Scaling Neural Tangent Kernels via Sketching and Random Features

Amir Zandieh[*]
Max-Planck-Institut für Informatik
azandieh@mpi-inf.mpg.de

Insu Han[*]
Yale University
insu.han@yale.edu

Haim Avron
Tel Aviv University
haimav@tauex.tau.ac.il

Neta Shoham
Tel Aviv University
shohamne@gmail.com

Chaewon Kim
KAIST
chaewonk@kaist.ac.kr

Jinwoo Shin
KAIST
jinwoos@kaist.ac.kr

Abstract

The Neural Tangent Kernel (NTK) characterizes the behavior of infinitely-wide neural networks trained under least squares loss by gradient descent. Recent works also report that NTK regression can outperform finitely-wide neural networks trained on small-scale datasets. However, the computational complexity of kernel methods has limited its use in large-scale learning tasks. To accelerate learning with NTK, we design a near input-sparsity time approximation algorithm for NTK, by sketching the polynomial expansions of arc-cosine kernels: our sketch for the convolutional counterpart of NTK (CNTK) can transform any image using a linear runtime in the number of pixels. Furthermore, we prove a spectral approximation guarantee for the NTK matrix, by combining random features (based on leverage score sampling) of the arc-cosine kernels with a sketching algorithm. We benchmark our methods on various large-scale regression and classification tasks and show that a linear regressor trained on our CNTK features matches the accuracy of exact CNTK on CIFAR-10 dataset while achieving 150× speedup.

1 Introduction

Recent results have shown that over-parameterized Deep Neural Networks (DNNs), generalize surprisingly well. In an effort to understand this phenomena, researchers have studied ultra-wide DNNs and shown that in the infinite width limit, a fully connected DNN trained by gradient descent under least-squares loss is equivalent to kernel regression with respect to the Neural Tangent Kernel (NTK) [5, 11, 22, 28]. This connection has shed light on DNNs' ability to generalize [10, 34] and optimize (train) their parameters efficiently [3, 4, 16]. More recently, Arora et al. [5] proved an analogous equivalence between convolutional DNNs with infinite number of channels and Convolutional NTK (CNTK). Beyond the aforementioned theoretical purposes, several papers have explored the algorithmic use of this kernel. Arora et al. [6] and Geifman et al. [19] showed that NTK based kernel models can outperform trained DNNs (of finite width). Additionally, CNTK kernel regression sets an impressive performance record on CIFAR-10 for kernel methods without trainable kernels [5]. The NTK has also been used in experimental design [39] and predicting training time [43].

However, the NTK-based approaches encounter the computational bottlenecks of kernel learning. In particular, for a dataset of n images $x_1, x_2, \dots x_n \in \mathbb{R}^{d\times d}$, only writing down the CNTK kernel matrix requires $\Omega\left(d^4 \cdot n^2\right)$ operations [5]. Running regression or PCA on the resulting kernel matrix takes additional cubic time in n, which is infeasible in large-scale setups.

[*]Equal contribution.

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

There is a rich literature on kernel approximations for large-scale learning. One of the most popular approaches is the *random features method* which works by randomly sampling the feature space of the kernel function, originally due to the seminal work of Rahimi and Recht [37]. Another popular approach which is developed in linear sketching literature [41], works by designing sketches that can be efficiently applied to the feature space of a kernel without needing to explicitly form the high dimensional feature space. This approach has been successful at designing efficient subspace embeddings for the polynomial kernel [7, 1]. In this paper, we propose solutions for scaling the NTK and CNTK by building on both of these kernel approximations techniques and designing efficient feature maps that approximate the NTK/CNTK evaluation. Consequently, we can simply transform the input dataset to these feature spaces, and then apply fast linear learning methods to approximate the answer of the corresponding nonlinear kernel method efficiently. The performance of such approximate methods is similar or sometimes better than the exact kernel methods due to implicit regularization effects of the approximation algorithms [37, 38, 23].

1.1 Overview of Our Contributions

• One of our results is an efficient random features construction for the NTK. Our starting point is the explicit NTK feature map suggested by Bietti and Mairal [9] based on tensor product of the feature maps of arc-cosine kernels. We obtain our random features, by sampling the feature space of arc-cosine kernels [12]. However, the naïve construction of the features would incur an exponential cost in the depth of the NTK, due to the tensor product of features generated in consecutive layers. We remedy this issue, by utilizing an efficient sketching algorithm for tensor products known as TENSORSRHT [1] which can effectively approximate the tensor products of vectors while preserving their inner products. We provide a rigorous error analysis of the proposed scheme in Theorem 2.

• Our next results are sketching methods for both NTK and CNTK using a runtime that is linearly proportional to the sparsity of the input dataset (or number of pixels of images). Our methods rely on the arc-cosine kernels' feature space defined by their Taylor expansion. By careful truncation of their Taylor series, we approximate the arc-cosine kernels with bounded-degree polynomial kernels. Because the feature space of a polynomial kernel is the tensor product of its input space, its dimensionality is exponential in the degree of the kernel. Fortunately, Ahle et al. [1] have developed a linear sketch known as POLYSKETCH that can reduce the dimensionality of high-degree tensor products very efficiently, therefore, we can sketch the resulting polynomial kernels using this technique. We then combine the transformed features from consecutive layers by further sketching their tensor products. In case of CNTK, we have an extra operation which sketches the direct sum of the features of neighbouring pixels at each layer that precisely corresponds to the convolution operation in CNNs. We carefully analyze the errors introduced by polynomial approximations and various sketching steps in our algorithms and also bound their runtimes in Theorems 1 and 4.

• Furthermore, we improve the arc-cosine random features to spectrally approximate the entire kernel matrix, which is advocated in recent literature for ensuring high approximation quality in downstream tasks [8, 32]. Our construction is based on leverage score sampling, which entertains better convergence bounds [8, 28, 29]. However, computing this distribution is as expensive as solving the kernel methods exactly. We propose a simple distribution that tightly upper bounds the leverage scores of arc-cosine kernels and for further efficiency, use Gibbs sampling to generate random features from our proposed distribution. We provide our spectral approximation guarantee in Theorem 3.

• Finally, we empirically benchmark our proposed methods on various classification/regression tasks and demonstrate that our methods perform similar to or better than exact kernel method with NTK and CNTK while running extremely faster. In particular, we classify CIFAR-10 dataset 150$\times$ faster than exact CNTK and at the same time achieve higher test accuracy.

1.2 Related Works

There has been a long line of work on the correspondence between DNN and kernel machines [26, 30, 35, 18, 42]. Furthermore, there has been many efforts in understanding a variety of NTK properties including optimization [27, 3, 16, 44], generalization [10], loss surface [31], etc.

Novak et al. [35] tried accelerating CNTK computations via Monte Carlo methods by taking the gradient of a randomly initialized CNN with respect to its weights. Although they do not theoretically bound the number of required features, the fully-connected version of this method is analyzed in [5].

Particularly, for the gradient features to approximate the NTK up to ε, the network width needs to be $\Omega(\frac{L^6}{\varepsilon^4}\log\frac{L}{\delta})$, thus, transforming a single vector $x \in \mathbb{R}^d$ requires $\Omega(\frac{L^{13}}{\varepsilon^8}\log^2\frac{L}{\delta} + \frac{L^6}{\varepsilon^4}\log\frac{L}{\delta}\cdot \mathrm{nnz}(x))$ operations, which is slower than our Theorem 1 by a factor of L^3/ε^2. Furthermore, [5] shows that the performance of these random gradients is worse than exact CNTK by a large margin, in practice. More recently, [28] proposed leverage score sampling for the NTK, however, their work is primarily theoretical and suggests no practical way of sampling the features. Another line of work on NTK approximation is an explicit feature map construction via tensor product proposed by Bietti and Mairal [9]. These explicit features can have infinite dimension in general and even if one uses a finite-dimensional approximation to the features, the computational gain of random features will be lost due to expensive tensor product operations.

A popular line of work on kernel approximation problem is based on the Fourier features method [37], which works well for shift-invariant kernels and with some modifications can embed the Gaussian kernel near optimally [8]. Other random feature constructions have been suggested for a variety of kernels, e.g., arc-cosine kernels [12], polynomial kernels [36]. In linear sketching literature, Avron et al. [7] proposed a subspace embedding for the polynomial kernel which was recently extended to general dot product kernels [20]. The runtime of this method, while nearly linear in sparsity of the input dataset, scales exponentially in kernel's degree. Recently, Ahle et al. [1] improved this exponential dependence to polynomial which enabled them to sketch high-degree polynomial kernels and led to near-optimal embeddings for Gaussian kernel. In fact, this sketching technology constitutes one of the main ingredients of our proposed methods. Additionally, combining sketching with leverage score sampling can improve the runtime of the polynomial kernel embeddings [40].

1.3 Preliminaries: POLYSKETCH and TENSORSRHT Transforms

Notations. We use $[n] := \{1, \ldots, n\}$. We denote the tensor (a.k.a. Kronecker) product by $\otimes$ and the element-wise (a.k.a. Hadamard) product of two vectors or matrices by $\odot$. Although tensor products are multidimensional objects, we often associate $x \otimes y$ with a single dimensional vector $(x_1y_1, x_2y_1, \ldots x_my_1, x_1y_2, \ldots x_my_2, \ldots x_my_n)$. For shorthand, we use the notation $x^{\otimes p}$ to denote $\underbrace{x \otimes \ldots \otimes x}_{p \text{ terms}}$, the p-fold self-tensoring of x. Another related operation that we use is the *direct sum* of vectors: $x \oplus y := \begin{bmatrix} x^\top, y^\top \end{bmatrix}^\top$. We need notation for *sub-tensors* of a tensor. For instance, for a 3-dimensional tensor $\boldsymbol{Y} \in \mathbb{R}^{m\times n\times d}$ and every $l \in [d]$, we denote by $\boldsymbol{Y}_{(:,:,l)}$ the $m \times n$ matrix that is defined as $\left[\boldsymbol{Y}_{(:,:,l)}\right]_{i,j} := \boldsymbol{Y}_{i,j,l}$ for $i \in [m], j \in [n]$. For square matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, we write $\boldsymbol{A} \preceq \boldsymbol{B}$ if $\boldsymbol{B} - \boldsymbol{A}$ is positive semi-definite. We also denote $\mathrm{ReLU}(x) = \max(x, 0)$ and consider this element-wise operation when the input is a matrix. We use $\mathrm{nnz}(x)$ to denote the number of nonzero entries in x. Given a positive semidefinite matrix $\boldsymbol{K}$ and $\lambda > 0$, the statistical dimension of $\boldsymbol{K}$ with λ is defined as $s_\lambda(\boldsymbol{K}) := \mathtt{tr}(\boldsymbol{K}(\boldsymbol{K} + \lambda \boldsymbol{I})^{-1})$. For two functions f and g we denote their twofold composition by $f \circ g$, defined as $f \circ g(\alpha) := f(g(\alpha))$.

The TENSORSRHT is a norm-preserving dimensionality reduction that can be applied to the tensor product of two vectors very quickly [1]. This transformation is a generalization of the Subsampled Randomized Hadamard Transform (SRHT) [2] and can be computed in near linear time using the FFT algorithm. The POLYSKETCH extends the idea behind TENSORSRHT to high-degree tensor products by recursively sketching pairs of vectors in a binary tree structure. This sketch preserves the norm of vectors in $\mathbb{R}^{d^p}$ with high probability and can be applied to tensor product vectors very quickly. The following Lemma, summarizes Theorems 1.2 and 1.3 of [1] and is proved in Appendix B.

Lemma 1 (POLYSKETCH)**.** *For every integers $p, d \geq 1$ and every $\varepsilon, \delta > 0$, there exists a distribution on random matrices $\boldsymbol{Q}^p \in \mathbb{R}^{m\times d^p}$, called degree p POLYSKETCH such that* **(1)** *for some $m = \mathcal{O}\left(\frac{p}{\varepsilon^2}\log^3\frac{1}{\varepsilon\delta}\right)$ and any $y \in \mathbb{R}^{d^p}$, $\Pr\left[\|\boldsymbol{Q}^p y\|_2^2 \in (1\pm\varepsilon)\|y\|_2^2\right] \geq 1 - \delta$;* **(2)** *for any $x \in \mathbb{R}^d$, if $e_1 \in \mathbb{R}^d$ is the standard basis vector along the first coordinate, the total time to compute $\boldsymbol{Q}^p(x^{\otimes(p-j)} \otimes e_1^{\otimes j})$ for all $j = 0, 1, \ldots, p$ is $\mathcal{O}\left(pm\log^2 m + \min\left\{\frac{p^{3/2}}{\varepsilon}\log\frac{1}{\delta}\,\mathrm{nnz}(x), pd\log d\right\}\right)$;* **(3)** *for any collection of vectors $v_1, \ldots, v_p \in \mathbb{R}^d$, the time to compute $\boldsymbol{Q}^p(v_1 \otimes \cdots \otimes v_p)$ is bounded by $\mathcal{O}\left(pm\log m + \frac{p^{3/2}}{\varepsilon}d\log\frac{1}{\delta}\right)$;* **(4)** *for any $\lambda > 0$ and any matrix $\boldsymbol{A} \in \mathbb{R}^{d^p\times n}$, where the statis-*

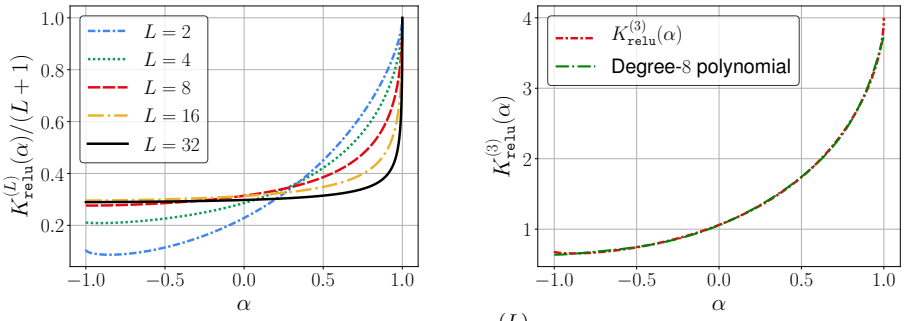


Figure 1: (Left) Normalized ReLU-NTK function $K^{(L)}_{\texttt{relu}}(\cdot)$ for $L = \{2, 4, 8, 16, 32\}$ and (Right) a degree-8 polynomial approximation of ReLU-NTK with $L = 3$.

tical dimension of $\boldsymbol{A}^\top \boldsymbol{A}$ *is* s_λ*, there exists some* $m = \mathcal{O}\left(\frac{p^4 s_\lambda}{\varepsilon^2} \log^3 \frac{n}{\varepsilon\delta}\right)$ *such that,*

$$\Pr\left[(1-\varepsilon)\left(\boldsymbol{A}^\top \boldsymbol{A} + \lambda \boldsymbol{I}\right) \preceq (\boldsymbol{Q}^p \boldsymbol{A})^\top (\boldsymbol{Q}^p \boldsymbol{A}) + \lambda \boldsymbol{I} \preceq (1+\varepsilon)\left(\boldsymbol{A}^\top \boldsymbol{A} + \lambda \boldsymbol{I}\right)\right] \geq 1 - \delta. \tag{1}$$

2 ReLU Neural Tangent Kernel

Arora et al. [5] showed how to exactly compute the NTK of a L-layer fully-connected network, denoted by $\Theta^{(L)}_{\texttt{ntk}}(y, z)$, for any pair of vectors $y, z \in \mathbb{R}^d$ using a dynamic program in $\mathcal{O}(d + L)$ time. However, it is hard to gain insight into the structure of this kernel using that the dynamic program expression which involves recursive applications of nontrivial expectations. Fortunately, for the important case of ReLU activation this kernel takes an extremely nice and highly structured form. The NTK in this case can be fully characterized by a univariate function $K^{(L)}_{\texttt{relu}} : [-1, 1] \to \mathbb{R}$ that we refer to as *ReLU-NTK*, which is the composition of the arc-cosine kernels [12] and was recently derived in [9]. Exploiting this special structure is the key to designing efficient sketching methods and random features for this kernel.

Definition 1 (ReLU-NTK function)**.** For every integer $L > 0$, the L-layer ReLU-NTK function $K^{(L)}_{\texttt{relu}} : [-1, 1] \to \mathbb{R}$ is defined via following procedure, for every $\alpha \in [-1, 1]$:

1. Let $\kappa_0(\alpha)$ and $\kappa_1(\alpha)$ be 0^{th} and 1^{st} order arc-cosine kernels [12] defined as follows,

$$\kappa_0(\alpha) := \frac{1}{\pi}\left(\pi - \arccos(\alpha)\right), \quad \text{and} \quad \kappa_1(\alpha) := \frac{1}{\pi}\left(\sqrt{1-\alpha^2} + \alpha \cdot (\pi - \arccos(\alpha))\right). \tag{2}$$

2. Let $\Sigma^{(0)}_{\texttt{relu}}(\alpha) := \alpha$ and for $\ell = 1, 2, \ldots L$, define $\Sigma^{(\ell)}_{\texttt{relu}}(\alpha)$ and $\dot{\Sigma}^{(\ell)}_{\texttt{relu}}(\alpha)$ as follows,

$$\Sigma^{(\ell)}_{\texttt{relu}}(\alpha) := \underbrace{\kappa_1 \circ \kappa_1 \circ \cdots \circ \kappa_1}_{\ell\text{-fold self composition}}(\alpha), \quad \text{and} \quad \dot{\Sigma}^{(\ell)}_{\texttt{relu}}(\alpha) := \kappa_0\left(\Sigma^{(\ell-1)}_{\texttt{relu}}(\alpha)\right). \tag{3}$$

3. Let $K^{(0)}_{\texttt{relu}}(\alpha) := \Sigma^{(0)}_{\texttt{relu}}(\alpha) = \alpha$ and for $\ell = 1, 2, \ldots L$, define $K^{(\ell)}_{\texttt{relu}}(\alpha)$ recursively as follows,

$$K^{(\ell)}_{\texttt{relu}}(\alpha) := K^{(\ell-1)}_{\texttt{relu}}(\alpha) \cdot \dot{\Sigma}^{(\ell)}_{\texttt{relu}}(\alpha) + \Sigma^{(\ell)}_{\texttt{relu}}(\alpha). \tag{4}$$

The connection between ReLU-NTK function $K^{(L)}_{\texttt{relu}}$ and the NTK kernel $\Theta^{(L)}_{\texttt{ntk}}$ is formalized bellow,

$$\Theta^{(L)}_{\texttt{ntk}}(y, z) \equiv \|y\|_2 \|z\|_2 \cdot K^{(L)}_{\texttt{relu}}\left(\frac{\langle y, z\rangle}{\|y\|_2 \|z\|_2}\right), \quad \text{for any } y, z \in \mathbb{R}^d. \tag{5}$$

This shows that the NTK is a *normalized dot-product kernel* which can be fully characterized by $K^{(L)}_{\texttt{relu}} : [-1, 1] \to \mathbb{R}$, plotted in Fig. 1. As shown in Fig. 1, this function is smooth and can be tightly approximated with a low-degree polynomial. It is evident that for larger values of L, $K^{(L)}_{\texttt{relu}}(\cdot)$ converges to a *knee shape*, i.e., it has a nearly constant value of roughly $0.3(L+1)$ on the interval $[-1, 1 - \mathcal{O}(L^{-1})]$, and on the interval $[1 - \mathcal{O}(L^{-1}), 1]$ its value sharply increases to $L + 1$ at $\alpha = 1$.

Algorithm 1 NTKSKETCH for fully-connected ReLU networks

input: vector $x \in \mathbb{R}^d$, network depth L, error and failure parameters $\varepsilon, \delta > 0$

Choose integers $s = \widetilde{\mathcal{O}}\left(\frac{L^2}{\varepsilon^2}\right)$, $n_1 = \widetilde{\mathcal{O}}\left(\frac{L^4}{\varepsilon^4}\right)$, $r = \widetilde{\mathcal{O}}\left(\frac{L^6}{\varepsilon^4}\right)$, $m = \widetilde{\mathcal{O}}\left(\frac{L^8}{\varepsilon^{\frac{16}{3}}}\right)$, and $s^* = \mathcal{O}\left(\frac{1}{\varepsilon^2}\log\frac{1}{\delta}\right)$ appropriately[†]

For $p = \left\lceil 2L^2/\varepsilon^{\frac{4}{3}} \right\rceil$ and $p' = \lceil 9L^2/\varepsilon^2 \rceil$, polynomials $P^{(p)}_{\texttt{relu}}(\cdot)$ and $\dot{P}^{(p')}_{\texttt{relu}}(\cdot)$ are defined as,

$$P^{(p)}_{\texttt{relu}}(\alpha) \equiv \sum_{j=0}^{2p+2} c_j \cdot \alpha^j := \frac{1}{\pi} + \frac{\alpha}{2} + \frac{1}{\pi}\sum_{i=0}^{p} \frac{(2i)! \cdot \alpha^{2i+2}}{2^{2i}(i!)^2(2i+1)(2i+2)},$$
$$\dot{P}^{(p')}_{\texttt{relu}}(\alpha) \equiv \sum_{j=0}^{2p'+1} b_j \cdot \alpha^j := \frac{1}{2} + \frac{1}{\pi}\sum_{i=0}^{p'} \frac{(2i)!}{2^{2i}(i!)^2(2i+1)} \cdot \alpha^{2i+1}. \tag{6}$$

$\phi^{(0)}(x) \leftarrow \|x\|_2^{-1} \cdot \boldsymbol{Q}^1 \cdot x$, where $\boldsymbol{Q}^1 \in \mathbb{R}^{r \times d}$ is a degree-1 POLYSKETCH as per Lemma 1

$\psi^{(0)}(x) \leftarrow \boldsymbol{V} \cdot \phi^{(0)}(x)$, where $\boldsymbol{V} \in \mathbb{R}^{s \times r}$ is an instance of SRHT [2]

for $\ell = 1$ to L **do**

Let $\boldsymbol{Q}^{2p+2} \in \mathbb{R}^{m \times r^{2p+2}}$ be a degree-$2p+2$ POLYSKETCH. Also, let $\boldsymbol{T} \in \mathbb{R}^{r \times (2p+3)\cdot m}$ be an instance of SRHT. For every $l = 0, 1, \ldots, 2p+2$, compute:

$$Z^{(\ell)}_l(x) \leftarrow \boldsymbol{Q}^{2p+2}\left(\left[\phi^{(\ell-1)}(x)\right]^{\otimes l} \otimes e_1^{\otimes 2p+2-l}\right),\ \phi^{(\ell)}(x) \leftarrow \boldsymbol{T} \cdot \bigoplus_{l=0}^{2p+2} \sqrt{c_l} Z^{(\ell)}_l(x) \tag{7}$$

Let $\boldsymbol{Q}^{2p'+1} \in \mathbb{R}^{n_1 \times r^{2p'+1}}$ be a degree-$2p'+1$ POLYSKETCH. Also, let $\boldsymbol{W} \in \mathbb{R}^{s \times (2p'+2)\cdot n_1}$ be an instance of SRHT. For every $l = 0, 1, \ldots, 2p'+1$, compute:

$$Y^{(\ell)}_l(x) \leftarrow \boldsymbol{Q}^{2p'+1}\left(\left[\phi^{(\ell-1)}(x)\right]^{\otimes l} \otimes e_1^{\otimes 2p'+1-l}\right),\ \dot{\phi}^{(\ell)}(x) \leftarrow \boldsymbol{W} \cdot \bigoplus_{l=0}^{2p'+1} \sqrt{b_l} Y^{(\ell)}_l(x) \tag{8}$$

Let $\boldsymbol{Q}^2 \in \mathbb{R}^{s \times s^2}$ be a degree-2 POLYSKETCH. Also, let $\boldsymbol{R} \in \mathbb{R}^{s \times (s+r)}$ be an SRHT. Compute:

$$\psi^{(\ell)}(x) \leftarrow \boldsymbol{R} \cdot \left(\boldsymbol{Q}^2\left(\psi^{(\ell-1)}(x) \otimes \dot{\phi}^{(\ell)}(x)\right) \oplus \phi^{(\ell)}(x)\right). \tag{9}$$

Let $\boldsymbol{G} \in \mathbb{R}^{s^* \times s}$ be a matrix of i.i.d. entries with distribution $\mathcal{N}(0, \frac{1}{s^*})$. Compute:

$$\Psi^{(L)}_{\texttt{ntk}}(x) \leftarrow \|x\|_2 \cdot \boldsymbol{G} \cdot \psi^{(L)}(x). \tag{10}$$

return $\Psi^{(L)}_{\texttt{ntk}}(x)$

3 Sketching and Random Features for NTK

The main results of this section are efficient oblivious sketching as well as random features for the fully-connected NTK. As shown in Definition 1 and Eq. (5), the NTK $\Theta^{(L)}_{\texttt{ntk}}$, is constructed by recursive composition of arc-cosine kernels $\kappa_1(\cdot)$ and $\kappa_0(\cdot)$. So, to design efficient sketches for the NTK we crucially need efficient methods for approximating these functions. Generally, there are two main approaches to approximating these functions; one is random features sampling and the other is truncated Taylor series expansion coupled with fast sketching. We design algorithms by exploiting both of these techniques.

3.1 NTK Sketch

Our main tool is approximating the arc-cosine kernels with low-degree polynomials, and then applying POLYSKETCH to the resulting polynomial kernels. The features for multi-layer NTK are the recursive tensor product of arc-cosine sketches at consecutive layers, which in turn can be sketched efficiently using POLYSKETCH. We present our oblivious sketch in Algorithm 1.

Now we present our main theorem on NTKSKETCH algorithm as follows.

Theorem 1. *For every integers $d \geq 1$ and $L \geq 2$, and any $\varepsilon, \delta > 0$, let $\Theta_{\texttt{ntk}}^{(L)} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be the L-layer NTK with ReLU activation as per Definition 1 and Eq. (5). Then there exists a randomized map $\Psi_{\texttt{ntk}}^{(L)} : \mathbb{R}^d \to \mathbb{R}^{s^*}$ for some $s^* = \mathcal{O}\left(\frac{1}{\varepsilon^2}\log\frac{1}{\delta}\right)$ such that the following invariants hold,*

1. *For any vectors $y, z \in \mathbb{R}^d$: $\Pr\left[\left|\left\langle \Psi_{\texttt{ntk}}^{(L)}(y), \Psi_{\texttt{ntk}}^{(L)}(z)\right\rangle - \Theta_{\texttt{ntk}}^{(L)}(y,z)\right| \leq \varepsilon \cdot \Theta_{\texttt{ntk}}^{(L)}(y,z)\right] \geq 1-\delta.$*
2. *For every vecor $x \in \mathbb{R}^d$, the time to compute $\Psi_{\texttt{ntk}}^{(L)}(x)$ is $\mathcal{O}\left(\frac{L^{11}}{\varepsilon^{6.7}}\log^3\frac{L}{\varepsilon\delta} + \frac{L^3}{\varepsilon^2}\log\frac{L}{\varepsilon\delta}\cdot \mathrm{nnz}(x)\right)$.*

For a proof, see Appendix C. One can observe that the runtime of our NTKSKETCH is faster than the gradient features of an ultra-wide random DNN, studied by Arora et al. [5], by a factor of L^3/ε^2.

3.2 NTK Random Features

The main difference between our random features construction and NTKSKETCH is the use of random features for approximating arc-cosine kernels κ_0 and κ_1 in Eq. (2). For any $x \in \mathbb{R}^d$, we denote

$$\Phi_0(x) := \sqrt{\frac{2}{m_0}}\,\mathrm{Step}\left(\left[w_1,\ldots,w_{m_0}\right]^\top x\right), \quad \Phi_1(x) := \sqrt{\frac{2}{m_1}}\,\mathrm{ReLU}\left(\left[w'_1,\ldots,w'_{m_1}\right]^\top x\right), \tag{11}$$

where $w_1,\ldots,w_{m_0},w'_1,\ldots,w'_{m_1} \in \mathbb{R}^d$ are i.i.d. samples from $\mathcal{N}(0,\boldsymbol{I}_d)$. Cho and Saul [12] showed that $\mathbb{E}[\langle\Phi_0(y),\Phi_0(z)\rangle] = \kappa_0\left(\frac{\langle y,z\rangle}{\|y\|_2\|z\|_2}\right)$ and $\mathbb{E}[\langle\Phi_1(y),\Phi_1(z)\rangle] = \|y\|_2\|z\|_2\cdot\kappa_1\left(\frac{\langle y,z\rangle}{\|y\|_2\|z\|_2}\right)$. The feature map for multi-layer NTK can be obtained by recursive tensoring of random feature maps for arc-cosine kernels at each layer of the network. However, one major drawback of such explicit tensoring is that the number of features, and thus the runtime, will be exponential in depth L. In order to make the feature map more compact, we utilize a degree-2 POLYSKETCH $\boldsymbol{Q}^2$ to reduce the dimension of the tensor products at each layer and get rid of exponential dependence on L. We present the performance guarantee of our random features, defined in Algorithm 2, in Theorem 2.

Theorem 2. *Given $y, z \in \mathbb{R}^d$ and $L \geq 2$, let $\Theta_{\texttt{ntk}}^{(L)}$ the L-layer fully-connected ReLU NTK. For $\varepsilon, \delta > 0$, there exist $m_0 = \mathcal{O}\left(\frac{L^2}{\varepsilon^2}\log\frac{L}{\delta}\right), m_1 = \mathcal{O}\left(\frac{L^6}{\varepsilon^4}\log\frac{L}{\delta}\right), m_s = \mathcal{O}\left(\frac{L^2}{\varepsilon^2}\log^3\frac{L}{\varepsilon\delta}\right)$, such that,*

$$\Pr\left[\left|\left\langle \Psi_{\texttt{rf}}^{(L)}(y), \Psi_{\texttt{rf}}^{(L)}(z)\right\rangle - \Theta_{\texttt{ntk}}^{(L)}(y,z)\right| \leq \varepsilon\cdot\Theta_{\texttt{ntk}}^{(L)}(y,z)\right] \geq 1-\delta, \tag{12}$$

where $\Psi_{\texttt{rf}}^{(L)}(y), \Psi_{\texttt{rf}}^{(L)}(z) \in \mathbb{R}^{m_1+m_s}$ are the outputs of Algorithm 2, using the same randomness.

The proof of Theorem 2 is provided in Appendix D. Arora et al. [5] proved that the gradient of randomly initialized ReLU network with finite width can approximate the NTK, but their feature dimension should be $\Omega\left(\frac{L^{13}}{\varepsilon^8}\log^2\frac{L}{\delta} + \frac{L^6}{\varepsilon^4}\cdot\log\frac{L}{\delta}\cdot d\right)$ which is larger than ours by a factor of $\frac{L^7}{\varepsilon^4}\log\frac{L}{\delta}$. In Section 5, we also empirically show that Algorithm 2 requires far fewer features than random gradients.

3.3 Spectral Approximation for NTK via Leverage Scores Sampling

Although the above NTK approximations can estimate the kernel function itself, it is still questionable how it affects the performance of downstream tasks. Several works on kernel approximation adopt spectral approximation bound with regularization $\lambda > 0$ and approximation factor $\varepsilon > 0$, that is,

$$(1-\varepsilon)(\boldsymbol{K}_{\texttt{ntk}}^{(L)} + \lambda\boldsymbol{I}) \preceq (\boldsymbol{\Psi}^{(L)})^\top\boldsymbol{\Psi}^{(L)} + \lambda\boldsymbol{I} \preceq (1+\varepsilon)(\boldsymbol{K}_{\texttt{ntk}}^{(L)} + \lambda\boldsymbol{I}), \tag{13}$$

where $\boldsymbol{\Psi}^{(L)} := \left[\Psi^{(L)}(x_1),\ldots,\Psi^{(L)}(x_n)\right]$ and $[\boldsymbol{K}_{\texttt{ntk}}^{(L)}]_{i,j} = \Theta_{\texttt{ntk}}^{(L)}(x_i,x_j)$. The spectral bound can provide rigorous guarantees for downstream applications including kernel ridge regression [8], clustering and PCA [32]. We first provide spectral bounds for arc-cosine kernels, then we present our spectral approximation bound for two-layer ReLU networks, which is the first in the literature.

[†] $\widetilde{\mathcal{O}}(\cdot)$ suppresses poly$(\log\frac{L}{\varepsilon\delta})$ factors.

Algorithm 2 Random Features for ReLU NTK via PolySketch

input: vector $x \in \mathbb{R}^d$, network depth L, feature dimensions m_0, m_1, and m_s
$\psi_{\mathtt{rf}}^{(0)}(x) \leftarrow x/\|x\|_2, \phi_{\mathtt{rf}}^{(0)}(x) \leftarrow x/\|x\|_2$
for $\ell = 1$ to L **do**
  $\dot{\phi}_{\mathtt{rf}}^{(\ell)}(x) \leftarrow \Phi_0\left(\phi_{\mathtt{rf}}^{(\ell-1)}(x)\right)$, where Φ_0 is defines as per Eq. (11) with m_0 features
  $\phi_{\mathtt{rf}}^{(\ell)}(x) \leftarrow \Phi_1\left(\phi_{\mathtt{rf}}^{(\ell-1)}(x)\right)$, where Φ_1 is defines as per Eq. (11) with m_1 features
  Draw a degree-2 PolySketch $\boldsymbol{Q}^2$ that maps to $\mathbb{R}^{m_s}$ and compute:

$$\psi_{\mathtt{rf}}^{(\ell)}(x) \leftarrow \phi_{\mathtt{rf}}^{(\ell)}(x) \oplus \boldsymbol{Q}^2 \cdot \left(\dot{\phi}_{\mathtt{rf}}^{(\ell)}(x) \otimes \psi_{\mathtt{rf}}^{(\ell-1)}(x)\right)$$

return $\Psi_{\mathtt{rf}}^{(L)}(x) \leftarrow \|x\|_2 \cdot \psi_{\mathtt{rf}}^{(L)}(x)$

To guarantee that the arc-cosine random features in Eq. (11) provide spectral approximation, we will use the leverage score sampling framework of [8, 28]. We reduce the variance of random features by performing importance sampling. The challenge is to find a proper modified distribution that certainly reduces the variance. It turns out that the original 0^{th} order arc-cosine random features has a small enough variance. More precisely, let $\boldsymbol{K}_0$ be the 0^{th} order arc-cosine kernel matrix, i.e., $[\boldsymbol{K}_0]_{i,j} = \kappa_0\left(\frac{\langle x_i, x_j\rangle}{\|x_i\|_2\|x_j\|_2}\right)$, and $\boldsymbol{\Phi}_0 := [\Phi_0(x_1), \ldots, \Phi_0(x_n)]$, where $\Phi_0(x)$ is defined in Eq. (11). If the number of features $m_0 \geq \frac{8}{3}\frac{n}{\lambda\varepsilon^2}\log\left(\frac{16 s_\lambda}{\delta}\right)$, then

$$\Pr\left[(1-\varepsilon)(\boldsymbol{K}_0 + \lambda \boldsymbol{I}) \preceq \boldsymbol{\Phi}_0^\top \boldsymbol{\Phi}_0 + \lambda \boldsymbol{I} \preceq (1+\varepsilon)(\boldsymbol{K}_0 + \lambda \boldsymbol{I})\right] \geq 1 - \delta. \tag{14}$$

Next, we consider spectral approximation of the 1^{st} order arc-cosine kernel. Unlike the previous case, modifications of the sampling distribution are required. Specifically, for any $x \in \mathbb{R}^d$, let

$$\widetilde{\Phi}_1(x) = \sqrt{\frac{2d}{m_1}}\mathrm{ReLU}\left(\left[\frac{w_1}{\|w_1\|_2}, \ldots, \frac{w_{m_1}}{\|w_{m_1}\|_2}\right]^\top x\right), \tag{15}$$

where $w_1, \ldots, w_{m_1} \in \mathbb{R}^d$ are i.i.d. samples from $p(w) := \frac{1}{(2\pi)^{d/2} d}\|w\|_2^2 \exp\left(-\frac{1}{2}\|w\|_2^2\right)$. For this modified features, let $\boldsymbol{X} \in \mathbb{R}^{d\times n}$ be the dataset, $\boldsymbol{K}_1$ be the 1^{st} order arc-cosine kernel matrix, i.e., $[\boldsymbol{K}_1]_{i,j} = \|x_i\|_2\|x_j\|_2 \cdot \kappa_1\left(\frac{\langle x_i, x_j\rangle}{\|x_i\|_2\|x_j\|_2}\right)$, and $\boldsymbol{\Phi}_1 := \left[\widetilde{\Phi}_1(x_1), \ldots, \widetilde{\Phi}_1(x_n)\right]$. If the number of features $m_1 \geq \frac{8}{3}\frac{d}{\varepsilon^2} \cdot \min\left\{\mathrm{rank}(\boldsymbol{X})^2, \frac{\|\boldsymbol{X}\|_2^2}{\lambda}\right\}\log\left(\frac{16 s_\lambda}{\delta}\right)$, then

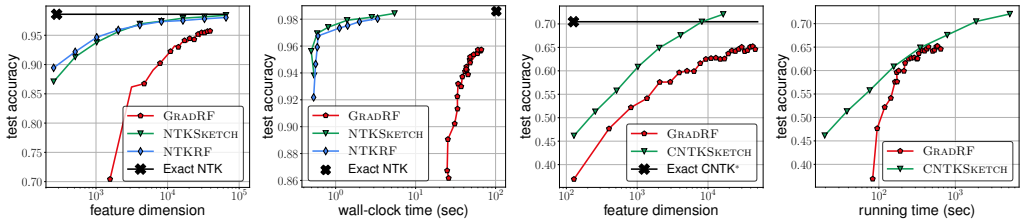
$$\Pr\left[(1-\varepsilon)(\boldsymbol{K}_1 + \lambda \boldsymbol{I}) \preceq \boldsymbol{\Phi}_1^\top \boldsymbol{\Phi}_1 + \lambda \boldsymbol{I} \preceq (1+\varepsilon)(\boldsymbol{K}_1 + \lambda \boldsymbol{I})\right] \geq 1 - \delta. \tag{16}$$

The details are provided in Appendix E.1 and Appendix E.2. We are now ready to state our spectral approximation bound for our modified random features.

