University of Tartu
Faculty of Science and Technology
Institute of Mathematics and Statistics

Karl-Joan Alesma

Quantitative convergence of convolutional neural networks with correlated weights

Mathematical Statistics
Master Thesis (30 ECTS)

Supervisors:
PhD candidate Eloy Mosig García, University of Pisa
prof. Andrea Agazzi, University of Bern
prof. Jüri Lember, University of Tartu

Tartu 2026

KORRELEERITUD KAALUDEGA KONVOLUTSIOONILISE NÄRVIVÕRGU KVANTITATIIVNE KOONDUMINE

Magistritöö

Karl-Joan Alesma

Lühikokkuvõte

Me üldistame Basteri ja Trevisani, 2024 tulemust, mis mõõdab juhusliku täisühendusega närvivõrgu ja Gaussi protsessi vahelist kaugust, juhuslikule korreleeritud kaaludega konvolutsioonilisele närvivõrgule. Selleks leiame ülemise tõkke ruut Wassersteini kaugusele juhusliku konvolutsioonilise närvivõrgu ja talle vastava Gaussi protsessi vahel. Saadud tõke näitab, kuidas arhitektuurilised parameetrid mõjutavad koondumiskiirust, kui kanalite arv kasvab. Erijuhul kui konvolutsioonilise närvivõrgu struktuur taandub tavaliseks närvivõrguks, saame tagasi tulemuse Basteri ja Trevisani, 2024 tööst.

CERCS teaduseriala: P160 Statistika, operatsioonianalüüs, programmeerimine, finants- ja kindlustusmatemaatika.

Märksõnad: Juhuslik konvolutsiooniline närvivõrk, närvivõrgu Gaussi protsess, Gaussi lähendamine, Wassersteini kaugus, lai piirjuht.

QUANTITATIVE CONVERGENCE OF CONVOLUTIONAL NEURAL NETWORKS WITH CORRELATED WEIGHTS

Master thesis

Karl-Joan Alesma

Abstract

We extend the quantitative Gaussian approximation of randomly initialized fully connected neural networks established in Basteri and Trevisan (2024) to convolutional neural networks with patch-wise correlated weights. Specifically, we derive an upper bound on the quadratic Wasserstein distance between the output distribution of a randomly initialized convolutional neural network and its associated neural network Gaussian process. The obtained bound reveals how the architectural parameters affect convergence in the wide limit. As a corollary, we recover the fully connected network bound of Basteri and Trevisan (2024) when the convolutional structure degenerates to a dense layer.

CERCS research specialisation: P160 Statistics, operations research, programming, financial and actuarial mathematics.

Key Words: Random convolutional neural network, neural network Gaussian process, Gaussian approximation, Wasserstein distance, wide limit.

Contents

Introduction

Deep learning neural networks are a family of parametric models that have been successfully applied to a wide range of real-world tasks, including image recognition and natural language processing (LeCun, Bengio, and Hinton, 2015; Goodfellow, Bengio, and Courville, 2016). This widespread success has led to an increasing interest in the theoretical foundations of such methods.

One approach to theoretical analysis is to consider the case where the number of network parameters tends to infinity. It was shown by Neal (1996) that the output of a single-layer neural network with randomly initialised parameters converges to a Gaussian process (GP) as the width goes to infinity. This was later extended to deeper networks by Lee et al. (2018) and Matthews et al. (2018), to convolutional networks by Garriga-Alonso, Rasmussen, and Aitchison (2019) and Novak et al. (2019), and Yang (2019) showed that networks of "any reasonable architecture" converge to a Gaussian process.

An important breakthrough in the analysis of neural networks was the Neural Tangent Kernel (NTK) (Jacot, Gabriel, and Hongler, 2018). They showed that the training of an infinitely wide neural network under gradient descent evolves approximately linearly around its initialization and can be understood as kernel regression with a fixed kernel (the NTK). This kernel only depends on the network architecture evaluated at initialisation. Lee et al. (2019) subsequently showed that gradient descent dynamics in the NTK regime converge to those of a linearised model.

Both the GP correspondence and the NTK characterise neural networks at initialization, which is important for several reasons. To perform Bayesian inference with neural networks, one needs to specify a prior distribution over the space of networks. This can be done by placing a prior distribution on the parameters, which in turn induces a distribution over the networks. The neural network Gaussian process correspondence enables exact Bayesian inference in the infinite-width limit (Neal, 1996; Lee et al., 2018). Beyond the Bayesian motivation, the choice of initialization has direct practical consequences. A proper initialization keeps the variance of activations and gradients controlled across layers, preventing the signal from exploding or vanishing as it propagates through the network (Hanin, 2018; Hanin and Rolnick, 2018).

The preceding convergence results are qualitative, which raises the natural question: how close is a finite-width, randomly initialized network to its infinite-width limit? Quantitative bounds make it possible to evaluate the

gap between the predictions of a finite-width network and its infinite-width NTK equivalent. In addition, they help understand how this gap depends on the network's architecture and training configuration.

Recently, a body of literature has emerged on this topic. Basteri and Trevisan (2024) and Trevisan (2023) established quantitative bounds in Wasserstein distance between the output distribution of a randomly initialized neural network and its corresponding Gaussian process. Favaro et al. (2025) obtained similar bounds for total variation and convex distances. Mosig, Agazzi, and Trevisan (2025) extended these results beyond the initialisation regime by providing explicit Wasserstein bounds for shallow networks at any positive training time under gradient descent.

In this thesis, we extend the results of Basteri and Trevisan (2024) to convolutional neural networks (CNN). We additionally allow the convolutional weights to be patch-wise correlated, as was done by Garriga-Alonso and Wilk (2021). This reveals how the CNN architecture affects the rate of convergence to its corresponding Gaussian process. Although Trevisan (2023) achieves a faster rate of $\mathcal{O}(1/n)$ in Wasserstein distance, the constants are left implicit and require the non-degeneracy of the Gaussian process kernel.

The first chapter provides an overview of the theoretical background, and the second chapter contains the proof of the main result.

1 Notation and preliminaries

In this section, we give an extensive overview of the notation and the theoretical background needed to understand Theorem 2.1. Subsection 1.1 borrows notation and concepts from Garriga-Alonso and Wilk (2021), subsections 1.2–1.7 from Basteri and Trevisan (2024) and Trevisan (2023), subsection 1.9 from Goodfellow, Bengio, and Courville (2016) and Garriga-Alonso and Wilk (2021).

1.1 Multi-index notation

Let $D \in \mathbb{N}$ and $\boldsymbol{F} = (F_1, \dots, F_D) \in \mathbb{N}^D$ be a tuple of positive integers. For a single integer $F \in \mathbb{N}$ we write $[\![F]\!] := \{1, \dots, F\}$, and for a tuple we set

$$[\![\boldsymbol{F}]\!] = [\![F_1]\!] \times \cdots \times [\![F_D]\!].$$

Elements of $[\![\boldsymbol{F}]\!]$ are called *multi-indices*. A sum indexed by $\boldsymbol{q} \in [\![\boldsymbol{F}]\!]$ is understood as the iterated sum

$$\sum_{\boldsymbol{q}=\boldsymbol{1}}^{\boldsymbol{F}} := \sum_{q_1=1}^{F_1} \cdots \sum_{q_D=1}^{F_D}.$$

We write $|\boldsymbol{F}| := \prod_{d=1}^{D} F_d$ for the total number of multi-indices in $[\![\boldsymbol{F}]\!]$.

1.2 Functions on index sets

Throughout the thesis, we work with real-valued functions defined on finite index sets. Since convolutional layers index the inputs and outputs by grids such as $(C^{(0)}, F_h^{(0)}, F_d^{(0)}) \in \mathbb{N}^3$ (as with RGB images), we adopt from the outset the abstract viewpoint where indices range over an arbitrary finite set S.

Given a set S, we write $\mathbb{R}^S$ for the set of real-valued functions $f\colon S \to \mathbb{R}$. To reduce notational clutter, we often use the bracket notation $f[s] = f_s$. When $S = [\![k]\!]$ we recover the space of k-dimensional vectors $\mathbb{R}^S = \mathbb{R}^k$ and when $S = [\![h]\!] \times [\![k]\!]$ we recover the space of matrices $\mathbb{R}^S = \mathbb{R}^{h \times k}$. If T is a set and $f : \mathbb{R}^S \to \mathbb{R}^T$, then for any $X \in \mathbb{R}^S$ and $t \in T$, we denote $f(X)[t] = (f(X))_t$ to index the output of f.

We view $\mathbb{R}^S$ as a vector space under the usual pointwise operations, $(f + g)[s] = f[s] + g[s]$ and $(\lambda f)[s] = \lambda f[s]$, together with the canonical basis $(e_s)_{s\in S}$ defined by $e_s[t] = \mathbf{1}_{\{s=t\}}$.

For $A \in \mathbb{R}^{S\times T}$, we denote by $A[s,:] = (A[s,t])_{t\in T} \in \mathbb{R}^T$ the slice of A obtained for $s \in S$.

1.3 Tensor products

Given sets S and T, the tensor product of $f \in \mathbb{R}^S$ and $g \in \mathbb{R}^T$ is the element $f \otimes g \in \mathbb{R}^{S\times T}$ defined by

$$(f \otimes g)[s,t] = f[s]\, g[t],$$

which is also associative. The induced canonical basis of $\mathbb{R}^{S\times T}$ is $(e_s \otimes e_t)_{s\in S,\, t\in T}$.

When S and T are finite, any $A \in \mathbb{R}^{S\times T}$ can be identified with the linear map $\mathbb{R}^T \to \mathbb{R}^S$ acting by

$$(Af)[s] = \sum_{t\in T} A[s,t]\, f[t].$$

The identity matrix $\mathrm{Id}_S \in \mathbb{R}^{S\times S}$ is defined as $\mathrm{Id}_S = \sum_{s\in S} e_s \otimes e_s$, that is $\mathrm{Id}_S[s,t] = \mathbf{1}_{\{s=t\}}$. When $S = [\![k]\!]$, we write Id_k for a matrix in $\mathbb{R}^{k\times k}$.

If $S = S_1 \times S_2$, $T = T_1 \times T_2$, and $A \in \mathbb{R}^{S_1\times T_1}$, $B \in \mathbb{R}^{S_2\times T_2}$, then, under the natural identification of $\mathbb{R}^{(S_1\times T_1)\times(S_2\times T_2)}$ with $\mathbb{R}^{(S_1\times S_2)\times(T_1\times T_2)}$, one has

$$(A \otimes B)(f \otimes g) = (Af) \otimes (Bg) \qquad \text{for all } f \in \mathbb{R}^{T_1},\ g \in \mathbb{R}^{T_2}.$$

1.4 Norms

When $S = T$, any matrix $A \in \mathbb{R}^{S\times S}$ is said to be *symmetric*, if A equals its transpose $A^{\mathsf{T}}[s,t] = A[t,s]$, and *positive semidefinite*, if for every finite $T \subseteq S$

$$\sum_{s,t\in T} v[s]A[s,t]v[t] \geq 0, \qquad \text{for all } v \in \mathbb{R}^T. \tag{1}$$

For finite S, the trace of $A \in \mathbb{R}^{S\times S}$ is $\operatorname{tr}(A) = \sum_{s\in S} A[s,s]$. This yields the inner product on $\mathbb{R}^S$,

$$\langle f, g\rangle = \operatorname{tr}(f \otimes g) = \sum_{s\in S} f[s]\, g[s],$$

with the induced norm

$$\|f\| = \sqrt{\operatorname{tr}(f \otimes f)} = \left(\sum_{s\in S} |f[s]|^2\right)^{1/2}. \tag{2}$$

This recovers the Euclidean norm on $\mathbb{R}^k$, given by $\|f\| = \sqrt{\sum_i f[i]^2}$, and the Frobenius (or Hilbert-Schmidt) norm on $\mathbb{R}^{h\times k}$, given by $\|A\| = \sqrt{\sum_{i,j} A[i,j]^2}$. The norm is multiplicative:

$$\|f \otimes g\| = \|f\| \cdot \|g\|.$$

The induced operator norm of a linear operator $A \in \mathbb{R}^{T\times S}$ will be denoted instead as

$$\|A\|_{\mathrm{op}} \coloneqq \sup_{\substack{v\in\mathbb{R}^S \\ \|v\|=1}} \|Av\|.$$

We will also use the entry-wise 1-norm

$$\|A\|_{1,1} \coloneqq \sum_{t\in T}\sum_{s\in S} |A[t,s]|.$$

1.5 Lipschitz functions

For finite sets S and T, a function $f\colon \mathbb{R}^S \to \mathbb{R}^T$ is called Lipschitz, if there exists a constant $\operatorname{Lip}(f) < \infty$, such that

$$\|f(z) - f(z')\| \le \operatorname{Lip}(f)\, \|z - z'\| \qquad \text{for all } z, z' \in \mathbb{R}^S.$$

If instead $f\colon \mathbb{R} \to \mathbb{R}$ and is applied element-wise, then for $A, B \in \mathbb{R}^{S\times T}$, it holds

$$\|f(A) - f(B)\|^2 \le \operatorname{Lip}(f)^2\, \|A - B\|^2.$$

Indeed,

$$\begin{aligned}\|f(A) - f(B)\|^2 &= \sum_{s,t} \big(f(A)[s,t] - f(B)[s,t]\big)^2 \le \sum_{s,t} \big(\operatorname{Lip}(f)|A[s,t] - B[s,t]|\big)^2 \\ &= \sum_{s,t} \operatorname{Lip}(f)^2 (A[s,t] - B[s,t])^2 = \operatorname{Lip}(f)^2\, \|A - B\|^2.\end{aligned}$$

1.6 Matrix square roots

Let S be finite and let $A \in \mathbb{R}^{S\times S}$ be symmetric and positive semidefinite. By the spectral theorem there exists an orthogonal $U \in \mathbb{R}^{S\times S}$ (meaning $UU^{\mathsf{T}} = \mathrm{Id}_S$) such that $U^{\mathsf{T}}AU = D = \sum_{s\in S} d_s e_s \otimes e_s$, where the eigenvalues satisfy $d_s \geq 0$. The square root of A is then

$$\sqrt{A} = U\sqrt{D}U^{\mathsf{T}}, \qquad \sqrt{D} = \sum_{s\in S} \sqrt{d_s}\, e_s \otimes e_s. \tag{3}$$

Moreover, $\sqrt{A}$ is itself symmetric and positive semidefinite.

We denote the smallest eigenvalue of A by $\lambda(A) = \min_{s\in S} d_s$ and the smallest strictly positive eigenvalue of A by $\lambda_+(A)$.

For symmetric matrix A and vector $v \in \mathbb{R}^S$ that is different for zero, it holds

$$\lambda(A) \leq \frac{v^{\mathsf{T}}Av}{v^{\mathsf{T}}v}. \tag{4}$$

If A and B are both symmetric and positive semidefinite, then the bound by Hemmen and Ando (1980, Proposition 3.2) gives

$$\left\|\sqrt{A} - \sqrt{B}\right\| \leq \frac{1}{\sqrt{\lambda(A)}}\|A - B\|, \tag{5}$$

where $\|\cdot\|$ is the Frobenius norm introduced above. When $\lambda(A) = 0$, the right-hand side is interpreted as $+\infty$.

Additionally, if $A \in \mathbb{R}^{S\times S}$ and $B \in \mathbb{R}^{T\times T}$ are both symmetric and positive semidefinite, then

$$\sqrt{A \otimes B} = \sqrt{A} \otimes \sqrt{B},$$

where both sides are viewed as elements of $\mathbb{R}^{(S\times T)\times(S\times T)}$. This follows from the fact that if $U^{\mathsf{T}}AU = D_A$ and $V^{\mathsf{T}}BV = D_B$ are spectral decompositions, then $(U\otimes V)^{\mathsf{T}}(A\otimes B)(U\otimes V) = D_A \otimes D_B$ is a spectral decomposition of $A\otimes B$, and $\sqrt{D_A \otimes D_B} = \sqrt{D_A} \otimes \sqrt{D_B}$ holds entry-wise since $\sqrt{d_s \cdot d_t} = \sqrt{d_s}\sqrt{d_t}$.

1.7 Random variables and Gaussian processes

Throughout this subsection, S and T are finite sets. Let X be a random variable taking values in $\mathbb{R}^S$, defined on some probability space $(\Omega, \mathcal{A}, \mathbb{P})$. We

denote by $\mathbb{P}_X$ the distribution (law) of X and write $X \overset{d}{=} Y$ if X and Y have the same distribution. For $p \geq 1$, define the L^p norm of X by

$$\|X\|_{L^p} := \left(\mathbb{E}[\|X\|^p]\right)^{1/p}.$$

The mean $\mathbb{E}[X] \in \mathbb{R}^S$ is taken coordinatewise,

$$\mathbb{E}[X][s] = \mathbb{E}\big[X[s]\big], \qquad s \in S.$$

The *second moment* of X is the matrix

$$\mathbb{E}[X \otimes X] = \big(\mathbb{E}[X[s]X[s']]\big)_{s,s' \in S} \in \mathbb{R}^{S \times S},$$

which is symmetric and positive semi-definite. We assume throughout, without further mention, that the moments written down are finite. We write $\Sigma(X)$ for the *covariance* of X, that is, the second moment of the centred variable $X - \mathbb{E}[X]$. For a centred random variable X, the second moment and the covariance coincide.

