Lossless Compression of Deep Neural Networks: A High-dimensional Neural Tangent Kernel Approach

Lingyu Gu[*1] **Yongqi Du**[*1] **Yuan Zhang**[2] **Di Xie**[2] **Shiliang Pu**[2] **Robert C. Qiu**[1]

Zhenyu Liao[†1]
[1]EIC, Huazhong University of Science and Technology, China
[2]Hikvision Research Institute, Hangzhou, China

Abstract

Modern deep neural networks (DNNs) are extremely powerful; however, this comes at the price of increased depth and having more parameters per layer, making their training and inference more computationally challenging. In an attempt to address this key limitation, efforts have been devoted to the compression (e.g., sparsification and/or quantization) of these large-scale machine learning models, so that they can be deployed on low-power IoT devices. In this paper, building upon recent research advances in neural tangent kernel (NTK) and random matrix theory, we provide a novel compression approach to wide and fully-connected *deep* neural nets. Specifically, we demonstrate that in the high-dimensional regime where the number of data points n and their dimension p are both large, and under a Gaussian mixture model for the data, there exists *asymptotic spectral equivalence* between the NTK matrices for a large family of DNN models. This theoretical result enables "lossless" compression of a given DNN to be performed, in the sense that the compressed network yields asymptotically the same NTK as the original (dense and unquantized) network, with its weights and activations taking values *only* in $\{0, \pm 1\}$ up to a scaling. Experiments on both synthetic and real-world data are conducted to support the numerical advantages of the proposed method.

1 Introduction

Modern deep neural networks (DNNs) are becoming increasingly over-parameterized, having more parameters than required to fit the also increasingly large, complex, and high-dimensional data. While the list of successful applications of these large-scale machine learning (ML) models is rapidly growing, the energy consumption of these models is also increasing, making them more challenging to deploy on close-to-user and low-power devices. To address this issue, compression techniques have been proposed that prune, sparsify, and/or quantize DNN models [12, 20], thereby yielding DNNs of a much smaller size that can still achieve satisfactory performance on a given ML task. As an illustrative example, it has been recently demonstrated that at least 90% of the weights in popular DNN models such as VGG19 and ResNet32 can be removed with virtually no performance loss [50].

Despite the remarkable progress achieved by various DNN model compression techniques, due to the nonlinear and highly non-convex nature of DNNs, our theoretical understanding of these large-scale ML models, as well as of their compression schemes, is progressing at a more modest pace. For example, it is unclear how much a given DNN model can be compressed *without* severe performance

[*]Equal contribution, listed in random order by rolling a dice on WeChat.

[†]Author to whom any correspondence should be addressed. Email: zhenyu_liao@hust.edu.cn

Preprint. Under review.

degradation; perhaps more importantly, on the degree to which such a *trade-off between performance and complexity* depends on the ML problem and the data also remains unknown.

In this respect, neural tangent kernels (NTKs) [25], provide a powerful tool for use in assessing the convergence and generalization properties of very wide (sometimes unrealistically so) DNNs by studying their corresponding NTK eigenspectra, which are *solely* dependent on the input data, the network activation function, and the random weights distribution.[†]

In this paper, building upon recent advances in random matrix theory (RMT) and high-dimensional statistics, we demonstrate that for data $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ drawn from a K-class Gaussian mixture model (GMM), in a high-dimensional and non-trivial classification regime where the input data dimension p and their size n are both large and comparable, the *eigenspectra* of both the NTK and the closely related conjugate kernel (CK) matrices at *any* layer $\ell \in \{1, \ldots, L\}$ are *independent* of the distribution of the i.i.d. entries of the (random) weight matrix $\mathbf{W}_\ell$, provided that they are normalized to have zero mean and unit variance, and *only* depend on the activation function $\sigma_\ell(\cdot)$ via *four* scalar parameters. In a sense, we establish, at least for GMM data, the asymptotic *spectral equivalence* between the NTK matrices of the corresponding network layers, and consequently of the whole network, for a large family of DNN models with possibly very different weights and activations, given that they have normalized weight entries and share the same few activation parameters.

Since the convergence and generalization properties of ultra-wide DNNs depend only on the eigenspectra (i.e., eigenvalue-eigenvector pairs; see also Remark 2 below) of the corresponding NTK matrices [25, 16], we further exploit the above theoretical results to propose a novel NTK-based approach that allows for the "lossless" compression of a given *fully-connected* DNN model, by designing a sparse and quantized DNN that (i) has asymptotically the *same* NTK eigenspectra as the original "dense and full precision" network, and (ii) has both weights and activations taking values in the set $\{-1, 0, +1\}$ before scaling, and can thus be stored and computed much more efficiently.

Despite being derived here for Gaussian mixture data, an unexpected close match is observed between our theory and the empirical results on real-world datasets, suggesting possibly wider applicability for the proposed NTK-based Lossless Compression (NTK-LC) approach. Looking forward, we expect that our analysis will open the door to improved analysis based on RMT and high-dimensional statistics, which will demystify the seemingly striking empirical observations in modern DNNs.

1.1 Our contributions

Our main results can be summarized as follows:

1. We provide, in Theorems 1 and 2 respectively, for GMM data and in the high-dimensional regime (Assumption 1), *precise* spectral characterizations of the CK and NTK matrices of fully-connected networks; furthermore, we show that the CK and NTK eigenspectra do *not* depend on the distribution of i.i.d. weights, and *depend* solely on the activation function via a few scalar parameters.
2. In Corollary 1 and Algorithm 1, we exploit these results to propose a novel DNN compression scheme, named NTK-based Lossless Compression (NTK-LC), with *sparsified and ternarized* weights and activations, which nevertheless do not affect (the spectral behavior of) the NTK matrices.
3. In Section 4, we present empirical evidence on (not so) wide DNNs trained with both synthetic Gaussian and popular real-world data such as MNIST [28] and CIFAR10 [27], and show a factor of 10^3 less memory is needed with the proposed NTK-LC approach, with virtually no performance loss.

1.2 Related work

Neural network model compression. The study of NN compression dates back to early 1990 [29], at which point, in the absence of the (possibly more than) sufficient computational power that we have today, compression techniques allowed neural networks to be empirically evaluated on computers with limited computational and/or storage resources [46]. Alongside the rapid growth of increasingly powerful computing devices, the development of more efficient NN architectures and training/inference protocols, and the need to implement NNs on mobile and low-power devices, (D)NN model compression has become an active research topic and many elegant and efficient

[†] In the remainder of this article, what we refer to as the "NTK matrix" is essentially the limiting (nonrandom) NTK matrix to which the random NTK converges under the infinite-wide limit.

compression approaches have been proposed over the years [19, 22, 24, 20]. However, due to the nonlinear and highly non-convex nature of DNNs, our theoretical understanding of these large-scale ML models, as well as of (e.g., the fundamental "performance and complexity" trade-off of) compressed DNNs, is somewhat limited [20].

Neural tangent kernel. Neural tangent kernel (NTK) theory recently proposed in [25], by considering the limit of infinitely wide DNNs, characterizes the convergence and generalization properties of very wide DNNs when trained using gradient descent with small steps. Initially proposed for fully-connected nets, the NTK framework has been subsequently extended to convolutional [3], graph [14], and recurrent [1] settings. The NTK theory, while having the advantage of being mathematically more tractable (via, e.g., the characterization of the associated reproducing kernel Hilbert space [6]), seems to diverge from the regime on which modern (and not so wide) DNNs operate, see [8, 35, 16].

Random matrix theory and neural networks. Random matrix theory (RMT), a powerful and flexible tool for assessing the behavior of large-scale systems with a large "degree of freedom," is gaining popularity in the field of NN analysis [41, 42], in both shallow [45, 33, 34] and deep [5, 16, 44] settings, and considers both homogeneous (e.g., standard normal) [45, 43] and mixture-type data [33, 2]. From a technical perspective, the most relevant paper is [2] , in which the authors proposed a RMT-inspired NN compression scheme, albeit only in the single-hidden-layer setting. This paper extends the analysis in [2] to multi-layer fully-connected DNNs by focusing on the associated NTK matrices, and proposes a novel sparsification and quantization scheme for fully-connected DNNs (which is in spirit similar to, although formally different from, that proposed in [2]).

1.3 Notations and organization of the paper

We denote scalars by lowercase letters, vectors by bold lowercase, and matrices by bold uppercase. We denote the transpose operator by $(\cdot)^\mathsf{T}$, and use $\|\cdot\|$ to denote the Euclidean norm for vectors and the spectral/operator norm for matrices. For a random variable z, $\mathbb{E}[z]$ denotes the expectation of z. We use $\mathbf{1}_p$ and $\mathbf{I}_p$ to represent an all-ones vector of dimension p and the identity matrix of size $p \times p$.

The remainder of this article is structured as follows. In Section 2, we present the DNN model under study, together with our working assumptions. Section 3 contains our main technical results on the eigenspectra of the conjugate kernel $\mathbf{K}_{\rm CK}$ and NTK matrix $\mathbf{K}_{\rm NTK}$, along with an account of how they apply to the compression of fully-connected deep neural nets with the proposed NTK-based Lossless Compression (NTK-LC) approach. Empirical evidence is provided in Section 4 to demonstrate the significant computation and storage savings, with virtually no performance degradation, that can be obtained using NTK-LC. Conclusion and future perspectives are placed in Section 5.

2 Preliminaries

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ be n data vectors independently drawn from one of the K-class Gaussian mixtures $\mathcal{C}_1, \ldots, \mathcal{C}_K$, and let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p \times n}$, with class $\mathcal{C}_a$ having cardinality n_a; that is,

$$\mathbf{x}_i \in \mathcal{C}_a \Leftrightarrow \mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}_a/\sqrt{p}, \mathbf{C}_a/p), \tag{1}$$

for the mean vector $\boldsymbol{\mu}_a \in \mathbb{R}^p$ and covariance matrix $\mathbf{C}_a \in \mathbb{R}^{p \times p}$ associated with class $\mathcal{C}_a$.

In the high-dimensional scenario where n, p are both large and comparable, we position ourselves in the following non-trivial classification setting, so that the K-class classification above is neither trivially easy nor impossible; see also [11] and [7, Section 2].

Assumption 1 (High-dimensional asymptotics)**.** *As $n \to \infty$, we have, for $a \in \{1, \ldots, K\}$ that (i) $p/n \to c \in (0, \infty)$ and $n_a/n \to c_a \in [0, 1)$; (ii) $\|\boldsymbol{\mu}_a\| = O(1)$; (iii) for $\mathbf{C}^\circ \equiv \sum_{a=1}^K \frac{n_a}{n} \mathbf{C}_a$ and $\mathbf{C}_a^\circ \equiv \mathbf{C}_a - \mathbf{C}^\circ$, we have $\|\mathbf{C}_a\| = O(1)$, $\operatorname{tr} \mathbf{C}_a^\circ = O(\sqrt{p})$ and $\operatorname{tr}(\mathbf{C}_a \mathbf{C}_b) = O(p)$ for $a, b \in \{1, \ldots, K\}$; and (iv) $\tau_0 \equiv \sqrt{\operatorname{tr} \mathbf{C}^\circ / p}$ converges in $(0, \infty)$.*

We consider using a *fully-connected* neural network model of depth L for the classification of the above K-class Gaussian mixture. Such a network can be parameterized by a sequence of weight matrices $\mathbf{W}_1 \in \mathbb{R}^{d_1 \times d_0}, \ldots, \mathbf{W}_L \in \mathbb{R}^{d_L \times d_{L-1}}$ (with $d_0 = p$), and nonlinear activation functions $\sigma_1, \ldots, \sigma_L$ that apply entry-wise, so that the network output is given by the following:

$$f(\mathbf{x}) = \frac{1}{\sqrt{d_L}} \mathbf{w}^\mathsf{T} \sigma_L \left(\frac{1}{\sqrt{d_{L-1}}} \mathbf{W}_L \sigma_{L-1} \left(\cdots \frac{1}{\sqrt{d_2}} \sigma_2 \left(\frac{1}{\sqrt{d_1}} \mathbf{W}_2 \sigma_1 (\mathbf{W}_1 \mathbf{x}) \right) \right) \right), \tag{2}$$

for an input data vector $\mathbf{x} \in \mathbb{R}^p$ and output vector $\mathbf{w} \in \mathbb{R}^{d_L}$. We denote $\mathbf{\Sigma}_\ell \in \mathbb{R}^{d_\ell \times n}$ the representations of the data matrix $\mathbf{X} \in \mathbb{R}^{p \times n}$ at layer $\ell \in \{1, \ldots, L\}$ defined as

$$\mathbf{\Sigma}_\ell = \frac{1}{\sqrt{d_\ell}} \sigma_\ell \left(\frac{1}{\sqrt{d_{\ell-1}}} \mathbf{W}_\ell \sigma_{\ell-1} \left(\ldots \frac{1}{\sqrt{d_2}} \sigma_2 \left(\frac{1}{\sqrt{d_1}} \mathbf{W}_2 \sigma_1 (\mathbf{W}_1 \mathbf{X}) \right) \right) \right). \tag{3}$$

The normalization by $\sqrt{d_\ell}$ follows from the NTK literature and ensures the consistent asymptotic behavior of the network in the high-dimensional setting in Assumption 1 and 2; see also [25, 6, 16].

The training and generalization performance of the neural network model defined in (2) are closely related to two types of kernel matrices: the Conjugate Kernel (CK) matrix and Neural Tangent Kernel (NTK) matrix, defined respectively for $\ell \in \{1, \ldots, L\}$ as follows:

$$\mathbf{K}_{\mathrm{CK},\ell} = \mathbb{E}[\mathbf{\Sigma}_\ell^\mathsf{T} \mathbf{\Sigma}_\ell] \in \mathbb{R}^{n \times n}, \tag{4}$$

with expectation taken with respect to the weights and $\mathbf{\Sigma}_\ell \in \mathbb{R}^{d_\ell \times n}$ the data representation at the output of layer ℓ defined in (3). In particular, CKs are known to satisfy the following recursive relation [25, 6]

$$[\mathbf{K}_{\mathrm{CK},\ell}]_{ij} = \mathbb{E}_{u,v}[\sigma_\ell(u)\sigma_\ell(v)], \text{ with } u, v \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} [\mathbf{K}_{\mathrm{CK},\ell-1}]_{ii} & [\mathbf{K}_{\mathrm{CK},\ell-1}]_{ij} \\ [\mathbf{K}_{\mathrm{CK},\ell-1}]_{ij} & [\mathbf{K}_{\mathrm{CK},\ell-1}]_{jj} \end{bmatrix} \right), \tag{5}$$

while for the NTK matrix $\mathbf{K}_{\mathrm{NTK},\ell} \in \mathbb{R}^{n \times n}$ of layer ℓ, we have:

$$\mathbf{K}_{\mathrm{NTK},\ell} = \mathbf{K}_{\mathrm{CK},\ell} + \mathbf{K}_{\mathrm{NTK},\ell-1} \circ \mathbf{K}'_{\mathrm{CK},\ell}, \quad \mathbf{K}_{\mathrm{NTK},0} = \mathbf{K}_{\mathrm{CK},0} = \mathbf{X}^\mathsf{T}\mathbf{X}, \tag{6}$$

where '$\mathbf{A} \circ \mathbf{B}$' denotes the Hadamard product between two matrices $\mathbf{A}, \mathbf{B}$ of the same size, and $\mathbf{K}'_{\mathrm{CK},\ell}$ denotes the CK matrix with nonlinear function σ'_ℓ instead of σ_ℓ as for $\mathbf{K}_{\mathrm{CK},\ell}$ defined in (5); that is, $[\mathbf{K}'_{\mathrm{CK},\ell}]_{ij} = \mathbb{E}_{u,v}[\sigma'_\ell(u)\sigma'_\ell(v)]$. Note in particular that for a given DNN model, the corresponding CK and NTK matrices depend *only* on the network structure (i.e., the number of layers and the activation function in each layer), the *distribution* of the random (initializations of the) weights to be integrated over (e.g., in the expectation in equation (4)), and the input data.

It has been shown in a series of previous efforts [25, 16, 23] that for very (and sometimes unrealistically) wide DNNs trained using gradient descent with a small step size, the time evolution of the residual errors and in-sample predictions of a given DNN are *explicit* functionals of the corresponding $\mathbf{K}_{\mathrm{NTK}}$ involving its eigenvalues and eigenvectors. In this respect, the NTK theory provides, via the eigenspectral behavior of $\mathbf{K}_{\mathrm{NTK}}$, precise characterizations of the convergence and generalization properties of DNNs [25, 6], by focusing on the impact of the network structure (e.g., the number of layers and the choice of activation functions), the input data, and the weight initialization schemes.

In this paper, we focus on fully-connected nets under the following assumption regarding the weights.

Assumption 2 (On random weights)**.** *The weight matrices $\mathbf{W}_1 \in \mathbb{R}^{d_1 \times p}, \ldots, \mathbf{W}_L \in \mathbb{R}^{d_L \times d_{L-1}}$ are independent and have i.i.d. entries of zero mean, unit variance, and finite fourth-order moment.*

Assumption 2, together with the $\sqrt{d_\ell}$ normalization, is compatible with *fully-connected* DNNs in (2), which are admittedly less interesting, from a practical perspective, compared to their convolutional counterparts. The proposed framework is envisioned to be extendable to a convolutional [3, 6] and more involved setting (e.g., graph NNs [14]) by considering (e.g., Toeplitz-type) structures on $\mathbf{W}$s.

Unlike most existing NTK literature [25, 6, 15], we do not assume the Gaussianity of the entries of $\mathbf{W}_\ell$s, but only that they are i.i.d. and "normalized" to have zero mean and unit variance. As it turns out, this assumption together with a (Lyapunov-type) central limit theorem argument, is sufficient to establish most existing results on the convergence and generalization of DNNs; see for example [31].

Assumption 3 (On activation functions)**.** *The activation functions $\sigma_1(\cdot), \ldots, \sigma_L(\cdot)$ in each layer are at least four-times differentiable with respect to standard normal measure, in the sense that $\max_{k \in \{0,1,2,3,4\}} \{|\mathbb{E}[\sigma_\ell^{(k)}(\xi)]|\}$ is finite for $\xi \sim \mathcal{N}(0,1)$ and $\ell \in \{1, \ldots, L\}$.*

Using the Gaussian integration by parts formula, one has $\mathbb{E}[\sigma'(\xi)] = \mathbb{E}[\xi\sigma(\xi)]$ for $\xi \sim \mathcal{N}(0,1)$, as long as the right-hand side expectation exists. As a result, it suffices to have $|\sigma_\ell|$ upper-bounded by some (high-degree) polynomial function for Assumption 3 to hold.

With these preliminaries, we can move on to present our main technical results on the spectral behavior of the CK and NTK matrices for a large family of fully-connected DNNs.

3 Main results

For a fully-connected DNN defined in (2), our first result is on the eigenspectral behavior of the corresponding CK matrices $\mathbf{K}_{\mathrm{CK}}$ defined in (4). More specifically, we show for Gaussian mixture data in (1) and in the high-dimensional setting of Assumption 1, that the $\mathbf{K}_{\mathrm{CK},\ell}$ of layer $\ell \in \{1, \ldots, L\}$ is asymptotically *spectrally equivalent* to another random matrix $\tilde{\mathbf{K}}_{\mathrm{CK},\ell}$, in the sense that their spectral norm difference $\|\mathbf{K}_{\mathrm{CK},\ell} - \tilde{\mathbf{K}}_{\mathrm{CK},\ell}\|$ vanishes as $n, p \to \infty$. This result is stated as follows, the proof of which is based on an induction on ℓ and is given in Section A.1 of the appendix.

