ISCTE IUL REPOSITÓRIO
INSTITUTO UNIVERSITÁRIO DE LISBOA

Repositório ISCTE-IUL

Deposited in *Repositório ISCTE-IUL*:
2025-04-30

Deposited version:
Accepted Version

Peer-review status of attached file:
Peer-reviewed

Citation for published item:
Carvalho, L., Costa, J. L., Mourão, J. & Oliveira, G. (2025). The positivity of the neural tangent kernel. SIAM Journal on Mathematics of Data Science. 7 (2), 495-515

Further information on publisher's website:
10.1137/24M1659534

Publisher's copyright statement:
This is the peer reviewed version of the following article: Carvalho, L., Costa, J. L., Mourão, J. & Oliveira, G. (2025). The positivity of the neural tangent kernel. SIAM Journal on Mathematics of Data Science. 7 (2), 495-515, which has been published in final form at https://dx.doi.org/10.1137/24M1659534. This article may be used for non-commercial purposes in accordance with the Publisher's Terms and Conditions for self-archiving.

Use policy

Creative Commons CC BY 4.0
The full-text may be used and/or reproduced, and given to third parties in any format or medium, without prior permission or charge, for personal research or study, educational, or not-for-profit purposes provided that:

- a full bibliographic reference is made to the original source
- a link is made to the metadata record in the Repository
- the full-text is not changed in any way

The full-text must not be sold in any format or medium without the formal permission of the copyright holders.

Serviços de Informação e Documentação, Instituto Universitário de Lisboa (ISCTE-IUL)
Av. das Forças Armadas, Edifício II, 1649-026 Lisboa Portugal
Phone: +(351) 217 903 024 | e-mail: administrador.repositorio@iscte-iul.pt
https://repositorio.iscte-iul.pt

THE POSITIVITY OF THE NEURAL TANGENT KERNEL

LUÍS CARVALHO[1,2], JOÃO L. COSTA[1,2], JOSÉ MOURÃO[2], AND GONÇALO OLIVEIRA[2]

[1]INSTITUTO UNIVERSITÁRIO DE LISBOA (ISCTE-IUL), AV. DAS FORÇAS ARMADAS, 1649-026 LISBOA, PORTUGAL

[2]CENTRO DE ANÁLISE MATEMÁTICA, GEOMETRIA E SISTEMAS DINÂMICOS, INSTITUTO SUPERIOR TÉCNICO, UNIVERSIDADE DE LISBOA, AV. ROVISCO PAIS 1, 1049-001 LISBOA, PORTUGAL

Abstract. The Neural Tangent Kernel (NTK) has emerged as a fundamental concept in the study of wide Neural Networks. In particular, it is known that the positivity of the NTK is directly related to the memorization capacity of sufficiently wide networks, i.e., to the possibility of reaching zero loss in training, via gradient descent. Here we will improve on previous works and obtain a sharp result concerning the positivity of the NTK of feedforward networks of any depth. More precisely, we will show that, for any non-polynomial activation function, the NTK is strictly positive definite. Our results are based on a novel characterization of polynomial functions which is of independent interest.

Contents

1. Introduction

Recently, the increase in the size of deep neural networks (DNNs), both in the number of trainable parameters and the amount of training data resources, has been in step with the astonishing success of using DNNs in practical applications. This motivates the theoretical study of wide DNNs. In such context, the Neural Tangent Kernel (NTK) [13] as emerged as a fundamental concept. In particular, it is known that the ability of sufficiently wide neural networks to memorize a given

2020 *Mathematics Subject Classification.* 68T07,68R01.
Key words and phrases. Wide neural networks, Neural Tangent Kernel, Memorization, Global minima.

training data set is related to the positivity of the NTK. More precisely, if the NTK is strictly positive definite then the quadratic loss will converge to zero, in the training via gradient descent, of an appropriately initialized and sufficiently wide feed-forward network (see for instance [11, 8], and references therein; we also refer to Section 1.1 of this paper for a sketch of what happens in the infinite width limit). The positivity of the NTK has also been related to the generalization performance of DNNs [7, 3, 9].

Consequently, understating which conditions lead to this positivity becomes a fundamental problem in machine learning, and several works, that we will review later, have tackled this question providing relevant and interesting partial results. However, all of these results require nontrivial extra assumptions, either at the level of the training set, for instance by assuming that the data lies in the unit sphere, or at the level of the architecture, for instance by using a specific activation function, or both. The goal of the current paper is to obtain a sharp result in the context of feedforward networks which requires no such extra assumptions. In fact, we will show that for any depth and any non-polynomial activation function the corresponding (infinite width limit) NTK is strictly positive definite (see Section 1.2).

Finally, the proofs we present here are self-contained and partially based on an interesting characterization of polynomial functions (see Section 3) which we were unable to locate in the literature and believe to have mathematical value in itself.

One should note that the results of our paper deal with the idealized setting of infinite width networks and therefore might be of limited use. In this regard we recall that there are empirical evidences that reveal that the finite and infinite networks have different performances [18]. Nonetheless, the simplicity of the limiting case studied in this paper provides a compelling conceptual framework that can help us interpret and explain the performance of networks even in practical applications (see for instance [20])

1.1. **Feedforward Neural Networks and the Neural tangent kernel.** Given $L \in \mathbb{Z}^+$, define a feedforward neural network with $L-1$ hidden layers to be the function $f_\theta = f_\theta^{(L)} : \mathbb{R}^{n_0} \to \mathbb{R}^{n_L}$ defined recursively by the relations

$$f_\theta^{(1)}(x) = \frac{1}{\sqrt{n_0}} W^{(0)} x + \beta b^{(0)} , \tag{1.1}$$

$$f_\theta^{(\ell+1)}(x) = \frac{1}{\sqrt{n_\ell}} W^{(\ell)} \sigma(f_\theta^{(\ell)}(x)) + \beta b^{(\ell)} , \tag{1.2}$$

where the networks parameters θ correspond to the collection of all weight matrices $W^{(\ell)} \in \mathbb{R}^{n_{\ell+1} \times n_\ell}$ and bias vectors $b^{(\ell)} \in \mathbb{R}^{n_{\ell+1}}$, $\sigma : \mathbb{R} \to \mathbb{R}$ is an activation function, that operates entrywise when applied to vectors, and $\beta \geq 0$ is a fixed/non-learnable parameter used to control the intensity of the bias.

We will assume that our networks are initiated with *iid* parameters satisfying:

$$W \sim \mathcal{N}(0, \rho_W^2) \quad \text{and} \quad b \sim \mathcal{N}(0, \rho_b^2) , \tag{1.3}$$

with non-vanishing variances ρ_W and ρ_b .

For $\mu = 1,\dots,n_L$, let $f^{(L)}_{\theta,\mu}$ be the μ-component of the output function $f^{(L)}_{\theta}$. It is well known [24, 17] from the central limit theorem that, in the (sequential) limit $n_1,\dots,n_{L-1}\to\infty$, i.e. when the number of all hidden neurons goes to infinity, the n_L components of the output function $f^{(L)}_{\theta,\mu}:\mathbb{R}^{n_0}\to\mathbb{R}$ converge in law to independent centered Gaussian processes $f^{(L)}_{\infty,\mu}:\mathbb{R}^{n_0}\to\mathbb{R}$ with covariance $\hat{\Sigma}^{(L)}:\mathbb{R}^{n_0}\times\mathbb{R}^{n_0}\to\mathbb{R}$, defined recursively by (compare with [27]):

$$\hat{\Sigma}^{(1)}(x,y) = \frac{\rho_W^2}{\sqrt{n_0}} x^\intercal y + \beta\,\rho_b^2\,, \tag{1.4}$$

$$\hat{\Sigma}^{(\ell+1)}(x,y) = \rho_W^2\,\mathbb{E}_{f\sim\hat{\Sigma}^{(\ell)}}[\sigma(f(x))\sigma(f(y))] + \rho_b^2\,\beta^2\,. \tag{1.5}$$

A centered Gaussian Process f with covariance Σ will be denoted by $f\sim\Sigma$. Thus, for any $\mu\in\{1,\dots,n_L\}$ we have $f^{(L)}_{\infty,\mu}\sim\hat{\Sigma}^{(L)}$. In particular, for $x,y\in\mathbb{R}^{n_0}$,

$$\begin{bmatrix} f^{(L)}_{\infty,\mu}(x) \\ f^{(L)}_{\infty,\mu}(y) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \hat{\Sigma}^{(L)}(x,x) & \hat{\Sigma}^{(L)}(x,y) \\ \hat{\Sigma}^{(L)}(y,x) & \hat{\Sigma}^{(L)}(y,y) \end{bmatrix}\right). \tag{1.6}$$

For a given neural network, defined as before, its Neural Tangent Kernel (NTK) is the matrix valued Kernel whose components $\Theta^{(L)}_{\mu\nu}:\mathbb{R}^{n_0}\times\mathbb{R}^{n_0}\to\mathbb{R}$ are defined by

$$\Theta^{(L)}_{\mu\nu}(x,y) = \sum_{\theta\in\mathcal{P}} \frac{\partial f^{(L)}_{\theta,\mu}}{\partial\theta}(x)\,\frac{\partial f^{(L)}_{\theta,\nu}}{\partial\theta}(y)\,, \tag{1.7}$$

with $\mathcal{P} = \{W^{(\ell)}_{ij}, b^{(\ell)}_k \mid 0\le\ell\le L-1,\ 1\le i,k\le n_{\ell+1},\ 1\le j\le n_\ell\}$ the set of all (learnable) parameters.