From (2), we have that $\|X\|^2 = \operatorname{tr}(X \otimes X)$ and using the linearity of the trace to change it with the expectation, we get $\mathbb{E}\left[\|X\|^2\right] = \operatorname{tr}(\mathbb{E}[X \otimes X])$. So it follows that

$$\mathbb{E}\left[\|X\|^2\right] = \operatorname{tr}\left(\Sigma(X)\right), \qquad \text{if } X \text{ is centered.} \tag{6}$$

When A takes values in $\mathbb{R}^{T \times S}$, the same definition gives a second moment

$$\mathbb{E}[A \otimes A] \in \mathbb{R}^{(T \times S) \times (T \times S)},$$

which it is sometimes convenient to rearrange into $\mathbb{R}^{(T \times T) \times (S \times S)}$.

Definition 1.1 (Centred Gaussian process)**.** Let $\mathcal{D}$ be a set. A **centered Gaussian process** indexed by $\mathcal{D}$ with values in $\mathbb{R}^T$ is a collection $G = (G(X))_{X \in \mathcal{D}}$ of $\mathbb{R}^T$-valued random variables such that for every finite $\mathcal{X} = \{X_1, \ldots, X_k\} \subseteq \mathcal{D}$, the random variable

$$G(\mathcal{X}) := \sum_{\alpha=1}^{k} G(X_\alpha) \otimes e_\alpha \in \mathbb{R}^{T \times k}$$

is a centred multivariate Gaussian. Its joint distribution is determined by the **covariance kernel**

$$K \colon (\mathcal{D} \times T) \times (\mathcal{D} \times T) \to \mathbb{R}, \qquad K[(X,t),(X',t')] = \mathbb{E}\left[G(X)[t]G(X')[t']\right],$$

which is symmetric and positive semi-definite. We write $G \sim \mathcal{N}(K)$.

In order to prove the Wasserstein property (13) for $p \geq 2$ in the later section, we first prove the following lemma. This result was given as equation 2.25 in Trevisan, 2023, equation (2.25), but without proof or explicit constants.

Lemma 1.1. Let $p \geq 2$, S be a finite set, and let X be a centred Gaussian random variable taking values in $\mathbb{R}^S$ with the covariance matrix K. Then

$$\|X\|_{L^p} \leq C_{\text{mom}} \sqrt{\text{tr}(K)} = C_{\text{mom}} \left\| \sqrt{K} \right\|, \tag{7}$$

where

$$C_{\text{mom}} = C_{\text{mom}}(p) = \left(\frac{2^{p/2} \Gamma\left(\frac{p+1}{2}\right)}{\sqrt{\pi}} \right)^{\frac{1}{p}}.$$

Proof. We begin the proof by finding a different representation for $\|X\|$. Since X is a centred Gaussian with covariance K, we may write $X = \sqrt{K}\tilde{Z}$, where $\tilde{Z} \sim \mathcal{N}(0, \text{Id}_S)$. Let $K = UDU^\mathsf{T}$ be the spectral decomposition of K as in (3). Set $Z \coloneqq U^\mathsf{T}\tilde{Z}$ and since U is an orthogonal matrix, we have $Z \sim \mathcal{N}(0, \text{Id}_S)$. Due to the orthogonality of U, we have $\|Uv\| = \|v\|$ for any $v \in \mathbb{R}^S$, therefore

$$\|X\|^2 = \left\| \sqrt{K}\tilde{Z} \right\|^2 = \left\| U\sqrt{D}U^\mathsf{T}\tilde{Z} \right\|^2 = \left\| U\sqrt{D}Z \right\|^2 = \left\| \sqrt{D}Z \right\|^2 = \sum_{s \in S} d_s Z_s^2.$$

Recall that $\text{tr}(K) = \sum_{s \in S} d_s$ and note that the function $t \mapsto t^{p/2}$ is convex on $[0, \infty)$. Applying the Jensen inequality, we have

$$\left(\sum_{s \in S} \frac{d_s}{\text{tr}(K)} Z_s^2 \right)^{p/2} \leq \sum_{s \in S} \frac{d_s}{\text{tr}(K)} Z_s^p.$$

Multiplying both sides by $\text{tr}(K)^{p/2}$ and taking the expectation gives

$$\begin{aligned} \mathbb{E}[\|X\|^p] = \mathbb{E}\left[\left(\sum_{s \in S} d_s Z_s^2 \right)^{p/2} \right] &\leq \text{tr}(K)^{p/2-1} \sum_{s \in S} d_s \, \mathbb{E}[|Z_s^p|] \\ &= \text{tr}(K)^{p/2-1} \, \mathbb{E}[|Z_{s_0}^p|] \sum_{s \in S} d_s = \text{tr}(K)^{p/2} \, \mathbb{E}[|Z_{s_0}^p|], \end{aligned} \tag{8}$$

where $s_0 \in S$ is arbitrary, since all Z_s are i.i.d for $s \in S$.

For a standard centered normal random variable, the absolute p-th moment is (Papoulis, 1991, p. 148)

$$\mathbb{E}[|Z_{s_0}|^p] = \frac{2^{p/2} \Gamma\left(\frac{p+1}{2}\right)}{\sqrt{\pi}}$$

and take

$$C_{\text{mom}}(p) \coloneqq \left(\frac{2^{p/2}\Gamma\left(\frac{p+1}{2}\right)}{\sqrt{\pi}} \right)^{\frac{1}{p}}.$$

To conclude, note that since K is symmetric, we have that

$$\operatorname{tr}(K) = \operatorname{tr}\left(\sqrt{K}\sqrt{K}\right) = \operatorname{tr}\left(\sqrt{K}\sqrt{K}^{\mathsf{T}}\right) = \operatorname{tr}\left(\sqrt{K} \otimes \sqrt{K}\right) = \|\sqrt{K}\|^2.$$

Taking the p-th roof of (8) and substituting in C_{mom} and $\operatorname{tr}(K) = \|\sqrt{K}\|^2$ gives

$$\|X\|_{L^p} \leq C_{\text{mom}} \sqrt{\operatorname{tr}(K)} = C_{\text{mom}} \left\|\sqrt{K}\right\|.$$

□

Remark 1.1. Note that for $p = 2$, we have $\Gamma(\frac{3}{2}) = \frac{\sqrt{\pi}}{2}$ and therefore $C_{\text{mom}}(2) = 1$. Then the inequality (7) is actually an equality as seen in (6).

1.8 Wasserstein distance

This section presents some preliminary results about the Wasserstein distance from Villani (2009).

Definition 1.2 (Wasserstein distance)**.** Let $p \in [1, \infty)$ and let μ, ν be two probability measures defined on a Polish space (M, d_M). Denote by $\Gamma(\mu, \nu)$ the collection of all *couplings* of μ and ν, that is all joint distributions γ on $M \times M$ whose first and second marginals are μ and ν respectively. The **p-Wasserstein distance** between μ and ν is given by

$$\mathcal{W}_p(\mu, \nu) = \inf \left\{ \left(\mathbb{E}\left[d_M(X, Y)^p\right]\right)^{1/p} \mid (X, Y) \sim \gamma \in \Gamma(\mu, \nu) \right\}.$$

With a slight abuse of notation, we will often write $\mathcal{W}_p(X, Y) = \mathcal{W}_p(\mu, \nu)$ for any $X \sim \mu$ and $Y \sim \nu$.

Remark 1.2. For the distance to be finite, the probability measures μ and ν need to have finite p-moments, that is for arbitrary $x_0 \in M$, it holds $\int_M d_M(x_0, x)^p \, \mu(dx) < \infty$.

Let S be a finite set. In the context of this thesis, we work on the set $M = \mathbb{R}^S$ and with the distance $d(x, y) = \|x - y\|$, $x, y \in \mathbb{R}^S$. The quadratic Wasserstein distance we are interested in is then

$$\mathcal{W}_2(\mu, \nu) = \inf \left\{ \sqrt{\mathbb{E}\left[\|X - Y\|^2\right]} \mid (X, Y) \sim \gamma \in \Gamma(\mu, \nu) \right\}.$$

We now list some properties of the Wasserstein distance that we will use later.

Lemma 1.2 (Properties of Wasserstein distance)**.** Let $p \in [1, \infty)$ and let X, Y and Z be random variables with values on $\mathbb{R}^S$. Then the following properties hold.

1. If X and Y are defined on the same probability space, then
$$\mathcal{W}_p(X, Y) \leq \left(\mathbb{E}\left[\|X - Y\|^p\right]\right)^{1/p} = \|X - Y\|_{L^p}. \tag{9}$$
2. The triangle inequality holds
$$\mathcal{W}_p(X, Z) \leq \mathcal{W}_p(X, Y) + \mathcal{W}_p(Y, Z). \tag{10}$$
3. If Z is independent of X and Y, then
$$\mathcal{W}_p(X + Z, Y + Z) \leq \mathcal{W}_p(X, Y). \tag{11}$$
4. Let Z take values on $\mathbb{R}^T$ where T is finite. The function $\mathcal{W}_p^p$ is convex
$$\mathcal{W}_p^p(X, Y) \leq \int_{\mathbb{R}^T} \mathcal{W}_p^p(\mathbb{P}_{X|Z=z}, \mathbb{P}_Y)\, d\,\mathbb{P}_Z(z). \tag{12}$$
5. Let $p \geq 2$. If X and Y are centred Gaussian random variables, then
$$\mathcal{W}_p(X, Y) \leq C_{\text{mom}}(p) \left\| \sqrt{\Sigma(X)} - \sqrt{\Sigma(Y)} \right\|. \tag{13}$$

Proof. For the proof of properties 1–4, refer to Villani (2009). To see that 5. holds, let Z be a centred standard Gaussian variable. Take $\widetilde{X} = \sqrt{\Sigma(X)}Z$ and $\widetilde{Y} = \sqrt{\Sigma(Y)}Z$. Then $\widetilde{X} \overset{d}{=} X$ and $\widetilde{Y} \overset{d}{=} Y$, which means that $(\widetilde{X}, \widetilde{Y})$ is a valid coupling. Now it follows that

$$\begin{aligned}\mathcal{W}_p(\widetilde{X}, \widetilde{Y}) &\overset{(9)}{\leq} \|\widetilde{X} - \widetilde{Y}\|_{L^p} = \left\| \left(\sqrt{\Sigma(X)} - \sqrt{\Sigma(Y)} \right) Z \right\|_{L^p} \\ &\overset{(7)}{\leq} C_{\text{mom}}(p) \left\| \sqrt{\Sigma(X)} - \sqrt{\Sigma(Y)} \right\|.\end{aligned}$$

□

Remark 1.3. By convexity in property 12, we mean that it is convex in its first argument over distributions:

$$\begin{aligned}\mathcal{W}_p^p(X, Y) = \mathcal{W}_p^p(\mathbb{P}_X, \mathbb{P}_Y) &= \mathcal{W}_p^p\left(\int_{\mathbb{R}^T} \mathbb{P}_{X|Z=z}\; d\,\mathbb{P}_Z(z), \mathbb{P}_Y \right) \\ &\leq \int_{\mathbb{R}^T} \mathcal{W}_p^p(\mathbb{P}_{X|Z=z}, \mathbb{P}_Y)\, d\,\mathbb{P}_Z(z).\end{aligned}$$

1.9 Deep neural networks

A classical *fully connected neural network* (FCNN) is a sequence of layers in which every output node of each layer is connected to every input node, with a different set of weights per output node. This is in contrast to a *deep convolutional neural network* (CNN), in which each output node is connected only to a small *patch* of input nodes and the weights are shared across output positions. As the name suggests, this architecture achieves this by performing a convolution at each layer (more precisely, a cross-correlation, as in most implementations, the weights associated with the convolution operation are not flipped).

Before we introduce the definition of CNN, we give a quick overview of the fully connected neural network. For both architectures, let $L \geq 1$ and $\phi : \mathbb{R} \to \mathbb{R}$ be an activation function that is applied element-wise. Note that the weights $W^{(\ell)}$ and biases $b^{(\ell)}$ in this section do not need to be random.

1.9.1 Fully connected neural network

Let $n^{(0)}, \ldots, n^{(L)} \in \mathbb{N}$. For every $\ell \in \{0, \ldots, L-1\}$, denote the weight matrix by $W^{(\ell)} \in \mathbb{R}^{n^{(\ell+1)} \times n^{(\ell)}}$ and the bias vector by $b^{(\ell)} \in \mathbb{R}^{n^{(\ell+1)}}$. The FCNN is defined as follows: for every input $X \in \mathbb{R}^{n^{(0)}}$ set

$$Z^{(1)}(X) = W^{(0)}X + b^{(0)},$$

and then recursively, for $\ell \in \{2, \ldots, L\}$,

$$Z^{(\ell)} = W^{(\ell-1)}\phi(Z^{(\ell-1)}) + b^{(\ell-1)}.$$

Here L is the number of *layers* and the superscript $^{(\ell)}$ denotes objects associated with layer ℓ. The vector $Z^{(\ell)}$ is called the *pre-activation* of layer ℓ and each entry $Z^{(\ell)}[i]$, for $i \in [\![n^{(\ell)}]\!]$, is called a *neuron*. The quantity $n^{(\ell)}$ denotes the number of neurons in layer ℓ. Note that each neuron has its own set of weight parameters stored in the corresponding column of $W^{(\ell)}$.

1.9.2 Convolutional neural network

Before defining the CNN, we give some intuition about the terms channel and patch, and introduce the formal definition later.

A *channel* of layer ℓ refers to a slice of the layer's output. For an RGB image at the input, the channels are the three colour components red, green, and blue, so $C^{(0)} = 3$. In deeper layers, each channel is produced by a different set of learned weights. As different weights learn to represent distinct features, the number of channels determines how many features the network can track at each layer.

To produce a single entry in a channel of the next layer, the convolution selects a *patch*, a small subregion of the input, and combines it with a weight matrix to generate a single output value. For example, in a two-dimensional convolution with a 3×3 patch, every output pixel is obtained by taking the inner product between a fixed 3×3 weight matrix and the corresponding 3×3 region of the input. Then the same weight matrix is reused across all output positions, which is referred to as *weight sharing*. Compare it to the FCNN, where each output neuron has its own set of weights, which are not shared with other output neurons.

Let $C^{(0)}, \dots, C^{(L)} \in \mathbb{N}$ denote the number of channel in each layer. Let $D \in \mathbb{N}$ be the *number of spatial dimensions* and $\boldsymbol{F}^{(\ell)} \in \mathbb{N}^D$ the *spatial dimensions* of the output of layer ℓ. For images, one typically has $D = 2$, whereas $D = 1$ models a one-dimensional sequence, such as audio. We write $\boldsymbol{P}^{(\ell)} \in \mathbb{N}^D$ for the *patch dimensions* of the convolution between layers ℓ and $\ell + 1$, and require $\boldsymbol{P}^{(\ell)}[d] \leq \boldsymbol{F}^{(\ell)}[d]$ for each $d \in [\![D]\!]$, so that the patch does not exceed the spatial extent of the input. Throughout, we use i and j to index channels, $\boldsymbol{q}$ to index spatial positions, and $\boldsymbol{p}$ to index positions within a patch.

To express the convolution at a single output location, we need a function that, given an output position $\boldsymbol{q} \in [\![\boldsymbol{F}^{(\ell)}]\!]$ and a patch index $\boldsymbol{p} \in [\![\boldsymbol{P}^{(\ell-1)}]\!]$, returns the corresponding spatial position $\tilde{\boldsymbol{q}}(\boldsymbol{p})$ in the previous layer. As $\boldsymbol{p}$ ranges over $[\![\boldsymbol{P}^{(\ell-1)}]\!]$, $\tilde{\boldsymbol{q}}(\boldsymbol{p})$ traces out the patch in the input (see figure 2). For now, we assume that this function exists and defer the formal definition to a later section (Definition 1.3).