Theorem 1 (Asymptotic equivalents for CK matrices)**.** *Let Assumptions 1–3 hold, and let $\tau_0, \tau_1, \ldots, \tau_L \geq 0$ be a sequence of non-negative numbers satisfying the following recursion:*

$$\tau_\ell = \sqrt{\mathbb{E}[\sigma_\ell^2(\tau_{\ell-1}\xi)]}, \quad \xi \sim \mathcal{N}(0,1), \quad \ell \in \{1, \ldots, L\}. \tag{7}$$

Further assume that the activation functions $\sigma_\ell(\cdot)$s are "centered," such that $\mathbb{E}[\sigma_\ell(\tau_{\ell-1}\xi)] = 0$. Then, for the CK matrix $\mathbf{K}_{\mathrm{CK},\ell}$ of layer $\ell \in \{1, \ldots, L\}$ defined in (4)*, as $n, p \to \infty$, one has that:*

$$\|\mathbf{K}_{\mathrm{CK},\ell} - \tilde{\mathbf{K}}_{\mathrm{CK},\ell}\| \to 0, \quad \tilde{\mathbf{K}}_{\mathrm{CK},\ell} \equiv \alpha_{\ell,1}\mathbf{X}^\mathsf{T}\mathbf{X} + \mathbf{V}\mathbf{A}_\ell\mathbf{V}^\mathsf{T} + (\tau_\ell^2 - \tau_0^2\alpha_{\ell,1} - \tau_0^4\alpha_{\ell,3})\mathbf{I}_n, \tag{8}$$

almost surely, with

$$\mathbf{V} = [\mathbf{J}/\sqrt{p},\ \boldsymbol{\psi}] \in \mathbb{R}^{n\times(K+1)}, \quad \mathbf{A}_\ell = \begin{bmatrix} \alpha_{\ell,2}\mathbf{t}\mathbf{t}^\mathsf{T} + \alpha_{\ell,3}\mathbf{T} & \alpha_{\ell,2}\mathbf{t} \\ \alpha_{\ell,2}\mathbf{t}^\mathsf{T} & \alpha_{\ell,2} \end{bmatrix} \in \mathbb{R}^{(K+1)\times(K+1)}, \tag{9}$$

for class label vectors $\mathbf{J} = [\mathbf{j}_1, \ldots, \mathbf{j}_K] \in \mathbb{R}^{n\times K}$, "second-order" data fluctuation vector $\boldsymbol{\psi} = \{\|\mathbf{x}_i - \boldsymbol{\mu}_a/\sqrt{p}\|^2 - \mathbb{E}[\|\mathbf{x}_i - \boldsymbol{\mu}_a/\sqrt{p}\|^2]\}_{i=1}^n \in \mathbb{R}^n$, second-order discriminative statistics $\mathbf{t} = \{\mathrm{tr}\,\mathbf{C}_a^\circ/\sqrt{p}\}_{a=1}^K \in \mathbb{R}^K$ and $\mathbf{T} = \{\mathrm{tr}\,\mathbf{C}_a\mathbf{C}_b/p\}_{a,b=1}^K \in \mathbb{R}^{K\times K}$ of the Gaussian mixture in (1)*, as well as non-negative scalars $\alpha_{\ell,1}, \alpha_{\ell,2}, \alpha_{\ell,3} \geq 0$ satisfying*

$$\alpha_{\ell,1} = \mathbb{E}[\sigma_\ell'(\tau_{\ell-1}\xi)]^2\alpha_{\ell-1,1}, \quad \alpha_{\ell,2} = \mathbb{E}[\sigma_\ell'(\tau_{\ell-1}\xi)]^2\alpha_{\ell-1,2} + \frac{1}{4}\mathbb{E}[\sigma_\ell''(\tau_{\ell-1}\xi)]^2\alpha_{\ell-1,4}^2, \tag{10}$$

$$\alpha_{\ell,3} = \mathbb{E}[\sigma_\ell'(\tau_{\ell-1}\xi)]^2\alpha_{\ell-1,3} + \frac{1}{2}\mathbb{E}[\sigma_\ell''(\tau_{\ell-1}\xi)]^2\alpha_{\ell-1,1}^2. \tag{11}$$

with $\alpha_{\ell,4} = \alpha_{\ell-1,4}\mathbb{E}\left[(\sigma_\ell'(\tau_{\ell-1}\xi))^2 + \sigma_\ell(\tau_{\ell-1}\xi)\sigma_\ell''(\tau_{\ell-1}\xi)\right]$ for $\xi \sim \mathcal{N}(0,1)$.

A few remarks on Theorem 1 are in order. The first remark is on the condition $\mathbb{E}[\sigma_\ell(\tau_{\ell-1}\xi)] = 0$.

Remark 1 (On activation centering)**.** *Note that the condition $\mathbb{E}[\sigma_\ell(\tau_{\ell-1}\xi)] = 0$, seemingly restrictive at first glance, in fact only subtracts an* identical constant *from all entries of the data representation $\boldsymbol{\Sigma}_\ell$ at layer ℓ as defined in* (3)*, and should* not *restrict the expressive power of the network under study, nor its performance on downstream ML tasks. For a given DNN model of interest, it suffices to "center" the output of each layer by subtracting a constant to satisfy Assumption 3, and to further apply our Theorem 1.*

Theorem 1 unveils the (possibly surprising) fact that, for the high-dimensional and non-trivial Gaussian mixture classification of (1), the spectral behavior of $\tilde{\mathbf{K}}_{\mathrm{CK},\ell}$ (and thus that of the CK matrix $\mathbf{K}_{\mathrm{CK},\ell}$) is (i) *independent* of the distribution of the (entries of the) weights $\mathbf{W}_\ell$ when they are "normalized" to have zero mean and unit variance, as demanded in Assumption 2, and (ii) depends on the activation function σ_ℓ *only* via four[†] scalar parameters $\alpha_{\ell,1}, \alpha_{\ell,2}, \alpha_{\ell,3}$ and τ_ℓ: such universal results have been previously observed in random matrix theory and high-dimensional statistics literature (see for example [10, 4, 53]) and indicate the wide applicability of our theoretical result.

On closer inspection of Theorem 1, we further observe that:

(i) for a given DNN, Theorem 1 characterizes, via the form of $\tilde{\mathbf{K}}_{\mathrm{CK},\ell}$ in (8) and the recursions in (10) and (11), how the linear (via $\alpha_{\ell,1}$, which is multiplied by $\mathbf{X}^\mathsf{T}\mathbf{X}$) and nonlinear (via $\alpha_{\ell,2}$ and $\alpha_{\ell,3}$ in $\mathbf{A}_\ell$, which respectively weight the second-order data statistics $\mathbf{t}$ and $\mathbf{T}$) data features "propagate," in a layer-by-layer fashion, as ℓ increases, as *quantitatively* measured by the corresponding α_ℓs; and

[†]It is worth noting that the parameter τ_ℓ appears in the CK eigenspectrum *only* by shifting all its eigenvalues (by τ_ℓ^2), thereby acting as an (implicit) ridge-type regularization in large-scale DNN models, see also [26, 13, 36].

(ii) for two DNNs with the same number of layers, but possibly different weights and activations, given the same input data $\mathbf{X}$ (so that the two nets have the same $\mathbf{K}_{\mathrm{CK},0}$), if they have asymptotically equivalent CK matrices $\mathbf{K}_{\mathrm{CK},\ell-1}$ at layer $\ell-1$ with the *same* $\alpha_{\ell-1,1}$, $\alpha_{\ell-1,2}$ and $\alpha_{\ell-1,3}$, then having the *same* $\mathbb{E}[\sigma_\ell'(\tau_{\ell-1}\xi)]^2$, $\mathbb{E}[\sigma_\ell''(\tau_{\ell-1}\xi)]^2$, and $\mathbb{E}[(\sigma_\ell^2(\tau_{\ell-1}\xi))'']$ (which *only* depends on the activation σ_ℓ of layer ℓ and $\tau_{\ell-1}$) will suffice for the two nets to have equivalent $\mathbf{K}_{\mathrm{CK},\ell}$ at layer ℓ.

It follows from the above item (ii) that for a given DNN of depth L, it is possible to design a novel DNN model that "matches" the original one – in the sense that both models will have asymptotically equivalent CK matrices *at each layer*, by using the following layer-by-layer matching strategy: Starting from the same $\mathbf{K}_{\mathrm{CK},0} = \mathbf{X}^\mathsf{T}\mathbf{X}$, one chooses the weights $\mathbf{W}_1$ of the novel DNN according to Assumption 2, and then select the first-layer activation σ_1 in such a way that the novel net has the same parameters $\alpha_{1,1}$, $\alpha_{1,2}$ and $\alpha_{1,3}$ as the original one, so that the first-layer CK matrices $\mathbf{K}_{\mathrm{CK},1}$ of the two nets are *spectrally* matched as per Theorem 1; one then proceeds similarly to match the second, the third, etc., and eventually the Lth layer of the two nets. As we shall see below, this layer-by-layer matching strategy facilitates the "lossless" compression of a given DNN.

Using the relation in (6), a similar result (as in Theorem 1 for CKs) can be established for NTK matrices, as shown in the following theorem. The proof is provided in Section A.2 of the appendix.

Theorem 2 (Asymptotic equivalent for NTK matrices)**.** *Let Assumptions 1–3 hold, let $\sigma_\ell(\cdot)$s be centered so that $\mathbb{E}[\sigma_\ell(\tau_{\ell-1}\xi)] = 0$ for τ_ℓs defined in* (7)*, and let $\kappa_1, \ldots, \kappa_L \geq 0$ be a sequence of non-negative numbers satisfying the following recursion:*

$$\kappa_\ell = \sqrt{\tau_\ell^2 + \kappa_{\ell-1}^2 \mathbb{E}[(\sigma_{\ell-1}'(\tau_{\ell-2}\xi))^2]}, \quad \xi \sim \mathcal{N}(0,1), \quad \ell \in \{1, \ldots, L\}. \tag{12}$$

Then, for the NTK matrix $\mathbf{K}_{\mathrm{NTK},\ell}$ of layer ℓ defined in (6)*, as $n, p \to \infty$ one has that*

$$\|\mathbf{K}_{\mathrm{NTK},\ell} - \tilde{\mathbf{K}}_{\mathrm{NTK},\ell}\| \to 0, \quad \tilde{\mathbf{K}}_{\mathrm{NTK},\ell} \equiv \beta_{\ell,1}\mathbf{X}^\mathsf{T}\mathbf{X} + \mathbf{V}\mathbf{B}_\ell\mathbf{V}^\mathsf{T} + (\kappa_\ell^2 - \tau_0^2\beta_{\ell,1} - \tau_0^4\beta_{\ell,3})\mathbf{I}_n, \tag{13}$$

almost surely, with $\mathbf{V} \in \mathbb{R}^{n\times(K+1)}$, $\mathbf{t} \in \mathbb{R}^K$, $\mathbf{T} \in \mathbb{R}^{K\times K}$ as defined in Theorem 1, and

$$\mathbf{B}_\ell \equiv \begin{bmatrix} \beta_{\ell,2}\mathbf{t}\mathbf{t}^\mathsf{T} + \beta_{\ell,3}\mathbf{T} & \beta_{\ell,2}\mathbf{t} \\ \beta_{\ell,2}\mathbf{t}^\mathsf{T} & \beta_{\ell,2} \end{bmatrix}, \tag{14}$$

as well as the non-negative constants $\beta_{\ell,1}, \beta_{\ell,2}, \beta_{\ell,3}, \beta_{\ell,4} \geq 0$, such that

$$\beta_{\ell,1} = \alpha_{\ell,1}, \quad \beta_{\ell,2} = \alpha_{\ell,2}, \quad \beta_{\ell,3} = \alpha_{\ell,3} + \beta_{\ell-1,1}\beta_{\ell,4}, \tag{15}$$

and $\beta_{\ell,4}$ satisfies the recursion $\beta_{\ell,4} = \mathbb{E}[\sigma_\ell''(\tau_{\ell-1}\xi)]^2\beta_{\ell-1,4}$.

Roughly speaking, Theorem 2 shows that a similar eigenspectral behavior established in Theorem 1 for CK matrices also holds for NTK matrices if we change the associated coefficients α_ℓs to β_ℓs. The remarks after Theorem 1 thus remain valid, at least in spirit, for NTK matrices.

Remark 2 (On spectral norm characterization)**.** *Note that the characterizations in Theorem 1 and 2 for CK and NTK matrices are provided in a spectral norm sense. It then follows from Weyl's inequality [21] and the Davis–Kahan theorem [57] that the difference between the eigenvalues (e.g., when listed in a decreasing order) and the associated eigenvectors (when the eigenvalues under study are "isolated") of $\mathbf{K}_{\mathrm{NTK}}$ and $\tilde{\mathbf{K}}_{\mathrm{NTK}}$ vanish asymptotically as $n, p \to \infty$. As such, the spectral norm guarantees in Theorem 1 and 2 provide more tractable access to the convergence and generalization properties of wide DNNs, at least for GMM data, via the spectral study of $\tilde{\mathbf{K}}_{\mathrm{NTK}}$ [25, 16].*

Despite being derived here for the Gaussian mixture model in (1), we conjecture that the results in Theorem 1 and 2 hold true beyond the Gaussian setting and can be extended, for example, to the family of concentrated random vectors [47, 30]. As previously discussed after Assumption 2 for the distribution of $\mathbf{W}$, universality commonly arises in random matrix theory and high-dimensional statistics [10, 4, 53]; we refer interested readers to Remark 4 in Appendix A for further discussions.

From a technical perspective, the results in Theorem 1 and 2 extend the single-hidden-layer CK analysis in [2, 33] to both CK and NTK matrices of fully-connected DNNs with an arbitrary number of layers. In particular, taking $\ell = 1$ in Theorem 1, one obtains [2, Theorem 1] as a special case.[†]

[†]Note that in [2, Theorem 1], the authors do *not* assume $\mathbb{E}[\sigma(\tau_0\xi)] = 0$ as in our Theorem 1, but instead "center" the CK matrices by pre- and post-multiplying $\mathbf{K}_{\mathrm{CK}}$ with $\mathbf{P} = \mathbf{I}_n - \mathbf{1}_n\mathbf{1}_n^\mathsf{T}/n$. This can be shown equivalent to taking $\mathbb{E}[\sigma(\tau_0\xi)] = 0$ in the single-hidden-layer setting; see Appendix A.3 for more discussions.

Remark 3 (On the equivalent CK and NTK matrices for two neural nets)**.** *It follows from Theorem 1 that for a given DNN model, it suffices to "match" the coefficients $\alpha_{\ell,1}, \alpha_{\ell,2}$ and $\alpha_{\ell,3}$ in a layer-by-layer manner to design a novel DNN with asymptotically equivalent CK. In doing so, one can choose the activation σ_ℓ at layer ℓ to satisfy a system of equations involving $\mathbb{E}[\sigma'_\ell(\tau_{\ell-1}\xi)]^2$, $\mathbb{E}[\sigma''_\ell(\tau_{\ell-1}\xi)]^2$, and $\mathbb{E}\left[(\sigma'_\ell(\tau_{\ell-1}\xi))^2 + \sigma_\ell(\tau_{\ell-1}\xi)\sigma''_\ell(\tau_{\ell-1}\xi)\right]$. Interestingly, it can then be checked from Theorem 2 that, by performing the above matching, if one has equivalent NTK matrices $\mathbf{K}_{\mathrm{NTK},\ell-1}$ at layer $\ell-1$ for the two nets under consideration, then the corresponding $\mathbf{K}_{\mathrm{NTK},\ell}$ at layer ℓ are also "matched" in a spectral sense, since the (additional) parameter $\mathbb{E}[\sigma''_\ell(\tau_{\ell-1}\xi)]^2$ in the recursion of $\beta_{\ell,4}$ in Theorem 2 has already been matched. Also, similarly to Footnote † for CK, matching (or nor) the parameter κ_ℓ only affects the level of regularization by "shifting" all eigenvalues of $\mathbf{K}_{\mathrm{NTK}}$.*

In the following corollary, we present a concrete example of how to apply the results in Theorem 1 and 2 in the design of a novel computationally and storage efficient DNN, that shares the same CK and NTK eigenspectra with any given fully-connected neural net having centered activation.

Corollary 1 (Sparse and quantized DNNs)**.** *For a given fully-connected DNN (referred to as* DNN1*) of depth L with centered activation such that $\mathbb{E}[\sigma_\ell(\tau_{\ell-1}\xi)] = 0$ for $\xi \sim \mathcal{N}(0,1)$, one is able to construct (again in a layer-by-layer manner) a sparse and quantized "equivalent" DNN model, of depth L and referred to as* DNN2*, such that the two nets have asymptotically the same eigenspectra for their CK (and thus NTK, as per Remark 3) matrices, by using the following ternary weights:*

$$[\mathbf{W}]_{ij} = 0 \text{ with proba } \varepsilon \in [0,1), \quad [\mathbf{W}]_{ij} = \pm(1-\varepsilon)^{-1/2} \text{ each with proba } 1/2 - \varepsilon/2, \tag{16}$$

as well as quantized activations (as visually displayed in Figure 1):

$$\sigma_T(t) = a \cdot (1_{t<s_1} + 1_{t>s_2}), \quad \sigma_Q(t) = b_1 \cdot (1_{t<r_1} + 1_{t>r_4}) + b_2 \cdot 1_{r_2 \le t \le r_3}. \tag{17}$$

We refer readers to Section A.4 in the appendix for the proof and discussions of Corollary 1, as well as the detailed expressions of $\mathbb{E}[\sigma'(\tau\xi)]$, $\mathbb{E}[\sigma''(\tau\xi)]$, and $\mathbb{E}[(\sigma^2(\tau\xi))'']$, of direct algorithmic use for both σ_T and σ_Q as functions of a, s_1, s_2 and $b_1, b_2, r_1, r_2, r_3, r_4$. The proposed NTK-LC compression approach based on Corollary 1, is described in detail below.

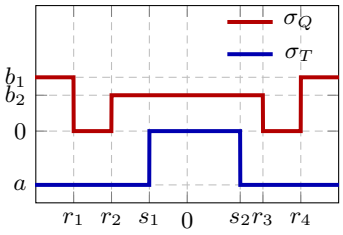
Algorithm 1 NTK-based "lossless" compression (NTK-LC)

Input: Input data $\mathbf{x}_1, \ldots, \mathbf{x}_n$, sparsity level $\varepsilon \in [0,1)$, and DNN1 with $\sigma_1, \ldots, \sigma_L$ activations
Output: Sparse and quantized DNN2 model with weights $\mathbf{W}_\ell$s and activations $\tilde{\sigma}_\ell$s.

