mathbf{Z}_t^{(L)+}(x)\mathbf{D}_{t+1}^{(L)}(x) + \mathbf{Z}_t^{(L)-}(x)\mathbf{D}_t^{(L)}(x)\right]\frac{1}{\sqrt{m}}\mathbf{W}^{(L)}(t)\left(o_{t+1}^{(L-1)}(x) - o_t^{(L-1)}(x)\right).$$

Intuitively, since the change of $o_t^{(\ell)}$ comes from the update of $\mathbf{W}^{(\ell)}(t)$ and $o_t^{(\ell-1)}$, we can obtain a recursive relation of $\Delta_{o_t^{(\ell)}}$. In particular, we show that for all ℓ,

$$\Delta_{o_t^{(\ell)}}(x) \leq \Delta_{o_t^{(\ell-1)}}(x) + \Delta_{\mathbf{W}^{(\ell)}(t)}(x) + \mathfrak{A}_t^{(\ell)}(x), \tag{158}$$

where

$$\begin{aligned} &\mathfrak{A}_t^{(\ell)}(x) \\ &\triangleq \left[z_t^{(\ell+1)}(x)\right]^\top \mathbf{Z}_t^{(\ell+1)+}(x)\left[\mathbf{D}_{t+1}^{(\ell)}(x) - \mathbf{D}_t^{(\ell)}(x)\right]\frac{1}{\sqrt{m}}\mathbf{W}^{(\ell+1)}(t)\left(o_{t+1}^{(\ell)}(x) - o_t^{(\ell)}(x)\right). \end{aligned} \tag{159}$$

Note that $\mathbf{Z}_t^{(\ell)+}(x) + \mathbf{Z}_t^{(\ell)-}(x) = \mathbf{I}$. Therefore, for any $1 \le \ell \le L$,

$$\mathbf{Z}_t^{(\ell)+}(x)\mathbf{D}_{t+1}^{(\ell)}(x) + \mathbf{Z}_t^{(\ell)-}(x)\mathbf{D}_t^{(\ell)}(x) = \mathbf{D}_t^{(\ell)}(x) + \mathbf{Z}_t^{(\ell)+}(x)\left(\mathbf{D}_{t+1}^{(\ell)}(x) - \mathbf{D}_t^{(\ell)}(x)\right). \tag{160}$$

Thus, we have

$$\begin{aligned}
&\Delta_{o_t^{(\ell)}}(x) \\
&= \left[z_t^{(\ell+1)}(x)\right]^\top \left[\mathbf{Z}_t^{(\ell+1)+}(x)\mathbf{D}_{t+1}^{(\ell+1)}(x) + \mathbf{Z}_t^{(\ell+1)-}(x)\mathbf{D}_t^{(\ell+1)}(x)\right]\frac{1}{\sqrt{m}}\mathbf{W}^{(\ell+1)}(t)\left(o_{t+1}^{(\ell)}(x) - o_t^{(\ell)}(x)\right) \\
&= \underbrace{\left[z_t^{(\ell+1)}(x)\right]^\top \mathbf{D}_t^{(\ell+1)}(x)\frac{1}{\sqrt{m}}\mathbf{W}^{(\ell+1)}(t)\left(o_{t+1}^{(\ell)}(x) - o_t^{(\ell)}(x)\right)}_{\text{(I)}} \\
&+ \underbrace{\left[z_t^{(\ell+1)}(x)\right]^\top \mathbf{Z}_t^{(\ell+1)+}(x)\left[\mathbf{D}_{t+1}^{(\ell)}(x) - \mathbf{D}_t^{(\ell)}(x)\right]\frac{1}{\sqrt{m}}\mathbf{W}^{(\ell+1)}(t)\left(o_{t+1}^{(\ell)}(x) - o_t^{(\ell)}(x)\right)}_{\mathfrak{A}_t^{(\ell)}(x)}.
\end{aligned}$$

From (50), we know $\left[z_t^{(\ell)}(x)\right]^\top = \left[z_t^{(\ell+1)}(x)\right]^\top \frac{1}{\sqrt{m}}\mathbf{D}_t^{(\ell+1)}(x)\mathbf{W}^{(\ell+1)}(t)$. Thus, we have

$$\begin{aligned}
\text{(I)} &= \left[z_t^{(\ell)}(x)\right]^\top \left(o_{t+1}^{(\ell)}(x) - o_t^{(\ell)}(x)\right) \\
&= \frac{1}{\sqrt{m}}\left[z_t^{(\ell)}(x)\right]^\top \left(\sigma\left(\mathbf{W}^{(\ell)}(t+1)o_{t+1}^{(\ell-1)}(x)\right) - \sigma\left(\mathbf{W}^{(\ell)}(t)o_t^{(\ell-1)}(x)\right)\right) \\
&\overset{\text{(i)}}{\le} \left[z_t^{(\ell)}(x)\right]^\top \left(\mathbf{Z}^{(\ell)+}\mathbf{D}_{t+1}^{(\ell)}(x) + \mathbf{Z}^{(\ell)-}\mathbf{D}_t^{(\ell)}(x)\right)\frac{1}{\sqrt{m}}\left(\mathbf{W}^{(\ell)}(t+1)o_{t+1}^{(\ell-1)}(x) - \mathbf{W}^{(\ell)}(t)o_t^{(\ell-1)}(x)\right) \\
&= \left[z_t^{(\ell)}(x)\right]^\top \left(\mathbf{Z}^{(\ell)+}\mathbf{D}_{t+1}^{(\ell)}(x) + \mathbf{Z}^{(\ell)-}\mathbf{D}_t^{(\ell)}(x)\right)\frac{1}{\sqrt{m}}\Big\{\left(\mathbf{W}^{(\ell)}(t+1) - \mathbf{W}^{(\ell)}(t)\right)o_{t+1}^{(\ell-1)}(x) \\
&\quad + \mathbf{W}^{(\ell)}(t)\left(o_{t+1}^{(\ell-1)}(x) - o_t^{(\ell-1)}(x)\right)\Big\} \\
&= \Delta_{\mathbf{W}^{(\ell)}(t)}(x) + \Delta_{o_t^{(\ell-1)}}(x),
\end{aligned}$$

where (i) holds by (156) and the last equality holds by the definition of $\Delta_{\mathbf{W}^{(\ell)}(t)}$ and $\Delta_{o_t^{(\ell-1)}}$. Hence, we get (158).