The relevance of the NTK comes from the fact that it codifies the learning dynamics in output space, if the learning is carried out using gradient descent with a quadratic loss function. To make this statement clear we need to formulate a precise supervised learning setup. To do that consider that we are given a training set composed of training inputs $\{x_1,\cdots,x_N\}$ and training labels $\{y_1,\cdots,y_N\}$, with each $x_i\in\mathbb{R}^{n_0}$ and each $y_i\in\mathbb{R}^{n_L}$. Then if we set our loss function to be the quadratic loss defined by

$$\mathcal{L}(\theta) = \frac{1}{2}\sum_{j=1}^{N} \|f_\theta(x_j) - y_j\|^2\,, \tag{1.8}$$

our goal is to find parameters θ, and a corresponding neural network f_θ, that minimize this loss. If we are given an initialization θ_0 of our parameters, say sampled according to (1.3), and use gradient descent to determine a learning trajectory in parameter space, i.e., if we evolve the parameters according to the ode

$$\frac{d\theta(t)}{dt} = -\nabla_\theta\mathcal{L}(\theta(t)), \tag{1.9}$$

with initial data $\theta(0)=\theta_0$, then it isn't hard to conclude [13, 8] that the learning dynamics $t\mapsto f_{\theta(t)}(x_i)$, of the output of the neural network associated to the training input x_i, satisfies the evolution equation

$$\frac{d}{dt}\Big(f_{\theta(t),\mu}(x_i)\Big) = \sum_{j=1}^{N}\sum_{\nu=1}^{n_L}\Big(y_{j,\nu} - f_{\theta(t),\nu}(x_j)\Big)\Theta^{(L)}_{\mu\nu}|_{\theta(t)}(x_i,x_j)\,, \tag{1.10}$$

where $y_{j,\nu}$ is the ν−component of the training label $y_j \in \mathbb{R}^{n_L}$.

A fundamental observation by [13], is that, under the initialization conditions (1.3), in the (sequential) infinite width limit (i.e., as $n_1,\dots,n_{L-1}\to\infty$) the NTK converges in law to a deterministic kernel. This result was significantly deepened in [27] [2]. The first of these references showed that the same holds if all the widths are sent to infinite at the same rate. Finally, the second reference [2] gives probabilistic bounds on the deviation of this limit in terms of $\min\{n_1,\dots,n_L\}$. Overall, we find that in such infinite width limits

$$\Theta^{(L)}_{\mu\nu} \to \Theta^{(L)}_{\infty,\mu\nu} = \Theta^{(L)}_{\infty}\delta_{\mu\nu}\,, \tag{1.11}$$

with the scalar kernel $\Theta^{(L)}_{\infty}:\mathbb{R}^{n_0}\times\mathbb{R}^{n_0}\to\mathbb{R}$ defined recursively by

$$\Theta^{(1)}_{\infty}(x,y) = \frac{1}{n_0}x^\intercal y + \beta^2\,, \tag{1.12}$$

$$\Theta^{(\ell+1)}_{\infty}(x,y) = \Theta^{(\ell)}_{\infty}(x,y)\,\dot{\Sigma}^{(\ell+1)}(x,y) + \Sigma^{(\ell+1)}(x,y)\,, \tag{1.13}$$

where, for $\ell\geq 1$,

$$\Sigma^{(\ell+1)}(x,y) = \mathbb{E}_{f\sim\hat{\Sigma}^{(\ell)}}\left[\sigma(f(x))\,\sigma(f(y))\right] + \beta^2\,, \tag{1.14}$$

$$\dot{\Sigma}^{(\ell+1)}(x,y) = \rho_W^2\,\mathbb{E}_{f\sim\hat{\Sigma}^{(\ell)}}\left[\dot{\sigma}(f(x))\,\dot{\sigma}(f(y))\right]\,. \tag{1.15}$$

A particularly relevant consequence of this last result is that, in the infinite width limit, the learning dynamics, obtained from (1.10) by replacing $\Theta^{(L)}$ by $\Theta^{(L)}_{\infty}$, is linear, since the infinite width NTK is constant in parameter space. In particular, this reveals that if $\Theta^{(L)}_{\infty}$ is strictly positive definite, then $f_{\theta(t),\mu}(x_i)\to y_{i,\mu}$, as $t\to\infty$, for all $i\in\{i,\dots,N\}$ and all $\mu\in\{1,\dots,n_L\}$, which implies that the loss function converges to zero, a global minimum, during training. The neural network is therefore able to *memorize* the entire training set.

1.2. **Main results.** Recall that a symmetric matrix $P\in\mathbb{R}^{N\times N}$ is strictly positive definite provided that $u^\intercal P u > 0$, for all $u\in\mathbb{R}^N\setminus\{0\}$. Recall also the following:

Definition 1. *A symmetric function*

$$K:\mathbb{R}^{n_0}\times\mathbb{R}^{n_0}\to\mathbb{R}$$

is a strictly positive definite Kernel provided that, for all choices of finite subsets of $\mathbb{R}^{n_0}$, $X=\{x_1,\dots,x_N\}$ *(thus without repeated elements), the matrix*

$$K_X := \Big[K(x_i,x_j)\Big]_{i,j\in\{1,\dots,N\}}\,, \tag{1.16}$$

is strictly positive definite.

We are now ready to state our main results.

Theorem 1 (Positivity of the NTK for networks with biases)**.** *Consider an architecture with activated biases, i.e. $\beta \neq 0$, and a continuous, almost everywhere differentiable and non-polynomial activation function σ. Then, the NTK $\Theta_\infty^{(L)}$ is (in the sense of Definition 1) a strictly positive definite Kernel for all $L \geq 2$.*

Remark 1. *Notice that the previous result is sharp in the following sense. First, the NTK matrix clearly degenerates if we have repeated training inputs, so, in practice, our result does not make any spurious restrictions at the level of the data. Second, it is also known, see for instance* [23, theorem 4.3] *and compare with* [13, Remark 5]*, that the minimum eigenvalue of an NTK matrix is zero if the activation function is polynomial and the data set is sufficiently large. Finally, the regularity assumption of almost everywhere differentiability is, in view of* (1.15)*, required to have a well defined NTK.*

For the sake of completeness we will also establish a positivity result for the case with no biases ($\beta = 0$). This situation calls for extra work and stronger, yet still reasonable, assumptions about the training set. This further emphasizes the well-established relevance of including biases in our models.

Theorem 2 (Positivity of the NTK for networks with no biases)**.** *Consider an architecture with deactivated biases ($\beta = 0$) and a continuous, almost everywhere differentiable and non-polynomial activation function σ. If the training inputs $\{x_1, \ldots, x_N\}$ are all pairwise non-proportional, then, for all $L \geq 2$, the matrix $\Theta_X^{(L)} = \left[\Theta_\infty^{(L)}(x_i, x_j)\right]_{i,j\in[N]}$ is strictly positive definite.*

Remark 2. *As it is well known, one of the main effects of adding bias terms corresponds, in essence, to adding a new dimension to the input space and embedding the inputs into the hyperplane $x_{n_0+1} = 1$. This has the effect of turning distinct inputs, in the original $\mathbb{R}^{n_0}$ space, into non-proportional inputs in $\mathbb{R}^{n_0+1}$. Hopefully this sheds some light into the distinctions between the last two theorems.*

Remark 3. *Our techniques can be adapted to prove that a more general statement handling networks whose activation functions change from layer to layer, as long as such activation functions satisfy the hypothesis of Theorems 1, 2. For $2 \leq \ell \leq N$, let σ_ℓ be the activation function being used on layer ℓ; changes in the activation functions from layer to layer will lead to modified inductive formulas in* (1.13)–(1.15)*. These new formulas will have σ replaced by the σ_ℓ being used in the corresponding layer. In section 4, which contains the proof of our main theorems, this then corresponds to having a new formula for equation (4.2) and the rest of the proof follows suit.*

The proofs of the previous theorems are the subject of section 4 (see Corollaries 2 and 4 respectively). They partially rely on the following interesting characterization of polynomial functions which we take the opportunity to highlight here:

Theorem 3. *Let $z = (z_i)_{i\in[N]}$, $w = (w_i)_{i\in[N]} \in \mathbb{R}^N$ be totally non-aligned, meaning that*

$$\begin{vmatrix} z_i & w_i \\ z_j & w_j \end{vmatrix} \neq 0 \ \text{, for all } i \neq j \ , \tag{1.17}$$

and let $\sigma : \mathbb{R} \to \mathbb{R}$ be continuous. If there exists $u \in \mathbb{R}^N \setminus \{0\}$, such that

$$\sum_{i=1}^{N} u_i\, \sigma(\theta_1 z_i + \theta_2 w_i) = 0 \,, \ \text{for every} \ (\theta_1, \theta_2) \in \mathbb{R}^2 \,, \tag{1.18}$$

then σ is a polynomial.

The previous result is an immediate consequence of Theorem 5 and Theorem 4, proven in the Section 3.

1.3. **Related Work.** In their original work, [13] already discussed the issue studied in the current paper and proved that, under the additional hypothesis that the training data lies in the unit sphere, the Neural Tangent Kernel (NTK) is strictly positive definite for Lipschitz activation functions. [10] made a further interesting contribution in the case where there are no biases. They found that if the activation function is analytic but non-polynomial and no two data points are parallel, then the minimum eigenvalue of an appropriate Gram matrix is positive; this, in particular, provides a positivity result for the NTK, under the described restrictions. As mentioned above we generalize this result by withdrawing these and other restrictions.

Later [1] worked with the specific case of ReLu activation functions, but were able to drop the very restrictive hypothesis that the data points all lie in the unit sphere. Instead, they provide a result showing that for ReLu activation functions, the minimum eigenvalue of the NTK is "large" under the assumption that the data is δ-separated (meaning that no two data points are very close). In a related work, [23] conducted a study on one hidden layer neural nets where only the input layer is trained. They made the assumption that the data points are on the unit sphere and satisfy a specific δ-separation condition. Their results are applicable to large networks where the number of neurons m increases with the number of data points. Moreover, if the activation function is polynomial the minimal eigenvalue of the NTK vanishes for large enough data sets, as illustrated in theorem 4.3 of the same reference. This shows that our conditions, at the level of the activation function, are also necessary so that our results are sharp.

