Asymptotic Spectral Representation of Linear Convolutional Layers

Xinping Yi, *Member, IEEE*

***Abstract*—By stacking a number of convolutional layers, convolutional neural networks (CNNs) have made remarkable performance boosts in many artificial intelligence applications. While the convolution operation is well-understood, it is still a mystery why repeated convolutions yield so good expressive power and generalization performance. Noting that the linear convolution operation can be represented as a matrix-vector product with the matrix being of a Toeplitz structure, we propose to inspect the individual convolutional layer through its asymptotic spectral representation - the spectral density matrix - by leveraging Toeplitz matrix theory. Thanks to such spectral representation, we are able to develop a simple singular value approximation method with improved accuracy, and spectral norm upper bounds with reduced computational complexity, compared with the state-of-the-art methods. Both the improved approximation and upper bounds can be employed as regularization techniques to further enhance the generalization performance of CNNs. By extensive experiments on well-deployed CNN models (e.g., ResNets), we also demonstrate that the approximation approach achieves higher accuracy and the upper bounds are effective spectral regularizers for generalization.**

***Index Terms*—Convolutional neural networks, spectral representation, Toeplitz matrix, regularization**

I. Introduction

The last decade has witnessed the success of deep learning models, in particular the most popular convolutional neural networks (CNNs), in various artificial intelligence applications, such as computer vision and natural language processing [1]–[3]. With a number of convolutional and fully-connected layers, the state-of-the-art CNN models (e.g., ResNets [4]) are able to learn an over-parameterized model from training data for representing any level of data abstraction, in such a way that the learned models achieve remarkable prediction performance on test data that are unseen during the training. These advantages are attributed to the excellent expressive power and generalization ability of CNN models.

Albeit promising from an empirical point of view, the advances of CNN models are mostly focused on the new network architecture design and parameter tuning. There is still a lack of thorough theoretical understandings as to how and why CNNs achieve so good performance [5]. As such, it is crucial to advance the fundamental understanding of deep learning for the reliable and secure deployment of CNN models in the safety-critical applications, e.g., autonomous driving and healthcare. Recent progress pushing forward this line of research includes a variety of regularization techniques for generalization (e.g., [6]–[9]), and generalization error bounds (e.g., [10]–[12]). It is suggested that a deeper understanding of CNN models requires investigating the structural properties of the individual convolutional layers.

In CNNs, each individual convolutional layer performs a linear transformation from its input to output representations, through a multi-channel multi-dimensional convolution operation. This convolution can be expressed as a matrix-vector product, where the linear transformation matrix comes from convolutional filters and the vector represents the layer's reshaped input. Motivated by spectral analysis of convolution theorem in signal processing using Fourier transform, one may wonder how spectral theory can be applied to linear operators with respect to convolution operations. Spectral analysis of linear operators consists in inspecting the set of their singular/eigenvalues – the spectra of the operators – of the associated matrices.

Spectral properties of weight matrices of fully-connected layers have been analyzed to understand the generalization performance of deep neural networks (DNNs) models, e.g., [13]–[17] for the entire singular value distribution, and e.g., [6], [7], [11], [18]–[22] for the largest singular value (a.k.a. spectral norm), to name just a few. In particular, it has been shown in [15] that the analysis of singular value distribution of layer weight matrices helps reveal the implicit self-regularization of DNN models; and in [20] that generalization error is upper bounded by spectral and Frobenius norms of the layer weight matrices, suggesting that suppressing singular values can therefore enhance the generalization performance.

Whilst such spectral analysis is applicable to CNNs, where the layer weight matrices are more structured due to convolution, recent advances in this thread lie in developing low-complexity spectral regularization methods for generalization by exploiting the structural properties of the linear transformation of convolution. For instance, spectral regularization has been also applied to convolutional layers so as to guide the training process by e.g., clipping singular values within an interval to avoid explosion or vanishing of gradients [16], and bounding spectral norms to enhance generalization performance and robustness against adversarial examples [6], [7], [22].

However, as the size of such linear transformation matrices grows with the input size of the layers, it is computationally challenging to compute the spectra, even for the largest singular values. The straightforward singular value decomposition (SVD) incurs huge computational burden, which is even worse when singular values are required to be computed during the training process to guide spectral regularization and normalization [6], [7]. Fortunately, the structures of the linear

X. Yi is with the Department of Electrical Engineering and Electronics, University of Liverpool, L69 3BX, United Kingdom. Email: xinping.yi@liverpool.ac.uk.

transformation matrices can be exploited to reduce the computational complexity of SVDs. In [16], [17], [22], the linear convolutional layer is treated as a circular one by a "wrapping round" operation. In doing so, the linear transformation matrices are endowed with a circulant structure, by which efficient methods were proposed to compute a circular approximation of the convolutional layers with substantially reduced complexity. To further reduce computational complexity, upper bounds of spectral norm of the circular convolutional layers were derived in [22] at the expense of degraded accuracy.

As a matter of fact, such a "wrapping round" operation is not always endowed in many convolutional layers, for which a linear, rather than circular, convolutional operation is applied. With such a linear convolution, the linear transformation matrix has a Toeplitz structure, which includes the circulant one as a special case. This has been pointed out by a number of previous works, e.g., [23]–[25], that the two-dimensional single-channel convolutional layer results in a doubly block Toeplitz matrix. A question then arises as to how close is the circular approximation to the exact linear Toeplitz case. While this gap has been studied for large Hermitian matrices (e.g., symmetric real matrices) [26], it is still unclear for the *non-Hermitian* matrices, including the linear transformation matrices of linear convolution, which are *asymmetric* real matrices. This motivates the study of the current work.

In this paper, we consider the linear convolutional layers, with main focus on the multi-channel two-dimensional linear convolution with stride size of 1, so that the linear transformation matrix is a block matrix with each block being a doubly Toeplitz matrix. By rows and columns permutation, we construct an alternative representation as a doubly block Toeplitz matrix with each element being a matrix, for which the singular values of both representations are identical. As such, we propose a spectral representation of the linear transformation matrix by a spectral density matrix, by which the spectral analysis of the former can be alternatively done on the latter. Specifically, the main contributions are three-fold:

- The singular value distribution of linear transformation matrix of CNNs is cast to that of its spectral density matrix, thanks to an extension of the celebrated Szegö Theorem for Hermitian Toeplitz matrices to non-Hermitian block doubly Toeplitz matrices. In doing so, the asymptotic spectral analysis of the linear convolutional layers can be alternatively done by inspecting the corresponding spectral density matrix. The circular convolution by "wrapping around" is a special case of such a spectral representation, by which the singular values can also be produced by uniformly sampling the spectral density matrix.
- By treating singular values of the spectral density matrix as random variables, the individual singular value distribution can be quantified by a quantile function. As such, we propose a simple yet effective algorithm to compute singular values of linear convolutional layers by subtly adjusting the singular value distribution obtained from the circular approximation.
- To upper-bound the spectral norm of the linear transformation matrix, we instead upper-bound that of its corresponding spectral density matrix. As a consequence, we come up with three spectral norm bounds that can be used for spectral regularization.

Experimental results demonstrate the superior accuracy of our singular value approximation method and the effectiveness of spectral norm bounds for regularization with respect to generalization in practical CNN models, e.g., ResNets. As a side remark, our proposed spectral representations is different from that in [27], in which the image representations in the pixel domain (i.e., the representations of data rather than the convolutional layers) are transformed to the frequency domain through the Discrete Fourier transforms (DFT). In addition, such a spectral representation of the convolutional layers has been found useful to tackle adversarial robustness for CNNs using Lipschitz regularization [28].

II. Notations and Preliminaries

A. Notations and Definitions

For two integers m and n satisfying $m < n$, define $[m] \triangleq \{1, 2, \dots, m\}$, $n - [m] \triangleq \{n-1, n-2, \dots, n-m\}$, and $[m : n] \triangleq \{m, m+1, \dots, n\}$. Define $\{a_j\}_j \triangleq \{a_j, \forall j\}$. $x \in [a, b]$ is such that $a \le x \le b$. $\jmath$ is the imaginary unit.

Denote by a, $\boldsymbol{a}$, $\boldsymbol{A}$ scalars, vectors, and matrices/tensors, respectively. $\boldsymbol{A}^{\mathsf{T}}$ and $\boldsymbol{A}^{\mathsf{H}}$ represent matrix transpose and Hermitian transpose of $\boldsymbol{A}$, respectively. A complex-valued matrix $\boldsymbol{A}$ is Hermitian if $\boldsymbol{A} = \boldsymbol{A}^{\mathsf{H}}$. If $\boldsymbol{A}$ is real-valued, $\boldsymbol{A}$ is Hermitian is equivalent to $\boldsymbol{A}$ is symmetric, i.e., $\boldsymbol{A} = \boldsymbol{A}^{\mathsf{T}}$. We denote by $\mathrm{blkdiag}(\boldsymbol{A}, \boldsymbol{B}, \dots)$ a block diagonal matrix with diagonal blocks being $\boldsymbol{A}, \boldsymbol{B}, \dots$, and by $\mathrm{circ}(a, b, \dots)$ a circulant matrix with elements in the first row being $a, b, \dots$. Likewise, $\mathrm{circ}(\boldsymbol{A}, \boldsymbol{B}, \dots)$ is the block circulant matrix with first row blocks being $\boldsymbol{A}, \boldsymbol{B}, \dots$. An $n \times n$ matrix $\boldsymbol{F}_n$ is called Discrete Fourier Transform (DFT) matrix, where $[\boldsymbol{F}_n]_{ik} = \frac{1}{\sqrt{n}} e^{-\jmath 2\pi (i-1)(k-1)/n}$ for $i, k \in [n]$. $\boldsymbol{I}_n$ is the $n \times n$ identity matrix. For a tensor $\boldsymbol{A}$, $\mathrm{vec}(\boldsymbol{A})$ denotes the vectorized version of $\boldsymbol{A}$, and for a 4-order tensor $\boldsymbol{A}$, $\boldsymbol{A}_{i,j,k,l}$ is used to index its elements.

Denote by $\otimes$ the Kronecker product between two matrices. For a scalar k, it holds $\boldsymbol{A} \otimes (k\boldsymbol{B}) = k(\boldsymbol{A} \otimes \boldsymbol{B})$ and $\boldsymbol{A} \otimes (\sum_i \boldsymbol{B}_i) = \sum_i \boldsymbol{A} \otimes \boldsymbol{B}_i$. For two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, $\boldsymbol{A} \otimes \boldsymbol{B}$ is permutation equivalent to $\boldsymbol{B} \otimes \boldsymbol{A}$, i.e., there exist permutation matrices $\boldsymbol{\Pi}_1$ and $\boldsymbol{\Pi}_2$ such that $\boldsymbol{B} \otimes \boldsymbol{A} = \boldsymbol{\Pi}_1 (\boldsymbol{A} \otimes \boldsymbol{B}) \boldsymbol{\Pi}_2$.

For a matrix $\boldsymbol{A} = (a_{ij})_{i,j=1}^{m,n}$ with $\mathrm{rank}(\boldsymbol{A}) = r$, we denote by $\{\sigma_j(\boldsymbol{A})\}_j$ the collection of singular values of $\boldsymbol{A}$ arranged in descending order, i.e., $\sigma_1(\boldsymbol{A}) \ge \sigma_2(\boldsymbol{A}) \ge \cdots \ge \sigma_r(\boldsymbol{A})$. The norm $\|\boldsymbol{A}\|_2 \triangleq \sigma_1(\boldsymbol{A})$ is called spectral norm. The Schatten p-norm is defined as $\|\boldsymbol{A}\|_p \triangleq (\sum_{j=1}^r \sigma_j^p(\boldsymbol{A}))^{\frac{1}{p}}$. When $p = 2$, it coincides with Frobenius norm $\|\boldsymbol{A}\|_{\mathrm{F}} \triangleq \sqrt{\sum_{i=1}^m \sum_{j=1}^n |a_{ij}|^2} = \sqrt{\sum_{j=1}^r \sigma_j^2(\boldsymbol{A})}$. The matrix ℓ_1 and ℓ_∞ norms are defined as $\|\boldsymbol{A}\|_1 \triangleq \max_j \sum_{i=1}^m |a_{ij}|$ and $\|\boldsymbol{A}\|_\infty \triangleq \max_i \sum_{j=1}^n |a_{ij}|$, respectively. $|a|$ is the absolute value or modulus of a scalar a.

A matrix-valued function $F : [a, b]^k \to \mathbb{C}^{m \times n}$ is such that $F(\boldsymbol{x}) \in \mathbb{C}^{m \times n}$ for $\boldsymbol{x} \in [a, b]^k$. F is Lebesgue measurable (resp. bounded, continuous) in $[a, b]^k$ if each of its element F_{ij} is Lebesgue measurable (resp. bounded, continuous) in $[a, b]^k$. For $F : [-\pi, \pi]^2 \to \mathbb{C}^{m \times n}$, its L_p-norm is

defined as $\|F\|_{L_p} \triangleq (\frac{1}{(2\pi)^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\|F(\omega_1,\omega_2)\|_{\mathrm{F}}^p d\omega_1 d\omega_2)^{\frac{1}{p}}$ for $p \geq 1$. In particular, for $\boldsymbol{\omega} = (\omega_1, \omega_2) \in [-\pi, \pi]^2$, $\|F\|_{L_\infty} = \sup_{\boldsymbol{\omega}}\|F(\boldsymbol{\omega})\|_{\mathrm{F}}$. When $F \in \mathcal{L}_2([-\pi,\pi]^2)$, it means $\|F\|_{L_2}^2 < +\infty$. Define $\|F\|_p \triangleq \sup_{\boldsymbol{\omega}}\|F(\boldsymbol{\omega})\|_p$ for all $p > 0$, e.g., $\|F\|_2 \triangleq \sup_{\boldsymbol{\omega}}\|F(\boldsymbol{\omega})\|_2$.

B. Toeplitz and Circulant Matrices

A Toeplitz matrix $\boldsymbol{T} = [t_{i-j}]_{i,j=1}^n$ is an $n \times n$ matrix for which the entries come from a sequence $\{t_k, k = 0, \pm1, \pm2, \ldots, \pm(n-1)\}$. A circulant matrix is a special Toeplitz matrix, where $\boldsymbol{C} = [t_{(i-j) \bmod n}]_{i,j=1}^n$. That is, $t_{-k} = t_{n-k}$ for $k = 1, 2, \ldots, n-1$. Denote the circulant matrix by $\boldsymbol{C} = \mathrm{circ}(t_0, t_{-1}, \ldots, t_{-(n-1)})$ using its first row, where the rest is cyclic shift of the first row with n times.

An $m \times m$ block Toeplitz matrix $\boldsymbol{B} = [\boldsymbol{A}_{i-j}]_{i,j=1}^m \in \mathbb{C}^{mp\times mq}$ is a Toeplitz matrix with each element being a $p \times q$ matrix. Similarly, the block circulant matrix $\boldsymbol{C}$ is such that $\boldsymbol{C} = [\boldsymbol{A}_{(i-j) \bmod m}]_{i,j=1}^m$ with $0 \bmod m = m \bmod m = 0$. That is, $\boldsymbol{A}_{-k} = \boldsymbol{A}_{m-k}$ for $k = 1, 2, \ldots, m-1$, such that $\boldsymbol{C} = \mathrm{circ}(\boldsymbol{A}_0, \boldsymbol{A}_{-1}, \ldots, \boldsymbol{A}_{-(m-1)})$ and the rest row blocks are block-wise cyclic shift of the first row block.