For every $\ell \in \{0, \dots, L-1\}$, denote the weights by $W^{(\ell)} \in \mathbb{R}^{C^{(\ell+1)} \times C^{(\ell)} \times \boldsymbol{P}^{(\ell)}}$ and the biases by $b^{(\ell)} \in \mathbb{R}^{C^{(\ell+1)}}$. For $\boldsymbol{q} \in [\![\boldsymbol{F}^{(1)}]\!]$ and $i \in [\![C^{(1)}]\!]$, we first define

$$Z^{(1)}(X)[i, \boldsymbol{q}] = \sum_{j=1}^{C^{(0)}} \sum_{\boldsymbol{p}=\boldsymbol{1}}^{\boldsymbol{P}^{(0)}} W^{(0)}[i, j, \boldsymbol{p}] X[j, \tilde{\boldsymbol{q}}(\boldsymbol{p})] + b^{(0)}[i], \tag{14}$$

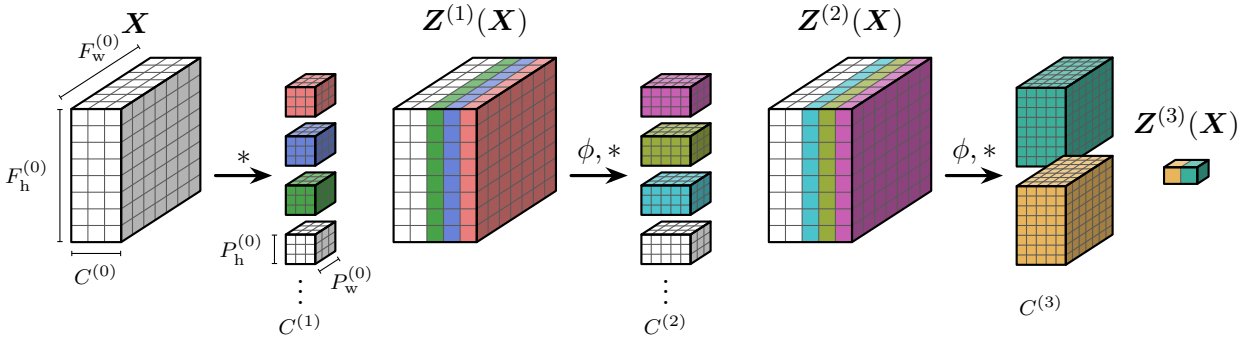

Figure 1: An example of a deep convolutional neural network inspired by Garriga-Alonso and Wilk (2021). This network has $D = 2$ spatial dimensions and $L = 3$ layers. The each coloured cube represents the weight slices $W^{(\ell)}[i, :, :]$, where the colour is different for each $i \in [\![C^{(\ell+1)}]\!]$. The information flows through the network as follows. The input $X \in \mathbb{R}^{C^{(0)} \times F_h^{(0)} \times F_w^{(0)}}$ is divided into possibly overlapping patches of size $\boldsymbol{P}^{(0)} = P_h^{(0)} \times P_w^{(0)}$. For each set of weights $W^{(\ell)}[i, :, :]$, we compute the inner product between the weights and all the patches, then add the bias term to the result. This produces one slice along a single channel in the next layer (shown with same colour as the weights). Performing this $C^{(1)}$ times for all sets of weights and concatenating the results gives us $Z^{(1)}$. We then apply the activation ϕ and repeat the process.

and then recursively, for $\ell \in \{2, \dots, L\}$, $\boldsymbol{q} \in [\![\boldsymbol{F}^{(\ell)}]\!]$, and $i \in [\![C^{(\ell)}]\!]$, define

$$Z^{(\ell)}(X)[i, \boldsymbol{q}] = \sum_{j=1}^{C^{(\ell-1)}} \sum_{\boldsymbol{p}=\boldsymbol{1}}^{\boldsymbol{P}^{(\ell-1)}} W^{(\ell-1)}[i, j, \boldsymbol{p}]\, \phi\big(Z^{(\ell-1)}(X)[j, \tilde{\boldsymbol{q}}(\boldsymbol{p})]\big) + b^{(\ell-1)}[i], \tag{15}$$

where $Z^{(\ell)}(X) \in \mathbb{R}^{C^{(\ell)} \times \boldsymbol{F}^{(\ell)}}$. See figure 1 for an illustration.

The vector $Z^{(\ell)}(X)$ is the *pre-activation* of layer ℓ, and the slices $Z^{(\ell)}[i, :]$ are the *channels* of layer ℓ.

1.9.3 Bias-free CNN and multiple inputs

It is convenient to separate the part of the pre-activation that does not depend on the current layer's bias. For the base case, write

$$Z_0^{(1)}(X)[i, \boldsymbol{q}] = \sum_{j=1}^{C^{(0)}} \sum_{\boldsymbol{p}=\boldsymbol{1}}^{\boldsymbol{P}^{(0)}} W^{(0)}[i, j, \boldsymbol{p}] X[j, \tilde{\boldsymbol{q}}(\boldsymbol{p})],$$

and for $\ell \in \{2, \dots, L\}$,

$$Z_0^{(\ell)}(X)[i, \boldsymbol{q}] = \sum_{j=1}^{C^{(\ell-1)}} \sum_{\boldsymbol{p}=\boldsymbol{1}}^{\boldsymbol{P}^{(\ell-1)}} W^{(\ell-1)}[i, j, \boldsymbol{p}]\, \phi\big(Z^{(\ell-1)}(X)[j, \tilde{\boldsymbol{q}}(\boldsymbol{p})]\big)\,,$$

so that

$$Z^{(\ell)}(X)[i, \boldsymbol{q}] = Z_0^{(\ell)}(X)[i, \boldsymbol{q}] + b^{(\ell-1)}[i].$$

Note that the activations in $Z_0^{(\ell)}$ are computed using the entirety of $Z^{(\ell-1)}$, including its bias term, while only the bias of the current layer is omitted.

Finally, if $\mathcal{X} = \{X_\alpha\}_{\alpha \in [\![k]\!]} \subset \mathbb{R}^{C^{(0)} \times F^{(0)}}$ is a finite subset of inputs, then we define

$$Z^{(\ell)}(\mathcal{X}) = \sum_{\alpha=1}^{k} Z^{(\ell)}(X_\alpha) \otimes e_\alpha \in \mathbb{R}^{C^{(\ell)} \times \boldsymbol{F}^{(\ell)} \times k}.$$

1.9.4 Patch function

We now make the patch function $\tilde{\boldsymbol{q}}$ precise. Often, convolutions are defined by subtracting the indices of one tensor from another. This can be quite cumbersome, even when $D = 2$. To abstract away these details and make the presentation cleaner, we similarly introduce the patch function as was done by Garriga-Alonso and Wilk (2021). The three standard parameters of a convolution operator enter as the parameters of $\tilde{\boldsymbol{q}}$:

- *padding* enlarges the input by $\pi_d^{(\ell)}$ extra positions on each side of dimension d, which would otherwise shrink the output;
- *stride* $s_d^{(\ell)} > 1$ means the patch is moved by $s_d^{(\ell)}$ input positions for every single step in the output;
- *dilation* $h_d^{(\ell)} > 1$ inserts holes between weight entries, so that consecutive patch indices select input positions that are $h_d^{(\ell)}$ apart.

We refer to Dumoulin and Visin (2016) for a detailed treatment of convolution arithmetic.

Definition 1.3 (Patch function)**.** For each spatial dimension $d \in [\![D]\!]$ and layer $\ell+1 \in [\![L]\!]$, fix the stride $s_d^{(\ell)} \in \mathbb{N}$, the dilation $h_d^{(\ell)} \in \mathbb{N}$, and the padding $\pi_d^{(\ell)} \in \mathbb{N}_0$ of the convolution operator. For a spatial location $\boldsymbol{q} \in [\![\boldsymbol{F}^{(\ell+1)}]\!]$

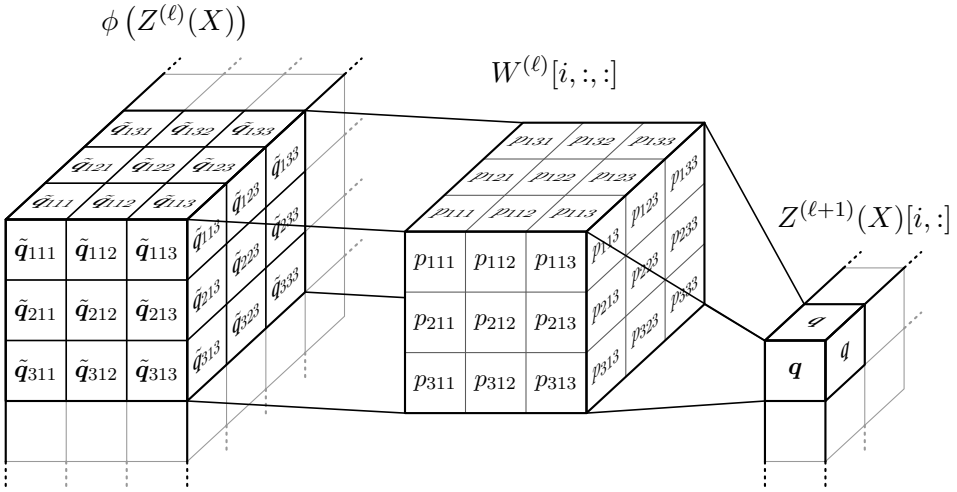

Figure 2: Consider an example where $D = 2$, $\boldsymbol{P}^{(\ell)} = (3, 3)$ and $C^{(\ell)} = 3$. As a notational shorthand, write $p_{abj} = (a, b, j)$ and $\tilde{\boldsymbol{q}}_{abj} = (\tilde{\boldsymbol{q}}(a, b), j)$, where $a, b, j \in [\![3]\!]$. Here $(a, b) \in [\![\boldsymbol{P}^{(\ell)}]\!]$ is patch index and j is a channel index. Given an output position $\boldsymbol{q}$ in layer $\ell + 1$ and a patch index $\boldsymbol{p}$, the patch function returns the corresponding spatial position $\tilde{\boldsymbol{q}}(\boldsymbol{p})$ in layer ℓ. As $\boldsymbol{p}$ varies over $[\![\boldsymbol{P}^{(\ell)}]\!]$ the function traces out the patch in the input. The same weight $W^{(\ell)}[i, :, :]$ is used for each $\boldsymbol{q} \in [\![\boldsymbol{F}^{(\ell+1)}]\!]$ to produce the elements of $Z^{(\ell+1)}(X)[i, :]$.

in the output of layer $\ell + 1$, the d-th component of the **patch function** $\tilde{q}_d : [\![P_d^{(\ell)}]\!] \to \mathbb{Z}$ is

$$\tilde{q}_d(p_d) \coloneqq s_d^{(\ell)}(q_d - 1) + h_d^{(\ell)}(p_d - 1) - \pi_d^{(\ell)} + 1. \tag{16}$$

Given a D-tuple, we write $\tilde{\boldsymbol{q}}(\boldsymbol{p}) \coloneqq (\tilde{q}_1(p_1), \ldots, \tilde{q}_D(p_D))$. We call the image $\tilde{\boldsymbol{q}}([\![\boldsymbol{P}^{(\ell)}]\!])$ a **patch**.

For any $A \in \mathbb{R}^{\boldsymbol{F}^{(\ell)}}$ we adopt the convention $A[\boldsymbol{r}] \coloneqq 0$ whenever $\boldsymbol{r} \notin [\![\boldsymbol{F}^{(\ell)}]\!]$, so that $A[\tilde{\boldsymbol{q}}(\boldsymbol{p})]$ is well-defined for every $\boldsymbol{p} \in [\![\boldsymbol{P}^{(\ell)}]\!]$ and every $\boldsymbol{q} \in [\![\boldsymbol{F}^{(\ell+1)}]\!]$.

Remark 1.4. Note that the function $\tilde{\boldsymbol{q}}(\cdot)$ also depends on the value $\boldsymbol{q}$.

Example 1.1. Consider a $D = 2$ convolution with input spatial size $\boldsymbol{F}^{(\ell)} = (5, 5)$, output spatial size $\boldsymbol{F}^{(\ell+1)} = (2, 2)$, patch size $\boldsymbol{P}^{(\ell)} = (3, 3)$, stride $\boldsymbol{s}^{(\ell)} = (2, 2)$, dilation $\boldsymbol{h}^{(\ell)} = (2, 2)$, and padding $\boldsymbol{\pi}^{(\ell)} = (1, 1)$. For simplicity, we'll drop layer superscripts and work with a single channel. Substituting into (16) gives for each spatial dimension $d \in \{1, 2\}$

$$\tilde{q}_d(p_d) = 2(q_d - 1) + 2(p_d - 1) - 1 + 1 = 2q_d + 2p_d - 4.$$

Consider the output position $\boldsymbol{q} = (1, 1)$. As $\boldsymbol{p}$ ranges over $[\![\boldsymbol{P}^{(\ell)}]\!]$, the patch function selects the nine input positions:

$$\begin{aligned}
&\tilde{\boldsymbol{q}}(1,1) = (0,0), \quad \tilde{\boldsymbol{q}}(1,2) = (0,2), \quad \tilde{\boldsymbol{q}}(1,3) = (0,4),\\
&\tilde{\boldsymbol{q}}(2,1) = (2,0), \quad \tilde{\boldsymbol{q}}(2,2) = (2,2), \quad \tilde{\boldsymbol{q}}(2,3) = (2,4),\\
&\tilde{\boldsymbol{q}}(3,1) = (4,0), \quad \tilde{\boldsymbol{q}}(3,2) = (4,2), \quad \tilde{\boldsymbol{q}}(3,3) = (4,4).
\end{aligned}$$

The dilation $h_d = 2$ causes consecutive patch indices $p_d \in \{1, 2\}$ to land on indices $\{0, 2\}$ rather than $\{0, 1\}$. With weights $W \in \mathbb{R}^{\boldsymbol{P}^{(\ell)}}$ and input $X \in \mathbb{R}^{\boldsymbol{F}^{(\ell)}}$, the convolution at $\boldsymbol{q} = (1, 1)$ reads as

$$\begin{aligned}
Y[1,1] &= \sum_{\boldsymbol{p} \in [\![\boldsymbol{P}^{(\ell)}]\!]} W[\boldsymbol{p}] X[\tilde{\boldsymbol{q}}(\boldsymbol{p})]\\
&= W[1,1]X[0,0] + W[1,2]X[0,2] + W[1,3]X[0,4]\\
&+ W[2,1]X[2,0] + W[2,2]X[2,2] + W[2,3]X[2,4]\\
&+ W[3,1]X[4,0] + W[3,2]X[4,2] + W[3,3]X[4,4]\\
&= W[2,2]X[2,2] + W[2,3]X[2,4] + W[3,2]X[4,2] + W[3,3]X[4,4].
\end{aligned}$$

Five of the nine terms are zero, because of the out-of-bounds convention in Definition 1.3.

Remark 1.5. The patch function allows us to simplify the notation and abstract away the details of the convolution. With the patch function, for every $D \in \mathbb{N}$ the convolution is just

$$Y[\boldsymbol{q}] = \sum_{\boldsymbol{p} \in [\![\boldsymbol{P}^{(\ell)}]\!]} W[\boldsymbol{p}]\, X[\tilde{\boldsymbol{q}}(\boldsymbol{p})], \qquad \boldsymbol{q} \in [\![\boldsymbol{F}^{(\ell)}]\!].$$

Furthermore, if we were to use circular padding instead of zero padding, the formula would remain the same. Compare it to the case without the patch function. Already for $D = 2$,

$$\begin{aligned}
Y[q_1, q_2] = \sum_{p_1=1}^{P_1^{(\ell)}} \sum_{p_2=1}^{P_2^{(\ell)}} W[p_1, p_2] X\Big[&s_1^{(\ell)}(q_1 - 1) + h_1^{(\ell)}(p_1 - 1) - \pi_1^{(\ell)} + 1,\\
&s_2^{(\ell)}(q_2 - 1) + h_2^{(\ell)}(p_2 - 1) - \pi_2^{(\ell)} + 1\Big],
\end{aligned}$$

and for $D = 3$, each summation, each weight slot, and each input slot acquires a further component.

Remark 1.6 (Comparison with Garriga-Alonso and Wilk (2021)). The patch function (16) differs from the one used by Garriga-Alonso and Wilk (2021), who define

$$\tilde{q}_d^{\mathrm{GAW}}(p_d) \coloneqq s_d^{(\ell)} q_d - h_d^{(\ell)} \left(p_d - \left\lceil P_d^{(\ell)}/2 \right\rceil \right).$$

The patch function $\tilde{\boldsymbol{q}}^{\mathrm{GAW}}$ uses *centred placement with flipping*, whereas $\tilde{\boldsymbol{q}}$ uses *corner placement without flipping*. Flipping makes the operation a true convolution, which otherwise is cross-correlation. Definition 1.3 also makes the padding parameter $\pi_d^{(\ell)}$ explicit.

The drawback of their convention appears when using global mean pooling at the last layer, where $\boldsymbol{F}^{(L)} = \mathbf{1}$ and $\boldsymbol{P}^{(L-1)} = \boldsymbol{F}^{(L-1)}$. In this case, $\tilde{q}^{\mathrm{GAW}}$ fails to recover every spatial position in $[\![\boldsymbol{F}^{(L-1)}]\!]$, so the final average sum does not include all elements. This can be remedied artificially by enlarging $\boldsymbol{P}^{(L-1)}$ beyond $\boldsymbol{F}^{(L-1)}$ in each dimension, but at the cost of a worse constant in Theorem 2.1.

This is illustrated by the following example. Take $D = 1$, $F^{(L-1)} = P^{(L-1)} = 3$, $s = h = 1$, $\pi = 0$, so $F^{(L)} = 1$ and $q = 1$. The two conventions give:

p	1	2	3
$\tilde{q}^{\mathrm{GAW}}(p) = 1 - (p-2)$	2	1	0
$\tilde{q}(p) = 1 + (p-1)$	1	2	3

The patch function $\tilde{q}^{\mathrm{GAW}}$ lands at index $0 \notin [\![F^{(L-1)}]\!]$ and never reaches index 3, whereas (16) traverses all elements in $[\![F^{(L-1)}]\!]$.

1.9.5 Patch counter

We now introduce a quantity that counts the number of overlapping patches in layer ℓ. This quantity will be used in the proof of Theorem 2.1.