Theorem 3. *Given a dataset $\boldsymbol{X} \in \mathbb{R}^{d\times n}$ with $\|\boldsymbol{X}_{(:,i)}\|_2 \leq 1$ for every $i \in [n]$, let $\boldsymbol{K}_{\mathtt{ntk}}, \boldsymbol{K}_0, \boldsymbol{K}_1$ be kernel matrices for two-layer ReLU NTK and arc-cosine kernels of 0^{th} and 1^{st} order, respectively. For any $\lambda > 0$, suppose s_λ is the statistical dimension of $\boldsymbol{K}_{\mathtt{ntk}}$. Modify Algorithm 2 by replacing $\Phi_1(\cdot)$ in line 5 with $\widetilde{\Phi}_1(\cdot)$ defined in Eq. (15). For any $\varepsilon, \delta > 0$, let $\boldsymbol{\Psi}_{\mathtt{rf}}^{(L)} \in \mathbb{R}^{(m_1+m_s)\times n}$ be the output matrix of this algorithm with $L = 1$. There exist $m_0 = \mathcal{O}\left(\frac{n}{\varepsilon^2\lambda}\log\frac{s_\lambda}{\delta}\right), m_1 = \mathcal{O}\left(\frac{d}{\varepsilon^2} \cdot \min\left\{\mathrm{rank}(\boldsymbol{X})^2, \frac{\|\boldsymbol{X}\|_2^2}{\lambda}\right\}\log\frac{s_\lambda}{\delta}\right), m_s = \mathcal{O}\left(\frac{1}{\varepsilon^2} \cdot \frac{n}{1+\lambda}\log^3\frac{n}{\varepsilon\delta}\right)$ such that,*

$$\Pr\left[(1-\varepsilon)\left(\boldsymbol{K}_{\mathtt{ntk}} + \lambda \boldsymbol{I}\right) \preceq \left(\boldsymbol{\Psi}_{\mathtt{rf}}^{(L)}\right)^\top \boldsymbol{\Psi}_{\mathtt{rf}}^{(L)} + \lambda \boldsymbol{I} \preceq (1+\varepsilon)\left(\boldsymbol{K}_{\mathtt{ntk}} + \lambda \boldsymbol{I}\right)\right] \geq 1 - \delta. \tag{17}$$

For a proof see Appendix E.3. To generalize the current proof technique to deeper networks, one needs a monotone property of arc-cosine kernels, i.e., $\kappa_1(\boldsymbol{X}) \preceq \kappa_1(\boldsymbol{Y})$ for $\boldsymbol{X} \preceq \boldsymbol{Y}$. However, this property does not hold in general and we leave the extension to deeper networks to future work.



(a) MNIST with NTK (b) CIFAR-10 with CNTK

Figure 2: Test accuracy of: (a) approximate NTK methods (GRADRF, NTKSKETCH and NTKRF) on MNIST and (b) approximate CNTK methods (GRADRF and CNTKSKETCH) on CIFAR-10.

4 Sketching Convolutional Neural Tangent Kernel

In this section, we design and analyze an efficient sketching method for the Convolutional Neural Tangent Kernel (CNTK). We focus mainly on CNTK with Global Average Pooling (GAP), which exhibits superior empirical performance compared to vanilla CNTK with no pooling [5], however, our techniques can be applied to the vanilla version, as well. Using the DP of Arora et al. [5], the number of operations needed for exact computation of the depth-L CNTK value $\Theta_{\texttt{cntk}}^{(L)}(y,z)$ for images $y,z\in\mathbb{R}^{d\times d}$ is $\Omega\left(d^4\cdot L\right)$, which is extremely slow particularly due to its quadratic dependence on the number of pixels of input images d^2. Fortunately, we are able to show that the CNTK for the important case of ReLU activation is a highly structured object that can be fully characterized in terms of tensoring and composition of arc-cosine kernels, and exploiting this special structure is key to designing efficient sketching methods for the CNTK. Unlike the fully-connected NTK, CNTK is not a simple dot-product kernel function like Eq. (5). The key reason being that CNTK works by partitioning its input images into patches and locally transforming the patches at each layer, as opposed to the NTK which operates on the entire input vectors. We present our derivation of the ReLU CNTK function and its main properties in Appendix F.

Similar to NTKSKETCH our method relies on approximating the arc-cosine kernels with low-degree polynomials via Taylor expansion, and then applying POLYSKETCH to the resulting polynomial kernels. Our sketch computes the features for each pixel of the input image, by tensor product of arc-cosine sketches at consecutive layers, which in turn can be sketched efficiently using POLYSKETCH . Additionally, the features of pixels that lie in the same patch get *locally combined* at each layer via direct sum operation. This precisely corresponds to the convolution operation in neural networks. We present our CNTKSKETCH algorithm in Appendix G and give its performance guarantee in the following theorem.

Theorem 4. *For every positive integers d_1,d_2,c and $L\geq 2$, and every $\varepsilon,\delta>0$, if we let $\Theta_{\texttt{cntk}}^{(L)}:\mathbb{R}^{d_1\times d_2\times c}\times\mathbb{R}^{d_1\times d_2\times c}\to\mathbb{R}$ be the L-layer CNTK with ReLU activation and GAP given in [5], then there exist a randomized map $\Psi_{\texttt{cntk}}^{(L)}:\mathbb{R}^{d_1\times d_2\times c}\to\mathbb{R}^{s^*}$ for some $s^*=\mathcal{O}\left(\frac{1}{\varepsilon^2}\log\frac{1}{\delta}\right)$ such that:*

1. *For any images $y,z\in\mathbb{R}^{d_1\times d_2\times c}$:*

$$\Pr\left[\left|\left\langle\Psi_{\texttt{cntk}}^{(L)}(y),\Psi_{\texttt{cntk}}^{(L)}(z)\right\rangle-\Theta_{\texttt{cntk}}^{(L)}(y,z)\right|\leq\varepsilon\cdot\Theta_{\texttt{cntk}}^{(L)}(y,z)\right]\geq 1-\delta.$$

2. *For every image $x\in\mathbb{R}^{d_1\times d_2\times c}$, time to compute $\Psi_{\texttt{cntk}}^{(L)}(x)$ is $\mathcal{O}\left(\frac{L^{11}}{\varepsilon^{6.7}}\cdot(d_1d_2)\cdot\log^3\frac{d_1d_2L}{\varepsilon\delta}\right)$.*

The proof is in Appendix G. Runtime of our CNTKSKETCH is only linear in the number of image pixels d_1d_2, which is in stark contrast to quadratic scaling of the exact CNTK computation [5].

5 Experiments

In this section, we empirically show that running least squares regression on the features generated by our methods is extremely fast and effective for learning with NTK and CNTK kernel machines. We run experiments on a system with an Intel E5-2630 CPU with 256 GB RAM and a single GeForce RTX 2080 GPUs with 12 GB RAM. Codes are available at `https://github.com/insuhan/ntk-sketch-rf`.

Table 1: Test accuracy and runtime to solve CNTK regression and its approximations on CIFAR-10. (*) means that the result is copied from Arora et al. [5].

	CNTKSKETCH (ours)			GRADRF			Exact CNTK	CNN
Feature dimension	4,096	8,192	16,384	9,328	17,040	42,816		
Test accuracy (%)	67.58	70.46	72.06	62.49	62.57	65.21	70.47*	63.81*
Time (s)	780	1,870	5,160	300	360	580	> 1,000,000	

Table 2: MSE and runtime on large-scale UCI datasets. We measure the entire time to solve kernel ridge regression. $(-)$ means Out-of-Memory error.

	MillionSongs		WorkLoads		CT		Protein	
# of data points (n)	467,315		179,585		53,500		39,617	
	MSE	Time (s)	MSE	Time (s)	MSE	Time (s)	MSE	Time (s)
RBF Kernel	–	–	–	–	35.37	59.23	18.96	46.45
RFF	109.50	231	4.034×10^4	53.0	48.20	15.2	19.72	12.1
NTK	–	–	–	–	30.52	72.10	20.24	76.93
NTKRF (ours)	94.27	95	3.554×10^4	35.7	46.91	2.12	20.51	4.3
NTKSKETCH (ours)	92.83	36	3.538×10^4	27.5	46.52	18.8	21.19	14.91

5.1 NTK Classification on MNIST

We first benchmark our proposed NTK approximation algorithms on MNIST [25] dataset and compare against gradient-based NTK random features [5] (GRADRF) as a baseline method. To apply our methods and GRADRF into classification task, we encode class labels into one-hot vectors with zero-mean and solve the ridge regression problem. We search the ridge parameter with a random subset of training set and choose the one that achieves the best validation accuracy. We use the ReLU network with depth $L = 1$. In Fig. 2a, we observe that our random features (NTKRF) achieves the best test accuracy. The NTKSKETCH narrowly follows the performance of NTKRF and the Grad-RF is the worst method which confirms the observations of Arora et al. [5], i.e., gradient of a finite width network degrades practical performances.

Remark 1 (Optimizing NTKSKETCH for Deeper Nets)**.** As shown in Eq. (5), the NTK is a normalized dot-product kernel characterized by the function $K^{(L)}_{\texttt{relu}}(\alpha)$. This function can be easily computed using $\mathcal{O}(L)$ operations at any desired $\alpha \in [-1, 1]$, therefore, we can efficiently fit a polynomial to this function using numerical methods (for instance, it is shown in Fig. 1 that a degree-8 polynomial can tightly approximate the depth-3 ReLU-NTK function $K^{(3)}_{\texttt{relu}}$). Then, we can efficiently sketch the resulting polynomial kernel using POLYSKETCH , as was previously done for Gaussian and general dot-product kernels [1, 40]. Therefore, we can accelerate our NTKSKETCH for deeper networks ($L > 2$), using this heuristic.

5.2 CNTK Classification on CIFAR-10

Next we test our CNTKSKETCH on CIFAR-10 dataset [24]. We choose a convolutional network of depth $L = 3$ and compare CNTKSKETCH and GRADRF for various feature dimensions. We borrow results of both CNTK and CNN from Arora et al. [5]. The results are provided in Fig. 2b and Table 1. Somewhat surprisingly, CNTKSKETCH even performs better than the exact CNTK regression by achieving 72.06% when feature dimension is set to $16{,}384$. The likely explanation is that CNTKSKETCH takes advantages of implicit regularization effects of approximate feature map and powerful expressiveness of the CNTK. Moreover, computing the CNTK matrix takes over 250 hours (12 days) under our setting which is at least $150\times$ slower than our CNTKSKETCH.

5.3 Regression on Large-scale UCI Datasets

We also demonstrate the computational efficiency of our NTKSKETCH and NTKRF using 4 large-scale UCI regression datasets [17] by comparing against exact NTK, RBF as well as Random Fourier Features (RFF). For our methods and RFF, we fix the output dimension to $m = 8{,}192$ for all datasets. In Table 2, we report the runtime to compute feature map or kernel matrix and evaluate the averaged mean squared errors (MSE) on the test set via 4-fold cross validation. The exact kernel methods face

Out-of-Memory error on larger datasets. The proposed NTK approximations are significantly faster than the exact NTK, e.g., NTKRF shows up to 30× speedup under CT dataset. We also verify that, except for Protein dataset, our methods outperform RFF.

6 Discussion and Conclusion

In this work, we propose efficient low-rank feature maps for the NTK and CNTK kernel matrices based on both sketching and random features. Computing NTK have been raised severe computational problems when they apply to practical applications. Our methods runs remarkably faster than the NTK with performance improvement.

Potential negative societal impact. This is a technical work proposing provable algorithms which stand alone independently of data, e.g., do not learn any private information of input data. We think there is no particular potential negative societal impact due to our work.

Limitations. This paper only considers fully-connected and convolutional neural networks, and our ideas are not directly applicable to scale up NTK of other deep networks, e.g., transformers [21].

Acknowledgments and Disclosure of Funding

Amir Zandieh was partially supported by the Swiss NSF grant No. P2ELP2_195140. Haim Avron and Neta Shoham were partially supported by BSF grant 2017698 and ISF grant 1272/17. Jinwoo Shin was partially supported by the Engineering Research Center Program through the National Research Foundation of Korea (NRF) funded by the Korean Government MSIT (NRF_2018R1A5A1059921) and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2019_0_00075, Artificial Intelligence Graduate School Program (KAIST)).

References

[1] Thomas D Ahle, Michael Kapralov, Jakob BT Knudsen, Rasmus Pagh, Ameya Velingker, David P Woodruff, and Amir Zandieh. Oblivious sketching of high-degree polynomial kernels. In *Symposium on Discrete Algorithms (SODA)*, 2020.

[2] Nir Ailon and Bernard Chazelle. The fast Johnson–Lindenstrauss transform and approximate nearest neighbors. *SIAM Journal on computing*, 2009.

[3] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning (ICML)*, 2019.

[4] Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning (ICML)*, 2019.

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

[6] Sanjeev Arora, Simon S Du, Zhiyuan Li, Ruslan Salakhutdinov, Ruosong Wang, and Dingli Yu. Harnessing the Power of Infinitely Wide Deep Nets on Small-data Tasks. In *International Conference on Learning Representations (ICLR)*, 2019.

[7] Haim Avron, Huy Nguyen, and David Woodruff. Subspace embeddings for the polynomial kernel. In *Neural Information Processing Systems (NeurIPS)*, 2014.

[8] Haim Avron, Michael Kapralov, Cameron Musco, Christopher Musco, Ameya Velingker, and Amir Zandieh. Random Fourier features for kernel ridge regression: Approximation bounds and statistical guarantees. In *International Conference on Machine Learning (ICML)*, 2017.

[9] Alberto Bietti and Julien Mairal. On the inductive bias of neural tangent kernels. In *Neural Information Processing Systems (NeurIPS)*, 2019.

[10] Yuan Cao and Quanquan Gu. Generalization bounds of stochastic gradient descent for wide and deep neural networks. In *Neural Information Processing Systems (NeurIPS)*, 2019.

[11] Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Neural Information Processing Systems (NeurIPS)*, 2019.

[12] Youngmin Cho and Lawrence Saul. Kernel methods for deep learning. In *Neural Information Processing Systems (NeurIPS)*, 2009.

[13] Michael B Cohen, Jelani Nelson, and David P Woodruff. Optimal Approximate Matrix Product in Terms of Stable Rank. In *International Colloquium on Automata, Languages, and Programming (ICALP)*, 2016.

[14] Amit Daniely, Roy Frostig, and Yoram Singer. Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity. In *Neural Information Processing Systems (NeurIPS)*, 2016.

[15] Sanjoy Dasgupta and Anupam Gupta. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures & Algorithms*.

[16] Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations (ICLR)*, 2019.

[17] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

[18] Adrià Garriga-Alonso, Carl Edward Rasmussen, and Laurence Aitchison. Deep Convolutional Networks as shallow Gaussian Processes. In *International Conference on Learning Representations (ICLR)*, 2018.

[19] Amnon Geifman, Abhay Yadav, Yoni Kasten, Meirav Galun, David Jacobs, and Ronen Basri. On the similarity between the laplace and neural tangent kernels. In *Neural Information Processing Systems (NeurIPS)*, 2020.

[20] Insu Han, Haim Avron, and Jinwoo Shin. Polynomial Tensor Sketch for Element-wise Function of Low-Rank Matrix. In *International Conference on Machine Learning (ICML)*, 2020.

[21] Jiri Hron, Yasaman Bahri, Jascha Sohl-Dickstein, and Roman Novak. Infinite attention: NNGP and NTK for deep attention networks. In *International Conference on Machine Learning (ICML)*, 2020.

[22] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Neural Information Processing Systems (NeurIPS)*, 2018.

[23] Arthur Jacot, Berfin Simsek, Francesco Spadaro, Clément Hongler, and Franck Gabriel. Implicit regularization of random feature models. In *International Conference on Machine Learning (ICML)*, 2020.

[24] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

[25] Yann LeCun, Corinna Cortes, and CJ Burges. MNIST handwritten digit database. 2010.

[26] Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep Neural Networks as Gaussian Processes. In *International Conference on Learning Representations (ICLR)*, 2018.

[27] Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In *Neural Information Processing Systems (NeurIPS)*, 2019.

[28] Jason Lee, Ruoqi Shen, Zhang Song, Mendi Wang, and Zheng Yu. Generalized Leverage Score Sampling for Neural Networks. In *Neural Information Processing Systems (NeurIPS)*, 2020.

[29] Zhu Li, Jean-Francois Ton, Dino Oglic, and Dino Sejdinovic. Towards a Unified Analysis of Random Fourier Features. *Journal of Machine Learning Research (JMLR)*, 2021.

[30] Alexander G de G Matthews, Jiri Hron, Mark Rowland, Richard E Turner, and Zoubin Ghahramani. Gaussian Process Behaviour in Wide Deep Neural Networks. In *International Conference on Learning Representations (ICLR)*, 2018.

[31] Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 2018.

[32] Cameron Musco and Christopher Musco. Recursive Sampling for the Nyström Method. In *Neural Information Processing Systems (NeurIPS)*, 2017.

[33] Jelani Nelson and Huy L Nguyên. OSNAP: Faster numerical linear algebra algorithms via sparser subspace embeddings. In *Foundations of Computer Science (FOCS)*, 2013.

[34] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. In *International Conference on Learning Representations (ICLR)*, 2019.

[35] Roman Novak, Lechao Xiao, Yasaman Bahri, Jaehoon Lee, Greg Yang, Jiri Hron, Daniel A Abolafia, Jeffrey Pennington, and Jascha Sohl-dickstein. Bayesian Deep Convolutional Networks with Many Channels are Gaussian Processes. In *International Conference on Learning Representations (ICLR)*, 2018.

[36] Jeffrey Pennington, Felix Xinnan X Yu, and Sanjiv Kumar. Spherical random features for polynomial kernels. In *Neural Information Processing Systems (NeurIPS)*, 2015.

[37] Ali Rahimi and Benjamin Recht. Random Features for Large-Scale Kernel Machines. In *Neural Information Processing Systems (NeurIPS)*, 2007.

[38] Alessandro Rudi and Lorenzo Rosasco. Generalization properties of learning with random features. In *Neural Information Processing Systems (NeurIPS)*, 2017.

[39] Neta Shoham and Haim Avron. Experimental Design for Overparameterized Learning with Application to Single Shot Deep Active Learning. In *arXiv preprint arXiv:2009.12820*, 2020.

[40] David P Woodruff and Amir Zandieh. Near Input Sparsity Time Kernel Embeddings via Adaptive Sampling. In *International Conference on Machine Learning (ICML)*, 2020.

[41] David P Woodruff et al. Sketching as a Tool for Numerical Linear Algebra. *Foundations and Trends® in Theoretical Computer Science*, 2014.

[42] Greg Yang. Scaling limits of wide neural networks with weight sharing: Gaussian process behavior, gradient independence, and neural tangent kernel derivation. *arXiv preprint arXiv:1902.04760*, 2019.

[43] Luca Zancato, Alessandro Achille, Avinash Ravichandran, Rahul Bhotika, and Stefano Soatto. Predicting training time without training. In *Neural Information Processing Systems (NeurIPS)*, 2020.

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

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes]
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [No]
   (c) Did you discuss any potential negative societal impacts of your work? [No]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See the "README.md" in supplement.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A ReLU-NTK Expression

Arora et al. [5] showed how to exactly compute the L-layer NTK with activation $\sigma : \mathbb{R} \to \mathbb{R}$ using the following dynamic program (DP):

1. For every $y, z \in \mathbb{R}^d$, let $\Sigma^{(0)}(y, z) := \langle y, z \rangle$ and for every layer $h = 1, 2, \ldots L$, recursively define the covariance $\Sigma^{(h)} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ as:

$$\begin{aligned} \Lambda^{(h)}(y, z) &:= \begin{pmatrix} \Sigma^{(h-1)}(y, y) & \Sigma^{(h-1)}(y, z) \\ \Sigma^{(h-1)}(z, y) & \Sigma^{(h-1)}(z, z) \end{pmatrix}, \\ \Sigma^{(h)}(y, z) &:= \frac{\mathbb{E}_{(u,v)\sim\mathcal{N}\left(0, \Lambda^{(h)}(y,z)\right)} \left[\sigma(u) \cdot \sigma(v)\right]}{\mathbb{E}_{x\sim\mathcal{N}(0,1)} \left[|\sigma(x)|^2\right]}. \end{aligned} \tag{18}$$

2. For $h = 1, 2, \ldots L$, define the derivative covariance as,

$$\dot{\Sigma}^{(h)}(y, z) := \frac{\mathbb{E}_{(u,v)\sim\mathcal{N}\left(0, \Lambda^{(h)}(y,z)\right)} \left[\dot{\sigma}(u) \cdot \dot{\sigma}(v)\right]}{\mathbb{E}_{x\sim\mathcal{N}(0,1)} \left[|\sigma(x)|^2\right]}. \tag{19}$$

3. Let $\Theta^{(0)}_{\texttt{ntk}}(y, z) := \Sigma^{(0)}(y, z)$ and for every integer $L \geq 1$, the depth-L NTK expression is defined recursively as:

$$\Theta^{(L)}_{\texttt{ntk}}(y, z) := \Theta^{(L-1)}_{\texttt{ntk}}(y, z) \cdot \dot{\Sigma}^{(L)}(y, z) + \Sigma^{(L)}(y, z). \tag{20}$$

While using this DP, one can compute the kernel value $\Theta^{(L)}_{\texttt{ntk}}(y, z)$ for any pair of vectors $y, z \in \mathbb{R}^d$ using $\mathcal{O}(d + L)$ operations, it is hard to gain insight into the structure of this kernel using the expression above. In particular, the NTK expression involves recursive computation of nontrivial expectations which makes it is unclear whether there exist efficient sketching solutions for this kernel in its current form. However, we show that for the important case of ReLU activation, this kernel takes an extremely nice and highly structured form. In fact, the NTK can be fully characterized by a *univariate function* $K^{(L)}_{\texttt{relu}} : [-1, 1] \to \mathbb{R}$, and exploiting this special structure is the key to designing efficient sketching methods for this kernel.

Now we proceed to prove Eq. (5), that is,

$$\Theta^{(L)}_{\texttt{ntk}}(y, z) \equiv \|y\|_2 \|z\|_2 \cdot K^{(L)}_{\texttt{relu}} \left(\frac{\langle y, z \rangle}{\|y\|_2 \|z\|_2} \right), \quad \text{for any } y, z \in \mathbb{R}^d. \tag{5}$$

First note that the main component of the DP given in Eq. (18), Eq. (19), and Eq. (20) is the *Activation Covariances*:

$$\mathbb{E}_{w\sim\mathcal{N}(0,\boldsymbol{I}_d)} \left[\sigma(w^\top y) \cdot \sigma(w^\top z)\right], \text{ and } \mathbb{E}_{w\sim\mathcal{N}(0,\boldsymbol{I}_d)} \left[\dot{\sigma}(w^\top y) \cdot \dot{\sigma}(w^\top z)\right] \quad \text{for every } y, z \in \mathbb{R}^d.$$

It is worth noting that the above activation covariance functions are positive definite and hence they both define valid kernel functions in $\mathbb{R}^d \times \mathbb{R}^d$. The connection between the ReLU activation covariance functions and arc-cosine kernel functions defined in Eq. (2) of Definition 1 is proved in Cho and Saul [12]. We restate this result as follows,

$$\begin{aligned} \mathbb{E}_{w\sim\mathcal{N}(0,\boldsymbol{I}_d)} \left[\mathrm{ReLU}(w^\top y) \cdot \mathrm{ReLU}(w^\top z)\right] &= \frac{\|y\|_2 \|z\|_2}{2} \cdot \kappa_1 \left(\frac{\langle y, z \rangle}{\|y\|_2 \|z\|_2} \right) \\ \mathbb{E}_{w\sim\mathcal{N}(0,\boldsymbol{I}_d)} \left[\mathrm{Step}(w^\top y) \cdot \mathrm{Step}(w^\top z)\right] &= \frac{1}{2} \cdot \kappa_0 \left(\frac{\langle y, z \rangle}{\|y\|_2 \|z\|_2} \right). \end{aligned} \tag{21}$$

Proof of Eq. (5): Consider the NTK expression given in Eq. (18), Eq. (19), and Eq. (20). We first prove by induction on $h = 0, 1, 2, \ldots L$ that the covariance function $\Sigma^{(h)}(y, z)$ defined in Eq. (18) satisfies:

$$\Sigma^{(h)}(y, z) = \|y\|_2 \|z\|_2 \cdot \Sigma^{(h)}_{\texttt{relu}} \left(\frac{\langle y, z \rangle}{\|y\|_2 \|z\|_2} \right).$$

The **base of induction** is trivial for $h = 0$ due to $\Sigma^{(0)}(y, z) = \langle y, z \rangle = \|y\|_2 \|z\|_2 \cdot \Sigma^{(0)}_{\texttt{relu}} \left(\frac{\langle y, z \rangle}{\|y\|_2 \|z\|_2} \right)$.

To prove the **inductive step**, suppose that the inductive hypothesis holds for $h-1$, i.e.,

$$\Sigma^{(h-1)}(y,z) = \|y\|_2\|z\|_2 \cdot \Sigma^{(h-1)}_{\texttt{relu}}\left(\frac{\langle y,z\rangle}{\|y\|_2\|z\|_2}\right)$$

The 2×2 covariance matrix $\Lambda^{(h)}(y,z)$, defined in Eq. (18), can be decomposed as $\Lambda^{(h)}(y,z) = \begin{pmatrix} f^\top \\ g^\top \end{pmatrix} \cdot \begin{pmatrix} f & g \end{pmatrix}$, where $f,g \in \mathbb{R}^2$. Now note that $\|f\|_2^2 = \Sigma^{(h-1)}(y,y)$, hence, by inductive hypothesis, we have, $\|f\|_2^2 = \|y\|_2^2 \cdot \Sigma^{(h-1)}_{\texttt{relu}}\left(\frac{\langle y,y\rangle}{\|y\|_2^2}\right) = \|y\|_2^2 \cdot \Sigma^{(h-1)}_{\texttt{relu}}(1) = \|y\|_2^2$.

Using a similar argument we have $\|g\|_2^2 = \|z\|_2^2$. Therefore, by Eq. (21), we can write

$$\begin{aligned}
\Sigma^{(h)}(y,z) &= \frac{1}{\mathbb{E}_{x\sim\mathcal{N}(0,1)}\left[|\sigma(x)|^2\right]} \cdot \mathbb{E}_{w\sim\mathcal{N}(0,I_2)}\left[\sigma(w^\top f)\cdot\sigma(w^\top g)\right] \\
&= \frac{2}{\kappa_1(1)} \cdot \mathbb{E}_{w\sim\mathcal{N}(0,I_2)}\left[\sigma(w^\top f)\cdot\sigma(w^\top g)\right] \\
&= \|f\|_2\|g\|_2 \cdot \kappa_1\left(\frac{\langle f,g\rangle}{\|f\|_2\|g\|_2}\right) = \|y\|_2\|z\|_2 \cdot \kappa_1\left(\frac{\langle f,g\rangle}{\|y\|_2\|z\|_2}\right).
\end{aligned}$$

Since we assumed that $\Lambda^{(h)}(y,z) = \begin{pmatrix} f^\top \\ g^\top \end{pmatrix} \cdot \begin{pmatrix} f & g \end{pmatrix}$, we have $\langle f,g\rangle = \Sigma^{(h-1)}(y,z)$. By inductive hypothesis along with Eq. (3), we find that

$$\Sigma^{(h)}(y,z) = \|y\|_2\|z\|_2 \cdot \kappa_1\left(\Sigma^{(h-1)}_{\texttt{relu}}\left(\frac{\langle y,z\rangle}{\|y\|_2\|z\|_2}\right)\right) = \|y\|_2\|z\|_2 \cdot \Sigma^{(h)}_{\texttt{relu}}\left(\frac{\langle y,z\rangle}{\|y\|_2\|z\|_2}\right),$$

which completes the induction and proves that for every $h = 0,1,\ldots,L$,

$$\Sigma^{(h)}(y,z) = \|y\|_2\|z\|_2 \cdot \Sigma^{(h)}_{\texttt{relu}}\left(\frac{\langle y,z\rangle}{\|y\|_2\|z\|_2}\right). \tag{22}$$

For obtaining the final NTK expression we also need to simplify the derivative covariance function defined in Eq. (19). Recall that we proved before, the covariance matrix $\Lambda^{(h)}(y,z)$ can be decomposed as $\Lambda^{(h)}(y,z) = \begin{pmatrix} f^\top \\ g^\top \end{pmatrix} \cdot \begin{pmatrix} f & g \end{pmatrix}$, where $f,g \in \mathbb{R}^2$ with $\|f\|_2 = \|y\|_2$ and $\|g\|_2 = \|z\|_2$. Therefore, by Eq. (21), we can write

$$\begin{aligned}
\dot\Sigma^{(h)}(y,z) &= \frac{1}{\mathbb{E}_{x\sim\mathcal{N}(0,1)}\left[|\sigma(x)|^2\right]} \cdot \mathbb{E}_{w\sim\mathcal{N}(0,I_2)}\left[\dot\sigma(w^\top f)\cdot\dot\sigma(w^\top g)\right] \\
&= \frac{2}{\kappa_1(1)} \cdot \mathbb{E}_{w\sim\mathcal{N}(0,I_2)}\left[\dot\sigma(w^\top f)\cdot\dot\sigma(w^\top g)\right] \\
&= \kappa_0\left(\frac{\langle f,g\rangle}{\|y\|_2\|z\|_2}\right).
\end{aligned}$$

Since we assumed that $\Lambda^{(h)}(y,z) = \begin{pmatrix} f^\top \\ g^\top \end{pmatrix} \cdot \begin{pmatrix} f & g \end{pmatrix}$, $\langle f,g\rangle = \Sigma^{(h-1)}(y,z) = \|y\|_2\|z\|_2 \cdot \Sigma^{(h-1)}_{\texttt{relu}}\left(\frac{\langle y,z\rangle}{\|y\|_2\|z\|_2}\right)$. Therefore, by Eq. (3), for every $h \in \{1,2,\ldots L\}$,

$$\dot\Sigma^{(h)}(y,z) = \kappa_0\left(\Sigma^{(h-1)}_{\texttt{relu}}\left(\frac{\langle y,z\rangle}{\|y\|_2\|z\|_2}\right)\right) = \dot\Sigma^{(h)}_{\texttt{relu}}\left(\frac{\langle y,z\rangle}{\|y\|_2\|z\|_2}\right). \tag{23}$$

Now we prove by induction on integer L that the NTK with L layers and ReLU activation given in Eq. (20) satisfies

$$\Theta^{(L)}_{\texttt{ntk}}(y,z) = \|y\|_2\|z\|_2 \cdot K^{(L)}_{\texttt{relu}}\left(\frac{\langle y,z\rangle}{\|y\|_2\|z\|_2}\right).$$

The **base of induction**, trivially holds because, by Eq. (22):

$$\Theta^{(0)}_{\texttt{ntk}}(y,z) = \Sigma^{(0)}(y,z) = \|y\|_2\|z\|_2 \cdot K^{(0)}_{\texttt{relu}}\left(\frac{\langle y,z\rangle}{\|y\|_2\|z\|_2}\right).$$

To prove the **inductive step**, suppose that the inducive hypothesis holds for $L-1$, that is $\Theta_{\texttt{ntk}}^{(L-1)}(y,z) = \|y\|_2\|z\|_2 \cdot K_{\texttt{relu}}^{(L-1)}\left(\frac{\langle y,z\rangle}{\|y\|_2\|z\|_2}\right)$. Now using the recursive definition of $\Theta_{\texttt{ntk}}^{(L)}(y,z)$ given in Eq. (20) along with Eq. (22) and Eq. (23) we can write,

$$\begin{aligned}
\Theta_{\texttt{ntk}}^{(L)}(y,z) &= \Theta_{\texttt{ntk}}^{(L-1)}(y,z)\cdot\dot{\Sigma}^{(L)}(y,z) + \Sigma^{(L)}(y,z) \\
&= \|y\|_2\|z\|_2 \cdot K_{\texttt{relu}}^{(L-1)}\left(\frac{\langle y,z\rangle}{\|y\|_2\|z\|_2}\right)\cdot\dot{\Sigma}_{\texttt{relu}}^{(h)}\left(\frac{\langle y,z\rangle}{\|y\|_2\|z\|_2}\right) + \|y\|_2\|z\|_2\cdot\Sigma_{\texttt{relu}}^{(h)}\left(\frac{\langle y,z\rangle}{\|y\|_2\|z\|_2}\right) \\
&\equiv \|y\|_2\|z\|_2 \cdot K_{\texttt{relu}}^{(L)}\left(\frac{\langle y,z\rangle}{\|y\|_2\|z\|_2}\right).
\end{aligned}$$

This completes the proof of Eq. (5). □

B Sketching Preliminaries: POLYSKETCH and SRHT

Our sketching algorithms use the Subsampled Randomized Hadamard Transform (SRHT) [2] to reduce the dimensionality of the intermediate vectors that arise in our computations. Next lemma gives the performance of SRHT sketches which is proved, for instance, in Theorem 9 of [13],

Lemma 2 (SRHT Sketch)**.** *For every positive integer d and every $\varepsilon,\delta>0$, there exists a distribution on random matrices $\boldsymbol{S}\in\mathbb{R}^{m\times d}$ with $m=\mathcal{O}\left(\frac{1}{\varepsilon^2}\cdot\log^2\frac{1}{\varepsilon\delta}\right)$, called* **SRHT sketch***, such that for any vector $x\in\mathbb{R}^d$, $\Pr\left[\|\boldsymbol{S}x\|_2^2\in(1\pm\varepsilon)\|x\|_2^2\right]\geq 1-\delta$. Moreover, $\boldsymbol{S}x$ can be computed in time $\mathcal{O}\left(\frac{1}{\varepsilon^2}\cdot\log^2\frac{1}{\varepsilon\delta}+d\log d\right)$.*

Now we restate the Lemma 1 and present the proof,

Lemma 1 (POLYSKETCH)**.** *For every integers $p,d\geq 1$ and every $\varepsilon,\delta>0$, there exists a distribution on random matrices $\boldsymbol{Q}^p\in\mathbb{R}^{m\times d^p}$, called degree p* POLYSKETCH *such that* **(1)** *for some $m=\mathcal{O}\left(\frac{p}{\varepsilon^2}\log^3\frac{1}{\varepsilon\delta}\right)$ and any $y\in\mathbb{R}^{d^p}$, $\Pr\left[\|\boldsymbol{Q}^p y\|_2^2\in(1\pm\varepsilon)\|y\|_2^2\right]\geq 1-\delta$;* **(2)** *for any $x\in\mathbb{R}^d$, if $e_1\in\mathbb{R}^d$ is the standard basis vector along the first coordinate, the total time to compute $\boldsymbol{Q}^p(x^{\otimes(p-j)}\otimes e_1^{\otimes j})$ for all $j=0,1,\ldots,p$ is $\mathcal{O}\left(pm\log^2 m+\min\left\{\frac{p^{3/2}}{\varepsilon}\log\frac{1}{\delta}\,\mathrm{nnz}(x), pd\log d\right\}\right)$;* **(3)** *for any collection of vectors $v_1,\ldots,v_p\in\mathbb{R}^d$, the time to compute $\boldsymbol{Q}^p(v_1\otimes\cdots\otimes v_p)$ is bounded by $\mathcal{O}\left(pm\log m+\frac{p^{3/2}}{\varepsilon}d\log\frac{1}{\delta}\right)$;* **(4)** *for any $\lambda>0$ and any matrix $\boldsymbol{A}\in\mathbb{R}^{d^p\times n}$, where the statistical dimension of $\boldsymbol{A}^\top\boldsymbol{A}$ is s_λ, there exists some $m=\mathcal{O}\left(\frac{p^4 s_\lambda}{\varepsilon^2}\log^3\frac{n}{\varepsilon\delta}\right)$ such that,*

$$\Pr\left[(1-\varepsilon)\left(\boldsymbol{A}^\top\boldsymbol{A}+\lambda\boldsymbol{I}\right)\preceq(\boldsymbol{Q}^p\boldsymbol{A})^\top(\boldsymbol{Q}^p\boldsymbol{A})+\lambda\boldsymbol{I}\preceq(1+\varepsilon)\left(\boldsymbol{A}^\top\boldsymbol{A}+\lambda\boldsymbol{I}\right)\right]\geq 1-\delta. \tag{1}$$

Proof of Lemma 1: The fourth statement of the lemma immediately follows from Theorem 1.3 of Ahle et al. [1]. Moreover, by invoking Theorem 1.2 of Ahle et al. [1], we find that there exists a random sketch $\boldsymbol{Q}^p\in\mathbb{R}^{m\times d^p}$ such that $m=C\cdot\frac{p}{\varepsilon^2}\log^3\frac{1}{\varepsilon\delta}$, for some absolute constant C, and for any $y\in\mathbb{R}^{d^p}$,

$$\Pr\left[\|\boldsymbol{Q}^p y\|_2^2\in(1\pm\varepsilon)\|y\|_2^2\right]\geq 1-\delta.$$

This immediately proves the first statement of the lemma.

As shown in [1], the sketch $\boldsymbol{Q}^p$ can be applied to tensor product vectors of the form $v_1\otimes v_2\otimes\ldots v_p$ by recursive application of $\mathcal{O}(p)$ independent instances of OSNAP transform [33] and a novel variant of the SRHT , proposed in [1] called TENSORSRHT, on vectors v_i and their sketched versions. The sketch $\boldsymbol{Q}^p$, as shown in Fig. 3, can be represented by a binary tree with p leaves where the leaves are OSNAP sketches and the internal nodes are the TENSORSRHT. The use of OSNAP in the leaves of this sketch structure ensures excellent runtime for sketching sparse input vectors. However, note that if the input vectors are not sparse, i.e., $\mathrm{nnz}(v_i)=\widetilde{\Omega}(d)$ for input vectors v_i, then we can simply remove the OSNAP transforms from the leaves of this structure and achieve improved runtime, without hurting the approximation guarantee. Therefore, the sketch $\boldsymbol{Q}^p$ that satisfies the statement of the lemma is exactly the one introduced in [1] for sparse input vectors and for non-sparse inputs is obtained by removing the OSNAP transforms from the leaves of the sketch structure given in Fig. 3.