Recursively plugging (158) into the right hand side of (157), we get

$$f(x;\mathbf{W}(t+1)) - f(x;\mathbf{W}(t)) \le \sum_{\ell=1}^{L}\Delta_{\mathbf{W}^{(\ell)}(t)}(x) + \sum_{\ell=1}^{L-1}\mathfrak{A}_t^{(\ell)}(x). \tag{161}$$

D.2.2 Decomposing $\Delta_{\mathbf{W}^{(\ell)}(t)}(x)$

Here we prove (153). Plugging (10) into (155) to replace $\mathbf{W}^{(\ell)}(t+1) - \mathbf{W}^{(\ell)}(t)$, we have

$$\begin{aligned}
&\Delta_{\mathbf{W}^{(\ell)}(t)} \\
&= \frac{1}{m}\left[z_t^{(\ell)}(x)\right]^\top \left[\mathbf{Z}_t^{(\ell)+}(x)\mathbf{D}_{t+1}^{(\ell)}(x) + \mathbf{Z}_t^{(\ell)-}(x)\mathbf{D}_t^{(\ell)}(x)\right]\langle o_{t+1}^{(\ell-1)}(x), o_t^{(\ell-1)}(X_t)\rangle \\
&\quad \eta_t\left(\Delta_t(X_t) + u_t\right)\mathbf{D}_t^{(\ell)}(X_t)z_t^{(\ell)}(X_t).
\end{aligned} \tag{162}$$

Note that

$$\begin{aligned}
&\left(\mathbf{Z}_t^{(\ell)+}(x)\mathbf{D}_{t+1}^{(\ell)}(x)+\mathbf{Z}_t^{(\ell)-}(x)\mathbf{D}_t^{(\ell)}(x)\right)\langle o_{t+1}^{(\ell-1)}(x),o_t^{(\ell-1)}(X_t)\rangle\\
&=\left(\mathbf{Z}_t^{(\ell)+}(x)\mathbf{D}_{t+1}^{(\ell)}(x)+\mathbf{Z}_t^{(\ell)-}(x)\mathbf{D}_t^{(\ell)}(x)\right)\langle o_{t+1}^{(\ell-1)}(x)-o_t^{(\ell-1)}(x)+o_t^{(\ell-1)}(x),o_t^{(\ell-1)}(X_t)\rangle\\
&\overset{(a)}{=}\left(\mathbf{Z}_t^{(\ell)+}(x)\mathbf{D}_{t+1}^{(\ell)}(x)+\mathbf{Z}_t^{(\ell)-}(x)\mathbf{D}_t^{(\ell)}(x)\right)\langle o_{t+1}^{(\ell-1)}(x)-o_t^{(\ell-1)}(x),o_t^{(\ell-1)}(X_t)\rangle\\
&+\mathbf{D}_t^{(\ell)}(x)\langle o_t^{(\ell-1)}(x),o_t^{(\ell-1)}(X_t)\rangle+\mathbf{Z}_t^{(\ell)+}(x)\left(\mathbf{D}_{t+1}^{(\ell)}(x)-\mathbf{D}_t^{(\ell)}(x)\right)\langle o_t^{(\ell)}(x),o_t^{(\ell-1)}(X_t)\rangle,
\end{aligned}$$

where (a) holds by (160).

Plugging the above equation into (162), we have

$$\begin{aligned}
\Delta_{\mathbf{W}^{(\ell)}(t)}=&\frac{1}{m}\eta_t\left(\Delta_t(X_t)+u_t\right)\Bigg\{\left[z_t^{(\ell)}(x)\right]^\top\mathbf{D}_t^{(\ell)}(x)\langle o_t^{(\ell-1)}(x),o_t^{(\ell-1)}(X_t)\rangle\\
&+\left[z_t^{(\ell)}(x)\right]^\top\mathbf{Z}_t^{(\ell)+}(x)\left(\mathbf{D}_t^{(\ell)}(x)-\mathbf{D}_t^{(\ell)}(x)\right)\langle o_t^{(\ell-1)}(x),o_t^{(\ell-1)}(X_t)\rangle\\
&+\left[z_t^{(\ell)}(x)\right]^\top\left(\mathbf{Z}_t^{(\ell)+}(x)\mathbf{D}_{t+1}^{(\ell)}(x)+\mathbf{Z}_t^{(\ell)-}(x)\mathbf{D}_t^{(\ell)}(x)\right)\langle o_{t+1}^{(\ell-1)}(x)-o_t^{(\ell-1)}(x),o_t^{(\ell-1)}(X_t)\rangle\Bigg\}\\
&\mathbf{D}_t^{(\ell)}(X_t)z_t^{(\ell)}(X_t)\\
=&\eta_t\left(\Delta_t(X_t)+u_t\right)\left(H_t^{(\ell)}(x,X_t)+\mathfrak{B}_t^{(\ell)}(x,X_t)+\mathfrak{R}_t^{(\ell)}(x,X_t)\right),
\end{aligned}\tag{163}$$

where

$$H_t^{(\ell)}(x,x')=\frac{1}{m}\left\langle\mathbf{D}_t^{(\ell)}(x)z_t^{(\ell)}(x)\left[o_t^{(\ell-1)}(x)\right]^\top,\mathbf{D}_t^{(\ell)}(x')z_t^{(\ell)}(x')\left[o_t^{(\ell-1)}(x')\right]^\top\right\rangle$$

from (39),

$$\begin{aligned}
\mathfrak{B}_t^{(\ell)}(x,X_t)\triangleq&\left[z_t^{(\ell)}(x)\right]^\top\mathbf{Z}_t^{(\ell)+}(x)\left(\mathbf{D}_{t+1}^{(\ell)}(x)-\mathbf{D}_t^{(\ell)}(x)\right)\frac{1}{m}\langle o_t^{(\ell-1)}(x),o_t^{(\ell-1)}(X_t)\rangle\\
&\mathbf{D}_t^{(\ell)}(X_t)z_t^{(\ell)}(X_t),
\end{aligned}\tag{164}$$

and

$$\begin{aligned}
\mathfrak{R}_t^{(\ell)}(x,X_t)\triangleq&\left[z_t^{(\ell)}(x)\right]^\top\left[\mathbf{Z}_t^{(\ell)+}(x)\mathbf{D}_{t+1}^{(\ell)}(x)+\mathbf{Z}_t^{(\ell)-}(x)\mathbf{D}_t^{(\ell)}(x)\right]\\
&\frac{1}{m}\langle o_{t+1}^{(\ell-1)}(x)-o_t^{(\ell-1)}(x),o_t^{(\ell-1)}(X_t)\rangle\mathbf{D}_t^{(\ell)}(X_t)z_t^{(\ell)}(X_t).
\end{aligned}\tag{165}$$

Intuitively, $\mathfrak{B}_t^{(\ell)}$ captures the error from the change in activation pattern of the ℓ-th hidden layer and $\mathfrak{R}_t^{(\ell)}$ captures the error from the change of the output $o_t^{(\ell-1)}$.

Plugging (163) back into (161), we have

$$\begin{aligned}
f(x;\mathbf{W}(t+1))-f(x;\mathbf{W}(t))\le&\eta_t\left(\Delta_t(X_t)+u_t\right)\sum_{\ell=1}^{L}\left(H_t^{(\ell)}(x,X_t)+\mathfrak{B}_t^{(\ell)}(x,X_t)+\mathfrak{R}_t^{(\ell)}(x,X_t)\right)\\
&+\sum_{\ell=1}^{L-1}\mathfrak{A}_t^{(\ell)}(x).
\end{aligned}\tag{166}$$

Recall the definition of ϵ_t from (152). For any $x \in \mathbb{S}^{d-1}$, we have

$$\begin{aligned}\epsilon_t(x) &= f(x;\mathbf{W}(t)) - f(x;\mathbf{W}(t+1)) + \eta_t H_t(x,X_t)(\Delta_t(X_t)+u_t)\\ &\geq -\eta_t\left(\Delta_t(X_t)+u_t\right)\sum_{\ell=1}^{L}\left(\mathfrak{B}_t^{(\ell)}(x,X_t)+\mathfrak{R}_t^{(\ell)}(x,X_t)\right) - \sum_{\ell=1}^{L-1}\mathfrak{A}_t^{(\ell)}.\end{aligned}$$

To bound ϵ_t, it remains to bound $\mathfrak{A}_t^{(\ell)}$, $\mathfrak{B}_t^{(\ell)}$ and $\mathfrak{R}_t^{(\ell)}$.

