DEEP BAYESIAN CONVOLUTIONAL NETWORKS WITH MANY CHANNELS ARE GAUSSIAN PROCESSES

Anonymous authors
Paper under double-blind review

ABSTRACT

There is a previously identified equivalence between wide fully connected neural networks (FCNs) and Gaussian processes (GPs). This equivalence enables, for instance, test set predictions that would have resulted from a fully Bayesian, infinitely wide trained FCN to be computed without ever instantiating the FCN, but by instead evaluating the corresponding GP. In this work, we derive an analogous equivalence for multi-layer convolutional neural networks (CNNs) both with and without pooling layers, and achieve state of the art results on CIFAR10 for GPs without trainable kernels. We also introduce a Monte Carlo method to estimate the GP corresponding to a given neural network architecture, even in cases where the analytic form has too many terms to be computationally feasible.

Surprisingly, in the absence of pooling layers, the GPs corresponding to CNNs with and without weight sharing are identical. As a consequence, translation equivariance in finite-channel CNNs trained with stochastic gradient descent (SGD) has no corresponding property in the Bayesian treatment of the infinite channel limit – a qualitative difference between the two regimes that is not present in the FCN case. We confirm experimentally, that while in some scenarios the performance of SGD-trained finite CNNs approaches that of the corresponding GPs as the channel count increases, with careful tuning SGD-trained CNNs can significantly outperform their corresponding GPs.

1 INTRODUCTION

Neural networks (NNs) demonstrate remarkable performance (He et al., 2016; Oord et al., 2016; Silver et al., 2017; Vaswani et al., 2017), but are still only poorly understood from a theoretical perspective (Goodfellow et al., 2015; Choromanska et al., 2015; Pascanu et al., 2014; Zhang et al., 2017). NN performance is often motivated in terms of model architectures, initializations, and training procedures together specifying biases, constraints, or implicit priors over the class of functions learned by a network. This induced structure in learned functions is believed to be well matched to structure inherent in many practical machine learning tasks, and in many real-world datasets. For instance, properties of NNs which are believed to make them well suited to modeling the world include: hierarchy and compositionality (Lin et al., 2017; Poggio et al., 2017), Markovian dynamics (Tiňo et al., 2004; 2007), and equivariances in time and space for RNNs (Werbos, 1988) and CNNs (Fukushima & Miyake, 1982; Rumelhart et al., 1985) respectively.

The recent discovery of an equivalence between deep neural networks and GPs (Lee et al., 2018; Matthews et al., 2018b) allow us to express an analytic form for the prior over functions encoded by deep NN architectures and initializations. This transforms an implicit prior over functions into an *explicit prior*, which can be analytically interrogated and reasoned about[1].

Previous work studying these Neural Network-equivalent Gaussian Processes (NN-GPs) has established the correspondence only for fully connected networks (FCNs). Additionally, previous work has not used analysis of NN-GPs to gain specific insights into the equivalent NNs.

In the present work, we extend the equivalence between NNs and NN-GPs to deep *Convolutional* Neural Networks (CNNs), both with and without pooling. CNNs are a particularly interesting architecture for study, since they are frequently held forth as a success of motivating NN design based on invariances and equivariances of the physical world (Cohen & Welling, 2016) – specifically, designing a NN to respect translation equivariance (Fukushima & Miyake, 1982; Rumelhart et al., 1985).

[1] While there is broad literature on empirical interpretation of finite CNNs (Zeiler & Fergus, 2014; Simonyan et al., 2014; Long et al., 2014; Olah et al., 2017), it is commonly only applicable to fully trained networks.

As we will see in this work, absent pooling, this quality can vanish in the Bayesian treatment of the infinite channel (number of convolutional filters) limit.

The specific novel contributions of the present work are:

1. We show analytically that CNNs with many channels, trained in a fully Bayesian fashion, correspond to an NN-GP (§2, §3). We show this for CNNs both with and without pooling, with arbitrary convolutional striding, and with both same and valid padding. We prove convergence as the number of channels in hidden layers approach infinity simultaneously (i.e. $\min\left\{n^1, \ldots, n^L\right\} \to \infty$, see §A.5.4 for details), strengthening and extending the result of Matthews et al. (2018a) under mild conditions on the nonlinearity derivative.
2. We show that in the absence of pooling, the NN-GP for a CNN and a Locally Connected Network (LCN) are identical (§5.1). An LCN has the same local connectivity pattern as a CNN, but without weight sharing or translation equivariance.
3. We experimentally compare trained CNNs and LCNs and find that under certain conditions both perform similarly to the respective NN-GP (Figure 3, b, c). Moreover, both architectures tend to perform better with increased channel count, suggesting that similarly to FCNs (Neyshabur et al., 2015; Novak et al., 2018) CNNs benefit from overparameterization (Figure 3, a, b), corroborating a similar trend observed in Canziani et al. (2016, Figure 2). However, we also show that careful tuning of hyperparameters allows finite CNNs trained with SGD to outperform their corresponding NN-GP by a significant margin. We experimentally disentangle and quantify the contributions stemming from local connectivity, equivariance, and invariance in a convolutional model in one such setting (Table 1).
4. We introduce a Monte Carlo method to compute NN-GP kernels for situations (such as CNNs with pooling) where evaluating the NN-GP is otherwise computationally infeasible (§4).

We stress that we do not evaluate finite width Bayesian networks nor do we make any claims about their performance relative to the infinite width GP studied here or finite width SGD-trained networks. While this is an interesting subject to pursue (see Matthews et al. (2018a); Neal (1994)), it is outside of the scope of this paper.

1.1 Related work

In early work on neural network priors, Neal (1994) demonstrated that, in a fully connected network with a single hidden layer, certain natural priors over network *parameters* give rise to a Gaussian process prior over *functions* when the number of hidden units is taken to be infinite. Follow-up work extended the conditions under which this correspondence applied (Williams, 1997; Le Roux & Bengio, 2007; Hazan & Jaakkola, 2015). An exactly analogous correspondence for infinite width, finite depth *deep* fully connected networks was developed recently in Lee et al. (2018); Matthews et al. (2018b), with Matthews et al. (2018a) extending the convergence guarantees to a more general class of nonlinearities and width growth rates.

The line of work examining signal propagation in random deep networks (Poole et al., 2016; Schoenholz et al., 2017; Yang & Schoenholz, 2017; Hanin & Rolnick, 2018; Chen et al., 2018) is related to the construction of the GPs we consider. They apply a mean field approximation in which the pre-activation signal is replaced with a Gaussian, and the derivation of the covariance function with depth is the same as for the kernel function of a corresponding GP. Recently, Xiao et al. (2017b; 2018) extended this to convolutional architectures without pooling. Xiao et al. (2018) also analyzed properties of the convolutional kernel at large depths to construct a *phase diagram* which will be relevant to NN-GP performance, as discussed in §A.2.

Compositional kernels coming from wide convolutional and fully connected layers also appeared outside of the GP context. Cho & Saul (2009) derived closed-form compositional kernels for rectified polynomial activations (including ReLU). Daniely et al. (2016) proved approximation guarantees between a network and its corresponding kernel, and show that empirical kernels will converge as the number of channels increases.

There is a line of work considering stacking of GPs, such as *deep GP*s (Lawrence & Moore, 2007; Damianou & Lawrence, 2013). These no longer correspond to GPs, though they can describe a

rich class of probabilistic models beyond GPs. Alternatively, *deep kernel learning* (Wilson et al., 2016b;a; Bradshaw et al., 2017) utilizes GPs with base kernels which take in features produced by a deep neural network (often a CNN), and train the resulting model end-to-end. Finally, van der Wilk et al. (2017) incorporates convolutional structure into GP kernels, with follow-up work stacking multiple such GPs (Kumar et al., 2018; Blomqvist et al., 2018; Anonymous, 2019) to produce a deep convolutional GP (which is no longer a GP). Our work differs from all of these in that our GP corresponds exactly to a fully Bayesian CNN in the infinite channel limit, when all layers are taken to be of infinite size. We remark that while alternative models, such as deep GPs, do include infinite-sized layers in their construction, they do not treat all layers in this way – for instance, through insertion of bottleneck layers which are kept finite. While it remains to be seen exactly which limit is applicable for understanding realistic CNN architectures in practice, the limit we consider is natural for a large class of CNNs, namely those for which all layers sizes are large and rather comparable in size.

Borovykh (2018) analyzes the convergence of CNN outputs at different spatial locations (or different timepoints for a temporal CNN) to a GP for a single input example. Thus, while they also consider a GP limit (and perform an experiment comparing posterior GP predictions to an SGD-trained CNN), they do not address the dependence of network outputs on multiple input examples, and thus their model is unable to generate predictions on a test set consisting of new input examples.

In concurrent work, Garriga-Alonso et al. (2018) derive an NN-GP kernel equivalent to one of the kernels considered in our work. In addition to explicitly specifying kernels corresponding to pooling and vectorizing, we also compare the NN-GP performance to finite-width SGD-trained CNNs and analyze the differences between the two models.

2 Many-channel Bayesian CNNs are Gaussian processes

2.1 Preliminaries

General setup. For simplicity of presentation we consider 1D convolutional networks with circularly-padded activations. Our analysis is straighforward to extend to higher dimensions, using zero ("same") or no ("valid") padding, and strided convolutions (§A.3). We consider a series of $L+1$ convolutional hidden layers, $l = 0, ..., L$.

Random weights and biases. The parameters of the network are the convolutional filters and biases, $\omega^l_{ij,\beta}$ and b^l_i, respectively, with outgoing (incoming) channel index i (j) and filter relative spatial location $\beta \in [\pm k] \equiv \{-k, \ldots, 0, \ldots, k\}$.[2] Assume a Gaussian prior on both the filter weights and biases,

$$\omega^l_{ij,\beta} \sim \mathcal{N}\left(0, v_\beta \frac{\sigma^2_\omega}{n^l}\right), \qquad b^l_i \sim \mathcal{N}\left(0, \sigma^2_b\right). \tag{1}$$

The weight and bias variances are $\sigma^2_\omega, \sigma^2_b$, respectively. n^l is the number of channels (filters) in layer l, $2k+1$ is the filter size, and v_β is the fraction of the receptive field variance at location β (with $\sum_\beta v_\beta = 1$). In experiments we utilize uniform $v_\beta = 1/(2k+1)$, but nonuniform $v_\beta \neq 1/(2k+1)$ should enable kernel properties that are better suited for ultra-deep networks, as in Xiao et al. (2018).

Inputs, pre-activations, and activations. Let $\mathcal{X}$ denote a set of input images (training set or validation set or both). The network has activations $y^l(x)$ and pre-activations $z^l(x)$ for each input image $x \in \mathcal{X} \subset \mathbb{R}^{n^0 d}$, with input channel count $n^0 \in \mathbb{N}$, number of pixels $d \in \mathbb{N}$, where

$$y^l_{i,\alpha}(x) \equiv \begin{cases} x_{i,\alpha} & l = 0 \\ \phi\left(z^{l-1}_{i,\alpha}(x)\right) & l > 0 \end{cases}, \qquad z^l_{i,\alpha}(x) \equiv \sum_{j=1}^{n^l} \sum_{\beta=-k}^{k} \omega^l_{ij,\beta} y^l_{j,\alpha+\beta}(x) + b^l_i. \tag{2}$$

We emphasize the dependence of $y^l_{i,\alpha}(x)$ and $z^l_{i,\alpha}(x)$ on the input x. $\phi : \mathbb{R} \to \mathbb{R}$ is a nonlinearity (with elementwise application to higher-dimensional inputs). Similarly to Xiao et al. (2018), y^l

[2] We will use Roman letters to index channels and Greek letters for spatial location. We use letters i, j, i', j', etc to denote channel indices, α, α', etc to denote spatial indices and β, β', etc for filter indices.

is assumed to be circularly-padded and the spatial size d hence remains constant throughout the network. See Figure 4 for a visual depiction of our notation.

Activation covariance. A recurring quantity in this work will be the empirical uncentered covariance matrix K^l of the activations y^l, defined as

$$\left[K^l\right]_{\alpha,\alpha'}(x,x') \equiv \frac{1}{n^l}\sum_{i=1}^{n^l} y^l_{i,\alpha}(x) y^l_{i,\alpha'}(x'). \tag{3}$$

K^l is therefore a random variable indexed by two inputs x, x' and two spatial locations α, α' (the dependence on layer widths $n^1, \ldots, n^l$, as well as weights and biases, is implied and by default not stated explicitly). K^0, the empirical uncentered covariance of inputs, is deterministic.

Shapes and indexing. Whenever an index is omitted, the variable is assumed to contain all possible entries along the respective dimension. E.g. y^0 is a vector of size $|\mathcal{X}|\, n^0 d$, $\left[K^l\right]_{\alpha,\alpha'}$ is a matrix of shape $|\mathcal{X}| \times |\mathcal{X}|$, z^l_j is a vector of size $|\mathcal{X}|\, d$, etc.

Our work concerns proving that the top-layer pre-activations z^L converge in distribution to an $|\mathcal{X}|\, n^{L+1} d$-variate normal random vector with a particular covariance matrix of shape $\left(|\mathcal{X}|\, n^{L+1} d\right) \times \left(|\mathcal{X}|\, n^{L+1} d\right)$. Therefore we always consider z^l, y^l as well as any of their indexed subsets like z^l_j, $y^l_{i,\alpha}$ to be 1D vector random variables, while K^l, as well as any of their indexed subsets (when applicable, e.g. $\left[K^l\right]_{\alpha,\alpha'}$, $\left[K^l\right](x,x')$) to be 2D matrix random variables.

2.2 Correspondence between Gaussian processes and Bayesian deep CNNs with infinitely many channels

We next consider the prior over outputs z^L computed by a CNN in the limit of infinitely many channels in the hidden (excluding input and output) layers, $\min\left\{n^1, \ldots, n^L\right\} \to \infty$, and derive its equivalence to a GP with a compositional kernel. This section provides a brief argument showing that z^L is normally distributed conditioned on previous layer activation covariance K^L, which itself becomes deterministic in the infinite limit. As we elaborate in §A.5.4, this allows to conclude convergence in distribution of the outputs to a Gaussian with the respective deterministic covariance limit.

2.2.1 A single convolutional layer is a GP conditioned on the uncentered covariance matrix of the previous layer's activations

As can be seen in Equation 2, the pre-activations z^l are an affine transformation of the multivariate Gaussian $\left\{\omega^l, b^l\right\}$, specified by the previous layer's activations y^l. An affine transformation of a multivariate Gaussian is itself a Gaussian with a covariance matrix that can be derived straightforwardly. Specifically,

$$\left(z^l | y^l\right) \sim \mathcal{N}\left(0, \mathcal{A}\left(K^l\right) \otimes I_{n^{l+1}}\right), \tag{4}$$

where $I_{n^{l+1}}$ is an $n^{l+1} \times n^{l+1}$ identity matrix, and $\mathcal{A}\left(K^l\right)$ is the covariance of the pre-activations z^l_i and is derived in Xiao et al. (2018). Precisely, $\mathcal{A} : \mathbb{R}^{|\mathcal{X}|d \times |\mathcal{X}|d} \to \mathbb{R}^{|\mathcal{X}|d \times |\mathcal{X}|d}$ is an affine transformation (a cross-correlation operator followed by a shifting operator) defined as follows:

$$\left[\mathcal{A}\left(K\right)\right]_{\alpha,\alpha'}(x,x') \equiv \sigma_b^2 + \sigma_\omega^2 \sum_\beta v_\beta \left[K\right]_{\alpha+\beta,\alpha'+\beta}(x,x'). \tag{5}$$

Notice that the covariance matrix in Equation 4 factorizes due to the fact that separate channels $\left\{z^l_i\right\}_{i=1}^{n^{l+1}}$ are i.i.d. conditioned on y^l, due to i.i.d. weights and biases $\left\{\omega^l_i, b^l_i\right\}_{i=1}^{n^{l+1}}$.

We further remark that per Equation 4 the normal distribution of $\left(z^l | y^l\right)$ only depends on K^l, hence the random variable $\left(z^l | K^l\right)$ has the same distribution by the law of total expectation (as long as K^l is positive semidefinite):

$$\left(z^l | K^l\right) \sim \mathcal{N}\left(0, \mathcal{A}\left(K^l\right) \otimes I_{n^{l+1}}\right). \tag{6}$$

2.2.2 ACTIVATION COVARIANCE MATRIX BECOMES DETERMINISTIC WITH INCREASING CHANNEL COUNT

It follows from Equation 6 that the summands in Equation 3 are i.i.d. conditioned on K^{l-1}. Subject to weak restrictions on the nonlinearity ϕ, we can apply the weak law of large number and conclude that the covariance matrix K^l becomes deterministic in the infinite channel limit in layer l (note that pre-activations z^l remain stochastic). Precisely,

$$\forall K^{l-1} \quad \left(K^l | K^{l-1}\right) \xrightarrow[n^l \to \infty]{P^3} \left(\mathcal{C} \circ \mathcal{A}\right)\left(K^{l-1}\right) \quad \text{(in probability)}, \tag{7}$$

where

$$\left[\mathcal{C}\left(K\right)\right]_{\alpha,\alpha'}\left(x, x'\right) \equiv \mathbb{E}_{z \sim \mathcal{N}(0,K)}\left[\phi\left(z_\alpha(x)\right)\phi\left(z_{\alpha'}(x')\right)\right]. \tag{8}$$

The decoupling of the kernel "propagation" into $\mathcal{C}$ and $\mathcal{A}$ is highly convenient since $\mathcal{A}$ is a simple affine transformation of the kernel (see Equation 5), and $\mathcal{C}$ is a well-studied map in literature (see §A.7.4), and for nonlinearities such as ReLU (Nair & Hinton, 2010) and the error function (erf) $\mathcal{C}$ can be computed in closed form as derived in Cho & Saul (2009) and Williams (1997) respectively. We refer the reader to Xiao et al. (2018, Lemma A.1) for complete derivation of the limiting value in Equation 7.