When $\{\boldsymbol{A}_k, k = 0, \pm1, \ldots, \pm(m-1)\}$ are also $n \times n$ Toeplitz/circulant matrices, $\boldsymbol{B}$ is a block Toeplitz/circulant matrix with Toeplitz/circulant blocks, which is also known as doubly Toeplitz/circulant matrix.

A banded (block) Toeplitz matrix is a special Toeplitz matrix $\boldsymbol{T}$ [resp. $\boldsymbol{B}$] such that $t_k = 0$ [resp. $\boldsymbol{A}_k = \boldsymbol{0}$] when $k > r$ or $k < -s$ for some $1 < r, s < n$ [resp. $1 < r, s < m$].

III. Convolutional Neural Networks

A. Linear Convolutional Layer

We consider multiple-channel two-dimensional *linear* convolutional layers with *arbitrary padding* schemes in CNNs before applying activation functions and pooling. For ease of presentation, we first consider the stride size 1, and the extension to larger stride size will be discussed in Section VI.

Let the input be $\boldsymbol{X} \in \mathbb{R}^{c_{in}\times n\times n}$ and the linear convolutional filter be $\boldsymbol{K} \in \mathbb{R}^{c_{out}\times c_{in}\times h\times w}$ with $h, w \leq n$, where n, h, w, c_{in}, c_{out} are input size, filter height, filter width, the numbers of input and output channels, respectively. For convenience, we let the output $\boldsymbol{Y}$ have the same size as the input $\boldsymbol{X}$ by arbitrary padding strategies, and reuse $\boldsymbol{X}$ as the input with padding. By applying linear convolution of the filter $\boldsymbol{K}$ to the input $\boldsymbol{X}$, the output $\boldsymbol{Y} \in \mathbb{R}^{c_{out}\times n\times n}$ is given by

$$\boldsymbol{Y}_{c,r,s} = \sum_{d=1}^{c_{in}}\sum_{p=1}^{n}\sum_{q=1}^{n}\boldsymbol{X}_{d,r+p,s+q}\boldsymbol{K}_{c,d,p,q} \tag{1}$$

for $r, s \in [n]$ and $c \in [c_{out}]$ where $\boldsymbol{K}_{c,d,p,q} = 0$ if p, q exceed the ranges of h, w. A compact form of the above input-output relation can be rewritten as

$$\mathrm{vec}(\boldsymbol{Y}) = \boldsymbol{A}\mathrm{vec}(\boldsymbol{X}), \tag{2}$$

where $\boldsymbol{A} \in \mathbb{R}^{c_{out}n^2\times c_{in}n^2}$ is the linear transformation matrix of the convolutional layer. For the general case with multiple-input and multiple-output channels, the linear transformation can be represented as a $c_{out} \times c_{in}$ block matrix, i.e.,

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{1,1} & \boldsymbol{A}_{1,2} & \cdots & \boldsymbol{A}_{1,c_{in}} \\ \boldsymbol{A}_{2,1} & \boldsymbol{A}_{2,2} & \cdots & \boldsymbol{A}_{2,c_{in}} \\ \vdots & \vdots & & \vdots \\ \boldsymbol{A}_{c_{out},1} & \boldsymbol{A}_{c_{out},2} & \cdots & \boldsymbol{A}_{c_{out},c_{in}} \end{bmatrix}, \tag{3}$$

where each block $\boldsymbol{A}_{c,d}$ is a banded block Toeplitz matrix, i.e., $[\boldsymbol{A}_{c,d}]_{i_1,j_1} = \boldsymbol{A}_{i_1-j_1}^{c,d}$ for $-h_1 \leq i_1 - j_1 \leq h_2$. Specifically,

$$\boldsymbol{A}_{c,d} = \begin{bmatrix} \boldsymbol{A}_0^{c,d} & \cdots & \boldsymbol{A}_{-h_1}^{c,d} & 0 & \ldots & 0 \\ \vdots & \boldsymbol{A}_0^{c,d} & \ddots & \ddots & \ddots & \vdots \\ \boldsymbol{A}_{h_2}^{c,d} & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \boldsymbol{A}_{-h_1}^{c,d} \\ \vdots & \ddots & \ddots & \ddots & \boldsymbol{A}_0^{c,d} & \vdots \\ 0 & \cdots & 0 & \boldsymbol{A}_{h_2}^{c,d} & \cdots & \boldsymbol{A}_0^{c,d} \end{bmatrix} \tag{4}$$

where h_1, h_2 depend on the size of padding in height subject to $h = h_1 + h_2 + 1$. Each block $\boldsymbol{A}_k^{c,d}$ is still a banded Toeplitz matrix, i.e., $[\boldsymbol{A}_k^{c,d}]_{i_2,j_2} = a_{k,i_2-j_2}^{c,d}$ for $-w_1 \leq i_2 - j_2 \leq w_2$. Specifically,

$$\boldsymbol{A}_k^{c,d} = \begin{bmatrix} a_{k,0}^{c,d} & \cdots & a_{k,-w_1}^{c,d} & 0 & \cdots & 0 \\ \vdots & a_{k,0}^{c,d} & \ddots & \ddots & \ddots & \vdots \\ a_{k,w_2}^{c,d} & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & a_{k,-w_1}^{c,d} \\ \vdots & \ddots & \ddots & \ddots & a_{k,0}^{c,d} & \vdots \\ 0 & \cdots & 0 & a_{k,w_2}^{c,d} & \cdots & a_{k,0}^{c,d} \end{bmatrix} \tag{5}$$

where w_1, w_2 subject to $w_1 + w_2 + 1 = w$ that are determined by the size of padding in width. The elements in $\boldsymbol{A}_k^{c,d}$ are weights in the filter [cf. (6)]. In matrix analysis, $\boldsymbol{A}$ is usually referred to as multi-block multi-level (doubly) Toeplitz matrix. For $k \in [-h_1 : h_2]$ and $l \in [-w_1 : w_2]$, we have

$$a_{k,l}^{c,d} = \boldsymbol{K}_{c,d,h_1+k+1,w_1+l+1}, \tag{6}$$

for all $c \in [c_{out}]$ and $d \in [c_{in}]$.

B. Alternative Representation

For ease of spectral analysis, we transform $\boldsymbol{A}$ into a multi-level block Toeplitz matrix (whose entries of the last level are matrices) via vec-permutation operation [29], for which the matrix spectrum keeps unchanged.

Denote by $\boldsymbol{T} \in \mathbb{R}^{c_{out}n^2\times c_{in}n^2}$ the alternative representation as a block Toeplitz matrix with $[\boldsymbol{T}]_{i_1,j_1} = \boldsymbol{T}_{i_1-j_1}$ for $i_1 \leq h_2 + 1$ and $j_1 \leq h_1 + 1$, that is,

$$\boldsymbol{T} = \begin{bmatrix} \boldsymbol{T}_0 & \cdots & \boldsymbol{T}_{-h_1} & 0 & \ldots & 0 \\ \vdots & \boldsymbol{T}_0 & \ddots & \ddots & \ddots & \vdots \\ \boldsymbol{T}_{h_2} & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \boldsymbol{T}_{-h_1} \\ \vdots & \ddots & \ddots & \ddots & \boldsymbol{T}_0 & \vdots \\ 0 & \cdots & 0 & \boldsymbol{T}_{h_2} & \cdots & \boldsymbol{T}_0 \end{bmatrix} \tag{7}$$

with each block $\boldsymbol{T}_k$ for all $k \in [-h_1 : h_2]$ being still a block Toeplitz matrix $[\boldsymbol{T}_k]_{i_2,j_2} = \boldsymbol{T}_{k,i_2-j_2}$ for $i_2 \leq w_2+1$ and $j_2 \leq w_1+1$, that is,

$$\boldsymbol{T}_k = \begin{bmatrix} \boldsymbol{T}_{k,0} & \cdots & \boldsymbol{T}_{k,-w_1} & 0 & \ldots & 0 \\ \vdots & \boldsymbol{T}_{k,0} & \ddots & \ddots & \ddots & \vdots \\ \boldsymbol{T}_{k,w_2} & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \boldsymbol{T}_{k,-w_1} \\ \vdots & \ddots & \ddots & \ddots & \boldsymbol{T}_{k,0} & \vdots \\ 0 & \cdots & 0 & \boldsymbol{T}_{k,w_2} & \cdots & \boldsymbol{T}_{k,0} \end{bmatrix} \tag{8}$$

where each block $\boldsymbol{T}_{k,l} \in \mathbb{R}^{c_{out} \times c_{in}}$ with $l \in [-w_1 : w_2]$ is

$$\boldsymbol{T}_{k,l} = \begin{bmatrix} t_{1,1}^{k,l} & t_{1,2}^{k,l} & \cdots & t_{1,c_{in}}^{k,l} \\ t_{2,1}^{k,l} & t_{2,2}^{k,l} & \cdots & t_{2,c_{in}}^{k,l} \\ \vdots & \ddots & \ddots & \vdots \\ t_{c_{out},1}^{k,l} & t_{c_{out},2}^{k,l} & \cdots & t_{c_{out},c_{in}}^{k,l} \end{bmatrix}. \tag{9}$$

By such an alternative representation, we have

$$t_{c,d}^{k,l} = \boldsymbol{K}_{c,d,h_1+k+1,w_1+l+1} = a_{k,l}^{c,d}, \tag{10}$$

for all $c \in [c_{out}]$ and $d \in [c_{in}]$.

In what follows, we show that the alternative representation $\boldsymbol{T}$ of the linear convolutional layers has the identical spectrum structure as the original form $\boldsymbol{A}$.

Lemma 1. $\{\sigma_j(\boldsymbol{T}),\ \forall j\} = \{\sigma_j(\boldsymbol{A}),\ \forall j\}$.

Proof. Let $\boldsymbol{e}_i$ be i-th column of identity matrix and $\boldsymbol{E}_{i,j} = \boldsymbol{e}_i \boldsymbol{e}_j^\top$ be a $c_{out} \times c_{in}$ matrix with only the (i,j)-th element being 1 and 0 elsewhere. Define $\boldsymbol{P}_k$ as an $n \times n$ matrix with $[\boldsymbol{P}_k]_{i,j} = 1$ if $i - j = k$ and 0 otherwise. Thus, the original linear transformation matrix $\boldsymbol{A}$ can be represented as

$$\begin{aligned} \boldsymbol{A} &= \sum_{c=1}^{c_{out}} \sum_{d=1}^{c_{in}} \boldsymbol{E}_{c,d} \otimes \boldsymbol{A}_{c,d} && (11) \\ &= \sum_{c=1}^{c_{out}} \sum_{d=1}^{c_{in}} \boldsymbol{E}_{c,d} \otimes \Big(\sum_{k=-h_1}^{h_2} \boldsymbol{P}_k \otimes \boldsymbol{A}_k^{c,d} \Big) && (12) \\ &= \sum_{c=1}^{c_{out}} \sum_{d=1}^{c_{in}} \boldsymbol{E}_{c,d} \otimes \Big(\sum_{k=-h_1}^{h_2} \boldsymbol{P}_k \otimes \Big(\sum_{l=-w_1}^{w_2} \boldsymbol{P}_l \otimes a_{k,l}^{c,d} \Big) \Big) && (13) \\ &= \sum_{c=1}^{c_{out}} \sum_{d=1}^{c_{in}} \sum_{k=-h_1}^{h_2} \sum_{l=-w_1}^{w_2} a_{k,l}^{c,d} \boldsymbol{E}_{c,d} \otimes \boldsymbol{P}_k \otimes \boldsymbol{P}_l && (14) \end{aligned}$$

where the last equality is because $a_{k,l}^{c,d}$ is a scalar. The alternative one $\boldsymbol{T}$ can be represented as

$$\begin{aligned} \boldsymbol{T} &= \sum_{k=-h_1}^{h_2} \boldsymbol{P}_k \otimes \boldsymbol{T}_k && (15) \\ &= \sum_{k=-h_1}^{h_2} \boldsymbol{P}_k \otimes \Big(\sum_{l=-w_1}^{w_2} \boldsymbol{P}_l \otimes \boldsymbol{T}_{k,l} \Big) && (16) \\ &= \sum_{k=-h_1}^{h_2} \boldsymbol{P}_k \otimes \Big(\sum_{l=-w_1}^{w_2} \boldsymbol{P}_l \otimes \Big(\sum_{c=1}^{c_{out}} \sum_{d=1}^{c_{in}} t_{c,d}^{k,l} \boldsymbol{E}_{c,d} \Big) \Big) && (17) \\ &= \sum_{c=1}^{c_{out}} \sum_{d=1}^{c_{in}} \sum_{k=-h_1}^{h_2} \sum_{l=-w_1}^{w_2} t_{c,d}^{k,l} \boldsymbol{P}_k \otimes \boldsymbol{P}_l \otimes \boldsymbol{E}_{c,d} && (18) \end{aligned}$$

where the last equality is because $t_{c,d}^{k,l}$ is a scalar.

According to [29], $\boldsymbol{P}_k \otimes \boldsymbol{P}_l \otimes \boldsymbol{E}_{c,d}$ is permutation equivalent to $\boldsymbol{E}_{c,d} \otimes \boldsymbol{P}_k \otimes \boldsymbol{P}_l$, for which there exist two permutation matrices $\boldsymbol{\Pi}_1$ and $\boldsymbol{\Pi}_2$, such that $\boldsymbol{P}_k \otimes \boldsymbol{P}_l \otimes \boldsymbol{E}_{c,d} = \boldsymbol{\Pi}_1 (\boldsymbol{E}_{c,d} \otimes \boldsymbol{P}_k \otimes \boldsymbol{P}_l) \boldsymbol{\Pi}_2$. Given the fact that $a_{k,l}^{c,d} = t_{c,d}^{k,l}$, it follows that

$$\boldsymbol{T} = \boldsymbol{\Pi}_1 \boldsymbol{A} \boldsymbol{\Pi}_2. \tag{19}$$

Because permutation matrices are also orthogonal matrices, and thus unitary, $\boldsymbol{T}$ and $\boldsymbol{A}$ have an identical set of singular values. This completes the proof. □

Lemma 1 says the block matrix with doubly Toeplitz matrix blocks (i.e., $\boldsymbol{A}$) has the same set of singular values as the block doubly Toeplitz matrix (i.e., $\boldsymbol{T}$). This holds for any Toeplitz matrices which are not necessarily banded, and for any multi-level case but not limited to doubly Toeplitz case.

C. Circular Approximation

The "wrapping around" operation makes linear transformation a circular convolution, which is deemed as a circular approximation of linear convolution. As $h, w \leq n$, we can construct a circulant matrix by "wrapping around" to assist the spectral analysis.