There are a number of other works which investigate these problems and come to interesting partial results. They all have some intersection with the above mentioned results, but given their relevance we shall briefly mention some of these below.

In [16] it is shown that the NTK is strictly positive definite for a two-layered neural net with ReLU activation. Later, [19] extended this result to multilayered nets, but maintained the ReLU activation restriction. [21] proves a lower bound on the first eigenvalue of the NTK assuming σ has polynomial growth and the training data lies on a sphere. [6] found a lower bound on the smallest eigenvalue of the empirical NTK for finite deep neural networks, where at least one hidden layer is large, with the number of neurons growing linearly with amount of data. They also require a Lipschitz activation function with Lipschitz derivative. Related results can also be found in [4] and references therein.

Moreover, [5] provides a proof of the global convergence of gradient flow in training neural networks in an appropriate limit, using the control of the smallest eigenvalue of the NTK. There the activation function is second-order differentiable, with bounded first and second derivatives, and the input data is drawn independently from a normal distribution $N(0, I_d)$, with the outputs being B^2-sub-Gaussian. This setup provides a controlled probabilistic framework for their results. [15] established new lower and upper bounds on the smallest eigenvalue of the NTK for spherical data with ReLU activation. [26] derived a lower bound for the smallest eigenvalue of the empirical NTK in the two-layer case. The activation function is Lipschitz and piecewise linear, or twice differentiable with a bounded second derivative. Furthermore, σ is centered and normalized with respect to $\mathcal{N}(0,1)$. They work in the ultra-wide regime, where each data point is close to unit norm, and all data points are approximately orthogonal. Using the same assumptions, [25] proved the sharp limits of the smallest and largest eigenvalues of the CK matrix, which is one of the components of the NTK.

Other interesting and relevant works which study the positivity of the NTK and/or its eigenvalues include [28, 12, 22].

1.4. **Paper overview.** This work is organized as follows: In Section 2 we consider, as a warm-up, the particular case of a one hidden layer network with a sufficiently smooth activation function; this provides an accessible introduction to the subject that allows to clarify some of the basic ideas behind the proof of the general case. Our main results are based on a novel characterization of polynomials which is fairly easy to establish in the smooth case considered in Section 2 but requires a lot more effort in the continuous category; this work is carried out in Section 3. Finally, in Section 4, we use the results of the previous section to establish the (strict) positive definiteness of the NTK in the general case of a neural network of any depth, with a continuous and almost everywhere differentiable activation function.

2. The Positivity of the NTK I: warm-up with an instructive special case

It might be instructive to first consider the simplest of cases: a one hidden layer network $L = 2$, with one-dimensional inputs $n_0 = 1$ and one-dimensional outputs $n_2 = 1$. This will allow us to clarify part of the strategy employed in the proof of the general case that will be presented in section 4; nonetheless, the more impatient reader can skip the present section.

In this special case we do not even need to use the recurrence relation (1.12)-(1.13), since we can compute the NTK directly from its definition (1.7). In order to do that it is convenient to introduce the *perceptron random variable*

(2.1) $$p(x) = W^{(1)}\sigma(W^{(0)}x + b^{(0)}) ,$$

with parameters $\theta \in \{W^{(0)}, b^{(0)}, W^{(1)}\}$ satisfying (1.3), and the *kernel random variable*

(2.2) $$\mathcal{K}_\theta(x,y) := \sum_{\theta\in\{W^{(0)},b^{(0)},W^{(1)}\}} \frac{\partial p}{\partial\theta}(x)\frac{\partial p}{\partial\theta}(y) .$$

Using perceptrons the networks under analysis in this section are functions $f_\theta^{(2)} : \mathbb{R} \to \mathbb{R}$ that can be written as

(2.3) $$f_\theta^{(2)}(x) = \frac{1}{\sqrt{n_1}}\sum_{k=1}^{n_1} p_k(x) + \beta\, b^{(1)} ,$$

where n_1 is the number of neurons in the hidden layer and where each perceptron has *iid* parameters $\theta_k \in \mathcal{P}_k = \{W_k^{(0)}, b_k^{(0)}, W_k^{(1)}\}$ and $b^{(1)}$ that satisfy (1.3). Moreover the corresponding NTK (1.7), which in this case is a scalar, satisfies

$$\begin{aligned}\Theta^{(2)}(x,y) &= \frac{1}{n_1}\sum_{k=1}^{n_1}\sum_{\theta_k\in\mathcal{P}_k}\frac{\partial p_k(x)}{\partial\theta_k}\frac{\partial p_k(y)}{\partial\theta_k} + \frac{\partial f_\theta^{(2)}(x)}{\partial b^{(1)}}\frac{\partial f_\theta^{(2)}(y)}{\partial b^{(1)}}\\ &= \frac{1}{n_1}\sum_{k=1}^{n_1}\mathcal{K}_{\theta_k}(x,y) + \beta^2\,.\end{aligned}$$

In the limit $n_1 \to \infty$, the Law of Large Numbers guarantees that it converges a.s. to

$$\Theta_\infty^{(2)}(x,y) = \mathbb{E}_\theta\left[\mathcal{K}_\theta(x,y)\right] + \beta^2\,. \tag{2.4}$$

If we denote the gradient of p, with respect to θ, at $x \in \mathbb{R}$, by

$$\begin{aligned}\nabla_\theta p(x)^\intercal &= \left[\frac{\partial p}{\partial W^{(0)}}(x)\,,\ \frac{\partial p}{\partial b^{(0)}}(x)\,,\ \frac{\partial p}{\partial W^{(1)}}(x)\right]\\ &= \left[xW^{(1)}\dot\sigma(W^{(0)}x + b^{(0)})\,,\ W^{(1)}\dot\sigma(W^{(0)}x+b^{(0)})\,,\ \sigma(W^{(0)}x+b^{(0)})\right]\,,\end{aligned} \tag{2.5}$$

we see that $\mathcal{K}_\theta(x,y) = \nabla_\theta p(y)^\intercal \nabla_\theta p(x)$. Then the Gram matrix $(\mathcal{K}_\theta)_X$, defined over the training set $X = \{x_1, \ldots, x_N\}$, is

$$(\mathcal{K}_\theta)_X := \Big[\mathcal{K}_\theta(x_i,x_j)\Big]_{i,j\in[n]} = \nabla_\theta p(X)^\intercal\nabla_\theta p(X)\,,$$

where we used the $3 \times N$ matrix $\nabla_\theta p(X)$ given by

$$\nabla_\theta p(X) = \Big(\nabla_\theta p(x_1) \quad \nabla_\theta p(x_2) \quad \cdots \quad \nabla_\theta p(x_N)\Big)\,. \tag{2.6}$$

The (infinite width) NTK matrix over X is defined by $\Theta_X^{(2)} := \Big[\Theta_\infty^{(2)}(x_i,x_j)\Big]_{i,j\in[n]}$ and, in view of (2.4), the two matrices are related by

$$\Theta_X^{(2)} = \mathbb{E}_\theta\left[(\mathcal{K}_\theta)_X\right] + \beta^2 ee^\intercal\,, \tag{2.7}$$

where $e := [1\cdots 1]^\intercal \in \mathbb{R}^N$. Now, given $u \in \mathbb{R}^N$ we have

$$\begin{aligned}u^\intercal\Theta_X^{(2)}u &= \mathbb{E}_\theta\left[u^\intercal\nabla_\theta p(X)^\intercal\nabla_\theta p(X)u\right] + \beta^2 u^\intercal ee^\intercal u\\ &= \mathbb{E}_\theta\left[\Big(\nabla_\theta p(X)\,u\Big)^\intercal\nabla_\theta p(X)\,u\right] + \beta^2(u^\intercal e)^2\\ &= \mathbb{E}_\theta\left[\Big\|\nabla_\theta p(X)\,u\Big\|^2\right] + \beta^2(u^\intercal e)^2 \geq 0\,,\end{aligned} \tag{2.8}$$

which shows that $\Theta_X^{(2)}$ is positive semi-definite.

Moreover, we can use the previous observations to achieve our main goal for this section by showing that, under slightly stronger assumptions, $\Theta_X^{(2)}$ is, in fact, strictly positive definite. To do that we will only need to assume that there are no repeated elements in the training set and that the activation function σ is continuous, non-polynomial and almost everywhere differentiable with respect to the Lebesgue measure. Under such conditions, assume there exists $u \neq 0$ such that $u^\intercal\Theta_X^{(2)}u = 0$, i.e., that the NTK matrix is not strictly positive definite. From (2.8) we see that this can only happen if $\beta u^\intercal e = 0$ and $\nabla_\theta p(X)u = 0$, for almost every θ, as measured by the parameter initialization. However, since the third component of the gradient is continuous and our parameters are sampled from probability measures with full support, we conclude that

$$\sum_{i=1}^{N} u_i \sigma(W^{(0)} x_i + b^{(0)}) = 0 \ , \text{ for all } (W^{(0)}, b^{(0)}) \in \mathbb{R}^2 \, .$$