A less obvious result is that, under slightly stronger assumptions on ϕ, the top-layer activation covariance K^L becomes unconditionally (dependence on observed deterministic inputs y^0 is implied) deterministic as channels in all hidden layers grow to infinity simultaneously:

$$K^L \xrightarrow[\min\{n^1,\ldots,n^L\} \to \infty]{P} K^L_\infty \equiv \left(\mathcal{C} \circ \mathcal{A}\right)^L\left(K^0\right), \tag{9}$$

i.e. K^L_∞ is $(\mathcal{C} \circ \mathcal{A})$ applied L times to K^0, the deterministic input covariance. We prove this in §A.5.4 (Theorem A.4).

2.2.3 A CONDITIONALLY NORMAL RANDOM VARIABLE BECOMES NORMAL IF ITS COVARIANCE BECOMES DETERMINISTIC

§2.2.1 established that $\left(z^L | K^L\right)$ is Gaussian, and §2.2.2 established that its covariance matrix $\mathcal{A}\left(K^L\right) \otimes I_{n^{L+1}}$ converges in probability to a deterministic $\mathcal{A}\left(K^L_\infty\right) \otimes I_{n^{l+1}}$ in the infinite channel limit (since $K^L \xrightarrow{P} K^L_\infty$, and $\mathcal{A}\left(\cdot\right) \otimes I_{n^{L+1}} : \mathbb{R}^{|\mathcal{X}|d \times |\mathcal{X}|d} \to \mathbb{R}^{n^{L+1}|\mathcal{X}|d \times n^{L+1}|\mathcal{X}|d}$ is continuous). As we establish in §A.5.4 (Theorem A.5), this is sufficient to conclude with the following result.

Result. If $\phi : \mathbb{R} \to \mathbb{R}$ has an exponentially-bounded derivative almost everywhere, i.e. $\exists\, a, b \in \mathbb{R} : \forall x \in \mathbb{R} \;\; |\phi'\left(x\right)| \leq a \exp\left(bx\right)$ a.e., then the following convergence in distribution holds:

$$\left(z^L | y^0\right) \xrightarrow[\min\{n^1,\ldots,n^L\} \to \infty]{\mathcal{D}} \mathcal{N}\left(0, \mathcal{A}\left(K^L_\infty\right) \otimes I_{n^{L+1}}\right). \tag{10}$$

3 TRANSFORMING A GP OVER SPATIAL LOCATIONS INTO A GP OVER CLASSES

In §2.2 we have shown that in the infinite channel limit a deep CNN is a GP indexed by input samples, output spatial locations, and output channel indices. Further, its covariance matrix K^L_∞ can be computed in closed form. Here we show that transformations to obtain class predictions that are common in CNN classifiers can be represented as either vectorization or projection (as long as we treat classification as regression, similarly to Lee et al. (2018)). Both of these operations preserve the GP equivalence and allow the computation of the covariance matrix of the respective GP (now indexed by input samples and target classes) as a simple transformation of K^L_∞.

[3] Weak law of large numbers allows convergence in probability of individual entries of K^l. However, due to a finite dimensionality of K^l, joint convergence in probability follows.

[4] Unlike Matthews et al. (2018a), we do not require each n^l to be strictly increasing.

In the following we will denote C as the number of target classes, $\bar{z}^{L+1}(x) \in \mathbb{R}^C$ as the outputs of the CNN-based classifier, and $\bar{\omega}^{L+1}$, $\bar{b}^{L+1}$ as the last (number $L+1$) layer weights and biases. For each class i we will denote the empirical uncentered covariance of the outputs as

$$\mathcal{K}_i \equiv \left(\bar{z}_i^{L+1}\right)\left(\bar{z}_i^{L+1}\right)^T, \tag{11}$$

a random variable of shape $|\mathcal{X}| \times |\mathcal{X}|$. We will show that in the scenarios considered below in §3.1 and §3.2, the outputs $\left(\bar{z}_i^{L+1}|y^0\right)$ will converge in distribution to a multivariate Gaussian (i.i.d. for every class i) as $\min\{n^1, \ldots, n^{L+1}\} \to \infty$ with covariance denoted as $\mathcal{K}_\infty \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{X}|}$, or, jointly, $\left(\bar{z}^{L+1}|y^0\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \mathcal{K}_\infty \otimes I_C\right)$. However, as in §2.2, we postpone the formal derivation until §A.5.4 (Theorem A.5).

3.1 Vectorization

One common readout strategy is to vectorize (flatten) the output of the last convolutional layer $z^L(x) \in \mathbb{R}^{n^{L+1}d}$ into a vector[5] $\mathrm{vec}\left[z^L(x)\right] \in \mathbb{R}^{n^{L+1}d}$ and stack a fully connected layer on top:

$$\bar{z}_i^{L+1}(x) \equiv \sum_{j=1}^{n^{L+1}d} \bar{\omega}_{ij}^{L+1} \phi\left(\mathrm{vec}\left[z^L(x)\right]\right)_j + \bar{b}_i^{L+1}, \tag{12}$$

where the weights $\bar{\omega}^{L+1} \in \mathbb{R}^{C\times n^{L+1}d}$ and biases $\bar{b}^{L+1} \in \mathbb{R}^C$ are i.i.d. Gaussian, $\bar{\omega}_{ij}^{L+1} \sim \mathcal{N}\left(0, \sigma_\omega^2/(n^{L+1}d)\right)$, $\bar{b}_i^{L+1} \sim \mathcal{N}\left(0, \sigma_b^2\right)$.

It is easy to verify that $\left(\bar{z}_i^{L+1}|K^{L+1}\right)$ is a multivariate Gaussian with the covariance:

$$\mathbb{E}\left[\mathcal{K}_i^{\mathrm{vec}}|K^{L+1}\right] = \mathbb{E}\left[\left(\bar{z}_i^{L+1}\right)\left(\bar{z}_i^{L+1}\right)^T|K^{L+1}\right] = \frac{\sigma_\omega^2}{d}\sum_\alpha \left[K^{L+1}\right]_{\alpha,\alpha} + \sigma_b^2. \tag{13}$$

The argument of §2.2 can then be extended to conclude that in the limit of infinite width $\left(\bar{z}_i^{L+1}|y^0\right)$ converges in distribution to a multivariate Gaussian (i.i.d. for each class i) with covariance

$$\mathcal{K}_\infty^{\mathrm{vec}} = \frac{\sigma_\omega^2}{d}\sum_\alpha \left[K_\infty^{L+1}\right]_{\alpha,\alpha} + \sigma_b^2. \tag{14}$$

A sample network using this readout strategy is depicted in Figure 4.

Note that as observed in Xiao et al. (2018), to compute any diagonal terms of $\left[K_\infty^{l+1}\right](x,x')$, one needs only the corresponding diagonal terms of $\left[K_\infty^l\right](x,x')$. Consequently, we only need to compute $\left\{\left[K_\infty^l\right]_{\alpha,\alpha}(x,x') : x,x' \in \mathcal{X}, \alpha \in \{1\ldots d\}\right\}_{l=0,\ldots,L}$ and the memory cost is $\mathcal{O}\left(|\mathcal{X}|^2 d\right)$ (or $\mathcal{O}(d)$ per covariance entry in an iterative or distributed setting). Note that this approach ignores pixel-pixel covariances and produces a GP corresponding to a locally-connected network (see §5.1).

3.2 Projection

Another approach is a projection collapsing the spatial dimensions. Let $h \in \mathbb{R}^d$ be a deterministic vector, $\bar{\omega}_{ij}^{L+1} \sim \mathcal{N}\left(0, \sigma_\omega^2/n^{L+1}\right)$, and $\bar{b}^{L+1}$ be the same as in §3.1.

Define the output to be

$$\bar{z}_i^{L+1}(x) \equiv \sum_{j=1}^{n^{L+1}} \bar{\omega}_{ij}^{L+1}\left(\phi\left(z^L(x)\right)h\right)_j + \bar{b}_i^{L+1}, \quad \text{leading to (analogously to §3.1)} \tag{15}$$

$$\mathcal{K}_\infty^h \equiv \sigma_\omega^2 \sum_{\alpha,\alpha'} h_\alpha h_{\alpha'}\left[K_\infty^{L+1}\right]_{\alpha,\alpha'} + \sigma_b^2. \tag{16}$$

Examples of this approach include

[5] Note that since per our notation described in §2.1 $z^L(x)$ is already considered a 1D vector, here this operation simply amounts to re-indexing $z^L(x)$ with one channel index j instead two (channel i and pixel α).

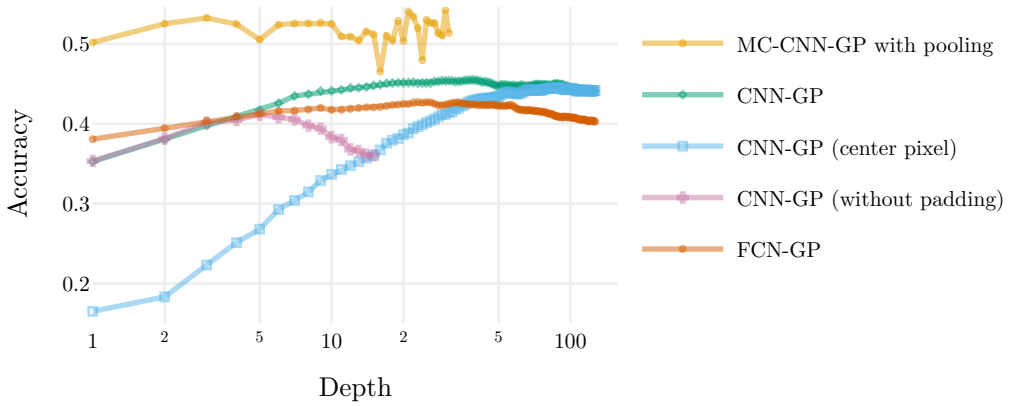

Figure 1: Different dimensionality collapsing strategies described in §3. Validation accuracy of an **MC-CNN-GP with pooling** (§3.2.1) is consistently better than other models due to translation invariance of the kernel. **CNN-GP with zero padding** (§3.1) outperforms an analogous **CNN-GP without padding** as depth increases. At depth 15 the spatial dimension of the output without padding is reduced to 1×1, making the **CNN-GP without padding** equivalent to the **center pixel selection strategy** (§3.2.2) – which also performs worse than the **CNN-GP** (we conjecture, due to overfitting to centrally-located features) but approaches the latter (right) in the limit of large depth, as information becomes more uniformly spatially distributed (Xiao et al., 2018). **CNN-GPs** generally outperform **FCN-GP**, presumably due to the local connectivity prior, but can fail to capture nonlinear interactions between spatially-distant pixels at shallow depths (left). Values are reported on a 2K/4K train/validation subsets of CIFAR10. See §A.7.3 for experimental details.

1. **Global average pooling:** take $h = \frac{1}{d}\mathbf{1_d}$. Then

$$\mathcal{K}_{\infty}^{\text{pool}} \equiv \frac{\sigma_\omega^2}{d^2} \sum_{\alpha,\alpha'} \left[K_\infty^{L+1}\right]_{\alpha,\alpha'} + \sigma_b^2. \tag{17}$$

   This approach corresponds to applying global average pooling right after the last convolutional layer.[6] This approach takes all pixel-pixel covariances into consideration and makes the kernel translation invariant. However, it requires $\mathcal{O}\left(|\mathcal{X}|^2 d^2\right)$ memory to compute the sample-sample covariance of the GP (or $\mathcal{O}\left(d^2\right)$ per covariance entry in an iterative or distributed setting). It is impractical to use this method to analytically evaluate the GP, and we propose to use a Monte Carlo approach (see §4).

2. **Subsampling one particular pixel:** take $h = e_\alpha$,

$$\mathcal{K}_{\infty}^{e_\alpha} \equiv \sigma_\omega^2 \left[K_\infty^{L+1}\right]_{\alpha,\alpha} + \sigma_b^2. \tag{18}$$

   This approach makes use of only one pixel-pixel covariance, and requires the same amount of memory as vectorization (§3.1) to compute.

We compare the performance of presented strategies in Figure 1. Note that all described strategies admit stacking additional FC layers on top while retaining the GP equivalence, using a derivation analogous to §2 (Lee et al., 2018; Matthews et al., 2018b).

4 Monte Carlo Evaluation of Intractable GP Kernels

We introduce a Monte Carlo estimation method for NN-GP kernels which are computationally impractical to compute analytically, or for which we do not know the analytic form. Similar in spirit to traditional random feature methods (Rahimi & Recht, 2007), the core idea is to instantiate many

[6] Spatially local average pooling in intermediary layers can be constructed in a similar fashion (§A.3). We focus on global average pooling in this work to more effectively isolate the effects of pooling from other aspects of the model like local connectivity or equivariance.

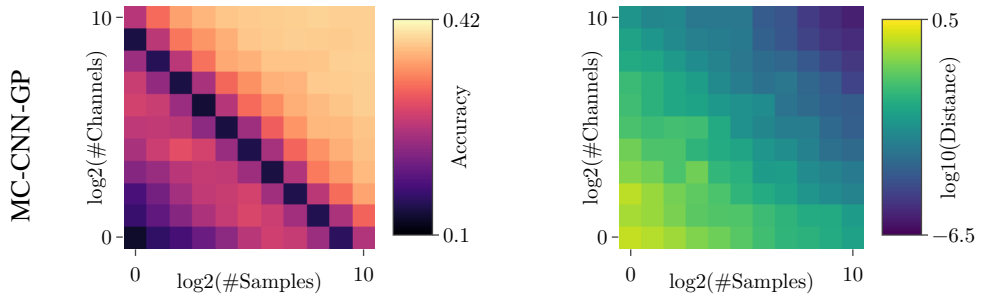

Figure 2: **Validation accuracy** (left) of an MC-CNN-GP increases with $n \times M$ (i.e. channel count times number of samples) and approaches that of the exact CNN-GP (not shown), while the **distance** (right) to the exact kernel decreases. The dark band in the left plot corresponds to ill-conditioning of $K^{L+1}_{n,M}$ when the number of outer products contributing to $K^{L+1}_{n,M}$ approximately equals its rank. Values reported are for a 3-layer model applied to a 2K/4K train/validation subset of CIFAR10 downsampled to 8×8. See Figure 7 for similar results with other architectures and §A.7.2 for experimental details.

random *finite* width networks and use the empirical uncentered covariances of activations to estimate the Monte Carlo-GP (MC-GP) kernel,

$$\left[K^l_{n,M}\right]_{\alpha,\alpha'}(x,x') \equiv \frac{1}{Mn}\sum_{m=1}^{M}\sum_{c=1}^{n} y^l_{c\alpha}\left(x;\theta_m\right) y^l_{c\alpha'}\left(x';\theta_m\right) \tag{19}$$

where θ consists of M draws of the weights and biases from their prior distribution, $\theta_m \sim p(\theta)$, and n is the width or number of channels in hidden layers. The MC-GP kernel converges to the analytic kernel with increasing width, $\lim_{n\to\infty} K^l_{n,M} = K^l_\infty$ in probability.

For finite width networks, the uncertainty in $K^l_{n,M}$ is $\mathrm{Var}\left[K^l_{n,M}\right] = \mathrm{Var}_\theta\left[K^l_n(\theta)\right]/M$. From Daniely et al. (2016), we know that $\mathrm{Var}_\theta\left[K^l_n(\theta)\right] \propto \frac{1}{n}$, which leads to $\mathrm{Var}_\theta[K^l_{n,M}] \propto \frac{1}{Mn}$. For finite n, $K^l_{n,M}$ is also a biased estimate of K^l_∞, where the bias depends solely on network width. We do not currently have an analytic form for this bias, but we can see in Figures 2 and 7 that for the hyperparameters we probe it is small relative to the variance. In particular, $\left\|K^l_{n,M}(\theta) - K^L_\infty\right\|^2_{\mathrm{F}}$ is nearly constant for constant Mn. We thus treat Mn as the effective sample size for the Monte Carlo kernel estimate. Increasing M and reducing n can reduce memory cost, though potentially at the expense of increased compute time and bias.