Given the doubly block Toeplitz matrix $\boldsymbol{T} = [\boldsymbol{T}_{i-j}]_{i,j=1}^n$ with $\boldsymbol{T}_k = 0$ if $k > h_2$ or $k < -h_1$ and $\boldsymbol{T}_k = [\boldsymbol{T}_{k,p-q}]_{p,q=1}^n$ with $\boldsymbol{T}_{k,l} = 0$ if $l > w_2$ or $l < -w_1$, the doubly block circulant matrix $\boldsymbol{C} = \mathrm{circ}(\boldsymbol{C}_0, \boldsymbol{C}_1, \ldots, \boldsymbol{C}_{n-1})$ is as follows

$$\boldsymbol{C}_k = \begin{cases} \boldsymbol{T}_{-k}, & k \in \{0\} \cup [h_1] \\ \boldsymbol{T}_{n-k}, & k \in n - [h_2] \\ 0, & \text{otherwise} \end{cases} \tag{20}$$

where $\boldsymbol{C}_k = \mathrm{circ}(\boldsymbol{C}_{k,0}, \boldsymbol{C}_{k,1}, \ldots, \boldsymbol{C}_{k,n-1})$ with

$$\boldsymbol{C}_{k,l} = \begin{cases} \boldsymbol{T}_{-k,-l}, & k \in \{0\} \cup [h_1],\ l \in \{0\} \cup [w_1] \\ \boldsymbol{T}_{-k,n-l}, & k \in \{0\} \cup [h_1],\ l \in n - [w_2] \\ \boldsymbol{T}_{n-k,-l}, & k \in n - [h_2],\ l \in \{0\} \cup [w_1] \\ \boldsymbol{T}_{n-k,n-l}, & k \in n - [h_2],\ l \in n - [w_2] \\ 0, & \text{otherwise} \end{cases} \tag{21}$$

where $\boldsymbol{T}_{k,l}$ is defined in (9).

In a similar way, the original block doubly Toeplitz matrix $\boldsymbol{A}$ can also have a corresponding block doubly circulant matrix $\boldsymbol{C}(\boldsymbol{A}) = [\boldsymbol{C}(\boldsymbol{A}_{c,d})]_{c,d=1}^{c_{out},c_{in}}$ where

$$\begin{aligned} \boldsymbol{C}(\boldsymbol{A}_{c,d}) = \mathrm{circ}(&\boldsymbol{C}(\boldsymbol{A}_0^{c,d}), \boldsymbol{C}(\boldsymbol{A}_{-1}^{c,d}), \ldots, \boldsymbol{C}(\boldsymbol{A}_{-h_1}^{c,d}), 0, \\ &\ldots, 0, \boldsymbol{C}(\boldsymbol{A}_{h_2}^{c,d}), \ldots, \boldsymbol{C}(\boldsymbol{A}_1^{c,d})) \end{aligned} \tag{22}$$

with $\boldsymbol{C}(\boldsymbol{A}_{c,d}) \in \mathbb{R}^{n^2 \times n^2}$ where

$$\boldsymbol{C}(\boldsymbol{A}_k^{c,d}) = \mathrm{circ}(a_{k,0}^{c,d}, \ldots, a_{k,-w_1}^{c,d}, 0, \ldots, 0, a_{k,w_2}^{c,d}, \ldots, a_{k,1}^{c,d})$$

with $\boldsymbol{C}(\boldsymbol{A}_k^{c,d}) \in \mathbb{R}^{n \times n}$. Similarly to Lemma 1, we have the following lemma.

Lemma 2. $\{\sigma_j(\boldsymbol{C}),\ \forall j\} = \{\sigma_j(\boldsymbol{C}(\boldsymbol{A})),\ \forall j\}$.

Proof. The proof is similar to that of Lemma 1 and thus omitted. The only difference is that, for the representation of $n \times n$ circulant matrices, we have $[\boldsymbol{P}_k]_{i,j} = 1$ if $(i-j) \bmod n = k$ and 0 otherwise. □

It can be easily verified that $\boldsymbol{C}(\boldsymbol{A})$ is essentially the linear transformation matrix of circular convolutional layers considered in [16]. As a byproduct of Lemma 2, we present an alternative calculation of the singular values for the circular convolutional layers that were characterized in [16].

Lemma 3. *The linear transformation matrix $\boldsymbol{C}(\boldsymbol{A})$ can be block-diagonalized as*

$$\boldsymbol{C} = (\boldsymbol{F}_n \otimes \boldsymbol{F}_n \otimes \boldsymbol{I}_{c_{out}}) \text{blkdiag}(\boldsymbol{B}_{1,1}, \boldsymbol{B}_{1,2}, \ldots, \boldsymbol{B}_{1,n}, \boldsymbol{B}_{2,1}, \ldots, \boldsymbol{B}_{n,n}) (\boldsymbol{F}_n \otimes \boldsymbol{F}_n \otimes \boldsymbol{I}_{c_{in}})^{\mathsf{H}} \quad (23)$$

where both $(\boldsymbol{F}_n \otimes \boldsymbol{F}_n \otimes \boldsymbol{I}_{c_{out}})$ and $(\boldsymbol{F}_n \otimes \boldsymbol{F}_n \otimes \boldsymbol{I}_{c_{in}})$ are unitary matrices. Thus, the singular values of $\boldsymbol{C}(\boldsymbol{A})$ are the collection of singular values of $\{\boldsymbol{B}_{i,k}\}_{i,k=1}^n$ where

$$\boldsymbol{B}_{i,k} = \sum_{p=0}^{n-1} \sum_{q=0}^{n-1} \boldsymbol{C}_{p,q} e^{-\jmath 2\pi \frac{p(i-1)+q(k-1)}{n}} \quad (24)$$

with $\boldsymbol{C}_{p,q}$ defined in (21).

Proof. By extending Lemma 5.1 in [30] from block circulant matrices to doubly block circulant matrices, we conclude that $\boldsymbol{C}$ can be block-diagonalized as in (23). As such, the singular values of $\boldsymbol{C}$ are the collection of singular values of n^2 matrices $\{\boldsymbol{B}_{i,k}\}_{i,k=1}^n$. By Lemma 5.1 in [30], for each $i,k \in [n]$, we compute $\boldsymbol{B}_{i,k} \in \mathbb{C}^{c_{out} \times c_{in}}$ by (24). The singular values of $\boldsymbol{B}_{i,k}$ can be therefore obtained by applying off-the-shelf singular-value decomposition algorithms. □

The computation of $\boldsymbol{B}_{i,k}$ can be seen as a two-dim DFT of $\boldsymbol{C}_{p,q}$. With hw non-zero submatrices $\{\boldsymbol{C}_{p,q}\}$, the computational complexity consists in hw FFTs and n^2 SVDs, which is identical to that in [16]. We also point out that this alternative approach essentially has the same flavor as that in [17].

Given Lemmas 1-3, we hereafter take $\boldsymbol{T}$ as the linear transformation matrix of the linear convolutional layer and $\boldsymbol{C}$ as its circular approximation, for asymptotic spectral analysis.

IV. Asymptotic Spectral Representation

For a convolutional layer with input size n and s input and r output channels, the corresponding linear transformation matrix $\boldsymbol{T} \in \mathbb{C}^{rn^2 \times sn^2}$ is challenging to analyze due to the high dimensionality as the input size n increases. For instance, a typical convolutional layer with filter size $64 \times 3 \times 3 \times 3$ and input size $3 \times 224 \times 224$ has $\boldsymbol{T}$ of size $3,211,264 \times 150,528$, for which matrix analysis is prohibitively intractable and expensive.

To make it more tractable, in what follows, we present an asymptotic spectral representation of $\boldsymbol{T}$, taking advantage of its Toeplitz structure [31]–[39].

Theorem 1. *Given the Toeplitz matrix $\boldsymbol{T} \in \mathbb{C}^{rn^2 \times sn^2}$, let a complex matrix-valued Lebesgue-measurable function $F : [-\pi,\pi]^2 \to \mathbb{C}^{r \times s}$ be the generating function such that*

$$\boldsymbol{T}_{k,l} = \frac{1}{(2\pi)^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} F(\omega_1,\omega_2) e^{-\jmath(k\omega_1 + l\omega_2)} d\omega_1 d\omega_2. \quad (25)$$

(Equal Spectral Distribution) *It follows that, for any continuous function Φ with compact support in $\mathbb{R}$, we have*

$$\lim_{n\to\infty} \frac{1}{n^2} \sum_{j=1}^{\min\{r,s\}n^2} \Phi(\sigma_j(\boldsymbol{T})) = \frac{1}{(2\pi)^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \sum_{j=1}^{\min\{r,s\}} \Phi(\sigma_j(F(\omega_1,\omega_2))) d\omega_1 d\omega_2, \quad (26)$$

where $\sigma_j(\boldsymbol{T})$ is the j-th singular value of $\boldsymbol{T}$ and $\sigma_j(F(\omega_1,\omega_2))$ is the j-th singular value function of F with respect to (ω_1,ω_2). As such, $\boldsymbol{T}$ is said to be equally distributed as $F(\omega_1,\omega_2)$ with respect to singular values, i.e., $\boldsymbol{T} \sim_\sigma F$.

(Spectral Representation) *For linear convolutional layers, $\boldsymbol{T}$ has doubly banded structures, so that the generating function can be explicitly written as*

$$F(\omega_1,\omega_2) = \sum_{k=-h_1}^{h_2} \sum_{l=-w_1}^{w_2} \boldsymbol{T}_{k,l} e^{\jmath(k\omega_1 + l\omega_2)}, \quad (27)$$

which is also referred to as the spectral density matrix of $\boldsymbol{T}$.

Proof. The proof is an extension of those in [34]–[37] that consider block Toeplitz matrices or doubly Toeplitz matrices. The main proof technique is to relate the non-Hermitian block doubly Toeplitz matrix to a properly constructed circulant counterpart. The construction of such non-Hermitian block doubly circulant matrix is due to the circulant approximation in Section III-C inspired by the circular convolution operation. In particular, we follow the footsteps of [35], [36] to extend the proofs to non-Hermitian block doubly Toeplitz matrices $\boldsymbol{T}$, by relating to the block doubly circulant matrices $\boldsymbol{C}$. The complete proof can be found in the arXiv version [40]. □

Remark 1. *The equal spectral distribution in Theorem 1 describes the identical collective behavior of singular values of $\boldsymbol{T}$ and $F(\omega_1,\omega_2)$ when n tends to infinity. That is, all $\min\{r,s\}n^2$ singular values of $\boldsymbol{T}$ converge to $\min\{r,s\}$ singular value distributions $\{\sigma_j(F(\omega_1,\omega_2))\}$ asymptotically. It is a generalization of the celebrated Szegö Theorem [31], which deals with real scalar-valued generating functions $F : [-\pi,\pi] \to \mathbb{R}$ that correspond to Hermitian Toeplitz matrices. It was extended to non-Hermitian matrices [32], [33], block Toeplitz matrices [35], and multi-level Toeplitz matrices [34], [37]. In CNNs, the matrix $\boldsymbol{T}$ is an asymmetric real matrix and hence non-Hermitian, with doubly block Toeplitz structure, which corresponds to a complex matrix-valued generating function $F : [-\pi,\pi]^2 \to \mathbb{C}^{r \times s}$. In particular, when $s = r = 1$, $\boldsymbol{T}$ reduces to a single-channel 2D convolutional layer matrix, for which $\boldsymbol{T} \sim_\sigma |F(\omega_1,\omega_2)|$; When it comes to a signal-channel 1D convolutional layer, Theorem 1 indicates $\boldsymbol{T} \sim_\sigma |F(\omega)|$, with similar spectral representations applicable.*

Theorem 1 endows the linear convolutional layer with an asymptotic spectral representation - the spectral density matrix $F(\omega_1,\omega_2)$ - by establishing the collective equivalence of their asymptotic singular value distributions. In particular, under the context of spectral analysis, the collection of convolutional filters $\{\boldsymbol{T}_{k,l} \in \mathbb{C}^{r \times s}\}$ and the spectral density

matrix $F(\omega_1,\omega_2)\in\mathbb{C}^{r\times s}$ are a pair of Fourier transform in the matrix form. That is, each element in $F(\omega_1,\omega_2)$ is the two-dimensional Fourier transform of the collection of the corresponding element in $\boldsymbol{T}_{k,l}$ for all k,l. As such, the analysis of singular values of $\boldsymbol{T}$ with size $rn^2\times sn^2$ can be alternatively done on its spectral representation $F(\omega_1,\omega_2)$ with size $r\times s$, which significantly reduces the computational complexity. For instance, the spectral analysis of the aforementioned $\boldsymbol{T}$ of size $3,211,264\times 150,528$ can be done on $F(\omega_1,\omega_2)$ of size 64×3.

Further, as singular values of $\boldsymbol{T}$ converges to singular value functions of $F(\omega_1,\omega_2)$ when n tends to infinity, we can interpret the former as the samples of the latter over $(\omega_1,\omega_2)\in[-\pi,\pi]^2$. The $\min\{r,s\}n^2$ singular values of $\boldsymbol{T}$ can be clustered into $\min\{r,s\}$ non-overlapping subsets, each of which contains roughly n^2 ones. When n is sufficiently large, the singular values in the j-th subset concentrate on $\sigma_j(F)$, where $\sigma_j(F)$ is the j-th singular value function of $F(\omega_1,\omega_2)$. It suggests that, the singular value of the circulant matrix $\boldsymbol{C}$, a circular approximation of the Toeplitz matrix $\boldsymbol{T}$, can be approximately obtained by sampling $\sigma_j(F)$, the j-th singular value function of $F(\omega_1,\omega_2)$, over a uniform gird in $[-\pi,\pi]^2$, for all $j\in[\min\{r,s\}]$. This will be formally stated in Theorem 2 as follows.

Theorem 2. *Consider the Toeplitz matrix $\boldsymbol{T}$ and its circular approximation $\boldsymbol{C}$ as in (20)-(21).*

***(Uniform Sampling)** The singular values of $\boldsymbol{C}$ are the collection of singular values of all spectral density matrices $F(\omega_1,\omega_2)$ sampled on a uniform grid*

$$(\omega_1,\omega_2)=(-\pi+\frac{2\pi j_1}{n},-\pi+\frac{2\pi j_2}{n}),\quad \forall j_1,j_2\in[n]-1. \tag{28}$$

***(Bounded Average Difference)** There exists a constant $c_1>0$ such that*

$$\lim_{n\to\infty}\frac{1}{n}\sum_{j=1}^{\min\{r,s\}n^2}|\sigma_j(\boldsymbol{T})-\sigma_j(\boldsymbol{C})|\le c_1, \tag{29}$$

which indicates that the average *difference of overall singular values of the circular approximation $\boldsymbol{C}$ from the exact ones of $\boldsymbol{T}$ is bounded by $O(\frac{1}{n})$,[1] and tends to zero as n increases.*

Proof. See Appendix A. □

Remark 2. *The block diagonal matrices $\boldsymbol{B}_{i,k}$ of $\boldsymbol{C}$ in (23) is essentially the matrix-valued function $F(\omega_1,\omega_2)$ with uniform sampling on grids as in (28), i.e.,*

$$\boldsymbol{B}_{j_1,j_2}=F\Big(\frac{2\pi(j_1-1)}{n},\frac{2\pi(j_2-1)}{n}\Big),\quad \forall j_1,j_2\in[n]. \tag{30}$$

This also confirms that the circular approximation could come from the uniform sampling on the spectral density matrix $F(\omega_1,\omega_2)$.