It follows from Theorem 3 that this condition implies that σ must be a polynomial, in contradiction with our assumptions. The proof of this theorem assuming only the continuity of σ is rather involved and will be postponed to Section 3. For now, in coherence with the pedagogical spirit of this section, we will content ourselves with a simple proof that holds for the case of an activation function which is C^{N-1}. Note, however, that this is insufficient for many application in deep learning, where one uses activation functions which fail to be differentiable at some points; the ReLu being the prime example. Let $u \neq 0$ satisfy (1.18). If u has any vanishing components these can be discarded from (1.18), so we can assume, without loss of generality, that all $u_i \neq 0$. Therefore we are allowed to rewrite (1.18) in the following form

$$\sigma(\theta_1 z_N + \theta_2 w_N) = \sum_{i=1}^{N-1} u_i^{(1)} \sigma(\theta_1 z_i + \theta_2 w_i) \, , \tag{2.9}$$

where all $u_i^{(1)} := u_i / u_N \neq 0$. Now we differentiate (2.9) with respect to θ_1 and with respect to θ_2 in order to obtain

$$z_N \dot{\sigma}(\theta_1 z_N + \theta_2 w_N) \ = \ \sum_{i=1}^{N-1} u_i^{(1)} z_i \, \dot{\sigma}(\theta_1 z_i + \theta_2 w_i) \, , \tag{2.10}$$

$$w_N \dot{\sigma}(\theta_1 z_N + \theta_2 w_N) \ = \ \sum_{i=1}^{N-1} u_i^{(1)} w_i \dot{\sigma}(\theta_1 z_i + \theta_2 w_i). \tag{2.11}$$

Then, we multiply equations 2.10 and 2.11 by w_N and z_N respectively and subtract them to derive

$$\sum_{i=1}^{N-1} (z_N w_i - w_N z_i) u_i^{(1)} \dot{\sigma}(\theta_1 z_i + \theta_2 w_i) = 0 \, .$$

Since $(z_N w_{N-1} - w_N z_{N-1}) u_{N-1}^{(1)} \neq 0$, we have

$$\dot{\sigma}(\theta_1 z_{N-1} + \theta_2 w_{N-1}) = \sum_{i=1}^{N-2} u_i^{(2)} \dot{\sigma}(\theta_1 z_i + \theta_2 w_i) \, ,$$

where

$$u_i^{(2)} := -\frac{(z_N w_i - w_N z_i) \, u_i^{(1)}}{(z_N w_{N-1} - w_N z_{N-1}) u_{N-1}^{(1)}} \neq 0.$$

Once again we differentiate with respect to θ_1 and θ_2 and equate the results to obtain

$$\sum_{i=1}^{N-2} (z_{N-1} w_i - w_{N-1} z_i) u_i^{(2)} \ddot{\sigma}(\theta_1 z_i + \theta_2 w_i) = 0 \, .$$

Under the assumed conditions we can keep on repeating this process until we arrive at

$$(z_2 w_1 - w_2 z_1) u_i^{(N-1)} \sigma^{(N-1)}(\theta_1 z_1 + \theta_2 w_1) = 0 .$$

Since the last equality holds for all $(\theta_1, \theta_2) \in \mathbb{R}^2$ we conclude that $\sigma^{(N-1)} \equiv 0$ which implies that σ is a polynomial.

3. Two characterizations of polynomial functions.

In the previous section we proved Theorem 3, in the simple case when σ is C^{N-1}, by showing that, in such case, the conditions of the theorem implied that $\sigma^{(N-1)} \equiv 0$, from which one immediately concludes that σ must be a polynomial. In this section, we will show how to extend this result to the case when σ is only continuous. For that we clearly need different techniques. Basically we will rely in the analysis of σ's finite differences and show that, under the conditions of the theorem, all finite differences of order $N-1$ vanish. Remarkably this also implies that σ is a polynomial.

More precisely, in theorem 5 we will show that, under the conditions of Theorem 3, we must have $\Delta_h^{N-1}\sigma(x) = 0$, for all x and h, and in Theorem 4 that if a continuous function σ satisfies this relation then σ must be a polynomial. We believe that this last result is already known, unfortunately we were unable to find it explicitly in the literature. The article [14] contains a related result which implies Theorem 4. Nonetheless, we present a complete proof here.

For a given function $f : \mathbb{R} \to \mathbb{R}$ let its finite differences be given by

$$(\Delta_h f)(x) = f(x+h) - f(x) . \tag{3.1}$$

Note that each finite difference Δ_h is a linear operator on the space of functions $\mathcal{M} := \mathrm{Map}(\mathbb{R}, \mathbb{R})$, $\Delta_h : \mathcal{M} \to \mathcal{M}$.

The finite differences of second order with increments $\boldsymbol{h} = (h_1, h_2)$ are defined by

$$\begin{aligned}\left(\Delta_{\boldsymbol{h}}^2 f\right)(x) =& \left(\Delta_{h_2}\left(\Delta_{h_1} f\right)\right)(x)\\ =& \left(\Delta_{h_1} f\right)(x+h_2) - \left(\Delta_{h_1} f\right)(x)\\ =& \Big(f(x+h_2+h_1) - f(x+h_2)\Big) - \Big(f(x+h_1) - f(x)\Big) .\end{aligned}$$

Note that Δ_{h_1} and Δ_{h_2} commute, that is $\Delta_{h_1}\left(\Delta_{h_2} f\right) = \Delta_{h_2}\left(\Delta_{h_1} f\right)$. Proceeding inductively we have that

$$\Delta_{(\boldsymbol{h}, h_{n+1})}^{n+1} f(x) = \Delta_{h_{n+1}}\left(\Delta_{\boldsymbol{h}}^n f\right)(x) = \left(\Delta_{\boldsymbol{h}}^n f\right)(x + h_{n+1}) - \left(\Delta_{\boldsymbol{h}}^n f\right)(x).$$

When $\boldsymbol{h} = (h, \dots, h)$ we have $\Delta_{\boldsymbol{h}}^n = \Delta_h^n$.

Theorem 4. *Let $f : \mathbb{R} \to \mathbb{R}$ be a continuous function that, for a given $n \in \mathbb{N}$, satisfies $\Delta_h^n f(x) = 0$, for all h and x. Then f is a polynomial of order at most $n-1$.*

Proof. First we will prove the following restricted version of the result: if $f : \mathbb{R} \to \mathbb{R}$ is such that, for a given $n \in \mathbb{N}$, we have $\Delta_h^n f(x) = 0$, for all $x > 0$ and $h > 0$, then $f|_{\mathbb{R}^+}$ is a polynomial of order $n-1$.

We will do so by induction on n:

The case $n = 1$ is obvious; $\Delta_h f(x) = 0$, that is $f(x+h) = f(x)$, for any $x, h > 0$, is the same as $f(y) = f(x)$, for any $y > x > 0$. In other words, f is constant in $\mathbb{R}^+$.

Now assume the result holds for n and that $\Delta_h^{n+1} f(x) = 0$, for all $x, h > 0$. Consider the function $p : \mathbb{R} \to \mathbb{R}$, defined by $p(x) = \dfrac{f(x) - f(0)}{x}$, when $x > 0$, and $p(x) = 0$, for $x \leq 0$. Note that since $xp(x) = f(x) - f(0)$, for any $x \geq 0$, we have $0 = \Delta_h^{n+1}(f(x) - f(0)) = \Delta_h^{n+1}(xp(x))$, for all $x, h > 0$. Then, by the particular case of Leibniz rule provided by the upcoming identity (3.7) we get (for $x, h > 0$)

$$\begin{aligned} 0 = \Delta_h^{n+1}\Big(xp(x)\Big) &= x\Delta_h^{n+1} p(x) + (n+1)h\Delta_h^n p(x+h) \\ &= x\Big(\Delta_h^n p(x+h) - \Delta_h^n p(x)\Big) + (n+1)h\Delta_h^n p(x+h) \\ &= \Big(x + (n+1)h\Big)\Delta_h^n p(x+h) - x\Delta_h^n p(x)\,, \end{aligned}$$

therefore, for $x, h > 0$,

$$\Big(x + (n+1)h\Big)\Delta_h^n p(x+h) = x\Delta_h^n p(x)\,. \tag{3.2}$$

Unfortunately we cannot evaluate the previous identity directly on $x = 0$. However, if we recall the well known general identity

$$\Delta_h^n g(x) = \sum_{k=0}^{n} (-1)^{n-k} \binom{n}{k} g(x+kh)\,, \tag{3.3}$$

and the definition of p, we get, for $x, h > 0$,

$$x\Delta_h^n p(x) = \sum_{k=0}^{n} (-1)^{n-k} \binom{n}{k} x\, \frac{f(x+kh) - f(0)}{x+kh}\,, \tag{3.4}$$

which converges to zero, when $x \to 0$. Therefore, it follows from (3.2) and the continuity of p, in $\mathbb{R}^+$, that $\Delta_h^n p(h) = 0$, for $h > 0$. Considering $x = (k-1)h$ in (3.2), for $k \geq 2$, we conclude that $(n-k)h\Delta_h^n p(kh) = (k-1)h\Delta_h^n p((k-1)h)$. Inductively we determine that $\Delta_h^n p(kh) = 0$, for all $k \in \mathbb{N}$. We have thus concluded that

$$\Delta_h^n p(x) = 0, \text{ for all } x > 0 \text{ and all } h \in \left\{\frac{x}{k} : k \in \mathbb{N}\right\}.$$

Additionally, when $h = x/k$, it also holds that $h = (x + jh)/(k + j)$, implying that

$$\Delta_h^n p(x + jh) = 0, \text{ for all } x > 0\,, \text{ all } h \in \left\{\frac{x}{k} : k \in \mathbb{N}\right\} \text{ and all } j \in \mathbb{N}_0\,. \tag{3.5}$$