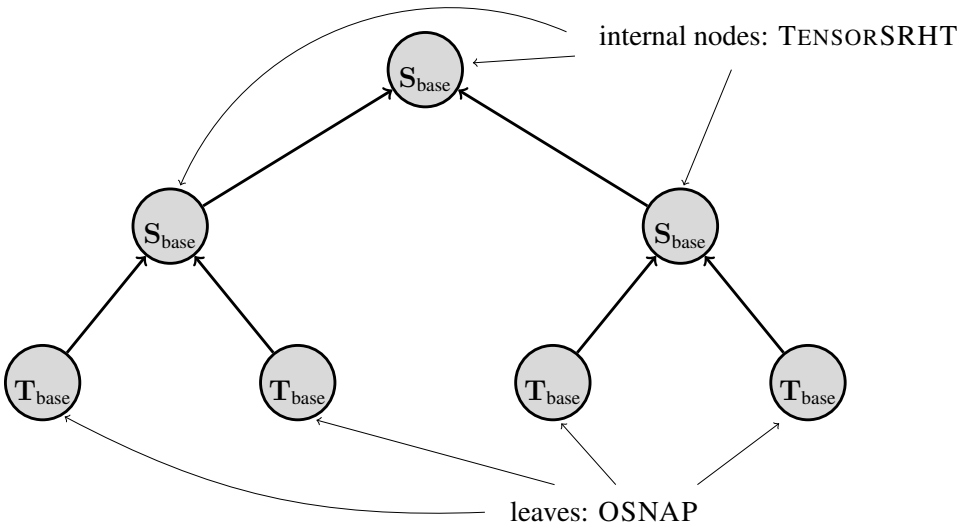


Figure 3: The structure of sketch $\boldsymbol{Q}^p$ proposed in Theorem 1.2 of [1]: the sketch matrices in nodes of the tree labeled with $\boldsymbol{S}_{\text{base}}$ and $\boldsymbol{T}_{\text{base}}$ are independent instances of TENSORSRHT and OSNAP, respectively.

Runtime analysis: By Theorem 1.2 of [1], for any vector $x \in \mathbb{R}^d$, $\boldsymbol{Q}^p x^{\otimes p}$ can be computed in time $\mathcal{O}\left(pm\log m + \frac{p^{3/2}}{\varepsilon}\log\frac{1}{\delta}\cdot\text{nnz}(x)\right)$. From the binary tree structure of the sketch, shown in Fig. 3, it follows that once we compute $\boldsymbol{Q}^p x^{\otimes p}$, then $\boldsymbol{Q}^p\left(x^{\otimes p-1}\otimes e_1\right)$ can be computed by updating the path from one of the leaves to the root of the binary tree which amounts to applying an instance of OSNAP transform on the e_1 vector and then applying $\mathcal{O}(\log p)$ instances of TENSORSRHT on the intermediate nodes of the tree. This can be computed in a total additional runtime of $\mathcal{O}(m\log m\log p)$. By this argument, it follows that $\boldsymbol{Q}^p\left(x^{\otimes p-j}\otimes e_1^j\right)$ can be computed sequentially for all $j = 0, 1, 2, \cdots p$ in total time $\mathcal{O}\left(pm\log p\log m + \frac{p^{3/2}}{\varepsilon}\log\frac{1}{\delta}\cdot\text{nnz}(x)\right)$. By plugging in the value $m = \mathcal{O}\left(\frac{p}{\varepsilon^2}\log^3\frac{1}{\varepsilon\delta}\right)$, this runtime will be $\mathcal{O}\left(\frac{p^2\log^2\frac{p}{\varepsilon}}{\varepsilon^2}\log^3\frac{1}{\varepsilon\delta} + \frac{p^{3/2}}{\varepsilon}\log\frac{1}{\delta}\cdot\text{nnz}(x)\right)$, which gives the second statement of the lemma for sparse input vectors x. If x is non-sparse, as we discussed in the above paragraph, we just need to omit the OSNAP transforms from our sketch construction which translates into a runtime of $\mathcal{O}\left(\frac{p^2\log^2\frac{p}{\varepsilon}}{\varepsilon^2}\log^3\frac{1}{\varepsilon\delta} + pd\log d\right)$. Therefore, the final runtime bound is $\mathcal{O}\left(\frac{p^2\log^2\frac{p}{\varepsilon}}{\varepsilon^2}\log^3\frac{1}{\varepsilon\delta} + \min\left\{\frac{p^{3/2}}{\varepsilon}\log\frac{1}{\delta}\cdot\text{nnz}(x), pd\log d\right\}\right)$, which proves the second statement of the lemma.

Furthermore, the sketch $\boldsymbol{Q}^p$ can be applied to tensor product of any collection of p vectors. The time to apply $\boldsymbol{Q}^p$ to the tensor product $v_1\otimes v_2\otimes\ldots v_p$ consists of time of applying OSNAP to each of the vectors $v_1, v_2\ldots v_p$ and time of applying $\mathcal{O}(p)$ instances of TENSORSRHT to intermediate vectors which are of size m. This runtime can be upper bounded by $\mathcal{O}\left(\frac{p^2\log\frac{p}{\varepsilon}}{\varepsilon^2}\log^3\frac{1}{\varepsilon\delta} + \frac{p^{3/2}}{\varepsilon}d\cdot\log\frac{1}{\delta}\right)$, which proves the third statement of the Lemma 1. □

C NTK Sketch: Claims and Invariants

We start by proving that the polynomials $P^{(p)}_{\texttt{relu}}(\cdot)$ and $\dot{P}^{(p')}_{\texttt{relu}}(\cdot)$ defined in Eq. (6) of Algorithm 1 closely approximate the arc-cosine functions $\kappa_1(\cdot)$ and $\kappa_0(\cdot)$ on the interval $[-1, 1]$.

Remark. Observe that $\kappa_0(\alpha) = \frac{d}{d\alpha}\left(\kappa_1(\alpha)\right)$.

Lemma 3 (Polynomial Approximations to κ_1 and κ_0)**.** *If we let $\kappa_1(\cdot)$ and $\kappa_0(\cdot)$ be defined as in Eq. (2) of Definition 1, then for any integer $p \geq \frac{1}{9\varepsilon^{2/3}}$, the polynomial $P^{(p)}_{\texttt{relu}}(\cdot)$ defined in Eq. (6) of*

Algorithm 1 satisfies,

$$\max_{\alpha\in[-1,1]}\left|P^{(p)}_{\texttt{relu}}(\alpha)-\kappa_1(\alpha)\right|\leq\varepsilon.$$

Moreover, for any integer $p'\geq\frac{1}{26\varepsilon^2}$, the polynomial $\dot{P}^{(p')}_{\texttt{relu}}(\cdot)$ defined as in Eq. (6) of Algorithm 1, satisfies,

$$\max_{\alpha\in[-1,1]}\left|\dot{P}^{(p')}_{\texttt{relu}}(\alpha)-\kappa_0(\alpha)\right|\leq\varepsilon.$$

Proof of Lemma 3: We start by Taylor series expansion of $\kappa_0(\cdot)$ around $\alpha=0$, $\kappa_0(\alpha)=\frac{1}{2}+\frac{1}{\pi}\cdot\sum_{i=0}^{\infty}\frac{(2i)!}{2^{2i}\cdot(i!)^2\cdot(2i+1)}\cdot\alpha^{2i+1}$. Therefore, we have

$$\begin{aligned}
\max_{\alpha\in[-1,1]}\left|\dot{P}^{(p')}_{\texttt{relu}}(\alpha)-\kappa_0(\alpha)\right| &= \frac{1}{\pi}\cdot\sum_{i=p'+1}^{\infty}\frac{(2i)!}{2^{2i}\cdot(i!)^2\cdot(2i+1)}\\
&\leq \frac{1}{\pi}\cdot\sum_{i=p'+1}^{\infty}\frac{e\cdot e^{-2i}\cdot(2i)^{2i+1/2}}{2\pi\cdot 2^{2i}\cdot e^{-2i}\cdot n^{2i+1}\cdot(2i+1)}\\
&= \frac{e}{\sqrt{2}\pi^2}\cdot\sum_{i=p'+1}^{\infty}\frac{1}{\sqrt{i}\cdot(2i+1)}\\
&\leq \frac{e}{\sqrt{2}\pi^2}\cdot\int_{p'}^{\infty}\frac{1}{\sqrt{x}\cdot(2x+1)}dx\\
&\leq \frac{e}{\sqrt{2}\pi^2}\cdot\frac{1}{\sqrt{p'}}\leq\varepsilon.
\end{aligned}$$

To prove the second part of the lemma, we consider the Taylor expansion of $\kappa_1(\cdot)$ at $\alpha=0$. Since $\kappa_0(\alpha)=\frac{d}{d\alpha}\left(\kappa_1(\alpha)\right)$, the Taylor series of $\kappa_1(\alpha)$ can be obtained from the Taylor series of $\kappa_0(\alpha)$ as follows,

$$\kappa_1(\alpha)=\frac{1}{\pi}+\frac{\alpha}{2}+\frac{1}{\pi}\cdot\sum_{i=0}^{\infty}\frac{(2i)!}{2^{2i}\cdot(i!)^2\cdot(2i+1)\cdot(2i+2)}\cdot\alpha^{2i+2}.$$

Hence, we have

$$\begin{aligned}
\max_{\alpha\in[-1,1]}\left|P^{(p)}_{\texttt{relu}}(\alpha)-\kappa_1(\alpha)\right| &= \frac{1}{\pi}\cdot\sum_{i=p+1}^{\infty}\frac{(2i)!}{2^{2i}\cdot(i!)^2\cdot(2i+1)\cdot(2i+2)}\\
&\leq \frac{1}{\pi}\cdot\sum_{i=p+1}^{\infty}\frac{e\cdot e^{-2i}\cdot(2i)^{2i+1/2}}{2\pi\cdot 2^{2i}\cdot e^{-2i}\cdot n^{2i+1}\cdot(2i+1)\cdot(2i+2)}\\
&= \frac{e}{\sqrt{2}\pi^2}\cdot\sum_{i=p+1}^{\infty}\frac{1}{\sqrt{i}\cdot(2i+1)\cdot(2i+2)}\\
&\leq \frac{e}{\sqrt{2}\pi^2}\cdot\int_{p}^{\infty}\frac{1}{\sqrt{x}\cdot(2x+1)\cdot(2x+2)}dx\\
&\leq \frac{e}{\sqrt{2}\pi^2}\cdot\frac{1}{6\cdot p^{3/2}}\leq\varepsilon.
\end{aligned}$$

This completes the proof of Lemma 3. □.

Therefore, it is possible to approximate the function $\kappa_0(\cdot)$ up to error ε using a polynomial of degree $\mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$. Also if we want to approximate $\kappa_1(\cdot)$ using a polynomial up to error ε on the interval $[-1,1]$, it suffices to use a polynomial of degree $\mathcal{O}\left(\frac{1}{\varepsilon^{2/3}}\right)$. One can see that since the Taylor expansions of κ_1 and κ_0 contain non-negative coefficients only, both of these functions are positive definite. Additionally, the polynomial approximations $P^{(p)}_{\texttt{relu}}$ and $\dot{P}^{(p')}_{\texttt{relu}}$ given in Eq. (6) of Algorithm 1 are positive definite functions.

In order to prove Theorem 1, we also need the following lemma on the error sensitivity of polynomials $P^{(p)}_{\texttt{relu}}$ and $\dot{P}^{(p')}_{\texttt{relu}}$,

Lemma 4 (Sensitivity of $P^{(p)}_{\texttt{relu}}$ and $\dot{P}^{(p)}_{\texttt{relu}}$)**.** *For any integer $p \geq 3$, any $\alpha \in [-1, 1]$, and any α' such that $|\alpha - \alpha'| \leq \frac{1}{6p}$, if we let the polynomials $P^{(p)}_{\texttt{relu}}(\alpha)$ and $\dot{P}^{(p)}_{\texttt{relu}}(\alpha)$ be defined as in Eq. (6) of Algorithm 1, then*

$$\left|P^{(p)}_{\texttt{relu}}(\alpha) - P^{(p)}_{\texttt{relu}}(\alpha')\right| \leq |\alpha - \alpha'|,$$

and

$$\left|\dot{P}^{(p)}_{\texttt{relu}}(\alpha) - \dot{P}^{(p)}_{\texttt{relu}}(\alpha')\right| \leq \sqrt{p} \cdot |\alpha - \alpha'|.$$

Proof of Lemma 4: Note that an α' that satisfies the preconditions of the lemm, is in the range $\left[-1 - \frac{1}{6p}, 1 + \frac{1}{6p}\right]$. Now we bound the derivative of the polynomial $\dot{P}^{(p)}_{\texttt{relu}}$ on the interval $\left[-1 - \frac{1}{6p}, 1 + \frac{1}{6p}\right]$,

$$\begin{aligned}
\max_{\alpha \in \left[-1-\frac{1}{6p}, 1+\frac{1}{6p}\right]} \left|\frac{d}{d\alpha}\left(\dot{P}^{(p)}_{\texttt{relu}}(\alpha)\right)\right| &= \frac{1}{\pi} \cdot \sum_{i=0}^{p} \frac{(2i)!}{2^{2i} \cdot (i!)^2} \cdot \left(1 + \frac{1}{6p}\right)^{2i} \\
&\leq \frac{1}{\pi} + \frac{e^{4/3}}{\sqrt{2}\pi^2} \cdot \sum_{i=1}^{p} \frac{1}{\sqrt{i}} \\
&\leq \frac{1}{\pi} + \frac{e^{4/3}}{\sqrt{2}\pi^2} \cdot \int_0^p \frac{1}{\sqrt{x}} dx \\
&\leq \sqrt{p},
\end{aligned}$$

therefore, the second statement of lemma holds.

To prove the first statement of lemma, we bound the derivative of the polynomial $P^{(p)}_{\texttt{relu}}$ on the interval $\left[-1 - \frac{1}{6p}, 1 + \frac{1}{6p}\right]$ as follows,

$$\begin{aligned}
\max_{\alpha \in \left[-1-\frac{1}{6p}, 1+\frac{1}{6p}\right]} \left|\frac{d}{d\alpha}\left(P^{(p)}_{\texttt{relu}}(\alpha)\right)\right| &= \frac{1}{\pi} \cdot \sum_{i=0}^{p} \frac{(2i)!}{2^{2i} \cdot (i!)^2 \cdot (2i+1)} \cdot \left(1 + \frac{1}{6p}\right)^{2i+1} \\
&\leq \frac{19}{18\pi} + \frac{e^{25/18}}{\sqrt{2}\pi^2} \cdot \sum_{i=1}^{p} \frac{1}{\sqrt{i} \cdot (2i+1)} \\
&\leq \frac{19}{18\pi} + \frac{e^{25/18}}{\sqrt{2}\pi^2} \cdot \int_0^p \frac{1}{\sqrt{x} \cdot (2x+1)} dx \\
&\leq 1,
\end{aligned}$$

therefore, the second statement of the lemma follows. This completes the proof of Lemma 4. □

For the rest of this section, we need two basic properties of tensor products and direct sums:

$$\langle x \otimes y, z \otimes w \rangle = \langle x, z \rangle \cdot \langle y, w \rangle, \quad \langle x \oplus y, z \oplus w \rangle = \langle x, z \rangle + \langle y, w \rangle \tag{24}$$

for vectors x, y, z, w with conforming sizes.

Now we are in a position to analyze the invariants that are maintained throughout the execution of NTKSKETCH (Algorithm 1):

Lemma 5 (Invariants of the NTKSKETCH algorithm)**.** *For every positive integers d and L, every $\varepsilon, \delta > 0$, every vectors $y, z \in \mathbb{R}^d$, if we let $\Sigma^{(h)}_{\texttt{relu}} : [-1, 1] \to \mathbb{R}$ and $K^{(h)}_{\texttt{relu}} : [-1, 1] \to \mathbb{R}$ be the functions defined in Eq. (3) and Eq. (4) of Algorithm 1, then with probability at least $1 - \delta$ the following invariants hold for every $h = 0, 1, 2, \ldots L$:*

1. *The mapping $\phi^{(h)}(\cdot)$ computed by* NTKSKETCH *in line 4 and Eq. (7) of Algorithm 1 satisfy*

$$\left|\left\langle \phi^{(h)}(y), \phi^{(h)}(z) \right\rangle - \Sigma^{(h)}_{\texttt{relu}}\left(\frac{\langle y, z \rangle}{\|y\|_2 \|z\|_2}\right)\right| \leq (h+1) \cdot \frac{\varepsilon^2}{60L^3}.$$

2. *The mapping $\psi^{(h)}(\cdot)$ computed by* NTKSKETCH *in line 5 and Eq. (9) of Algorithm 1 satisfy*

$$\left|\left\langle \psi^{(h)}(y), \psi^{(h)}(z) \right\rangle - K^{(h)}_{\texttt{relu}}\left(\frac{\langle y, z \rangle}{\|y\|_2 \|z\|_2}\right)\right| \leq \varepsilon \cdot \frac{h^2 + 1}{10L}.$$

Proof of Lemma 5: The proof is by induction on the value of $h = 0, 1, 2, \dots L$. More formally, consider the following statements for every $h = 0, 1, 2, \dots L$:

$\mathbf{P_1(h)}$:

$$\left|\left\langle \phi^{(h)}(y), \phi^{(h)}(z)\right\rangle - \Sigma^{(h)}_{\texttt{relu}}\left(\frac{\langle y, z\rangle}{\|y\|_2\|z\|_2}\right)\right| \le (h+1)\cdot\frac{\varepsilon^2}{60L^3},$$
$$\left|\left\|\phi^{(h)}(y)\right\|_2^2 - 1\right| \le (h+1)\cdot\frac{\varepsilon^2}{60L^3}, \text{ and } \left|\left\|\phi^{(h)}(z)\right\|_2^2 - 1\right| \le (h+1)\cdot\frac{\varepsilon^2}{60L^3}.$$

$\mathbf{P_2(h)}$:

$$\left|\left\langle \psi^{(h)}(y), \psi^{(h)}(z)\right\rangle - K^{(h)}_{\texttt{relu}}\left(\frac{\langle y, z\rangle}{\|y\|_2\|z\|_2}\right)\right| \le \varepsilon\cdot\frac{h^2+1}{10L},$$
$$\left|\left\|\psi^{(h)}(y)\right\|_2^2 - K^{(h)}_{\texttt{relu}}(1)\right| \le \varepsilon\cdot\frac{h^2+1}{10L}, \text{ and } \left|\left\|\psi^{(h)}(z)\right\|_2^2 - K^{(h)}_{\texttt{relu}}(1)\right| \le \varepsilon\cdot\frac{h^2+1}{10L}.$$

We prove that the following holds,

$$\Pr[P_1(0)] \ge 1 - \mathcal{O}\left(\frac{\delta}{L}\right), \quad \text{and} \quad \Pr[P_2(0)|P_1(0)] \ge 1 - \mathcal{O}\left(\frac{\delta}{L}\right). \tag{25}$$

Additionally, by induction, we prove that for every $h = 1, 2, \dots L$,

$$\Pr\left[P_1(h)|P_1(h-1)\right] \ge 1 - \mathcal{O}\left(\frac{\delta}{L}\right), \quad \text{and}$$
$$\Pr\left[P_2(h)|P_2(h-1), P_1(h), P_1(h-1)\right] \ge 1 - \mathcal{O}\left(\frac{\delta}{L}\right). \tag{26}$$

(1) Base of induction ($h = 0$): By line 4 of Algorithm 1, $\phi^{(0)}(y) = \frac{1}{\|y\|_2}\cdot \boldsymbol{S}\cdot\boldsymbol{Q}^1\cdot y$ and $\phi^{(0)}(z) = \frac{1}{\|z\|_2}\cdot \boldsymbol{S}\cdot\boldsymbol{Q}^1\cdot z$, thus, Lemma 2 implies the following

$$\Pr\left[\left|\left\langle \phi^{(0)}(y), \phi^{(0)}(z)\right\rangle - \frac{\left\langle \boldsymbol{Q}^1 y, \boldsymbol{Q}^1 z\right\rangle}{\|y\|_2\|z\|_2}\right| \le \mathcal{O}\left(\frac{\varepsilon^2}{L^3}\right)\cdot\frac{\left\|\boldsymbol{Q}^1 y\right\|_2\left\|\boldsymbol{Q}^1 z\right\|_2}{\|y\|_2\|z\|_2}\right] \ge 1 - \mathcal{O}\left(\frac{\delta}{L}\right). \tag{27}$$

By using the above together with Lemma 1 and union bound as well as triangle inequality, we have

$$\Pr\left[\left|\left\langle \phi^{(0)}(y), \phi^{(0)}(z)\right\rangle - \frac{\langle y, z\rangle}{\|y\|_2\|z\|_2}\right| \le \mathcal{O}\left(\frac{\varepsilon^2}{L^3}\right)\right] \ge 1 - \mathcal{O}\left(\frac{\delta}{L}\right). \tag{28}$$

Similarly, we can prove that

$$\Pr\left[\left|\left\|\phi^{(0)}(y)\right\|_2^2 - 1\right| \le \mathcal{O}\left(\frac{\varepsilon^2}{L^3}\right)\right] \ge 1 - \mathcal{O}\left(\frac{\delta}{L}\right), \text{ and}$$
$$\Pr\left[\left|\left\|\phi^{(0)}(z)\right\|_2^2 - 1\right| \le \mathcal{O}\left(\frac{\varepsilon^2}{L^3}\right)\right] \ge 1 - \mathcal{O}\left(\frac{\delta}{L}\right).$$

Using union bound, this proves the base of induction for statement $P_1(0)$, i.e.,

$$\Pr[P_1(0)] \ge 1 - \mathcal{O}\left(\frac{\delta}{L}\right). \tag{29}$$

Moreover, by line 5 of Algorithm 1, $\psi^{(0)}(y) = \boldsymbol{V}\cdot\phi^{(0)}(y)$ and $\psi^{(0)}(z) = \boldsymbol{V}\cdot\phi^{(0)}(z)$, thus, Lemma 2 implies that,

$$\Pr\left[\left|\left\langle \psi^{(0)}(y), \psi^{(0)}(z)\right\rangle - \left\langle \phi^{(0)}(y), \phi^{(0)}(z)\right\rangle\right| \le \mathcal{O}\left(\frac{\varepsilon}{L}\right)\cdot\left\|\phi^{(0)}(y)\right\|_2\left\|\phi^{(0)}(z)\right\|_2\right] \ge 1 - \mathcal{O}\left(\frac{\delta}{L}\right). \tag{30}$$

By conditioning on $P_1(0)$ and using the above together with triangle inequality it follows that,

$$\Pr\left[\left|\left\langle \psi^{(0)}(y), \psi^{(0)}(z)\right\rangle - K^{(0)}_{\texttt{relu}}\left(\frac{\langle y,z\rangle}{\|y\|_2\|z\|_2}\right)\right| \le \frac{\varepsilon}{10L}\right] \ge 1-\mathcal{O}\left(\frac{\delta}{L}\right). \tag{31}$$

Similarly we can prove that with probability $1-\mathcal{O}\left(\frac{\delta}{L}\right)$ we have $\left|\left\|\psi^{(0)}(y)\right\|_2^2 - K^{(0)}_{\texttt{relu}}(1)\right| \le \frac{\varepsilon}{10L}$ and $\left|\left\|\psi^{(0)}(z)\right\|_2^2 - K^{(0)}_{\texttt{relu}}(1)\right| \le \frac{\varepsilon}{10L}$, which proves the base of induction for the second statement, i.e., $\Pr[P_2(0)|P_1(0)] \ge 1-\mathcal{O}(\delta/L)$. This completes the base of induction.

(2) Inductive step: Assume that the inductive hypothesis holds for $h-1$. First, note that by Lemma 2 and using Eq. (7) we have the following,

$$\Pr\left[\left|\left\langle \phi^{(h)}(y), \phi^{(h)}(z)\right\rangle - \sum_{j=0}^{2p+2} c_j \left\langle Z_j^{(h)}(y), Z_j^{(h)}(z)\right\rangle\right| \le \mathcal{O}\left(\frac{\varepsilon^2}{L^3}\right)\cdot A\right] \ge 1-\mathcal{O}\left(\frac{\delta}{L}\right), \tag{32}$$

where $A := \sqrt{\sum_{j=0}^{2p+2} c_j \|Z_j^{(h)}(y)\|_2^2}\cdot\sqrt{\sum_{j=0}^{2p+2} c_j \|Z_j^{(h)}(z)\|_2^2}$ and the collection of vectors $\left\{Z_j^{(h)}(y)\right\}_{j=0}^{2p+2}$ and $\left\{Z_j^{(h)}(z)\right\}_{j=0}^{2p+2}$ and coefficients $\{c_j\}_{j=0}^{2p+2}$ are defined as per Eq. (7) and Eq. (6), respectively.

By Lemma 1 together with union bound, the following inequalities simultaneously hold for all $j=0,\ldots,2p+2$, with probability at least $1-\mathcal{O}(\delta/L)$:

$$\begin{aligned}
&\left|\left\langle Z_j^{(h)}(y), Z_j^{(h)}(z)\right\rangle - \left\langle \phi^{(h-1)}(y), \phi^{(h-1)}(z)\right\rangle^j\right| \le \mathcal{O}\left(\frac{\varepsilon^2}{L^3}\right)\cdot\left\|\phi^{(h-1)}(y)\right\|_2^j\left\|\phi^{(h-1)}(z)\right\|_2^j\\
&\left\|Z_j^{(h)}(y)\right\|_2^2 \le \frac{11}{10}\cdot\left\|\phi^{(h-1)}(y)\right\|_2^{2j}\\
&\left\|Z_j^{(h)}(z)\right\|_2^2 \le \frac{11}{10}\cdot\left\|\phi^{(h-1)}(z)\right\|_2^{2j}
\end{aligned} \tag{33}$$

Therefore, by plugging Eq. (33) to Eq. (32) and using union bound, triangle inequality and Cauchy–Schwarz inequality we find that,

$$\Pr\left[\left|\left\langle \phi^{(h)}(y), \phi^{(h)}(z)\right\rangle - P^{(p)}_{\texttt{relu}}\left(\left\langle \phi^{(h-1)}(y), \phi^{(h-1)}(z)\right\rangle\right)\right| \le \mathcal{O}\left(\frac{\varepsilon^2}{L^3}\right)\cdot B\right] \ge 1-\mathcal{O}\left(\frac{\delta}{L}\right), \tag{34}$$

where $B := \sqrt{P^{(p)}_{\texttt{relu}}\left(\|\phi^{(h-1)}(y)\|_2^2\right)\cdot P^{(p)}_{\texttt{relu}}\left(\|\phi^{(h-1)}(z)\|_2^2\right)}$ and $P^{(p)}_{\texttt{relu}}(\alpha) = \sum_{j=0}^{2p+2} c_j\cdot\alpha^j$ is the polynomial defined in Eq. (6). Using the inductive hypothesis $P_1(h-1)$, we have that

$$\left|\left\|\phi^{(h-1)}(y)\right\|_2^2 - 1\right| \le h\cdot\frac{\varepsilon^2}{60L^3}, \quad \text{and} \quad \left|\left\|\phi^{(h-1)}(z)\right\|_2^2 - 1\right| \le h\cdot\frac{\varepsilon^2}{60L^3}. \tag{35}$$

Therefore, by Lemma 4 we have $\left|P^{(p)}_{\texttt{relu}}\left(\|\phi^{(h-1)}(y)\|_2^2\right) - P^{(p)}_{\texttt{relu}}(1)\right| \le h\cdot\frac{\varepsilon^2}{60L^3}$ and $\left|P^{(p)}_{\texttt{relu}}\left(\|\phi^{(h-1)}(z)\|_2^2\right) - P^{(p)}_{\texttt{relu}}(1)\right| \le h\cdot\frac{\varepsilon^2}{60L^3}$. Consequently, since $P^{(p)}_{\texttt{relu}}(1) \le P^{(+\infty)}_{\texttt{relu}}(1) = 1$, we obtain that $B \le \frac{11}{10}$. By plugging this into Eq. (34) we have,

$$\Pr\left[\left|\left\langle \phi^{(h)}(y), \phi^{(h)}(z)\right\rangle - P^{(p)}_{\texttt{relu}}\left(\left\langle \phi^{(h-1)}(y), \phi^{(h-1)}(z)\right\rangle\right)\right| \le \mathcal{O}\left(\frac{\varepsilon^2}{L^3}\right)\right] \ge 1-\mathcal{O}\left(\frac{\delta}{L}\right). \tag{36}$$

Furthermore, the inductive hypothesis $P_1(h-1)$ implies that

$$\left|\left\langle \phi^{(h-1)}(y), \phi^{(h-1)}(z)\right\rangle - \Sigma^{(h-1)}_{\texttt{relu}}\left(\frac{\langle y,z\rangle}{\|y\|_2\|z\|_2}\right)\right| \le h\cdot\frac{\varepsilon^2}{60L^3}. \tag{37}$$

Hence, using Lemma 4 we find that,

$$\left|P^{(p)}_{\texttt{relu}}\left(\left\langle \phi^{(h-1)}(y), \phi^{(h-1)}(z)\right\rangle\right) - P^{(p)}_{\texttt{relu}}\left(\Sigma^{(h-1)}_{\texttt{relu}}\left(\frac{\langle y,z\rangle}{\|y\|_2\|z\|_2}\right)\right)\right| \le h\cdot\frac{\varepsilon^2}{60L^3}. \tag{38}$$

By incorporating the above inequality into Eq. (36) using triangle inequality we find that,

$$\Pr\left[\left|\left\langle\phi^{(h)}(y),\phi^{(h)}(z)\right\rangle - P^{(p)}_{\texttt{relu}}\left(\Sigma^{(h-1)}_{\texttt{relu}}\left(\frac{\langle y,z\rangle}{\|y\|_2\|z\|_2}\right)\right)\right| \le \frac{h\cdot\varepsilon^2}{60L^3} + \mathcal{O}\left(\frac{\varepsilon^2}{L^3}\right)\right] \ge 1-\mathcal{O}\left(\frac{\delta}{L}\right). \quad (39)$$

Now, by invoking Lemma 3 and using the fact that $p = \lceil 2L^2/\varepsilon^{4/3}\rceil$ we have,

$$\left|P^{(p)}_{\texttt{relu}}\left(\Sigma^{(h-1)}_{\texttt{relu}}\left(\frac{\langle y,z\rangle}{\|y\|_2\|z\|_2}\right)\right) - \kappa_1\left(\Sigma^{(h-1)}_{\texttt{relu}}\left(\frac{\langle y,z\rangle}{\|y\|_2\|z\|_2}\right)\right)\right| \le \frac{\varepsilon^2}{76L^3}. \quad (40)$$

By combining the above inequality with Eq. (39) using triangle inequality and using the fact that $\kappa_1\left(\Sigma^{(h-1)}_{\texttt{relu}}\left(\frac{\langle y,z\rangle}{\|y\|_2\|z\|_2}\right)\right) = \Sigma^{(h)}_{\texttt{relu}}\left(\frac{\langle y,z\rangle}{\|y\|_2\|z\|_2}\right)$ (by Eq. (3)), we get the following bound,

$$\Pr\left[\left|\left\langle\phi^{(h)}(y),\phi^{(h)}(z)\right\rangle - \Sigma^{(h)}_{\texttt{relu}}\left(\frac{\langle y,z\rangle}{\|y\|_2\|z\|_2}\right)\right| \le (h+1)\cdot\frac{\varepsilon^2}{60L^3}\right] \ge 1-\mathcal{O}\left(\frac{\delta}{L}\right). \quad (41)$$

Similarly, we can prove that

$$\Pr\left[\left|\left\|\phi^{(h)}(y)\right\|_2^2 - 1\right| > \frac{(h+1)\cdot\varepsilon^2}{60L^3}\right] \le \mathcal{O}\left(\frac{\delta}{L}\right), \text{ and}$$
$$\Pr\left[\left|\left\|\phi^{(h)}(z)\right\|_2^2 - 1\right| > \frac{(h+1)\cdot\varepsilon^2}{60L^3}\right] \le \mathcal{O}\left(\frac{\delta}{L}\right).$$

This is sufficient to prove the inductive step by union bound, i.e., $\Pr[P_1(h)|P_1(h-1)] \ge 1-\mathcal{O}\left(\frac{\delta}{L}\right)$.

Now we prove the inductive step for statement $P_2(h)$, that is, we prove that conditioned on $P_2(h-1), P_1(h), P_1(h-1)$, statement $P_2(h)$ holds with probability at least $1-\mathcal{O}\left(\frac{\delta}{L}\right)$. First, note that by Lemma 2 and using Eq. (8) we have,

$$\Pr\left[\left|\left\langle\dot\phi^{(h)}(y),\dot\phi^{(h)}(z)\right\rangle - \sum_{j=0}^{2p'+1} b_j\left\langle Y_j^{(h)}(y),Y_j^{(h)}(z)\right\rangle\right| \le O\left(\frac{\varepsilon}{L}\right)\cdot\widehat{A}\right] \ge 1-\mathcal{O}\left(\frac{\delta}{L}\right), \quad (42)$$

where $\widehat{A} := \sqrt{\sum_{j=0}^{2p'+1} b_j\|Y_j^{(h)}(y)\|_2^2}\cdot\sqrt{\sum_{j=0}^{2p'+1} b_j\|Y_j^{(h)}(z)\|_2^2}$ and the collection of vectors $\left\{Y_j^{(h)}(y)\right\}_{j=0}^{2p'+1}$ and $\left\{Y_j^{(h)}(z)\right\}_{j=0}^{2p'+1}$ and coefficients $\{b_j\}_{j=0}^{2p'+1}$ are defined as per Eq. (8) and Eq. (6), respectively. By invoking Lemma 1 along with union bound, with probability at least $1-\mathcal{O}\left(\frac{\delta}{L}\right)$, the following inequalities hold true simultaneously for all $j = 0,1,\dots 2p'+1$

$$\begin{aligned}
&\left|\left\langle Y_j^{(h)}(y),Y_j^{(h)}(z)\right\rangle - \left\langle\phi^{(h-1)}(y),\phi^{(h-1)}(z)\right\rangle^j\right| \le \mathcal{O}\left(\frac{\varepsilon}{L}\right)\cdot\left\|\phi^{(h-1)}(y)\right\|_2^j\left\|\phi^{(h-1)}(z)\right\|_2^j\\
&\left\|Y_j^{(h)}(y)\right\|_2^2 \le \frac{11}{10}\cdot\left\|\phi^{(h-1)}(y)\right\|_2^{2j}\\
&\left\|Y_j^{(h)}(z)\right\|_2^2 \le \frac{11}{10}\cdot\left\|\phi^{(h-1)}(z)\right\|_2^{2j}
\end{aligned} \quad (43)$$

Therefore, by plugging Eq. (43) into Eq. (42) and using union bound, triangle inequality and Cauchy–Schwarz inequality we find that,

$$\Pr\left[\left|\left\langle\dot\phi^{(h)}(y),\dot\phi^{(h)}(z)\right\rangle - \dot P^{(p')}_{\texttt{relu}}\left(\left\langle\phi^{(h-1)}(y),\phi^{(h-1)}(z)\right\rangle\right)\right| \le \mathcal{O}\left(\frac{\varepsilon}{L}\right)\cdot\widehat{B}\right] \ge 1-\mathcal{O}\left(\frac{\delta}{L}\right), \quad (44)$$

where $\widehat{B} := \sqrt{\dot P^{(p')}_{\texttt{relu}}\left(\|\phi^{(h-1)}(y)\|_2^2\right)\cdot\dot P^{(p')}_{\texttt{relu}}\left(\|\phi^{(h-1)}(z)\|_2^2\right)}$ and $\dot P^{(p')}_{\texttt{relu}}(\alpha) = \sum_{j=0}^{2p'+1} b_j\cdot\alpha^j$ is the polynomial defined in Eq. (6). By inductive hypothesis $P_1(h-1)$ we have $\left|\left\|\phi^{(h-1)}(y)\right\|_2^2 - 1\right| \le h\cdot\frac{\varepsilon^2}{60L^3}$ and $\left|\left\|\phi^{(h-1)}(z)\right\|_2^2 - 1\right| \le h\cdot\frac{\varepsilon^2}{60L^3}$. Therefore, using the fact that $p' = \lceil 9L^2/\varepsilon^2\rceil$ and Lemma 4, $\left|\dot P^{(p')}_{\texttt{relu}}\left(\|\phi^{(h-1)}(y)\|_2^2\right) - \dot P^{(p')}_{\texttt{relu}}(1)\right| \le \frac{h\cdot\varepsilon}{20L^2}$ and $\left|\dot P^{(p')}_{\texttt{relu}}\left(\|\phi^{(h-1)}(z)\|_2^2\right) - \dot P^{(p')}_{\texttt{relu}}(1)\right| \le \frac{h\cdot\varepsilon}{20L^2}$. Consequently, since $\dot P^{(p')}_{\texttt{relu}}(1) \le \dot P^{(+\infty)}_{\texttt{relu}}(1) = 1$, we find that $\widehat{B} \le \frac{11}{10}$. By plugging this into Eq. (44) we have,

$$\Pr\left[\left|\left\langle\dot\phi^{(h)}(y),\dot\phi^{(h)}(z)\right\rangle - \dot P^{(p')}_{\texttt{relu}}\left(\left\langle\phi^{(h-1)}(y),\phi^{(h-1)}(z)\right\rangle\right)\right| \le \mathcal{O}\left(\frac{\varepsilon}{L}\right)\right] \ge 1-\mathcal{O}\left(\frac{\delta}{L}\right). \quad (45)$$

Furthermore, inductive hypothesis $P_1(h-1)$ implies $\left|\left\langle \phi^{(h-1)}(y), \phi^{(h-1)}(z)\right\rangle - \Sigma_{\texttt{relu}}^{(h-1)}\left(\frac{\langle y,z\rangle}{\|y\|_2\|z\|_2}\right)\right| \leq h \cdot \frac{\varepsilon^2}{60L^3}$, hence, by invoking Lemma 4 we find that,

$$\left|\dot{P}_{\texttt{relu}}^{(p')}\left(\left\langle \phi^{(h-1)}(y), \phi^{(h-1)}(z)\right\rangle\right) - \dot{P}_{\texttt{relu}}^{(p')}\left(\Sigma_{\texttt{relu}}^{(h-1)}\left(\frac{\langle y,z\rangle}{\|y\|_2\|z\|_2}\right)\right)\right| \leq \frac{h\cdot\varepsilon}{20L^2}.$$

By plugging the above inequality into Eq. (45) using triangle inequality, we find that,

$$\Pr\left[\left|\left\langle \dot{\phi}^{(h)}(y), \dot{\phi}^{(h)}(z)\right\rangle - \dot{P}_{\texttt{relu}}^{(p')}\left(\Sigma_{\texttt{relu}}^{(h-1)}\left(\frac{\langle y,z\rangle}{\|y\|_2\|z\|_2}\right)\right)\right| \leq \frac{h\cdot\varepsilon}{20L^2} + \mathcal{O}\left(\frac{\varepsilon}{L}\right)\right] \geq 1 - \mathcal{O}\left(\frac{\delta}{L}\right). \quad (46)$$