Estimate τ_0 from input data as $\tau_0 = \sqrt{\frac{1}{n}\sum_{i=1}^n \|\mathbf{x}_i\|^2}$. Set $\tau = \tau_0$, $\tilde{\tau} = \tau_0$.
for $\ell = 1, \ldots, L-1$ **do**
  Compute $\mathbb{E}[\sigma'_\ell(\tau\xi)]^2$, $\mathbb{E}[\sigma''_\ell(\tau\xi)]^2$, and $\mathbb{E}[(\sigma^2_\ell(\tau\xi))'']$, and use them to solve for the coefficients a, s_1, s_2 in the novel activation $\tilde{\sigma}_\ell$, with $\tilde{\tau}$, Corollary 1, and expressions in Appendix A.4.
  Set $\tau = \sqrt{\mathbb{E}[\sigma^2_\ell(\tau\xi)]}$, $\tilde{\tau} = \sqrt{\mathbb{E}[\sigma^2_\ell(\tilde{\tau}\xi)]}$.
end for
For layer $\ell = L$, compute $\sqrt{\mathbb{E}[\sigma^2_L(\tau\xi)]}$, $\mathbb{E}[\sigma'_L(\tau\xi)]^2$, $\mathbb{E}[\sigma''_L(\tau\xi)]^2$, and $\mathbb{E}[(\sigma^2_L(\tau\xi))'']$, and use them to solve for the coefficients $b_1, b_2, r_1, r_2, r_3, r_4$ in the novel activation $\tilde{\sigma}_L$, with $\tilde{\tau}$, Corollary 1 and detailed expressions in Appendix A.4.
Draw independently the i.i.d. entries of $\mathbf{W}_1, \ldots, \mathbf{W}_L$ according to (16) with sparsity level ε.
return DNN2 model with weights $\mathbf{W}_\ell$ and activations $\tilde{\sigma}_\ell$, $\ell = 1, \ldots, L$.

Note that if a given layer (say the ℓth) of the original network DNN1 uses an odd (or even) activation function, then the corresponding layer in the sparse and quantized network DNN2 *must* also have odd (or even, respectively) activation, as a consequence of the fact that some of $\mathbb{E}[\sigma'_\ell(\xi)], \mathbb{E}[\sigma''_\ell(\xi)], \mathbb{E}[(\sigma^2_\ell(\xi))'']$ must be zero, depending on whether σ_ℓ is even or odd. Further note that, the "sign" of activations does not really matter in the design of computationally efficient DNNs described in Algorithm 1 above, in the sense that the key parameters $\alpha_{\ell,1}, \alpha_{\ell,2}$ and $\alpha_{\ell,3}$ for CKs, as well as $\beta_{\ell,1}, \beta_{\ell,2}$ and $\beta_{\ell,3}$ for NTKs, remain unchanged when $-\sigma_\ell(t)$ is used instead of $\sigma_\ell(t)$.

Before embarking on the detailed numerical experiments in Section 4, we would like to bring the readers' attention to the recent line of works [32, 50, 54, 51, 17] showing that for wide and deep NN models, very efficient sparse sub-networks can be found that almost match the performance of the original dense nets *with little or even no training*, for instance by uniformly pruning the network weights [50]. To develop a theoretical grasp of these (extremely counterintuitive) empirical successes,



$$\mathbb{E}[\sigma_T(\tau\xi)] = \frac{a}{2}\left(\operatorname{erf}\left(\frac{s_1}{\sqrt{2\tau}}\right) - \operatorname{erf}\left(\frac{s_2}{\sqrt{2\tau}}\right)\right) + a$$

$$\mathbb{E}[\sigma_Q(\tau\xi)] = \frac{b_1}{2}\left(\operatorname{erf}\left(\frac{r_1}{\sqrt{2\tau}}\right) - \operatorname{erf}\left(\frac{r_4}{\sqrt{2\tau}}\right)\right) + b_1 + \frac{b_2}{2}\left(\operatorname{erf}\left(\frac{r_3}{\sqrt{2\tau}}\right) - \operatorname{erf}\left(\frac{r_2}{\sqrt{2\tau}}\right)\right)$$

Figure 1: Visual representations of activations σ_T and σ_Q defined in (17) **(left)** and the detailed expressions of $\mathbb{E}[\sigma_T(\tau\xi)]$ and $\mathbb{E}[\sigma_Q(\tau\xi)]$**(right)**, with specifically $r_1 - r_2 = r_3 - r_4$ here.

a few attempts have been made, for example, to carefully prune the network weights to retain the same (limiting) NTK [37], or to show that randomly pruned sparse nets have the same (limiting) NTK as the original net up to a scaling factor [56]. Instead, our work, by considering the *statistical structure* of the input data, leverages tools from RMT to "compress" both the weights and activations (per Theorem 2 and Corollary 1) without affecting the NTK eigenstructure.

4 Numerical experiments

In this section, we provide numerical experiments to (i) validate the asymptotic characterizations in Theorem 1 and 2, on both synthetic GMM and real-world data (such as MNIST and CIFAR10) of (in fact not so) large sizes and dimensions; and to (ii) show how these results can be used to sparsify and quantize fully-connected DNNs, leading to huge savings in computational and storage resources (up to a factor of 10^3 in memory and a level of sparsity $\varepsilon - 90\%$) without significant performance degradation. We refer readers to Section B in the appendix for further experiments and discussions.

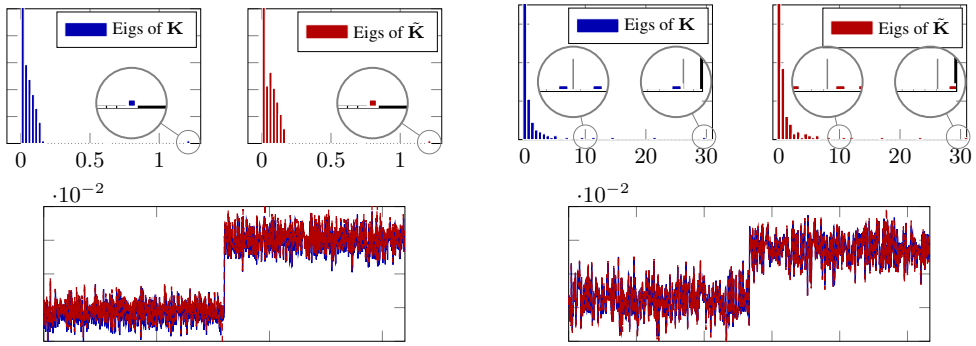


Figure 2: Eigenvalue histograms (**top**) and dominant eigenvectors (**bottom**) of last-layer CK matrices $\mathbf{K}_{\mathrm{CK}}$ (**blue**) defined in (4) (with the expectation estimated from 1000 independent realizations of $\mathbf{W}$s) and the asymptotic equivalent $\tilde{\mathbf{K}}_{\mathrm{CK}}$ (**red**) matrices. (**Left**) Gaussian $\mathbf{W}$ on two-class GMM data, with $p = 8\,000$, $n = 2\,000$, $\boldsymbol{\mu}_a = [\mathbf{0}_{8(a-1)};\ 8;\ \mathbf{0}_{p-8a+7}]$, $\mathbf{C}_a = (1 + 8(a-1)/\sqrt{p})\mathbf{I}_p$, $a \in \{1, 2\}$ using $[\mathrm{ReLU},\ \mathrm{ReLU},\ \mathrm{ReLU}]$ activations; and (**right**) symmetric Bernoulli $\mathbf{W}$ on MNIST data (number 6 versus 8) [28], with $p = 784$, $n = 3\,200$, using $[\mathrm{poly},\ \mathrm{ReLU},\ \mathrm{ReLU}]$ activations. $\mathbf{x}_1, \ldots, \mathbf{x}_{n/2} \in \mathcal{C}_1$ and $\mathbf{x}_{n/2+1}, \ldots, \mathbf{x}_n \in \mathcal{C}_2$ in both cases.

Figure 2 compares the eigenvalues and dominant eigenvectors of the CK matrices $\mathbf{K}_{\mathrm{CK}}$ defined in (4) versus those of their asymptotic approximations $\tilde{\mathbf{K}}_{\mathrm{CK}}$ given in Theorem 1, in the case of fully-connected DNNs having three hidden layers (of width $d_1 = 2\,000, d_2 = 2\,000, d_3 = 1\,000$ for each layer). For different types of activations ($\mathrm{poly}(t) = 0.2t^2 + t$ and $\mathrm{ReLU}(t) = \max(t, 0)$), different weight distributions (Gaussian and symmetric Bernoulli), and on synthetic GMM as well as MNIST data, we consistantly observe a close match between the eigenvalues and the dominant eigenvectors of $\mathbf{K}_{\mathrm{CK}}$ and $\tilde{\mathbf{K}}_{\mathrm{CK}}$, as a consequence of the spectral norm convergence in Theorem 1

(and Remark 2), suggesting a possibly wider applicability of the proposed results beyond GMM data.[†]

Figure 3 depicts the test accuracies of (i) the original dense and unquantized network with three fully-connected layers, (ii) the proposed NTK-LC approach described in Corollary 1 and Algorithm 1, and (iii) two "heuristic" compression approaches: (iii-i) sparsification by uniformly zeroing out 80% of the weights (we *cannot* do more, as the resultant performance is too poor to be compared with other curves in Figure 3), and (iii-ii) binarization using $\sigma(t) = 1_{t<-1} + 1_{t>1}$, for different choices of width per layer, and the ten-class classification problems of MNIST and CIFAR10. In particular, we observe that the proposed NTK-LC approach occupies (up to) a factor of 10^3 less memory, and produces significantly sparser networks (up to 90% of weights set to zero) with minimal performance loss, when compared to the original or the "heuristically" compressed nets.

In Figure 3, the neural networks before and after compression have three fully-connected layers, and the original network uses ReLU activations for all its three layers. The classification is performed on the output of a trainable classification layer that takes the features from the fully-connected layers as input. For the MNIST dataset, raw data are taken as the network input; for the CIFAR10 dataset, we take the flattened output of the 16th convolutional layer of VGG19 [49] as the input of the fully-connected layers (to be compressed).

We perform DNN quantization together with sparsification at levels of sparsity $\varepsilon \in \{0\%, 50\%, 90\%\}$. We see that the sparsity level ε has limited impact on the classification accuracy, which is in line with our theory. These experimental results show that the proposed NTK-LC approach achieves a better *performance-complexity trade-off* than commonly used heuristic DNN compression methods.

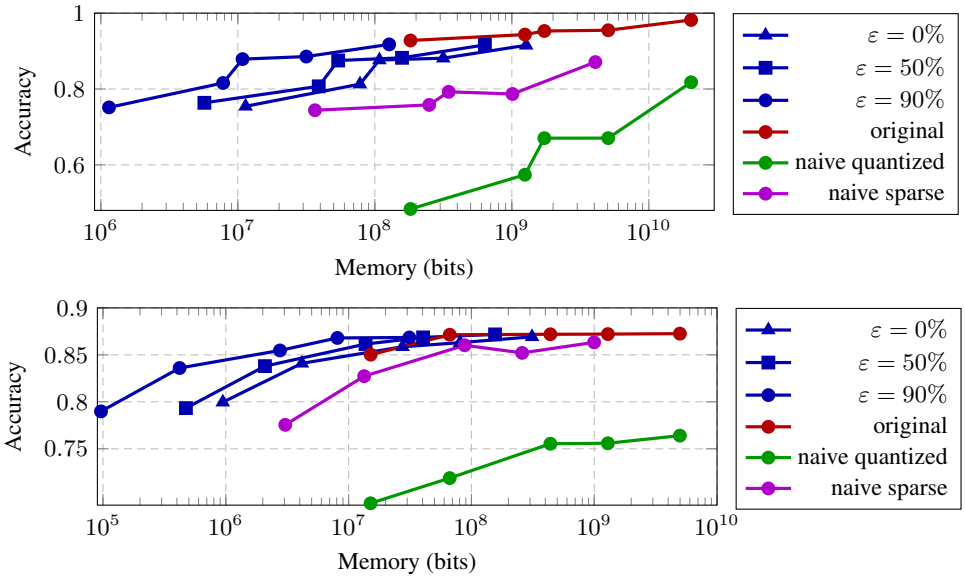


Figure 3: Test accuracy of classification on MNIST [28] (**top**) and CIFAR10 [27] (**bottom**) datasets. **Blue** curves represent the proposed NTK-LC approach with different levels of sparsity $\varepsilon \in \{0\%, 50\%, 90\%\}$, **purple** curves represent the heuristic sparsification approach by uniformly zeroing out 80% of the weights, **green** curves represent the heuristic quantization approach using the binary activation $\sigma(t) = 1_{t<-1} + 1_{t>1}$ (only applied to the first two layers, otherwise the performance is too poor to be compared to other curves), and **red** curves represent the original network. All nets have three fully-connected layers, and the original (dense and unquantized) network uses ReLU activations for all layers. Memory varies due to the **change of layer width** of the network.

[†]Zero eigenvalues of $\mathbf{K}_{\mathrm{CK}}, \tilde{\mathbf{K}}_{\mathrm{CK}}$ are removed from the top displays in Figure 2 for better visualization.

5 Conclusion and perspectives

In this paper, built upon recent advances in random matrix theory and high-dimensional statistics, we provide *precise* characterizations of the eigenspectra of both conjugate kernel and neural tangent kernel matrices, for high-dimensional Gaussian mixture data and fully-connected multi-layer neural nets. These results further allows us to sparsify and quantize fully-connected deep nets, resulting in a factor of 10^3 less memory consumption with virtually no performance degradation.

Acknowledgments and Disclosure of Funding

ZL would like to acknowledge the CCF-Hikvision Open Fund (20210008), the National Natural Science Foundation of China (NSFC-12141107), the Fundamental Research Funds for the Central Universities of China (2021XXJS110), the Key Research and Development Program of Hubei (2021BAA037), and the Key Research and Development Program of Guangxi (GuiKe-AB21196034) for providing partial support.

References

[1] Sina Alemohammad, Zichao Wang, Randall Balestriero, and Richard Baraniuk. The Recurrent Neural Tangent Kernel. *arXiv*, 2020.

[2] Hafiz Tiomoko Ali, Zhenyu Liao, and Romain Couillet. Random matrices in service of ML footprint: ternary random features with no performance loss. In *International Conference on Learning Representations*, 2022.

[3] Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang. On Exact Computation with an Infinitely Wide Neural Net. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[4] Zhidong Bai and Jack W. Silverstein. *Spectral Analysis of Large Dimensional Random Matrices*, volume 20 of *Springer Series in Statistics*. Springer-Verlag New York, 2 edition, 2010.

[5] Lucas Benigni and Sandrine Péché. Eigenvalue distribution of nonlinear models of random matrices. *arXiv*, 2019.

[6] Alberto Bietti and Julien Mairal. On the Inductive Bias of Neural Tangent Kernels. In *Advances in Neural Information Processing Systems*, volume 32 of *NIPS'19*, pages 12893–12904. Curran Associates, Inc., 2019.

[7] Avrim Blum, John Hopcroft, and Ravindran Kannan. *Foundations of Data Science*. Cambridge University Press, 2020.

[8] Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On Lazy Training in Differentiable Programming. In *Advances in Neural Information Processing Systems*, volume 32 of *NIPS'19*, pages 2937–2947. Curran Associates, Inc., 2019.

[9] Romain Couillet and Florent Benaych-Georges. Kernel spectral clustering of large dimensional data. *Electronic Journal of Statistics*, 10(1):1393–1454, 2016.

[10] Romain Couillet and Zhenyu Liao. *Random Matrix Methods for Machine Learning*. Cambridge University Press.

[11] Romain Couillet, Zhenyu Liao, and Xiaoyi Mai. Classification Asymptotics in the Random Matrix Regime. In *2018 26th European Signal Processing Conference (EUSIPCO)*, volume 00 of *EUSIPCO'18*, pages 1875–1879, 2018.

[12] By Lei Deng, Guoqi Li, Song Han, Luping Shi, and Yuan Xie. Model Compression and Hardware Acceleration for Neural Networks: A Comprehensive Survey. *Proceedings of the IEEE*, 108(4):485–532, 2020.

[13] Michal Derezinski, Feynman T Liang, and Michael W Mahoney. Exact expressions for double descent and implicit regularization via surrogate random design. In *Advances in Neural Information Processing Systems*, volume 33, pages 5152–5164. Curran Associates, Inc., 2020.

[14] Simon S Du, Kangcheng Hou, Russ R Salakhutdinov, Barnabas Poczos, Ruosong Wang, and Keyulu Xu. Graph Neural Tangent Kernel: Fusing Graph Neural Networks with Graph Kernels. 32.

[15] Zhou Fan and Andrea Montanari. The spectral norm of random inner-product kernel matrices. *Probability Theory and Related Fields*, 173(1-2):27–85, 2019.

[16] Zhou Fan and Zhichao Wang. Spectra of the Conjugate Kernel and Neural Tangent Kernel for linear-width neural networks. In *Advances in Neural Information Processing Systems*, volume 33, pages 7710–7721. Curran Associates, Inc., 2020.

[17] Jonathan Frankle, Gintare Karolina Dziugaite, Daniel Roy, and Michael Carbin. Pruning neural networks at initialization: Why are we missing the mark? In *International Conference on Learning Representations*, 2021.

[18] J. Han and C. Moraga. The influence of the sigmoid function parameters on the speed of backpropagation learning. 1995.

[19] Song Han, Huizi Mao, and William J Dally. Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding. *arXiv*, 2015.

[20] Torsten Hoefler, Dan Alistarh, Tal Ben-Nun, Nikoli Dryden, and Alexandra Peste. Sparsity in deep learning: Pruning and growth for efficient inference and training in neural networks. *Journal of Machine Learning Research*, 22(241):1–124, 2021.

[21] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 2 edition, 2012.

[22] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv*, 2017.

[23] Jiaoyang Huang and Horng-Tzer Yau. Dynamics of Deep Neural Networks and Neural Tangent Hierarchy. 119:4542–4551, 2020-13.

[24] Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Binarized Neural Networks. In *Advances in Neural Information Processing Systems*, volume 29 of *NIPS'16*, pages 4107–4115. Curran Associates, Inc., 2016.

[25] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural Tangent Kernel: Convergence and Generalization in Neural Networks. In *Advances in Neural Information Processing Systems*, volume 31 of *NIPS'18*, pages 8571–8580. Curran Associates, Inc., 2018.

[26] Arthur Jacot, Berfin Simsek, Francesco Spadaro, Clement Hongler, and Franck Gabriel. Implicit regularization of random feature models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4631–4640. PMLR, 13–18 Jul 2020.

[27] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[28] Yann LeCun, Leon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[29] Yann LeCun, John Denker, and Sara Solla. Optimal Brain Damage. In *Advances in Neural Information Processing Systems*, volume 2 of *NIPS'90*, pages 598–605. Morgan-Kaufmann, 1990.

[30] Michel Ledoux. *The Concentration of Measure Phenomenon*. Mathematical Surveys and Monographs. 2005.

[31] Jaehoon Lee, Jascha Sohl-dickstein, Jeffrey Pennington, Roman Novak, Sam Schoenholz, and Yasaman Bahri. Deep Neural Networks as Gaussian Processes. In *International Conference on Learning Representations*, 2018.

[32] Namhoon Lee, Thalaiyasingam Ajanthan, and Philip Torr. SNIP: Single-shot network pruning based on connection sensitivity. In *International Conference on Learning Representations*, 2018.

[33] Zhenyu Liao and Romain Couillet. On the Spectrum of Random Features Maps of High Dimensional Data. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3063–3071, Stockholmsmässan, Stockholm Sweden, 2018. PMLR.

[34] Zhenyu Liao and Romain Couillet. The Dynamics of Learning: A Random Matrix Approach. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3072–3081, Stockholmsmässan, Stockholm Sweden, 2018. PMLR.

[35] Chaoyue Liu, Libin Zhu, and Misha Belkin. On the linearity of large non-linear models: when and why the tangent kernel is constant. 33:15954–15964.