Moreover, given $m \in \mathbb{N}$, using (3.5) and the upcoming identity (3.6) we conclude that $\Delta_{mh}^n p(x) = 0$, provided $h = x/k$ and $m, k \in \mathbb{N}$, i.e.,

$$\Delta_h^n p(x) = 0, \text{ for } \text{ for all } x > 0 \text{ and } h \in \{xQ : Q \in \mathbb{Q} \cap \mathbb{R}^+\},$$

where $\mathbb{Q}$ denotes the rational numbers. By continuity of p, in $\mathbb{R}^+$, we finally know that

$$\Delta_h^n p(x) = 0, \text{ for all } x, h > 0\,.$$

By the induction hypothesis, when restricted to $\mathbb{R}^+$, p is a polynomial of order $n-1$ and, therefore, there exists a polynomial $q : \mathbb{R} \to \mathbb{R}$, of order n, such that

$$f(x) = q(x), \text{ for all } x > 0\,.$$

If we now apply the general identity (3.3), to f, at the point $x = -h/2$, with $h > 0$, and take into consideration that, for $k \in \mathbb{N}_0$, $-h/2 + kh < 0 \Leftrightarrow k = 0$, we get

$$
\begin{aligned}
\Delta_h^{n+1} f(-h/2) &= \sum_{k=0}^{n+1} (-1)^{n+1-k} \binom{n+1}{k} f(-h/2 + kh) \\
&= f(-h/2) + \sum_{k=1}^{n+1} (-1)^{n+1-k} \binom{n+1}{k} q(-h/2 + kh) \\
&= f(-h/2) - q(-h/2) + \sum_{k=0}^{n+1} (-1)^{n+1-k} \binom{n+1}{k} q(-h/2 + kh) \\
&= f(-h/2) - q(-h/2) + \Delta_h^{n+1} q(-h/2) \,.
\end{aligned}
$$

Since, for all x, $\Delta_h^{n+1} f(x) = 0$ by hypothesis, and $\Delta_h^{n+1} q(x) = 0$, because q is a polynomial of order n, we conclude that $f(-h/2) = q(-h/2)$, for all $h > 0$. By continuity $f = q$, in the real line. □

In the proof of the previous theorem we relied on:

Lemma 1. *Let $n, k \in \mathbb{N}$. There exist real coefficients $\{a_j^{(n)}\}_{j \in [k]}$ such that, for any function $p : \mathbb{R} \to \mathbb{R}$, the following identity holds*

$$
\Delta_{kh}^n p(y) = \sum_{j=0}^{n(k-1)} a_j^{(n)} \Delta_h^n p(y + jh) \,. \tag{3.6}
$$

Proof. If $n = 1$,

$$
\Delta_{kh} p(y) = p(y + kh) - p(y) = \sum_{j=0}^{k-1} p\Big(y + (j+1)h\Big) - p(y + jh) = \sum_{j=0}^{k-1} \Delta_h p(y + jh).
$$

In this case $a_j^{(1)} = 1$, for $j = 1, \ldots, k-1$. Assuming the result for n we will prove it for $n+1$.

$$
\Delta_{kh}^{n+1} p(y) = \Delta_{kh}\Big(\Delta_{kh}^n p(y)\Big) = \Delta_{kh}\left(\sum_{j=0}^{n(k-1)} a_j^{(n)} \Delta_h^n p(y + jh)\right) = \sum_{j=0}^{n(k-1)} a_j^{(n)} \Delta_{kh}\Big(\Delta_h^n p(y + jh)\Big)
$$

Now, using the result for $n = 1$, that is $\Delta_{kh} p(y) = \sum_{i=0}^{k-1} \Delta_h p(y + ih)$, we get

$$
\begin{aligned}
\Delta_{kh}^{n+1} p(y) &= \sum_{j=0}^{n(k-1)} a_j^{(n)} \sum_{i=0}^{k-1} \Delta_h^{(n+1)} p(y + jh + ih) \\
&= \sum_{m=0}^{(n+1)(k-1)} \left(\sum_{\substack{i+j=m \\ 0 \le i \le k-1 \\ 0 \le j \le n(k-1)}} a_j^{(n)} \right) \Delta_h^{(k+1)} p(y + mh) \\
&= \sum_{m=0}^{(n+1)(k-1)} a_m^{(n+1)} \Delta_h^{(k+1)} p(y + mh) \,,
\end{aligned}
$$

where the second equality arises from the change of variables $(i,j) \mapsto (m = i+j, j)$, and the last corresponds to the recursive definition of the coefficients $a_m^{(n+1)}$. □

In the proof of the last theorem we also used the following special case of the well known Leinbiz rule for finite difference, the proof of which we present here for the sake of completeness.

Lemma 2. *For any function $g : \mathbb{R} \to \mathbb{R}$,*

$$\Delta_h^{n+1}\Big(xg(x)\Big) = x\Delta_h^{n+1}\Big(g(x)\Big) + (n+1)h\Delta_h^n\Big(g(x+h)\Big). \tag{3.7}$$

Proof. For $n = 0$,

$$\Delta_h\Big(xg(x)\Big) = (x+h)g(x+h) - xg(x) = x\Delta_h\Big(g(x)\Big) + hg(x+h).$$

Assume the identity is valid for n. Then

$$\begin{aligned}
\Delta_h^{n+1}\Big(xg(x)\Big) &= \Delta_h^n\Big(\Delta_h\Big(xg(x)\Big)\Big) = \Delta_h^n\Big((x+h)g(x+h)\Big) - \Delta_h^n\Big(xg(x)\Big) \\
&= \Big((x+h)\Delta_h^n g(x+h) + (nh)\Delta_h^{n-1} g(x+2h)\Big) - \Big(x\Delta_h^n g(x) + (nh)\Delta_h^{n-1} g(x+h)\Big) \\
&= x\Delta_h^{n+1} g(x) + h\Delta_h^n g(x+h) + (nh)\Delta_h^n g(x+h) \\
&= x\Delta_h^{n+1} g(x) + (n+1)h\Delta_h^n g(x+h)\,.
\end{aligned}$$

□

For our next result we will also need a simple version of the chain rule for finite differences. To state it, we need to recall that given $g : \mathbb{R}^2 \to \mathbb{R}$ we can define the variations with respect to the second variable by

$$\frac{\Delta_h g}{\Delta y}(x,y) := g(x, y+h) - g(x,y)\,.$$

It is then easy to see that, given $f : \mathbb{R} \to \mathbb{R}$ and $\alpha, \beta \in \mathbb{R}$, we can apply this to the function $g(x,y) = f(\alpha x + \beta y)$ yielding

$$\frac{\Delta_h}{\Delta y}\left[f(\alpha x + \beta y)\right] = \Big(\Delta_{\beta h} f\Big)(\alpha x + \beta y)\,. \tag{3.8}$$

We now have all we need to state and prove the final result of this section:

Theorem 5. *Let $z = (z_i)$ and $w = (w_i)$ be totally non-aligned, meaning that*

$$\begin{vmatrix} z_i & w_i \\ z_j & w_j \end{vmatrix} \neq 0 \ \text{, for all } i \neq j\ , \tag{3.9}$$

and let $\sigma : \mathbb{R} \to \mathbb{R}$ be continuous. If there exists $u \in \mathbb{R}^N$, with all components non-vanishing, such that

$$\sum_{i=1}^N u_i\, \sigma(\theta_1 z_i + \theta_2 w_i) = 0\ ,\ \text{ for every } \ (\theta_1, \theta_2) \in \mathbb{R}^2\ , \tag{3.10}$$

then $\Delta_{\boldsymbol{h}}^{N-1}\sigma(x) = 0$, for all $x \in \mathbb{R}$ and all $\boldsymbol{h} \in \mathbb{R}^{N-1}$.

Proof. The totally non-aligned condition implies, in particular, that no more than one z_i can vanish. Therefore, by rearranging the indices, we can guarantee that $z_i \neq 0$, for all $i \in [N-1]$. Then, since

$u_1 \neq 0$, we rewrite (3.10) as

$$\sigma(\theta_1 z_1 + \theta_2 w_1) = \sum_{i=2}^{N} u_i^{(1)} \sigma(\theta_1 z_i + \theta_2 w_i)\,, \text{ for every } (\theta_1, \theta_2) \in \mathbb{R}^2\,, \tag{3.11}$$

where $u_i^{(1)} := -u_i/u_1 \neq 0$.

Next we consider the change of variables, $(\theta_1, \theta_2) \mapsto (x_1, y_1)$, defined by

$$\begin{cases} x_1 = z_1\theta_1 + w_1\theta_2 \\ y_1 = \theta_2\,, \end{cases} \tag{3.12}$$

which is clearly a bijection since $z_1 \neq 0$. In the new variables we have

$$\theta_1 z_i + \theta_2 w_i = \frac{z_i}{z_1}(\theta_1 z_1 + \theta_2 w_1) + \left(w_i - \frac{z_i w_1}{z_1}\right)\theta_2 = \alpha_1^i x_1 + \beta_1^i y_1\,,$$

where

$$\alpha_1^i := \frac{z_i}{z_1} \quad \text{and} \quad \beta_1^i := \frac{z_1 w_i - z_i w_1}{z_1}\,.$$

Applying these to (3.11) gives

$$\sigma(x_1) = \sum_{i=2}^{N} u_i^{(1)} \sigma(\alpha_1^i x_1 + \beta_1^i y_1)\,, \text{ for all } (x_1, y_1) \in \mathbb{R}^2\,. \tag{3.13}$$

It turns out that by taking variations with respect to the second variable we can iterate this process. In fact, we will now prove that, for all $0 \leq k \leq N-2$ and all non-vanishing h_i, $i \in [N-1]$, the following recursive identity holds:

$$\left(\Delta^k_{\boldsymbol{h}_k^{k+1}} \sigma\right)(x_{k+1}) = \sum_{i=k+2}^{N} u_i^{(k+1)} \left(\Delta^k_{\boldsymbol{h}_k^i} \sigma\right)(\alpha_{k+1}^i x_{k+1} + \beta_{k+1}^i y_{k+1})\,, \text{ for all } (x_{k+1}, y_{k+1}) \in \mathbb{R}^2\,, \tag{3.14}$$

where the coefficients are determined by

$$u_i^{(k+1)} = -u_i^{(k)}/u_{k+1}^{(k)} \neq 0\,, \tag{3.15}$$

$$\alpha_j^i = \frac{z_i}{z_j} \quad \text{and} \quad \beta_j^i = \frac{z_j w_i - z_i w_j}{z_j}\,, \tag{3.16}$$

the change of variables is defined by

$$\begin{cases} x_{k+1} = \alpha_k^{k+1} x_k + \beta_k^{k+1} y_k \\ y_{k+1} = y_k\,, \end{cases} \tag{3.17}$$

and the increment vectors are set according to

$$\boldsymbol{h}_k^i = \left(\beta_k^i h_k, \beta_{k-1}^i h_{k-1}, \ldots, \beta_1^i h_1\right)\,. \tag{3.18}$$

Notice that all components of $\boldsymbol{h}_k^i$ are non-vanishing.