For $\boldsymbol{r} \in [\![\boldsymbol{F}^{(\ell)}]\!]$, let

$$N^{(\ell)}(\boldsymbol{r}) \coloneqq \left| \left\{ (\boldsymbol{q}, \boldsymbol{p}) \in [\![\boldsymbol{F}^{(\ell+1)}]\!] \times [\![\boldsymbol{P}^{(\ell)}]\!] : \boldsymbol{r} = \tilde{\boldsymbol{q}}(\boldsymbol{p}) \right\} \right| \tag{17}$$

be the number of patches that overlap at location $\boldsymbol{r}$. Denote the maximum for layer ℓ as

$$C_{\mathrm{patch}}^{(\ell)} \coloneqq \max_{\boldsymbol{r} \in [\![\boldsymbol{F}^{(\ell)}]\!]} N^{(\ell)}(\boldsymbol{r}). \tag{18}$$

Example 1.2. Consider a one-dimensional convolution from layer ℓ to layer $\ell+1$ with $F^{(\ell)} = 6$, with $F^{(\ell+1)} = 4$, patch size $P^{(\ell)} = 3$, and parameters $s^{(\ell)} = h^{(\ell)} = 1$, $\pi^{(\ell)} = 0$. The patch function simplifies to $\tilde{q}(p) = q + p - 1$, and the four patches in layer ℓ are

q	patch
1	$\{1, 2, 3\}$
2	$\{2, 3, 4\}$
3	$\{3, 4, 5\}$
4	$\{4, 5, 6\}$

Counting how many patches contain each input position $r \in [\![F^{(\ell)}]\!] = \{1, \dots, 6\}$ gives

$$\big(N^{(\ell)}(r)\big)_{r=1}^{6} = (1, 2, 3, 3, 2, 1),$$

so $C^{(\ell)}_{\text{patch}} = 3$.

1.10 Neural network Gaussian process (NNGP)

We now define the Gaussian process approximating a wide random CNN, as was done by Garriga-Alonso and Wilk (2021). To obtain the Gaussian process, we iterate layer-by-layer and for each layer $\ell \in \{1, \dots, L-1\}$, we take the number of channels $C^{(\ell)}$ to infinity, essentially applying the central limit theorem at each layer (channels $C^{(0)}$ and $C^{(L)}$ are kept fixed). The process $G^{(\ell)}$, which takes values on $\mathbb{R}^{\boldsymbol{F}^{(\ell)}}$, approximates a single channel of $Z^{(\ell)}$, while $\mathcal{G}^{(\ell)}$, which takes values on $\mathbb{R}^{C^{(\ell)} \times \boldsymbol{F}^{(\ell)}}$, denotes a process approximating the full output of $Z^{(\ell)}$.

Alternative approaches for deriving the kernel include taking the limit of all channel widths simultaneously, as done by Matthews et al. (2018), or using the tensor framework of Yang (2019). We chose the recursive approach because it most closely mirrors the structure of the proof of the main theorem. Regardless of the method used, all approaches yield the same Gaussian process.

As before, for each layer ℓ, we have weights $W^{(\ell)} \in \mathbb{R}^{C^{(\ell+1)} \times C^{(\ell)} \times \boldsymbol{P}^{(\ell)}}$ and biases $b^{(\ell)} \in \mathbb{R}^{C^{(\ell+1)}}$, which we assume now to follow the prior distribution

$$W^{(\ell)} \sim \mathcal{N}\left(0, \frac{1}{C^{(\ell)}} \operatorname{Id}_{C^{(\ell+1)}} \otimes \operatorname{Id}_{C^{(\ell)}} \otimes \Sigma^{(\ell)}\right) \quad \text{and} \quad b^{(\ell)} \sim \mathcal{N}\left(0, c_b^{(\ell)} \operatorname{Id}_{C^{(\ell+1)}}\right), \tag{19}$$

where $\Sigma^{(\ell)} \in \mathbb{R}^{\boldsymbol{P}^{(\ell)} \times \boldsymbol{P}^{(\ell)}}$ is a positive semi-definite covariance matrix and $c_b^{(\ell)} > 0$. Specifically, for any two patch positsions $\boldsymbol{p}, \boldsymbol{p}' \in [\![\boldsymbol{P}^{(\ell)}]\!]$,

$$\mathbb{E}\left[W^{(\ell)}[i,j,\boldsymbol{p}]\, W^{(\ell)}[i',j',\boldsymbol{p}']\right] = \frac{\Sigma^{(\ell)}[\boldsymbol{p},\boldsymbol{p}']}{C^{(\ell)}} \mathbf{1}_{\{i=i'\}} \mathbf{1}_{\{j=j'\}}. \tag{20}$$

The weights are thus independent across output channels i and input channels j, but may be correlated across patch positions $\boldsymbol{p}$ within the same filter. All weights and biases are mutually independent across layers.

From now on, we consider $Z^{(\ell)}(\mathcal{X})$ as a random variable, where the randomness arises from the weights $(W^{(\ell)})_{\ell=0}^{L-1}$ and biases $(b^{(\ell)})_{\ell=0}^{L-1}$ drawn according to (19). Crucially, the *same* weights and biases are used to evaluate $Z^{(\ell)}$ at every input in $\mathcal{X}$, so the outputs $\left(Z^{(\ell)}(X_\alpha)\right)_{\alpha \in [\![k]\!]}$ are dependent random variables, as opposed to being an i.i.d. family indexed by α.

Garriga-Alonso and Wilk (2021) highlights three important cases for $\Sigma^{(\ell)}$.

1. **Independent case**: $\Sigma^{(\ell)} = c_w^{(\ell)} \operatorname{Id}_{\boldsymbol{P}^{(\ell)}}$ for $c_w^{(\ell)} > 0$.
2. **Mean-pooling case**: $\Sigma^{(\ell)} = c^{(\ell)} \mathbf{1}_{\boldsymbol{P}^{(\ell)}} \mathbf{1}_{\boldsymbol{P}^{(\ell)}}^{\intercal}$, where typically $c^{(\ell)} = 1/\left|\boldsymbol{F}^{(\ell+1)}\right|^2$.
3. **Spatially correlated case**: $\Sigma^{(\ell)}[\boldsymbol{p},\boldsymbol{p}'] = \exp(-\left\|\boldsymbol{p}-\boldsymbol{p}'\right\|/l)$ for $l \in (0,\infty)$.

In Case 1, the resulting covariance kernel sums components across patches independently. In Case 2, the weights are fully correlated across all spatial positions, which forces them to be identical and is equivalent to computing the mean response over all spatial locations (also called global average pooling). This shows that mean pooling, which is ordinarily implemented in a CNN as a separate layer, can instead be realised solely through the weight covariance structure. Case 3 interpolates between the two extremes by letting the distance between patch positions control the strength of correlation. By taking $l \to 0$, we recover Case 1, and by taking $l \to \infty$, we recover Case 2.

1.10.1 Derivation of the kernel

To determine the Gaussian process, we need to know its mean and covariance function. Since in (19) the parameters are centred, the mean of $Z^{(\ell)}(X)$ is zero at every layer. Therefore, the Gaussian process is completely determined by its kernel alone.

Furthermore, observe that since the weights $W^{(\ell-1)}[i,:,:]$ are independent across channels i and the activations $\phi(Z^{(\ell-1)}(X)[j,:])$ in (15) do not depend on i, means that $Z^{(\ell)}(X)[1,\boldsymbol{q}],\ldots,Z^{(\ell)}(X)[C^{(\ell)},\boldsymbol{q}]$ are i.i.d for every fixed $\boldsymbol{q}$ and X. This fact will be used when deriving the NNGP kernel and is assumed in Lemma 1.3. Note that Lemma 1.3 establishes the inductive case while the base case $K^{(1)}$ will be provided in section 1.10.2.

Lemma 1.3 (NNGP kernel recursion)**.** Assume the correlated weight prior (19). Suppose that, for a fixed layer $\ell \geq 1$ and for all inputs $X \in \mathbb{R}^{C^{(0)}\times \boldsymbol{F}^{(0)}}$, the channels $\left(Z^{(\ell)}(X)[i,:]\right)_{i\in[\![C^{(\ell)}]\!]}$ are i.i.d. copies of a centered Gaussian process $G^{(\ell)}(X)$ with the covariance kernel $K^{(\ell)}$. Then for layer $\ell+1$ it holds that

$$\mathbb{E}\left[Z^{(\ell+1)}(X_\alpha)[i,\boldsymbol{q}]Z^{(\ell+1)}(X_{\alpha'})[i',\boldsymbol{q}']\right] = \mathbf{1}_{\{i=i'\}}K^{(\ell+1)}\left[(\boldsymbol{q},X_\alpha),(\boldsymbol{q}',X_{\alpha'})\right], \tag{21}$$

where

$$K^{(\ell+1)}\left[(\boldsymbol{q},X_\alpha),(\boldsymbol{q}',X_{\alpha'})\right] \coloneqq c_b^{(\ell)} + \sum_{\boldsymbol{p},\boldsymbol{p}'=\mathbf{1}}^{\boldsymbol{P}^{(\ell)}} \Sigma^{(\ell)}[\boldsymbol{p},\boldsymbol{p}']V^{(\ell)}\left[(\tilde{\boldsymbol{q}}(\boldsymbol{p}),X_\alpha),(\tilde{\boldsymbol{q}}'(\boldsymbol{p}'),X_{\alpha'})\right], \tag{22}$$

and

$$V^{(\ell)}\left[(\boldsymbol{r},X_\alpha),(\boldsymbol{r}',X_{\alpha'})\right] \coloneqq \mathbb{E}\left[\phi\left(G^{(\ell)}(X_\alpha)[\boldsymbol{r}]\right)\phi\left(G^{(\ell)}(X_{\alpha'})[\boldsymbol{r}']\right)\right]. \tag{23}$$

Moreover, $K^{(\ell+1)}$ is symmetric and positive semi-definite.

Proof. Fix $i,i' \in [\![C^{(\ell+1)}]\!]$, $\boldsymbol{q},\boldsymbol{q}' \in [\![\boldsymbol{F}^{(\ell+1)}]\!]$, and two inputs $X_\alpha, X_{\alpha'}$. Since the weights, biases, and previous-layer outputs are mutually independent, we have

$$\begin{aligned}\mathbb{E}\left[Z^{(\ell+1)}(X_\alpha)[i,\boldsymbol{q}]Z^{(\ell+1)}(X_{\alpha'})[i',\boldsymbol{q}']\right] &= \mathbb{E}\left[b^{(\ell)}[i]b^{(\ell)}[i']\right] \\ &\quad + \mathbb{E}\left[Z_0^{(\ell+1)}(X_\alpha)[i,\boldsymbol{q}]Z_0^{(\ell+1)}(X_{\alpha'})[i',\boldsymbol{q}']\right].\end{aligned}$$

Based on the biases' prior (19), we have $\mathbb{E}\left[b^{(\ell)}[i]b^{(\ell)}[i']\right] = \mathbf{1}_{\{i=i'\}}c_b^{(\ell)}$. To get the covariance of $Z_0^{(\ell+1)}$, we apply the result (24) from Lemma 2.1. For this

set $A[j, \boldsymbol{q}, \alpha] = \phi(Z^{(\ell)}(X_\alpha)[j, \boldsymbol{q}])$ and condition $Z_0^{(\ell+1)}$ on $Z^{(\ell)}$. Then

$$\begin{aligned}
&\mathbb{E}\left[Z_0^{(\ell+1)}(X_\alpha)[i, \boldsymbol{q}] Z_0^{(\ell+1)}(X_{\alpha'})[i', \boldsymbol{q}']\right] \\
&\quad = \mathbb{E}\left[\mathbb{E}\left[Z_0^{(\ell+1)}(X_\alpha)[i, \boldsymbol{q}] Z_0^{(\ell+1)}(X_{\alpha'})[i', \boldsymbol{q}'] \middle| Z^{(\ell)}\right]\right] \\
&\quad = \mathbf{1}_{\{i=i'\}} \frac{1}{C^{(\ell)}} \sum_{j=1}^{C^{(\ell)}} \sum_{\boldsymbol{p}, \boldsymbol{p}'=\mathbf{1}}^{\boldsymbol{P}^{(\ell)}} \Bigg(\Sigma^{(\ell)}[\boldsymbol{p}, \boldsymbol{p}'] \\
&\qquad\qquad \mathbb{E}\left[\phi\left(Z^{(\ell)}(X_\alpha)[j, \tilde{\boldsymbol{q}}(\boldsymbol{p})]\right) \phi\left(Z^{(\ell)}(X_{\alpha'})[j, \tilde{\boldsymbol{q}}'(\boldsymbol{p}')]\right) \middle| Z^{(\ell)}\right] \Bigg).
\end{aligned}$$

By the assumptions of the lemma, for each j, the random variable

$$Z^{(\ell)}(X_\alpha)[j, :] Z^{(\ell)}(X_{\alpha'})[j, :]$$

has the same joint distribution as

$$G^{(\ell)}(X_\alpha) G^{(\ell)}(X_{\alpha'}).$$

Thus every expectation in the sum is of the form $V^{(\ell)}\left[(\tilde{\boldsymbol{q}}(\boldsymbol{p}), X_\alpha), (\tilde{\boldsymbol{q}}'(\boldsymbol{p}'), X_{\alpha'})\right]$ and doesn't depend on j, whence

$$\begin{aligned}
&\mathbb{E}\left[Z_0^{(\ell+1)}(X_\alpha)[i, \boldsymbol{q}] Z_0^{(\ell+1)}(X_{\alpha'})[i', \boldsymbol{q}']\right] \\
&\qquad = \mathbf{1}_{\{i=i'\}} \sum_{\boldsymbol{p}, \boldsymbol{p}'=\mathbf{1}}^{\boldsymbol{P}^{(\ell)}} \Sigma^{(\ell)}[\boldsymbol{p}, \boldsymbol{p}'] V^{(\ell)}\left[(\tilde{\boldsymbol{q}}(\boldsymbol{p}), X_\alpha), (\tilde{\boldsymbol{q}}'(\boldsymbol{p}'), X_{\alpha'})\right].
\end{aligned}$$

Combining both results gives us (21) and (22):

$$\begin{aligned}
&\mathbb{E}\left[Z^{(\ell+1)}(X_\alpha)[i, \boldsymbol{q}] Z^{(\ell+1)}(X_{\alpha'})[i', \boldsymbol{q}']\right] \\
&= \mathbf{1}_{\{i=i'\}} \left(c_b^{(\ell)} + \sum_{\boldsymbol{p}, \boldsymbol{p}'=\mathbf{1}}^{\boldsymbol{P}^{(\ell)}} \Sigma^{(\ell)}[\boldsymbol{p}, \boldsymbol{p}'] V^{(\ell)}\left[(\tilde{\boldsymbol{q}}(\boldsymbol{p}), X_\alpha), (\tilde{\boldsymbol{q}}'(\boldsymbol{p}'), X_{\alpha'})\right] \right) \\
&= \mathbf{1}_{\{i=i'\}} K^{(\ell+1)}\left[(\boldsymbol{q}, X_\alpha), (\boldsymbol{q}', X_{\alpha'})\right].
\end{aligned}$$

It remains to show that $K^{(\ell+1)}$ is symmetric and positive semidefinite. Symmetry follows immediately from the fact that $V^{(\ell)}$ and $\Sigma^{(\ell)}$ are symmetric.