[36] Fanghui Liu, Zhenyu Liao, and Johan Suykens. Kernel regression in high dimensions: Refined analysis beyond double descent. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 649–657. PMLR, 13–15 Apr 2021.

[37] Tianlin Liu and Friedemann Zenke. Finding trainable sparse networks through neural tangent transfer. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6336–6347. PMLR, 13–18 Jul 2020.

[38] Cosme Louart and Romain Couillet. Concentration of Measure and Large Random Matrices with an application to Sample Covariance Matrices. *arXiv*, 2018.

[39] Andrew L. Maas, Awni Y. Hannun, and Andrew Y. Ng. Rectifier Nonlinearities Improve Neural Network Acoustic Models. In *ICML Workshop on Deep Learning for Audio, Speech and Language Processing*, ICML Workshop, page 3, 2013.

[40] Vladimir A Marcenko and Leonid Andreevich Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457, 1967.

[41] Charles H. Martin and Michael W. Mahoney. Traditional and Heavy Tailed Self Regularization in Neural Network Models. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4284–4293. PMLR, 2019.

[42] Charles H. Martin and Michael W. Mahoney. Heavy-Tailed Universality Predicts Trends in Test Accuracies for Very Large Pre-Trained Deep Neural Networks. In *Proceedings of the 2020 SIAM International Conference on Data Mining (SDM)*, SDM'20, pages 505–513. SIAM, 2020.

[43] Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv*, 2019.

[44] Leonid Pastur. On Random Matrices Arising in Deep Neural Networks. Gaussian Case. *arXiv*, 2020.

[45] Jeffrey Pennington and Pratik Worah. Nonlinear random matrix theory for deep learning. In *Advances in Neural Information Processing Systems*, volume 30 of *NIPS'17*, pages 2637–2646. Curran Associates, Inc., 2017.

[46] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, 2015.

[47] Mohamed El Amine Seddik, Cosme Louart, Mohamed Tamaazousti, and Romain Couillet. Random Matrix Theory Proves that Deep Learning Representations of GAN-data Behave as Gaussian Mixtures. In *Proceedings of the 37th International Conference on Machine Learning*, Proceedings of Machine Learning Research, pages 8573–8582. PMLR, 2020.

[48] Mohamed El Amine Seddik, Mohamed Tamaazousti, and Romain Couillet. A Kernel Random Matrix-Based Approach for Sparse PCA. In *International Conference on Learning Representations*, ICLR'19, 2019.

[49] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, ICLR'14, 2014.

[50] Jingtong Su, Yihang Chen, Tianle Cai, Tianhao Wu, Ruiqi Gao, Liwei Wang, and Jason D Lee. Sanity-Checking Pruning Methods: Random Tickets can Win the Jackpot. In *Advances in Neural Information Processing Systems*, volume 33, pages 20390–20401. Curran Associates, Inc.

[51] Hidenori Tanaka, Daniel Kunin, Daniel L Yamins, and Surya Ganguli. Pruning neural networks without any data by iteratively conserving synaptic flow. In *Advances in Neural Information Processing Systems*, volume 33, pages 6377–6389. Curran Associates, Inc., 2020.

[52] Terence Tao and Van Vu. Random matrices: the circular law. *Communications in Contemporary Mathematics*, 10(02):261–307, 2008.

[53] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

[54] Chaoqi Wang, Guodong Zhang, and Roger Grosse. Picking winning tickets before training by preserving gradient flow. In *International Conference on Learning Representations*, 2020.

[55] Eugene P. Wigner. Characteristic Vectors of Bordered Matrices with Infinite Dimensions. *The Annals of Mathematics*, 62(3):548, 1955.

[56] Hongru Yang and Zhangyang Wang. On the Neural Tangent Kernel Analysis of Randomly Pruned Wide Neural Networks. *arXiv*, 2022.

[57] Y. Yu, T. Wang, and R. J. Samworth. A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.

Checklist

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? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]

(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]
   (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? [No]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (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? [No]
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]

Supplementary Material

Lossless Compression of Deep Neural Networks:
A High-dimensional Neural Tangent Kernel Approach

A Proofs and auxiliary results

A.1 Proof of Theorem 1

In this section, we provide the detailed proof of Theorem 1. Before going into details of the proof, we first recall our system model and working assumptions as follow.

We consider n data vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ independently drawn from one of the K-class Gaussian mixtures $\mathcal{C}_1, \ldots, \mathcal{C}_K$ and denote $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p \times n}$, with class $\mathcal{C}_a$ having cardinality n_a; that is, for $\mathbf{x}_i \in \mathcal{C}_a$ we have

$$\mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}_a/\sqrt{p}, \mathbf{C}_a/p), \tag{18}$$

for mean vector $\boldsymbol{\mu}_a \in \mathbb{R}^p$ and non-negative definite covariance $\mathbf{C}_a \in \mathbb{R}^{p \times p}$ associated with class $\mathcal{C}_a$.

We positive ourselves in the high-dimensional and non-trivial classification regime as stated in Assumption 1, that is: As $n \to \infty$, we have, for $a \in \{1, \ldots, K\}$ that,

(i) $p/n \to c \in (0, \infty)$ and $n_a/n \to c_a \in [0, 1)$; and

(ii) $\|\boldsymbol{\mu}_a\| = O(1)$; and

(iii) for $\mathbf{C}^\circ \equiv \sum_{a=1}^K \frac{n_a}{n} \mathbf{C}_a$ and $\mathbf{C}_a^\circ \equiv \mathbf{C}_a - \mathbf{C}^\circ$, we have $\|\mathbf{C}_a\| = O(1)$, $\operatorname{tr} \mathbf{C}_a^\circ = O(\sqrt{p})$ and $\operatorname{tr}(\mathbf{C}_a \mathbf{C}_b) = O(p)$ for $a, b \in \{1, \ldots, K\}$; and

(iv) $\tau_0 \equiv \sqrt{\operatorname{tr} \mathbf{C}^\circ / p}$ converges in $(0, \infty)$.

Remark 4 (Beyond Gaussian mixture data)**.** *Despite derived here for Gaussian mixture data, we conjecture that our results hold more generally beyond the Gaussian setting. As concrete examples, many results in random matrix theory and high dimensional statistics such as the popular Marčenko-Pastur [40], the semicircular [55], as well as the circular laws [52], have all been shown* universal *in the sense that they do* not *depend on the distribution of the (independent entries of the) data, as long as they are normalized to have zero mean and unit variance. In a machine learning context, such* universal *behavior are observed to hold beyond the above models, and extends to* nonlinear *model such as kernel matrices [48] and neural nets [47], in the sense that for data drawn from the family of concentrated random vectors [30, 38] (so* not *necessarily Gaussian), the performance on those ML models are the* same*, in the lager n, p setting,* as if they were mere Gaussian mixtures *with the same means and covariances. We refer the interested readers to [10, Chapter 8] for more discussions on this point.*

We consider the fully-connected neural network model of depth L and of successive widths $d_1, \ldots, d_L$ as defined in (2), and denote $\mathbf{W}_\ell \in \mathbb{R}^{d_\ell \times d_{\ell-1}}$ as well as $\sigma_\ell(\cdot)$ the weight matrix and activation at layer $\ell \in \{1, \ldots, L\}$, respectively.

In this section, we focus on the Conjugate Kernel (CK) matrix defined via the following recursive relation [25, 6]

$$[\mathbf{K}_{\mathrm{CK},\ell}]_{ij} = \mathbb{E}_{u,v}[\sigma_\ell(u)\sigma_\ell(v)], \quad \mathbf{K}_{\mathrm{CK},0} = \mathbf{X}^\mathsf{T}\mathbf{X}, \tag{19}$$

with

$$u, v \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} [\mathbf{K}_{\mathrm{CK},\ell-1}]_{ii} & [\mathbf{K}_{\mathrm{CK},\ell-1}]_{ij} \\ [\mathbf{K}_{\mathrm{CK},\ell-1}]_{ij} & [\mathbf{K}_{\mathrm{CK},\ell-1}]_{jj} \end{bmatrix}\right). \tag{20}$$

The assessment of the closely related neural tangent kernel (NTK) matrix is given in Section A.2.

We assume the following conditions hold for the random weight matrices $\mathbf{W}_\ell$s and the activation σ_ℓs for $\ell \in \{1, \ldots, L\}$, as demanded in Assumption 2 and 3:

(i) The weight matrices $\mathbf{W}_\ell$s are independent and have i.i.d. entries of zero mean, unit variance, and finite fourth-order moment.

(ii) The activations σ_ℓs are at least four-times differentiable with respect to standard normal distribution, in the sense that $\max_{k\in\{0,1,2,3,4\}}\{|\mathbb{E}[\sigma_\ell^{(k)}(\xi)]|\}$ is finite for $\xi\sim\mathcal{N}(0,1)$.

Let Assumptions 1–3 hold, and let $\tau_0,\tau_1,\ldots,\tau_L\geq 0$ be a sequence of non-negative numbers recursively defined via

$$\tau_\ell=\sqrt{\mathbb{E}[\sigma_\ell^2(\tau_{\ell-1}\xi)]} \tag{21}$$

as in (7), and assume the activation functions $\sigma_\ell(\cdot)$s are "centered" such that $\mathbb{E}[\sigma_\ell(\tau_{\ell-1}\xi)]=0$. (This assumption, as we shall see, plays a central role in our proof.)

Then, to prove Theorem 1 it suffices to show that,

(i) the CK matrix $\mathbf{K}_{\mathrm{CK},\ell}$ of layer $\ell\in\{1,\ldots,L\}$ defined in (4) satisfies

$$\|\mathbf{K}_{\mathrm{CK},\ell}-\tilde{\mathbf{K}}_{\mathrm{CK},\ell}\|\to 0, \tag{22}$$

almost surely as $n,p\to\infty$, with $\tilde{\mathbf{K}}_{\mathrm{CK},\ell}$ taking the ("unified") form

$$\tilde{\mathbf{K}}_{\mathrm{CK},\ell}\equiv\alpha_{\ell,1}\mathbf{X}^\mathsf{T}\mathbf{X}+\mathbf{V}\mathbf{A}_\ell\mathbf{V}^\mathsf{T}+(\tau_\ell^2-\tau_0^2\alpha_{\ell,1}-\tau_0^4\alpha_{\ell,3})\mathbf{I}_n, \tag{23}$$

for all $\ell\in\{1\ldots,L\}$, $\mathbf{J}=[\mathbf{j}_1,\ldots,\mathbf{j}_K]\in\mathbb{R}^{n\times K}$, random vector $\boldsymbol{\psi}=\{\|\mathbf{x}_i-\boldsymbol{\mu}_a/\sqrt{p}\|^2-\mathbb{E}[\|\mathbf{x}_i-\boldsymbol{\mu}_a/\sqrt{p}\|^2]\}_{i=1}^n\in\mathbb{R}^n$, $\mathbf{t}=\{\operatorname{tr}\mathbf{C}_a^\circ/\sqrt{p}\}_{a=1}^K\in\mathbb{R}^K$, $\mathbf{T}=\{\operatorname{tr}\mathbf{C}_a\mathbf{C}_b/p\}_{a,b=1}^K\in\mathbb{R}^{K\times K}$, and

$$\mathbf{V}=[\mathbf{J}/\sqrt{p},\ \boldsymbol{\psi}]\in\mathbb{R}^{n\times(K+1)},\quad \mathbf{A}_\ell=\begin{bmatrix}\alpha_{\ell,2}\mathbf{t}\mathbf{t}^\mathsf{T}+\alpha_{\ell,3}\mathbf{T} & \alpha_{\ell,2}\mathbf{t}\\ \alpha_{\ell,2}\mathbf{t}^\mathsf{T} & \alpha_{\ell,2}\end{bmatrix}; \tag{24}$$

(ii) the coefficients $\alpha_{\ell,1},\alpha_{\ell,2},\alpha_{\ell,3}$ are non-negative and satisfy

$$\alpha_{\ell,1}=\mathbb{E}[\sigma_\ell'(\tau_{\ell-1}\xi)]^2\alpha_{\ell-1,1},\quad \alpha_{\ell,2}=\mathbb{E}[\sigma_\ell'(\tau_{\ell-1}\xi)]^2\alpha_{\ell-1,2}+\frac{1}{4}\mathbb{E}[\sigma_\ell''(\tau_{\ell-1}\xi)]^2\alpha_{\ell-1,4}^2,$$
$$\alpha_{\ell,3}=\mathbb{E}[\sigma_\ell'(\tau_{\ell-1}\xi)]^2\alpha_{\ell-1,3}+\frac{1}{2}\mathbb{E}[\sigma_\ell''(\tau_{\ell-1}\xi)]^2\alpha_{\ell-1,1}^2.$$
with $\alpha_{\ell,4}=\alpha_{\ell-1,4}\mathbb{E}\left[(\sigma_\ell'(\tau_{\ell-1}\xi))^2+\sigma_\ell(\tau_{\ell-1}\xi)\sigma_\ell''(\tau_{\ell-1}\xi)\right]$ for $\xi\sim\mathcal{N}(0,1)$.

We prove the above results by induction on $\ell\in\{1,\ldots,L\}$: For $\ell=1$, we have $\mathbf{K}_{\mathrm{CK},0}=\mathbf{X}^\mathsf{T}\mathbf{X}$ so that

$$\mathbf{K}_{\mathrm{CK},0}=\tilde{\mathbf{K}}_{\mathrm{CK},0}=\mathbf{X}^\mathsf{T}\mathbf{X}, \tag{25}$$

with $\alpha_{1,1}=1$, $\alpha_{1,2}=0$, and $\alpha_{1,3}=0$.

We then assume $\|\mathbf{K}_{\mathrm{CK},\ell-1}-\tilde{\mathbf{K}}_{\mathrm{CK},\ell-1}\|\to 0$ holds at layer $\ell-1$ with

$$\tilde{\mathbf{K}}_{\mathrm{CK},\ell-1}\equiv\alpha_{\ell-1,1}\mathbf{X}^\mathsf{T}\mathbf{X}+\mathbf{V}\mathbf{A}_{\ell-1}\mathbf{V}^\mathsf{T}+(\tau_{\ell-1}^2-\tau_0^2\alpha_{\ell-1,1}-\tau_0^4\alpha_{\ell-1,3})\mathbf{I}_n,$$

for $\mathbf{A}_{\ell-1}=\begin{bmatrix}\alpha_{\ell-1,2}\mathbf{t}\mathbf{t}^\mathsf{T}+\alpha_{\ell-1,3}\mathbf{T} & \alpha_{\ell-1,2}\mathbf{t}\\ \alpha_{\ell-1,2}\mathbf{t}^\mathsf{T} & \alpha_{\ell-1,2}\end{bmatrix}$, and work on the CK matrix $\mathbf{K}_{\mathrm{CK},\ell}$ at layer ℓ.

We first introduce the following notations that will be consistently used in the proof: for $\mathbf{x}_i,\mathbf{x}_j\in\mathbb{R}^p$ with $i\neq j$, let

$$\mathbf{x}_i=\boldsymbol{\mu}_i/p+\mathbf{z}_i/p,\quad \mathbf{x}_j=\boldsymbol{\mu}_j/p+\mathbf{z}_j/p, \tag{26}$$

so that $\mathbf{z}_i\sim\mathcal{N}(\mathbf{0},\mathbf{C}_i)$, $\mathbf{z}_j\sim\mathcal{N}(\mathbf{0},\mathbf{C}_j)$, and

$$A_{ij}\equiv\underbrace{\frac{1}{p}\mathbf{z}_i^\mathsf{T}\mathbf{z}_j}_{O(p^{-1/2})}+\underbrace{\frac{1}{p}\boldsymbol{\mu}_i^\mathsf{T}\boldsymbol{\mu}_j+\frac{1}{p}(\boldsymbol{\mu}_i^\mathsf{T}\mathbf{z}_j+\boldsymbol{\mu}_j^\mathsf{T}\mathbf{z}_i)}_{O(p^{-1})},$$
$$t_i\equiv\frac{1}{p}\operatorname{tr}\mathbf{C}_i^\circ=O(p^{-1/2}),\quad \psi_i=\frac{1}{p}\|\mathbf{z}_i\|^2-\frac{1}{p}\operatorname{tr}\mathbf{C}_i=O(p^{-1/2}),$$
$$\tau_0\equiv\sqrt{\frac{1}{p}\operatorname{tr}\mathbf{C}^\circ}=O(1),$$
$$\chi_i\equiv\underbrace{t_i+\psi_i}_{O(p^{-1/2})}+\underbrace{\|\boldsymbol{\mu}_i\|^2/p+2\boldsymbol{\mu}_i^\mathsf{T}\mathbf{z}_i/p}_{O(p^{-1})}=\|\mathbf{x}_i\|^2-\tau_0,$$

where we note that the notations τ_0, ψ_i and t_i (with a slight abuse of notation to denote $\mathbf{C}_i = \mathbf{C}_a$ for $\mathbf{x}_i \in \mathcal{C}_a$) are in line with those defined in Assumption 1 and Theorem 1, and we denote S_{ij} terms of the form

$$S_{ij} = \alpha \frac{1}{p} \mathbf{z}_i^\mathsf{T} \mathbf{z}_j \left(\beta(t_i + \psi_i) + \gamma(t_j + \psi_j) \right), \tag{27}$$

for random or deterministic scalars $\alpha, \beta, \gamma = O(1)$ (with high probability when being random), we have $S_{ij} = O(p^{-1})$ and perhaps surprisingly, it leads to, in matrix form, a matrix of spectral norm order $O(p^{-1/2})$, see [9]. This spectral norm result will be exploited in the remainder of the proof.

By definition in (5), using the Gram-Schmidt orthogonalization procedure for standard Gaussian random variable as in [16, 2], we write

$$u = \sqrt{[\mathbf{K}_{\mathrm{CK},\ell-1}]_{ii}} \cdot \xi_i, \quad v = \frac{[\mathbf{K}_{\mathrm{CK},\ell-1}]_{ij}}{\sqrt{[\mathbf{K}_{\mathrm{CK},\ell-1}]_{ii}}} \cdot \xi_i + \sqrt{[\mathbf{K}_{\mathrm{CK},\ell-1}]_{jj} - \frac{[\mathbf{K}_{\mathrm{CK},\ell-1}]_{ij}^2}{[\mathbf{K}_{\mathrm{CK},\ell-1}]_{ii}}} \cdot \xi_j, \tag{28}$$

for *independent* $\xi_i, \xi_j \sim \mathcal{N}(0,1)$ so that at layer ℓ, we have

$$[\mathbf{K}_{\mathrm{CK},\ell}]_{ii} = \mathbb{E}\left[\sigma_\ell^2 \left(\sqrt{[\mathbf{K}_{\mathrm{CK},\ell-1}]_{ii}} \cdot \xi_i \right) \right] \tag{29}$$

$$\begin{aligned} [\mathbf{K}_{\mathrm{CK},\ell}]_{ij} = \mathbb{E}\Bigg[& \sigma_\ell \left(\sqrt{[\mathbf{K}_{\mathrm{CK},\ell-1}]_{ii}} \cdot \xi_i \right) \\ & \times \sigma_\ell \left(\frac{[\mathbf{K}_{\mathrm{CK},\ell-1}]_{ij}}{\sqrt{[\mathbf{K}_{\mathrm{CK},\ell-1}]_{ii}}} \cdot \xi_i + \sqrt{[\mathbf{K}_{\mathrm{CK},\ell-1}]_{jj} - \frac{[\mathbf{K}_{\mathrm{CK},\ell-1}]_{ij}^2}{[\mathbf{K}_{\mathrm{CK},\ell-1}]_{ii}}} \cdot \xi_j \right) \Bigg], \end{aligned} \tag{30}$$

where we note that the expectations are taken with respect to (the now *independent*) random variables ξ_i and ξ_j, so conditioned on the random vectors $\mathbf{x}_i$ and $\mathbf{x}_j$.