Here, we claim that with probability at least $1-\exp\left(-\Omega(C_0^{-L}m^{1/36})\right)$ over the randomness of the weight $\mathbf{W}(0)$ and the outer weight a, for any sample path $\{(X_s,y_s)\}_{s=0}^{t-1}$,

$$\sup_x |\mathfrak{A}_t^{(\ell)}(x)| = O\left(\frac{\eta_t \ell C_1^L}{m^{1/36}}\left|\Delta_t(X_t)+u_t\right|\right), \tag{167}$$

$$\sup_{x,x'}\left|\mathfrak{B}_t^{(\ell)}(x,x')\right| = O\left(\frac{C_2^{2L}}{m^{1/36}}\right) \tag{168}$$

$$\sup_{x,x'}|\mathfrak{R}_t^{(\ell)}(x,x')| = O\left(\frac{C_3^L}{m^{1/6}}\right). \tag{169}$$

With the above claims, we have with probability at least $1-\exp\left(-\Omega(C_0^{-L}m^{1/36})\right)$, for any x and sample path $\{(X_s,y_s)\}_{s=0}^{t-1}$,

$$\epsilon_t(x) \geq -C_5\eta_t\left|\Delta_t(X_t)+u_t\right|\frac{LC_4^{2L}}{m^{1/36}},$$

for some constant C_4 and C_5.

Analogously, we get with probability at least $1-\exp\left(-\Omega(C_0^{-L}m^{1/36})\right)$,

$$\epsilon_t(x) \leq C_5\eta_t|\Delta_t(X_t)+u_t|\frac{LC_4^{2L}}{m^{1/36}}.$$

As a result, we have

$$\begin{aligned}\mathbb{E}\left[\|\epsilon_t\|_2\,|\mathbf{W}(0),a\right] &\leq \sqrt{\mathbb{E}\left[\|\epsilon_t\|_2^2\,\middle|\mathbf{W}(0),a\right]} = \sqrt{\mathbb{E}_{(X_s,u_s)_{s=0}^t}\left[\mathbb{E}_X\left[\epsilon_t^2(X)\right]\middle|\mathbf{W}(0),a\right]}\\ &\leq \sqrt{\mathbb{E}_{(X_s,u_s)_{s=0}^t}\left[\sup_x \epsilon_t^2(x)\middle|\mathbf{W}(0),a\right]}\\ &= O\left(\frac{\eta_t LC^L}{m^{1/36}}\right)\sqrt{\mathbb{E}_{(X_s,u_s)_{s=0}^t}\left[\left(\Delta_t(X_t)+u_t\right)^2\middle|\mathbf{W}(0),a\right]} = \frac{\eta_t LC^L\sigma_t}{m^{1/36}},\end{aligned} \tag{170}$$

where the last equality holds since

$$\begin{aligned}\mathbb{E}_{(X_s,u_s)_{s=0}^t}\left[\left(\Delta_t(X_t)+u_t\right)^2\middle|\mathbf{W}(0),a\right] &= \mathbb{E}_{(X_s,u_s)_{s=0}^t}\left[\Delta_t^2(X_t)\middle|\mathbf{W}(0),a\right] + \mathbb{E}\left[u_t^2\right]\\ &= \mathbb{E}_{(X_s,u_s)_{s=0}^{t-1}}\left[\|\Delta_t\|_2^2\,|\mathbf{W}(0),a\right] + \tau^2 = \sigma_t^2.\end{aligned}$$

In the following, we prove (167)–(169). Throughout the remaining of Section D.2, we assume the conclusions of Lemma 10–12 hold which is guaranteed to occur with probability at least $1-\exp\left(-\Omega(C_0^{-L}m^{1/36})\right)$.

D.2.3 Bounding $\mathfrak{A}_t^{(\ell)}$

Recall the definition of $\mathfrak{A}_t^{(\ell)}$ from (159). Here, we bound $\sup_x \left|\mathfrak{A}_t^{(\ell)}(x)\right|$. Fix arbitrary $x \in \mathbb{S}^{d-1}$. Note that

$$|\mathfrak{A}_t^{(\ell)}(x)| \le \underbrace{\left\| \left[z_t^{(\ell+1)}(x)\right]^\top \mathbf{Z}_t^{(\ell+1)+}(x) \left[\mathbf{D}_{t+1}^{(\ell)}(x) - \mathbf{D}_t^{(\ell)}(x)\right] \frac{1}{\sqrt{m}} \mathbf{W}^{(\ell+1)}(t) \right\|_2}_{\text{(I)}} \underbrace{\left\| o_{t+1}^{(\ell)}(x) - o_t^{(\ell)}(x) \right\|_2}_{\text{(II)}}. \tag{171}$$

We first bound (II). Note that the change of $o_t^{(\ell)}$ comes from the change of $\mathbf{W}(t)$. Intuitively, since $\mathbf{W}(t)$ does not change much by Lemma 10, we expect that o_t does not change much. By Lipschitz property of ReLU function and the triangle inequality, we obtain the following layer-wise recursive relation of $\left\| o_{t+1}^{(\ell)}(x) - o_t^{(\ell)}(x) \right\|_2$:

$$\begin{aligned}
\left\| o_{t+1}^{(\ell)}(x) - o_t^{(\ell)}(x) \right\|_2 &= \frac{1}{\sqrt{m}} \left\| \sigma\left(\mathbf{W}^{(\ell)}(t+1) o_{t+1}^{(\ell-1)}(x)\right) - \sigma\left(\mathbf{W}^{(\ell)}(t) o_t^{(\ell-1)}(x)\right) \right\|_2 \\
&\le \frac{1}{\sqrt{m}} \left\| \mathbf{W}^{(\ell)}(t+1) o_{t+1}^{(\ell-1)}(x) - \mathbf{W}^{(\ell)}(t) o_t^{(\ell-1)}(x) \right\|_2 \\
&\le \frac{1}{\sqrt{m}} \left\| \left(\mathbf{W}^{(\ell)}(t+1) - \mathbf{W}^{(\ell)}(t)\right) \right\|_2 \left\| o_{t+1}^{(\ell-1)}(x) \right\|_2 \\
&\quad + \frac{1}{\sqrt{m}} \left\| \mathbf{W}^{(\ell)}(t+1) \right\|_2 \left\| o_{t+1}^{(\ell-1)}(x) - o_t^{(\ell-1)}(x) \right\|_2.
\end{aligned}$$