For positive semidefiniteness, fix $\boldsymbol{q}_\alpha \in [\![\boldsymbol{F}^{(\ell+1)}]\!]$, inputs X_α and $a_\alpha \in \mathbb{R}$ for $\alpha \in [\![n]\!]$. Since $\Sigma^{(\ell)}$ is positive semidefinite, we can write it as $\Sigma^{(\ell)} = BB^\mathsf{T}$

for $B \in \mathbb{R}^{\boldsymbol{P}^{(\ell)} \times d}$ and $d \in \mathbb{N}$. Then

$$\sum_{\alpha,\beta=1}^{n} a_\alpha a_\beta K^{(\ell+1)}[(\boldsymbol{q}_\alpha, X_\alpha), (\boldsymbol{q}_\beta, X_\beta)]$$

$$= c_b^{(\ell)} \sum_{\alpha,\beta=1}^{n} a_\alpha a_\beta + \sum_{\boldsymbol{p},\boldsymbol{p}'=\boldsymbol{1}}^{\boldsymbol{P}^{(\ell)}} \sum_{\alpha,\beta=1}^{n} (BB^\mathsf{T})[\boldsymbol{p},\boldsymbol{p}'] a_\alpha a_\beta V^{(\ell)}\left[(\tilde{\boldsymbol{q}}_\alpha(\boldsymbol{p}), X_\alpha), (\tilde{\boldsymbol{q}}_\beta(\boldsymbol{p}'), X_\beta)\right].$$

Since $c_b^{(\ell)} > 0$, the first term is non-negative. For the second term, we expand

$$\sum_{k=1}^{d} \sum_{\alpha,\beta=1}^{n} \sum_{\boldsymbol{p},\boldsymbol{p}'=\boldsymbol{1}}^{\boldsymbol{P}^{(\ell)}} B[\boldsymbol{p},k] B[\boldsymbol{p}',k] a_\alpha a_\beta V^{(\ell)}\left[(\tilde{\boldsymbol{q}}_\alpha(\boldsymbol{p}), X_\alpha), (\tilde{\boldsymbol{q}}_\beta(\boldsymbol{p}'), X_\beta)\right]$$

$$= \sum_{k=1}^{d} \sum_{\alpha,\beta=1}^{n} \sum_{\boldsymbol{p},\boldsymbol{p}'=\boldsymbol{1}}^{\boldsymbol{P}^{(\ell)}} \tilde{a}_{\alpha,\boldsymbol{p},k} \tilde{a}_{\beta,\boldsymbol{p}',k} V^{(\ell)}\left[(\tilde{\boldsymbol{q}}_\alpha(\boldsymbol{p}), X_\alpha), (\tilde{\boldsymbol{q}}_\beta(\boldsymbol{p}'), X_\beta)\right],$$

where $\tilde{a}_{\alpha,\boldsymbol{p},k} \coloneqq a_\alpha B[\boldsymbol{p},k] \in \mathbb{R}$. Because $V^{(\ell)}$ is a covariance kernel and hence positive semidefinite, then for every pair $(\boldsymbol{p},\boldsymbol{p}')$, the function $((\boldsymbol{q},X),(\boldsymbol{q}',X')) \mapsto V^{(\ell)}[(\tilde{\boldsymbol{q}}(\boldsymbol{p}),X),(\tilde{\boldsymbol{q}}'(\boldsymbol{p}'),X')]$ is positive semidefinite. Consequently, the second term is also non-negative. Combining the two expressions, we conclude that $K^{(\ell+1)}$ is positive semidefinite. □

1.10.2 Definition of the NNGP

Using Lemma 1.3, we can now define the NNGP recursively. For the first layer, $Z^{(1)}$ is Gaussian for any number of channels $C^{(1)}$ (we will see this in the proof of Theorem 2.1). Computing its covariance for a single channel, we have

$$K^{(1)}\left[(\boldsymbol{q},X),(\boldsymbol{q}',X')\right] = c_b^{(0)} + \frac{1}{C^{(0)}} \sum_{j=1}^{C^{(0)}} \sum_{\boldsymbol{p},\boldsymbol{p}'=\boldsymbol{1}}^{\boldsymbol{P}^{(0)}} \Sigma^{(0)}[\boldsymbol{p},\boldsymbol{p}'] X[j,\tilde{\boldsymbol{q}}(\boldsymbol{p})] X'[j,\tilde{\boldsymbol{q}}'(\boldsymbol{p}')].$$

We let $G^{(1)} \sim \mathcal{N}(K^{(1)})$. For $\ell = 2,\ldots,L$, having already defined $G^{(\ell-1)}$, to obtain $G^{(\ell)}$, we take $C^{(\ell-1)} \to \infty$, turning $Z^{(\ell)}$ into a Gaussian (for a formal justification we refer to Garriga-Alonso and Wilk (2021)). Based on Lemma 1.3, the covariance for a single channel is

$$K^{(\ell)}[(\boldsymbol{q},X),(\boldsymbol{q}',X')] = c_b^{(\ell-1)} + \sum_{\boldsymbol{p},\boldsymbol{p}'=\boldsymbol{1}}^{\boldsymbol{P}^{(\ell-1)}} \Sigma^{(\ell-1)}[\boldsymbol{p},\boldsymbol{p}'] V^{(\ell-1)}[(\tilde{\boldsymbol{q}}(\boldsymbol{p}),X),(\tilde{\boldsymbol{q}}'(\boldsymbol{p}'),X')]$$

and we define $G^{(\ell)} \sim \mathcal{N}(K^{(\ell)})$.

We additionally define the bias-free kernel

$$K_0^{(\ell)}\left[(\boldsymbol{q}, X), (\boldsymbol{q}', X')\right] \coloneqq \sum_{\boldsymbol{p},\boldsymbol{p}'=\boldsymbol{1}}^{\boldsymbol{P}^{(\ell-1)}} \Sigma^{(\ell-1)}[\boldsymbol{p}, \boldsymbol{p}']\, V^{(\ell-1)}\left[(\tilde{\boldsymbol{q}}(\boldsymbol{p}), X),\, (\tilde{\boldsymbol{q}}'(\boldsymbol{p}'), X')\right],$$

so that $K^{(\ell)} = c_b^{(\ell-1)} + K_0^{(\ell)}$. This will be used in the proof of Theorem 2.1.

1.10.3 Multi-channel NNGP

The Gaussian process approximating the ℓ-th layer of the network's output is $\mathcal{G}^{(\ell)} = (\mathcal{G}^{(\ell)}(X))_{X \in \mathbb{R}^{C^{(0)} \times \boldsymbol{F}^{(0)}}}$ with values in $\mathbb{R}^{C^{(\ell)} \times \boldsymbol{F}^{(\ell)}}$, defined as $C^{(\ell)}$ independent copies of $G^{(\ell)}$, with $\mathcal{G}^{(\ell)}(\cdot)[i, \boldsymbol{q}]$ denoting the value of the i-th copy at spatial position $\boldsymbol{q}$, for $i \in [\![C^{(\ell)}]\!]$ and $\boldsymbol{q} \in [\![\boldsymbol{F}^{(\ell)}]\!]$. That is

$$\mathbb{E}\left[\mathcal{G}^{(\ell)}(X)[i, \boldsymbol{q}]\mathcal{G}^{(\ell)}(X')[i', \boldsymbol{q}']\right] = \mathbf{1}_{\{i=i'\}} K^{(\ell)}\left[(\boldsymbol{q}, X), (\boldsymbol{q}', X')\right].$$

For k inputs $\mathcal{X} = \{X_\alpha\}_{\alpha \in [\![k]\!]} \subset \mathbb{R}^{C^{(0)} \times \boldsymbol{F}^{(0)}}$, write $\mathcal{G}^{(\ell)}(\mathcal{X}) = \sum_{\alpha=1}^{k} \mathcal{G}^{(\ell)}(X_\alpha) \otimes e_\alpha$ which takes values in $\mathbb{R}^{C^{(\ell)} \times \boldsymbol{F}^{(\ell)} \times k}$. If we define the restriction of kernel $K^{(\ell)}$ to the dataset $\mathcal{X}$ as

$$K^{(\ell)}(\mathcal{X}) = \left(K^{(\ell)}[(\boldsymbol{q}, X), (\boldsymbol{q}', X')]\right)_{\substack{\boldsymbol{q},\boldsymbol{q}' \in [\![\boldsymbol{F}^{(\ell)}]\!] \\ X, X' \in \mathcal{X}}} \in \mathbb{R}^{(\boldsymbol{F}^{(\ell)} \times k) \times (\boldsymbol{F}^{(\ell)} \times k)},$$

the covariance of $\mathcal{G}^{(\ell)}(\mathcal{X})$ is

$$\Sigma\left(\mathcal{G}^{(\ell)}(\mathcal{X})\right) = \mathrm{Id}_{C^{(\ell)}} \otimes K^{(\ell)}(\mathcal{X}).$$

1.10.4 Non-nullity of $K_0^{(\ell+1)}$

In the proof of Theorem 2.1, we require that $K_0^{(\ell+1)}(\mathcal{X})$ is not null for $\ell \in [\![L-1]\!]$. We'll show now that this at least holds for the two out of three important cases of $\Sigma^{(\ell)}$, whenever $K^{(\ell)}(\mathcal{X})$ is not null.

Lemma 1.4. Let $\ell \in [\![L-1]\!]$. Suppose that ϕ is not identically zero and $\Sigma^{(\ell)}$ falls under either Case 1 or Case 3. Assume furthermore that $K^{(\ell)}(\mathcal{X})$ is not null with $K^{(\ell)}[(\boldsymbol{q}^*, X), (\boldsymbol{q}^*, X)] > 0$ for some $\boldsymbol{q}^* \in [\![\boldsymbol{F}^{(\ell)}]\!]$ and $X \in \mathcal{X}$. If $N^{(\ell)}(\boldsymbol{q}^*) \geq 1$, then $K_0^{(\ell+1)}(\mathcal{X})$ is not null.

Remark 1.7. It is quite reasonable to assume that $N^{(\ell)}(\boldsymbol{q}^*) \geq 1$, that is $\boldsymbol{q}^*$ lies in the range of the patch function for some output position. This, for example, holds whenever the convolution parameters (patch size, stride, dilation) are such that the convolution covers the full input, which is the typical setting in practice. The assumption that $K^{(\ell)}(\mathcal{X})$ is not null reduces to the case that $K^{(1)}(\mathcal{X})$ is not null, which is valid to assume, as this depends on the input $\mathcal{X}$. From $K_0^{(\ell+1)}(\mathcal{X})$ is not null follows that $K^{(\ell+1)}(\mathcal{X})$ is not null, since $c_b^{(\ell)} > 0$.

Proof. Common preliminary. Fix $X \in \mathcal{X}$ and $\boldsymbol{q}^* \in [\![\boldsymbol{F}^{(\ell)}]\!]$ such that

$$K^{(\ell)}[(\boldsymbol{q}^*, X), (\boldsymbol{q}^*, X)] > 0.$$

Then $G^{(\ell)}(X)[\boldsymbol{q}^*]$ is not identically null. Since ϕ is not identically null, it follows that $\phi\left(G^{(\ell)}(X)[\boldsymbol{q}^*]\right)$ is not almost surely zero either. Whence

$$V^{(\ell)}\left[(\boldsymbol{q}^*, X), (\boldsymbol{q}^*, X)\right] > 0.$$

Furthermore, fix $\boldsymbol{q}' \in [\![\boldsymbol{F}^{(\ell+1)}]\!]$ and $\boldsymbol{p}^* \in \boldsymbol{P}^{(\ell)}$ such that $\tilde{\boldsymbol{q}}'(\boldsymbol{p}^*) = \boldsymbol{q}^*$, which exists by the assummption $N^{(\ell)}(\boldsymbol{q}^*) \geq 1$.

Case 1: $\Sigma^{(\ell)} = c\,\mathrm{Id}_{\boldsymbol{P}^{(\ell)}}$. For the output position $\boldsymbol{q}'$ chosen above,

$$K_0^{(\ell+1)}\left[(\boldsymbol{q}', X), (\boldsymbol{q}', X)\right] = c \sum_{\boldsymbol{p}=\boldsymbol{1}}^{\boldsymbol{P}^{(\ell)}} V^{(\ell)}\left[(\tilde{\boldsymbol{q}}'(\boldsymbol{p}), X), (\tilde{\boldsymbol{q}}'(\boldsymbol{p}), X)\right].$$

Each summand is nonnegative and the term at $\boldsymbol{p} = \boldsymbol{p}^*$ equals

$$c \cdot V^{(\ell)}[(\boldsymbol{q}^*, X), (\boldsymbol{q}^*, X)] > 0.$$

Hence $K_0^{(\ell+1)}(\mathcal{X})$ is not null.

Case 3: $\Sigma^{(\ell)}[\boldsymbol{p}, \boldsymbol{p}'] = \exp(-\left\|\boldsymbol{p} - \boldsymbol{p}'\right\| / l)$. Since $\Sigma^{(\ell)}$ is positive definite (Rasmussen and Williams, 2006), its smallest eigenvalue satisfies $\lambda(\Sigma^{(\ell)}) > 0$. Let $v[\boldsymbol{p}] \coloneqq \phi(G^{(\ell)}(X)[\tilde{\boldsymbol{q}}'(\boldsymbol{p})])$ as shorthand. Then by (4), we have

$$\begin{aligned} K_0^{(\ell+1)}\left[(\boldsymbol{q}', X), (\boldsymbol{q}', X)\right] &= \mathbb{E}\left[v^\mathsf{T} \Sigma^{(\ell)} v\right] \geq \mathbb{E}\left[\lambda\left(\Sigma^{(\ell)}\right) v^\mathsf{T} v\right] \\ &= \lambda\left(\Sigma^{(\ell)}\right) \sum_{\boldsymbol{p}=\boldsymbol{1}}^{\boldsymbol{P}^{(\ell)}} V^{(\ell)}\left[(\tilde{\boldsymbol{q}}'(\boldsymbol{p}), X), (\tilde{\boldsymbol{q}}'(\boldsymbol{p}), X)\right]. \end{aligned}$$

The right-hand side is strictly positive, as shown in Case 1 with $c = \lambda(\Sigma^{(\ell)}) > 0$. Hence $K_0^{(\ell+1)}(\mathcal{X})$ is not null. □

2 Proof of the main theorem

Recall that we previously defined the deep convolutional network and the prior distributions for the parameters

$$W^{(\ell)} \sim \mathcal{N}\left(0, \frac{1}{C^{(\ell)}} \mathrm{Id}_{C^{(\ell+1)}} \otimes \mathrm{Id}_{C^{(\ell)}} \otimes \Sigma^{(\ell)}\right) \quad \text{and} \quad b^{(\ell)} \sim \mathcal{N}\left(0, c_b^{(\ell)} \mathrm{Id}_{C^{(\ell+1)}}\right). \tag{19}$$

The proof of the main theorem follows the same structure as in Basteri and Trevisan (2024). Lemmas 2.1 and 2.2 generalise the corresponding Lemmas 3.3 and 3.4 from Basteri and Trevisan (2024), respectively. We now state these lemmas and postpone their proofs to the next section.

Lemma 2.1. Fix a layer index ℓ. Let $W^{(\ell)} \in \mathbb{R}^{C^{(\ell+1)} \times C^{(\ell)} \times \boldsymbol{P}^{(\ell)}}$ be the convolution weight matrix defined in (19). Let $A \in \mathbb{R}^{C^{(\ell)} \times \boldsymbol{F}^{(\ell)} \times k}$ and B be a random variable with values in $\mathbb{R}^{C^{(\ell+1)} \times \boldsymbol{F}^{(\ell+1)} \times k}$ and satisfying

$$B[i, \boldsymbol{q}, \alpha] = \sum_{j=1}^{C^{(\ell)}} \sum_{\boldsymbol{p}=\boldsymbol{1}}^{\boldsymbol{P}^{(\ell)}} W^{(\ell)}[i, j, \boldsymbol{p}] A[j, \tilde{\boldsymbol{q}}(\boldsymbol{p}), \alpha].$$

Then B is a centred Gaussian random variable with the covariance

$$\Sigma(B) = \mathrm{Id}_{C^{(\ell+1)}} \otimes \bar{A}, \tag{24}$$

where $\bar{A} \in \mathbb{R}^{(\boldsymbol{F}^{(\ell+1)} \times k) \times (\boldsymbol{F}^{(\ell+1)} \times k)}$ is given by

$$\bar{A}[(\boldsymbol{q}, \alpha), (\boldsymbol{q}', \alpha')] \coloneqq \frac{1}{C^{(\ell)}} \sum_{j=1}^{C^{(\ell)}} \sum_{\boldsymbol{p}, \boldsymbol{p}'=\boldsymbol{1}}^{\boldsymbol{P}^{(\ell)}} \Sigma^{(\ell)}[\boldsymbol{p}, \boldsymbol{p}'] A[j, \tilde{\boldsymbol{q}}(\boldsymbol{p}), \alpha] A[j, \tilde{\boldsymbol{q}}'(\boldsymbol{p}'), \alpha'].$$

Moreover,

$$\mathbb{E}\left[\|B\|^2\right] \leq \frac{\|\Sigma^{(\ell)}\|_{\mathrm{op}} C_{\mathrm{patch}}^{(\ell)} C^{(\ell+1)}}{C^{(\ell)}} \|A\|^2. \tag{25}$$

Remark 2.1. This lemma extends to the discrete convolution setting, the fact that an affine transformation of a Gaussian random vector is again Gaussian: if $X \sim \mathcal{N}(\mu, \Sigma)$ and $Y = a + BX$, then $Y \sim \mathcal{N}(a + B\mu, B\Sigma B^{\mathsf{T}})$.

Lemma 2.2. Let S, T and I be finite sets and let $X = (X[i])_{i=1}^n$ be independent and identically distributed random variables with values in $\mathbb{R}^S$ and finite fourth moment. Furthermore, let $M = \mathbb{E}[X[1] \otimes X[1]]$ and define the following random variable, taking values in $\mathbb{R}^{S\times S}$,

$$M_n = \frac{1}{n}\sum_{i=1}^{n} X[i] \otimes X[i].$$

Let $\Sigma \in \mathbb{R}^{I\times I}$ be a symmetric and positive semidefinite matrix and let $L : \mathbb{R}^{S\times S} \to \mathbb{R}^{T\times T}$ be a linear operator that can be written as

$$L(N) \coloneqq \sum_{i,j\in I} \Sigma[i,j] B_i^{\mathsf{T}} N B_j, \tag{26}$$

where $N \in \mathbb{R}^{S\times S}$ and $B_i \in \{0,1\}^{S\times T}$ is a partial permutation matrix, that is, for every $s \in S$ we have $\sum_{t\in T} B_i[s,t] \in \{0,1\}$, and for every $t \in T$ we have $\sum_{s\in S} B_i[s,t] \in \{0,1\}$. Moreover, assume that $L(M)$ is nonzero, symmetric and positive semidefinite.