Now, by invoking Lemma 3 and using the fact that $p' = \lceil 9L^2/\varepsilon^2 \rceil$ we have,

$$\left|\dot{P}_{\texttt{relu}}^{(p')}\left(\Sigma_{\texttt{relu}}^{(h-1)}\left(\frac{\langle y,z\rangle}{\|y\|_2\|z\|_2}\right)\right) - \kappa_0\left(\Sigma_{\texttt{relu}}^{(h-1)}\left(\frac{\langle y,z\rangle}{\|y\|_2\|z\|_2}\right)\right)\right| \leq \frac{\varepsilon}{15L}. \quad (47)$$

By combining the above inequality with Eq. (46) using triangle inequality and using the fact that $\kappa_0\left(\Sigma_{\texttt{relu}}^{(h-1)}\left(\frac{\langle y,z\rangle}{\|y\|_2\|z\|_2}\right)\right) = \dot{\Sigma}_{\texttt{relu}}^{(h)}\left(\frac{\langle y,z\rangle}{\|y\|_2\|z\|_2}\right)$ (by Eq. (3)), we get the following bound,

$$\Pr\left[\left|\left\langle \dot{\phi}^{(h)}(y), \dot{\phi}^{(h)}(z)\right\rangle - \dot{\Sigma}_{\texttt{relu}}^{(h)}\left(\frac{\langle y,z\rangle}{\|y\|_2\|z\|_2}\right)\right| \leq \frac{\varepsilon}{8L}\right] \geq 1 - \mathcal{O}\left(\frac{\delta}{L}\right). \quad (48)$$

Similarly we can show that,

$$\Pr\left[\left|\left\|\dot{\phi}^{(h)}(y)\right\| - 1\right| > \frac{\varepsilon}{8L}\right] \leq \mathcal{O}\left(\frac{\delta}{L}\right), \text{ and } \Pr\left[\left|\left\|\dot{\phi}^{(h)}(z)\right\| - 1\right| > \frac{\varepsilon}{8L}\right] \leq \mathcal{O}\left(\frac{\delta}{L}\right). \quad (49)$$

Now let $f := \psi^{(h-1)}(y) \otimes \dot{\phi}^{(h)}(y)$ and $g := \psi^{(h-1)}(z) \otimes \dot{\phi}^{(h)}(z)$. Then by Lemma 2 and using Eq. (9) we have the following,

$$\Pr\left[\left|\left\langle \psi^{(h)}(y), \psi^{(h)}(z)\right\rangle - \left\langle \boldsymbol{Q}^2 f \oplus \phi^{(h)}(y), \boldsymbol{Q}^2 g \oplus \phi^{(h)}(z)\right\rangle\right| \leq \mathcal{O}\left(\frac{\varepsilon}{L}\right)\cdot D\right] \geq 1 - \mathcal{O}\left(\frac{\delta}{L}\right), \quad (50)$$

where $D := \left\|\boldsymbol{Q}^2 f \oplus \phi^{(h)}(y)\right\|_2 \left\|\boldsymbol{Q}^2 g \oplus \phi^{(h)}(z)\right\|_2$. By the fact that we conditioned on $P_1(h)$,

$$D \leq \sqrt{\|\boldsymbol{Q}^2 f\|_2^2 + \frac{11}{10}} \cdot \sqrt{\|\boldsymbol{Q}^2 g\|_2^2 + \frac{11}{10}}.$$

By Lemma 1, we can further obtain an upper bound:

$$D \leq \frac{11}{10} \cdot \sqrt{\|f\|_2^2 + 1} \cdot \sqrt{\|g\|_2^2 + 1}.$$

Now note that because we conditioned on $P_2(h-1)$ and using Eq. (49), with probability at least $1 - \mathcal{O}\left(\frac{\delta}{L}\right)$ the following holds:

$$\|f\|_2^2 = \left\|\psi^{(h-1)}(y)\right\|_2^2 \left\|\dot{\phi}^{(h)}(y)\right\|_2^2 \leq \frac{11}{10} \cdot K_{\texttt{relu}}^{(h-1)}(1) = \frac{11}{10}h.$$

Similarly, $\|g\|_2^2 \leq \frac{11}{10}h$ with probability at least $1 - \mathcal{O}\left(\frac{\delta}{L}\right)$, thus, by union bound:

$$\Pr[D \leq 2(h+1) | P_2(h-1), P_1(h), P_1(h-1)] \geq 1 - \mathcal{O}\left(\frac{\delta}{L}\right).$$

Therefore, by combining the above with Eq. (50) via union bound we find that,

$$\Pr\left[\left|\left\langle \psi^{(h)}(y), \psi^{(h)}(z)\right\rangle - \left\langle \boldsymbol{Q}^2 f \oplus \phi^{(h)}(y), \boldsymbol{Q}^2 g \oplus \phi^{(h)}(z)\right\rangle\right| \leq \mathcal{O}\left(\frac{\varepsilon h}{L}\right)\right] \geq 1 - \mathcal{O}\left(\frac{\delta}{L}\right), \quad (51)$$

Now note that $\left\langle \boldsymbol{Q}^2 f \oplus \phi^{(h)}(y), \boldsymbol{Q}^2 g \oplus \phi^{(h)}(z)\right\rangle = \left\langle \boldsymbol{Q}^2 f, \boldsymbol{Q}^2 g\right\rangle + \left\langle \phi^{(h)}(y), \phi^{(h)}(z)\right\rangle$. We proceed by bounding the term $\left|\left\langle \boldsymbol{Q}^2 f, \boldsymbol{Q}^2 g\right\rangle - \langle f, g\rangle\right|$ using Lemma 1, as follows,

$$\Pr\left[\left|\left\langle \boldsymbol{Q}^2 f, \boldsymbol{Q}^2 g\right\rangle - \langle f, g\rangle\right| \leq \mathcal{O}\left(\frac{\varepsilon}{L}\right)\cdot\|f\|_2\|g\|_2\right] \geq 1 - \mathcal{O}\left(\frac{\delta}{L}\right). \quad (52)$$

We proved that conditioned on $P_2(h-1)$ and $P_1(h-1)$, $\|f\|_2^2 \le 11h/10$ and $\|g\|_2^2 \le 11h/10$ with probability at least $1 - \mathcal{O}\left(\frac{\delta}{L}\right)$. Hence, by union bound we find that,

$$\Pr\left[\left|\left\langle \boldsymbol{Q}^2 f, \boldsymbol{Q}^2 g\right\rangle - \langle f, g\rangle\right| \le \mathcal{O}\left(\frac{\varepsilon h}{L}\right) \middle| P_2(h-1), P_1(h-1)\right] \ge 1 - \mathcal{O}\left(\frac{\delta}{L}\right). \tag{53}$$

Note that $\langle f, g\rangle = \left\langle \psi^{(h-1)}(y), \psi^{(h-1)}(z)\right\rangle \cdot \left\langle \dot{\phi}^{(h)}(y), \dot{\phi}^{(h)}(z)\right\rangle$, thus by conditioning on inductive hypothesis $P_2(h-1)$ and Eq. (48) we have,

$$\left|\langle f, g\rangle - K_{\texttt{relu}}^{(h-1)}\left(\frac{\langle y, z\rangle}{\|y\|_2\|z\|_2}\right) \cdot \dot{\Sigma}_{\texttt{relu}}^{(h)}\left(\frac{\langle y, z\rangle}{\|y\|_2\|z\|_2}\right)\right| \le \frac{\varepsilon}{8L}\left(h + \varepsilon \cdot \frac{(h-1)^2+1}{10L}\right) + \varepsilon \cdot \frac{(h-1)^2+1}{10L}$$

By combining the above inequality with Eq. (53), $P_1(h)$, and Eq. (51) using triangle inequality and union bound we get the following inequality,

$$\Pr\left[\left|\left\langle \psi^{(h)}(y), \psi^{(h)}(z)\right\rangle - K_{\texttt{relu}}^{(h-1)}\left(\frac{\langle y, z\rangle}{\|y\|\|z\|}\right) \cdot \dot{\Sigma}_{\texttt{relu}}^{(h)}\left(\frac{\langle y, z\rangle}{\|y\|\|z\|}\right) - \Sigma_{\texttt{relu}}^{(h)}\left(\frac{\langle y, z\rangle}{\|y\|\|z\|}\right)\right| > \varepsilon \cdot \frac{h^2+1}{10L}\right] \le O\left(\frac{\delta}{L}\right).$$

By noting that $K_{\texttt{relu}}^{(h-1)}(\alpha) \cdot \dot{\Sigma}_{\texttt{relu}}^{(h)}(\alpha) + \Sigma_{\texttt{relu}}^{(h)}(\alpha) = K_{\texttt{relu}}^{(h)}(\alpha)$ (see Eq. (4)) we have proved that

$$\Pr\left[\left|\left\langle \psi^{(h)}(y), \psi^{(h)}(z)\right\rangle - K_{\texttt{relu}}^{(h)}\left(\frac{\langle y, z\rangle}{\|y\|_2\|z\|_2}\right)\right| \le \varepsilon \cdot \frac{h^2+1}{10L}\right] \ge 1 - \mathcal{O}\left(\frac{\delta}{L}\right). \tag{54}$$

Similarly we can prove the following inequalities hold with probability at least $1 - \mathcal{O}\left(\frac{\delta}{L}\right)$,

$$\left|\left\|\psi^{(h)}(y)\right\|_2^2 - K_{\texttt{relu}}^{(h)}(1)\right| \le \varepsilon \cdot \frac{h^2+1}{10L}, \text{ and } \left|\left\|\psi^{(h)}(z)\right\|_2^2 - K_{\texttt{relu}}^{(h)}(1)\right| \le \varepsilon \cdot \frac{h^2+1}{10L}.$$

This proves the inductive step for the statement $P_2(h)$ follows, i.e.,

$$\Pr[P_2(h) | P_2(h-1), P_1(h), P_1(h-1)] \ge 1 - \mathcal{O}\left(\frac{\delta}{L}\right).$$

Therefore, by union bounding over all $h = 0, 1, 2, \ldots L$, it follows that the statements of the lemma hold simultaneously for all h with probability at least $1-\delta$. This completes the proof of Lemma 5. □

We now analyze the runtime of the NTKSKETCH algorithm:

Lemma 6 (Runtime of NTKSKETCH)**.** *For every positive integers d and L, every $\varepsilon, \delta > 0$, every vector $x \in \mathbb{R}^d$, the time to compute* NTKSKETCH $\Psi_{\texttt{ntk}}^{(L)}(x) \in \mathbb{R}^{s^*}$*, for $s^* = \mathcal{O}\left(\frac{1}{\varepsilon^2} \cdot \log\frac{1}{\delta}\right)$, using the procedure given in Algorithm 1 is bounded by,*

$$\mathcal{O}\left(\frac{L^{11}}{\varepsilon^{6.7}} \cdot \log^3\frac{L}{\varepsilon\delta} + \frac{L^3}{\varepsilon^2} \cdot \log\frac{L}{\varepsilon\delta} \cdot \text{nnz}(x)\right).$$

Proof of Lemma 6: There are three main components to the runtime of this procedure that we have to account for. The first is the time to apply the sketch $\boldsymbol{Q}^1$ to x in line 4 of Algorithm 1. By Lemma 1, the runtime of computing $\boldsymbol{Q}^1 \cdot x$ is $\mathcal{O}\left(\frac{L^6}{\varepsilon^4} \cdot \log^3\frac{L}{\varepsilon\delta} + \frac{L^3}{\varepsilon^2} \cdot \log\frac{L}{\varepsilon\delta} \cdot \text{nnz}(x)\right)$. The second heavy operation corresponds to computing vectors $Z_j^{(h)}(x) = \boldsymbol{Q}^{2p+2} \cdot \left(\left[\phi^{(h-1)}(x)\right]^{\otimes j} \otimes e_1^{\otimes 2p+2-j}\right)$ for $j = 0, 1, 2, \ldots 2p+2$ and $h = 1, 2, \ldots L$ in Eq. (7). By Lemma 1, the time to compute $Z_j^{(h)}(x)$ for a fixed h and all $j = 0, 1, 2, \ldots 2p+2$ is bounded by,

$$\mathcal{O}\left(\frac{L^{10}}{\varepsilon^{20/3}} \cdot \log^2\frac{L}{\varepsilon} \cdot \log^3\frac{L}{\varepsilon\delta} + \frac{L^8}{\varepsilon^{16/3}} \cdot \log^3\frac{L}{\varepsilon\delta}\right) = \mathcal{O}\left(\frac{L^{10}}{\varepsilon^{6.7}} \cdot \log^3\frac{L}{\varepsilon\delta}\right).$$

The total time to compute vectors $Z_j^{(h)}(x)$ for all $h = 1, 2, \ldots L$ and all $j = 0, 1, 2, \ldots 2p+2$ is thus $\mathcal{O}\left(\frac{L^{11}}{\varepsilon^{6.7}} \cdot \log^3\frac{L}{\varepsilon\delta}\right)$. Finally, the last computationally expensive operation is computing vectors $Y_j^{(h)}(x) = \boldsymbol{Q}^{2p'+1} \cdot \left(\left[\phi^{(h-1)}(x)\right]^{\otimes j} \otimes e_1^{\otimes 2p'+1-j}\right)$ for $j = 0, 1, 2, \ldots 2p'+1$ and $h = 1, 2, \ldots L$ in

Eq. (8). By Lemma 1, the runtime of computing $Y_j^{(h)}(x)$ for a fixed h and all $j = 0, 1, 2, \ldots 2p' + 1$ is bounded by,

$$\mathcal{O}\left(\frac{L^6}{\varepsilon^6} \cdot \log^2 \frac{L}{\varepsilon} \cdot \log^3 \frac{L}{\varepsilon\delta} + \frac{L^8}{\varepsilon^6} \cdot \log^3 \frac{L}{\varepsilon\delta}\right) = \mathcal{O}\left(\frac{L^8}{\varepsilon^6} \cdot \log^3 \frac{L}{\varepsilon\delta}\right).$$

Hence, the total time to compute vectors $Y_j^{(h)}(x)$ for all $h = 1, 2, \ldots L$ and all $j = 0, 1, 2, \ldots 2p' + 1$ is $\mathcal{O}\left(\frac{L^9}{\varepsilon^6} \cdot \log^3 \frac{L}{\varepsilon\delta}\right)$. The total runtime of the NTK Sketch is obtained by summing up these three contributions. This completes the proof of Lemma 6. □

Now we are ready to prove the main theorem on NTKSKETCH.

Theorem 1. *For every integers $d \geq 1$ and $L \geq 2$, and any $\varepsilon, \delta > 0$, let $\Theta_{\texttt{ntk}}^{(L)} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be the L-layer NTK with ReLU activation as per Definition 1 and Eq. (5). Then there exists a randomized map $\Psi_{\texttt{ntk}}^{(L)} : \mathbb{R}^d \to \mathbb{R}^{s^*}$ for some $s^* = \mathcal{O}\left(\frac{1}{\varepsilon^2} \log \frac{1}{\delta}\right)$ such that the following invariants hold,*

1. *For any vectors $y, z \in \mathbb{R}^d$:* $\Pr\left[\left|\left\langle \Psi_{\texttt{ntk}}^{(L)}(y), \Psi_{\texttt{ntk}}^{(L)}(z)\right\rangle - \Theta_{\texttt{ntk}}^{(L)}(y, z)\right| \leq \varepsilon \cdot \Theta_{\texttt{ntk}}^{(L)}(y, z)\right] \geq 1 - \delta.$
2. *For every vecor $x \in \mathbb{R}^d$, the time to compute $\Psi_{\texttt{ntk}}^{(L)}(x)$ is $\mathcal{O}\left(\frac{L^{11}}{\varepsilon^{6.7}} \log^3 \frac{L}{\varepsilon\delta} + \frac{L^3}{\varepsilon^2} \log \frac{L}{\varepsilon\delta} \cdot \text{nnz}(x)\right)$.*

Proof of Theorem 1: Let $\psi^{(L)} : \mathbb{R}^d \to \mathbb{R}^s$ for $s = \mathcal{O}\left(\frac{L^2}{\varepsilon^2} \cdot \log^2 \frac{L}{\varepsilon\delta}\right)$ be the mapping defined in Eq. (9) of Algorithm 1. By Eq. (10), the NTK Sketch $\Psi_{\texttt{ntk}}^{(L)}(x)$ is defined as

$$\Psi_{\texttt{ntk}}^{(L)}(x) := \|x\|_2 \cdot \boldsymbol{G} \cdot \psi^{(L)}(x).$$

Because $\boldsymbol{G}$ is a matrix of i.i.d normal entries with $s^* = C \cdot \frac{1}{\varepsilon^2} \cdot \log \frac{1}{\delta}$ rows, for a large enough constant C, $\boldsymbol{G}$ is a JL transform [15] and hence $\Psi_{\texttt{ntk}}^{(L)}$ satisfies the following,

$$\Pr\left[\left|\left\langle \Psi_{\texttt{ntk}}^{(L)}(y), \Psi_{\texttt{ntk}}^{(L)}(z)\right\rangle - \|y\|_2\|z\|_2 \cdot \left\langle \psi^{(L)}(y), \psi^{(L)}(z)\right\rangle\right| \leq \mathcal{O}(\varepsilon) \cdot A\right] \geq 1 - \mathcal{O}(\delta),$$

where $A := \|y\|_2\|z\|_2 \left\|\psi^{(L)}(y)\right\|_2 \left\|\psi^{(L)}(z)\right\|_2$. By Lemma 5 and using the fact that $K_{\texttt{relu}}^{(L)}(1) = L + 1$, the following bounds hold with probability at least $1 - \mathcal{O}(\delta)$:

$$\left\|\psi^{(L)}(y)\right\|_2^2 \leq \frac{11}{10} \cdot (L + 1), \text{ and } \left\|\psi^{(L)}(z)\right\|_2^2 \leq \frac{11}{10} \cdot (L + 1).$$

Therefore, by union bound we find that,

$$\Pr\left[\left|\left\langle \Psi_{\texttt{ntk}}^{(L)}(y), \Psi_{\texttt{ntk}}^{(L)}(z)\right\rangle - \|y\|_2\|z\|_2 \cdot \left\langle \psi^{(L)}(y), \psi^{(L)}(z)\right\rangle\right| \leq \mathcal{O}(\varepsilon L) \cdot \|y\|_2\|z\|_2\right] \geq 1 - \mathcal{O}(\delta).$$

Additionally, by Lemma 5, the following holds with probability at least $1 - \mathcal{O}(\delta)$:

$$\left|\left\langle \psi^{(L)}(y), \psi^{(L)}(z)\right\rangle - K_{\texttt{relu}}^{(L)}\left(\frac{\langle y, z\rangle}{\|y\|_2\|z\|_2}\right)\right| \leq \frac{\varepsilon(L + 1)}{10}.$$

Hence by union bound and triangle inequality we have,

$$\Pr\left[\left|\left\langle \Psi_{\texttt{ntk}}^{(L)}(y), \Psi_{\texttt{ntk}}^{(L)}(z)\right\rangle - \|y\|_2\|z\|_2 \cdot K_{\texttt{relu}}^{(L)}\left(\frac{\langle y, z\rangle}{\|y\|_2\|z\|_2}\right)\right| \leq \frac{\varepsilon(L + 1)}{9} \cdot \|y\|_2\|z\|_2\right] \geq 1 - \mathcal{O}(\delta).$$

Now note that by Eq. (5), $\|y\|_2\|z\|_2 \cdot K_{\texttt{relu}}^{(L)}\left(\frac{\langle y,z\rangle}{\|y\|_2\|z\|_2}\right) = \Theta_{\texttt{ntk}}^{(L)}(y, z)$, and also note that for every $L \geq 2$ and any $\alpha \in [-1, 1]$, $K_{\texttt{relu}}^{(L)}(\alpha) \geq (L + 1)/9$, therefore,

$$\Pr\left[\left|\left\langle \Psi_{\texttt{ntk}}^{(L)}(y), \Psi_{\texttt{ntk}}^{(L)}(z)\right\rangle - \Theta_{\texttt{ntk}}^{(L)}(y, z)\right| \leq \varepsilon \cdot \Theta_{\texttt{ntk}}^{(L)}(y, z)\right] \geq 1 - \delta.$$

Remark on the fact that $K_{\texttt{relu}}^{(L)}(\alpha) \geq (L + 1)/9$ for every $L \geq 2$ and any $\alpha \in [-1, 1]$. Note that from the definition of $\Sigma_{\texttt{relu}}^{(h)}$ in Eq. (3), we have that for any $\alpha \in [-1, 1]$: $\Sigma_{\texttt{relu}}^{(0)}(\alpha) \geq -1$,

$\Sigma^{(1)}_{\texttt{relu}}(\alpha) \geq 0$, $\Sigma^{(2)}_{\texttt{relu}}(\alpha) \geq \frac{1}{\pi}$, and $\Sigma^{(h)}_{\texttt{relu}}(\alpha) \geq \frac{1}{2}$ for every $h \geq 3$ because $\kappa_1(\cdot)$ is a monotonically increasing function on the interval $[-1, 1]$. Moreover, using the definition of $\dot{\Sigma}^{(h)}_{\texttt{relu}}$ in Eq. (3), we have that for any $\alpha \in [-1, 1]$: $\dot{\Sigma}^{(1)}_{\texttt{relu}}(\alpha) \geq 0$, $\dot{\Sigma}^{(2)}_{\texttt{relu}}(\alpha) \geq \frac{1}{2}$, and $\dot{\Sigma}^{(h)}_{\texttt{relu}}(\alpha) \geq \frac{3}{5}$ for every $h \geq 3$ because $\kappa_0(\cdot)$ is a monotonically increasing function on the interval $[-1, 1]$. By an inducive proof and using the definition of $K^{(L)}_{\texttt{relu}}$ in Eq. (4), we can show that $K^{(L)}_{\texttt{relu}}(\alpha) \geq (L+1)/9$ for every $L \geq 2$ and any $\alpha \in [-1, 1]$

Runtime analysis: By Lemma 6, runtime to compute the NTKSKETCH is

$$\mathcal{O}\left(\frac{L^{11}}{\varepsilon^{6.7}} \log^3 \frac{L}{\varepsilon\delta} + \frac{L^3}{\varepsilon^2} \log \frac{L}{\varepsilon\delta} \cdot \text{nnz}(x)\right). \tag{55}$$

This completes the proof of Theorem 1. □

D NTK Random Features: Claims and Proofs

D.1 Proof of Theorem 2

In this section we prove Theorem 2. We first restate the theorem:

Theorem 2. *Given $y, z \in \mathbb{R}^d$ and $L \geq 2$, let $\Theta^{(L)}_{\texttt{ntk}}$ the L-layer fully-connected ReLU NTK. For $\varepsilon, \delta > 0$, there exist $m_0 = \mathcal{O}\left(\frac{L^2}{\varepsilon^2} \log \frac{L}{\delta}\right), m_1 = \mathcal{O}\left(\frac{L^6}{\varepsilon^4} \log \frac{L}{\delta}\right), m_s = \mathcal{O}\left(\frac{L^2}{\varepsilon^2} \log^3 \frac{L}{\varepsilon\delta}\right)$, such that,*

$$\Pr\left[\left|\left\langle \Psi^{(L)}_{\texttt{rf}}(y), \Psi^{(L)}_{\texttt{rf}}(z)\right\rangle - \Theta^{(L)}_{\texttt{ntk}}(y, z)\right| \leq \varepsilon \cdot \Theta^{(L)}_{\texttt{ntk}}(y, z)\right] \geq 1 - \delta, \tag{12}$$

where $\Psi^{(L)}_{\texttt{rf}}(y), \Psi^{(L)}_{\texttt{rf}}(z) \in \mathbb{R}^{m_1+m_s}$ are the outputs of Algorithm 2, using the same randomness.

In the proof of this theorem we use the following results from the literature,

Lemma 7 (Corollary 16 in [14])**.** *Given integer $\ell > 0$ and $x, x' \in \mathbb{R}^d$ such that $\|y\|_2 = \|z\|_2 = 1$, let $\phi^{(\ell)}_{\texttt{rf}}(y), \phi^{(\ell)}_{\texttt{rf}}(z)$ be defined as per line 5 of Algorithm 2. For $\delta_1, \varepsilon_1 \in (0, 1)$, there exists a constant $C_1 > 0$ such that for any $m_1 \geq C_1 \frac{L^2}{\varepsilon_1^2} \log\left(\frac{L}{\delta_1}\right)$ the following holds:*

$$\Pr\left[\left|\left\langle \phi^{(\ell)}_{\texttt{rf}}(y), \phi^{(\ell)}_{\texttt{rf}}(z)\right\rangle - \Sigma^{(\ell)}_{\texttt{relu}}(\langle y, z\rangle)\right| \leq \varepsilon_1\right] \geq 1 - \delta_1. \tag{56}$$

We also need the following lemma,

Lemma 8 (Lemma E.5 in [5])**.** *Given $x, x' \in \mathbb{R}^d$ with $\|x\|_2 = \|x'\|_2 = 1$, integer $\ell > 0$, let $\phi^{(\ell)}_{\texttt{rf}}(x), \phi^{(\ell)}_{\texttt{rf}}(x')$ be defined as per line 5 of Algorithm 2. For any $\varepsilon_2 \in (0, 1)$ assume that*

$$\left|\left\langle \phi^{(\ell)}_{\texttt{rf}}(x), \phi^{(\ell)}_{\texttt{rf}}(x')\right\rangle - \Sigma^{(\ell)}_{\texttt{relu}}(\langle x, x'\rangle)\right| \leq \frac{\varepsilon_2^2}{2}. \tag{57}$$

Then, for $\dot{\phi}^{(\ell)}_{\texttt{rf}}(x), \dot{\phi}^{(\ell)}_{\texttt{rf}}(x')$ defined as in line 4 of Algorithm 2, and any $\delta_2 > 0$ the following holds:

$$\Pr\left[\left|\left\langle \dot{\phi}^{(\ell)}_{\texttt{rf}}(x), \dot{\phi}^{(\ell)}_{\texttt{rf}}(x')\right\rangle - \dot{\Sigma}^{(\ell)}_{\texttt{relu}}(\langle x, x'\rangle)\right| \leq \varepsilon_2 + \sqrt{\frac{2}{m_0} \log\left(\frac{6}{\delta_2}\right)}\right] \geq 1 - \delta_2. \tag{58}$$

Proof of Theorem 2: For fixed $y, z \in \mathbb{R}^d$ and $\ell = 0, \ldots, L$, we denote the estimation error as

$$\Delta_\ell := \max_{(x,x') \in \{(y,z),(y,y),(z,z)\}} \left|\left\langle \psi^{(\ell)}_{\texttt{rf}}(x), \psi^{(\ell)}_{\texttt{rf}}(x')\right\rangle - K^{(\ell)}_{\texttt{relu}}\left(\frac{\langle x, x'\rangle}{\|x\|_2 \|x'\|_2}\right)\right|$$

and note that $\Delta_0 = 0$. Recall that, by Definition 1 and Eq. (5):

$$\Theta^{(\ell)}_{\texttt{ntk}}(x, x') = \|x\|_2 \|x'\|_2 \cdot K^{(\ell)}_{\texttt{relu}}\left(\frac{\langle x, x'\rangle}{\|x\|_2 \|x'\|_2}\right), \tag{59}$$

where for every $\alpha \in [-1,1]$ and $\ell = 1,\dots,L$:

$$K^{(\ell)}_{\texttt{relu}}(\alpha) := K^{(\ell-1)}_{\texttt{relu}}(\alpha)\cdot \dot{\Sigma}^{(\ell)}_{\texttt{relu}}(\alpha) + \Sigma^{(\ell)}_{\texttt{relu}}(\alpha),$$
$$\dot{\Sigma}^{(\ell)}_{\texttt{relu}}(\alpha) := \kappa_0\left(\Sigma^{(\ell-1)}_{\texttt{relu}}(\alpha)\right),$$
$$\Sigma^{(\ell)}_{\texttt{relu}}(\alpha) := \kappa_1\left(\Sigma^{(\ell-1)}_{\texttt{relu}}(\alpha)\right).$$

We use the recursive relation to approximate:

$$\begin{aligned}\left\langle \psi^{(\ell)}_{\texttt{rf}}(x), \psi^{(\ell)}_{\texttt{rf}}(x')\right\rangle &= \left\langle \phi^{(\ell)}_{\texttt{rf}}(x), \phi^{(\ell)}_{\texttt{rf}}(x')\right\rangle \\ &\quad + \left\langle \boldsymbol{Q}^2\cdot\left(\dot{\phi}^{(\ell)}_{\texttt{rf}}(x)\otimes\psi^{(\ell-1)}_{\texttt{rf}}(x)\right), \boldsymbol{Q}^2\cdot\left(\dot{\phi}^{(\ell)}_{\texttt{rf}}(x')\otimes\psi^{(\ell-1)}_{\texttt{rf}}(x')\right)\right\rangle \\ &\approx \left\langle \phi^{(\ell)}_{\texttt{rf}}(x), \phi^{(\ell)}_{\texttt{rf}}(x')\right\rangle + \left\langle \dot{\phi}^{(\ell)}_{\texttt{rf}}(x)\otimes\psi^{(\ell-1)}_{\texttt{rf}}(x), \dot{\phi}^{(\ell)}_{\texttt{rf}}(x')\otimes\psi^{(\ell-1)}_{\texttt{rf}}(x')\right\rangle \\ &\approx \left(\Sigma^{(\ell)}_{\texttt{relu}} + \dot{\Sigma}^{(\ell)}_{\texttt{relu}}\cdot K^{(\ell-1)}_{\texttt{relu}}\right)\left(\frac{\langle x,x'\rangle}{\|x\|_2\|x'\|_2}\right) = K^{(\ell)}_{\texttt{relu}}\left(\frac{\langle x,x'\rangle}{\|x\|_2\|x'\|_2}\right).\end{aligned}$$

For notational simplicity, we define the following events:

$$\mathcal{E}^{(\ell)}_{\phi}(\varepsilon) := \left\{\left|\left\langle \phi^{(\ell)}_{\texttt{rf}}(x), \phi^{(\ell)}_{\texttt{rf}}(x')\right\rangle - \Sigma^{(\ell)}_{\texttt{relu}}\left(\frac{\langle x,x'\rangle}{\|x\|_2\|x'\|_2}\right)\right| \le \varepsilon : \forall (x,x') \in \{(y,z),(y,y),(z,z)\}\right\}, \quad (60)$$

$$\dot{\mathcal{E}}^{(\ell)}_{\phi}(\varepsilon) := \left\{\left|\left\langle \dot{\phi}^{(\ell)}_{\texttt{rf}}(x), \dot{\phi}^{(\ell)}_{\texttt{rf}}(x')\right\rangle - \dot{\Sigma}^{(\ell)}_{\texttt{relu}}\left(\frac{\langle x,x'\rangle}{\|x\|_2\|x'\|_2}\right)\right| \le \varepsilon : \forall (x,x') \in \{(y,z),(y,y),(z,z)\}\right\}. \quad (61)$$

Our proof is based on the following claims:

First we claim that, there exists a constant $C_1 > 0$ such that for any $m_1 \ge C_1 \frac{L^6}{\varepsilon^4}\log\left(\frac{L}{\delta}\right)$:

$$\Pr\left[\mathcal{E}^{(\ell)}_{\phi}\left(\frac{\varepsilon^2}{100L^2}\right)\right] \ge 1 - \frac{\delta}{3L}. \quad (62)$$

which directly follows by invoking Lemma 7 with $(x,x') \in \left\{\left(\frac{y}{\|y\|_2},\frac{z}{\|z\|_2}\right),\left(\frac{y}{\|y\|_2},\frac{y}{\|y\|_2}\right),\left(\frac{z}{\|z\|_2},\frac{z}{\|z\|_2}\right)\right\}$ and setting $\varepsilon_1 = \frac{\varepsilon^2}{100L^2}$ and $\delta_1 = \frac{\delta}{9L}$ and applying union bound over choices of (x,x').

Our second claim is that there exists a constant $C_0 > 0$ such that if $m_0 \ge C_0\frac{L^2}{\varepsilon^2}\log\left(\frac{L}{\delta}\right)$ then

$$\Pr\left[\dot{\mathcal{E}}^{(\ell)}_{\phi}\left(\frac{\varepsilon}{8L}\right)\middle|\,\mathcal{E}^{(\ell)}_{\phi}\left(\frac{\varepsilon^2}{100L^2}\right)\right] \ge 1 - \frac{\delta}{3L}. \quad (63)$$

The above statement is a direct consequence of invoking Lemma 8 with $(x,x') \in \left\{\left(\frac{y}{\|y\|_2},\frac{z}{\|z\|_2}\right),\left(\frac{y}{\|y\|_2},\frac{y}{\|y\|_2}\right),\left(\frac{z}{\|z\|_2},\frac{z}{\|z\|_2}\right)\right\}$ and union bounding over choices of (x,x'). In Lemma 8, we choose $\varepsilon_2 = \frac{\varepsilon}{10L}, \delta_2 = \frac{\delta}{9L}$ and $m_0 \ge \frac{3200L^2}{\varepsilon^2}\log\left(\frac{54L}{\delta}\right)$ for $\varepsilon,\delta \in (0,1)$ to obtain Eq. (63) by union bound.

Our next claim is the following,

Claim 1. *Let $f(x) := \dot{\phi}^{(\ell)}_{\texttt{rf}}(x)\otimes\psi^{(\ell-1)}_{\texttt{rf}}(x)$ for every $x \in \mathbb{R}^d$. There exists a constant $C_2 > 0$ such that if $m_s \ge C_2\frac{L^2}{\varepsilon^2}\log\frac{L}{\varepsilon\delta}$ then conditioned on the events $\dot{\mathcal{E}}^{(\ell)}_{\phi}\left(\frac{\varepsilon}{8L}\right)$ and $\mathcal{E}^{(\ell)}_{\phi}\left(\frac{\varepsilon^2}{100L^2}\right)$ the following holds with probability at least $1-\frac{\delta}{3L}$, for all $(x,x') \in \left\{\left(\frac{y}{\|y\|_2},\frac{z}{\|z\|_2}\right),\left(\frac{y}{\|y\|_2},\frac{y}{\|y\|_2}\right),\left(\frac{z}{\|z\|_2},\frac{z}{\|z\|_2}\right)\right\}$,*

$$\left|\left\langle \boldsymbol{Q}^2\cdot f(x), \boldsymbol{Q}^2\cdot f(x')\right\rangle - \langle f(x), f(x')\rangle\right| \le \frac{\varepsilon}{100L}(\ell + \Delta_{\ell-1}).$$

The proof of the above claim is provided in Appendix D.2. Now, by combining Eq. (62), Eq. (63) and Claim 1, we can prove the following claim which provides a recursive relation for bounding Δ_ℓ.

Claim 2. *If the events in Eq. (62), Eq. (63) and Claim 1 hold, then*

$$\Delta_\ell \le \left(1 + \frac{\varepsilon}{7L}\right)\Delta_{\ell-1} + \frac{\varepsilon}{7L}\ell. \tag{64}$$

The proof of Claim 2 is provided in Appendix D.2.