Since $o_{t+1}^{(0)}(x) - o_t^{(0)}(x) = 0$, by recursively applying the above inequality, we have

$$\left\| o_{t+1}^{(\ell)}(x) - o_t^{(\ell)}(x) \right\|_2 \le \frac{1}{\sqrt{m}} \sum_{s=1}^{\ell} \left\| \prod_{r=s+1}^{\ell} \frac{1}{\sqrt{m}} \mathbf{W}^{(s)}(t+1) \right\|_2 \left\| \mathbf{W}^{(s)}(t+1) - \mathbf{W}^{(s)}(t) \right\|_2 \left\| o_{t+1}^{(s-1)}(x) \right\|_2.$$

Plugging (126) into the above displayed equation to replace $\left\|\mathbf{W}^{(s)}(t+1) - \mathbf{W}^{(s)}(t)\right\|_2$, we have

$$\begin{aligned}
&\left\| o_{t+1}^{(\ell)}(x) - o_t^{(\ell)}(x) \right\|_2 \\
&\le \frac{1}{\sqrt{m}} \sum_{s=1}^{\ell} \left\| \left(\prod_{r=s+1}^{\ell} \frac{1}{\sqrt{m}} \mathbf{W}^{(s)}(t+1) \right) \right\|_2 \left\| \eta_t \left(\Delta_t(X_t) + u_t\right) \mathbf{V}_{L,t}^{(\ell)}(x) a \left[o_t^{(\ell-1)}(X_t)\right]^\top \right\|_2 \\
&\quad \left\| o_{t+1}^{(s-1)}(x) \right\|_2,
\end{aligned} \tag{172}$$

where $\mathbf{V}_{L,t}^{(\ell)}(x)$ is defined in (8).

From (127), we have $\frac{1}{\sqrt{m}} \left\|\mathbf{W}^{(\ell)}(t+1)\right\|_2 \le C$, and hence,

$$\left\| \mathbf{V}_{L,t}^{(\ell)}(x) \right\|_2 \le C^{L-\ell}/\sqrt{m}. \tag{173}$$

Plugging (127), (129) and (173) into the right hand side of (172), we get

$$\left\|o_{t+1}^{(\ell)}(x)-o_t^{(\ell)}(x)\right\|_2 = O\left(\frac{\ell C^{L+\ell}\eta_t}{\sqrt{m}}\left|\Delta_t(X_t)+u_t\right|\right). \tag{174}$$

Note that although X_t does not explicitly appear on the left hand side of (174), the evolution of $o_t^{(\ell)}(x)$ depends on X_t and u_t through the update of $\mathbf{W}(t)$.

Now we bound (I) on the right hand side of (171). Note that

$$\begin{aligned}
&\left\|\left[z_t^{(\ell+1)}(x)\right]^\top \mathbf{Z}_t^{(\ell+1)+}(x)\left[\mathbf{D}_{t+1}^{(\ell)}(x)-\mathbf{D}_t^{(\ell)}(x)\right]\frac{1}{\sqrt{m}}\mathbf{W}^{(\ell+1)}(t)\right\|_2\\
&\le \left\|\left[z_t^{(\ell+1)}(x)\right]^\top \mathbf{Z}_t^{(\ell+1)+}(x)\left[\mathbf{D}_{t+1}^{(\ell)}(x)-\mathbf{D}_t^{(\ell)}(x)\right]\right\|_2\left\|\frac{1}{\sqrt{m}}\mathbf{W}^{(\ell+1)}(t)\right\|_2\\
&\overset{(a)}{\le} C\left\|\left[z_t^{(\ell+1)}(x)\right]^\top \mathbf{Z}_t^{(\ell+1)+}(x)\left[\mathbf{D}_{t+1}^{(\ell)}(x)-\mathbf{D}_t^{(\ell)}(x)\right]\right\|_2\\
&\overset{(b)}{\le} C\left\|\left(\mathbf{D}_{t+1}^{(\ell)}(x)-\mathbf{D}_t^{(\ell)}(x)\right)z_t^{(\ell+1)}(x)\right\|_2\\
&\le C\left\|\left(\mathbf{D}_{t+1}^{(\ell)}(x)-\mathbf{D}_0^{(\ell)}(x)\right)z_t^{(\ell+1)}(x)\right\|_2 + C\left\|\left(\mathbf{D}_t^{(\ell)}(x)-\mathbf{D}_0^{(\ell)}(x)\right)z_t^{(\ell+1)}(x)\right\|_2,
\end{aligned} \tag{175}$$

where (a) holds by (127), (b) holds since $\mathbf{Z}_t^{(\ell+1)+}$ is a diagonal matrix with diagonal entries 0 or 1 and the last inequality holds by the triangle inequality.

Here we bound

$$\left\|\left(\mathbf{D}_{t+1}^{(\ell)}(x)-\mathbf{D}_0^{(\ell)}(x)\right)z_t^{(\ell+1)}(x)\right\|_2. \tag{176}$$

By Lemma 11, we know $\mathbf{D}_{t+1}^{(\ell)}(x)-\mathbf{D}_0^{(\ell)}(x)$ has very few non-zero diagonal coordinates. However, if $z_t^{(\ell+1)}(x)$ has large values on those coordinates, (176) can still be large. To show such situation does not occur, we crucially decompose the coordinates of $z_t^{(\ell+1)}(x)$ into $\mathcal{M}$ and $\mathcal{M}^c$ where

$$\mathcal{M}=\left\{i\in[m]:\left|\left[z_t^{(\ell+1)}(x)\right]_i\right|<2m^{1/36}\right\}.$$

For coordinate $i\in\mathcal{M}^c$, since $\left|\left[z_t^{(\ell+1)}(x)\right]_i\right|\ge 2m^{1/36}$ and $\sup_x\left\|z_0^{(\ell+1)}(x)\right\|_\infty\le m^{1/36}$, we know

$$\left|\left[z_t^{(\ell+1)}(x)\right]_i\right|\le 2\left|\left[z_t^{(\ell+1)}(x)\right]_i-\left[z_0^{(\ell+1)}(x)\right]_i\right|.$$

Intuitively, the above displayed equation says that for coordinate i of $z_t^{(\ell+1)}(x)$ with large absolute value, since the initial value $\left|\left[z_0^{(\ell+1)}(x)\right]_i\right|$ is small, the magnitude of $\left[z_t^{(\ell+1)}(x)\right]_i$ is of the same order of its deviation from the initial value. With the bound on $\left\|z_t^{(\ell+1)}(x)-z_0^{(\ell+1)}(x)\right\|_2$ in (49) from Lemma 12, we are able to control the contribution of coordinates in $\mathcal{M}^c$ on (176) as follows:

$$\begin{aligned}
\sum_{j\in\mathcal{M}^c}\left[\left(\mathbf{D}_{t+1}^{(\ell)}(x)-\mathbf{D}_0^{(\ell)}(x)\right)z_t^{(\ell+1)}(x)\right]_j^2 &\le \sum_{j\in\mathcal{M}^c}\left[z_t^{(\ell+1)}(x)\right]_i^2 \le 4\left\|z_t^{(\ell+1)}(x)-z_0^{(\ell+1)}(x)\right\|_2^2\\
&= O(C_1^{4L-2\ell-2}m^{17/18}),
\end{aligned} \tag{177}$$

for some constant C_1.