The proof follows by induction: The $k = 0$ case corresponds to (3.13). So let us assume that the identity holds for $0 \leq k \leq N-3$. Then, by taking variations of (3.14) with respect to y_{k+1} and increment $h = h_{k+1} \neq 0$, we can use the chain rule (3.8) to obtain

$$0 = \sum_{i=k+2}^{N} u_i^{(k+1)} \left(\Delta^{k+1}_{\boldsymbol{h}^i_{k+1}} \sigma\right)(\alpha^i_{k+1} x_{k+1} + \beta^i_{k+1} y_{k+1}) \,, \ \text{ for all } \ (x_{k+1}, y_{k+1}) \in \mathbb{R}^2 \,, \tag{3.19}$$

where $\boldsymbol{h}^i_{k+1} = (\beta^i_{k+1} h_{k+1}, \boldsymbol{h}^i_k)$, which can be rewritten as

$$\left(\Delta^{k+1}_{\boldsymbol{h}^{k+2}_{k+1}} \sigma\right)(\alpha^{k+2}_{k+1} x_{k+1} + \beta^{k+2}_{k+1} y_{k+1}) = \sum_{i=k+3}^{N} u_i^{(k+2)} \left(\Delta^{k+1}_{\boldsymbol{h}^i_k} \sigma\right)(\alpha^i_{k+1} x_{k+1} + \beta^i_{k+1} y_{k+1}) \,, \tag{3.20}$$

with $u_i^{(k+2)} = -u_i^{(k+1)} / u_{k+2}^{(k+1)} \neq 0$.

Following the iterative procedure, consider the change of variables $(x_{k+1}, y_{k+1}) \mapsto (x_{k+2}, y_{k+2})$ defined by

$$\begin{cases} x_{k+2} = \alpha^{k+2}_{k+1} x_{k+1} + \beta^{k+2}_{k+1} y_{k+1} \\ y_{k+2} = y_{k+1} \,, \end{cases} \tag{3.21}$$

and observe that it is a bijection, since $k+2 \leq N-1$ implies that $z_{k+2} \neq 0 \Leftrightarrow \alpha^{k+2}_{k+1} \neq 0$. In these new variables

$$\begin{aligned} \alpha^i_{k+1} x_{k+1} + \beta^i_{k+1} y_{k+1} &= \frac{\alpha^i_{k+1}}{\alpha^{k+2}_{k+1}} \left(\alpha^{k+2}_{k+1} x_{k+1} + \beta^{k+2}_{k+1} y_{k+1}\right) + \left(\beta^i_{k+1} - \beta^{k+2}_{k+1} \frac{\alpha^i_{k+1}}{\alpha^{k+2}_{k+1}}\right) y_{k+1} \\ &= \alpha^i_{k+2} x_{k+2} + \beta^i_{k+2} y_{k+2} \,, \end{aligned}$$

with the last identity requiring some algebraic manipulations to be established. So we see that, in the new variables, (3.20) becomes

$$\left(\Delta^{k+1}_{\boldsymbol{h}^{k+2}_{k+1}} \sigma\right)(x_{k+2}) = \sum_{i=k+3}^{N} u_i^{(k+2)} \left(\Delta^{k+1}_{\boldsymbol{h}^i_k} \sigma\right)(\alpha^i_{k+2} x_{k+2} + \beta^i_{k+2} y_{k+2}) \,, \ \text{ for all } \ (x_{k+2}, y_{k+2}) \in \mathbb{R}^2 \,, \tag{3.22}$$

as desired. This closes the induction proof that establishes the validity of (3.14) for all $0 \leq k \leq N-2$. By choosing $k = N-2$ in that identity we obtain

$$\left(\Delta^{N-2}_{\boldsymbol{h}^{N-1}_{N-2}} \sigma\right)(x_{N-1}) = u_N^{(N)} \left(\Delta^{N-2}_{\boldsymbol{h}^N_{N-2}} \sigma\right)(\alpha^N_{N-1} x_{N-1} + \beta^N_{N-1} y_{N-1}) \,, \ \text{ for all } \ (x_{N-1}, y_{N-1}) \in \mathbb{R}^2 \,. \tag{3.23}$$

Finally, if we take variations of the last equation with respect to y_{N-1} and increment $h = h_{N-1} \neq 0$, and recall that $u_N^{(N)} \neq 0$, we arrive at

$$0 = \left(\Delta^{N-1}_{\boldsymbol{h}^N_{N-1}} \sigma\right)(\alpha^N_{N-1} x_{N-1} + \beta^N_{N-1} y_{N-1}) \,, \ \text{ for all } \ (x_{N-1}, y_{N-1}) \in \mathbb{R}^2 \,, \tag{3.24}$$

where $\boldsymbol{h}^N_{N-1} = (\beta^N_{N-1} h_{N-1}, \boldsymbol{h}^N_{N-2})$. Since, in view of (3.9), all β^i_j, with $i \neq j$, are non-vanishing we conclude that $\left(\Delta^{N-1}_{\boldsymbol{h}} \sigma\right)(x) = 0$, for all $x \in \mathbb{R}$ and all $\boldsymbol{h} \in \mathbb{R}^{N-1}$. □

Theorem 3 is now a direct consequence of the main results of this section, namely Theorem 5 and Theorem 4.

4. The Positivity of the NTK II: the general case

After establishing the necessary technical results of the previous section, we now return to the study of the sign of the NTK. More precisely, in this section we will consider general networks (1.1) – in terms of activation function, and the number of inputs, outputs and hidden layers – and we will show that, under very general assumptions, the (infinite width limit) NTK, $\Theta_\infty^{(L)}$, is strictly positive definite, for all $L \geq 2$ (at least one hidden layer). This will be achieved by studying the positive definiteness of various symmetric matrices related to the recurrence formulas (1.12)-(1.13).

Given a symmetric function $K : \mathbb{R}^{n_0} \times \mathbb{R}^{n_0} \to \mathbb{R}$, and a training set $X = \{x_1, \dots, x_N\} \subset \mathbb{R}^{n_0}$, we define its matrix over X by

$$K_X = \Big[K_{ij} := K(x_i, x_j)\Big]_{i,j\in[N]} , \tag{4.1}$$

where we use the notation $[N] = \{1, \dots, N\}$. Furthermore, to clarify our terminology, recall that the symmetric matrix K_X is strictly positive definite when $u^\intercal K_X u > 0$, for all $u \in \mathbb{R}^N \setminus \{0\}$.

Inspired by the recurrence structure in both (1.4), (1.5) and (1.12), (1.13), we consider two Kernel matrices over X related by the identity

$$K_{ij}^{(2)} = \mathbb{E}_{f\sim K^{(1)}}\Big[\sigma\Big(f(x_i)\Big)\sigma\Big(f(x_j)\Big)\Big] + \beta^2 . \tag{4.2}$$

In the following, given $f : \mathbb{R}^{n_0} \to \mathbb{R}$, we will write $Y = f(X) = [f(x_1)\cdots f(x_N)]^\intercal \in \mathbb{R}^N$ and, as before, we will also use the notation $e := [1\cdots 1]^\intercal \in \mathbb{R}^N$. Recall that the notation $\sim K^{(1)}$, introduced after (1.6), is a shorthand for a centered Gaussian Process with covariance function $K^{(1)}$. Analogously, $\sim K_X^{(1)}$ refers to the centered normal distribution with covariance matrix $K_X^{(1)}$. So, when $f \sim K^{(1)}$ then $f(X) \sim K_X^{(1)}$. We are assuming $K_X^{(1)}$ is positive semi-definite.

We see that, for $i, j \in [N]$, the (i,j) entry of $K_X^{(2)}$ is given by (4.2), and so we can write it as

$$\begin{aligned}
K_X^{(2)} &= \mathbb{E}_{f(X)\sim K_X^{(1)}}\Big(\sigma\Big(f(X)\Big)\sigma\Big(f(X)^\intercal\Big)\Big) + \beta^2 e e^\intercal \\
&= \mathbb{E}_{Y\sim K_X^{(1)}}\Big(\sigma(Y)\sigma(Y)^\intercal\Big) + \beta^2 e e^\intercal \\
&= \mathbb{E}_{Y\sim K_X^{(1)}}\left(\begin{bmatrix}\sigma(Y) & \beta e\end{bmatrix}\begin{bmatrix}\sigma(Y)^\intercal \\ \beta e^\intercal\end{bmatrix}\right),
\end{aligned}$$

where $\sigma\Big(f(X)\Big)$ and $\sigma(Y)$ are $N \times 1$ matrices defined by

$$\sigma\Big(f(X)\Big) = \Big[\sigma(f(x_1))\cdots\sigma(f(x_N))\Big]^\intercal \quad \text{and} \quad \sigma(Y) = \Big[\sigma(y_1)\cdots\sigma(y_N)\Big]^\intercal .$$