Note that by union bound, with probability $1 - \frac{\delta}{L}$, all preconditions of Claim 2 hold and thus Eq. (64) holds with probability $1 - \frac{\delta}{L}$. Applying union bound on Eq. (64) for all $\ell \in [L]$ and solving the recurrence, we obtain that with probability at least $1 - \delta$, the following bound holds

$$\Delta_\ell \le \frac{\varepsilon}{9L} \cdot \ell^2. \tag{65}$$

When $L \ge 2$, we showed in the proof of Theorem 1 that $K^{(L)}_{\texttt{relu}}(\cdot) \ge \frac{L+1}{9}$, therefore,

$$\left|\left\langle \psi^{(L)}_{\texttt{rf}}(x), \psi^{(L)}_{\texttt{rf}}(x')\right\rangle - K^{(L)}_{\texttt{relu}}\left(\frac{\langle x, x'\rangle}{\|x\|_2\|x'\|_2}\right)\right| \le \varepsilon \cdot K^{(L)}_{\texttt{relu}}\left(\frac{\langle x, x'\rangle}{\|x\|_2\|x'\|_2}\right). \tag{66}$$

Since $\Psi^{(L)}_{\texttt{rf}}(x) = \|x\|_2 \cdot \psi^{(L)}(x)$ and $\Psi^{(L)}_{\texttt{rf}}(x') = \|x'\|_2 \cdot \psi^{(L)}(x')$, this implies that,

$$\Pr\left[\left|\left\langle \Psi^{(L)}_{\texttt{rf}}(x), \Psi^{(L)}_{\texttt{rf}}(x')\right\rangle - \Theta^{(L)}_{\texttt{ntk}}(x, x')\right| \le \varepsilon \cdot \Theta^{(L)}_{\texttt{ntk}}(x, x')\right] \ge 1 - \delta.$$

This completes the proof of Theorem 2. □

D.2 Proof of Auxiliary Claims

Claim 1. *Let $f(x) := \dot{\phi}^{(\ell)}_{\texttt{rf}}(x) \otimes \psi^{(\ell-1)}_{\texttt{rf}}(x)$ for every $x \in \mathbb{R}^d$. There exists a constant $C_2 > 0$ such that if $m_s \ge C_2 \frac{L^2}{\varepsilon^2}\log\frac{L}{\varepsilon\delta}$ then conditioned on the events $\dot{\mathcal{E}}^{(\ell)}_\phi\left(\frac{\varepsilon}{8L}\right)$ and $\mathcal{E}^{(\ell)}_\phi\left(\frac{\varepsilon^2}{100L^2}\right)$ the following holds with probability at least $1 - \frac{\delta}{3L}$, for all $(x, x') \in \left\{\left(\frac{y}{\|y\|_2}, \frac{z}{\|z\|_2}\right), \left(\frac{y}{\|y\|_2}, \frac{y}{\|y\|_2}\right), \left(\frac{z}{\|z\|_2}, \frac{z}{\|z\|_2}\right)\right\}$,*

$$\left|\langle \boldsymbol{Q}^2 \cdot f(x), \boldsymbol{Q}^2 \cdot f(x')\rangle - \langle f(x), f(x')\rangle\right| \le \frac{\varepsilon}{100L}(\ell + \Delta_{\ell-1}).$$

Proof of Claim 1: The proof is based on Lemma 1 that provides an upper bound on variance of the PolySketch. By using the definition of $f(x), f(x')$ and Lemma 1, with probability at least $1 - \frac{\delta}{9L}$, we have

$$\begin{aligned}
&\left|\langle \boldsymbol{Q}^2 \cdot f(x), \boldsymbol{Q}^2 \cdot f(x')\rangle - \langle f(x), f(x')\rangle\right| \\
&\quad\le \frac{\varepsilon}{200L}\left\|\dot{\phi}^{(\ell)}_{\texttt{rf}}(x)\right\|_2\left\|\dot{\phi}^{(\ell)}_{\texttt{rf}}(x')\right\|_2\left\|\psi^{(\ell-1)}_{\texttt{rf}}(x)\right\|_2\left\|\psi^{(\ell-1)}_{\texttt{rf}}(x')\right\|_2 \\
&\quad\le \frac{\varepsilon}{200L}\left(1 + \frac{\varepsilon}{8L}\right)\left\|\psi^{(\ell-1)}_{\texttt{rf}}(x)\right\|_2\left\|\psi^{(\ell-1)}_{\texttt{rf}}(x')\right\|_2 \\
&\quad\le \frac{\varepsilon}{100L}\left(\ell + \Delta_{\ell-1}\right)
\end{aligned}$$

where the second inequality follows from the assumption that $\dot{\mathcal{E}}^{(\ell)}_\phi\left(\frac{\varepsilon}{8L}\right)$ holds and the third one follows from the fact that $K^{(\ell-1)}_{\texttt{relu}}(\alpha) \le \ell$ for any $\alpha \in [-1, 1]$ and

$$\left\|\psi^{(\ell-1)}_{\texttt{rf}}(x)\right\|_2^2 \le K^{(\ell-1)}_{\texttt{relu}}(x, x') + \Delta_{\ell-1} \le \ell + \Delta_{\ell-1}. \tag{67}$$

Union bounding over the choices of (x, x') completes the proof of Claim 1. □

Claim 2. *If the events in Eq. (62), Eq. (63) and Claim 1 hold, then*

$$\Delta_\ell \le \left(1 + \frac{\varepsilon}{7L}\right)\Delta_{\ell-1} + \frac{\varepsilon}{7L}\ell. \tag{64}$$

Proof of Claim 2: Recall that

$$\Delta_\ell := \max_{(x,x')\in\{(y,z),(y,y),(z,z)\}} \left| \left\langle \psi_{\mathtt{rf}}^{(\ell)}(x), \psi_{\mathtt{rf}}^{(\ell)}(x') \right\rangle - K_{\mathtt{relu}}^{(\ell)} \left(\frac{\langle x, x' \rangle}{\|x\|_2 \|x'\|_2} \right) \right|.$$

Observe that the estimation error Δ_ℓ can be decomposed into three parts:

$$\begin{aligned} \left| \left\langle \psi_{\mathtt{rf}}^{(\ell)}(x), \psi_{\mathtt{rf}}^{(\ell)}(x') \right\rangle - K_{\mathtt{relu}}^{(\ell)} \left(\frac{\langle x, x' \rangle}{\|x\|_2 \|x'\|_2} \right) \right| &\le \left| \left\langle \phi_{\mathtt{rf}}^{(\ell)}(x), \phi_{\mathtt{rf}}^{(\ell)}(x') \right\rangle - \Sigma_{\mathtt{relu}}^{(\ell)} \left(\frac{\langle x, x' \rangle}{\|x\|_2 \|x'\|_2} \right) \right| \\ &+ \left| \langle \boldsymbol{Q}^2 \cdot f(x), \boldsymbol{Q}^2 \cdot f(x') \rangle - \langle f(x), f(x') \rangle \right| \\ &+ \left| \left\langle \dot{\phi}_{\mathtt{rf}}^{(\ell)}(x) \otimes \psi_{\mathtt{rf}}^{(\ell-1)}(x), \dot{\phi}_{\mathtt{rf}}^{(\ell)}(x') \otimes \psi_{\mathtt{rf}}^{(\ell-1)}(x') \right\rangle - \dot{\Sigma}_{\mathtt{relu}}^{(\ell)} \left(\frac{\langle x, x' \rangle}{\|x\|_2 \|x'\|_2} \right) K_{\mathtt{relu}}^{(\ell-1)} \left(\frac{\langle x, x' \rangle}{\|x\|_2 \|x'\|_2} \right) \right|. \end{aligned} \tag{68}$$

for $(x, x') \in \{(y, z), (y, y), (z, z)\}$. By the assumption that Eq. (62) holds, the definition of $\mathcal{E}_\phi^{(\ell)} \left(\frac{\varepsilon^2}{100L^2} \right)$ implies that,

$$\left| \left\langle \phi_{\mathtt{rf}}^{(\ell)}(x), \phi_{\mathtt{rf}}^{(\ell)}(x') \right\rangle - \Sigma_{\mathtt{relu}}^{(\ell)} \left(\frac{\langle x, x' \rangle}{\|x\|_2 \|x'\|_2} \right) \right| \le \frac{\varepsilon^2}{100L^2} \tag{69}$$

Furthermore, the assumption that Claim 1 holds, implies that,

$$\left| \langle \boldsymbol{Q}^2 \cdot f(x), \boldsymbol{Q}^2 \cdot f(x') \rangle - \langle f(x), f(x') \rangle \right| \le \frac{\varepsilon}{100L} (\ell + \Delta_{\ell-1}). \tag{70}$$

For the third part in Eq. (68), we observe that

$$\begin{aligned} &\left| \left\langle \dot{\phi}_{\mathtt{rf}}^{(\ell)}(x) \otimes \psi_{\mathtt{rf}}^{(\ell-1)}(x), \dot{\phi}_{\mathtt{rf}}^{(\ell)}(x') \otimes \psi_{\mathtt{rf}}^{(\ell-1)}(x') \right\rangle - \dot{\Sigma}_{\mathtt{relu}}^{(\ell)} \left(\frac{\langle x, x' \rangle}{\|x\|_2 \|x'\|_2} \right) K_{\mathtt{relu}}^{(\ell-1)} \left(\frac{\langle x, x' \rangle}{\|x\|_2 \|x'\|_2} \right) \right| \\ &\le \left| \left\langle \dot{\phi}_{\mathtt{rf}}^{(\ell)}(x), \dot{\phi}_{\mathtt{rf}}^{(\ell)}(x') \right\rangle \right| \cdot \left| \left\langle \psi_{\mathtt{rf}}^{(\ell-1)}(x), \psi_{\mathtt{rf}}^{(\ell-1)}(x') \right\rangle - K_{\mathtt{relu}}^{(\ell-1)} \left(\frac{\langle x, x' \rangle}{\|x\|_2 \|x'\|_2} \right) \right| \\ &\quad + K_{\mathtt{relu}}^{(\ell-1)} \left(\frac{\langle x, x' \rangle}{\|x\|_2 \|x'\|_2} \right) \cdot \left| \left\langle \dot{\phi}_{\mathtt{rf}}^{(\ell)}(x), \dot{\phi}_{\mathtt{rf}}^{(\ell)}(x') \right\rangle - \dot{\Sigma}_{\mathtt{relu}}^{(\ell)} \left(\frac{\langle x, x' \rangle}{\|x\|_2 \|x'\|_2} \right) \right| \\ &\le \left(1 + \frac{\varepsilon}{8L} \right) \cdot \Delta_{\ell-1} + \ell \cdot \frac{\varepsilon}{8L}, \end{aligned} \tag{71}$$

where the second inequality comes from that the assumption that Eq. (63) holds along with $\left| \dot{\Sigma}_{\mathtt{relu}}^{(\ell)}(\cdot) \right| \le 1$ and $\left| K_{\mathtt{relu}}^{(\ell-1)}(\cdot) \right| \le \ell$. Putting Eq. (69), Eq. (70) and Eq. (71) into Eq. (68), we have

$$\Delta_\ell \le \frac{\varepsilon^2}{100L^2} + \frac{\varepsilon}{100L} (\ell + \Delta_{\ell-1}) + \left(1 + \frac{\varepsilon}{8L} \right) \Delta_{\ell-1} + \frac{\ell \cdot \varepsilon}{8L} = \left(1 + \frac{\varepsilon}{7L} \right) \Delta_{\ell-1} + \frac{\varepsilon}{7L} \ell. \tag{72}$$

This completes the proof of Claim 2. □

E Spectral Approximation via Leverage Scores Sampling

E.1 Zeroth Order Arc-Cosine Kernels

The proofs here rely on Theorem 3.3 in [28] which states spectral approximation bounds of random features for general kernels equipped with the leverage score sampling. This result is a generalization of [8] on the Random Fourier Features.

Theorem 5 (Theorem 3.3 in [28])**.** *Suppose $\boldsymbol{K} \in \mathbb{R}^{n \times n}$ is a kernel matrix with statistical dimension s_λ for some $\lambda \in (0, \|\boldsymbol{K}\|_2)$. Let $\boldsymbol{\Phi}(\boldsymbol{w}) \in \mathbb{R}^n$ be a feature map with a random vector $\boldsymbol{w} \sim p(\boldsymbol{w})$ satisfying that $\boldsymbol{K} = \mathbb{E}_{\boldsymbol{w}} \left[\boldsymbol{\Phi}(\boldsymbol{w}) \boldsymbol{\Phi}(\boldsymbol{w})^\top \right]$. Define $\tau_\lambda(\boldsymbol{w}) := p(\boldsymbol{w}) \cdot \boldsymbol{\Phi}(\boldsymbol{w})^\top (\boldsymbol{K} + \lambda \boldsymbol{I})^{-1} \boldsymbol{\Phi}(\boldsymbol{w})$. Let $\widetilde{\tau}(\boldsymbol{w})$ be any measurable function such that $\widetilde{\tau}(\boldsymbol{w}) \ge \tau_\lambda(\boldsymbol{w})$ for all $\boldsymbol{w}$. Assume that $s_{\widetilde{\tau}} := \int \widetilde{\tau}(\boldsymbol{w}) d\boldsymbol{w}$ is finite. Consider random vectors $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_m$ sampled from $q(\boldsymbol{w}) := \widetilde{\tau}(\boldsymbol{w}) / s_{\widetilde{\tau}}$ and define that*

$$\overline{\boldsymbol{\Phi}} := \frac{1}{\sqrt{m}} \left[\sqrt{\frac{p(\boldsymbol{w}_1)}{q(\boldsymbol{w}_1)}} \boldsymbol{\Phi}(\boldsymbol{w}_1), \ \ldots \ , \sqrt{\frac{p(\boldsymbol{w}_m)}{q(\boldsymbol{w}_m)}} \boldsymbol{\Phi}(\boldsymbol{w}_m) \right]^\top. \tag{73}$$

If $m \geq \frac{8}{3}\varepsilon^{-2} s_{\tilde{\tau}} \log\left(16 s_\lambda / \delta\right)$ *then*

$$(1-\varepsilon)\left(\boldsymbol{K}+\lambda \boldsymbol{I}\right) \preceq \overline{\boldsymbol{\Phi}}^\top \overline{\boldsymbol{\Phi}} + \lambda \boldsymbol{I} \preceq (1+\varepsilon)\left(\boldsymbol{K}+\lambda \boldsymbol{I}\right) \tag{74}$$

holds with probability at least $1-\delta$.

We now ready to provide spectral approximation guarantee for arc-cosine kernels of order zero.

Theorem 6. *Given dataset* $\boldsymbol{X} \in \mathbb{R}^{d\times n}$*, let* $\boldsymbol{K}_0 \in \mathbb{R}^{n\times n}$ *be the arc-cosine kernel matrix of* 0*-th order with* $\boldsymbol{X}$ *and denote* $\boldsymbol{\Phi}_0 := \sqrt{\frac{2}{m}}\mathrm{Step}(\boldsymbol{W}\boldsymbol{X}) \in \mathbb{R}^{m\times n}$ *where each entry in* $\boldsymbol{W} \in \mathbb{R}^{m\times d}$ *is an i.i.d. sample from* $\mathcal{N}(0,1)$*. For* $\lambda \in (0, \|\boldsymbol{K}_0\|_2)$*, let* s_λ *be the statistical dimension of* $\boldsymbol{K}_0$*. Given* $\varepsilon \in (0,1/2)$ *and* $\delta \in (0,1)$*, if* $m \geq \frac{8}{3}\frac{n}{\lambda\varepsilon^2}\log\left(\frac{16 s_\lambda}{\delta}\right)$*, then it holds that*

$$(1-\varepsilon)(\boldsymbol{K}_0+\lambda\boldsymbol{I}) \preceq \boldsymbol{\Phi}_0^\top\boldsymbol{\Phi}_0 + \lambda\boldsymbol{I} \preceq (1+\varepsilon)(\boldsymbol{K}_0+\lambda\boldsymbol{I})$$

with probability at least $1-\delta$.

Proof of Theorem 6: Let $\boldsymbol{\Phi}_0(w) := \sqrt{2}\,\mathrm{Step}(\boldsymbol{X}^\top w) \in \mathbb{R}^n$ for $w \in \mathbb{R}^d$ and $p(w)$ be the probability density function of the standard normal distribution. As studied in [12], $\boldsymbol{\Phi}_0(w)$ is a random feature of $\boldsymbol{K}_0$ such that

$$\boldsymbol{K}_0 = \mathbb{E}_{w\sim p(w)}\left[\boldsymbol{\Phi}_0(w)\boldsymbol{\Phi}_0(w)^\top\right]. \tag{75}$$

In order to utilize Theorem 5, we need an upper bound of $\tau_\lambda(w)$ as below:

$$\begin{aligned}
\tau_\lambda(w) &:= p(w)\cdot\boldsymbol{\Phi}_0(w)^\top\left(\boldsymbol{K}_0+\lambda\boldsymbol{I}\right)^{-1}\boldsymbol{\Phi}_0(w) && (76)\\
&\leq p(w)\left\|\left(\boldsymbol{K}_0+\lambda\boldsymbol{I}\right)^{-1}\right\|_2\|\boldsymbol{\Phi}_0(w)\|_2^2 && (77)\\
&\leq p(w)\frac{\|\boldsymbol{\Phi}_0(w)\|_2^2}{\lambda} && (78)\\
&\leq p(w)\frac{2n}{\lambda} && (79)
\end{aligned}$$

where the inequality in second line holds from the definition of matrix operator norm and the inequality in third line follows from the fact that smallest eigenvalue of $\boldsymbol{K}_0+\lambda\boldsymbol{I}$ is equal to or greater than λ. The last inequality is from that $\|\mathrm{Step}(x)\|_2^2 \leq n$ for any $x \in \mathbb{R}^n$. Note that $\int_{\mathbb{R}^d} p(w)\frac{2n}{\lambda}dw = \frac{2n}{\lambda}$ and since it is constant the modified random features correspond to the original ones. Putting all together into Theorem 5, we can obtain the result. This completes the proof of Theorem 6. □

E.2 First Order Arc-Cosine Kernels

Theorem 7. *Given dataset* $\boldsymbol{X} \in \mathbb{R}^{d\times n}$*, let* $\boldsymbol{K}_1 \in \mathbb{R}^{n\times n}$ *be the arc-cosine kernel matrix of* 1*-th order with* $\boldsymbol{X}$ *and* $v_1,\ldots,v_m \in \mathbb{R}^d$ *be i.i.d. random vectors from probability distribution* $q(v) = \frac{1}{(2\pi)^{d/2}d}\|v\|_2^2\exp\left(-\frac{1}{2}\|v\|_2^2\right)$*. Denote* $\boldsymbol{\Phi}_1 := \sqrt{\frac{2d}{m}}\left[\frac{\mathrm{ReLU}(\boldsymbol{X}^\top v_1)}{\|v_1\|_2},\ \ldots\ ,\frac{\mathrm{ReLU}(\boldsymbol{X}^\top v_m)}{\|v_m\|_2}\right]^\top$ *and for* $\lambda \in (0, \|\boldsymbol{K}_1\|_2)$*, let* s_λ *be the statistical dimension of* $\boldsymbol{K}_1$*. Given* $\varepsilon \in (0,1/2)$ *and* $\delta \in (0,1)$*, if* $m \geq \frac{8}{3}\frac{d}{\varepsilon^2}\min\left\{\mathrm{rank}(\boldsymbol{X})^2, \frac{\|\boldsymbol{X}\|_2^2}{\lambda}\right\}\log\left(\frac{16 s_\lambda}{\delta}\right)$*, then it holds that*

$$(1-\varepsilon)(\boldsymbol{K}_1+\lambda I) \preceq \boldsymbol{\Phi}_1^\top\boldsymbol{\Phi}_1 + \lambda\boldsymbol{I} \preceq (1+\varepsilon)(\boldsymbol{K}_1+\lambda I)$$

with probability at least $1-\delta$.

Proof of Theorem 7: Let $\boldsymbol{\Phi}_1(w) := \sqrt{2}\,\mathrm{ReLU}(\boldsymbol{X}^\top w) \in \mathbb{R}^n$ for $w \in \mathbb{R}^d$ and $p(w)$ be the probability density function of standard normal distribution. Cho and Saul [12] also showed that $\boldsymbol{\Phi}_1(w)$ is a random feature of $\boldsymbol{K}_1$ such that

$$\boldsymbol{K}_1 = \mathbb{E}_{w\sim p(w)}\left[\boldsymbol{\Phi}_1(w)\boldsymbol{\Phi}_1(w)^\top\right]. \tag{80}$$

Then,

$$\begin{aligned}
\tau_\lambda(w) &:= p(w) \cdot \boldsymbol{\Phi}_1(w)^\top (\boldsymbol{K}_1 + \lambda \boldsymbol{I})^{-1} \boldsymbol{\Phi}_1(w) \\
&\le p(w) \left\| (\boldsymbol{K}_1 + \lambda \boldsymbol{I})^{-1} \right\|_2 \left\| \boldsymbol{\Phi}_1(w) \right\|_2^2 \\
&= 2\, p(w) \frac{\left\| \mathrm{ReLU}(\boldsymbol{X}^\top w) \right\|_2^2}{\lambda} \\
&\le 2\, p(w) \frac{\left\| \boldsymbol{X}^\top w \right\|_2^2}{\lambda} \\
&\le 2\, p(w) \left\| w \right\|_2^2 \frac{\left\| \boldsymbol{X} \right\|_2^2}{\lambda}
\end{aligned}$$

where the inequality in fourth line holds from the fact that $\|\mathrm{ReLU}(x)\|_2^2 \le \|x\|_2^2$ for any vector x.

On the other hand, if we write the ReLU function in terms of a matrix form, i.e., $\boldsymbol{\Phi}_1(w) = \sqrt{2}\,\mathrm{ReLU}(\boldsymbol{X}^\top w) = \sqrt{2}\,\boldsymbol{S}\boldsymbol{X}^\top w$ where $\boldsymbol{S}$ is a diagonal matrix such that $\boldsymbol{S}_{ii} = 1$ if $[\boldsymbol{X}^\top w]_i > 0$ else $\boldsymbol{S}_{ii} = 0$ for $i \in [n]$, then we can obtain that

$$\begin{aligned}
\tau_\lambda(w) &:= p(w) \cdot \boldsymbol{\Phi}_1(w)^\top (\boldsymbol{K}_1 + \lambda \boldsymbol{I})^{-1} \boldsymbol{\Phi}_1(w) \\
&= 2\, p(w) \cdot w^\top \boldsymbol{X}\boldsymbol{S} (\boldsymbol{K}_1 + \lambda \boldsymbol{I})^{-1} \boldsymbol{S}\boldsymbol{X}^\top w. \qquad (81)
\end{aligned}$$

Now note that by definition of $\boldsymbol{K}_1$, we have $[\boldsymbol{K}_1]_{i,j} = \|x_i\|_2 \|x_j\|_2 \cdot \kappa_1\left(\frac{\langle x_i, x_j \rangle}{\|x_i\|_2 \|x_j\|_2}\right)$. Therefore, using the Taylor expansion of the function $\kappa_1(\alpha) = \frac{1}{\pi} + \frac{\alpha}{2} + \frac{1}{\pi} \cdot \sum_{i=0}^{\infty} \frac{(2i)!}{2^{2i} \cdot (i!)^2 \cdot (2i+1) \cdot (2i+2)} \cdot \alpha^{2i+2}$ and the fact that its Taylor coefficients are all non-negative, we have,

$$\frac{1}{2} \boldsymbol{X}^\top \boldsymbol{X} \preceq \boldsymbol{K}_1.$$

Using the above inequality along with Eq. (81), we can write,

$$\begin{aligned}
\tau_\lambda(w) &\le 2\, p(w) \cdot w^\top \boldsymbol{X}\boldsymbol{S} \left(\frac{1}{2} \boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I} \right)^{-1} \boldsymbol{S}\boldsymbol{X}^\top w \\
&\le 2\, p(w) \cdot \|w\|_2^2 \cdot \left\| \boldsymbol{X}\boldsymbol{S} \left(\frac{1}{2} \boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I} \right)^{-1} \boldsymbol{S}\boldsymbol{X}^\top \right\|_2 \qquad (82)
\end{aligned}$$

To obtain an upper bound of the third term in Eq. (82), we consider the singular value decomposition of $\boldsymbol{X} = \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^\top$. And we have

$$\begin{aligned}
\left\| \boldsymbol{X}\boldsymbol{S} \left(\frac{1}{2} \boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I} \right)^{-1} \boldsymbol{S}\boldsymbol{X}^\top \right\|_2 &= \left\| \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^\top \boldsymbol{S} \left(\frac{1}{2} \boldsymbol{U}\boldsymbol{\Sigma}^2\boldsymbol{U}^\top + \lambda \boldsymbol{I} \right)^{-1} \boldsymbol{S}\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top \right\|_2 \\
&= \left\| \boldsymbol{\Sigma}\boldsymbol{U}^\top \boldsymbol{S} \left(\frac{1}{2} \boldsymbol{U}\boldsymbol{\Sigma}^2\boldsymbol{U}^\top + \lambda \boldsymbol{I} \right)^{-1} \boldsymbol{S}\boldsymbol{U}\boldsymbol{\Sigma} \right\|_2 \\
&= \left\| \boldsymbol{\Sigma}\boldsymbol{U}^\top \boldsymbol{S}\boldsymbol{U} \left(\frac{1}{2} \boldsymbol{\Sigma}^2 + \lambda \boldsymbol{I} \right)^{-1} \boldsymbol{U}^\top \boldsymbol{S}\boldsymbol{U}\boldsymbol{\Sigma} \right\|_2 \qquad (83)
\end{aligned}$$

Now we observe that

$$[\boldsymbol{U}^\top \boldsymbol{S}\boldsymbol{U}]_{ij} \le \|u_i\|_2 \|u_j\|_2 \qquad (84)$$

which leads us to $\left\|\boldsymbol{U}^\top \boldsymbol{S}\boldsymbol{U}\right\|_F \le \mathrm{rank}(\boldsymbol{U}) = \mathrm{rank}(\boldsymbol{X})$. Since $\boldsymbol{U}^\top \boldsymbol{S}\boldsymbol{U}$ is a positive semi-definite matrix we have,

$$0 \preceq \boldsymbol{U}^\top \boldsymbol{S}\boldsymbol{U} \preceq \mathrm{rank}(\boldsymbol{X}) \cdot \boldsymbol{I}.$$

Therefore, plugging this into Eq. (83) and Eq. (82) gives,

$$\tau_\lambda(w) \le 2\, p(w) \cdot \|w\|_2^2 \cdot \mathrm{rank}(\boldsymbol{X})^2.$$

Algorithm 3 Gibbs Sampling for Approximating Eq. (86) via Inverse Transformation Method

Input: $\boldsymbol{X} \in \mathbb{R}^{d\times n}$, feature dimension m, Gibbs iterations T
Draw i.i.d. $w_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$ for $i \in [m]$
for $i = 1$ to m **do**
  $q(x, z) \leftarrow$ inverse of $\frac{\mathrm{erf}\left(x/\sqrt{2}\right)+1}{2} - \frac{x\exp\left(-x^2/2\right)}{\sqrt{2\pi}(z+1)}$ (i.e., CDF of $\Pr([w_i]_j|[w_i]_{\setminus\{j\}})$)
  for $t = 1$ to T **do**
    for $j = 1$ to d **do**
      $u \leftarrow$ sample from $[0, 1]$ at uniformly random
      $[w_i]_j \leftarrow q\left(u, \sum_{k\in[d]\setminus\{j\}}[w_i]_k^2\right)$
return $\sqrt{\frac{2d}{m}}\left[\frac{\mathrm{ReLU}(\boldsymbol{X}^\top w_1)}{\|w_1\|_2}, \ldots, \frac{\mathrm{ReLU}(\boldsymbol{X}^\top w_m)}{\|w_m\|_2}\right]$

Denote $\widetilde{\tau}(w) := 2\,p(w)\,\|w\|_2^2 \min\left\{\mathrm{rank}(\boldsymbol{X})^2, \frac{\|\boldsymbol{X}\|_2^2}{\lambda}\right\}$ and it holds that

$$\int_{\mathbb{R}^d} \widetilde{\tau}(w)dw = 2d\min\left\{\mathrm{rank}(\boldsymbol{X})^2, \frac{\|\boldsymbol{X}\|_2^2}{\lambda}\right\} \tag{85}$$

since $\int_{\mathbb{R}^d} p(w)\,\|w\|_2^2 = \mathtt{tr}(\boldsymbol{I}_d) = d$ for $w \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$. We define the modified distribution as

$$q(w) := \frac{\widetilde{\tau}(w)}{\int_{\mathbb{R}^d}\widetilde{\tau}(w)dw} = p(w)\frac{\|w\|_2^2}{d} = \frac{1}{(2\pi)^{d/2}d}\|w\|_2^2\exp\left(-\frac{1}{2}\|w\|_2^2\right) \tag{86}$$

and recall that the modified random features are defined as

$$\boldsymbol{\Phi}_1 = \frac{1}{\sqrt{m}}\left[\sqrt{\frac{p(w_1)}{q(w_1)}}\,\boldsymbol{\Phi}_1(w_1), \ldots, \sqrt{\frac{p(w_m)}{q(w_m)}}\,\boldsymbol{\Phi}_1(w_m)\right]^\top \tag{87}$$

$$= \sqrt{\frac{2d}{m}}\left[\frac{\mathrm{ReLU}(\boldsymbol{X}^\top w_1)}{\|w_1\|_2}, \ldots, \frac{\mathrm{ReLU}(\boldsymbol{X}^\top w_m)}{\|w_m\|_2}\right]^\top. \tag{88}$$

Putting all together into Theorem 5, we derive the result. This completes the proof of Theorem 7. □

Approximate sampling. It is not trivial to sample a vector $w \in \mathbb{R}^d$ from the distribution $q(\cdot)$ defined in Eq. (86). Thus, we suggest to perform an approximate sampling via Gibbs sampling. The algorithm starts with a random initialized vector w and then iteratively replaces $[w]_i$ with a sample

$$q([w]_i|[w]_{\setminus\{i\}}) \propto \frac{\|w\|_2^2}{1+\|w\|_2^2-[w]_i^2}\exp\left(-\frac{[w]_i^2}{2}\right) \tag{89}$$

for $i \in [d]$ and repeats this process for T iterations. Sampling from $q([w]_i|[w]_{\setminus\{i\}})$ can be done via the inverse transformation method.[‡] We empirically verify that $T = 1$ is enough for promising performances. The running time of Gibbs sampling becomes $\mathcal{O}(m_1 dT)$ where m_1 corresponds to the number of independent samples from $q(v)$. This is negligible compared to the feature map construction with PolySketch for $T = \mathcal{O}(1)$. The pseudo-code for the modified random features of A_1 using Gibbs sampling is outlined in Algorithm 3.

E.3 Proof of Theorem 3

Our proof relies on spectral approximation bounds of PolySketch given in the fourth part of Lemma 1.

Theorem 3. *Given a dataset $\boldsymbol{X} \in \mathbb{R}^{d\times n}$ with $\|\boldsymbol{X}_{(:,i)}\|_2 \le 1$ for every $i \in [n]$, let $\boldsymbol{K}_{\mathtt{ntk}}, \boldsymbol{K}_0, \boldsymbol{K}_1$ be kernel matrices for two-layer ReLU NTK and arc-cosine kernels of 0^{th} and 1^{st} order, respectively. For any $\lambda > 0$, suppose s_λ is the statistical dimension of $\boldsymbol{K}_{\mathtt{ntk}}$. Modify Algorithm 2 by*

[‡]It requires the CDF of $q([w]_i|[w]_{\setminus\{i\}})$ which is equivalent to $\frac{\mathrm{erf}\left([w]_i/\sqrt{2}\right)+1}{2} - \frac{[w]_i\exp\left(-[w]_i^2/2\right)}{\sqrt{2\pi}(1+\|w\|_2^2-[w]_i^2)}$.

replacing $\Phi_1(\cdot)$ *in line 5 with* $\widetilde{\Phi}_1(\cdot)$ *defined in Eq. (15). For any* $\varepsilon, \delta > 0$*, let* $\mathbf{\Psi}_{\mathtt{rf}}^{(L)} \in \mathbb{R}^{(m_1+m_s)\times n}$ *be the output matrix of this algorithm with* $L = 1$*. There exist* $m_0 = \mathcal{O}\left(\frac{n}{\varepsilon^2\lambda}\log\frac{s_\lambda}{\delta}\right), m_1 = \mathcal{O}\left(\frac{d}{\varepsilon^2}\cdot\min\left\{\mathrm{rank}(\boldsymbol{X})^2, \frac{\|\boldsymbol{X}\|_2^2}{\lambda}\right\}\log\frac{s_\lambda}{\delta}\right), m_s = \mathcal{O}\left(\frac{1}{\varepsilon^2}\cdot\frac{n}{1+\lambda}\log^3\frac{n}{\varepsilon\delta}\right)$ *such that,*

$$\Pr\left[(1-\varepsilon)\left(\boldsymbol{K}_{\mathtt{ntk}}+\lambda\boldsymbol{I}\right) \preceq \left(\mathbf{\Psi}_{\mathtt{rf}}^{(L)}\right)^\top \mathbf{\Psi}_{\mathtt{rf}}^{(L)} + \lambda\boldsymbol{I} \preceq (1+\varepsilon)\left(\boldsymbol{K}_{\mathtt{ntk}}+\lambda\boldsymbol{I}\right)\right] \geq 1-\delta. \tag{17}$$

Proof of Theorem 3: Note that the NTK of two-layer ReLU network can be formulated as

$$\boldsymbol{K}_{\mathtt{ntk}} = \boldsymbol{K}_1 + \boldsymbol{K}_0 \odot (\boldsymbol{X}^\top\boldsymbol{X}) \tag{90}$$

where $\boldsymbol{K}_0$ and $\boldsymbol{K}_1$ are the arc-cosine kernel matrices of order 0 and 1 with dataset $\boldsymbol{X}$, respectively.

Let $\mathbf{\Phi}_0$ and $\mathbf{\Phi}_1$ be the random features of $\boldsymbol{K}_0$ and $\boldsymbol{K}_1$, defined as per Theorem 6 and Theorem 7, respectively. Also let $\mathbf{\Psi}_{\mathtt{rf}}$ be the feature matrix that Algorithm 2 outputs, that is each column of this matrix is obtained by applying this algorithm on the dataset $\boldsymbol{X}$. By basic properties of tensor products we have,

$$\mathbf{\Psi}_{\mathtt{rf}} := \mathbf{\Phi}_1 \oplus \boldsymbol{Q}^2\cdot(\mathbf{\Phi}_0\otimes\boldsymbol{X}). \tag{91}$$

Our proof is a combination of spectral analysis of $\mathbf{\Phi}_0^\top\mathbf{\Phi}_0, \mathbf{\Phi}_1^\top\mathbf{\Phi}_1$ and $\left(\boldsymbol{Q}^2(\mathbf{\Phi}_0\otimes\boldsymbol{X})\right)^\top\cdot\boldsymbol{Q}^2(\mathbf{\Phi}_0\otimes\boldsymbol{X})$ which are stated in Theorem 6, Theorem 7 and Lemma 1, respectively.

From Theorem 7, if $m_1 \geq \frac{16}{3}\frac{d}{\varepsilon^2}\min\left\{\mathrm{rank}(\boldsymbol{X})^2, \frac{\|\boldsymbol{X}\|_2^2}{\lambda}\right\}\log\left(\frac{48 s_\lambda}{\delta}\right)$ then with probability at least $1-\frac{\delta}{3}$ the following holds,

$$(1-\varepsilon)\left(\boldsymbol{K}_1+\frac{\lambda}{2}\boldsymbol{I}\right) \preceq \mathbf{\Phi}_1^\top\mathbf{\Phi}_1 + \frac{\lambda}{2}\boldsymbol{I} \preceq (1+\varepsilon)\left(\boldsymbol{K}_1+\frac{\lambda}{2}\boldsymbol{I}\right). \tag{92}$$

From Theorem 6, if $m_0 \geq 48\frac{n}{\lambda\varepsilon^2}\log\left(\frac{48 s_\lambda}{\delta}\right)$ then with probability at least $1-\frac{\delta}{3}$ it holds that,

$$\left(1-\frac{\varepsilon}{3}\right)\left(\boldsymbol{K}_0+\frac{\lambda}{2}\boldsymbol{I}\right) \preceq \mathbf{\Phi}_0^\top\mathbf{\Phi}_0 + \frac{\lambda}{2}\boldsymbol{I} \preceq \left(1+\frac{\varepsilon}{3}\right)\left(\boldsymbol{K}_0+\frac{\lambda}{2}\boldsymbol{I}\right) \tag{93}$$

Rearranging Eq. (93), we get

$$\mathbf{\Phi}_0^\top\mathbf{\Phi}_0 \preceq \left(1+\frac{\varepsilon}{3}\right)\boldsymbol{K}_0 + \frac{\varepsilon}{6}\lambda\boldsymbol{I}.$$

Now we bound the trace of $(\mathbf{\Phi}_0\otimes\boldsymbol{X})^\top\cdot(\mathbf{\Phi}_0\otimes\boldsymbol{X}) = \mathbf{\Phi}_0^\top\mathbf{\Phi}_0\odot\boldsymbol{X}^\top\boldsymbol{X}$:

$$\begin{aligned}
\mathtt{tr}\left(\mathbf{\Phi}_0^\top\mathbf{\Phi}_0\odot\boldsymbol{X}^\top\boldsymbol{X}\right) &= \sum_{j\in[n]}[\mathbf{\Phi}_0^\top\mathbf{\Phi}_0]_{j,j}\cdot[\boldsymbol{X}^\top X]_{j,j} \\
&\leq \sum_{j\in[n]}[\mathbf{\Phi}_0^\top\mathbf{\Phi}_0]_{j,j} \\
&\leq n.
\end{aligned}$$

Now note that, we can write,

$$s_\lambda\left((\mathbf{\Phi}_0\otimes\boldsymbol{X})^\top(\mathbf{\Phi}_0\otimes\boldsymbol{X})\right) \leq \frac{\mathtt{tr}\left(\mathbf{\Phi}_0^\top\mathbf{\Phi}_0\odot\boldsymbol{X}^\top\boldsymbol{X}\right)}{\mathtt{tr}\left(\mathbf{\Phi}_0^\top\mathbf{\Phi}_0\odot\boldsymbol{X}^\top\boldsymbol{X}\right)/n+\lambda} \leq \frac{n}{1+\lambda}. \tag{94}$$

To guarantee spectral approximation of $\left(\boldsymbol{Q}^2(\mathbf{\Phi}_0\otimes\boldsymbol{X})\right)^\top\cdot\boldsymbol{Q}^2(\mathbf{\Phi}_0\otimes\boldsymbol{X})$, we will use the result of Lemma 1. Using Eq. (94) along with Lemma 1 and the fact that $m_s \geq \frac{C}{\varepsilon^2}\cdot\frac{n}{1+\lambda}\log^3\frac{n}{\varepsilon\delta}$ for some

constant C, and union bound, with probability at least $1-\frac{\delta}{2}$, we have

$$\begin{aligned}
\left(\boldsymbol{Q}^2\left(\boldsymbol{\Phi}_0 \otimes \boldsymbol{X}\right)\right)^\top \boldsymbol{Q}^2\left(\boldsymbol{\Phi}_0 \otimes \boldsymbol{X}\right) + \frac{\lambda}{2}\boldsymbol{I} &\preceq \left(1+\frac{\varepsilon}{3}\right)\left(\boldsymbol{\Phi}_0^\top \boldsymbol{\Phi}_0 \odot \boldsymbol{X}^\top \boldsymbol{X} + \frac{\lambda}{2}\boldsymbol{I}\right) \\
&\preceq \left(1+\frac{\varepsilon}{3}\right)\left(\left[\left(1+\frac{\varepsilon}{3}\right)\boldsymbol{K}_0 + \frac{\varepsilon}{6}\lambda \boldsymbol{I}\right] \odot \boldsymbol{X}^\top \boldsymbol{X} + \frac{\lambda}{2}\boldsymbol{I}\right) \\
&= \left(1+\frac{\varepsilon}{3}\right)\left(\left(1+\frac{\varepsilon}{3}\right)\left(\boldsymbol{K}_0 \odot \boldsymbol{X}^\top \boldsymbol{X}\right) + \frac{\varepsilon}{6}\lambda(\boldsymbol{I} \odot \boldsymbol{X}^\top \boldsymbol{X}) + \frac{\lambda}{2}\boldsymbol{I}\right) \\
&\preceq \left(1+\frac{\varepsilon}{3}\right)\left(1+\frac{\varepsilon}{3}\right)\left(\boldsymbol{K}_0 \odot \boldsymbol{X}^\top \boldsymbol{X} + \frac{\lambda}{2}\boldsymbol{I}\right) \\
&\preceq (1+\varepsilon)\left(\boldsymbol{K}_0 \odot \boldsymbol{X}^\top \boldsymbol{X} + \frac{\lambda}{2}\boldsymbol{I}\right) \qquad (95)
\end{aligned}$$

where the inequality in second line follows from Lemma 9 and the fourth line follows from the assumption $\left\|\boldsymbol{X}_{(:,i)}\right\|_2 \leq 1$ for all $i \in [n]$ which leads that $\boldsymbol{I} \odot (\boldsymbol{X}^\top \boldsymbol{X}) \preceq \boldsymbol{I}$. The last inequality holds since $\varepsilon \in (0, 1/2)$.