Remark 3. *Theorem 2 shows that the singular values of the circular approximation of the linear convolution can be alternatively obtained by computing singular values on the uniformly sampled spectral density matrix $F(\omega_1,\omega_2)$ over $(\omega_1,\omega_2)\in[-\pi,\pi]^2$. To collect all singular values (resp. spectral norm) of $\boldsymbol{C}$, it requires to compute singular values (resp. spectral norm) of n^2 matrices with size $r\times s$ each, with the computational complexity identical to those in [16], [22]. Although there is a vanishing average difference of all singular values between $\boldsymbol{T}$ and $\boldsymbol{C}$, the accuracy of circular approximation for each individual singular value is not guaranteed, where the difference could scale as n.*

As a side remark, the bounded difference of singular values between a specific family of Toeplitz matrices and their circular approximations has been observed in the literature, whilst Theorem 2 generalizes it to a wider family of Toeplitz matrices. In particular, as [26] dealt with eigenvalues of Hermitian Toeplitz matrices that correspond to real scalar-valued generating functions, the bounded difference results in [26, Theorem 2] can not be taken as granted to justify that of linear convolutional layers with non-Hermitian block doubly Toeplitz matrices. For non-Hermitian matrices, Theorem 2 only guarantees the bounded *average* difference of all singular values but not the difference of each *individual* singular value between Toeplitz and circulant matrices.

V. Applications of Spectral Representation

To demonstrate the usefulness of the asymptotic spectral representations in Theorem 1, in this section, we present two applications to (1) approximate the singular values of convolutional layers, and (2) to upper bound spectral norm of convolutional layers for the use of spectral norm regularization during training. The effectiveness will be verified with experiments in Section VII.

A. Singular Value Approximation

As Theorem 2 implies that the simple circular approximation of singular values with uniform sampling may not guarantee bounded difference of individual singular values, one may think of a non-uniform sampling of the spectral density matrix for a better approximation.

Before proceeding further, let us first see what the equal spectral distribution in Theorem 1 implies with respect to sampling. Collecting all singular values $\{\sigma_j(F)\}_j$ according to the uniform sampling grids as in (28), we sort them in non-decreasing order as $(\kappa_1,\kappa_2,\ldots,\kappa_N)$. Let $\psi:[0,1]\to\mathbb{R}$ be a piece-wise linear non-decreasing function that interpolates the samples $(\kappa_1,\kappa_2,\ldots,\kappa_N)$ over the nodes $(0,\frac{1}{N},\frac{2}{N},\ldots,1)$ such that $\psi(\frac{i}{N})=\kappa_i$ for all $i\in\{0\}\cup[N]$ and $\psi(\cdot)$ is linear between any two consecutive nodes. Then we have

$$\frac{1}{(2\pi)^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\sum_{j=1}^{\min\{r,s\}}\Phi(\sigma_j(F(\omega_1,\omega_2)))d\omega_1 d\omega_2 = \int_0^1\Phi(\psi(t))dt. \tag{31}$$

It means the singular values of $\boldsymbol{T}$ can be approximately obtained by sampling the function $\psi(t)$ in $[0,1]$, which can be interpreted

[1] The big O notation $O(n)$ follows the standard Bachmann–Landau notation, meaning that there exists a positive constant $c>0$ such that the term is upper-bounded by cn.

as the probability density function of singular values. This motivates a new approximation method as follows.

From a probabilistic perspective, Theorem 1 implies that the statistical average of the singular values of $\boldsymbol{T}$ converges to that of the singular values of the corresponding spectral density matrix F in distribution with any continuous functions Φ [39]. A detailed justification can be found in Appendix B. Inspired by this, we treat $\{\sigma_j(F)\}_j$ as independent continuous random variables and propose a simple approach to approximate $\{\sigma_j(\boldsymbol{T})\}_j$ through non-uniform sampling on the probability distributions of $\{\sigma_j(F)\}_j$. The main idea is that, the singular values of both $\boldsymbol{T}$ and $\boldsymbol{C}$ can be regarded as samples of the random variables $\{\sigma_j(F)\}_j$, from which those of $\boldsymbol{C}$ are uniformly sampled on the underlying probability distribution. As such, we can produce an approximation of $\{\sigma_j(\boldsymbol{T})\}_j$ by (1) estimating the quantile and cumulative distribution functions (CDFs) of $\{\sigma_j(F)\}_j$ through their uniform samples $\{\sigma_j(\boldsymbol{C})\}_j$ by circular approximation, and (2) non-uniformly sampling the CDF by adjusting the quantiles, given the fact that quantile functions are the inverse of CDFs.

Algorithm 1 presents a simple implementation to approximate $\{\sigma_j(\boldsymbol{T})\}_j$ through quantile estimation and interpolation from the uniform samples (i.e., singular values of the circular approximation $\boldsymbol{C}$). Specifically, Algorithm 1 consists of two parts. The first part (Lines 4-11) is to produce an initialization of all singular values by e.g., circular approximation. Here we use the uniform sampling method (Line 7) to generate the block diagonal matrices of the circulant matrix $\boldsymbol{C}$, and then compute all singular values (Line 9) using singular value decomposition (SVD) for each block diagonal matrix. The second part (Lines 13-20) is to adjust the obtained singular values according to their underlying distribution with quantile interpolation. All singular values obtained in the first part are clustered into $\min\{c_{out}, c_{in}\}$ groups (Line 14), each of which contains n^2 singular values being arranged in descending order (Line 15). For each cluster, the underlying quantile function $Q_{\phi_j}(u)$ can be estimated (Line 16) by n^2 singular values with e.g., linear interpolation. Then, quantile interpolation using e.g., kernel smoothing, can be applied to make the estimated quantile function $\hat{Q}_{\phi_j}(u)$ more smooth (Line 17). Given the interpolated smooth quantile function $\hat{Q}_{\phi_j}(u)$, we sample the quantiles at points $u = \{\frac{j-\gamma_j}{n^2}\}_{j=1}^{n^2}$, where $\gamma_j \in (0,1)$ can be randomly generated for simplicity (Line 18). Finally, we compute the value on the interpolated quantile function estimate at the sampling point u, and the resulting values $\{\hat{Q}_{\phi_j}(u)\}_u$ are taken as new singular value approximations (Line 19).

This approach can be proven to have bounded approximation error for each individual singular value, as shown in Theorem 3. It has an improved accuracy and performance guarantee compared to the circular approximation.

Theorem 3. *Let $\phi_j : [-\pi,\pi]^2 \to \mathbb{R}_+$ be a continuous random variable corresponding to the j-th singular value function of $F(\boldsymbol{\omega})$ with $\boldsymbol{\omega} \triangleq (\omega_1,\omega_2)$ and $\sigma_k^{(j)}(\boldsymbol{T})$ be k-th singular value of j-th cluster when all singular values are evenly divided into $\min\{r,s\}$ clusters. It follows that*

$$\sup_{u\in(\frac{k-1}{n^2},\frac{k}{n^2}]} |\sigma_k^{(j)}(\boldsymbol{T}) - Q_{\phi_j}(u)| \le \frac{c_2}{n},$$

Algorithm 1 Singular Values via Quantile Interpolation

1: **Input:** Convolutional filter $\boldsymbol{K} \in \mathbb{R}^{c_{out}\times c_{in}\times h\times w}$
2: Initialize h_1, h_2, w_1, w_2
3: %Compute singular values with circular approximation.
4: Construct $\boldsymbol{T}_{k,l}$ from $\boldsymbol{K}$ according to (9)
5: **for** $j_1 = 1$ **to** n **do**
6: **for** $j_2 = 1$ **to** n **do**
7: Set $(\omega_1,\omega_2) = (-\pi + \frac{2\pi j_1}{n}, -\pi + \frac{2\pi j_1}{n})$
8: Compute $F(\omega_1,\omega_2)$ by (27)
9: Compute SVD of $F(\omega_1,\omega_2)$
10: **end for**
11: **end for**
12: %Adjust singular value sampling via quantile.
13: **for** $j = 1$ **to** $\min\{c_{out}, c_{in}\}$ **do**
14: Collect singular values $\{\sigma_j(F(\omega_1,\omega_2))\}_{\omega_1,\omega_2}$
15: Arrange $\{\sigma_j(F(\omega_1,\omega_2))\}_{\omega_1,\omega_2}$ in descending order
16: Estimate quantile $\hat{Q}_{\phi_j}$ by $\{\sigma_j(F(\omega_1,\omega_2))\}_{\omega_1,\omega_2}$
17: Interpolate quantile using e.g., kernel smoothing
18: Select proper $u = \{\frac{j-\gamma_j}{n^2}\}_{j=1}^{n^2}$ with $\gamma_j \in (0,1)$
19: Compute $\{\hat{Q}_{\phi_j}(u)\}_u$ as singular value estimates
20: **end for**
21: **Output:** Singular values $\{\{\hat{Q}_{\phi_j}(u)\}_u\}_j$

$$\forall 1 \le k \le n^2, 1 \le j \le \min\{r,s\} \tag{32}$$

where $c_2 > 0$ is a constant that only depends on $F(\boldsymbol{\omega})$, and

$$Q_{\phi_j}(u) = \inf\{v \in \mathbb{R} : u \le G_{\phi_j}(v)\} \tag{33}$$

$$G_{\phi_j}(v) = \frac{1}{(2\pi)^2}\mu\{\boldsymbol{\omega} \in [-\pi,\pi]^2 : \phi_j(\boldsymbol{\omega}) \le v\} \tag{34}$$

are quantile and cumulative distribution functions for $\phi_j(\boldsymbol{\omega})$, respectively, and μ is Lebesgue measure.

Proof. See Appendix B. □

Theorem 3 reveals that non-uniform sampling achieves stronger results on the bounded difference of individual singular values than uniform sampling, by leveraging the relation between quantile and cumulative distribution functions for non-uniform sampling. In particular, the singular values $\{\sigma_j(\boldsymbol{T})\}_j$ can be approximated by sampling the quantile functions of $\{\phi_j(\boldsymbol{\omega})\}_j$ within each interval $(\frac{k-1}{n^2}, \frac{k}{n^2}]$. The approximation error of each *individual* singular value is bounded with a finite n and approaches zero as n tends to infinity. Specifically, if the estimation of the quantile function is perfect, this approach approximates each *individual* singular value with gap to the exact one within $O(\frac{1}{n})$. This is in sharp contrast to the circular approximation with the performance guarantee only for the *average* difference of *all* singular values.

Remark 4. *In practice, it is challenging to compute the closed-form expression of the singular value function[2] $\phi_j(\boldsymbol{\omega})$ from $F(\boldsymbol{\omega})$. As such, the estimation of quantile functions is also a challenging task. A feasible and practical way is, as Algorithm 1 shows, to estimate the quantile function $Q_{\phi_j}(u)$*

[2]As $F(\omega_1,\omega_2)$ is a Laurent polynomial matrix w.r.t. $e^{\jmath\omega_1}$ and $e^{\jmath\omega_2}$, the singular value functions $\{\phi_j(\boldsymbol{\omega})\}_j$ can be computed efficiently by, e.g., [41].

from some easily attainable samples, e.g., $\{\sigma_j(\boldsymbol{C})\}_j$, which are the uniform sampling of $\sigma_j(F)$ on $[-\pi,\pi]^2$, followed by quantile interpolation/extrapolation with e.g., kernel smoothing tricks. As such, the singular value approximation can be done by properly sampling the interpolated quantile function. In this way, the approximation accuracy of $\{\sigma_j(\boldsymbol{T})\}_j$ depends on (1) the accuracy of quantile estimation from the samples, (2) the smoothing factors of quantile interpolation, and (3) the sampling grid in $(\frac{k-1}{n^2}, \frac{k}{n^2}]$.

In Algorithm 1, it is expected that the performance is better than that of the circular approximation, because the approximation is built upon the singular value distributions obtained from the circular approximation with some adjusted quantiles, as shown in the first part of Algorithm 1 (Lines 4-11). As such, if the circular approximation is substantially inaccurate, the improvement of Algorithm 1 will be also restricted. To relieve such restriction, a possible way is to apply the second part of Algorithm 1 multiple times to fine-tune singular values, at the cost of increased computational complexity. This is particularly useful for smaller n.

For quantile interpolation, a simple way is linear interpolation, which uses linear polynomials to interpolate new values between two consecutive data points. Kernel density estimation can be used to smooth interpolation. Some other interpolation methods, such as t-Digests [42], are also available in Python and MATLAB from 2019b onward.

B. Spectral Norm Bounding

Although the spectral norm regularization has been successful in enhancing generalization and adversarial robustness, the cost of its exact computation is too expensive to be applied to training because of the high-dimensionality of weight matrices of fully-connected and/or convolutional layers. Instead of computing the exact spectral norm of high-dimensional matrices, existing methods (e.g., [6], [7], [22]) favor differentiable upper bounds that are of lower computational complexity.

Thanks to the spectral representation, spectral norm bounding on the high-dimensional matrix $\boldsymbol{T} \in \mathbb{C}^{n^2 c_{out} \times n^2 c_{in}}$ can be alternatively done on the much lower-dimensional spectral density matrix $F(\boldsymbol{\omega}) \in \mathbb{C}^{c_{out} \times c_{in}}$ with $\boldsymbol{\omega} \in [-\pi,\pi]^2$. Specifically, to upper-bound spectral norm of $\boldsymbol{T}$, we can instead upper-bound it on F due to the following lemma.

Lemma 4. $\|\boldsymbol{T}\|_2 \leq \|F\|_2 \triangleq \sup_{\boldsymbol{\omega}} \|F(\boldsymbol{\omega})\|_2$.

Proof. See Appendix C. □

As the computational complexity of spectral norm (using e.g., power iteration method as in [6], [7], [22]) scales as the size of the matrix, Lemma 4 allow us to compute spectral norm of a low-dimensional matrix with substantially reduced complexity, which is independent of the layer's input size n.

Built upon Lemma 4, the spectral norm of $\boldsymbol{T}$ can be further upper-bounded in different ways.

Theorem 4. *The spectral norm $\|F\|_2$ can be bounded by*

$$\|F\|_2 \leq \min\left\{\sqrt{hw}\|\boldsymbol{R}\|_2, \sqrt{hw}\|\boldsymbol{L}\|_2\right\}, \tag{35}$$

$$\|F\|_2 \leq \max_{\boldsymbol{\omega}} \sqrt{\|F(\boldsymbol{\omega})\|_1 \|F(\boldsymbol{\omega})\|_\infty}, \tag{36}$$

$$\|F\|_2 \leq \sum_{k=-h_1}^{h_2} \sum_{l=-w_1}^{w_2} \|\boldsymbol{T}_{k,l}\|_2, \tag{37}$$

where $\boldsymbol{R} \in \mathbb{R}^{hc_{out} \times wc_{in}}$ and $\boldsymbol{L} \in \mathbb{R}^{wc_{out} \times hc_{in}}$ are $c_{out} \times c_{in}$ block matrices with (c,d)-th block being $\boldsymbol{K}_{c,d,:,:} \in \mathbb{R}^{h \times w}$, and $\boldsymbol{K}^{\intercal}_{c,d,:,:} \in \mathbb{R}^{w \times h}$, respectively.

Proof. See Appendix D. □

Theorem 4 provides a principled way to upper-bound spectral norm of convolutional layers, which substantially reduces the computational complexity of spectral norm approximation. In particular, the first upper bound (35) recovers that in [22], however the derivation here is different as we directly work on F, while the bounds in [22] is for the circulant approximation. It is worth noting that, this bound is different from those used for spectral norm regularizatoin in [6] and spectral normalization in [7], where the 4D filter $\boldsymbol{K}$ is reshaped as a matrix in a heuristic way and spectral norm is computed thereby for the reshaped matrix.