Based on the induction hypothesis on the layer $\ell - 1$, we have

$$[\mathbf{K}_{\mathrm{CK},\ell-1}]_{ij} = \alpha_{\ell-1,1} A_{ij} + \alpha_{\ell-1,2}(t_i + \psi_i)(t_j + \psi_j) + \alpha_{\ell-1,3} \left(\frac{1}{p} \mathbf{z}_i^\mathsf{T} \mathbf{z}_i \right)^2 + S_{ij} + O(p^{-3/2}), \tag{31}$$

for $i \neq j$, and

$$[\mathbf{K}_{\mathrm{CK},\ell-1}]_{ii} = \tau_{\ell-1}^2 + \alpha_{\ell-1,4} \chi_i + \alpha_{\ell-1,5}(t_i + \psi_i)^2 + O(p^{-3/2}). \tag{32}$$

The objective is then to derive the expression/approximation of $[\mathbf{K}_{\mathrm{CK},\ell}]_{ij}$ and $[\mathbf{K}_{\mathrm{CK},\ell}]_{ii}$ at layer ℓ, both to terms of order $O(p^{-3/2})$, and to subsequently derive the relation between the key coefficients of layer $\ell - 1$:

$$\{\alpha_{\ell-1,1}, \alpha_{\ell-1,2}, \alpha_{\ell-1,3}, \alpha_{\ell-1,4}, \alpha_{\ell-1,5}\}, \tag{33}$$

and those of layer ℓ

$$\{\alpha_{\ell,1}, \alpha_{\ell,2}, \alpha_{\ell,3}, \alpha_{\ell,4}, \alpha_{\ell,5}\}. \tag{34}$$

To this end, we first focus on the diagonal entries by evaluating $[\mathbf{K}_{\mathrm{CK},\ell}]_{ii}$, and then on the off-diagonal terms $[\mathbf{K}_{\mathrm{CK},\ell}]_{ij}$ for $i \neq j$, we conclude the proof by putting everything into matrix form.

On the diagonal. We start with the diagonal entries of $\mathbf{K}_{\mathrm{CK},\ell}$, which, as per its definition in (29), depends on the diagonal entries $[\mathbf{K}_{\mathrm{CK},\ell-1}]_{ii}$ at layer $\ell - 1$ as defined in (32), so that by Taylor-expanding $\sqrt{t}$ around $t \simeq \tau_{\ell-1}^2 = O(1)$, one gets

$$\begin{aligned} \sqrt{[\mathbf{K}_{\mathrm{CK},\ell-1}]_{ii}} &= \sqrt{\tau_{\ell-1}^2 + \alpha_{\ell-1,4}\chi_i + \alpha_{\ell-1,5}(t_i + \psi_i)^2 + O(p^{-3/2})} \\ &= \tau_{\ell-1} + \frac{1}{2\tau_{\ell-1}} \left(\alpha_{\ell-1,4}\chi_i + \alpha_{\ell-1,5}(t_i + \psi_i)^2 \right) - \frac{\alpha_{\ell-1,4}^2}{8\tau_{\ell-1}^3}(t_i + \psi_i)^2 + O(p^{-3/2}) \\ &= \tau_{\ell-1} + \frac{1}{2\tau_{\ell-1}} \alpha_{\ell-1,4}\chi_i + \frac{4\tau_{\ell-1}^2 \alpha_{\ell-1,5} - \alpha_{\ell-1,4}^2}{8\tau_{\ell-1}^3}(t_i + \psi_i)^2 + O(p^{-3/2}), \end{aligned}$$

and therefore by Taylor-expanding $\sigma_\ell^2(\sqrt{[\mathbf{K}_{\mathrm{CK},\ell-1}]_{ii}}\cdot\xi_i) = f(\sqrt{[\mathbf{K}_{\mathrm{CK},\ell-1}]_{ii}}\cdot\xi_i)$ around $\tau_{\ell-1}\xi_i$,

$$\begin{aligned}
[\mathbf{K}_{\mathrm{CK},\ell}]_{ii} &= \mathbb{E}\left[\sigma_\ell^2\left(\sqrt{[\mathbf{K}_{\mathrm{CK},\ell-1}]_{ii}}\cdot\xi\right)\right] = \mathbb{E}\left[f\left(\sqrt{[\mathbf{K}_{\mathrm{CK},\ell-1}]_{ii}}\cdot\xi\right)\right] \\
&= \mathbb{E}\left[f(\tau_{\ell-1}\xi) + f'(\tau_{\ell-1}\xi)\xi\left(\frac{1}{2\tau_{\ell-1}}\alpha_{\ell-1,4}\chi_i + \frac{4\tau_{\ell-1}^2\alpha_{\ell-1,5}-\alpha_{\ell-1,4}^2}{8\tau_{\ell-1}^3}(t_i+\psi_i)^2\right)\right] \\
&\quad + \mathbb{E}\left[\frac{1}{2}f''(\tau_{\ell-1}\xi)\xi^2\right]\frac{\alpha_{\ell-1,4}^2}{4\tau_{\ell-1}^2}(t_i+\psi_i)^2 + O(p^{-3/2}) \\
&= \mathbb{E}[f(\tau_{\ell-1}\xi)] + \mathbb{E}[f''(\tau_{\ell-1}\xi)]\left(\frac{1}{2}\alpha_{\ell-1,4}\chi_i + \frac{4\tau_{\ell-1}^2\alpha_{\ell-1,5}-\alpha_{\ell-1,4}^2}{8\tau_{\ell-1}^2}(t_i+\psi_i)^2\right) \\
&\quad + \mathbb{E}\left[\frac{1}{2}f''(\tau_{\ell-1}\xi)\xi^2\right]\frac{\alpha_{\ell-1,4}^2}{4\tau_{\ell-1}^2}(t_i+\psi_i)^2 + O(p^{-3/2}) \\
&= \mathbb{E}[f(\tau_{\ell-1}\xi)] + \frac{\alpha_{\ell-1,4}}{2}\mathbb{E}[f''(\tau_{\ell-1}\xi)]\chi_i + \frac{4\alpha_{\ell-1,5}+\alpha_{\ell-1,4}^2\mathbb{E}[f''''(\tau_{\ell-1}\xi)]}{8}(t_i+\psi_i)^2 \\
&\quad + O(p^{-3/2}),
\end{aligned}$$

where we denote the shortcut $f(x) = \sigma_\ell^2(x)$ and used the facts that

$$\mathbb{E}[f'(\tau_{\ell-1}\xi)\xi] = \tau_{\ell-1}\mathbb{E}[f''(\tau_{\ell-1}\xi)], \quad \mathbb{E}[f''(\tau_{\ell-1}\xi)(\xi^2-1)] = \tau_{\ell-1}^2\mathbb{E}[f''''(\tau_{\ell-1}\xi)], \tag{35}$$

for $\xi\sim\mathcal{N}(0,1)$, as a consequence of the Gaussian integration by parts formula.

As a consequence, we obtain the following relation

$$\begin{aligned}
\tau_\ell &= \sqrt{\mathbb{E}[\sigma_\ell^2(\tau_{\ell-1}\xi)]}, \quad \alpha_{\ell,4} = \alpha_{\ell-1,4}\mathbb{E}\left[(\sigma_\ell'(\tau_{\ell-1}\xi))^2 + \sigma_\ell(\tau_{\ell-1}\xi)\sigma_\ell''(\tau_{\ell-1}\xi)\right], \\
\alpha_{\ell,5} &= \alpha_{\ell-1,5}\mathbb{E}\left[(\sigma_\ell'(\tau_{\ell-1}\xi))^2 + \sigma_\ell(\tau_{\ell-1}\xi)\sigma_\ell''(\tau_{\ell-1}\xi)\right] \\
&\quad + \frac{\alpha_{\ell-1,4}^2}{4}\mathbb{E}\left[\sigma_\ell(\tau_{\ell-1}\xi)\sigma_\ell''''(\tau_{\ell-1}\xi) + 4\sigma_\ell'(\tau_{\ell-1}\xi)\sigma_\ell'''(\tau_{\ell-1}\xi) + 3(\sigma_\ell''(\tau_{\ell-1}\xi))^2\right].
\end{aligned}$$

Off the diagonal. We now move on to the non-diagonal (and more involved) entries of $\mathbf{K}_{\mathrm{CK},\ell}$. First note, for $i\neq j$, that

$$\begin{aligned}
\frac{[\mathbf{K}_{\mathrm{CK},\ell-1}]_{ij}}{\sqrt{[\mathbf{K}_{\mathrm{CK},\ell-1}]_{ii}}} &= \frac{[\mathbf{K}_{\mathrm{CK},\ell-1}]_{ij}}{\tau_{\ell-1} + \frac{1}{2\tau_{\ell-1}}\alpha_{\ell-1,4}\chi_i + \frac{4\tau_{\ell-1}^2\alpha_{\ell-1,5}-\alpha_{\ell-1,4}^2}{8\tau_{\ell-1}^3}(t_i+\psi_i)^2 + O(p^{-3/2})} \\
&= [\mathbf{K}_{\mathrm{CK},\ell-1}]_{ij}\left(\frac{1}{\tau_{\ell-1}} - \frac{1}{\tau_{\ell-1}^2}\left(\frac{\alpha_{\ell-1,4}}{2\tau_{\ell-1}}(t_i+\psi_i) + O(p^{-1})\right)\right) + O(p^{-3/2}) \\
&= \frac{1}{\tau_{\ell-1}}[\mathbf{K}_{\mathrm{CK},\ell-1}]_{ij} - \frac{\alpha_{\ell-1,4}\alpha_{\ell-1,1}}{2\tau_{\ell-1}^3}(t_i+\psi_i)\frac{1}{p}\mathbf{z}_i^\mathsf{T}\mathbf{z}_j + O(p^{-3/2}) \\
&= \frac{1}{\tau_{\ell-1}}[\mathbf{K}_{\mathrm{CK},\ell-1}]_{ij} + O(p^{-3/2}) = O(p^{-1/2})
\end{aligned}$$

with

$$\begin{aligned}
[\mathbf{K}_{\mathrm{CK},\ell-1}]_{ij} &= \alpha_{\ell-1,1}A_{ij} + \alpha_{\ell-1,2}(t_i+\psi_i)(t_j+\psi_j) + \alpha_{\ell-1,3}\left(\frac{1}{p}\mathbf{z}_i^\mathsf{T}\mathbf{z}_i\right)^2 + S_{ij} + O(p^{-3/2}) \\
&= O(p^{-1/2})
\end{aligned}$$

as per (31), where we recall that $S_{ij} = O(p^{-1})$ represents a matrix of the form $\alpha\frac{1}{p}\mathbf{z}_i^\mathsf{T}\mathbf{z}_j\left(\beta(t_i+\psi_i)+\gamma(t_j+\psi_j)\right)$ and of vanishing spectral norm as defined in (27).

Then,

$$\sqrt{[\mathbf{K}_{\mathrm{CK},\ell-1}]_{jj} - \frac{[\mathbf{K}_{\mathrm{CK},\ell-1}]_{ij}^2}{[\mathbf{K}_{\mathrm{CK},\ell-1}]_{ii}}} = \sqrt{\tau_{\ell-1}^2 + \alpha_{\ell-1,4}\chi_j + \alpha_{\ell-1,5}(t_j+\psi_j)^2 - \frac{\alpha_{\ell-1,1}^2}{\tau_{\ell-1}}(\frac{1}{p}\mathbf{z}_i^\mathsf{T}\mathbf{z}_j)^2 + O(p^{-3/2})}$$
$$= \tau_{\ell-1} + \frac{1}{2\tau_{\ell-1}}\left(\alpha_{\ell-1,4}\chi_j + \alpha_{\ell-1,5}(t_j+\psi_j)^2 - \frac{\alpha_{\ell-1,1}^2}{\tau_{\ell-1}}(\frac{1}{p}\mathbf{z}_i^\mathsf{T}\mathbf{z}_j)^2\right) - \frac{\alpha_{\ell-1,4}^2(t_j+\psi_j)^2}{8\tau_{\ell-1}^3} + O(p^{-3/2})$$
$$= \tau_{\ell-1} + \frac{1}{2\tau_{\ell-1}}\left(\alpha_{\ell-1,4}\chi_j - \frac{\alpha_{\ell-1,1}^2}{\tau_{\ell-1}}(\frac{1}{p}\mathbf{z}_i^\mathsf{T}\mathbf{z}_j)^2\right) + \frac{4\tau_{\ell-1}^2\alpha_{\ell-1,5} - \alpha_{\ell-1,4}^2}{8\tau_{\ell-1}^3}(t_j+\psi_j)^2 + O(p^{-3/2}).$$

As a consequence, we get, again by Taylor expansion that

$$\sigma_\ell\left(\sqrt{[\mathbf{K}_{\mathrm{CK},\ell-1}]_{ii}}\cdot\xi_i\right)\sigma_\ell\left(\frac{[\mathbf{K}_{\mathrm{CK},\ell-1}]_{ij}}{\sqrt{[\mathbf{K}_{\mathrm{CK},\ell-1}]_{ii}}}\cdot\xi_i + \sqrt{[\mathbf{K}_{\mathrm{CK},\ell-1}]_{jj} - \frac{[\mathbf{K}_{\mathrm{CK},\ell-1}]_{ij}^2}{[\mathbf{K}_{\mathrm{CK},\ell-1}]_{ii}}}\cdot\xi_j\right)$$
$$= \sigma_\ell\left(\tau_{\ell-1}\xi_i + \frac{1}{2\tau_{\ell-1}}\alpha_{\ell-1,4}\chi_i\xi_i + \frac{4\tau_{\ell-1}^2\alpha_{\ell-1,5} - \alpha_{\ell-1,4}^2}{8\tau_{\ell-1}^3}(t_i+\psi_i)^2\xi_i + O(p^{-3/2})\right)$$
$$\times\,\sigma_\ell\left(\frac{1}{\tau_{\ell-1}}[\mathbf{K}_{\mathrm{CK},\ell-1}]_{ij}\xi_i + \tau_{\ell-1}\xi_j + \frac{1}{2\tau_{\ell-1}}\left(\alpha_{\ell-1,4}\chi_j - \frac{\alpha_{\ell-1,1}^2}{\tau_{\ell-1}}(\frac{1}{p}\mathbf{z}_i^\mathsf{T}\mathbf{z}_j)^2\right)\xi_j\right.$$
$$\left.+\frac{4\tau_{\ell-1}^2\alpha_{\ell-1,5} - \alpha_{\ell-1,4}^2}{8\tau_{\ell-1}^3}(t_j+\psi_j)^2\xi_j + O(p^{-3/2})\right)$$
$$= \left(\sigma_\ell(\tau_{\ell-1}\xi_i) + \sigma_\ell'(\tau_{\ell-1}\xi_i)\xi_i\left(\frac{1}{2\tau_{\ell-1}}\alpha_{\ell-1,4}\chi_i + \frac{4\tau_{\ell-1}^2\alpha_{\ell-1,5} - \alpha_{\ell-1,4}^2}{8\tau_{\ell-1}^3}(t_i+\psi_i)^2\right)\right.$$
$$\left.+\sigma_\ell''(\tau_{\ell-1}\xi_i)\xi_i^2\frac{\alpha_{\ell-1,4}^2}{8\tau_{\ell-1}^2}(t_i+\psi_i)^2\right)$$
$$\times\left(\sigma_\ell(\tau_{\ell-1}\xi_j) + \sigma_\ell'(\tau_{\ell-1}\xi_j)\left(\frac{[\mathbf{K}_{\mathrm{CK},\ell-1}]_{ij}}{\tau_{\ell-1}}\xi_i + \frac{1}{2\tau_{\ell-1}}\left(\alpha_{\ell-1,4}\chi_j - \frac{\alpha_{\ell-1,1}^2}{\tau_{\ell-1}}(\frac{1}{p}\mathbf{z}_i^\mathsf{T}\mathbf{z}_j)^2\right)\xi_j\right)\right.$$
$$\left.+\sigma_\ell'(\tau_{\ell-1}\xi_j)\frac{4\tau_{\ell-1}^2\alpha_{\ell-1,5} - \alpha_{\ell-1,4}^2}{8\tau_{\ell-1}^3}(t_j+\psi_j)^2\xi_j + \frac{1}{2}\sigma_\ell''(\tau_{\ell-1}\xi_j)\left(\frac{\alpha_{\ell-1,1}\frac{1}{p}\mathbf{z}_i^\mathsf{T}\mathbf{z}_j}{\tau_{\ell-1}}\xi_i + \frac{\alpha_{\ell-1,4}(t_j+\psi_j)}{2\tau_{\ell-1}}\xi_j\right)^2\right)$$
$$+ O(p^{-3/2})$$
$$\equiv (\sigma_\ell(\tau_{\ell-1}\xi_i) + T_{1,i} + T_{2,i})\,(\sigma_\ell(\tau_{\ell-1}\xi_j) + T_{3,ij} + T_{3,j} + T_{4,ij} + T_{4,j} + S_{ij}) + O(p^{-3/2}),$$

where we denote the shortcuts:

$$T_{1,i} = \sigma_\ell'(\tau_{\ell-1}\xi_i)\xi_i\cdot\frac{\alpha_{\ell-1,4}}{2\tau_{\ell-1}}\chi_i = O(p^{-1/2}),$$
$$T_{2,i} = \left(\frac{\alpha_{\ell-1,5}\sigma_\ell'(\tau_{\ell-1}\xi_i)\xi_i}{2\tau_{\ell-1}} + \alpha_{\ell-1,4}^2\frac{\sigma_\ell''(\tau_{\ell-1}\xi_i)\xi_i^2\tau_{\ell-1} - \sigma_\ell'(\tau_{\ell-1}\xi_i)\xi_i}{8\tau_{\ell-1}^3}\right)(t_i+\psi_i)^2 = O(p^{-1}),$$

that *only* depend on ξ_i; and

$$T_{3,ij} = \sigma_\ell'(\tau_{\ell-1}\xi_j)\xi_i\cdot\frac{[\mathbf{K}_{\mathrm{CK},\ell-1}]_{ij}}{\tau_{\ell-1}} = O(p^{-1/2}),$$
$$T_{4,ij} = \frac{1}{2}\sigma_\ell''(\tau_{\ell-1}\xi_j)\xi_i^2\cdot\frac{\alpha_{\ell-1,1}^2}{\tau_{\ell-1}^2}(\frac{1}{p}\mathbf{z}_i^\mathsf{T}\mathbf{z}_j)^2 = O(p^{-1}),$$

that depend on both ξ_i and ξ_j; and

$$T_{3,j} = \sigma'_\ell(\tau_{\ell-1}\xi_j)\xi_j \cdot \left(\frac{\alpha_{\ell-1,4}}{2\tau_{\ell-1}}\chi_j - \frac{\alpha_{\ell-1,1}^2}{2\tau_{\ell-1}^2}(\frac{1}{p}\mathbf{z}_i^\mathsf{T}\mathbf{z}_j)^2 + \frac{4\tau_{\ell-1}^2\alpha_{\ell-1,5} - \alpha_{\ell-1,4}^2}{8\tau_{\ell-1}^3}(t_j+\psi_j)^2 \right) = O(p^{-1/2}),$$

$$T_{4,j} = \frac{1}{2}\sigma''_\ell(\tau_{\ell-1}\xi_j)\xi_j^2 \cdot \frac{\alpha_{\ell-1,4}^2}{4\tau_{\ell-1}^2}(t_j+\psi_j)^2 = O(p^{-1}),$$

that *only* depend on ξ_j, where we particularly note that the cross terms are of the form S_{ij}.