Next, we show the contribution on (176) from $\mathcal{M}$ is small. This is true since all coordinates in $\mathcal{M}$ have small values and the number of coordinates having nonzero $\mathbf{D}_{t+1}^{(\ell)}(x) - \mathbf{D}_0^{(\ell)}(x)$ is small. In particular, we have

$$\begin{aligned}
&\sum_{j\in\mathcal{M}} \left[\left(\mathbf{D}_{t+1}^{(\ell)}(x) - \mathbf{D}_0^{(\ell)}(x)\right) z_t^{(\ell+1)}(x)\right]_j^2 \\
&\le 4m^{1/18} \sum_{j\in\mathcal{M}} \left(\mathbf{1}_{\left\{\langle w_j^{(\ell)}(t+1), o_t^{(\ell-1)}(x)\rangle \ge 0\right\}} - \mathbf{1}_{\left\{\langle w_j^{(\ell)}(0), o_0^{(\ell-1)}(x)\rangle \ge 0\right\}}\right)^2 \\
&\le 4m^{1/18} \sup_x S_{t+1}^{(\ell)}(x) = O(C_2^\ell m^{17/18}),
\end{aligned} \tag{178}$$

for some constant C_2 where the last equality holds by (47) from Lemma 11.

Combining (177) and (178), we have

$$\left\|\left(\mathbf{D}_{t+1}^{(\ell)}(x) - \mathbf{D}_0^{(\ell)}(x)\right) z_t^{(\ell+1)}(x)\right\|_2 = O(C_3^{2L-\ell-1} m^{17/36}) \tag{179}$$

for some constant C_3.

Similarly, we can get $\left\|\left(\mathbf{D}_t^{(\ell)}(x) - \mathbf{D}_0^{(\ell)}(x)\right) z_t^{(\ell+1)}(x)\right\|_2 = O(C_3^{2L-\ell-1} m^{17/36})$.

Plugging the above bound on $\left\|\left(\mathbf{D}_t^{(\ell)}(x) - \mathbf{D}_0^{(\ell)}(x)\right) z_t^{(\ell+1)}(x)\right\|_2$ and (179) into (175), we get

$$\begin{aligned}
&\left\|\left[z_t^{(\ell+1)}(x)\right]^\top \mathbf{Z}_t^{(\ell+1)+}(x) \left[\mathbf{D}_{t+1}^{(\ell)}(x) - \mathbf{D}_t^{(\ell)}(x)\right] \frac{1}{\sqrt{m}} \mathbf{W}^{(\ell+1)}(t)\right\|_2 \\
&= C \left\|\left(\mathbf{D}_{t+1}^{(\ell)}(x) - \mathbf{D}_t^{(\ell)}(x)\right) z_t^{(\ell+1)}(x)\right\|_2 \\
&= O(C_4^L m^{17/36})
\end{aligned} \tag{180}$$

for some constant C_4.

Combining (174) and (180), we have for any $x \in \mathbb{S}^{d-1}$,

$$|\mathfrak{A}_t^{(\ell)}(x)| = O\left(\frac{\eta_t \ell C^L}{m^{1/36}} |\Delta_t(X_t) + u_t|\right),$$

for some constant C.

D.2.4 Bounding $\mathfrak{B}_t^{(\ell)}$ and $\mathfrak{R}_t^{(\ell)}$

As is mentioned in Section D.2.2, $\mathfrak{B}_t^{(\ell)}$ captures the error caused by the change of activation pattern. To bound $\left|\mathfrak{B}_t^{(\ell)}(x, x')\right|$, we crucially apply (180) which bounds $\left\|\left(\mathbf{D}_{t+1}^{(\ell)}(x) - \mathbf{D}_0^{(\ell)}(x)\right) z_t^{(\ell)}(x)\right\|_2$. To bound $\mathfrak{R}_t^{(\ell)}$ which captures the error from the change of the output $o_t^{(\ell-1)}$, we apply (45) from Lemma 10 which bounds the deviation of $o_t^{(\ell-1)}$.

Bounding $\left|\mathfrak{B}_t^{(\ell)}\right|$: Recall that

$$\mathfrak{B}_t^{(\ell)}(x,x') = \left[z_t^{(\ell)}(x)\right]^\top \mathbf{Z}_t^{(\ell)+}(x)\left(\mathbf{D}_{t+1}^{(\ell)}(x) - \mathbf{D}_t^{(\ell)}(x)\right) \frac{1}{m}\langle o_t^{(\ell-1)}(x), o_t^{(\ell-1)}(x')\rangle \mathbf{D}_t^{(\ell)}(x') z_t^{(\ell)}(x').$$

Fix any x and $x' \in \mathbb{S}^{d-1}$. By Cauchy-Schwartz inequality, we have

$$\begin{aligned}\left|\mathfrak{B}_t^{(\ell)}(x,x')\right| &\le \left\|\left[z_t^{(\ell)}(x)\right]^\top \mathbf{Z}_t^{(\ell)+}(x)\left(\mathbf{D}_{t+1}^{(\ell)}(x) - \mathbf{D}_t^{(\ell)}(x)\right)\right\|_2 \\ &\quad \left\|\frac{1}{m}\langle o_t^{(\ell-1)}(x), o_t^{(\ell-1)}(x')\rangle \mathbf{D}_t^{(\ell)}(x') z_t^{(\ell)}(x')\right\|_2 \\ &= O\left(C^L m^{17/36}\right)\left\|\frac{1}{m}\langle o_t^{(\ell-1)}(x), o_t^{(\ell-1)}(x')\rangle \mathbf{D}_t^{(\ell)}(x') z_t^{(\ell)}(x')\right\|_2\end{aligned}$$

where the last equality holds by (180).

Moreover, applying Cauchy-Schwartz inequality again, we get

$$\begin{aligned}\left\|\frac{1}{m}\langle o_t^{(\ell-1)}(x), o_t^{(\ell-1)}(x')\rangle \mathbf{D}_t^{(\ell)}(X_t) z_t^{(\ell)}(x')\right\|_2 &\le \frac{1}{m}\left\|o_t^{(\ell-1)}(x)\right\|_2 \left\|o_t^{(\ell-1)}(x')\right\|_2 \left\|z_t^{(\ell)}(x')\right\|_2 \\ &= O\left(\frac{C_1^L}{\sqrt{m}}\right).\end{aligned}$$

where the last equality holds by (129) and (146).

As a result, for any $x, x' \in \mathbb{S}^{d-1}$, we have

$$\left|\mathfrak{B}_t^{(\ell)}(x,x')\right| \le O\left(\frac{C_2^{2L}}{m^{1/36}}\right)$$

for some constant C_2.