Then,

$$\mathbb{E}\left[\left\|\sqrt{L(M_n)} - \sqrt{L(M)}\right\|^2\right] \leq \frac{\|\Sigma\|_{1,1}^2 \cdot \mathbb{E}\left[\|X[1]\otimes X[1] - M\|^2\right]}{n\lambda^+(L(M))}$$

where $\lambda^+(L(M)) > 0$ denotes the smallest strictly positive eigenvalue of $L(M)$, which exists since $L(M) \neq 0$. The right-hand side is finite by the finite fourth moment assumption.

Remark 2.2. Here M and M_n represent the theoretical and empirical second moments, respectively. In the context of the theorem, this will be the second moment of the activations. The index set I takes the role of all possible patch positions and the matrix $B_{\boldsymbol{p}}$ is the matrix encoding of the patch function (16). Under these identifications, $L(M) = K_0^{(\ell+1)}(\mathcal{X})$ is the NNGP kernel of the next layer, and $L(M_n) = \hat{K}_0(\mathcal{G}^{(\ell)}(\mathcal{X}))$ is its empirical counterpart, formed by averaging over channels. The lemma therefore bounds the distance between the square roots of the empirical and theoretical kernels, with the channel width $C^{(\ell)}$ playing the role of the sample size n.

We now have all the ingredients necessary to state and prove the main result of this thesis.

Theorem 2.1. Consider a deep convolutional neural network $Z^{(L)}$ with randomly initialized parameters as in (19) and Lipschitz activation function ϕ that is applied element-wise. Then, for every collection of k inputs $\mathcal{X} = (X_\alpha)_{\alpha\in[\![k]\!]} \subset \mathbb{R}^{C^{(0)}\times \boldsymbol{F}^{(0)}}$, such that $K_0^{(\ell+1)}(\mathcal{X})$ is nonzero for all $\ell \in [\![L-1]\!]$, there exist finite constants $\left(C_{\text{ker}}^{(\ell)}, C_{\text{patch}}^{(\ell-1)}\right)_{\ell\in[\![L]\!]}$, such that for every layer $\ell \in [\![L]\!]$, the distribution of random variable $Z^{(\ell)}(\mathcal{X})$ is quantitatively close to the associated Gaussian process $\mathcal{G}^{(\ell)}(\mathcal{X})$ and satisfies the bound

$$\mathcal{W}_2\left(Z^{(\ell)}(\mathcal{X}), \mathcal{G}^{(\ell)}(\mathcal{X})\right) \leq \sqrt{C^{(\ell)}} \sum_{i=1}^{\ell-1} \frac{\left\|\Sigma^{(i)}\right\|_{1,1} C_{\text{ker}}^{(i+1)} \operatorname{Lip}(\phi)^{\ell-1-i} \prod_{j=i+1}^{\ell-1} \sqrt{C_{\text{patch}}^{(j)} \|\Sigma^{(j)}\|_{\text{op}}}}{\sqrt{C^{(i)}}}. \quad (27)$$

Each finite constant $C_{\text{ker}}^{(\ell)}$ depends only on the activation function ϕ, the input set $\mathcal{X}$ and the parameter variances $(\Sigma^{(i)}, c_b^{(i)})_{i\in[\![\ell-1]\!]}$, while each finite constant $C_{\text{patch}}^{(\ell)}$ depends only on the convolution parameters between layers ℓ and $\ell+1$.

Remark 2.3. Note that the constants don't depend on the channel widths. Therefore, by taking the hidden channel widths $C^{(1)}, \ldots, C^{(L-1)}$ to infinity while keeping the input and output widths $C^{(0)}$ and $C^{(L)}$ fixed, we recover the convergence of the CNN to its corresponding Gaussian process.

Furthermore, the theorem reveals how the network architecture influences the emergence of Gaussian behaviour. In particular, deeper networks require larger channel widths for the Gaussian behaviour to emerge. Additionally, larger overlaps between patches and stronger correlations between patches (making the matrix $\Sigma^{(i)}$ low-rank) also increase the required channel widths.

Corollary 2.1 (Recovery of the bound in Basteri and Trevisan (2024))**.** If in Theorem 2.1 we take $\boldsymbol{P}^{(\ell-1)} = \mathbf{1}_D$ and $\Sigma^{(\ell-1)}[\mathbf{1}_D, \mathbf{1}_D] = c_w^{(\ell-1)} > 0$ for $\ell \in [\![L]\!]$, and additionally assume that $\boldsymbol{F}^{(0)} = \ldots = \boldsymbol{F}^{(L)} = \mathbf{1}_D$, then the CNN simplifes to a fully connected neural network and we recover the bound from Basteri and Trevisan (2024):

$$\mathcal{W}_2\left(Z^{(\ell)}(\mathcal{X}), \mathcal{G}^{(\ell)}(\mathcal{X})\right) \leq \sqrt{C^{(\ell)}} \sum_{i=1}^{\ell-1} \frac{c_w^{(i)} C_{\text{ker}}^{(i+1)} \operatorname{Lip}(\phi)^{\ell-1-i} \prod_{j=i+1}^{\ell-1} \sqrt{c_w^{(j)}}}{\sqrt{C^{(i)}}}.$$

Remark 2.4. Note that the bound in Corollary 2.1 differs slightly from the one stated in Basteri and Trevisan (2024), equation (3.2), where the factor $c_w^{(\ell)}$ is not tracked through the application of Lemma 3.4.

Proof of the main theorem. The proof proceeds by induction on $\ell \in [\![L]\!]$.

Base case. We note that

$$\begin{aligned}
\Sigma\left(Z^{(1)}(\mathcal{X})\right) &= \Sigma\left(Z_0^{(1)}(\mathcal{X}) + b^{(0)} \otimes \mathbf{1}_{\boldsymbol{F}^{(1)}} \otimes \mathbf{1}_k\right) \\
&= \Sigma\left(Z_0^{(1)}(\mathcal{X})\right) + \Sigma\left(b^{(0)} \otimes \mathbf{1}_{\boldsymbol{F}^{(1)}} \otimes \mathbf{1}_k\right) && \text{(independence)} \\
&= \mathrm{Id}_{C^{(1)}} \otimes \bar{\mathcal{X}} + c_b^{(0)}\, \mathrm{Id}_{C^{(1)}} \otimes \mathbf{1}_{\boldsymbol{F}^{(1)}} \otimes \mathbf{1}_k \otimes \mathbf{1}_{\boldsymbol{F}^{(1)}} \otimes \mathbf{1}_k && \text{(lemma 2.1)} \\
&= \mathrm{Id}_{C^{(1)}} \otimes K^{(1)}(\mathcal{X}), && \text{(definition)}
\end{aligned}$$

which is exactly as the kernel of the first layer's Gaussian process, therefore $\mathcal{W}_2(Z^{(1)}(\mathcal{X}), \mathcal{G}^{(1)}(\mathcal{X})) = 0$. The lemma 2.1 was applied here with $A = \mathcal{X}$.

Induction step. Assume that (27) holds for some $\ell \in [\![L-1]\!]$. We will show that the following recursive bound holds

$$\begin{aligned}
&\mathcal{W}_2\left(Z^{(\ell+1)}(\mathcal{X}), \mathcal{G}^{(\ell+1)}(\mathcal{X})\right) \leq \\
&\sqrt{\frac{C^{(\ell+1)}}{C^{(\ell)}}}\left(\mathrm{Lip}(\phi)\sqrt{C_{\text{patch}}^{(\ell)}\|\Sigma^{(\ell)}\|_{\text{op}}}\,\mathcal{W}_2\left(Z^{(\ell)}(\mathcal{X}), \mathcal{G}^{(\ell)}(\mathcal{X})\right) + \left\|\Sigma^{(\ell)}\right\|_{1,1} C_{\text{ker}}^{(\ell+1)}\right).
\end{aligned} \tag{28}$$

Replacing $\mathcal{W}_2\left(Z^{(\ell)}(\mathcal{X}), \mathcal{G}^{(\ell)}(\mathcal{X})\right)$ with the inductive hypothesis gives us (27).

Fix a probability space on which $Z^{(\ell)}(\mathcal{X})$ and $\mathcal{G}^{(\ell)}(\mathcal{X})$ are jointly defined. WLOG, assume that the parameters $W^{(\ell)}$ and $b^{(\ell)}$ are defined on the same space and independent of both $Z^{(\ell)}(\mathcal{X})$ and $\mathcal{G}^{(\ell)}(\mathcal{X})$. If this is not the case, we may enlarge the probability space to include independent copies of the parameters.

Define the auxiliary random variables:

$$\begin{aligned}
h^{(\ell+1)}[i, \boldsymbol{q}, \alpha] &= \sum_{j=1}^{C^{(\ell)}} \sum_{\boldsymbol{p}=\mathbf{1}}^{\boldsymbol{P}^{(\ell)}} W^{(\ell)}[i, j, \boldsymbol{p}]\phi\left(\mathcal{G}^{(\ell)}(\mathcal{X})[j, \tilde{\boldsymbol{q}}(\boldsymbol{p}), \alpha]\right), \\
g^{(\ell+1)}[i, \boldsymbol{q}, \alpha] &= h^{(\ell+1)}[i, \boldsymbol{q}, \alpha] + b^{(\ell)}[i].
\end{aligned}$$

To account for biases, let $H^{(\ell+1)}$ be a centred Gaussian random variable that is defined on the same space, taking values in $\mathbb{R}^{C^{(\ell+1)} \times \boldsymbol{F}^{(\ell+1)} \times k}$ and with covariance

$$\mathrm{Id}_{C^{(\ell+1)}} \otimes K_0^{(\ell+1)}(\mathcal{X}) \in \mathbb{R}^{(C^{(\ell+1)} \times C^{(\ell+1)}) \times (\boldsymbol{F}^{(\ell+1)} \times \boldsymbol{F}^{(\ell+1)} \times k \times k)}.$$

Similarly, WLOG, assume that $H^{(\ell+1)}$ is independent of $b^{(\ell)}$, so that

$$H^{(\ell+1)} + b^{(\ell)} \otimes \mathbf{1}_{\boldsymbol{F}^{(\ell+1)}} \otimes \mathbf{1}_k \stackrel{d}{=} \mathcal{G}^{(\ell+1)}(\mathcal{X}).$$

By the triangle inequality (10), we now split

$$\begin{aligned}\mathcal{W}_2\left(Z^{(\ell+1)}(\mathcal{X}), \mathcal{G}^{(\ell+1)}(\mathcal{X})\right) \leq \\ \mathcal{W}_2\left(Z^{(\ell+1)}(\mathcal{X}), g^{(\ell+1)}(\mathcal{X})\right) + \mathcal{W}_2\left(g^{(\ell+1)}(\mathcal{X}), \mathcal{G}^{(\ell+1)}(\mathcal{X})\right)\end{aligned}$$

and bound the two terms separately.

First term. To begin with, note that

$$\begin{aligned}\mathcal{W}_2^2(Z^{(\ell+1)}(\mathcal{X}), g^{(\ell+1)}(\mathcal{X})) &= \mathcal{W}_2^2\big(Z_0^{(\ell+1)}(\mathcal{X}) + b^{(\ell)} \otimes \mathbf{1}_{\boldsymbol{F}^{(\ell+1)}} \otimes \mathbf{1}_k, \\ &\qquad h^{(\ell+1)}(\mathcal{X}) + b^{(\ell)} \otimes \mathbf{1}_{\boldsymbol{F}^{(\ell+1)}} \otimes \mathbf{1}_k\big) \\ &\leq \mathcal{W}_2^2(Z_0^{(\ell+1)}(\mathcal{X}), h^{(\ell+1)}(\mathcal{X})) \qquad \text{(independence)} \\ &\leq \mathbb{E}\left[\|Z_0^{(\ell+1)}(\mathcal{X}) - h^{(\ell+1)}(\mathcal{X})\|^2\right], \qquad \text{(property (9))}\end{aligned}$$

and the bias term plays no role in the bound.

Furthermore, we write

$$\begin{aligned}&Z_0^{(\ell+1)}(\mathcal{X})[i, \boldsymbol{q}, \alpha] - h^{(\ell+1)}(\mathcal{X})[i, \boldsymbol{q}, \alpha] \\ &= \sum_{j=1}^{C^{(\ell)}} \sum_{\boldsymbol{p}=\mathbf{1}}^{\boldsymbol{P}^{(\ell)}} W^{(\ell)}[i, j, \boldsymbol{p}] \left(\phi\left(Z^{(\ell)}(\mathcal{X})[j, \tilde{\boldsymbol{q}}(\boldsymbol{p}), \alpha]\right) - \phi\left(G^{(\ell)}(\mathcal{X})[j, \tilde{\boldsymbol{q}}(\boldsymbol{p}), \alpha]\right)\right).\end{aligned}$$

Conditioning on $Z^{(\ell)}(\mathcal{X})$ and $\mathcal{G}^{(\ell)}(\mathcal{X})$, and applying (25) from Lemma 2.1 with $A = \phi(Z^{(\ell)}) - \phi(\mathcal{G}^{(\ell)})$, we obtain

$$\begin{aligned}&\mathbb{E}\left[\|Z_0^{(\ell+1)}(\mathcal{X}) - h^{(\ell+1)}(\mathcal{X})\|^2 \middle| Z^{(\ell)}(\mathcal{X}), \mathcal{G}^{(\ell)}(\mathcal{X})\right] \\ &\qquad \leq \frac{\|\Sigma^{(\ell)}\|_{\mathrm{op}} C_{\mathrm{patch}}^{(\ell)} C^{(\ell+1)}}{C^{(\ell)}} \|\phi(Z^{(\ell)}(\mathcal{X})) - \phi(\mathcal{G}^{(\ell)}(\mathcal{X}))\|^2.\end{aligned}$$

Finally, we use the Lipschitz property

$$\|\phi(Z^{(\ell)}(\mathcal{X})) - \phi(\mathcal{G}^{(\ell)}(\mathcal{X}))\|^2 \leq \mathrm{Lip}(\phi)^2 \|Z^{(\ell)}(\mathcal{X}) - \mathcal{G}^{(\ell)}(\mathcal{X})\|^2.$$

Combining everything yields the bound on the first term

$$\begin{aligned}\mathcal{W}_2^2(Z^{(\ell+1)}(\mathcal{X}), g^{(\ell+1)}(\mathcal{X})) &\leq \mathbb{E}\left[\mathbb{E}\left[\|Z_0^{(\ell+1)}(\mathcal{X}) - h^{(\ell+1)}(\mathcal{X})\|^2 \middle| Z^{(\ell)}(\mathcal{X}), \mathcal{G}^{(\ell)}(\mathcal{X})\right]\right] \\ &\leq \frac{C_{\text{patch}}^{(\ell)} C^{(\ell+1)}}{C^{(\ell)}} \|\Sigma^{(\ell)}\|_{\text{op}} \operatorname{Lip}(\phi)^2 \,\mathbb{E}\left[\|Z^{(\ell)}(\mathcal{X}) - \mathcal{G}^{(\ell)}(\mathcal{X})\|^2\right].\end{aligned}$$

Second term. We start off the same, as we did for the first term

$$\begin{aligned}\mathcal{W}_2^2(g^{(\ell+1)}(\mathcal{X}), \mathcal{G}^{(\ell+1)}(\mathcal{X})) &= \mathcal{W}_2^2(h^{(\ell+1)}(\mathcal{X}) + b^{(\ell)} \otimes \mathbf{1}_{\boldsymbol{F}^{(\ell+1)}} \otimes \mathbf{1}_k, \\ &\qquad H^{(\ell+1)}(\mathcal{X}) + b^{(\ell)} \otimes \mathbf{1}_{\boldsymbol{F}^{(\ell+1)}} \otimes \mathbf{1}_k) \\ &\leq \mathcal{W}_2^2(h^{(\ell+1)}(\mathcal{X}), H^{(\ell+1)}(\mathcal{X})).\end{aligned}$$

Next, we condition $h^{(\ell+1)}(\mathcal{X})$ on $\mathcal{G}^{(\ell)}(\mathcal{X}) = z \in \mathbb{R}^{C^{(\ell)} \times \boldsymbol{F}^{(\ell)} \times k}$. By lemma 2.1, $h^{(\ell+1)}(\mathcal{X})$ has a centered Gaussian law with the covariance $\mathrm{Id}_{C^{(\ell+1)}} \otimes \hat{K}_0(z)$, where $\hat{K}_0(z) \in \mathbb{R}^{(\boldsymbol{F}^{(\ell+1)} \times k) \times (\boldsymbol{F}^{(\ell+1)} \times k)}$ is given by

$$\hat{K}_0(z)[(\boldsymbol{q}, \alpha), (\boldsymbol{q}', \alpha')] \coloneqq \frac{1}{C^{(\ell)}} \sum_{j=1}^{C^{(\ell)}} \sum_{\boldsymbol{p}, \boldsymbol{p}' = \mathbf{1}}^{\boldsymbol{P}^{(\ell)}} \Sigma^{(\ell)}[\boldsymbol{p}, \boldsymbol{p}'] \phi(z[j, \tilde{\boldsymbol{q}}(\boldsymbol{p}), \alpha]) \phi(z[j, \tilde{\boldsymbol{q}}'(\boldsymbol{p}'), \alpha']).$$