As such, we have

$$\begin{aligned}
&\sigma_\ell\left(\sqrt{[\mathbf{K}_{\mathrm{CK},\ell-1}]_{ii}}\cdot\xi_i\right)\sigma_\ell\left(\frac{[\mathbf{K}_{\mathrm{CK},\ell-1}]_{ij}}{\sqrt{[\mathbf{K}_{\mathrm{CK},\ell-1}]_{ii}}}\cdot\xi_i + \sqrt{[\mathbf{K}_{\mathrm{CK},\ell-1}]_{jj} - \frac{[\mathbf{K}_{\mathrm{CK},\ell-1}]_{ij}^2}{[\mathbf{K}_{\mathrm{CK},\ell-1}]_{ii}}}\cdot\xi_j\right)\\
&= \sigma_\ell(\tau_{\ell-1}\xi_i)\sigma_\ell(\tau_{\ell-1}\xi_j) + \sigma_\ell(\tau_{\ell-1}\xi_i)(T_{3,ij}+T_{4,ij}) + \sigma_\ell(\tau_{\ell-1}\xi_i)(T_{3,j}+T_{4,j})\\
&\quad + \sigma_\ell(\tau_{\ell-1}\xi_j)(T_{1,i}+T_{2,i}) + T_{1,i}(T_{3,ij}+T_{3,j}) + S_{ij} + O(p^{-3/2}),
\end{aligned}$$

with in particular

$$\begin{aligned}
T_{1,i}(T_{3,ij}+T_{3,j}) &= \sigma'_\ell(\tau_{\ell-1}\xi_i)\xi_i\cdot\frac{\alpha_{\ell-1,4}}{2\tau_{\ell-1}}(t_i+\psi_i)\cdot\sigma'_\ell(\tau_{\ell-1}\xi_j)\xi_i\frac{\alpha_{\ell-1,1}\frac{1}{p}\mathbf{z}_i^\mathsf{T}\mathbf{z}_j}{\tau_{\ell-1}}\\
&\quad + \sigma'_\ell(\tau_{\ell-1}\xi_i)\xi_i\cdot\frac{\alpha_{\ell-1,4}}{2\tau_{\ell-1}}(t_i+\psi_i)\cdot\sigma'_\ell(\tau_{\ell-1}\xi_j)\xi_j\frac{\alpha_{\ell-1,4}}{2\tau_{\ell-1}}(t_j+\psi_j) + O(p^{-3/2})\\
&= \sigma'_\ell(\tau_{\ell-1}\xi_i)\xi_i\sigma'_\ell(\tau_{\ell-1}\xi_j)\xi_j\frac{\alpha_{\ell-1,4}^2}{4\tau_{\ell-1}^2}(t_i+\psi_i)(t_j+\psi_j) + S_{ij} + O(p^{-3/2}).
\end{aligned}$$

We thus conclude that, for $\mathbb{E}[\sigma_\ell(\tau_{\ell-1}\xi)] = 0$ with $\xi\sim\mathcal{N}(0,1)$,

$$\begin{aligned}
[\mathbf{K}_{\mathrm{CK},\ell}]_{ij} &= \mathbb{E}\left[\sigma_\ell\left(\sqrt{[\mathbf{K}_{\mathrm{CK},\ell-1}]_{ii}}\cdot\xi_i\right)\sigma_\ell\left(\frac{[\mathbf{K}_{\mathrm{CK},\ell-1}]_{ij}}{\sqrt{[\mathbf{K}_{\mathrm{CK},\ell-1}]_{ii}}}\cdot\xi_i + \sqrt{[\mathbf{K}_{\mathrm{CK},\ell-1}]_{jj} - \frac{[\mathbf{K}_{\mathrm{CK},\ell-1}]_{ij}^2}{[\mathbf{K}_{\mathrm{CK},\ell-1}]_{ii}}}\cdot\xi_j\right)\right]\\
&= \mathbb{E}[\sigma_\ell(\tau_{\ell-1}\xi_i)(T_{3,ij}+T_{4,ij})] + \mathbb{E}\left[\sigma'_\ell(\tau_{\ell-1}\xi_i)\xi_i\sigma'_\ell(\tau_{\ell-1}\xi_j)\xi_j\right]\frac{\alpha_{\ell-1,4}^2}{4\tau_{\ell-1}^2}(t_i+\psi_i)(t_j+\psi_j)\\
&\quad + S_{ij} + O(p^{-3/2})\\
&= \mathbb{E}\left[\sigma_\ell(\tau_{\ell-1}\xi_i)\sigma'_\ell(\tau_{\ell-1}\xi_j)\xi_i\frac{[\mathbf{K}_{\mathrm{CK},\ell-1}]_{ij}}{\tau_{\ell-1}}\right] + \frac{1}{2}\mathbb{E}\left[\sigma_\ell(\tau_{\ell-1}\xi_i)\sigma''_\ell(\tau_{\ell-1}\xi_j)\xi_i^2\frac{\alpha_{\ell-1,1}^2}{\tau_{\ell-1}^2}\left(\frac{1}{p}\mathbf{z}_i^\mathsf{T}\mathbf{z}_j\right)^2\right]\\
&\quad + \mathbb{E}[\sigma'_\ell(\tau_{\ell-1}\xi_i)\xi_i]\mathbb{E}[\sigma'_\ell(\tau_{\ell-1}\xi_j)\xi_j]\frac{\alpha_{\ell-1,4}^2}{4\tau_{\ell-1}^2}(t_i+\psi_i)(t_j+\psi_j) + S_{ij} + O(p^{-3/2})\\
&= \mathbb{E}[\sigma'_\ell(\tau_{\ell-1}\xi)]^2[\mathbf{K}_{\mathrm{CK},\ell-1}]_{ij} + \frac{\alpha_{\ell-1,1}^2}{2}\mathbb{E}[\sigma''_\ell(\tau_{\ell-1}\xi)]^2\left(\frac{1}{p}\mathbf{z}_i^\mathsf{T}\mathbf{z}_j\right)^2\\
&\quad + \frac{\alpha_{\ell-1,4}^2}{4}\mathbb{E}[\sigma''_\ell(\tau_{\ell-1}\xi)]^2(t_i+\psi_i)(t_j+\psi_j) + S_{ij} + O(p^{-3/2}),
\end{aligned}$$

where we used again the fact that

$$\mathbb{E}[\xi f(\tau\xi)] = \tau\mathbb{E}[f'(\tau\xi)], \quad \mathbb{E}[\xi^2 f(\tau\xi)] = \mathbb{E}[(\xi^2-1)f(\tau\xi)] = \tau^2\mathbb{E}[f''(\tau\xi)], \tag{36}$$

if $\mathbb{E}[f(\tau\xi)] = 0$.

This allows us to conclude that

$$\alpha_{\ell,1} = \mathbb{E}[\sigma'_\ell(\tau_{\ell-1}\xi)]^2\alpha_{\ell-1,1}, \quad \alpha_{\ell,2} = \mathbb{E}[\sigma'_\ell(\tau_{\ell-1}\xi)]^2\alpha_{\ell-1,2} + \frac{1}{4}\mathbb{E}[\sigma''_\ell(\tau_{\ell-1}\xi)]^2\alpha_{\ell-1,4}^2, \tag{37}$$

$$\alpha_{\ell,3} = \mathbb{E}[\sigma'_\ell(\tau_{\ell-1}\xi)]^2\alpha_{\ell-1,3} + \frac{1}{2}\mathbb{E}[\sigma''_\ell(\tau_{\ell-1}\xi)]^2\alpha_{\ell-1,1}^2. \tag{38}$$

Assembling in matrix form. Following the discussion above, we have, uniformly for $i \neq j \in \{1, \ldots, n\}$ that,

$$[\mathbf{K}_{\mathrm{CK},\ell}]_{ij} = \alpha_{\ell,1} A_{ij} + \alpha_{\ell,2}(t_i + \psi_i)(t_j + \psi_j) + \alpha_{\ell,3}(\frac{1}{p}\mathbf{z}_i^\mathsf{T}\mathbf{z}_j)^2 + S_{ij} + O(p^{-3/2}), \tag{39}$$

and

$$[\mathbf{K}_{\mathrm{CK},\ell}]_{ii} = \tau_\ell^2 + O(p^{-1/2}), \tag{40}$$

so that in matrix form (by using the fact that $\|\mathbf{A}\| \leq n\|\mathbf{A}\|_\infty$ for $\mathbf{A} \in \mathbb{R}^{n\times n}$ with $\|\mathbf{A}\|_\infty = \max_{ij} |\mathbf{A}|_{ij}$ and $\{S_{ij}\}_{i,j} = O_{\|\cdot\|}(p^{-\frac{1}{2}})$, see [9]):

$$\mathbf{K}_{\mathrm{CK},\ell} = \alpha_{\ell,1}\mathbf{X}^\mathsf{T}\mathbf{X} + \mathbf{V}\mathbf{A}_\ell\mathbf{V}^\mathsf{T} + (\tau_\ell^2 - \tau_0^2\alpha_{\ell,1} - \tau_0^4\alpha_{\ell,3})\mathbf{I}_n + O_{\|\cdot\|}(p^{-\frac{1}{2}}) \tag{41}$$

where $O_{\|\cdot\|}(p^{-\frac{1}{2}})$ denotes matrices of spectral norm order $O(p^{-\frac{1}{2}})$ as $n, p \to \infty$, with

$$\mathbf{V} = [\mathbf{J}/\sqrt{p},\ \boldsymbol{\psi}] \in \mathbb{R}^{n\times(K+1)}, \quad \mathbf{A}_\ell = \begin{bmatrix} \alpha_{\ell,2}\mathbf{t}\mathbf{t}^\mathsf{T} + \alpha_{\ell,3}\mathbf{T} & \alpha_{\ell,2}\mathbf{t} \\ \alpha_{\ell,2}\mathbf{t}^\mathsf{T} & \alpha_{\ell,2} \end{bmatrix}, \tag{42}$$

and

$$\mathbf{T} = \left\{\frac{1}{p}\operatorname{tr}\mathbf{C}_a\mathbf{C}_b\right\}_{a,b=1}^K, \quad \mathbf{t} = \left\{\frac{1}{\sqrt{p}}\operatorname{tr}\mathbf{C}_a^\circ\right\}_{a=1}^K, \tag{43}$$

as the statement of Theorem 1. This thus concludes the proof of Theorem 1.

A.2 Proof of Theorem 2

In this section, we provide detailed proof of Theorem 2. We follows the same notations and working assumptions as in the proof of Theorem 1 in Appendix A.1.

As already mentioned in (6), the NTK matrices $\mathbf{K}_{\mathrm{NTK},\ell}$ of layer ℓ can be defined, in an iterative manner, via the CK matrices $\mathbf{K}_{\mathrm{CK},\ell}$ and $\mathbf{K}'_{\mathrm{CK},\ell}$ as follows [25]:

$$\begin{aligned} \mathbf{K}_{\mathrm{NTK},0} &= \mathbf{K}_{\mathrm{CK},0} = \mathbf{X}^\mathsf{T}\mathbf{X}, \\ \mathbf{K}_{\mathrm{NTK},\ell} &= \mathbf{K}_{\mathrm{CK},\ell} + \mathbf{K}_{\mathrm{NTK},\ell-1} \circ \mathbf{K}'_{\mathrm{CK},\ell} \end{aligned}$$

where '$\mathbf{A} \circ \mathbf{B}$' denotes the Hadamard product between two matrices $\mathbf{A}, \mathbf{B}$, and $\mathbf{K}'_{\mathrm{CK},\ell} \in \mathbb{R}^{n\times n}$ denotes a CK matrix with nonlinear function $\sigma'_\ell(\cdot)$ instead of $\sigma_\ell(\cdot)$ (as for $\mathbf{K}_{\mathrm{CK},\ell}$ in (4)), that is

$$[\mathbf{K}'_{\mathrm{CK},\ell}]_{ij} = \mathbb{E}_{u,v}[\sigma'_\ell(u)\sigma'_\ell(v)], \quad u, v \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} [\mathbf{K}'_{\mathrm{CK},\ell-1}]_{ii} & [\mathbf{K}'_{\mathrm{CK},\ell-1}]_{ij} \\ [\mathbf{K}'_{\mathrm{CK},\ell-1}]_{ij} & [\mathbf{K}'_{\mathrm{CK},\ell-1}]_{jj} \end{bmatrix}\right). \tag{44}$$

As in the proof of Theorem 1 in Appendix A.1, we follow the three-step proof strategy to work on the non-diagonal, the diagonal, and eventually the matrix form of $\mathbf{K}_{\mathrm{NTK},\ell}$.

We first write the non-diagonal entries of $\mathbf{K}_{\mathrm{NTK},1}$ as

$$\begin{aligned} [\mathbf{K}_{\mathrm{NTK},1}]_{ij} &= [\mathbf{K}_{\mathrm{CK},1}]_{ij} + [\mathbf{K}_{\mathrm{NTK},0}]_{ij}[\mathbf{K}'_{\mathrm{CK},1}]_{ij} = \alpha_{1,1}A_{ij} + \alpha_{1,2}(t_i + \psi_i)(t_j + \psi_j) \\ &\quad + \alpha_{1,3}\left(\frac{1}{p}\mathbf{z}_i^\mathsf{T}\mathbf{z}_i\right)^2 + \beta_{0,1}\frac{1}{p}\mathbf{z}_i^\mathsf{T}\mathbf{z}_j \cdot \alpha'_{1,1}\frac{1}{p}\mathbf{z}_i^\mathsf{T}\mathbf{z}_j + S_{ij} + O(p^{-3/2}) \\ &= \alpha_{1,1}A_{ij} + \alpha_{1,2}(t_i + \psi_i)(t_j + \psi_j) + (\alpha_{1,3} + \beta_{0,1}\alpha'_{1,1})\left(\frac{1}{p}\mathbf{z}_i^\mathsf{T}\mathbf{z}_j\right)^2 + S_{ij} + O(p^{-3/2}) \end{aligned}$$

where we denote $\alpha'_{1,1}, \alpha'_{2,1}, \alpha'_{3,1}$ the associated key coefficients of $\mathbf{K}'_{\mathrm{CK},1}$, with

$$[\mathbf{K}'_{\mathrm{CK},\ell}]_{ij} = \alpha'_{\ell,1}A_{ij} + \alpha'_{\ell,2}(t_i + \psi_i)(t_j + \psi_j) + \alpha'_{\ell,3}\left(\frac{1}{p}\mathbf{z}_i^\mathsf{T}\mathbf{z}_j\right)^2 + S_{ij} + O(p^{-3/2}) \tag{45}$$

so that $[\mathbf{K}_{\mathrm{NTK},1}]_{ij}$ (and thus $[\mathbf{K}_{\mathrm{NTK},\ell}]_{ij}$ for $\ell \in \{1, \ldots, L\}$) *must* also take the form

$$[\mathbf{K}_{\mathrm{NTK},\ell}]_{ij} = \beta_{\ell,1}A_{ij} + \beta_{\ell,2}(t_i + \psi_i)(t_j + \psi_j) + \beta_{\ell,3}\left(\frac{1}{p}\mathbf{z}_i^\mathsf{T}\mathbf{z}_j\right)^2 + S_{ij} + O(p^{-3/2}) \tag{46}$$

with

$$
\begin{aligned}
[\mathbf{K}_{\mathrm{NTK},\ell}]_{ij} &= [\mathbf{K}_{\mathrm{CK},\ell}]_{ij} + [\mathbf{K}_{\mathrm{NTK},\ell-1}]_{ij}[\mathbf{K}'_{\mathrm{CK},\ell}]_{ij} \\
&= \alpha_{\ell,1}A_{ij} + \alpha_{\ell,2}(t_i+\psi_i)(t_j+\psi_j) + \alpha_{\ell,3}\left(\frac{1}{p}\mathbf{z}_i^\mathsf{T}\mathbf{z}_j\right)^2 + \\
&\left(\beta_{\ell-1,1}A_{ij} + \beta_{\ell-1,2}(t_i+\psi_i)(t_j+\psi_j) + \beta_{\ell-1,3}\left(\frac{1}{p}\mathbf{z}_i^\mathsf{T}\mathbf{z}_j\right)^2\right) \times \\
&\left(\alpha'_{\ell,1}A_{ij} + \alpha'_{\ell,2}(t_i+\psi_i)(t_j+\psi_j) + \alpha'_{\ell,3}\left(\frac{1}{p}\mathbf{z}_i^\mathsf{T}\mathbf{z}_j\right)^2\right) + S_{ij} + O(p^{-3/2}) \\
&= \alpha_{\ell,1}A_{ij} + \alpha_{\ell,2}(t_i+\psi_i)(t_j+\psi_j) + (\alpha_{\ell,3} + \beta_{\ell-1,1}\alpha'_{\ell,1})\left(\frac{1}{p}\mathbf{z}_i^\mathsf{T}\mathbf{z}_j\right)^2 + S_{ij} + O(p^{-3/2}),
\end{aligned}
$$

so that

$$
\beta_{\ell,1} = \alpha_{\ell,1}, \quad \beta_{\ell,2} = \alpha_{\ell,2}, \quad \beta_{\ell,3} = \alpha_{\ell,3} + \beta_{\ell-1,1}\alpha'_{\ell,1}. \tag{47}
$$

Note, $\alpha'_{\ell,1}$ of $\mathbf{K}'_{\mathrm{CK},\ell}$ is indeed what we referred to as $\beta_{\ell,4}$ in the statement of Theorem 2.

We next evaluate the diagonal entries of $\mathbf{K}_{\mathrm{NTK},\ell}$ by letting

$$
[\mathbf{K}_{\mathrm{NTK},\ell-1}]_{ii} = \kappa_{\ell-1}^2 + O(p^{-1/2}), \tag{48}
$$

where we only expand to terms of order $O(1)$. Note that

$$
[\mathbf{K}_{\mathrm{NTK},\ell}]_{ii} = [\mathbf{K}_{\mathrm{CK},\ell}]_{ii} + [\mathbf{K}_{\mathrm{NTK},\ell-1}]_{ii} \cdot [\mathbf{K}'_{\mathrm{CK},\ell}]_{ii} = \tau_\ell^2 + O(p^{-1/2}),
$$

for

$$
[\mathbf{K}'_{\mathrm{CK},\ell-1}]_{ii} = (\tau'_{\ell-1})^2 + \alpha'_{\ell-1,4}\chi_i + \alpha'_{\ell-1,5}(t_i+\psi_i)^2 + O(p^{-3/2}), \tag{49}
$$

so that we obtain the following relation

$$
\kappa_\ell^2 = \tau_\ell^2 + \kappa_{\ell-1}^2(\tau'_{\ell-1})^2 = \tau_\ell^2 + \kappa_{\ell-1}^2\mathbb{E}[(\sigma'_{\ell-1}(\tau_{\ell-2}\xi))^2], \tag{50}
$$

with $\kappa_0 = \tau_0 = \sqrt{\operatorname{tr}\mathbf{C}^\circ/p}$ as in Assumption 1.