In a non-distributed setting, the MC-GP reduces the memory requirements to compute $\mathcal{GP}^{\text{pool}}$ from $\mathcal{O}\left(|\mathcal{X}|^2 d^2\right)$ to $\mathcal{O}\left(|\mathcal{X}|^2 + n^2 + nd\right)$, making the evaluation of CNN-GPs with pooling practical.

5 Discussion

5.1 Bayesian CNNs with many channels are identical to locally connected networks, in the absence of pooling

Locally Connected Networks (LCNs) (Fukushima, 1975; Lecun, 1989) are CNNs without weight sharing between spatial locations. LCNs preserve the connectivity pattern, and thus topology, of a CNN. However, they do not possess the equivariance property of a CNN – if an input is translated, the latent representation in an LCN will be completely different, rather than also being translated.

The CNN-GP predictions without spatial pooling in §3.1 and §3.2.2 depend only on sample-sample covariances, and do not depend on pixel-pixel covariances. LCNs destroy pixel-pixel covariances: $\left[K^L_\infty\right]^{\text{LCN}}_{\alpha,\alpha'}(x,x') = 0$, for $\alpha \neq \alpha'$ and all $x, x' \in \mathcal{X}$ and $L > 0$. However, LCNs preserve the covariances between input examples at every pixel: $\left[K^L_\infty\right]^{\text{LCN}}_{\alpha,\alpha}(x,x') = \left[K^L_\infty\right]^{\text{CNN}}_{\alpha,\alpha}(x,x')$. As a result, in the absence of pooling, LCN-GPs and CNN-GPs are identical. Moreover, LCN-GPs with

pooling are identical to CNN-GPs with vectorization of the top layer (under suitable scaling of y^{L+1}). We confirm these findings experimentally in trained networks in the limit of large width in Figure 3 (b), as well as by demonstrating convergence of MC-GPs of the respective architectures to the same CNN-GP (modulo scaling of y^{L+1}) in Figures 2 and 7.

5.2 Pooling leverages equivariance to provide invariance

The only kernel leveraging pixel-pixel covariances is that of the CNN-GP with pooling. This enables the predictions of this GP and the corresponding CNN to be invariant to translations (modulo edge effects) – a beneficial quality for an image classifier. We observe strong experimental evidence supporting the benefits of invariance throughout this work (Figures 1, 2, 3 (b); Tables 1, 2), in both CNNs and CNN-GPs.

5.3 Finite-channel SGD-trained CNNs can outperform infinite-channel Bayesian CNNs, in the absence of pooling

In the absence of pooling, the benefits of equivariance and weight sharing are more challenging to explain in terms of Bayesian priors on class predictions (since without pooling equivariance is not a property of the outputs, but only of intermediary representations). Indeed, in this work we find that the performance of finite-width SGD-trained CNNs often approaches that of their CNN-GP counterpart (Figure 3, b, c)[7], suggesting that in those cases equivariance does not play a beneficial role in SGD-trained networks.

However, as can be seen in Tables 1, 2 and Figure 3 (c), the best CNN *overall* outperforms the best CNN-GP by a significant margin – an observation specific to CNNs and not FCNs or LCNs. We observe this gap in performance especially in the case of ReLU networks trained with a large learning rate. In Table 1 we demonstrate this large gap in performance by evaluating different models with equivalent architecure and hyperparameter settings, chosen for good SGD-trained CNN performance.

We conjecture that equivariance, a property lacking in LCNs and the Bayesian treatment of the infinite channel CNN limit, contributes to the performance of SGD-trained finite-channel CNNs with the correct settings of hyperparameters. Nonetheless, more work is needed to disentangle and quantify the separate contributions of stochastic optimization and finite width effects[8] to differences in performance between CNNs with weight sharing and their corresponding CNN-GPs.

6 Conclusion

In this work we have derived a Gaussian process that corresponds to a deep fully Bayesian CNN with infinitely many channels. The covariance of this GP can be efficiently computed either in closed form or by using Monte Carlo sampling, depending on the architecture.

The CNN-GP achieves state of the art results for GPs without trainable kernels on CIFAR10. It can perform competitively with CNNs (that fit the training set) of equivalent architecture and weight priors, which makes it an appealing choice for small datasets, as it eliminates all training-related hyperparameters. However, we found that the best *overall* performance is achieved by finite SGD-trained CNNs and not by their infinite Bayesian counterparts. We hope our work stimulates future research into disentangling the contributions of the two qualities (Bayesian treatment and infinite width) to the performance gap observed.

[7] This observation is conditioned on the respective NN fitting the training set to 100%. Underfitting breaks the correspondance to an NN-GP, since train set predictions of such a network no longer correspond to the true training labels. Properly tuned underfitting often also leads to better generalization (Table 2).

[8] We remark that concerns about GPs not being able to learn hierarchical representations have been raised in the literature (Matthews et al., 2018a, Section 7), (Neal, 1995, Chapter 5), (MacKay, 2003, Section 45.7). However, practical impact of these assertions have not been carefully investigated empirically or theoretically, and we hope that our work stimulates research in this direction.

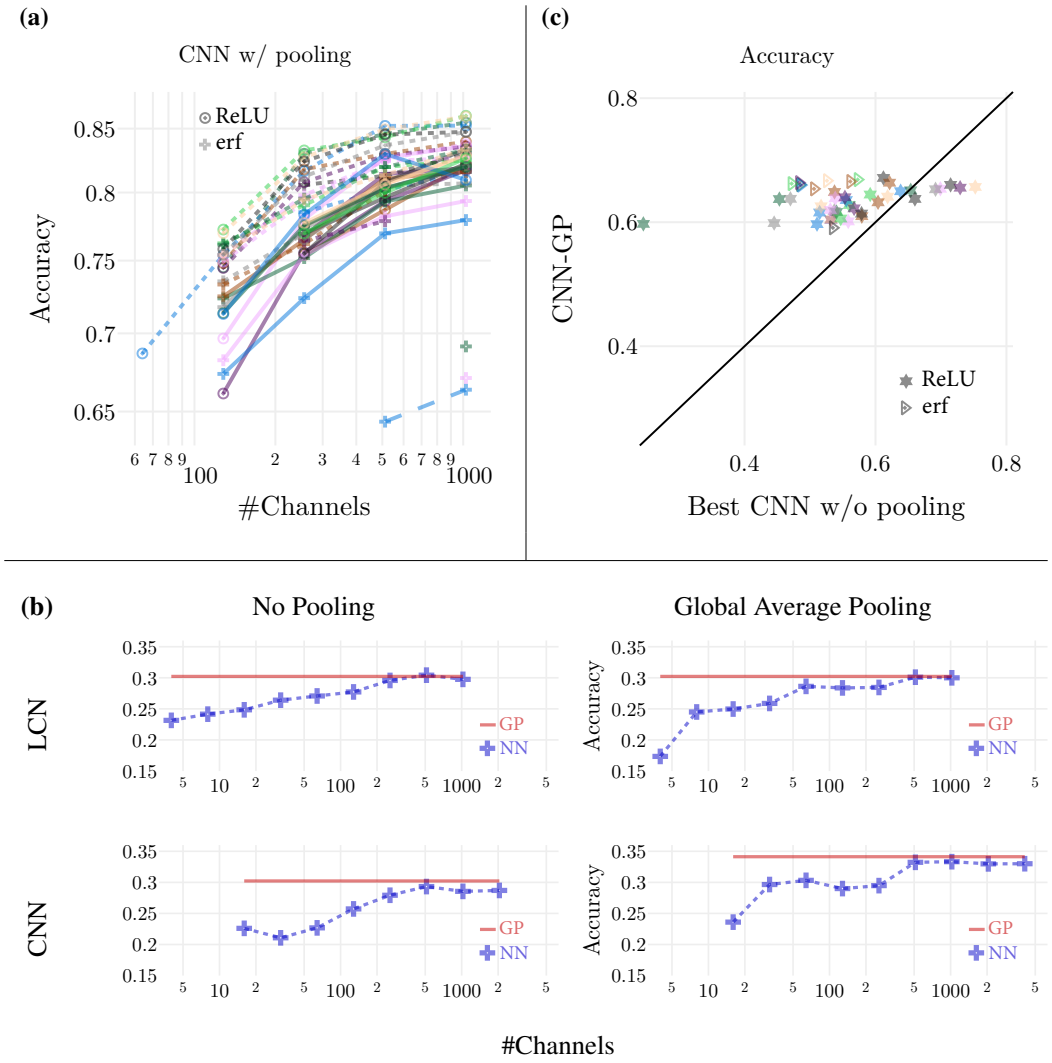

Figure 3: **(a): SGD-trained CNNs often perform better with increasing number of channels.** Each line corresponds to a particular choice of architecture and initialization hyperparameters, with best learning rate and weight decay selected independently for each number of channels (x-axis). **(b): SGD-trained CNNs often approach the performance of their corresponding CNN-GP with increasing number of channels.** All models have the same architecture except for pooling and weight sharing, as well as training-related hyperparameters such as learning rate, weight decay and batch size, which are selected for each number of channels (x-axis) to maximize validation performance (y-axis) of a neural network. As the number of channels grows, best validation accuracy increases and approaches accuracy of the respective GP (solid horizontal line). **(c): However, the best-performing SGD-trained CNNs can outperform their corresponding CNN-GPs.** Each point corresponds to the test accuracy of: (y-axis) a specific CNN-GP; (x-axis) the best (on validation) CNN with the same architectural hyper-parameters selected among the 100%-accurate models on the full training CIFAR10 dataset with different learning rates, weight decay and number of channels. While CNN-GP appears competitive against 100%-accurate CNNs (above the diagonal), the best CNNs *overall* outperform CNN-GPs by a significant margin (below the diagonal, right). For further analysis of factors leading to similar or diverging behavior between SGD-trained finite CNNs and infinite Bayesian CNNs see Tables 1 and 2. **Experimental details:** all networks have reached 100% training accuracy on CIFAR10. Values in (b) are reported on an 0.5K/4K train/validation subset downsampled to 8×8 for computational reasons. See §A.7.5 and §A.7.1 for full experimental details of (a, c) and (b) plots respectively.

Quality:	Compositionality		Local connectivity		Equivariance	Invariance
Model:	FCN	FCN-GP	LCN (w/ pooling)	CNN-GP	CNN	CNN w/ pooling
Error:	46.26	41.45	36.52 (36.23)	36.71	19.93	**16.54**

Table 1: **Disentangling the role of network topology, equivariance, and invariance on test performance, for SGD-trained and infinite width Bayesian networks.** Test error (%) on CIFAR10 of different models of the same depth, nonlinearity, and weight and bias variances. LCN and CNN-GP have a hierarchical local topology, beneficial for image recognition tasks and outperform fully connected models (FCN and FCN-GP). As predicted in §5.1: (i) weight sharing has no effect in the Bayesian treatment of an infinite width CNN (CNN-GP performs similarly to an LCN, a CNN without weight sharing), and (ii) pooling has no effect on generalization of an LCN model (LCN and LCN with pooling perform nearly identically). Local connectivity combined with equivariance (CNN) is enabled by weight sharing in an SGD-trained finite model, allowing for a significant improvement. Finally, invariance enabled by weight sharing and pooling allows for the best performance. Values are reported for 8-layer ReLU models. See §A.7.6 for experimental details and Table 2 for more model comparisons.

Model	CIFAR10	MNIST	Fashion-MNIST
CNN with pooling	**14.85 (15.65)**	–	–
CNN with ReLU and large learning rate	24.76 (17.64)	–	–
CNN-GP	32.86	0.88	**7.40**
CNN with small learning rate	33.31(22.89)	–	–
CNN with erf (any learning rate)	33.31(22.17)	–	–
Convolutional GP (van der Wilk et al., 2017)	35.40	1.17	–
ResNet GP (Garriga-Alonso et al., 2018)	–	**0.84**	–
Residual CNN-GP (Garriga-Alonso et al., 2018)	–	0.96	–
CNN-GP (Garriga-Alonso et al., 2018)	–	1.03	–
FCN-GP	41.06	1.22	8.22
FCN-GP (Lee et al., 2018)	44.34	1.21	–
FCN	45.52 (44.73)	–	–

Table 2: **Aspects of architecture and inference influencing test performance.** Test error (%) for best model within each model family, maximizing validation accuracy over depth, width, and training and initialization hyperpameters. Except where indicated by parentheses, all models achieve 100% training accuracy. For SGD-trained CNNs, numbers in parentheses correspond to the same model family, but without restriction on training accuracy. CNN-GP achieves state of the art results on CIFAR10 for GPs without trainable kernels and outperforms SGD models optimized with a small learning rate to 100% train accuracy. When SGD optimization is allowed to underfit the training set, there is a significant improvement in generalization. Further, when ReLU nonlinearities are paired with large learning rates, the performance of SGD-trained models again improves relative to CNN-GPs, suggesting a beneficial interplay between ReLUs and fast SGD training. These differences in performance between CNNs and CNN-GPs are not observed between FCNs and FCN-GPs, or between LCNs and LCN-GPs (Table 1), suggesting that *equivariance* is the underlying factor responsible for the improved performance of finite SGD-trained CNNs relative to infinite Bayesian CNNs without pooling. See §A.7.5 for experimental details.

REFERENCES

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

Anonymous. Deep convolutional gaussian process. In *Submitted to International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HyeUPi09Y7. under review.

Patrick Billingsley. *Convergence of probability measures*. John Wiley & Sons, 1999.

Kenneth Blomqvist, Samuel Kaski, and Markus Heinonen. Deep convolutional gaussian processes. *arXiv preprint arXiv:1810.03052*, 2018.

Anastasia Borovykh. A gaussian process perspective on convolutional neural networks. *arXiv preprint arXiv:1810.10798*, 2018.

John Bradshaw, Alexander G de G Matthews, and Zoubin Ghahramani. Adversarial examples, uncertainty, and transfer testing robustness in gaussian process hybrid deep networks. *arXiv preprint arXiv:1707.02476*, 2017.

Alfredo Canziani, Adam Paszke, and Eugenio Culurciello. An analysis of deep neural network models for practical applications. *arXiv preprint arXiv:1605.07678*, 2016.

Minmin Chen, Jeffrey Pennington, and Samuel Schoenholz. Dynamical isometry and a mean field theory of RNNs: Gating enables signal propagation in recurrent neural networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 873–882, Stockholmsmssan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL http://proceedings.mlr.press/v80/chen18i.html.

Youngmin Cho and Lawrence K Saul. Kernel methods for deep learning. In *Advances in neural information processing systems*, pp. 342–350, 2009.

Anna Choromanska, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, and Yann LeCun. The loss surfaces of multilayer networks. In *Artificial Intelligence and Statistics*, pp. 192–204, 2015.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pp. 2990–2999, 2016.

Andreas Damianou and Neil Lawrence. Deep gaussian processes. In *Artificial Intelligence and Statistics*, pp. 207–215, 2013.

Amit Daniely, Roy Frostig, and Yoram Singer. Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity. In *Advances In Neural Information Processing Systems*, pp. 2253–2261, 2016.

Kunihiko Fukushima. Cognitron: A self-organizing multilayered neural network. *Biological cybernetics*, 20(3-4):121–136, 1975.

Kunihiko Fukushima and Sei Miyake. Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and cooperation in neural nets*, pp. 267–285. Springer, 1982.

Adrià Garriga-Alonso, Laurence Aitchison, and Carl Edward Rasmussen. Deep convolutional networks as shallow Gaussian processes. *arXiv preprint arXiv:1808.05587*, aug 2018. URL https://arxiv.org/abs/1808.05587.

Daniel Golovin, Benjamin Solnik, Subhodeep Moitra, Greg Kochanski, John Karro, and D Sculley. Google vizier: A service for black-box optimization. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1487–1495. ACM, 2017.

Ian J Goodfellow, Oriol Vinyals, and Andrew M Saxe. Qualitatively characterizing neural network optimization problems. *International Conference on Learning Representations*, 2015.

Boris Hanin and David Rolnick. How to start training: The effect of initialization and architecture. *arXiv preprint arXiv:1803.01719*, 2018.

Tamir Hazan and Tommi Jaakkola. Steps toward deep kernel methods from infinite neural networks. *arXiv preprint arXiv:1508.05133*, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *3rd International Conference for Learning Representations*, 2015.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

Vinayak Kumar, Vaibhav Singh, PK Srijith, and Andreas Damianou. Deep gaussian processes with convolutional kernels. *arXiv preprint arXiv:1806.01655*, 2018.

Neil D Lawrence and Andrew J Moore. Hierarchical gaussian process latent variable models. In *Proceedings of the 24th international conference on Machine learning*, pp. 481–488. ACM, 2007.

Nicolas Le Roux and Yoshua Bengio. Continuous neural networks. In *Artificial Intelligence and Statistics*, pp. 404–411, 2007.

Yann Lecun. Generalization and network design strategies. In *Connectionism in perspective*. Elsevier, 1989.