Applying (13), we obtain

$$\begin{aligned}\mathcal{W}_2^2(\mathbb{P}_{h^{(\ell+1)}(\mathcal{X}) | \mathcal{G}^{(\ell)}(\mathcal{X}) = z}, \mathbb{P}_{H^{(\ell+1)}(\mathcal{X})}) &\leq \left\| \sqrt{\mathrm{Id}_{C^{(\ell+1)}} \otimes \hat{K}_0(z)} - \sqrt{\mathrm{Id}_{C^{(\ell+1)}} \otimes K_0^{(\ell+1)}(\mathcal{X})} \right\|^2 \\ &= \left\| \mathrm{Id}_{C^{(\ell+1)}} \otimes \sqrt{\hat{K}_0(z)} - \mathrm{Id}_{C^{(\ell+1)}} \otimes \sqrt{K_0^{(\ell+1)}(\mathcal{X})} \right\|^2 \\ &= \left\| \mathrm{Id}_{C^{(\ell+1)}} \otimes \left(\sqrt{\hat{K}_0(z)} - \sqrt{K_0^{(\ell+1)}(\mathcal{X})} \right) \right\|^2 \\ &= C^{(\ell+1)} \left\| \sqrt{\hat{K}_0(z)} - \sqrt{K_0^{(\ell+1)}(\mathcal{X})} \right\|^2.\end{aligned}$$

Using the property (12) of Wasserstein distance, we now get

$$\mathcal{W}_2^2(h^{(\ell+1)}(\mathcal{X}), H^{(\ell+1)}(\mathcal{X})) \leq C^{(\ell+1)} \,\mathbb{E}\left[\left\| \sqrt{\hat{K}_0(\mathcal{G}^{(\ell)}(\mathcal{X}))} - \sqrt{K_0^{(\ell+1)}(\mathcal{X})} \right\|^2 \right].$$

We now apply Lemma 2.2. Set $X[i] = \phi(\mathcal{G}^{(\ell)}(\mathcal{X})[i])$, where $\mathcal{G}^{(\ell)}(\mathcal{X})[i] = \mathcal{G}^{(\ell)}(\mathcal{X})[i, :, :] \in \mathbb{R}^{\boldsymbol{F}^{(\ell)} \times k}$, and take

$$n = C^{(\ell)}, \quad S = [\![\boldsymbol{F}^{(\ell)}]\!] \times [\![k]\!], \quad T = [\![\boldsymbol{F}^{(\ell+1)}]\!] \times [\![k]\!], \quad I = [\![\boldsymbol{P}^{(\ell)}]\!].$$

For $N \in \mathbb{R}^{S\times S}$, define the linear operator as

$$L(N)[(\boldsymbol{q},\alpha),(\boldsymbol{q}',\alpha')] = \sum_{\boldsymbol{p},\boldsymbol{p}'=\mathbf{1}}^{\boldsymbol{P}^{(\ell)}} \Sigma^{(\ell)}[\boldsymbol{p},\boldsymbol{p}']N[(\tilde{\boldsymbol{q}}(\boldsymbol{p}),\alpha),(\tilde{\boldsymbol{q}}'(\boldsymbol{p}'),\alpha')].$$

If we also define,

$$B_{\boldsymbol{p}}[\boldsymbol{r},\alpha,\boldsymbol{q},\alpha'] = \mathbf{1}_{\{\boldsymbol{r}=\tilde{\boldsymbol{q}}(\boldsymbol{p})\}} \cdot \mathbf{1}_{\{\alpha=\alpha'\}} \in \mathbb{R}^{\boldsymbol{F}^{(\ell)}\times k\times \boldsymbol{F}^{(\ell+1)}\times k},$$

then the linear operator can be expressed as $L(N) = \sum_{\boldsymbol{p},\boldsymbol{p}'=\mathbf{1}}^{\boldsymbol{P}^{(\ell)}} \Sigma^{(\ell)}[\boldsymbol{p},\boldsymbol{p}']B_{\boldsymbol{p}}^{\mathsf{T}} N B_{\boldsymbol{p}'}$. In particular $L(M_n) = \hat{K}_0(G^{(\ell)}(\mathcal{X}))$ and $L(M) = K_0^{(\ell+1)}(\mathcal{X})$. Indeed,

$$\begin{aligned}
L(N)[(\boldsymbol{q},\alpha),(\boldsymbol{q}',\alpha')] &= \sum_{\boldsymbol{p},\boldsymbol{p}'=\mathbf{1}}^{\boldsymbol{P}^{(\ell)}} \Sigma^{(\ell)}[\boldsymbol{p},\boldsymbol{p}'](B_{\boldsymbol{p}}^T N B_{\boldsymbol{p}'})[(\boldsymbol{q},\alpha),(\boldsymbol{q}',\alpha')] \\
&= \sum_{\boldsymbol{p},\boldsymbol{p}'=\mathbf{1}}^{\boldsymbol{P}^{(\ell)}} \sum_{\boldsymbol{r},\boldsymbol{r}'=\mathbf{1}}^{\boldsymbol{F}^{(\ell)}} \sum_{\beta,\beta'=1}^{k} \Sigma^{(\ell)}[\boldsymbol{p},\boldsymbol{p}']B_{\boldsymbol{p}}^T[(\boldsymbol{q},\alpha),(\boldsymbol{r},\beta)]N[(\boldsymbol{r},\beta),(\boldsymbol{r}',\beta')]B_{\boldsymbol{p}'}[(\boldsymbol{r}',\beta'),(\boldsymbol{q}',\alpha')] \\
&= \sum_{\boldsymbol{p},\boldsymbol{p}'=\mathbf{1}}^{\boldsymbol{P}^{(\ell)}} \sum_{\boldsymbol{r},\boldsymbol{r}'=\mathbf{1}}^{\boldsymbol{F}^{(\ell)}} \sum_{\beta,\beta'=1}^{k} \Sigma^{(\ell)}[\boldsymbol{p},\boldsymbol{p}']\mathbf{1}_{\{\boldsymbol{r}=\tilde{\boldsymbol{q}}(\boldsymbol{p})\}}\mathbf{1}_{\{\beta=\alpha\}}N[(\boldsymbol{r},\beta),(\boldsymbol{r}',\beta')]\mathbf{1}_{\{\boldsymbol{r}'=\tilde{\boldsymbol{q}}'(\boldsymbol{p}')\}}\mathbf{1}_{\{\beta'=\alpha'\}} \\
&= \sum_{\boldsymbol{p},\boldsymbol{p}'=\mathbf{1}}^{\boldsymbol{P}^{(\ell)}} \Sigma^{(\ell)}[\boldsymbol{p},\boldsymbol{p}']N[(\tilde{\boldsymbol{q}}(\boldsymbol{p}),\alpha),(\tilde{\boldsymbol{q}}'(\boldsymbol{p}'),\alpha')].
\end{aligned}$$

The independence assumption required by the lemma is satisfied since $\mathcal{G}^{(\ell)}(\mathcal{X})$ is centered Gaussian with covariance $\mathrm{Id}_{C^{(\ell)}} \otimes K^{(\ell)}(\mathcal{X})$, hence its coordinates are independent across channel indices, and this property is preserved under ϕ. Since $\hat{K}_0(z)$ and $K_0^{(\ell+1)}(\mathcal{X})$ are both covariance matricies, they are positive semidefinite. Furthermore, by assumption, $K_0^{(\ell+1)}(\mathcal{X})$ is nonzero. Therefore, the lemma yields

$$\mathbb{E}\left[\left\|\sqrt{\hat{K}_0\left(\mathcal{G}^{(\ell)}(\mathcal{X})\right)} - \sqrt{K_0^{(\ell+1)}(\mathcal{X})}\right\|^2\right] \leq \frac{\left\|\Sigma^{(\ell)}\right\|_{1,1}^2 \left(C_{\mathrm{ker}}^{(\ell+1)}\right)^2}{C^{(\ell)}},$$

where the finite constant $C_{\mathrm{ker}}^{(\ell+1)}$ is explicitly

$$C_{\mathrm{ker}}^{(\ell+1)} := \sqrt{\mathbb{E}\left[\left\|\phi\left(\mathcal{G}^{(\ell)}(\mathcal{X})\right)[1] \otimes \phi\left(\mathcal{G}^{(\ell)}(\mathcal{X})\right)[1] - V^{(\ell)}(\mathcal{X})\right\|^2\right] / \lambda^+\left(K_0^{(\ell+1)}(\mathcal{X})\right)}, \tag{29}$$

and depends uniquely on the activation function ϕ, the input set $\mathcal{X}$ and the weights variances $(\Sigma^{(i)}, c_b^{(i)})_{i=1}^{\ell}$.

Combining the above estimates, we obtain the following bound on the second term:

$$\begin{aligned}\mathcal{W}_2^2(g^{(\ell+1)}(\mathcal{X}), \mathcal{G}^{(\ell+1)}(\mathcal{X})) &\leq \mathcal{W}_2^2(h^{(\ell+1)}(\mathcal{X}), H^{(\ell+1)}(\mathcal{X})) \\ &\leq C^{(\ell+1)}\,\mathbb{E}\left[\left\|\sqrt{\hat{K}_0(G^{(\ell)}(\mathcal{X}))} - \sqrt{K_0^{(\ell+1)}(\mathcal{X})}\right\|^2\right] \\ &\leq \frac{\left\|\Sigma^{(\ell)}\right\|_{1,1}^2 C^{(\ell+1)}}{C^{(\ell)}}\left(C_{\mathrm{ker}}^{(\ell+1)}\right)^2.\end{aligned}$$

Combining the terms. Combining the bounds for the first and second terms gives us

$$\begin{aligned}&\mathcal{W}_2\left(Z^{(\ell+1)}(\mathcal{X}), \mathcal{G}^{(\ell+1)}(\mathcal{X})\right) \leq \\ &\sqrt{\frac{C^{(\ell+1)}}{C^{(\ell)}}}\left(\mathrm{Lip}(\phi)\sqrt{C_{\mathrm{patch}}^{(\ell)}\|\Sigma^{(\ell)}\|_{\mathrm{op}}\,\mathbb{E}\left(\|Z^{(\ell)}(\mathcal{X}) - \mathcal{G}^{(\ell)}(\mathcal{X})\|^2\right)} + \left\|\Sigma^{(\ell)}\right\|_{1,1} C_{\mathrm{ker}}^{(\ell+1)}\right).\end{aligned}$$

Since we considered an arbitrary probability space, where $Z^{(\ell)}(\mathcal{X})$ and $\mathcal{G}^{(\ell)}(\mathcal{X})$ were jointly defined, we consider the infimum over such spaces in the above inequality, which yields (28). □

2.1 Proof of lemma 2.1

Lemma 2.1. Fix a layer index ℓ. Let $W^{(\ell)} \in \mathbb{R}^{C^{(\ell+1)} \times C^{(\ell)} \times \boldsymbol{P}^{(\ell)}}$ be the convolution weight matrix defined in (19). Let $A \in \mathbb{R}^{C^{(\ell)} \times \boldsymbol{F}^{(\ell)} \times k}$ and B be a random variable with values in $\mathbb{R}^{C^{(\ell+1)} \times \boldsymbol{F}^{(\ell+1)} \times k}$ and satisfying

$$B[i, \boldsymbol{q}, \alpha] = \sum_{j=1}^{C^{(\ell)}} \sum_{\boldsymbol{p}=\boldsymbol{1}}^{\boldsymbol{P}^{(\ell)}} W^{(\ell)}[i, j, \boldsymbol{p}] A[j, \tilde{\boldsymbol{q}}(\boldsymbol{p}), \alpha].$$

Then B is a centred Gaussian random variable with the covariance

$$\Sigma(B) = \mathrm{Id}_{C^{(\ell+1)}} \otimes \bar{A}, \tag{24}$$

where $\bar{A} \in \mathbb{R}^{(\boldsymbol{F}^{(\ell+1)} \times k) \times (\boldsymbol{F}^{(\ell+1)} \times k)}$ is given by

$$\bar{A}[(\boldsymbol{q}, \alpha), (\boldsymbol{q}', \alpha')] \coloneqq \frac{1}{C^{(\ell)}} \sum_{j=1}^{C^{(\ell)}} \sum_{\boldsymbol{p}, \boldsymbol{p}'=\boldsymbol{1}}^{\boldsymbol{P}^{(\ell)}} \Sigma^{(\ell)}[\boldsymbol{p}, \boldsymbol{p}'] A[j, \tilde{\boldsymbol{q}}(\boldsymbol{p}), \alpha] A[j, \tilde{\boldsymbol{q}}'(\boldsymbol{p}'), \alpha'].$$

Moreover,

$$\mathbb{E}\left[\|B\|^2\right] \leq \frac{\|\Sigma^{(\ell)}\|_{\mathrm{op}} C^{(\ell)}_{\mathrm{patch}} C^{(\ell+1)}}{C^{(\ell)}} \|A\|^2. \tag{25}$$

Proof. We first compute the covariance. For any pair of indices $(i, \boldsymbol{q}, \alpha)$ and $(i', \boldsymbol{q}', \alpha')$,

$$\begin{aligned}
&\Sigma(B)[(i, \boldsymbol{q}, \alpha), (i', \boldsymbol{q}', \alpha')] = \mathbb{E}\left[B[i, \boldsymbol{q}, \alpha] B[i', \boldsymbol{q}', \alpha']\right] \\
&= \sum_{j, j', \boldsymbol{p}, \boldsymbol{p}'} \mathbb{E}\left[W^{(\ell)}[i, j, \boldsymbol{p}] A[j, \tilde{\boldsymbol{q}}(\boldsymbol{p}), \alpha] W^{(\ell)}[i', j', \boldsymbol{p}'] A[j', \tilde{\boldsymbol{q}}'(\boldsymbol{p}'), \alpha']\right] \\
&= \frac{1}{C^{(\ell)}} \sum_{j, j', \boldsymbol{p}, \boldsymbol{p}'} \mathbf{1}_{\{i=i'\}} \mathbf{1}_{\{j=j'\}} \Sigma^{(\ell)}[\boldsymbol{p}, \boldsymbol{p}'] A[j, \tilde{\boldsymbol{q}}(\boldsymbol{p}), \alpha] A[j', \tilde{\boldsymbol{q}}'(\boldsymbol{p}'), \alpha'] \\
&= \mathbf{1}_{\{i=i'\}} \frac{1}{C^{(\ell)}} \sum_{j, \boldsymbol{p}, \boldsymbol{p}'} \Sigma^{(\ell)}[\boldsymbol{p}, \boldsymbol{p}'] A[j, \tilde{\boldsymbol{q}}(\boldsymbol{p}), \alpha] A[j, \tilde{\boldsymbol{q}}'(\boldsymbol{p}'), \alpha'].
\end{aligned}$$

For the second part, define

$$\boldsymbol{v}[j, \boldsymbol{q}, \alpha] = (A[j, \tilde{\boldsymbol{q}}(\boldsymbol{p}), \alpha])_{\boldsymbol{p} \in [\![\boldsymbol{P}^{(\ell)}]\!]} \in \mathbb{R}^{\boldsymbol{P}^{(\ell)}}.$$

Since $\Sigma^{(\ell)}$ is positive semidefinite,

$$\begin{aligned}
0 &\leq \boldsymbol{v}^{\mathsf{T}}[j, \boldsymbol{q}, \alpha] \Sigma^{(\ell)} \boldsymbol{v}[j, \boldsymbol{q}, \alpha] \leq \|\boldsymbol{v}[j, \boldsymbol{q}, \alpha]\| \left\|\Sigma^{(\ell)} \boldsymbol{v}[j, \boldsymbol{q}, \alpha]\right\| \\
&\leq \|\Sigma^{(\ell)}\|_{\mathrm{op}} \|\boldsymbol{v}[j, \boldsymbol{q}, \alpha]\|^2,
\end{aligned} \tag{30}$$

where we used the Cauchy-Schwarz inequality and a property of the operator norm.