Bounding $\left|\mathfrak{R}_t^{(\ell)}\right|$: Recall that

$$\mathfrak{R}_t^{(\ell)}(x,x') = \left[z_t^{(\ell)}(x)\right]^\top \left[\mathbf{Z}_t^{(\ell)+}(x)\mathbf{D}_{t+1}^{(\ell)}(x) + \mathbf{Z}_t^{(\ell)-}(x)\mathbf{D}_t^{(\ell)}(x)\right] \frac{1}{m}\langle o_{t+1}^{(\ell-1)}(x) - o_t^{(\ell-1)}(x), o_t^{(\ell-1)}(x')\rangle \mathbf{D}_t^{(\ell)}(x') z_t^{(\ell)}(x').$$

By Cauchy-Schwartz inequality, we have

$$\begin{aligned}|\mathfrak{R}_t^{(\ell)}(x,x')| &\le \left\|\left[\mathbf{Z}_t^{(\ell)+}(x)\mathbf{D}_{t+1}^{(\ell)}(x) + \mathbf{Z}_t^{(\ell)-}(x)\mathbf{D}_t^{(\ell)}(x)\right] z_t^{(\ell)}(x)\right\|_2 \\ &\quad \left\|\frac{1}{m}\langle o_{t+1}^{(\ell-1)}(x) - o_t^{(\ell-1)}(x), o_t^{(\ell-1)}(x')\rangle \mathbf{D}_t^{(\ell)}(x') z_t^{(\ell)}(x')\right\|_2.\end{aligned}$$

By (146) we have

$$\left\|\left[z_t^{(\ell)}(x)\right]^\top \left[\mathbf{Z}_t^{(\ell)+}(x)\mathbf{D}_{t+1}^{(\ell)}(x) + \mathbf{Z}_t^{(\ell)-}(x)\mathbf{D}_t^{(\ell)}(x)\right]\right\|_2 \le \left\|z_t^{(\ell)}(x)\right\|_2 = O(C^{L-\ell}\sqrt{m}).$$

Further note that

$$\begin{aligned}
&\left\|\frac{1}{m}\langle o_{t+1}^{(\ell-1)}(x)-o_t^{(\ell-1)}(x),o_t^{(\ell-1)}(x')\rangle\mathbf{D}_t^{(\ell)}(x')z_t^{(\ell)}(x')\right\|_2\\
&\le\frac{1}{m}\left\|o_{t+1}^{(\ell-1)}(x)-o_t^{(\ell-1)}(x)\right\|_2\left\|o_t^{(\ell-1)}(x')\right\|_2\left\|z_t^{(\ell)}(x')\right\|_2\\
&\le\frac{C^{2L-1}}{\sqrt{m}}\left(\left\|o_{t+1}^{(\ell-1)}(x)-o_0^{(\ell-1)}(x)-o_t^{(\ell-1)}(x)+o_0^{(\ell-1)}(x)\right\|_2\right)\\
&\le\frac{C^{2L-1}}{\sqrt{m}}\left(\left\|o_{t+1}^{(\ell-1)}(x)-o_0^{(\ell-1)}(x)\right\|_2+\left\|o_t^{(\ell-1)}(x)-o_0^{(\ell-1)}(x)\right\|_2\right)\\
&=O\left(\frac{C^{2L+\ell-2}}{m^{1/2+1/6}}\right).
\end{aligned}$$

where the second inequality holds by (146) and (129), the third one holds by the triangle inequality, and the last equality holds by (45) from Lemma 10.

As a result, for any $x,x'\in\mathbb{S}^{d-1}$,

$$|\mathfrak{R}_t^{(\ell)}(x,x')|=O\left(\frac{C_3^L}{m^{1/6}}\right),$$

for some constant C_3.

D.3 Proof of Lemma 16

Throughout the proof, we assume the conclusions of Lemma 15 holds. For the ease of presentation, we use $\mathbb{E}[\cdot]$ to denote the conditional expectation $\mathbb{E}[|\mathbf{W}(0),a]$.

Recall the definition of v_t from (54). We first show

$$\mathbb{E}\left[\left\|\sum_{s=0}^{t}\prod_{r=s+1}^{t}\mathsf{Q}_r\circ v_s\right\|_2^2\right]\le\sum_{s=0}^{t}\mathbb{E}\left[\|v_s\|_2^2\right]. \tag{181}$$

For notation simplicity, denote F_t as the filtration of $\{X_1,\cdots,X_t\}$. Let $q_t=\sum_{r=0}^{t}\prod_{i=r+1}^{t}\mathsf{Q}_i\circ v_r$ and $h_t=\mathsf{Q}_t\circ q_{t-1}$. Thus, $q_t=v_t+h_t$. Then

$$\mathbb{E}\left[\|q_t\|_2^2\right]=\mathbb{E}\left[\|v_t+h_t\|_2^2\right]\overset{(a)}{=}\mathbb{E}\left[\|v_t\|_2^2\right]+\mathbb{E}\left[\|h_t\|_2^2\right]\overset{(b)}{\le}\mathbb{E}\left[\|v_t\|_2^2\right]+\mathbb{E}\left[\|q_{t-1}\|_2^2\right],$$

where (a) uses the fact that $\mathbb{E}[\langle v_t,h_t\rangle]=\mathbb{E}[\mathbb{E}[\langle v_t,h_t\rangle|F_{t-1}]]=\mathbb{E}[\langle\mathbb{E}[v_t|F_{t-1}],h_t\rangle]=0$; (b) follows from $\|\mathsf{Q}_t\|_2\le1$ by Lemma 14.

Recursively applying the last displayed equation yields that

$$\mathbb{E}\left[\|q_t\|_2^2\right]\le\sum_{r=0}^{t}\mathbb{E}\left[\|v_r\|_2^2\right].$$

Next, we bound $\mathbb{E}\left[\|v_s\|_2^2\right]$. Recall from (56) that $\sigma_s^2 = \mathbb{E}\left[\|\Delta_s\|_2^2\right] + \tau^2$. Note that

$$\begin{aligned}\mathbb{E}\left[\|v_s\|_2^2\right] &= \eta_s^2\mathbb{E}\left[\left(H_s(x,X_s)\left(\Delta_s(X_s)+u_s\right)^2 - \mathbb{E}_{X_s}\left[H_s(x,X_s)\Delta_s(X_s)\right]\right)^2\right]\\ &= \eta_s^2\mathbb{E}\left[H_s^2(x,X_s)\left(\Delta_s(X_s)+u_s\right)^2\right] - \eta_s^2\left(\mathbb{E}_{X_s}\left[H_s(x,X_s)\Delta_s(X_s)\right]\right)^2\\ &\le \eta_s^2L^2\left(\mathbb{E}\left[\|\Delta_s\|_2^2\right]+\tau^2\right) = \eta_s^2\frac{4L^2}{9}\sigma_s^2, \end{aligned} \tag{182}$$

where the inequality holds by (151) that gives $\|H_t\|_2 \le \|H_t\|_\infty \le \frac{2L}{3}$.

Therefore, to control $\mathbb{E}\left[\|v_s\|_2^2\right]$, we need to bound σ_t^2. We now claim that

$$\sigma_{t+1}^2 \le \prod_{s=0}^{t}(1+\frac{\sqrt{44}L}{9}\eta_t)^2\sigma_0^2,$$

when $m = \Omega(\exp(L^2))$ and $\eta_t \le \frac{3}{2L}$ for all t.