Now, given $u \in \mathbb{R}^N$, we have

$$\begin{aligned} u^\intercal K_X^{(2)} u &= \mathbb{E}_{Y\sim K_X^{(1)}}\left(u^\intercal \begin{bmatrix} \sigma(Y) & \beta e \end{bmatrix} \begin{bmatrix} \sigma(Y)^\intercal \\ \beta e^\intercal \end{bmatrix} u \right) \\ &= \mathbb{E}_{Y\sim K_X^{(1)}}\left(\begin{bmatrix} u^\intercal \sigma(Y) & \beta u^\intercal e \end{bmatrix} \begin{bmatrix} \sigma(Y)^\intercal u \\ \beta e^\intercal u \end{bmatrix} \right) \\ &= \mathbb{E}_{Y\sim K_X^{(1)}}\left(\left\| [\sigma(Y)^\intercal u \, , \, \beta e^\intercal u] \right\|^2 \right) . \end{aligned} \tag{4.3}$$

We conclude that $u^\intercal K_X^{(2)} u \geq 0$, that is, $K_X^{(1)}$ positive semi-definite implies that $K_X^{(2)}$ is also positive semi-definite. As already observed in [13], it turns out that we can easily strengthen this relation:

Proposition 1 (Induction step)**.** *Assume that the activation function σ is continuous and not a constant. If $K_X^{(1)}$ is strictly positive definite, then $K_X^{(2)}$, defined by* (4.2)*, is also strictly positive definite.*

Proof. Assume that under the prescribed assumptions $K_X^{(2)}$ is not strictly positive definite. Then there exists $u \in \mathbb{R}^N \setminus \{0\}$ such that $u^\intercal K_X^{(2)} u = 0$. In view of (4.3) this implies that $\sigma\big(Y\big)^\intercal u = 0$, $\mathcal{N}(0, K_X^{(1)})$-almost everywhere, but since $K_X^{(1)}$ is, by assumption, strictly positive definite, the corresponding Gaussian measure has full support in $\mathbb{R}^N$ and, by continuity, we must have

$$\sigma\big(y\big)^\intercal u = 0 \ \text{, for all } y \in \mathbb{R}^N . \tag{4.4}$$

By rearranging the components of u we can assume that $u_N \neq 0$. Then, the last identity, applied to a vector of the form $y = (0, \ldots, 0, x)$, $x \in \mathbb{R}$, would imply that $\sigma(x) = -\sigma(0) \sum_{i=1}^{N-1} \frac{u_i}{u_N}$, i.e., σ is a constant. Since this contradicts our assumptions we conclude that $K_X^{(2)}$ is strictly positive definite. □

While Proposition 1 offers the necessary induction step to propagate the favorable sign to the matrices $\hat{\Sigma}_X^{(\ell+1)}$ and $\Sigma_X^{(\ell+1)}$, it is insufficient to assert that these matrices are strictly positive definite. Since $\hat{\Sigma}_X^{(1)}$ typically does not exhibit this property. Nonetheless, we will now show that under suitable and relatively mild conditions related to the training set and activation function, the desired positivity for $\hat{\Sigma}_X^{(2)}$ and $\Sigma_X^{(2)}$ emerges from the recurrence relations (1.5) and (1.14).

4.1. **Networks with biases.** We will first deal with the case with biases ($\beta \neq 0$). Our strategy, inspired by the special case studied in Section 2, will be to steer towards Theorem 3 to obtain the desired conclusion.

Theorem 6. *Assume that the training inputs x_i are all distinct, and that the activation function σ is continuous and non-polynomial. If*

$$K^{(1)}(x, y) = \alpha^2 x^\intercal y + \beta^2 \, , \tag{4.5}$$

with $\alpha\,\beta \neq 0$, then $K_X^{(2)}$, as defined by (4.2) *is strictly positive definite.*

Proof. As in the proof of Proposition 1, if $K_X^{(2)}$ is not strictly positive definite, then there exists a non-vanishing $u \in \mathbb{R}^N$ such that $\sigma\big(Y\big)^\intercal u = 0$, $\mathcal{N}(0, K_X^{(1)})$-almost everywhere. Let X also denote the matrix whose columns are the training inputs. Then, for $\tilde{X} = \begin{bmatrix}\alpha X^\intercal & \beta e\end{bmatrix}$ we can write

$$K_X^{(1)} = \alpha^2 X^\intercal X + \beta^2 e e^\intercal = \begin{bmatrix}\alpha X^\intercal & \beta e\end{bmatrix}\begin{bmatrix}\alpha X \\ \beta e^\intercal\end{bmatrix} = \tilde{X}\tilde{X}^\intercal .$$

Let $\mathrm{rank}(\tilde{X}) = r \geq 1$ and $\tilde{X}_{(r)}$ be a $r \times (n_0 + 1)$ matrix containing r linearly independent rows of $\tilde{X}$. We assume without loss of generality that $\tilde{X}_{(r)}$ consists of the first r rows of $\tilde{X}$. Then, there exists an $N \times r$ matrix B such that

$$\tilde{X} = B\tilde{X}_{(r)} . \tag{4.6}$$

The distribution, over $\mathbb{R}^N$, of $Y = (Y_1 \dots, Y_r, \dots, Y_N) \sim \mathcal{N}(0, K_X^{(1)})$, in general, has a degenerated covariance, however the distribution over the first r components $Y_{(r)} := (Y_1, \dots, Y_r) \sim \mathcal{N}(0, \tilde{X}_{(r)}\tilde{X}_{(r)}^\intercal)$ has a non degenerated covariance matrix and

$$\mathrm{Cov}(BY_{(r)}) = B\,\mathrm{Cov}(Y_{(r)})B^\intercal = B\tilde{X}_{(r)}\tilde{X}_{(r)}^\intercal B^\intercal = \tilde{X}\tilde{X}^\intercal = K_X^{(1)}.$$

Thus the fact that there exists $u \in \mathbb{R}^N \setminus \{0\}$ such that $\sigma\Big(Y\Big)^\intercal u = 0$, for $\mathcal{N}(0, K_X^{(1)})$-almost every $Y \in \mathbb{R}^N$, is equivalent to $\sigma\Big(BY_{(r)}\Big)^\intercal u = 0$, for $\mathcal{N}(0, \tilde{X}_{(r)}\tilde{X}_{(r)}^\intercal)$-almost every $Y_{(r)} \in \mathbb{R}^r$. An advantage of the last formulation is that the corresponding measure has full support in $\mathbb{R}^r$ so, by continuity, we conclude that

$$\sigma\Big(By\Big)^\intercal u = 0 \text{ , for all } y \in \mathbb{R}^r . \tag{4.7}$$

To proceed we will need the following

Lemma 3. *Assume that B is an $N \times r$ matrix with no repeated rows. Then there exists $y^{\neq} \in \mathbb{R}^r$ such that $z^{\neq} = By^{\neq}$ is a vector in $\mathbb{R}^N$ with pairwise distinct entries.*

Proof. Let $y^{\neq}(x) = (1, x, \dots, x^{r-1})$, for $x \in \mathbb{R}$. Then $z^{\neq}(x) := By^{\neq}(x)$ is a vector whose entries are polynomials $p_i(x) = \sum_j B_{ij}x^{j-1}$, $i \in [N]$, which (as polynomials) are pairwise distinct since the rows of B are also pairwise distinct.

Now consider the set $\mathcal{I}$, of real numbers where at least two of the polynomials coincide, and the sets $\mathcal{I}_{ij}$, corresponding to the solutions of $p_i(x) = p_j(x)$, for a specific pair of indices $i \neq j$. Clearly $\mathcal{I} \subset \cup_{i\neq j}\mathcal{I}_{ij}$, and therefore

$$\#\mathcal{I} \leq \sum_{i\neq j} \#\mathcal{I}_{ij} \leq \binom{N}{2} \times (r-1) < \#\mathbb{R} .$$

In conclusion, we can choose $x \in \mathbb{R}$ such that all the entries of $z^{\neq}(x) = (p_1(x), \dots, p_N(x))$ are pairwise distinct, and we are done. □

Now, since X has no repeated elements, B has no repeated rows. Then, for any given (θ_1, θ_2), let $y = \theta_1 y^{\neq} + \theta_2\,\beta\,e$, with $y^{\neq}$ as in the previous lemma, for which $By = \theta_1 z^{\neq} + \theta_2\,\beta\,e$. In such case the equality (4.7) becomes$\sum_{i=1}^N u_i\sigma(\theta_1 z_i + \theta_2\,\beta) = 0$, for all $(\theta_1, \theta_2) \in \mathbb{R}^2$, with the z_i pairwise distinct and $\beta \neq 0$. But since a vector $z^{\neq} \in \mathbb{R}^N$ with all entries pairwise distinct is, in the sense of (1.17), *totally non-aligned* with the bias vector $e = (1, \dots, 1)$, we can apply Theorem 3 to conclude that σ must be a polynomial. This contradicts our assumptions, therefore $K_X^{(2)}$ must be strictly positive definite. □

An immediate consequence of the preceding result is that $\hat{\Sigma}_X^{(2)}$ and $\Sigma_X^{(2)}$ are strictly positive definite. The induction provided by Proposition 1 then leads to the following conclusion.