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

Jaehoon Lee, Yasaman Bahri, Roman Novak, Sam Schoenholz, Jeffrey Pennington, and Jascha Sohl-dickstein. Deep neural networks as gaussian processes. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=B1EA-M-0Z.

Henry W Lin, Max Tegmark, and David Rolnick. Why does deep and cheap learning work so well? *Journal of Statistical Physics*, 168(6):1223–1247, 2017.

Jonathan L Long, Ning Zhang, and Trevor Darrell. Do convnets learn correspondence? In *Advances in Neural Information Processing Systems*, pp. 1601–1609, 2014.

David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

Alexander G de G Matthews, Mark Rowland, Jiri Hron, Richard E Turner, and Zoubin Ghahramani. Gaussian process behaviour in wide deep neural networks. *arXiv preprint arXiv:1804.11271*, 9 2018a.

Alexander G. de G. Matthews, Jiri Hron, Mark Rowland, Richard E. Turner, and Zoubin Ghahramani. Gaussian process behaviour in wide deep neural networks. In *International Conference on Learning Representations*, 4 2018b. URL https://openreview.net/forum?id=H1-nGgWC-.

Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pp. 807–814, 2010.

Radford M. Neal. Priors for infinite networks (tech. rep. no. crg-tr-94-1). *University of Toronto*, 1994.

Radford M Neal. *BAYESIAN LEARNING FOR NEURAL NETWORKS*. PhD thesis, University of Toronto, 1995.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. *Proceeding of the international Conference on Learning Representations workshop track*, abs/1412.6614, 2015.

Roman Novak, Yasaman Bahri, Daniel A. Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. Sensitivity and generalization in neural networks: an empirical study. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=HJC2SzZCW.

Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2(11):e7, 2017.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

Razvan Pascanu, Yann N Dauphin, Surya Ganguli, and Yoshua Bengio. On the saddle point problem for non-convex optimization. *arXiv preprint arXiv:1405.4604*, 2014.

Tomaso Poggio, Hrushikesh Mhaskar, Lorenzo Rosasco, Brando Miranda, and Qianli Liao. Why and when can deep-but not shallow-networks avoid the curse of dimensionality: A review. *International Journal of Automation and Computing*, 14(5):503–519, Oct 2017. ISSN 1751-8520. doi: 10.1007/s11633-017-1054-2. URL https://doi.org/10.1007/s11633-017-1054-2.

Ben Poole, Subhaneil Lahiri, Maithra Raghu, Jascha Sohl-Dickstein, and Surya Ganguli. Exponential expressivity in deep neural networks through transient chaos. In *Advances In Neural Information Processing Systems*, pp. 3360–3368, 2016.

Joaquin Quiñonero-Candela and Carl Edward Rasmussen. A unifying view of sparse approximate gaussian process regression. *Journal of Machine Learning Research*, 6(Dec):1939–1959, 2005.

Ali Rahimi and Ben Recht. Random features for large-scale kernel machines. In *In Neural Information Processing Systems*, 2007.

Carl Edward Rasmussen and Christopher KI Williams. *Gaussian processes for machine learning*, volume 1. MIT press Cambridge, 2006.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

Samuel S Schoenholz, Justin Gilmer, Surya Ganguli, and Jascha Sohl-Dickstein. Deep information propagation. *ICLR*, 2017.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *Nature*, 550(7676):354, 2017.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *ICLR Workshop*, 2014.

Peter Tiňo, Michal Cernansky, and Lubica Benuskova. Markovian architectural bias of recurrent neural networks. *IEEE Transactions on Neural Networks*, 15(1):6–15, 2004.

Peter Tiňo, Barbara Hammer, and Mikael Bodén. Markovian bias of neural-based architectures with feedback connections. In *Perspectives of neural-symbolic integration*, pp. 95–133. Springer, 2007.

Michalis Titsias. Variational learning of inducing variables in sparse gaussian processes. In *Artificial Intelligence and Statistics*, pp. 567–574, 2009.

Mark van der Wilk, Carl Edward Rasmussen, and James Hensman. Convolutional gaussian processes. In *Advances in Neural Information Processing Systems 30*, pp. 2849–2858, 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Paul J Werbos. Generalization of backpropagation with application to a recurrent gas market model. *Neural networks*, 1(4):339–356, 1988.

Christopher KI Williams. Computing with infinite networks. In *Advances in neural information processing systems*, pp. 295–301, 1997.

Andrew G Wilson, Zhiting Hu, Ruslan R Salakhutdinov, and Eric P Xing. Stochastic variational deep kernel learning. In *Advances in Neural Information Processing Systems*, pp. 2586–2594, 2016a.

Andrew Gordon Wilson, Zhiting Hu, Ruslan Salakhutdinov, and Eric P Xing. Deep kernel learning. In *Artificial Intelligence and Statistics*, pp. 370–378, 2016b.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017a.

Lechao Xiao, Yasaman Bahri, Sam Schoenholz, and Jeffrey Pennington. Training ultra-deep cnns with critical initialization. In *NIPS Workshop*, 2017b.

Lechao Xiao, Yasaman Bahri, Jascha Sohl-Dickstein, Samuel Schoenholz, and Jeffrey Pennington. Dynamical isometry and a mean field theory of CNNs: How to train 10,000-layer vanilla convolutional neural networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5393–5402, Stockholmsmssan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL http://proceedings.mlr.press/v80/xiao18a.html.

Ge Yang and Samuel Schoenholz. Mean field residual networks: On the edge of chaos. In *Advances in neural information processing systems*, pp. 7103–7114, 2017.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pp. 818–833. Springer, 2014.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

A Appendix

A.1 Additional Figures

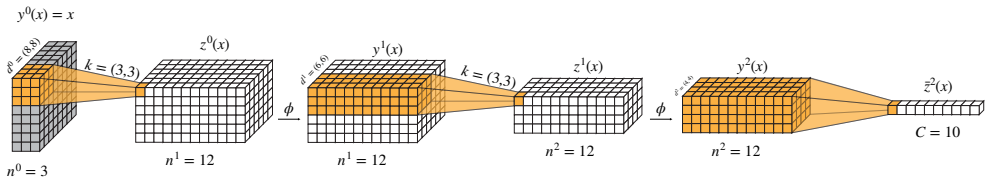

Figure 4: A sample CNN classifier annotated according to notation in §2.1, §3. The network transforms $n^0 \times d^0 = 3 \times 8 \times 8$-dimensional inputs $y^0(x) = x \in \mathcal{X}$ into $C = 10$-dimensional logits $\bar{z}^2(x)$. Model has two convolutional layers with $k = (3, 3)$-shaped filters, nonlinearity ϕ, and a fully-connected layer at the top ($y^2 \to \bar{z}^2(x)$, §3.1). Hidden (pre-)activations have $n^1 = n^2 = 12$ filters. As $\min\left\{n^1, n^2\right\} \to \infty$, the prior of this CNN will approach that of a GP indexed by inputs x and target class indices from 1 to $C = 10$. The covariance of such GP can be computed as $\left[\frac{\sigma_\omega^2}{6\times 6}\sum_{\alpha=(1,1)}^{d^2=(6,6)}\left[\left(\mathcal{C}\circ\mathcal{A}\right)^2\left(K^0\right)\right]_{\alpha,\alpha} + \sigma_b^2\right] \otimes I_C$, where the sum is over the $\{1, \ldots, 6\}^2$ hypercube (see §2.2, §3.1, §A.3). Presented is a CNN with stride 1 and no ("valid") padding, i.e. the spatial shape of the input shrinks as it propagates through it $\left(d^0 = (8, 8) \to d^1 = (6, 6) \to d^2 = (4, 4)\right)$. Note that for notational simplicity 1D CNN and circular padding with $d^0 = d^1 = d^2 = d$ is assumed in the text, yet our formalism easily extends to the model displayed (§A.3). Further, while displayed (pre-)activations have 3D shapes, in the text we treat them as 1D vectors (§2.1).

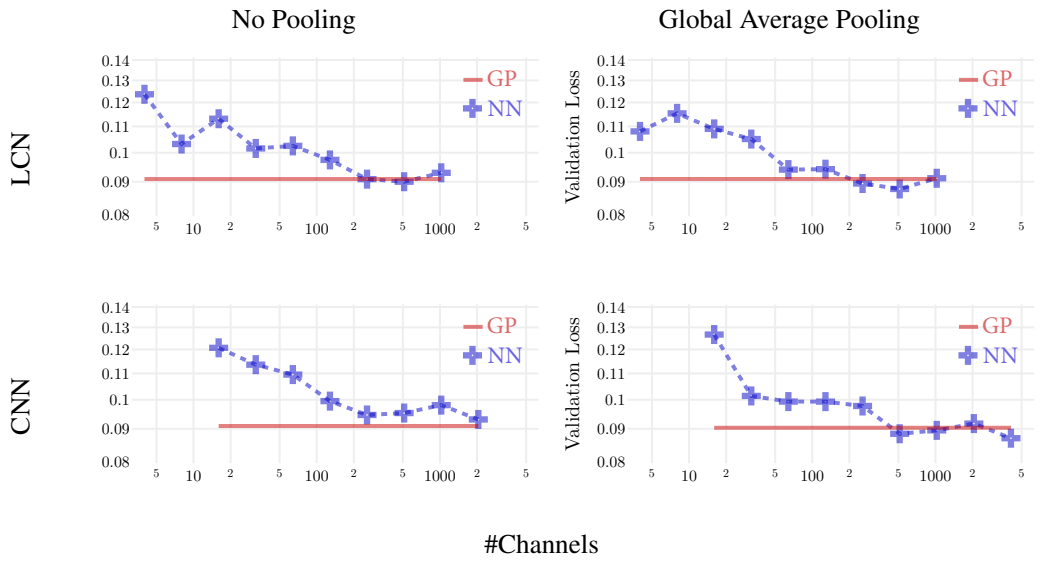

Figure 5: Best validation loss (vertical axis) of **trained neural networks** (dashed line) as the number of channels increases (horizontal axis) approaches that of a respective **(MC-)CNN-GP** (solid horizontal line). See Figure 3 (b) for validation accuracy, Figure 6 for training loss and §A.7.1 for experimental details.

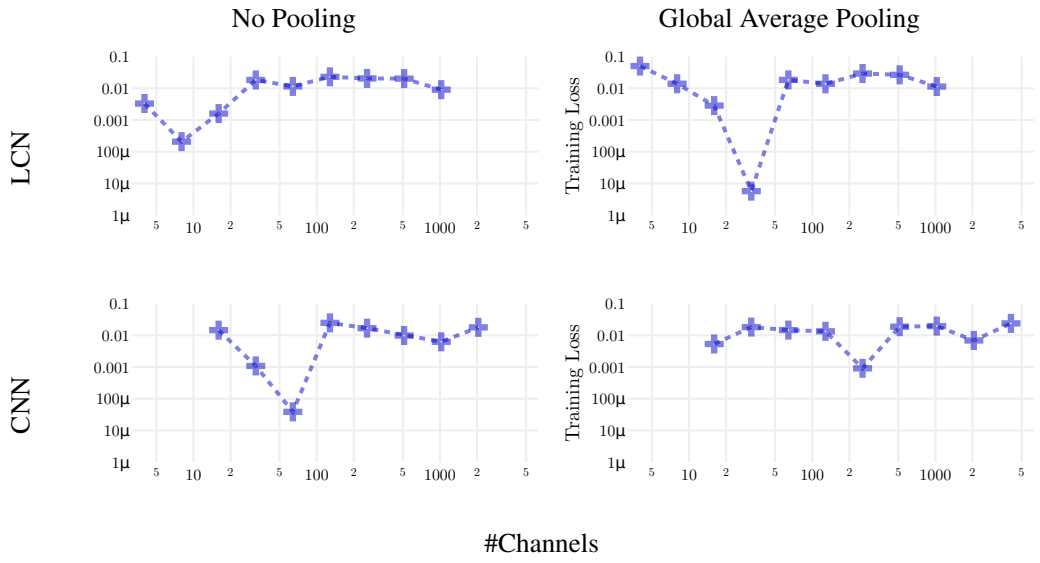

Figure 6: **Training loss** (vertical axis) of best (in terms of validation loss) neural networks as the number of channels increases (horizontal axis). While perfect 0 loss is not achieved (but 100% accuracy is), we observe no consistent improvement when increasing the capacity of the network (left to right). This eliminates underfitting as a possible explanation for why small models perform worse in Figure 3 (b). See Figure 5 for validation loss and §A.7.1 for experimental details.

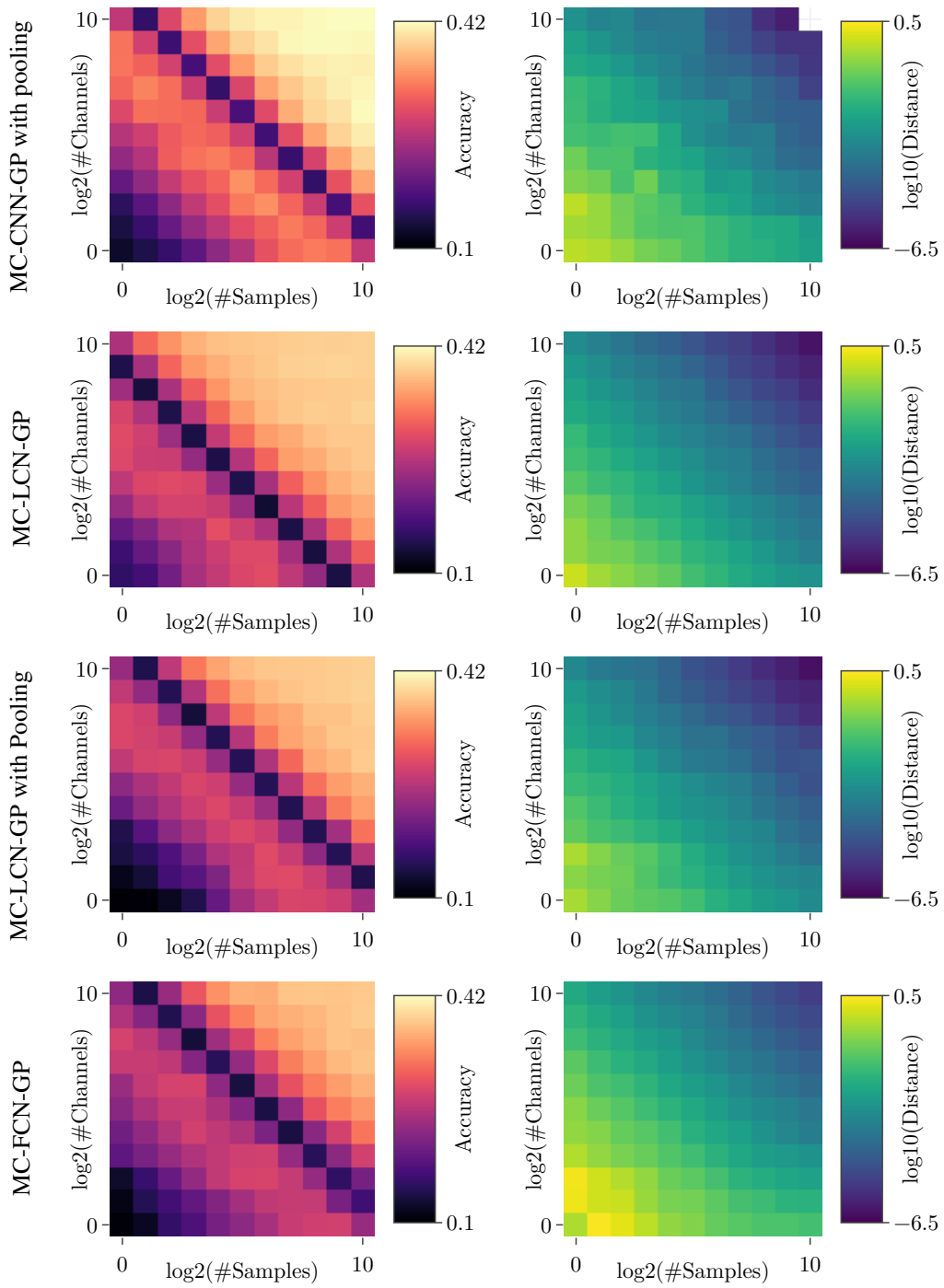

Figure 7: As in Figure 2, **validation accuracy** (left) of MC-GPs increases with $n \times M$ (i.e. width times number of samples), while the **distance** (right) to the the respective exact GP kernel (or the best available estimate in the case of CNN-GP with pooling, top row) decreases. We remark that when using shared weights, convergence is slower as smaller number of independent random parameters are being used. For example a single-layer MC-LCN-GP kernel is expected to converge approximately $\mathrm{Var}[K_{\text{CNN}}]/\mathrm{Var}[K_{\text{LCN}}] \sim \sqrt{\#\text{ LCN params}/\#\text{ CNN params}} = \sqrt{\text{spatial size of the output layer}}$ times faster than MC-CNN-GP, which is in agreement with our results obtained in the second row and Figure 2. I.e. the geometric mean of the ratios of the kernel distance from (3-layer) MC-CNN-GP and MC-LCN-GP to the respective CNN-GP is $\approx 2.2 > \sqrt{\text{spatial size of the output layer}} = \sqrt{2 \times 2} = 2$). See §A.7.2 for experimental details.