Similarly, we can obtain the following lower bound:

$$\left(\boldsymbol{Q}^2\left(\boldsymbol{\Phi}_0 \otimes \boldsymbol{X}\right)\right)^\top \boldsymbol{Q}^2\left(\boldsymbol{\Phi}_0 \otimes \boldsymbol{X}\right) + \frac{\lambda}{2}\boldsymbol{I} \succeq (1-\varepsilon)\left(\boldsymbol{K}_0 \odot \boldsymbol{X}^\top \boldsymbol{X} + \frac{\lambda}{2}\boldsymbol{I}\right). \qquad (96)$$

Combining Eq. (92), Eq. (95) and Eq. (96) gives

$$(1-\varepsilon)\left(\boldsymbol{K}_{\texttt{ntk}} + \lambda \boldsymbol{I}\right) \preceq \boldsymbol{\Psi}_{\texttt{rf}}^\top \boldsymbol{\Psi}_{\texttt{rf}} + \lambda \boldsymbol{I} \preceq (1+\varepsilon)\left(\boldsymbol{K}_{\texttt{ntk}} + \lambda \boldsymbol{I}\right). \qquad (97)$$

Furthermore, by taking a union bound over all events, Eq. (97) holds with probability at least $1-\delta$. This completes the proof of Theorem 3. □

E.4 Auxiliary Lemmas

Lemma 9. *If $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ are positive semi-definite matrices of conforming sizes such that $\boldsymbol{B} \preceq \boldsymbol{C}$, then,*

$$\boldsymbol{A} \odot \boldsymbol{B} \preceq \boldsymbol{A} \odot \boldsymbol{C}.$$

Proof of Lemma 9: We want to show that for any vector v, $v^\top \boldsymbol{A} \odot \boldsymbol{B} v \leq v^\top \boldsymbol{A} \odot \boldsymbol{C} v$. Because the matrices $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ are PSD, there exist matrices $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$ of appropriate sizes such that we can decompose these matrices as follows,

$$\boldsymbol{A} = \boldsymbol{X}^\top \boldsymbol{X}, \quad \boldsymbol{B} = \boldsymbol{Y}^\top \boldsymbol{Y}, \quad \boldsymbol{C} = \boldsymbol{Z}^\top \boldsymbol{Z}.$$

Using this and basic properties of tensor products, we have the following for any vector v,

$$\begin{aligned}
v^\top \boldsymbol{A} \odot \boldsymbol{B} v &= \|\boldsymbol{X} \otimes \boldsymbol{Y} v\|_2^2 \\
&= \|\boldsymbol{X} \cdot \texttt{diag}(v) \cdot \boldsymbol{Y}^\top\|_F^2 \\
&= \sum_i (\boldsymbol{X}_{(i,:)} \odot v)^\top \cdot \boldsymbol{B} \cdot (\boldsymbol{X}_{(i,:)} \odot v) \\
&\leq \sum_i (\boldsymbol{X}_{(i,:)} \odot v)^\top \cdot \boldsymbol{C} \cdot (\boldsymbol{X}_{(i,:)} \odot v) \\
&= v^\top \boldsymbol{A} \odot \boldsymbol{C} v.
\end{aligned}$$

This completes the proof of Lemma 9. □

F ReLU-CNTK: Expression and Main Properties

In this section we prove that the depth-L CNTK corresponding to ReLU activation is highly structured and can be fully characterized in terms of tensoring and composition of arc-cosine kernel functions $\kappa_1(\cdot)$ and $\kappa_0(\cdot)$. We refer to this kernel function as **ReLU-CNTK**. First we start by restating the DP proposed by Arora et al. [5] for computing the L-layer CNTK kernel corresponding to an arbitrary activation function $\sigma : \mathbb{R} \to \mathbb{R}$ and convolutional filters of size $q \times q$, with GAP:

1. Let $y, z \in \mathbb{R}^{d_1 \times d_2 \times c}$ be two input images, where c is the number of channels ($c = 3$ for RGB images). Define $\Gamma^{(0)} : \mathbb{R}^{d_1 \times d_2 \times c} \times \mathbb{R}^{d_1 \times d_2 \times c} \to \mathbb{R}^{d_1 \times d_2 \times d_1 \times d_2}$ and $\Sigma^{(0)} : \mathbb{R}^{d_1 \times d_2 \times c} \times \mathbb{R}^{d_1 \times d_2 \times c} \to \mathbb{R}^{d_1 \times d_2 \times d_1 \times d_2}$ as follows for every $i, i' \in [d_1]$ and $j, j' \in [d_2]$:

$$\begin{aligned} \Gamma^{(0)}(y, z) &:= \sum_{l=1}^{c} y_{(:,:,l)} \otimes z_{(:,:,l)}, \\ \Sigma^{(0)}_{i,j,i',j'}(y, z) &:= \sum_{a=-\frac{q-1}{2}}^{\frac{q-1}{2}} \sum_{b=-\frac{q-1}{2}}^{\frac{q-1}{2}} \Gamma^{(0)}_{i+a,j+b,i'+a,j'+b}(y, z). \end{aligned} \tag{98}$$

2. For every layer $h = 1, 2, \ldots, L$ of the network and every $i, i' \in [d_1]$ and $j, j' \in [d_2]$, define $\Gamma^{(h)} : \mathbb{R}^{d_1 \times d_2 \times c} \times \mathbb{R}^{d_1 \times d_2 \times c} \to \mathbb{R}^{d_1 \times d_2 \times d_1 \times d_2}$ recursively as:

$$\begin{aligned} \Lambda^{(h)}_{i,j,i',j'}(y, z) &:= \begin{pmatrix} \Sigma^{(h-1)}_{i,j,i,j}(y, y) & \Sigma^{(h-1)}_{i,j,i',j'}(y, z) \\ \Sigma^{(h-1)}_{i',j',i,j}(z, y) & \Sigma^{(h-1)}_{i',j',i',j'}(z, z) \end{pmatrix}, \\ \Gamma^{(h)}_{i,j,i',j'}(y, z) &:= \frac{1}{q^2 \cdot \mathbb{E}_{w \sim \mathcal{N}(0,1)}\left[|\sigma(w)|^2\right]} \cdot \mathbb{E}_{(u,v) \sim \mathcal{N}\left(0, \Lambda^{(h)}_{i,j,i',j'}(y,z)\right)}\left[\sigma(u) \cdot \sigma(v)\right], \\ \Sigma^{(h)}_{i,j,i',j'}(y, z) &:= \sum_{a=-\frac{q-1}{2}}^{\frac{q-1}{2}} \sum_{b=-\frac{q-1}{2}}^{\frac{q-1}{2}} \Gamma^{(h)}_{i+a,j+b,i'+a,j'+b}(y, z), \end{aligned} \tag{99}$$

3. For every $h = 1, 2, \ldots L$, every $i, i' \in [d_1]$ and $j, j' \in [d_2]$, define $\dot{\Gamma}^{(h)}(y, z) \in \mathbb{R}^{d_1 \times d_2 \times d_1 \times d_2}$ as:

$$\dot{\Gamma}^{(h)}_{i,j,i',j'}(y, z) := \frac{1}{q^2 \cdot \mathbb{E}_{w \sim \mathcal{N}(0,1)}\left[|\sigma(w)|^2\right]} \cdot \mathbb{E}_{(u,v) \sim \mathcal{N}\left(0, \Lambda^{(h)}_{i,j,i',j'}(y,z)\right)}\left[\dot{\sigma}(u) \cdot \dot{\sigma}(v)\right]. \tag{100}$$

4. Let $\Pi^{(0)}(y, z) := 0$ and for every $h = 1, 2, \ldots, L-1$, every $i, i' \in [d_1]$ and $j, j' \in [d_2]$, define $\Pi^{(h)} : \mathbb{R}^{d_1 \times d_2 \times c} \times \mathbb{R}^{d_1 \times d_2 \times c} \to \mathbb{R}^{d_1 \times d_2 \times d_1 \times d_2}$ recursively as:

$$\begin{aligned} \Pi^{(h)}_{i,j,i',j'}(y, z) &:= \sum_{a=-\frac{q-1}{2}}^{\frac{q-1}{2}} \sum_{b=-\frac{q-1}{2}}^{\frac{q-1}{2}} \left[\Pi^{(h-1)}(y, z) \odot \dot{\Gamma}^{(h)}(y, z) + \Gamma^{(h)}(y, z)\right]_{i+a,j+b,i'+a,j'+b}, \\ \Pi^{(L)}(y, z) &:= \Pi^{(L-1)}(y, z) \odot \dot{\Gamma}^{(L)}(y, z). \end{aligned} \tag{101}$$

5. The final CNTK expressions is defined as:

$$\Theta^{(L)}_{\texttt{cntk}}(y, z) := \frac{1}{d_1^2 d_2^2} \cdot \sum_{i,i' \in [d_1]} \sum_{j,j' \in [d_2]} \Pi^{(L)}_{i,j,i',j'}(y, z). \tag{102}$$

Now we show how to recursively compute the ReLU-CNTK as follows,

Definition 2 (ReLU-CNTK)**.** For every positive integers q, L, the L-layer CNTK for ReLU activation function and convolutional filter size of $q \times q$ is defined as follows

1. For $x \in \mathbb{R}^{d_1 \times d_2 \times c}$, every $i \in [d_1]$ and $j \in [d_2]$ let $N^{(0)}_{i,j}(x) := q^2 \cdot \sum_{l=1}^{c} |x_{i+a,j+b,l}|^2$, and for every $h \geq 1$, recursively define,

$$N^{(h)}_{i,j}(x) := \frac{1}{q^2} \cdot \sum_{a=-\frac{q-1}{2}}^{\frac{q-1}{2}} \sum_{b=-\frac{q-1}{2}}^{\frac{q-1}{2}} N^{(h-1)}_{i+a,j+b}(x). \tag{103}$$

2. Define $\Gamma^{(0)}(y, z) := \sum_{l=1}^{c} y_{(:,:,l)} \otimes z_{(:,:,l)}$. Let $\kappa_1 : [-1, 1] \to \mathbb{R}$ be the function defined in Eq. (2) of Definition 1. For every layer $h = 1, 2, \ldots, L$, every $i, i' \in [d_1]$ and $j, j' \in [d_2]$, define $\Gamma^{(h)} : \mathbb{R}^{d_1 \times d_2 \times c} \times \mathbb{R}^{d_1 \times d_2 \times c} \to \mathbb{R}^{d_1 \times d_2 \times d_1 \times d_2}$ recursively as:

$$\Gamma^{(h)}_{i,j,i',j'}(y, z) := \frac{\sqrt{N^{(h)}_{i,j}(y) \cdot N^{(h)}_{i',j'}(z)}}{q^2} \cdot \kappa_1\left(\frac{\sum_{a=-\frac{q-1}{2}}^{\frac{q-1}{2}} \sum_{b=-\frac{q-1}{2}}^{\frac{q-1}{2}} \Gamma^{(h-1)}_{i+a,j+b,i'+a,j'+b}(y, z)}{\sqrt{N^{(h)}_{i,j}(y) \cdot N^{(h)}_{i',j'}(z)}}\right). \tag{104}$$

3. Let $\kappa_0 : [-1, 1] \to \mathbb{R}$ be the function defined in Eq. (2) of Definition 1. For every $h = 1, 2, \ldots L$, every $i, i' \in [d_1]$ and $j, j' \in [d_2]$, define $\dot{\Gamma}^{(h)}(y, z) \in \mathbb{R}^{d_1 \times d_2 \times d_1 \times d_2}$ as:

$$\dot{\Gamma}^{(h)}_{i,j,i',j'}(y, z) := \frac{1}{q^2} \cdot \kappa_0 \left(\frac{\sum_{a=-\frac{q-1}{2}}^{\frac{q-1}{2}} \sum_{b=-\frac{q-1}{2}}^{\frac{q-1}{2}} \Gamma^{(h-1)}_{i+a,j+b,i'+a,j'+b}(y, z)}{\sqrt{N^{(h)}_{i,j}(y) \cdot N^{(h)}_{i',j'}(z)}} \right). \tag{105}$$

4. Let $\Pi^{(0)}(y, z) := 0$ and for every $h = 1, 2, \ldots, L-1$, every $i, i' \in [d_1]$ and $j, j' \in [d_2]$, define $\Pi^{(h)} : \mathbb{R}^{d_1 \times d_2 \times c} \times \mathbb{R}^{d_1 \times d_2 \times c} \to \mathbb{R}^{d_1 \times d_2 \times d_1 \times d_2}$ recursively as:

$$\Pi^{(h)}_{i,j,i',j'}(y, z) := \sum_{a=-\frac{q-1}{2}}^{\frac{q-1}{2}} \sum_{b=-\frac{q-1}{2}}^{\frac{q-1}{2}} \left[\Pi^{(h-1)}(y, z) \odot \dot{\Gamma}^{(h)}(y, z) + \Gamma^{(h)}(y, z) \right]_{i+a,j+b,i'+a,j'+b}. \tag{106}$$

Furthermore, for $h = L$ define:

$$\Pi^{(L)}(y, z) := \Pi^{(L-1)}(y, z) \odot \dot{\Gamma}^{(L)}(y, z). \tag{107}$$

5. The final CNTK expressions for ReLU activation is:

$$\Theta^{(L)}_{\texttt{cntk}}(y, z) := \frac{1}{d_1^2 d_2^2} \cdot \sum_{i,i' \in [d_1]} \sum_{j,j' \in [d_2]} \Pi^{(L)}_{i,j,i',j'}(y, z). \tag{108}$$

In what follows we prove that the procedure in Definition 2 precisely computes the CNTK kernel function corresponding to ReLU activation and additionally, we present useful corollaries and consequences of this fact.

Lemma 10. *For every positive integers d_1, d_2, c, odd integer q, and every integer $h \geq 0$, if the activation function is ReLU, then the tensor covariances $\Gamma^{(h)}, \dot{\Gamma}^{(h)}(y, z) : \mathbb{R}^{d_1 \times d_2 \times c} \times \mathbb{R}^{d_1 \times d_2 \times c} \to \mathbb{R}^{d_1 \times d_2 \times d_1 \times d_2}$ defined in Eq. (99) and Eq. (100), are precisely equal to the tensor covariances defined in Eq. (104) and Eq. (105) of Definition 2, respectively.*

Proof of Lemma 10: To prove the lemma, we first show by induction on $h = 1, 2, \ldots$ that $N^{(h)}_{i,j}(x) \equiv \Sigma^{(h-1)}_{i,j,i,j}(x, x)$ for every $x \in \mathbb{R}^{d_1 \times d_2 \times c}$ and every $i \in [d_1]$ and $j \in [d_2]$, where $\Sigma^{(h-1)}(x, x)$ is defined as per Eq. (98) and Eq. (99). The **base of induction** trivially holds for $h = 1$ because by definition of $N^{(1)}(x)$ and Eq. (98) we have,

$$N^{(1)}_{i,j}(x) = \sum_{a=-\frac{q-1}{2}}^{\frac{q-1}{2}} \sum_{b=-\frac{q-1}{2}}^{\frac{q-1}{2}} \sum_{l=1}^{c} |x_{i+a,j+b,l}|^2 \equiv \Sigma^{(0)}_{i,j,i,j}(x, x).$$

To prove the **inductive step**, suppose that the inductive hypothesis $N^{(h-1)}_{i,j}(x) = \Sigma^{(h-2)}_{i,j,i,j}(x, x)$ holds for some $h \geq 2$. Now we show that conditioned on the inductive hypothesis, the inductive claim

holds. By Eq. (99), we have,

$$\begin{aligned}
\Sigma^{(h-1)}_{i,j,i,j}(x,x) &= \sum_{a=-\frac{q-1}{2}}^{\frac{q-1}{2}} \sum_{b=-\frac{q-1}{2}}^{\frac{q-1}{2}} \Gamma^{(h-1)}_{i+a,j+b,i+a,j+b}(x,x) \\
&= \sum_{a=-\frac{q-1}{2}}^{\frac{q-1}{2}} \sum_{b=-\frac{q-1}{2}}^{\frac{q-1}{2}} \frac{\mathbb{E}_{(u,v)\sim\mathcal{N}\left(0,\Lambda^{(h-1)}_{i+a,j+b,i+a,j+b}(x,x)\right)}\left[\sigma(u)\cdot\sigma(v)\right]}{q^2\cdot\mathbb{E}_{w\sim\mathcal{N}(0,1)}\left[|\sigma(w)|^2\right]} \\
&= \sum_{a=-\frac{q-1}{2}}^{\frac{q-1}{2}} \sum_{b=-\frac{q-1}{2}}^{\frac{q-1}{2}} \frac{\mathbb{E}_{u\sim\mathcal{N}\left(0,\Sigma^{(h-2)}_{i+a,j+b,i+a,j+b}(x,x)\right)}\left[|\max(0,u)|^2\right]}{q^2\cdot\mathbb{E}_{w\sim\mathcal{N}(0,1)}\left[|\max(0,w)|^2\right]} \\
&= \sum_{a=-\frac{q-1}{2}}^{\frac{q-1}{2}} \sum_{b=-\frac{q-1}{2}}^{\frac{q-1}{2}} \frac{1}{q^2}\cdot\Sigma^{(h-2)}_{i+a,j+b,i+a,j+b}(x,x) \\
&= \sum_{a=-\frac{q-1}{2}}^{\frac{q-1}{2}} \sum_{b=-\frac{q-1}{2}}^{\frac{q-1}{2}} \frac{1}{q^2}\cdot N^{(h-1)}_{i+a,j+b}(x) \equiv N^{(h)}_{i,j}(x). \qquad \text{(by Eq. (103))}
\end{aligned}$$

Therefore, this proves that $N^{(h)}_{i,j}(x) \equiv \Sigma^{(h-1)}_{i,j,i,j}(x,x)$ for every x and every integer $h \geq 1$.

Now, note that the 2×2 covariance matrix $\Lambda^{(h)}_{i,j,i',j'}(y,z)$, defined in Eq. (99), can be decomposed as $\Lambda^{(h)}_{i,j,i',j'}(y,z) = \begin{pmatrix} f^\top \\ g^\top \end{pmatrix}\cdot\begin{pmatrix} f & g \end{pmatrix}$, where $f,g\in\mathbb{R}^2$. Also note that $\|f\|_2^2 = \Sigma^{(h-1)}_{i,j,i,j}(y,y)$ and $\|g\|_2^2 = \Sigma^{(h-1)}_{i',j',i',j'}(z,z)$, hence, by what we proved above, we have,

$$\|f\|_2^2 = N^{(h)}_{i,j}(y), \text{ and } \|g\|_2^2 = N^{(h)}_{i',j'}(z).$$

Therefore, by Eq. (21), we can write:

$$\begin{aligned}
\Gamma^{(h)}_{i,j,i',j'}(y,z) &= \frac{1}{q^2\cdot\mathbb{E}_{w\sim\mathcal{N}(0,1)}\left[|\sigma(w)|^2\right]}\cdot\mathbb{E}_{(u,v)\sim\mathcal{N}\left(0,\Lambda^{(h)}_{i,j,i',j'}(y,z)\right)}\left[\sigma(u)\cdot\sigma(v)\right] \\
&= \frac{1}{q^2\cdot\mathbb{E}_{w\sim\mathcal{N}(0,1)}\left[|\sigma(w)|^2\right]}\cdot\mathbb{E}_{u\sim\mathcal{N}(0,I_d)}\left[\sigma(u^\top f)\cdot\sigma(u^\top g)\right] \\
&= \frac{2\cdot\|f\|_2\cdot\|g\|_2}{q^2\cdot\kappa_1(1)}\cdot\frac{1}{2}\cdot\kappa_1\left(\frac{\langle f,g\rangle}{\|f\|_2\cdot\|g\|_2}\right) \\
&= \frac{\sqrt{N^{(h)}_{i,j}(y)\cdot N^{(h)}_{i',j'}(z)}}{q^2}\cdot\kappa_1\left(\frac{\Sigma^{(h-1)}_{i,j,i',j'}(y,z)}{\sqrt{N^{(h)}_{i,j}(y)\cdot N^{(h)}_{i',j'}(z)}}\right) \\
&= \frac{\sqrt{N^{(h)}_{i,j}(y)\cdot N^{(h)}_{i',j'}(z)}}{q^2}\cdot\kappa_1\left(\frac{\sum_{a=-\frac{q-1}{2}}^{\frac{q-1}{2}}\sum_{b=-\frac{q-1}{2}}^{\frac{q-1}{2}}\Gamma^{(h-1)}_{i+a,j+b,i'+a,j'+b}(y,z)}{\sqrt{N^{(h)}_{i,j}(y)\cdot N^{(h)}_{i',j'}(z)}}\right),
\end{aligned}$$

where the third line follows from Eq. (21) and fourth line follows because we have $\langle f,g\rangle = \Sigma^{(h-1)}_{i,j,i',j'}(y,z)$. The fifth line above follows from Eq. (99). This proves the equivalence between the tensor covariance defined in Eq. (99) and the one defined in Eq. (104) of Definition 2. Similarly, by

using Eq. (21), we can prove the statement of the lemma about $\dot{\Gamma}^{(h)}_{i,j,i',j'}(y,z)$ as follows,

$$\begin{aligned}
\dot{\Gamma}^{(h)}_{i,j,i',j'}(y,z) &= \frac{1}{q^2 \cdot \mathbb{E}_{w\sim\mathcal{N}(0,1)}\left[|\sigma(w)|^2\right]} \cdot \mathbb{E}_{(u,v)\sim\mathcal{N}\left(0,\Lambda^{(h)}_{i,j,i',j'}(y,z)\right)}\left[\dot{\sigma}(u)\cdot\dot{\sigma}(v)\right] \\
&= \frac{1}{q^2 \cdot \mathbb{E}_{w\sim\mathcal{N}(0,1)}\left[|\sigma(w)|^2\right]} \cdot \mathbb{E}_{u\sim\mathcal{N}(0,I_d)}\left[\dot{\sigma}(u^\top f)\cdot\dot{\sigma}(u^\top g)\right] \\
&= \frac{2}{q^2\cdot\kappa_1(1)}\cdot\frac{1}{2}\cdot\kappa_0\left(\frac{\langle f,g\rangle}{\|f\|_2\cdot\|g\|_2}\right) \\
&= \frac{1}{q^2}\cdot\kappa_0\left(\frac{\Sigma^{(h-1)}_{i,j,i',j'}(y,z)}{\sqrt{N^{(h)}_{i,j}(y)\cdot N^{(h)}_{i',j'}(z)}}\right) \\
&= \frac{1}{q^2}\cdot\kappa_0\left(\frac{\sum_{a=-\frac{q-1}{2}}^{\frac{q-1}{2}}\sum_{b=-\frac{q-1}{2}}^{\frac{q-1}{2}}\Gamma^{(h-1)}_{i+a,j+b,i'+a,j'+b}(y,z)}{\sqrt{N^{(h)}_{i,j}(y)\cdot N^{(h)}_{i',j'}(z)}}\right).
\end{aligned}$$

This completes the proof of Lemma 10. □

Corollary 1 (Consequence of Lemma 10)**.** *Consider the preconditions of Lemma 10. For every* $x\in\mathbb{R}^{d_1\times d_2\times c}$, $N^{(h)}_{i,j}(x)\equiv\sum_{a=-\frac{q-1}{2}}^{\frac{q-1}{2}}\sum_{b=-\frac{q-1}{2}}^{\frac{q-1}{2}}\Gamma^{(h-1)}_{i+a,j+b,i+a,j+b}(x,x)$.

We describe some of the basic properties of the function $\Gamma^{(h)}(y,z)$ defined in Eq. (104) in the following lemma,

Lemma 11 (Properties of $\Gamma^{(h)}(y,z)$)**.** *For every images* $y,z\in\mathbb{R}^{d_1\times d_2\times c}$*, every integer* $h\geq 0$ *and every* $i,i'\in[d_1]$ *and* $j,j'\in[d_2]$ *the following properties are satisfied by functions* $\Gamma^{(h)}$ *and* $N^{(h)}$ *defined in Eq. (104) and Eq. (103) of Definition 2:*

1. **Cauchy–Schwarz inequality:** $\left|\Gamma^{(h)}_{i,j,i',j'}(y,z)\right|\leq\frac{\sqrt{N^{(h)}_{i,j}(y)\cdot N^{(h)}_{i',j'}(z)}}{q^2}$.
2. **Norm value:** $\Gamma^{(h)}_{i,j,i,j}(y,y)=\frac{N^{(h)}_{i,j}(y)}{q^2}\geq 0$.

Proof of Lemma 11: We prove the lemma by induction on h. The **base of induction** corresponds to $h=0$. In the base case, by Eq. (103) and Eq. (104) and Cauchy–Schwarz inequality, we have

$$\begin{aligned}
\left|\Gamma^{(0)}_{i,j,i',j'}(y,z)\right| &\equiv \left|\sum_{l=1}^{c} y_{i,j,l}\cdot z_{i',j',l}\right| \\
&\leq \sqrt{\sum_{l=1}^{c}|y_{i,j,l}|^2\cdot\sum_{l=1}^{c}|z_{i',j',l}|^2} \\
&= \frac{\sqrt{N^{(0)}_{i,j}(y)\cdot N^{(0)}_{i',j'}(z)}}{q^2}.
\end{aligned}$$

This proves the base for the first statement. Additionally we have, $\Gamma^{(0)}_{i,j,i,j}(y,y)=\sum_{l=1}^{c}y^2_{i,j,l}=\frac{N^{(0)}_{i,j}(y)}{q^2}\geq 0$ which proves the base for the second statement of the lemma. Now, in order to prove the inductive step, suppose that statements of the lemma hold for $h-1$, where $h\geq 1$. Then, conditioned on this, we prove that the lemma holds for h. First note that by conditioning on the inductive hypothesis, applying Cauchy–Schwarz inequality, and using the definition of $N^{(h)}$ in Eq. (103), we can write

$$\begin{aligned}
\left|\frac{\sum_{a=-\frac{q-1}{2}}^{\frac{q-1}{2}}\sum_{b=-\frac{q-1}{2}}^{\frac{q-1}{2}}\Gamma^{(h-1)}_{i+a,j+b,i'+a,j'+b}(y,z)}{\sqrt{N^{(h)}_{i,j}(y)\cdot N^{(h)}_{i',j'}(z)}}\right| &\leq \frac{\sum_{a=-\frac{q-1}{2}}^{\frac{q-1}{2}}\sum_{b=-\frac{q-1}{2}}^{\frac{q-1}{2}}\sqrt{\frac{N^{(h-1)}_{i+a,j+b}(y)}{q^2}\cdot\frac{N^{(h-1)}_{i'+a,j'+b}(z)}{q^2}}}{\sqrt{N^{(h)}_{i,j}(y)\cdot N^{(h)}_{i',j'}(z)}} \\
&\leq 1.
\end{aligned}$$

Thus, by monotonicity of the function $\kappa_1 : [-1, 1] \to \mathbb{R}$, we can write,

$$\begin{aligned}\left|\Gamma^{(h)}_{i,j,i',j'}(y,z)\right| &\equiv \frac{\sqrt{N^{(h)}_{i,j}(y)\cdot N^{(h)}_{i',j'}(z)}}{q^2}\cdot\kappa_1\left(\frac{\sum_{a=-\frac{q-1}{2}}^{\frac{q-1}{2}}\sum_{b=-\frac{q-1}{2}}^{\frac{q-1}{2}}\Gamma^{(h-1)}_{i+a,j+b,i'+a,j'+b}(y,z)}{\sqrt{N^{(h)}_{i,j}(y)\cdot N^{(h)}_{i',j'}(z)}}\right)\\ &\leq \frac{\sqrt{N^{(h)}_{i,j}(y)\cdot N^{(h)}_{i',j'}(z)}}{q^2}\cdot\kappa_1(1)\\ &= \frac{\sqrt{N^{(h)}_{i,j}(y)\cdot N^{(h)}_{i',j'}(z)}}{q^2},\end{aligned}$$

where the second line above follows because of the fact that $\kappa_1(\cdot)$ is a monotonically increasing function. This completes the inductive step for the first statement of lemma. Now we prove the inductive step for the second statement as follows,

$$\begin{aligned}\Gamma^{(h)}_{i,j,i,j}(y,y) &\equiv \frac{N^{(h)}_{i,j}(y)}{q^2}\cdot\kappa_1\left(\frac{\sum_{a=-\frac{q-1}{2}}^{\frac{q-1}{2}}\sum_{b=-\frac{q-1}{2}}^{\frac{q-1}{2}}\Gamma^{(h-1)}_{i+a,j+b,i+a,j+b}(y,y)}{N^{(h)}_{i,j}(y)}\right)\\ &= \frac{N^{(h)}_{i,j}(y)}{q^2}\cdot\kappa_1(1)\\ &= \frac{N^{(h)}_{i,j}(y)}{q^2}\geq 0,\end{aligned}$$

where we used Corollary 1 to conclude that $\sum_{a=-\frac{q-1}{2}}^{\frac{q-1}{2}}\sum_{b=-\frac{q-1}{2}}^{\frac{q-1}{2}}\Gamma^{(h-1)}_{i+a,j+b,i+a,j+b}(y,y) = N^{(h)}_{i,j}(y)$ and then used the fact that $N^{(h)}_{i,j}(y)$ is non-negative. This completes the inductive proof of the lemma. This completes the proof of Lemma 11. □

We also describe some of the main properties of the function $\dot{\Gamma}^{(h)}(y,z)$ defined in Eq. (105) in the following lemma,

Lemma 12 (Properties of $\dot{\Gamma}^{(h)}(y,z)$)**.** *For every images $y, z \in \mathbb{R}^{d_1\times d_2\times c}$, every integer $h \geq 0$ and every $i, i' \in [d_1]$ and $j, j' \in [d_2]$ the following properties are satisfied by function $\dot{\Gamma}^{(h)}$ defined in Eq. (105) of Definition 2:*

1. **Cauchy–Schwarz inequality:** $\left|\dot{\Gamma}^{(h)}_{i,j,i',j'}(y,z)\right| \leq \frac{1}{q^2}$.
2. **Norm value:** $\dot{\Gamma}^{(h)}_{i,j,i,j}(y,y) = \frac{1}{q^2} \geq 0$.

Proof of Lemma 12: First, note that by Lemma 11 and the definition of $N^{(h)}$ in Eq. (103) we have,

$$\begin{aligned}\left|\frac{\sum_{a=-\frac{q-1}{2}}^{\frac{q-1}{2}}\sum_{b=-\frac{q-1}{2}}^{\frac{q-1}{2}}\Gamma^{(h-1)}_{i+a,j+b,i'+a,j'+b}(y,z)}{\sqrt{N^{(h)}_{i,j}(y)\cdot N^{(h)}_{i',j'}(z)}}\right| &\leq \frac{\sum_{a=-\frac{q-1}{2}}^{\frac{q-1}{2}}\sum_{b=-\frac{q-1}{2}}^{\frac{q-1}{2}}\sqrt{\frac{N^{(h-1)}_{i+a,j+b}(y)}{q^2}\cdot\frac{N^{(h-1)}_{i'+a,j'+b}(z)}{q^2}}}{\sqrt{N^{(h)}_{i,j}(y)\cdot N^{(h)}_{i',j'}(z)}}\\ &\leq 1.\end{aligned}$$

Thus, by monotonicity of function $\kappa_0 : [-1, 1] \to \mathbb{R}$ and using Eq. (105), we can write, $\dot{\Gamma}^{(h)}_{i,j,i',j'}(y,z) \leq \frac{1}{q^2}\cdot\kappa_0(1) = \frac{1}{q^2}$. Moreover, the equality is achieved when $y = z$ and $i = i'$ and $j = j'$. This proves both statements of the lemma. □

We also need to use some properties of $\Pi^{(h)}(y,z)$ defined in Eq. (106) and Eq. (107). We present these propertied in the next lemma,

Lemma 13 (Properties of $\Pi^{(h)}$)**.** *For every images $y, z \in \mathbb{R}^{d_1\times d_2\times c}$, every integer $h \geq 0$ and every $i \in [d_1]$ and $j \in [d_2]$ the following properties are satisfied by the function $\Pi^{(h)}$ defined in Eq. (106) and Eq. (107) of Definition 2:*

1. **Cauchy–Schwarz inequality:** $\Pi^{(h)}_{i,j,i',j'}(y,z) \le \sqrt{\Pi^{(h)}_{i,j,i,j}(y,y) \cdot \Pi^{(h)}_{i',j',i',j'}(z,z)}$.

2. **Norm value:** $\Pi^{(h)}_{i,j,i,j}(y,y) = \begin{cases} h \cdot N^{(h+1)}_{i,j}(y) & \text{if } h < L \\ \frac{L-1}{q^2} \cdot N^{(L)}_{i,j}(y) & \text{if } h = L \end{cases}$.

Proof of Lemma 13: The proof is by induction on h. The base of induction corresponds to $h = 0$. By definition of $\Pi^{(0)}_{i,j,i,j} \equiv 0$ in Eq. (106), the base of induction for both statements of the lemma follow immediately.

Now we prove the inductive hypothesis. Suppose that the lemma statement holds for $h-1$. We prove that conditioned on this, the statements of the lemma hold for h. There are two cases. The first case corresponds to $h < L$. In this case, by definition of $\Pi^{(h)}_{i,j,i,j}(x,x)$ in Eq. (106) and using Lemma 11 and Lemma 12 we can write,

$$\begin{aligned}
\left|\Pi^{(h)}_{i,j,i',j'}(y,z)\right| &\equiv \left| \sum_{a=-\frac{q-1}{2}}^{\frac{q-1}{2}} \sum_{b=-\frac{q-1}{2}}^{\frac{q-1}{2}} \left[\Pi^{(h-1)}(y,z) \odot \dot{\Gamma}^{(h)}(y,z) + \Gamma^{(h)}(y,z)\right]_{i+a,j+b,i'+a,j'+b} \right| \\
&\le \sum_{a=-\frac{q-1}{2}}^{\frac{q-1}{2}} \sum_{b=-\frac{q-1}{2}}^{\frac{q-1}{2}} \frac{\sqrt{\Pi^{(h-1)}_{i+a,j+b,i+a,j+b}(y,y) \cdot \Pi^{(h-1)}_{i'+a,j'+b,i'+a,j'+b}(z,z)}}{q^2} + \frac{\sqrt{N^{(h)}_{i+a,j+b}(y) \cdot N^{(h)}_{i'+a,j'+b}(z)}}{q^2} \\
&\le \sum_{a=-\frac{q-1}{2}}^{\frac{q-1}{2}} \sum_{b=-\frac{q-1}{2}}^{\frac{q-1}{2}} \sqrt{\frac{\Pi^{(h-1)}_{i+a,j+b,i+a,j+b}(y,y) + N^{(h)}_{i+a,j+b}(y)}{q^2}} \cdot \sqrt{\frac{\Pi^{(h-1)}_{i'+a,j'+b,i'+a,j'+b}(z,z) + N^{(h)}_{i'+a,j'+b}(z)}{q^2}} \\
&\le \sqrt{\Pi^{(h)}_{i,j,i,j}(y,y)} \cdot \sqrt{\Pi^{(h)}_{i',j',i',j'}(z,z)},
\end{aligned}$$

where the second line above follows from inductive hypothesis along with Lemma 11 and Lemma 12. The third and fourth lines above follow by Cauchy–Schwarz inequality. The second case corresponds to $h = L$. In this case, by definition of $\Pi^{(L)}_{i,j,i',j'}(y,z)$ in Eq. (107) and using Lemma 11 and Lemma 12 along with the inductive hypothesis we can write,

$$\begin{aligned}
\left|\Pi^{(L)}_{i,j,i',j'}(y,z)\right| &\equiv \left|\Pi^{(L-1)}_{i,j,i',j'}(y,z) \cdot \dot{\Gamma}^{(L)}_{i,j,i',j'}(y,z)\right| \\
&\le \frac{\sqrt{\Pi^{(L-1)}_{i,j,i,j}(y,y) \cdot \Pi^{(L-1)}_{i',j',i',j'}(z,z)}}{q^2} \\
&= \sqrt{\Pi^{(L)}_{i,j,i,j}(y,y)} \cdot \sqrt{\Pi^{(L)}_{i',j',i',j'}(z,z)},
\end{aligned}$$

where the second line above follows from inductive hypothesis along with Lemma 12. This completes the inductive step and in turn proves the first statement of the lemma.

To prove the inductive step for the second statement of lemma we consider two cases again. The first case is $h < L$. In this case, note that by using inductive hypothesis together with Lemma 11 and Lemma 12 we can write,

$$\begin{aligned}
\Pi^{(h)}_{i,j,i,j}(y,y) &\equiv \sum_{a=-\frac{q-1}{2}}^{\frac{q-1}{2}} \sum_{b=-\frac{q-1}{2}}^{\frac{q-1}{2}} \left[\Pi^{(h-1)}(y,y) \odot \dot{\Gamma}^{(h)}(y,y) + \Gamma^{(h)}(y,y)\right]_{i+a,j+b,i+a,j+b} \\
&= \sum_{a=-\frac{q-1}{2}}^{\frac{q-1}{2}} \sum_{b=-\frac{q-1}{2}}^{\frac{q-1}{2}} \frac{(h-1) \cdot N^{(h)}_{i+a,j+b}(y)}{q^2} + \frac{N^{(h)}_{i+a,j+b}(y)}{q^2} \\
&= h \cdot \sum_{a=-\frac{q-1}{2}}^{\frac{q-1}{2}} \sum_{b=-\frac{q-1}{2}}^{\frac{q-1}{2}} \frac{N^{(h)}_{i+a,j+b}(y)}{q^2} \\
&= h \cdot N^{(h+1)}_{i,j}(y),
\end{aligned}$$

where the last line above follows from definition of $N^{(h)}$ in Eq. (103). The second case corresponds to $h = L$. In this case, by inductive hypothesis together with Lemma 11 and Lemma 12 we can write,

$$\begin{aligned}\Pi^{(L)}_{i,j,i,j}(y,y) &\equiv \Pi^{(L-1)}_{i,j,i,j}(y,y) \cdot \dot{\Gamma}^{(L)}_{i,j,i,j}(y,y) \\ &= \frac{(L-1)\cdot N^{(L)}_{i,j}(y)}{q^2}.\end{aligned}$$

This completes the inductive step for the second statement and in turn proves the second statement of the lemma. This completes the proof of Lemma 13. □

G CNTK Sketch: Algorithm, Claims and Invariants

In this section we give our sketching algorithm for the CNTK kernel and prove our main theorem for this algorithm, i.e., Theorem 4. We start by introducing our CNTKSKETCH algorithm in the following definition:

Definition 3 (CNTKSKETCH Algorithm)**.** For every image $x \in \mathbb{R}^{d_1 \times d_2 \times c}$, we compute the CNTKSKETCH , $\Psi^{(L)}_{\mathtt{cntk}}(x)$, recursively as follows,

• Let $s = \widetilde{\mathcal{O}}\left(\frac{L^2}{\varepsilon^2}\right)$, $r = \widetilde{\mathcal{O}}\left(\frac{L^6}{\varepsilon^4}\right)$, $n_1 = \widetilde{\mathcal{O}}\left(\frac{L^4}{\varepsilon^4}\right)$, $m = \widetilde{\mathcal{O}}\left(\frac{L^8}{\varepsilon^{16/3}}\right)$, and $s^* = \mathcal{O}(\frac{1}{\varepsilon^2}\log\frac{1}{\delta})$ and $P^{(p)}_{\mathtt{relu}}(\alpha) = \sum_{l=0}^{2p+2} c_l \cdot \alpha^l$ and $\dot{P}^{(p')}_{\mathtt{relu}}(\alpha) = \sum_{l=0}^{2p'+1} b_l \cdot \alpha^l$ be the polynomials defined in Eq. (6).