With respect to computational complexity, the first bound (35) requires to compute two spectral norms with sizes $hc_{out} \times wc_{in}$ and $wc_{out} \times hc_{in}$ respectively. The complexity of the second bound (36) depends on the sampling complexity of $\boldsymbol{\omega}$, which usually takes as n^2. As such, it requires to compute n^2 times of ℓ_1 and ℓ_∞ norms with size $c_{out} \times c_{in}$. The third bound (37) requires to compute hw spectral norms with size $c_{out} \times c_{in}$. As spectral norm is usually computed using power method, whose complexity is $O(mn)$ for an $m \times n$ matrix, the computational complexity of all three bounds is $O(hwc_{out}c_{in})$.

It is worth noting that the first and the last bounds are irrelevant to $\boldsymbol{\omega}$ and are differentiable with respect to weights, so that they are good candidates for regularizers.

VI. Extensions and Discussions

To complement the above common settings, some more general cases are discussed with respect to larger stride size, higher dimensional linear convolution, and multiple convolutional layers in linear networks without activation functions and pooling layers.

A. Stride Larger Than 1

In previous sections, we focused on linear convolution with stride size 1. When the stride size g is larger than 1, i.e., $g > 1$, the linear transformation matrix $\boldsymbol{T}$ becomes a block g-Toeplitz matrix, denoted by $\boldsymbol{T}^g$. For simplicity, we consider the same stride side on both horizontal and vertical directions. Thus, we have $\boldsymbol{T}^g = [\boldsymbol{T}_{gk}]_{k=0}^{n-1}$ where $\boldsymbol{T}_{gk} = [\boldsymbol{T}_{gk,gl}]_{l=0}^{n-1}$ with $\boldsymbol{T}_{k,l}$ defined in (9).

According to [43], we have an analogous result to Theorem 1. Let $F : [-\pi,\pi]^2 \to \mathbb{C}^{r \times s}$ be a matrix-valued function, subject to $F \in \mathcal{L}^2([-\pi,\pi]^2)$. The linear transformation matrix $\boldsymbol{T}^g$ with stride g converges to the generating function F, i.e., $\boldsymbol{T}^g \sim_\sigma F(\boldsymbol{\omega}, \boldsymbol{m})$, where

$$F(\boldsymbol{\omega}, \boldsymbol{m}) = \sqrt{\frac{1}{g^2} \sum_{m_1=0}^{g-1} \sum_{m_2=0}^{g-1} f^2(\boldsymbol{\omega}, \boldsymbol{m})} \tag{38}$$

if $\boldsymbol{m} = (m_1, m_2) \in [0, \frac{1}{g}]^2$ and 0 otherwise, with

$$f(\boldsymbol{\omega}, \boldsymbol{m}) = \sum_k \sum_l \boldsymbol{T}_{gk,gl} e^{\jmath \frac{1}{g}(k(\omega_1 + 2\pi m_1) + l(\omega_2 + 2\pi m_2))}.$$

By this, the singular value distribution of $\boldsymbol{T}^g$ can be alternatively studied on the generating function $F(\boldsymbol{\omega}, \boldsymbol{m})$.

B. Higher Dimensional Convolution

According to [44], a block multi-level Toeplitz matrix $\boldsymbol{T} = \{\boldsymbol{T}_{\boldsymbol{i}-\boldsymbol{j}}\}_{\boldsymbol{i},\boldsymbol{j}=\boldsymbol{1}}^{\boldsymbol{n}}$ with $\boldsymbol{i} = (i_1, \ldots, i_d)$, $\boldsymbol{j} = (j_1, \ldots, j_d)$, and $\boldsymbol{n} = (n_1, \ldots, n_d)$, it can be alternatively represented as

$$\boldsymbol{T} = \sum_{|k_1|<n_1} \cdots \sum_{|k_d|<n_d} [\boldsymbol{J}_{n_1}^{(k_1)} \otimes \cdots \otimes \boldsymbol{J}_{n_d}^{(k_d)}] \otimes \boldsymbol{T}_{\boldsymbol{k}} \tag{39}$$

where $\boldsymbol{J}_{n_j}^{(k_j)}$ is a $n_j \times n_j$ binary matrix with (p, q)-th entry being 1 of $p - q = k_j$ and 0 elsewhere, and

$$\boldsymbol{T}_{\boldsymbol{k}} = \frac{1}{(2\pi)^d} \int_\Omega F(\boldsymbol{\omega}) e^{-\jmath <\boldsymbol{k}, \boldsymbol{\omega}>} d\boldsymbol{\omega} \tag{40}$$

with $\Omega = [-\pi, \pi]^d$, $\boldsymbol{k} = (k_1, \ldots, k_d)$, $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_d)$ and $< \boldsymbol{k}, \boldsymbol{\omega} >= \sum_{j=1}^d k_j \omega_j$. Then it follows that Theorem 1 can be generalized to d-dim linear convolutional layers

$$\begin{aligned} &\lim_{\boldsymbol{n}\to\infty} \frac{1}{N} \sum_{j=1}^{\min\{r,s\}N} \Phi(\sigma_j(\boldsymbol{T})) \\ &= \frac{1}{(2\pi)^d} \int_\Omega \sum_{j=1}^{\min\{r,s\}} \Phi(\sigma_j(F(\boldsymbol{\omega}))) d\boldsymbol{\omega} \end{aligned} \tag{41}$$

with $N = \prod_{i=1}^d n_i$, for which the asymptotic singular value distribution of higher dimensional linear convolutional layers can be studied through $F : [-\pi, \pi]^d \to \mathbb{C}^{r\times s}$.

C. Multiple Linear Convolutional Layers

The collective effect of multiple linear convolutional layers without activation function or pooling layers in CNNs can be seen as the product of the linear transformation matrices of multiple convolutional layers.

For convolutional layers, denote by $\boldsymbol{T}(F_i)$ the linear transformation matrix generated from the spectral density matrix $F_i : [-\pi, \pi]^2 \to \mathbb{C}^{r\times s}$, for $i = 1, \ldots, M$. It follows from [45, Theorem 2.46] that

$$\lim_{n\to\infty} \frac{1}{n^2} \|\prod_{i=1}^M \boldsymbol{T}(F_i) - \boldsymbol{T}(\prod_{i=1}^M F_i)\|_1 = 0 \tag{42}$$

which means that the product of Toeplitz matrices is asymptotically equal to the Toeplitz matrix generated by the product of all generating functions associated to each linear convolutional layer. By this, the spectral analysis of M linear convolutional layers can be alternatively studied on the product of generating functions $\prod_{i=1}^M F_i$.

VII. Experiments

A. Singular Value Approximation

To verify the singular value approximation in Section IV, we conduct experiments with respect to four different methods on singular values calculation. The weights of filters are extracted from either the pre-trained networks, e.g., GoogLeNet [46], with ImageNet dataset or from the training process of ResNet-20 [4] on CIFAR-10 dataset. More experimental results using randomly generated weights and weights from pre-trained networks can be found in the arXiv version [40].

- Exact Method: A block doubly Toeplitz matrix $\boldsymbol{T}$ is generated from the convolutional filter $\boldsymbol{K}$ according to (10). The exact singular values of linear convolutional layers are computed by applying SVD to $\boldsymbol{T}$ directly.
- Circular Approximation: A block doubly circulant matrix $\boldsymbol{C}$ is constructed according to (20)-(21). The singular values are computed by applying SVD on $\boldsymbol{C}$ directly.
- Uniform Sampling: The block diagonal matrices $\boldsymbol{B}_{j_1,j_2}$ is produced by uniformly sampling the spectral density matrix $F(\omega_1, \omega_2)$ with sampling grids $(\omega_1, \omega_2) = (-\pi + \frac{2\pi j_1}{n}, -\pi + \frac{2\pi j_1}{n})$ for all $j_1, j_2 \in [n]$. The singular values are obtained by collecting all singular values of $\{\boldsymbol{B}_{j_1,j_2}\}_{j_1,j_2=1}^n$. This corresponds to Lines 4-11 in Algorithm 1.
- Quantile Interpolation: The singular values obtained from uniform sampling are arranged for each $1 \le j \le \min\{c_{in}, c_{out}\}$ in descending order. By quantile estimation using linear interpolation methods, the singular values are recomputed by selecting properly shifted sampling grids as outlined in Algorithm 1.

The experiments are conducted on MATLAB 2020a, which is more friendly to matrix computation. For simplicity, we set $h_1 = h_2$ and $w_1 = w_2$, and the input size per channel is set to 10×10. Figure 1 presents the $(i-1)n+1$-th largest singular values ($i \in [n]$) of four methods with four different filter sizes. The first two filters are from the pre-trained GoogLeNet, and the last two are from the training process of ResNet-20. It can be observed that (1) both circular approximation and uniform sampling have identical singular values for different filter sizes, (2) quantile interpolation improves accuracy of the singular values over the circular approximation with negligible extra running time (see Section [40, Section 9.1]), and (3) during the training process the improvement of the largest singular value approximation is dominant, while for the well-trained networks, the improvement is mainly due to that on smaller singular values. This might be attributed to implicit regularization during training.

B. Spectral Norm Bounding

In what follows, we compare the different spectral norm bounds with respect to the running time and the generalization performance when applied as regularizers during training. Compared with the heuristic approach in [6], [7], extensive experimental results in [22] show that the first bound (35) is more accurate and leads to better generalization performance and adversarial robustness using spectral norm regularization.

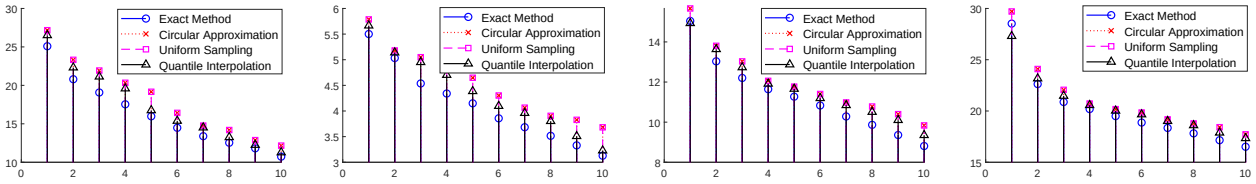

Fig. 1: Exact and approximated singular values of linear convolutional layers arranged in descending order. Input size per channel is set to 10×10. For illustration, only 10 singular values are plotted. Four types of convolutional filters are considered from left to right with sizes $64 \times 3 \times 7 \times 7$ (pre-trained GoogLeNet conv1), $32 \times 16 \times 5 \times 5$ (pre-trained GoogLeNet inception), $16 \times 3 \times 3 \times 3$ (ResNet-20 conv1 after 10 training epochs), and $16 \times 3 \times 3 \times 3$ (ResNet-20 conv1 after 100 training epochs), respectively.

TABLE I: Comparison of spectral norm bounds (a/b: accuracy ratio/running time).

FILTER SIZE	(35)	(36)	(37)
$64 \times 3 \times 7 \times 7$	3.00/12.84	2.14/51.51	4.33/1.146
$64 \times 64 \times 3 \times 3$	1.63/77.68	3.21/54.27	2.20/5.427
$128 \times 64 \times 3 \times 3$	1.48/155.3	3.52/102.3	2.10/8.981
$256 \times 256 \times 3 \times 3$	1.27/1285	4.66/671.7	1.56/68.74
$512 \times 256 \times 3 \times 3$	1.10/2516	4.72/2010	1.27/124.6
$512 \times 512 \times 3 \times 3$	1.13/7232	4.51/3215	1.26/288.5

Therefore, we place our focus on the comparison between the first bound (35) and the third bound (37).

1) Running Time: To verify the accuracy and running time of different spectral norm bounds, we conduct experiments on the pre-trained ResNet-18 model with ImageNet dataset on MATLAB 2020a on HP EliteBook. For the accuracy, we use the circular approximation as the reference and present the ratios to it. Table I summarizes the results for different filters, where the numbers "a/b" read as a times of the circular approximation in accuracy and b milliseconds (ms) in running time. We observe that (1) the first bound (35) usually has the best accuracy except for the larger filter size, e.g., 7×7, while the second bound (36) works better for large filter size; (2) the third bound (37) has comparable accuracy as the first one (35), yet accounting for less than 10% running time of the latter.

2) Regularization: We use spectral norm bounds as regularizers during the training of ResNet-20 model on CIFAR-10 dataset. The ResNet-20 model has 20 convolutional layers, most of which have a 3×3 filter. The CIFAR-10 dataset consists of 50,000 training and 10,000 testing images with size 32×32 in 10 classes. The batch size is 128, and the learning rate is initialized as 0.1 and changed to 0.01 after 100 training epochs. The optimizer is SGD and the momentum is 0.9.

According to the accuracy and running time of different spectral norm bounds in Table I, we place our focus on the first (35) and the third (37) bounds for spectral regularization. Given the training data samples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ drown from an unknown distribution of $(\boldsymbol{x}, y)$ for training an L-layer deep neural network model $y = f_\Theta(\boldsymbol{x})$ with parameters Θ, the spectral regularization is to minimize the following objective function with spectral norm bounds as a regularization term

$$\min_{\Theta} \ \mathbb{E}_{(\boldsymbol{x},y)} \ell(f_\Theta(\boldsymbol{x}), y) + \beta \sum_{j=1}^{L} R_j^u \tag{43}$$

where $\ell(f)$ is the loss function of the model for training, R_j^u is a regularization term using the spectral norm upper bounds of the j-th layer, e.g., (35)-(37), and $\beta > 0$ is a constant to balance between the loss function and the spectral norm regularizer.

In the experiments, the cross entropy function is chosen as the loss function. For the j-th convolutional layer, the regularization term R_j^u is the spectral norm upper bounds chosen from (35) with $R_j^u = \sqrt{hw} \min\{\|\boldsymbol{R}\|_2, \|\boldsymbol{L}\|_2\}$, and from (37) with $R_j^u = \sum_k \sum_l \|\boldsymbol{T}_{k,l}\|_2$, respectively. For the fully-connected layers, R_j^u is directly chosen as the exact spectral norm of the weight matrices. As both the upper bounds in (35) and (37) are in the form of spectral norm, we adopt power iteration method to compute it in the forward propagation. As shown in the proof of Theorem 4, $\boldsymbol{R}$ and $\boldsymbol{L}$ are reshaped matrices of the convolutional filter $\boldsymbol{K}$ with sizes $hc_{out} \times wc_{in}$ and $wc_{out} \times hc_{in}$, respectively, in contrast to the set of hw matrices $\{\boldsymbol{T}_{k,l}\}$ with size $c_{out} \times c_{in}$ each rearranged from $\boldsymbol{K}$.

For a matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, the computational complexity of power iteration method is $O(mn)$. While both bounds (35) and (37) have the same level of computational complexity $O(hwc_{out}c_{in})$, it turns out computing (37) is much faster as the matrices have smaller size. In the backward propagation, the derivative of spectral norms of a matrix $\boldsymbol{A}$ can be computed as $\nabla_{\boldsymbol{A}} \|\boldsymbol{A}\|_2 = \boldsymbol{u}_1 \boldsymbol{v}_1^\mathsf{T}$ where $\boldsymbol{u}_1$ and $\boldsymbol{v}_1$ are the left and right singular vectors corresponding to the largest singular value, respectively, which can be approximately obtained by the power iteration method, as in [6], [7], [22]. Such a derivative is used to update weights for SGD in the backward propagation.