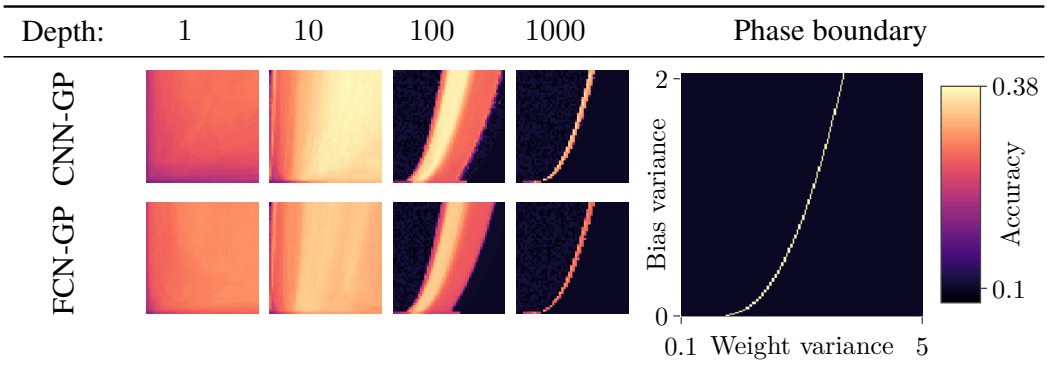

Table 3: **Validation accuracy** of CNN- and FCN- GPs as a function of weight (σ_ω^2, horizontal axis) and bias (σ_b^2, vertical axis) variances. As predicted in §A.2, the regions of good performance concentrate around the critical line (phase boundary, right) as the depth increases (left to right). All plots share common axes ranges and employ the erf nonlinearity. See §A.7.2 for experimental details.

A.2 Relationship to Deep Signal Propagation

The recurrence relation linking the GP kernel at layer $l+1$ to that of layer l following from Equation 10 (i.e. $K_\infty^{l+1} = (\mathcal{C} \circ \mathcal{A}) \left(K_\infty^l\right)$) is precisely the *covariance map* examined in a series of related papers on signal propagation (Xiao et al., 2018; Poole et al., 2016; Schoenholz et al., 2017; Lee et al., 2018) (modulo notational differences; denoted as F, $\mathcal{C}$ or e.g. $\mathcal{A} \star \mathcal{C}$ in Xiao et al. (2018)). In those works, the action of this map on hidden-state covariance matrices was interpreted as defining a dynamical system whose large-depth behavior informs aspects of trainability. In particular, as $l \to \infty$, $K_\infty^{l+1} = (\mathcal{C} \circ \mathcal{A}) \left(K_\infty^l\right) \approx K_\infty^l \equiv K_\infty^*$, i.e. the covariance approaches a fixed point K_∞^*. The convergence to a fixed point is problematic for learning because the hidden states no longer contain information that can distinguish different pairs of inputs. It is similarly problematic for GPs, as the kernel becomes pathological as it approaches a fixed point. Precisely, in the chaotic regime outputs of the GP become asymptotically decorrelated and therefore independent, while in the ordered regime they approach perfect correlation of 1. Either of these scenarios captures no information about the training data in the kernel and makes learning infeasible.

This problem can be ameliorated by judicious hyperparameter selection, which can reduce the rate of exponential convergence to the fixed point. For hyperpameters chosen on a critical line separating two untrainable phases, the convergence rates slow to polynomial, and very deep networks can be trained, and inference with deep NN-GP kernels can be performed – see Table 3.

A.3 Strided convolutions and average pooling in intermediate layers

Our analysis in the main text can easily be extended to cover average pooling and strided convolutions (applied before the pointwise nonlinearity). Recall that conditioned on K^l the pre-activation $z_j^l(x) \in \mathbb{R}^{d_1}$ is a zero-mean multivariate Gaussian. Let $B \in \mathbb{R}^{d_2 \times d_1}$ denote a linear operator. Then $Bz_j^l(x) \in \mathbb{R}^{d_2}$ is a zero-mean Gaussian, and the covariance is

$$\mathbb{E}_{\{\omega^l, b^l\}}\left[\left(Bz_j^l(x)\right)\left(Bz_j^l(x')\right)^T \middle| K^l\right] = B\mathbb{E}_{\{\omega^l, b^l\}}\left[z_j^l(x)\, z_j^l(x')^T \middle| K^l\right] B^T. \tag{20}$$

One can easily see that $\left\{Bz_j^l \middle| K^l\right\}_j$ are i.i.d. multivariate Gaussian as well.

Strided convolution. Strided convolution is equivalent to a non-strided convolution composed with subsampling. Let $s \in \mathbb{N}$ denote size of the stride. Then the strided convolution is equivalent to choosing B as follows: $B_{ij} = \delta(is - j)$ for $i \in \{0, 1, \ldots (d_2 - 1)\}$.

Average pooling. Average pooling with stride s and window size ws is equivalent to choosing $B_{ij} = 1/ws$ for $i = 0, 1, \ldots (d_2 - 1)$ and $j = is, \ldots, (is + ws - 1)$.

ND convolutions. Note that our analysis in the main text (1D) easily extends to higher-dimensional convolutions by replacing integer pixel indices and sizes d, α, β with tuples (see also Figure 4).

In Equation 2 β values would have to span the hypercube $[\pm k]^N = \{-k, \ldots, k\}^N$ in the pre-activation definition. Similarly, in §3 the normalizing factor d (d^2) should be the product (squared) of its entries, and summations over α, β should span the $[d_0] \times \cdots \times [d_N]$ hypercube as well. The definition of the kernel propagation operator $\mathcal{A}$ in Equation 5 will remain exactly the same, so long as β is summed over the hypercube, and the variance weights remain respectively normalized $\sum_\beta v_\beta = 1$.

A.4 Review of exact Bayesian regression with GPs

Our discussion in the paper has focused on model *priors*. A crucial benefit we derive by mapping to a GP is that Bayesian inference is straightforward to implement and can be done *exactly* for regression (Rasmussen & Williams, 2006, chapter 2), requiring only simple linear algebra. Let $\mathcal{X}$ denote training inputs $x_1, ..., x_{|\mathcal{X}|}$, $\mathbf{t}^T = (t_1, ..., t_{|\mathcal{X}|})$ training targets, and collectively $\mathcal{D}$ for the training set. The integral over the posterior can be evaluated analytically to give a posterior predictive distribution on a test point y_* which is Normal, $(z^*|\mathcal{D}, y^*) \sim \mathcal{N}\left(\mu_*, \sigma_*^2\right)$, with

$$\mu_* = \mathcal{K}(x^*, \mathcal{X})(\mathcal{K}(\mathcal{X}, \mathcal{X}) + \sigma_\varepsilon^2 \mathbb{I}_{|\mathcal{X}|})^{-1}\mathbf{t}, \tag{21}$$

$$\sigma_*^2 = \mathcal{K}(x^*, x^*) - \mathcal{K}(x^*, \mathcal{X})(\mathcal{K}(\mathcal{X}, \mathcal{X}) + \sigma_\varepsilon^2 \mathbb{I}_{|\mathcal{X}|})^{-1}\mathcal{K}(\mathcal{X}, x^*). \tag{22}$$

We use the shorthand $\mathcal{K}(\mathcal{X}, \mathcal{X})$ to denote the $|\mathcal{X}| \times |\mathcal{X}|$ matrix formed by evaluating the GP covariance on the training inputs, and likewise $\mathcal{K}(x^*, \mathcal{X})$ is a $|\mathcal{X}|$-length vector formed from the covariance between the test input and training inputs. Computationally, the costly step in GP posterior predictions comes from the matrix inversion, which in all experiments were carried out exactly, and typically scales as $\mathcal{O}(|\mathcal{X}|^3)$ (though algorithms scaling as $\mathcal{O}(|\mathcal{X}|^{2.4})$ exist for sufficiently large matrices). Nonetheless, there is a broad literature on approximate Bayesian inference with GPs which can be utilized for efficient implementation (Rasmussen & Williams, 2006, chapter 8); (Quiñonero-Candela & Rasmussen, 2005; Titsias, 2009).

A.5 Equivalence between randomly initialized NNs and GPs

In this section, we present two different approaches, the *sequential limit* (§A.5.3) and *simultaneous limit* (§A.5.4), to illustrate the relationship between many-channels Bayesian CNNs and GPs.

Sequential limit (§A.5.3) involves taking the infinite channel limit in hidden layers in a sequence, starting from bottom (closest to inputs) layers and going upwards (to the outputs), i.e. $n^1 \to \infty, \ldots, n^L \to \infty$. Note that this approach in fact *constructs* a GP using a NN architecture to define its covariance, and does not provide guarantees on actual convergence of large but finite Bayesian CNNs to GPs, which is of most practical interest. However, it has the following benefits:

1. Weak assumptions on the NN activation function ϕ and on the distribution of the NN parameters.
2. The arguments can be easily extended to more complicated network architectures, e.g. architectures with max pooling, dropout, etc.
3. This approach provides a straightforward and intuitive way to compute the covariance of the Gaussian process without diving into mathematical details.

Simultaneous limit (§A.5.4) considers growing the number of channels in hidden layers uniformly, i.e. $\min\left\{n^1, \ldots, n^L\right\} \to \infty$. This approach establishes convergence of finite-channel Bayesian CNNs to GPs and is thus a more practically relevant result. However, it makes stronger assumptions, and the proof is more involved.

We hightlight that the GPs obtained by the two approaches are identical.

In both sections, we only provide the arguments for CNNs. It is straightforward (and in fact simpler) to extend them to LCNs and FCNs. Indeed, an FCN is a particular case of a CNN where the inputs and filters have singular spatial dimensions ($d = 1$, $k = 0$). For LCNs, the proof goes through in an identical fashion if we replace $\mathcal{A}$ with $\mathcal{A}^{\text{LCN}}$ defined as $\left[\mathcal{A}^{\text{LCN}}(K)\right]_{\alpha,\alpha'}(x, x') \equiv \delta_{\alpha,\alpha'}\left[\mathcal{A}(K)\right]_{\alpha,\alpha'}(x, x')$.

A.5.1 SETUP

Probability space. Let $\mathscr{P}$ be a collection of countably many mutually independent random variables (R.V.s) defined on some probability space $(\Omega, \mathcal{F}, P)$, where $\mathcal{F}$ is a σ-algebra defined on Ω and P is a probability measure. Here $\mathscr{P} \equiv \mathscr{W} \cup \mathscr{B} \cup \mathscr{H}$ is the collection of parameters used to define neural networks:

(i) **Weights.** $\mathscr{W} = \bigcup_{l \in \mathbb{N}} \mathscr{W}^l$ and $\mathscr{W}^l = \left\{\omega^l_{ij,\beta} : i, j \in \mathbb{N}, \beta \in [\pm k]\right\}$, where $[\pm k] \equiv [-k, k] \cap \mathbb{Z}$. We assume $\omega^l_{ij,\beta}$ are i.i.d. R.V.s with mean zero and finite variance $0 < \sigma^2_\omega < \infty$. When $l = 0$, we further assume they are Gaussian distributed.

(ii) **Biases**. $\mathscr{B} = \bigcup_{l \in \mathbb{N}} \mathscr{B}^l$ and $\mathscr{B}^l = \left\{b^l_j : j \in \mathbb{N}\right\}$. We assume b^l_j are i.i.d. Gaussian with mean zero and variance $0 \leq \sigma^2_b < \infty$.

(iii) **Place-holder.** $\mathscr{H}$ is a place-holder to store extra (if needed) R.V.s , e.g. parameters coming from the final dense layer.

Inputs. We will consider a fixed $\mathcal{X} \subseteq \mathbb{R}^{n^0 \times d}$ to denote the inputs, with input channel count n^0, number of pixels d. Assume $x \neq 0$ for all $x \in \mathcal{X}$ and $|\mathcal{X}|$, the cardinality of the inputs, is finite.

However, the *simultaneous limit* (§A.5.4) result can be extended to a countably-infinite input indexing spaces $\mathcal{X}$ for certain topologies via an argument presented in Matthews et al. (2018b, section 2.2), allowing to infer weak convergence on $\mathcal{X}$ from convergence on any finite subset (which is the case we consider in this text; see also Billingsley (1999, page 19) for details).

Notation, shapes, and indexing. We adopt the notation, shape, and indexing convention similar to §2.1, which the reader is encouraged to review. We emphasize that whenever an index is omitted, the variable is assumed to contain all possible entries along the respective dimension (e.g. whole $\mathcal{X}$ if x is omitted, or all n^l channels if the channel i is omitted). In addition, to simplify the notation, we set $v_\beta = \frac{1}{2k+1}$.

A.5.2 PRELIMINARY

We will use the following well-known theorem.

Theorem A.1. Let $X, \{X_n\}_{n \in \mathbb{N}}$ be R.V.s in $\mathbb{R}^m$. The following are equivalent:

(i) $X_n \xrightarrow{\mathcal{D}} X$ (converges in distribution / converges weakly),

(ii) (Portmanteau Theorem) For all bounded continuous function $f : \mathbb{R}^m \to \mathbb{R}$,

$$\lim_{n \to \infty} \mathbb{E}\left[f(X_n)\right] = \mathbb{E}\left[f(X)\right], \tag{23}$$

(iii) (Lévy's Continuity Theorem) The characteristic functions of X_n, i.e. $\mathbb{E}\left[e^{\mathbb{i} t^T X_n}\right]$ converge to that of X pointwisely, i.e. for all $t \in \mathbb{R}^m$,

$$\lim_{n \to \infty} \mathbb{E}\left[e^{\mathbb{i} t^T X_n}\right] = \mathbb{E}\left[e^{\mathbb{i} t^T X}\right], \tag{24}$$

where $\mathbb{i}$ denotes the imaginary unit.

Using the equivalence between (i) and (iii), it is straightforward to show that

$$X_n \xrightarrow{\mathcal{D}} X \quad \Longleftrightarrow \quad a^T X_n \xrightarrow{\mathcal{D}} a^T X \quad \text{for all } a \in \mathbb{R}^m. \tag{25}$$

In particular,

$$X_n \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma) \quad \Longleftrightarrow \quad a^T X_n \xrightarrow{\mathcal{D}} \mathcal{N}(0, a^T \Sigma a) \quad \text{for all } a \in \mathbb{R}^m. \tag{26}$$

A.5.3 SEQUENTIAL LIMIT

In this section, we construct a Gaussian process using an infinite channel CNN with the limits taking sequentially. We will show that pre-activation functions of different channels are i.i.d. Gaussian with mean zero and covariance that can be computed recursively.

Let Ψ_1 denote the following space of functions:
Uniformly square-integrable: for every $R \geq 1$,

$$C_2(R, \phi) \equiv \sup_{1/R \leq r \leq R} \mathbb{E}_{x \sim \mathcal{N}(0,r)} |\phi(x)|^2 < \infty. \tag{27}$$

Let $\mathbb{N}^* = \mathbb{N} \backslash \{0\}$. We construct Gaussian processes recursively using an infinite width neural network architecture with the following procedure:

(i) If $l > 0$, assume $\{z_i^{l-1,\infty}\}_{i \in \mathbb{N}^*}$ are i.i.d. Gaussian with mean zero and variance K_∞^{l-1} and independent from future events (i.e. independent from the σ-alegbra generated by all R.V.s with layer index greater than $(l-1)$). For $n \in \mathbb{N}^*$, let

$$y_{i,\alpha}^l(x) \equiv \begin{cases} x_{i,\alpha} & l = 0 \\ \phi\left(z_{i,\alpha}^{l-1,\infty}(x)\right) & l > 0 \end{cases}, \tag{28}$$

$$z_{i,\alpha}^{l,n}(x) \equiv \begin{cases} \frac{1}{\sqrt{n^0}} \sum_{j \in [n^0]} \frac{1}{\sqrt{2k+1}} \sum_{\beta \in [\pm k]} \omega_{ij,\beta}^l y_{j,\alpha+\beta}^l(x) + b_i^l & l = 0 \\ \frac{1}{\sqrt{n}} \sum_{j \in [n]} \frac{1}{\sqrt{2k+1}} \sum_{\beta \in [\pm k]} \omega_{ij,\beta}^l y_{j,\alpha+\beta}^l(x) + b_i^l & l > 0 \end{cases}, \tag{29}$$

where

$$[n] \equiv \{1, \ldots, n\} \quad \text{and} \quad [\pm k] \equiv \{-k, \ldots, 0, \ldots k\}. \tag{30}$$

(ii) Prove that for any $m \geq 1$, $\left[z_i^{l,n}\right]_{i \in [m]} \subseteq \mathbb{R}^{|\mathcal{X}|d}$ converges in distribution to a multivariate normal with mean zero and covariance $\mathcal{A}\left(K_\infty^l\right) \otimes I_m$, where K_∞^l is defined identically to Equation 10.