Given the claim, we have

$$\begin{aligned}\eta_r\sigma_r &\le \frac{\theta}{r+1}\prod_{k=0}^{r-1}\left(1+\frac{\sqrt{44}L\theta}{9(k+1)}\right)\sigma_0\\ &\le \frac{\theta}{r+1}\exp\left(\frac{\sqrt{44}L\theta}{9}\left(\log(r+1)+1\right)\right)\sigma_0\\ &\le \theta\left(r+1\right)^{\sqrt{44}L\theta/9-1}e^{\sqrt{44}L\theta/9}\sigma_0. \end{aligned} \tag{183}$$

Combining (183) and (182) into (181), we have

$$\begin{aligned}\mathbb{E}\left[\left\|\sum_{s=0}^{t}\prod_{r=s+1}^{t}\mathsf{Q}_r\circ v_s\right\|_2^2\right] &\le L^2\sum_{s=0}^{t}\eta_s^2\sigma_s^2\\ &\le L^2\sum_{r=0}^{t}\theta^2(r+1)^{2\sqrt{44}\theta L/9-2}e^{2\sqrt{44}L\theta/9}\sigma_0^2\\ &\le L^2\theta^2e^{2\sqrt{44}L\theta/9}\sigma_0^2\left(\frac{1}{1-2\sqrt{44}L\theta/9}+1\right) = c_2^2\sigma_0^2.\end{aligned}$$

By Cauchy-Schwartz inequality, we have

$$\mathbb{E}\left[\left\|\sum_{s=0}^{t}\prod_{r=s+1}^{t}\mathsf{Q}_r\circ v_s\right\|_2\right] \le \sqrt{\mathbb{E}\left[\left\|\sum_{s=0}^{t}\prod_{r=s+1}^{t}\mathsf{Q}_r\circ v_s\right\|_2^2\middle|\mathbf{W}(0),a\right]} = c_2\sigma_0,$$

which completes the proof.

Now we prove the claim. Recall $\Delta_{t+1} = \mathsf{Q}_t\cdot\Delta_t - v_t + \epsilon_t$. Therefore,

$$\begin{aligned}\|\Delta_{t+1}\|_2^2 &= \|\mathsf{Q}_t\cdot\Delta_t - v_t + \epsilon_t\|_2^2\\ &= \|\mathsf{Q}_t\cdot\Delta_t\|_2^2 + \|v_t\|_2^2 + \|\epsilon_t\|_2^2 - 2\langle\mathsf{Q}_t\cdot\Delta_t, v_t\rangle - 2\langle v_t,\epsilon_t\rangle + 2\langle Q_t\cdot\Delta_t,\epsilon_t\rangle\\ &\le \|\Delta_t\|_2^2 + \|v_t\|_2^2 + \|\epsilon_t\|_2^2 + 2\|\Delta_t\|_2\|v_t\|_2 + 2\|v_t\|_2\|\epsilon_t\|_2 + 2\|\Delta_t\|_2\|\epsilon_t\|_2, \end{aligned} \tag{184}$$

where the last inequality holds by $\|\mathsf{Q}_t\|_2 \leq 1$ whenever $\eta_t \leq 2/L$ and Cauchy-Schwartz inequality.

From (170), for m satisfying (41), we can get

$$\mathbb{E}\left[\|\epsilon_t\|_2^2\right] \leq O\left(\frac{L^2 C^L \eta_t^2}{m^{1/18}}\sigma_t^2\right) \leq \frac{L^2\eta_t^2}{81}\sigma_t^2. \tag{185}$$

Taking conditional expectation on both hand sides of (184), we have

$$\begin{aligned}
\sigma_{t+1}^2 &\leq \sigma_t^2 + \frac{4L^2\eta_t^2}{9}\sigma_t^2 + \frac{L^2\eta_t^2}{81}\sigma_t^2 + 2\mathbb{E}\left[\|\Delta_t\|_2 \|v_t\|_2\right] + 2\mathbb{E}\left[\|v_t\|_2 \|\epsilon_t\|_2\right] + 2\mathbb{E}\left[\|\Delta_t\|_2 \|\epsilon_t\|_2\right] \\
&\leq \left(1 + \frac{37L^2\eta_t^2}{81}\right)\sigma_t^2 + 2\sqrt{\mathbb{E}\left[\|\Delta_t\|_2^2\right]}\sqrt{\mathbb{E}\left[\|v_t\|_2^2\right]} \\
&+ 2\sqrt{\mathbb{E}\left[\|v_t\|_2^2\right]}\sqrt{\mathbb{E}\left[\|\epsilon_t\|_2^2\right]} + 2\sqrt{\mathbb{E}\left[\|\Delta_t\|_2^2\right]}\sqrt{\mathbb{E}\left[\left\|\epsilon_t^2\right\|_2\right]} \\
&\leq \left(1 + \frac{37L^2\eta_t^2}{81}\right)\sigma_t^2 + \frac{2}{9}\eta_t\sigma_t^2 + \frac{2L^2\eta_t^2}{27}\sigma_t^2 + \frac{2L\eta_t}{9}\sigma_t^2 \\
&= (1 + \frac{\sqrt{44}L}{9}\eta_t)^2\sigma_t^2
\end{aligned}$$

where the first and the third inequalities hold by (182) and (185) for m satisfying (41) and the second inequality holds by Cauchy-Schwartz inequality.

Appendix E. Proof of Corollary 4

We first show a key intermediate step to prove Corollary 4.

Define the space of homogeneous harmonic polynomials of order ℓ on the sphere as

$$\mathcal{H}_\ell = \left\{ P : \mathbb{S}^{d-1} \to \mathbb{R} : P(x) = \sum_{|\alpha|=\ell} c_\alpha x^\alpha,\ \Delta P = 0 \right\}$$

where $x^\alpha = x_1^{\alpha_1}\cdots x_d^{\alpha_d}$, $|\alpha| = \sum_{i=1}^d \alpha_i$, $c_\alpha \in \mathbb{R}$ and $\Delta = \sum_{i=1}^d \frac{\partial^2}{\partial x_i^2}$ is the Laplacian operator.

Denote for all $\ell \geq 0$, $\{Y_{\ell,i}\}_{i=1}^{N_\ell}$ as some orthonormal basis of $\mathcal{H}_\ell$ where N_ℓ is the dimension of $\mathcal{H}_\ell$, i.e., $\langle Y_{\ell,i}, Y_{\ell,j}\rangle = 0$ for $i \neq j$. Moreover, from Dai and Xu (2013, Theorem 1.1.2) for $\ell \neq \ell'$, $\mathcal{H}_\ell$ and $\mathcal{H}_{\ell'}$ are orthogonal. Hence, $\{Y_{\ell,i}\}$ are orthogonal across different ℓ as well.

We now derive in Theorem 36 an expansion for functions with the form $\mathcal{K}(x,y) = h(\langle x, y\rangle),\ x, y \in \mathbb{S}^{d-1}, d \geq 3$ in terms of $\{Y_{\ell,i}\}, 1 \leq i \leq N_\ell, \ell \geq 0$. A similar result is obtained in Su and Yang (2019) without a full proof. We provide a proof in Appendix E.1 for completeness.

Theorem 36 *Suppose the function $\mathcal{K}$ has the form $\mathcal{K}(x,y) = h(\langle x, y\rangle)$ where h is analytic on $[-1,1]$, $x, y \in \mathbb{S}^{d-1}$ and $d \geq 3$. Then*

$$\mathcal{K}(x,y) = \sum_{\ell\geq 0} \beta_\ell(h) \sum_{i=1}^{N_\ell} Y_{\ell,i}(x) Y_{\ell,i}(y)$$

where

$$\beta_\ell(h) = \frac{d-2}{2} \sum_{m=0}^{\infty} \frac{h_{\ell+2m}}{2^{\ell+2m} m! (\frac{d-2}{2})_{\ell+m+1}} \tag{186}$$

with $h_{\ell+2m}$ is the $(\ell+2m)$-th derivative of h at 0 and $(\cdot)_n$ is the Pochhammer symbol recursively defined as $(a)_0 = 1$, $(a)_k = (a+k-1)(a)_{k-1}$ for $k \geq 1$.