To demonstrate how spectral norm affects generalization, we consider both cases with or without weight decay.

a) Without weight decay: When the weight decay is set to 0, we collect the final prediction accuracy after in total 150 training epochs. For comparison, we use the case with no regularization ($\beta = 0$), which has a test accuracy 89.67%, as the reference. Both spectral norm regularizers have improvement, 1.1% with (35) as the regularizer and 0.8% with (37) as the regularizer, over the the one with no regularizer, which demonstrates the effective of spectral regularization in enhancing generalization performance. The regularizer using (35) has a higher accuracy (0.3%) than (37), due to the more tighter upper bound. Although test accuracy does matter in generalization, we argue that the regularizer (37) would be more preferable as it substantially reduces the computational complexity (with more than 30% running time saving) at the expense of slight performance degradation.

b) With weight decay: when the weight decay is set to 1e-4, we collect the final prediction accuracy after in total

TABLE II: Comparison of test accuracy with spectral norm regularization and with weight decay.

β	0.001	0.0012	0.0014	0.0016	0.0018
(35)	91.72%	91.85%	91.67%	91.52%	**92.23%**
(37)	91.87%	**92.02%**	91.91%	91.67%	91.83%

200 training epochs, where the learning rate is further reduced to 0.001 after 150 epochs. Table II collects the test accuracy with regularization using both spectral norm bounds (35) and (37) with different values of β. We observe that different values of β make different trade-off between loss and spectral regularization, and the choices of $\beta = 0.0018$ for (35), and $\beta = 0.0012$ for (37) yield the best generalization performance, respectively.

To conclude, the spectral norm bound (37) appears more favorable because of its lower computational complexity at the cost of negligible performance degradation.

VIII. Conclusion

In this paper, we proposed to use spectral density matrices to represent the linear convolutional layers in CNNs, for which the linear transformation matrices are block doubly Toeplitz matrices constructed from the convolutional filters. By doing so, spectral analysis of linear convolutional layers with high-dimensionality can be alternatively done on the corresponding spectral density matrices with much lower-dimensionality. Such a spectral representation has been demonstrated to be useful in singular value approximation and spectral norm bounding, which can be used as regularizers to enhance generalization performance with substantially reduced computational complexity. This spectral representation is expected to offer a different approach to understand linear convolutional layers and network architectures, through analyzing the spectral density matrices associated to linear transformation.

Appendix

A. Proof of Theorem 2

Given the generating function $F(\omega_1, \omega_2)$ defined in (27), we introduce an auxiliary matrix $\boldsymbol{C}(F)$ generated by F in the following form

$$\boldsymbol{C}(F) = (\boldsymbol{F}_n \times \boldsymbol{F}_n \times \boldsymbol{I}_r)\text{blkdiag}\Big(\{F(\omega_1,\omega_2), (\omega_1,\omega_2) \in \mathcal{M}\}\Big)(\boldsymbol{F}_n \times \boldsymbol{F}_n \times \boldsymbol{I}_s)^{\mathsf{H}} \quad (44)$$

where $\mathcal{M}$ is the uniform sampling over $[-\pi, \pi]^2$ defined as

$$\mathcal{M} \triangleq \Big\{(\omega_1,\omega_2) = \Big(-\pi + \frac{2\pi j_1}{n}, -\pi + \frac{2\pi j_2}{n}\Big), \forall\, j_1, j_2 \in [n] - 1\Big\}. \quad (45)$$

It can be readily verified that $\boldsymbol{C}(F)$ is also a block doubly circulant matrix, similar to $\boldsymbol{C}$.

First, we show $\boldsymbol{C}(F)$ and $\boldsymbol{C}$ are identical, and thus uniform sampling F yields singular values of $\boldsymbol{C}$. Denote by $[\boldsymbol{C}(F)]_{p,q} \in \mathbb{C}^{r\times s}$ the (p,q)-th block of $\boldsymbol{C}(F)$, where p and q indicate the indices of the first and second levels of circulant blocks, similar to the definition of $\boldsymbol{C}_{p,q}$ in (21). Therefore, we have

$$[\boldsymbol{C}(F)]_{p,q} = \frac{1}{n^2}\sum_{j_1=0}^{n-1}\sum_{j_2=0}^{n-1} F(\frac{2\pi j_1}{n}, \frac{2\pi j_2}{n}) e^{-\jmath 2\pi \frac{pj_1+qj_2}{n}} \quad (46)$$

$$= \frac{1}{n^2}\sum_{j_1=0}^{n-1}\sum_{j_2=0}^{n-1}\sum_{k=-h_1}^{h_2}\sum_{l=-w_1}^{w_2} \boldsymbol{T}_{k,l} e^{\jmath\frac{2\pi}{n}((k-p)j_1+(l-q)j_2)} \quad (47)$$

$$= \frac{1}{n^2}\sum_{k=-h_1}^{h_2}\sum_{l=-w_1}^{w_2} \boldsymbol{T}_{k,l}\sum_{j_1=0}^{n-1} e^{\jmath\frac{2\pi j_1}{n}(k-p)}\sum_{j_2=0}^{n-1} e^{\jmath\frac{2\pi j_2}{n}(l-q)} \quad (48)$$

$$\overset{(a)}{=} \sum_{m_1=-\infty}^{\infty}\sum_{m_2=-\infty}^{\infty} \boldsymbol{T}_{-p+nm_1,-q+nm_2} \quad (49)$$

$$\overset{(b)}{=} \begin{cases} \boldsymbol{T}_{-p,-q}, & p \in \{0\}\cup[h_1],\ q \in \{0\}\cup[w_1] \\ \boldsymbol{T}_{-p,n-q}, & p \in \{0\}\cup[h_1],\ q \in n-[w_2] \\ \boldsymbol{T}_{n-p,-q}, & q \in n-[h_2],\ q \in \{0\}\cup[w_1] \\ \boldsymbol{T}_{n-p,n-q}, & p \in n-[h_2],\ q \in n-[w_2] \\ \boldsymbol{0}, & \text{otherwise} \end{cases} \quad (50)$$

$$\overset{(c)}{=} \boldsymbol{C}_{p,q} \quad (51)$$

for $p, q \in [n] - 1$, where (a) is due to

$$\sum_{j=0}^{n-1} e^{\jmath\frac{2\pi j}{n}(k-p)} = \begin{cases} n, & (k-p) \mod n = 0 \\ 0, & \text{otherwise} \end{cases}, \quad (52)$$

(b) is due to $\boldsymbol{T}_{p,q} = \boldsymbol{0}$ if $p \notin [-h_1 : h_2]$ or $q \notin [-w_1 : w_2]$, and (c) is from (21).

For each $p, q \in [n] - 1$, the (p,q)-th blocks of $\boldsymbol{C}(F)$ and $\boldsymbol{C}$ are identical. Thus, we have

$$\boldsymbol{C}(F) = \boldsymbol{C}. \quad (53)$$

Therefore, by Lemma 3, we conclude that the singular values of $\boldsymbol{C}$ can be given by those of $F(\omega_1, \omega_2)$ with uniform sampling on $[-\pi, \pi]^2$, i.e.,

$$\{\sigma_j(F(\omega_1,\omega_2)) : (\omega_1,\omega_2) \in \mathcal{M}\}, \quad (54)$$

where $\mathcal{M}$ is the uniform sampling grids defined in (45).

Second, we show that the accumulated difference of the singular values between $\boldsymbol{C}$ and $\boldsymbol{T}$ is upper-bounded.

Lemma 5. *Given the banded block doubly Toeplitz and circulant matrices $\boldsymbol{T}$ and $\boldsymbol{C}$, it follows that*

$$\|\boldsymbol{C} - \boldsymbol{T}\|_p^p \le O(n). \quad (55)$$

where $\|\boldsymbol{A}\|_p \triangleq (\sum_{i=1}^n \sum_{j=1}^n |\boldsymbol{A}_{i,j}|^p)^{\frac{1}{p}}$ for $1 \le p < \infty$. When $p = 2$, $\|\boldsymbol{A}\|_p$ boils down to the Frobenius norm $\|\boldsymbol{A}\|_{\mathrm{F}}$.

Proof. Given $\boldsymbol{T}$ and $\boldsymbol{C}$, we define the difference of the (k,l)-th block $\boldsymbol{\Delta}_{k,l} \in \mathbb{C}^{r\times s}$, where $k \in [-n+1 : n-1]$ and $l \in [-n+1 : n-1]$ are indices of two levels of Toeplitz and circulant matrices but not the indices of rows and columns, in the following way

$$\boldsymbol{\Delta}_{k,l} \triangleq [\boldsymbol{C} - \boldsymbol{T}]_{k,l} \quad (56)$$

$$\overset{(a)}{=} \sum_{m_1=-1}^{1}\sum_{m_2=-1}^{1} \boldsymbol{T}_{k+nm_1,l+nm_2}(1 - \delta(m_1,m_2)) \quad (57)$$

where $\delta(m_1,m_2)=1$ if and only if $m_1=m_2=0$, and (a) is due to the banded structure of circulant matrix as in (21). It can be easily verified that $\boldsymbol{C}-\boldsymbol{T}$ is still a block doubly Toeplitz matrix with blocks $\{\boldsymbol{\Delta}_{k,l}\}_{k,l}$. Thus, we have

$$\begin{aligned}
&\|\boldsymbol{C}-\boldsymbol{T}\|_p^p \\
&\overset{(a)}{=} \sum_k\sum_l (n-|k|)(n-|l|)\|\boldsymbol{\Delta}_{k,l}\|_p^p \\
&\overset{(b)}{\le} \sum_k\sum_l\sum_{m_1}\sum_{m_2}(n-|k|)(n-|l|)(1-\delta(m_1,m_2)) \\
&\qquad\qquad \cdot \|\boldsymbol{T}_{k+nm_1,l+nm_2}\|_p^p \\
&\overset{(c)}{=} \sum_{(k,l)\in\mathcal{B}_{12}} (n-|k|)(n-|l|)\|\boldsymbol{T}_{k,l+n}\|_p^p \\
&\quad + \sum_{(k,l)\in\mathcal{B}_{13}} (n-|k|)(n-|l|)\|\boldsymbol{T}_{k,l-n}\|_p^p \\
&\quad + \sum_{(k,l)\in\mathcal{B}_{21}} (n-|k|)(n-|l|)\|\boldsymbol{T}_{k+n,l}\|_p^p \\
&\quad + \sum_{(k,l)\in\mathcal{B}_{22}} (n-|k|)(n-|l|)\|\boldsymbol{T}_{k+n,l+n}\|_p^p \\
&\quad + \sum_{(k,l)\in\mathcal{B}_{23}} (n-|k|)(n-|l|)\|\boldsymbol{T}_{k+n,l-n}\|_p^p \\
&\quad + \sum_{(k,l)\in\mathcal{B}_{31}} (n-|k|)(n-|l|)\|\boldsymbol{T}_{k-n,l}\|_p^p \\
&\quad + \sum_{(k,l)\in\mathcal{B}_{32}} (n-|k|)(n-|l|)\|\boldsymbol{T}_{k-n,l+n}\|_p^p \\
&\quad + \sum_{(k,l)\in\mathcal{B}_{33}} (n-|k|)(n-|l|)\|\boldsymbol{T}_{k-n,l-n}\|_p^p \\
&\overset{(d)}{\le} hw_1^2C_pn + hw_2^2C_pn + h_1^2wC_pn + h_1^2w_1^2C_p \\
&\quad + h_1^2w_2^2C_p + h_2^2wC_pn + h_2^2w_1^2C_p + h_2^2w_2^2C_p \\
&\overset{(e)}{=} an+b
\end{aligned}$$

where $k,l\in[-n+1:n-1]$ and $m_1,m_2\in\{-1,0,1\}$, (a) is due the definition of the element-wise p-norm, (b) is due to the sub-additivity of matrix norms, in (c) we define

$$\begin{aligned}
\mathcal{B}_{11} &= \{(k,l): k\in[-h_1:h_2], l\in[-w_1:w_2]\} \\
\mathcal{B}_{12} &= \{(k,l): k\in[-h_1:h_2], l\in[-(n-1):-(n-w_1)]\} \\
\mathcal{B}_{13} &= \{(k,l): k\in[-h_1:h_2], l\in[(n-w_2):(n-1)]\} \\
\mathcal{B}_{21} &= \{(k,l): k\in[-(n-1):-(n-h_1)], l\in[-w_1:w_2]\} \\
\mathcal{B}_{22} &= \{(k,l): k\in[-(n-1):-(n-h_1)], \\
&\qquad\qquad l\in[-(n-1):-(n-w_1)]\} \\
\mathcal{B}_{23} &= \{(k,l): k\in[-(n-1):-(n-h_1)], \\
&\qquad\qquad l\in[(n-w_2):(n-1)]\} \\
\mathcal{B}_{31} &= \{(k,l): k\in[(n-h_2):(n-1)], l\in[-w_1:w_2]\} \\
\mathcal{B}_{32} &= \{(k,l): k\in[(n-h_2):(n-1)], \\
&\qquad\qquad l\in[-(n-1):-(n-w_1)]\} \\
\mathcal{B}_{33} &= \{(k,l): k\in[(n-h_2):(n-1)], \\
&\qquad\qquad l\in[(n-w_2):(n-1)]\}
\end{aligned}$$

for which $\boldsymbol{T}_{k+nm_1,l+nm_2}\neq\boldsymbol{0}$ in $\mathcal{B}_{11}$ if and only if $m_1=m_2=0$ which invokes $\delta(m_1,m_2)=1$, (d) is due to $\|\boldsymbol{T}_{k,l}\|_p^p$ is upper-bounded by a constant, say C_p for all k,l, and in (e), $a=C_p(h(w_1^2+w_2^2)+(h_1^2+h_2^2)w)$ and $b=C_p(h_1^2+h_2^2)(w_1^2+w_2^2)$. This completes the proof. □

By inspecting $\boldsymbol{T}$ and $\boldsymbol{C}$, we find from Lemma 5 that $\boldsymbol{\Delta}_{k,l}=0$ if and only if $(k,l)\in\mathcal{B}_{11}$. The number of rows and columns with indices outside $\mathcal{B}_{11}$ scales as n. As such, invoking Theorem 3.1 in [38], we conclude that

$$\lim_{n\to\infty}\frac{1}{n}\sum_{j=1}^{\min\{r,s\}n^2}|\sigma_j(\boldsymbol{T})-\sigma_j(\boldsymbol{C})|=O(1) \tag{58}$$

This completes the proof.