(iii) Define random variable $\left[z_i^{l,\infty}\right]_{i \in [m]}$ in $(\Omega, \mathcal{F}, P)$ with the same distribution as the limit in (ii) so that they are also independent from any future events, i.e. independent from the σ-algebra generated by all R.V.s with layer index greater than l.

In (iii), we implicitly assume the probability space $(\Omega, \mathcal{F}, P)$ has enough capacity to fit in all R.V.s generated by the above procedure, if not we will extend the sample space Ω using product measures. In what follows, we use the central limit theorem to prove (ii).

Theorem A.2. If $\phi \in \Psi_1$, then for every $l \geq 0$ and every $m \geq 1$, $\left[z_i^{l,n}\right]_{i \in [m]} \subseteq \mathbb{R}^{|\mathcal{X}|d}$ converges in distribution to a multivariate normal with mean zero and covariance $\mathcal{A}\left(K_\infty^l\right) \otimes I_m$.

Proof. We proceed by induction. This is obvious in the base case $l = 0$, since the weights and biases are assumed to be independent Gaussian. Now we assume the theorem holds for $l - 1$. This implies $\{y_i^l\}_{i \in \mathbb{N}^*} = \left\{\phi(z_i^{l-1,\infty})\right\}_{i \in \mathbb{N}^*}$ are i.i.d. random variables.

Choose a vector $a \equiv [a_i(x)]_{i \in [m],\, x \in \mathcal{X}}^T \in \mathbb{R}^{m|\mathcal{X}|}$. Then

$$\sum_{\substack{i \in [m] \\ x \in \mathcal{X}}} a_i(x) z_i^{l,n}(x) = \sum_{\substack{i \in [m] \\ x \in \mathcal{X}}} a_i(x) \left(\frac{1}{\sqrt{n}} \sum_{j \in [n]} \frac{1}{\sqrt{2k+1}} \sum_{\beta \in [\pm k]} \omega_{ij,\beta}^l y_{j,\alpha+\beta}^l(x) + b_i^l \right) \tag{31}$$

$$= \frac{1}{\sqrt{n}} \sum_{j \in [n]} \sum_{\substack{i \in [m] \\ x \in \mathcal{X}}} a_i(x) \frac{\sum_{\beta \in [\pm k]} \omega_{ij,\beta}^l y_{j,\alpha+\beta}^l(x)}{\sqrt{2k+1}} + \sum_{\substack{i \in [m] \\ x \in \mathcal{X}}} a_i(x) b_i^l \tag{32}$$

$$\equiv \frac{1}{\sqrt{n}} \sum_{j \in [n]} u_j + v. \tag{33}$$

It is not difficult to see that u_j are i.i.d. and v is Gaussian. Then we can use the central limit theorem to conclude that the above converges in distribution to a Gaussian, once we verify the second moment of u_j is finite. Using the fact that $\left\{\omega_{ij,\beta}^l\right\}_{i,\beta}$ is a collection of independent R.V.s and integrating over these R.V.s first, we get

$$\mathbb{E}u_j^2 = \frac{1}{2k+1}\mathbb{E}\left[\sum_{i\in[m]\,,\,x\in\mathcal{X}} a_i(x)\sum_{\beta\in[\pm k]}\omega_{ij,\beta}^l y_{j,\alpha+\beta}^l(x)\right]^2 \tag{34}$$

$$= \frac{\sigma_\omega^2}{2k+1}\sum_{i\in[m]}\sum_{\beta\in[\pm k]}\mathbb{E}\left[\sum_{x\in\mathcal{X}} a_i(x) y_{j,\alpha+\beta}^l(x)\right]^2 \tag{35}$$

$$= \frac{1}{2k+1}\sum_{i\in[m]} a_i^T \bar{\mathcal{A}}\left(K_\infty^l\right) a_i \tag{36}$$

where by $\bar{\mathcal{A}}$ we denote the linear transformation part of $\mathcal{A}$ from Equation 5, i.e. $\mathcal{A}$ without the translation term σ_b^2:

$$\left[\bar{\mathcal{A}}\left(K\right)\right]_{\alpha,\alpha'}(x,x') \equiv \sigma_\omega^2 \sum_\beta \left[K\right]_{\alpha+\beta,\alpha'+\beta}(x,x')\,. \tag{37}$$

To prove finiteness of the second moment in Equation 36, it is sufficient to show that all the diagonal terms of K_∞^l are finite. This easily follows from the assumption $\phi \in \Psi_1$ and the definition of $K_\infty^l = \left(\mathcal{C}\circ\mathcal{A}\right)^l\left(K^0\right)$. Together with the distribution of v (whose covariance is straightforward to compute), the joint distribution of $\left[z_i^{l,n}\right]_{i\in[m]}$ converges weakly to a mean zero Gaussian with covaraince matrix $\mathcal{A}\left(K_\infty^l\right)\otimes I_m$ by Theorem A.1 (Equation 26). □

Remark A.1. The results of Theorem A.2 can be strengthened / extended in many directions.

(1) Same analysis carries over (and the covariance matrix can be computed without much extra effort) if we stack a channel-wise deterministic affine transform after the convolution operator. Note that average pooling (not max pooling, which is not affine), global average pooling and the convolutional striding are particular examples of such affine tranforms. Moreover, valid padding (i.e. no padding) convolution can be regarded as subsampling (a linear projection) operator composed with the regular circular padding.

(2) Same analysis applies to max pooling, but computing the covariance may require non-trivial effort. Let $\mathfrak{m}$ denote the max pooling operator and assume it is applied right after the activation function. The assumption $\left\{y_i^l\right\}_{i\in[n]} = \left\{\phi(z_i^{l-1,\infty})\right\}_{i\in[n]}$ are i.i.d. implies $\left\{\mathfrak{m}\left(y_i^l\right)\right\}_{i\in[n]}$ are also i.i.d. Then we can proceed exactly as above except for verifying the finiteness of second moment of $\mathfrak{m}(y_i^l)$ with the following trivial estimate:

$$\mathbb{E}\max\left\{y_{i,\alpha}^l\right\}_{\alpha\in[s]}^2 \le \mathbb{E}\sum_{\alpha\in[s]}\left|y_{i,\alpha}^l\right|^2 \tag{38}$$

where s is the window size of the max pooling.

In general, one can stack a channel-wise deterministic operator $\mathfrak{op}$ on $\{y_i^l\}$ so long as the second moment of $\{\mathfrak{op}(y_i^l)\}$ is finite. One can also stack a stochastic operator (e.g. dropout), so long as the outputs are still channel-wisely i.i.d. and have finite second moments.

A.5.4 Simultaneous Limit

In this section, we present a sufficient condition on the activation function ϕ so that the neural networks converge to a Gaussian process as all the widths approach to infinity simultaneously. Precisely, let $t\in\mathbb{N}$ and for each $l\ge 0$, let $n^l:\mathbb{N}\to\mathbb{N}$ be the width function at layer l (by convention $n^0(t)=n^0$). We are interested in the simultaneous limit $n^l = n^l(t)\to\infty$ as $t\to\infty$, i.e., for any fixed $L\ge 1$

$$\min\left\{n^1(t),\ldots,n^L(t)\right\} \xrightarrow[t\to\infty]{} \infty. \tag{39}$$

Define a sequence of finite channel CNNs as follows:

$$y_{i,\alpha}^{l,t}(x) \equiv \begin{cases} x_{i,\alpha} & l=0 \\ \phi\left(z_{i,\alpha}^{l-1,t}(x)\right) & l>0 \end{cases}, \tag{40}$$

$$z_{i,\alpha}^{l,t}(x) \equiv \begin{cases} \frac{1}{\sqrt{n^0}} \sum_{j\in[n^0]} \frac{1}{\sqrt{2k+1}} \sum_{\beta\in[\pm k]} \omega_{ij,\beta}^{l} y_{j,\alpha+\beta}^{l,t}(x) + b_i^l\,, & l=0 \\ \frac{1}{\sqrt{n^l(t)}} \sum_{j\in[n^l(t)]} \frac{1}{\sqrt{2k+1}} \sum_{\beta\in[\pm k]} \omega_{ij,\beta}^{l} y_{j,\alpha+\beta}^{l,t}(x) + b_i^l\,, & l>0 \end{cases}. \tag{41, 42}$$

This network induces a sequence of covariances matrices K_t^l (which are R.V.s): for $l \geq 0$ and $t \geq 0$, for $x, x' \in \mathcal{X}$

$$\left[K_t^l\right]_{\alpha,\alpha'}(x,x') \equiv \frac{1}{n^l(t)} \sum_{i=1}^{n^l(t)} y_{i,\alpha}^{l,t}(x) y_{i,\alpha'}^{l,t}(x'). \tag{43}$$

We make an extra assumption on the parameters.

Assumption: all R.V.s in $\mathscr{W}$ are Gaussian distributed.

Notation. Let PSD_m denote the set of $m \times m$ positive semi-definite matrices and for $R \geq 1$, define

$$\mathrm{PSD}_m(R) \equiv \{\Sigma \in \mathrm{PSD}_m : 1/R \leq \Sigma_{\alpha,\alpha} \leq R \quad \text{for} \quad 1 \leq \alpha \leq m\}\,. \tag{44}$$

Further let $\mathcal{T}_\infty : \mathrm{PSD}_2 \to \mathbb{R}$ be a function given by

$$\mathcal{T}_\infty(\Sigma) \equiv \mathbb{E}_{(x,y)\sim\mathcal{N}(0,\Sigma)}\left[\phi(x)\phi(y)\right], \tag{45}$$

and $C_k(\phi, R)$ (may equal to ∞) denotes the uniform upper bound for the k-th moment

$$C_k(\phi, R) \equiv \sup_{1/R \leq r \leq R} \mathbb{E}_{x\sim\mathcal{N}(0,r)} \left|\phi(x)\right|^k. \tag{46}$$

Let Ψ denotes the space of measurable functions with the following properties:

1. **Uniformly square-integrable:** $C_2(\phi, R) < \infty$.
2. **Lipschitz continuity:** for every $R \geq 1$, there exists $\beta = \beta(\phi, R) > 0$ such that for all $\Sigma, \Sigma' \in \mathrm{PSD}_2(R)$,
$$\left|\mathcal{T}_\infty(\Sigma) - \mathcal{T}_\infty(\Sigma')\right| \leq \beta \left\|\Sigma - \Sigma'\right\|_\infty; \tag{47}$$
3. **Uniform convergence in probability:** for every $R \geq 1$ and every $\varepsilon > 0$ there exists a positive sequence $\rho_n(\phi, \varepsilon, R)$ with $\rho_n(\phi, \varepsilon, R) \to 0$ as $n \to \infty$ such that for every $\Sigma \in \mathrm{PSD}_2(R)$ and any $\{(x_i, y_i)\}_{i=1}^n$ i.i.d. $\sim \mathcal{N}(0, \Sigma)$
$$P\left(\left|\frac{1}{n}\sum_{i=1}^n \phi(x_i)\phi(y_i) - \mathcal{T}_\infty(\Sigma)\right| > \varepsilon\right) \leq \rho_n(\phi, \varepsilon, R). \tag{48}$$

We will also use Ψ_1, Ψ_2 and Ψ_3 to denote the spaces of functions satisfying property 1, property 2 and property 3, respectively. It is not difficult to see that for every i, Ψ_i is a vector space, and so is $\Psi = \cap_i \Psi_i$.

Finally, we say that a function $f : \mathbb{R} \to \mathbb{R}$ is exponentially bounded if there exist $a, b > 0$ such that

$$|f(x)| \leq a e^{b|x|} \quad \text{a.e. (almost everywhere)} \tag{49}$$

We now prove our main result presented in §2.2.3 through the following three theorems.

Theorem A.3. If ϕ' is exponentially bounded then $\phi \in \Psi$.

Theorem A.4. If $\phi \in \Psi$, then for $l \geq 0$, $K_t^l \xrightarrow{P} K_\infty^l$.

Theorem A.5. If for $l \geq 0$, $K_t^l \xrightarrow{P} K_\infty^l$ and $m \geq 1$, the joint distribution of $\left[z_j^{l,t}\right]_{j\in[m]}$ converges in distribution to a multivariate normal distribution with mean zero and covariance $\mathcal{A}\left(K_\infty^l\right) \otimes I_m$.

The proofs of Theorems A.3, A.4, and A.5 can be found in §A.5.7, §A.5.6, and §A.5.5 respectively.

A.5.5 PROOF OF THEOREM A.5

It suffices to prove that for any vector $[a_i(x)]_{i\in[m],\, x\in\mathcal{X}} \in \mathbb{R}^{m|\mathcal{X}|}$,

$$\sum_{\substack{i\in[m]\\ x\in\mathcal{X}}} a_i(x) z_i^{l,t}(x) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sum_{i\in[m],\, x\in\mathcal{X}} a_i(x)^T \mathcal{A}\left(K_\infty^l\right) a_i(x)\right). \tag{50}$$

Indeed, the characteristic function

$$\mathbb{E}\exp\left(\mathbb{i} \sum_{i\in[m],\, x\in\mathcal{X}} a_i(x) z_i^{l,t}(x)\right) \tag{51}$$

$$=\mathbb{E}\exp\left(\mathbb{i} \sum_{\substack{i\in[m]\\ x\in\mathcal{X}}} a_i(x) \left(\frac{1}{\sqrt{n^l(t)}} \sum_{j\in[n^l(t)]} \frac{1}{\sqrt{2k+1}} \sum_{\beta\in[\pm k]} \omega_{ij,\beta}^l y_{j,\alpha+\beta}^{l,t}(x) + b_i^l\right)\right). \tag{52}$$

Note that conditioned on $\left\{y_j^{l,t}\right\}_{j\in[n^l(t)]}$, the exponent in the above expression is just a linear combination of independent Gaussian R.V.s $\left\{\omega_{ij,\beta}^l,\, b_i^l\right\}$, which is also a Gaussian. Integrating out these R.V.s using the formula of the characteristic function of Gaussian distribution yields

$$\mathbb{E}\exp\left(\mathbb{i} \sum_{i\in[m],\, x\in\mathcal{X}} a_i(x) z_i^{l,t}(x)\right) \tag{53}$$

$$= \mathbb{E}\exp\left(-\frac{1}{2} \sum_{i\in[m]} a_i^T \left(\mathcal{A}\left(K_t^l\right)\right) a_i\right) \tag{54}$$

$$\longrightarrow \exp\left(-\frac{1}{2} \sum_{i\in[m]} a_i^T \left(\mathcal{A}\left(K_\infty^l\right)\right) a_i\right) \quad \text{as} \quad t\to\infty\,, \tag{55}$$

where have used $K_t^l \xrightarrow{P} K_\infty^l$ in the last step. Therefore, Equation 50 is true by Theorem A.1 (iii).

□

Remark A.2. We briefly comment how to handle the cases when stacking an average pooling, a subsampling or a dense layer after flattening the activations in the last layer.

(i) **Global Average pooling / subsampling.** Let $B \in \mathbb{R}^{1\times d}$ be any deterministic *linear* functional defined on $\mathbb{R}^d$. The fact that

$$K_t^l \xrightarrow{P} K_\infty^l \tag{56}$$

implies that the empirical covariance of $\left\{B y_j^{l,t}\right\}$

$$\frac{1}{n^l(t)} \sum_{j\in n^l(t)} B^{\otimes|\mathcal{X}|} y_j^{l,t} \left(B^{\otimes|\mathcal{X}|} y_j^{l,t}\right)^T \xrightarrow{P} B^{\otimes|\mathcal{X}|} K_\infty^l \left(B^{\otimes|\mathcal{X}|}\right)^T \tag{57}$$

where $B^{\otimes|\mathcal{X}|} \in \mathbb{R}^{1\times d|\mathcal{X}|}$, $|\mathcal{X}|$ copies of B. Invoking the same "characteristic function" arguments as above, it is not difficult to show that stacking a dense layer (assuming the weights and biases are drawn from i.i.d. Gaussian with mean zero and variances σ_ω^2 and σ_b^2, and are properly normalized) on top of $\left\{B y_j^{l,t}\right\}$ the outputs are i.i.d. Gaussian with mean zero and covariance $\sigma_\omega^2 B^{\otimes|\mathcal{X}|} K_\infty^l \left(B^{\otimes|\mathcal{X}|}\right)^T + \sigma_b^2$.

Taking $B = \left(\frac{1}{d}, \ldots, \frac{1}{d}\right) \in \mathbb{R}^{1\times d}$ $\left(\text{or } B = e_\alpha \in \mathbb{R}^{1\times d}\right)$ implies the result of global average pooling (§3.2.1, Equation 17), or subsampling (§3.2.2, Equation 18).