Using (6), we see that

$$\begin{aligned}
\mathbb{E}\left[\|B\|^2\right] &= \operatorname{tr}\left(\mathrm{Id}_{C^{(\ell+1)}} \otimes \bar{A}\right) \\
&= \sum_{i=1}^{C^{(\ell+1)}} \sum_{\boldsymbol{q}=\mathbf{1}}^{\boldsymbol{F}^{(\ell+1)}} \sum_{\alpha=1}^{k} \mathbf{1}_{\{i=i\}} \frac{1}{C^{(\ell)}} \sum_{j=1}^{C^{(\ell)}} \sum_{\boldsymbol{p},\boldsymbol{p}'=\mathbf{1}}^{\boldsymbol{P}^{(\ell)}} \Sigma^{(\ell)}[\boldsymbol{p},\boldsymbol{p}'] A[j,\tilde{\boldsymbol{q}}(\boldsymbol{p}),\alpha] A[j,\tilde{\boldsymbol{q}}(\boldsymbol{p}'),\alpha] \\
&= \frac{C^{(\ell+1)}}{C^{(\ell)}} \sum_{j=1}^{C^{(\ell)}} \sum_{\boldsymbol{q}=\mathbf{1}}^{\boldsymbol{F}^{(\ell+1)}} \sum_{\alpha=1}^{k} \boldsymbol{v}^{\mathsf{T}}[j,\boldsymbol{q},\alpha] \Sigma^{(\ell)} \boldsymbol{v}[j,\boldsymbol{q},\alpha] \\
&\leq \frac{\|\Sigma^{(\ell)}\|_{\mathrm{op}} C^{(\ell+1)}}{C^{(\ell)}} \sum_{j=1}^{C^{(\ell)}} \sum_{\boldsymbol{q}=\mathbf{1}}^{\boldsymbol{F}^{(\ell+1)}} \sum_{\alpha=1}^{k} \|\boldsymbol{v}[j,\boldsymbol{q},\alpha]\|^2 \\
&= \frac{\|\Sigma^{(\ell)}\|_{\mathrm{op}} C^{(\ell+1)}}{C^{(\ell)}} \sum_{j=1}^{C^{(\ell)}} \sum_{\boldsymbol{q}=\mathbf{1}}^{\boldsymbol{F}^{(\ell+1)}} \sum_{\alpha=1}^{k} \sum_{\boldsymbol{p}=\mathbf{1}}^{\boldsymbol{P}^{(\ell)}} A[j,\tilde{\boldsymbol{q}}(\boldsymbol{p}),\alpha]^2 \\
&= \frac{\|\Sigma^{(\ell)}\|_{\mathrm{op}} C^{(\ell+1)}}{C^{(\ell)}} \sum_{j=1}^{C^{(\ell)}} \sum_{\boldsymbol{q}=\mathbf{1}}^{\boldsymbol{F}^{(\ell)}} \sum_{\alpha=1}^{k} N^{(\ell)}(\boldsymbol{q}) A[j,\boldsymbol{q},\alpha]^2 \\
&\leq \frac{\|\Sigma^{(\ell)}\|_{\mathrm{op}} C^{(\ell)}_{\mathrm{patch}} C^{(\ell+1)}}{C^{(\ell)}} \sum_{j=1}^{C^{(\ell)}} \sum_{\boldsymbol{q}'=\mathbf{1}}^{\boldsymbol{F}^{(\ell)}} \sum_{\alpha=1}^{k} A[j,\boldsymbol{q}',\alpha]^2 \\
&= \frac{\|\Sigma^{(\ell)}\|_{\mathrm{op}} C^{(\ell)}_{\mathrm{patch}} C^{(\ell+1)}}{C^{(\ell)}} \|A\|^2,
\end{aligned}$$

where $N^{(\ell)}$ was defined as (17) and $C^{(\ell)}_{\mathrm{patch}}$ as (18). The term $N^{(\ell)}$ takes into consideration the fact, that there might exists different indices $\boldsymbol{q}, \boldsymbol{q}' \in [\![\boldsymbol{F}^{(\ell+1)}]\!]$ and $\boldsymbol{p}, \boldsymbol{p}' \in [\![\boldsymbol{P}^{(\ell)}]\!]$, such that $\tilde{\boldsymbol{q}}(\boldsymbol{p})$ and $\tilde{\boldsymbol{q}}'(\boldsymbol{p}')$ index the same element, that is $\tilde{\boldsymbol{q}}(\boldsymbol{p}) = \tilde{\boldsymbol{q}}'(\boldsymbol{p}')$. □

2.2 Proof of lemma 2.2

Lemma 2.2. Let S, T and I be finite sets and let $X = (X[i])_{i=1}^n$ be independent and identically distributed random variables with values in $\mathbb{R}^S$ and finite fourth moment. Furthermore, let $M = \mathbb{E}[X[1] \otimes X[1]]$ and define the following random variable, taking values in $\mathbb{R}^{S \times S}$,

$$M_n = \frac{1}{n} \sum_{i=1}^{n} X[i] \otimes X[i].$$

Let $\Sigma \in \mathbb{R}^{I\times I}$ be a symmetric and positive semidefinite matrix and let $L : \mathbb{R}^{S\times S} \to \mathbb{R}^{T\times T}$ be a linear operator that can be written as

$$L(N) \coloneqq \sum_{i,j\in I} \Sigma[i,j] B_i^\mathsf{T} N B_j, \tag{26}$$

where $N \in \mathbb{R}^{S\times S}$ and $B_i \in \{0,1\}^{S\times T}$ is a partial permutation matrix, that is, for every $s \in S$ we have $\sum_{t\in T} B_i[s,t] \in \{0,1\}$, and for every $t \in T$ we have $\sum_{s\in S} B_i[s,t] \in \{0,1\}$. Moreover, assume that $L(M)$ is nonzero, symmetric and positive semidefinite.

Then,

$$\mathbb{E}\left[\left\|\sqrt{L(M_n)} - \sqrt{L(M)}\right\|^2\right] \le \frac{\|\Sigma\|_{1,1}^2 \cdot \mathbb{E}\left[\|X[1]\otimes X[1] - M\|^2\right]}{n\lambda^+(L(M))}$$

where $\lambda^+(L(M)) > 0$ denotes the smallest strictly positive eigenvalue of $L(M)$, which exists since $L(M) \neq 0$. The right-hand side is finite by the finite fourth moment assumption.

Proof. The proof consists of three parts: applying (5), computing $\|L(M_n) - L(M)\|^2$ and $\|M_n - M\|^2$.

Part 1. We start off by showing that

$$\|L(M_n) - L(M)\| \le \|M_n - M\| \, \|\Sigma\|_{1,1}. \tag{31}$$

Let f_i be a function that maps a column index of B_i to a row index, such that the element at that row and column is nonzero, that is

$$f_i(s) = t \iff B_i[t,s] = 1.$$

Then for $i, i' \in I$ and $N \in \mathbb{R}^{S\times S}$

$$\begin{aligned}
\|B_i^\mathsf{T} N B_{i'}\|^2 &= \sum_{s,t} (B_i^\mathsf{T} N B_{i'})^2[s,t] \\
&= \sum_{s,t} \left(\sum_{u,v} B_i[u,s] N[u,v] B_{i'}[v,t] \right)^2 \\
&= \sum_{s\in \mathrm{dom}(f_i)} \sum_{t\in \mathrm{dom}(f_{i'})} \left(B_i[f_i(s),s] N[f_i(s), f_{i'}(t)] B_{i'}[f_{i'}(t),t]\right)^2 \\
&= \sum_{s\in \mathrm{dom}(f_i)} \sum_{t\in \mathrm{dom}(f_{i'})} N^2[f_i(s), f_{i'}(t)] \\
&\le \sum_{s,t} N^2[s,t] = \|N\|^2.
\end{aligned}$$

Using the triangle inequality for the Frobenius norm and applying the previous result, we get

$$\begin{aligned}\|L(M_n) - L(M)\| = \|L(M_n - M)\| &\leq \sum_{i,j\in I} \left\|\Sigma[i,j] B_i^{\mathsf{T}} (M_n - M) B_j\right\| \\ &= \sum_{i,j\in I} |\Sigma[i,j]| \left\|B_i^{\mathsf{T}} (M_n - M) B_j\right\| \\ &\leq \|M_n - M\| \sum_{i,j\in I} |\Sigma[i,j]| \\ &= \|M_n - M\| \, \|\Sigma\|_{1,1} .\end{aligned}$$

Part 2. Next, we calculate $\|M_n - M\|^2$. Note that by the i.i.d. assumption

$$M = \mathbb{E}\left[M_n\right], \quad \text{and} \quad \Sigma\left(M_n\right) = \Sigma\left(\frac{1}{n}\sum_{i=1}^{n} X[i] \otimes X[i]\right) = \frac{\Sigma(X[1] \otimes X[1])}{n},$$

so that, by (6),

$$\mathbb{E}\left[\|M_n - M\|^2\right] = \operatorname{tr}\left(\Sigma\left(M_n\right)\right) = \frac{\operatorname{tr}\left(\Sigma(X[1] \otimes X[1])\right)}{n} = \frac{\mathbb{E}\left[\|X[1] \otimes X[1] - M\|^2\right]}{n}. \tag{32}$$

Part 3. To conclude the proof, we want to apply (5) to $\left\|\sqrt{L(M_n)} - \sqrt{L(M)}\right\|^2$. However, we need to take into account that $L(M)$ may not be strictly positive, hence $\lambda(L(M))$ could be null. To this aim, we set $M' = L(M)$, $M'_n = L(M_n)$ and diagonalize $M' = UDU^{\mathsf{T}}$ with U orthogonal and $D = \sum_{t\in T} d_t e_t \otimes e_t$ diagonal. Let $T^+ = \{t \in T : d_t > 0\}$ (non-empty since $L(M)$ is not null by assumption) and $T^0 = \{t \in T : d_t = 0\}$ be a partition of T.

Define $D_n = U^{\mathsf{T}} M'_n U$. Since M'_n is positive semidefinite, then D_n is also positive semidefinite. Therefore, for each $t \in T$, we have $D_n[t,t] \geq 0$. For $t \in T^0$, its expectation equals

$$\mathbb{E}\left[D_n[t,t]\right] = e_t^{\mathsf{T}} U^{\mathsf{T}} \, \mathbb{E}[M'_n] U e_t = e_t^{\mathsf{T}} U^{\mathsf{T}} M' U e_t = e_t^{\mathsf{T}} D e_t = d_t = 0.$$

A non-negative random variable with zero expectation is zero almost surely, hence $D_n[t,t] = 0$ almost surely for $t \in T^0$. For a positive semidefinite matrix, a zero diagonal entry forces the entire corresponding row and column to be also zero. Therefore, both D and D_n have a block structure where entries

outside $T^+ \times T^+$ are null. It then follows that the square root $\sqrt{D_n}$ has the same block structure and by appling (5) only to these blocks we obtain

$$\left\|\sqrt{D} - \sqrt{D_n}\right\| \leq \frac{\|D - D_n\|}{\sqrt{\lambda^+(D)}},$$

where $\lambda^+(D) = \lambda^+(M')$ is strictly positive.

Note that $D_n = U^\mathsf{T} M_n' U$ implies $M_n' = U D_n U^\mathsf{T}$ and by (3)

$$\sqrt{M'} = U\sqrt{D}U^\mathsf{T} \quad \text{and} \quad \sqrt{M_n'} = U\sqrt{D_n}U^\mathsf{T}.$$

Since U is orthogonal, the norm satisfies $\left\|UAU^\mathsf{T}\right\| = \|A\|$ for every $A \in \mathbb{R}^{T\times T}$, hence

$$\left\|\sqrt{M'} - \sqrt{M_n'}\right\| = \left\|\sqrt{D} - \sqrt{D_n}\right\| \leq \frac{\|M' - M_n'\|}{\sqrt{\lambda^+(M')}}. \tag{33}$$

Concolusion. Combining everything together, we get

$$\begin{aligned}
\mathbb{E}\left[\left\|\sqrt{L(M_n)} - \sqrt{L(M)}\right\|^2\right] &\overset{(33)}{\leq} \frac{\mathbb{E}\left[\|L(M) - L(M_n)\|^2\right]}{\lambda^+(L(M))} \\
&\overset{(31)}{\leq} \frac{\|\Sigma\|_{1,1}^2\, \mathbb{E}\left[\|M_n - M\|^2\right]}{\lambda^+(L(M))} \\
&\overset{(32)}{=} \frac{\|\Sigma\|_{1,1}^2 \cdot \mathbb{E}\left[\|X[1] \otimes X[1] - M\|^2\right]}{n\lambda^+(L(M))}.
\end{aligned}$$

□

Conclusion

In this thesis, we established a quantitative bound on the quadratic Wasserstein distance between the output distribution of a randomly initialized convolutional neural network and its associated neural network Gaussian process by extending the result of Basteri and Trevisan (2024) to the convolutional setting. We derived this in the case where the weights are possibly patch-wise correlated and showed how it depends on the network's architectural parameters, such as depth, channel width, patch size, and the structure of patch correlations. In the special case where the CNN architecture reduces to a fully connected neural network, we recover the result of Basteri and Trevisan (2024).

We now discuss the main limitations of this work and possible future research directions.

1. The thesis does not include numerical experiments. The theoretical bounds should be empirically validated and compared with the observed convergence rates for concrete architectures and correlation matrices.

2. The current analysis relies on $L(N)$ in Lemma 2.2 or equivalently on $K_0^{(\ell+1)}$ being nonzero. In the proof of Lemma 1.4, we verified this condition for the independent weights and spatially correlated cases, but the mean-pooling case remains unresolved and would need a separate argument.

3. The result should generalise to arbitrary $\mathcal{W}_p$ Wasserstein distances. Lemma 1.1 was introduced precisely to establish property (13) in this generality, but a corresponding generalisation of Lemma 2.2 is still needed. Trevisan (2023) addressed this in an analogous step via Rosenthal's inequality, but this leaves the constants implicit.

4. As noted by Basteri and Trevisan (2024), the approach should in principle extend to other architectures. The present thesis carries this out for CNNs, but a natural step would be to place the result within the tensor program framework of Yang (2019). The main obstacle to this extension is the presence of architectures in which the same weight matrix appears more than once. Such weight tying renders the conditioning arguments used in the proof of Theorem 2.1 invalid.

References

Basteri, Andrea and Dario Trevisan (2024). "Quantitative Gaussian approximation of randomly initialized deep neural networks". In: *Machine Learning* 113.9, pp. 6373–6393. DOI: 10.1007/s10994-024-06578-z.

Dumoulin, Vincent and Francesco Visin (2016). "A guide to convolution arithmetic for deep learning". In: *arXiv preprint arXiv:1603.07285*.

Favaro, Stefano, Boris Hanin, Domenico Marinucci, Ivan Nourdin, and Giovanni Peccati (2025). "Quantitative CLTs in deep neural networks". In: *Probability Theory and Related Fields* 191.3, pp. 933–977.

Garriga-Alonso, Adrià, Carl Edward Rasmussen, and Laurence Aitchison (2019). "Deep Convolutional Networks as shallow Gaussian Processes". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=Bklfsi0cKm.

Garriga-Alonso, Adrià and Mark van der Wilk (2021). "Correlated weights in infinite limits of deep convolutional neural networks". In: *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*. Vol. 161. Proceedings of Machine Learning Research. PMLR, pp. 1998–2007.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep Learning*. http://www.deeplearningbook.org. MIT Press.

Hanin, Boris (2018). "Which neural net architectures give rise to exploding and vanishing gradients?" In: *Advances in neural information processing systems* 31.

Hanin, Boris and David Rolnick (2018). "How to start training: The effect of initialization and architecture". In: *Advances in neural information processing systems* 31.

Hemmen, J. L. van and T. Ando (1980). "An inequality for trace ideals". In: *Communications in Mathematical Physics* 76.2, pp. 143–148. DOI: 10.1007/BF01212822.

Jacot, Arthur, Franck Gabriel, and Clément Hongler (2018). "Neural tangent kernel: Convergence and generalization in neural networks". In: *Advances in neural information processing systems* 31.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton (2015). "Deep learning". In: *Nature* 521.7553, pp. 436–444. DOI: 10.1038/nature14539.

Lee, Jaehoon, Yasaman Bahri, Roman Novak, Samuel S. Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein (2018). "Deep Neural Networks as Gaussian Processes". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=B1EA-M-0Z.

Lee, Jaehoon, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington (2019). "Wide neural

networks of any depth evolve as linear models under gradient descent". In: *Advances in neural information processing systems* 32.

Matthews, Alexander G. de G., Jiri Hron, Mark Rowland, Richard E. Turner, and Zoubin Ghahramani (2018). "Gaussian Process Behaviour in Wide Deep Neural Networks". In: International Conference on Learning Representations. URL: https://openreview.net/forum?id=H1-nGgWC-.

Mosig, Eloy, Andrea Agazzi, and Dario Trevisan (2025). "Quantitative convergence of trained single layer neural networks to Gaussian processes". In: *arXiv preprint arXiv:2509.24544*.

Neal, Radford M. (1996). "Priors for Infinite Networks". In: *Bayesian Learning for Neural Networks*. Springer New York, pp. 29–53. ISBN: 978-1-4612-0745-0. DOI: 10.1007/978-1-4612-0745-0_2.

Novak, Roman, Lechao Xiao, Yasaman Bahri, Jaehoon Lee, Greg Yang, Daniel A. Abolafia, Jeffrey Pennington, and Jascha Sohl-dickstein (2019). "Bayesian Deep Convolutional Networks with Many Channels are Gaussian Processes". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=B1g30j0qF7.

Papoulis, Athanasios (1991). *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, p. 148.

Rasmussen, Carl Edward and Christopher K. I. Williams (2006). *Gaussian Processes for Machine Learning*. MIT Press. Chap. 4.

Trevisan, Dario (2023). "Wide deep neural networks with Gaussian weights are very close to Gaussian processes". In: *arXiv preprint arXiv:2312.11737*.

Villani, Cédric (2009). *Optimal Transport: Old and New*. Vol. 338. Grundlehren der mathematischen Wissenschaften. Springer. DOI: 10.1007/978-3-540-71050-9.

Yang, Greg (2019). "Wide Feedforward or Recurrent Neural Networks of Any Architecture are Gaussian Processes". In: *Advances in Neural Information Processing Systems*. Vol. 32.

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Karl-Joan Alesma,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose „Quantitative convergence of convolutional neural networks with correlated weights", mille juhendajad on Eloy Mosig García, Andrea Agazzi ja Jüri Lember, reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 4.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.

3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.

4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Karl-Joan Alesma
19.05.2026