B. Proof of Theorem 3

Without loss of generality, we let $r\le s$, i.e., $r=\min\{r,s\}$. We divide all $\{\sigma_j(\boldsymbol{T})\}_{j=1}^{rn^2}$ into r clusters $\{\sigma_k^{(j)}(\boldsymbol{T}),k\in[n^2]\}_{j=1}^r$ according to their localization, each of which is arranged in ascending order, i.e.,

$$\sigma_1^{(j)}(\boldsymbol{T})\le\sigma_2^{(j)}(\boldsymbol{T})\le\cdots\le\sigma_{n^2}^{(j)}(\boldsymbol{T}),\quad\forall j\in[r]. \tag{59}$$

From Theorem 1, we have

$$\begin{aligned}
&\frac{1}{r}\sum_{j=1}^r\lim_{n\to\infty}\frac{1}{n^2}\sum_{k=1}^{n^2}\Phi(\sigma_k^{(j)}(\boldsymbol{T})) \\
&\quad=\frac{1}{r}\sum_{j=1}^r\frac{1}{(2\pi)^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\Phi(\sigma_j(F(\omega_1,\omega_2)))d\omega_1d\omega_2.
\end{aligned} \tag{60}$$

Let $\phi_j:[-\pi,\pi]^2\to\mathbb{R}_+$ be the j-th singular value function of $F(\boldsymbol{\omega})$, i.e., $\phi_j(\boldsymbol{\omega})=\sigma_j(F(\omega_1,\omega_2))$. When taking $\boldsymbol{\omega}$ as a multivariate random variable with uniform distribution on $[-\pi,\pi]^2$, we can treat $\phi_j(\boldsymbol{\omega})$ as a continuous random variable, such that the right-hand side of (60) can be interpreted as

$$\frac{1}{r}\sum_{j=1}^r\mathbb{E}_{\boldsymbol{\omega}}[\Phi(\phi_j(\boldsymbol{\omega}))]$$

Similarly, we can treat $\{\sigma_k^{(j)}(\boldsymbol{T})\}_{k=1}^{n^2}$ as realizations of discrete random variable $X_n^{(j)}$ with equal probability $\Pr(X_n^{(j)}=\sigma_k^{(j)}(\boldsymbol{T}))=\frac{1}{n^2}$, and interpret the left-hand side of (60) as

$$\frac{1}{r}\sum_{j=1}^r\lim_{n\to\infty}\mathbb{E}_{X_n^{(j)}}[\Phi(X_n^{(j)})]$$

Thus, from a probabilistic perspective, Theorem 1 says, for the sequence of random variables $\{X_1^{(j)},X_2^{(j)},\ldots,X_n^{(j)},\ldots\}$, $\mathbb{E}_{X_n^{(j)}}[\Phi(X_n^{(j)})]$ converges to $\mathbb{E}_{\boldsymbol{\omega}}[\Phi(\phi_j(\boldsymbol{\omega}))]$ in distribution for any continuous function Φ.

For both random variables $X_n^{(j)}$ and $\phi_j(\boldsymbol{\omega})$, let us define the cumulative distribution and quantile functions as

$$G_{X_n^{(j)}}(v)=\frac{1}{n^2}\max\{k\in[n^2]:\sigma_k^{(j)}(\boldsymbol{T})\le v\} \tag{61}$$

$$Q_{X_n^{(j)}}(u)=\inf\{v\in\mathbb{R}:u\le G_{X_n^{(j)}}(v)\} \tag{62}$$

$$G_{\phi_j}(v)=\frac{1}{(2\pi)^2}\mu\{\boldsymbol{\omega}\in[-\pi,\pi]^2:\phi_j(\boldsymbol{\omega})\le v\} \tag{63}$$

$$Q_{\phi_j}(u)=\inf\{v\in\mathbb{R}:u\le G_{\phi_j}(v)\} \tag{64}$$

where μ is the Lebesgue measure of $\boldsymbol{\omega}$ on $[-\pi,\pi]^2$. As $\{\sigma_k^{(j)}(\boldsymbol{T})\}_{k=1}^{n^2}$ is ordered and $G_{X_n^{(j)}}(v)$ is right continuous and non-decreasing over v, it follows from [39, Proposition 2.5] that

$$Q_{X_n^{(j)}}(\frac{k}{n^2}) = \sigma_k^{(j)}(\boldsymbol{T}). \tag{65}$$

By Portmanteau Lemma [39, Lemma 3.1], the fact that $\mathbb{E}_{X_n^{(j)}}[\Phi(X_n^{(j)})]$ converges to $\mathbb{E}_{\boldsymbol{\omega}}[\Phi(\phi_j(\boldsymbol{\omega}))]$ in distribution for any continuous function Φ leads to (1) $G_{X_n^{(j)}}(v)$ converges to $G_{\phi_j}(v)$ for every $v\in\mathbb{R}$ at which G_{ϕ_j} is continuous, and (2) $Q_{X_n^{(j)}}(u)$ converges to $Q_{\phi_j}(u)$ for every $u\in(0,1]$ at which Q_{ϕ_j} is continuous.

Inspired by [38, Theorem 3.2, Corollary 3.3], we can further bound the gap between $G_{X_n^{(j)}}(v)$ and $G_{\phi_j}(v)$.

Lemma 6. *There exists a constant c_1 such that*

$$\max_{j\in[r]} |G_{X_n^{(j)}}(v) - G_{\phi_j}(v)| \le \frac{c_1}{n} \tag{66}$$

for every $n>1$.

Proof. Due to Theorem 2, the singular values of $\boldsymbol{C}$ can be given by those of $F(\boldsymbol{\omega})$ with uniform sampling on $[-\pi,\pi]^2$, i.e.,

$$\{\sigma_k^{(j)}(\boldsymbol{C})\}_{k=1}^{n^2} = \{\sigma_j(F(\omega_1,\omega_2)) : (\omega_1,\omega_2)\in\mathcal{M}\} \tag{67}$$

for $j\in[r]$. Following the same footsteps of [38, Theorem 2.2], we have

$$|\sum_{k=1}^{n^2}\sigma_k^{(j)}(\boldsymbol{C}) - \frac{n^2}{(2\pi)^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\sigma_j(F)d\omega_1 d\omega_2| \le c_0' n \tag{68}$$

where $c_0'>0$ is a constant that does not depend on n. Due to Theorem 2, there must exist a constant $c_0>0$ such that

$$\sum_{k=1}^{n^2}|\sigma_k^{(j)}(\boldsymbol{T}) - \sigma_k^{(j)}(\boldsymbol{C})| \le c_0 n. \tag{69}$$

It follows that, there exists a constant $c_1>0$ that does not depend on n such that

$$|\sum_{k=1}^{n^2}\sigma_k^{(j)}(\boldsymbol{T}) - \frac{n^2}{(2\pi)^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\sigma_j(F)d\omega_1 d\omega_2| \le c_1 n \tag{70}$$

By [38, Corollary 3.3], for a real value v, we have

$$|G_{X_n^{(j)}}(v) - G_{\phi_j}(v)| \le \frac{c_1}{n} \tag{71}$$

for all j, where $\phi_j(\boldsymbol{\omega})$ takes values of $\sigma_j(F(\boldsymbol{\omega}))$ that are upper bounded given the fact that $F(\boldsymbol{\omega})$ is a Laurent polynomial matrix with respect to $e^{\jmath\boldsymbol{\omega}}$, each element of which is a Laurent polynomial. This completes the proof. □

Let $\epsilon=\frac{c_1}{n}$ and $\frac{k-1}{n^2}<u\le\frac{k}{n^2}$. By [39, Proposition 2.2], we have $u\le G_{\phi_j}(Q_{\phi_j}(u))$. Together with Lemma 6, we have

$$\begin{aligned} u = u+\epsilon-\epsilon &\le G_{\phi_j}(Q_{\phi_j}(u+\epsilon)) - \epsilon \\ &\le G_{X_n^{(j)}}(Q_{\phi_j}(u+\epsilon)) \end{aligned} \tag{72}$$

Let $\delta=c\epsilon$ with $c>0$ being a constant. Given the fact that $Q_{\phi_j}(u-\epsilon)\ge Q_{\phi_j}(u-\epsilon)-\delta$, we have

$$\begin{aligned} u-\epsilon &\ge G_{\phi_j}(Q_{\phi_j}(u-\epsilon)-\delta) \\ &\ge G_{X_n^{(j)}}(Q_{\phi_j}(u-\epsilon)-\delta) - \epsilon \end{aligned} \tag{73}$$

Thus, due to the fact that $u\le G_{X_n^{(j)}}(v)$ if and only if $Q_{X_n^{(j)}}(u)\le v$, we have

$$Q_{X_n^{(j)}}(u) \le Q_{\phi_j}(u+\epsilon) \tag{74}$$

$$Q_{X_n^{(j)}}(u) \ge Q_{\phi_j}(u-\epsilon) - \delta. \tag{75}$$

Before proceeding further, we investigate the Lipschitz continuity of ϕ_j.

Lemma 7. *The singular value function $\phi_j(\boldsymbol{\omega})=\sigma_j(F(\boldsymbol{\omega}))$ is Lipschitz continuous for every j.*

Proof. According to the generalized Hoffman-Wielandt theorem for singular values [47, Theorem 5] and [48, Theorem 5.1], we have

$$\sqrt{\sum_{j=1}^{r}|\sigma_j(F(\boldsymbol{\omega})) - \sigma_j(F(\boldsymbol{\omega}'))|^2} \tag{76}$$

$$\le \|F(\boldsymbol{\omega}) - F(\boldsymbol{\omega}')\|_{\mathrm{F}} \tag{77}$$

$$= \|\sum_{k_1=-h_1}^{h_2}\sum_{k_2=-w_1}^{w_2}\boldsymbol{T}_{k_1,k_2}(e^{\jmath\boldsymbol{k}^{\mathsf{T}}\boldsymbol{\omega}} - e^{\jmath\boldsymbol{k}^{\mathsf{T}}\boldsymbol{\omega}'})\|_{\mathrm{F}} \tag{78}$$

$$\overset{(a)}{\le} \sum_{k_1=-h_1}^{h_2}\sum_{k_2=-w_1}^{w_2}\|\boldsymbol{T}_{k_1,k_2}\|_{\mathrm{F}}|e^{\jmath\boldsymbol{k}^{\mathsf{T}}\boldsymbol{\omega}} - e^{\jmath\boldsymbol{k}^{\mathsf{T}}\boldsymbol{\omega}'}| \tag{79}$$

$$\overset{(b)}{\le} \sum_{k_1=-h_1}^{h_2}\sum_{k_2=-w_1}^{w_2}\|\boldsymbol{T}_{k_1,k_2}\|_{\mathrm{F}}|\boldsymbol{k}^{\mathsf{T}}(\boldsymbol{\omega}-\boldsymbol{\omega}')| \tag{80}$$

$$\overset{(c)}{\le} \sum_{k_1=-h_1}^{h_2}\sum_{k_2=-w_1}^{w_2}\|\boldsymbol{k}\|\|\boldsymbol{T}_{k_1,k_2}\|_{\mathrm{F}}\|\boldsymbol{\omega}-\boldsymbol{\omega}'\| \tag{81}$$

where (a) is due to the triangle inequality of matrix norm, (b) is due to the non-negativity of matrix norms and the following inequality

$$|e^{\jmath\boldsymbol{k}^{\mathsf{T}}\boldsymbol{\omega}} - e^{\jmath\boldsymbol{k}^{\mathsf{T}}\boldsymbol{\omega}'}| = |\int_{\boldsymbol{\omega}'}^{\boldsymbol{\omega}}\jmath e^{\jmath\boldsymbol{k}^{\mathsf{T}}\boldsymbol{t}}\boldsymbol{k}^{\mathsf{T}}d\boldsymbol{t}| \tag{82}$$

$$\le |\int_{\boldsymbol{\omega}'}^{\boldsymbol{\omega}}|\jmath e^{\jmath\boldsymbol{k}^{\mathsf{T}}\boldsymbol{t}}|\boldsymbol{k}^{\mathsf{T}}d\boldsymbol{t}| \tag{83}$$

$$\le |\boldsymbol{k}^{\mathsf{T}}\int_{\boldsymbol{\omega}'}^{\boldsymbol{\omega}}d\boldsymbol{t}| \tag{84}$$

$$\le |\boldsymbol{k}^{\mathsf{T}}(\boldsymbol{\omega}-\boldsymbol{\omega}')| \tag{85}$$

and (c) is due to Cauchy-Schwarz inequality.

Let $K=\sum_{k_1=-h_1}^{h_2}\sum_{k_2=-w_1}^{w_2}\|\boldsymbol{k}\|\|\boldsymbol{T}_{k_1,k_2}\|_{\mathrm{F}}$, which is a positive constant that does not depend on $\boldsymbol{\omega}$. Thus, we have

$$|\sigma_j(F(\boldsymbol{\omega})) - \sigma_j(F(\boldsymbol{\omega}'))| \le K\|\boldsymbol{\omega}-\boldsymbol{\omega}'\| \tag{86}$$

for all j, which means that $\sigma_j(F(\boldsymbol{\omega}))$ is K-Lipschitz continuous, so is $\phi_j(\boldsymbol{\omega})$ by definition. □

Provided Lemma 7, following the same footsteps in [39, Proposition 2.7], we conclude that $Q_{\phi_j}(u)$ is also Lipschitz continuous, i.e.,

$$|Q_{\phi_j}(u_1) - Q_{\phi_j}(u_2)| \le L|u_1-u_2| \tag{87}$$

for all $u_1,u_2\in(0,1]$.

Now, equipped with the Lipschitz continuity, by (65) and (74), we have

$$\sigma_k^{(j)}(\boldsymbol{T}) = Q_{X_n^{(j)}}(u) \le Q_{\phi_j}(u+\epsilon) \le Q_{\phi_j}(u) + L\epsilon \tag{88}$$

$$\sigma_k^{(j)}(\boldsymbol{T}) = Q_{X_n^{(j)}}(u) \ge Q_{\phi_j}(u-\epsilon) - \delta \ge Q_{\phi_j}(u) - L\epsilon - \delta \tag{89}$$

for $u \in (\frac{k-1}{n^2}, \frac{k}{n^2}]$. This implies that

$$|\sigma_k^{(j)}(\boldsymbol{T}) - Q_{\phi_j}(u)| \le L\epsilon + \delta \triangleq \frac{c_2}{n} \tag{90}$$

for all $k \in [n^2]$ and $j \in [r]$. This completes the proof.

C. Proof of Lemma 4

Inspired by [35, Theorem 4.1], we extend the proof from block Toeplitz to doubly block Toeplitz matrices.