Putting everything together in matrix form, we obtain, as in the proof of Theorem 1 in Appendix A.1 that

$$
\mathbf{K}_{\mathrm{NTK},\ell} = \beta_{\ell,1}\mathbf{X}^\mathsf{T}\mathbf{X} + \mathbf{V}\mathbf{B}_\ell\mathbf{V}^\mathsf{T} + (\kappa_\ell^2 - \tau_0^2\beta_{\ell,1} - \tau_0^4\beta_{\ell,3})\mathbf{I}_n + O_{\|\cdot\|}(p^{-\frac{1}{2}}), \tag{51}
$$

with

$$
\mathbf{V} = [\mathbf{J}/\sqrt{p},\ \boldsymbol{\psi}] \in \mathbb{R}^{n\times(K+1)}, \quad \mathbf{B}_\ell = \begin{bmatrix} \beta_{\ell,2}\mathbf{t}\mathbf{t}^\mathsf{T} + \beta_{\ell,3}\mathbf{T} & \beta_{\ell,2}\mathbf{t} \\ \beta_{\ell,2}\mathbf{t}^\mathsf{T} & \beta_{\ell,2} \end{bmatrix}, \tag{52}
$$

which concludes the proof of Theorem 2.

A.3 Two equivalent centering approaches in the single-hidden-layer case

In this section, we aim to show that "centering" the CK matrices $\mathbf{K}_{\mathrm{CK}}$ by pre- and post-multiplying $\mathbf{P} = \mathbf{I}_n - \mathbf{1}_n\mathbf{1}_n^\mathsf{T}/n$ performed in [2, Theorem 1] is *equivalent* to take $\mathbb{E}[\sigma(\tau_0\xi)] = 0$ as in our Theorem 1 in the single-hidden-layer $\ell = 1$ setting, in the sense that one has

$$
\|\mathbf{P}(\mathbf{K}_{\mathrm{CK},1} - \tilde{\mathbf{K}}_{\mathrm{CK},1})\mathbf{P}\| \to 0 \tag{53}
$$

almost surely as $n, p \to \infty$, for the *same* $\tilde{\mathbf{K}}_{\mathrm{CK},1}$ as defined in Theorem 1 and an *arbitrary* choice of $\mathbb{E}[\sigma(\tau_0\xi)]$ (so in particular, one may freely take $\mathbb{E}[\sigma(\tau_0\xi)] \neq 0$ which is different from the setting of our Theorem 1). The proof is as follows.

First note that the assumption $\mathbb{E}[\sigma(\tau_0\xi)] = 0$ is *only* used for the off-diagonal entries of the CK matrix $\mathbf{K}_{\mathrm{CK},1}$, so we focus, in the sequel, only on the off-diagonal terms, while the discussions on the on-diagonal entries are the same as in Appendix A.1.

By its definition in (5) and the fact that $\mathbf{K}_{\mathrm{CK},0} = \mathbf{X}^\mathsf{T}\mathbf{X}$, one has

$$
[\mathbf{K}_{\mathrm{CK},1}]_{ij} = \mathbb{E}_{u,v}[\sigma_1(u)\sigma_1(v)], \text{ with } u, v \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \|\mathbf{x}_i\|^2 & \mathbf{x}_i^\mathsf{T}\mathbf{x}_j \\ \mathbf{x}_i^\mathsf{T}\mathbf{x}_j & \|\mathbf{x}_j\|^2 \end{bmatrix}\right), \tag{54}
$$

so by performing a Gram-Schmidt orthogonalization procedure as in the proof of Theorem 1 in Appendix A.1, one has

$$u = \|\mathbf{x}_i\| \cdot \xi_i, \quad v = \|\mathbf{x}_j\| \left(\angle_{ij} \cdot \xi_i + \sqrt{1 - \angle_{ij}^2} \cdot \xi_j \right) \tag{55}$$

for two *independent* standard Gaussian random variables ξ_i and ξ_j, where we denote the shortcut $\angle_{ij} \equiv \frac{\mathbf{x}_i^\mathsf{T} \mathbf{x}_j}{\|\mathbf{x}_i\| \cdot \|\mathbf{x}_j\|}$ for the "angle" between data vectors $\mathbf{x}_i$ and $\mathbf{x}_j$.

It can be checked, for $\mathbf{x}_i = \boldsymbol{\mu}_i/\sqrt{p} + \mathbf{z}_i/\sqrt{p}$ with $\mathbb{E}[\mathbf{z}_i] = \mathbf{0}$ and $\mathbb{E}[\mathbf{z}_i \mathbf{z}_i^\mathsf{T}] = \mathbf{C}_i$ that

$$\begin{aligned} \|\mathbf{x}_i\|^2 &= \frac{1}{p}(\boldsymbol{\mu}_i + \mathbf{z}_i)^\mathsf{T}(\boldsymbol{\mu}_i + \mathbf{z}_i) = \frac{1}{p}\|\boldsymbol{\mu}_i\|^2 + \frac{2}{p}\boldsymbol{\mu}_i^\mathsf{T}\mathbf{z}_i + \frac{1}{p}\mathbf{z}_i^\mathsf{T}\mathbf{z}_i \\ &= \frac{1}{p}\|\boldsymbol{\mu}_i\|^2 + \underbrace{\frac{2}{p}\boldsymbol{\mu}_i^\mathsf{T}\mathbf{z}_i}_{O(p^{-1})} + \underbrace{\frac{1}{p}\operatorname{tr}\mathbf{C}^\circ}_{\equiv \tau_0^2 = O(1)} + \underbrace{\frac{1}{p}\operatorname{tr}\mathbf{C}_i^\circ}_{\equiv t_i = O(p^{-1/2})} + \underbrace{\psi_i}_{O(p^{-1/2})} \end{aligned}$$

where we recall the definition $\psi_i \equiv \frac{1}{p}\|\mathbf{z}_i\|^2 - \frac{1}{p}\operatorname{tr}\mathbf{C}_i = O(p^{-1/2})$. As such, by Taylor-expanding $\sqrt{\|\mathbf{x}_i\|^2}$ around $\|\mathbf{x}_i\|^2 \simeq \tau_0^2 = O(1)$, we get

$$\begin{aligned} \|\mathbf{x}_i\| &= \tau_0 + \frac{1}{2\tau_0}(\|\boldsymbol{\mu}_i\|^2/p + 2\boldsymbol{\mu}_i^\mathsf{T}\mathbf{z}_i/p + t_i + \psi_i) - \frac{1}{8\tau_0^3}(t_i + \psi_i)^2 + O(p^{-3/2}) \\ &\equiv \tau_0 + \theta_i + O(p^{-3/2}), \end{aligned}$$

where we denote the shortcut

$$\theta_i \equiv \frac{1}{2\tau_0}(\|\boldsymbol{\mu}_i\|^2/p + 2\boldsymbol{\mu}_i^\mathsf{T}\mathbf{z}_i/p + t_i + \psi_i) - \frac{1}{8\tau_0^3}(t_i + \psi_i)^2 = O(p^{-1/2}), \tag{56}$$

so that

$$\begin{aligned} \|\mathbf{x}_j\|\angle_{ij} &= \frac{\frac{1}{p}(\boldsymbol{\mu}_i + \mathbf{z}_i)^\mathsf{T}(\boldsymbol{\mu}_j + \mathbf{z}_j)}{\|\mathbf{x}_i\|} \\ &= \frac{\frac{1}{p}\boldsymbol{\mu}_i^\mathsf{T}\boldsymbol{\mu}_j + \frac{1}{p}(\boldsymbol{\mu}_i^\mathsf{T}\mathbf{z}_j + \boldsymbol{\mu}_j^\mathsf{T}\mathbf{z}_i) + \frac{1}{p}\mathbf{z}_i^\mathsf{T}\mathbf{z}_j}{\tau_0 + \frac{1}{2\tau_0}(t_i + \psi_i) + O(p^{-1})} + O(p^{-3/2}) \\ &= \left(\frac{1}{\tau_0} - \frac{t_i + \psi_i}{2\tau_0^3} + O(p^{-1})\right)\left(\frac{1}{p}\boldsymbol{\mu}_i^\mathsf{T}\boldsymbol{\mu}_j + \frac{1}{p}(\boldsymbol{\mu}_i^\mathsf{T}\mathbf{z}_j + \boldsymbol{\mu}_j^\mathsf{T}\mathbf{z}_i) + \frac{1}{p}\mathbf{z}_i^\mathsf{T}\mathbf{z}_j\right) + O(p^{-3/2}) \\ &= \frac{1}{\tau_0}\left(\frac{1}{p}\mathbf{z}_i^\mathsf{T}\mathbf{z}_j + \frac{1}{p}\boldsymbol{\mu}_i^\mathsf{T}\boldsymbol{\mu}_j + \frac{1}{p}(\boldsymbol{\mu}_i^\mathsf{T}\mathbf{z}_j + \boldsymbol{\mu}_j^\mathsf{T}\mathbf{z}_i) + S_{ij}\right) + O(p^{-3/2}) \\ &= \frac{1}{\tau_0}A_{ij} + S_{ij} + O(p^{-3/2}). \end{aligned}$$

Therefore, again by Taylor-expansion,

$$\begin{aligned} \sqrt{\|\mathbf{x}_j\|^2 - (\|\mathbf{x}_j\|\angle_{ij})^2} &= \sqrt{(\|\boldsymbol{\mu}_j\|^2/p + 2\boldsymbol{\mu}_j^\mathsf{T}\mathbf{z}_j/p + \tau_0^2 + t_j + \psi_j) - (A_{ij}/\tau_0 + S_{ij})^2} \\ &= \tau_0 + \frac{1}{2\tau_0}(\|\boldsymbol{\mu}_i\|^2/p + 2\boldsymbol{\mu}_i^\mathsf{T}\mathbf{z}_i/p + t_i + \psi_i) - \frac{1}{8\tau_0^3}(t_i + \psi_i)^2 \\ &\quad - \frac{1}{2\tau_0^3}\left(\frac{1}{p}\mathbf{z}_i^\mathsf{T}\mathbf{z}_j\right)^2 + S_{ij} + O(p^{-3/2}) \\ &= \tau_0 + \theta_j - \frac{1}{2\tau_0^3}\left(\frac{1}{p}\mathbf{z}_i^\mathsf{T}\mathbf{z}_j\right)^2 + S_{ij} + O(p^{-3/2}). \end{aligned}$$

Following the same idea, we again Taylor-expand $\sigma_1(\cdot)$ in the definition of $\mathbf{K}_{\mathrm{CK},1}$ as

$$\sigma_1(u) = \sigma_1(\tau_0\xi_i) + \sigma_1'(\tau_0\xi_i)\xi_i\theta_i + \frac{1}{8\tau_0^2}\sigma_1''(\tau_0\xi_i)\xi_i^2(t_i + \psi_i)^2 + O(p^{-3/2}),$$

and

$$\begin{aligned}\sigma_1(v) &= \sigma_1\left(\|\mathbf{x}_j\|\angle_{ij}\xi_j + \|\mathbf{x}_j\|\sqrt{1-\angle_{ij}^2}\xi_i\right)\\ &= \sigma_1\left(\tau_0\xi_j + \xi_j\theta_j - \xi_j\frac{1}{2\tau_0^3}\left(\frac{1}{p}\mathbf{z}_i^\mathsf{T}\mathbf{z}_j\right)^2 + \xi_i\frac{1}{\tau_0}A_{ij} + S_{ij} + O(p^{-3/2})\right)\\ &= \sigma_1(\tau_0\xi_j) + \sigma_1'(\tau_0\xi_j)\xi_j\theta_j + \frac{1}{8\tau_0^2}\sigma_1''(\tau_0\xi_j)\xi_j^2(t_j+\psi_j)^2 + X_{ij} + O(p^{-3/2}),\end{aligned}$$

with

$$X_{ij} = \frac{1}{\tau_0}\xi_i\sigma_1'(\tau_0\xi_j)A_{ij} + \frac{1}{2\tau_0^2}\left(\frac{1}{p}\mathbf{z}_i^\mathsf{T}\mathbf{z}_j\right)^2\left(\xi_i^2\sigma_1''(\tau_0\xi_j) - \frac{1}{\tau_0}\xi_j\sigma_1'(\tau_0\xi_j)\right) + S_{ij} = O(p^{-1/2}),$$

where we recall the definition

$$A_{ij} = \underbrace{\frac{1}{p}\mathbf{z}_i^\mathsf{T}\mathbf{z}_j}_{O(p^{-1/2})} + \underbrace{\frac{1}{p}\boldsymbol{\mu}_i^\mathsf{T}\boldsymbol{\mu}_j + \frac{1}{p}(\boldsymbol{\mu}_i^\mathsf{T}\mathbf{z}_j + \boldsymbol{\mu}_j^\mathsf{T}\mathbf{z}_i)}_{O(p^{-1})} = \mathbf{x}_i^\mathsf{T}\mathbf{x}_j. \tag{57}$$

For independent ξ_i and ξ_j, we denote the following coefficients

$$p_0 = \mathbb{E}[\sigma_1(\tau_0\xi)], \quad p_1 = \mathbb{E}[\sigma_1'(\tau_0\xi)], \quad p_2 = \mathbb{E}[\sigma_1''(\tau_0\xi)], \quad p_3 = \mathbb{E}[\sigma_1'''(\tau_0\xi)], \tag{58}$$

so that

$$\mathbb{E}[\xi\sigma_1(\tau_0\xi)] = \tau_0\mathbb{E}[\sigma_1'(\tau_0\xi)] = \tau_0 p_1, \quad \mathbb{E}[\xi\sigma_1'(\tau_0\xi)] = \tau_0 p_2, \tag{59}$$

as well as

$$\mathbb{E}[\xi^2\sigma_1''(\tau_0\xi)] = \mathbb{E}[(\xi^2-1)\sigma_1''(\tau_0\xi)] + p_2 = \tau_0^2\mathbb{E}[\sigma_1''''(\tau_0\xi)] + p_2 \equiv \tau_0^2 p_4 + p_2, \tag{60}$$

for $p_4 = \mathbb{E}[\sigma_1''''(\tau_0\xi)]$.

This further allows us to write, for $A_{ij} = O(p^{-1/2})$ and $\theta_j = O(p^{-1/2})$ that

$$\begin{aligned}&[\mathbf{K}_{\mathrm{CK},1}]_{ij} = \mathbb{E}_{u,v}[\sigma_1(u)\sigma_1(v)]\\ =&\mathbb{E}\left[\sigma_1(\tau_0\xi_i) + \sigma_1'(\tau_0\xi_i)\xi_i\theta_i + \frac{1}{8\tau_0^2}\sigma_1''(\tau_0\xi_i)\xi_i^2(t_i+\psi_i)^2\right]\\ &\times\mathbb{E}\left[\sigma_1(\tau_0\xi_j) + \sigma_1'(\tau_0\xi_j)\xi_j\theta_j + \frac{1}{8\tau_0^2}\sigma_1''(\tau_0\xi_j)\xi_j^2(t_j+\psi_j)^2\right]\\ &+\mathbb{E}\left[\left(\sigma_1(\tau_0\xi_i) + \sigma_1'(\tau_0\xi_i)\xi_i\theta_i + \frac{1}{8\tau_0^2}\sigma_1''(\tau_0\xi_i)\xi_i^2(t_i+\psi_i)^2\right)X_{ij}\right] + O(p^{-3/2})\\ =&\left(p_0 + \tau_0 p_2\theta_i + \frac{\tau_0^2 p_4 + p_2}{8\tau_0^2}(t_i+\psi_i)^2\right)\left(p_0 + \tau_0 p_2\theta_j + \frac{\tau_0^2 p_4 + p_2}{8\tau_0^2}(t_j+\psi_j)^2\right)\\ &+\mathbb{E}\left[\sigma_1(\tau_0\xi_i)X_{ij}\right] + S_{ij} + O(p^{-3/2}),\end{aligned}$$

where the expectation is taken with respect to the *independent* ξ_i and ξ_j (so, in fact, conditioned on $\mathbf{x}_i, \mathbf{x}_j$), so that

$$
\begin{aligned}
&[\mathbf{K}_{\mathrm{CK},1}]_{ij} = \mathbb{E}_{u,v}[\sigma_1(u)\sigma_1(v)] \\
=&\left(p_0 + \tau_0 p_2 \theta_i + \frac{\tau_0^2 p_4 + p_2}{8\tau_0^2}(t_i + \psi_i)^2\right)\left(p_0 + \tau_0 p_2 \theta_j + \frac{\tau_0^2 p_4 + p_2}{8\tau_0^2}(t_j + \psi_j)^2\right) \\
&+ \mathbb{E}\left[\frac{1}{\tau_0}\xi_i \sigma_1(\tau_0\xi_i)\sigma_1'(\tau_0\xi_j)A_{ij} + \frac{1}{2\tau_0^2}\left(\frac{1}{p}\mathbf{z}_i^\mathsf{T}\mathbf{z}_j\right)^2\left(\xi_i^2\sigma_1(\tau_0\xi_i)\sigma_1''(\tau_0\xi_j) - \frac{1}{\tau_0}\sigma_1(\tau_0\xi_i)\xi_j\sigma_1'(\tau_0\xi_j)\right)\right] \\
&+ S_{ij} + O(p^{-3/2}) \\
=&\left(p_0 + \frac{p_2}{2}\chi_i + \frac{p_4}{8}(t_i + \psi_i)^2\right)\left(p_0 + \frac{p_2}{2}\chi_j + \frac{p_4}{8}(t_j + \psi_j)^2\right) \\
&+ p_1^2 A_{ij} + \frac{1}{2\tau_0^2}\left(\frac{1}{p}\mathbf{z}_i^\mathsf{T}\mathbf{z}_j\right)^2 \cdot p_2\left(\mathbb{E}[(\xi^2 - 1)\sigma_1(\tau_0\xi)]\right) + S_{ij} + O(p^{-3/2}) \\
=&p_0^2 + \frac{p_0 p_2}{2}(\chi_i + \chi_j) + \frac{p_0 p_4}{8}\left((t_i + \psi_i)^2 + (t_j + \psi_j)^2\right) + \frac{p_2^2}{4}(t_i + \psi_i)(t_j + \psi_j) \\
&+ p_1^2 A_{ij} + \frac{p_2^2}{2}\left(\frac{1}{p}\mathbf{z}_i^\mathsf{T}\mathbf{z}_j\right)^2 + S_{ij} + O(p^{-3/2}),
\end{aligned}
$$

where we recall the shortcut

$$
\theta_i \equiv \frac{1}{2\tau_0}(\|\boldsymbol{\mu}_i\|^2/p + 2\boldsymbol{\mu}_i^\mathsf{T}\mathbf{z}_i/p + t_i + \psi_i) - \frac{1}{8\tau_0^2\tau_0}(t_i + \psi_i)^2 \equiv \frac{\chi_i}{2\tau_0} - \frac{(t_i + \psi_i)^2}{8\tau_0^2\tau_0} = O(p^{-1/2}), \tag{61}
$$

with

$$
\chi_i \equiv t_i + \psi_i + \|\boldsymbol{\mu}_i\|^2/p + 2\boldsymbol{\mu}_i^\mathsf{T}\mathbf{z}_i/p = \|\mathbf{x}_i\|^2 - \tau_0. \tag{62}
$$

This gives, in matrix form,

$$
\begin{aligned}
\mathbf{K}_{\mathrm{CK},1} =& p_0^2 \mathbf{1}_n\mathbf{1}_n^\mathsf{T} + p_1^2\left(\frac{1}{p}\mathbf{Z}^\mathsf{T}\mathbf{Z} + \frac{1}{p}\mathbf{J}\mathbf{M}^\mathsf{T}\mathbf{M}\mathbf{J}^\mathsf{T} + \frac{1}{p}(\mathbf{J}\mathbf{M}^\mathsf{T}\mathbf{Z} + \mathbf{Z}^\mathsf{T}\mathbf{M}\mathbf{J}^\mathsf{T})\right) \\
&+ \frac{p_0 p_2}{2}(\boldsymbol{\chi}\mathbf{1}_n^\mathsf{T} + \mathbf{1}_n\boldsymbol{\chi}^\mathsf{T}) + \frac{p_0 p_4}{8}\left((\{t_a\mathbf{1}_{n_a}\}_{a=1}^K + \boldsymbol{\psi})^2\mathbf{1}_n^\mathsf{T} + \mathbf{1}_n[(\{t_a\mathbf{1}_{n_a}\}_{a=1}^K + \boldsymbol{\psi})^2]^\mathsf{T}\right) \\
&+ \frac{p_2^2}{4}(\{t_a\mathbf{1}_{n_a}\}_{a=1}^K + \boldsymbol{\psi})(\{t_a\mathbf{1}_{n_a}\}_{a=1}^K + \boldsymbol{\psi})^\mathsf{T} + \frac{p_2^2}{2}\left(\frac{1}{p}\mathbf{Z}^\mathsf{T}\mathbf{Z}\right)^{\circ 2} \\
&+ (\mathbb{E}[\sigma_1^2(\tau_0\xi)] - p_0^2 - \tau_0^2 p_1^2)\mathbf{I}_n + O_{\|\cdot\|}(p^{-1/2})
\end{aligned}
$$

where we denote $\boldsymbol{\chi} \equiv \{\chi_i\}_{i=1}^n \in \mathbb{R}^n$, $\mathbf{A}^{\circ 2}$ the *entry-wise* square of the matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, i.e., $[\mathbf{A}^{\circ 2}]_{ij} = [\mathbf{A}_{ij}]^2$, and use again the fact that $\|\mathbf{A}\| \le n\|\mathbf{A}\|_\infty$ for $\mathbf{A} \in \mathbb{R}^{n\times n}$ with $\|\mathbf{A}\|_\infty = \max_{ij}|\mathbf{A}|_{ij}$, $\{S_{ij}\}_{i,j} = O_{\|\cdot\|}(p^{-\frac{1}{2}})$ as well as $\left(\frac{1}{p}\mathbf{Z}^\mathsf{T}\mathbf{Z}\right)^{\circ 2} = \frac{1}{p}\mathbf{J}\mathbf{T}\mathbf{J}^\mathsf{T} + O_{\|\cdot\|}(p^{-1/2})$ according to [9].