(ii) **Vectorization and a dense layer.** Let $\left\{\omega^l_{ij,\alpha}\right\}_{i\in[m],j\in[n^l(t)],\alpha\in[d]}$ be the weights of the dense layer, $\omega^l_{ij,\alpha}$ represents the weight connecting the α-th pixel of the j-channel to the i-th output. Note that the range of α is $[d]$ not $[\pm k]$ because there is no weight sharing. Define the outputs to be

$$f_i(x) = \frac{1}{\sqrt{dn^l(t)}} \sum_{\alpha\in[d]} \sum_{j\in[n^l(t)]} \omega^l_{ij,\alpha} y^{l,t}_{j,\alpha}(x) + b^l_i \tag{58}$$

Now let $[a_i(x)]_{i\in[m],x\in\mathcal{X}} \in \mathbb{R}^{m|\mathcal{X}|}$ and compute the characteristic function of

$$\sum_{i,x} a_i(x) f_i(x). \tag{59}$$

Using the fact $\mathbb{E}\omega_{ij,\alpha}\omega_{i'j',\alpha'} = 0$ unless $(ij,\alpha) = (i'j',\alpha')$ and integrating out the R.V.s of the dense layer, the characteristic function is equal to

$$\mathbb{E}\exp\left(-\frac{1}{2}\sum_{i\in[m]}\sum_{x,x'\in\mathcal{X}} a_i(x)a_i(x')\left(\frac{1}{dn^l(t)}\sum_{\substack{j\in[n^l(t)]\\ \alpha\in[d]}} \sigma^2_\omega y^{l,t}_{j,\alpha}(x) y^{l,t}_{j,\alpha}(x') + \sigma^2_b\right)\right) \tag{60}$$

$$= \mathbb{E}\exp\left(-\frac{1}{2}\sum_{i\in[m]}\sum_{x,x'\in\mathcal{X}} a_i(x)a_i(x')\bar{\text{tr}}\left(\sigma^2_\omega K^l_t(x,x') + \sigma^2_b\right)\right) \tag{61}$$

$$\longrightarrow \exp\left(-\frac{1}{2}\sum_{i\in[m]}\sum_{x,x'\in\mathcal{X}} a_i(x)a_i(x')\bar{\text{tr}}\left(\sigma^2_\omega K^l_\infty(x,x') + \sigma^2_b\right)\right), \tag{62}$$

where $\bar{\text{tr}}$ denotes the mean trace operator acting on the pixel by pixel matrix, i.e. the functional computing the mean of the diagonal terms of the pixel by pixel matrix. Therefore $[f_i]_{i\in[m]}$ converges weakly to a mean zero Gaussian with covariance $\left[\sigma^2_\omega \bar{\text{tr}}\left(K^l_\infty(x,x')\right) + \sigma^2_b\right]_{x,x'\in\mathcal{X}} \otimes I_m$ in the case of vectorization (§3.1).

A.5.6 Proof of Theorem A.4

Proof. We recall K^L_t and K^L_∞ to be random matrices in $\mathbb{R}^{d|\mathcal{X}|\times d|\mathcal{X}|}$, and we will prove convergence $K^L_t \xrightarrow[t\to\infty]{P} K^L_\infty$ with respect to $\|\cdot\|_\infty$, the pointwise ℓ^∞-norm (i.e. $\|K\|_\infty = \max_{x,x',\alpha,\alpha'} |K_{\alpha,\alpha'}(x,x')|$). Note that due to finite dimensionality of K^L, convergence w.r.t. all other norms follows.

We first note that the affine transform $\mathcal{A}$ is σ^2_ω-Lipschitz and property 2 of Ψ implies that the $\mathcal{C}$ operator is β-Lipschitz. Indeed, if we consider

$$\Sigma \equiv \begin{pmatrix} [K]_{\alpha,\alpha}(x,x) & [K]_{\alpha,\alpha'}(x,x') \\ [K]_{\alpha',\alpha}(x',x) & [K]_{\alpha',\alpha'}(x',x') \end{pmatrix}, \tag{63}$$

then $[\mathcal{C}(K)]_{\alpha,\alpha'}(x,x') = \mathcal{T}_\infty(\Sigma)$. Thus $\mathcal{C}\circ\mathcal{A}$ is $\sigma^2_\omega\beta$-Lipschitz.

We now prove the theorem by induction. Assume $K^l_t \xrightarrow{P} K^l_\infty$ as $t\to\infty$ (obvious for $l=0$).

Let $\varepsilon > 0$ be sufficiently small so that the $\frac{\varepsilon}{2\beta}$-neighborhood of $\mathcal{A}(K^l_\infty)$ is contained in $\text{PSD}_{|\mathcal{X}|d}(R)$, where we take R to be large enough for K^l_∞ to be an interior point of $\text{PSD}_{|\mathcal{X}|d}(R)$.

Since

$$\left\|K^{l+1}_\infty - K^{l+1}_t\right\|_\infty \le \left\|K^{l+1}_\infty - \mathcal{C}\circ\mathcal{A}\left(K^l_t\right)\right\|_\infty + \left\|\mathcal{C}\circ\mathcal{A}\left(K^l\right) - K^{l+1}_t\right\|_\infty \tag{64}$$

$$= \left\|\mathcal{C}\circ\mathcal{A}\left(K^l_\infty\right) - \mathcal{C}\circ\mathcal{A}\left(K^l_t\right)\right\|_\infty + \left\|\mathcal{C}\circ\mathcal{A}\left(K^l_t\right) - K^{l+1}_t\right\|_\infty, \tag{65}$$

to prove $K^{l+1}_t \xrightarrow{P} K^{l+1}_\infty$, it suffices to show that for every $\delta > 0$, there is a t^* such that for all $t > t^*$,

$$P\left(\left\|\mathcal{C}\circ\mathcal{A}\left(K^l_\infty\right) - \mathcal{C}\circ\mathcal{A}\left(K^l_t\right)\right\|_\infty > \frac{\varepsilon}{2}\right) + P\left(\left\|\mathcal{C}\circ\mathcal{A}\left(K^l_t\right) - K^{l+1}_t\right\|_\infty > \frac{\varepsilon}{2}\right) < \delta. \tag{66}$$

By our induction assumption, there is a t^l such that for all $t > t^l$

$$P\left(\left\|K_\infty^l - K_t^l\right\|_\infty > \frac{\varepsilon}{2\sigma_\omega^2\beta}\right) < \frac{\delta}{3}. \tag{67}$$

Since $\mathcal{C} \circ \mathcal{A}$ is $\sigma_\omega^2\beta$-Lipschitz, then

$$P\left(\left\|\mathcal{C} \circ \mathcal{A}\left(K_\infty^l\right) - \mathcal{C} \circ \mathcal{A}\left(K_t^l\right)\right\|_\infty > \frac{\varepsilon}{2}\right) < \frac{\delta}{3}. \tag{68}$$

To bound the second term in Equation 66, let $U(t)$ denotes the event

$$U(t) \equiv \left\{\mathcal{A}\left(K_t^l\right) \in \mathrm{PSD}_{|\mathcal{X}|d}(R)\right\} \tag{69}$$

and $U(t)^c$ its complement. For all $t > t^l$ it's probability is

$$\begin{aligned} P\left(U(t)^c\right) &< P\left(\left\|\mathcal{A}\left(K_\infty^l\right) - \mathcal{A}\left(K_t^l\right)\right\|_\infty > \frac{\varepsilon}{2\beta}\right) && \text{[assumption on small } \varepsilon] && (70)\\ &< P\left(\sigma_\omega^2\left\|K_\infty^l - K_t^l\right\|_\infty > \frac{\varepsilon}{2\beta}\right) && [\mathcal{A} \text{ is } \sigma_\omega^2\text{-Lipshitz}] && (71)\\ &= P\left(\left\|K_\infty^l - K_t^l\right\|_\infty > \frac{\varepsilon}{2\sigma_\omega^2\beta}\right) < \frac{\delta}{3}. && \text{[Equation 67]} && (72) \end{aligned}$$

Finally, denote

$$\left[V(t)\right]_{\alpha,\alpha'}(x,x') \equiv \left\{\left|\left[\mathcal{C} \circ \mathcal{A}\left(K_t^l\right)\right]_{\alpha,\alpha'}(x,x') - \left[K_t^{l+1}\right]_{\alpha,\alpha'}(x,x')\right| > \frac{\varepsilon}{2}\right\}. \tag{73}$$

The fact

$$\left\{\left\|\mathcal{C} \circ \mathcal{A}\left(K_t^l\right) - K_t^{l+1}\right\|_\infty > \frac{\varepsilon}{2}\right\} \subseteq U(t)^c \bigcup \left(\bigcup_{x,x',\alpha,\alpha'} \left[V(t)\right]_{\alpha,\alpha'}(x,x') \bigcap U(t)\right)$$

implies

$$P\left(\left\{\left\|\mathcal{C} \circ \mathcal{A}\left(K_t^l\right) - K^{l+1}(t)\right\|_\infty > \frac{\varepsilon}{2}\right\}\right) \leq \frac{\delta}{3} + |\mathcal{X}|^2 d^2 \max_{x,x',\alpha,\alpha'} P\left(\left[V(t)\right]_{\alpha,\alpha'}(x,x') \cap U(t)\right), \tag{74}$$

Where the maximum is taken over all $(x, x', \alpha, \alpha') \in \mathcal{X}^2 \times [d]$.

Consider a fixed $\kappa \in \mathrm{PSD}_{|\mathcal{X}|d}$, and define

$$\Sigma(\kappa,\alpha,\alpha',x,x') \equiv \begin{pmatrix} \left[\mathcal{A}(\kappa)\right]_{\alpha,\alpha}(x,x) & \left[\mathcal{A}(\kappa)\right]_{\alpha,\alpha'}(x,x') \\ \left[\mathcal{A}(\kappa)\right]_{\alpha',\alpha}(x',x) & \left[\mathcal{A}(\kappa)\right]_{\alpha',\alpha'}(x',x') \end{pmatrix}, \tag{75}$$

a deterministic matrix in PSD_2. Then

$$\left[\mathcal{C} \circ \mathcal{A}(\kappa)\right]_{\alpha,\alpha'}(x,x') = \mathcal{T}_\infty\left(\Sigma(\kappa,\alpha,\alpha',x,x')\right), \tag{76}$$

and, conditioned on K_t^l,

$$\left[K_t^{l+1} | K_t^l = \kappa\right]_{\alpha,\alpha'}(x,x') = \frac{1}{n^{l+1}(t)} \sum_{i=1}^{n^{l+1}(t)} \phi\left(z_{i,\alpha}^{l,t}(x)\right)\phi\left(z_{i,\alpha'}^{l,t}(x')\right), \tag{77}$$

where $\left\{\left(z_{i,\alpha}^{l,t}(x), z_{i,\alpha'}^{l,t}(x')\right) | K_t^l = \kappa\right\}_{i\in[n^{l+1}(t)]}$ are i.i.d. $\sim \mathcal{N}\left(0, \Sigma(\kappa,\alpha,\alpha',x,x')\right)$.

Then if $\Sigma(\kappa,\alpha,\alpha',x,x') \in \mathrm{PSD}_2(R)$ we can apply property 3 of Ψ to conclude that:

$$P\left(\left(\left[V(t)\right]_{\alpha,\alpha'}(x,x') \cap U(t)\right) \middle| K_t^l = \kappa\right) < \rho_{n^{l+1}(t)}\left(\phi, R, \frac{\varepsilon}{2}\right). \tag{78}$$

However, if $\Sigma(\kappa,\alpha,\alpha',x,x') \notin \mathrm{PSD}_2(R)$, then necessarily $\mathcal{A}(\kappa) \notin \mathrm{PSD}_{|\mathcal{X}|d}(R)$ (since $\Sigma(\kappa,\alpha,\alpha',x,x') \in \mathrm{PSD}_2$), ensuring that $P\left(U(t) | K_t^l = \kappa\right) = 0$. Therefore Equation 78 holds for any $\kappa \in \mathrm{PSD}_{|\mathcal{X}|d}$, and for any $(x, x', \alpha, \alpha') \in \mathcal{X}^2 \times [d]^2$.

We further remark that $\rho_{n^{l+1}(t)}\left(\phi, R, \frac{\varepsilon}{2}\right)$ is deterministic and does not depend on $(\kappa, x, x', \alpha, \alpha')$. Marginalizing out K_t^l and maximizing over (x, x', α, α') in Equation 78 we conclude that

$$\max_{x,x',\alpha,\alpha'} P\left([V(t)]_{\alpha,\alpha'}(x,x') \cap U(t)\right) < \rho_{n^{l+1}(t)}\left(\phi, R, \frac{\varepsilon}{2}\right). \tag{79}$$

Since $\rho_n\left(\phi, R, \frac{\varepsilon}{2}\right) \to 0$ as $n \to \infty$, there exists n such that for any $n^{l+1}(t) \geq n$,

$$\max_{x,x',\alpha,\alpha'} P\left([V(t)]_{\alpha,\alpha'}(x,x') \cap U(t)\right) < \rho_{n^{l+1}(t)}\left(\phi, R, \frac{\varepsilon}{2}\right) \leq \frac{\delta}{3|\mathcal{X}|^2 d^2}, \tag{80}$$

and, substituting this bound in Equation 74,

$$P\left(\left\{\left\|\mathcal{C} \circ \mathcal{A}\left(K_t^l\right) - K_t^{l+1}\right\|_\infty > \frac{\varepsilon}{2}\right\} \cap U(t)\right) < \frac{2\delta}{3}. \tag{81}$$

Therefore we just need to choose $t^{l+1} > t^l$ so that $n^{l+1}(t) \geq n$ for all $t > t^{l+1}$.

□

Remark A.3. We list some directions to strengthen / extent the results of Theorem A.4 (and thus Theorem A.5) using the above framework.

1. Consider stacking a deterministic channel-wise linear operator right after the convolutional layer. Again, strided convolution, convolution with no ("valid") padding and (non-global) average pooling are particular examples of this category. Let $B \in \mathbb{R}^{d' \times d}$ denote a linear operator. Then the recurrent formula between two consecutive layers is

$$K_\infty^{l+1} = \mathcal{C} \circ \mathcal{B} \circ \mathcal{A}(K_\infty^l) \tag{82}$$

   where $\mathcal{B}$ is the linear operator on the covariance matrix induced by B. Conditioned on K_t^l, since the outputs after applying the linear operator B are still i.i.d. Gaussian (and the property 3 is applicable), the analysis in the above proof can carry over with $\mathcal{A}$ replaced by $\mathcal{B} \circ \mathcal{A}$.

2. More generally, one may consider inserting an operator $\mathfrak{op}$ (e.g. max-pooling, dropout and more interestingly, normalization) in some hidden layer.

3. It is possible to drop the Gaussianity assumption on the weight parameters by incorporating the proof of the central limit theorem using characteristic function into the above arguments.

A.5.7 Proof of Theorem A.3

Note that exponentially bounded functions contain all polynomials, and are closed under multiplication and integration in the sense that for any constant C the function

$$\int_0^x \phi(t)dt + C \tag{83}$$

is also exponentially bounded. Theorem A.3 is a consequence of the following lemma.

Lemma A.6. The following is true:

1. for $k \geq 1$, $C_k(\phi, R) < \infty$ if ϕ is exponentially bounded.
2. $\phi \in \Psi_2$ if ϕ' exists a.e. and is exponentially bounded.
3. $\phi \in \Psi_3$ if $C_4(\phi, R) < \infty$.

Indeed, if ϕ' is exponentially bounded, then ϕ is also exponentially bounded. By the above lemma, $\phi \in \Psi$.

Proof of Lemma A.6. **1.** [RN: TODO:check] We prove the first statement. Assume $|\phi(x)| \leq ae^{b|x|}$.

$$\mathbb{E}_{x\sim\mathcal{N}(0,r)} |\phi(x)|^k = \mathbb{E}_{x\sim\mathcal{N}(0,1)} \left|\phi\left(\sqrt{r}x\right)\right|^k \leq \mathbb{E}_{x\sim\mathcal{N}(0,1)} \left|ae^{\sqrt{r}b|x|}\right|^k \leq 2a^k e^{k^2b^2r/2}. \tag{84}$$

Thus

$$C_k(\phi, R) = \sup_{1/R\leq r\leq R} \mathbb{E}_{x\sim\mathcal{N}(0,r)} |\phi(x)|^k \leq 2a^k e^{k^2b^2R/2}. \tag{85}$$

2. To prove the second statement, let $\Sigma, \Sigma' \in \mathrm{PSD}_2(R)$ and define A (similarly for A'):

$$A \equiv \begin{pmatrix} \sqrt{\Sigma_{11}} & 0 \\ \frac{\Sigma_{12}}{\sqrt{\Sigma_{11}}} & \sqrt{\frac{\Sigma_{22}\Sigma_{11}-\Sigma_{12}^2}{\Sigma_{11}}} \end{pmatrix}. \tag{86}$$

Then $AA^T = \Sigma$ (and $A'A'^T = \Sigma'$). Let

$$A(t) \equiv (1-t)A + tA', \quad t \in [0,1] \tag{87}$$

and

$$f(w) \equiv \phi(x)\phi(y) \quad \text{where} \quad w \equiv (x,y)^T. \tag{88}$$

Since ϕ' is exponentially bounded, ϕ is also exponentially bounded. In addition, $p(\|w\|_2)\|\nabla f(w)\|_2$ is exponentially bounded for any polynomial $p(\|w\|_2)$.