Given a singular value of $\boldsymbol{T} \in \mathbb{R}^{rn^2 \times sn^2}$, say $\sigma(\boldsymbol{T})$, there exist $\boldsymbol{u} \in \mathbb{R}^{rn^2}$ and $\boldsymbol{v} \in \mathbb{R}^{sn^2}$ subject to $\|\boldsymbol{u}\|_2 = \|\boldsymbol{v}\|_2 = 1$ such that $\sigma(\boldsymbol{T}) = \boldsymbol{u}^{\mathsf{T}}\boldsymbol{T}\boldsymbol{v}$, where $\boldsymbol{u} = [\boldsymbol{u}_{k,l}]_{k,l}$ and $\boldsymbol{v} = [\boldsymbol{v}_{k,l}]_{k,l}$, with the (k,l)-th block vector $\boldsymbol{u}_{k,l} \in \mathbb{R}^{r\times s}$ and $\boldsymbol{v}_{k,l} \in \mathbb{R}^{r\times s}$ corresponding to $\boldsymbol{T}_{k,l}$. Thus, we have

$$\sigma(\boldsymbol{T}) = \frac{1}{(2\pi)^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} u(\omega_1,\omega_2)^{\mathsf{T}} F(\omega_1,\omega_2) v(\omega_1,\omega_2) d\omega_1 d\omega_2 \tag{91}$$

where $u(\omega_1,\omega_2)$ and $v(\omega_1,\omega_2)$ are Fourier transforms of $\boldsymbol{u}_{k,l}$ and $\boldsymbol{v}_{k,l}$, respectively, i.e.,

$$u(\omega_1,\omega_2) = \sum_{k=1}^{n}\sum_{l=1}^{n} \boldsymbol{u}_{k,l} e^{\jmath(k\omega_1 + l\omega_2)}, \tag{92}$$

$$v(\omega_1,\omega_2) = \sum_{k=1}^{n}\sum_{l=1}^{n} \boldsymbol{v}_{k,l} e^{\jmath(k\omega_1 + l\omega_2)}. \tag{93}$$

Thus, we have

$$\sigma(\boldsymbol{T}) \overset{(a)}{\le} \frac{1}{(2\pi)^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} \sigma_{\max}(F)\|u(\boldsymbol{\omega})\|_2\|v(\boldsymbol{\omega})\|_2 d\omega_1 d\omega_2 \tag{94}$$

$$\overset{(b)}{\le} \sigma_{\max}(F)\frac{1}{(2\pi)^2}\sqrt{\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\|u(\boldsymbol{\omega})\|_2^2 d\omega_1 d\omega_2} \cdot \sqrt{\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\|v(\boldsymbol{\omega})\|_2^2 d\omega_1 d\omega_2} \tag{95}$$

$$\overset{(c)}{=} \sigma_{\max}(F)\|\boldsymbol{u}\|_2\|\boldsymbol{v}\|_2 \tag{96}$$

$$= \sigma_{\max}(F) \tag{97}$$

where (a) is from the definition of the largest singular value, i.e., $\sigma_{\max}(F) = \sup_{\boldsymbol{\omega}} \frac{u^{\mathsf{T}}Fv}{\|u\|_2\|v\|_2}$, (b) is due to Cauchy inequality, and (c) is resulted directly from the computation of integrals. Thus, it follows immediately that $\|\boldsymbol{T}\|_2 \le \|F\|_2$.

D. Proof of Theorem 4

Let $z_1 = e^{\jmath\omega_1}$ and $z_2 = e^{\jmath\omega_2}$. The (c,d)-th element of the spectral density matrix $F(\omega_1,\omega_2)$ can be rewritten as

$$F_{c,d}(z_1,z_2) = \sum_{k=-h_1}^{h_2}\sum_{l=-w_1}^{w_2} t_{c,d}^{k,l} z_1^k z_2^l. \tag{98}$$

which is a polynomial with respect to z_1 and z_2.

Let $\boldsymbol{R}_{c,d} = [t_{c,d}^{k,l}]_{k,l} \in \mathbb{R}^{h\times w}$, $\boldsymbol{z}_1 = [z_1^{-h_2},\ldots,z_1^{h_1}]$ and $\boldsymbol{z}_2 = [z_2^{-w_2},\ldots,z_2^{w_1}]$. Thus, we can represent $F_{c,d}$ in the following two ways.

$$F_{c,d} = \boldsymbol{z}_1\boldsymbol{R}_{c,d}\boldsymbol{z}_2^{\mathsf{T}} = \boldsymbol{z}_2\boldsymbol{R}_{c,d}^{\mathsf{T}}\boldsymbol{z}_1^{\mathsf{T}}. \tag{99}$$

Hence, the spectral density matrix F can be represented as

$$F = (\boldsymbol{I}_r \otimes \boldsymbol{z}_1)\boldsymbol{R}(\boldsymbol{I}_s \otimes \boldsymbol{z}_2^{\mathsf{T}}) \tag{100}$$

$$= (\boldsymbol{I}_r \otimes \boldsymbol{z}_2)\boldsymbol{L}(\boldsymbol{I}_s \otimes \boldsymbol{z}_1^{\mathsf{T}}) \tag{101}$$

where

$$\boldsymbol{R} = \begin{bmatrix} \boldsymbol{R}_{1,1} & \boldsymbol{R}_{1,2} & \cdots & \boldsymbol{R}_{1,s} \\ \boldsymbol{R}_{2,1} & \cdots & \cdots & \boldsymbol{R}_{2,s} \\ \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{R}_{r,1} & \cdots & \cdots & \boldsymbol{R}_{r,s} \end{bmatrix}, \tag{102}$$

$$\boldsymbol{L} = \begin{bmatrix} \boldsymbol{R}_{1,1}^{\mathsf{T}} & \boldsymbol{R}_{1,2}^{\mathsf{T}} & \cdots & \boldsymbol{R}_{1,s}^{\mathsf{T}} \\ \boldsymbol{R}_{2,1}^{\mathsf{T}} & \cdots & \cdots & \boldsymbol{R}_{2,s}^{\mathsf{T}} \\ \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{R}_{r,1}^{\mathsf{T}} & \cdots & \cdots & \boldsymbol{R}_{r,s}^{\mathsf{T}} \end{bmatrix}, \tag{103}$$

with $\boldsymbol{R} \in \mathbb{R}^{rh\times sw}$ and $\boldsymbol{L} \in \mathbb{R}^{rw\times sh}$. Note that

$$(\boldsymbol{I}_r \otimes \boldsymbol{z}_1)(\boldsymbol{I}_r \otimes \boldsymbol{z}_1)^{\mathsf{H}} = h\boldsymbol{I}_r \tag{104}$$

$$(\boldsymbol{I}_s \otimes \boldsymbol{z}_2)(\boldsymbol{I}_s \otimes \boldsymbol{z}_2)^{\mathsf{H}} = w\boldsymbol{I}_s \tag{105}$$

where the columns are orthogonal. So, we have

$$\|F\|_2 \le \sqrt{hw}\|\boldsymbol{R}\|_2, \quad \|F\|_2 \le \sqrt{hw}\|\boldsymbol{L}\|_2. \tag{106}$$

This gives us the first bound.

For the second bound, given any $\boldsymbol{\omega} \in [-\pi,\pi]^2$, we have

$$\|F(\boldsymbol{\omega})\|_2^2 \le \|F(\boldsymbol{\omega})\|_1\|F(\boldsymbol{\omega})\|_\infty. \tag{107}$$

As $\|F\|_2 = \sup_{\boldsymbol{\omega}}\|F(\boldsymbol{\omega})\|_2$, we have the second spectral norm bound.

For the third bound, we have

$$\|F(\omega_1,\omega_2)\|_2 = \|\sum_k\sum_l \boldsymbol{T}_{k,l} e^{\jmath(k\omega_1+l\omega_2)}\|_2 \tag{108}$$

$$\le \sum_k\sum_l \|\boldsymbol{T}_{k,l}\|_2 |e^{\jmath(k\omega_1+l\omega_2)}| \tag{109}$$

$$= \sum_k\sum_l \|\boldsymbol{T}_{k,l}\|_2, \tag{110}$$

where the inequality is due to Cauchy–Schwarz inequality.

References

[1] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[2] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[5] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning (still) requires rethinking generalization," *Communications of the ACM*, vol. 64, no. 3, pp. 107–115, 2021.

[6] Y. Yoshida and T. Miyato, "Spectral norm regularization for improving the generalizability of deep learning," *arXiv:1705.10941*, 2017.

[7] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," in *International Conference on Learning Representations*, 2018.

[8] Y. Jiang, B. Neyshabur, H. Mobahi, D. Krishnan, and S. Bengio, "Fantastic generalization measures and where to find them," in *International Conference on Learning Representations*, 2020.

[9] G. Jin, X. Yi, L. Zhang, L. Zhang, S. Schewe, and X. Huang, "How does weight correlation affect generalisation ability of deep neural networks?" *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[10] S. Arora, R. Ge, B. Neyshabur, and Y. Zhang, "Stronger generalization bounds for deep nets via a compression approach," in *International Conference on Machine Learning*, PMLR, 2018, pp. 254–263.

[11] P. M. Long and H. Sedghi, "Generalization bounds for deep convolutional neural networks," in *International Conference on Learning Representations*, 2020.

[12] S. Lin and J. Zhang, "Generalization bounds for convolutional neural networks," *arXiv:1910.01487*, 2019.

[13] J. Pennington, S. S. Schoenholz, and S. Ganguli, "Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice," in *Proceedings of International Conference on Neural Information Processing Systems*, 2017, pp. 4788–4798.

[14] J. Pennington, S. Schoenholz, and S. Ganguli, "The emergence of spectral universality in deep networks," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2018, pp. 1924–1932.

[15] C. H. Martin and M. W. Mahoney, "Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning," *Journal of Machine Learning Research*, vol. 22, no. 165, pp. 1–73, 2021.

[16] H. Sedghi, V. Gupta, and P. M. Long, "The singular values of convolutional layers," in *International Conference on Learning Representations*, 2019.

[17] A. Bibi, B. Ghanem, V. Koltun, and R. Ranftl, "Deep layers as stochastic solvers," in *International Conference on Learning Representations*, 2019.

[18] P. L. Bartlett, D. J. Foster, and M. J. Telgarsky, "Spectrally-normalized margin bounds for neural networks," in *Advances in Neural Information Processing Systems*, 2017, pp. 6240–6249.

[19] F. Farnia, J. M. Zhang, and D. Tse, "Generalizable adversarial training via spectral normalization," *arXiv:1811.07457*, 2018.

[20] B. Neyshabur, S. Bhojanapalli, and N. Srebro, "A PAC-Bayesian approach to spectrally-normalized margin bounds for neural networks," *arXiv:1707.09564*, 2017.

[21] K. Roth, Y. Kilcher, and T. Hofmann, "Adversarial training generalizes data-dependent spectral norm regularization," *arXiv preprint arXiv:1906.01527*, 2019.

[22] S. Singla and S. Feizi, "Bounding singular values of convolution layers," *arXiv:1911.10258*, 2019.

[23] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[24] J. Wang, Y. Chen, R. Chakraborty, and S. X. Yu, "Orthogonal convolutional neural networks," *arXiv:1911.12207*, 2019.

[25] R. Appuswamy, T. Nayak, J. Arthur, S. Esser, P. Merolla, J. Mckinstry, T. Melano, M. Flickner, and D. Modha, "Structured convolution matrices for energy-efficient deep learning," *arXiv:1606.02407*, 2016.

[26] Z. Zhu and M. B. Wakin, "On the asymptotic equivalence of circulant and Toeplitz matrices," *IEEE Transactions on Information Theory*, vol. 63, no. 5, pp. 2975–2992, May 2017.

[27] O. Rippel, J. Snoek, and R. P. Adams, "Spectral representations for convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 2449–2457.

[28] A. Araujo, B. Negrevergne, Y. Chevaleyre, and J. Atif, "On Lipschitz regularization of convolutional layers using Toeplitz matrix theory," *Thirty-Fifth AAAI Conference on Artificial Intelligence*, 2021.

[29] H. V. Henderson and S. R. Searle, "The vec-permutation matrix, the vec operator and Kronecker products: A review," *Linear and Multilinear Algebra*, vol. 9, no. 4, pp. 271–288, 1981.

[30] J. Gutiérrez-Gutiérrez and P. M. Crespo, "Block Toeplitz matrices: Asymptotic results and applications," *Foundations and Trends® in Communications and Information Theory*, vol. 8, no. 3, pp. 179–257, 2012.

[31] R. Gray, "On the asymptotic eigenvalue distribution of Toeplitz matrices," *IEEE Transactions on Information Theory*, vol. 18, no. 6, pp. 725–730, November 1972.

[32] F. Avram, "On bilinear forms in Gaussian random variables and Toeplitz matrices," *Probability Theory and Related Fields*, vol. 79, no. 1, pp. 37–45, 1988.

[33] S. V. Parter, "On the distribution of the singular values of Toeplitz matrices," *Linear Algebra and its Applications*, vol. 80, pp. 115–130, 1986.

[34] P. A. Voois, "A theorem on the asymptotic eigenvalue distribution of Toeplitz-block-Toeplitz matrices," *IEEE Transactions on Signal Processing*, vol. 44, no. 7, pp. 1837–1841, 1996.

[35] P. Tilli, "Singular values and eigenvalues of non-Hermitian block Toeplitz matrices," *Linear Algebra and its Applications*, vol. 272, no. 1, pp. 59 – 89, 1998.

[36] M. Miranda and P. Tilli, "Asymptotic spectra of Hermitian block Toeplitz matrices and preconditioning results," *SIAM Journal on Matrix Analysis and Applications*, vol. 21, no. 3, pp. 867–881, 2000.

[37] E. E. Tyrtyshnikov, "A unifying approach to some old and new theorems on distribution and clustering," *Linear Algebra and its Applications*, vol. 232, pp. 1 – 43, 1996.

[38] P. Zizler, R. A. Zuidwijk, K. F. Taylor, and S. Arimoto, "A finer aspect of eigenvalue distribution of selfadjoint band Toeplitz matrices," *SIAM J. Matrix Anal. Appl.*, vol. 24, no. 1, p. 59–67, Jan. 2002.

[39] J. Bogoya, A. Böttcher, S. Grudsky, and E. Maximenko, "Maximum norm versions of the Szegö and Avram–Parter theorems for Toeplitz matrices," *Journal of Approximation Theory*, vol. 196, pp. 79 – 100, 2015.

[40] X. Yi, "Asymptotic singular value distribution of linear convolutional layers," *arXiv:2006.07117*, 2020.

[41] J. A. Foster, J. G. McWhirter, M. R. Davies, and J. A. Chambers, "An algorithm for calculating the QR and singular value decompositions of polynomial matrices," *IEEE Transactions on Signal Processing*, vol. 58, no. 3, pp. 1263–1274, 2009.

[42] T. Dunning and O. Ertl, "Computing extremely accurate quantiles using t-digests," *arXiv:1902.04023*, 2019.

[43] E. Ngondiep, S. Serra-Capizzano, and D. Sesana, "Spectral features and asymptotic properties for g-circulants and g-Toeplitz sequences," *SIAM Journal on Matrix Analysis and Applications*, vol. 31, no. 4, pp. 1663–1687, 2010.

[44] M. Oudin and J. P. Delmas, "Asymptotic generalized eigenvalue distribution of block multilevel Toeplitz matrices," *IEEE Transactions on Signal Processing*, vol. 57, no. 1, pp. 382–387, Jan 2009.

[45] G. Barbarino, C. Garoni, and S. Serra-Capizzano, "Block generalized locally Toeplitz sequences: theory and applications in the multidimensional case," *Electronic Transactions on Numerical Analysis*, vol. 53, pp. 113–216, 2020.

[46] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[47] L. Mirsky, "Symmetric gauge functions and unitarily invariant norms," *The Quarterly Journal of Mathematics*, vol. 11, no. 1, pp. 50–59, 1960.

[48] J.-G. Sun, "Perturbation analysis for the generalized singular value problem," *SIAM Journal on Numerical Analysis*, vol. 20, no. 3, pp. 611–625, 1983.

Xinping Yi (Member, IEEE) received the Ph.D. degree in electronics and communications from Télécom ParisTech, Paris, France, in 2015. He is currently a Lecturer (Assistant Professor) with the Department of Electrical Engineering and Electronics, University of Liverpool, U.K. Prior to Liverpool, he was a Research Associate with Technische Universität Berlin, Berlin, Germany, from 2014 to 2017; a Research Assistant with EURECOM, Sophia Antipolis, France, from 2011 to 2014; and a Research Engineer with Huawei Technologies, Shenzhen, China, from 2009 to 2011. His main research interests include information theory, graph theory, machine learning, and their applications in wireless communications and artificial intelligence.