1. For every $i \in [d_1]$, $j \in [d_2]$, and $h = 0, 1, 2, \ldots L$ compute $N^{(h)}_{i,j}(x)$ as per Eq. (103) of Definition 2.

2. Let $\boldsymbol{S} \in \mathbb{R}^{r\times c}$ be an SRHT. For every $i \in [d_1]$ and $j \in [d_2]$, compute $\phi^{(0)}_{i,j}(x) \in \mathbb{R}^r$ as,

$$\phi^{(0)}_{i,j}(x) \leftarrow \boldsymbol{S} \cdot x_{(i,j,:)}. \tag{109}$$

3. Let $\boldsymbol{Q}^{2p+2} \in \mathbb{R}^{m\times\left(q^2r\right)^{2p+2}}$ be a degree-$2p+2$ POLYSKETCH , and $\boldsymbol{T} \in \mathbb{R}^{r\times(2p+3)\cdot m}$ be an SRHT. For every $h \in [L]$, every $i \in [d_1]$ and $j \in [d_2]$, and $l = 0, 1, 2, \ldots 2p+2$ compute:

$$\begin{aligned}\mu^{(h)}_{i,j}(x) &\leftarrow \bigoplus_{a=-\frac{q-1}{2}}^{\frac{q-1}{2}} \bigoplus_{b=-\frac{q-1}{2}}^{\frac{q-1}{2}} \frac{\phi^{(h-1)}_{i+a,j+b}(x)}{\sqrt{N^{(h)}_{i,j}(x)}}, \\ \left[Z^{(h)}_{i,j}(x)\right]_l &\leftarrow \boldsymbol{Q}^{2p+2}\cdot\left(\left[\mu^{(h)}_{i,j}(x)\right]^{\otimes l} \otimes e_1^{\otimes 2p+2-l}\right), \\ \phi^{(h)}_{i,j}(x) &\leftarrow \frac{\sqrt{N^{(h)}_{i,j}(x)}}{q}\cdot\boldsymbol{T}\cdot\left(\bigoplus_{l=0}^{2p+2}\sqrt{c_l}\left[Z^{(h)}_{i,j}(x)\right]_l\right).\end{aligned} \tag{110}$$

4. Let $\boldsymbol{Q}^{2p'+1} \in \mathbb{R}^{n_1\times\left(q^2r\right)^{2p'+1}}$ be a degree-$2p'+1$ POLYSKETCH , and $\boldsymbol{W} \in \mathbb{R}^{s\times(2p'+2)\cdot n_1}$ be an SRHT. For every $h \in [L], i \in [d_1], j \in [d_2]$, and $l = 0, 1, \ldots 2p'+1$ compute:

$$\begin{aligned}\left[Y^{(h)}_{i,j}(x)\right]_l &\leftarrow \boldsymbol{Q}^{2p'+1}\left(\left[\mu^{(h)}_{i,j}(x)\right]^{\otimes l} \otimes e_1^{\otimes 2p'+1-l}\right), \\ \dot{\phi}^{(h)}_{i,j}(x) &\leftarrow \frac{1}{q}\cdot\boldsymbol{W}\left(\bigoplus_{l=0}^{2p'+1}\sqrt{b_l}\left[Y^{(h)}_{i,j}(x)\right]_l\right).\end{aligned} \tag{111}$$

5. Let $\boldsymbol{Q}^2 \in \mathbb{R}^{s\times s^2}$ be a degree-2 POLYSKETCH , and $\boldsymbol{R} \in \mathbb{R}^{s\times q^2(s+r)}$ be an SRHT. Let $\psi^{(0)}_{i,j}(x) \leftarrow 0$ and for every $h \in [L-1]$, and $i \in [d_1], j \in [d_2]$, compute $\psi^{(h)}_{i,j}(x) \in \mathbb{R}^s$ as:

$$\begin{aligned}\eta^{(h)}_{i,j}(x) &\leftarrow \boldsymbol{Q}^2\left(\psi^{(h-1)}_{i,j}(x) \otimes \dot{\phi}^{(h)}_{i,j}(x)\right) \oplus \phi^{(h)}_{i,j}(x), \\ \psi^{(h)}_{i,j}(x) &\leftarrow \boldsymbol{R}\left(\bigoplus_{a=-\frac{q-1}{2}}^{\frac{q-1}{2}} \bigoplus_{b=-\frac{q-1}{2}}^{\frac{q-1}{2}} \eta^{(h)}_{i+a,j+b}(x)\right).\end{aligned} \tag{112}$$

$$\text{(For } h = L\text{:)} \qquad \psi^{(L)}_{i,j}(x) \leftarrow \boldsymbol{Q}^2 \cdot \left(\psi^{(L-1)}_{i,j}(x) \otimes \dot{\phi}^{(L)}_{i,j}(x)\right). \tag{113}$$

6. Let $\boldsymbol{G} \in \mathbb{R}^{s^* \times s}$ be a random matrix of i.i.d. normal entries with distribution $\mathcal{N}(0, 1/s^*)$. The CNTKSKETCH is the following:

$$\Psi^{(L)}_{\texttt{cntk}}(y,z) := \frac{1}{d_1 d_2} \cdot \boldsymbol{G} \cdot \left(\sum_{i\in[d_1]} \sum_{j\in[d_2]} \psi^{(L)}_{i,j}(x)\right). \tag{114}$$

In the following lemma, we analyze the correctness of the CNTKSKETCH algorithm by giving the invariants that the algorithm maintains at all times,

Lemma 14 (Invariants of the CNTKSKETCH). *For every positive integers d_1, d_2, c, and L, every $\varepsilon, \delta > 0$, every images $y, z \in \mathbb{R}^{d_1\times d_2\times c}$, if we let $N^{(h)} : \mathbb{R}^{d_1\times d_2\times c} \to \mathbb{R}^{d_1\times d_2}$, $\Gamma^{(h)}(y,z) \in \mathbb{R}^{d_1\times d_2\times d_1\times d_2}$ and $\Pi^{(h)}(y,z) \in \mathbb{R}^{d_1\times d_2\times d_1\times d_2}$ be the tensor functions defined in Eq. (103), Eq. (104), Eq. (106), and Eq. (107) of Definition 2, respectively, then with probability at least $1-\delta$ the following invariants are maintained simultaneously for all $i, i' \in [d_1]$ and $j, j' \in [d_2]$ and every $h = 0, 1, 2, \ldots L$:*

1. *The mapping $\phi^{(h)}_{i,j}(\cdot)$ computed by the CNTK Sketch algorithm in Eq. (109) and Eq. (110) of Definition 3 satisfy the following,*

$$\left|\left\langle \phi^{(h)}_{i,j}(y), \phi^{(h)}_{i',j'}(z)\right\rangle - \Gamma^{(h)}_{i,j,i',j'}(y,z)\right| \le (h+1)\cdot\frac{\varepsilon^2}{60L^3}\cdot\frac{\sqrt{N^{(h)}_{i,j}(y)\cdot N^{(h)}_{i',j'}(z)}}{q^2}.$$

2. *The mapping $\psi^{(h)}_{i,j}(\cdot)$ computed by the CNTK Sketch algorithm in Eq. (112) and Eq. (113) of Definition 3 satisfy the following,*

$$\left|\left\langle \psi^{(h)}_{i,j}(y), \psi^{(h)}_{i',j'}(z)\right\rangle - \Pi^{(h)}_{i,j,i',j'}(y,z)\right| \le \begin{cases} \frac{\varepsilon}{10}\cdot\frac{h^2}{L+1}\cdot\sqrt{N^{(h+1)}_{i,j}(y)\cdot N^{(h+1)}_{i',j'}(z)} & \text{if } h < L \\ \frac{\varepsilon}{10}\cdot\frac{L-1}{q^2}\cdot\sqrt{N^{(L)}_{i,j}(y)\cdot N^{(L)}_{i',j'}(z)} & \text{if } h = L \end{cases}.$$

Proof of Lemma 14: The proof is by induction on the value of $h = 0, 1, 2, \ldots L$. More formally, consider the following statements for every $h = 0, 1, 2, \ldots L$:

$\mathbf{P_1(h)}$: Simultaneously for all $i, i' \in [d_1]$ and $j, j' \in [d_2]$:

$$\left|\left\langle \phi^{(h)}_{i,j}(y), \phi^{(h)}_{i',j'}(z)\right\rangle - \Gamma^{(h)}_{i,j,i',j'}(y,z)\right| \le (h+1)\cdot\frac{\varepsilon^2}{60L^3}\cdot\frac{\sqrt{N^{(h)}_{i,j}(y)\cdot N^{(h)}_{i',j'}(z)}}{q^2},$$

$$\left|\left\|\phi^{(h)}_{i,j}(y)\right\|_2^2 - \Gamma^{(h)}_{i,j,i,j}(y,y)\right| \le \frac{(h+1)\cdot\varepsilon^2}{60L^3}\cdot\frac{N^{(h)}_{i,j}(y)}{q^2},$$

$$\left|\left\|\phi^{(h)}_{i',j'}(z)\right\|_2^2 - \Gamma^{(h)}_{i',j',i',j'}(z,z)\right| \le \frac{(h+1)\cdot\varepsilon^2}{60L^3}\cdot\frac{N^{(h)}_{i',j'}(z)}{q^2}.$$

$\mathbf{P_2(h)}$: Simultaneously for all $i, i' \in [d_1]$ and $j, j' \in [d_2]$:

$$\left|\left\langle \psi^{(h)}_{i,j}(y), \psi^{(h)}_{i',j'}(z)\right\rangle - \Pi^{(h)}_{i,j,i',j'}(y,z)\right| \le \begin{cases} \frac{\varepsilon}{10}\cdot\frac{h^2}{L+1}\cdot\sqrt{N^{(h+1)}_{i,j}(y)\cdot N^{(h+1)}_{i',j'}(z)} & \text{if } h < L \\ \frac{\varepsilon}{10}\cdot\frac{L-1}{q^2}\cdot\sqrt{N^{(L)}_{i,j}(y)\cdot N^{(L)}_{i',j'}(z)} & \text{if } h = L \end{cases},$$

$$\text{(only for } h < L\text{)}: \quad \left|\left\|\psi^{(h)}_{i,j}(y)\right\|_2^2 - \Pi^{(h)}_{i,j,i,j}(y,y)\right| \le \frac{\varepsilon}{10}\cdot\frac{h^2}{L+1}\cdot N^{(h+1)}_{i,j}(y),$$

$$\text{(only for } h < L\text{)}: \quad \left|\left\|\psi^{(h)}_{i',j'}(z)\right\|_2^2 - \Pi^{(h)}_{i',j',i',j'}(z,z)\right| \le \frac{\varepsilon}{10}\cdot\frac{h^2}{L+1}\cdot N^{(h+1)}_{i',j'}(z).$$

We prove that probabilities $\Pr[P_1(0)]$ and $\Pr[P_2(0)|P_1(0)]$ are both greater than $1-\mathcal{O}(\delta/L)$. Additionally, for every $h=1,2,\dots L$, we prove that the conditional probabilities $\Pr[P_1(h)|P_1(h-1)]$ and $\Pr[P_2(h)|P_2(h-1),P_1(h),P_1(h-1)]$ are greater than $1-\mathcal{O}(\delta/L)$.

The **base of induction** corresponds to $h=0$. By Eq. (109), $\phi^{(0)}_{i,j}(y)=\boldsymbol{S}\cdot y_{(i,j,:)}$ and $\phi^{(0)}_{i',j'}(z)=\boldsymbol{S}\cdot z_{(i',j',:)}$, thus, Lemma 2 implies the following

$$\Pr\left[\left|\left\langle\phi^{(0)}_{i,j}(y),\phi^{(0)}_{i',j'}(z)\right\rangle-\left\langle y_{(i,j,:)},z_{(i',j',:)}\right\rangle\right|\le\mathcal{O}\left(\varepsilon^2/L^3\right)\cdot\|y_{(i,j,:)}\|_2\|z_{(i',j',:)}\|_2\right]\ge 1-\mathcal{O}\left(\frac{\delta}{d_1^2d_2^2L}\right),$$

therefore, by using Eq. (103) and Eq. (104) we have

$$\Pr\left[\left|\left\langle\phi^{(0)}_{i,j}(y),\phi^{(0)}_{i',j'}(z)\right\rangle-\Gamma^{(0)}_{i,j,i',j'}(y,z)\right|\le\mathcal{O}\left(\varepsilon^2/L^3\right)\cdot\frac{\sqrt{N^{(0)}_{i,j}(y)\cdot N^{(0)}_{i',j'}(z)}}{q^2}\right]\ge 1-\mathcal{O}\left(\frac{\delta}{d_1^2d_2^2L}\right).$$

Similarly, we can prove that with probability at least $1-\mathcal{O}\left(\frac{\delta}{d_1^2d_2^2L}\right)$, the following hold

$$\left|\left\|\phi^{(0)}_{i,j}(y)\right\|_2^2-\Gamma^{(0)}_{i,j,i,j}(y,y)\right|\le\mathcal{O}\left(\varepsilon^2/L^3\right)\cdot\frac{N^{(0)}_{i,j}(y)}{q^2},$$
$$\left|\left\|\phi^{(0)}_{i',j'}(z)\right\|_2^2-\Gamma^{(0)}_{i',j',i',j'}(z,z)\right|\le\mathcal{O}\left(\varepsilon^2/L^3\right)\cdot\frac{N^{(0)}_{i',j'}(z)}{q^2}.$$

By union bounding over all $i,i'\in[d_1]$ and $j,j'\in[d_2]$, this proves the base of induction for statement $P_1(h)$, i.e., $\Pr[P_1(0)]\ge 1-\mathcal{O}(\delta/L)$.

Moreover, by Eq. (112), we have that $\psi^{(0)}_{i,j}(y)=0$ and $\psi^{(0)}_{i',j'}(z)=0$, thus, by Eq. (106), it trivially holds that $\Pr[P_2(0)|P_1(0)]=1\ge 1-\mathcal{O}(\delta/L)$. This completes the base of induction.

Now, we proceed to prove the **inductive step**. That is, by assuming the inductive hypothesis for $h-1$, we prove that statements $P_1(h)$ and $P_2(h)$ hold. More precisely, first we condition on the statement $P_1(h-1)$ being true for some $h\ge 1$, and then prove that $P_1(h)$ holds with probability at least $1-\mathcal{O}(\delta/L)$. Next we show that conditioned on statements $P_2(h-1),P_1(h),P_1(h-1)$ being true, $P_2(h)$ holds with probability at least $1-\mathcal{O}(\delta/L)$. This will complete the induction.

First, note that by Lemma 2, union bound, and using Eq. (110), the following holds simultaneously for all $i,i'\in[d_1]$ and all $j,j'\in[d_2]$, with probability at least $1-\mathcal{O}\left(\frac{\delta}{L}\right)$,

$$\left|\left\langle\phi^{(h)}_{i,j}(y),\phi^{(h)}_{i',j'}(z)\right\rangle-\frac{\sqrt{N^{(h)}_{i,j}(y)N^{(h)}_{i',j'}(z)}}{q^2}\cdot\sum_{l=0}^{2p+2}c_l\left\langle\left[Z^{(h)}_{i,j}(y)\right]_l,\left[Z^{(h)}_{i',j'}(z)\right]_l\right\rangle\right|\le\mathcal{O}\left(\frac{\varepsilon^2}{L^3}\right)\cdot A,\quad(115)$$

where $A:=\frac{\sqrt{N^{(h)}_{i,j}(y)N^{(h)}_{i',j'}(z)}}{q^2}\cdot\sqrt{\sum_{l=0}^{2p+2}c_l\left\|\left[Z^{(h)}_{i,j}(y)\right]_l\right\|_2^2}\cdot\sqrt{\sum_{l=0}^{2p+2}c_l\left\|\left[Z^{(h)}_{i',j'}(z)\right]_l\right\|_2^2}$ and the collection of vectors $\left\{\left[Z^{(h)}_{i,j}(y)\right]_l\right\}_{l=0}^{2p+2}$ and $\left\{\left[Z^{(h)}_{i',j'}(z)\right]_l\right\}_{l=0}^{2p+2}$ and coefficients $c_0,c_1,c_2,\dots c_{2p+2}$ are defined as per Eq. (110) and Eq. (6), respectively. Additionally, by Lemma 1 and union bound, the following inequalities hold, with probability at least $1-\mathcal{O}\left(\frac{\delta}{L}\right)$, simultaneously for all $l=0,1,2,\dots 2p+2$, all $i,i'\in[d_1]$ and all $j,j'\in[d_2]$:

$$\begin{aligned}
&\left|\left\langle\left[Z^{(h)}_{i,j}(y)\right]_l,\left[Z^{(h)}_{i',j'}(z)\right]_l\right\rangle-\left\langle\mu^{(h)}_{i,j}(y),\mu^{(h)}_{i',j'}(z)\right\rangle^l\right|\le O\left(\frac{\varepsilon^2}{L^3}\right)\left\|\mu^{(h)}_{i,j}(y)\right\|_2^l\left\|\mu^{(h)}_{i',j'}(z)\right\|_2^l\\
&\left\|\left[Z^{(h)}_{i,j}(y)\right]_l\right\|_2^2\le\frac{11}{10}\cdot\left\|\mu^{(h)}_{i,j}(y)\right\|_2^{2l}\qquad(116)\\
&\left\|\left[Z^{(h)}_{i',j'}(z)\right]_l\right\|_2^2\le\frac{11}{10}\cdot\left\|\mu^{(h)}_{i',j'}(z)\right\|_2^{2l}
\end{aligned}$$

Therefore, by plugging Eq. (116) back to Eq. (115) and using union bound and triangle inequality as well as Cauchy–Schwarz inequality, we find that with probability at least $1-\mathcal{O}\left(\frac{\delta}{L}\right)$, the following holds simultaneously for all $i,i'\in[d_1]$ and $j,j'\in[d_2]$

$$\left|\left\langle\phi^{(h)}_{i,j}(y),\phi^{(h)}_{i',j'}(z)\right\rangle-\frac{\sqrt{N^{(h)}_{i,j}(y)N^{(h)}_{i',j'}(z)}}{q^2}\cdot P^{(p)}_{\texttt{relu}}\left(\left\langle\mu^{(h)}_{i,j}(y),\mu^{(h)}_{i',j'}(z)\right\rangle\right)\right|\le\mathcal{O}\left(\frac{\varepsilon^2}{L^3}\right)\cdot B,\quad(117)$$

where $B := \frac{\sqrt{N^{(h)}_{i,j}(y)N^{(h)}_{i',j'}(z)}}{q^2} \cdot \sqrt{P^{(p)}_{\texttt{relu}}\left(\|\mu^{(h)}_{i,j}(y)\|_2^2\right) \cdot P^{(p)}_{\texttt{relu}}\left(\|\mu^{(h)}_{i',j'}(z)\|_2^2\right)}$ and $P^{(p)}_{\texttt{relu}}(\alpha) = \sum_{l=0}^{2p+2} c_l \cdot \alpha^l$ is the polynomial defined in Eq. (6). By using the definition of $\mu^{(h)}_{i,j}(\cdot)$ in Eq. (110) we have,

$$\begin{aligned}
\left\langle \mu^{(h)}_{i,j}(y), \mu^{(h)}_{i',j'}(z) \right\rangle &= \frac{\sum_{a=-\frac{q-1}{2}}^{\frac{q-1}{2}} \sum_{b=-\frac{q-1}{2}}^{\frac{q-1}{2}} \left\langle \phi^{(h-1)}_{i+a,j+b}(y), \phi^{(h-1)}_{i'+a,j'+b}(z) \right\rangle}{\sqrt{N^{(h)}_{i,j}(y)N^{(h)}_{i',j'}(z)}}, \\
\left\|\mu^{(h)}_{i,j}(y)\right\|_2^2 &= \frac{\sum_{a=-\frac{q-1}{2}}^{\frac{q-1}{2}} \sum_{b=-\frac{q-1}{2}}^{\frac{q-1}{2}} \left\|\phi^{(h-1)}_{i+a,j+b}(y)\right\|_2^2}{N^{(h)}_{i,j}(y)}, \\
\left\|\mu^{(h)}_{i',j'}(z)\right\|_2^2 &= \frac{\sum_{a=-\frac{q-1}{2}}^{\frac{q-1}{2}} \sum_{b=-\frac{q-1}{2}}^{\frac{q-1}{2}} \left\|\phi^{(h-1)}_{i'+a,j'+b}(z)\right\|_2^2}{N^{(h)}_{i,j}(z)}.
\end{aligned} \tag{118}$$

Hence, by conditioning on the inductive hypothesis $P_1(h-1)$ and using Eq. (118) and Corollary 1 we have,

$$\left|\left\|\mu^{(h)}_{i,j}(y)\right\|_2^2 - 1\right| \le h \cdot \frac{\varepsilon^2}{60L^3}, \text{ and } \left|\left\|\mu^{(h)}_{i',j'}(z)\right\|_2^2 - 1\right| \le h \cdot \frac{\varepsilon^2}{60L^3}.$$

Therefore, by invoking Lemma 4, it follows that $\left|P^{(p)}_{\texttt{relu}}\left(\|\mu^{(h)}_{i,j}(y)\|_2^2\right) - P^{(p)}_{\texttt{relu}}(1)\right| \le h \cdot \frac{\varepsilon^2}{60L^3}$ and $\left|P^{(p)}_{\texttt{relu}}\left(\|\mu^{(h)}_{i',j'}(z)\|_2^2\right) - P^{(p)}_{\texttt{relu}}(1)\right| \le h \cdot \frac{\varepsilon^2}{60L^3}$. Consequently, because $P^{(p)}_{\texttt{relu}}(1) \le P^{(+\infty)}_{\texttt{relu}}(1) = 1$, we find that

$$B \le \frac{11}{10} \cdot \frac{\sqrt{N^{(h)}_{i,j}(y)N^{(h)}_{i',j'}(z)}}{q^2}.$$

For shorthand we use the notation $\beta := \frac{\sqrt{N^{(h)}_{i,j}(y)N^{(h)}_{i',j'}(z)}}{q^2}$. By plugging this into Eq. (117) and using the notation β, we find that the following holds simultaneously for all $i, i' \in [d_1]$ and all $j, j' \in [d_2]$, with probability at least $1 - \mathcal{O}\left(\frac{\delta}{L}\right)$,

$$\left|\left\langle \phi^{(h)}_{i,j}(y), \phi^{(h)}_{i',j'}(z) \right\rangle - \beta \cdot P^{(p)}_{\texttt{relu}}\left(\left\langle \mu^{(h)}_{i,j}(y), \mu^{(h)}_{i',j'}(z) \right\rangle\right)\right| \le \mathcal{O}\left(\frac{\varepsilon^2}{L^3}\right) \cdot \beta. \tag{119}$$

Furthermore, by conditioning on the inductive hypothesis $P_1(h-1)$ and combining it with Eq. (118) and applying Cauchy–Schwarz inequality and invoking Corollary 1 we find that,

$$\begin{aligned}
&\left|\left\langle \mu^{(h)}_{i,j}(y), \mu^{(h)}_{i',j'}(z) \right\rangle - \frac{\sum_{a=-\frac{q-1}{2}}^{\frac{q-1}{2}} \sum_{b=-\frac{q-1}{2}}^{\frac{q-1}{2}} \Gamma^{(h-1)}_{i+a,j+b,i'+a,j'+b}(y,z)}{\sqrt{N^{(h)}_{i,j}(y) \cdot N^{(h)}_{i',j'}(z)}}\right| \\
&\le \frac{\sum_{a=-\frac{q-1}{2}}^{\frac{q-1}{2}} \sum_{b=-\frac{q-1}{2}}^{\frac{q-1}{2}} \sqrt{N^{(h-1)}_{i+a,j+b}(y) \cdot N^{(h-1)}_{i'+a,j'+b}(z)}}{q^2 \cdot \sqrt{N^{(h)}_{i,j}(y) \cdot N^{(h)}_{i',j'}(z)}} \cdot h \cdot \frac{\varepsilon^2}{60L^3} \\
&\le \frac{\sqrt{\sum_{a=-\frac{q-1}{2}}^{\frac{q-1}{2}} \sum_{b=-\frac{q-1}{2}}^{\frac{q-1}{2}} \frac{N^{(h-1)}_{i+a,j+b}(y)}{q^2}} \cdot \sqrt{\sum_{a=-\frac{q-1}{2}}^{\frac{q-1}{2}} \sum_{b=-\frac{q-1}{2}}^{\frac{q-1}{2}} \frac{N^{(h-1)}_{i'+a,j'+b}(z)}{q^2}}}{\sqrt{N^{(h)}_{i,j}(y) \cdot N^{(h)}_{i',j'}(z)}} \cdot \frac{h \cdot \varepsilon^2}{60L^3} \\
&= h \cdot \frac{\varepsilon^2}{60L^3},
\end{aligned} \tag{120}$$

where the last line follows from Eq. (103).

For shorthand, we use the notation $\gamma := \frac{\sum_{a=-\frac{q-1}{2}}^{\frac{q-1}{2}} \sum_{b=-\frac{q-1}{2}}^{\frac{q-1}{2}} \Gamma^{(h-1)}_{i+a,j+b,i'+a,j'+b}(y,z)}{\sqrt{N^{(h)}_{i,j}(y) \cdot N^{(h)}_{i',j'}(z)}}$. Note that by Lemma 11 and Eq. (103), $-1 \le \gamma \le 1$. Hence, we can invoke Lemma 4 and use Eq. (120) to find

that,

$$\left|P^{(p)}_{\texttt{relu}}\left(\left\langle\mu^{(h)}_{i,j}(y),\mu^{(h)}_{i',j'}(z)\right\rangle\right)-P^{(p)}_{\texttt{relu}}(\gamma)\right|\leq h\cdot\frac{\varepsilon^2}{60L^3}.$$

By incorporating the above inequality into Eq. (119) using triangle inequality we find that, with probability at least $1-\mathcal{O}\left(\frac{\delta}{L}\right)$, the following holds simultaneously for all $i,i'\in[d_1]$ and all $j,j'\in[d_2]$:

$$\left|\left\langle\phi^{(h)}_{i,j}(y),\phi^{(h)}_{i',j'}(z)\right\rangle-\beta\cdot P^{(p)}_{\texttt{relu}}(\gamma)\right|\leq\left(\mathcal{O}\left(\frac{\varepsilon^2}{L^3}\right)+\frac{h\cdot\varepsilon^2}{60L^3}\right)\cdot\beta. \tag{121}$$

Additionally, since $-1\leq\gamma\leq 1$, we can invoke Lemma 3 and use the fact that $p=\lceil 2L^2/\varepsilon^{4/3}\rceil$ to conclude,

$$\left|P^{(p)}_{\texttt{relu}}(\gamma)-\kappa_1(\gamma)\right|\leq\frac{\varepsilon^2}{76L^3}.$$

By combining the above inequality with Eq. (121) via triangle inequality and using the fact that, by Eq. (104), $\beta\cdot\kappa_1(\gamma)\equiv\Gamma^{(h)}_{i,j,i',j'}(y,z)$ we get the following inequality, with probability at least $1-\mathcal{O}\left(\frac{\delta}{L}\right)$,

$$\left|\left\langle\phi^{(h)}_{i,j}(y),\phi^{(h)}_{i',j'}(z)\right\rangle-\Gamma^{(h)}_{i,j,i',j'}(y,z)\right|\leq(h+1)\cdot\frac{\varepsilon^2}{60L^3}\cdot\frac{\sqrt{N^{(h)}_{i,j}(y)N^{(h)}_{i',j'}(z)}}{q^2}.$$

Similarly, we can prove that with probability at least $1-\mathcal{O}\left(\frac{\delta}{L}\right)$ the following hold, simultaneously for all $i,i'\in[d_1]$ and $j,j'\in[d_2]$,

$$\left|\left\|\phi^{(h)}_{i,j}(y)\right\|_2^2-\Gamma^{(h)}_{i,j,i,j}(y,y)\right|\leq\frac{(h+1)\varepsilon^2}{60L^3}\cdot\frac{N^{(h)}_{i,j}(y)}{q^2},$$

$$\left|\left\|\phi^{(h)}_{i',j'}(z)\right\|_2^2-\Gamma^{(h)}_{i',j',i',j'}(z,z)\right|\leq\frac{(h+1)\varepsilon^2}{60L^3}\cdot\frac{N^{(h)}_{i',j'}(z)}{q^2}.$$

This is sufficient to prove the inductive step for statement $P_1(h)$, i.e., $\Pr[P_1(h)|P_1(h-1)]\geq 1-\mathcal{O}(\delta/L)$.

Now we prove the inductive step for statement $P_2(h)$. That is, we prove that conditioned on $P_2(h-1)$, $P_1(h)$, and $P_1(h-1)$, $P_2(h)$ holds with probability at least $1-\mathcal{O}(\delta/L)$. First, note that by Lemma 2 and using Eq. (111) and union bound, we have the following simultaneously for all $i,i'\in[d_1]$ and all $j,j'\in[d_2]$, with probability at least $1-O\left(\frac{\delta}{L}\right)$,

$$\left|\left\langle\dot\phi^{(h)}_{i,j}(y),\dot\phi^{(h)}_{i',j'}(z)\right\rangle-\frac{1}{q^2}\sum_{l=0}^{2p'+1}b_l\left\langle\left[Y^{(h)}_{i,j}(y)\right]_l,\left[Y^{(h)}_{i',j'}(z)\right]_l\right\rangle\right|\leq\mathcal{O}\left(\frac{\varepsilon}{L}\right)\widehat{A}, \tag{122}$$

where $\widehat{A}:=\frac{1}{q^2}\cdot\sqrt{\sum_{l=0}^{2p'+1}b_l\left\|\left[Y^{(h)}_{i,j}(y)\right]_l\right\|_2^2}\cdot\sqrt{\sum_{l=0}^{2p'+1}b_l\left\|\left[Y^{(h)}_{i',j'}(z)\right]_l\right\|_2^2}$ and the collection of vectors $\left\{\left[Y^{(h)}_{i,j}(y)\right]_l\right\}_{l=0}^{2p'+1}$ and $\left\{\left[Y^{(h)}_{i',j'}(z)\right]_l\right\}_{l=0}^{2p'+1}$ and coefficients $b_0,b_1,b_2,\ldots b_{2p'+1}$ are defined as per Eq. (111) and Eq. (6), respectively. By Lemma 1 and union bound, with probability at least $1-\mathcal{O}\left(\frac{\delta}{L}\right)$, the following inequalities hold true simultaneously for all $l\in\{0,1,2,\ldots 2p'+1\}$, all $i,i'\in[d_1]$ and all $j,j'\in[d_2]$,

$$\begin{aligned}
&\left|\left\langle\left[Y^{(h)}_{i,j}(y)\right]_l,\left[Y^{(h)}_{i',j'}(z)\right]_l\right\rangle-\left\langle\mu^{(h)}_{i,j}(y),\mu^{(h)}_{i',j'}(z)\right\rangle^l\right|\leq\mathcal{O}\left(\frac{\varepsilon}{L}\right)\cdot\left\|\mu^{(h)}_{i,j}(y)\right\|_2^l\left\|\mu^{(h)}_{i',j'}(z)\right\|_2^l\\
&\left\|\left[Y^{(h)}_{i,j}(y)\right]_l\right\|_2^2\leq\frac{11}{10}\cdot\left\|\mu^{(h)}_{i,j}(y)\right\|_2^{2l}\\
&\left\|\left[Y^{(h)}_{i',j'}(z)\right]_l\right\|_2^2\leq\frac{11}{10}\cdot\left\|\mu^{(h)}_{i',j'}(z)\right\|_2^{2l}
\end{aligned} \tag{123}$$

Therefore, by plugging Eq. (123) into Eq. (122) and using union bound and triangle inequality as well as Cauchy–Schwarz inequality, we find that with probability at least $1-\mathcal{O}\left(\frac{\delta}{L}\right)$, the following holds simultaneously for all $i,i'\in[d_1]$ and $j,j'\in[d_2]$

$$\left|\left\langle\dot\phi^{(h)}_{i,j}(y),\dot\phi^{(h)}_{i',j'}(z)\right\rangle-\frac{1}{q^2}\cdot\dot P^{(p')}_{\texttt{relu}}\left(\left\langle\mu^{(h)}_{i,j}(y),\mu^{(h)}_{i',j'}(z)\right\rangle\right)\right|\leq\mathcal{O}\left(\frac{\varepsilon}{L}\right)\cdot\widehat{B}, \tag{124}$$

where $\widehat{B} := \frac{1}{q^2} \cdot \sqrt{\dot{P}^{(p')}_{\texttt{relu}}\left(\|\mu^{(h)}_{i,j}(y)\|_2^2\right) \cdot \dot{P}^{(p')}_{\texttt{relu}}\left(\|\mu^{(h)}_{i',j'}(z)\|_2^2\right)}$ and $\dot{P}^{(p)}_{\texttt{relu}}(\alpha) = \sum_{l=0}^{2p'+1} b_l \cdot \alpha^l$ is the polynomial defined in Eq. (6). By conditioning on the inductive hypothesis $P_1(h-1)$ and using Eq. (118) and Corollary 1 we have $\left|\left\|\mu^{(h)}_{i,j}(y)\right\|_2^2 - 1\right| \le h \cdot \frac{\varepsilon^2}{60L^3}$ and $\left|\left\|\mu^{(h)}_{i',j'}(z)\right\|_2^2 - 1\right| \le h \cdot \frac{\varepsilon^2}{60L^3}$. Therefore, using the fact that $p' = \lceil 9L^2/\varepsilon^2 \rceil$ and by invoking Lemma 4, it follows that $\left|\dot{P}^{(p')}_{\texttt{relu}}\left(\|\mu^{(h)}_{i,j}(y)\|_2^2\right) - \dot{P}^{(p')}_{\texttt{relu}}(1)\right| \le \frac{h\cdot\varepsilon}{20L^2}$ and $\left|\dot{P}^{(p')}_{\texttt{relu}}\left(\|\mu^{(h)}_{i',j'}(z)\|_2^2\right) - \dot{P}^{(p')}_{\texttt{relu}}(1)\right| \le \frac{h\cdot\varepsilon}{20L^2}$. Consequently, because $\dot{P}^{(p')}_{\texttt{relu}}(1) \le \dot{P}^{(+\infty)}_{\texttt{relu}}(1) = 1$, we find that

$$\widehat{B} \le \frac{11}{10 \cdot q^2}.$$

By plugging this into Eq. (124) we get the following, with probability at least $1 - \mathcal{O}\left(\frac{\delta}{L}\right)$,

$$\left|\left\langle \dot{\phi}^{(h)}_{i,j}(y), \dot{\phi}^{(h)}_{i',j'}(z)\right\rangle - \frac{1}{q^2} \cdot \dot{P}^{(p')}_{\texttt{relu}}\left(\left\langle \mu^{(h)}_{i,j}(y), \mu^{(h)}_{i',j'}(z)\right\rangle\right)\right| \le \mathcal{O}\left(\frac{\varepsilon}{q^2 \cdot L}\right). \tag{125}$$

Furthermore, recall the notation $\gamma = \frac{\sum_{a=-\frac{q-1}{2}}^{\frac{q-1}{2}} \sum_{b=-\frac{q-1}{2}}^{\frac{q-1}{2}} \Gamma^{(h-1)}_{i+a,j+b,i'+a,j'+b}(y,z)}{\sqrt{N^{(h)}_{i,j}(y) \cdot N^{(h)}_{i',j'}(z)}}$ and note that by Lemma 11 and Eq. (103), $-1 \le \gamma \le 1$. Hence, we can invoke Lemma 4 and use the fact that $p' = \lceil 9L^2/\varepsilon^2 \rceil$ to find that Eq. (120) implies the following,

$$\left|\dot{P}^{(p')}_{\texttt{relu}}\left(\left\langle \mu^{(h)}_{i,j}(y), \mu^{(h)}_{i',j'}(z)\right\rangle\right) - \dot{P}^{(p')}_{\texttt{relu}}(\gamma)\right| \le \frac{h \cdot \varepsilon}{20L^2}.$$

By incorporating the above inequality into Eq. (125) using triangle inequality, we find that, with probability at least $1 - \mathcal{O}\left(\frac{\delta}{L}\right)$, the following holds simultaneously for all $i, i' \in [d_1]$ and all $j, j' \in [d_2]$:

$$\left|\left\langle \dot{\phi}^{(h)}_{i,j}(y), \dot{\phi}^{(h)}_{i',j'}(z)\right\rangle - \frac{1}{q^2} \cdot \dot{P}^{(p')}_{\texttt{relu}}(\gamma)\right| \le \mathcal{O}\left(\frac{\varepsilon}{q^2 L^2}\right) + \frac{h}{q^2} \cdot \frac{\varepsilon}{20L^2}. \tag{126}$$

Since $-1 \le \gamma \le 1$, we can invoke Lemma 3 and use the fact that $p' = \lceil 9L^2/\varepsilon^2 \rceil$ to conclude,

$$\left|\dot{P}^{(p')}_{\texttt{relu}}(\gamma) - \kappa_0(\gamma)\right| \le \frac{\varepsilon}{15L}.$$

By combining above inequality with Eq. (126) via triangle inequality and using the fact that, by Eq. (105), $\frac{1}{q^2} \cdot \kappa_0(\gamma) \equiv \dot{\Gamma}^{(h)}_{i,j,i',j'}(y,z)$ we get the following bound simultaneously for all $i, i' \in [d_1]$ and all $j, j' \in [d_2]$, with probability at least $1 - \mathcal{O}\left(\frac{\delta}{L}\right)$:

$$\left|\left\langle \dot{\phi}^{(h)}_{i,j}(y), \dot{\phi}^{(h)}_{i',j'}(z)\right\rangle - \dot{\Gamma}^{(h)}_{i,j,i',j'}(y,z)\right| \le \frac{1}{q^2} \cdot \frac{\varepsilon}{8L}. \tag{127}$$

Similarly we can prove that with probability at least $1 - \mathcal{O}\left(\frac{\delta}{L}\right)$, the following hold simultaneously for all $i, i' \in [d_1]$ and all $j, j' \in [d_2]$,

$$\left|\left\|\dot{\phi}^{(h)}_{i,j}(y)\right\|_2^2 - \dot{\Gamma}^{(h)}_{i,j,i,j}(y,y)\right| \le \frac{1}{q^2} \cdot \frac{\varepsilon}{8L}, \text{ and } \left|\left\|\dot{\phi}^{(h)}_{i',j'}(z)\right\|_2^2 - \dot{\Gamma}^{(h)}_{i',j',i',j'}(z,z)\right| \le \frac{1}{q^2} \cdot \frac{\varepsilon}{8L}. \tag{128}$$

We will use Eq. (127) and Eq. (128) to prove the inductive step for $P_2(h)$.