Remark 37 *The case $d = 2$ can be analyzed using Fourier analysis. Since this is not of particular interest in our study, we do not provide the analysis here. One can refer to (Dai and Xu, 2013, Section 1.6) if interested.*

Proof [Proof of Corollary 4] From (Wang, 2010, Theorem 7.4), we know the polynomial of degree ℓ^* can be projected onto the direct sum of the spaces of homogeneous harmonic polynomials up to degree $\ell^* + 1$. Now we claim Φ can be expanded in the space of homogeneous harmonic polynomials. With the claim, we have $\mathcal{R}(f^*, \ell^* + 1) = 0$ which completes the proof.

It remains to prove the claim. Recall the definition of $\Phi^{(\ell)}$ in (21). Here, we show $\Phi^{(\ell)}(x, x')$ is analytic and can be viewed as a function of $\langle x, x' \rangle$ only by analyzing $\mathbb{E}\left[\sigma(U^{(\ell)}(x))\sigma(U^{(\ell)}(x')\right]$ and $q_L^{(\ell)}(x, x')$.

We begin with analyzing $\mathbb{E}\left[\sigma(U^{(\ell)}(x))\sigma(U^{(\ell)}(x')\right]$. By (15), we get $(U^{(\ell)}(x), U^{(\ell)}(x'))$ depends on $\mathrm{Cov}\left(\sigma(U^{(\ell)}(x)), \sigma(U^{(\ell)}(x'))\right)$. Since $\Sigma^{(0)}$ only depends on $\langle x, x' \rangle$, we know the joint distribution of $(U^{(1)}(x), U^{(1)}(x'))$ only depends on $\langle x, x' \rangle$. Hence, $\mathrm{Cov}\left(\sigma(U^{(1)}(x)), \sigma(U^{(1)}(x')\right)$ only depends on $\langle x, x' \rangle$. Following the recursive relationship of $U^{(\ell)}$, we get the joint distribution of $(U^{(\ell)}(x), U^{(\ell)}(x'))$ for all $\ell \geq 1$ only depends on $\langle x, x' \rangle$. Hence, $\mathbb{E}\left[\sigma(U^{(\ell)}(x))\sigma(U^{(\ell)}(x')\right]$ only depends on $\langle x, x' \rangle$. Note that a product of two ReLU functions is analytic. By Fubini Theorem and Lebniz integral rule, we know $\mathbb{E}\left[\sigma(U^{(\ell)}(x))\sigma(U^{(\ell)}(x')\right]$ is analytic.

Next, we study $q_L^{(\ell)}(x, x')$ which is defined in (19). We have shown the numerator of $\rho^{(\ell)}(x, x')$ only depends on $\langle x, x' \rangle$. By (38), we know the denominator of $\rho^{(k)}(x, x')$ is some constant independent of x and x'. Therefore, $\rho^{(k)}(x, x')$ and hence $q_L^{(\ell)}(x, x')$ only depends on $\langle x, x' \rangle$. Since a composition of analytic functions is analytic and arccos function is analytic, we know $q_L^{(\ell)}$ is analytic.

Since for any ℓ, $\Phi^{(\ell)}$ is analytic and can be viewed as a function of $\langle x, x' \rangle$, we know $\Phi = \sum_{\ell=1}^{L} \Phi^{(\ell)}$ is also analytic and is a function of $\langle x, x' \rangle$. ■

E.1 Proof of Theorem 36

We begin with a key result that will be used in the proof of Theorem 36.

Proposition 38 *(Cantero and Iserles, 2012, Theorem 2, eq (2.1)) Let h be analytic in $[-1, 1]$. Letting $h_n = h^{(n)}(0)$ be n-th order derivative, then for any $\alpha > -1, \alpha \neq -\frac{1}{2}$,*

$$h(x) = \sum_{n=0}^{\infty} \widetilde{h}_n C_n^{\alpha+1/2}(x), \ x \in [-1, 1] \tag{187}$$

where

$$C_n^{\alpha+1/2}(x) = \frac{(2\alpha+1)_n}{n!}\sum_{k=0}^{n}(-1)^k\binom{n}{k}\frac{(n+2\alpha+1)_k}{(\alpha+1)_k}\left(\frac{1-x}{2}\right)^k,$$

is the Gegenbauer polynomial, and

$$\widetilde{h}_n = (\alpha+n+1/2)\sum_{m=0}^{\infty}\frac{h_{n+2m}}{2^{n+2m}m!(\alpha+1/2)_{n+m+1}}, \tag{188}$$

with $h_{n+2m} = h^{(n+2m)}(0)$*, the* $n+2m$*-th derivative of* h *at* 0*.*

Remark 39 *Gegenbauer polynomials are orthogonal across different* n*, i.e., for* $m \neq n$*,* $d \geq 3$ *and any fixed* $y \in \mathbb{S}^{d-1}$*,* $\left\langle C_n^{\frac{d-2}{2}}(\langle\cdot, y\rangle), C_m^{\frac{d-2}{2}}(\langle\cdot, y\rangle)\right\rangle_{\mathbb{S}^{d-1}} = 0$*. The proof is based on the orthogonality of* $\mathcal{H}_\ell$*. One can check Dai and Xu (2013, Corollary 2.8) for a detailed proof.*

The form of $\beta_\ell(h)$ in (186) depends on the specific function h. For the ease of presentation, we abbreviate $\beta_\ell(h)$ as β_ℓ. Now we proceed to the proof of Theorem 36.

Proof [Proof of Theorem 36] From Dai and Xu (2013, eq(2.8)), we know for any $l \geq 0$,

$$\frac{\ell+\lambda}{\lambda}C_\ell^\lambda(\langle x, y\rangle) = \sum_{i=1}^{N_\ell}Y_{\ell,i}(x)Y_{\ell,i}(y) \tag{189}$$

where $\lambda = \frac{d-2}{2}$, $x, y \in \mathbb{S}^{d-1}$.

Plugging (189) into (187) and note that $\alpha + 1/2 = \lambda = \frac{d-2}{2}$, we get

$$h(\langle x, y\rangle) = \sum_{\ell\geq 0}\widetilde{h}_\ell\frac{\lambda}{\ell+\lambda}\sum_{i=1}^{N_\ell}Y_{\ell,i}(x)Y_{\ell,i}(y) = \beta_\ell\sum_{i=1}^{N_\ell}Y_{\ell,i}(x)Y_{\ell,i}(y)$$

where

$$\beta_\ell = \widetilde{h}_\ell\frac{\lambda}{\ell+\lambda} = \frac{d-2}{2}\sum_{m=0}^{\infty}\frac{h_{\ell+2m}}{2^{\ell+2m}m!(\frac{d-2}{2})_{\ell+m+1}}.$$

■