Corollary 1. *Under the conditions of Theorem 6, and for all $\ell \geq 2$, $\hat{\Sigma}_X^{(\ell)}$ and $\Sigma_X^{(\ell)}$, defined by* (1.4), (1.5) *and* (1.14), *are strictly positive definite.*

We are now ready to achieve our main goal for this section, which is the proof of Theorem 1 which for convenience we restate here as follows:

Corollary 2 (Theorem 1)**.** *Under the conditions of Theorem 6, and for σ continuous and differentiable almost everywhere, the matrix $\Theta_X^{(\ell)}$ is strictly positive definite, for all $\ell \geq 2$.*

Proof. We start by noticing that, if we assume that σ is differentiable almost everywhere, a computation similar to the one used for equation (4.3), shows that $\dot{\Sigma}_X^{(\ell)} = [\dot{\Sigma}^{(\ell)}(x_i, x_j)]_{i,j \in \{1,\ldots,N\}}$ is positive semi-definite, for all $\ell \geq 2$. On the other hand $\Theta_X^{(1)}$ as defined in remark 1 is positive semi-definite. We recall that due to the Schur product theorem, the Hadamard product of two positive semi-definite matrices remains positive semi-definite. Thus, the matrix $\Theta_X^{(1)} \odot \dot{\Sigma}_X^{(2)} = \left[\Theta_\infty^{(1)}(x_i, x_j)\dot{\Sigma}_\infty^{(2)}(x_i, x_j)\right]_{i,j \in [N]}$ is positive semi-definite. Since the sum of a strictly positive definite matrix with a positive semi-definite matrix gives rise to a strictly positive definite matrix, we can conclude that $\Theta_X^{(2)} = \Theta_X^{(1)} \odot \dot{\Sigma}_X^{(2)} + \Sigma_X^{(2)}$ is strictly positive definite. The statement then follows from the recurrence (1.13) and Corollary 1. □

4.2. **Networks with no biases.** We can also deal with the case with no biases, i.e., $\beta = 0$, but this case requires more effort and stronger (although still mild) assumptions on the training set; another reason in favor of the well known importance of including biases in our models.

Theorem 7. *Assume that the training inputs are all pairwise non-proportional and that the activation function σ is continuous and non-polynomial. If $K^{(1)}(x, y) = \alpha^2 x^\intercal y$, with $\alpha \neq 0$, then $K_X^{(2)}$, as defined by* (4.2) *is strictly positive definite.*

Proof. Just as in the proof of Theorem 6 we can construct a ranked r matrix $\tilde{X}_{(r)}$ and a matrix B such that (4.6) holds, but now, with the removal of the biases column from $\tilde{X} \in \mathbb{R}^{N \times n_0}$. Although we do not have the helpful bias column, our assumptions on the training set guarantee that the rows of B are pairwise non-proportional which allows to prove the following:

Lemma 4. *Assume the rows of B are all pairwise non-proportional, then there exists $y_1, y_2 \in \mathbb{R}^r$, such that the $\mathbb{R}^N$ vectors $z = By_1$ and $w = By_2$ are* totally non-aligned, *meaning that* (1.17) *holds.*

Proof. As before, consider the polynomials defined by $p_i(x) = \sum_j B_{ij} x^{j-1}$, which (as polynomials) are pairwise non-proportional, in view of the assumptions on B. Now choose x_1 such that $w = (w_i := p_i(x_1))_{i \in [N]}$ has all non zero entries and consider the polynomials $q_i = p_i / w_i$ which (as polynomials) are distinct, since the p_i are pairwise non-proportional. We then see that

$$\begin{vmatrix} p_i(x_2) & w_i \\ p_j(x_2) & w_j \end{vmatrix} = w_j p_i(x_2) - w_i p_j(x_2) \neq 0 \Leftrightarrow q_i(x_2) \neq q_j(x_2)\,. \tag{4.8}$$

So we can construct the desired z by setting $z_i = p_i(x_2)$, where x_2 is such that all $q_i(x_2)$ are distinct. □

Given (θ_1, θ_2), let $y = \theta_1 y_1 + \theta_2 y_2$, with the y_i as in the previous lemma. Then, $By = \theta_1 z + \theta_2 w$ and the equality (4.7) becomes $\sum_{i=1}^N u_i \sigma(\theta_1 z_i + \theta_2 w_i) = 0$, for all $(\theta_1, \theta_2) \in \mathbb{R}^2$, with $z = (z_i)$ and $w = (w_i)$ totally non-aligned. In view of Theorem 3 σ must be a polynomial. Once more, this contradicts our assumptions, therefore $K_X^{(2)}$ must be strictly positive definite. □

An immediate consequence of Proposition 1 and the previous result is the following

Corollary 3. *Under the conditions of Theorem 7, for all $\ell \geq 2$, $\hat{\Sigma}_X^{(\ell)}$ and $\Sigma_X^{(\ell)}$, defined by* (1.4), (1.5), *and* (1.14), *with $\beta = 0$, are strictly positive definite.*

Finally, as in the case in which $\beta \neq 0$, we are now ready to conclude that:

Corollary 4 (Theorem 2)**.** *Under the conditions of Theorem 7, assume moreover that σ is differentiable almost everywhere and $\beta = 0$. Then, $\Theta_X^{(\ell)}$ is strictly positive definite, for all $\ell \geq 2$.*

Acknowledgement

This work was partially supported by FCT/Portugal through CAMGSD, IST-ID , projects UIDB/04459/2020 and UIDP/04459/2020.

References

[1] Z. Allen-Zhu, Y. Li, and Z. Song, *A convergence theory for deep learning via over-parameterization*, in International conference on machine learning, PMLR, 2019, pp. 242–252.

[2] S. Arora, S. S. Du, W. Hu, Z. Li, R. Salakhutdinov, and R. Wang, *On exact computation with an infinitely wide neural net*, Advances in Neural Information Processing Systems, 32 (2018), p. 8141–8150.

[3] S. Arora, S. S. Du, W. Hu, Z. Li, and R. Wang, *Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks*, 36th International Conference on Machine Learning, (2019), pp. 477–502.

[4] A. Banerjee, P. Cisneros-Velarde, L. Zhu, and M. Belkin, *Neural tangent kernel at initialization: linear width suffices*, 216 (2023), pp. 110–118.

[5] P. L. Bartlett, A. Montanari, and A. Rakhlin, *Deep learning: a statistical viewpoint*, Acta numerica, 30 (2021), pp. 87–201.

[6] S. Bombari, M. H. Amani, and M. Mondelli, *Memorization and optimization in deep neural networks with minimum over-parameterization*, (2022).

[7] Y. Cao and Q. Gu, *Generalization bounds of stochastic gradient descent for wide and deep neural networks*, Advances in neural information processing systems, 32 (2019), p. 10836–10846.

[8] L. Carvalho, J. L. Costa, J. Mourão, and G. Oliveira, *Wide neural networks: From non-Gaussian random fields at initialization to the NTK geometry of training*, (2023).

[9] Z. Chen, Y. Cao, Q. Gu, and T. Zhang, *A generalized neural tangent kernel analysis for two-layer neural networks*, Advances in Neural Information Processing Systems, 33 (2020).

[10] S. Du, J. D. Lee, H. Li, L. Wang, and X. Zhai, *Gradient descent finds global minima of deep neural networks*, International conference on machine learning, (2019).

[11] S. Du, X. Zhai, B. Poczos, and A. Singh, *Gradient descent provably optimizes over-parameterized neural networks*, International Conference on Learning Representations, (2018).

[12] Z. Fan and Z. Wang, *Spectra of the conjugate kernel and neural tangent kernel for linear-width neural networks*, Advances in neural information processing systems, 33 (2020), pp. 7710–7721.

[13] A. Jacot, G. Franck, and C. Hongler, *Neural tangent kernel: Convergence and generalization in neural networks*, Advances in Neural Information Processing Systems, (2018), p. 8580–8589.

[14] M. A. Jones, *A difference equation approach to finite differences of polynomials*, The College Mathematics Journal, 51 (2020), pp. 375–377.

[15] K. Karhadkar, M. Murray, and G. Montúfar, *Bounds for the smallest eigenvalue of the ntk for arbitrary spherical data of arbitrary dimension*, arXiv preprint arXiv:2405.14630, (2024).

[16] J. Lai, M. Xu, R. Chen, and Q. Lin, *Generalization ability of wide neural networks on* $\mathbb{R}$, (2023).

[17] J. Lee, Y. Bahri, R. Novak, S. Schoenholz, J. Pennington, and J. Sohl-Dickstein, *Deep neural networks as gaussian processes*, International Conference on Learning Representations, (2018).

[18] J. Lee, S. Schoenholz, J. Pennington, B. Adlam, L. Xiao, R. Novak, and J. Sohl-Dickstein, *Finite versus infinite neural networks: an empirical study*, Advances in Neural Information Processing Systems, 33 (2020), pp. 15156–15172.

[19] Y. Li, Z. Yu, G. Chen, and Q. Lin, *Statistical optimality of deep wide neural networks*, (2023).

[20] S. Lin, Y. Ma, J. Chen, and B. Wen, *Compressed event sensing (ces) volumes for event cameras*, International Journal of Computer Vision, (2024), pp. 1–21.

[21] A. Montanari and Y. Zhong, *The interpolation phase transition in neural networks: Memorization and generalization under lazy training*, (2023).

[22] Q. Nguyen, M. Mondelli, and G. Montufar, *Tight bounds on the smallest eigenvalue of the neural tangent kernel for deep relu networks*, (2020).

[23] A. Panigrahi, A. Shetty, and N. Goyal, *Effect of activation functions on the training of overparametrized neural nets*, (2020).

[24] N. Radford, *Priors for infinite networks*, tech. rep. no. crg-tr-94-1, University of Toronto, 1994.

[25] Z. Wang, D. Wu, and Z. Fan, *Nonlinear spiked covariance matrices and signal propagation in deep neural networks*, arXiv preprint arXiv:2402.10127, (2024).

[26] Z. Wang and Y. Zhu, *Deformed semicircle law and concentration of nonlinear random matrices for ultra-wide neural networks*, The Annals of Applied Probability, 34 (2024), pp. 1896–1947.

[27] G. Yang, *Scaling limits of wide neural networks with weight sharing: Gaussian process behavior, gradient independence, and neural tangent kernel derivation*, ArXiv, abs/1902.04760 (2019).

[28] Z. Zhu, F. Liu, G. G. Chrysos, and V. Cevher, *Generalization properties of nas under activation and skip connection search*, (2022).

Email address: JLCA@ISCTE-IUL.PT