Applying the Mean Value Theorem (we use the notation $\lesssim$ to hide the dependence on R and other absolute constants)

$$|\mathcal{T}_\infty(\Sigma) - \mathcal{T}_\infty(\Sigma')| = \frac{1}{2\pi}\left|\int (f(Aw) - f(A'w)) \exp\left(-\|w\|_2^2/2\right) dw\right| \tag{89}$$

$$= \frac{1}{2\pi}\left|\int\int_{[0,1]} (\nabla f(A(t)w))((A'-A)w) \exp\left(-\|w\|_2^2/2\right) dtdw\right| \tag{90}$$

$$\lesssim \int_{[0,1]}\int \|(A'-A)w\|_2 \|\nabla f(A(t)w)\|_2 \exp\left(-\|w\|_2^2/2\right) dwdt \tag{91}$$

$$\leq \int_{[0,1]}\int \|A'-A\|_{\mathrm{op}} \|w\|_2 \|\nabla f(A(t)w)\|_2 \exp\left(-\|w\|_2^2/2\right) dwdt. \tag{92}$$

Note that the operator norm is bounded by the infinity norm (up to a multiplicity constant) and $\|w\|_2 \|\nabla f(A(t)w)\|_2$ is exponentially bounded. There is a constant a (hidden in $\lesssim$) and b such that the above is bounded by

$$\int_{[0,1]}\int \|A'-A\|_\infty \exp(b\|A(t)\|_\infty \|w\|_2) \exp\left(-\|w\|_2^2/2\right) dwdt \tag{93}$$

$$\lesssim \|A'-A\|_\infty \int_{[0,1]}\int \exp\left(b\sqrt{R}\|w\|_2 - \|w\|_2^2/2\right) dwdt \tag{94}$$

$$\lesssim \|A'-A\|_\infty \tag{95}$$

$$\lesssim \|\Sigma'-\Sigma\|_\infty. \tag{96}$$

Here we have applied the facts

$$\|A'-A\|_\infty \lesssim \|\Sigma-\Sigma'\|_\infty \quad \text{and} \quad \|A(t)\|_\infty \leq \sqrt{R}. \tag{97}$$

3. Chebyshev's inequality implies

$$P\left(\left|\frac{1}{n}\sum_{i=1}^n \phi(x_i)\phi(y_i) - \mathcal{T}_\infty(\Sigma)\right| > \varepsilon\right) \tag{98}$$

$$\leq \frac{1}{n\epsilon^2}\mathrm{Var}(\phi(x_i)\phi(y_i)) \leq \frac{1}{n\epsilon^2}\mathbb{E}|\phi(x_i)\phi(y_i)|^2 \tag{99}$$

$$\leq \frac{1}{n\epsilon^2}C_4(\phi, R) \to 0 \quad \text{as} \quad n\to\infty. \tag{100}$$

□

Remark A.4. In practice the $1/n$ decay bound obtained by Chebyshev's inequality in equation 100 is often too weak to be useful. However, if ϕ is linearly bounded, then one can obtain an exponential decay bound via the following concentration inequality:

Lemma A.7. If $|\phi(x)| \leq a + b|x|$ a.e., then there is an absolute constant $c > 0$ and a constant $\kappa = \kappa(a, b, R) > 0$ such that property 3 (Equation 48) holds with

$$\rho_n(\phi, \epsilon, R) = 2 \exp\left(-c \min\left\{\frac{n^2\varepsilon^2}{(2\kappa)^2}, \frac{n\varepsilon}{2\kappa}\right\}\right). \tag{101}$$

Proof. We postpone the proof of the following claim.

Claim A.8. Assume $|\phi(x)| \leq a + b|x|$. Then there is a $\kappa = \kappa(a, b, R)$ such that for all $\Sigma \in \mathrm{PSD}_2(R)$ and all $p \geq 1$,

$$\left(\mathbb{E}_{(x,y)\sim\mathcal{N}(0,\Sigma)}|\phi(x)\phi(y)|^p\right)^{1/p} \leq \kappa p. \tag{102}$$

Claim A.8 and the triangle inequality imply

$$\left(\mathbb{E}_{(x,y)\sim\mathcal{N}(0,\Sigma)}|\phi(x)\phi(y) - \mathbb{E}\phi(x)\phi(y)|^p\right)^{1/p} \leq 2\kappa p. \tag{103}$$

We can apply Bernstein-type inequality (Vershynin, 2010, Lemma 5.16) to conclude that there is a $c > 0$ such that for every $\Sigma \in \mathrm{PSD}_2(R)$ and any $\{(x_i, y_i)\}_{i=1}^n$ i.i.d. $\sim \mathcal{N}(0, \Sigma)$

$$P\left(\left|\frac{1}{n}\sum_{i=1}^n \phi(x_i)\phi(y_i) - \mathcal{T}_\infty(\Sigma)\right| > \varepsilon\right) \leq 2\exp\left(-c\min\left\{\frac{n^2\varepsilon^2}{(2\kappa)^2}, \frac{n\varepsilon}{2\kappa}\right\}\right). \tag{104}$$

It remains to prove Claim A.8. For $p \geq 1$,

$$\begin{aligned}
\left(\mathbb{E}_{(x,y)\sim\mathcal{N}(0,\Sigma)}|\phi(x)\phi(y)|^p\right)^{1/p} &\leq \left(\mathbb{E}_{x\sim\mathcal{N}(0,\Sigma_{11})}|\phi(x)|^{2p}\right)^{1/2p}\left(\mathbb{E}_{y\sim\mathcal{N}(0,\Sigma_{22})}|\phi(y)|^{2p}\right)^{1/2p} && (105)\\
&\leq \left(a + b\left(\mathbb{E}|x|^{2p}\right)^{1/2p}\right)\left(a + b\left(\mathbb{E}|y|^{2p}\right)^{1/2p}\right) && (106)\\
&\leq \left(a + b\sqrt{R}\left(\mathbb{E}_{u\sim\mathcal{N}(0,1)}|u|^{2p}\right)^{1/2p}\right)^2 && (107)\\
&\leq \left(a + b\sqrt{R}\left(c'^{2p}p^p\right)^{1/2p}\right)^2 && (108)\\
&\leq \left(a + bc'^2\sqrt{R}\right)^2 p && (109)\\
&\equiv \kappa p. && (110)
\end{aligned}$$

We applied Cauchy-Schwarz' inequality in the first inequality, the triangle inequality in the second one, the fact $\Sigma_{11}, \Sigma_{22} \leq R$ in the third one, absolute moments estimate of standard Gaussian in the fourth one, where c' is a constant such that

$$\left(\mathbb{E}_{u\sim\mathcal{N}(0,1)}|u|^p\right)^{1/p} \leq c'\sqrt{p}. \tag{111}$$

□

A.6 Glossary

We use the following shorthands in this work:

1. NN - neural network;
2. CNN - convolutional neural network;
3. LCN - locally-connected network, a.k.a. convolutional network without weight sharing;
4. FCN - fully connected network, a.k.a. multilayer perceptron (MLP);
5. GP - Gaussian process;
6. X-GP - a GP equivalent to a Bayesian infinitely wide neural network of architecture X (§2).
7. MC-(X-)-GP - a Monte Carlo estimate (§4) of the X-GP.

8. Width, (number of) filters, (number of) channels represent the same property for CNNs and LCNs.
9. Pooling - referring to architectures as "with" or "without pooling" means having a single global average pooling layer (collapsing the spatial dimensions of the activations y^{L+1}) before the final linear FC layer giving the regression outputs z^{L+1}.
10. Invariance and equivariance are always discussed w.r.t. translations in the spatial dimensions of the inputs.

A.7 Experimental Setup

Throughout this work we only consider 3×3 (possibly unshared) convolutional filters with stride 1 and no dilation.

All inputs are normalized to have zero mean and unit variance, i.e. lie on the $d-$dimensional sphere of radius $\sqrt{d}$, where d is the total dimensionality of the input.

All labels are treated as regression targets with zero mean. I.e. for a single-class classification problem with C classes targets are $C-$dimensional vectors with $-1/C$ and $(C-1)/C$ entries in incorrect and correct class indices respectively.

If a subset of a full dataset is considered for computational reasons, it is randomly selected in a balanced fashion. No data augmentation is used.

All experiments were implemented in Tensorflow (Abadi et al., 2016) and executed with the help of Vizier (Golovin et al., 2017).

All neural networks are trained using Adam (Kingma & Ba, 2015) minimizing the mean squared error loss.

A.7.1 Many-channel CNNs and LCNs

Relevant Figures: 3 (b), 5, 6.

We use a training and validation subsets of CIFAR10 of sizes 500 and 4000 respectively. All images are bilinearly downsampled to 8×8 pixels.

All models have 3 hidden layers with an erf nonlinearity. No ("valid") padding is used.

Weight and bias variances are set to $\sigma_\omega^2 \approx 1.7562$ and $\sigma_b^2 \approx 0.1841$, corresponding to the pre-activation variance fixed point $q^* = 1$ (Poole et al., 2016) for the erf nonlinearity.

NN training proceeds for 2^{19} gradient updates, but aborts if no progress on training loss is observed for the last 100 epochs. If the training loss does not reduce by at least 10^{-4} for 20 epochs, the learning rate is divided by 10.

All computations are done with 32-bit precision.

The following NN parameters are considered[9]:

1. Architecture: CNN or LCN.
2. Pooling: no pooling or a single global average pooling (averaging over spatial dimensions) before the final FC layer.
3. Number of channels: 2^k for k from 0 to 12.
4. Initial learning rate: 10^{-k} for k from 0 to 15.
5. Weight decay: 0 and 10^{-k} for k from 0 to 8.
6. Batch size: 10, 25, 50, 100, 200.

For NNs, all models are filtered to only 100%-accurate ones on the training set and then for each configuration of $\{$architecture, pooling, number of channels$\}$ the model with the lowest validation loss is selected among the configurations of $\{$learning rate, weight decay, batch size$\}$.

[9]Due to time and memory limitations certain large configurations could not be evaluated. We believe this did not impact the results of this work in a qualitative way.

For GPs, the same CNN-GP is plotted against CNN and LCN networks without pooling. For LCN with pooling, inference was done with an appropriately rescaled CNN-GP kernel, i.e. $\left(\mathcal{K}_\infty^{\text{vec}} - \sigma_b^2\right)/d + \sigma_b^2$, where d is the spatial size of the penultimate layer. For CNNs with pooling, a Monte Carlo estimate was computed (see §4) with $n = 2^{12}$ filters and $M = 2^6$ samples.

For GP inference, the initial diagonal regularization term applied to the training convariance matrix is 10^{-10}; if the cholesky decompisition fails, the regularization term is increased by a factor of 10 until it either succeeeds or reaches the value of 10^5, at which point the trial is considered to have failed.

A.7.2 Monte Carlo Evaluation of Intractable GP Kernels

Relevant Figures: 2, 7.

We use the same setup as in §A.7.1, but training and validation sets of sizes 2000 and 4000 respectively.

For MC-GPs we consider the number of channels n (width in FCN setting) and number of NN instantiations M to accept values of 2^k for k from 0 to 10.

Kernel distance is computed as:

$$\frac{\|\mathcal{K}_\infty - K_{n,M}\|_{\text{F}}^2}{\|\mathcal{K}_\infty\|_{\text{F}}^2}, \tag{112}$$

where $\mathcal{K}_\infty$ is substituted with $K_{2^{10},2^{10}}$ for the CNN-GP pooling case (due to impracticality of computing the exact $\mathcal{K}_\infty^{\text{pool}}$). GPs are regularized in the same fashion as in §A.7.1, but the regularization factor starts at 10^{-4} and ends at 10^{10} and is multiplied by the mean of the training covariance diagonal.

A.7.3 Transforming a GP over spatial locations into a GP over classes

Relevant Figure: 1.

We use the same setup as in §A.7.2, but rescale the input images to size of 31×31, so that at depth 15 the spatial dimension collapses to a 1×1 patch if no padding is used (hence the curve of the CNN-GP without padding halting at that depth).

For MC-CNN-GP with pooling, we use samples of networks with $n = 16$ filters. Due to computational complexity we only consider depths up to 31 for this architecture. The number of samples M was selected independently for each depth among $\left\{2^k\right\}$ for k from 0 to 15 to maximize the validation accuracy on a separate 500-points validation set. This allowed us to avoid the poor conditioning of the kernel. GPs are regularized in the same fashion as in §A.7.1, but for MLP-GP the multiplicative factor starts at 10^{-4} and ends at 10^{10}.

A.7.4 Relationship to Deep Signal Propagation

Relevant Table: 3.

We use a training and validation subsets of CIFAR10 of sizes 500 and 1000 respectively.

We use the erf nonlinearity. For CNN-GP, images are zero-padded ("same" padding) to maintain the spatial shape of the activations as they are propagated through the network.

Weight and bias variances (horizontal axis σ_ω^2 and vertical axis σ_b^2 respectively) are sampled from a uniform grid of size 50×50 on the range $[0.1, 5] \times [0, 2]$ including the endpoints.

All computations are done with 64-bit precision. GPs are regularized in the same fashion as in §A.7.1, but the regularization factor is multiplied by the mean of the training covariance diagonal. If the experiment fails due to numerical reasons, 0.1 (random chance) validation accuracy is reported.

A.7.5 CNN-GP on full datasets

Relevant Table: 2, Figure 3 (a, c).

We use full training, validation, and test sets of sizes 50000, 10000, and 10000 respectively for MNIST (LeCun et al., 1998) and Fashion-MNIST (Xiao et al., 2017a), 45000, 5000, and 10000 for CIFAR10 (Krizhevsky, 2009). We use validation accuracy to select the best configuration for each model (we do not retrain on valdiation sets).

GPs are computed with 64-bit precision, and NNs are trained with 32-bit precision. GPs are regularized in the same fashion as in §A.7.4.

Zero-padding ("same") is used.

The following parameters are considered:

1. Architecture: CNN or FCN.
2. Nonlinearity: erf or ReLU.
3. Depth: 2^k for k from 0 to 4 (and up to 2^5 for MNIST and Fashion-MNIST datasets).
4. Weight and bias variances. For erf: q^* from $\{0.1, 1, 2, \ldots, 8\}$. For ReLU: a fixed weight variance $\sigma_\omega^2 = 2 + 4e^{-16}$ and bias variance σ_b^2 from $\{0.1, 1, 2, \ldots, 8\}$.

On CIFAR10, we additionally train NNs for 2^{18} gradient updates with a batch size of 128 with corresponding parameters in addition to[10]

1. Pooling: no pooling or a single global average pooling (averaging over spatial dimensions) before the final FC layer (only for CNNs).
2. Number of channels or width: 2^k for k from 1 to 9 (and up to 2^{10} for CNNs with pooling in Figure 3, a).
3. Learning rate: $10^{-k} \times 2^{16} / (\text{width} \times q^*)$ for k from 5 to 9, where width is substituted with the number of channels for CNNs and q^* is substituted with σ_b^2 for ReLU networks. "Small learning rate" in Table 2 refers to $k \in \{8, 9\}$.
4. Weight decay: 0 and 10^{-k} for k from 0 to 5.

For NNs, all models are filtered to only 100%-accurate ones on the training set (expect for values in parentheses in Table 2). The reported values are then reported for models that achieve the best validation accuracy.

A.7.6 Model comparison on CIFAR10

Relevant Table: 1.

We use the complete CIFAR10 dataset as described in §A.7.5 and consider 8-layer ReLU models with weight and bias variances of $\sigma_\omega^2 = 2$ and $\sigma_b^2 = 0.01$. The number of channels / width is set to 2^5, 2^{10} and 2^{12} for LCN, CNN, and FCN respectively.

GPs are computed with 64-bit precision, and NNs are trained with 32-bit precision.

No padding ("valid") is used.

NN training proceeds for 2^{18} gradient updates with batch size 64, but aborts if no progress on training loss is observed for the last 10 epochs. If the training loss does not reduce by at least 10^{-4} for 2 epochs, the learning rate is divided by 10.

Values for NNs are reported for the best validation accuracy over different learning rates (10^{-k} for k from 2 to 12) and weight decay values (0 and 10^{-k} for k from 2 to 7). For GPs, validation accuracy is maximized over initial diagonal regularization terms applied to the training convariance matrix: $10^{-k} \times$ [mean of the diagonal] for k among 2, 4 and 9 (if the cholesky decompisition fails, the regularization term is increased by a factor of 10 until it succeeds or k reaches the value of 10).

[10] Due to time and compute limitations certain large configurations could not be evaluated. We believe this did not impact the results of this work in a qualitative way.