Finally, using the fact that for $\mathbf{P} = \mathbf{I}_n - \mathbf{1}_n\mathbf{1}_n^\mathsf{T}/n$, we have $\mathbf{1}_n^\mathsf{T}\mathbf{P} = \mathbf{0}, \mathbf{P}\mathbf{1}_n = \mathbf{0}$, we conclude the proof of (53) with the same expression of $\tilde{\mathbf{K}}_{\mathrm{CK},1}$ as in the statement of our Theorem 1, *without* the assumption $\mathbb{E}[\sigma_1(\tau_0\xi)] = 0$. This, however, no longer holds in the multi-layer setting with a number of layers $L \ge 1$.

A.4 Proof and discussions of Corollary 1

To prove Corollary 1, it can be easily checked that the i.i.d. entries of the weights $\mathbf{W}$ defined in (16) have zero mean and unit variance. So we focus on the design of the activations.

To ensure that the activation functions $\sigma_\ell(\cdot)$s are "centered" and satisfy $\mathbb{E}[\sigma_\ell(\tau_{\ell-1}\xi)] = 0$, we define, with a slight abuse of notation, for the non-negative sequence $\tau_1, \dots, \tau_L$ defined in Theorem 1,

$$
\sigma_T(t) = a \cdot (1_{t<s_1} + 1_{t>s_2}), \quad \sigma_Q(t) = b_1 \cdot (1_{t<r_1} + 1_{t>r_4}) + b_2 \cdot 1_{r_2\le t\le r_3}, \tag{63}
$$

and take $\alpha_{\ell,0} \equiv \mathbb{E}[\sigma_T(\tau_{\ell-1}\xi)]$, $\sigma_T(\tau_{\ell-1}\xi) \equiv \tilde{\sigma}_T(\tau_{\ell-1}\xi) = \sigma_T(\tau_{\ell-1}\xi) - \alpha_{\ell,0}$, which serves as the activation of the first $\ell = 1, \dots, L-1$ layers, and a, s_1 and s_2 satisfying the following equations

$$
\mathbb{E}[\sigma_T'(\tau_{\ell-1}\xi)] = \frac{a}{\sqrt{2\pi}\tau_{\ell-1}} \cdot \left(e^{-s_2^2/(2\tau_{\ell-1}^2)} - e^{-s_1^2/(2\tau_{\ell-1}^2)}\right) \tag{64}
$$

$$\mathbb{E}[\sigma_T''(\tau_{\ell-1}\xi)] = \frac{a}{\sqrt{2\pi}\tau_{\ell-1}^3} \cdot \left(s_2 e^{-s_2^2/(2\tau_{\ell-1}^2)} - s_1 e^{-s_1^2/(2\tau_{\ell-1}^2)}\right) \tag{65}$$

$$\mathbb{E}[(\sigma_T^2(\tau_{\ell-1}\xi))''] = \frac{a^2 - 2a\cdot\alpha_{\ell,0}}{\sqrt{2\pi}\tau_{\ell-1}^3} \cdot \left(s_2 e^{-s_2^2/(2\tau_{\ell-1}^2)} - s_1 e^{-s_1^2/(2\tau_{\ell-1}^2)}\right) \tag{66}$$

$$\mathbb{E}[\sigma_T^2(\tau_{\ell-1}\xi)] = \frac{a^2}{2}\left(\mathrm{erf}\left(\frac{s_1}{\sqrt{2}\tau_{\ell-1}}\right) - \mathrm{erf}\left(\frac{s_2}{\sqrt{2}\tau_{\ell-1}}\right) + 2\right) - \alpha_{\ell,0}^2 \tag{67}$$

and $\alpha_{L,0} \equiv \mathbb{E}[\sigma_Q(\tau\xi)]$, $\sigma_T(\tau\xi) \equiv \tilde{\sigma}_T(\tau\xi) = \sigma_T(\tau\xi) - \alpha_{L,0}$, which serves as the activation of the last and Lth layer, and b_1, b_2, r_1, r_2, r_3 and r_4 satisfying the following equations

$$\mathbb{E}[\sigma_Q'(\tau\xi)] = \frac{b_1\left(e^{-r_4^2/(2\tau^2)} - e^{-r_2^2/(2\tau^2)}\right)}{\sqrt{2\pi}\tau} + \frac{b_2\left(e^{-r_2^2/(2\tau^2)} - e^{-r_2^3/(2\tau^2)}\right)}{\sqrt{2\pi}\tau} \tag{68}$$

$$\mathbb{E}[\sigma_Q''(\tau\xi)] = \frac{b_1\left(r_4 e^{-r_4^2/(2\tau^2)} - r_1 e^{-r_1^2/(2\tau^2)}\right)}{\sqrt{2\pi}\tau^3} + \frac{b_2\left(r_2 e^{-r_2^2/(2\tau^2)} - r_3 e^{-r_2^3/(2\tau^2)}\right)}{\sqrt{2\pi}\tau^3} \tag{69}$$

$$\begin{aligned}\mathbb{E}[(\sigma_Q^2(\tau\xi))''] = &\frac{b_1^2\left(r_4 e^{-r_4^2/(2\tau^2)} - r_1 e^{-r_1^2/(2\tau^2)}\right)}{\sqrt{2\pi}\tau^3} + \frac{b_2^2\left(r_2 e^{-r_2^2/(2\tau^2)} - r_3 e^{-r_2^3/(2\tau^2)}\right)}{\sqrt{2\pi}\tau^3} \\ &- 2\alpha_{L,0}\mathbb{E}[(\sigma_Q''(\tau\xi))]\end{aligned} \tag{70}$$

$$\begin{aligned}\mathbb{E}[(\sigma_Q^2(\tau\xi))] = &\frac{b_1^2}{2}\left(\mathrm{erf}\left(\frac{r_1}{\sqrt{2}\tau}\right) - \mathrm{erf}\left(\frac{r_4}{\sqrt{2}\tau}\right)\right) + b_1^2 + \frac{b_2^2}{2}\left(\mathrm{erf}\left(\frac{r_2}{\sqrt{2}\tau}\right) - \mathrm{erf}\left(\frac{r_3}{\sqrt{2}\tau}\right)\right) \\ &- \alpha_{L,0}^2\end{aligned} \tag{71}$$

with $\tau = \tau_{L-1}$.

A few remarks on Corollary 1 and Algorithm 1 are as follows.

On the numerical determinations of σ_T and σ_Q. The above system of nonlinear equations does not admit explicit solutions, but can be solved efficiently using, for example, a (numerical) least squares method. Precisely, we use the numerical least squares method (the optimize.minimize function of SciPy library) to solve the above system of equations, and run for $1\,000$ times with random and independent initializations to get $1\,000$ solutions, among which we choose the optimal parameters to determine σ_Q and σ_T.

On the two activations. Note that in Algorithm 1 we use the activation σ_T and σ_Q respectively for the first $\ell = 1, \ldots, L-1$ and the final and Lth layer, since we *only* need to match the key parameters $\alpha_{\ell,1}$, $\alpha_{\ell,2}$ and $\alpha_{\ell,3}$ for the first $\ell = 1, \ldots, L-1$ layer, and the additional parameter τ_ℓ for the last Lth, so as to obtain spectrally equivalent CK and NTK matrices for the whole network of depth L. Also note that the proposed activation functions σ_T and σ_Q have respectively three and five (in fact six parameters with the symmetric constraint $r_1 - r_2 = r_3 - r_4$ as in Figure 1) parameters that are freely tunable. And we have respectively three and four (nonlinear) equations to determine these parameters in the system of equations above.

B Additional experiments

In this section, we provide additional experiments to demonstrate the advantageous performance of the proposed NTK-LC approach. Figure 4 depicts the classification accuracies using three different neural networks: (i) the original "dense and unquantized" nets with three fully-connected layers of ReLU activations, (ii) the proposed sparse and quantized NTK-LC as per Algorithm 1, and (iii) the "heuristically" compressed networks by (iii-i) uniformly and randomly zeroing out 90% of the weights, as well as (iii-ii) natively binarizing using $\sigma(t) = 1_{t<-1} + 1_{t>1}$, on two tasks of MNIST data classification [28] having five classes (digits $0, 1, 2, 3, 4$) and two classes (digits 6 versus 8). This allows us to have a more qualitative assessment of the impact of data and task on the performance of the proposed NTK-LC approach. We see, as in Figure 3 for ten-class MNIST and ten-class

CIFAR10, that the proposed NTK-LC approach significantly outperform the two "naive" compression approaches, and can achieve a memory compression rate of 10^3 and a level of sparsity up to 90%, with virtually no performance loss. Also note that the experimental settings of Figure 4 is almost the *same* as those of Figure 3 in Section 4, except that the former networks have less neurons per layer and slightly higher level of sparsity (90% here instead of 80% in the setting of Figure 3), to solve the simpler two-class or five-class classification problems.

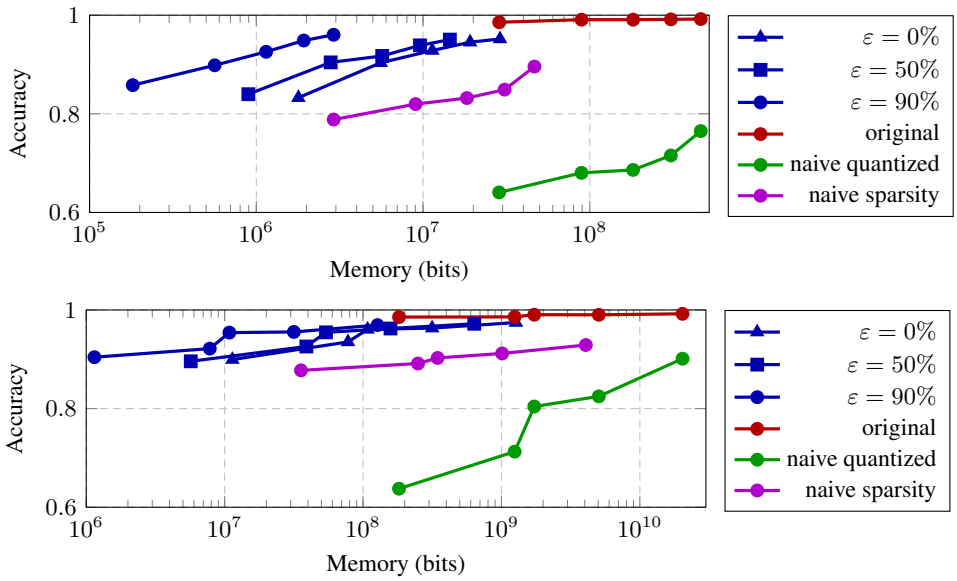


Figure 4: Test accuracy of classification on 2-class (**top**) MNIST dataset - digits 6 versus 8 and 5-class (**bottom**) MNIST dataset - digits $(0, 1, 2, 3, 4)$. **Blue** curves represent the proposed NTK-LC approach with different levels of sparsity $\varepsilon \in \{0\%, 50\%, 90\%\}$, **purple** curves represent the heuristic sparsification approach by uniformly zeroing out 90% of the weights, **green** curves represent the heuristic quantization approach using the binary activation $\sigma(t) = 1_{t<-1} + 1_{t>1}$ (only applied on the first two layers, otherwise the performance is too poor to be compared to other curves), and **red** curves represent the original (dense and unquantized) network. All nets have three fully-connected layers, and the original network uses ReLU activations for all layers. Memory varies due to the **change of layer width** of the network.

In Table 1 and 2, we evaluate the impact of activation functions on the classification performance on data of *different nature*, on a set of fully-connected DNN models having three hidden layers (of width $d_1 = 3\,000, d_2 = 3\,000, d_3 = 1\,000$ in each layer) and use the *same* activation $\sigma(\cdot)$ for all layers.

More precisely, Table 1 depicts the classification accuracy and the values of the key parameters α_1, α_2, α_3 and τ for different activations $\sigma(\cdot)$ in the asymptotic equivalent CK matrix $\tilde{\mathbf{K}}_{\mathrm{CK}}$ defined in Theorem 1 of the third and final layer of the network, on a binary classification of MNIST data (class 6 versus 8). Similarly, Table 2 compares the classification accuracy and α_1, α_2, α_3, τ for different activations, on two-class GMM data with identical mean $\boldsymbol{\mu}_a = \mathbf{0}_p$ and different covariance $\mathbf{C}_a = (1 + 8(a-1)/\sqrt{p})\mathbf{I}_p$, $a \in \{1, 2\}$, The numerical experiments are performed on a training set of size $12\,000$, a test set of $1\,800$, with $\mathbf{x}_1, \ldots, \mathbf{x}_{n/2} \in \mathcal{C}_1$ and $\mathbf{x}_{n/2+1}, \ldots, \mathbf{x}_n \in \mathcal{C}_2$, for standard Gaussian $\mathbf{W}$ on both MNIST and GMM data.

We observe from Table 1 and 2 that:

(i) while in theory, the key parameters α_1, α_2, α_3 and τ (in Theorem 1) depend on both (the statistics of) the data and the activation, the impact of the activation σ appears much more significant;

(ii) by using some σ (with the corresponding α_1, α_2 and/or α_3 being zero), one asymptotically "discards" either the first-order ($\boldsymbol{\mu}_a$) or the second-order ($\mathbf{t}, \mathbf{T}$) statistics of the data (which, per Theorem 1, are respectively weighted by the key parameter α_1, α_2 and α_3), resulting in performance degradation;

(iii) precisely, we divide commonly used activations in Table 1 and 2 into the following three categories:

1. covariance-oriented activations with $\alpha_1 = 0$: this includes $\cos(t)$ and $|t|$; and
2. mean-oriented activations with $\alpha_2 = 0$ and $\alpha_3 = 0$: this includes $1_{t\geq 0}$, $\mathrm{sign}(t)$, $\frac{1}{1+e^{-x}}$ [18], $\sin(t)$, linear function, and the Gaussian error function $\mathrm{erf(t)}$; and
3. balanced activations with nonzero $\alpha_1, \alpha_2, \alpha_3$: this includes ReLU activation $\mathrm{ReLU}(t) = \max(t, 0)$ and Leaky ReLU activation [39].

The above classification of activation functions is reminiscent of that proposed in [33], which is, however, only valid in a single-hidden-layer setting. In line with the observations made in [33], we see in Table 1 that covariance-oriented activations behave poorly in the classification of MNIST data (that are known to have very different first-order statistics, see for example [33, Table 3]), while mean-oriented activations yield unsatisfactory performance on GMM data having different covariance structure in Table 2. In a sense, the parameter α_1 characterizes the "ability" of a given net to extract first-order data statistics and α_2, α_3 the "ability" to extract second-order statistics from the input data, respectively.

Table 1: Classification accuracy and values of α_1, α_2, α_3 and τ at the third and final layer, on MNIST data (digits 6 versus 8).

$\sigma(t)$	α_1	α_2	α_3	τ	Accuracy
$\max(0, t)$	0.0156	0.0105	0.0112	0.1994	0.971
$0.1t \cdot 1_{t<0} + t \cdot 1_{t\geq 0}$	0.0083	0.0097	0.0081	0.1750	0.9654
$1_{t\geq 0}$	0.0642	0	0	0.5	0.9665
$\mathrm{sign}(t)$	0.1779	0	0	0.4689	0.9715
$1/(1 + e^{-x})$	0.0002	0	0	0.0129	0.9637
$\sin(t)$	0.1779	0	0	0.4689	0.9749
t	1	0	0	1.0021	0.981
$\mathrm{erf}(t)$	0.2166	0	0	0.5053	0.9788
$\cos(t)$	0	0.0003	0	0.0116	0.5257
$\lvert t\rvert$	0	0.0209	0	0.2195	0.5709

Table 2: Classification accuracy and values of α_1, α_2, α_3 and τ at the third and final layer, on GMM data.

$\sigma(t)$	α_1	α_2	α_3	τ	Accuracy
$\max(0, t)$	0.0156	0.0092	0.0099	0.2128	0.8945
$0.1t \cdot 1_{t<0} + t \cdot 1_{t\geq 0}$	0.0083	0.0085	0.0071	0.1867	0.9079
$1_{t\geq 0}$	0.0564	0	0	0.5	0.5028
$\mathrm{sign}(t)$	0.2256	0	0	1	0.4916
$1/(1 + e^{-x})$	0.0002	0	0	0.0135	0.5173
$\sin(t)$	0.1512	0	0	0.4729	0.5025
t	1	0	0	1.0693	0.5045
$\mathrm{erf}(t)$	0.1912	0	0	0.51	0.4989
$\cos(t)$	0	0.0003	0	0.015	0.9598
$\lvert t\rvert$	0	0.0184	0	0.2342	0.9302