Next, we consider two cases for the value of h. When $h < L$, the vectors $\psi^{(h)}_{i,j}(y), \psi^{(h)}_{i',j'}(z)$ are defined in Eq. (112) and when $h = L$, these vectors are defined differently in Eq. (113). First we consider the case of $h < L$. Note that in this case, if we let $\eta^{(h)}_{i,j}(y)$ and $\eta^{(h)}_{i',j'}(z)$ be the vectors defined in Eq. (112), then by Lemma 2 and union bound, the following holds simultaneously for all $i, i' \in [d_1]$ and all $j, j' \in [d_2]$, with probability at least $1 - \mathcal{O}\left(\frac{\delta}{L}\right)$:

$$\left|\left\langle \psi^{(h)}_{i,j}(y), \psi^{(h)}_{i',j'}(z)\right\rangle - \sum_{a=-\frac{q-1}{2}}^{\frac{q-1}{2}} \sum_{b=-\frac{q-1}{2}}^{\frac{q-1}{2}} \left\langle \eta^{(h)}_{i+a,j+b}(y), \eta^{(h)}_{i'+a,j'+b}(z)\right\rangle\right| \le \mathcal{O}(\varepsilon/L) \cdot D, \tag{129}$$

where $D := \sqrt{\sum_{a=-\frac{q-1}{2}}^{\frac{q-1}{2}} \sum_{b=-\frac{q-1}{2}}^{\frac{q-1}{2}} \|\eta^{(h)}_{i+a,j+b}(y)\|_2^2} \cdot \sqrt{\sum_{a=-\frac{q-1}{2}}^{\frac{q-1}{2}} \sum_{b=-\frac{q-1}{2}}^{\frac{q-1}{2}} \|\eta^{(h)}_{i'+a,j'+b}(z)\|_2^2}$. Now, if we let $f_{i,j} := \psi^{(h-1)}_{i,j}(y) \otimes \dot{\phi}^{(h)}_{i,j}(y)$ and $g_{i',j'} := \psi^{(h-1)}_{i',j'}(z) \otimes \dot{\phi}^{(h)}_{i',j'}(z)$, then by Eq. (112), $\eta^{(h)}_{i,j}(y) = \left(\boldsymbol{Q}^2 \cdot f_{i,j}\right) \oplus \phi^{(h)}_{i,j}(y)$ and $\eta^{(h)}_{i',j'}(z) = \left(\boldsymbol{Q}^2 \cdot g_{i',j'}\right) \oplus \phi^{(h)}_{i',j'}(z)$. Thus by Lemma 1 and union bound, with probability at least $1 - \mathcal{O}\left(\frac{\delta}{L}\right)$, we have the following inequalities simultaneously for all $i, i' \in [d_1]$ and $j, j' \in [d_2]$:

$$\begin{aligned}
&\left|\left\langle \eta^{(h)}_{i,j}(y), \eta^{(h)}_{i',j'}(z)\right\rangle - \langle f_{i,j}, g_{i',j'}\rangle - \left\langle \phi^{(h)}_{i,j}(y), \phi^{(h)}_{i',j'}(z)\right\rangle\right| \le \mathcal{O}\left(\frac{\varepsilon}{L}\right) \cdot \|f_{i,j}\|_2 \|g_{i',j'}\|_2 \\
&\left\|\eta^{(h)}_{i,j}(y)\right\|_2^2 \le \frac{11}{10} \cdot \|f_{i,j}\|_2^2 + \left\|\phi^{(h)}_{i,j}(y)\right\|_2^2 \\
&\left\|\eta^{(h)}_{i',j'}(z)\right\|_2^2 \le \frac{11}{10} \cdot \|g_{i',j'}\|_2^2 + \left\|\phi^{(h)}_{i',j'}(z)\right\|_2^2
\end{aligned} \tag{130}$$

Therefore, if we condition on inductive hypotheses $P_1(h)$ and $P_2(h-1)$, then by using Corollary 1, Lemma 13, the inequality Eq. (128) and Lemma 12 along with the fact that $\|f_{i,j}\|_2^2 = \|\psi^{(h-1)}_{i,j}(y)\|_2^2 \cdot \|\dot{\phi}^{(h)}_{i,j}(y)\|_2^2$, we have:

$$\begin{aligned}
&\sum_{a=-\frac{q-1}{2}}^{\frac{q-1}{2}} \sum_{b=-\frac{q-1}{2}}^{\frac{q-1}{2}} \|\eta^{(h)}_{i+a,j+b}(y)\|_2^2 \\
&\quad\le \sum_{a=-\frac{q-1}{2}}^{\frac{q-1}{2}} \sum_{b=-\frac{q-1}{2}}^{\frac{q-1}{2}} \frac{11}{10}\|f_{i+a,j+b}\|_2^2 + \Gamma^{(h)}_{i+a,j+b,i+a,j+b}(y,y) + \frac{N^{(h)}_{i+a,j+b}(y)}{10q^2} \\
&\quad= \frac{11}{10} \cdot \sum_{a=-\frac{q-1}{2}}^{\frac{q-1}{2}} \sum_{b=-\frac{q-1}{2}}^{\frac{q-1}{2}} \|\psi^{(h-1)}_{i+a,j+b}(y)\|_2^2 \cdot \|\dot{\phi}^{(h)}_{i+a,j+b}(y)\|_2^2 + \Gamma^{(h)}_{i+a,j+b,i+a,j+b}(y,y) \\
&\quad\le \frac{12}{10} \sum_{a=-\frac{q-1}{2}}^{\frac{q-1}{2}} \sum_{b=-\frac{q-1}{2}}^{\frac{q-1}{2}} \Pi^{(h-1)}_{i+a,j+b,i+a,j+b}(y,y) \cdot \dot{\Gamma}^{(h)}_{i+a,j+b,i+a,j+b}(y,y) + \Gamma^{(h)}_{i+a,j+b,i+a,j+b}(y,y) \\
&\quad= \frac{12}{10} \cdot \Pi^{(h)}_{i,j,i,j}(y,y) = \frac{12}{10} \cdot h \cdot N^{(h+1)}_{i,j}(y),
\end{aligned}$$

where the fourth line above follows from the inductive hypothesis $P_2(h-1)$ along with Eq. (128) and Lemma 12 and Lemma 13. The last line above follows from Eq. (106) and Lemma 13. Similarly we can prove, $\sum_{a=-\frac{q-1}{2}}^{\frac{q-1}{2}} \sum_{b=-\frac{q-1}{2}}^{\frac{q-1}{2}} \|\eta^{(h)}_{i'+a,j'+b}(z)\|_2^2 \le \frac{12}{10} \cdot h \cdot N^{(h+1)}_{i',j'}(z)$, thus conditioned on $P_2(h-1), P_1(h), P_1(h-1)$, with probability at least $1 - \mathcal{O}\left(\frac{\delta}{L}\right)$:

$$D \le \frac{12}{10} \cdot h \cdot \sqrt{N^{(h+1)}_{i,j}(y) \cdot N^{(h+1)}_{i',j'}(z)}.$$

By incorporating this into Eq. (129) it follows that if we condition on $P_2(h-1), P_1(h), P_1(h-1)$, then, with probability at least $1 - \mathcal{O}\left(\frac{\delta}{L}\right)$, the following holds simultaneously for all $i, i' \in [d_1]$ and all $j, j' \in [d_2]$,

$$\begin{aligned}
&\left|\left\langle \psi^{(h)}_{i,j}(y), \psi^{(h)}_{i',j'}(z)\right\rangle - \sum_{a=-\frac{q-1}{2}}^{\frac{q-1}{2}} \sum_{b=-\frac{q-1}{2}}^{\frac{q-1}{2}} \left\langle \eta^{(h)}_{i+a,j+b}(y), \eta^{(h)}_{i'+a,j'+b}(z)\right\rangle\right| \\
&\qquad\le \mathcal{O}\left(\varepsilon h/L\right) \cdot \sqrt{N^{(h+1)}_{i,j}(y) \cdot N^{(h+1)}_{i',j'}(z)}.
\end{aligned} \tag{131}$$

Now we bound the term $\left|\left\langle \eta^{(h)}_{i,j}(y), \eta^{(h)}_{i',j'}(z)\right\rangle - \langle f_{i,j}, g_{i',j'}\rangle - \left\langle \phi^{(h)}_{i,j}(y), \phi^{(h)}_{i',j'}(z)\right\rangle\right|$ using Eq. (130), Eq. (128), and Lemma 12 along with inductive hypotheses $P_2(h-1)$ and Lemma 13. With probability

at least $1-\mathcal{O}\left(\frac{\delta}{L}\right)$ the following holds simultaneously for all $i,i'\in[d_1]$ and all $j,j'\in[d_2]$:

$$\begin{aligned}
&\left|\left\langle \eta_{i,j}^{(h)}(y),\eta_{i',j'}^{(h)}(z)\right\rangle - \langle f_{i,j},g_{i',j'}\rangle - \left\langle \phi_{i,j}^{(h)}(y),\phi_{i',j'}^{(h)}(z)\right\rangle\right| \\
&\quad\le \mathcal{O}\left(\frac{\varepsilon}{L}\right)\cdot\sqrt{\Pi_{i,j,i,j}^{(h-1)}(y,y)\cdot\dot{\Gamma}_{i,j,i,j}^{(h)}(y,y)\cdot\Pi_{i',j',i',j'}^{(h-1)}(z,z)\cdot\dot{\Gamma}_{i',j',i',j'}^{(h)}(z,z)} \\
&\quad= \mathcal{O}\left(\frac{\varepsilon\cdot h}{L}\right)\cdot\frac{\sqrt{N_{i,j}^{(h)}(y)\cdot N_{i',j'}^{(h)}(z)}}{q^2},
\end{aligned}$$

where the last line above follows from Lemma 13 together with the fact that $\dot{\Gamma}_{i,j,i,j}^{(h)}(y,y) = \dot{\Gamma}_{i',j',i',j'}^{(h)}(z,z) = \frac{1}{q^2}$.

By combining the above with inductive hypotheses $P_1(h), P_2(h-1)$ and Eq. (127) via triangle inequality and invoking Lemma 13 we get that the following holds simultaneously for all $i,i'\in[d_1]$ and all $j,j'\in[d_2]$, with probability at least $1-\mathcal{O}\left(\frac{\delta}{L}\right)$,

$$\begin{aligned}
&\left|\left\langle \eta_{i,j}^{(h)}(y),\eta_{i',j'}^{(h)}(z)\right\rangle - \Pi_{i,j,i',j'}^{(h-1)}(y,z)\cdot\dot{\Gamma}_{i,j,i',j'}^{(h)}(y,z) - \Gamma_{i,j,i',j'}^{(h)}(y,z)\right| \\
&\le \frac{\varepsilon}{10}\cdot\frac{(h-1)^2}{L+1}\cdot\sqrt{N_{i,j}^{(h)}(y)\cdot N_{i',j'}^{(h)}(z)}\cdot\left(\left|\dot{\Gamma}_{i,j,i',j'}^{(h)}(y,z)\right| + \frac{1}{q^2}\cdot\frac{\varepsilon}{8L}\right) + \frac{1}{q^2}\cdot\frac{\varepsilon}{8L}\cdot\left|\Pi_{i,j,i',j'}^{(h-1)}(y,z)\right| \\
&+ \frac{(h+1)\cdot\varepsilon^2}{60L^3}\cdot\frac{\sqrt{N_{i,j}^{(h)}(y)\cdot N_{i',j'}^{(h)}(z)}}{q^2} + \mathcal{O}\left(\frac{\varepsilon\cdot h}{L}\right)\cdot\frac{\sqrt{N_{i,j}^{(h)}(y)\cdot N_{i',j'}^{(h)}(z)}}{q^2} \\
&\le \frac{\varepsilon}{10}\cdot\frac{(h-1)^2}{L+1}\cdot\frac{\sqrt{N_{i,j}^{(h)}(y)\cdot N_{i',j'}^{(h)}(z)}}{q^2}\cdot\left(1+\frac{\varepsilon}{8L}\right) + \frac{h-1}{q^2}\cdot\frac{\varepsilon}{8L}\cdot\sqrt{N_{i,j}^{(h)}(y)\cdot N_{i',j'}^{(h)}(z)} \\
&+ \left(\frac{(h+1)\cdot\varepsilon^2}{60L^3} + \mathcal{O}\left(\frac{\varepsilon\cdot h}{L}\right)\right)\cdot\frac{\sqrt{N_{i,j}^{(h)}(y)\cdot N_{i',j'}^{(h)}(z)}}{q^2} \\
&\le \frac{\varepsilon}{10}\cdot\frac{h^2-h/2}{L+1}\cdot\frac{\sqrt{N_{i,j}^{(h)}(y)\cdot N_{i',j'}^{(h)}(z)}}{q^2}.
\end{aligned}$$

By plugging the above bound into Eq. (131) using triangle inequality and using Eq. (106) we get the following, with probability at least $1-\mathcal{O}\left(\frac{\delta}{L}\right)$:

$$\begin{aligned}
&\left|\left\langle \psi_{i,j}^{(h)}(y),\psi_{i',j'}^{(h)}(z)\right\rangle - \Pi_{i,j,i',j'}^{(h)}(y,z)\right| \\
&\le \mathcal{O}\left(\varepsilon h/L\right)\cdot\sqrt{N_{i,j}^{(h+1)}(y)\cdot N_{i',j'}^{(h+1)}(z)} \\
&\quad + \frac{\varepsilon}{10}\cdot\frac{h^2-h/2}{L+1}\cdot\sum_{a=-\frac{q-1}{2}}^{\frac{q-1}{2}}\sum_{b=-\frac{q-1}{2}}^{\frac{q-1}{2}}\frac{\sqrt{N_{i+a,j+b}^{(h)}(y)\cdot N_{i'+a,j'+b}^{(h)}(z)}}{q^2} \\
&\le \mathcal{O}\left(\varepsilon h/L\right)\cdot\sqrt{N_{i,j}^{(h+1)}(y)\cdot N_{i',j'}^{(h+1)}(z)} \\
&\quad + \frac{\varepsilon}{10}\cdot\frac{h^2-h/2}{L+1}\cdot\sqrt{\sum_{a=-\frac{q-1}{2}}^{\frac{q-1}{2}}\sum_{b=-\frac{q-1}{2}}^{\frac{q-1}{2}}\frac{N_{i+a,j+b}^{(h)}(y)}{q^2}}\cdot\sqrt{\sum_{a=-\frac{q-1}{2}}^{\frac{q-1}{2}}\sum_{b=-\frac{q-1}{2}}^{\frac{q-1}{2}}\frac{N_{i'+a,j'+b}^{(h)}(z)}{q^2}} \\
&\le \frac{\varepsilon}{10}\cdot\frac{h^2}{L+1}\cdot\sqrt{N_{i,j}^{(h+1)}(y)\cdot N_{i',j'}^{(h+1)}(z)}.
\end{aligned} \tag{132}$$

Similarly, we can prove that with probability at least $1-\mathcal{O}\left(\frac{\delta}{L}\right)$ the following hold simultaneously for all $i,i'\in[d_1]$ and all $j,j'\in[d_2]$,

$$\begin{aligned}
\left|\left\|\psi_{i,j}^{(h)}(y)\right\|_2^2 - \Pi_{i,j,i,j}^{(h)}(y,y)\right| &\le \frac{\varepsilon}{10}\cdot\frac{h^2}{L+1}\cdot N_{i,j}^{(h+1)}(y), \\
\left|\left\|\psi_{i',j'}^{(h)}(z)\right\|_2^2 - \Pi_{i',j',i',j'}^{(h)}(z,z)\right| &\le \frac{\varepsilon}{10}\cdot\frac{h^2}{L+1}\cdot N_{i',j'}^{(h+1)}(z).
\end{aligned}$$

This is sufficient to prove the inductive step for statement $P_2(h)$, in the case of $h < L$, i.e., $\Pr[P_2(h)|P_2(h-1), P_1(h), P_1(h-1)] \geq 1 - \mathcal{O}(\delta/L)$.

Now we prove the inductive step for $P_2(h)$ in the case of $h = L$. Similar to before, if we let $f_{i,j} := \psi_{i,j}^{(L-1)}(y) \otimes \dot{\phi}_{i,j}^{(L)}(y)$ and $g_{i',j'} := \psi_{i',j'}^{(L-1)}(z) \otimes \dot{\phi}_{i',j'}^{(L)}(z)$, then by Eq. (113), we have $\psi_{i,j}^{(L)}(y) = \boldsymbol{Q}^2 \cdot f_{i,j}$ and $\psi_{i',j'}^{(L)}(z) = \boldsymbol{Q}^2 \cdot g_{i',j'}$. Thus by Lemma 1 and union bound, we find that, with probability at least $1 - \mathcal{O}\left(\frac{\delta}{L}\right)$, the following inequality holds simultaneously for all $i, i' \in [d_1]$ and $j, j' \in [d_2]$:

$$\left|\left\langle \psi_{i,j}^{(L)}(y), \psi_{i',j'}^{(L)}(z) \right\rangle - \langle f_{i,j}, g_{i',j'} \rangle\right| \leq \mathcal{O}\left(\frac{\varepsilon}{L}\right) \cdot \|f_{i,j}\|_2 \|g_{i',j'}\|_2 .$$

Therefore, using Eq. (128) and Lemma 12 along with inductive hypotheses $P_2(L-1)$ and Lemma 13, with probability at least $1 - \mathcal{O}\left(\frac{\delta}{L}\right)$, the following holds simultaneously for all $i, i' \in [d_1]$ and $j, j' \in [d_2]$,

$$\begin{aligned}\left|\left\langle \psi_{i,j}^{(L)}(y), \psi_{i',j'}^{(L)}(z) \right\rangle - \langle f_{i,j}, g_{i',j'} \rangle\right| &\leq \mathcal{O}\left(\frac{\varepsilon}{L}\right) \sqrt{\Pi_{i,j,i,j}^{(L-1)}(y,y) \cdot \dot{\Pi}_{i,j,i,j}^{(L)}(y,y) \cdot \Pi_{i',j',i',j'}^{(L-1)}(z,z) \cdot \dot{\Pi}_{i',j',i',j'}^{(L)}(z,z)} \\ &= \mathcal{O}(\varepsilon) \cdot \frac{\sqrt{N_{i,j}^{(L)}(y) \cdot N_{i',j'}^{(L)}(z)}}{q^2}.\end{aligned}$$

By combining the above with inductive hypotheses $P_1(L), P_2(L-1)$ and Eq. (127) via triangle inequality and invoking Lemma 13 and also using the definition of $\Pi^{(L)}(y,z)$ given in Eq. (107), we get that the following holds, simultaneously for all $i, i' \in [d_1]$ and $j, j' \in [d_2]$, with probability at least $1 - \mathcal{O}\left(\frac{\delta}{L}\right)$,

$$\begin{aligned}&\left|\left\langle \psi_{i,j}^{(L)}(y), \psi_{i',j'}^{(L)}(z) \right\rangle - \Pi_{i,j,i',j'}^{(L)}(y,z)\right| \\ &\leq \frac{\varepsilon}{10} \cdot \frac{(L-1)^2}{L+1} \cdot \sqrt{N_{i,j}^{(L)}(y) \cdot N_{i',j'}^{(L)}(z)} \cdot \left(\left|\dot{\Pi}_{i,j,i',j'}^{(L)}(y,z)\right| + \frac{1}{q^2} \cdot \frac{\varepsilon}{8L}\right) + \frac{1}{q^2} \cdot \frac{\varepsilon}{8L} \cdot \left|\Pi_{i,j,i',j'}^{(L-1)}(y,z)\right| \\ &+ \frac{(L+1) \cdot \varepsilon^2}{60L^3} \cdot \frac{\sqrt{N_{i,j}^{(L)}(y) \cdot N_{i',j'}^{(L)}(z)}}{q^2} + \mathcal{O}(\varepsilon) \cdot \frac{\sqrt{N_{i,j}^{(L)}(y) \cdot N_{i',j'}^{(L)}(z)}}{q^2} \\ &\leq \frac{\varepsilon}{10} \cdot \frac{(L-1)^2}{L+1} \cdot \frac{\sqrt{N_{i,j}^{(L)}(y) \cdot N_{i',j'}^{(L)}(z)}}{q^2} \cdot \left(1 + \frac{\varepsilon}{8L}\right) + \frac{\varepsilon}{8q^2} \cdot \sqrt{N_{i,j}^{(L)}(y) \cdot N_{i',j'}^{(L)}(z)} \\ &+ \left(\frac{(L+1) \cdot \varepsilon^2}{60L^3} + \mathcal{O}(\varepsilon)\right) \cdot \frac{\sqrt{N_{i,j}^{(L)}(y) \cdot N_{i',j'}^{(L)}(z)}}{q^2} \\ &\leq \frac{\varepsilon \cdot (L-1)}{10} \cdot \frac{\sqrt{N_{i,j}^{(L)}(y) \cdot N_{i',j'}^{(L)}(z)}}{q^2}.\end{aligned}$$

This proves the inductive step for statement $P_2(h)$, in the case of $h = L$, i.e., $\Pr[P_2(L)|P_2(L-1), P_1(L), P_1(L-1)] \geq 1 - \mathcal{O}(\delta/L)$. The induction is complete and hence the statements of lemma are proved by union bounding over all $h = 0, 1, 2, \ldots L$. This completes the proof of Lemma 14. □

In the following lemma we analyze the runtime of the CNTK Sketch algorithm,

Lemma 15 (Runtime of the CNTK Sketch)**.** *For every positive integers d_1, d_2, c, and L, every $\varepsilon, \delta > 0$, every image $x \in \mathbb{R}^{d_1 \times d_2 \times c}$, the time to compute the CNTK Sketch $\Psi_{\text{cntk}}^{(L)}(x) \in \mathbb{R}^{s^*}$, for $s^* = \mathcal{O}\left(\frac{1}{\varepsilon^2} \cdot \log \frac{1}{\delta}\right)$, using the procedure given in Definition 3 is bounded by $\mathcal{O}\left(\frac{L^{11}}{\varepsilon^{6.7}} \cdot (d_1 d_2) \cdot \log^3 \frac{d_1 d_2 L}{\varepsilon \delta}\right)$.*

Proof of Lemma 15: First note that the total time to compute $N_{i,j}^{(h)}(x)$ for all $i \in [d_1]$ and $j \in [d_2]$ and $h = 0, 1, \ldots L$ as per Eq. (103) is bounded by $\mathcal{O}\left(q^2 L \cdot d_1 d_2\right)$. Besides the time to compute $N_{i,j}^{(h)}(x)$, there are two other main components to the runtime of this procedure. The first heavy operation corresponds to computing vectors $\left[Z_{i,j}^{(h)}(x)\right]_l = \boldsymbol{Q}^{2p+2} \cdot \left(\left[\mu_{i,j}^{(h)}(x)\right]^{\otimes l} \otimes e_1^{\otimes 2p+2-l}\right)$ for $l = 0, 1, 2, \ldots 2p+2$ and $h = 1, 2, \ldots L$ and all indices $i \in [d_1]$ and $j \in [d_2]$, in Eq. (110).

By Lemma 1, the time to compute $\left[Z_{i,j}^{(h)}(x)\right]_l$ for a fixed h, fixed $i \in [d_1]$ and $j \in [d_2]$, and all $l = 0, 1, 2, \ldots 2p+2$ is bounded by,

$$\mathcal{O}\left(\frac{L^{10}}{\varepsilon^{20/3}} \cdot \log^2 \frac{L}{\varepsilon} \cdot \log^3 \frac{d_1 d_2 L}{\varepsilon\delta} + q^2 \cdot \frac{L^8}{\varepsilon^{16/3}} \cdot \log^3 \frac{d_1 d_2 L}{\varepsilon\delta}\right) = \mathcal{O}\left(\frac{L^{10}}{\varepsilon^{6.7}} \cdot \log^3 \frac{d_1 d_2 L}{\varepsilon\delta}\right).$$

The total time to compute vectors $\left[Z_{i,j}^{(h)}(x)\right]_l$ for all $h = 1, 2, \ldots L$ and all $l = 0, 1, 2, \ldots 2p+2$ and all indices $i \in [d_1]$ and $j \in [d_2]$ is thus bounded by $\mathcal{O}\left(\frac{L^{11}}{\varepsilon^{6.7}} \cdot (d_1 d_2) \cdot \log^3 \frac{d_1 d_2 L}{\varepsilon\delta}\right)$. The next computationally expensive operation is computing vectors $\left[Y_{i,j}^{(h)}(x)\right]_l$ for $l = 0, 1, 2, \ldots 2p'+1$ and $h = 1, 2, \ldots L$, and all indices $i \in [d_1]$ and $j \in [d_2]$, in Eq. (111). By Lemma 1, the runtime of computing $\left[Y_{i,j}^{(h)}(x)\right]_l$ for a fixed h, fixed $i \in [d_1]$ and $j \in [d_2]$, and all $l = 0, 1, 2, \ldots 2p'+1$ is bounded by,

$$\mathcal{O}\left(\frac{L^6}{\varepsilon^6} \cdot \log^2 \frac{L}{\varepsilon} \log^3 \frac{d_1 d_2 L}{\varepsilon\delta} + \frac{q^2 \cdot L^8}{\varepsilon^6} \cdot \log^3 \frac{d_1 d_2 L}{\varepsilon\delta}\right) = \mathcal{O}\left(\frac{L^8}{\varepsilon^6} \log^2 \frac{L}{\varepsilon} \cdot \log^3 \frac{d_1 d_2 L}{\varepsilon\delta}\right).$$

Hence, the total time to compute vectors $\left[Y_{i,j}^{(h)}(x)\right]_l$ for all $h = 1, 2, \ldots L$ and $l = 0, 1, 2, \ldots 2p'+1$ and all indices $i \in [d_1]$ and $j \in [d_2]$ is $\mathcal{O}\left(\frac{L^9}{\varepsilon^6} \log^2 \frac{L}{\varepsilon} \cdot (d_1 d_2) \cdot \log^3 \frac{d_1 d_2 L}{\varepsilon\delta}\right)$. The total runtime bound is obtained by summing up these three contributions. This completes the proof of Lemma 15. □

Now we are ready to prove Theorem 4.

Theorem 4. *For every positive integers d_1, d_2, c and $L \geq 2$, and every $\varepsilon, \delta > 0$, if we let $\Theta_{\mathtt{cntk}}^{(L)} : \mathbb{R}^{d_1 \times d_2 \times c} \times \mathbb{R}^{d_1 \times d_2 \times c} \to \mathbb{R}$ be the L-layer CNTK with ReLU activation and GAP given in [5], then there exist a randomized map $\Psi_{\mathtt{cntk}}^{(L)} : \mathbb{R}^{d_1 \times d_2 \times c} \to \mathbb{R}^{s^*}$ for some $s^* = \mathcal{O}\left(\frac{1}{\varepsilon^2} \log \frac{1}{\delta}\right)$ such that:*

1. *For any images $y, z \in \mathbb{R}^{d_1 \times d_2 \times c}$:*

$$\Pr\left[\left|\left\langle \Psi_{\mathtt{cntk}}^{(L)}(y), \Psi_{\mathtt{cntk}}^{(L)}(z)\right\rangle - \Theta_{\mathtt{cntk}}^{(L)}(y,z)\right| \leq \varepsilon \cdot \Theta_{\mathtt{cntk}}^{(L)}(y,z)\right] \geq 1 - \delta.$$

2. *For every image $x \in \mathbb{R}^{d_1 \times d_2 \times c}$, time to compute $\Psi_{\mathtt{cntk}}^{(L)}(x)$ is $\mathcal{O}\left(\frac{L^{11}}{\varepsilon^{6.7}} \cdot (d_1 d_2) \cdot \log^3 \frac{d_1 d_2 L}{\varepsilon\delta}\right)$.*

Proof of Theorem 4: Let $\psi^{(L)} : \mathbb{R}^{d_1 \times d_2 \times c} \to \mathbb{R}^{d_1 \times d_2 \times s}$ for $s = \mathcal{O}\left(\frac{L^4}{\varepsilon^2} \cdot \log^3 \frac{d_1 d_2 L}{\varepsilon\delta}\right)$ be the mapping defined in Eq. (113) of Definition 3. By Eq. (114), the CNTK Sketch $\Psi_{\mathtt{cntk}}^{(L)}(x)$ is defined as

$$\Psi_{ntk}^{(L)}(x) := \frac{1}{d_1 d_2} \cdot \boldsymbol{G} \cdot \left(\sum_{i \in [d_1]} \sum_{j \in [d_2]} \psi_{i,j}^{(L)}(x)\right).$$

The matrix $\boldsymbol{G}$ is defined in Eq. (114) to be a matrix of i.i.d. normal entries with $s^* = C \cdot \frac{1}{\varepsilon^2} \cdot \log \frac{1}{\delta}$ rows for large enough constant C. [15] shows that $\boldsymbol{G}$ is a JL transform and hence $\Psi_{\mathtt{cntk}}^{(L)}$ satisfies the following,

$$\Pr\left[\left|\left\langle \Psi_{\mathtt{cntk}}^{(L)}(y), \Psi_{\mathtt{cntk}}^{(L)}(z)\right\rangle - \frac{1}{d_1^2 d_2^2} \cdot \sum_{i,i' \in [d_1]} \sum_{j,j' \in [d_2]} \left\langle \psi_{i,j}^{(L)}(y), \psi_{i',j'}^{(L)}(z)\right\rangle\right| \leq \mathcal{O}(\varepsilon) \cdot A\right] \geq 1 - \mathcal{O}(\delta),$$

where $A := \frac{1}{d_1^2 d_2^2} \cdot \left\|\sum_{i \in [d_1]} \sum_{j \in [d_2]} \psi_{i,j}^{(L)}(y)\right\|_2 \cdot \left\|\sum_{i \in [d_1]} \sum_{j \in [d_2]} \psi_{i,j}^{(L)}(z)\right\|_2$. By triangle inequality together with Lemma 14 and Lemma 13, the following bounds hold with probability at least $1 - \mathcal{O}(\delta)$:

$$\left\|\sum_{i \in [d_1]} \sum_{j \in [d_2]} \psi_{i,j}^{(L)}(y)\right\|_2 \leq \frac{11}{10} \cdot \frac{\sqrt{L-1}}{q} \cdot \sum_{i \in [d_1]} \sum_{j \in [d_2]} \sqrt{N_{i,j}^{(L)}(y)},$$

$$\left\|\sum_{i \in [d_1]} \sum_{j \in [d_2]} \psi_{i,j}^{(L)}(z)\right\|_2 \leq \frac{11}{10} \cdot \frac{\sqrt{L-1}}{q} \cdot \sum_{i \in [d_1]} \sum_{j \in [d_2]} \sqrt{N_{i,j}^{(L)}(z)},$$

Therefore, by union bound we find that, with probability at least $1 - \mathcal{O}(\delta)$:

$$\left|\left\langle \Psi^{(L)}_{\texttt{cntk}}(y), \Psi^{(L)}_{\texttt{cntk}}(z)\right\rangle - \frac{1}{d_1^2 d_2^2} \cdot \sum_{i,i' \in [d_1]} \sum_{j,j' \in [d_2]} \left\langle \psi^{(L)}_{i,j}(y), \psi^{(L)}_{i',j'}(z)\right\rangle\right|$$
$$\leq \mathcal{O}\left(\frac{\varepsilon L}{q^2 \cdot d_1^2 d_2^2}\right) \cdot \sum_{i,i' \in [d_1]} \sum_{j,j' \in [d_2]} \sqrt{N^{(L)}_{i,j}(y) \cdot N^{(L)}_{i',j'}(z)}.$$

Be combining the above with Lemma 14 using triangle inequality and union bound and also using Eq. (108), the following holds with probability at least $1 - \mathcal{O}(\delta)$:

$$\left|\left\langle \Psi^{(L)}_{\texttt{cntk}}(y), \Psi^{(L)}_{\texttt{cntk}}(z)\right\rangle - \Theta^{(L)}_{\texttt{cntk}}(y,z)\right| \leq \frac{\varepsilon \cdot (L-1)}{9q^2 \cdot d_1^2 d_2^2} \cdot \sum_{i,i' \in [d_1]} \sum_{j,j' \in [d_2]} \sqrt{N^{(L)}_{i,j}(y) \cdot N^{(L)}_{i',j'}(z)}. \tag{133}$$

Now we prove that $\Theta^{(L)}_{\texttt{cntk}}(y,z) \geq \frac{L-1}{9q^2 d_1^2 d_2^2} \cdot \sum_{i,i' \in [d_1]} \sum_{j,j' \in [d_2]} \sqrt{N^{(L)}_{i,j}(y) \cdot N^{(L)}_{i',j'}(z)}$ for every $L \geq 2$. First note that, it follows from Eq. (104) that $\Gamma^{(1)}_{i,j,i',j'}(y,z) \geq 0$ for any i, i', j, j' because the function κ_1 is non-negative everywhere on $[-1,1]$. This also implies that $\Gamma^{(2)}_{i,j,i',j'}(y,z) \geq \frac{\sqrt{N^{(2)}_{i,j}(y) \cdot N^{(2)}_{i',j'}(z)}}{\pi \cdot q^2}$ because $\kappa_1(\alpha) \geq \frac{1}{\pi}$ for every $\alpha \in [0,1]$. Since, $\kappa_1(\cdot)$ is a monotone increasing function, by recursively using Eq. (103) and Eq. (104) along with Lemma 11, we can show that for every $h \geq 1$, the value of $\Gamma^{(h)}_{i,j,i',j'}(y,z)$ is lower bounded by $\frac{\sqrt{N^{(h)}_{i,j}(y) \cdot N^{(h)}_{i',j'}(z)}}{q^2} \cdot \Sigma^{(h)}_{\texttt{relu}}(-1)$, where $\Sigma^{(h)}_{\texttt{relu}} : [-1,1] \to \mathbb{R}$ is the function defined in Eq. (3).

Furthermore, it follows from Eq. (105) that $\dot{\Gamma}^{(1)}_{i,j,i',j'}(y,z) \geq 0$ for any i, i', j, j' because the function κ_0 is non-negative everywhere on $[-1,1]$. Additionally, $\dot{\Gamma}^{(2)}_{i,j,i',j'}(y,z) \geq \frac{1}{2q^2}$ because $\kappa_0(\alpha) \geq \frac{1}{2}$ for every $\alpha \in [0,1]$. By using the inequality $\Gamma^{(h)}_{i,j,i',j'}(y,z) \geq \frac{\sqrt{N^{(h)}_{i,j}(y) \cdot N^{(h)}_{i',j'}(z)}}{q^2} \cdot \Sigma^{(h)}_{\texttt{relu}}(-1)$ that we proved above along with the fact that $\kappa_0(\cdot)$ is a monotone increasing function and recursively using Eq. (105) and Lemma 11, it follows that for every $h \geq 1$, we have $\dot{\Gamma}^{(h)}_{i,j,i',j'}(y,z) \geq \frac{1}{q^2} \cdot \dot{\Sigma}^{(h)}_{\texttt{relu}}(-1)$.

By using these inequalities and Definition of $\Pi^{(h)}$ in Eq. (106) together with Eq. (103), recursively, it follows that, for every i, i', j, j' and $h = 2, \ldots L-1$:

$$\Pi^{(h)}_{i,j,i',j'}(y,z) \geq \frac{h}{4} \cdot \sqrt{N^{(h+1)}_{i,j}(y) \cdot N^{(h+1)}_{i',j'}(z)},$$

Therefore, using this inequality and Eq. (107) we have that for every $L \geq 2$:

$$\Pi^{(L)}_{i,j,i',j'}(y,z) \geq \frac{L-1}{4} \cdot \sqrt{N^{(L)}_{i,j}(y) \cdot N^{(L)}_{i',j'}(z)} \cdot \frac{\dot{\Sigma}^{(L)}_{\texttt{relu}}(-1)}{q^2}$$
$$\geq \frac{L-1}{9q^2} \cdot \sqrt{N^{(L)}_{i,j}(y) \cdot N^{(L)}_{i',j'}(z)}.$$

Now using this inequality and Eq. (108), the following holds for every $L \geq 2$:

$$\Theta^{(L)}_{\texttt{cntk}}(y,z) \geq \frac{L-1}{9q^2 d_1^2 d_2^2} \cdot \sum_{i,i' \in [d_1]} \sum_{j,j' \in [d_2]} \sqrt{N^{(L)}_{i,j}(y) \cdot N^{(L)}_{i',j'}(z)}.$$

Therefore, by incorporating the above into Eq. (133) we get that,

$$\Pr\left[\left|\left\langle \Psi^{(L)}_{\texttt{cntk}}(y), \Psi^{(L)}_{\texttt{cntk}}(z)\right\rangle - \Theta^{(L)}_{\texttt{cntk}}(y,z)\right| \leq \varepsilon \cdot \Theta^{(L)}_{\texttt{cntk}}(y,z)\right] \geq 1 - \delta.$$

Runtime analysis: By Lemma 15, time to compute the CNTK Sketch is $\mathcal{O}\left(\frac{L^{11}}{\varepsilon^{6.7}} \cdot (d_1 d_2) \cdot \log^3 \frac{d_1 d_2 L}{\varepsilon \delta}\right)$.

This completes the proof of Theorem 4. □